ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

"ANÁLISIS MULTIVARIANTE DE LOS SINIESTROS DE TRÁNSITO DEL CANTÓN LA JOYA DE LOS SACHAS, 2015-2020"

Trabajo de titulación

Tipo: Proyecto de investigación

Presentado para obtener el grado académico de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

AUTORA:

ANDREA ESTHEFANIA CARRIÓN ALVARADO

Riobamba – Ecuador

2022

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

“ANÁLISIS MULTIVARIANTE DE LOS SINIESTROS DE TRÁNSITO DEL CANTÓN LA JOYA DE LOS SACHAS, 2015-2020”

Trabajo de titulación

Tipo: Proyecto de investigación

Presentado para obtener el grado académico de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

AUTORA:

ANDREA ESTHEFANIA CARRIÓN ALVARADO

DIRECTORA:

ING. NANCY ELIZABETH CHARIGUAMÁN MAURISACA Mgs.

Riobamba – Ecuador

2022

Yo, ANDREA ESTHEFANIA CARRIÓN ALVARADO, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 13 de diciembre de 2022

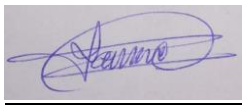
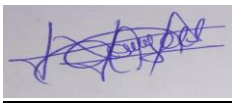
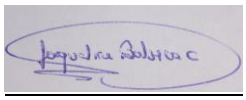
Andrea Esthefania Carrión Alvarado
220039362-3

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE ESTADÍSTICA

El Tribunal del Trabajo de Titulación, certifica que: El Trabajo de Titulación: Tipo: Proyecto de Investigación: **"ANÁLISIS MULTIVARIANTE DE LOS SINIESTROS DE TRÁNSITO DEL CANTÓN LA JOYA DE LOS SACHAS, 2015-2020"**, realizado por la señorita: **ANDREA ESTHEFANIA CARRIÓN ALVARADO**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación. El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	Firma	**Fecha**
Ing. Johanna Enith Aguilar Reyes Mgs. **PRESIDENTE DEL TRIBUNAL**		2022-12-13
Ing. Nancy Elizabeth Chariguamán Maurisaca Mgs. **DIRECTOR DEL TRABAJO DE TITULACION**		2022-12-13
Dra. Jaqueline Elizabeth Balseca Castro Mgs. **ASESORA DEL TRABAJO DE TITULACION**		2022-12-13

DEDICATORIA

Este trabajo va dedicado a mi padre Pedro Manuel Carrión Gaona, a mi madre Miriam Betty Alvarado Rojas, y a mis hermanos Bryan y Juan.

Andrea

AGRADECIMIENTO

Agradezco a mi familia, especialmente a mis padres por motivarme a continuar con mis estudios y apoyarme económicamente para culminar la carrera, además de ser la única razón por la que he resistido a lo largo de la carrera universitaria, estoy eternamente agradecida con ustedes por permitirme cumplir una meta familiar y entregarles un motivo más de felicidad.

De igual manera, agradezco a la Ing. Nancy Chariguamán por colaborarme como tutora en el presente trabajo, y a la Dra. Jaqueline Balseca por su ayuda como miembro para la elaboración de este trabajo, por su tiempo y consejos brindados.

Agradezco a la Agencia Nacional de Tránsito del Ecuador por brindar la información respectiva para el análisis.

Andrea

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE ECUACIONES

ÍNDICE DE ANEXOS

RESUMEN

El objetivo de esta investigación fue aplicar técnicas multivariantes a los siniestros de tránsito del cantón La Joya de los Sachas, 2015-2020, para identificar la relación y similitud entre ellos y que faciliten al GAD Municipal en la toma de decisiones; por lo que al realizar el análisis descriptivo y exploratorio de los datos se encontró que los siniestros de tránsito se desarrollaron con mayor frecuencia los días miércoles, viernes, sábados y domingos con un promedio de 14.28% de los cuales el 56.76% no reporto víctimas mortales a causa del siniestro de tránsito, utilizando la técnica de la detección de los datos atípicos se puede indicar que la numerosidad del colectivo cambia a 65 debido a que existieron 8 datos atípicos, luego se procedió a realizar el análisis de correspondencias múltiples utilizando las 5 variables cualitativas (feriado, zona, causa, clase, día, número de fallecidos y número de lesionados) para identificar las variables que tengan alta frecuencia y que indiquen por qué lo de los siniestros de tránsito en el cantón la Joya de los Sachas mismas que son: Clase de atropello y volcamiento, al realizar el análisis de conglomerados se tomaron en cuenta a las variables cuantitativas y para su debida agrupación se trabajó con el método de linkage completo. Finalmente se puede indicar que al aplicar la técnica del análisis de datos conglomerados o clusters se pueden identificar el número de clusters que se van a formar mediante un dendograma de tal manera que se forman 2 grupos, el primero conformado de 0 a 1 fallecido con un total de 1 a 3 lesionados, el segundo conformado por 0 a 1 fallecido y 0 a 1 lesionado por accidentes de tránsito e indicando que estos grupos están bien agrupados ya que se realizó la prueba de correlación cofenética.

Palabras clave: <TRÁNSITO>, <ANÁLISIS MULTIVARIANTE>, <ANÁLISIS DE CORRESPONDENCIAS> <ANÁLISIS DE CONGLOMERADOS>, < LA JOYA DE LOS SACHAS (CANTÓN)>

SUMMARY

The objective of this research was to apply multivariate techniques to traffic accidents in the La Joya de los Sachas canton, 2015-2020, to identify the relationship and similarity between them and to facilitate decision-making in the Municipal GAD; Therefore, when performing the descriptive and exploratory analysis of the data, it was found that traffic accidents occurred more frequently on Wednesdays, Fridays, Saturdays and Sundays with an average of 14.28% of which 56.76% did not report fatalities due to the traffic accident, using the technique of detecting atypical data, it can be indicated that the number of the group changes to 65 because there were 8 atypical data, then the multiple correspondence analysis was done using the 5 qualitative variables (holiday, zone, cause, class, day, number of deaths and number of injured) to identify the variables that have a high frequency and indicate the reason why the traffic accidents in the canton of La Joya de los Sachas which are as follows: running over and overturning, when performing the cluster analysis, the quantitative variables were taken into account and for their proper grouping, full linkage method was used. Finally, it can be indicated that by applying the conglomerate or cluster data analysis technique, the number of clusters that are going to be formed can be identified by means of a dendrogram in such a way that 2 groups were formed, the first one 0 to 1 deceased with a total of 1 to 3 injured, the second one 0 to 1 deceased and 0 to 1 injured by traffic accidents and indicating that these groups are well grouped since the cophenetic correlation test was performed.

Keywords: <TRAFFIC>, <MULTIVARIATE ANALYSIS>, <CORRESPONDENCE ANALYSIS> <CLUSTER ANALYSIS>, <LA JOYA DE LOS SACHAS (CANTON)>.

Edgar Mesias Jaramillo Moyano
0603497397

INTRODUCCIÓN

Con el aumento de la cantidad de automotores en general, los siniestros de tránsito se han convertido en un problema social importante y una grave amenaza para la salud y la vida de las personas. Según el Informe sobre la situación mundial de la seguridad vial 2015 de la Organización Mundial de la Salud, los traumatismos causados por el tránsito se consideran actualmente la novena causa principal de muerte (OMS, 2015, p. 8). Además, se prevé que el número total de muertes y lesiones en accidentes de tráfico aumente en aproximadamente un 65% hasta 2030 si no se hacen esfuerzos para mejorar la seguridad (Leone, et al., 2015, pp. 15-31) (Golman, et al., 2014, pp. 1-8).

Generalmente, la mayoría de los siniestros de tránsito son causados por el comportamiento inadecuado de los conductores. Las lesiones por siniestros de tránsito se encuentran entre las diez causas principales de muerte en todo el mundo y son la principal causa de muerte entre los adultos jóvenes de 15 a 29 años (WHO, 2013, p. 5). Además del trágico impacto en las vidas humanas, las lesiones por accidentes de tráfico tienen un efecto significativo en la economía mundial. Un estudio estima que los siniestros de tránsito costarán a la economía mundial cerca de US$1.8 billones de dólares en el período 2015-2030 (Chen, et al., 2019, p. 9).

La Unión Europea estima que el impacto representa el 2% de su PIB. Estos costos incluyen los costos de daños a la propiedad, costos de los seguros, costos de administración, costos de hospital, y las pérdidas de productividad, entre otros (Chen, et al., 2019, p. 9) (CDC, 2020, pp. 2-5) (García & Pérez, 2007, pp. 65-68).

En el mismo sentido, en el Ecuador los siniestros de tránsito es una problemática que afecta en la salud pública, económico, social. Actualmente, se ha evidenciado un alto número de accidentes lo que conlleva a afrontar altas cantidades de fallecidos y lesionados, tras estos sucesos. El número de accidentes de tránsito se mantiene elevado y tiende a aumentar por año. Debido a estos enormes impactos, los gobiernos y el sector privado realizan grandes esfuerzos para reducir estas cifras. Hoy en día, como resultado al trabajo desplegado, se desarrolla sistemas de GPS basados en sistemas que proporcionan información en tiempo real sobre el tránsito y las condiciones climáticas. Estos datos y otras estadísticas útiles generalmente son proporcionados por los gobiernos. Sin embargo, esta información no es suficiente para reducir la incidencia de los siniestros de tránsito, por lo tanto, es importante estudiar más a fondo la problemática y poner a disposición los resultados de este estudio.

De esta manera, conocer las condiciones bajo las cuáles se producen los accidentes y dónde se originan es una información muy poderosa que puede ser utilizada para tomar medidas para evitarlos. Por ejemplo, empresas de logística puede utilizar esta información para evitar rutas específicas, las compañías de seguros pueden compartir esta información con sus clientes, y las ciudades pueden asignar a la policía de tránsito (o agentes de tránsito) a los puntos de mayor incidencia de siniestros de tránsito.

Antecedentes

Los accidentes de tráfico por lo general son aleatorios en el espacio y el tiempo debido al entorno subyacente de los accidentes de tráfico, como las redes de carreteras, los volúmenes de tráfico y, en última instancia, las actividades humanas, a menudo exhibe patrones espaciales y temporales discernibles. En el modelo de accidentes de tránsito es indispensable la aplicación de modelos multivariantes de los casos, los accidentes de tráfico para formar grupos en el espacio geográfico (Xie & Yan, 2008, pp. 396-406).

Existen modelos estadísticos sobre el análisis de accidentes de tráfico (Black & Thomas, 1998, pp. 23-31) (Li, et al., 2007, pp. 357-375) (Li, et al., 2007, pp. 274-285), entre ellos el análisis de datos atípicos que permiten examinar e involucrar la consideración de la eficiencia de los valores perdidos, el manejo de los mismos y la complejidad resultante en el análisis para distinguir dichos valores atípicos, y el sesgo entre los valores extraviados y los observados (Knorr, 2002, pp. 407-411). Una de las medidas para el análisis de datos atípicos es la distancia de Cook el cual consiste en la detección para el estudio de valores atípicos multivariantes (Kannan & Manoj, 2015). En función de lo planteado, el análisis de variable redundantes es un método de reducción de dimensionalidad que permiten mejorar la precisión de los datos y analizarlos permitiendo identificar cuantas variables pueden ser explicadas por otras dentro de un conjunto de variables (Kubus, 2018, p. 12).

Mediante el análisis de correspondencia múltiple se permite estudiar un conjunto de datos y puede resumir en diferentes dimensiones, por ello se requiere considerar variables categóricas u ordinales (De la Fuente Fernández, 2011, pp. 1-9), como es el caso de los siniestros del cantón "La Joya de los Sachas" valiéndose de las variables día, zona, feriado, causa y clase.

Por lo tanto, el análisis de conglomerados identificará mediante una partición los datos, en este caso los accidentes de tránsito en grupos con características análogas entre ellos, se requiere iniciar desde una matriz conocida como distancias de similaridades o disimilaridades entre los individuos u objetos de estudio (López, 2018).

Por todo lo expuesto anteriormente, para los departamentos de tránsito, los tribunales, los GAD municipales y para los tomadores de decisiones, es fundamental analizar las similitudes o detectar grupos específicos de siniestros de tránsito, como choques, estrellamiento, volcamiento, arrollamiento, atropello, colisión, etc. De igual manera, indicar la relación con respecto a los días en que estos ocurren, en temporadas con o sin presencia de feriados al igual que las zonas en que ocurrieron los siniestros.

En la presente tesis, se detalla un análisis descriptivo que identifica las zonas de alto riesgo de accidentes de tránsito, un análisis de variables redundantes de los siniestros de tránsito, un análisis de correspondencias para medir la asociación y un análisis de conglomerados para la detección de grupos según sus similitudes. El análisis se centra en el cantón La Joya de los Sachas (Ecuador) con información secundaria tomada del catálogo de datos abiertos de la Agencia Nacional de Tránsito del periodo 2015-2020.

Planteamiento del problema

• Enunciado del problema

La Agencia Nacional de Tránsito del Ecuador registra día a día accidentes de tránsito en las vías, las mismas que ocasionan pérdidas de vidas, lesiones leves o graves que requieran tratamientos o cuidados de por vida, daños en los vehículos e infraestructuras públicas o privadas, económicas debido a sanciones legales y reconstrucción de los bienes implicados, esto se evidencia en las cifras de ANT que existe aumento de siniestros en el cantón La Joya de los Sachas. Esta entidad enumera los sucesos, ubicación, tipo y las causas por las que acontecen, este comportamiento se debe a errores por parte de los conductores o hacer caso omiso a las señales de tránsito. En el cantón "La Joya de los Sachas" perteneciente a la provincia de Orellana, se evidencia un aumento en el número de accidentes de tránsito lo mismo que conlleva a que la provincia se situé como la segunda provincia con más letalidad, se deduce que al ocurrir 1000 accidentes existen 31 fallecidos (Lahuathe, 2018, pp. 1-15), debido a que se evidencian las siguientes causas: las víctimas sobrepasan el límite máximo de velocidad, conducen bajo los efectos del alcohol, no usar el cinturón de seguridad o el casco de motociclista (OMS, 2013, p. 6). Todo lo mencionado anteriormente recalca la importancia de analizar y detectar relaciones entre los siniestros para visualizar la existencia de similitud en grupos, de este modo tomar decisiones que conlleven a la disminución de pérdidas de vidas, recursos, e incluso lesiones en los involucrados.

- **Formulación (incógnita)**

¿Qué relación existe entre los siniestros del cantón La Joya de los Sachas?, ¿Qué similitud presentan los siniestros del cantón La Joya de los Sachas?

Justificación

En Ecuador se registraron un total de 338.442 accidentes de tránsito entre 2000 y 2015, resultando en 233.794 heridos y 26.811 muertos. La presente investigación se enfocará en el estudio de siniestros de tránsito ocurridos en el Cantón La Joya de los Sachas, provincia de Orellana, donde se estiman las tasas de brutalidad del cantón a nivel provincial donde el 6.8% representa a los siniestros, 3.4% con respecto a lesionados y el 10% a fallecidos. Además, las cifras reportadas por la Agencia Nacional de Tránsito se evidencian hasta la actualidad un incremento constante de accidentes debido a diversas causas que los acarrean, por tanto, la entidad ANT tiene el compromiso de regulación, planificación y control del transporte terrestre, tránsito y seguridad vial en el territorio nacional.

Se considera un análisis detallado de los accidentes de tránsito ocurridos durante el período 2015-2020, mediante técnicas multivariantes, para así analizar las variables: número de lesionados y fallecidos, zona, día, feriado, causa y clase que acontecen estos sucesos, permitirá aclarar e identificar la relación y asociación de los siniestros que ocurren desde hace mucho tiempo. De tal manera, esta investigación es de vital importancia para el Gobierno Autónomo Descentralizado ya que contribuirá a diseñar y mejorar políticas públicas para regular, planificar y controlar el tránsito del cantón y así generar seguridad vial a los ciudadanos.

OBJETIVOS

Objetivo general

- Aplicar técnicas multivariantes a los siniestros de tránsito del cantón La Joya de los Sachas, 2015-2020, con la finalidad de identificar la relación y similitud entre ellos y que faciliten al GAD Municipal en la toma de decisiones.

Objetivos específicos

- Análisis estadístico descriptivo de cada una de las variables en estudio.
- Depurar la base de datos mediante datos atípicos y aplicar variables redundantes para la selección de variables adecuadas para este estudio.
- Aplicar análisis de correspondencias para medir la asociación que tienen los siniestros.
- Aplicar análisis de conglomerados para detectar grupos de siniestros en función de las similitudes entre ellos.

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL

1.1. Bases conceptuales

1.1.1. Análisis exploratorio de datos

El análisis exploratorio de datos (AED) fue promovido por el estadístico John Tukey en su libro de 1977, "Análisis de datos exploratorios". El objetivo general del AED ayuda a formular y refinar hipótesis que conducirán a análisis informativos o una mayor recopilación de datos. Los objetivos centrales de EDA son:

- Sugerir hipótesis sobre las causas de los fenómenos observados.
- Orientar la selección de herramientas y técnicas estadísticas adecuadas.
- Evaluar los supuestos en los que se basará el análisis estadístico.
- Proporcionar una base para una mayor recopilación de datos.

AED implica una combinación de métodos de análisis numéricos y visuales. En ocasiones, los métodos estadísticos se utilizan para complementar el AED, pero su objetivo principal es facilitar la comprensión antes de sumergirse en el modelado estadístico formal.
Incluso si se piensa que ya se sabe qué tipo de análisis se debe realizar, siempre es una buena idea explorar un conjunto de datos antes de sumergirnos en el análisis. Como mínimo, esto ayudará a determinar si los planes son razonables o no. Muy a menudo, se descubre nuevos patrones y conocimientos. En este apartado de la tesis se examina algunos conceptos básicos que sustentan el AED como:

- Clasificación de los diferentes tipos de datos y
- Distinción entre poblaciones y muestras.

1.1.2. Variable

Una variable es un componente esencial de cualquier dato estadístico. Es una característica de un miembro de una muestra o población dada, que es única y puede diferir en cantidad o cantidad de otro miembro de la misma muestra o población. Las variables son las cantidades primarias de interés o actúan como sustitutos prácticos de las mismas. La importancia de las variables es que ayudan en la operacionalización de conceptos para la recolección de datos. Por ejemplo, si quiere hacer un experimento basado en la severidad de la urticaria, una opción sería medir la severidad usando una escala para calificar la severidad de la picazón. Esto se convierte en una variable

operativa. Para que una variable sea catalogada como "buena", debe tener algunas propiedades, como buena confiabilidad y validez, bajo sesgo, factibilidad/practicidad, bajo costo, objetividad, claridad y aceptación. (Kaliyadan & Kulkarni, 2019, p. 83).

1.1.3. Clasificación de variables

Las variables se pueden clasificar de varias maneras, como se explica y detallan a continuación:

a) Variables numéricas

Son aquellas que tienen valores que describen una cantidad medible como un número, como "cuántos" o "cuánto". Las variables numéricas también se denominan variables cuantitativas; los datos recopilados que contienen variables numéricas se denominan datos cuantitativos. Las variables numéricas se pueden describir con más detalle como continuas o discretas:

b) Variable numérica continua

Las observaciones pueden tomar cualquier valor entre un determinado conjunto de números reales, es decir, números representados con decimales. Los ejemplos de variables continuas son masa corporal, edad, tiempo y temperatura. Aunque en teoría las variables continuas pueden admitir cualquier número en el conjunto de números posibles, en la práctica los valores dados a una observación pueden estar acotados y solo pueden incluir valores tan pequeños como lo permita el protocolo de medición.

c) Variable numérica discreta

Las observaciones pueden tomar un valor basado en un recuento de un conjunto de valores completos; por ejemplo, 1, 2, 3, 4, 5, etc. Una variable discreta no puede tomar el valor de una fracción entre un valor y el siguiente valor más cercano. Los ejemplos de variables discretas incluyen el número de individuos en una población, el número de descendientes producidos y el número de individuos infectados en un experimento. Todos estos se miden como unidades completas.

1.1.4. Escalas de medición

a) Escala de intervalo

Esto permite el grado de diferencia entre los elementos de datos, pero no la relación entre ellos. Este tipo de escala no tiene un valor cero único y no arbitrario. Sin embargo, se puede comparar la relación de diferencias en una escala de intervalo. Un buen ejemplo de una escala de intervalo es la fecha, que se mide en relación con una época arbitraria.

b) Escala de razón

Esta escala posee un valor cero significativo. Toma su nombre del hecho de que una medida en esta escala representa una relación entre una medida de la magnitud de una cantidad y una unidad del mismo tipo.

1.1.4.1. Variables categóricas

Son aquellas que tienen valores que describen una característica de una unidad de datos, como "qué tipo" o "qué categoría". Las variables categóricas se clasifican en categorías mutuamente excluyentes (en una categoría u otra) y exhaustivas (incluyen todas las opciones posibles). Por lo tanto, las variables categóricas son variables cualitativas y tienden a estar representadas por un valor no numérico. Los datos recopilados para una variable categórica son datos cualitativos. Las variables categóricas pueden describirse además como ordinales o nominales:

a) Variable ordinal

Las observaciones pueden tomar un valor que se puede ordenar o clasificar de manera lógica. Las categorías asociadas con las variables ordinales se pueden clasificar por encima o por debajo de otra, pero no necesariamente establecen una diferencia numérica entre cada categoría. Ejemplos de variables categóricas ordinales incluyen calificaciones académicas (A, B, C), clase de tamaño de una planta (pequeña, mediana, grande) y comportamiento.

b) Variable nominal

Las observaciones pueden tomar un valor que no se puede organizar en una secuencia lógica. Ejemplos de variables categóricas nominales incluyen sexo, tipo de negocio, color de ojos, religión y marca. Una segunda forma de clasificar variables numéricas se relaciona con la escala

en la que se miden. La escala de medición es importante porque determina cómo se interpretan cosas como las diferencias, las proporciones y la variabilidad.

1.1.5. Poblaciones, muestras y distribuciones

El AED se ocupa principalmente de las propiedades de las muestras: su objetivo es caracterizar la muestra en cuestión sin tratar de decir demasiado sobre la población más amplia de la que se deriva. El objetivo al explorar la distribución muestral de una variable es responder preguntas como, ¿Cuáles son los valores más comunes de la variable? y ¿En qué se diferencian las observaciones entre sí? Hay dos maneras de hacer esto:

1.1.5.1. Mediante estadísticas descriptivas

Las estadísticas descriptivas se utilizan para cuantificar las características básicas de una distribución de muestra. Proporcionan resúmenes simples sobre la muestra que pueden usarse para hacer comparaciones y sacar conclusiones preliminares.
La estadística descriptiva involucra varios métodos que reducen grandes conjuntos de datos que se presentan en forma de tablas o gráficos para especificar las características de su distribución y se describen como sumas, promedios, relaciones y diferencias. Se miden en términos de ubicación central y de dispersión. Las estadísticas descriptivas no están orientadas a la "decisión". Los estudios pilotos, por ejemplo, son descriptivos (Ali , et al., 2019, p. 120).

1.1.5.2. Mediante resúmenes gráficos

Las estadísticas descriptivas no son de mucha utilidad por sí solas, por la sencilla razón de que unos pocos números no pueden capturar todos los aspectos de una distribución de muestra en la que se tiene interés. Los resúmenes gráficos son un complemento ideal para la estadística descriptiva porque permiten presentar mucha información sobre la distribución de una muestra en un solo lugar y de una manera fácil de entender para las personas.

1.1.6. Indicadores descriptivos

Cada vez que se recopila datos, se obtiene un grupo de puntajes en una o más variables que formen parte del estudio. Si se toma los puntajes de una variable y se ordenan los mismos de menor a mayor, lo que se obtiene es una distribución de puntuaciones. Los investigadores a menudo quieren saber acerca de las características de estas distribuciones de puntuaciones, como la forma de la distribución, qué tan dispersas están las puntuaciones, cuál de las puntuaciones es la más

común, y así sucesivamente. En los datos no agrupados, se resumen según las medidas de posición y medidas de dispersión (Urdan, 2010, p. 14).

1.1.7. Medidas de posición de tendencia central

La medida de tendencia central se define como la medida estadística que identifica un solo valor como representativo de toda una distribución en estudio. Su objetivo es proporcionar una descripción precisa de la totalidad de los datos. Es el valor único que es más típico representativa de los datos recogidos. El término "procesamiento de números" se utiliza para ilustrar este aspecto de la descripción de datos (Manikandan, 2011, p. 140).

Los estadísticos usan una gran variedad de medidas o índices numéricos para de esta manera resumir los datos de una manera concisa pero informativa. (Williams, 1984, p. 51). Un conjunto de características de distribución que suele interesar a los analistas es la tendencia central. Este conjunto está formado por la media, la mediana y la moda.

- **Media aritmética**

La media se utiliza para resumir datos de intervalo. Se sabe que la media aritmética puede estar influenciada por puntos de datos atípicos, es mejor utilizarla como una medida de tendencia central cuando los datos normalmente están simétricamente distribuidos. Aunque se definen varias medias diferentes, la media aritmética es la más utilizada (Lalkhen & McCluskey , 2007, p. 127). Consiste en obtener un valor central, para ello se realiza la sumatoria de todos los datos de la muestra y se divide para *n* observaciones siendo el número total de las observaciones, se define matemáticamente como:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (1\text{-}1)$$

donde:

$\boldsymbol{n}$: cantidad de datos

$\boldsymbol{x_i}$:posición de cada dato de la variable X.

- **Mediana**

La mediana se define como el dato central cuando todos los datos están ordenados (clasificados) en orden numérico. Como tal, es una medida literal de tendencia central. Cuando hay un número par de datos, la media de los dos puntos centrales de datos se toma como la mediana. La mediana

se puede utilizar para datos categóricos ordinales y para datos de intervalo. Cuando se analizan datos de intervalo, se prefiere la mediana a la media cuando los datos no se distribuyen normalmente (simétricamente), ya que es menos sensible a la influencia de los valores atípicos (Lalkhen & McCluskey , 2007, p. 127).

Representa un valor en el cual esté indica que el 50% de los datos están por encima o debajo. Está definida como:

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & si\ n\ es\ impar \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & si\ n\ es\ par, \end{cases} \qquad (2\text{-}1)$$

donde:

$\boldsymbol{n}$: cantidad de datos

- **Moda**

Es el valor que es más constante o el dato más repetitivo en toda la muestra que se desea analizar. Sin duda, la moda es estrictamente una medida del valor más popular en un conjunto de datos y, a menudo, no es un buen indicador de tendencia central. A pesar de sus limitaciones, la moda es el único valor de medir la tendencia central en un conjunto de datos que contiene valores categóricos nominales (Lalkhen & McCluskey , 2007, p. 127).

- **Relación entre media, mediana y moda**

Para la distribución unimodal, la relación entre las tres centrales medidas de tendencia viene dada por:

1. Si la distribución presenta simetría, entonces Media = Mediana = Moda
2. Si la distribución tiene una asimetría positiva, entonces Moda < Mediana < Media
3. Si la distribución tiene asimetría negativa, entonces Media < Mediana < Moda

De esta manera, se puede verificar la relación entre ellas (Weisberg & Weisberg, 1992, p. 5).

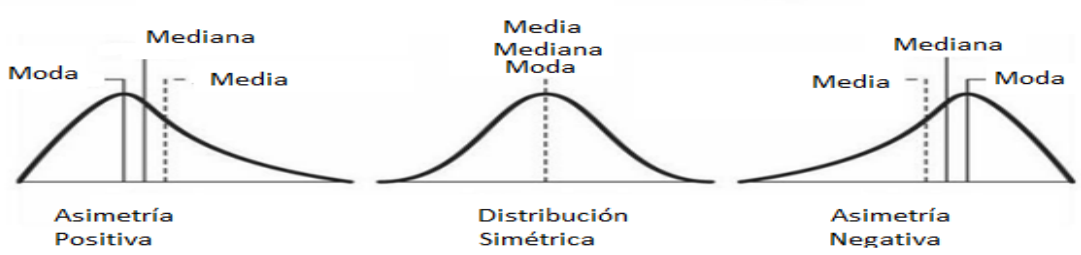


Ilustración 1-1: Relación entre las medidas de tendencia central

Fuente: (Weisberg & Weisberg, 1992, p. 5)

1.1.7.1. Gráficas de dispersión

Las medidas absolutas de dispersión se presentan en las mismas unidades que la unidad de distribución. Por ejemplo, si las unidades que se están utilizando son dólares, metros, años, etc., la medida de dispersión se presentará en términos de dólares, metros y años, respectivamente.
Las medidas absolutas de dispersión son útiles para comparar dos distribuciones donde la unidad de datos sigue siendo la misma.
En los casos en que los dos conjuntos de valores de datos se expresan en diferentes unidades, esta medida de dispersión no es útil. Por ejemplo, no puede haber ninguna comparación entre metros y dólares (Bajpai, 2019, p. 119).
Permite describir que tan parecidos o distintos son los datos entre sí de una población en estudio. Los autores (Rendón, et al., 2016, pp. 397-407) parten de la idea de los desvíos que tienen los datos con respecto a la media aritmética, entre ellos se tiene:

- **Varianza**

Es una medida de cómo los puntos de datos varían de la media y la varianza está definida por:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (3\text{-}1)$$

donde:
$\boldsymbol{n}$: cantidad de datos,
$\boldsymbol{x_i}$: posición de cada dato de la variable X,
$\boldsymbol{\bar{x}}$: representa a la media aritmética de la variable X.

- **Desviación estándar**

La desviación estándar (SD) es la medida más utilizada de dispersión. Es una medida de la dispersión de datos sobre el significar. La SD es la raíz cuadrada de la suma de la desviación al cuadrado de la media dividida por el número de observaciones (Manikandan, 2011, p. 315). La desviación estándar está definida como la raíz de la varianza:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \qquad (4\text{-}1)$$

donde:

$\boldsymbol{n}$: cantidad de datos

$\boldsymbol{x_i}$: posición de cada dato de la variable X

$\overline{\boldsymbol{x}}$: representa a la media aritmética de la variable X

- **Covarianza**

Al momento de graficar se obtiene como una nube de puntos como si se obtuviera una recta de regresión (Navarro, 2003, pp. 11-122). Se la puede calcular por:

$$cov_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y} \tag{5-1}$$

donde:

$\boldsymbol{n}$: cantidad de datos,

$\boldsymbol{x_i}$: posición de cada dato de la variable X,

$\overline{\boldsymbol{x}}$: representa a la media aritmética de la variable X.

- **El rango**

La medida más simple de la variabilidad de los datos es el rango, definido simplemente como el intervalo entre los valores más alto y más bajo en una distribución. Tiene un uso práctico limitado en el análisis estadístico, ya que obviamente está profundamente influenciado por valores atípicos extremos (Lalkhen & McCluskey , 2007, p. 128).

Esta medida de dispersión consiste en obtener la diferencia entre el dato más mayor y el más menor de la muestra de datos y a su vez añadir la unidad (Rendón, et al., 2016, pp. 397-407), y se lo calcula mediante la fórmula siguiente.

$$RANGO = X_{max} - X_{min} \tag{6-1}$$

donde:

$\boldsymbol{X_{max}}$**:** es el máximo valor de la variable X

$\boldsymbol{X_{min}}$**:** es el mínimo valor de la variable X

1.1.7.2. Medidas de posición

- **Percentiles**

Cuando un conjunto de datos se organiza en orden de magnitud, se puede dividir en 100 puntos de corte separados (percentiles). El percentil "x" se define como un punto de corte tal que el x% de la muestra tiene un valor igual o menor que el punto de corte. Por ejemplo, el percentil 35 divide los datos en dos grupos que contienen, respectivamente, el 35% y el 65 % de los datos.
Los cuartiles se usan con mayor frecuencia, es decir, inferior (percentil 25), medio (percentil 50 o mediana) y superior (percentil 75). Dividen los datos en cuatro grupos iguales. El rango intercuartílico (IQR) se cita a menudo cuando se refiere a datos de intervalo que no se distribuyen normalmente. Si el valor del percentil 25 (cuartil inferior) de un conjunto de datos es 10 y el valor del percentil 75 (cuartil superior) es 40, el IQR puede expresarse como 10–40 o simplemente como 30. Los percentiles y los cuartiles pueden estimarse a partir de un valor acumulativo curva de frecuencia. Una representación gráfica útil de la distribución de intervalo datos es el diagrama de caja y bigotes (Lalkhen & McCluskey , 2007, p. 128).

- **Cuartiles**

Los cuartiles dividen un conjunto de datos ordenados por rango en cuatro partes iguales. Los valores que dividen cada parte se denominan cuartiles primero, segundo y tercero; y se denotan por Q1, Q2 y Q3, respectivamente. En términos de percentiles, los cuartiles se pueden definir de la siguiente manera:

Tabla 1-1: Cuartiles

$Q_1 = P_{25}$
$Q_2 = P_{50}$ = Mediana
$Q_3 = P_{75}$.

Elaborado por: Carrión, Andrea, 2021.

Fuente: (Rockinson , 2013, p. 119).

1.1.8. Descripción para variables cualitativas

La información se puede resumir de dos maneras, la primera consiste en que los datos se presentan como frecuencias simples lo cual nos indica el conteo de estos y la segunda representación mediante frecuencias relativas esto es la división de cada conteo para el total de datos y a su vez se puede multiplicar por 100 indicando en porcentaje para mayor facilidad de interpretación (Rendón, et al., 2016, pp. 397- 407).

1.1.9. Gráficas

Las gráficas complementan las presentaciones tabulares de las estadísticas descriptivas. En general, los gráficos son más adecuados que las tablas para identificar patrones en los datos, de una o más variables de interés de manera visual, mientras que las tablas son mejores para proporcionar grandes cantidades de datos con un alto grado de detalle numérico. (Fisher & Marshall, 2009, pp. 93-97).

1.1.10. Gráficas para una variable

Los gráficos han sido una herramienta esencial para el análisis y la comunicación de datos estadísticos durante unos 200 años. A pesar de su uso generalizado e importancia en la ciencia, los negocios y muchos otros ámbitos de la vida, se sabe relativamente poco sobre cómo las personas perciben y procesan los gráficos estadísticos. (Lewandowsky & Spence, 1989, pp. 200-242).
El análisis de datos gráficos es útil para la limpieza de datos, la exploración de la estructura de datos, detección de valores atípicos y grupos inusuales dentro de un conjunto de datos, identificación de tendencias y grupos, detección de patrones locales, la evaluación de la salida del modelo y presentación de resultados (Unwin, 2015, p. 310).

- **Diagrama de barras**

En este gráfico se requiere de barras para representar a cada modalidad de la variable analizada, la proporción de cada caso permite la construcción de la altura de las barras, además se describe las variables numéricas cuando estás tienen pocos valores (Arteaga, et al., 2009, pp. 18-93).

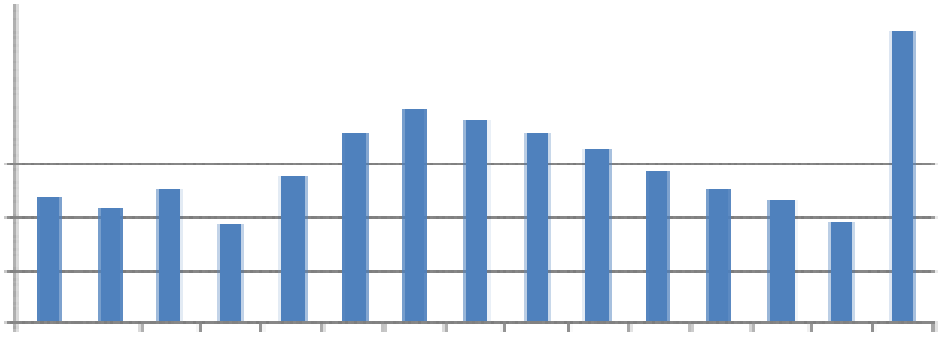


Ilustración 2-1: Representación de un diagrama de barras

Fuente: (Arteaga, 2008, p. 5)

- **Diagrama de cajas y bigotes**

Este gráfico es muy útil ya que se puede visualizar características como: la tendencia central, variabilidad, asimetría y datos alejados de la distribución, para realizar esta gráfica se requiere una caja y bigotes. Para la construcción de la caja de forma rectangular, aquí se representan el recorrido intercuartílico donde se aprecia el valor de la mediana, cuartiles, por lo tanto; la línea izquierda de la caja contiene el primer cuartil y la línea derecha contiene el tercer cuartil, entre ellas, la línea que divide en dos partes la caja contiene el valor de la mediana. Los bigotes, en el segmento izquierdo se ubica el valor mínimo y en la derecha el valor máximo de los datos. Se puede visualizar valores sospechosos de ser anómalos, estos se detectan visualmente por ser datos muy menores o mayores que están alejados o sobrepasan a los bigotes de la caja, y se requiere una prueba analítica para comprobar si los valores realmente son anómalos (López, 2007, pp. 1-379).

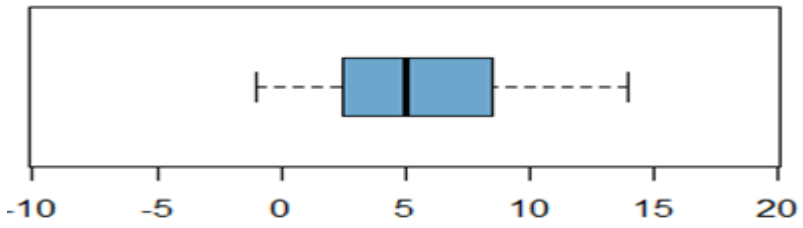


Ilustración 3-1: Representación de un diagrama de Caja y bigotes
Elaborado por: Carrión, Andrea, 2021.

1.1.11. Gráficas para dos variables

- **Diagrama de dispersión**

Los diagramas de dispersión transmiten un mensaje implícito, aunque sutil, sobre la causalidad, ya sea que se observen funciones de una variable en matemáticas puras, gráficas de medidas experimentales en función de las condiciones experimentales o gráficas de dispersión de variables predictoras y de respuesta, por convención se supone que el valor graficado en el eje vertical está determinado o influenciado por el valor en el eje horizontal (Bergstrom & West, 2018, p. 1).

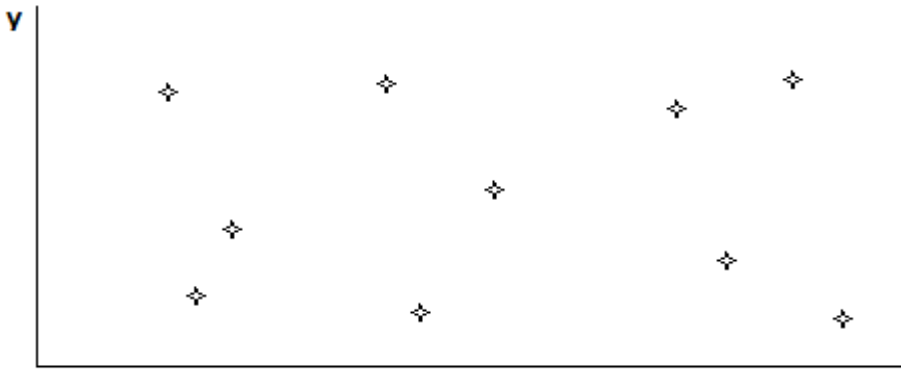


Ilustración 4-1: Representación de un diagrama de dispersión

Elaborado por: Carrión, Andrea, 2022.

Fuente: (Bergstrom & West, 2018).

1.1.12. Matriz de varianzas y covarianzas

La estimación de la matriz de varianzas y covarianzas es útil e inevitable para muchos métodos estadísticos, como en los casos de: la prueba t-student, el análisis de componentes principales y el análisis discriminante lineal, los datos de alta dimensión, como los datos de micromatrices de expresión génica y los datos financieros (Tong, et al., 2014, pp. 255-264).

Las estimaciones de las covarianzas entre variables se obtienen comúnmente a partir de un análisis de pares de variables en estudio medidas en los mismos individuos. Las covarianzas se pueden obtener a partir de una función simple entre las varianzas de cada variable y la suma de los dos mudables en interés (Schaeffer, 1986, p. 187).

1.1.13. Detección de datos atípicos

1.1.13.1. Influencia de los datos atípicos en la información

Todos aquellos valores perdidos y considerados como atípicos se suelen encontrar con regularidad durante el proceso de recopilación de datos de los estudios observacionales o experimentales empleados en alguno de los campos de las ciencias naturales y sociales.

Los valores faltantes pueden brotar de la pérdida de información, así como de abandonos y falta de recopilación de respuesta de los individuos que deben proporcionar dicha respuesta. La falta de presencia de estos valores genera una disminución de tamaño de muestra de lo supuesto y complica o dificulta la confiabilidad de los resultados del estudio previsto.

También puede producir resultados sesgados cuando las inferencias sobre una población se basan en dicha muestra, lo que socava la confiabilidad de los datos. Como parte del proceso de pretratamiento, los datos faltantes se ignoran a favor de la simplicidad o se reemplazan con valores

sustituidos estimados con un método estadístico. En general, el análisis de los valores perdidos involucra la consideración de la eficiencia, el manejo de los datos perdidos y la complejidad resultante en el análisis, y el sesgo entre los valores perdidos y los observados.

El otro problema es el de los valores atípicos, que se refieren a valores extremos que se encuentran anormalmente fuera del patrón general de una distribución de variables. Cuando se recopilan datos de peso, un valor de 250 kg no puede encajar en la distribución normal de pesos; por lo tanto, representa un valor atípico. Los valores atípicos resultan de varios factores, incluidos los errores de respuesta de los participantes y los errores de ingreso de datos.

En una distribución de variables, los valores atípicos se encuentran lejos de la mayoría de los otros puntos de datos ya que los valores correspondientes son extremos o anormales. Los valores atípicos contenidos en los datos de la muestra introducen sesgos en las estimaciones estadísticas, como los valores medios, lo que conduce a valores resultantes subestimados o sobreestimados. Tratar con valores atípicos es esencial antes del análisis del conjunto de datos que contiene valores atípicos. Esto implica modificar los valores atípicos después de identificar sus fuentes o reemplazarlos con valores sustituidos.

Los diferentes enfoques para manejar los valores perdidos y los valores atípicos pueden cambiar drásticamente los resultados del análisis de datos. Por lo tanto, el tratamiento adecuado de los datos faltantes y los valores atípicos es crucial para el análisis. (Kwak & Kim, 2017, p. 407).

1.1.14. Proceso para detección de datos atípicos

Al iniciar, lo más adecuado es representar de manera gráfica a cada una de las variables que serán tomadas en cuenta como objeto de estudio del análisis, siendo adecuado optar por un histograma o un diagrama de caja, las cuales permiten visualizar el comportamiento de datos y se puede observar si existen anomalías.

Es común considerar el diagrama de box-plot esquemático ("completo") de Tukey como una prueba informal de la existencia de valores atípicos. Si bien el procedimiento es útil, debe usarse con precaución, ya que al menos el 30 % de las muestras de una población distribuida normalmente de cualquier tamaño se marcarán como que contienen un valor atípico, mientras que para muestras pequeñas ($N < 10$), incluso los valores atípicos extremos indican poco (Dawson, 2011, p. 1).

1.1.14.1. Definición

Los datos atípicos hacen referencia a datos o valores que forman parte de un conjunto de información en estudio, pero se evidencia como algo inesperado o anormal según las especificaciones de las variables. La presencia de solo un dato atípico puede provocar alteraciones

con respecto a la media aritmética, desviaciones típicas y deshacerlas relaciones que existen entre las variables. Uno de los enfoques clásicos para encontrar valores atípicos es la distancia de Mahalanobis, el cual se define como:

$$d_M^2(X_i, \bar{X}_M) = (X_i - \bar{X}_M) S_M^{-1} (X_i - \bar{X}_M)' \quad (7\text{-}1)$$

Donde:

$\boldsymbol{X_i}$: representa al individuo

$\boldsymbol{\bar{X}_M}$: es el vector de medias

$\boldsymbol{S_M^{-1}}$: es la inversa de la matriz de varianzas y covarianzas

1.1.14.2. ***Distancia cook***

Para evitar el enmascaramiento de datos inesperados ya que esta distancia adolece del efecto de enmascaramiento por el cual múltiples valores atípicos no necesariamente tienen una gran distancia de Mahalanobis. Por lo tanto, Dennis Cook (1977) introdujo una medida de distancia para las estimaciones de uso común de la influencia de un punto de datos al realizar un análisis de regresión de mínimos cuadrados. Para ello, se emplea la fórmula:

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}^{-i}\right)^T (X^T X) \left(\hat{\beta} - \hat{\beta}^{-i}\right)}{(1+p)s^2} \quad (8\text{-}1)$$

Donde:

$\boldsymbol{\hat{\beta}}$: mínimos cuadrados estimados de β

$\boldsymbol{\hat{\beta}^{-i}}$**:** vector de coeficientes calculado después de eliminar la i-ésima observación

$\boldsymbol{X}$: matriz de datos (diseño)

$\boldsymbol{X^T}$: matriz de datos transpuestos

$\boldsymbol{p}$**:** número de parámetros

$\boldsymbol{s^2}$**:** cuadrado medio del error

Una observación que representa a un valor extremo de una variable predictora se considera como punto con gran distorsión, el cual tiene influencia en los análisis estadísticos aleja a los resultados de los mismos.

En la regresión lineal, la identificación de los puntos u observaciones distorsionadoras puede resultar un trabajo bastante difícil de detectar. En el modelo de regresión lineal, la puntuación de comparación para la i-ésima observación se encuentran entre 0 y 1. Por tanto, si se considera y se

investiga las observaciones identificadas como valores anómalos se debe considerar aquellas que tienen distancias superiores a cuatro veces la media pueden clasificarse como datos o valores influyentes (Kannan & Manoj, 2015, p. 2319).

1.1.15. Análisis de variables redundantes

Se define redundancia en el sentido amplio de la posibilidad de que atributos únicos posean una causa y una referencia semántica compartida sustantiva ejemplo: redacción de elementos similares, contenido de elementos similares; y tendencia de respuesta general (Christensen, et al., 2020, pp. 1-36).

En la teoría de pruebas clásica y moderna, generalmente se piensa que la redundancia es beneficiosa para la construcción de pruebas porque aumenta la homogeneidad de la prueba y minimiza las causas idiosincrásicas, es decir, reduce el error de medición. Esta visión de la redundancia a menudo está sesgada hacia una causa sustantiva compartida y descarta otras causas potenciales (Christensen, et al., 2020, pp. 1-36).

Los analistas de datos que están preocupados por obtener buenos ajustes a las respuestas tienen una tendencia natural a incluir cualquier variable de predicción que pueda influir en la respuesta en una ecuación de predicción. Esto es especialmente cierto cuando se realizan estudios piloto o estudios referentes a esfuerzos de investigación donde no existe un modelo teórico o evidencia experimental pasada que sugiera cómo se debe especificar el modelo.

En algunos casos, el miedo al sesgo en el predictor lleva a los investigadores a recopilar datos sobre tantas variables como sea posible, las ecuaciones de predicción formadas de esta manera pueden sufrir una especificación excesiva. Demasiadas variables predictoras en el modelo pueden crear problemas cuando el objetivo de la investigación es hacer inferencias sobre los parámetros del modelo. Esto se debe a que existe un gran peligro de redundancia entre las variables predictoras. Si varias variables independientes repiten información, es difícil determinar cuál de las variables redundantes debe retenerse en el modelo y cuál debe eliminarse (Richard F. Gunst, 1980, pp. 170-500).

Los estudios observacionales adolecen con frecuencia de una redundancia inevitable: demasiadas variables aumentan o disminuyen de forma lineal aproximada con el tiempo. Sin embargo, equilibrar la necesidad de poseer suficientes variables para asegurar que no se produzcan grandes sesgos en la ecuación de predicción con el peligro de introducir redundancias innecesarias no es una tarea fácil. Como precaución mínima, el analista de datos debe tener cuidado de no incluir variables predictoras solo porque estén disponibles (Gunst & Mason, 1980, pp. 170-500).

Aunque la mayoría de los textos de validación instructiva sugieren muestrear todo el contenido que es útil o relevante para el atributo objetivo, gran parte de la literatura sobre pruebas para el desarrollo y la validación sugieren el uso de las variables que conducen a estimaciones de alta

consistencia interna para cada prueba unidimensional. Esta recomendación no es explícita, pero es parte del desarrollo de la escala y proceso de validación en el que las variables con baja consistencia interna se eliminan del grupo final de variables. De hecho, las medidas de coherencia interna son el método más utilizado para selección de variables.

En el año 1988 Comrey fue uno de los defensores del uso de variables similares para la identificación de cada factor unidimensional a través de sus dimensiones de ítem homogéneas factorizadas. Para construir estas dimensiones, él sugirió que el desarrollo de escalas de cuatro ítems debería resumirse para formar una "escala homogénea", sugiriendo implícitamente el uso de ítems redundantes (Christensen, et al., 2020, pp. 1-36).

En las perspectivas más recientes se ve a la redundancia de manera menos favorable, particularmente en áreas donde los atributos son amplios y más allá de las razones teóricas, la redundancia puede tener consecuencias para las interpretaciones de cualquier estudio (Christensen, et al., 2020, pp. 1-36).

1.1.16. Problemas potenciales de redundancia

La dimensionalidad, es un área que probablemente se ve afectada por relaciones suficientemente fuertes entre subconjuntos de variables para formar factores menores o factores de residuales correlacionados con varianza o varianza única (Christensen, et al., 2020, pp. 1-36).

La redundancia también puede tener consecuencias similares para los modelos factoriales y de variables latentes. Inherente a los modelos de variables latentes se encuentra el principio de independencia local, es decir, después de condicionar las variables a una variable latente, estas se convierten en estadísticamente independientes.

La redundancia, independientemente de cómo o por qué, a menudo quebrantará este principio, conduciendo a modelos que se ajustan mal y que confunden la interpretación de los puntajes de las pruebas estadísticas (Christensen, et al., 2020, pp. 1-36).

Para mitigar los efectos de un ajuste deficiente, la práctica común es simplemente correlacionar la varianza residual en el modelo. Este enfoque solo evita el problema de la posible redundancia y no aborda la cuestión de la interpretación de las puntuaciones de las pruebas (Christensen, et al., 2020, pp. 1-36).

Por ello, en vista de que en ocasiones las bases de datos cuentan con variables de mala calidad que ralentizan y hacen difícil el tiempo de análisis, se opta por seleccionar un nuevo conjunto de variables que representan la misma información de las variables redundantes existentes, o bien realizar una combinación de las variables redundantes que maximice la información contenida en ellas reduciendo significativamente el número de estas (Cestero & Caballero, 2018, pp. 5-276).

1.1.16.1. Coeficiente de correlación de Pearson

Esta prueba se aplica cuando las variables en estudio son normalmente distribuidas, y se sabe que las mismas presentan una asociación entre ellas (Ali , et al., 2019, p. 125). Este coeficiente de correlación se puede utilizar como criterio de optimización para derivar diferentes filtros de reducción de ruido óptimos, pero es aún más útil para analizar estos filtros óptimos por su rendimiento de reducción de ruido (Benesty, et al., 2009, pp. 1-4).

Los coeficientes de correlación son índices que representa o miden la fuerza de la relación estadística entre dos variables aleatorias como objeto de estudio que obedecen a una distribución de probabilidad conjunta. En general, los coeficientes de correlación se requieren que sean grandes y positivos ya que si existe una alta probabilidad de que valores grandes (pequeños) de una variable se presenten junto con valores grandes (pequeños) de otra; y debe ser grande y negativo si la dirección se invierte.

Existen algunos métodos de análisis de correlación, entre ellos se tiene; el coeficiente de correlación conocido como producto-momento de Pearson, también la tau de Kendall y la rho de Spearman los cuales son los más conocidos y comúnmente usados, los cuales son ideales para medir el grado de asociación entre las variables de interés. (Ma , et al., 2012).

El coeficiente de correlación (***ρ)*** producto-momento de Pearson es una medida de la dependencia lineal entre dos variables aleatorias. Su versión muestreada, comúnmente denotada por ***r***, ha sido bien estudiada por los fundadores de la estadística moderna como Galton, Pearson y Fisher.

Sin embargo, con base en conocimientos geométricos, Fisher en el transcurso de los años de 1915 y 1921 consiguió derivar la distribución muestral exacta de este coeficiente ***r*** y estableció que esta distribución muestral converge a una distribución normal a medida que aumenta el tamaño de la muestra (Li, et al., 2018, pp. 4-13).

Este estadístico se calcula mediante la ecuación que relaciona dos conjuntos de puntuaciones de diferentes medidas. Esta ecuación genera un solo valor conocido como coeficiente de correlación y se designa la letra ***r*** (Emerson, 2015, p. 242). La fórmula es:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \qquad (9\text{-}1)$$

donde:

n: es el tamaño de las muestras

X: variable X en estudio

Y: variable Y en estudio

Determinar la existencia o no de una relación entre dos variables de interés es de mucha utilidad. De aquí se puede determinar, ¿qué tan significativa o fuerte es esta asociación entre las dos variables? Por ejemplo, ¿existe una relación entre la variable "los años de servicio como ecografista" y la mudable "los puntajes obtenidos en el examen de registro"? Sin embargo, mediante el coeficiente de correlación o coeficiente ***r*** es el estadístico que se utiliza y permite medir el grado o la fuerza de este tipo de relación o asociación (Richard, 1990, p. 36).

El valor representativo del coeficiente de correlación se encuentra en el rango de -1 a +1, representando el valor de 0 que no hay o no existe de ninguna manera una asociación lineal o monótona entre las variables, y la relación se fortalece y finalmente se aproxima a una línea recta o una curva que aumenta o disminuye constantemente a medida que el valor obtenido representativo del coeficiente se aproxima a un valor absoluto de 1. Además, mediante las pruebas de hipótesis y los intervalos de confianza se pueden utilizar para abordar la importancia estadística de los resultados y para estimar la fuerza de la relación en la población de la que se tomaron muestras de los datos (Schober, et al., 2018, pp. 1763-176).

1.1.17. Prueba del coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson representa el grado en que los sujetos tienen el mismo orden en dos variables. Por ejemplo, si los sujetos más altos fueran más pesados y los sujetos más bajos fueran más livianos, entonces tendríamos una correlación positiva entre el peso y la altura: cuanto más alto, más alto, mayor el peso.

a) Hipótesis

H_0: *No existe una relación lineal entre las variables X e Y*

H_1: *Existe una relación lineal entre las variables X e Y*

b) Coeficiente de correlación

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \qquad (10\text{-}1)$$

Donde:

X: representa a los datos de la variable X

Y: representa a los datos de la variable Y

n: es el tamaño de la muestra

1.1.18. Coeficiente de correlación significativa

Se considera la expresión $p < 0.05$, o probabilidad de error inferior al 5% de probabilidad de error cuando existe una relación. Una correlación estadísticamente significativa significa que en muestras similares se encuentra una correlación distinta de cero entre dos variables. Se puede inferir el hecho de esta relación, no su magnitud.

1.1.19. Magnitud del coeficiente de correlación

Una vez que se ha determinado que un coeficiente de correlación es estadísticamente significativo, como los valores mínimo y máximo son 0 y ±1, si la magnitud es baja, por ejemplo: 0.20 será intuitivamente una relación baja y si la magnitud es alta o alrededor de 0.7 indicará una relación que se considera grande (Dagnino, 2014).

1.1.20. Análisis de correspondencias múltiples

El análisis de correspondencias múltiples (ACM), también se conoce como un análisis de correspondencia de la tabla de Burt $\boldsymbol{B} = \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z}$, la matriz de todas las asociaciones que se forman por pares entre las variables en estudio, aquí se incluyen las asociaciones diagonales entre cada variable y así misma. En esta tabla, la información sobre los individuos no está disponible y solo está presente la información sobre las relaciones entre categorías.

La representación de la nube de categorías para este análisis tiene que coincidir exactamente con la representación de la nube de baricentro. No obstante, es posible representar a los individuos en el gráfico de categorías y así obtener la representación del gráfico medio proyectando la matriz de indicadores como elementos complementarios. En esta ilustración, es posible realizar la interpretación de las distancias entre un individuo y una categoría analizada, siendo la misma clasificada como un conjunto de individuos. Además, esta gráfica visual es la representación óptima de los individuos en cuanto al ajuste de la nube de individuos, y del baricentro de los individuos en cuanto al ajuste de la tabla de Burt (Husson & Josse, 2014, p. 177).

En el análisis de correspondencias se requiere construir una tabla de contingencia que es equivalente al análisis de correspondencias simples en la cual se considerandos variables nominales (es decir, una para las filas y otra para las columnas), este es una extensión de análisis de correspondencias simples (AC) que permite analizar un conjunto de observaciones descritas por un conjunto de variables nominales. Cada variable nominal se compone de varios niveles, y cada uno de estos se codifica como una variable binaria. Por ejemplo, el género (F vs M) es una variable nominal con dos niveles. El patrón para un encuestado en este ejemplo se codificaría de

la siguiente manera: masculino tomará valores de [0,1] y los valores de [1,0] indicaría el caso para un individuo de género de identidad mujer.

La tabla de datos completa está compuesta por columnas binarias con una y sólo una columna tomando el valor "1" por variable nominal. En el ACM también se puede codificar variables cuantitativas recodificándolas. Por ejemplo, una puntuación con un rango de −5 a +5 se podría denotar como una variable nominal con tres niveles: menor que 0, igual a 0, o mayor que 0. El esquema de codificación de ACM implica que cada fila tiene el mismo total, lo que para AC implica que cada fila tiene el mismo peso. Se puede demostrar que el esquema de codificación binaria utilizado en el ACM crea factores artificiales y, por lo tanto, se reduce artificialmente la inercia explicada en los primeros factores del análisis (Abdi & Valentin, 2007, pp. 3-10).

Antes de comenzar con el desarrollo correspondiente, se debe cubrir algunas preguntas generales que son relevantes para cada análisis de correspondencia. La primera pregunta importante es: ¿qué buscar? Por ejemplo, si hay cuatro opciones de interés: pesca, exceso de simplicidad, exceso de complejidad y escasez de datos. Se considera brevemente cada uno a su vez. Por pesca, se entiende a la selección de manera arbitraria (o casi arbitraria) de los factores con la esperanza de encontrar correlaciones. El análisis de correspondencias es una herramienta para identificar correlaciones entre los objetos de estudio, una herramienta que necesita ser utilizada de manera razonada. No tiene sentido establecer correlaciones entre el uso de características del lenguaje que no tienen una correlación interpretable en la realidad, o peor aún, tienen una correlación interpretable, pero son solo el resultado de unas pocas ocurrencias fortuitas (Glynn & Robinson, 2014, p. 450).

El exceso de simplicidad es un problema menos grave, pero el cual debe tenerse en cuenta. No sería nada útil aplicar un análisis de correspondencias para identificar una correlación que un simple gráfico circular o un histograma, combinado con una prueba de significación, haría aún mejor. Del mismo modo, las correlaciones obvias pueden dominar los resultados a expensas de resultados menos obvios y, por lo tanto, más interesantes.

El problema radica en que, si estos dos factores se localizan entre una gama más compleja de componentes, la asociación presente entre ellas obviamente podría anular u ocultar otras asociaciones que se perciban. Por esto, a veces es necesario el tener que incluir correlaciones tan obvias en un análisis, en tal caso si es posible evitarlo, lo factible sería evitarlo. Es decir, las correlaciones conocidas como obvias corren el riesgo de esconder los resultados más interesantes del estudio. En otras palabras, la gráfica identificará lo que está más fuertemente correlacionado en lugar de las correlaciones visualizadas como las más sutiles, pero analíticamente éstas serán las más importantes.

La complicación repercuta en el análisis de correspondencias binarias cuando se usan tablas relacionadas y en el caso del análisis de correspondencias múltiples cuando se examinan demasiados factores de manera simultánea (Glynn & Robinson, 2014, p. 450).

1.1.21. Notación

Hay *k* variables nominales, cada variable nominal tiene J_k niveles y la suma de losJ_k es igual a J. Existen I observaciones, la matriz de indicadores IxJ se denota con **X**. Al momento de realizar un AC en la matriz de indicadores se proporcionará dos conjuntos de puntajes de factores: uno de ellos con respecto a las filas y otra para corresponderá a las columnas. Estos puntajes factoriales son, en general, escalados de tal manera que su varianza es igual a su correspondiente valor propio.

El gran total de la tabla se anota N, y el primer paso del análisis consiste en calcular la matriz de probabilidad $\boldsymbol{Z = N^{-1}X}$. Las puntuaciones factoriales se las logra obtener a partir de la descomposición de valores singulares por medio de la fórmula: (Abdi & Valentin, 2007, pp. 3-10).

$$D_r^{-\frac{1}{2}} = (Z - rc^T)D_c^{-\frac{1}{2}} = P\Delta Q^T \qquad (11\text{-}1)$$

Donde:

$\boldsymbol{Z}$**:** matriz de probabilidad

$\boldsymbol{c}$**:** el vector de los totales de las columnas

$\boldsymbol{r}$: el vector de los totales de las filas de Z

$\boldsymbol{D_c}$**:** diag{c}

$\boldsymbol{D_r}$: diag{r}

1.1.21.1. Biplots

El análisis de correspondencia ha permitido el cálculo de valores de proximidad que se requiere para la combinación de las celdas en las filas y columnas para la tabla de contingencia, los cuales se pueden trazar en un espacio gráfico. En cada dimensión del gráfico se representará un cierto porcentaje de la estructuración de la variación de datos, o "inercia", de aquí se debe trazar una sola dimensión, una línea simple (representando el eje x), en el cual se colocará los puntos de datos en esta línea en diferentes distancias entre sí. Sin embargo, en la mayoría de las situaciones, esto representará pobremente las relaciones entre esas características.

Si se añade una segunda dimensión, el eje y, se obtendrá un biplot bidimensional, típico del análisis de correspondencias y un rango de otras técnicas que permiten la reducción de espacio. Esto permitirá representar con una magnífica precisión a una gran parte de la estructura en la variación de los datos. Matemáticamente, se tiene que el número de dimensiones posibles es igual al número de filas o columnas (la que sea menor) menos uno. Por tanto, para visualizar una tabla con cinco filas y ocho columnas, se necesitarían cuatro dimensiones. Las puntuaciones de la

inercia (o variación explicada) suelen darse para estas dos primeras dimensiones; los ejes x e y del biplot (Glynn & Robinson, 2014, p. 447).

Ilustración 5-1: Representación de un biplot.

Fuente: (Blasius & Greenacre, 2008, p. 141)

1.1.22. Análisis de conglomerados

Los métodos de análisis de conglomerados acarrean una larga historia, a través de los años 1911, el antropólogo Czekanowski dio a conocer los primeros procedimientos, posteriormente años más tarde, Driver y Kroeber lo sugieren específicamente en el año 1932.Años más tarde, estas ideas fueron recogidas y desarrolladas en el ámbito de la psicología. Por ejemplo, Zubin en el año 1938 propuso un método bastante simple para clasificar una matriz de correlación que detectaría grupos. Por otro lado, Stephenson durante el año 1936 sugirió el uso del análisis factorial invertido para encontrar grupos de individuos. (Blashfield & Aldenderfer, 1988, p. 447).

La utilización del análisis de conglomerados es ideal cuando el objetivo es individualizar en un conjunto de grupos de objetos similares para interpretarlos como miembros de una "categoría". El análisis para agrupamiento también es óptimo para aplicar a las variables, para encontrar grupos de variables similares, frecuentemente con el objetivo de selección de variables para técnicas de calibración (Forina, et al., 2002, p. 13).

El análisis de correspondencias también es conocido como “análisis clúster”, mediante esta técnica multivariante se logra clasificar objetos descritos como datos. Una de las razones por las que el análisis de conglomerados es tan útil es que los investigadores en todos los campos necesitan hacer y revisar clasificaciones continuamente. Además, es el método más básico para estimar similaridades.

Los métodos de análisis de conglomerados siguen un conjunto prescrito de pasos, siendo los principales:

• Se inicia constituyendo una matriz de datos en la cual cuyas columnas se representa el objeto que se va está analizando y en las filas se ubica los atributos que describen al objeto en estudio.

• Los datos deben reducir su dimensionalidad mediante la estandarización de los datos, cabe mencionar que este paso es opcional según el criterio del analista conforme a la información que se esté trabajando.

• Se requiere calcular los valores de un coeficiente de semejanza que el analista haya seleccionado, y la cual permitirá medir las similitudes entre todos los pares de objetos.

• Usar un método de agrupamiento para procesar los valores de los coeficientes de semejanzas, el cual permite generar un diagrama llamado árbol de cual se pueden leer los grupos.

PASO 1: OBTENER LA MATRIZ DE DATOS

Las columnas de la matriz de datos representan los objetos considerados al análisis, cuyas similitudes entre sí que se requiere estimar, en las filas se representan los atributos, las propiedades de los objetos. Por ejemplo, los objetos podrían ser personas; los atributos, un conjunto de sus respuestas a un test psiquiátrico. O los objetos podrían ser parcelas de tierra en un bosque; los atributos, recuentos de las especies de árboles que crecen en las parcelas. O los objetos podrían ser fósiles; los atributos, sus dimensiones. El hecho de que los objetos y atributos puedan ser casi cualquier cosa le da al análisis de conglomerados una variedad infinita. De esta manera, el analista logra el objetivo de conglomerados es averiguar qué objetos son similares y diferentes entre sí.

PASO 2: ESTANDARIZACIÓN DE LA MATRIZ DE DATOS

La estandarización es opcional según lo disponga el analista. La estandarización de la matriz de datos convierte los atributos originales en nuevos atributos sin unidades, para esto se requiere sólo de aritmética simple, por lo tanto, de esta manera se transforma la matriz de datos original en una matriz de datos estandarizados. La nueva matriz al igual que la matriz original, tiene la misma cantidad de objetos y atributos. Pero sus valores han cambiado y algunos incluso pueden ser negativos.

Los cálculos principales del análisis de conglomerados se pueden realizar en la matriz de datos o en la matriz de datos estandarizados. La estandarización refunde las unidades de medida de los atributos como unidades adimensionales.

La base de datos original trabaja bajo ciertas unidades de medición como: pulgadas, metros, libras, gramo, etc.

La estandarización facilita eliminar la unidad de medición de cada atributo en estudio, cambiando su valor numérico y reformulando en forma adimensional. Por ahora, se analizará la matriz de datos original en lugar de la matriz de datos estandarizada.

PASO 3: CALCULAR LA MATRIZ DE SEMEJANZAS

Un coeficiente de semejanza permite medir la semejanza general o el grado de similitud, entre cada par de objetos. Un coeficiente de semejanza es una fórmula matemática, el cual se calcula considerando para un par dado de objetos, los valores de sus columnas en la matriz de datos.

La fórmula da un valor que representa qué tan similares son el par de objetos. Por lo cual, se debe de optar por uno de ellos para proceder con los cálculos. Se debe calcular *n* cantidad de valores del coeficiente de semejanza, es decir, un valor para cada par de objetos.

En cada columna se identifica el primer objeto de un par; cada fila identifica el segundo objeto. La celda formada por la intersección de una columna y una fila contiene el valor del coeficiente de semejanza para el par de objetos considerado. Además, el concepto de semejanza es simétrico, por esta razón, solo la mitad inferior izquierda de la matriz contiene valores ya que los restantes solo es cuestión de ubicarlos según corresponda.

Un coeficiente de semejanza es siempre uno de dos tipos:

- un coeficiente de disimilitud
- un coeficiente de similitud

La diferencia de los coeficientes es una cuestión de "dirección". Cuanto más pequeño es el valor de un coeficiente de disimilitud, más similares son los dos objetos. Cuanto mayor es su valor, más diferentes son éstos. Por otro lado, cuanto mayor sea el valor de un coeficiente de similitud, más similares serán los dos objetos. Y cuanto menor es su valor, más diferentes lo son entre ellos.

La matriz nueva contiene los valores de un coeficiente de disimilitud denominado coeficiente de distancia euclidiana, el cual permite medir la distancia literal entre dos objetos cuando se ven como puntos en el espacio bidimensional formado por sus atributos. El coeficiente de similitud o disimilitud se utiliza para medir la asociación de los clústeres en la tecnología de grupo. (Sarker & Islam, 1999, p. 769).

Las medidas de similitudes o distancias son componentes centrales utilizados por los algoritmos de agrupación en clústeres que se basan en la distancia para la agrupación de los puntos de datos

similares en los mismos clústeres, mientras que los puntos de datos diferentes o distantes se colocan en diferentes clústeres según las particularidades semejantes.

El desempeño de las medidas de similitud se aborda principalmente en espacios bidimensionales o tridimensionales, más allá de los cuales, no existe ningún estudio empírico que haya revelado el comportamiento de las medidas de similitud cuando se trata de conjuntos de datos de alta dimensión.

Por lo cual en este estudio se propone un marco técnico para analizar, comparar y comparar la influencia de diferentes medidas de similitud en los resultados de los algoritmos de agrupamiento basados en la distancia. Con fines de reproducibilidad, se utilizaron quince conjuntos de datos disponibles públicamente para este estudio y, en consecuencia, las medidas de distancia futuras se pueden evaluar y comparar con los resultados de las medidas discutidas en este trabajo.

Estos conjuntos de datos se clasificaron en categorías como baja y alta dimensión para estudiar el rendimiento de cada medida frente a cada categoría (Shirkhorshidi, et al., 2015, p. 1).

Las medidas de similitud se utilizan para comparar diferentes tipos de datos, lo que es fundamentalmente importante para la clasificación de patrones, el agrupamiento y los problemas de recuperación de información. Las relaciones de similitud generalmente han estado dominadas por modelos geométricos en los que los objetos están representados por puntos en un espacio euclidiano. La semejanza se define como "Tener las mismas o casi las mismas características", mientras que la distancia métrica se define como "La propiedad creada por el espacio entre dos objetos o puntos" (Santisteban & Carcamo, 2015, p. 23).

- **DISTANCIA EUCLÍDEA**

La geometría de distancia euclidiana es el estudio de la geometría euclidiana basada en el concepto de distancia. Esto es útil en varias aplicaciones donde los datos de entrada consisten en un conjunto incompleto de distancias y la salida es un conjunto de puntos en el espacio euclidiano que realizan esas distancias dadas. (Liberti, et al., 2014, p. 1).

$$d(A,B) = \sqrt{(X_B - X_A)^2 - (Y_B - Y_A)^2} \qquad (12\text{-}1)$$

Donde:

$\boldsymbol{X_B}$: representa al individuo en la variable X

$\boldsymbol{Y_B}$: representa al individuo en la variable Y

$\boldsymbol{X_A}$: media aritmética de la variable X

$\boldsymbol{Y_A}$: media aritmética de la variable Y

La distancia o métrica euclidiana tiene las siguientes propiedades (EcuRed, 2019, pp. 242-413).

- $d(A,B) \geq 0$
- $d(A,B) = d(B,A)$
- $d(A,A) = 0$
- $d(A,B) \leq d(A,C) + d(C,B)$
- $Si\ d(A,B) = 0, \Rightarrow\ A = B$

PASO 4: APLICAR EL MÉTODO DE CLÚSTER

Luego de haber encontrado los valores del coeficiente de distancia euclidiana para todos los pares de objetos, se puede construir una especie de mapa, conocido como árbol, en el cual se visualizará los grados de similitud entre todos los pares de objetos. Por tanto, la matriz de semejanza se convierte en un árbol usando un método de agrupamiento el cual contiene una serie de pasos que gradualmente "descomponen" la matriz de semejanza, reduciéndola de tamaño.

Un clúster es un conjunto de uno o más objetos que se dispone a llamar similares entre sí. Un grupo puede ser tan pequeño como un objeto, si no se dispone que se deba llamar a otros objetos similares a ese objeto. O puede ser tantos como todos los objetos en la matriz de datos, si se disponen para llamarlos a todos similares entre sí.

Al comienzo de los pasos del método de agrupamiento, cada objeto es considerado como parte de un grupo separado; entonces, como se dispone el ejemplo con cinco objetos, habrá cinco grupos al principio. Cada paso de agrupación fusionará los dos clústeres más similares que existen al comienzo del paso, reduciendo la cantidad de clústeres en uno.

Por medio del análisis de conglomerados permitirá comprender cómo se pueden determinar estos puntos de compromiso óptimo para cada grupo de objetos. En resumen, este paso avanza al principal objetivo agrupar y representar mediante el árbol para visualizar las agrupaciones (Oksanen, 2012, p. 4).

• MÉTODO DEL AMALGAMIENTO COMPLETO O COMPLETE LINKAGE

Este método de agrupamiento se basa en el criterio de escoger la distancia máxima entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la máxima de las distancias (López, 2018). Se lo obtiene por:

$$d\big((i,j),(k)\big) = \ máx\{\big(d(i,k),(j,k)\big)\} \qquad (13\text{-}1)$$

PASO 5: REORDENAR LOS DATOS Y LAS MATRICES DE SEMEJANZA

De manera arbitraria se ordenan los números de identificación a los objetos y, por lo tanto, no hay manera de saber si los objetos más similares estaban o no adyacentes entre sí en estas matrices. Sin embargo, los objetos más similares tienden a tener posiciones adyacentes en la parte inferior del árbol. Por lo tanto, se debe reorganizar los datos y las matrices de semejanza es cuestión de escribir los números según su orden que estos mantienen en el árbol.

En la matriz de datos reorganizados, los pares de objetos que son más similares tienden a estar cerca de la diagonal; la similitud entre pares tiende a disminuir alejándose de la diagonal. Esta organización ayuda a resaltar las similitudes entre objetos en los datos mismos. Si bien el árbol sugiere similitudes, se requiere volver a la matriz de datos original y la matriz de semejanza original, y reorganizarlas para confirmar la naturaleza de las similitudes.

PASO 6: CALCULAR EL COEFICIENTE DE CORRELACIÓN COFENÉTICA

Mediante el árbol que permite visualizar las similitudes, no es exactamente como la matriz de datos que se representa. Se debe verificar qué tan bien el árbol representa la matriz de datos.

El coeficiente de correlación cofenética, r_{xy} permite conocer una respuesta parcial: qué tan bien mide el árbol y la matriz de semejanza "dicen lo mismo". Si bien sería mejor si se pudiera comparar el árbol (la salida de un análisis de conglomerados) con la matriz de datos (la entrada de un análisis de conglomerados), esta comparación no es factible. Así que se hace lo siguiente mejor y se compara el árbol y la matriz de semejanza.

Para calcular r_{xy} se comienza convirtiendo el árbol en su equivalente, la matriz cofenética. Luego se hace la comparación entre la matriz de semejanza y la matriz cofenética. El árbol y su matriz cofenética son dos formas diferentes de una misma cosa. Para afectar la comparación, se separa la matriz de semejanza y la matriz cofenética en dos listas, para cada una para la matriz, donde cada fila corresponde a la misma. Para calcular el coeficiente de correlación conocida como "producto-momento" de Pearson, r_{xy} entre las listas $X\ e\ Y$. Su fórmula es:

$$r_{x,y} = \frac{\sum xy - (\frac{1}{n})\sum x \sum y}{\{[\sum x^2 - \left(\frac{1}{n}\right)(\sum x)^2][\sum y^2 - \left(\frac{1}{n}\right)(\sum y)^2]\}^{1/2}} \qquad (14\text{-}1)$$

Donde:

***x*:** valores de la matriz de distancias iniciales

y: valores de la matriz cofenética

n: tamaño de la muestra

Esto es un poco menos que una correlación perfecta $r_{xy} = 1$; pero está muy por encima del punto de no correlación $r_{xy} = 0$; y está fuera del rango de una concordancia negativa $-1 \leq r_{xy} \leq 0$. Este valor permitirá indicar que tan bien el árbol representa la estructura de similitud entre objetos inherente a la matriz de semejanza.

El coeficiente r_{xy} es un índice que indica cuánto distorsiona el método de agrupamiento la información en su entrada para producir su salida. No se puede establecer pautas exactas que definan cuánta distorsión es tolerable. Sin embargo, la mayoría de los campos aceptarían que cuando r_{xy} es grande se aproxima a la correlación perfecta, cuando se considera un valor de 0,7 o más, la distorsión no es grande. (Romesburg, 2004, pp. 2-27).

La distancia estimada entre dos puntos es el nivel en el que se fusionan en el dendrograma, o la altura de la raíz. Un buen método de agrupamiento reproduce correctamente las diferencias reales. La distancia estimada a partir de un dendrograma se denomina distancia cofenética, que hace eco de los orígenes de la agrupación jerárquica en la taxonomía numérica antigua (Oksanen, 2012, p. 4).

1.1.23. Dendrograma de similitud

El formato de dendrograma conciso permite la comparación visual de varias clasificaciones diferentes, como los producidos por diferentes medidas de similitud o algoritmos de agrupamiento a diferencia de la mayoría ordenaciones, los dendrogramas representan la separación de clases y la compacidad directamente en el original unidades de la medida de similitud elegida (Van, 1997, pp. 370-388).

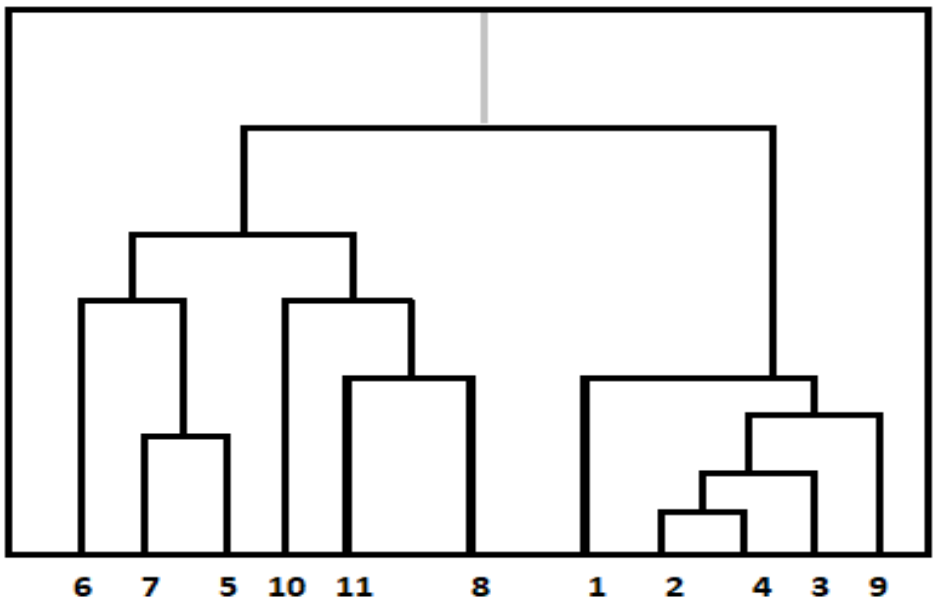


Ilustración 6-1: Representación de un dendrograma

Fuente: (Forina, et al., 2002, p. 15)

1.1.24. Proyección sobre variables de interpretación

La significación estadística de los conglomerados puede ser evaluada mediante una prueba de Fisher. A pesar de que se han sugerido otros métodos, pero no se ha encontrado una solución satisfactoria debido a la estructura muy diferente de los grupos encontrados en problemas prácticos.

En la interpretación se basa en que se compara los grupos sugeridos por el dendrograma con información externa, no utilizada para calcular el dendrograma. Con frecuencia, el problema real sugiere que existen algunas categorías y el paso de interpretación es comparar el número y la composición de los conglomerados con estas categorías. El índice de categoría es una variable discreta. En otros casos la interpretación utiliza una o más variables continuas.

La proyección directa del dendrograma sobre una o dos variables de interpretación no presenta dificultades matemáticas. En el caso de una variable de interpretación, la abscisa es la variable de interpretación y la ordenada es la similitud. En el caso de dos variables de interpretación el dendrograma simula un espacio tridimensional.

En lugar del procedimiento inverso del dendrograma parte de los objetos y continúa exactamente como el procedimiento aglomerativo. Cuando dos objetos se fusionan, su posición en la abscisa es la media de los valores de la variable de interpretación para los dos objetos (Forina, et al., 2002, p. 17).

1.2. Bases teóricas

Varios son los conceptos y percepciones acerca de los accidentes de tránsito y su tipología por lo que seguido se presentan definiciones que permitirán establecer claridad de entendimiento en la investigación.

1.2.1. Accidente o siniestro de tránsito

Un siniestro de tránsito es un suceso fortuito o acción inconsciente, derivado de una o varias causas que suceden en vías, caminos o espacios destinados al uso público o privado, a consecuencia se observan individuos muertos, personas con lesiones de distinto nivel de gravedad, adicional durante mencionadas colisiones se generan daños significativos en vehículos, vías o infraestructura, de igual manera pérdidas económicas e incluso consecuencias en el ámbito de salud según el caso (Córdova & Paucar, 2014, p. 2).

Los accidentes de tránsito comúnmente causan lesiones catastróficas a los pasajeros porque muchos de estos vehículos carecen de equipo de seguridad estándar. Según el Instituto Nacional

de Estadísticas y Censos (INEC) y la cita tomada de la Agencia Nacional de Transito (ANT, 2021) existen dos tipos de siniestros

1.2.2. Tipos de siniestros

- **Siniestro mortal:** en este suceso se produce al menos el fallecimiento de una persona involucrada.
- **Siniestro no mortal:** en este suceso involucra a más de una víctima, pero sin dejar personas fallecidas.

1.2.3. Clases de Siniestros

Se requiere define los siguientes conceptos (INEC, 2010, como se citó en ANT, 2021):

- **Choque:** es cuando dos vehículos que se encuentran en movimiento se impactan, se detallan los siguientes:
- **Choque posterior o por alcance:** es el impacto que existe entre un vehículo y el vehículo que le antecede.
- **Choque frontal longitudinal:** dos vehículos tienen un impacto de frente, es decir formando una paralela.
- **Choque frontal excéntrico:** dos vehículos tienen un impacto de frente, los ejes longitudinales no coinciden en forma de una recta.
- **Choque lateral angular:** ocurre cuando en vehículo impacta de manera frontal contra la parte lateral de otro vehículo formando un ángulo diferente de 90 grados.
- **Choque lateral perpendicular:** ocurre cuando en vehículo impacta de manera frontal contra la parte lateral de otro vehículo formando un ángulo 90 grados.
- **Estrellamiento:** es cuando el vehículo en movimiento choca a otro vehículo que no está en movimiento u objeto fijo.
- **Pérdida de carril:** es cuando el vehículo sale de su carril perteneciente de circular.
- **Pérdida de pista:** es cuando el vehículo sale de la carretera de circulación.
- **Roce:** se denomina así a la fricción de las partes laterales de los vehículos dañando la carrocería de los mismos.
- **Roce negativo:** impacto lateral de los vehículos que circulan en sentidos iguales.
- **Roce positivo:** impacto lateral de los vehículos que circulan en sentidos diferentes.
- **Rozamiento:** es el rozamiento de forma lateral de un vehículo en movimiento con otro vehículo o un obstáculo que permanece estático.
- **Volcamiento:** este accidente es cuando el vehículo se invierte o cae lateralmente.

- **Volcamiento lateral:** el vehículo pierde la posición por una de sus laterales.
- **Volcamiento longitudinal:** la posición del vehículo se cambia con respecto al sentido de su eje longitudinal en cualquier medida o dirección generando moviendo de ángulos o giros completos.
- **Arrollamiento:** las ruedas del vehículo en movimiento pasan por encima del cuerpo de una de las víctimas.
- **Atropello:** es cuando vehículo en movimiento se impacta contra una víctima causando daños.
- **Caída del pasajero:** el pasajero pierde el equilibrio por alguna maniobra al subir o bajar desde el interior del vehículo o en la parte del estribo.
- **Colisión: s**e considera colisión cuando dos o más vehículos tienen un impacto entre sí.
- **Atípico:** es todo aquel evento o suceso que no se detalla en ninguno de los anteriores casos.

1.2.4. Víctimas involucradas

Se define los siguientes conceptos (INEC, 2010, como se citó en ANT, 2021):
Víctima: se considera a la persona que fallece o sufre una lesión tras ocurrir un siniestro.
Fallecido: se lo considera a la persona que muere al instante o después de 30 días del accidente.
Lesionado: es la persona que sobreviva más de 24 horas tras un siniestro y se encuentre con lecciones, estas personas requieren de tratamientos médicos.

1.2.5. Factores influyentes en los siniestros de tránsito

1.2.5.1. Exceso de velocidad

Esta actividad adherente al conductor aumenta las probabilidades de que se suscite un siniestro de tránsito, generando así que los daños humanos y materiales sean aún más graves. Los peatones y los ciclistas son los más vulnerables dentro del sistema vial; a nivel mundial solo un poco más de la mitad de los países consideran que el límite de velocidad debe ser 50 km/h en zonas urbanas (OMS, 2013, p. 6).

1.2.5.2. Conducción bajo los efectos del alcohol

La conducción bajo los efectos del alcohol incrementa la probabilidad de accidentes en todo el mundo y junto a ello la gravedad de las lesiones, con miras a mantener líneas de precaución los conductores deberían estar frente al volante con un valor máximo de 0,05 g/dl de alcohol para reducir el número de accidentes ocasionados por el alcohol (OMS, 2013, p. 6).

1.2.5.3. Cinturón de seguridad

La falta de uso de los cinturones de seguridad por parte de los ocupantes de un vehículo es determinante en una colisión fatal pues de ello depende el nivel de la lesión o la muerte del pasajero, con el uso del cinturón de seguridad se reduce entre el 40 % y 50 % del riesgo de una lesión mortal de todos quienes se encuentran en la parte delantera y entre el 25 % y 75 % de quienes se encuentran en la parte trasera (OMS, 2013, p. 6).

1.2.5.4. Casco de motociclista

El uso de cascos homologados cuando conducen motocicletas evita un 40 % el riesgo de muerte y un 70 % lesiones o traumatismos graves durante siniestros viales de lo contrario este representa uno de los factores más significativos de muerte alrededor del mundo (OMS, 2013, p. 6).

1.2.5.5. Asientos para niños

Estos asientos son también llamados sistemas de retención que protegen a niños pequeños y lactantes contra las lesiones importantes durante un accidente, por tal motivo más de la mitad de los países han aplicado leyes sobre el uso de sistemas de retención para niños (OMS, 2013, p. 6), (Menon, 2014, p. 1).

CAPÍTULO II

2. MARCO METODOLÓGICO

2.1. Tipo de la investigación

De acuerdo con el método de investigación fue mixto porque que se utilizó variables de tipo cualitativas y cuantitativas para el análisis del presente trabajo.

En cuanto al objetivo fue aplicada, ya que permitió proponer sugerencias para un problema social como lo son los siniestros de tránsito, de esta manera se podrá ayudar en la toma de decisiones para la disminución de casos.

Según el nivel de profundización en el objeto de estudio fue descriptiva y explicativa, ya que se detalla las características de las variables en estudio y se pretende resolver la problemática que consiste en el aumento de siniestros en el cantón.

Debido a la manipulación de variables se trabajó con un diseño no experimental porque la información procede de la fuente secundaria del banco de datos libres por parte de la Agencia Nacional de Tránsito del Ecuador quienes registran los eventos.

Con respecto al tipo de inferencia fue deductiva que permitió confirmar o refutar premisas, entre ellas la relación y similitud de los casos.

En cuanto al el periodo temporal fue transversal en el cual se considera los años 2015-2020 para el análisis de las variables del fenómeno en estudio (Sampieri, et al., 2014, pp. 15-634), (Patten & Newhart, 2017, pp. 18-352).

2.2. Diseño de la investigación

Se respaldó en un diseño no experimental ya que no existió manipulación deliberada de las variables en estudio y solo se observó la ocurrencia de los siniestros en su contexto natural en el periodo 2015-2020 (Escamilla, 2020, pp. 2-3).

2.2.1. Localización de estudio

El análisis de los siniestros de tránsito fue analizado en el cantón la Joya de los Sachas, la cual pertenece a la provincia de Orellana y forma parte de la Región Amazónica de la República de Ecuador, este territorio se caracteriza por encontrarse en las llanuras amazónicas, en donde predomina el bosque húmedo tropical la misma que acoge una diversidad de especies de flora y fauna silvestre, no obstante se aprovecha de esta gran ventaja para realizar actividades de turismo ecológico, agroturismo y turismo cultural que permite un gran turismo.

El cantón se encuentra limitado al norte con la Provincia de Sucumbíos, al sur con el Cantón Francisco de Orellana, al este con la Provincia de Sucumbíos y al oeste con el Cantón Francisco de Orellana (EcuRed, 2018, pp. 5-9).

Ilustración 1-2: Ubicación de la provincia de Orellana

Fuente: Wikipedia, 2021

Se localiza su posición según sus coordenadas de latitud -76.8571063, y de longitud -0.3013626. A través de datos emitidos por el Censo del 2010, el 77,9 % de su población reside en el área rural; se caracteriza por ser una población joven, ya que el 51,5% de los habitantes son menores de 20 años. (EcuRed, 2018, pp. 5-9).

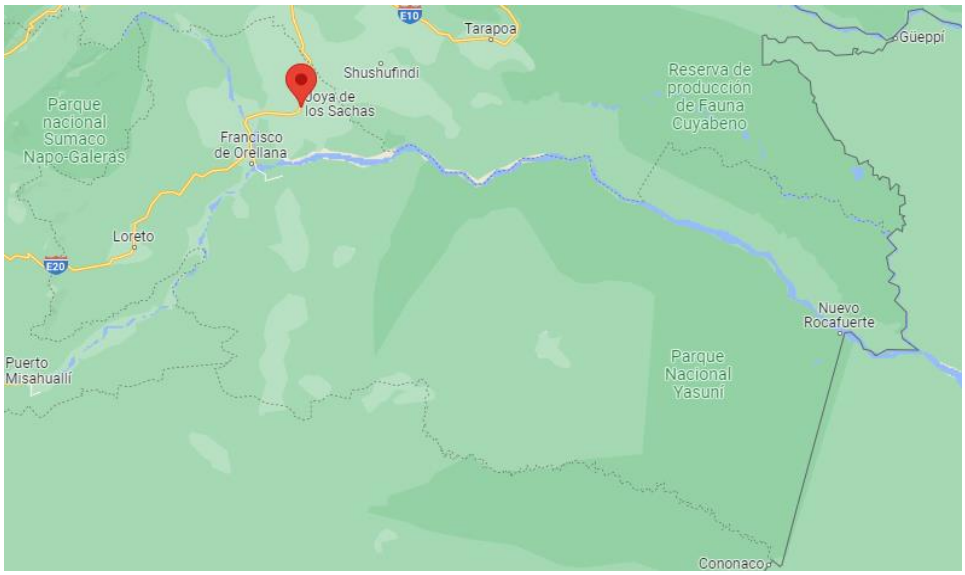


Ilustración 2-1: Ubicación del cantón La Joya de los Sachas.

Fuente: Google Maps, 2022.

La corteza terrestre de este cantón comúnmente se caracteriza por ser firme constituyendo así una zona plana, donde sobresale el suelo de tipo arcilloso de estructurada por capas delgadas y de ferruginosas. La altitud de este cantón no sobrepasa los 270 m.s.n.m. Generalmente, se puede constatar año tras año que existe mayor abundancia de lluvia en los meses de mayo a noviembre. Se caracteriza por su clima húmedo tropical, además de las constantes brisas, intensa evaporación y altas temperaturas con nubosidad media de seis octavos y su temperatura normal es de 28ºC, con climas no menor a 18ºC y logra alcanzar una temperatura máxima de 34ºC.

El cantón La Joya de los Sachas (cabecera cantonal) pertenece como la única parroquia urbana, mientras que San Sebastián del Coca, Pompeya, Enokanqui, San Carlos, Unión Milagreña, Lago San Pedro, Rumipamba, Tres de Noviembre son parroquias rurales. (EcuRed, 2018, pp. 5-9).

2.2.2. Población de estudio

Se consideró como parte de la población de estudio a todos los siniestros de tránsito suscitados en el cantón La Joya de los Sachas durante el período 2015 y 2020.

2.2.3. Método de muestreo

El catálogo de eventos fue adquirido directamente de la Agencia Nacional de Tránsito por lo que no se aplicó ninguna técnica de muestreo.

2.2.4. Tamaño de la muestra

El estudio se realizó con el conteo de 73 siniestros de tránsito registrados por la Agencia Nacional de Tránsito durante el período 2015 y 2020.

2.2.5. Técnica de recolección de datos

La información adquirida formó parte de una fuente secundaria de información publicada en el banco de datos abiertos de la Agencia Nacional de Tránsito del Ecuador, por tanto, no se necesitó emplear alguna técnica para la recolección de la información.

2.2.6. Identificación de variables

Para el análisis propuesto se ha considerado un total de siete variables, de las cuales cinco variables son cualitativas y dos cuantitativas, las que se detallan a continuación:
Las variables cualitativas para el análisis fueron:

- Día
- Zona
- Feriado
- Causa
- Clase

Las variables cuantitativas para el análisis fueron:

- Número de fallecidos
- Número de lesionados

2.2.7. Modelo estadístico

Los modelos estadísticos utilizados en la investigación fueron: un estudio descriptivo para conocer las características generales de las variables: zona, día, presencia de feriado, causa, clase, número de fallecidos y número de lesionados en los siniestros, la selección de variables adecuadas mediante análisis de variables redundantes para evidenciar el aporte de la información presente en la matriz de datos, un análisis de correspondencias para detectar asociaciones utilizando las variables cualitativas (día, zona, feriado, causa, clase) y un análisis de conglomerados para la detección de grupos con similitudes entre ellos utilizando las variables numéricas (número de siniestros, número de lesionados).

2.3. Variables en estudio

2.3.1. Operacionalización de variables

Para cada variable en estudio se detalla el nombre de la misma, se indica en qué consiste o describe cada una, además el tipo de variable y la escala de medición a la que pertenece.

Tabla 1-2: Operacionalización de variables

Nombre de la variable	Descripción	Tipo de variable	Escala de medición
Feriado	Afirmación de la presencia o no de feriados decretados por el gobierno nacional	Cualitativa dicotómica	Nominal
Zona	Identificación de la zona ya sea urbana o rural donde ocurre el siniestro de tránsito.	Cualitativa dicotómica	Nominal
Causa	Listas de señales de tránsito que han sido desobedecidas por los peatones o vehículos y que hayan contribuido al desenlace de un siniestro de tránsito	Cualitativa politómica	Nominal
Clase	Acciones que generan daños puntuales en los vehículos como motivo de causas de siniestros de tránsito.	Cualitativa politómica	Nominal
Día	Tiempo que emplea la Tierra en dar una vuelta sobre sí misma.	Cualitativa politómica	Nominal
Número de fallecidos	Conteo de fallecidos reportados en un siniestro de tránsito.	Cuantitativa discreta	Razón
Número de lesionados	Conteo de lesionados reportados en un siniestro de tránsito.	Cuantitativa discreta	Razón

Fuente: ANT, 2021.

Realizado por: Carrión, Andrea, 2022.

CAPÍTULO III

3. MARCO DE RESULTADOS Y DISCUSIÓN DE RESULTADOS

A partir de los datos observados de los siniestros de tránsito. Se muestra a continuación, los resultados del análisis descriptivo de datos, para identificar zonas de alto riesgo de accidentes de tránsito del cantón La Joya de los Sachas, 2015- 2020. Luego, se desarrolla un análisis de variables redundantes entre las variables número de fallecidos y lesionados de siniestros de tránsito. Después, un análisis de correspondencias para identificar asociaciones entre las variables. Finalmente, se aplica un análisis de conglomerados para identificación de grupos.

3.1. Análisis Descriptivo

3.1.1. Variables Cualitativas

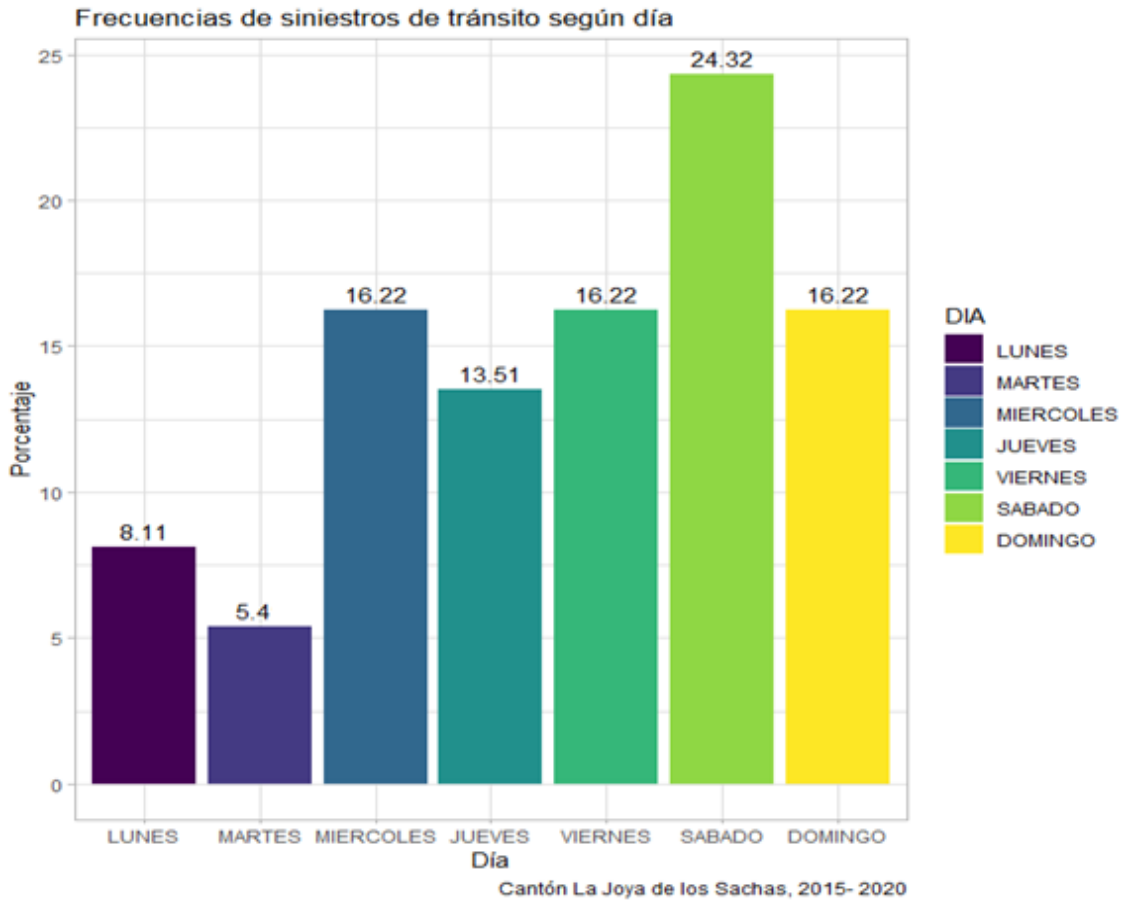


Ilustración 1-3: Gráfica de frecuencias de siniestros de tránsito según Día del siniestro del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

El porcentaje promedio de los siniestros de tránsito por día en la Joya de los Sachas fue de 14,28% y fueron los días miércoles, viernes, sábados y domingos aquellos que generaron una mayor accidentabilidad en la matriz, el elevado índice puede deberse a que los días miércoles y viernes existen ferias de abastecimiento mientras que los sábados y domingos por tratarse de fines de semana la gente frecuenta bares y discotecas, consumo de alcohol lo que genera perdida de la conciencia e incrementa la posibilidad de accidentes de tránsito.

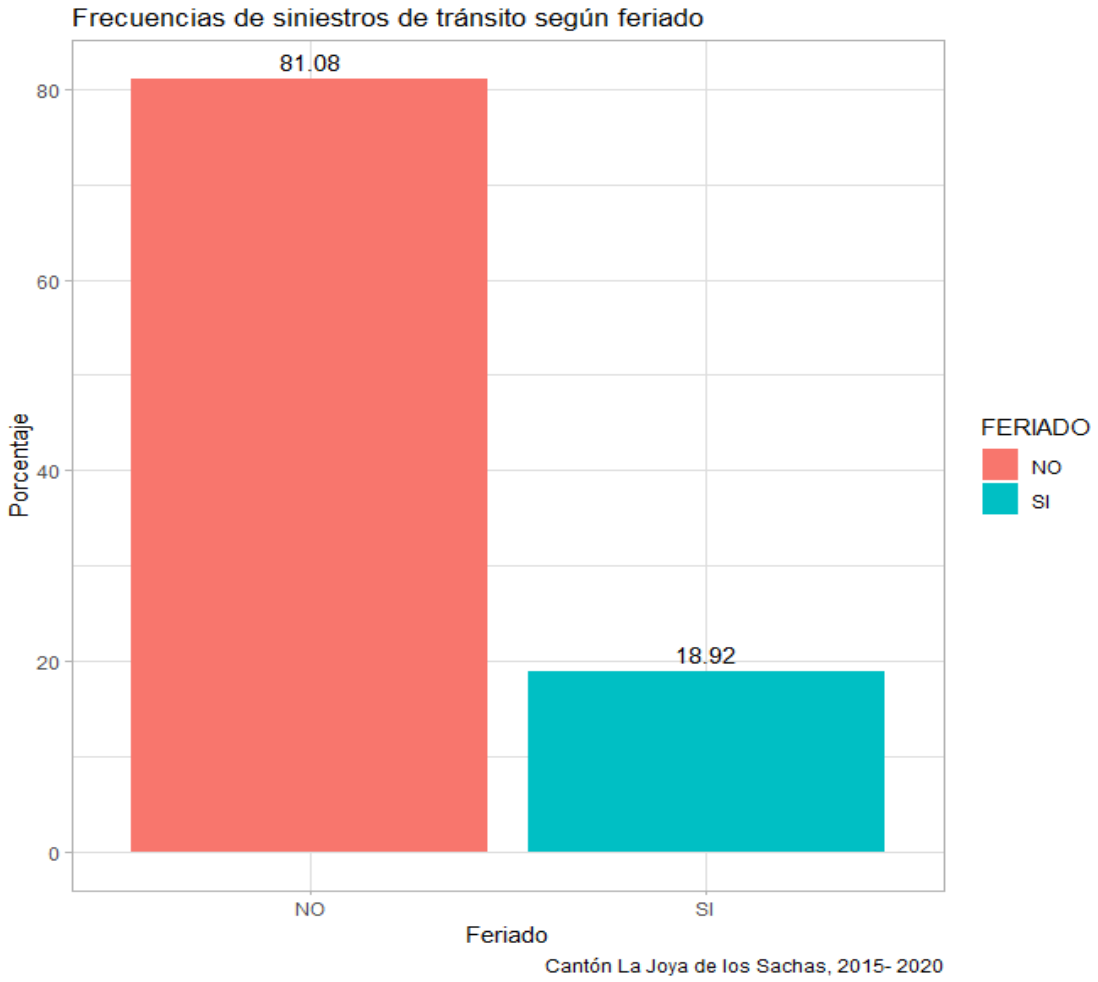


Ilustración 2-3: Gráfica de frecuencias de siniestros de tránsito según Feriado del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

Con relación al porcentaje de siniestros durante los feriados el 81,08% se registraron en ausencia de una festividad, apenas el 18,92% sucedió cuando se desarrollaba algún festejo.

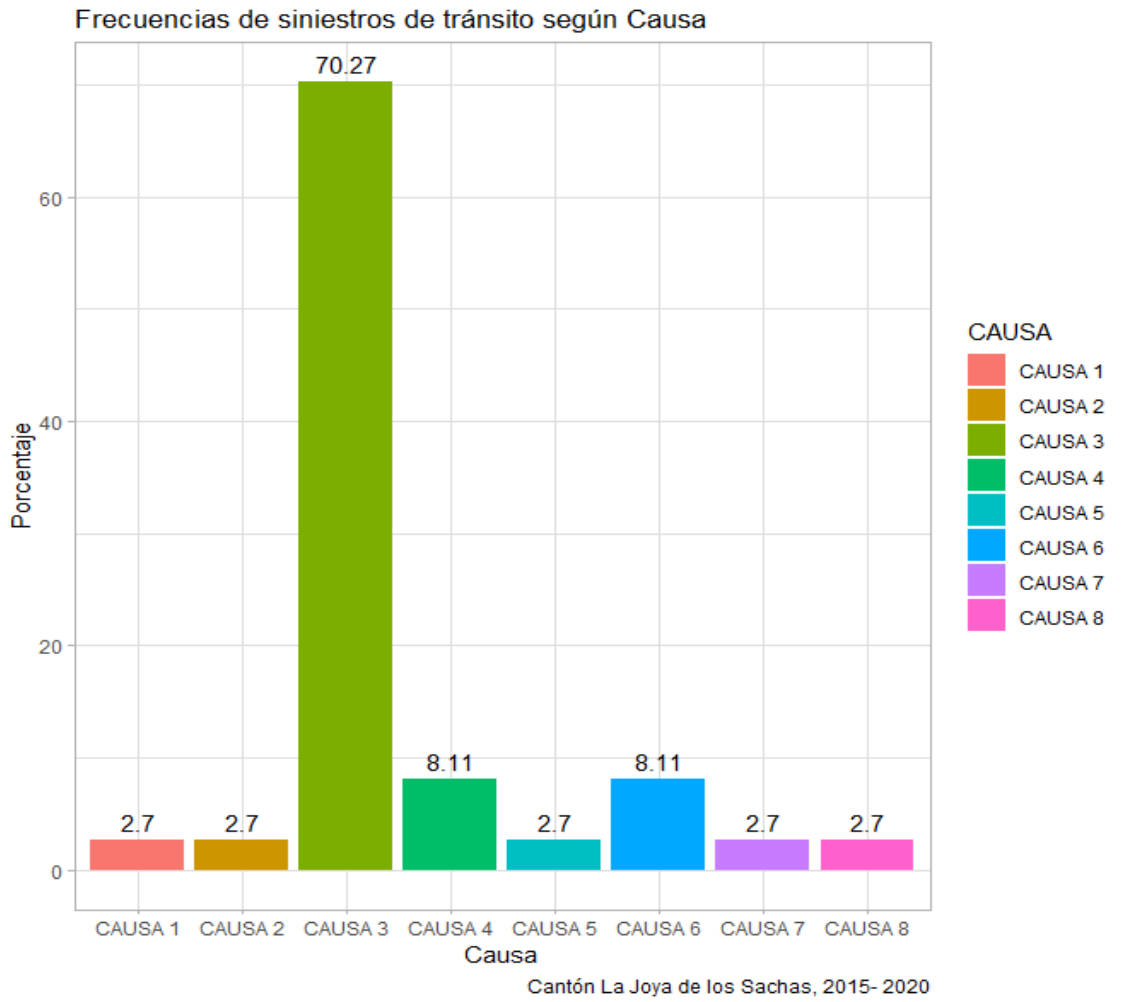


Ilustración 3-1: Gráfico de frecuencias de siniestros de tránsito según Causa del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

En torno a la causa de los siniestros de tránsito el 70,27% representó a las causas fusionadas en el grupo 3 como es el caso de conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor), seguido por un 8,11% tanto en grupo 4 como en el 6. En el caso del grupo 4 se registraron los siniestros reportados a causa de conducir vehículos superando los límites máximos de velocidad y en el grupo 6 por no respetar las señales reglamentarias de tránsito, y para los grupos 1, 2, 5, 7, 8 se tiene para cada uno el 2.7%, donde los siniestros registrados del primer grupo son por caso fortuito o fuerza mayor, condiciones ambientales y/o atmosféricas, grupo 2 por conducir bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos y por influencia de las condiciones ambientales y/o atmosféricas, grupo 5 debido a las malas condiciones de la vía y/o configuración (iluminación y diseño) y no ceder el derecho de vía o preferencia de paso al peatón, grupo 7 no transitar por las aceras o zonas de seguridad destinadas para el efecto y no guardar la distancia

entre vehículos, y finalmente el grupo 8 por causa de presencia de agentes externos en la vía, además por la presencia de agentes externos en la vía.

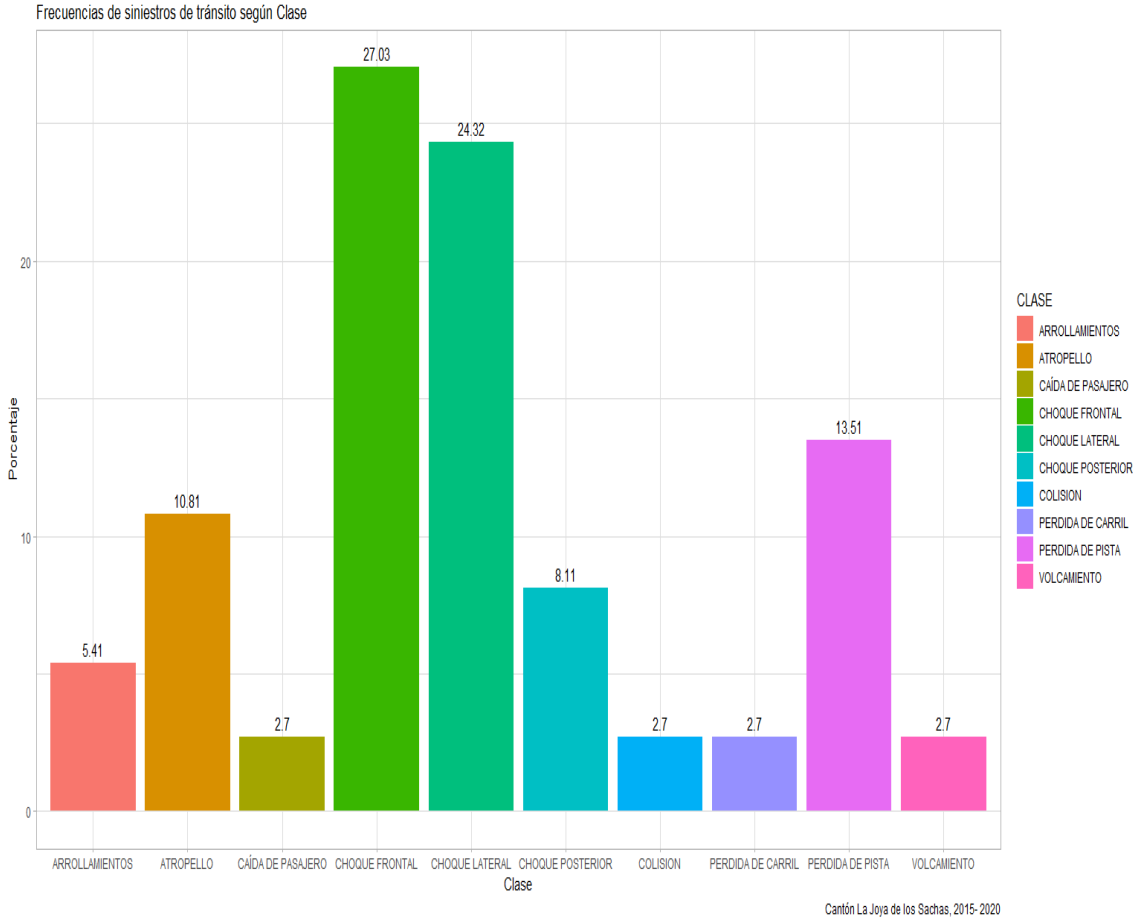


Ilustración 4-3: Gráfica de frecuencias de siniestros de tránsito según la clase de siniestros del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

En cuanto a la clase de siniestro de tránsito el 51,35% corresponde a choques frontales y laterales, seguidos por un 24,32% de pérdida de pistas y atropellos, con porcentajes inferiores al 6% se mostraron los arrollamientos, caídas de pasajeros de transportes públicos o privados, colisiones, pérdidas de carril y volcamientos.

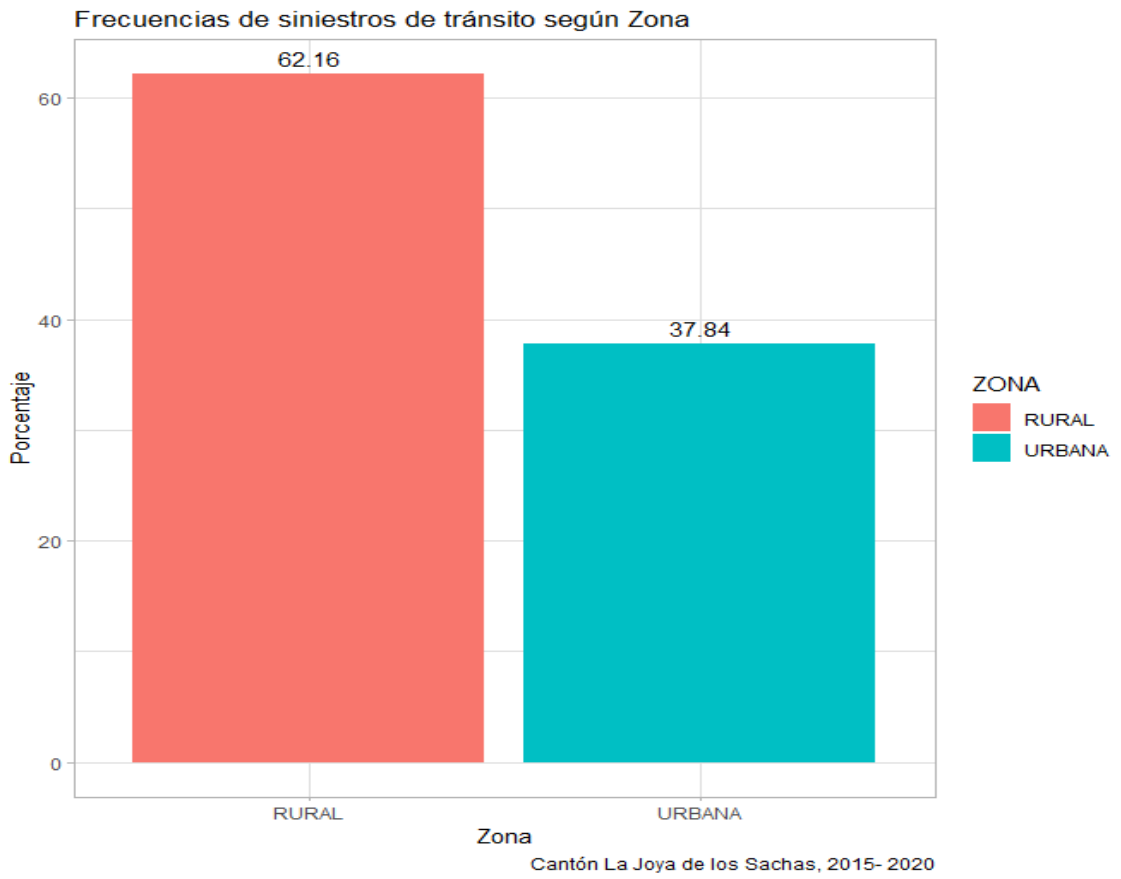


Ilustración 5-3: Gráfica de frecuencias de siniestros de tránsito según la zona del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

El 62,16% de los siniestros de tránsito en la Joya de los Sachas se ubicaron en zonas rurales pues los caminos que conectan a las vías urbanas o a los propios caminos secundarios se encuentran en mal estado no por falta de atención de las autoridades sino por los cambios ambientales que mantiene la zona en los últimos años.

3.1.2. Variables Cuantitativas

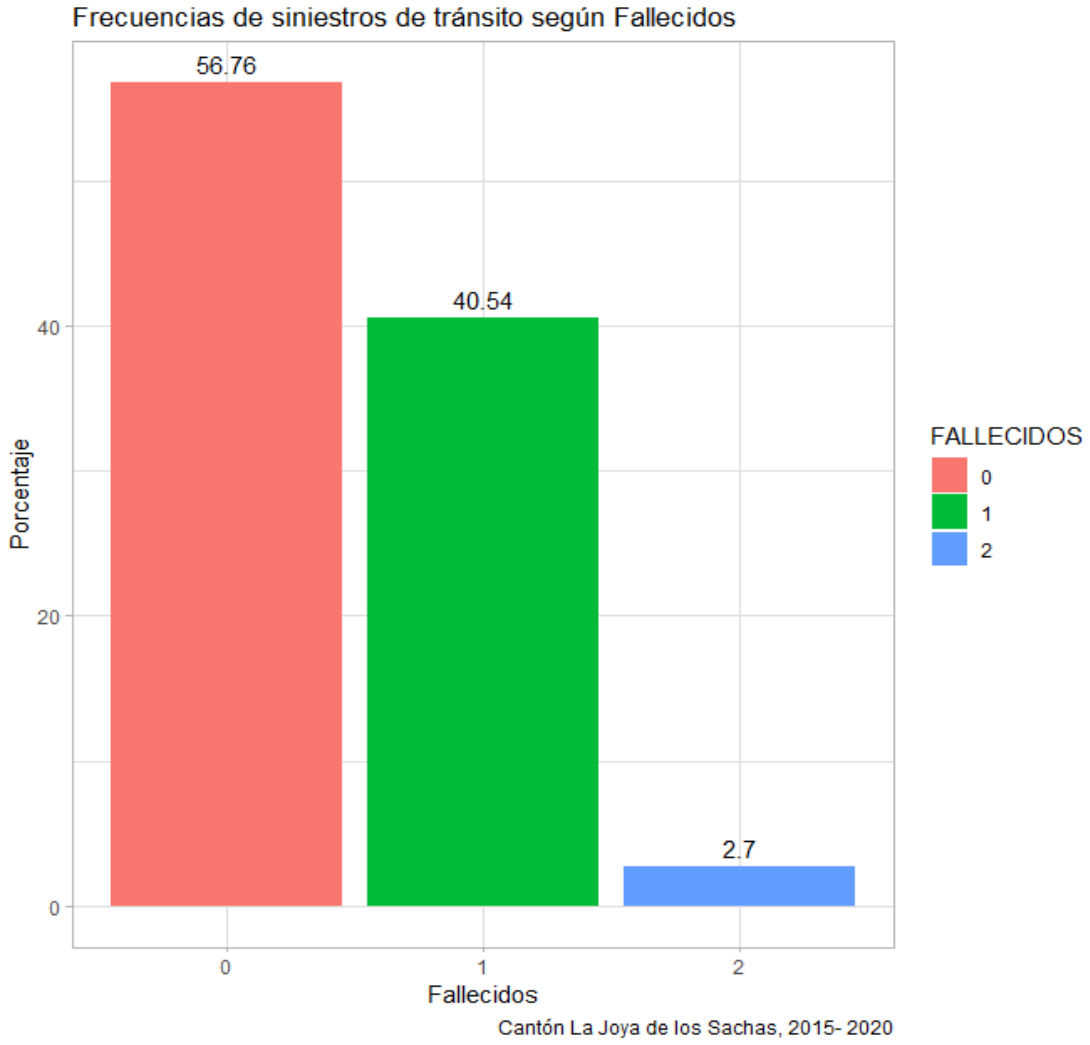


Ilustración 6-3: Gráfica de frecuencias de siniestros de tránsito según el número de fallecidos del siniestro del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

El 56,76% no reportó víctimas mortales a causa de un siniestro de tránsito en la Joya de los Sachas, el 40,54% por su parte advirtió de la presencia de un fallecido, y se evidencia un 2.7% de dos fallecidos, a pesar de que las víctimas mortales son pocas las autoridades y ciudadanía misma busca mediante campañas de concientización generar mayor cuidado al momento de conducir, pues la vida de las personas es irremplazable.

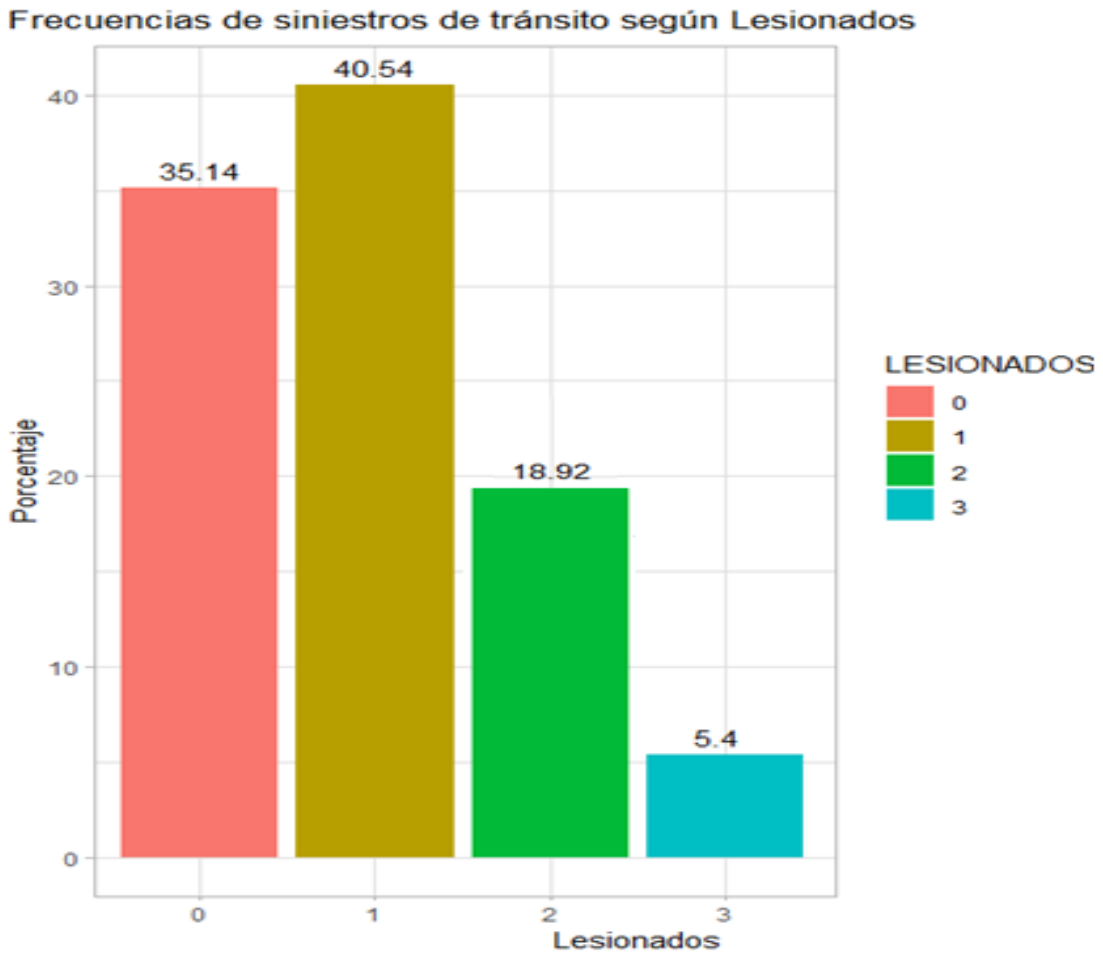


Ilustración 7-3: Gráfica de frecuencias de siniestros de tránsito según número de lesionados del siniestro del cantón La Joya de los Sachas, 2015- 2020

Realizado por: Carrión, Andrea, 2022.

En relación a los siniestros de tránsito que registraron lesionados durante los accidentes en La Joya de los Sachas, el 40,54% evidenció la presencia de un lesionado, el 35,14% ningún lesionada, el 18,92% 2 lesionados y el 5.4% de 3 lesionados.

Tabla 1-3: Medidas de tendencia central

Medida de tendencia central	Número de fallecidos	Número de lesionados
Media	0.469697	0.9090909
Mediana	0	1
Moda	0	1

Realizado por: Carrión, Andrea, 2022.

En el cantón la Joya de los Sachas durante el periodo 2015 al 2020, el promedio del número de fallecidos es aproximadamente de 1 persona. Además, el promedio del número de lesionados es aproximadamente de 1persona. El 50% de los siniestros están por debajo de 0 fallecidos y el 50%

de los accidentes ocurridos están por debajo de 1 persona lesionada. Por otra parte, el número de fallecidos en siniestros con más frecuencia es de 0 personas. Además, el número de lesionados con más frecuencia es de 1 persona.

Mediante la **Tabla 1-3**, se puede conocer la relación que tiene las variables con respecto a las tres centrales medidas de tendencia. Para el caso de la métrica número de fallecido: la distribución tiene una asimetría positiva, ya que el valor de la moda y mediana son menores a la media. Para el caso de la métrica número de lesionado: la distribución tiene una asimetría negativa, ya que el valor de la media es menor al valor de mediana y moda.

Tabla 2-3: Medidas de dispersión

Medida de dispersión	Número de fallecidos	Número de lesionados
Desviación estándar	0.5607604	0.8897238
Varianza	0.3144522	0.7916084
Rango	2	3

Realizado por: Carrión, Andrea, 2022

En el cantón la Joya de los Sachas durante el periodo 2015 al 2020, la desviación del número de fallecidos con respecto a su media es de 1 persona, mientras que para el número de lesionados la desviación es de 1 personas con respecto a su media. Además, se pudo obtener que el rango por accidentes de tránsito para el número de fallecidos es de 2 personas, mientras que para el número de lesionados es de 3 personas.

Tabla 3-3: Medidas de posición

Medidas de posición		Número de fallecidos	Número de lesionados
Cuantiles	**25%**	0	0
	50%	0	1
	75%	1	1
Cuartiles	**25%**	0	0
	50%	0	1
	75%	1	1

Realizado por: Carrión, Andrea, 2022.

En el cantón la Joya de los Sachas durante el periodo 2015 al 2020, se pudo obtener para la variable número de fallecidos: el 25% de los casos están bajo 0 fallecidos, de igual manera el 50% de los datos están entre 0 fallecidos, y el 50% de los siniestros cuentan con 1 persona fallecida. Para la variable número de fallecidos: el 25% de los casos están bajo 1 lesionados, de igual manera el 50% de los datos están entre 1 lesionados, y el 50% de los siniestros cuentan con 1 persona lesionada.

3.2. Detección de datos atípicos

Previo al análisis, es importante comenzar representado los datos en un diagrama de cajas y bigotes, de esta manera se analiza el comportamiento que tienen y a su vez se puede identificar datos que se sobresalgan e influyan en la información.

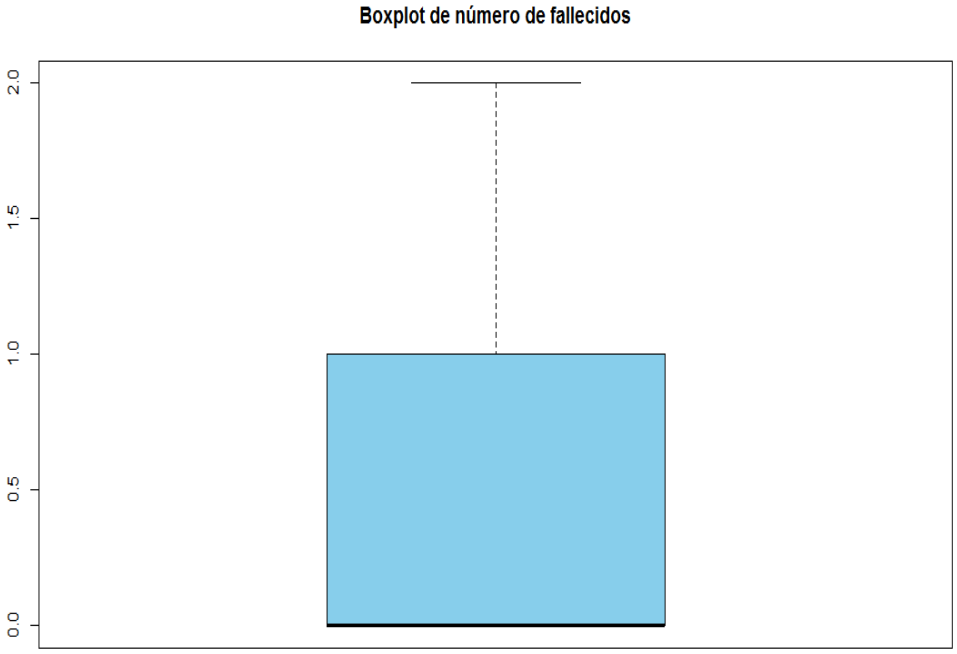


Ilustración 8-3: Gráfica de boxplot de la variable número de fallecidos antes del análisis de datos atípicos

Realizado por: Carrión, Andrea, 2022.

En esta gráfica se realizó un box-plot para la variable número de fallecidos de los datos, se puede ver que el valor de la mediana se sitúa hacia el lado izquierdo, lo que nos indica que los datos presentan asimetría hacia la derecha. En este caso, no se ha logrado identificar algún valor que sobresalga de los bigotes del box-plot, de esta manera se puede indicar que no existen datos que afecten directamente a posteriores cálculos.

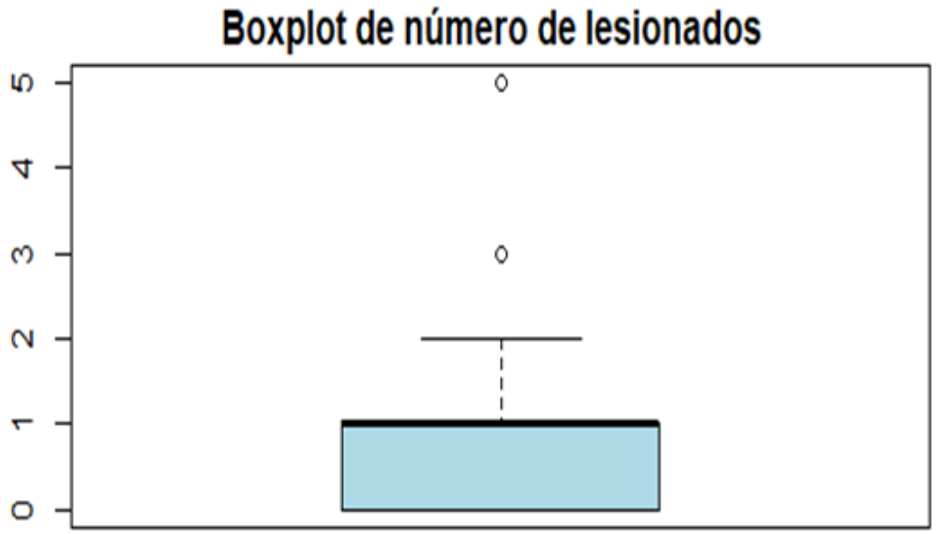


Ilustración 9-3: Gráfica de boxplot de la variable número de lesionados antes del análisis de datos atípicos

Realizado por: Carrión, Andrea, 2022.

En esta gráfica se realizó un box-plot para la variable número de lesionados de todos los datos originales, se puede ver que el valor de la mediana se sitúa en el centro de los datos, lo que nos indica que los datos se distribuyen simétricamente. En este caso, se ha logrado identificar algún valor que sobresalga de los bigotes del box-plot, de esta manera se puede indicar que se obtiene que un dato atípico el cual se encuentra bastante alejado de los demás, por lo tanto, es importante detectarlo y comprobar si es realmente influyente y evitar que perjudique en los resultados. Por ello, se realiza la siguiente técnica.

En el análisis de datos atípicos se consideran las variables: número de fallecidos y lesionados, para detectar valores anómalos y evitar distorsión en los análisis. Por lo tanto, se ha calculado las distancias de Cook para todas las observaciones. De esta manera, se considera el criterio que indica 4 veces el promedio de las distancias calculadas para los 73 datos de la muestra de estudio.

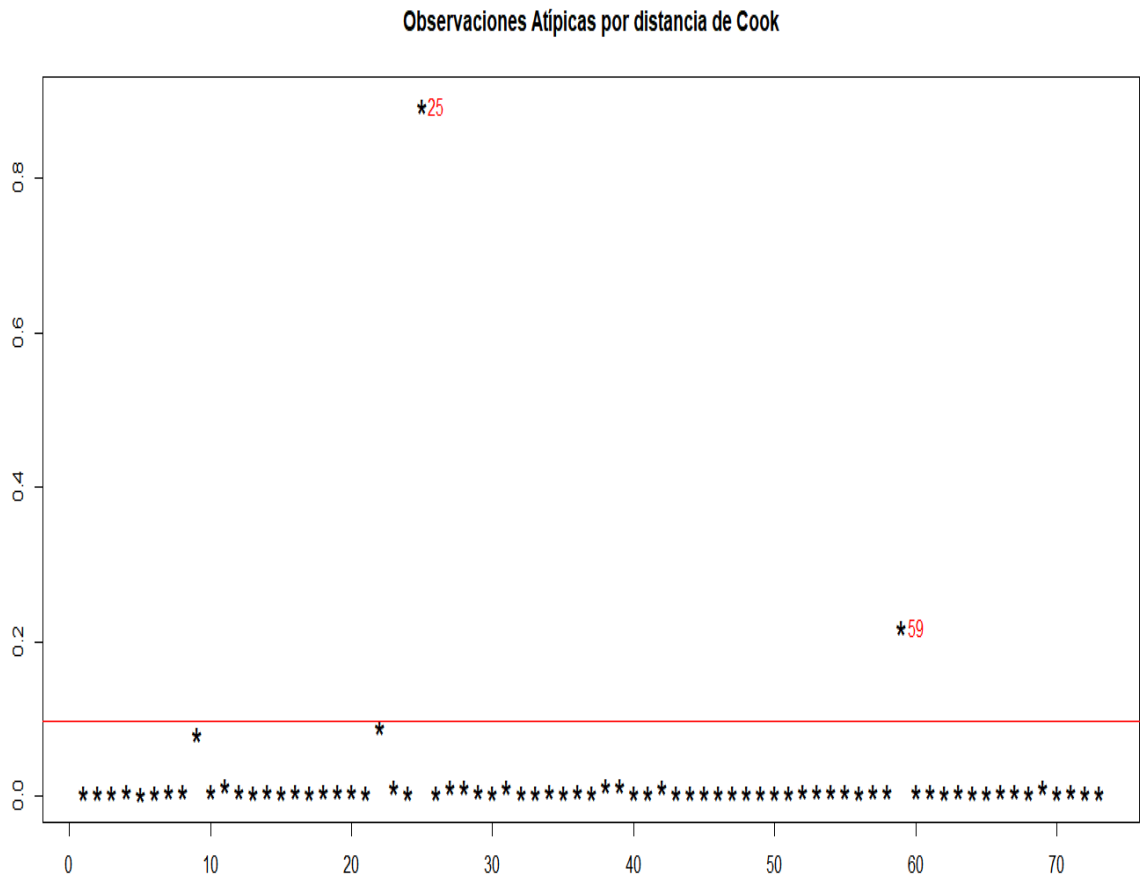


Ilustración 10-3: Gráfica de la primera corrida de datos atípicos según distancias Cook

Realizado por: Carrión, Andrea, 2022.

Una vez conocido el valor que permite delimitar a los individuos como atípicos, se logra verificar el gráfico los puntos que sobrepasan al valor de distancia calculada igual a 0.09631064. Mediante este método gráfico se identificó que el punto 25 y 59 son influyentes, por lo tanto, se retira la información de las filas que se ha logrado localizar.

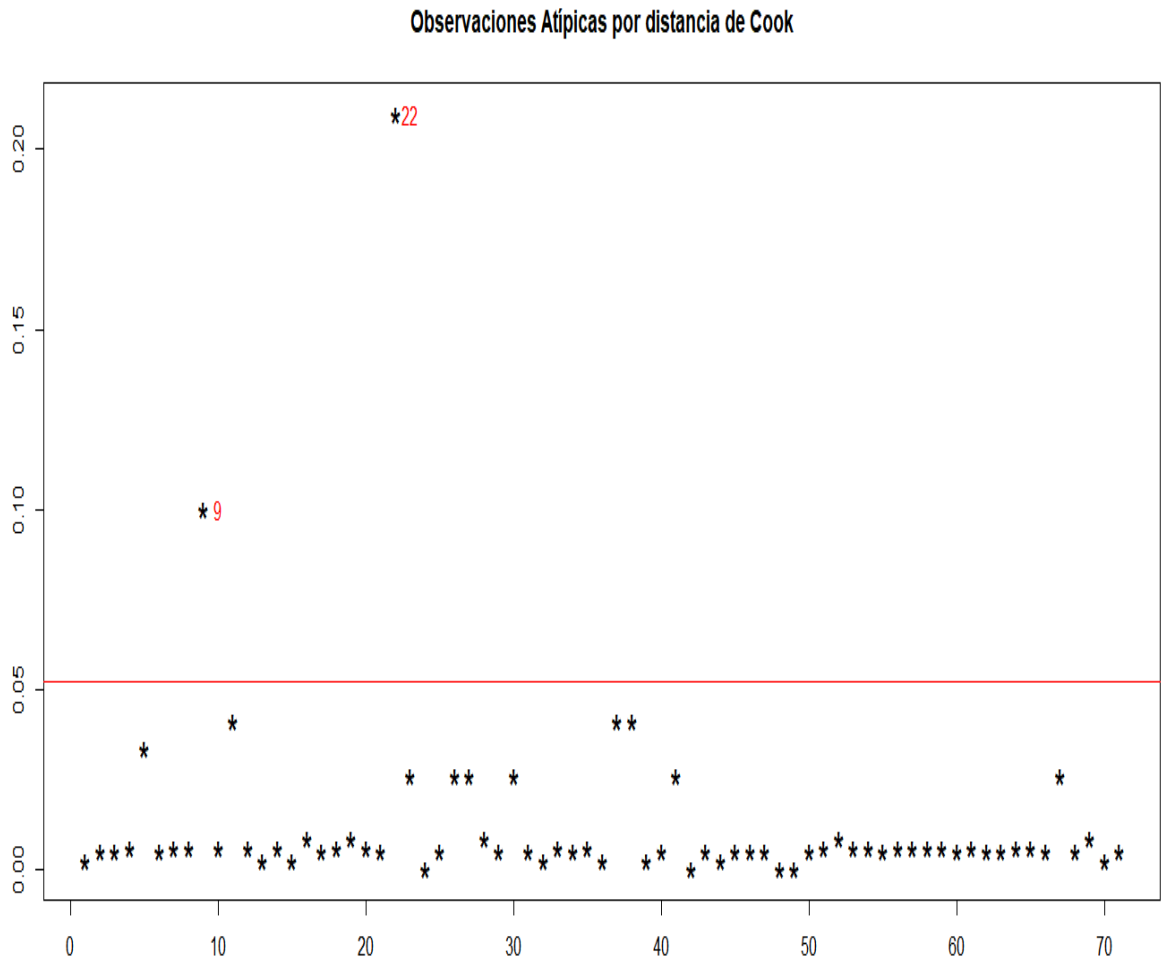


Ilustración 11-3: Gráfica de la segunda corrida de datos atípicos según distancias Cook

Realizado por: Carrión, Andrea, 2022.

Luego de retirar las filas 25 y 29 detectados como influyentes. Se recalcula las distancias de la nueva matriz de datos y se obtiene el valor de distancia igual 0.05227635 y se visualiza las observaciones que sobrepasen a dicho valor. Se detecta que los puntos 9 y 22 son los que sobrepasan al nuevo límite, por lo tanto, se procede a retirarlos ya que son observaciones atípicas.

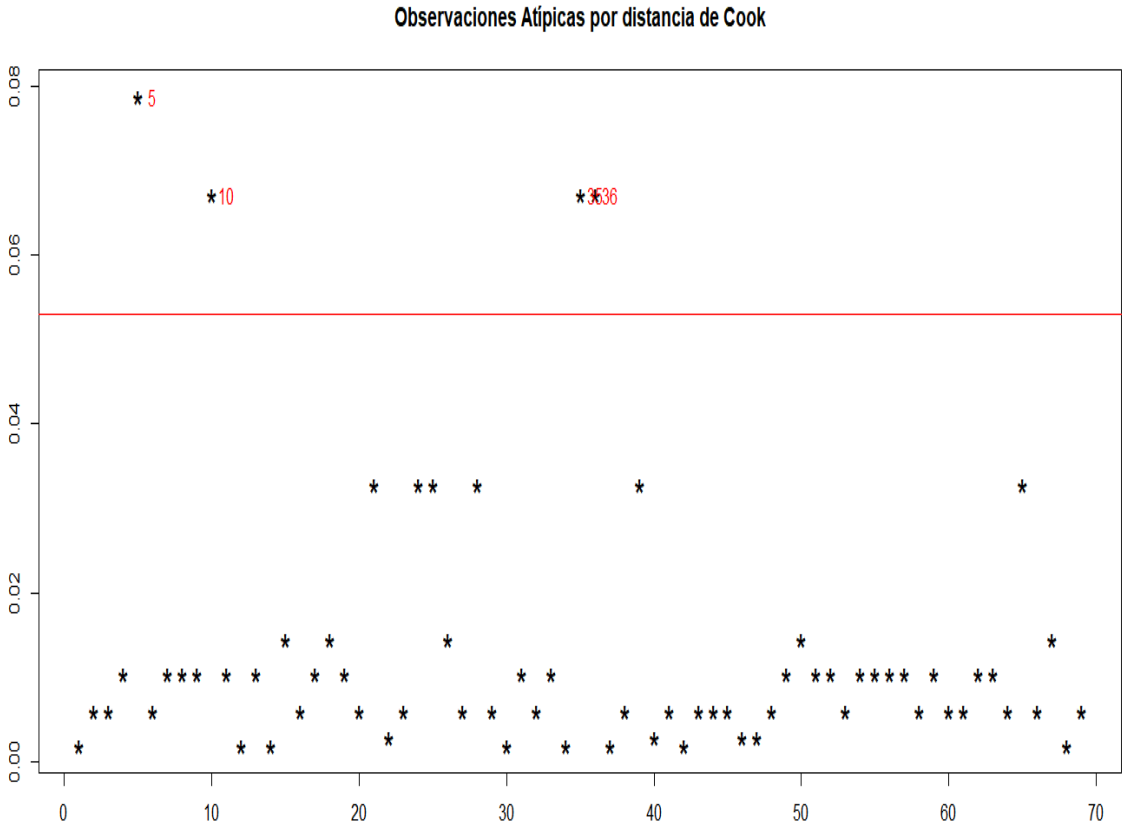


Ilustración 12-3: Gráfica de la tercera corrida de datos atípicos según distancias Cook

Realizado por: Carrión, Andrea, 2022.

Luego de retirar las filas 9 y 22 que conforman a valores anómalos. Se recalcula el nuevo valor a limitar los datos restantes, en esta ocasión la distancia toma el valor de 0.05303702. En esta gráfica se detecta que los valores 5, 10, 35 y 36, son valores influyentes ya que sobre pasan la distancia Cook establecida, por lo tanto, se procede a retirarlos.

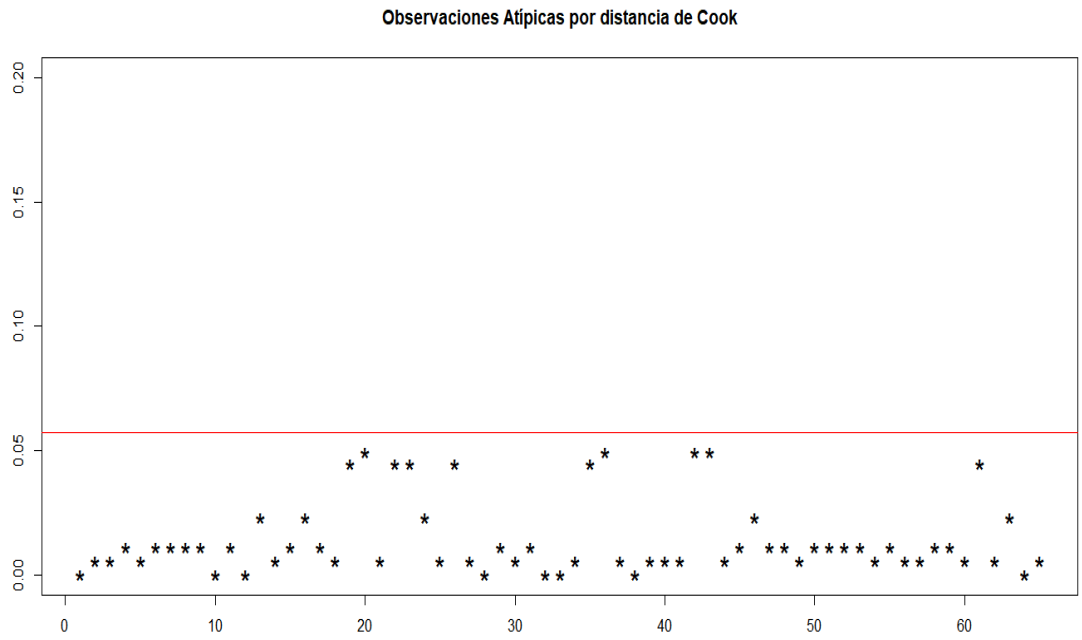


Ilustración 13-2: Gráfica de la cuarta corrida de datos atípicos según distancias Cook

Realizado por: Carrión, Andrea, 2022.

Posteriormente de haber retíralo los valores atípicos, se calcula la nueva distancia de Cook que determina un valor de distancia igual a 0.05732768, en la gráfica se visualiza los nuevos valores que sobrepasen a este limitante y no se obtienen nuevas observaciones para declararlos como influyentes. Por consiguiente, no se efectúan nuevas corridas. De este modo, el proceso de detección de puntos influyentes de las variables en estudio termina.

Finalmente, se indica en la siguiente tabla los 8 datos de las filas o posiciones que resultaron como atípicos: 25, 59, 9, 22, 5, 10, 35 y 36, los mismos que se han retirado de la base original, de esta manera se evita distorsiones en posteriores cálculos.

Tabla 4-3: Datos atípicos retirados de la base de datos

Número de Fallecidos	Número de Lesionados
0	4
1	5
2	5
2	5
0	4
1	5
1	4
1	5

Realizado por: Carrión, Andrea, 2022.

Una vez terminado el análisis de detección de datos atípicos y eliminar a los mismos, se presenta nuevamente un boxplot para de esta manera constatar que las distribuciones de variables se visualizan de mejor manera, y que no existe presencia de puntos anómalos.

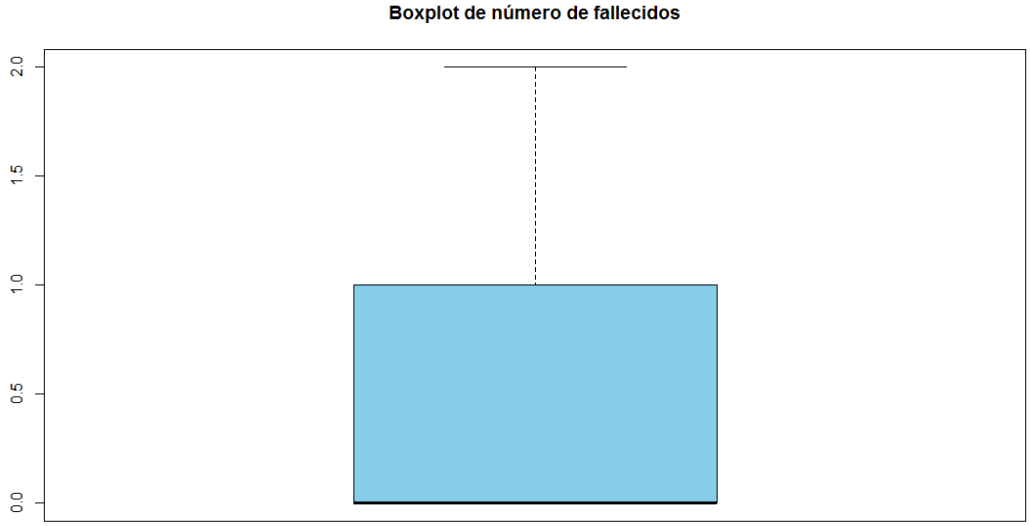


Ilustración 14-3: Gráfica de boxplot de la variable número de fallecidos posterior del dato atípico

Realizado por: Carrión, Andrea, 2022.

En esta gráfica se realizó un box-plot para la variable número de fallecidos de los datos, se puede ver que el valor de la mediana se sitúa hacia el lado izquierdo, lo que nos indica que los datos presentan asimetría hacia la derecha. En este caso, no se ha logrado identificar algún valor que sobresalga de los bigotes del box-plot, de esta manera se puede indicar que no existen datos que afecten directamente a posteriores cálculos.

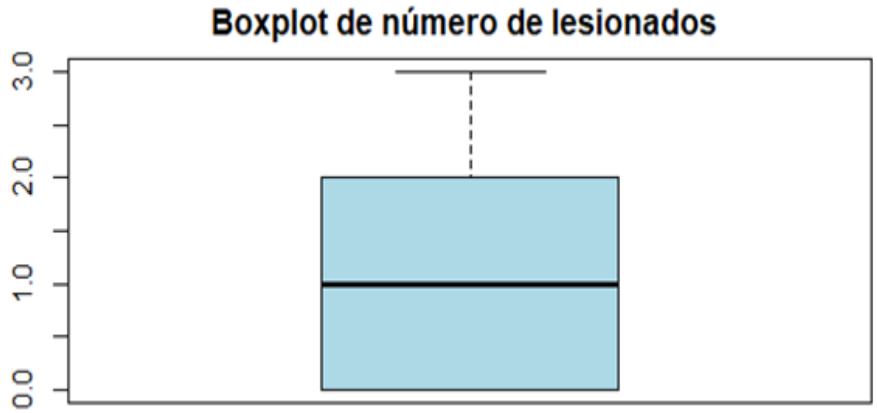


Ilustración 15-3: Gráfica de boxplot de la variable número de lesionados

Realizado por: Carrión, Andrea, 2022.

En esta gráfica se realizó un box-plot para la variable número de lesionados, se puede ver que el valor de la mediana se sitúa hacia el lado derecho, lo que nos indica que los datos presentan asimetría hacia la izquierda. En este caso, no se ha logrado identificar algún valor que sobresalga de los bigotes del box-plot, de esta manera se puede indicar que no existen datos que afecten directamente a posteriores cálculos.

En definitiva, mediante los box-plots después del análisis de datos atípicos se identifica que no existen valores que influyan en la base de datos.

3.3. Análisis de variables redundantes

3.3.1. Gráfico de dispersión y matriz de correlación

En primer lugar, se estable las hipótesis de interés bajo las cuales se trabaja para este análisis:

H_0: No existe una relación lineal entre las variables número de fallecidos y lesionados.

H_1: Existe una relación lineal entre las variables número de fallecidos y lesionados.

Luego, se realizó un gráfico de dispersión entre las variables número de fallecidos y lesionados en siniestros de tránsito. También, se realizó una matriz de correlación de las variables y prueba de hipótesis.

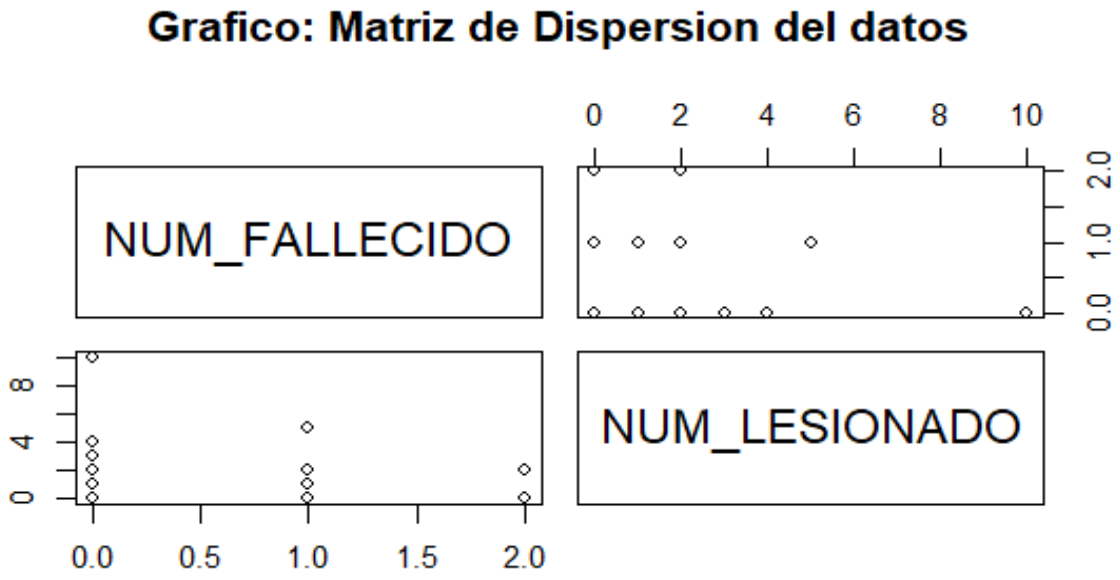


Ilustración 16-3: Gráfica de matriz de dispersión entre número de fallecido y lesionado

Realizado por: Carrión, Andrea, 2022.

Tabla 5-3: Matriz de correlación de Pearson

	Fallecidos	Lesionados
Fallecidos	1.00	-0.31
Lesionados	-0.31	1.00

Realizado por: Carrión, Andrea, 2022.

Tabla 6-3: Pruebas de hipótesis de correlación Pearson (p-valor)

	Fallecidos	Lesionados
Fallecidos	-	0.06
Lesionados	0.06	-

Realizado por: Carrión, Andrea, 2022.

En general, se observó en la **Tabla 5-3** y **Tabla 6-3** que no existe una relación significativa entre las variables números de Fallecidos y Lesionados los cuales presentan valores de p superiores a 0.05). Por lo tanto, no existen variables redundantes.

3.4. Análisis de correspondencias múltiples

Para este estudio se presentan 5 variables cualitativas (día, zona, feriado, causa y clase de siniestro), por lo tanto, se opta por análisis de correspondencias múltiples. En las siguientes gráficas se visualizan las categorías de cada una de ellas para una mejor identificación.
Para la variable día se tienen 7 categorías las cuales son:

Tabla 7-3: Codificación de la variable día

CODIFICACIÓN	CATEGORÍAS
1	DOMINGO
2	JUEVES
3	LUNES
4	MARTES
5	MIÉRCOLES
6	SÁBADO
7	VIERNES

Realizado por: Carrión, Andrea, 2022.

Para la variable Zona se indica dos categorías y entre ellas se tiene:

Tabla 8-3: Codificación de la variable Zona

CODIFICACIÓN	CATEGORÍAS
8	RURAL
9	URBANA

Realizado por: Carrión, Andrea, 2022.

Para la variable Feriado se tiene las siguientes opciones:

Tabla 9-3: Codificación de la variable Feriado

CODIFICACIÓN	CATEGORÍAS
10	NO
11	SI

Realizado por: Carrión, Andrea, 2022.

Dentro de las opciones de la variable Causa se tiene las siguientes codificaciones:

Tabla 10-3: Codificación de la variable Causa

CODIFICACIÓN	CATEGORÍAS
12	Caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.,)
13	Condiciones ambientales y/o atmosféricas (niebla, neblina, granizo, lluvia)
14	Conduce bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos
16	conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor)
17	Conducir vehículo superando los límites máximos de velocidad
18	Daños mecánicos previsibles
19	Dejar o recoger pasajeros en lugares no permitidos
20	Mal estacionado (el conductor que detenga o estacione vehículos en sitios o zonas que entrañen peligro, tales como zona de seguridad, curvas, puentes, túneles, pendientes)
21	Malas condiciones de la vía y/o configuración, (iluminación y diseño),
22	No ceder el derecho de vía o preferencia de paso al peatón
24	No guardar la distancia lateral mínima de seguridad entre vehículos
25	No mantener la distancia prudencial con respecto al vehículo que le antecede
27	No respetar las señales reglamentarias de tránsito (pare, ceda el paso, luz roja del semáforo, etc)
28	No transitar por las aceras o zonas de seguridad destinadas para el efecto
30	Presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc,)

Realizado por: Carrión, Andrea, 2022.

Finalmente, para la variable Clase se ha codificado de la siguiente manera:

Tabla 11-3: Codificación de la variable Clase

CATEGORÍAS	CODIFICACIÓN
31	Arrollamientos
32	Atropellos
33	Caída de pasajero
34	Choque frontal
35	Choque lateral
36	Choque posterior
37	Colisión
38	Estrellamiento
39	Pérdida de carril
40	Pérdida de pista
41	Volcamientos

Realizado por: Carrión, Andrea, 2022.

En base a las anteriores codificaciones y facilitando de esa manera su interpretación, se obtiene la siguiente gráfica:

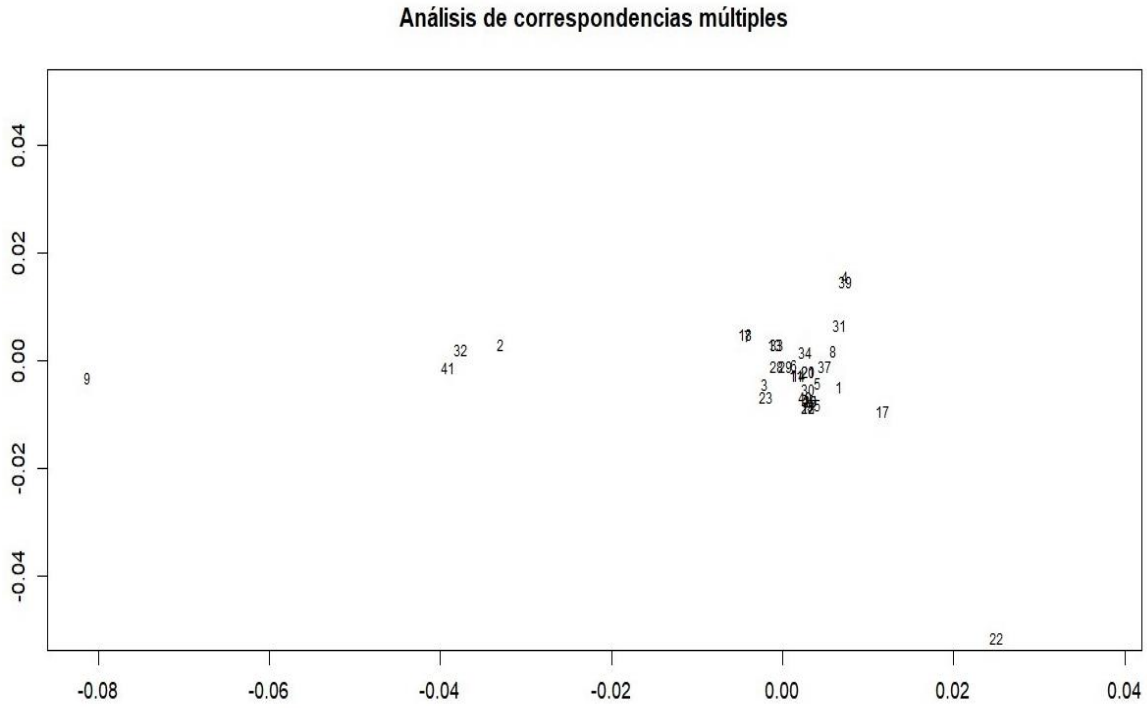


Ilustración 17-3: Gráfica de análisis de correspondencias múltiples

Realizado por: Carrión, Andrea, 2022.

En la gráfica se visualiza la nube de puntos que representan a las modalidades de las variables en estudio, se considera que existe mayor fuerza de asociación a todas las categorías que se encuentre alejadas del punto de origen y todas las modalidades cercanas al origen no tienen fuerza de asociación. En el caso 32 está lejanos al origen al igual que la categoría 41, por tanto, se puede decir que existe una fuerte asociación entre las categorías entre los siniestros de clase atropello y

volcamiento que ocurrieron en el cantón La Joya de los Sachas durante el periodo 2015-2020. De igual manera, se pude considerar el caso 2 que son los días jueves que presenta cercanía o distancia similar a siniestros de causa los atropellos (32), al igual para el caso 41 referente a los siniestros debido a volcamientos. Además, la modalidad 9 que hace referencia los siniestros ocurridos en zonas urbanas se encuentra bastante distante del origen y no presenta similitud con ninguna otra categoría, esto se debe a que cerca de esta no se encuentra otra modalidad representada. Por otra parte, se tiene a la causa 22 de cuyos siniestros ocurrieron por no ceder el derecho de vía o preferencia de paso al peatón bastante alejado del origen y sin otra modalidad a la cual agruparse. Lo mismo ocurre para la modalidad 26 que conlleva a los siniestros que se provocaron por no mantener la distancia prudencial con respecto al vehículo que le antecede.

Se tiene que las categorías más cercanas entre sí presentan una alta asociación entre ellas, como se refleja en los casos 8 y 37 que consisten en zonas rurales donde ocurrieron siniestros de clase colisión, además presenta una similitud moderada con respecto al caso 1 que se deben a siniestros que ocurrieron en días lunes, de igual manera presenta una similitud moderada respecto a la modalidad 31 que corresponde a los siniestros por arrollamientos. Para las modalidades 7 y 18 se observa que la representación de sus distancias son las mismas por lo que se sobreponen entre sí, esto nos indica que existe una excelente asociación entre los días viernes y daños mecánicos previsibles que ocasionan un accidente. Al igual en el caso 17 de los siniestros que ocurrieron por conducir un vehículo superando los límites máximos de velocidad.

Entre las siguientes modalidades de los siniestros que ocurrieron y cuyas características se especifican como: 6 día sábado, 11 presencia de feriado, 20 mal estacionamiento por parte del conductor que detenga o estacione vehículos en sitios o zonas que entrañen peligro, tales como zona de seguridad, curvas, puentes, túneles, pendientes, existe una asociación entre ellas.
Se observa una gran similitud en términos de frecuencias entre las modalidades:10 sin feriado, 12 caso fortuito o fuerza mayor (explosión de neumático nuevo, derrumbe, inundación, caída de puente, árbol, presencia intempestiva e imprevista de semovientes en la vía, etc.,), 13 condiciones ambientales y/o atmosféricas (niebla, neblina, granizo, lluvia), 14 conduce bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos, 16 conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor), 21 malas condiciones de la vía y/o configuración, (iluminación y diseño), 25 no mantener la distancia prudencial con respecto al vehículo que le antecede, 27 no respetar las señales reglamentarias de tránsito (pare, ceda el paso, luz roja del semáforo, etc.),29 no transitar por las aceras o zonas de seguridad destinadas para el efecto, 33 caída de pasajero, 34 choque frontal, 35 choque lateral, 36 choque posterior, 37 estrellamiento, 40 perdida de pista, todas estas clases se sobreponen entre sí en el gráfico o están muy próximas entre sí evidenciando la alta relación entre ellas.

De igual manera, se tiene casos que están muy cercanos entre sí como: la modalidad 3 y 23 que representa similitud entre los siniestros ocurridos en el día lunes y la causa de accidente de no guardar la distancia lateral mínima de seguridad entre vehículos. Para el par de modalidades 4 y 39, existe relación entre los siniestros que se presentan en los días martes y los mismos que se clasifican como clase de pérdida de carril, estos mismo presentan una menor similitud a la clase 19 y 28, dejar o recoger pasajeros en lugares no permitidos, no transitar por las aceras o zonas de seguridad destinadas para el efecto. Otro caso, se tiene a 5 y 30 donde los siniestros ocurridos en día miércoles y siendo su causa por presencia de agentes externos en la vía (agua, aceite, piedra, lastre, escombros, maderos, etc.,) presentan similitud entre ellos.

Existe una asociación negativa entre las categorías opuestas en el origen como se visualiza en las categorías ubicados en el primer cuadrante y su opuesto de origen que son las modalidades del tercer cuadrante, al ser opuestos presentan una asociación negativa. De igual manera, las modalidades que se encuentran en el cuadrante dos presentarán y los del cuadrante cuatro tendrán una asociación negativa. Tomando en cuenta el criterio anterior, dentro de cada cuadrante se observará que tan cercano o lejano se encuentran del origen para así conocer que tan mayor o menor es la fuerza de asociación entre las categorías.

3.5. Análisis de conglomerados

Para llevar a cabo este análisis, se consideran las variables cuantitativas en estudio (número de fallecidos y numero de lesionados) además la técnica seleccionada para agrupamiento de clúster es el Método de linkage completo o amalgamamiento completo. De esta manera, se procede con el cálculo de las distancias euclídeas y se obtiene la siguiente gráfica para visualizar los grupos.

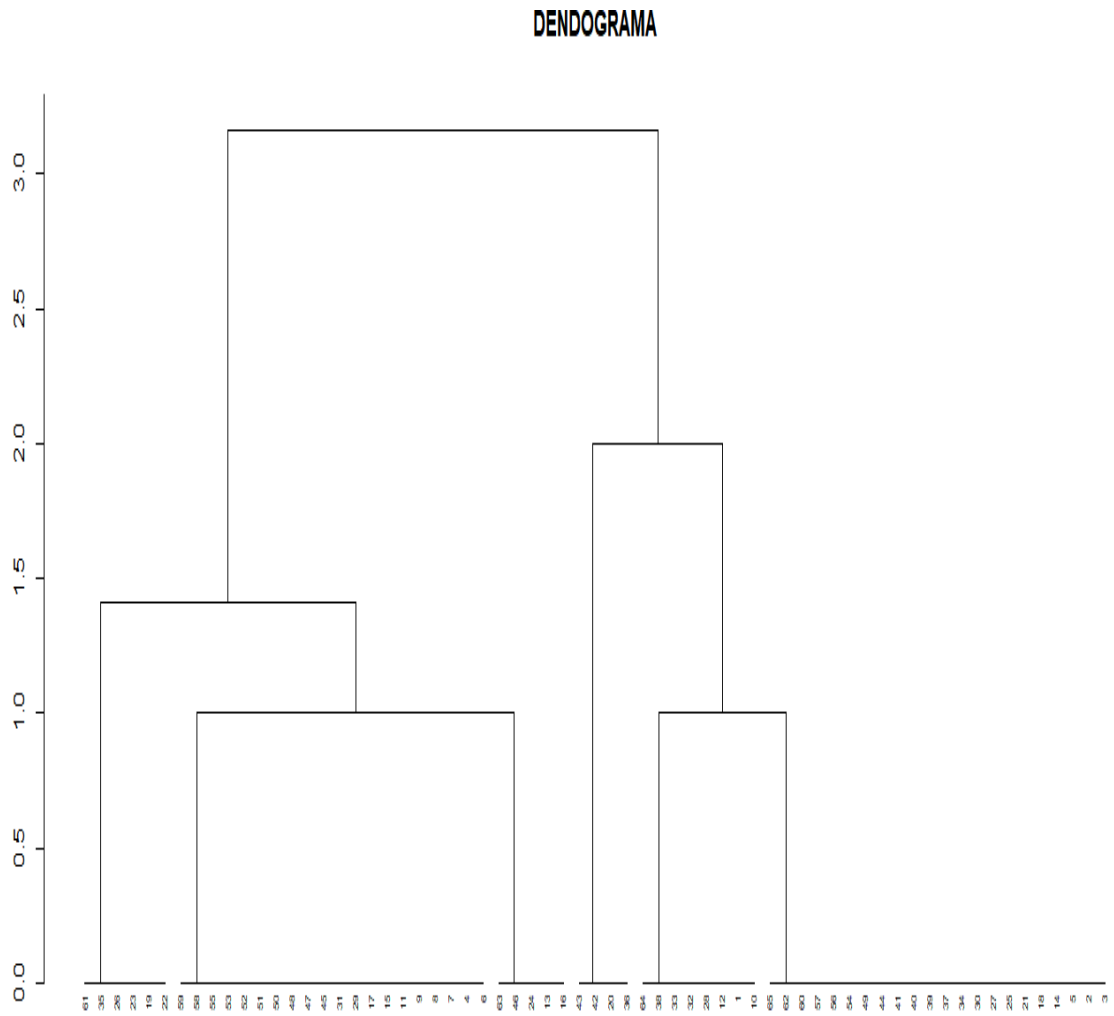


Ilustración 18-3: Dendrograma para detectar el número de clústeres

Realizado por: Carrión, Andrea, 2022.

A partir del dendrograma, se traza una línea horizontal con la finalidad de dividir los enlaces iniciales para identificar el número de clústers que se necesita formar, y se obtiene los grupos los cuales se forman basándose en la distancia máximas entre sus componentes individuales, dando como resultado un total de dos clústeres conformados de la siguiente manera:

Tabla 12-3: Conjunto de datos para el clúster 1

CLÚSTER 1		
SINIESTRO	**NÚMERO DE FALLECIDO**	**NÚMERO DE LESIONADO**
1	0	2
2	0	1
3	0	1
5	0	1
10	0	2
12	0	2
14	0	1

18	0	1
20	0	3
21	0	1
25	0	1
27	0	1
28	0	2
30	0	1
32	0	2
33	0	2
34	0	1
36	0	3
37	0	1
38	0	2
39	0	1
40	0	1
41	0	1
42	0	3
43	0	3
44	0	1
49	0	1
54	0	1
56	0	1
57	0	1
60	0	1
62	0	1
64	0	2
65	0	1

Elaborado por: Carrión, Andrea, 2022.

A continuación, se detalla los siniestros que pertenecen al grupo 2

Tabla 13-3: Conjunto de datos para el clúster 2

CLÚSTER 2		
SINIESTRO	NÚMERO DE FALLECIDO	NÙMERO DE LESIONADO
4	1	0
6	1	0
7	1	0
8	1	0
9	1	0
11	1	0
13	1	1
15	1	0
16	1	1
17	1	0
19	0	0
22	0	0
23	0	0
24	1	1

26	0	0
29	1	0
31	1	0
35	0	0
45	1	0
46	1	1
47	1	0
48	1	0
50	1	0
51	1	0
52	1	0
53	1	0
55	1	0
58	1	0
59	1	0
61	0	0
63	1	1

Elaborado por: Carrión, Andrea, 2022.

Finalmente, de los 65 datos se clasifican en dos grupos, el primer clúster contiene a 34 siniestros y el segundo clúster contiene 31 casos de siniestros.

Ilustración 19-3: Gráfica para detectar las características de los clústeres

Realizado por: Carrión, Andrea, 2022.

A partir de la información estudiada, se encontró un total de dos clústeres cada uno diferenciado por un color como se indica en la gráfica, el primer clúster está conformado por aquellos eventos

correspondiente a un total de 0 fallecidos y un total de 1 a 3 lesionados, el segundo clúster está conformado de 0 a 1 fallecidos y 0 a 1 lesionados por accidentes de tránsito.

Validación de los grupos

Finalmente, tras el análisis de conglomerados se obtiene dos grupos, por lo tanto, se calcula el coeficiente de correlación cofenética para medir la precisión del agrupamiento de los clústeres encontrados:

```
> cor(cophenetic(hc1), d)
[1] 0.780543
```

Ilustración 20-3: Coeficiente cofenético

zRealizado por: Carrión, Andrea, 2022.

Se obtiene un coeficiente de correlación cofenético de 0.789543, se observa que un valor bastante cercano o próximo a 1, por lo tanto, se considera las características de los individuos que forman parte de los grupos formados tras el análisis de conglomerados es bastante bueno.

DISCUSIÓN

Los siniestros de tránsito son una de las principales amenazas para el bienestar y la seguridad humana, y la identificación de las causas y de zonas críticas de los siniestros de tránsito es esencial para la asignación de recursos y la generación de políticas públicas para mejorar la seguridad del transporte, por parte de las autoridades. Dentro de este contexto, esta investigación genera dos resultados importantes, el primer resultado, evidencia que las clases de siniestros de accidentes de tránsito más predominantes en las zonas del cantón La Joya de los Sachas, fue el conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor) con un 70.27%. Con menor porcentaje se presenta la causa de conducir vehículos superando los límites máximos de velocidad y por malas condiciones de la vía y/o configuración como iluminación y diseño (ambos con el 8.11%). Este importante resultado, muestra que se podría disminuir la cantidad de accidentes en el cantón de estudio, enfocándose en campañas de concientización, para que no se utilice mientras se conduzca el celular, pantallas de video o cualquier otro elemento distractor y que tampoco se realice actividades como comer o maquillarse.

Al respecto, la Agencia Nacional de Tránsito del Ecuador en sus publicaciones periódicas indican que para diciembre de 2020 la causa más probable de siniestro de tránsito a nivel nacional fue conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor) con una frecuencia relativa absoluta de 25.60%, esto corrobora el primer resultado importante de este trabajo. Por lo tanto, una reacción natural del Gobierno Ecuatoriano ha sido reformar la Ley de Tránsito, donde desde agosto 2020, los conductores que hablen por celular al conducir o excedan los límites de velocidad serán sancionados con penas económicas (Puentes, 2021, p. 15).

El segundo resultado más importante, considerando las variables cualitativas en estudio (zona, feriado, causa, clase y día), se encontraron varias agrupaciones de modalidades sin embargo el par de caso que más destacó fueron a los siniestros que ocurrieron como causa de atropellos y clase volcamientos. De igual manera, se evidenció casos muy lejanos con respecto a otras y sin llegar a conformar asociaciones con otras como en los casos: siniestros ocurridos en zonas urbanas, causados por no ceder el derecho de vía o preferencia de paso al peatón bastante alejado del origen y sin otra modalidad a la cual agruparse, y los que se provocaron por no mantener la distancia prudencial con respecto al vehículo que le antecede.

Por último, se ha determinado mediante un análisis de conglomerados que los accidentes de tránsito del cantón La Joya de los Sachas, se ha clasificado en dos grupos, el primer clúster conforma los siniestros que tienen 0 fallecidos y un total de 1 a 3 lesionados, y el segundo clúster agrupa a los siniestros que tienen 0 a 1 fallecidos y 0 a 1 lesionados.

CONCLUSIONES

- Se determinó que conducir desatento es la causa más predominante con un porcentaje aproximadamente de 70,27% con respecto a los siniestros de tránsito que ocurrieron en el cantón.
- El análisis de datos atípicos dio como resultado un total de 8 observaciones atípicas correspondientes a mediciones que contaban con una mayor cantidad de fallecidos y lesionados por siniestros de tránsito. Por otra parte, a partir del análisis de variables redundantes se determinó que no existe relación lineal significativa entre las mudables propuestas.
- El análisis de correspondencias permitió identificar que las modalidades con mayor fuerza de asociación, resultando ser los siniestros que se identificaron como atropellos y volcamientos.
- Existen dos grupos de clasificación, el primer grupo engloba aquellos siniestros con un número de 0 fallecidos y 1-3 lesionados y el segundo grupo incluye a los siniestros que tienen similitud entre 0–1 fallecidos y 0-1 lesionados.

RECOMENDACIONES

- Utilizar nuevas variables como género o edad para realizar una comparación con las que se han estudiado.
- Efectuar un estudio al conjunto de datos resultantes como atípicos e indagar en su origen.
- Asociar las ubicaciones específicas de los siniestros más recurrentes considerando las relaciones más predominantes.
- Comparar entre varios métodos de agrupamiento y analizar los resultados de cada uno de ellos.

BIBLIOGRAFÍA

ABDI, Hervé & VALENTIN, Dominique. "Multiple correspondence analysis". *Encyclopedia of measurement and statistics* [en línea], 2007. 2(4), pp. 651-657. [Consulta: 11 diciembre de 2021]. Disponible en: https://personal.utdallas.edu/~Herve/Abdi-MCA2007-pretty.pdf

ALI, Zulfiqar; et al. "Descriptive statistics: Measures of central tendency, dispersion, correlation and regression". *Airway* [en línea], 2019**.** 2(3), pp. 120-125. [Consulta: 20 noviembre 2021]. Disponible en: https://www.arwy.org/article.asp?issn=26659425;year=2019;volume=2;issue=3;spage=120;epage=125;aulast=Ali

ARTEAGA, Pedro; et al. *Evaluación de conocimientos sobre gráficos estadísticos y* Departamento de Didáctica de la Matemática. España: ed. Granada, pp. 18-93.

BAJPAI, Naval. *Business Statistics.* India: Pearson Education India, 2019. P.119.

BERGSTROM, Carl; & WEST, Jevin. "Why scatter plots suggest causality, and what we can do about it". *arXiv* [en línea], 2018. 1(1), pp. 1-4. [Consulta: 22 diciembre 2010. Disponible en: https://arxiv.org/abs/1809.09328

BENESTY, Jacob; et al. Pearson Correlation Coefficient. *Springer - Berlin: Topics in Signal Processing*, 2(1), 2009. pp. 1-4.

BLACK, William; & THOMAS, Isabelle. "Accidents on Belgium's motorways: a network autocorrelation analysis". *Journal of Transport Geography* [en línea], 1998. 6(1), pp. 23-31. [Consulta: 20 noviembre 2020]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0966692397000379

BLASHFIELD, Roger; & ALDENDERFER, Mark. "*The Methods and Problems of Cluster Analysis.* Springer-Boston: Fayyad, 1988. pp. 47-473.

BLASIUS, Jorg; & GREENACRE, Michael. *Visualization and verbalization of data.* Boca Raton: Taylor & Francis Group, 2014. P.141

CHEN, Simiao; et al. "*The global macroeconomic burden of road injuries: estimates and projections for 166 countries*". The Lancet Planetary Health [en línea], 2019. 3(9), p. 9. [Consulta: 20 noviembre 2021]. Disponible en: https://www.sciencedirect.com/science/article/pii/S2542519619301706

CHRISTENSEN, Alexander; & et. al. *Unique Variable Analysis:A Novel Approach for Detecting Redundant Variables in Multivariate Data*. PsyArXi [en línea], 2020. pp. 1-36. [Consulta: 10 noviembre 2021]. Disponible en: https://files.osf.io/v1/resources/4kra2/providers/osfstorage/5fe3ceb6e3acd100224a9243?format=pdf&action=download&direct&version=3

Center for disease control and prevention. *Division of unintentional injury prevention,* 2020. pp.2-5.

CESTERO, Eloy Vicente; et.al. *Data science y redes complejas.* Madrid-España: Centro de Estudios Ramón Areces SA, 2018. pp. 5-276.

CÓRDOVA GUZMÁN, Luis Antonio; & PAUCAR FLORES, Christian Rómulo. *Análisis de los indicadores de seguridad vial para la disminución de accidentes de tránsito en el Ecuador.* Cuenca-Ecuador: Tesis de Licenciatura. 2014. P.120.

DAGNINO, Jorge. Coeficiente de correlación lineal de Pearson. *Chil Anest*, 2014, vol. 43, no 1, p. 150-153.

DAWSON, Robert. "How Significant is a Boxplot Outlier?". *Journal of Statistics Education* [en línea], 2011. 19(2), pp. 1-13. [Consulta: 17 noviembre 2021]. Disponible en: https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889610

DE LA FUENTE FERNÁNDEZ, Santiago. *Madrid:* Análisis de correspondencias simples y múltiples.2011. pp. 1-9.

EcuRed. ***Joya de los Sachas.*** [En línea], 2018. [Consulta: 16 diciembre 2021]. Disponible en: https://www.ecured.cu/Cant%C3%B3n_La_Joya_de_los_Sachas_(Ecuador)

EcuRed. ***Distancia euclidea.*** [En línea], *2018.* [Consulta: 01 01 2022]. Disponible en: https://www.ecured.cu/Distancia_eucl%C3%ADdea

EMERSON, Robert Wall. "Causation and Pearson's Correlation Coefficient". *Journal of Visual Impairment & Blindness* [en línea], 2015. 109(3), p. 242. [Consulta: 03 enero de 2022]. Disponible en: https://journals.sagepub.com/doi/abs/10.1177/0145482X1510900311

ESCAMILLA, Marisela Dzul. "Aplicación básica de los métodos científicos". *Universidad Estatal del estado de Hidalgo* [en línea], **2020.** pp. 2-3. [Consulta: 15 diciembre de 2021]. Disponible en: http://repository.uaeh.edu.mx/bitstream/handle/123456789/8046/discover?filtertype_0=subject&filter_0=TECHNOLOGY&filter_relational_operator_0=equals&filtertype=author&filter_relational_operator=equals&filter=Dzul+Escamilla%2C+Marisela

FISHER, Murray; & MARSHALL, Andrea. "Understanding descriptive statistics". *ScienceDirect* [en línea],*2009.* 22(3), pp. 93-97. [Consulta: 13 enero de 2022]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1036731408001732

FORINA, M; & et. al. Clustering with dendrograms on interpretation variables. *ScienceDirect* [en línea], 2002. 454(1), pp. 13-19. [Consulta: 14 febrero 2022]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0003267001015173

GARCÍA ALTÉS, A; & PÉREZ, Katherine. "The economic cost of road traffic crashes in an urban setting". *Injury prevention* [en línea], 200**7.** 13, pp. 65-68. [Consulta: 10 diciembre de 2021]. Disponible en: https://injuryprevention.bmj.com/content/13/1/65.short

GLYNN, Dylan; ROBINSON, Justyna A. *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy.* Amsterdam: John Benjamins, 2014. P.450.

GOLMAN, Adam; et. al. *"*Injury prediction in a side impact crash using human body model simulation". *Accid Anal Prev* [en línea], *2014.* 64(1), pp. 1-8. [Consulta: 7 noviembre de 2021]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0001457513004363

GREENACRE, Michael. *La práctica del análisis de correspondencia* .Plaza de San Nicolás: Fundación BBVA. 2008. P.447.

GREENACRE, Michael. *La práctica del análisis de correspondencia.* Plaza de San Nicolás: Fundación BBVA. 2008. P.144.

GUNST, Richard; & MASON, Robert. *Regression Analysis and Its Application: a Data-Oriented Approach.* Boca Raton: CRC Press, 2018. P.170-500.

KALIYADAN, Feroze; & KULKARNI, Vinay. "Types of Variables, Descriptive Statistics, and Sample Size". *Indian dermatology online journal* [en línea], 2019. 10(1), p. 82–86. [Consulta: 20 febrero de 2022]. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc6362742/

KANNAN, Senthamarai; & MANOJ, K. "*Outlier detection in multivariate da*ta". *Applied Mathematical Sciences* [en línea], 2015. 47(9), pp. 2317-2324. [Consulta: 04 marzo de 2022] Disponible en: http://m-hikari.com/ams/ams-2015/ams-45-48-2015/13manojAMS45-48-2015-96.pdf

KNORR, Edwin M. *Outliers and data mining: finding exceptions in data*. Tesis Doctoral. University of British Columbia. 2002. Pp. 407-411

KRISP, Jukka Matthias; & DUROT, Sara. "Segmentation of lines based on point densities an optimisation of wildlife warning sign placement in southern Finland". *Accident Analysis & Prevention* [en línea], 2007. 39(1), pp. 38-46. [Consulta: 01 diciembre de 2021]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0001457506001011

KUBUS, Mariusz. "The problem of redundant variables in random forests". *Repozytorium Uniwersytetu Łódzkiego*, vol. 15, nº 6, 339(2018), (Polonia) pp.12.

KWAK, Sang Kyu; & KIM, Jong Hae. "Statistical data preparation: management of missing values and outliers". *Korean J Anesthesiol* [en línea], 2017. 70(4), pp. 407-411. [Consulta: 20 noviembre 2021]. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/

LAHUATHE ALARCÓN, Gabriela. Análisis exploratorio espacial de los accidentes de tránsito en las provincias y cantones del Ecuador (Trabajo de titulación). (Maestría) Universidad Internacional SEK, Quito, Ecuador. 2018. pp.1.

LALKHEN, Abdul Ghaaliq; & MCCLUSKEY, Anthony. "Statistics II: Central tendency and. Continuing Education in Anaesthesia", *Critical Care and Pain* [en línea], 2007. 7(4), pp. 127-130. [Consulta: 20 febrero de 2022]. Disponible en: https://www.bjaed.org/article/S1743-1816(17)30353-0/fulltext

LEONE, Nicola, et al. *Occupant protection performance in side impact collisions preceded by pre-crash deployment of on-board safety systems*. Germany: *Proceedings of the 24th International Technical Conference on the Enhanced Safety of Vehicles (ESV), 2015.* pp. 15-31.

LEWANDOWSKY, Stephan; & SPENCE, Ian. "The perception of statistical graphs". *Sociological Methods & Research* [en línea], 1989. 188(2-3), pp. 200-242. [Consulta: 02 marzo de 2022. Disponible en: https://journals.sagepub.com/doi/abs/10.1177/0049124189018002002

LIBERTI, Leo, et al. "Euclidean Distance Geometry and Applications". *Society for Industrial and Applied Mathematics* [en línea], 2014. 56(1), pp. 1-5. [Consulta: 20 febrero 2022]. Disponible en: https://epubs.siam.org/doi/abs/10.1137/120875909

LI, Hongfei; & et. al. "Beyond Moran's : Testing for Spatial Dependence Based on the Spatial Autoregressive Model". *Geographical analysis* [en línea], 2007. 39(4), pp. 357-375. [Consulta: 06 enero de 2022]. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.2007.00708.x

LI, Linhua; ZHU, Li; SUI, Daniel Z. "GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes". *J. Transp. Geog* [en línea], 2007.5(4), pp. 274-285. [Consulta: 20 diciembre 2021]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0966692306000901

LÓPEZ, Ana María. Análisis de conglomerados (cluster analysis). *Área de Metodología de las Ciencias del Comportamiento Departamento de Psicología Experimental*. 2018. pp. 3.

LÓPEZ MONTERO, José María, *Estadística Descriptiva.* Madrid: Paraninfo. 2007. P. 1-379

LY, Alexander; & MARSMAN, Maarten. *Analytic posteriors for Pearson's correlation coefficient.* Amsterdam: arXis. 2018. P.140

MA, Rubao, et al. "Asymptotic Properties of Pearson's Rank-Variate Correlation Coefficient under Contaminated Gaussian Model". *PLOS ONE* [en línea], 2012. 9(11). [Consulta: 06 enero de 2022]. Disponible en: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112215

MANIKANDAN, S. "Measures of central tendency: The mean". *Journal of Pharmacology and Pharmacotherapeutics* [en línea], *2011.* 2(2), pp. 140-142. [Consulta: 06 enero de 2022]. Disponible en: https://www.proquest.com/openview/b6d16b1943d675c7148493f4c1ae9b14/1?pqorigsite=gscholar&cbl=226473

MARTINEZ, Wendy; & MARTINEZ, Angel. *Computational statistics handbook with MATLAB.* New York: Chapman and Hall/CRC, 2001. P.10.

MENON, Shashi. *Transmission Pipeline Calculations and Simulations Manual. Waltham-USA:*Gulf Proffesional Publishing, 2014. P.15.

MOONS, Elke; & et. al. *Wets Improving Moran's index to identify hot spots in traffic safety.* Springer-Verlag Berlin Heidelberg: Geocomputation Urban Plan, 2009. pp. 117-132.

NAVARRO, Horra, et al. *Estadística aplicada*. 3º ed. Madrid: Ediciones Díaz de Santos, 2003. pp.11-122.

OKSANEN, Jari. *Cluster Analysis: Tutorial with R.* Oulu: University of Oulu, 2012. P.4

ORGANIZACIÓN MUNDIAL DE LA SALUD, 2013. *Violence and Injury Prevention, and World Health Organization.* [Arte] (Organización Mundial de la Salud).

ORGANIZACIÓN MUNDIAL DE LA SALUD, 2015. *Global status report on road safety 2015.* [En línea] [Consulta: 16 de diciembre 2021]. Disponible en: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

PATTEN, Mildred; & NEWHART, Michelle. *Understanding Research Methods.* 10 ed. New York: Routledge, 2017. pp. 18-352.

PUENTES, D., 2021. Así quedaron los cambios en la Ley de Tránsito aplicadas desde el 10 de agosto. *El Comercio*, agosto, p. 15.

RENDÓN, MACÍAS, Mario Enrique; & et. al. "Estadística descriptiva". *Revista Alergia México* [en línea], *2016.* 63(4), pp. 397-407. [Consulta: 06 enero de 2022]. Disponible en: http://revistaalergia.mx/ojs/index.php/ram/article/view/230

RICHARD, Taylor. "Interpretation of the Correlation Coefficient: A Basic Review". *Journal of Diagnostic Medical Sonography* [en línea], 1990**.** 6(1), pp. 35-39. [Consulta: 06 enero de 2022]. Disponible en: https://journals.sagepub.com/doi/abs/10.1177/875647939000600106

ROCKINSON-SZAPKIW, Amanda. *Statistics Guide.* s.l.: Retrieved from., 2013. P.119

ROMESBURG, Charles. *Cluster analysis for researchers.* North Calorina: LULU PRESS, 2004. pp. 2-27.

SAMPIERI, HERANDEZ Roberto; & et. al. *Metodología de la Investigación.* Sexta ed. México: McGRAW-HILL, 2014**.** P.127.

SANTISTEBAN, Julio; & TEJADA-CÁRCAMO, Javier. *Unilateral Jaccard Similarity Coefficient.* Santiago - Chile: GSB@ SIGIR, 2015. pp. 23-27.

SARKER, Bhaba R; & ISLAM, Khan M. Saiful. "Relative performances of similarity and dissimilarity measures". *SciencDirect*, [en línea], 1999**.** 37(4), pp. 769-807. [Consulta: 20 enero de 2022]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0360835200000115

SCHAEFFER, L. Estimation of Variances and Covariances Within the Allowable Parameter Space. *ScienceDirect* [en línea],*1986.* 69(1), pp. 87-194. [Consulta: 04 marzo de 2022]. Disponible en: https://www.sciencedirect.com/science/article/pii/S002203028680385X

SCHOBER, Patrick; et. al. "Correlation Coefficients: Appropriate Use and Interpretation". *Wolters Kluwer* [en línea], 2018. 126(5), pp. 1763-1768. [Consulta: 04 marzo de 2022]. Disponible en: https://www.ingentaconnect.com/content/wk/ane/2018/00000126/00000005/art00051

SHIRKHORSHIDI, Ali Seyed; AGHABOZORGI, Saeed; WAH, Teh Ying. A comparison study on similarity and dissimilarity measures in clustering continuous data". *PloS one* [en línea], 2015. 10(12), pp. 1-20. [Consulta: 04 marzo de 2022]. Disponible en: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144059

TONG, Tiejun; & et. al. "Estimation of variances and covariances for high-dimensional data: a selective review". *Wiley Interdisciplinary Reviews* [en línea], 2014**.** 6(4), pp. 255-264.

UNWIN, Antony. *Graphical Data Analysis with R.* Boca Raton: Chapman and Hall/CRC, 2015. P. 310.

URDAN, Timothy. *Statistics in Plain English.* 3rd Edition ed. New York: Behavioral Sciences, Education, Social Sciences, 2010.P. 13.

VAN SICKL, John. "Using Mean Similarity Dendrograms to Evaluate Classifications". *JOURNAL ARTICLE* [en línea], 1997. 2(4), pp. 370-388. [Consulta: 04 marzo de 2022]. Disponible en: https://www.jstor.org/stable/1400509

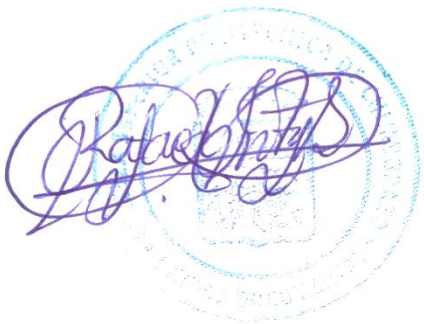
WEISBERG, Herbert; & WEISBERG, Herbert. *Central tendency and variability*. Londo: Sage Publications, 1992. P.5.

WORLD HEALTH ORGANIZATION, et al. *Global status report on road safety 2013: supporting a decade of action: summary.* World Health Organization, 2013.

WILLIAMS, R. B. G. *Introduction to Statistics for Geographers and Earth Scientists.* Palgrave, London: Macmillan Publishers Limited, 1984. P.51.

XIE, Zhixiao; & YAN, Jun. "Kernel density estimation of traffic accidents in a network space Comput". *Computers, environment and urban systems [en línea],2008.* 35(5), pp. 396-406. [Consulta: 04 marzo de 2022]. Disponible en: https://link.springer.com/chapter/10.1007/978-1-349-06815-9_6

ZANI, Sergio; & at. al. "Robust bivariate boxplots and multiple outlier detection". *ScienceDirect [en línea],* 1998. 28(3), pp. 257-270. [Consulta:10 enero de 2022]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0167947398000401



ANEXOS

ANEXO A: MATRIZ DE DATOS DE LA AGENCIA NACIONAL DE TRÁNSITO

DATOS	DÍA	ZONA	…	NÚMERO DE FALLECIDO	NÚMERO DE LESIONADO
1	JUEVES	URBANA	…	0	2
2	MARTES	RURAL	…	0	1
3	MARTES	RURAL	…	0	1
4	DOMINGO	RURAL	…	1	0
5	SÁBADO	RURAL	…	0	4
6	DOMINGO	URBANA	…	0	1
7	DOMINGO	RURAL	…	1	0
8	DOMINGO	RURAL	…	1	0
9	MIÉRCOLES	RURAL	…	2	0
10	DOMINGO	RURAL	…	1	0
11	SÁBADO	RURAL	…	1	2
12	MIÉRCOLES	RURAL	…	1	0
13	LUNES	RURAL	…	0	2
14	MARTES	RURAL	…	1	0
15	MIÉRCOLES	RURAL	…	0	2
16	SÁBADO	RURAL	…	1	1
17	JUEVES	RURAL	…	0	1
18	DOMINGO	RURAL	…	1	0
19	DOMINGO	RURAL	…	1	1
20	MARTES	RURAL	…	1	0
21	MIÉRCOLES	RURAL	…	0	1
22	JUEVES	RURAL	…	2	2
23	MIÉRCOLES	URBANO	…	0	0
24	SÁBADO	RURAL	…	0	3
25	LUNES	RURAL	…	0	10
26	LUNES	URBANO	…	0	1
27	JUEVES	URBANO	…	0	0
28	SÁBADO	RURAL	…	0	0
29	DOMINGO	URBANO	…	1	1
30	LUNES	RURAL	…	0	1
31	SÁBADO	RURAL	…	0	0
32	LUNES	RURAL	…	0	1
33	MIÉRCOLES	URBANO	…	0	2
34	JUEVES	URBANO	…	1	0
35	VIERNES	RURAL	…	0	1
36	JUEVES	RURAL	…	1	0
37	SÁBADO	RURAL	…	0	2
38	LUNES	RURAL	…	1	2
39	VIERNES	RURAL	…	1	2
40	SÁBADO	RURAL	…	0	2

41	MIÉRCOLES	RURAL	**…**	0	1
42	MARTES	RURAL	**…**	0	0
43	MIÉRCOLES	RURAL	**…**	0	3
44	JUEVES	URBANA	**…**	0	1
45	JUEVES	URBANA	**…**	0	2
46	JUEVES	URBANA	**…**	0	1
47	VIERNES	URBANA	**…**	0	1
48	MIÉRCOLES	URBANA	**…**	0	1
49	SÁBADO	URBANA	**…**	0	3
50	DOMINGO	URBANA	**…**	0	3
51	DOMINGO	RURAL	**…**	0	1
52	LUNES	RURAL	**…**	1	0
53	LUNES	URBANA	**…**	1	1
54	MARTES	URBANA	**…**	1	0
55	SÁBADO	URBANA	**…**	1	0
56	DOMINGO	RURAL	**…**	0	1
57	DOMINGO	RURAL	**…**	1	0
58	VIERNES	RURAL	**…**	1	0
59	SÁBADO	URBANA	**…**	1	5
60	VIERNES	RURAL	**…**	1	0
61	DOMINGO	URBANA	**…**	1	0
62	LUNES	URBANA	**…**	0	1
63	MIÉRCOLES	RURAL	**…**	1	0
64	SÁBADO	URBANA	**…**	0	1
65	SÁBADO	RURAL	**…**	0	1
66	SÁBADO	RURAL	**…**	1	0
67	MIÉRCOLES	RURAL	**…**	1	0
68	LUNES	RURAL	**…**	0	1
69	MIÉRCOLES	RURAL	**…**	0	0
70	JUEVES	RURAL	**…**	0	1
71	VIERNES	RURAL	**…**	1	1
72	JUEVES	RURAL	**…**	0	2
73	VIERNES	RURAL	**…**	0	1

ANEXO B: CÓDIGO EN R, ANÁLISIS DESCRIPTIVO DE VARIABLES CUANTITATIVAS

```
#Codigo para generar boxplot de las variables en estudio

boxplot(datos$NUM_LESIONADO, horizontal = FALSE,

col = "light blue", main = "Boxplot de número de lesionados")

#Fin

#Histograma

sachas.1 <- sachas
sachas.1$VEHICULOS <- as.character(sachas.1$VEHICULOS)
a <- sachas.1 %>% count(VEHICULOS)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%
  ggplot(aes(x=VEHICULOS,
 y=f,
fill=VEHICULOS)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn VehÃ-culo",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "VehÃ-culo",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

sachas.1 <- sachas
sachas.1$FALLECIDOS <- as.character(sachas.1$FALLECIDOS)

a <- sachas.1 %>% count(FALLECIDOS)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%

ggplot(aes(x=FALLECIDOS,
y=f, fill=FALLECIDOS)) +
  geom_bar(stat="identity") +
```

```
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Fallecidos",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Fallecidos",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

sachas.1 <- sachas
sachas.1$LESIONADOS <- as.character(sachas.1$LESIONADOS)
a <- sachas.1 %>% count(LESIONADOS)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%
  ggplot(aes(x=LESIONADOS,
y=f,
fill=LESIONADOS)) +
geom_bar(stat="identity") +
geom_text(aes(label=f),
vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Lesionados",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Lesionados",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

#Fin
```

ANEXO C: CÓDIGO EN R, ANÁLISIS DESCRIPTIVO DE VARIABLES CUANLITATIVAS

```
#### Cargar de datos ----

#boxplot

#librerias para el analisis

library(gstat)
library(tidyverse)
library(readxl)

sachas <- read_excel("Data/Andrea C/sachas.xlsx",
col_types = c("text", "date", "date",
"text", "numeric", "text",
"text", "text", "text",
"text", "text", "text",
"text", "numeric",
"numeric", "numeric",
 "numeric", "numeric"))

sachas <- sachas %>% filter(!is.na(x))
sachas <- sachas %>% filter(x < -0.2)
sachas <- sachas %>% arrange(y)

sachas[c(6),"PARROQUIA"] <- "SAN SEBASTIAN DEL COCA"
sachas[sachas$PARROQUIA == "CARRETERA",
"PARROQUIA"] <- "SACHAS"

#### FIN Cargar de datos ----

#### Paso 1 AED ----

a <- sachas %>% count(PARROQUIA)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%
  ggplot(aes(x=PARROQUIA,
 y=f,
```

```
fill=PARROQUIA)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Parroquia",
   caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Parroquia",
  y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(DIA)
a$DIA <- factor(a$DIA,
levels=c("LUNES","MARTES",
"MIERCOLES", "JUEVES",
"VIERNES","SABADO",
"DOMINGO"),
ordered = TRUE)
a <- a[order(a$DIA),]

a %>% mutate(f=round(prop.table(n)*100,2)) %>%
  ggplot(aes(x=DIA, y=f, fill=DIA)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),
vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn dÃ-a",
   caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
   x = "DÃ-a",
   y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

ano <- as.numeric(substr(sachas$FECHA, 1,4))
ano <- as.tibble(x = ano)
ano$value <- as.character(ano$value)
```

```
a <- ano %>% count(value)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%
ggplot(aes(x=value,
y=f, fill=value)) +
 geom_bar(stat="identity") +
 geom_text(aes(label=f),
vjust=-0.5) +
labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn AÃ±o",
   caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
   x = "AÃ±o",
   y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

ano <- as.numeric(substr(sachas$FECHA, 6,7))
ano <- as.tibble(x = ano)

a <- ano %>% count(value)
names(a) <- c("MES","n")

a$MES <- c("ENERO","FEBRERO",
"MARZO","ABRIL","MAYO",
"JUNIO","JULIO","AGOSTO",
"SEPTIEMBRE","OCTUBRE",
"NOVIEMBRE","DICIEMBRE")
a$MES <- factor(a$MES, levels=c("ENERO",
"FEBRERO","MARZO",
"ABRIL","MAYO","JUNIO",
"JULIO","AGOSTO","SEPTIEMBRE",
"OCTUBRE","NOVIEMBRE","DICIEMBRE"),
ordered = TRUE)

a <- a[order(a$MES),]
a %>% mutate(f=round(prop.table(n)*100,2)) %>%

  ggplot(aes(x=MES, y=f, fill=MES)) +
```

```
  geom_bar(stat="identity") +
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Mes",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Mes",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(HORA_D)
a$HORA_D<- factor(a$HORA_D,
levels=c("MADRUGADA",
"MAÃ‘ANA",
"TARDE",
"NOCHE"),
ordered = TRUE)
a <- a[order(a$HORA_D),]
a %>% mutate(f=round(prop.table(n)*100,2)) %>%

ggplot(aes(x=HORA_D,
y=f, fill=HORA_D)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn hora del dÃ-a",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Hora del dÃ-a",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(FERIADO)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%

ggplot(aes(x=FERIADO,
y=f, fill=FERIADO)) +
```

```
  geom_bar(stat="identity") +
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn feriado",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Feriado",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(CAUSA)

a$CAUSA <- c("CAUSA 1","CAUSA 2",
"CAUSA 3","CAUSA 4","CAUSA 5",
"CAUSA 6","CAUSA 7","CAUSA 8")

a %>% mutate(f=round(prop.table(n)*100,2)) %>%

  ggplot(aes(x=CAUSA, y=f, fill=CAUSA)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),
vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Causa",
    caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
    x = "Causa",
    y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(CLASE)

a %>% mutate(f=round(prop.table(n)*100,2)) %>%

ggplot(aes(x=CLASE, y=f, fill=CLASE)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),
```

```
vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Clase",
       caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
       x = "Clase",
       y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(CONSECUENCIA)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%

ggplot(aes(x=CONSECUENCIA,
y=f,
fill=CONSECUENCIA)) +
geom_bar(stat="identity") +
geom_text(aes(label=f),
vjust=-0.5) +
labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Consecuencia",
caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
x = "Consecuencia",
y = "Porcentaje") +
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()

a <- sachas %>% count(ZONA)
a %>% mutate(f=round(prop.table(n)*100,2)) %>%

ggplot(aes(x=ZONA, y=f,
 fill=ZONA)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=f),vjust=-0.5) +
  labs(title = "Frecuencias de siniestros de trÃ¡nsito segÃºn Zona",
  caption = "CantÃ³n La Joya de los Sachas, 2015- 2020",
  x = "Zona",
  y = "Porcentaje") +
```

```
  #  theme_minimal() +
  #  theme_dark() +
  theme_light()


#### FIN
```

ANEXO D: CÓDIGO EN R, ANÁLISIS DE DATOS ATÍPICOS

```
library(readxl)

data <- read_xlsx("BASE LIMPIA (1).xlsx",
sheet = "A USAR")
data_conglomerados <- data[,8:9]

##Datos atipicos

library(readxl)
library(mvoutlier)

data <- read_xlsx("BASE LIMPIA (1).xlsx",
sheet = "A USAR")

data_conglomerados <- data[,8:9]
modelo.regresion <- lm(NUM_FALLECIDO ~ .,
data=data_conglomerados)

cooksd <- cooks.distance(modelo.regresion)



# Gráfica de la distancia de Cook

plot(cooksd, pch="*",
cex=2,
main="Observaciones Atípicas por distancia de Cook")

# Superponemos el límite definido
abline(h = 4*mean(cooksd, na.rm=T), col="red")

# Agregamos etiquetas de identificación para observaciones atípicas

text(x=1:length(cooksd)+1,
y=cooksd,
labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),
```

```
names(cooksd),""),
col="red")
data_2 <- data_conglomerados

for(i in 1:3){
modelo.regresion <- lm(NUM_FALLECIDO ~ .,
data=data_2,na.action=na.omit)
cooksd <- cooks.distance(modelo.regresion)
atipicas <- which((cooksd>4*mean(cooksd, na.rm=T)) == "TRUE")
data_2 <- data_2[-atipicas,]
i <- i+1
}

modelo.regresion <- lm(NUM_FALLECIDO ~ ., data=data_2)
cooksd <- cooks.distance(modelo.regresion)

# Gráfica de la distancia de Cook

plot(cooksd,
pch="*",
cex=2,
main="Observaciones Atípicas por distancia de Cook"
,ylim = c(0,0.06))

# Limite definido

abline(h = 4*mean(cooksd,
na.rm=T),
col="red")

# Agregamos etiquetas de identificación para observaciones atípicas

text(x=1:length(cooksd)+1,
 y=cooksd,
labels=ifelse(cooksd>4*mean(cooksd,
na.rm=T),
names(cooksd),""),
```

```
col="red")

#### Variables redundantes

datos <- data_conglomerados
names(datos)

# Grafico entre las variables explicativas

pairs(NUM_FALLECIDO ~ NUM_LESIONADO,
data =datos ,
main="Grafico: Matriz de Dispersion del datos")

#Matriz de correlaciones

cor(datos,method = "spearman")

cor.test(x=datos$NUM_FALLECIDO,
y=datos$NUM_LESIONADO,

#Coeficiente de Spearman
method = "spearman")
var(datos)
```

ANEXO E: CÓDIGO EN R, ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES

```
##ANALISIS DE CORRESPONDENCIAS MULTIPLES

#librerias para el análisis

library(readxl)
library(fastDummies)

data <- read_xlsx("BASE LIMPIA.xlsx",
sheet = "A USAR")

data_correspondencia <- data[,2:7]
data_conglomerados <- data[,8:9]

dummy <- dummy_cols(data_correspondencia,
colnames(data_correspondencia))
dummy <- dummy[,-c(1:6)]

p <- dummy/sum(dummy)
p <- as.matrix(p)

rr <- margin.table(p,1)
cc <- margin.table(p,2)
S <- diag(rr^(0.5))%%(p-rr %% t(cc)) %*% diag(cc^(-0.5))

u <- svd(S)$u
u
v <- svd(S)$v
v
Da <- diag(svd(S)$d)

# Coordenadas principales de filas

FF <- diag(rr^(-0.5))%%u%%Da
```

```
# Coordenadas principales de columnas

GG <- diag(cc^(-0.5))%% v %% Da
# Gráfico 1

plot(GG[,1], GG[,2],
type = "n")
text(GG[,1], GG[,2],
labels = names(data_correspondencia),
cex=0.7)
```

ANEXO F: CÓDIGO EN R, ANÁLISIS DE CONGLOMERADOS

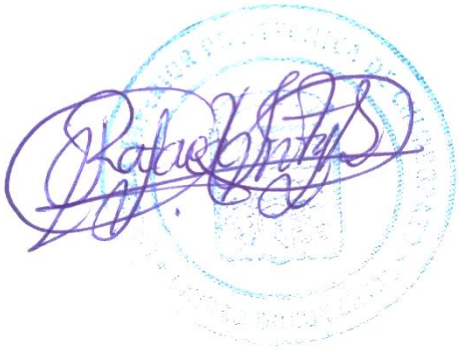
```
#Inicio del Analisis de conglomerados

library(readxl)
library(rgl)

data <- read.csv("Sin Atipicos.csv")
data <- data[,2:3]
plot(data)

d   <- dist(data,
method = "euclidean")

hc1 <- hclust(d,
method = "complete" )

plot(hc1,
cex=.5,
hang=-1)
c1 <-cutree(hc1, 4)

plot(data,
 col = c1)

legend("topright",
legend=paste("Cluster",
unique(c1)),
col=unique(c1),
pch=rep(c(16,18),
each=4),

bty="n",
ncol=2,
cex=0.7,
pt.cex=0.7)
```

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 19 / 01 / 2023

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: Andrea Esthefania Carrión Alvarado
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Estadística
Título a optar: Ingeniera en Estadística Informática
f. responsable: Ing. Rafael Inty Salto Hidalgo



0093-DBRA-UPT-2023