

## **Accidentabilidad Barranquilla 2021- 2025**

**Candy Manrique**

**Gina Ortega**

**Karen Sanjuan**

**Lina Piocuda**

**Martha Coronel**

**Marcela Luna**

**Sharon Varelo**

**Stefany Lozano**

**Tina Varela**

**Vanessa Gonzalez**

## **Análisis de Datos – Conexión Mujeres TIC**

**Tutor: Santiago Giraldo**

**Fecha de entrega: Junio 12 del 2025**

**Barranquilla, Atlántico**

## Índice

### Introducción

#### 1. Objetivos del análisis

#### 2. Hipótesis de trabajo

#### 3. Metodología

- 3.1. Fuentes de datos
- 3.2. Herramientas utilizadas
- 3.3. Técnicas de análisis aplicadas

#### 4. Pasos preliminares realizados

- 4.1. Obtención de datos
- 4.2. Carga de datos en MySQL
- 4.3. Conexión con Visual Studio Code (Python)
- 4.4. Exploración y limpieza de datos

#### 5. Análisis de datos, resultados y hallazgos

#### 6. Interpretaciones de la evidencia y conclusiones

- 6.1. Hipótesis 1 (H1)
- 6.2. Hipótesis 2 (H2)

#### 7. Modelos

##### 7.1 Random forest

- 7.2 Conclusiones y recomendaciones Random forest
- 7.3 Modelo K-means
- 7.4 Conclusión del modelo K-means

#### 8. Referencias

## Introducción

Este proyecto realizado por el equipo de la compañía **DATA WIZARDS**, busca analizar y visualizar los patrones de accidentalidad vial en la ciudad de Barranquilla durante el periodo comprendido entre 2021 y 2025.

Barranquilla ha experimentado fluctuaciones significativas en las tasas de accidentabilidad, motivadas por múltiples factores socio-demográficos y urbanos. En este análisis, se examinan variables clave como el sexo y la edad de las personas involucradas, el tipo de accidente como colisiones vehiculares, atropellos, caídas, accidentes de motos o bicicletas, el día de la semana en que ocurren y las zonas geográficas de la ciudad con mayor incidencia. Esta combinación de variables permite identificar patrones y correlaciones, esenciales para comprender quiénes están más expuestos y en qué sectores se concentran los siniestros.

La integración de estas variables brinda una visión integral de la siniestralidad vial en Barranquilla. El sexo y la edad permiten desagregar la población en grupos vulnerables, mientras que el tipo de accidente ayuda a diseñar intervenciones específicas.

A lo largo del estudio se presentan los principales hallazgos derivados del análisis de datos, con el fin de ofrecer una visión más clara del contexto vial de la ciudad y evaluar la efectividad de las medidas de seguridad implementadas en los últimos años.

## 1. Objetivo del análisis

Analizar y visualizar los patrones de accidentalidad vial en Barranquilla entre 2021 y 2025. El objetivo es identificar factores que influyen en la gravedad de los accidentes (muertes y heridos), generar conocimiento con enfoque estadístico y predictivo, y proponer alertas estratégicas que ayuden a reducir su impacto.

## **2. Hipótesis principal**

La severidad de los accidentes viales en Barranquilla se ve influenciada por variables sociodemográficas y contextuales, tales como el sexo, la edad, el tipo de siniestro, el día de la semana y la ubicación geográfica del incidente.

“La gravedad de los accidentes viales en Barranquilla está influenciada por variables como el sexo, la edad, el tipo de accidente, el día de la semana y las zonas”.

### **2.1. Sub hipótesis 1:**

Los accidentes viales con mayor número de víctimas mortales se concentran en zonas específicas de Barranquilla.

### **2.2. Sub hipótesis 2:**

La gravedad de los accidentes de tránsito está fuertemente relacionada con la edad y el rol de la víctima, donde los motociclistas jóvenes (18-30 años) y peatones mayores (más de 60 años) presentan un riesgo significativamente mayor de resultar con lesiones graves o fatales.

### 3. Metodología

#### 3.1 Fuente de datos

La base de datos fue obtenida de los datos abiertos del gobierno de Colombia (<https://www.datos.gov.co/>) donde encontramos información pública dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 de la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional.

#### 3.2 Herramientas utilizadas

Para llevar a cabo el análisis de los datos de accidentalidad vial en Barranquilla (2021–2025), se emplearon herramientas tecnológicas enfocadas en el procesamiento, análisis y visualización de datos. A continuación, se detallan las principales:

- **MySQL:** Utilizado para la carga estructurada de los datos en una base relacional. Permitió la importación masiva de archivos .csv, así como la organización, depuración inicial y consultas específicas mediante sentencias SQL.
- **Python (en Visual Studio Code):** Lenguaje central para el procesamiento y análisis de los datos. Se emplearon múltiples librerías especializadas:
  - **Pandas:** para limpieza, transformación y manipulación de estructuras tabulares.
  - **Matplotlib y seaborn:** para la generación de gráficos estadísticos y visualizaciones exploratorias.
  - **Scikit-learn:** para el desarrollo de modelos de machine learning aplicados a la predicción y segmentación de patrones de siniestralidad.
- **Power BI:** Herramienta de inteligencia de negocios usada para construir dashboards interactivos. Permitió integrar visualmente los hallazgos del análisis, facilitando su

interpretación y comunicación a distintos públicos.

- **Técnicas de machine learning:** Se implementaron modelos de aprendizaje supervisado y no supervisado con el fin de:
  - Predecir la gravedad de los accidentes (modelo de clasificación con Random Forest).
  - Agrupar zonas críticas según patrones espacio-temporales (modelo de clustering K-Means).

### 3.3 Técnicas de análisis aplicadas

- a. **Análisis de datos descriptivo:** Se calcularon estadísticas básicas (media, mediana) para entender la distribución de los datos.
- b. **Análisis temporal:** Se analizó la evolución de las multas a lo largo del tiempo, realizando agrupaciones por año y mes.
- c. **Segmentación y agrupación:** Se agruparon datos por variables para identificar patrones y diferencias significativas entre años, además del tipo de vehículo y evaluar diferencias entre los distintos grupos.

#### 4. Pasos preliminares realizados

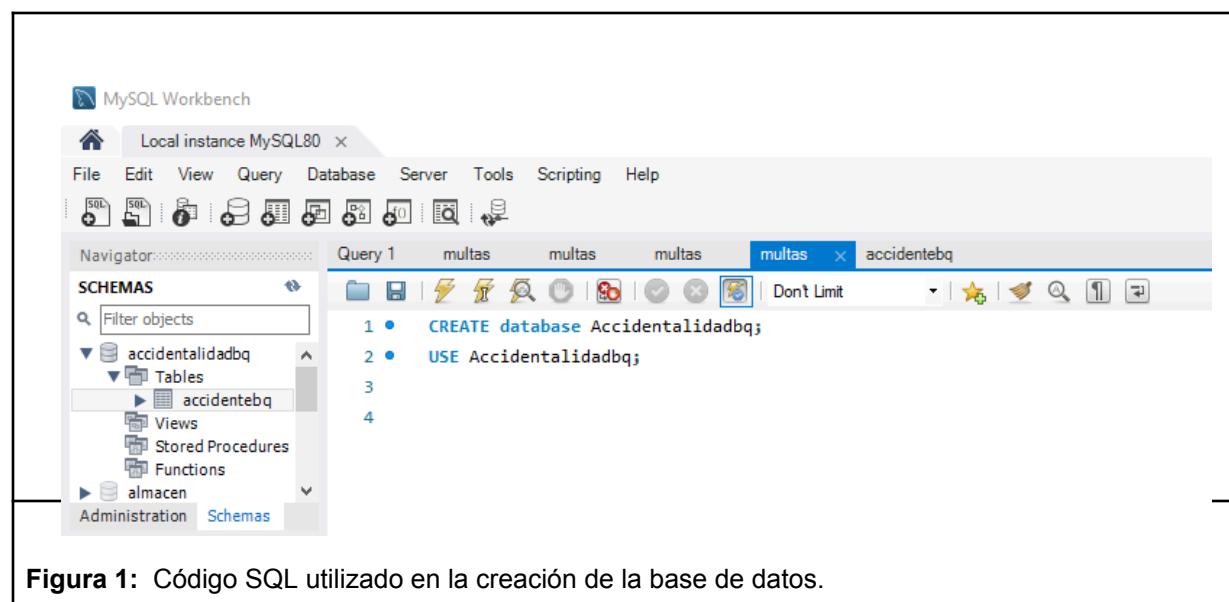
##### 4.1 Obtención de datos:

Se descargaron los datos en formato .csv desde el portal oficial datos.gov.co, específicamente el conjunto de datos titulado “Accidentalidad en Barranquilla - Víctimas”.

La información contenida incluía variables como: Fecha\_Accidente, Direccion\_Accidente, Condicion\_Victima, Gravedad\_Accidente, Clase\_Accidente, Sexo\_Victima, Edad\_victima, Cantidad\_Victimas.

##### 4.2 Carga de datos en MySQL:

Para estructurar los datos, se creó una base de datos en MySQL llamada Accidentalidadadbq. A continuación, se utilizó la herramienta de importación de archivos CSV de MySQL Workbench para cargar los datos directamente en una tabla denominada accidentebq.



The screenshot shows the MySQL Workbench interface. The title bar says "MySQL Workbench" and "Local instance MySQL80". The menu bar includes File, Edit, View, Query, Database, Server, Tools, Scripting, and Help. The toolbar has various icons for database management. The Navigator pane on the left shows "SCHEMAS" with "accidentalidadadbq" expanded, revealing "Tables", "accidentebq", "Views", "Stored Procedures", and "Functions". Below it is "almacen". The "Administration" tab is selected. The main area has three tabs: "Query 1", "multas", and "multas". The "Query 1" tab contains the following SQL code:

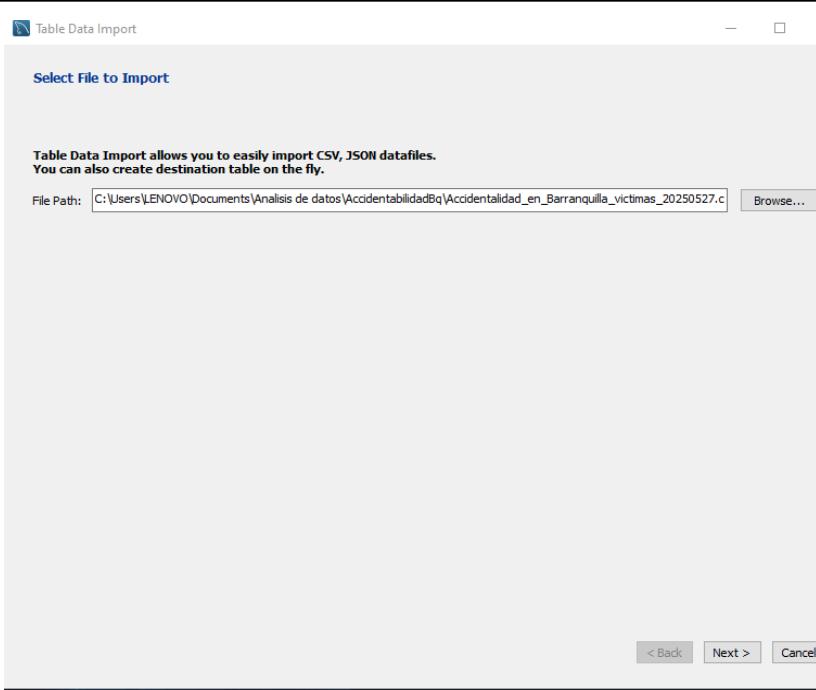
```
1 • CREATE database Accidentalidadadbq;
2 • USE Accidentalidadadbq;
3
4
```

A search bar at the top right contains the text "multas" and "accidentebq".

**Figura 1:** Código SQL utilizado en la creación de la base de datos.



**DATA  
WIZARDS**



**Figura 2:** Proceso de importación del archivo .csv.

A screenshot of MySQL Workbench. The title bar says 'MySQL Workbench - Local instance MySQL80'. The main interface shows the 'accidentebq' table from the 'accidentabilidadBq' schema. The table has columns: Fecha\_Accidente, DIRECCION\_ACCIDENTE, CONDICION\_VICTIMA, GRAVEDAD\_ACCIDENTE, CLASE\_ACCIDENTE, SEXO\_VICTIMA, EDAD\_VICTIMA, and CANTIDAD\_VICTIMAS. The 'Result Grid' shows several rows of data. The first few rows are: 01/01/2018 12:00:00 AM CL 87 9H 24 Peaton herido Atropello M 57 1; 01/01/2018 12:00:00 AM CLLE 119B CRA 11B Peaton herido Atropello F 13 1; 01/01/2018 12:00:00 AM CR 8 CL 41 Acompañante herido Choque F 32 1; etc. The bottom status bar says 'accidentebq 1' and 'Read Only'.

**Figura 3:** Visualización de los datos cargados en la tabla accidentebq.

#### 4.3 Conexión de base de datos desde MySQL con Visual Studio Code

Se realizó la conexión a la base de datos MySQL desde Python utilizando la librería pymysql. Esto permitió extraer los datos necesarios desde la tabla accidentebq para su posterior análisis. La conexión y consulta se llevaron a cabo en el entorno de desarrollo Visual Studio Code, como se muestra en la **Figura 5**.

```
# Conexion a una base de datos MySQL
import pymysql

conn = pymysql.connect(
    host="localhost",
    user="root",
    password="Mysql123#",
    database="Accidentalidadbq"
)

cursor = conn.cursor()

cursor.execute("SELECT * FROM accidentebq")

resultados = cursor.fetchall()
```

**Figura 5.** Conexión de la base de datos MySQL con Visual Studio Code mediante Python e importación de la librerías.

#### 4.4. Exploración y limpieza de datos con Python

Se llevó a cabo un proceso de exploración y limpieza de datos utilizando la librería **pandas**. Inicialmente se revisaron los tipos de datos y se realizaron las conversiones necesarias para garantizar un tratamiento adecuado de la información. Se separaron las columnas de fecha y hora, y se filtraron los registros correspondientes al periodo 2021-2025. Además, se identificaron y eliminaron registros duplicados, así como aquellos con valores nulos o inconsistentes. Como parte del enriquecimiento de los datos, se generaron nuevas columnas para determinar el día de la semana del accidente y el rango de edad de la víctima. Al finalizar este proceso, la información tratada fue exportada a un archivo .csv para su posterior análisis. Ver código en la Figura 6 y resultados en la Figura 7.

```

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

print(" Análisis de accidentalidad vial en la ciudad de Barranquilla 2022-2025")

columnas = ["Fecha_Accidente", "DIRECCION_ACCIDENTE", "CONDICION_VICTIMA", "GRAVEDAD_ACCIDENTE", "CLASE_ACCIDENTE",
            "SEXO_VICTIMA", "EDAD_VICTIMA", "CANTIDAD_VICTIMA"]
df = pd.DataFrame(resultados, columns=columnas)

print(df)
print(df.shape)
print(df.dtypes)

#Separar la columna Fecha_Accidente en dos columnas: FECHA y HORA
df[['FECHA', 'HORA']] = df['Fecha_Accidente'].str.split(' ', n=1, expand=True)
df['FECHA'] = pd.to_datetime(df['FECHA'], format='%m/%d/%Y', errors='coerce')
#Filtrar los años 2021 a 2025
df = df[(df['FECHA'].dt.year >= 2021) & (df['FECHA'].dt.year <= 2025)]

df.drop('HORA', axis=1, inplace=True)
df.drop('Fecha_Accidente', axis=1, inplace=True)

df = df.astype (
    {
        'DIRECCION_ACCIDENTE': 'string',
        'CONDICION_VICTIMA': 'string',
        'GRAVEDAD_ACCIDENTE': 'string',
        'CLASE_ACCIDENTE': 'string',
        'SEXO_VICTIMA': 'string',
    }
)
print(df.info())
#Antes de eliminar
print(f"\n Duplicados antes de limpiar: {df.duplicated().sum()}")
print(df.duplicated())

# Eliminar duplicados
df = df.drop_duplicates()
print(f"\n Duplicados después de limpiar: {df.duplicated().sum()}")

print("Valores nulos antes de limpieza:\n", df.isnull().sum())

df['EDAD_VICTIMA'] = pd.to_numeric(df['EDAD_VICTIMA'], errors='coerce')
df['CANTIDAD_VICTIMA'] = pd.to_numeric(df['CANTIDAD_VICTIMA'], errors='coerce')

# Eliminar registros con valores cero o nulos o vacíos
df = df.dropna()
df = df[df['CANTIDAD_VICTIMA'] > 0]
df = df[(df['EDAD_VICTIMA'] > 0) & (df['EDAD_VICTIMA'] <= 100)]
df['SEXO_VICTIMA'] = df['SEXO_VICTIMA'].str.strip()
df = df[df['SEXO_VICTIMA'] != '']
df = df.dropna(subset=['SEXO_VICTIMA'])

print("Valores nulos después de limpieza:\n", df.isnull().sum())
print(f"Filas después de limpiar: {len(df)}")

# Nueva columna: Día de la semana
df['DIA_SEMANA'] = df['FECHA'].dt.day_name()

# Nueva columna: Rango de edad
bins = [0, 17, 30, 45, 60, 200]
labels = ['<18', '18-30', '31-45', '46-60', '60+']
df['RANGO_EDAD'] = pd.cut(df['EDAD_VICTIMA'], bins=bins, labels=labels)

#Guardar base de datos limpia con los cambios aplicados
df.to_csv('accidentesbq_limpio.csv', index=False)

```

**Figura 6. Proceso de exploración y limpieza de datos en Python.**

Análisis de accidentalidad vial en la ciudad de Barranquilla 2022-2025							
	Fecha_Accidente	DIRECCION ACCIDENTE \			EDAD_VICTIMA	CANTIDAD_VICTIMA	
0	01/01/2018 12:00:00 AM	CL 87 9H 24			0	57	1
1	01/01/2018 12:00:00 AM	CLLE 119B CRA 11B			1	13	1
2	01/01/2018 12:00:00 AM	CR 8 CL 41			2	32	1
3	01/01/2018 12:00:00 AM	CR 8 CL 41			3	26	1
4	01/01/2018 12:00:00 AM	CR 8 CL 41			4	53	1
...	...	...			...	...	...
16639	04/29/2025 12:00:00 AM	CALLE 110 CON CARRERA 44			16639	27	1
16640	04/29/2025 12:00:00 AM	CALLE 110 CON CARRERA 44			16640	25	1
16641	04/29/2025 12:00:00 AM	CALLE 110 CRA 43			16641	35	1
16642	04/30/2025 12:00:00 AM	CALLE 48 CARRERA 46			16642	29	1
16643	04/30/2025 12:00:00 AM	CALLE 87 CARRERA 73 BARRANQUILLA			16643	25	1
[16644 rows x 8 columns]							
(16644, 8)							
Fecha_Accidente object							
DIRECCION_ACCIDENTE object							
CONDICION_VICTIMA object							
GRAVEDAD_ACCIDENTE object							
CLASE_ACCIDENTE object							
SEXO_VICTIMA object							
EDAD_VICTIMA int64							
CANTIDAD_VICTIMA int64							
dtype: object							
<class 'pandas.core.frame.DataFrame'>							
Index: 10325 entries, 6319 to 16643							
Data columns (total 8 columns):							
# Column Non-Null Count Dtype							
---							
0	DIRECCION_ACCIDENTE	10325	non-null	string			
1	CONDICION_VICTIMA	10325	non-null	string			
2	GRAVEDAD_ACCIDENTE	10325	non-null	string			
3	CLASE_ACCIDENTE	10325	non-null	string			
4	SEXO_VICTIMA	10325	non-null	string			
5	EDAD_VICTIMA	10325	non-null	int64			
6	CANTIDAD_VICTIMA	10325	non-null	int64			
7	FECHA	10325	non-null	datetime64[ns]			
dtypes: datetime64[ns](1), int64(2), string(5)							
memory usage: 726.0 KB							
None							
Duplicados antes de limpiar: 0							
5319	False						
5320	False						
5321	False						
5322	False						
5323	False						
...							
16639	False						
16640	False						
16641	False						
16642	False						
16643	False						
length: 10325, dtype: bool							

**Figura 7. Resultados preliminares de la exploración y limpieza de datos: duplicados, nulos y tipos de datos corregidos.**

## 5. Análisis de datos, resultados y hallazgos

Una vez finalizado el proceso de limpieza, se procedió al análisis estadístico de los datos de accidentalidad con el objetivo de responder las preguntas de investigación y comprobar las hipótesis planteadas. Entre los cálculos realizados se encuentran: el conteo de muertes por tipo de accidente, el cálculo de la tasa de mortalidad por tipo de accidente, y la identificación de las direcciones con mayor número de muertes y accidentes. Asimismo, se determinó el top 10 de zonas con mayor tasa de mortalidad (considerando aquellas con al menos cinco accidentes registrados). Este análisis permitió establecer patrones geográficos y tipológicos en los accidentes de tránsito. **Ver código en la Figura 8.**

```

# Calculos
print("\nConteo de muertes por tipo de accidente")
muertes_por_clase = df[df['GRAVEDAD_ACCIDENTE'] == 'muerto']['CLASE_ACCIDENTE'].value_counts()
print(muertes_por_clase)

print("\nTasa de mortalidad por tipo de accidente - compara cuántas muertes hay en proporción al total de accidentes de cada tipo")
# Total de accidentes por tipo
total_por_clase = df['CLASE_ACCIDENTE'].value_counts()
print(total_por_clase)

# Tasa de mortalidad
tasa_mortalidad = (muertes_por_clase / total_por_clase).sort_values(ascending=False)
print(tasa_mortalidad)

# Asegurar formato uniforme
df['GRAVEDAD_ACCIDENTE'] = df['GRAVEDAD_ACCIDENTE'].str.strip().str.lower() # herido / muerto
df['DIRECCION_ACCIDENTE'] = df['DIRECCION_ACCIDENTE'].str.strip().str.upper()
# Agrupar por dirección y tipo de gravedad (muerto/herido)
zonas = df.groupby(['DIRECCION_ACCIDENTE', 'GRAVEDAD_ACCIDENTE']).size().unstack(fill_value=0)

# Asegurar que las columnas 'muerto' y 'herido' existen (por si no hay datos)
for col in ['muerto', 'herido']:
    if col not in zonas.columns:
        zonas[col] = 0

# Total y tasa de mortalidad:
zonas['TOTAL_ACCIDENTES'] = zonas['muerto'] + zonas['herido'] #TOTAL_ACCIDENTES: suma de muertos y heridos por dirección
zonas['TASA_MORTALIDAD_%'] = (zonas['muerto'] / zonas['TOTAL_ACCIDENTES']) * 100
zonas['TASA_MORTALIDAD_%'] = zonas['TASA_MORTALIDAD_%'].round(2)

# Top 10 direcciones con más muertes
top_muertos = zonas.sort_values('muerto', ascending=False).head(10)
print("\n Zonas con más víctimas fatales:")
print(top_muertos[['muerto', 'TOTAL_ACCIDENTES']])

# Top 10 direcciones con más accidentes (totales)
top_accidentes = zonas.sort_values('TOTAL_ACCIDENTES', ascending=False).head(10)
print("\n Zonas con más accidentes totales:")
print(top_accidentes[['TOTAL_ACCIDENTES', 'muerto']])

# Top zonas con mayor tasa de mortalidad (%)
top_tasa = zonas[zonas['TOTAL_ACCIDENTES'] >= 5].sort_values('TASA_MORTALIDAD_%', ascending=False).head(10)
print("\n Zonas con mayor tasa de mortalidad (%):")
print(top_tasa[['TASA_MORTALIDAD_%', 'TOTAL_ACCIDENTES']])

```

**Figura 8. Análisis de datos y cálculos estadísticos realizados en Python.**

## 6. Interpretaciones de la evidencia y conclusiones:

A partir de los resultados obtenidos durante el análisis de los datos de accidentalidad vial en Barranquilla, se realizaron las siguientes interpretaciones en relación con las hipótesis planteadas:

**6.1. Hipótesis 1 (H1):** “Los accidentes con mayor número de víctimas mortales se concentran en zonas específicas de Barranquilla”.

Para comprobar esta hipótesis, se abordaron preguntas clave como:

- ¿Cuántas muertes se registran por tipo de accidente?
- ¿Qué zonas concentran mayor cantidad de víctimas mortales?
- ¿Cuáles son las direcciones con más accidentes y mayor tasa de mortalidad?

Se desarrolló un análisis agrupando los accidentes por dirección y gravedad (muerto/herido), lo que permitió identificar las zonas con mayor número de muertes y calcular la tasa de mortalidad por ubicación. Esto facilitó la identificación de las áreas críticas con alta incidencia de fatalidades visualizadas en las figuras **Figura 9 y 10**.

```

# Visualización para Top 10 zonas con más muertes:
plt.figure(figsize=(12, 6))
bars = plt.barh(top_muertos.index, top_muertos['muerto'], color='pink')
plt.xlabel('Cantidad de fallecidos')
plt.title('Zonas con mayor cantidad de víctimas fatales')
plt.gca().invert_yaxis()
plt.tight_layout()

# Etiquetas numéricas al final de cada barra
for bar in bars:
    width = bar.get_width()
    plt.text(width + 1, bar.get_y() + bar.get_height()/2,
             str(int(width)), va='center')

plt.show()

# Visualización para la cantidad total de accidentes:
plt.figure(figsize=(12, 6))
bars_total = plt.barh(top_accidentes.index, top_accidentes['TOTAL_ACCIDENTES'], color='skyblue')
plt.xlabel('Cantidad total de accidentes')
plt.title('Zonas con mayor cantidad total de accidentes')
plt.gca().invert_yaxis()
plt.tight_layout()

# Etiquetas numéricas
for bar in bars_total:
    width = bar.get_width()
    plt.text(width + 1, bar.get_y() + bar.get_height()/2,
             str(int(width)), va='center')

plt.show()

# Visualización de la tasa de mortalidad:
plt.figure(figsize=(12, 6))
bars_tasa = plt.barh(top_tasa.index, top_tasa['TASA_MORTALIDAD_%'], color='orange')
plt.xlabel('Tasa de mortalidad (%)')
plt.title('Zonas con mayor tasa de mortalidad (con al menos 5 accidentes)')
plt.gca().invert_yaxis()
plt.tight_layout()

# Etiquetas con porcentaje
for bar in bars_tasa:
    width = bar.get_width()
    plt.text(width + 0.5, bar.get_y() + bar.get_height()/2, f'{width:.2f}%', va='center')

plt.show()

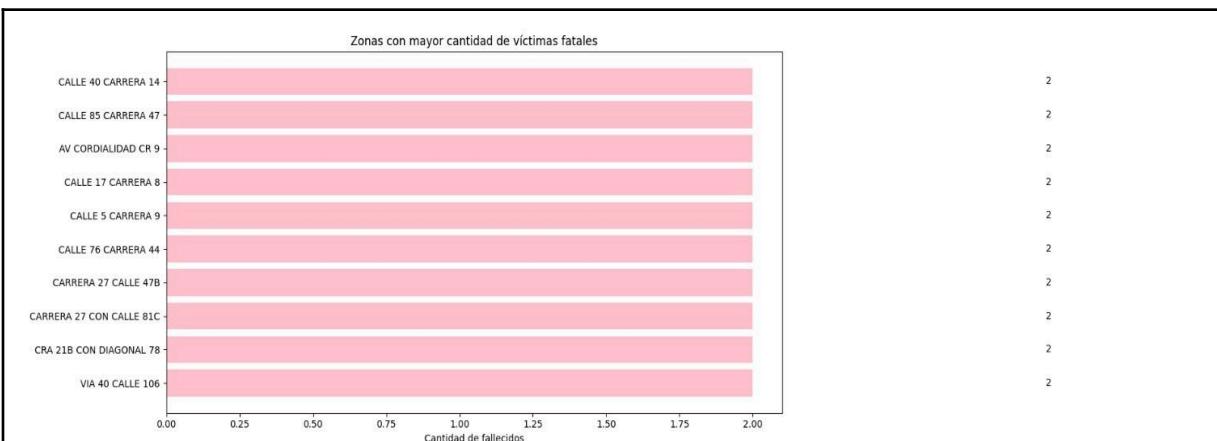
```

**Figura 9. Código para visualización de la Hipótesis 1.**

Conteo de muertes por tipo de accidente		
CLASE_ACCIDENTE		
Choque	115	
Atropello	53	
Otro	7	
Caida Ocupante	3	
Volcamiento	3	
Name: count, dtype: Int64		
Tasa de mortalidad por tipo de accidente - compara cuántas muertes hay en proporción al total de accidentes de cada tipo		
CLASE_ACCIDENTE		
Otro	0.107692	Zonas con mayor tasa de mortalidad (%):
Atropello	0.048007	GRAVEDAD_ACCIDENTE TASA_MORTALIDAD_% TOTAL_ACCIDENTES
Volcamiento	0.026549	DIRECCION ACCIDENTE
Caida Ocupante	0.018868	CALLE 40 CARRERA 14 28.57 7
Choque	0.013002	CALLE 5 CARRERA 9 22.22 9
Incendio	<NA>	CALLE 17 CARRERA 8 22.22 9
Name: count, dtype: Float64		CALLE 30 CARRERA 36 20.00 5
Zonas con más víctimas fatales:		
GRAVEDAD_ACCIDENTE	muerto	TOTAL_ACCIDENTES
DIRECCION ACCIDENTE		
CALLE 40 CARRERA 14	2	7
CALLE 85 CARRERA 47	2	12
AV CORDIALIDAD CR 9	2	2
CALLE 17 CARRERA 8	2	9
CALLE 5 CARRERA 9	2	9
CALLE 76 CARRERA 44	2	10
CARRERA 27 CALLE 47B	2	2
CARRERA 27 CON CALLE 81C	2	3
CRA 21B CON DIAGONAL 78	2	14
VIA 40 CALLE 106	2	4
Zonas con más accidentes totales:		
GRAVEDAD_ACCIDENTE		TOTAL_ACCIDENTES
DIRECCION ACCIDENTE		
AVENIDA CIRCUNVALAR CALLE 96	33	0
CALLE 110 CARRERA 43	27	1
CL 110 CR 9G	23	1
CARRERA 43 CALLE 110	23	0
CALLE 110 CARRERA 9G	21	0
CALLE 47 CARRERA 20	21	0
CALLE 84B CON CARRERA 40A	20	0
CALLE 31 CARRERA 24	20	0
CALLE 110 CRA 9G	19	0
CL 31 CR 24	19	0

**Figura 10. Análisis de Datos y Hallazgos de H1**

A continuación, se muestran las gráficas elaboradas en función de las preguntas planteadas en la **hipótesis 1**



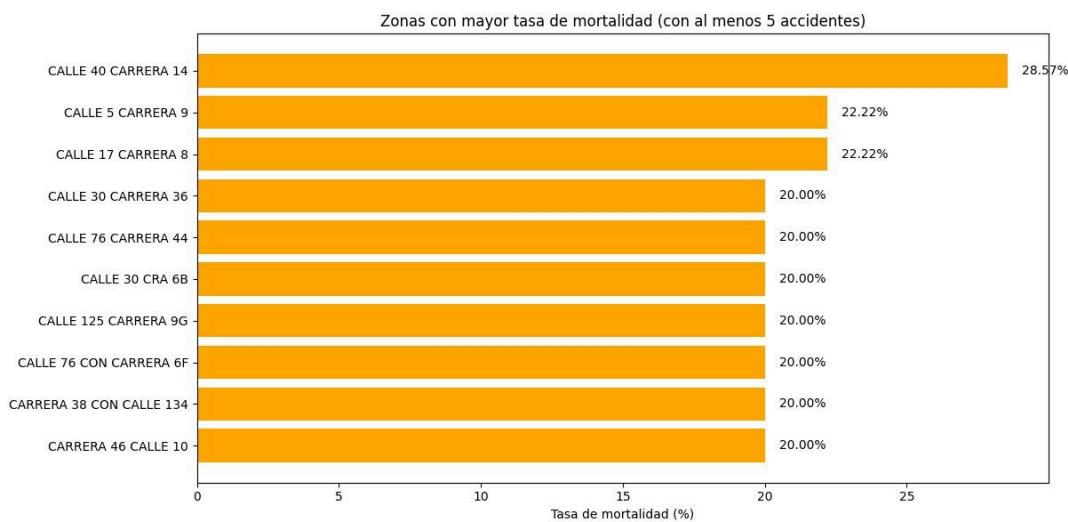
La figura muestra zonas con mayor número de muertes registradas, todas con dos víctimas fatales. Aunque no son las más accidentadas, destacan por su alta letalidad, lo que las convierte en puntos críticos para la intervención.

**Figura 12. Gráfico de Muertes Registradas por tipo de accidente.**



Las zonas con más accidentes reportados se concentran en vías principales, como Avenida Circunvalar, destacando por su alta frecuencia, aunque no necesariamente por su letalidad. Estas áreas requieren medidas de prevención y control del flujo vehicular.

**Figura 13. Gráfico de las zonas con mayor cantidad de accidentes.**



Este gráfico resalta zonas donde, pese a no tener tantos accidentes, la proporción de muertes es alta. Lugares como Calle 40 con Carrera 14 presentan tasas superiores al 28%, lo que indica alta letalidad y necesidad de intervención prioritaria.

**Figura 14. Gráfico de las zonas con mayor tasa de mortalidad.**

A continuación, se presenta un dashboard general con los principales indicadores relacionados con las víctimas de accidentes de tránsito, desagregados por tipo de accidente, condición de la víctima y zonas con mayor concentración de muertes.

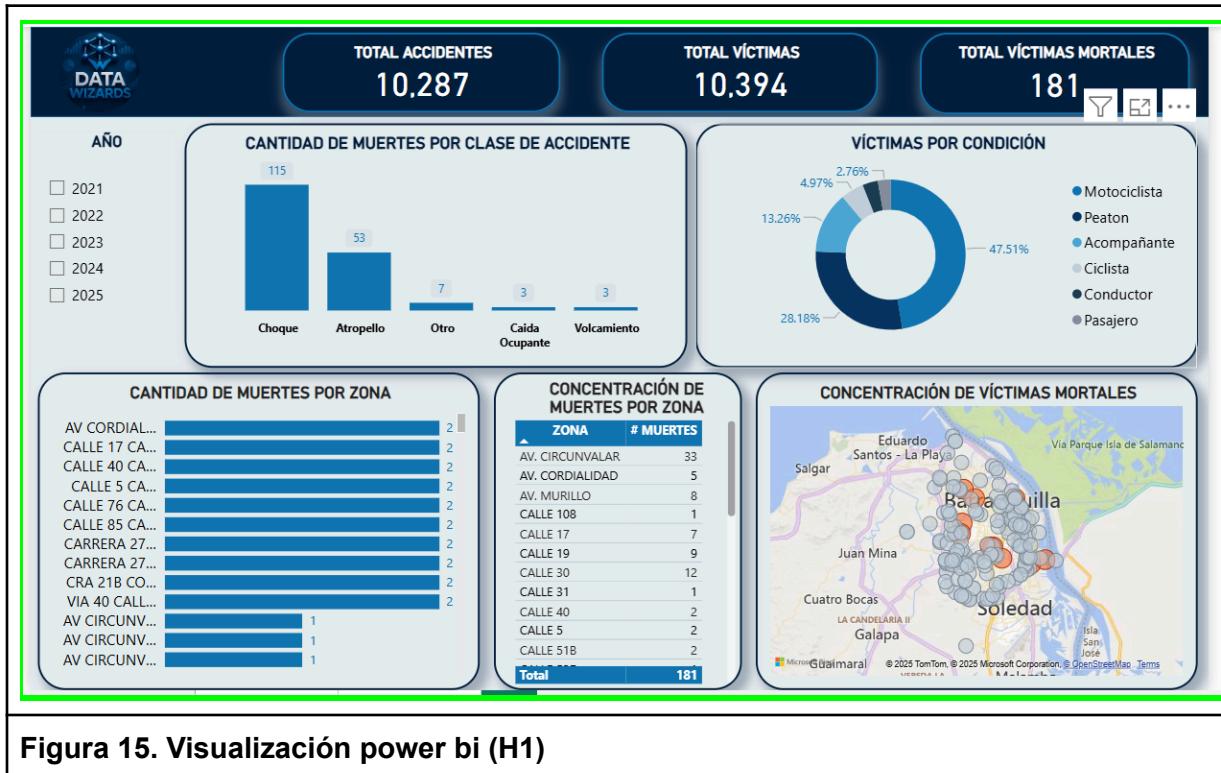


Figura 15. Visualización power bi (H1)

**Conclusión de la Subhipótesis 1 (H1):** “Los accidentes viales con mayor número de víctimas mortales se concentran en zonas específicas de Barranquilla.”

- Los choques son los accidentes más comunes, pero con baja mortalidad.
- Atropellos y "Otros" tienen menor frecuencia, pero mayor tasa de mortalidad.
- Zonas como AV CORDIALIDAD CR 9 y CRA 21B CON DIAGONAL 78 presentan alta letalidad con pocos accidentes.
- Las zonas más transitadas no son siempre las más letales.
- Es clave priorizar zonas críticas para reducir muertes, más que solo el volumen de accidentes.
- Requieren intervenciones distintas: **Alta mortalidad:** Control y respuesta rápida.  
**Alta frecuencia :** Prevención, cultura vial y rediseño urbano.

## Síntesis

Las zonas con más muertes no son necesariamente las más accidentadas, lo que confirma la Hipótesis 1. Se necesita un enfoque diferenciado según el tipo de riesgo.

## 6.2 Hipótesis (H2) Subhipótesis 2:

La gravedad de los accidentes de tránsito está fuertemente relacionada con la edad y el rol de la víctima, donde los motociclistas jóvenes (18-30 años) y peatones mayores (más de 60 años) presentan un riesgo significativamente mayor de resultar con lesiones graves o fatales.

Además del análisis por zonas y tipos de accidente, se realizó un estudio detallado sobre la variable edad de la víctima, con el fin de evaluar su relación con la gravedad del accidente. Para ello, se formularon las siguientes preguntas clave:

- ¿Cuál es la edad promedio de las víctimas de accidentes?
- ¿Cómo varía la edad entre víctimas mortales y heridas?
- ¿La edad de las víctimas sigue una distribución normal?

El análisis se desarrolló mediante scripts en Python implementados en Visual Studio Code. Se calcularon medidas estadísticas generales (media, mediana, moda, desviación estándar y cuartiles) para toda la muestra, y luego se compararon los valores entre víctimas heridas y fallecidas. Adicionalmente, se generó un histograma con distribución KDE diferenciando entre niveles de gravedad, y se aplicaron pruebas estadísticas (Shapiro-Wilk) para evaluar si los datos se ajustaban a una distribución normal.

Este análisis permitió evidenciar patrones de edad asociados a mayor letalidad, como una mayor concentración de muertes en ciertos rangos de edades, especialmente en motociclistas jóvenes y peatones mayores, apoyando así la hipótesis planteada.

**Ver código en la Figura 15 y resultados en la Figura 16.**

```

# 1. Medidas generales
# Estadísticas generales de EDAD_VICTIMA
print("Media:", df['EDAD_VICTIMA'].mean())
print("Mediana:", df['EDAD_VICTIMA'].median())
print("Moda:", df['EDAD_VICTIMA'].mode()[0])
print("Desviación estándar:", df['EDAD_VICTIMA'].std())
print("Cuartiles:\n", df['EDAD_VICTIMA'].quantile([0.25, 0.5, 0.75]))
# 2. Comparar estadísticas entre muertos y heridos:
df.groupby('GRAVEDAD_ACCIDENTE')['EDAD_VICTIMA'].describe()

# Análisis de distribuciones
# Histograma de edades de las víctimas según gravedad del accidente
import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(data=df, x='EDAD_VICTIMA', hue='GRAVEDAD_ACCIDENTE', kde=True)
plt.title("Distribución de edades según gravedad")
plt.show()

# Se ajusta a una distribución normal?
from scipy.stats import shapiro, normaltest

# Solo edades
stat, p = shapiro(df['EDAD_VICTIMA'])
print("Shapiro-Wilk p-value:", p)

```

**Figura 16.** Código para analizar la relación entre edad y gravedad del accidente..

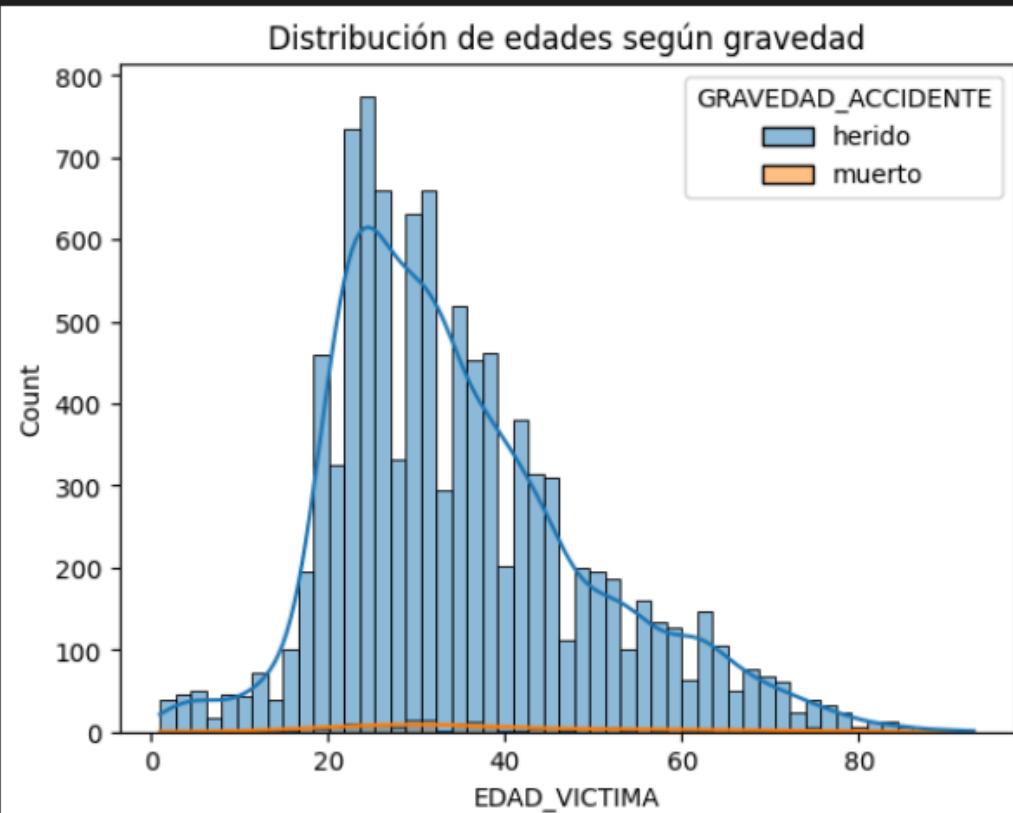
Se calcularon estadísticas descriptivas de la edad de las víctimas, se visualiza su distribución según la gravedad del accidente y se aplica la prueba de normalidad Shapiro-Wilk.

## Análisis de Datos y Hallazgos H2:

```

Media: 34.99727811801303
Mediana: 32.0
Moda: 24
Desviación estándar: 14.618068200895955
Cuartiles:
  0.25    24.0
  0.50    32.0
  0.75    43.0
Name: EDAD_VICTIMA, dtype: float64

```



Shapiro-Wilk p-value: 8.13400190474322e-49

La figura muestra que la mayoría de heridos tienen entre 18 y 40 años, mientras que las muertes aumentan en adultos mayores, especialmente desde los 60 años, evidenciando su mayor vulnerabilidad.

**Figura 17. Resultado obtenido en Visual Studio Code para la H2.**

```

# Visualización de datos

# 1 Distribución de gravedad
plt.figure()
sns.countplot(data=df, x='GRAVEDAD_ACCIDENTE')
plt.title('Distribución de Gravedad del Accidente')
plt.show()

# 2 Gravedad vs Rango de Edad
plt.figure(figsize=(8,5))
sns.countplot(data=df, x='RANGO_EDAD', hue='GRAVEDAD_ACCIDENTE', palette='Set2')
plt.title("Gravedad del accidente por rango de edad")
plt.xlabel("Rango de edad")
plt.ylabel("Cantidad")
plt.legend(title='Gravedad')
plt.tight_layout()
plt.show()

# 3 Gravedad vs Condición de la víctima
plt.figure(figsize=(10,5))
sns.countplot(data=df, x='CONDICION_VICTIMA', hue='GRAVEDAD_ACCIDENTE', palette='Set1')
plt.title("Gravedad del accidente por tipo de víctima")
plt.xlabel("Condición de la víctima")
plt.ylabel("Cantidad")
plt.xticks(rotation=45)
plt.legend(title='Gravedad')
plt.tight_layout()
plt.show()

# 4 Gravedad según edad y condición
plt.figure()
sns.boxplot(data=df[df['GRAVEDAD_ACCIDENTE'] != 'ilesos'], x='CONDICION_VICTIMA', y='EDAD_VICTIMA', hue='GRAVEDAD_ACCIDENTE')
plt.title("Edad y Gravedad según condición")
plt.xticks(rotation=45)
plt.show()

# 5 Mapa de calor
heatmap_data = df.pivot_table(
    values='GRAVEDAD_ACCIDENTE',
    index='CONDICION_VICTIMA',
    columns='RANGO_EDAD',
    aggfunc=lambda x: (x == 'muerto').mean()
)

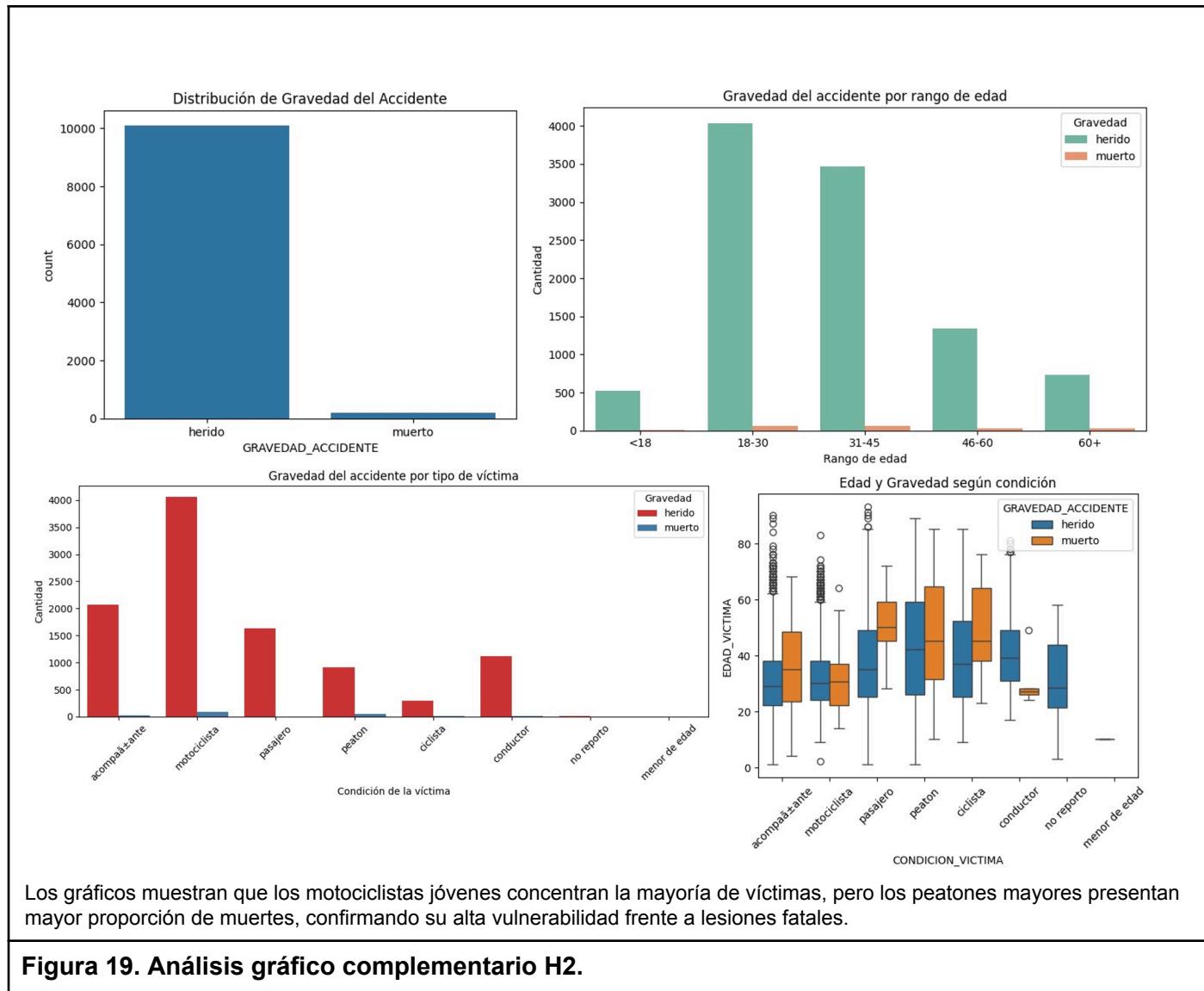
plt.figure(figsize=(10, 6))
sns.heatmap(heatmap_data, annot=True, cmap='Reds', fmt=".2f")
plt.title('Proporción de fallecidos por condición y rango de edad')
plt.ylabel('Condición de la víctima')
plt.xlabel('Rango de edad')
plt.tight_layout()
plt.show()

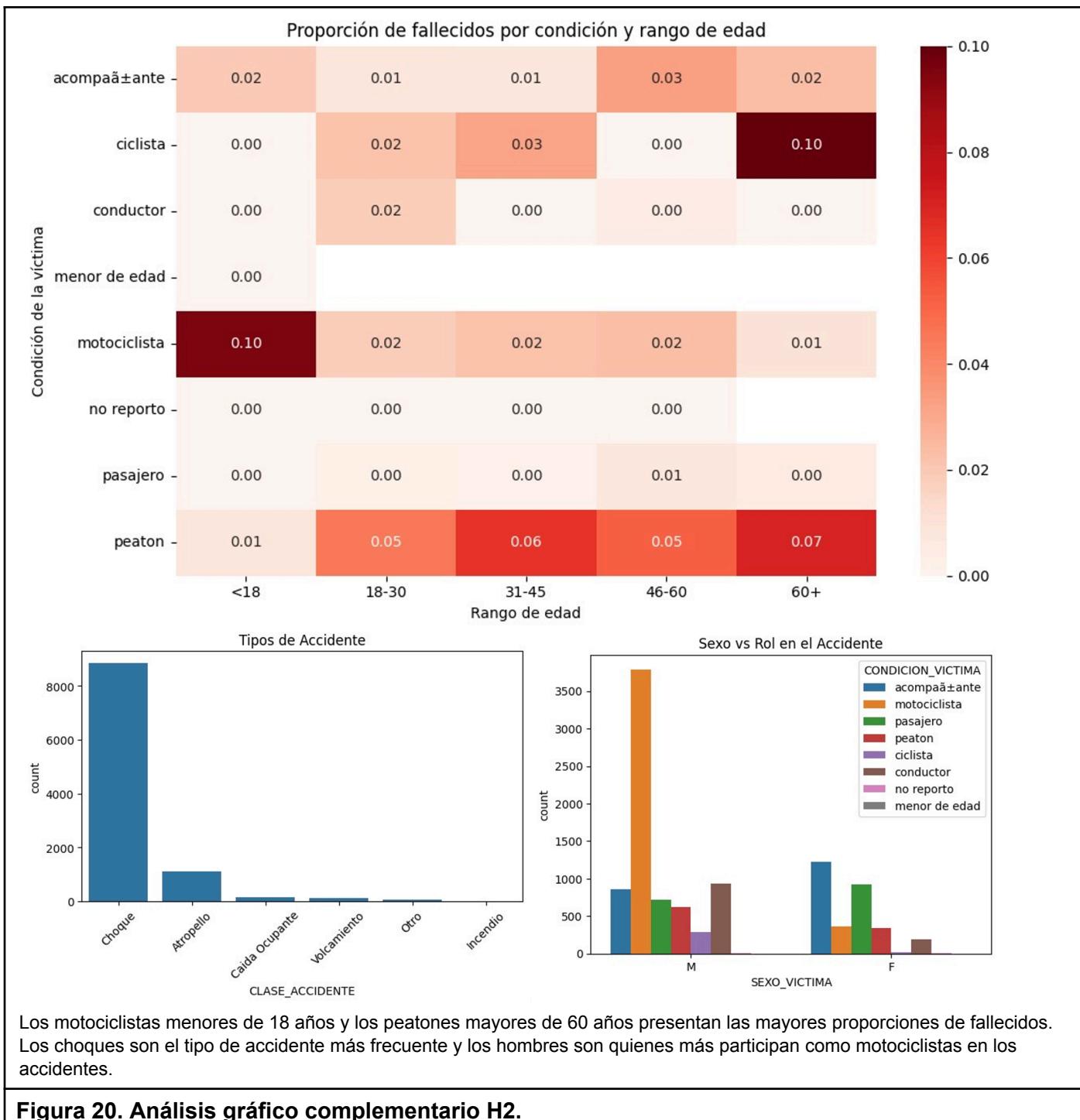
# 6 Distribución por sexo y condición
plt.figure()
sns.countplot(data=df, x='SEXO_VICTIMA', hue='CONDICION_VICTIMA')
plt.title('Sexo vs Rol en el Accidente')
plt.show()

# 7 Distribución por tipo de accidente
plt.figure()
sns.countplot(data=df, x='CLASE_ACCIDENTE', order=df['CLASE_ACCIDENTE'].value_counts().index)
plt.title('Tipos de Accidente')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

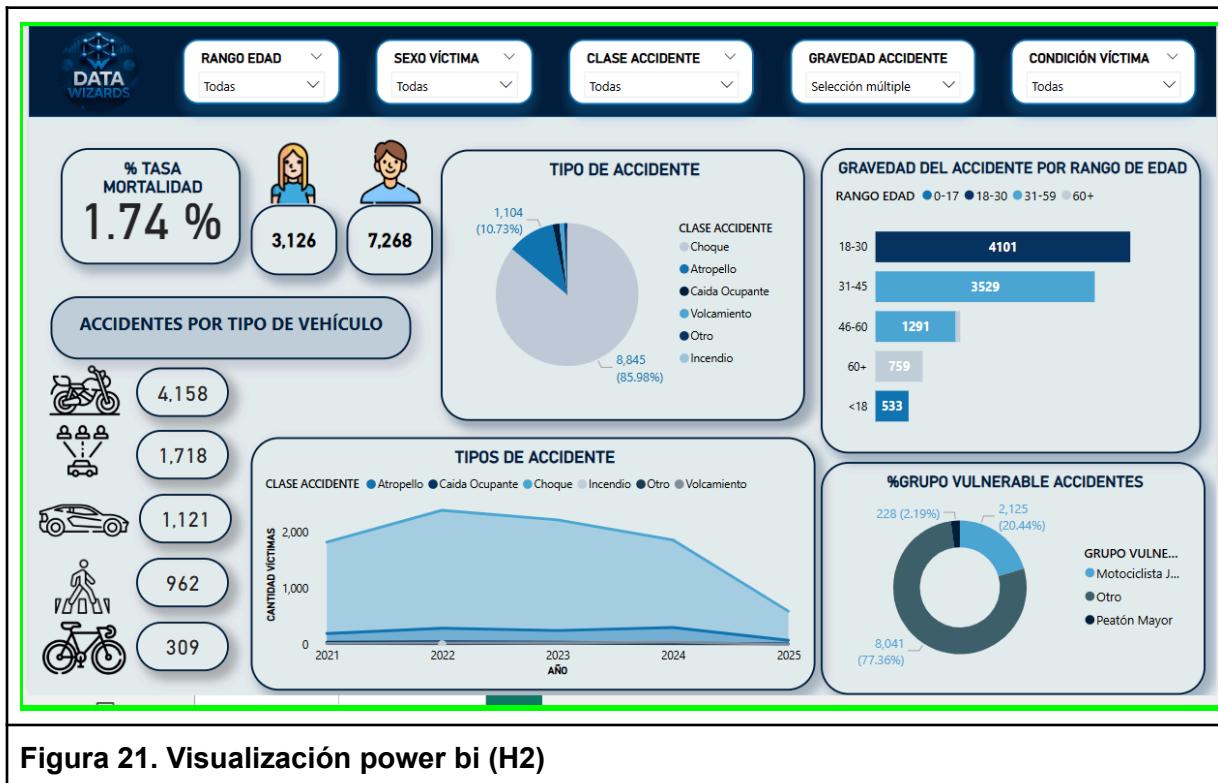
```

**Figura 18.** Script para obtener gráficas de la H2.





A continuación, se presenta el dashboard de análisis de víctimas por características demográficas y tipo de accidente. Este panel permite explorar la relación entre la gravedad de los accidentes y variables como la edad, el rol de la víctima, el tipo de accidente y el tipo de vehículo involucrado. Además, se destaca la participación de grupos vulnerables como motociclistas jóvenes y peatones mayores.



**Subhipótesis 2:** "La gravedad de los accidentes está relacionada con la edad y el rol de la víctima: los motociclistas jóvenes (18-30) y peatones mayores (60+) tienen mayor riesgo de lesiones graves o fatales".

### Conclusiones generales – Subhipótesis 2

- Motociclistas jóvenes (18-30): Tasa de mortalidad del 1.84%, apenas sobre el promedio (1.76%).
- Peatones mayores de 60 años: Tasa de 7.05%, muy superior al resto. Son el grupo más vulnerable.
- Por condición de víctima:
  - Peatones y ciclistas tienen mayores tasas de mortalidad.
  - Conductores y pasajeros presentan las más bajas.
- Por edad:
  - El grupo 60+ años lidera con 3.16%.
  - Le siguen 46-60 y menores de 18 años (posiblemente por fragilidad física).

### La hipótesis se confirma:

- La edad avanzada y la condición de peatón implican alto riesgo de muerte.
- Motociclistas jóvenes también son vulnerables, pero en menor grado.
- Se requieren estrategias diferenciadas para cada perfil de riesgo.

## 7. Modelos

### 7.1 Modelo Random Forest.

Se implementa un modelo de clasificación supervisada utilizando Random Forest para predecir la gravedad de los accidentes de tránsito. Se transforman variables categóricas, se dividen los datos en entrenamiento y prueba, y se evalúa el desempeño del modelo con métricas como accuracy y reporte de clasificación como se muestra en la **Figura 21**.

```
# Modelo 1 Random Forest para predecir la gravedad del accidente (modelo de clasificación supervisada)

# Importar librerías necesarias
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

#Cargar y limpiar datos,
df['GRAVEDAD_BINARIA'] = df['GRAVEDAD_ACCIDENTE'].apply(lambda x: 1 if x == 'muerto' else 0)

#Variables predictoras (puedes ajustar según tus datos),
variables = ['SEXO_VICTIMA', 'EDAD_VICTIMA', 'CONDICION_VICTIMA', 'CLASE_ACCIDENTE', 'DIA_SEMANA']

#Quitar filas con valores faltantes,
df_modelo = df[variables + ['GRAVEDAD_BINARIA']].dropna()

#convertimos las variables de texto a numero (columna binaria 0 o 1),
df_modelo = pd.get_dummies(df_modelo, columns=['SEXO_VICTIMA', 'CONDICION_VICTIMA', 'CLASE_ACCIDENTE', 'DIA_SEMANA'], drop_first=True)

#Separar X e y,
X = df_modelo.drop('GRAVEDAD_BINARIA', axis=1) #todo menos la columna de gravedad
y = df_modelo['GRAVEDAD_BINARIA']

#Dividir en conjunto de entrenamiento y prueba,
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#entrenamiento del modelo
modelo = RandomForestClassifier(n_estimators=100, random_state=42)
modelo.fit(X_train, y_train)

#Hacer predicciones y evaluar el modelo
y_pred = modelo.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nReporte de clasificación:\n", classification_report(y_test, y_pred))
```

**Figura 22. Modelo Random Forest para predecir la gravedad del accidente.**

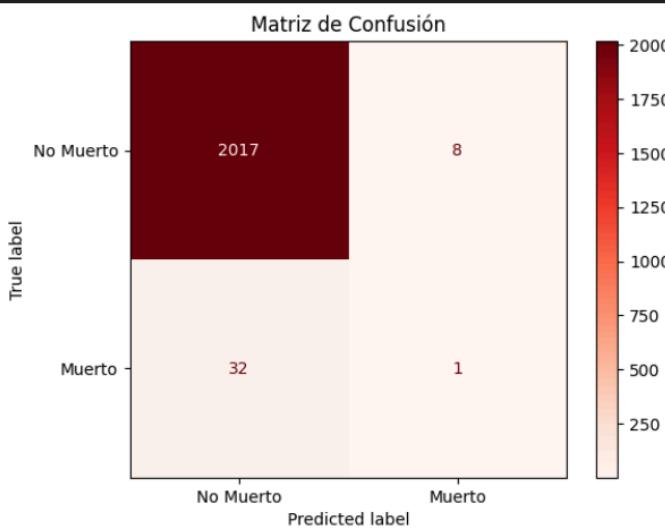
Accuracy: 0.9805636540330418

#### Reporte de clasificación:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	2025
1	0.11	0.03	0.05	33
accuracy			0.98	2058
macro avg	0.55	0.51	0.52	2058
weighted avg	0.97	0.98	0.98	2058

**Figura 23. Resultados random forest**

```
from sklearn.metrics import ConfusionMatrixDisplay
ConfusionMatrixDisplay.from_estimator(modelo, X_test, y_test, display_labels=["No Muerto", "Muerto"], cmap="Reds")
plt.title("Matriz de Confusión")
plt.show()
```



**Figura 24. Matriz de confusión**

## 7.2 Conclusiones y recomendaciones Random forest

El modelo de Random Forest fue desarrollado para predecir la gravedad de un accidente de tránsito en Barranquilla, clasificando los casos como 'herido' (0) o 'muerto' (1), usando variables como sexo, edad, condición de la víctima, tipo de accidente y día de la semana.

### Evaluación del Modelo (Resultados Reales)

- Exactitud general del modelo: 98.06%
- La clase 'herido' (0) fue predicha con muy alta precisión (0.98), recall (1.00) y f1-score (0.99).
- La clase 'muerto' (1), sin embargo, tiene un bajo desempeño: precisión 0.11, recall 0.03 y f1-score 0.05.
- Esto indica un problema de **desbalance de clases**, ya que los fallecidos son una minoría en los datos.

### Variables más relevantes

- Aunque no se imprimió el ranking exacto de importancia, por el tipo de modelo se estima que las variables con mayor influencia son:
- Edad de la víctima
- Condición de la víctima (peatón, conductor, ciclista, etc.)
- Clase del accidente

### Conclusiones

- Exactitud (accuracy): 98.06%: A primera vista, parece excelente. Sin embargo, esta métrica puede ser engañosa en casos con clases desbalanceadas.
- El modelo casi no identifica correctamente a los muertos.
  - Clasifica prácticamente todos los casos como "heridos".
  - Esto ocurre porque el número de fallecidos (33) es muy inferior al de heridos (2025).

### Recomendaciones

- Aplicar campañas preventivas focalizadas en los grupos más vulnerables, como peatones mayores y motociclistas jóvenes.
- Reforzar medidas de control y vigilancia en los días con mayor riesgo de accidentes graves.
- Para mejorar el modelo: balancear las clases mediante técnicas como sobre muestreo (SMOTE) o penalización de clases.

- Utilizar este modelo como una herramienta de alerta temprana para identificar condiciones críticas y prevenir accidentes fatales.

#### Análisis Matriz de Confusión:

Adicional realizamos un análisis implementando la matriz de Confusión donde vemos el resumen del rendimiento de un modelo de clasificación al mostrar la cantidad de predicciones correctas e incorrectas para cada clase arrojándonos lo siguiente:

- Predijo correctamente 2017 heridos, pero solo 1 muerto.
- Falló en 32 casos donde la víctima real fue un fallecido.

#### Conclusiones

- El modelo tiene buen desempeño para la clase dominante (herido), pero falla en la clase crítica (muerto).
- Esta situación es común en datasets desbalanceados, y puede generar falsas expectativas en contextos reales.
- Aunque útil para describir patrones generales, no debe implementarse sin ajustes si se desea prevenir muertes.

#### Recomendaciones Técnicas

- Aplicar técnicas de balanceo como SMOTE o `class_weight='balanced'`.
- Medir el rendimiento con métricas como `recall` o `ROC-AUC`, no solo `accuracy`.
- Probar modelos alternativos como Logistic Regression o XGBoost con penalización por clase.

#### Recomendaciones Prácticas

- No utilizar el modelo como sistema de alerta sin ajustar su sensibilidad a muertes.
- Complementar con análisis geoespacial o clustering para identificar zonas de riesgo.
- Reentrenar el modelo con más datos o múltiples niveles de gravedad.

### 7.3 Modelo K-means

Este código implementa un modelo de clustering K-means que agrupa zonas de accidentes según la frecuencia y el día de la semana, con el objetivo de identificar patrones temporales y espaciales en la ocurrencia de siniestros viales. Utiliza datos preprocesados y visualiza los resultados en un gráfico de dispersión coloreado por clúster.

```
# Modelo 2 K-means: Clustering de zonas por frecuencia de accidentes

from sklearn.cluster import KMeans
import numpy as np

# Preparamos los datos: agrupamos por dirección y día de la semana
zonas_frecuencia = df.groupby(['DIRECCION_ACCIDENTE', 'DIA_SEMANA']).size().reset_index(name='FRECUENCIA')

# Convertimos los días a números para el modelo
dias_map = {'Monday':0, 'Tuesday':1, 'Wednesday':2, 'Thursday':3, 'Friday':4, 'Saturday':5, 'Sunday':6}
zonas_frecuencia['DIA_NUM'] = zonas_frecuencia['DIA_SEMANA'].map(dias_map)

# Seleccionamos las variables para clustering
X_cluster = zonas_frecuencia[['FRECUENCIA', 'DIA_NUM']]

# Elegimos el número de clusters (puedes ajustar este valor)
kmeans = KMeans(n_clusters=4, random_state=42)
zonas_frecuencia['CLUSTER'] = kmeans.fit_predict(X_cluster)

# Visualización de los clusters
plt.figure(figsize=(10,6))
sns.scatterplot(data=zonas_frecuencia, x='DIA_NUM', y='FRECUENCIA', hue='CLUSTER', palette='Set2')
plt.title('Clusters de frecuencia de accidentes por día y zona')
plt.xlabel('Día de la semana (0=Lunes)')
plt.ylabel('Frecuencia de accidentes')
plt.show()

# Mostrar las zonas más críticas de cada cluster
for c in zonas_frecuencia['CLUSTER'].unique():
    top_zonas = zonas_frecuencia[zonas_frecuencia['CLUSTER'] == c].sort_values('FRECUENCIA', ascending=False).head(3)
    print(f"\nCluster {c} - Zonas y días más críticos:")
    print(top_zonas[['DIRECCION_ACCIDENTE', 'DIA_SEMANA', 'FRECUENCIA']])
```

**Figura 25. Modelo K-means**

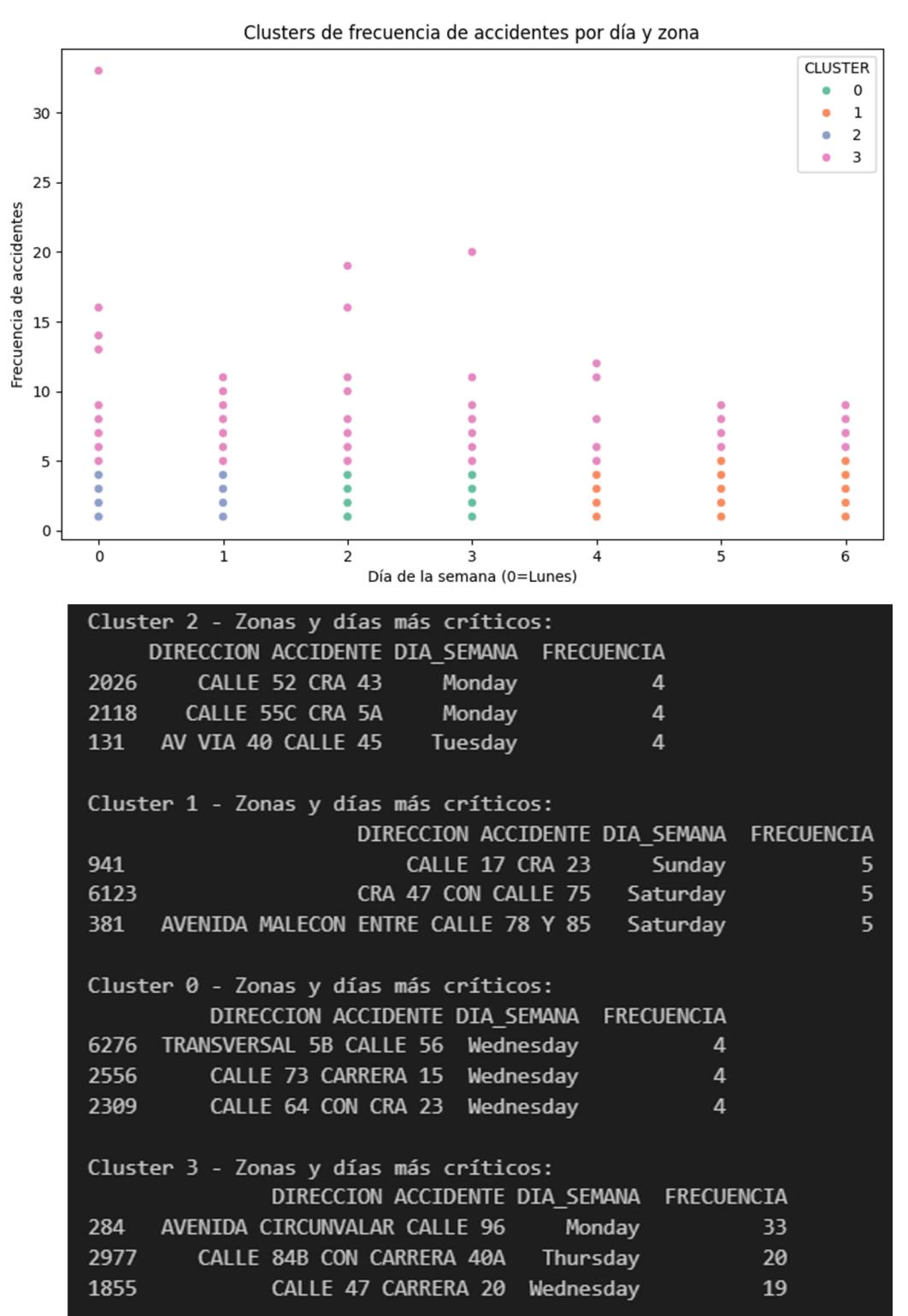


Figura 26. Hallazgos Kmeans

## 7.4 Conclusión y recomendaciones del modelo K-means

Este modelo usa K-Means para clasificar zonas de Barranquilla según la frecuencia de accidentes y el día de la semana en que ocurren.

Se agruparon los datos por dirección y día, luego se transformaron los días en valores numéricos. Se aplicó K-Means con 4 clusters.

El objetivo es identificar grupos de zonas con patrones similares para priorizar acciones de prevención vial.

### Conclusiones Principales

- El modelo identificó 4 clusters de comportamiento.
- El Cluster 3 contiene las zonas con más frecuencia de accidentes, especialmente en días laborales.
- El Cluster 1 destaca por accidentes en fines de semana.
- El Cluster 0 agrupa zonas críticas los miércoles.
- El Cluster 2 refleja zonas con menor frecuencia de accidentes.

### Recomendaciones Generales por Clusters

- Priorizar las zonas del Cluster 3 para intervención inmediata.
- Reforzar vigilancia en fines de semana para zonas del Cluster 1.
- Implementar control de tráfico los miércoles para zonas del Cluster 0.
- Monitorear constantemente la evolución de estos clusters para anticipar riesgos.

### Recomendación Automática

El cluster más crítico es el 3, con mayor frecuencia de accidentes.

Se recomienda priorizar las siguientes zonas y días para intervenciones:

- Zona: AVENIDA CIRCUNVALAR CALLE 96 | Día: Monday | Frecuencia: 33
- Zona: CALLE 84B CON CARRERA 40A | Día: Thursday | Frecuencia: 20
- Zona: CALLE 47 CARRERA 20 | Día: Wednesday | Frecuencia: 19
- Zona: CARRERA 43 CALLE 110 | Día: Monday | Frecuencia: 16
- Zona: VIA 40 CALLE 66 | Día: Wednesday | Frecuencia: 16

### Estrategias Sugeridas

- Aumentar la vigilancia y controles de tránsito en estos puntos y días.
- Implementar campañas de prevención y educación vial dirigidas a los usuarios de estas zonas.
- Mejorar la señalización y condiciones viales en los lugares identificados.

### Recomendaciones

Se recomienda priorizar las siguientes zonas y días:

- AVENIDA CIRCUNVALAR CALLE 96, Lunes, Frecuencia: 33
- CALLE 84B CON CARRERA 40A, Jueves, Frecuencia: 20
- CALLE 47 CARRERA 20, Miércoles, Frecuencia: 19

## 8. Referencias

**Datos Abiertos Colombia.** (2025). Accidentalidad en Barranquilla – Víctimas (2021–2025). Base de datos descargada del portal oficial del Gobierno de Colombia.  
<https://www.datos.gov.co>

**Alcaldía de Barranquilla.** Accidentalidad Barranquilla – Detalle de Vehículos. Datos Abiertos Colombia. Disponible en:

<https://www.datos.gov.co/Transporte/Accidentalidad-Barranquilla-Detalle-de-Vehiculos/9nhr-4agv>

**Data Wizards.** (2025). Análisis de accidentalidad vial en Barranquilla (2021–2025). Proyecto académico desarrollado en el marco del programa **Conexión Mujeres TIC**.