Minería de datos: modelos no supervisados

Sesión 1: Introducción y Definiciones

David Bacca Morales

hbaccamo@poligran.edu.co

Objetivos de la sesión

- Comprender la clasificación de Machine Learning en supervisado y no supervisado.
- Familiarizarse con los conceptos clave de Machine Learning no supervisado.
- Identificar y formular problemas de minería de datos.

Introducción a Machine Learning y su clasificación (supervisado vs no supervisado).

¿Qué es el Machine Learning?

Machine Learning es una rama de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de la experiencia sin ser explícitamente programados. Utiliza algoritmos para identificar patrones en datos y hacer predicciones o tomar decisiones basadas en esos patrones.

Método Supervisado. Utilizan datos etiquetados para entrenar un modelo. Se les proporcionan ejemplos con sus correspondientes etiquetas para que el modelo aprenda a predecir la etiqueta de nuevos datos.

Método No Supervisado. No utilizan datos etiquetados. El modelo intenta encontrar estructuras o patrones ocultos en los datos sin orientación explícita.

Aprendizaje por refuerzo. Técnica en la que un agente aprende a tomar decisiones a través de ensayo y error, recibiendo recompensas o castigos por sus acciones, con el objetivo de maximizar una recompensa acumulada a lo largo del tiempo.

Introducción a Machine Learning y su clasificación (supervisado vs no supervisado).

Método Supervisado: Clasificación de correos electrónicos como spam o no spam

• **Descripción**: Etiquetas “spam” o “no spam”.

• **Entrenamiento**: Identificación de patrones y características:
- palabras clave,
- estructura del correo,
- remitente, etc.

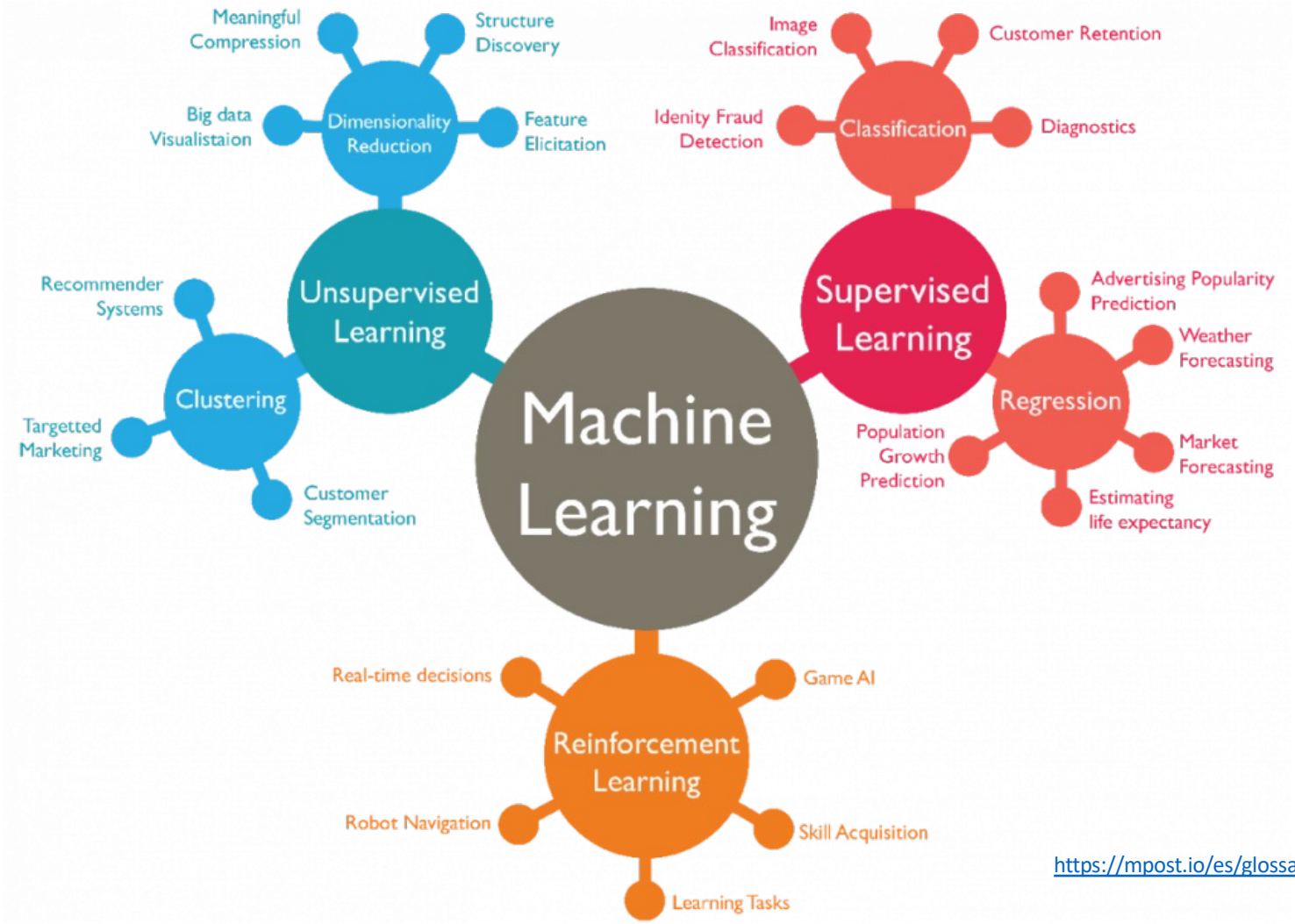
• **Aplicación**: Clasifica nuevo correo como “spam” o “no spam” con lo que aprendió.

Método No Supervisado: Agrupamiento de clientes en una tienda en línea

• **Descripción**: Información sin etiquetas, solo datos del comportamiento de compra de los clientes
- historial de compras,
- frecuencia de visitas,
- monto gastado, etc.

• **Entrenamiento**: Buscar patrones y similitudes en los datos para agrupar a los clientes en diferentes segmentos. Clientes con características como:
- productos de alta gama
- económicos ocasionalmente.

• **Aplicación**: Personalización de estrategias de marketing, promociones específicas a cada grupo según sus patrones de comportamiento.



https://mpost.io/es/glossary/machine-learning/

Programación tradicional vs. Machine learning

Programación Tradicional

Proceso: El programador escribe explícitamente las reglas y lógica para que el sistema siga

Entrada: Datos y reglas.
Salida: Resultados.

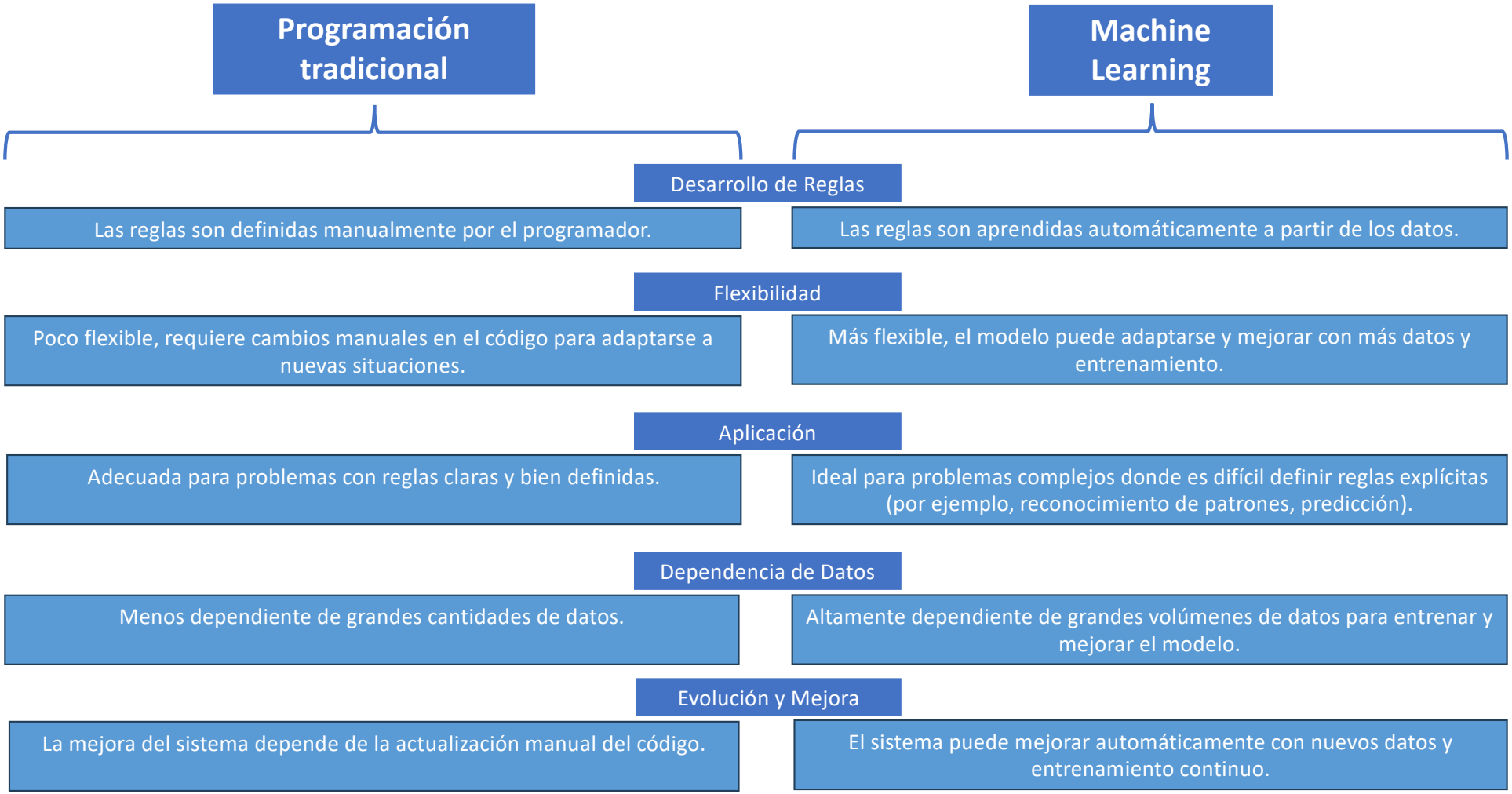
Ejemplo: Crear un programa que calcule el impuesto sobre las ventas. El programador define las reglas de cálculo (porcentaje del impuesto, etc.).

Machine Learning

Proceso: El sistema aprende patrones y reglas a partir de ejemplos y datos.

Entrada: Datos y resultados esperados (para supervisado) o solo datos (para no supervisado).
Salida: Modelo que puede hacer predicciones o identificar patrones.

Ejemplo: Entrenar un modelo para clasificar correos electrónicos como spam o no spam. El sistema aprende de ejemplos etiquetados de correos para hacer predicciones sobre nuevos correos.


Programación tradicional
Machine Learning
Desarrollo de Reglas
Las reglas son definidas manualmente por el programador.
Las reglas son aprendidas automáticamente a partir de los datos.
Flexibilidad
Poco flexible, requiere cambios manuales en el código para adaptarse a nuevas situaciones.
Más flexible, el modelo puede adaptarse y mejorar con más datos y entrenamiento.
Aplicación
Adecuada para problemas con reglas claras y bien definidas.
Ideal para problemas complejos donde es difícil definir reglas explícitas (por ejemplo, reconocimiento de patrones, predicción).
Dependencia de Datos
Menos dependiente de grandes cantidades de datos.
Altamente dependiente de grandes volúmenes de datos para entrenar y mejorar el modelo.
Evolución y Mejora
La mejora del sistema depende de la actualización manual del código.
El sistema puede mejorar automáticamente con nuevos datos y entrenamiento continuo.

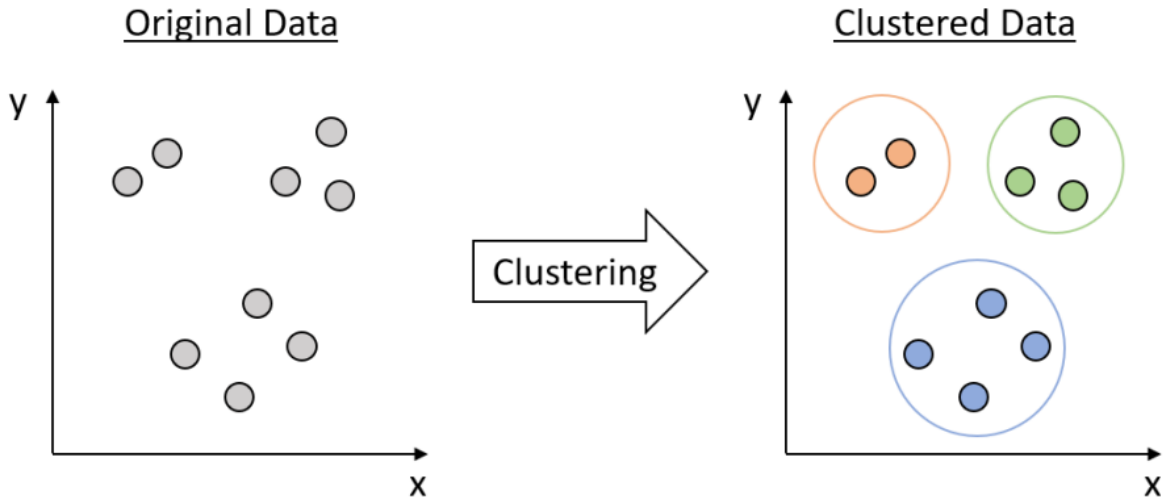
Definiciones clave en Machine Learning no supervisado.

Clustering (Agrupamiento)

Definición: Técnica para agrupar datos en grupos (clusters) basados en la similitud. Los elementos dentro de un mismo cluster son más similares entre sí que con los elementos de otros clusters.

Ejemplos

- En marketing, agrupar clientes según su comportamiento de compra para identificar segmentos de mercado.
- Agrupar áreas geográficas en función de características como demografía, ingresos, y niveles educativos.



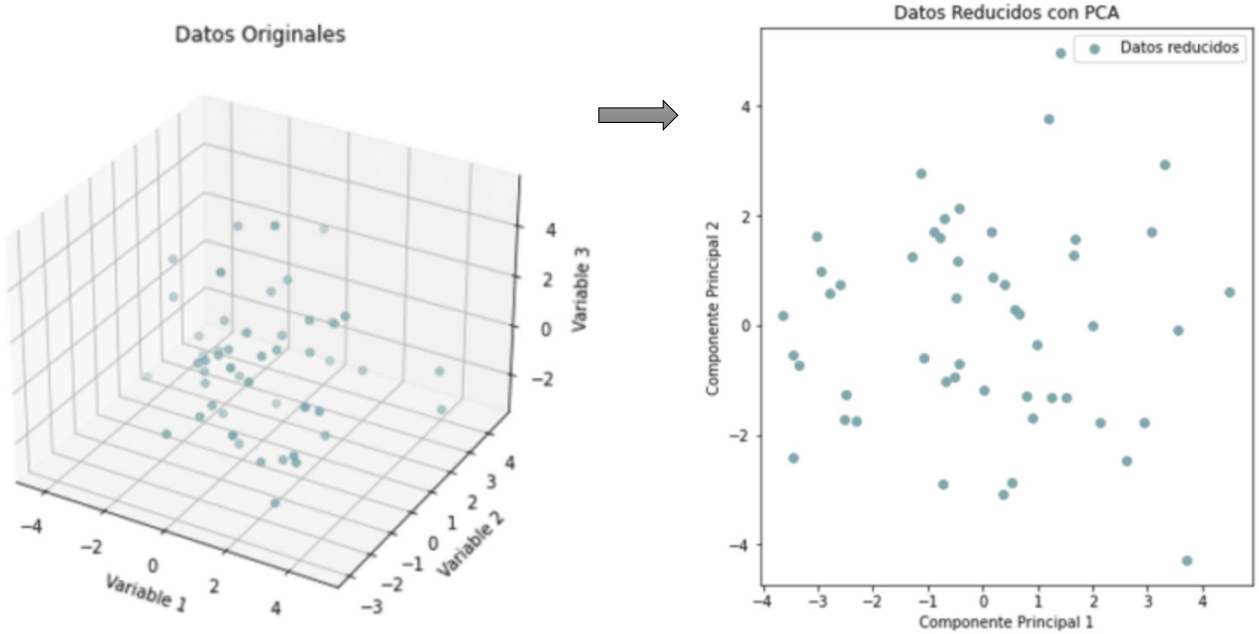
Definiciones clave en Machine Learning no supervisado.

Reducción de Dimensionalidad

Definición: Técnicas para reducir el número de variables (dimensiones) en un dataset, manteniendo la mayor cantidad de información relevante posible. Esto facilita la visualización y el análisis de datos complejos.

Ejemplos

- En biología, reducir el número de características genéticas para clasificar especies de manera más eficiente.
- Utilizar PCA para reducir el número de píxeles (dimensiones) en imágenes, manteniendo las características más importantes.



Definiciones clave en Machine Learning no supervisado.

Reglas de Asociación

Definición: Métodos para descubrir relaciones interesantes entre variables en grandes bases de datos. Utilizado para identificar patrones de co-ocurrencia.

Ejemplos

- En retail, identificar productos que se compran juntos frecuentemente (como pan y leche) para mejorar las estrategias de venta cruzada.

- Descubrir patrones en las preferencias de visualización de los usuarios para recomendar series o películas.

Análisis Exploratorio de Datos (EDA) en Minería de Datos

Definición

El Análisis Exploratorio de Datos (EDA) es un enfoque crítico en la minería de datos que permite a los analistas comprender mejor la estructura, las características y las relaciones en un conjunto de datos antes de aplicar técnicas de modelado más avanzadas.

Objetivos del EDA

1. Comprender las Características del Conjunto de Datos: Identificar distribuciones, valores atípicos, y patrones generales.
2. Detección de Anomalías: Identificar datos anómalos o errores que puedan afectar el análisis.
3. Resumen de Relaciones entre Variables: Explorar correlaciones y dependencias entre diferentes variables.
4. Preparación de Datos para Modelado: Guiar el proceso de limpieza y transformación de datos, incluyendo la selección de características relevantes.

Pasos en el EDA

1. Describir los Datos
    - Estadísticas Descriptivas
    - Distribuciones

2. Visualización de Datos
    - Gráficos de Barras y Histogramas
    - Diagramas de Caja (Boxplots)
    - Diagramas de Dispersión (Scatter Plots)
    - Mapas de Calor (Heatmaps)

3. Detección de Valores Atípicos y Anomalías
    - Identificación de Outliers
    - Análisis de Consistencia de Datos

4. Análisis de Correlaciones
    - Matrices de Correlación
    - Pruebas de Correlación

5. Transformación y Limpieza de Datos
    - Tratamiento de Datos Faltantes
    - Normalización y Escalado

Caso práctico – Análisis exploratorio de datos

Accidentes de tránsito en New York

En este caso práctico haremos la exploración de un conjunto de datos aplicando una de las etapas de los procesos de ciencia de datos. Haremos transformación y limpieza de los datos. Entenderemos cómo extraer valor de los datos desde una perspectiva exploratoria.

El caso estará estructurado así:

1. Explorar la estructura de los datos.
2. Hacer un análisis exploratorio y descriptivo de las principales variables.
3. Sacar conclusiones e hipótesis de análisis a partir de los hallazgos.

Caso práctico – Análisis exploratorio de datos

Contexto: New York ha tenido un incremento en el número de accidentes de tránsito constante y se requiere analizar datos de estos accidentes del periodo Enero 2018 - Agosto 2019.

Problema: Identificar patrones en los datos que permita tomar decisiones informadas dirigidas hacia la planeación de políticas públicas de carácter preventivo para disminuir el número o gravedad de los accidentes.

Algunas de las preguntas que se desea responder con el conjunto de datos son:

- ¿Se ha incrementado el número de accidentes en el periodo de observación?
- ¿Qué podemos concluir acerca del número de accidentes por mes?
- ¿Existen patrones horarios en los accidentes?
- ¿Existen patrones en los accidentes según el día de la semana?
- ¿Existen patrones accidentales por vecindario?
- ¿Existen patrones horarios en la accidentalidad por vecindario?
- ¿Cuáles son las causas más comunes de los accidentes?
- ¿Cuáles son los tipos de vehículos involucrados en más accidentes?
- ¿Pueden existir patrones, de factores o vehículos, diferencial para vecindarios?

Caso práctico – Análisis exploratorio de datos

Se tienen las variables:

BOROUGH: Vecindario donde ocurrió el accidente

COLLISION_ID: ID del accidente

CONTRIBUTING FACTOR VEHICLE (1, 2, 3, 4, 5): Razones del accidente

CROSS STREET NAME: La calle cruzada más cercana en la que pasó el accidente

DATE: Fecha del accidente

TIME: Hora del accidente

LATITUDE: Latitud del accidente

LONGITUDE: Longitud del accidente

NUMBER OF (CYCLISTS, MOTORISTS, PEDESTRIANS) INJURED: Número de heridos de cada tipo

NUMBER OF (CYCLISTS, MOTORISTS, PEDESTRIANS) KILLED: Número de muertos de cada tipo

ON STREET NAME: Calle de accidente

VEHICLE TYPE CODE (1, 2, 3, 4, 5): Tipo de vehículo involucrado en el accidente

ZIP CODE: Código zip del accidente

Ejercicios

1. ¿Qué podemos decir sobre el número de personas heridas y muertes en los accidentes? Exploremos el comportamiento del total de heridos y total de muertes en todos los accidentes. (Crear una nueva columna que sume todos los herido y otra que sume todas las muertes).
2. Exploremos las muertes y heridas por peatón, ciclista y motociclista. ¿Existen diferencias en estos? ¿En qué casos mueren más personas durante los accidentes? ¿En qué casos resultan más heridos?
3. ¿Existen patrones en las muertes y heridos en accidentes por vecindario? ¿En cuál vecindario se presentaron más muertes? ¿En cuál menos?
4. A partir de los análisis anteriores, ¿qué recomendaciones sobre la planeación de políticas preventivas frente a la accidentalidad se podrían hacer para reducir los accidentes y su gravedad?
5. ¿Qué otras variables o análisis exploratorios podríamos hacer para complementar este análisis?
6. ¿Cuáles podrían ser los siguientes pasos para el análisis de este problema en búsqueda de soluciones?

Conclusiones

- La exploración y descripción inicial de los datos es una gran herramienta para encontrar patrones de interés del problema.
- Encontramos patrones en los datos que ayudan a generar hipótesis sobre el comportamiento de los accidentes.
- Resultan ser relevantes factores cómo el día de la semana, hora y vecindario para analizar el comportamiento de la accidentalidad.
- Esta descripción puede ser una guía importante para definir pasos siguientes en un análisis con mayor profundidad.