Instituto Politécnico Nacional Escuela Superior de Cómputo

Data Mining

Práctica 2: Limpieza de datos y exploración básica

Graciano Herrera Gabriel
Meza Zamora Abraham Manuel

22 de marzo de 2022

Índice

1. Introducción

1.1. Explicación de la propuesta de solución

Comprender el alcance del análisis exploratorio de datos y la limpieza de datos, la visualización de datos como herramienta para identificar hallazgos en una muestra de datos por arriba de los 10 mil registros.

1.2. Introducción

Los repositorios de datos permiten la compartición, reutilización y localización de los datos para el aprendizaje y descubrimiento del nuevo conocimiento dentro de las organizaciones. Los datos deben estar limpios y en el formato idóneo para ser analizados.

En la era actual la informatización ha llevado al desarrollo de sistemas de información que se caracterizan por el manejo de grandes volúmenes de datos. Tal acumulación de datos propicia la ocurrencia de anomalías e impurezas, fenómeno que distorsiona los resultados obtenidos de la interpretación y análisis de los datos, y provoca, como consecuencia, la elevación de los costos y la disminución de los beneficios de su procesamiento.

La anomalía es una propiedad de los valores de los datos, que ocasiona una representación errónea del mini mundo que estos reflejan. Los datos que contienen anomalías, se denominan datos erróneos o sucios y su presencia puede obstaculizar el uso efectivo y eficiente de la información.

Una anomalía que aparece en grandes conjuntos de datos muy frecuentemente es la ausencia de valor en algunos de los campos de los registros de datos, fenómeno conocido en la literatura como valor ausente, en inglés missing value. La ausencia de valores en los datos se debe a diferentes causas, las más comunes son: ausencia de respuesta del cliente (por ejemplo, en una encuesta), fallas en la transcripción de datos, fallas en el soporte físico de los datos, mal funcionamiento de los sistemas de adquisición de datos, no aplicabilidad del valor del campo al registro de información, entre otras.

La limpieza de datos es un proceso de gran importancia cuando se quiere asegurar la calidad de los mismos. La limpieza de datos, también llamada lavado de datos (data scrubbing), trata de detectar y eliminar los errores e inconsistencias en los datos con el objetivo de mejorar su calidad.

El proceso de limpieza y construcción del repositorio se concibe alineado completamente con los procesos de minería de datos descritos a continuación:

- Selección: Tiene como objetivo la preparación de las fuentes de datos y la selección de las mismas.
- Preprocesamiento y transformación: Se aplican técnicas como limpieza de datos, la integración y transformación de datos, la reducción de datos y la selección de atributos.
- Aplicar las técnicas de minería: Es el proceso del cálculo de resúmenes y valores derivados. En esta etapa se perfeccionan constantemente las técnicas y algoritmos que se encargan de extraer y representar el conocimiento de forma adecuada para la toma de decisiones. Se combinan técnicas potenciando las ventajas de cada una y atenuando sus debilidades.
- Interpretación y evaluación: En este paso se procede al análisis de los resultados descubiertos. Incluye a su vez la resolución de posibles inconsistencias con otros conocimientos anteriores a la investigación.

2. Valores nulos y errores en los formatos de tipo dato

Reporte y documente los hallazgos de datos inconsistentes. Proceda a eliminarlos de la base (solo en caso que la inconsistencia de los datos afecte al interpretación de cada registro). Revise todas las columnas, pero comience y ponga especial atención en las siguientes que ya fueron analizadas en la práctica 1 (de hecho se sugiere utilice los hallazgos identificados de la práctica 1):

- `fecha_creacion`
- `año_cierre` y `hora_cierre` (todos los relacionados al cierre")
- `incidente_c4`
- `tipo_entrada`
- `clas_con_f_alarma`
- `delegación`

2.1. Responda a las siguientes preguntas

1. ¿Cuantos registros inconsistentes encontró?

   Al analizar los datos se encontraron 153 entradas inconsistentes, entre valores nulos y valores con punto decimal en la fecha.

2. ¿Cuántos registros después de la limpieza obtuvo como total en la muestra de datos?

   Después de limpiar los datos se obtuvieron 32,918 registros.

3. Análisis en Tableau

Realice el análisis correspondiente en Tableau, se recomienda usar el procedimiento de la clase "exploración básica de datos con Tableau". Documente el resultado a fin de responder a las siguientes preguntas de exploración de datos (realice las gráficas según corresponda):

- A ¿Cuál es la frecuencia de ocurrencia de cada incidente vial? ¿Cual es el más y el menos frecuente en la muestra de datos proporcionada?

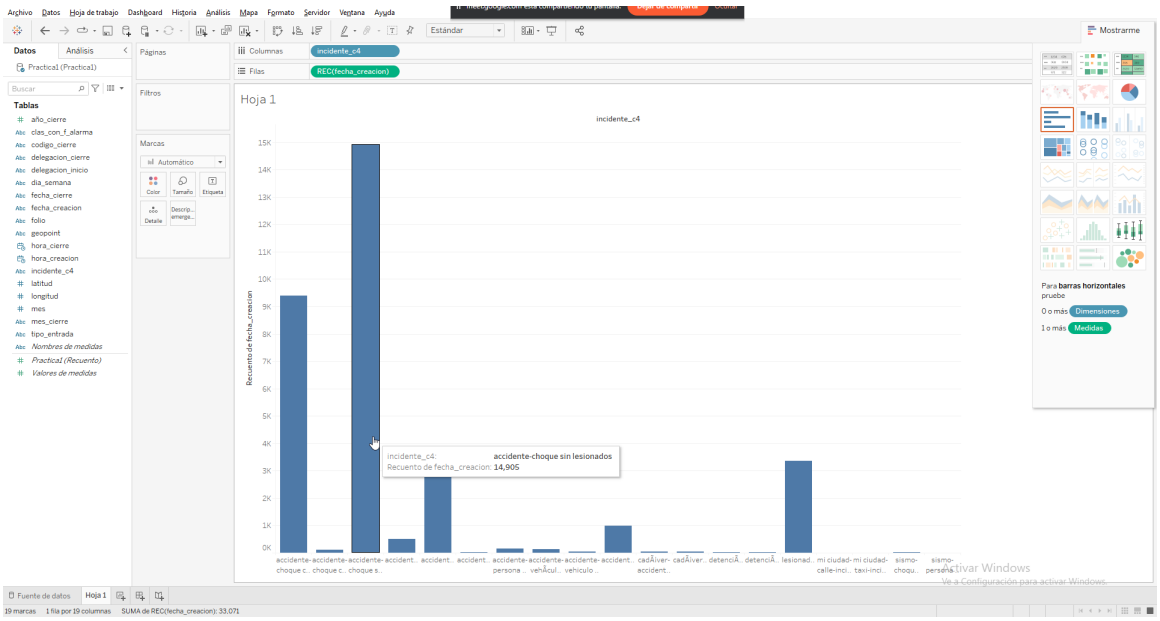

Figura 1: Vemos que el accidente que más ocurre es el choque sin lesionados con una frecuencia de 14905.

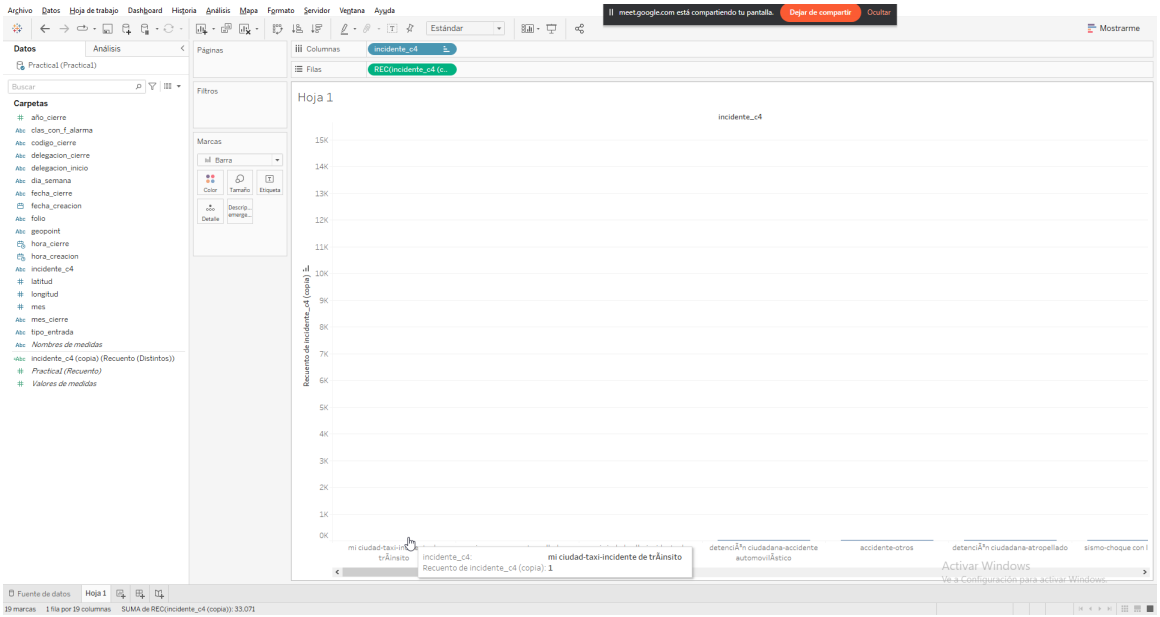

Figura 2: El menos frecuente es el accidente de taxi con 1 sola ocurrencia en todo 2020.

- B ¿Cuál es el `día_semana` con la mayor cantidad de incidentes viales?

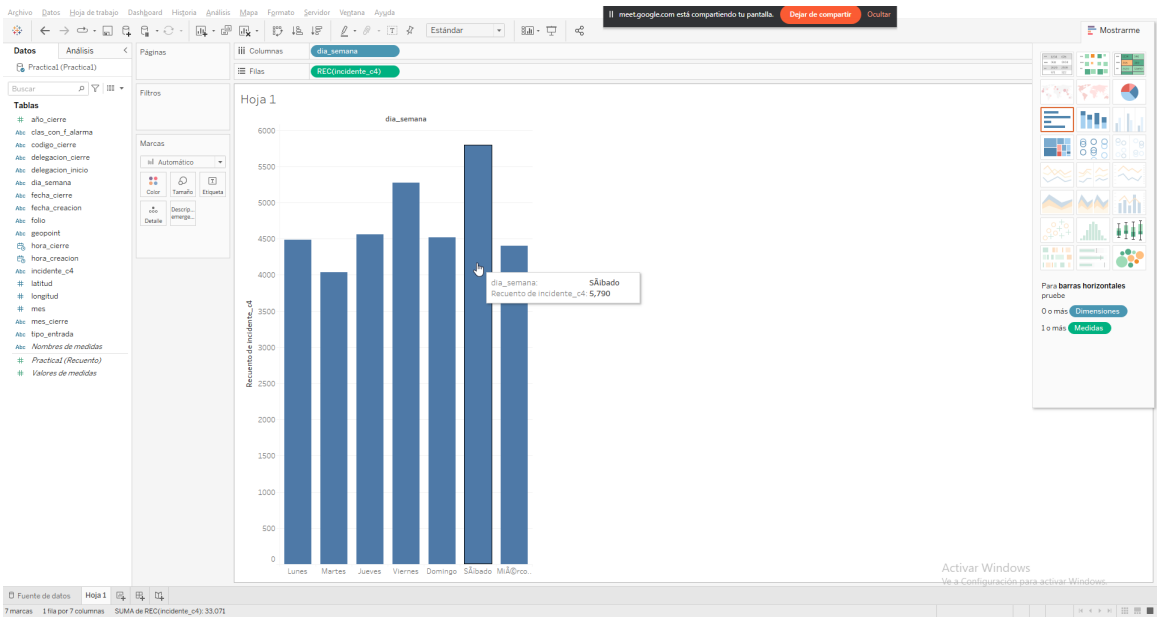

Figura 3: Vemos que el día con mayor cantidad de incidentes viales es el sábado.

- C ¿Cuál es el mes (`fecha_creacion`) con la mayor cantidad de incidentes viales?

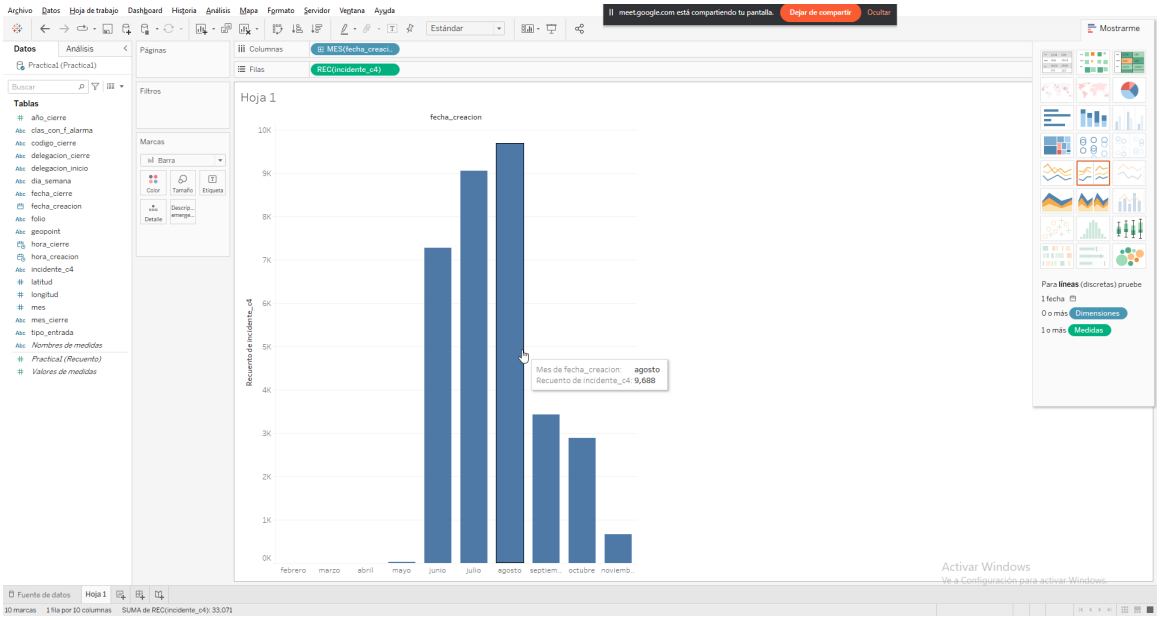

Figura 4: El mes con mayor antiad de incidentes viales es agosto.

- D ¿Cuál es la `hora_creacion` con la mayor cantidad de incidentes viales?

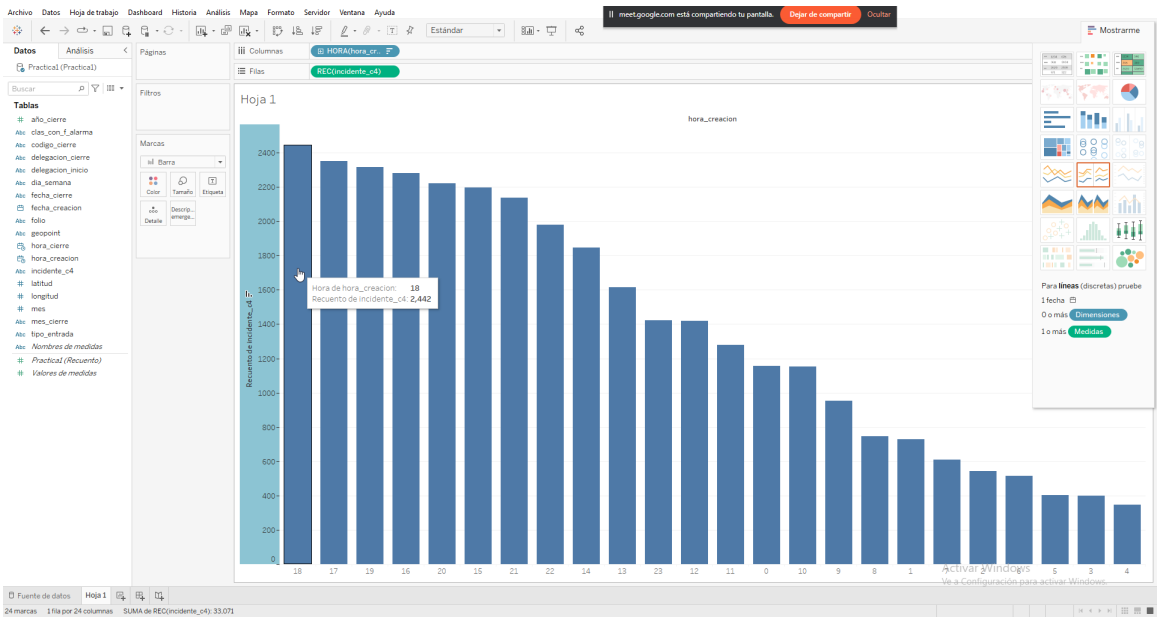

Figura 5: Vemos que la hora con mayor cantidad de incidentes viales es a las 6:00 PM.

- E ¿Cuál es la `delegación_inicio` con la mayor cantidad de incidentes viales?

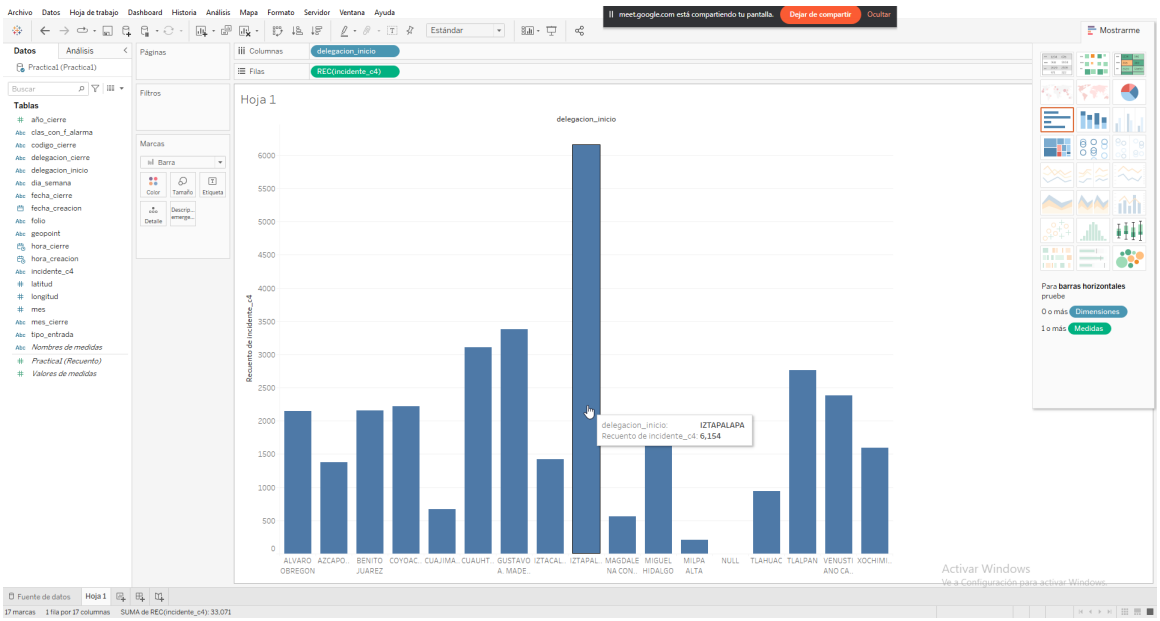

Figura 6: Iztapalapa es la delegación con la mayor cantidad de incidentes viales.

- F ¿Cuál es la `clas_con_f_alarma` con la mayor cantidad de incidentes viales?

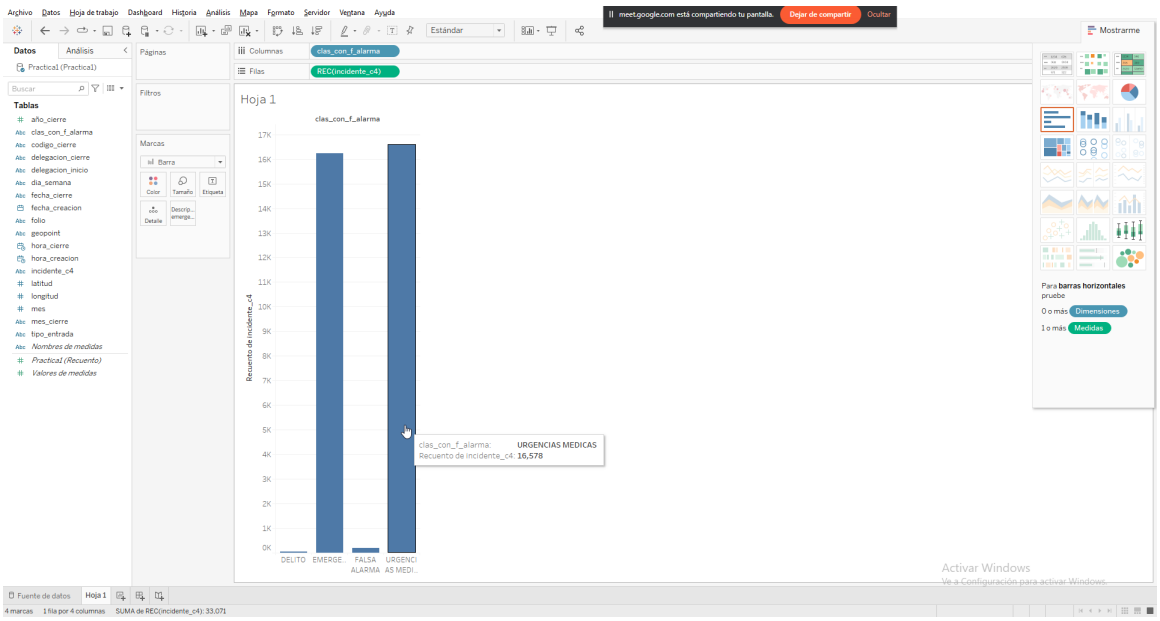

Figura 7: Las urgencias médicas cuenta con la mayor cantidad de incidentes viales.

- G ¿Cuál es el `tipo_entrada` con la mayor cantidad de incidentes viales?

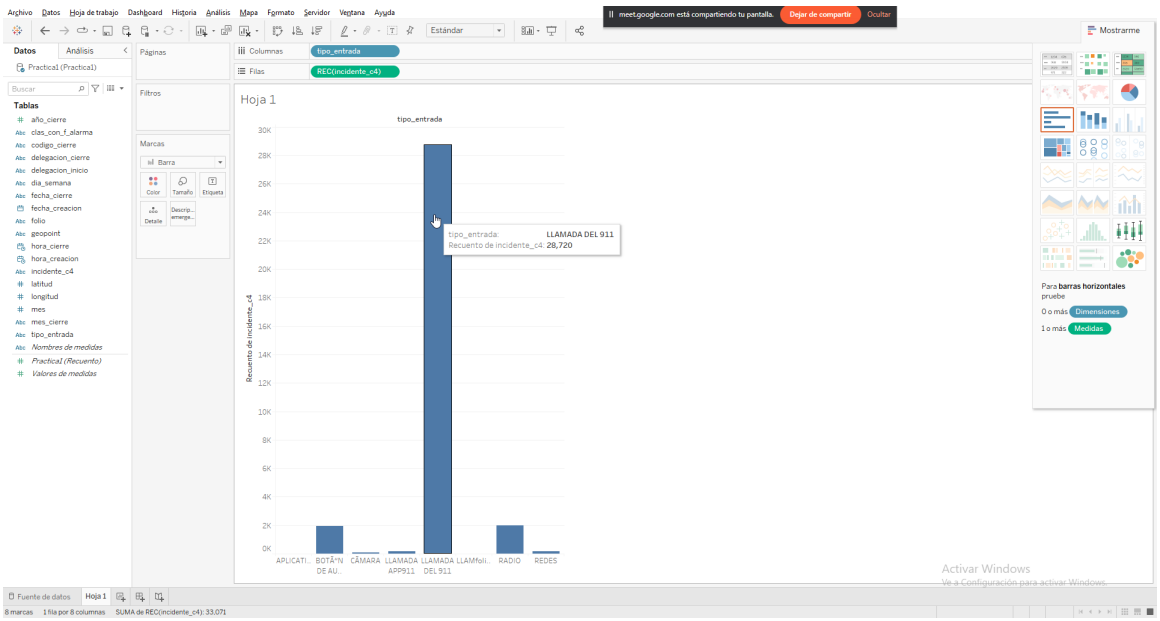

Figura 8: Llamada de 911 es la que tiene el tipo de entrada con la mayor cantidad de incidentes viales.

- H ¿Cuál es el `codigo_cierre` con la mayor cantidad de incidentes viales?

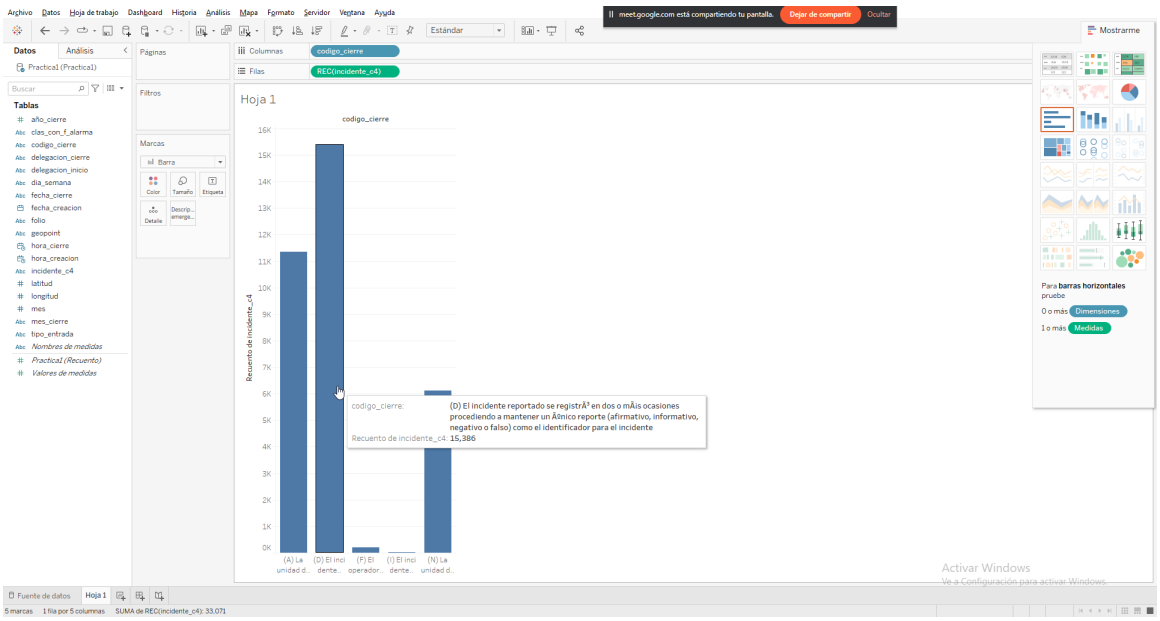

Figura 9: Vemos que aquellos incidentes que se registran una o más veces tienen un único reporte y es aquel el que tiene la mayor cantidad de incidentes viales

- I Considerando el incidente vial más y menos común, ¿cual es la frecuencia de ocurrencia de estos dos incidentes por `hora_cierre`?

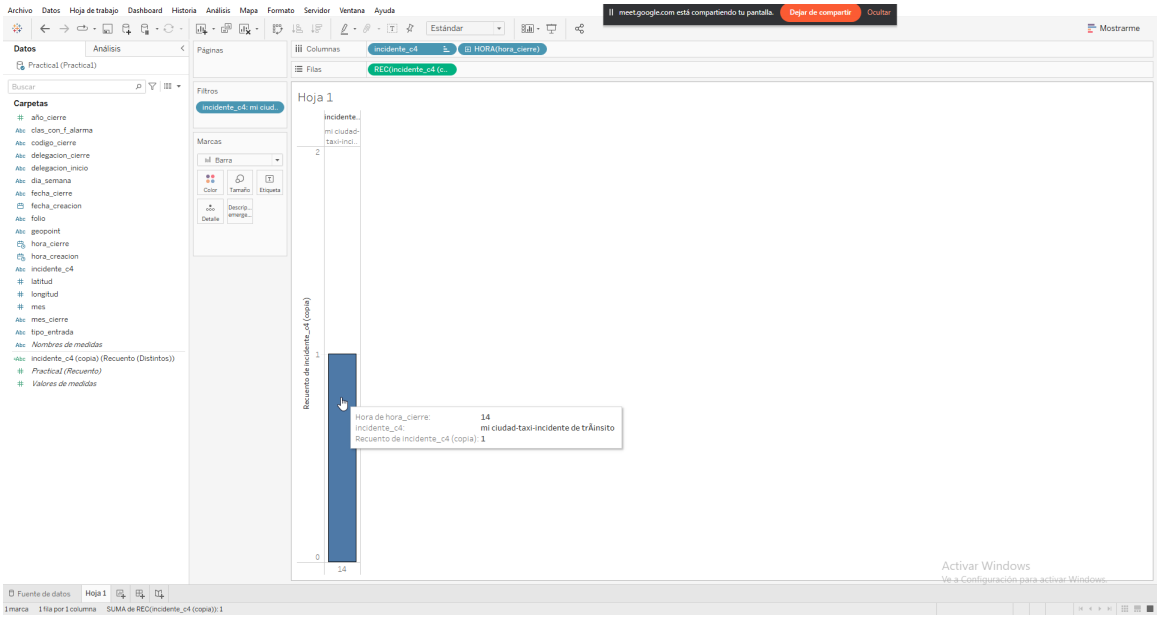

Figura 10: El menos común por hora de cierre es el incidente de taxi a las 2:00 PM.

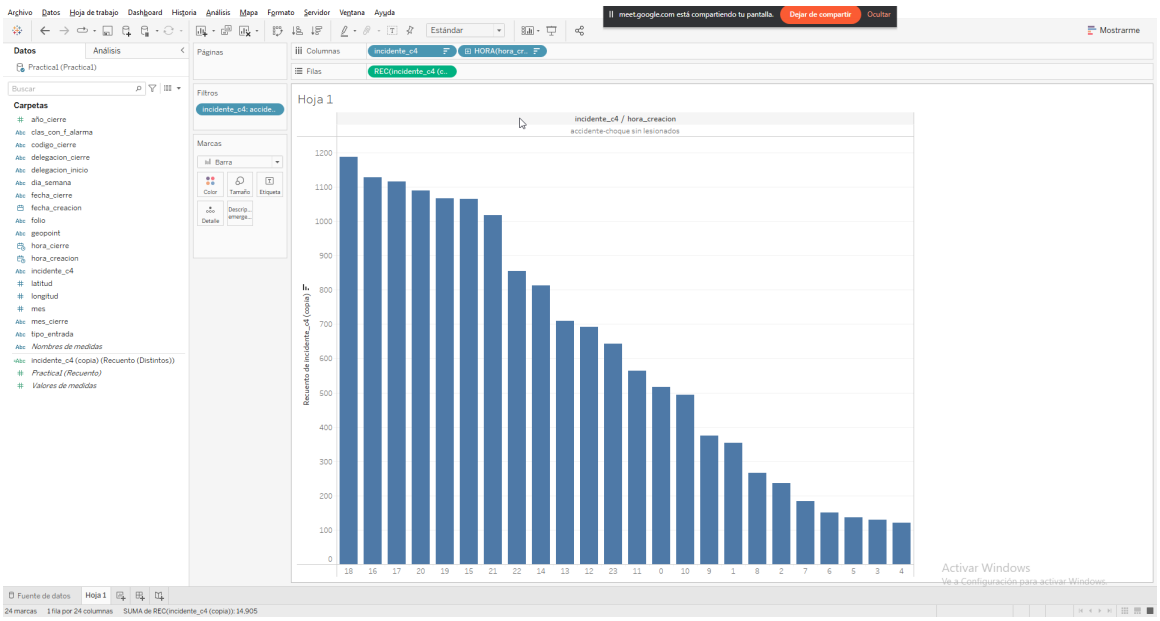

Figura 11: El más común se da a las 6:00 PM.

- J Considerando el incidente vial más frecuente, ¿cuál es la frecuencia de ocurrencia por delegación?

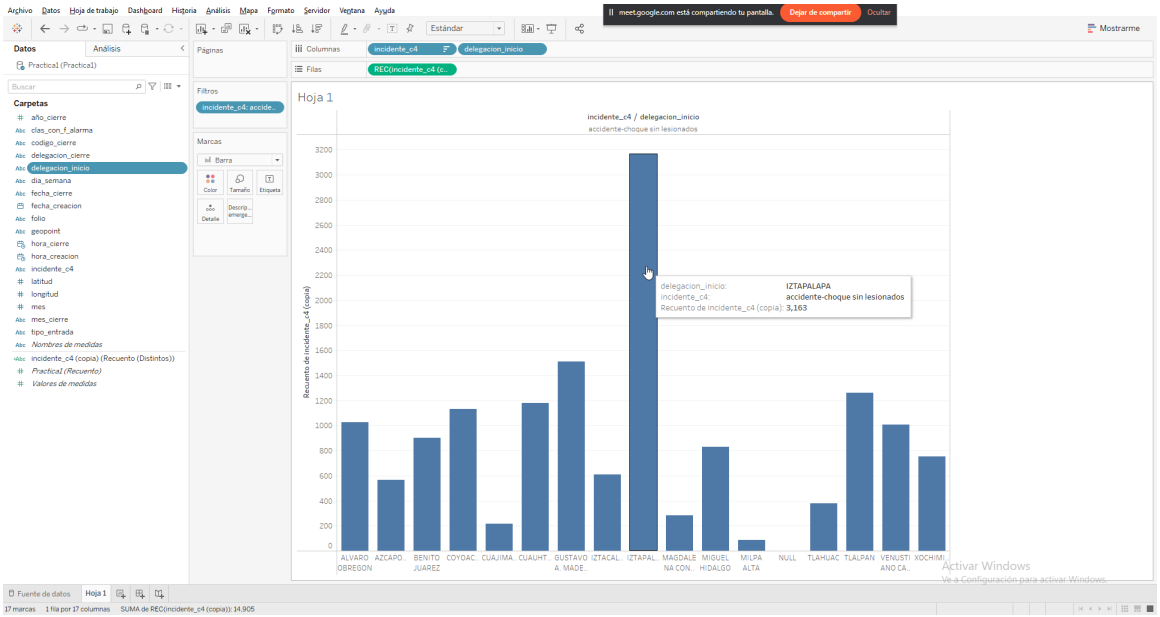

Figura 12: La que presenta mayor frecuencia de nuevo es la delegación Iztapalapa.

- K Considerando el incidente vial más frecuente, ¿cuál es la frecuencia de ocurrencia por `tipo_entrada`?

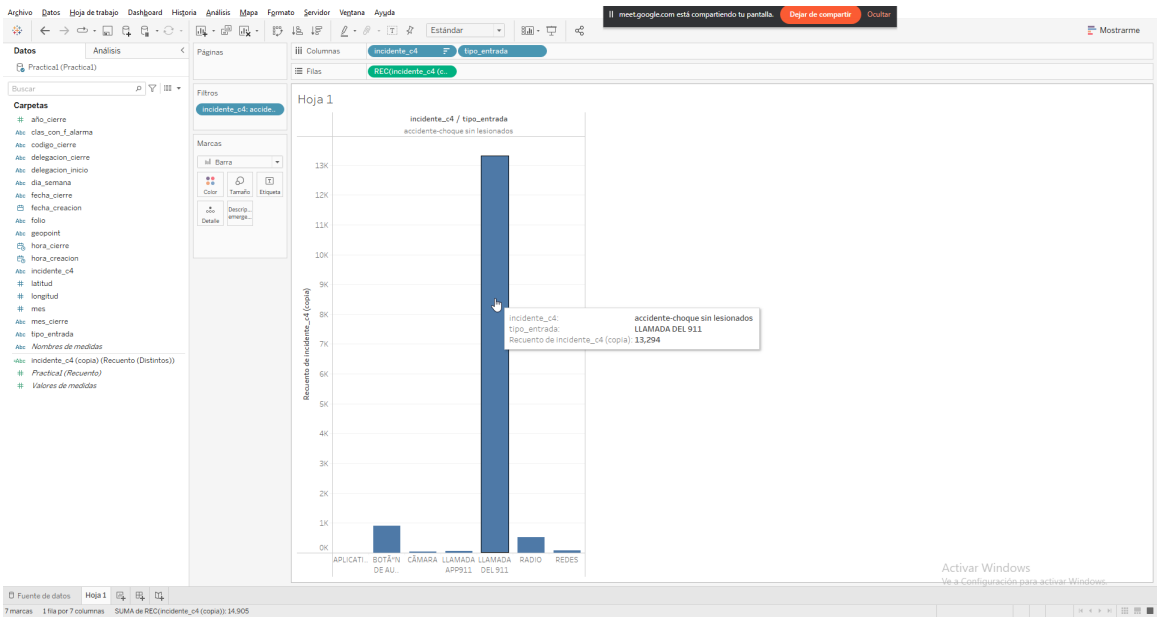

Figura 13: Accidente choque sin lesionados es aquel que tiene la mayor frecuencia de ocurrencia por tipo de entrada de llamada al 911.

4. Conclusiones

- **Graciano Herrera Gabriel**
  La limpieza de los datos es un proceso de suma importancia, aunque este llega a ser un poco tedioso, por la necesidad de encontrar todos aquellos datos inconsistentes, sin embargo, si no realizáramos este proceso nuestros posteriores análisis saldrían con ruido y provocaría que estos fueran deficientes. Así mismo la herramienta de Tableau muestra ser poderosa y eficiente para un primer análisis exploratorio de los datos dándonos la facilidad de interpretar estos datos e graficas de una manera muy sencilla.

- **Meza Zamora Abraham Manuel**
  En esta práctica comprendimos la importancia de tener nuestros datos congruentes y sin inconsistencias, ya que de lo contrario se producen errores y anomalías al momento de querer realizar un análisis. Este proceso puede ser tardado y tedioso, pero es indispensable si queremos realizar análisis precisos y coherentes.