Instituto Politécnico Nacional

Escuela Superior de Cómputo

Data Mining

Practica 2

Limpieza de datos y exploración básica.

Grupo: 3CV18

Integrantes.
Cazares Martínez Maximiliano
Ramos Nieves Adrián

Profesor:
Roberto Zagal

Introducción

¿Qué es la limpieza de datos?

Como su nombre lo indica es la depuración de datos erróneos en una tabla o base de datos. Esta acción permite identificar datos incorrectos, incompletos o poco relevantes para tu empresa o proyecto. Después de la limpieza, se sustituyen, modifican eliminan por completo los datos inservibles. Asegura que los datos con los que cuentas sean confiables y que la información que se obtenga de ellos era mucho más precisa y útil para tu empresa o proyecto.

¿Por qué es importante la limpieza de datos?

Los datos de calidad pueden variar dependiendo de cuál sea su cualidad, entre las principales se encuentran:

- Exactitud: todos los datos dentro de tu empresa deben ser precisos. Una forma de comprobar su exactitud es comparándolos con otras fuentes
- Coherencia: la coherencia de los datos te permite saber si la información de contacto que tienes de una persona u organización
- Validez: todos los datos deben cumplir con reglas o restricciones definidas.
- Uniformidad: es importante que todos los datos dentro de tus bases tengan los mismos valores o unidades.

Por lo tanto, el objetivo de esta práctica es 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. En el procedimiento que se llevara a cabo se realizara un análisis del data set de accidentes viales de la práctica 1, en el cual con el uso de la herramienta Tableau se realizaran las gráficas de cada una de las preguntas planteadas en el ejercicio y se limpiara la base de datos con los datos erróneos o inconsistencias.

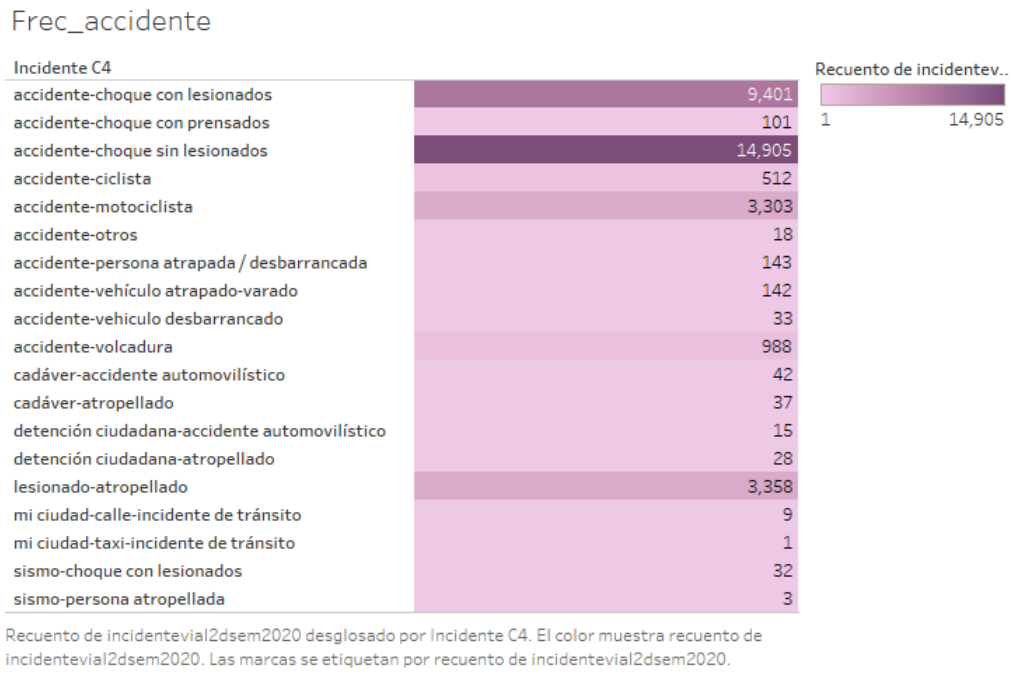
Desarrollo

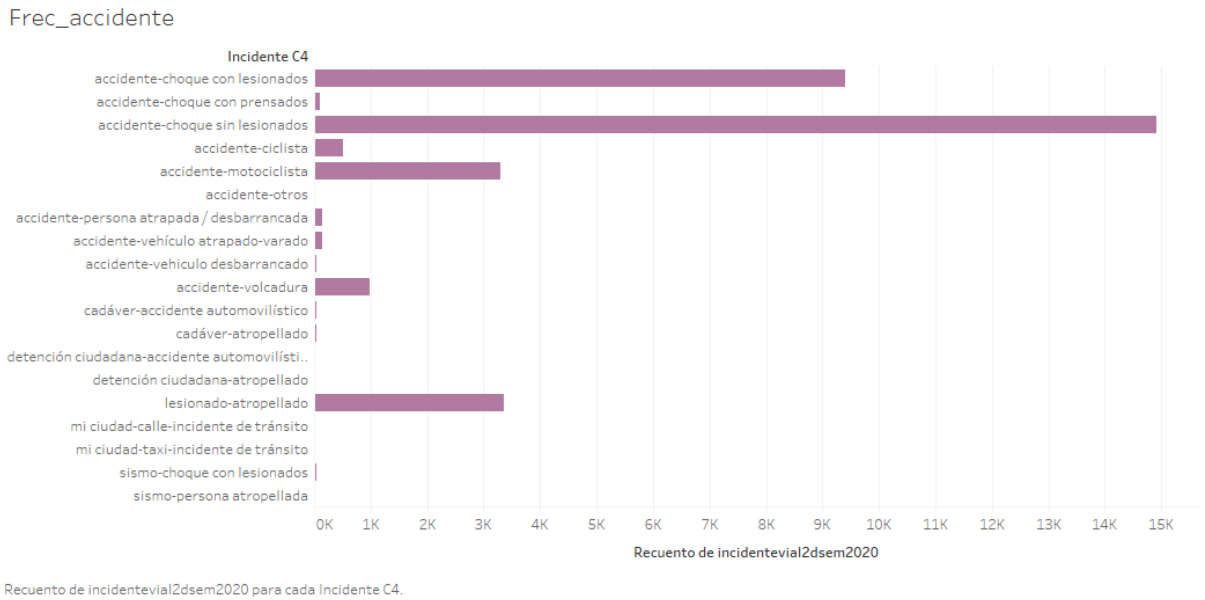
¿Cuántos registros inconsistentes encontró? ¿Cuántos registros después de la limpieza obtuvo como total en la muestra de datos?

Se realizaron las consultas necesarias para encontrar nulos e inconsistencias de la práctica de la cual se encontraron 4 nulos en las columnas de delegación, fecha y una extra que no estaba dentro de las principales. Datos inconsistentes no fueron encontrados, pero creemos importante resaltar que en algunas de las filas solo se tenía el dato de 1 registro como almacenado, se tendrá que ver si puede llegar a ser un caso aislado y si valdría la pena el tomarlo en cuenta como casos mínimos o máximos de algunas preguntas de la practica posteriores, ya que o bien podría ser una inconsistencia o solamente un caso aislado que tal vez se podría llegar a considerar eliminarlo para tener un análisis más preciso de casos más típicos.

Análisis del data set mediante graficas

A. ¿Cuál es la frecuencia de ocurrencia de cada incidente vial? ¿Cuál es el más y el menos frecuente en la muestra de datos proporcionada? Podemos observar que se tiene un índice mayor de personas con accidentes de choque sin lesionados, que viéndolo en retrospectiva es una cifra muy alta en comparación con los otros tipos de accidentes y aunque no existan lesionados en estos casos, es impresionante que la mayoría de los casos se encuentren tipificados como choques, se tendría que hacer un análisis de los lugares para sacar conclusiones de que medidas preventivas se podría llegar a aplicar.





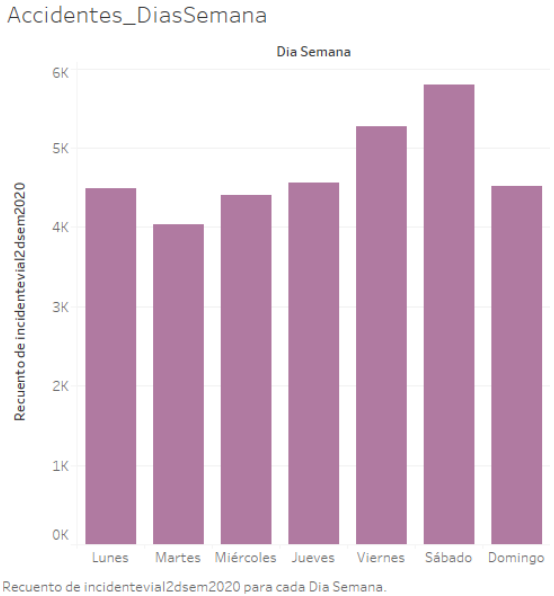
B. ¿Cuál es el **día_semana** con la mayor cantidad de incidentes viales?

En esta grafica podemos observar que el día con mayor frecuencia de incidentes es el sábado, algunas suposiciones o casos son probables que sucedan, son por alcoholismo o descuidos de conductores, ya que al ser fines de semana es más comprensible que las personas quieran salir a divertirse y pueda llegar a causar accidentes

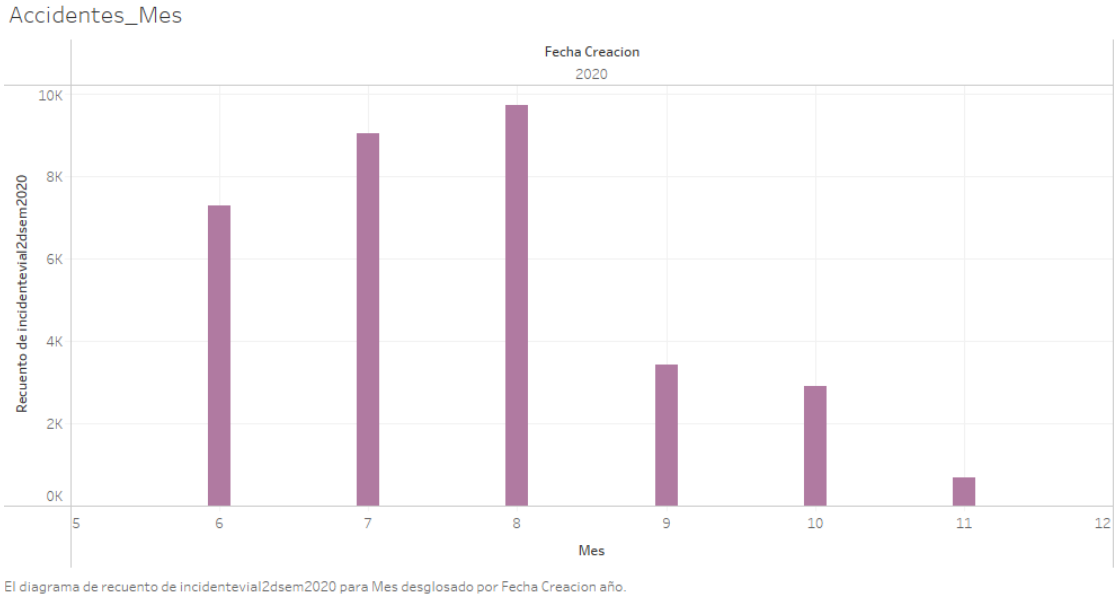
Accidente-s_DiasSema-na

Dia Semana	Recuento de incidentev..
Lunes	4,485
Martes	4,039
Miércoles	4,403
Jueves	4,558
Viernes	5,275
Sábado	5,789
Domingo	4,522

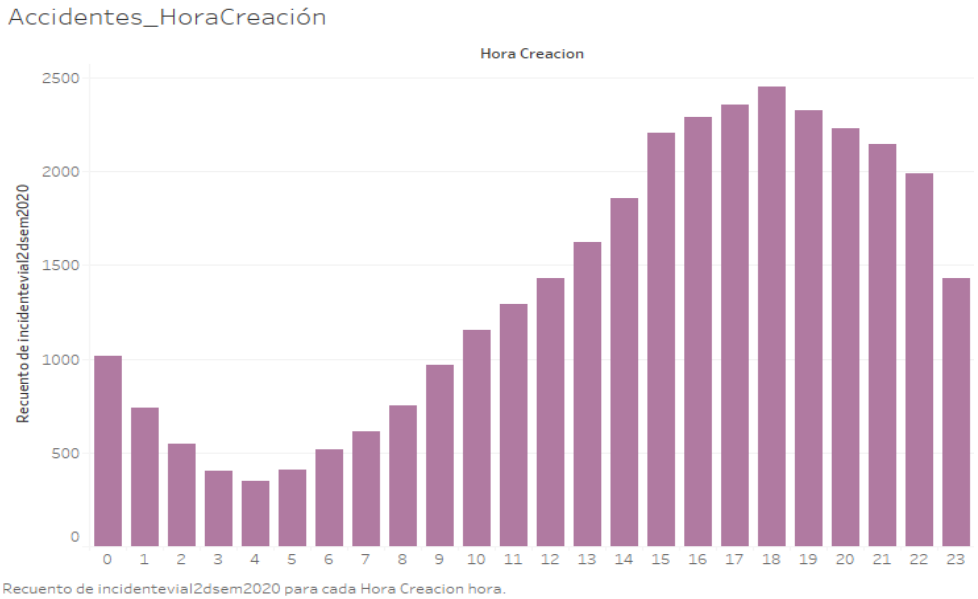
4,039 5,789



C. ¿Cuál es el mes (**fecha_creacion)** con la mayor cantidad de incidentes viales?
Se encontró con Tableau que el mes con más accidentes registrados es agosto, podríamos decir que en estos meses generalmente las personas que estudian tienen vacaciones, por lo tanto, es el tiempo perfecto en que muchos quieren salir de viaje o pasar sus vacaciones fuera de casa, lo que implica un mayor riesgo para accidentes viales.

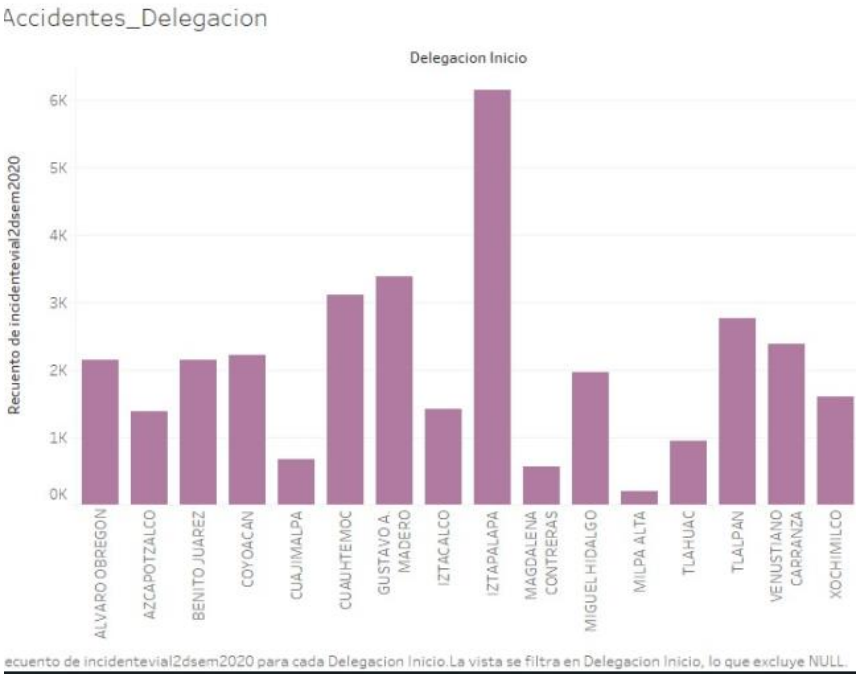


D. ¿Cuál es la **hora_creacion** con la mayor cantidad de incidentes viales?
La hora con más accidentes registradas en el Data set, es a las 6 de la tarde, planteando una hipótesis consideramos que es debido a la hora pico en donde muchas personas salen de trabajar y se dirigen a casa, es el momento del día al igual que la mañana que se tiene mucho disturbio vial.



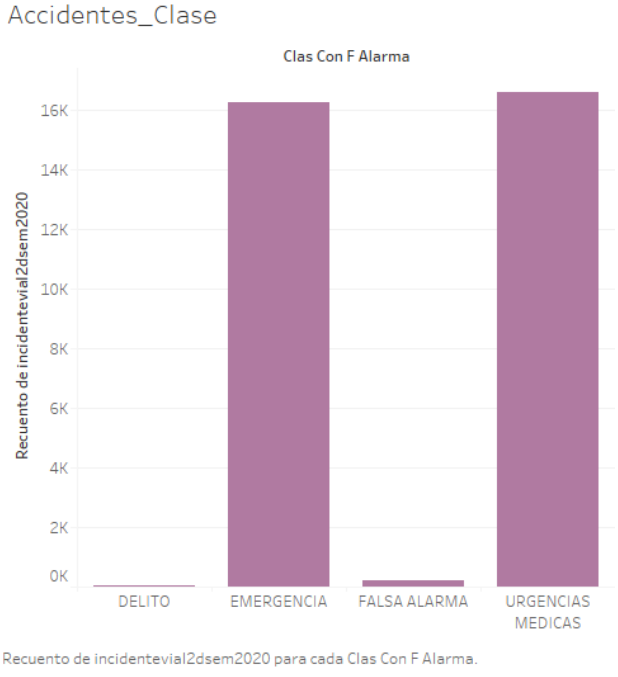
E. ¿Cuál es la **delegación_inicio** con la mayor cantidad de incidentes viales?

La delegación con mayor cantidad de incidentes viales es Iztapalapa, se tendría que plantear si esto tiene alguna correlación con las altas tasas de delincuencia en la Alcaldía o solo existen más razones por las cuales se presente un incremento considerable a comparación con otras delegaciones.



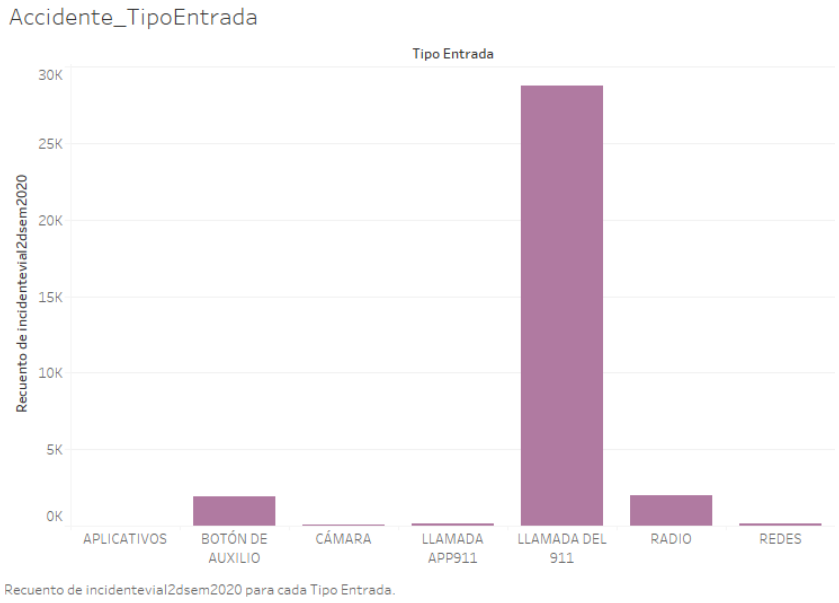
F. ¿Cuál es la **clas_con_f_alarma** con la mayor cantidad de incidentes viales?

La clase de Alarma que se utiliza más en accidentes viales es por urgencias y emergencias, esto podría ir a la mano con el tipo de accidente C4 que representa la mayor cantidad de registros con las que cuenta la base de datos.



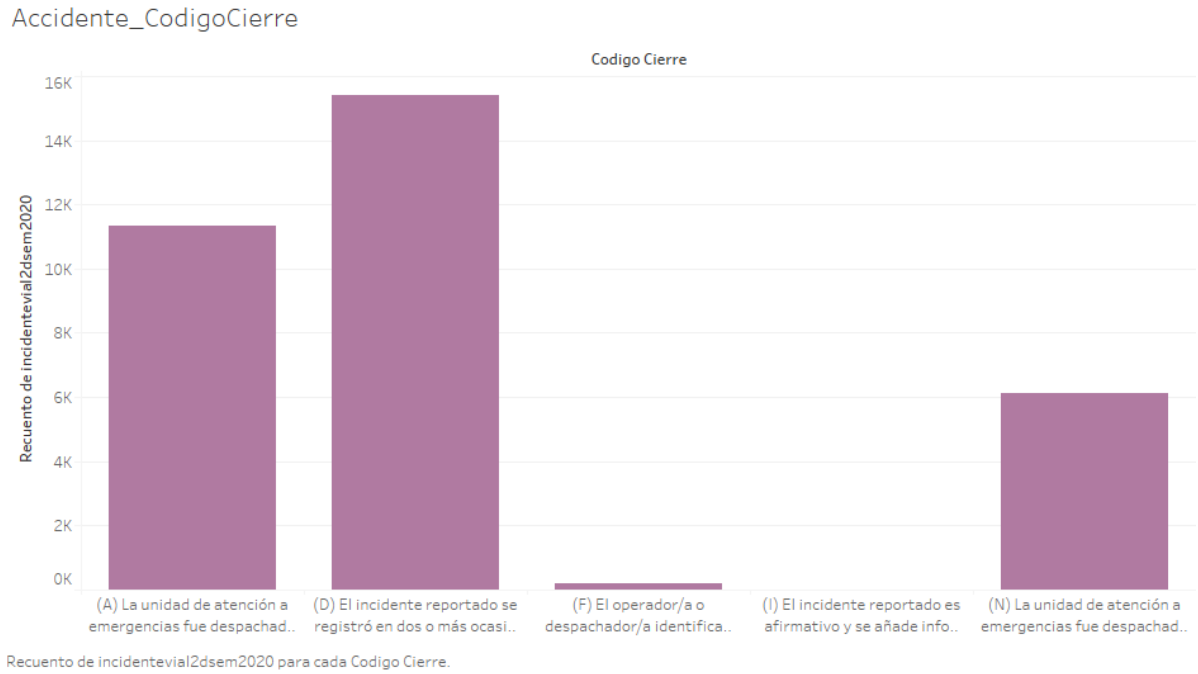
G. ¿Cuál es el **tipo_entrada** con la mayor cantidad de incidentes viales?

El tipo de entrada más común para este tipo de accidentes observamos que comprende un marco de llamadas al 911, que es la columna con más registros a comparación de las otras opciones presentes.

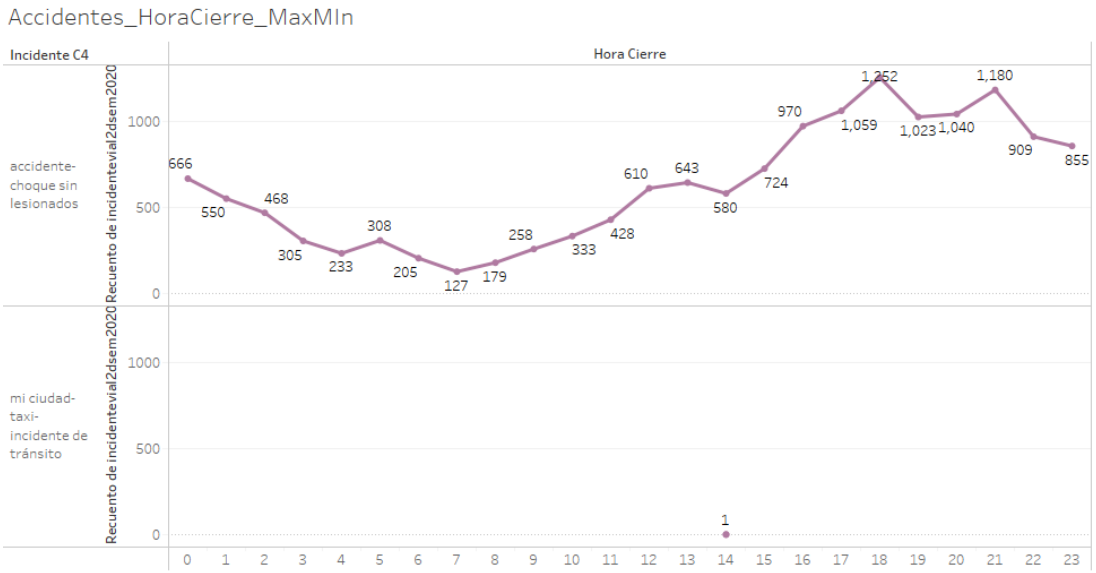


H. ¿Cuál es el **codigo_cierre** con la mayor cantidad de incidentes viales?

El incidente reportado se registró en más de dos ocasiones es la respuesta más frecuente en los registros del Data set.



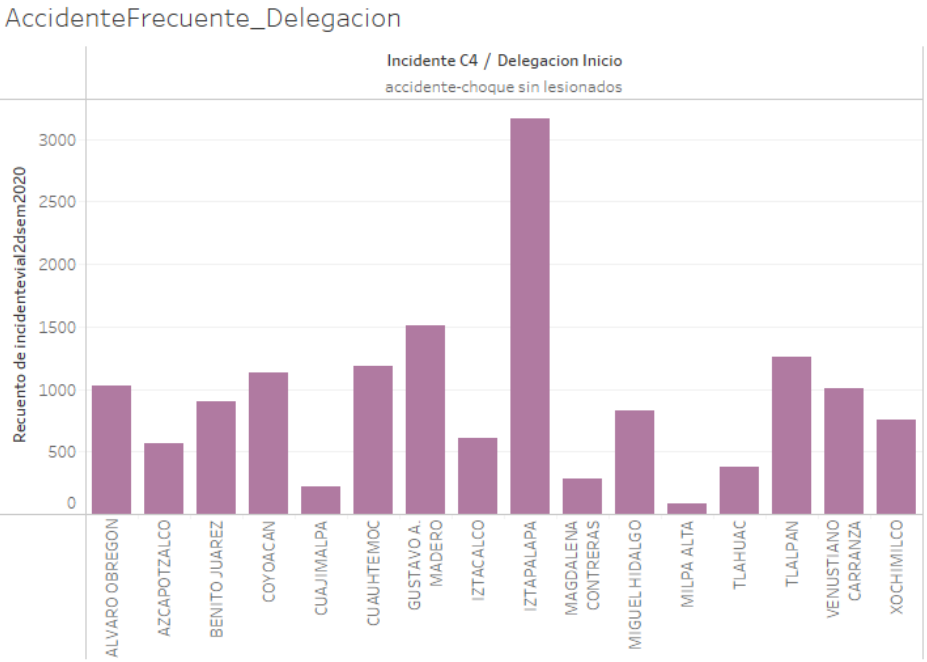
I. Considerando el incidente **vial más y menos común**, ¿cuál es la frecuencia de ocurrencia de estos dos incidentes por **hora_cierre**?



La tendencia de recuento de incidentevial2dsem2020 para Hora Cierre hora desglosada por Incidente C4. Las marcas se etiquetan por recuento de incidentevial2dsem2020.La vista se filtra en Incidente C4, lo que conserva accidente-choque sin lesionadosymi ciudad-taxi-incidente de tránsito.

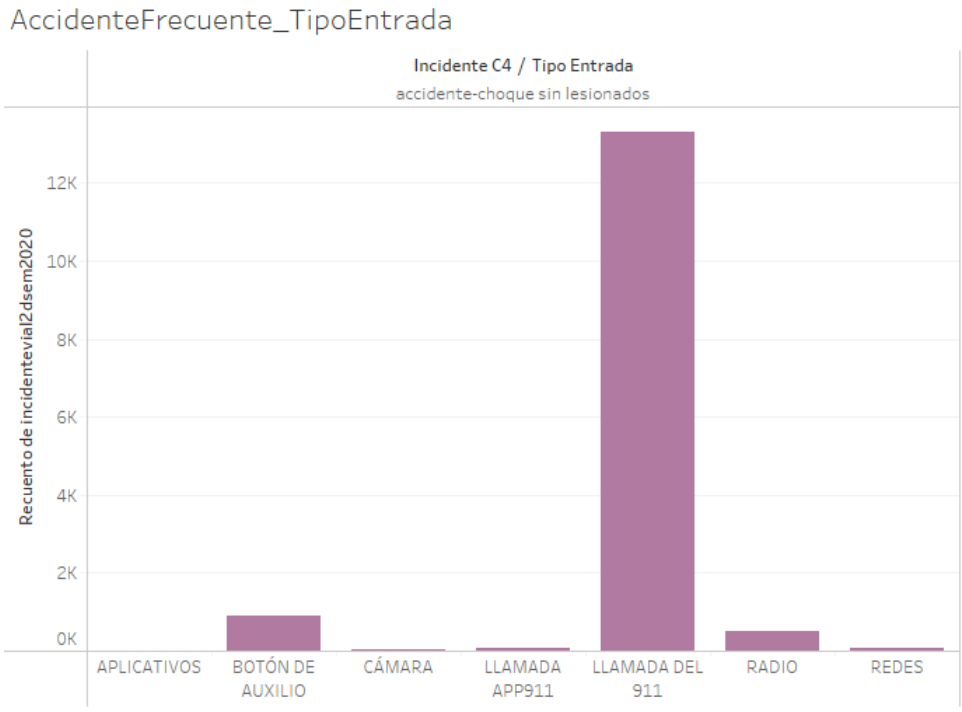
J. Considerando el incidente vial **más frecuente**, ¿cuál es la frecuencia de ocurrencia por **delegación**?

Se puede observar que la primera alcaldía con los mayores registros de incidentes viales se encuentra en Iztapalapa, como se comentó anteriormente son choques sin lesionados, pero se tendría que ver que zonas son las afectadas dentro de Iztapalapa, ya que sus registros son muy superiores en comparación con otras alcaldías.



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

Podemos analizar que sigue siendo el 911 el número uno en el tipo de entrada de reporte del accidente en el caso específico choque sin lesionados.



Recuento de incidentevial2dsem2020 para cada Tipo Entrada desglosado por Incidente C4.La vista se filtra en Incidente C4, lo que conserva accidente-choque sin lesionados.

Conclusiones

En esta práctica se tomó en cuenta las clases de limpieza de datos, que con el material visto en clase se comprendió la importancia del tema para los proyectos y la empresa, ya que se puede comprender como podemos quitar valores inconsistentes, así como nulos de nuestra Data set. Posteriormente con la parte dos de la práctica analizamos los datos con la herramienta de trabajo Tableu, la cual nos ayuda a manejar los datos para tener una mejor comprensión del tema, como por ejemplo los lugares con más accidentes viales, y cuál es el más común de todos ellos. Es importante tener una noción de lo que vamos a describir porque así podemos tener una mejor comprensión de los temas, como es la tendencia de las cosas, o planes a llevar a cabo para contrarrestar las diferentes problemáticas.