Instituto Politécnico Nacional
Escuela Superior de Cómputo

Minería de Datos

Vianey Maravilla Pérez

5AM1

Roberto Zagal

DATA IMPORT

Práctica 2: Limpieza de datos y exploración básica

Introducción:

En esta práctica se realizó una limpieza y exploración básica de nuestros datos, poniendo a prueba nuestras capacidades dentro del software de “Tableau”.

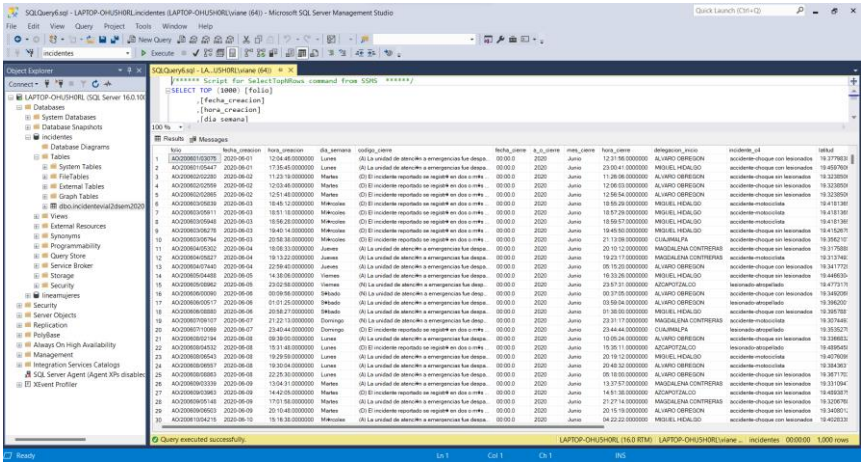
Dentro de este software podremos analizar y tener una visualización más gráfica de nuestros datos y consultas.

Tableau es un software con una visualización de datos interactivos para la inteligencia empresarial, siendo así, cada que queramos interpretar se podrá hacer de una manera más rápida y precisa.

Objetivo: 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.

Desarrollos:

1.- Utilice el dataset de incidentes viales de la práctica 1

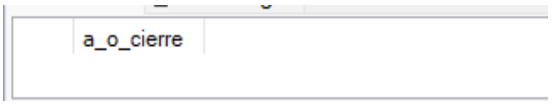


2.- Identifique valores NULOS y errores en los formatos de tipo de datos, reporte y documente los hallazgos de datos consistentes. Proceda a eliminarlos de la base (solo en caso que las inconsistencias de los datos afecten a la interpretación de cada registro). Revise todas las columas, pero comience y ponga especial atención en las siguientes que ya fueron analizadas en la practica 1 (de hecho, se sugiere utilice los hallazgos identificados de la práctica 1).

Al analizar todas las columnas, una por una y hacer un par de consultas nos dimos cuenta de lo siguiente:

La hora de inicio y cierre son los que más valores nulos nos arroja, se hicieron las siguientes consultas de todas las columnas para poder hallar con más precisión los valores nulos de toda la base de datos:

```
use incidentes
-- Año cierre valores nulos
select incidentes.a_o_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.a_o_cierre Is NULL ;
```



```sql
-- Clas con f alarma valores nulos
select incidentes.clas_con_f_alarma
from incidentevial2dsem2020 as incidentes
where incidentes.clas_con_f_alarma Is NULL;
```

clas_con_f_alarma

```sql
-- Código de cierre valores nulos
select incidentes.codigo_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.codigo_cierre Is NULL;
```

codigo_cierre

```sql
-- Delegación cierre valores nulos
select incidentes.delegacion_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.delegacion_cierre Is NULL;
```

	delegacion_cierre
1	NULL

```sql
-- Delegación inicio valores nulos
select incidentes.delegacion_inicio
from incidentevial2dsem2020 as incidentes
where incidentes.delegacion_inicio Is NULL;
```

delegacion_inicio

```sql
-- Día semana valores nulos
select incidentes.dia_semana
from incidentevial2dsem2020 as incidentes
where incidentes.dia_semana Is NULL;
```

dia_semana

```sql
-- Fecha cierre valores nulos
select incidentes.fecha_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.fecha_cierre Is NULL;
```

fecha_cierre

```sql
-- Fecha creación valores nulos
select incidentes.fecha_creacion
from incidentevial2dsem2020 as incidentes
where incidentes.fecha_creacion Is NULL;
```

fecha_creacion

```sql
-- Folio valores nulos
select incidentes.folio
from incidentevial2dsem2020 as incidentes
where incidentes.folio Is NULL;
```

folio

```sql
-- Geopoint valors nulos
select incidentes.geopoint
from incidentevial2dsem2020 as incidentes
where incidentes.geopoint Is NULL;
```

	geopoint
1	NULL

```sql
-- Hora cierre valores nulos
select incidentes.hora_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.hora_cierre Is NULL;
```

	hora_cierre
1	NULL
2	NULL
3	NULL
4	NULL
5	NULL

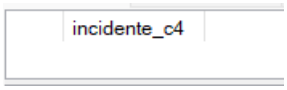
151 rows

```sql
-- Hora creación valores nulos
select incidentes.hora_creacion
from incidentevial2dsem2020 as incidentes
where incidentes.hora_creacion Is NULL;
```

	hora_creacion
1	NULL
2	NULL
3	NULL
4	NULL
5	NULL

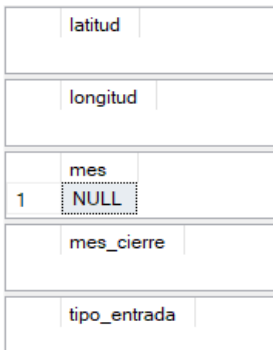
151 rows

```sql
-- Incidente c4 valores nulos
select incidentes.incidente_c4
from incidentevial2dsem2020 as incidentes
where incidentes.incidente_c4 Is NULL;
```



	incidente_c4

```sql
-- Latitud valores nulos
select incidentes.latitud
from incidentevial2dsem2020 as incidentes
where incidentes.latitud Is NULL;

-- Longitud valores nulos
select incidentes.longitud
from incidentevial2dsem2020 as incidentes
where incidentes.longitud Is NULL;

-- Mes valores nulos
select incidentes.mes
from incidentevial2dsem2020 as incidentes
where incidentes.mes Is NULL;
```



	latitud

	longitud

	mes
1	NULL

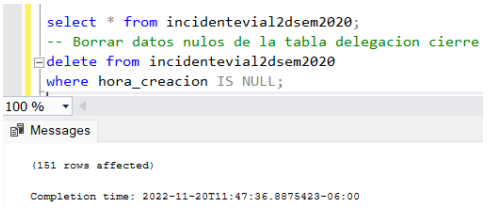
	mes_cierre

	tipo_entrada

```sql
-- Mes cierre valores nulos
select incidentes.mes_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.mes_cierre Is NULL;

-- Tipo entrada valores nulos
select incidentes.tipo_entrada
from incidentevial2dsem2020 as incidentes
where incidentes.tipo_entrada Is NULL;
```

LIMPIEZA DE DATOS NULOS

```sql
-- Borrar datos nulos de la tabla hora cierre
delete from incidentevial2dsem2020
where hora_creacion IS NULL;
```



```
select * from incidentevial2dsem2020;
-- Borrar datos nulos de la tabla delegacion cierre
delete from incidentevial2dsem2020
where hora_creacion IS NULL;
```

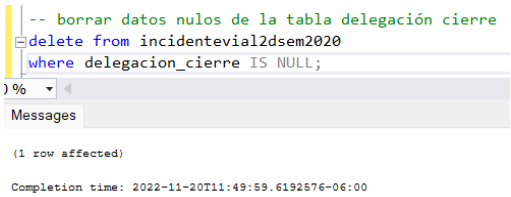
100 %

Messages

(151 rows affected)

Completion time: 2022-11-20T11:47:36.8875423-06:00

```sql
-- borrar datos nulos de la tabla delegación cierre
delete from incidentevial2dsem2020
where delegacion_cierre IS NULL;
```



```
-- borrar datos nulos de la tabla delegación cierre
delete from incidentevial2dsem2020
where delegacion_cierre IS NULL;
```

) %

Messages

(1 row affected)

Completion time: 2022-11-20T11:49:59.6192576-06:00

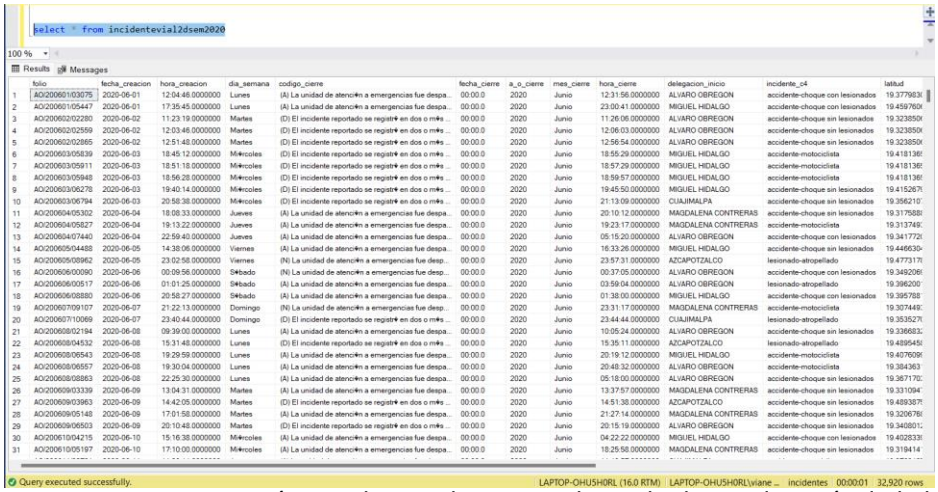
2.1.- Responda a las siguientes preguntas:

a) ¿Cuántos registros inconsistentes encontró?

Entre los valores de hora de cierre y hora de creación se encontraron 151 renglones con valores nulos cada uno, es decir, no existe una hora registrada en 151 casos de la base de datos.

Por otro lado, en los campos de “delegación cierre”, “geopoint”, y “mes”, se encontró solo un registro con valor nulo.

b) ¿Cuántos registros después de la limpieza obtuvo como total en la muestra de datos?



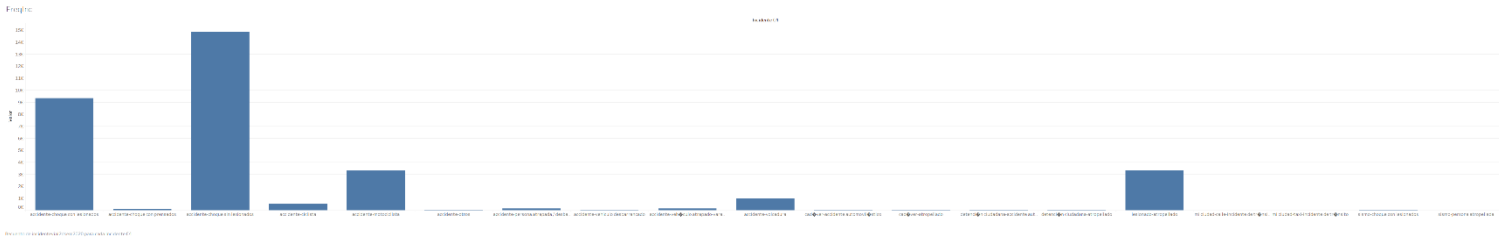
De los 33072 registros que teníamos dentro de nuestra base de datos, después de la limpieza como vemos nos quedaron 32920 datos, de los cuales ya podemos ver que no hay datos nulos y podremos hacer una mejor consulta dentro de Tableau

Al hacer de nuevo nuestras consultas nos encontramos con lo siguiente

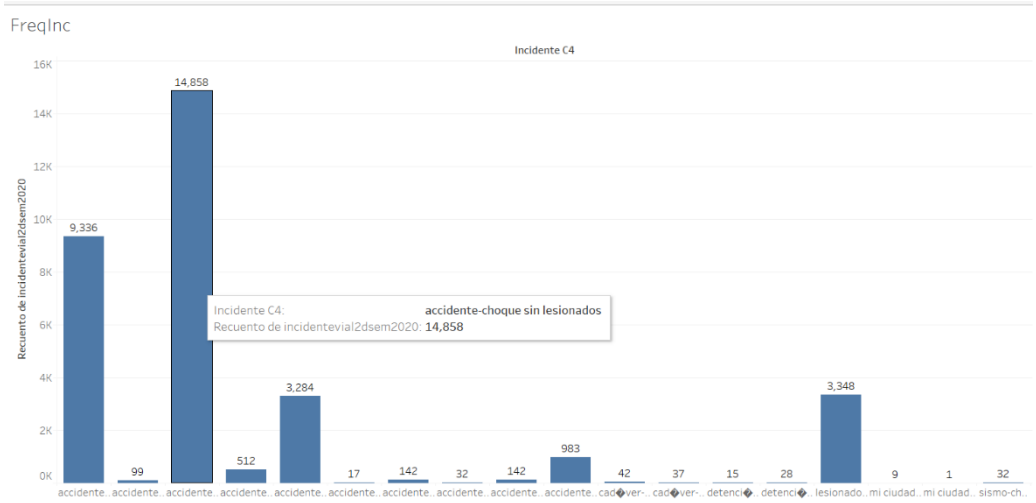
```
-- Código de cierre valores nulos
select incidentes.codigo_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.codigo_cierre Is NULL;

-- Delegación cierre valores nulos
select incidentes.delegacion_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.delegacion_cierre Is NULL;

-- Delegación inicio valores nulos
select incidentes.delegacion_inicio
from incidentevial2dsem2020 as incidentes
where incidentes.delegacion_inicio Is NULL;

-- Día semana valores nulos
select incidentes.dia_semana
from incidentevial2dsem2020 as incidentes
where incidentes.dia_semana Is NULL;

-- Fecha cierre valores nulos
select incidentes.fecha_cierre
from incidentevial2dsem2020 as incidentes
where incidentes.fecha_cierre Is NULL;
```

3- 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 responde a las siguientes preguntas de exploración de datos (realice las gráficas según corresponda):

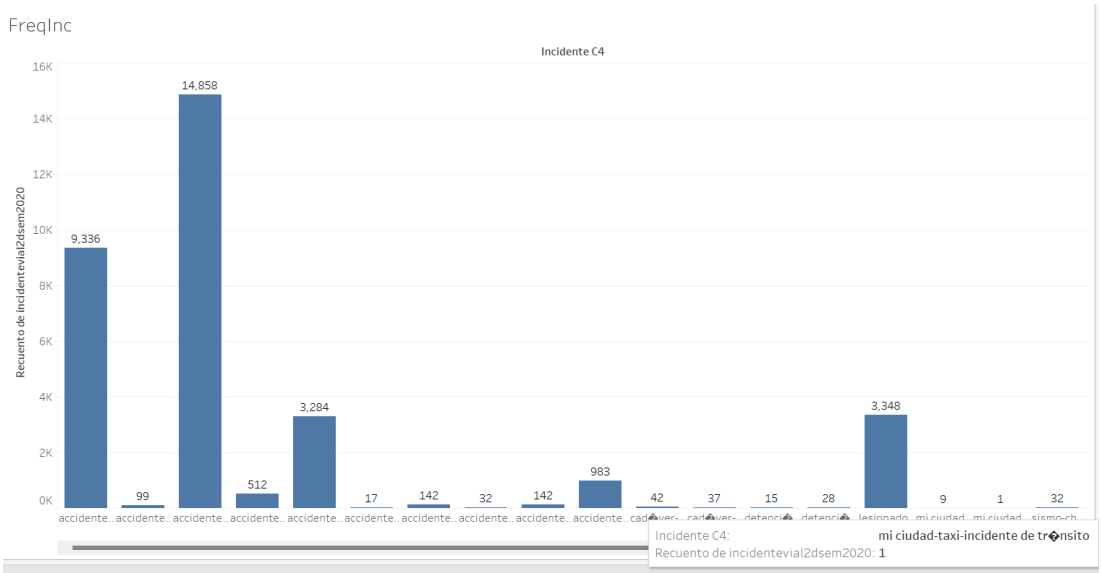
a) ¿Cuál es la frecuencia de ocurrencia de cada incidente vial? ¿Cuál es el mas y el menos frecuente en la muestra de datos proporcionada?



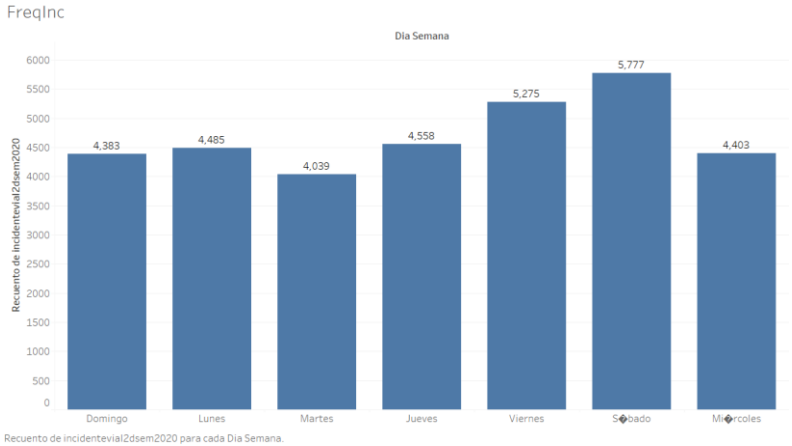
Como vemos el accidente más frecuente o el que más cifras tiene es el choque sin lesionados con 14, 858 accidentes



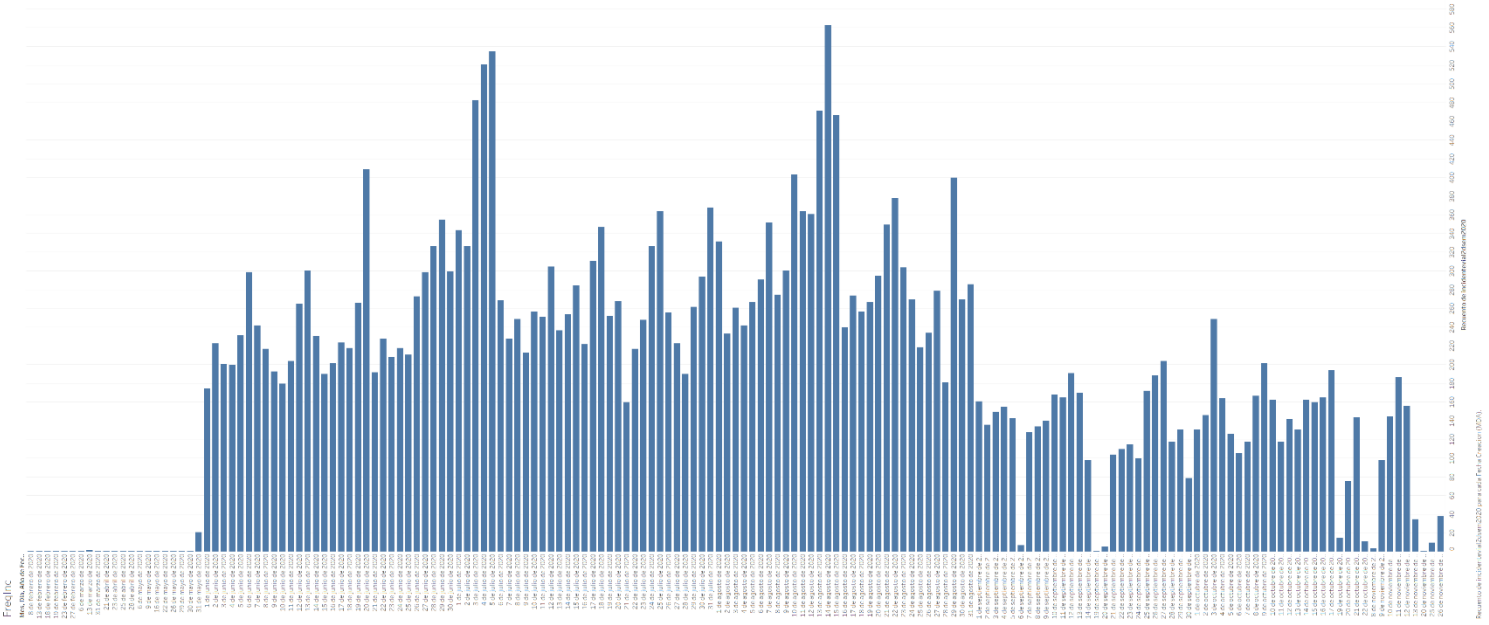
Los accidentes con menor frecuencia son los accidentes de transito dentro de un taxi con un solo accidente.



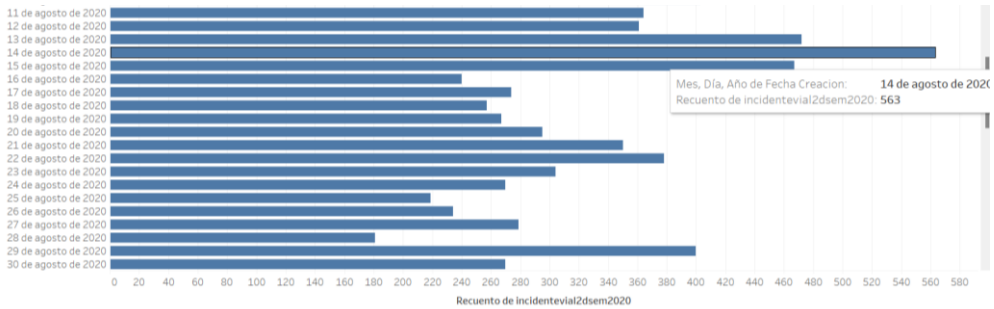
b) ¿Cuál es el dia_semana con la mayor cantidad de incidentes viales?
El día que más tiene o cuando más ocurren incidentes viales son los sábados con registros de 5,777 accidentes



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

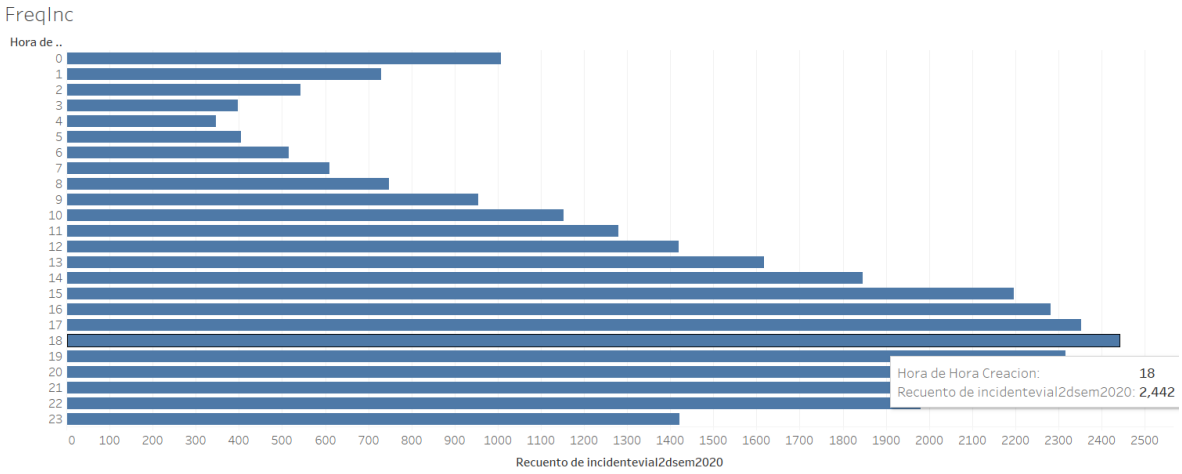


Si nos damos cuenta la mayor cantidad de accidentes que tiene un día es el 14 de agosto del 2020 con 563 accidentes durante ese día:



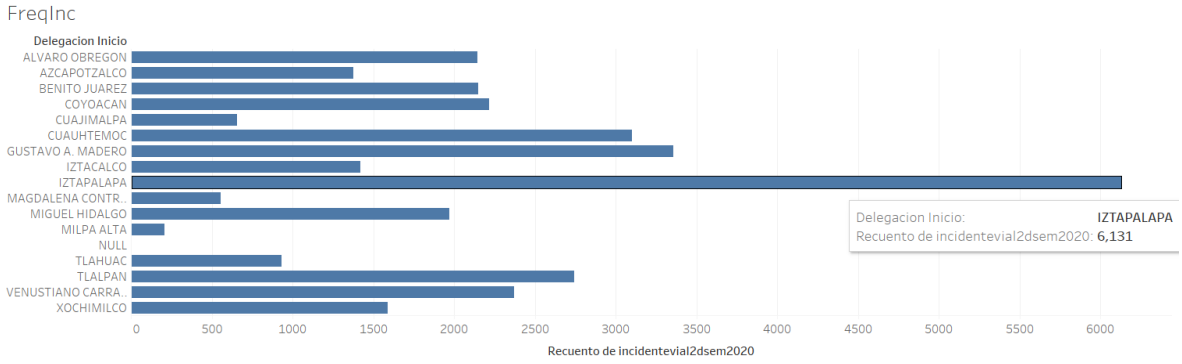
d) ¿Cuál es la hora_creacion con la mayor cantidad de incidentes viales?

La hora con más accidentes viales son las 6 de la tarde con 2,442 registros como vemos a continuación



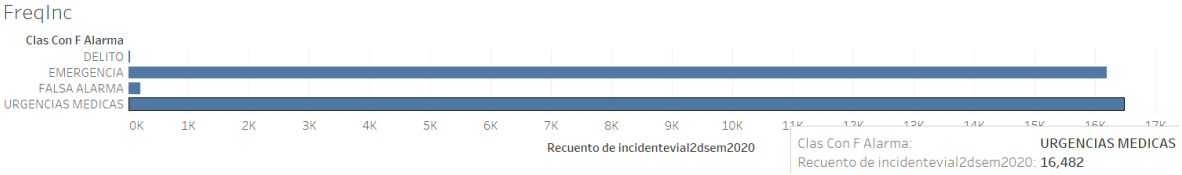
e) ¿Cuál es la delegación_inicio con la mayor cantidad de incidentes viales?

La alcaldía / delegación con mayor cantidad de incidentes viales es Iztapalapa con al menos 6,131 registros.



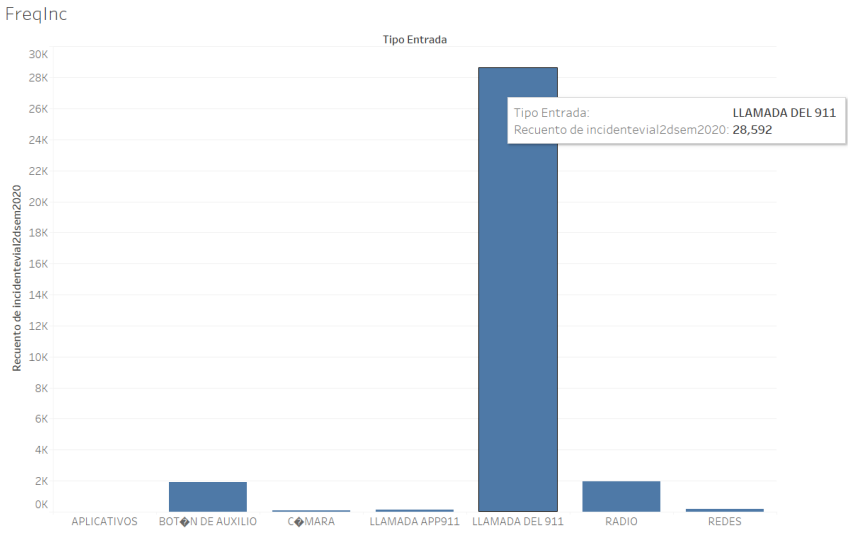
f) ¿Cuál es la clas_con_f_alarma con la mayor cantidad de incidentes viales?

La clas con f alarma con mayor cantidad de incidentes viales es Urgencias Médicas, teniendo al menos 16,482 registros y siento Delito con 43 registros

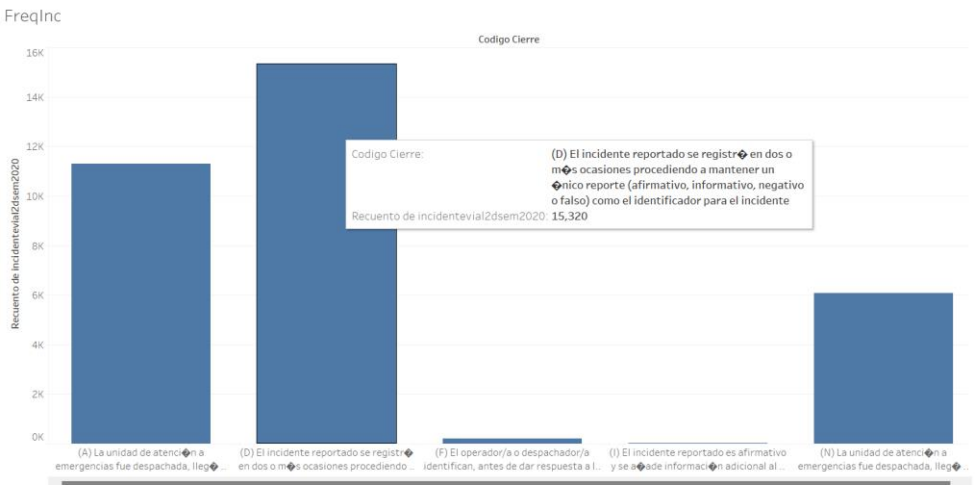


g) ¿Cuál es el tipo_entrada con la mayor cantidad de incidentes viales?

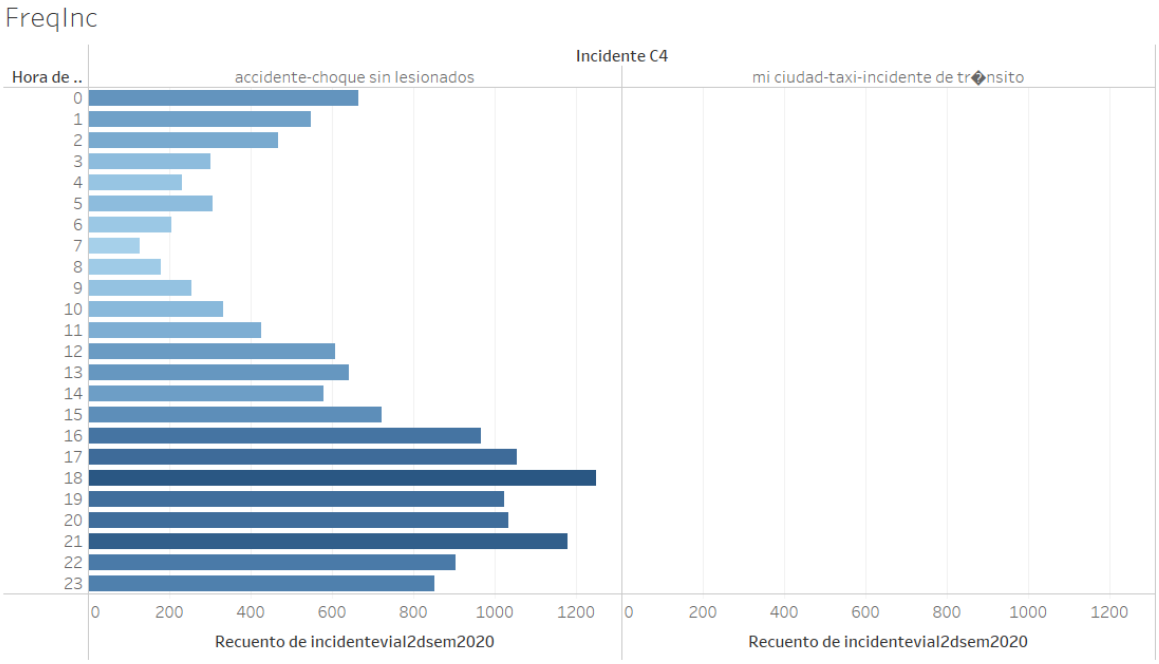
El medio o por donde mas se llamó para reportar estos incidentes viales fue por una llamada al 911 el cual cuenta con 28,592 registros a través del cual se comunicaron.



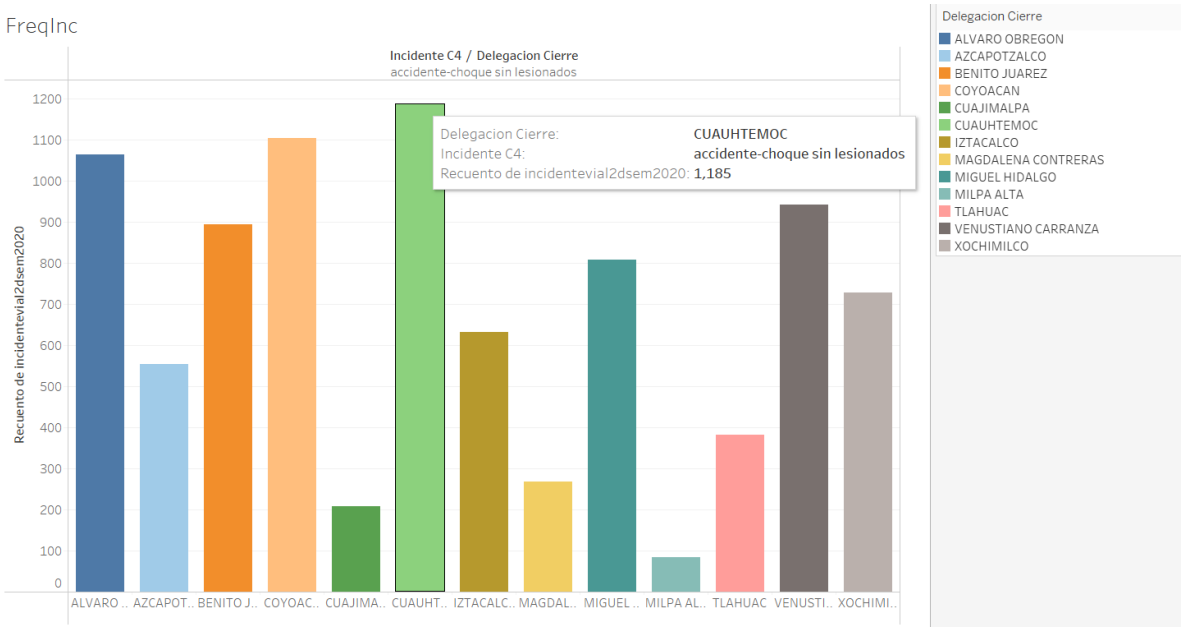
h) ¿Cuál es el código_cierre con la mayor cantidad de incidentes viales?
El código de cierre con la mayor cantidad de incidentes viales o el mayor recuento de incidentes viales es cuando el incidente reportado se registró en dos o más ocasiones procediendo a mantener un único reporte (afirmativo, informativo, negativo o falso) como el identificador para el incidente con 15, 320 registros



i) Considerando el incidente vial más y menos común, ¿cuál es la frecuencia de ocurrencia de estos dos incidentes por hora_cierre?
Nos fijamos en la siguiente imagen que la frecuencia de ocurrencia del choque sin lesionados es entre las 6 de la tarde y 9 de la noche, por otro lado, el incidente de transito de los taxis, vemos que como solo hay un registro, ni siquiera esta listo para poner las horas dentro del gráfico.

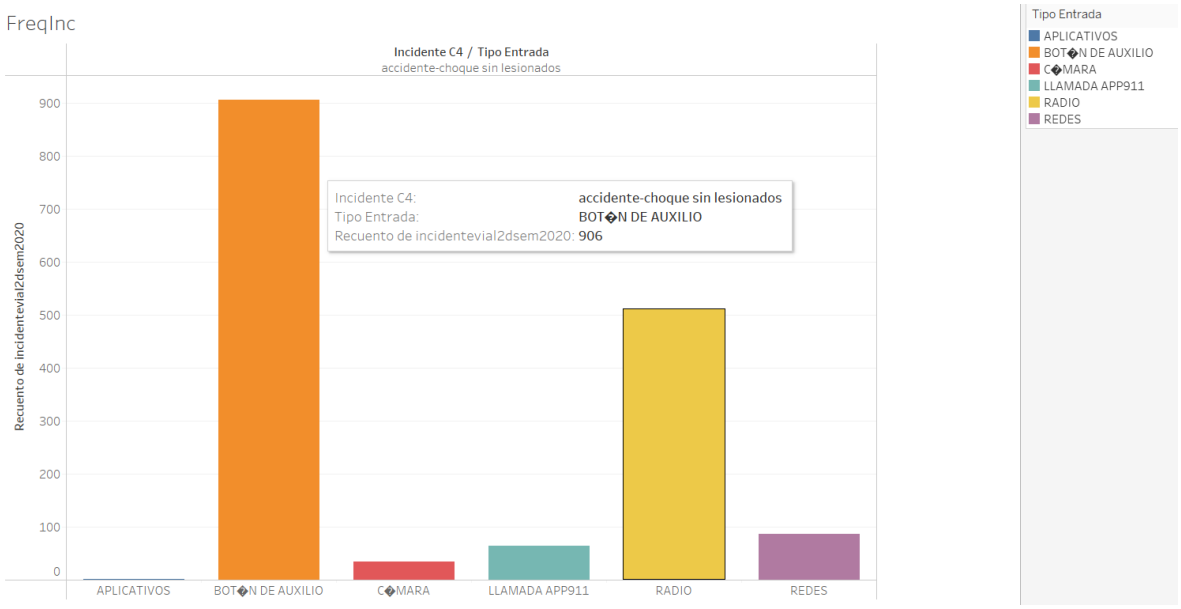


j) Considerando el incidente vial más frecuente, ¿cuál es la frecuencia de ocurrencia por delegación?
La delegación donde más ocurre el choque sin lesionados es la alcaldía / delegación Cuauhtémoc con 1,185 registros de este accidente dentro de la demarcación, por otro lado, la alcaldía con menos ocurrencia es en Milpa Alta con tan solo 84 registros.



k) Considerando el incidente vial más frecuencia, ¿cuál es la frecuencia de ocurrencia por tipo_entrada?

La manera más recurrente en la que los ciudadanos reportaron un accidente de un choque sin lesionados fue a través de un botón de auxilio con 906 registros de este reporte.
Por otro lado, el que menos tuvo reportes por su medio son los aplicativos con 2 registros.



Conclusiones:

Durante el desarrollo de esta práctica pudimos explorar de forma diferente y más gráficamente los datos de incidentes viales durante el segundo semestre del 2020.

Por otro lado, comprendimos de manera más eficiente los datos, ya que Tableau es una herramienta que facilita las interpretaciones al momento de hacer consultas, este software nos ayudó a precisar información y categorizar de manera más directa.

Nos percatamos que los incidentes viales más comunes son los choques sin lesionados y se han reportado mayormente en las horas de 6 de la tarde y 9 de la noche, siendo reportada a través de un botón de auxilio y la alcaldía que tiene más accidentes reportados es Cuauhtémoc.

Estas interpretaciones logramos hacerlas de manera más eficiente por la facilidad de comprensión del software, por ahora. La base de datos limpia también nos proporciono facilidad al hacer dichas consultas, sin embargo, se esta aprendiendo para cuando llegue en el proyecto una base de datos con una dimensión de registros más grandes.