

# Instituto Politécnico Nacional

Escuela Superior de Cómputo

Data Mining

Profesor: Zagal Flores Roberto Eswart

Autor: Velasco Huerta Angel Eduardo

Práctica no. -2

18/09/2021

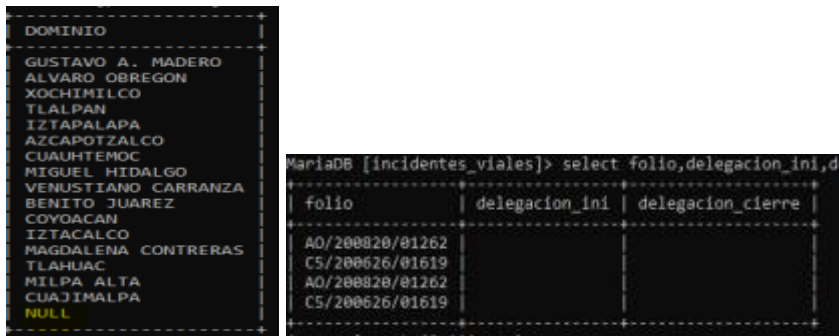
## Identificación de datos incompletos o “sucios”.

Antes de realizar análisis de datos y de explorar las diferentes variables presentes en la información, es indispensable que limpiemos los datos, o que revisemos que estos tengan coherencia.

Para ello, primero limpiaremos aquellos datos que presenten información Nula o vacía, como se está trabajando con un proyecto académico, los borraremos, si se trabajara en un entorno profesional, deberíamos de contextualizar la falta, o error de esos datos.

Estos datos podrían mermar el análisis que se realizara posteriormente.

En la práctica anterior, identificamos 2 filas con valores NULOS de la delegación, sin embargo, cuando se importaron los datos a SQL SERVER, no se mostraban como nulos, curiosamente, después de intentar de todas formas, encontré que los valores estaban definidos como una cadena “NULL”.

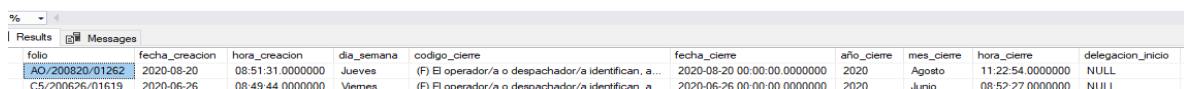


```
DOMINIO
GUSTAVO A. MADERO
ALVARO OBREGON
XOCHIMILCO
TLALPAN
IZTAPALAPA
AZCAPOTZALCO
CUAUHTEMOC
MIGUEL HIDALGO
VENUSTIANO CARRANZA
BENITO JUAREZ
COYOACAN
IZTACALCO
MAGDALENA CONTRERAS
TLAHUAC
MILPA ALTA
CUAJIMALPA
NULL
```

```
MariaDB [incidentes_viales]> select folio,delegacion_ini,delegacion_cierre
+-----+-----+-----+
| folio | delegacion_ini | delegacion_cierre |
+-----+-----+-----+
| AO/200820/01262 | NULL | NULL |
| C5/200626/01619 | NULL | NULL |
+-----+-----+-----+
```

Por lo que los identificamos en SQL SERVER de la siguiente manera:

```
SELECT *
FROM [incidentes2020].[dbo].[incidentevial2dsem2020]
WHERE delegacion_inicio = 'NULL' OR delegacion_cierre = 'NULL';
```



folio	fecha_creacion	hora_creacion	dia_semana	codigo_cierre	fecha_cierre	año_cierre	mes_cierre	hora_cierre	delegacion_inicio
AO/200820/01262	2020-08-20	08:51:31.0000000	Jueves	(F) El operador/a o despachador/a identifican, a...	2020-08-20 00:00:00.0000000	2020	Agosto	11:22:54.0000000	NULL
C5/200626/01619	2020-06-26	08:49:44.0000000	Viernes	(F) El operador/a o despachador/a identifican, a...	2020-06-26 00:00:00.0000000	2020	Junio	08:52:27.0000000	NULL

Los eliminamos:

```
DELETE
FROM [incidentes2020].[dbo].[incidentevial2dsem2020]
WHERE delegacion_inicio = 'NULL' OR delegacion_cierre = 'NULL';
```

En la práctica pasada también identificamos algunos registros repetidos, esto podía ser un error del manejador utilizado, pero al momento de importar los datos en SQL SERVER, identificamos el folio como una primary key, por lo que ya no existe la posibilidad de que existan elementos repetidos.

Una vez terminado esto, podemos contabilizar los registros con los que realizaremos el análisis.

```
select count(*) as TOTAL from incidentevia12dsem2020;
```

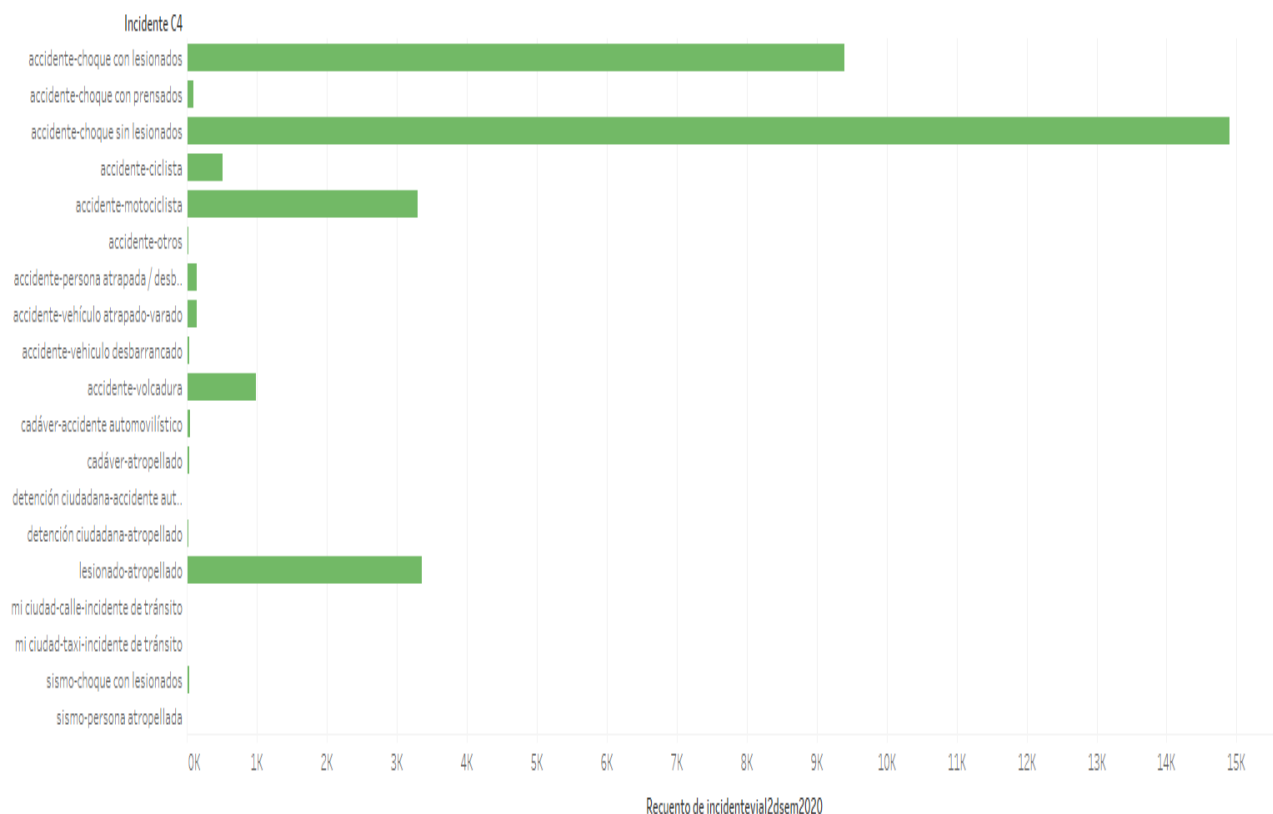
	TOTAL
1	33069

## Análisis con Tableau.

Se utilizará el software Tableau, pues nos permite realizar vistas de nuestra información, que son más fáciles de interpretar.

**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?**

frec\_inci

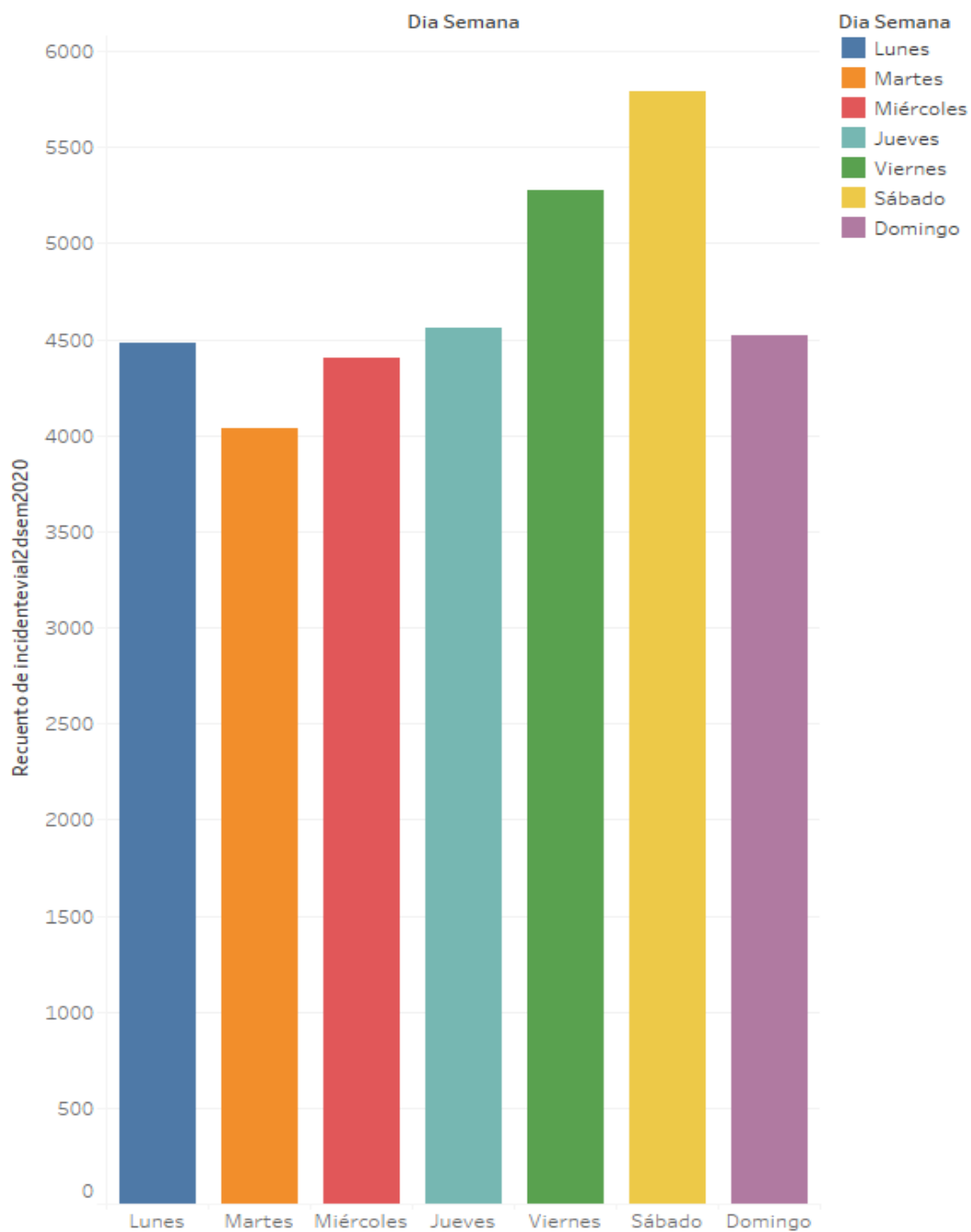


Recuento de incidente via12dsem2020 para cada Incidente C4.

El más repetido es accidente-choque sin lesionados  
El menos repetido es mi-ciudad taxi incidente de tráfico.

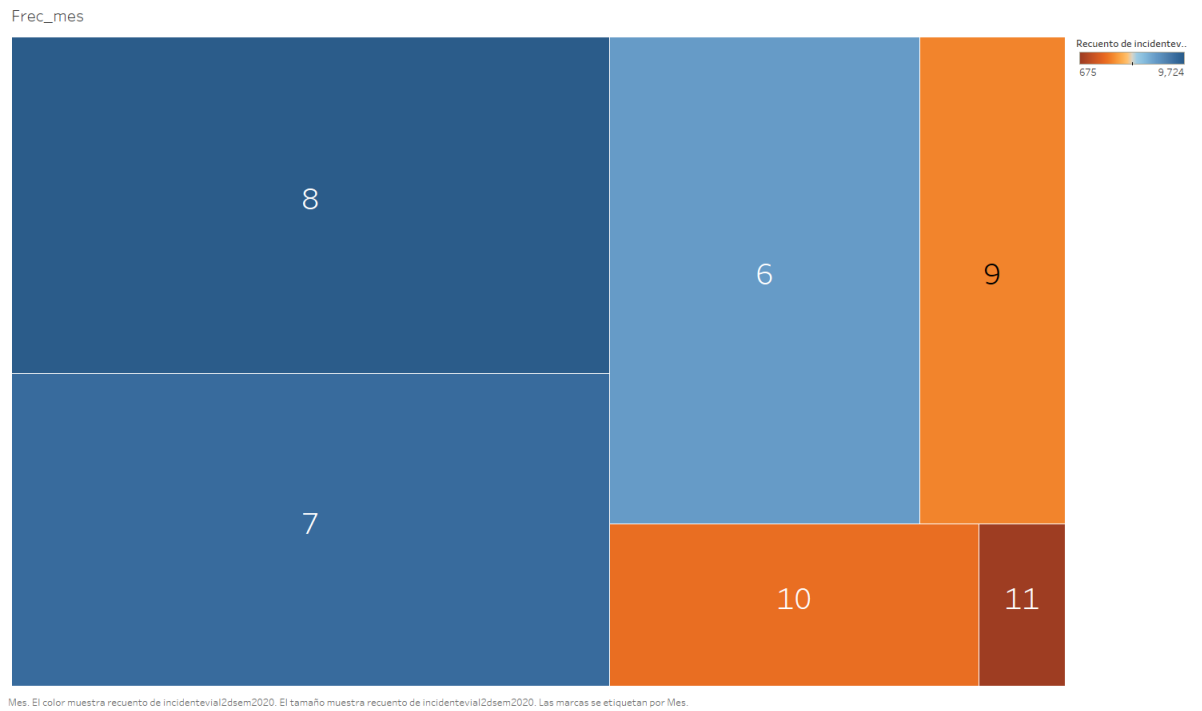
B- ¿Cuál es el día\_semana con la mayor cantidad de incidentes viales?

frec\_inci

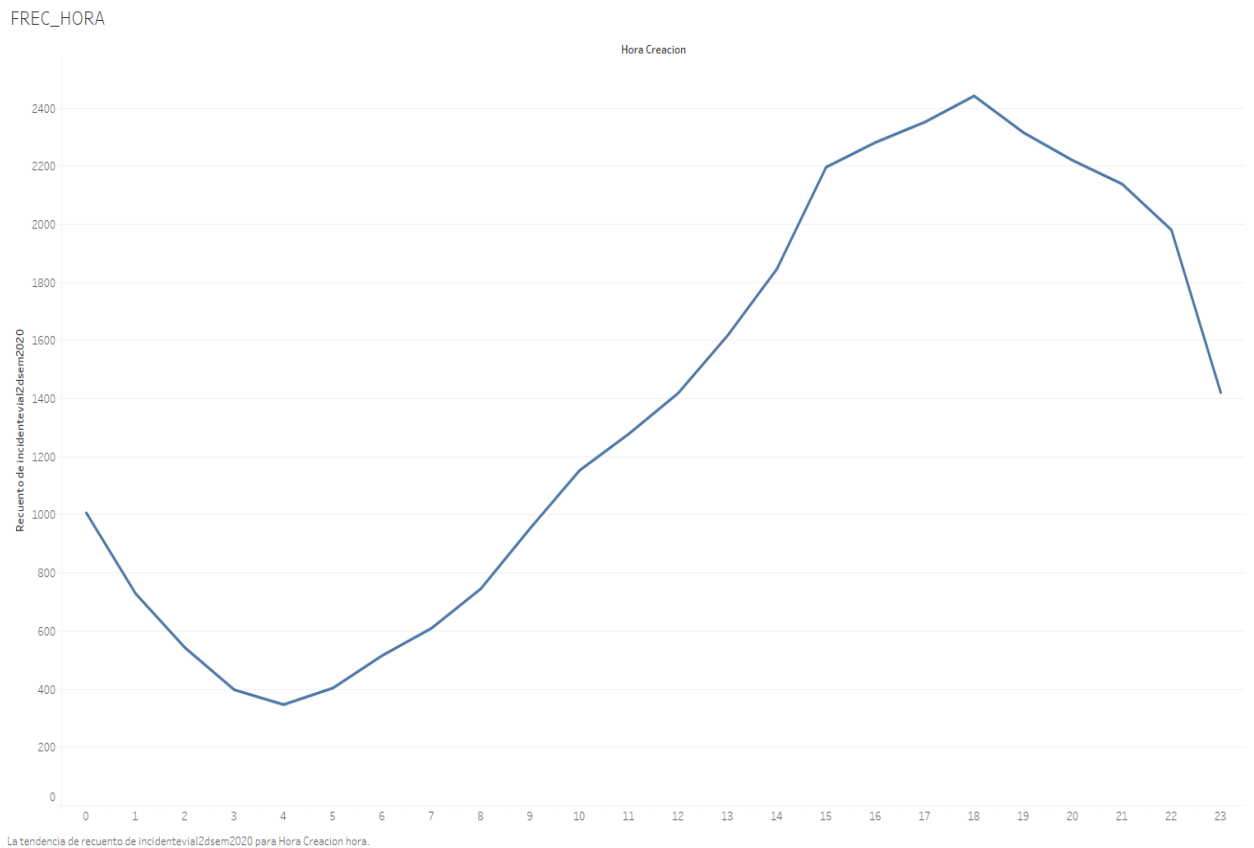


Recuento de incidentevial2dsem2020 para cada Dia Semana. El color muestra detalles acerca de Dia Semana.

C- ¿Cuál es el mes (**fecha\_creacion**) con la mayor cantidad de incidentes viales?



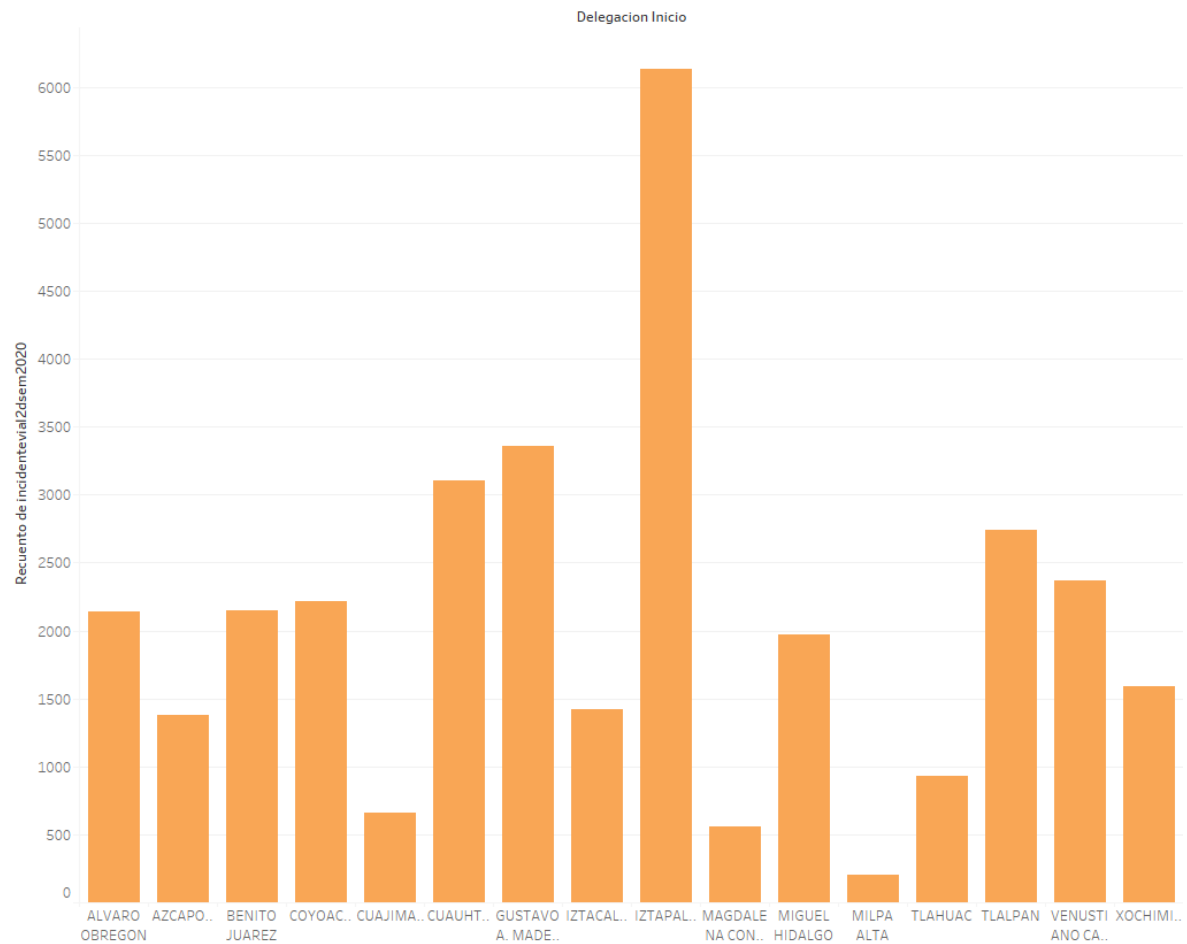
D- ¿Cuál es la **hora\_creacion** con la mayor cantidad de incidentes viales?



A. ¿Cuál es la **delegación\_inicio** con la mayor cantidad de incidentes viales?

E- ¿Cuál es la **delegación\_inicio** con la mayor cantidad de incidentes viales?

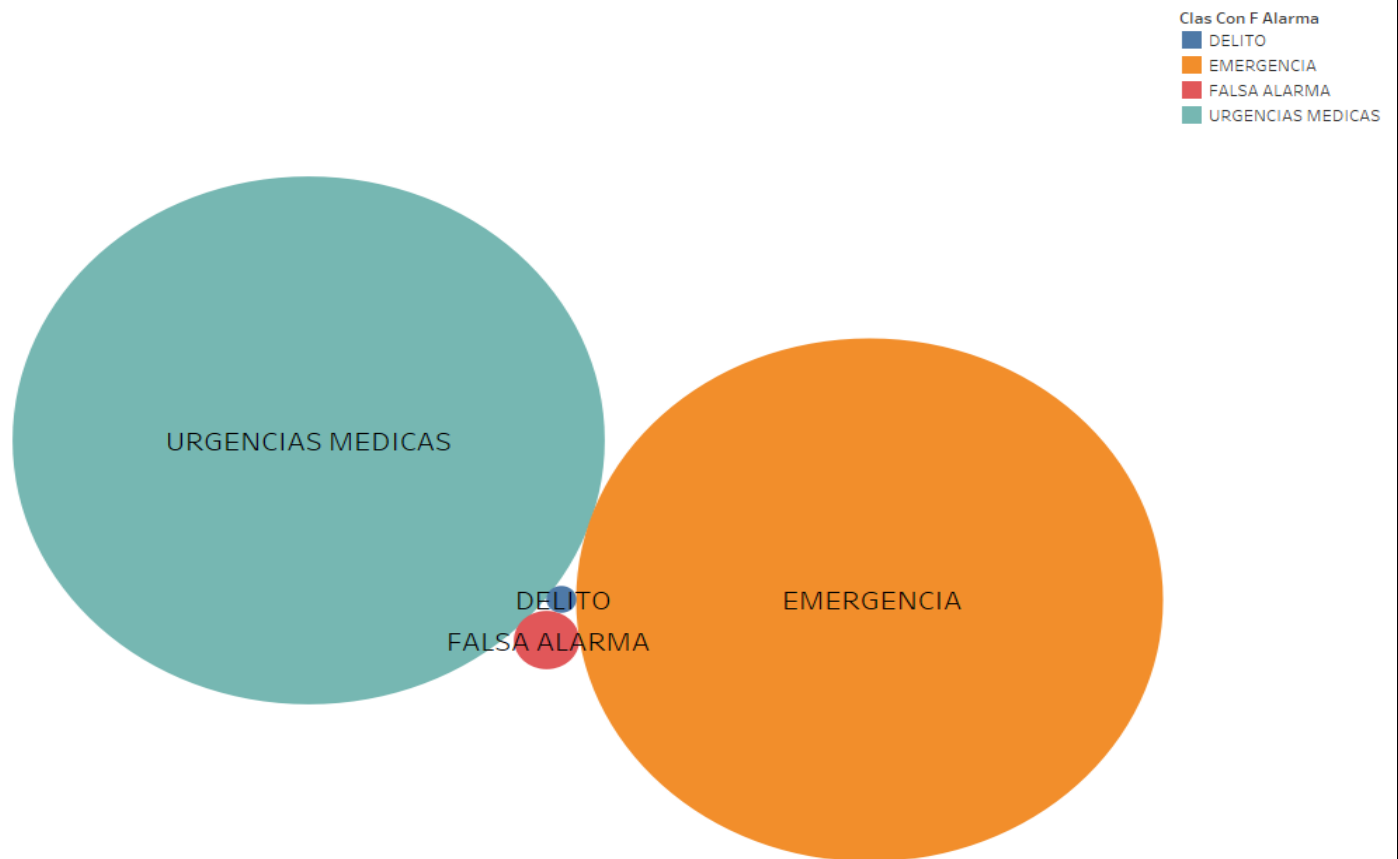
DELINI\_FREC



Recuento de incidentes viales 2dsem2020 para cada Delegacion Inicio.

F- ¿Cuál es la **clas\_con\_f\_alarma** con la mayor cantidad de incidentes viales?

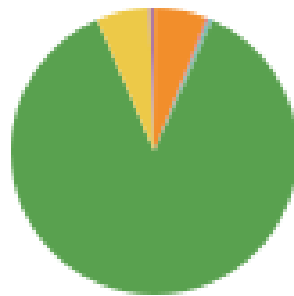
ALARMA\_ACCIDENTE



Clas Con F Alarma. El color muestra detalles acerca de Clas Con F Alarma. El tamaño muestra recuento de incidente viales 2dsem2020. Las marcas se etiquetan por Clas Con F Alarma.

G- ¿Cuál es el **tipo\_entrada** con la mayor cantidad de incidentes viales?

ENTRADA\_FREC



Tipo Entrada

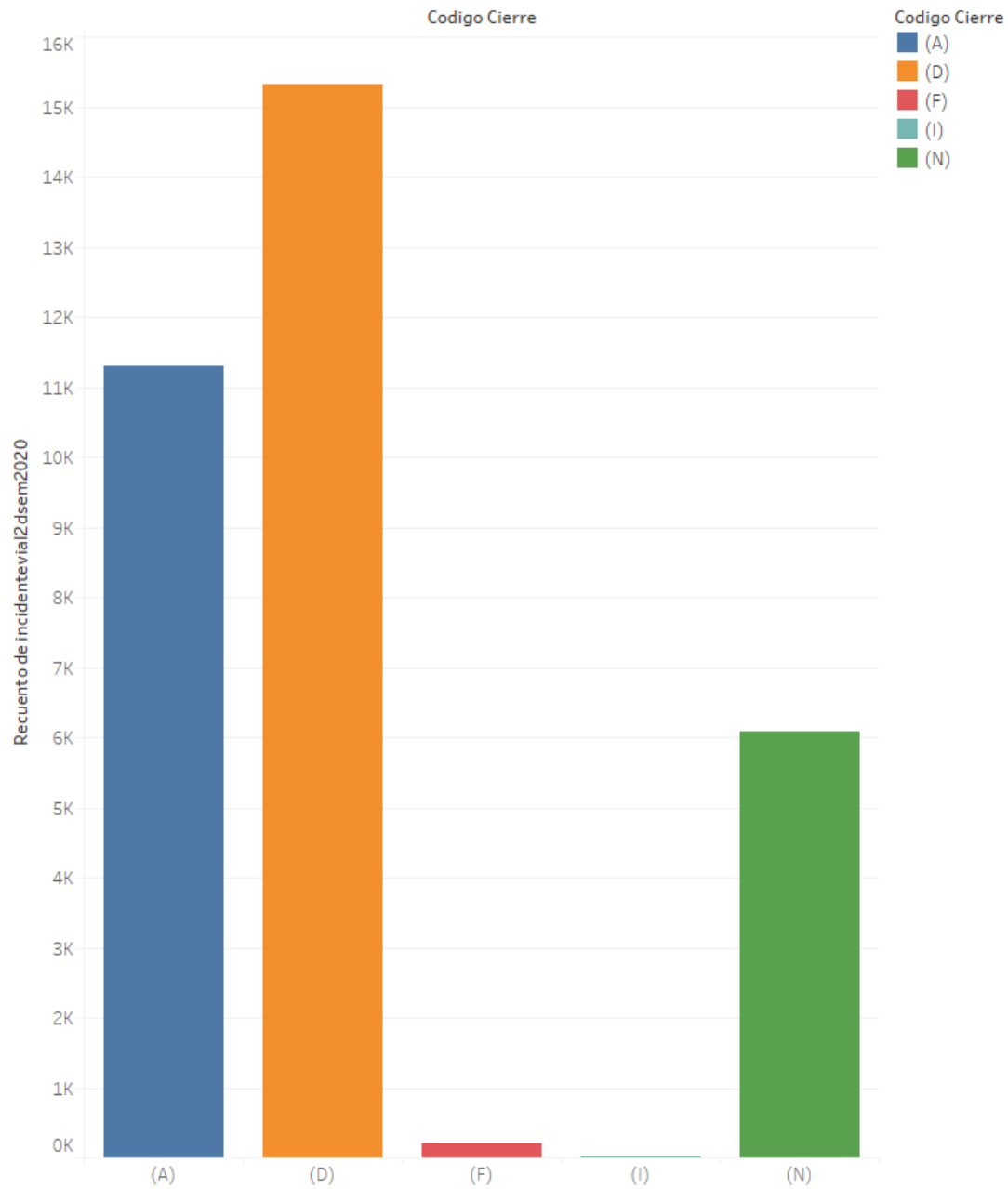


Tipo Entrada (color).



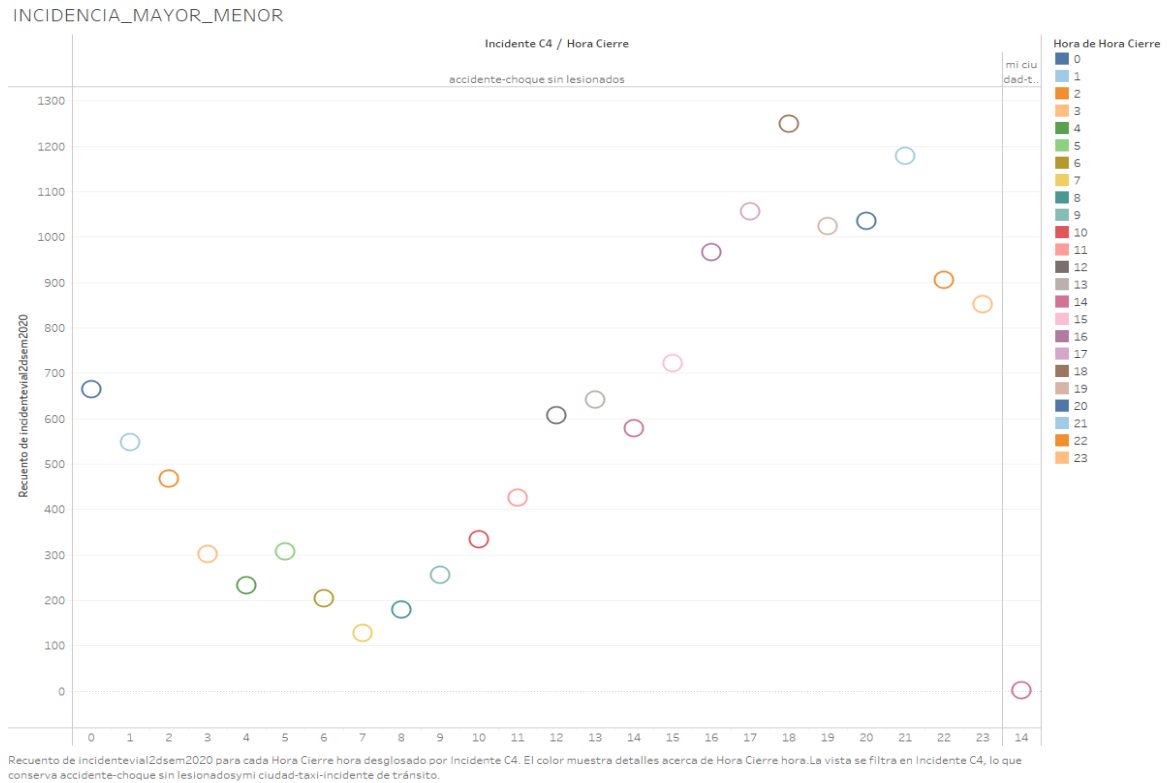
¿Cuál es el **codigo\_cierre** con la mayor cantidad de incidentes viales?

COD\_CIERRE

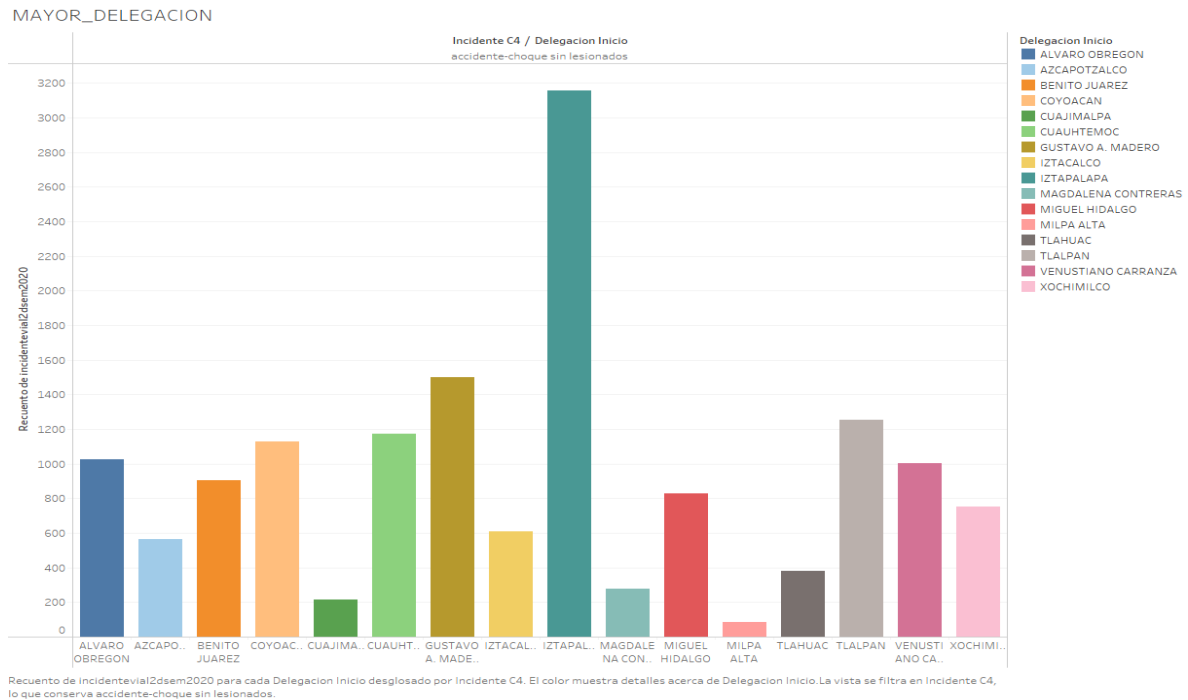


Recuento de incidente viales 2dsem2020 para cada Codigo Cierre. El color muestra detalles acerca de Codigo Cierre.

Considerando el incidente vial más y menos común, ¿cuál es la frecuencia de ocurrencia de estos dos incidentes por hora\_cierre?

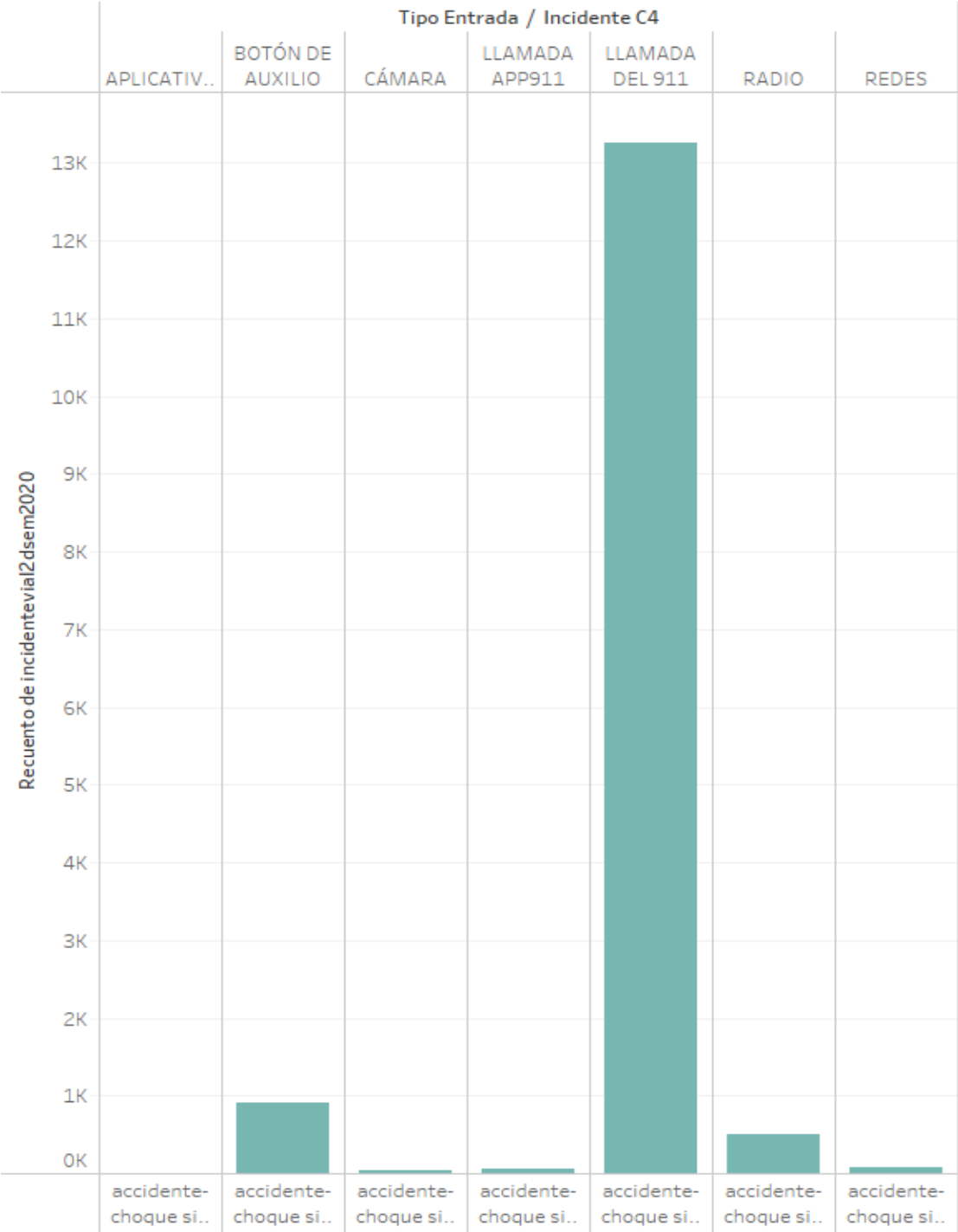


Considerando el incidente vial más frecuente, ¿cuál es la frecuencia de ocurrencia por delegación?



Considerando el incidente vial **más frecuente**, ¿cuál es la frecuencia de ocurrencia por **tipo\_entrada**?

MAYOR\_DELEGACION



Recuento de incidente vial 2dsem2020 para cada Incidente C4 desglosado por Tipo Entrada. La vista se filtra en Incidente C4, lo que conserva accidente-choque sin lesionados.

## CONCLUSIONES:

En esta práctica, pudimos realizar la limpieza de los datos con los que estamos trabajando, asimismo, realizamos un análisis básico en Tableau, que si bien, se puede realizar algo similar con Queries en SQL (para obtener los datos) el graficar esta información, hace que visualizar la misma sea extremadamente sencillo, pues a un simple vistazo podemos identificar los de mayor y menor frecuencia.

También aprendimos a limpiar nuestros datos ya al momento de representar esta información, pues cuando realizamos ciertos análisis, necesitábamos solo cierta información, por ejemplo, el accidente más repetido, y para ello aplicamos mas filtros que limpien la información.

Finalmente puedo concluir que, gracias a esta práctica, pude entender mejor el porque de la limpieza de datos en bases medianamente grandes como estas, pues muchas veces si la información no está representada como debería de, algunas estadísticas no resultarían como las pensamos.