



Instituto Politécnico Nacional

Escuela Superior de Cómputo

Grupo 3CV19 Data Mining

Profesor Zagal Flores Roberto Eswart

1er Parcial Practica #2: Limpieza de datos y exploración básica.

Alumno García González Aarón Antonio

Martes 23 de marzo de 2021



Índice

Objetivo3

Introducción3

Desarrollo.....4

Conclusiones.....15

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.

Introducción

Explicación de la propuesta de solución

La manera en que resolveremos esta tarea consta de las siguientes etapas:

- Importación de datos a motor de base de datos SQL SERVER
- Limpieza de datos
- Exportación de datos limpios
- Importación de datos limpios a software de visualización
- Visualización e interpretación del requerimiento de datos en tableau

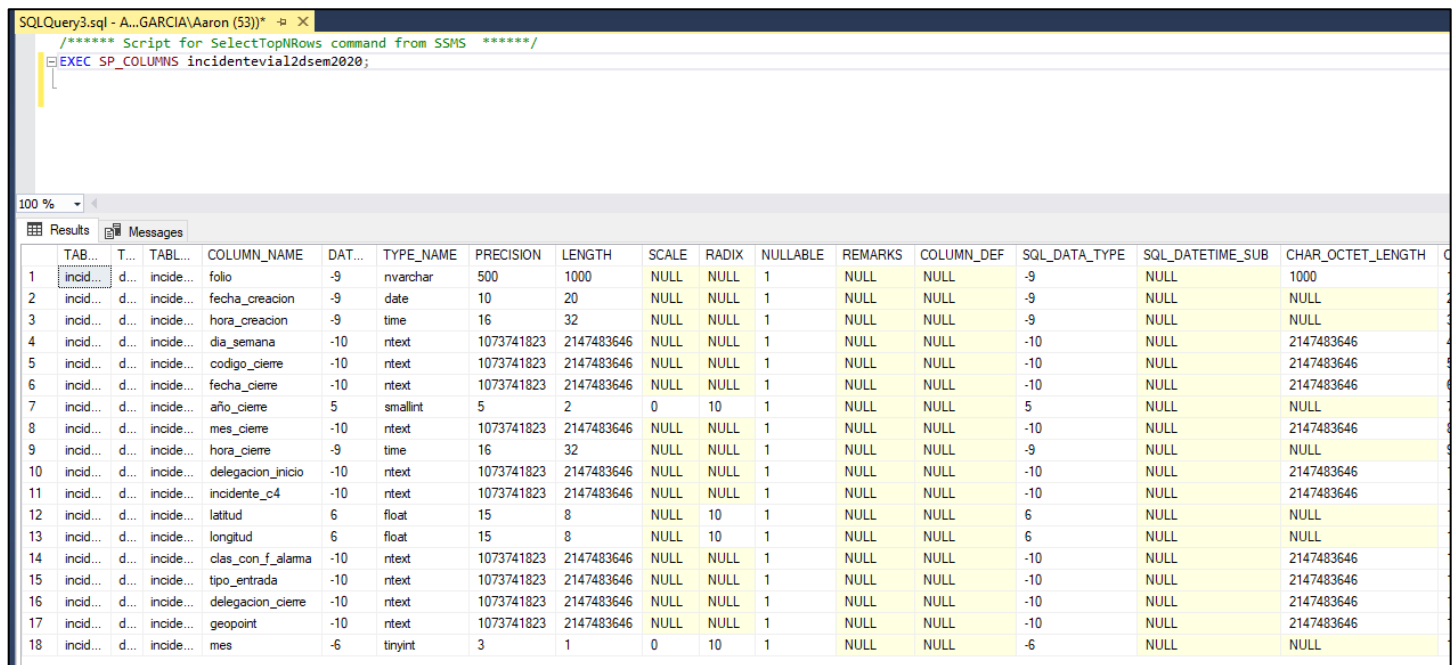
Previo a todo ello se solicito la licencia de estudiante en tableau.com

Desarrollo

Procedimiento: Realizar un análisis exploratorio de datos a nivel básico a la base de datos de incidentes viales.

0. Revise la clase que corresponda al tema “exploración básica de datos con Tableau” y el tema de “limpieza de datos”. ✓
1. Utilice el dataset de incidentes viales de la práctica 1 ✓
2. Identifique valores NULOS y errores en los formatos de tipo datos, reporte y documente los hallazgos de datos inconsistentes. Proceda a eliminarlos de la base (solo en caso de que la inconsistencia de los datos afecte a la interpretación de cada registro).

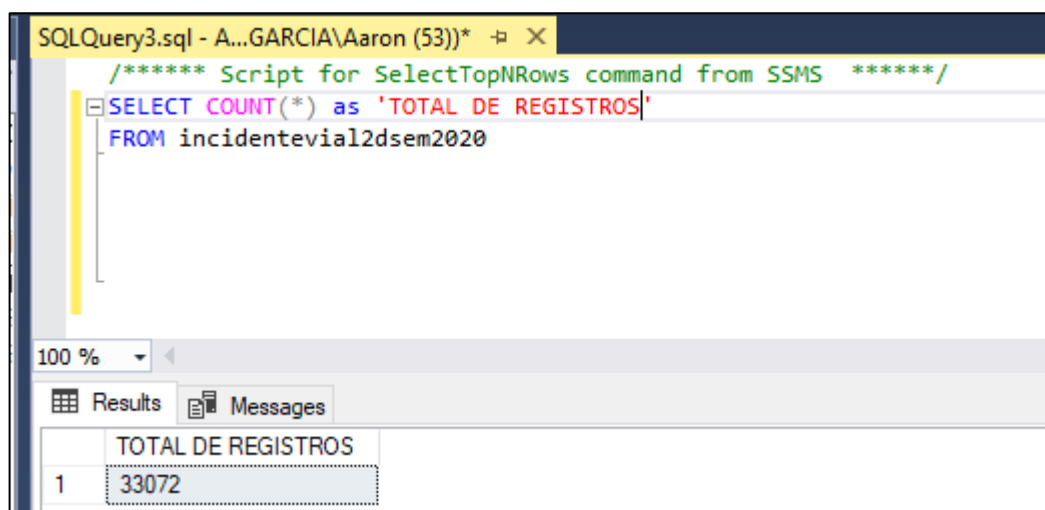
Para empezar, realicé un equivalente a DESCRIBE TABLE de mysql, pero ahora en SQL server con el objetivo de



TAB...	T...	TABL...	COLUMN_NAME	DAT...	TYPE_NAME	PRECISION	LENGTH	SCALE	RADIX	NULLABLE	REMARKS	COLUMN_DEF	SQL_DATA_TYPE	SQL_DATETIME_SUB	CHAR_OCTET_LENGTH
1	incid...	d...	folio	-9	nvarchar	500	1000			1			-9		1000
2	incid...	d...	fecha_creacion	-9	date	10	20			1			-9		
3	incid...	d...	hora_creacion	-9	time	16	32			1			-9		
4	incid...	d...	dia_semana	-10	ntext	1073741823	2147483646			1			-10		2147483646
5	incid...	d...	codigo_cierre	-10	ntext	1073741823	2147483646			1			-10		2147483646
6	incid...	d...	fecha_cierre	-10	ntext	1073741823	2147483646			1			-10		2147483646
7	incid...	d...	año_cierre	5	smallint	5	2	0	10	1			5		
8	incid...	d...	mes_cierre	-10	ntext	1073741823	2147483646			1			-10		2147483646
9	incid...	d...	hora_cierre	-9	time	16	32			1			-9		
10	incid...	d...	delegacion_inicio	-10	ntext	1073741823	2147483646			1			-10		2147483646
11	incid...	d...	incidente_c4	-10	ntext	1073741823	2147483646			1			-10		2147483646
12	incid...	d...	latitud	6	float	15	8		10	1			6		
13	incid...	d...	longitud	6	float	15	8		10	1			6		
14	incid...	d...	clas_con_f_alama	-10	ntext	1073741823	2147483646			1			-10		2147483646
15	incid...	d...	tipo_entrada	-10	ntext	1073741823	2147483646			1			-10		2147483646
16	incid...	d...	delegacion_cierre	-10	ntext	1073741823	2147483646			1			-10		2147483646
17	incid...	d...	geopoint	-10	ntext	1073741823	2147483646			1			-10		2147483646
18	incid...	d...	mes	-6	tinyint	3	1	0	10	1			-6		

conocer el tipo de dato de cada columna o atributo de la tabla de hechos.

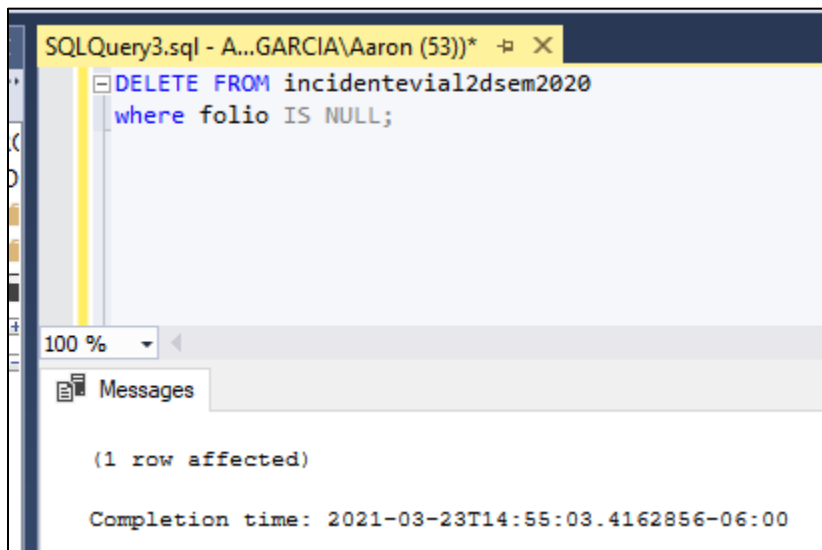
Antes de realizar ninguna alteración, consultaremos el total de tuplas que incluye en data set proporcionado.



TOTAL DE REGISTROS	
1	33072

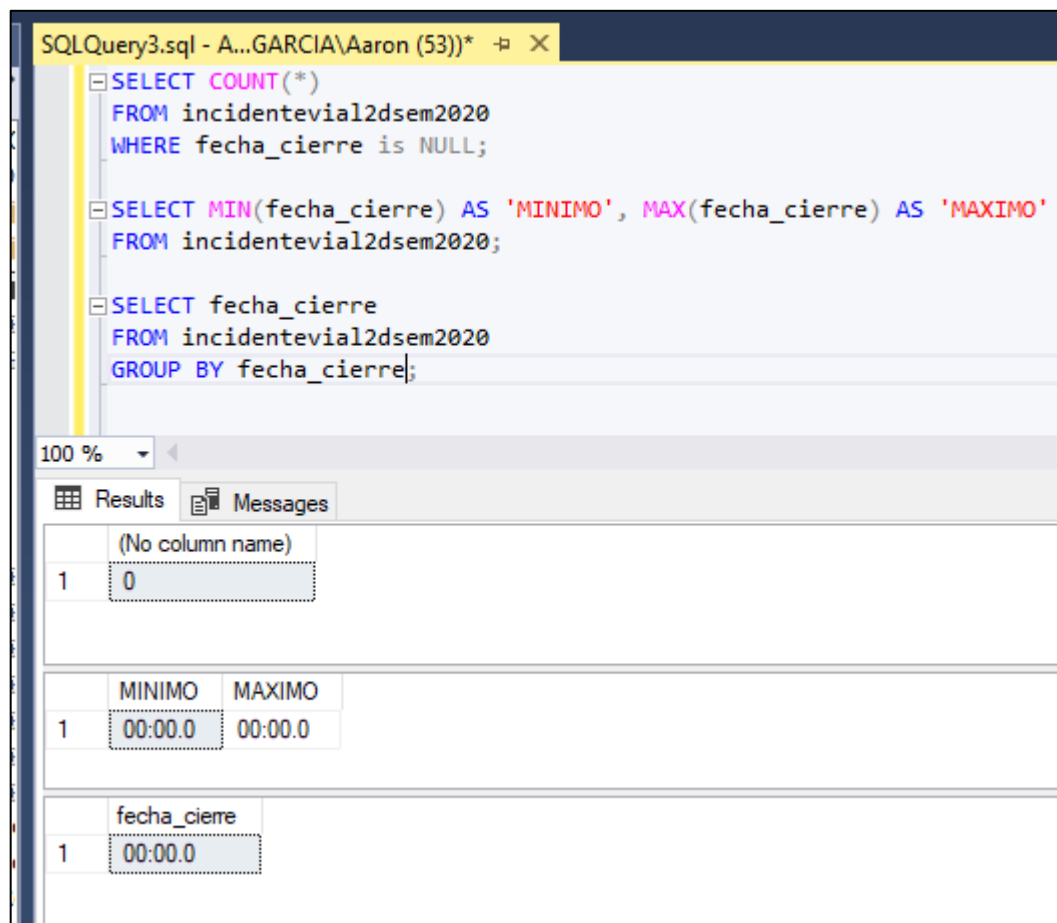
Ahora buscaré valores nulos y errores en los formatos de tipo datos por cada atributo.

En el atributo “folio”, encontramos un registro nulo, por lo que lo eliminamos



The screenshot shows a SQL query window titled "SQLQuery3.sql - A...GARCIA\Aaron (53))". The query is: `DELETE FROM incidentevisual2dsem2020 where folio IS NULL;`. Below the query, the "Messages" pane shows the result: `(1 row affected)` and the completion time: `2021-03-23T14:55:03.4162856-06:00`.

En el caso del atributo “fecha_cierre”, no encontramos valores nulos, pero el catálogo o rango de valores que puede tomar solo incluye un valor: “00:00.0”, esto además debería de ser de tipo DATE.



The screenshot shows three SQL queries in a window titled "SQLQuery3.sql - A...GARCIA\Aaron (53))".

- Query 1: `SELECT COUNT(*) FROM incidentevisual2dsem2020 WHERE fecha_cierre is NULL;`
- Query 2: `SELECT MIN(fecha_cierre) AS 'MINIMO', MAX(fecha_cierre) AS 'MAXIMO' FROM incidentevisual2dsem2020;`
- Query 3: `SELECT fecha_cierre FROM incidentevisual2dsem2020 GROUP BY fecha_cierre;`

The "Results" pane shows the following data:

	(No column name)
1	0

	MINIMO	MAXIMO
1	00:00.0	00:00.0

	fecha_cierre
1	00:00.0

En el caso del atributo “año_cierre”, solo hay un valor: 2020, por lo que no se podría variar o filtrar datos por este atributo.

SQLQuery1.sql - A...GARCIA\Aaron (51))*

```

SELECT COUNT(*)
FROM incidente12dsem2020
WHERE año_cierre is NULL;

SELECT MIN(año_cierre) AS 'MINIMO', MAX(año_cierre) AS 'MAXIMO'
FROM incidente12dsem2020;

SELECT año_cierre
FROM incidente12dsem2020
GROUP BY año_cierre;

```

100 %

Results Messages

(No column name)
1 0

MINIMO	MAXIMO
1 2020	2020

año_cierre
1 2020

En el caso del atributo “hora_cierre”, encontramos 151 valores nulos, en este caso que estamos haciendo una exploración inicial por lo que requerimos que todos los atributos estén lo más correctos y menos distorsionados en conjunto, se eliminarán dichas tuplas.

SQLQuery1.sql - A...GARCIA\Aaron (51))*

```

SELECT COUNT(*) AS 'VALORES NULOS'
FROM incidente12dsem2020
WHERE hora_cierre is NULL;

SELECT *
FROM incidente12dsem2020
WHERE hora_cierre is NULL;

SELECT MIN(hora_cierre) AS 'MINIMO', MAX(hora_cierre) AS 'MAXIMO'
FROM incidente12dsem2020;

SELECT hora_cierre
FROM incidente12dsem2020
GROUP BY hora_cierre;

```

100 %

Results Messages

VALORES NULOS
1 151

folio	fecha_creacion	hora_creacion	día_semana	codigo_cierre	fecha_cierre	año_cierre	mes_cierre	hora_cierre	delegacion_inicio	incidente_c4	latitud
1 C5/200906/01906	2020-09-06	NULL	Domingo	(A) La unidad de atención a emergencias fue desp...	00:00:0	2020	Septiembre	NULL	GUSTAVO A. MADERO	accidente-choque sin lesionados	19.
2 C5/200906/04920	2020-09-06	NULL	Domingo	(D) El incidente reportado se registró en dos o más ...	00:00:0	2020	Septiembre	NULL	GUSTAVO A. MADERO	accidente-choque sin lesionados	19.
3 C5/200906/08022	2020-09-06	NULL	Domingo	(D) El incidente reportado se registró en dos o más ...	00:00:0	2020	Septiembre	NULL	IZTACALCO	accidente-choque sin lesionados	19.
4 C5/200906/07607	2020-09-06	NULL	Domingo	(N) La unidad de atención a emergencias fue desp...	00:00:0	2020	Septiembre	NULL	COYOACAN	accidente-choque sin lesionados	19.
5 C5/200906/00673	2020-09-06	NULL	Domingo	(N) La unidad de atención a emergencias fue desp...	00:00:0	2020	Septiembre	NULL	MAGDALENA CONTRERAS	accidente-choque con lesionados	19.
6 C5/200906/04915	2020-09-06	NULL	Domingo	(N) La unidad de atención a emergencias fue desp...	00:00:0	2020	Septiembre	NULL	GUSTAVO A. MADERO	accidente-choque sin lesionados	19.
7 C5/200906/04918	2020-09-06	NULL	Domingo	(D) El incidente reportado se registró en dos o más ...	00:00:0	2020	Septiembre	NULL	GUSTAVO A. MADERO	accidente-choque sin lesionados	19.
8 C5/200906/08136	2020-09-06	NULL	Domingo	(A) La unidad de atención a emergencias fue desp...	00:00:0	2020	Septiembre	NULL	TLAHUAC	lesionado-atropellado	19.

MINIMO	MAXIMO
1 00:00:00.00000000	23:59:59.00000000

hora_cierre
1 04:58:22.00000000
2 00:00:00.00000000
3 05:24:50.00000000
4 01:27:54.00000000

En el caso del atributo “delegación_inicio”, encontramos 2 valores nulos de manera varchar o cadena, igual los eliminaremos.

SQLQuery1.sql - A...GARCIA\Aaron (511)*

```

SELECT COUNT(*) AS 'VALORES NULOS'
FROM incidente_vial2dsem2020
WHERE delegacion_inicio IS NULL;

SELECT COUNT(*) AS 'VALORES NULOS STRING'
FROM incidente_vial2dsem2020
WHERE delegacion_inicio = 'NULL';

SELECT *
FROM incidente_vial2dsem2020
WHERE delegacion_inicio IS NULL
OR delegacion_inicio = 'NULL';

SELECT MIN(delegacion_inicio) AS 'MINIMO', MAX(delegacion_inicio) AS 'MAXIMO'
FROM incidente_vial2dsem2020;

SELECT delegacion_inicio
FROM incidente_vial2dsem2020
GROUP BY delegacion_inicio;

```

100 %

Results Messages

1	2
folio	fecha_creacion
AO/200820/01262	2020-08-20
C5/200626/01619	2020-06-26
MINIMO	MAXIMO
ALVARO OBREGON	XOCHIMILCO
delegacion_inicio	
MIGUEL HIDALGO	
MILPA ALTA	
AZCAPOTZALCO	
CUAJIMALPA	
TLALPAN	
NULL	
VENUSTIANO C...	
IZTACALCO	

En el caso del atributo “delegacion_cierre”, encontré 2 valores nulos de manera varchar o cadena, igual lo eliminaremos.

SQLQuery1.sql - A...GARCIA\Aaron (511)*

```

SELECT COUNT(*) AS 'VALORES NULOS'
FROM incidente_vial2dsem2020
WHERE delegacion_cierre IS NULL;

SELECT COUNT(*) AS 'VALORES NULOS STRING'
FROM incidente_vial2dsem2020
WHERE delegacion_cierre = 'NULL';

SELECT *
FROM incidente_vial2dsem2020
WHERE delegacion_cierre IS NULL
OR delegacion_cierre = 'NULL';

SELECT MIN(delegacion_cierre) AS 'MINIMO', MAX(delegacion_cierre) AS 'MAXIMO'
FROM incidente_vial2dsem2020;

SELECT delegacion_cierre
FROM incidente_vial2dsem2020
GROUP BY delegacion_cierre;

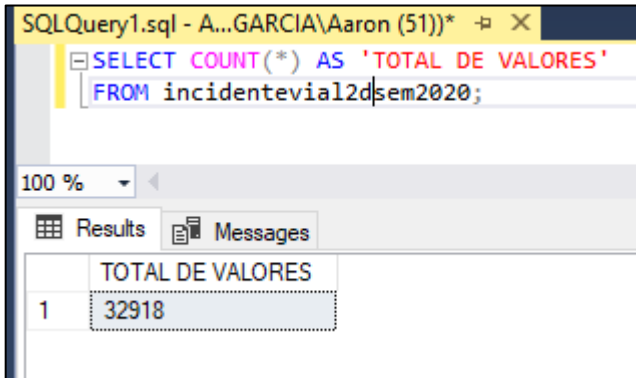
```

100 %

Results Messages

1	2
VALORES NULOS	
0	
VALORES NULOS STRING	
2	
folio	fecha_creacion
AO/200820/01262	2020-08-20
C5/200626/01619	2020-06-26
MINIMO	MAXIMO
ALVARO OBREGON	XOCHIMILCO
delegacion_cierre	
CUAJIMALPA	
TLALPAN	
NULL	
VENUSTIANO C...	
IZTACALCO	

Finalmente, después de eliminar todas las tuplas con datos nulos, obtenemos 32,918 registros de los 33,072.



The screenshot shows a SQL query window titled 'SQLQuery1.sql - A...GARCIA\Aaron (51))'. The query is: `SELECT COUNT(*) AS 'TOTAL DE VALORES' FROM incidentevia12dsem2020;`. Below the query, the 'Results' tab is active, displaying a single row with the column 'TOTAL DE VALORES' and the value '32918'.

TOTAL DE VALORES
32918

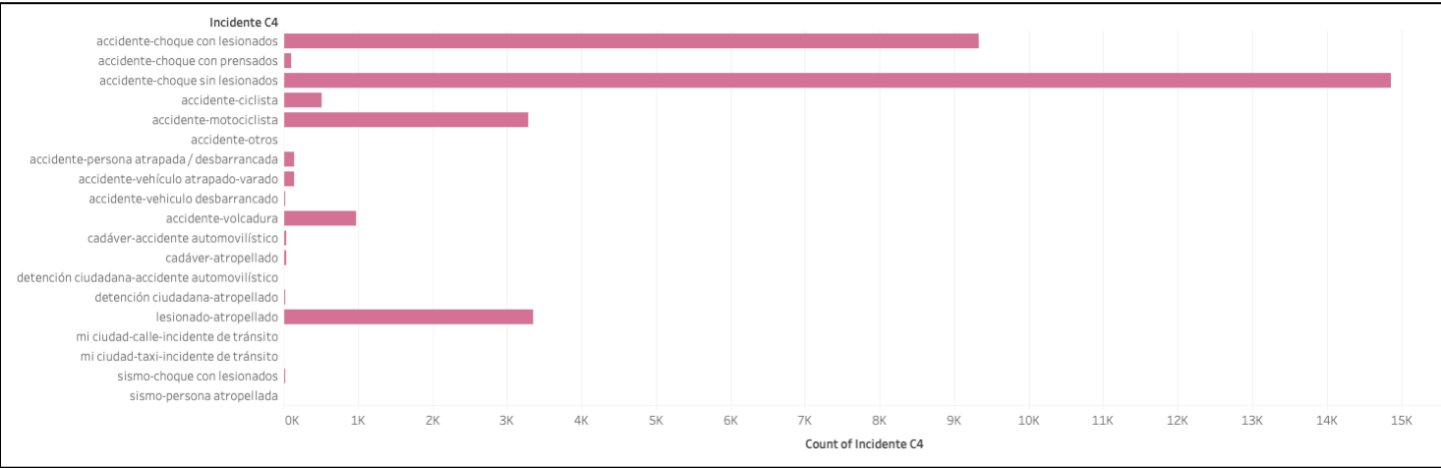
Revise todas las columnas, pero comience y ponga especial atención en las siguientes que ya fueron analizadas en la práctica 1 (de hecho, se sugiere utilice los hallazgos identificados de la práctica 1):

- ⇒ Fecha_creacion ✓
- ⇒ Año_cierre y hora_cierre (todos los relacionados al cierre) ✓
- ⇒ Incidente_c4 ✓
- ⇒ Tipo_entrada ✓
- ⇒ Clas_con_f_alarma ✓
- ⇒ Delegación ✓

¿Cuántos registros inconsistentes encontró? ¿Cuántos registros después de la limpieza obtuvo como total en la muestra de datos? **32,918 de 33,072**

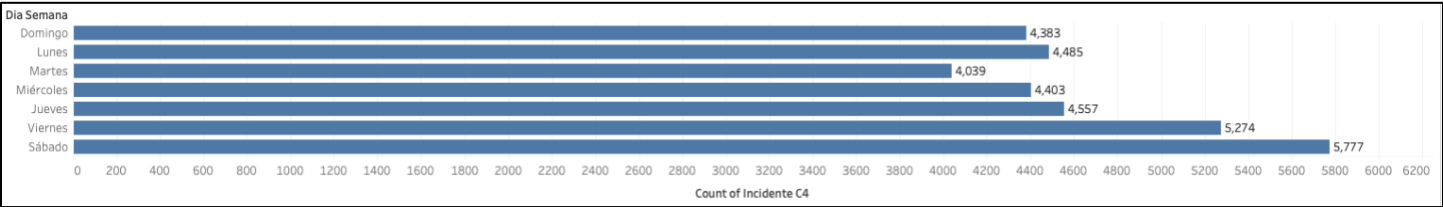
3. Realice el análisis correspondiente en Tableau, se recomienda usar el procedimiento de la clase ““exploración básica de datos con Tableau”. Documente el resultado a fin de responder a las siguientes preguntas de exploración de datos (realice las gráficas según corresponda):
- (Después de realizar la limpieza anterior, exporte esa nueva tabla de hechos a csv, la cual ahora importe a tableau)
- 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?

El incidente vial más frecuente es choque sin lesionados, el incidente vial menos frecuente es “mi ciudad taxi incidente de tránsito”.

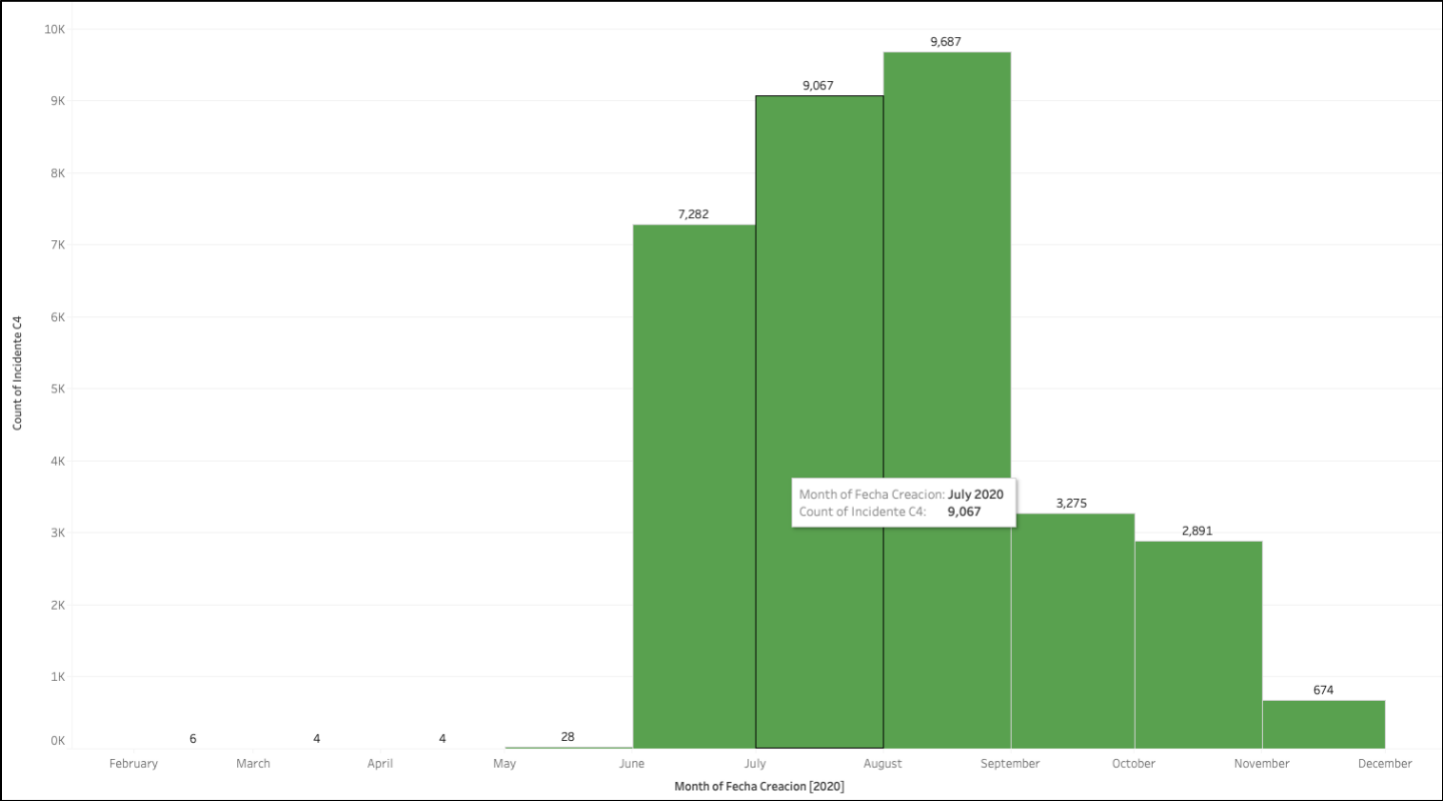


Incidente C4	
accidente-choque con lesionados	9,336
accidente-choque con prensados	99
accidente-choque sin lesionados	14,857
accidente-ciclista	512
accidente-motociclista	3,283
accidente-otros	17
accidente-persona atrapada / desbarrancada	142
accidente-vehículo atrapado-varado	142
accidente-vehículo desbarrancado	32
accidente-volcadura	983
cadáver-accidente automovilístico	42
cadáver-atropellado	37
detención ciudadana-accidente automovilístico	15
detención ciudadana-atropellado	28
lesionado-atropellado	3,348
mi ciudad-calle-incidente de tránsito	9
mi ciudad-taxi-incidente de tránsito	1
sismo-choque con lesionados	32
sismo-persona atropellada	3

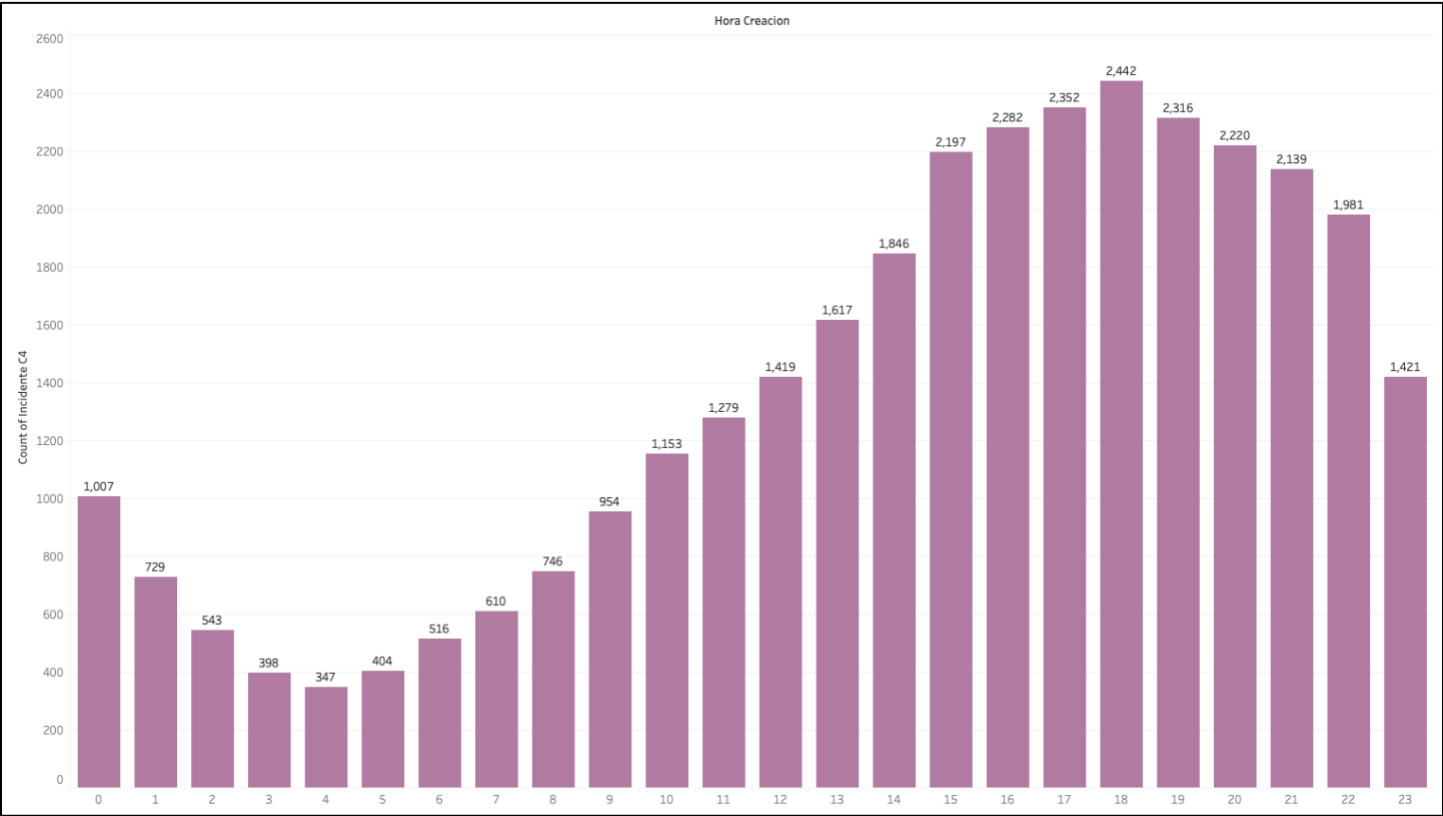
b) ¿Cuál es el día_semana con la mayor cantidad de incidentes viales? El sábado



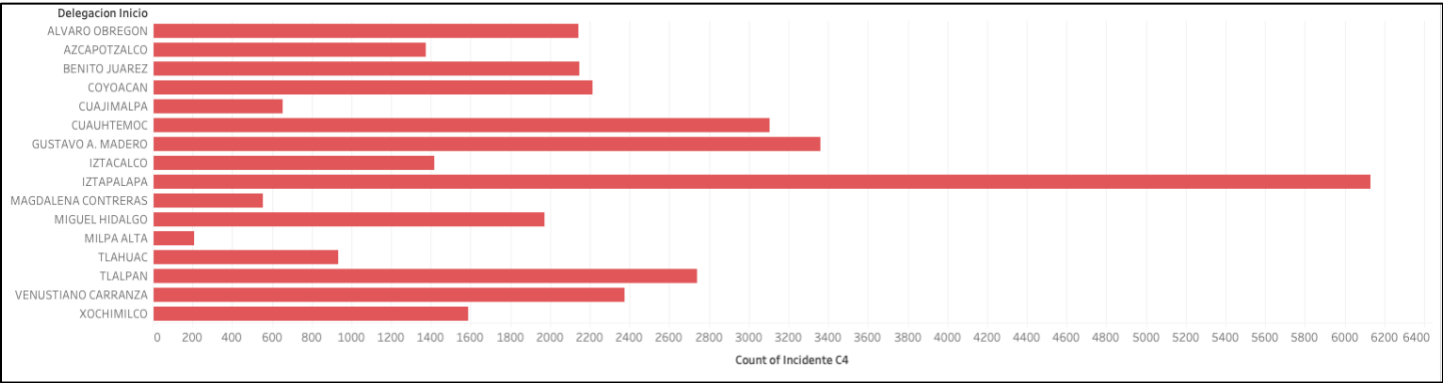
c) ¿Cuál es el mes (fecha_creacion) con la mayor cantidad de incidentes viales? En el periodo dado, fue Agosto con 9,687 incidentes viales



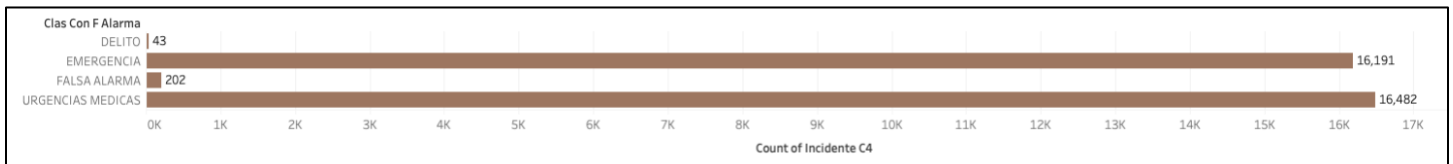
d) ¿Cuál es la hora_creacion con la mayor cantidad de incidentes viales? A las 6 de la tarde



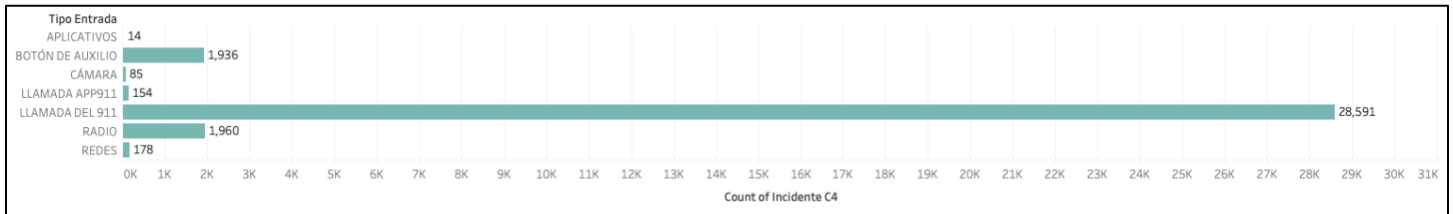
e) ¿Cuál es la delegación_inicio con la mayor cantidad de incidentes viales? Iztapalapa



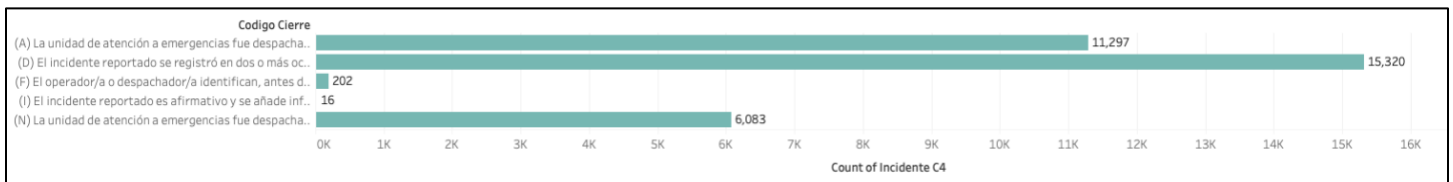
f) ¿Cuál es la clas_con_f_alarma con la mayor cantidad de incidentes viales? Urgencias medicas



g) ¿Cuál es el tipo_entrada con la mayor cantidad de incidentes viales? Llamada del 911

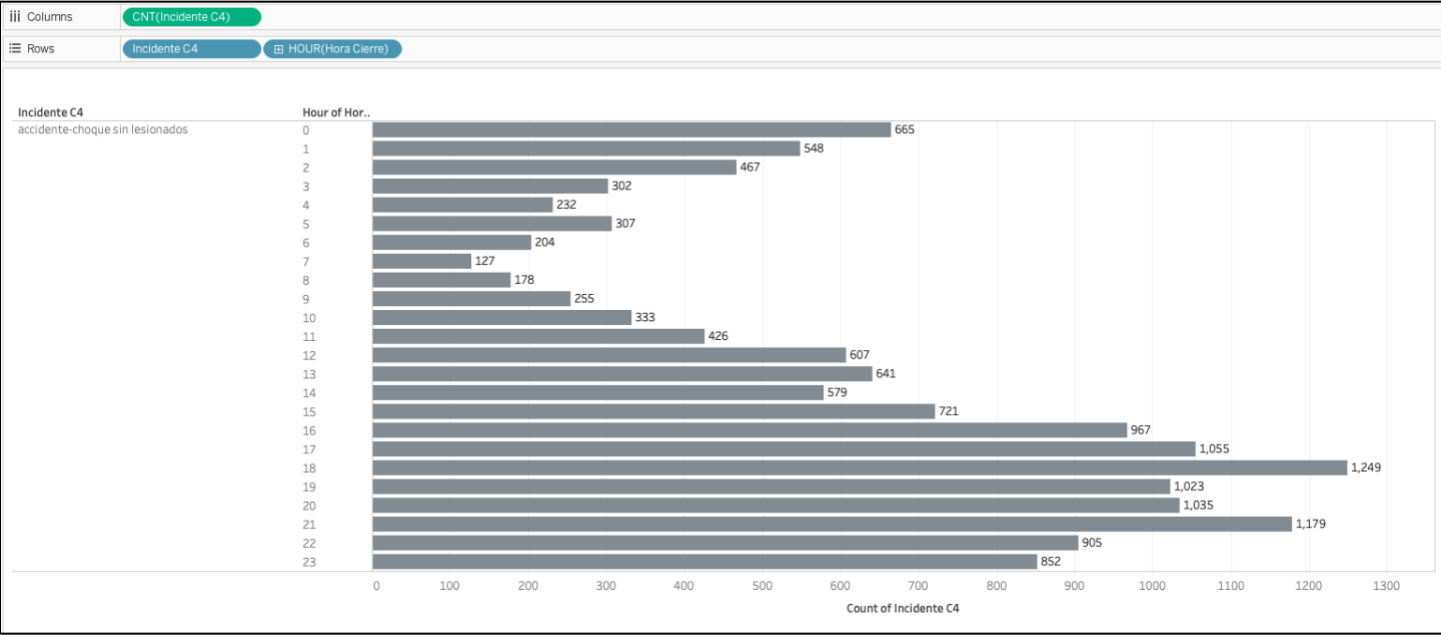


h) ¿Cuál es el codigo_cierre con la mayor cantidad de incidentes viales? (D) El incidente reportado se registró en dos o más ocasiones procediendo a mantener un único reporte (afirmativo, informativo, negativo o falso) como el identificador para el incidente.

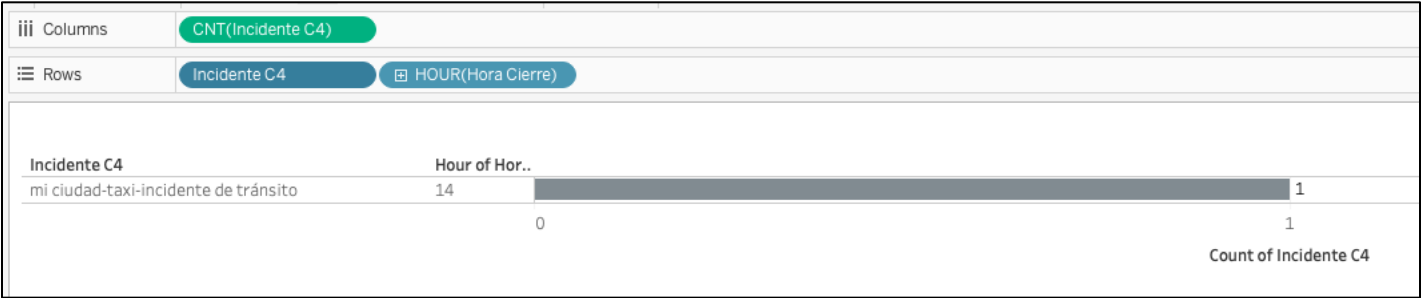


i) Considerando el incidente vial más y menos común, ¿cual es la frecuencia de ocurrencia de estos dos incidentes por hora_cierre?

Frecuencia para el incidente vial más frecuente (accidente-choque sin lesionados)

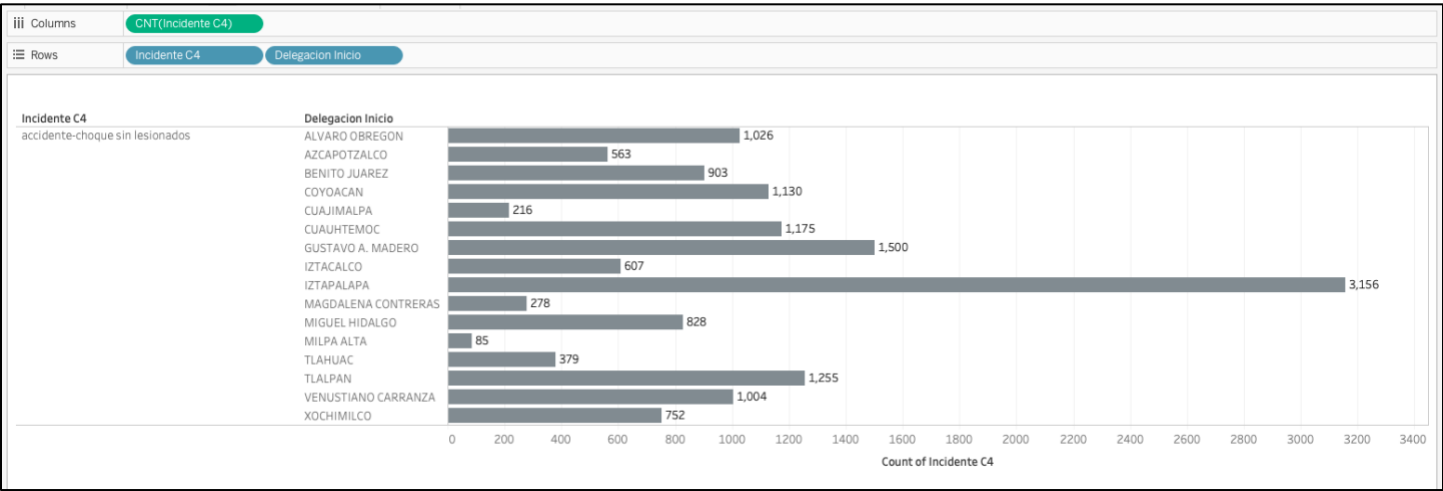


Frecuencia para el incidente vial menos frecuente (mi ciudad taxi incidente de tránsito), solo ocurrió una vez a las 2 de la tarde.

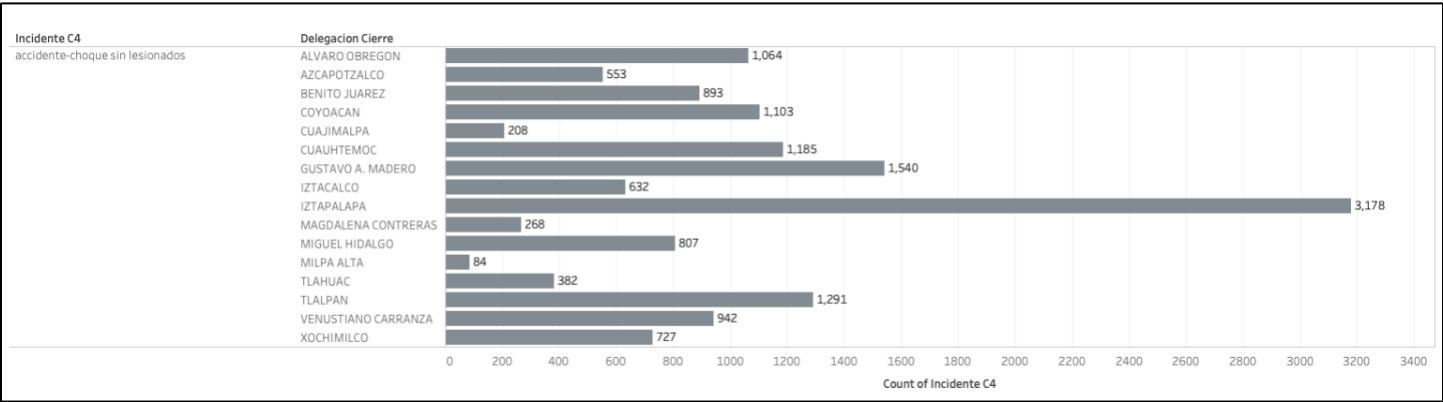


j) Considerando el incidente vial más frecuente, ¿cuál es la frecuencia de ocurrencia por delegación?

Frecuencia de ocurrencia por delegación para el incidente vial más frecuente (accidente-choque sin lesionados) por delegación de inicio.

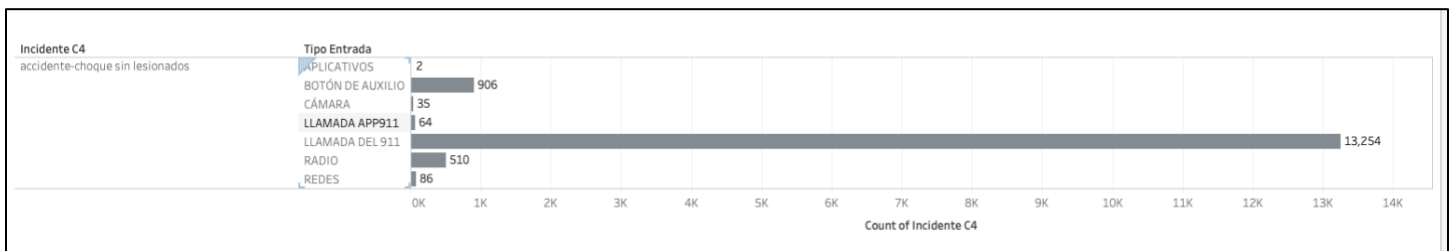


Frecuencia de ocurrencia por delegación para el incidente vial más frecuente (accidente-choque sin lesionados) por delegación de cierre.



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

Frecuencia de ocurrencia por tipo_entrada para el incidente vial más frecuente (accidente-choque sin lesionados) por delegación de inicio.



Conclusiones

Fue una práctica bastante interesante, comienzo por decir que en teoría el data set es sobre incidentes viales del 2020 en el segundo semestre del mismo año, aún encuentro datos que corresponden a los meses dentro del primer semestre del 2020, cuando se grafican datos por temporalidad la diferencia es abismal, ya que meses del primer semestre tienen muy pocas muestras a comparación con los meses del segundo semestre.

Después, volví a analizar mi base de datos, es decir como en la primera práctica, pero a diferencia de aquella vez, ahora elimine tuplas que incluían algún atributo nulo, fueron alrededor de 160 tuplas las que elimine, posteriormente esa tabla limpia la exporte a CSV, una vez exportado el archivo, lo importe a tableau, eso fue sencillo.

Finalmente familiarizarme con el software es bastante sencillo, el error que noté a lo largo de algunas preguntas, por ejemplo, cuando a partir del incidente más frecuente, graficar la hora de cierre por cada una de estas horas, hay que poner atención en lo que se nos solicita, ya que en un primer intento la gráfica que obtuve fue por cada hora de cierre obtener el incidente vial más frecuente, cuando era al revés casi casi.