

## 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

En la presente práctica se continúa con el trabajo presentado durante la práctica 1, en dónde identificamos los datos los que estábamos trabajando, analizando el significado de los mismos y la relación entre las columnas, sin embargo también pudimos identificar que algunos de estos datos presentaban problemas al momento de importarse en la base de datos o simplemente eran incoherentes entre sí, por lo que existía la necesidad de limpiar la base de datos antes de poder trabajar con ella. Es por ello que a continuación se enlistan los cambios que se realizaron durante la práctica 1 para poder manejar correctamente los datos.

- Se eliminó la última fila de los registros, debido a que presentaban datos incompletos.
- Se adecuaron los tipos de datos de las columnas para que coincidieran con los del sistema.

## Desarrollo

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 que la inconsistencia de los datos afecte la 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**

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

Procedemos a verificar cada columna:

100 %

Results		Messages
	fecha_creacion	
51	2020-08-10	
52	2020-09-28	
53	2020-06-03	
54	2020-09-05	
55	2020-08-16	
56	2020-02-18	
57	2020-05-22	
58	2020-10-12	
59	2020-07-10	
60	2020-08-22	
61	2020-07-30	
62	2020-09-14	
63	2020-07-24	
64	2020-04-21	
65	2020-06-20	
66	2020-03-06	
67	2020-09-22	
68	2020-06-14	
69	2020-05-31	
70	2020-11-13	
71	2020-06-23	
72	2020-09-25	
73	2020-09-02	
74	2020-10-21	
75	2020-08-13	
76	2020-10-01	
77	2020-07-21	
78	2020-09-08	
79	2020-07-01	
80	2020-07-13	

✓ Query executed successfully.

No se encuentran datos inconsistentes en la fecha de creación.

100 %	
Results	Messages
cantidadNulos	FechaCreacion

No se encuentran valores nulos en la fecha de creación.

Results	Messages
codigo_cierre	
1	(V) La unidad de atención a emergencias fue despachada, llegó al lugar de los hechos y confirmó la emergencia reportada
2	(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
3	(I) El incidente reportado es afirmativo y se añade información adicional al evento
4	(F) El operador/a o despachador/a identificó, antes de dar respuesta a la emergencia, que ésta es falsa. O al ser despachada una unidad de atención a emergencias en el lugar de los hechos se percata que el incidente no corresponde al reportado inicialmente
5	(N) La unidad de atención a emergencias fue despachada, llegó al lugar de los hechos, pero en el sitio del evento nadie solicitó el apoyo de la unidad

Las opciones de código de cierre son distintas y coherentes.

100 %	
Results	Messages
codigo_cierre	

No se localizan valores nulos

100 %

Results		Messages
	folio	
33053	IZ/201020/05477	
33054	IZ/201021/00070	
33055	IZ/201021/03808	
33056	IZ/201109/02209	
33057	IZ/201109/04681	
33058	IZ/201110/03929	
33059	IZ/201110/05679	
33060	IZ/201111/01251	
33061	IZ/201111/01375	
33062	IZ/201111/03863	
33063	IZ/201111/03967	
33064	IZ/201111/04002	
33065	IZ/201111/04097	
33066	IZ/201111/05205	
33067	IZ/201111/05585	
33068	IZ/201112/03956	
33069	IZ/201112/04240	
33070	IZ/201112/04608	
33071	IZ/201112/05331	

Query executed successfully

Se desglosan todos los registros del folio con lo que en total son 33071

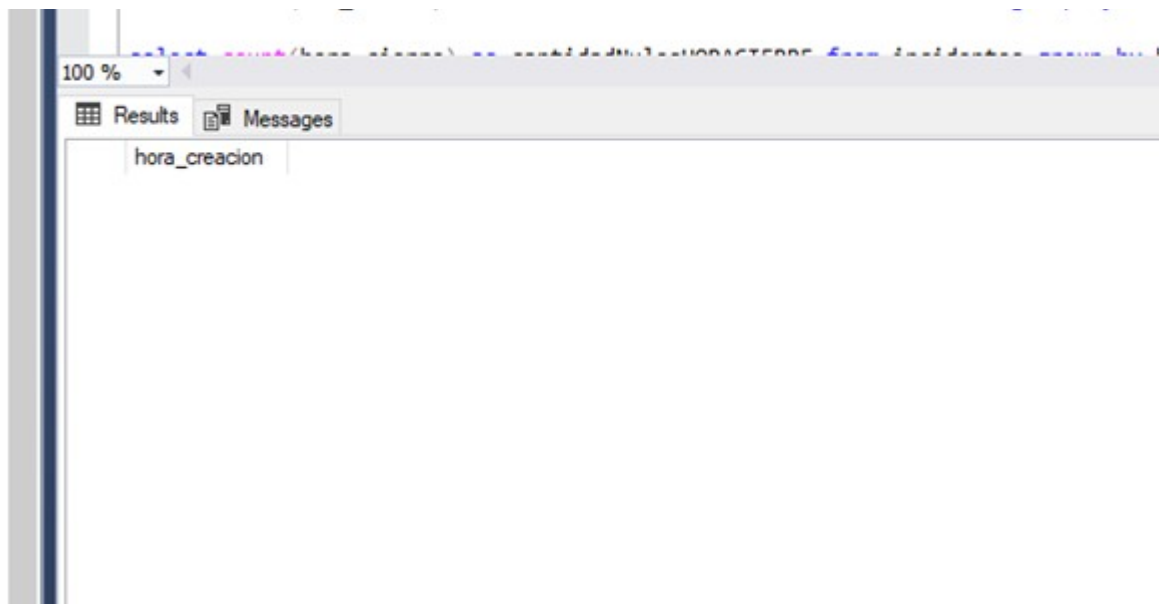
```
select count(mes_cierre) as cantidadNulosMESCIERRE from incidentes group by mes_c
```

```
select count(hora_cierre) as cantidadNulosHORACIERRE from incidentes group by hora
```

100 %

Results		Messages
	folio	

Se rastrean los folios que podrían ser inconsistentes, pero no hay llaves primarias nulas.

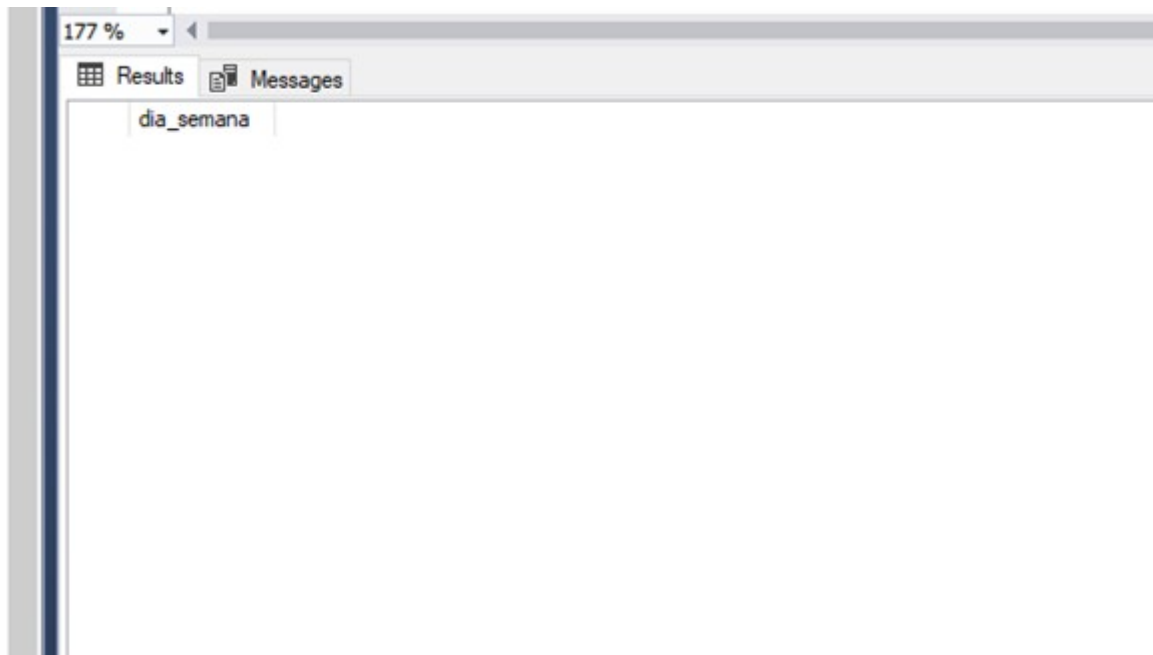


No existen valores nulos en la hora de creación.

A screenshot of the SQL Server Enterprise Manager interface showing a query result. The 'Results' tab is active, displaying a table with two columns: an implicit ID column and 'hora\_creacion'. The data is sorted in ascending order by the 'hora\_creacion' values. A status bar at the bottom indicates 'Query executed successfully.'

	hora_creacion
26048	23:59:03.0000000
26049	23:59:04.0000000
26050	23:59:11.0000000
26051	23:59:13.0000000
26052	23:59:14.0000000
26053	23:59:15.0000000
26054	23:59:16.0000000
26055	23:59:18.0000000
26056	23:59:22.0000000
26057	23:59:27.0000000
26058	23:59:29.0000000
26059	23:59:30.0000000
26060	23:59:36.0000000
26061	23:59:45.0000000
26062	23:59:47.0000000
26063	23:59:49.0000000
26064	23:59:54.0000000
26065	23:59:55.0000000
26066	23:59:57.0000000

Ordenamos la hora de creación de manera ascendente para revisar de mejor manera y no se encuentran valores inconsistentes.



Ejecutamos la consulta para corroborar que no existan valores nulos en días de la semana.

A screenshot of a SQL query results window. The window has a title bar with a zoom level of 177%. Below the title bar, there are two tabs: 'Results' and 'Messages'. The 'Results' tab is active, showing a table with two columns: an index and 'dia\_semana'. The table contains seven rows of data.

	dia_semana
1	Domingo
2	Jueves
3	Lunes
4	Martes
5	Miércoles
6	Sábado
7	Viernes

Solo existen 7 días de la semana que corresponden a los que conocemos por lo tanto son datos consistentes.

177 %

Results Messages

año_cierre
------------

No existen valores nulos en el año de cierre.

177 %

Results Messages

año_cierre
1 2020

Sólo nos arroja que todos los registros se encuentran en el año 2020 por lo que no es inconsistente.

177 %

Results Messages

mes_cierre
------------

No existen valores nulos en el mes de cierre.

177 %

Results Messages

	mes_cierre
1	Julio
2	Octubre
3	Septiembre
4	Agosto
5	Junio
6	Noviembre

Se localizan 6 meses distintos de los que existen en un año por lo que es coherente.

177 %

Results Messages

hora_cierre
-------------

En la hora de cierre no existen valores nulos.

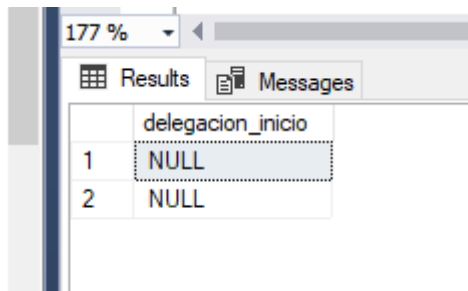
177 %

Results Messages

	hora_cierre
2...	23:59:30.0000000
2...	23:59:31.0000000
2...	23:59:35.0000000
2...	23:59:37.0000000
2...	23:59:39.0000000
2...	23:59:42.0000000
2...	23:59:43.0000000
2...	23:59:49.0000000
2...	23:59:50.0000000
2...	23:59:55.0000000
2...	23:59:59.0000000

En la hora de cierre tenemos un intervalo de 00:00:00 a 23:59:59 lo que corresponde a un día completo por lo que no se encuentran inconsistencias.






177 %

Results Messages

	delegacion_inicio
1	NULL
2	NULL

Se localizan dos registros con valores nulos.



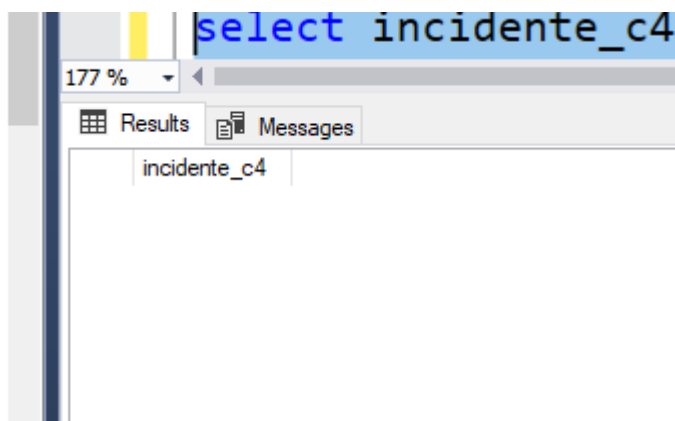
177 %

Results Messages

	delegacion_inicio
1	ALVARO OBREGON
2	AZCAPOTZALCO
3	BENITO JUAREZ
4	COYOACAN
5	CUAJIMALPA
6	CUAUHTEMOC
7	GUSTAVO A. MADERO
8	IZTACALCO
9	IZTAPALAPA
10	MAGDALENA CONTRERAS
11	MIGUEL HIDALGO
12	MILPA ALTA
13	NULL
14	TLAHUAC
15	TLALPAN
16	VENUSTIANO CARRANZA
17	XOCHIMILCO

Query executed successfully.

Se detectan inconsistencias ya que no existen delegaciones que se llamen NULL



177 %

Results Messages

	incidente_c4
--	--------------

No se encuentran valores nulos.

Results		Messages
	incidente_c4	
1	accidente-choque con lesionados	
2	accidente-choque con prensados	
3	accidente-choque sin lesionados	
4	accidente-ciclista	
5	accidente-motociclista	
6	accidente-otros	
7	accidente-persona atrapada / desbarrancada	
8	accidente-vehículo atrapado-varado	
9	accidente-vehículo desbarrancado	
10	accidente-volcadura	
11	cadáver-accidente automovilístico	
12	cadáver-atropellado	
13	detención ciudadana-accidente automovilístico	
14	detención ciudadana-atropellado	
15	lesionado-atropellado	
16	mi ciudad-calle-incidente de tránsito	
17	mi ciudad-taxi-incidente de tránsito	
18	sismo-choque con lesionados	
19	sismo-persona atropellada	

Query executed successfully

Tenemos 19 distintos tipos de incidentes por lo que no es inconsistente.

null)

null)

177 %

Results		Messages
	latitud	

No se localizan valores nulos en la latitud.

177 %

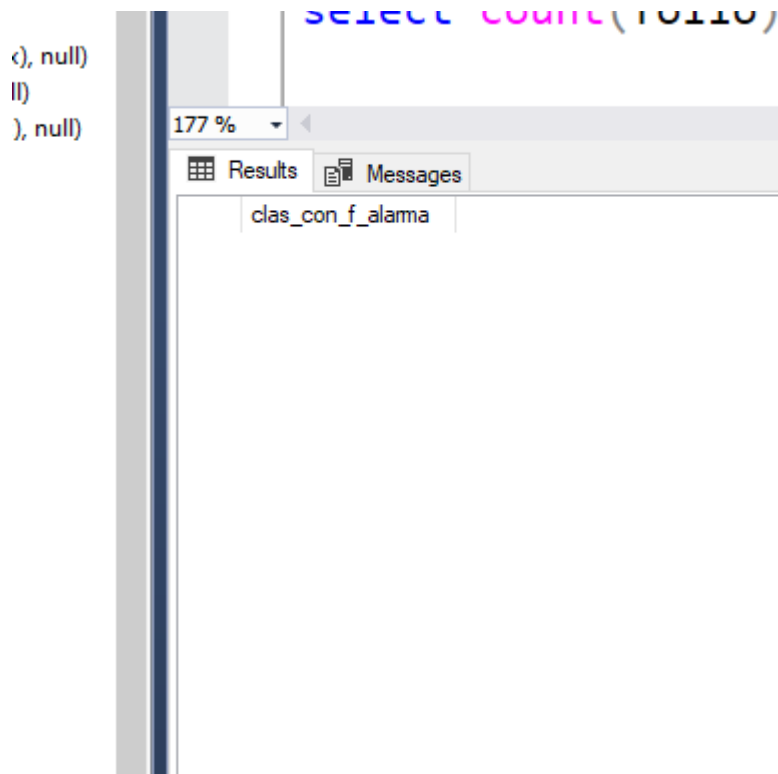
Results Messages

	latitud
1...	19.56602
1...	19.56606
1...	19.56611
1...	19.56662
1...	19.56699
1...	19.56709
1...	19.56864
1...	19.56885
1...	19.56889
1...	19.56894
1...	19.57023
1...	19.57025
1...	19.57096
1...	19.57105
1...	19.57109
1...	19.574963
1...	19.57514
1...	19.57534
1...	19.57671

Query executed successfully

;

No se identifican inconsistencias en la latitud.



No se visualizan valores nulos en clas\_con\_f\_alarma.

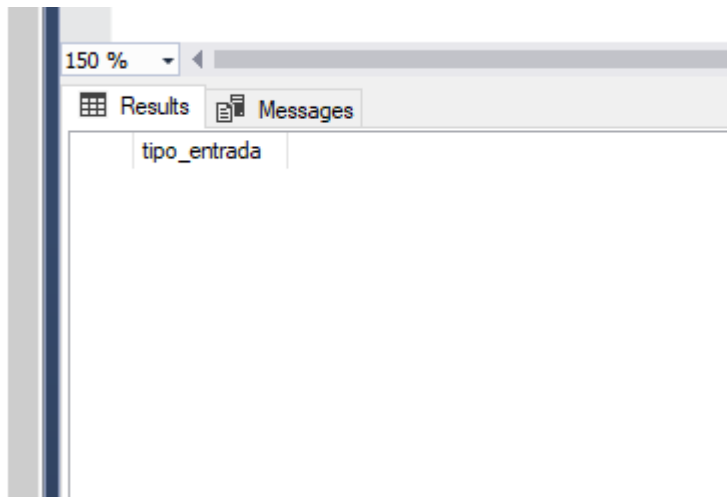
)

177 %

Results Messages

	clas_con_f_alarma	cant
1	DELITO	43
2	EMERGENCIA	16246
3	FALSA ALARMA	204
4	URGENCIAS MEDICAS	16578

Son 4 distintos tipos de alarmas por lo que son valores consistentes.



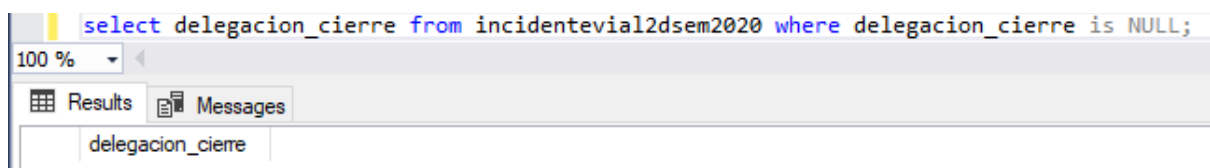
No se localizan tipos de entrada nulas.

A screenshot of a SQL query results window. The window has a zoom level of 150%. The 'Results' tab is active, showing a table with 7 rows of data. The columns are 'tipo\_entrada' and '(No column name)'. The data is as follows:

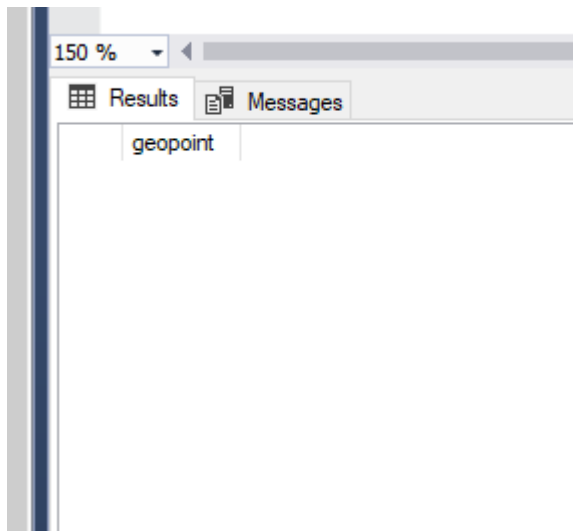
	tipo_entrada	(No column name)
1	BOTÓN DE AUXILIO	1944
2	LLAMADA APP911	155
3	APLICATIVOS	14
4	RADIO	1974
5	LLAMADA DEL 911	28721
6	REDES	178
7	CÁMARA	85

Below the table, there is a button that says 'Click to select the whole column'.

Tenemos 7 distintos tipos de entradas coherentes.



No se detectan valores nulos en la delegación de cierre.



No existen valores nulos en geopoint.

A screenshot of a software interface showing a zoom level of 150%. The 'Results' tab is active, displaying a table with 17 rows of data. The table has three columns: an index, 'geopoint', and 'conteo'. The first row is highlighted with a dashed border.

	geopoint	conteo
1	19.30431996,-99.08714016	59
2	19.37168001,-99.08024004	58
3	19.36316997,-99.05739984	55
4	19.35663003,-99.08655984	47
5	19.24151,-99.147148	38
6	19.30545999,-99.20580984	35
7	19.37941002,-99.09559008	35
8	19.29339,-99.10607004	34
9	19.29272004,-99.12555	32
10	19.347021,-99.180646	32
11	19.35257004,-99.01496988	31
12	19.256075,-99.178482	31
13	19.43129997,-99.20873016	30
14	19.3446,-99.06129	27
15	19.38306996,-99.07611984	27
16	19.385469,-99.035416	25
17	19.36311003,-99.05676984	25

Identificamos distintos puntos de localización, se puede observar puntos repetidos pero existe la posibilidad de que exista accidentes en los mismos puntos en tiempos distintos por lo que no se considera inconsistente.

150 %

Results Messages

mes
-----

Podemos identificar que no existen valores nulos.

150 %

Results Messages

	mes
1	11
2	10
3	9
4	8
5	7
6	6

Tenemos registros de 6 meses distintos a partir de junio a noviembre por lo que no son inconsistentes.

150 %

Results Messages

longitud
----------

No tenemos valores nulos en la columna de longitud.

150 %

Results Messages

	longitud	conteo
1	-99.08714	59
2	-99.08024	58
3	-99.0574	55
4	-99.08656	47
5	-99.2391	42
6	-99.12555	39
7	-99.147148	38
8	-99.10607	37
9	-99.20581	35
10	-99.09559	35
11	-99.07612	34
12	-99.01497	34
13	-99.180646	32
14	-99.178482	31
15	-99.20873	30
16	-99.07818	30
17	-99.13746	30
18	-99.05677	28
19	-99.06129	27

Podemos observar que tenemos valores repetidos en longitud, pero puede ocurrir más de un accidente en la misma longitud en diferentes tiempos por lo que no se considera inconsistente.

#### LIMPIEZA DE DATOS.

Se muestra todos los campos de los valores nulos identificados, en los atributos delegación inicio y delegación cierre, corresponden a las misma tuplas de información inconsistentes,



por lo que tenemos un total de 2 tuplas con inconsistencia de datos y se muestran a continuación:

The first screenshot shows a table with columns: folio, fecha\_creacion, hora\_creacion, dia\_semana, codigo\_cieme, afo\_cieme, mes\_cieme, hora\_cieme, delegacion\_inicio, incidente\_cif, latitud, longitud, clas\_con\_f\_alama, tipo\_entrada. It contains two rows:

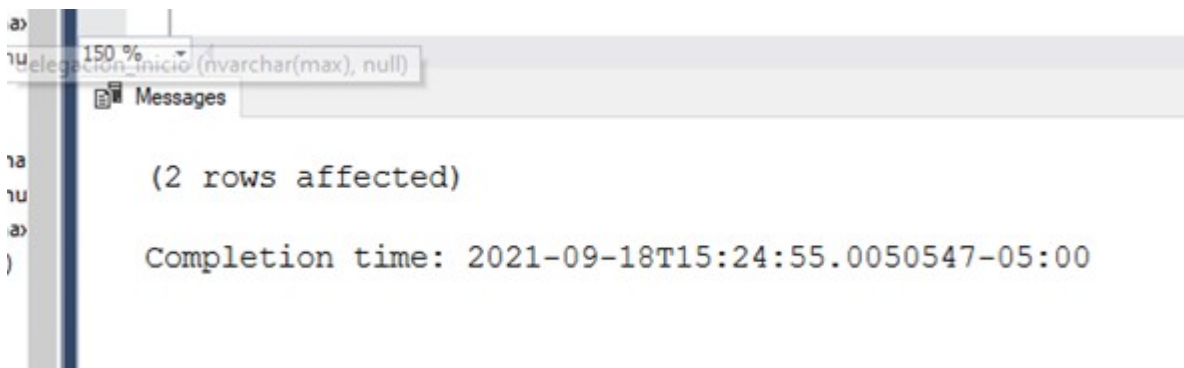
folio	fecha_creacion	hora_creacion	dia_semana	codigo_cieme	afo_cieme	mes_cieme	hora_cieme	delegacion_inicio	incidente_cif	latitud	longitud	clas_con_f_alama	tipo_entrada
AO-200625-01632	2020-08-20	08:51:31.0000000	Juernes	(f) El operador/a o despachador/a identifican a...	2020	Agosto	15:22:54.0000000	NULL	accidente-choque en lesionados	19.36997	-99.29422	FALSA ALARMA	RADIO
CS-200625-01619	2020-08-26	08:49:44.0000000	Viernes	(f) El operador/a o despachador/a identifican a...	2020	Junio	08:52:27.0000000	NULL	accidente-motocicleta	19.423016	-99.085213	FALSA ALARMA	LLAMADA DEL 911

The second screenshot shows a similar table with columns: hora\_creacion, dia\_semana, codigo\_cieme, afo\_cieme, mes\_cieme, hora\_cieme, delegacion\_inicio, incidente\_cif, latitud, longitud, clas\_con\_f\_alama, tipo\_entrada, delegacion\_cieme, geopoint, mes. It contains two rows:

hora_creacion	dia_semana	codigo_cieme	afo_cieme	mes_cieme	hora_cieme	delegacion_inicio	incidente_cif	latitud	longitud	clas_con_f_alama	tipo_entrada	delegacion_cieme	geopoint	mes
08:51:31.0000000	Juernes	(f) El operador/a o despachador/a identifican a...	2020	Agosto	15:22:54.0000000	NULL	accidente-choque en lesionados	19.36997	-99.29422	FALSA ALARMA	RADIO	NULL	19.36997001;-99.29422008	8
08:49:44.0000000	Viernes	(f) El operador/a o despachador/a identifican a...	2020	Junio	08:52:27.0000000	NULL	accidente-motocicleta	19.423016	-99.085213	FALSA ALARMA	LLAMADA DEL 911	NULL	19.423016;-99.085213	6

Ambas imágenes corresponden a las 2 tuplas.

Se procede a eliminar las tuplas.



Se muestra la consulta después de la primera limpieza.

The screenshot shows a table with columns: folio, fecha\_creacion, hora\_creacion, dia\_semana, codigo\_cieme, afo\_cieme, mes\_cieme, hora\_cieme, delegacion\_inicio, incidente\_cif, latitud, longitud, clas\_con\_f\_alama, tipo\_entrada, delegacion\_cieme, geopoint, mes. It contains one row:

folio	fecha_creacion	hora_creacion	dia_semana	codigo_cieme	afo_cieme	mes_cieme	hora_cieme	delegacion_inicio	incidente_cif	latitud	longitud	clas_con_f_alama	tipo_entrada	delegacion_cieme	geopoint	mes

Se muestran los registros totales después de la limpieza.

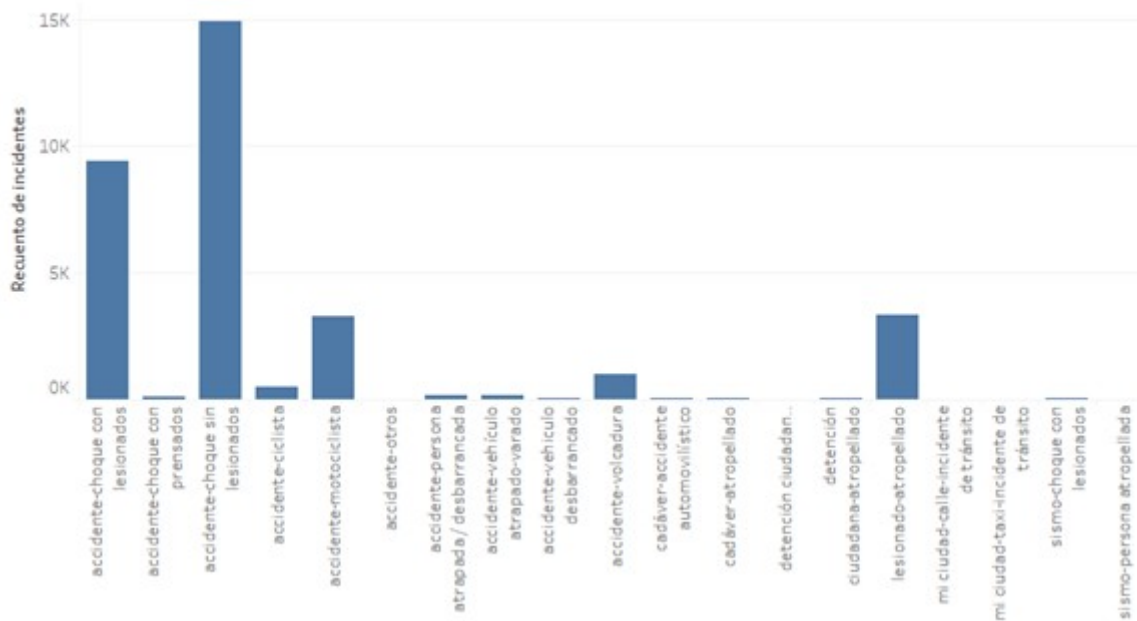
The screenshot shows a table with columns: despuesDeLimpieza. It contains one row:

despuesDeLimpieza
33069

**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):

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

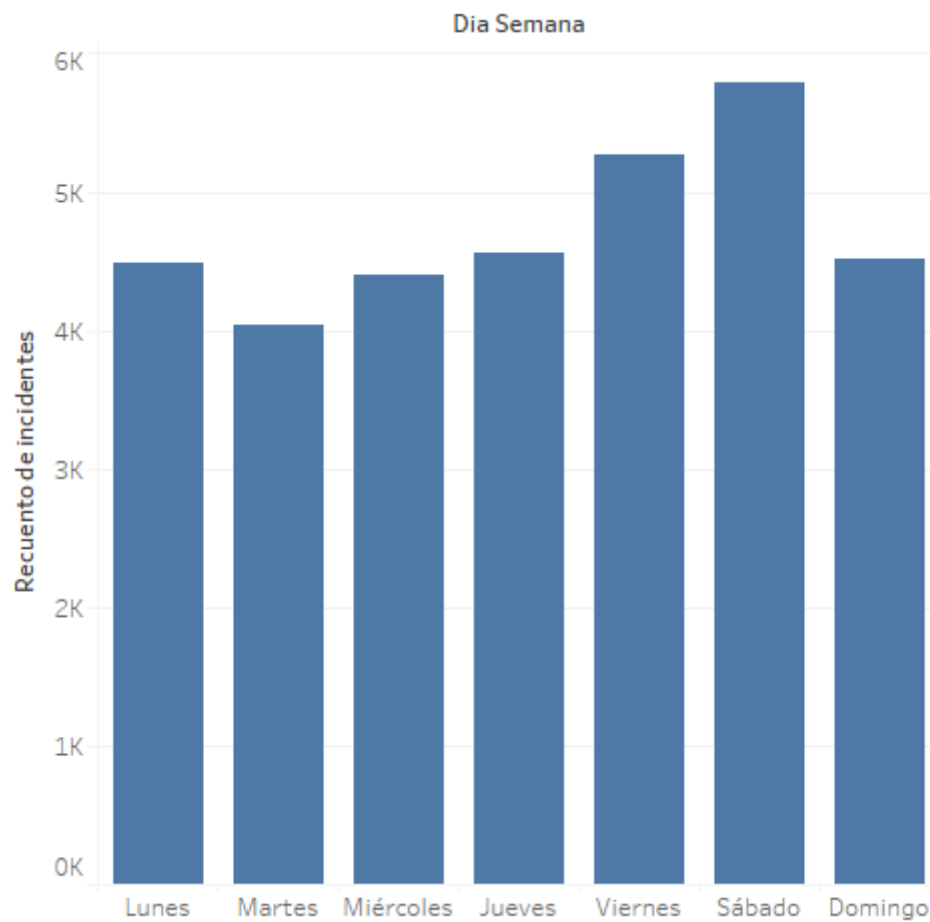
	incidente_c4	cant
1	mi ciudad-taxi-incidente de tránsito	1
2	sismo-persona atropellada	3
3	mi ciudad-calle-incidente de tránsito	9
4	detención ciudadana-accidente automovilístico	15
5	accidente-otros	18
6	detención ciudadana-atropellado	28
7	sismo-choque con lesionados	32
8	accidente-vehículo desbarrancado	33
9	cadáver-atropellado	37
10	cadáver-accidente automovilístico	42
11	accidente-choque con prensados	101
12	accidente-vehículo atrapado-varado	142
13	accidente-persona atrapada / desbarrancada	143
14	accidente-ciclista	512
15	accidente-volcadura	988
16	accidente-motociclista	3302
17	lesionado-atropellado	3358
18	accidente-choque con lesionados	9401
19	accidente-choque sin lesionados	14904



El mas frecuente es: accidente choque sin lesionados.

El menos frecuente es: Mi ciudad taxi incidente de transito.

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

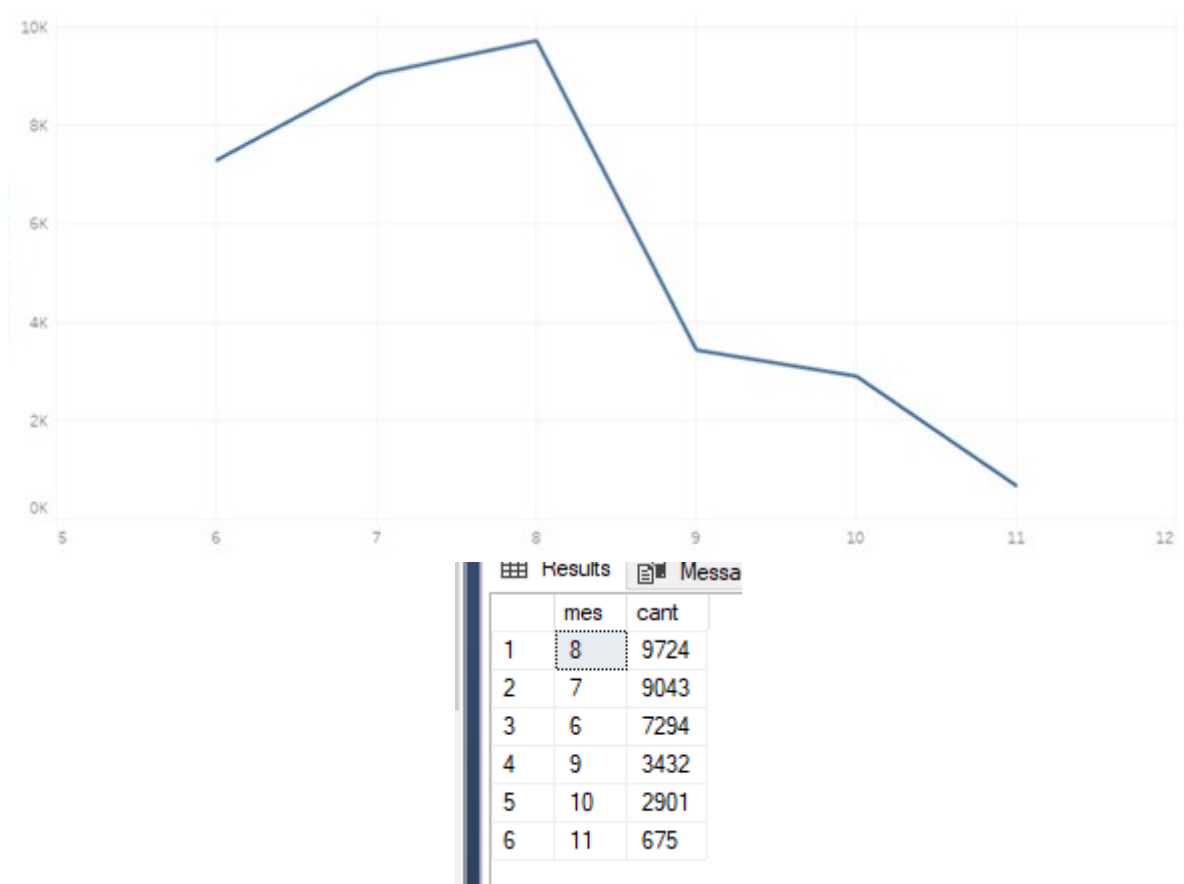


150 %

Results		Messages
	dia_semana	cant
1	Sábado	5789
2	Viernes	5274
3	Jueves	4557
4	Domingo	4522
5	Lunes	4485
6	Miércoles	4403
7	Martes	4039

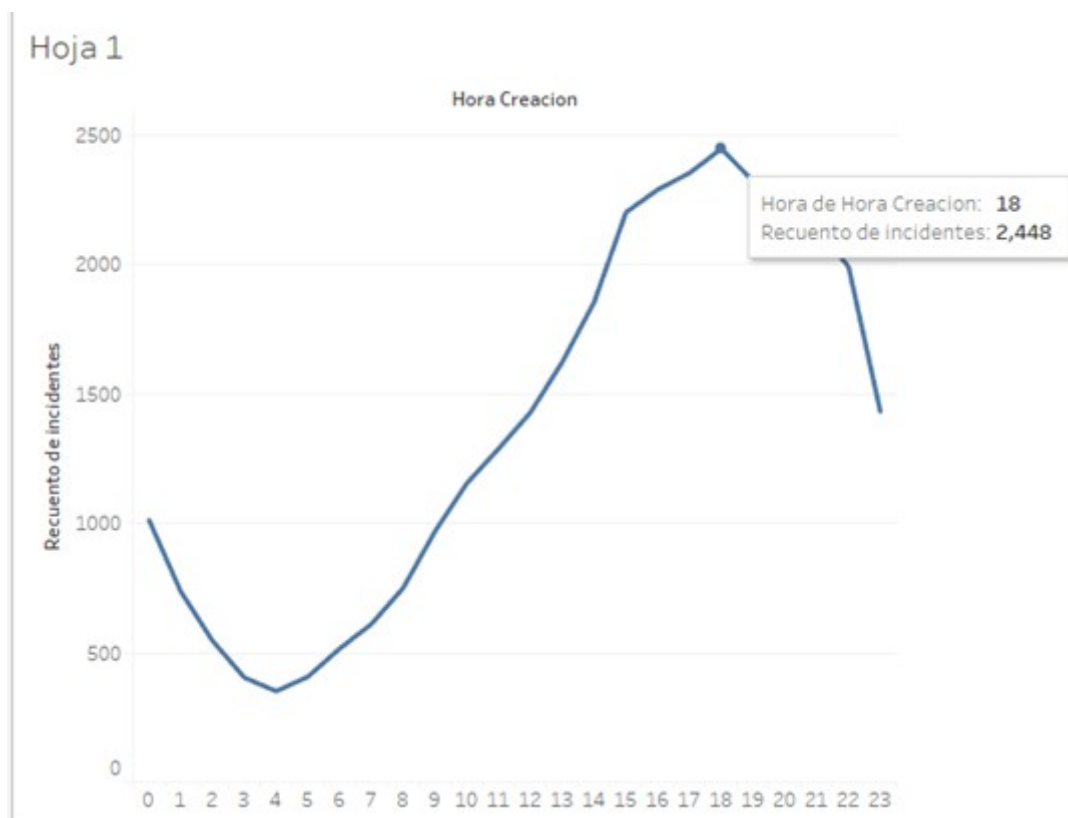
El día sábado

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

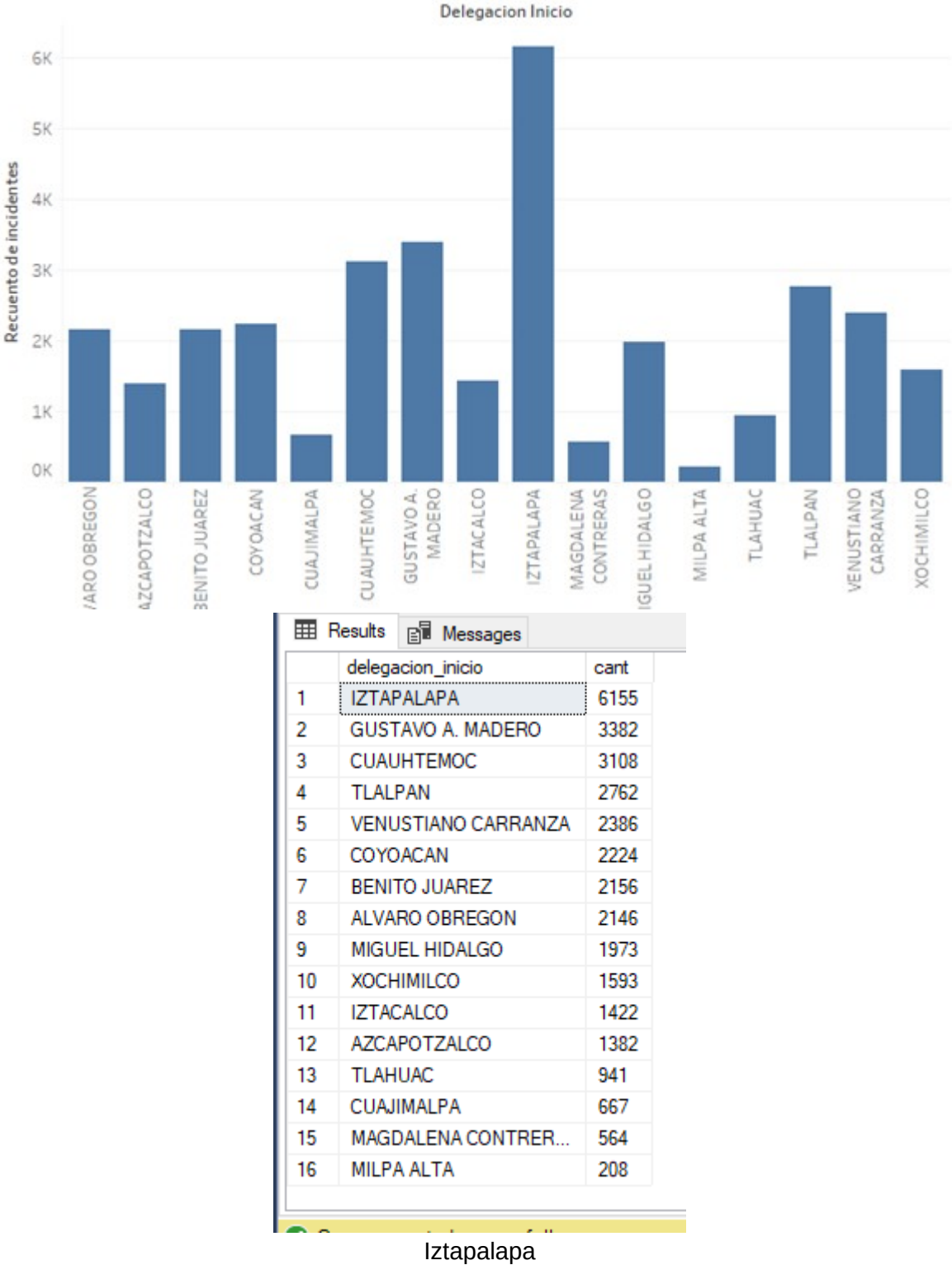


Agosto

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

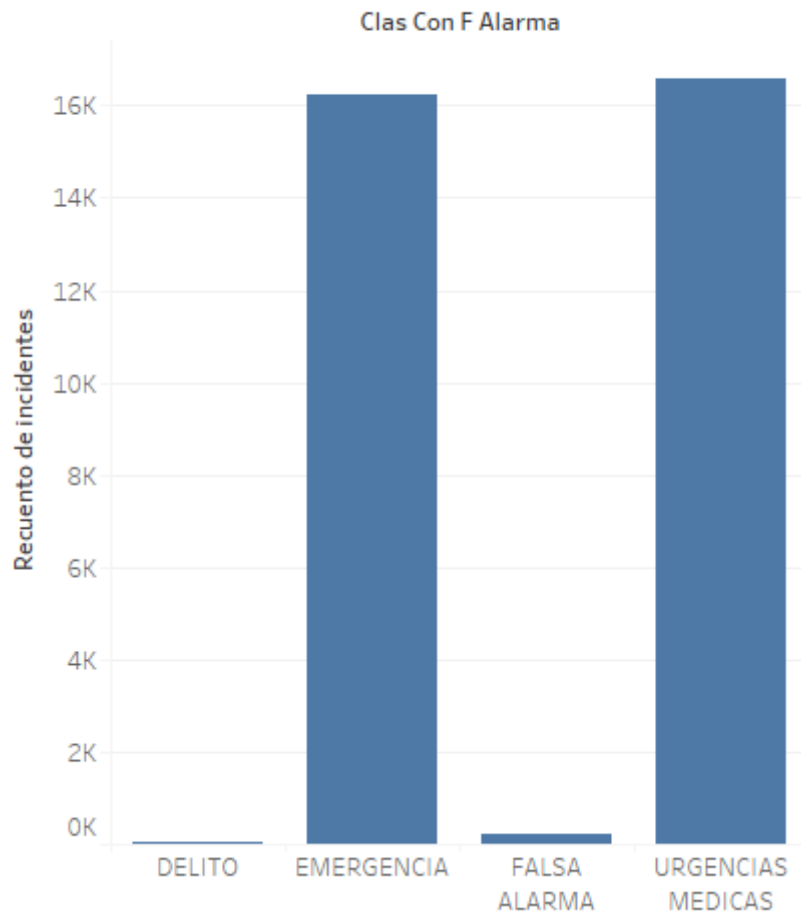


E. ¿Cuál es la **delegación\_inicio** con la mayor cantidad de incidentes viales?



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

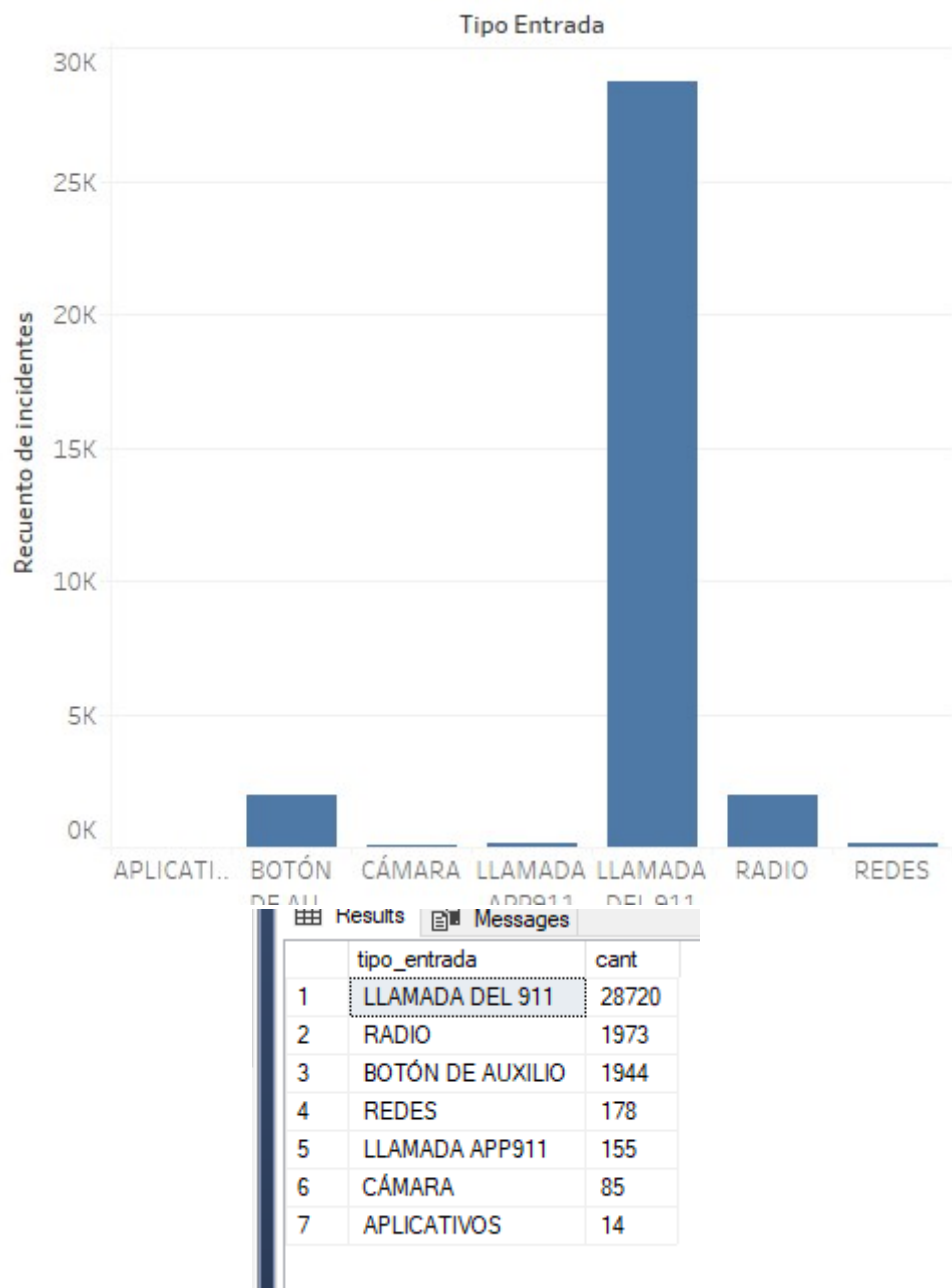
## Hoja 1



Results			Messages	
	clas_con_f_alarma	cant		
1	URGENCIAS MEDICAS	16578		
2	EMERGENCIA	16246		
3	FALSA ALARMA	202		
4	DELITO	43		

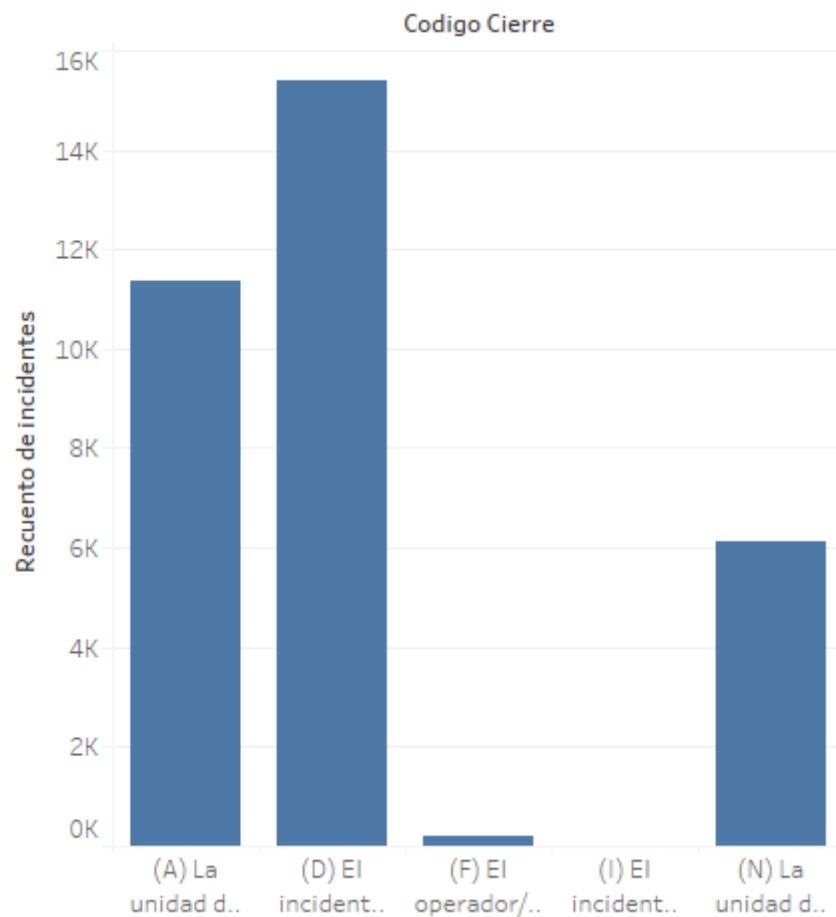
Urgencias médicas

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



Llamada del 911

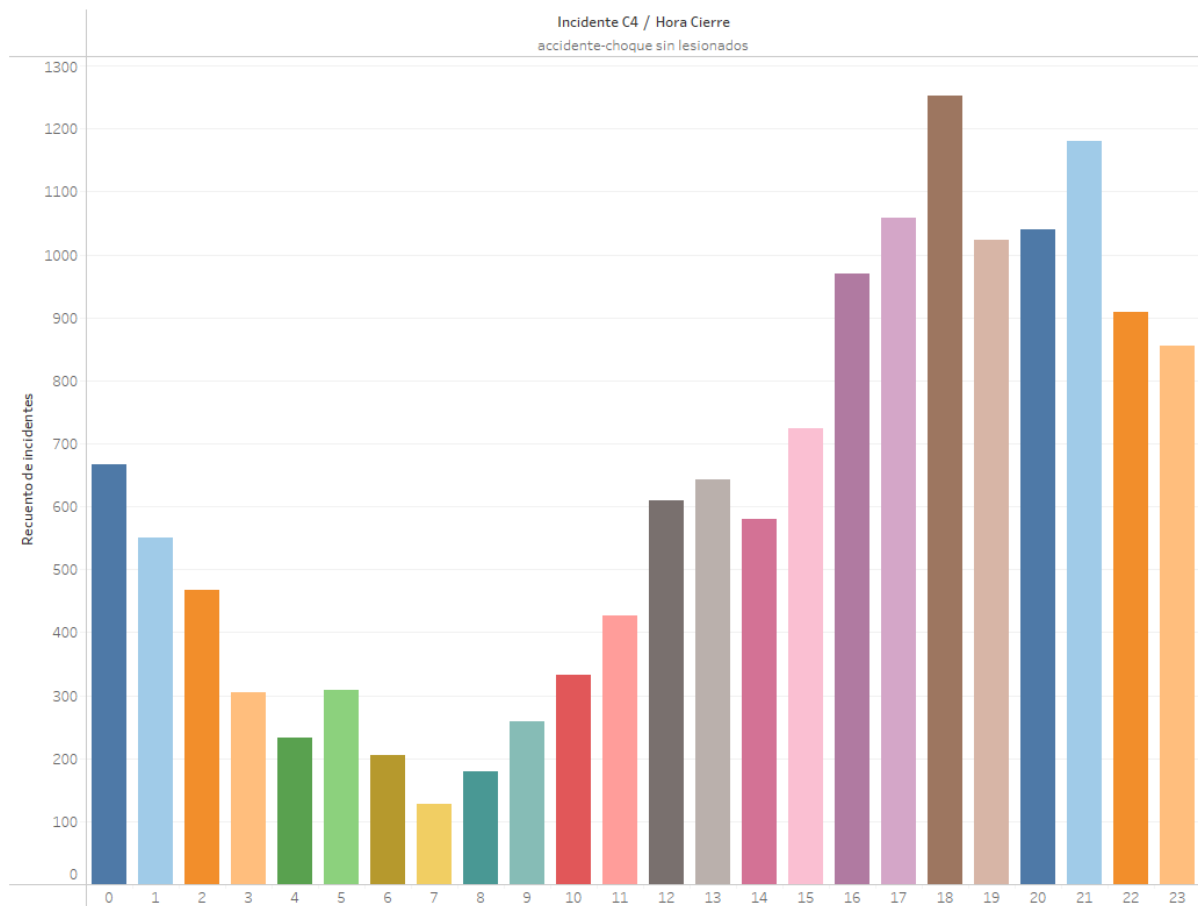
H. ¿Cuál es el **codigo\_cierre** con la mayor cantidad de incidentes viales?



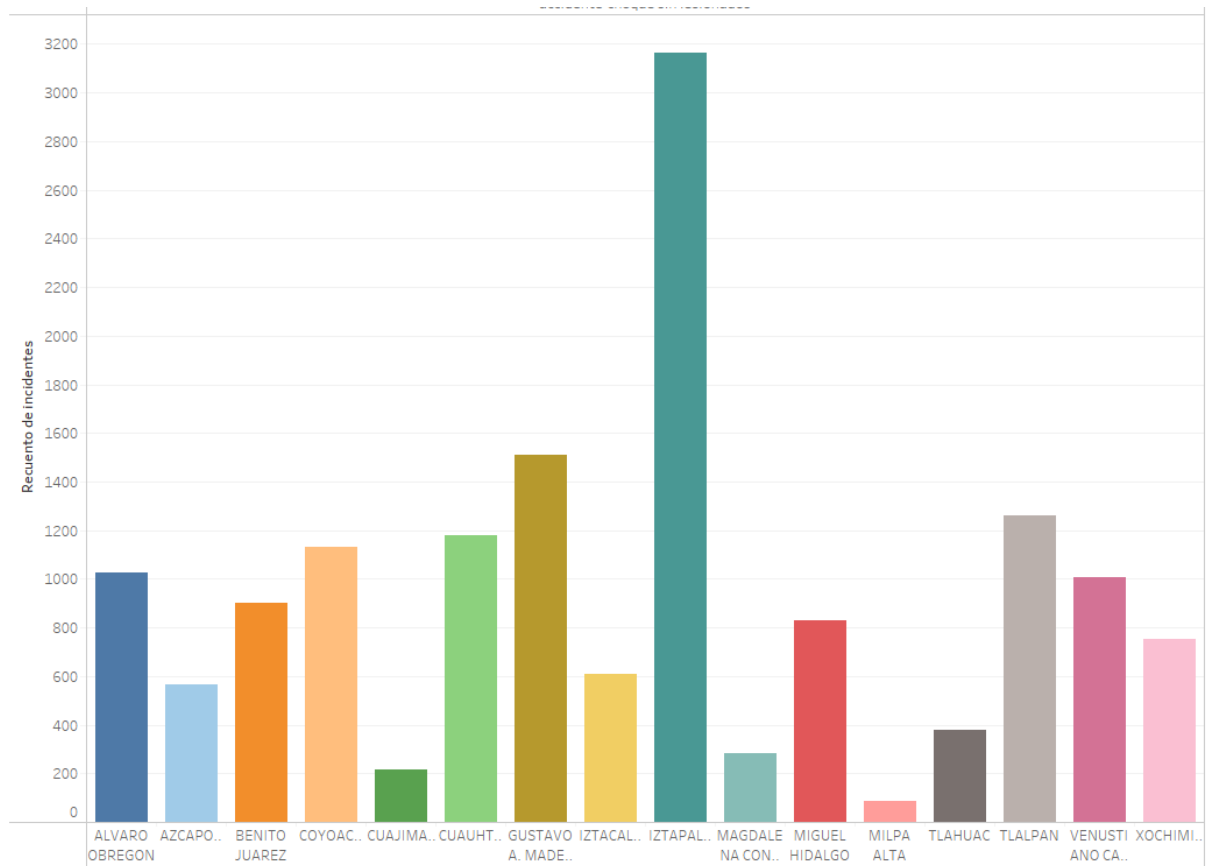
Results		Messages
	codigo_cierre	cant
1	(D) El incidente reportado se registró en dos o más ocasiones procediendo a mantener un único reporte (afirmativo, informativo, negativo o falso) como el identificador para el incidente	15386
2	(A) La unidad de atención a emergencias fue despachada, llegó al lugar de los hechos y confirmó la emergencia reportada	11356
3	(N) La unidad de atención a emergencias fue despachada, llegó al lugar de los hechos, pero en el sitio del evento nadie solicitó el apoyo de la unidad	6109
4	(F) El operador/a o despachador/a identifican, antes de dar respuesta a la emergencia, que ésta es falsa. O al ser despachada una unidad de atención a emergencias en el lugar de...	202
5	(I) El incidente reportado es afirmativo y se añade información adicional al evento	16

- I. Considerando el incidente **vial más y menos común**, ¿cual es la frecuencia de ocurrencia de estos dos incidentes por **hora\_cierre**?

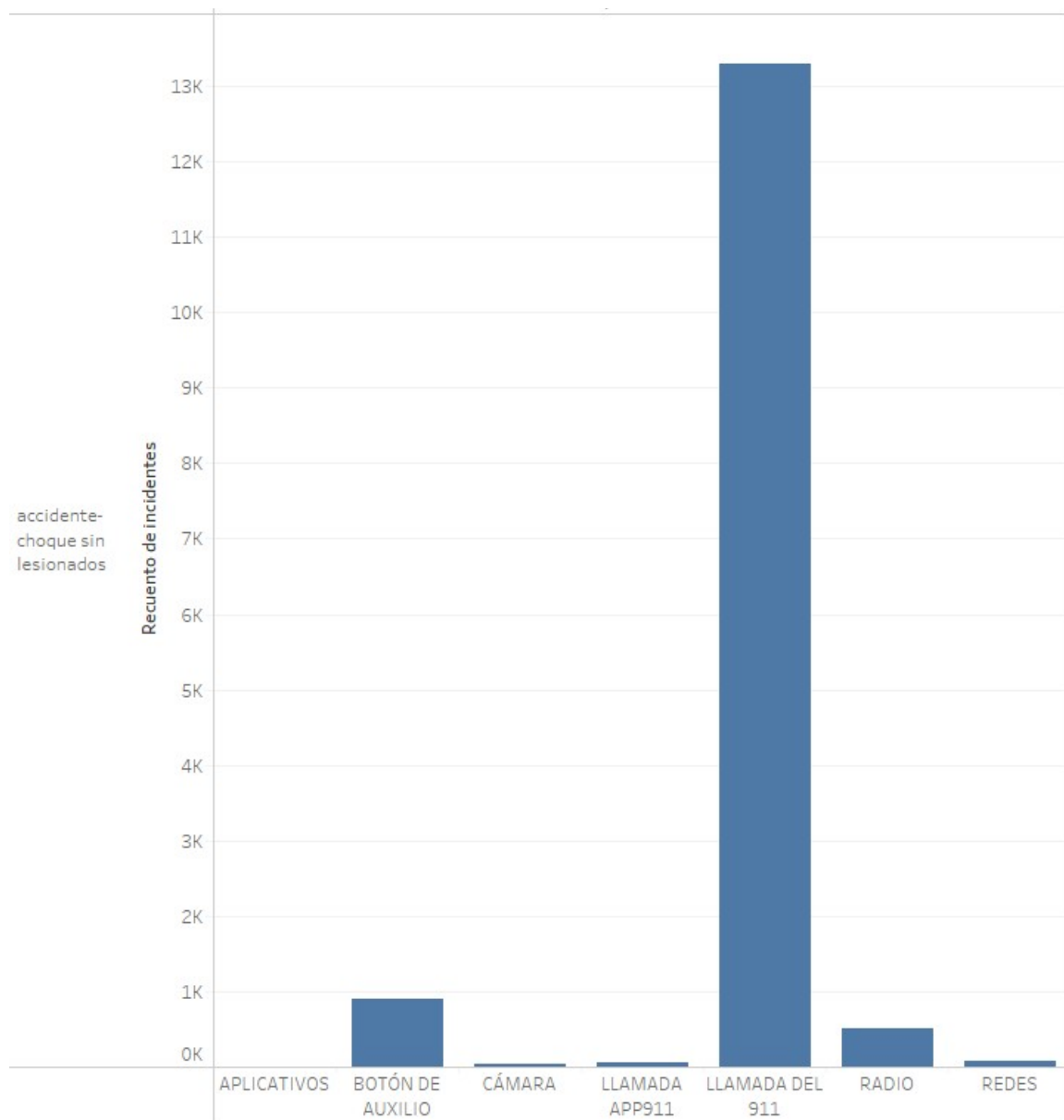




- J. Considerando el incidente vial **más frecuente**, ¿cuál es la frecuencia de ocurrencia por **delegación**?



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



## Conclusiones

Logramos encontrar valores inconsistentes como las delegaciones con nombre null, se identificaron y procedieron a eliminar.

También se observó que existen demasiados casos de incidentes duplicados, pensamos que se debe a que en el momento del incidente más de una persona reporta en incidente.

Otra observación es que la aplicación para reportar los incidentes es muy poco utilizada y la marcación rápida es muy práctica, además que cualquier usuario en uso de sus facultades físicas puede hacerlo sin tener demasiado conocimiento de un smartphone.