Incidentes Viales CDMX: Exploración de los datos

Enrique Ortiz
Carolina Acosta
Leonardo Ceja

Nota: En el notebook de GEDA todas las gráficas tienen título. Por fines de claridad, decidimos recortarlos aquí para escribirlos en cada diapositiva.

I. *Data Profiling*

- El *data profiling* se muestra sobre la información final, sobre la que estaremos trabajando durante el resto del proyecto.
- Se muestran las tablas y los 3 puntos más relevantes de los siguientes tipos de variables:
  - Variables numéricas.
  - Variables categóricas.
  - Variables de fecha.
  - Variables geoespaciales.
- Éste *data profiling* está realizado sobre la totalidad de las llamadas, con etiquetas 0 y 1. El desglose detallado por etiquetas se aprecia con mayor claridad en el GEDA.
- Para más detalle sobre el data profiling, referirse al html en la carpeta de documentos.

I. *Data Profiling* - Variables numéricas

	metric	label
0	25%	0.000000
1	75%	0.000000
2	kurtosis	0.155930
3	max	1.000000
4	mean	0.204103
5	median	0.000000
6	min	0.000000
7	prop_missings	0.000000
8	skewness	1.468308
9	stdv	0.403045
10	top1_repeated	0.000000
11	top2_repeated	1.000000
12	uniques	2.000000

Puntos más relevantes:

1. La única variable numérica con la que estaremos trabajando es “label”.
    - Esta etiqueta tiene un valor de 1 para llamadas con códigos de cierre (F - falso) o (N - negativo). (Llamadas falsas)
2. Se tiene un 20.41% de llamadas falsas (etiqueta 1).
3. Se generaron etiquetas para el 100% de los datos.

I. *Data Profiling* - Variables categóricas

	metric	hora_creacion	dia_semana	codigo_cierre	delegacion_inicio	incidente_c4	clas_con_f_alarma	tipo_entrada	delegacion_cierre	año_creacion	mes_creacion	dia_creacion	mes_creacion_str	hora_simple	espacio_del_dia
0	num_categories	105887	7	5	17	26	4	9	17	8	12.0	31.0	12	34	8
1	missings	0	0	0	0	0	0	0	0	0	0.0	0.0	0	0	0
2	top1_repeated	20:44:00	Viernes	A	IZTAPALAPA	accidente-choque sin lesionados	EMERGENCIA	LLAMADA DEL 911	IZTAPALAPA	2018	10.0	14.0	Octubre	19	6-8 p.m.
3	top2_repeated	19:16:00	Sábado	D	GUSTAVO A. MADERO	accidente-choque con lesionados	URGENCIAS MEDICAS	LLAMADA DEL 066	GUSTAVO A. MADERO	2019	8.0	15.0	Agosto	20	3-5 p.m.
4	top3_repeated	18:38:00	Jueves	N	CUAUHTEMOC	lesionado-atropellado	FALSA ALARMA	BOTÓN DE AUXILIO	CUAUHTEMOC	2017	9.0	13.0	Septiembre	18	9-11 p.m.

Puntos más relevantes:

1. No se cuenta con registros faltantes en las columnas categóricas.
2. La delegación con más llamadas es Iztapalapa, y el tipo de entrada más común es la “Llamada 911”.
3. En cuanto a temporalidad, el mes con más registros es Octubre, el día de la semana con más registros es el Viernes, y el espacio del día con más registros es de 6 a 8pm.

Todos estos detalles se aprecian con mayor claridad en el GEDA.

I. *Data Profiling* - Variables de fecha

	metric	fecha_creacion	fecha_cierre
0	mode	[2020-02-14T00:00:00.000000000]	[2020-02-14T00:00:00.000000000]
1	num_dates	2497	2496
2	max	2020-10-31 00:00:00	2020-10-31 00:00:00
3	min	2013-12-31 00:00:00	2014-01-01 00:00:00
4	mean	2017-07-12 21:28:36.224410880	2017-07-12 23:02:39.774077952
5	uniques	2497	2496
6	missings	0	0
7	top1_repeated	2020-02-14 00:00:00	2020-02-14 00:00:00
8	top2_repeated	2018-10-26 00:00:00	2017-12-02 00:00:00
9	top3_repeated	2017-12-08 00:00:00	2017-12-08 00:00:00

Puntos más relevantes:

1. La información cubre los años 2014-2020.
2. La fecha más reciente es 31/Oct/2020.
    - Originalmente se tuvieron problemas de lectura con meses y días invertidos, que se corrigieron posteriormente en la función de transformación de fechas.
3. La fecha más repetida es el 14/Feb/2020.

I. *Data Profiling* - Variables geoespaciales

	metric	latitud	longitud
0	mode	[19.30431996]	[-99.08024004]
1	max	195.303	-98.9454
2	min	19.094	-991.764
3	mean	19.3839	-99.1436
4	stdv	0.266638	2.39968
5	25%	19.3369	-99.1793
6	median	19.3841	-99.1402
7	75%	19.435	-99.096
8	kurtosis	400835	138084
9	skewness	611.376	-371.477
10	uniques	82501	78984
11	count_missings	443	435
12	prop_missings	0.0320286	0.0314502
13	top1_repeated	19.3043	-99.0802
14	top2_repeated	19.3717	-99.0871

Puntos más relevantes:

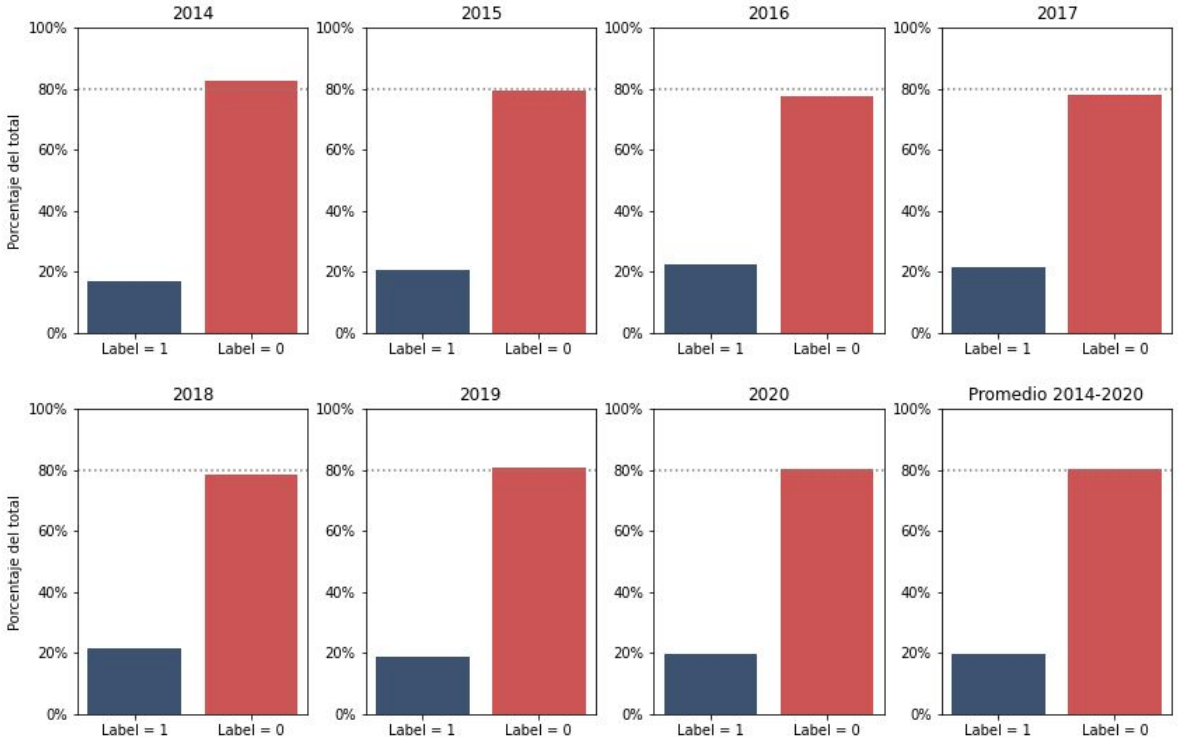
1. Se encontró aproximadamente un 3% de valores faltantes de latitud y longitud, que se trabajarán como valores nulos.
2. Se encontraron 13 registros con errores de captura (parecería un decimal movido, basándose en el promedio de los datos). Estos puntos se trabajarán como nulos.
3. La ubicación más repetida es (19.3042, -99.0802), su ubicación se muestra en el mapa incluido en el GEDA.

II. EDA / GEDA

- Las siguientes diapositivas muestran las gráficas más representativas de nuestro análisis.

- La mayoría de las gráficas se explican por sí mismas, pero se agrega un breve comentario con las conclusiones principales de cada una de ellas.

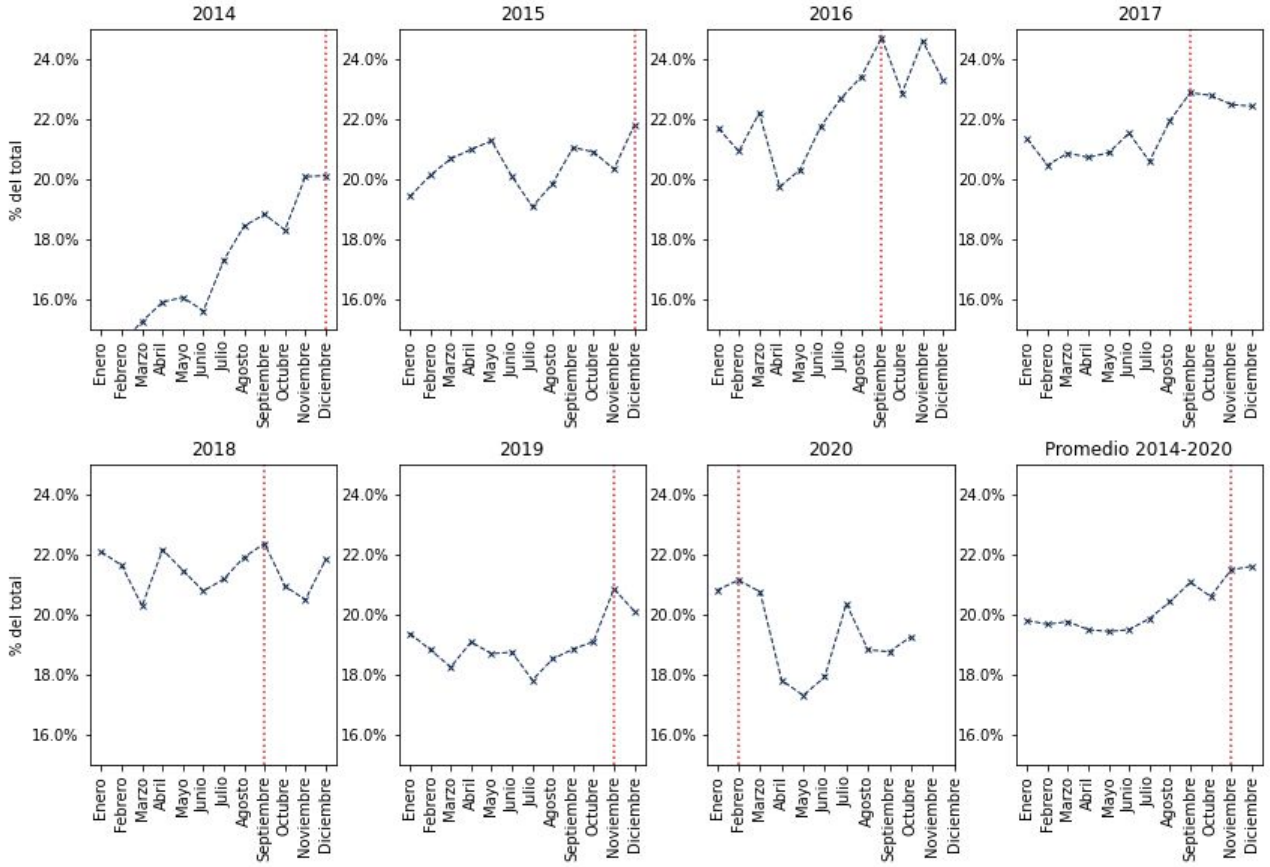
Proporción de la etiqueta en el dataset: *



Se observa una proporción de aproximadamente 20% de llamadas falsas y 80% de llamadas verdaderas a lo largo de los años.

*Nota: Label = 1 implica llamadas falsas o negativas y label = 0 todas las demás.

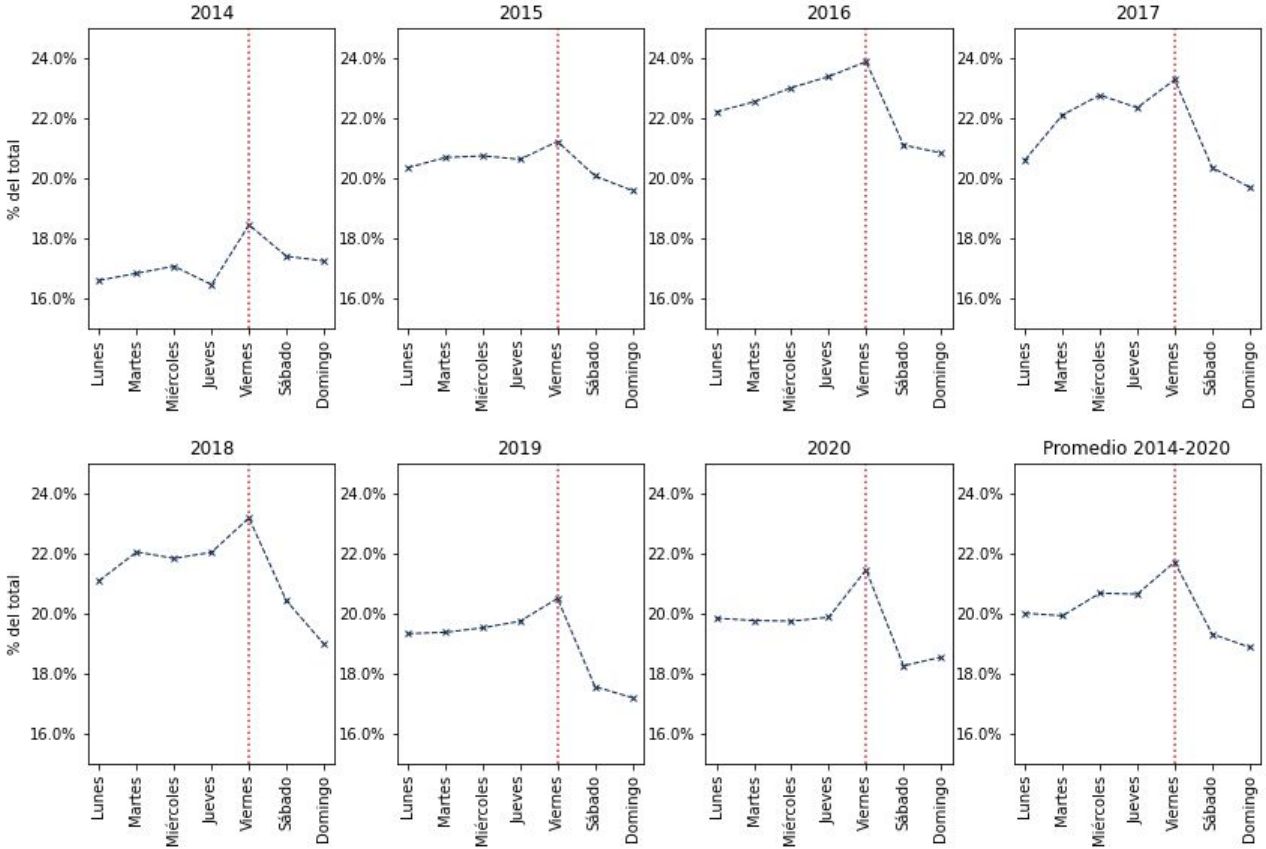
Proporción de llamadas falsas o negativas a lo largo del año:



No se aprecia una tendencia clara respecto a la distribución de llamadas falsas a lo largo del año.

Podemos notar una disminución en llamadas falsas durante el periodo de la contingencia sanitaria en 2020.

Proporción de llamadas falsas o negativas por día de la semana (2014-2020):

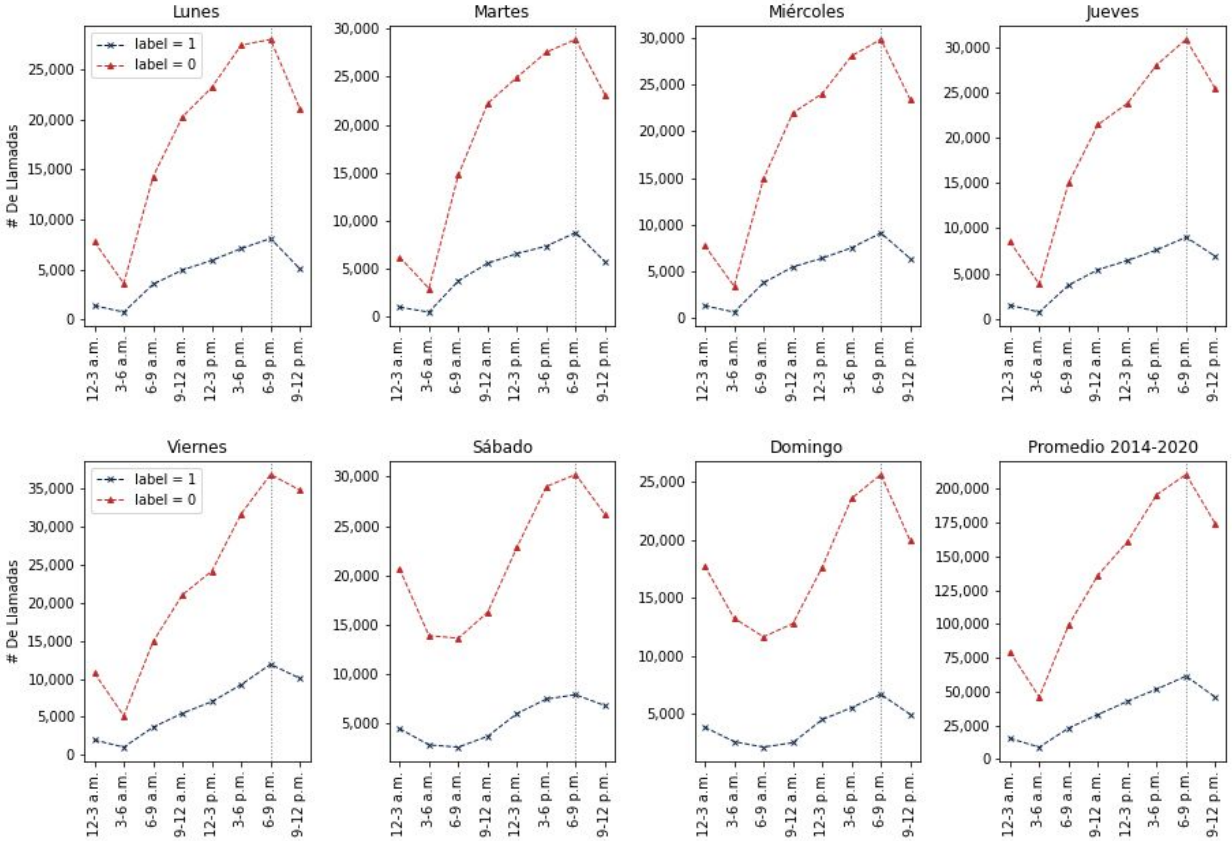


El viernes es el día de la semana con la proporción más alta de llamadas falsas.

Proporción de llamadas falsas o negativas por horario (2014-2020):

De 6 a 9pm cada día, se registra la mayor proporción de llamadas falsas.

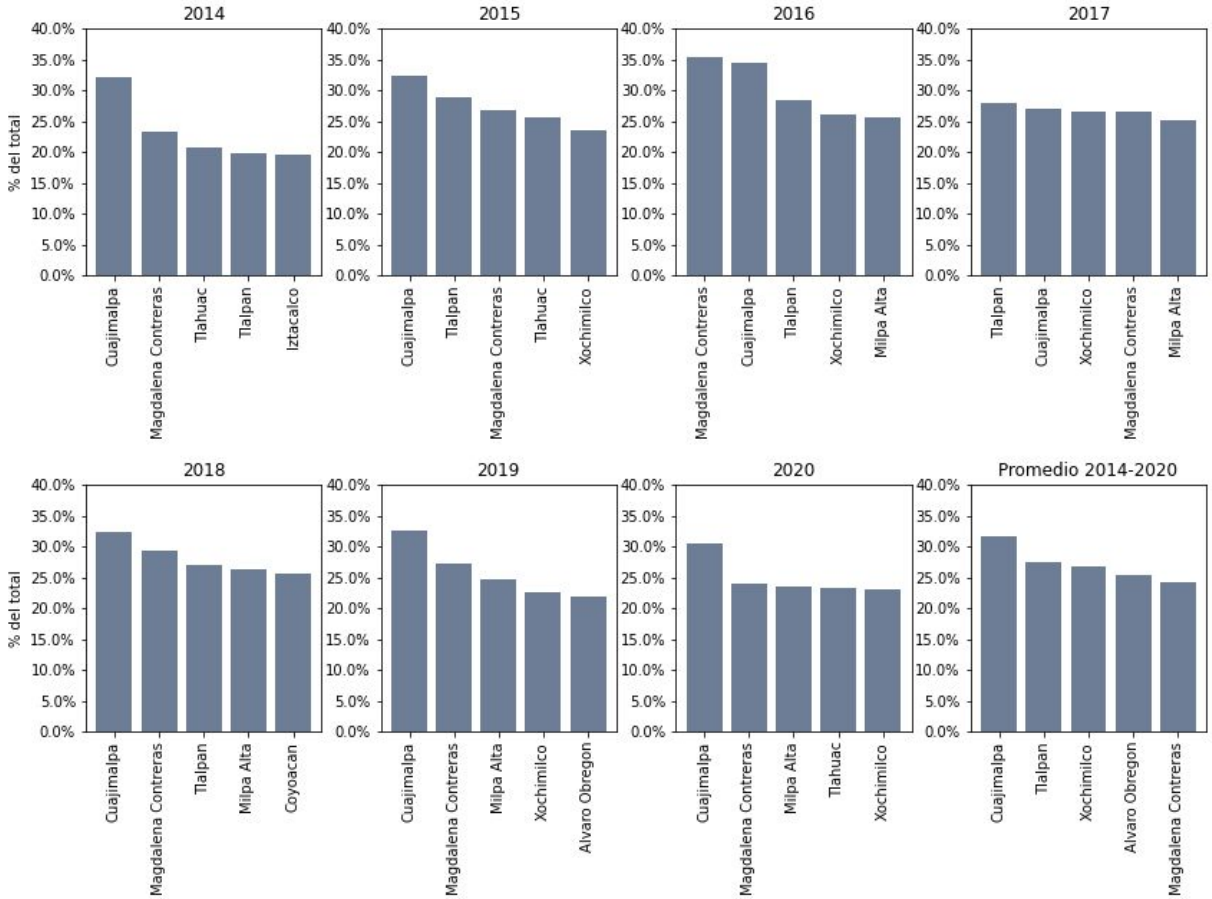
Número de llamadas a lo largo del día:



De 6 a 9pm cada día, se registra el mayor número de llamadas diarias, tanto falsas como verdaderas.

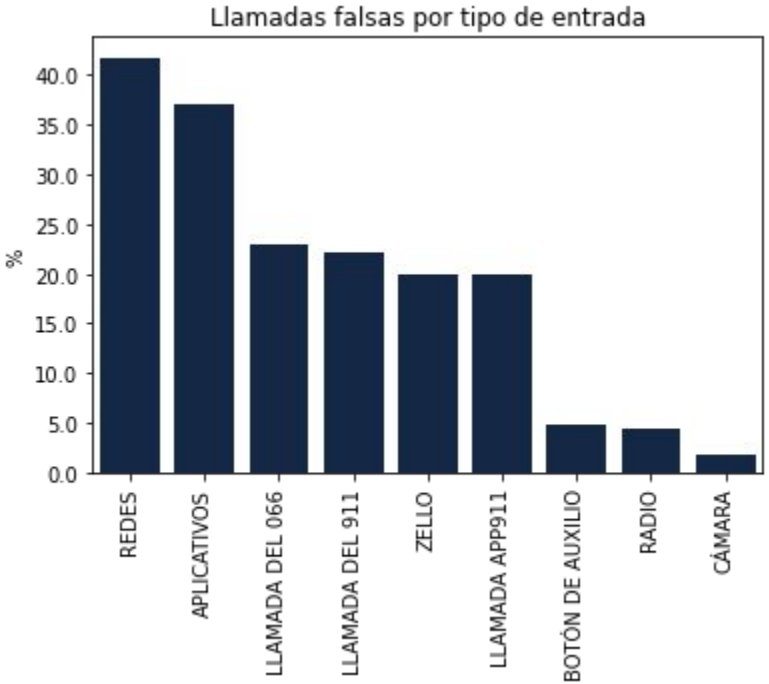
*Nota: Label = 1 implica llamadas falsas o negativas y label = 0 todas las demás.

Top 5 Delegaciones con mayor proporción de llamadas falsas:



Cuajimalpa es la delegación con la mayor proporción de llamadas falsas durante la mayoría de los años.

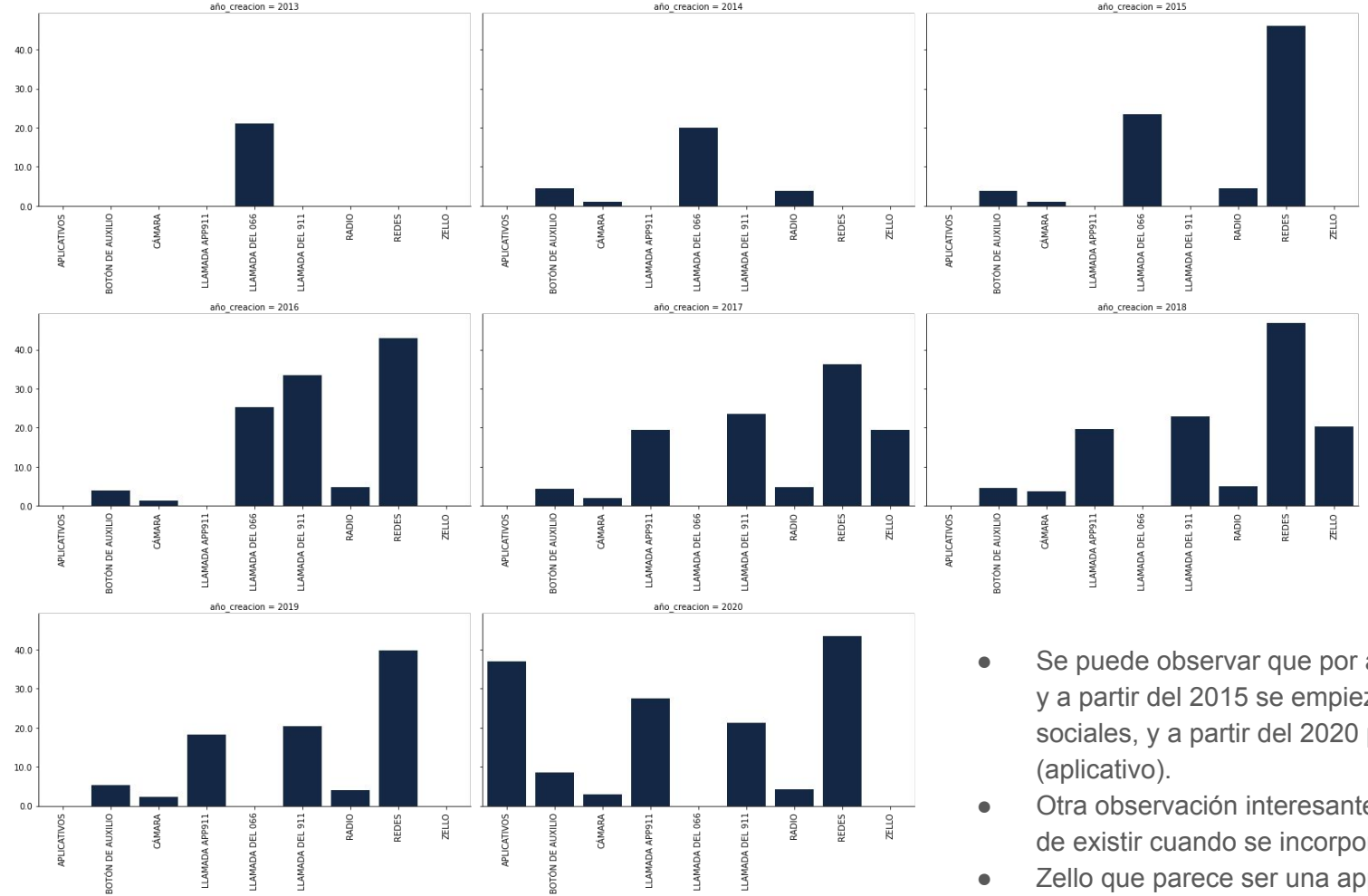
Llamadas falsas por tipo de entrada (1/2)



	tipo_entrada	label	num_llamadas	prop
0	REDES	1	1963	41.792634
1	APLICATIVOS	1	30	37.037037
2	LLAMADA DEL 066	1	106624	23.007475
3	LLAMADA DEL 911	1	163284	22.085133
4	ZELLO	1	1290	19.996900
5	LLAMADA APP911	1	1889	19.919857
6	BOTÓN DE AUXILIO	1	3795	4.826402
7	RADIO	1	3370	4.334461
8	CÁMARA	1	58	1.770452

- Redes es el tipo de entrada de llamada falsa más común, seguido de las aplicaciones, aunque aplicativo tiene pocas entradas (30 llamadas falsas)

Llamadas falsas por tipo de entrada (2/2)



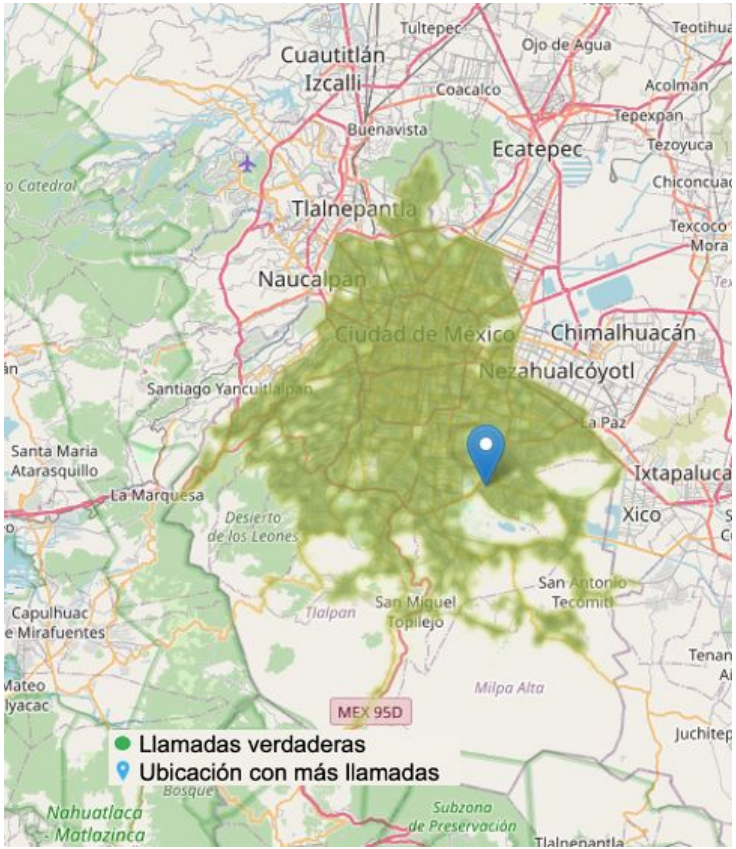
- Se puede observar que por años se comporta diferente y a partir del 2015 se empieza a ver más las redes sociales, y a partir del 2020 por medio de aplicaciones (aplicativo).
- Otra observación interesante es la llamada al 066 deja de existir cuando se incorpora llamar al 911 en México.
- Zello que parece ser una app tipo Walkie Talkie cuando las llamadas no entran, sólo aparece en el 2017 y 2018.

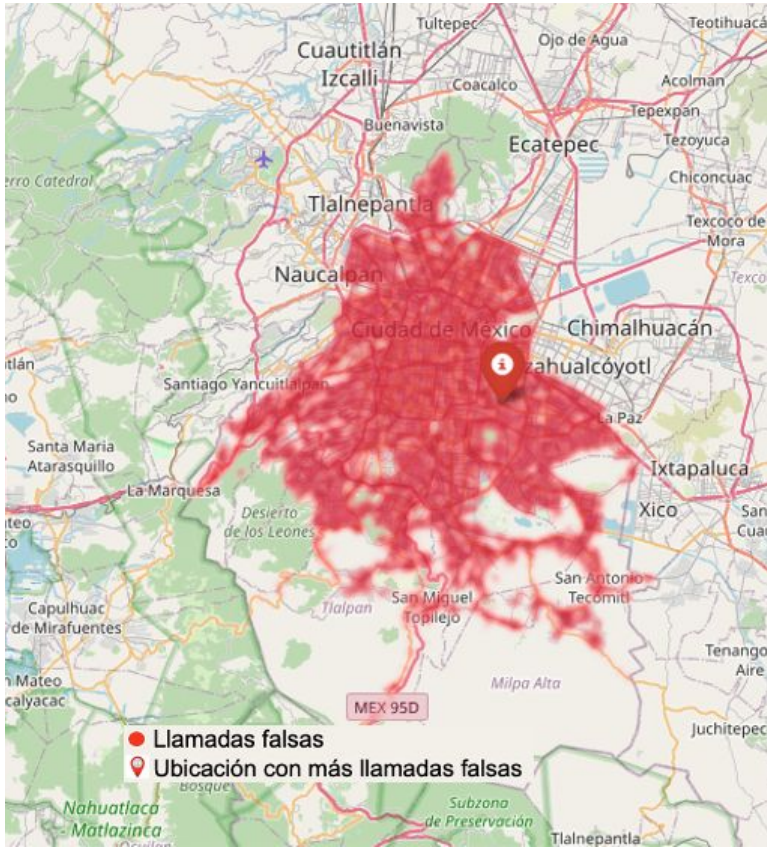
Llamadas falsas por tipo de incidente

Top 5

incidente_c4	label	num_llamadas	prop
mi ciudad-taxi-incidente de tránsito	1	1.0	100.000000
mi ciudad-calle-incidente de tránsito	1	28.0	50.909091
accidente-vehiculo atrapado	1	314.0	28.010705
accidente-vehículo atrapado-varado	1	232.0	27.230047
accidente-choque sin lesionados	1	198694.0	26.126553

Mapa de llamadas falsas y verdaderas: *





*Nota: Debido a la gran cantidad de puntos, se separan los mapas por llamadas falsas y verdaderas.