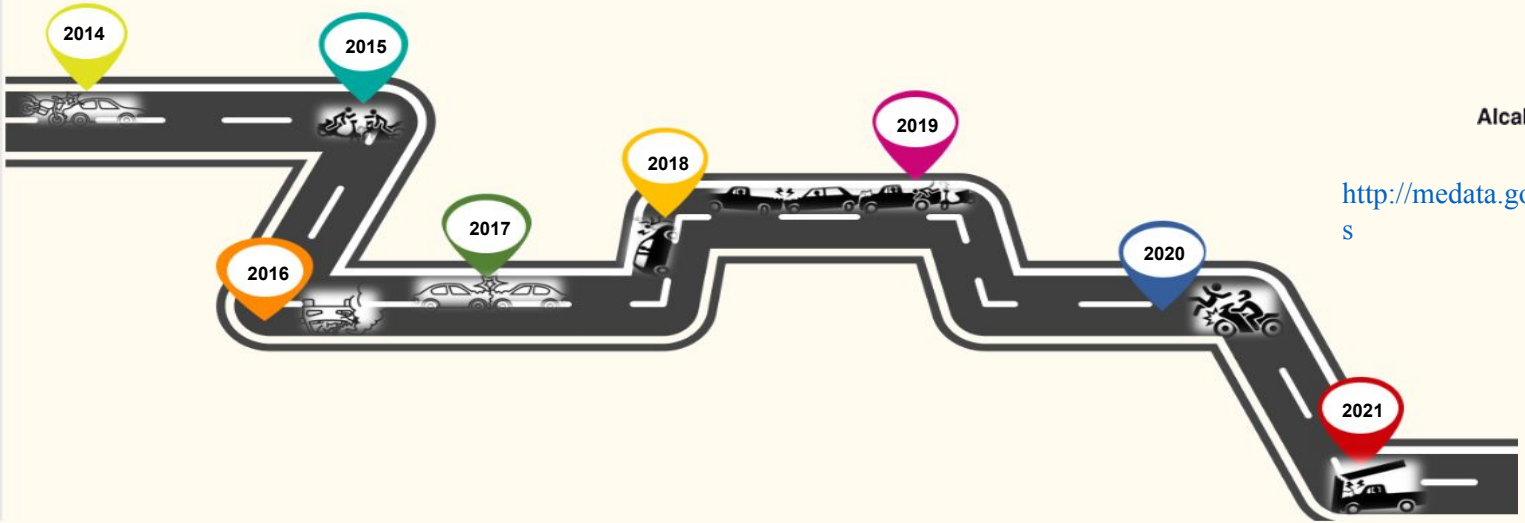
DIPLOMATURA EN CIENCIA DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS APLICACIONES.
COHORTE 2022 VIRTUAL

M08 – VÍCTIMAS DE INCIDENTES VIALES

Profesora: Isabel Mejía

Grupo formado por:
* Giovine – Spitale - Venchiarutti

Análisis Inicial de Datos



Alcaldía de Medellín

http://medata.gov.co/dataset/incidentes-viales

Objetivo

Estudiar, analizar y comprender los datos y su complejidad para obtener informacion sobre la ocurrencia de los incidente viales ayudando a la disminución y prevención de los mismos

Tipos Datos		Cantidad	Nulos
Gravedad_victima	object	235843	0
Fecha_incidente	object	235843	0
Hora_incidente	object	235843	0
Clase_incidente	object	235843	0
Direccion_incidente	object	235831	12
Sexo	object	235843	0
Edad	object	235335	508
Condicion	object	235843	0
Mes	object	235843	0
Dia	object	235843	0
Num_dia	int64	235843	0
Hora	object	235843	0
Grupo_edad	object	235843	0
Año	int64	235843	0
Radicado	object	235838	5
Latitud	object	235843	0
Longitud	object	235843	0
Comuna	object	235843	0
Barrio	object	235225	618

Variables Categóricas

Ordinales

Fecha_incidente = ['1/1/2014' '2/1/2014' '3/1/2014' ... '28/9/2021' '29/9/2021' '30/9/2021']
Hora_incidente = ['00:15:00' '00:30:00' '00:37:00' ... '01:18:00' '03:53:00' '02:07:00']
Mes = ['Ene' 'Feb' 'Mar' 'Abr' 'May' 'Jun' 'Jul' 'Ago' 'Sept' 'Oct' 'Nov' 'Dic' 'Sep']
Dia = ['Mié' 'Jue' 'Vie' 'Sáb' 'Dom' 'Lun' 'Mar']
Grupo_edad = ['oct-19' '20 - 29' '30 - 39' '40 - 49' '0 - 9' '50 - 59' 'Sin Inf' '60 - 69' '70 - 79' '80 o más']

Nominales

Gravedad_victima = ['Heridos' 'Muertos']
Clase_incidente = ['Otro' 'Atropello' 'Choque' 'Caida Ocupante' 'Volcamiento' 'Incendio']
Direccion_incidente = ['CR 49 CL 72' 'CR 46 CL 98' 'CL 32 CR 84' ... 'CR 49 DG 50' 'DG 75 B CL 76' 'CL 28 A CR 65 A']
Sexo = ['M' 'F' 'Sin Inf' 'Sin inf'] ']
Condicion = ['Motociclista' 'Peatón' 'Acompañante de Motocicleta' 'Conductor' 'Ciclista' 'Pasajero' 'Acompañante de motocicleta']
Comuna = Compuesto por 22 Comunas identificadas por un Numero y su Nombre
Barrio =

Variables Cuantitativas

Discretas

Edad = ['17' '20' '18' '19' '39' '44' '7' '35' '51' '30' 'Sin Inf' '34' '26' '29' '27' '32' '33' '24' '23' '36' '25' '28' '52' '38' '61' '58' '22' '73' '21' '5' '31' '4' '14' '63' '50' '49' '59' '54' '85' '6' '46' '62' '15' '41' '16' '2' '47' '37' '83' '55' '13' '65' '3' '72' '57' '9' '45' '12' '82' '43' '1' '40' '53' '56' '0' '8' '76' '71' '42' '11' '64' '67' '70' '66' '77' '48' '78' '68' '74' '10' '60' '79' '75' '69' '91' '81' '88' '89' '86' '90' '84' '80' '87' '92' '98' '95' '94' '97' '93' '96' '118' '106' '108' '107' '104' '105' '119' '30-35' '109' '45-50' '137' '102' '30 - 35' '20 - 29' '99' '110' nan '120' '100' '121' '111
Num_dia = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 0]
Hora = [0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 '9' '10' '11' '12' '13' '14' '15' '16' '17' '18' '19' '20' '21' '22' '0' '5' '6' '7' '8' '23' '3' '1' '2' '4' 'Sin Inf']
Año = [2014 2015 2016 2017 2018 2019 2020 2021]
Radicado = ['1423940' '1423921' '1423849' ... 1763968 1764133 1763946]

Continuas

Latitud = ['6,26691466' '6,289353458' '6,234327372' ... '-75,57582422' '-75,53631071' '-75,54867484'] **Longitud =** ['-75,5590994' '-75,55329197' '-75,60761079' ... '6,2178952' '6,23426695' '6,272697']

Variables Categóricas

Mes = ['Ene' 'Feb' 'Mar' 'Abr' 'May' 'Jun' 'Jul' 'Ago' 'Sept' 'Oct' 'Nov' 'Dic' 'Sep']

Grupo_edad = ['oct-19' '20 - 29' '30 - 39' '40 - 49' '0 - 9' '50 - 59' 'Sin Inf' '60 - 69' '70 - 79' '80 o más']

Sexo = ['M' 'F' 'Sin Inf' 'Sin inf']

Condicion = ['Motociclista' 'Peatón' 'Acompañante de Motocicleta' 'Conductor' 'Ciclista' 'Pasajero' 'Acompañante de motocicleta']

Corregimos los datos atípicos marcados y dejamos el valor 'Sin Inf' para variables categóricas.

Variables Cuantitativas

Edad = ['17' '20' '18' '19' '39' '44' '7' '35' '51' '30' 'Sin Inf' '34' '26' '29' '27' '32' '33' '24' '23' '36' '25' '28' '52' '38' '61' '58' '22' '73' '21' '5' '31' '4' '14' '63' '50' '49' '59' '54' '85' '6' '46' '62' '15' '41' '16' '2' '47' '37' '83' '55' '13' '65' '3' '72' '57' '9' '45' '12' '82' '43' '1' '40' '53' '56' '0' '8' '76' '71' '42' '11' '64' '67' '70' '66' '77' '48' '78' '68' '74' '10' '60' '79' '75' '69' '91' '81' '88' '89' '86' '90' '84' '80' '87' '92' '98' '95' '94' '97' '93' '96' '118' '106' '108' '107' '104' '105' '119' '30-35' '109' '45-50' '137' '102' '30 - 35' '20 - 29' '99' '110' nan '120' '100' '121' '111']

Num_dia = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 0]

Hora = [0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 '9' '10' '11' '12' '13' '14' '15' '16' '17' '18' '19' '20' '21' '22' '0' '5' '6' '7' '8' '23' '3' '1' '2' '4' 'Sin Inf']

Radicado = [1423940. 1423921. 1423849. ... 1763968. 1764133. 1763946.]

Latitud = [6.26691466 6.28935346 6.23432737 ... -75.57582422 -75.53631071 -75.54867484]

Longitud = [-75.5590994 -75.55329197 -75.60761079 ... 6.2178952 6.23426695 6.272697]

Se transformaron de tipo **Object** a **Float** y los **Sin Inf** a nulos **NaN.** Corregimos los datos atípicos marcados y eliminamos **outliers**.

Data Set Final

```
RangeIndex: 235843 entries, 0 to 235842
Data columns (total 19 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Gravedad_victima     235843 non-null  object
 1   Fecha_incidente      235843 non-null  datetime64[ns]
 2   Hora_incidente       235843 non-null  object
 3   Clase_incidente      235843 non-null  object
 4   Direccion_incidente  235831 non-null  object
 5   Sexo                 235843 non-null  object
 6   Edad                 233429 non-null  float64
 7   Condicion            235843 non-null  object
 8   Mes                  235843 non-null  object
 9   Dia                  235843 non-null  object
 10  Num_dia              235842 non-null  float64
 11  Hora                 235836 non-null  float64
 12  Grupo_edad           235843 non-null  object
 13  Año                  235843 non-null  int64
 14  Radicado             235794 non-null  float64
 15  Latitud              214998 non-null  float64
 16  Longitud             214998 non-null  float64
 17  Comuna               235843 non-null  object
 18  Barrio               235225 non-null  object
dtypes: datetime64[ns](1), float64(6), int64(1), object(11)
```

Visualización de Datos

El incidente más frecuente es Choque

Análisis detallado por día según la Clase de Incidente

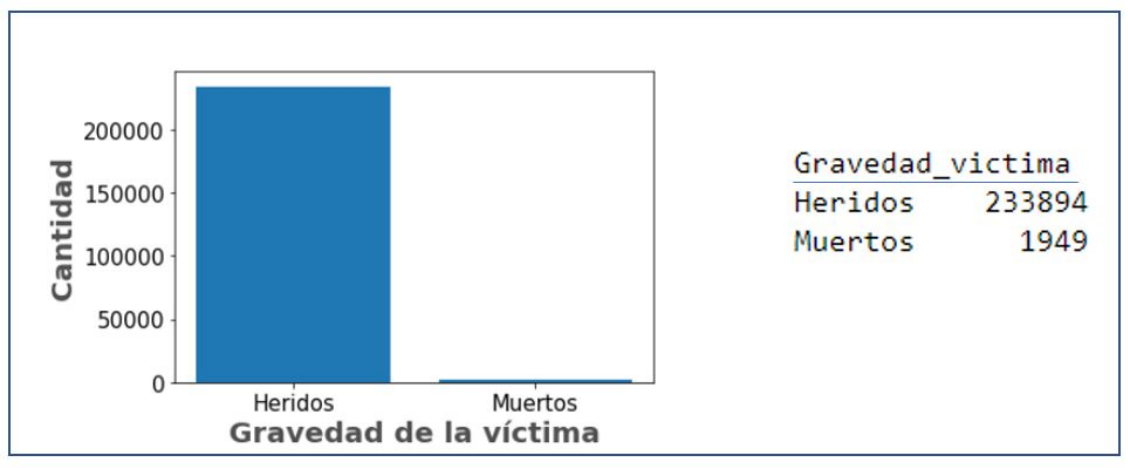
Dia	Dom	Jue	Lun	Mar	Mié	Sáb	Vie	All
Clase_incidente								
Atropello	5109	5625	5259	5622	5686	6311	5843	39455
Caida Ocupante	4287	5274	5429	5228	5008	4726	5146	35098
Choque	12308	14845	14820	15186	15523	15265	15436	103383
Incendio	3	5	8	2	2	1	3	24
Otro	4913	6347	6093	6195	6171	5414	6029	41162
Volcamiento	1950	2472	2469	2517	2524	2390	2399	16721
All	28570	34568	34078	34750	34914	34107	34856	235843

La ocurrencia de incidentes se distribuye de manera uniforme para cualquier día de la semana con una leve baja los días domingos.

Identificamos en el gráfico de barras una distribución casi uniforme de cantidad de incidentes para cualquier mes del año, levemente sobresalen los meses de Agosto y Septiembre.

En cuanto a la frecuencia de incidentes por sexo, la ocurrencia de los mismos en el sexo Masculino es más del doble que los incidentes que sufre el sexo Femenino.

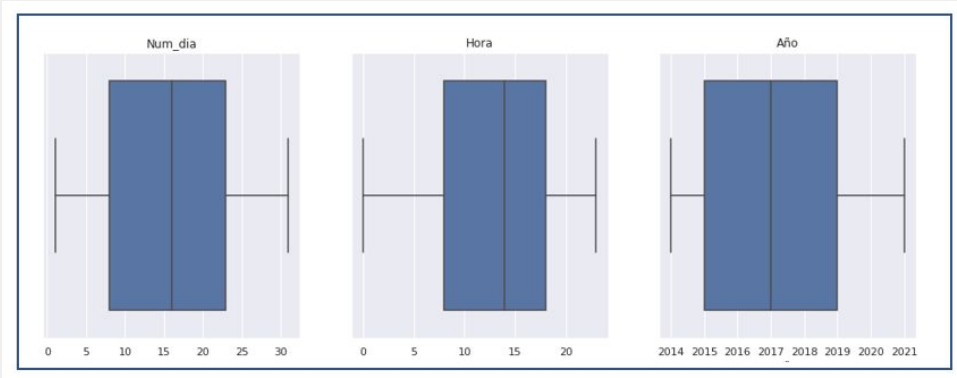
	Sexo	Cantidad_Incidentes
0	F	71763
1	M	161603
2	Sin Inf	2477

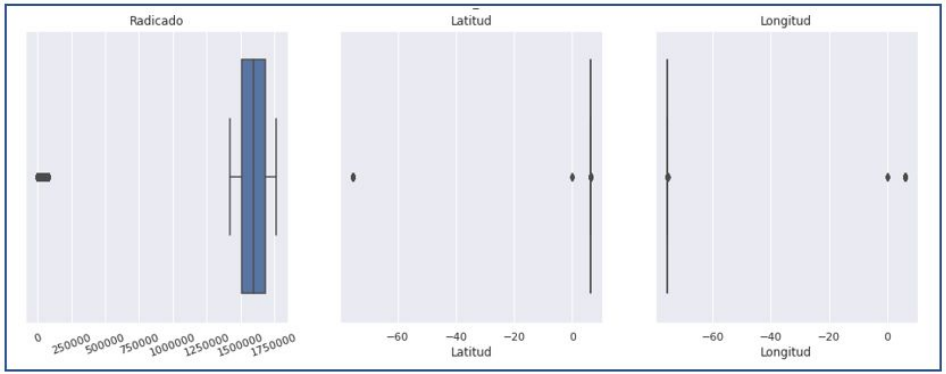


Gravedad_victima	
Heridos	233894
Muertos	1949

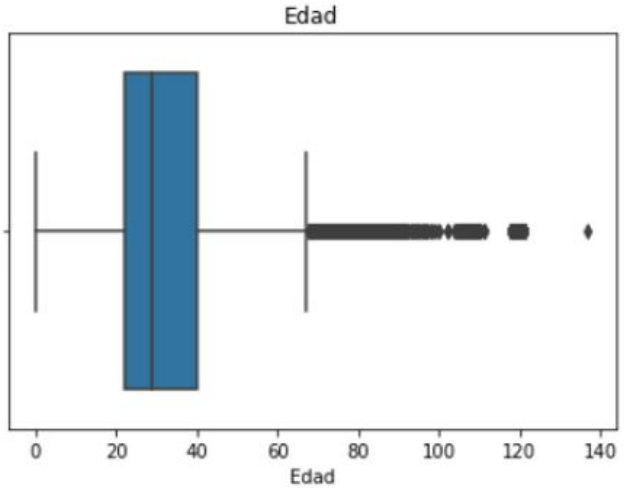
Variable: ‘Gravedad_victima’
Conteo de la cantidad de ‘Heridos’ y ‘Muertos’ en los incidentes viales.

Boxplot de variables numéricas

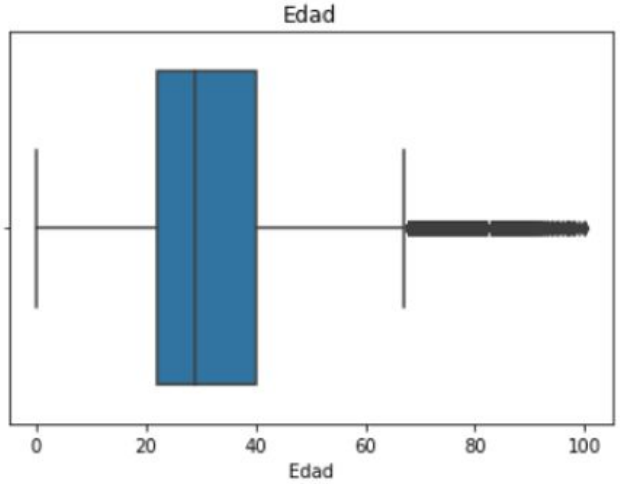




```
count    233429.000000
mean         32.400344
std          14.849399
min           0.000000
25%          22.000000
50%          29.000000
75%          40.000000
max         137.000000
Name: Edad, dtype: float64
```




```
count    232875.000000
mean         32.203255
std          14.303002
min           0.000000
25%          22.000000
50%          29.000000
75%          40.000000
max         100.000000
Name: Edad, dtype: float64
```



En la columna ‘Edad’ se acotó la edad a menores de 100 años.

Son 556 registros mayores a 100 años.

Clase_incidente / Comuna	Atropello	Caida Ocupante	Choque	Incendio	Otro	Volcamiento
01 - Popular	1520	747	1045	1	770	365
02 - Santa Cruz	1191	544	1157	0	585	273
03 - Manrique	2390	1419	3342	1	1545	724
04 - Aranjuez	2852	2052	6398	2	2326	904
05 - Castilla	2850	3551	10761	0	4173	1548
06 - Doce de Octubre	1921	1787	2504	0	1499	478
07 - Robledo	2158	3291	6696	0	3445	1103
08 - Villa Hermosa	1647	1248	2892	1	1295	630
09 - Buenos Aires	1514	1351	3921	0	1837	809
10 - La Candelaria	7744	4434	17472	3	5176	2156
11 - Laureles Estadio	2588	2540	10186	2	3398	1208
12 - La América	974	946	3433	1	1081	365
13 - San Javier	1131	872	1741	0	840	430
14 - El Poblado	1152	1381	6803	1	2017	891
15 - Guayabal	1782	1715	7480	8	2347	1059
16 - Belén	1848	1817	6692	2	2292	992
50 - Corregimiento de San Sebastián de Palmitas	4	0	15	0	2	6
60 - Corregimiento de San Cristóbal	530	613	1048	0	628	275
70 - Corregimiento de Altavista	149	119	237	0	147	65
80 - Corregimiento de San Antonio de Prado	833	442	1568	0	535	224
90 - Corregimiento de Santa Elena	90	116	184	0	125	99
Sin Inf	2587	4113	7808	2	5099	2117

Detalle de la frecuencia de la Clase de Incidente por Comuna.

- La Comuna con más cantidad de incidentes viales por atropello, choque y caída de ocupante es “La Candelaria”.

- La Comuna con menos incidentes es “Corregimiento de San Sebastián de Palmitas”.

Exploración y Curación de Datos

Valores faltantes

- Presencia de valor ‘Sin Inf’ y sus variantes en algunas variables categóricas: **Sexo**, **Grupo_edad**, **Comuna**, **Barrio**.

Primeras decisiones

- Reemplazar a valor nulo
- Excepción de la variable Sexo

```
Proporción de valores nulos en Direccion_incidente = 0.0001
Proporción de valores nulos en Edad = 0.0102
Proporción de valores nulos en Num_dia = 0.0000
Proporción de valores nulos en Hora = 0.0000
Proporción de valores nulos en Grupo_edad = 0.0107
Proporción de valores nulos en Radicado = 0.0002
Proporción de valores nulos en Latitud = 0.0884
Proporción de valores nulos en Longitud = 0.0884
Proporción de valores nulos en Comuna = 0.0921
Proporción de valores nulos en Barrio = 0.0943
```

```
'Sin Inf', '15Sin Inf3', '1Si
'1Sin Inf18', '1Sin Inf19', '
'15Sin Inf4', 'Sin Inf31Sin I
'Sin Inf312', 'Sin Inf9Sin In
'151Sin Inf', 'Sin Inf2Sin In
'Sin Inf4Sin Inf2', '16Sin In
'Sin Inf5Sin Inf5', '11Sin In
'1Sin InfSin Inf1', 'Sin Inf5
'Sin Inf4Sin Inf9', 'Sin Inf8
'7Sin InfSin Inf2', 'Sin Inf3
'Sin Inf2Sin Inf8', 'Sin Inf3
'Sin Inf1Sin Inf5', 'Sin Inf6
'Sin Inf815', 'Sin Inf912', '
'12Sin Inf6', 'Sin Inf413', '
'11Sin Inf3', 'Sin Inf813', '
'Sin Inf819', 'Sin Inf914', '
'5Sin InfSin Inf2', '9Sin Inf
```

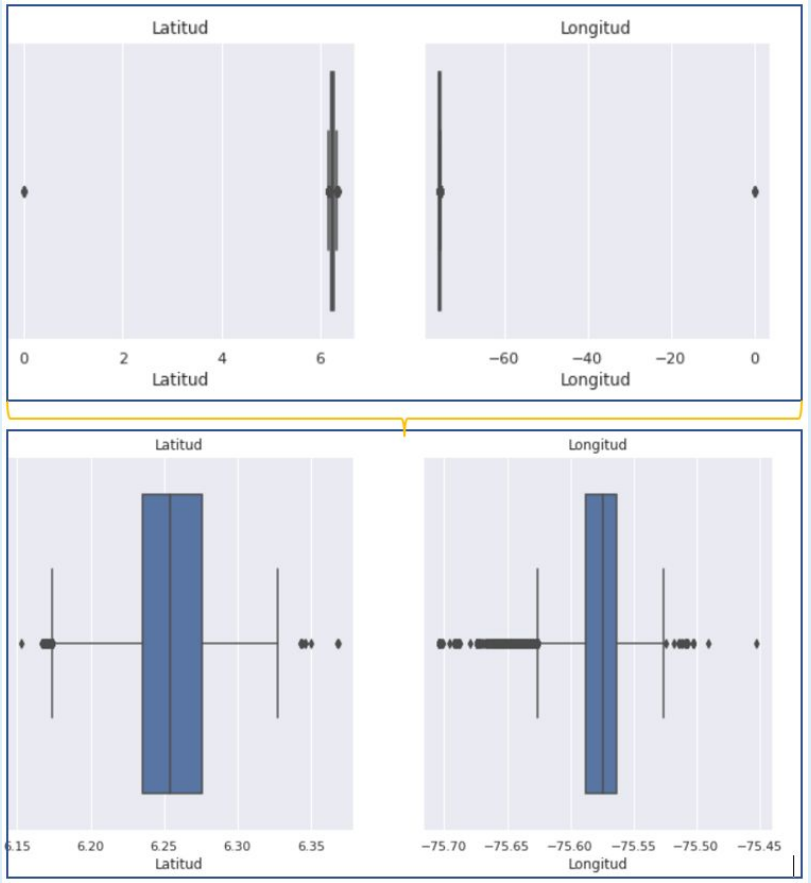
¿Se podría imputar Latitud y Longitud a partir de inferencia de otros datos?

¿Se podría imputar Barrio y Comuna siguiendo la misma idea?

Datos erróneos/atípicos

- Valores invertidos de las coordenadas de los incidentes, considerando las coordenadas de Medellín (**Lat** 6.217, **Lon** -75.567). 1949 registros.

	Latitud	Longitud		Latitud	Longitud
0	6.266915	-75.559099	0	6.266915	-75.559099
1	6.289353	-75.553292	1	6.289353	-75.553292
2	6.289353	-75.553292	2	6.289353	-75.553292
3	6.234327	-75.607611	3	6.234327	-75.607611
4	6.234327	-75.607611	4	6.234327	-75.607611
...	...	...	...	...	...
235838	-75.536311	6.234267 →	235838	6.234267	-75.536311
235839	-75.536311	6.234267 →	235839	6.234267	-75.536311
235840	-75.548675	6.272697 →	235840	6.272697	-75.548675
235841	NaN	NaN	235841	NaN	NaN
235842	NaN	NaN	235842	NaN	NaN



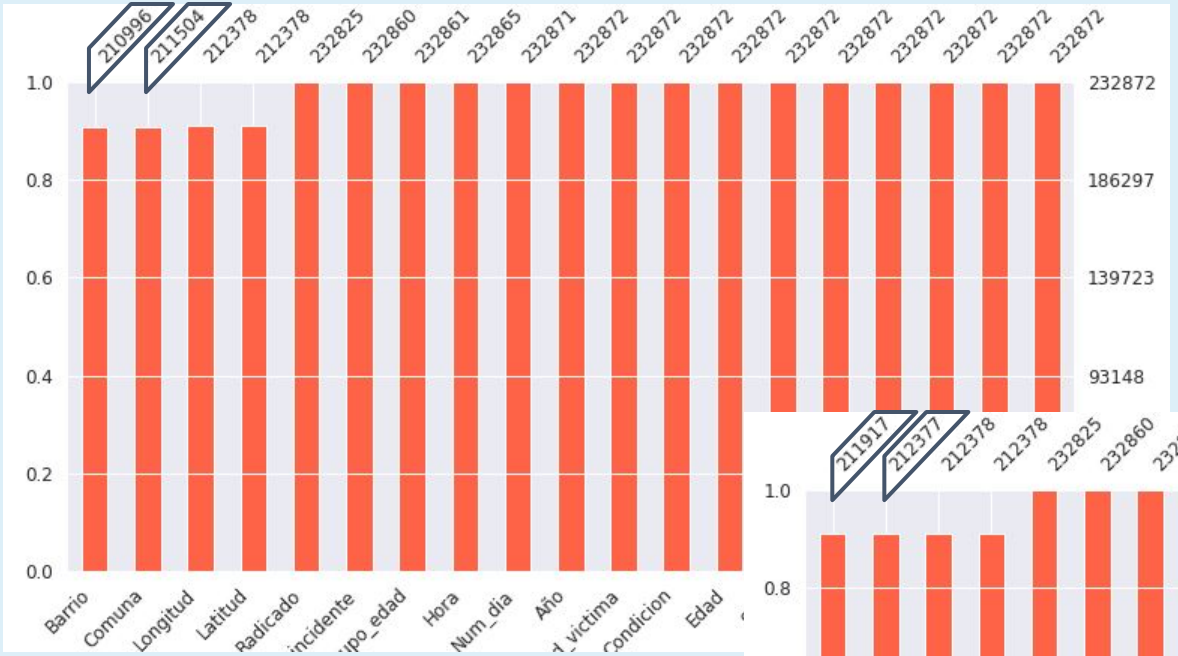
- Exploración de Radicado

```
1656780.0    34
1549032.0    33
1659473.0    24
1510395.0    24
1678108.0    18
             ..
1549065.0     1
1549037.0     1
1549133.0     1
1549054.0     1
1763946.0     1
Name: Radicado,
```

Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	43.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	34.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	39.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	24.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	49.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	30.0	Conductor	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	16.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	62.0	Pasajero	Ene
Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	40.0	Pasajero	Ene

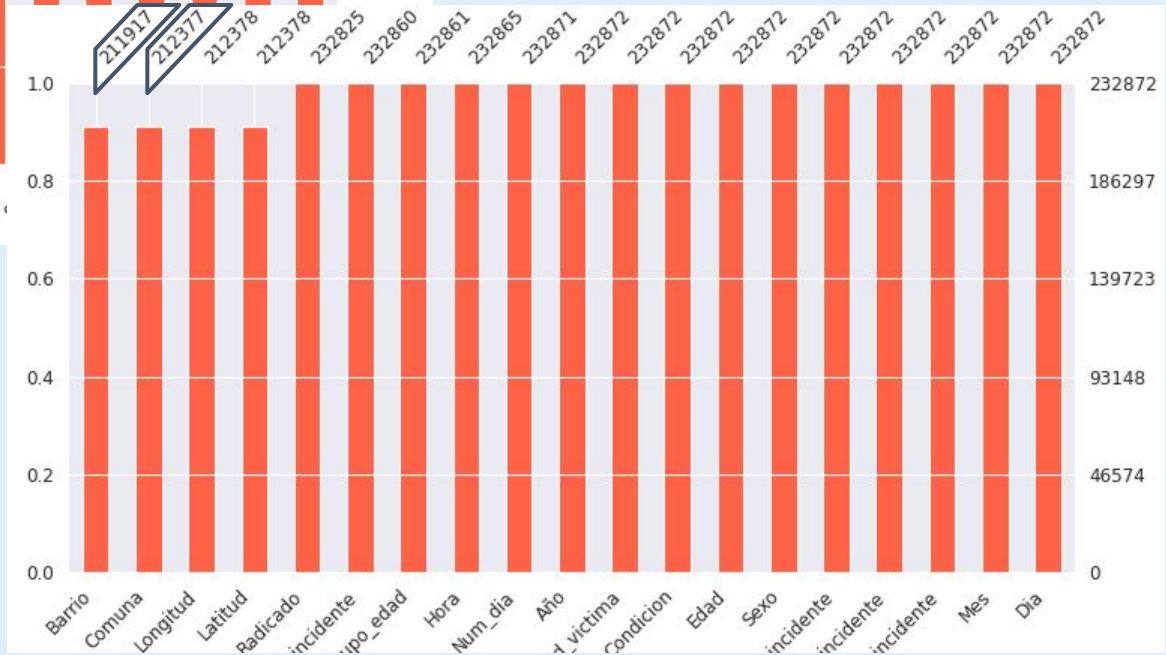
- Análisis de Sexo por Condición

- Obtener sólo un registro de cada grupo de radicado repetido, y unirlos con los de un único valor
- Agregar nuevas columnas a partir de esta



Imputación de Barrio y Comuna

Se utilizó la función **Nominatim** de la librería **geopy.geocoders** y se imputaron los registros que contaban con 'Latitud' y 'Longitud' como información de ubicación.



- 873/875 registros con coordenadas no nulas imputados para Comuna
- 921/1382 registros con coordenadas no nulas imputados para Barrio

En esta primera etapa hemos analizado, visualizado y curado nuestros datos.

Posteriormente, aplicaremos un modelo de Machine Learning que nos permita predecir algunas características de accidentes viales, que tienen que ver con la ubicación, el tiempo o ciertos atributos de las víctimas. Lo cual puede ser de ayuda para generar estrategias, controles de tránsito y otras cuestiones que permitan disminuir los índices de accidentes.

Muchas Gracias.

Candela
Gustavo
Carina.