



DIPLOMATURA EN

CIENCIA DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS

APLICACIONES

COHORTE 2022 VIRTUAL

MENTORÍA M08-Víctimas de Incidentes Viales.

Mentora: Isabel Mejía.

Grupo formado por:

- * Giovine, Carina;
- * Spitale, Candela;
- * Venchiarutti, Gustavo.

Tabla de contenido

[INTRODUCCIÓN](#)

[ANÁLISIS DE LOS DATOS](#)

[Descripción de cada columna que forma parte del Dataset](#)

[Clasificación de Variables](#)

[VISUALIZACIÓN DE DATOS](#)

[CURACIÓN DE DATOS](#)

INTRODUCCIÓN

En el presente trabajo vamos a plasmar todo lo aprendido en las materias de la Diplomatura analizando un set de datos de la Alcaldía de Medellín.

Los datos utilizados son datos abiertos, o sea datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona. Los mismos se obtuvieron de la página <http://medata.gov.co/dataset/incidentes-viales> . Esta base de datos se genera del reporte de víctimas entregado por UNE (Secretaría de Movilidad-UNE Subsecretaría Técnica).

Se entiende por Incidente de tránsito: evento, generalmente involuntario, generado al menos por un vehículo en movimiento, que causa daños a personas y bienes involucrados en él, e igualmente afecta la normal circulación de los vehículos que se movilizan por la vía o vías comprendidas en el lugar o dentro de la zona de influencia del hecho.

Se han hecho incontables esfuerzos a nivel mundial para reducir el número de accidentes, y los países han adoptado leyes y reglamentos sobre los factores de riesgo fundamentales. Tras este panorama, es necesario continuar la búsqueda de soluciones y medidas que permitan la prevención de accidentes.

A medida que nuestros sistemas de información evolucionan, se ha facilitado la recolección, administración y almacenamiento de información acerca de estos eventos.

La Alcaldía de Medellín ha dado un paso adelante al liberar estos datos para su análisis. Una vez que estas prácticas comienzan a implementarse y reproducirse, es posible plantear nuevas formas de abordar esta problemática basada en datos, que permitan hacer análisis más precisos que lleven a tomar acciones más acertadas para reducir la cantidad de estos trágicos eventos.

Nuestro objetivo se centra en estudiar, analizar y comprender los datos y su complejidad, con el objeto de obtener información sobre la ocurrencia de los incidentes viales ayudando a la disminución y prevención de los mismos. Con esta información podemos aplicar modelos de machine learning para detectar patrones y los factores determinantes de los tipos de hechos de accidentes de tránsito; permitiendo apoyar la toma de decisiones frente a los mismos, como pueden ser implementando planes de prevención de accidentes vehiculares o rápidas acciones para el tratamiento oportuno de las víctimas de los siniestros.

Este documento es soporte de los archivos ejecutables con código Python que se encuentran disponibles en el siguiente repositorio <https://github.com/Knd9/mentoríaDD2022>

ANÁLISIS DE LOS DATOS

Nuestro Dataset original tiene 235.843 filas y 19 columnas. Contiene datos de siniestros desde el 01 enero de 2014 al 30 de septiembre de 2021.

En detalle las columnas que contiene son: 'Gravedad_victima', 'Fecha_incidente', 'Hora_incidente', 'Clase_incidente', 'Direccion_incidente', 'Sexo', 'Edad', 'Condicion', 'Mes', 'Dia', 'Num_dia', 'Hora', 'Grupo_edad', 'Año', 'Radicado', 'Latitud', 'Longitud', 'Comuna', 'Barrio'.

Descripción de cada columna que forma parte del Dataset:

Gravedad_victima: Representa valores tales como 'heridos' y 'muertos'.

Fecha_incidente: Representa la fecha en la que ocurrió el evento.

Hora_incidente: Indica la hora en que se produce del evento.

Clase_incidente: Representa el tipo de incidente y toma los valores: 'Otro', 'Atropello', 'Choque', 'Caída Ocupante', 'Volcamiento', 'Incendio'.

Direccion_incidente: Indica el lugar donde se produjo el accidente, especificando CR(carretera) y CL(calle).

Sexo: Indica el sexo de la persona involucrada en el accidente. Toma valores tales como 'hombre', 'mujer' y 'sin información'.

Edad: Indica el sexo de la persona involucrada en el accidente. Variable que contiene numeros, algunos con valor 0 y otros registros con rangos de edades.

Condicion: Indica la condición de la persona accidentada. La variable toma los siguientes valores: 'Motociclista', 'Peatón', 'Acompañante de Motocicleta', 'Conductor', 'Ciclista', 'Pasajero', 'Acompañante de motocicleta'.

Mes: Indica el mes que se produce el accidente. La variable toma los siguientes valores: 'Ene', 'Feb', 'Mar', 'Abr', 'May', 'Jun', 'Jul', 'Ago', 'Sept', 'Oct', 'Nov', 'Dic', 'Sep'.

Dia: Identifica el día de la semana en que se produce el accidente. La variable toma los siguientes valores: 'Mié', 'Jue', 'Vie', 'Sáb', 'Dom', 'Lun', 'Mar'.

Num_dia: Identifica el número del día del mes que se produce el accidente. La variable toma valores enteros del 1 al 31.

Hora: Identifica la hora del día en la que se produce el accidente. La variable toma valores numéricos que van de 0 a 23 y también valores caracteres que van desde '0' a '23' y 'Sin Inf'.

Grupo_edad: Identifica el rango etario de la persona involucrada en el incidente. La variable toma los siguientes valores: 'oct-19', '20 - 29', '30 - 39', '40 - 49', '0 - 9', '50 - 59', 'Sin Inf', '60 - 69', '70 - 79', '80 o más'.

Año: Representa el año en que se produjo el incidente. La variable toma los siguientes valores: 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021.

Radicado: Representa el número de Expediente que se genera por cada accidente y puede ser 1 o varios para identificar los damnificados el mismo accidente.

Latitud: Identifica una latitud del incidente.

Longitud: Identifica una longitud del incidente.

Comuna: Identifica la comuna donde sucedió el incidente. La variable toma los siguientes valores:

'04 - Aranjuez', '01 - Popular', '16 - Belén', '10 - La Candelaria', '03 - Manrique', '07 - Robledo', '11 - Laureles Estadio', 'Sin Inf', '14 - El Poblado', '15 - Guayabal', '09 - Buenos Aires', '06 - Doce de Octubre', '05 - Castilla', '12 - La América', '08 - Villa Hermosa', '13 - San Javier', '60 - Corregimiento de San Cristóbal', '02 - Santa Cruz', '90 - Corregimiento de Santa Elena', '70 - Corregimiento de Altavista', '80 - Corregimiento de San Antonio de Prado', '50 - Corregimiento de San Sebastián de P.'

Barrio: Representa el barrio donde se produce el accidente.

Los tipos de datos de nuestras variables son:

Gravedad_victima	object
Fecha_incidente	object
Hora_incidente	object
Clase_incidente	object
Direccion_incidente	object
Sexo	object
Edad	object
Condicion	object
Mes	object
Dia	object
Num_dia	int64
Hora	object
Grupo_edad	object
Año	int64
Radicado	object
Latitud	object
Longitud	object
Comuna	object
Barrio	object

Visualización de las primeras 5 celdas de nuestro conjunto de datos:

	Gravedad_victima	Fecha_incidente	Hora_incidente	Clase_incidente	Direccion_incidente	Sexo	Edad	Condicion	Mes	Dia	Num_dia	Hora	Grupo_edad	Año	Radicado	Latitud	Longitud	Comuna	Barrio
0	Heridos	1/1/2014	00:15:00	Otro	CR 49 CL 72	M	17	Motociclista	Ene	Mié	1	0	oct-19	2014	1423940	6,26691466	-75,5590994	04 - Aranjuez	Manrique Central No. 1
1	Heridos	1/1/2014	00:30:00	Atropello	CR 46 CL 98	M	20	Motociclista	Ene	Mié	1	0	20 - 29	2014	1423921	6,289353458	-75,55329197	01 - Popular	Moscu No. 2
2	Heridos	1/1/2014	00:30:00	Atropello	CR 46 CL 98	F	18	Peatón	Ene	Mié	1	0	oct-19	2014	1423921	6,289353458	-75,55329197	01 - Popular	Moscu No. 2
3	Heridos	1/1/2014	00:37:00	Atropello	CL 32 CR 84	M	19	Motociclista	Ene	Mié	1	0	oct-19	2014	1423849	6,234327372	-75,60761079	16 - Belén	Las Mercedes
4	Heridos	1/1/2014	00:37:00	Atropello	CL 32 CR 84	M	39	Peatón	Ene	Mié	1	0	30 - 39	2014	1423849	6,234327372	-75,60761079	16 - Belén	Las Mercedes

Visualización de las últimas líneas del set de datos

235838	Heridos	30/9/2021	22:00:00	Otro	CL 54 CR 9 A	M	32	Motociclista	Sep	Jue	30	22	30 - 39	2021	1764135	-75,53631071	6,23426695	08 - Villa Hermosa	Las Estancias
235839	Heridos	30/9/2021	22:00:00	Otro	CL 54 CR 9 A	F	29	Acompañante de motociclista	Sep	Jue	30	22	20 - 29	2021	1764135	-75,53631071	6,23426695	08 - Villa Hermosa	Las Estancias
235840	Heridos	30/9/2021	22:00:00	Caida Ocupante	CL 81 CR 39	M	41	Acompañante de motociclista	Sep	Jue	30	22	40 - 49	2021	1763968	-75,54867484	6,272697	03 - Manrique	Santa Inés
235841	Heridos	30/9/2021	23:00:00	Atropello	CR 63 CL 32	F	51	Peatón	Sep	Jue	30	23	50 - 59	2021	1764133	Sin Inf	Sin Inf	Sin Inf	NaN
235842	Heridos	30/9/2021	23:00:00	Otro	CR 107 CL 65	M	23	Motociclista	Sep	Jue	30	23	20 - 29	2021	1763946	Sin Inf	Sin Inf	Sin Inf	NaN

235843 rows x 19 columns

La cantidad de registros faltantes por columnas se muestra en el siguiente detalle:

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

Si conocemos que nuestro dataset tiene un total de 235.843 registros, en el detalle superior se puede observar que las columnas 'Direccion_incidente', 'Edad', 'Radicado' y 'Barrio' tienen datos faltantes.

En cuanto a los datos únicos que componen cada una de las columnas se muestran:

Gravedad_victima = ['Heridos' 'Muertos']

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']

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']

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']

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

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

Dia = ['Mié' 'Jue' 'Vie' 'Sáb' 'Dom' 'Lun' 'Mar']

Num_día = [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']

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

Año = [2014 2015 2016 2017 2018 2019 2020 2021]

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

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']

Comuna=['04 - Aranjuez' '01 - Popular' '16 - Belén'
 '10 - La Candelaria'
 '03 - Manrique' '07 - Robledo' '11 - Laureles Estadio' 'Sin Inf'
 '14 - El Poblado' '15 - Guayabal' '09 - Buenos Aires'
 '06 - Doce de Octubre' '05 - Castilla' '12 - La América'
 '08 - Villa Hermosa' '13 - San Javier'
 '60 - Corregimiento de San Cristóbal' '02 - Santa Cruz'
 '90 - Corregimiento de Santa Elena' '70 - Corregimiento de Altavista'
 '80 - Corregimiento de San Antonio de Prado'
 '50 - Corregimiento de San Sebastián de Palmitas']

Barrio = ['Manrique Central No. 1' 'Moscú No. 2' 'Las Mercedes'
 'Jesús Nazareno' 'Manrique Oriental' 'Villa Flora'
 'U.D. Atanasio Girardot' 'Sin Inf' 'Villa Carlota' 'Loma de los Bernal']

Clasificación de Variables

De acuerdo a la teoría estadística las variables se clasifican en:

 **Variables categóricas:** no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).

1. **Ordinales:** Aquellas que sugieren una ordenación. Por ejemplo: nivel de estudio, posición de los ganadores de un concurso.
2. **Nominales:** Aquellas que sólo admiten una mera ordenación alfabética, pero no establecen orden por su contenido. Por ejemplo: género, estado civil, color de cabello.

📌 **Variables cuantitativas:** tienen valor numérico (edad, precio de un producto, ingresos anuales). Por su parte, las variables cuantitativas se pueden clasificar en:

1. **Discretas:** sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3,..., etc., pero, por ejemplo, nunca podrá ser 3.45).
2. **Continuas:** pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 90.4 km/h, 94.57 km/h...etc.

Aplicando algunas correcciones de tipo de datos que nos servirán para análisis posteriores, la clasificación de nuestras variables sería:

Categorías

- ➔ **Ordinales:** 'Fecha_incidente', 'Hora_incidente', 'Mes', 'Dia', 'Grupo_edad'.
- ➔ **Nominales:** 'Gravedad_victima', 'Clase_incidente', 'Direccion_incidente', 'Sexo', 'Condicion', 'Comuna', 'Barrio'.

Cuantitativas

- ➔ **Discretas:** 'Edad', 'Num_dia', 'Hora', 'Año', 'Radicado'.
- ➔ **Continuas:** 'Latitud', 'Longitud'.

En cuanto a datos atípicos, pudimos detectar los siguientes:

Se detallan por columna:

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

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

➔ 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']


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

```
→Latitud = ['6,26691466' '6,289353458' '6,234327372' ... '-75,57582422' '-75,53631071' '-75,54867484']
```

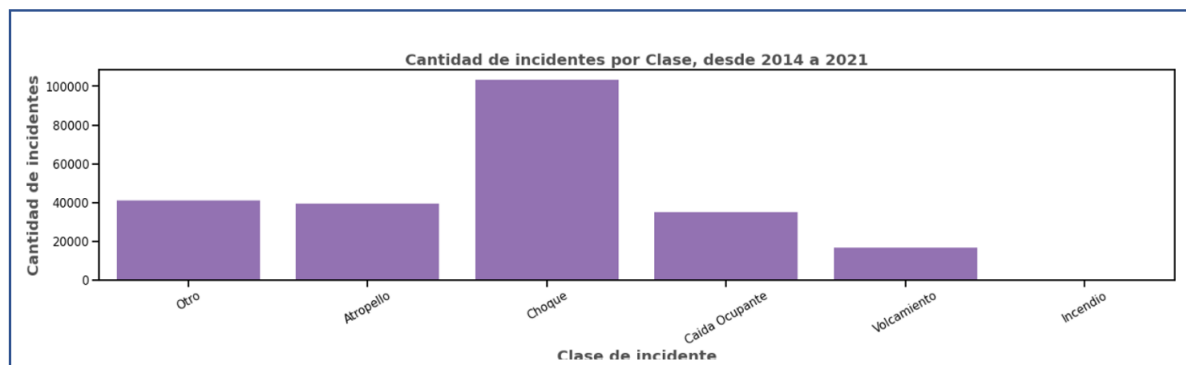
```
→Longitud = ['-75,5590994' '-75,55329197' '-75,60761079' ... '6,2178952' '6,23426695' '6,272697']
```

En este último caso se detectaron valores invertidos en los campos de Latitud y Longitud.

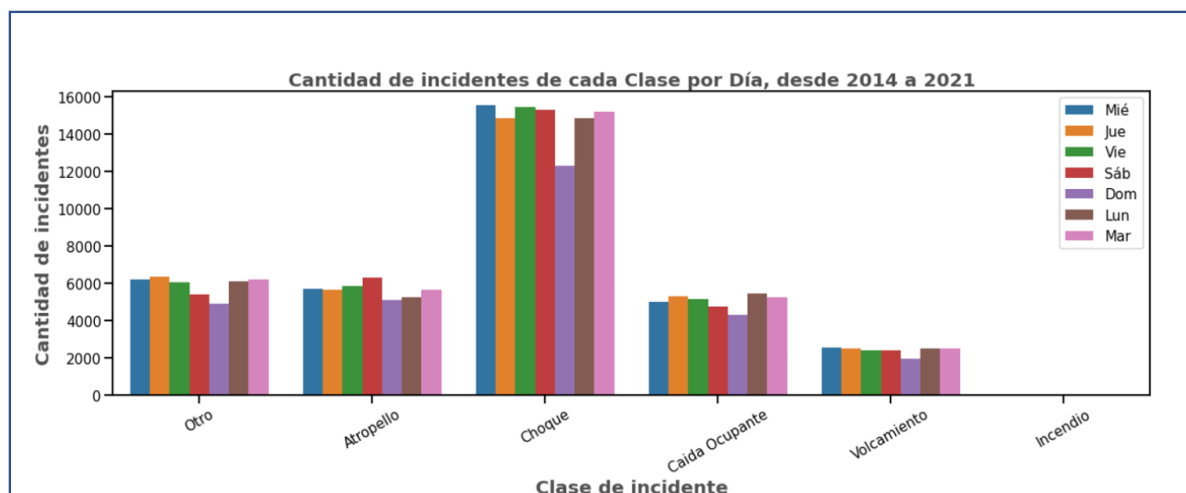
Son algunos de los casos que encontramos y que posteriormente trabajaremos en la curación de los mismos.

VISUALIZACIÓN DE DATOS

Este gráfico muestra la frecuencia de la clase de incidentes.



Cantidad de incidentes por clase en dónde cada barra muestra la frecuencia de acuerdo al día de la semana.



A través de la siguiente tabla se pueden observar en detalle las frecuencias mostradas en el gráfico superior.

	Dia	Dom	Jue	Lun	Mar	Mié	Sáb	Vie	All
Clase_incidente									
Atropello		5109	5625	5259	5622	5686	6311	5843	39455
Caída 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

Cantidad de incidentes por clase en dónde cada barra muestra la frecuencia por año, recordar que, en el caso del año 2021, los datos son hasta el mes de septiembre inclusive.

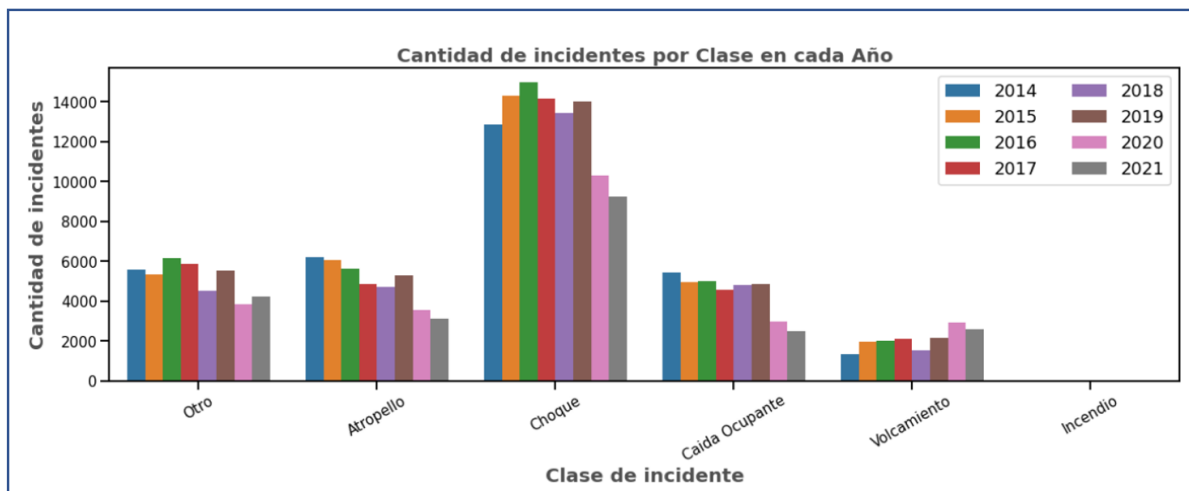


Gráfico donde se muestra la frecuencia según la condición del accidente.

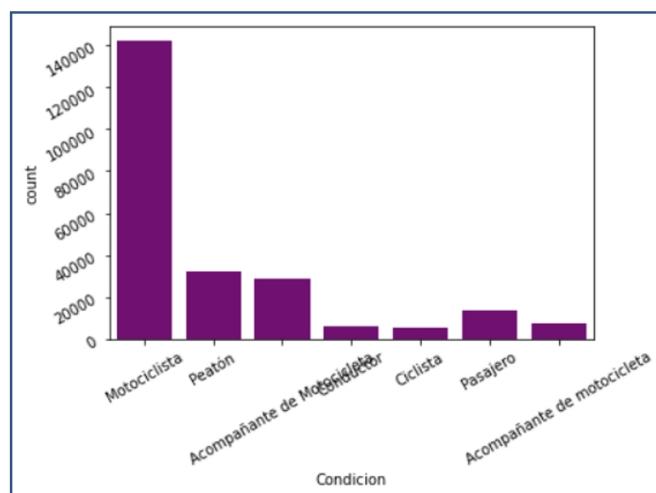
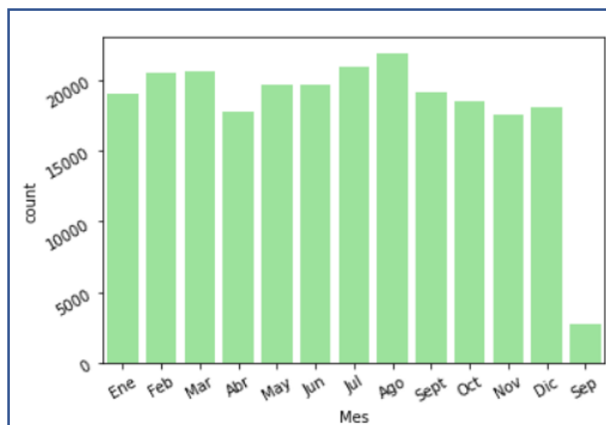


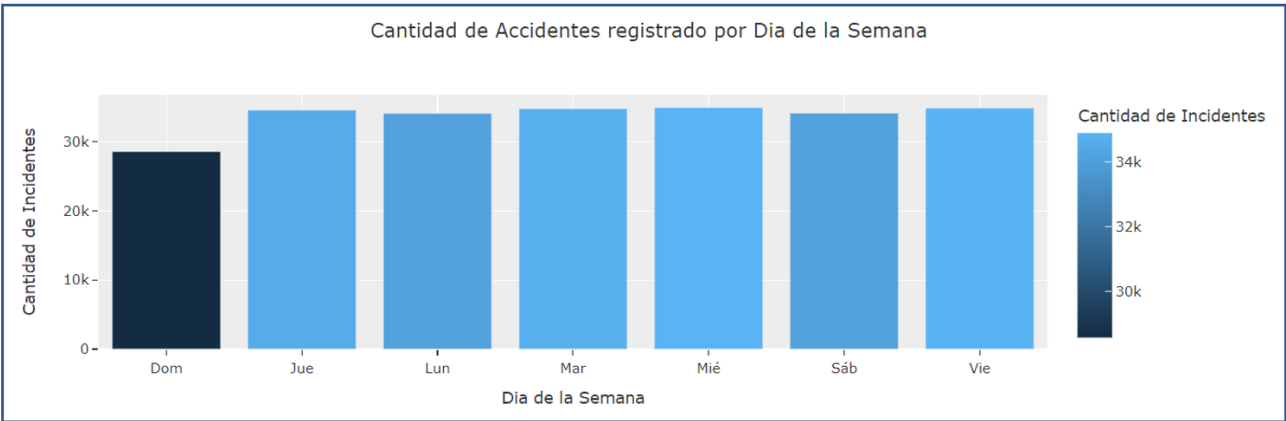
Gráfico de incidentes por mes.



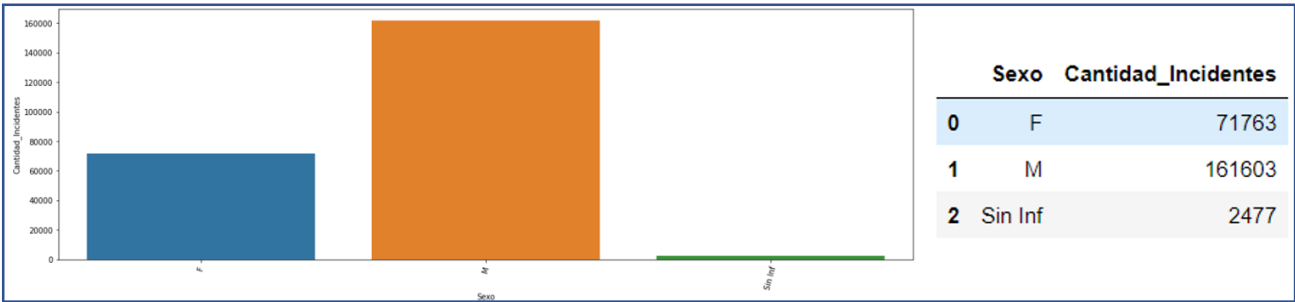
La siguiente tabla informa la cantidad total de incidentes de todos los registros del set de datos agrupados por mes.

	Mes	Cantidad_Incidentes
0	Abr	17713
1	Ago	21893
2	Dic	18115
3	Ene	18981
4	Feb	20513
5	Jul	20951
6	Jun	19612
7	Mar	20560
8	May	19649
9	Nov	17519
10	Oct	18515
11	Sept	21822

Cantidad de Incidentes por día de semana.



En la gráfica y tabla siguiente se detalla la cantidad de incidentes ocurridos de acuerdo al sexo, es notablemente superior la ocurrencia de incidentes con protagonistas masculinos.



En cuanto a los incidentes ocurridos según la comuna, se muestra un detalle de los mismos en la siguiente tabla.

Clase_incidente	Atropello	Caída Ocupante	Choque	Incendio	Otro	Volcamiento
Comuna						
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

De acuerdo a lo observado podemos determinar que en la comuna 'La Candelaria' se producen la mayor cantidad de Choques, le siguen la comuna de 'Castilla' y 'Laureles Estadio'.

Son algunos de los gráficos/visualizaciones realizadas para lograr mejor representación de los datos.

CURACIÓN DE DATOS

Cómo primera medida se analiza y se modifica el tipo de dato de algunas columnas.

#	Column	Non-Null	Dtype
0	Gravedad_victima	235843	object
1	Fecha_incidente	235843	datetime64[ns]
2	Hora_incidente	235843	object
3	Clase_incidente	235843	object
4	Direccion_incidente	235831	object
5	Sexo	235843	object
6	Edad	233429	float64
7	Condicion	235843	object
8	Mes	235843	object
9	Dia	235843	object
10	Num_dia	235842	float64
11	Hora	235836	float64
12	Grupo_edad	235843	object
13	Año	235843	int64
14	Radicado	235794	float64
15	Latitud	214998	float64
16	Longitud	214998	float64
17	Comuna	235843	object
18	Barrio	235225	object

Se observó que en varias columnas existía el valor 'Sin Inf' y variantes con esa cadena como se muestran a continuación:

```
(['Sin Inf', '15Sin Inf3', '1Sin Inf2Sin Inf', '14Sin Inf8',
 '1Sin Inf18', '1Sin Inf19', '15Sin Inf7', 'Sin Inf3Sin Inf7',
 '15Sin Inf4', 'Sin Inf31Sin Inf', '1Sin InfSin Inf7', '1Sin Inf11',
 'Sin Inf312', 'Sin Inf9Sin Inf7', '12Sin Inf4', 'Sin Inf514',
 '151Sin Inf', 'Sin Inf2Sin Inf5', 'Sin Inf4Sin Inf8',
 'Sin Inf4Sin Inf2', '16Sin Inf3', 'Sin Inf1Sin Inf3', 'Sin Inf717',
 'Sin Inf5Sin Inf5', '11Sin Inf2', '12Sin Inf2', 'Sin Inf3Sin Inf2',
 '1Sin InfSin Inf1', 'Sin Inf5Sin Inf9', '1Sin InfSin Inf4',
 'Sin Inf4Sin Inf9', 'Sin Inf811', 'Sin Inf712', '12Sin Inf5',
 '7Sin InfSin Inf2', 'Sin Inf3Sin Inf1', '1Sin InfSin Inf3',
 'Sin Inf2Sin Inf8', 'Sin Inf3Sin Inf9', 'Sin Inf3Sin Inf3',
 'Sin Inf1Sin Inf5', 'Sin Inf6Sin Inf6', 'Sin Inf3Sin Inf6',
 'Sin Inf815', 'Sin Inf912', 'Sin Inf3Sin Inf4', 'Sin Inf1Sin Inf9',
 '12Sin Inf6', 'Sin Inf413', 'Sin Inf8Sin Inf9', 'Sin Inf716',
 '11Sin Inf3', 'Sin Inf813', 'Sin Inf211', '13Sin Inf6',
 'Sin Inf819', 'Sin Inf914', '11Sin Inf1', '16Sin Inf5',
 '5Sin InfSin Inf2', '9Sin InfSin Inf8'], dtype=object)
```

Dichas columnas son: 'Sexo', 'Grupo_edad', 'Comuna', 'Barrio'.

Se reemplazan esos valores, menos en la columna 'Sexo', por valores NaN.

Gravedad_victima	0	Gravedad_victima	0
Fecha_incidente	0	Fecha_incidente	0
Hora_incidente	0	Hora_incidente	0
Clase_incidente	0	Clase_incidente	0
Direccion_incidente	12	Direccion_incidente	12
Sexo	0	Sexo	0
Edad	2414	Edad	2414
Condicion	0	Condicion	0
Mes	0	Mes	0
Dia	0	Dia	0
Num_dia	1	Num_dia	1
Hora	7	Hora	7
Grupo_edad	0	Grupo_edad	2534
Año	0	Año	0
Radicado	49	Radicado	49
Latitud	20845	Latitud	20845
Longitud	20845	Longitud	20845
Comuna	0	Comuna	21726
Barrio	618	Barrio	22243
dtype: int64		dtype: int64	

Luego de este cambio se observa como cambiaron los valores NaN en las columnas 'Comuna' y 'Barrio' ya que estas columnas eran las que más campos con 'Sin Inf' y sus variantes tenían.

Se realiza mediante consulta la cantidad de datos nulos en las columnas con el siguiente resultado:

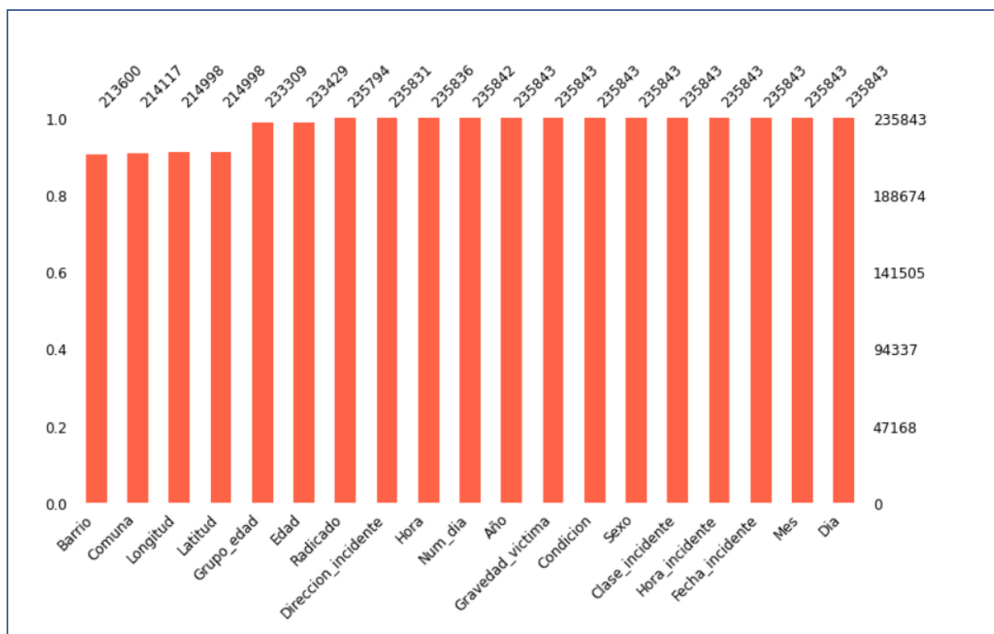
```
Cantidad de valores nulos en Direccion_incidente = 12/235843
Cantidad de valores nulos en Edad = 2414/235843
Cantidad de valores nulos en Num_dia = 1/235843
Cantidad de valores nulos en Hora = 7/235843
Cantidad de valores nulos en Grupo_edad = 2534/235843
Cantidad de valores nulos en Radicado = 49/235843
Cantidad de valores nulos en Latitud = 20845/235843
Cantidad de valores nulos en Longitud = 20845/235843
Cantidad de valores nulos en Comuna = 21726/235843
Cantidad de valores nulos en Barrio = 22243/235843
```

La proporción que representan esos datos en nuestro dataset es de:

```
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
```

Las variables 'Direccion_incidente', 'Num_dia', 'Hora', 'Radicado', 'Edad' y 'Grupo_edad', poseen muy poco porcentaje de valores nulos.

En el gráfico se muestra los datos nulos de las columnas (Direccion_incidente, Edad, Num_dia, Hora, Grupo_edad ,Radicado, Latitud, Longitud, Comuna, Barrio) en orden descendente:



Lo que se detalla a continuación, son algunas de las acciones ejecutadas para la corrección de los datos atípicos.

Se hace con código Python la unificación de los valores correspondientes como se muestra a continuación:

```
# Debemos unificar los registros con el valor 'Sep' con los de valor 'Sept'. Aparecerá sólo 'Sept'
invia_df.Mes = invia_df.Mes.replace(['Sep'], 'Sept')
print('\n Mes = {}'.format(invia_df.Mes.unique()))
```

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

```
# En este caso tenemos que corregir el 'oct-19' y pasar todos los registros a '10-19'
invia_df.Grupo_edad = invia_df.Grupo_edad.replace(['oct-19'], '10 - 19')
print('Grupo_edad = {}'.format(invia_df.Grupo_edad.unique()))
```

```
Grupo_edad = ['10 - 19' '20 - 29' '30 - 39' '40 - 49' '0 - 9' '50 - 59' 'Sin Inf'
              '60 - 69' '70 - 79' '80 o más']
```

```
# En este caso tenemos que corregir el 'Sin inf' y pasar todos los registros a 'Sin Inf'
invia_df.Sexo = invia_df.Sexo.replace(['Sin inf'], 'Sin Inf')
print('Sexo = {}'.format(invia_df.Sexo.unique()))
```

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



```
# En este caso tenemos que corregir el 'Acompañante de motocicleta' y pasar todos los registros a 'Acompañante de Motocicleta'
invia_df.Condicion = invia_df.Condicion.replace(['Acompañante de motocicleta'], 'Acompañante de Motocicleta')
print('Condicion = {}'.format(invia_df.Condicion.unique()))
```

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

```
invia_df.Edad = invia_df.Edad.replace(['30-35'], '33')
invia_df.Edad = invia_df.Edad.replace(['45-50'], '47')
invia_df.Edad = invia_df.Edad.replace(['30 - 35'], '33')
invia_df.Edad = invia_df.Edad.replace(['20 - 29'], '25')
```

En el caso de la columna 'Edad' había valores que tenían un rango de edades, entonces el criterio que se tomó para reemplazar esos valores fue tomar algún valor significativo contenido dentro de ese rango cargado originalmente.

Las coordenadas de dónde se obtuvo el conjunto de datos es Medellín, Colombia, cuyas coordenadas geográficas son: Latitud: 6.217 - Longitud: -75.567 (Latitud: 6° 13' 1" Norte Longitud: 75° 34' 1" Oeste).

Se modificaron los valores de Latitud - Longitud que estaban invertidos, según se muestra a continuación:

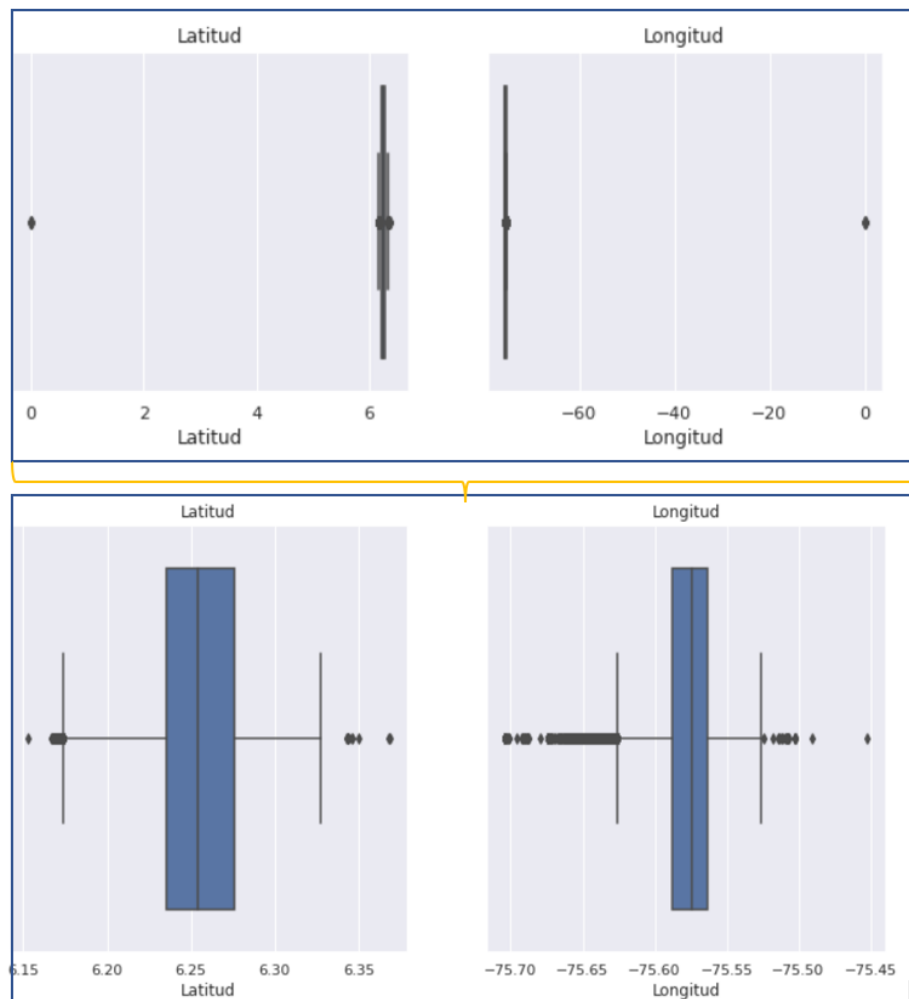
	Latitud	Longitud		Latitud	Longitud
0	6.266915	-75.559099	0	6.266915	-75.559099
1	6.289353	-75.553292	1	6.289353	-75.553292
2	6.289353	-75.553292	2	6.289353	-75.553292
3	6.234327	-75.607611	3	6.234327	-75.607611
4	6.234327	-75.607611	4	6.234327	-75.607611
...
235838	-75.536311	6.234267	→	235838	6.234267 -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

En total había 1949 registros invertidos en dichas columnas.

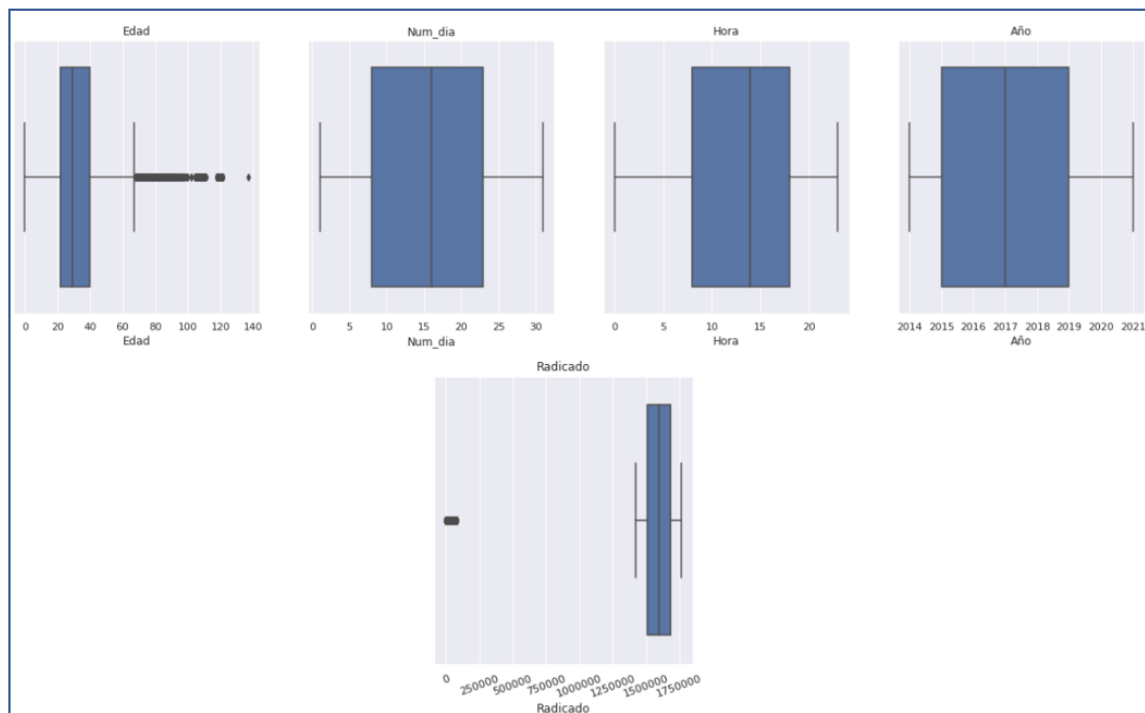
Se notará también que hay 3 registros con valor cero tanto en 'Longitud' como en 'Latitud', que a su vez tienen nulo 'Comuna' y 'Barrio'. Los eliminaremos ya que no son representativos y, siendo las coordenadas datos muy puntuales, no nos parece apropiado ni fácil de imputar, ya que no poseemos conocimiento sobre el relevamiento de estos datos.

Gravedad_victima	Fecha_incidente	Hora_incidente	Clase_incidente	Direccion_incidente	Sexo	Edad	Condicion	Mes	Dia	Num_dia	Hora	Grupo_edad	Año	Radicado	Latitud	Longitud	Comuna	Barrio	
227498	Heridos	2021-06-29	23:25:00	Choque	DG 80 CR 78	M	20.0	Motociclista	Jun	Mar	29.0	23.0	20 - 29	2021	1752580.0	0.0	0.0	NaN	NaN
227530	Heridos	2021-06-30	13:15:00	Choque	CL 12 CR 49	M	41.0	Motociclista	Jun	Miê	30.0	13.0	40 - 49	2021	1752484.0	0.0	0.0	NaN	NaN
227531	Heridos	2021-06-30	13:50:00	Choque	CR 63 CL 72	M	36.0	Motociclista	Jun	Miê	30.0	13.0	30 - 39	2021	1752585.0	0.0	0.0	NaN	NaN

A través de las siguientes gráficas boxplot se puede visualizar la irregularidad de los datos y como quedan los valores unificados luego de la estandarización de los mismos.



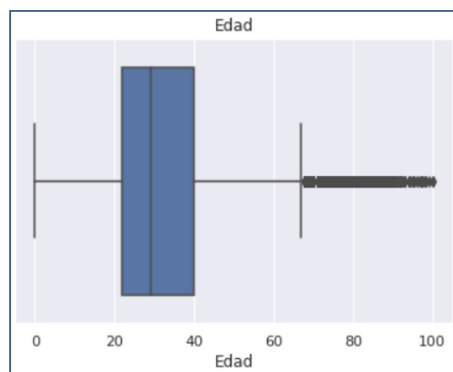
Usamos boxplot para detectar posibles outliers, o valores atípicos de nuestras variables numéricas, quedando los mismos de la siguiente manera:



Radicado si bien tiene valores extremos no la acotaremos ya que es solamente un identificador del incidente.

Las otras columnas numéricas no tienen valores extremos.

En la columna 'Edad' se acotó la edad a menores de 100 años quedando la distribución de edad como se muestra:



Son 556 registros mayores a 100 años, es un porcentaje muy reducido, lo cual nos parece conveniente ya que deseamos adoptar un enfoque conservador para la eliminación de los datos.

En la columna 'Sexo' dichos valores no queremos convertirlo a nulos, pues, puede pertenecer a otro género o no se completó el campo. La cantidad de estos registros son:

M	161603
F	71763
Sin Inf	2477

Análisis de la columna 'Radicado'

Representa el número de Expediente que se genera por cada accidente y puede ser 1 o varios para identificar los damnificados del mismo, por lo tanto, es un valor duplicado cuando intervienen por ejemplo dos actores sea auto-auto, auto-motocicleta, motocicleta-motocicleta entre otros.

La cantidad de registros únicos es de : 186954.

```
invia_copy_df.Radicado.value_counts()
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, Length: 186954, dtype: int64
```

Se muestra por ejemplo, el detalle del número de Radicado 1656780.0, que está repetido 34 veces y se presume de acuerdo a un análisis visual, de alguno de sus campos, que debe ser el volcamiento de un colectivo o similar porque hay gran cantidad de pasajeros heridos y un conductor.

MENTORÍA 08- VÍCTIMA DE INCIDENTES VIALES

Gravedad_victima	Fecha_incidente	Hora_incidente	Clase_incidente	Direccion_incidente	Sexo	Edad	Condicion	Mes	Dia	Num_dia	Hora	G
159519	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	37.0	Pasajero	Ene	Mié	16.0	5.0
159520	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	31.0	Pasajero	Ene	Mié	16.0	5.0
159521	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	29.0	Pasajero	Ene	Mié	16.0	5.0
159522	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	27.0	Pasajero	Ene	Mié	16.0	5.0
159523	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	31.0	Pasajero	Ene	Mié	16.0	5.0
159524	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	16.0	Pasajero	Ene	Mié	16.0	5.0
159525	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	37.0	Pasajero	Ene	Mié	16.0	5.0
159526	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	15.0	Pasajero	Ene	Mié	16.0	5.0
159527	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	44.0	Pasajero	Ene	Mié	16.0	5.0
159528	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	56.0	Pasajero	Ene	Mié	16.0	5.0
159529	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	51.0	Pasajero	Ene	Mié	16.0	5.0
159530	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	48.0	Pasajero	Ene	Mié	16.0	5.0
159531	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	43.0	Pasajero	Ene	Mié	16.0	5.0
159532	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	34.0	Pasajero	Ene	Mié	16.0	5.0
159533	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	39.0	Pasajero	Ene	Mié	16.0	5.0
159534	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	24.0	Pasajero	Ene	Mié	16.0	5.0
159535	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	49.0	Pasajero	Ene	Mié	16.0	5.0
159536	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	30.0	Conductor	Ene	Mié	16.0	5.0
159537	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	16.0	Pasajero	Ene	Mié	16.0	5.0
159538	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	62.0	Pasajero	Ene	Mié	16.0	5.0
159539	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	40.0	Pasajero	Ene	Mié	16.0	5.0
159540	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	59.0	Pasajero	Ene	Mié	16.0	5.0
159541	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	24.0	Pasajero	Ene	Mié	16.0	5.0
159542	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	F	20.0	Pasajero	Ene	Mié	16.0	5.0
159543	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	24.0	Pasajero	Ene	Mié	16.0	5.0
159544	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	29.0	Pasajero	Ene	Mié	16.0	5.0
159545	Heridos	2019-01-16	05:10:00	Volcamiento	CR 28 CL 107	M	46.0	Pasajero	Ene	Mié	16.0	5.0

Análisis de 'Sexo' en la 'Condicion' que refiera a una persona a cargo de un vehículo

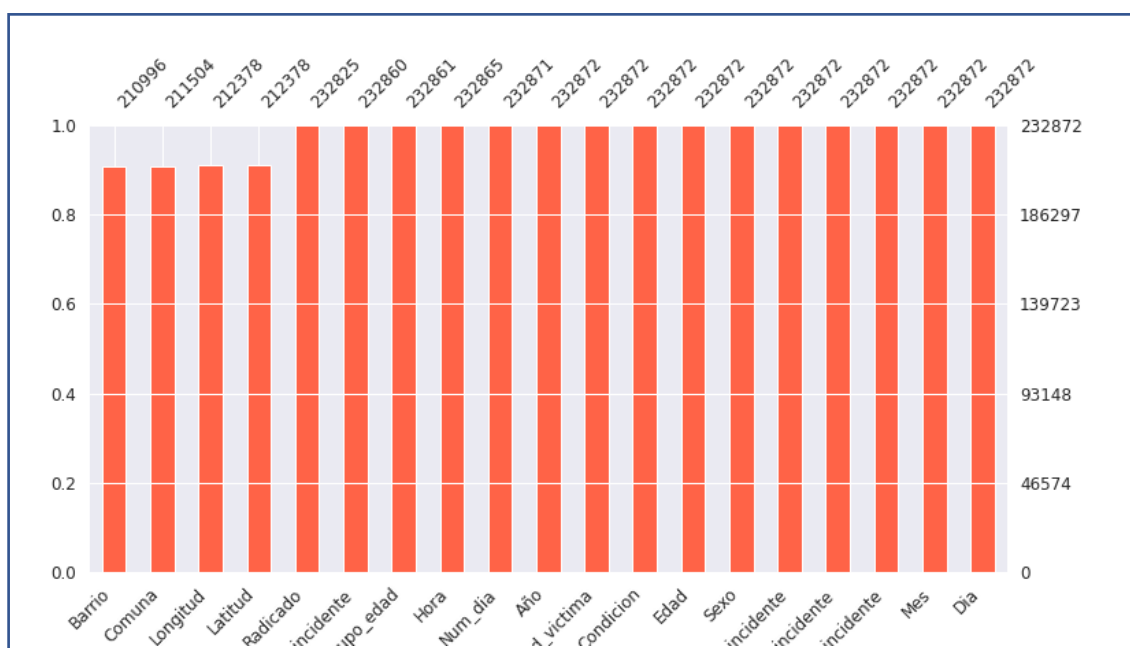
Se consulta la cantidad de 'M', 'F' y 'Sin Inf':

M	122210
F	29708
Sin Inf	1844

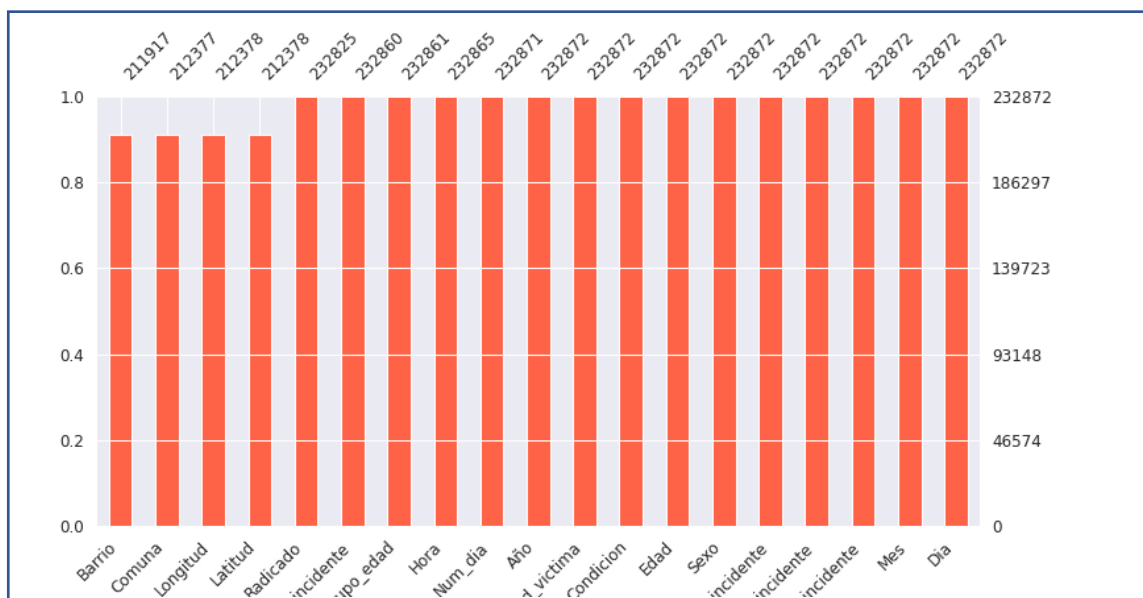
Imputación de Barrio y Comuna a partir de los valores no nulos de Latitud y Longitud

Considerando las modificaciones que se mencionaron anteriormente en las columnas '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.

El gráfico siguiente muestra los datos con los valores nulos iniciales.



Aquí se visualizan los valores nulos con los datos ya curados.



Comparando las visualizaciones podemos notar que Comuna y Barrio presentan menores datos nulos que inicialmente.