CLUSTERING MEDIANTE TECNICA DE APRENDIZAJE NO SUPERVISADO K-MEANS PARA SEGMENTAR LOS HECHOS DE TRÁNSITO QUE REGISTRA LA SSC (SECRETARIA DE SEGURIDAD CIUDADANA) DENTRO DE LA CDMX

Minería de Datos

Chavez Clavellina Angel Uriel

FASE 1. Entendimiento del caso

Planteamiento del Problema:

No es ningún misterio saber que las grandes ciudades, son las que albergan los mayores problemas de tránsito. Según el Universal, La CDMX es la segunda ciudad nacional y la número 22 en la escala global, que mas se ve afectada por ello.

Una de las razones por las que estos problemas han aumentado desde principios del 2022, es la pandemia ocurrida por COVID-19, dado el aislamiento al que todos nos vimos inmersos durante casi 2.5 años, se crearon ciertas necesidades sociales, tales como salir y/o transportarse.

La CDMX en el censo aplicado en el año 2020, registró una población de 9.2 millones, de la cual, hasta diciembre del 2022, se hizo un conteo de 35,883,179 automóviles, aproximadamente 4 autos por persona, dato sorprendente para el tamaño de la ciudad. Esto seguramente es consecuencia de tener que cubrir estas necesidades.

El aumento de esto implica que también aumenten los accidentes de tránsito, dado que ya hay mas personas en circulación. De Abril a Junio del 2022 se reportaron 22 mil incidentes, una alza del 26% respecto al año pasado. Diciembre es el mes que mas se dispara la cantidad de incidentes de tránsito.

La Comisión Nacional de Seguridad indica que las causas de los accidentes en las carreteras federales, alrededor del 80% de las veces se deben al conductor, 7% al vehículo, 9% a los agentes naturales y solo el 4% al camino

Los principales factores que provocan que causan accidentes carreteros

Factor Humano:

Conducir bajo los efectos del alcohol, medicinas y estupefacientes.

Realizar maniobras imprudentes y de omisión por parte del conductor, por ejemplo; no respetar los señalamientos viales.

Conducir a exceso de velocidad (produciendo vuelcos, salida del automóvil de la carretera, derrapes).

Salud física del conductor (ceguera, daltonismo, sordera).

Conducir con fatiga, cansancio o con sueño.

Factor Mecánico:

Vehículo en condiciones no adecuadas para su operación (sistemas averiados de frenos, eléctrico, dirección o suspensión).

Mantenimiento inadecuado del vehículo.

Factor Climatológico:

Niebla, humedad, derrumbes, zonas inestables, hundimientos.

Factor estructural de tránsito:

Errores de señalamientos viales.

Carreteras en mal estado o sin mantenimiento (baches, hoyos, pavimento deteriorado).

La falta de pintura y reflejantes en las líneas centrales y laterales de la carretera.

Justificación

De acuerdo con fuentes oficiales, los mas vulnerables en accidentes de tránsito, ni siquiera son los automovilistas que conducen, sino que se trata de peatones, ciclistas y motociclistas, considerados por autoridades y especialistas como usuarios vulnerables de la vía pública, por exponerse a los accidentes más que los propios automovilistas.

Sobre las víctimas más jóvenes, la subsecretaria de Planeación de la Secretaría de Movilidad, Laura Ballesteros, explica que las personas que más fallecen en accidentes de tránsito en la capital son quienes tienen entre 5-30 años de edad, además de que una gran cantidad de personas que viven con alguna discapacidad, la obtuvieron a causa de incidentes viales.

Conocer las características que suelen tener, podría ser una ventaja social bastante interesante, para el diseño de políticas de prevención que impacten positivamente a la reducción de accidentes de tránsito. Reconocer a la población más vulnerable segmentándolos

La idea es cuestionar las posibles causas que estén provocando que esto esté sucediendo, no cuestionar directamente los accidentes de tránsito.

Si logramos identificar a la población mas vulnerable tendríamos que relacionar las posibles causas que estén provocando que esta población sea la mas vulnerable respecto a accidentes de tránsito.

Hipótesis:

Los accidentes de tránsito pueden segmentarse de acuerdo con las características en común que presentan, identificando a la población mas vulnerable.

FASE 2. Entendimiento de los datos.

Recopilación de los datos

La información acerca de accidentes de Tránsito es registrada por la Secretaría de Seguridad Ciudadana de la CDMX (SSC), datos recopilados desde el año 2018 hasta la última actualización en abril del 2023. A partir de Julio del 2018 se enriqueció la recolección de información, llevando registros en dos series de datos.

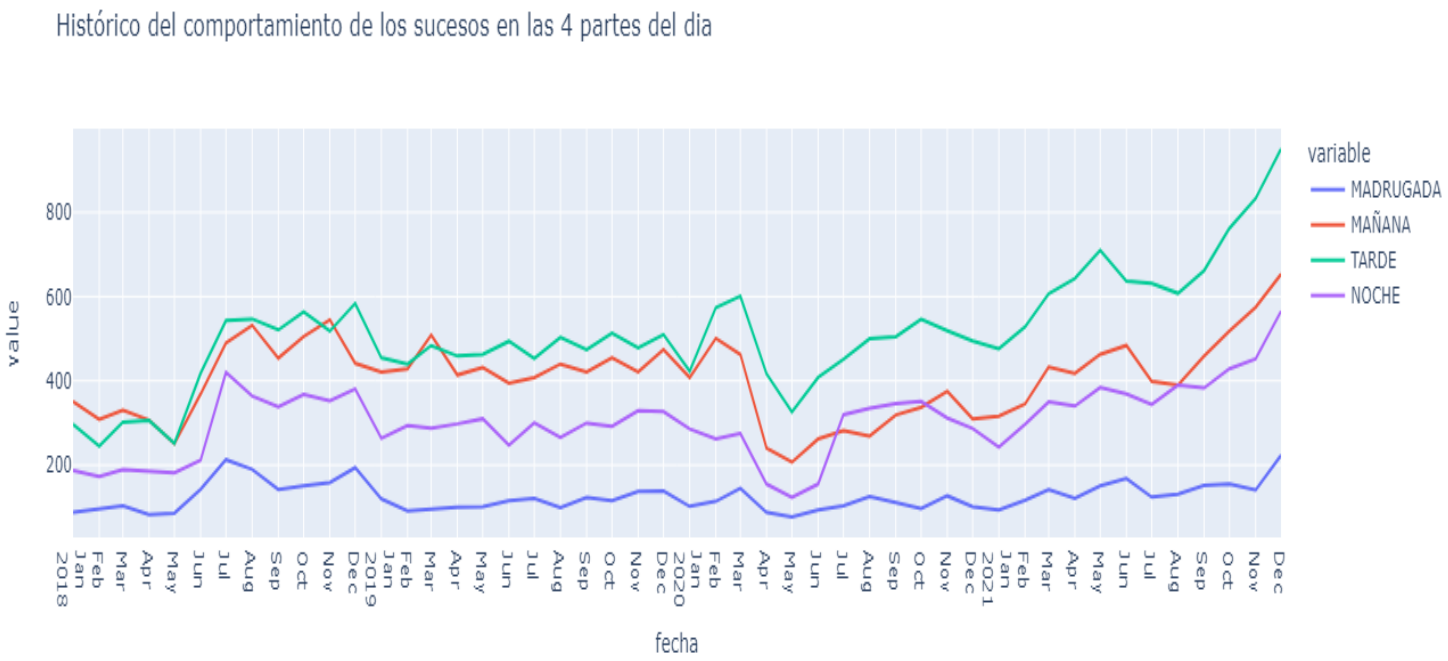
Esta información recopila las características asociadas a los accidentes de tránsito ocurridos en todas las alcaldías de la CDMX respondiendo a las preguntas; ¿Qué?, ¿Cómo?, ¿Cuándo?, ¿Por qué?, ¿Cuántos?, etc.

Descripción de los datos:

- mes: mes en el que ocurrió el suceso.
- hora: hora del día en el que sucedió el suceso.
- dia: dia en el que ocurrieron los hechos.
- tipo_de_evento: Hecho de transito que sucedió.
- punto_1, punto_2, colonia, alcaldía: Localización de donde suceden los hechos.
- tipo, color y marca del vehículo: Características del transporte asociado.
- interseccion_semaforizada, clasificacion_de_la_vialidad, sentido_de_circulacion: Características del accidente.
- Involucrado: Identifica al protagonista del accidente.
- Total, de Lesionados: Cantidad de Lesionados por Involucrado.
- Total, de Occisos: Cantidad de Occisos por Involucrado.

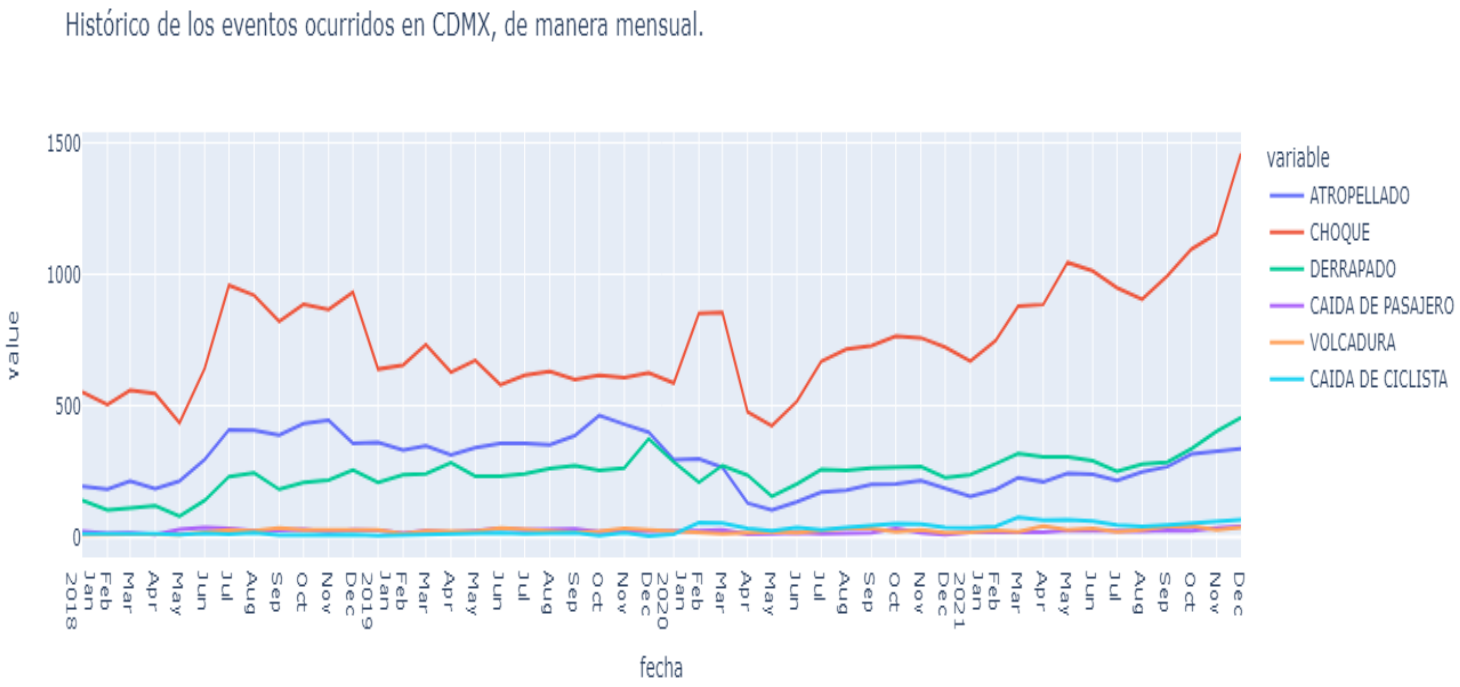
Exploración de los datos

Para la exploración de datos, se utilizó la librería de Python ‘seaborn’, permite crear gráficas y criterios visuales, especificando ciertos parámetros.



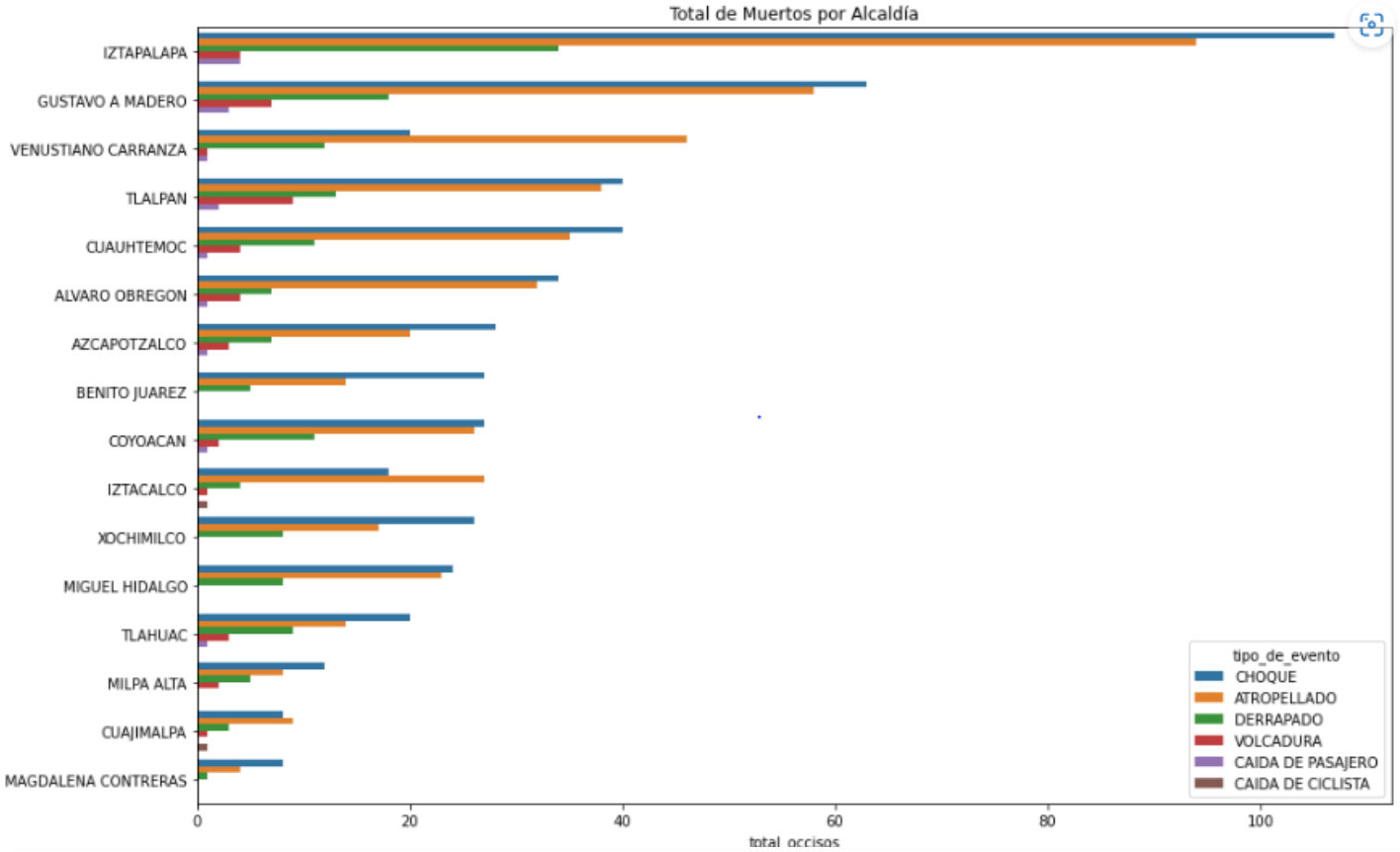
Se puede observar que de manera general hay una distribución muy homogénea de cómo se comportan los sucesos en las 4 partes del dia, por las madrugadas es cuando menos accidentes ocurren entre la 1 y 7 am, suelen ser bajos los sucesos activos.

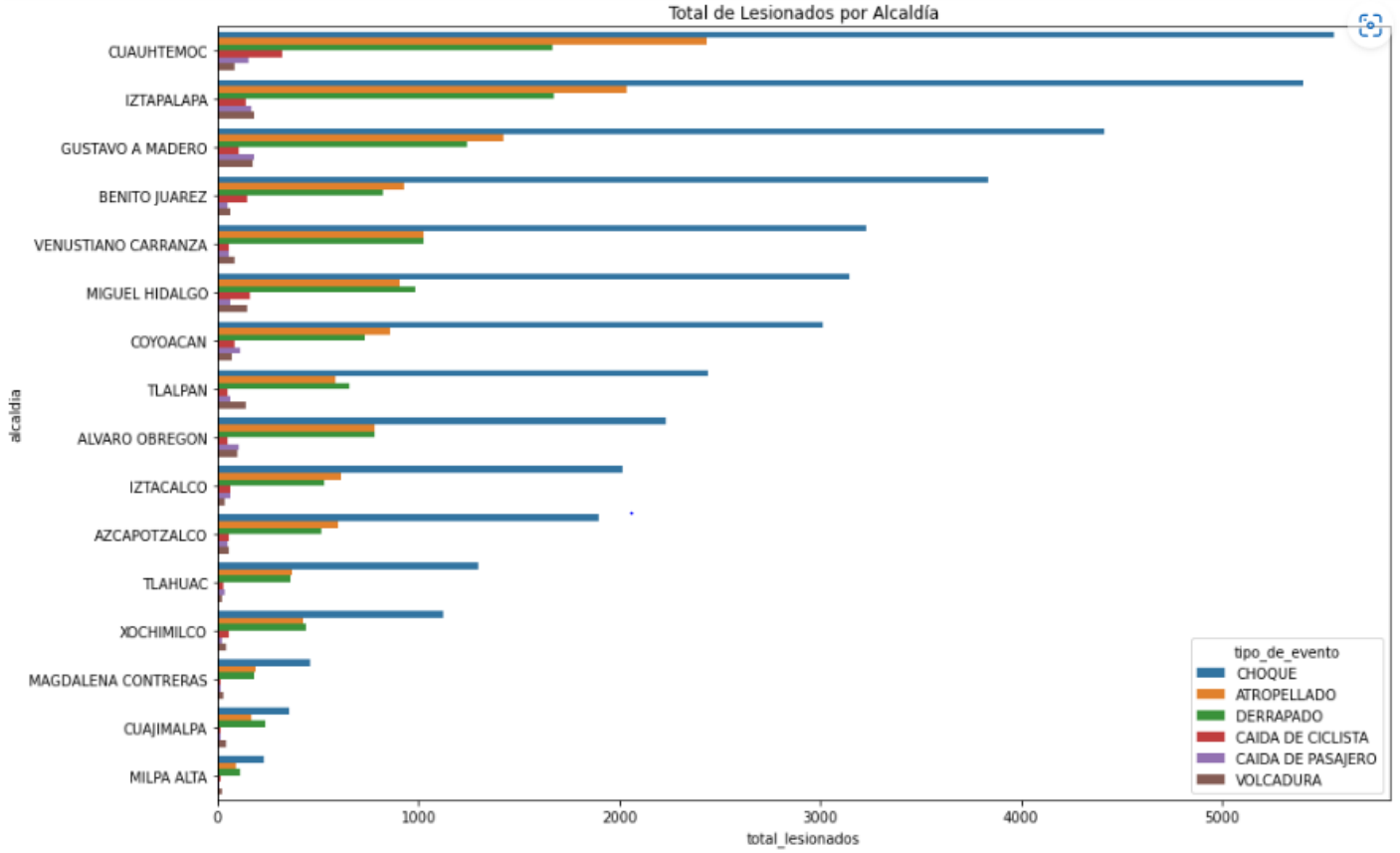
De manera histórica suelen haber más accidentes por las tardes, su comportamiento es muy homogéneo también, aunque en el último trimestre del 2022 hubo un alza en las 4 partes del dia, a comparación de otros periodos.

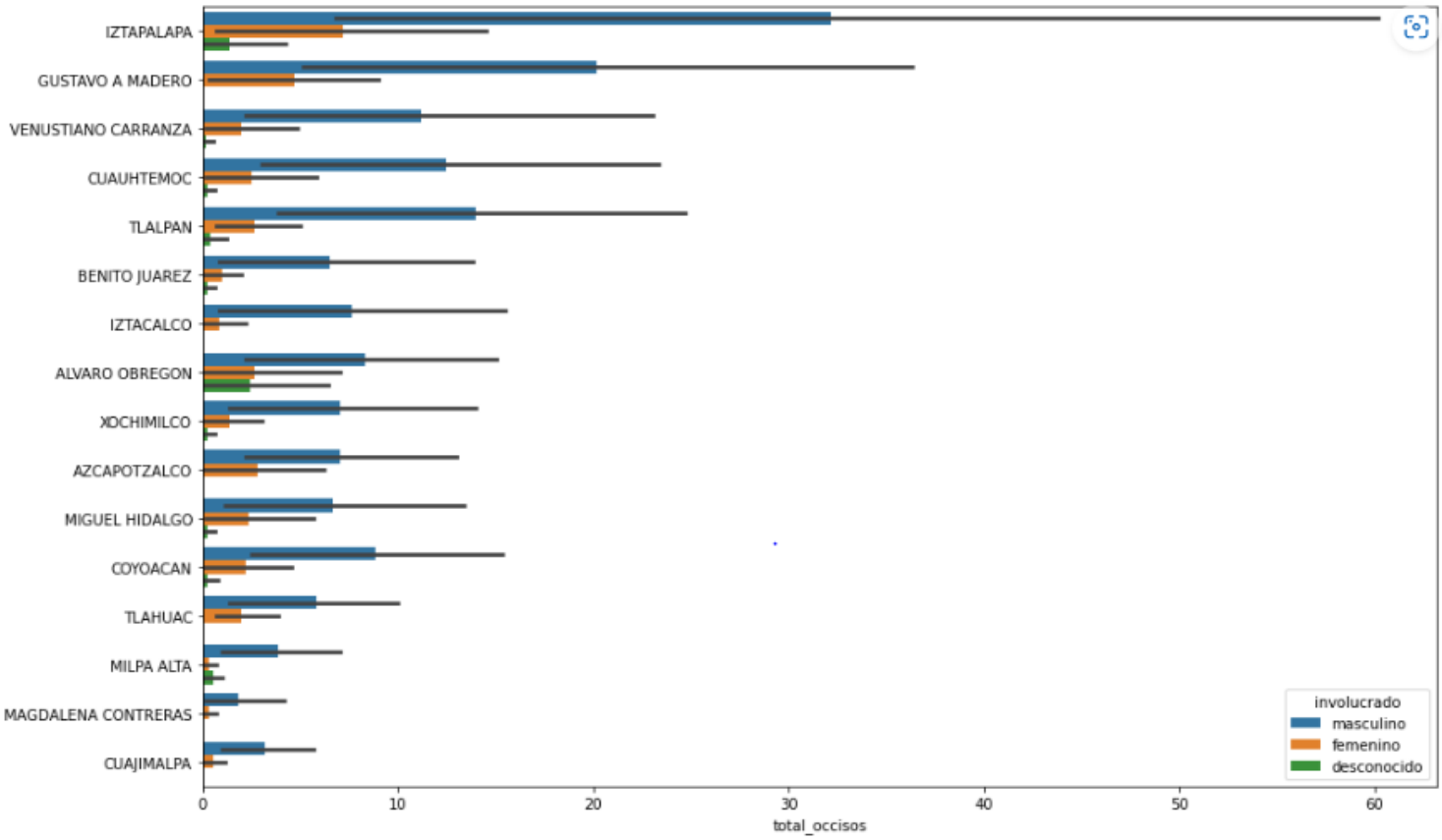


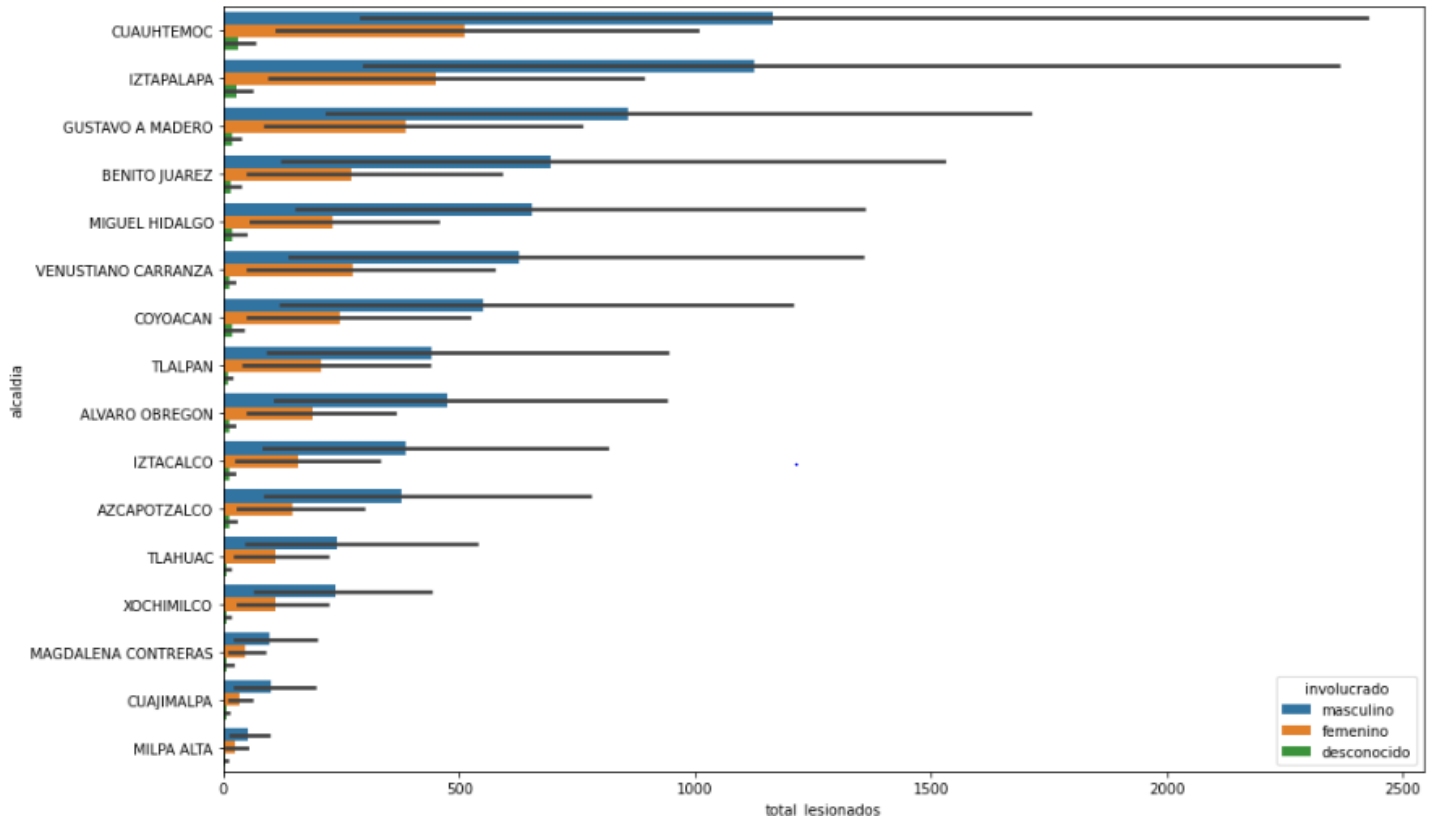
Se observa que, por encima de todos, los CHOQUES ocurren en mayor medida, cada uno de los eventos ha seguido las mismas tendencias, con pequeñas excepciones donde hay pequeños o grandes aumentos.

Se destaca que los CHOQUES ocurrieron en mayor proporción en el último trimestre del 2022 respecto a su histórico, pero no es un comportamiento cíclico, en años anteriores en el mismo trimestre nunca se había visto este aumento tan significativo, los DERRAPOS y los ATROPPELLOS también tuvieron un comportamiento similar en este último trimestre del 2022.


Total de Muertos por Alcaldía
IZTAPALAPA
GUSTAVO A MADERO
VENUSTIANO CARRANZA
TLALPAN
CUAUHTEMOC
ALVARO OBREGON
AZCAPOTZALCO
BENITO JUAREZ
COYOACAN
IZTACALCO
XOCHIMILCO
MIGUEL HIDALGO
TLAHUAC
MILPA ALTA
CUAJIMALPA
MAGDALENA CONTRERAS
0
20
40
60
80
100
total_occisos
tipo_de_evento
CHOQUE
ATROPELLADO
DERRAPADO
VOLCADURA
CAIDA DE PASAJERO
CAIDA DE CICLISTA


Total de Lesionados por Alcaldía
CUAUHTEMOC
IZTAPALAPA
GUSTAVO A MADERO
BENITO JUAREZ
VENUSTIANO CARRANZA
MIGUEL HIDALGO
COYOACAN
TLALPAN
ALVARO OBREGON
IZTACALCO
AZCAPOTZALCO
TLAHUAC
XOCHIMILCO
MAGDALENA CONTRERAS
CUAJIMALPA
MILPA ALTA
alcaldia
0
1000
2000
3000
4000
5000
total_lesionados
tipo_de_evento
CHOQUE
ATROPELLADO
DERRAPADO
CAIDA DE CICLISTA
CAIDA DE PASAJERO
VOLCADURA


IZTAPALAPA
GUSTAVO A MADERO
VENUSTIANO CARRANZA
CUAUHTEMOC
TLALPAN
BENITO JUAREZ
IZTACALCO
ALVARO OBREGON
XOCHIMILCO
AZCAPOTZALCO
MIGUEL HIDALGO
COYOACAN
TLAHUAC
MILPA ALTA
MAGDALENA CONTRERAS
CUAJIMALPA
0
10
20
30
40
50
60
total_occisos
involucrado
masculino
femenino
desconocido


CUAUHTEMOC
IZTAPALAPA
GUSTAVO A MADERO
BENITO JUAREZ
MIGUEL HIDALGO
VENUSTIANO CARRANZA
COYOACAN
TLALPAN
ALVARO OBREGON
IZTACALCO
AZCAPOTZALCO
TLAHUAC
XOCHIMILCO
MAGDALENA CONTRERAS
CUAJIMALPA
MILPA ALTA
alcaldia
0
500
1000
1500
2000
2500
total_lesionados
involucrado
masculino
femenino
desconocido

Muertos:

- Tlalpan, Iztapalapa y Gustavo A. Madero son las alcaldías que más muertes presentan.

- Las muertes por atropellos sobrepasan en Venustiano Carranza e Iztacalco.

* Suelen morir más Hombres que Mujeres en todas las alcaldías.

Lesionados:

- Cuauhtémoc, Iztapalapa y Gustavo A. Madero son las alcaldías que más lesiones presentan.

- De manera general en todas alcaldías, ocurren más las lesiones por CHOQUES.

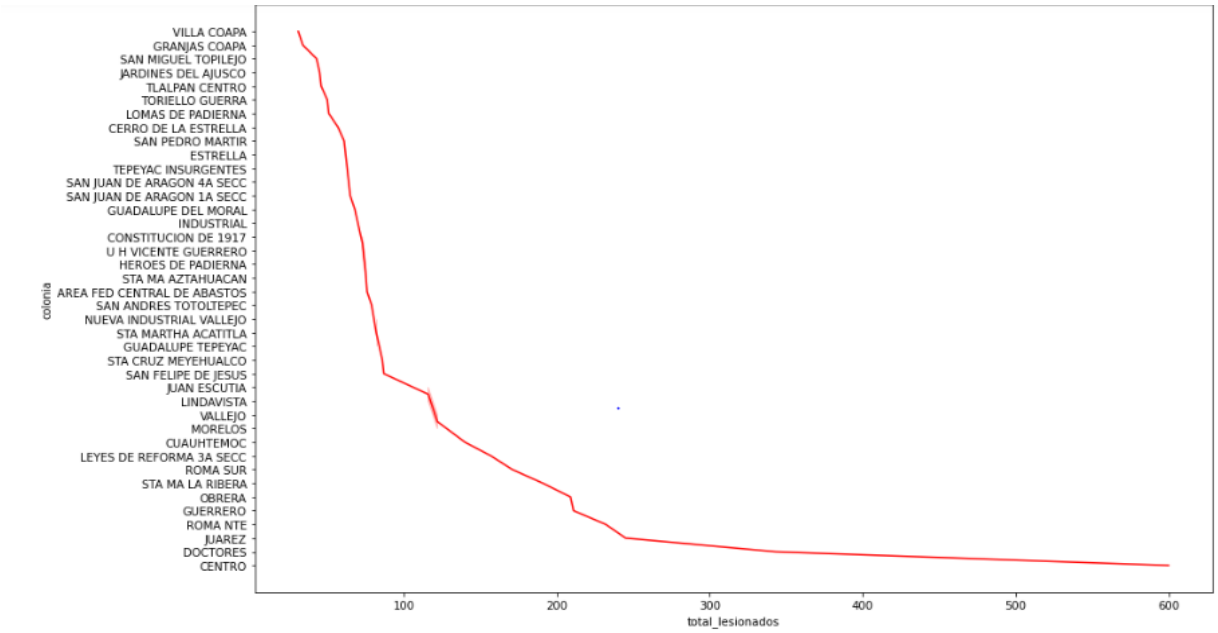
- Atropellos y Derrapos suelen ocurrir con la misma frecuencia.

- Caída de Ciclista, Volcadura y Caída de Pasajero ocurren de igual manera con la misma frecuencia y en menores proporciones

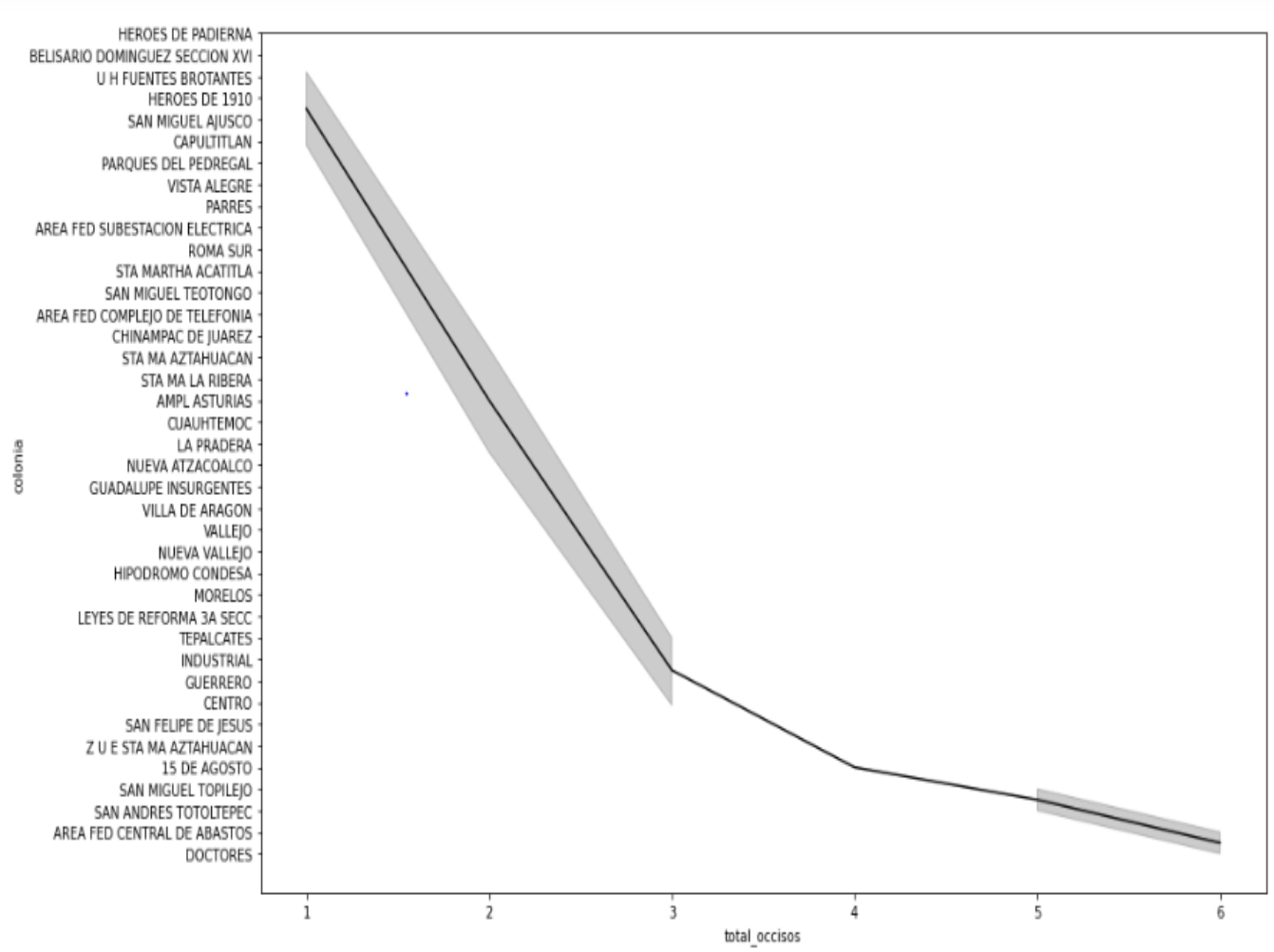
* Suelen lesionarse más Hombres que Mujeres en todas las alcaldías.

target 1 = ¿Cuáles son las condiciones en las que mueren y se lesionan más los Hombres por choques en Iztapalapa, Gustavo A. Madero, Tlalpan y Cuauhtémoc? **

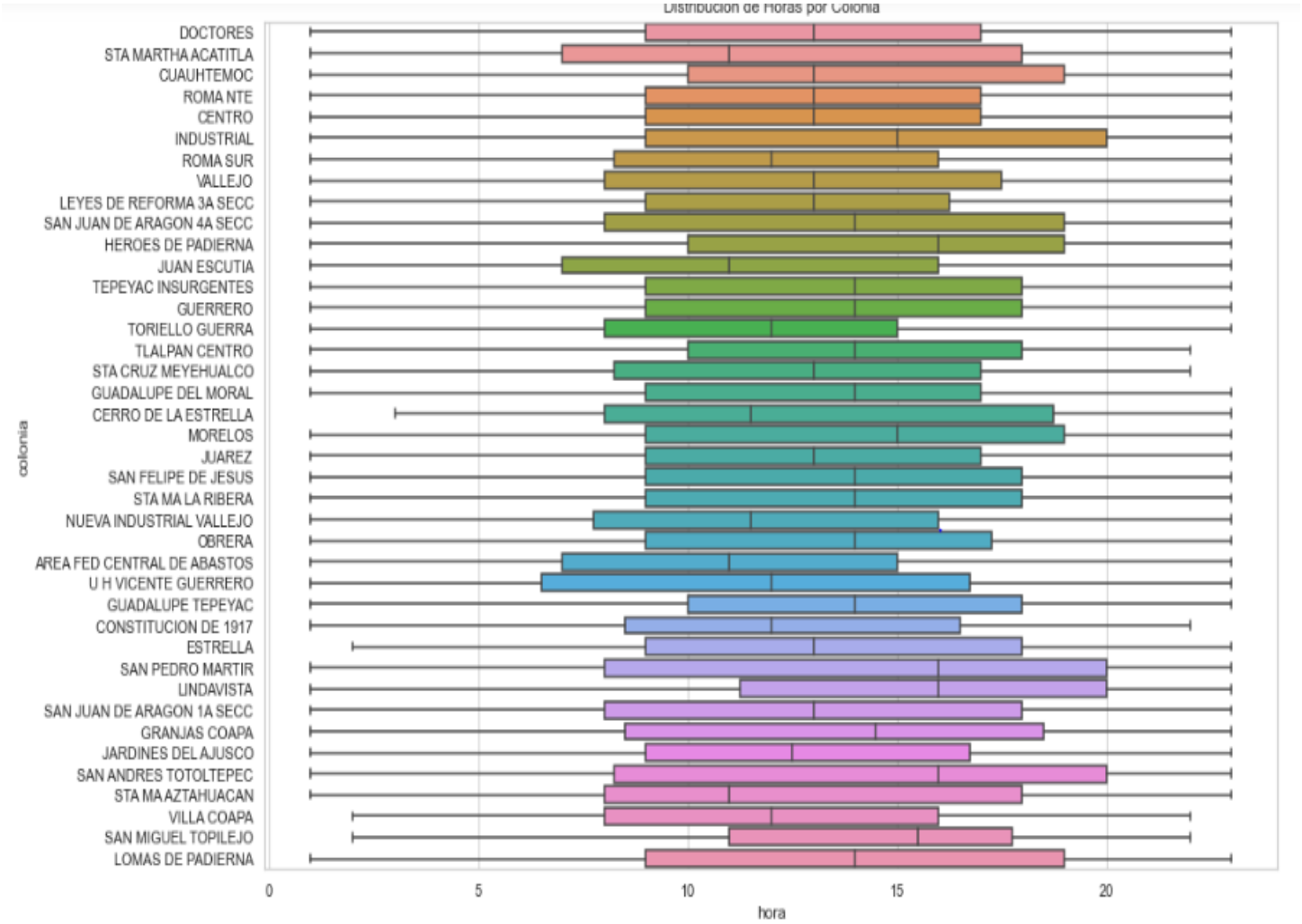
target 2 = ¿Cuáles son las condiciones por las que mueren más Hombres atropellados en Venustiano Carranza e Iztacalco? **



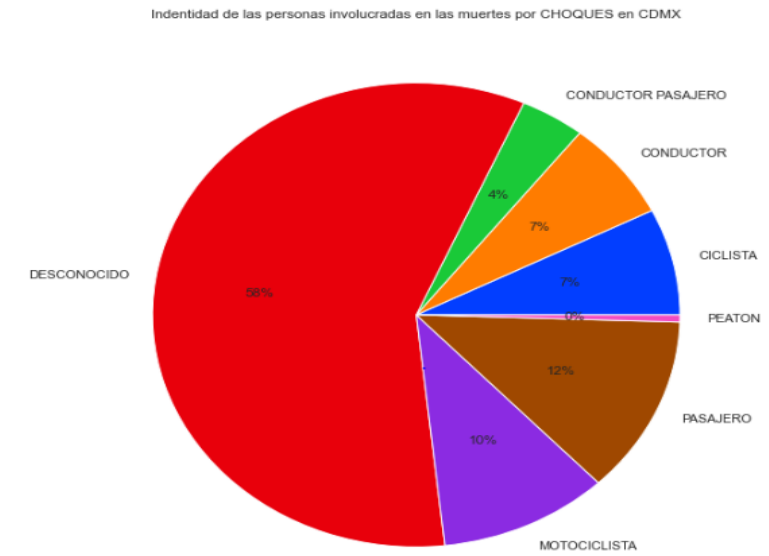
Juan Escutia, Roma Sur, Leyes de Reforma 3a Secc, Sta María la Ribera, Guerrero, Obrera, Roma Nte, Doctores y Centro son todas colonias donde los hombres se lesionan más por CHOQUES > 200.

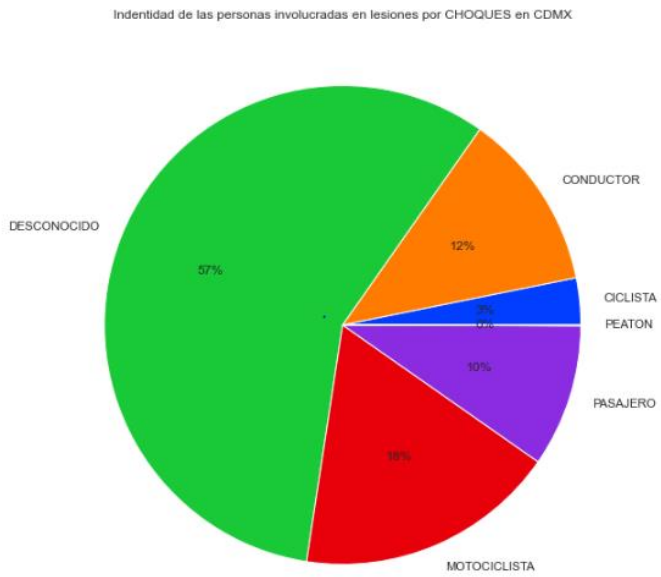


Leyes de Reforma 3a Sección, El Paraiso, Morelos, San Felipe de Jesús, Juan Escutia, Area Fed Central de Abastos, San Andrés Totoltepec, San Miguel Topilejo, Centro, Rustica Tlalpan y Doctores son las colonias en donde más hombres mueren a causa de un choque > 5.



Se observa que la hora en la que suceden los hechos, suelen ser muy similares en las colonias con mayor presencia Muertos y Lesiones por Choques, incluso tomando las horas mínimas y máximas. Todas poseen una mediana entre las 12 y las 16 horas ósea el 50% de los sucesos de cada alcaldía al menos ocurren entre casi las mismas horas.





Los Pasajeros son quienes más fallecen por CHOQUES en CDMX correspondiente a un 12%, seguido de los Motociclistas con un 10% y en tercer lugar los conductores con un 7%.

Los Motociclistas son las personas que más se lesionan por CHOQUES, correspondiente al 18% de lesiones, seguido de los Conductores con un 12% y al final los Pasajeros con un 10%.

FASE 3. Preparación de los datos.

Selección y Limpieza de los datos

La información a como es recopilada por la SSC viene de formas no adecuadas para poder hacer un modelo directamente, trae los nombres de las identidades asociadas a cada accidente en una misma columna, al igual que las edades, etc.

- Eliminación de variables nulas: Se descartaron aquellas variables con un porcentaje mayor al 70% de valores nulos.
- Eliminación de variables similares: Se descartaron las variables que desde el inicio aportaban poca o nada de información dado que ya se tenia otra similar.
- Eliminación de variables de tiempo: Se descartaron las variables relacionadas al año en que habían ocurrido los sucesos, para fines del clustering.
- Limpieza de variables:
  1. Tenemos una serie de variables donde se describe la cantidad de lesionados y occisos por accidente, pero no es la forma adecuada para poder hacer un análisis, la idea es reducir estas columnas y solo tener dos; 'Total_occisos y 'Total_lesionados.
  2. Variable del involucrado/a, explícitamente no viene una variable que identifique si la persona lesionada o muerta es hombre o mujer, pero dada una serie de condiciones se puede deducir.
  3. Variable identidad relacionada al involucrado/a en el hecho de tránsito, los valores vienen de esta manera: "CONDUCTOR PASAJERO CICLISTA" sin hacer una distinción, la idea es separar estas etiquetas en etiquetas individuales dependiendo de quien sea la persona y la condición, Ejemplo: Femenino – Lesionado – Conductor o Masculino – Occiso – Motociclista, se iteró a través de un algoritmo en Python y dadas ciertas condiciones como el número de muertos por Hombre y Mujer.

Input:

Y	Z	AA	AB	AC	AD	AE	AF	AG
total_occisos	occisos_femeninos	occisos_masculinos	occiso_se_desconoce	total_lesionados	lesionados_femeninos	lesionados_masculinos	lesionado_se_desconoce	identidad
1	1	0	0	2	0	2	0	CONDUCTOR PASAJERO PASAJERO
1	0	1	0	3	0	3	0	CONDUCTOR PASAJERO PASAJERO F
3	1	1	1	4	0	0	4	CONDUCTOR PASAJERO PASAJERO
3	1	2	0	0	0	0	0	CONDUCTOR MOTOCICLISTA PASAJ
4	0	0	4	12	1	4	7	CONDUCTOR PASAJERO PASAJERO F
3	0	3	0	0	0	0	0	CONDUCTOR PASAJERO PASAJERO

Output:

	total_lesionados	total_occisos	involucrado	lesionado	occiso
60356	1	0	masculino	PEATON	DESCONOCIDO
50703	1	0	masculino	MOTOCICLISTA	DESCONOCIDO
7181	1	0	femenino	PEATON	DESCONOCIDO
7182	1	0	femenino	PEATON	DESCONOCIDO
61063	1	0	masculino	PEATON	DESCONOCIDO
...	...	...	...	...	...
60492	1	0	masculino	PASAJERO	DESCONOCIDO
82740	0	1	masculino	DESCONOCIDO	DESCONOCIDO
82475	1	0	masculino	DESCONOCIDO	DESCONOCIDO
71077	2	0	masculino	DESCONOCIDO	DESCONOCIDO
115050	0	1	desconocido	DESCONOCIDO	CONDUCTOR

4. Dado que el 80% de nuestras variables son categóricas, el 50% de ellas tienen valores nulos, como, por ejemplo, tipo y marca del vehículo 2 y 3, Si tratamos de eliminar todas las variables con valores nulos nos quedamos sin datos, para fines de este análisis se procedió a reemplazar en todas, estos valores Nulos por strings de 'DESCONOCIDO'.

Se hicieron filtros adecuados para quitar filas que no aportaban información, dependiendo de la hora en que habían ocurrido los hechos, había accidentes registrados tal que su hora era mayor a las 24 horas que dura un día.

Preprocesamiento de Datos

Todas nuestras variables son Categóricas a excepción del número de Lesionados y Occisos.

Una vez hecha la limpieza y selección de las variables oportunas se emplearán técnicas de preprocesamiento de datos.

- Ordinal Encoding: Los meses, Los días y las prioridades de cada accidente (ALTA, MEDIA, BAJA), dada su naturaleza pueden ordenarse secuencialmente.

	mes	mes_aux	dia	dia_aux	prioridad	prioridad_aux
60356	DICIEMBRE	12	DOMINGO	7	BAJA	1
50703	FEBRERO	2	MARTES	2	BAJA	1
7181	FEBRERO	2	MARTES	2	BAJA	1
7182	FEBRERO	2	MARTES	2	BAJA	1
61063	FEBRERO	2	MARTES	2	BAJA	1
...	...	...	...	...	...	...
60492	DICIEMBRE	12	LUNES	1	BAJA	1
82740	SEPTIEMBRE	9	VIERNES	5	ALTA	3
82475	SEPTIEMBRE	9	VIERNES	5	BAJA	1
71077	MARZO	3	VIERNES	5	BAJA	1
115050	MARZO	3	JUEVES	4	ALTA	3

- One Hot Encoding: Algunas variables como la alcaldía donde sucedió el accidente, el involucrado en el accidente, el tipo de evento y si fue un accidente en intersección semaforizada, les crearemos variables dummies

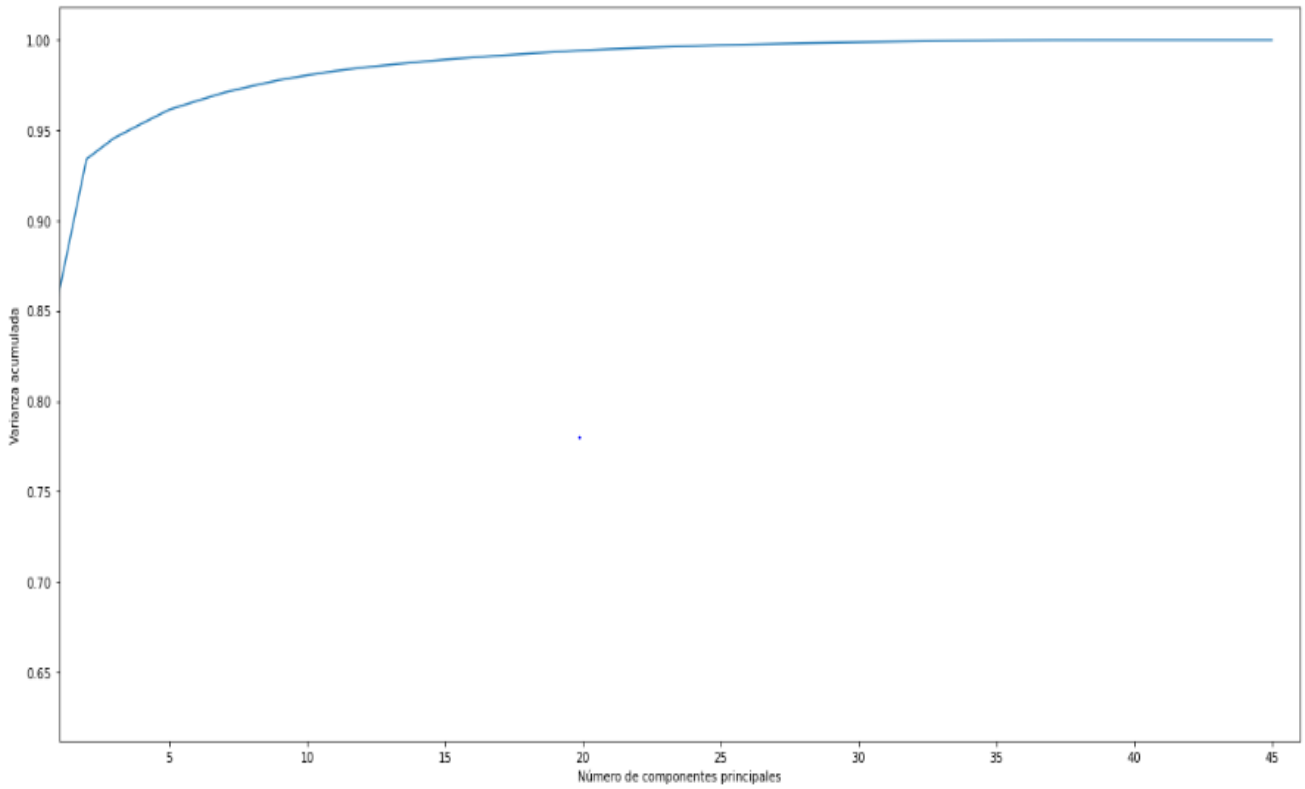
tipo_de_evento_ATROPELLADO	tipo_de_evento_CAIDA DE CICLISTA	tipo_de_evento_CAIDA DE PASAJERO	tipo_de_evento_CHOQUE
1	0	0	0
0	0	0	1
1	0	0	0
1	0	0	0
1	0	0	0
...	...	...	...
0	0	0	1
1	0	0	0
0	0	0	1
0	0	0	1
0	0	0	1

- Frecuency Encoding: Como no tenemos una variable dependiente de la cual podamos obtener promedios, etc, obtendremos la normalización de las frecuencias de algunas categorías, pues estamos trabajando también con variables dummie.

colonia	tipo_de_interseccion	unidad_medica_de_apoyo	occiso	lesionado	tipo_de_vehiculo_1	marca_de_vehiculo_1
PAULINO NAVARRO	CRUZ	PM	DESCONOCIDO	PEATON	SD	SD
JOSE MA PINO SUAREZ	Y	PC	DESCONOCIDO	MOTOCICLISTA	AUTOMOVIL	GOLF
GUADALUPE TEPEYAC	CRUZ	PC	DESCONOCIDO	PEATON	TAXI	TSURU
STA FE	T	CRUZ ROJA	DESCONOCIDO	PEATON	AUTOMOVIL	TSURU
ZACATEPEC	T	ERUM	DESCONOCIDO	PEATON	AUTOMOVIL	SD
...	...	...	...	...	...	...
SAN JUAN JOYA	T	ERUM	DESCONOCIDO	PASAJERO	AUTOMOVIL	POINTER
STA CRUZ AVIACION	CRUZ	CCO/SAMU	DESCONOCIDO	DESCONOCIDO	AUTOMOVIL	BEAT
SANTIAGO CENTRO	T	PC	DESCONOCIDO	DESCONOCIDO	MOTOCICLETA	SD
EJERCITO CONSTITUCIONALISTA	T	PC	DESCONOCIDO	DESCONOCIDO	MOTOCICLETA	SD
BOSQUE DE CHAPULTEPEC 1A SECC	RECTA	CCO/1ER CONTACTO	CONDUCTOR	DESCONOCIDO	AUTOMOVIL	FORD 150

colonia_frequency_encouded	tipo_de_interseccion_frequency_encouded	unidad_medica_de_apoyo_frequency_encouded	occiso_frequency_encouded
0.001019	0.505926	0.046145	0.991357
0.000648	0.039123	0.150935	0.991357
0.002901	0.505926	0.150935	0.991357
0.001960	0.332058	0.172588	0.991357
0.000355	0.332058	0.308908	0.991357
...	...	...	...
0.000293	0.332058	0.308908	0.991357
0.000340	0.505926	0.000864	0.991357
0.000432	0.332058	0.150935	0.991357
0.000216	0.332058	0.150935	0.991357
0.002732	0.056207	0.000015	0.000864

- Reducción de la dimensionalidad (ACP)
  Tenemos un dataframe con 46 columnas, entonces debemos de reducir la dimensionalidad para poder hacer mas eficiente la etapa del modelamiento, recordemos que con ACP, se busca el numero óptimo de componentes principales tal que posean la mejor varianza acumulada para representar al conjunto de datos, para esto ocupamos un método que nos identifica el número óptimo de PCA´s dada su varianza acum.



Se ha identificado que con solo 2 Componentes Principales obtenemos el 86% de la varianza explicada, para fines de este análisis y para hacer más fácil la visualización de los PCA´s nos quedaremos con ellos.

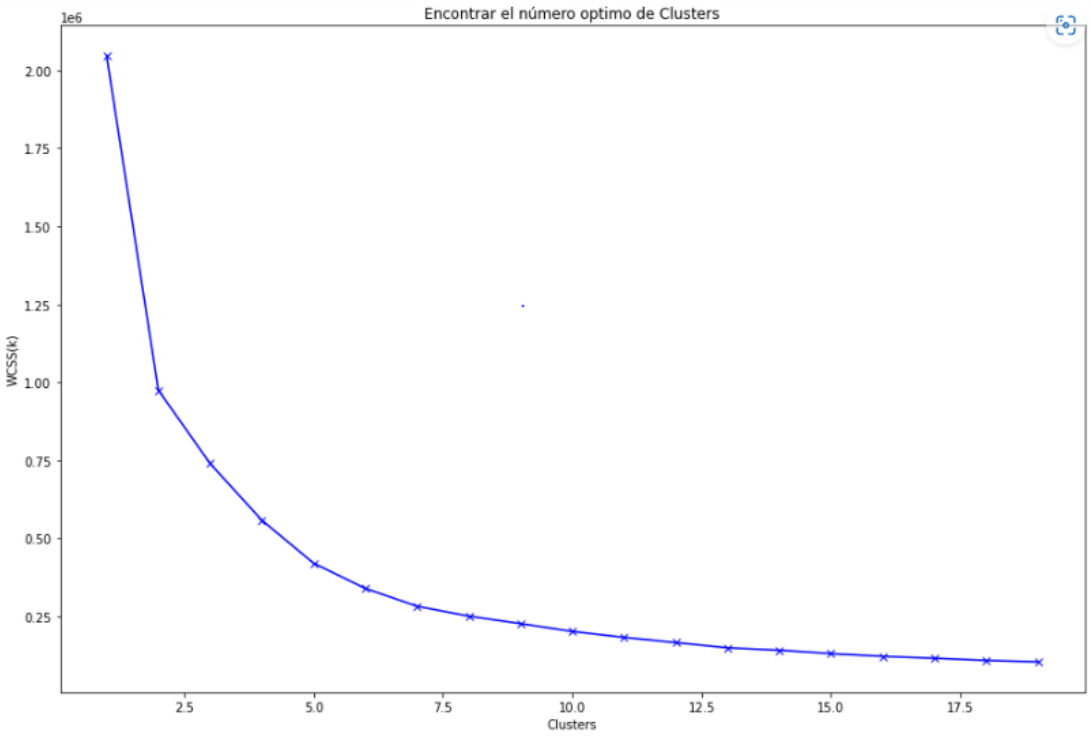
	PCA1	PCA2
0	9.268618	-5.292337
1	6.389935	4.772971
2	5.380567	4.790938
3	5.381919	4.789387
4	5.384572	4.790201
...	...	...
64791	3.171929	-5.164589
64792	9.297204	-2.281119
64793	-0.705469	-2.084013
64794	5.417478	3.792155
64795	11.411394	3.676169

FASE 4. Modelamiento

En esta fase se ocupó un modelo de Aprendizaje No Supervisado, Clustering, implementando el algoritmo más famoso, K-Means, implementado en un Notebook de Anaconda con la librería Scikit-Learn.

Dividiremos El dataframe en entrenamiento y validación

Y aplicando el método del CODO para encontrar el número óptimo de clústeres sobre el conjunto de entrenamiento obtenemos



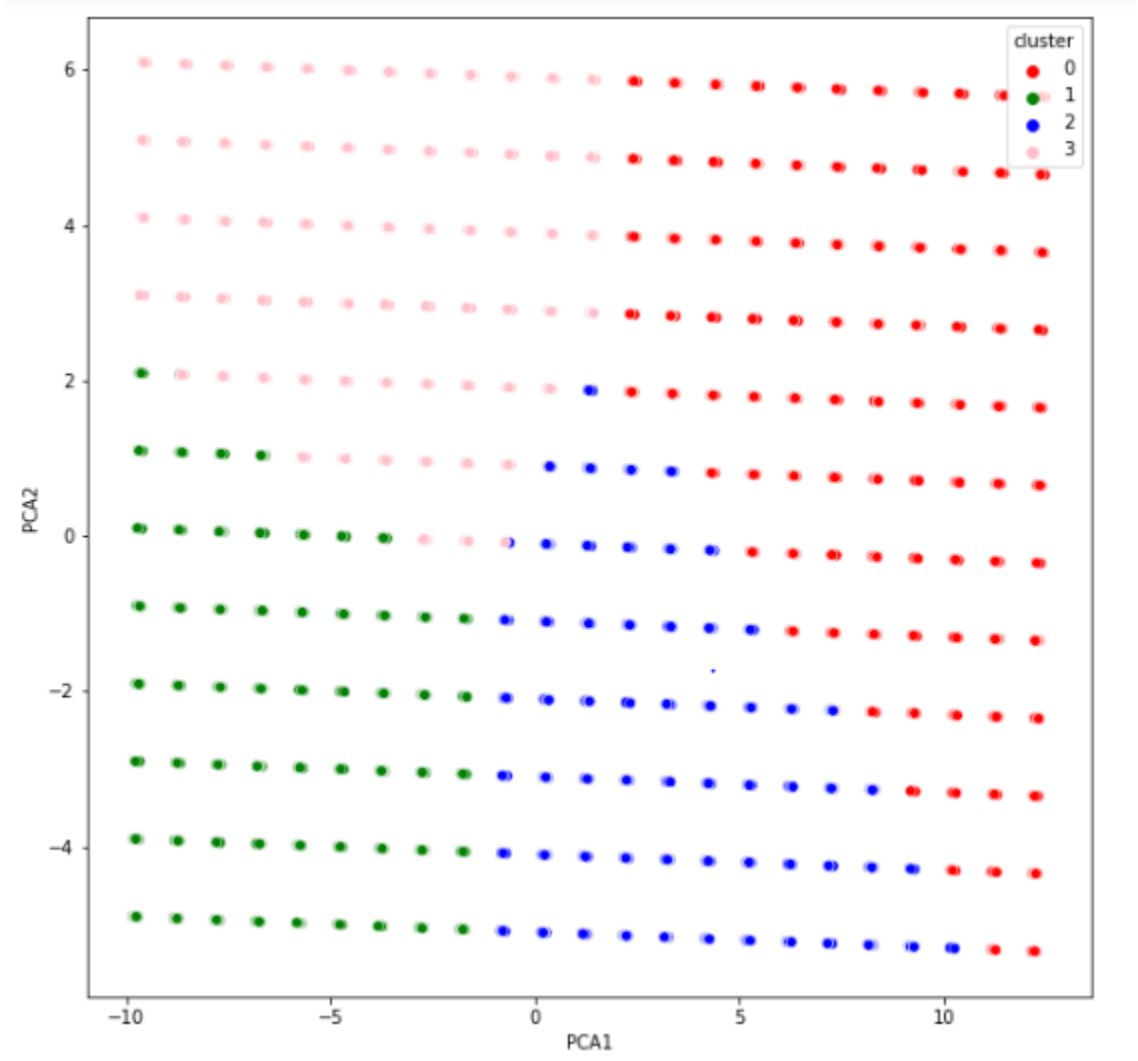
Se observa que en 4 Clústers es en donde cambia de dirección el codo, aunque desde el cluster 2 esta tendencia ya va decreciendo, tomaremos 4 Clústers.

Para cada uno de los 4 clúster identificamos sus centros y coordenadas respectivas.

	PCA1	PCA2
0	6.974376	2.052254
1	-5.727676	-2.429981
2	3.034960	-2.778213
3	-3.472294	3.469605

Visualizando los clústeres:

	PCA1	PCA2	cluster
0	9.268618	-5.292337	2
1	6.389935	4.772971	0
2	5.380567	4.790938	0
3	5.381919	4.789387	0
4	5.384572	4.790201	0
...	...	...	...
64791	3.171929	-5.164589	2
64792	9.297204	-2.281119	0
64793	-0.705469	-2.084013	2
64794	5.417478	3.792155	0
64795	11.411394	3.676169	0



FASE 5. Evaluación

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.4014	42510.1526	0.7907	0	0	0

El Coeficiente de Silhouette: 0.4, tenemos un modelo que relativamente esta agrupando correctamente.

Coeficiente de Davies: 0.8, es un indicador muy bajo, por ende nuestros clusters están bien compactos cuyos centros separados.

FASE 6. Validación

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.4006	18153.1630	0.7908	0	0	0

Los resultados son similares o los mismos para el conjunto de validación, por ende, es un buen modelo.

Dados estos resultados, podemos decir que el modelo obtenido no se ve afectado por la selección de los conjuntos de entrenamiento y validación.

BIBLIOGRAFIA

https://towardsdatascience.com/clustering-made-easy-with-pycaret-656316c0b080

https://datos.cdmx.gob.mx/dataset/hechos-de-transito-reportados-por-ssc-base-ampliada-no-comparativa/resource/3ea0519c-9690-4cfa-ab46-b84dccba5886

https://www.eluniversal.com.mx/metropoli/aumentan-los-accidentes-viales-en-la-cdmx/

https://www.animalpolitico.com/sociedad/accidentes-viales-diciembre-cdmx

Repositorio de GitHub: https://github.com/RiemanNClav/Hechos-de-Transito-registrados-por-la-SSC