DEPARTAMENTO DE MATEMÁTICAS Y FÍSICA

Ciencia de datos e inteligencia de negocios

Proyecto de aplicación: Manejo de Datos, Similitud y Clustering

Flavio Cesar Palacios Salas

ITESO, Universidad Jesuita de Guadalajara

Departamento de Matemáticas y Física

Ciencia de datos e inteligencia de negocios

Proyecto de aplicación: Manejo de Datos, Similitud y Clustering

Introducción

Actualmente, con los avances en las tecnologías de la información, la generación de datos de diversos tipos en un solo día tiene volúmenes muy altos y con tendencia creciente. Con el fin del aprovechamiento de la información valiosa que pueda estar oculta en los datos que se generan, se requieren de tener conocimientos básicos de manejo de información y de análisis exploratorio de datos.

De forma general, a menos que la persona sea una experta en el fenómeno en el cual se están generando los datos, el ingeniero que se disponga al análisis de los datos generados debe de realizar un análisis exploratorio para rescatar las características básicas que poseen los datos. Además de realizar un agrupamiento de los mismos datos en base a una característica de interés.

Objetivo:

El objetivo de este proyecto de aplicación se puede separar en tres fases:

1.- La limpieza y extracción de la información estadística básica que tienen los datos que se están analizando.

2.- Realización de un agrupamiento (“Clustering”) de los datos en base a una característica de interés.

3.- La obtención o formulación de conclusiones sobre el fenómeno del cual provienen los datos en base a los resultados de los análisis anteriores

Actividades

1. Obtención de una base de datos que fuera generada por un fenómeno de interés (La orientación o tema de los datos será especificada para cada equipo por el profesor).

2. Aplicar el estudio de calidad de los datos para determinar el tipo de datos, categorías e información estadística básica.
3. Realizar una limpieza de datos y obtener un análisis exploratorio de datos (EDA) donde se muestren gráficas y conclusiones acerca del análisis. Al menos obtener 5 insights.
4. En base al estudio anterior, realizar un análisis de similitud entre variables y entre muestras disponibles en su base de datos.
5. Crear agrupamientos o "clusters" basados en el algoritmo "hierarchical clustering" ó "Kmeans", y presentar sus resultados (si los datos lo permiten).
6. Basados en los análisis anteriores, formular conclusiones sobre la información importante que se haya logrado encontrar de los datos analizados.

Desarrollo

La base de datos como indica la descripción contiene accidentes registrados por el C4 un sistema mexicano que registra todos los incidentes de tráfico, en este caso en particular contienen datos entre 2017 y 2019 y fueron obtenidos a través de la siguiente liga: https://datos.cdmx.gob.mx/explore/dataset/incidentes-viales-c5/table/?disjunctive.incidente_c4

Además, contiene las siguientes explicaciones para cada una de las variables:

- folio: una identificación única para cada registro
- fecha_creacion: fecha de creación
- hora_creacion: hora de creación
- dia_semana: día de la semana en que ocurre el incidente
- codigo_cierre: clasificación interna. La columna puede contener los siguientes códigos.
    - R: Afirmativo, si el incidente es confirmado por el equipo de emergencias.
    - N: Negativo, si el equipo de emergencias no confirma el incidente en el punto de ubicación.
    - I: Informativo, en caso de que los equipos de atención quieran agregar información extra.
    - F: Falso, si el informe inicial no coincide con los eventos reales.
    - D: Duplicados, registros con código de cierre afirmativo, negativo o falso pero los operadores los identifican
- fecha_cierre: fecha de cierre, la fecha en la que se resolvió el incidente
- año_cierre: año de cierre
- mes_cierre: mes de cierre
- hora_cierre: tiempo de cierre
- delegacion_inicio: entidad dentro de la Ciudad de México donde se registró el incidente
- incidente_c4: una breve explicación sobre el incidente.
- latitud: latitud del accidente
- longitud: longitud del accidente
- clasconf_alarma: código que identifica la gravedad de la situación
- tipo_entrada: cómo se informó el incidente
- delegacion_cierre: entidad dentro de la Ciudad de México donde se cerró el incidente
- geopoint: columnas de latitud y longitud combinadas

- mes: mes en el que se informó el incidente

Data Quality Report

Se realizaron los Data Quality Report de cada una de las bases de datos por año con resultados como el siguiente:

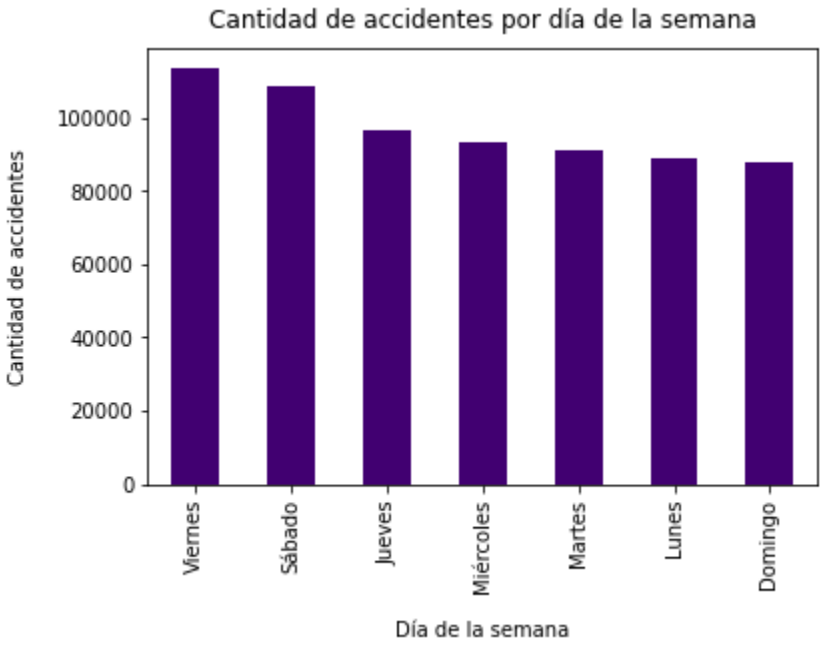
	Nombres	Data_Types	Missing_values	Present_values	Unique_values	Min	Max
folio	folio	object	0	207634	207634	AO/190101/03499	IZ/191130/09782
fecha_creacion	fecha_creacion	object	0	207634	475	01/01/2019	31/12/2018
hora_creacion	hora_creacion	object	0	207634	82567	00:00:00	9:59:58
dia_semana	dia_semana	object	0	207634	7	Domingo	Viernes
codigo_cierre	codigo_cierre	object	0	207634	5	(A) La unidad de atención a emergencias fue de...	(N) La unidad de atención a emergencias fue de...
fecha_cierre	fecha_cierre	object	0	207634	459	01/01/2019	31/10/2019
año_cierre	año_cierre	int64	0	207634	1	2019	2019
mes_cierre	mes_cierre	object	0	207634	11	Abril	Septiembre
hora_cierre	hora_cierre	object	0	207634	83826	00:00:01	9:59:55
delegacion_inicio	delegacion_inicio	object	19	207615	16	NaN	NaN
incidente_c4	incidente_c4	object	0	207634	23	Detención ciudadana-accidente automovilístico	sismo-persona atropellada
latitud	latitud	float64	0	207634	30216	19.0954	19.5786
longitud	longitud	float64	0	207634	29013	-99.3535	-98.9454
clas_con_f_alarma	clas_con_f_alarma	object	0	207634	4	DELITO	URGENCIAS MEDICAS
tipo_entrada	tipo_entrada	object	0	207634	6	BOTÓN DE AUXILIO	REDES
delegacion_cierre	delegacion_cierre	object	19	207615	16	NaN	NaN
geopoint	geopoint	object	0	207634	43484	19.095427,-99.209295	19.57857003,-99.13017996
mes	mes	int64	0	207634	11	1	11

Se puede comprobar que los datos de todos los años contienen las mismas columnas y que se pueden usar indistintamente los datos o inclusive se pueden combinar las tablas, una de las ventajas de los datos es que en la mayoría de estos hay apenas una pequeña cantidad de datos perdidos, alrededor de 500 lo que facilitará el análisis además de que se cuentan con aproximadamente 680,000 datos y no se perderá una parte considerable de la información, en algunos casos alguna de las columnas pueden contener información que puede ser repetitiva o innecesaria debido a que otra de las columnas lo contiene previamente como por ejemplo la columna de geopoint pues contiene los mismos datos que latitud y longitud.

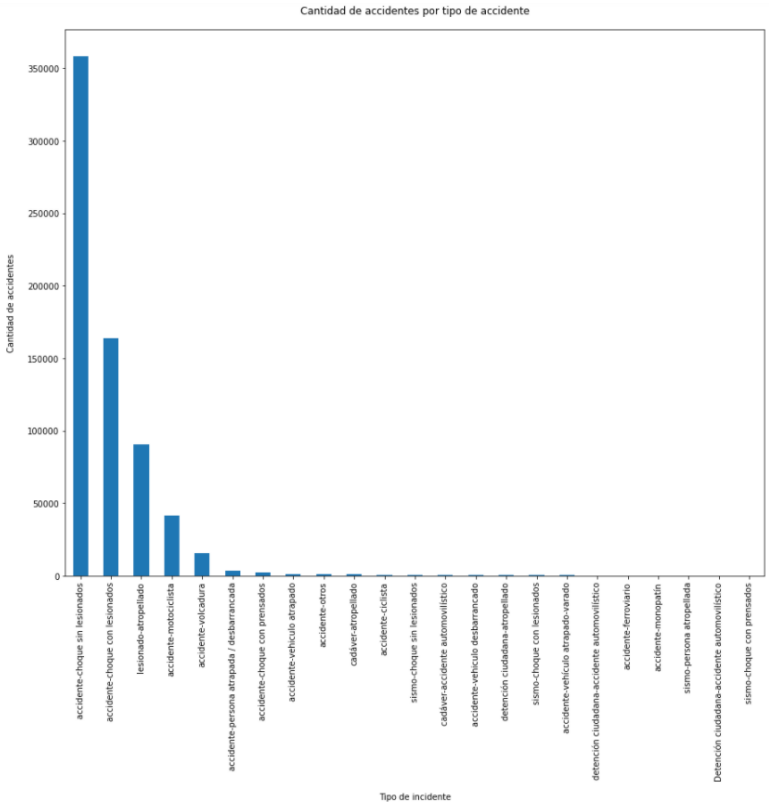
A continuación, se unieron los datasets para trabajar únicamente con una sola base de datos unida y se eliminó todas las variables que no fueran necesarias.

Dado que el dataset estaba en general limpio no se necesitó limpiar los datos demasiado a excepción de las variables que consideramos importantes de analizar o usar tanto para el EDA como los índices de similitud y el cluster.

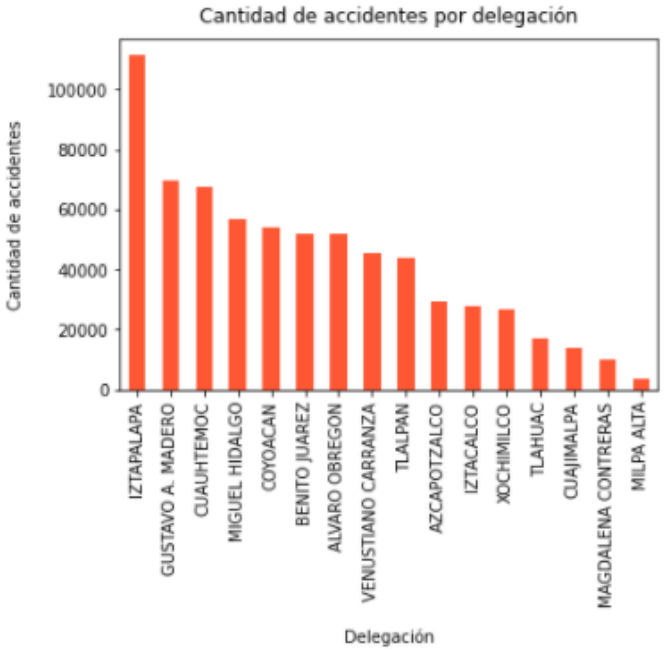
EDA Exploratory Data Analysis



Aquí se puede observar, como los días de la semana que más accidentes tiene son el viernes y el sábado. Se puede observar como a medida que la semana avanza (comienza en domingo), la cantidad de accidentes aumenta.

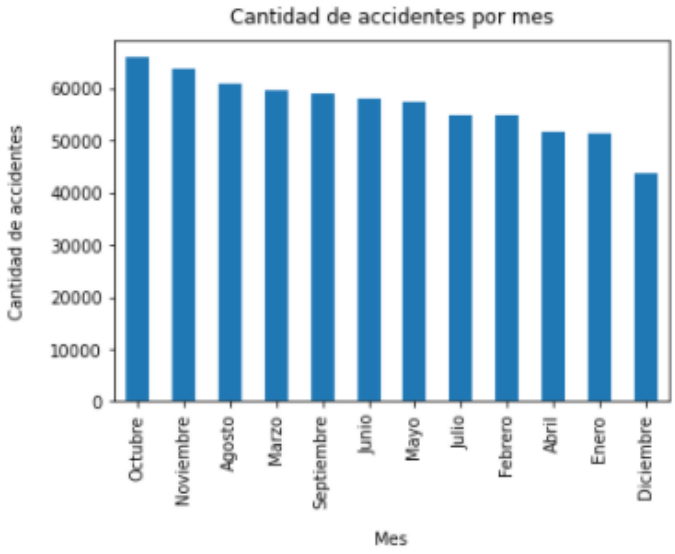


Se puede observar que la gran mayoría de los accidentes son choques sin lesionados. Después sigue accidentes con choque con lesionados, pero la diferencia es bastante grande. Esto significa que la gran mayoría de los choques son choques leves. También hubo bastante atropellados. Fuera de eso los incidentes se vuelven menos comunes siguiendo un comportamiento logarítmico y apenas hay un par de incidentes para las ultimas categorías como en el caso de los sismos.



Aquí podemos observar cuales son las delegaciones con más accidentes. La delegación de Iztapalapa es el lugar con más accidentes y supera a los demás significativamente, en contraste milpa alta es la delegación que tiene menos incidentes.

El tipo de entrada que se usó para reportar el accidente fue una llamada al 911 por mucho.

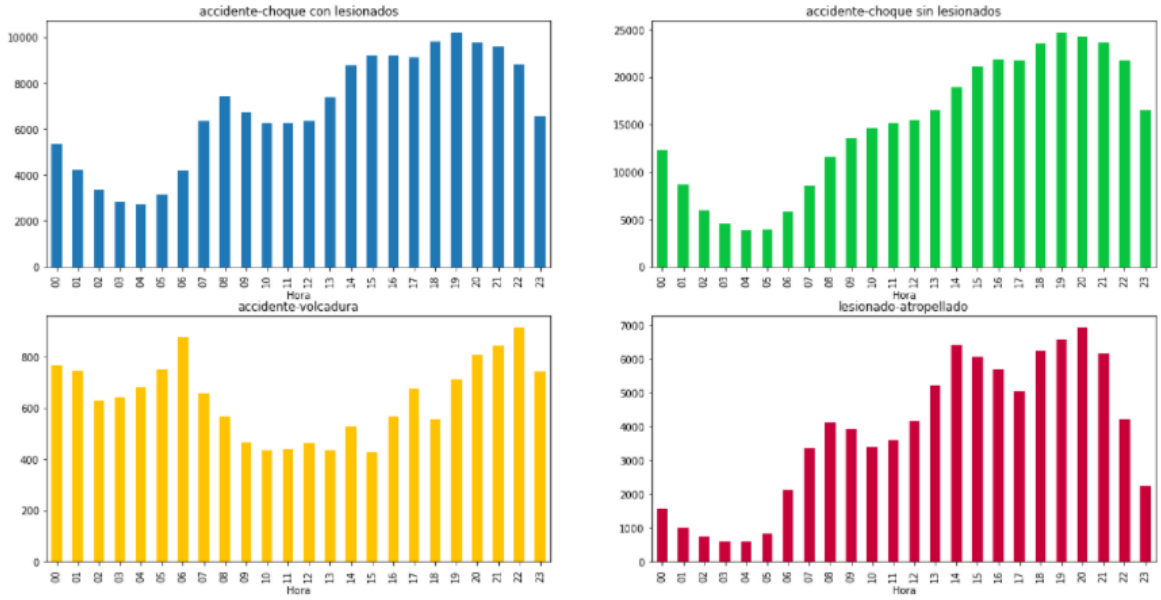


El mes en el que más accidentes hubo fue el de Octubre. Sorprendentemente diciembre fue el mes con menos accidentes. Debido a que la mayoría de los accidentes habían sido por accidentes automovilísticos el sentido común nos indicaba lo contrario porque con las posadas y las fiestas se generarían condiciones para que más gente chocara por ir bajo la influencia del alcohol.

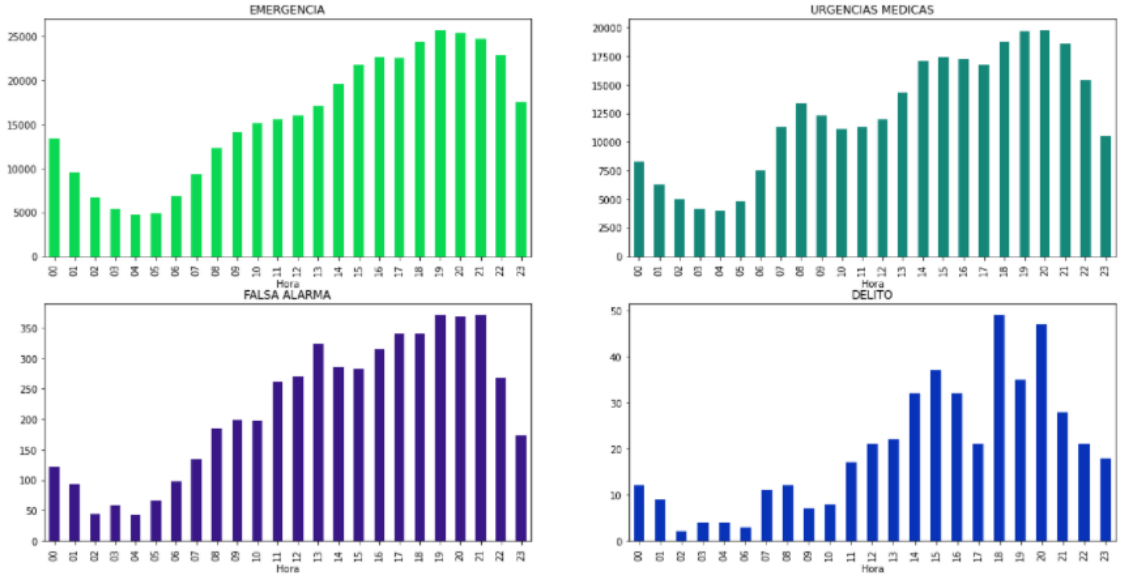
Podemos observar como la mayoría de los accidentes ocurrieron a las 19, 20, 21, 18 y 16 horas. Esto fue sorpresivo ya que se podía esperar que la mayoría de los accidentes ocurrieran a más altas horas de la noche cuando la gente sale de bares y antros. Pero no, ocurrieron a las horas en las que la gente va saliendo del trabajo y el tráfico aumenta.

Dada la conclusión anterior resultaba muy particular por lo que al comparar los principales tipos de accidentes se puede observar una mayor tendencia que corresponde a los pensamientos derivados del sentido común, si tomamos los accidentes choque sin lesionados como nuestro control y nos enfocamos en el número de incidentes proporcionales entonces se puede observar que durante las horas de la madrugada se incrementan los accidentes con lesionados y volcadura principalmente aunque en el primer caso sin ser tan drástico como en el del caso de volcaduras, los atropellados no

se ven influenciados, posiblemente por la baja movilidad en esas horas y que hay un aporte por el alcohol que sea de relevancia.



De igual manera quisimos checar si la hora influía en alguna manera en la clasificación de la alarma del incidente, en general todos se comportan de manera similar con una disminución del tipo de alarma hacia la madrugada, pero con diferencias en sobre la media tarde, hay un pico de urgencias a las 8 de la mañana mientras que las falsas alarmas tienen el pico hacia la una de la tarde y los delitos a las 3 de la tarde, todas hacia el final de día es donde tienen los valores máximos.



Similitud entre los datos

Si consideramos que la mayor parte de la información está contenida dentro de los datos que tuvieron alguna de las clasificaciones del C4 como los accidentes o los lesionados, por ello vamos a trabajar con ese subset para poder hacer la medición de la similitud de los datos

Vamos a considerar como información relevante para comparar nuestros datos el tipo incidente según el código del C4, la hora, la latitud, longitud y la delegación donde fue reportado el incidente, esto también considerando como en algunas otras variables que pudieran ser relevantes a partir del EDA se pudo observar que su distribución es aproximadamente uniforme indistintamente de la condición, debido a la cantidad de datos, el considerar más variables podría hacer que los cálculos se tarden o no se puedan hacer.

Debido a que vamos a trabajar con datos mixtos (categóricos y numéricos) procedimos a crear variables dummies para los datos categóricos y finalmente estandarizar los datos para poder sacar la similitud entre las variables. Finalmente aplicamos a las medidas de similitud de distancia euclidiana y correlación de los datos obteniendo resultados similares en los primeros datos que eran más afines.

Matriz de similitud resultante

	0	1	2	3	4	5	6	7	8	9	...
0	0.000000	11.011672	11.134514	10.662400	11.295041	12.184000	10.778318	12.446406	10.339974	11.539185	...
1	11.011672	0.000000	6.149774	5.610629	6.343125	8.015154	5.654624	8.242294	5.011761	6.667178	...
2	11.134514	6.149774	0.000000	4.464125	5.377716	7.558106	3.699655	6.870450	4.342055	4.497195	...
3	10.662400	5.610629	4.464125	0.000000	5.298067	7.396298	3.397489	7.232998	2.254685	5.005785	...
4	11.295041	6.343125	5.377716	5.298067	0.000000	6.475262	4.824259	6.131132	4.575857	2.547920	...
...	...	...	...	...	...	...	...	...	...	...	...
9996	11.376314	6.454599	4.092799	4.248587	5.263387	7.635483	4.326857	6.753715	4.438866	4.096422	...
9997	11.152218	6.175040	2.630024	3.936104	4.736187	7.136201	4.459634	6.733179	4.325470	4.643670	...
9998	11.476539	6.439787	4.656315	4.439558	4.698461	7.233709	4.298212	6.762871	4.680063	4.568696	...
9999	11.724297	6.825434	5.536608	5.921970	4.918118	7.375420	5.090479	7.165697	5.248264	5.319908	...
10000	11.114874	2.297486	6.525223	5.373620	7.149204	8.565068	6.016926	8.652955	5.592715	7.178148	...

level_0	0
index	0
folio	C5/170622/06952
fecha_creacion	22/06/2017
hora_creacion	1234.75
dia_semana	Jueves
codigo_cierre	(D) El incidente reportado se registró en dos ...
fecha_cierre	22/06/2017
año_cierre	2017
mes_cierre	Junio
hora_cierre	20:39:42
delegacion_inicio	MILPA ALTA
incidente_c4	accidente-choque sin lesionados
latitud	19.2014
longitud	-99.0073
clas_con_f_alarma	EMERGENCIA
tipo_entrada	LLAMADA DEL 911
delegacion_cierre	MILPA ALTA
geopoint	19.20137004,-99.00731988
mes	6
Hora	20

level_0	7788
index	7789
folio	C5/170814/05602
fecha_creacion	14/08/2017
hora_creacion	1091.95
dia_semana	Lunes
codigo_cierre	(D) El incidente reportado se registró en dos ...
fecha_cierre	14/08/2017
año_cierre	2017
mes_cierre	Agosto
hora_cierre	18:29:37
delegacion_inicio	MILPA ALTA
incidente_c4	accidente-choque sin lesionados
latitud	19.1952
longitud	-99.0273
clas_con_f_alarma	EMERGENCIA
tipo_entrada	LLAMADA DEL 911
delegacion_cierre	MILPA ALTA
geopoint	19.19521998,-99.02727
mes	8
Hora	18

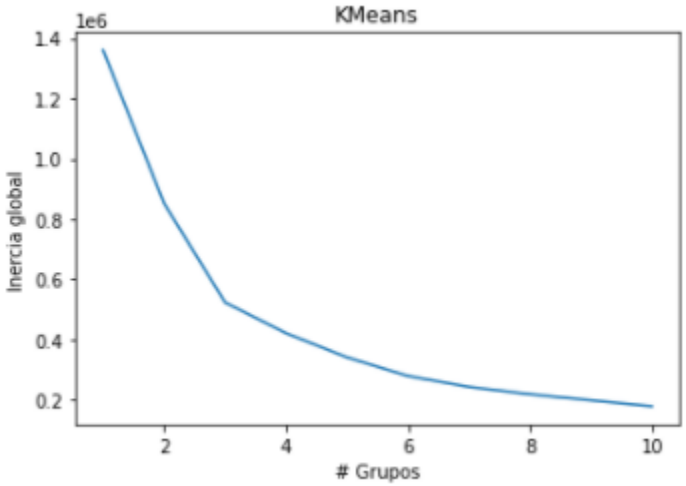
Con esta comparativa se puede observar cómo se obtienen resultados similares por lo menos dentro de los primeros valores que son los más relevantes por ser los más cercanos si se utilizan tanto la métrica euclidiana como la correlación.

Clustering

La mayor parte de los datos son variables categóricas por lo que hacer clustering con cualquiera de las técnicas aprendidas dentro del curso como clustering jerárquico o K-means no podría realizarse ya que esos algoritmos están reservados exclusivamente para datos categóricos, dentro de las variables numéricas con las que si podemos contar de las que más destacan son la longitud y latitud de donde se inician las llamadas de emergencia por lo que serán las variables que utilizaremos para el clustering.

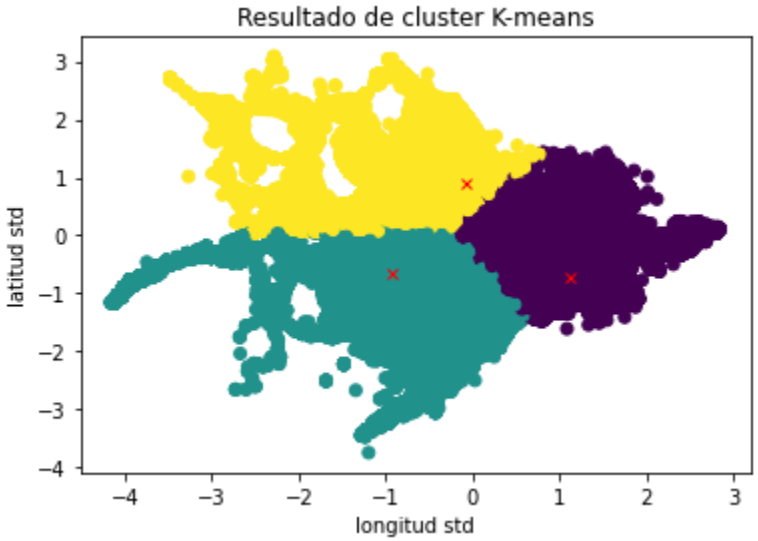
Decidimos de utilizar el algoritmo de clustering de K-means debido a la gran cantidad de datos ya que hacer el algoritmo de clustering jerárquico al ser un método exhaustivo tardaría demasiado tiempo en computarse y posiblemente la memoria se desbordaría.

Para delimitar el número optimo de grupos en base a nuestros datos utilizamos el criterio del codo el cuál nos quedo de la siguiente forma:



Como se puede ver en la inercia global cuando se alcanzan 3 grupos la tendencia cambia y por tanto decidimos seleccionar ese valor como nuestro número de grupos para hacer el clustering.

Después de realizar el clustering utilizando como inicialización K-means ++ para evitar que los centroides inicien en un punto alejado y los resultados no sean los más adecuados se obtuvo el siguiente resultado.



Conclusiones

Pudimos hacer un análisis de la base de datos de accidentes que se nos proporcionó. Fue muy interesante trabajar con datos reales y una base tan grande e ir adentrándose en los datos que albergaba, su composición y entenderlos. Estaba bastante completa por lo que no fue necesario limpiarla tanto. Hicimos el EDA, de donde creemos que encontramos interesantes conclusiones que posteriormente nos sirvió para encontrar los índices de similitud e hicimos el clustering. Con todo esto pudimos ver que había tres clusters en los cuáles podíamos organizar parte de la información.

En el análisis de los grupos que nos tocó, el problema fue que se tuvieron tomar la latitud y longitud, que eran de los datos más relevantes para tomar en cuenta para el clustering y que se perdió información valiosa ya que no se toman en cuenta todas las variables que habíamos visto anteriormente en el EDA y en la similitud de los datos. Consideramos que en líneas generales fue un muy buen proyecto, ya que aplicamos todo lo que hemos aprendido en los módulos.