ZAGAL FLORES ROBERTO ESWART

DATA MINING

ALUMNO:
LÓPEZ MORALES MIGUEL ÁNGEL

3CV19

Práctica 3:

Definición del proyecto de datos semestral. Carga y exploración de datos.

CONTENIDO

1. OBJETIVO

Verificar la factibilidad del proyecto de datos semestral.

2. INTRODUCCIÓN

Una base de datos es un "almacén" que permite guardar grandes cantidades de información de forma organizada, para luego poder usarlo fácilmente. Y estos bases de datos se pueden crear y diseñar usando diferentes sistemas de Gestor de Bases de Datos, conocido con las siglas SGBD, que es un software que actúa como interfaz, entre los datos almacenados y el usuario que desea manejar tales datos.

MySQL Workbench es un software creado por la empresa Sun Microsystems, esta herramienta permite modelar diagramas de Entidad-Relación para bases de datos MySQL.

Con esta herramienta se puede elaborar una representación visual de las tablas, vistas, procedimientos almacenados y claves foráneas de la base de datos. Además, es capaz de sincronizar el modelo en desarrollo con la base de datos real. Se puede realizar una ingeniería directa e ingeniería inversa para exportare e importar el esquema de una base de datos ya existente el cual haya sido guardado o hecho copia de seguridad con MySQL Administrador.

MySQL Workbench puede generar también el guion necesario para crear la base de datos que se ha dibujado en el esquema; es compatible con los modelos de base de datos de DBDesigner 4 y soporta las novedades incorporadas en MySQL 5.x

La versión utilizada en esta práctica es la 8.0.12.

Tableau.
Tableau es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios (más conocida como Business Intelligence). Simplifica los datos en bruto en un formato muy fácil de entender.

La esencia de Tableau es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos. Y esto lo consigue ofreciendo a los usuarios toda una selección de herramientas útiles e intuitivas de inteligencia de negocios.

A través de funciones simples como la de arrastrar y soltar, cualquier persona puede acceder y analizar de forma sencilla datos, e incluso, crear informes y compartir esta información con otros usuarios.

Herramientas
Tableau funciona a través de 3 medios principales:

- Escritorio (Tableau Desktop): el escritorio conecta y analiza los datos e informaciones.

- Servidor (Tableau Server): gracias al servidor podremos colaborar de forma segura y compartir la información a partir de los datos que hayamos subido a través de Tableau Desktop (la versión de escritorio del software).

- En línea (Tableau Online): se trata de una versión de Tableau Server alojada en la nube. De esta forma, podremos acceder a nuestros datos sin necesidad de tener que pasar por un tedioso proceso de instalación.

Además de estas tres herramientas principales, Tableau integra otras para proporcionar una experiencia lo más completa posible a los usuarios:

- Tableau Mobile: se trata de una aplicación complementaria gratuita para Tableau Server o Tableau Online que permite un acceso a los datos y la información guardada en nuestra cuenta.

- Tableau Public: es una versión completamente gratuita de Tableau Desktop y Tableau Online para mostrar los datos que se desean compartir de forma pública.

- Tableau Prep: se trata de una solución que nos permitirá combinar, limpiar y preparar nuestros datos de una forma fácil y sencilla.

Características y funcionalidades de Tableau

Ya sabemos que estamos ante un programa excelente que nos ayudará a tener nuestros datos y documentos correctamente archivados y controlados. Pero, ¿cómo lo consigue? Estas son las funcionalidades más interesantes de Tableau:

Numerosas conexiones de datos: puede conectarse a varias fuentes de datos sin necesidad de ninguna programación, como por ejemplo Redshift, Cloudera Hadoop, SQL Server, Salesforce, Google Analytics y Google Sheets, MongoDB, archivos PDF, Dropbox, Amazon Athena, entre otros.

Datos en vivo y almacenados en memoria: puedes cambiar fácilmente entre datos extraídos y conexiones en vivo, configurando las actualizaciones automáticas de extracción y recibiendo notificaciones cuando falle una conexión de datos.

Colaboración segura: gracias a Tableau Server y Tableau Online, podrás compartir y colaborar de forma segura sin preocuparte por filtraciones de datos o informaciones relevantes.

Diseños optimizados para dispositivos móviles: Device Designer es una herramienta que permite a los usuarios diseñar, personalizar y publicar cuadros de mandos a escala que se optimizan según el dispositivo sin importar si lo estamos visualizando desde un ordenador, un móvil o una tableta.

Tableros integrados: podrás integrar paneles en tus aplicaciones existentes, como Salesforce, SharePoint y Jive, consiguiendo un análisis rápido de forma práctica.

Modo “arrastrar y soltar”: gracias al modo “arrastrar y soltar”, podrás integrar de forma sencilla todo tipo de datos y crear elementos visuales para identificar patrones gracias a unos pocos clics.

3. DESARROLLO

Definir el alcance del proyecto semestral de datos, realizando un primer reconocimiento a la muestra de datos a elegir obtenida en el repositorio de la Ciudad de México.

I. Revise la clase que corresponda al tema "exploración básica de datos con Tableau" y el tema de "limpieza de datos".

II. Explore las diferentes categorías de los conjuntos de datos abiertos de la Ciudad de México: https://datos.cdmx.gob.mx/

- El dataset elegido es:

Carpetas de investigación PGJ de la Ciudad de México
Esta base de datos contiene las Carpetas de investigación de delitos a nivel de calle de la Procuraduría General de Justicia de la Ciudad de México desde enero de 2016 hasta junio de 2019.

Esta base de datos dejó de actualizarse a partir de julio de 2019 debido a que contiene la información de carpetas de investigación de 2018 previa a la reclasificación que realizó la PGJ-CDMX con supervisión de la Oficina de Naciones Unidas contra la Droga y el Delito (UNDOC) y la validación del Centro Nacional de Información (CNI) del Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP)

III. Seleccione un conjunto de datos (dataset) que cumpla con las siguientes condiciones:

- Tener al menos tres años de registros o tuplas. Que el dataset se pueda reducir su tamaño haciendo filtros por año (filtrando al año más reciente), en caso que no sea posible procesar todos los registros.

- Los datos están registrados en años desde 2016 hasta 2019, esto en la columna que indica el año en el que se inició el reporte.

Consulta:

```
SELECT ao_inicio FROM practica_3.carpetas_de_investigacion_pgj_de_cdmx
group by ao_inicio;
```

	ao_inicio
▶	2016
	2017
	2018
	2019

- El dataset debe contener al menos en la dimensión del tiempo “año” y “mes” como dimensión mínima de temporalidad.

- El dataset contiene dimensiones del tiempo y estas están divididas en dos, la primera es la fecha en las que se reportaron los hechos y la segunda es la fecha en las que sucedieron los hechos, ambas están dadas en año, mes, día, hora y minuto.

	Field	Type	Null	Key
▶	id	int(11)	YES	
	ao_hechos	int(11)	YES	
	mes_hechos	text	YES	
	fecha_hechos	text	YES	
	delito	text	YES	
	categoria_delito	text	YES	
	fiscalia	text	YES	
	agencia	text	YES	
	unidad_investigacion	text	YES	
	colonia_hechos	text	YES	
	alcaldia_hechos	text	YES	
	fecha_inicio	text	YES	
	mes_inicio	text	YES	
	ao_inicio	int(11)	YES	

	id	ao_hechos	mes_hechos	fecha_hechos
▶	0	2016	Enero	05/01/2016 17:00
	2	2017	Noviembre	01/11/2017 16:40
	3	2015	Diciembre	30/12/2015 20:00
	4	2018	Mayo	16/05/2018 16:00
	5	2018	Mayo	21/05/2018 20:40
	7	2016	Enero	02/01/2016 21:20
	8	2017	Noviembre	01/11/2017 17:15
	10	2017	Noviembre	01/11/2017 09:00
	11	2016	Enero	06/01/2016 11:00
	12	2017	Noviembre	01/11/2017 14:30
	15	2017	Octubre	28/10/2017 20:30
	17	2017	Octubre	15/10/2017 12:00
	18	2017	Noviembre	01/11/2017 23:30
	19	2018	Mayo	22/05/2018 14:35

- La dimensión de espacio, al menos deben contener “delegación o alcaldía” y “coordenadas (latitud-longitud)”.

- El dataset cumple con dichas condiciones debido a que contienen campos como: “alcaldía”, “dirección”, “longitud”, “latitud”, entre otras.

Consulta:

show columns from carpetas_de_investigacion_pgj_de_cdmx;

Field	Type	Null	Key	Default	Extra
id	int(11)	YES		NULL	
ao_hechos	int(11)	YES		NULL	
mes_hechos	text	YES		NULL	
fecha_hechos	text	YES		NULL	
delito	text	YES		NULL	
categoria_delito	text	YES		NULL	
fiscalia	text	YES		NULL	
agencia	text	YES		NULL	
unidad_investigacion	text	YES		NULL	
colonia_hechos	text	YES		NULL	
alcaldia_hechos	text	YES		NULL	
fecha_inicio	text	YES		NULL	
mes_inicio	text	YES		NULL	
ao_inicio	int(11)	YES		NULL	
calle_hechos	text	YES		NULL	
calle_hechos2	text	YES		NULL	
longitud	double	YES		NULL	
latitud	double	YES		NULL	
geopoint	text	YES		NULL	

- Que la cantidad de registros mínima del dataset debe ser 3 veces mayor al de incidentes viales usado en prácticas anteriores; es decir aproximadamente mayor a 90 mil registros. En caso de que el dataset en su tamaño original no pueda ser procesado, filtre los datos hasta que el dataset cumpla con este requisito. ES IMPORTANTE IMPORTAR LOS DATOS AL MANEJADOR DE SU PREFERENCIA PARA CONOCER SI ESTE REQUISITO SE CUMPLE.

- El dataset cumple con dichos requerimientos debido a que contiene alrededor de 809,000 registros, pero para un mejor manejo de los datos se decidió reducir dicho dataset eliminando datos nulos en “latitud”, “longitud” y “calle_hechos2”

Después de reducir el dataset deshaciéndonos de dichos datos, se encontraron un total de 314,281 registros.

Consulta:

SELECT count(*) as Total FROM practica_3.carpetas_de_investigacion_pgj_de_cdmx;

	Total
▶	314281

- Buscar una aplicación o caso de estudio de valor adicional del dataset elegido si este se complementa o se le integra información sobre el perfil de la población en la CDMX, está información será obtenida desde el sitio oficial del INEGI.

- El caso de estudio de este dataset sería el averiguar cuál es la alcaldía con una mayor tasa de delincuencia en la CDMX.

IV. Adicionalmente, explique cuántas dimensiones temáticas identificó en el dataset. Es importante identificar si el dataset cuenta con diccionario de datos. Por ejemplo, tipo de incidente vial, clasificación del origen del reporte, etc.

- Se identificaron dos dimensiones temáticas, las cuales son: “delito” y “categoría_delito”.

V. Ponga especial atención en describir a detalle la granularidad temporal y espacial. Por ejemplo, el nivel de descripción del tiempo: día, mes, año minuto, segundo, etc.

- La granularidad temporal esta detallada en año, mes, día, hora y minuto.
  La granularidad espacial esta detallada en alcaldía, calle y coordenadas de latitud y longitud.
  La granularidad temporal está dada en dos tipos de granularidad debido a que están guardadas correspondiendo a la fecha en los que ocurrieron los hechos y la fecha en los cuales fueron reportados los hechos.

VI. Explique a detalle el caso de estudio adicional donde el dataset elegido se pueda complementar con datos del perfil población de la CDMX. Por ejemplo, en el dataset de incidentes viales integrando el perfil poblacional, podemos conocerla relación entre la cantidad de incidentes y la cantidad de población que vive en la delegación Coyoacán.

- Este dataset integrado al perfil poblacional de alguna alcaldía podría utilizarse para comparar la cantidad de delitos ocurridos en alguna alcaldía con la población de dicha alcaldía.

VII. Explique las razones o los motivos por las que ha elegido el dataset.

- Interés por conocer la tasa de delincuencia de la alcaldía en donde se encuentra nuestra escuela, es decir, la alcaldía Gustavo A. Madero.

VIII. Explique el problema que quiere resolver al explorar los datos. El alcance del proyecto: es decir explicar cuál es el conocimiento que espero descubrir al estudiar el dataset.

- Comparar la cantidad de población con la cantidad de delitos ocurridos en la alcaldía Gustavo A. Madero para conocer la tasa de delincuencia que ahí ocurre.

IX. Realice el análisis exploratorio básico usando Tableau, contestando las siguientes preguntas generales. Responda aplicando su propio criterio, es decir filtrando la información como considere conveniente. Agregue los resultados en el reporte.

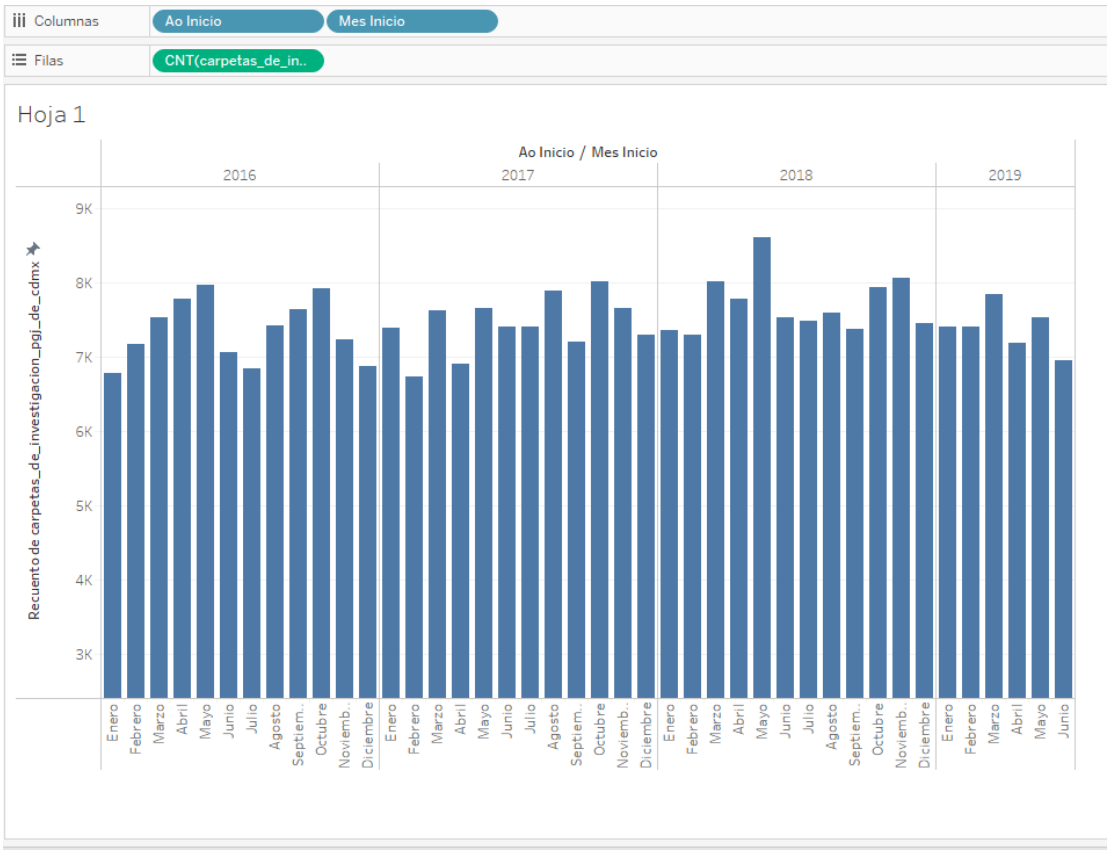
- ¿Cuál es la distribución de la dimensión categorial o temática (el tema del dataset) más importante (del fenómeno que es descrito por el dataset)? Ej. La distribución general de incidentes viales.

- La dimensión temática está centrada en el delito ocurrido en cierta delegación de la CDMX, dado por el dato “delito”.

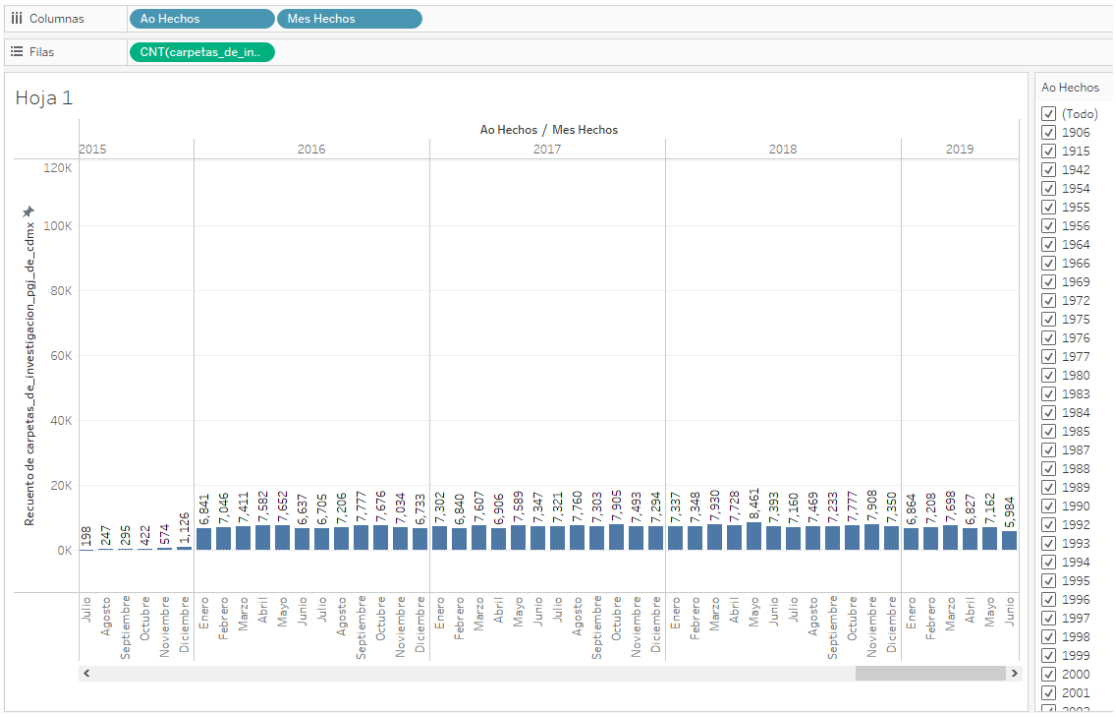
fecha_hechos	delito	categoria_delito
05/01/2016 17:00	LESIONES CULPOSAS	DELITO DE BAJO IMPACTO
01/11/2017 16:40	ROBO A TRANSEUNTE EN VIA PUBLICA CON VI...	ROBO A TRANSEUNTE EN VÃA PÃšBLICA CON Y...
30/12/2015 20:00	FRAUDE	DELITO DE BAJO IMPACTO
16/05/2018 16:00	ROBO A REPARTIDOR CON VIOLENCIA	ROBO A REPARTIDOR CON Y SIN VIOLENCIA
21/05/2018 20:40	ROBO A TRANSEUNTE DE CELULAR CON VIOLE...	DELITO DE BAJO IMPACTO
02/01/2016 21:20	ROBO DE DINERO	DELITO DE BAJO IMPACTO
01/11/2017 17:15	LESIONES INTENCIONALES	DELITO DE BAJO IMPACTO
01/11/2017 09:00	ABUSO DE AUTORIDAD	DELITO DE BAJO IMPACTO
06/01/2016 11:00	DAÃ‘O EN PROPIEDAD AJENA CULPOSA POR T...	DELITO DE BAJO IMPACTO
01/11/2017 14:30	ROBO DE VEHICULO DE SERVICIO PÃšBLICO SI...	ROBO DE VEHÃCULO CON Y SIN VIOLENCIA
28/10/2017 20:30	ROBO DE OBJETOS	DELITO DE BAJO IMPACTO
15/10/2017 12:00	ROBO DE OBJETOS	DELITO DE BAJO IMPACTO
01/11/2017 23:30	VIOLENCIA FAMILIAR	DELITO DE BAJO IMPACTO
22/05/2018 14:35	NARCOMENUDEO POSESION SIMPLE	DELITO DE BAJO IMPACTO

- ¿Cuál es la distribución del fenómeno que mide el dataset en el tiempo?, explorar la mayor cantidad de los niveles de granularidad de tiempo. Ej. La distribución anual de incidentes viales por mes.

- El tiempo que mide el dataset en cuestión de la fecha en la que se reportaron los hechos de algún delito muestra datos desde el 2016 hasta el 2019 mostrado en la siguiente imagen.

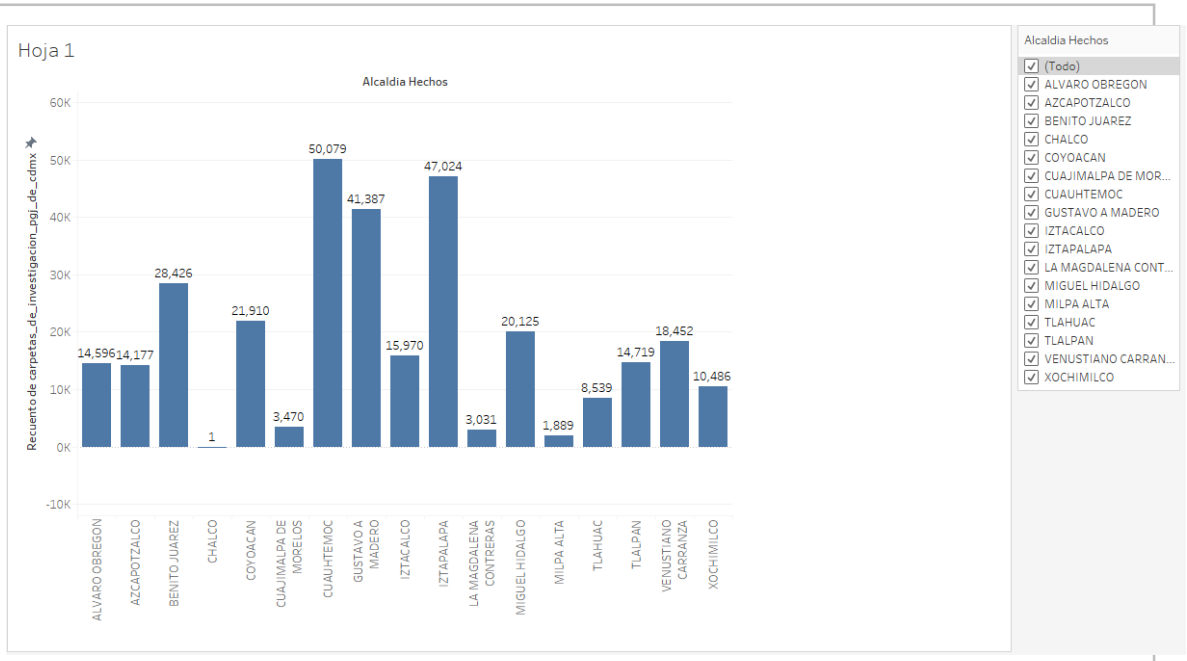


- El tiempo que mide el dataset en cuestión de la fecha en la que se surgieron los hechos de algún delito muestra datos desde el 1906 hasta el 2019 mostrado en la siguiente imagen.



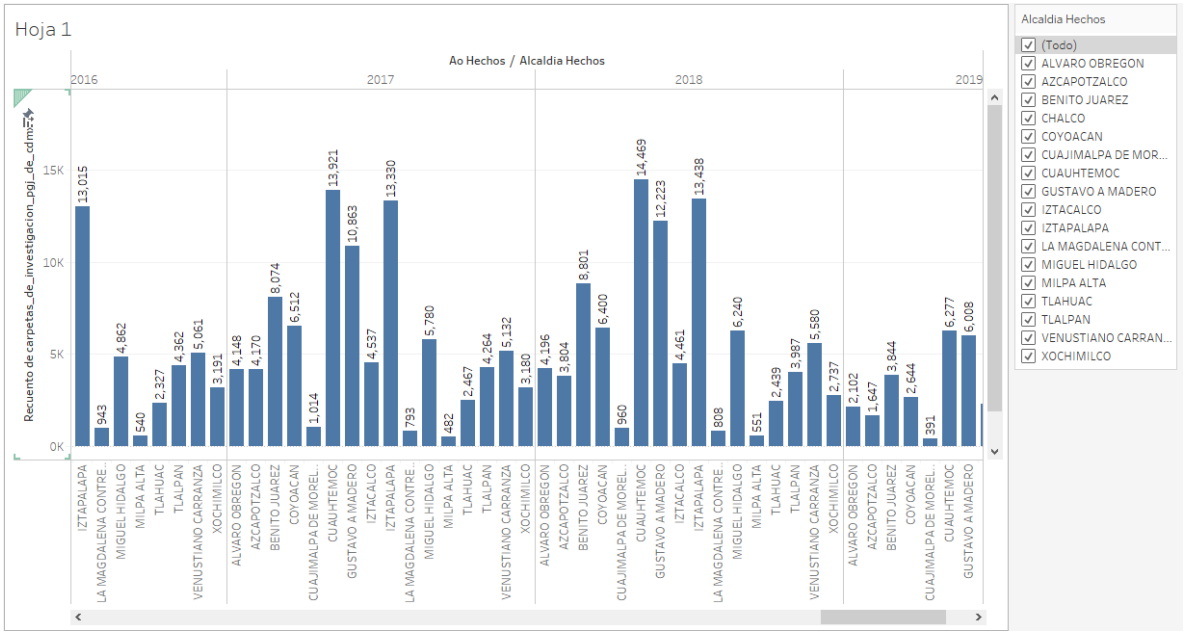
- ¿Cuál es la distribución del fenómeno que mide el dataset en el espacio?, explorar la mayor cantidad de los niveles de granularidad. Ej. La distribución anual de incidentes viales en la delegación Coyoacán.

- La distribución del número de delitos con respecto al lugar está dada en las alcaldías Álvaro Obregón, Azcapotzalco, Benito Juárez, entre otras.



- ¿Cuál es la distribución del fenómeno que mide el dataset en el tiempo y en el espacio?, explorar la mayor cantidad de los niveles de granularidad. Ej. La distribución anual de incidentes viales en la delegación Coyoacán en el 2012.

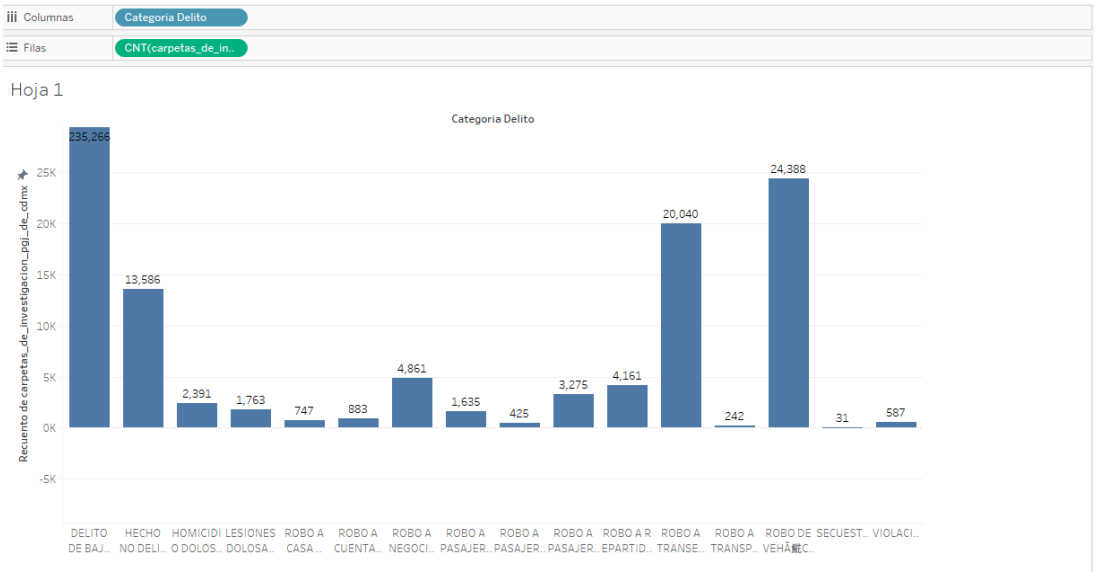
- La distribución de los delitos en el tiempo y espacio se da desde el año 1906 hasta el 2019 y en las alcaldías Álvaro Obregón, Azcapotzalco, Benito Juárez, entre otras.



- ¿Cuál es la distribución de otras dimensiones temáticas (que consideren importante) del dataset? Ej. La distribución de los medios por los que son reportados los incidentes viales.

- Una dimensión temática seria el campo de categoría de delito, el cual se muestra en la siguiente imagen.

-



- ¿Encontró valores atípicos en el dataset o valores inconsistentes?

- Se encontraron valores nulos en campos como “latitud”, “longitud” y “calle_hechos2” los cuales fueron eliminados para facilitar la consulta de datos.

- Verifique si las preguntas se pueden procesar con todos los registros originales del dataset o explique si el dataset fue recortado o filtrado por tiempo u otra variable.

- El dataset fue recortado por valores inconsistentes y también debido a que tardaba un tiempo excesivo en realizar las consultas.

4. CONCLUSIÓN

- **Miguel Ángel López Morales**

“En esta práctica pudimos rectificar nuestros conocimientos adquiridos en la práctica pasada, pero ahora con un dataset de nuestra elección, el cual fue de un tamaño mayor que lo antes visto y que ahora las consultas serán más tardadas de obtener, además de que se tuvo que realizar una limpieza de este dataset como lo hicimos en la práctica pasada pero a un grado mayor, hablando en el total de inconsistencias encontradas ya que el dataset se redujo casi a la mitad debido a la gran cantidad de datos nulos encontrados, con esto se logró reducir bastante el tiempo de la obtención de las consultas ya que llego a tardar mas de 30 minutos en el dataset original y en el reducido lo mas que tardó muchisimo menos tiempo, e incluso podria llegarse a reducir a un más al filtrar por año dicho dataset.

Por estas razones concluyo que el dataset es factible para seguir trabajando en un futuro con él”.

5. REFERENCIAS

I. Colaboradores de Wikipedia. (2020, 19 noviembre). MySQL Workbench. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/MySQL_Workbench

II. Nelson Aranibar, Monografias.com. (s. f.). MySQL WorkBench. Monografias.com. Recuperado 3 de marzo de 2021, de https://www.monografias.com/trabajos88/mysql-worckbench/mysql-worckbench.shtml

III. Tableau, una de las principales herramientas de Inteligencia de Negocios. (s. f.). SPnet. Recuperado 28 de marzo de 2021, de https://softwarepara.net/tableau/