INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Data Mining

Grupo: 3CV14

Guzmán Hernández Luis Daniel
López Sánchez Kevin Ian

Práctica 6. Definición del proyecto de datos semestral. Carga y exploración de datos.

03 de mayo de 2022

INTRODUCCIÓN

En la presente práctica, definiremos el alcance de nuestro proyecto semestral, realizando una búsqueda, selección y reconocimiento de una muestra o conjunto de datos, los cuáles serán obtenidos en el repositorio proporcionado por la Ciudad de México.

Para poder realizar una propuesta y posteriormente la selección de la muestra de datos, deberemos seguir algunas condiciones impuestas relacionadas con formatos de espacio y tiempo, encontrar un problema que se pueda resolver procesando los datos obtenidos y así poder verificar la factibilidad de que el proyecto sea realizado.

Por último, realizaremos una exploración básica utilizando el software de Tableau para así obtener las distribuciones de la dimensión categórica y del fenómeno y encontrar valores atípicos o datos inconsistentes.

DESARROLLO

Se seleccionó entre las diferentes opciones de conjuntos de datos, las denuncias realizadas ante la PAOT (Procuraduría Ambiental y del Ordenamiento Territorial de la CDMX). Cabe mencionar que el dataset seleccionado no cuenta con un diccionario de datos o metadatos. Al cargar el archivo al manejador de base de datos se obtuvo una tabla con los siguientes campos:

Columna	Tipo de Dato	Descripción
id	numeric	Identificador de la denuncia
expediente	text	Identificador en el expendiente
tipo_de_denuncia	text	Tipo de la denuncia cometida
estatus	text	Estado de la denuncia
tema	text	Tema de la denuncia
colonia	text	Colonia en donde se registró la denuncia
cp	numeric	Código postal en donde se registró la denuncia
denunciante	text	Tipo de denunciante
lat (latitud)	float	Latitud registrada en la denuncia
anio_de_recepcion	numeric	Año en que se recibió la denuncia
mes_de_recepcion	text	Mes en que se recibió la denuncia
geopoint	text	Coordenadas registradas en la denuncia
medio_de_recepcion	text	Medio por el que se recibió la denuncia
fecha_de_recepcion	date	Fecha en que se recibió la denuncia

fecha_de_admision	text	Fecha en que se admitió la denuncia
alcaldia	text	Alcaldía en la que se registró la denuncia
long (longitud)	float	Longitud registrada en la denuncia

1. El dataset cumple con las siguientes condiciones:

1. El dataset tiene 18 años de registros, los cuales van desde el año 2002 hasta el año 2019, pudiéndose filtrar los datos por año como se muestra a continuación:

	TOTAL	AÑO
1	49	2002
2	457	2003
3	610	2004
4	874	2005
5	1079	2006
6	1198	2007
7	1447	2008
8	1528	2009
9	2654	2010
10	2397	2011
11	2357	2012
12	3198	2013
13	3425	2014
14	3466	2015
15	3559	2016
16	4294	2017
17	5236	2018
18	4018	2019

2. Los campos “fecha_de_recepcion” y “fecha_de_admision” tienen día, mes y año, por lo que la dimensión del tiempo posee granularidad a nivel de “año”, “mes” y “día”.

	fecha_de_recepcion	fecha_de_admision
1	30/01/2002	06/02/2002
2	02/05/2002	08/05/2002
3	02/05/2002	07/05/2002
4	04/07/2002	05/07/2002
5	04/07/2002	05/06/2002
6	12/09/2002	27/09/2002
7	09/10/2002	21/10/2002
8	09/10/2002	22/10/2002
9	15/10/2002	31/10/2002
10	23/10/2002	01/11/2002
11	18/11/2002	29/11/2002
12	19/11/2002	02/12/2002
13	25/11/2002	02/12/2002
14	04/12/2002	17/12/2002
15	13/01/2003	14/01/2003
16	15/01/2003	17/01/2003
17	04/09/2018	08/11/2018
18	16/01/2003	21/01/2003
19	16/01/2003	27/01/2003
20	20/01/2003	31/01/2003
21	23/01/2003	07/02/2003
22	27/01/2003	29/01/2003
23	04/02/2003	12/02/2003
24	04/02/2003	18/02/2003
25	04/02/2003	10/02/2003

3. La dimensión espacio posee granularidad a nivel de “colonia”, alcaldía” y “geopoint”.

	colonia	alcaldia	lat	long
1	DEL VALLE	Benito Juárez	19.3720512390137	-99.1735992431641
2	MIGUEL HIDALGO AMP	Tlalpan	19.2781848907471	-99.2014846801758
3	POPULAR RASTRO	Venustiano Carranza	19.4533805847168	-99.1168823242188
4	NARVARTE	Benito Juárez	19.4010276794434	-99.1476593017578
5	MORELOS	Cuauhtémoc	19.4463005065918	-99.1348266601563
6	ARTES GRAFICAS	Venustiano Carranza	19.4118347167969	-99.127685546875
7	SAN LORENZO HUIPULCO	Tlalpan	19.2966384887695	-99.1520004272461
8	ROMA NORTE	Cuauhtémoc	19.4223670959473	-99.167594909668
9	DEL VALLE	Benito Juárez	19.3683414459229	-99.175651550293
10	OBRERA	Cuauhtémoc	19.4161434173584	-99.1366958618164
11	SAN RAFAEL	Cuauhtémoc	19.4367828369141	-99.1659851074219
12	LAS AGUILAS AMP	Álvaro Obregón	19.349910736084	-99.2203369140625
13	NONOALCO	Benito Juárez	19.3776626586914	-99.1904220581055
14	ACUEDUCTO DE GUADALUPE	Gustavo A. Madero	19.5230274200439	-99.1525650024414
15	MERCED GOMEZ	Álvaro Obregón	19.3664798736572	-99.2039794921875
16	EL PRADO	Iztapalapa	19.3594665527344	-99.1383209228516
17	CHIMALCOYOTL	Tlalpan	19.2749519348145	-99.1706619262695
18	SANTA CATARINA	Azcapotzalco	19.4931049346924	-99.1743469238281
19	EJERCITO CONSTITUCIONALISTA	Iztapalapa	19.3848762512207	-99.0410461425781
20	SALVADOR DIAZ MIRON	Gustavo A. Madero	19.5230274200439	-99.1525650024414
21	NUEVA ATZACOALCO	Gustavo A. Madero	19.5230274200439	-99.1525650024414

4. Identificamos las siguientes dimensiones temáticas y sus diferentes posibles valores:

Dimensión	Posibles valores
Tipo de denuncia	• Ciudadana • De oficio
Estatus	• Concluida • No admitida • No presentada • En investigación • En proceso de admisión • No ratificada
Tema	• Áreas Verdes (en suelo urbano) • Animales • Uso de Suelo Urbano • Ruido y Vibraciones • Suelo de Conservación • Agua • Barrancas • Aire • Gases Olores y Vapores • Residuos • Contaminación Visual • Energía Lumínica y Térmica • Áreas Naturales Protegidas • Áreas de Valor Ambiental
Denunciante	• Persona física • Persona moral • Autoridad
Medio de recepción	• Escrita • Personal • Medios Informativos • Inspección Ocular • Teléfono • Archivo PAOT • Internet • Fax • Módulos Delegacionales • Consejo Ciudadano de Seguridad Publica y Procuración de Justicia • Radio • Modulo Móvil • Aplicación móvil (App) • Correo Postal

5. Importamos el dataset en el manejador de bases de datos Microsoft SQL Server. Obtuvimos un total de registros de 41,846 registros, por lo que cumple con la condición de tener un número mayor a 20 mil registros.

	TOTAL
1	41846

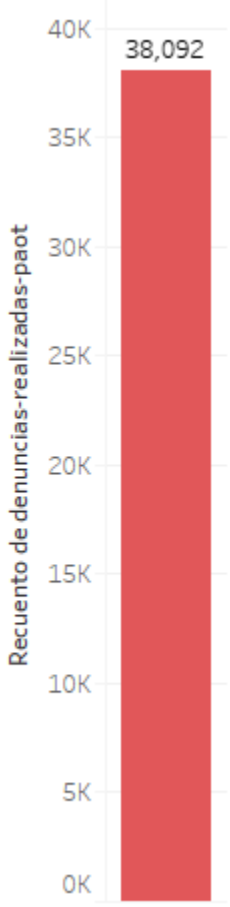
6. Una vez realizada una limpieza rápida de valores nulos, el total de registros del dataset se redujo a 38,092 registros, por lo que cumple con la condición de tener un número menor a 40 mil registros y sigue siendo mayor a los 20 mil registros.

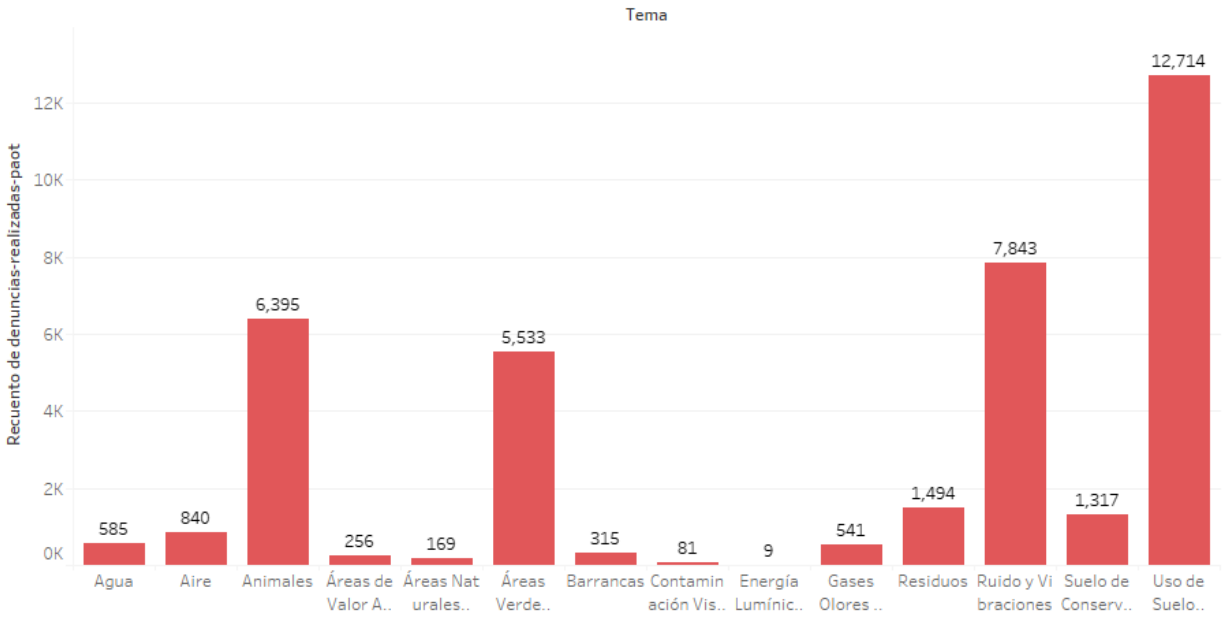
```
DELETE FROM dbo.[denuncias-realizadas-paot] where id is null
DELETE FROM dbo.[denuncias-realizadas-paot] where expediente is null
DELETE FROM dbo.[denuncias-realizadas-paot] where tipo_de_denuncia is null
DELETE FROM dbo.[denuncias-realizadas-paot] where estatus is null
DELETE FROM dbo.[denuncias-realizadas-paot] where tema is null
DELETE FROM dbo.[denuncias-realizadas-paot] where colonia is null
DELETE FROM dbo.[denuncias-realizadas-paot] where cp is null
DELETE FROM dbo.[denuncias-realizadas-paot] where denunciante is null
DELETE FROM dbo.[denuncias-realizadas-paot] where lat is null
DELETE FROM dbo.[denuncias-realizadas-paot] where anio_de_recepcion is null
DELETE FROM dbo.[denuncias-realizadas-paot] where mes_de_recepcion is null
DELETE FROM dbo.[denuncias-realizadas-paot] where geopoint is null
DELETE FROM dbo.[denuncias-realizadas-paot] where medio_de_recepcion is null
DELETE FROM dbo.[denuncias-realizadas-paot] where fecha_de_recepcion is null
DELETE FROM dbo.[denuncias-realizadas-paot] where fecha_de_admision is null
DELETE FROM dbo.[denuncias-realizadas-paot] where alcaldia is null
DELETE FROM dbo.[denuncias-realizadas-paot] where long is null
```

	TOTAL
1	38092

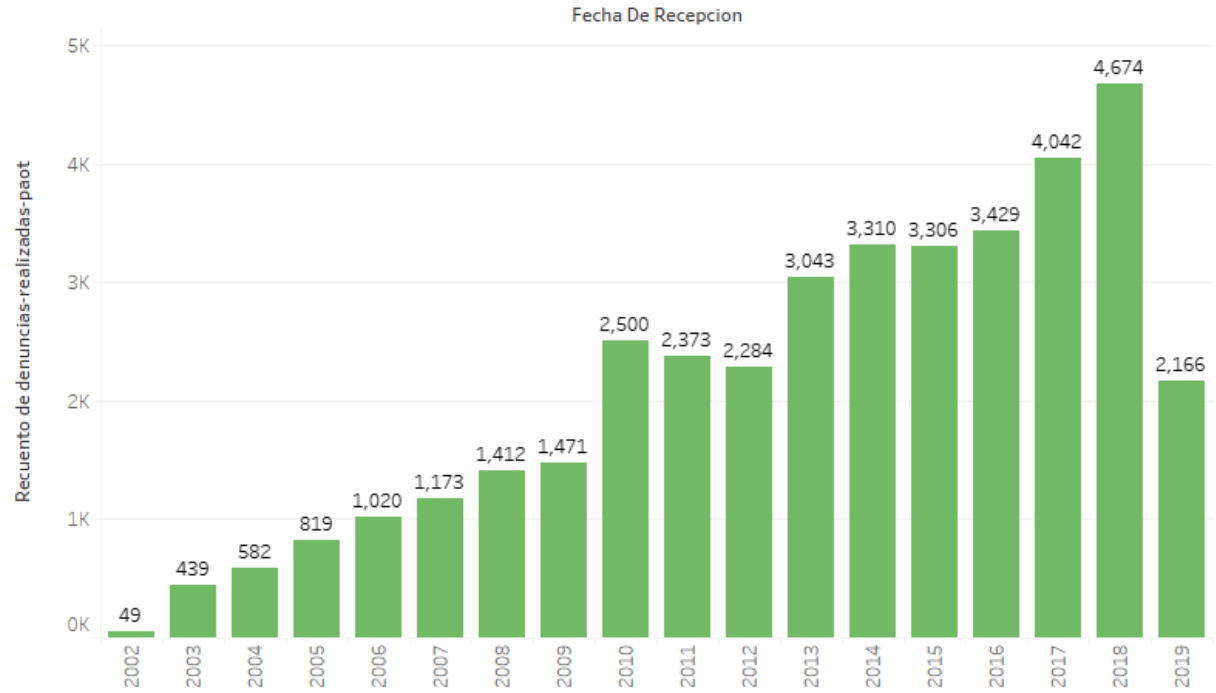
7. Una vez realizando un análisis básico al cargar los datos nos percatamos de que existe una gran cantidad de problemas que se pueden resolver y de que las aplicaciones que le encontramos a la manipulación del dataset elegido son variadas, sin embargo, para definir el alcance de nuestro proyecto nos centraremos en una parte más general del dataset para intentar encontrar **la tendencia en los tipos de problemas que se han reportado en cada alcaldía y colonia a través de los días, meses y años**, y así, poder concluir y determinar si estos se han atendido o han sido ignorados.

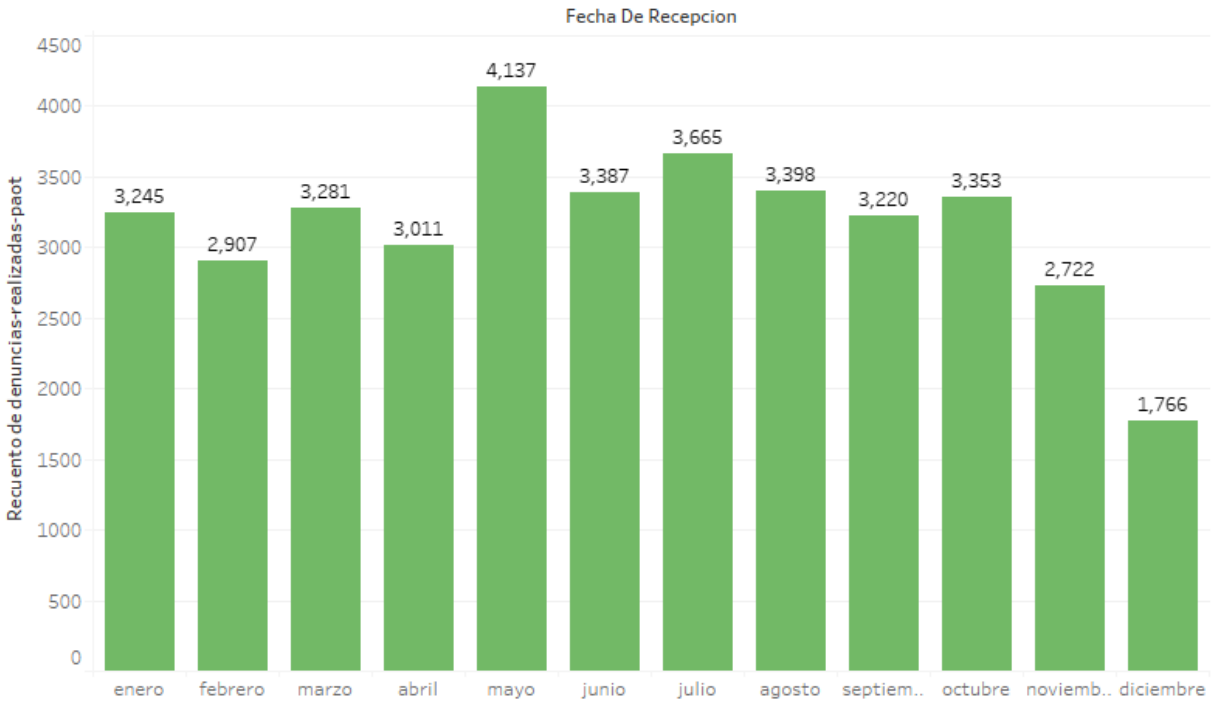
2. **Realice el análisis exploratorio *básico* usando Tableau,** contestando las siguientes preguntas generales. Responda aplicando su propio criterio, es decir filtrando la información como considere conveniente. Agregue los resultados en el reporte.
    1. ¿Cuál es la distribución de la **dimensión categorica ó temática (el tema del dataset)** más importante (del fenómeno que es descrito por el dataset)?. Ej. La distribución general de incidentes viales.

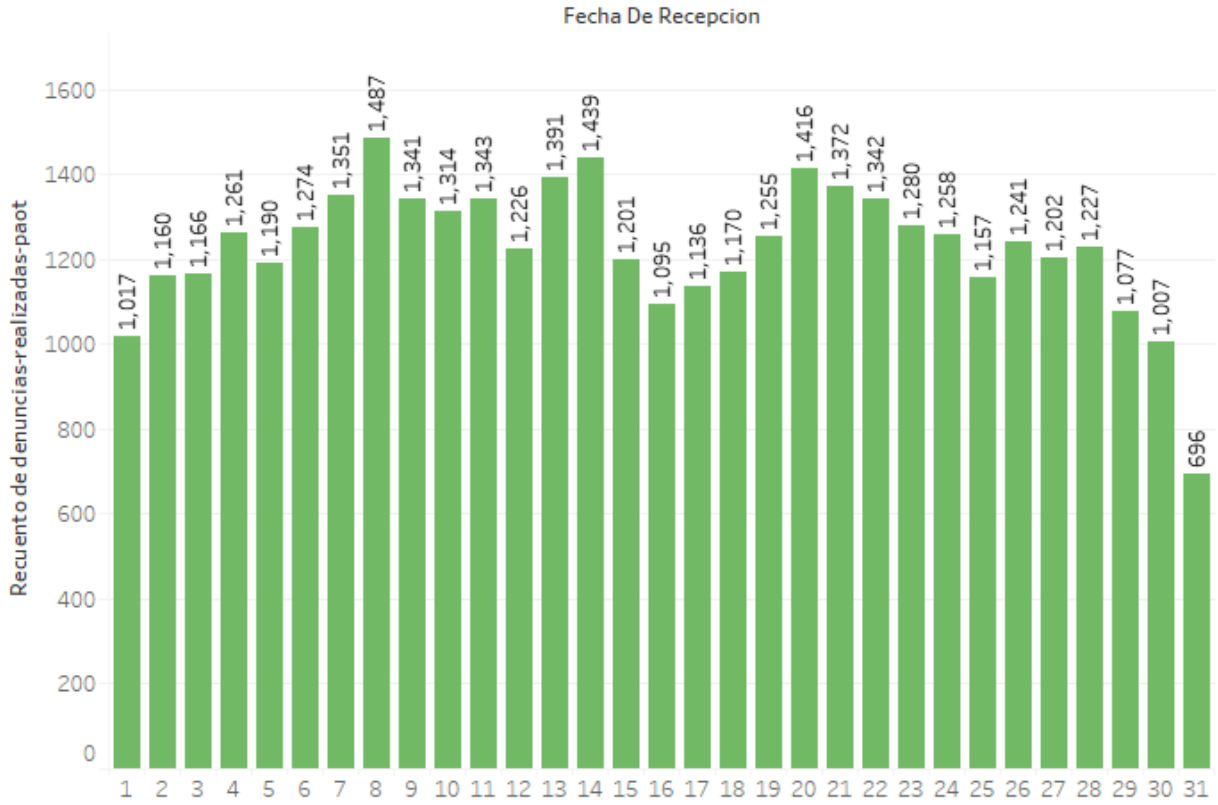




2. ¿Cuál es la **distribución del fenómeno** que mide el dataset en el tiempo?, ***explorar la mayor cantidad de los niveles de granularidad de tiempo***. Ej. La distribución anual de incidentes viales por mes.




Fecha De Recepcion
Recuento de denuncias-realizadas-paot
4500
4000
3500
3000
2500
2000
1500
1000
500
0
3,245
2,907
3,281
3,011
4,137
3,387
3,665
3,398
3,220
3,353
2,722
1,766
enero
febrero
marzo
abril
mayo
junio
julio
agosto
septiem..
octubre
noviemb..
diciembre


Fecha De Recepcion
Recuento de denuncias-realizadas-paot
1600
1400
1200
1000
800
600
400
200
0
1,017
1,160
1,166
1,261
1,190
1,274
1,351
1,487
1,341
1,314
1,343
1,226
1,391
1,439
1,201
1,095
1,136
1,170
1,255
1,416
1,372
1,342
1,280
1,258
1,157
1,241
1,202
1,227
1,077
1,007
696
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

3. **¿Encontró valores atípicos en el dataset o valores inconsistentes? (verificar el diccionario de datos).**

Se encontraron alrededor de 3000 valores nulos, principalmente en los campos de código postal, alcaldía, colonia, geopoint, latitud y longitud, es decir, que algunos de los registros no contaban con información acerca de la ubicación en donde se realizaron las denuncias.

También encontramos algunas irregularidades en el tipo de datos de las columnas, ya que no permitía, por ejemplo, ingresar el valor de la fecha de admisión como tipo de dato “date”.

El dataset seleccionado **no cuenta con diccionario de datos**, sin embargo, esto no fue un impedimento para realizar la exploración del dataset, ya que los nombres de las columnas cuentan con un nombre muy explícito acerca de lo que contiene.

4. **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.

Para realizar una exploración básica utilizando el software de Tableau y contestar las preguntas anteriores, tuvimos que realizar una limpieza rápida de los valores nulos encontrados, así que no pudimos procesarlas con todos los registros originales.

CONCLUSIONES

El procesamiento del dataset es bastante factible, ya que, no es demasiado grande, la cantidad de registros del dataset está dentro del rango de los otros datasets que hemos procesado. La granularidad que tiene el dataset en la dimensión de espacio es bastante buena, ya que ofrece información hasta colonia y punto geográfico, lo que puede enriquecer el análisis, la granularidad de la dimensión del tiempo no es tanta como la de la dimensión del espacio, pero es suficiente para lo que vamos a desarrollar.

Teniendo en cuenta los factores anteriores concluimos que en el dataset existe una cantidad de datos considerable que cumple con las condiciones para realizar un análisis exploratorio del mismo, y así desarrollar nuestro proyecto semestral encontrando una aplicación favorable al conjunto de datos a través de su manipulación.