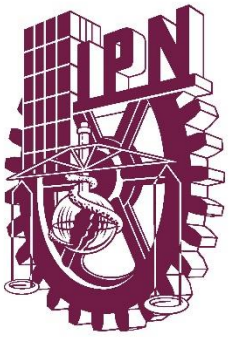

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE COMPUTO

3CV19

MINERÍA DE DATOS

RAMIREZ OLVERA GUILLERMO

PRÁCTICA 3

DEFINICIÓN DEL PROYECTO DE DATOS SEMESTRAL

CARGA Y EXPLORACIÓN DE DATOS

FECHA DE ENTREGA:

28/03/2021

INSTITUTO POLITÉCNICO NACIONAL
"La Técnica al Servicio de la Patria"

Para la realización de la practica 3 se escogió el dataset de víctimas en carpeta de investigación, esta base tiene alrededor de 514000 registros por lo que se le procede la aplicación de un filtro, en este caso se escogió los primeros registros más resientes, por motivos de carga en el tiempo también se descartaron los registro que no tuvieran datos en los campos de altitud o longitud, esto debido a que en nuestro manejador de base de datos, al recibir el dato del dataset que indicaba que el campo se encontraba vacío, en este caso era NA, el manejador de base de datos trataba ese campo como text haciendo que la carga de los registros sea realmente lenta, alrededor de 3 segundos por registro, haciendo imposible la carga de un numero significativo de estos, por lo que al final se quedo con un n**úmero total de registros de 112345.**

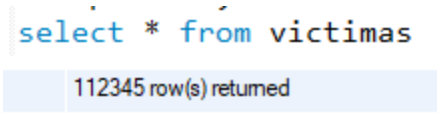

Figura 1: Número de registro de la tabla victimas

A continuación, se presenta el diccionario de datos del dataset, este será adjuntado en el zip de la entrega de la práctica.

Nombre de variat	Definición	Catálogo
idCarpeta	Número entero que representa el identificador único usado por PGJ asociado a cada carpeta de investigación dentro de su sistema.	
Delito	Es la conducta, a	Vehiculo
Categoria	Las carpetas de	Impacto
FechaHecho	Día y hora en que se cometió el delito, según el reporte de la víctima.	
HoraHecho	Hora y minuto en que se cometió el delito, según el reporte de la víctima.	
Fechalnicio	Día y hora en que se hizo la denuncia para iniciar la carpeta de investigación.	
Horalnicio	Hora y minuto en que se hizo la denuncia para iniciar la carpeta de investigación.	
Año_hecho	Año en que se c	1968
Mes_hecho	Mes en que se cometió el delito, según el reporte de la víctima.	
Año_inicio	Año en que se hizo la denuncia para inciar la carpeta de investigación.	
Mes_inicio	Mes en que se hizo la denuncia para inciar la carpeta de investigación.	
Sexo	Sexo de la víctim	Femenino
Edad	Edad de la víctima del delito reportado en la carpeta de investigación.	
TipoPersona	Cómo se reconoc	Moral
CalidadJuridica	Título con el que	(Personas que,
Competencia	Variable categóri	COMUN
lon	Longitud de la geolocalización, uno de dos elemento que componen la referencia angular que permite localizar el lugar donde se cometió el delito. WGS84	
lat	Latitud de la geolocalización, uno de dos elemento que componen la referencia angular que permite localizar el lugar donde se cometió el delito. WGS84	
AlcaldiaHechos	Alcaldía en que se cometió el delito, según el reporte de la víctima. Notar que puede ser fuera de la CDMX.	
ColoniaHechos	Colonia en que se cometió el delito, según el reporte de la víctima. Notar que puede ser fuera de la CDMX.	
Calle_hechos	Calle en que se cometió el delito, según el reporte de la víctima.	
Calle_hechos2	Calle secundaria en que se cometió el delito, según el reporte de la víctima.	

Figura 2: Diccionario de datos del dataset.

Las **dimensiones temáticas** más importantes que se pueden apreciar en este dataset es el tipo de delito que se cometió, su categoría, Conducta jurídica y el sexo.

Como se puede apreciar se cuenta con la **dimensión de tiempo** año y mes, además que la granularidad temporal está muy bien detallada, dándonos información del tiempo hasta segundos de cuando ocurrió el delito y cuando se denunció el delito.

Con o que respecta a la **dimensión del espacio**, contamos con alcaldía, colonia, calle de los hechos, latitud y longitud, esto nos permite un análisis a detalle de la información permitiéndonos manejar con mucha facilidad los datos para obtener alguna información específica.

Un **caso de estudio** que se puede realizar al conocer estos datos conjunto a los datos de la población del INEGI, es la relación del número de delitos en algún área en particular conjunto al numero de habitantes dentro de dicha área, esto con el fin de poder distribuir de una mejor forma los cuerpos policiales para que al con su presencia pueda mermar estos delitos o hacer prota su aparición ante los hechos. Otro caso interesante puede ser que, sabiendo la edad de las víctimas, con la adición del número de pobladores cerca de los hechos, podamos crear centros de atención psicológica y de ayuda para las víctimas, con profundizar en estos datos junto con la información de la dimensión del espacio, podemos crear dichos centros con la mayor eficiencia posible, donde los centros puedan estar cerca del mayor número de afectados posibles con relación a la proporción de la población de esa área.

Los motivos que me motivaron a escoger este dataset fue su buena granularidad en la dimensión del tiempo, además de que contaba con datos muy precisos en la dimensión del espacio como son las coordenadas, esto último fue muy difícil de encontrar en la pagina de datos de la CDMX, debido a que muchos dataset se limitaban a lo mucho al CP, pero el motivo más interesante fue que anteriormente realice con unos compañeros un proyecto para intentar prevenir los robos con violencia en la CDMX, y en su día no teníamos estos conocimientos para extraer la información, por lo que me parece interesante ver los datos sobre las víctimas de la CDMX.

Análisis exploratorio usando tableau

Para poder analizar la distribución más importante, se realizo un filtro al número de victimas para poder visualizar la información de una manera mucho más sencilla, dicho filtro es que muestre aquellos datos que tengan más de 2000 registros.

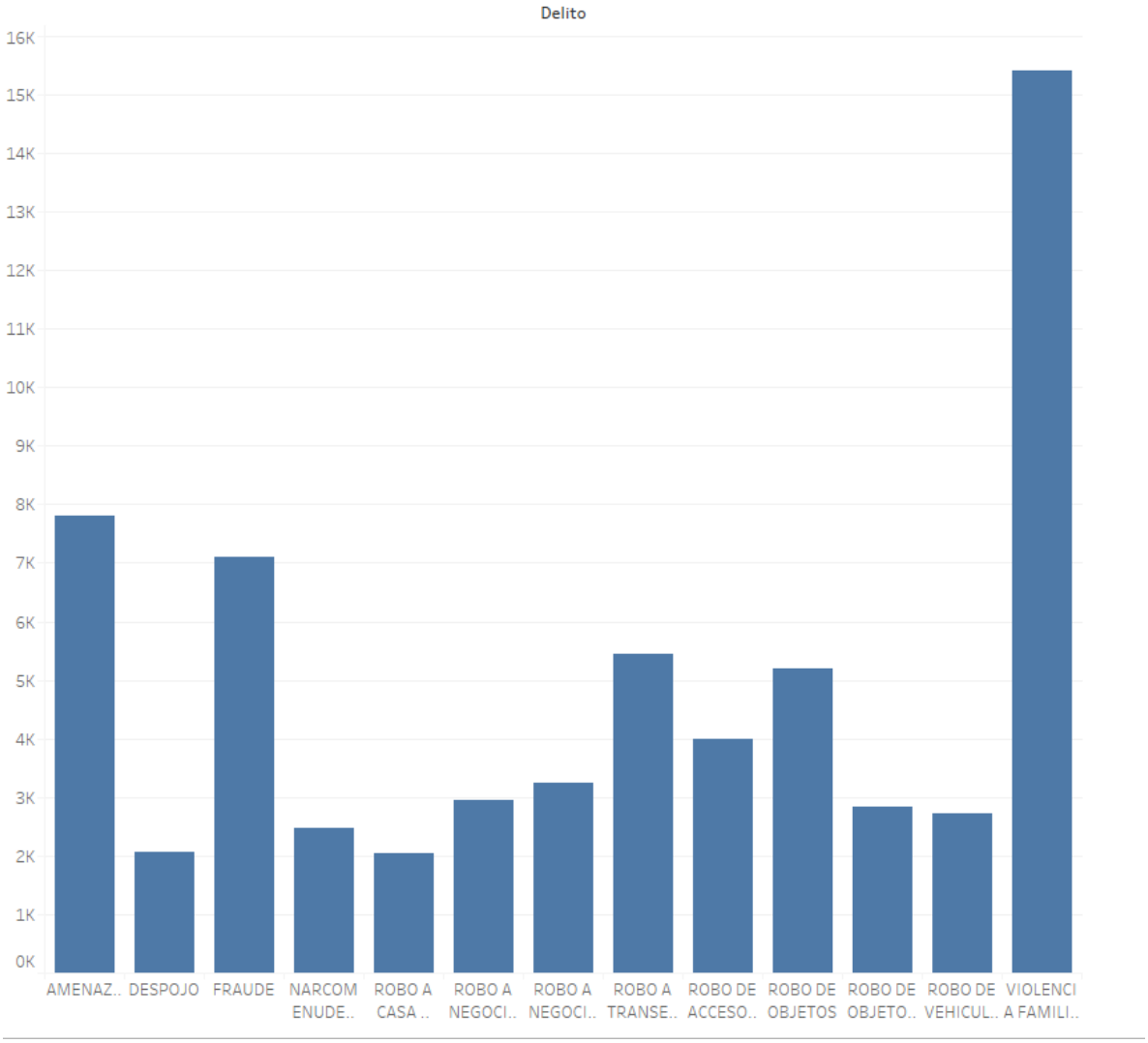

Figura 3: Grafica con el número de registros x delito

Para el análisis en la dimensión del tiempo primero se exploro con el mes de los hechos, dándonos a continuación la siguiente gráfica:

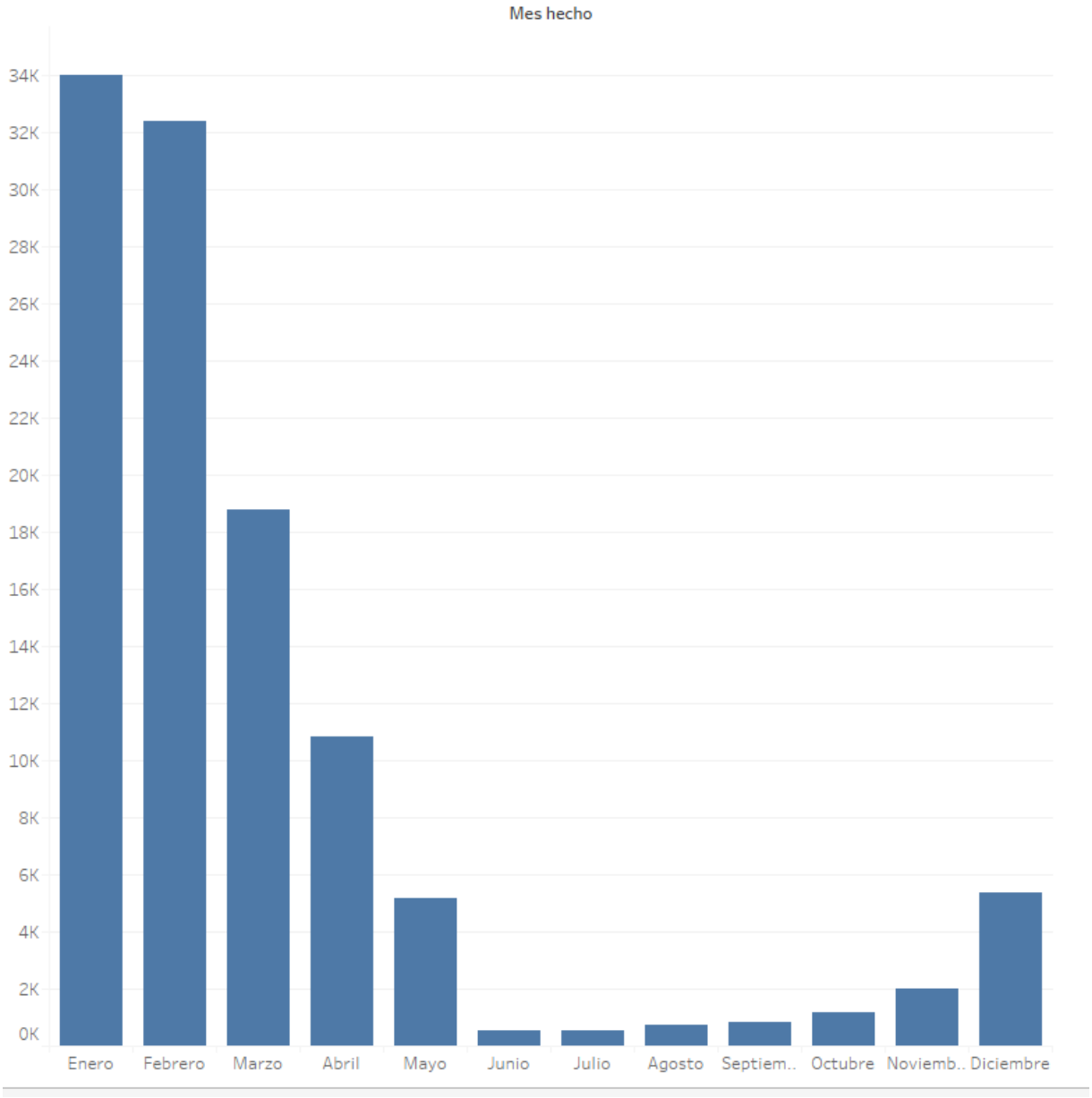

Figura 4: Grafica con el número de registros x mes

Al seguir analizando los datos en función del tiempo se opto por ver a que los registros por hora de los hechos, para esto se aplico un filtro al número de víctimas, donde este debía ser mayor a 500, lo que nos dio la siguiente grafica.

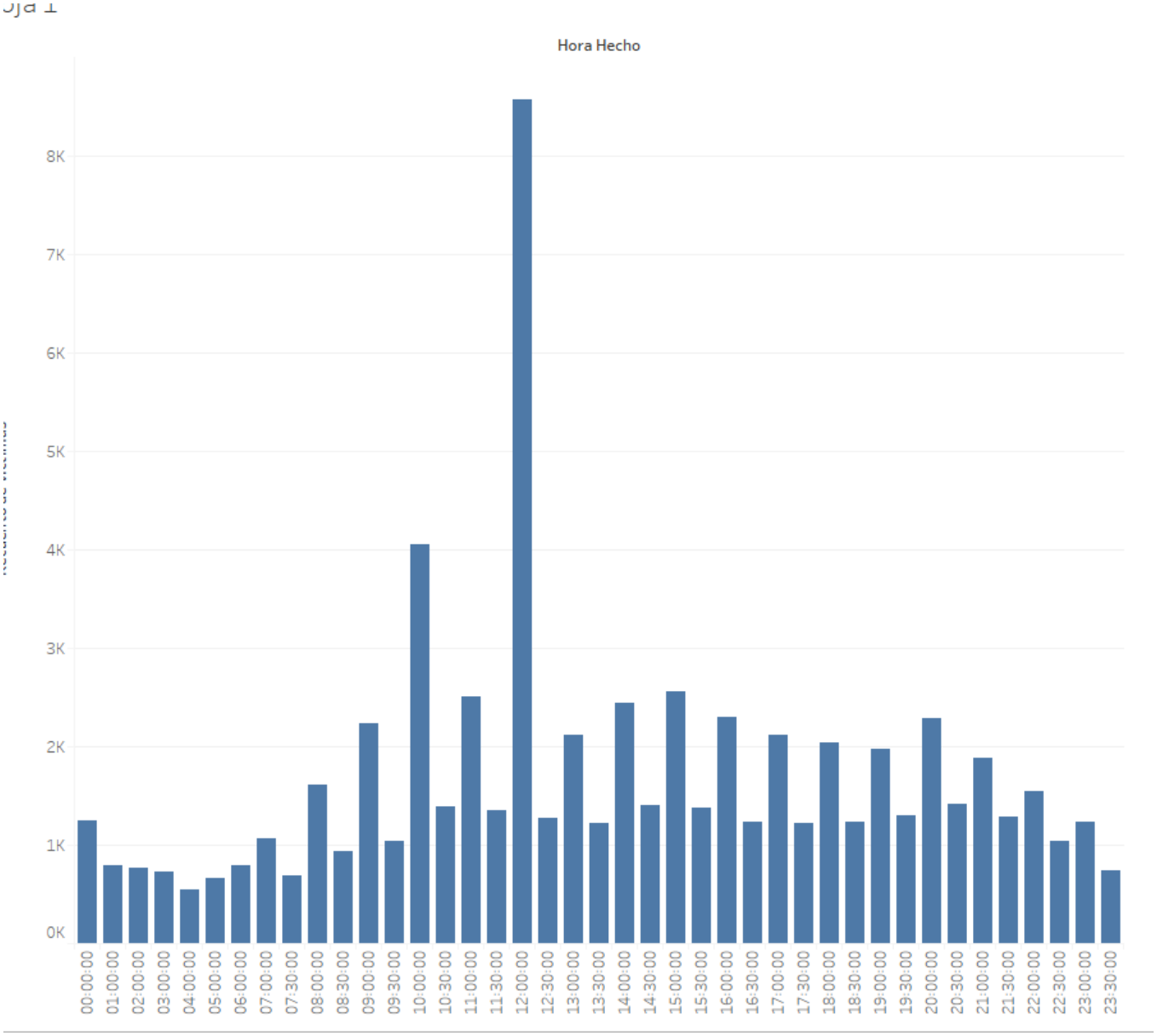

Figura 5: Grafica con el número de registros x hora de los hechos

Podemos apreciar que la hora en donde hay mas llamadas son las 12:00, siendo de este un gran dato que nos puede ayudar en nuestro caso de estudio haciendo que el patrullaje del cuerpo de la policía sea más riguroso en esa hora.

Para el análisis de los hechos en el espacio, se opto por ver la alcaldía de los hechos, dándonos la siguiente gráfica:

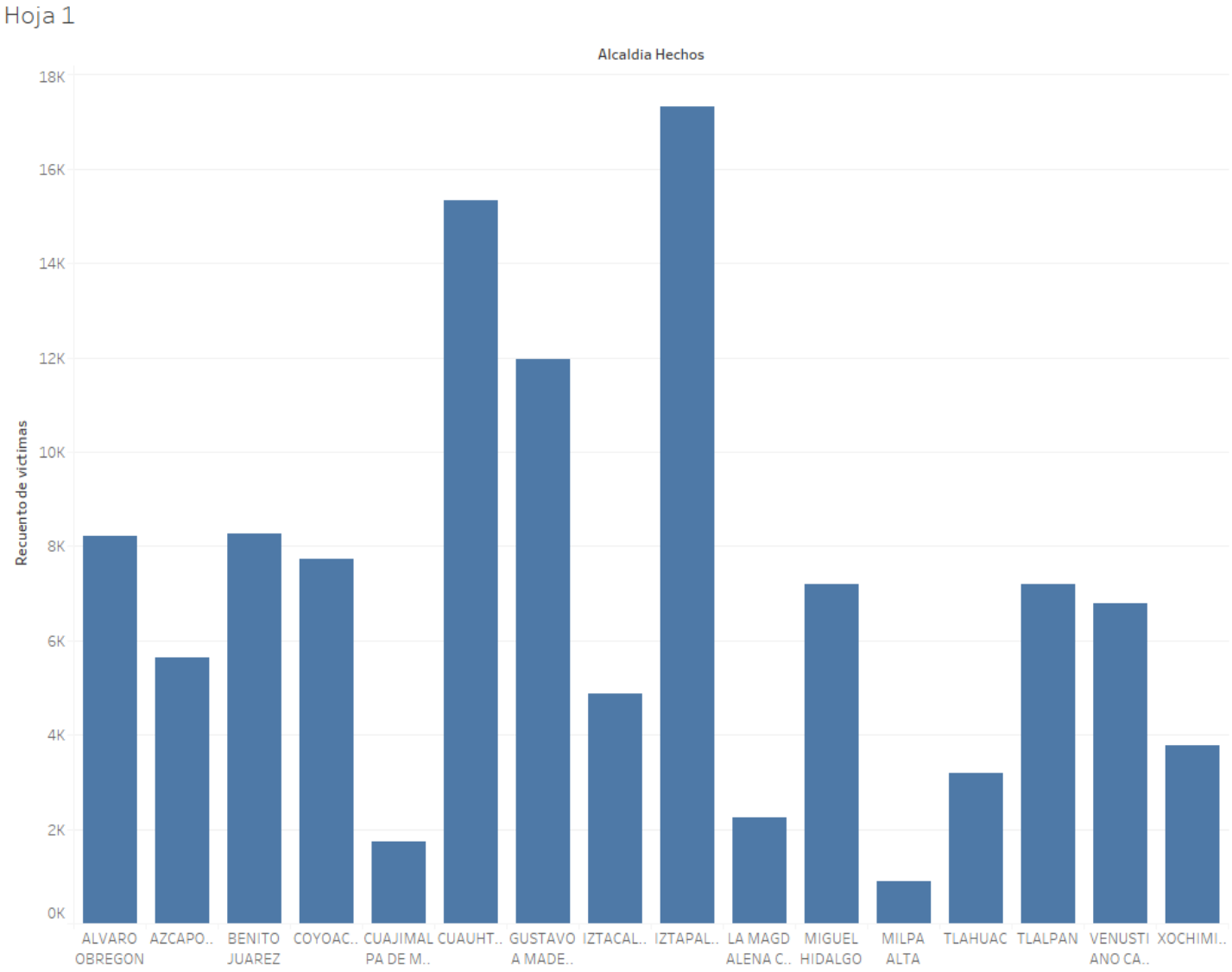

Figura 6: Grafica con el número de registros x alcaldía hechos

Se puede apreciar un gran número de hechos en la alcaldía de Iztapalapa, esto también nos puede ser de mucha ayuda en nuestro caso de estudio, dando una primera pauta, de en donde hay que enfocar la una mayor parte de nuestro cuerpo policial.

Analizando el dataset en el tiempo y en el espacio, se opto por la granularidad de la hora de los hechos, dándonos como resultado la siguiente gráfica:

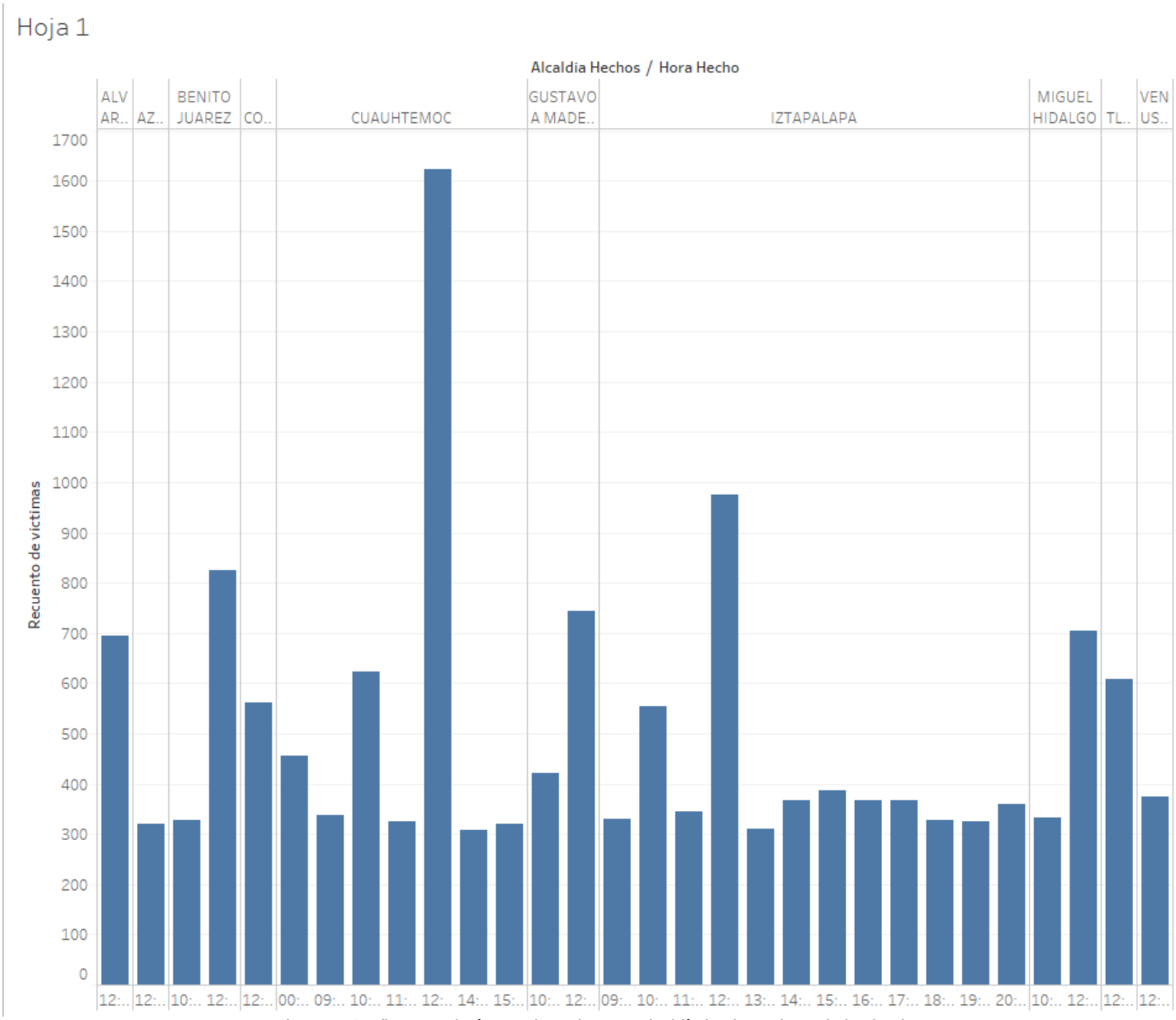

Figura 7: Grafica con el número de registros x alcaldía hechos y hora de los hechos

Esta es una información muy útil para nuestro caso de estudio, donde se puede notar que en las alcaldías Cuauhtémoc e Iztapalapa, coinciden las horas de los hechos, por lo que nos da una información mucho más precisa de lo que pude hacer nuestro cuerpo policiaco, dando una especial atención en esas zonas y horas.

Otra distribución de una dimensión temática importante es la categoría:

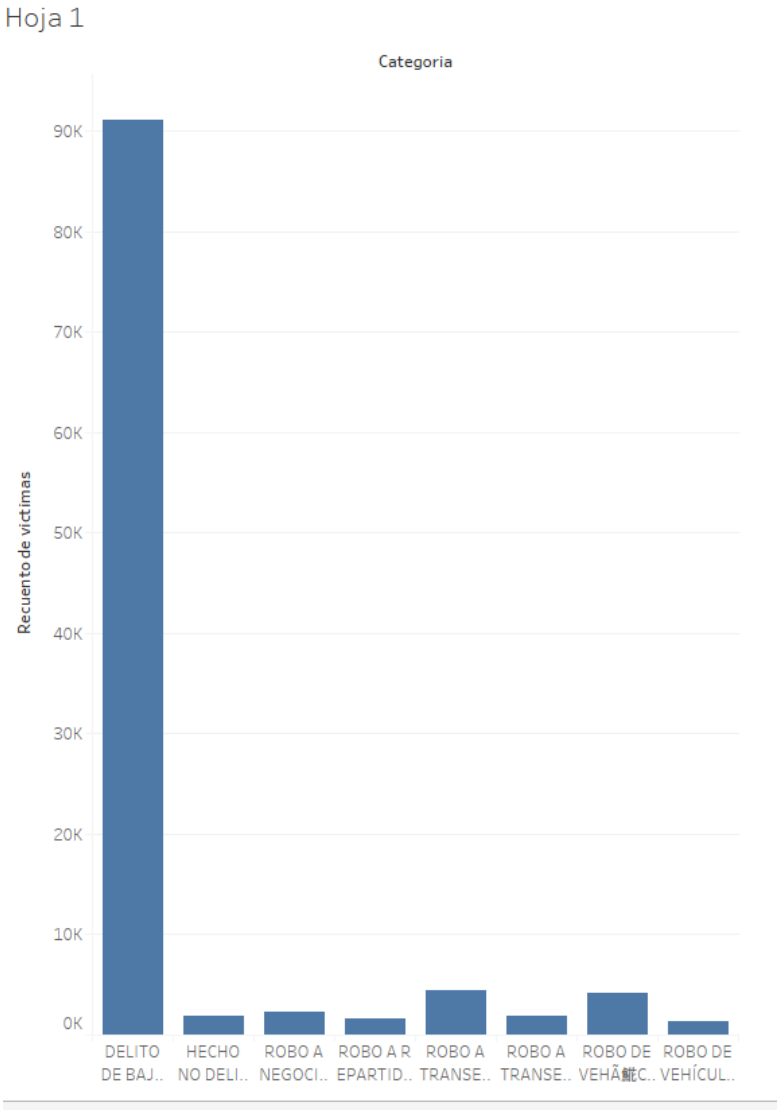

Figura 8: Grafica con el número de registros x categoría

Podemos apreciar que la más recurrente es el delito de bajo impacto, también notamos un símbolo raro en el robo de vehículo, afortunadamente es no hay información alterada, por lo que para solucionarlo basta con un update en el manejador de la base de datos.

Por ultimo analizado la inconsistencia de los datos, analizaremos el siguiente caso:

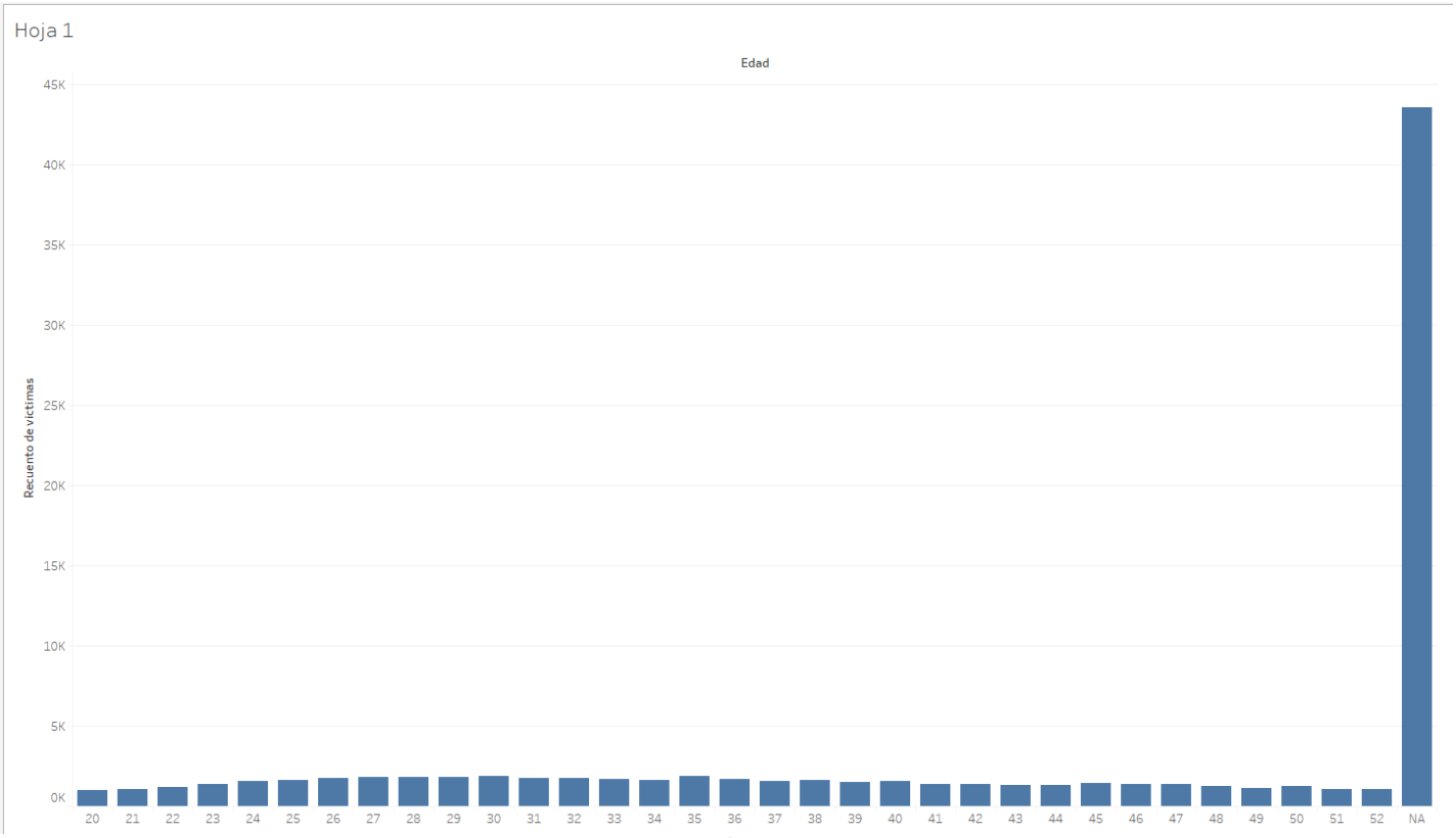

Figura 9: Grafica con el número de registros x edad

Se puede ver claramente el gran número de NA que existen en la base de datos, por lo tanto no es una opción viable deshacernos de estos, pero nos da una premisa interesante, ya que es sorprendente que no se conozca la edad de las victimas siendo un factor decisivo en alguna sentencia, por lo que es recomendable contactar con el proveedor de la información para corroborar la fidelidad de este dato.

Al final los únicos datos inconsistentes preocupantes son los caracteres raros que se encuentran, pero afortunadamente no interfieren con los propios datos al no dividirse sus valores o crear duplas extras.

Me gustaría finalizar haciendo hincapié en lo difícil que fue la exportación de los datos, realmente el tiempo fue mucho, pero al final se pudo lograr de una manera satisfactoria, al ser tan grande el dataset, en la exportación se me apago la computadora 3 veces por lo que tuve que dividir el dataset en 3 secciones viendo cual fue el ultimo registro, para no duplicar algún dato. El procesamiento del dataset es factible y nos puede dar mucha información interesante, en especial para nuestro caso propuesto, ya que al obtener una buena granularidad en el espacio y en el tiempo, es posible tomar acciones al respecto.

Por último le comparto el link de donde se obtuvo el dataset.

https://datos.cdmx.gob.mx/dataset/victimas-en-carpetas-de-investigacion-fgj