
**Universidad Nacional Autónoma de México**
**(UNAM)**

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
(IIMAS)

# CDMX
# Ciudad de Víctimas

Materia:
*Visualización de la Información*

Profesor:
*Dr. Luis Miugel De La Cruz Salas*

Autor:
*Ortega Ibarra Jaime Jesús*

Mayo 29, 2020

# Índice



# Índice de figuras



# Índice de tablas

# 1. Resumen

En los últimos años se ha visto un crecimiento exponencial en cuanto a la cantidad de delitos cometidos dentro de la Ciudad de México, en la actualidad se viven altos niveles de inseguridad, pero ¿A qué se debe dicha situación? Hoy en día muchas personas salen a las calles con el miedo de ser víctimas de robo, violencia, secuestro y en ciertas ocasiones hasta homicidio. Existe una gran inconformidad hacia las autoridades por falta de atención ante dicha situación. En este proyecto se buscará analizar y detectar cuales son los índices de la situación actual, es decir, Alcaldías con mayores índices delictivos, posibles vulnerabilidades para ser motivo de víctima, entre otras características.

# 2. Introducción

Mediante la base de datos obtenida se desea realizar un análisis de estos con la finalidad de obtener datos con mayor impacto para que un delito suceda, es decir, evaluar los distintos atributos con los que se cuenta para detectar posibles correlaciones entre ellos con el fin de identificar los valores más influyentes para la ocurrencia de algún delito.

Mediante este análisis obtendremos datos exactos tales como:

- Índices delictivos por colonia.
- Porcentaje de delitos de acuerdo al género de la víctima.
- Correlación entre los diferentes posibles delitos y el género de la víctima.

# 3. Fuente de datos

Para la manipulación de los datos, se ha decidido dividir en tres diferentes etapas:

- Obtención.
- Limpieza.
- Transformación.

De esta manera se logra identificar y organizar de manera eficiente el desarrollo de este proyecto.

## 3.1. Obtención de los datos

La base de datos fue conseguida a través del portal de Datos Abiertos de la CDMX, cuyo sitio es el siguiente:

`https://datos.cdmx.gob.mx/pages/home/`

Dentro de este, se pueden encontrar todos los datos de la Ciudad de México, desde temas jurídicos hasta temas ambientales y de transporte. En esta ocasión nos enfocamos en el apartado de Justicia y Seguridad, en el cual encontramos los datos Carpetas de investigación de la FGJ de la Ciudad de México, el cual contiene los registros de todas las denuncias que han sido realizadas en las distintas alcaldías dentro de la Ciudad de México y Zonas Metropolitanas. Los datos cuentan con un total de 330,240 registros, cada uno de estos con 23 Atributos diferentes, los cuales pueden ser descargados en distintos formatos (Json, Excel, CSV), en mi caso, por facilidad de manejo, decidí descargarlos como un archivo CSV, para poder ser procesados mediante

Python, específicamente realizando la lectura mediante la librería *Pandas*.
En la Tabla 1 podemos visualizar los datos mostrados desde el sitio web.


Tabla 1: Datos desde sitio web

## 3.2. Limpieza de los datos

Dentro de nuestros datos, encontramos ciertos registros incompletos, lo cual nos generaba valores nulos, para ello decidimos reemplazar dichos valores con el texto **No se especifica**, ya que para el procesamiento es importante no contar con valores nulos y estos no pueden ser eliminados, pues nos generaría ciertas inconsistencias al momento de realizar los análisis.
En la Tabla 2 podemos visualizar una muestra de los datos con los valores reemplazados.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VICTIMA | ... | 23/05/2019 | 23:30 | 00:04 | 19.36099,-99.04954 | IZTAPALAPA | UNIDAD VICENTE GUERRERO | ROSENDO SALAZAR | No se especifica | 19.360990 | -99.049540 |
| VICTIMA Y NUNCIANTE | ... | 23/05/2019 | 23:30 | 00:04 | 19.36099,-99.04954 | IZTAPALAPA | UNIDAD VICENTE GUERRERO | ROSENDO SALAZAR | No se especifica | 19.360990 | -99.049540 |
| VICTIMA Y NUNCIANTE | ... | 23/05/2019 | 22:20 | 00:05 | 19.35486,-99.10394 | IZTAPALAPA | SANTA ISABEL INDUSTRIAL | CALZADA ERMITA IZTAPALAPA | No se especifica | 19.354860 | -99.103940 |

Tabla 2: Datos limpios

## 3.3. Transformación

Para la transformación de los datos he decidido realizar un método de codificación tal como lo es OneHotEncoder ya que queremos transformar nuestras variables categóricas a numéricas, dado que el análisis que se desea realizar requiere el cálculo de la correlación entre los distintos atributos.

### 3.3.1. OneHotEncoder

Este método de codificación, lo podemos obtener mediante la librería *Scikitlearn*, dicha función codifica las características categóricas como una matriz numérica. La entrada a este transformador debe ser un conjunto de cadenas, que denotan los valores adquiridos por las características categóricas (discretas), estas se codifican utilizando un esquema de codificación $one - hot$, creando así una columna binaria para cada categoría y devolviendo una matriz dispersa según los parámetros.
En la Tabla 3 podemos visualizar nuestra tabla resultante al aplicar *OneHotEncoder*.

| | Sexo_Femenino | Sexo_Masculino | Sexo_No se especifica | Delito_VIOLENCIA FAMILIAR | Delito_ROBO A NEGOCIO SIN VIOLENCIA | Delito_FRAUDE | Delito_AMENAZAS | Delito_ROBO A TRANSEUNTE EN VIA PUBLICA CON VIOLENCIA | Delito_ABUSO SEXUAL |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **3** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Tabla 3: Datos resultantes de OneHotEncoder

Una vez obtenida dicha tabla, podemos realizar el cálculo de la correlación entre los distintos atributos, pero antes debemos saber ¿Qué es la correlación?
La correlación de dos atributos, se puede definir como el número que mide el grado de intensidad y el sentido de la relación entre dos variables. Esta correlación se define en términos de la varianza $S^2$ de las variables $x$ y $y$, así como la covarianza de estas, por lo tanto es una medida de la variación conjunta de ambas variables $cov(x, y)$, la cual se puede calcular mediante:

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(yi - \bar{y})}{N - 1} \tag{1}$$

Dentro de *python*, mediante la función $corr()$, podemos obtener el cálculo dentro de las variables de un conjunto de datos, devolviendo una matriz vista como una tabla indicando la correlación entre cada atributo.

| | Femenino | Masculino | Violencia Familiar | Robo A Negocio | Fraude | Amenazas | Abuso Sexual |
|---|---|---|---|---|---|---|---|
| **Femenino** | 1.000000 | -0.650125 | 0.300999 | -0.174246 | -0.001815 | 0.110579 | 0.146856 |
| **Masculino** | -0.650125 | 1.000000 | -0.144094 | -0.202536 | 0.064935 | -0.008085 | -0.086654 |
| **Violencia Familiar** | 0.300999 | -0.144094 | 1.000000 | -0.093833 | -0.085382 | -0.084056 | -0.045049 |
| **Robo A Negocio** | -0.174246 | -0.202536 | -0.093833 | 1.000000 | -0.066999 | -0.065958 | -0.035349 |
| **Fraude** | -0.001815 | 0.064935 | -0.085382 | -0.066999 | 1.000000 | -0.060018 | -0.032166 |
| **Amenazas** | 0.110579 | -0.008085 | -0.084056 | -0.065958 | -0.060018 | 1.000000 | -0.031666 |
| **Abuso Sexual** | 0.146856 | -0.086654 | -0.045049 | -0.035349 | -0.032166 | -0.031666 | 1.000000 |

Tabla 4: Tabla de correlación

Para mayor claridad se ha decidido observar visualizando creando un mapa de calor utilizando la librería *seaborn*, dando los siguietnes resultados:

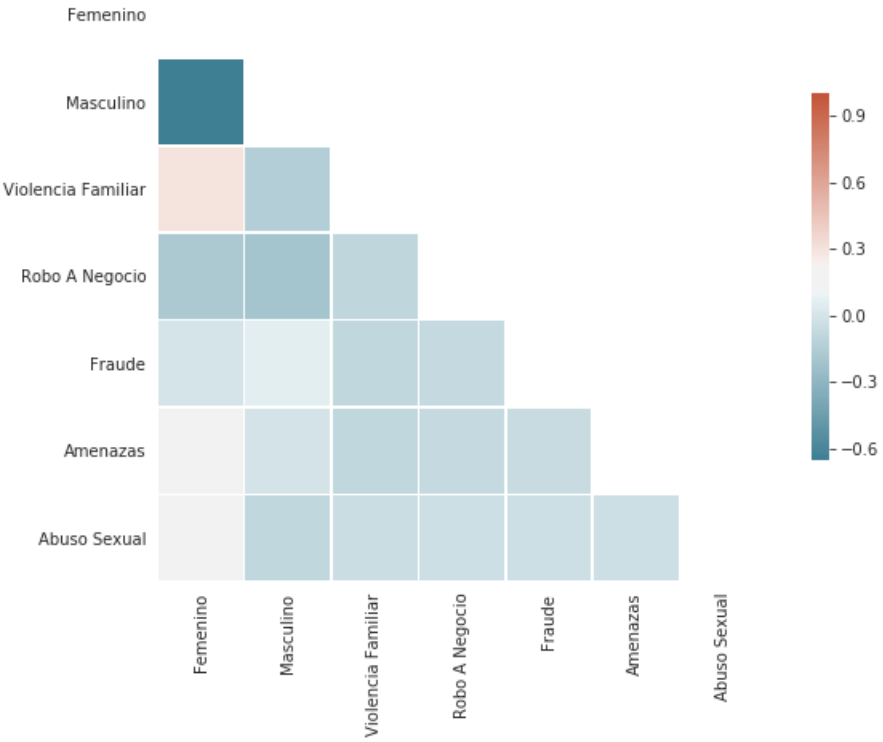


Figura 1: Mapa de calor correlaciones

Como se puede observar, encontramos mayor correlación entre el género Femenino y delitos tales como Violencia Familiar, Amenazas y Abuso sexual, mientras las víctimas pertenecientes al género Masculino, presentan mayor correlación con delitos como Fraude y Robo.

# 4. Cálculos matemáticos

Para la estimación de nuestros resultados, utilizamos Estadística Descriptiva, esto nos ayudará al realizar conteos de información específica para cada alguna de las zonas, estimaciones para definir intervalos y categorías, tales como:

Media

$Media(x) = \frac{\sum X_i}{N}$, Donde $(X_1, X_2, ..., X_n)$ representan el conjunto de observaciones.

Varianza

$s_x^2 = \frac{\sum (X_i - \bar{x})^2}{N-1}$, Siendo $(X_1, X_2, ..., X_n$ un conjunto de datos y $\bar{x}$ la media.

Y Probabilidad Clásica para dar un aproximado de riesgo en ciertas formas.

$P(E) = \frac{CasosFavorables}{CasosTotales}$

# 5. Gráficos

Para comenzar con nuestro análisis, determinamos la cantidad de denuncias que se realizan en cada mes, dentro de la siguiente figura, podemos observar dichas cantidades, de Mayo a Diciembre del 2019 y de Enero a Abril del 2020.

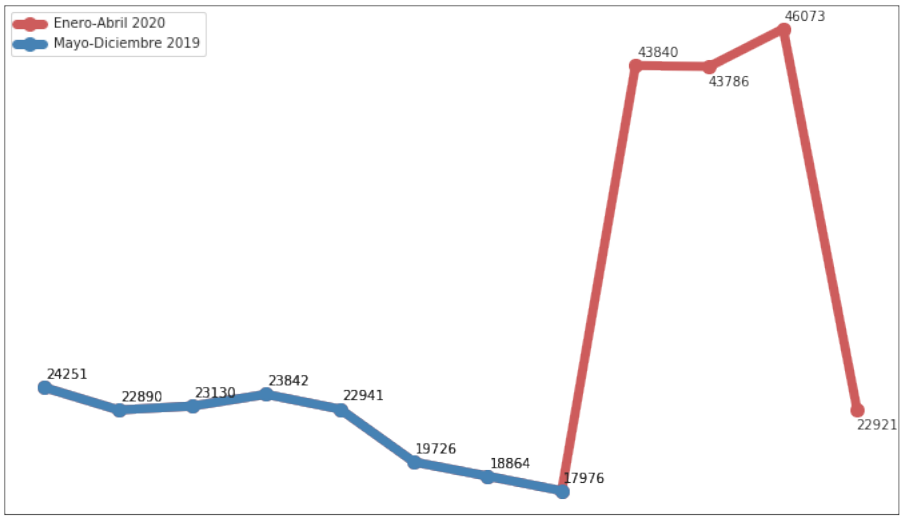


Figura 2: Delitos por mes

Una vez analizado esto, tomando en cuenta las correlaciones obtenidas entre los diferentes delitos con respecto al genero, decidimos obtener el porcentaje de delitos dividido en géneros, en la siguiente gráfica de pastel podemos observar los resultados obtenidos:

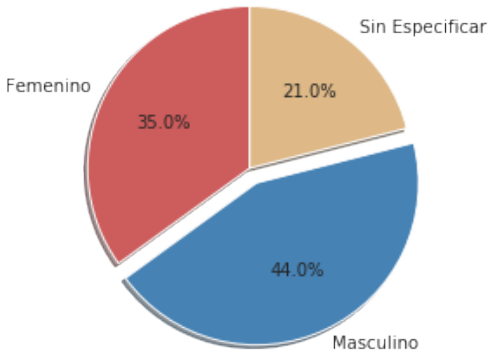


Figura 3: Delitos por género

Dentro de nuestro planteamiento inicial, se deseaba identificar las delegaciones con mayores índices delictivos, para ello realizamos un conteo de todos los delitos que se realizaron dentro de cada delegación durante el año 2019 y 2020, en la siguiente gráfica de barras podemos visualizar las 5 delegaciones con mayores índices delictivos, de acuerdo a la información obtenida.

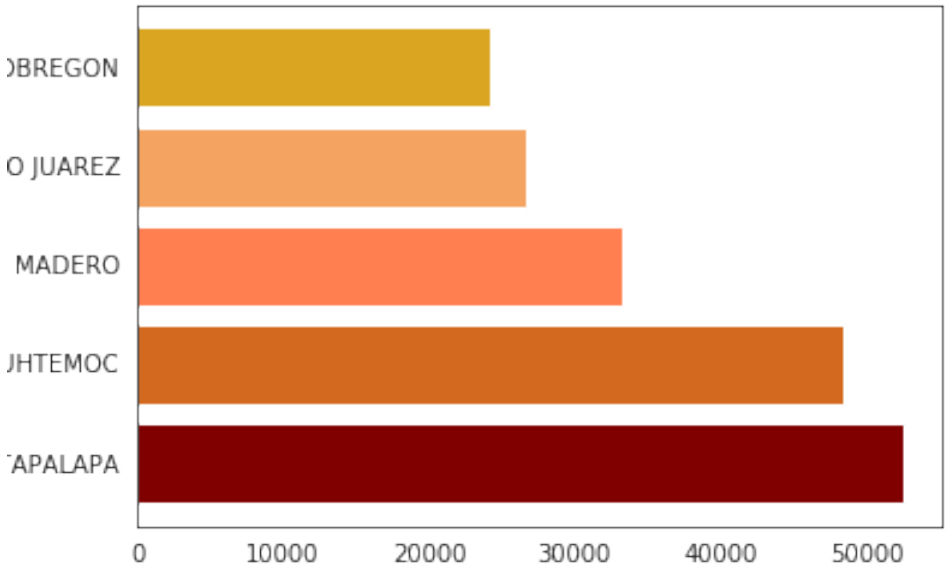


Figura 4: Delegaciones más delictivas

Otra manera de observar los índices de cada delegación, puede ser mediante un mapa de calor.

Mediante la librería de *Geopandas* y *Matplotlib*, podemos realizar la visualización de los puntos geográficos obtenidos de nuestro conjunto de datos, cabe mencionar que para dicha gráfica, hemos extraído nuestros datos como un formato *geojson* para mayor facilidad, posteriormente recortamos nuestros datos, para quedarnos únicamente con los atributos necesarios.

En la siguiente Tabla podemos observar los datos por graficar.

| | alcaldiahechos | geometry | Total |
|---|---|---|---|
| **0** | CUAUHTEMOC | POINT (-99.14867 19.45325) | 48355 |
| **1** | ALVARO OBREGON | POINT (-99.22126 19.34908) | 24271 |
| **2** | IZTAPALAPA | POINT (-99.00483 19.34981) | 52509 |
| **3** | COYOACAN | POINT (-99.16046 19.31976) | 20976 |
| **4** | COYOACAN | POINT (-99.16046 19.31976) | 20976 |

Tabla 5: Datos GeoJson

Una vez realizado esto, procedemos a graficar los diferentes puntos de acuerdo a su localización, lo cual nos devolverá un mapa de calor, pues cada punto tendrá una tonalidad de acuerdo a la cantidad de delitos que se han realizado dentro de esa misma delegación.

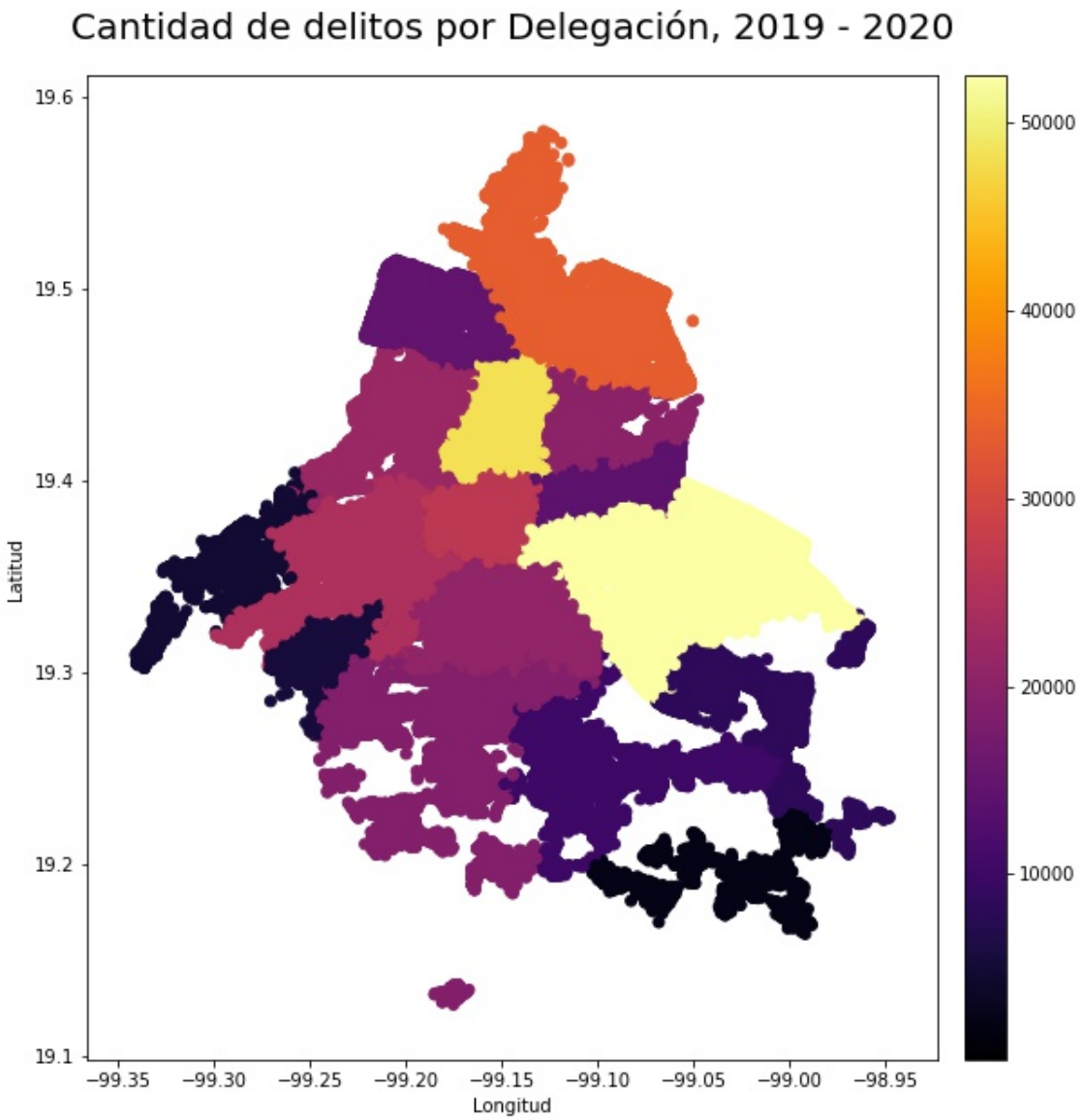


Figura 5: Mapa de Calor

# 6. Conclusión

Dentro de nuestros datos pudimos detectar que la delegación con mayor riesgo de acuerdo a sus índices delictivos es Iztapalapa, por lo cual se puede realizar algún plan de estrategia para solucionar dicho problema, además a pesar de tener mayor cantidad de casos cuyo género es Masculino, el género Femenino suele ser más vulnerable a distintos delitos, mayormente de violencia.

# 7. Referencias


- Pablo Vinuesa. (2016). Tema 8 - Correlación: teoría y práctica. 14 de Octubre, de CCG-UNAM Sitio web: www.ccg.unam.mx/ vinuesa/R4biosciences/docs/Tema8$_c orrelacion.html/introduccion-el-concepto-de-correlacion$.
- Reúl Estévez. (2019). Mapas con Python utilizando geopandas y matplotlib. 1 de Agosto, de Geomapik Sitio web: http://www.geomapik.com/desarrollo-programacion-gis/mapas-con-python-geopandas-matplotlib/
- No Identificado. (2019). Gráfico de correlación básica. Junio, de Tutorial Pedia Sitio web: https://riptutorial.com/es/seaborn/example/31922/grafico-de-correlacion-basica
- Desarrolladores Scikit-Learn. (2007-2019). sklearn.preprocessing.OneHotEncode. 2020, de Scikit-learn Sitio web: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html