PRACTICA 2

By

César Iván Rodríguez Sánchez

Ciclo de Vida de los Datos

PREGUNTAS

PREGUNTAS A RESPONDER EN LA PRACTICA

DESCRIPCION DEL DATASET

Los datos describen el hurto de automotores del 01 de enero al 31 de diciembre de 2018, reportados por la Dijin de la Policia Nacional.

El conjunto se encuentra publicado en formato CSV, el cual cuenta con 9532 filas y 25 columnas.

Campo	**Descripción**
Fecha	Fecha del hurto
Departamento	Departamento de Colombia
Municipio	Municipio dentro del departamento
Día	Día de la semana del hurto
Hora	Hora del hurto
Barrio	Barrio dentro del municipio donde se cometió el hurto
Zona	Rural , Urbana u Otra
Clase de sitio	Tipos de sitios especificos donde se realizó el hurto por ejemplo Via Publica, Aeropuerto, Centro Comercial, etc.
Arma empleada	Arma empleada en el hurto
Móvil Agresor	Medio transporte utilizado por el agresor
Móvil Victima	Medio transporte utilizados por la víctima
Edad	Edad de la víctima
Sexo	Genero de la Víctima
Estado civil	Estado civil de la víctima
País de nacimiento	Pais de nacimiento de la victima
Clase de empleado	Actividad económica de la víctima
Profesión	Profesión o rama académica de la víctima
Escolaridad	Nivel académico de la víctima
Código DANE	Código DANE del municipio
CLASE	Clase del automotor o vehiculo hurtado por ejemplo automovil, bus , motocicleta.
MARCA	Marca del vehiculo
LINEA	Linea del vehiculo hurtado
MODELO	Año de fabricación del vehiculo hurtado
COLOR	Color del vehiculo hurtado
Cantidad	Cantidad de vehiculos hurtados

Importancia

Este Dataset es importante por que permite describir el problema del hurto de vehículos en Colombia, se pueden identificar tendencias de hurto por sitios, por municipios, por tipos de vehículos.

Problema a responder

Qué patrones se pueden identificar en los hurtos de vehículos en Colombia.

2. INTEGRACION Y SELECCION DE LOS DATOS

En esta fase de acuerdo con los datos seleccionados no hay necesidad de integración a partir de otras fuentes sin embargo si es posible realizar análisis de redundancia y correlación de los atributos de los data set, para identificar atributos redundantes. Igualmente se pueden revisar filas para encontrar observaciones repetidas, y teniendo en cuenta el problema decidir los atributos a analizar.

Se verificó el dataset y no se encontraron datos repetidos.

Teniendo en cuenta la descripción de los campos se encontró redundancia entre el campo Municipio y Código DANE, ya que cada código pertenence a un municipio por lo tanto se elimina este atributo.

El resto de atributos nos pueden ayudar a entender los hurtos de automotores.

3. LIMPIEZA DE LOS DATOS

Los datos contienen ceros o elementos vacíos?

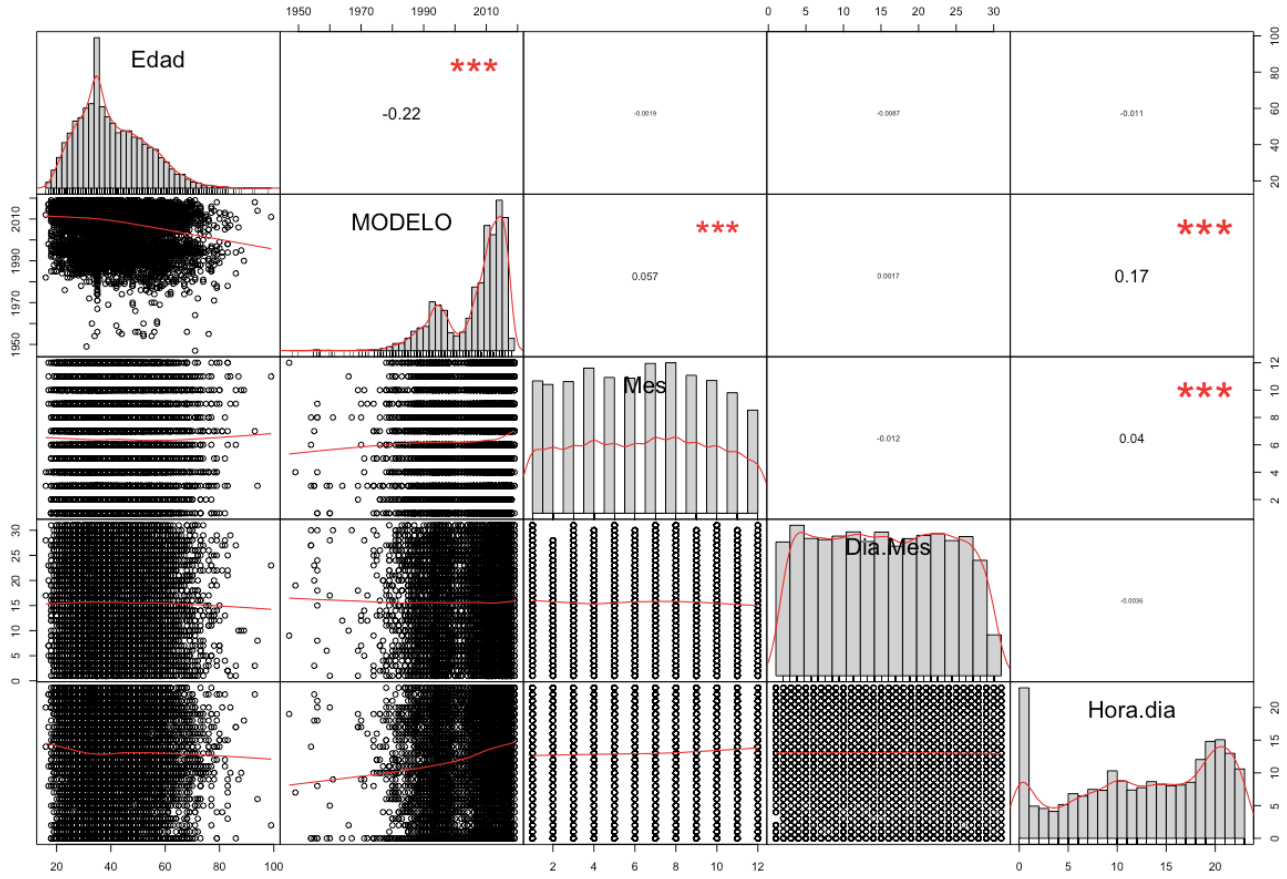
El atributo fecha tenia un valor invalido en la última fila, se eliminó ya que correspondía a una fila de total vacía.

Los atributos Departamento, Municipio, Hora, Zona no tienen datos vacios.

El atributo Barrio tiene 5 filas en “-” las cuales se colocan en DESCONOCIDO, para que sea más claro, en el momento que se analicen.

##

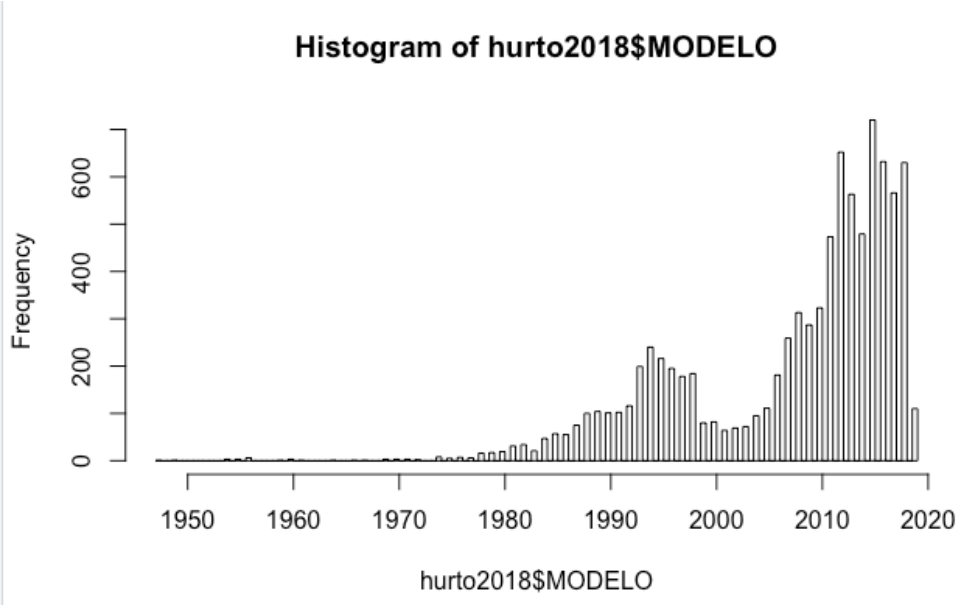
4.ANALISIS DE LOS DATOS



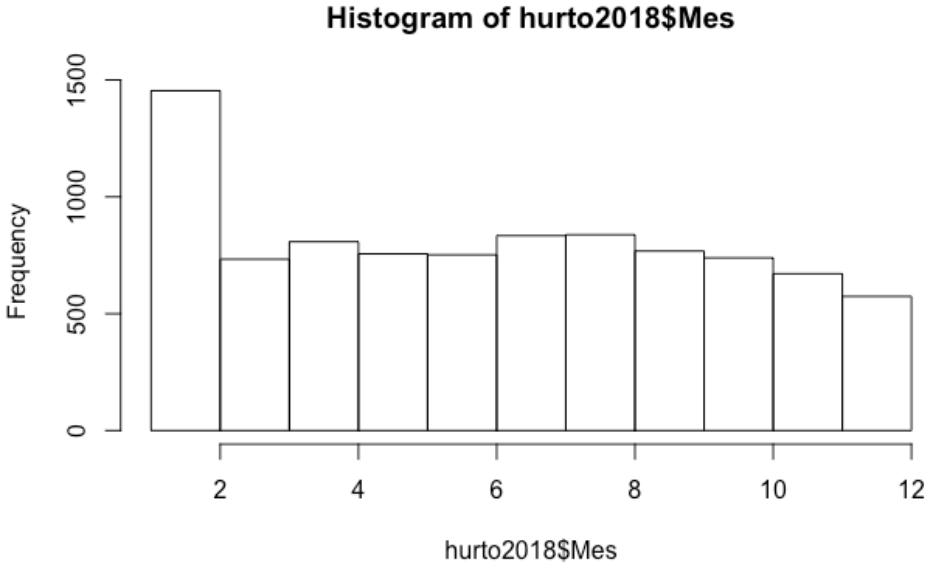
Edad	MODELO	Mes	Dia.Mes	Hora.dia
Min. :16.00	Min. :1947	Min. : 1.000	Min. : 1.00	Min. : 0.00
1st Qu.:32.00	1st Qu.:1999	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: 7.00
Median :38.00	Median :2011	Median : 6.000	Median :16.00	Median :14.00
Mean :40.75	Mean :2007	Mean : 6.386	Mean :15.61	Mean :12.83
3rd Qu.:49.00	3rd Qu.:2015	3rd Qu.: 9.000	3rd Qu.:23.00	3rd Qu.:20.00
Max. :99.00	Max. :2019	Max. :12.000	Max. :31.00	Max. :23.00

En cuanto al análisis descriptivo de las variables estadísticas se puede ver de manera general que la correlación entre las misma es muy baja entre todas las variables.

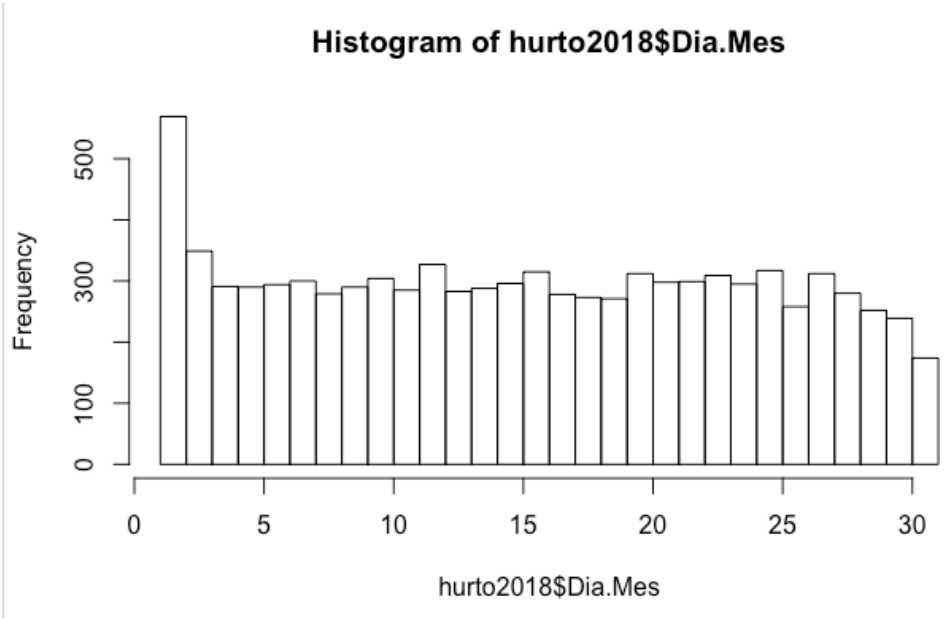
La edad tiene una concentración alta hacia los 30 y 35 años, con una mayor tendencia hacia las edades menores.



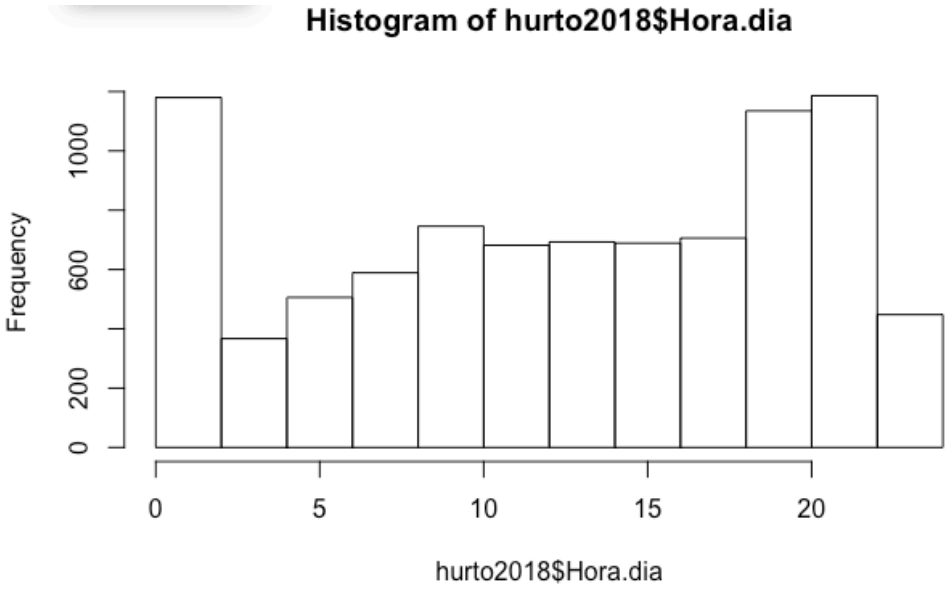
El modelo tiene una tendencia hacia los modelos mas recientes y otra entre los modelos 1993 a 1998. Hay un cambio notorio entre los modelos 1998 y 1999, el número de automotores hurtados disminuyó de 184 a 80 y asi se mantuvo el valor similar para los siguientes 6 años.



El mes del año tiene una tendencia predominante en Enero.

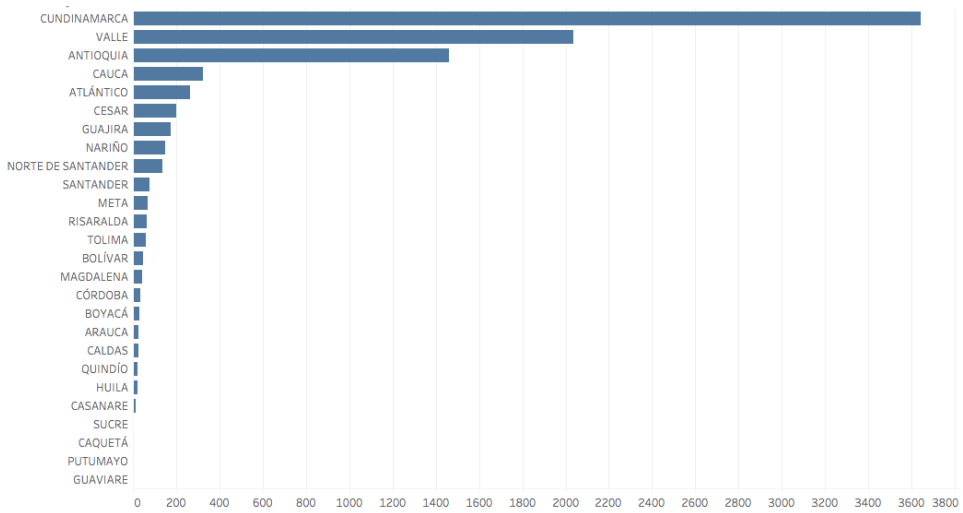


El dia del mes tiene un dia predominante el primer dia del mes.



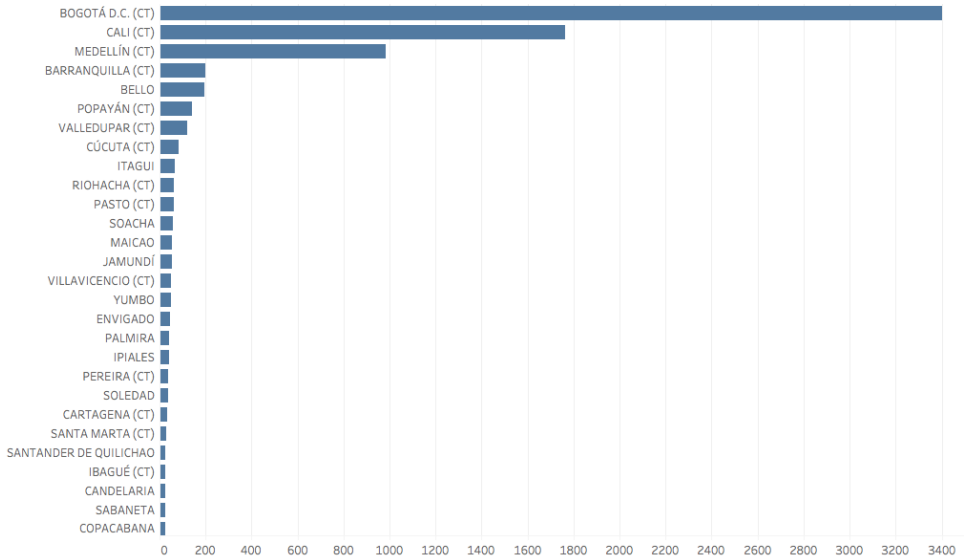
En cuanto a la hora del día hay un pico a las 0 horas y otro de 7 a 9 pm. El pico a las 0 horas no es muy explicable, por lo tanto puede causarse por la calidad de los datos.

5. REPRESENTACIÓN DE LOS DATOS A PARTIR DE TABLAS Y GRÁFICAS

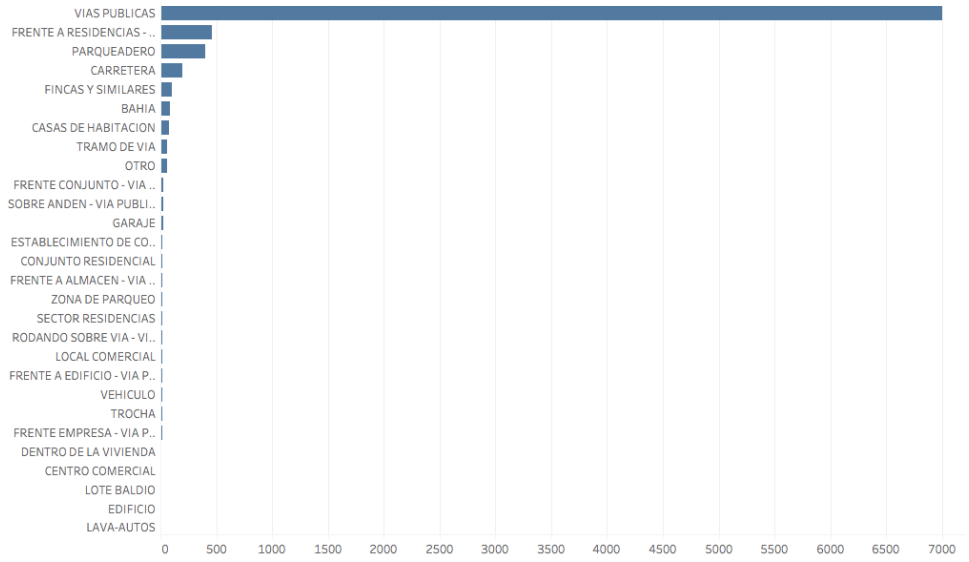


Se puede observar que la mayoría de hurtos se realizaron en el departamento de Cundinamarca seguido por Valle del Cauca.

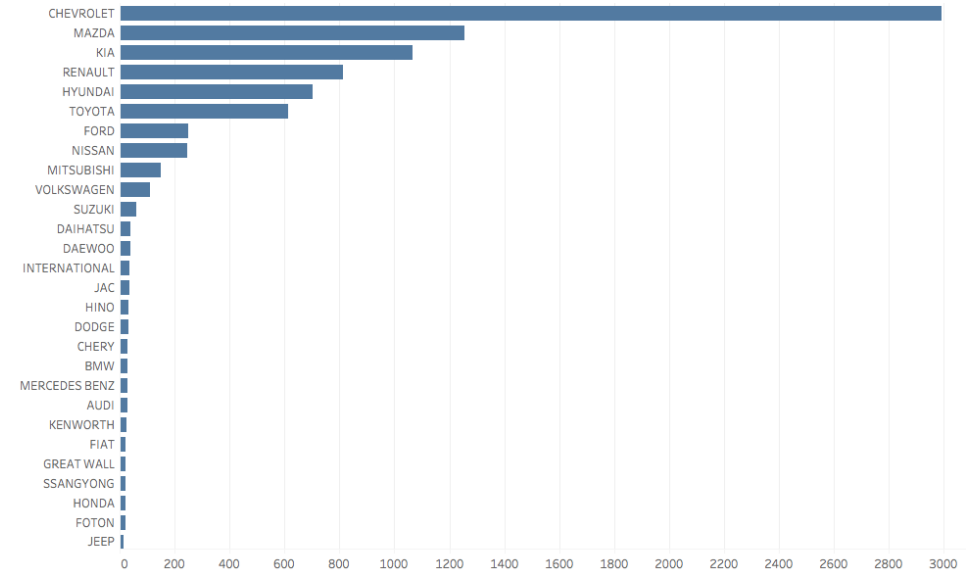
El segundo puesto de Valle del Cauca es un resultado notorio ya que no es el segundo departamento con mayor población de Colombia.



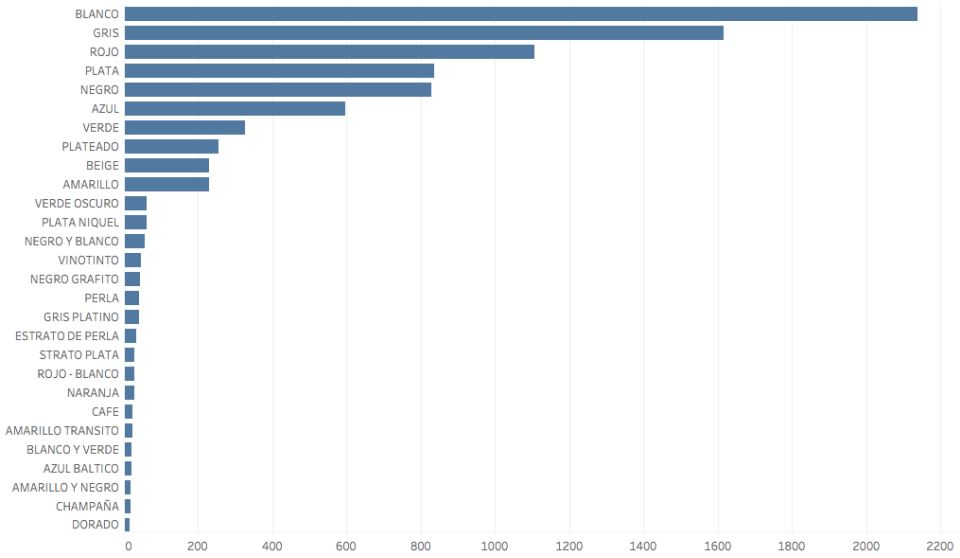
El municipio con mayores hurtos fue Bogotá como se espera por su mayor población, sin embargo, le sigue Cali que no es la segunda ciudad en población.



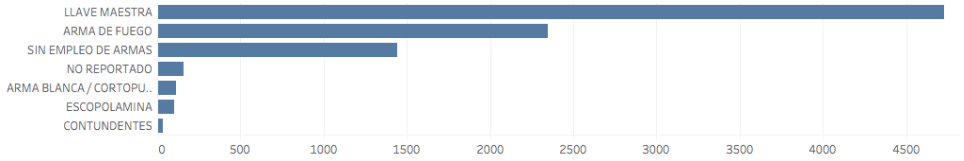
La clase de sitio donde mas se reportaron hurtos fue en la vías públicas siendo casi el 80% de los robos.



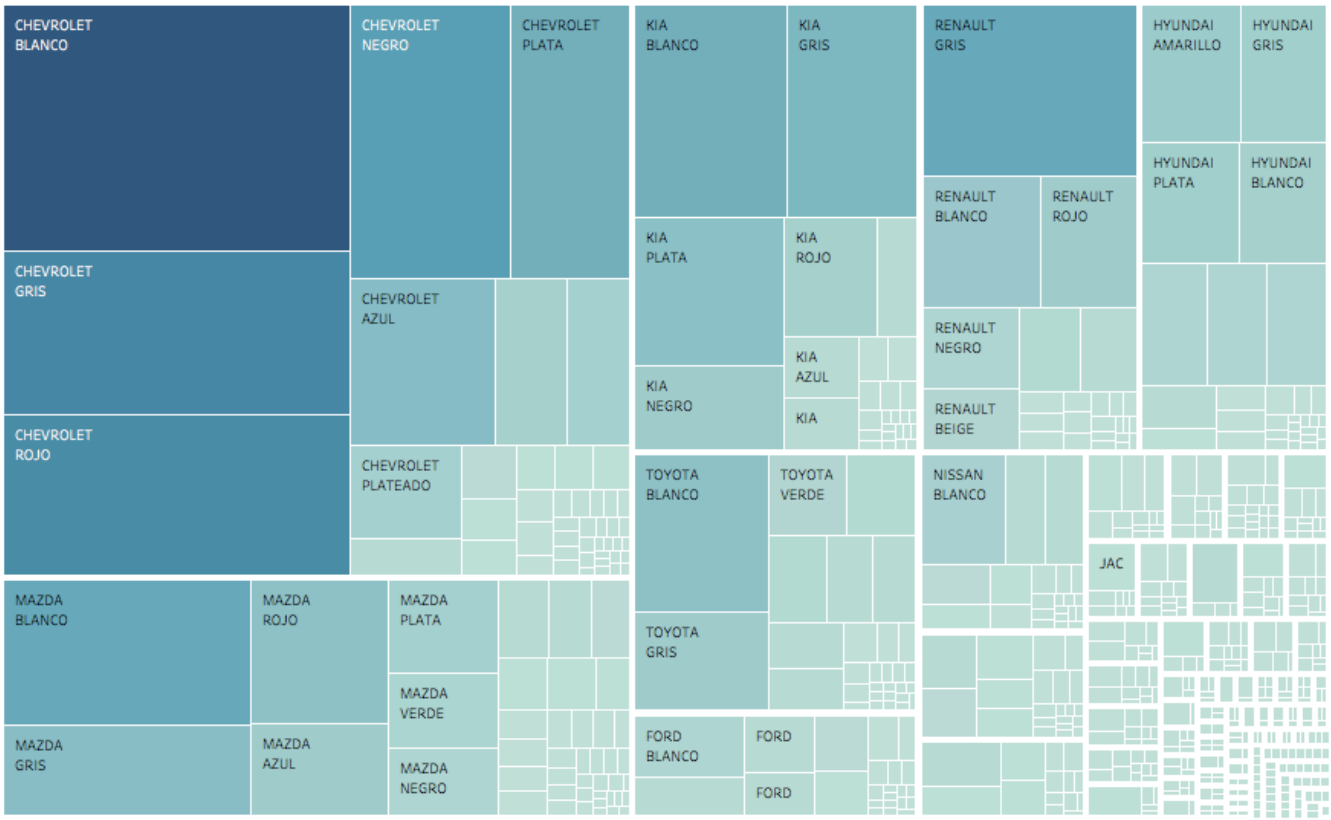
La Marca con mayores hurtos registrados es Chevrolet, con un 34% seguido por Mazda con 14% y Kia con 12%



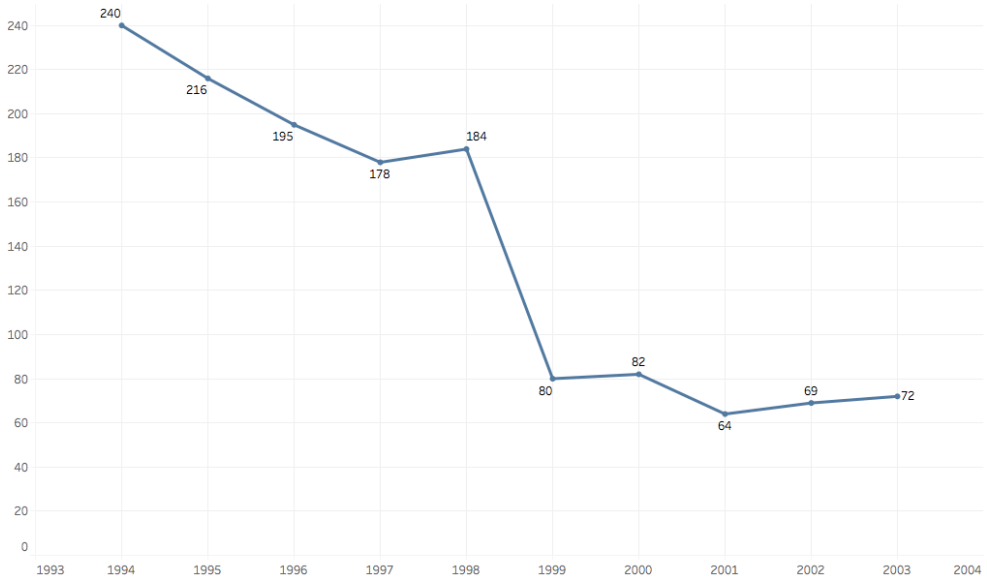
En cuanto al color del automotor el blanco, gris, rojo, plata y negro fueron los más hurtados en el 2018. Si unimos el gris con el plata, tanto el gris como el blanco serian los mas hurtados.



El arma empleada mas reportada fue la llave maestra con un 53% seguido por Arma de fuego con 26%, si se suman los robos realizados sin arma (llave maestra, Sin empleo de armas, y no reportado) se puede observar que aproximadamente un 71% de los hurtos se realizó sin armas.



Combinando la marca con el color se puede observar que el más hurtado fue el Chevrolet blanco con un 8% de todos los hurtos en 2018.



Esta gráfica muestra el detalle del cambio espontáneo que hay en el hurto de autos anteriores a 1998 y posteriores a 1999, se puede observar que el hurto de

autos posteriores a 1999 descendió a menos de la mitad, de 184 modelo 1998 a 80 modelo 1990.

Este comportamiento en los datos sugiere que puede existir una característica en los automotores modelos 1999 y anteriores que es más atractiva para los delincuentes, o que alguna característica se implementó en estos que mejoró su seguridad disminuyendo su probabilidad de robo.

RESOLUCIÓN DEL PROBLEMA

Se identificaron varios patrones:

1. El rango de modelos de automotores de 1993 a 1998 tuvieron una mayor frecuencia de Hurto en 2018, esto puede sugerir que estos modelos son más atractivos para los delincuentes que los modelos de años similares 1990-2005.
2. La mayoría de hurtos se realizan sin la utilización de armas, por lo que se podrían atribuir a descuidos del propietario, o deficiencias en la seguridad de los vehículos. Este hallazgo podría generar políticas para disminuir los hurtos por descuidos o deficiencias en seguridad.
3. Los vehículos con mayor hurto fueron los Chevrolet de color blanco, gris y rojo, sumando entre ellos el 19% de todos los automotores hurtados en 2018.
4. Cali a pesar de ser la tercera ciudad en población de Colombia tuvo en 2018 la segunda mayor cantidad de hurtos de automotores, con casi el doble de hurtos que Medellín la segunda ciudad más grande de Colombia.
5. Cartagena a pesar de ser la quinta Ciudad en Población esta 22 en la cantidad de hurto de automotores.
6. Por el contrario Popayán que es el municipio numero 24 en población de Colombia, es sexta en la cantidad de hurto de vehiculos.

Aunque se identificaron varios patrones útiles para entender y describir el problema de hurto de automotores en Colombia un análisis más detallados de los datos podrían descubrir patrones menos evidentes. Incluso en la realización de la conclusiones se pudo observar la necesidad de integrar otros datos como los de población por municipios y parque automotor. Para poder analizar el problema en términos de densidad poblacional y proporcionales a los vehículos matriculados.

REFERENCIAS

"Megan Squire (2015). Clean Data. Packt Publishing Ltd"

"Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann"

"Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews"

Contribuciones	Firma
Investigación Previa	CIRS
Redacción de las respuestas	CIRS
Desarrollo código	CIRS