Programa de Ciencia de Datos
Curso: Big Data
Profesor: Juan Manuel Esquivel
Estudiante: María Yorleni Alfaro Alfaro

Proyecto de Big Data

Investigación Preliminar – Primer Problema

Fuentes de datos analizadas

Para esta investigación se procedió a buscar y a analizar información de diferentes fuentes tanto a nivel laboral como fuentes de datos abiertas.

A nivel laboral no fue posible obtener información dado que la mayor parte de la información es de carácter confidencial, y la información considerada como pública es información que ya se encuentra agregada por lo cual no es funcional por los requerimientos del proyecto.

También se analizaron diferentes fuentes de datos de instituciones como el INEC, Programa Estado de la Nación, CCSS, diferentes ministerios del gobierno, etc.

Producto de este análisis se decidió escoger las siguientes dos fuentes de datos, para el desarrollo del proyecto:

- Base de datos de educación primaria, tomada del Programa Estado de la Nación, que muestra una serie de características de las escuelas públicas y privadas en Costa Rica.
- Set de datos de los códigos geográficos de Costa Rica, tomado del INEC, que muestra la división del territorio de Costa Rica por provincia, cantón, distrito, según código.

El motivo por el cual se eligió la fuentes de datos de educación primaria es porque considero que contiene información muy valiosa que puede servir para realizar análisis de información de diferentes ámbitos como excelencia en la educación primaria (ya que muestra información de aprobados, reprobados, repitentes, etc), problemas sociales (estudiantes que han abandonado la escuela, embarazos en estudiantes, casos atendidos por violencia, robos, destrucción de materiales, etc), infraestructura en las escuelas (cantidad de aulas, laboratorios, comedores, bibliotecas, sodas, computadoras, etc). Todo lo anterior nos permite ver de una manera más detallada la realidad de Costa Rica en estos ámbitos.

La base de datos de educación primaria del Programa Estado de la Nación contiene los códigos geográficos de la provincia, cantón y distrito de cada centro educativo, pero no el nombre de estos. Es por esta razón que se decidió escoger como segunda fuente de datos la información de los códigos geográficos de Costa Rica de manera que se pueda conocer a qué provincia, cantón y distrito corresponde cada centro educativo, lo cual ayuda a enriquecer de gran manera la información de la base de datos de educación primaria, ya que nos podría permitir realizar análisis de información de los diferentes ámbitos mencionados anteriormente, pero a nivel de provincias, cantones o incluso distritos.

Descripción detallada de los datos

Educación Primaria

El set de datos de educación primaria, del Programa Estado de la Nación, contiene información de diferentes características de las escuelas en Costa Rica, desde el año 2000 hasta el año 2016, por lo tanto, se decidió tomar la información únicamente del año 2015, eliminando todo el resto de las columnas que no corresponden a este año. Además muchas de las columnas estaban dadas por año y por grado (1º , 2º , 3º , 4º, 5º y 6º) , por lo que se decidió utilizar únicamente las columnas que contienen el total por centro educativo (eliminando las que contienen los total por cada grado).

En la siguiente tabla se muestra el detalle de cada uno de los campos del set de datos de educación primaria:

Columna	Tipo de dato	Descripción
llave	Numérico	Llave única del registro
nombre_ins	String	Nombre del centro educativo
creacion00	Numérico	Año de creación de la institución
direg15	Numérico	Dirección Regional Educativa
cdpr15	Numérico	Código de Provincia
cdcan15	Numérico	Código de Cantón
cddis15	Numérico	Código Distrital
regplan15	Numérico	Región de Planificación
zona15	Numérico	Código de Zona
mit_15	Numérico	Matrícula Inicial Total en el año 2015
mih_15	Numérico	Matrícula Inicial de Hombres en el año 2015
rt_15	Numérico	Repitentes Total 2015
rh_15	Numérico	Repitentes Hombres 2015
aprobt_15	Numérico	Aprobados Total 2015
aprobh_15	Numérico	Aprobados Hombres 2015
reprot_15	Numérico	Reprobados Total 2015
reproh_15	Numérico	Reprobados Hombres 2015
desa_15	Numérico	Abandono Total 2015
desah_15	Numérico	Abandono Hombres 2015
desert_15	Numérico	Exclusión intra-anual Total 2015
deserh_15	Numérico	Exclusión intra-anual Hombres 2015
act_15	Numérico	Estudiantes con Adecuación de Acceso Total 2015
ach_15	Numérico	Estudiantes con Adecuación de Acceso Hombre 2015
nst_15	Numérico	Estudiantes con Adecuación No Significativa 2015
nsh_15	Numérico	Estudiantes con Adecuación No Significativa Hombres 2015
sit_15	Numérico	Estudiantes con Adecuación Significativa Total 2015
sih_15	Numérico	Estudiantes con Adecuación Significativa Hombre 2015

embt_15	Numérico	Estudiantes Embarazadas total 2015
embmenor_15	Numérico	Estudiantes Embarazadas Menor de 18 años 2015
embmayor_15	Numérico	Estudiantes Embarazadas con 18 años o más 2015
aat15	String	Aulas para impartir lecciones I y II ciclos total 2015
aab15	String	Aulas para impartir lecciones I y II ciclos buenas 2015
apt15	Numérico	Aulas para impartir lecciones Educación Preescolar total 2015
apb15	Numérico	Aulas para impartir lecciones Educación Preescolar buenas 2015
aest15	Numérico	Aulas para impartir lecciones Aula Integrada total 2015
aesb15	Numérico	Aulas para impartir lecciones Aula Integrada 2015
aaet15	Numérico	Aulas para impartir lecciones Aula Edad total 2015
aaeb15	Numérico	Aulas para impartir lecciones Aula Edad buenas 2015
anat15	Numérico	Aulas que no se utilizan para impartir lecciones total 2015
anab15	Numérico	Aulas que no se utilizan para impartir lecciones buenas 2015
inft15	Numérico	Laboratorio de Informática Total 2015
infb15	Numérico	Laboratorio de Informática Bueno 2015
olat15	Numérico	Otro Laboratorio Total 2015
olab15	Numérico	Otro Laboratorio Total Bueno 2015
salt15	Numérico	Sala de Profesores Total 2015
salb15	Numérico	Sala de Profesores Bueno 2015
comt15	Numérico	Comedor Total en 2015
comb15	Numérico	Comedor Bueno en 2015
bibt15	Numérico	Biblioteca Total en 2015
bibb15	Numérico	Biblioteca Bueno en 2015
gimt15	Numérico	Gimnasio en 2015
gimb15	Numérico	Gimnasio Bueno en 2015
talt_ai15	Numérico	Taller de Artes Industriales Total 2015
talb_ai15	Numérico	Taller de Artes Industriales Bueno en 2015
otalt15	Numérico	Otros Talleres Total 2015
otalb15	Numérico	Otros Talleres Bueno en 2015
sodt15	Numérico	Soda Total 20015
sodb15	Numérico	Soda Bueno 20015
indt15	Numérico	Inodoros Total 2015
indb15	Numérico	Inodoros Bueno 2015
lavt15	Numérico	Lavatorios Total 2015
lavb15	Numérico	Lavatorios Bueno 2015
sant15	Numérico	Servicio Sanitario Accesible Total 2015
sanb15	Numérico	Servicio Sanitario Accesible Bueno 2015
tvt15	Numérico	Televisión Total 2015
tvb15	Numérico	Televisión Bueno 2015
vbt15	Numérico	Proyector de video (Video Beam) Total 2015
vbb15	Numérico	Proyector de video (Video Beam) Bueno 2015
dvdt15	Numérico	DVD Total 2015
dvdb15	Numérico	DVD Bueno 2015

cetoi15	Numérico	Computadoras escritorio con internet
cetos15	Numérico	Computadoras escritorio sin internet
cepei15	Numérico	Computadoras escritorio con internet Pedagógico
cepes15	Numérico	Computadoras escritorio sin internet Pedagógico
cepai15	Numérico	Computadoras escritorio con internet Pedagógico y Administrativo
cepas15	Numérico	Computadoras escritorio sin internet Pedagógico y Administrativo
ceadi15	Numérico	Computadoras escritorio con internet Administrativo
ceads15	Numérico	Computadoras escritorio sin internet Administrativo
cptoi15	Numérico	Computadoras portátil con internet
cptos15	Numérico	Computadoras Portátil sin internet
cppei15	Numérico	Computadoras portátil con internet Pedagógico
cppes15	Numérico	Computadoras portátil sin internet Pedagógico
cppai15	Numérico	Computadoras portátil con internet Pedagógico y Administrativo
cppas15	Numérico	Computadoras portátil sin internet Pedagógico y Administrativo
cpadi15	Numérico	Computadoras portátil con internet Administrativo
cpads15	Numérico	Computadoras portátil sin internet Administrativo
bib15	Numérico	Servicio de Biblioteca
sal15	Numérico	Servicio de Salud
pla15	Numérico	Servicio de Planes de Emergencia
aux15	Numérico	Servicio de Primeros Auxilios
serv_int15	Numérico	Servicio de Internet
expto_15	Numérico	Expulsiones por agresión total 2015
expdef_15	Numérico	Expulsiones por agresión definitivas 2015
exptem_15	Numérico	Expulsiones por agresión temporales 2015
agrve_15	Numérico	Casos atendidos entre estudiantes de violencia verbal 2015
agrvep_15	Numérico	Casos atendidos entre estudiantes y docentes de violencia verbal 2015
agrveo_15	Numérico	Casos atendidos entre estudiantes y otro personal de violencia verbal 2015
agrfe_15	Numérico	Casos atendidos entre estudiantes de violencia física 2015
agrfep_15	Numérico	Casos atendidos entre estudiantes y docentes de violencia física 2015
agrfeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de violencia física 2015
agree_15	Numérico	Casos atendidos entre estudiantes de violencia escrita 2015
agreep_15	Numérico	Casos atendidos entre estudiantes y docentes de violencia escrita 2015
agreeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de violencia escrita 2015
agrre_15	Numérico	Casos atendidos entre estudiantes de robos 2015
agrrep_15	Numérico	Casos atendidos entre estudiantes y docentes de robos 2015
agrreo_15	Numérico	Casos atendidos entre estudiantes y otro personal de robos 2015
agrde_15	Numérico	Casos atendidos entre estudiantes de destrucción de materiales 2015
agrdep_15	Numérico	Casos atendidos entre estudiantes y docentes de destrucción de materiales 2015
agrdeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de destrucción de materiales 2015
agroe_15	Numérico	Casos atendidos entre estudiantes de otros tipos de violencia 2015

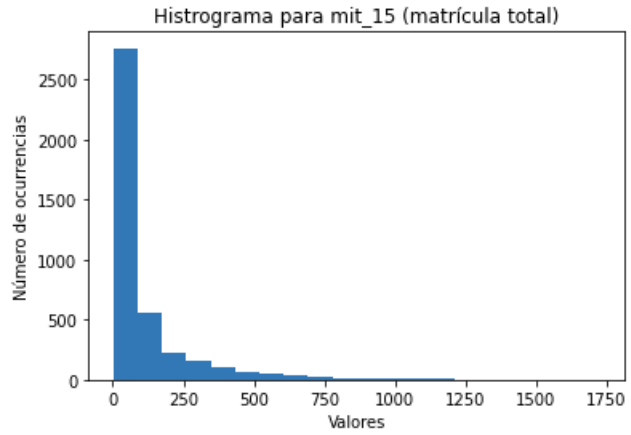
agroep_15	Numérico	Casos atendidos entre estudiantes y docentes de otros tipos de violencia 2015
agroeo_15	Numérico	Casos atendidos entre estudiantes y otro personal de otros tipos de violencia 2015
int_15	Numérico	Matrícula Inicial en Inglés Total 2015
inht_15	Numérico	Matrícula Inicial en Inglés Hombre 2015
rit_15	Numérico	Matrícula Inicial en Inglés Radio Interactiva Total 2015
rih_15	Numérico	Matrícula Inicial en Inglés Radio Interactiva Hombre 2015
frt_15	Numérico	Matrícula Inicial en Francés Total 2015
frh_15	Numérico	Matrícula Inicial en Francés Hombre 2015
itt_15	Numérico	Matrícula Inicial en Italiano Total 2015
ith_15	Numérico	Matrícula Inicial en Italiano Hombre 2015
extrant_15	Numérico	Matrícula Inicial Alumnos Extranjeros 2015
extranh_15	Numérico	Matrícula Inicial Alumnos Extranjeros Hombres 2015

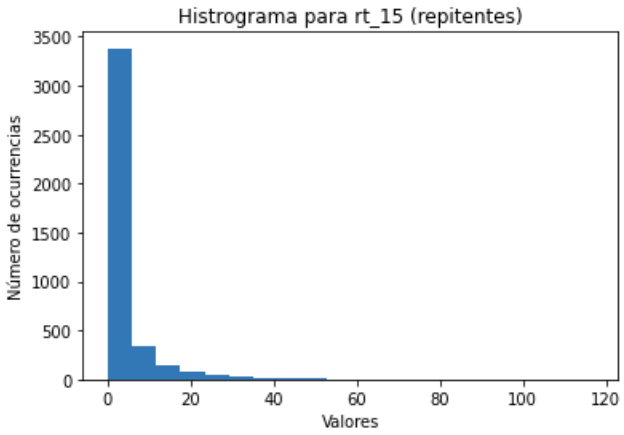
Cantidad de datos de la fuente: 4266 filas.

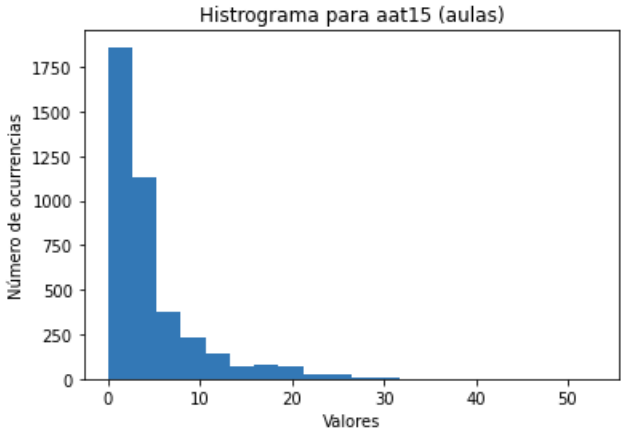
La fuente de datos fue tomada directamente de la página web del Programa Estado de la Nación (https://estadonacion.or.cr/base-datos/), específicamente la que tiene el nombre “Educación Primaria” (la cual corresponde a un archivo csv).

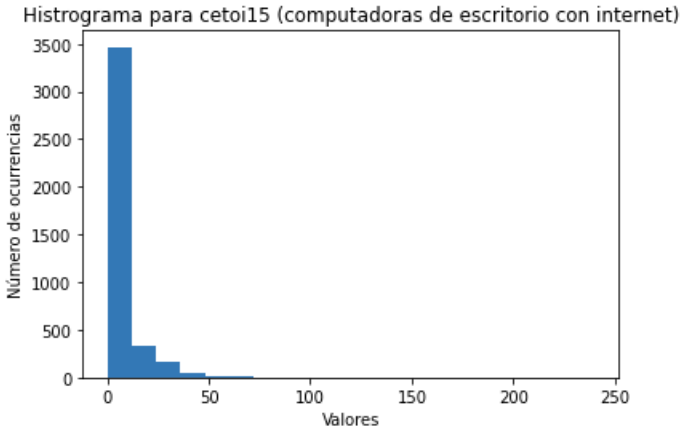
Histogramas

Los siguientes histogramas permiten entender mejor la distribución de algunas de las columnas:









Códigos geográficos

En la siguiente tabla se muestra el detalle de cada uno de los campos del set de datos de códigos geográficos de Costa Rica:

Columna	Tipo de dato	Descripción
CodigoProvincia	Numérico	Código de la provincia

Provincia	String	Nombre de la provincia
CodigoCanton	Numérico	Código del cantón
Canton	String	Nombre del cantón
CodigoDistrito	Numérico	Código del distrito
Distrito	String	Nombre del distrito

Cantidad de datos de la fuente: 474 filas.

La fuente de datos fue tomada directamente de la página web del INEC (https://www.inec.cr/sites/default/files/documentos/inec_institucional/metodologias/documentos_metodologicos/3_clasificacion_codigos_geograficos.pdf). Este archivo está en formato pdf pero se trasladó dicha información a un archivo csv.

Unión de ambos conjuntos de datos

Ambos conjuntos de datos (escuelas primarias y códigos geográficos) se unirán por el código de distrito, específicamente a través de la columna "cddis15" (que contiene el código de distrito) de la base de datos de escuelas primarias, con la columna "CodigoDistrito" del set de datos de códigos geográficos.

No se considera necesario unir ambos conjuntos de datos por el código de provincia y el código de cantón, dado que el código de distrito está conformado por ambos campos.

Objetivo predictivo

Dado que la recomendación es que el objetivo sea de predicción binaria, se agregará una nueva columna al set de datos, la cual indica si el centro educativo es "Dirección 2 o superior" (Sí o No); esta columna se calculará a partir de los datos de la columna "mit_15" (que contiene la matrícula inicial total para cada centro educativo), de manera que si la cantidad de estudiantes matriculados es mayor a 90 el valor de la columna es SI (es decir, sí es de tipo Dirección 2 o superior), y si la cantidad de estudiantes matriculados es menor o igual a 90 entonces el valor de la columna es NO.

Por lo tanto, el atributo que se utilizará como variable objetivo del modelo de aprendizaje automático, será esta nueva columna llamada "Direccion2_o_Superior".

Para poder determinar la clasificación de instituciones de educación primaria por tipo de dirección, se tomó la siguiente información del Programa Estado de la Nación, en donde se indica la cantidad de alumnos por cada tipo de dirección:

Instituciones en I y II ciclos por tipo de dirección
Unidocente (hasta 30 alumnos)
Dirección 1 (de 31 a 90 alumnos)
Dirección 2 (de 91 a 200 alumnos)
Dirección 3 (de 201 a 400 alumnos)
Dirección 4 (de 401 a 800 alumnos)
Dirección 5 (más de 800 alumnos)

Tomado de https://www.estadonacion.or.cr/educacion2017/assets/parte-1-capitulo-3.pdf

La predicción de este atributo se realizará basado en la información que se tiene en el dataset, referente en primera instancia, a la infraestructura de cada centro educativo (cantidad de aulas, laboratorios, comedor, biblioteca, gimnasio, sodas, computadoras, inodoros, lavatorios, etc.), además de información de aprobados, reprobados, repitentes, etc.

Investigación Preliminar – Segundo Problema

Fuentes de datos analizadas

Como parte de la investigación y el análisis de distintas fuentes de datos, se tuvo acceso, a un set de datos que contiene el inventario de carros de una compañía estadounidense llamada J.D.Power que vende carros en línea (https://www.jdpower.com/) para la cual mi esposo brinda servicios de TI. Cabe resaltar que, para la utilización de esta información se cuenta con el permiso del dueño de los datos, únicamente para los fines académicos de este proyecto.

Esta fuente de datos contiene información detallada de los carros que ellos venden (información de los distribuidores, características de los carros: marca, modelo, año, color, precio, millaje, etc).

Posteriormente se procedió a buscar otra fuente de datos que se pudiera unir a este conjunto de datos y enriquecer dicha información. A raíz de esto se procedió a buscar una fuente de datos que contenga los zip codes de Estados Unidos e información relevante acerca de estos, dado que el inventario de carros de J.D.Power contiene una columna con el zip code del distribuidor. Esta fuente de datos se logró obtener de una página web (https://simplemaps.com/data/us-zips).

Producto de este análisis se decidió escoger esas dos fuentes de datos (inventario de carros de J.D.Power y el conjunto de datos de los códigos postales de USA) para el desarrollo del proyecto.

El motivo por el cual se eligió la fuente de datos del inventario de carros de J.D.Power es porque se adapta a los requerimientos del proyecto en diferentes aspectos:

- Contiene gran cantidad de registros.
- Corresponden a datos de la vida real y cotidiana.
- Contiene columnas con información relevante que pueden ser utilizadas para el análisis predictivo.
- Permite cruzar los datos con otra fuente de datos

El motivo por el cual se escogió la fuente de datos de los códigos postales de USA es porque, considero que es de mucha relevancia poder unir ambas fuentes de datos para conocer información relacionada con la ubicación (nombre del estado, ciudad, etc) del distribuidor, además de otra información de importancia como por ejemplo la cantidad de habitantes para cada código postal, lo cual permitiría realizar un análisis más detallado de la información.

Descripción detallada de los datos

Inventario de carros

En la siguiente tabla se muestra el detalle de cada uno de los campos del set de datos con el inventario de los carros:

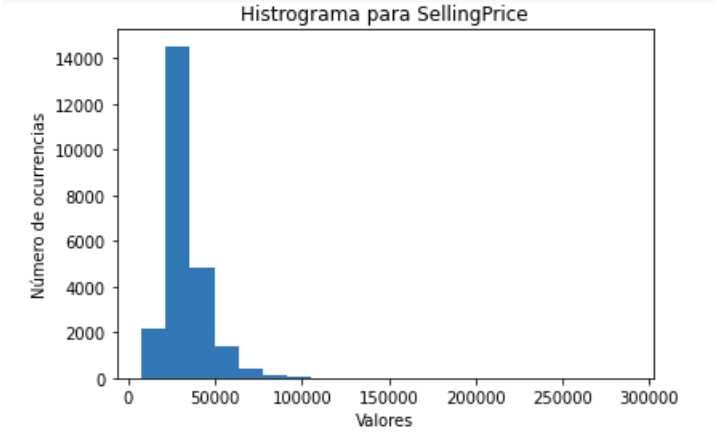
Columna	Tipo de dato	Descripción
ICCID	Numérico	Identificador del distribuidor
DealerName	String	Nombre de distribuidor (dealer)
DealerZip	Numérico	Código postal del dealer
VIN	String	Corresponde al número identificador del carro
Year	Numérico	Año del carro
Make	String	Marca del carro
Model	String	Modelo del carro
Trim	String	Versión del modelo
BodyStyle	String	Estilo del carro
BodyType	String	Tipo de carro
ModelCode	String	Código del modelo
MSRP	Numérico	Valor de venta sugerido por el fabricante
SellingPrice	Numérico	Precio de venta
Mileage	Numérico	Millaje del carro
ChromeStyleID	Numérico	Identificador de las características del vehículo
FuelType	String	Tipo de combustible
Exteriorcolor	String	Color exterior del carro
Exteriorcolorcode	String	Código del color exterior del carro
Interiorcolor	String	Color interior del carro
Interiorcolorcode	String	Código del color interior del carro
Interiormaterial	String	Material del interior del carro
DoorCount	Numérico	Cantidad de puertas
EngineDisplacement	String	Desplazamiento del motor
Cylinders	Numérico	Cantidad de cilindros
Engine	String	Indica características del tipo de motor
Drivetrain	String	Indica si el carro es 4x4, 4x2, etc
Transmission	String	Transmisión del vehículo
TransmissionSpeed	String	Cantidad de velocidades de la transmisión
CityMPG	Numérico	Millas por galón en cuidad
HwyMPG	Numérico	Millas por galón en autopista

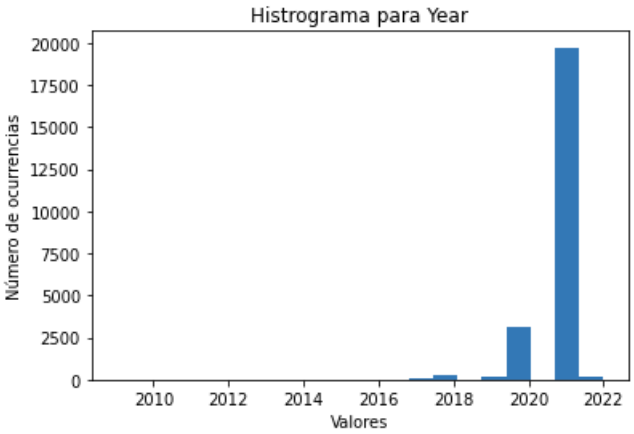
Cantidad de datos de la fuente: 23522 filas.

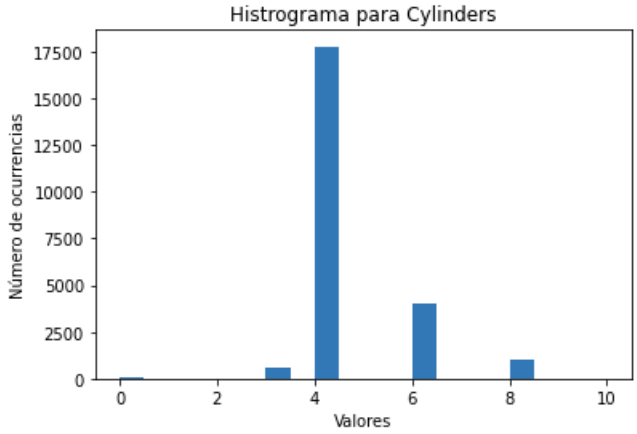
La fuente de datos fue tomada directamente de J.D.Power y corresponde a un archivo csv.

Histogramas

Los siguientes histogramas permiten entender mejor la distribución de algunas de las columnas:







Zip Codes de USA

En la siguiente tabla se muestra el detalle de cada uno de los campos del set de datos que contiene los códigos postales de USA:

Columna	Tipo de dato	Descripción
zip	Numérico	Código postal

city	String	Nombre de la ciudad
state_id	String	Identificador del estado
state_name	String	Nombre del estado
population	Numérico	Población (cantidad de habitantes)
timezone	String	Zona horaria

Cantidad de datos de la fuente: 33121 filas.

La fuente de datos fue tomada de la página web https://simplemaps.com/data/us-zips y corresponde a un archivo csv.

Unión de ambos conjuntos de datos

Ambos conjuntos de datos se unirán por el código zip, específicamente a través de la columna “DealerZip” (que contiene el código zip del dealer) de la fuente de datos del inventario de carros, con la columna “zip” de la fuente de datos de los zip codes de USA.

Objetivo predictivo

El atributo de los datos que se utilizará como variable objetivo del modelo de aprendizaje automático será la columna “SellingPrice”, con el fin de poder predecir el precio de venta de un vehículo a partir de una serie de características de este (año, precio de venta sugerido por el fabricante, millaje, cantidad de puertas, cantidad de cilindros, etc).