Seminario vacacional
“Big data, Analítica de datos y gestión de la información”

Laboratorio #3.

Estudiante:
Joseph Oswald Quiroz Mejia.

Código:
69794.

Profesor:
Elias Buitrago Bolivar.

Universidad:
ECCI.

Fecha:
01/07/2024.

Desarrollo del laboratorio.

Recolección de Datos:

Open in Colab

Scraping Used Car Web Data: Case Study tucarrro.com (Colab version)

Author: Elias Buitrago Bolivar

This jupyter notebook depicts a python based web scraping algorithm to obtain data to train a price car prediction machine learning algorithm. Used cars web data are extracted from Tu Carro. The code presented here is functional and was tested by scraping real data. This code version is compatible with Colab. *Updated: Jun 20, 2024*

Install required libraries

```
[ ] !pip install lxml
    !pip install scrapy
    !pip3 install requests-html
    !pip3 install selenium
```

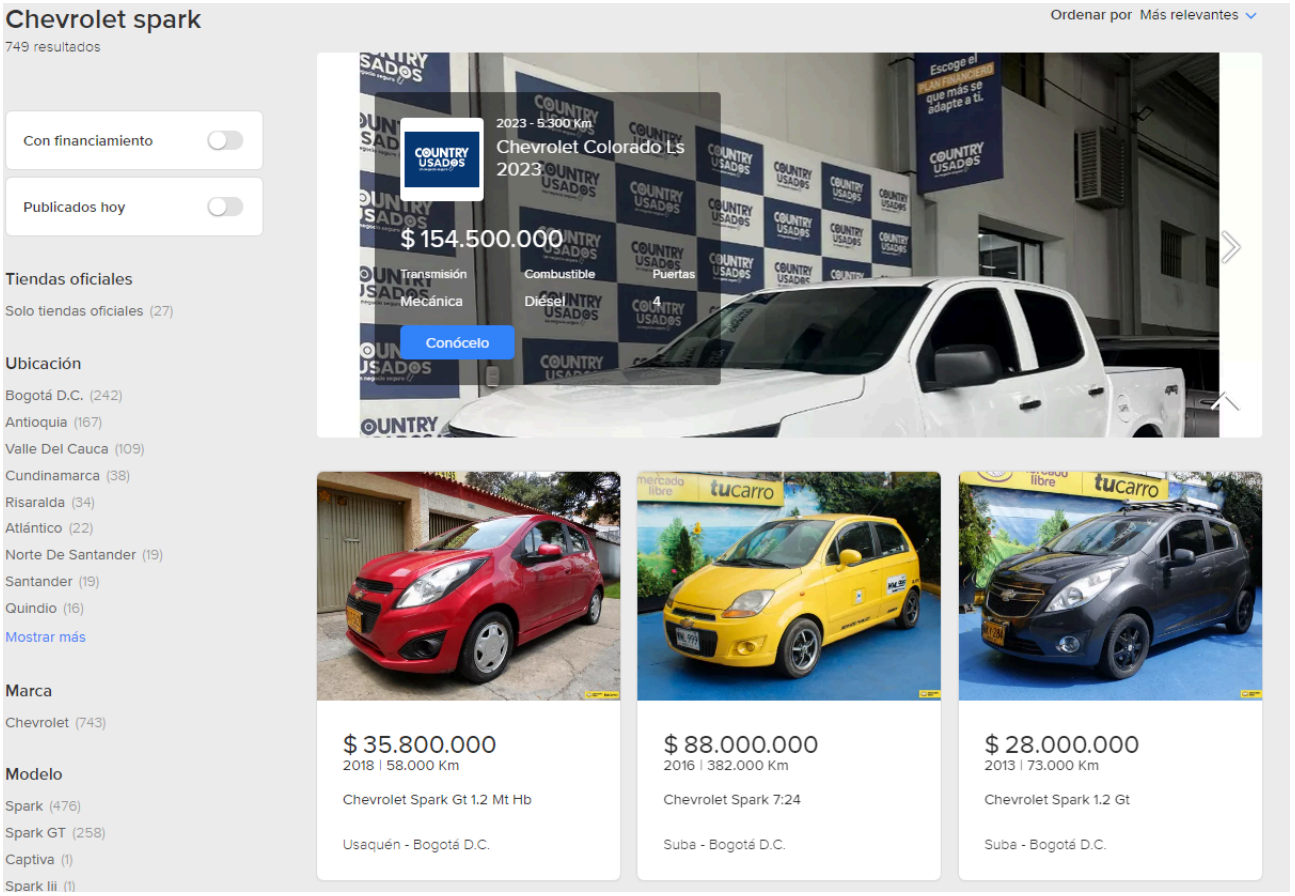
+ Code + Text

```
[ ] %%shell
    # Install chromedriver
    # Credits: https://medium.com/@MinatoNamikaze02/running-selenium-on-google-colab-a118d10ca5f8
    sudo apt -y update
    sudo apt install -y wget curl unzip
    wget http://archive.ubuntu.com/ubuntu/pool/main/libu/libu2f-host/libu2f-udev_1.1.4-1_all.deb
    dpkg -i libu2f-udev_1.1.4-1_all.deb
    wget https://dl.google.com/linux/direct/google-chrome-stable_current_amd64.deb
    dpkg -i google-chrome-stable_current_amd64.deb

    wget -N https://edgedl.me.gvt1.com/edgedl/chrome/chrome-for-testing/120.0.6099.62/linux64/chromedriver-linux64.zip -P /tmp/
    unzip -o /tmp/chromedriver-linux64.zip -d /tmp/
    chmod +x /tmp/chromedriver-linux64/chromedriver
    mv /tmp/chromedriver-linux64/chromedriver /usr/local/bin/chromedriver

    pip install selenium chromedriver_autoinstaller
```

```
Scrapping 47 / 48 ...
Este es el valor de p[i]: ['Chevrolet Spar Gt 2013 Contacto 3136629446', '$26.000.000', '2013', '120.000']
**********
WEB SCRAPING FROM SEARCH PAGE #8
Href obtained:  28
Scrapping 0 / 28 ...
Este es el valor de p[i]: ['Chevrolet Spark Gt', '$28.000.000', '2012', '159.000']
Scrapping 1 / 28 ...
Este es el valor de p[i]: ['Vendo Chevrolet Spart Gt Modelo 2016', '$32.000.000', '2016', '119.000']
Scrapping 2 / 28 ...
Este es el valor de p[i]: ['Chevrolet Spark Gt 2012', '$25.800.000', '2012', '176.000']
Scrapping 3 / 28 ...
Este es el valor de p[i]: ['Chevrolet Spark Gt Active La Más Full', '$41.000.000', '2020', '74.000']
Scrapping 4 / 28 ...
Este es el valor de p[i]: ['Chevrolet Spark Gt  2014', '$30.000.000', '2014', '108.000']
Scrapping 5 / 28 ...
Este es el valor de p[i]: ['Chevrolet Spark GT 1.2 Lt', '$42.000.000', '2019', '45.000']
```



Se realiza por medio del programa web scraping en el cual se vincula la url de la página tu carro y de donde se extraen los datos.

El carro seleccionado es el “Chevrolet Spark Gt”, los datos recolectados son las variables de, modelo, precio, año del modelo y kilometraje.

355	Chevrolet Spark Gt 2011 1.2 90.000 Km	$26.900.000	2011	92.000
356	Chevrolet Spark Gt Fe 1200	$30.000.000	2016	90.000
357	Chevrolet Spark GT 1.2 Gt M300 Ltz	$36.500.000	2019	56.300
358	Chevrolet Spark Gt Ltz	$33.500.000	2017	87.100
359	Chevrolet Spark Gt Ab Abs Lt Full Equipo 2017	$33.900.000	2017	36.500
360	Chevrolet Spark Gt	$29.900.000	2016	63.000
361	Chevrolet Spark Gt Ltz 2014	$23.900.000	2014	178.000
362	Chevrolet Spark Gt Único Dueño	$24.000.000	2012	96.000
363	Chevrolet Spark Gt Ltz 2015	$28.999.999	2015	89.000
364	Chevrolet Spark Gt 1200 Cc M/t Aa 2019	$33.000.000	2019	77.000
365	Chevrolet Spark 1.2 Gt M300	$20.000.000	2013	198.000

Se logran obtener un total de 365 datos en total a los cuales se les aplicará un filtro.

car_model	price	year_model	kms
Chevrolet Spark Gt Ltz Mt 1.2 Cc 2014	$28.900.000	2014	Publicado
	0	0	0
Chevrolet Spark Gt Ltz	$34.000.000	2015	90
Chevrolet Spark Gt Gt	$30.100.000	2014	115
Chevrolet Spark Gt Gt Ltz	$31.000.000	2015	999.999

Se realiza un filtro sobre los valores más lejanos o menos comunes, en total 5 datos son eliminados ya que poseen números muy alejados del promedio general

355	Chevrolet Spark Gt	$29.900.000	2016	63.000
356	Chevrolet Spark Gt Ltz 2014	$23.900.000	2014	178.000
357	Chevrolet Spark Gt Único Dueño	$24.000.000	2012	96.000
358	Chevrolet Spark Gt Ltz 2015	$28.999.999	2015	89.000
359	Chevrolet Spark Gt 1200 Cc M/t Aa 2019	$33.000.000	2019	77.000
360	Chevrolet Spark 1.2 Gt M300	$20.000.000	2013	198.000

para resolver las siguientes preguntas se tomaron 360 datos en total.

Preguntas (recolección de datos).

1. **¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?**

   Modelo: Ya que este nos ayuda a diferenciar entre los distintos tipos de Spark GT.

   Año: nos ayuda a determinar la depreciación y el valor basado en la antigüedad del vehículo.

   Kilometraje: Es la más importante porque nos ayuda a evaluar el desgaste y el uso del vehículo.

2. **¿Qué variables parecen irrelevantes y pueden ser excluidas?**

   Como nos pudimos dar cuenta en los filtros las variables más despreciadas pueden ser las que contengan letras en vez de número como en el caso del kilometraje que sale "publicado", variables muy alejadas de los valores más común, además la columna de precios considero que no es importante ya que se puede determinar gracias al modelo, el año del modelo y su kilometraje

3. **¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?**

   La tabla cuenta con alrededor de 24 tipos de vehículos, lo cual puede ser suficiente para obtener tendencias primerizas, en lo personal considero que no son suficientes datos para predicciones muy precisas.

4. **¿Hay demasiadas variables para el método de modelado de su elección?**

   No, según lo visto en clase el obtener cuatro variables es manejable y adecuado para la mayoría de los métodos de modelado vistos.

5. **¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?**

   No, ya que todos los datos son extraídos de una fuente que posee cada variable extraída de un mismo lugar, por ende la probabilidad de que se unan es muy baja.

6. **¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?**

   No, pero si considero que hay una forma de hacer este manejo de datos más eficiente y preciso.

Preguntas (Descripción de datos)

1. **¿Cuál es el formato de los datos?**

   Los datos al ser descargado del Google Colab se hallan en formato CSV, una vez cargados en la hoja de cálculo de google se hallan en algo similar a una hoja de cálculo.

2. **¿Cuál es el método utilizado para capturar los datos?**

   Se utilizó una programación llamada WEBscrapping.

3. **¿Qué tamaño tiene la base de datos (en número de filas y columnas)?**

   La base de datos tiene 360 filas (registros) y 4 columnas (variables) y 24 tipos de nombres de vehículos.

4. **¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?**

   Considero que las variables de modelo, precio, año y kilometraje son relevantes para analizar el mercado de los vehículos Chevrolet Spark GT ya que nos dan a grandes rasgos una idea de cómo está el carro en general.

5. **¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?**

   Simbólicos o cualitativos:: Modelo (palabras).

   cuantitativos o numéricos: Precio, Año, Kilometraje.(numeros)

6. **¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?**

   No, se pueden calcular en la hoja de cálculo de google pero no se ha realizado hasta el momento.

7. **¿Es capaz de priorizar las variables relevantes? Si no es así, ¿hay analistas de negocio disponibles para proporcionar más información?**

   Las variables relevantes que se pueden tomar en cuenta pueden ser precio, año y kilometraje, ya que estas nos pueden ayudar a analizar el mercado y la depreciación de los vehículos.

Exploración de los datos

1. ¿Qué tipo de hipótesis se ha formado sobre los datos?

Según los datos tomados se cree que los factores que más influyen en el precio de venta son el kilometraje y el año del vehículo, ya que existen registros del mismo vehículo con variaciones en estos aspectos que afectan directamente su valor.

2. ¿Qué variables parecen prometedoras para un análisis más profundo?

Sería interesante analizar variables adicionales como variaciones del modelo, si el vehículo tiene o no modificaciones, y si tiene turbo, entre otros.

3. ¿Sus exploraciones han revelado nuevas características sobre los datos?

No

4. ¿Cómo han cambiado estas exploraciones su hipótesis inicial?

No las han cambiado

5. ¿Considera que debería reformular el alcance del proyecto?

No

6. ¿Esta exploración ha alterado los objetivos?

No

7. ¿Puede identificar subconjuntos particulares de datos para su uso posterior?

No

Verificación de los datos

1. ¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?

Sí, se identificaron registros en los cuales en la variable "precio" se indicaba "publicado". Esto se debe a que el vendedor no describió el precio del vehículo en la etiqueta correspondiente, por lo que estos datos se eliminaron.

2. ¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?

Sí, al cargar el archivo *.csv directamente a Excel, todas las palabras que contenían tilde (como "Automático", "clásico", "mecánico") presentaban el carácter donde se situaba la tilde cambiado automáticamente por “??”. Fue necesario filtrar y reemplazar estas palabras para evitar conflictos posteriores.

3. ¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?

Sí, se consideró descartar los vehículos que en la variable kilometraje tienen un valor de cero, o un valor como 999.999 ya que su valor se compara con el de un concesionario o un carro nuevo.

4. ¿Cada registro contiene el mismo número de campos?

Sí, después de aplicar toda la metodología y tener finalmente los datos limpios, cada registro contiene la misma cantidad de campos.