Universitat Oberta de Catalunya
Master Universitario en Ciencia de Datos
Tipología y Ciclo de vida de los Datos

Dataset: Características de Vehículos nuevos en venta en Bogotá, Colombia, mediante WebScraping.

Es el resultado de la primera práctica de la asignatura Tipología y ciclo de vida de los datos, en la cual se hace uso de diferentes técnicas de scraping, todo bajo el lenguaje de programación Python, y su librería Scrapy. Así, entonces se obtiene un dataset con información de la página http://www.carroya.com, la cual brinda información de los vehículos en el mercado colombiano, donde se obtienen características como país de ensamblaje, precio, nombre, modelo, número de puertas, entre otros.

Carlos Alfonso Cuervo Rodríguez

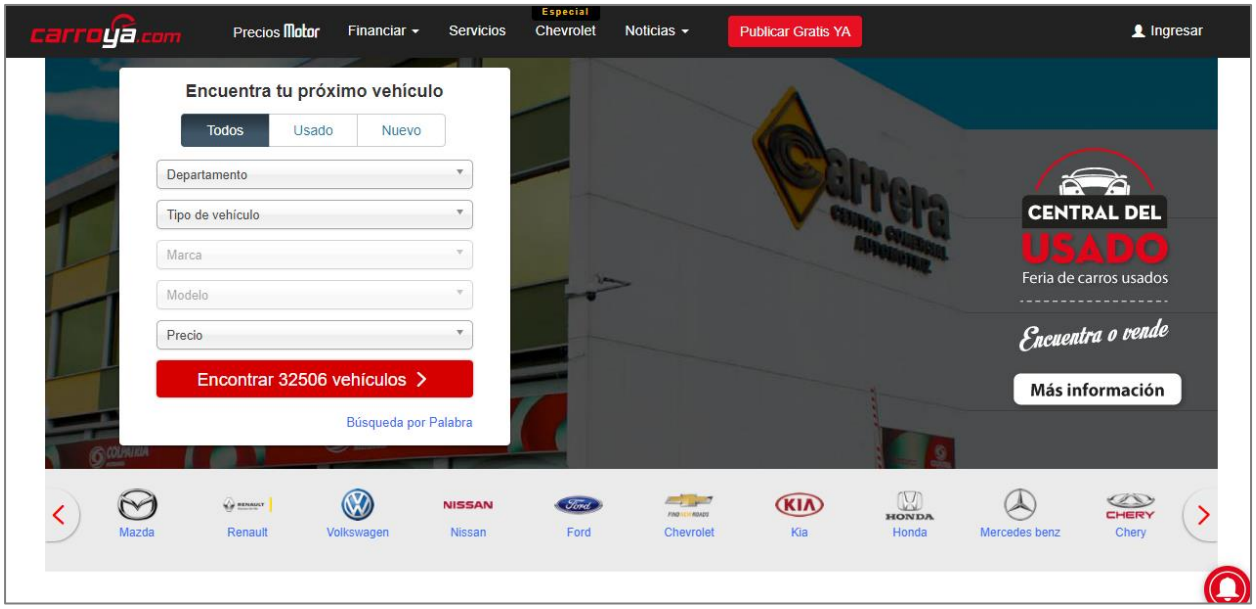

Fig1: Pantallazo de la página Web: carroya.com

El mercado de vehículos en Colombia:

La venta de vehículos nuevos en Colombia es un mercado bastante grande y maduro, donde es posible encontrar vehículos de casi todas las nacionalidades, que pueden ir desde la gama más baja y básica, hasta los carros más costosos y lujosos como son Mercedez, BMW, Audi, Jaguar, Porche entre muchos otros. Por este dinamismo, debido a estrategias comerciales, los precios de los vehículos pueden cambiar de un momento a otro, por lo cual es de suma importancia para el consumidor y para la competencia estar al tanto de los últimos precios de todos los vehículos que se ofrecen en su ciudad.

Contenido:

El dataset contiene las principales características que un vehículo puede tener, y que pueden ser importantes para un comprador al momento de querer adquirir un vehículo nuevo, entre ellas:

- Nombre: Nombre comercial del vehículo. Es el primer identificador del mismo.
- Referencia: Como se sabe puede haber diferentes modelos con el mismo nombre, por lo cual se hace importante saber la referencia del mismo.
- Precio: Precio en pesos de cada uno de los vehículos.
- No. Puertas: Dependiendo si es un sedan, una camioneta, una pick-up, coupe, el vehículo puede tener diferente número de puertas, lo cual también puede influir en su precio.
- Tipo de caja: La caja puede ser mecánica, automática, secuencial, entre otras.
- Lugar de ensamble: Este campo es bastante importante para muchos consumidores, ya que le atribuyen gran parte de la calidad del vehículo al lugar donde este fue ensamblado. Siempre son muy apreciados los vehículos ensamblados en Europa o EE. UU.
- Tipo Dirección: La dirección puede ser mecánica, asistida, hidráulica, entre otros tipos.
- Combustible: Finalmente el tipo de combustible, para identificar si es a base de gasolina o diesel en algunos casos de carros más grandes.

Estos datos son obtenidos de la página web: www.carroya.com , mediante la técnica de web scraping. Es de anotar que la cantidad de referencias que se pueden obtener en este sitio son miles, por lo cual se ha reducido la búsqueda a vehículos únicamente de la ciudad de Bogotá y que se encuentren nuevos, ya que es posible encontrar vehículos en todo el país y en cualquier estado, sea nuevo o usado.
Esta data es una "foto" del día en que se ha corrido el procedimiento, por lo cual no guarda información histórica. Si se deseara ver la evolución de los precios de cierta referencia, sería necesario cada día correr el procedimiento e ir almacenando la información en una BBDD. Además, la página tampoco guarda información histórica por lo que esta sería la única manera.

Motivación:

Según el contexto y el contenido del dataset, ya es posible darse una idea de la utilidad que puede tener el presente ejercicio, ya que por ejemplo un concesionario lo puede utilizar para monitorear el precio de los vehículos nuevos en la ciudad de Bogotá y así establecer sus estrategias de venta de cada segmento objetivo, o por ejemplo ofrecer descuentos de alguna de las referencias.

De esta misma manera un ciudadano del común puede usarlo para monitorear el precio de los vehículos dentro de un lapso determinado para decidir cuál vehículo comprar dadas unas características de su interés.

Por otra parte, esta data junto con otra podría ser usada para realizar data mining, y por ejemplo determinar si hay relación existente entre el precio de los vehículos y el precio del dólar o del acero, dependiendo del lugar de ensamble y si este comportamiento es estacionario o no. Así entre muchos otros estudios podrían realizarse gracias a este dataset.

Licenciamiento:

El licenciamiento adecuado para este desarrollo, se considera que es el **CC BY-SA 4.0 License** (Atrribution -ShareAlike 4.0 International), ya que permite la libre circulación del contenido en cualquier medio, con diferentes propósitos incluso comercial (que claramente se indicó en la motivación), pero que además se debe dar crédito al dueño de los datos, y por último permite seguir con desarrollos futuros, ya que si se cambia debe distribuirse bajo la misma licencia[1].

Agradecimientos:

Todos los datos obtenidos en esta práctica son provenientes del dominio carroya.com. Los cuales fueron accedidos por medio de webscraping, haciendo uso de la librería Scrapy de Python.

Referencias:

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 8. Scrapy.
- Documentación de Scrapy, recuperada del sitio web: https://doc.scrapy.org/en/latest/

[1] Información obtenida de la página Web: https://creativecommons.org/licenses/by-sa/4.0/