LABORATORIO 3 LABORATORIO EN CLASE COMPRENSIÓN DE LOS DATOS

Andrés Felipe Esquivel Ruiz; Código 12967.

Nathaly Daniela Mejia Meléndez; Código 84588.

Universidad ECCI

Seminario Big Data

Elías Buitrago Bolívar

julio de 2024

Recolección de datos

La recolección de datos inicio en la pagina web de Tu Carro la cual gracias al código dado por el docente en el entorno de Colab nos permitirá visualizar los datos de las marcas de los carros que se encuentran vendiendo en esta página, para esta actividad hemos decidido analizar los datos de Marca Renault Logan la cual cuenta con 894 resultados:

```
car_brand = 'renault'  # Brand car name. Ej: chevrolet, renault, kia.
car_model = 'logan'  # Model car name. Ej: duster, onix, rio.
data = scrapebyPages(car_brand,car_model,1,13)
# scrapebyPages(1,2)
```

Descripción de los datos

En este database encontramos 574 filas y 6 columnas, de estas los datos que más se destacan son el precio, el modelo y el kilometraje del carro, aunque al analizar los datos estas 6 son bastantes relevantes para hacer la adquisición de un vehículo. El formato descargado del código es un CSV, los datos son alfanuméricos y símbolos, ya que contienen precios y kilometrajes de los vehículos gracias a esta información se puede calcular datos como cuantos modelos hay, que kilometraje y a qué precio esta cada uno según sus variables.

Exploración de los datos

- Según lo analizado en el database podemos pensar que estos datos si podrían ser suficientes para la adquisición de los vehículos, ya que se puede realizar comparaciones, en un análisis preliminar podemos observar que hay más 574 modelos de Renault Logan pero hay que tener en cuenta que como es una página donde las personas editan estos datos pueden variar por comas, puntos, números y demás símbolos que pueden alterar el mismo modelo.
- El Modelo de carro Renault Logan en venta es desde el año 2000 hasta el 2024, el modelo que se encuentra más en venta es el modelo 2023 con 84 publicaciones. El rango de precios del Renault Logan es de 13.000.000 hasta \$72.900.000.
- Adicionalmente se encontró 3 tipos de combustible Gasolina, Gasolina y gas e hibrido.

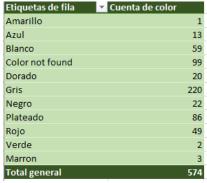
Verificar la calidad de los datos

- Se identifico 3 filas que tenían valor 0 las cuales luego de una revisión se procedió a eliminarlas
- En la identificar errores tipográficos en los datos encontramos los que se muestran en la siguiente imagen, son errores que podrían ser que el formato no lee las tildes y las transcribe de esta manera, adicional se evidencia uso de símbolos para una marca que no tiene en su nombre el símbolo

Mendant Logari 110 Line				
Renault Logan 1.6 life + automatico				
Renault Logan 1.6 Life Automatico				
Renault Logan 1.6 Life MecÃinico				
Renault Logan 1.6 Life MecÃinico				
Renault Logan 1.6 life Mecanico				
car_model 🚚	price year_model	kms 🔻	color 🍱	fueltype
Renault Logan 1.6 Authentique	31000000 20	17 90	Marrón	Gasolina
Renault Logan 1.6 Expression	32000000 20	16 80	Marrón	Gasolina
Renault Logan 1.6 Expression Fii	24900000 20	12 115	Marrón	Gasolina



Al finalizar realice una tabla dinámica donde identifique que este era un error común y que 99 de los datos no tenían el color del vehículo.



Respondiendo a las preguntas sugeridas por IBM

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

Si, como lo comenté en la parte de verificar los datos había muchas palabras con errores porque tienen tildes, adicionalmente las publicaciones tienen errores ortográficos por lo cual al momento que se evidenciaron se realizó la modificación correspondiente sin tildes para seguir teniendo un archivo plano.

¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente

Se realizo un análisis por medio de tablas dinámicas en Excel ya que permite hacer una visualización más fácil de los datos y con la cual pudimos evidenciar los errores.

¿Cada registro contiene el mismo número de campos?

No dado que en el análisis que se realizó se evidencio que el color no siempre lo trae la publicación si no el vendedor coloca las otras características