DSC Group 4

Milestone 1 – Dataset

1/26/22

Title of dataset and original goal of its collection: Title of dataset is called Used Cars Dataset. The original goal of this collection is to help bring together all of Craigslist's posts for used cars into one dataset. Author turned this dataset from a school project and expanded its reach to where every few months his program will scrape Craigslist and update the dataset for every new used vehicle entry in the United States.

Source of dataset:
https://www.kaggle.com/austinreese/craigslist-carstrucks-data?select=vehicles.csv

Number of rows in your dataset: 426,881 rows

Complete list of all variables:

Continuous numeric count: 10

Other variable types: 17

id (Continuous numerical): A specific unique id number sequence associated with the specific post car each user has put up.

url (Categorical): URL to the specific car link posting. Not clear yet what this may help later in the project but included in the dataset.

region (Categorical): This is the region from where in Craigslist this the post is coming from.

region_url (Categorical): URL to the specific region the car belongs to. Not clear yet what this may help later in the project but included in the dataset.

price (Continuous numerical): Price for the used car listed on Craigslist.

year (Continuous numerical): Year the car was manufactured in.

manufacturer (Categorical): What company made the car in the posting.

model (Categorical): Model of car in the post.

condition (Ordinal): Has different types of conditions the car is in ex. (good, excellent, fair, like new, etc.)

cylinders (Contiguous numerical): Number of cylinders the used car has in it

fuel (Ordinal): The type of fuel the used car runs on ex. (gas, diesel, other)

odometer (Continuous numerical): Number of miles the car has at the moment of when the car was posted

title_status (Ordinal): Vehicle history of the car when it was posted to check how insurance sees the car ex. (clean, rebuilt, lien, salvage, etc.)

transmission: (Ordinal): Type of transition the car runs on ex. (automatic, manual, other)

VIN (Continuous numerical): Vehicle Identification Number unique code of numbers and letters per each car that has where the car was made, model year, type of vehicle, etc.

drive (Categorical): Amount of wheel drive the car runs on

size (Ordinal): How big the vehicle is on Craigslist ex. (mid-size, full-size, compact, etc.)

type (Ordinal): Type of car from the post ex. (pickup, truck. Other, coupe, mini-van, etc.)

paint_color (Categorical): Color of the specific used car posted

image_url (Categorical): Image URL to the specific car posting. Not clear yet what this may help later in the project but included in the dataset.

description (Categorical): Short descriptions (a sentence or two max) users posted about their car.

county (Categorical): Existing column in the dataset but not filled in the dataset for any of the rows. This is version 10 of the dataset. Maybe in previous versions it may be filled in and would have to be checked later.

state (Categorical): The state where the car post came from and located in.

lat (Continuous numerical): Latitude coordinates from where the car was posted from.

long (Continuous numerical): Longitude coordinates from where the car was posted from.

posting_date (Continuous numerical): Date (YYYY-MM-DD) and time the car post was posted at.

Any other relevant, interesting, or potentially useful information you have about the dataset: None at the moment. Have not been able to find any yet.