Milestone 1

Car Sales Data Preparation and Visualization

In this project, the datasets of interest are the car datasets which are all available at kaggle.com. Precisely three different datasets have been used in which the relationship between them exists by the car type, year, fuel type columns. In other words, the dataset can be joined together by the standard columns and explored together. Due to the varying characteristics of the car's year, fuel type is more consistent and did not have much variation and will be used to join the datasets.

In subsequent milestones of the project, I would be leveraging the option to pull the data from Kaggle API/ carroya.com and apply data wrangling techniques that I have learned throughout the course. And as a part of data visualization, I would be using matplotlib and ggplot2.

Datasets & Sources:

Data Source 1: Flat file data source, Kaggle
https://www.kaggle.com/hellbuoy/car-price-prediction

Data Source 2: Data Pull from API, Kaggle
API download from avikasliwal/used-cars-price-prediction

Data Source 3: Website Data, Carroya
https://www.carroya.com

Relationship between Datasets:

The key elements or grain of the dataset are identified as Car Make Year, Model and Fuel Type.

Interpretation of Data and next steps for upcoming milestones :

Dataset1:

The first dataset was car price which is available at https://www.kaggle.com/hellbuoy/car-price-prediction and had 26 variables. The variables in the dataset were both continuous and categorical.

Data Dictionary:

	Variable	Description
1	Car_ID	Unique id of each observation (Integer)

2	Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
3	carCompany	Name of car company (Categorical)
4	fueltype	Car fuel type i.e gas or diesel (Categorical)
5	aspiration	Aspiration used in a car (Categorical)
6	doornumber	Number of doors in a car (Categorical)
7	carbody	body of Car (Categorical)
8	drivewheel	type of drive wheel (Categorical)
9	enginelocation	Location of car engine (Categorical)
10	wheelbase	Wheelbase of Car (Numeric)
11	carlength	Length of Car (Numeric)
12	carwidth	Width of Car (Numeric)
13	carheight	height of Car (Numeric)
14	curbweight	The weight of a car without occupants or baggage. (Numeric)
15	enginetype	Type of engine. (Categorical)
16	cylindernumber	cylinder placed in the Car (Categorical)
17	enginesize	Size of Car (Numeric)
18	fuelsystem	Fuel system of Car (Categorical)
19	boreratio	Boreratio of car (Numeric)
20	stroke	Stroke or volume inside the engine (Numeric)
21	compressionratio	compression ratio of Car (Numeric)
22	horsepower	Horsepower (Numeric)
23	peakrpm	car peak rpm (Numeric)
24	citympg	Mileage in city (Numeric)
25	highwaympg	Mileage on highway (Numeric)
26	price(Dependent variable)	Price of Car (Numeric)

Dataset2:

The second dataset was US car price which is available at https://www.kaggle.com/avikasliwal/used-cars-price-prediction and had 13 columns. Similarly, the dataset had both numerical and categorical variables.

Variable	Description
Name	Car Make
Location	Car Location
Year	Car Make Year
Kilometers_Driven	Total Mileage
Fuel_Type	Categorizing Car w.r.t its fuel type
Transmission	Auto or manual
Owner_Type	to identify Car is pre-owned or new
Mileage	Average Mileage
Engine	Engine Capacity
Power	Horse power
Seats	Number of seats
New_Price	Car Price

Dataset3: HTML source from carroya.com website

URL: https://www.carroya.com/buscar/vehiculos/t4e0.do#paginaActual=4

I would be scraping the url to pull the used car details

Variable	Description
Name	Car Brand
Year	Make Year
Mileage	Total Mileage
Price	Current car value

To Summarize, I would like to clean the data as needed by dropping off columns which may not be required in further data analysis/visualizations, adding new columns/updating the data elements in selected columns to maintain consistent relationships between various data sources using the country code and then create the required visualizations.