UDACITY DATA SCIENTIST CAPSTONE PROJECT SUBMISSION

Analyzing and modelling 2022 Fuel Consumption Ratings

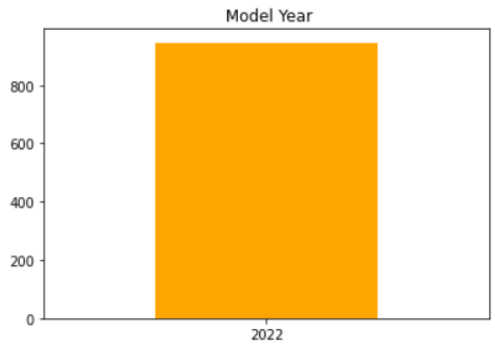
Submitted by: Prajita Chowdhury

Section 1: Project Definition

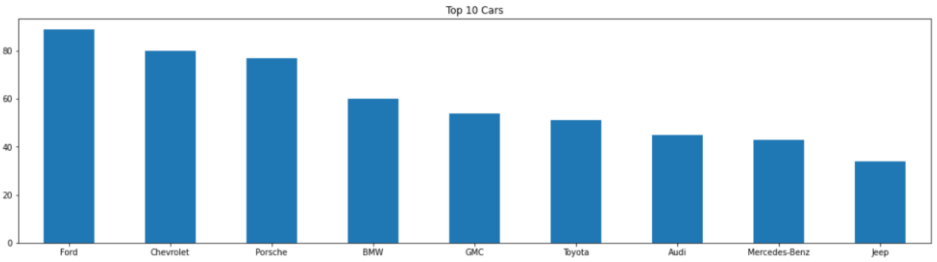
- Project Overview:
  - This project involves analysis of model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada in 2022.
  - Fuel economy is one of the critical metrics of consideration for vehicle buyers and environmental policy makers alike. When we burn less gas, we cut global warming emissions and produce less pollution, while spending less on gas—a lot less.
  - The measure of a vehicle's energy efficiency, given as a ratio of distance traveled per unit of fuel consumed. It is dependent on several factors like engine efficiency, transmission design (manual/automatic) and tire design.
  - In most countries, fuel economy is stated as "fuel consumption" in liters per 100 kilometers (L/100 km) or kilometers per liter (km/L or kmpl). In a number of countries still using other systems, fuel economy is expressed in miles per gallon (mpg), for example in the US and usually also in the UK (imperial gallon).
  - Data Source: https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings
  - Github Link for Python Notebook: Link
- Problem Statement
  - To build a regression model to calculate the fuel consumption of any car's given parameters.
- Metrics:
  - Root Mean Squared Error to identify the accuracy of the model.

Section 2: Analysis

- Data Exploration:
  - The dataset has 946 records have 15 features.
  - **The features are explained as follows**:
    - Model Year: Vehicle Model Year
    - Make: Vehicle Make
    - Model: Vehicle Model Type
    - Vehicle Class: Car Vehicle Class
    - Engine Size(L): Engine Size in Liters
    - Cylinders: No. Of cylinders
    - Transmission: Type of Transmission available in the vehicle (
    - Fuel Type: Fuel Type supported in the vehicle
    - Fuel Consumption (City (L/100 km): City fuel consumption ratings are shown in liters per 100 kilometers (L/100 km)
    - Fuel Consumption(Hwy (L/100 km)): Highway fuel consumption ratings are shown in liters per 100 kilometers (L/100 km)
  - **Data Distributions/ Visualizations**:
    - Model Year:

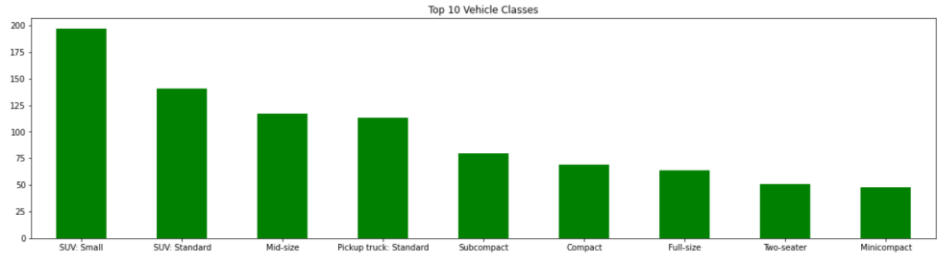


- Make :

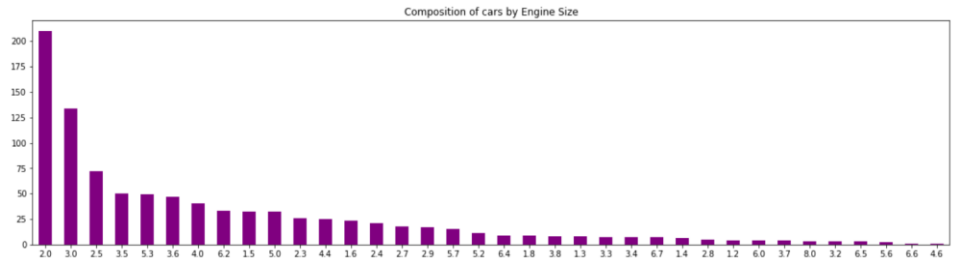


- Model:

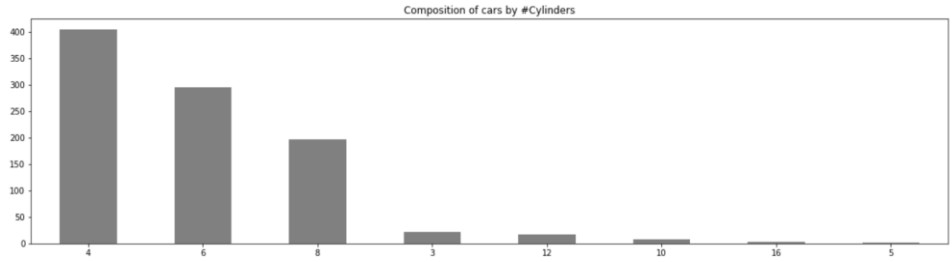
- Vehicle Class:



- Engine Size(L):

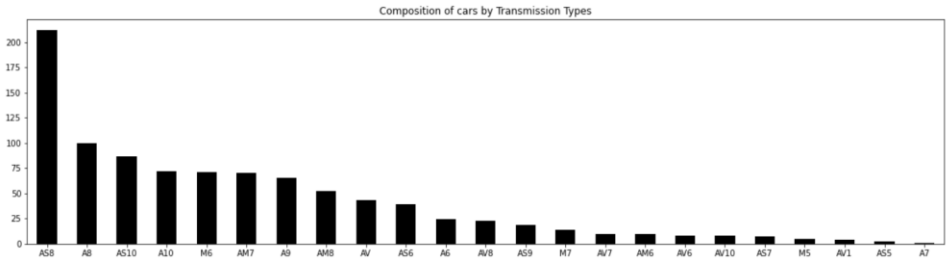


- Cylinders:

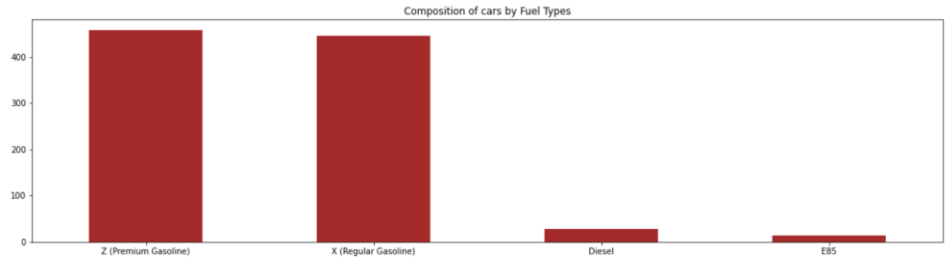


- Transmission Type:

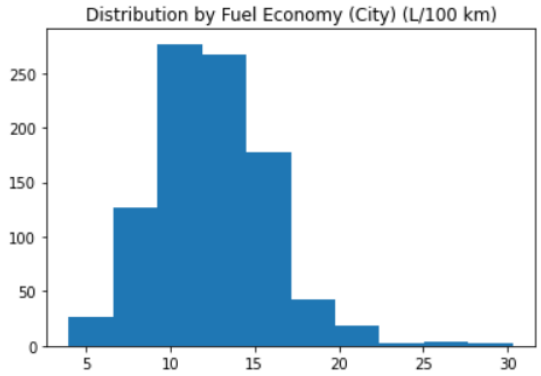
Transmission: A = automatic; AM = automated manual; AS = automatic with select shift; AV = continuously variable; M = manual; 3 – 10 = Number of gears.



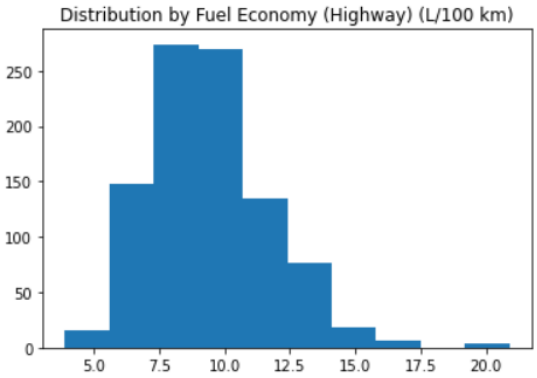
- Fuel Type:



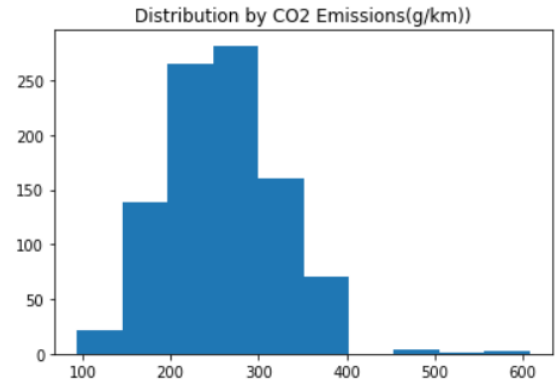
- Fuel Consumption (City (L/100 km):



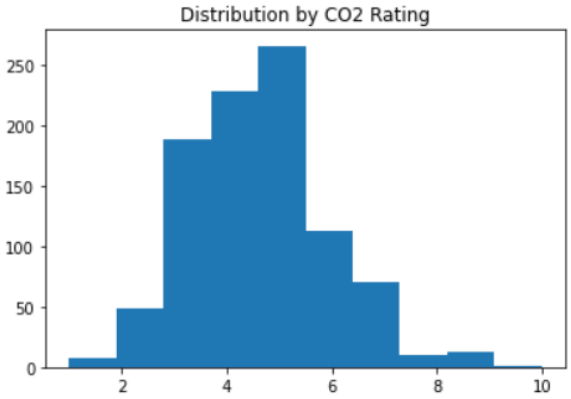
- Fuel Consumption(Hwy (L/100 km)



- CO2 Emission (g/km)



- CO2 Rating:



Top 5 cars with Highest Average Fuel Economy

Make	Average Fuel Economy (L/100 km)
Bugatti	29.133333
Lamborghini	21.412500
Rolls-Royce	19.985714
Bentley	17.437500
Dodge	15.766667

Bottom 5 cars with lowest Average Fuel Economy

Make	Average Fuel Economy (L/100 km)
Hyundai	8.813793
Kia	8.952000
Honda	9.008696
Mitsubishi	9.014286
Toyota	9.366667

Section 3: Methodology

- Data Preprocessing:
  - Checking for Missing Data: Dataset has no null values

```
print(f"Total no. of null values in dataset : {df.isnull().sum().sum()}")
```

```
Total no. of null values in dataset : 0
```

  - Checking for Data Duplication: Dataset has no duplicate rows

```
print(f'No. of duplicated rows = {df.duplicated().sum()}')
```

```
No. of duplicated rows = 0
```

- Implementation:
  - Selection of Feature Variable to train the model: Following features are selected for modelling:
    - Engine Size(L)
    - Transmission
    - CO2 Emissions(g/km)
    - Smog Rating
  - 'Transmission' is a categorical variable, hence OrdinalEncoder() used to transform it into numerical values.

```
cat_trans = df['Transmission']
encoder = OrdinalEncoder()
cat_trans = enc.fit_transform(cat_trans)
df['Transmission'] = pd.DataFrame(cat_trans)
```

  - Creating a new feature data frame X to store the above-mentioned features.
  - Selection of Target Variable (Y): 'Fuel Consumption (Comb (L/100 km))'
  - Splitting the dataset into training and test sub-sets
  - Building the Linear Regression Model
  - Using the Model to fit the training data
  - Using the Model to predict the output of test input data

```
X = df[['Engine Size(L)', 'Transmission', 'CO2 Emissions(g/km)', 'Smog Rating']]
y = df['Fuel Consumption(Comb (L/100 km))']
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)

reg = linear_model.LinearRegression()
reg.fit (X_train, y_train)
y_pred=reg.predict(X_test)
```

  - Calculating the Mean Squared Error and R2 Score for the model

```
# The mean squared error
print("Mean squared error: %.2f" % mean_squared_error(y_test, y_pred))
# The coefficient of determination: 1 is perfect prediction
print("Coefficient of determination: %.2f" % r2_score(y_test, y_pred))
```

  - Identifying the Coefficient and Intercept Values to define the modelling equation.

```
i=0
coeffs = []
intercept = round(reg.intercept_,2)

for x in reg.coef_:
    #print('Coeff.# ',i+1,'=',round(x,2))
    coeffs.append(round(reg.coef_[i],2))
    i+=1
print(f"Optimum Parameters: {coeffs}")
print(f"Intercept = {intercept}")
print(f"Linear Regression Model:\n h(x) = {intercept} + ({coeffs[0]})*Engine Size(L) + ({coeffs[1]})*Transmission + ({coeffs[2]})*CO2 Emissions(g/km) +
({coeffs[1]})*Smog Rating")
```

Section 4: Results

- Model Evaluation and Validation: discuss the models and parameters used in the methodology. If no model is used, students can discuss the methodology using data visualizations and other means.
- **Following are the parameters of the linear regression model developed**:

Parameter	Value
Intercept	-0.43
Parameter_1 (Engine Size(L))	0.12
Parameter_2 (Transmission)	0.01
Parameter_3 (CO2 Emissions(g/km))	0.04
Parameter_4 (Smog Rating)	0.03

- **Mean Squared Error and R2 Score:**

- Mean Squared Error = 0.35
- R2 Score (Coefficient of determination) = 0.96

Section 5: Conclusion

- We selected 4 features of a car to build a linear regression model for establishing a relation between engine size, transmission type, CO2 Emisison and Smog rating with the fuel consumption metric of the car. Out of the four, engine size carries the highest weight in determine the fuel consumption rate, but in no way we can conclude that it is an important determining factor in the same.

```
--------------------
Optimum Parameters: [0.12, 0.01, 0.04, 0.03]
Intercept = -0.43
Linear Regression Model:
 h(x) = -0.43 + (0.12)*Engine Size(L) + (0.01)*Transmission + (0.04)*CO2 Emissions(g/km) + (0.01)*Smog Rating
--------------------
```

- Improvement: The data contains categorical features as well. The model could be fine-tuned more by hot-encoding the categorical features and including them to build a model.