**Analyzing fuel economy based on automobile features in vehicles from 1984 to 2020**

Group 12

Matt McDonnell, Ji Ma, Zhongwei Chen

**Objective**

According to the Huffington Post, an average American spends over $2,000 a year on gas for their vehicle; this is a considerable amount for many Americans considering there are other costs associated with driving, such as insurance, registration, and tax. Through the work done in this project we will aim to look into the relationships between a plethora of variables that we believe to be responsible for the annualized fuel cost and/or potential CO2 tailpipe emissions of certain vehicles. The goal is to ultimately help two stakeholders who could benefit from this analysis: customers looking to buy vehicles and manufacturers looking to produce more efficient vehicles. This could help customers best match preferences in a vehicle to an optimal fuel cost for their living budget. On the other hand, while it is unlikely that diesel gas powered vehicles will become obsolete anytime in the next few decades, the emphasis for manufacturers to create engines that reduce the carbon footprint is more evident today than ever before. With states like California and Oregon implementing LCFS credits to regulate and incentivize this issue in refineries, it would be extremely beneficial if manufacturers could figure out how to best create cars that are desirable yet lower in fuel cost or tailpipe emission. This could keep consumers happy, while significantly lowering the consumption of gasoline and thus the demand for gas. It would position a certain manufacturer well ahead of the curve in this regard. The EPA even warns that transportation contributes to about 28% of all greenhouse gas emissions. Nonetheless, it is crucial to emphasis that this project will primarily aim to enhance the efficiency and value in the selection process for consumers who care about both cost and carbon footprint. Fortunately, the data set also has a time series component that will allow us to explore changes over time in fuel cost and emissions in vehicles.

**Data Set Description**

* Overview/Description: As an overview, the data set contains a variety of variables involved in measuring the efficiency of a vehicle in regard to fuel consumption and emissions. There are many columns that are less relevant in exploring this and we plan to remove or subset in order to mitigate for this. The variables we are trying to predict are fuel cost and tailpipe emissions. Some variables may point to explaining these in laments terms, but we want to provide the most statistically significant and accurate overview possible to allow for feature pairing and creation in aligning consumers and manufacturers respectively to get ahead of the current trend in whatever feature combination is most suitable.
* Rows: 42,377
* Columns: 83
* Sample Predictors: fuel cost, CO2 tailpipe emission, save/spend
* Web link: <https://www.fueleconomy.gov/feg/ws/index.shtml>
* Anything Interesting: time series implication, displacement, electric features in gas powered vehicles

A close up of a screen

Description automatically generated**Preliminary Data Exploration**

Since the dataset includes fuel economy information on both gasoline and electronic vehicles, decided to eliminate those dealing with electronic vehicles from our dataset because they are not relevant to the problem we are looking into. From the pair plot above, the positive linear relationships seen in the tailpipe emission(co2TailpipeGpm), fuel cost, and annual petroleum consumption (barrels08) variables suggest vehicles with large CO2 emissions are more likely to consume more fuel; therefore, vehicle owners are more likely to spend more money on fuel. Moreover, the number of cylinders and engine displacement in liters (displ) appear to have the highest potential to affect fuel cost as well as the cost on vehicle in 5 years (youSaveSpend). Consequentially, we want to use these implications to approach the goal of helping consumers find the most suitable vehicles and manufacturers help create designs that best meet these consumers’ needs.

**Predictions**

* Save/Spend Over 5 years (youSaveSpend)
  + Predict the amount of cost/saving that will possibly generate based on vehicle features that consumers prefer. For instance, if a consumer would prefer an Automatic 3-spd vehicle with 4-door, 4 cylinders, and displacement larger than 3.0, then he will most likely to spend extra 8500 dollars over 5 years.
* Fuel Cost (fuelCost08)
  + Predict annual fuel cost based on vehicle model, manufacture, displacement and other factors. Since we know there is potential relationship between fuel cost and displacement, annual petroleum consumption and cylinders, we want to predict fuel cost based on these factors so that we can give consumers a general prediction when they are uncertain about the vehicle they want to purchase.
* Tailpipe Emission (co2TailpipeGpm)
  + Predict CO2 tailpipe emission based on displacement, annual petroleum consumption and other factors such as manufacturers. This predication targets on consumers who are environmentally aware and want to take tailpipe emission into consideration when purchasing a vehicle.

Overall, we are creating a predictive model as a guideline for a customer who is looking for a vehicle. If a customer has preferences on certain features of his vehicle, we can predict the annual fuel cost and emission by using this model.

**Inference**

* The rate at which fuel cost and emissions decline over time
* The effect of displacement on fuel cost and emissions
  + In order to predict the fuel cost, we will use the displacement, cylinders numbers, gear box type, driving type, etc. Element as independent variables, to see what element will have most impact on fuel cost and emissions.
* Is the size of the car dependent on the predictors?
  + 2 doors vs. 4 doors cars explain size or sport?
* Model car that is known to be the most fuel efficient and cost effective
  + Has this been consistent over the year?
* Manufacturers that have the most room to catch up and improve fuel efficiency
  + Potential area to profit (2 doors, 6 cylinders, guzzler, etc.)

**Non Spark Packages**

Numpy: numerical calculations and np array creation.

Pandas: pandas data frame; change data type; detect NAs, null values, merge two data frames.

Matplotlib: basic plotting system for visualization

Seaborn: additional tool for the visualization, if the graphs require multiple elements that cannot execute on matplotlib.