Correlation between Model Year and it Fuel Efficiency (MPG)

Aden Letchworth, Tanay Shah, and Malik Nasla

***Abstract*—The automobile industry is constantly evolving with technology, one of the most important advancements is fuel efficiency. This is tied in with a combined effort to curb the use of fossil fuels and reduce the total amount of emissions. This study will examine the correlation between MPG (fuel efficiency) and Model Year. We are using a data set of 398 automobiles including 9 features. This data set was analyzed using several methods such as Simple Linear Regression (1) and Multiple Linear Regression (2). The results of (1) indicated a slight positive correlation between the two, indicating that a higher MPG is slightly associated with a higher Model Year. However the R^2 of this model is roughly 34%, which indicates that our model only explains roughly 34% of the variation in our data. However with the introduction of a second independent variable, Displacement, we were able to make a more robust model (2). This model was much more accurate giving us roughly a 220% increase in our R^2 value. These results indicate that Model Year isn't the strongest indicator for MPG. This means that focusing on Model Year may not be sufficient for improvements in fuel efficiency however adding additional features can help us predict it.**

I. Introduction

As the automobile industry advances we have seen a push towards environmental friendly options. One major component of this environmental conscious trend is fuel efficiency. It is a way of building on our current infastructure, gasoline, and making it more efficient and environment friendly. We see this obvious trend today but we were curious if this was always the case. With our dataset containing data on automobiles from 1970-1981, an 11 year span, we can investigate if this trend existed. This can be done by analyzing the correlation between our MPG feature and our Model Year feature. We will analyze this correlation using several methods such as data visualizations, correlation analysis, and linear regression models.

II. Dataset

Our dataset is from UCI Machine Learning Library where it was acquired and modified from StatLib library maintained by Carnegie Mellon University. It contains 9 attributes including: MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration, Model Year, Origin and Car Name. The dataset has 398 datapoints for each attribute with 6 missing values for horsepower.

The variable MPG is our variable of interest. It indicates Miles Per Galon which is the distance, in miles, that an automobile can travel per gallon of fuel.

The next variable of interest is Model Year which is the year the automobile was manafactured. This is our ideal predictor or independent variable, as we want to see the change of mpg, our dependent variable, over time.

The other 7 attributes in this dataset are also potential predictors for mpg and can be defined as the following. Cylinders refers to the amount of cylinders in the automobiles engine. Displacement is the total volume of air displaced by the automobiles engine pistons. Horsepower is a unit of measurement for the engines power output. Acceleration refers to the automobiles ability to accelerate or increase its velocity. Origin refers to the country it was manafactured in, with a value of 1 indicating United States, 2 indicating Europe, and 3 indicating Japan.

III. Data Investigation

Acknowledgements

Acknowledgements text goes here.