Correlation between Model Year and its Fuel Efficiency (MPG)

Aden Letchworth, Tanay Shah, and Malik Nasla

***Abstract*—The automobile industry is constantly evolving with technology, one of the most importance 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 over 33,000 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 1.7%, which indicated that out model only explains roughly 1.7% of the variation in our data. However with the introduction of a second independent variable, we were able to make a more robust model (2). This model was much more accurate giving us roughly a 799% increase in out R^2 value. These results indicate that Model Year isn't the strongest indicator for MPG. This means 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 infrastructure, gasoline, and making it more efficient and environmentally 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 1970-2024, a 54 year span, we can investigate if this trend existed. This can be don't 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

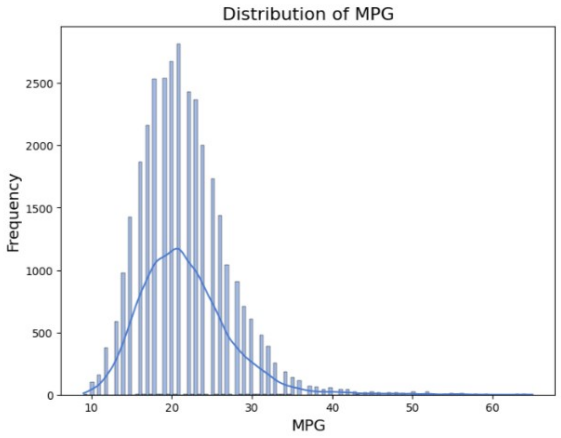
Our dataset consists of two different datasets merged together. Our first dataset is from UCI's Machine Learning Library where it was quired and modified from StatLib library maintained by Carnegie Mellon University and it has 9 attributed including: MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration, Model Year, Origin, and Car Name. The dataset has 398 datapoints for each attribute with 6 missing values for horsepower. Our second dataset came from fueleconomy.gov which is an official U.S government website that is guaranteed accurate since the Department of Energy has approved all the cars that are on the road. We were able to get about 33,000 data points from this dataset. We cleaned these datasets by removing unwanted columns and structured them by standardizing the data types and measurements uniformly. We then merged them into one dataset.

The variable MPG is our variable of interest. It indicated Miles Per Gallon 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 manufactured. This is our ideal predicator or independent variable, as we want to see the change of mpg, our dependent variable, overtime.

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 vehicle's engine. Displacement is the total volume of air displaced by the automobiles engine pistons. Horsepower is a unit or 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 manufactured in, with a value of 1 indicating United States, 2 indicating Europe, and 3 Indicating Japan.

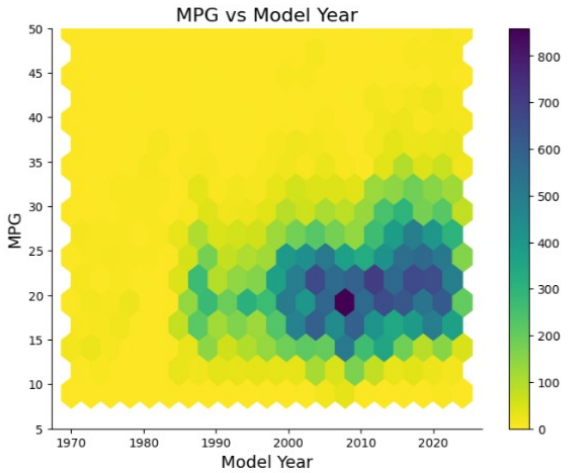
III. Data Investigation

The first thing we did in our analysis was take a closer look at our dependent variable (MPG).

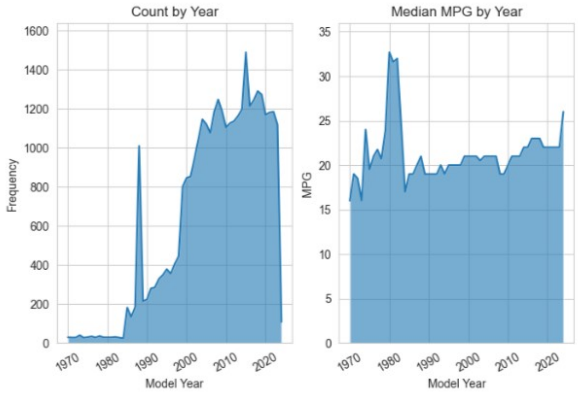


We wanted to see the distribution of MPG so we used a histogram to analyze and the chart above was the result of that. We were able to see the data visually and notices it is unimodal and slightly skewed to the right and that there seems to be a fair amount of extreme values. Noticing this, we calculated the skewness and kurtosis values to confirm our assumptions. The skewness value turned out to be 1.253741 and the kurtosis value was 3.951499. The skewness value indicated that we have a positive skew meaning our data falls slightly above the median and the high kurtosis value indicated that we have a lot of extreme values confirming what the graph showed. The high kurtosis value caused us to want to further investigate the extreme values and after doing so, we realized that they are actually intrinsic to our data. These values are not extreme because of a mistake in the data so they are important for us to keep in the dataset.

We then made a hexbin plot to show the relation between MPG and Model Year. We used this kind of plot because it showed the densities very well, helping us visually see what is going on.

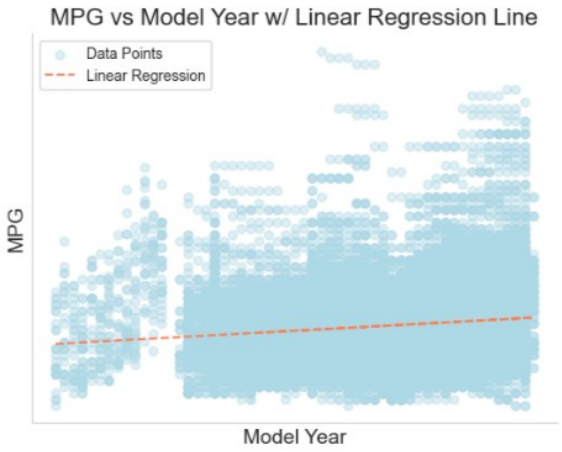


We can see in the graph that there seems to be a slight positive trend. It also shows that most of our data is concentrated between 2000-2020 and 15-25 MPG, indicated by the densities. Since this chart only gives us a rough idea of what is going on we decided to calculate the correlation coefficient which turned out to be 0.13710078939342699. This is a pretty small correlation coefficient indicating that there is little to no linear relationship between the two variables.

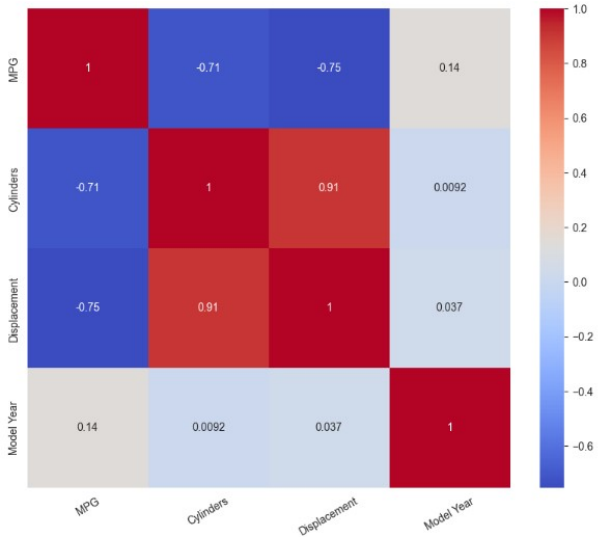


These next two graphs helped us further investigate the correlation between MPG and Model Year. Graph 1, on the left, is similar to our very first graph, it is just visualizing the frequency of Model Years. Graph 2, on the right, is showing the median MPG for each year. For the first 20 years, the data looks pretty unstable in graph 2 and we assume that is because there was very limited data from those years as shown by graph 1. As the years went on, there was much more datapoints for us to work with and that is when the median MPG per year graph mellowed down and showed relatively no increase in the median mpg from about 1990 to present day.

To further investigate the relationship between the two variables we implemented a Simple Linear Regression because it will give us insight to the relationship between our variables and help us identify any issues with our data. It will also give us a nice visualization as we can see the slope of our line in accordance to our data.

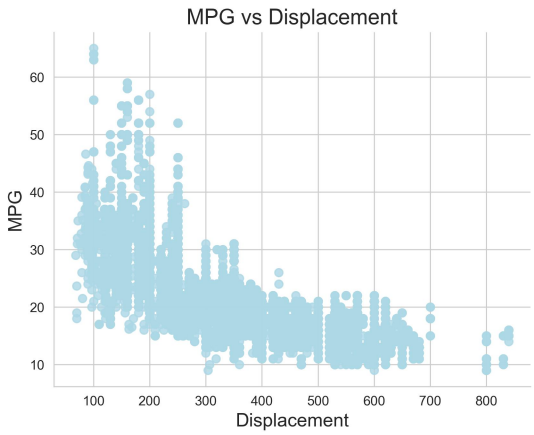


This graph provided a great visual for seeing our predicted value versus our actual values. However there is obviously a lot of variability in this graph. It would be beneficial to use a more robust regression analysis. Considering all the variability of this graph, we wanted to find a model that fits our data better. A good fit would be a multiple linear regression, but we need to find another feature. The feature we are looking for should have a strong correlation with MPG and it shouldn't be linearly correlated with Model Year. To find the feature we are looking for we implemented a heat map that shows the correlation strength of the variables with each other.

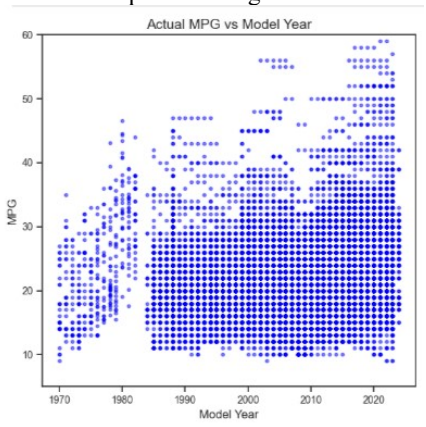


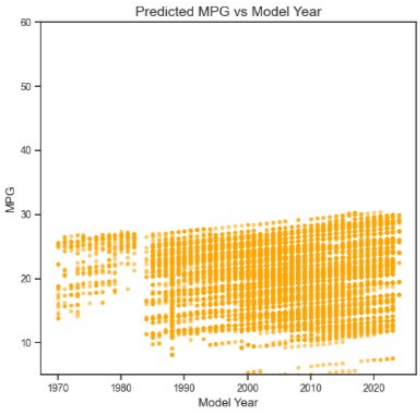
As I just mentioned, the purpose of the heat map is to see the correlation between variables. If the value is greater zero there is a positive correlation while if the value is less than zero there is a negative correlation. From this heat map we can see that displacement and cylinders have the strongest correlation with MPG and both are strongly negative meaning as they increase, MPG decreases. Since these two attributes satisfy the conditions we were looking for, we had to check the multicollinearity which is evaluated using a metric called VIF. What we want to look for is which of the two attributes has a lower value. The VIF score for cylinders is 11.429758 and for displacement it was 7.146666. From this we can see that Displacement has the lowest VIF score which is our best candidate for multiple linear regression. We already know that Displacement and MPG have a strong

correlation from the correlation matrix heatmap, however it is still useful to visualize it so we can see any possible outliers and analyze our curve type.

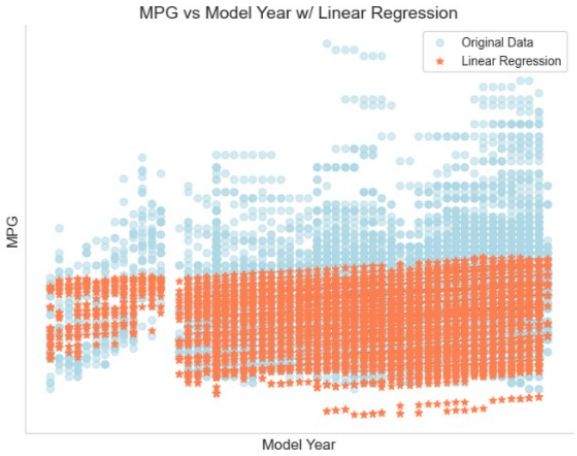


With this visualization we see that there seems to be a strong correlation, possibly exponential decay, between MPG and Weight. We can also see some possible outliers. Through the multiple linear regression, we were able to get a chart of our predicted MPG vs Model Year which is displayed. We compared it to our actual MPG vs Model Year data and as you can see it obviously fits the data much better than our simple linear regression.





To help further see the two compared to each other, we overlapped them. This really helped visualize that our predicted values were relatively accurate. Obviously we weren't able to predict the top half of the graph, but we were relatively accurate when it came to the bottom half.



To get an understanding of the of the magnitude between the simple linear regression and multiple linear regression we calculated the R^2 score because it is a good value to use for comparing correlation between two different things. The R^2 score for our simple lineaer regression was about 0.017 and the R^2 score for our multiple linear regression was about .60. This shows that out multiple linear regression explains our variation more by a magniture of 3363.

IV. Challenges and Solutions

We ran into a few challenges while trying to test our hypothesis. The first challenge we ran into was enriching the data with other datasets. It was difficult to find relevant automobile data that would work with our existing dataset. Integrating the new data was also a challenge because the data we found had each year in a separate file and every few years, the format of the file would change like the column names. The solution we figured out was to use a combination of pandas and sqlite3. This allowed us to process each data set year by year and also standardize them to a final data frame. Creating an accurate model was also causing us some trouble at first. We initially used a simple linear regression to try and predict the correlation but that turned out to be very inaccurate. To solve this we included more attributes to our predictor and were able to improve our model's performance despite the weak correlations between the original set of attributes and the target value.

V. Conclusion

After our intensive investigation, we concluded that our hypothesis was incorrect. To our surprise, model year is not the best predictor for MPG or fuel efficiency. This is an unexpected result due to the advancements in technology and societal pushes for more fuel efficiency over the years. Despite this result we were still able to predict MPG with the help of better correlated predictors. This allowed for our model to get from 0% accuracy to that of 60% accuracy.