# DS210 Final Project

* Link to Tableau: <https://public.tableau.com/views/DS210FinalProject/EffectsofWeightonMPG?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>
* Tableau:
  + Worksheet 1 – The Effects of Weight on Miles per Gallon – The first graph is a simple scatterplot graph, depicting the effect of weight on MPG. The scatterplot graph shows a clear negative relationship between weight and MPG – meaning, the more a car weighs, the lower the miles per gallon it will have. Therefore, we can assume heavier cars will have lower fuel efficiency.
  + Worksheet 2 – Cylinder Effects of MPG, Horsepower, and Acceleration – This next graph is a triple line graph showing the relationship between a car’s cylinder number and MPG, Horsepower, and Acceleration:
    - Cylinder and MPG: Average MPG decreases as the number of cylinders increases.
    - Cylinder and Horsepower: As the number of cylinders increases, the average horsepower seems to rise significantly.
    - Cylinder and Acceleration: Average acceleration seems to improves with the increase in the number of cylinders; however, the relationship is not as linear as horsepower.
  + Worksheet 3 – MPG Over the Years – This graph shows the trend of MPG over the years and can be filtered by three regions: Europe, Japan, and the USA.

Generally, there is an increasing trendline, showing that over the years, cars have increased MPG. However, when filtering by region, you can see that this applies to each region differently. Specifically, Japan and the USA seem to have much steeper slopes, indicating a more drastic increase in MPG, while Japan, though still showing an overall increase, has a much smaller rate of change.

* Worksheet 4 – MPG by Car – This is a clear, simple bar chart depicting each individual car type’s MPG, showcasing Toyota Corolla at the very top with the highest MPG, all the way down to hi 1200d, with the very lowest MPG.
* Part 1:

First, checked the dataset structure using the str() function in R. Converted the horsepower variable to numeric (otherwise, it would have been interpreted as null, and thus removed from the dataset). Checked to see if there was any missing data, and as there was, removed any rows with missing data using the na.omit() function. Split the data into first 300 samples, and last 98 samples.

First 300 samples:

Simple Linear Regression:

Multiple R-Squared = 0.7714

Adjusted R-Squared = 0.7706

Linear Regression Equation: y = 40.6 +-0.00629 \* weight

Multiple Linear Regression

Multiple R-Squared=0.7804

Adjusted R-Squared=0.7782

Multiple Linear Regression Equation: y = 39.6+ - 0.0226 \* horsepower + -0.0048 \* weight + -0.00525 \* displacement

Displacement did not show a statistically significant effect on mpg, as its p-value was greater than 0.05. Based on this, I removed displacement from the model, so the new model became:

Multiple R-Squared = 0.7796

Adjusted R-Squared = 0.7781

Multiple Linear Regression Equation: y = 40.3 + -0.0278 \* horsepower + -0.0052 \* weight

Last 98 samples:

Linear Regression:

Multiple R-Squared = 0.5721

Adjusted R-Squared = 0.5625

Linear Regression Equation: y = 57.9 + -0.125 \* horsepower + -0.00646 \* weight

To get the residuals - calculated predicted mpg, then subtracted that from actual mpg to get residuals.

Created a residual plot, which showed that the residuals were randomly scattered around zero, indicating that the model's assumptions were met and there were no obvious patterns.

A histogram of the residuals indicated that the residuals were approximately normally distributed, with a general bell shape, showing that the model fit the data well.

Analysis:

Based on my analysis, I found that weight and horsepower are significant predictors of mpg. The simple linear regression model, which used only weight as a predictor, explained about 77% of the variation in mpg. The multiple linear regression model, which included both horsepower and weight, showed a slightly better fit. However, adding displacement did not significantly improve the model, so I removed it from the analysis.

The model performed well on the first 300 samples of the data, but when applied to the last 98 samples, the fit was a little weaker, which could be due to differences between the two parts of the dataset. The residual analysis showed no major issues, suggesting that the model’s assumptions were satisfied.

* Part 2:
  + Question 1 – What is the most common reason for calls?

To answer this question, I created a bar plot, where the x-axis represents the different reasons for calls, and the y-axis represents the number of calls for each reason.

The bar plot clearly shows the distribution of calls across different reasons. While payments and service outage both seem to be minor reasons for calls, billing questions take up a large majority of the calls. Thus, the reasons with the highest number of calls (billing questions), can now be targeted for further investigation.

* + Question 2 – What is the relationship between response time and customer satisfaction score (CSAT)?

To answer this question, I used a boxplot, where the x-axis shows the response times in minutes, while the y-axis shows CSAT scores.

There are three SLA categories: Below SLA, Within SLA, and Above SLA. Below SLA refers to responses faster than the agreed time, Within SLA means responses within the agreed time, and Above SLA indicates slower responses. The boxplot shows that faster response times (Below SLA) lead to higher customer satisfaction (high CSAT), while slower response times (Above SLA) result in lower CSAT. However, there are some outliers where customers with longer response times were still satisfied, indicating that response time is not the sole factor influencing customer happiness.

* + Question 3 – Which states have the highest number of calls?

To answer this question, I created a bar plot, with the x-axis representing the states and the y-axis representing the call count.

The bar plot shows that certain states (such as California and Florida) have a much higher volume of calls, which could simply be because of their larger populations or could indicate more problems in those regions.

States with fewer calls, like Wyoming or Maine, could indicate either smaller populations or less problems in those regions. These insights could be used to adjust service strategies based on regional demand.