DS-210 Final Project Report

Batsheva Levin

**Part 1 - Tableau**

Sheet 1: Car Names and Average mpg

This graph shows the car names in alphabetical order and each one’s average mpg . The filter to the left of the graph can be used to display only cars with an mpg within a certain range.

Sheet 2: Weigt of Cars Over the Years

This graph shows how the average weight of cars changes over the years. There are many factors that contribute to the change in weight of cars. New safety and security features add to the car’s weight, while the usage of lighter materials helps to decrease the weight.

Sheet 3: How Horsepower and Weight Impact Acceleration

This shows two scatterplots side by side depicting how weight and horsepower affect the cars’ acceleration. A car’s weight and horsepower both seem to have a (weak) inverse relationship with its acceleration. However, there are other factors that contribute to the cars’ acceleration as well.

**Part 1 – RStudio**

Firstly, I converted the horsepower column from characters into numeric values so that it can be used in regression. I also removed all NA values from the dataset.

Simple linear regression (with mpg and weight)

**R-squared:** 0.7741

**Adjusted R-squared:** 0.7733

**Complete linear regression equation:** mpg = 40.388 - 0.006 \* weight

I created a scatterplot showing the relationship between mpg and weight. We see that there is a strong inverse relationship between them.

Multiple linear regression (full model)

**R-squared:** 0.823

**Adjusted R-squared:** 0.8187

**Complete linear regression equation:** mpg = 5.8112 - 0.456(cylinder) + 0.01(displacement) – 0.017(horsepower) – 0.005(weight) – 0.028(acceleration) + 0.444(model.year) + 0.993(origin)

After running the full model, I removed all variables with a p-value above the significance level (0.05). My reduced model contained only variables that are statistically significant: origin, model year, and weight.

Multiple linear regression (reduced model)

**R-squared:** 0.8206

**Adjusted R-squared:** 0.8188

**Complete linear regression equation:** mpg = 3.229 + 0.85(origin) + 0.459(model.year) – 0.006(weight)

I calculated my reduced model’s predicted mpg values and created a data frame with the actual and predicted values for each of the last 98 cars. I then subtracted the predicted values from the actual values to get the residuals

In the residual plot, the points are randomly scattered, indicating that the model is a good fit. However, the points seem to be positively skewed and not completely centered around zero, suggesting that although the model is a good fit, it may not be perfect.

The histogram has an approximate symmetric bell shape, which means that the data is normally distributed. However, this graph is also slightly right-skewed.

**Part 2**

The first question that I analyzed was: “Why do customers call?” I created a table that counts how many calls there were for each reason (billing question, payments, or service outage). I then created a pie chart showing the percentage of calls for each reason out of the total calls received by the call center. Each color in the chart represents a different reason, as shown in the graph’s legend.

The next question was, “Where do most customers call from?” I wanted to find out the top 5 states that the call center receives calls from. I created a table of all states and the number of calls for each and sorted it from greatest to least. I then printed out the first 5 states, which are the states with the highest number of calls. The top 5 states were: California, Texas, Florida, New York, and Virginia.

Last, I wanted to observe, “How do customers rate our service?” Customers have an option of rating the company as “very positive”, “positive”, “neutral”, “negative”, or “very negative”. I created a bar chart of customer ratings (sentiment) to show the distribution of customer satisfaction responses for the call center.