Shalva Stulberg

Data Science Final Project: Report

**Part One:**

After loading the auto\_mpg dataset, I split the dataset on the 300th row and used the first 300 rows for my regression models. Afterward, I made the column ‘horsepower’ from a string to an integer, enabling me to run regression analysis on it.

To find a multiple linear regression model with a good fit, I first ran a linear regression model with all continuous variables from the dataset.

*equation for the full model:*

*mpg = 5.8 - .46 \* cylinder + .01 \* displacement - .02 \* horsepower - .01 \* weight + .99 \* origin + .44 \* model.year - .03 \* acceleration*

*R squared = .823*

*adjusted R-squared = .8187*

After running that model and seeing which variables were statistically significant to the mpg, and after playing around with variable selection, the multiple linear regression model with roughly the best fit, was the one containing all of the statistically significant variables.

*equation for the multiple linear regression model:*

*mpg = 3.223 - .006 \* weight + .46 \* model year + .85 \* origin*

*R-squared = .82*

*adjusted R-squared = .819*

I picked the variable with the lowest p-value (which means it has the highest correlation to mpg) and used it in my simple linear regression model.

*equation for the simple model:*

*mpg = 46.3 - .008 \* weight*

*R-squared = .69*

*adjusted R-squared = .69*

I then turned to the last 98 rows of the dataset and used the multiple linear regression model to predict the last 98 rows of data.

To judge the goodness of the model’s fit, I made a residual plot, by first calculating the residuals (the difference between the actual mpg and predicted mpg) and then plotting them. I put a line at y = 0 so we can see if the spread of residuals is normal, and randomly distributed above and below the x-axis. In this plot, the points are randomly scattered, and follow no clear pattern. It seems that not all the data is equally above and below the x-axis, making this graph slightly positively skewed, indicating that the fit is good, but not perfect.

In addition, I made a histogram of the residuals, and we can see the same thing here- the plot is not perfectly normally distributed, with a bell-shaped curve. It seems that the graph is skewed slightly to the right.

**Part Two:**

Because this dataset contains primarily categorical data, I needed to transform the columns that I wanted to use, into a table. Once the column was in tabular form, it was split into counts of each distinct value within the column.

The first question I wanted to answer is why our customers call us. Knowing this information is important for the company, so they can take preemptive measures to clarify and troubleshoot the most common questions and complaints. To achieve this, I made a table from the Reason column in the call data. I then made a pie chart from the Reason table. Looking at the pie chart, you can clearly see that billing questions account for almost 75% of all customer calls.

The second question I addressed was how do our customers contact us. Knowing this can help the company with future decisions, like where to invest more time and energy to ensure quick and efficient responses. After making the Channel column into a table, I turned it into a bar graph. By looking at this graph, one can see clearly that the call- center is the most commonly used channel, and they should maybe think about hiring more staff for the call- center.

The third, and most important question is how do our customers feel about us. This includes overall customer satisfaction and gives us some insight as to what we might expect the customer churn rate to be. I took the Sentiment column from the dataset and made it into a table, and turned the table into a dot chart. By looking at this graph, we can tell immediately that many customers have negative sentiments. This is definitely something the company should be looking into.