Esti Rosenbaum

DS210 Final Project

Part 1

Tableau:

After creating many scatter plots of the data, it became clear that the greatest correlation was between weight and mpg. I therefore went ahead and ran a simple linear regression model on weight and mpg.

Simple Linear Regression: (Explaining how *weight* affects the variability of *mpg*.)

I began preparing the data by changing any missing values to null values and converting all data types to integers. I then split the data into training and testing sets. From the training set I ran a simple linear regression model measuring the effect that weight had on the mpg of a car. The R-squared value equaled 0.77. This can be interpreted that 77% of the variance of mpg (dependent variable) can be explained by the weight of the car (independent variable). This R value is relatively high, indicating that a great proportion of the variance is based on the given variable.

Multiple Linear Regression: (Showing the relationship between multiple predictors, not just weight, and mpg)

The adjusted R-squared value is 0.78. The adjusted R-square value tells us how well the independent variables explain the variance in the dependent variable. So, in context this means that 78% of the variance of the mpg is explained based on horsepower, weight, acceleration and displacement. 78% is a pretty high value. It tells us that this model is a pretty good fit and that the independent variables explain a good portion of the variance of mpg.

If you look at the p-value for each predictor, one can see which of the independent variables are significant to the variance of the mpg. Both the p-value for horsepower and weight are less than 0.05. This indicates that they are statistically significant in explaining the variance of mpg. The opposite is true for acceleration and displacement. Because the p-value of both is greater than 0.05, they are less significant in explaining the variance in mpg.

(I interpreted the multiple R-squared value for the simple linear regression and the adjusted R-square value for the multiple linear regression. This is because the adjusted R-square value considers the number of predictors while the multiple R-square value does not. Therefore, when measuring the multiple R-squared value, which takes in many predictors, the adjusted R-squared value is a more accurate measure to use. )

To test the data, I first ran a multiple linear regression using the test data and then made prediction for the mpg. I then created a data frame that showed the difference of the actual mpg and the predicted mpg. With those diffeences, I created a residual plot.

Interpretation:

The residual plot graphs the difference between the actual and predicted values. There was no pattern seen in the plot and all the data points are scattered around the zero line. This indicates that there are no problems in the model and it is a good fit.

The same holds true with the histogram. In the histogram plot the residuals forms a bell shape which indicates that the residuals are normally distributed.

Because both the residual plot and the histogram look just as they are supposed to, we can say that the model that we created is a good model to predict mpg.

I was not able to compile a report of the R code so I screenshotted the code and the output.

Here are the linear regression eqautions:

model <- lm(mpg ~ weight, data = train\_data)

model2 <- lm(mpg ~ horsepower + weight + acceleration + displacement, data = train\_data)

model3 <- lm(mpg ~ horsepower + weight + acceleration + displacement, data = test\_data)

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A close up of a text

Description automatically generated

A screenshot of a data

Description automatically generated

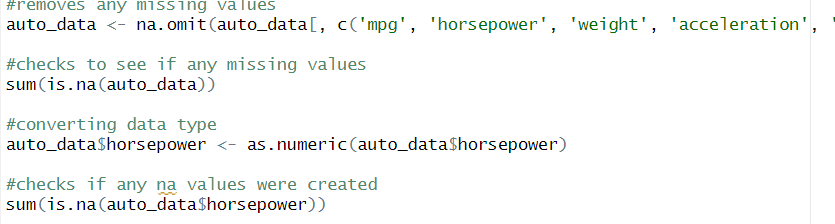
A screenshot of a computer screen

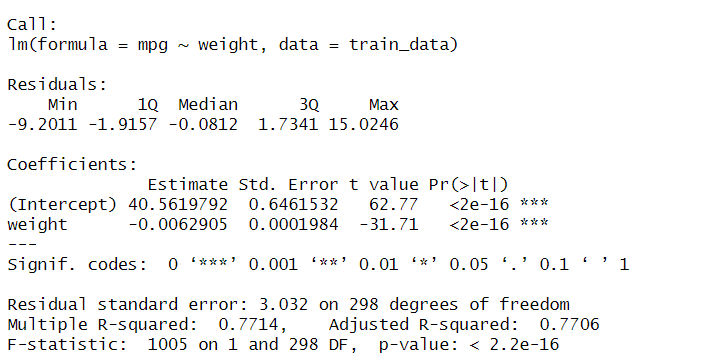
Description automatically generated



A screenshot of a computer code

Description automatically generated

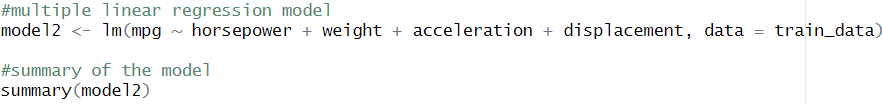


A computer code with text

Description automatically generated with medium confidence

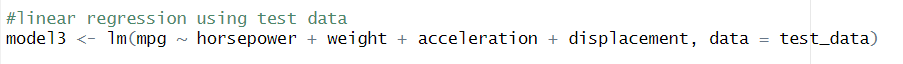
A diagram of a graph

Description automatically generated with medium confidence



A screenshot of a computer

Description automatically generated





A screenshot of a computer

Description automatically generated

A computer code with black and white text

Description automatically generated

A table of numbers with black text

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A diagram of red dots

Description automatically generated

A screenshot of a computer program

Description automatically generatedA graph of a number of individuals

Description automatically generated with medium confidence

Part 2

Question 1:

Is there a correlation between the customers satisfaction score and the sentiments, channel, reason, call duration, city and state of the customer?

Linear Regression equation:

model <- lm(Csat.Score ~ Call.Duration.In.Minutes + Sentiment + Channel + Reason + City + State, data = call\_data)

(The company would like to know how they can increase their customers satisfaction score. With the regression analysis, they can see which factor highly affects the customer satisfaction score. By knowing which factor/s affect the satisfaction score the company can work on improving those specific variables.)

To prepare the data, I removed missing values and converted the data to necessary data types.

By looking at the output of the model one can see that the p-value for all the sentiments are very low. This indicates that sentiments have a very significant impact on the variation of the costumer’s satisfaction score.

Because channel, call duration and reason have high p-values they do not significantly influence satisfaction.

The p-value of the cities varies. Some cities have a high p-value while others have a low p-value. From this we can conclude that the city is not a strong determining factor in terms of customer satisfaction.

It can be concluded that to improve customer satisfaction it would be most effective to improve the sentiments of the conversation with customers. Other factors such as city, call duration and channels are less likely to have an impact on customer satisfaction.

A white background with black text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

…

A screenshot of a computer

Description automatically generated

…

A screenshot of a computer program

Description automatically generated

Question 2

Is there a difference in the mean of the call duration based on the different reasons of customer calls?

(A company might want to know about how long a costumer service call will take based on the reason for the call. Is there a difference in expected time based on different reasons.)

I performed an ANOVA test where the independent variable was the reason for the call and the dependent variable was the call duration.

Null hypothesis: there is no difference in the mean of the call duration based on the reason for the call.

Alternative hypothesis: at least one reason has a mean call duration that is different from the others.

Anova code: anova\_result <- aov(Call.Duration.In.Minutes ~ Reason, data = call\_data)

Interpretation:

Because the p-value is higher than 0.05, we fail to reject the null hypothesis, and we cannot conclude that there is a significant effect on call duration based on the reason.

A close up of a text

Description automatically generated

A black text on a white background

Description automatically generated

Question 3

Is there an association between the city and channel?

I performed a chi-square test of association. (This tests if there is an association between two categorical variables.)

Interpretation:

Because the p-value is greater than 0.05, there is no significant association between the channel and city.

I wanted to see if there would be any two variables that did have an association between each other, so I performed another chi square test, this time between channel and reason. And surprisingly there were significant results! The p-value was less than 0.05, which means that there is a statistically significant association between the channel and the reason. The reason of the call was somehow related to by which means the customer contacted the company.

This can be important information for the company, to know which type of customer request is expected to come through which type of communication.

A computer code with black text

Description automatically generated

A black text on a white background

Description automatically generated

A screen shot of a computer code

Description automatically generated

A black text on a white background

Description automatically generated