Zach Hinz, Joshua Jansen-Montoya
Prof. Talithia Williams
Math158: Statistical Linear Models

Data Set Analysis

For our dataset, we chose to utilize the data set, “Elon Musk Tweet Sentiment (classified via RoBERTa)” from the dataset website, Kaggle. In the dataset, we have 17 variables of interest, Datetime (time at which the tweet was sent), Tweet ID (what number the tweet is), Text (the text content of the tweet), Username (who tweeted [Elon Musk for all tweets]), Location (where the tweet was sent from), Reply Count (number of replies), Retweet Count (number of retweets), Like Count (number of likes), Language (language of the tweet), Twitter Access Point (type of device used to tweet), Follower Count (number of followers the tweet sender has), Friends Count (number of users that the tweeter follows), Verified (whether or not the tweeter has been verified), Date (date at which tweet was sent), Mentions (other Twitter users mentioned in the tweet), and Sentiment (the RoBERTa sentiment classification consisting of either negative, neutral, or positive, and the magnitude of that sentiment from 0-1).

In our investigation, we are looking to investigate the question of which variables best predict the number of likes, retweets, or replies that Elon Musk will get on a tweet. We plan on performing statistical analysis using the variables in our dataset that are not our desired response variable as our predictive variables. Furthermore, from our data exploration, we can see that we have a couple of non-numeric variables that we will be attempting to utilize (such as the sentiment), and from which, we will be constructing a transformation of said variables to make them more conducive to our statistical analysis.