Kyle Hum

Professor Rich Yueh

MGT 290

**MGT 290 Final Project - Machine Learning in R**

My data set consists of information regarding the evaluation of burritos in San Diego, California. The data was gathered mostly from the neuroscience graduate department at UC San Diego. With a total of 385 rows, the dataset contains both burrito evaluation criteria and general information about the individual establishment, such as address, website, etc.. The burritos were evaluated based on 10 core dimensions: Volume, Tortilla quality (Temperature), Meat quality, Non-meat filling quality, Meat-to-filling ratio, uniformity, salsa quality, flavor synergy, and wrap integrity. The dimensions were rated on a scale from 0 to 5. 0 being horrible and 5 being ideal.

My project will consist of two machine learning algorithms. The first algorithm utilizes linear regression to predict the overall rating of burritos using a highly correlated variable of non-meat fillings. The second algorithm utilizes logistic regression to predict a binary classification variable. Factoring in all 10 core dimensions that the burritos are evaluated on, the model will predict whether or not the reviewers will recommend it or not recommend it.

The dataset will be loaded into R using this command:

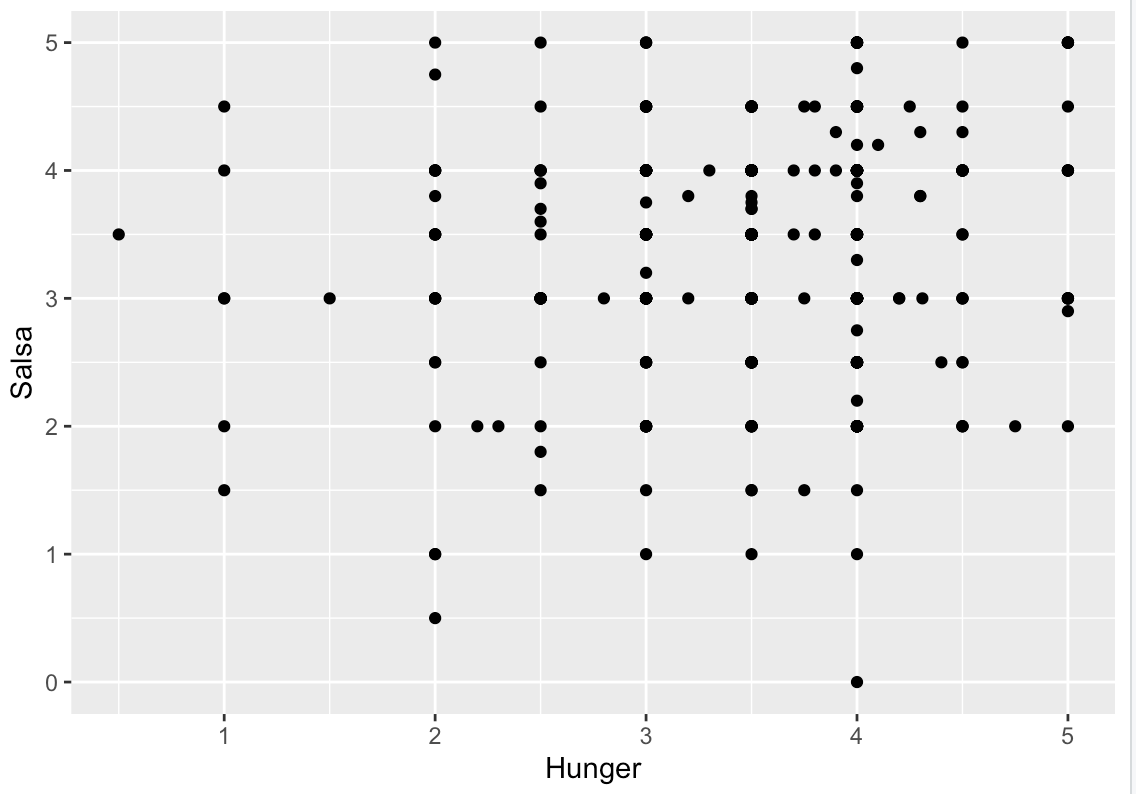
library(readr)

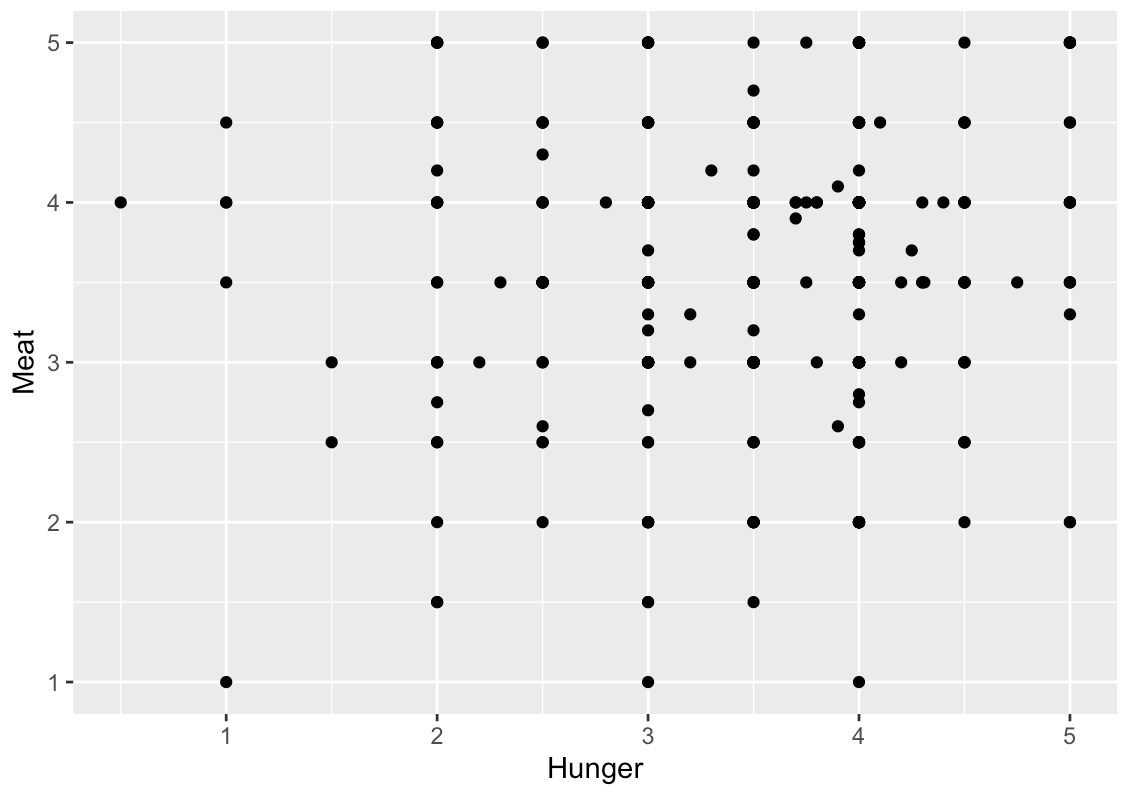
burritos <- read\_csv("burritos\_01022018.csv")

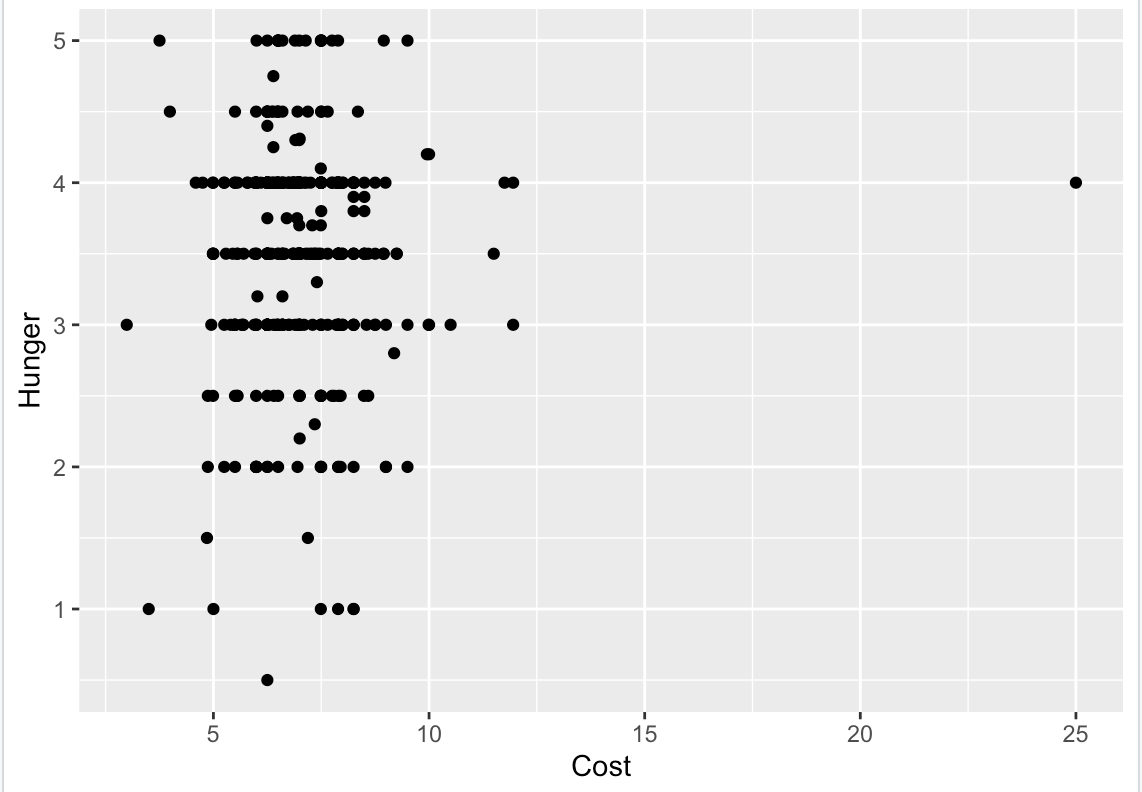
read(burritos)

**Linear Regression**

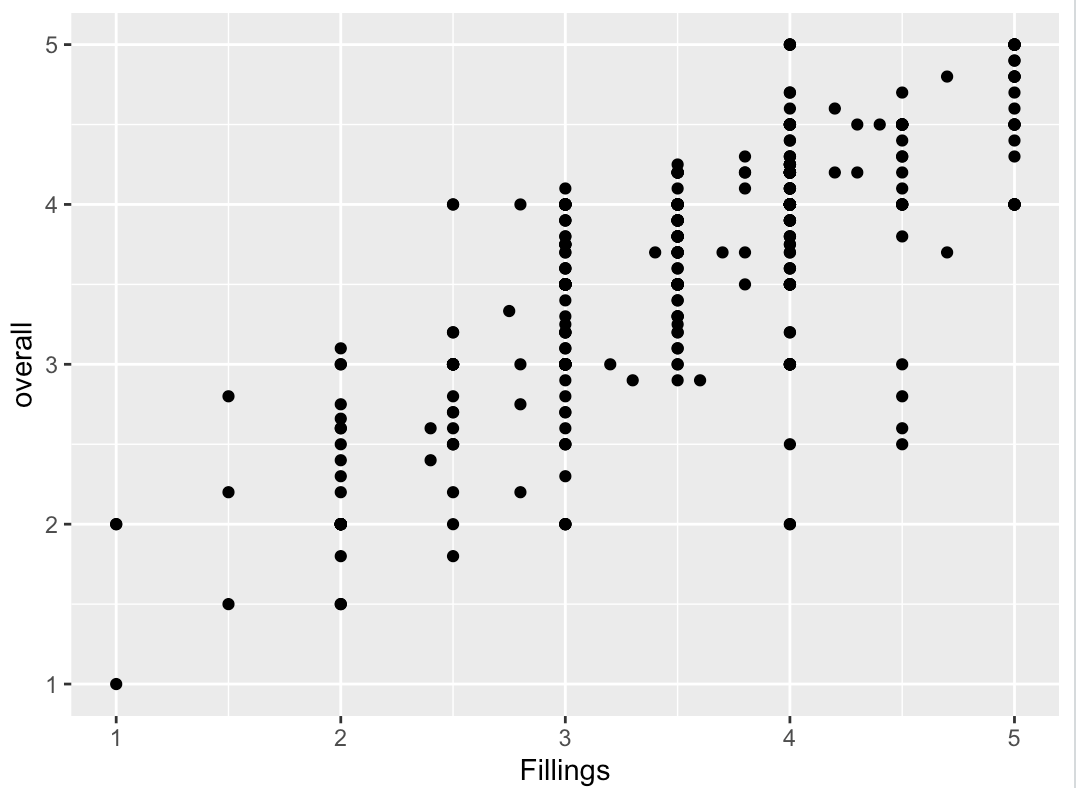
Once the data is loaded into R, I will identify a pattern to base my algorithm off of. I will utilize the ggplot2 package with a scatterplot to visually identify any correlations between variables. Listed below are a few sample scatterplots:



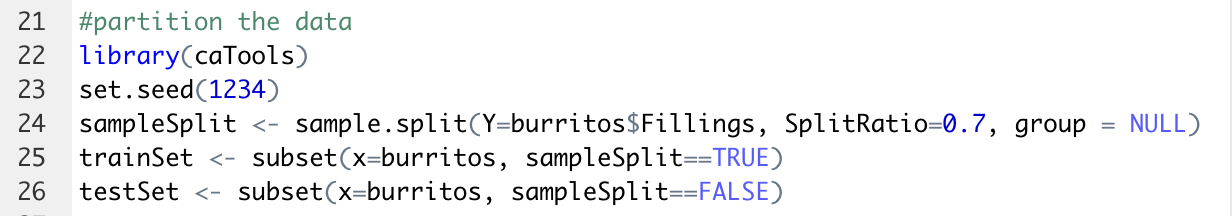




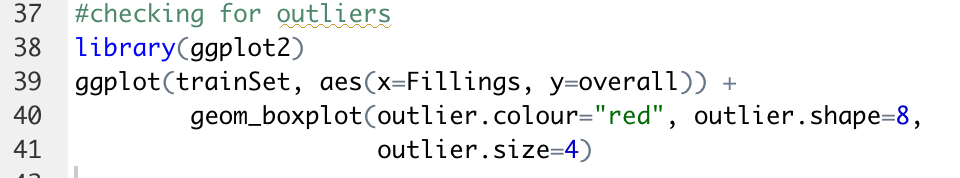
After running multiple scatterplots, I have identified a high correlation between Fillings (non-meat fillings) and overall score.

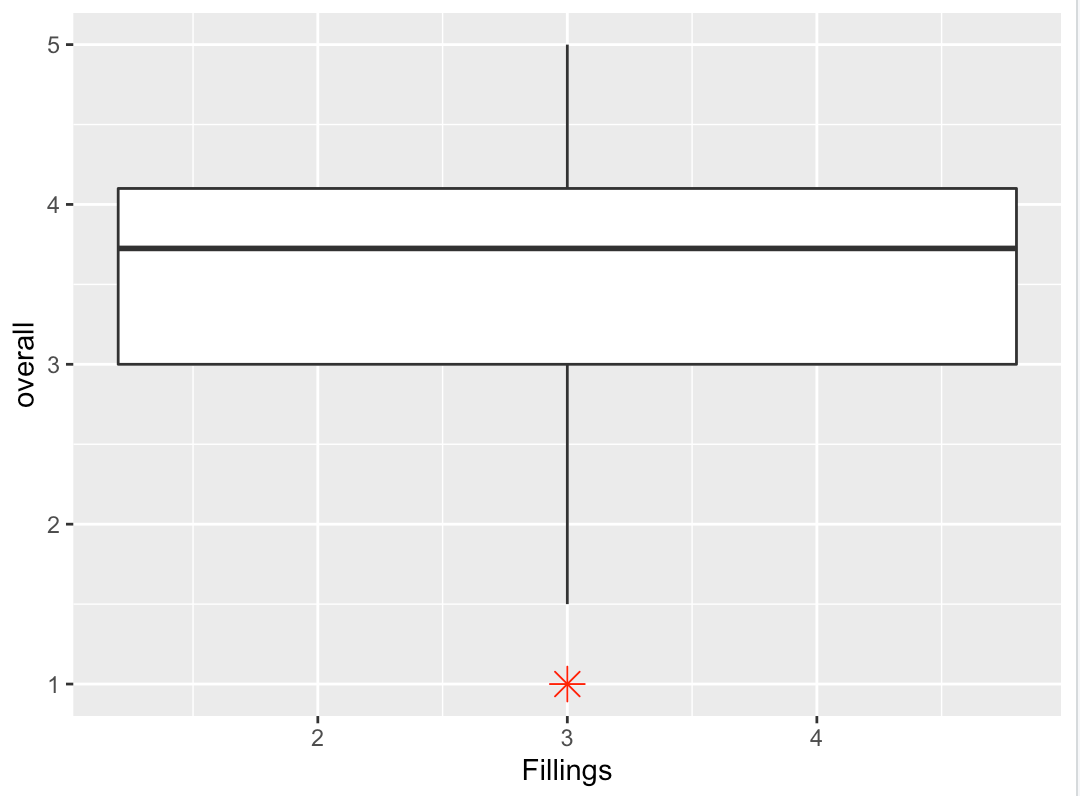


With a visual correlation identified, I can partition the data into training and test sets. The training set will be used to develop the predictor model on. While the test set will be used to evaluate the predictor model.

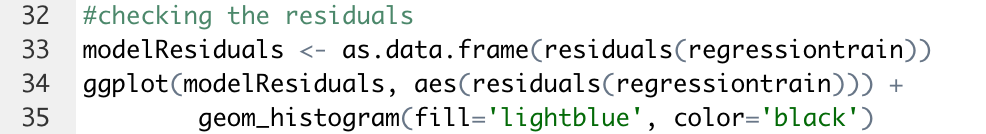


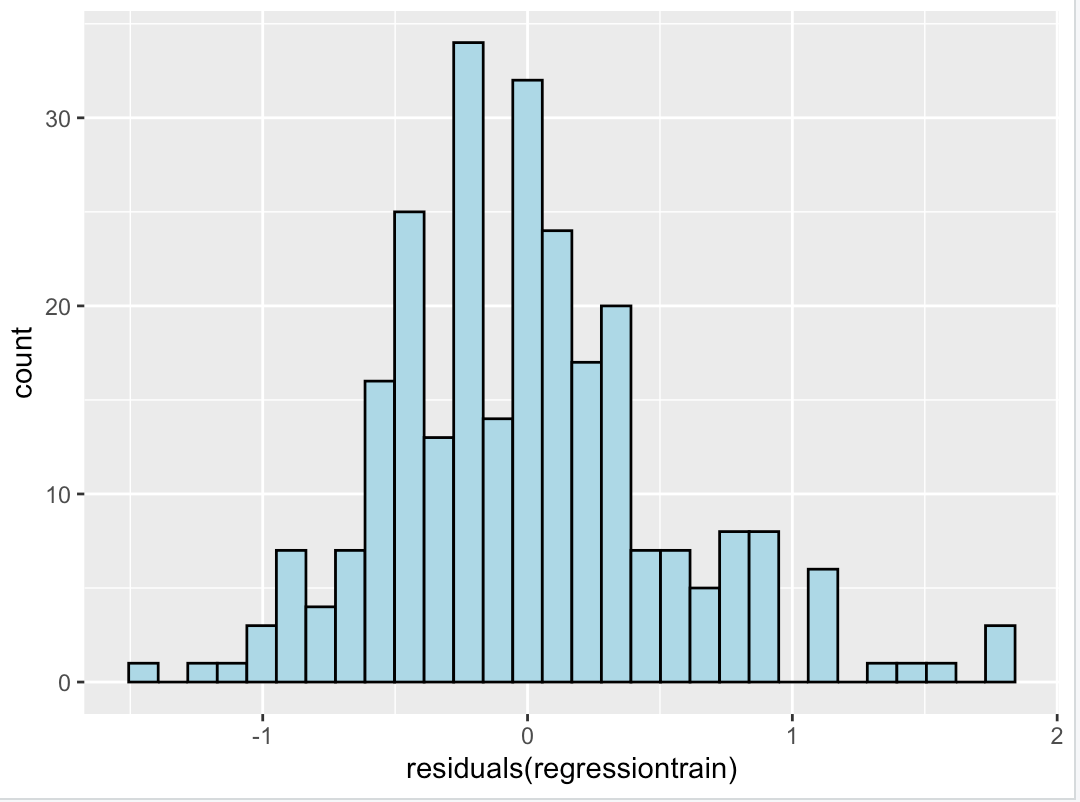
Before developing a model for prediction, I will run a few analysis to check the validity of the dataset. Creating a boxplot will identify any outliers within the training set. With only 1 outlier identified, my training set appears to be valid



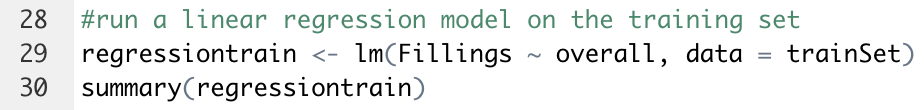


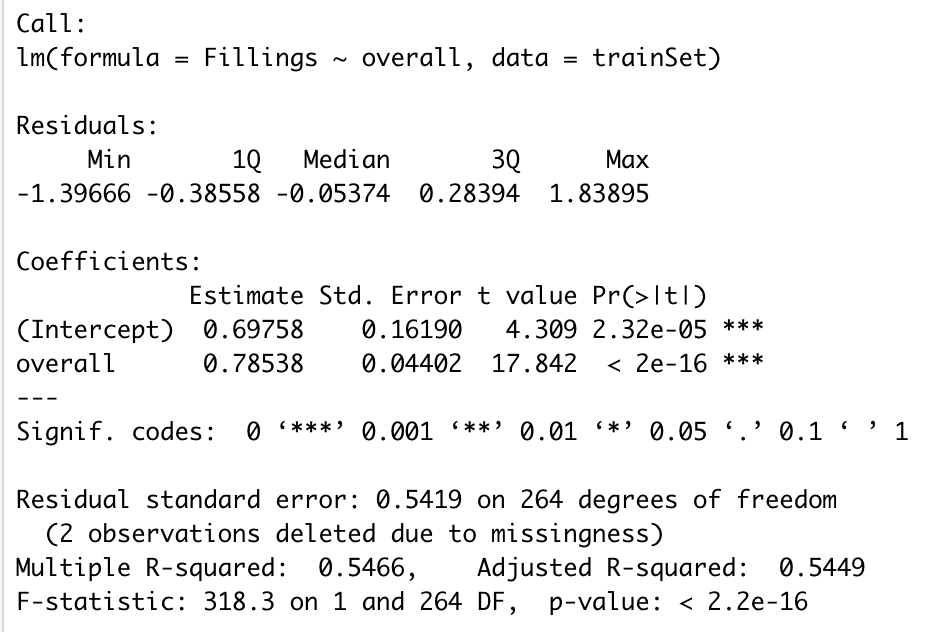
Next, I will create a histogram of a residual plot. This will determine if there is any unexplained variance within the data set. As shown in the histogram, the dataset appears to have a bell shaped curve and can be determined as having a normal distribution



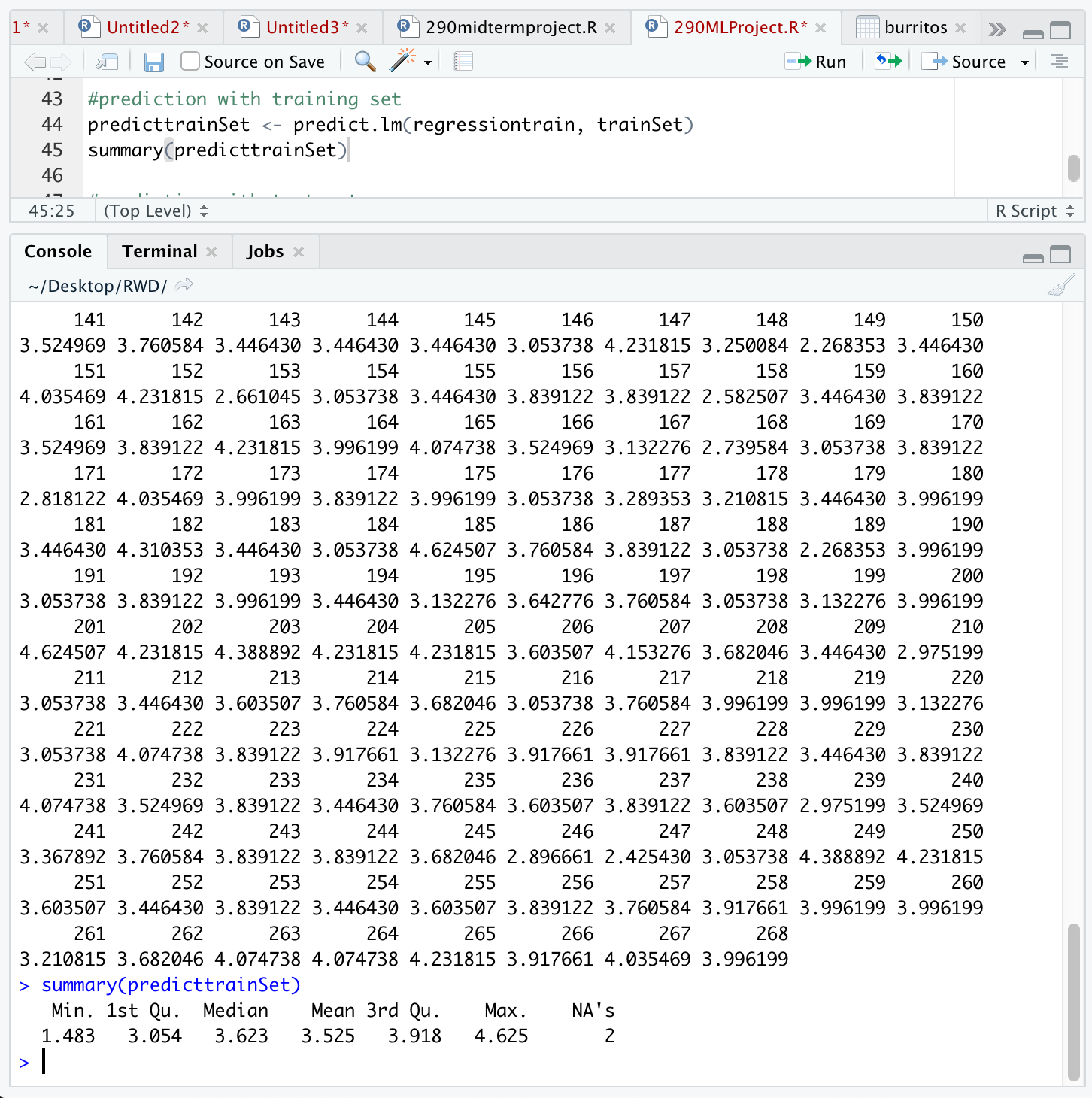


With the validity of the dataset confirmed, I will run a linear regression on the training set

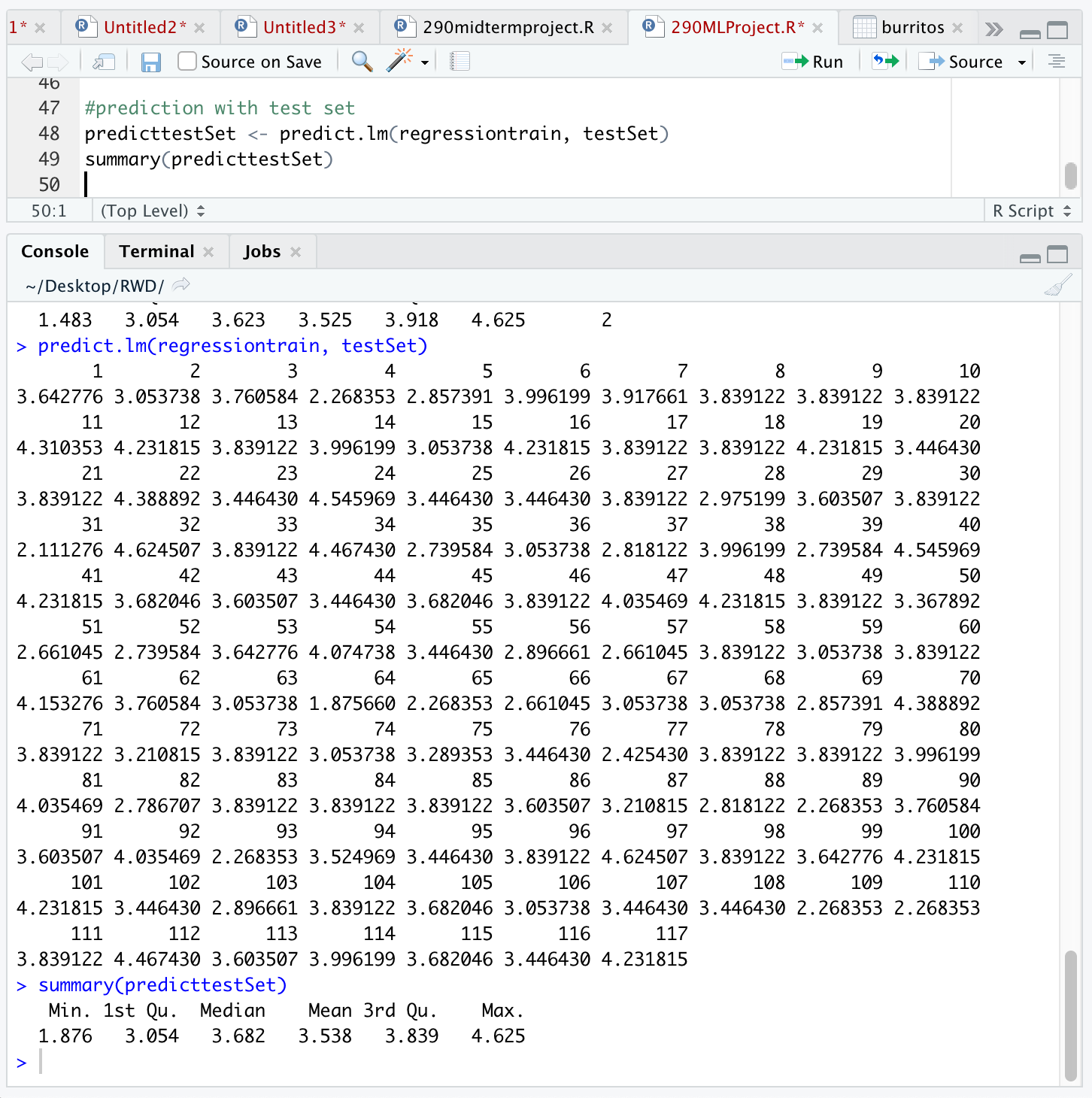




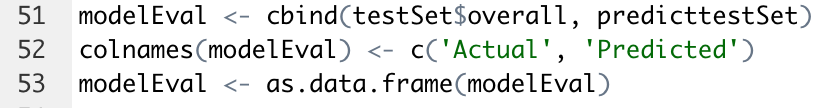
Once the linear regression model has been established, I can use the predict.lm function to predict the overall score of burritos in correlation with non-meat fillings. The first prediction is run on the training set.

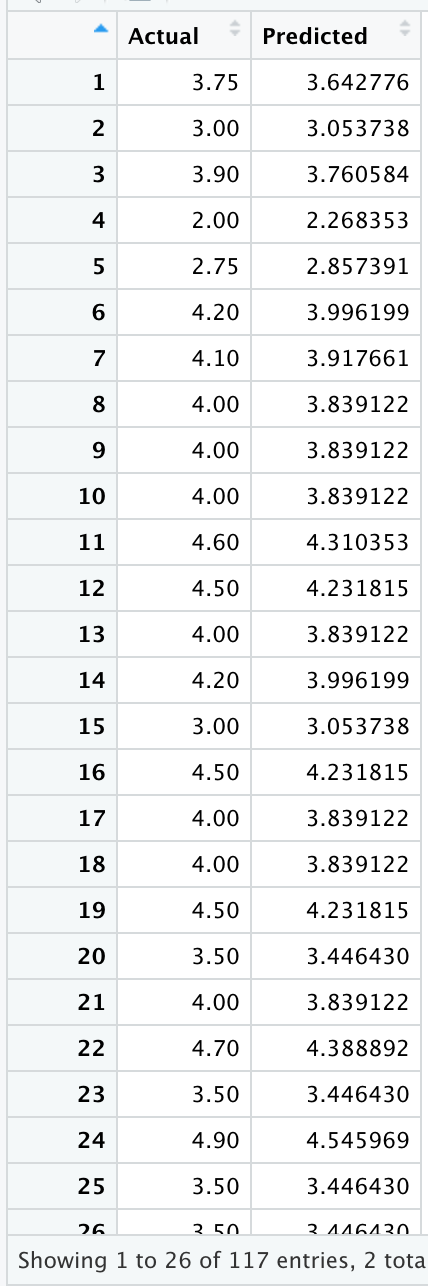


Once the model is fitted to the training set, I can fit the model to the test set



To compare the actual ratings of “overall” and the predicted values, in an easy and intuitive manner, I column binded the actual ratings with the predicted values into a new data frame

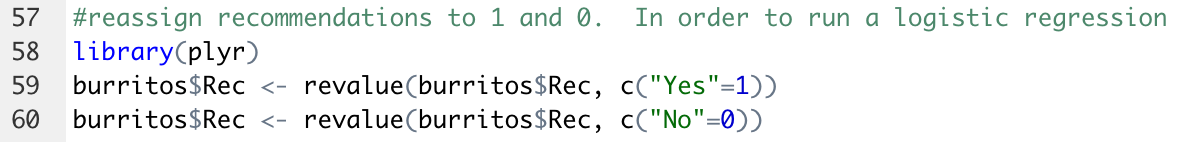




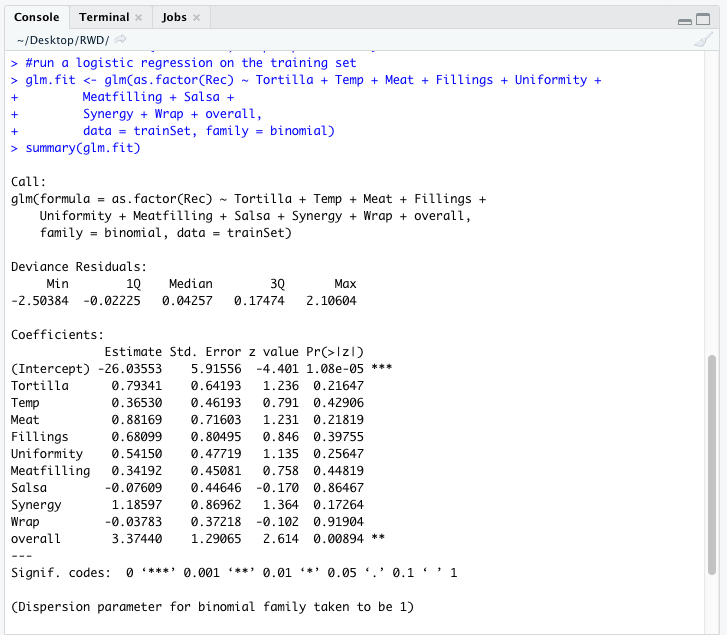
**Logistic Regression**

A logistic regression analysis has the ability to determine and predict whether or not the reviewers will recommend each burrito restaurant. In the original dataset, the reviewers recommended the restaurant by inputting an uppercase X. A lowercase x signified the reviewers would not recommend the restaurant. However, most of the reviewers did not enter in a value, resulting in a NA value. Therefore, my algorithm will solve this problem.

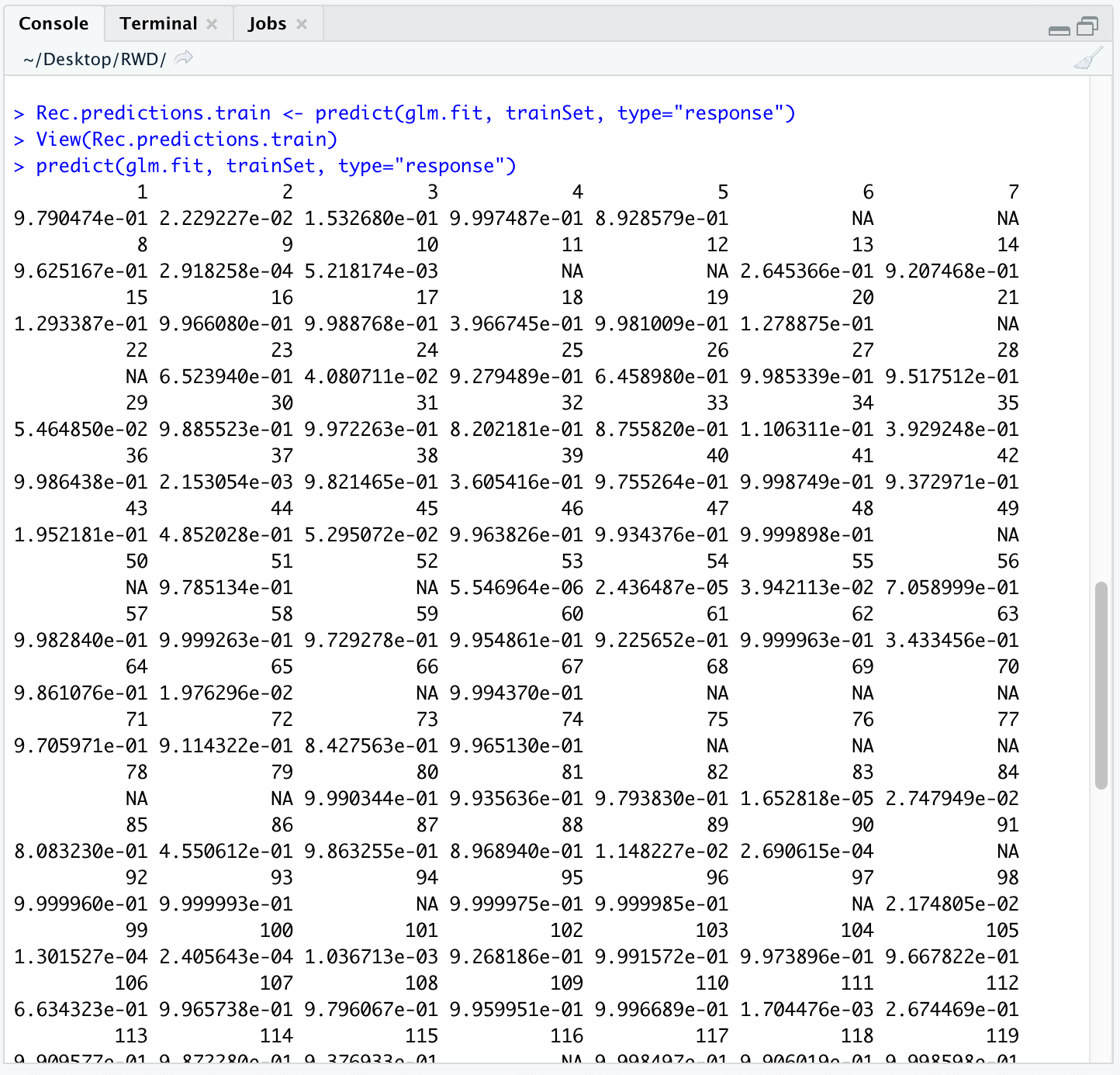
In order to run a logistic regression, the dependent variable must be valued as binary classification variable, assigned as 0 and 1. Therefore, I will use the plyr package to revalue a “Yes” to 1 and a “No” to 0.



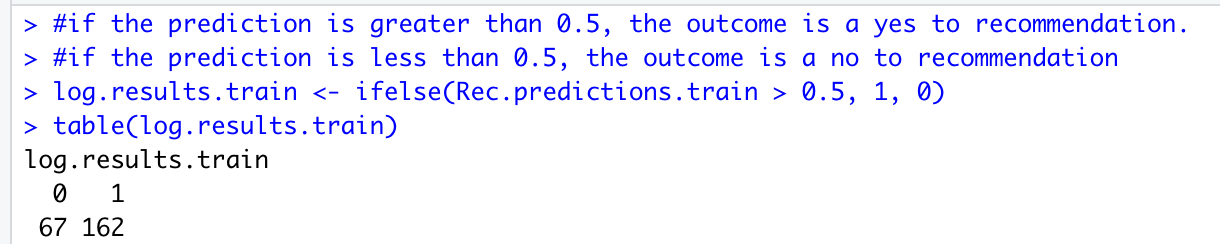
Once the values are reassigned, I can run a logistic regression with “Rec” as the dependent variable and all 10 dimensions as the independent variables. As my data set was previously partitioned from the linear regression, I will utilize the same training set.



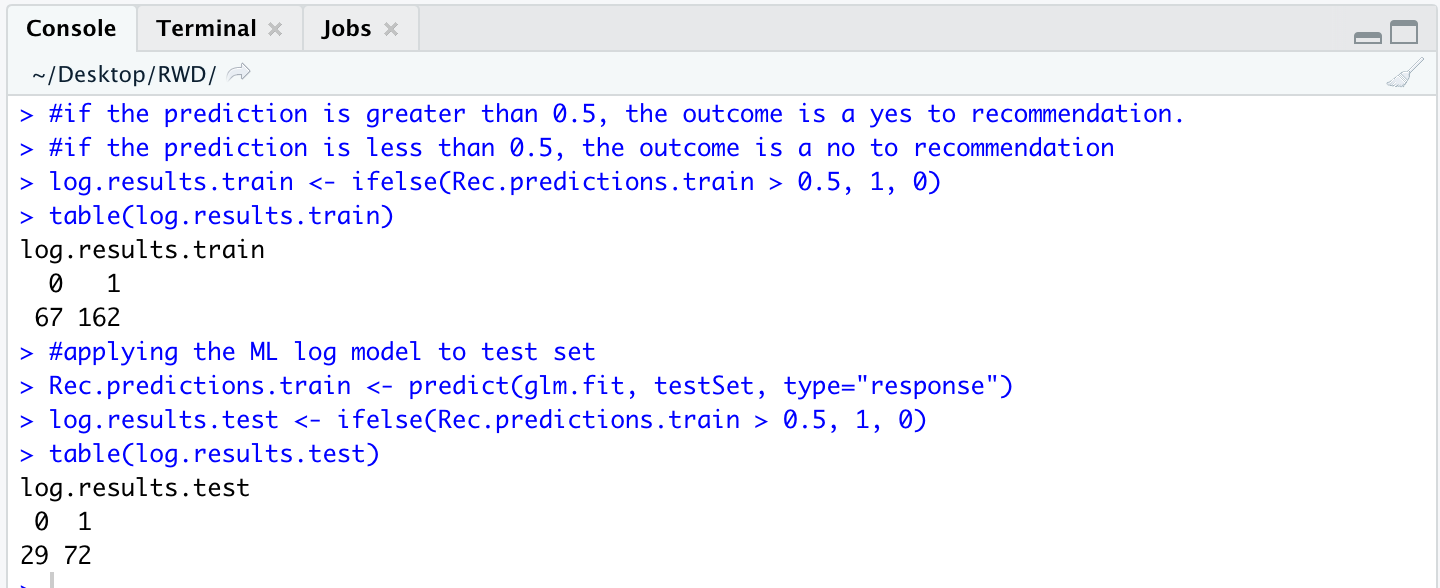
Once the logistic regression is run, the model can be applied to predict function. As shown below, this prediction results in an output that is difficult to read and interpret



As shown above, this prediction results in an output that is difficult to read and interpret. To combat this, I will create a table that orders the predictions as a yes or no to the recommendation. If the prediction is greater than 0.5, the prediction will be ordered in the 1 or “Yes” category. If the prediction is less than 0.5, the outcome will be ordered in the 0 or “No” category.



Once the model is ran through my training set, I applied it to the test set



As shown above, the model has predicted a total of 234 burrito restaurants that the reviewers would recommend and 96 establishments that the reviewers would not recommend