**ECON 386 Group Project: Walmart Sales Data**

Ryan Haugh, Tati Acosta, Eric Boose, William Harrington, Luis Romero

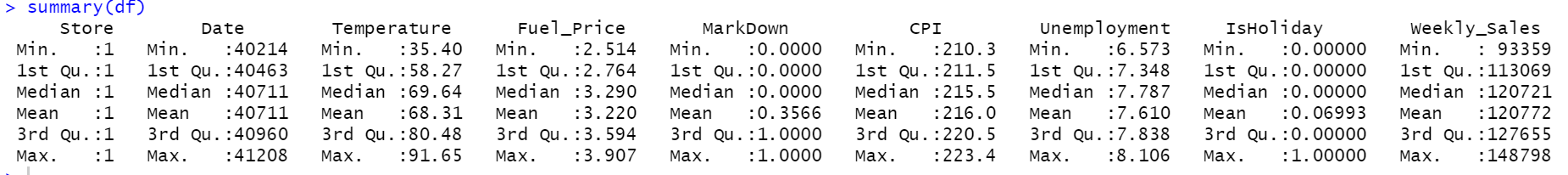
***Executive Summary***

Our data was provided to us from the Walmart corporation via a corporate recruiting competition published on the data science website Kaggle. The data provided historical sales data for 45 Walmart stores and their departments, and this data was collected from February 2010 through November 2012. For learning purposes we decided to focus on one department within one store over the time period.To create the dataset required to isolate one store and one department we used MS Excel as we received the data as Excel CSV files.

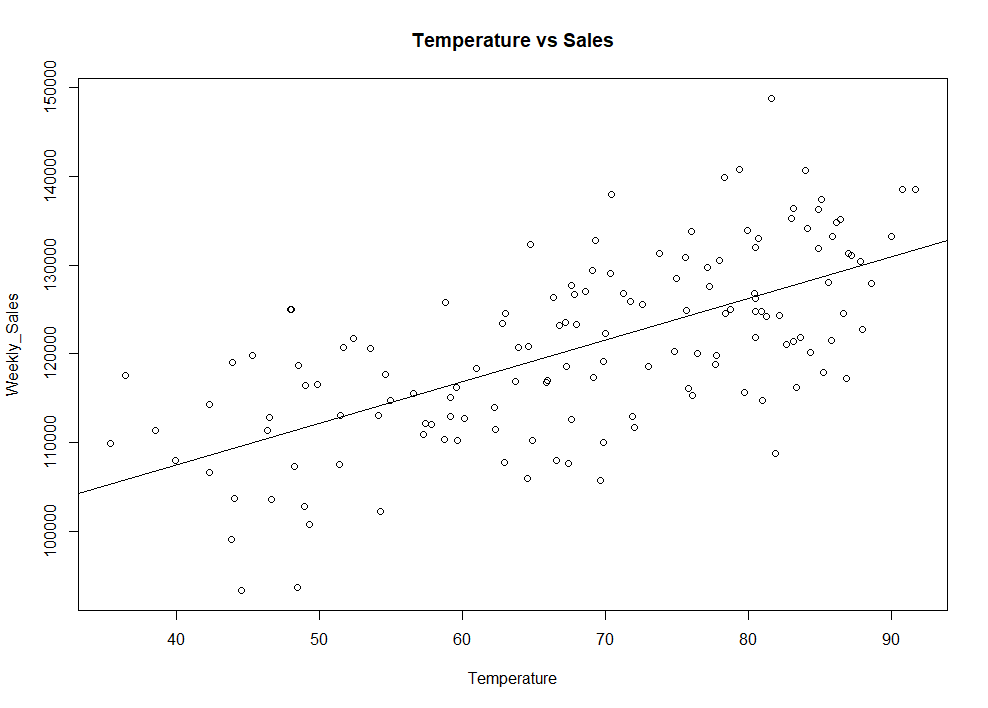
The data provided include the following variables: Weekly Sales, Temperature, Fuel Price, Markdown, Holiday, CPI, Unemployment Rate. Our goal as a group was to use the provided variables to see if we could predict the Weekly Sales values in our linear regressions and predict if there was or was not a weekly markdown in our logistic regressions. The data was provided via 5 Excel CSV files which needed to be stitched together to allow for our stated problem.

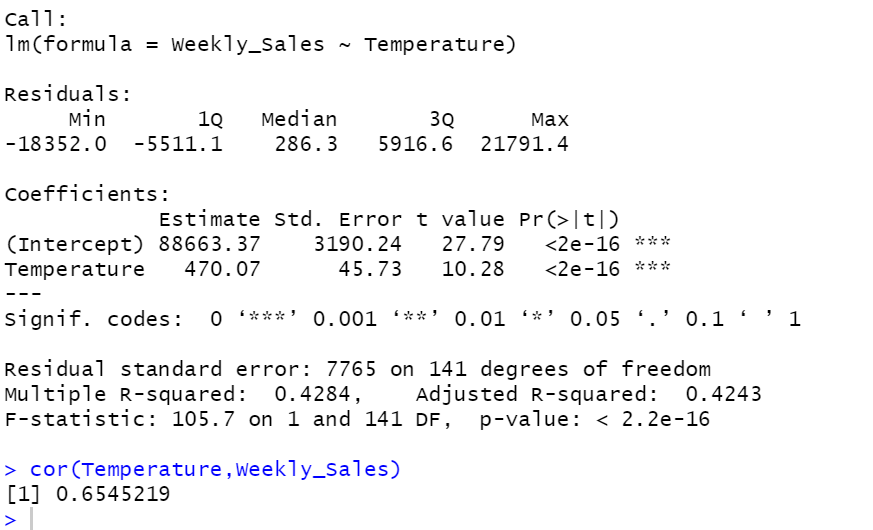
***Preprocessing Data***

After downloading all five CSV files from Kaggle, we used MS Excel to format a file that would easily import to R Studio in TIDY format. Once imported to R studio we began the process of exploratory analysis and developing hypotheses in regards to our models. Below you will find the summary statistics of our TIDY data.



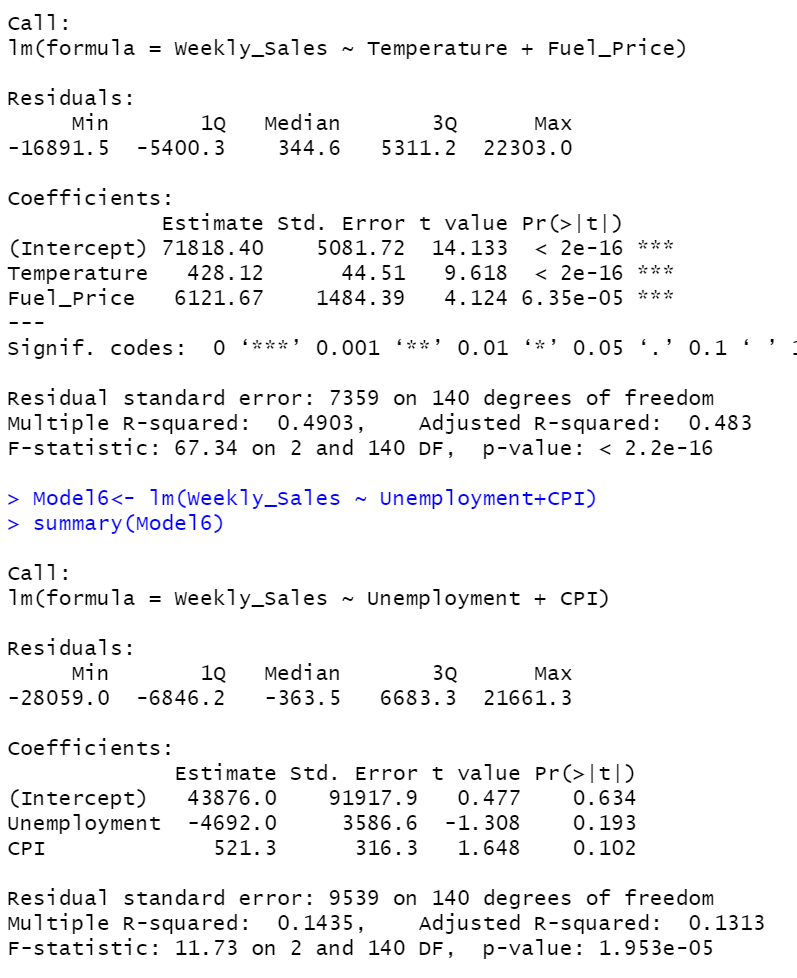
After observing our overall dataset we began to create simple linear regressions to observe the relationships with our “y-variable” which was Weekly\_Sales. Common practice to observe these relationships was the use of simple linear regression models of two variables, scatter plots with the AB line overlays, and observing the correlation between the two variables using the “cor” function in R studio. Below you will find a scatter plot, summary statistics, and correlation matrix values for how Temperature interacts with Weekly\_Sales.





Our goal during exploratory analysis was to discover which variables would be statistically significant in helping to explain the variation in Weekly Sales at the store in question. You can see in the summary statistics above that in our data the weekly average Temperature can explain 42.43% of the variation in weekly sales and has a correlation value of 0.655 suggesting a high positive correlation.

These interactions were performed for Temperature, Fuel Price, CPI, and Unemployment rate with varying degrees of correlation and variation. A theme that stood out to the group was whether or not microeconomic factors such as the temperature or fuel price would provide greater insight into weekly sales than the macroeconomic variables such as unemployment rate and CPI would provide. Below you will find the summary statistics for two regressions testing this curiosity.

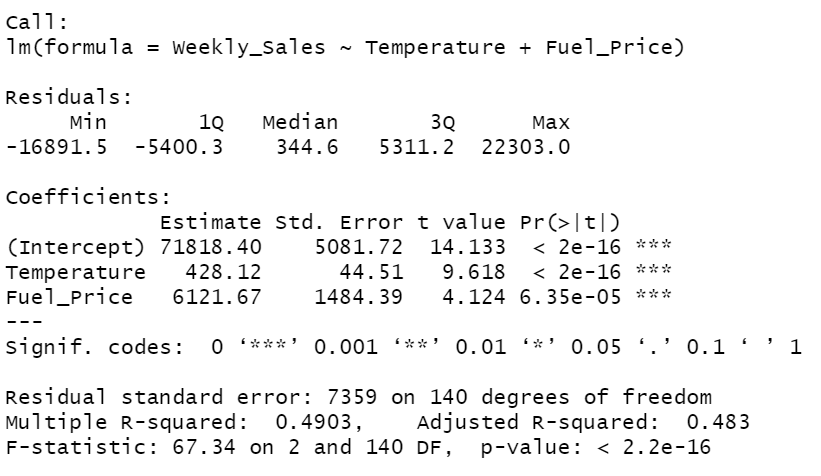


***Linear Regression Model Proposals***

Below you will find the linear models proposed by each individual group member.

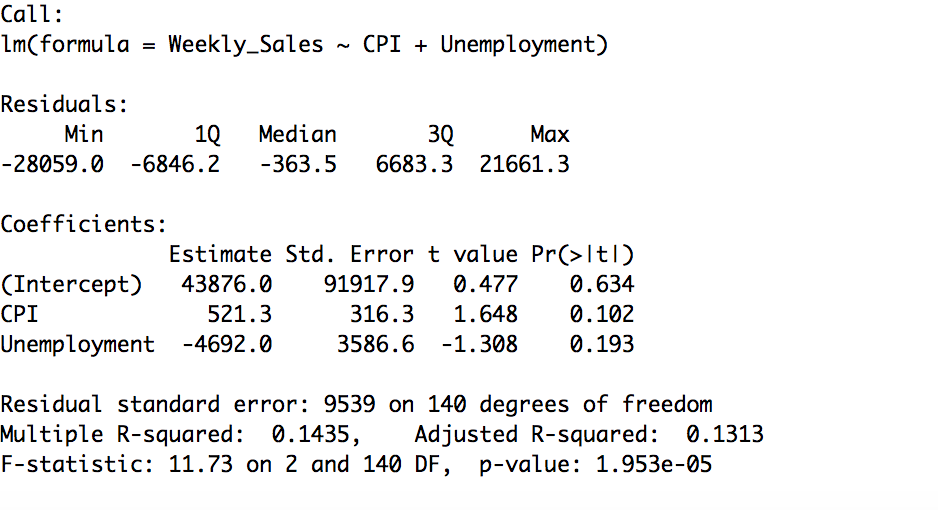
Ryan Haugh

Model3<- lm(Weekly\_Sales ~ Temperature+Fuel\_Price)

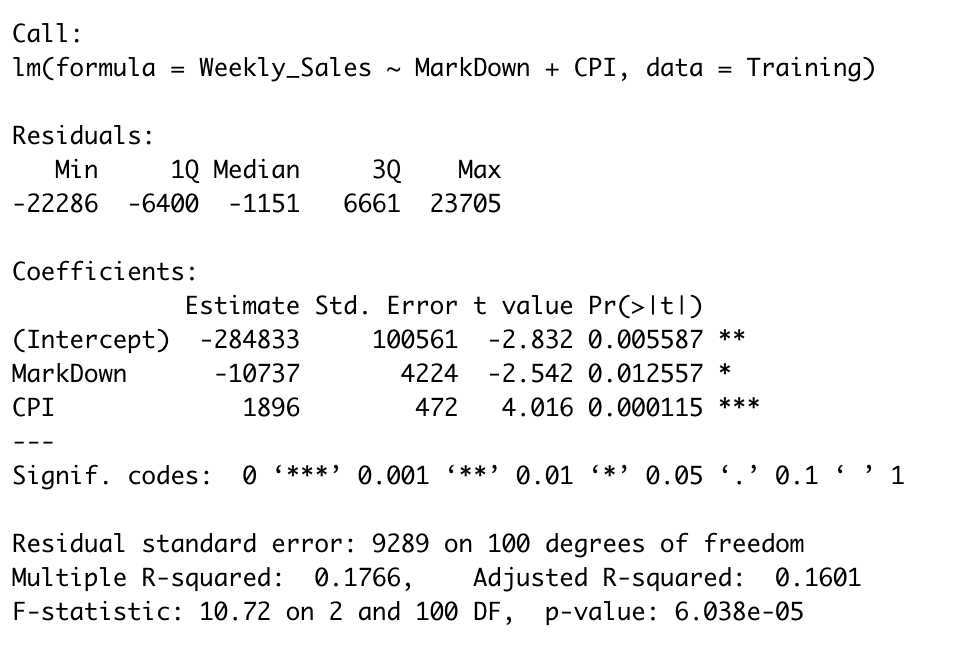


Tati Acosta

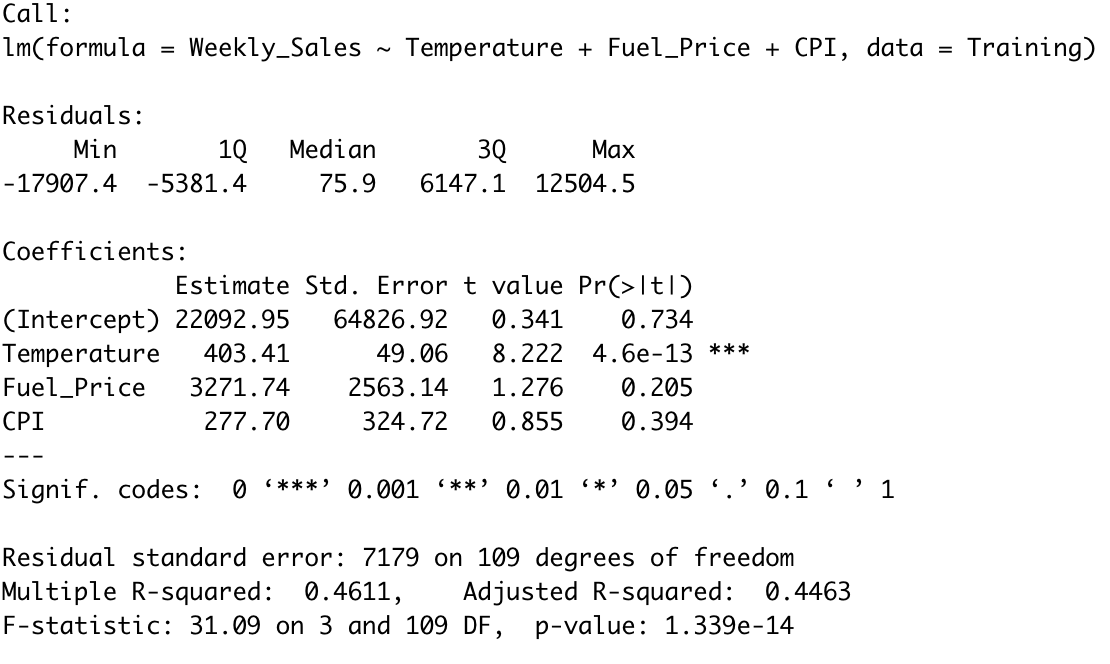
Model3<- lm(Weekly\_Sales ~ CPI+Unemployment)

Eric Boose

M\_Linear<-lm(Weekly\_Sales~MarkDown+CPI)

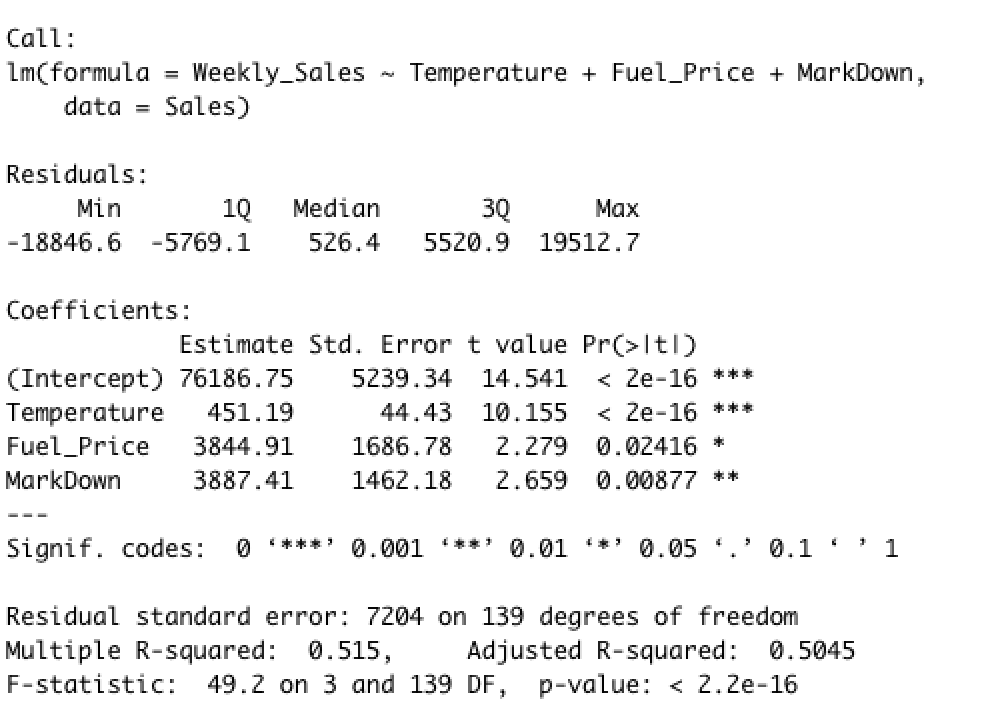


William Harrington



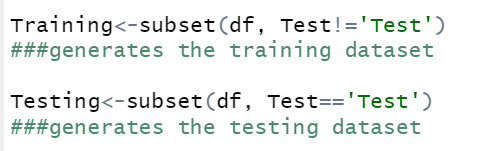
Luis Romero

M5 ← lm(Weekly\_Sales~Temperature + Fuel\_Price + MarkDown, Sales)



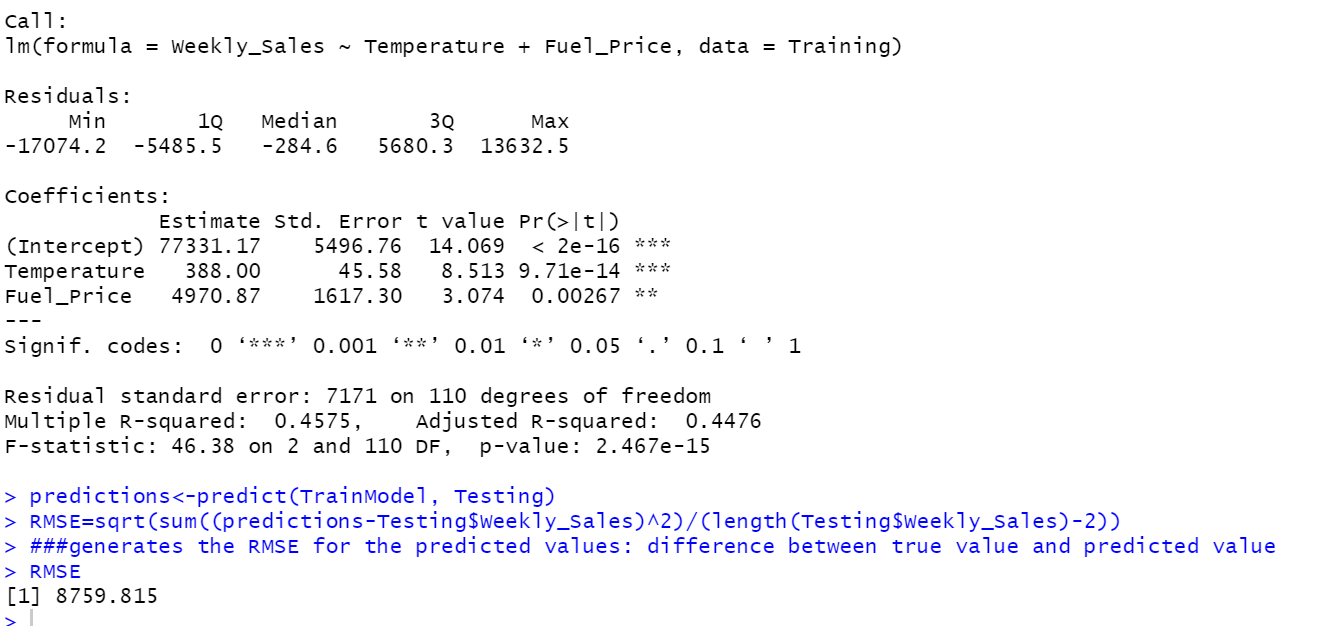
***Regression Model Validation***

Below you will find the code used to partition our testing and training data. Before coding for the partition a random dummy variable was assigned to 20% of the observations to create a 80/20 split in the data for testing and training purposes. After creating that variable, the following code was used to create the two distinct sets.

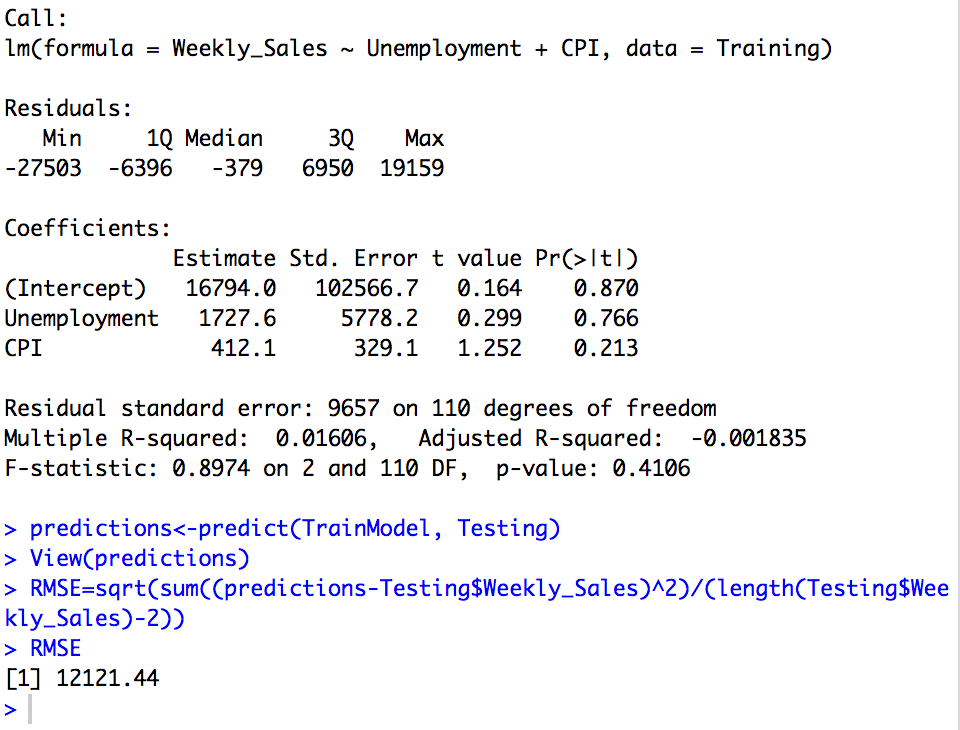


After partitioning the data, we ran our original models using only the training data and developed the summary statistics for the training data. After deriving the standard error of the training model, we generated the predictions for the testing data, and used those predictions to derive the RMSE of those predictions to compare our in sampler error with our new out of sample error. The code for these actions and the results can be found below for all models.

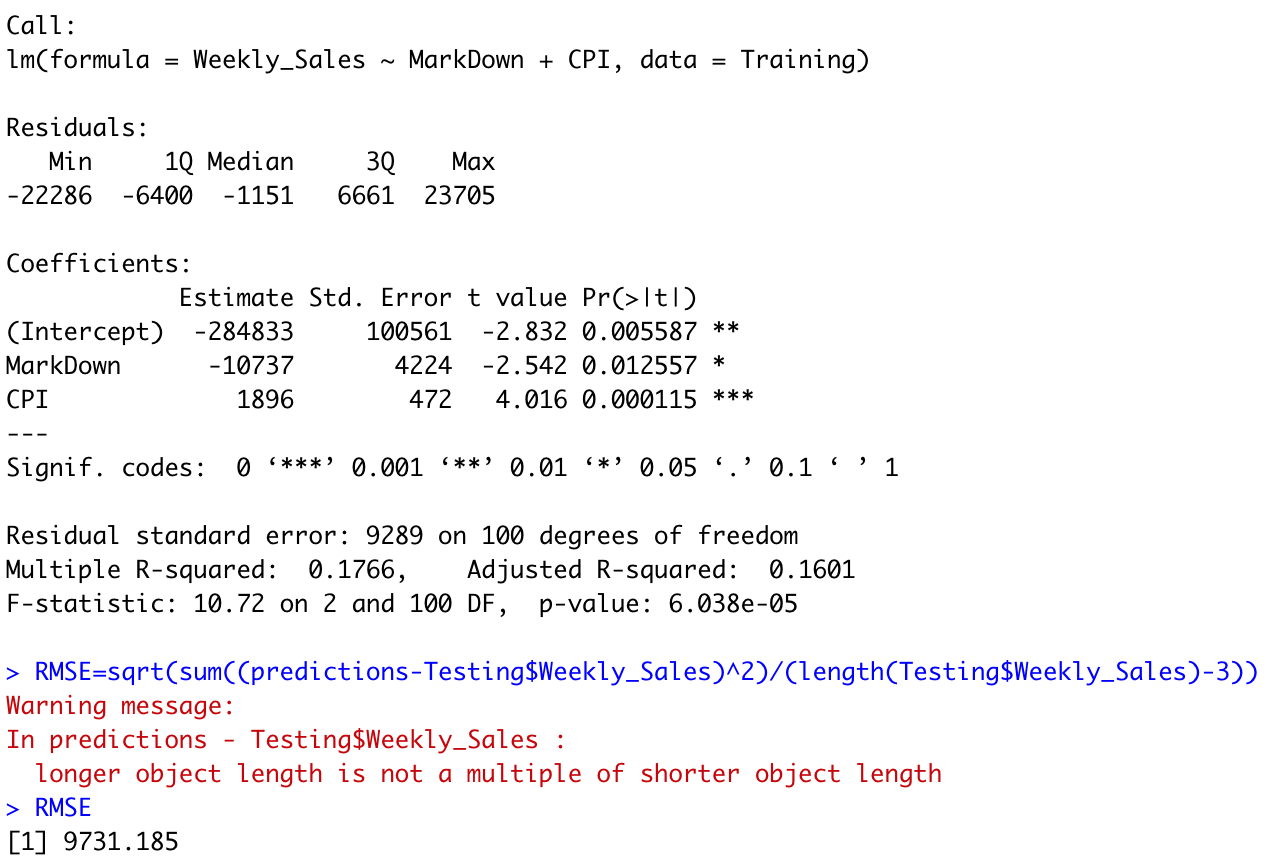
Ryan Haugh



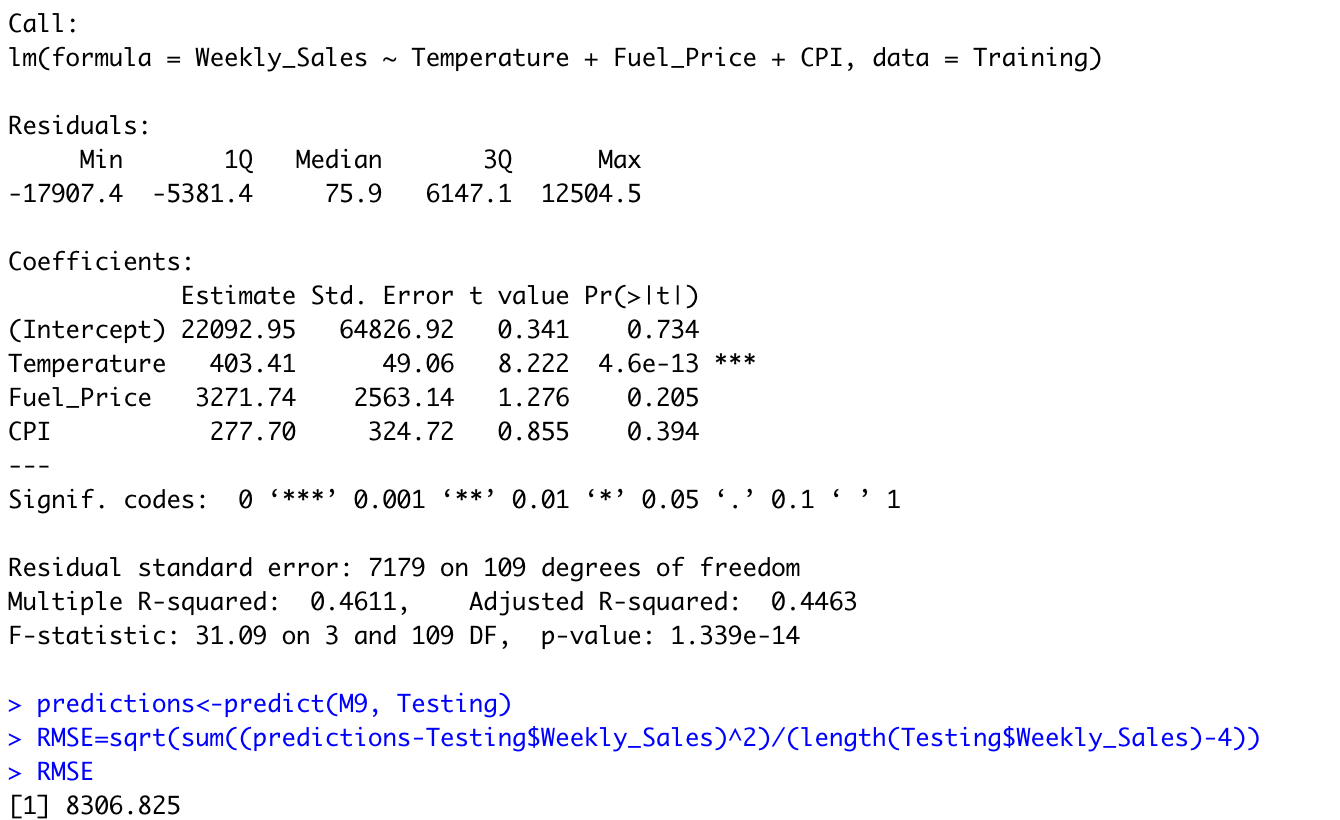
Tati Acosta



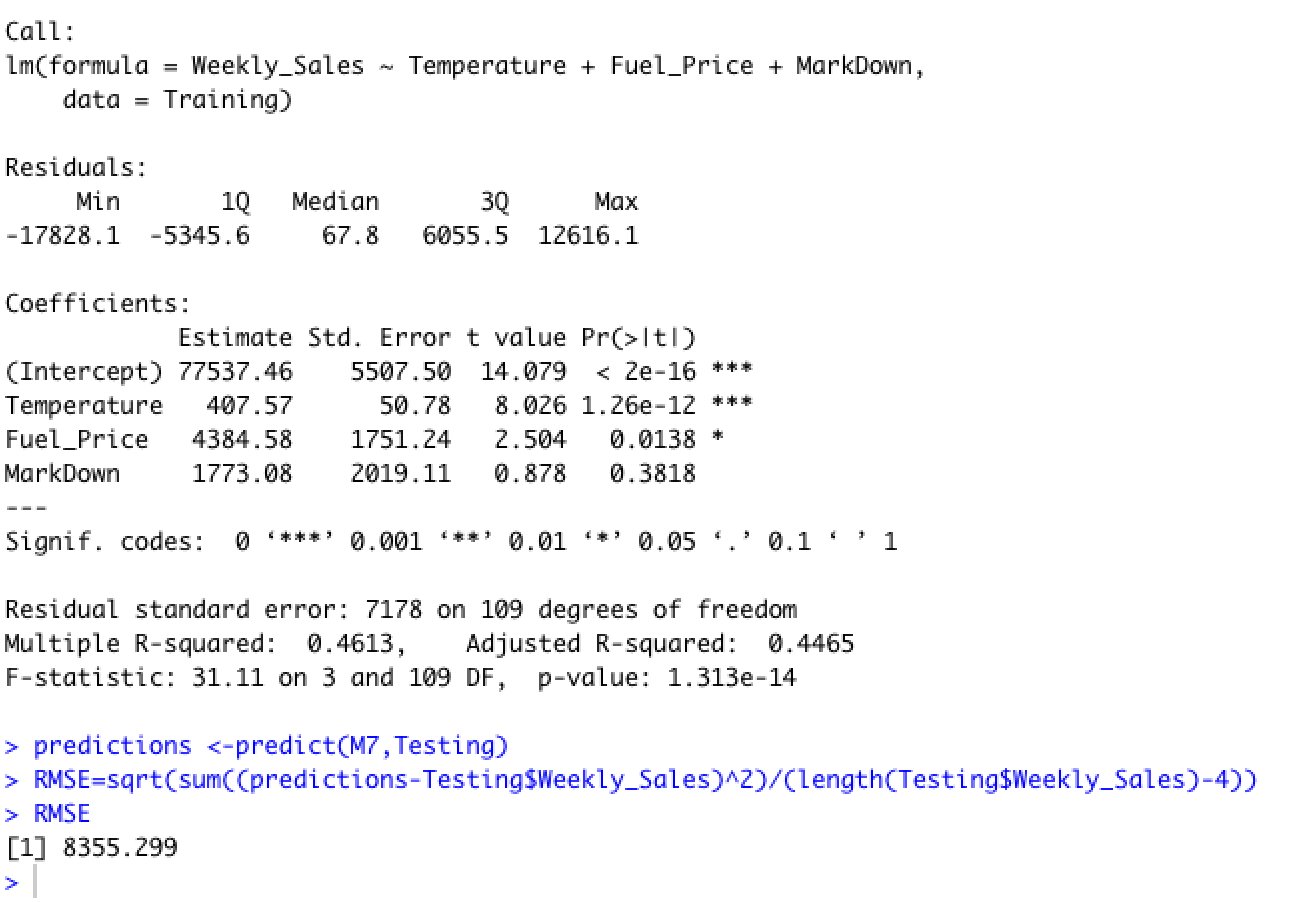
Eric Boose

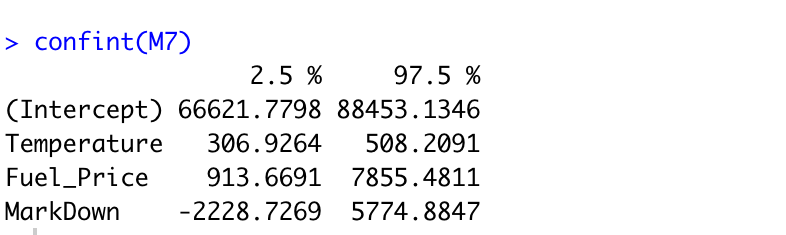


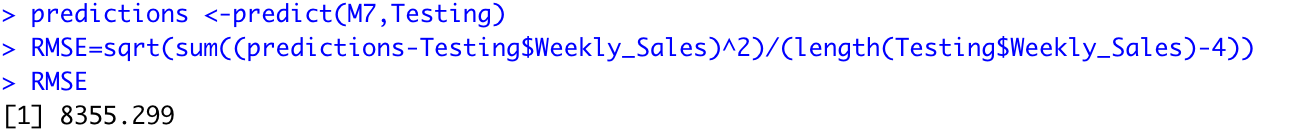
William Harrington



Luis Romero





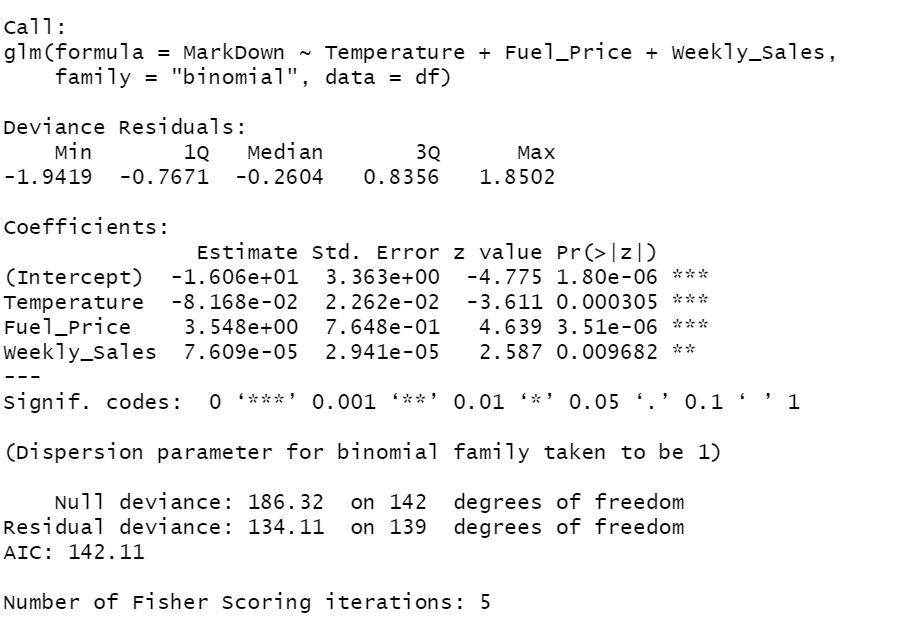


***Classification (Logistic) Model Proposals***

Below you will find the logistic regression models proposed by each individual group member.

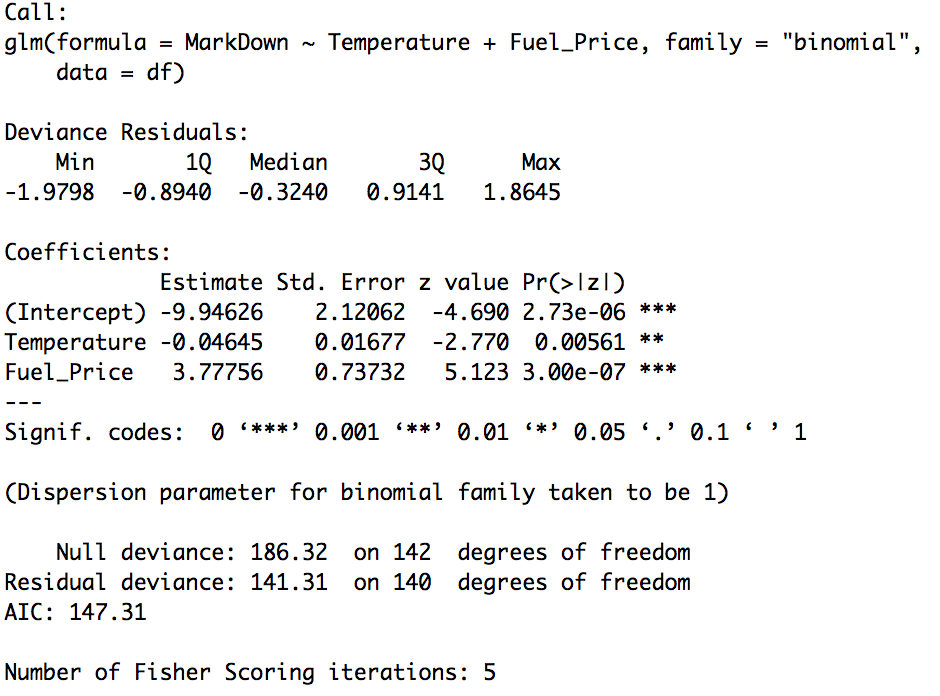
Ryan Haugh

Model1.1<- glm(MarkDown ~ Temperature+Fuel\_Price+Weekly\_Sales, data = df, family = "binomial")



Tati Acosta

Model1.1<- glm(MarkDown ~ Temperature+Fuel\_Price, data = df, family = "binomial")



Eric Boose

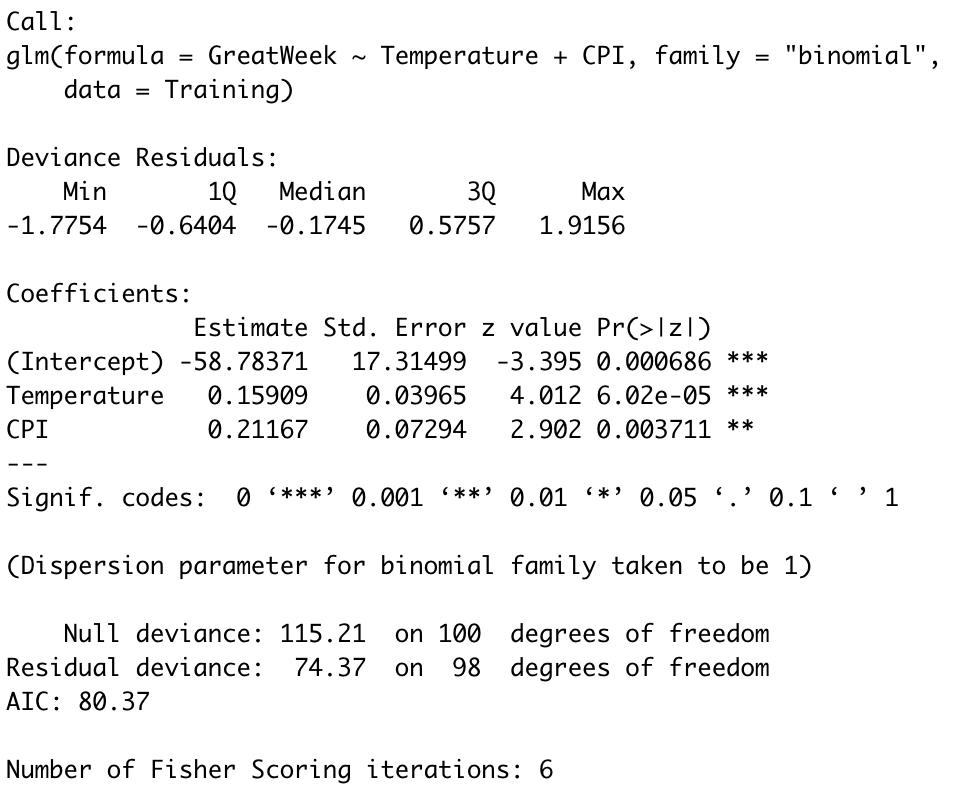
M\_Logistic<-glm(GreatWeek~Temperature+CPI,data=Training,family='binomial')

To create this model, I built a dummy variable to describe the goodness of each week’s sales. If the weekly sales value falls in the 75th percentile or higher, the “Great Week” variable takes the value of 1. If not, the variable has a value of 0. The code to create this dummy variable is as follows:

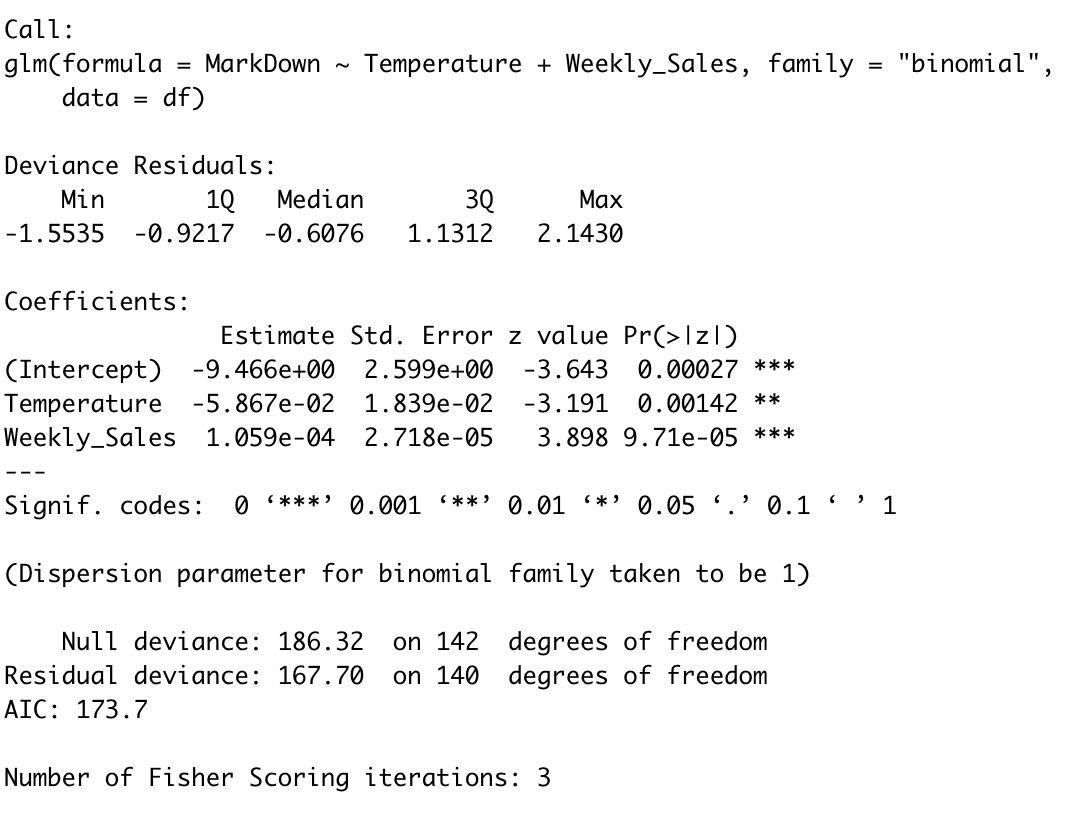
df$GreatWeek[df$Weekly\_Sales>=127654]<-1

df$GreatWeek[df$Weekly\_Sales<127654]<-0

##created dummy variable "Great Week" if Weekly Sales was in 75th percentile or higher

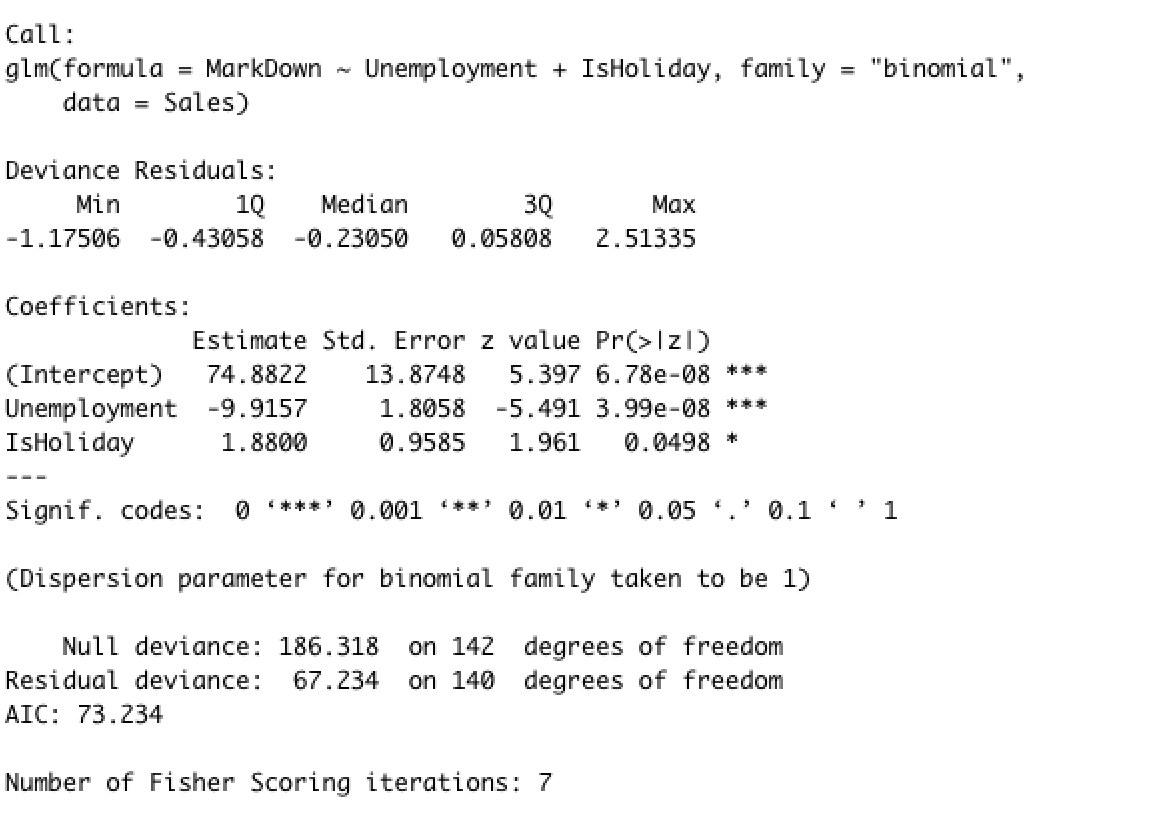


William Harrington



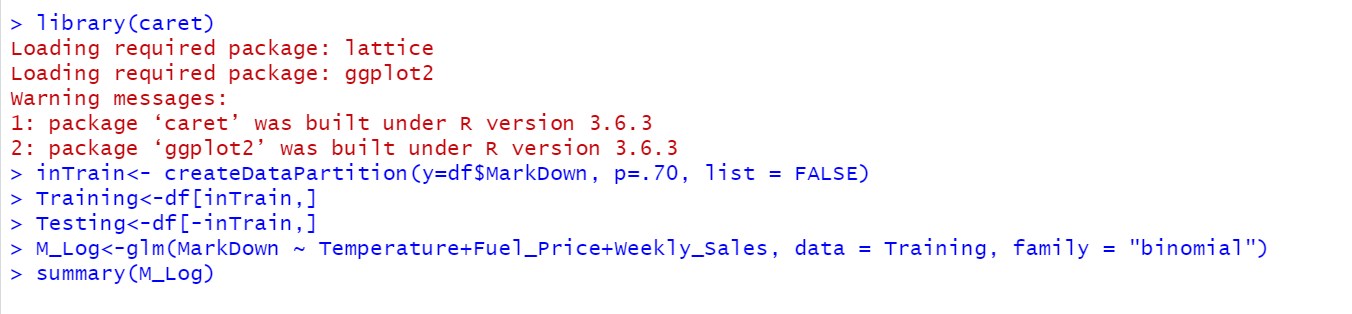
Luis Romero

M\_LOG ← glm(formula = MarkDown ~ Unemployment + IsHoliday, family = “binomial”, data = Sales)

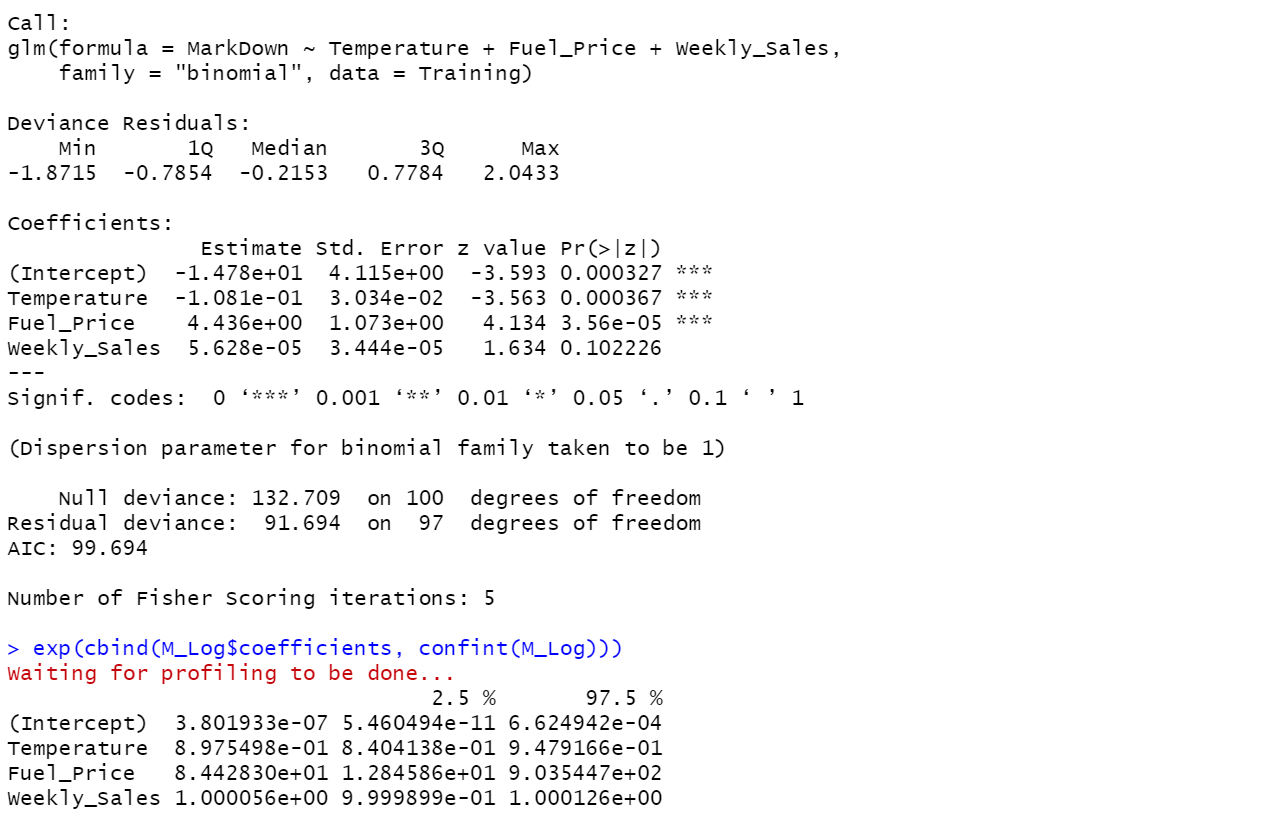


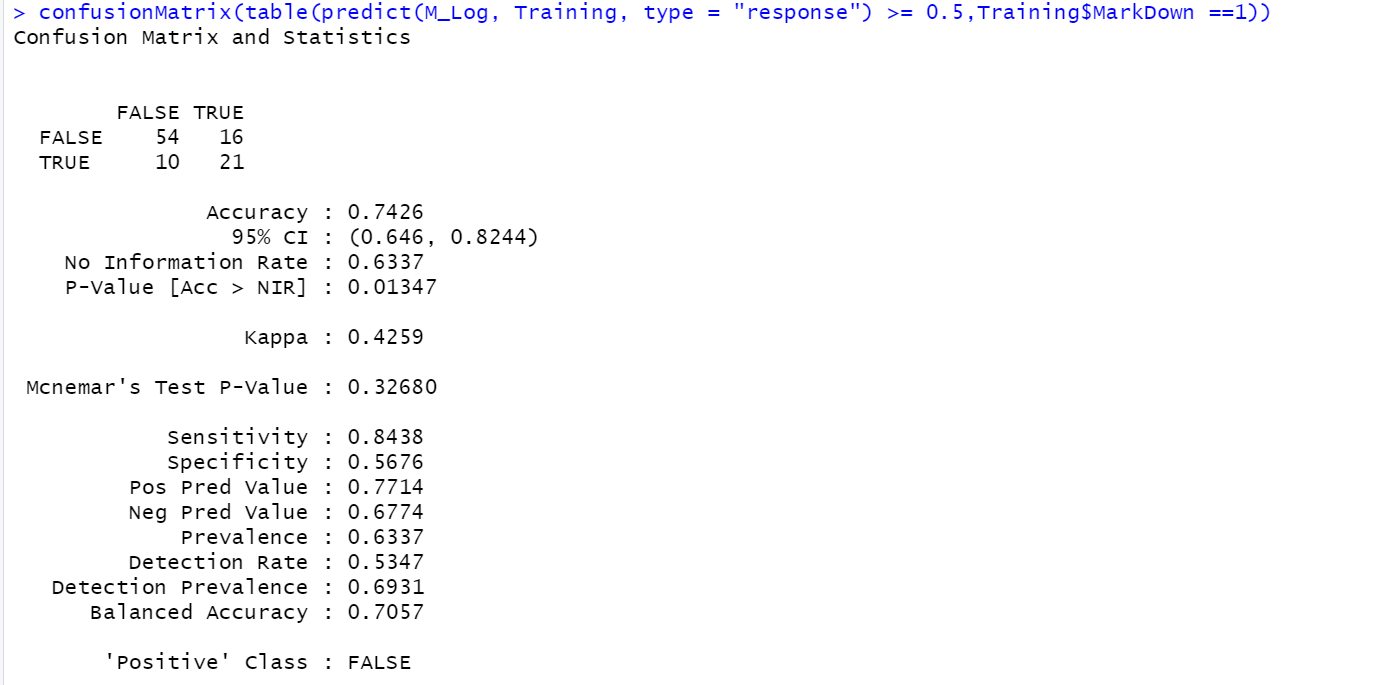
***Classification Model Validation***

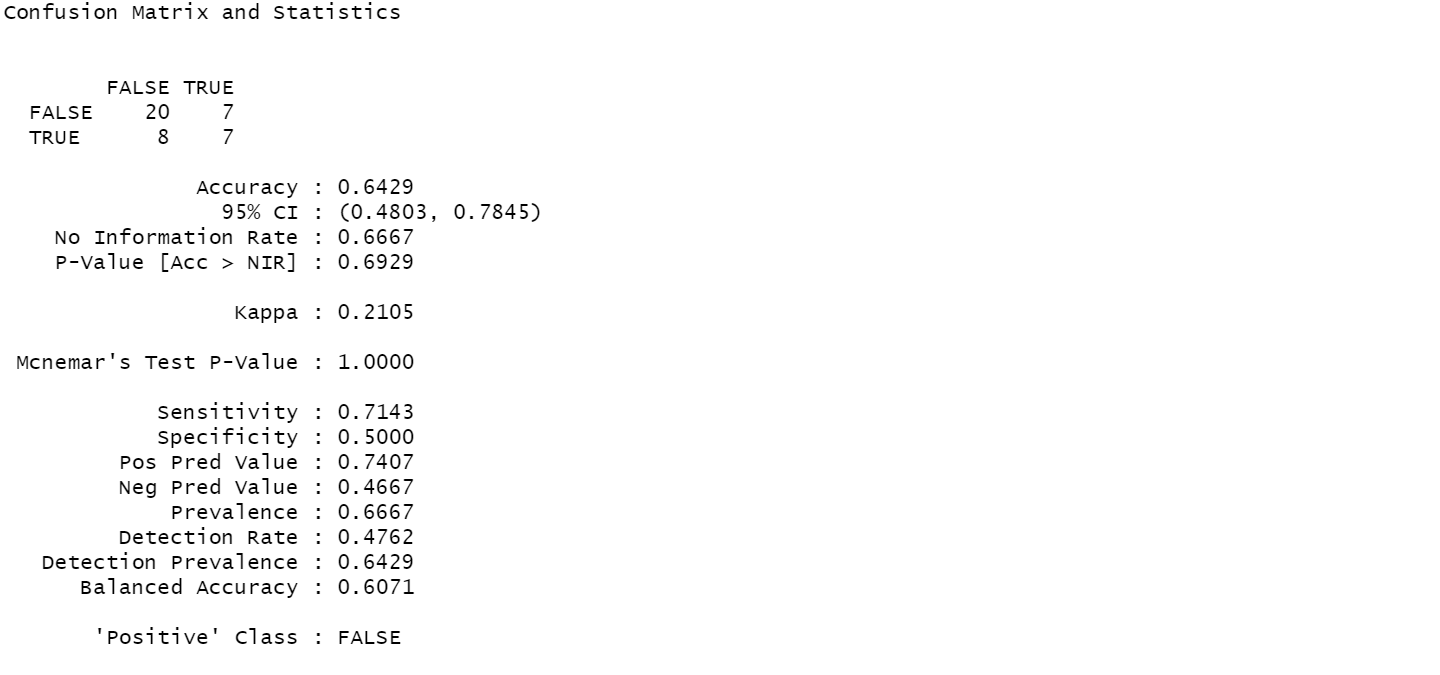
Below you will find the code for how the data was partitioned for training and testing, it was split 70/30 for training and testing respectively.



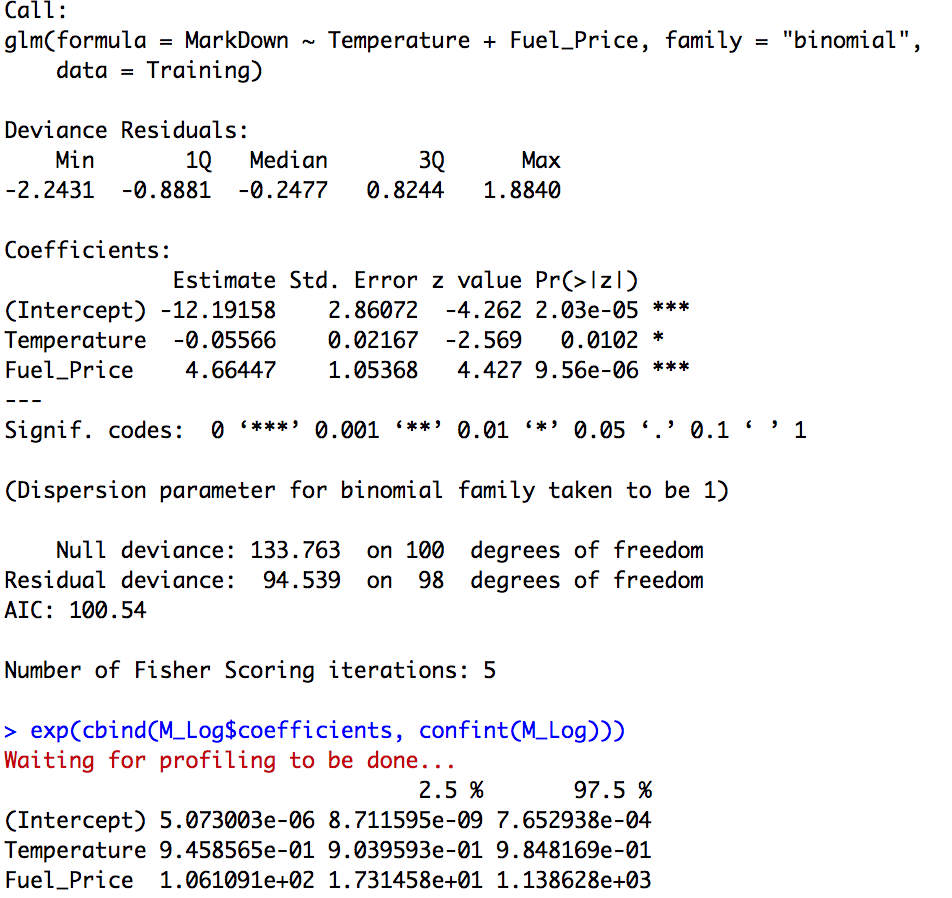
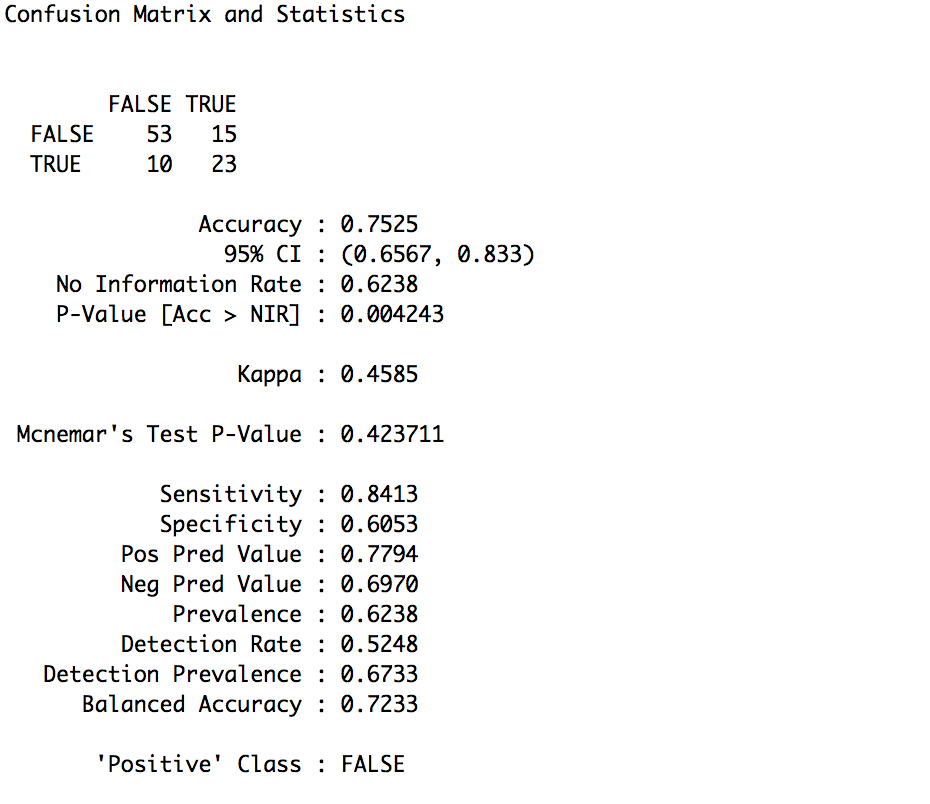
After partitioning the data, we then created new models using the training data, and developed confusion matrices to show its performance on the training and testing datasets. The code and outputs for each individual can be seen below.

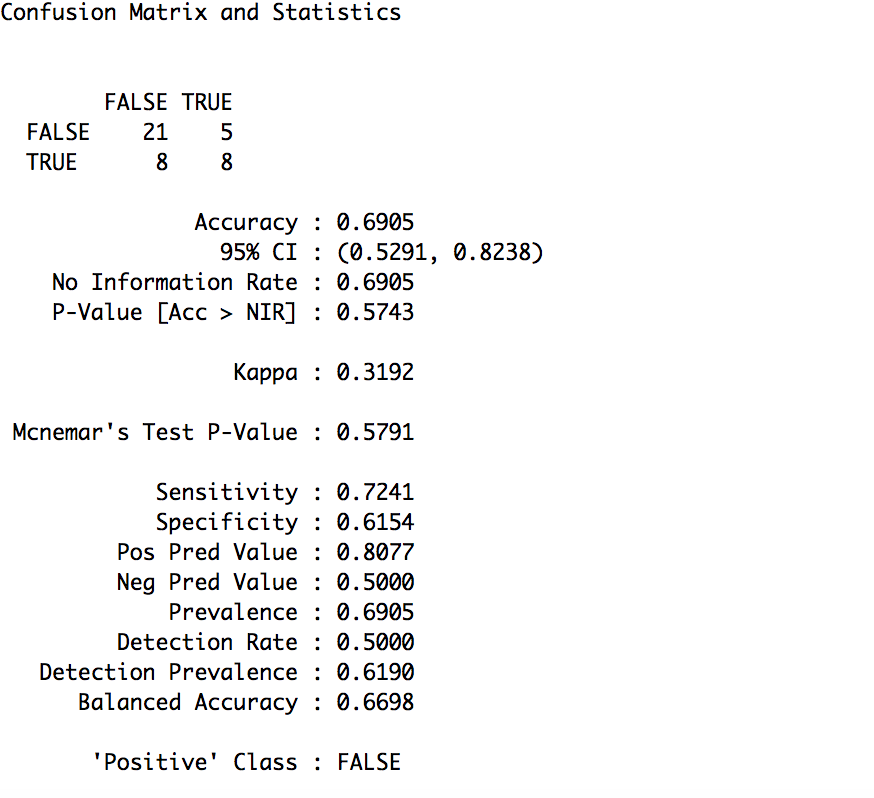
Ryan Haugh

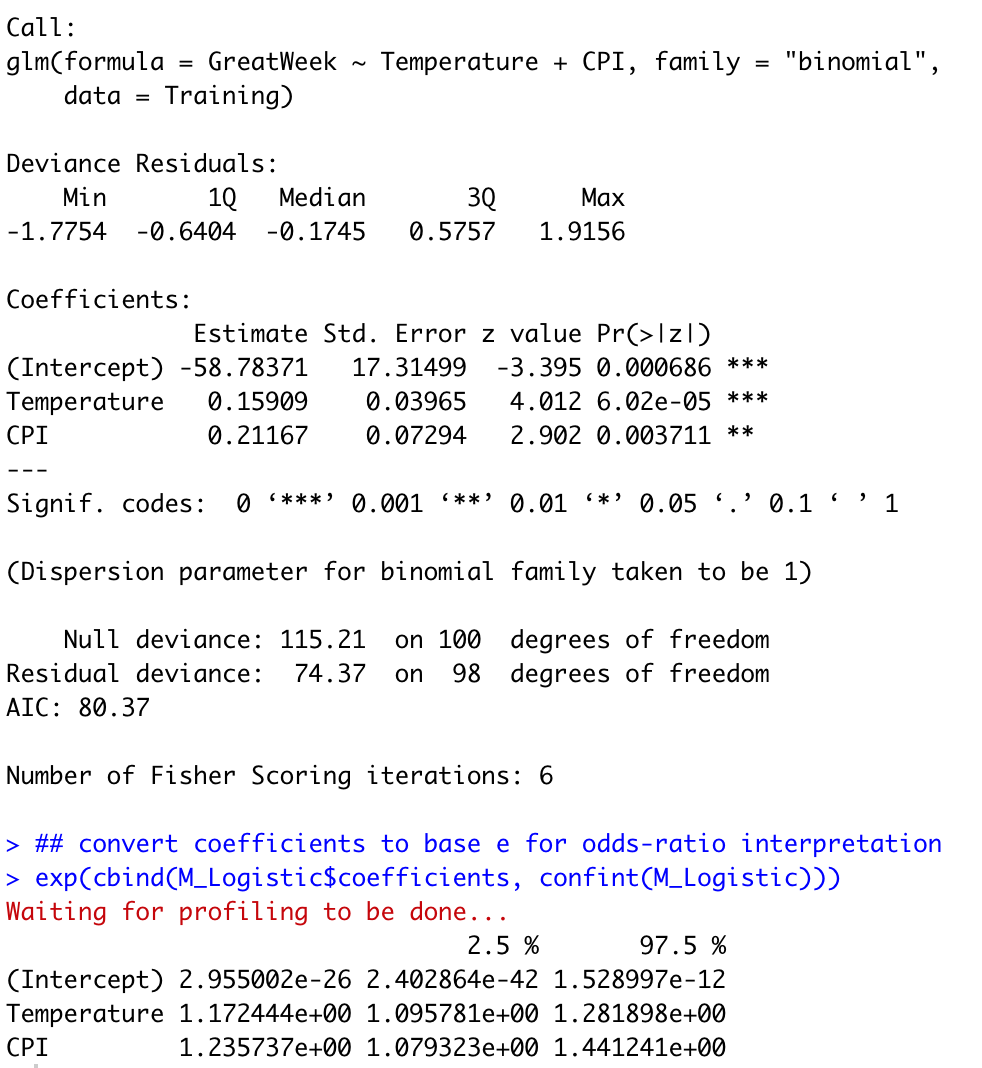


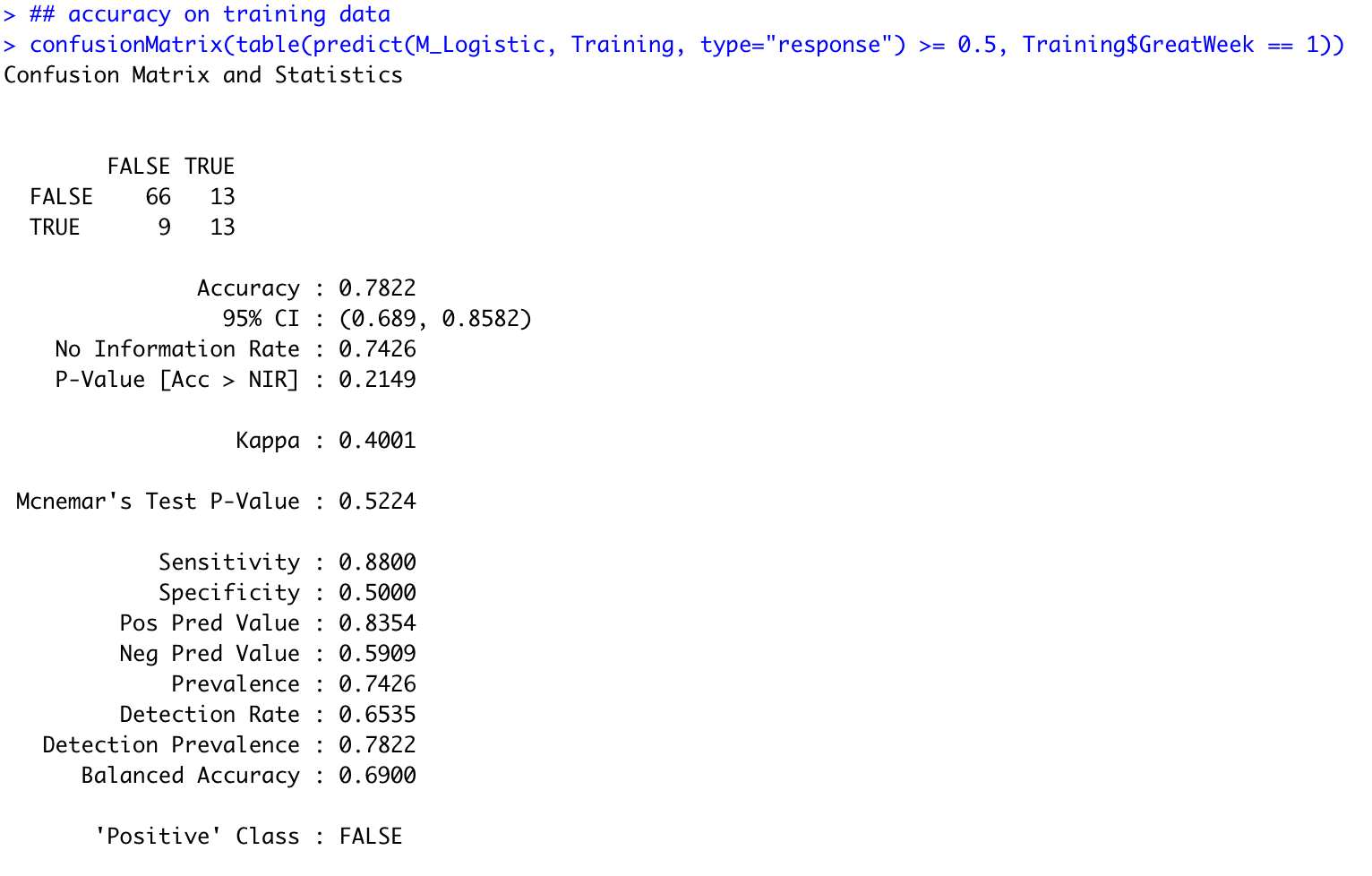


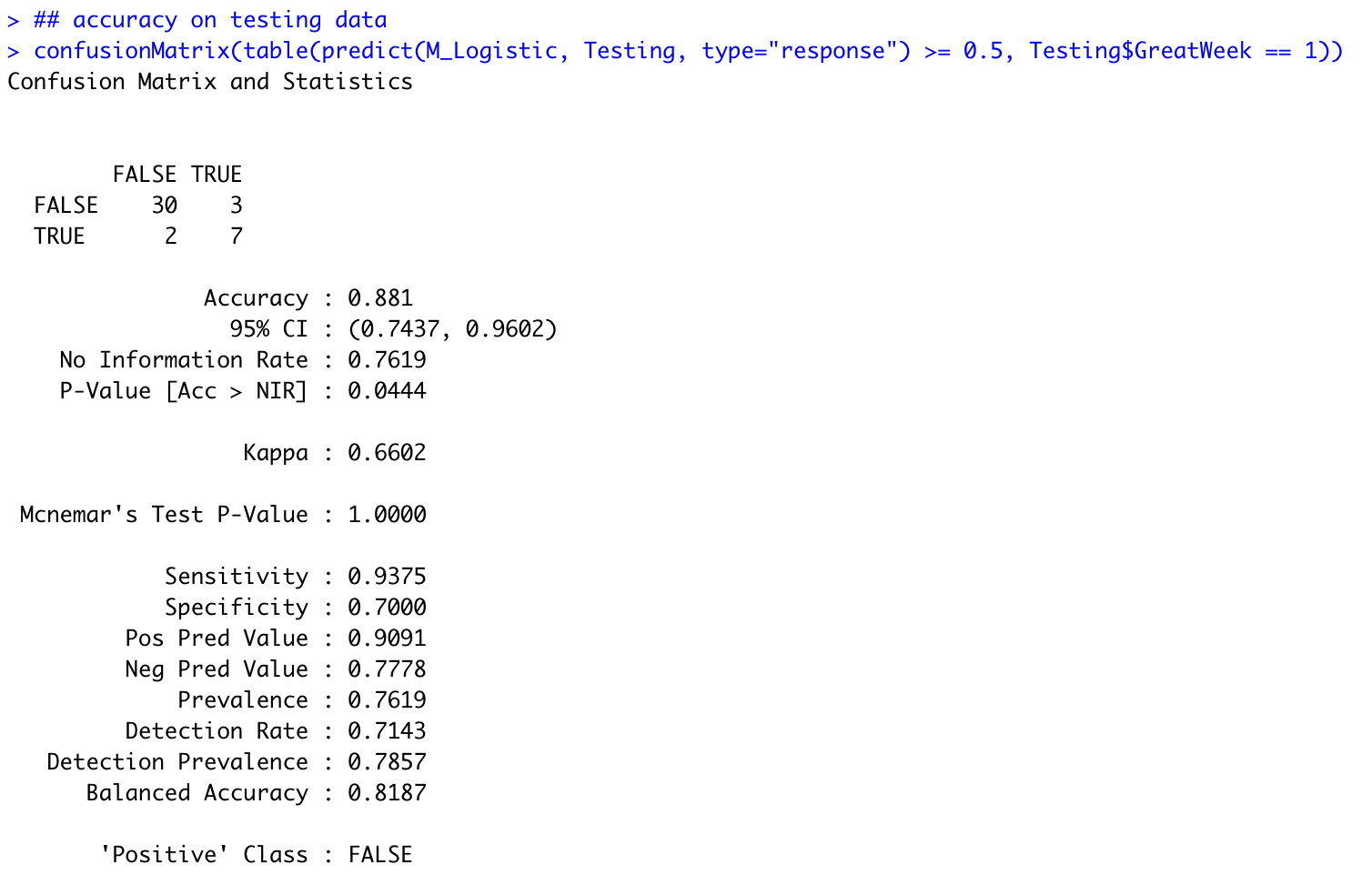
Tati Acosta



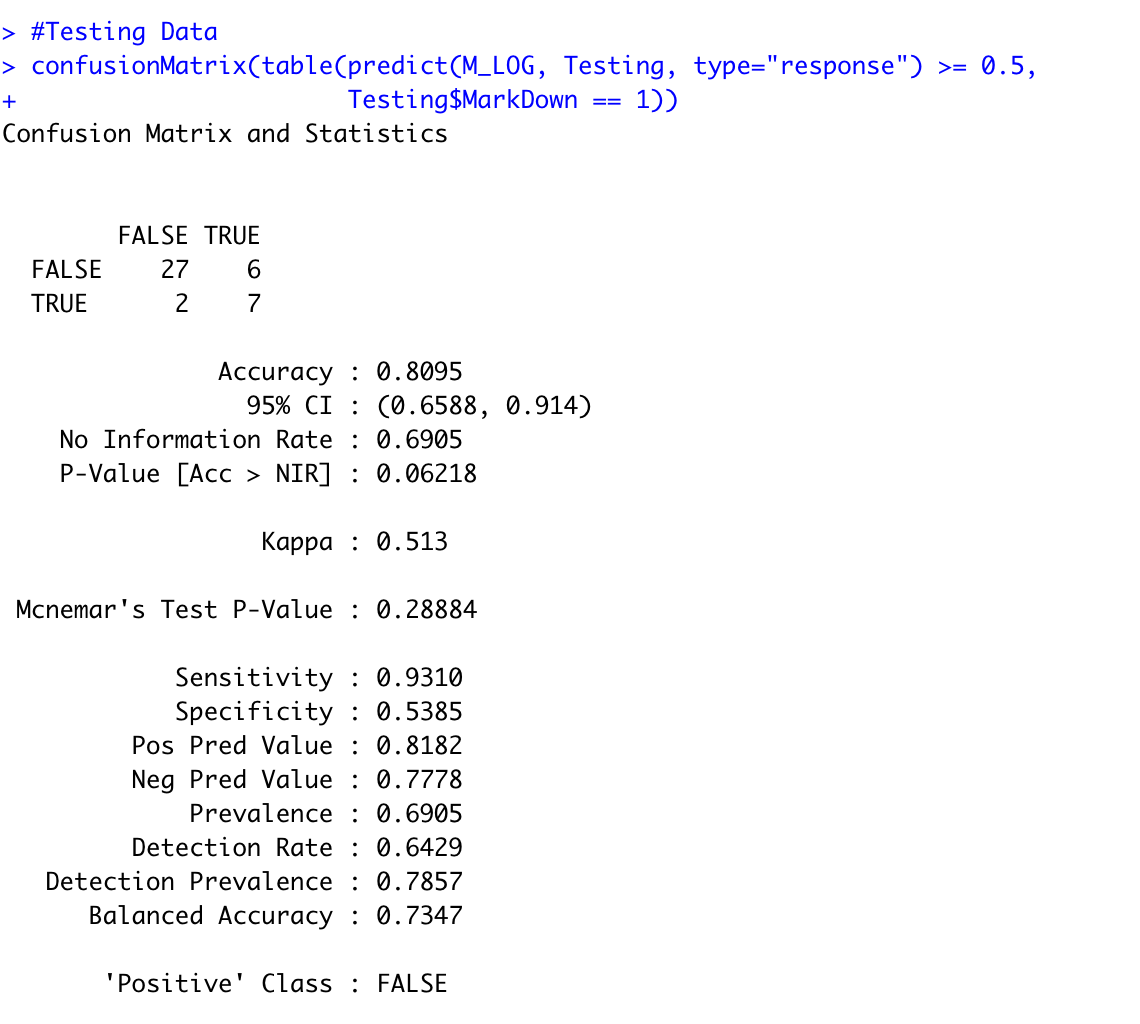
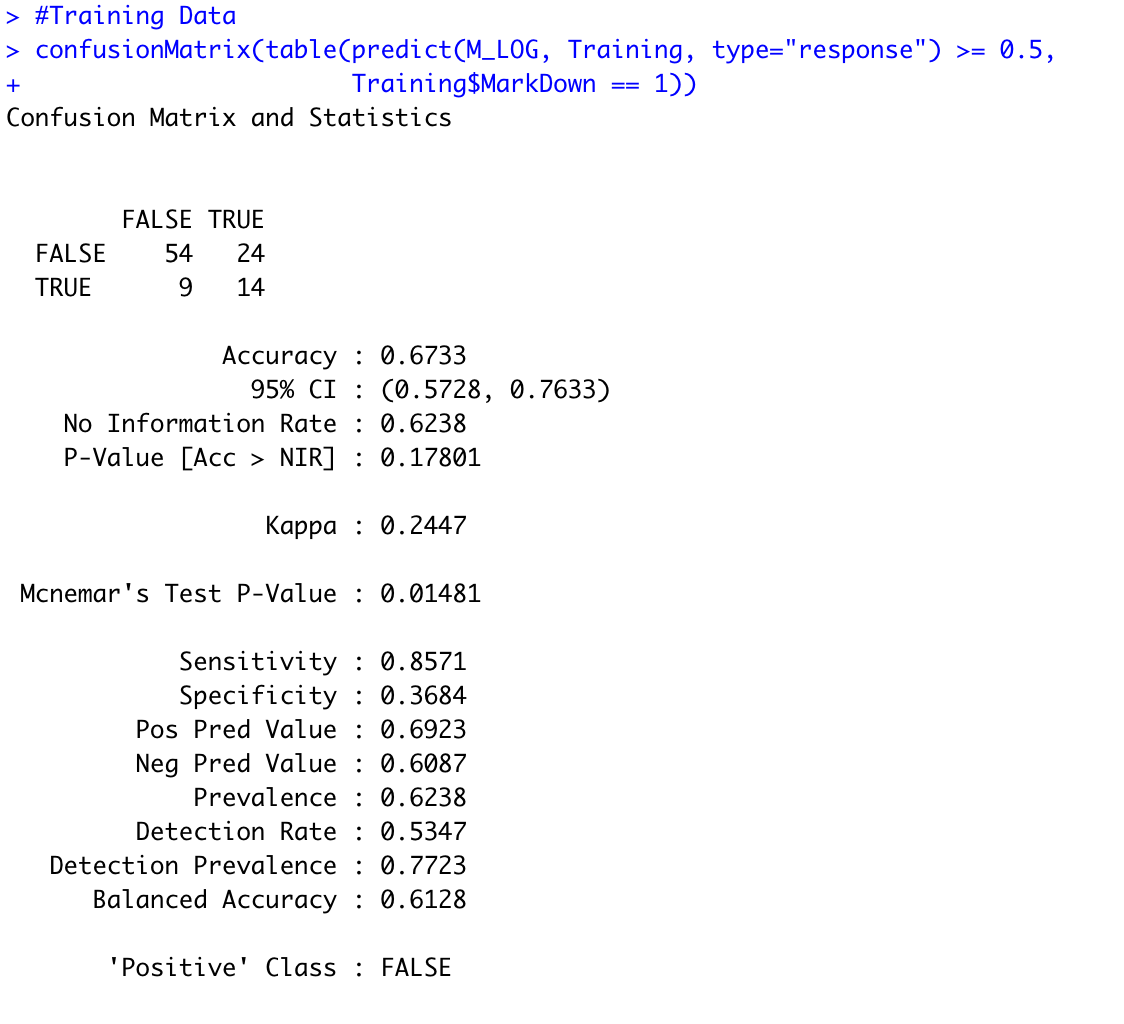
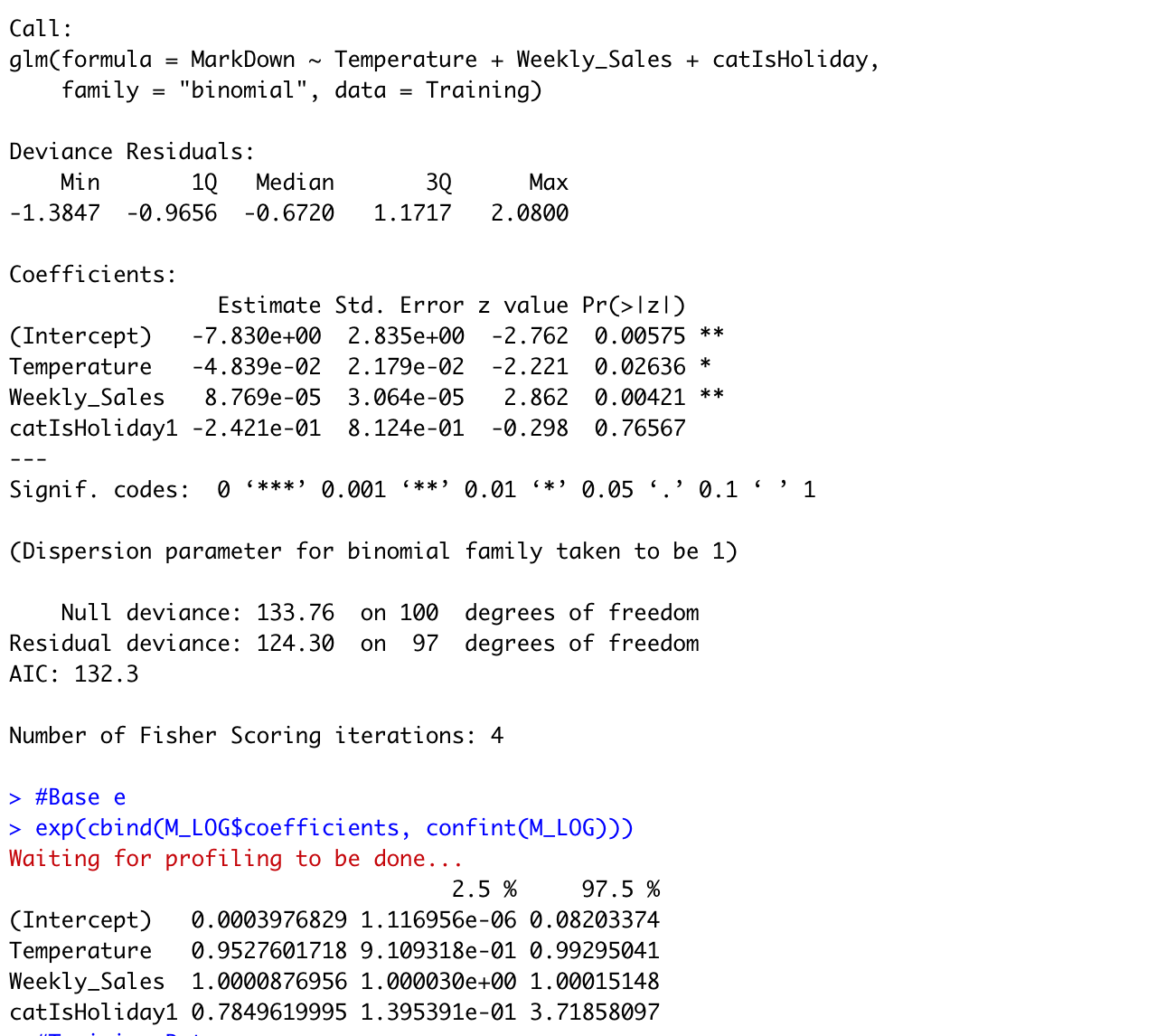
Eric Boose



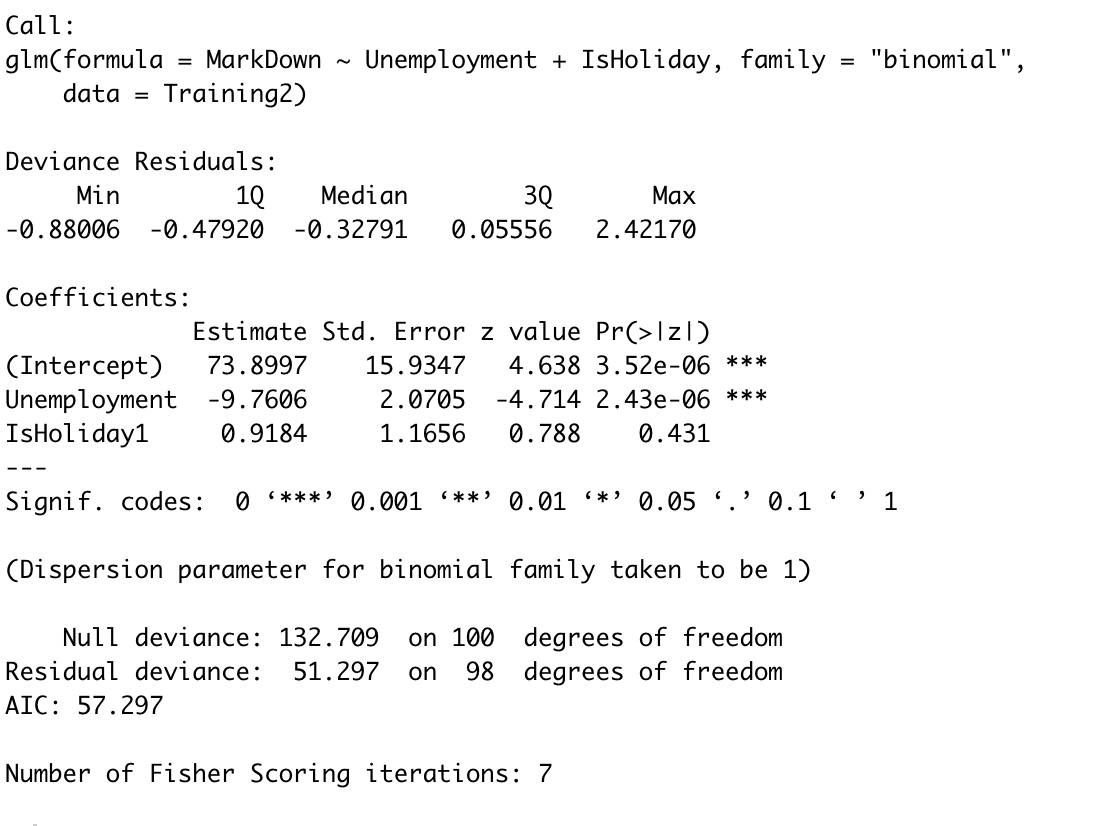


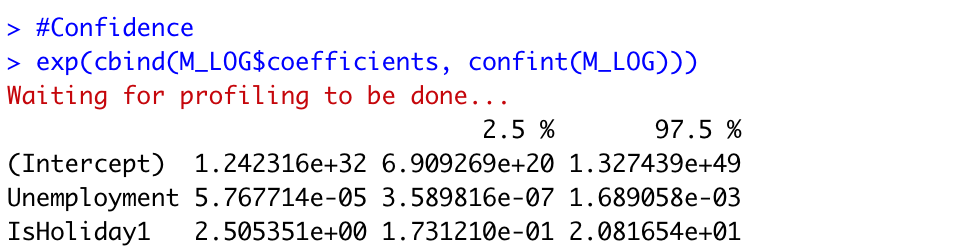


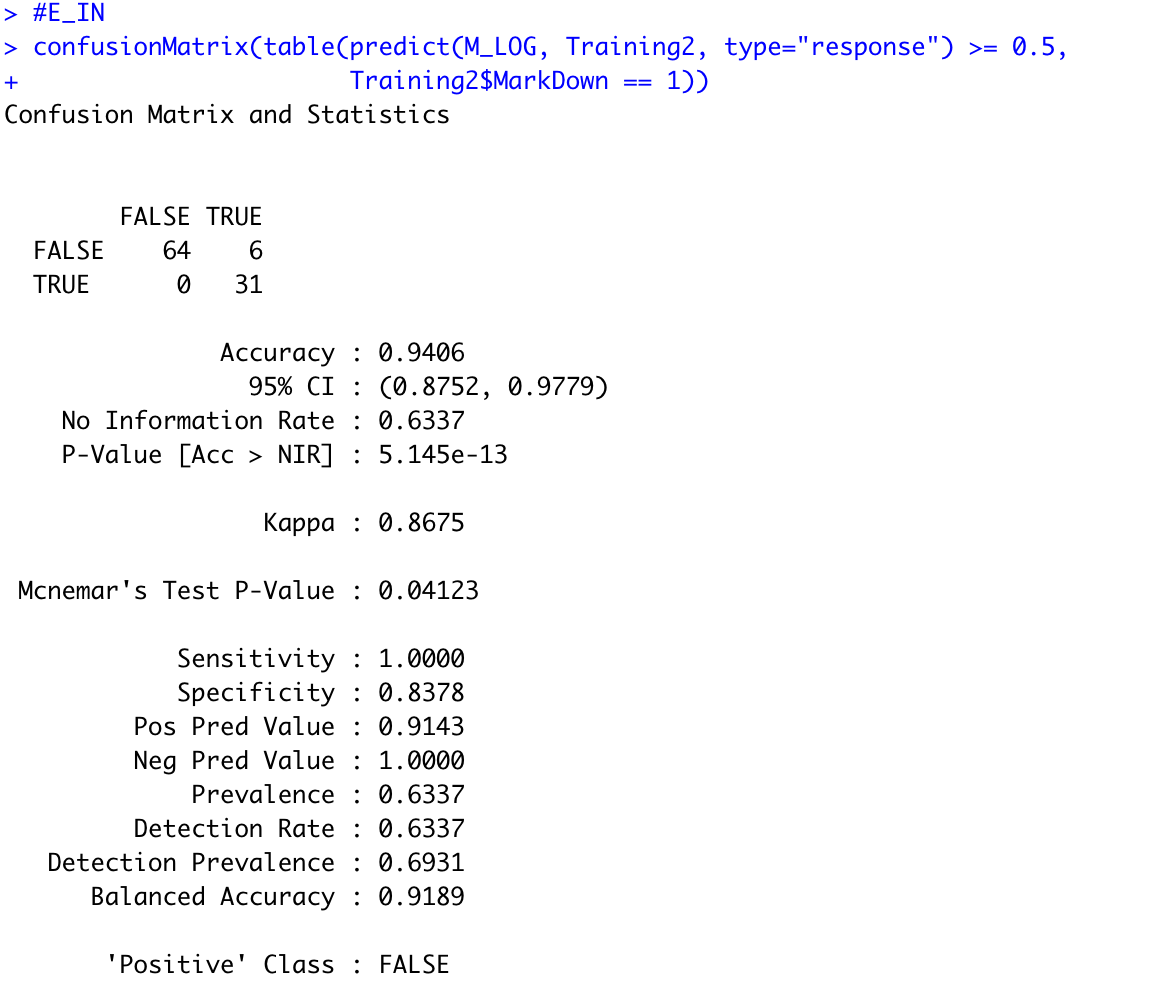
William Harrington

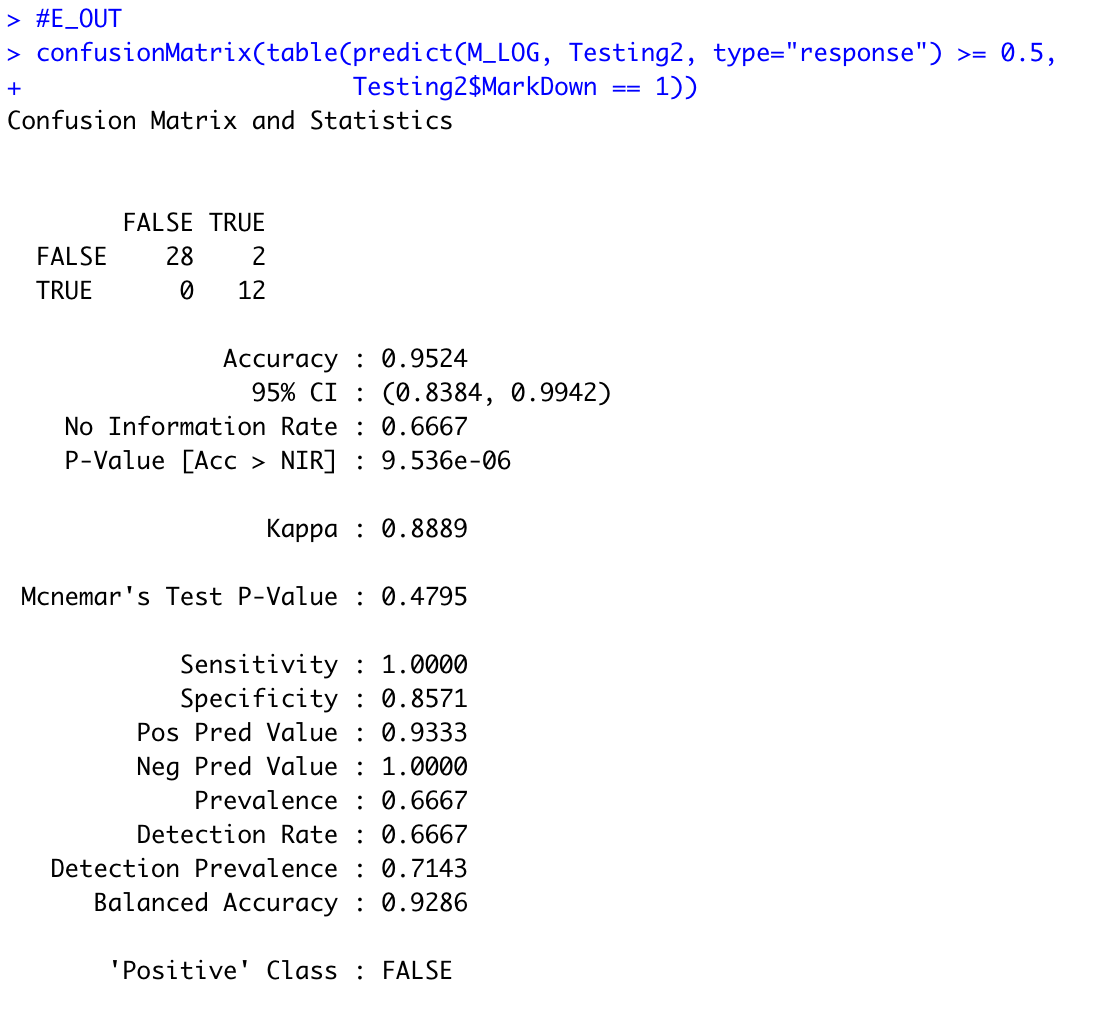


Luis Romero









***Conclusion***

Our most effective models were William’s linear regression model and Luis’ logistic regression. William’s linear regression, which predicts weekly sales using the temperature, fuel price, and CPI variables, explained 44.63% of the variance in the data. More importantly, the model had the lowest out of sample error of the group at 7007. The combination of one of the highest adjusted R-squared values and lowest out of sample error value in the group is a large part of why we believe William’s linear model is the best of the five.

Luis’s logistic regression used the Unemployment and IsHoliday variables to predict whether there was a markdown on a given week. The model has an AIC of 73.234, a training accuracy of 94.06%, and a testing accuracy of 95.24%. We were impressed by the accuracy of Luis’s logistic model, and we believe that it is the best logistic model that our group produced.

While working on this project, we were surprised that temperature seemed to be the best predictor of good weekly sales in the stores. In addition, the holiday and markdown variables were less effective in predicting sales than we expected them to be, and we found them to be the least effective of all the predictive variables.