Irene Huang

Kathleen Gendotti

Black Friday Sale

General Question and Motivation

Through this project, we wish to learn more about the spending behaviors of customers on Black Friday. We want to know which variables affect the sales the most and which model is the best to predict the amount in dollars that customers spend. Additionally, we wanted to gain a better understanding of the characteristics of people who spend a lot of black friday.

The reason we choose this topic is not only because it is an interesting topic but also because it relates to both of our work experience in different industries. By analyzing the data, we will be able to connect the findings with our experience and gain more insights about the study.

Summary Statistic

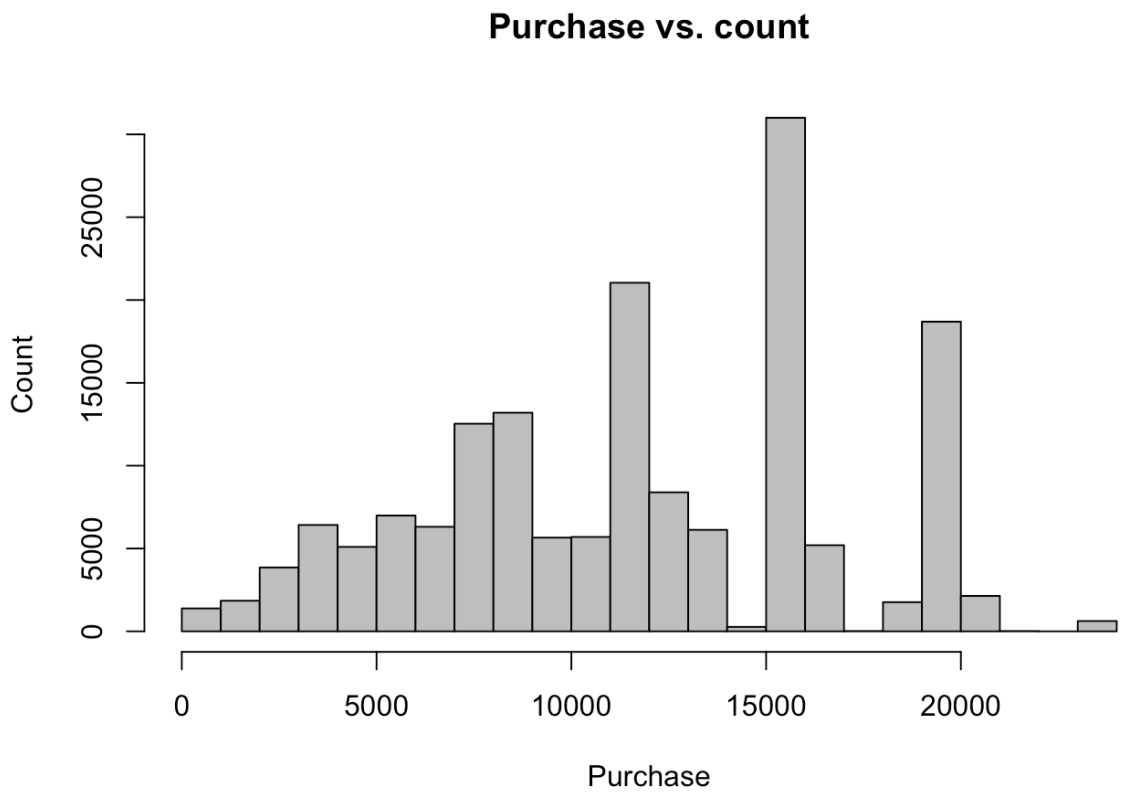
For the analysis, we use a dataset that we found on Kaggle called Black Friday. The data contains 12 variables and 550,000 observations. In order to make the analyzing process easier, we decide to take out variables User_ID and Product_ID because the numbers do not actually have meanings to use for this analysis and we are not able to know what products each Product_ID represent. We also took out Product Categories for variables because most of the values are missing in Product_Category_B and Product_Category_C and again, we are not able to know what each product category represents. Additionally, we broke apart the variables age, occupation, number of years living in current city to their factor levels instead of integer. We also set the seed to 1861 so our data analysis would be easily reproducible. Lastly, we used

options (scipen=15) to change the of the scientific notation outputs in more easily understood numbers.

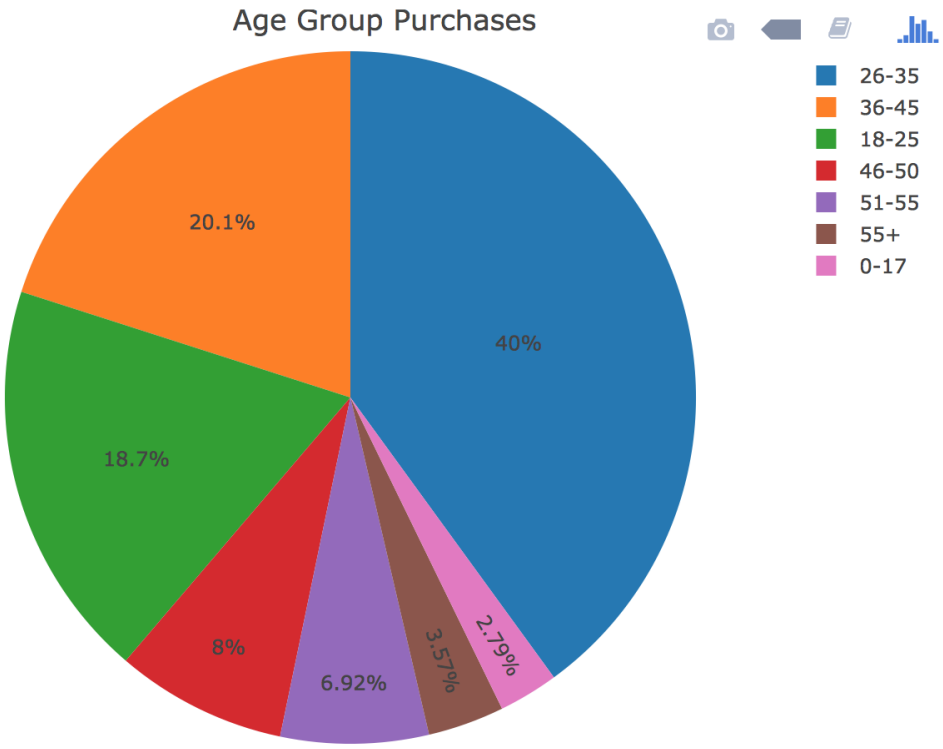
After making these adjustments, we did a basic summary over the dataset to get a better idea of the variables we were working with. Through this summary we found that there were 127,346 males that data was taken from while only 36,932 females. We also found the purchase values (before doing the log transformation) minimum value to be $185 and a maximum value of $23,950.

Analysis

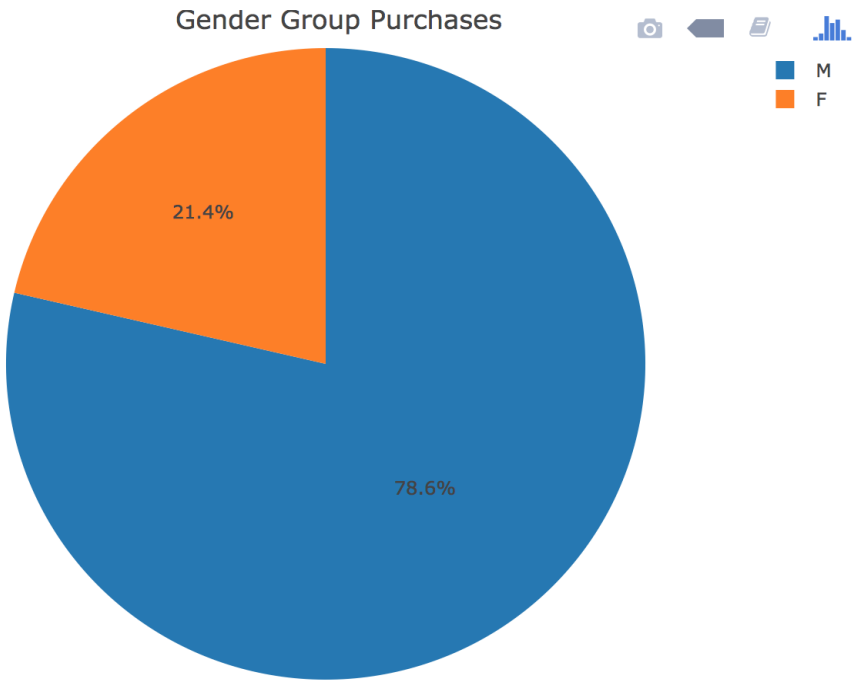
After getting a better understanding of the variables, we moved on to creating plots based on the data we found most interesting.



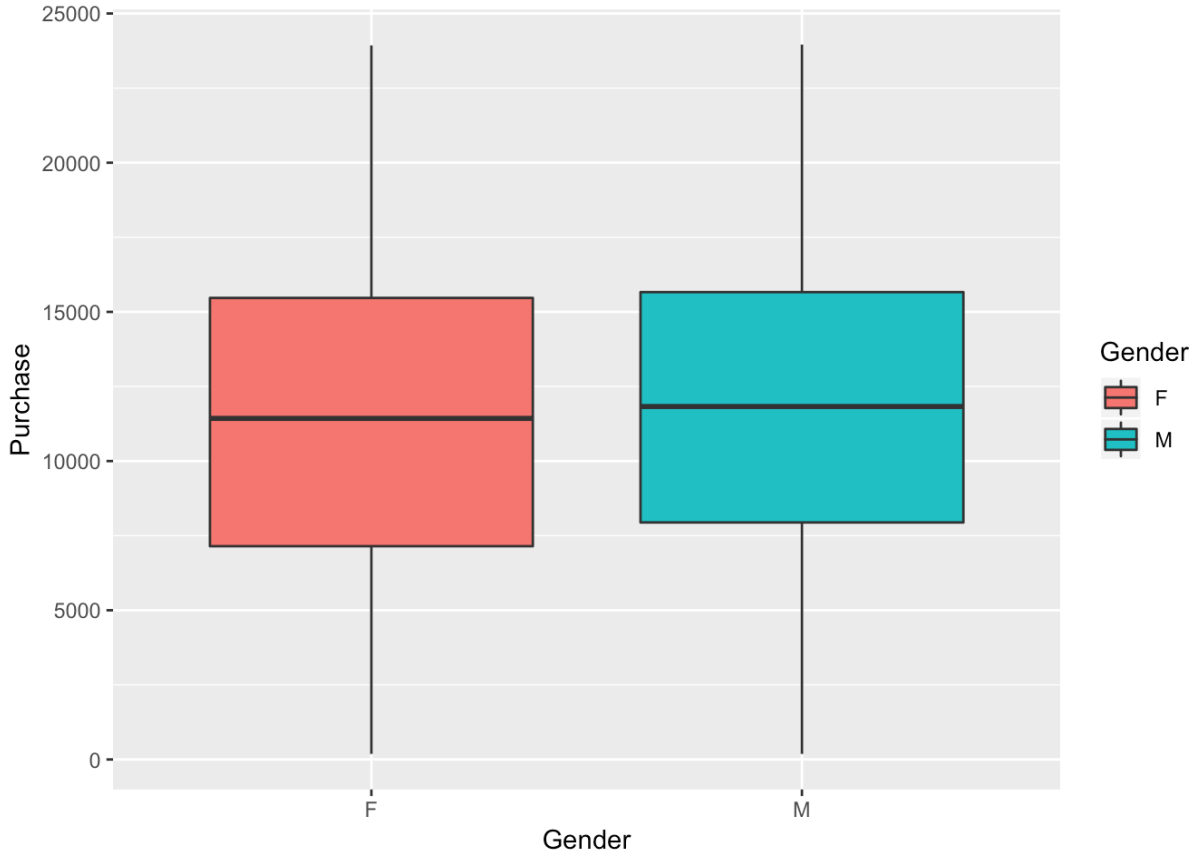
This first plot is showing a histogram of the total purchase value compared to the number of people who spent that value. As you can see the data the biggest column, accounting for around 30,000 people matches with around $16,000.



This plot shows the distribution of the age groups of shoppers in a pie chart. As you can see, most of the shoppers came from the age group of 26 to 35 year old accounting for 40% of the data. While the smallest amount of shoppers came from the ages 55 and over and people 0 to 17 years old and which accounted for only 3.57% and 2.79% of the data respectively.



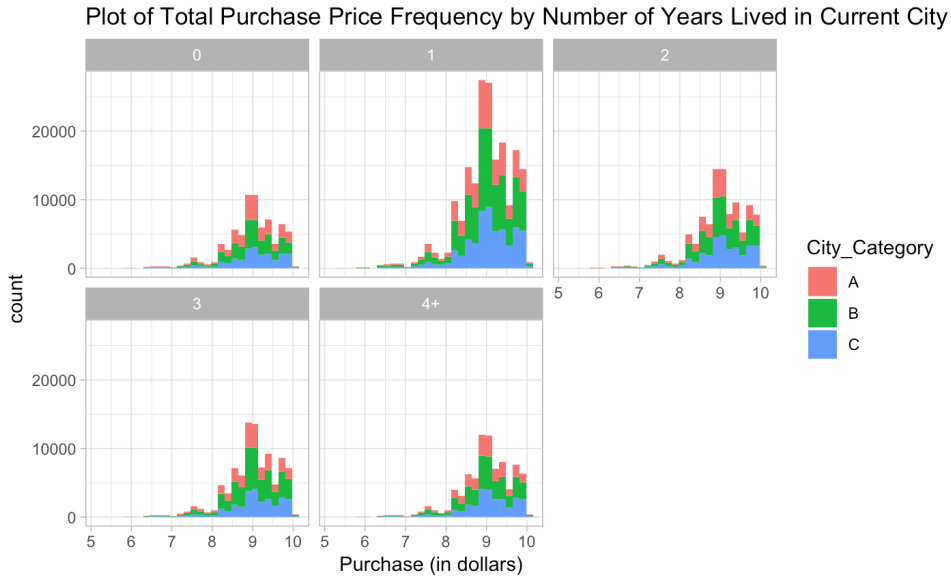
This next pie chart shows that the majority of the total purchases were done by males at a total of 78.6%. This is consistent with our summary statistics which indicated that the majority of data was taken from males. Meaning that because the majority of data was taken from males, it makes sense that this group also displayed the highest purchases.



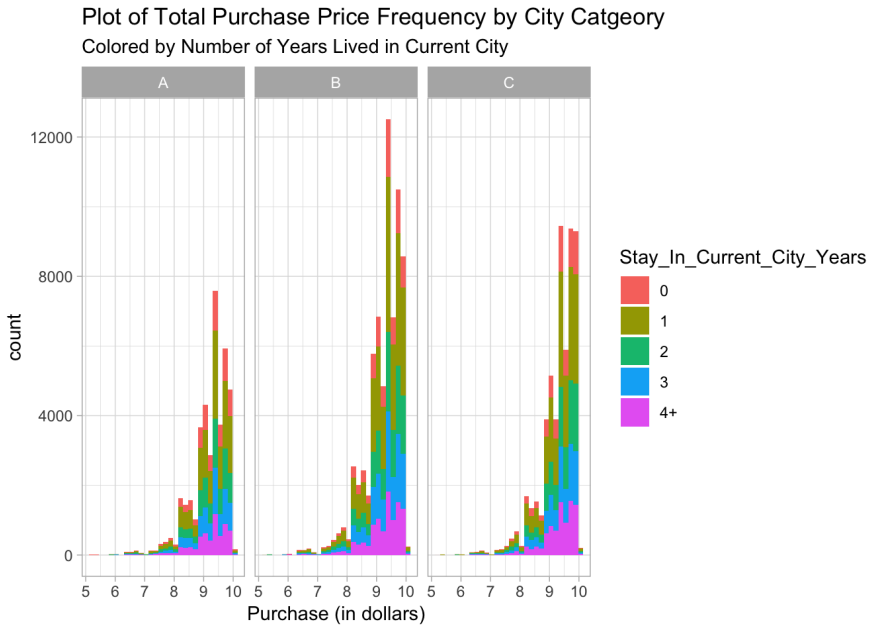
These boxplots indicate that although the median for both genders is practically the same, at a value of about $12,000. This is consistent with the summary of values which indicates the overall median of purchases is $11,757. But, males have a slightly larger distance between median and the third quartile and females have a larger distance between the first quartile and the median.

For the rest of the plots and the data analysis, we used the log of purchase value to reduce the skewness and to make the values easier to interpret.

This next shows histograms of the purchase total values compared to the 20 factor levels of the variable occupation. As we can see, all of the plots are skewed left. Additionally, the occupations 0,1,4,7,12, and 17 look to have the most activity in the number of shoppers they have. However, the downside is that we do not know what the occupation levels actually represent so we cannot draw conclusions about specific professions spending or shopping more than others.



This plot show the total purchase frequencies split up by the number of years people have lived in their city, colored in by which city category they live in (A, B, or C). From this plot we can conclude that the people who have lived in the same place for a year tend to have the highest amount of shoppers.



This plot displays the purchase totals broken up by the city categories and then colored in by the number of years people have lived in their current city. It shows that most of the shoppers

come from city category B. Additionally, the longer people live in the area, the less they tend to shop. We thought this could be related to how close the store is to each city, since we do not know what the cities stand for, or what type of store this is like is it a stand alone or in a mall. All of these things could play a role in why we are seeing these trends.

Models

As for the models we created using this dataset, we used linear regression, generalized linear, lasso, ridge, and forward stepwise models to analyze the data.

Starting with linear regression, we used the variables, gender, age, occupation, city category, number of year lived in current city and marital status to predict purchase. We chose to use these variables because they were all of the ones in the dataset (after excluding the variables we mentioned taking out earlier). With this linear regression model, the most statistically significant variables were gender, ages 51-55, occupations 5, 6, 7, 12, 14, 15, 16, 17, 19, and 20, city categories B and C, and staying in current city for 2 years. This model also gave an adjusted r-squared of 0.01038, meaning that only about 10% of the variability of the purchase value around the mean within the model is being accounted for in this model.

Next, we created a generalized linear model. For this we created a binary variable using the median purchase value. We then used the same variables as the linear regression model and the to predict the higher purchase values (1) or low purchase values (0). This model gave similar statistically significant variables, the only difference was that is indicated occupations 1, 4, 5, 6, 7, 8, 12, 14, 15, 17, 18, 19, and 20 instead. Although this model does not provide a adjusted r-squared, we still wanted another way to determine the goodness of fit. So, we looked into the deviance values. Deviance is a measure of goodness of fit of a generalized linear model. The

higher the numbers, the worse the fit. The null deviance shows how well the response is predicted by a model with nothing but an intercept while the residual tells us how well we can predict our output using the intercept and our variables. Additionally, the bigger the difference between the null deviance and residual deviance is, the more helpful our input variables were for predicting the output variables. As you can see, the difference between the two is substantial at about 4,350 and the numbers themselves are pretty large. Therefore we can conclude that the overall fit is okay.

The third and fourth model we created were lasso and ridge. We used these to see if we could build upon making better models. Although the resulting coefficients for the min of both models were fairly similar, the 1se of both were quite different. For the 1se of the Lasso model, it chose only the variables gender and city category C. When comparing the r-squared of both models, the min values were again similar at values of 0.0105828 and 0.0205729. However, the r-squared values for the 1se of both models indicated that the lasso model was slightly better with a value of 0.00443 while ridge had a value of 0.00356.

Although the findings of lasso and ridge were interesting, we also wanted to do a forward stepwise model to see if the variables it chose were consistent with those chosen by the lasso model. Our forward stepwise model, with a nvmax set to 8 (consistent with the default), chose the variables gender, age 51-55, occupations 1, 10, 19, 20, city category B and city category C. As you can see, this is quite a few more variables than the lasso model, because if the nvmax, and is actually more similar to the the linear regression and generalized linear model created. So, we decided that we wanted to see how a linear model with these variables would perform. So, we created different variables based on these factor levels chosen and added them to the dataset. A summary over the linear regression model we created out of these variables showed that there

were all considered to be statistically significant. However, the r-squared value of these was only at 0.00919.

Results

```
##                                    RMSE                     MSE
## Linear Regression  0.00000000000004791599 0.00000000000004791599
## Generalized Linear 0.00000000000062269982 0.00000000000066568344
## Lasso              0.00000000000003284116 0.35263327463169907849
## Ridge              0.35263327463169907849 0.35482611630906329969
## Forward Stepwise   0.35513665466722271757 0.00000000000003284116
```

Since we could not rely on the adjusted r-squared values for these models, given that it is not the best value to go off of. Adjusted r-squared does not determine if a linear model is adequate at fitting the data. So, we decided to calculate the root mean squared error (RMSE) and mean squared error (MSE) for all of the models in order to compare them. We put these values into a table so they are easier to read and compare, as seen above. As you can see, the smallest RMSE is given by the Lasso model and the smallest MSE is provided by the linear regression model using the variables selected by the forward stepwise model.

Conclusion

Based on the result, we are able to conclude that Gender and City_Category are the variables that relate to the amount of purchase the most. In terms of the characteristics of shoppers, people who are male, between age 26-35, work in occupation 4, have lived in their current city for 1 year, and live in City category B are the people who tend to spend more money on Black Friday. For further analysis we would be interesting in obtaining the meaning of these variables, specifically occupation, city categories and product categories, in order to gain a better understanding of what this all means.