Morganne Peak

Data Practicum I

Final Project Write-Up

Dataset:

The dataset was retrieved from an online database, Kaggle. The dataset includes the information of black Friday shoppers. The data was collected from a company that was looking to learn more about the demographics of their customers and maybe determine customer behavior based on the information gathered. The original dataset has 12 variables and 537,578 observations.

Variables:

**User\_ID** – The number assigned to each individual customer

**Product\_ID** – The identifier for the particular product purchased

**Gender** – Male = 1, Female = 0

**Age** – 0-17=0, 18-25=1, 26-35=2, 36-45=3, 46-50=4, 51-55=5, 55+=6

**Occupation** – 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

**City\_Category** - A=1, B=2, C=3

**Stay\_In\_Current\_City\_Years** – 0, 1, 2, 3, 4+=4

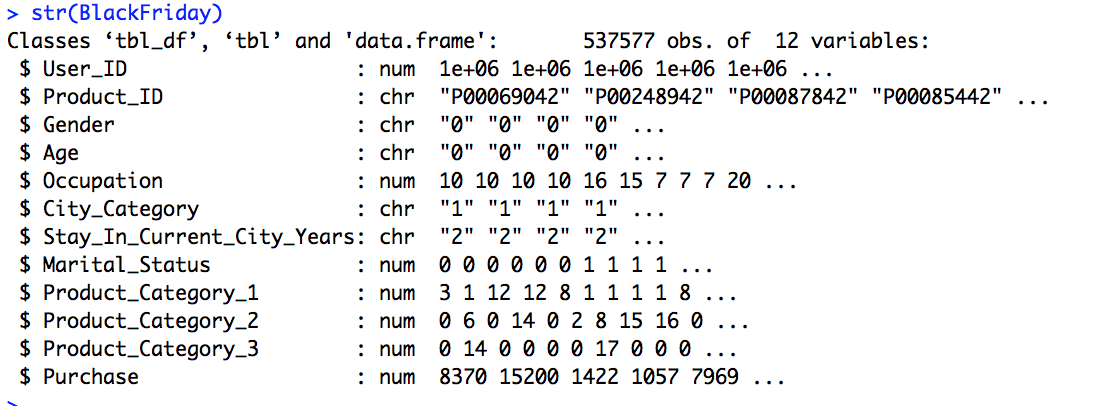
**Marital\_Status** – Single = 0, Married = 1

**Product\_Category\_1** – This column captures the number of items the customers purchased from this category.

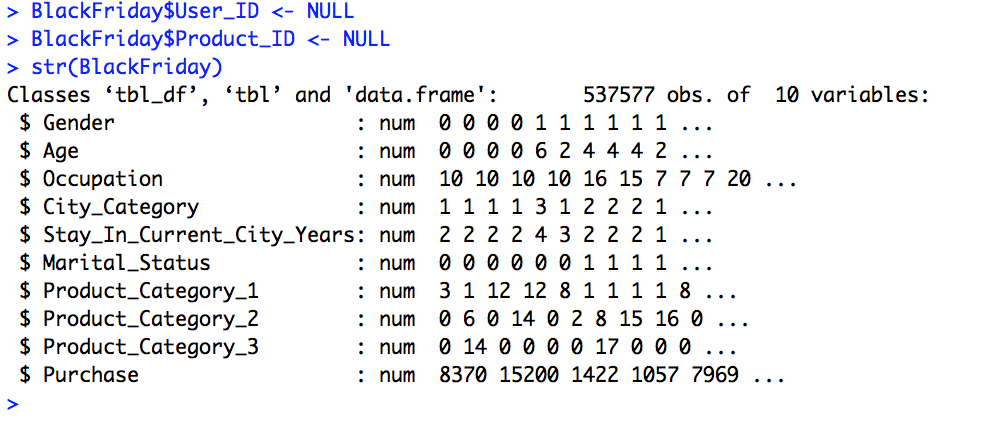
**Product\_Category\_2 –** This column captures the number of items the customers purchased from this category.

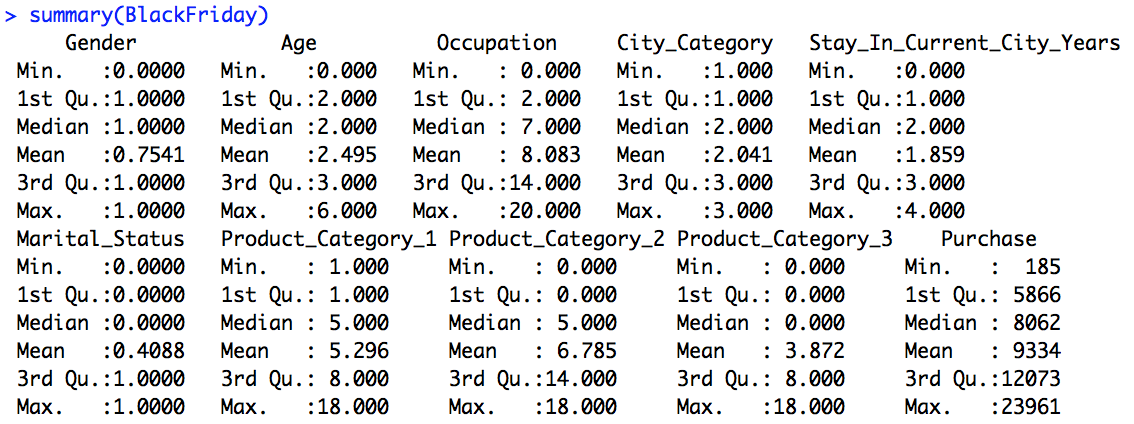
**Product\_Category\_3 -** This column captures the number of items the customers purchased from this category.

**Purchase –** The amount each consumer spent.



Here is the way that the data came into R. There are some variables that came into R as a ‘character’ variable. I will be changing those variables to be numeric. I decided to remove both the User\_ID and Product\_ID columns because they do not have much significance to the dataset as a whole.

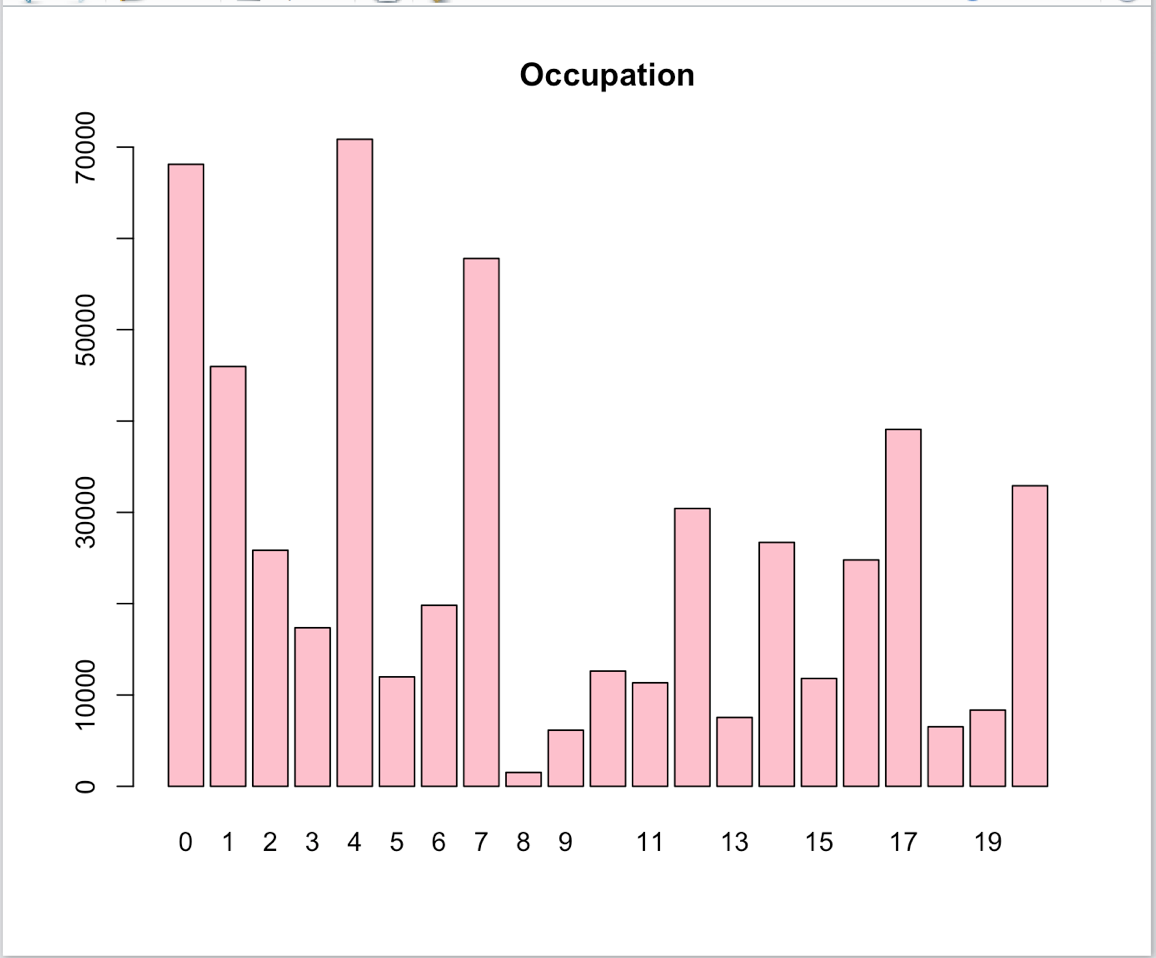
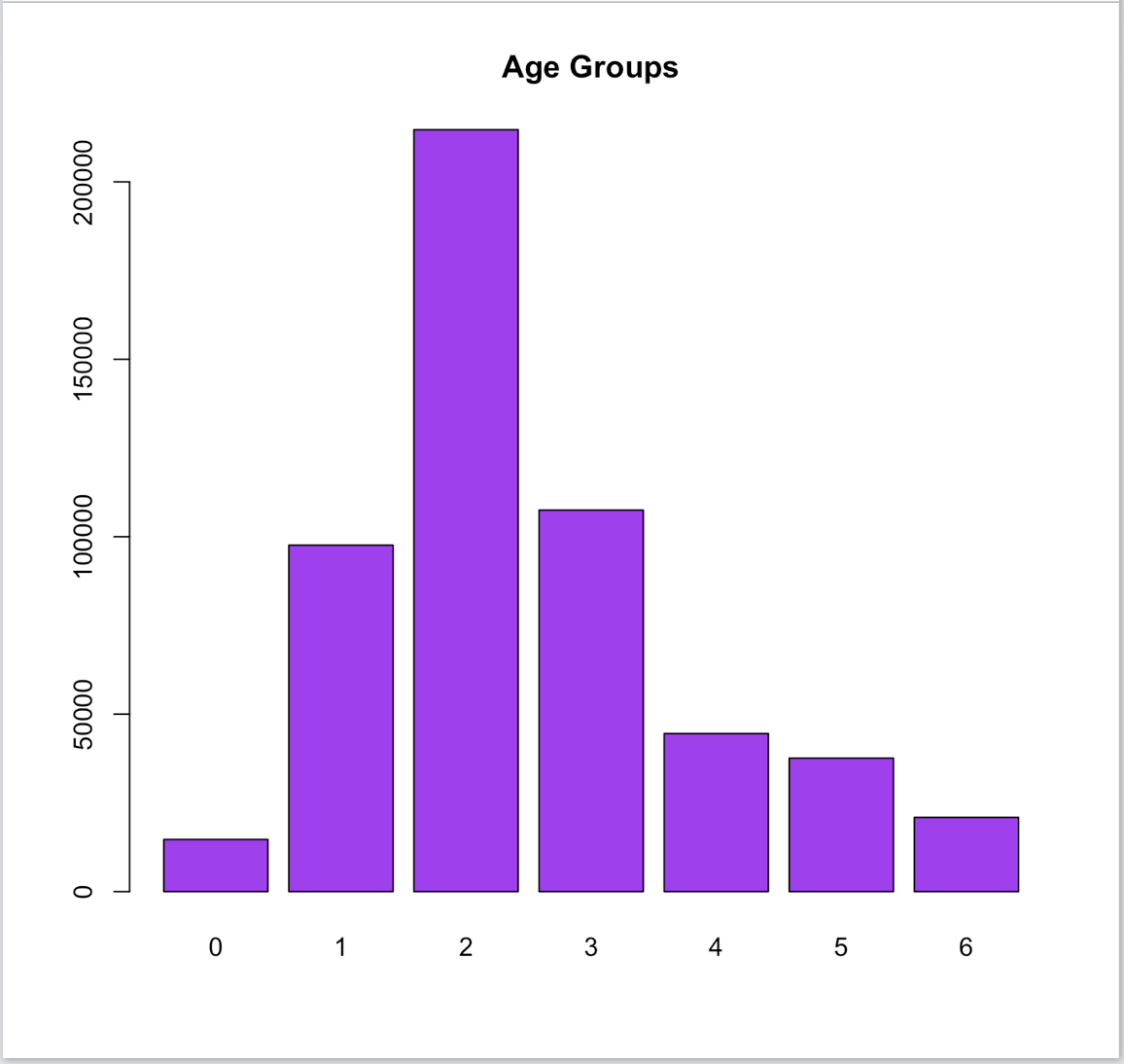


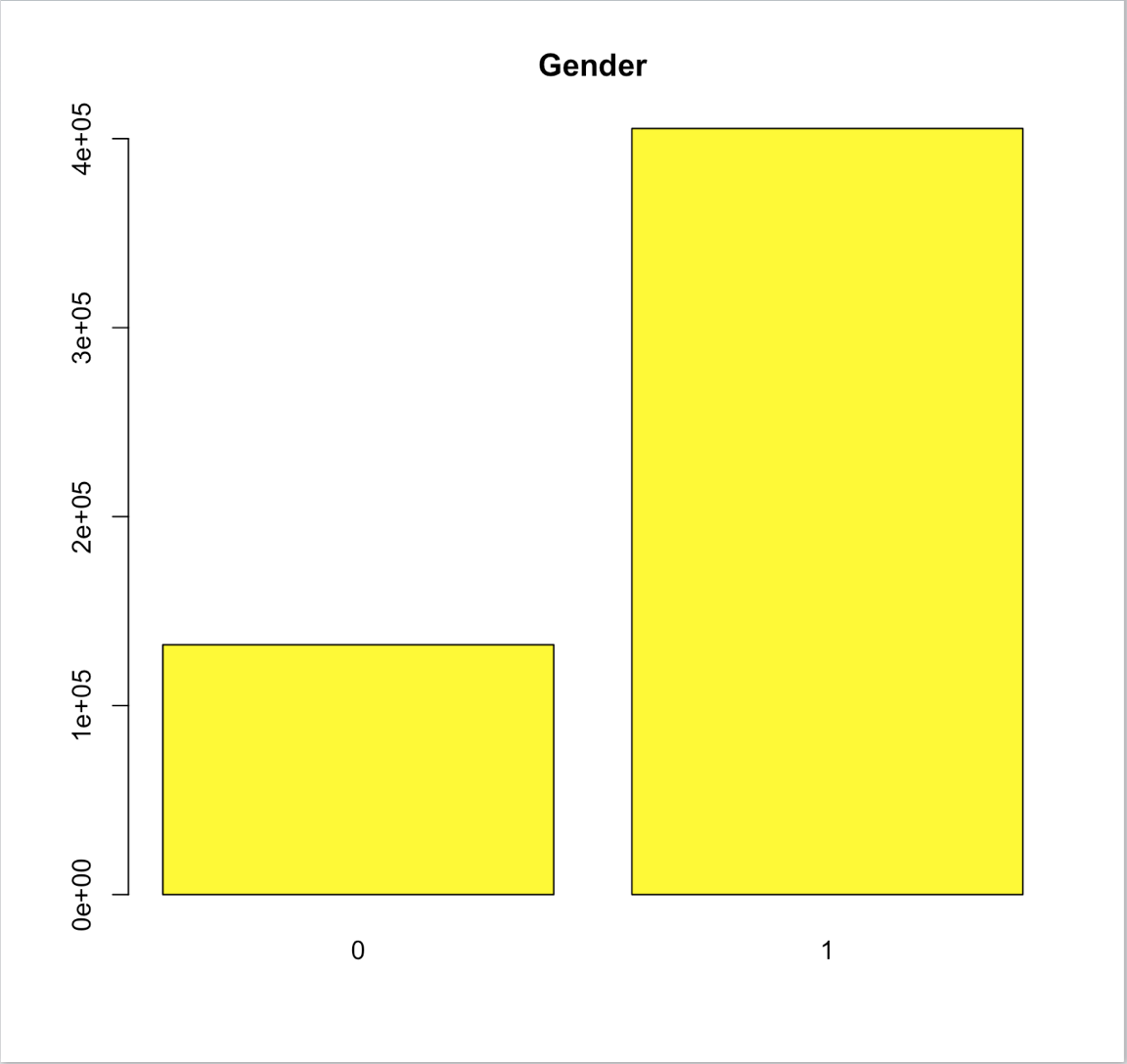
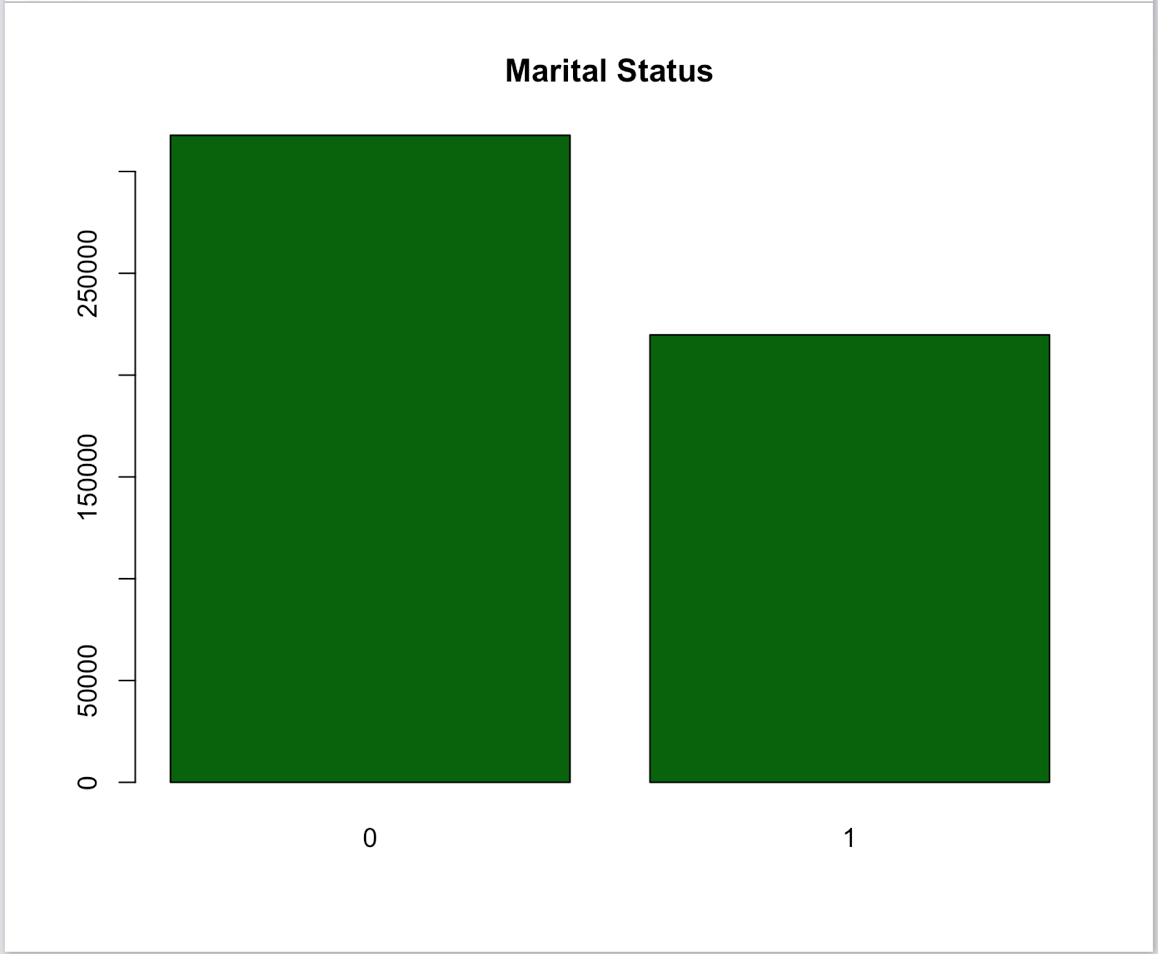


After all changes to the data have been made there are 10 variables with 537577 variables. The summary report of the data shows that there are not really any outliers that could somehow deviate the outcome of a model ran. It will be good to move forward with the data since there are not any huge outliers.

Data Visualization:

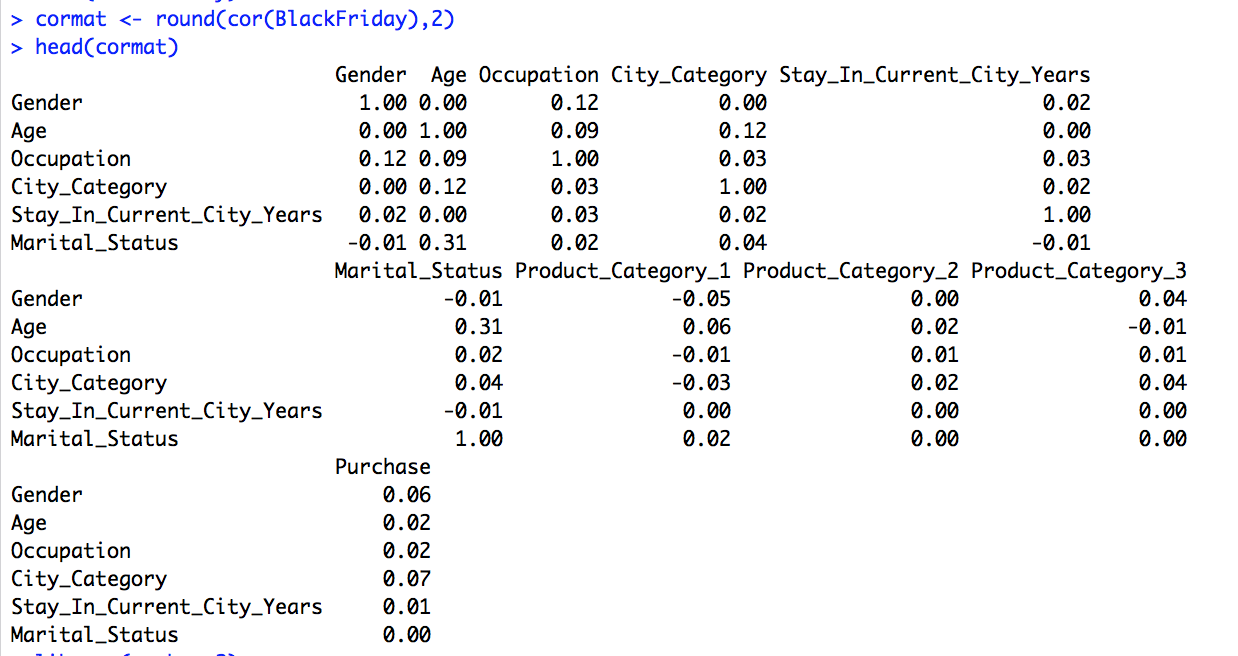
I created a few graphs of the data in order to learn a little more about it. The first graph created is of the age groups in the data. According to the graph, the age group 26-35 had the highest number of people participate in black Friday shopping. The next bar graph is of the different occupations of the consumers. Occupations 4 and 0 were the highest groups. Unfortunately, the data did not include a description of what the different occupations are so there aren’t any conclusions that can be drawn there. The last two graphs are of the gender and marital status variables. More single people participated than married people and more males participated than females.

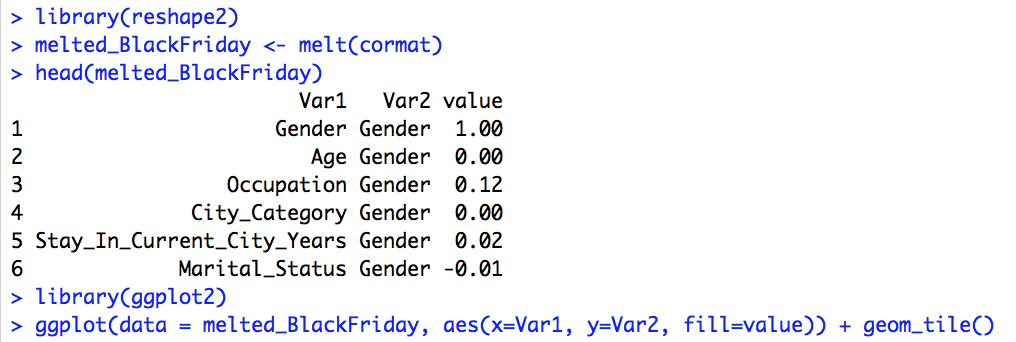


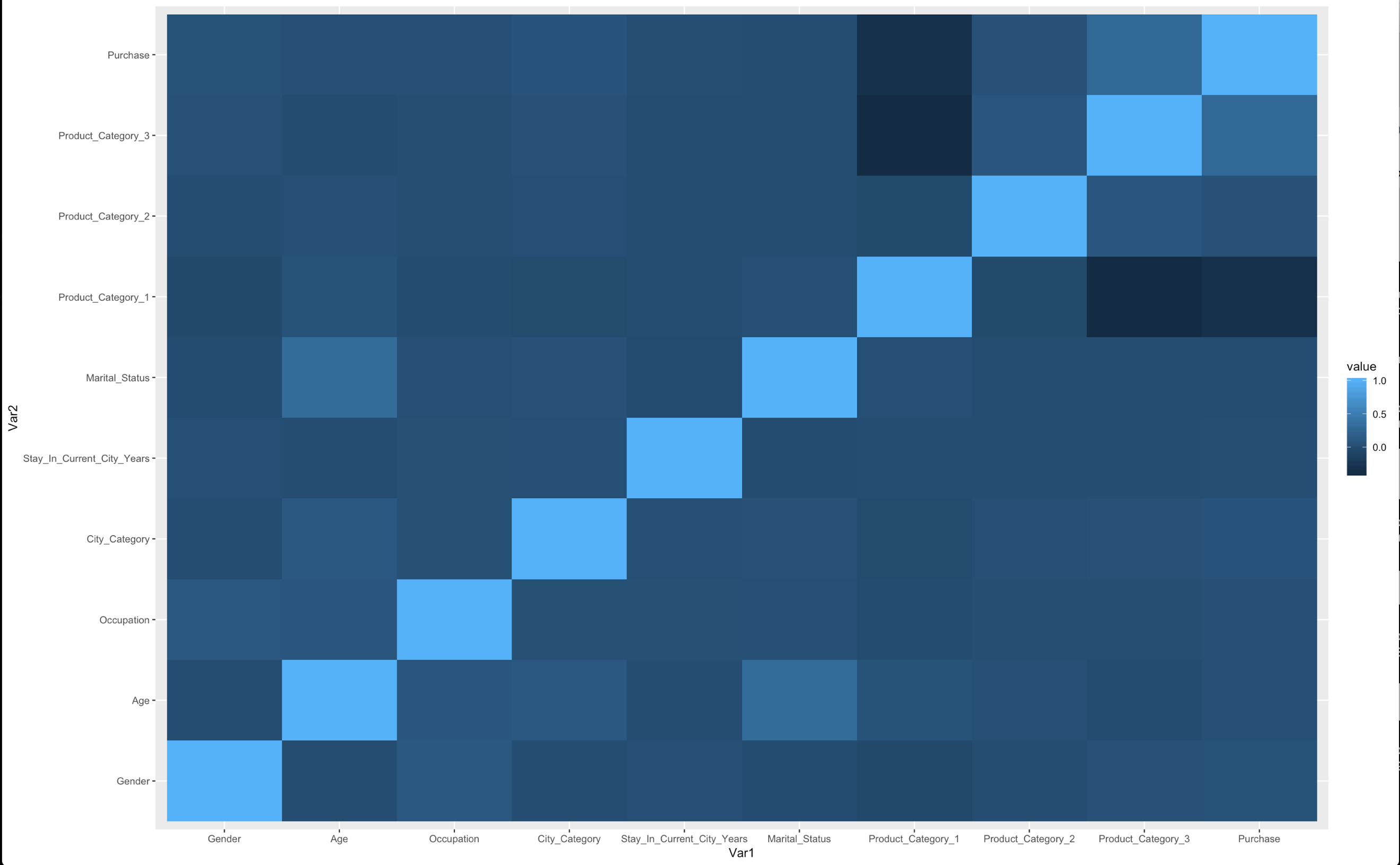


Model:

After much thought and trial and error, I decided to use a regression model for this dataset to help predict the purchase amount given the other variables. The first step was to create a correlation matrix in order to see if any of the variables are closely correlated to the ‘Purchase’ variable. The coding below shows the correlation by creating a heatmap.

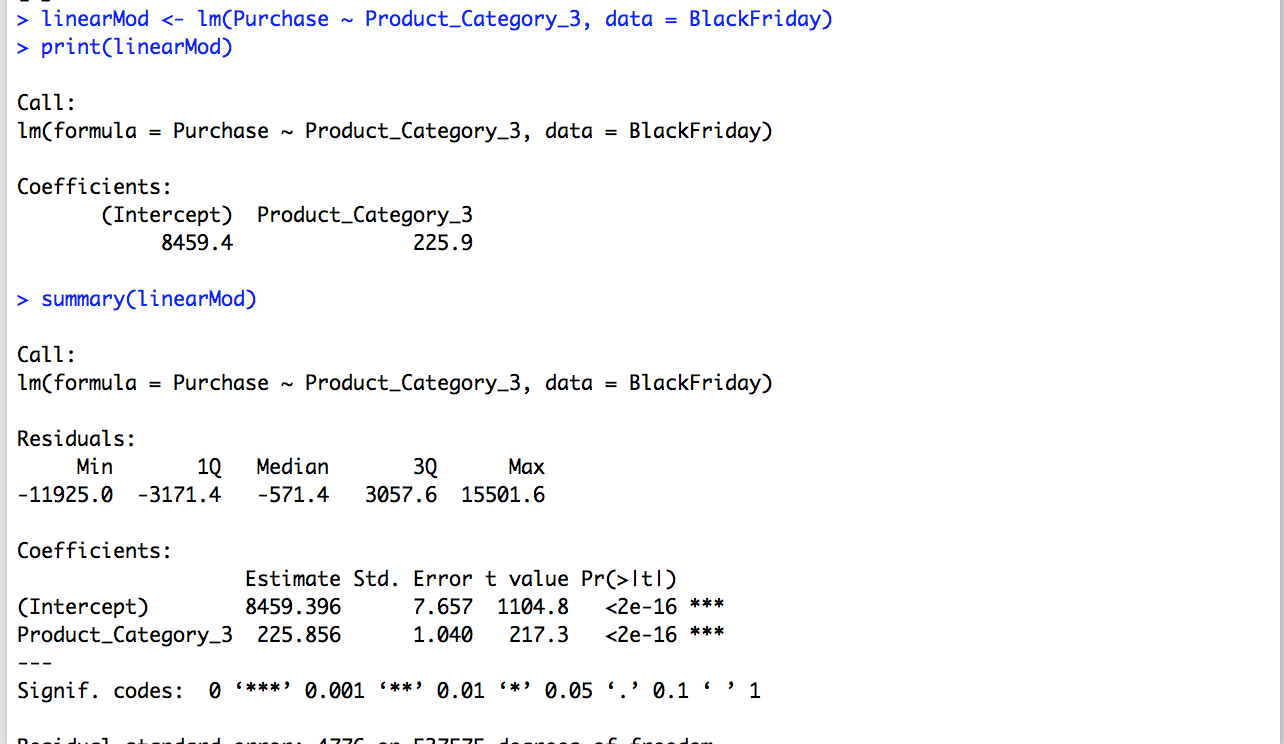


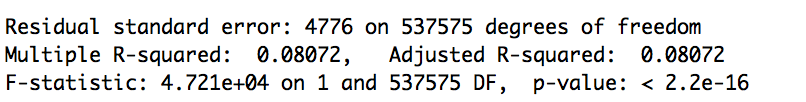




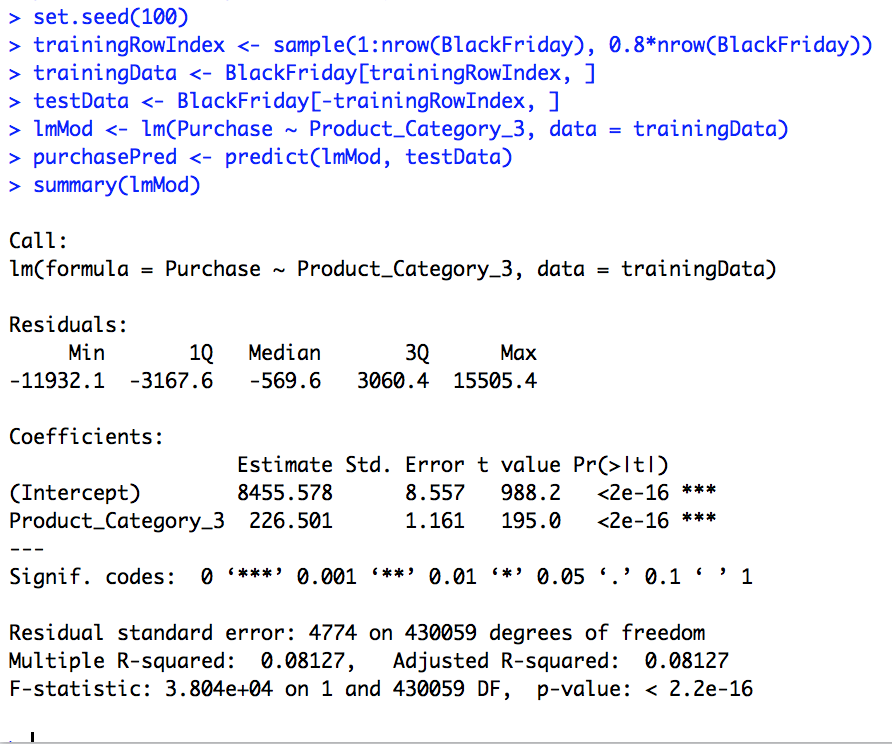
From the heatmap above, you can tell that there are not any variables that are closely related to the purchase amount. You can also see that ‘Product Category 3’ is the only variable that shows any potential correlation with ‘Purchase’. Thus, the relationship between these two variables will be used to create the regression model. As another note you can see the ‘Age’ and ‘Marital Status’ are correlated and perhaps this correlation could be used for another prediction model.

The first step in the process is to create the linear regression model. A linear equation with the slope and y-intercept in create when the model is created. In order to know in the model is statically significant, i.e. a good model to use to predict the dependent variable, you can look at the p-values, r-squared, and r-adjusted. After reviewing those numbers, it can be concluded that the model is a good predictor or good fit for the data because the p-value is less than 0.05. The r-squared values are a little low but will still work for this model.





Now that we know it is possible to make a good predictor model from the dataset, it is time to use a model to predict the dependent variable. The next step is to split the Black Friday data into a train and test datasets. The data in this case was spilt 80/20, 80% of the data is now for training the model and 20% of the data is for testing the model. As in the model before, the p-values are under 0.05.



Conclusion:

The results of the model show that it is possible to create a regression model to show the relationship between  Product Category 3 and Purchase amount. However, the correlation between the two variables may not be strong enough to create a strong linear model. I think that the strong the correlation between two variables, the stronger the regression will be. I don't think that it is the worst regression model but it could definitely be better.

References:

Dagdoug, M. (2018, July 25). Black Friday. Retrieved from

<https://www.kaggle.com/mehdidag/black-friday/home>

Ggplot2 : Quick correlation matrix heatmap - R software and data visualization. (n.d.). Retrieved from [http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-](http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization)

P. (n.d.). Eval(ez\_write\_tag([[728,90],'r\_statistics\_co-box-3','ezslot\_4']));Linear Regression.

Retrieved from <http://r-statistics.co/Linear-Regression.html>