Black Friday Project

Dave Anderson

December 6, 2018

Abstract

This report outlines a study of Black Friday sales data for a single company. The company itself is masked, as are many of the categories for variables. The general purpose of studying this data is to help maximize profits for the company. This is accomplished by first analyzing the demographics and general spending trends of the customers. Then I developed multiple models in hopes of predicting how much various customers are expected to spend and gain more insight into the customer base. Data limitations, including the masked categories, mysterious purchase amounts, and large variations prevented the development of a trustworthy model. At the same time, I was able to use the models to draw important conclusions about the company's target demographic.

Introduction

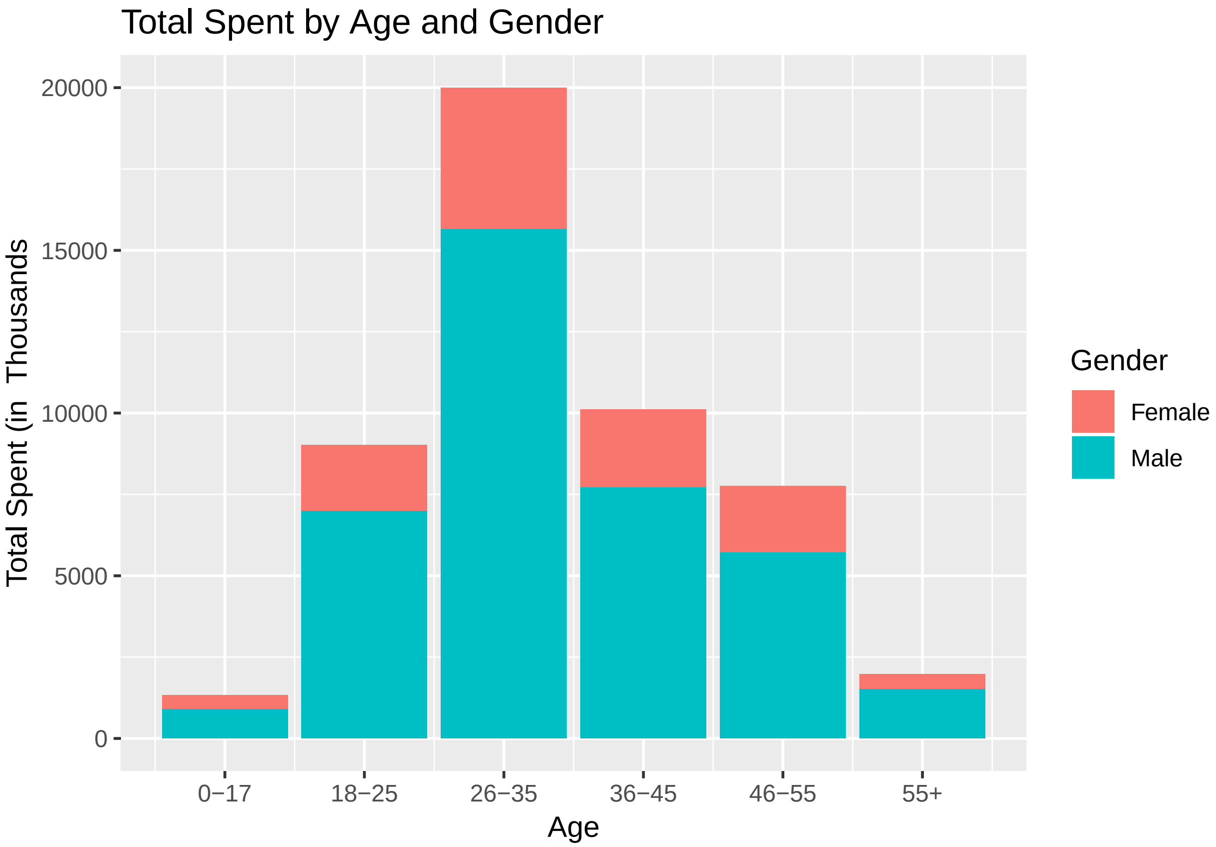
I find quantifying human behavior to be both incredibly interesting and powerful. One of the best ways to understand our society is to examine consumer data. Like it or not, money is a critical aspect of our lives, and how we choose to spend it is an important decision. I am fascinated by the ability of companies to predict who their customers will be and what they will buy, and I would be honored to work in a position where I perform similar analysis to help a company succeed. To begin to understand this field, I chose to analyze a company's black Friday sales report. The data includes about 500,000 transactions. Variables include a customer code, age category, gender, occupation category, product category, and city information. Most of the data is masked as we do not know what the store is or what the categories represent. At first, the missing information deterred me from this dataset, but it was interesting to attempt to guess the nature of the categories as I investigated the data. The overall goal of my analysis is to build a model that will help predict spending of customers in order to maximize profit and marketing resources for this company. I will begin by organizing and visualizing the data at a basic level and check for relationships between variables.

Method

EDA

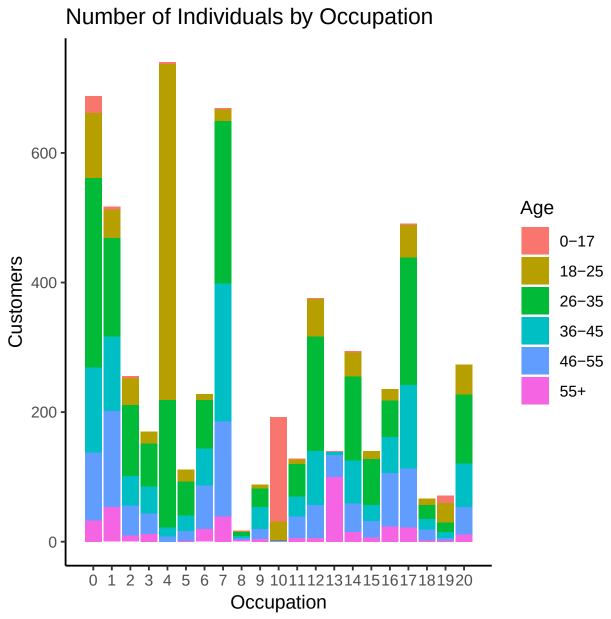
Demographics

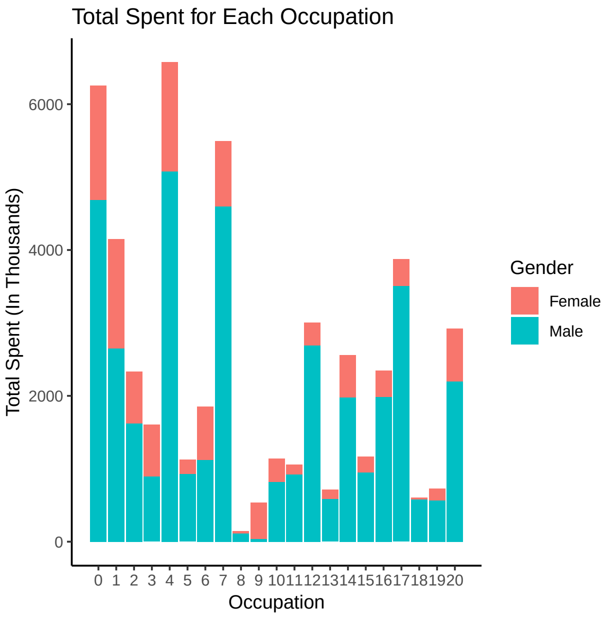
The original dataset was found on Kaggle and comes from a competition hosted by Analytics Vidhya. The original form will be useful in analyzing the consumer data by looking into what products certain people are buying. But first, I wanted to learn about who the customers are. I created a dataset with each individual as a row, including average purchase, total purchase amount, and number of purchases as variables. There are 5,891 customers with number of purchases ranging from 5 to 1,025. From the first plot, we can see that the target demographic of this store is 26-35 year old males. Males tend to buy more expensive products (95 to 88), more items (222 to 192), and there are many more male customers in general (4,225 to 1,666).



Occupations

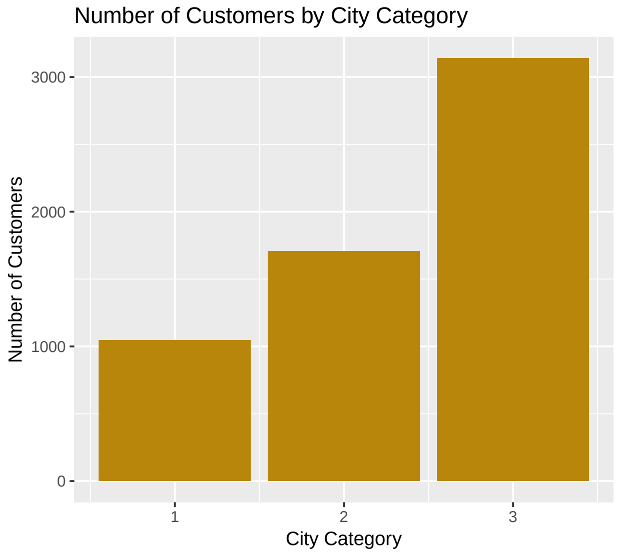
Occupation is most likely another key factor. From the two plots, we can see that occupations 0, 4 and 7 have the most customers and contribute the most total revenue. Occupations 12, 15, and 17 buy the most expensive products, on average, but that metric does not vary much (87-99). Occupation 9 is the only one with a female majority. Occupation 10 is where most of the customers under the age of 18 are, which makes me believe this is unemployed or student. Occupation 4 could potentially be college students, with most of the 18-25 year-olds in this category. The company's target age, 26-35, is present across multiple occupations.

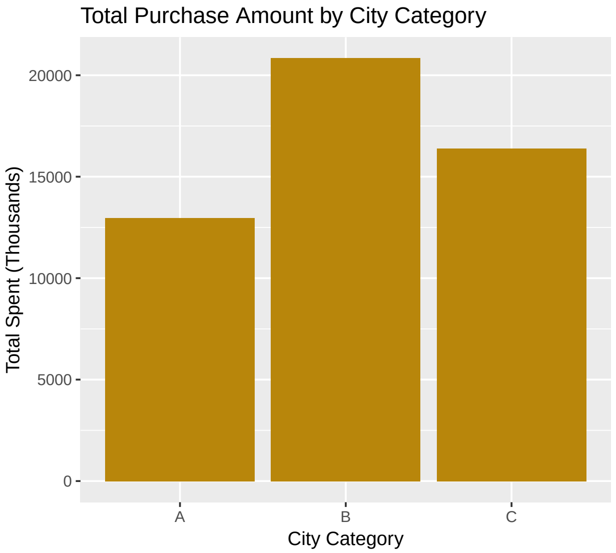




Cities

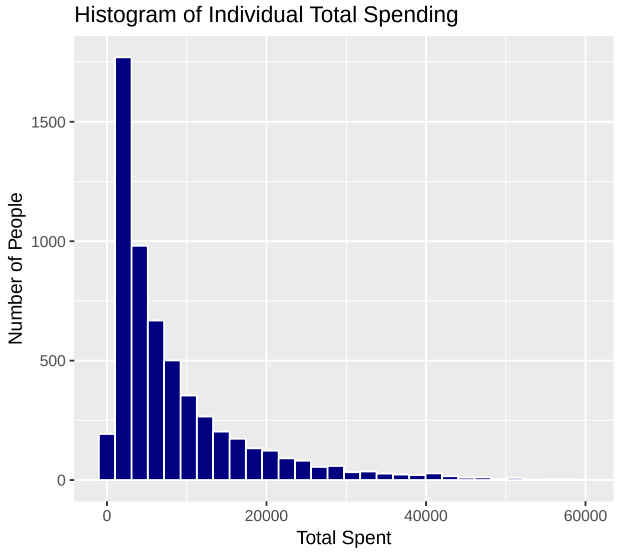
The city category may be difficult to draw conclusions from as there are only three categories, and we don't know what they represent. The two bar plots below show that the largest portion of customers come from city category 3, but total purchases are highest in category 2. We could guess that city category 3 is a low-income area. Age and gender are evenly spread across each city category and length of stay in the current city.

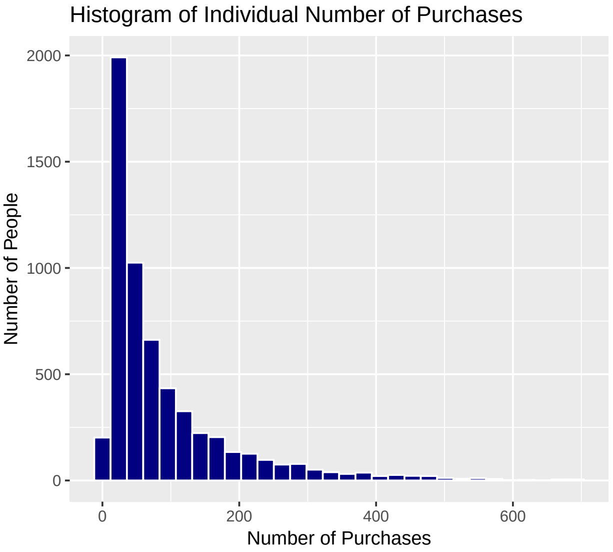




Purchases at Customer Level

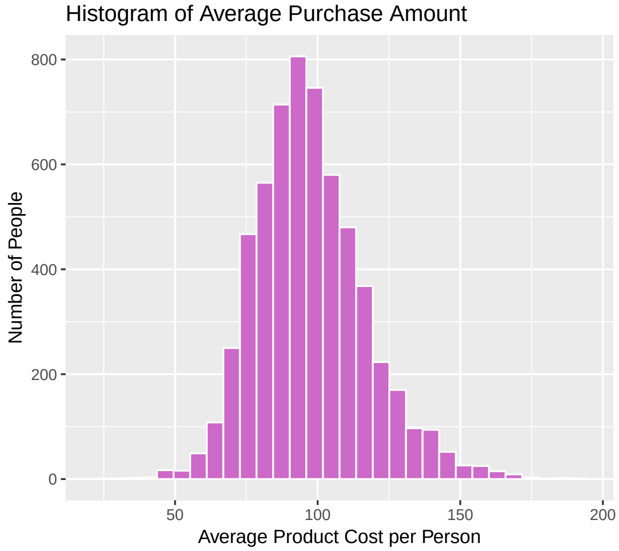
As previously stated, the goal of studying this data are to help maximize profits for the company. Therefore, the key outcome variables for our models will be purchase amounts for the individual, number of purchases for each individual, and product purchases. To investigate these variables, I began with a histogram of total amount spent by each individual. For visual purposes, I excluded 13 individuals over 60,000, with the largest value being about 100,000. Plotting the distribution of the number of purchases for each customer shows a similar pattern.

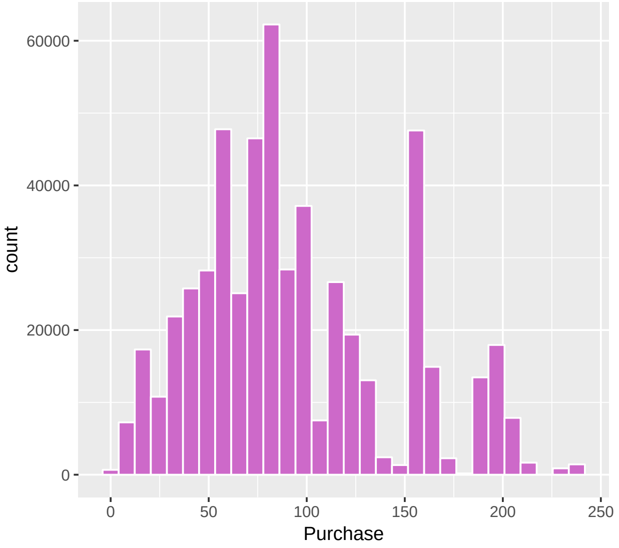




Product Purchases

Continuing to examine our potential response variables, I chose to look at the average purchase amounts for each individual. As expected, the distribution of these averages looks approximately normal. The plot on the right shows the distribution of all 537577 purchase amounts from the original dataset. I assume the peaks in the histogram are from certain products that are especially popular.





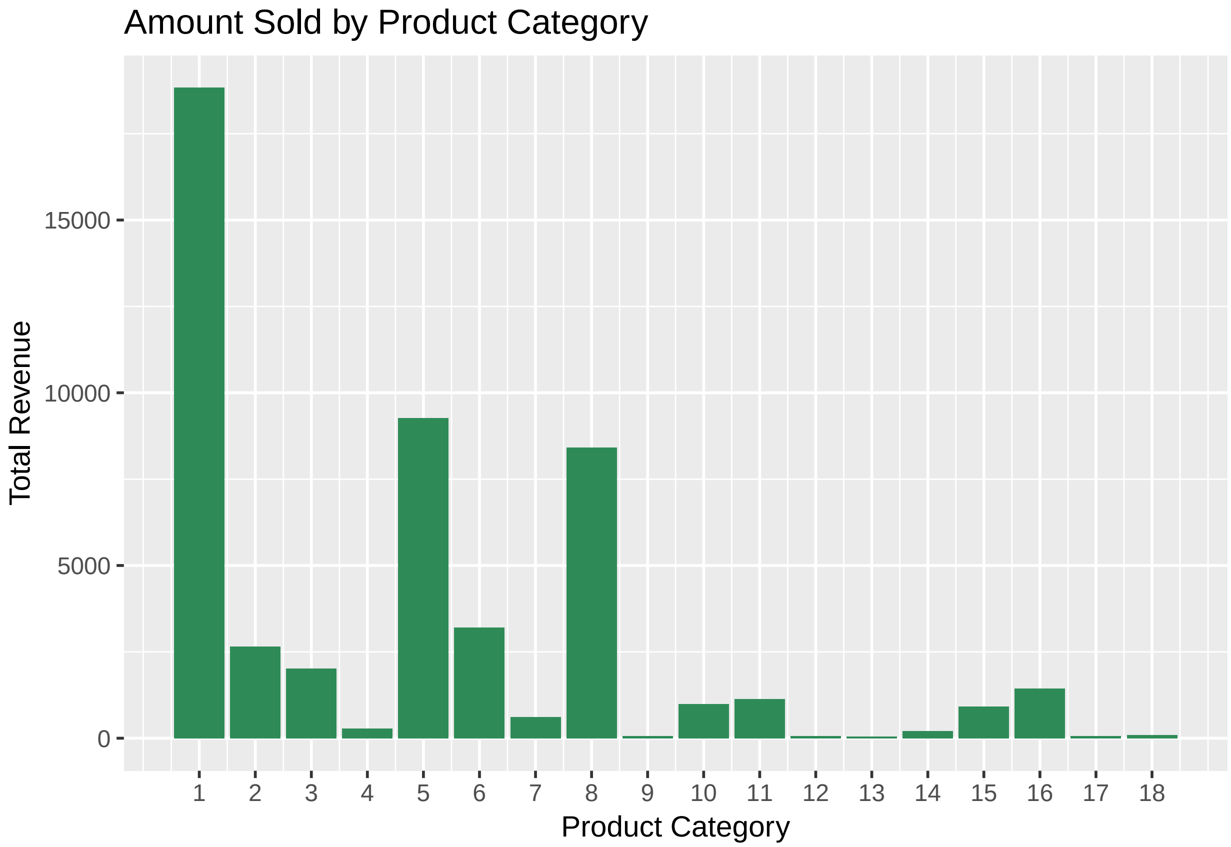
Products

Product categories range from 1 to 18. Some observations have one category listed, while others have two or three. With so many categories of products, I am tempted to think this store is a large department store. On the other hand, we can see from the bar plot that most of the sales fall into a only a few categories. Grouping

Table 1: Top Products

Product_ID	Number	Male Ratio	Average Age Category	Category
P00265242	1858	0.73	2.38	5
P00110742	1591	0.78	2.32	1
P00025442	1586	0.78	2.41	1
P00112142	1539	0.78	2.26	1
P00057642	1430	0.82	2.32	1

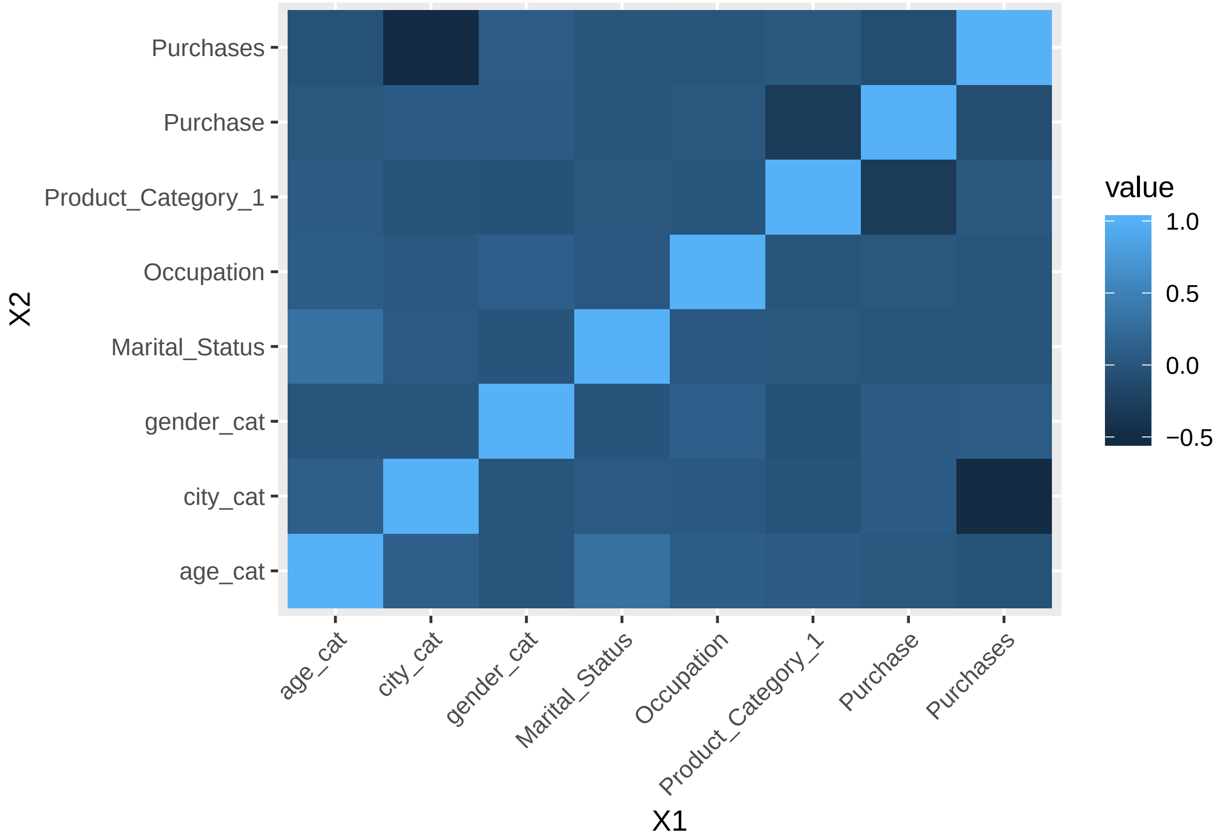
by product ID, we can see that there are some very popular items. The strange pattern of sales from categories, combined with the large variation of prices within categories, makes me skeptical about its help in modeling. As we would expect, the average age and gender of these top-selling products is in line with our target demographic.



EDA Conclusions

I had hoped to have gained some insight into the categories and the store itself by this point. The target demographic of this store is clearly 25-36 year old males. A few occupations contribute more total revenue, but individual spending habits do not differ among occupations. It is difficult to draw any more conclusions as to what the occupation categories may represent. The large number of product categories being sold makes me think this is a large department store. When thinking of stores that would target young males, electronics and athletics come to mind, but I would expect to see more female customers at a large chain such as BestBuy or Dicks Sporting Goods. The biggest issue I am seeing with the dataset is the purchase amounts. The original source claims these are in dollars, but I had a hard time believing the store was selling $20,000 items. Therefore I divided the amounts by 100, but it still didn't seem right. The purchase amounts actually vary among individual products. With the purchase amounts being the focus of my models and

conclusions, these are troubling assumptions to keep in mind.
The large number of observations combined with mainly categorical variables makes plotting scatterplots difficult. Based on EDA and the following correlation plot, it seems as though we will need to try utilize almost all variables to create a predictive model. I will not include marital status, as this seems to be closely related with age.



Models Chosen

The company would most likely want to predict how much an individual of a certain discription will spend in their store. I will attempt to create a simple linear model with a log transformation on the individual total purchase amounts as the outcome variable. This model will probably not be the best fit model, but it will allow for easily interperatable effects of each category. Because I do not completely trust the purchase amounts, I also ran a ordinal multinomial regression. I have created categories for each purchase. Categories are set from 0-3, every $50 is a new category. I attempted multiple combinations of variables as predictors. None of the models stood out as more effective than the others. In fact, using most of the variables seemed to produce the more interesting results.

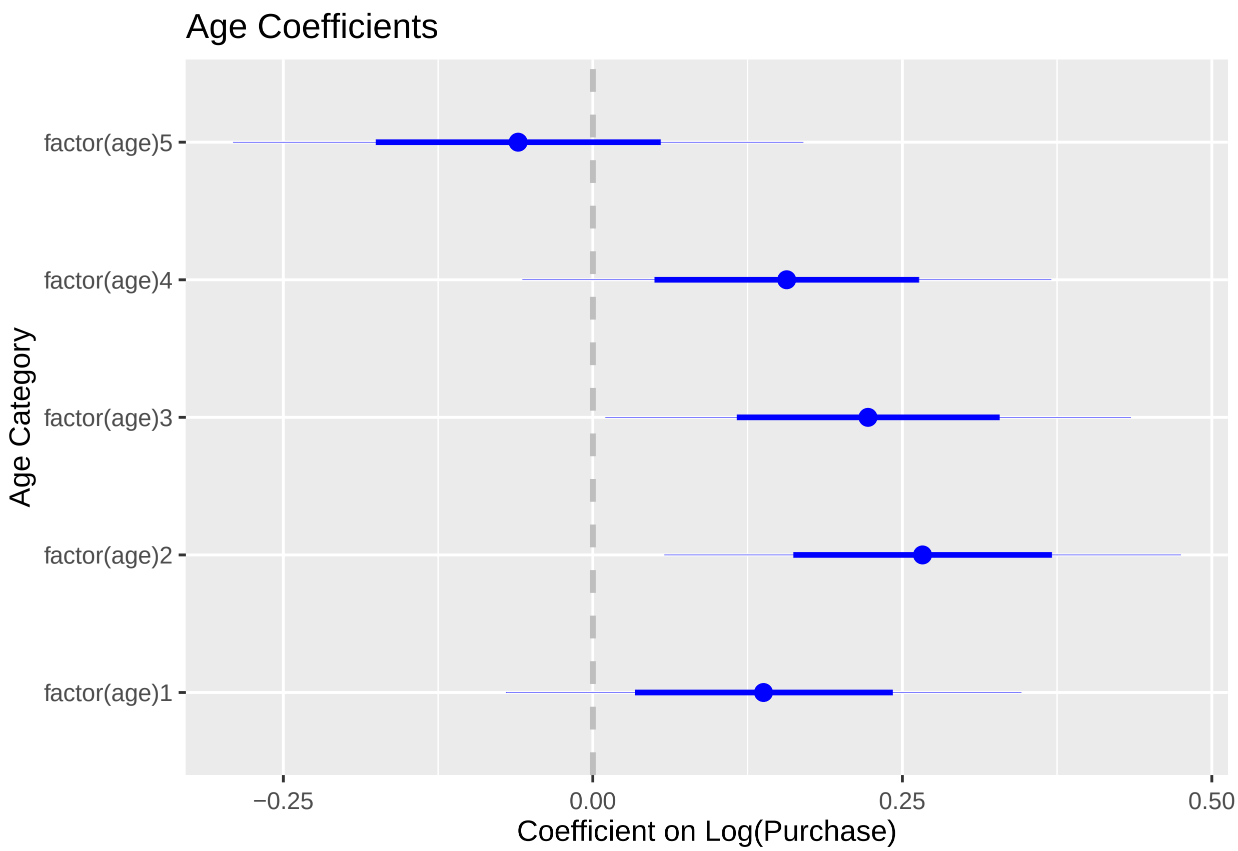
Results

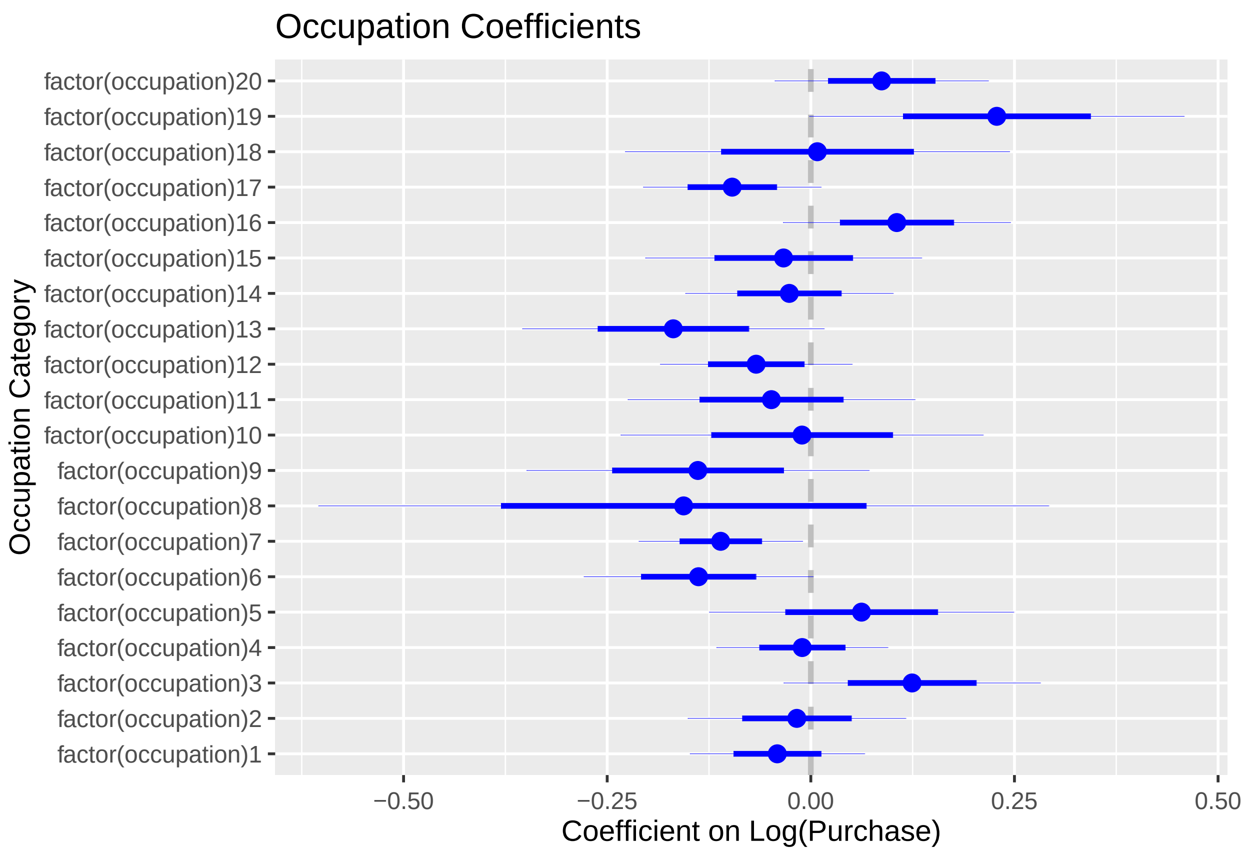
Model Outputs

Linear

```
##
## Call:
```

```
## lm(formula = log(total) ~ factor(gender) + factor(age) + factor(city) +
##     factor(occupation), data = lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81555 -0.69622  0.03695  0.72486  2.49057
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.45843    0.10946  77.277  < 2e-16 ***
## factor(gender)1       0.28456    0.02774  10.260  < 2e-16 ***
## factor(age)1          0.13786    0.10401   1.325   0.1851
## factor(age)2          0.26631    0.10420   2.556   0.0106 *
## factor(age)3          0.22225    0.10601   2.097   0.0361 *
## factor(age)4          0.15659    0.10671   1.467   0.1423
## factor(age)5         -0.06037    0.11505  -0.525   0.5998
## factor(city)2         0.14131    0.03612   3.912 9.26e-05 ***
## factor(city)3        -0.55980    0.03324 -16.841  < 2e-16 ***
## factor(occupation)1  -0.04119    0.05361  -0.768   0.4424
## factor(occupation)2  -0.01721    0.06690  -0.257   0.7970
## factor(occupation)3   0.12429    0.07877   1.578   0.1146
## factor(occupation)4  -0.01059    0.05256  -0.202   0.8403
## factor(occupation)5   0.06222    0.09344   0.666   0.5055
## factor(occupation)6  -0.13790    0.07038  -1.959   0.0501 .
## factor(occupation)7  -0.11069    0.05026  -2.202   0.0277 *
## factor(occupation)8  -0.15618    0.22413  -0.697   0.4859
## factor(occupation)9  -0.13868    0.10509  -1.320   0.1870
## factor(occupation)10 -0.01085    0.11123  -0.098   0.9223
## factor(occupation)11 -0.04841    0.08813  -0.549   0.5828
## factor(occupation)12 -0.06714    0.05893  -1.139   0.2546
## factor(occupation)13 -0.16889    0.09263  -1.823   0.0683 .
## factor(occupation)14 -0.02642    0.06376  -0.414   0.6786
## factor(occupation)15 -0.03350    0.08479  -0.395   0.6928
## factor(occupation)16  0.10555    0.06977   1.513   0.1303
## factor(occupation)17 -0.09655    0.05457  -1.769   0.0769 .
## factor(occupation)18  0.00800    0.11799   0.068   0.9459
## factor(occupation)19  0.22828    0.11500   1.985   0.0472 *
## factor(occupation)20  0.08688    0.06555   1.325   0.1851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.912 on 5849 degrees of freedom
## Multiple R-squared:  0.1492, Adjusted R-squared:  0.1451
## F-statistic: 36.63 on 28 and 5849 DF,  p-value: < 2.2e-16
```


Age Coefficients
factor(age)5
factor(age)4
factor(age)3
factor(age)2
factor(age)1
Age Category
−0.25
0.00
0.25
0.50
Coefficient on Log(Purchase)



####Categorical

```
## Call:
## polr(formula = blackfriday$`Purchase Cat` ~ factor(Age) + factor(Occupation) +
##     factor(City_Category) + factor(Stay_In_Current_City_Years) +
##     factor(Product_Category_1) + factor(Gender), data = blackfriday)
##
## Coefficients:
##                      factor(Age)18-25                     factor(Age)26-35
##                          -0.101116650                         -0.037003297
##                      factor(Age)36-45                     factor(Age)46-50
##                           0.056532918                          0.045903647
##                      factor(Age)51-55                       factor(Age)55+
##                           0.210799843                          0.149640060
##                   factor(Occupation)1                  factor(Occupation)2
##                          -0.011198038                          0.018118631
##                   factor(Occupation)3                  factor(Occupation)4
##                           0.152465809                          0.072052776
##                   factor(Occupation)5                  factor(Occupation)6
##                           0.012752338                          0.134577608
##                   factor(Occupation)7                  factor(Occupation)8
##                           0.068278782                         -0.175863645
##                   factor(Occupation)9                 factor(Occupation)10
##                           0.079819232                         -0.046337338
##                  factor(Occupation)11                 factor(Occupation)12
##                           0.044980883                          0.158714750
##                  factor(Occupation)13                 factor(Occupation)14
##                          0.078731814                          0.102497604
##                 factor(Occupation)15                 factor(Occupation)16
##                          0.229078399                          0.081177579
##                 factor(Occupation)17                 factor(Occupation)18
##                          0.115481147                          0.034300709
##                 factor(Occupation)19                 factor(Occupation)20
##                         -0.197612897                         -0.081796971
##               factor(City_Category)B               factor(City_Category)C
##                          0.092942830                          0.319103560
##  factor(Stay_In_Current_City_Years)1  factor(Stay_In_Current_City_Years)2
##                         -0.006883512                          0.020519626
##  factor(Stay_In_Current_City_Years)3 factor(Stay_In_Current_City_Years)4+
##                         -0.005933184                          0.009722108
##          factor(Product_Category_1)2          factor(Product_Category_1)3
##                         -1.447131161                         -1.620360682
##          factor(Product_Category_1)4          factor(Product_Category_1)5
##                        -61.172533089                         -4.536437060
##          factor(Product_Category_1)6          factor(Product_Category_1)7
##                          0.728660676                          0.918443862
##          factor(Product_Category_1)8          factor(Product_Category_1)9
##                         -4.041803637                         -0.300983106
##         factor(Product_Category_1)10         factor(Product_Category_1)11
##                          1.274706060                         -6.245947652
##         factor(Product_Category_1)12         factor(Product_Category_1)13
##                        -33.314362440                        -19.815175790
##         factor(Product_Category_1)14         factor(Product_Category_1)15
##                         -0.539615140                         -0.142310263
##         factor(Product_Category_1)16         factor(Product_Category_1)17
##                          0.174489912                         -1.493492352
##         factor(Product_Category_1)18                      factor(Gender)M
##                        -35.000370000                         -0.038738476
##
## Intercepts:
##        0|1        1|2        2|3
## -5.6868056 -1.4790812  0.1210516
##
## Residual Deviance: 868297.00
## AIC: 868403.00
```

Multilevel

Results hidden for faster knitting

Interpretation

Linear

Many of our coefficients are not statistically significant, and the R squared value is a very low 0.194. The model does confirm previous observations. According to this model, individuals in the 25-36 age range have a 30% increase in total spending. Customers living in city category 2 are expected to spend 15% more, holding other variables constant. Occupation 19 actually shows the greatest increase in overall spending, although this coefficient is not statistically significant.

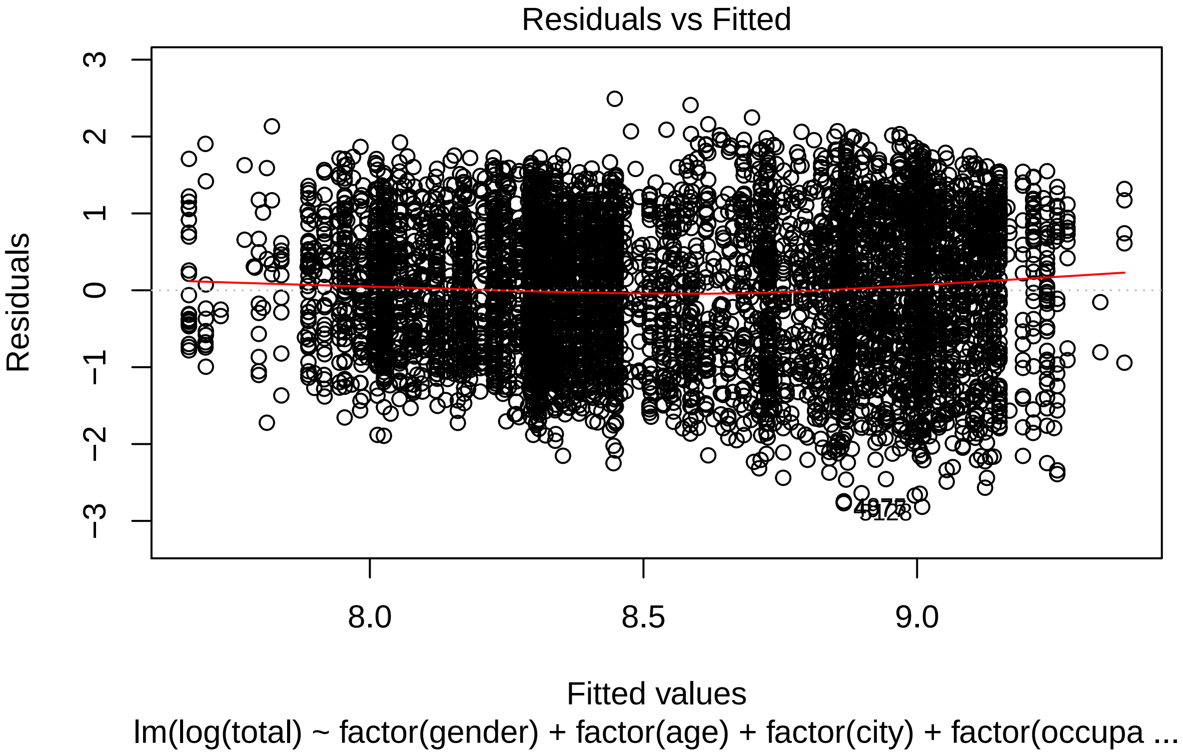
Categorical/Multilevel

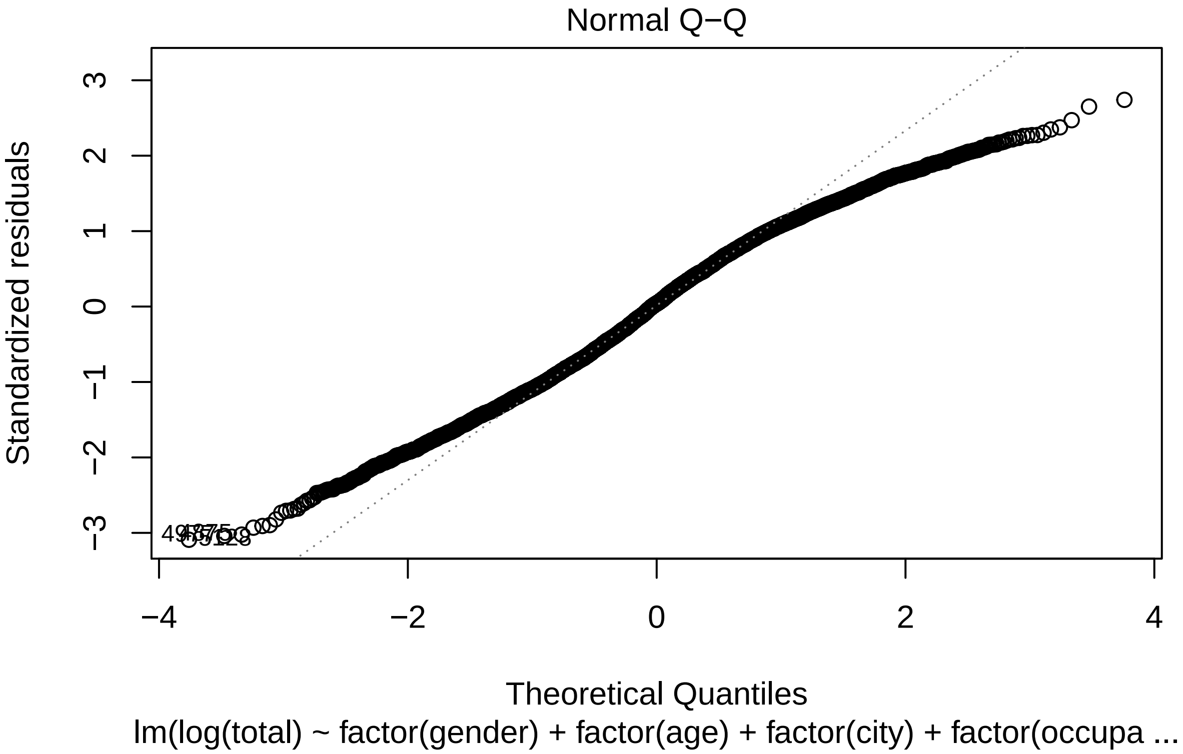
Although the majority of the revenue for this company comes from younger populations, these two models both show that the older customers are the ones more likely to purchase more expensive products. Occupation 15 also tends to buy more expensive products, which fits our previous assumptions. In general, coefficients from these models show similar patterns to the linear model.

Model Checking

Linear

We can see from the residuals and normal qq plots that this model does not fit very well. I decided to split the data into a train and test set for the model. 70% of the data was used to generate a new model using the same variables. The model was then used to predict the total purchase amounts for the rest of the dataset. On average, the model predictions were about 40% higher than the actual amounts spent.





```
## [1] 1.463877
```

Categorical

Using a similar strategy as the linear model, I split the original dataset into a train and test set to test the accuracy of the model choice. The model accuately predicted the purchase amount category for 70% of the purchases. This seems like a pretty good amount, but we have to remember that these categories span a $50 range. Determining who will buy more expensive products is important, but I think the company would want a more precise model than this.

Discussion

I think it is easy to draw conclusions on the customer base of this company. The majority of their customers and revenue comes from a younger, male population. Our models predicting individual item costs show that the older populations will buy more expensive products. We can see this in other variables as well as a certain few occupations drive the majority of business, but a few are more likely to buy more expensive products. The number of products that people purchase can vary drastically. I am going to assume the customer who purchased over a thousand items was either buying something in bulk for an organization, or the entry was a typo. The data poses many of these questions, especially with the actual purchase prices. They follow a strange pattern and the range does not seem realistic with a store of this magnitude. We are also limited in our conclusions with masked variable categories. As far as the model goes, I was not able to create a model that fit well enough to make confident predictions moving forward. That said, I was able to draw conclusions about our population from the models. If I were to move forward, I would like to have information on the product categories and potentially build models to predict what types of products individuals would buy. I would also advice this company to advertise more outside of their target demographic. Although most of their profits come from young males, older customers and females are also willing to spend similar amounts of money when they do shop at this store. There seems to be room to grow in this market. My final guess of what this store is? No idea… but I will guess some type of electronics store.

Acknowledgement

I can't thank my professors enough for a great first semester. I have been a little slow to pick up on skills and concepts, and I appreciate their support and patience through the learning process. Thank you to my family, friends, and former collegues who are supporting my passion and career change. Thank you to Americans for spending crazy amounts of money on black friday, allowing me to showcase my skills as a data scientist.

Reference

Data Source: https://www.kaggle.com/mehdidag/black-friday