Marketing Analytics-Report

Zilei_Zhang&Weicheng_Zhang&Maura_Oray&Gloria_Zheng

12/3/2018

1. Introduction

Every year in America, there is one day where retail stores make the most money. In fact, stores made over $5 billion online from customers on this day in 2017. This date is always the day after Thanksgiving, and it's called "Black Friday." In our dataset, a particular company wants to gain a better understanding about customer purchasing behavior, because Black Friday is the most profitable day of the year. The company also aims to compare different products to see which ones customers are buying more, and also the demographics of the customers. Having this information could be useful for marketing so that they can target particular customer segments for particular products, and also affect company purchasing decisions for the next year.

In this project, we aim to:
* Look at the distribution of customer demographics
* Determine which variables are highly correlated with purchase
* Identify which products are high-selling

The dataset comes from a competition hosted by Analytics Vidhya.

Group Github url: https://github.com/MauraO/Marketing-Analytics-Project (https://github.com/MauraO/Marketing-Analytics-Project)

2. Data Description

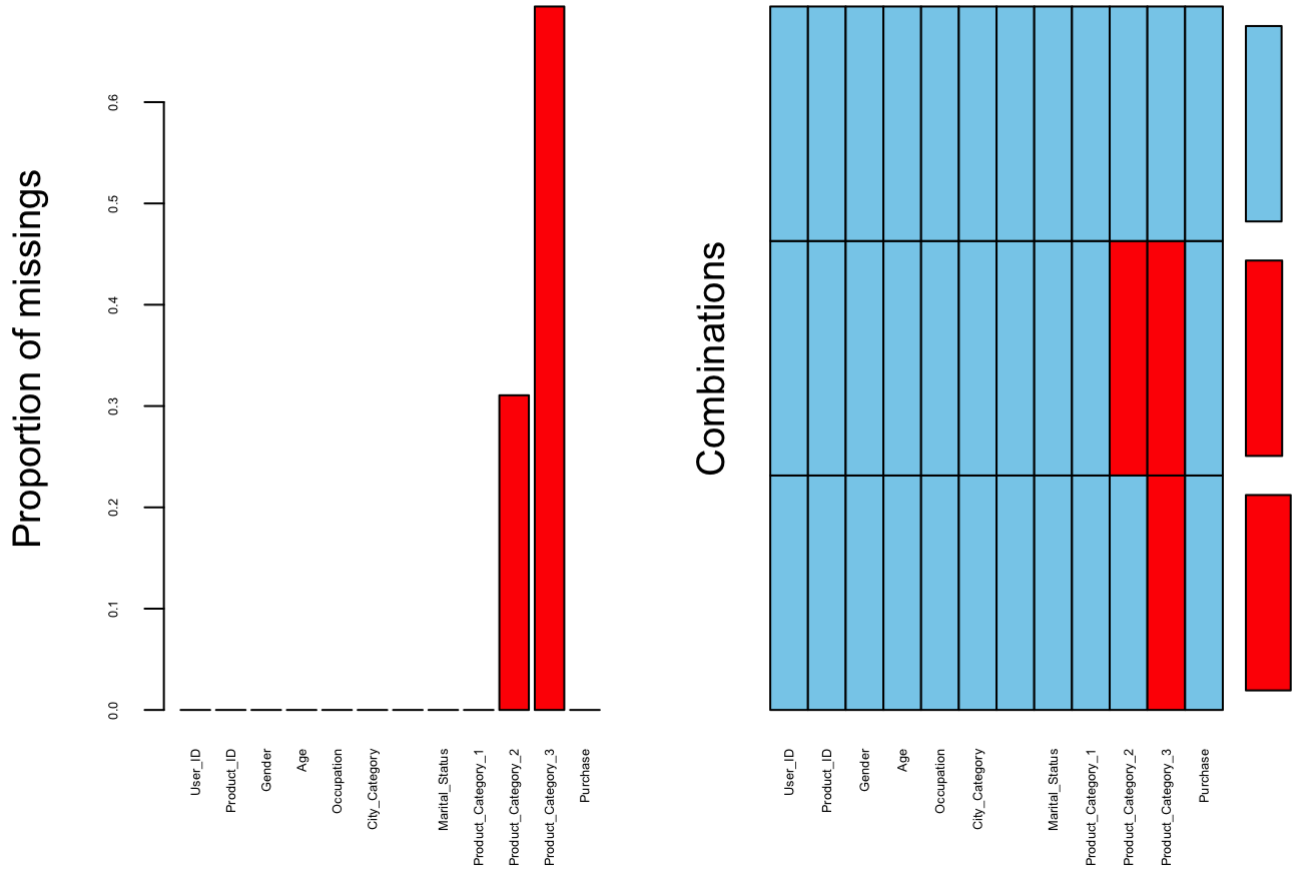
a. Describe the conceptual measure types of the different variables

User_ID (discrete, nominal), Product_ID (discrete, nominal), Gender (discrete, nominal), Age (discrete, ordinal), Occupation (discrete, nominal), City_Category (discrete, nominal), Stay_in_Current_City_Years (continuous, ratio), Marital_Status (discrete, nominal), Product_Category_1 (discrete, nominal), Product_Category_2 (discrete, nominal), Product_Category_3 (discrete, nominal), and Purchase (continuous, ratio).

b. Mention all the steps you took to clean the data

Step 1. Check if the data contains missing values

```
sum(complete.cases(BF))
aggr(BF,cex.axis = .4)
```



In this visualization, we wanted to see which variables had the most NA values. We see that the NA values were only concentrated in Product Category 2 and Product Category 3 (red indicates the location of NA values). Because these two variables are sub categories of Product Category 1 and our further analysis does not use them, we deleted them and made a new data set later.

Step 2. Check the structure of the data

```
summary(BF)
str(BF)
```

In this part, we wanted to see the type of variables for later use (group_by). We found that there were no factors, which leads us to the next step.

Step 3. Type coercion

```
BF[1:11] = lapply(BF[1:11], factor)
```

The majority of the variables were either integers or characters. We decided to change the variables into factors (all except purchase) because the majority of them are categorical, and coercing to factors makes for easier and clearer analysis (group_by).

Step 4. Tidy the data

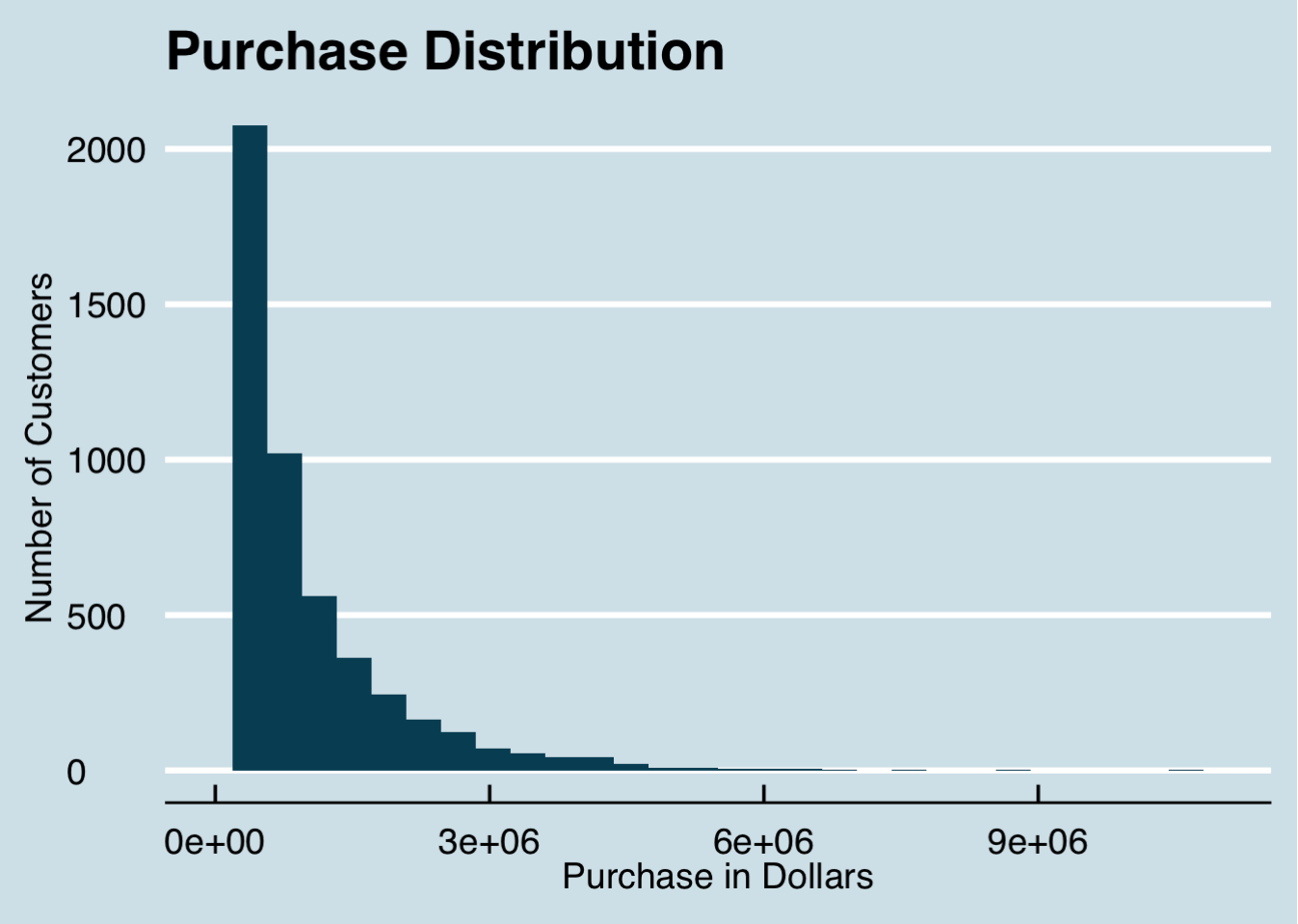
```
#Select demographic data
demo = BF %>%
  select(-starts_with("Product")) %>%
  group_by(User_ID) %>%
  mutate(Purchase = sum(Purchase))
demo = demo[!duplicated(demo$User_ID), ]
```

Because we decided to focus more on customer analysis instead of the product, and that Product Category 2 and Product Category 3 had a lot of missing data, we deleted all of the product category columns to make a clearer customer data set. Moreover, we found that the Customer IDs were repeated many times which is redundant and we needed a purchase sum. Therefore, we used customers' whole purchase amount to replace the individual product purchase and deleted the redundant customer IDs.

3. Summary statistics and Data Visualizations Customer Demographics

Visualization 1. Distribution of Purchase

```
par(mfrow=c(2,2))
demo %>%
  ggplot(aes(Purchase)) +
  geom_histogram(fill = "#014d64") +
  scale_x_continuous(limits = c(0, 11000000)) +
  theme_economist(base_size=14)+
  scale_fill_economist()+
  labs(x = "Purchase in Dollars", y = "Number of Customers", title = "Purchase Distri
bution")
```



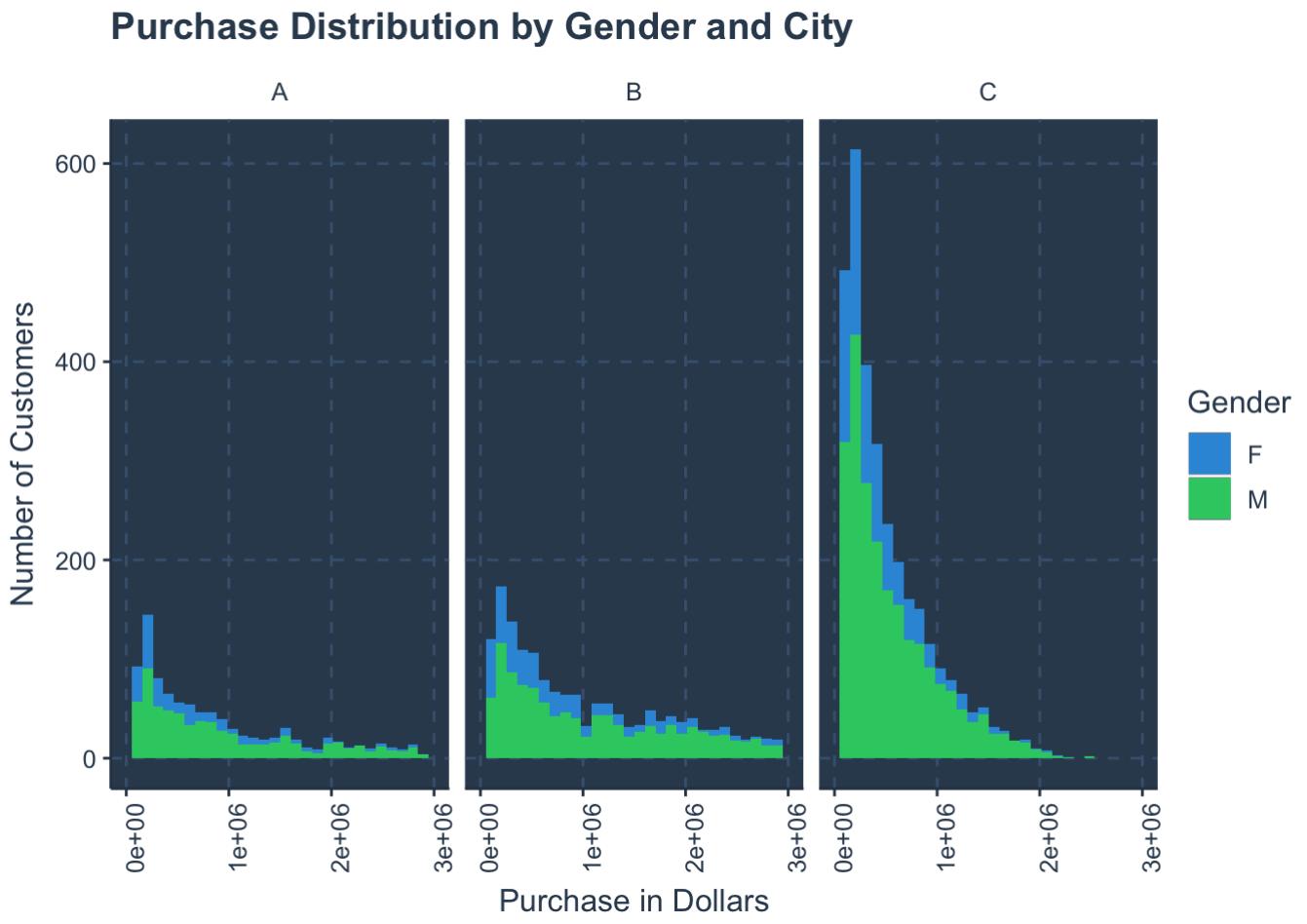
Here we selected the variable purchase. We decided to use a histogram to view the distribution of customers and purchase because it is the best way to display continuous data.

From the graph, it shows a downward slope between number of customers and purchase, which illustrates that a larger number of customers spend less than a smaller number of customers who spend more.

[1]Glorious Christian, Black Friday Analysis https://www.kaggle.com/gloriousc/black-friday-analysis (https://www.kaggle.com/gloriousc/black-friday-analysis)

Visualization 2. Purchase Distribution by Gender and City

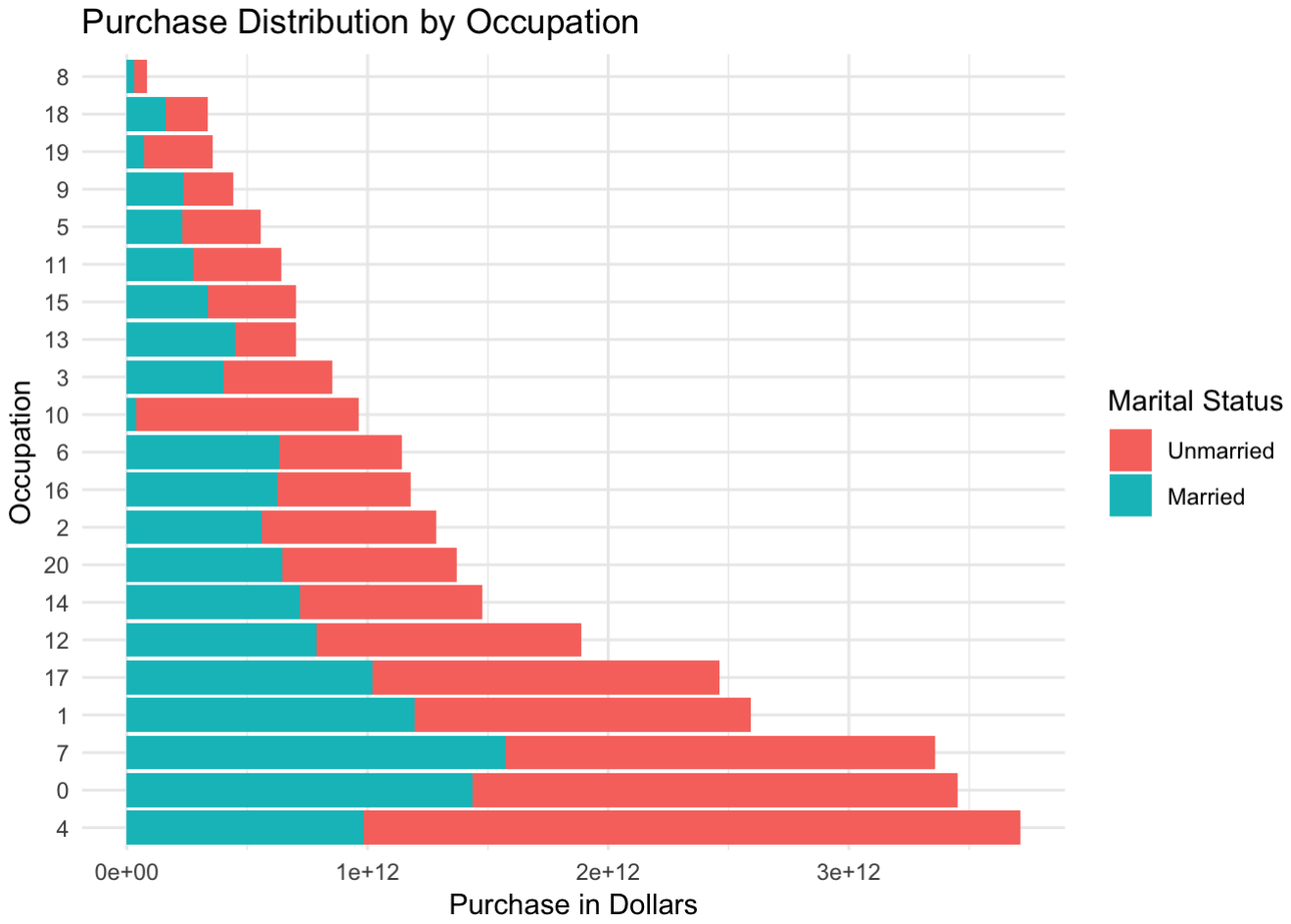
```
ggthemr('flat dark')
demo %>%
  ggplot(aes(x = Purchase, fill = Gender)) +
  geom_histogram() +
  facet_wrap(~City_Category) +
  scale_x_continuous(limits = c(0, 3000000)) +
  labs(x = "Purchase in Dollars", y = "Number of Customers", title = "Purchase Distri
bution by Gender and City")+
  theme(axis.text.x = element_text(angle = 90, hjust = 0.5))
```



Here we selected the variables gender and city in order to view the purchasing distribution of different genders in different cities. We used the histogram visualization because purchase is continuous. We were interested in comparing the purchasing pattern in different cities and genders. We also improved the plot by adding facets which divides purchasing into the 3 cities. It shows that at the same price, there are more customers in city B, and fewer customers in city A. Moreover, at the same price, men spend more than women.

Visualization 3. Purchase Distribution by Occupation

```
#Create new data frame grouping by occupation and marital status, and taking the mean
 purchase of each occupation and status.
ggthemr_reset()
demo %>%
  group_by(Occupation) %>%
  ggplot(aes(x = fct_infreq(Occupation), y = sum(Purchase), fill = Marital_Status)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  coord_flip() +
  labs(x = "Occupation", y = "Purchase in Dollars", title = "Purchase Distribution by
 Occupation", fill="Marital Status") +
    scale_fill_discrete(labels=c("Unmarried", "Married"))
```

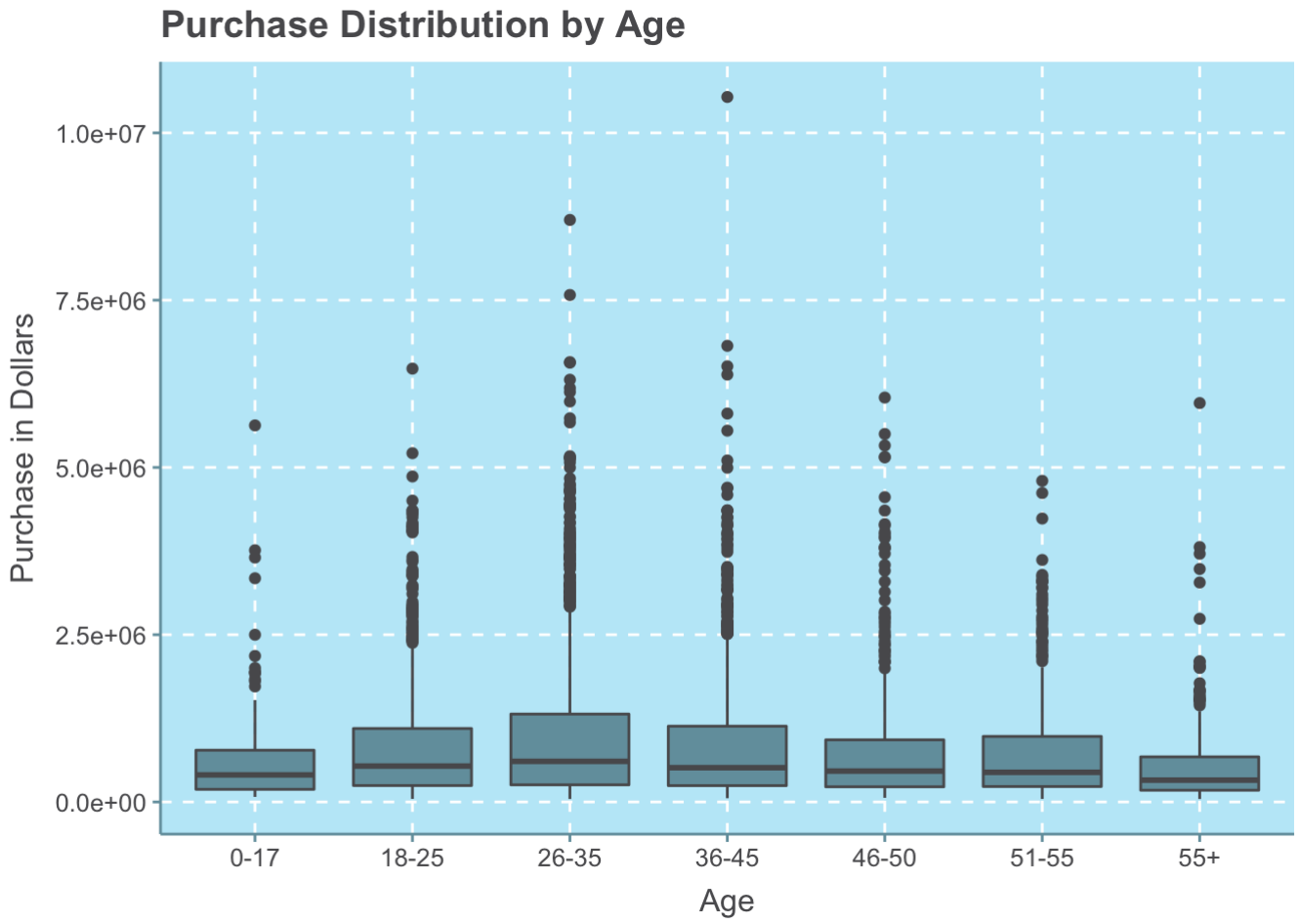


We were interested in determining which occupations spend more and whether married customers spend more than unmarried customers. Since marital status and occupation are factors (discrete) variables, and purchase is a continuous variable, we used a barplot to show total purchase for each occupation, split by matital status. Without looking at the marital status, this barplot shows the purchase distribution among 21 occupations and can answer questions such as “How much do customers from occupation x spend on Black Friday?” Taking into account marital status, the plot answers whether married or unmarried customers spend more. We improved the plot by flipping the coordinates and reordering the purchase by occupation in ascending order.

Generally, unmarried customers spend more on Black Friday and this holds true for most occupations. Specifically, occupations 17, 1, 7, 0 and 4 spend the most money, over $2 trillion, out of which occupation 4 purchases the most. Interestingly, for occupation 10, married customers hardly spend money on Black Friday.

Visualization 4. Purchase Distribution by Age

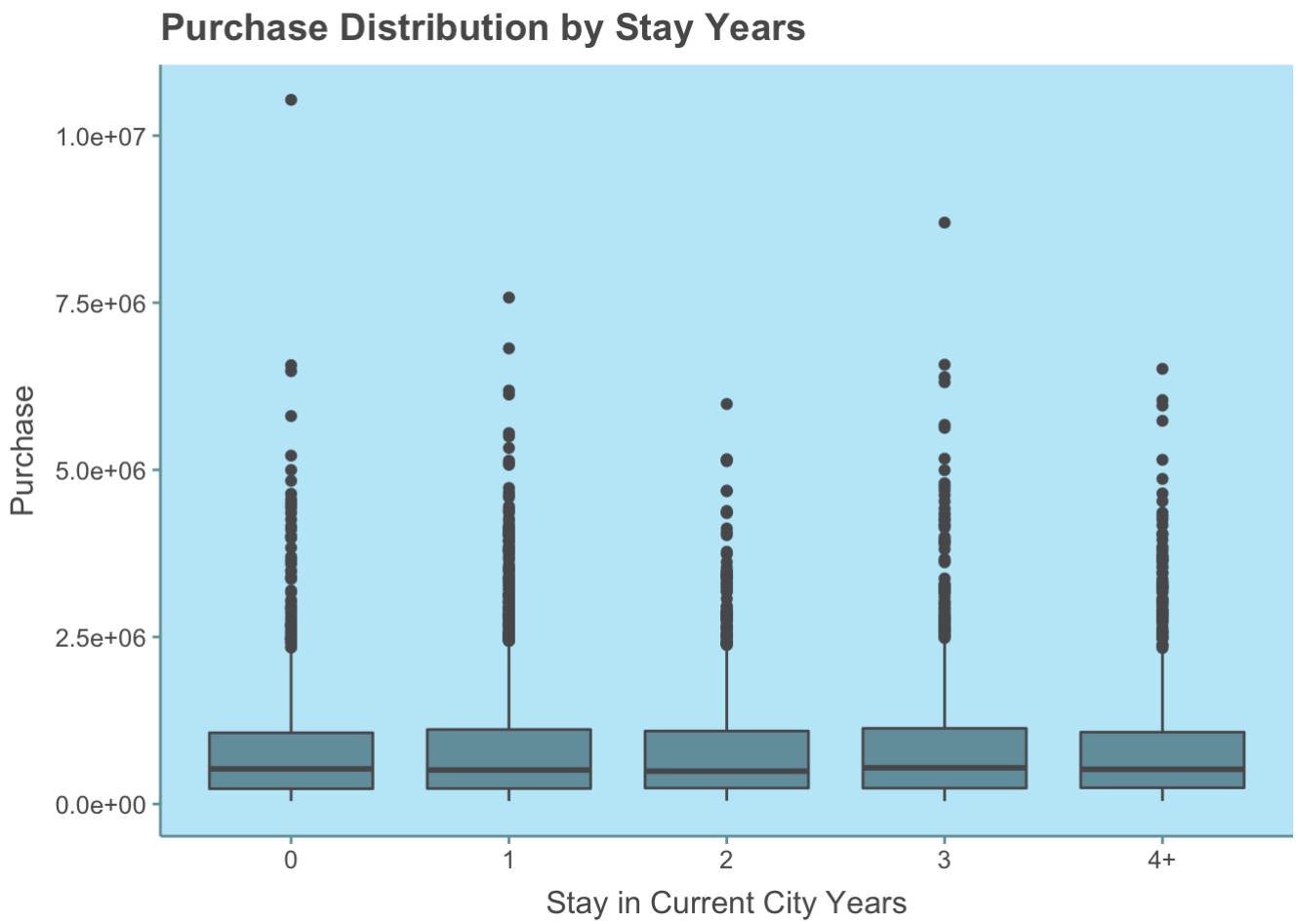
```
ggthemr('sky')
demo %>%
  ggplot(aes(x = Age, y = Purchase)) +
  geom_boxplot() +
  labs(x = "Age", y = "Purchase in Dollars", title = "Purchase Distribution by Age")
```



Here we wanted to determine if there are similar purchasing patterns in different age ranges. A boxplot is the most appropriate visualization because age is discrete and purchase is continuous. We improved the plot by setting the theme as sky. The boxplot can answer the question if is there a difference in median purchase among the 7 groups. Customers aged from 18-55 have similar median purchases except for group 26-35, which has the highest median purchase and range. Customers who are under 17 or over 55, however seem to spend a little bit less.

Visualization 5. Purchase Distribution by Stay in City Years

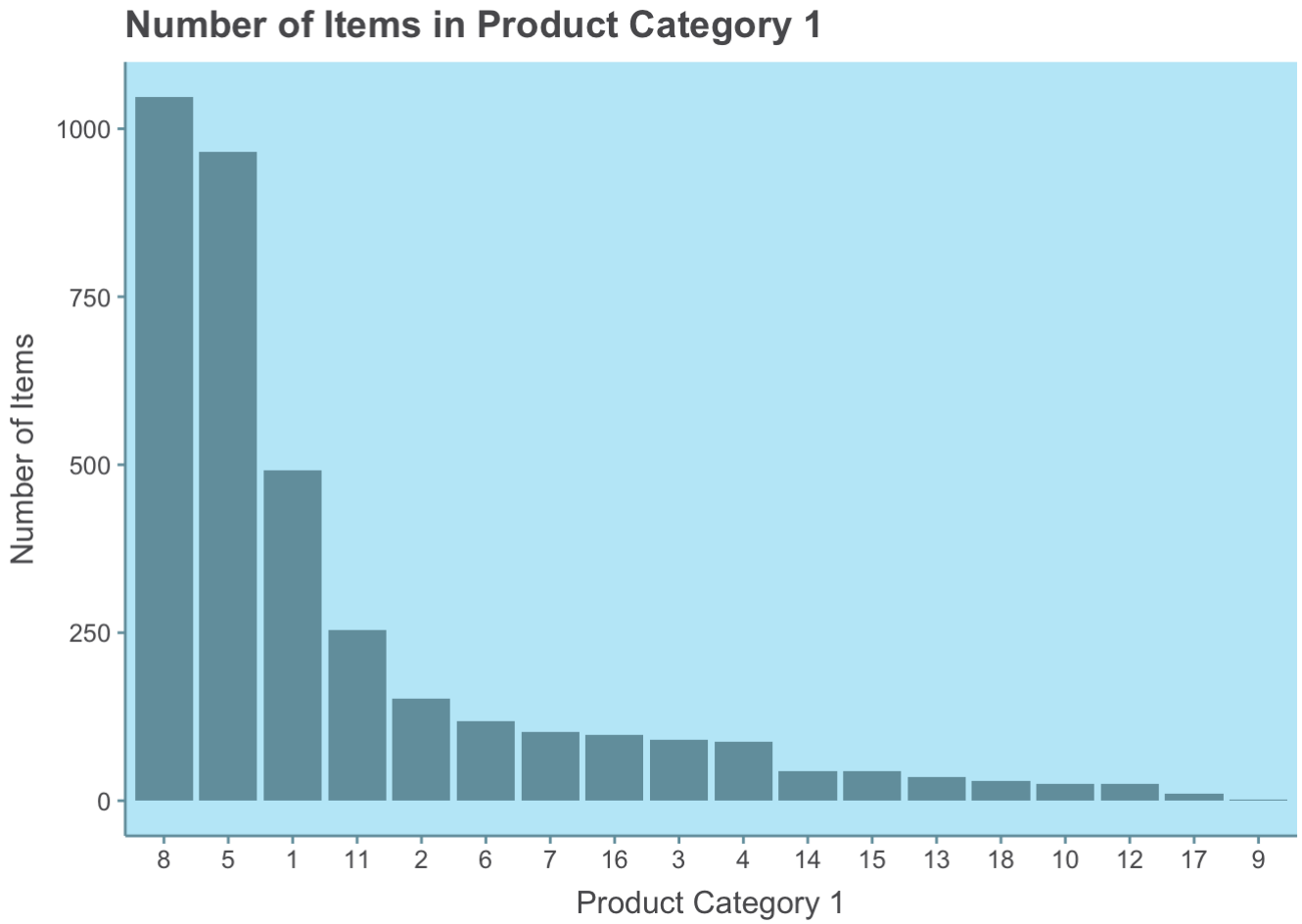
```
ggthemr('sky', layout = 'minimal')
demo %>%
  ggplot(aes(x = Stay_In_Current_City_Years, y = Purchase)) +
  geom_boxplot()+
  labs(x = "Stay in Current City Years", y = "Purchase", title = "Purchase Distributi
on by Stay Years")
```



In the above plot, we wanted to find the purchase distribution by stay in current city years. Because stay in current city years is discrete and purchase is continuous, we chose a boxplot visualization. We used a different theme to make the plot more visually appealing and added labels and titles to make it clearer. From this plot, we found that there may be no difference among the purchase distribution of different current city years stay. But we wanted to have more evidence to draw such conclusion. We used the ANOVA test in the next part to determine whether this difference is significant. Based on the above plot, we drew the conclusion that stay in current city years may not influence the purchase and the store may not need to take this factor into consideration when evaluating marketing decisions.

Visualization 6. Number of Items in Each Product Category

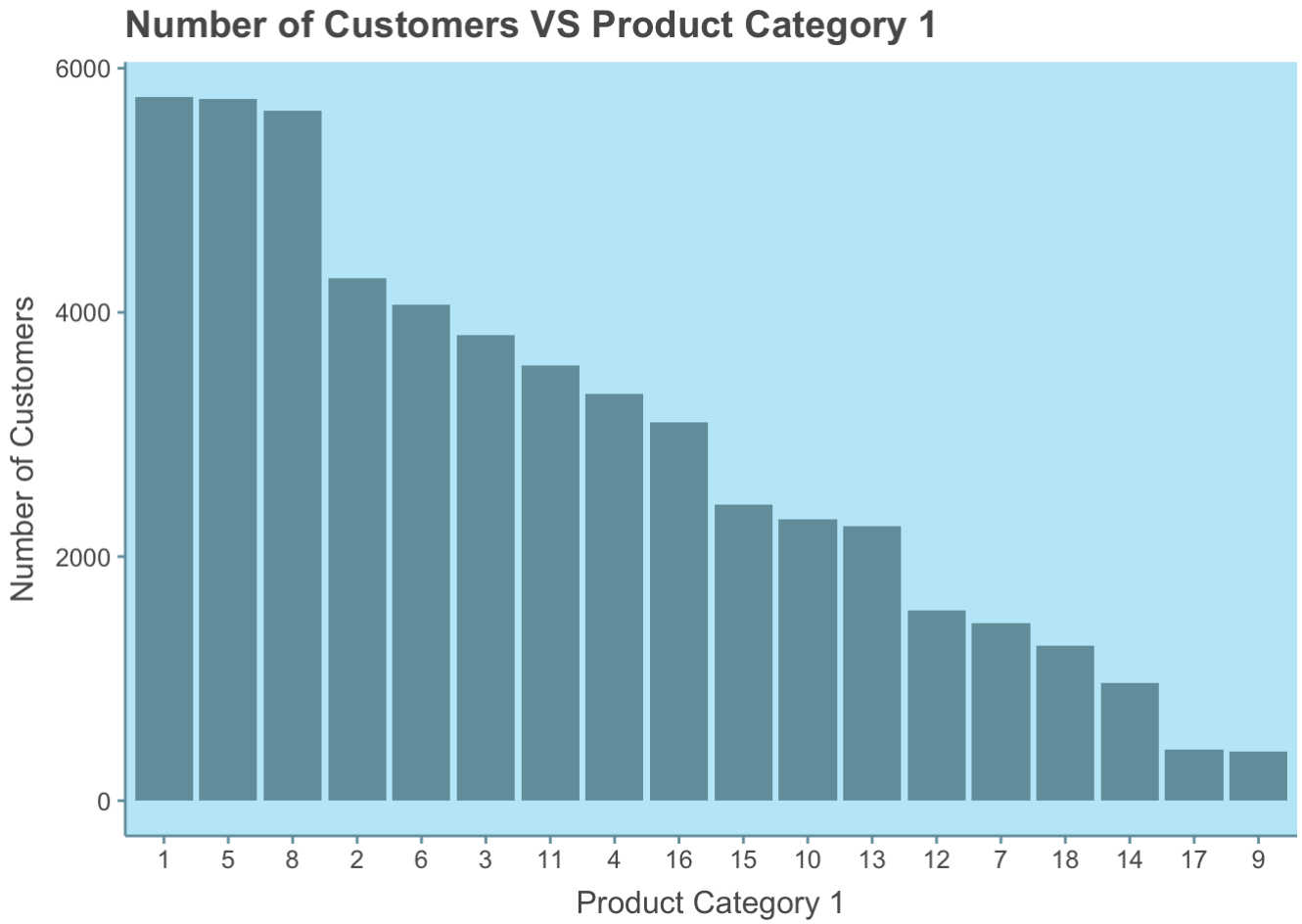
```
# Number of Items VS Product Category 1
BF %>% group_by(Product_Category_1) %>%
  summarise(num_items=n_distinct(Product_ID)) %>%
  ggplot(aes(x = reorder(factor(Product_Category_1), -num_items), y=num_items))+
  geom_bar(stat = 'identity')+
  labs(x ="Product Category 1" , y = "Number of Items", title = "Number of Items in P
roduct Category 1")
```



The assumptions here were that the quantity of distinct products within a certain category lead to popularity of this category and cheaper products have more substitutes. To verify our aforementioned guess, we came to the question that how many products are contained in each category. Therefore, we plotted the quantity of distinct products on the vertical axis. The output showed that category 8 stored contained the most products, followed by category 5 and 1. Now we saw why categories 1, 5 and 8 were so popular!

Visualization 7. The Number of Customers in Product Category

```
# Number of Customers VS Product Category 1
BF %>% group_by(Product_Category_1) %>%
  summarise(num_customer=n_distinct(User_ID)) %>%
  ggplot(aes(x = reorder(factor(Product_Category_1), -num_customer), y=num_customer))
+
  geom_bar(stat = 'identity')+
  labs(x ="Product Category 1" , y = "Number of Customers", title = "Number of Custom
ers VS Product Category 1")
```

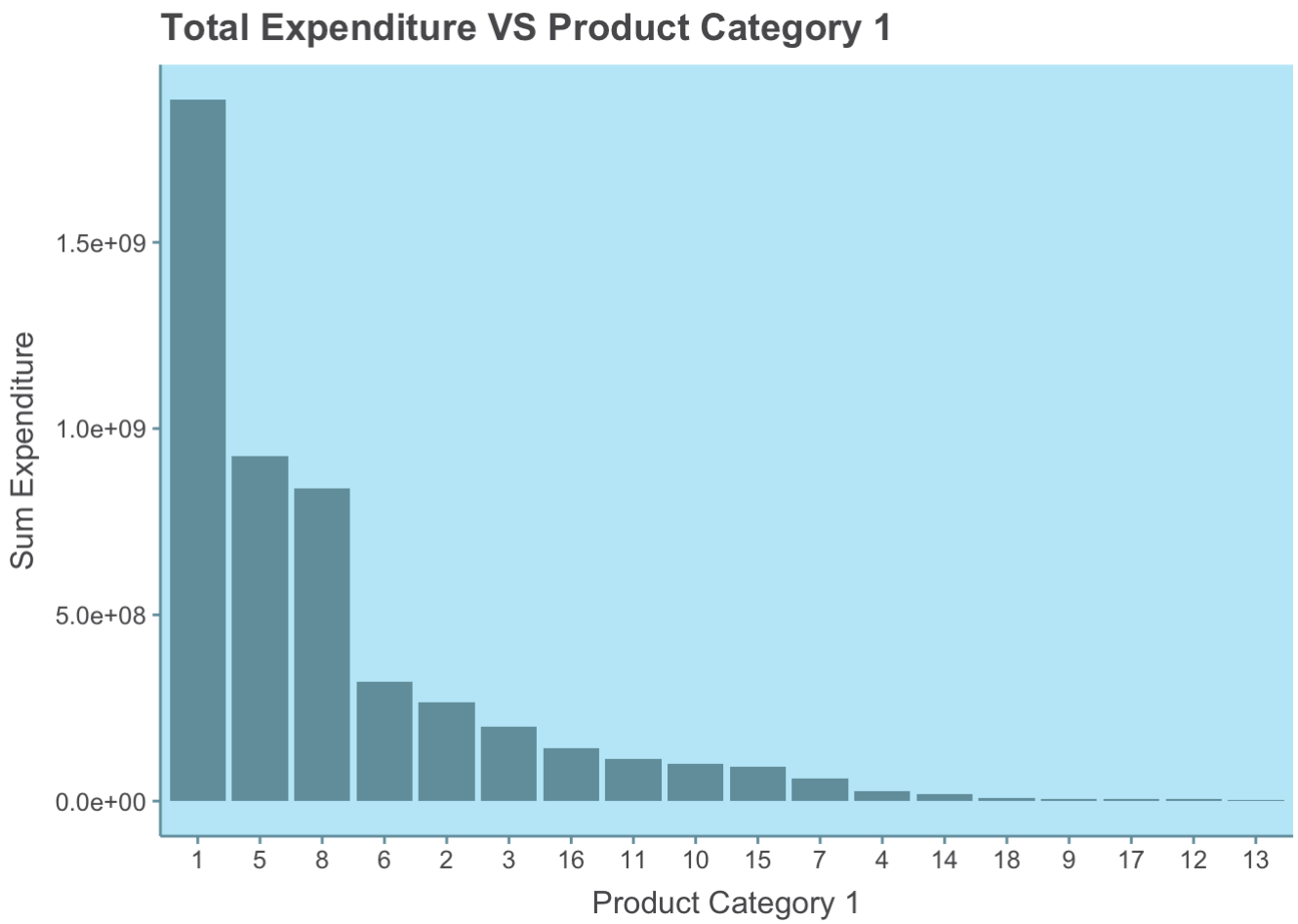


We were also interested in the customer distribution versus product category, e.g., how many distinct customers have bought from a specific category. The number of distinct customers were counted for each category and plotted as the y variable. The plot showed that categories 1,5 and 8 were still the three categories that had the most visits.

Bases on the series of plots, our conclusion is that the three most popular product categories have the most product items and customers.

Visualization 8. The Most Profitable Product Category

```
# Average Expenditure VS Product Category 1
BF %>% group_by(Product_Category_1) %>%
  summarise(sum_purchase = sum(Purchase)) %>%
  ggplot(aes(x = reorder(factor(Product_Category_1), -sum_purchase), y=sum_purchase))
+
  geom_bar(stat = 'identity')+
  labs(x ="Product Category 1" , y = "Sum Expenditure", title = "Total Expenditure VS
 Product Category 1")
```



Profitability, measured by the sum purchase spent in each category, was another concern in addition to the number of items and the number of customers in each category. Therefore, we summed the purchased amount for each category and plotted it against each category. The above bar plot showed that the top 3 categories that customers spent the most money on were still categories 5, 1 and 8. Hence, we came to the conclusion that categories 5,1 and 8 were still the most profitable categories.

4. Preliminary statistical analyses

Test 1. Do men spend more money on Black Friday than women?

```
#Independent t-test
t.test(demo[demo$Gender == "M", ]$Purchase,
       demo[demo$Gender == "F", ]$Purchase,
       paired = FALSE,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  demo[demo$Gender == "M", ]$Purchase and demo[demo$Gender == "F", ]$Purchase
## t = 8.6542, df = 3709.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  172432.4      Inf
## sample estimates:
## mean of x mean of y
##  911963.2  699054.0
```

In the above code, we wanted to determine whether gender influences purchase. In table 1 and plot 2, we find that there are more men than women and that men make more purchases than women. Because gender is discrete and purchase is continuous, we chose to apply a t-test to determine whether there is a difference between men and women in purchase.

Based on the evidence from this data set and the result from the t-test (p-value < 2.2e-16), we rejected the null hypothesis and accepted the alternative hypothesis: that men have a larger mean purchase amount than women.

From the statistical analysis, we recommend that the store should focus more on men because they have higher purchasing power than women and for the store to have more methods of promotion towards men in order to increase their revenue.

Test 2. Do unmarried customers spend more money on Black Friday than married?

```
#Independent t-test
t.test(demo[demo$Marital_Status == 0, ]$Purchase,
       demo[demo$Marital_Status == 1, ]$Purchase,
       paired = FALSE,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  demo[demo$Marital_Status == 0, ]$Purchase and demo[demo$Marital_Status ==
1, ]$Purchase
## t = 1.5859, df = 5392.3, p-value = 0.05641
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1454.306       Inf
## sample estimates:
## mean of x mean of y
##  868097.6  829175.0
```

In this instance, we wanted to determine whether marital status influenced purchase. In table 1 and visualization 3, we found that there were more unmarried people than married, and that unmarried customers seemed to make more purchases than married customers. Because marital status is categorical and purchase is continuous, we applied an independent t-test to test whether this difference was significant.

Based on the evidence from this data set and the result from the t-test (p-value = 0.056), we drew the conclusion that in the 95% confidence interval, we cannot reject the null hypothesis. However, in the 90% confidence interval, we can reject the null hypothesis and accept the alternative hypothesis: that unmarried customers have larger mean purchase amounts than married customers.

Because the mean purchase amount was higher for unmarried people than married people, the company's marketing department could focus on targeting single people in order to increase purchase amounts. However, it would be wise not to focus all marketing dollars on single people, since the difference between the mean purchase amount for unmarried versus married is not stark.

Test 3. Is there an association between city and purchase?

```
# Mutate a new column
demo = demo %>%
  mutate(purchase_category = cut(Purchase, breaks = c(-Inf, 304987.6, 826809.6, Inf),
  labels = c("Low", "Medium", "High")))
table(demo$City_Category, demo$purchase_category)
```

```
##
##       Low Medium High
##   A  276    287  482
##   B  360    460  887
##   C 1308   1197  634
```

```
# Apply Chi-Square test
chisq.test(table(demo$City_Category, demo$purchase_category))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(demo$City_Category, demo$purchase_category)
## X-squared = 595.38, df = 4, p-value < 2.2e-16
```

In the above code, we aimed to see whether city stay influences purchase amounts. In visualization 2, we saw the differences in purchase distribution. Because city stay is a categorical variable, it was appropriate to use a Chi-Square test to determine whether there was significant differences in purchase between cities.

Based on the evidence from this data set and the result from the Chi-Square test (p-value < 2.2e-16), we believe that the probability that there is no difference among the three cities is less than 5%, and that the probability that there is a difference among the three cities is greater than 95%. Therefore, we rejected the null hypothesis and accepted the alternative hypothesis: that there is a difference in purchase distribution between the three cities. Therefore, from this data set, we conclude that there is an association between city and purchase.

From a business perspective, it would be wise for the company to analyze which cities are more profitable, so that it can appropriately direct marketing dollars towards either the more profitable or less profitable cities, depending on the marketing campaign.

Test 4. Is there an association between Age and purchase?

```
# Chi-Square test
# Create purchase categories
demo = demo %>%
  mutate(purchase_category = cut(Purchase, breaks = c(-Inf, 304987.6, 826809.6, Inf),
  labels = c("low", "medium", "high")))
# Create two-way frequency table
table(demo$Age, demo$purchase_category)
```

```
##
##         low medium high
##   0-17   90     75   53
##   18-25 345    365  359
##   26-35 599    640  814
##   36-45 380    378  409
##   46-50 182    197  152
##   51-55 171    164  146
##   55+   177    125   70
```

```
# Apply Chi-Square test
chisq.test(table(demo$Age, demo$purchase_category))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(demo$Age, demo$purchase_category)
## X-squared = 102.05, df = 12, p-value < 2.2e-16
```

In visualization 4, there was a slight difference between average purchase amounts in the 7 age bins. Therefore, we decided to test if the difference was signficant. Since both age and purchase categories are discrete variables with more than 2 categories, we applied the Chi-Square test. This test would be able to determine if there was an association between age range and purchase levels. We added another variable, purchase category, which would show purchase amounts as Low, Medium, or High. The result showed that the p-value for this test was 0 so we rejected the null hypothesis. Therefore, we conclude that age range did affect purchase levels on Black Friday. For example, in group 0-17, the higher the purchase level, the less amount of customers, while in group 26-35 the number of custmers increased with the price levels.

Test 5. Is there an association between staying years and purchase?

```
stay = lm(Purchase ~ Stay_In_Current_City_Years, data = demo)
anova(stay)
```

```
## Analysis of Variance Table
##
## Response: Purchase
##                              Df     Sum Sq    Mean Sq F value Pr(>F)
## Stay_In_Current_City_Years    4 3.3427e+12 8.3566e+11    0.96 0.4282
## Residuals                  5886 5.1238e+15 8.7051e+11
```

```
pairwise.t.test(demo$Purchase, demo$Stay_In_Current_City_Years, p.adjust.method = "bo
nferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  demo$Purchase and demo$Stay_In_Current_City_Years
##
##    0    1    2    3
## 1  1.00 -    -    -
## 2  1.00 1.00 -    -
## 3  1.00 1.00 0.65 -
## 4+ 1.00 1.00 1.00 1.00
##
## P value adjustment method: bonferroni
```

The boxplot of stay in current city years exhibited highly similar patterns for each subgroup. Therefore, we were interested in testing the relationship between stay in a certain city with Black Friday purchase amounts at the population level. Stay is a factor variable with 5 categories and purchase is a continuous variable, so we used analysis of variance (ANOVA), which is the extension of the t-test and can be used to test whether there are significant differences in more than 2 population means (Ralph, 2010). However, since ANOVA does not tell how the mean of one population's group is different than the others', we still needed to use a pairwise t-test. A pairwise t-test enabled us to compare multiple groups' means with the correlation pairwisely. The p-value table showed that the differences in average purchase amounts were not significantly different regardless of

how long the customer had been living there. We deduced this because the all the p-values were greater than 0.05. Therefore, we conclude that stay in current city years did not influence customers' purchasing behavior on Black Friday.

5. Regression Analyses

a. The Baseline Regression Model:

In part 4, we determined that gender, city, and age were three significant factors in purchase, and that marital status was marginally significant. Therefore, we used the first three variables as the first baseline regressors and built our model.

b. Dependent and Independent Variables

The dependent variable in our model was log(purchase), and independent variables were: gender, age, city category.

c. Data Aggregations

We used total purchase amounts for each distinct user ID. The cleaned data set "demo" (as shown previously) was the data aggregation where we looked at overall expenditures for each customer. We then used customers' total purchase to replace the individual product purchase.

d. Results of the Regression

```
lm_multiple_1=lm(log(Purchase)~Gender+Age+City_Category,data=demo)
summary(lm_multiple_1)
```

```
##
## Call:
## lm(formula = log(Purchase) ~ Gender + Age + City_Category, data = demo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81970 -0.70895  0.02868  0.72659  2.60097
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.09432    0.07028 186.321  < 2e-16 ***
## GenderM         0.28326    0.02662  10.642  < 2e-16 ***
## Age18-25        0.13192    0.06840   1.929 0.053832 .
## Age26-35        0.24104    0.06565   3.672 0.000243 ***
## Age36-45        0.19182    0.06782   2.829 0.004691 **
## Age46-50        0.14561    0.07387   1.971 0.048750 *
## Age51-55        0.09652    0.07501   1.287 0.198219
## Age55+         -0.12681    0.07838  -1.618 0.105749
## City_CategoryB  0.10912    0.03613   3.020 0.002541 **
## City_CategoryC -0.59685    0.03313 -18.016  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9182 on 5881 degrees of freedom
## Multiple R-squared:  0.1458, Adjusted R-squared:  0.1445
## F-statistic: 111.5 on 9 and 5881 DF,  p-value: < 2.2e-16
```

The output of the baseline regression model showed that gender, age and city category had a significant effect on purchase amount.

Men purchased 28.30% on Black Friday than women; customers aged from 26-35 bought the most on Black Friday, followed by customers in the age ranges of 36-45 and 45-50; customers from city B purchased the most followed by customers from city A and then C. Due to the low R-square coefficient (0.1458), we decided to focus more on the qualitative analysis rather than interpreting the coefficients of the regressors.

e. Analyzing Controlling for omitted variable biases

The definition of an omitted variable is that: (1) the omitted variable is correlated with the included regressors and (2) the omitted variable is a determinant of the dependent variable. To verify the presence of omitted variable biases, we added one more non-included variable to see if the aforementioned effects would occur. The variable we added was used as the control, which can be defined as the omitted variable if it significantly differs with 0 and affects the coefficients of other regressors.

Firstly, we added marital status as the control variable and re-performed the original regression:

```
lm_multiple_2=lm(log(Purchase)~Gender+Age+City_Category+Marital_Status,data=demo)
summary(lm_multiple_2)
```

```
##
## Call:
## lm(formula = log(Purchase) ~ Gender + Age + City_Category + Marital_Status,
##     data = demo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82117 -0.71035  0.02976  0.72647  2.59951
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.094247   0.070286 186.300  < 2e-16 ***
## GenderM          0.283218   0.026620  10.639  < 2e-16 ***
## Age18-25         0.132847   0.068676   1.934 0.053111 .
## Age26-35         0.242624   0.066465   3.650 0.000264 ***
## Age36-45         0.193409   0.068606   2.819 0.004832 **
## Age46-50         0.148422   0.076121   1.950 0.051245 .
## Age51-55         0.099383   0.077308   1.286 0.198650
## Age55+          -0.124248   0.080149  -1.550 0.121143
## City_CategoryB   0.109212   0.036143   3.022 0.002525 **
## City_CategoryC  -0.596725   0.033142 -18.005  < 2e-16 ***
## Marital_Status1 -0.003977   0.025974  -0.153 0.878312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9183 on 5880 degrees of freedom
## Multiple R-squared:  0.1458, Adjusted R-squared:  0.1444
## F-statistic: 100.4 on 10 and 5880 DF,  p-value: < 2.2e-16
```

According to the result, we found that marital status was not significant in this model. In other words, marital status was not a determinent of log(Purchase). Moreover, the R-square had no change, meaning that marital status can not explain the variance in dependent variable. Therefore, we can draw the conclusion that the marital status was not an omitted variable.

Secondly, we added stay in current city years as the control variable and re-performed the regression:

```
lm_multiple_3=lm(log(Purchase)~Gender+Age+City_Category+Stay_In_Current_City_Years,da
ta=demo)
summary(lm_multiple_3)
```

```
##
## Call:
## lm(formula = log(Purchase) ~ Gender + Age + City_Category + Stay_In_Current_City_Y
ears,
##     data = demo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82014 -0.71155  0.02898  0.72778  2.60202
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.094371   0.076510 171.147  < 2e-16 ***
## GenderM                      0.283344   0.026631  10.640  < 2e-16 ***
## Age18-25                     0.131228   0.068442   1.917 0.055241 .
## Age26-35                     0.240271   0.065682   3.658 0.000256 ***
## Age36-45                     0.190651   0.067850   2.810 0.004972 **
## Age46-50                     0.145197   0.073899   1.965 0.049483 *
## Age51-55                     0.097094   0.075069   1.293 0.195926
## Age55+                      -0.127698   0.078415  -1.628 0.103475
## City_CategoryB               0.109312   0.036157   3.023 0.002511 **
## City_CategoryC              -0.596424   0.033146 -17.994  < 2e-16 ***
## Stay_In_Current_City_Years1 -0.006991   0.038709  -0.181 0.856689
## Stay_In_Current_City_Years2 -0.008210   0.042796  -0.192 0.847876
## Stay_In_Current_City_Years3  0.025185   0.044237   0.569 0.569157
## Stay_In_Current_City_Years4+ 0.001077   0.044967   0.024 0.980896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9184 on 5877 degrees of freedom
## Multiple R-squared:  0.1459, Adjusted R-squared:  0.1441
## F-statistic: 77.25 on 13 and 5877 DF,  p-value: < 2.2e-16
```

According to the result, we found that stay in current city years was not significant in this model. In other words, stay in current city years was not a determinant of log(Purchase). Moreover, the R-square value was unchanged, indicating that stay did not explain the variance in the dependent variable. Therefore, we could say that marital status was not an omitted variable.

Next, we added occupation as the control variable and re-performed the original regression:

```
lm_multiple_4=lm(log(Purchase)~Gender+Age+City_Category+Occupation,data=demo)
summary(lm_multiple_4)
```

```
##
## Call:
## lm(formula = log(Purchase) ~ Gender + Age + City_Category + Occupation,
##     data = demo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.7012  0.0340  0.7249  2.4684
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.084834   0.110007 118.946  < 2e-16 ***
## GenderM         0.289802   0.027873  10.397  < 2e-16 ***
## Age18-25        0.134545   0.104566   1.287  0.19825
## Age26-35        0.268525   0.104748   2.564  0.01039 *
## Age36-45        0.226224   0.106563   2.123  0.03380 *
## Age46-50        0.177767   0.110631   1.607  0.10814
## Age51-55        0.131437   0.111639   1.177  0.23911
## Age55+         -0.064950   0.115676  -0.561  0.57449
## City_CategoryB  0.113109   0.036178   3.126  0.00178 **
## City_CategoryC -0.588464   0.033269 -17.688  < 2e-16 ***
## Occupation1    -0.043020   0.053883  -0.798  0.42467
## Occupation2    -0.022637   0.067243  -0.337  0.73640
## Occupation3     0.134589   0.078992   1.704  0.08847 .
## Occupation4    -0.005742   0.052788  -0.109  0.91339
## Occupation5     0.060227   0.093929   0.641  0.52142
## Occupation6    -0.126730   0.070661  -1.793  0.07295 .
## Occupation7    -0.106728   0.050483  -2.114  0.03455 *
## Occupation8    -0.158636   0.225338  -0.704  0.48147
## Occupation9    -0.135811   0.105643  -1.286  0.19864
## Occupation10   -0.009926   0.111820  -0.089  0.92927
## Occupation11   -0.050837   0.088592  -0.574  0.56610
## Occupation12   -0.072188   0.059225  -1.219  0.22295
## Occupation13   -0.161096   0.093329  -1.726  0.08438 .
## Occupation14   -0.022368   0.064006  -0.349  0.72675
## Occupation15   -0.037977   0.085234  -0.446  0.65593
## Occupation16    0.115777   0.070050   1.653  0.09843 .
## Occupation17   -0.090546   0.054776  -1.653  0.09838 .
## Occupation18    0.038598   0.117809   0.328  0.74320
## Occupation19    0.226482   0.115609   1.959  0.05016 .
## Occupation20    0.091990   0.065808   1.398  0.16221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9168 on 5861 degrees of freedom
## Multiple R-squared:  0.1512, Adjusted R-squared:  0.147
## F-statistic: 36.01 on 29 and 5861 DF,  p-value: < 2.2e-16
```

Compared with the result of the original regression, we found that some occupations (3,6,7,13,16,17,19) were significant at the 90% significance level. That is to say, occupation was correlated with purchase amount to some degree. More importantly, customers aged from 18-25 and 46-50 were not as marginally significant as previously stated. Occupation was also correlated with age in some sense, so we can say that occupation was an omitted variable. This is also cemented by the increase in the R-square value, from 0.1458 to 0.1512.

f. Additional Data Aggregations

Another data aggregation level we used to do regressions was the user-product category level. The product purchase amount in the same category was aggregated for each customer. The purchase amount at the user

level was customer-specific while at the user-category level the purchase amount was customer- and product-category specific. At this data aggregation level we can answer questions from a more pointed and directed perspective: given a certain product category, how do demographic factors affect the purchase amount? Which product category do men from city A aged from 18-25 purchase more? Or given a particular category, which group of customers bought more?

```
BF_cat = BF %>%
  group_by(Product_Category_1, User_ID) %>%
  mutate(sumpp = sum(Purchase))
BF_cat = BF_cat[-c(2)]
BF_cat = BF_cat[-c(9:11)]
BF_cat = BF_cat[!duplicated(BF_cat), ]
re_f <- lm(log(sumpp)~Age*Product_Category_1+Gender*Product_Category_1+City_Category*
Product_Category_1+Occupation*Product_Category_1, data = BF_cat)
#summary(re_f)
```

Here we took product category 5 as an example because it was a popular category. The result shows that customers who were older than 55 purchased significantly less than customers in other age ranges, and that customers who were from city C spent less money in this category than their counterparts.

We were also interested in product category 9. The output shows that all the demographic regressors were not significant. We interpreted this as that regardless of their origin city, or whether they were young or old or male or female, there was no significant difference in purchase amount for this category. Therefore, we assumed that this product could be a basic necessity that all customers would buy, such as a toothbrush or toothpaste.

6. Segmentation:

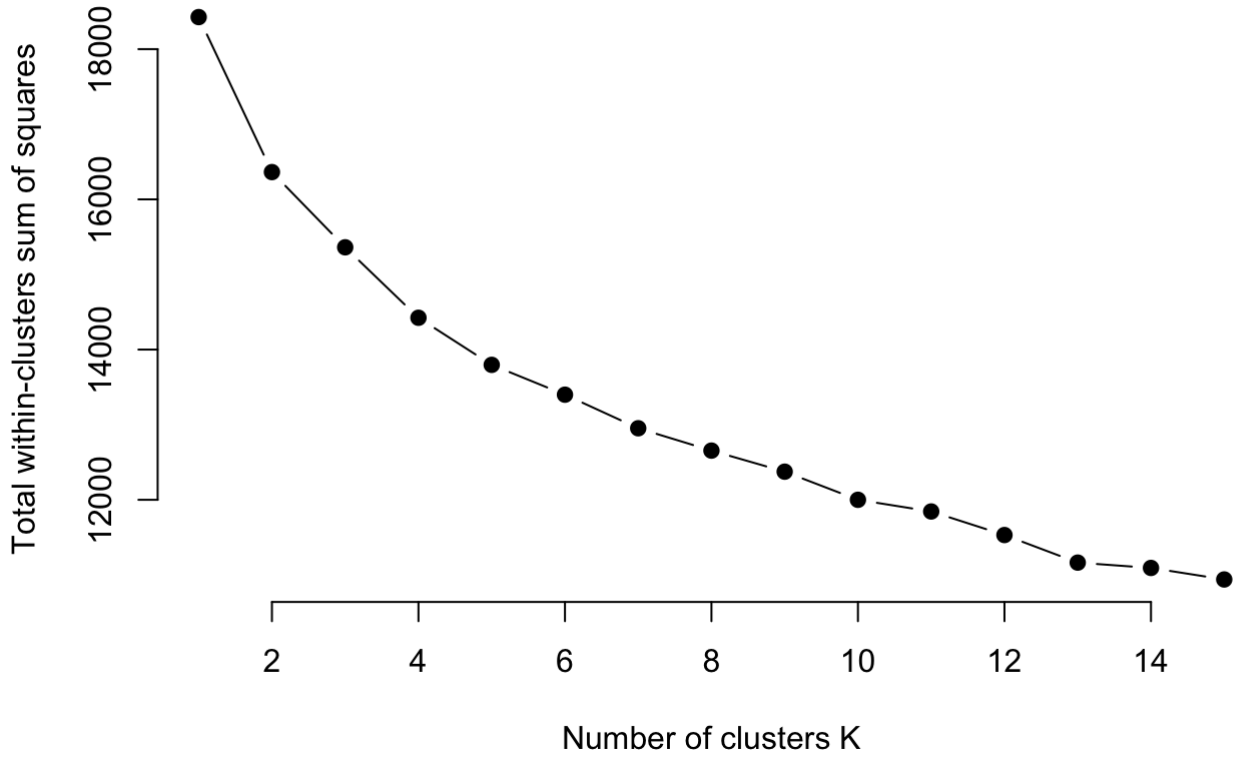
a. Demographics

The final model we used for segmentation regressed log(Purchase) on age, gender, city and occupation. From this, we could segment customers into men and women, age range, origin city, or occupation.

As mentioned previously, we found that men purchased significantly more than women during Black Friday; customers in the age range 26-35 spent the most on Black Friday, followed by customers aged 36-45; customers from city B spent the most in terms of city category; and that the top 3 segments in terms of occupation were occupations 19, 3 and 16.

b. Post-Hoc Segmentation

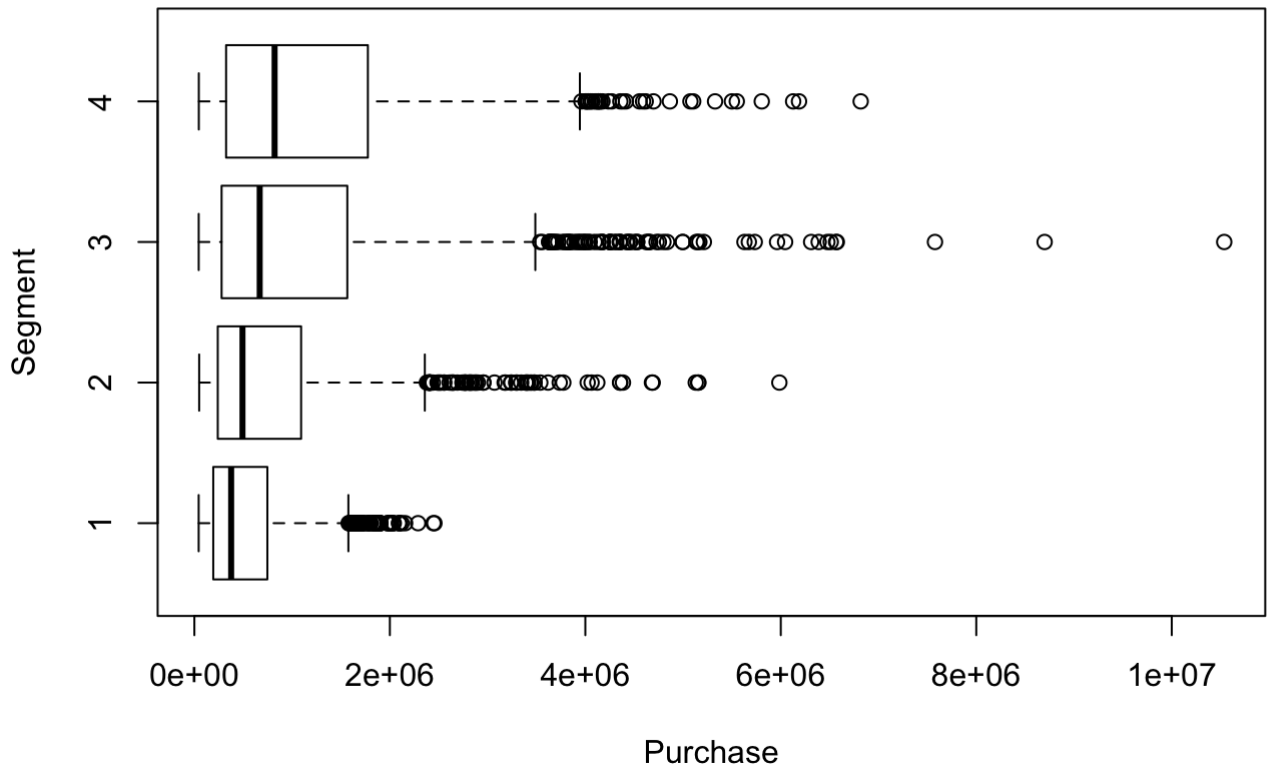
```
# Expand factors into a set of dummy variables
data = model.matrix(~Gender+Age+Occupation+City_Category+Marital_Status+Stay_In_Curre
nt_City_Years,data = demo)
# Scale the data
scaled_data = scale(data,scale=FALSE)
k.max <- 15
wss <- sapply(1:k.max,
              function(k){kmeans(scaled_data, k, nstart=50,iter.max = 20 )$tot.within
ss})
# Plot out scree plot, determine the number of clusters
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```



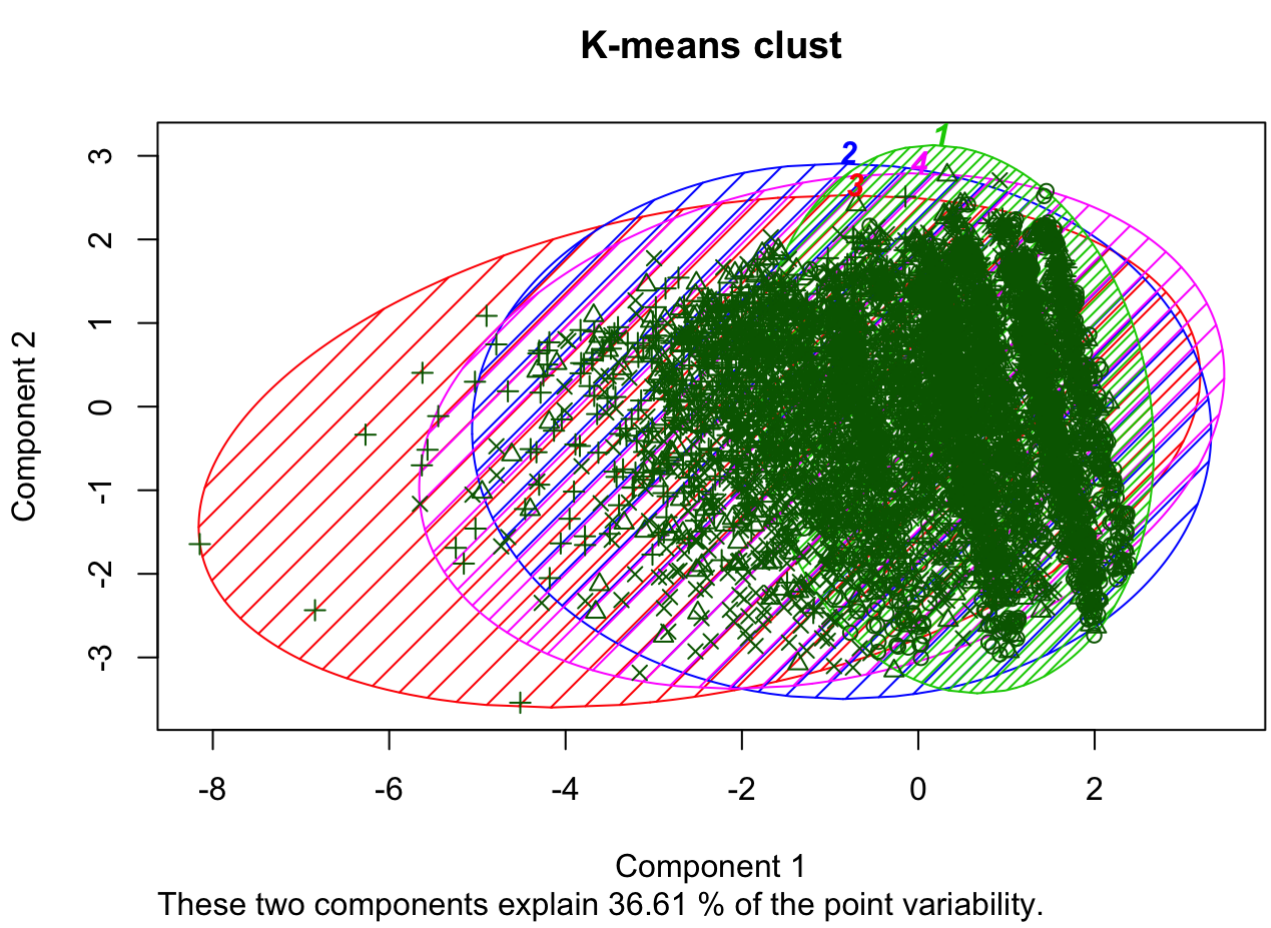

```
seg.summ <- function(data, groups) {
  aggregate(data, list(groups), function(x) mean(as.numeric(x)))
}
set.seed(96743)
# Perform cluster analysis usinng Kmeans()
seg.k4 <- kmeans(scaled_data, centers=4)
seg.summ(data[,-1], seg.k4$cluster)
```

```
##   Group.1   GenderM  Age18-25  Age26-35  Age36-45   Age46-50   Age51-55
## 1       1 0.7168889 0.1840000 0.2200000 0.2306667 0.12133333 0.10444444
## 2       2 0.7135371 0.1877729 0.3353712 0.2069869 0.08908297 0.06812227
## 3       3 0.7274131 0.1359073 0.6694981 0.1003861 0.02934363 0.02316602
## 4       4 0.7102415 0.2198168 0.2556203 0.2339717 0.09825146 0.11490425
##       Age55+ Occupation1 Occupation2 Occupation3 Occupation4 Occupation5
## 1 0.08888889  0.08933333  0.03111111  0.03111111   0.1053333  0.01955556
## 2 0.07161572  0.09432314  0.04192140  0.02969432   0.1449782  0.01572052
## 3 0.02702703  0.08108108  0.06100386  0.02702703   0.1382239  0.01698842
## 4 0.04579517  0.08576187  0.04912573  0.02581182   0.1315570  0.02248127
##   Occupation6 Occupation7 Occupation8 Occupation9 Occupation10
## 1  0.04000000   0.1186667 0.002666667 0.018222222   0.04355556
## 2  0.03493450   0.1187773 0.001746725 0.013100437   0.03668122
## 3  0.03552124   0.1073359 0.004633205 0.009266409   0.01544402
## 4  0.04329725   0.1057452 0.002497918 0.016652789   0.02664446
##   Occupation11 Occupation12 Occupation13 Occupation14 Occupation15
## 1   0.02133333   0.05822222  0.035555556   0.05200000   0.02177778
## 2   0.02008734   0.05851528  0.022707424   0.04366812   0.02707424
## 3   0.02239382   0.07567568  0.008494208   0.05791506   0.02471042
## 4   0.02331391   0.06661116  0.019150708   0.04329725   0.02331391
##   Occupation16 Occupation17 Occupation18 Occupation19 Occupation20
## 1   0.04800000   0.09111111  0.013777778   0.01288889   0.03244444
## 2   0.04279476   0.07860262  0.013973799   0.01135371   0.04366812
## 3   0.02625483   0.08339768  0.003861004   0.01235521   0.04787645
## 4   0.03663614   0.07327227  0.012489592   0.01082431   0.07327227
##   City_CategoryB City_CategoryC Marital_Status1
## 1      0.0000000      1.0000000       0.4915556
## 2      0.2986900      0.5414847       0.4209607
## 3      0.3606178      0.2077220       0.1915058
## 4      0.7477102      0.0000000       0.5312240
##   Stay_In_Current_City_Years1 Stay_In_Current_City_Years2
## 1                  0.49244444                           0
## 2                  0.00000000                           1
## 3                  0.07181467                           0
## 4                  0.73688593                           0
##   Stay_In_Current_City_Years3 Stay_In_Current_City_Years4+
## 1                  0.18177778                   0.17955556
## 2                  0.00000000                   0.00000000
## 3                  0.37528958                   0.34208494
## 4                  0.06994172                   0.05162365
```

```
# Comparing groups on 1 variable
boxplot(demo$Purchase ~ seg.k4$cluster,
        xlab="Purchase", ylab="Segment", horizontal=TRUE)
```




```
# Visualizing the overall clusters
library(cluster)
clusplot(demo, seg.k4$cluster, color=TRUE, shade=TRUE,
         labels=4, lines=0, main="K-means clust")
```



The above code shows a visualization of the sum of squared distance for k in range 1-15. The 'elbow' of the visualization shows the optimal number of clusters, which in this dataset is 4. The 4 segments differ in age, occupation, city category, marital status and stay in current city years.

- Cluster 1: Customers in this group were all from city C and had the lowest purchase amount on Black Friday. Most customers were male (71.69%), aged 18-45 (64%) and were a resident in their current city for less than 1 year (64%). They were almost evenly distributed in terms of marital status.
- Cluster 2: Customers in this group had the second lowest purchase amount on Black Friday. The majority of them were male (71.35%) and unmarried (58%), aged from 18-45 (77%). All customers had been in their current city for 2 years and most were from city C (54.15%), followed by city B (29.87%) and city A (15.98%).
- Cluster 3: Customers in this group spent the second most on Black Friday. Most customers were male (72.74%) and married (80.85%). They are also young, i.e., two thirds were aged 26-35. Roughly half of total customers were from city A and most (72) had been in their current city for over 3 years.
- Cluster 4: Customers in this group bought the most on Black Friday. Most customers were male (71%), aged 18-45 (71%), from city B (75%) and had been living in their current city for less than 1 year (88%). They were almost evenly distributed in terms of marital status.

Regression after Segmentation

```
demo$Cluster = seg.k4$cluster
cluster_multiple=lm(log(Purchase)~Gender*Cluster+Age*Cluster+Occupation*Cluster+Stay_
In_Current_City_Years*Cluster+Marital_Status*Cluster,data=demo)
summary(cluster_multiple)
```

```
##
## Call:
## lm(formula = log(Purchase) ~ Gender * Cluster + Age * Cluster +
##     Occupation * Cluster + Stay_In_Current_City_Years * Cluster +
##     Marital_Status * Cluster, data = demo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.87141 -0.71704  0.00902  0.72646  2.66348
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       12.241366   0.221599  55.241  < 2e-16
## GenderM                            0.226880   0.062159   3.650 0.000265
## Cluster                            0.246785   0.096879   2.547 0.010880
## Age18-25                           0.233813   0.209272   1.117 0.263926
## Age26-35                           0.385507   0.214789   1.795 0.072734
## Age36-45                           0.245473   0.213366   1.150 0.249993
## Age46-50                           0.265727   0.223539   1.189 0.234596
## Age51-55                           0.285714   0.225672   1.266 0.205542
## Age55+                             0.185469   0.233236   0.795 0.426530
## Occupation1                       -0.091744   0.120624  -0.761 0.446941
## Occupation2                       -0.001865   0.163971  -0.011 0.990925
## Occupation3                        0.078897   0.174004   0.453 0.650261
## Occupation4                        0.062593   0.123351   0.507 0.611864
## Occupation5                       -0.121081   0.205847  -0.588 0.556414
## Occupation6                       -0.397429   0.156183  -2.545 0.010965
## Occupation7                       -0.206035   0.113011  -1.823 0.068332
## Occupation8                       -0.792883   0.530652  -1.494 0.135186
## Occupation9                       -0.271556   0.219984  -1.234 0.217091
## Occupation10                       0.257881   0.225531   1.143 0.252902
## Occupation11                       0.016398   0.198344   0.083 0.934114
## Occupation12                       0.164116   0.135721   1.209 0.226630
## Occupation13                       0.218551   0.186597   1.171 0.241548
## Occupation14                      -0.047646   0.143423  -0.332 0.739746
## Occupation15                      -0.099720   0.194783  -0.512 0.608701
## Occupation16                      -0.048265   0.149180  -0.324 0.746302
## Occupation17                      -0.053127   0.121308  -0.438 0.661439
## Occupation18                      -0.203736   0.242755  -0.839 0.401353
## Occupation19                       0.238380   0.253643   0.940 0.347347
## Occupation20                       0.091517   0.158523   0.577 0.563754
## Stay_In_Current_City_Years1       -0.012923   0.084387  -0.153 0.878291
## Stay_In_Current_City_Years2        0.076111   0.045310   1.680 0.093058
## Stay_In_Current_City_Years3        0.018879   0.100768   0.187 0.851391
## Stay_In_Current_City_Years4+       0.069871   0.101428   0.689 0.490928
## Marital_Status1                   -0.034816   0.058964  -0.590 0.554908
## GenderM:Cluster                    0.026095   0.024541   1.063 0.287666
## Cluster:Age18-25                  -0.049272   0.092780  -0.531 0.595398
## Cluster:Age26-35                  -0.061984   0.094164  -0.658 0.510402
## Cluster:Age36-45                  -0.012931   0.094287  -0.137 0.890924
## Cluster:Age46-50                  -0.045741   0.098760  -0.463 0.643276
## Cluster:Age51-55                  -0.078624   0.098949  -0.795 0.426885
## Cluster:Age55+                    -0.151711   0.104468  -1.452 0.146495
## Cluster:Occupation1                0.018492   0.047773   0.387 0.698714
## Cluster:Occupation2               -0.007410   0.061003  -0.121 0.903316
## Cluster:Occupation3                0.034482   0.070470   0.489 0.624638
## Cluster:Occupation4               -0.021819   0.047472  -0.460 0.645817
```

```
## Cluster:Occupation5                    0.082440   0.079660   1.035 0.300761
## Cluster:Occupation6                    0.108448   0.061061   1.776 0.075773
## Cluster:Occupation7                    0.035416   0.044951   0.788 0.430799
## Cluster:Occupation8                    0.232515   0.203269   1.144 0.252723
## Cluster:Occupation9                    0.039371   0.089072   0.442 0.658496
## Cluster:Occupation10                  -0.146229   0.099881  -1.464 0.143239
## Cluster:Occupation11                  -0.025616   0.077269  -0.332 0.740269
## Cluster:Occupation12                  -0.103334   0.052518  -1.968 0.049164
## Cluster:Occupation13                  -0.208217   0.082957  -2.510 0.012102
## Cluster:Occupation14                   0.011158   0.057121   0.195 0.845130
## Cluster:Occupation15                   0.028011   0.076438   0.366 0.714044
## Cluster:Occupation16                   0.078597   0.061767   1.272 0.203255
## Cluster:Occupation17                  -0.029186   0.048711  -0.599 0.549086
## Cluster:Occupation18                   0.073200   0.101290   0.723 0.469910
## Cluster:Occupation19                  -0.009202   0.102231  -0.090 0.928283
## Cluster:Occupation20                   0.001274   0.057194   0.022 0.982229
## Cluster:Stay_In_Current_City_Years1    0.006307   0.031283   0.202 0.840220
## Cluster:Stay_In_Current_City_Years2          NA         NA      NA       NA
## Cluster:Stay_In_Current_City_Years3    0.024580   0.038958   0.631 0.528100
## Cluster:Stay_In_Current_City_Years4+  -0.010684   0.039840  -0.268 0.788580
## Cluster:Marital_Status1                0.019299   0.023201   0.832 0.405558
##
## (Intercept)                           ***
## GenderM                               ***
## Cluster                               *
## Age18-25
## Age26-35                              .
## Age36-45
## Age46-50
## Age51-55
## Age55+
## Occupation1
## Occupation2
## Occupation3
## Occupation4
## Occupation5
## Occupation6                           *
## Occupation7                           .
## Occupation8
## Occupation9
## Occupation10
## Occupation11
## Occupation12
## Occupation13
## Occupation14
## Occupation15
## Occupation16
## Occupation17
## Occupation18
## Occupation19
## Occupation20
## Stay_In_Current_City_Years1
## Stay_In_Current_City_Years2           .
## Stay_In_Current_City_Years3
## Stay_In_Current_City_Years4+
## Marital_Status1
## GenderM:Cluster
## Cluster:Age18-25
```

```
## Cluster:Age26-35
## Cluster:Age36-45
## Cluster:Age46-50
## Cluster:Age51-55
## Cluster:Age55+
## Cluster:Occupation1
## Cluster:Occupation2
## Cluster:Occupation3
## Cluster:Occupation4
## Cluster:Occupation5
## Cluster:Occupation6                  .
## Cluster:Occupation7
## Cluster:Occupation8
## Cluster:Occupation9
## Cluster:Occupation10
## Cluster:Occupation11
## Cluster:Occupation12                 *
## Cluster:Occupation13                 *
## Cluster:Occupation14
## Cluster:Occupation15
## Cluster:Occupation16
## Cluster:Occupation17
## Cluster:Occupation18
## Cluster:Occupation19
## Cluster:Occupation20
## Cluster:Stay_In_Current_City_Years1
## Cluster:Stay_In_Current_City_Years2
## Cluster:Stay_In_Current_City_Years3
## Cluster:Stay_In_Current_City_Years4+
## Cluster:Marital_Status1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9373 on 5826 degrees of freedom
## Multiple R-squared:  0.1182, Adjusted R-squared:  0.1085
## F-statistic:  12.2 on 64 and 5826 DF,  p-value: < 2.2e-16
```

In cluster 1 where customers came from city C, men spent 23.3% more than women. Customers aged 26-35 bought the most, 31% more than customers under 17 years old, followed by those who were in the age range 46-50 and 51-55 (at the 90% significance level). We also found that people in occupation 6 and 7 purchased significantly less than those in occupation 0.

In cluster 2 where all customers were new in their current city, men spent 32.3% more than women. There was no difference in purchase among various age ranges. Only occupation 13 spent significantly less (53%) than occupation 0. It is noticeable that, customers from city B purchased marginally (14.5%) less than those from city A and that customers from city C spent significantly (72.4%) less than customers from city A. We also found that marital status was not correlated with purchase on Black Friday.

In cluster 3, men spent 39% more than women. Age, marital status and stay in current city years had no effect on the purchase amount on Black Friday. Occupation 19 had the highest purchase amount, followed by occupation 3. Customers from city B spent the most followed by those who came from city A and C.

Finally, men in the last cluster bought more than women. Age, marital status, and stay in current city years (at the 95% significance level) did not affect purchase. Customers from city B spent more than customers from city A.

In general, marital status is not correlated with purchase amount on Black Friday. Likewise, age has no effect on total spending except in cluster 1.

7. Predictive Model

a.Prediction Problem

The objective of predictive analysis is to predict the final purchase for individual customer on Black Friday.

b.State the data aggregation and split the dataset

```
set.seed(12345678)
validation.index = sample(seq(1,nrow(demo)),
                          floor(0.10*nrow(demo)))
validation.data = demo[validation.index, ]
training.data = demo[-validation.index, ]
```

In this part, we continued to use the overall expenditure for customers, which is user-level data.

c.Run at least 15 predictive models

```
# Compare AIC and BIC of regression with all the categorical variables with regressio
n with statistically significant variable.
lm_age_gender_occupation_city = lm(log(Purchase)~Age*factor(Cluster)+Gender*factor(Cl
uster)+Occupation*factor(Cluster)+City_Category*factor(Cluster), data = training.dat
a)
model_sel = data.frame(Adj_r = c(summary(cluster_multiple)$adj.r.squared,
                                 summary(lm_age_gender_occupation_city)$adj.r.squar
ed),
                       AIC = c(AIC(cluster_multiple),
                               AIC(lm_age_gender_occupation_city)),
                       BIC = c(BIC(cluster_multiple),
                               BIC(lm_age_gender_occupation_city)))
rownames(model_sel) = c("Model 1", "Model 2")
```

We first needed to decide which variables to put into the models. To determine this, we used AIC and BIC, which penalizes for having excessive regressors in the model. The lower AIC and BIC scores of “Model 2” indicates that regression with only significant variables is better. Our next issue was determining the different combinations of these variables to create the best model.To measure the best model, we used the error metric Root Mean Squared Error (RMSE). The model with lowest RMSE will be the best predictive model!

```
# RMSE
prediction.error = function(lm_model, validation.data){
  predicted.units = predict(lm_model, validation.data)
  error = sqrt(mean((predicted.units-log(validation.data$Purchase))^2))
}
# Predictive models
lm_cluster = lm(log(Purchase)~factor(Cluster), data = training.data)
lm_cluster_age = lm(log(Purchase)~Age*factor(Cluster), data = training.data)
lm_cluster_gender = lm(log(Purchase)~Gender*factor(Cluster), data = training.data)
lm_cluster_occupation = lm(log(Purchase)~Occupation*factor(Cluster), data = training.
data)
lm_cluster_city = lm(log(Purchase)~City_Category*factor(Cluster), data = training.dat
a)

lm_age_gender = lm(log(Purchase)~Age*factor(Cluster)+Gender*factor(Cluster), data = t
raining.data)
lm_age_occupation = lm(log(Purchase)~Age*factor(Cluster)+Occupation*factor(Cluster),
 data = training.data)
lm_age_city = lm(log(Purchase)~Age*factor(Cluster)+City_Category*factor(Cluster), dat
a = training.data)
lm_gender_occupation = lm(log(Purchase)~Gender*factor(Cluster)+Occupation*factor(Clus
ter), data = training.data)
lm_gender_city = lm(log(Purchase)~Gender*factor(Cluster)+City_Category*factor(Cluste
r), data = training.data)
lm_occupation_city = lm(log(Purchase)~Occupation*factor(Cluster)+City_Category*factor
(Cluster), data = training.data)

lm_age_gender_occupation = lm(log(Purchase)~Age*factor(Cluster)+Gender*factor(Cluste
r)+Occupation*factor(Cluster)+City_Category*factor(Cluster), data = training.data)
lm_age_gender_city = lm(log(Purchase)~Age*factor(Cluster)+Gender*factor(Cluster)+City
_Category*factor(Cluster), data = training.data)
lm_age_occupation_city = lm(log(Purchase)~Age*factor(Cluster)+Occupation*factor(Clust
er)+City_Category*factor(Cluster), data = training.data)
lm_gender_occupation_city = lm(log(Purchase)~Gender*factor(Cluster)+Occupation*factor
(Cluster)+City_Category*factor(Cluster), data = training.data)

lm_age_gender_occupation_city = lm(log(Purchase)~Age*factor(Cluster)+Gender*factor(Cl
uster)+Occupation*factor(Cluster)+City_Category*factor(Cluster), data = training.dat
a)

# Errors of predictive models
error_model1 = prediction.error(lm_cluster, validation.data)
error_model2 = prediction.error(lm_cluster_age, validation.data)
error_model3 = prediction.error(lm_cluster_gender, validation.data)
error_model4 = prediction.error(lm_cluster_occupation, validation.data)
error_model5 = prediction.error(lm_cluster_city, validation.data)

error_model6 = prediction.error(lm_age_gender, validation.data)
error_model7 = prediction.error(lm_age_occupation, validation.data)
error_model8 = prediction.error(lm_age_city, validation.data)
error_model9 = prediction.error(lm_gender_occupation, validation.data)
error_model10 = prediction.error(lm_gender_city, validation.data)
error_model11 = prediction.error(lm_occupation_city, validation.data)

error_model12 = prediction.error(lm_age_gender_occupation, validation.data)
error_model13 = prediction.error(lm_age_gender_city, validation.data)
error_model14 = prediction.error(lm_age_occupation_city, validation.data)
```

```
error_model15 = prediction.error(lm_gender_occupation_city, validation.data)

error_model16 = prediction.error(lm_age_gender_occupation_city, validation.data)

data.frame(Names = c("cluster", "age","gender", "occupation", "city","age_gender","ag
e_occupatin","age_city","gender_occupation","gender_city","occupation_city","age_gend
er_occupation","age_gender_city","age_occupation_city","gender_occupation_city","age_
gender_occupation_city"),
           Error = c(error_model1,error_model2, error_model3, error_model4, error_mod
el5, error_model6, error_model7, error_model8, error_model9, error_model10, error_mod
el11, error_model12,error_model13, error_model14, error_model15, error_model16))
```

```
##                         Names     Error
## 1                     cluster 0.9595668
## 2                         age 0.9547564
## 3                      gender 0.9453166
## 4                  occupation 0.9593373
## 5                        city 0.9263621
## 6                  age_gender 0.9409337
## 7               age_occupatin 0.9550684
## 8                   age_city 0.9245644
## 9           gender_occupation 0.9462851
## 10                gender_city 0.9111179
## 11            occupation_city 0.9268711
## 12      age_gender_occupation 0.9119683
## 13            age_gender_city 0.9099173
## 14        age_occupation_city 0.9261304
## 15     gender_occupation_city 0.9118157
## 16 age_gender_occupation_city 0.9119683
```

From the above table we can see that the model “lm_age_gender_city” has the smallest RMSE. That is our best model!

d.choose the best model and show its summary output

```
summary(lm_age_gender_city)
```

```
##
## Call:
## lm(formula = log(Purchase) ~ Age * factor(Cluster) + Gender *
##     factor(Cluster) + City_Category * factor(Cluster), data = training.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.88413 -0.71109  0.03447  0.72787  2.48395
##
## Coefficients: (3 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   13.283716   0.122742 108.224  < 2e-16 ***
## Age18-25                       0.163990   0.101795   1.611  0.10724
## Age26-35                       0.170459   0.099985   1.705  0.08828 .
## Age36-45                       0.133842   0.099803   1.341  0.17996
## Age46-50                       0.147743   0.107473   1.375  0.16928
## Age51-55                       0.079683   0.109619   0.727  0.46731
## Age55+                         0.014064   0.113297   0.124  0.90122
## factor(Cluster)2              -0.055247   0.203269  -0.272  0.78579
## factor(Cluster)3               0.181162   0.255311   0.710  0.47800
## factor(Cluster)4              -0.550476   0.205306  -2.681  0.00736 **
## GenderM                        0.213223   0.045582   4.678 2.97e-06 ***
## City_CategoryB                 0.197463   0.064215   3.075  0.00212 **
## City_CategoryC                -0.708882   0.077275  -9.173  < 2e-16 ***
## Age18-25:factor(Cluster)2     -0.114752   0.187890  -0.611  0.54140
## Age26-35:factor(Cluster)2      0.009641   0.181929   0.053  0.95774
## Age36-45:factor(Cluster)2      0.017033   0.185681   0.092  0.92692
## Age46-50:factor(Cluster)2     -0.133775   0.203358  -0.658  0.51068
## Age51-55:factor(Cluster)2     -0.178158   0.211845  -0.841  0.40040
## Age55+:factor(Cluster)2       -0.207364   0.210138  -0.987  0.32379
## Age18-25:factor(Cluster)3     -0.537600   0.255340  -2.105  0.03530 *
## Age26-35:factor(Cluster)3     -0.315060   0.247383  -1.274  0.20287
## Age36-45:factor(Cluster)3     -0.274871   0.258270  -1.064  0.28725
## Age46-50:factor(Cluster)3     -0.619427   0.291247  -2.127  0.03348 *
## Age51-55:factor(Cluster)3     -0.528570   0.306727  -1.723  0.08490 .
## Age55+:factor(Cluster)3       -0.724749   0.298288  -2.430  0.01514 *
## Age18-25:factor(Cluster)4      0.312988   0.193655   1.616  0.10611
## Age26-35:factor(Cluster)4      0.422821   0.190673   2.218  0.02663 *
## Age36-45:factor(Cluster)4      0.476941   0.191619   2.489  0.01284 *
## Age46-50:factor(Cluster)4      0.432114   0.207264   2.085  0.03713 *
## Age51-55:factor(Cluster)4      0.413596   0.206468   2.003  0.04521 *
## Age55+:factor(Cluster)4       -0.114628   0.229139  -0.500  0.61692
## factor(Cluster)2:GenderM       0.140539   0.078607   1.788  0.07385 .
## factor(Cluster)3:GenderM       0.149552   0.075875   1.971  0.04877 *
## factor(Cluster)4:GenderM      -0.011774   0.076824  -0.153  0.87820
## factor(Cluster)2:City_CategoryB -0.300303   0.109188  -2.750  0.00597 **
## factor(Cluster)3:City_CategoryB -0.083553   0.088535  -0.944  0.34535
## factor(Cluster)4:City_CategoryB        NA         NA      NA       NA
## factor(Cluster)2:City_CategoryC  0.004566   0.112546   0.041  0.96764
## factor(Cluster)3:City_CategoryC        NA         NA      NA       NA
## factor(Cluster)4:City_CategoryC        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9177 on 5265 degrees of freedom
## Multiple R-squared:  0.1496, Adjusted R-squared:  0.1438
## F-statistic: 25.73 on 36 and 5265 DF,  p-value: < 2.2e-16
```

e&f. predict the mean outcome and 95% confidence interval for the major segments

```
# Creat the test data
test.data = demo %>% group_by(Cluster, Gender, Age, City_Category)%>%summarize(actua
l.purchase = log(mean(Purchase)))

# Prediction of mean log purchase for each customer
predicted.y = predict(lm_age_gender_city, test.data, se.fit = TRUE)
test.data$predicted.purchase = predicted.y$fit
test.data$lowerlimit =test.data$predicted.purchase - qnorm(0.975)*predicted.y$se.fit
test.data$upperlimit =test.data$predicted.purchase + qnorm(0.975)*predicted.y$se.fit

# Convert log purchase into purchase
test.data[5:8]=exp(test.data[5:8])


# Prediction of Clusters
test.data %>% group_by(Cluster) %>%
  summarise(predicted.purchase=mean(predicted.purchase),pred.left=mean(lowerlimit),pr
ed.right=mean(upperlimit))
```

```
## # A tibble: 4 x 4
##   Cluster predicted.purchase pred.left pred.right
##     <int>              <dbl>     <dbl>      <dbl>
## 1       1            358847.   317231.    406259.
## 2       2            550360.   445981.    681148.
## 3       3            652353.   503195.    858004.
## 4       4            632035.   525136.    763116.
```

```
# Prediction of Age Categories
test.data %>% group_by(Age) %>%
  summarise(predicted.purchase=mean(predicted.purchase),pred.left=mean(lowerlimit),pr
ed.right=mean(upperlimit))
```

```
## # A tibble: 7 x 4
##   Age   predicted.purchase pred.left pred.right
##   <fct>              <dbl>     <dbl>      <dbl>
## 1 0-17             574290.   404729.    820174.
## 2 18-25            584594.   499077.    685157.
## 3 26-35            641155.   561397.    732679.
## 4 36-45            670961.   570424.    789855.
## 5 46-50            580082.   461923.    730754.
## 6 51-55            543028.   423785.    699900.
## 7 55+              413654.   319245.    537598.
```

```
# Prediction of Gender Categories
test.data %>% group_by(Gender) %>%
  summarise(predicted.purchase=mean(predicted.purchase),pred.left=mean(lowerlimit),pr
ed.right=mean(upperlimit))
```

```
## # A tibble: 2 x 4
##   Gender predicted.purchase pred.left pred.right
##   <fct>               <dbl>     <dbl>      <dbl>
## 1 F                 487180.   391494.    610608.
## 2 M                 660303.   537841.    817519.
```

```
# Prediction of City Categories
test.data %>% group_by(City_Category) %>%
  summarise(predicted.purchase=mean(predicted.purchase),pred.left=mean(lowerlimit),pr
ed.right=mean(upperlimit))
```

```
## # A tibble: 3 x 4
##   City_Category predicted.purchase pred.left pred.right
##   <fct>                      <dbl>     <dbl>      <dbl>
## 1 A                         631095.   501536.    799942.
## 2 B                         675640.   544848.    845188.
## 3 C                         350788.   300800.    410260.
```

For the predicted top 10 highest average actual purchases from customers: 100% were male, 60% were from city B, and 70% were from the age range of 26-45.

Conclusion

From our preliminary data analysis, we were able to determine customer demographics that were associated with higher purchase amounts. In particular, we found that men spend more money on Black Friday than women, the purchase amounts are higher in different cities (City C has higher purchase amounts), unmarried people spend more than married (but only slightly more), certain occupation categories spend more than others, and interestingly, that there was not a huge difference in mean purchase by age (26-35 category spends slightly more). Also interestingly, there is almost no perceivable difference in mean purchase by years in city. For the marital status, its effect on purchase is ambiguous.

We next furthered the analysis by use of regressions. Based on previous takeaways, we built a baseline model, which regressed log(purchase) on gender, age and city category. In the regression section, we found that occupation was an omitted variable and then were able to compile our final regression model. We used four significant factors: gender, age, city category and occupation as the independent variables to figure out the unknown effect on the dependent variable - log(purchase). Based on the result from the final model, we found that certain occupations (3,6,7,13,16,17,19) were correlated with purchase. Moreover, men purchased 28.98% more than women; customers aged 26-35 bought the most on Black Friday and customers from city B also purchased the most.

Customer segmentation provided more insight into customers by grouping them based on similar characteristics or purchasing behavior, which can be classified as either a priori or post-hoc. The result of a priori segmentation was based on customer demographics, while post-hoc segmentation placed customers into 4 clusters. By re-performing the regressions for each cluster, we found that marital status was not correlated with purchase amount on Black Friday, which matched the result of the regression part. We also found that age had no effect on total spending except in cluster 1.

The goal of our predictive analysis was to predict the purchase amount of an individual customer. The lower AIC and BIC scores justified excluding stay in current city years and marital status as the independent variables in our predictive models. User-level data was split into a training set (90%), used to train the models and validation set (10%), and finally used to evaluate the performance of the models. Regressing purchase on cluster, age, gender and city resulted in the the smallest RMSE, indicating that this was the best model. Equipped with these results, we could now predict the purchase amount for each customer given corresponding demographics.

According to the predictive analysis, cluster 3 spent the most; customers in the age of range 36-45 spent the most; male customers spent the most; and customers from city B spent the most. We also noticed that customers who were predicted to have the highest purchase amounts had certain characteristics: 100% were men, 60% were from city B, 40% were from city A, and 70% were in the age range of 26-45. It is important to note, however, that while this segment is crucial, it is important to not direct all marketing dollars towards it. This is because the actual data showed more of a mix: some women, some older, most from city A. Therefore, we suggest that the company direct more marketing campaign on the up-and-coming cluster 3, and continue more traditional, “tried but true” campaigns for the other clusters.