Marketing Analytics Project: Black Friday

Weicheng Zhang, Zilei Zhang, Maura Oray, Gloria Zheng

1. Introduction

The dataset here is a sample of the transactions made in a retail store during "Black Friday." The store wants to gain a better understanding about customer purchase behavior compared to different products. This is useful for the company to make purchasing decisions for the next year. In this project, we want to: (i) Look at the distribution of customer demographics; (ii) Determine which variables are highly correlated with purchase; (iii) 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

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 the visualization in the appendix, 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 coerce 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

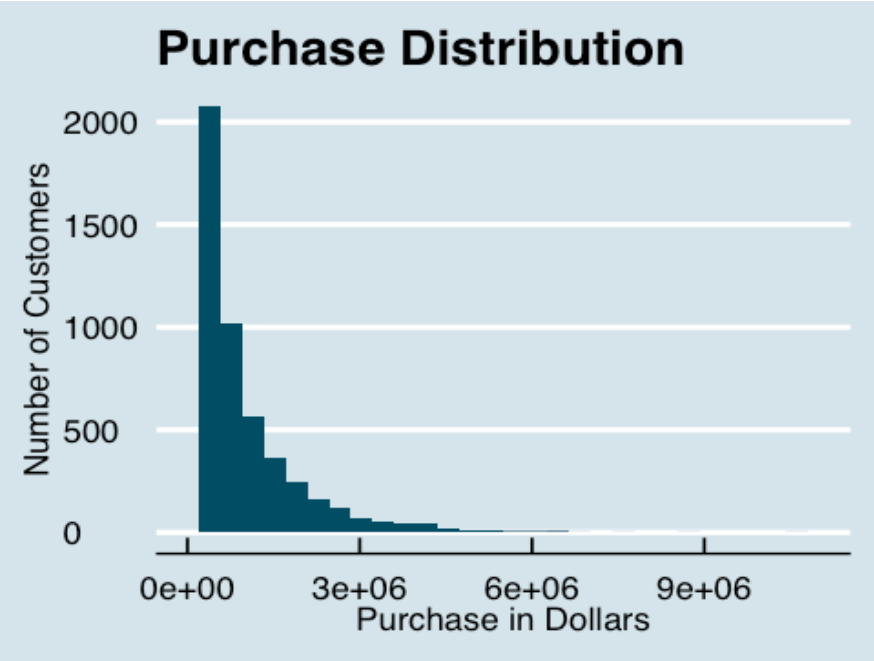
```
#Select demographic data
demo = BF %>%
  select(-starts_with("Product")) %>%
  group_by(User_ID) %>%
  mutate(Purchase = sum(Purchase))
demo = demo[!duplicated(demo$User_ID), ]
```

Because we decided to focus more on customer analysis instead of the product, and that Product Category 2 and Product Category 3 have 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

Plot 1. Distribution of Purchase

```
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 Distribution")
```

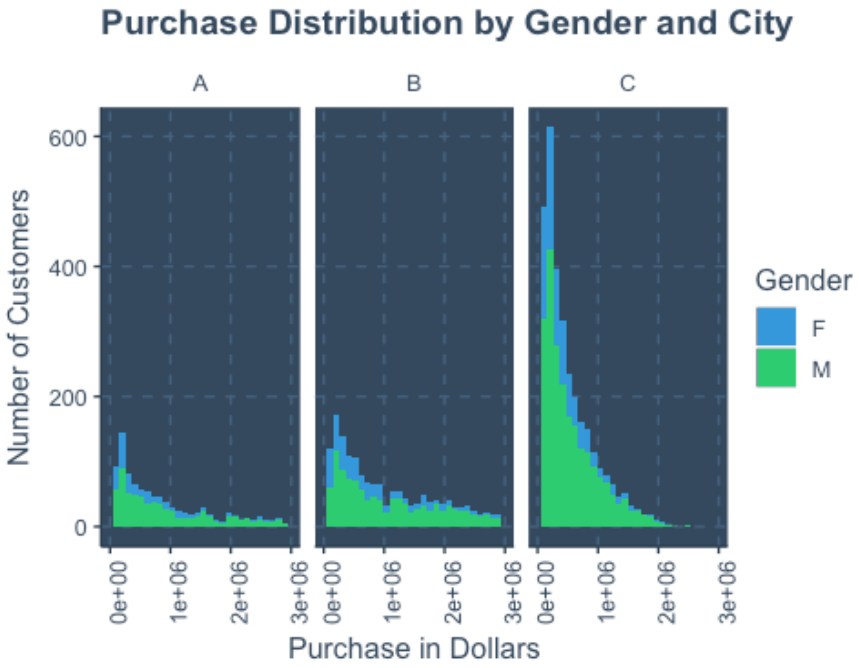


Here we selected the variable “Purchase.” We decided to use a histogram to view the distribution of customers and purchase because it is the best way to display continuous data. From the graph, it shows a downward slope between number of customers and purchase, which illustrates that a larger number 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

Plot 2. Purchase Distribution by Gender and City

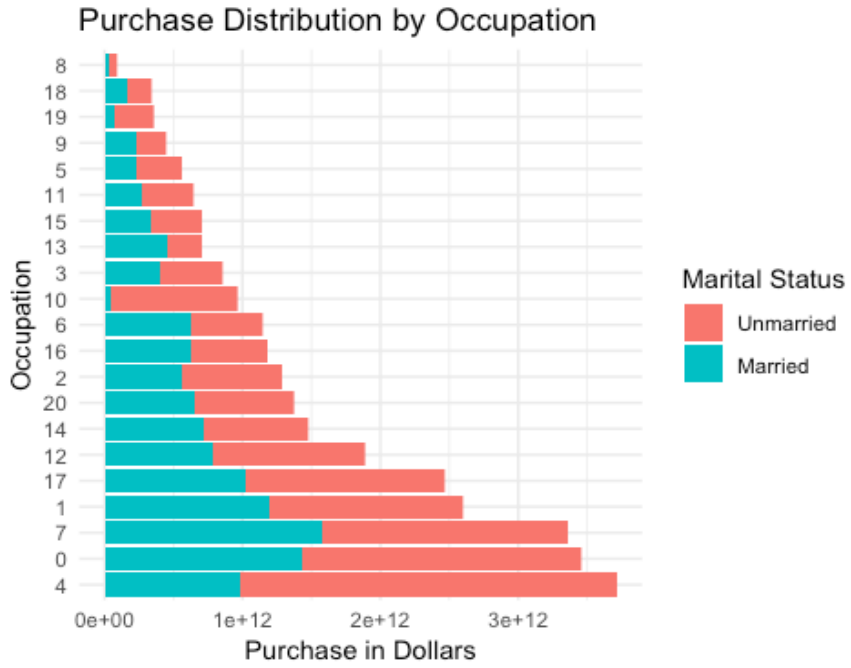
```
ggthemr('flat dark')
demo %>%
  ggplot(aes(x = Purchase, fill = Gender)) +
  geom_histogram() +
  facet_wrap(~City_Category) +
  scale_x_continuous(limits = c(0, 3000000)) +
  labs(x = "Purchase in Dollars", y = "Number of Customers", title = "Purchase Distribution by Gender an
d City")+
  theme(axis.text.x = 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 are interested in comparing the purchasing pattern in different cities and different 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.

Plot 3. Purchase Distribution by Occupation

```r
#Create new data frame grouping by occupation and marital status, and taking the mean purchase of eac
h 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="M
arital Status") +
    scale_fill_discrete(labels=c("Unmarried", "Married"))
```

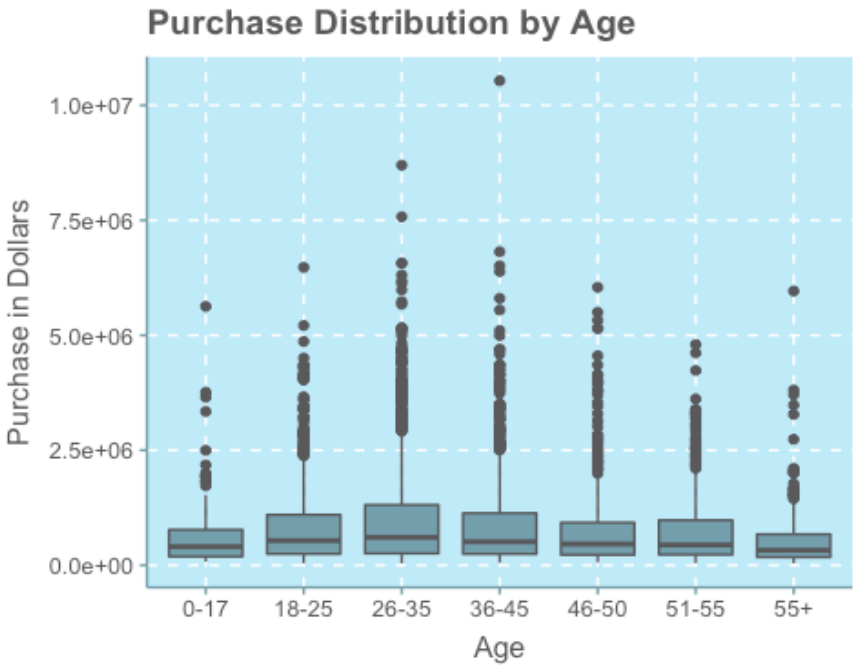


We are interested in determining which occupations spend more and whether married customers spend more than unmarried customers. Since marital status and occupation are factors (discrete) variables, and purchase is a continuous variable, we used a barplot to show total purchase for each occupation, split 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.

Plot 4. Purchase Distribution by Age

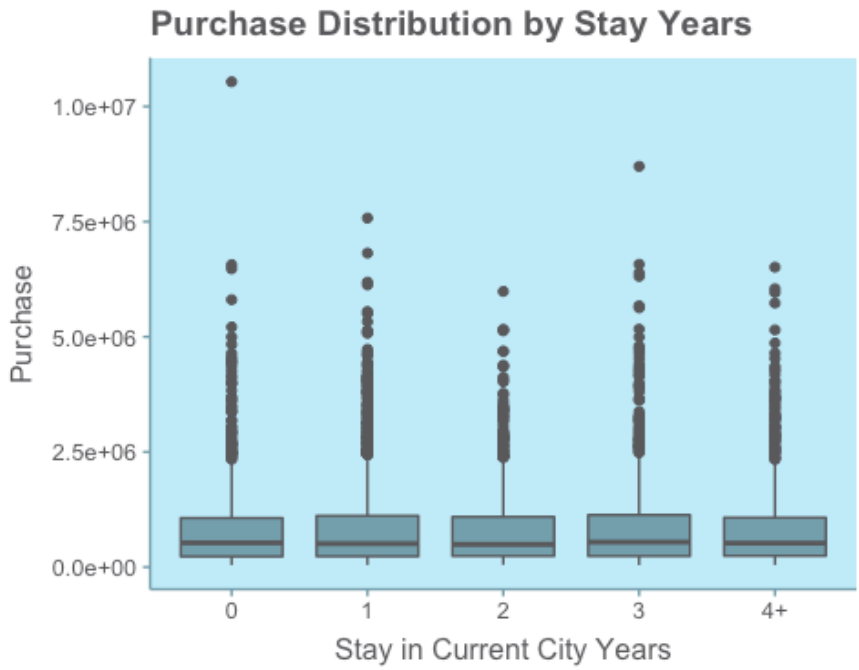
```
ggthemr('sky')
demo %>%
  ggplot(aes(x = Age, y = Purchase)) +
  geom_boxplot() +
  labs(x = "Age", y = "Purchase in Dollars", title = "Purchase Distribution by Age")
```



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 36-35, which has the highest median purchase and range. Customers who are under 17 or over 55, however seem to spend a little bit less.

Plot 5. Purchase distribution by stay in city years

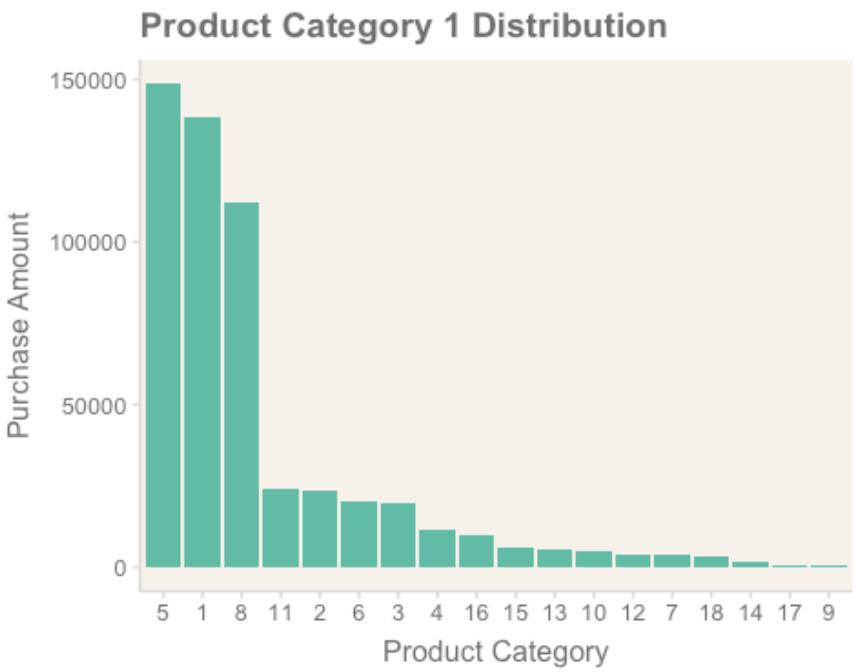
```
ggthemr('sky', layout = 'minimal')
demo %>%
  ggplot(aes(x = Stay_In_Current_City_Years, y = Purchase)) +
  geom_boxplot()+
  labs(x = "Stay in Current City Years", y = "Purchase", title = "Purchase Distribution by Stay Years")
```



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.

Plot 6. The most popular product category

```
ggthemr('light', layout = 'clean')
BF_prt = BF %>%
  group_by(Product_ID) %>%
  mutate(sumpp = sum(Purchase), avgpp = mean(Purchase))
BF_prt = BF_prt[-c(3:8)]
BF_prt%>%
  ggplot(aes(x = Product_Category_1))+
  geom_bar(aes(x = fct_infreq(as.factor(Product_Category_1))))+
  labs(x ="Product Category" , y = "Purchase Amount", title = "Product Category 1 Distribution")
```



In the above visualization, we decided to choose Product Category 1 and Purchase variables in order to see which products were the highest selling. We decided to omit product 2 and 3 categories because many of the values are zero. Using a bar plot visualization is the most appropriate because it shows continuous data on the y-axis (purchase, using geom_bar) and discrete on the x-axis. We improved the plot by reordering the bars by highest to lowest and changing the theme colors.
We can easily see that the top 3 product 1 categories are 5, 1, and 8, by far. This visualization can help the business because they can direct more marketing dollars toward promoting the top 3 product categories. On the other hand, the company could use this plot to target certain products that it wants to sell more and direct marketing dollars towards those categories.

4. Preliminary statistical analyses

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 have 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 males and females in purchase.

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

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

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 want to determine whether marital status influences purchase. In table 1 and plot 3, we find that there are more unmarried people than married, and that unmarried customers seem to make more purchases than married customers. Because maritial status is category and purchase is continuous, we use an independent t-test to test whether this difference is significant. Based on the evidence from this data set and the result from the t-test (p-value = 0.056), we draw the conclusion that in the 95% significant level, we cannot reject the null hypothesis. But in the 90% significant level, however, we can reject the null hypothesis and accept the alternative hypothesis: that unmarried customers have larger mean purchase than married customers.

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$Age, demo$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 the above code, we want to determine whether city stay influences purchase. In Plot 2, we find differences in purchase distribution. Therefore, we want to use a Chi-Square test to see whether this finding is significant.
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 the probability that there is a difference among the three cities is bigger than 95%. Therefore, we reject the null hypothesis, and accept the alternative hypothesis: that there is a difference in purchase distribution between the three cities. Therefore, from this data set, we draw the conclusion that there is an association between city and purchase.
In terms of 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.

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 different in the 3 cities, certain occupation categories spend more than others, there was a certain difference in mean purchase by age. Interestingly, there is almost no perceivable difference in mean purchase by years in city. For the marital status, its effect on purchase is ambiguous. Based on the current data, we cannot confidentially say that there exists difference between married and unmarried customers. We need more data to analyze this problem later.
By using statistical tests and visualizations, we are able to easily pinpoint particular demographics on which the company should direct its marketing dollars. Further analysis will illuminate these important demographics.