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 know better about the customer purchase behaviour against different products which is useful for them to make the purchasing decision 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 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, therefore, we come 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**

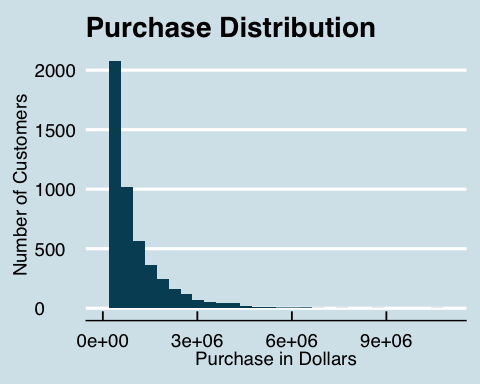
#Select demographic data  
demo = BF %>%   
 select(-starts\_with("Product")) %>%  
 group\_by(User\_ID) %>%  
 mutate(Purchase = sum(Purchase))  
demo = demo[!duplicated(demo$User\_ID), ]

Because we decided to focus more on customer analysis instead of the product and Product Category 2 and Product Category 3 have a lot of missing data, we delete all of the product category columns to make a clearer customer data set. Moreover, we find that the Customer IDs are repeated for many times which is redundant and we need a sum purchase number. Therefore, we use the customer all purchase to replace the individual per product purchase and delete the redundant customer IDs.

**3. Summary statistics and Data Visualizations** **Customer Demographics**

**Plot 1. Distribution of Purchase**

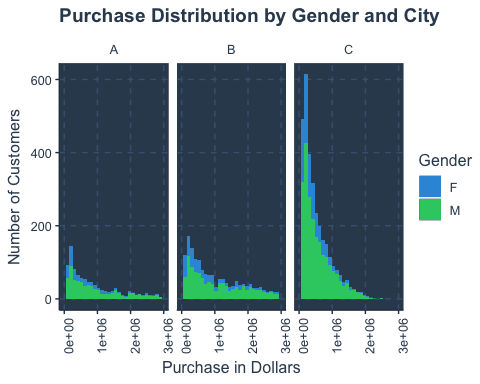
demo %>%  
 ggplot(aes(Purchase)) +  
 geom\_histogram(fill = "#014d64") +  
 scale\_x\_continuous(limits = c(0, 11000000)) +  
 theme\_economist(base\_size=14)+  
 scale\_fill\_economist()+  
 labs(x = "Purchase in Dollars", y = "Number of Customers", title = "Purchase Distribution")



We select variable “Purchase”. We use histogram to see the distribution of customers in amount of purchasing because it is the best way for continuous data. So, we put “Purchase” in x axis, and see how many customers spend in the different prices. From the graph, it shows a downward slope between number of customers and purchase in dollars, which illustrates more numbers of customers spend in lower price and fewer number of customers spend in higher price. [1]Glorious Christian, Black Friday Analysis <https://www.kaggle.com/gloriousc/black-friday-analysis>

**Plot 2. Purchase Distribution by Gender and City**

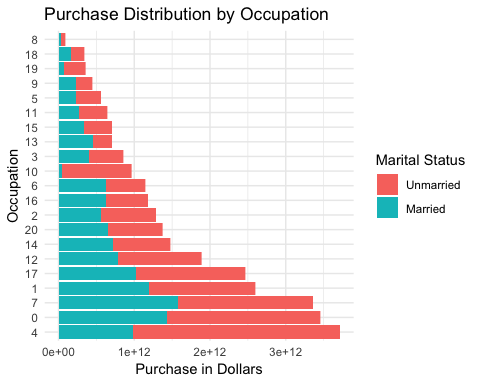
ggthemr('flat dark')  
demo %>%  
 ggplot(aes(x = Purchase, fill = Gender)) +  
 geom\_histogram() +  
 facet\_wrap(~City\_Category) +  
 scale\_x\_continuous(limits = c(0, 3000000)) +  
 labs(x = "Purchase in Dollars", y = "Number of Customers", title = "Purchase Distribution by Gender and City")+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5))



We select variable “Gender”, “City” to see the purchasing distribution of different gender in different cities. We also use histogram because variable “Purchase” is a continuous data. We are interested in comparing the purchasing pattern in different cities and different gender. We use facets which divide purchasing pattern to three cities. It shows that at the same price more customers spend in city “C”, and fewer customers spend in city “A”. Moreover, in each city, at the same price male purchase more than female.

**Plot 3. Purchase Distribution by Occupation**

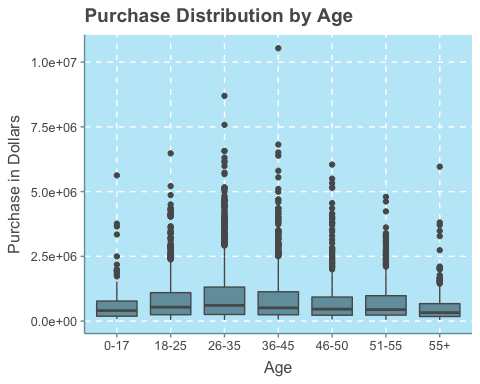
#Create new data frame grouping by occupation and marital status, and taking the mean purchase of each occupation and status.  
ggthemr\_reset()  
demo %>%  
 group\_by(Occupation) %>%  
 ggplot(aes(x = fct\_infreq(Occupation), y = sum(Purchase), fill = Marital\_Status)) +  
 geom\_bar(stat = "identity") +  
 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 are intrested in which occupation spend more and whether the married spending more or the unmarried on BlackFriday. Since marital status and occupation being factor (discrete) variables, and purchase bieng continuous variable, we use 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 like “Hou much does customers from occupation x spend on BlackFriday”. Taking into account th marital status the plot answers whether the married or the unmarried spend more on BlackFriday. We make the plot more readable by flipping the coordinates and reordring the purchasing by occupation in ascending order. Generally, unmarried customers spend more on BlackFriday and this hold true for most occupations. Specifically, occupation 17, 1, 7, 0 and 4 are the five occupations that spend the most money, over 2 trillions, of which occupation 4 purchases the most. Interestingly, for occupation 10, the married customers hardly spend money on BlackFriday.

**Plot 4. Purchase Distribution by Age**

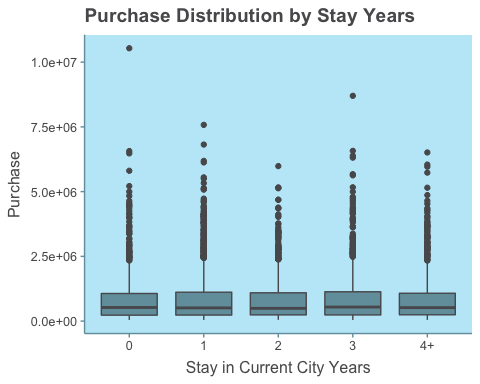
ggthemr('sky')  
demo %>%  
 ggplot(aes(x = Age, y = Purchase)) +  
 geom\_boxplot() +  
 labs(x = "Age", y = "Purchase in Dollars", title = "Purchase Distribution by Age")



Here,by boxplot, we want to have a look on whether there are similar purchasing patterns in different age ranges. Boxplot is used because age is a factor variable and purchase is a continuous variable. We make the plot fancier by setting the theme as sky. The boxplot can answer question that is there a difference in median purchase among the 7 groups. To conclude, the customers aged from 18-55 has the similar median purchase except for group 36-35, which has the higest median purchase and range. Whereas customers who are aged under 17 or over 55 seem to spend a little bit less.

**Plot 5. Purchase distribution by stay in city years**

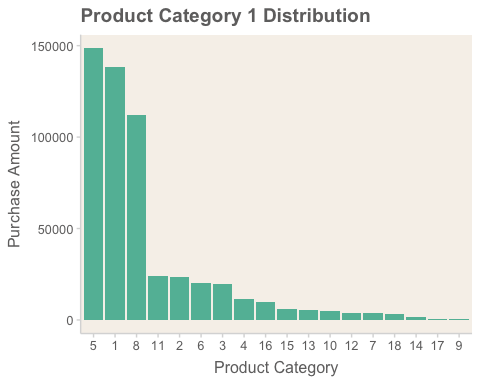
ggthemr('sky', layout = 'minimal')  
demo %>%  
 ggplot(aes(x = Stay\_In\_Current\_City\_Years, y = Purchase)) +  
 geom\_boxplot()+  
 labs(x = "Stay in Current City Years", y = "Purchase", title = "Purchase Distribution by Stay Years")



In the above plot, we want to find the purchase distribution by stay in current city years. Because the stay in current city years is catergory and the purchase is continuous, we choose boxplot to visualize the relationship. We use different theme to make the plot more beautiful and add labels and titles to make it clearer. From this plot, we find there may be no difference among the purchase distribution of different years. But we want to have more confidence to draw such conclusion. So we use anova test in the next part to find whether it is significant. Based on the above plot, we draw 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 making promotion decision.

**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 the gender is category and the purchase is continuous, we choose t-test for figuring out whether there is a difference between male and female 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: men have larger mean purchase than women.  
From the statistic analysis, we recommend the store needs to focus more on male customers because they have stronger purchase power than female and have more methods of promotion towards male customers 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.