Project 1

Kevil Khadka

4/10/2019

Project: Black Friday Analysis

1. Introduction

Black Friday knows as the best Friday of the year for those people who like shopping a lot. It is an informal name for the Friday following the Thanksgiving day in the United States.

The Black Friday data set is a sample of the transactions made in a retail store. Retailers like Amazon, E bay, Macy's, Walmart, Best buy and other many stores look for this day every year with the hope of many customers which will take advantage of door-busting deals.

With the help of this data set, I am going to do random experiments to get the answer of my questions. All questions are listed below:

Question_1. Is there any duplicate User id in the dataset? If yes, how many customers are registered in the USER_ID after removing duplicate id?

Question_2. How much each customer (each User_ID) spending on the black friday?

Question_3. Who spends the average amount of money during the black friday? Male or Female?

Question_4. According to the Age category, what age people are registered in the store? and what the location of their residence?

Thrugh the questions, I try to analyize what population age group purchase more on the black friday. According to simple hypothesis, it is clear that most adult will do shopping on the black friday, and age above 55+ would do less shopping. So, I am analysing the gender variable with other variables to find any differences.

2. Body

The black Friday data set is composed with 537577 observations with 12 different columns. Some of variables are categorical and quantitative variables.

Let's do quick look at the data set: Black Friday

```
library(readr)
BlackFriday <- read_csv("/Volumes/College/SPRING 2019/STAT 266/Project 1/BlackFrida
y.csv")
```

```
## Parsed with column specification:
## cols(
##   User_ID = col_double(),
##   Product_ID = col_character(),
##   Gender = col_character(),
##   Age = col_character(),
##   Occupation = col_double(),
##   City_Category = col_character(),
##   Stay_In_Current_City_Years = col_character(),
##   Marital_Status = col_double(),
##   Product_Category_1 = col_double(),
##   Product_Category_2 = col_double(),
##   Product_Category_3 = col_double(),
##   Purchase = col_double()
## )
```

```
summary(BlackFriday)
```

```
##     User_ID         Product_ID           Gender
##  Min.   :1000001   Length:537577      Length:537577
##  1st Qu.:1001495   Class :character   Class :character
##  Median :1003031   Mode  :character   Mode  :character
##  Mean   :1002992
##  3rd Qu.:1004417
##  Max.   :1006040
##
##      Age              Occupation     City_Category
##  Length:537577      Min.   : 0.000   Length:537577
##  Class :character   1st Qu.: 2.000   Class :character
##  Mode  :character   Median : 7.000   Mode  :character
##                     Mean   : 8.083
##                     3rd Qu.:14.000
##                     Max.   :20.000
##
##  Stay_In_Current_City_Years Marital_Status   Product_Category_1
##  Length:537577              Min.   :0.0000   Min.   : 1.000
##  Class :character           1st Qu.:0.0000   1st Qu.: 1.000
##  Mode  :character           Median :0.0000   Median : 5.000
##                             Mean   :0.4088   Mean   : 5.296
##                             3rd Qu.:1.0000   3rd Qu.: 8.000
##                             Max.   :1.0000   Max.   :18.000
##
##  Product_Category_2 Product_Category_3    Purchase
##  Min.   : 2.00      Min.   : 3.0       Min.   :  185
##  1st Qu.: 5.00      1st Qu.: 9.0       1st Qu.: 5866
##  Median : 9.00      Median :14.0       Median : 8062
##  Mean   : 9.84      Mean   :12.7       Mean   : 9334
##  3rd Qu.:15.00      3rd Qu.:16.0       3rd Qu.:12073
##  Max.   :18.00      Max.   :18.0       Max.   :23961
##  NA's   :166986     NA's   :373299
```

```
head(BlackFriday)
```

```
## # A tibble: 6 x 12
##   User_ID Product_ID Gender Age   Occupation City_Category Stay_In_Current…
##     <dbl> <chr>      <chr>  <chr>      <dbl> <chr>         <chr>
## 1 1000001 P00069042  F      0-17          10 A             2
## 2 1000001 P00248942  F      0-17          10 A             2
## 3 1000001 P00087842  F      0-17          10 A             2
## 4 1000001 P00085442  F      0-17          10 A             2
## 5 1000002 P00285442  M      55+           16 C             4+
## 6 1000003 P00193542  M      26-35         15 A             3
## # … with 5 more variables: Marital_Status <dbl>, Product_Category_1 <dbl>,
## #   Product_Category_2 <dbl>, Product_Category_3 <dbl>, Purchase <dbl>
```

This data set contains 12 different columns, each representing a corresponding variable.

1. User_ID - identification of the customer; quantitative variable

2. Product_ID - identification code for the product; categorical variable
3. Gender - Sex of customer; M(male) and F(Female); categorical variable
4. Age - Age of customer; divided on categorical variable of age group; maybe categorical variable
5. Occuption - Occuption of customer; given on quatitative variable
6. City_Category - Residence of customer; Catergorical Variable
7. Stay_In_Current_City_Years - Number of years customers stay in current city; Quantitative variable
8. Marital_Status - 0 = Single and 1 = Married; Categorical Variable
9. Product_Category_1 - Product may belong to category 1
10. Product_Category_2 - Product may belong to category 2
11. Product_Category_3 - Product may belong to category 3
12. Purchase: Purchase amount of product by customer; Quantitative variable

Let's find out the confidence interval of 95% for the mean. We are going to find out how much each customer purcharse the product during the Black Friday.

```
library(readr)
BlackFriday <- read_csv("/Volumes/College/SPRING 2019/STAT 266/Project 1/BlackFrida
y.csv")
```

```
## Parsed with column specification:
## cols(
##   User_ID = col_double(),
##   Product_ID = col_character(),
##   Gender = col_character(),
##   Age = col_character(),
##   Occupation = col_double(),
##   City_Category = col_character(),
##   Stay_In_Current_City_Years = col_character(),
##   Marital_Status = col_double(),
##   Product_Category_1 = col_double(),
##   Product_Category_2 = col_double(),
##   Product_Category_3 = col_double(),
##   Purchase = col_double()
## )
```

```
summary(BlackFriday)
```

```
##     User_ID         Product_ID          Gender
##  Min.   :1000001   Length:537577      Length:537577
##  1st Qu.:1001495   Class :character   Class :character
##  Median :1003031   Mode  :character   Mode  :character
##  Mean   :1002992
##  3rd Qu.:1004417
##  Max.   :1006040
##
##      Age              Occupation     City_Category
##  Length:537577      Min.   : 0.000   Length:537577
##  Class :character   1st Qu.: 2.000   Class :character
##  Mode  :character   Median : 7.000   Mode  :character
##                     Mean   : 8.083
##                     3rd Qu.:14.000
##                     Max.   :20.000
##
##  Stay_In_Current_City_Years Marital_Status   Product_Category_1
##  Length:537577              Min.   :0.0000   Min.   : 1.000
##  Class :character           1st Qu.:0.0000   1st Qu.: 1.000
##  Mode  :character           Median :0.0000   Median : 5.000
##                             Mean   :0.4088   Mean   : 5.296
##                             3rd Qu.:1.0000   3rd Qu.: 8.000
##                             Max.   :1.0000   Max.   :18.000
##
##  Product_Category_2 Product_Category_3    Purchase
##  Min.   : 2.00      Min.   : 3.0       Min.   :  185
##  1st Qu.: 5.00      1st Qu.: 9.0       1st Qu.: 5866
##  Median : 9.00      Median :14.0       Median : 8062
##  Mean   : 9.84      Mean   :12.7       Mean   : 9334
##  3rd Qu.:15.00      3rd Qu.:16.0       3rd Qu.:12073
##  Max.   :18.00      Max.   :18.0       Max.   :23961
##  NA's   :166986     NA's   :373299
```

```
length(BlackFriday$Gender)
```

```
## [1] 537577
```

```
mean(BlackFriday$Purchase)
```

```
## [1] 9333.86
```

```
sd(BlackFriday$Purchase)
```

```
## [1] 4981.022
```

Mean for the Purchase variable : 9333.86 Standard Deviation for the purchase variable : 4981.022

Now we can calculate an error for the mean

```
error <- qt(0.975,df=length(BlackFriday$Gender)-1)*sd(BlackFriday$Purchase)/sqrt(len
gth(BlackFriday$Purchase))
error
```

```
## [1] 13.31518
```

The error for the mean : 13.31518 which is pretty high.

We can find the confidence interval by adding and subtracting the error from the mean:

```
left <- mean(BlackFriday$Purchase)-error
right <- mean(BlackFriday$Purchase)+error
left
```

```
## [1] 9320.545
```

```
right
```

```
## [1] 9347.175
```

The confidence interval for the mean is: (9320.545, 9347.175).

2.1. Question_1

After looking at the dataset, we can find the number of duplicates data for different variables.

First, Let's find the gender of customer stored in the dataset by using User_ID to remove duplicates.

```
customer_gender = BlackFriday %>%
  select(User_ID, Gender) %>%
  group_by(User_ID) %>%
  distinct()
#View(customer_gender)
head(customer_gender)
```

```
## # A tibble: 6 x 2
## # Groups:   User_ID [6]
##   User_ID Gender
##     <dbl> <chr>
## 1 1000001 F
## 2 1000002 M
## 3 1000003 M
## 4 1000004 M
## 5 1000005 M
## 6 1000006 F
```

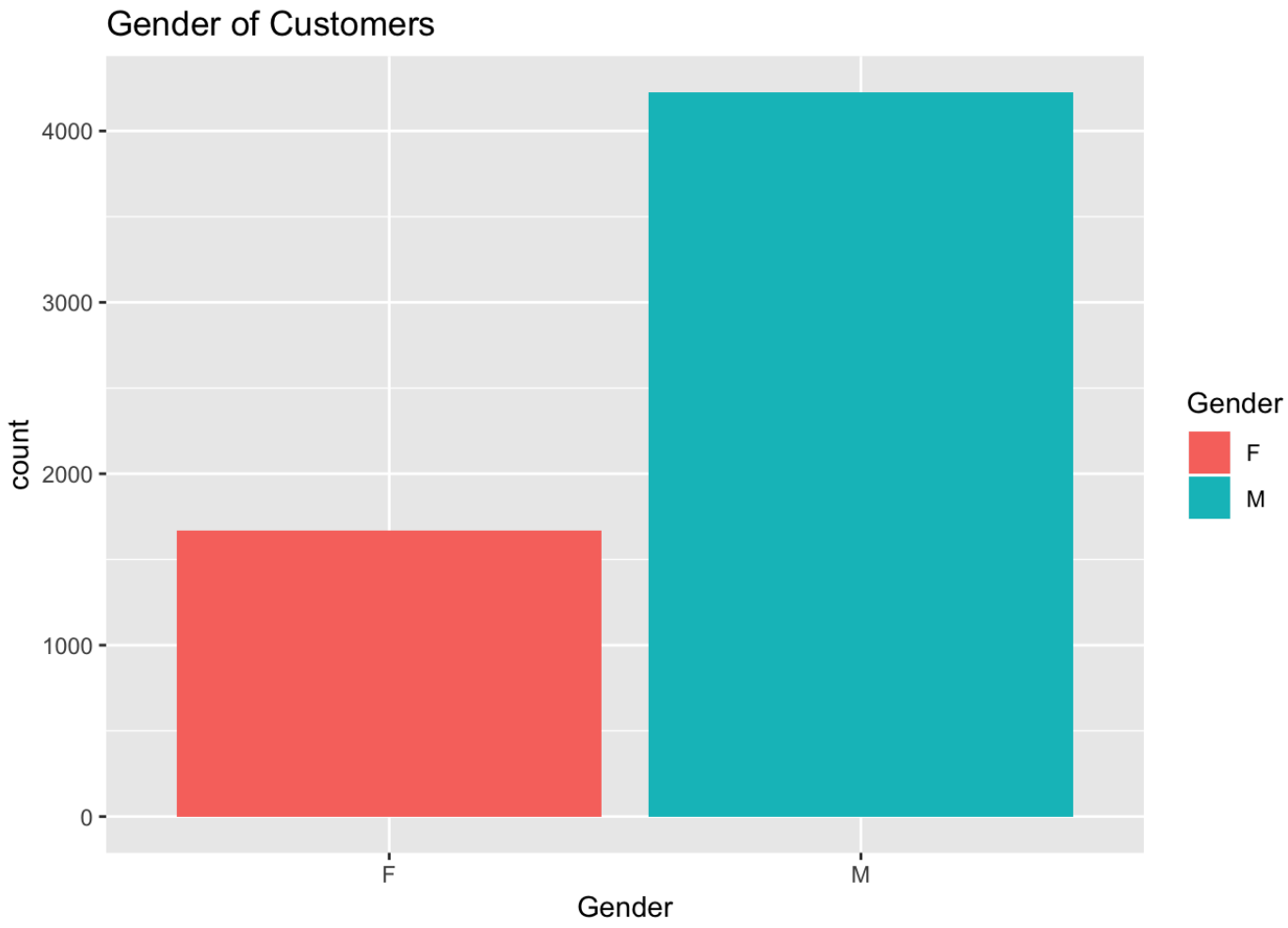
```
summary(customer_gender$Gender)
```

```
##    Length     Class      Mode
##      5891 character character
```

Conclusion: We found out that there are 5891 customers registered in the retail stores. With same User ID, the number of males seem greater compare with female number.

We can plot the dataset of gender using ggplot.

```
ggplot(data = customer_gender) + geom_bar(aes(x = Gender, y=..count.., fill = Gende
r)) +
                labs(title = 'Gender of Customers')
```



2.2. Question_2

Let's find out the amount of purchase did by each customers according to their USER_ID and gender.

Amount of purchase made by each customers

```
each_customer_purchase = BlackFriday %>%
  select(User_ID, Gender, Purchase) %>%
  group_by(User_ID) %>%
  arrange(User_ID) %>%
  summarise(total_amount_purchase = sum(Purchase)) %>%
  arrange(desc(total_amount_purchase))

gender_of_customer_purchase = full_join(each_customer_purchase, customer_gender, by=
"User_ID")

#View(gender_of_customer_purchase)
head(gender_of_customer_purchase)
```

```
## # A tibble: 6 x 3
##   User_ID total_amount_purchase Gender
##     <dbl>                 <dbl> <chr>
## 1 1004277              10536783 M
## 2 1001680               8699232 M
## 3 1002909               7577505 M
## 4 1001941               6817493 M
## 5 1000424               6573609 M
## 6 1004448               6565878 M
```

```
summary(gender_of_customer_purchase)
```

```
##     User_ID        total_amount_purchase    Gender
##  Min.   :1000001   Min.   :   44108      Length:5891
##  1st Qu.:1001518   1st Qu.:  234914      Class :character
##  Median :1003026   Median :  512612      Mode  :character
##  Mean   :1003025   Mean   :  851752
##  3rd Qu.:1004532   3rd Qu.: 1099005
##  Max.   :1006040   Max.   :10536783
```

Conclusion: we see that each user id is spending the high amount of money during the black friday. We can see the mean for the total amount of each customer purchase is 851752.

2.3. Question_3

Now, we can try to find out the average spending gender who usaully spends more money on the BlackFriday.

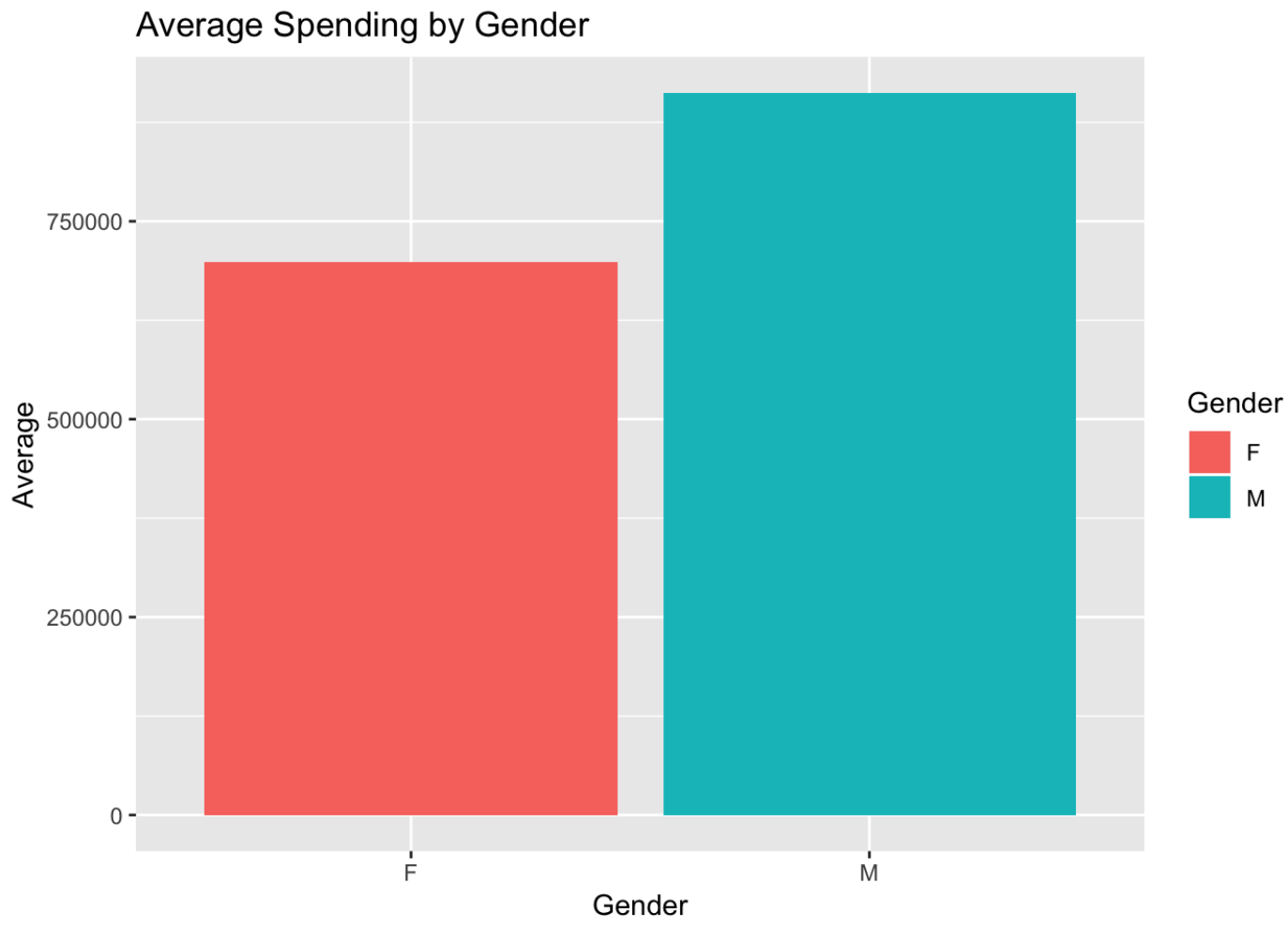
```
average_spending_gender = gender_of_customer_purchase %>%
  group_by(Gender) %>%
  summarize(Purchase = sum(as.numeric(total_amount_purchase)), Count = n(),
            Average = Purchase/Count) %>%
  arrange(desc(Purchase))

head(average_spending_gender)
```

```
## # A tibble: 2 x 4
##   Gender   Purchase Count Average
##   <chr>       <dbl> <int>   <dbl>
## 1 M      3853044357  4225 911963.
## 2 F      1164624021  1666 699054.
```

We can conclude that during black friday, male customers purchase more amount of products than female. We can plot the graph to have better understanding of above result.

```
ggplot(data = average_spending_gender) + geom_bar(aes(x = Gender, y=Average, fill =
Gender), stat = 'identity') + labs(title = 'Average Spending by Gender')
```



2.4. Question_4

The dataset contains the age column to distinguish each customer in different age group.

```
customer_age <- BlackFriday %>%
  select(User_ID, Age) %>%
  distinct() %>%
  count(Age)
head(customer_age)
```

```
## # A tibble: 6 x 2
##   Age       n
##   <chr> <int>
## 1 0-17    218
## 2 18-25  1069
## 3 26-35  2053
## 4 36-45  1167
## 5 46-50   531
## 6 51-55   481
```

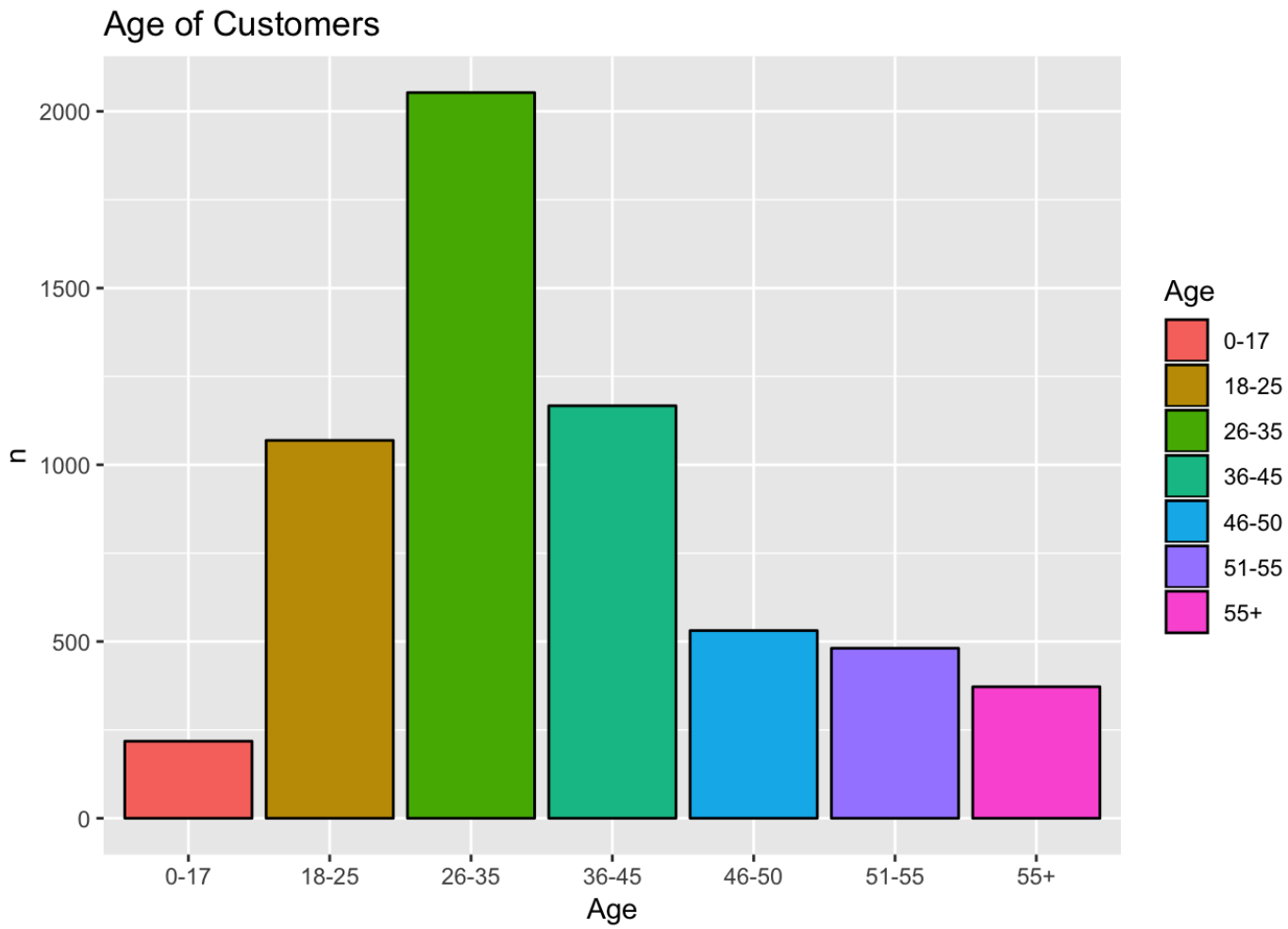
```
summary(customer_age)
```

```
##      Age                  n
##  Length:7           Min.   : 218.0
##  Class :character   1st Qu.: 426.5
##  Mode  :character   Median : 531.0
##                     Mean   : 841.6
##                     3rd Qu.:1118.0
##                     Max.   :2053.0
```

It is found out that the highest number of customers registerd in the store are from the age group of 26-35. Probabily, family people have done alot of shopping during the black friday. We thought adult group which is 18-25 age would have highest number for shopping activity. But it came on third place after age group of 36-45.

Let’s plot the data on the bar graph to have clear visualization data.

```
ggplot(data =customer_age) + geom_bar(aes(x=Age, y=n,fill=Age), color='black', stat=
'identity')+labs(title = "Age of Customers")
```



3. Conclusion

Through the help of dataset, i figure out how the shopping activity is different according to the gender and age group. The males are spending more money for purchasing more product than female. In future, i would like to find out more about marital status of the customer. I want to find out if marital status effects purchase variable or not.