INFSCI 2160 DATA MINING - Homework 1

Jing Pang

1/8/2019

In this assignment, we will do exploratory analysis on the Black Friday data from Kaggle: https://www.kaggle.com/mehdidag/black-friday in order to complete several questions.

First of all, we have to load the dataset into RStudio for preparation of the further study.

```
blackfriday <- read.csv("~/R-workspace/BlackFriday.csv", header=TRUE)
```

Question 1

- How many observations are in the dataset? How many features?

```
nrow(blackfriday)
```

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

```
ncol(blackfriday)
```

```
## [1] 12
```

There is total number of 537577 observations in the dataset, and there is 12 features in total.

- How many nulls are in the dataset?

```
sum(is.na(blackfriday))
```

```
## [1] 540285
```

There is 540285 nulls in the dataset.

- Summarize the dataset.

```
summary(blackfriday)
```

```
##     User_ID            Product_ID     Gender        Age
##  Min.   :1000001   P00265242:  1858   F:132197   0-17 : 14707
##  1st Qu.:1001495   P00110742:  1591   M:405380   18-25: 97634
##  Median :1003031   P00025442:  1586              26-35:214690
##  Mean   :1002992   P00112142:  1539              36-45:107499
##  3rd Qu.:1004417   P00057642:  1430              46-50: 44526
##  Max.   :1006040   P00184942:  1424              51-55: 37618
##                    (Other)  :528149              55+  : 20903
##    Occupation     City_Category Stay_In_Current_City_Years
##  Min.   : 0.000   A:144638      0 : 72725
##  1st Qu.: 2.000   B:226493      1 :189192
##  Median : 7.000   C:166446      2 : 99459
##  Mean   : 8.083                 3 : 93312
##  3rd Qu.:14.000                 4+: 82889
##  Max.   :20.000
##
##  Marital_Status   Product_Category_1 Product_Category_2 Product_Category_3
##  Min.   :0.0000   Min.   : 1.000     Min.   : 2.00      Min.   : 3.0
##  1st Qu.:0.0000   1st Qu.: 1.000     1st Qu.: 5.00      1st Qu.: 9.0
```

```
##  Median :0.0000   Median : 5.000     Median : 9.00      Median :14.0
##  Mean   :0.4088   Mean   : 5.296     Mean   : 9.84      Mean   :12.7
##  3rd Qu.:1.0000   3rd Qu.: 8.000     3rd Qu.:15.00      3rd Qu.:16.0
##  Max.   :1.0000   Max.   :18.000     Max.   :18.00      Max.   :18.0
##                                      NA's   :166986     NA's   :373299
##     Purchase
##  Min.   :  185
##  1st Qu.: 5866
##  Median : 8062
##  Mean   : 9334
##  3rd Qu.:12073
##  Max.   :23961
##
```

- Find the min, max, 1st quartile, 3rd quartile, median, and mean of the 'Product_Category_1' column.

```
summary(blackfriday$Product_Category_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   5.000   5.296   8.000  18.000
```

- What datatype is the 'Age' column?

```
class(blackfriday$Age)
```

```
## [1] "factor"
```

Question 2

- Convert the "Marital_Status" column to a factor

```
blackfriday$Marital_Status <- as.factor(blackfriday$Marital_Status)
class(blackfriday$Marital_Status)
```

```
## [1] "factor"
```

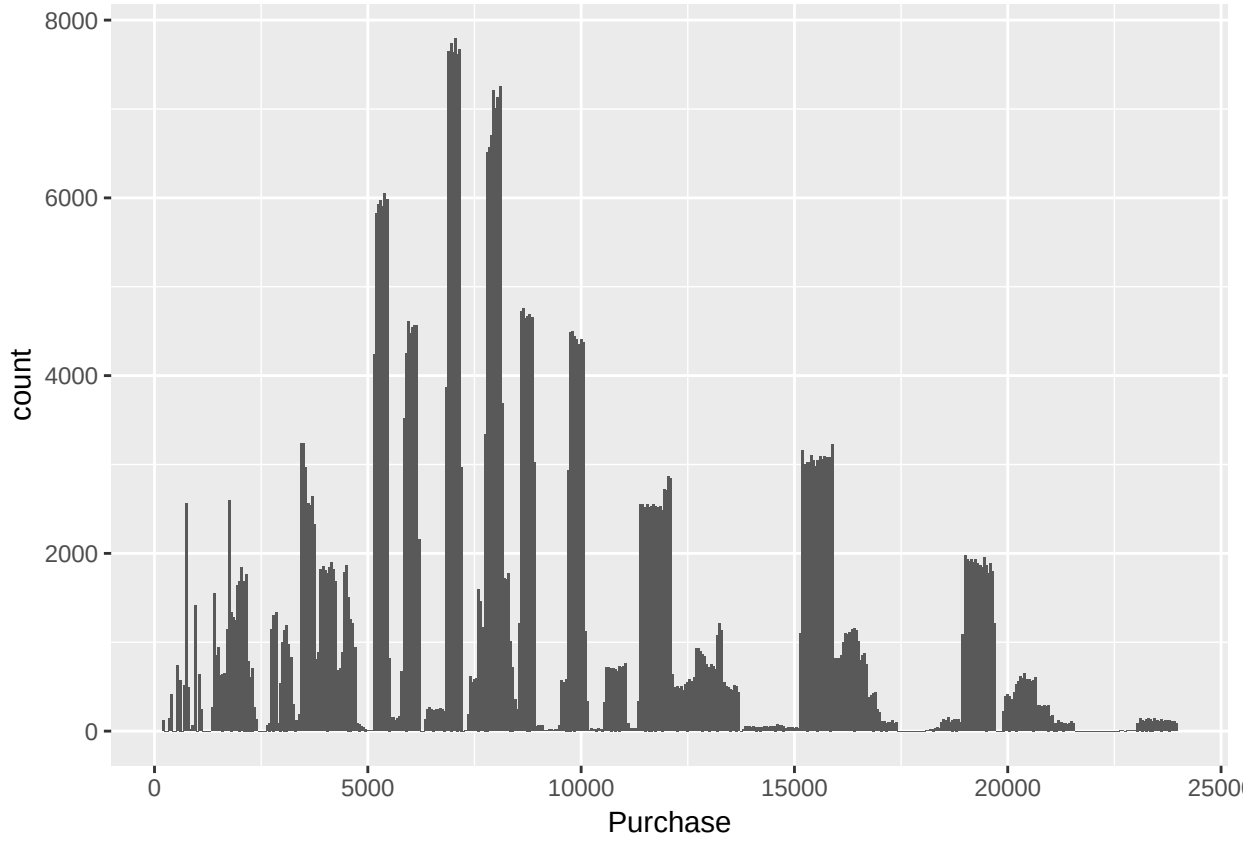
Question 3

- Create a histogram of the 'Purchase' column using ggplot2

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
ggplot(blackfriday, aes(x = Purchase)) + geom_histogram(binwidth = 50)
```



Question 4

- Create a table to analyze the 'City_Category' column.

```
table(blackfriday$City_Category)
```

```
##
##      A      B      C
## 144638 226493 166446
```

Question 5

- Filter the dataset where Gender = M and Marital_Status = 1. How many observations are there?

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
male_marital_data <- blackfriday %>%
  filter(Gender == "M") %>%
  filter(Marital_Status == 1)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
summary(male_marital_data)
```

```
##     User_ID            Product_ID     Gender        Age
##  Min.   :1000004   P00265242:   566   F:     0   0-17 :    0
##  1st Qu.:1001457   P00110742:   495   M:164537   18-25:14524
##  Median :1002989   P00025442:   490              26-35:64207
##  Mean   :1002965   P00112142:   477              36-45:32392
##  3rd Qu.:1004446   P00057642:   465              46-50:22397
##  Max.   :1006033   P00184942:   458              51-55:20829
##                    (Other)  :161586              55+  :10188
##    Occupation     City_Category Stay_In_Current_City_Years Marital_Status
##  Min.   : 0.000   A:41744       0 :21139                   0:     0
##  1st Qu.: 2.000   B:69683       1 :59421                   1:164537
##  Median : 7.000   C:53110       2 :30551
##  Mean   : 8.577                 3 :26846
##  3rd Qu.:15.000                 4+:26580
##  Max.   :20.000
##
##  Product_Category_1 Product_Category_2 Product_Category_3    Purchase
##  Min.   : 1.000     Min.   : 2.00      Min.   : 3.00      Min.   :  187
##  1st Qu.: 1.000     1st Qu.: 5.00      1st Qu.: 9.00      1st Qu.: 5909
##  Median : 5.000     Median : 9.00      Median :15.00      Median : 8108
##  Mean   : 5.296     Mean   : 9.88      Mean   :12.83      Mean   : 9485
##  3rd Qu.: 8.000     3rd Qu.:15.00      3rd Qu.:16.00      3rd Qu.:12420
##  Max.   :18.000     Max.   :18.00      Max.   :18.00      Max.   :23961
##                     NA's   :51440      NA's   :113641
```

- Make a table of the age column. Which age group has the most observations?

```
AgeTable <- table(blackfriday$Age)
AgeFrame <- as.data.frame(AgeTable)
names(AgeFrame) <- c("Age", "Freq")
library(dplyr)
AgeFrame %>%
  filter(Freq == max(Freq))
```

```
##     Age   Freq
## 1 26-35 214690
```