R Case Study 1 - Retail

Question 1. Merge the Customer, prod_cat_info and Transactions Files as Customer_Final

a. Using base merge()

```
# Q1 Merge the 3 Files as customer_final(Base merge)

Customer_Final <- merge(Transactions,Customer, by.x = "cust_id",
                        by.y = "customer_Id",all.x = T)

Customer_Final <- merge(Customer_Final,prod_cat_info, by.x = "prod_cat_code",
                        by.y = "prod_cat_code", all.x = T)
```

b. Using dplyr

```
# Q1 Merge the 3 Files as customer_final (Dplyr merge)

Customer_Final <- dplyr::left_join(Transactions,Customer, by = c("cust_id" = "customer_Id"))

Customer_Final <- dplyr::left_join(Customer_Final,prod_cat_info,
                                   by = "prod_cat_code")
```

Question 2 Summary report

a.Variable names and their data types of Customer_Final dataset

```
#Q2 a. column names and corresponding data types
str(Customer_Final)
```

```
## ’data.frame':    99293 obs. of  16 variables:
##  $ transaction_id   : chr  "80712190438" "80712190438" "80712190438" "29258453508" ...
##  $ cust_id          : chr  "270351" "270351" "270351" "270384" ...
##  $ tran_date        : Date, format: "2014-02-28" "2014-02-28" ...
##  $ prod_subcat_code : chr  "1" "1" "1" "5" ...
##  $ prod_cat_code    : chr  "1" "1" "1" "3" ...
##  $ Qty              : int  -5 -5 -5 -5 -5 -5 -5 -5 -2 -2 ...
##  $ Rate             : int  -772 -772 -772 -1497 -1497 -1497 -1497 -1497 -791 -791 ...
##  $ Tax              : num  405 405 405 786 786 ...
##  $ total_amt        : num  -4265 -4265 -4265 -8271 -8271 ...
##  $ Store_type       : chr  "e-Shop" "e-Shop" "e-Shop" "e-Shop" ...
##  $ DOB              : Date, format: "1981-09-26" "1981-09-26" ...
##  $ Gender           : chr  "M" "M" "M" "F" ...
##  $ city_code        : chr  "5" "5" "5" "8" ...
##  $ prod_cat         : chr  "Clothing" "Clothing" "Clothing" "Electronics" ...
##  $ prod_sub_cat_code: chr  "4" "1" "3" "4" ...
##  $ prod_subcat      : chr  "Mens" "Women" "Kids" "Mobiles" ...
```

b. Top 10 records of Customer_Final dataset

```
#Q2 b.Top 10 observations
head(Customer_Final,10)
```

```
##    transaction_id cust_id  tran_date prod_subcat_code prod_cat_code Qty  Rate
## 1     80712190438  270351 2014-02-28                1             1  -5  -772
## 2     80712190438  270351 2014-02-28                1             1  -5  -772
## 3     80712190438  270351 2014-02-28                1             1  -5  -772
## 4     29258453508  270384 2014-02-27                5             3  -5 -1497
## 5     29258453508  270384 2014-02-27                5             3  -5 -1497
## 6     29258453508  270384 2014-02-27                5             3  -5 -1497
## 7     29258453508  270384 2014-02-27                5             3  -5 -1497
## 8     29258453508  270384 2014-02-27                5             3  -5 -1497
## 9     51750724947  273420 2014-02-24                6             5  -2  -791
## 10    51750724947  273420 2014-02-24                6             5  -2  -791
##        Tax total_amt Store_type        DOB Gender city_code    prod_cat
## 1  405.300 -4265.300     e-Shop 1981-09-26      M         5    Clothing
## 2  405.300 -4265.300     e-Shop 1981-09-26      M         5    Clothing
## 3  405.300 -4265.300     e-Shop 1981-09-26      M         5    Clothing
## 4  785.925 -8270.925     e-Shop 1973-05-11      F         8 Electronics
## 5  785.925 -8270.925     e-Shop 1973-05-11      F         8 Electronics
## 6  785.925 -8270.925     e-Shop 1973-05-11      F         8 Electronics
## 7  785.925 -8270.925     e-Shop 1973-05-11      F         8 Electronics
## 8  785.925 -8270.925     e-Shop 1973-05-11      F         8 Electronics
## 9  166.110 -1748.110   TeleShop 1992-07-27      M         8       Books
## 10 166.110 -1748.110   TeleShop 1992-07-27      M         8       Books
##    prod_sub_cat_code         prod_subcat
## 1                  4                Mens
## 2                  1               Women
## 3                  3                Kids
## 4                  4             Mobiles
## 5                  5           Computers
## 6                  8 Personal Appliances
## 7                  9             Cameras
## 8                 10     Audio and video
## 9                  7             Fiction
## 10                12            Academic
```

c. 5-number summary of continuous variables

```
#Q2 c.5 number summary for continuous variables

numericvariables <- Customer_Final[ ,sapply(Customer_Final, is.numeric)]
summary(numericvariables)
```

```
##       Qty              Rate              Tax            total_amt
##  Min.   :-5.000   Min.   :-1499.0   Min.   :  7.35   Min.   :-8270.9
##  1st Qu.: 1.000   1st Qu.:  313.0   1st Qu.: 98.28   1st Qu.:  762.5
##  Median : 3.000   Median :  713.0   Median :199.92   Median : 1761.4
##  Mean   : 2.438   Mean   :  637.9   Mean   :248.87   Mean   : 2114.6
##  3rd Qu.: 4.000   3rd Qu.: 1109.0   3rd Qu.:366.98   3rd Qu.: 3585.7
##  Max.   : 5.000   Max.   : 1500.0   Max.   :787.50   Max.   : 8287.5
```

d. Frequency tables for Categorical variables

```
# Q2 d.Frequency table for all categorical variables

categoricalvariables <- Customer_Final[ ,sapply(Customer_Final, is.character)]
categoricalvars_F <- categoricalvariables[,-c(1,2)] #removing id variables
#Frequency table for categorical variables


Freq1<- data.frame(dplyr::group_by(categoricalvars_F,prod_subcat_code)%>%
                     summarise(Count = n())%>%arrange(prod_subcat_code))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
Freq1
```

```
##    prod_subcat_code Count
## 1                 1  7847
## 2                10 14932
## 3                11 10302
## 4                12 10050
## 5                 2  4028
## 6                 3 12294
## 7                 4 13073
## 8                 5  4790
## 9                 6  5934
## 10                7  6258
## 11                8  4860
## 12                9  4925
```

```
Freq2<- data.frame(dplyr::group_by(categoricalvars_F,prod_cat_code)%>%
                     summarise(Count = n()))%>%arrange(prod_cat_code)
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
Freq2
```

```
##   prod_cat_code Count
## 1             1  8880
## 2             2  8997
## 3             3 24490
## 4             4  3996
## 5             5 36414
## 6             6 16516
```

```
Freq3<-data.frame(dplyr::group_by(categoricalvars_F,Store_type)%>%
                    summarise(Count = n()))%>%arrange(Store_type)
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
Freq3
```

```
##       Store_type Count
## 1         e-Shop 40185
## 2 Flagship store 19814
## 3            MBR 19974
## 4       TeleShop 19320
```

```
Freq4<- data.frame(dplyr::filter(categoricalvars_F,Gender != "")
                   %>%group_by(Gender)
                   %>%summarise(Count = n()))%>%arrange(Gender)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Freq4
```

```
##   Gender Count
## 1      F 48202
## 2      M 51051
```

```
Freq5<- data.frame(dplyr::filter(categoricalvars_F,city_code != "" )%>%
                     group_by(city_code)%>%
                     summarise(Count = n()))%>%arrange(city_code)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Freq5
```

```
##    city_code Count
## 1          1  9717
## 2         10  9976
## 3          2  9843
## 4          3 10467
## 5          4 10571
## 6          5 10116
## 7          6  9130
## 8          7 10258
## 9          8  9965
## 10         9  9214
```

```
Freq6 <-data.frame(dplyr::group_by(categoricalvars_F,prod_cat)%>%
                     summarise(Count = n()))%>%arrange(prod_cat)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Freq6
```

```
##           prod_cat Count
## 1             Bags  3996
## 2            Books 36414
## 3         Clothing  8880
## 4      Electronics 24490
## 5         Footwear  8997
## 6 Home and kitchen 16516
```

```
Freq7<- data.frame(dplyr::group_by(categoricalvars_F,prod_sub_cat_code)
                         %>%summarise(Count = n())
                   %>%arrange(prod_sub_cat_code))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Freq7
```

```
##    prod_sub_cat_code Count
## 1                  1  7957
## 2                 10 15096
## 3                 11 10198
## 4                 12 10198
## 5                  2  4129
## 6                  3 12028
## 7                  4 12855
## 8                  5  4898
## 9                  6  6069
## 10                 7  6069
## 11                 8  4898
## 12                 9  4898
```

```
Freq8<- data.frame(dplyr::group_by(categoricalvars_F,prod_subcat)%>%
                       summarise(Count = n()))%>%arrange(prod_subcat)
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

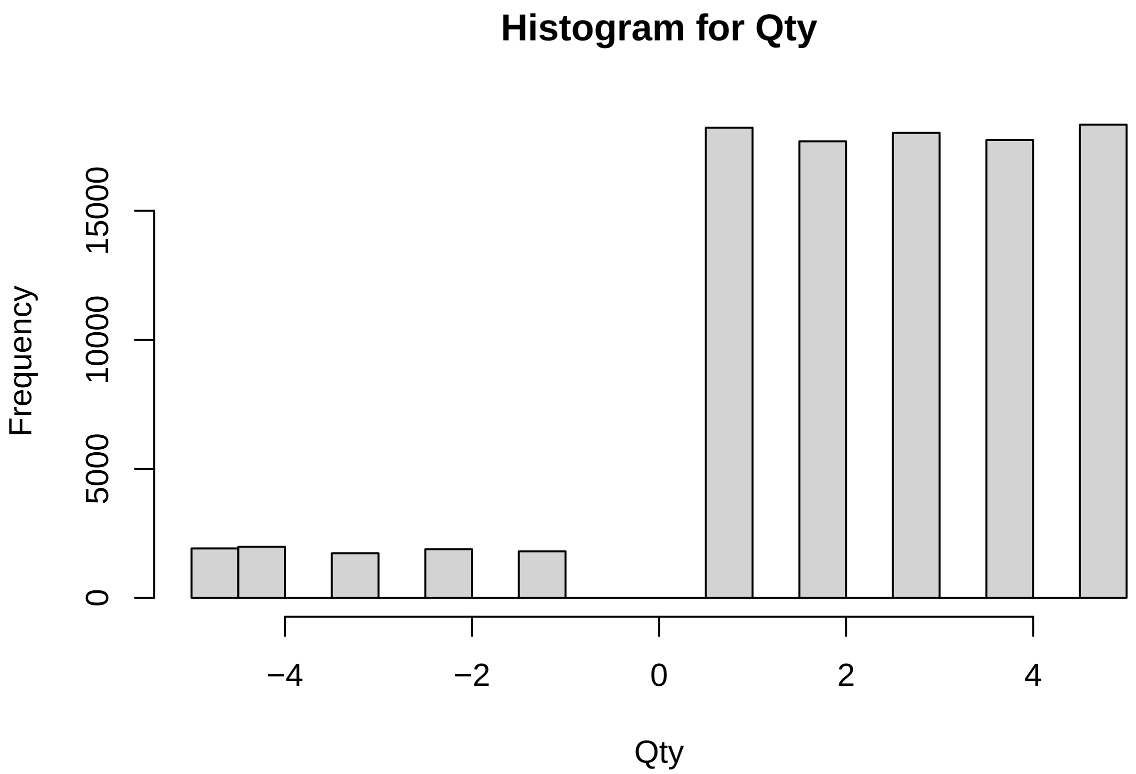
```
Freq8
```

```
##            prod_subcat Count
## 1             Academic  6069
## 2      Audio and video  4898
## 3                 Bath  4129
## 4              Cameras  4898
## 5             Children  6069
## 6               Comics  6069
## 7            Computers  4898
## 8                  DIY  6069
## 9              Fiction  6069
## 10          Furnishing  4129
## 11                Kids  5959
## 12             Kitchen  4129
## 13                Mens  7957
## 14             Mobiles  4898
## 15         Non-Fiction  6069
## 16 Personal Appliances  4898
## 17               Tools  4129
## 18               Women  7957
```

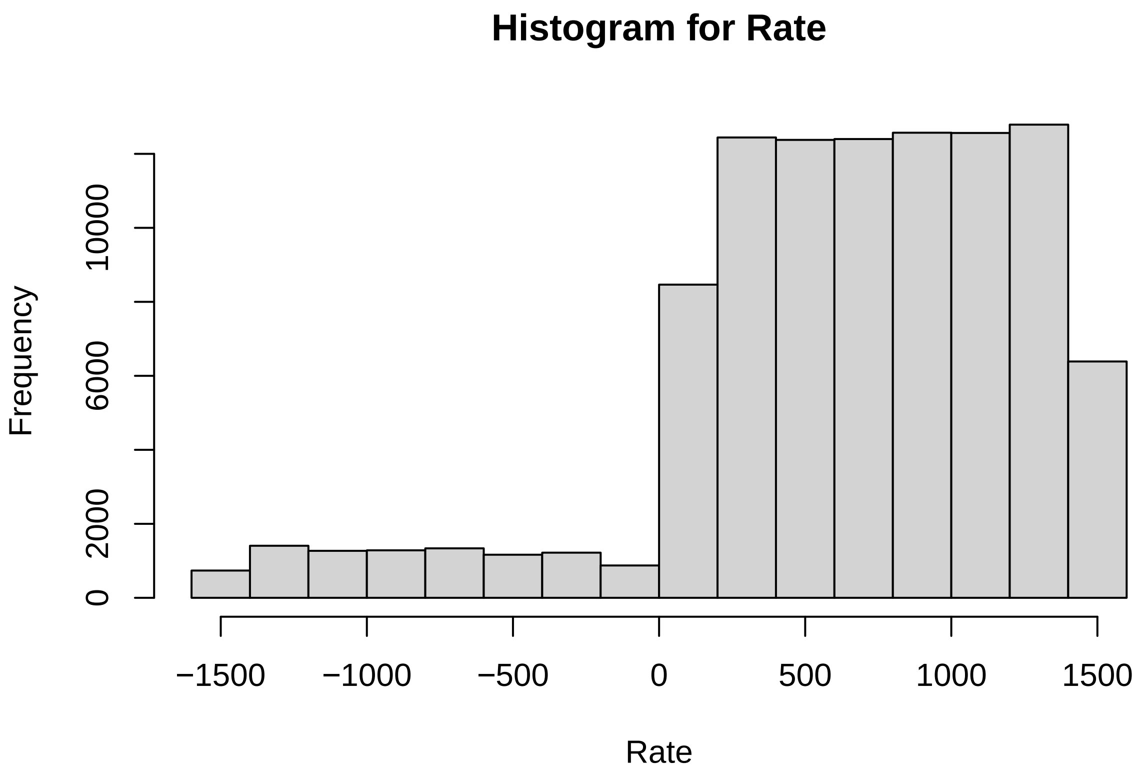
Question 3. Graphical representation of variables

a. Histogram for numeric variables

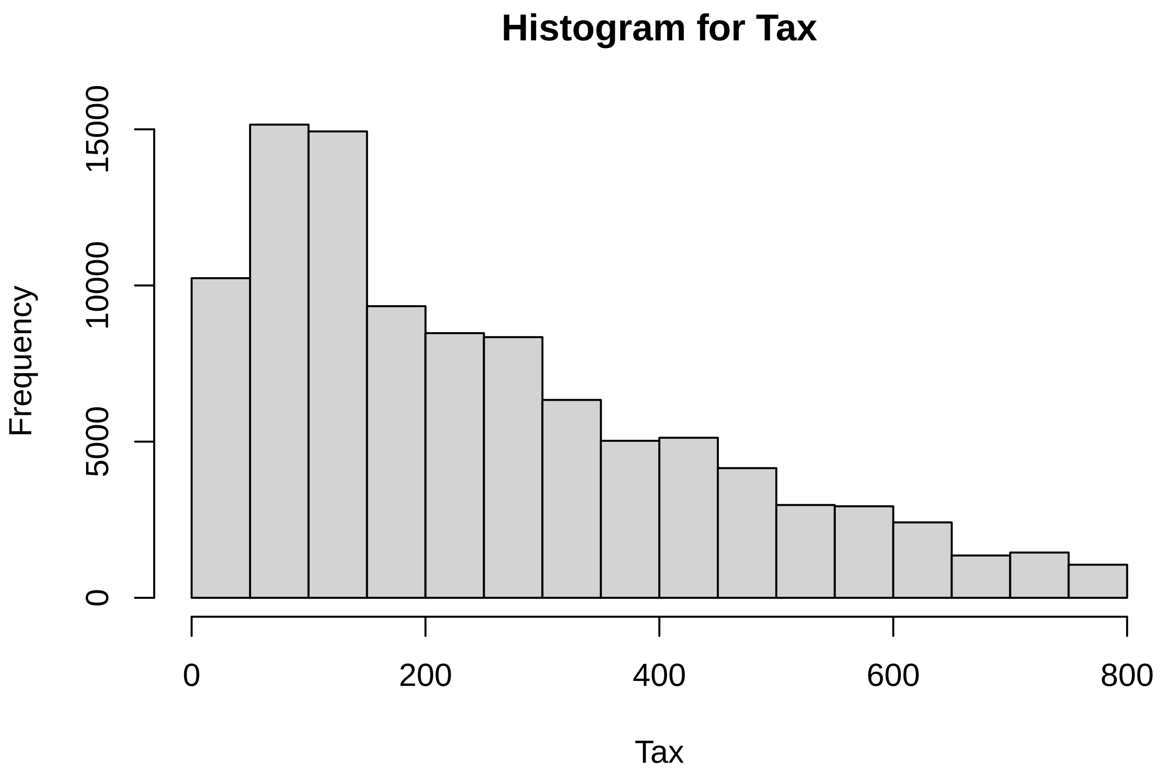
```
#Q3 1.Histogram for all the numeric variables

hist1 <- hist(numericvariables$Qty,xlab = "Qty",main="Histogram for Qty")
```



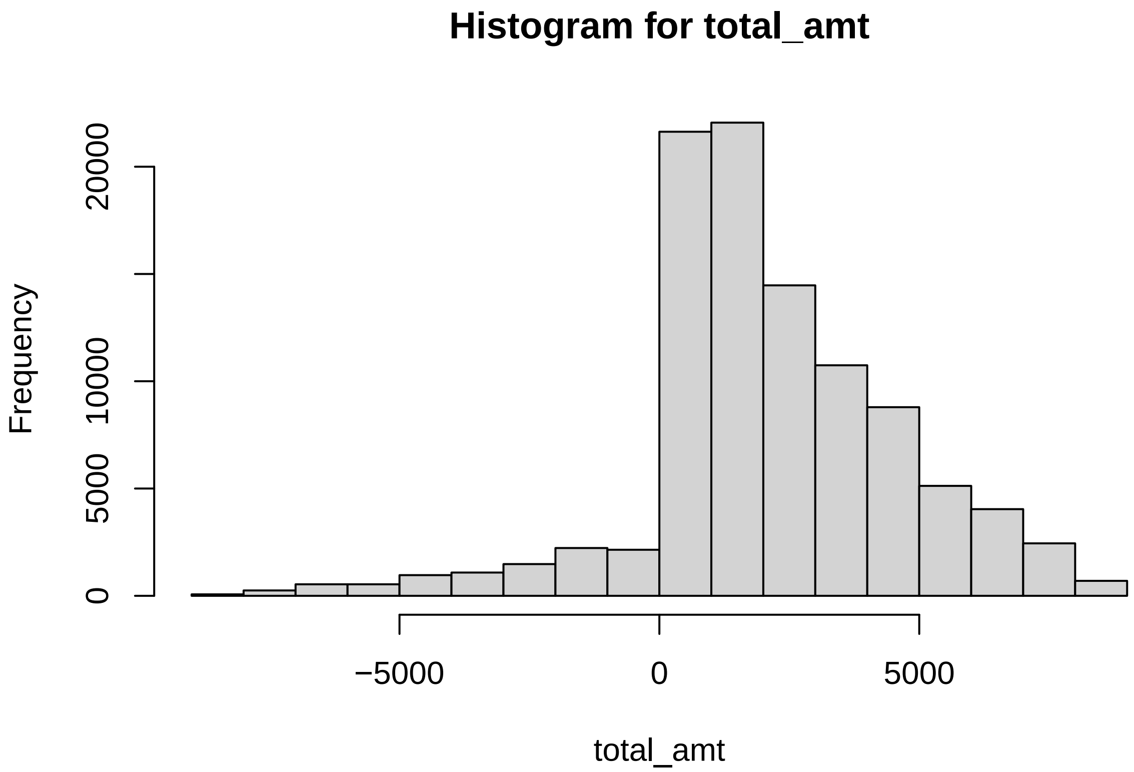
```
hist2 <- hist(numericvariables$Rate,xlab = "Rate",main="Histogram for Rate")
```



```
hist3 <- hist(numericvariables$Tax,xlab = "Tax",main="Histogram for Tax")
```

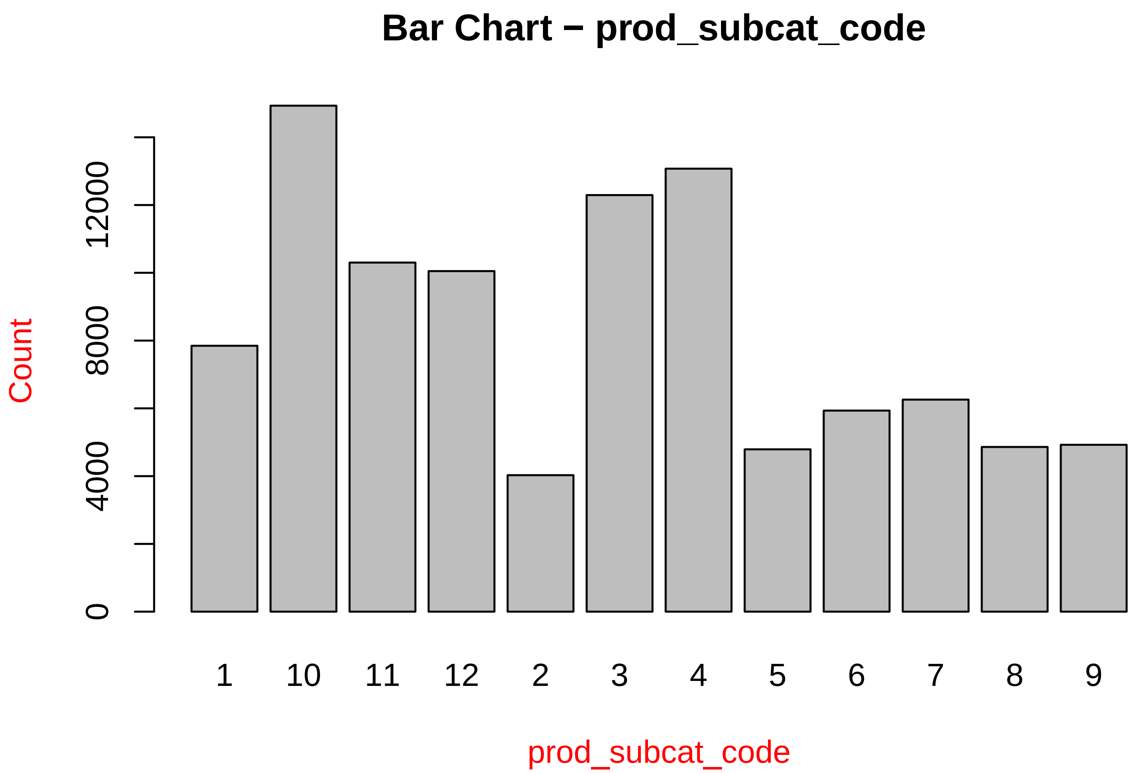


```
hist2 <- hist(numericvariables$total_amt,xlab = "total_amt",
              main="Histogram for total_amt")
```

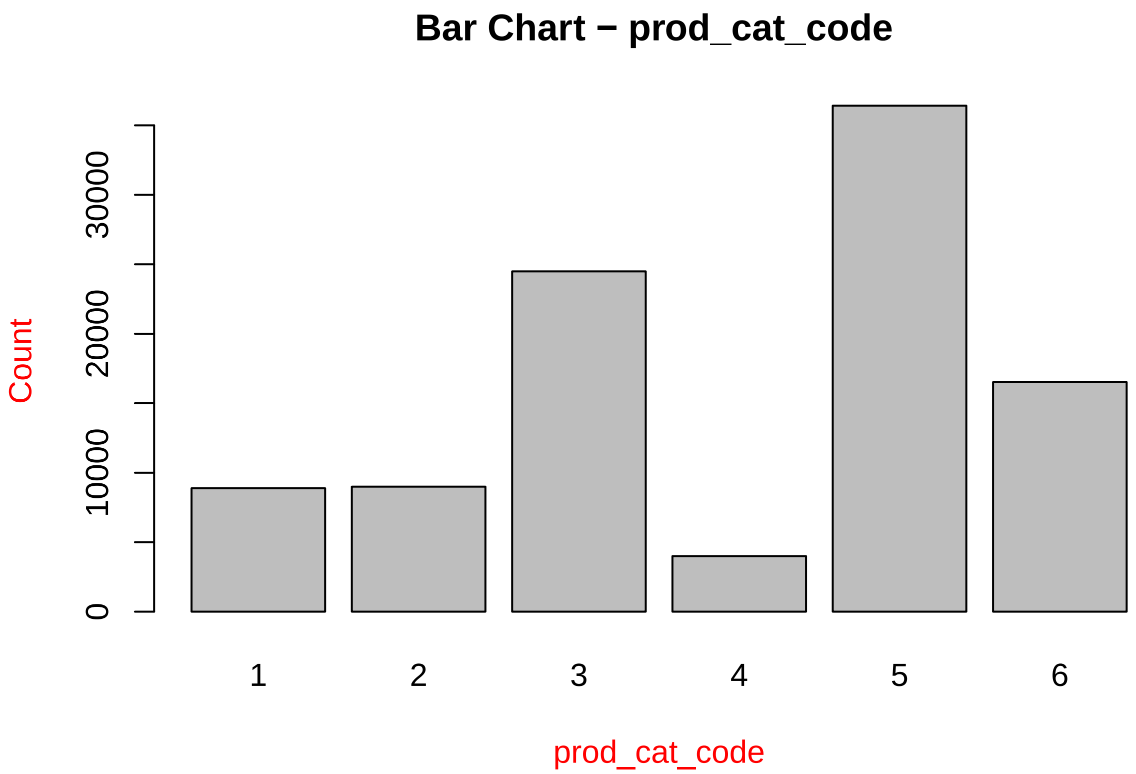


b. Bar Graph for categorical variables

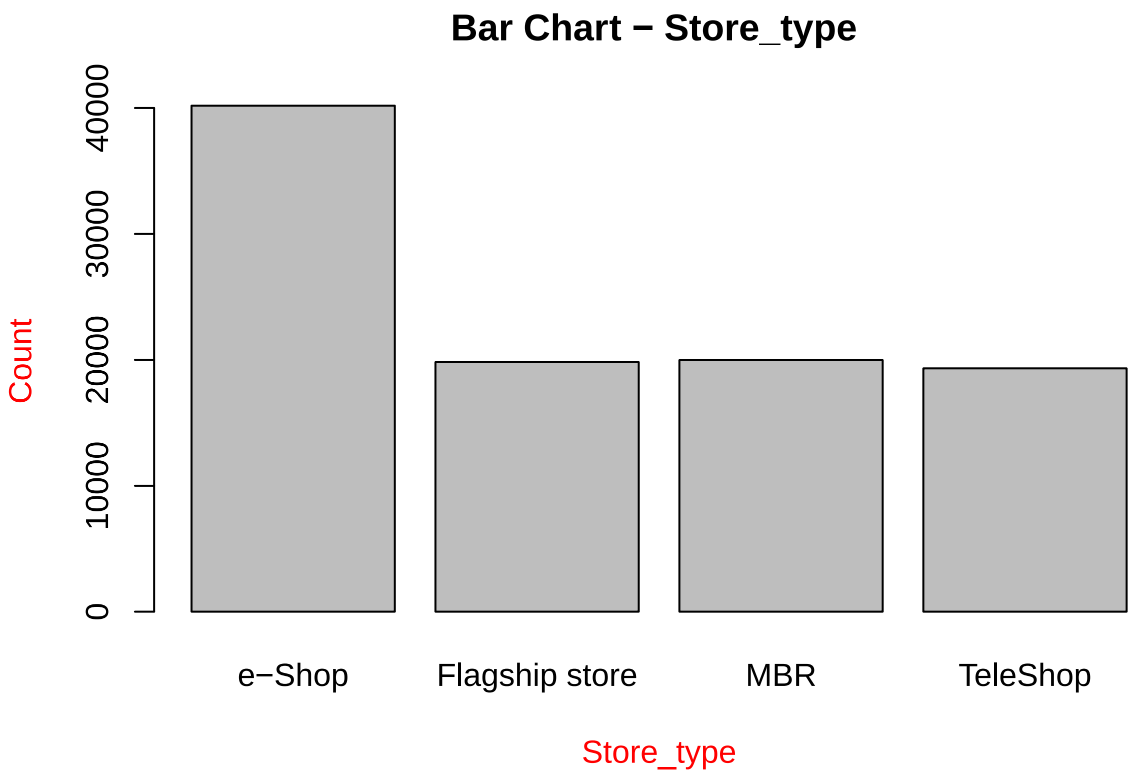
```
#Bar Graph for all the Categorical variables

freqbar1<- barplot(Count~prod_subcat_code,Freq1,xlab = "prod_subcat_code",
                   col.lab = "Red",main = "Bar Chart - prod_subcat_code ")
```



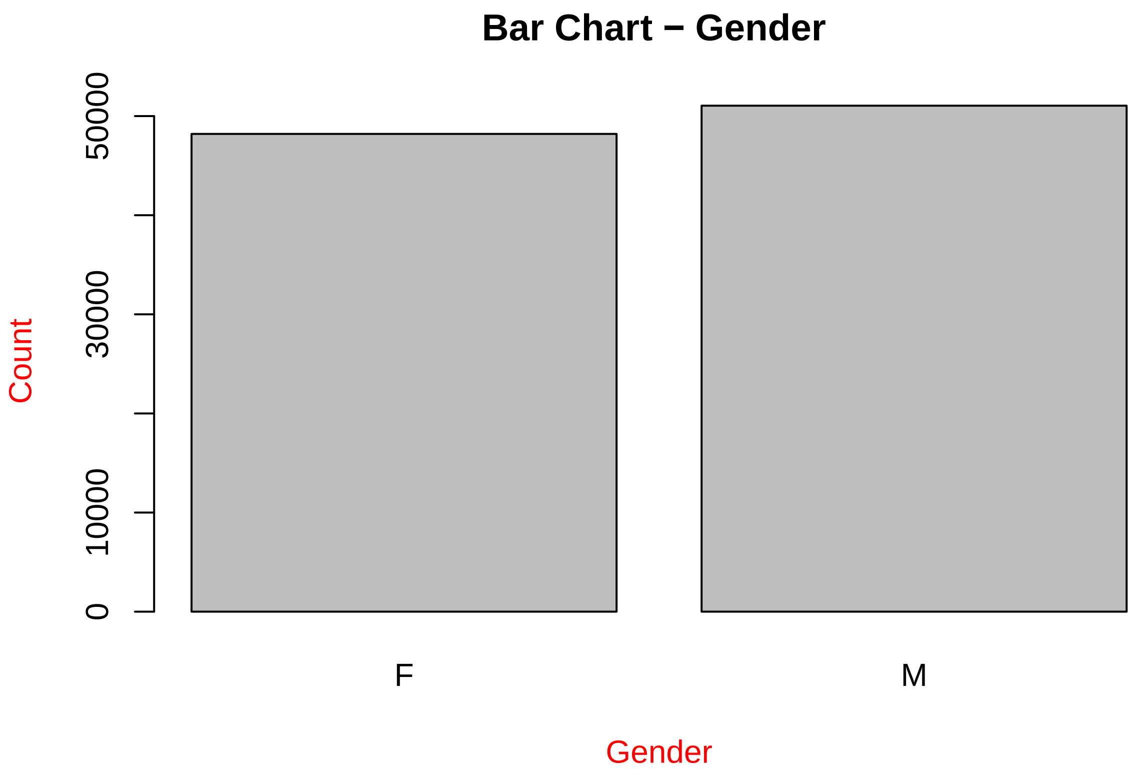
```
freqbar2 <- barplot(Count~prod_cat_code,Freq2,xlab = "prod_cat_code",
                    col.lab = "Red",main = "Bar Chart - prod_cat_code ")
```



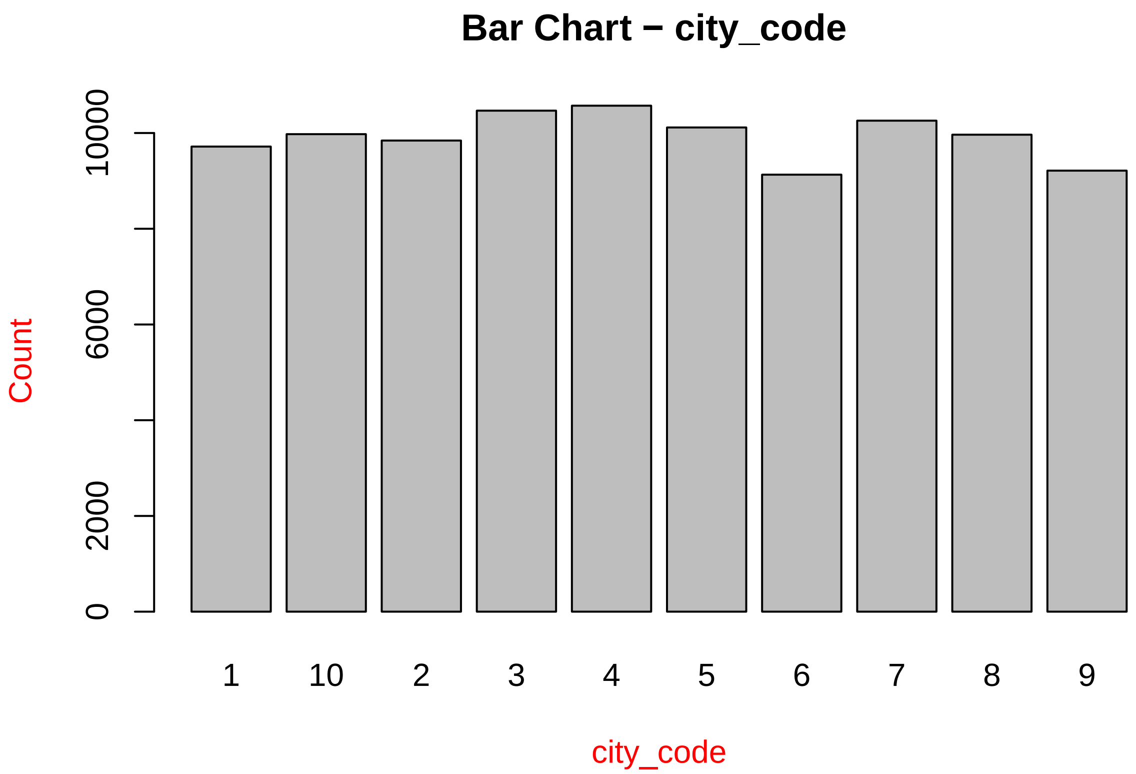
```r
freqbar3 <- barplot(Count~Store_type,Freq3,xlab = "Store_type",col.lab = "Red",
                    main = "Bar Chart - Store_type ")
```



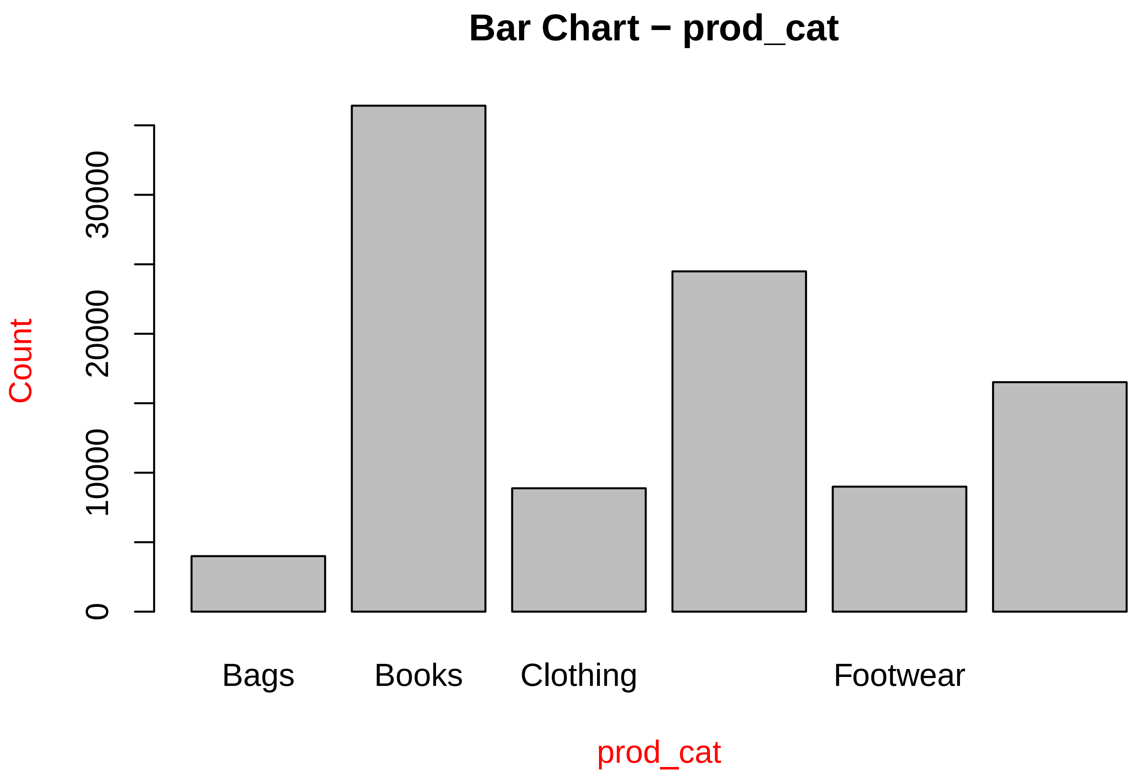
```
freqbar4 <- barplot(Count~Gender,Freq4,xlab = "Gender",col.lab = "Red",
                    main = "Bar Chart - Gender ")
```



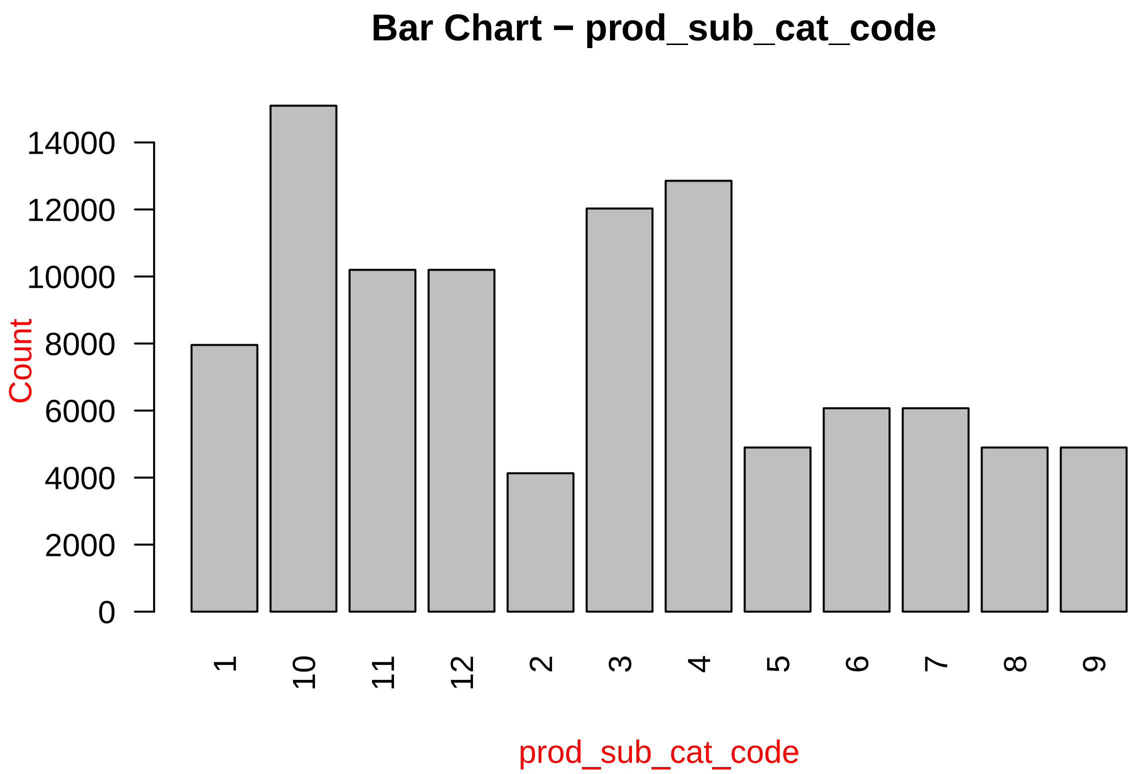
```
freqbar5 <- barplot(Count~city_code,Freq5,xlab = "city_code",col.lab = "Red",
                    main = "Bar Chart - city_code ")
```



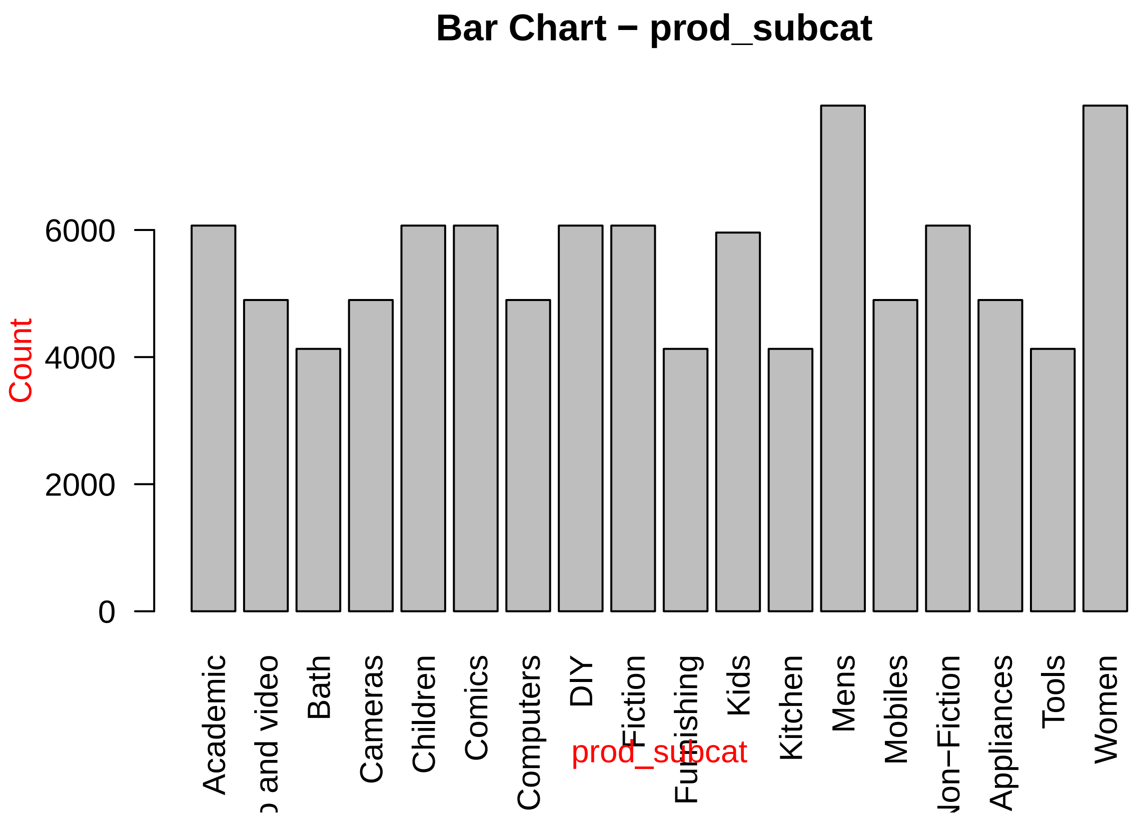
```
freqbar6 <- barplot(Count~prod_cat,Freq6,xlab = "prod_cat",col.lab = "Red",
                    main = "Bar Chart - prod_cat ")
```



```
freqbar7 <- barplot(Count~prod_sub_cat_code,Freq7,xlab = "prod_sub_cat_code",las=2,
                    col.lab = "Red",main = "Bar Chart - prod_sub_cat_code ")
```



```
freqbar8 <- barplot(Count~prod_subcat,Freq8,xlab = "prod_subcat",las=2,
                    col.lab = "Red",main = "Bar Chart - prod_subcat ")
```



Question 4. Information from data

a.Time period for the transaction data

```
#Q4 a. Calculate the following time period of the available transaction data

firstdate <- min(Customer_Final$tran_date)
lastdate <- max(Customer_Final$tran_date)
paste(as.numeric(lastdate - firstdate),"days")
```

```
## [1] "1130 days"
```

b. Number of transactions where the total amount of the transactions are negative

```
#Q4 b. Count the transactions where the total amount of the transactions
#are negative

dplyr::filter(Customer_Final,total_amt< 0)%>%summarise(Count = n())
```

```
##   Count
## 1  9294
```

Question 5. Product categorie/s that are more popular amongst F customers than male customers

```r
#Q5 Analyze which product categories are more popular among female vs male
#customers
Categories <- data.frame(dplyr::filter(Customer_Final,Gender != "")%>%
                           group_by(prod_cat,Gender)%>%
                           summarise(frequency = n()))
```

```
## ‘summarise()‘ regrouping output by ’prod_cat’ (override with ‘.groups‘ argument)
```

```r
ProdGender <- reshape2::dcast(Categories,prod_cat~Gender)
```

```
## Using frequency as value column: use value.var to override.
```

```r
ProdGender[ProdGender$F>ProdGender$M,1]
```

```
## [1] "Footwear"
```

Question 6. City Code that has the maximum customers and the percentage of customers from that city.

```r
#Q6 Which city code has the maximum customers and what was the percentage of
#customers from that city.
sub1<- data.frame(dplyr::filter(Customer_Final,city_code != "")%>%
                    group_by(city_code,cust_id)%>%
                    summarise(frequency = n()))
```

```
## ‘summarise()‘ regrouping output by ’city_code’ (override with ‘.groups‘ argument)
```

```r
sub1<-data.frame(group_by(sub1,city_code)%>%
                   summarise(noofcustomers = n()))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```r
sub1$percentage <- round((sub1$noofcustomers/sum(sub1$noofcustomers))*100,2)
sub1[sub1$noofcustomers == max(sub1$noofcustomers),]
```

```
##   city_code noofcustomers percentage
## 4         3           576      10.47
```

Question 7. Store type that sells maximum products by value and by quantity

```r
#Q7Which store type sells maximum product by value and by quantity
#By Quantity
MAxprodbyQTY <- dplyr::group_by(Customer_Final,Store_type)%>%
  summarise(TProdSold = sum(Qty))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
MAxprodbyQTY[MAxprodbyQTY$TProdSold == max(MAxprodbyQTY$TProdSold),1]
```

```
## # A tibble: 1 x 1
##   Store_type
##   <chr>
## 1 e-Shop
```

```
#By Value
MAxprodbyVALUE <- dplyr::group_by(Customer_Final,Store_type)%>%
  summarise(TProdSold = sum(total_amt))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
MAxprodbyVALUE[MAxprodbyVALUE$TProdSold == max(MAxprodbyVALUE$TProdSold),1]
```

```
## # A tibble: 1 x 1
##   Store_type
##   <chr>
## 1 e-Shop
```

Question 8. Total amount earned for Electronics and Clothing categories from Flagship stores.

```
#Q8What was the total amount earned from electronics and clothing categories
#from Flagship stores
ProdTamt <- data.frame(dplyr::filter(Customer_Final,Store_type == "Flagship store")
                       %>% group_by(prod_cat)
                       %>% summarise(Totalamtearned=sum(total_amt)))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
ProdTamt[ProdTamt$prod_cat=="Electronics"|ProdTamt$prod_cat=="Clothing",]
```

```
##      prod_cat Totalamtearned
## 3    Clothing        3583270
## 4 Electronics       11075680
```

Question 9. Total amount earned from Male customers under the Electronics category

```
#Q9 What was the total amount earned from male customers under the electronics
#category
data.frame(dplyr::filter(Customer_Final,prod_cat=="Electronics")%>%
             group_by(Gender)%>%
             summarise(Totalamtearned=sum(total_amt)))%>%
  filter(Gender == "M")
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
##   Gender Totalamtearned
## 1      M       28515547
```

Question 10.Number of customers who have more than 10 unique transactions after removing all transactions which have any negative amounts

```
#Q10 How many customers have more than 10 unique transactions after removing all transactions which have a
A <- data.frame(dplyr::filter(Customer_Final,total_amt >0)%>%
                  group_by(cust_id,transaction_id)%>%
                  summarise(count = n()))
```

```
## `summarise()` regrouping output by 'cust_id' (override with `.groups` argument)
```

```
B <- data.frame(table(A$cust_id))
length( B[B$Freq>10,1])
```

```
## [1] 6
```

Question 11. For all customers ages 25-35

a. Total amount spend for “Electronics and”Books" product categories

```
#Q11 a.For all customers ages 25-35 calculate what was the total amount spent
#for "Electronics and "Books" product categories
Customer_Final$Age <- round(as.numeric(Sys.Date() - Customer_Final$DOB )/365.25)

ProdTamt2 <- data.frame(dplyr::filter(Customer_Final, Age >= 25 & Age <= 35)
                        %>%group_by(prod_cat)%>%
                          summarise(TotalamtSpent=sum(total_amt)))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ProdTamt2[ProdTamt2$prod_cat=="Electronics"|ProdTamt$prod_cat=="Books",]
```

```
##      prod_cat TotalamtSpent
## 2       Books      25260936
## 4 Electronics      18466384
```

b. Total amount spent by these customers between 1st Jan,2014 to 1st Mar,2014

```
#Q11 b.For all customers ages 25-35 calculate what was the total amount spent by
#these customers between 1st Jan,2014 to 1st Mar,2014
data.frame(dplyr::filter(Customer_Final, Age >= 25 & Age <= 35,
                         tran_date >= "2014-01-01"& tran_date <= "2014-03-01")
           %>%summarise(TotalamtSpent=sum(total_amt)))
```

```
##   TotalamtSpent
## 1       3458965
```