Retail Case Study

#Import CSV Data Files

getwd()

## [1] "C:/Users/Pawan Sharma/Desktop/AnalytixLabs/Case Studies/R case study/Retail data analysis/R - Retail Case study/R case study 1 (Retail)/Customer Analysis for Retail"

customer<-read.csv("Customer.csv")  
View(customer)  
  
prod\_cat\_info<-read.csv("prod\_cat\_info.csv")  
View(prod\_cat\_info)  
  
Transactions<-read.csv("Transactions.csv")  
View(Transactions)

#Q(1) ## (a) Merge the data set with the help of the merge()

customer\_df<-merge(customer, Transactions,by.x="customer\_Id",by.y="cust\_id")  
View(customer\_df)  
  
Customer\_Final<-merge(customer\_df,prod\_cat\_info, by.x="prod\_subcat\_code",by.y = "prod\_sub\_cat\_code")  
View(Customer\_Final)  
  
head(Customer\_Final,3)

## prod\_subcat\_code customer\_Id DOB Gender city\_code transaction\_id  
## 1 1 275166 05-11-1970 M 10 31795262740  
## 2 1 275166 05-11-1970 M 10 31795262740  
## 3 1 275166 05-11-1970 M 10 31795262740  
## tran\_date prod\_cat\_code.x Qty Rate Tax total\_amt Store\_type  
## 1 1/2/2011 4 4 531 223.02 2347.02 e-Shop  
## 2 1/2/2011 4 4 531 223.02 2347.02 e-Shop  
## 3 1/2/2011 4 4 531 223.02 2347.02 e-Shop  
## prod\_cat\_code.y prod\_cat prod\_subcat  
## 1 2 Footwear Mens  
## 2 1 Clothing Women  
## 3 4 Bags Mens

##(b) Merge the data set with the help of dplyr package

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## 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

customer\_df<-inner\_join(customer,Transactions, by=c("customer\_Id"="cust\_id"))  
View(customer\_df)  
  
Customer\_Final<-inner\_join(customer\_df,prod\_cat\_info,by=c("prod\_subcat\_code"="prod\_sub\_cat\_code"))  
View(Customer\_Final)  
  
head(Customer\_Final,3)

## customer\_Id DOB Gender city\_code transaction\_id tran\_date  
## 1 268408 02-01-1970 M 4 87243835584 13-01-2014  
## 2 268408 02-01-1970 M 4 16197868036 16-12-2013  
## 3 268408 02-01-1970 M 4 16197868036 16-12-2013  
## prod\_subcat\_code prod\_cat\_code.x Qty Rate Tax total\_amt Store\_type  
## 1 7 5 5 187 98.175 1033.175 TeleShop  
## 2 1 4 -5 -210 110.250 -1160.250 Flagship store  
## 3 1 4 -5 -210 110.250 -1160.250 Flagship store  
## prod\_cat\_code.y prod\_cat prod\_subcat  
## 1 5 Books Fiction  
## 2 1 Clothing Women  
## 3 2 Footwear Mens

#Q(2) Prepare a summary report for the marged data set. ##(a) Get the column names and their corresponding data types.

colnames(Customer\_Final)

## [1] "customer\_Id" "DOB" "Gender" "city\_code"   
## [5] "transaction\_id" "tran\_date" "prod\_subcat\_code" "prod\_cat\_code.x"   
## [9] "Qty" "Rate" "Tax" "total\_amt"   
## [13] "Store\_type" "prod\_cat\_code.y" "prod\_cat" "prod\_subcat"

#OR  
names(Customer\_Final)

## [1] "customer\_Id" "DOB" "Gender" "city\_code"   
## [5] "transaction\_id" "tran\_date" "prod\_subcat\_code" "prod\_cat\_code.x"   
## [9] "Qty" "Rate" "Tax" "total\_amt"   
## [13] "Store\_type" "prod\_cat\_code.y" "prod\_cat" "prod\_subcat"

##(b)Top/Bottom 10 observations

head(Customer\_Final,10)

## customer\_Id DOB Gender city\_code transaction\_id tran\_date  
## 1 268408 02-01-1970 M 4 87243835584 13-01-2014  
## 2 268408 02-01-1970 M 4 16197868036 16-12-2013  
## 3 268408 02-01-1970 M 4 16197868036 16-12-2013  
## 4 268408 02-01-1970 M 4 16197868036 16-12-2013  
## 5 268408 02-01-1970 M 4 16197868036 12/12/2013  
## 6 268408 02-01-1970 M 4 16197868036 12/12/2013  
## 7 268408 02-01-1970 M 4 16197868036 12/12/2013  
## 8 268408 02-01-1970 M 4 28810141075 2/9/2013  
## 9 268408 02-01-1970 M 4 28810141075 2/9/2013  
## 10 268408 02-01-1970 M 4 28810141075 2/9/2013  
## prod\_subcat\_code prod\_cat\_code.x Qty Rate Tax total\_amt Store\_type  
## 1 7 5 5 187 98.175 1033.175 TeleShop  
## 2 1 4 -5 -210 110.250 -1160.250 Flagship store  
## 3 1 4 -5 -210 110.250 -1160.250 Flagship store  
## 4 1 4 -5 -210 110.250 -1160.250 Flagship store  
## 5 1 4 5 210 110.250 1160.250 Flagship store  
## 6 1 4 5 210 110.250 1160.250 Flagship store  
## 7 1 4 5 210 110.250 1160.250 Flagship store  
## 8 10 6 2 813 170.730 1796.730 TeleShop  
## 9 10 6 2 813 170.730 1796.730 TeleShop  
## 10 10 6 2 813 170.730 1796.730 TeleShop  
## prod\_cat\_code.y prod\_cat prod\_subcat  
## 1 5 Books Fiction  
## 2 1 Clothing Women  
## 3 2 Footwear Mens  
## 4 4 Bags Mens  
## 5 1 Clothing Women  
## 6 2 Footwear Mens  
## 7 4 Bags Mens  
## 8 3 Electronics Audio and video  
## 9 5 Books Non-Fiction  
## 10 6 Home and kitchen Kitchen

tail(Customer\_Final,10)

## customer\_Id DOB Gender city\_code transaction\_id tran\_date  
## 57157 269626 27-12-1992 F 5 94134051896 7/7/2011  
## 57158 269626 27-12-1992 F 5 94134051896 7/7/2011  
## 57159 274308 29-12-1992 F 5 26208690928 22-10-2012  
## 57160 274308 29-12-1992 F 5 26208690928 22-10-2012  
## 57161 274308 29-12-1992 F 5 26208690928 22-10-2012  
## 57162 274308 29-12-1992 F 5 57358822237 5/9/2012  
## 57163 274308 29-12-1992 F 5 57358822237 5/9/2012  
## 57164 274308 29-12-1992 F 5 57358822237 5/9/2012  
## 57165 274308 29-12-1992 F 5 58160544337 4/7/2012  
## 57166 274308 29-12-1992 F 5 58160544337 4/7/2012  
## prod\_subcat\_code prod\_cat\_code.x Qty Rate Tax total\_amt Store\_type  
## 57157 3 1 3 1078 339.57 3573.57 MBR  
## 57158 3 1 3 1078 339.57 3573.57 MBR  
## 57159 1 2 2 898 188.58 1984.58 MBR  
## 57160 1 2 2 898 188.58 1984.58 MBR  
## 57161 1 2 2 898 188.58 1984.58 MBR  
## 57162 1 2 4 876 367.92 3871.92 e-Shop  
## 57163 1 2 4 876 367.92 3871.92 e-Shop  
## 57164 1 2 4 876 367.92 3871.92 e-Shop  
## 57165 12 5 5 588 308.70 3248.70 e-Shop  
## 57166 12 5 5 588 308.70 3248.70 e-Shop  
## prod\_cat\_code.y prod\_cat prod\_subcat  
## 57157 2 Footwear Women  
## 57158 5 Books Comics  
## 57159 1 Clothing Women  
## 57160 2 Footwear Mens  
## 57161 4 Bags Mens  
## 57162 1 Clothing Women  
## 57163 2 Footwear Mens  
## 57164 4 Bags Mens  
## 57165 5 Books Academic  
## 57166 6 Home and kitchen Tools

##(c) “Five-Number summary”, for continuous variables.

Customer\_Final %>% select(Qty, Rate,Tax, total\_amt)%>%summary()

## Qty Rate Tax total\_amt   
## Min. :-5.00 Min. :-1499 Min. : 7.35 Min. :-8270.9   
## 1st Qu.: 1.00 1st Qu.: 312 1st Qu.: 98.49 1st Qu.: 764.7   
## Median : 3.00 Median : 709 Median :199.34 Median : 1759.2   
## Mean : 2.43 Mean : 635 Mean :248.97 Mean : 2102.1   
## 3rd Qu.: 4.00 3rd Qu.: 1109 3rd Qu.:364.98 3rd Qu.: 3553.7   
## Max. : 5.00 Max. : 1500 Max. :787.50 Max. : 8287.5

#OR  
summary(Customer\_Final$Qty)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -5.00 1.00 3.00 2.43 4.00 5.00

summary(Customer\_Final$Rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1499 312 709 635 1109 1500

summary(Customer\_Final$Tax)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.35 98.49 199.34 248.97 364.98 787.50

summary(Customer\_Final$total\_amt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -8270.9 764.7 1759.2 2102.1 3553.7 8287.5

##(d) Frequency tables for all the categorical variables.

data.frame(table(Customer\_Final$Gender))

## Var1 Freq  
## 1 F 27953  
## 2 M 29213

data.frame(table(Customer\_Final$city\_code))

## Var1 Freq  
## 1 1 5582  
## 2 2 5616  
## 3 3 5886  
## 4 4 5961  
## 5 5 5941  
## 6 6 5295  
## 7 7 5761  
## 8 8 5771  
## 9 9 5538  
## 10 10 5800

data.frame(table(Customer\_Final$prod\_subcat\_code))

## Var1 Freq  
## 1 1 8850  
## 2 2 1007  
## 3 3 9201  
## 4 4 16008  
## 5 5 958  
## 6 6 989  
## 7 7 1043  
## 8 8 972  
## 9 9 985  
## 10 10 8979  
## 11 11 4116  
## 12 12 4058

data.frame(table(Customer\_Final$prod\_cat\_code.x))

## Var1 Freq  
## 1 1 9847  
## 2 2 10006  
## 3 3 9895  
## 4 4 6989  
## 5 5 12141  
## 6 6 8288

data.frame(table(Customer\_Final$Store\_type))

## Var1 Freq  
## 1 e-Shop 23087  
## 2 Flagship store 11324  
## 3 MBR 11644  
## 4 TeleShop 11111

data.frame(table(Customer\_Final$prod\_cat\_code.y))

## Var1 Freq  
## 1 1 10019  
## 2 2 10019  
## 3 3 9910  
## 4 4 6952  
## 5 5 12179  
## 6 6 8087

data.frame(table(Customer\_Final$prod\_cat))

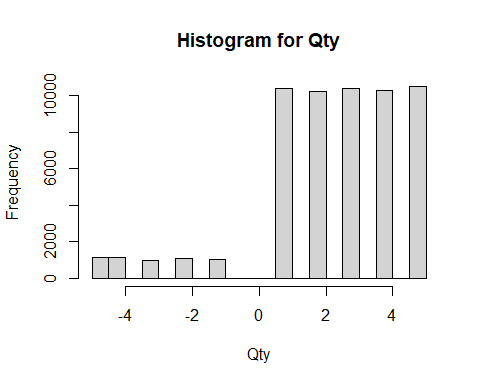
## Var1 Freq  
## 1 Bags 6952  
## 2 Books 12179  
## 3 Clothing 10019  
## 4 Electronics 9910  
## 5 Footwear 10019  
## 6 Home and kitchen 8087

data.frame(table(Customer\_Final$prod\_subcat))

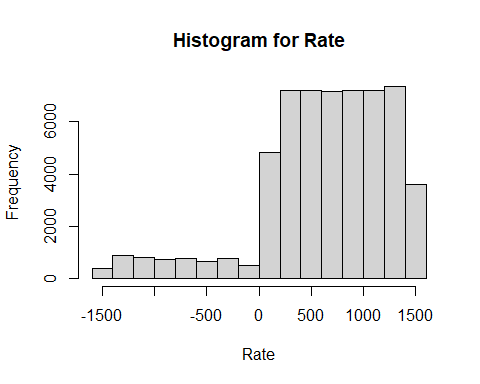
## Var1 Freq  
## 1 Academic 2029  
## 2 Audio and video 2993  
## 3 Bath 2058  
## 4 Cameras 985  
## 5 Children 2058  
## 6 Comics 3067  
## 7 Computers 958  
## 8 DIY 989  
## 9 Fiction 1043  
## 10 Furnishing 1007  
## 11 Kids 7069  
## 12 Kitchen 2993  
## 13 Mens 9902  
## 14 Mobiles 4002  
## 15 Non-Fiction 2993  
## 16 Personal Appliances 972  
## 17 Tools 2029  
## 18 Women 10019

#Q(3) Generate Histogram for all the continuous Variables and frequency bars for categorical variables.

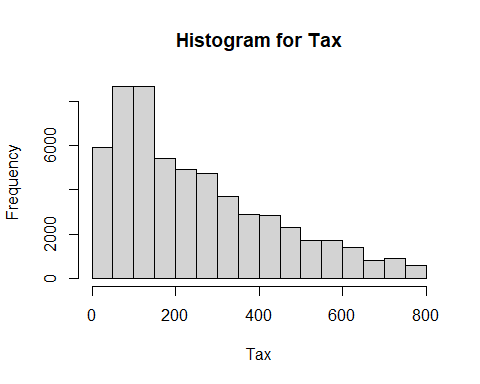
hist(Customer\_Final$Qty, xlab = "Qty", ylab = "Frequency", main = "Histogram for Qty")



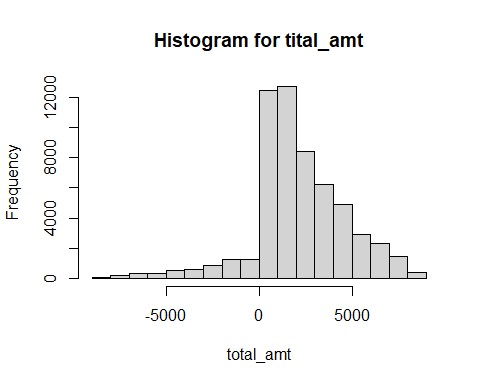
hist(Customer\_Final$Rate,xlab = "Rate", ylab = "Frequency", main = "Histogram for Rate")



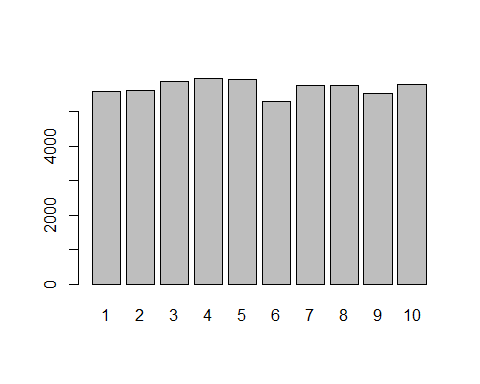
hist(Customer\_Final$Tax,xlab = "Tax", ylab = "Frequency", main = "Histogram for Tax")



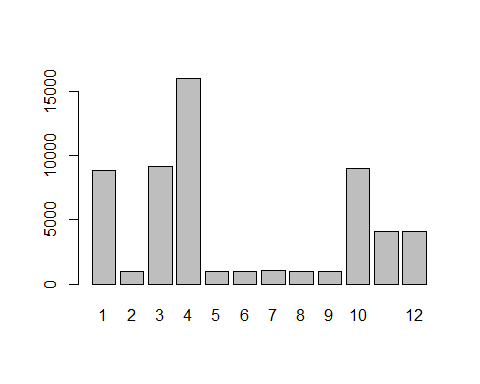
hist(Customer\_Final$total\_amt,xlab = "total\_amt", ylab = "Frequency", main = "Histogram for tital\_amt")



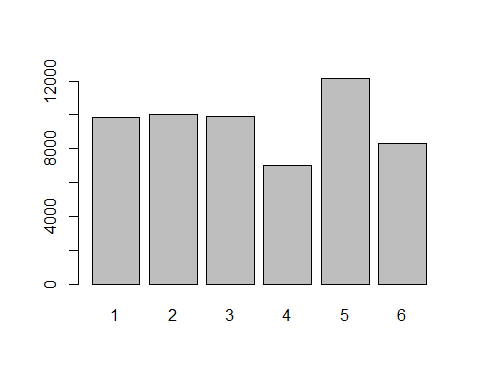
#barplot(data.frame(table(Customer\_Final$Gender)))  
barplot(table(Customer\_Final$city\_code))



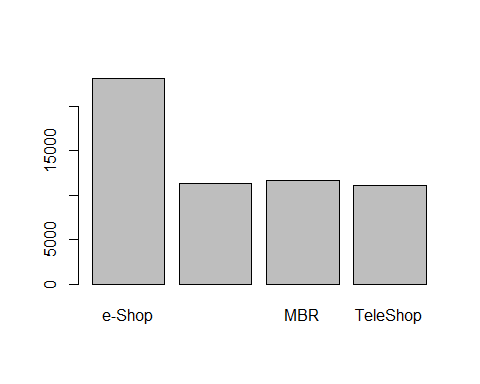
barplot(table(Customer\_Final$prod\_subcat\_code))



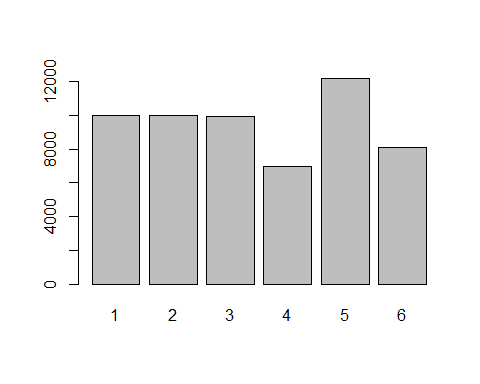
barplot(table(Customer\_Final$prod\_cat\_code.x))



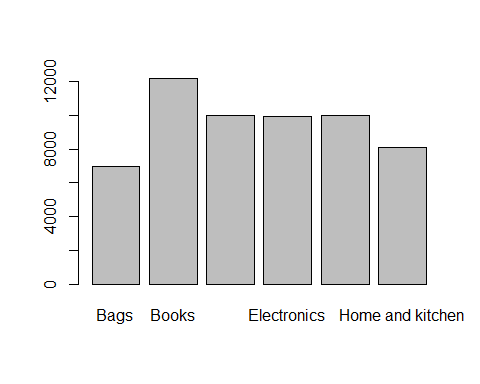
barplot(table(Customer\_Final$Store\_type))



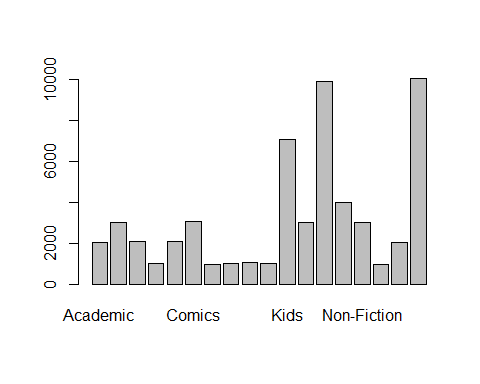
barplot(table(Customer\_Final$prod\_cat\_code.y))



barplot(table(Customer\_Final$prod\_cat))



barplot(table(Customer\_Final$prod\_subcat))



#Q(4) Calculate the following information using the merged dataset

library(lubridate)

## Warning: package 'lubridate' was built under R version 4.0.5

##   
## Attaching package: 'lubridate'

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

df<-distinct(select(Customer\_Final, transaction\_id,tran\_date))  
dmy=dmy(df$tran\_date)  
head(dmy,20)

## [1] "2014-01-13" "2013-12-16" "2013-12-12" "2013-09-02" "2013-05-06"  
## [6] "2013-01-30" "2012-10-14" "2012-09-07" "2012-08-14" "2012-04-26"  
## [11] "2011-07-12" "2012-04-15" "2012-04-08" "2011-09-18" "2013-10-03"  
## [16] "2013-03-31" "2013-01-30" "2012-12-24" "2012-07-20" "2012-02-06"

mdy=mdy(df$tran\_date)

## Warning: 13918 failed to parse.

head(mdy,20)

## [1] NA NA "2013-12-12" "2013-02-09" "2013-06-05"  
## [6] NA NA "2012-07-09" NA NA   
## [11] "2011-12-07" NA "2012-08-04" NA "2013-03-10"  
## [16] NA NA NA NA "2012-06-02"

mdy[is.na(mdy)]<-dmy[is.na(mdy)]  
df$tran\_date<-mdy  
head(df,5)

## transaction\_id tran\_date  
## 1 87243835584 2014-01-13  
## 2 16197868036 2013-12-16  
## 3 16197868036 2013-12-12  
## 4 28810141075 2013-02-09  
## 5 64633435931 2013-06-05

#Time Period  
years<-year(max(df$tran\_date))-year(min(df$tran\_date))  
years

## [1] 3

months<-month(max(df$tran\_date))-month(min(df$tran\_date))  
months

## [1] 11

day<-day(max(df$tran\_date))-month(min(df$tran\_date))  
day

## [1] 1

##(b) Count the transactions where the total amount of transaction was negative.

Customer\_Final%>%select(transaction\_id,total\_amt)%>%filter(total\_amt<0)%>%count(transaction\_id)%>%head()

## transaction\_id n  
## 1 87125650 3  
## 2 95570369 2  
## 3 156697409 2  
## 4 265535637 3  
## 5 268218855 4  
## 6 277893283 3

#Q(5) Analyze which product categories are more popular among females vs male customers.

Customer\_Final%>%select(prod\_cat, Gender)%>%group\_by(prod\_cat)%>%count(Gender)%>%arrange(desc(n))

## # A tibble: 12 x 3  
## # Groups: prod\_cat [6]  
## prod\_cat Gender n  
## <chr> <chr> <int>  
## 1 Books M 6175  
## 2 Books F 6004  
## 3 Electronics M 5115  
## 4 Clothing M 5103  
## 5 Footwear M 5103  
## 6 Clothing F 4916  
## 7 Footwear F 4916  
## 8 Electronics F 4795  
## 9 Home and kitchen M 4128  
## 10 Home and kitchen F 3959  
## 11 Bags M 3589  
## 12 Bags F 3363

prod\_cat “Books” is more popular among females and male.

#Q(6) Which City code has the maximum customers and what was the percentage of customers from the city?

max\_cust<-Customer\_Final%>%select(customer\_Id,city\_code)%>%distinct()%>%count(city\_code)  
max\_cust

## city\_code n  
## 1 1 535  
## 2 2 546  
## 3 3 576  
## 4 4 569  
## 5 5 570  
## 6 6 516  
## 7 7 563  
## 8 8 551  
## 9 9 532  
## 10 10 546  
## 11 NA 2

#Identify NA's.  
which(is.na(Customer\_Final$city\_code))

## [1] 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155 1547 1548 1549 1550

#Replace NA's with city code 3.  
Customer\_Final$city\_code[is.na(Customer\_Final$city\_code)]<-3  
  
#City code wise Maximum customers.  
max\_cust<-Customer\_Final%>%select(customer\_Id,city\_code)%>%distinct()%>%count(city\_code)%>%arrange(desc(n))  
max\_cust

## city\_code n  
## 1 3 578  
## 2 5 570  
## 3 4 569  
## 4 7 563  
## 5 8 551  
## 6 2 546  
## 7 10 546  
## 8 1 535  
## 9 9 532  
## 10 6 516

City code 3 has maximum customers.

#Percentage of the customers of each city.  
max\_cust$Percentage<-round((max\_cust$n/sum(max\_cust$n))\*100,2)  
max\_cust

## city\_code n Percentage  
## 1 3 578 10.50  
## 2 5 570 10.35  
## 3 4 569 10.33  
## 4 7 563 10.23  
## 5 8 551 10.01  
## 6 2 546 9.92  
## 7 10 546 9.92  
## 8 1 535 9.72  
## 9 9 532 9.66  
## 10 6 516 9.37

#Q(7)Which store type sells the maximum products by value and by quantity?

Customer\_Final%>% select(Store\_type,total\_amt,Qty)%>%group\_by(Store\_type)%>%summarise(Grand\_sales=sum(total\_amt),Grand\_Qty=sum(Qty))%>%arrange(desc(Grand\_Qty))

## # A tibble: 4 x 3  
## Store\_type Grand\_sales Grand\_Qty  
## <chr> <dbl> <int>  
## 1 e-Shop 49096589. 56372  
## 2 MBR 24018167. 27869  
## 3 Flagship store 24072659. 27600  
## 4 TeleShop 22980349. 27059

Store type “e-shope” sells the maximum products by value and by quantity.

#Q(8) What was the total amount earned from the “Electronics” and “Clothing” categories from Flagship stores?

Customer\_Final%>%select(Store\_type,prod\_cat, total\_amt)%>%filter(Store\_type=="Flagship store",prod\_cat==c("Electronics","Clothing"))%>%group\_by(prod\_cat)%>%summarise(Grand\_Total=sum(total\_amt))

## # A tibble: 2 x 2  
## prod\_cat Grand\_Total  
## <chr> <dbl>  
## 1 Clothing 2149593.  
## 2 Electronics 2011226.

#Q(9) What was the total amount earned from “Male” customers under the “Electronic” category?

Customer\_Final %>% select(prod\_cat,Gender,total\_amt)%>%filter(Gender=="M",prod\_cat=="Electronics")%>%summarise(prod\_cat="Electronics",Gender="M",Grand\_Total=sum(total\_amt))

## prod\_cat Gender Grand\_Total  
## 1 Electronics M 10953414

#Q(10) How many customers have more than 10 unique transactions, after removing all the transaction which have any negative amounts?

Customer\_Final%>% select(transaction\_id,customer\_Id,total\_amt)%>%filter(total\_amt>0)%>%distinct(transaction\_id, customer\_Id)%>%group\_by(customer\_Id)%>%count(customer\_Id)%>%arrange(desc(n))%>%filter(n > 10)

## # A tibble: 6 x 2  
## # Groups: customer\_Id [6]  
## customer\_Id n  
## <int> <int>  
## 1 266794 11  
## 2 270535 11  
## 3 270803 11  
## 4 272741 11  
## 5 273014 11  
## 6 274227 11

There are six customers having more than 10 unique transactions, after removing all the transaction which have any negative amounts.

#Q(11) For all customers aged between 25 - 35, find out: ##(a) What was the total amount spent for “Electronics” and “Books” product categories?

#Changing DOB Date format %Y-%m-%d as the R standard.  
Customer\_Final$DOB<-strptime(as.character(Customer\_Final$DOB),"%d-%m-%Y")  
Customer\_Final$DOB<-format(Customer\_Final$DOB,"%Y-%m-%d")  
  
#Adding Age column.  
Customer\_Final$Age<-year(Sys.Date())-year(Customer\_Final$DOB)  
  
Customer\_Final%>%select(prod\_cat,Age,total\_amt)%>%filter(prod\_cat==c("Electronics","Books"), Age==25:35)%>%group\_by(prod\_cat)%>%summarise(sum(total\_amt))

## Warning in Age == 25:35: longer object length is not a multiple of shorter  
## object length

## # A tibble: 2 x 2  
## prod\_cat `sum(total\_amt)`  
## <chr> <dbl>  
## 1 Books 363547.  
## 2 Electronics 295582.

#Q(b) What was the total amount spent by these customers between 1st Jan,2014 to 1st Mar, 2014?

#Converting the date format of tran\_date into the same format.  
library(lubridate)  
dmy=dmy(Customer\_Final$tran\_date)  
head(dmy,20)

## [1] "2014-01-13" "2013-12-16" "2013-12-16" "2013-12-16" "2013-12-12"  
## [6] "2013-12-12" "2013-12-12" "2013-09-02" "2013-09-02" "2013-09-02"  
## [11] "2013-05-06" "2013-05-06" "2013-05-06" "2013-05-06" "2013-01-30"  
## [16] "2012-10-14" "2012-10-14" "2012-10-14" "2012-09-07" "2012-09-07"

mdy=mdy(Customer\_Final$tran\_date)

## Warning: 34590 failed to parse.

head(mdy,20)

## [1] NA NA NA NA "2013-12-12"  
## [6] "2013-12-12" "2013-12-12" "2013-02-09" "2013-02-09" "2013-02-09"  
## [11] "2013-06-05" "2013-06-05" "2013-06-05" "2013-06-05" NA   
## [16] NA NA NA "2012-07-09" "2012-07-09"

mdy[is.na(mdy)]<-dmy[is.na(mdy)]  
Customer\_Final$tran\_date<-mdy  
  
#Calculating total amount.   
Customer\_Final%>%select(Age,tran\_date,total\_amt)%>%filter(Age==25:35,tran\_date>="2014-01-01" & tran\_date <= "2014-03-01" )%>%arrange(tran\_date) %>%summarise(sum(total\_amt))

## Warning in Age == 25:35: longer object length is not a multiple of shorter  
## object length

## sum(total\_amt)  
## 1 96515.12