#Report of Case study 1

Importing Data and Packages

```
d1.df = read.csv("C:\\Users\\Kanishka\\Documents\\BA360\\R case study 1 (Retail)\\Customer.csv", header
d2.df = read.csv("C:\\Users\\Kanishka\\Documents\\BA360\\R case study 1 (Retail)\\Transactions.csv", hea
d3.df = read.csv("C:\\Users\\Kanishka\\Documents\\BA360\\R case study 1 (Retail)\\prod_cat_info.csv", he

library(dplyr)
```

```
##
## 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
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
library(ggplot2)
```

Question 1 and Data Preparation

```
colnames(d1.df) [colnames(d1.df)=="customer_Id"] <- "cust_id"
colnames(d3.df) [colnames(d3.df)=="prod_sub_cat_code"] <- "prod_subcat_code"

f1.df <- merge(x = d2.df,y = d3.df,by=c("prod_cat_code","prod_subcat_code"),all.x = TRUE)
Customer_final.df <- merge(x=f1.df,y=d1.df,by="cust_id",all.x = TRUE) #using merge

f2.df <- left_join(d2.df,d3.df,by=c("prod_cat_code","prod_subcat_code"))
Customer_final_1.df <- left_join(f2.df,d1.df,by="cust_id") #using left join

## Data Preparation

df = subset(Customer_final_1.df, select = -c(prod_subcat_code) )
df$tran_date <- dmy(df$tran_date)
df$DOB <- dmy(df$DOB)

##For Question 4
max_date <- max(df$tran_date)
```

```r
min_date <- min(df$tran_date)

##For Question 11
mydate1 <- as.Date("2014-01-01")
mydate2 <- as.Date("2014-03-01")

df$Age = round(as.numeric(difftime(max_date, df$DOB, units = "weeks"))/52.25) #calculating Age

df$cust_id <- as.factor(df$cust_id)
df$transaction_id <- as.factor(df$transaction_id)
df$city_code <- as.factor(df$city_code)
df$prod_subcat <- as.factor(df$prod_subcat)
df$Gender <- as.factor(df$Gender)
df$prod_cat <- as.factor(df$prod_cat)
df$Store_type <- as.factor(df$Store_type)
```

Question 2

```r
#Q2.a Datatypes
str(df)
```

```
## 'data.frame':    23053 obs. of  15 variables:
##  $ transaction_id: Factor w/ 20878 levels "3268991","7073244",..: 16784 6148 10773 19468 10773 20330
##  $ cust_id       : Factor w/ 5506 levels "266783","266784",..: 2309 2328 4295 3063 4295 3608 4457 305
##  $ tran_date     : Date, format: "2014-02-28" "2014-02-27" ...
##  $ prod_cat_code : int  1 3 5 6 5 3 6 6 1 3 ...
##  $ Qty           : int  -5 -5 -2 -3 -2 -2 -1 -1 -3 -4 ...
##  $ Rate          : int  -772 -1497 -791 -1363 -791 -824 -1450 -1225 -908 -581 ...
##  $ Tax           : num  405 786 166 429 166 ...
##  $ total_amt     : num  -4265 -8271 -1748 -4518 -1748 ...
##  $ Store_type    : Factor w/ 4 levels "e-Shop","Flagship store",..: 1 1 4 1 4 4 1 4 3 1 ...
##  $ prod_cat      : Factor w/ 6 levels "Bags","Books",..: 3 4 2 6 2 4 6 6 3 4 ...
##  $ prod_subcat   : Factor w/ 18 levels "Academic","Audio and video",..: 18 7 8 3 8 16 3 17 11 16 ...
##  $ DOB           : Date, format: "1981-09-26" "1973-05-11" ...
##  $ Gender        : Factor w/ 3 levels "","F","M": 3 2 3 3 3 2 3 3 2 2 ...
##  $ city_code     : Factor w/ 10 levels "1","2","3","4",..: 5 8 8 3 8 6 9 9 8 3 ...
##  $ Age           : num  32 41 22 33 22 31 33 43 42 34 ...
```

```r
#Q2.b Top 10 rows
df[1:10,]
```

```
##    transaction_id cust_id  tran_date prod_cat_code Qty  Rate     Tax total_amt
## 1     80712190438  270351 2014-02-28             1  -5  -772 405.300 -4265.300
## 2     29258453508  270384 2014-02-27             3  -5 -1497 785.925 -8270.925
## 3     51750724947  273420 2014-02-24             5  -2  -791 166.110 -1748.110
## 4     93274880719  271509 2014-02-24             6  -3 -1363 429.345 -4518.345
## 5     51750724947  273420 2014-02-23             5  -2  -791 166.110 -1748.110
## 6     97439039119  272357 2014-02-23             3  -2  -824 173.040 -1821.040
## 7     45649838090  273667 2014-02-22             6  -1 -1450 152.250 -1602.250
## 8     22643667930  271489 2014-02-22             6  -1 -1225 128.625 -1353.625
## 9     79792372943  275108 2014-02-22             1  -3  -908 286.020 -3010.020
## 10    50076728598  269014 2014-02-21             3  -4  -581 244.020 -2568.020
```

```
##    Store_type         prod_cat         prod_subcat        DOB Gender city_code
## 1      e-Shop         Clothing               Women 1981-09-26      M         5
## 2      e-Shop      Electronics           Computers 1973-05-11      F         8
## 3    TeleShop            Books                 DIY 1992-07-27      M         8
## 4      e-Shop Home and kitchen                Bath 1981-06-08      M         3
## 5    TeleShop            Books                 DIY 1992-07-27      M         8
## 6    TeleShop      Electronics Personal Appliances 1982-10-09      F         6
## 7      e-Shop Home and kitchen                Bath 1981-05-29      M         9
## 8    TeleShop Home and kitchen               Tools 1971-04-21      M         9
## 9         MBR         Clothing                Kids 1971-11-04      F         8
## 10     e-Shop      Electronics Personal Appliances 1979-11-27      F         3
##    Age
## 1   32
## 2   41
## 3   22
## 4   33
## 5   22
## 6   31
## 7   33
## 8   43
## 9   42
## 10  34
```

```
#Q2.c Summary Of variables Total amount and Quantity
summarise(df, Median_qty = median(Qty, na.rm = T))
```

```
##   Median_qty
## 1          3
```

```
summarise(df, Median_amt = median(total_amt, na.rm = F))
```

```
##   Median_amt
## 1    1754.74
```

```
summarise(df, Min_amt = min(total_amt, na.rm = T))
```

```
##     Min_amt
## 1 -8270.925
```

```
summarise(df, Min_qty = min(Qty, na.rm = T))
```

```
##   Min_qty
## 1      -5
```

```
summarise(df, Max_amt = max(total_amt, na.rm = T))
```

```
##   Max_amt
## 1  8287.5
```

```r
summarise(df, Max_qty = max(Qty, na.rm = T))
```

```
##   Max_qty
## 1       5
```

```r
Q2c <- select(df,Qty,total_amt)
quantile(Q2c$Qty)
```

```
##   0%  25%  50%  75% 100%
##   -5    1    3    4    5
```

```r
quantile(Q2c$total_amt)
```

```
##        0%       25%       50%       75%      100%
## -8270.925   762.450  1754.740  3569.150  8287.500
```

```r
#Q2.d Freq Table
table(df$city_code)
```

```
##
##    1    2    3    4    5    6    7    8    9   10
## 2258 2270 2411 2422 2360 2127 2356 2330 2178 2333
```

```r
table(df$Gender)
```

```
##
##           F     M
##     9 11233 11811
```

```r
table(df$prod_cat)
```

```
##
##             Bags            Books         Clothing      Electronics
##             1998             6069             2960             4898
##         Footwear Home and kitchen
##             2999             4129
```

```r
table(df$prod_subcat)
```

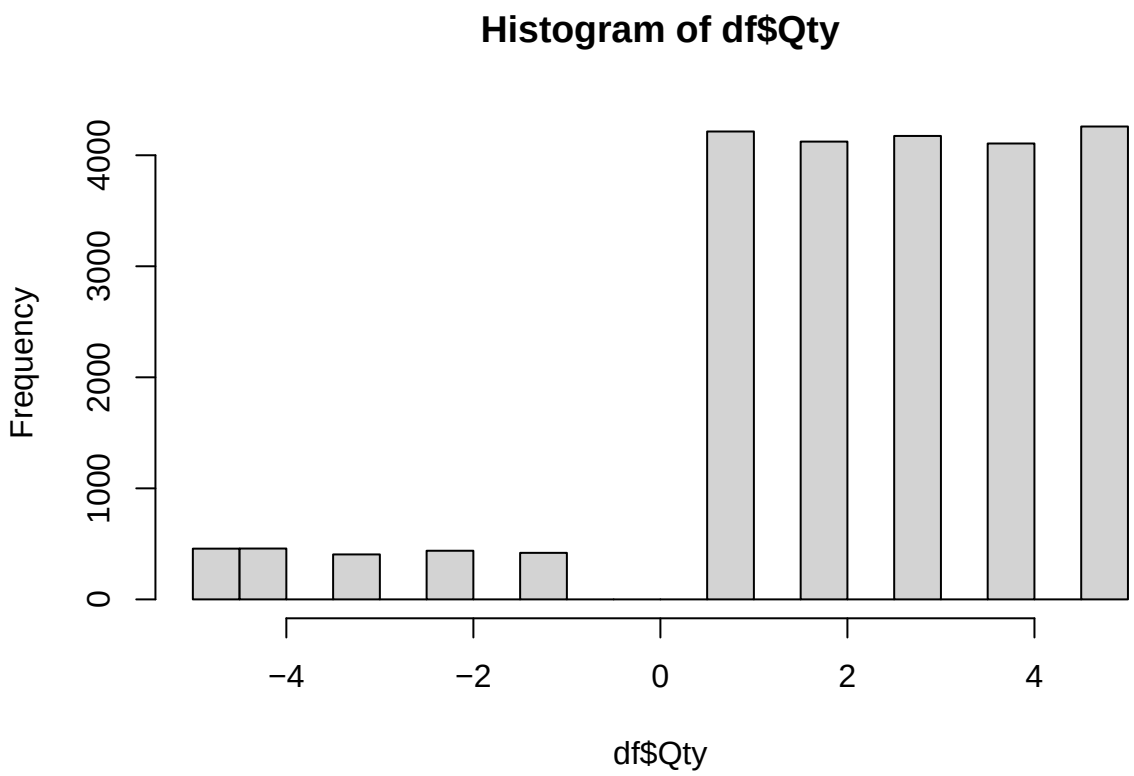
```
##
##            Academic     Audio and video                Bath             Cameras
##                 967                 952                1023                 985
##            Children              Comics           Computers                 DIY
##                1035                1031                 958                 989
##             Fiction          Furnishing                Kids             Kitchen
##                1043                1007                1997                1037
##                Mens             Mobiles         Non-Fiction Personal Appliances
##                2912                1031                1004                 972
##               Tools               Women
##                1062                3048
```

```
table(df$Store_type)
```

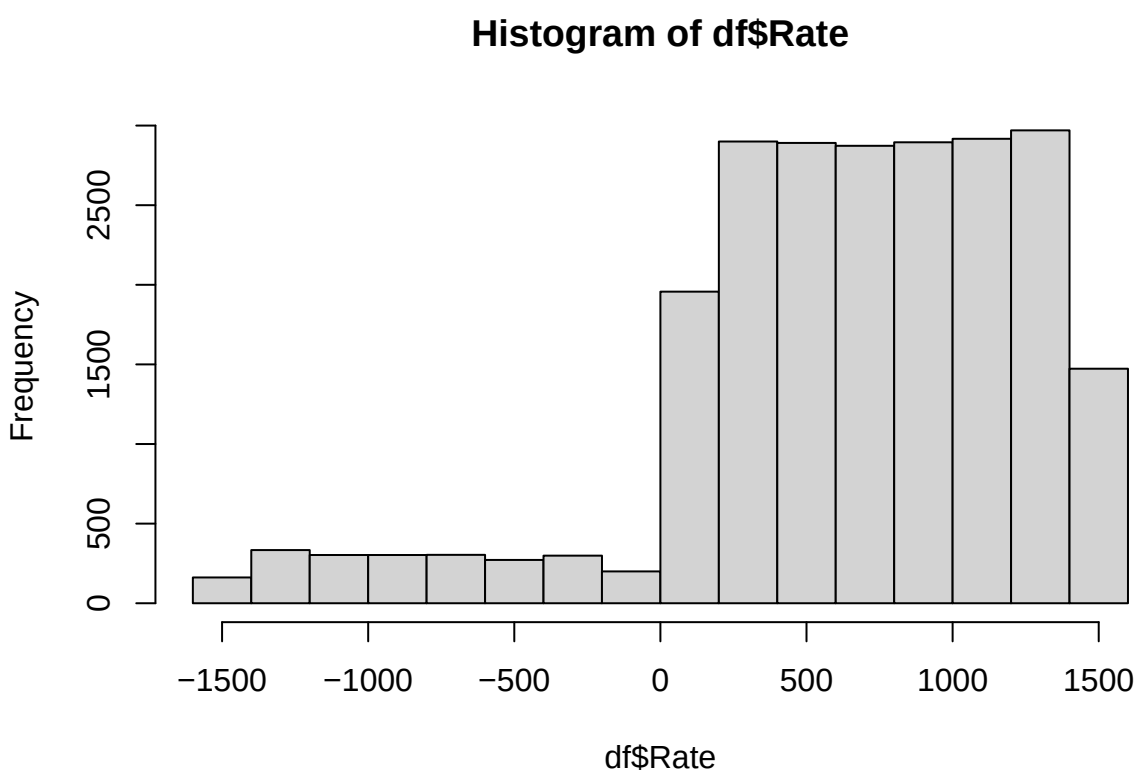
```
##
##         e-Shop Flagship store            MBR       TeleShop
##            9311            4577            4661            4504
```

Question 3

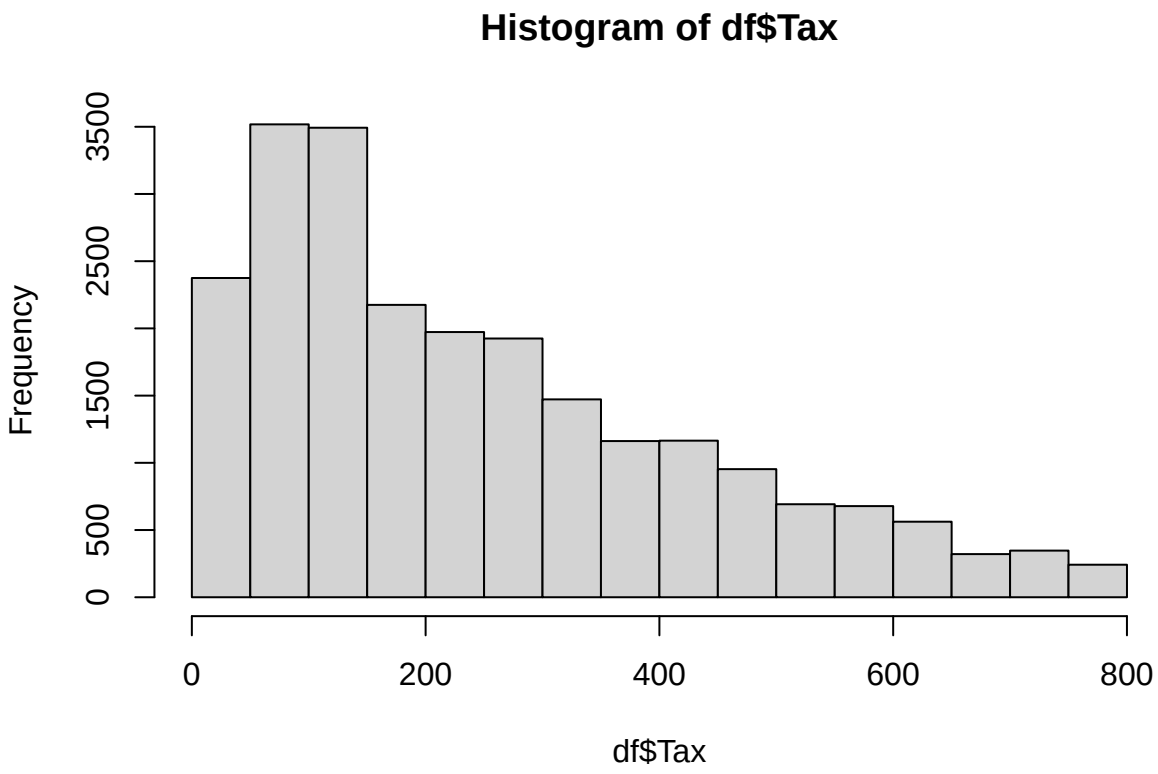
```
#Q3.
hist(df$Qty)
```



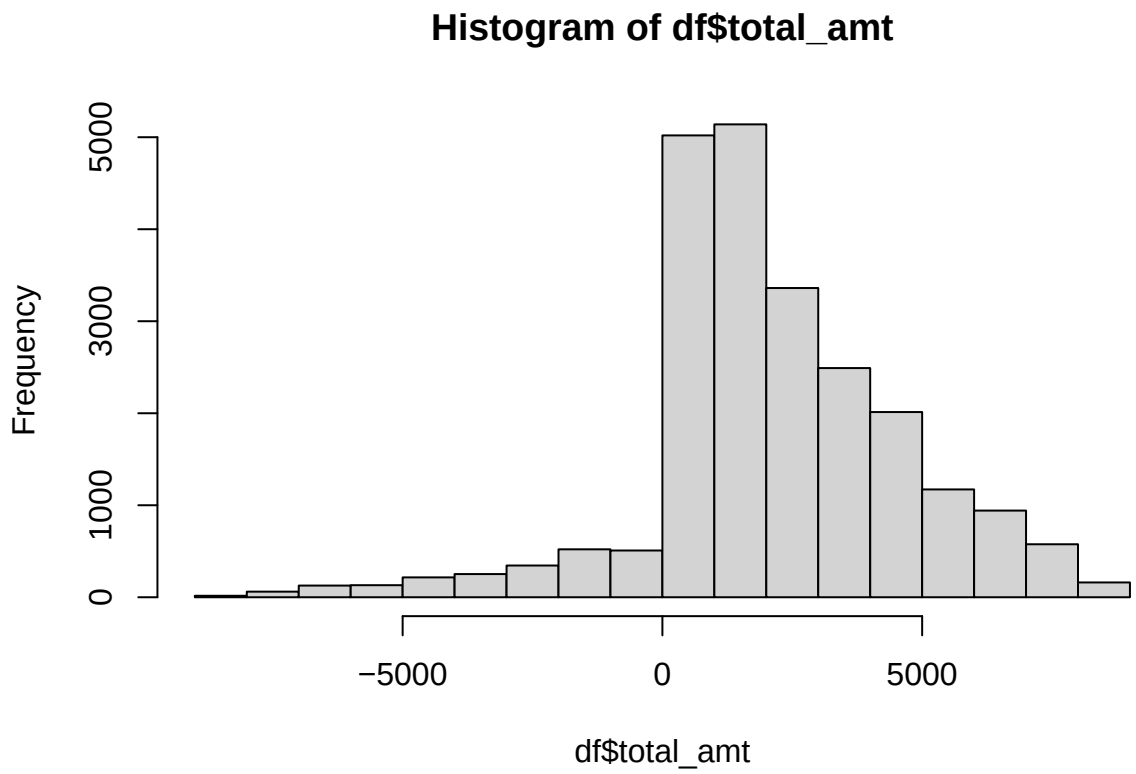

```
hist(df$Rate)
```


Histogram of df$Rate
Frequency
0
500
1500
2500
−1500
−1000
−500
0
500
1000
1500
df$Rate

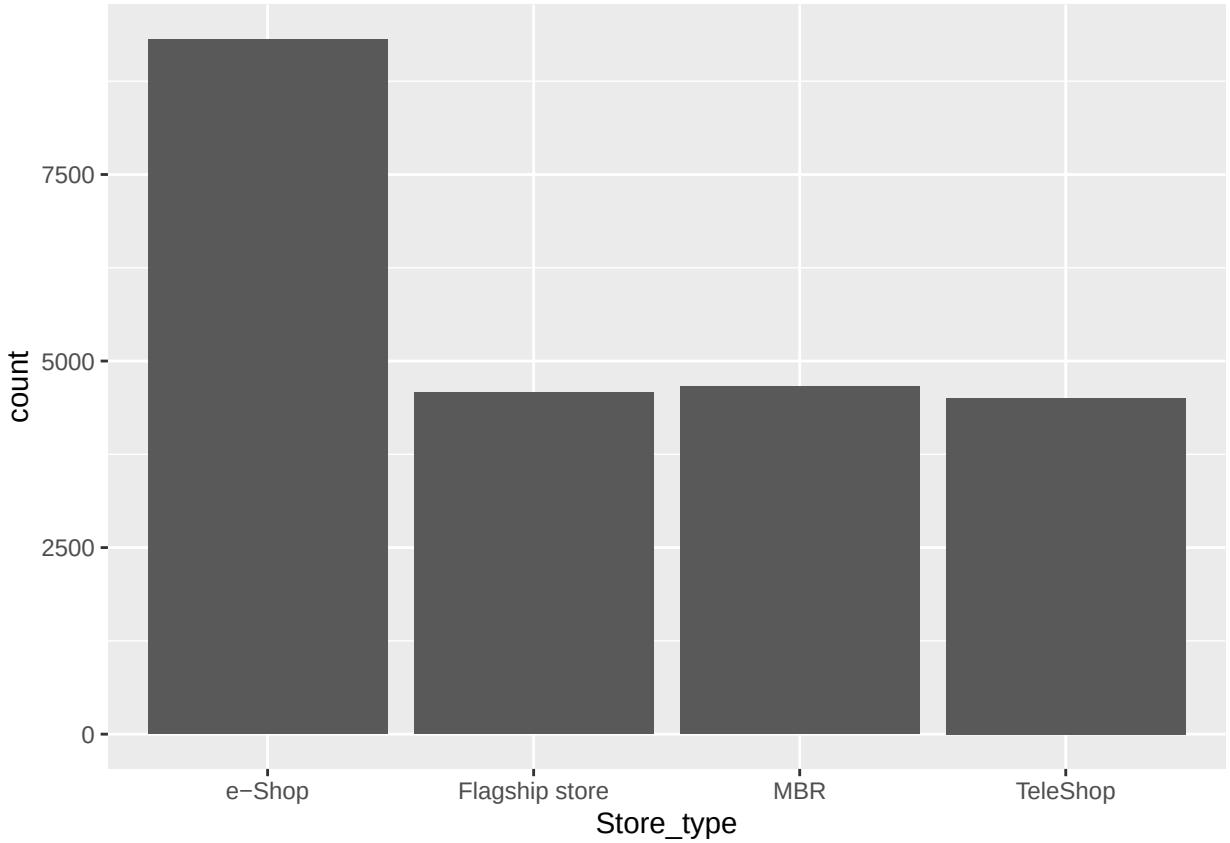
```
hist(df$Tax)
```



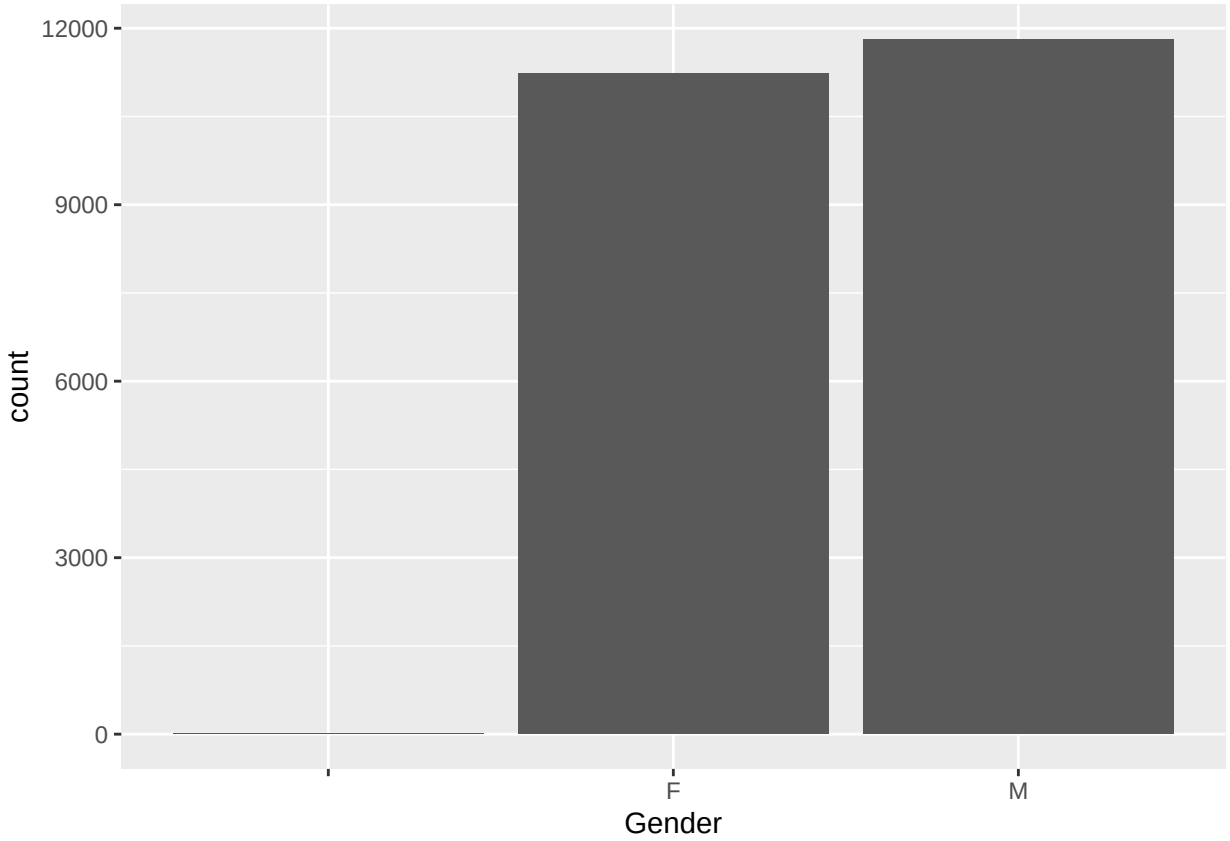

```
hist(df$total_amt)
```



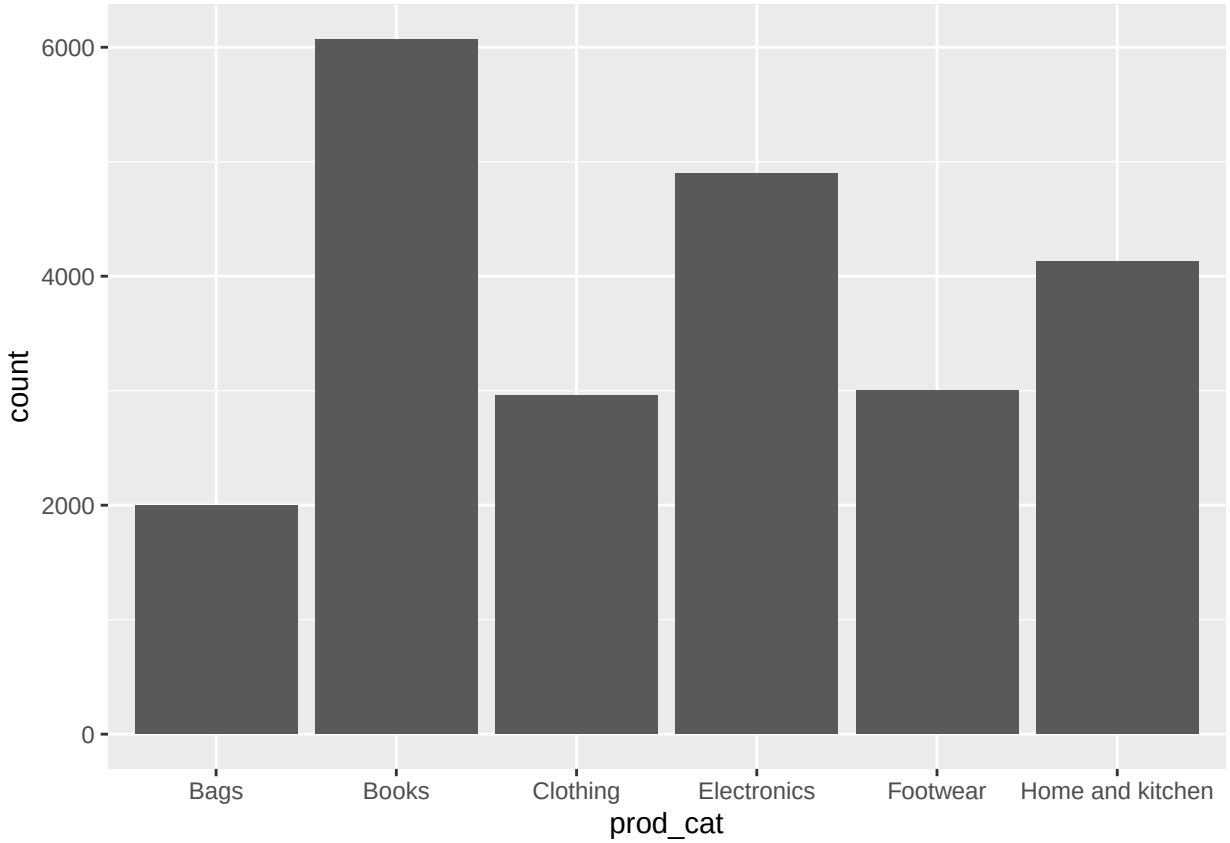

```
ggplot(df) + geom_bar(aes(x=Store_type))
```



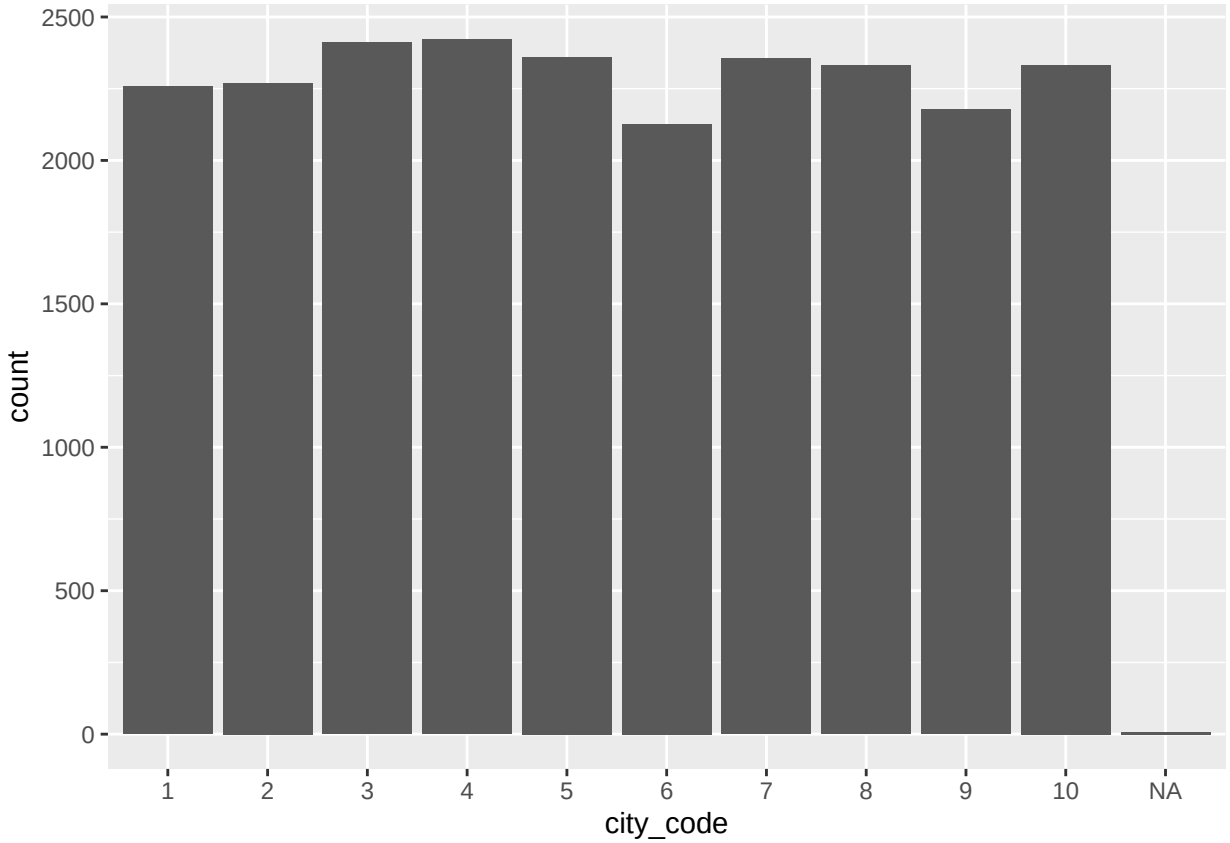

```
ggplot(df) + geom_bar(aes(x=Gender))
```




```
ggplot(df) + geom_bar(aes(x=prod_cat))
```




```
ggplot(df) + geom_bar(aes(x=city_code))
```



Question 4

```
#Q4. Time Range
cat(round(as.numeric(difftime(max_date,min_date,units="weeks"))/52.25), "years")
```

```
## 3 years
```

```
cat(round(as.numeric(difftime(max_date,min_date,units="days"))/(365.25/12)), "months")
```

```
## 37 months
```

```
cat(difftime(max_date,min_date), "days")
```

```
## 1130 days
```

```
#Q4. Count of -ve Transactions
Q4 <- df %>% group_by(transaction_id) %>% summarise(Amt=sum(total_amt))
Q4.1 <- Q4 %>% select(transaction_id,Amt) %>% filter(Amt < 0)
paste0("There are ",count(Q4.1)," customers")
```

```
## [1] "There are 117 customers"
```

Question 5

```
#Q5.
Q5<-df %>% select(Gender,prod_cat,Qty)
Q5a <- Q5 %>% group_by(Gender,prod_cat)  %>% summarise(Total=sum(Qty))
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the `.groups` argument.
```

```
Q5a %>% select(Gender,prod_cat,Total) %>% filter(Total==max(Total))
```

```
## # A tibble: 3 x 3
## # Groups:   Gender [3]
##   Gender prod_cat Total
##   <fct>  <fct>    <int>
## 1 ""     Books       12
## 2 "F"    Books     7070
## 3 "M"    Books     7587
```

```
print("Books are most popular in Males as well as Females")
```

```
## [1] "Books are most popular in Males as well as Females"
```

Question 6

```
#Q6.
Q6 <- df %>% group_by(city_code) %>% summarise(count=n())
Q6a <- mutate(Q6, percentage = count/sum(count)*100)
Q6a %>% select(city_code,count,percentage) %>% filter(count==max(count))
```

```
## # A tibble: 1 x 3
##   city_code count percentage
##   <fct>     <int>      <dbl>
## 1 4          2422       10.5
```

```
print("City code 4 has max percentage")
```

```
## [1] "City code 4 has max percentage"
```

Question 7

```
#Q7.
Q7 <- df %>% group_by(Store_type) %>% summarise(Total_qty=sum(Qty),Total_revenue=sum(total_amt))
Q7 %>% select(Store_type,Total_qty,Total_revenue) %>% filter(Total_qty==max(Total_qty) & Total_revenue==
```

```
## # A tibble: 1 x 3
##   Store_type Total_qty Total_revenue
##   <fct>          <int>         <dbl>
## 1 e-Shop         22763     19824816.
```

Question 8

```r
#Q8.
Q8 <- df %>% filter(Store_type=="Flagship store" & (prod_cat %in% c("Electronics","Clothing")))
Q8 %>% group_by(prod_cat) %>% summarise(Total_revenue=sum(total_amt))
```

```
## # A tibble: 2 x 2
##   prod_cat    Total_revenue
##   <fct>               <dbl>
## 1 Clothing         1194423.
## 2 Electronics      2215136.
```

Question 9

```r
#Q9.
Q9 <- df %>% filter(Gender=="M" & prod_cat=="Electronics")
paste0("The total amount is ", sum(Q9$total_amt))
```

```
## [1] "The total amount is 5703109.425"
```

Question 10

```r
Q10 <- df %>% select(cust_id,transaction_id,total_amt) %>% filter(total_amt > 0)
Q10a <- Q10 %>% group_by(cust_id,transaction_id) %>% summarise(count=n())
```

```
## `summarise()` has grouped output by 'cust_id'. You can override using the `.groups` argument.
```

```r
Q10b <- Q10a %>% group_by(cust_id) %>% summarise(count=n_distinct(transaction_id)) %>% filter(count > 1
cat("There are",nrow(Q10b),"customers")
```

```
## There are 6 customers
```

Question 11

```r
Q11 <- df %>% select(cust_id,prod_cat,prod_subcat,Age,tran_date,total_amt) %>% filter(prod_cat %in% c("
Q11a <- Q11 %>% filter(Age > 25 & Age < 35)
Q11a %>% group_by(prod_cat) %>% summarise(Net=sum(total_amt))
```

```
## # A tibble: 2 x 2
##   prod_cat         Net
##   <fct>          <dbl>
## 1 Books       4978408.
## 2 Electronics 4249648.
```

```r
Q11b <- Q11a %>% filter(tran_date>mydate1 & tran_date<mydate2)
Q11b %>% group_by(prod_cat) %>% summarise(Net=sum(total_amt))
```

```
## # A tibble: 2 x 2
##   prod_cat        Net
##   <fct>         <dbl>
## 1 Books       168527.
## 2 Electronics 164252.
```