retail-analysis-case-study.R

Arun Kumar Prasad

2021-07-30

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
install.packages("contrib.url")
```

```
## Installing package into 'C:/Users/Arun Kumar Prasad/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## Warning: package 'contrib.url' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```
install.packages('tinytex')
```

```
## Installing package into 'C:/Users/Arun Kumar Prasad/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## package 'tinytex' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Arun Kumar Prasad\AppData\Local\Temp\RtmpSQug7X\downloaded_packages
```

```
tinytex::install_tinytex()


install.packages("dplyr")
```

```
## Installing package into 'C:/Users/Arun Kumar Prasad/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'dplyr'

## Warning in file.copy(savedcopy, lib, recursive = TRUE):
## problem copying C:\Users\Arun Kumar Prasad\Documents\R\win-
## library\4.0\00LOCK\dplyr\libs\x64\dplyr.dll to C:\Users\Arun Kumar
## Prasad\Documents\R\win-library\4.0\dplyr\libs\x64\dplyr.dll: Permission denied
```

```
## Warning: restored 'dplyr'
```

```
##
## The downloaded binary packages are in
##  C:\Users\Arun Kumar Prasad\AppData\Local\Temp\RtmpSQug7X\downloaded_packages
```

```
install.packages("ggplot2")
```

```
## Installing package into 'C:/Users/Arun Kumar Prasad/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Arun Kumar Prasad\AppData\Local\Temp\RtmpSQug7X\downloaded_packages
```

```
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
```

```
#To merge using base merge and dplyr package
Customer <- read.csv("D:/praneeta/praneeta/R/R case study 1 (Retail)/Customer.csv")
Transactions <- read.csv("D:/praneeta/praneeta/R/R case study 1 (Retail)/Transactions.csv")
prod_cat_info <- read.csv("D:/praneeta/praneeta/R/R case study 1 (Retail)/prod_cat_info.csv")
customer_trans <- merge(Customer,Transactions,by.x = 'customer_Id',by.y='cust_id')
final_merge <- merge(customer_trans, prod_cat_info, by = 'prod_cat_code', all = T)
View(final_merge)

#by dplyr
custtrans <- full_join(Customer,Transactions,by=c("customer_Id" = "cust_id"))
f_data <- full_join(custtrans,prod_cat_info,by = 'prod_cat_code')
View(f_data)
final_merge$Gender = as.character(final_merge$Gender)
final_merge$DOB = as.Date(final_merge$DOB, format = "%d-%m-%Y")
final_merge$tran_date = as.Date(final_merge$tran_date, format = "%d-%m-%Y")

#question2

#Names of the columns in the dataset and their corresponding data types
str(final_merge) #column names with datatypes
```

```
## 'data.frame':    99293 obs. of  16 variables:
##  $ prod_cat_code    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ customer_Id      : int  266783 266783 266783 266783 266783 266783 270772 270772 270772 268032 ...
##  $ DOB              : Date, format: "1974-05-01" "1974-05-01" ...
##  $ Gender           : chr  "M" "M" "M" "M" ...
##  $ city_code        : int  4 4 4 4 4 4 4 4 4 1 ...
##  $ transaction_id   : num  8.41e+09 8.41e+09 8.41e+09 9.85e+10 9.85e+10 ...
##  $ tran_date        : Date, format: "2013-02-20" "2013-02-20" ...
##  $ prod_subcat_code : int  4 4 4 4 4 4 1 1 1 3 ...
##  $ Qty              : int  1 1 1 3 3 3 4 4 4 5 ...
##  $ Rate             : int  869 869 869 93 93 93 211 211 211 572 ...
##  $ Tax              : num  91.2 91.2 91.2 29.3 29.3 ...
##  $ total_amt        : num  960 960 960 308 308 ...
##  $ Store_type       : chr  "e-Shop" "e-Shop" "e-Shop" "TeleShop" ...
##  $ prod_cat         : chr  "Clothing" "Clothing" "Clothing" "Clothing" ...
##  $ prod_sub_cat_code: int  4 1 3 4 1 3 4 1 3 4 ...
##  $ prod_subcat      : chr  "Mens" "Women" "Kids" "Mens" ...
```

```
#Top 10 and bottom 10 observations
head(final_merge,10)
```

```
##    prod_cat_code customer_Id        DOB Gender city_code transaction_id
## 1              1      266783 1974-05-01      M         4     8410316370
## 2              1      266783 1974-05-01      M         4     8410316370
## 3              1      266783 1974-05-01      M         4     8410316370
## 4              1      266783 1974-05-01      M         4    98477711300
## 5              1      266783 1974-05-01      M         4    98477711300
## 6              1      266783 1974-05-01      M         4    98477711300
## 7              1      270772 1978-07-06      M         4    13147305211
## 8              1      270772 1978-07-06      M         4    13147305211
## 9              1      270772 1978-07-06      M         4    13147305211
## 10             1      268032 1979-02-17      F         1    10912587061
##     tran_date prod_subcat_code Qty Rate     Tax total_amt Store_type prod_cat
## 1  2013-02-20               4   1  869  91.245   960.245     e-Shop Clothing
## 2  2013-02-20               4   1  869  91.245   960.245     e-Shop Clothing
## 3  2013-02-20               4   1  869  91.245   960.245     e-Shop Clothing
## 4  2012-10-21               4   3   93  29.295   308.295   TeleShop Clothing
## 5  2012-10-21               4   3   93  29.295   308.295   TeleShop Clothing
## 6  2012-10-21               4   3   93  29.295   308.295   TeleShop Clothing
## 7  2012-06-21               1   4  211  88.620   932.620        MBR Clothing
## 8  2012-06-21               1   4  211  88.620   932.620        MBR Clothing
## 9  2012-06-21               1   4  211  88.620   932.620        MBR Clothing
## 10 2011-11-24               3   5  572 300.300  3160.300     e-Shop Clothing
##    prod_sub_cat_code prod_subcat
## 1                  4       Mens
## 2                  1      Women
## 3                  3       Kids
## 4                  4       Mens
## 5                  1      Women
## 6                  3       Kids
## 7                  4       Mens
## 8                  1      Women
## 9                  3       Kids
## 10                 4       Mens
```

```
tail(final_merge,10)
```

```
##       prod_cat_code customer_Id        DOB Gender city_code transaction_id
## 99284             6      266820 1978-04-16      F         5    58350344910
## 99285             6      266820 1978-04-16      F         5    58350344910
## 99286             6      270419 1981-05-12      F         2    27576087298
## 99287             6      270419 1981-05-12      F         2    27576087298
## 99288             6      270419 1981-05-12      F         2    27576087298
## 99289             6      270419 1981-05-12      F         2    27576087298
## 99290             6      275265 1990-01-01      M         3    24113900219
## 99291             6      275265 1990-01-01      M         3    24113900219
## 99292             6      275265 1990-01-01      M         3    24113900219
## 99293             6      275265 1990-01-01      M         3    24113900219
##       tran_date prod_subcat_code Qty Rate     Tax total_amt     Store_type
## 99284      <NA>               10   1  447  46.935   493.935 Flagship store
## 99285      <NA>               10   1  447  46.935   493.935 Flagship store
## 99286      <NA>               11   2  856 179.760  1891.760           MBR
## 99287      <NA>               11   2  856 179.760  1891.760           MBR
## 99288      <NA>               11   2  856 179.760  1891.760           MBR
## 99289      <NA>               11   2  856 179.760  1891.760           MBR
## 99290      <NA>                2   3  719 226.485  2383.485 Flagship store
## 99291      <NA>                2   3  719 226.485  2383.485 Flagship store
## 99292      <NA>                2   3  719 226.485  2383.485 Flagship store
## 99293      <NA>                2   3  719 226.485  2383.485 Flagship store
##               prod_cat prod_sub_cat_code prod_subcat
## 99284 Home and kitchen                11        Bath
## 99285 Home and kitchen                12       Tools
## 99286 Home and kitchen                 2  Furnishing
## 99287 Home and kitchen                10     Kitchen
## 99288 Home and kitchen                11        Bath
## 99289 Home and kitchen                12       Tools
## 99290 Home and kitchen                 2  Furnishing
## 99291 Home and kitchen                10     Kitchen
## 99292 Home and kitchen                11        Bath
## 99293 Home and kitchen                12       Tools
```

```
#Min, Q1, median, Q3 and max of the continuous variables.
summary(final_merge$Qty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -5.000   1.000   3.000   2.438   4.000   5.000
```

```
summary(final_merge$Rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1499.0   313.0   713.0   637.9  1109.0  1500.0
```

```
summary(final_merge$Tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.35   98.28  199.92  248.87  366.98  787.50
```

```r
summary(final_merge$total_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -8270.9   762.5  1761.4  2114.6  3585.7  8287.5
```

```r
#Frequency tables for all the categorical variables
head(table(factor(final_merge$customer_Id),exclude=NULL))
```

```
##
## 266783 266784 266785 266788 266794 266799
##     18     17     28     14     48     19
```

```r
head(table(final_merge$DOB,exclude=NULL))
```

```
##
## 1970-01-02 1970-01-07 1970-01-08 1970-01-10 1970-01-11 1970-01-15
##         39         15         22         48         12         20
```

```r
head(table(final_merge$Gender,exclude=NULL),10)
```

```
##
##           F     M
##    40 48202 51051
```

```r
head(table(factor(final_merge$city_code),exclude = NULL),10)
```

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

```r
head(table(factor(final_merge$transaction_id),exclude = NULL),10)
```

```
##
##  3268991  7073244 10861359 15741026 16165359 18629385 29740699 33156503
##        4        6        2        6        3        4        6        3
## 38816402 41453307
##        6        6
```

```r
head(table(final_merge$tran_date, exclude=NULL),10)
```

```
##
## 2011-01-25 2011-01-26 2011-01-27 2011-01-28 2011-01-29 2011-01-30 2011-01-31
##         89         95         84         61         81        121        103
## 2011-02-13 2011-02-14 2011-02-15
##         90         73        101
```

```r
head(table(factor(final_merge$prod_subcat_code), exclude = NULL),10)
```

```
##
##     1     2     3     4     5     6     7     8     9    10
##  7847  4028 12294 13073  4790  5934  6258  4860  4925 14932
```

```r
head(table(factor(final_merge$prod_cat_code),exclude=NULL),10)
```

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

```r
head(table(final_merge$Store_type, exclude=NULL),10)
```

```
##
##         e-Shop Flagship store            MBR       TeleShop
##          40185          19814          19974          19320
```

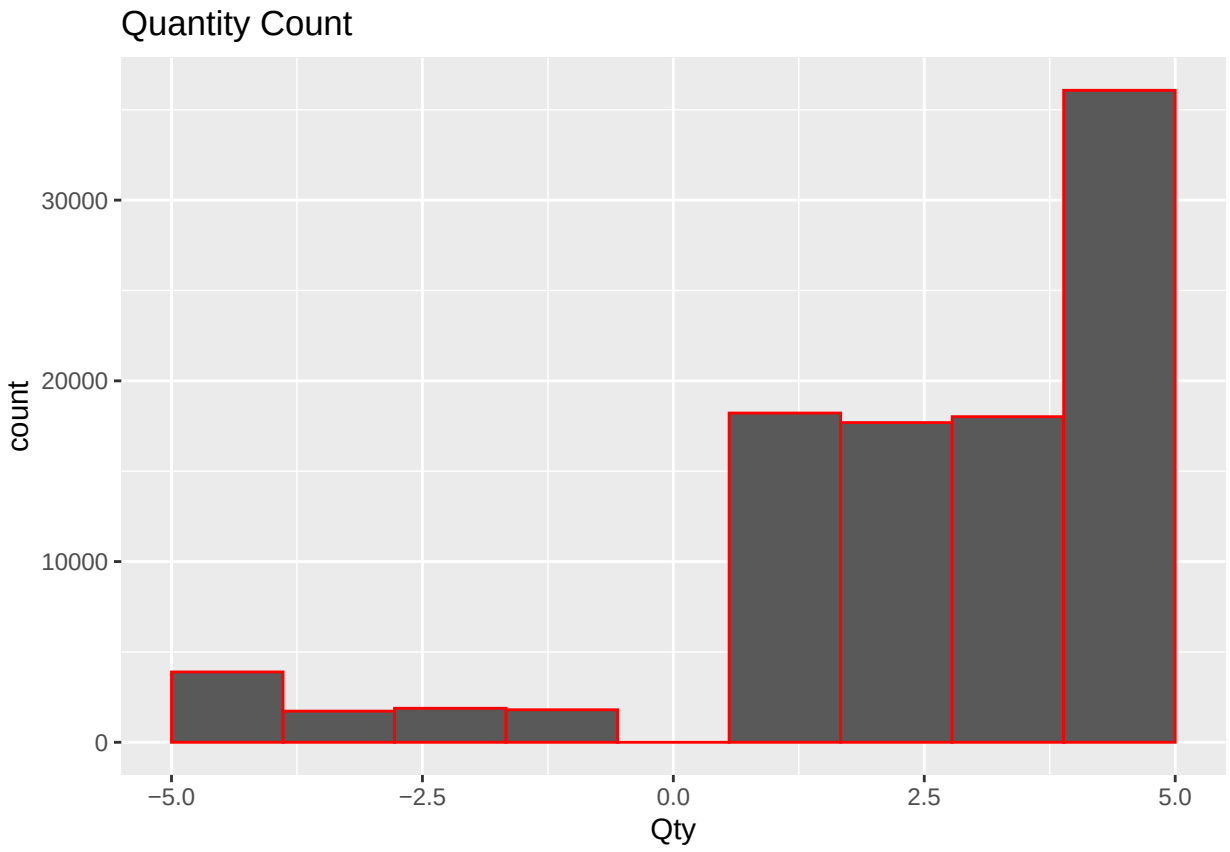
```r
head(table(final_merge$prod_cat,exclude=NULL),10)
```

```
##
##             Bags            Books         Clothing      Electronics
##             3996            36414             8880            24490
##         Footwear Home and kitchen
##             8997            16516
```

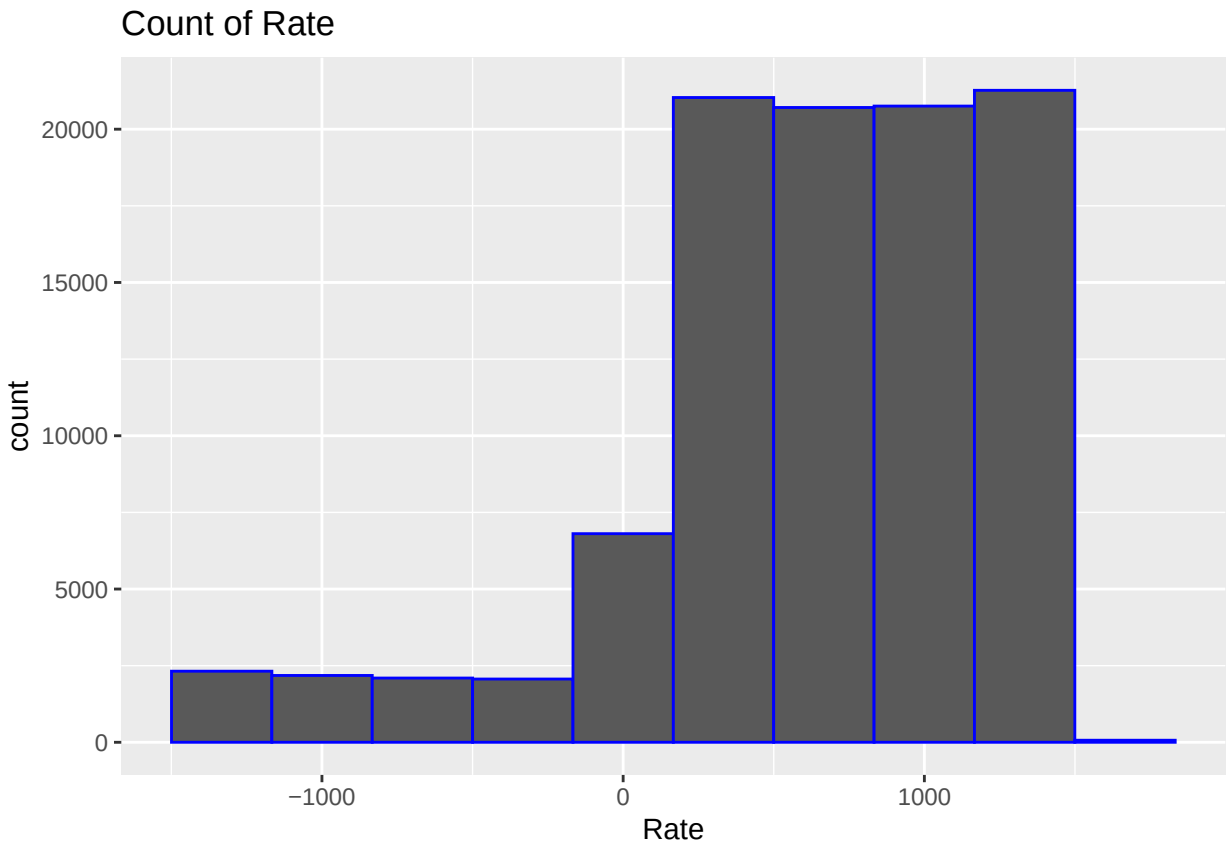
```r
head(table(final_merge$prod_subcat, exclude=NULL),10)
```

```
##
##        Academic Audio and video            Bath         Cameras        Children
##            6069            4898            4129            4898            6069
##          Comics       Computers             DIY         Fiction      Furnishing
##            6069            4898            6069            6069            4129
```

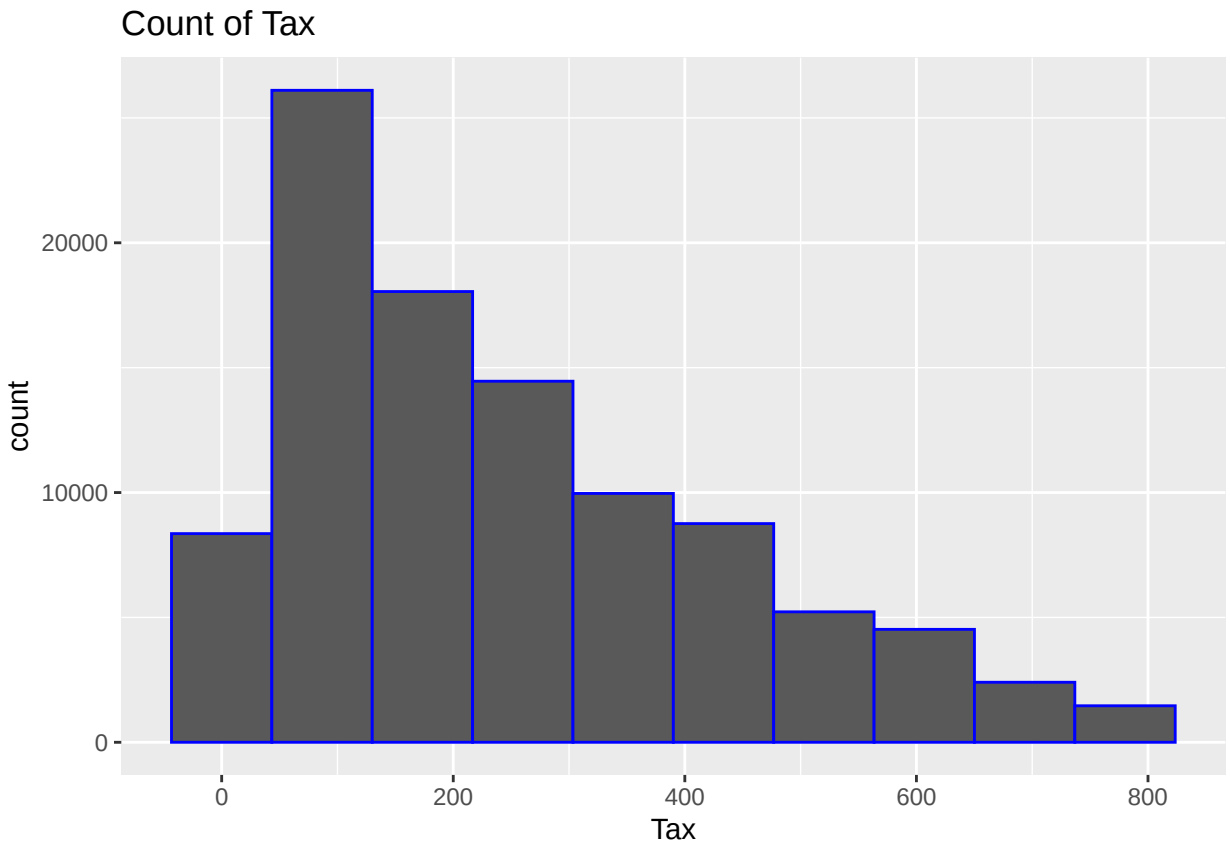
```r
#histograms for continuous variables
library(ggplot2)
ggplot(data = final_merge,aes(x=Qty))+geom_histogram(color="red",bins=10)+ggtitle("Quantity Count")
```



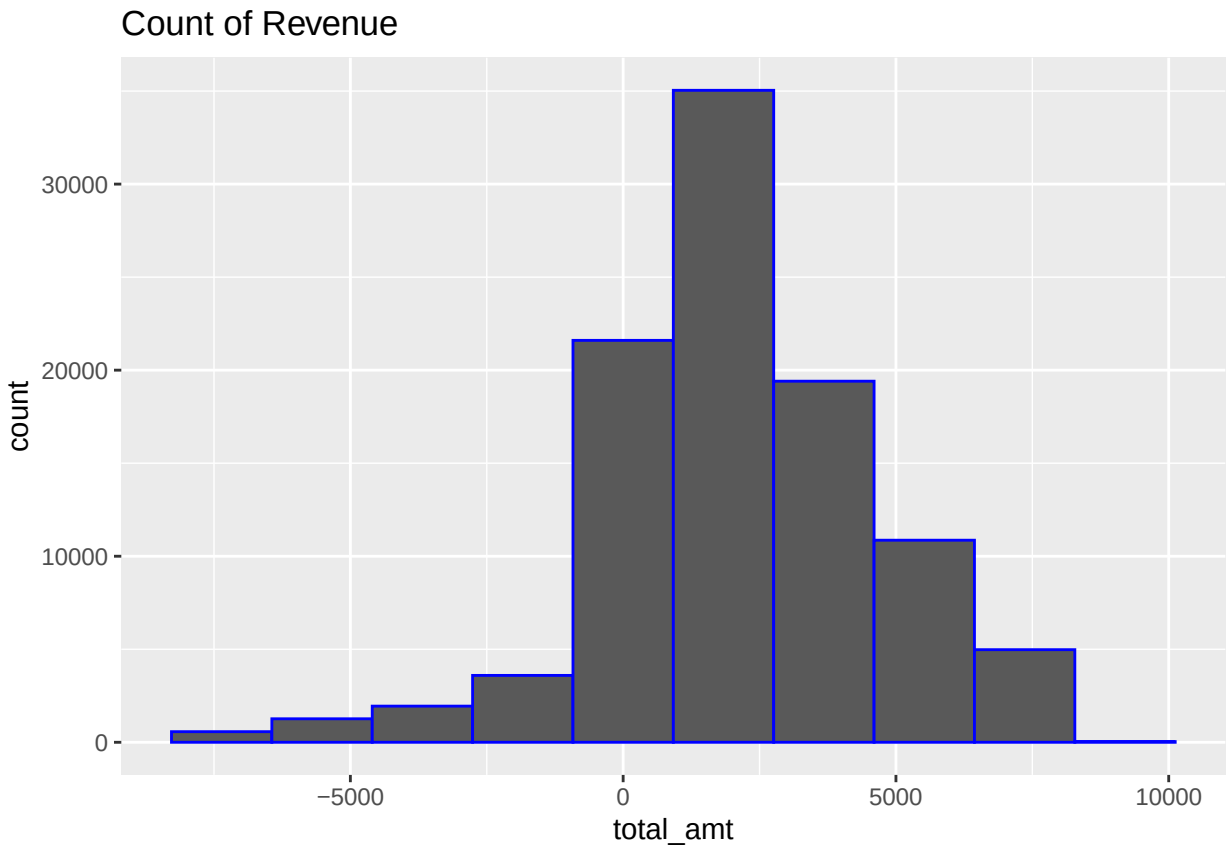

```
ggplot(data=final_merge,aes(x=Rate))+geom_histogram(color="blue",bins=10)+ggtitle("Count of Rate")
```



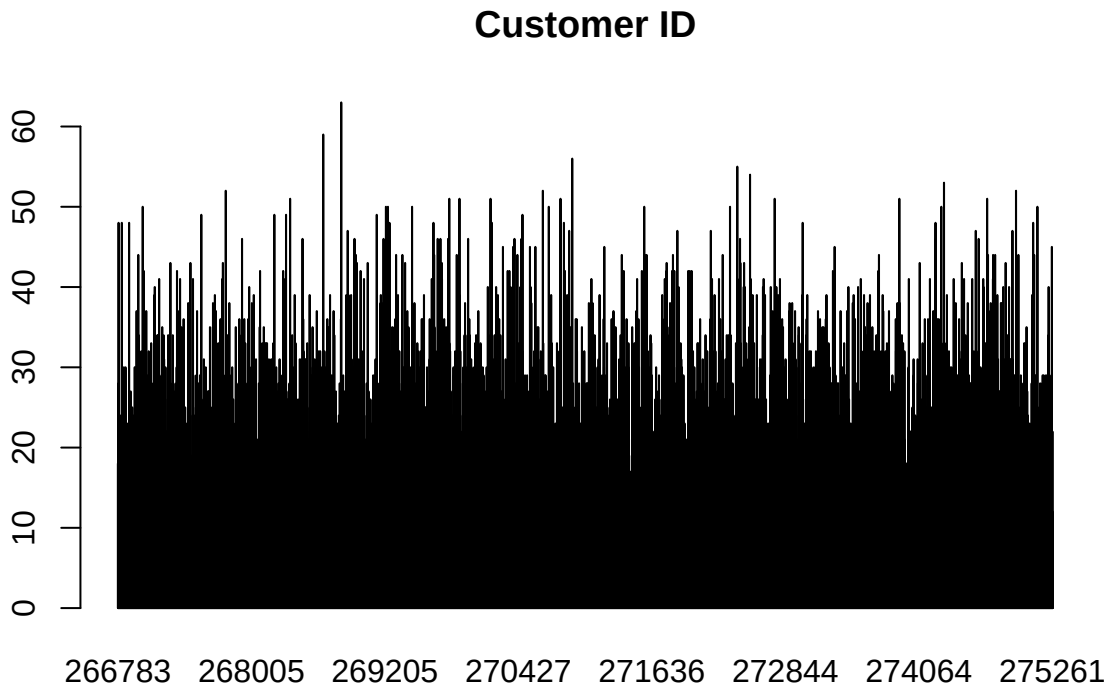

```
ggplot(data=final_merge,aes(x=Tax))+geom_histogram(color="blue",bins=10)+ggtitle("Count of Tax")
```



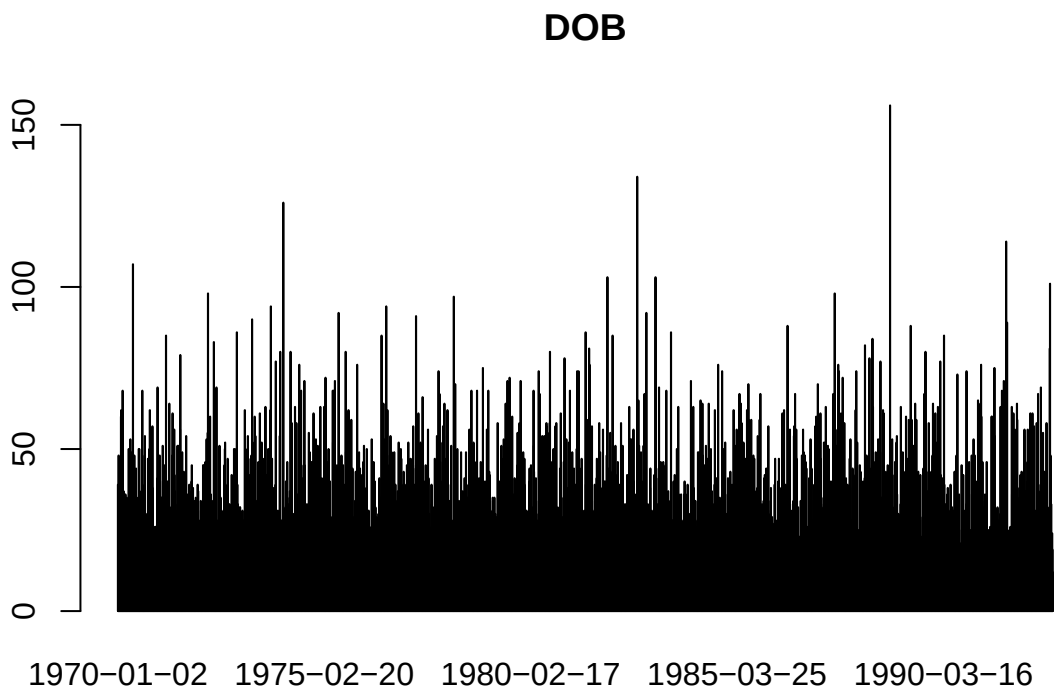

```
ggplot(data=final_merge,aes(x=total_amt))+geom_histogram(color="blue",bins=10)+ggtitle("Count of Revenue
```



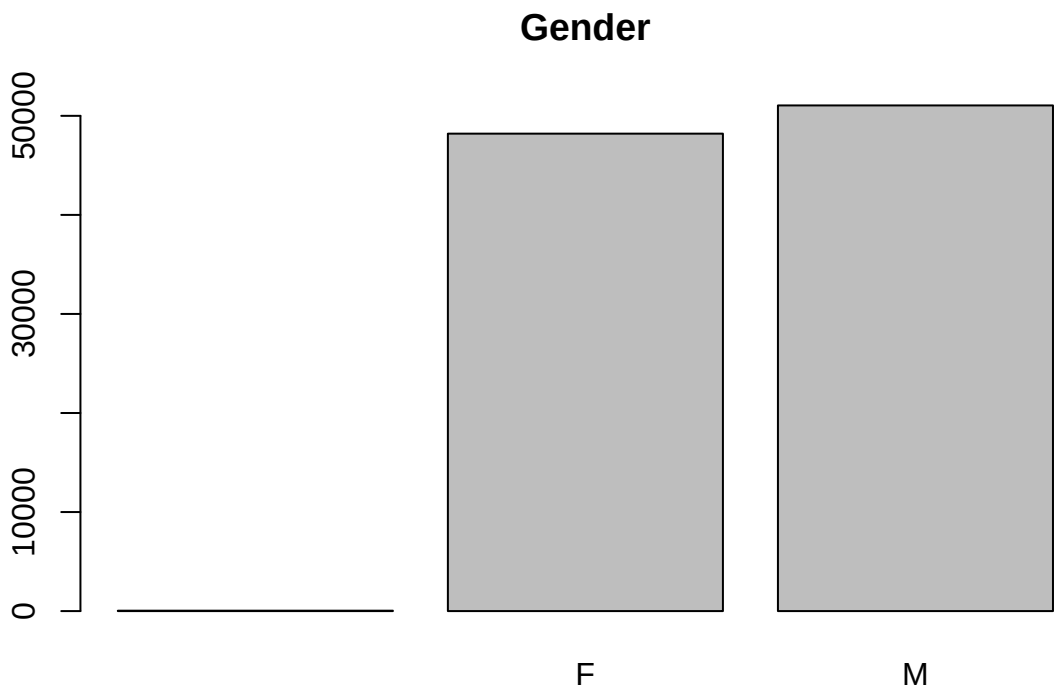

```
#frequency tables for categorical variables
barplot(table(factor(final_merge$customer_Id),exclude=NULL),main="Customer ID")
```



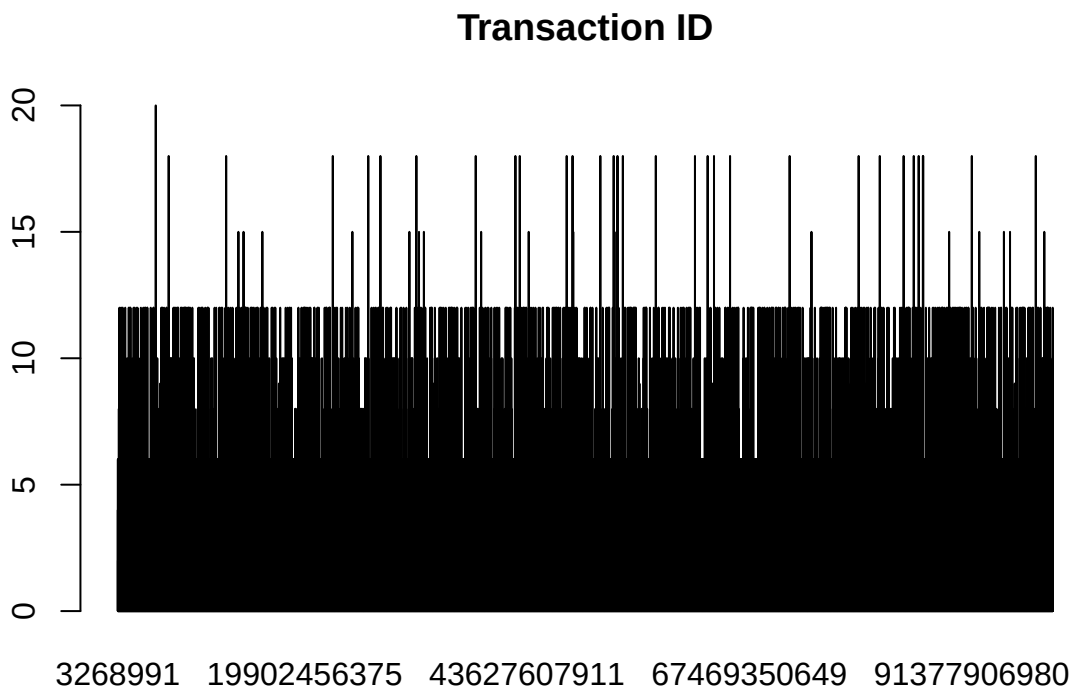

```
barplot(table(final_merge$DOB,exclude=NULL),main = "DOB")
```




```
barplot(table(final_merge$Gender,exclude=NULL),main = "Gender")
```

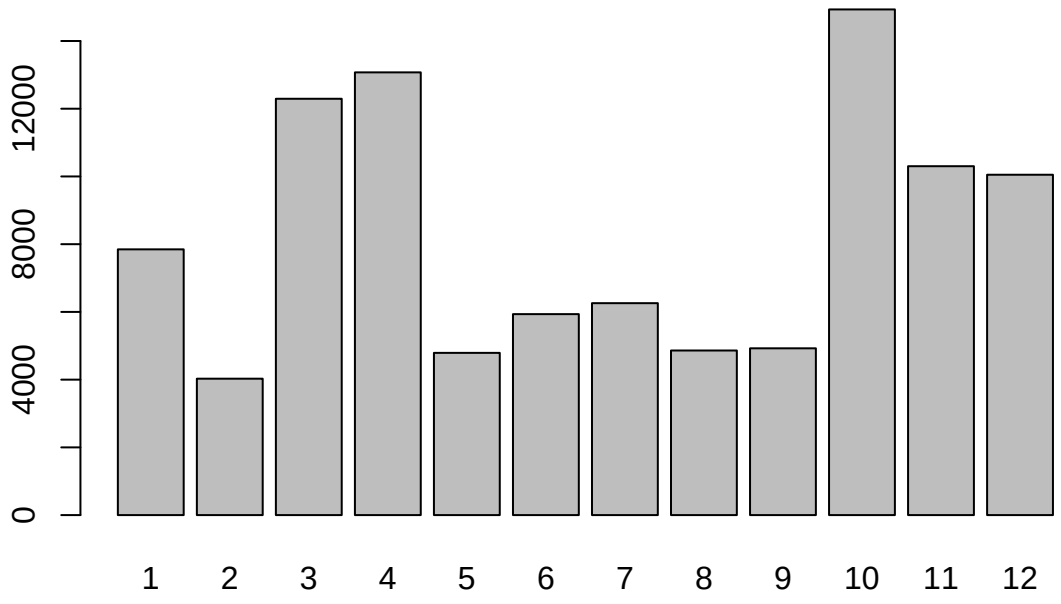



```
barplot(table(factor(final_merge$transaction_id),exclude = NULL),main = "Transaction ID")
```

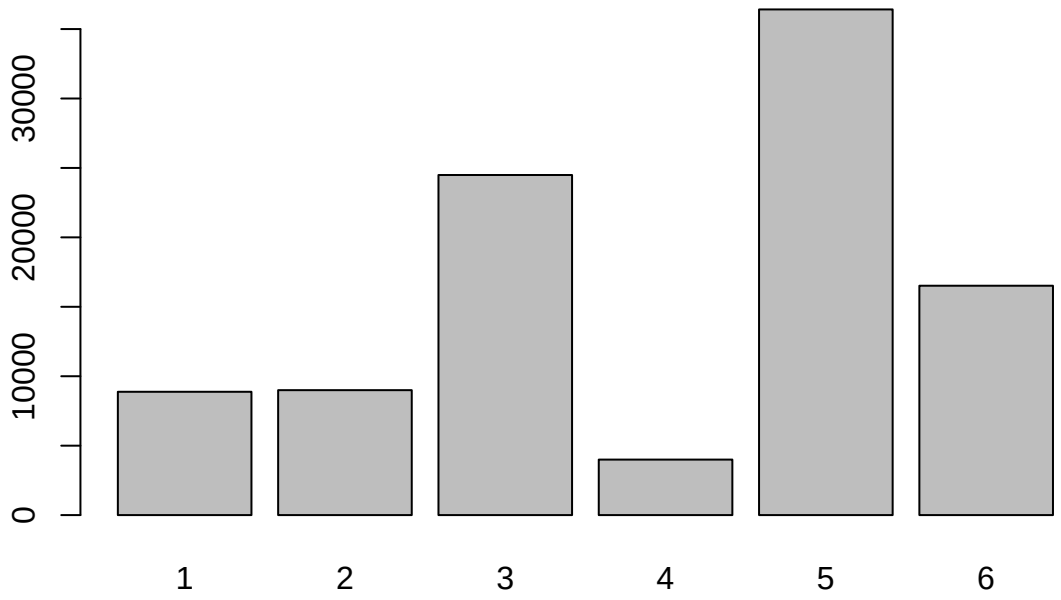



```
barplot(table(factor(final_merge$city_code),exclude = NULL), main = "City Code")
```


```
barplot(table(factor(final_merge$prod_subcat_code), exclude = NULL),main = "Product Subcategory Code")
```

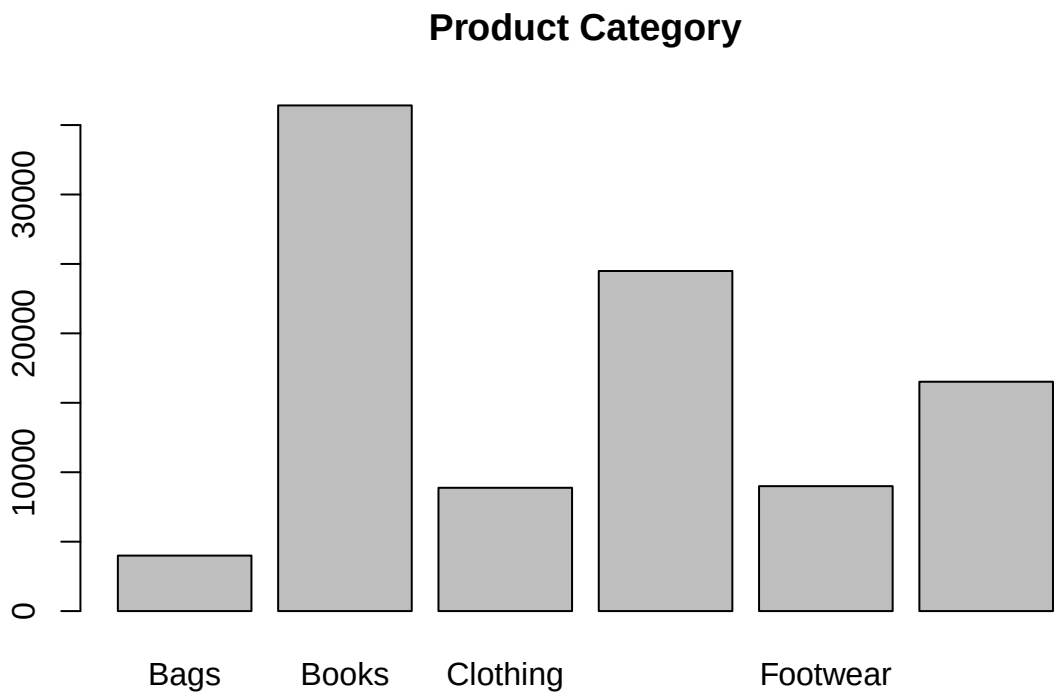



```
barplot(table(factor(final_merge$prod_cat_code),exclude=NULL),main = "Product Category Code")
```

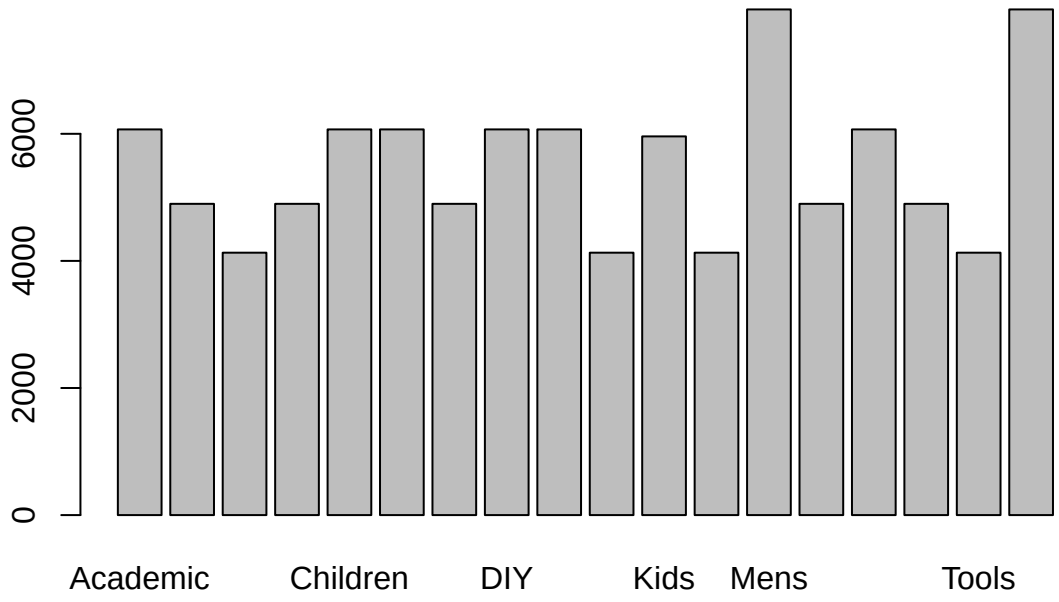



```
barplot(table(final_merge$Store_type, exclude=NULL), main = "Store Type")
```


```
barplot(table(final_merge$prod_cat,exclude=NULL),main = "Product Category")
```




```
barplot(table(final_merge$prod_subcat, exclude=NULL),main="Product Subcategory")
```




```
#Time period of the transaction data
min_date <- min(final_merge$tran_date,na.rm=TRUE)
max_date <- max(final_merge$tran_date,na.rm=TRUE)
range_date <- max_date - min_date
range_date
```

```
## Time difference of 1130 days
```

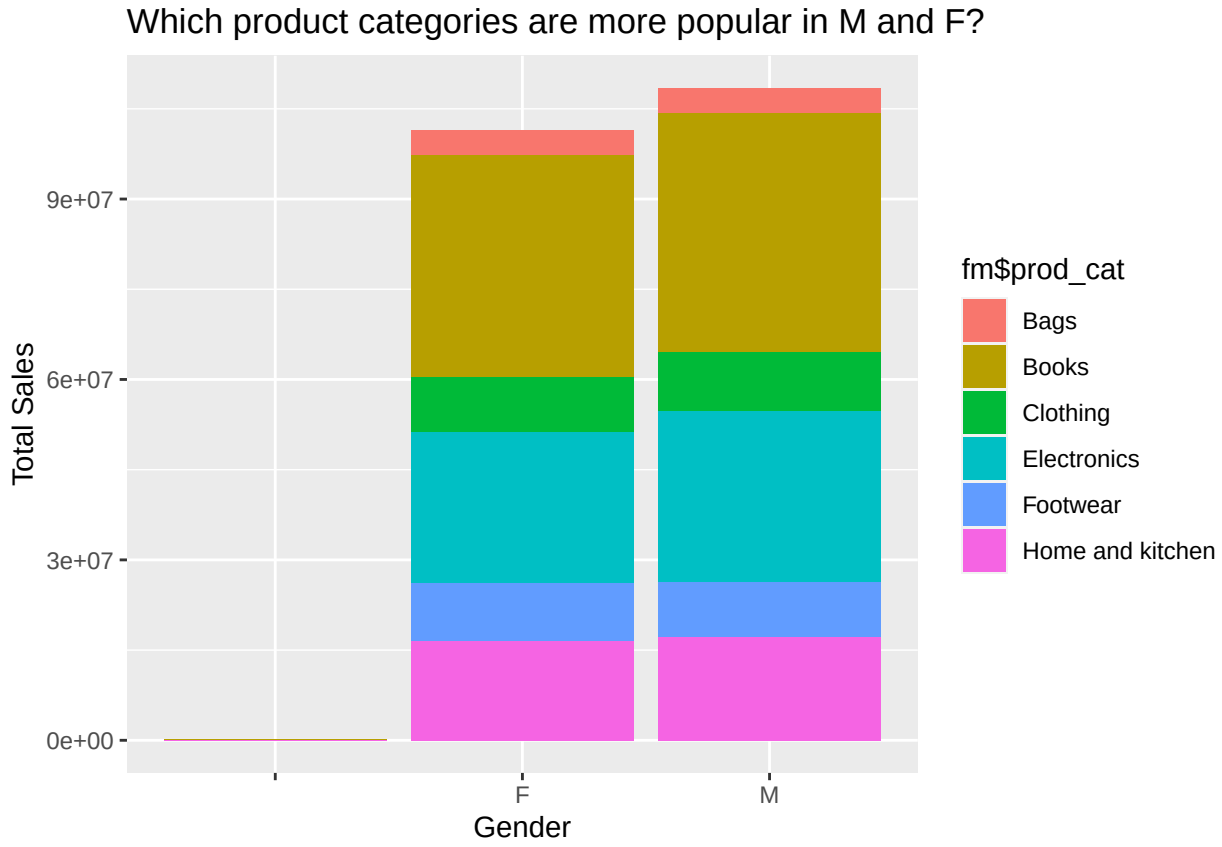
```
#Count of transactions with negative total amount
final_merge %>% filter(final_merge$total_amt < 0) %>% nrow()
```

```
## [1] 9294
```

```
#Comparison of sales of product categories among males and females
fm <- final_merge %>%
  group_by(Gender,prod_cat) %>%
  summarise(Totalsales = sum(total_amt,na.rm = TRUE))
```

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

```
ggplot(fm, aes(fill=fm$prod_cat, y=fm$Totalsales, x=fm$Gender)) +
  geom_bar(position="stack", stat="identity") +
  labs(x="Gender", y= "Total Sales", colour = "Product Category") +
  ggtitle("Which product categories are more popular in M and F?")
```




```
#Find the city with the maximum customers and the percentage of customers from that city
a <- final_merge %>% group_by(final_merge$city_code) %>% summarise(n=length(city_code))
max_city <- a[which.max(a$n),]
max_city
```

```
## # A tibble: 1 x 2
##   `final_merge$city_code`     n
##                     <int> <int>
## 1                       4 10571
```

```
paste("max no. of customers are from city code ",max_city$`final_merge$city_code`,":",max_city$n) #perc
```

```
## [1] "max no. of customers are from city code  4 : 10571"
```

```
paste("Percentage of customers:",round((max_city$n/sum(a$n))*100,2),"%") #percentage
```

```
## [1] "Percentage of customers: 10.65 %"
```

```
#Total amount earned in the store type "Flagship Stores" from the categories:
#a) Electronics
#b) Clothing
dataa <- final_merge %>%
  group_by(Store_type,prod_cat) %>%
  summarise(TotalRevenue = sum(final_merge$total_amt,na.rm=T))
```

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

```
dataa[dataa$Store_type =='Flagship store' & dataa$prod_cat == 'Clothing',]
```

```
## # A tibble: 1 x 3
## # Groups:   Store_type [1]
##   Store_type     prod_cat TotalRevenue
##   <chr>          <chr>           <dbl>
## 1 Flagship store Clothing   209966608.
```

```
dataa[dataa$Store_type =='Flagship store' & dataa$prod_cat == 'Electronics',]
```

```
## # A tibble: 1 x 3
## # Groups:   Store_type [1]
##   Store_type     prod_cat    TotalRevenue
##   <chr>          <chr>             <dbl>
## 1 Flagship store Electronics   209966608.
```

```
#Total amount earned from category "Male" under the "Electronics" category
datab <- final_merge %>%
  group_by(Gender,prod_cat) %>%
  summarise(TotalRevenue = sum(final_merge$total_amt,na.rm=T))
```

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

```
datab[datab$Gender =='M' & datab$prod_cat == 'Electronics',]
```

```
## # A tibble: 1 x 3
## # Groups:   Gender [1]
##   Gender prod_cat    TotalRevenue
##   <chr>  <chr>              <dbl>
## 1 M      Electronics   209966608.
```

```
#No. of customers with more 10 unique non-negative transactions
t1 <- final_merge %>%
  group_by(customer_Id,total_amt) %>%
  summarise()
```

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

```
t2 <- t1 %>%
  group_by(customer_Id) %>%
  summarise(nonneg = length(which(total_amt>0)))
t2
```

```
## # A tibble: 5,506 x 2
##    customer_Id nonneg
##          <int>  <int>
##  1      266783      4
##  2      266784      3
```

```
##  3      266785      7
##  4      266788      4
##  5      266794     11
##  6      266799      3
##  7      266803      1
##  8      266804      1
##  9      266805      1
## 10      266806      6
## # ... with 5,496 more rows
```

```
t2 %>% summarise(numberofcust = length(which(t2$nonneg > 10)))
```

```
## # A tibble: 1 x 1
##   numberofcust
##          <int>
## 1            6
```

```
#For all customers aged 25-35 find Total amount spent in "Electronics" and "Books" categories
cust_age <- ((final_merge$tran_date - final_merge$DOB)/365.25)
final_merge$age_grp <- ifelse(cust_age >=25 & cust_age <=35,"O","Y")

age <- final_merge %>%
  group_by(age_grp,prod_cat) %>%
  summarise(Totalrevenue = sum(total_amt,na.rm=T))
```

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

```
age[age$age_grp == "O" & age$prod_cat == "Electronics",]
```

```
## # A tibble: 2 x 3
## # Groups:   age_grp [2]
##   age_grp prod_cat    Totalrevenue
##   <chr>   <chr>              <dbl>
## 1 O       Electronics    13718498.
## 2 <NA>    <NA>                 NA
```

```
age[age$age_grp == "O" & age$prod_cat == "Books",]
```

```
## # A tibble: 2 x 3
## # Groups:   age_grp [2]
##   age_grp prod_cat Totalrevenue
##   <chr>   <chr>           <dbl>
## 1 O       Books       20290465.
## 2 <NA>    <NA>               NA
```

```
#Total amount spent between January 1, 2014 and March 1, 2014
tran1 <- final_merge %>%
  group_by(tran_date,age_grp) %>%
  summarise(totalrevenue = sum(total_amt,na.rm = T))
```

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

```
tran1[tran1$tran_date >= '2014-01-01' & tran1$tran_date <= '2014-03-01' & tran1$age_grp == "0",]
```

```
## # A tibble: 33 x 3
## # Groups:   tran_date [33]
##    tran_date  age_grp totalrevenue
##    <date>     <chr>          <dbl>
##  1 2014-01-13 0              55042.
##  2 2014-01-14 0              86200.
##  3 2014-01-15 0              97930.
##  4 2014-01-16 0              70087.
##  5 2014-01-17 0             152713.
##  6 2014-01-18 0             115147.
##  7 2014-01-19 0              46619.
##  8 2014-01-20 0              68657.
##  9 2014-01-21 0              57393.
## 10 2014-01-22 0             110321.
## # ... with 23 more rows
```