R\_CaseStudy\_Retail

Pranay

16/11/2020

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

##   
## 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(ggplot2)  
  
Customer = read.csv('D:\\Study\\AnalytixLabs\\Case Studies\\R\\R case study 1 (Retail)\\Customer.csv')  
prod\_cat\_info = read.csv('D:\\Study\\AnalytixLabs\\Case Studies\\R\\R case study 1 (Retail)\\prod\_cat\_info.csv')  
Transactions = read.csv('D:\\Study\\AnalytixLabs\\Case Studies\\R\\R case study 1 (Retail)\\Transactions.csv')  
  
  
  
  
############################## Q1 ################################  
# a  
  
Customer\_Final = merge(Customer, Transactions, by.x = 'customer\_Id', by.y = 'cust\_id' ) %>%  
 merge(prod\_cat\_info, by.x = c('prod\_subcat\_code', 'prod\_cat\_code'), by.y = c('prod\_sub\_cat\_code' , 'prod\_cat\_code'))  
  
# b  
Customer\_Final\_2 = left\_join(Transactions,Customer,  
 by=c("cust\_id"="customer\_Id"),  
 all.x=T)   
  
  
  
  
  
############################## Q2 ################################  
# a :-  
str(Customer\_Final)

## 'data.frame': 23053 obs. of 15 variables:  
## $ prod\_subcat\_code: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ prod\_cat\_code : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ customer\_Id : int 274066 268597 274484 268701 267134 268161 274260 274926 268126 274618 ...  
## $ DOB : chr "26-10-1976" "01-05-1986" "13-11-1976" "22-02-1978" ...  
## $ Gender : chr "M" "F" "F" "M" ...  
## $ city\_code : int 8 10 9 5 2 4 3 2 2 7 ...  
## $ transaction\_id : num 3.81e+10 7.69e+10 1.88e+10 1.06e+10 9.78e+10 ...  
## $ tran\_date : chr "2013/03/05" "2013/01/19" "2011/08/20" "2012/01/30" ...  
## $ Qty : int 5 4 2 5 5 3 2 5 1 4 ...  
## $ Rate : int 1239 853 324 1342 461 1309 630 224 1249 561 ...  
## $ Tax : num 650 358 68 705 242 ...  
## $ total\_amt : num 6845 3770 716 7415 2547 ...  
## $ Store\_type : chr "Flagship store" "MBR" "Flagship store" "MBR" ...  
## $ prod\_cat : chr "Clothing" "Clothing" "Clothing" "Clothing" ...  
## $ prod\_subcat : chr "Women" "Women" "Women" "Women" ...

# b :-  
head(Customer\_Final, 10)

## prod\_subcat\_code prod\_cat\_code customer\_Id DOB Gender city\_code  
## 1 1 1 274066 26-10-1976 M 8  
## 2 1 1 268597 01-05-1986 F 10  
## 3 1 1 274484 13-11-1976 F 9  
## 4 1 1 268701 22-02-1978 M 5  
## 5 1 1 267134 14-09-1992 M 2  
## 6 1 1 268161 25-06-1979 M 4  
## 7 1 1 274260 10-11-1991 M 3  
## 8 1 1 274926 20-10-1986 M 2  
## 9 1 1 268126 12-09-1975 F 2  
## 10 1 1 274618 09-12-1992 F 7  
## transaction\_id tran\_date Qty Rate Tax total\_amt Store\_type prod\_cat  
## 1 38110368883 2013/03/05 5 1239 650.475 6845.475 Flagship store Clothing  
## 2 76949751078 2013/01/19 4 853 358.260 3770.260 MBR Clothing  
## 3 18838561375 2011/08/20 2 324 68.040 716.040 Flagship store Clothing  
## 4 10577817962 2012/01/30 5 1342 704.550 7414.550 MBR Clothing  
## 5 97764417070 2011/12/09 5 461 242.025 2547.025 MBR Clothing  
## 6 44627697269 2012/07/24 3 1309 412.335 4339.335 Flagship store Clothing  
## 7 78340973279 2012/01/30 2 630 132.300 1392.300 e-Shop Clothing  
## 8 41963699352 2012/10/29 5 224 117.600 1237.600 Flagship store Clothing  
## 9 11950520948 2012/08/11 1 1249 131.145 1380.145 TeleShop Clothing  
## 10 95974831788 2014/01/13 4 561 235.620 2479.620 Flagship store Clothing  
## prod\_subcat  
## 1 Women  
## 2 Women  
## 3 Women  
## 4 Women  
## 5 Women  
## 6 Women  
## 7 Women  
## 8 Women  
## 9 Women  
## 10 Women

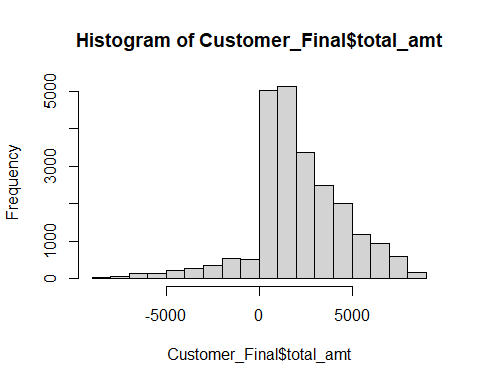
tail(Customer\_Final, 10)

## prod\_subcat\_code prod\_cat\_code customer\_Id DOB Gender city\_code  
## 23044 9 3 275137 08-03-1988 M 9  
## 23045 9 3 272193 26-04-1986 M 4  
## 23046 9 3 273977 25-07-1985 F 7  
## 23047 9 3 269633 11-05-1985 M 7  
## 23048 9 3 272252 06-03-1991 F 10  
## 23049 9 3 272427 09-04-1981 F 1  
## 23050 9 3 271275 01-07-1976 M 8  
## 23051 9 3 269798 28-02-1973 F 4  
## 23052 9 3 271714 01-01-1990 M 1  
## 23053 9 3 273490 07-06-1974 M 4  
## transaction\_id tran\_date Qty Rate Tax total\_amt Store\_type  
## 23044 65114820194 2013/09/19 2 1398 293.580 3089.580 e-Shop  
## 23045 81904564239 2012/05/21 3 776 244.440 2572.440 TeleShop  
## 23046 35607073389 2013/10/01 4 760 319.200 3359.200 Flagship store  
## 23047 73909207905 2013/08/26 1 233 24.465 257.465 e-Shop  
## 23048 94626119963 2013/03/07 5 217 113.925 1198.925 TeleShop  
## 23049 43083030068 2013/01/05 2 1069 224.490 2362.490 TeleShop  
## 23050 93775847705 2013/07/21 -3 -658 207.270 -2181.270 e-Shop  
## 23051 5513185492 2013/07/26 2 1039 218.190 2296.190 e-Shop  
## 23052 95736165852 2013/12/01 1 171 17.955 188.955 Flagship store  
## 23053 48695915351 2011/11/29 -5 -560 294.000 -3094.000 TeleShop  
## prod\_cat prod\_subcat  
## 23044 Electronics Cameras  
## 23045 Electronics Cameras  
## 23046 Electronics Cameras  
## 23047 Electronics Cameras  
## 23048 Electronics Cameras  
## 23049 Electronics Cameras  
## 23050 Electronics Cameras  
## 23051 Electronics Cameras  
## 23052 Electronics Cameras  
## 23053 Electronics Cameras

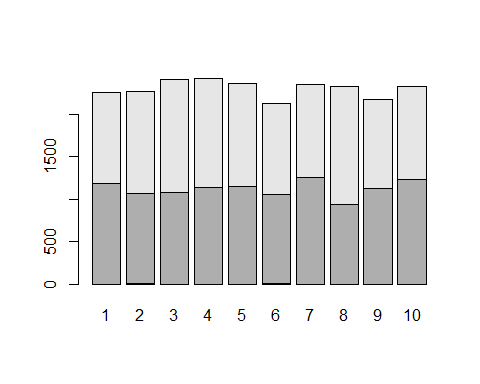
# c :-  
summary(Customer\_Final$Qty, Customer\_Final$Rate, Customer\_Final$Tax, Customer\_Final$total\_amt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -5.000 1.000 3.000 2.432 4.000 5.000

# d :-  
Frq\_Tbl\_Cat\_Var = table(Gen = Customer\_Final$Gender, City\_Code = Customer\_Final$city\_code, Store\_Type = Customer\_Final$Store\_type,  
 Prod\_Cat = Customer\_Final$prod\_cat, Prod\_Sub\_Cat = Customer\_Final$prod\_subcat)  
  
  
  
  
  
############################## Q3 ################################  
  
hist(Customer\_Final$total\_amt)



barplot(table(Gen = Customer\_Final$Gender, City\_Code = Customer\_Final$city\_code))



############################## Q4 ################################  
# a :-  
diff\_in\_Days = difftime(max(Customer\_Final$tran\_date), min(Customer\_Final$tran\_date), units = "auto")  
  
# b :-  
  
Negative\_Transactions = Customer\_Final[Customer\_Final$total\_amt < 0, ]  
  
  
  
  
############################## Q5 ################################  
  
Ans5 = table(Customer\_Final$Gender, Customer\_Final$prod\_cat)  
  
  
  
  
  
############################## Q6 ################################  
  
Ans6 = c('City-Code max Customers = ', Customer[max(table(Customer$city\_code)),'city\_code'],  
 'Percentage of customers from this city = ', round(max(table(Customer$city\_code))/sum(table(Customer$city\_code)) \* 100))  
  
  
  
  
############################## Q7 ################################  
  
Ans7 = group\_by(Customer\_Final, Store\_type) %>%  
 summarise(sum(total\_amt), sum(Qty))

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

############################## Q8 ################################  
  
Ans8 = Customer\_Final[(Customer\_Final$prod\_cat == c('Electronics','Clothing')) &   
 (Customer\_Final$Store\_type == 'Flagship store'),] %>%  
 group\_by(prod\_cat) %>%  
 summarise(sum(total\_amt))

## Warning in Customer\_Final$prod\_cat == c("Electronics", "Clothing"): longer  
## object length is not a multiple of shorter object length

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

############################## Q9 ################################  
  
Ans9 = Customer\_Final[(Customer\_Final$Gender == 'M') & (Customer\_Final$prod\_cat == 'Electronics'),] %>%  
 group\_by(Gender, prod\_cat) %>%  
 summarise(sum(total\_amt))

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

############################## Q10 ################################  
  
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

## Warning: package 'RSQLite' was built under R version 4.0.3

Ans10 = sqldf('select customer\_Id, (case when total\_amt > 0 then count(customer\_Id) end )  
 from Customer\_Final  
 group by customer\_Id  
 having (case when total\_amt > 0 then count(customer\_Id) end ) > 10')  
  
  
  
  
  
############################## Q11 ################################  
  
Customer\_Final$Age\_Yrs = difftime(as.Date(Sys.Date(), format="%d-%m-%Y")  
 , as.Date(Customer\_Final$DOB, format="%d-%m-%Y"),  
 tz, units = c("auto", "secs", "mins", "hours","days", "weeks"))/365  
  
  
x = Customer\_Final[Customer\_Final$Age\_Yrs >= 25 & Customer\_Final$Age\_Yrs <= 35,]  
  
y = x[x$prod\_cat == c('Electronics','Books'),]  
  
  
  
Ans11\_a = group\_by(y, prod\_cat) %>%  
 summarise(sum(total\_amt))

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

Ans11\_b = Customer\_Final %>%  
 filter(tran\_date > "2014-01-01" & tran\_date <"2014-03-01") %>%  
 summarise(ttl\_amt1=sum(total\_amt))