---

title: "Assignment1"

name : Aravind

output: html\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

# Reading the csv file

```{r}

Retail\_Management<-read.csv("Online\_Retail.csv")

head(Retail\_Management)

library(dplyr)

library(ISLR)

```

1. Show the breakdown of the number of transactions by countries i.e. how many transactions are in the dataset for each country (consider all records including cancelled transactions). Show this in total number and also in percentage. Show only countries accounting for more than 1% of the total transactions.

```{r}

Retail\_Management %>%

group\_by(Country) %>%

tally(sort=TRUE) %>% summarise(Country, counts=n,percent= n/sum(n)\*100) %>% filter (percent > 1)

```

# 2. Create a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables. Add this variable to the dataframe.

```{r}

Retail\_Management<- mutate(Retail\_Management,TransactionValue = Quantity \* UnitPrice)

head(Retail\_Management[,9])

#3. Using the newly created variable, TransactionValue, show the breakdown of transaction values by countries i.e. how much money in total has been spent each country. Show this in total sum of transaction values. Show only countries with total transaction exceeding 130,000 British Pound.

Retail\_Management %>%

group\_by(Country) %>%

summarise(TransValuesum =sum (TransactionValue)) %>% filter(TransValuesum > 130000) %>% arrange(desc(TransValuesum))

#4. This is an optional question which carries additional marks (golden questions). In this question, we are dealing with the InvoiceDate variable. The variable is read as a categorical when you read data from the file. Now we need to explicitly instruct R to interpret this as a Date variable. "POSIXlt" and "POSIXct" are two powerful object classes in R to deal with date and time

Temp<- strptime(Retail\_Management$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')

head(Temp)

Retail\_Management$New\_Invoice\_Date<-as.Date(Temp)

Retail\_Management$Invoice\_Day\_week <- weekdays(Retail\_Management$New\_Invoice\_Date)

Retail\_Management$New\_Invoice\_Hour <-as.numeric (format(Temp,"%H"))

Retail\_Management$New\_Invoice\_Month <- as.numeric(format(Temp, "%m"))

head(Retail\_Management)

#a) Show the percentage of transactions (by numbers) by days of the week

Retail\_Management %>%

group\_by(Invoice\_Day\_week) %>%

tally(sort=TRUE) %>%

summarise(Invoice\_Day\_week,TransactionCounts = n ,percent = n/sum(n)\*100) %>%

arrange(desc(TransactionCounts))

#b) Show the percentage of transactions (by transaction volume) by days of the week

Retail\_Management %>%

group\_by(Invoice\_Day\_week) %>%

summarise(TransValueSum = sum(TransactionValue)) %>%

mutate(TransValuepercent= TransValueSum/sum(TransValueSum))%>%

arrange(desc(TransValueSum))

#c) Show the percentage of transactions (by transaction volume) by month of the year

Retail\_Management %>%

group\_by(New\_Invoice\_Month) %>%

summarise(TransValueSum = sum(TransactionValue)) %>%

mutate(TransValuePercent=TransValueSum/sum(TransValueSum)) %>%

arrange(desc(TransValuePercent))

#d) What was the date with the highest number of transactions from Australia?

Retail\_Management %>%

filter(Country == "Australia") %>%

group\_by(InvoiceDate) %>%

tally(sort = TRUE) %>%

filter(n == max(n))

# e) The company needs to shut down the website for two consecutive hours for maintenance. What would be the hour of the day to start this so that the distribution is at minimum for the customers? The responsible IT team is available from 7:00 to 20:00 every day.

Retail\_Management %>%

group\_by(New\_Invoice\_Hour) %>%

tally(sort=TRUE) %>%

filter(New\_Invoice\_Hour>=7 & New\_Invoice\_Hour<=20) %>%

arrange(n) %>%

head(5)

#5. Plot the histogram of transaction values from Germany. Use the hist() function to plot.

Retail\_Management %>%

group\_by(Country) %>%

filter(Country == "Germany") %>%

group\_by(New\_Invoice\_Month) %>%

summarise(TransValueSum = sum(TransactionValue)) -> Germany

hist(Germany$TransValueSum, border = "deepskyblue3", main = "Germany Transaction Value", xlab = "Transaction Value Sum per Month", ylab ="Frequency", col ="deepskyblue")

#6. Which customer had the highest number of transactions? Which customer is most valuable (i.e. highest total sum of transactions)?

Retail\_Management %>%

group\_by(CustomerID) %>%

tally(sort=TRUE) %>%

filter(!is.na(CustomerID)) %>%

filter(n==max(n))

Retail\_Management %>%

group\_by(CustomerID) %>%

summarise(TransValueSum = sum(TransactionValue)) %>%

filter(is.na(CustomerID)) %>%

filter(TransValueSum == max(TransValueSum))

# 7. Calculate the percentage of missing values for each variable in the dataset (5 marks). Hint colMeans():

colMeans(is.na(Retail\_Management))

#8. What are the number of transactions with missing CustomerID records by countries?

Retail\_Management %>%

filter(is.na(CustomerID)) %>%

group\_by(Country)%>%

summarise(CustomerID) %>%

tally(sort=TRUE)

#9. On average, how often the costumers comeback to the website for their next shopping?

Retail\_Management %>%

select(CustomerID, New\_Invoice\_Date) %>%

group\_by(CustomerID) %>%

distinct(New\_Invoice\_Date) %>%

arrange(desc(CustomerID)) %>%

mutate(DaysBetween = New\_Invoice\_Date - lag(New\_Invoice\_Date))->

custDaysBtwVisit

custDaysBtwVisit %>%

filter(!is.na(DaysBetween)) -> RetcustDaysBtwVisits

mean(RetcustDaysBtwVisits$DaysBetween)

# 10. In the retail sector, it is very important to understand the return rate of the goods purchased by customers. In this example, we can define this quantity, simply, as the ratio of the number of transactions cancelled (regardless of the transaction value) over the total number of transactions. With this definition, what is the return rate for the French customers? (10 marks). Consider the cancelled transactions as those where the ‘Quantity’ variable has a negative value.

Retail\_Management %>%

group\_by(Country) %>%

filter(Country=="France") %>%

select(Country,Quantity) %>%

filter(Quantity < 0 ) -> FrenchReturns

Retail\_Management %>%

group\_by(Country) %>%

filter(Country== "France") %>%

select(Quantity, Country) %>%

filter(Quantity > 0 ) ->FrenchPurchases

FRReturns<- sum(FrenchReturns$Quantity)

FRTransactions<-sum(FrenchPurchases$Quantity)

FRReturns/FRTransactions \*100

# 11. What is the product that has generated the highest revenue for the retailer? (i.e. item with the highest total sum of ‘TransactionValue’).

Retail\_Management %>%

group\_by(StockCode) %>%

summarise(Trans\_Value\_Tot=sum(TransactionValue)) %>%

arrange(desc(Trans\_Value\_Tot)) %>%

filter(StockCode !="Dot") %>%

filter(Trans\_Value\_Tot == max(Trans\_Value\_Tot))

# 12. How many unique customers are represented in the dataset? You can use unique() and length() functions.

Retail\_Management %>%

group\_by(CustomerID) %>%

distinct(CustomerID) -> UniqueCustomers

length(UniqueCustomers$CustomerID)

```