---

title: "BA Assignment"

author: "ISHIKA NISHA"

date: "10/25/2021"

output: html\_document

---

getwd()

#Reading the file

```{r}

getwd()

Assign <-read.csv("Online\_Retail (1).csv")

```

```{r}

summary(Assign)

```

#1 Showing 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). Showing this

in total number and also in percentage. Showing only countries accounting for more than 1% of the

total transactions.

```{r}

Assign %>%

group\_by(Country)%>%

summarise(trans =n())%>%

mutate(percent=(trans/541909)\*100)%>% #541909 is number of observation

filter(TT1 <- percent >1)

```

#2 Creation of a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and

‘UnitPrice’ variables

```{r}

Assign <- mutate(Assign, "TransactionValue"=TransactionValue <-Assign$Quantity \* Assign$UnitPrice)

```

#cross cheacking if the new variable is formed

```{r}

summarize(Assign)

colnames(Assign)

```

library(zoo)

#3 showing the breakdown of transaction values

by countries, total sum of transaction values and countries with total transaction exceeding 130,000 British Pound.

```{r}

Assign%>%

group\_by(Country)%>%

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

filter(TSTV>130000)

```

#5 histogram of Transaction values from Germany

```{r}

hist(x=(Assign$TransactionValue[Assign$Country=="Germany"]),xlab="TransactionValue",ylab="numbers")

```

#6 Customer with highest number of transaction is CustomerID =17841, CustomerTransaction 7983

```{r}

TRAN<- Assign %>%

group\_by(CustomerID)%>%

summarise(CustomerTransaction = n())%>%

filter(CustomerID != "NA")%>%

filter(CustomerTransaction ==max(CustomerTransaction) )

```

#Most Valuable or highest total sum of transactions is with CustomerID 14646, TTEC= 279489

```{r}

TRAN2<- Assign%>%

group\_by(CustomerID)%>%

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

arrange(TTEC)%>%

filter(CustomerID != "NA")%>%

filter(TTEC ==max(TTEC) )

```

#7Calculating the percentage of missing values for each variable in the dataset using colMeans()

```{r}

Missing\_Value<-colMeans(is.na(Assign))

```

```{r}

print(Missing\_Value)

```

#As only customerID has missing values So its value in percentage

`````{r}

print(Missing\_Value['CustomerID']\*100)

```

# 8 the number oftransactions with missing CustomerID recordsby countries is

```{r}

Assign%>%

group\_by(Country)%>%

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

summarise(NMCT=n())

```

10. The return rate for the French customers, considering all the factors is

```{r}

RRV<-nrow(Assign%>%

group\_by(CustomerID)%>%

filter((Country=='France')&(TransactionValue<0)&(CustomerID != 'Na')))

French\_CX<-nrow(Assign%>%

group\_by(CustomerID)%>%

filter((Country=='France')&(CustomerID != 'Na')))

```

#11.What is the product that has generated the highest revenue for the retailer? (i.e. item with the

highest total sum of ‘TransactionValue’). (10 marks)

```{r}

CC<-Assign%>%

group\_by(Description,StockCode)%>%

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

arrange(desc(n))

TT<- CC[CC['n']==max(CC['n']),]

print(TT$Description)# highest Revenue

print(TT$StockCode)#Stock Code

```

#12 The number of unique customers that are represented in the dataset using unique() and length() functions is 4373

```{r}

length(unique(Assign$CustomerID))

```

#### EXTRA POINTS

#4 #4This 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. Click here

for more information. First let’s convert ‘InvoiceDate’ into a POSIXlt object:

Now,let’s convert ‘InvoiceDate’ into a POSIXltobject:

```{r}

CNVT=strptime(Assign$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')

head(CNVT)

```

#requirements to solve

```{r}

Assign$NID<-as.Date(CNVT)

Assign$IWD<-weekdays(Assign$NID)

Assign$NIH<-as.numeric(format(CNVT,"%H"))

Assign$NIM<-as.numeric(format(CNVT,"%H"))

head(Assign)

```

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

```{r}

Assign%>%

group\_by(IWD)%>%

tally(sort=TRUE)%>%

summarise(IWD,TransactionNum=n,Percentage= n/sum(n)\*100)%>%

arrange(desc(TransactionNum))

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

Assign%>%

group\_by(NIH)%>%

summarise(Sum\_trans=sum(TransactionValue))%>%

mutate(Percentage\_trans=Sum\_trans\*100/sum(Sum\_trans))

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

Assign%>%

group\_by(NIM)%>%

summarise(Sum\_trans\_month=sum(TransactionValue))%>%

mutate(Percentage\_trans\_month=Sum\_trans\_month\*100/sum(Sum\_trans\_month))

#d What was the date with the highest number of transactions from Australia? (Date= 6/15/2011 13:37 n=139)

Assign%>%

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

group\_by(InvoiceDate)%>%

tally(sort=TRUE)%>%

filter(n== max(n))

#eThe 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.(NIH=7 n=383 is the minimum derived from the table)

Assign%>%

group\_by(NIH)%>%

tally(sort=TRUE)%>%

filter(NIH>=7 & NIH<=20)

```

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

is the average number of days between consecutive shopping) (Optional/Golden question: 18

additional marks!) Hint: 1. A close approximation is also acceptable and you may find diff()

function useful.

```{r}

CC<-Assign%>%

group\_by(CustomerID)%>%

summarise(DIC= diff(NID))%>%

filter(DIC>0)

print(mean(CC$DIC))

```