~~~~
---
title: "Assignment1_BA"
author: "Ram"
date: "10/22/2021"
output:
  word_document: default
  pdf_document: default
---



Reading the file

```{r}
Online_Retail <- read.csv("C:/Users/ramne/Downloads/Online_Retail.csv")
Online_Retail
```


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}
library(dplyr)
library(scales)
library(gapminder)
colnames(Online_Retail)
table(Online_Retail$Country)
OR<- summarise(group_by(Online_Retail,Country), count=n())
OR1<-as.data.frame(OR)
OR2<-select(OR1, count)
Percent<-(OR2/sum(OR2))*100
OR3<-cbind(OR1,Percent)
names(OR3)[3]<-"percentage"
filter(OR3, 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.

creating new variable which is a product of quantity and Unitprice

```{r}
TransactionValue<- Online_Retail$Quantity*Online_Retail$UnitPrice
TV <- cbind(Online_Retail, TransactionValue)
head(TV)
```

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.
~~~~

~~~
```{r}
library(gapminder)
OR4<-
(summarise(group_by(TV,Country),totalsum=sum(TransactionValue))%>%filter(
totalsum>13000))
OR5<- as.data.frame(OR4)
OR5
```
~~~

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. First let's convert 'InvoiceDate' into a POSIXlt object:
Temp=strptime(Online_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')

~~~
```{r}
Temp=strptime(Online_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',
tz='GMT')
Online_Retail$New_Invoice_Date<- as.Date(Temp)
Online_Retail$New_Invoice_Date[20000]- Online_Retail$New_Invoice_Date[10]
```
```{r}
Online_Retail$Invoice_Day_Week= weekdays(Online_Retail$New_Invoice_Date)
Online_Retail$Invoice_Day_Hour= as.numeric(format(Temp, '%H'))
Online_Retail$Invoice_Day_Month= as.numeric(format(Temp, '%H'))
```
~~~

a) Show the percentage of transactions (by numbers) by days of the week (extra 2 marks)

~~~
```{r}
Percentbyweek<-summarise(group_by(Online_Retail,
Online_Retail$Invoice_Day_Week), count=n())
Percentbyweekdata<- as.data.frame(Percentbyweek)
Percentbyweekdata$count/sum(Percentbyweekdata$count)*100
```
~~~

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

~~~
```{r}
tapply(Online_Retail$Quantity,Online_Retail$Invoice_Day_Week,sum)/sum(Onl
ine_Retail$Quantity)*100
```
~~~

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

~~~
```{r}
PercentagebyMonth<- summarise(group_by(Online_Retail,
Online_Retail$Invoice_Day_Month), Sum=sum(Quantity))
PercentagebyMonthdata<- as.data.frame(PercentagebyMonth)
PercentagebyMonth$Sum/sum(PercentagebyMonth$Sum)*100
```
~~~

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

~~~
```{r}
HighestNumber<- Online_Retail%>%filter(Country=='Australia')
HighestNumberdata<- summarise(group_by(HighestNumber,Country), high =
max(Quantity))
H<-as.data.frame(HighestNumberdata)
H
filter<- filter(HighestNumber, Quantity==1152)
select(filter,Invoice_Day_Week)
```
~~~

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.

~~~
```{r}
library(zoo)
Maintenance<- table(Online_Retail$Invoice_Day_Hour)
Maintenance
rollapply(Maintenance, 2, sum)
```
~~~

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

~~~
```{r}
library(ISLR)
TVGermany<-select(TV,8,9)%>%filter(Country=="Germany")
hist(log(TVGermany$TransactionValue),main = "Germany Transactions",
     xlab= "Transaction Values",col = "red")
```
~~~

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

~~~
```{r}
Cust.Val<- tapply(TV$TransactionValue, TV$CustomerID, length)
Cust.Val[which.max(Cust.Val)]

Cust.Value<-tapply(TV$TransactionValue, TV$CustomerID, sum)
Cust.Value[which.max(Cust.Value)]
```
~~~

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

~~~
```{r}
colMeans(is.na(Online_Retail))*100
```
~~~

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

~~~
```{r}
~~~

~~~~
Missing<- function(x){
  z<-sum(is.na(x))
  return(z)
}
tapply(TV$CustomerID,TV$Country,Missing)
```

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)
Hint: 1. A close approximation is also acceptable and you may find diff()
function useful.

```{r}
Comeback<- select(TV, c(5,7))


```



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?
 Consider the cancelled transactions as those where the 'Quantity'
variable has a negative value.
```{r}
ReturnRate<- TV%>%filter(Country=="France")
NegativeReturnRate<- filter(ReturnRate, Quantity<0)
nrow(NegativeReturnRate)/nrow(ReturnRate)*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').

```{r}
Revenue<- summarise(group_by(TV, Description),
highvalue=sum(TransactionValue))
HighestRevenue <- as.data.frame(Revenue)
HighestRevenue[which.max(HighestRevenue$highvalue),]
```

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

```{r}
length(unique(TV$CustomerID))
```
~~~~