Ba -Assignment

TILAK KUMAR BONALA

2022-10-31

```
Online_data = read.csv("C://Users//Hello//Desktop//BA//assignment3//Online_Retail.csv")
View(Online_data)
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

#Question 1 #Show the breakdown of the number of transactions by countries i.e. how many transactions are in the dataset for each country in total number and also in percentage. Show only countries acco unting for more than 1% of the total transactions.

```
Totaltransactions <- table(Online_data$Country)
transaction_percentage <- round(100*prop.table(Totaltransactions))
percentage <- cbind(Totaltransactions, transaction_percentage)
Transaction_table <- subset(percentage, transaction_percentage >1)
Transaction_table
```

```
##                Totaltransactions transaction_percentage
## EIRE                        8196                      2
## France                      8557                      2
## Germany                     9495                      2
## United Kingdom            495478                     91
```

```
#Question 2
#Creating  a  new  variable  'T_Value'   for value of the transaction to product  of existing  'Quantit
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
```

```
T_value <- Online_data$Quantity * Online_data$UnitPrice
Online_data <- Online_data %>% mutate(T_value)
summary(Online_data$T_value)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -168469.60       3.40       9.75      17.99      17.40  168469.60
```

```
#Question 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.
data <- summarise(group_by(Online_data, Country), sum1 = sum(T_value))
Transaction <- filter(data, sum1>130000)
Transaction
```

```
## # A tibble: 6 x 2
##   Country            sum1
##   <chr>             <dbl>
## 1 Australia       137077.
## 2 EIRE            263277.
## 3 France          197404.
## 4 Germany         221698.
## 5 Netherlands     284662.
## 6 United Kingdom 8187806.
```

```
#question 4
Time=strptime(Online_data$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
head(Time)
```

```
## [1] "2010-12-01 08:26:00 GMT" "2010-12-01 08:26:00 GMT"
## [3] "2010-12-01 08:26:00 GMT" "2010-12-01 08:26:00 GMT"
## [5] "2010-12-01 08:26:00 GMT" "2010-12-01 08:26:00 GMT"
```

```
Online_data$New_Invoice_Date <- as.Date(Time)
Online_data$New_Invoice_Date[20000]-Online_data$New_Invoice_Date[10]
```

```
## Time difference of 8 days
```

```
## Time difference of 8 days
Online_data$Invoice_Day_Week= weekdays(Online_data$New_Invoice_Date)
Online_data$New_Invoice_Hour = as.numeric(format(Time, "%H"))
Online_data$New_Invoice_Month = as.numeric(format(Time, "%m"))
```

```
#question 4 a) Percentage of transactions in days of the week
Table<-summarise(group_by(Online_data,Invoice_Day_Week),T_value=n_distinct(InvoiceNo))
Table1<-mutate(Table, transaction_percent=(T_value/sum(T_value))*100)
Table1
```

```
## # A tibble: 6 x 3
```

```
##   Invoice_Day_Week T_value transaction_percent
##   <chr>              <int>               <dbl>
## 1 Friday              4184               16.2
## 2 Monday              4138               16.0
## 3 Sunday              2381                9.19
## 4 Thursday            5660               21.9
## 5 Tuesday             4722               18.2
## 6 Wednesday           4815               18.6
```

```
#question 4 b: percentage  of volume of   transactions
Table_4b<-summarise(group_by(Online_data,Invoice_Day_Week),T_Volume=sum(T_value))
Table_4b1<-mutate(Table_4b,percentage=(T_Volume/sum(T_Volume))*100)
Table_4b1
```

```
## # A tibble: 6 x 3
##   Invoice_Day_Week T_Volume percentage
##   <chr>               <dbl>      <dbl>
## 1 Friday           1540611.      15.8
## 2 Monday           1588609.      16.3
## 3 Sunday            805679.       8.27
## 4 Thursday         2112519       21.7
## 5 Tuesday          1966183.      20.2
## 6 Wednesday        1734147.      17.8
```

```
# Question 4 c) Show the percentage of volume of transactions in  month of the year
Table_c1<-summarise(group_by(Online_data,New_Invoice_Month),T_Volume=sum(T_value))
Table_c2<-mutate(Table_c1,percentage=(T_Volume/sum(T_Volume))*100)
Table_c2
```

```
## # A tibble: 12 x 3
##    New_Invoice_Month T_Volume percentage
##                <dbl>    <dbl>      <dbl>
##  1                 1  560000.       5.74
##  2                 2  498063.       5.11
##  3                 3  683267.       7.01
##  4                 4  493207.       5.06
##  5                 5  723334.       7.42
##  6                 6  691123.       7.09
##  7                 7  681300.       6.99
##  8                 8  682681.       7.00
##  9                 9 1019688.      10.5
## 10                10 1070705.      11.0
## 11                11 1461756.      15.0
## 12                12 1182625.      12.1
```

```
# Question 4 d) What was the date with the highest number of transactions from Australia?

Online_data <- Online_data %>% mutate(T_value= Quantity * UnitPrice)
Online_data %>% filter(Country == 'Australia') %>% group_by(New_Invoice_Date) %>% summarise(max=max(T_va
```

```
## # A tibble: 49 x 2
##    New_Invoice_Date     max
```

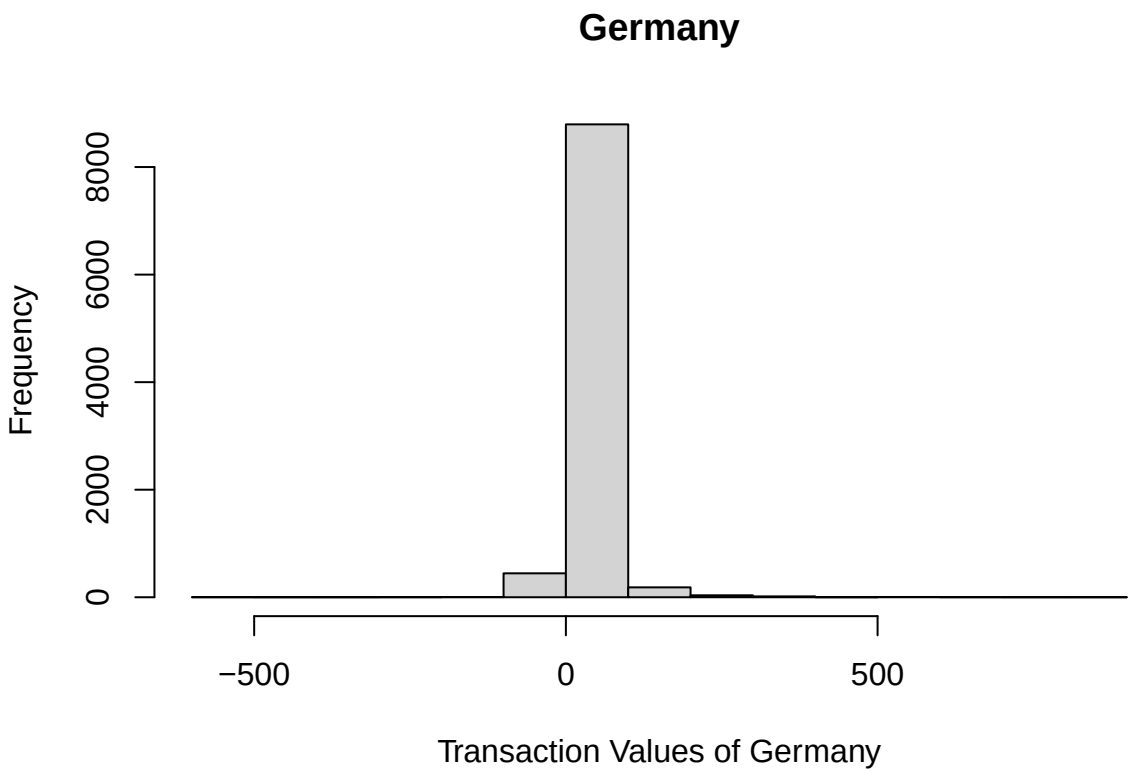
```
##    <date>              <dbl>
##  1 2010-12-01          51
##  2 2010-12-08          71.4
##  3 2010-12-14          -6.25
##  4 2010-12-17         148.
##  5 2011-01-06        1020
##  6 2011-01-10          81.6
##  7 2011-01-11          35.4
##  8 2011-01-14         142.
##  9 2011-01-17          47.4
## 10 2011-01-19          38.2
## # ... with 39 more rows
```

```
#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 custome

data_e1<-summarise(group_by(Online_data,New_Invoice_Hour),Transaction_min=n_distinct(InvoiceNo))
data_e1<-filter(data_e1,New_Invoice_Hour>=7&New_Invoice_Hour<=20)
data_e2<-rollapply(data_e1$Transaction_min,3,sum)
data_e3<-which.min(data_e2)
data_e3
```

```
## [1] 12
```

```
#Question5
#Ploting the histogram of transaction values from Germany by using the hist() function to plot.
Germany_data <- subset(Online_data$T_value, Online_data$Country == "Germany")
hist(Germany_data, xlab = "Transaction Values of Germany", main = "Germany")
```




```
#Question 6
#Finding Which customer had the highest number of transactions? Which customer is most valuable (i.e. h
Online_data1 <- na.omit(Online_data)
result_data1 <- summarise(group_by(Online_data1,CustomerID), sum2= sum(T_value))
result_data1[which.max(result_data1$sum2),]
```

```
## # A tibble: 1 x 2
##   CustomerID    sum2
##        <int>   <dbl>
## 1      14646 279489.
```

```
data_2 <- table(Online_data$CustomerID)
data_2 <- as.data.frame(data_2)
result_data2 <- data_2[which.max(data_2$Freq),]
result_data2
```

```
##       Var1 Freq
## 4043 17841 7983
```

```
#Question 7
#Calculate the percentage of missing values for each variable in the dataset
missing_values_data <- colMeans(is.na(Online_data)*100)
missing_values_data
```

```
##         InvoiceNo         StockCode       Description          Quantity
##           0.00000           0.00000           0.00000           0.00000
##       InvoiceDate         UnitPrice        CustomerID           Country
##           0.00000           0.00000          24.92669           0.00000
##           T_value  New_Invoice_Date  Invoice_Day_Week  New_Invoice_Hour
##           0.00000           0.00000           0.00000           0.00000
## New_Invoice_Month
##           0.00000
```

```
#Question 8
#The number of transactions with missing CustomerID records by countries
Online_data2 <- Online_data %>% filter(is.na(CustomerID)) %>% group_by(Country)
summary(Online_data2$Country)
```

```
##    Length     Class      Mode
##    135080 character character
```

```
#Question 9
#On average the costumers comeback to the website for their next shopping? #(i.e. what is  the  average
Online_NA_removed <- na.omit(Online_data)
Online_NA_removed <- subset(Online_NA_removed, Quantity > 0)
Online_subset <- Online_NA_removed[,c("CustomerID","New_Invoice_Date")]
Online_subset_distinct <- distinct(Online_subset)
Online_subset_distinct %>%
group_by(CustomerID) %>%
arrange(New_Invoice_Date)%>%
summarise(avg= mean(diff(New_Invoice_Date))) %>%
na.omit()%>%
summarise(avg_days_between_shopping = mean(avg))
```

```
## # A tibble: 1 x 1
##   avg_days_between_shopping
##   <drtn>
## 1 78.42025 days
```

#The customers come back after on an average of 78 days after shopping.

```
#Question 10
#what is the return rate for the French customers?
#A) Consider the cancelled transactions as those where the 'Quantity'
#variable has a negative value.
Online_table <- filter(Online_data,Country=="France")
total_row <- nrow(Online_table)
cancelled_transactions <- nrow(subset(Online_table,T_value<0))
cancelled_transactions
```

```
## [1] 149
```

```
## [1] 149
notcancel <- total_row- cancelled_transactions
notcancel
```

```
## [1] 8408
```

```
## [1] 8408
Q_10 =(cancelled_transactions/8408)*100
Q_10
```

```
## [1] 1.772122
```

```
#Question11
# The product that has generated the highest revenue for the retailer? .
T_value <- tapply(Online_data$T_value, Online_data$StockCode , sum)
T_value[which.max(T_value)]
```

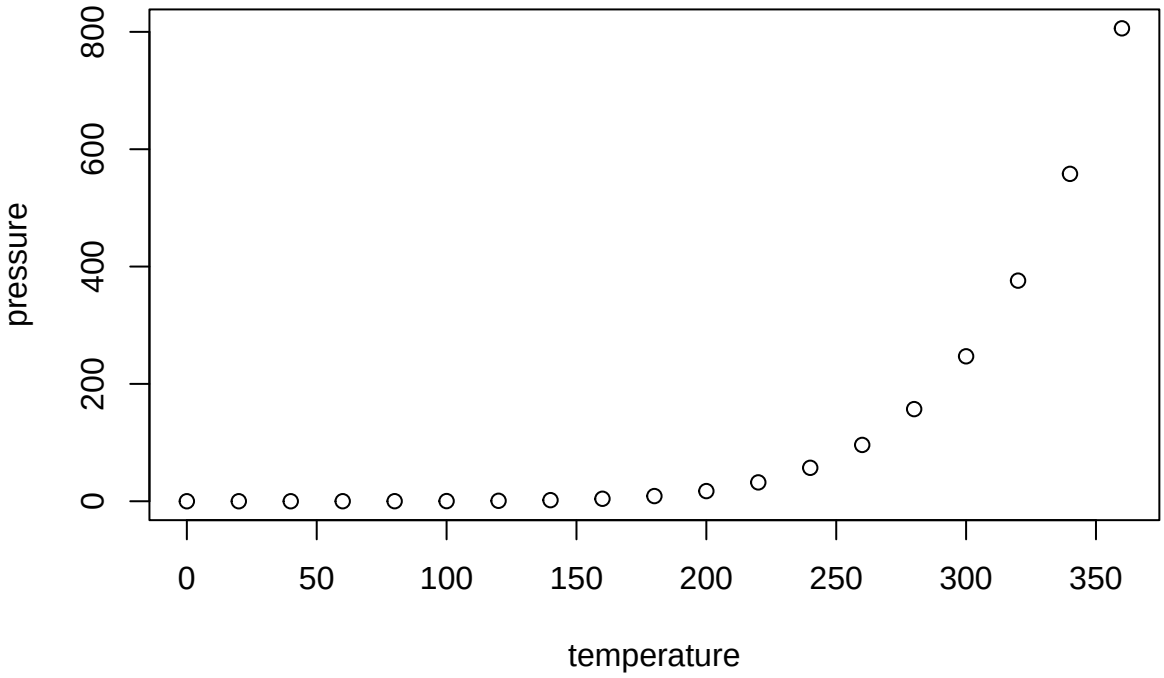
```
##      DOT
## 206245.5
```

```
#Question12
#unique customers are represented by using unique() and length() functions.
unique_customers <- unique(Online_data$CustomerID)
length(unique_customers)
```

```
## [1] 4373
```

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```


pressure
0
200
400
600
800
0
50
100
150
200
250
300
350
temperature