Business Analytics Assignment-2

LOKESH JETANGI

2023-03-12

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
Online_Retail<- read.csv("C:/Users/jetan/Downloads/Online_Retail.csv")
str(Online_Retail)
```

```
## 'data.frame':    541909 obs. of  8 variables:
##  $ InvoiceNo  : chr  "536365" "536365" "536365" "536365" ...
##  $ StockCode  : chr  "85123A" "71053" "84406B" "84029G" ...
##  $ Description: chr  "WHITE HANGING HEART T-LIGHT HOLDER" "WHITE METAL LAN
TERN" "CREAM CUPID HEARTS COAT HANGER" "KNITTED UNION FLAG HOT WATER BOTTLE"
...
##  $ Quantity   : int  6 6 8 6 6 2 6 6 6 32 ...
##  $ InvoiceDate: chr  "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:26" "1
2/1/2010 8:26" ...
##  $ UnitPrice  : num  2.55 3.39 2.75 3.39 3.39 7.65 4.25 1.85 1.85 1.69 ...
##  $ CustomerID : int  17850 17850 17850 17850 17850 17850 17850 17850 17850
13047 ...
##  $ Country    : chr  "United Kingdom" "United Kingdom" "United Kingdom" "U
nited Kingdom" ...
```

```
#Task-1
country_totaltran <-table(Online_Retail$Country)
transaction_percent<- round(100*prop.table(country_totaltran))
percentage <- cbind(country_totaltran, transaction_percent)
solution <-subset(percentage, transaction_percent >1)
solution
```

```
##                country_totaltran transaction_percent
## EIRE                        8196                   2
## France                      8557                   2
## Germany                     9495                   2
## United Kingdom            495478                  91
```

```r
#Task-2
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

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

Transactionvalue <-Online_Retail$Quantity * Online_Retail$UnitPrice
Online_Retail<- Online_Retail%>% mutate(Transactionvalue)
summary(Online_Retail$Transactionvalue)

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

summary(Online_Retail)

##   InvoiceNo          StockCode         Description           Quantity
##  Length:541909      Length:541909      Length:541909      Min.   :-80995.0
0
##  Class :character   Class :character   Class :character   1st Qu.:     1.0
0
##  Mode  :character   Mode  :character   Mode  :character   Median :     3.0
0
##                                                           Mean   :     9.5
5
##                                                           3rd Qu.:    10.0
0
##                                                           Max.   : 80995.0
0
##
##  InvoiceDate          UnitPrice           CustomerID       Country
##  Length:541909      Min.   :-11062.06   Min.   :12346    Length:541909
##  Class :character   1st Qu.:     1.25   1st Qu.:13953    Class :character
##  Mode  :character   Median :     2.08   Median :15152    Mode  :character
##                     Mean   :     4.61   Mean   :15288
##                     3rd Qu.:     4.13   3rd Qu.:16791
##                     Max.   : 38970.00   Max.   :18287
##                                         NA's   :135080
##  Transactionvalue
```

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

```
#Task-3
country_transaction_values <- Online_Retail %>%
group_by(Country) %>%summarise(Total_Transaction_Value = sum(Transactionvalue
))
country_transaction_values_above_130k <-subset(country_transaction_values, To
tal_Transaction_Value > 130000)
country_transaction_values_above_130k
```

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

```
#Task-4
# Creates temporary variable that formats transaction date into mm/dd/yyyy fo
rmat
Temp=strptime(Online_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
head(Temp)
```

```
## [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"
```

```
# Formats the New_Invoice_Date column into a date format from the Temp variab
le
Online_Retail$New_Invoice_Date <- as.Date(Temp)

# Example of how dates can be subtracted from each other and return the diffe
rence in values
Online_Retail$New_Invoice_Date[20000]- Online_Retail$New_Invoice_Date[10]
```

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

```
# Convert dates to days of the week and assigns column title to Invoice_Day_W
eek
Online_Retail$Invoice_Day_Week= weekdays(Online_Retail$New_Invoice_Date)

# Create a new column with the transaction hour assigned to New_Invoice_Hour
```

```
Online_Retail$New_Invoice_Hour = as.numeric(format(Temp, "%H"))

# Create a new column with the transaction month assigned to New_Invoice_Mont
h
Online_Retail$New_Invoice_Month = as.numeric(format(Temp, "%m"))

#(4a)Show the percentage of transactions (by numbers) by days of the week

Online_Retail%>% group_by(Invoice_Day_Week)%>% summarise(Number.of.transactio
n=(n()))%>% mutate(Number.of.transaction,'percent'=(Number.of.transaction*100
)/sum(Number.of.transaction))
```

```
## # A tibble: 6 × 3
##   Invoice_Day_Week Number.of.transaction percent
##   <chr>                            <int>   <dbl>
## 1 Friday                           82193    15.2
## 2 Monday                           95111    17.6
## 3 Sunday                           64375    11.9
## 4 Thursday                        103857    19.2
## 5 Tuesday                         101808    18.8
## 6 Wednesday                        94565    17.5
```

```
#(4b) Show the percentage of transactions (by transaction volume) by days of
the week

Online_Retail%>%
  group_by(Invoice_Day_Week)%>% summarise(Volume.of.transaction=(sum(Transact
ionvalue)))%>% mutate(Volume.of.transaction,'percent'=(Volume.of.transaction*
100)/sum(Volume.of.transaction))
```

```
## # A tibble: 6 × 3
##   Invoice_Day_Week Volume.of.transaction percent
##   <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
```

```
#(4c) Show the percentage of transactions (by transaction volume) by month of
the year

Online_Retail%>% group_by(New_Invoice_Month)%>% summarise(Volume.By.Month=sum
(Transactionvalue))%>% mutate(Volume.By.Month,'Percent'=(Volume.By.Month*100)
/sum(Volume.By.Month))
```

```
## # A tibble: 12 × 3
##    New_Invoice_Month Volume.By.Month Percent
##                <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
```

```
#(4d) What was the date with the highest number of transactions from Australi
a
```

```
subset(Online_Retail, Country == "Australia") %>% group_by(New_Invoice_Date)
%>% summarise(n_transactions = n()) %>% top_n(3)
```

```
## Selecting by n_transactions
```

```
## # A tibble: 3 × 2
##   New_Invoice_Date n_transactions
##   <date>                    <int>
## 1 2011-06-15                  139
## 2 2011-07-19                  137
## 3 2011-08-18                   97
```

```
#(4e) The company needs to shut down the website for two consecutive hours fo
r maintenance. What would be the hour of the day to start this so that the di
stribution is at minimum for the customers? The responsible IT team is availa
ble from 7:00 to 20:00 every day.
```

```
Online_Retail %>% group_by(New_Invoice_Hour) %>%
  summarise(percent_of_transactions = 100*(n()/nrow(Online_Retail))) %>% arra
nge(percent_of_transactions)
```

```
## # A tibble: 15 × 2
##    New_Invoice_Hour percent_of_transactions
##               <dbl>                   <dbl>
##  1                6                 0.00757
##  2                7                 0.0707
##  3               20                 0.161
##  4               19                 0.684
##  5               18                 1.47
##  6                8                 1.64
##  7               17                 5.26
##  8                9                 6.34
##  9               10                 9.05
## 10               16                10.1
## 11               11                10.6
## 12               14                12.5
```

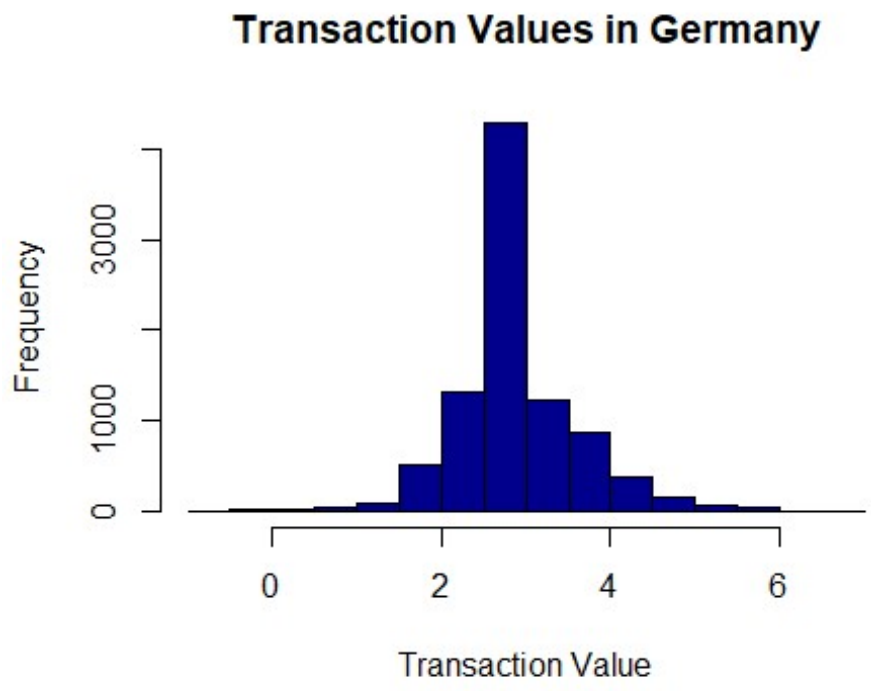
```
## 13                13                13.3
## 14                15                14.3
## 15                12                14.5
```

```
#Task-5
#Plot the histogram of transaction values from Germany. Use the hist() functi
on to plot.

# Filter the data for transactions made in Germany
germany_data <- filter(Online_Retail, Country == "Germany")

# Plot the histogram of transaction values from Germany
hist(x=log(Online_Retail$Transactionvalue[Online_Retail$Country=="Germany"]),
     xlab = "Transaction Value",
     col = 'dark blue' ,
     main = 'Transaction Values in Germany',
     ylab = 'Frequency')
```

```
## Warning in log(Online_Retail$Transactionvalue[Online_Retail$Country ==
## "Germany"]): NaNs produced
```




```
#Task-6
# Assumption 1: Considering the no. of transactions to calculate highest No.
of transactions(valuable customer)

TransactionwithNA<-Online_Retail%>%  group_by(CustomerID) %>%
  summarise(Highest_no_of_Trans_with_NAValues=n()) %>% arrange(desc(Highest_n
```

```
o_of_Trans_with_NAValues)) %>%
  top_n(3)

## Selecting by Highest_no_of_Trans_with_NAValues

## Selecting by Highest_no_of_Trans_with_NAValues
as.data.frame(TransactionwithNA)

##   CustomerID Highest_no_of_Trans_with_NAValues
## 1         NA                            135080
## 2      17841                              7983
## 3      14911                              5903

# Assumption 2 : Omitted NA Values and checked for the valuable customer

TransactionwithoutNA<-Online_Retail%>% na.omit() %>%
  group_by(CustomerID) %>% summarise(Highest_no_of_Trans=n()) %>% arrange(des
c(Highest_no_of_Trans)) %>%
  top_n(1)

## Selecting by Highest_no_of_Trans

## Selecting by Highest_no_of_Trans
as.data.frame(TransactionwithoutNA)

##   CustomerID Highest_no_of_Trans
## 1      17841                7983

# Assumption 3: Considering the total sum of transactions(Transaction Volume)
to calculate valuable cutomer
TransvolwithNA<-Online_Retail%>% group_by(CustomerID) %>%
  summarise(Highest_Trans_Volume_with_NAValues=sum(Transactionvalue)) %>%
  arrange(desc(Highest_Trans_Volume_with_NAValues)) %>% top_n(3)

## Selecting by Highest_Trans_Volume_with_NAValues

## Selecting by Highest_Trans_Volume_with_NAValues
as.data.frame(TransvolwithNA)

##   CustomerID Highest_Trans_Volume_with_NAValues
## 1         NA                           1447682.1
## 2      14646                            279489.0
## 3      18102                            256438.5

# Assumption 4: Omitted NA Values and checked for the valuable customer
TransvolwithoutNA <-Online_Retail%>% na.omit() %>% group_by(CustomerID) %>%
  summarise(Highest_Trans_Volume=sum(Transactionvalue)) %>% arrange(desc(High
est_Trans_Volume)) %>% top_n(1)

## Selecting by Highest_Trans_Volume

## Selecting by Highest_Trans_Volume
as.data.frame(TransvolwithoutNA)
```

```
##   CustomerID Highest_Trans_Volume
## 1      14646               279489

## customer 14646 had the highest number of transactions i.e., 279489

#Task-7

# Calculate the percentage of missing values for each variable
missing_percent <- colMeans(is.na(Online_Retail)) * 100
missing_percent

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

#The output data frame shows that CustomerID column has 24.92% of missing val
ues.

#Task-8
missing_transactions_by_country <- Online_Retail %>%
  filter(is.na(CustomerID)) %>%
  group_by(Country) %>%
  summarise(missing_transactions = n())
missing_transactions_by_country

## # A tibble: 9 × 2
##   Country        missing_transactions
##   <chr>                         <int>
## 1 Bahrain                           2
## 2 EIRE                            711
## 3 France                           66
## 4 Hong Kong                       288
## 5 Israel                           47
## 6 Portugal                         39
## 7 Switzerland                     125
## 8 United Kingdom                133600
## 9 Unspecified                      202

#Task-9
# Creating a new data frame with all "NA" CustomerIDs removed.

retail_Retail_NA_Removed <- na.omit(Online_Retail)

# Creating a new data frame with cancelled transactions removed.

retail_Retail_NA_Neg_Removed <- subset(retail_Retail_NA_Removed, Quantity > 0
```

```
)

# Creating a new dataframe that only has CustomerID and transaction date

retail_Retail_Subset <- retail_Retail_NA_Neg_Removed[,c("CustomerID","New_Inv
oice_Date")]

# Convert the transaction date to a numeric data type.
retail_Retail_Subset$New_Invoice_Date <- as.numeric(retail_Retail_Subset$New_
Invoice_Date)

# Creating a new data frame to remove multiple invoices from same customer on
same day

retail_Retail_Subset_Distinct <- distinct(retail_Retail_Subset)

# Groups data set by Customer ID, arranges them by date, and finds the averag
e time between consecutive transactions for each customer

# Removes CustomerIDs that result in an NA value (i.e. only have one distinct
transaction)

retail_Retail_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 × 1
##   avg_days_between_shopping
##                       <dbl>
## 1                      78.4

#Task-10

France_Transaction <- Online_Retail%>% select(Quantity,Country) %>% filter (C
ountry == "France") # French transaction count
Length_French_Orders <- length(France_Transaction$Quantity)
 #If the quantity value is less than 0, then we can consider it as a cancelle
d transaction
Cancelled_Transactions <- Online_Retail%>% select(Quantity,Country) %>% filte
r (Country == "France",Quantity<0)
French_Cancelled <-length(Cancelled_Transactions$Quantity)
#We perform cancelled order divided by total orders for France
Percentage_France <-  French_Cancelled / Length_French_Orders
Percentage_France

## [1] 0.01741264
```

```
data.frame(Length_French_Orders,French_Cancelled,Percentage_France)
```

```
##   Length_French_Orders French_Cancelled Percentage_France
## 1                 8557              149        0.01741264
```

```
#Task-11
Transactionvalue <- tapply(Online_Retail$Transactionvalue, Online_Retail$Stoc
kCode, sum)
Transactionvalue[which.max(Transactionvalue)]
```

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

```
#Task-12
length(unique(Online_Retail$CustomerID))
```

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