Assignment – Online Retail Analytics

Venkata Naga Siddartha Gutha

2022-10-30

```
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
```

```
# importing the data
data<-read.csv("C:/Users/sidda/Downloads/Online_Retail.csv")
head(data)
```

```
##   InvoiceNo StockCode                         Description Quantity
## 1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6
## 2    536365     71053                 WHITE METAL LANTERN        6
## 3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8
## 4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6
## 5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6
## 6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2
##      InvoiceDate UnitPrice CustomerID        Country
## 1 12/1/2010 8:26      2.55      17850 United Kingdom
## 2 12/1/2010 8:26      3.39      17850 United Kingdom
## 3 12/1/2010 8:26      2.75      17850 United Kingdom
## 4 12/1/2010 8:26      3.39      17850 United Kingdom
## 5 12/1/2010 8:26      3.39      17850 United Kingdom
## 6 12/1/2010 8:26      7.65      17850 United Kingdom
```

```
#Descriptive statistics
summary(data)
```

```
##   InvoiceNo          StockCode         Description          Quantity
##  Length:541909      Length:541909      Length:541909      Min.   :-80995.00
##  Class :character   Class :character   Class :character   1st Qu.:     1.00
##  Mode  :character   Mode  :character   Mode  :character   Median :     3.00
##                                                            Mean   :     9.55
##                                                            3rd Qu.:    10.00
##                                                            Max.   : 80995.00
##
##   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
```

Question 1

```
#Total number of transactions by each country with more than 1% transactions
Country_transactions<-data%>%group_by(Country)%>%
  summarise(number_of_transactions=n(),percentage=100*(n()/nrow(data)))   %>%filter(percentage>0.1)%>%a

Country_transactions
```

```
## # A tibble: 15 x 3
##    Country         number_of_transactions percentage
##    <chr>                            <int>      <dbl>
##  1 United Kingdom                  495478     91.4
##  2 Germany                           9495      1.75
##  3 France                            8557      1.58
##  4 EIRE                              8196      1.51
##  5 Spain                             2533      0.467
##  6 Netherlands                       2371      0.438
##  7 Belgium                           2069      0.382
##  8 Switzerland                       2002      0.369
##  9 Portugal                          1519      0.280
## 10 Australia                         1259      0.232
## 11 Norway                            1086      0.200
## 12 Italy                              803      0.148
## 13 Channel Islands                    758      0.140
## 14 Finland                            695      0.128
## 15 Cyprus                             622      0.115
```

Question 2 Adding a new variable 'Transaction Value' to the dataframe

```
# Adding new variable Transaction value to dataframe
data<-data%>%mutate(Transaction_value=Quantity*UnitPrice)
head(data)
```

```
##   InvoiceNo StockCode                         Description Quantity
## 1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6
## 2    536365     71053                 WHITE METAL LANTERN        6
## 3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8
## 4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6
## 5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6
## 6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2
##      InvoiceDate UnitPrice CustomerID        Country Transaction_value
## 1 12/1/2010 8:26      2.55      17850 United Kingdom             15.30
## 2 12/1/2010 8:26      3.39      17850 United Kingdom             20.34
## 3 12/1/2010 8:26      2.75      17850 United Kingdom             22.00
## 4 12/1/2010 8:26      3.39      17850 United Kingdom             20.34
## 5 12/1/2010 8:26      3.39      17850 United Kingdom             20.34
## 6 12/1/2010 8:26      7.65      17850 United Kingdom             15.30
```

Question 3 The breakdown of transaction values by countries with total transaction exceeding 130,000 British Pound.

```
Total_transaction_country<-data%>%group_by(Country)%>%
  summarise(Total_sum_of_Transactions=sum(Transaction_value))%>%
  filter(Total_sum_of_Transactions>13000)
Total_transaction_country
```

```
## # A tibble: 17 x 2
##    Country        Total_sum_of_Transactions
##    <chr>                              <dbl>
##  1 Australia                        137077.
##  2 Belgium                           40911.
##  3 Channel Islands                   20086.
##  4 Denmark                           18768.
##  5 EIRE                             263277.
##  6 Finland                           22327.
##  7 France                           197404.
##  8 Germany                          221698.
##  9 Italy                             16891.
## 10 Japan                             35341.
## 11 Netherlands                      284662.
## 12 Norway                            35163.
## 13 Portugal                          29367.
## 14 Spain                             54775.
## 15 Sweden                            36596.
## 16 Switzerland                       56385.
## 17 United Kingdom                  8187806.
```

Question 4

```
Temp=strptime(data$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"
```

```
#let's  separate  date,  day  of  the  week  and  hour  components dataframe with names as New_Invoice_
data$New_Invoice_Date <- as.Date(Temp)
# the difference between the two dates in terms of the number days
data$New_Invoice_Date[20000]- data$New_Invoice_Date[10]
```

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

```
#Converting dates to days
data$Invoice_Day_Week=weekdays(data$New_Invoice_Date)
#converting hour into numeric value
data$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
#converting month into numeric value
data$New_Invoice_Month = as.numeric(format(Temp, "%m"))
head(data)
```

```
##   InvoiceNo StockCode                         Description Quantity
## 1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6
## 2    536365     71053                 WHITE METAL LANTERN        6
## 3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8
## 4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6
## 5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6
## 6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2
##      InvoiceDate UnitPrice CustomerID        Country Transaction_value
## 1 12/1/2010 8:26      2.55      17850 United Kingdom             15.30
## 2 12/1/2010 8:26      3.39      17850 United Kingdom             20.34
## 3 12/1/2010 8:26      2.75      17850 United Kingdom             22.00
## 4 12/1/2010 8:26      3.39      17850 United Kingdom             20.34
## 5 12/1/2010 8:26      3.39      17850 United Kingdom             20.34
## 6 12/1/2010 8:26      7.65      17850 United Kingdom             15.30
##   New_Invoice_Date Invoice_Day_Week New_Invoice_Hour New_Invoice_Month
## 1       2010-12-01        Wednesday                8                12
## 2       2010-12-01        Wednesday                8                12
## 3       2010-12-01        Wednesday                8                12
## 4       2010-12-01        Wednesday                8                12
## 5       2010-12-01        Wednesday                8                12
## 6       2010-12-01        Wednesday                8                12
```

a) Percentage of transactions (by numbers) by days of the week

```
transactions_per_days_of_week<-data %>% group_by(Invoice_Day_Week) %>%
  summarise(Percent_of_transactions_per_days_of_week = 100*(n()/nrow(data)))

transactions_per_days_of_week
```

```
## # A tibble: 6 x 2
##   Invoice_Day_Week Percent_of_transactions_per_days_of_week
##   <chr>                                               <dbl>
## 1 Friday                                               15.2
## 2 Monday                                               17.6
## 3 Sunday                                               11.9
## 4 Thursday                                             19.2
## 5 Tuesday                                              18.8
## 6 Wednesday                                            17.5
```

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

```
Transactions_Volume_by_week<-data %>% group_by(Invoice_Day_Week) %>%
  summarise(Percent_of_Transactions_Volume_by_week=100*(sum(Transaction_value)/sum(data$Transaction_valu
Transactions_Volume_by_week
```

```
## # A tibble: 6 x 2
##   Invoice_Day_Week Percent_of_Transactions_Volume_by_week
##   <chr>                                             <dbl>
## 1 Friday                                            15.8
## 2 Monday                                            16.3
## 3 Sunday                                             8.27
## 4 Thursday                                          21.7
## 5 Tuesday                                           20.2
## 6 Wednesday                                         17.8
```

c) Percentage of transactions (by transaction volume) by month of the year

```
Percentage_Transactions_by_Month<-data %>% group_by(New_Invoice_Month) %>%
  summarise(Percentage_Transactions_by_Month=100*(sum(Transaction_value)/sum(data$Transaction_value)))
Percentage_Transactions_by_Month
```

```
## # A tibble: 12 x 2
##    New_Invoice_Month Percentage_Transactions_by_Month
##                <dbl>                            <dbl>
##  1                 1                             5.74
##  2                 2                             5.11
##  3                 3                             7.01
##  4                 4                             5.06
##  5                 5                             7.42
##  6                 6                             7.09
##  7                 7                             6.99
##  8                 8                             7.00
##  9                 9                            10.5
## 10                10                            11.0
## 11                11                            15.0
## 12                12                            12.1
```

d) Date with the highest number of transactions from Australia

```
Aus<-filter(data,Country=="Australia") %>% group_by(InvoiceDate) %>%
  summarise(Australia_highest_no_transactions=n())
Aus[which.max(Aus$Australia_highest_no_transactions),]
```

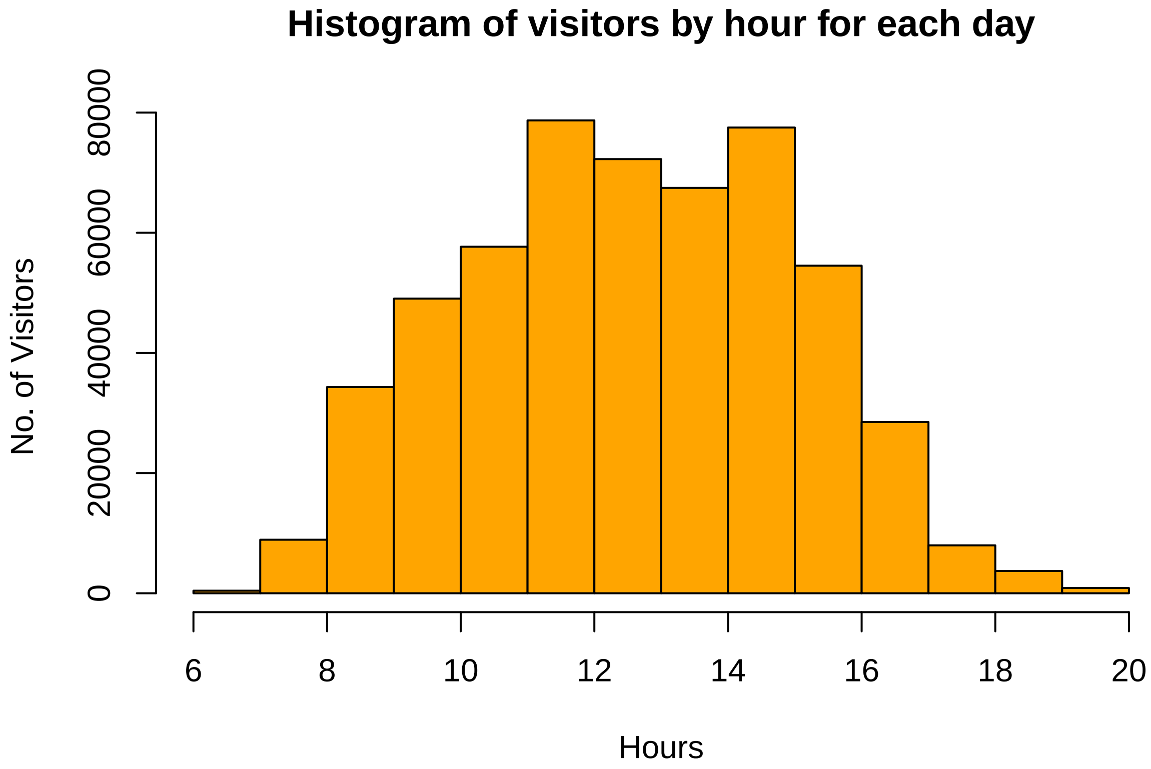
```
## # A tibble: 1 x 2
##   InvoiceDate     Australia_highest_no_transactions
##   <chr>                                       <int>
## 1 6/15/2011 13:37                               139
```

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

```
distribution<-data %>% group_by(New_Invoice_Hour)%>%
  summarise(No_Of_Transactions=n(),Percentage=100*(n()/nrow(data))) %>%
  filter(New_Invoice_Hour >=7 & New_Invoice_Hour <= 20)
distribution
```

```
## # A tibble: 14 x 3
##    New_Invoice_Hour No_Of_Transactions Percentage
##               <dbl>              <int>      <dbl>
##  1                7                383     0.0707
##  2                8               8909     1.64
##  3                9              34332     6.34
##  4               10              49037     9.05
##  5               11              57674    10.6
##  6               12              78709    14.5
##  7               13              72259    13.3
##  8               14              67471    12.5
##  9               15              77519    14.3
## 10               16              54516    10.1
## 11               17              28509     5.26
## 12               18               7974     1.47
## 13               19               3705     0.684
## 14               20                871     0.161
```

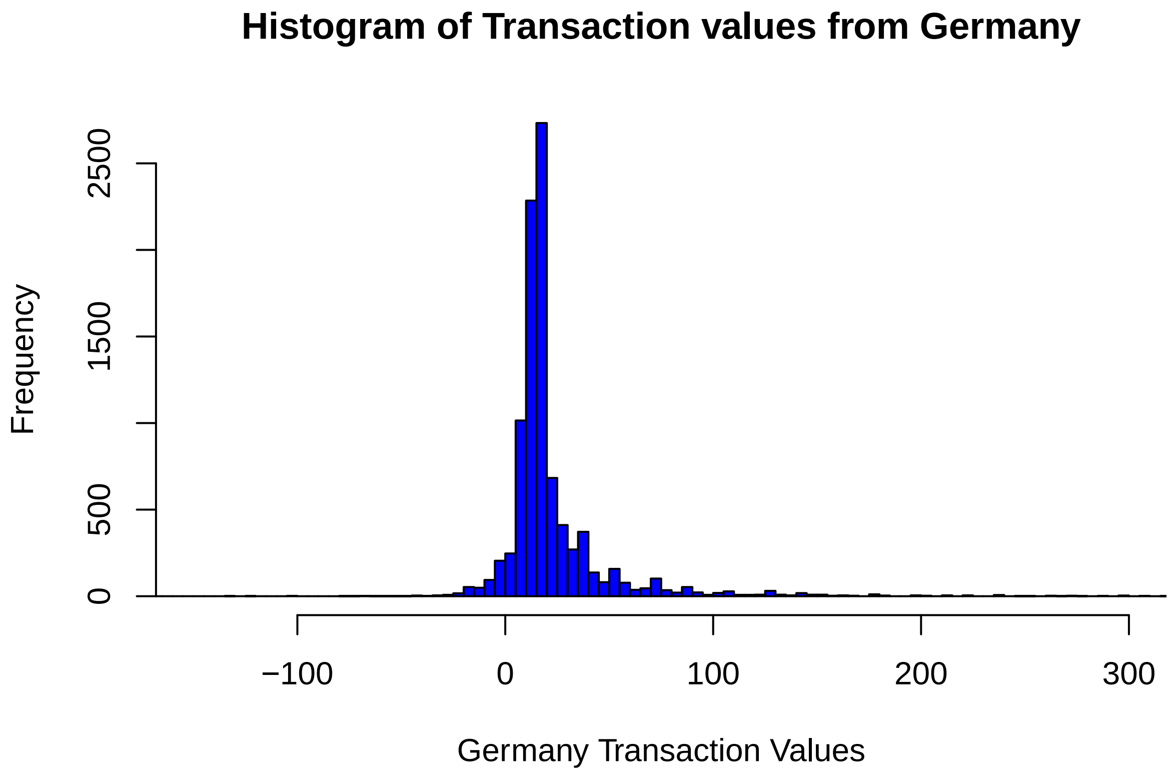
```
#Plotting a graph to show the website visitors for transactions per hour
hist(data$New_Invoice_Hour,
     main="Histogram of visitors by hour for each day",
     col = "orange",
     xlab = "Hours",
     ylab= "No. of Visitors",
     breaks = 12
     )
```



It can be seen from the graph that the best time for maintenance shutdown would be 6:00 am and 20:00 pm and it is also mentioned that responsible IT team is available from 7:00 to 20:00 every day.

Question 5 Plotting the histogram of transaction values from Germany.

```
Transactions_Germany<-filter(data,Country=="Germany")
hist(Transactions_Germany$Transaction_value,
     main = "Histogram of Transaction values from Germany",
     col = 'Blue',
     xlab = "Germany Transaction Values",
     ylab="Frequency",
     xlim = c(-150,300),
     breaks=500)
```



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

```
# highest No. of transactions(valuable customer)
Customer_high_transactions_withNA<-data %>%  group_by(CustomerID) %>%
  summarise(Highest_no_of_Trans_with_NAValues=n()) %>% arrange(desc(Highest_no_of_Trans_with_NAValues))
  top_n(1)
```

```
## Selecting by Highest_no_of_Trans_with_NAValues
```

```
Customer_high_transactions_withNA
```

```
## # A tibble: 1 x 2
##   CustomerID Highest_no_of_Trans_with_NAValues
##        <int>                             <int>
## 1         NA                            135080
```

```
# highest No. of transactions without NA
Customer_high_transactions_without_NA<-data %>% na.omit() %>%
  group_by(CustomerID) %>% summarise(Highest_no_of_Trans=n()) %>% arrange(desc(Highest_no_of_Trans)) %>
  top_n(1)
```

```
## Selecting by Highest_no_of_Trans
```

```
Customer_high_transactions_without_NA
```

```
## # A tibble: 1 x 2
##   CustomerID Highest_no_of_Trans
##        <int>               <int>
## 1      17841                7983
```

```
# Considering the Transaction Value
#with NA Values
Customer_high_transactionvalue_with_NA<-data %>% group_by(CustomerID) %>%
  summarise(Highest_Trans_Volume_with_NAValues=sum(Transaction_value)) %>%
  arrange(desc(Highest_Trans_Volume_with_NAValues)) %>% top_n(1)
```

```
## Selecting by Highest_Trans_Volume_with_NAValues
```

```
Customer_high_transactionvalue_with_NA
```

```
## # A tibble: 1 x 2
##   CustomerID Highest_Trans_Volume_with_NAValues
##        <int>                              <dbl>
## 1         NA                           1447682.
```

```
# without NA values
 Customer_high_transactionvalue_without_NA<- data %>% na.omit() %>% group_by(CustomerID) %>%
  summarise(Highest_Trans_Volume=sum(Transaction_value)) %>% arrange(desc(Highest_Trans_Volume)) %>% top
```

```
## Selecting by Highest_Trans_Volume
```

```
Customer_high_transactionvalue_without_NA
```

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

Question 7 The percentage of missing values for each variable in the data set

```
#Percentage of missing values in the data
Percentage_Missing_Values<-colMeans(is.na(data))
Percentage_Missing_Values
```

```
##         InvoiceNo         StockCode       Description          Quantity
##         0.0000000         0.0000000         0.0000000         0.0000000
##       InvoiceDate         UnitPrice        CustomerID           Country
##         0.0000000         0.0000000         0.2492669         0.0000000
## Transaction_value  New_Invoice_Date  Invoice_Day_Week  New_Invoice_Hour
##         0.0000000         0.0000000         0.0000000         0.0000000
## New_Invoice_Month
##         0.0000000
```

Data has 24.92% of missing Customer ID values.

Question 8 The number of transactions with missing Customer ID records by countries

```
#No. of transactions with missing Customer ID records by countries
data%>%filter(is.na(data$CustomerID)) %>% group_by(Country) %>%
  summarise(No_of_missing_ID=n()) %>% arrange(desc(No_of_missing_ID))
```

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

Question 9 On average, how often the customers comeback to the website for their next shopping?

```
# The average number of days between consecutive shopping per customer (with all the transactions)
data_without_NA<- data %>% na.omit()
Avg_days_Per_Customer<- select(data_without_NA,CustomerID,New_Invoice_Date) %>%
  distinct(CustomerID,New_Invoice_Date) %>%   group_by(CustomerID) %>%
  arrange(New_Invoice_Date) %>% summarise(avg=mean(diff(New_Invoice_Date))) %>%
  na.omit()

#The average number of days between shopping per customer (with out cancelled transactions)
Avg_days_Per_Cust_without_Cancelled_trans<- select(data_without_NA,CustomerID,New_Invoice_Date) %>%
  filter(data_without_NA$Quantity>0) %>% distinct(CustomerID,New_Invoice_Date) %>%
  group_by(CustomerID) %>% arrange(New_Invoice_Date) %>% summarise(avg=mean(diff(New_Invoice_Date))) %>%
  na.omit()
head(Avg_days_Per_Cust_without_Cancelled_trans)
```

```
## # A tibble: 6 x 2
##   CustomerID avg
##        <int> <drtn>
## 1      12347  60.83333 days
## 2      12348  94.33333 days
## 3      12352  43.33333 days
## 4      12356 151.50000 days
## 5      12358 149.00000 days
## 6      12359  91.33333 days
```

```
#Average number of days between consecutive shopping for all the customers
Avg_days_Per_Cust_without_Cancelled_trans%>% summarise(avg_days_between_shopping = mean(avg))
```

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

Question 10 n 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.

```
#Calculation of return rate for the french customers
Transactions_France<-filter(data,Country=='France')
Cancelled_Transactions_France<-filter(data,Country=='France'& Quantity<0)
Return_rate_France<- (nrow(Cancelled_Transactions_France)/nrow(Transactions_France))*100
Return_rate_France
```

```
## [1] 1.741264
```

The return rate for the customers in France is 1.741264

Question 11 Product that has generated the highest revenue for the retailer

```
Product_Revenue<-data %>% group_by(Description) %>% summarise(Product_Revenue=sum(Transaction_value)) %
```

```
## Selecting by Product_Revenue
```

```
as.data.frame(Product_Revenue)
```

```
##      Description Product_Revenue
## 1 DOTCOM POSTAGE        206245.5
```

Question 12 unique customers in the dataset

```
Unique_Customers<-length(unique(data$CustomerID))
Unique_Customers
```

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

There are 4373 unique customers in the data set