Business Analytics Assignment 1

Rakhee Moolchandani

10/31/2020

Install package if not already installed

```
#install.packages("lattice")
#install.packages("ggplot2")
#install.packages("dplyr")
```

load all the required libraries

```
library(readr)
library(gmodels)
library(dplyr)
```

Import the Universal Bank Dataset

```
# Read the CSV file
OnlineRetail <- read.csv("Online_Retail.csv")
# Print first few rows of the data set
head(OnlineRetail)
```

```
##   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
```

```
# O/p the summary of the data set
summary(OnlineRetail)
```

```
##   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
```

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.

```
# Total transactions in number and percentage
OnlineRetail %>% group_by(Country) %>% summarise(TotalNumber=n(), Percentage =n()*100/nrow(OnlineRetail)
```

```
## # A tibble: 4 x 3
##   Country        TotalNumber Percentage
##   <chr>                <int>      <dbl>
## 1 EIRE                  8196       1.51
## 2 France                8557       1.58
## 3 Germany               9495       1.75
## 4 United Kingdom      495478      91.4
```

The above table shows the total number and percentage of transactions for all the countries having transaction % more then 1%. United Kingdom seems to have the highest percentage of transactions 91.4%

2) Create a new variable 'TransactionValue' that is the product of the exising 'Quantity' and 'UnitPrice' variables. Add this variable to the dataframe

```
# Create new variable 'TransactionValue'
OnlineRetail$TransactionValue <- OnlineRetail$Quantity * OnlineRetail$UnitPrice
# Look at the data frame with new variable
head(OnlineRetail)
```

```
##   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 TransactionValue
## 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
```

```
# Look at the summary of the new variable
summary(OnlineRetail$TransactionValue)
```

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

A new variable 'TransactionValue' has been created and added to the dataframe.

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.

```
# Breakdown of Transaction Values by countries
OnlineRetail %>% group_by(Country) %>% summarize(TotalVolume=sum(TransactionValue)) %>% filter(TotalVolu
```

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

The above table shows the breakdown of sum of transaction values by countries. United Kingdom has the highest value.

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.

```
# let's convert InvoiceDate into a POSIXlt object
Temp=strptime(OnlineRetail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
# Check the variable
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"
```

```
# Now, let's separate date, day of the week and hour components
OnlineRetail$New_Invoice_Date <- as.Date(Temp)
# Convert dates to days of the week
OnlineRetail$Invoice_Day_Week= weekdays(OnlineRetail$New_Invoice_Date)
# Convert the Hour into normal numeric value
OnlineRetail$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
# Define month as seperate numeric variable
OnlineRetail$New_Invoice_Month = as.numeric(format(Temp, "%m"))
# See the data frame and its variables
head(OnlineRetail)
```

```
##   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 TransactionValue
## 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
```

New variables Date, Days of the week, Hour and month of the Invoice are generated and added in the dtataframe.

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

```
# Percentage of Transactions by numbers for Days of the week
OnlineRetail %>% group_by(Invoice_Day_Week) %>% summarise(TransactionPercentageByNumber=n()*100/nrow(On
```

```
## # A tibble: 6 x 2
##   Invoice_Day_Week TransactionPercentageByNumber
##   <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
```

The above table shows the percentage of transactions by numbers for all the days of the week.

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

```
# Percentage of Transactions by Volume for Days of the week
OnlineRetail %>% group_by(Invoice_Day_Week) %>% summarise(TransactionPercentageByVolume=sum(Transaction
```

```
## # A tibble: 6 x 2
##   Invoice_Day_Week TransactionPercentageByVolume
##   <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
```

The above table shows the percentage of transactions by volume for all the days of the week.

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

```
# Percentage of Transactions by Volume for months of the year
OnlineRetail %>% group_by(New_Invoice_Month) %>% summarise(TransactionPercentageByVolume=sum(Transactio
```

```
## # A tibble: 12 x 2
##    New_Invoice_Month TransactionPercentageByVolume
##                <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
```

The above table shows the percentage of transactions for every month of the year.

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

```
# Date with highest number of transactions in Australia
OnlineRetail %>% filter(Country=='Australia') %>% group_by(New_Invoice_Date) %>% summarise(n=n()) %>% a
```

```
## # A tibble: 49 x 2
##    New_Invoice_Date     n
##    <date>           <int>
##  1 2011-06-15         139
##  2 2011-07-19         137
##  3 2011-08-18          97
##  4 2011-03-03          84
##  5 2011-10-05          82
##  6 2011-05-17          73
##  7 2011-02-15          69
##  8 2011-01-06          48
##  9 2011-07-14          35
## 10 2011-09-16          34
## # ... with 39 more rows
```

2011-06-15 was the date with highest number of transactions 139 from Australia.

4e) 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.

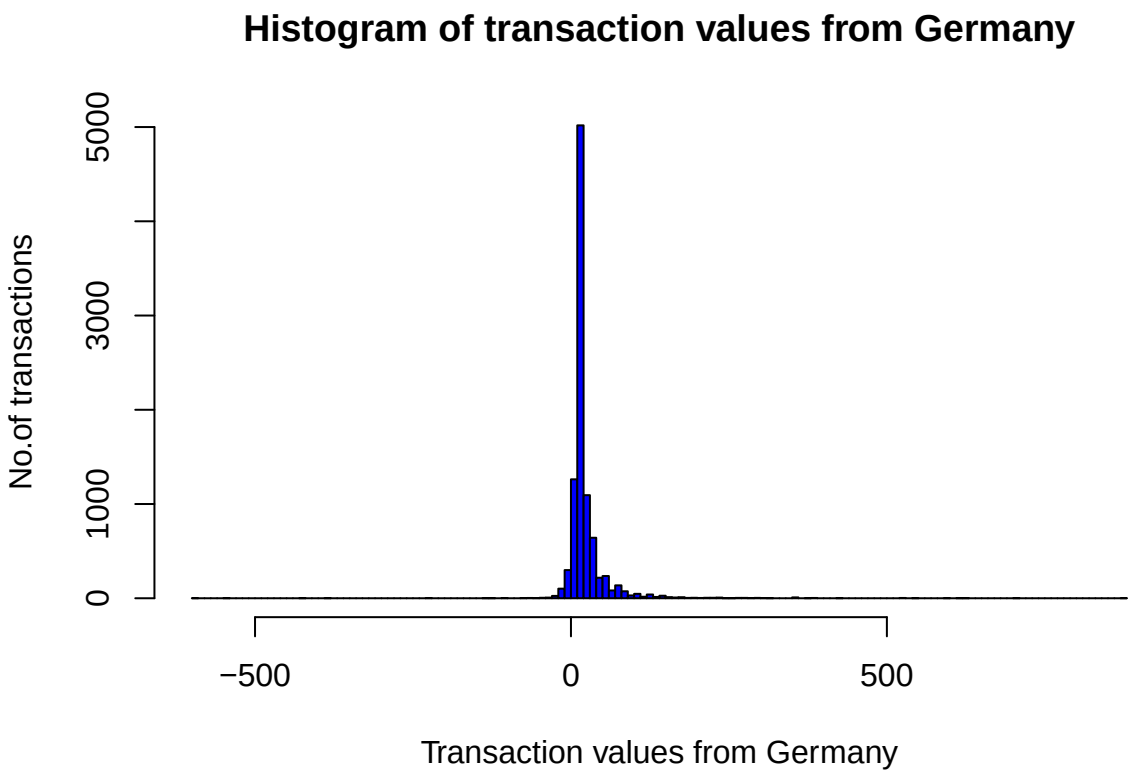
```
# Find the sum of invoice for hours from 7:00 to 20:00
OnlineRetail %>% group_by(New_Invoice_Hour) %>% summarise(n=n())
```

```
## # A tibble: 15 x 2
##    New_Invoice_Hour     n
##               <dbl> <int>
##  1                6    41
##  2                7   383
##  3                8  8909
##  4                9 34332
##  5               10 49037
##  6               11 57674
##  7               12 78709
##  8               13 72259
##  9               14 67471
## 10               15 77519
## 11               16 54516
## 12               17 28509
## 13               18  7974
## 14               19  3705
## 15               20   871
```

The sum is lowest for the hours 6:00 & 7:00 but since the responsible IT team is available only between 7:00 to 20:00, therefore the shut down hours will be from 18:00 till 20:00.

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

```
# Filter data with country Germany
Data <- filter(OnlineRetail, Country=='Germany')
#Plot the Graph
hist (Data$TransactionValue, n=200, xlab= "Transaction values from Germany", ylab= "No.of transactions"
```



The above graph shows the transaction values for Germany.

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

```
# Customer with highest number of transactions
OnlineRetail %>% group_by(CustomerID) %>% summarise(n=n()) %>% arrange(desc(n))
```

```
## # A tibble: 4,373 x 2
##    CustomerID      n
##         <int>  <int>
##  1         NA 135080
##  2      17841   7983
##  3      14911   5903
##  4      14096   5128
##  5      12748   4642
##  6      14606   2782
##  7      15311   2491
##  8      14646   2085
##  9      13089   1857
## 10      13263   1677
## # ... with 4,363 more rows
```

```
# Most valuable customer
OnlineRetail %>% group_by(CustomerID) %>% summarise(Sum=sum(TransactionValue)) %>% arrange(desc(Sum))
```

```
## # A tibble: 4,373 x 2
##    CustomerID      Sum
##         <int>    <dbl>
##  1         NA 1447682.
```

```
##  2      14646  279489.
##  3      18102  256438.
##  4      17450  187482.
##  5      14911  132573.
##  6      12415  123725.
##  7      14156  113384.
##  8      17511   88125.
##  9      16684   65892.
## 10      13694   62653.
## # ... with 4,363 more rows
```

Customer ID 17841 has the highent number of transactions i.e. 7983.

Customer ID 14646 is the most valuable cutomer. This customer has highest total sum of transactions of value 279489.020.

7) Calculate the percentage of missing values for each variable in the dataset

```
# Percentage of missing values
colMeans(is.na(OnlineRetail))*100
```

```
##         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
```

Only the Customer ID varible has the missing values of approximately 25%.

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

```
# Number of missing customer Ids for each country
OnlineRetail %>% filter(is.na(CustomerID)) %>% group_by(Country) %>% summarise(n=n()) %>% arrange(desc(n
```

```
## # A tibble: 9 x 2
##   Country             n
##   <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
```

The above table shows the number of transactions with missing customer ID for all countries.

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)

```
# Calculate no of days between consecutive shopping for the customers
OrdersData <- OnlineRetail %>% arrange(CustomerID, New_Invoice_Date)
Days = diff(OrdersData$New_Invoice_Date,1);
Days = Days[Days>0]
mean(Days)
```

```
## Time difference of 41.6156 days
```

On an average, in 41.6 days customers comeback to the website for their next shopping.

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.

```
# Calculate the return rate of the French Customers
FrenchTransactions <- filter(OnlineRetail, Country=='France')
FrenchTransactionsCancelled <- filter(FrenchTransactions, Quantity<0)
nrow(FrenchTransactionsCancelled)/nrow(FrenchTransactions)*100
```

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

The return rate for the french customers is 1.74%.

11) What is the product that has generated the highest revenue for the retailer? (i.e. item with the highest total sum of 'TransactionValue')

```
# Find the product that generated highest revenue
OnlineRetail %>% group_by(StockCode) %>% summarise(TransValue=sum(TransactionValue)) %>% arrange(desc(T
```

```
## # A tibble: 4,070 x 2
##    StockCode TransValue
##    <chr>          <dbl>
##  1 DOT          206245.
##  2 22423        164762.
##  3 47566         98303.
##  4 85123A        97894.
##  5 85099B        92356.
##  6 23084         66757.
##  7 POST          66231.
##  8 22086         63792.
##  9 84879         58960.
## 10 79321         53768.
## # ... with 4,060 more rows
```

Stockcode Dot has the highest revenue for the retailer.

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

```
# Find unique customers in the data set
length(unique(OnlineRetail$CustomerID))
```

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

There are 4373 unique customers in the data set.