Business Analytics Assignment 2

Nikitha Chigurupati

10/30/2022

#Importing the Dataset

```
Online_Retail<- read.csv("C:/Users/Nikitha/Downloads/Online_Retail.csv")
summary(Online_Retail)
```

```
##   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
```

#Loading the Packages

```
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## 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
```

```
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(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.

```
Online_Retail %>% group_by(Country) %>% summarise(n())
```

```
## # A tibble: 38 x 2
##    Country         `n()`
##    <chr>           <int>
##  1 Australia        1259
##  2 Austria           401
##  3 Bahrain            19
##  4 Belgium          2069
##  5 Brazil             32
##  6 Canada            151
##  7 Channel Islands   758
##  8 Cyprus            622
##  9 Czech Republic     30
## 10 Denmark           389
## # ... with 28 more rows
```

```
Online_Retail %>% group_by(Country) %>% summarise(percent =100 *n()/nrow(Online_Retail))
```

```
## # A tibble: 38 x 2
##    Country         percent
##    <chr>             <dbl>
##  1 Australia       0.232
##  2 Austria         0.0740
##  3 Bahrain         0.00351
##  4 Belgium         0.382
##  5 Brazil          0.00591
##  6 Canada          0.0279
##  7 Channel Islands 0.140
##  8 Cyprus          0.115
##  9 Czech Republic  0.00554
## 10 Denmark         0.0718
## # ... with 28 more rows
```

```
Online_Retail %>% group_by(Country) %>% summarise(percent =100 *n()/nrow(Online_Retail)) %>% fil
ter(Country>0.01)
```

```
## # A tibble: 38 x 2
##    Country         percent
##    <chr>             <dbl>
##  1 Australia       0.232
##  2 Austria         0.0740
##  3 Bahrain         0.00351
##  4 Belgium         0.382
##  5 Brazil          0.00591
##  6 Canada          0.0279
##  7 Channel Islands 0.140
##  8 Cyprus          0.115
##  9 Czech Republic  0.00554
## 10 Denmark         0.0718
## # ... with 28 more rows
```

#QUESTION 2: Create a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables. Add this variable to the dataframe.

```
TransactionValue <-  Online_Retail$Quantity * Online_Retail$UnitPrice

Online_Retail <- cbind(Online_Retail, TransactionValue)
head(Online_Retail)
```

```
##   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
```

```
colnames(Online_Retail)
```

```
## [1] "InvoiceNo"        "StockCode"        "Description"      "Quantity"
## [5] "InvoiceDate"      "UnitPrice"        "CustomerID"       "Country"
## [9] "TransactionValue"
```

#QUESTION 3: Using the newly created variable,TransactionValue, show the breakdown of transactionvaluesby 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.

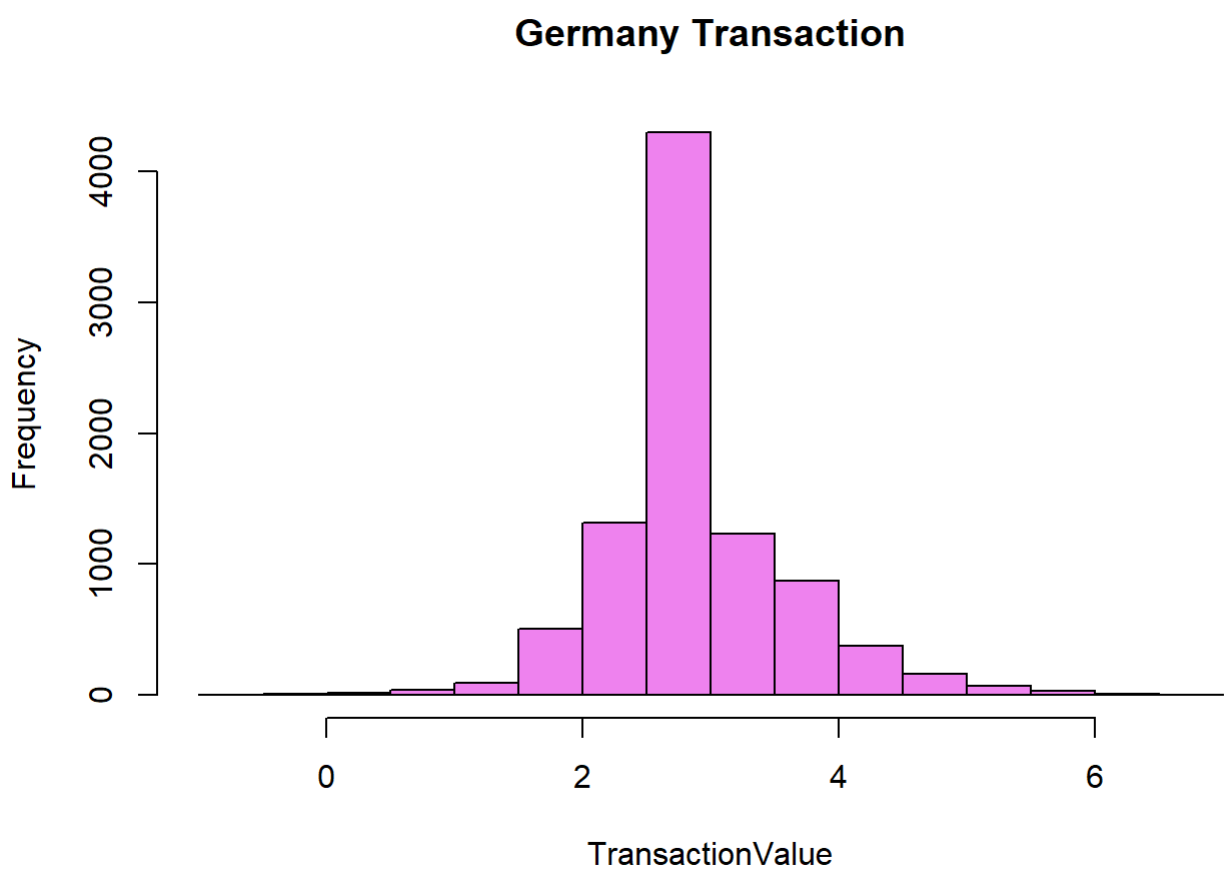
```
Online_Retail%>%group_by(Country)%>%summarise(Sum_of_Transaction_values = sum(TransactionValu
e))%>%filter(Sum_of_Transaction_values >130000)
```

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

#QUESTION 5: Plot the histogramof transaction values from Germany.

```
hist(x=log(Online_Retail$TransactionValue[Online_Retail$Country=="Germany"]),xlab = "Transaction
Value",col = 'violet' ,main = 'Germany Transaction',ylab = 'Frequency')
```

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



#QUESTION 6: Which customer had the highest number of transactions? Which customer is most valuable i.e. highest total sum of transactions

```
#The customer that had the highest number  of transanctions.
Online_Retail %>%group_by(CustomerID)%>%summarise(CustomerTransaction = n())%>%filter(CustomerID
!= "NA")%>%filter(CustomerTransaction ==max(CustomerTransaction))
```

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

```
#The most valuable customer that had the highest total sum of transanction.
Online_Retail%>%group_by(CustomerID)%>%summarise(total.transaction.by.each.customer = sum(Transa
ctionValue))%>%arrange(desc(total.transaction.by.each.customer))%>%filter(CustomerID != "NA")%>%
filter(total.transaction.by.each.customer ==max(total.transaction.by.each.customer) )
```

```
## # A tibble: 1 x 2
##   CustomerID total.transaction.by.each.customer
##        <int>                              <dbl>
## 1      14646                           279489.
```

#QUESTION 7 :Calculate the percentage of missing values for each variable in the dataset

```
colMeans(is.na(Online_Retail))
```

```
##        InvoiceNo        StockCode      Description         Quantity
##        0.0000000        0.0000000        0.0000000        0.0000000
##      InvoiceDate        UnitPrice       CustomerID          Country
##        0.0000000        0.0000000        0.2492669        0.0000000
## TransactionValue
##        0.0000000
```

#QUESTION 8 :What are the number oftransactions withmissing CustomerID records by countries?

```
Online_Retail%>% group_by(Country)%>%filter(is.na(CustomerID))%>%summarise(Missing_CustomerID=n
())
```

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

#QUESTION 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 numberof 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 hasa negative value.

```
Returns <-nrow(Online_Retail%>%group_by(CustomerID)%>%filter((Country=='France')&(TransactionVal
ue<0)&(CustomerID != 'Na')))

Totalfrenchcustomer<-nrow(Online_Retail%>%group_by(CustomerID)%>%filter((Country=='France')&(Cus
tomerID != 'Na')))

Returns/Totalfrenchcustomer*100
```

```
## [1] 1.754799
```

#QUESTION 11: What is the product that has generated the highest revenue for the retailer?

```
Total_customer1<-Online_Retail%>%group_by(Description,StockCode)%>%summarise(n=sum(TransactionVa
lue))%>%arrange(desc(n))
```

```
## `summarise()` has grouped output by 'Description'. You can override using the
## `.groups` argument.
```

```
Total_customer1[Total_customer1['n']==max(Total_customer1['n']),]
```

```
## # A tibble: 1 x 3
## # Groups:   Description [1]
##   Description    StockCode       n
##   <chr>          <chr>       <dbl>
## 1 DOTCOM POSTAGE DOT       206245.
```

#QUESTION 12: How many unique customers are represented in the dataset?

```
length(unique(Online_Retail$CustomerID))
```

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

#GOLDEN QUESTION: QUESTION- 4

```
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]
```

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

```
Online_Retail$Invoice_Day_Week=weekdays(Online_Retail$New_Invoice_Date)
Online_Retail$New_Invoice_Hour =as.numeric(format(Temp,"%H"))
Online_Retail$New_Invoice_Month = as.numeric(format(Temp, "%m"))
```

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

Online_Retail%>%group_by(Invoice_Day_Week)%>%summarise(No.of.transaction=(n()))%>%
mutate(No.of.transaction,'percent'=(No.of.transaction*100)/sum(No.of.transaction))
```

```
## # A tibble: 6 x 3
##   Invoice_Day_Week No.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
```

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

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

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

```
#c)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 x 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
```

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

No_of_Trans_Aust<-Online_Retail%>%group_by(New_Invoice_Date,Country)%>%filter(Country=='Australi
a')%>%
summarise(Number=sum(Quantity),amount=sum(TransactionValue))%>%arrange(desc(Number))
```

```
## `summarise()` has grouped output by 'New_Invoice_Date'. You can override using
## the `.groups` argument.
```

```
No_of_Trans_Aust
```

```
## # A tibble: 49 x 4
## # Groups:   New_Invoice_Date [49]
##    New_Invoice_Date Country   Number amount
##    <date>           <chr>      <int>  <dbl>
##  1 2011-06-15       Australia  15241 23427.
##  2 2011-08-18       Australia  12196 21880.
##  3 2011-03-03       Australia  10162 16558.
##  4 2011-02-15       Australia   8384 14023.
##  5 2011-05-17       Australia   8268 11925.
##  6 2011-10-05       Australia   7135 16472.
##  7 2011-01-06       Australia   4802  7154.
##  8 2011-07-13       Australia   4332  2796.
##  9 2011-11-15       Australia   3130  5355.
## 10 2011-09-01       Australia   2836  2942.
## # ... with 39 more rows
```

```
No_of_Trans_Aust<-No_of_Trans_Aust[No_of_Trans_Aust['Number']==max(No_of_Trans_Aust['Number']),]
No_of_Trans_Aust
```

```
## # A tibble: 1 x 4
## # Groups:   New_Invoice_Date [1]
##   New_Invoice_Date Country   Number amount
##   <date>           <chr>      <int>  <dbl>
## 1 2011-06-15       Australia  15241 23427.
```

#e)The company needs to shut down the website for two consecutivehours 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.

```
Fp=Online_Retail%>%group_by(New_Invoice_Hour)%>%summarise(Total.transaction= n())
Fp
```

```
## # A tibble: 15 x 2
##    New_Invoice_Hour Total.transaction
##               <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
```

```
Sp<-rollapply(Fp['Total.transaction'],2,sum)%>%index(min(Sp))
Sp
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14
```

```
print('The best time to shut down the website for two consecutive hours is between 7am-9am')
```

```
## [1] "The best time to shut down the website for two consecutive hours is between 7am-9am"
```

#GOLDEN QUESTION: QUESTION 9 :On average, how often the costumers comeback to the website for their next shopping?

```
Avg<-Online_Retail%>%group_by(CustomerID)%>%summarise(diff_consecutivedays= diff(New_Invoice_Dat
e))%>%filter(diff_consecutivedays>0)
```

```
## `summarise()` has grouped output by 'CustomerID'. You can override using the
## `.groups` argument.
```

```
print(paste('The average number of days between consecutive shopping is',mean(Avg$diff_con
secutivedays)))
```

```
## [1] "The average number of days between consecutive shopping is 38.4875"
```