Assignment - 2

Harika

2023-10-15

```
OnlineRetail <- read.csv("c:/Users/Harika/Downloads/Online_Retail.csv")
```

```
###The file is loaded into an R DataFrame by the above command.
```

```
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
```

```
###The above data represents the summary for the given dataset.
```

#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.

```
Country_total_number <- table(OnlineRetail$Country)
Country_total_number
```

```
##
##            Australia              Austria              Bahrain
##                 1259                  401                   19
##              Belgium               Brazil               Canada
##                 2069                   32                  151
##      Channel Islands               Cyprus       Czech Republic
##                  758                  622                   30
##              Denmark                 EIRE   European Community
##                  389                 8196                   61
##              Finland               France              Germany
##                  695                 8557                 9495
##               Greece            Hong Kong              Iceland
##                  146                  288                  182
##               Israel                Italy                Japan
##                  297                  803                  358
##              Lebanon            Lithuania                Malta
##                   45                   35                  127
##          Netherlands               Norway               Poland
##                 2371                 1086                  341
##             Portugal                  RSA         Saudi Arabia
##                 1519                   58                   10
##            Singapore                Spain               Sweden
##                  229                 2533                  462
##          Switzerland United Arab Emirates       United Kingdom
##                 2002                   68               495478
##          Unspecified                  USA
##                  446                  291
```

```
###The data above represents the breakdown of the number of transactions by country from the given data
```

```
transaction_percent <- round(100*prop.table(Country_total_number),digits = 2)
transaction_percent
```

```
## 
##            Australia              Austria              Bahrain
##                 0.23                 0.07                 0.00
##              Belgium               Brazil               Canada
##                 0.38                 0.01                 0.03
##      Channel Islands               Cyprus       Czech Republic
##                 0.14                 0.11                 0.01
##              Denmark                 EIRE   European Community
##                 0.07                 1.51                 0.01
##              Finland               France              Germany
##                 0.13                 1.58                 1.75
##               Greece            Hong Kong              Iceland
##                 0.03                 0.05                 0.03
##               Israel                Italy                Japan
##                 0.05                 0.15                 0.07
##              Lebanon            Lithuania                Malta
##                 0.01                 0.01                 0.02
##          Netherlands               Norway               Poland
##                 0.44                 0.20                 0.06
##             Portugal                  RSA         Saudi Arabia
##                 0.28                 0.01                 0.00
##            Singapore                Spain               Sweden
##                 0.04                 0.47                 0.09
##          Switzerland United Arab Emirates       United Kingdom
##                 0.37                 0.01                91.43
##          Unspecified                  USA
##                 0.08                 0.05
```

```
###The above data represents the percentage of transactions for each country.
```

```
total <- data.frame(Country=names(Country_total_number),
                    TotalNumber=Country_total_number,
                    percentage=transaction_percent)
total
```

```
##                 Country     TotalNumber.Var1 TotalNumber.Freq
## 1             Australia            Australia             1259
## 2               Austria              Austria              401
## 3               Bahrain              Bahrain               19
## 4               Belgium              Belgium             2069
## 5                Brazil               Brazil               32
## 6                Canada               Canada              151
## 7       Channel Islands      Channel Islands              758
## 8                Cyprus               Cyprus              622
## 9        Czech Republic       Czech Republic               30
## 10              Denmark              Denmark              389
## 11                 EIRE                 EIRE             8196
## 12   European Community   European Community               61
## 13              Finland              Finland              695
## 14               France               France             8557
## 15              Germany              Germany             9495
## 16               Greece               Greece              146
## 17            Hong Kong            Hong Kong              288
## 18              Iceland              Iceland              182
## 19               Israel               Israel              297
## 20                Italy                Italy              803
## 21                Japan                Japan              358
## 22              Lebanon              Lebanon               45
## 23            Lithuania            Lithuania               35
## 24                Malta                Malta              127
## 25          Netherlands          Netherlands             2371
## 26               Norway               Norway             1086
## 27               Poland               Poland              341
## 28             Portugal             Portugal             1519
## 29                  RSA                  RSA               58
## 30         Saudi Arabia         Saudi Arabia               10
## 31            Singapore            Singapore              229
## 32                Spain                Spain             2533
## 33               Sweden               Sweden              462
## 34          Switzerland          Switzerland             2002
## 35 United Arab Emirates United Arab Emirates               68
## 36       United Kingdom       United Kingdom           495478
## 37          Unspecified          Unspecified              446
## 38                  USA                  USA              291
##         percentage.Var1 percentage.Freq
## 1             Australia            0.23
## 2               Austria            0.07
## 3               Bahrain            0.00
## 4               Belgium            0.38
## 5                Brazil            0.01
## 6                Canada            0.03
## 7       Channel Islands            0.14
## 8                Cyprus            0.11
## 9        Czech Republic            0.01
## 10              Denmark            0.07
## 11                 EIRE            1.51
## 12   European Community            0.01
## 13              Finland            0.13
## 14               France            1.58
## 15              Germany            1.75
## 16               Greece            0.03
## 17            Hong Kong            0.05
## 18              Iceland            0.03
## 19               Israel            0.05
## 20                Italy            0.15
## 21                Japan            0.07
## 22              Lebanon            0.01
## 23            Lithuania            0.01
## 24                Malta            0.02
## 25          Netherlands            0.44
## 26               Norway            0.20
## 27               Poland            0.06
## 28             Portugal            0.28
## 29                  RSA            0.01
## 30         Saudi Arabia            0.00
## 31            Singapore            0.04
## 32                Spain            0.47
## 33               Sweden            0.09
## 34          Switzerland            0.37
## 35 United Arab Emirates            0.01
## 36       United Kingdom           91.43
## 37          Unspecified            0.08
## 38                  USA            0.05
```

```
###The data above combines the total number and percentage of transactions into a table.
```

```
total <- subset(total,transaction_percent>1)
total
```

```
##           Country TotalNumber.Var1 TotalNumber.Freq percentage.Var1
## 11           EIRE             EIRE             8196            EIRE
## 14         France           France             8557          France
## 15        Germany          Germany             9495         Germany
## 36 United Kingdom   United Kingdom           495478  United Kingdom
##    percentage.Freq
## 11            1.51
## 14            1.58
## 15            1.75
## 36           91.43
```

```
###The data above represents a subset of the table, showing only countries that account for more than 1
```

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

```
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
```

```
###This command calls 'dplyr' library.
```

```
OnlineRetail <- OnlineRetail %>% mutate(TransactionValue= Quantity*UnitPrice)
summary(OnlineRetail$TransactionValue)
```

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

```
###The above data represents the product of the 'Quantity' and 'UnitPrice' variables and assigned the r
```

#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.

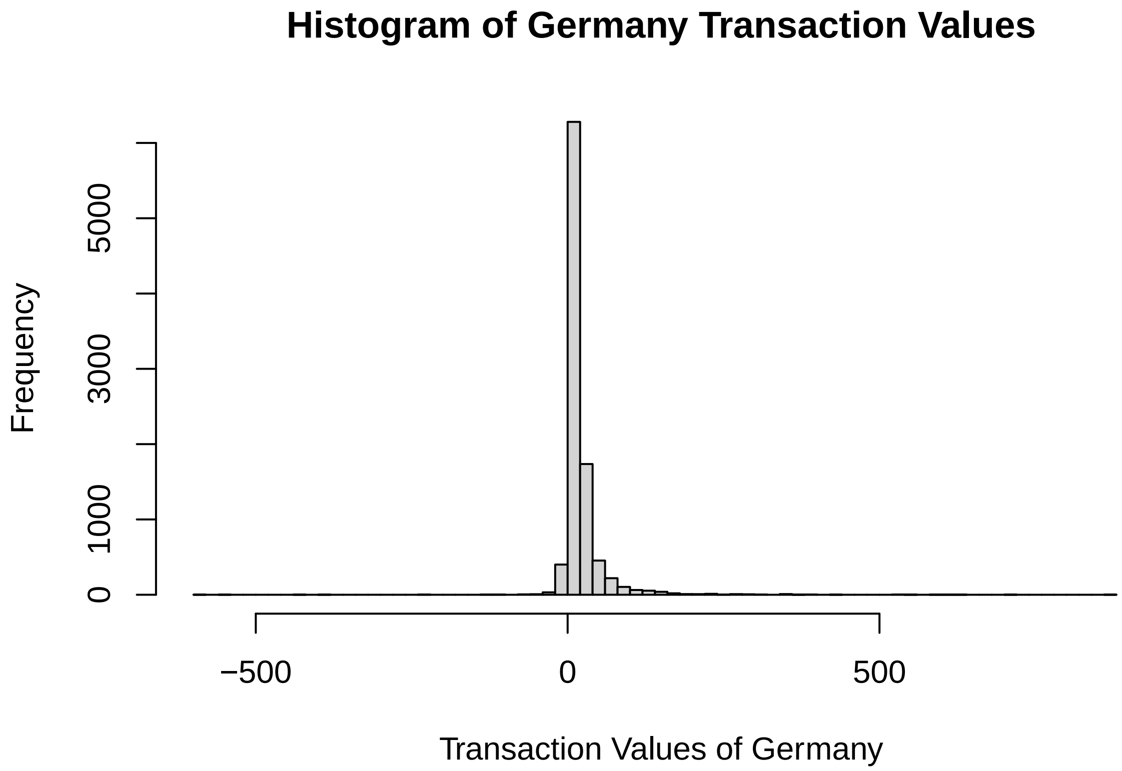
```
data <- summarise(group_by(OnlineRetail,Country),sum_1= sum(TransactionValue))
Transaction <- filter(data,sum_1 > 130000)
Transaction
```

```
## # A tibble: 6 x 2
##   Country           sum_1
##   <chr>             <dbl>
## 1 Australia       137077.
## 2 EIRE            263277.
## 3 France          197404.
## 4 Germany         221698.
## 5 Netherlands     284662.
## 6 United Kingdom 8187806.
```

```
###The data above shows the total transaction values for each country. It includes only those countries
```

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

```
Germany_data <- subset(OnlineRetail,Country == "Germany")
hist(Germany_data$TransactionValue, xlim = c(-600,900),breaks=100, xlab = "Transaction Values of German
```



```
###This is a representation of the histogram for transaction values in Germany.
```

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

```
OnlineRetail1 <- na.omit(OnlineRetail)
result1 <- summarise(group_by(OnlineRetail1,CustomerID), sum2 = sum(TransactionValue))
result1[which.max(result1$sum2),]
```

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

```
data2 <- table(OnlineRetail$CustomerID)
data2 <- as.data.frame(data2)
result2 <- data2[which.max(data2$Freq),]
result2
```

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

```
### This data represents the customer with the highest number of transactions and the most valuable cus
```

#7 > Calculate the percentage of missing values for each variable in the dataset.

```
missing_values <- colMeans(is.na(OnlineRetail)*100)
missing_values
```

```
##        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
##          0.00000
```

```
###The data above represents Missing Values in the given dataset.
```

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

```
OnlineRetail2 <- OnlineRetail %>% filter(is.na(CustomerID)) %>% group_by(Country)
summary(OnlineRetail2$Country)
```

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

```
###The data above shows the number of transactions with missing CustomerID records, records by countrie
```

#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.

```
OnlineRetail_Table <- filter(OnlineRetail,Country == "France")
Total_Row <- nrow(OnlineRetail_Table)
Cancel <- nrow(subset(OnlineRetail_Table,TransactionValue<0))
Cancel
```

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

```
NotCancel <- Total_Row-Cancel
NotCancel
```

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

```
Return_Rate <- Cancel / Total_Row
Return_Rate
```

```
## [1] 0.01741264
```

```
###The data above represents values for both cancelled and not cancelled transactions, as well as the r
```

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

```
Transaction_Value <- tapply(OnlineRetail$TransactionValue, OnlineRetail$StockCode, sum)
Transaction_Value[which.max(Transaction_Value)]
```

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

```
###The data above represents the item with the highest total sum of 'Transaction Value'.
```

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

```
unique_customers <- unique(OnlineRetail$CustomerID)
length(unique_customers)
```

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

```
###The above value represents number of unique customers that are present in the given dataset.
```