BA ASSIGNMENT 2

Zachariah Alex

2022-10-29

```
#Loading Library Functions

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
```

```
#Reading data from CSV file

data1<-read.csv("Online_Retail.csv")

head(data1)
```

```
##   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
```

```
# 1- Breakdown of the number of transactions by countries

data2<-data1%>%group_by(Country)%>%count()%>%mutate(percent=n/nrow(data1)*100)%>%filter(percent>1)
head(data2)
```

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

```
#2- Creating a new variable 'TransactionValue'

data1$TransactionValue<-c(data1$Quantity*data1$UnitPrice)
head(data1)
```

```
##   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
```

```
#3- Breakdown of transaction values by countries exceeding 13000 pounds

data3<-data1%>%group_by(Country)%>% summarize(x = sum(TransactionValue)) %>%filter(x>13000)
head(data3)
```

```
## # A tibble: 6 x 2
##   Country              x
##   <chr>            <dbl>
## 1 Australia      137077.
## 2 Belgium         40911.
## 3 Channel Islands 20086.
## 4 Denmark         18768.
## 5 EIRE           263277.
## 6 Finland         22327.
```

```
#4 -Golden Question
data1_new<-data1
Temp=strptime(data1$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"
```

```r
#New invoice date

data1_new$New_Invoice_Date<- as.Date(Temp)

data1_new$New_Invoice_Date[20000]-data1_new$New_Invoice_Date[10]
```

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

```r
#separating date, day of the week and hour components

#converting dates to days

data1_new$Invoice_Day_Week= weekdays(data1_new$New_Invoice_Date)

#Invoice Hour to numeric

data1_new$New_Invoice_Hour = as.numeric(format(Temp, "%H"))

#Invoice Month to numeric

data1_new$New_Invoice_Month = as.numeric(format(Temp, "%m"))

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

percent<-data1_new %>% group_by(Invoice_Day_Week) %>% summarise(count=n()) %>% mutate(x=(count/sum(coun

head(percent)
```

```
## # A tibble: 6 x 3
##   Invoice_Day_Week  count     x
##   <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
```

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

percent_week<-data1_new%>% group_by(Invoice_Day_Week) %>%summarise(Value=sum(TransactionValue)) %>% muta
head(percent_week)
```

```
## # A tibble: 6 x 3
##   Invoice_Day_Week    Value     x
##   <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
```

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

percent_month<-data1_new %>% group_by(New_Invoice_Month) %>% summarise(Value=sum(TransactionValue)) %>%
head(percent_month)
```

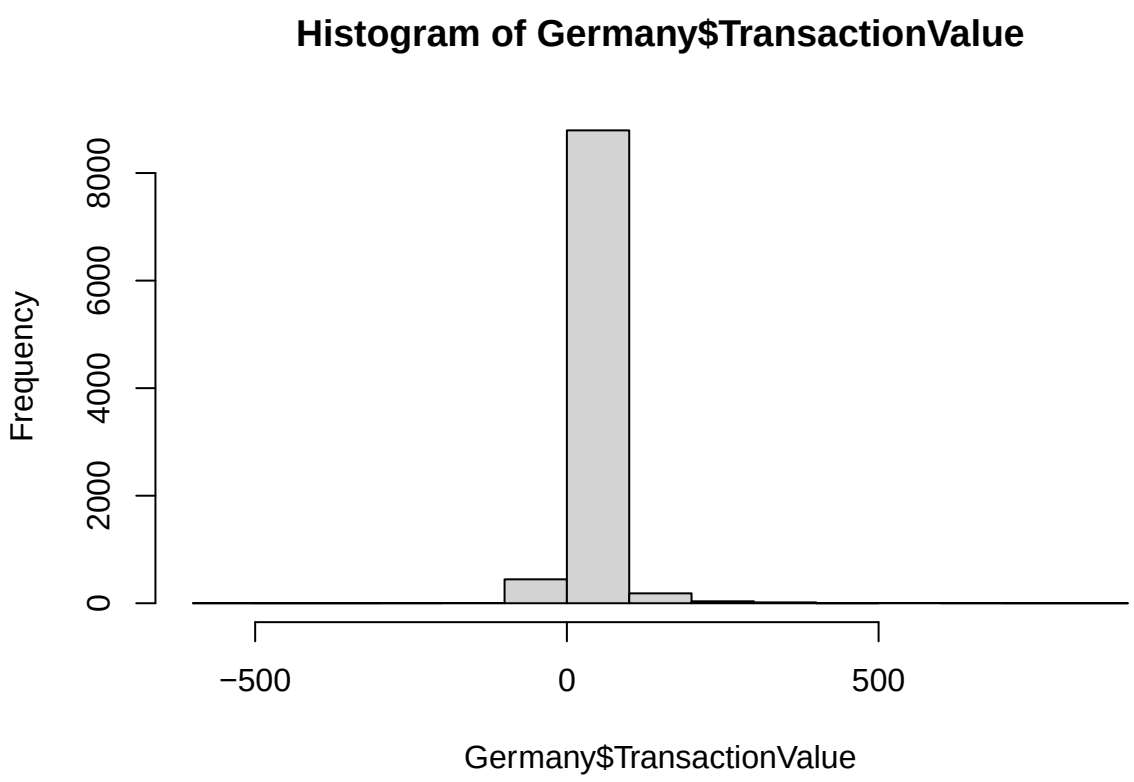
```
## # A tibble: 6 x 3
##   New_Invoice_Month   Value     x
##               <dbl>   <dbl> <dbl>
## 1                 1 560000.  5.74
## 2                 2 498063.  5.11
## 3                 3 683267.  7.01
## 4                 4 493207.  5.06
## 5                 5 723334.  7.42
## 6                 6 691123.  7.09
```

```r
#4(d) The date with the highest number of transactions from Australia

tran_date<-data1_new%>% filter(Country == 'Australia') %>% group_by(New_Invoice_Date) %>% summarise(cou

head(tran_date)
```

```
## # A tibble: 6 x 2
##   New_Invoice_Date count
##   <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
```

```r
#Highest number of transactions were on 15/06/2011
```

```r
#5- Plotting the histogram of transaction values from Germany
Germany<-data1%>%group_by(Country)%>%filter(Country=="Germany")
hist(Germany$TransactionValue)
```




```
#6- Customer having the highest number of transactions and  most valuable customer

data4<-data1%>%group_by(CustomerID)%>%summarise(ct=n())%>%arrange(desc(ct))
head(data4)
```

```
## # A tibble: 6 x 2
##   CustomerID     ct
##        <int>  <int>
## 1         NA 135080
## 2      17841   7983
## 3      14911   5903
## 4      14096   5128
## 5      12748   4642
## 6      14606   2782
```

```
#Customer 17841 is having the highest number of transactions

data1%>%group_by(CustomerID)%>%summarise(y=sum(TransactionValue))%>%arrange(desc(y))
```

```
## # A tibble: 4,373 x 2
##    CustomerID        y
##         <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
```

```
#The most valuable customer is 14646


#7- Percentage of missing values for each variable in the dataset

data1%>%is.na()%>%colMeans()*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
##          0.00000
```

```
#Only customer ID Coloumn are missing values with a total of 24.92669%
```

```
#8 -Number of transactions with missing customer id by countries

data4<-data1%>%filter(is.na(CustomerID))%>%group_by(Country)%>%count()
head(data4)
```

```
## # A tibble: 6 x 2
## # Groups:   Country [6]
##   Country        n
##   <chr>      <int>
## 1 Bahrain        2
## 2 EIRE         711
## 3 France        66
## 4 Hong Kong    288
## 5 Israel        47
## 6 Portugal      39
```

```
#9 -The average number of days between consecutive shopping

ndays<-data1_new%>% select(CustomerID,New_Invoice_Date) %>% group_by(CustomerID) %>% distinct(New_Invoi
head(ndays)
```

```
## # A tibble: 6 x 3
## # Groups:   CustomerID [2]
##   CustomerID New_Invoice_Date days
##        <int> <date>           <drtn>
## 1      18287 2011-10-12       143 days
## 2      18287 2011-10-28        16 days
```

```
## 3      18283 2011-01-23        17 days
## 4      18283 2011-02-28        36 days
## 5      18283 2011-04-21        52 days
## 6      18283 2011-05-23        32 days
```

```
#Finding average number of days

mean(ndays$days)
```

```
## Time difference of 38.4875 days
```

```
#10 - The return rate for the French customers

total<-data1%>%filter(Country=="France")%>%count()


cancelled<-data1%>%filter(Country=="France" & Quantity<0)%>% summarize(count = n())


returnRate = (cancelled/total)*100

View(returnRate)

#The return rate for French Customers= 1.741264
```

```
#11 - The product that has generated the highest revenue for the retailer

revenue<-data1%>%group_by(StockCode)%>%summarise(x=sum(TransactionValue))%>%arrange(desc(x))

head(revenue)
```

```
## # A tibble: 6 x 2
##   StockCode       x
##   <chr>       <dbl>
## 1 DOT       206245.
## 2 22423     164762.
## 3 47566      98303.
## 4 85123A     97894.
## 5 85099B     92356.
## 6 23084      66757.
```

```
#DOTCOM POSTAGE with Stockcode DOT has generated the highest revenue for the retailer
```

```
#12 - Number of unique customers represented in the dataset

unique(data1$CustomerID)%>%length()
```

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

```
#Number of unique customers represented in the dataset =4373
```