BA_Assignment_2

Karthik Badiganti

2022-10-10

Loading Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Importing & Cleaning Data

We are Importing Data from CSV file and cleaning

```
Online_retail <- read.csv("C://Users//gbkar//Documents//R Scripts//Online_Retail.csv")
```

Question 1

```
Online_retail %>%
  group_by(Country)  %>%
  summarise(percentage=(n()/nrow(Online_retail))*100, Total=n()) %>%
  filter(percentage>1)
```

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

Question 2

```
Online_retail <- Online_retail %>%
  mutate(TransactionValue = Quantity*UnitPrice
    )
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
```

Question 3

```
Online_retail %>% group_by(Country)  %>%
  summarise(Total_sum_Transaction=sum(TransactionValue))%>%
  filter(Total_sum_Transaction>130000)
```

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

Question 4 prep

```
Temp=strptime(Online_retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
Online_retail$New_Invoice_Date <- as.Date(Temp)
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"))
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
##   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
```

Question 4

```
# a

Online_retail %>% group_by(Invoice_Day_Week)  %>%
  summarise(percentage_by_num=(sum(TransactionValue)/sum(Online_retail$TransactionValue)*100))
```

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

```r
# b
Online_retail %>% group_by(Invoice_Day_Week)  %>%
  summarise(percentage_by_volume=(n()/nrow(Online_retail))*100)
```

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

```r
# c
Online_retail %>% group_by(New_Invoice_Month)%>%
  summarise(month_percentage=(n()/nrow(Online_retail))*100)
```

```
## # A tibble: 12 x 2
##    New_Invoice_Month month_percentage
##                <dbl>            <dbl>
##  1                 1             6.49
##  2                 2             5.11
##  3                 3             6.78
##  4                 4             5.52
##  5                 5             6.83
##  6                 6             6.80
##  7                 7             7.29
##  8                 8             6.51
##  9                 9             9.27
## 10                10            11.2
## 11                11            15.6
## 12                12            12.5
```

```r
# d
Online_retail %>% filter(Country=='Australia')%>%
  group_by(New_Invoice_Date) %>% summarise(max_Trans_count=n())%>%
  filter(max_Trans_count==max(max_Trans_count))
```

```
## # A tibble: 1 x 2
##   New_Invoice_Date max_Trans_count
##   <date>                    <int>
## 1 2011-06-15                  139
```

```r
# e
Hour1<-as.data.frame(Online_retail %>%
                       filter(New_Invoice_Hour>6 & New_Invoice_Hour<21 )%>%
                       group_by(New_Invoice_Hour)%>%
                       summarise(Trans_count=n()))
Hour2<-which.min((rollapply(Hour1$Trans_count,2,sum)))
print("The consecutive 2 hours where the downtime can be done is:")
```

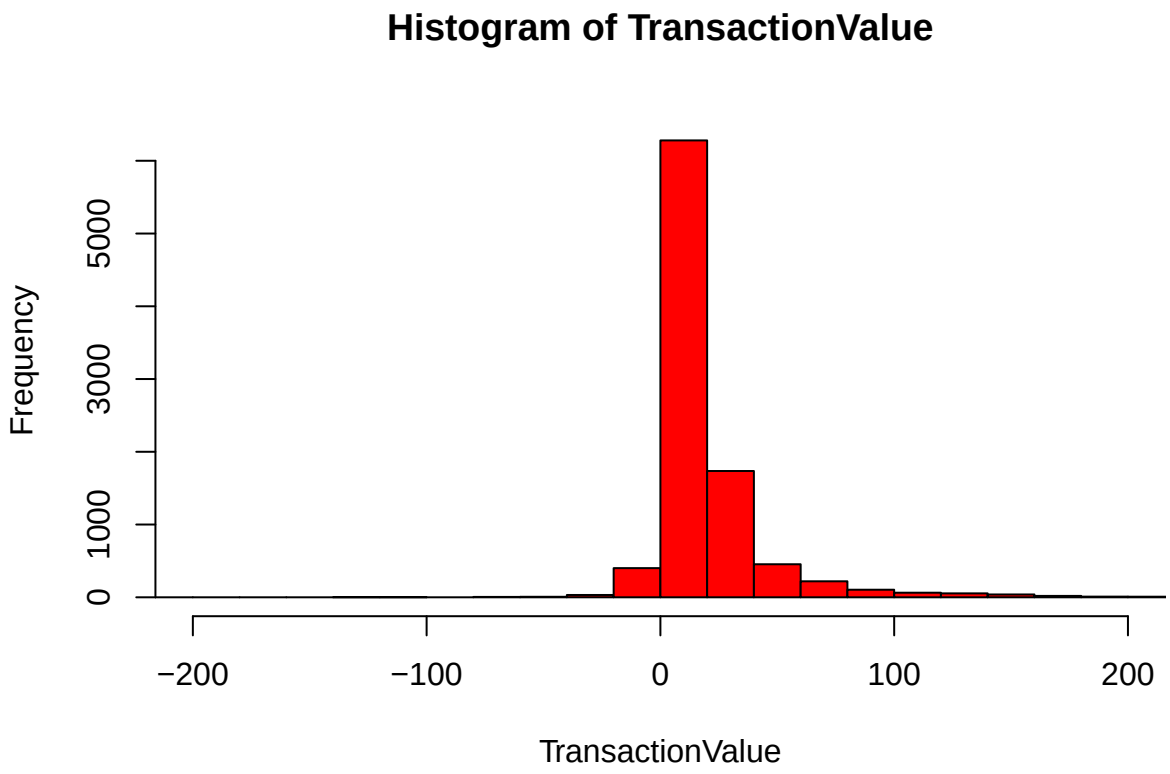
```
## [1] "The consecutive 2 hours where the downtime can be done is:"
```

```
Hour1[c(Hour2,Hour2+1),1]
```

```
## [1] 19 20
```

Question 5

```
Online_retail%>%
  filter(Country=='Germany')  %>%
  with(hist(TransactionValue,breaks=100,ylim=c(0,6500),xlim=c(-200,200),col='red'))
```



Question 6

```
Online_retail%>% group_by(CustomerID)%>%
  summarise(max_num_trans=n())%>%
  filter(!is.na(CustomerID))%>%
  filter(max_num_trans==max(max_num_trans))
```

```
## # A tibble: 1 x 2
```

```
##   CustomerID max_num_trans
##        <int>         <int>
## 1      17841          7983
```

```
Online_retail%>% group_by(CustomerID)%>%
  filter(!is.na(CustomerID))%>%
  summarise(max_sum_trans=sum(TransactionValue))%>%
  filter(max_sum_trans==max(max_sum_trans))
```

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

Question 7

```
missing_values<-as.data.frame(Online_retail%>%
                                sapply(function(x) sum(is.na(x))))
colnames(missing_values)[1] ="Percentage"
(missing_values/nrow(Online_retail))*100
```

```
##                   Percentage
## InvoiceNo            0.00000
## StockCode            0.00000
## Description          0.00000
## Quantity             0.00000
## InvoiceDate          0.00000
## UnitPrice            0.00000
## CustomerID          24.92669
## Country              0.00000
## TransactionValue     0.00000
## New_Invoice_Date     0.00000
## Invoice_Day_Week     0.00000
## New_Invoice_Hour     0.00000
## New_Invoice_Month    0.00000
```

Question 8

```
Online_retail%>%
  filter(is.na(CustomerID)) %>%
  group_by(Country) %>%
  summarise(no_of_missing=n())
```

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

```
print(paste("The average days difference between customer transactions is",(Online_retail%>%
  group_by(New_Invoice_Date) %>%
  summarise(n())%>%
  summarise(Average_diff_between_transaction=mean(diff(New_Invoice_Date))))))
```

```
## [1] "The average days difference between customer transactions is 1.22697368421053"
```

Question 10

```
Online_retail%>% group_by(Country) %>% filter(Quantity<0 & Country=='France')%>%
  summarise(return_rate=n()/nrow(Online_retail))
```

```
## # A tibble: 1 x 2
##   Country return_rate
##   <chr>         <dbl>
## 1 France     0.000275
```

Question 11

```
print("The product which is purchased maximum is:")
```

```
## [1] "The product which is purchased maximum is:"
```

```
Online_retail%>% group_by(Description) %>% summarise(product_Value=(sum(TransactionValue)))%>%
  filter(product_Value==max(product_Value))
```

```
## # A tibble: 1 x 2
##   Description    product_Value
##   <chr>                  <dbl>
## 1 DOTCOM POSTAGE       206245.
```

Question 12

```
Customer<-Online_retail%>%filter(!is.na(CustomerID))
       print(paste("No. of Unique customers are: ",(length(unique(Customer$CustomerID)))))
```

```
## [1] "No. of Unique customers are:  4372"
```