Business Analytics-Assignment-3

shiva gadila

2023-03-12

#Importing the Dataset

Online.Retail<- read.csv("~/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)  
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

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 country, or how many transactions are in each country’s dataset (take into account all records, including transactions that were cancelled). Give this as a percentage and a total number. Only countries that account for more than one per cent of all transactions should be shown.

Online.Retail %>% group\_by(Country) %>% summarise(n())

## # A tibble: 38 × 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 × 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)) %>% filter(Country>0.01)

## # A tibble: 38 × 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 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.

Online.Retail%>%group\_by(Country) %>% summarise(Sum\_of\_Transaction\_values = sum(TransactionValue)) %>% filter(Sum\_of\_Transaction\_values > 130000)

## # A tibble: 6 × 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 4:

Retail<- strptime(Online.Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')  
head(Retail)

## [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"

Online.Retail$New\_Invoice\_Date<-as.Date(Retail)  
  
Online.Retail$Invoice\_Day\_week <- weekdays(Online.Retail$New\_Invoice\_Date)  
Online.Retail$New\_Invoice\_Hour <-as.numeric (format(Retail,"%H"))  
Online.Retail$New\_Invoice\_Month <- as.numeric(format(Retail, "%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

#A.Show the percentage of transactions (by numbers) by days of the week

Online.Retail %>%  
group\_by(Invoice\_Day\_week) %>%  
tally(sort=TRUE) %>%  
summarise(Invoice\_Day\_week,TransactionCounts = n ,percent = n/sum(n)\*100) %>%  
arrange(desc(TransactionCounts))

## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in  
## dplyr 1.1.0.  
## ℹ Please use `reframe()` instead.  
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`  
## always returns an ungrouped data frame and adjust accordingly.

## # A tibble: 6 × 3  
## Invoice\_Day\_week TransactionCounts percent  
## <chr> <int> <dbl>  
## 1 Thursday 103857 19.2  
## 2 Tuesday 101808 18.8  
## 3 Monday 95111 17.6  
## 4 Wednesday 94565 17.5  
## 5 Friday 82193 15.2  
## 6 Sunday 64375 11.9

#B.Show the percentage of transactions (by transaction volume) by days of the week

Online.Retail %>%  
group\_by(Invoice\_Day\_week) %>%  
summarise(TransValueSum = sum(TransactionValue)) %>%  
mutate(TransValuepercent= TransValueSum/sum(TransValueSum))%>%  
arrange(desc(TransValueSum))

## # A tibble: 6 × 3  
## Invoice\_Day\_week TransValueSum TransValuepercent  
## <chr> <dbl> <dbl>  
## 1 Thursday 2112519. 0.217   
## 2 Tuesday 1966183. 0.202   
## 3 Wednesday 1734147. 0.178   
## 4 Monday 1588609. 0.163   
## 5 Friday 1540611. 0.158   
## 6 Sunday 805679. 0.0827

#C. Show the percentage of transactions (by transaction volume) by month of the year

Online.Retail %>%  
group\_by(New\_Invoice\_Month) %>%  
summarise(TransValueSum = sum(TransactionValue)) %>%  
mutate(TransValuePercent=TransValueSum/sum(TransValueSum)) %>%  
arrange(desc(TransValuePercent))

## # A tibble: 12 × 3  
## New\_Invoice\_Month TransValueSum TransValuePercent  
## <dbl> <dbl> <dbl>  
## 1 11 1461756. 0.150   
## 2 12 1182625. 0.121   
## 3 10 1070705. 0.110   
## 4 9 1019688. 0.105   
## 5 5 723334. 0.0742  
## 6 6 691123. 0.0709  
## 7 3 683267. 0.0701  
## 8 8 682681. 0.0700  
## 9 7 681300. 0.0699  
## 10 1 560000. 0.0574  
## 11 2 498063. 0.0511  
## 12 4 493207. 0.0506

#D. What was the date with the highest number of transactions from Australia?

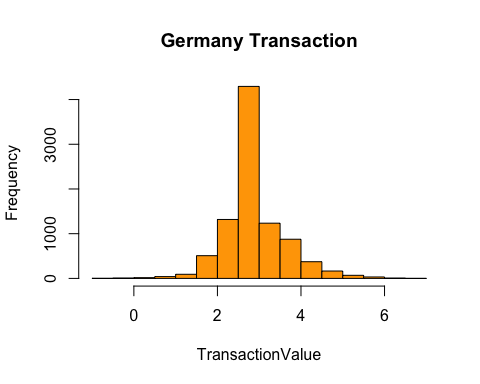
Online.Retail %>%  
filter(Country == "Australia") %>%  
group\_by(InvoiceDate) %>%  
tally(sort = TRUE) %>%  
filter(n == max(n))

## # A tibble: 1 × 2  
## InvoiceDate n  
## <chr> <int>  
## 1 6/15/2011 13:37 139

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

hist(x=log(Online.Retail$TransactionValue[Online.Retail$Country=="Germany"]),xlab = "TransactionValue",col = 'orange' ,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 who carried out the most transactions.  
Online.Retail %>%group\_by(CustomerID)%>%summarise(CustomerTransaction = n())%>%filter(CustomerID != "NA")%>%filter(CustomerTransaction ==max(CustomerTransaction))

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

#The customer with the highest total transaction sum and highest value.  
Online.Retail%>%group\_by(CustomerID)%>%summarise(total.transaction.by.each.customer = sum(TransactionValue))%>%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 × 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 New\_Invoice\_Date Invoice\_Day\_week New\_Invoice\_Hour   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## New\_Invoice\_Month   
## 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 × 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

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

Online.Retail %>%  
select(CustomerID, New\_Invoice\_Date) %>%  
group\_by(CustomerID) %>%  
distinct(New\_Invoice\_Date) %>%  
arrange(desc(CustomerID)) %>%  
mutate(DaysBetween = New\_Invoice\_Date - lag(New\_Invoice\_Date))->  
custDaysBtwVisit  
  
custDaysBtwVisit %>%  
filter(!is.na(DaysBetween)) -> RetcustDaysBtwVisits  
mean(RetcustDaysBtwVisits$DaysBetween)

## Time difference of 38.4875 days

#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')&(TransactionValue<0)&(CustomerID != 'Na')))  
Totalfrenchcustomer<-nrow(Online.Retail%>%group\_by(CustomerID)%>%filter((Country=='France')&(CustomerID != '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(TransactionValue))%>%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 × 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

#There are 4373 unique customers represented in the dataset.