Online retailing

shivani

2022-10-30

***Installing Required Packages***

library(dplyr)  
library(zoo)  
set.seed(150)  
library(readxl)  
Online.Retail=read.csv("C:/Users/shiva/Downloads/Online\_Retail.csv")

***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(transactions=n())%>%mutate(percentage=(transactions/541909)\*100)%>%arrange(desc(transactions))%>%filter(data<-percentage>1)

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

Online.Retail<- mutate(Online.Retail, "TransactionValue"=TransactionValue<- Online.Retail$Quantity \* Online.Retail$UnitPrice)  
colnames(Online.Retail)

***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(total.sum.of.transaction.values=sum(TransactionValue))%>%arrange(desc(total.sum.of.transaction.values))%>%filter(total.sum.of.transaction.values>130000)

***4***

#creating a POSIXltobject from "InvoiceDate":  
Temp=strptime(Online.Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')  
  
#spliting the dataframe's components for the date, day of the week, and hour under the labels New Invoice Date, Invoice Day Week, and New Invoice Hour:  
Online.Retail$New\_Invoice\_Date<-as.Date(Temp)  
  
#Knowing two date values gives you the ability to determine how many days between the two dates.  
Online.Retail$New\_Invoice\_Date[20000]-Online.Retail$New\_Invoice\_Date[10]  
  
#Dates can be converted to weekdays. For that, let's create a new variable.  
Online.Retail$Invoice\_Day\_Week=weekdays(Online.Retail$New\_Invoice\_Date)  
  
#Let's just turn the hour into a standard numerical value for the hour (ignore the minute):  
Online.Retail$New\_Invoice\_Hour =as.numeric(format(Temp,"%H"))  
  
#defining the month as a separate numeric variable too:  
Online.Retail$New\_Invoice\_Month = as.numeric(format(Temp, "%m"))

***Now answer the flowing questions*** ***4.a) Show the percentage of transactions (by numbers) by days of the week (extra 1% of total points)***

Online.Retail%>%  
 group\_by(Invoice\_Day\_Week)%>%  
 summarise(Number.of.transaction=(n()))%>%  
 mutate(Number.of.transaction,'percent'=(Number.of.transaction\*100)/sum(Number.of.transaction))

***4.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(TransactionValue)))%>%  
 mutate(Volume.of.transaction,'percent'=(Volume.of.transaction\*100)/sum(Volume.of.transaction))

***4.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))

***4.d)What was the date with the highest number of transactions from Australia?***

Online.Retail <- Online.Retail %>% mutate(Transactionvalue= Quantity \* UnitPrice)  
Online.Retail %>% filter(Country == 'Australia') %>% group\_by(New\_Invoice\_Date) %>% summarise(max=max(Transactionvalue))

***4.e)The company needs to shut down the website for twoconsecutivehours 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***

H<-summarise(group\_by(Online.Retail,New\_Invoice\_Hour),Transaction\_min=n\_distinct(InvoiceNo))  
H<-filter(H,New\_Invoice\_Hour>=7&New\_Invoice\_Hour<=20)  
H2<-rollapply(H$Transaction\_min,2,sum)  
H3<-which.min(H2)  
H3

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

hist(x=log(Online.Retail$TransactionValue[Online.Retail$Country=="Germany"]),xlab = "Transaction Value of Germany",col = 'Dark green' ,main = 'Germany Transaction',ylab = 'Frequency')

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

OnlineCustomer <- na.omit(Online.Retail)  
result.data <- summarise(group\_by(OnlineCustomer,CustomerID), sum.data= sum(Transactionvalue))  
result.data[which.max(result.data$sum.data),]  
  
Customer.data <- table(Online.Retail$CustomerID)  
Customer.data <- as.data.frame(Customer.data)  
result.data.2 <- Customer.data[which.max(Customer.data$Freq),]  
result.data.2

#With a total of 7983 transactions, CustomerID 17841 led all other customers in terms of transaction volume.and #The most valuable customer is CustomerID 14646, who spent the most money (279,489.020 British Sterling Pounds).

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

missing.values<-colMeans(is.na(Online.Retail))  
print(paste('Online customerID column in dataset lacks values i.e.',missing.values['CustomerID']\*100,'% of whole data'))

***8.What are the number of transactions with missing CustomerID records by countries?***

Online.Retail%>%group\_by(Country)%>%filter(is.na(CustomerID))%>%summarise(No.of.missing.CustomerID=n())  
#The United Kingdom has the most NA records, with 133,600 rows, out of the eight nations and one unnamed country that had NA values in the dataset.

***9. On average, how often the costumers comeback to the website for their next shopping? (i.e. what is the average number of days between consecutive shopping)***

average<-Online.Retail%>%group\_by(CustomerID)%>%summarise(difference.in.consecutivedays=diff(New\_Invoice\_Date))%>%filter(difference.in.consecutivedays>0)  
print(paste('the average number of days between consecutive shopping is',mean(average$difference.in.consecutivedays)))

***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?***

return.value<-nrow(Online.Retail%>%group\_by(CustomerID)%>%filter((Country=='France')&(TransactionValue<0)&(CustomerID != 'Na')))  
total.fcustomer<-nrow(Online.Retail%>%group\_by(CustomerID)%>%filter((Country=='France')&(CustomerID != 'Na')))  
print(paste('For French customers, the return rate is provided as',((return.value)/(total.fcustomer))\*100,'Percent'))

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

TransactionValue <- tapply(Online.Retail$TransactionValue, Online.Retail$StockCode , sum)  
TransactionValue[which.max(TransactionValue)]

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

unique.customers <- unique(Online.Retail$CustomerID)  
length(unique.customers)