Assignment 1

Anuraag Vasal

10/22/2021

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(readr)  
Retailonline <- read.csv("Online\_Retail.csv")

#The breakdown of the number of transactions by countries

Retailonline\_country <- as.data.frame(table(Retailonline$Country))  
Retailonline\_country["Percentage"] <- Retailonline\_country$Freq/NROW(Retailonline) \* 100  
names(Retailonline\_country) <- c("Country","Count","Percentage")  
Retailonline\_country[Retailonline\_country$Percentage > 1,]

## Country Count Percentage  
## 11 EIRE 8196 1.512431  
## 14 France 8557 1.579047  
## 15 Germany 9495 1.752139  
## 36 United Kingdom 495478 91.431956

#Countries accounting for more than 1% of the total transactions EIRE,France,Germany and United Kingdom are the 4 countries that accounts for more than 1 % of total transactions.

#Adding new variable ‘TransactionValue’ which is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables.

Retailonline["TransactionValue"] <- Retailonline$Quantity \* Retailonline$UnitPrice  
head(Retailonline)

## 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

#Using the newly created variable, TransactionValue, showing the breakdown of transaction values by countries.

Retailonline%>%group\_by(Country)%>%summarise(Total=sum(TransactionValue))

## # A tibble: 38 x 2  
## Country Total  
## <chr> <dbl>  
## 1 Australia 137077.  
## 2 Austria 10154.  
## 3 Bahrain 548.  
## 4 Belgium 40911.  
## 5 Brazil 1144.  
## 6 Canada 3666.  
## 7 Channel Islands 20086.  
## 8 Cyprus 12946.  
## 9 Czech Republic 708.  
## 10 Denmark 18768.  
## # ... with 28 more rows

Retailonline\_transactionwise <- Retailonline%>%group\_by(Country)%>%summarise(Total=sum(TransactionValue))

#Countries with total transaction exceeding 130,000 British Pound.

Retailonline\_transactionwise[Retailonline\_transactionwise$Total>130000,]

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

#Converting ‘InvoiceDate’ into a POSIXlt object

Temp <- strptime(Retailonline $ InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')  
Retailonline$New\_Invoice\_Date <- as.Date(Temp)  
Retailonline$New\_Invoice\_Date[20000]- Retailonline$New\_Invoice\_Date[10]

## Time difference of 8 days

Retailonline$Invoice\_Day\_Week= weekdays(Retailonline$New\_Invoice\_Date)  
Retailonline$New\_Invoice\_Hour = as.numeric(format(Temp, "%H"))  
Retailonline$New\_Invoice\_Month = as.numeric(format(Temp, "%m"))

#Percentage of transactions (by numbers) by days of the week:

Retailonline%>%group\_by(Invoice\_Day\_Week)%>%summarise(count=n())%>%mutate(percentage=count/nrow(Retailonline)\* 100)

## # A tibble: 6 x 3  
## Invoice\_Day\_Week count percentage  
## <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

#The percentage of transactions (by transaction volume) by days of the week:

Retailonline%>%group\_by(Invoice\_Day\_Week)%>%summarise(Total=sum(TransactionValue))%>%mutate(percentage=Total/sum(Total)\*100)

## # A tibble: 6 x 3  
## Invoice\_Day\_Week Total percentage  
## <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

#The percentage of transactions (by transaction volume) by month of the year

Retailonline%>%group\_by(New\_Invoice\_Month)%>%summarise(Total=sum(TransactionValue))%>%mutate(percentage=Total/sum(Total)\*100)

## # A tibble: 12 x 3  
## New\_Invoice\_Month Total percentage  
## <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  
## 7 7 681300. 6.99  
## 8 8 682681. 7.00  
## 9 9 1019688. 10.5   
## 10 10 1070705. 11.0   
## 11 11 1461756. 15.0   
## 12 12 1182625. 12.1

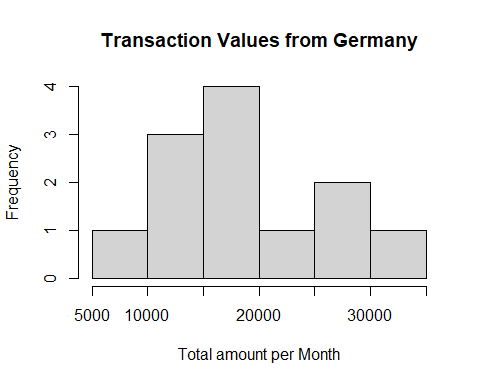
#The date with the highest number of transactions from Australia

Retailonline%>%  
filter(Country=="Australia")%>%  
group\_by(New\_Invoice\_Date)%>%  
tally(sort = TRUE)%>%   
filter(n==max(n))

## # A tibble: 1 x 2  
## New\_Invoice\_Date n  
## <date> <int>  
## 1 2011-06-15 139

#The histogram of transaction values from Germany.

Retailonline%>%  
group\_by(Country)%>%  
filter(Country=="Germany")%>%  
group\_by(New\_Invoice\_Month)%>%  
summarise(Total = sum(TransactionValue))-> Germany  
hist(Germany$Total,main = "Transaction Values from Germany", xlab = "Total amount per Month", ylab = "Frequency")

 #Customer having the highest number of transactions:

Retailonline%>%  
group\_by(CustomerID)%>%  
tally(sort = TRUE)%>%  
filter(!is.na(CustomerID))%>%   
filter(n==max(n))

## # A tibble: 1 x 2  
## CustomerID n  
## <int> <int>  
## 1 17841 7983

Retailonline%>%  
group\_by(CustomerID)%>%  
summarise(Total=sum(TransactionValue))%>%  
filter(!is.na(CustomerID))%>%   
filter(Total == max(Total))

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

Customer with CustomerID 17841 has the highest number of transactions,and most valuable customer is CustomerID 14646 with spending of 279,489 Pounds.

#The percentage of missing values for each variable in the dataset:

colMeans(is.na(Retailonline) \*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 New\_Invoice\_Date Invoice\_Day\_Week New\_Invoice\_Hour   
## 0.00000 0.00000 0.00000 0.00000   
## New\_Invoice\_Month   
## 0.00000

Here , we can see that the missing values are present in the column of Customer ID and Description. The percentage of missing values in Customer ID column is 24.92669 % and for Description it is 0.2683107 %.

#The total number of transactions with missing CustomerID records by countries :

NROW(Retailonline[is.na(Retailonline$CustomerID),])

## [1] 135080

The total number of rows in Customer ID column which have missing value is 135080

Retailonline[is.na(Retailonline$CustomerID),] %>% group\_by(Country) %>% summarise(count = n())

## # A tibble: 9 x 2  
## Country count  
## <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

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

## Time difference of 38.4875 days

#The return rate of the goods purchased by customers from France:

Retailonline\_france\_cancel <- Retailonline%>%filter(Country=="France",Quantity<0)%>%count  
Retailonline\_france\_total <- Retailonline%>%filter(Country=="France")%>%count  
Retailonline\_ratio <- Retailonline\_france\_cancel$n / Retailonline\_france\_total$n  
Retailonline\_ratio \* 100

## [1] 1.741264

The return rate for Customers of France is 1.741264 %

#The product that has generated the highest revenue for the retailer:

Retailonline%>%group\_by(Description)%>%summarise(Total=sum(TransactionValue))%>%arrange((desc(Total)))

## # A tibble: 4,224 x 2  
## Description Total  
## <chr> <dbl>  
## 1 "DOTCOM POSTAGE" 206245.  
## 2 "REGENCY CAKESTAND 3 TIER" 164762.  
## 3 "WHITE HANGING HEART T-LIGHT HOLDER" 99668.  
## 4 "PARTY BUNTING" 98303.  
## 5 "JUMBO BAG RED RETROSPOT" 92356.  
## 6 "RABBIT NIGHT LIGHT" 66757.  
## 7 "POSTAGE" 66231.  
## 8 "PAPER CHAIN KIT 50'S CHRISTMAS " 63792.  
## 9 "ASSORTED COLOUR BIRD ORNAMENT" 58960.  
## 10 "CHILLI LIGHTS" 53768.  
## # ... with 4,214 more rows

Doctom Postage is the product that has generated the highest revenue for the retailer.

#Unique customers are represented in the dataset

Retailonline\_selectCustID <- Retailonline%>%select(CustomerID)%>%unique %>% nrow  
Retailonline\_selectCustID

## [1] 4373

The total number of unique customers in the dataset is 4373