Online\_Retail\_Yash

Yash\_Patel

2022-10-31

# loading the package  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

# adding dataset  
or <- read.csv("C:/Users/YASH/Downloads/Online\_Retail.csv")

# Q1.

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.

#grouped by the countries and showing the countries with more than 1% of total transaction  
or\_1 <- group\_by(or,Country) %>% summarise( Total\_Transactions=length(InvoiceNo)) %>% filter(Total\_Transactions > ((nrow(or)) \* 1/100))  
or\_1

## # A tibble: 4 × 2  
## Country Total\_Transactions  
## <chr> <int>  
## 1 EIRE 8196  
## 2 France 8557  
## 3 Germany 9495  
## 4 United Kingdom 495478

# Q2.

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

#adding transaction value column  
or\_2 <- mutate(or,TransactionValue= Quantity \* UnitPrice) #%>% select(Quantity,UnitPrice,Description) %>% filter( UnitPrice < 0)  
head(select(or\_2, TransactionValue))

## TransactionValue  
## 1 15.30  
## 2 20.34  
## 3 22.00  
## 4 20.34  
## 5 20.34  
## 6 15.30

# Q3.

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.

#adding TotalTransactionValue column and filtering it for the transaction greater than 130000  
or\_3 <- group\_by(or\_2,Country) %>% summarise(TotalTransactionValue=sum(TransactionValue)) %>% filter(TotalTransactionValue>130000)  
or\_3

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

# Q4.

Temp=strptime(or$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')   
#head(Temp)  
  
or$New\_Invoice\_Date <- as.Date(Temp)  
#or$New\_Invoice\_Date[20000]- or$New\_Invoice\_Date[10]   
or$Invoice\_Day\_Week= weekdays(or$New\_Invoice\_Date)   
or$New\_Invoice\_Hour = as.numeric(format(Temp, "%H"))   
or$New\_Invoice\_Month = as.numeric(format(Temp, "%m"))   
  
# a) Show the percentage of transactions (by numbers) by days of the week   
  
or\_4.1 <- group\_by(or,Invoice\_Day\_Week) %>% drop\_na() %>% count()  
or\_4.1$perc <- (or\_4.1$n/sum(or\_4.1$n)) \* 100   
or\_4.1

## # A tibble: 6 × 3  
## # Groups: Invoice\_Day\_Week [6]  
## Invoice\_Day\_Week n perc  
## <chr> <int> <dbl>  
## 1 Friday 56127 13.8  
## 2 Monday 66382 16.3  
## 3 Sunday 63237 15.5  
## 4 Thursday 82374 20.2  
## 5 Tuesday 68110 16.7  
## 6 Wednesday 70599 17.4

# b) Show the percentage of transactions (by transaction volume) by days of the week   
  
or\_4.2 <- group\_by(or,Invoice\_Day\_Week) %>% select(Invoice\_Day\_Week,Quantity) %>% drop\_na() %>% count()  
or\_4.2$perc <- (or\_4.2$n/sum(or\_4.2$n)) \* 100   
or\_4.2

## # A tibble: 6 × 3  
## # Groups: Invoice\_Day\_Week [6]  
## Invoice\_Day\_Week n perc  
## <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

# c) Show the percentage of transactions (by transaction volume) by month of the year   
  
or\_4.3 <- group\_by(or,New\_Invoice\_Month) %>% select(New\_Invoice\_Month, Quantity) %>% drop\_na() %>% count()  
or\_4.3$perc <- (or\_4.3$n/sum(or\_4.3$n)) \* 100   
or\_4.3

## # A tibble: 12 × 3  
## # Groups: New\_Invoice\_Month [12]  
## New\_Invoice\_Month n perc  
## <dbl> <int> <dbl>  
## 1 1 35147 6.49  
## 2 2 27707 5.11  
## 3 3 36748 6.78  
## 4 4 29916 5.52  
## 5 5 37030 6.83  
## 6 6 36874 6.80  
## 7 7 39518 7.29  
## 8 8 35284 6.51  
## 9 9 50226 9.27  
## 10 10 60742 11.2   
## 11 11 84711 15.6   
## 12 12 68006 12.5

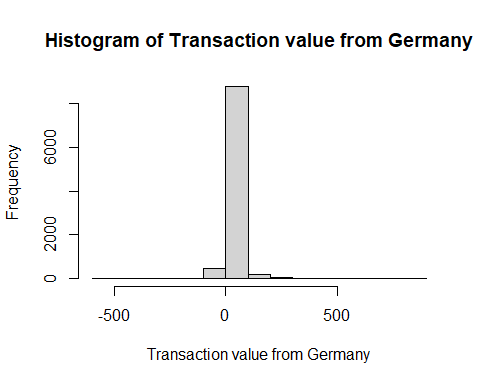
# d) What was the date with the highest number of transactions from Australia?   
or\_4.4 <- group\_by(or, New\_Invoice\_Date) %>% drop\_na()%>% select(New\_Invoice\_Date, Country) %>% filter( Country == "Australia" ) %>% count()  
max3 <- max(or\_4.4$n)  
or\_4.4.1 <- or\_4.4 %>% filter(n == max3)  
or\_4.4.1

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

# Q5.

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

#filtering or\_2 dataset to select Germany as Country and then selecting Country TransactionValue column   
or\_5 <- filter(or\_2, Country == "Germany" ) %>% select(Country, TransactionValue)  
  
#assigning variable for histogram   
"Transaction value from Germany" <- or\_5$TransactionValue  
  
#creating histogram using hist function  
hist(`Transaction value from Germany`)



# Q6.

1.Which customer had the highest number of transactions? 2.Which customer is most valuable (i.e. highest total sum of transactions)?

#1  
#counting the total transaction by CustomerID and droping the observation without ColumnID  
or\_6.0 <- group\_by(or\_2,CustomerID) %>% summarise( Total\_Transactions=length(TransactionValue)) %>% drop\_na()  
  
#finding the largest number  
max0<- max(or\_6.0$Total\_Transactions)  
  
#filtering the data to get the CustomerID with Maximum transaction  
or\_6.0\_max <- filter(or\_6.0,Total\_Transactions== max0)  
  
or\_6.0\_max

## # A tibble: 1 × 2  
## CustomerID Total\_Transactions  
## <int> <int>  
## 1 17841 7983

#2  
#totaling the TotalTransactionValue by CustomerID and droping the observation without ColumnID  
or\_6.1 <- group\_by(or\_2,CustomerID) %>% summarise(TotalTransactionValue=sum(TransactionValue)) %>% drop\_na()  
  
#finding the largest number  
max1 <- max(or\_6.1$TotalTransactionValue)  
  
#filtering the data to get the CustomerID with Maximum TotalTransactionValue   
or\_6.1\_max <- filter(or\_6.1,TotalTransactionValue== max1)  
  
or\_6.1\_max

## # A tibble: 1 × 2  
## CustomerID TotalTransactionValue  
## <int> <dbl>  
## 1 14646 279489.

# Q7.

Calculate the percentage of missing values for each variable in the dataset

colMeans(is.na(or\_2)) #25%(24.92669)

## 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   
## 0.0000000

#or  
  
summary(or)

## 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   
## New\_Invoice\_Date Invoice\_Day\_Week New\_Invoice\_Hour New\_Invoice\_Month  
## Min. :2010-12-01 Length:541909 Min. : 6.00 Min. : 1.000   
## 1st Qu.:2011-03-28 Class :character 1st Qu.:11.00 1st Qu.: 5.000   
## Median :2011-07-19 Mode :character Median :13.00 Median : 8.000   
## Mean :2011-07-04 Mean :13.08 Mean : 7.553   
## 3rd Qu.:2011-10-19 3rd Qu.:15.00 3rd Qu.:11.000   
## Max. :2011-12-09 Max. :20.00 Max. :12.000   
##

#only ColumnID variable has missing value  
  
or\_7 <- (135080\*100)/541909  
  
or\_7 # ~25%(24.92669)

## [1] 24.92669

# Q8.

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

#Filtering the Missing CustomeID records and Counting it by Countries under new Column "Total\_Transaction"  
or\_8 <- group\_by(or\_2,Country ) %>% filter(is.na(CustomerID),Country !="Unspecified") %>% summarise( Total\_Transactions=length(CustomerID))  
  
#First Arranging the Total\_Transaction in descending order and adding Percentage column showing Total\_Transaction in Percentage.  
or\_8.1 <- or\_8 %>% arrange(desc(Total\_Transactions)) %>% mutate(percentage = Total\_Transactions / sum(Total\_Transactions) \* 100)  
or\_8.1

## # A tibble: 8 × 3  
## Country Total\_Transactions percentage  
## <chr> <int> <dbl>  
## 1 United Kingdom 133600 99.1   
## 2 EIRE 711 0.527   
## 3 Hong Kong 288 0.214   
## 4 Switzerland 125 0.0927   
## 5 France 66 0.0489   
## 6 Israel 47 0.0348   
## 7 Portugal 39 0.0289   
## 8 Bahrain 2 0.00148

# Q9.

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)

or\_9<- or %>% select(CustomerID,New\_Invoice\_Date) %>% group\_by(CustomerID) %>% distinct(New\_Invoice\_Date) %>% arrange(desc(CustomerID)) %>% mutate(Days\_Between = New\_Invoice\_Date-lag(New\_Invoice\_Date)) %>% filter(!is.na(Days\_Between))   
mean(or\_9$Days\_Between)

## Time difference of 38.4875 days

# Q10.

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? Consider the cancelled transactions as those where the ‘Quantity’ variable has a negative value

#Return rate for the French customers  
  
#counting the total number of transaction   
x.0 <-filter(or\_2,Country == "France") %>% count()  
#counting the transaction with negative quantity  
y.0 <-filter(or\_2,Country == "France",Quantity < 0) %>% count()  
#finally finding the return rate for the french customers dividing the number of transactions cancelled over the total number of transactions.   
or\_10 <-summarise(or\_2, Return\_rate\_for\_the\_French\_customers= y.0/ x.0 \* 100)  
or\_10 #(1.741264)

## n  
## 1 1.741264

# Q11.

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

#grouping the dataset by StockCode and using summarise function to find the Total Transaction Value  
  
or\_11 <- group\_by(or\_2, StockCode) %>% summarise( Total\_Transac\_val = sum(TransactionValue))  
#finding the largest Value  
max2 <- max(or\_11$Total\_Transac\_val)  
#filtering the dataset to get the Stockcode which has the largest Value  
or\_11.1 <- filter(or\_11,Total\_Transac\_val==max2)  
or\_11.1

## # A tibble: 1 × 2  
## StockCode Total\_Transac\_val  
## <chr> <dbl>  
## 1 DOT 206245.

# Q12.

How many unique customers are represented in the dataset?

#finding the unique customer using the unique() function and droping the empty observations and then counting the actual unique customers using the count function.  
or\_12 <- data.frame(unique(or\_2$CustomerID)) %>% drop\_na() %>% count("Total\_Unique\_customers")  
or\_12

## "Total\_Unique\_customers" n  
## 1 Total\_Unique\_customers 4372