Online_Retail_Yash

Yash_Patel

2022-10-31

```
# loading the package
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()
```

```
# 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_Tran
or_1
```

```
## # A tibble: 4 x 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) %
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(To
or_3
```

```
## # A tibble: 6 x 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 x 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 x 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() %>%  cou
or_4.3$perc <- (or_4.3$n/sum(or_4.3$n)) * 100
or_4.3
```

```
## # A tibble: 12 x 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(
max3 <- max(or_4.4$n)
or_4.4.1 <- or_4.4 %>% filter(n == max3)
or_4.4.1
```

```
## # A tibble: 1 x 2
## # Groups:   New_Invoice_Date [1]
##   New_Invoice_Date     n
##   <date>           <int>
## 1 2011-06-15         139
```

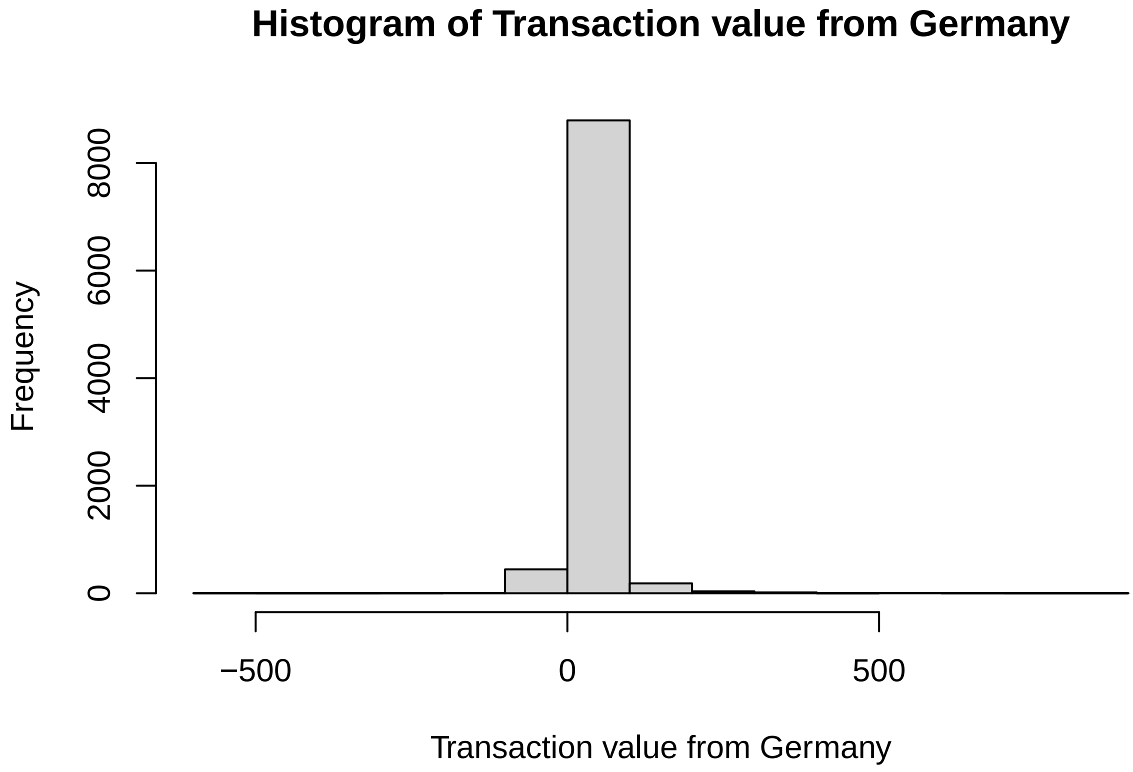
Q5.

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

#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 x 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

#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 x 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_Transacti
or_8 <- group_by(or_2,Country ) %>% filter(is.na(CustomerID),Country !="Unspecified") %>% summarise( To

#First Arranging the Total_Transaction in descending order and adding Percentage column showing Total_T
or_8.1 <- or_8 %>% arrange(desc(Total_Transactions)) %>% mutate(percentage = Total_Transactions / sum(T
or_8.1
```

```
## # A tibble: 8 x 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
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 cancell
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 x 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 co
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
```