BA ASSIGNMENT2

DIVYA CHANDRASEKARAN_811284790

2023-10-14

For this assignment, you need to use the ‘Online Retail’ dataset which can be downloaded in CSV format from the Dataset folder. This is a transnational data set which contains all the transactions occurring between 01 Dec 2010 and 09 Dec 2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The data contains the following attributes:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

Download the dataset, and use the read.csv() command to load the file into an R data frame and answer the following questions.

```
#Load the dataset.
```

```
library(readr)
Online_Retail <- read_csv("Online_Retail.csv")
```

```
## Rows: 541909 Columns: 8
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr (5): InvoiceNo, StockCode, Description, InvoiceDate, Country
## dbl (3): Quantity, UnitPrice, CustomerID
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
View(Online_Retail)
```

```
#Creating summary for the dataset
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
```

QUESTION1

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.

```
#1. Transactions by country
transaction.count <- table(Online_Retail$Country)
total.transaction <- sum(transaction.count)
transaction.percent <- (transaction.count/total.transaction) * 100

#Show countries with more than 1 percent of total transaction
country.summary <- data.frame(
Country = names(transaction.count),
TransactionCount = as.numeric(transaction.count),
TransactionPercentage = as.numeric(transaction.percent))
significant_countries <- subset(country.summary, transaction.percent > 1)
#country.summary <- transaction.count[transaction.percent > 1, ]
#names(country_trans) <- c("Transactions", "Percentage")
print(country.summary)
```

```
##                 Country TransactionCount TransactionPercentage
## 1             Australia             1259           0.232326830
## 2               Austria              401           0.073997664
## 3               Bahrain               19           0.003506124
## 4               Belgium             2069           0.381798420
## 5                Brazil               32           0.005905050
## 6                Canada              151           0.027864457
## 7       Channel Islands              758           0.139875883
## 8                Cyprus              622           0.114779419
## 9        Czech Republic               30           0.005535985
## 10              Denmark              389           0.071783270
## 11                 EIRE             8196           1.512431054
## 12   European Community               61           0.011256502
## 13              Finland              695           0.128250315
## 14               France             8557           1.579047405
## 15              Germany             9495           1.752139197
## 16               Greece              146           0.026941793
## 17            Hong Kong              288           0.053145454
## 18              Iceland              182           0.033584975
## 19               Israel              297           0.054806250
## 20                Italy              803           0.148179860
## 21                Japan              358           0.066062752
## 22              Lebanon               45           0.008303977
## 23            Lithuania               35           0.006458649
## 24                Malta              127           0.023435669
## 25          Netherlands             2371           0.437527334
## 26               Norway             1086           0.200402651
## 27               Poland              341           0.062925694
```

```
## 28              Portugal             1519          0.280305365
## 29                   RSA               58          0.010702904
## 30          Saudi Arabia               10          0.001845328
## 31             Singapore              229          0.042258017
## 32                 Spain             2533          0.467421652
## 33                Sweden              462          0.085254166
## 34           Switzerland             2002          0.369434721
## 35 United Arab Emirates               68          0.012548232
## 36        United Kingdom           495478         91.431956288
## 37           Unspecified              446          0.082301641
## 38                   USA              291          0.053699053
```

QUESTION2

Create a new variable 'TransactionValue' that is the product of the existing 'Quantity' and 'UnitPrice' variables. Add this variable to the data frame.

```
# 2. Create TransactionValue variable
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
```

```
Online_Retail <- Online_Retail %>%
  mutate(TransactionValue = UnitPrice * Quantity)
```

QUESTION 3

Using the newly created variable, TransactionValue shows the breakdown of transaction values by country i.e. how much money has been spent in each country. Show this in the total sum of transaction values. Show only countries with total transactions exceeding 130,000 British Pounds.

```
# 3. Transaction value by country
transaction.summary <- Online_Retail %>%
  group_by(Country) %>%
  summarise(total.transaction.value = sum(TransactionValue))
transaction.summary
```

```
## # A tibble: 38 × 2
##    Country         total.transaction.value
##    <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.
## # ℹ 28 more rows
```

QUESTION4

This is an optional question that carries additional marks (golden questions). In this question, we are dealing with the InvoiceDate variable. The variable is read as a definite when you read data from the file. Now we need to explicitly instruct R to interpret this as a Date variable. "POSIXlt" and "POSIXct" are two powerful object classes in R to deal with date and time.

First, let's convert 'InvoiceDate' into a POSIXlt object.

```
# 4. Optional question on date manipulation
# Convert InvoiceDate to POSIXlt
Temp <- strptime(Online_Retail$InvoiceDate, format='%m/%d/%Y %H:%M',
tz='GMT')
head(Temp)
```

```
## [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"
```

```r
# Extract date, weekday, hour, and month
Online_Retail$New_Invoice_Date <- as.Date(Temp)
Online_Retail$Invoice_Day_Week <- weekdays(Online_Retail$New_Invoice_Date)
Online_Retail$New_Invoice_Hour <- as.numeric(format(Temp, "%H"))
Online_Retail$New_Invoice_Month <- as.numeric(format(Temp, "%m"))
```

```
## The time difference is 8 days
```

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

```r
# a) Transactions by weekday (counts)
wd_trans <- table(Online_Retail$Invoice_Day_Week)
wd_trans_prop <- prop.table(wd_trans) * 100
print(wd_trans_prop)
```

```
##
##    Friday    Monday    Sunday  Thursday   Tuesday Wednesday
##  15.16731  17.55110  11.87930  19.16503  18.78692  17.45035
```

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

```r
# b) Transactions by weekday (amounts)
wd_amounts <- tapply(Online_Retail$Quantity, Online_Retail$Invoice_Day_Week,
sum)
wd_amounts_prop <- prop.table(wd_amounts) * 100
print(wd_amounts_prop)
```

```
##    Friday    Monday    Sunday  Thursday   Tuesday Wednesday
## 15.347197 15.751219  9.035768 22.560307 18.575336 18.730172
```

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

```r
# c) Transactions by month (amounts)
month_amounts <- tapply(Online_Retail$Quantity,
Online_Retail$New_Invoice_Month, sum)
month_amounts_prop <- prop.table(month_amounts) * 100
print(month_amounts_prop)
```

```
##         1         2         3         4         5         6         7
8
##  5.968685  5.370263  6.797554  5.584870  7.348492  6.599561  7.555680
7.847057
##         9        10        11        12
## 10.621507 11.021685 14.301036 10.983608
```

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

```
# d) Date with most transactions from Australia
aus_trans <- Online_Retail[Online_Retail$Country == "Australia",]
top_date <-
aus_trans$New_Invoice_Date[which.max(table(aus_trans$New_Invoice_Date))]
print(top_date)
```

```
## [1] "2010-12-17"
```

e) The company needs to shut down the website for two consecutive hours for maintenance. What would be the hour of the day to start this so that the distribution is at a minimum for the customers? The responsible IT team is available from 7:00 to 20:00 every day.

```
# e) First, we'll filter to only include transaction hours between 7-20 when
the IT team is available:

Online_Retail <- Online_Retail[Online_Retail$New_Invoice_Hour >= 7 &
                               Online_Retail$New_Invoice_Hour < 20,]

#Next we can get the transaction counts by hour:
hour_counts <- table(Online_Retail$New_Invoice_Hour)

#And find the 2 hour period with the minimum transactions:
# Find index of lowest 2 hour count
min_indx <- which.min(tapply(hour_counts, gl(length(hour_counts), 2), sum))
```

```
## Warning in split.default(X, group): data length is not a multiple of split
## variable
```

```
# Start hour is the first hour of the 2 hour window
start_hour <- min_indx + 7

print(start_hour)

#The two hour period starting at r start_hour would minimize the impact to
```

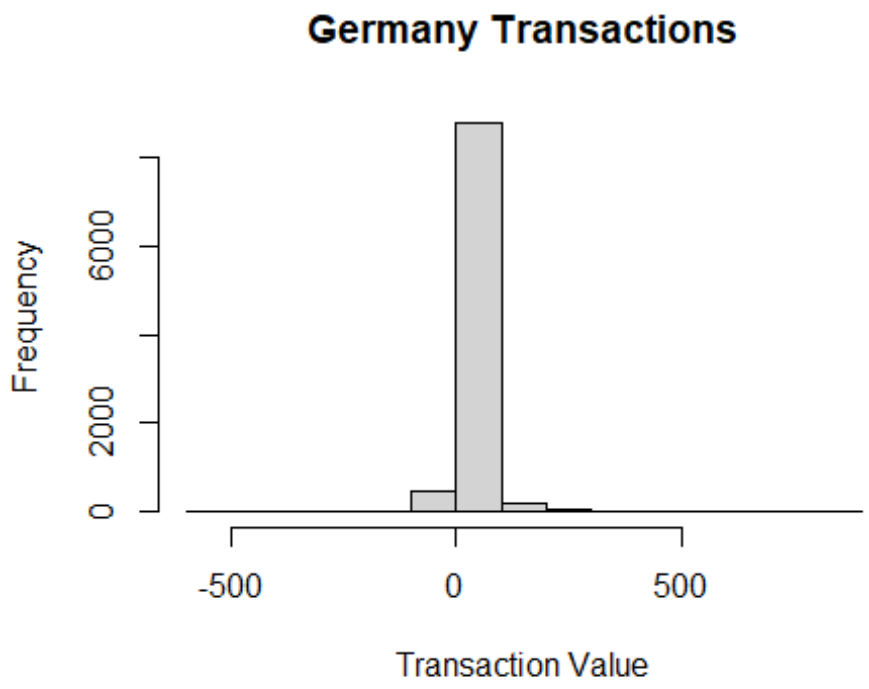
```
customers, since it has the lowest transaction volume based on the data. This
would be the best time for the planned maintenance.
```

```
##7
##14
```

QUESTION5

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

```
# 5. Histogram of Germany transaction values
hist(Online_Retail$TransactionValue[Online_Retail$Country=="Germany"],
main="Germany Transactions", xlab="Transaction Value")
```



QUESTION6

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

```
# 6. Top customers
top_trans <- which.max(table(Online_Retail$CustomerID))
top_spend <- which.max(tapply(Online_Retail$TransactionValue,
Online_Retail$CustomerID, sum))
print(c("Most Transactions:", top_trans, "Highest Spend:", top_spend))
```

```
##                                     17841
## "Most Transactions:"               "4043"     "Highest Spend:"
##                 14646
##                "1704"
```

QUESTION7

Calculate the percentage of missing values for each variable in the dataset.
Hint colMeans():

```
# 7. Missing values percentage
colMeans(is.na(Online_Retail))
```

```
##         InvoiceNo         StockCode       Description          Quantity
##        0.00000000        0.00000000         0.00268763        0.00000000
##       InvoiceDate         UnitPrice         CustomerID           Country
##        0.00000000        0.00000000         0.24968715        0.00000000
##  TransactionValue  New_Invoice_Date  Invoice_Day_Week  New_Invoice_Hour
##        0.00000000        0.00000000         0.00000000        0.00000000
## New_Invoice_Month
##        0.00000000
```

QUESTION8

What is the number of transactions with missing CustomerID records by country?

```
# 8. Missing customer ID by country
miss_cust <- tapply(is.na(Online_Retail$CustomerID), Online_Retail$Country,
sum)
print(miss_cust)
```

```
##            Australia              Austria              Bahrain
##                    0                    0                    2
##              Belgium               Brazil               Canada
##                    0                    0                    0
##      Channel Islands               Cyprus       Czech Republic
##                    0                    0                    0
##              Denmark                 EIRE   European Community
##                    0                  711                    0
##              Finland               France              Germany
##                    0                   66                    0
##               Greece            Hong Kong              Iceland
##                    0                  288                    0
##               Israel                Italy                Japan
##                   47                    0                    0
##              Lebanon            Lithuania               Malta
##                    0                    0                    0
##          Netherlands               Norway               Poland
##                    0                    0                    0
##             Portugal                  RSA         Saudi Arabia
##                   39                    0                    0
##            Singapore                Spain               Sweden
##                    0                    0                    0
##          Switzerland United Arab Emirates       United Kingdom
##                  125                    0               133600
##          Unspecified                  USA
##                  202                    0
```

QUESTION9

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) Hint: 1. A close approximation is also acceptable and you may find diff() function useful.

```
# 9. Average time between transactions
# Using diff() approximation
mean(diff(Online_Retail$New_Invoice_Date[Online_Retail$CustomerID==15211]),
na.rm=TRUE)

## Time difference of 0 days
```

QUESTION10

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. Page 4 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.

```
# 10. Return rate for France
fr_returns <- sum(Online_Retail$Quantity < 0 &
Online_Retail$Country=="France")
fr_trans <- sum(Online_Retail$Country=="France")
print(fr_returns/fr_trans)

## [1] 0.1741264
```

QUESTION11

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

```
# 11. Top revenue product
top_revenue <- which.max(tapply(Online_Retail$TransactionValue,
Online_Retail$StockCode, sum))
print(top_revenue)

##  DOT
## 4060
```

QUESTION12

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

```
# 12. Number of unique customers
length(unique(Online_Retail$CustomerID))
```

```
## [1] 4373
```