HW2

William Kelly

4/22/2021

Install and or load all the required packages here

```
 #install.packages("dplyr") # remove "#" if the package is not already installed.
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
```

Q1 – Read in the ities.csv datafile as a dataframe object, df.

```
df <- read.csv('ities.csv')     # Read "ities.csv" file and assign to data frame variable "df"
```

No descriptive answer is needed.

Q2 – Display the number of rows and columns in the dataset using an appropriate R function. Below the output, identify which numbers from the output correspond to the number of rows and columns.

```
dim(df)   # Reading dimension of dataset
```

```
## [1] 438151     13
```

This dataframe contains 438151 as number of rows and 13 as number of columns

Q3 – Display the structure of the dataframe, df. Below the output, briefly summarize one or two main points about the dataframe structure.

```
str(df)      # Function to display structure of data frame
```

```
## 'data.frame':    438151 obs. of  13 variables:
##  $ Date            : chr  "7/18/2016" "7/18/2016" "7/18/2016" "7/18/2016" ...
##  $ OperationType   : chr  "SALE" "SALE" "SALE" "SALE" ...
##  $ CashierName     : chr  "Wallace Kuiper" "Wallace Kuiper" "Wallace Kuiper" "Wallace Kuiper" ...
##  $ LineItem        : chr  "Salmon and Wheat Bran Salad" "Fountain Drink" "Beef and Squash Kabob" "Sa
##  $ Department      : chr  "Entrees" "Beverage" "Kabobs" "Salad" ...
```

```
##  $ Category         : chr  "Salmon and Wheat Bran Salad" "Fountain" "Beef" "general" ...
##  $ RegisterName     : chr  "RT149" "RT149" "RT149" "RT149" ...
##  $ StoreNumber      : chr  "AZ23501305" "AZ23501289" "AZ23501367" "AZ23501633" ...
##  $ TransactionNumber: chr  "002XIIC146121" "002XIIC146121" "00PG9FL135736" "00Z3B4R37335" ...
##  $ CustomerCode     : chr  "CWM11331L8O" "CWM11331L8O" "CWM11331L8O" "CWM11331L8O" ...
##  $ Price            : num  66.22 2.88 12.02 18.43 18.43 ...
##  $ Quantity         : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ TotalDue         : num  66.22 2.88 24.04 18.43 18.43 ...
```

This data frame has 13 columns (variables) and 438151 rows (observations) The data frame has 13 unique categories of data & contains chr, num, int types

Q4 – What is the unit of observation in the dataset?

Display at least one calculation in the code chunk below. Below the calculation(s), briefly explain how the output of your calculation(s) supports your conclusion.

```
unique(df$Category)           # Function displays unique observations in the Category column
```

```
##  [1] "Salmon and Wheat Bran Salad"     "Fountain"
##  [3] "Beef"                            "general"
##  [5] "Chicken"                         "Beef and Apple Burgers"
##  [7] "Beef and Broccoli"               "Rice"
##  [9] "Naan"                            "Glass Bottle"
## [11] "Yogurt"                          "Chutney"
## [13] "Roll"                            "Aubergine and Chickpea Vindaloo"
## [15] "Lamb Chops"                      "Lamb"
## [17] "Beef Stew"                       "Chips"
## [19] "Pork"                            "Non Food"
## [21] "Curry"
```

```
length(unique(df$Category))   # Function length presents number of unique observations in Category
```

```
## [1] 21
```

The number of observations for 'Category' is 21 based on the length function

Q5 – If we were to consider each transaction as the unit of observation, comment on whether the provided data is in a tidy format. (No code is needed)

The Data is Tidy because each column contains a variable corresponding to a unique piece of information i.e. Date, Price, Quantity, Total Due And each column represents a feature of the observations

Q6 – Display the summaries of the Price, Quantity and TotalDue columns. Below the output, provide a brief interpretation of the output.

```
summary(df[,c('Price','Quantity','TotalDue')])    # Show summaries of Price Quantity and TotalDue
```

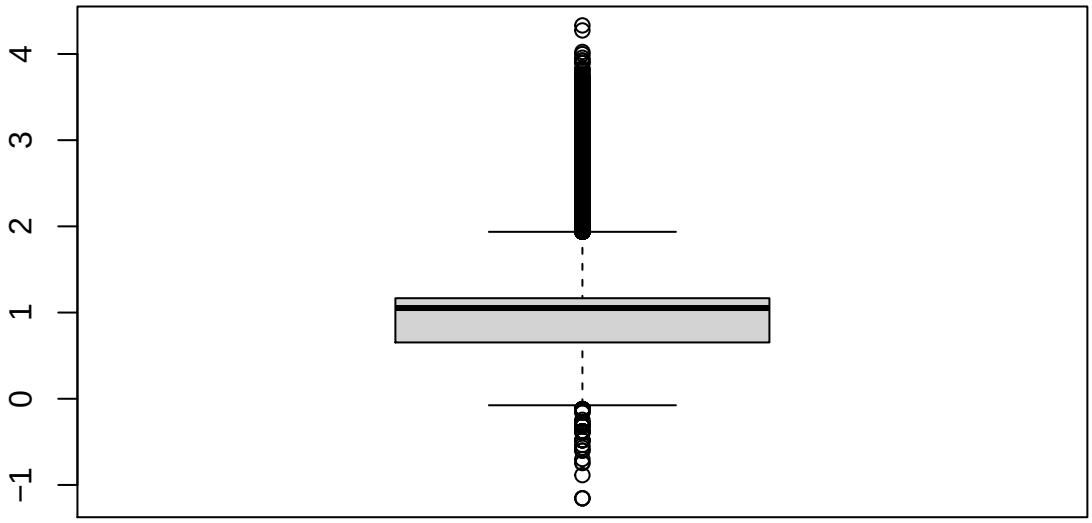
```
##      Price             Quantity          TotalDue
##  Min.   :-5740.51   Min.   :  1.000   Min.   :-5740.51
##  1st Qu.:    4.50   1st Qu.:  1.000   1st Qu.:    4.50
##  Median :   11.29   Median :  1.000   Median :   11.80
##  Mean   :   14.36   Mean   :  1.177   Mean   :   15.26
##  3rd Qu.:   14.68   3rd Qu.:  1.000   3rd Qu.:   15.04
##  Max.   :21449.97   Max.   :815.000   Max.   :21449.97
##  NA's   :12                           NA's   :12
```

Minimum Price is $-5740.51 and Max price is $21449.97 The average quantity is 1.177 and max quantity is 815 Median total due is $11.80 and max total due is $21449.97

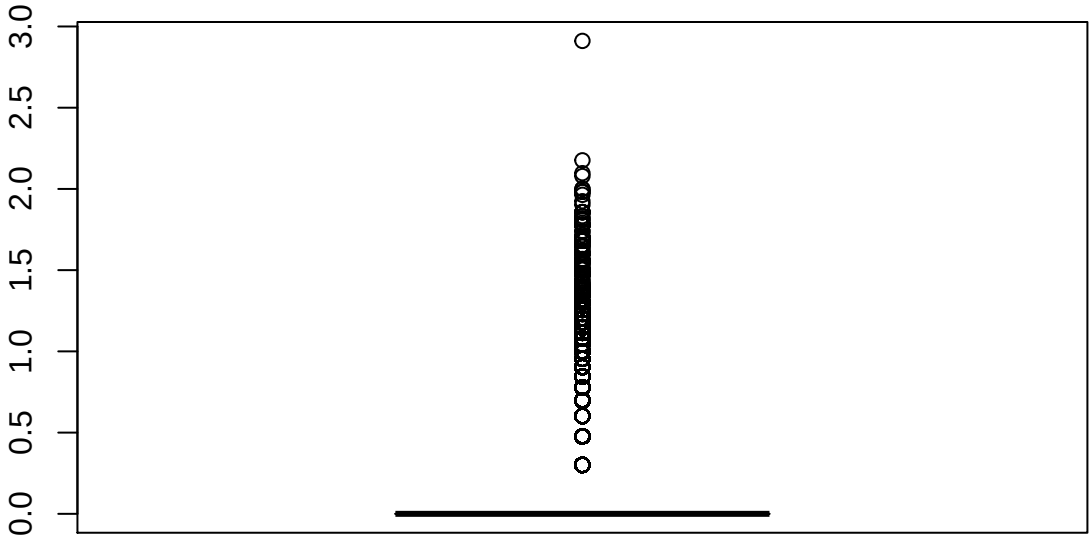
Q7 – Display the boxplots of the log values for the Price, Quantity and Total Due columns. Below the output, provide a brief interpretation of the output.

```
boxplot(log10(df$Price))    # Displays box plot of Price, Quantity, TotalDue
```

```
## Warning in boxplot(log10(df$Price)): NaNs produced
```

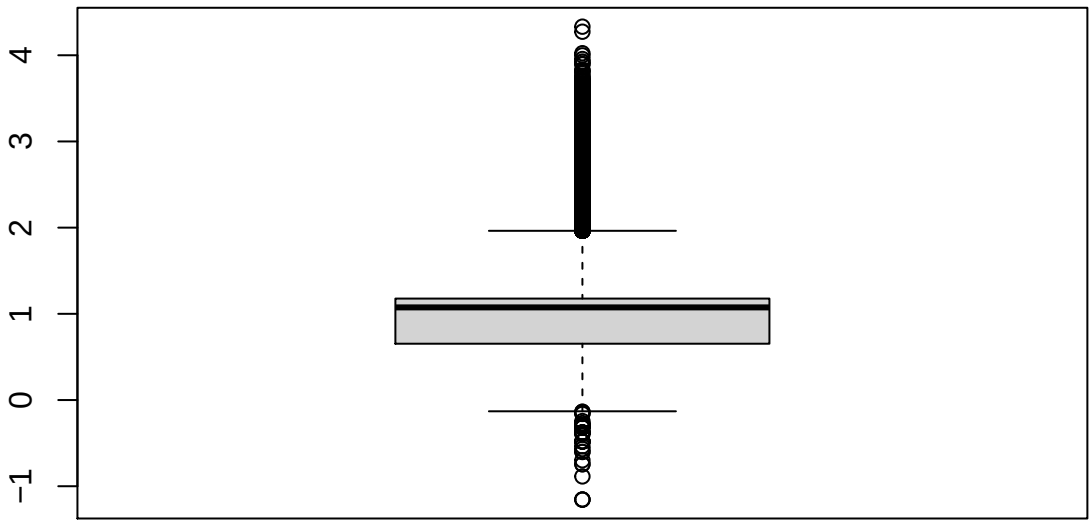



```
boxplot(log10(df$Quantity))
```




```
boxplot(log10(df$TotalDue))
```

```
## Warning in boxplot(log10(df$TotalDue)): NaNs produced
```



Price and Total Due box plots have similar distribution of data Median value falls at 1 and outliers range from -1 and 4 Min w/o outliers is 0 and Max w/o outliers is 2 Slight indication of right skew towards larger values

The quantity box plot has a smaller distribution and median IQR at 0 Outliers ranging from 0 to 3 and skew towards larger values