Plotting and Analyizng Grocery Data

Stephen Bernardo

8/22/2023

Reading the data file

```
df <- read.csv('ities.csv')
```

Displaying the row and column count

```
nrow(df)
```

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

```
ncol(df)
```

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

This dataframe has 438151 rows and 13 columns.

Displaying the structure of the dataframe, df

```
str(df)
```

```
## '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 ...
```

The dataframe structure shows that there are 8899 observations (rows) and 13 variables (columns). The dataframe structure shows that one variable, Quantity, is an integer (int), 2 variables have numerical values (num), and the remaining 9 variables are characters (chr).

Checking the length of unique dates and cashier name.

True or False: Every transaction is summarized in one row of the dataframe. Display at least one calculation in the code chunk below. Below the calculation(s), clearly indicate whether

the statement is true or false and explain how the output of your calculation(s) supports your conclusion.

```
length(unique(df$Date))
```

```
## [1] 1021
```

```
length(unique(df$CashierName))
```

```
## [1] 56
```

True. The transactions can be summarized into one row of the data frame by grouping similar transactions after. We can see from the sample calculation above that using:

- Date: There are 1,021 unique dates for the 438,151 transactions (rows) in the dataframe. This means that the transactions can be grouped according to 1,021 unique dates.
- CashierName: There are 56 cashiers for the 438,151 transactions (rows) in the dataframe. This means that the transactions can be grouped according to the 56 unique cashiers.

Displaying the summaries of the Price, Quantity, and TotalDue columns

```
summary(df[,c('Price','Quantity','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
```

Price and TotalDue:

- There were 12 null values (indicated by NA)
- The min value is -5740.51 and the max is 21449.57, which indicates that there is a high variability between the observation data, as compared to the 1st quartile, median, mean, and 3rd quartile.

Quantity:

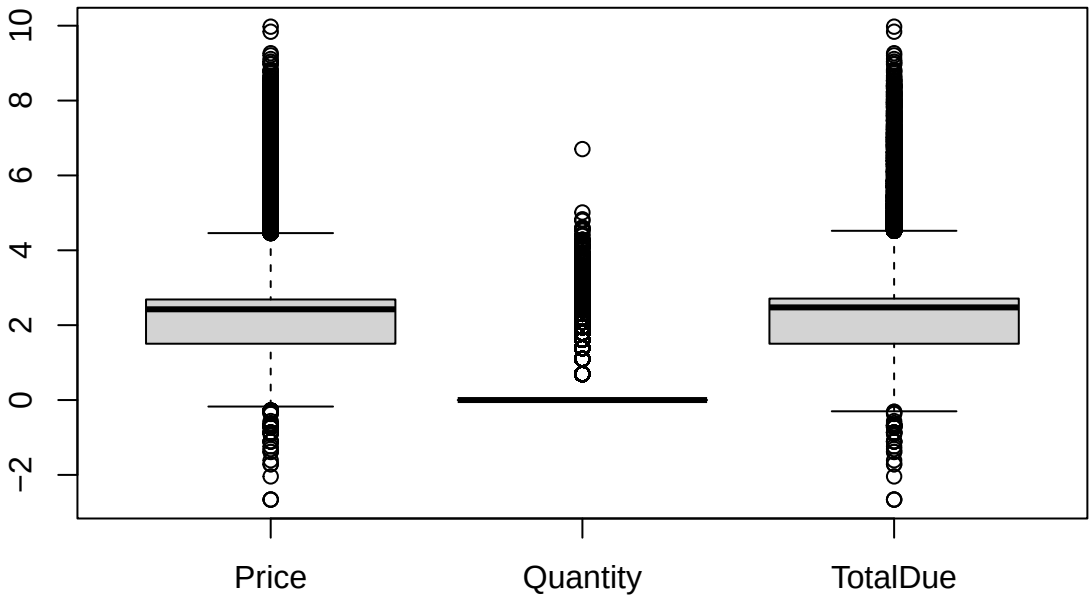
- The min, 1st. quartile, median, and 3rd quartile is 1.0000 while the mean is 1.177, which indicates that majority of the observations are 1.000.
- The max value is 815.000, compared to a mean value of 1.177, a min value of 1.000 and median value of 1.000. This implies that majority of the observations is 1.000.

Displaying the boxplots of the log values for the Price, Quantity and TotalDue columns

```
boxplot(log(df[,c('Price','Quantity','TotalDue')]))
```

```
## Warning in FUN(X[[i]], ...): NaNs produced
```

```
## Warning in FUN(X[[i]], ...): NaNs produced
```



Three insights:

1. The boxplot shows that there are a number of outliers, as indicated by the dots (observation data) that are beyond the 1st and 4th quartiles, as well as the interquartile range. This is consistent with the output from task 5 because for all variables (Price, Quantity, and TotalDue), the min and max values are significantly different from the 1st quartile, mean, and median, and the 3rd quartile. In addition, there were also null data (indicated by NaNs) that were not included in this plot.

2. All outliers in Quantity are above the median (indicated by the whisker), which implies that all values are above 0. This is consistent with the task 5 results because we saw that the min. is 1, which indicates that there are no negative values. In contrast, there are a number of outliers in Price and TotalDue that are below 0, which indicates that there were negative prices in the dataset. This is consistent with task 5 because we see that the min. value for both Price and Total due is -5740.51.

3. The interquartile range (indicated by the gray area that surrounds the whisker, the median) is wider in the Price and TotalDue columns than in the quantity column. This is consistent with task 5 because we see that the 1st quartile, median, and 3rd quartile for Quantity are all 1.000, while the values for the same statistic for Price and TotalDue are different.