Group Homework 1 Solutions

STAT011-F23

Introduction and Purpose

The data set you will analyse in this homework records information about different items on the menu at a local Burger King restaurant. One purpose of this assignment is to practice using software to explore and summarize qualities of these data consisting of 122 observations and 17 columns (the first column identifies the item name, the next 16 columns are different variables). Another purpose of this assignment is to begin to practice working with data through software. Most of your time on this assignment will be spent getting your software set up and on learning how to do basic exploratory statistics with either Excel or RStudio.

Required Tech

Excel

The skills necessary to complete this assignment in Excel are covered in the following seven videos:

- Excel 2016 with Data Analysis Toolpak Introduction to Excel 2016 with Data Analysis Toolpak (1:52)
- Excel 2016 with Data Analysis Toolpak Introduction to Excel 2016 with Data Analysis Toolpak: Common Procedures (3:08)
- Excel 2016 with Data Analysis Toolpak Descriptive Statistics and Confidence Intervals for a Mean (2:57)
- Excel 2016 with Data Analysis Toolpak Histogram (3:08)

RStudio

The skills necessary to complete this assignment in RStudio are covered in the following seven videos:

- R Studio Video Introduction to R and RStudio (1:52)
- R Studio Video Getting Started (3:51)
- R Studio Video Working with Data Objects 1 (3:29)
- R Studio Video Working with Data Objects 2 (5:15)
- R Studio Video Importing Data (4:36)
- R Studio Video Descriptive Statistics (3:33)
- R Studio Video Plotting - Histograms, Bar Charts, Boxplots, Scatterplots (4:08)

Instructions

If you are analyzing this data in Excel you first need to download the data set for HW 2 from our Stat 11 Github Data page. Do this by right clicking on the link "View Raw" and save the link with the name `Burger_King_items.csv`.

If you are analyzing this data in RStudio, you will import the data with the following command

```
Burger_King_items <- read.csv(
    "https://raw.githubusercontent.com/ProfSuzy/Stat11/main/Data/Burger_King_items.csv")
```

The data object is called `Burger_King_items`.

Once you have access to the data set, complete all parts of the five problems in this assignment. You are encouraged to work with your classmates on this assignment but you must hand in your own, unique write up of the solutions. In a Word document, clearly label each problem's solution. Most solutions will include graphics which can be copied from Excel or RStudio and pasted into your solution document. All solutions require a written component. When you are ready to submit your assignment, save the Word document as a PDF and upload it to the Moodle link for Group Homework #1.

Problem 1

This is a three part question. For this data set, what constitutes an observational unit? What are the different variables being collected? Finally, which of the variables are quantitative and which are categorical and are there any variables that could be either/both?

Solution 1

- Observational units - items on a BK menu
- Variables - 16 total

numeric - serving size, calories, fat calories, protein, fat, saturated fat, trans fat, cholesterol, sodium, carbs, fiber, sugar

categorical - meat breakfast, not breakfast

Note, "carbs times meat" is the last column which is the multiplication of meat indicator times carbs amount (not an original variable technically)

- Technically, any of the numeric varbs could be treated as categorical varbs. (More realistically, we'd probably consider treating trans fat or fiber or sugar as categorical since there seem to be a few values of these variables that are repeated quite often.)

For graders: it is ok if they say there are 17 variables

Potentially useful R Code:

```
head(Burger_King_items)
```

```
##                  Item Serving.size Calories Fat.Cal Protein.g. Fat.g.
## 1           Hamburger          109      260      90         13     10
## 2        Cheeseburger          121      300     130         16     14
## 3    Double_Hamburger          146      360     160         22     18
## 4 Double_Cheeseburger          171      450     230         26     26
## 5         Buck_Double          158      410     200         24     22
## 6  Rodeo_Cheeseburger          128      350     160         16     17
##   Sat.Fat.g. Trans.fat.g. Chol.mg. Sodium.mg. Carbs.g. Fiber.g. Sugar.g. Meat
## 1          4          0.0       35        490       28        1        6    1
## 2          6          0.0       45        710       28        1        6    1
## 3          8          0.0       70        520       28        1        6    1
## 4         12          1.0       95        960       29        1        6    1
## 5         10          0.5       85        740       28        1        6    1
## 6          7          0.0       45        600       37        2        9    1
##   Breakfast
```

```
## 1         0
## 2         0
## 3         0
## 4         0
## 5         0
## 6         0
```

```
summary(Burger_King_items)
```

```
##      Item            Serving.size      Calories         Fat.Cal
##  Length:122         Min.   : 43.0   Min.   :  25.0   Min.   :  0.0
##  Class :character   1st Qu.:113.5   1st Qu.: 310.0   1st Qu.:127.5
##  Mode  :character   Median :158.0   Median : 410.0   Median :190.0
##                     Mean   :167.7   Mean   : 452.1   Mean   :205.7
##                     3rd Qu.:216.5   3rd Qu.: 550.0   3rd Qu.:267.0
##                     Max.   :487.0   Max.   :1310.0   Max.   :650.0
##                     NA's   :11                       NA's   :6
##    Protein.g.        Fat.g.        Sat.Fat.g.      Trans.fat.g.
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 7.00   1st Qu.:14.25   1st Qu.: 4.000   1st Qu.:0.0000
##  Median :15.50   Median :22.00   Median : 7.000   Median :0.0000
##  Mean   :17.93   Mean   :24.78   Mean   : 8.701   Mean   :0.2992
##  3rd Qu.:24.75   3rd Qu.:33.00   3rd Qu.:12.000   3rd Qu.:0.5000
##  Max.   :71.00   Max.   :82.00   Max.   :32.000   Max.   :2.0000
##
##     Chol.mg.        Sodium.mg.        Carbs.g.         Fiber.g.
##  Min.   :  0.00   Min.   :   0.0   Min.   :  2.00   Min.   :0.000
##  1st Qu.: 20.00   1st Qu.: 490.0   1st Qu.: 27.00   1st Qu.:1.000
##  Median : 52.50   Median : 905.0   Median : 34.50   Median :1.000
##  Mean   : 92.75   Mean   : 918.8   Mean   : 39.37   Mean   :1.905
##  3rd Qu.:143.75   3rd Qu.:1245.0   3rd Qu.: 50.00   3rd Qu.:3.000
##  Max.   :455.00   Max.   :2490.0   Max.   :134.00   Max.   :9.000
##                                                     NA's   :6
##     Sugar.g.           Meat          Breakfast
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 6.000   Median :1.0000   Median :0.0000
##  Mean   : 9.975   Mean   :0.5984   Mean   :0.3689
##  3rd Qu.:10.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :58.000   Max.   :1.0000   Max.   :1.0000
##
```

Problem 2

This is a three part question. Create a histogram for the variable `Sugar.g.` and determine if it looks like this sample could have been drawn from a Normal/Gaussian population. Then compute a five number summary of the amount of sugar in these items that includes the mean, median, minimum, maximum, lower 25% quantile, and lower 75% quantile. Which measure of location (mean or median) is a more appropriate and why?

Solution 2

The distribution of the variable `Sugar.g.` is heavily skewed and does not look like it could be from a Normal population. The five number summary for `Sugar.g.` is:

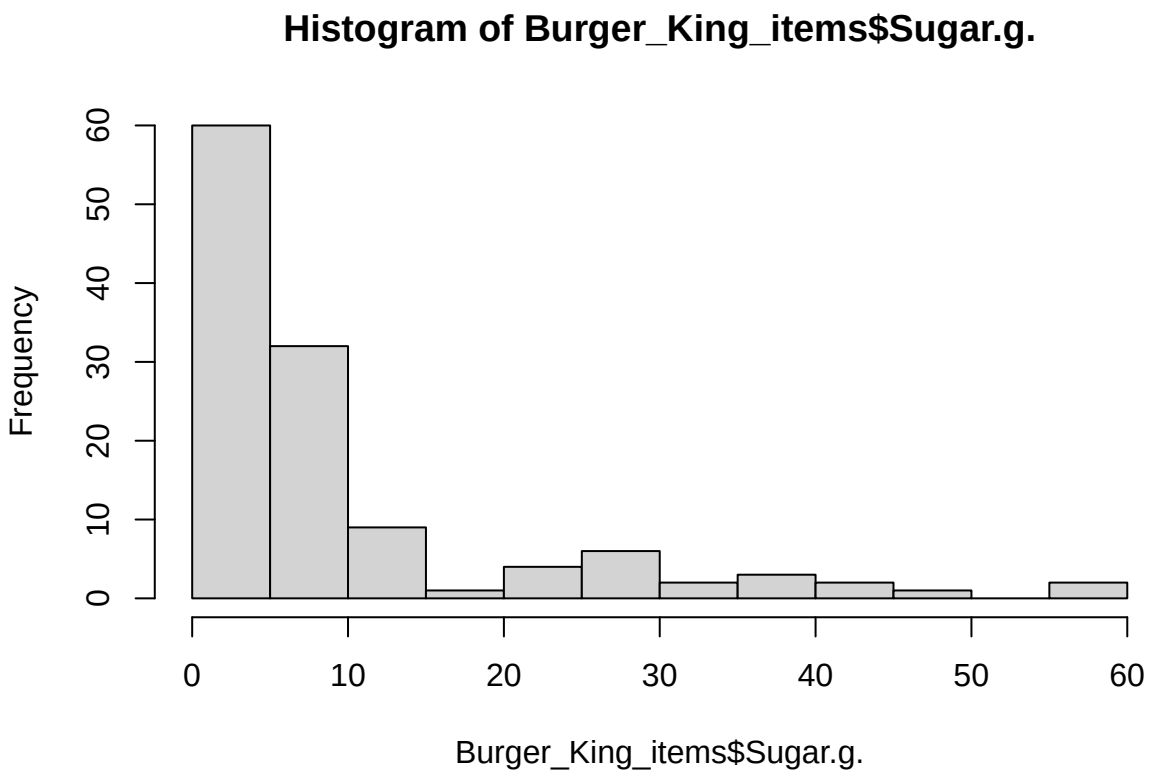
- minimum 0g
- 1st quartile 3g
- median 6g
- 3rd quartile 10g
- maximum 58g

Because of the skew, the median is a more appropriate measure of centrality/location.

```
summary(Burger_King_items$Sugar.g.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   6.000   9.975  10.000  58.000
```

```
hist(Burger_King_items$Sugar.g., breaks=20)
```



Problem 3

This is a three part question. Create a histogram for the variable `Sodium.mg.` and determine if it looks like this sample could have been drawn from a Normal/Gaussian population. Then, calculate the mean, standard deviation, and variance of the amount of sodium in each of the sampled items. How many standard deviations away from the mean sodium amount is the Sausage Egg & Cheese Biscuit?

Solution 3

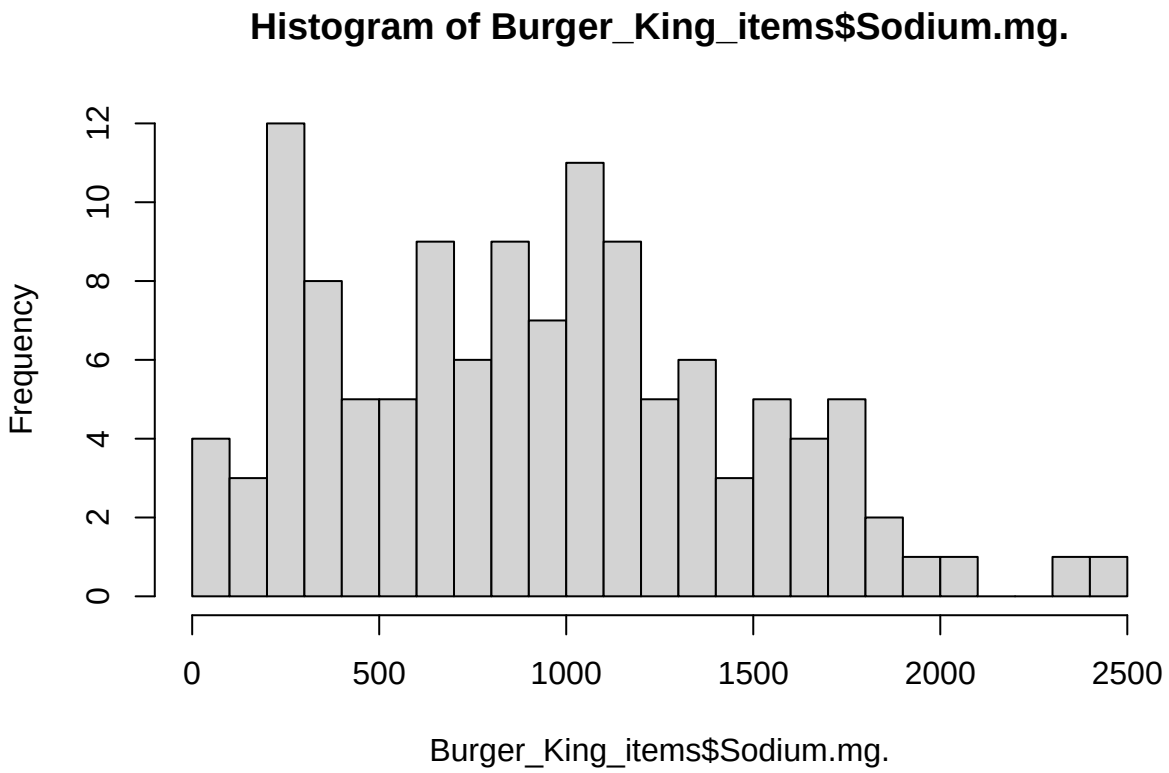
```
mean(Burger_King_items$Sodium.mg.); sd(Burger_King_items$Sodium.mg.); var(Burger_King_items$Sodium.mg.)
```

```
## [1] 918.7705
```

```
## [1] 537.2755
```

```
## [1] 288665
```

```
hist(Burger_King_items$Sodium.mg., breaks = 20)
```



The variable `Sodium.mg.` does not look very symmetric so it probably doesn't come from a Normal population. However, if we specify the number of breaks (or bins) in the histogram, we can get a plot that looks a bit more symmetric. In any case, `Sodium.mg.` is more plausibly Normally distributed than `Sugar.g.` but neither seem to fit the symmetry criteria very well.

The mean sodium content is $918.77mg$ with a standard deviation of $537.28mg$, or a variance of $288,665mg^2$.

The Sausage Egg & Cheese biscuit is observational unit # 84 and it has a sodium content of $1520mg$. This is $\frac{1520-918.77}{537.28} = 1.12$ standard deviations above the mean sodium content.

Problem 4

This is a three part question. What is the proportion (or percent) of menu items have more than the daily recommended intake of $2300mg$ of sodium? What is the proportion (or percent) of menu items that have between $500mg$ and $1000mg$ of sodium? Which item(s) on the menu is (are) at the lower 20^{th} percentile for the amount of sodium they contain?

Solution 4

Less than 10% of the items on the BK menu have more than the daily recommended intake of sodium. Roughly, only about a fourth of the items have between $500mg$ and $1000mg$ of sodium. Items at the lower 20^{th} percentile/quantile of the distribution have about $400 - 500mg$ of sodium. Some items in this range include: CHICKENTENDERS (6pc), PancakePlatterw/Sausage&BreakfastSyrup, HashBrownsSmall, FrenchToastSticks(5piece), MuffinBlueberry.

For graders: Note that eyeballed answers, based on a reasonable histogram in # 2, are OK. Students may or may not use code to answer these questions. Please be generous with grading.

Potentially useful R Code:

```
sum(Burger_King_items[,10]>=2300)/length(Burger_King_items[,10])
```

```
## [1] 0.01639344
```

```
sum((Burger_King_items[,10]>=500)&(Burger_King_items[,10]<=1000))/length(Burger_King_items[,10])
```

```
## [1] 0.295082
```

```
quantile(Burger_King_items[,10], probs = 0.2)
```

```
## 20%
## 380
```

Problem 5

Sometimes we may want to **standardize** a set of quantitative data. This process transforms the original data, x_i, into a *unitless* variable, z_i, and forces the mean of the standardized data to be zero and the standard deviation of the standardized data to be one using this formula:

$$z_i = \frac{x_i - \bar{x}}{sd(x)}, \quad \text{for } i = 1, \ldots, n.$$

For example, directly comparing sodium content to sugar content could be challenging as they are measured in different units (mg and g, respectively). If we standardize these variables then we get rid of the units for each of these variables and can compare the sodium and sugar content on the same (unitless) scale, centered at zero with standard deviation one. Sometimes these standardized values are called "z-scores".

This is a two part question. Suppose we decide to standardize the variable for sodium content. The Double Whopper contains 980mg of sodium. What is the standardized (unitless) amount of sodium in a Double Whopper? If the standardized sodium content of an item is -1.23, what was the sodium content in mg?

Solution 5

First, recall that the mean amount of sodium in all of these menu items is $918.77mg$ with a standard deviation of $537.28mg$.

The z-score for the Double Whopper is: $\frac{980-918.77mg}{537.28mg} = 0.114$

If the z-score for an item on the menu is -1.23 then the original sodium content was: $-1.23 \times 537.28mg + 918.77mg = 257.92mg$

For graders: For full credit, the answer to the first part must not have any units but the answer to the second part must be in mg.