STAT 216 Homework 0

Drusilla Talawa

07/09/2020

The ## symbols make the text larger and in bold when the text appears on a separate line.

** = italics *word and phrase*

backticks produce code-like text `backticks`

My Main Major Interest

My current major is Data Science with a concentration in Political Science. I have taken a few statistics, political science, epidemiology, math and computer science classes for my major which have all been great. I especially enjoy the interdisciplinary nature of my major, and all the challenges that come with it. Most people actually refer to data science as the computer programming version of statistics, so the two are very much related.

```
2 + 2
```

```
## [1] 4
```

```
10/2
```

```
## [1] 5
```

```
17*5
```

```
## [1] 85
```

```
3^3
```

```
## [1] 27
```

Importing a Dataset

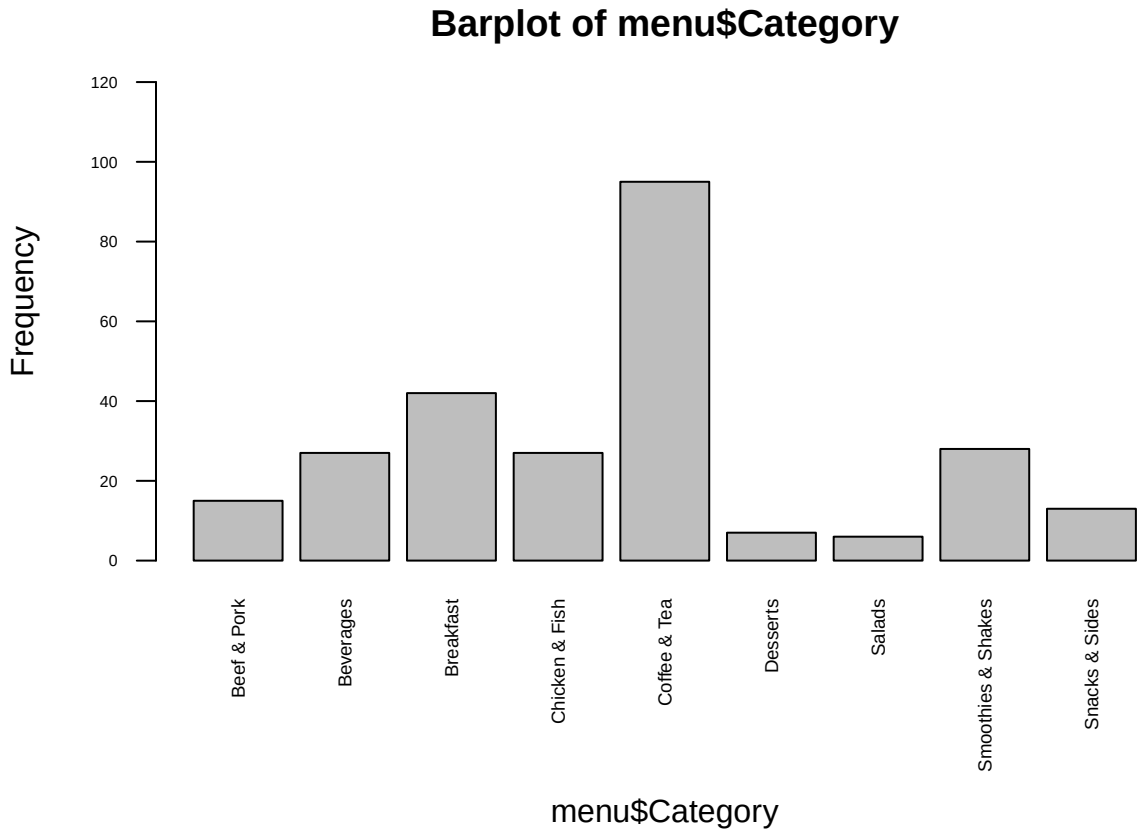
This is the code for importing a CSV file:

```
setwd("C:/Users/Drusilla/Documents/STAT216")
menu <- read.csv("menu.csv", header=TRUE)
```

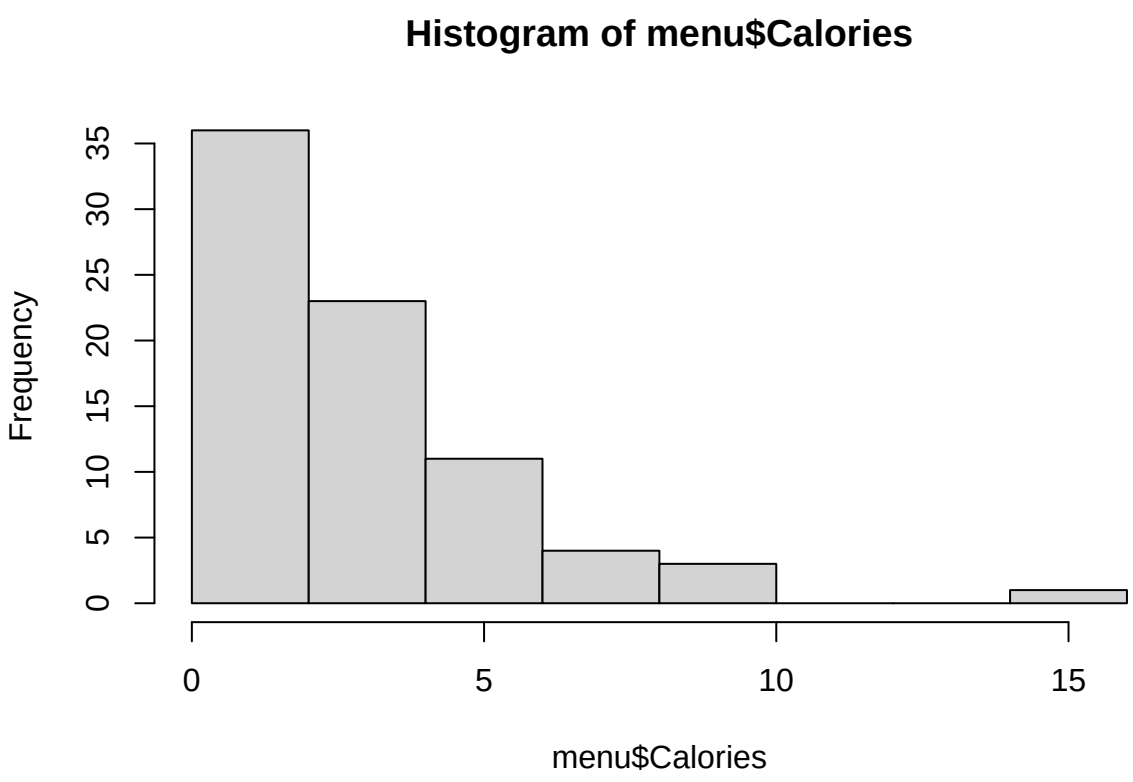
Exploring the Data

Barchart of menu categories

```
categories <- table(menu$Category)
par(mar= c(8,5,1,1))
barplot(categories,
        ylab = "Frequency",
        ylim = c(0,130),
        main="Barplot of menu$Category",
        las=2,
        cex.names = 0.60, #decrease font for axis names
        cex.axis=0.50) #decrease font for axis labels
mtext(side=1, text= "menu$Category", line = 6)
```



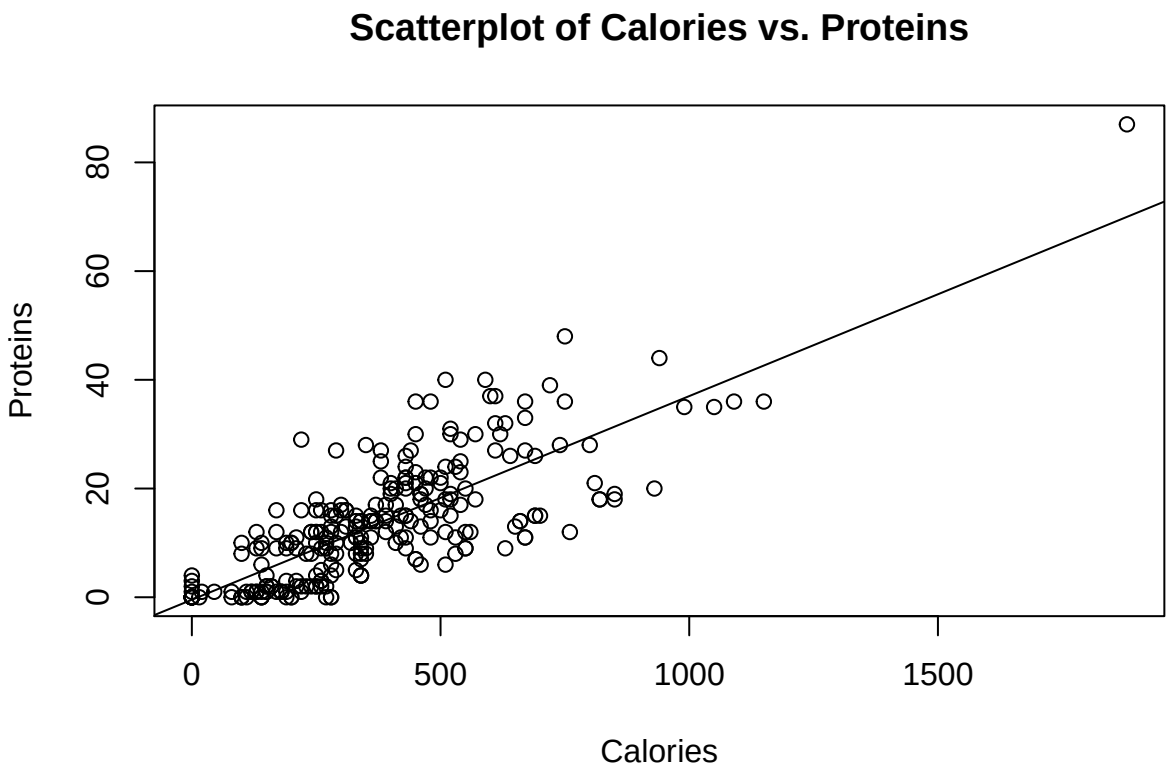

```
calories <- table(menu$Calories)
hist(calories,
     ylab = "Frequency",
     xlab = "menu$Calories",
     breaks = 6, #of bins
     main="Histogram of menu$Calories")
```




```
#pearson correlation for 2 variables
cor(menu$Calories, menu$Protein,
    method = c("pearson"))
```

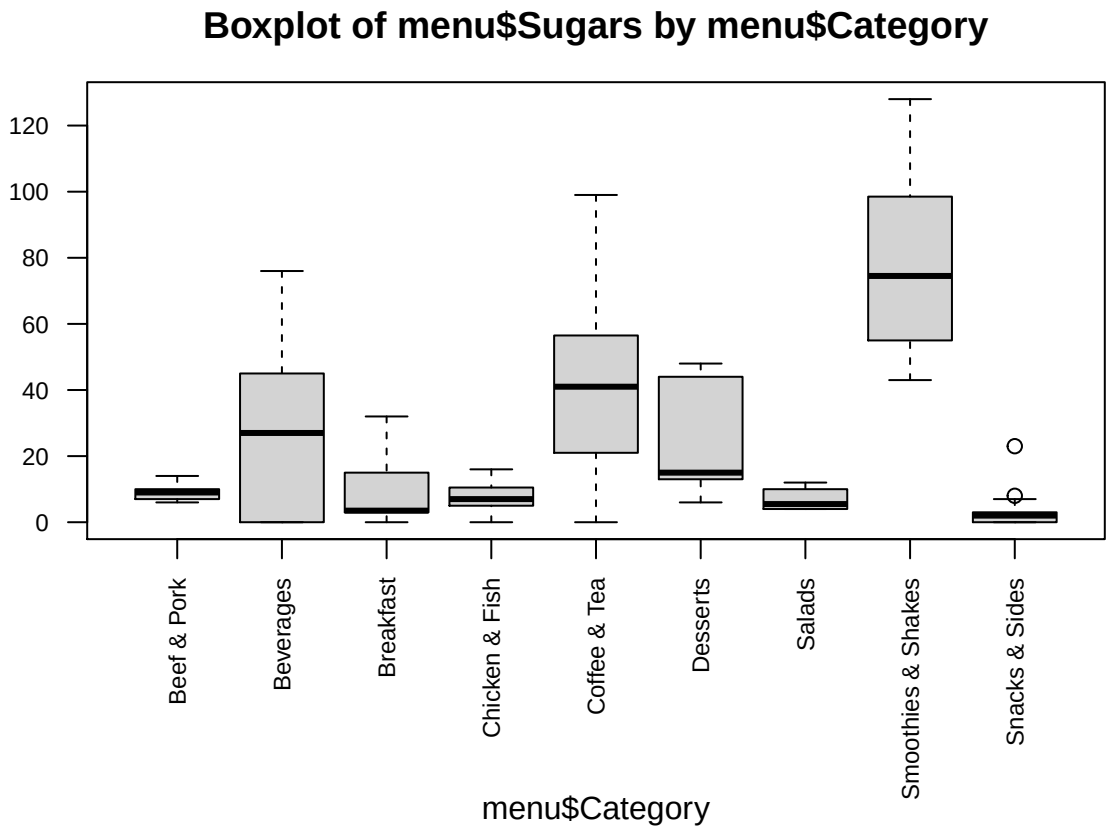
```
## [1] 0.7878475
```

```
x <- menu$Calories
y <- menu$Protein
plot(x, y, main = "Scatterplot of Calories vs. Proteins",
     xlab = "Calories", ylab = "Proteins",frame = TRUE)
abline(lm(y~x, data = menu), col="black")
```




```r
sugar <- menu$Sugars
category <- menu$Category

#showing the x-axis labels
par(mar=c(7.6, 3, 3, 3),
    cex.axis = 0.75)
boxplot(sugar~category, main = "Boxplot of menu$Sugars by menu$Category",
        xlab = "", #leave empty for mtext() to push down xlab by 6.5
        ylab = "menu$Sugars",
        las=2)
mtext(side=1, text= "menu$Category", line = 6.5)
```




```
#calculating mean for each category by sugar content

tapply(menu$Sugars, menu$Category, mean)
```

```
##        Beef & Pork          Beverages          Breakfast     Chicken & Fish
##           8.800000          27.851852           8.261905           7.333333
##       Coffee & Tea           Desserts             Salads Smoothies & Shakes
##          39.610526          26.142857           6.833333          77.892857
##     Snacks & Sides
##           4.076923
```

```
#alternative but more lines of code
Breakfast <- subset(menu, Category == "Breakfast") #subset "Breakfast"
mean(Breakfast$Sugars) #mean for "Breakfast" by Sugars
```

```
## [1] 8.261905
```

```
#then repeat above for each category
```

Discussion

a) Using the Histogram of menu$Calories, it is evident that a greater area of the histogram lies within the first three bins, and continues to decrease. Hence, we can infer that the distribution for calories is j-shaped and is unimodal with a right-skew.

b) Since we are trying to plot for sugar content by category, this is a case C(menu categories),Q(sugar content) and the independent variable 'category' is what we are testing against the dependent variable of 'sugar content'.

c) The correlation coefficient obtained for calories vs. proteins is 0.7878 which shows that there is a strong positive linear relationship between the two variables.

d) Most of the coordinates are also scattered close to the line of best fit, with one outlier that is located at a far extreme of the graph but generally follows the trend between X and Y. Perhaps we need more data to make a definite conclusion about whether we are to treat it as an outlier or not.

e) The mean values for each subset of the Category variable generally make sense. For instance, on average the Smoothies & Shakes have the highest sugar content while Salads have the 2nd lowest sugar content. While the snacks and sides have the lowest sugar content but are also an unhealthy food option, they might have higher sodium levels (mainly fries) instead of sugar (like in the fruit n' yogurt parfait) which is the only value pulling the mean to the right (more positive).