CMTH 642 Data Analytics: Advanced Methods

Assignment 2 (10%)

Tasdeed Aziz

Section: DHB & ID:500945638

DUE:4th June, 2020

```
USDA_Clean <- read.csv('/Users/ayon/Desktop/CMTH642/Assignments/Assignment2  /USDA_Clean.csv')
head(USDA_Clean,n=10)
```

1. Read the csv file (USDA_Clean.csv) in the folder and assign it to a data frame. (3 points)

```
##     X   ID              Description Calories Protein TotalFat Carbohydrate
## 1   1 1001         BUTTER,WITH SALT      717    0.85    81.11         0.06
## 2   2 1002 BUTTER,WHIPPED,WITH SALT      717    0.85    81.11         0.06
## 3   3 1003     BUTTER OIL,ANHYDROUS      876    0.28    99.48         0.00
## 4   4 1004             CHEESE,BLUE      353   21.40    28.74         2.34
## 5   5 1005            CHEESE,BRICK      371   23.24    29.68         2.79
## 6   6 1006             CHEESE,BRIE      334   20.75    27.68         0.45
## 7   7 1007         CHEESE,CAMEMBERT      300   19.80    24.26         0.46
## 8   8 1008           CHEESE,CARAWAY      376   25.18    29.20         3.06
## 9   9 1009           CHEESE,CHEDDAR      403   24.90    33.14         1.28
## 10 10 1010          CHEESE,CHESHIRE      387   23.37    30.60         4.78
##    Sodium Cholesterol    Sugar Calcium Iron Potassium VitaminC VitaminE
## 1     714         215 0.060000      24 0.02        24        0 2.320000
## 2     827         219 0.060000      24 0.16        26        0 2.320000
## 3       2         256 0.000000       4 0.00         5        0 2.800000
## 4    1395          75 0.500000     528 0.31       256        0 0.250000
## 5     560          94 0.510000     674 0.43       136        0 0.260000
## 6     629         100 0.450000     184 0.50       152        0 0.240000
## 7     842          72 0.460000     388 0.33       187        0 0.210000
## 8     690          93 8.229355     673 0.64        93        0 1.474358
## 9     621         105 0.520000     721 0.68        98        0 0.290000
## 10    700         103 8.229355     643 0.21        95        0 1.474358
##     VitaminD HighSodium HighCals HighSugar HighProtein HighFat
## 1  1.5000000          1        1         0           0       1
## 2  1.5000000          1        1         0           0       1
## 3  1.8000000          0        1         0           0       1
## 4  0.5000000          1        1         0           1       1
## 5  0.5000000          1        1         0           1       1
## 6  0.5000000          1        1         0           1       1
```

```
## 7  0.4000000          1        1         0           1       1
## 8  0.5770542          1        1         1           1       1
## 9  0.6000000          1        1         0           1       1
## 10 0.5770542          1        1         1           1       1
```

```
str(USDA_Clean)
```

2. Check the datatypes of the attributes. (3 points)

```
## 'data.frame':    6310 obs. of  21 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...
##  $ Description : Factor w/ 6306 levels "ABALONE,MIXED SPECIES,RAW",..: 1240 1239 1235 1972 1973 1974
##  $ Calories    : int  717 717 876 353 371 334 300 376 403 387 ...
##  $ Protein     : num  0.85 0.85 0.28 21.4 23.24 ...
##  $ TotalFat    : num  81.1 81.1 99.5 28.7 29.7 ...
##  $ Carbohydrate: num  0.06 0.06 0 2.34 2.79 0.45 0.46 3.06 1.28 4.78 ...
##  $ Sodium      : int  714 827 2 1395 560 629 842 690 621 700 ...
##  $ Cholesterol : int  215 219 256 75 94 100 72 93 105 103 ...
##  $ Sugar       : num  0.06 0.06 0 0.5 0.51 ...
##  $ Calcium     : int  24 24 4 528 674 184 388 673 721 643 ...
##  $ Iron        : num  0.02 0.16 0 0.31 0.43 0.5 0.33 0.64 0.68 0.21 ...
##  $ Potassium   : int  24 26 5 256 136 152 187 93 98 95 ...
##  $ VitaminC    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VitaminE    : num  2.32 2.32 2.8 0.25 0.26 ...
##  $ VitaminD    : num  1.5 1.5 1.8 0.5 0.5 ...
##  $ HighSodium  : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ HighCals    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HighSugar   : int  0 0 0 0 0 0 0 1 0 1 ...
##  $ HighProtein : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ HighFat     : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
names(USDA_Clean)
```

```
##  [1] "X"            "ID"           "Description"  "Calories"     "Protein"
##  [6] "TotalFat"     "Carbohydrate" "Sodium"       "Cholesterol"  "Sugar"
## [11] "Calcium"      "Iron"         "Potassium"    "VitaminC"     "VitaminE"
## [16] "VitaminD"     "HighSodium"   "HighCals"     "HighSugar"    "HighProtein"
## [21] "HighFat"
```

```
mydata <- USDA_Clean[,c('Calories','Protein','TotalFat','Carbohydrate','Sodium','Cholesterol')]
print(head(mydata, n = 5))
```

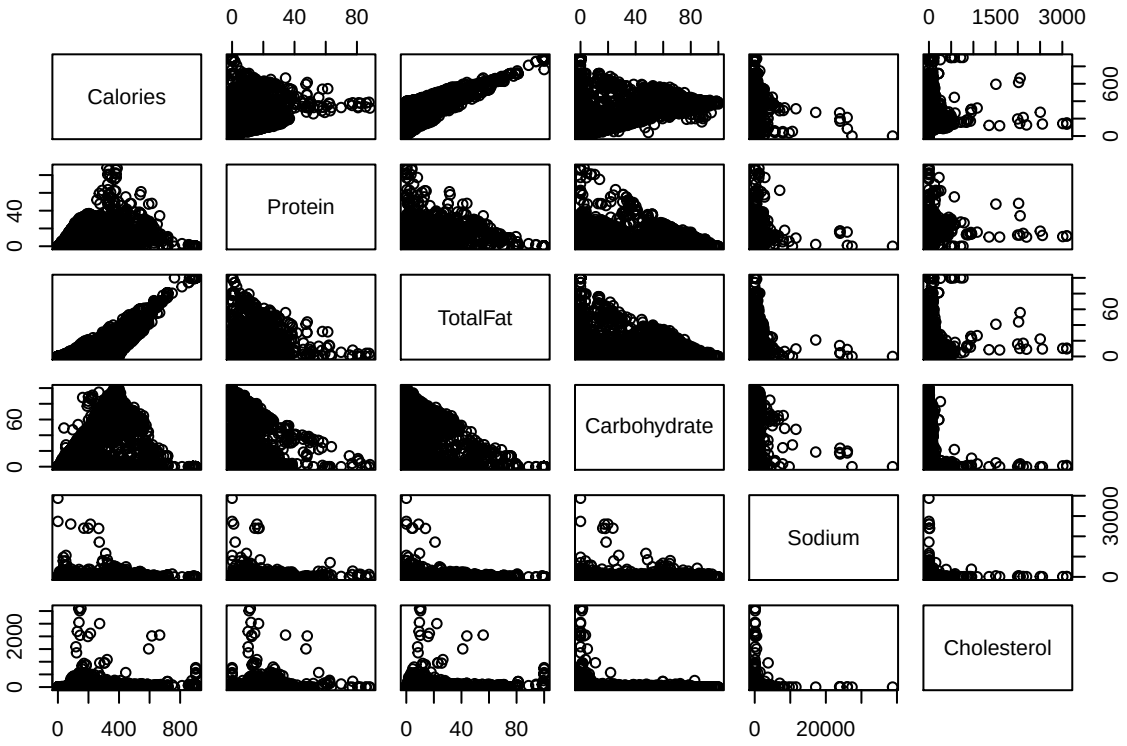
3. Visualize the correlation among Calories, Protein, Total Fat, Carbohydrate, Sodium and Cholesterol. (7 points)

```
##   Calories Protein TotalFat Carbohydrate Sodium Cholesterol
## 1      717    0.85    81.11         0.06    714         215
## 2      717    0.85    81.11         0.06    827         219
## 3      876    0.28    99.48         0.00      2         256
## 4      353   21.40    28.74         2.34   1395          75
## 5      371   23.24    29.68         2.79    560          94
```

```
cor(mydata)
```

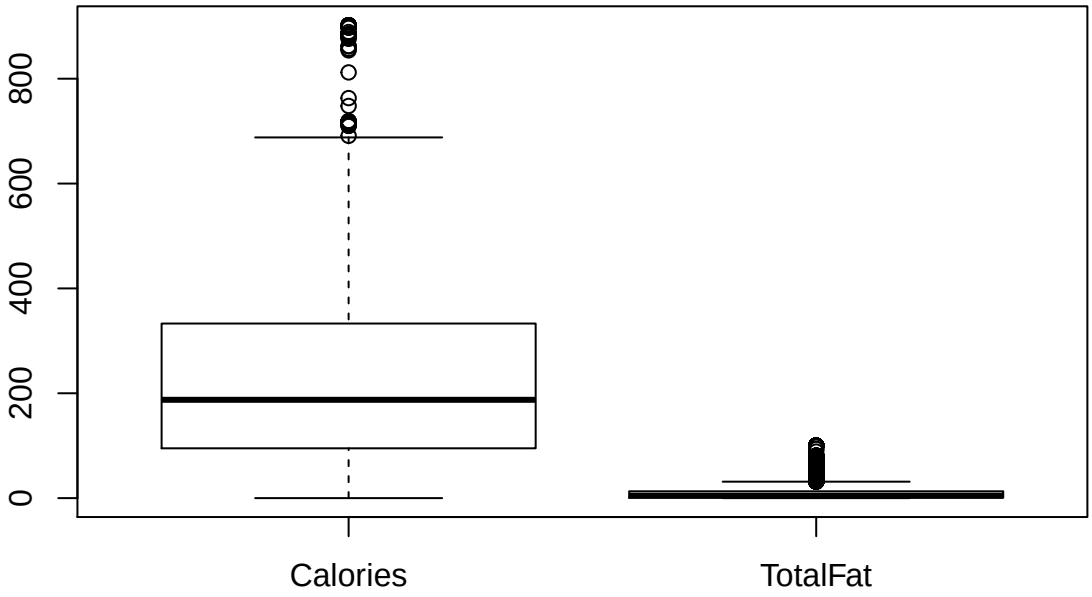
```
##                Calories      Protein     TotalFat Carbohydrate       Sodium
## Calories     1.00000000  0.122122537  0.804495022   0.42460618  0.032321026
## Protein      0.12212254  1.000000000  0.057035611  -0.30471117 -0.003489485
## TotalFat     0.80449502  0.057035611  1.000000000  -0.12434291  0.002916089
## Carbohydrate 0.42460618 -0.304711167 -0.124342914   1.00000000  0.046838692
## Sodium       0.03232103 -0.003489485  0.002916089   0.04683869  1.000000000
## Cholesterol  0.02391933  0.269854840  0.093289601  -0.21937986 -0.017774863
##               Cholesterol
## Calories       0.02391933
## Protein        0.26985484
## TotalFat       0.09328960
## Carbohydrate -0.21937986
## Sodium       -0.01777486
## Cholesterol   1.00000000
```

```
plot((mydata))
```



```r
attach(USDA_Clean)
#Outlier Check to determine the correlation method
boxplot(Calories, TotalFat, names = c('Calories','TotalFat'))
```

4. Is the correlation between Calories and Total Fat statistically significant? Why? (7 points)




```r
#Outlier present, so will apply spearman correlation to see the whether/not variables are significant
cor.test(Calories,TotalFat, method='spearman', exact = F)
```

```
##
##  Spearman's rank correlation rho
##
## data:  Calories and TotalFat
## S = 1.1303e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7300579
```

```r
print('Satistically significant as p-value is less than 0.05')
```

```
## [1] "Satistically significant as p-value is less than 0.05"
```

```r
model_lm <- lm(Calories~Protein + TotalFat + Carbohydrate + Sodium + Cholesterol)
print(summary(model_lm))
```

5. Create a Linear Regression Model, using Calories as the dependent variable Protein, Total Fat, Carbohydrate, Sodium and Cholesterol as the independent variables. (7 points)

```
##
## Call:
## lm(formula = Calories ~ Protein + TotalFat + Carbohydrate + Sodium +
##     Cholesterol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -191.087   -3.832    0.426    5.147  291.011
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.9882753  0.4832629   8.253  < 2e-16 ***
## Protein      3.9891994  0.0233550 170.807  < 2e-16 ***
## TotalFat     8.7716980  0.0143291 612.158  < 2e-16 ***
## Carbohydrate 3.7432001  0.0091404 409.522  < 2e-16 ***
## Sodium       0.0003383  0.0002189   1.545    0.122
## Cholesterol  0.0110138  0.0019861   5.545 3.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.92 on 6304 degrees of freedom
## Multiple R-squared:  0.9877, Adjusted R-squared:  0.9877
## F-statistic: 1.009e+05 on 5 and 6304 DF,  p-value: < 2.2e-16
```

```
coef(model_lm)
```

6. Write the Linear Regression Equation, using Calories as the dependent variable whereas Protein, TotalFat, Carbohydrate, Sodium and Cholesterol as the independent variables. (7 points)

```
##  (Intercept)      Protein     TotalFat Carbohydrate       Sodium  Cholesterol
## 3.9882752613 3.9891994394 8.7716980068 3.7432000604 0.0003383021 0.0110138110
```

```
print('equation = 3.9882753 + 3.9891994*Protein + 8.7716980*TotalFat + 3.7432001*Carbohydrate +
      0.0003383*Sodium + 0.0110138*Cholesterol')
```

```
## [1] "equation = 3.9882753 + 3.9891994*Protein + 8.7716980*TotalFat + 3.7432001*Carbohydrate + \n
```

```
#Will use Anova to check the significance of independent variable
anova(model_lm)
```

7. Which independent variable is the least significant? Why? (7 points)

```
## Analysis of Variance Table
##
## Response: Calories
##                Df    Sum Sq   Mean Sq    F value    Pr(>F)
```

```
## Protein         1   2728899   2728899 7.6197e+03 < 2.2e-16 ***
## TotalFat        1 116762840 116762840 3.2603e+05 < 2.2e-16 ***
## Carbohydrate    1  61215495  61215495 1.7093e+05 < 2.2e-16 ***
## Sodium          1       789       789 2.2031e+00    0.1378
## Cholesterol     1     11014     11014 3.0753e+01  3.05e-08 ***
## Residuals    6304   2257685       358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print('Sodium is least significant independent variable as from the anova table its noted
      the p-value is greater than 0.05.')
```

```
## [1] "Sodium is least significant independent variable as from the anova table its noted \n      the
```

```
x = data.frame (Protein = 0.1,
TotalFat = 35,
Carbohydrate = 405,
Sodium = 440,
Cholesterol = 70)

predict_calories = predict(model_lm,x)
print(predict_calories)
```

8. A new product is just produced with the following data: Protein=0.1, TotalFat=35, Carbohydrate=405, Sodium=440, Cholesterol=70, Sugar=NA, Calcium=35, Iron=NA, Potassium=35, VitaminC=10, VitaminE=NA, VitaminD=NA. Based on the model you created, what is the predicted value for Calories? (7 points)

```
##        1
## 1828.312
```

```
Actual = predict_calories
print(Actual)
```

9. If the Sodium amount increases from 440 to 44440 (10000% increase), how much change will occur on Calories in percent? Explain why? (7 points)

```
##        1
## 1828.312
```

```
Percent = 100
sodium_new = 44440
Newvalue = 3.9882753 + 3.9891994*x$Protein + 8.7716980*x$TotalFat + 3.7432001*x$Carbohydrate + 0.000338
print(Newvalue)
```

```
## [1] 1843.198
```

```
Change = Newvalue - Actual
print(Change)
```

```
##        1
## 14.88521
```

```
percentchange = round((Change/Actual) * Percent, digits = 2)
cat(percentchange,'% Change in Calories')
```

```
## 0.81 % Change in Calories
```

```
print(',explained by the workings as shown. ')
```

```
## [1] ",explained by the workings as shown. "
```

10. A study of primary education asked elementaty school students to retell two book articles that they read earlier in the week. The first (Article 1) had no pictures, and the second (Article 2) was illustrated with pictures. An expert listened to recordings of the students retelling each article and assigned a score for certain uses of language. Higher scores are better. Here are the data for five readers in this study:

Article 1 0.40 0.72 0.00 0.36 0.55

Article 2 0.77 0.49 0.66 0.28 0.38

```
print( '$H_0$ the mean score of students are equal')
```

A) What are H_0 and H_a ? (5 points)

```
## [1] "$H_0$ the mean score of students are equal"
```

```
print( '$H_a$ the mean score of students are not equal.')
```

```
## [1] "$H_a$ the mean score of students are not equal."
```

```
print('paired as booth student reads the articles')
```

B) Is this a paired or unpaired experiment? (5 points)

```
## [1] "paired as booth student reads the articles"
```

```
print('Wilcox rank sum test')
```

C) Based on your previous answer, which nonparametric test statistic would you use to compare the means of Article 1 and Article 2. (5 points)

```
## [1] "Wilcox rank sum test"
```

```
article1 <- c( 0.40,  0.72,   0.00,   0.36,   0.55)
article2 <- c(0.77,   0.49,   0.66,   0.28,   0.38)
print(article1)
```

D) Use a nonparametric test statistic to check if there is a statistically significant difference between the means of Article 1 and Article 2. (5 points)

```
## [1] 0.40 0.72 0.00 0.36 0.55
```

```
print(article2)
```

```
## [1] 0.77 0.49 0.66 0.28 0.38
```

```
wilcox.test(article1,article2)
```

```
##
##  Wilcoxon rank sum test
##
## data:  article1 and article2
## W = 10, p-value = 0.6905
## alternative hypothesis: true location shift is not equal to 0
```

```
print('There is no satistically significant different between the means of the article 1 and 2 as
the p-value is greater than 0.05 so not enough evidence to reject null hypothesis')
```

```
## [1] "There is no satistically significant different between the means of the article 1 and 2 as \nth
```

```
print('The p-value is greater than confidence level=0.05 so not statistically significant and not enoug
      to reject null hypothesis. Illustration does not improve how the student retell an article.')
```

E) Will you accept or reject your Null Hypothesis? ($\alpha = 0.05$) Do illustrations improve how the students retell an article or not? Why? (5 points)

```
## [1] "The p-value is greater than confidence level=0.05 so not statistically significant and not enou
```

11. Two companies selling toothpastes with the lable of 100 grams per tube on the package. We randomly bought eight toothpastes from each company A and B from random stores. Afterwards, we scaled them using high precision scale. Our measurements are recorded as follows:

Company A: 97.1 101.3 107.8 101.9 97.4 104.5 99.5 95.1

Company B: 103.5 105.3 106.5 107.9 102.1 105.6 109.8 97.2

```
CompanyA <- c(97.1,   101.3,   107.8,   101.9,   97.4,    104.5,   99.5,    95.1)
CompanyB <- c(103.5,  105.3,   106.5,   107.9,   102.1,   105.6,   109.8,   97.2)
CompanyA
```

A) Is this a paired or unpaired experiment? (5 points)

```
## [1]  97.1 101.3 107.8 101.9  97.4 104.5  99.5  95.1
```

```
CompanyB
```

```
## [1] 103.5 105.3 106.5 107.9 102.1 105.6 109.8  97.2
```

```
print('This is a unpaired experiment.')
```

```
## [1] "This is a unpaired experiment."
```

```
print('Will use Wilcox rank sum test to compare the means of Company A and Company B)')
```

B) Based on your previous answer, which nonparametric test statistic would you use to compare the means of Company A and Company B. (5 points)

```
## [1] "Will use Wilcox rank sum test to compare the means of Company A and Company B)"
```

```
wilcox.test(CompanyA,CompanyB)
```

C) Use a nonparametric test statistic to check if there is a statistically significant difference between the means of Company A and Company B. (5 points)

```
##
##  Wilcoxon rank sum test
##
## data:  CompanyA and CompanyB
## W = 13, p-value = 0.04988
## alternative hypothesis: true location shift is not equal to 0
```

```
print('The p-value is less than 0.05, there is statistically sigficance that the means
      of Company A and Company B is different')
```

```
## [1] "The p-value is less than 0.05, there is statistically sigficance that the means \n      of Compa
```

```
print('Will reject null hypothesis as the p-value is less than alpha 0.05,
      packaging process is different based on weight measurement')
```

D) Will you accept or reject your Null Hypothesis? ($\alpha = 0.05$) Are packaging process similar or different based on weight measurements? Why? (5 points)

```
## [1] "Will reject null hypothesis as the p-value is less than alpha 0.05, \n      packaging process is
```

This is the end of Assignment 2

Ceni Babaoglu, PhD