CMTH 642 - Assignment 2

USDA Clean Data

We uplodaded the clean csv file generated from Assignment 1 (USDA_Clean.csv). Please download and load it to your workspace.

```
#USDAclean = read.csv("USDA_Clean.csv")
#attach(USDA_Clean) ## Optional
# attch() function helps you to access USDA_Clean without the need of menioning it.
# For example, you can use Calories instead of USDA_Clean$Calories
#View(USDA_Clean)
#str(USDAclean)
```

Visualization of Feature Relationships

We have used a function panel.cor() inside pair() to show the correlations among different features. The only line you should complete is the line that you assign a value to **USDA_Selected_Featuers**. Research how can you select multiple columns from a dataframe to use it inside pair() function.

A) Show the relationship among *Calories*, *Carbohydrate*, *Protein*, *Total Fat* and *Sodium*. **(5 p)**

B) Describe the correlations among **Calories** and other features. **(5 p)**

Hint: We usually interpret the absolute value of correlation as follows:

.00-.19 *very weak*

.20-.39 *weak*

.40-.59 *moderate*

.60-.79 *strong*

.80-1.0 *very strong*

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

    test <- cor.test(x,y)
    # borrowed from printCoefmat
    Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                  cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                  symbols = c("***", "**", "*", ".", " "))

    text(0.5, 0.5, txt, cex = cex * r)
    text(.8, .8, Signif, cex=cex, col=2)
}
# Assign a value USDA_Selected_Featuers that represents
# "Calories","Carbohydrate","Protein","TotalFat", "Sodium" columns
#####################################################
##### Complete code here and uncomment it
#USDA_Selected_Featuers <-
#####################################################

#### Uncomment the following line when you assign USDA_Selected_Featuers to show the results
#pairs(USDA_Selected_Featuers, lower.panel=panel.smooth, upper.panel=panel.cor)

# Explain what you can conclude from this visualization as a comment here
```

Regression Model on USDA Clean Data

Create a Linear Regression Model (lm), using **Calories** as the dependent variable, and *Carbohydrate*, *Protein*, *Total Fat* and *Sodium* as independent variables. **(10 p)**

```
# Write your code here
```

Analyzing Regression Model

A) In the above example, which independent feature is less significant? (Hint: Use ANOVA) **(5 p)**

```
# Write your code here and explain answers as a comment
```

B) Which independent variable has the strongest positive predictive power in the model? (Hint: Look at the coefitients calculated for each independant variable) **(5 p)**

```
# Write your code here and explain answers as a comment
```

Calories Prediction

A new product is just produced with the following data:

“Protein” “TotalFat” “Carbohydrate” “Sodium” “Cholesterol”

0.1 40 425 430 75

“Sugar” “Calcium” “Iron” “Potassium” “VitaminC” “VitaminE” “VitaminD”

NA 42 NA 35 10 0.0 NA

A) Based on the model you created, what is the predicted value for **Calories** ? **(5 p)**

B) If the *Sodium* amount increases 101 times from 430 to 43430 (10000% increase), how much change will occur on Calories in percent? Can you explain why? **(5 p)**

```
# Write your code here and explain answers as a comment
```

Wilcoxon Tests

Research Question: Does illustrations improve memorization?

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 picutres, and the second (Article 2) 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 a this study:

Student 1 2 3 4 5

Article 1 0.40 0.72 0.00 0.36 0.55

Article 2 0.77 0.49 0.66 0.28 0.38

We wonder if illustrations improve how the students retell an article.

What is H_0 and H_a ?

(10 p)

```
# Write your Answer as a comment here
```

Paired or Independent design?

Based on your answer, which Wilcoxon test should you use? **(5 p)**

```
# Write your Answer as a comment here
```

Will you accept or reject your Null Hypothesis? ($\alpha = 0.05$)

Do illustrations improve how the students retell an article or not? **(5 p)**

```
# Write your code here
```

Packaging Problem

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

Distribution Analysis

Are the distributions of package weights similar for these companies? Are they normally distributed or skewed? **(10 p)** (Hint: Use boxplot)

```
# Write your code here
```

Are packaging process similar or different based on weight measurements?

Can we be at least 95% confident that there is no difference between packaging of these two companies? **(5 p)**

Can we be at least 99% confident? **(5 p)**

Please explain.

```
# Write your code here and explain answers as a comment
```

Correlation

Plot and see the relationship between “cylinder” (cyl) and “weight” (wt) of the cars from mtcars dataset. A) Can you see any patterns of correlation between these two variable? **(5 p)**

```
# Write your code here and explain answers as a comment
```

B) What is the best description for “cyl” and “wt” variables? (Ratio, Ordinal, Interval, or Categorical) **(5 p)**

```
# Explain answers as a comment here
```

C) Based on the description of the “cyl” and “wt” variables, should you use “Pearson” or “Spearman” correlation? Find the correlation between these two variables. **(10 p)**

```
# Write your code here and explain answers as a comment
```