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*

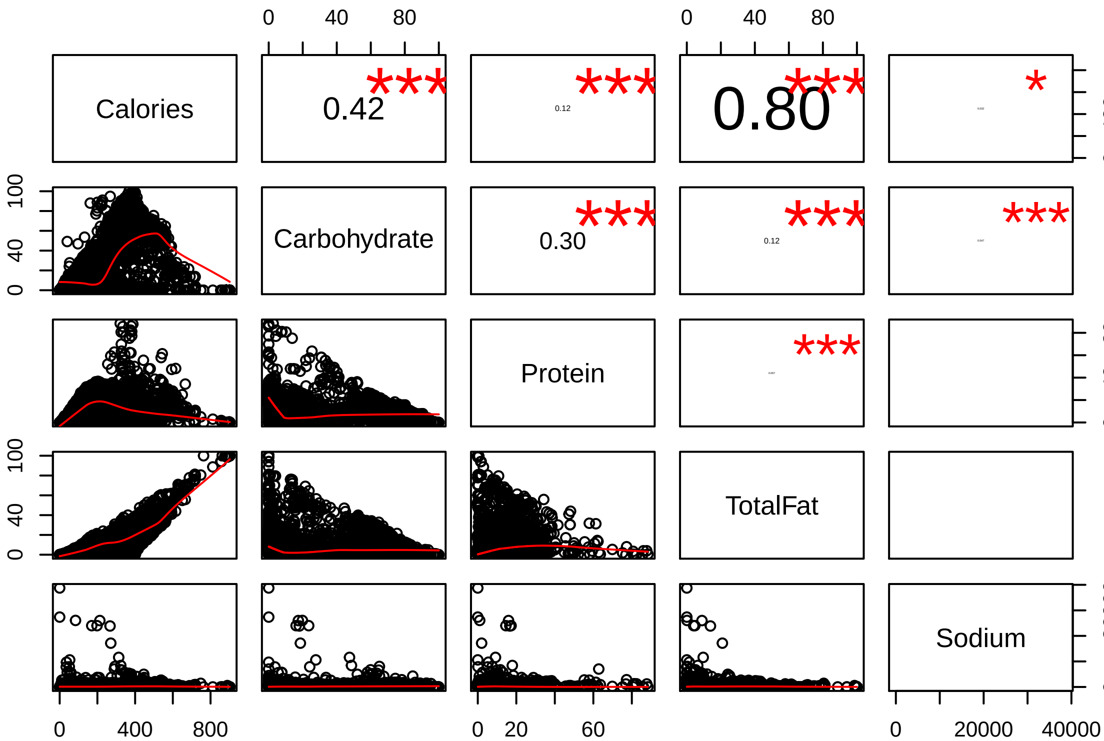
```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

    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
```

```r
# "Calories","Carbohydrate","Protein","TotalFat", "Sodium" columns
####################################################
##### Complete code here and uncomment it
USDA_Selected_Featuers <- USDAclean[,c('Calories','Carbohydrate','Protein','TotalFat','Sodium')]
####################################################

#### 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)
```




```r
# Explain what you can conclude from this visualization as a comment here
# Calories and TotalFat have the highest correlation: (0.8). Second is Calroies and Carbohydrate (0.42)
```

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)**

```r
# Write your code here
food.lm <- lm(Calories ~ Carbohydrate + Protein + TotalFat + Sodium, data=USDA_Selected_Featuers)

# for(attr in c('Carbohydrate','Protein','TotalFat','Sodium')){
  # USDA_Selected_Featuers[,attr] <- as.factor(USDA_Selected_Featuers[,attr])
# }

# Below is a shorthand (since we have already subsetted the dataframe):
# food.lm <- lm(Calories ~ ., data=USDA_Selected_Featuers)
```

Analyzing Regression Model

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

```r
# Write your code here and explain answers as a comment
anova(food.lm)
```

```
## Analysis of Variance Table
##
## Response: Calories
##               Df    Sum Sq   Mean Sq    F value Pr(>F)
## Carbohydrate   1  32988948  32988948 9.1680e+04 <2e-16 ***
## Protein        1  12758767  12758767 3.5458e+04 <2e-16 ***
## TotalFat       1 134959519 134959519 3.7507e+05 <2e-16 ***
## Sodium         1       789       789 2.1927e+00 0.1387
## Residuals   6305   2268698       360
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Sodium is least significant (has highest p-value)
```

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)**

```r
# Write your code here and explain answers as a comment
summary(food.lm)
```

```
##
## Call:
## lm(formula = Calories ~ Carbohydrate + Protein + TotalFat + Sodium,
##     data = USDA_Selected_Featuers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -191.521   -3.917    0.596    5.126  290.787
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.2126623  0.4827009   8.727   <2e-16 ***
## Carbohydrate 3.7360470  0.0090703 411.901   <2e-16 ***
## Protein      4.0174012  0.0228483 175.830   <2e-16 ***
## TotalFat     8.7768988  0.0143321 612.394   <2e-16 ***
## Sodium       0.0003249  0.0002194   1.481    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.97 on 6305 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9876
## F-statistic: 1.256e+05 on 4 and 6305 DF,  p-value: < 2.2e-16
```

```r
# I'm guessing that TotalFat has the highest predictive power because it has the largest co-efficient
```

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)**

```
predict(food.lm, data.frame(Protein=0.1,TotalFat=40,Carbohydrate=425,Sodium=430,Cholesterol=75),interva
```

```
##       fit      lwr      upr
## 1 1943.65 1905.774 1981.526
```

```
# Predicted Value: 1943.65
```

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

predict(food.lm, data.frame(Protein=0.1,TotalFat=40,Carbohydrate=425,Sodium=43430,Cholesterol=75),inter
```

```
##        fit      lwr     upr
## 1 1957.622 1915.605 1999.64
```

```
# Increasing the sodium from 430 to 43430 increases the predicted calories from 1943.65 to 1957.622
# This is because Sodium has the lowest coefficient in the model (0.0003 vs 3.736 for the next lowest,
```

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

# Independent
# H0: The distributions for populations 1 and 2 are identical
# Ha: There is a difference in the distributions

# Paired
```

```r
# H0: The two population relative frequency distributions are identical
# Ha: The two population relative frequency distributinos differ in location
```

Paired or Independent design?

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

```r
# Write your Answer as a comment here
# Paired deisgn: Same article with or without pictures? Or else the pairing is the student
```

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

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

```r
# Write your code here
A1 <- c(0.4, 0.72, 0, 0.36, 0.55)
A2 <- c(0.77, 0.49, 0.66, 0.28, 0.38)
wilcox.test(A1,A2, paired=T)
```

```
##
##  Wilcoxon signed rank test
##
## data:  A1 and A2
## V = 6, p-value = 0.8125
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Reject H0: Illustrations do not improve how students retell an article (p > 0.05)
```

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:

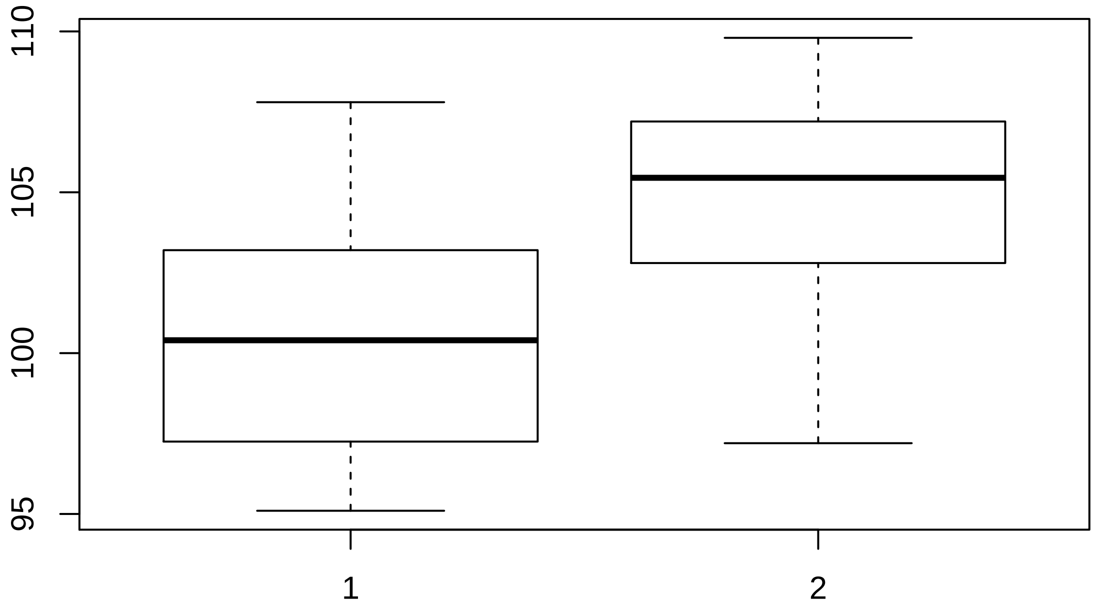
Company A: 97.1 101.3 107.8 101.9 97.4 104.5 99.5 95.1

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)

```r
# Write your code here
Ca <- c( 97.1 , 101.3 , 107.8 , 101.9 , 97.4,    104.5,   99.5,    95.1)
Cb <- c( 103.5 ,105.3  ,106.5  ,107.9  ,102.1  ,105.6 , 109.8,  97.2)
boxplot(Ca,Cb)
```




```
# Distributions appear different: Ca has a slight right skew while Cb has a left skew
# Normality assumption appears to be violated, as does common variance
```

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
wilcox.test(Ca,Cb,paired=T)
```

```
##
##  Wilcoxon signed rank test
##
## data:  Ca and Cb
## V = 2, p-value = 0.02344
## alternative hypothesis: true location shift is not equal to 0
```

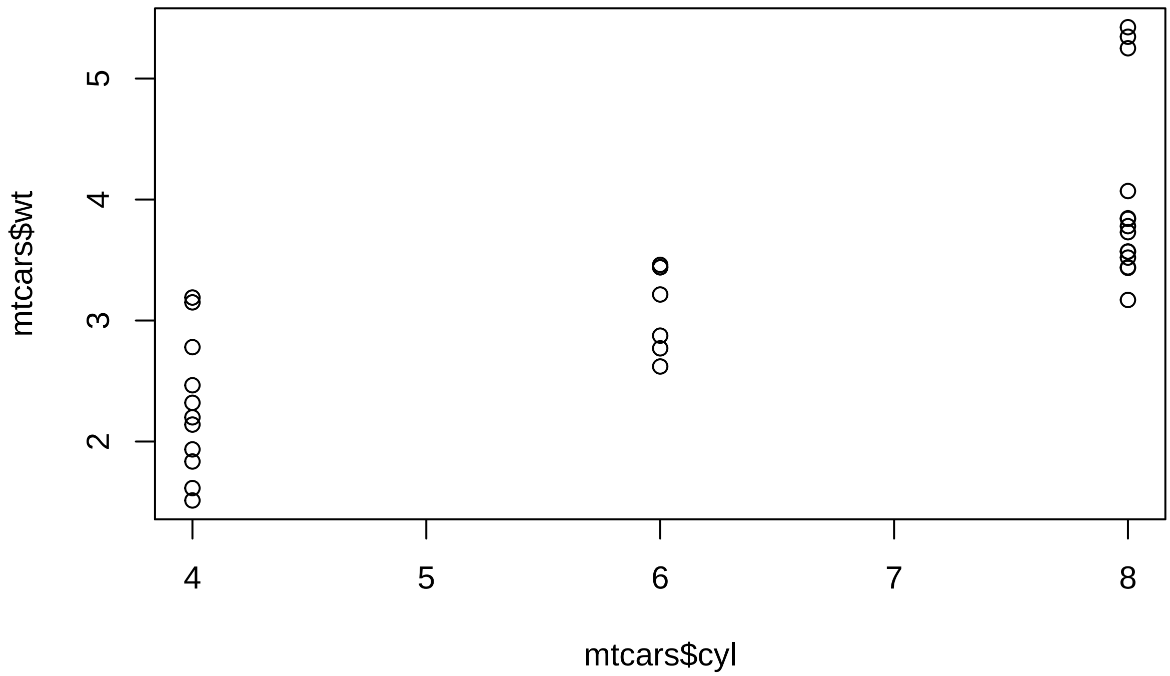
```
# H0: The two population relative frequency distributions are identical
# Ha: The two population relative frequency distributinos differ in location

# You can reject H0 with 95% confidence but not 99%
# p-value = 0.023 (0.95 <= (1 - a) <= 0.99)
```

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
plot(mtcars$cyl, mtcars$wt)
```




```r
# There appears to be a slight positive correlation
```

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

```r
# Explain answers as a comment here
# cyl: categorical
# wt: ratio (continuous)

# Weight is unambiguously a continuous ratio. I chose categorical for cylinders
# because there aren't really any ranks or order between the number of cylinders.
# Four cylinders is neither greater nor less than six, it is more used as a
# description of what kind of engine the vehicle has (V6, Straight-6, Flathead V8 ETC)
```

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)**

```r
# Write your code here and explain answers as a comment
# I would use Pearson's correlation because the model accounts for categorical variables
# (using dummy values). Since ranking cylinders doesn't make sense I wouldn't use Spearman.
cor(mtcars$cyl, mtcars$wt) #0.78 strong positive correlation
```

```
## [1] 0.7824958
```