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.

```
loc = getwd()
USDAclean = read.csv(paste(loc, "USDA_Clean.csv", sep = "/"))
attach(USDAclean) # Optional
# attach() function helps you to access USDA_Clean without the need of
mentioning 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_Features**. 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*

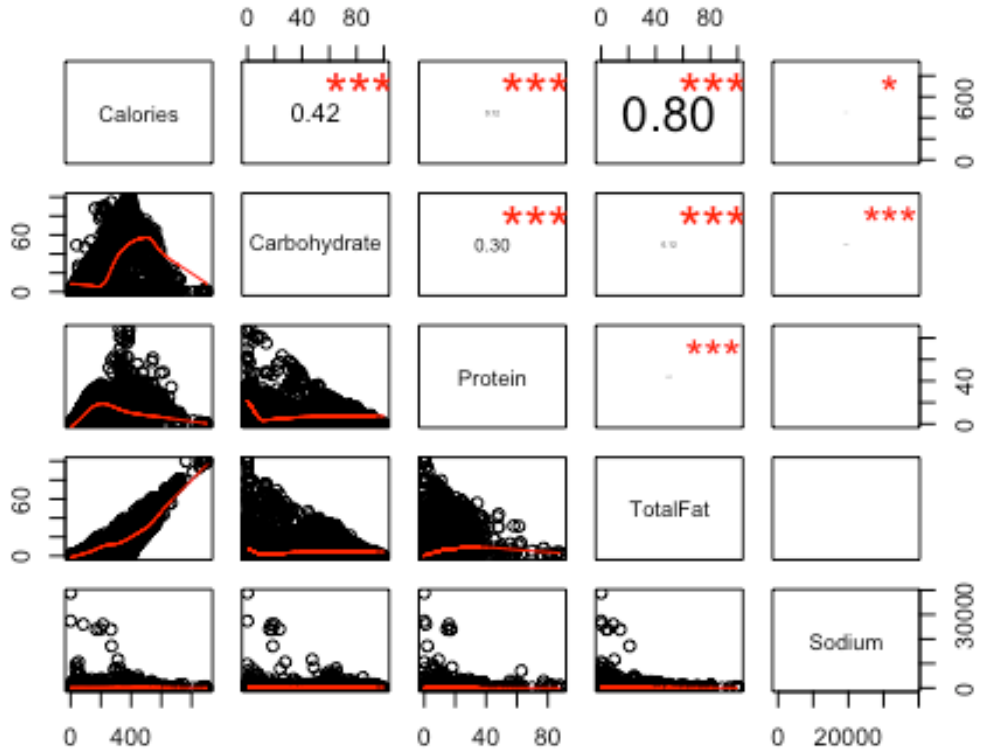
```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

    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_Features that represents
# "Calories","Carbohydrate","Protein","TotalFat", "Sodium" columns
#####################################################
# Complete code here and uncomment it
USDA_Selected_Features <- USDAclean[, c('Calories', 'Carbohydrate',
'Protein', 'TotalFat', 'Sodium')]
#####################################################

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




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

# We can tell that all attributes are highly correlated (due to the
```

```
plots and the presence of three astericks), with the exception of
sodium with all other features.
```

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)**

```
model = lm(Calories ~ Carbohydrate + Protein + TotalFat + Sodium)
```

Analyzing Regression Model

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

```
modelaov = anova(model)
modelaov

## 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

# The Sodium feature is less significant. We can deduce this from the
Analysis of Variance Table. The p-value for sodium is 0.1387, which is
not very significant, especially compared to the p-values of the other
features, which are all much smaller and less than 2e-16.
```

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

```
coef(model)

##  (Intercept) Carbohydrate      Protein     TotalFat       Sodium
## 4.2126623348 3.7360469886 4.0174011556 8.7768987924 0.0003249363

# The independent variable \'TotalFat\' has the strongest positive
predictive power in the model, due to it having the largest
coefficient: 8.78.
```

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)**

```
newrecord = data.frame(Carbohydrate = 425, Protein = 0.1, TotalFat =
40, Sodium = 430)
cat("Predicted value for calories is:", predict(model, newrecord))

## Predicted value for calories is: 1943.65

# If the value of Sodium increased 101 times from 430 to 43430, the
value for Calories would change by 13.97226. This represents a 0.72%
change in the value for Calories from when Sodium was equal to 430. To
get this result we multiply the difference in the Sodium value from
before to after (43430 - 430 = 43000) by the coefficient for Sodium
from the model, which is 0.0003249363. We use this value in our
regression calculation. The coefficient describes the change in the
dependent variable for each unit of change in the Sodium variable.
```

Wilcoxon Tests

Research Question: Does illustrations improve memorization?

A study of primary education asked elementary 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)

```
# H_0: The two population distributions are the same / the mean of the
scores of the students from both groups are equal.
# H_a: The two populations are different / The mean of the scores of
the students from both groups are not equal.
```

Paired or Independent design?

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

```
# The test should not be paired; two different articles are being
compared. We should use the Mann-Whitney rank sum test.
```

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

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

```
article1 = c(0.40, 0.72, 0, 0.36, 0.55)
article2 = c(0.77, 0.49, 0.66, 0.28, 0.38)
wilcox.test(article1, article2, paired = 0)

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

# Illustrations do not improve how well the students retell an article.
We can discern this from the fact that the p-value for the Rank Sum
test is 0.6905, which is much greater than the confidence level of
0.05.
```

Packaging Problem

Two companies selling toothpastes with the label 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:

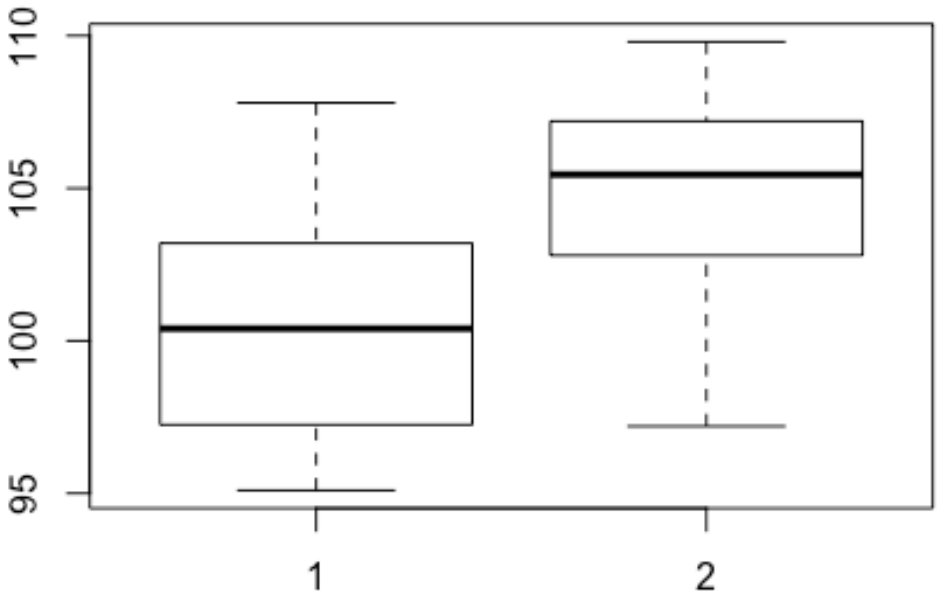
Company A: 97.1 101.3 107.8 101.9 97.4 104.5 99.5 95.1

Company B: 103.5 105.3 106.5 107.9 102.1 105.6 109.8 97.2

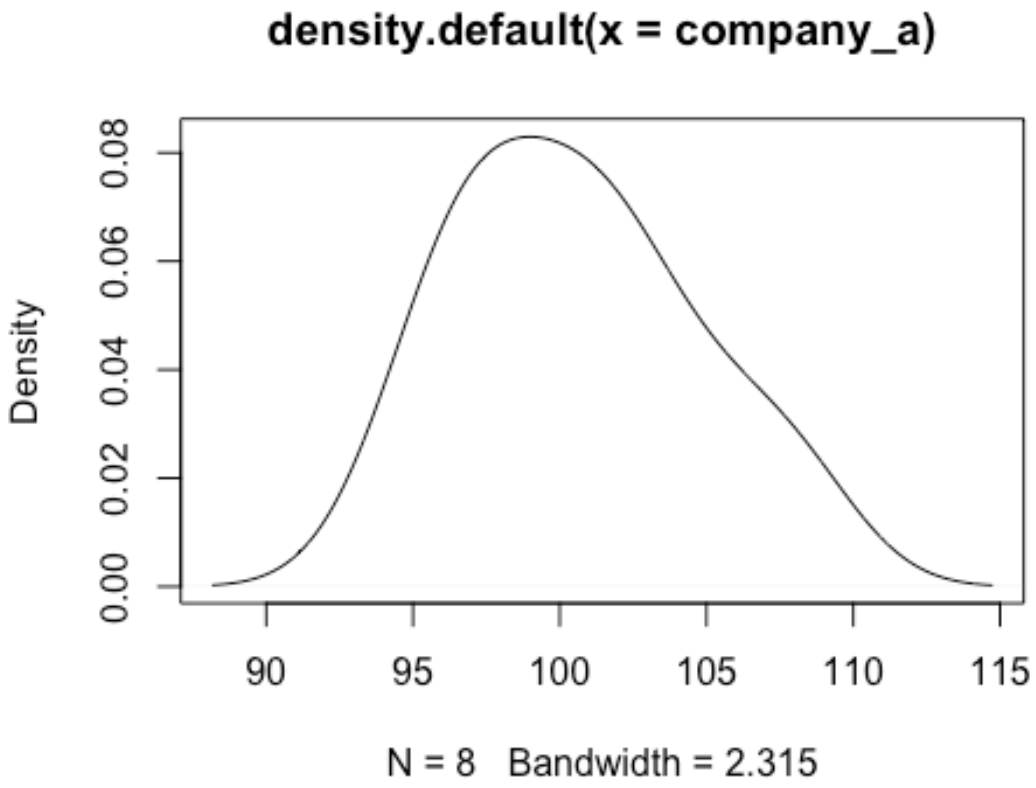
Distribution Analysis

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

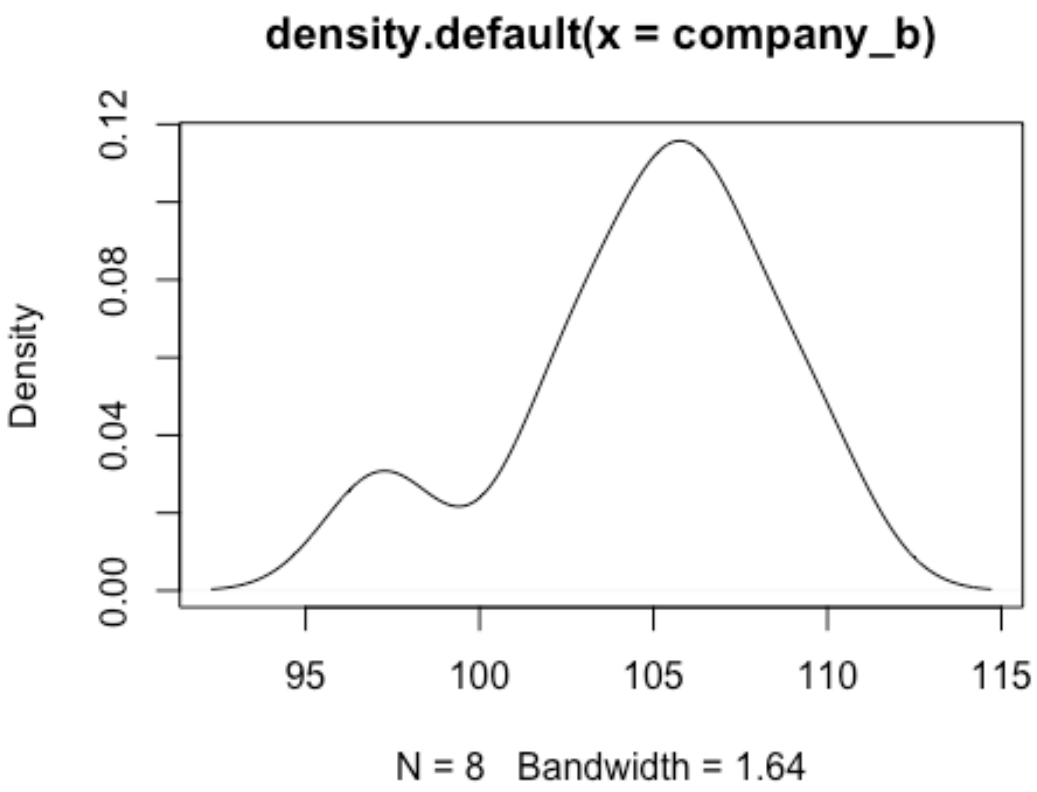
```
company_a = c(97.1, 101.3, 107.8, 101.9, 97.4, 104.5, 99.5, 95.1)
company_b = c(103.5, 105.3, 106.5, 107.9, 102.1, 105.6, 109.8, 97.2)
boxplot(company_a, company_b)
```




```
plot(density(company_a))
```




```r
plot(density(company_b))
```




```
wilcox.test(company_a, company_b, paired = 0)

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

# From the p-value of 0.04988 we conclude that the package weights of
Company A and Company B are different. Looking at the boxplots and the
density plots for each company, we can also conclude that the
distributions are slightly skewed and not perfectly normal, mainly
Company B.
```

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)**

```
t.test(company_a, company_b)
```

```
##
##  Welch Two Sample t-test
##
## data:  company_a and company_b
## t = -2.0617, df = 13.913, p-value = 0.05844
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.4953497  0.1703497
## sample estimates:
## mean of x mean of y
##  100.5750  104.7375
```

```
# We can be at least 95% confident, because our p-value is lower than
0.05.
```

```
t.test(company_a, company_b, conf.level = 0.99)
```

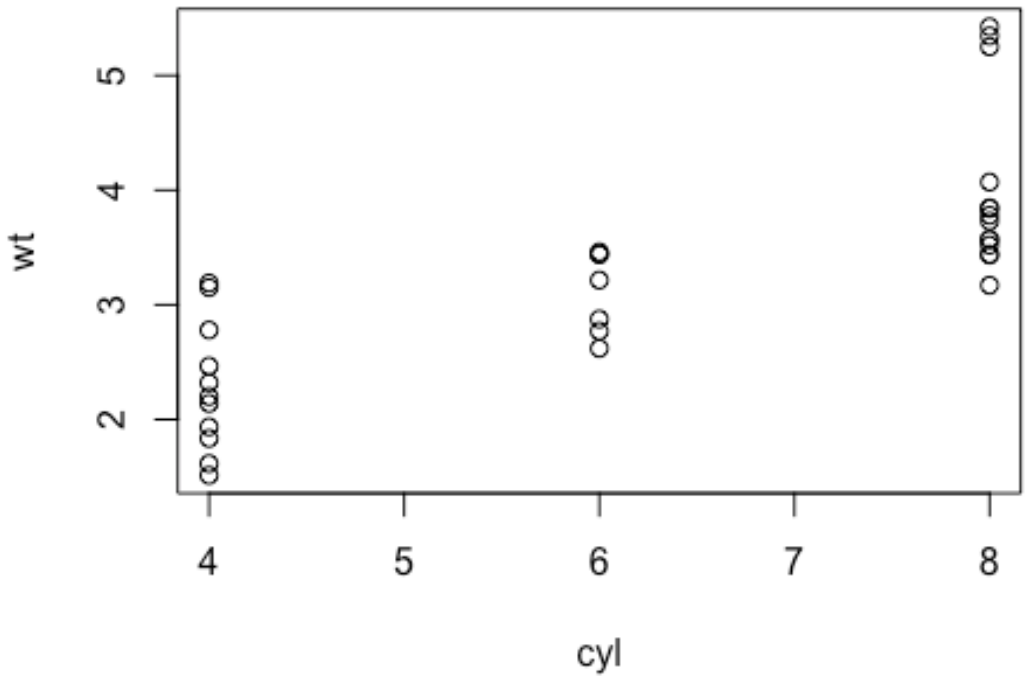
```
##
##  Welch Two Sample t-test
##
## data:  company_a and company_b
## t = -2.0617, df = 13.913, p-value = 0.05844
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -10.17849   1.85349
## sample estimates:
## mean of x mean of y
##  100.5750  104.7375
```

```
# We cannot be 99% confident, because the p-value is not less than
0.01.
```

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)**

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




```
# A pattern that can be observed is that as the number of cylinders
increases, the weight increases.
```

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

```
# 'cyl' is an ordinal variable.
# 'wt' is a ratio variable.
```

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)**

```
# Because 'cyl' is an ordinal variable, it is not necessarily normally
distributed. Because of this, we should use Spearman correlation.
cor(mtcars$cyl, mtcars$wt, method = "spearman")

## [1] 0.8577282
```