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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
USDAclean = read.csv("C:\\Users\\Derick\\Desktop\\Things to Keep\\USDAclean.csv")
attach(USDAclean) ## 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(USDAclean)
str(USDAclean)
```

```
## 'data.frame':    6310 obs. of  22 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 ...
##  $ HighCalories: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HighProtein : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ HighSugar   : int  0 0 0 0 0 0 0 1 0 1 ...
##  $ HighFat     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HealthCheck : Factor w/ 2 levels "Fail","Pass": 2 2 2 2 2 2 2 1 2 1 ...
```

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*

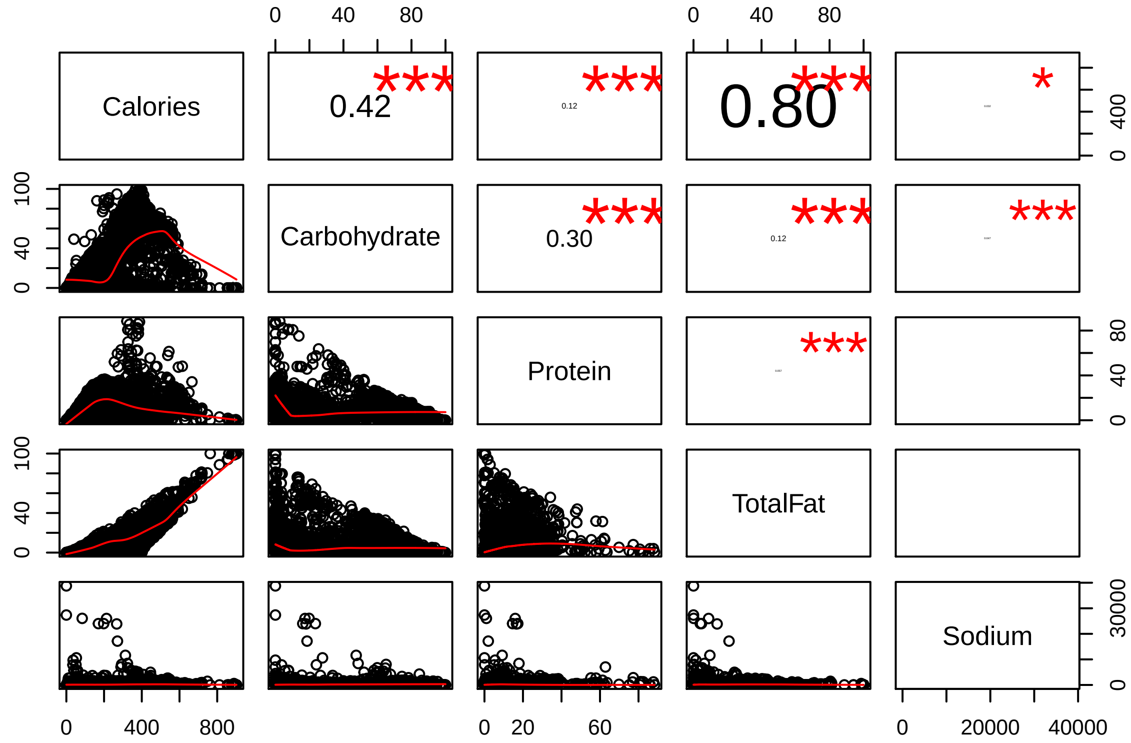
.20-.39 *weak*

.40-.59 *moderate*

.60-.79 *strong*

.80-1.0 *very strong*

```r
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 = data.frame(select(USDAclean, 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)
```



```
# There is a positive correlation between between Calories and TotalFat. The scatterplot shows the data
# The next highest correlation value would be Calories vs Carbohydrates at 0.42 which is also highly si
# A third relationship that should be looked at would be Carbohydrate vs Protein. It is siginificant at
# This would make sense because generally fat contains more calories than carbohydrates or protein and
# Another thing to note is that sodium has almost no correlation to any of the other variables, it is a
```

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)**

```
CalorieRegression = lm(Calories ~ Carbohydrate + Protein + TotalFat + Sodium, data = USDAclean)
summary(CalorieRegression)
```

```
##
## Call:
## lm(formula = Calories ~ Carbohydrate + Protein + TotalFat + Sodium,
##     data = USDAclean)
##
## 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
```

Analyzing Regression Model

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

```
anova(CalorieRegression)
```

```
## 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
```

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)**

```
# The independent feature that is less significant is Sodium. The value with the strongest positive pre
```

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)**

```
predict(CalorieRegression, data.frame(Protein = 0.1, TotalFat = 40, Carbohydrate = 425, Sodium = 430))
```

```
##       1
## 1943.65
```

```
predict(CalorieRegression, data.frame(Protein = 0.1, TotalFat = 40, Carbohydrate = 425, Sodium = 43430)
```

```
##        1
## 1957.622
```

```
# The predicted value for Calories would be 1943.
# The percentage change in calories if we increased Sodium to 43430 is 0.72%. This is equivalent to the
```

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)

```
# $H_0$ : The mean score for illustrations and no illustrations is the same in the population.
# $H_a$ : The mean score for illustrations and no illustrations is not the same in the population.
```

Paired or Independent design?

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

```
# Since the study uses the same students for the first article and the second article, the samples are
```

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
Article1 = c(0.4, 0.72, 0.0, 0.36, 0.55)
Article2 = c(0.77, 0.49, 0.66, 0.28, 0.38)

IllustrationTest = wilcox.test(Article1, Article2, paired = TRUE, alternative = "two.sided")
IllustrationTest
```

```
##
##  Wilcoxon signed rank test
##
## data:  Article1 and Article2
## V = 6, p-value = 0.8125
## alternative hypothesis: true location shift is not equal to 0
```

```r
# The P-value of the test is 0.8125 which is higher than the than the significance level alpha = 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:

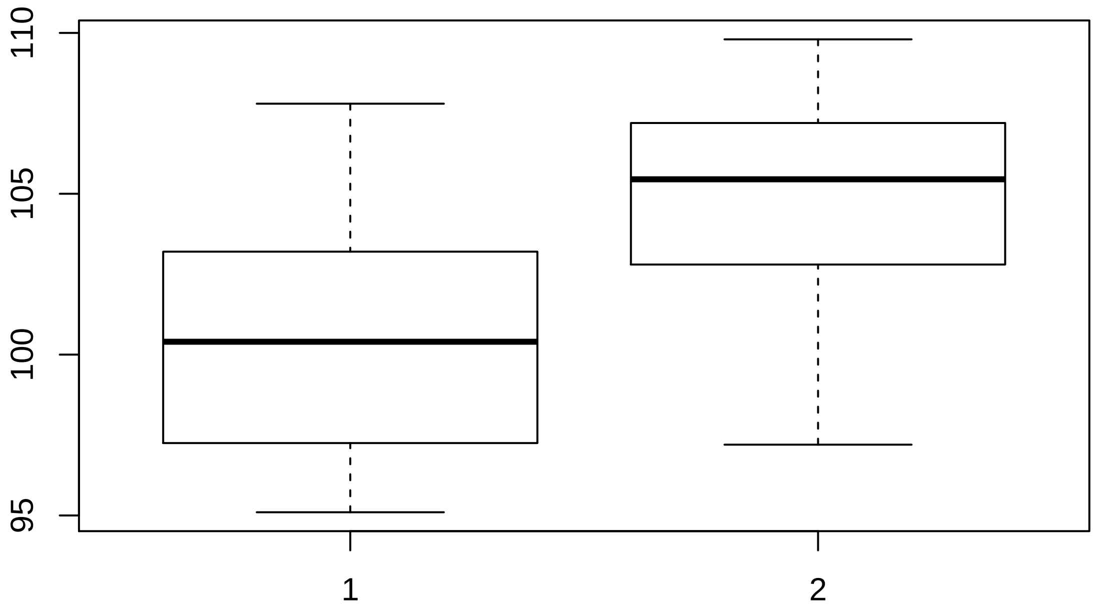
Company A: 97.1 101.3 107.8 101.9 97.4 104.5 99.5 95.1

Company B: 103.5 105.3 106.5 107.9 102.1 105.6 109.8 97.2

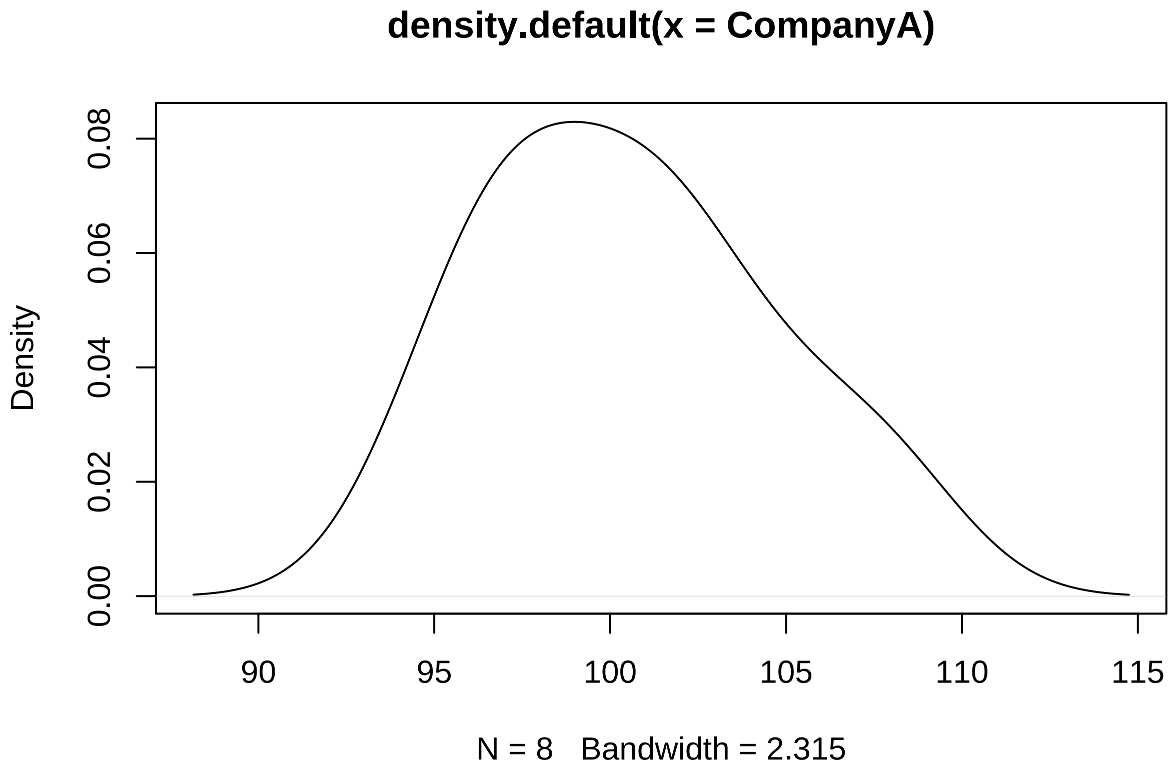
Distribution Analysis

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

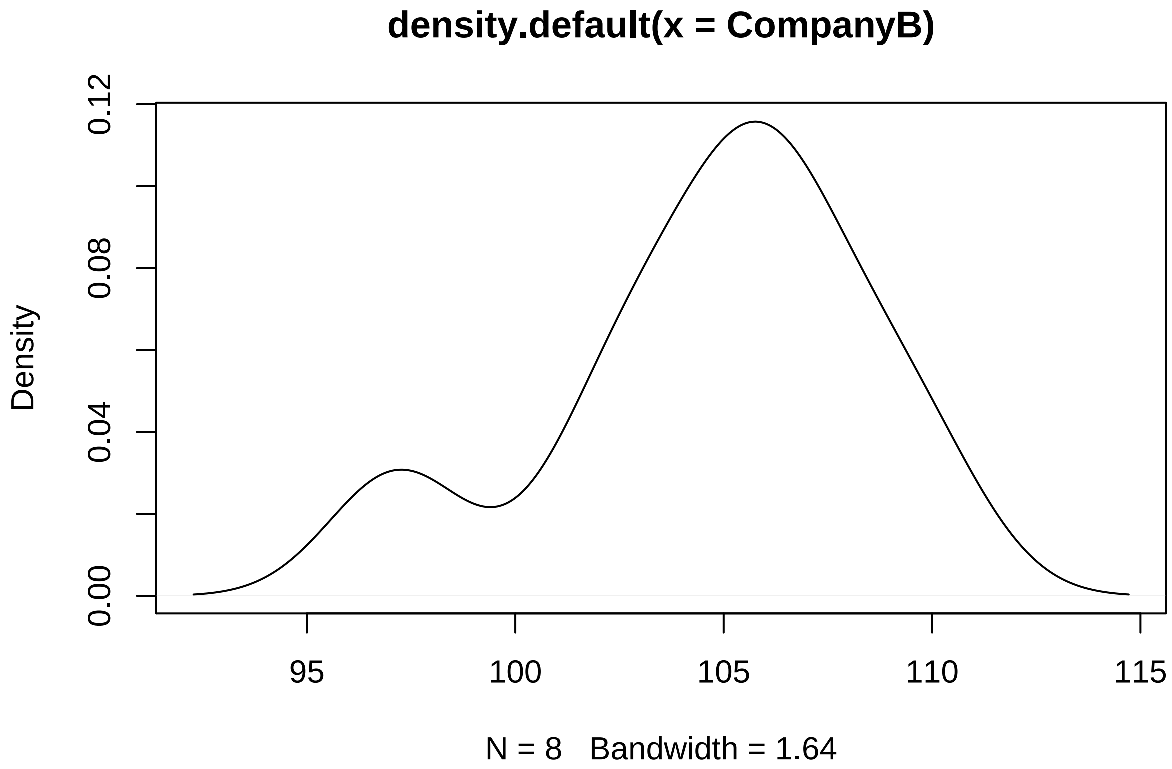
```r
CompanyA = c(97.1, 101.3, 107.8, 101.9, 97.4, 104.5, 99.5, 95.1)
CompanyB = c(103.5, 105.3, 106.5, 107.9, 102.1, 105.6, 109.8, 97.2)

boxplot(CompanyA, CompanyB)
```



```
plot(density(CompanyA))
```



```
plot(density(CompanyB))
```



```
# The distributions are different for these two companies. Company A is normally distributed and Compan
```

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.

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

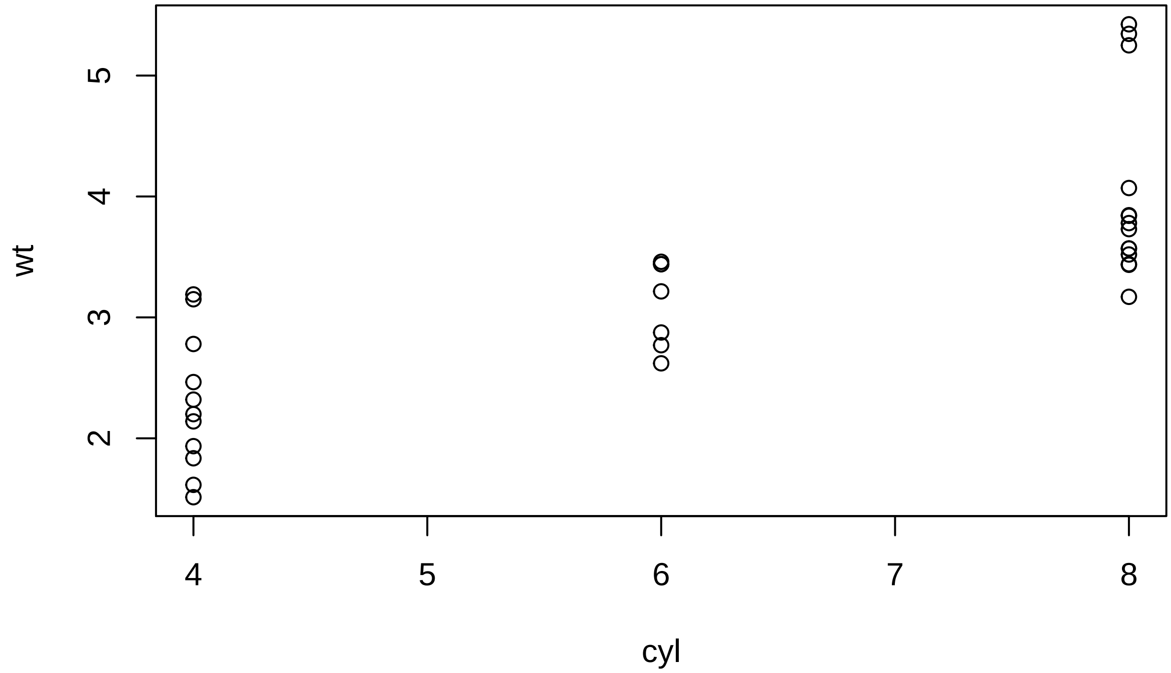
```
##
##  Welch Two Sample t-test
##
## data:  CompanyA and CompanyB
## t = -2.0617, df = 13.913, p-value = 0.05844
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.4953497  0.1703497
## sample estimates:
## mean of x mean of y
##  100.5750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

```
# We can not be at least 95% confident because at the 5% significance level, the p-value is not statisi
```

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)**

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



```
# Based on the plot, the correlation between these two variables is that the more cylinders a car has,
```

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

```
# Weight would be an interval variable. This is because weight carries an order and the differences bet
```

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
# Because cyl is an ordinal variable, it would be better to use Spearman correlation between these two
# The spearman correlation between these two variables is 0.8577.
cor(cyl, wt, method = "spearman")
```

```
## [1] 0.8577282
```