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(file="C:\\Users\\Paul\\Desktop\\USDA_Clean.csv", header=T, sep=",")
#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(USDAClean)
str(USDAClean)
```

```
## '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 ...
```

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*

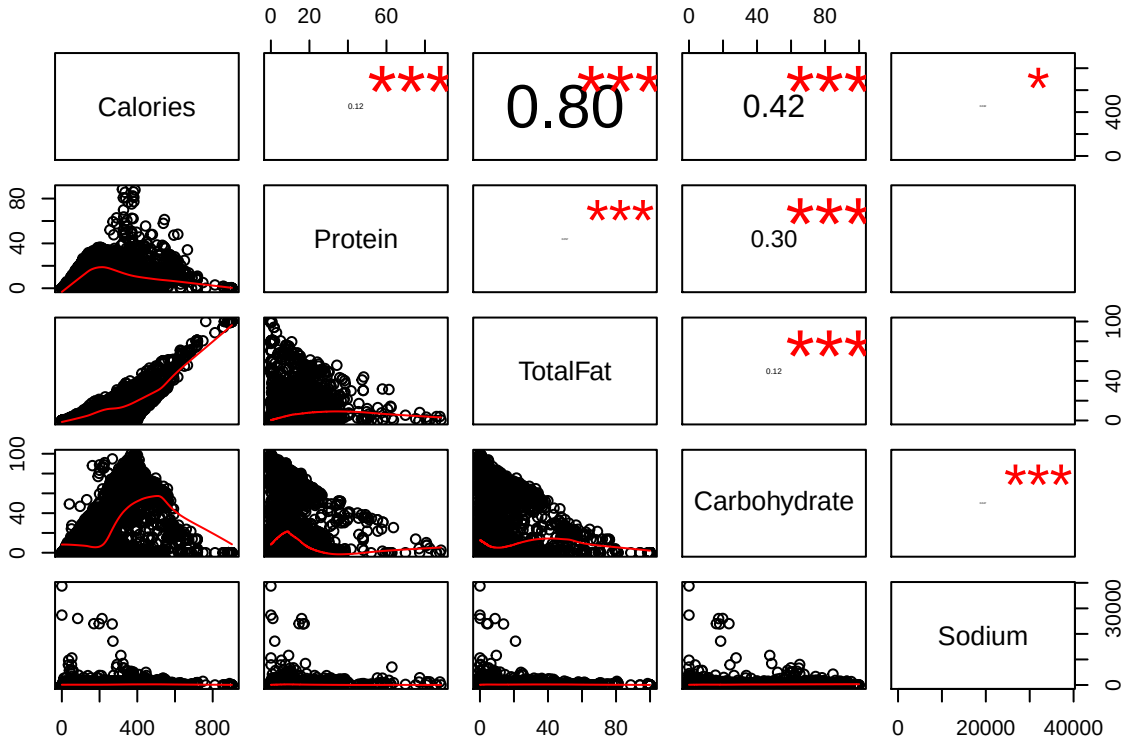
.80-1.0 *very strong*

```
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<-subset(USDAClean, select=c(4:8))
########################################################

#### 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)
```



The function outputs a graphic showing visual and numerical correlations between the variables. The lower triangular represents the graphs, while the upper triangular represents actual correlations. It seems that the stronger a correlation between two variables are, it will display a larger text. For example, Calories and TotalFat have a very strong correlation, Calories and Carbohydrate have a moderate correlation, Protein and Carbohydrate have a weak correlation, and the rest of the relationships have very weak or almost no correlations at all.

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)**

```
calories_dep<-lm(Calories ~ Carbohydrate+Protein+TotalFat+Sodium, data= USDAClean)
calories_dep
```

```
##
## Call:
## lm(formula = Calories ~ Carbohydrate + Protein + TotalFat + Sodium,
##     data = USDAClean)
##
## Coefficients:
##  (Intercept)  Carbohydrate       Protein      TotalFat        Sodium
##    4.2126623     3.7360470     4.0174012     8.7768988     0.0003249
```

Analyzing Regression Model

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

```
anova(calories_dep)
```

```
## 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
```

From the ANOVA table, it is obvious that the Sodium variable is less significant as it has a p-value of 0.1387, in which we cannot predict any change to the data, i.e, the null hypothesis is not rejected.

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)**

```
calories_dep$coefficients
```

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

TotalFat has the strongest positive predictive power, i.e, it has the strongest relationship with calories acting as a dependant variable.

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)**

Based on the model, we determine the linear regression equation to be y=3.736x_1+4.0174x_2+8.7769x_3+0.00032494x_4+4
with x_1, x_2, x_3, x_4 being Carbs, Protein, TotalFat, Sodium, respectively. Thus, inputting the values from the product, the predicted value for Calories would be 1944 calories.

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_calories<-data.frame(Carbohydrate=425, Protein=0.1, TotalFat=40, Sodium=430);
predict_calories_new<-data.frame(Carbohydrate=425, Protein=0.1, TotalFat=40, Sodium=43430);
increase<-predict(calories_dep,predict_calories_new)-predict(calories_dep,predict_calories)
percent_increase<-(increase/predict(calories_dep,predict_calories))*100
increase
```

```
##        1
## 13.97226
```

```
percent_increase
```

```
##         1
## 0.7188671
```

Calories will increase by 0.72%. A small increase since the Sodium variable does not have a high correlation with Calories variable # 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 ?

The null hypothesis states that illustrations do not improve how students retell the article. In other words, Article 2 has lower scores than Article 1. The alternative hypothesis states that illustrations do improve how students retell the article. In other words, Article 2 scores are greater than Article 1 scores. **(10 p)**

Paired or Independent design?

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

Based on the data and my answer, I would not pair the data. There are two differnt articles being tested, and thus, the Mann-Whitney rank sum test should be used.

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

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

```
one<-c(0.4,0.72,0.00,0.36,0.55);
two<-c(0.77,0.49,0.66,0.28,0.38);
wilcox.test(one,two, paired=TRUE)
```

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

Based on the test, our p-value is 0.8125. We cannot reject the null hyopthesis and state that illustrations do not improve scores overall.

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:

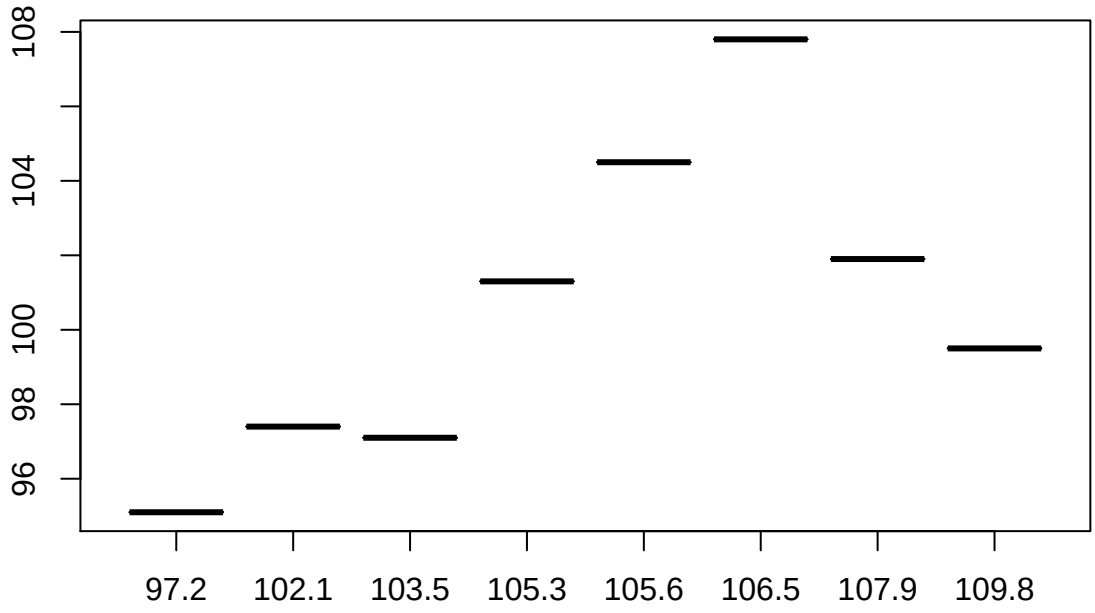
Company A: 97.1 101.3 107.8 101.9 97.4 104.5 99.5 95.1

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)

```
comp_a<-c(97.1, 101.3, 107.8, 101.9, 97.4, 104.5, 99.5, 95.1);
comp_b<-c(103.5, 105.3, 106.5, 107.9, 102.1, 105.6, 109.8, 97.2);
companies<-data.frame(comp_a,comp_b)
boxplot(comp_a~comp_b, data=companies)
```



From the above boxplot, the distribution is not normal. It is left-skewed.

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.

```
wilcox.test(comp_a, comp_b, conf.int = TRUE, conf.level = 0.95)
```

```
##
##  Wilcoxon rank sum test
##
## data:  comp_a and comp_b
## W = 13, p-value = 0.04988
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -8.5 -0.1
## sample estimates:
## difference in location
##                  -4.65
```

```r
wilcox.test(comp_a, comp_b, conf.int = TRUE, conf.level = 0.99)
```

```
##
##  Wilcoxon rank sum test
##
## data:  comp_a and comp_b
## W = 13, p-value = 0.04988
## alternative hypothesis: true location shift is not equal to 0
## 99 percent confidence interval:
##  -10.5   2.4
## sample estimates:
## difference in location
##                  -4.65
```

The 95% confidence interval associated with this dataset is between -8.5 and -0.1. Thus, we cannot be 95% confident. However, we can be at least 99% confident with an interval between -10.5 and 2.4.

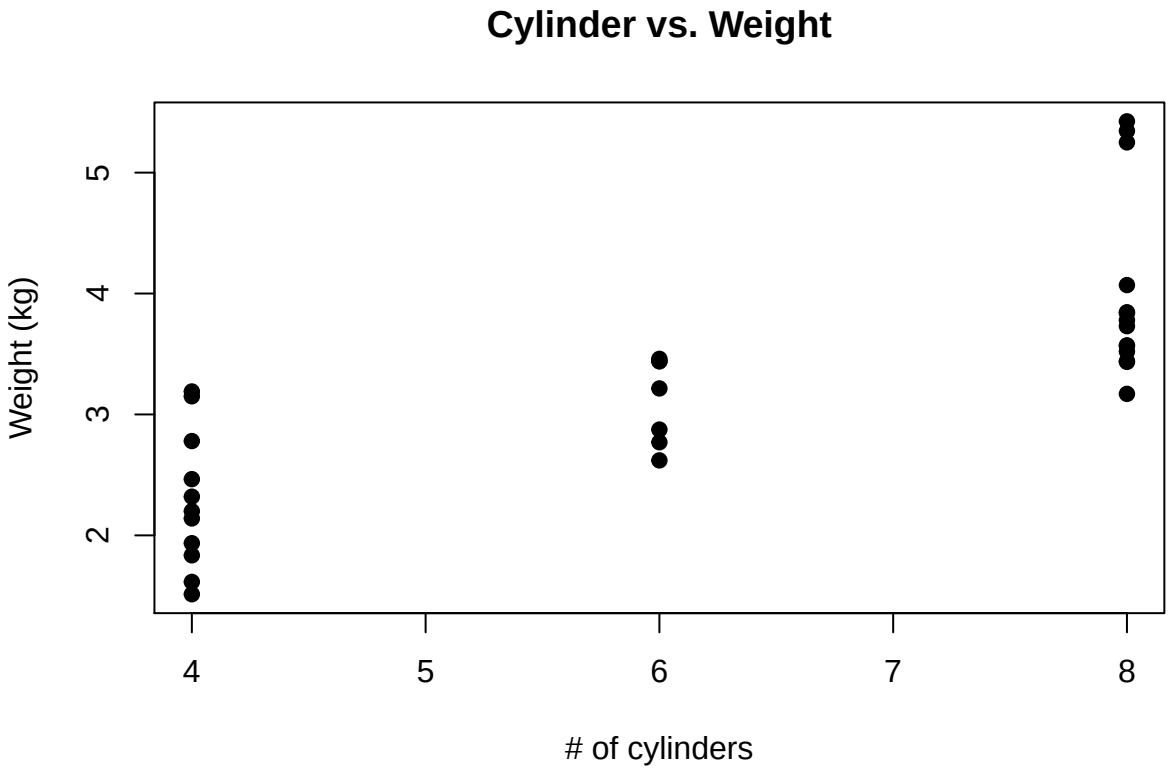
Correlation

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)**

```r
cor(mtcars$cyl, mtcars$wt)
```

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

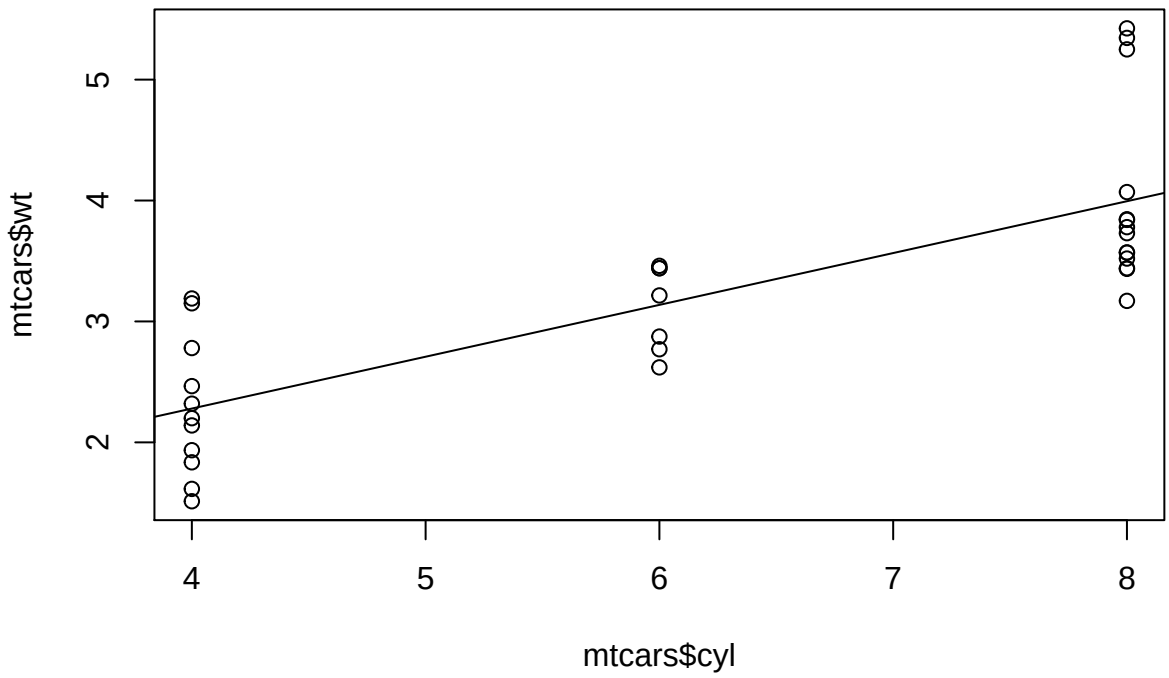
```r
plot(mtcars$cyl, mtcars$wt, main="Cylinder vs. Weight",
xlab="# of cylinders", ylab="Weight (kg)", pch=19)
```




```
asd<-lm(mtcars$wt~mtcars$cyl)
summary(asd)
```

```
##
## Call:
## lm(formula = mtcars$wt ~ mtcars$cyl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8243 -0.4293 -0.1518  0.3031  1.4297
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56462    0.40062   1.409    0.169
## mtcars$cyl   0.42871    0.06228   6.883 1.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6193 on 30 degrees of freedom
## Multiple R-squared:  0.6123, Adjusted R-squared:  0.5994
## F-statistic: 47.38 on 1 and 30 DF,  p-value: 1.218e-07
```

```
plot(mtcars$wt~mtcars$cyl)
abline(asd)
```



Based on these graphics and information, the correlation between these variables is strong positive. It is obvious from the plot that the weight of the car increases with the number of cylinders.

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

Weight is classified as a ratio variable and cylinder can be classified as ordinal.

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)**

```
cor.test(mtcars$cyl, mtcars$wt, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$cyl and mtcars$wt
## t = 6.8833, df = 30, p-value = 1.218e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5965795 0.8887052
## sample estimates:
##       cor
## 0.7824958
```

```r
cor.test(mtcars$cyl, mtcars$wt, method="spearman")
```

```
## Warning in cor.test.default(mtcars$cyl, mtcars$wt, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  mtcars$cyl and mtcars$wt
## S = 776.24, p-value = 3.574e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8577282
```

The correlation assumes a monotonic relationship, as it is obvious that the increase of cylinders increases the weight. Based on the plots, the relationship is not linear and the spearman correlation produced a stronger result. Thus, we use the Spearman correlation in this case.