Code ▾

Homework 4

Izzi Grasso Due: 4/13/2020

Problem 1

Create df

Hide

```
df <- data.frame(
  Method = c(rep(x = "MethodA", times = 6),
             rep(x = "MethodB", times = 7),
             rep(x = "MethodC", times = 5)),
  Value = c(77, 54, 67, 74, 71, 66,
            60, 41, 59, 65, 62, 64, 52,
            49, 52, 69, 47, 56)
)
```

Null hypothesis: Distributions are identical. Alt: Not null.

α = 0.05.

Kruskal Wallis

Hide

```
kruskal.test(Value ~ Method, data = df)
```

```

	Kruskal-Wallis rank sum test

data:  Value by Method
Kruskal-Wallis chi-squared = 6.6731, df = 2, p-value = 0.03556
```

Since p < 0.05, we reject the null hypothesis at α = 0.05 and conclude there is a difference in efficacy of corrosion reduction methods.

Post-hoc

Hide

```
pairwise.wilcox.test(df$Value, df$Method,
                 p.adjust.method = "bonferroni")
```

```
cannot compute exact p-value with ties
```

```

    Pairwise comparisons using Wilcoxon rank sum test 

data:  df$Value and df$Method 

        MethodA MethodB
MethodB 0.066   -      
MethodC 0.156   1.000  

P value adjustment method: bonferroni 
```

We can conclude Method A and Method B are significantly different.

Problem 2

Data

Hide

```
  Polluted <- c(21.3, 18.7, 23.0, 17.1, 16.8, 20.9, 19.7)
  Unpolluted  <-  c(14.2, 18.3, 17.2, 18.4, 20.0)
```

Test for homogeneity

Null hypothesis: $\sigma_1^2 = \sigma_2^2$ Alternative hyp: $\sigma_1^2 \neq \sigma_2^2$

Hide

```
var.test(Polluted, Unpolluted)
```

```

    F test to compare two variances

data:  Polluted and Unpolluted
F = 1.112, num df = 6, denom df = 4, p-value = 0.9618
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.120901 6.924380
sample estimates:
ratio of variances 
          1.111964 
```

At $\alpha = 0.05$, we fail to reject null hypothesis.

Check for normality (Shapiro)

Null Hypothesis: normally distributed Alternative Hyp: Not normally distributed

Hide

```
shapiro.test(Polluted)
```

```


    Shapiro-Wilk normality test

data:  Polluted
W = 0.95574, p-value = 0.7815
```

Hide

```
shapiro.test(Unpolluted)
```

```


    Shapiro-Wilk normality test

data:  Unpolluted
W = 0.92311, p-value = 0.5502
```

Fail to reject null hypothesis. Assumption of normality holds.

Pooled t-test

Null Hypothesis: $\mu_1 = \mu_2$ Althernative Hyp: $\mu_1 > \mu_2$

Hide

```
t.test(Polluted, Unpolluted, alternative = "greater", var.equal = TRUE)
```

```


    Two Sample t-test

data:  Polluted and Unpolluted
t = 1.5505, df = 10, p-value = 0.07603
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.3417755        Inf
sample estimates:
mean of x mean of y 
 19.64286  17.62000 
```

We fail to reject the null hypothesis at $\alpha = 0.05$. There is not significant evidence that the true fluoride concentration for livestock grazing is higher in the polluted region than the unpolluted region.

Problem 3

Data

Hide

```
Day1 <- c(5.0, 4.8, 5.1, 5.1, 4.8, 5.1, 4.8, 4.8, 5.0, 5.2, 4.9, 4.9, 5.0)
Day2 <- c(5.8, 4.7, 4.7, 4.9, 5.1, 4.9, 5.4, 5.3, 5.3, 4.8, 5.7, 5.1, 5.7)
```

Test for homogeneity

Null Hypothesis: $\sigma_1^2 = \sigma_2^2$ Alternative Hyp: $\sigma_1^2 < \sigma_2^2$

Hide

```
var.test(Day1, Day2, alternative = "less", conf.level = .99)
```

```

    F test to compare two variances

data:  Day1 and Day2
F = 0.12987, num df = 12, denom df = 12,
p-value = 0.0006359
alternative hypothesis: true ratio of variances is less than 1
99 percent confidence interval:
 0.0000000 0.5396439
sample estimates:
ratio of variances 
         0.1298701 
```

At $\alpha = 0.01$, we reject the null hypothesis. There is significant evidence that the variability of the process is greater on the second day than on the first.

Problem 4

Data

Hide

```
Men <- c(5, 10, 2, 0, 6, 4, 5, 15)
Women <- c(8, 9, 3, 5, 0, 4, 15)
```

Test for normality (shapiro)

Null Hypothesis: normally distributed Alternative Hyp: Not normally distributed

Hide

```
shapiro.test(Men)
```

```


    Shapiro-Wilk normality test

data:  Men
W = 0.92306, p-value = 0.4552
```

Hide

```
shapiro.test(Women)
```

```


    Shapiro-Wilk normality test

data:  Women
W = 0.95919, p-value = 0.8117
```

Fail to reject null hypothesis. Assumption of normality holds.

Test for homogeneity

Null Hypothesis: $\sigma_1^2 = \sigma_2^2$ Alternative Hyp: $\sigma_1^2 \neq \sigma_2^2$

Hide

```
var.test(Men, Women)
```

```


    F test to compare two variances

data:  Men and Women
F = 0.92555, num df = 7, denom df = 6,
p-value = 0.9082
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1625059 4.7375059
sample estimates:
ratio of variances
         0.9255478
```

Fail to reject null hypothesis. We can assume homogeneity.

Two sample t-test

Hide

```
t.test(Men, Women, alternative = "two.sided", paired = FALSE, var.equal = TRUE, conf.
level = 0.95)
```

```


    Two Sample t-test

data:  Men and Women
t = -0.16566, df = 13, p-value = 0.871
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.766671  4.945242
sample estimates:
mean of x mean of y 
 5.875000  6.285714 
```

At $\alpha = 0.05$, we fail to reject the null hypothesis. There is not significant evidence that there is a difference in sick days between men and women.

Problem 5

Data frame

Hide

```
df3 <- data.frame(
  F1 = c(58, 29, 37, 40, 44, 37, 49, 49, 38),
  F2 = c(68, 67, 69, 58, 62, 48, 62, 76, 66),
  F3 = c(96, 90, 90, 103, 100, 91, 100, 114, 94),
  F4 = c(101, 110, 90, 103, 100, 91, 100, 114, 94),
  F5 = c(124, 114, 111, 113, 114, 102, 114, 112, 103)
)
```

Test for normality

Hide

```
shapiro.test(df3$F1)
```

```

    Shapiro-Wilk normality test

data:  df3$F1
W = 0.95931, p-value = 0.7911
```

Hide

```
shapiro.test(df3$F2)
```

```

    Shapiro-Wilk normality test

data:  df3$F2
W = 0.94681, p-value = 0.6551
```

Hide

```
shapiro.test(df3$F3)
```

```

    Shapiro-Wilk normality test

data:  df3$F3
W = 0.88471, p-value = 0.1759
```

Hide

```
shapiro.test(df3$F4)
```

```

    Shapiro-Wilk normality test

data:  df3$F4
W = 0.94141, p-value = 0.5968
```

Hide

```
shapiro.test(df3$F5)
```

```

	Shapiro-Wilk normality test

data:  df3$F5
W = 0.8834, p-value = 0.1705
```

Fail to reject null hypothesis. Normality assumption holds.

ANOVA

Null hypothesis: $\mu_1 = \mu_2 = .. = \mu_5$ Alternative hyp: Not null hypothesis

Hide

```
df4 <- df3 %>% gather(Fertilizer, Value)
model <- aov(Value ~ Fertilizer, df4)
summary(model)
```

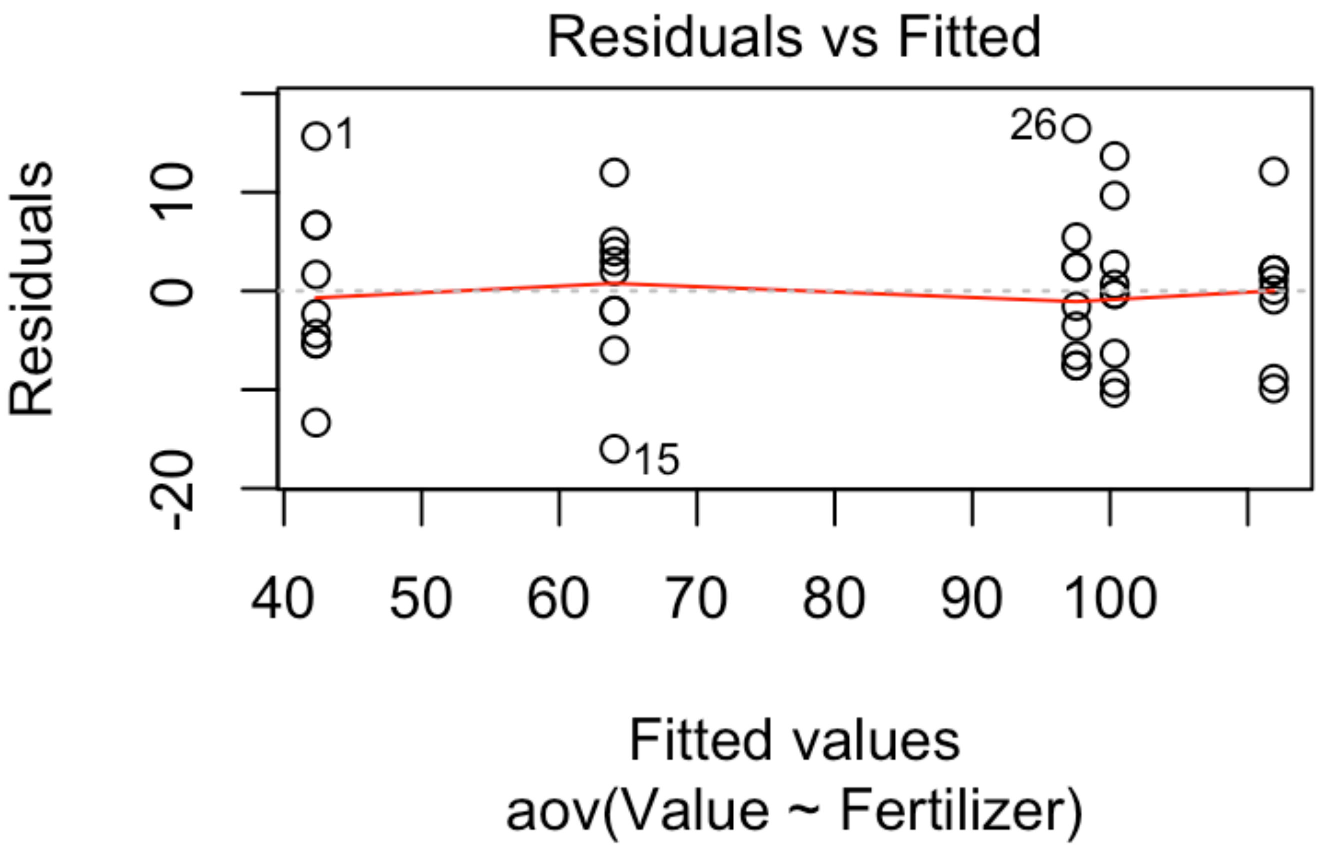
```
            Df Sum Sq Mean Sq F value Pr(>F)
Fertilizer   4  30253    7563     124 <2e-16
Residuals   40   2439      61               

Fertilizer  ***
Residuals      
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

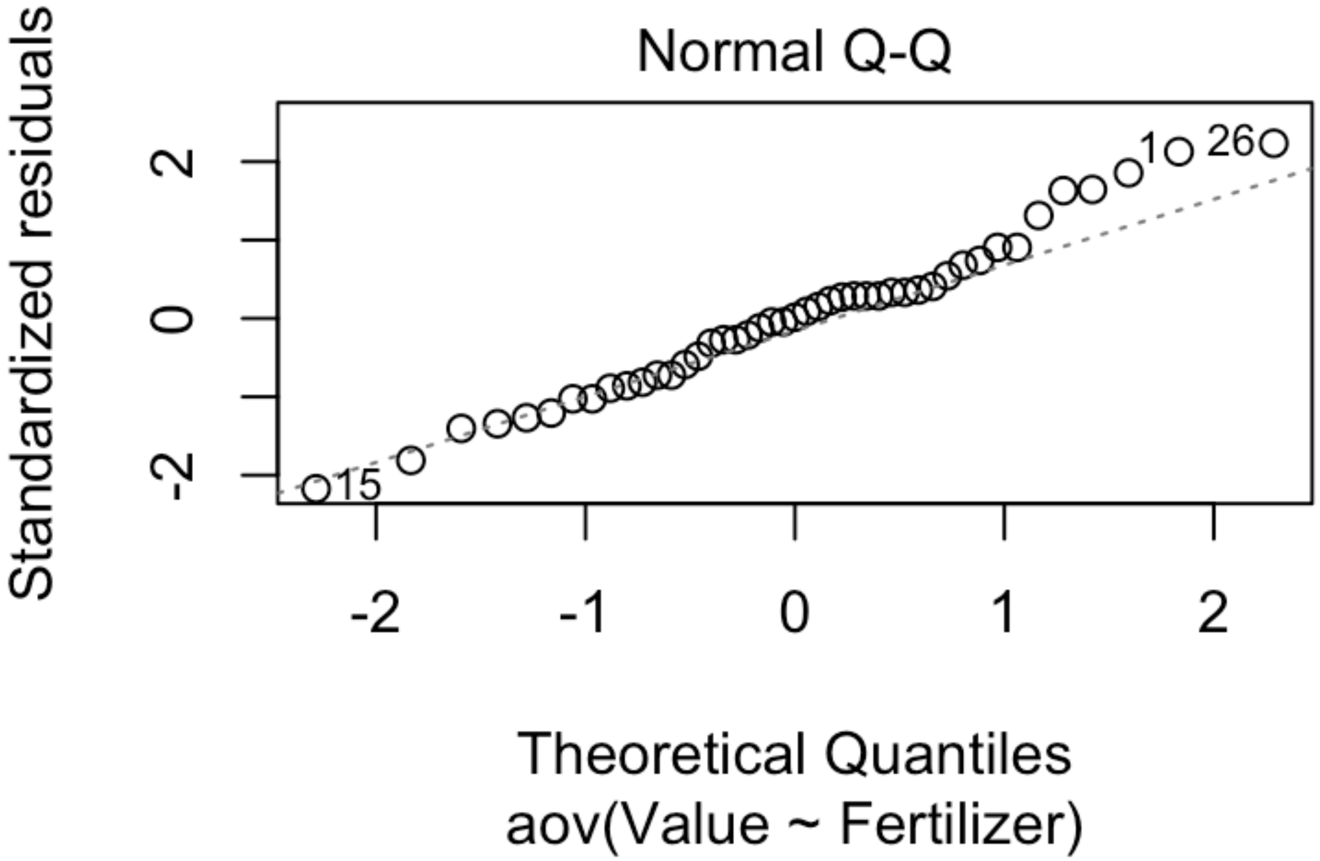
At $\alpha = 0.01$, we reject the null hypothesis. There is evidence that there is difference in yield by fertilizer.

Residual analysis

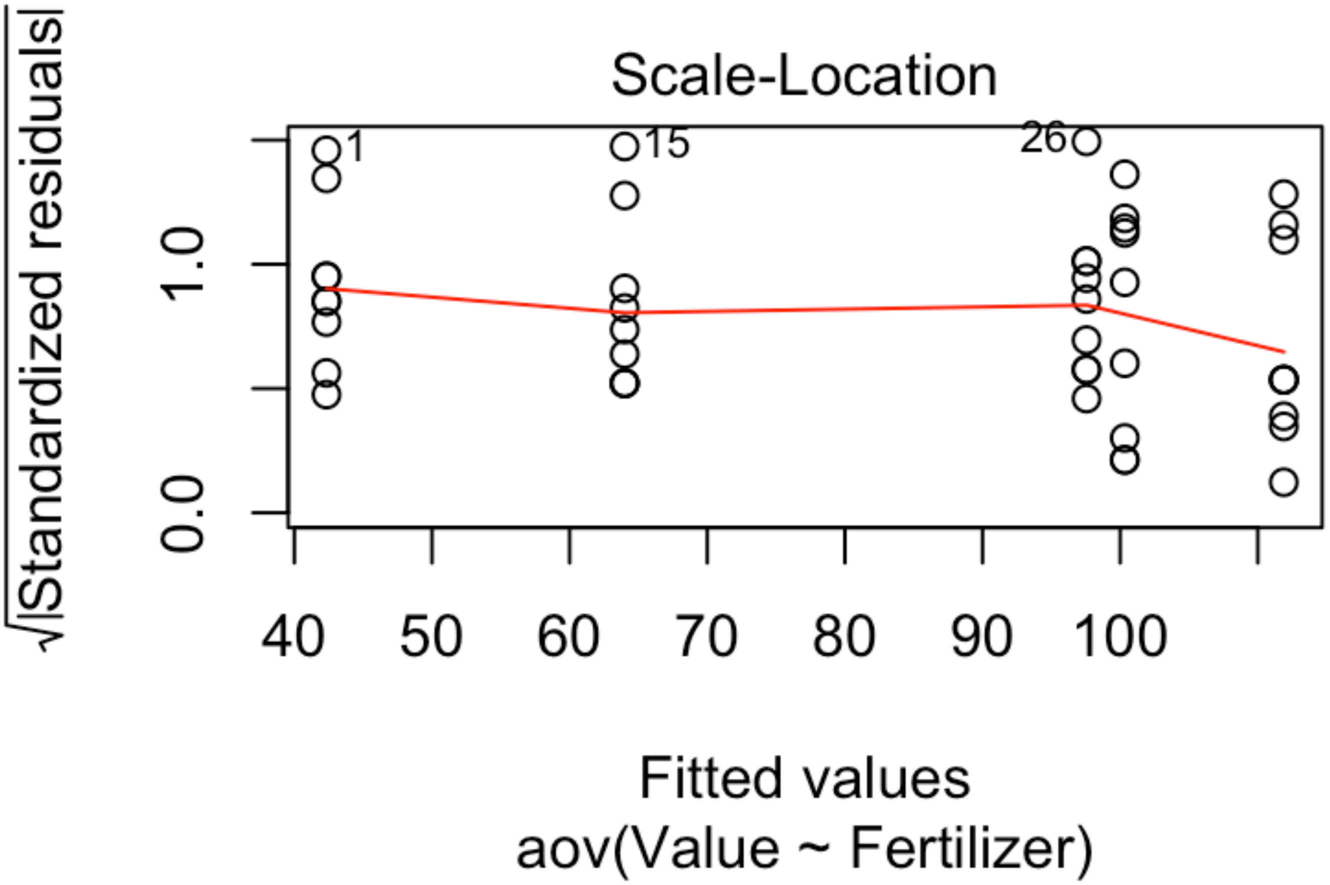
Hide

```
plot(model)
```


Residuals vs Fitted
Residuals
10
0
-20
1
26
15
40
50
60
70
80
90
100
Fitted values
aov(Value ~ Fertilizer)



```
hat values (leverages) are all = 0.1111111
 and there are no factor predictors; no plot no. 5
```

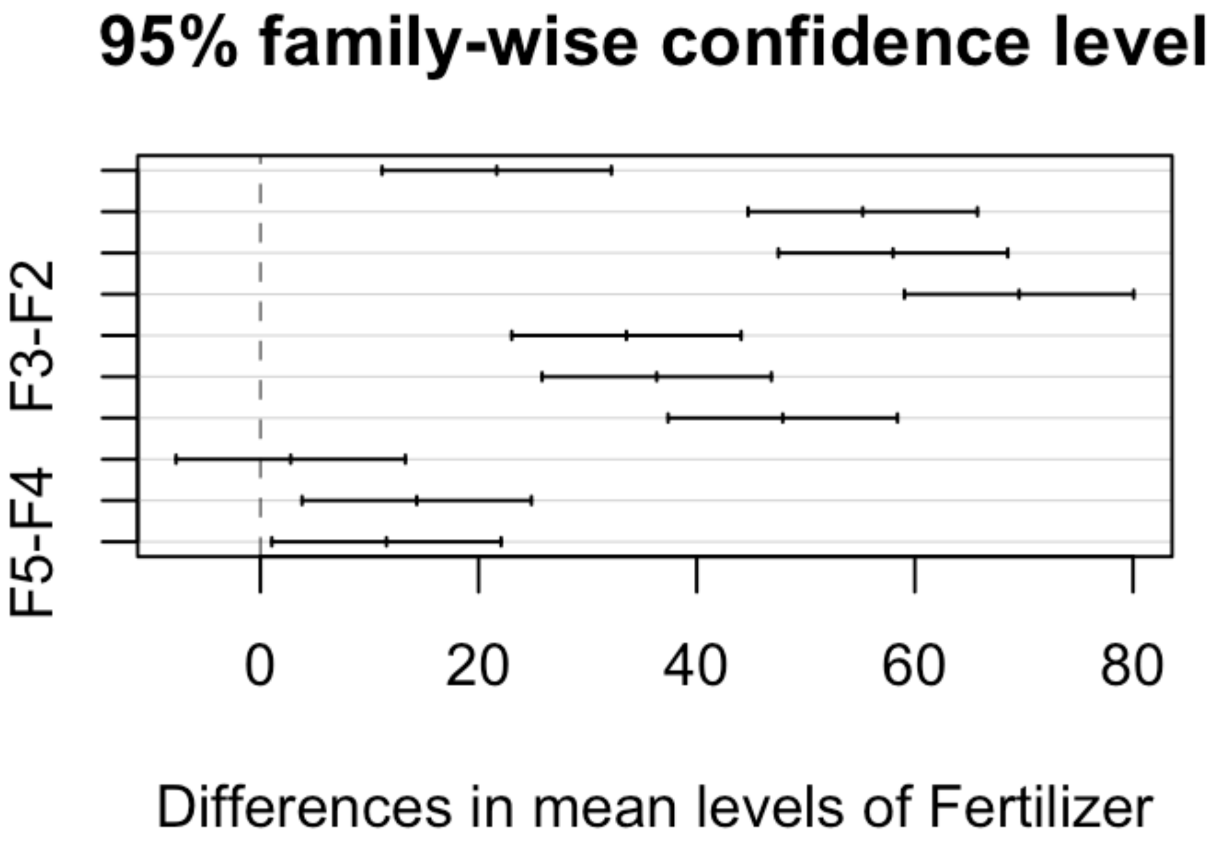

Scale-Location
√|Standardized residuals|
0.0
1.0
1
15
26
40
50
60
70
80
90
100
Fitted values
aov(Value ~ Fertilizer)

Normality assumption holds, and homogeneity assumption holds.

Tukey test

Hide

```
plot(TukeyHSD(model))
```



All yields are different from others except F4-F3. F5 produces the largest yield.

Kruskal Wallis

Null hypothesis: $\mu_1 = \mu_2 = .. = \mu_5$ Alternative hyp: Not null hypothesis

Hide

```
kruskal.test(Value ~ as.factor(Fertilizer), data = df4)
```

```

	Kruskal-Wallis rank sum test

data:  Value by as.factor(Fertilizer)
Kruskal-Wallis chi-squared = 37.545, df =
4, p-value = 1.391e-07
```

At $\alpha = 0.01$, we reject the null hypothesis. There is evidence that there is a difference of yield by fertilizer type.

Post hoc

Hide

```
pairwise.wilcox.test(df4$Value, df4$Fertilizer,
                 p.adjust.method = "bonferroni")
```

```
cannot compute exact p-value with tiescannot compute exact p-value with tiescannot co
mpute exact p-value with tiescannot compute exact p-value with tiescannot compute exa
ct p-value with tiescannot compute exact p-value with tiescannot compute exact p-valu
e with tiescannot compute exact p-value with tiescannot compute exact p-value with ti
escannot compute exact p-value with ties
```

```


    Pairwise comparisons using Wilcoxon rank sum test 

data:  df4$Value and df4$Fertilizer 

   F1     F2     F3     F4    
F2 0.0124 -      -      -     
F3 0.0040 0.0040 -      -     
F4 0.0040 0.0041 1.0000 -     
F5 0.0040 0.0040 0.0444 0.0768

P value adjustment method: bonferroni 
```

At α = 0.01, F3-F4, F4-F5, F3-F5, and F1-F2 do not have a significant difference.

Problem 6

Data

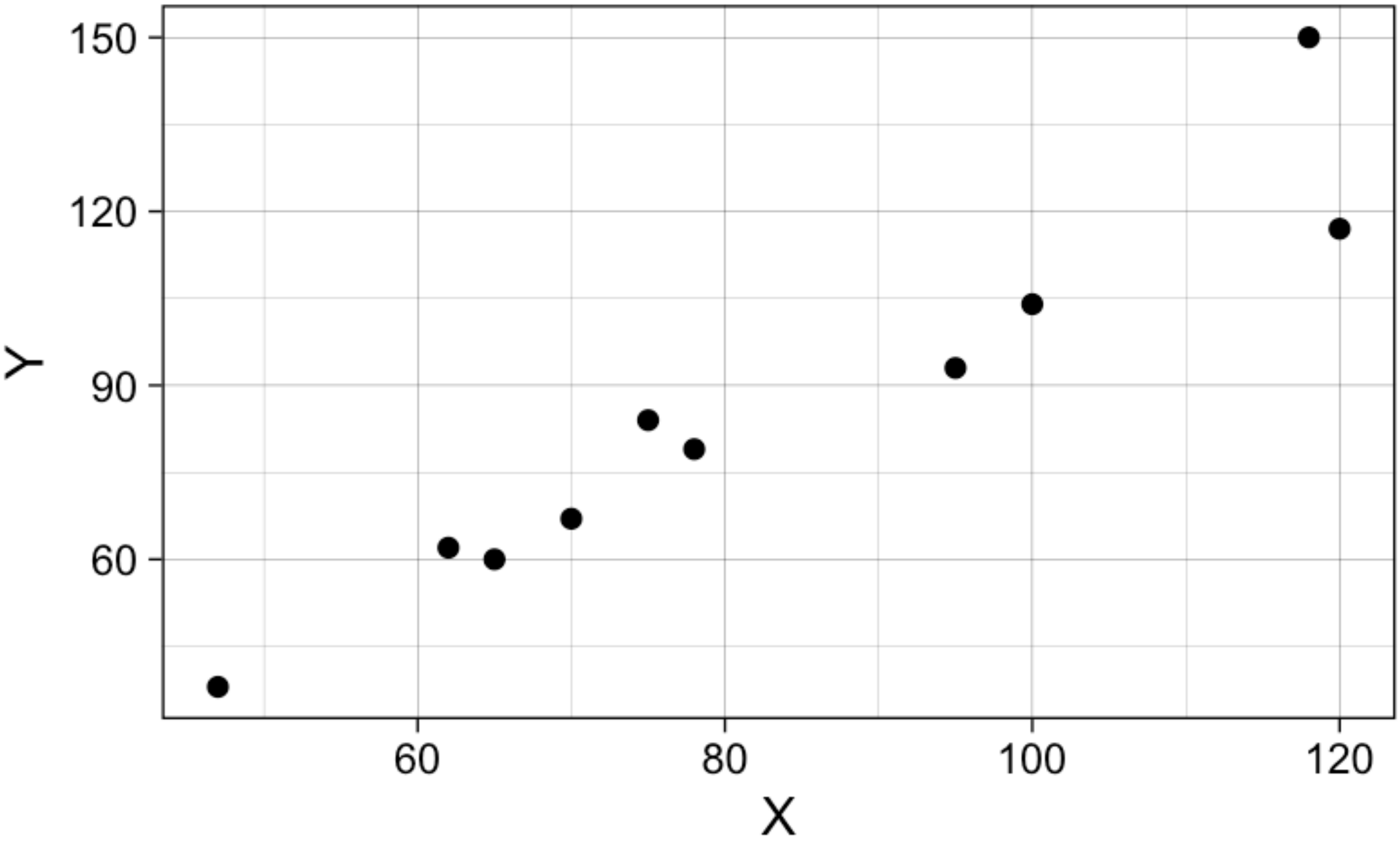
Hide

```
x <- c(47, 62, 65, 70, 75, 78, 95, 100, 120, 118)
y <- c(38, 62, 60, 67, 84, 79, 93, 104, 117, 150)
df5 <- data.frame(
  X = x,
  Y = y
)
```

Plot

Hide

```
ggplot(df5, aes(x = X, y = Y)) +
  geom_point() +
  theme_linedraw()
```



Data looks linear, correlation analysis is reasonable.

Spearman Rank

Null hypothesis: ρ = 0 Alternative hyp: $\rho \neq 0$

Hide

```
cor.test(x, y, method = "spearman")
```

```


    Spearman's rank correlation rho

data:  x and y
S = 6, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho 
0.9636364 
```

At α = 0.05 we reject the null hypothesis. There is a strong positive correlation between x and y.