STAT 511 Hw7

Amy Fox

10/20/2019

Load necessary packages

```
library(tidyverse)
library(emmeans)
library(dunn.test)
library(car)
```

Question 1

Read Problem 8.23 which concerns power plant reliability. This data is available from the Ott&Longnecker companion site. Note that the data is in "wide" format. In order to "transpose" the data to "long" format you can use the following code (assumes the original data is called InData after importing):

** Note for parts C and D: plots do NOT need to be included in your assignment, just discuss. However, p-values from Levene's test and Shapiro-Wilk test should be included in your assignment.**

```
power_plant_data <- read.csv("../Data/OTT_Final/ASCII-comma/CH08/ex8-23.txt") %>%
  rename(A = X.A.,
         B = X.B.,
         C = X.C.,
         D = X.D.,
         E = X.E.,
         F = X.F.,
         G = X.G.) %>%
  gather(key = "Plant", value = "EDG") %>%
  mutate(Plant = as.factor(Plant))  %>%
 mutate(Plant = fct_relevel(Plant, "G"))
```

A. Create side-by-side boxplots and include then in your assignment.

```
ggplot(power_plant_data, aes(x = Plant, y = EDG)) +
  geom_boxplot() +
  ggtitle("Power Plant Boxplot")
```

Power Plant Boxplot
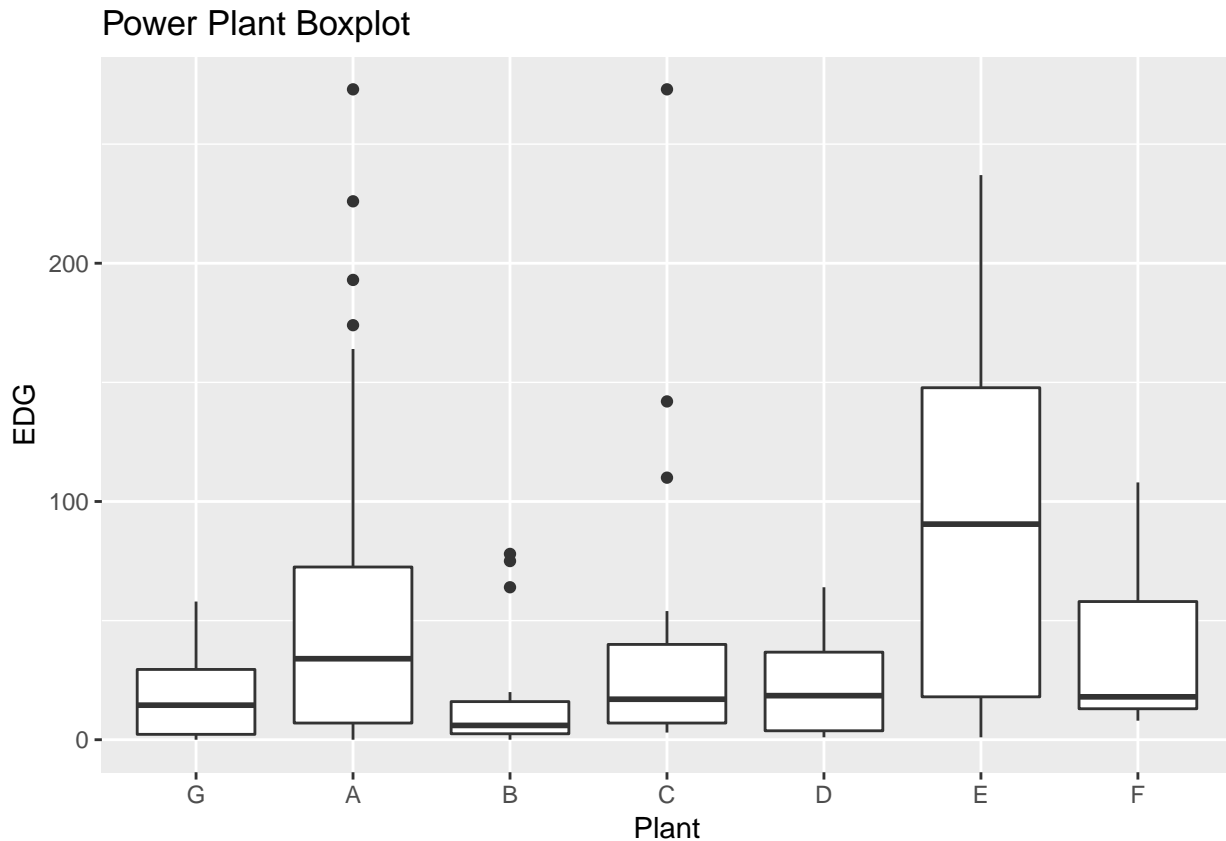
B. Run the one-way ANOVA on the original scale. Include the ANOVA table in your assignment.

```
One_way_aov_power_plant <- lm(EDG ~ Plant, data = power_plant_data)

anova(One_way_aov_power_plant)
```

```
## Analysis of Variance Table
##
## Response: EDG
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Plant      6  58745  9790.9  2.6761 0.01912 *
## Residuals 96 351233  3658.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

C. Do the assumptions for one-way ANOVA appear to be satisfied for these data? In order to justify your response, consider residual diagnostic plots, Levene's test (using default center = "median") and Shapiro-Wilk test results. (4 pts)

```
# residual diagnostic plots
par(mfrow = c(2,2))
plot(One_way_aov_power_plant)
```
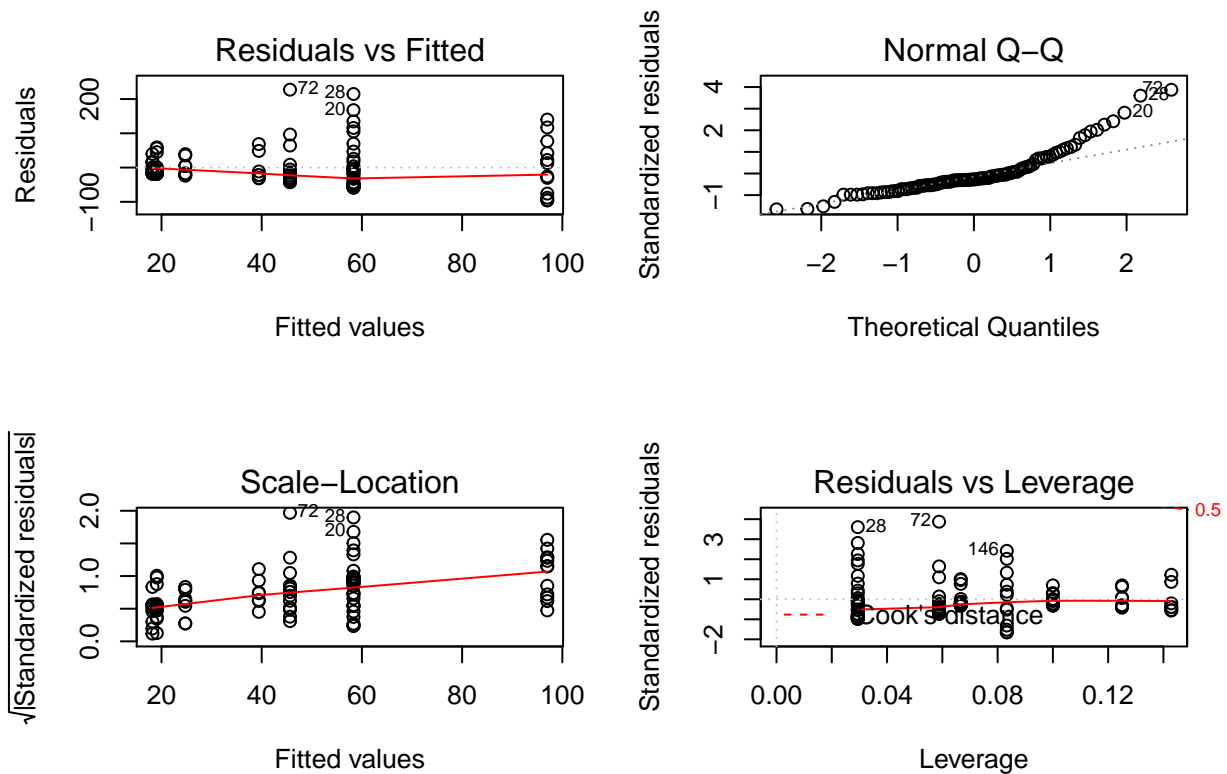
Based on the Residuals vs. Fitted, it does not appear as if there are equal variances. The Q-Q plots also show that the residuals are not normally distributed. Therefore, the assumptions for on-way ANOVA are not satisfied with this data.

```r
# levene test
leveneTest(EDG ~ Plant, data = power_plant_data, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##       Df F value  Pr(>F)
## group  6  2.3122 0.03963 *
##       96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value $= 0.040 < \alpha = 0.05$, we reject H0 and assume that the variances are not equal.

```r
shapiro.test(residuals(One_way_aov_power_plant ))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(One_way_aov_power_plant)
## W = 0.85649, p-value = 1.444e-08
```

Because the p-value $= 2.2$ x $10^{-12} < \alpha = 0.05$, we reject H0 and can assume that the data is NOT normally distributed.

In conclusions, all of these tests agree that the data is neither normally distributed, nor does it have equal variances.

D. Because the data are counts of the number of successes for the EDGs, a square root transform may be an alternative to analysis on the original scale. Run the one-way ANOVA after square root transforming EDG. Include the ANOVA table in your assignment.

3

```
power_plant_data <- power_plant_data %>%
  mutate(sqrt_EDG = sqrt(EDG))

One_way_aov_power_plant_sqrt <- lm(sqrt_EDG ~ Plant, data = power_plant_data)

anova(One_way_aov_power_plant_sqrt)
```

```
## Analysis of Variance Table
##
## Response: sqrt_EDG
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Plant      6  252.49  42.082  2.6817 0.0189 *
## Residuals 96 1506.46  15.692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

E. Do the assumptions for one-way ANOVA appear to be satisfied for after square root transform? In order to justify your response, consider residual diagnostic plots, Levene's test (using default center = "median") and Shapiro-Wilk test results. (4 pts)

```
# residual diagnostic plots
par(mfrow = c(2,2))
plot(One_way_aov_power_plant_sqrt)
```
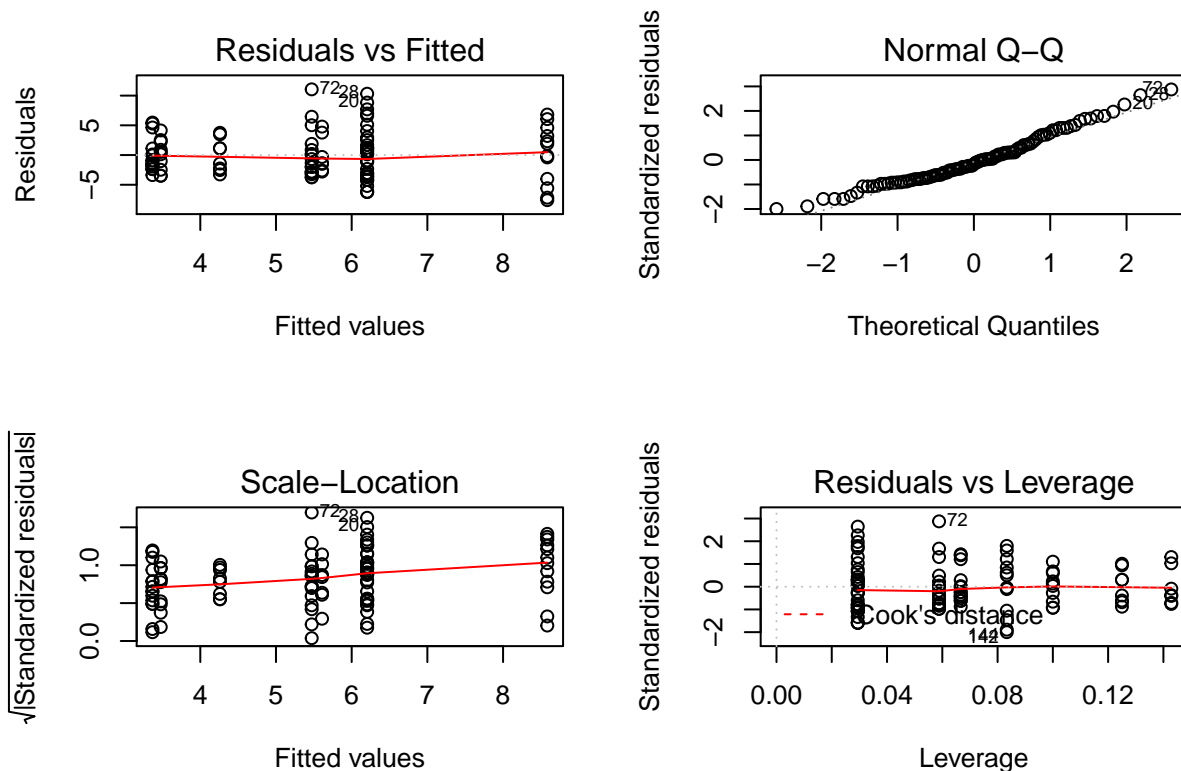

After the square-root tranfrom, based on the Residuals vs. Fitted, it appear as if there are equal variances. The Q-Q plots also show that the residuals are now normally distributed.

```
# levene test
leveneTest(sqrt_EDG ~ Plant, data = power_plant_data, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##       Df F value Pr(>F)
```

```
## group   6  1.6464 0.1428
##        96
```

Because the p-value $= 0.143 > \alpha = 0.05$, we fail to reject H0 and assume that the variances are equal.

shapiro.test(power_plant_data**$**sqrt_EDG)

```
##
##  Shapiro-Wilk normality test
##
## data:  power_plant_data$sqrt_EDG
## W = 0.91961, p-value = 1.02e-05
```

Because the p-value $= 1.02$ x $10^{-5} < \alpha = 0.05$, we reject H0 and can assume that the data is NOT normally distributed.

In conclusions, although the residual plots show that the data looks normal and has equal variances after the square root transformation, the Shapiro test shows that the data is not normally distributed. (Levene's agrees with the residual plots - the variances are equal)

F. Now run the Kruskal-Wallis test on the original scale (not square root transformed). Include the p-value and conclusion in your assignment.

kruskal.test(power_plant_data**$**EDG, power_plant_data**$**Plant)

```
##
##  Kruskal-Wallis rank sum test
##
## data:  power_plant_data$EDG and power_plant_data$Plant
## Kruskal-Wallis chi-squared = 12.537, df = 6, p-value = 0.051
```

The p-value $= 0.051 > 0.05 = \alpha$, therefore, we fail to reject H0. Therefore, we assume that the groups have identifcal distributions.

G. Use Dunn's test to run pairwise comparisons after Kruskal-Wallis. Include the pairwise comparisons in your assignment. Which plants show evidence of a difference from plant G at the alpha $= 0.05$ level? (4pts)

dunn.test(x = power_plant_data**$**EDG, g = power_plant_data**$**Plant, method = "none")

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 12.5372, df = 6, p-value = 0.05
##
##
##                            Comparison of x by group
##                                 (No adjustment)
## Col Mean-|
## Row Mean |          A          B          C          D          E          F
## ---------+-------------------------------------------------------------------
##        B |   2.286906
##          |     0.0111*
##          |
##        C |   0.455880  -1.618775
##          |     0.3242     0.0527
##          |
##        D |   0.994053  -0.726924   0.595221
##          |     0.1601     0.2336     0.2758
```

5

```
##         |
##      E |  -1.385870  -3.031778  -1.593366  -1.875307
##         |     0.0829     0.0012*     0.0555     0.0304
##         |
##      F |  -0.118471  -1.656043  -0.411029  -0.849750   0.875049
##         |     0.4528     0.0489     0.3405     0.1977     0.1908
##         |
##      G |   1.596615  -0.329450   1.101424   0.387375   2.428228   1.265278
##         |     0.0552     0.3709     0.1354     0.3492     0.0076*     0.1029
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

E is the only power plant that shows a difference in average EDG output because the p-value for E - G $<$ $0.05 = \alpha$.

Question 2

Read Problem 9.13 which concerns a weight loss study.

```
# tidy data
# reorder so S (standard is first for Dunnett's comparison)

weight_loss_data <- read.csv("../Data/OTT_Final/ASCII-comma/CH09/ex9-13.txt") %>%
  rename(A1 = X.A1.,
         A2 = X.A2.,
         A3 = X.A3.,
         A4 = X.A4.,
         S = X.S.) %>%
  gather(key = "Trt", value = "Loss") %>%
  mutate(Trt = as.factor(Trt) )%>%
 mutate(Trt = fct_relevel(Trt, "S"))
```

A. Calculate a table of summary statistics including sample size, mean, sd and se by Trt group. (4 pts)

```
Sum_stats_weight_loss <- weight_loss_data %>%
  group_by(Trt) %>%
  summarise(n  = n(),
            mean = mean(Loss),
            sd = sd(Loss),
            se = sd/sqrt(n)) %>%
  print()
```

```
## # A tibble: 5 x 5
##   Trt       n  mean    sd    se
##   <fct> <int> <dbl> <dbl> <dbl>
## 1 S        10  9.27  1.16  0.366
## 2 A1       10 12.0   0.829 0.262
## 3 A2       10 11.0   1.12  0.355
## 4 A3       10 10.3   1.03  0.325
## 5 A4       10 12.2   0.756 0.239
```

B. Run the one-way ANOVA and include the ANOVA table in your assignment. Note: I will not formally ask you to evaluate assumptions for this group of questions but based on the residual diagnostic plots, I think the data looks OK. There does seem to be an outlier for one

of the groups.

```
One_way_aov_weight_loss <- lm(Loss ~ Trt, data = weight_loss_data)

anova(One_way_aov_weight_loss)
```

```
## Analysis of Variance Table
##
## Response: Loss
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Trt        4 61.618 15.4045  15.681 4.164e-08 ***
## Residuals 45 44.207  0.9824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

C. Calculate unadjusted p-values for pairwise comparisons of means.

```
emmeans(One_way_aov_weight_loss, pairwise ~ Trt, adjust = "none")
```

```
## $emmeans
##  Trt emmean    SE df lower.CL upper.CL
##  S     9.27 0.313 45     8.64      9.9
##  A1   12.05 0.313 45    11.42     12.7
##  A2   11.02 0.313 45    10.39     11.7
##  A3   10.27 0.313 45     9.64     10.9
##  A4   12.24 0.313 45    11.61     12.9
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  S - A1      -2.78 0.443 45  -6.272  <.0001
##  S - A2      -1.75 0.443 45  -3.948  0.0003
##  S - A3      -1.00 0.443 45  -2.256  0.0290
##  S - A4      -2.97 0.443 45  -6.700  <.0001
##  A1 - A2      1.03 0.443 45   2.324  0.0247
##  A1 - A3      1.78 0.443 45   4.016  0.0002
##  A1 - A4     -0.19 0.443 45  -0.429  0.6702
##  A2 - A3      0.75 0.443 45   1.692  0.0976
##  A2 - A4     -1.22 0.443 45  -2.752  0.0085
##  A3 - A4     -1.97 0.443 45  -4.444  0.0001
```

D. Calculate Tukey adjusted p-value for pairwise comparisons of means.

```
# Tukey is done by default
Tukey_adjust <- emmeans(One_way_aov_weight_loss, pairwise ~ Trt)
Tukey_adjust
```

```
## $emmeans
##  Trt emmean    SE df lower.CL upper.CL
##  S     9.27 0.313 45     8.64      9.9
##  A1   12.05 0.313 45    11.42     12.7
##  A2   11.02 0.313 45    10.39     11.7
##  A3   10.27 0.313 45     9.64     10.9
##  A4   12.24 0.313 45    11.61     12.9
##
## Confidence level used: 0.95
```

```
## 
## $contrasts
##  contrast estimate     SE df t.ratio p.value
##  S - A1       -2.78 0.443 45 -6.272  <.0001
##  S - A2       -1.75 0.443 45 -3.948  0.0024
##  S - A3       -1.00 0.443 45 -2.256  0.1784
##  S - A4       -2.97 0.443 45 -6.700  <.0001
##  A1 - A2       1.03 0.443 45  2.324  0.1563
##  A1 - A3       1.78 0.443 45  4.016  0.0020
##  A1 - A4      -0.19 0.443 45 -0.429  0.9927
##  A2 - A3       0.75 0.443 45  1.692  0.4490
##  A2 - A4      -1.22 0.443 45 -2.752  0.0618
##  A3 - A4      -1.97 0.443 45 -4.444  0.0005
## 
## P value adjustment: tukey method for comparing a family of 5 estimates
```

E. Comparing unadjusted and Tukey adjusted results, how many comparisons yield pvalues less than 0.05? Just count the number of p-values less than 0.05.

Unadjusted: 8

Tukey Adjusted: 5

F. Calculate the Tukey HSD0.05 value (95% Tukey ME). In R, use qtukey(0.95, 5, 45) to find the exact table value.

TukeyME = q_a_(t_df) x sqrt(MSResid/n)

MSResid found in ANOVA table

```r
qtukey(0.95, 5, 45)*sqrt(0.9824/10)
```

```
## [1] 1.259503
```

The Tukey ME is 1.260

G. Create a "cld" display with Tukey adjustment.

```r
CLD(Tukey_adjust$emmeans)
```

```
##  Trt emmean     SE df lower.CL upper.CL .group
##  S     9.27 0.313 45     8.64      9.9  1
##  A3   10.27 0.313 45     9.64     10.9  12
##  A2   11.02 0.313 45    10.39     11.7   23
##  A1   12.05 0.313 45    11.42     12.7    3
##  A4   12.24 0.313 45    11.61     12.9    3
## 
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 5 estimates
## significance level used: alpha = 0.05
```

H. Calculate Dunnett adjusted p-values to compare each of the "A" treatments versus "S" (standard). Summarize your conclusions from the Dunnett adjusted pairwise comparisons. Which Trts show evidence of differences as compared to the standard at the alpha = 0.05 level? (4pts)

```r
emmeans(One_way_aov_weight_loss, dunnett ~ Trt)
```

```
## $emmeans
##  Trt emmean     SE df lower.CL upper.CL
##  S     9.27 0.313 45     8.64      9.9
```

```
## A1   12.05 0.313 45    11.42    12.7
## A2   11.02 0.313 45    10.39    11.7
## A3   10.27 0.313 45     9.64    10.9
## A4   12.24 0.313 45    11.61    12.9
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  A1 - S       2.78 0.443 45 6.272   <.0001
##  A2 - S       1.75 0.443 45 3.948   0.0010
##  A3 - S       1.00 0.443 45 2.256   0.0961
##  A4 - S       2.97 0.443 45 6.700   <.0001
##
## P value adjustment: dunnettx method for 4 tests
```

A1, A2, and A4 have p-values < 0.05 causing us to reject H0 and conclude that there is a true difference in means between these treatments and the standard.