Individual Project

The SoCAL International League of Triple-A minor league baseball consists of 14 teams organized into three divisions: North, South, and West. The following data: Triple-A.xlsx shows the average attendance for the 14 teams in the league. Also shown are the teams' records: W denotes the number of games won, L denotes the number of games lost, and PCT is the proportion of games played that were won.

A. Use alpha = 0.05 to test for any difference in the mean attendance for the three divisions

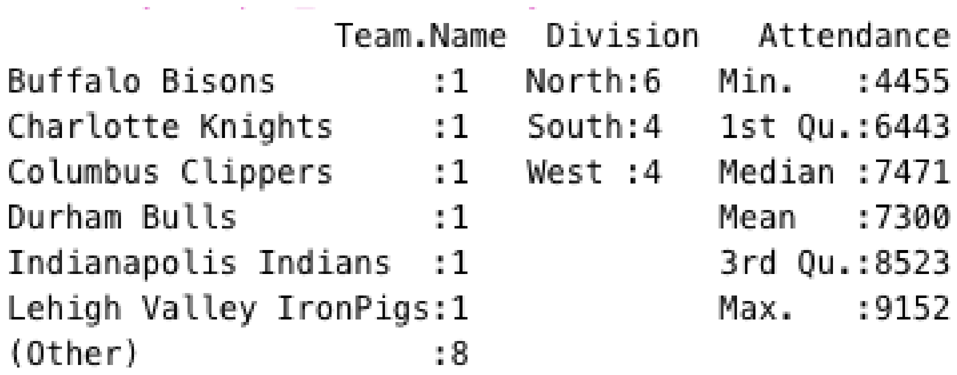
b. Use Fisher's LSD procedure to determine where the differences occur. Use alpha = 0.05

Solution:

Step 1: define hypothesis

Ho: u1 = u2 = u3 (all population means are equal)

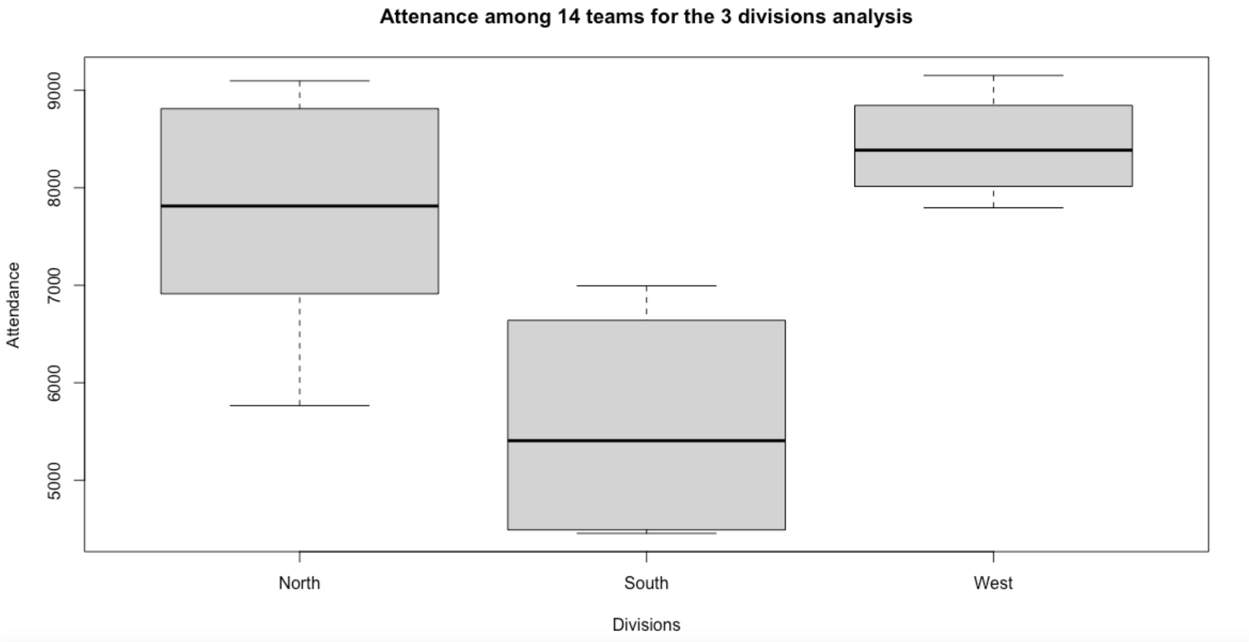
Ha: not all population means are equal



```
                  Team.Name  Division    Attendance
 Buffalo Bisons        :1   North:6   Min.   :4455
 Charlotte Knights     :1   South:4   1st Qu.:6443
 Columbus Clippers     :1   West :4   Median :7471
 Durham Bulls          :1             Mean   :7300
 Indianapolis Indians  :1             3rd Qu.:8523
 Lehigh Valley IronPigs:1             Max.   :9152
 (Other)               :8
```

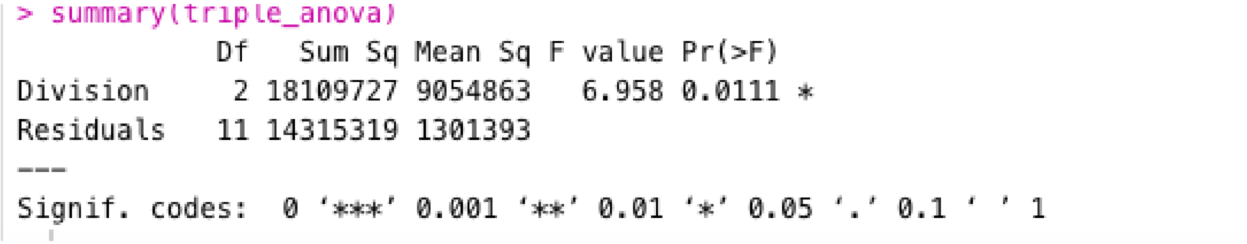
Interpretation: the categorical variable (Team Name's Division North, South, and West) has 6, 4, and 4 observations, respectively. The quantitative variable Attendance has a numeric summary with mean = 7300 and median = 7471. The average attendance for the 3 divisions is 7300.

Step 2: data visualization using boxplot



Interpretation: the boxplot gives us a visual summary of the data. It shows the median, outliers, quartiles, maximum and minimum values. It is observed from the plot that the median values of all three divisions (7800, 5400, 8400) are different. We can visually see the variation in the data from the boxplot, which shows high-within group variance and high among group-variance. Additional interpretation: we see that not all the notches in the boxplots overlap and we can conclude that with 95% confidence, that the true medians do differ.

Step 3: ANOVA test



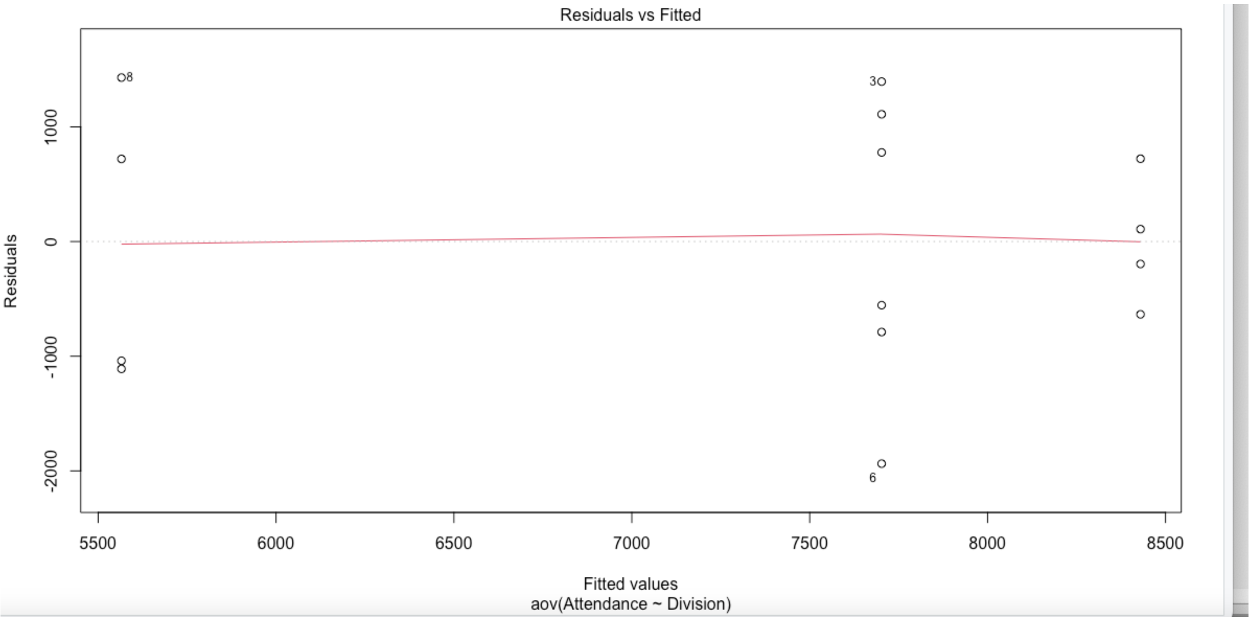
```
> summary(triple_anova)
            Df   Sum Sq Mean Sq F value Pr(>F)  
Division     2 18109727 9054863   6.958 0.0111 *
Residuals   11 14315319 1301393                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: the variance (MSE) is 1301393, the p-value is less than 0.05, so we reject the null hypothesis

Step 4: conclusion: we can conclude that there are significant differences between the mean attendance among the 3 divisions

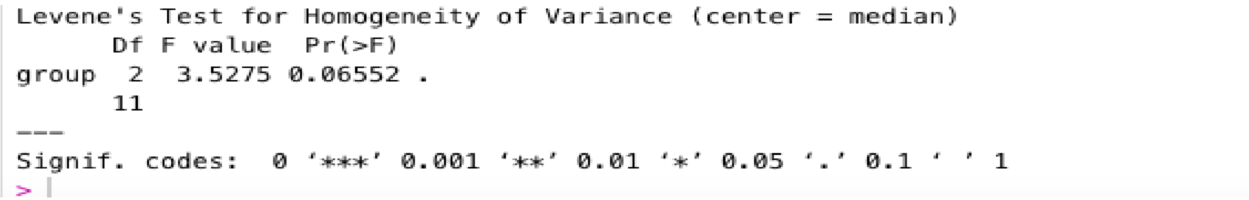
Step 5: assumption test

The first assumption is to check for the homogeneity of variance (i.e., are the pop.variance the same). Graphical analysis to check for variance using the Residuals vs Fitted plot



Interpretation: points 8, 3, and 6 are detected as outliers, which can severely affect the normality and variance assumptions. It can be useful to remove outliers from the data to meet the test assumptions.

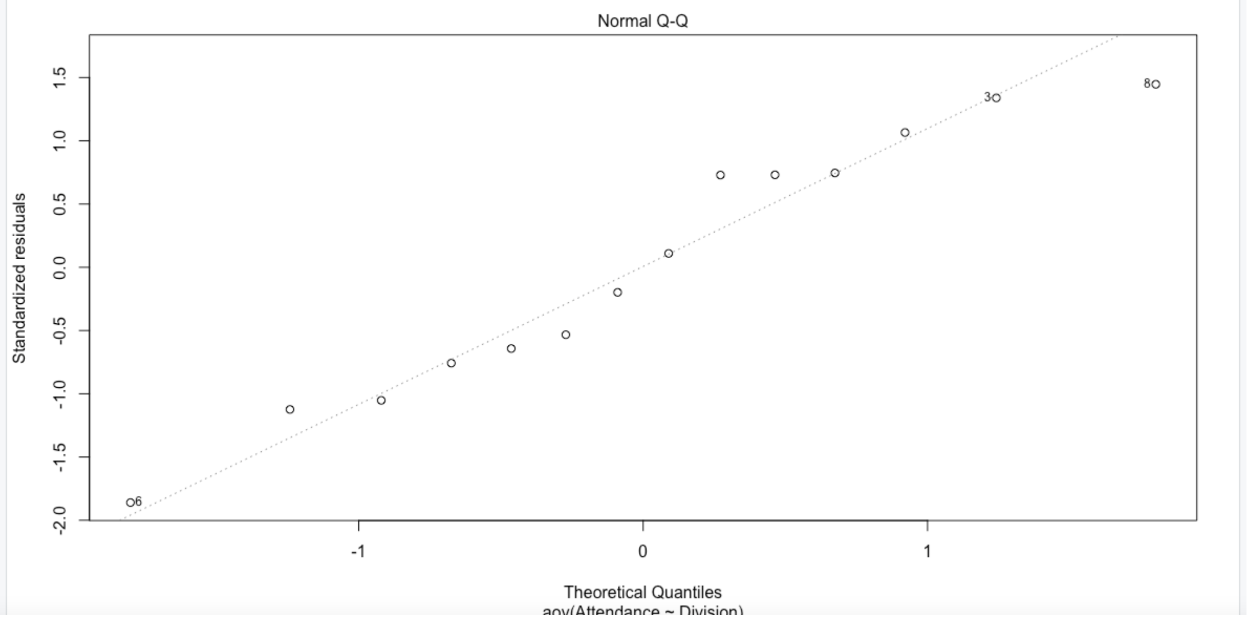
We can use a test called the Levene's test to check for variance:



```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)  
group  2  3.5275 0.06552 .
      11                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

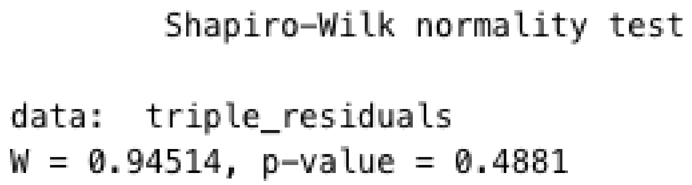
Interpretation: from the output we can see that the p-value is not less than the significant level of 0.05. This means that there is no evidence to suggest that the variance across methods is significantly different. Therefore, we can assume the homogeneity of variances in the different division groups

The second assumption is to check for normality (i.e. is our dataset normally distributed), the normal probability plot of residuals (Normal Quantile plot) is used to check for normality, the points on the plot should approximately follow a straight line:



Interpretation: the points on the plot do not really follow a straight line, so we cannot assume normality. We may have to perform another test to confirm normality

The Shapiro-Wilk test on the ANOVA residuals is used to confirm normality

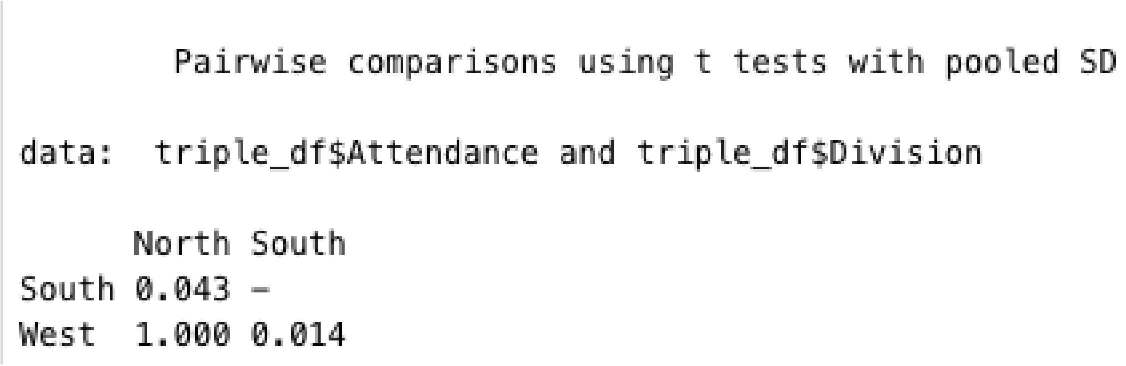


```
        Shapiro-Wilk normality test

data:  triple_residuals
W = 0.94514, p-value = 0.4881
```

Interpretation: W = 0.95, p-value = 0.5 which indicates that the normality assumption is not violated because p-value is greater than 0.05

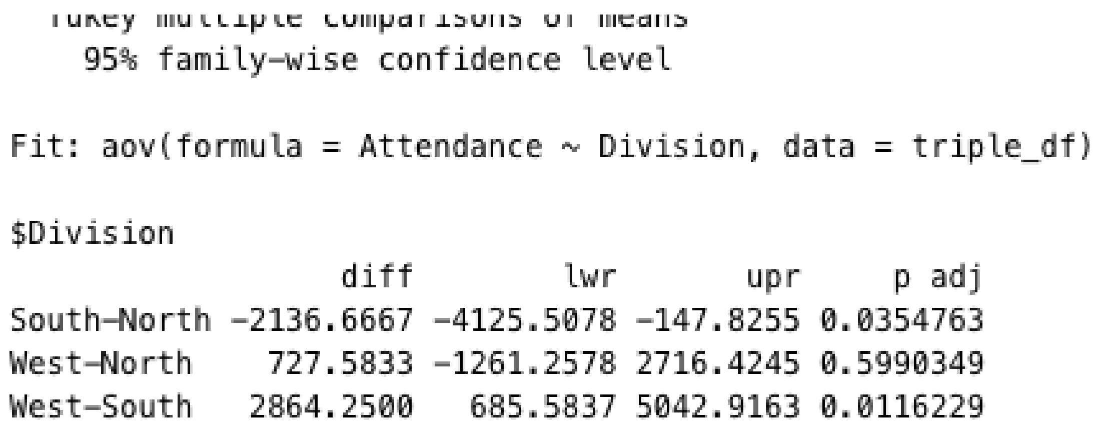
Step 6: multiple comparison procedure using t test



```
        Pairwise comparisons using t tests with pooled SD

data:  triple_df$Attendance and triple_df$Division

      North South
South 0.043 -
West  1.000 0.014
```



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Attendance ~ Division, data = triple_df)

$Division
                 diff        lwr       upr     p adj
South-North -2136.6667 -4125.5078 -147.8255 0.0354763
West-North    727.5833 -1261.2578 2716.4245 0.5990349
West-South   2864.2500   685.5837 5042.9163 0.0116229
```

Interpretation North: with an alpha value of 0.05 (p-value = 0.6 > 0.05), we FTR Ho. The population mean attendance for division North is equal to the population mean attendance for division West

Interpretation South: with an alpha value of 0.05 (p-value = 0.012 < 0.05), we can reject Ho. The population mean attendance for division South is not equal to division West

Overall interpretation: in effect, our conclusion is that the population mean attendance for division South differs from West but it is the same among North and West