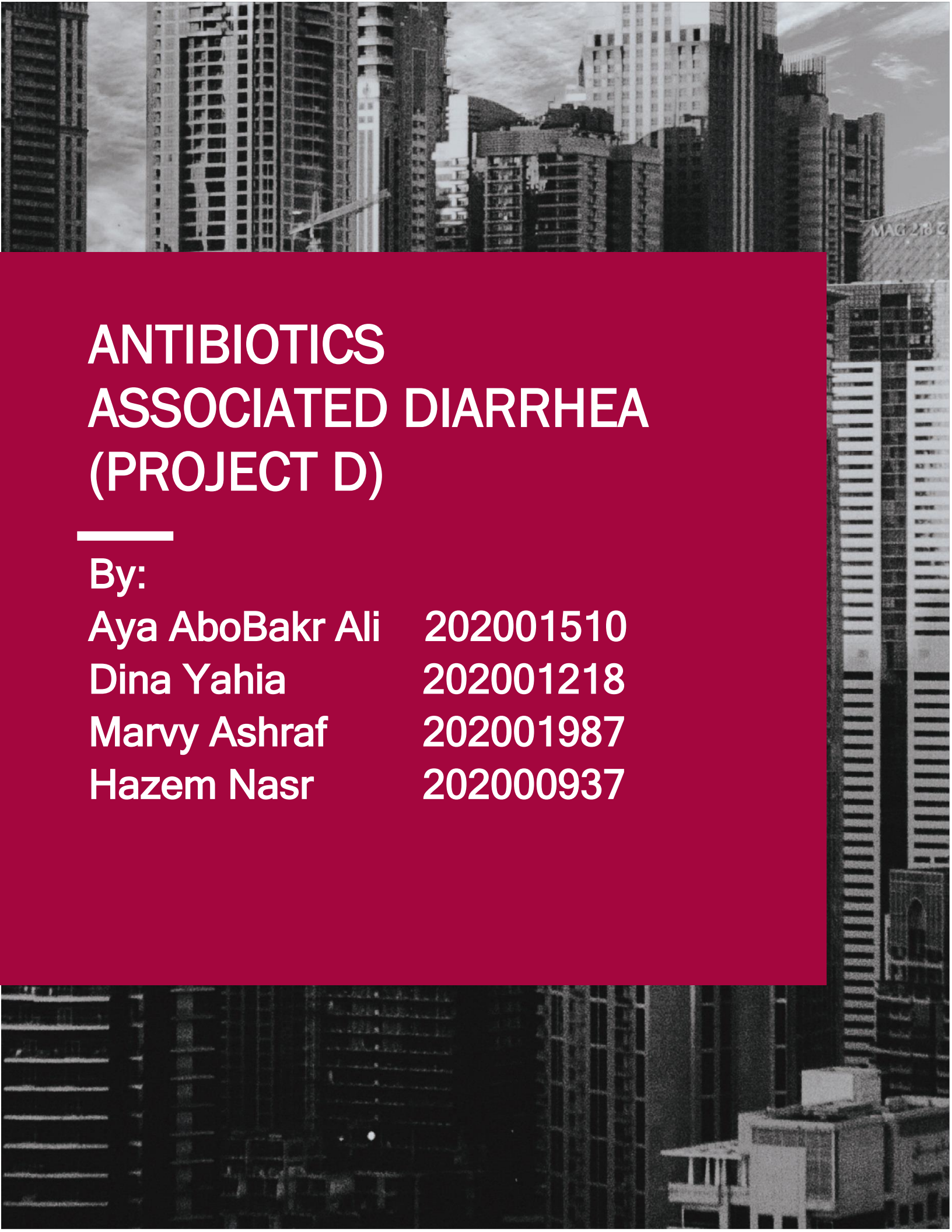


ANTIBIOTICS ASSOCIATED DIARRHEA (PROJECT D)

By:

Aya AboBakr Ali	202001510
Dina Yahia	202001218
Marvy Ashraf	202001987
Hazem Nasr	202000937

Table of Contents

0. Data reading

First of all, we read the data “AAD” and then assigned it into “mydata” variable. The data includes 8 columns, 3 character (categorical) and 5 numeric.

1. Descriptive Statistics

1.1. Summarize your data:

We summarize the data by using function summary(). Then we checked for any nulls in the data and found none.

```
> summary(mydata)
  Patient.ID        Antibiotic.class   D1.Shannon.diversity D6.Shannon.diversity D1.Chao1.diversity D6.Chao1.diversity
 Length:335         Length:335         Min.   :0.1276       Min.   :0.07041      Min.   : 25.14      Min.   : 36.15
 Class :character   Class :character   1st Qu.:2.8984       1st Qu.:2.52770      1st Qu.:138.46      1st Qu.:118.18
 Mode  :character   Mode  :character   Median :3.3923       Median :3.07407      Median :189.96      Median :169.00
                                       Mean   :3.2493       Mean   :2.86372      Mean   :200.84      Mean   :174.42
                                       3rd Qu.:3.7653       3rd Qu.:3.48406      3rd Qu.:247.91      3rd Qu.:222.03
                                       Max.   :4.4653       Max.   :4.46100      Max.   :552.93      Max.   :422.75
 D1.D6.Jaccard.distance   Outcome
 Min.   :0.2448         Length:335
 1st Qu.:0.5352         Class :character
 Median :0.6598         Mode  :character
 Mean   :0.6540
 3rd Qu.:0.7879
 Max.   :0.9485
```

1.2. Calculate the following:

(1) Mean: By adding up each value in the variable and dividing it by the total number of values, the mean is calculated. The mean() function can be used to calculate the mean.

```
> #  Mean
> mean(mydata$D1.Chao1.diversity)
[1] 200.8354
> mean(mydata$D1.D6.Jaccard.distance)
[1] 0.6540263
> mean(mydata$D1.Shannon.diversity)
[1] 3.249321
> mean(mydata$D6.Shannon.diversity)
[1] 2.863724
> mean(mydata$D6.Chao1.diversity)
[1] 174.4217
```

(2) Median: is the middle value in a sorted list of values. If there is an odd number of values, the median is the middle value. If there is an even number of values, the median is the average of the two middle values. The median() function is used to calculate it.

```
> # Median
> median(mydata$D1.Chao1.diversity)
[1] 189.9565
> median(mydata$D1.D6.Jaccard.distance)
[1] 0.659763
> median(mydata$D1.Shannon.diversity)
[1] 3.392265
> median(mydata$D6.Shannon.diversity)
[1] 3.074067
> median(mydata$D6.Chao1.diversity)
[1] 169
```

(3) Minimum: is the smallest value in the sorted list of values. The min() is used to calculate it.

```
> # Min
> min(mydata$D1.Chao1.diversity)
[1] 25.14286
> min(mydata$D1.D6.Jaccard.distance)
[1] 0.244755
> min(mydata$D1.Shannon.diversity)
[1] 0.127635
> min(mydata$D6.Shannon.diversity)
[1] 0.070407
> min(mydata$D6.Chao1.diversity)
[1] 36.15385
```

(4) Maximum: is the largest value in the sorted list of values. The max() is used to calculate it.

```
> # Max
> max(mydata$D1.Chao1.diversity)
[1] 552.9333
> max(mydata$D1.D6.Jaccard.distance)
[1] 0.948454
> max(mydata$D1.Shannon.diversity)
[1] 4.465318
> max(mydata$D6.Shannon.diversity)
[1] 4.461002
> max(mydata$D6.Chao1.diversity)
[1] 422.75
```

(5) 1st & 3rd Quantile: We calculate the 1st and 3rd quantile by using the function quantile(0.25,0.75); it will give us 2 values the first value being the 25th percentile and the second value being the 75th percentile of the data. So why do we get quantiles? So, we can understand the data's distribution better. The interquartile range (IQR), which is the difference between the third and first quartiles, is a measure of the spread of the data that is less sensitive to outliers than the range. The IQR can be used to recognize potential outliers or to determine the variability certain data sets are.

```
> # Quantiles
> quantile(mydata$D1.Chao1.diversity,prob=c(.25,.75))
     25%      75%
138.4643 247.9141
> quantile(mydata$D1.D6.Jaccard.distance,prob=c(.25,.75))
      25%       75%
0.5352305 0.7878620
> quantile(mydata$D1.Shannon.diversity,prob=c(.25,.75))
     25%      75%
2.898378 3.765255
> quantile(mydata$D6.Shannon.diversity,prob=c(.25,.75))
     25%      75%
2.527697 3.484056
> quantile(mydata$D6.Chao1.diversity,prob=c(.25,.75))
     25%      75%
118.1801 222.0333
```

1.3. Calculate a frequency table:

We use the table() function to calculate the frequency table of both “Outcome” & “Antibiotic.class”. It is helpful for figuring out how the data are distributed and for comparing the frequencies of different categories. Additionally, they can be used to calculate percentages, proportions, and other summary statistics.

```
> # Freguency Table
> table(mydata$Outcome)

AAD CDI  ND
 22   5 308
> table(mydata$Antibiotic.class)

FQN OBL PBL
 56 111 168
```

1.4. Calculate the correlation coefficient:

The correlation coefficient is a useful tool for understanding the relationship between two variables and can be used to inform decision-making in many different fields. So, we calculated the correlation coefficient between (D1 Shannon and D6 Shannon) and the result was “0.2208003” therefore it is weak positive linear relationship between two variables. And also, we calculated the correlation coefficient between (D1 Chao and D6 Chao) and the result was “0.3026013” therefore it is weak positive linear relationship between two variables.

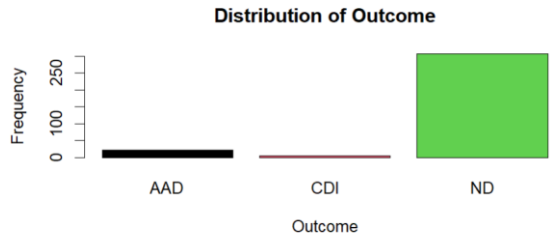
```
> # Correlation Coefficient
> cor(mydata$D1.Shannon.diversity,mydata$D6.Shannon.diversity, use="complete.obs")
[1] 0.2208003
> cor(mydata$D1.Chao1.diversity,mydata$D6.Chao1.diversity, use="complete.obs")
[1] 0.3026013
```

2. Graphics

2.1 Generate a bar chart of a categorical variable for the Outcome (AAD, CDI ,ND):

AAD, CDI, and ND are the categories in the "Outcome" section. Each unique value of "Outcome" is turned into a table of counts via the table() method. Then, using the counts from the table, the

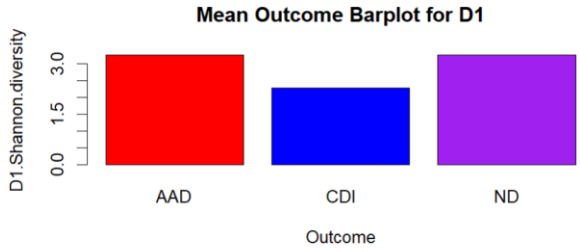
barplot() function generates a bar chart with labels for the x-axis, y-axis, and title, as well as unique colors for each category.

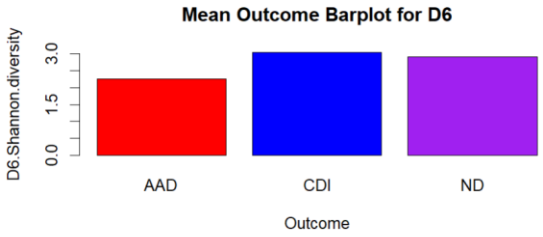


.

2.2 Generate a bar chart graph with mean Outcome in D1 and D6 Shannon:

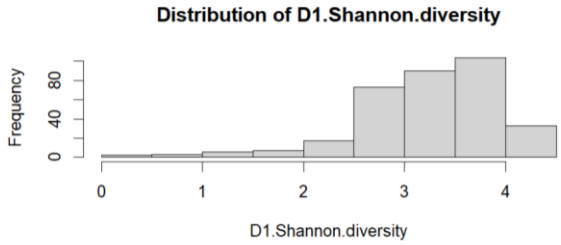
The mean values of "D1.Shannon.diversity" by "Outcome" and "D6.Shannon.diversity" by "Outcome" are represented by two different bar charts created using this code. The data are grouped by "Outcome" using the tapply() function, and the mean of "D1.Shannon.diversity" or "D6.Shannon.diversity" is determined for each group.

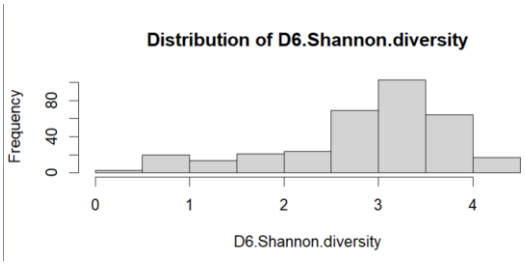




2.3 Make a histogram of a continuous variable: “D1 Shannon” as well as “D6 Shannon”:

The distribution of "D1.Shannon.diversity" and "D6.Shannon.diversity" are each represented by a distinct histogram generated by this code. Using the values of the relevant variables, the hist() function generates a histogram with labels for the x-axis, y-axis, and title.

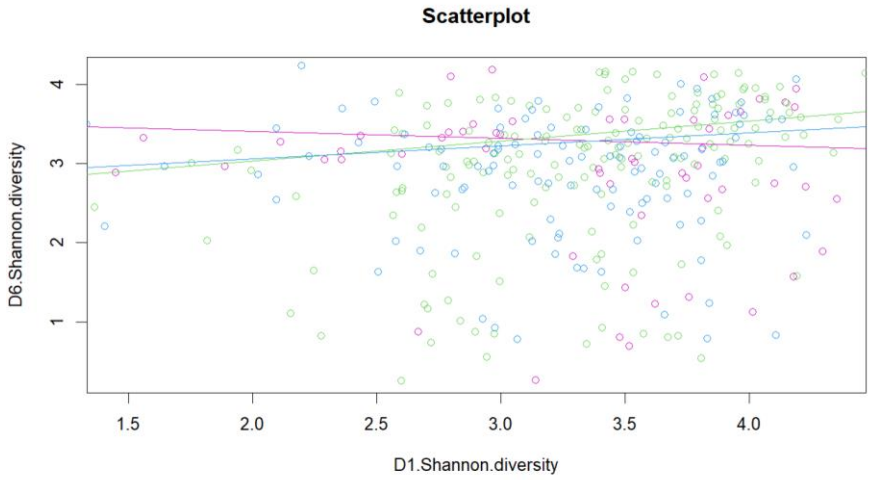




2.4 Make a scatterplot of 2 continuous variables D1 Shannon and D6 Shannon, and add the regression lines for each antibiotic:

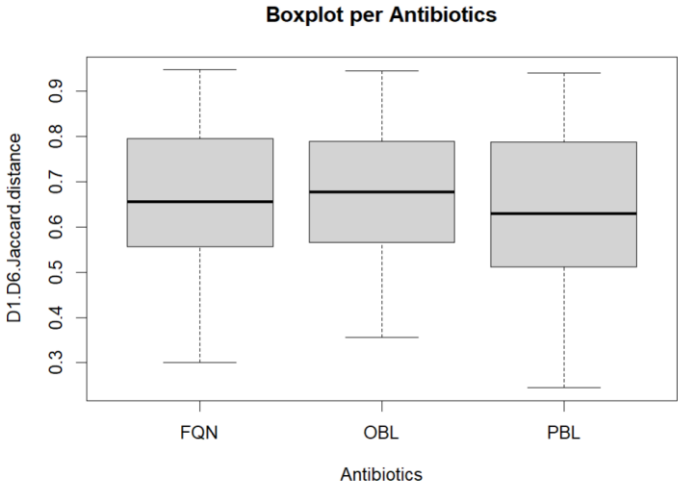
The "D1.Shannon.diversity" versus "D6.Shannon.diversity" scatterplot is generated by this code for each antibiotic class. The scatterplot for the antibiotic class "FQN" is generated by the plot() function, with the points' various colors. The scatterplot for the antibiotic classes "PBL" and "OBL" is then added using the points() method, with the points' colors varying. Only the observations for each antibiotic class are included in the data set by the [mydata$Antibiotic.class == "FQN"], [mydata$Antibiotic.class == "PBL"], and [mydata$Antibiotic.class == "OBL"].

The scatterplot for each antibiotic class includes regression lines are generated by this code. The lm() function is used to fit a linear regression model with "D1.Shannon.diversity" as the dependent variable and "D6.Shannon.diversity" as the independent variable. The abline() function then adds a regression line to the scatterplot for each antibiotic class.



2.5 Make a boxplot of Jacard distance and a separate boxplots per Antibiotics:

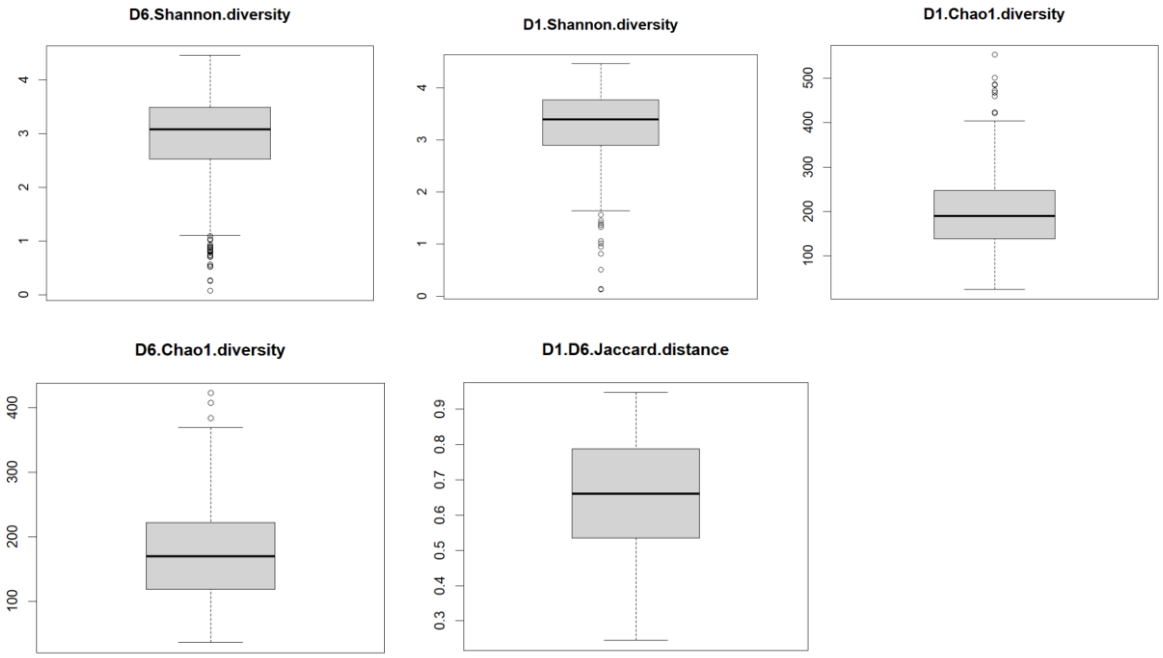
A boxplot of "D1.D6.Jaccard.distance" is generated by this code for each antibiotic class in the dataset "mydata" separately. Using the values of the relevant variables, the boxplot() function generates a boxplot with labels for the x-axis, y-axis, and title, as well as unique boxplots for each antibiotic class.



3. Outlier Detection

3.1 Detect Outliers Visually using Boxplot:

Each boxplot specifically shows how each "D6.Chao1.diversity," "D6.Shannon.diversity," "D1.Shannon.diversity," "D1.D6.Jaccard.distance," and "D1.Chao1.diversity" are distributed. The boxplot() function generates a box-and-whisker plot with a title for each variable, the variable names supplied as inputs.



3.2 Detect Outliers Statistically:

We displayed all the outliers in the data using $out.

```
> # Displaying all the outliers
> c(
+   boxplot.stats(mydata$D6.Chao1.diversity)$out,
+   boxplot.stats(mydata$D6.Shannon.diversity)$out,
+   boxplot.stats(mydata$D1.Shannon.diversity)$out,
+   boxplot.stats(mydata$D1.D6.Jaccard.distance)$out,
+   boxplot.stats(mydata$D1.Chao1.diversity)$out
+ )
 [1] 407.028571 383.659574 422.750000   0.537671   1.011686   0.805769   0.070407   0.519615   0.931239
[10]   0.878636   1.092522   0.822535   0.847067   0.927137   0.698893   0.874226   0.901139   0.738428
[19]   0.821437   0.854022   0.789490   1.039348   0.259205   0.262554   0.716793   0.777983   0.836249
[28]   0.558723   0.804012   1.010512   1.364614   0.127635   0.507176   1.063788   1.561207   0.816340
[37]   1.448330   1.331024   0.950594   1.404700   0.143722 459.441176 422.897059 500.544554 484.575342
[46] 552.933333 485.321429 420.588235 472.122449 466.520000
```

Then extracting outliers in each continuous columns and detecting their length.

```
> length(boxplot.stats(mydata$D1.Shannon.diversity)$out)
[1] 12
> length(boxplot.stats(mydata$D6.Shannon.diversity)$out)
[1] 26
> length(boxplot.stats(mydata$D1.Chao1.diversity)$out)
[1] 9
> length(boxplot.stats(mydata$D6.Chao1.diversity)$out)
[1] 3
> length(boxplot.stats(mydata$D1.D6.Jaccard.distance)$out)
[1] 0
```

4. Testing For Normality

4.1 Checking Normality using Q-Q plot, Histogram & Shapiro:

According to the 3 tests done on the numerical variables, data are found to be not normal, where according to shapiro-test, the p-value of all of them are less than 0.05 so there is a significant difference, and we have enough evidence to reject the null hypothesis (normality).

```
> shapiro.test(mydata$D1.Shannon.diversity)

        Shapiro-Wilk normality test

data:  mydata$D1.Shannon.diversity
W = 0.91168, p-value = 4.134e-13

> shapiro.test(mydata$D6.Shannon.diversity)

        Shapiro-Wilk normality test

data:  mydata$D6.Shannon.diversity
W = 0.91549, p-value = 8.795e-13

> shapiro.test(mydata$D1.Chao1.diversity)

        Shapiro-Wilk normality test

data:  mydata$D1.Chao1.diversity
W = 0.94928, p-value = 2.489e-09

> shapiro.test(mydata$D6.Chao1.diversity)

        Shapiro-Wilk normality test

data:  mydata$D6.Chao1.diversity
W = 0.96965, p-value = 1.776e-06
```

```
> shapiro.test(mydata$D1.D6.Jaccard.distance)

        Shapiro-Wilk normality test

data:  mydata$D1.D6.Jaccard.distance
W = 0.97893, p-value = 8.022e-05
```

4.2 Checking Homoscedasticity using Bartlett test & F-test:

According to the bartlett and F tests between D1 and D6 Shannon, the p-value = 2.533e-06, which means that there is a significant difference, so we have enough evidence to reject the null hypothesis, where the variance of both of them are not similar.

```
> # Bartlett test and F-test
> bartlett.test(list(mydata$D1.Shannon.diversity,mydata$D6.Shannon.diversity))

        Bartlett test of homogeneity of variances

data:  list(mydata$D1.Shannon.diversity, mydata$D6.Shannon.diversity)
Bartlett's K-squared = 22.141, df = 1, p-value = 2.533e-06

> var.test(mydata$D1.Shannon.diversity,mydata$D6.Shannon.diversity)

        F test to compare two variances

data:  mydata$D1.Shannon.diversity and mydata$D6.Shannon.diversity
F = 0.5956, num df = 334, denom df = 334, p-value = 2.533e-06
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4804468 0.7383570
sample estimates:
ratio of variances
         0.5956016
```

5. Statistical Inference

5.1 Calculate the 90%, 95%, 99% confidence interval for the means of Jacard distance each Antibiotic:

The CI interval for each percent is given by the equation: mean ± SE × Z score (or) T score. Z is the critical value from the standard normal distribution. Margin of error tells you how many percentage points your results will differ from the real population value. (Difference between uppermost limit of interval and sample mean)

1. We calculated the means of Jacard distance for each Antibiotics.

```
> meann=tapply(mydata$D1.D6.Jaccard.distance,list(antibiotics =mydata$Antibiotic.class),mean)
> meann
antibiotics
       FQN        OBL        PBL
0.6539343 0.6722537 0.6420139
```

2. Then, we computed the standard error to calculate the confidence interval: $SE=\sigma/\sqrt{N}$

```
> N = length(mydata$D1.D6.Jaccard.distance)
> N
[1] 335
> sigma = tapply(mydata$D1.D6.Jaccard.distance,list(antibiotics =mydata$Antibiotic.class),sd)
> sigma
antibiotics
      FQN       OBL       PBL
0.1525906 0.1404564 0.1668146
> SE <- sigma/sqrt(N)
> SE
antibiotics
        FQN         OBL         PBL
0.008336917 0.007673955 0.009114053
```

3. We used the Zscore method for each interval 90%, 95%, 99% which equals to 1.645, 1.96, 2.576 respectively. Then to calculate each margin error, we multiplied it by the standard error.

```
199  ME90 = SE * 1.645     # Z(0.9)
200  ME95 = SE * 1.96      # Z(0.95)
201  ME99 = SE * 2.576     # Z(0.99)
202
202:1  (Top Level)
Console  Terminal  Background Jobs
R 4.2.2 · ~/
> ME90
antibiotics
       FQN         OBL         PBL
0.01371423 0.01262366 0.01499262
> ME95
antibiotics
       FQN         OBL         PBL
0.01634036 0.01504095 0.01786354
> ME99
antibiotics
       FQN         OBL         PBL
0.02147590 0.01976811 0.02347780
>
```

4. Last step, we calculated lower and upper bound that represents the interval for each confidence interval.

```
> # 90% C.I
> lower.bound = meann - ME90
> upper.bound = meann + ME90
> df90
    lower.bound upper.bound
FQN   0.6402201   0.6676485
OBL   0.6596301   0.6848774
PBL   0.6270213   0.6570065
```

```
> # 95% C.I
> lower.bound = meann - ME95
> upper.bound = meann + ME95
> df95
    lower.bound upper.bound
FQN   0.6375939   0.6702747
OBL   0.6572128   0.6872947
PBL   0.6241503   0.6598774
```

```
> # 99% C.I
> lower.bound = meann - ME99
> upper.bound = meann + ME99
> df99
    lower.bound upper.bound
FQN   0.6324584   0.6754102
OBL   0.6524856   0.6920218
PBL   0.6185361   0.6654917
```

5.2 How would you describe those inferences and what do you observe in terms of the interval width when request higher confidence:

We can see that all the antibiotics have narrow confidence intervals which means that we are confident that our estimate is close to our true population value. By increasing our C.I, the width increases due to the process of achieving high degree of certainty. So, we know that the resulting intervals will be wider and less precise.

6. Hypothesis Testing

6.1 We hypothesis that Chao/Shannon at day 6 different between CDI vs ND:

1. H_0: Chao/Shannon at day 6 is the same for CDI vs ND.
   H_A: Chao/Shannon at day 6 different for CDI vs ND.
2. Assuming normality and homoscedasticity, we used Two-sample t-test for Chao and Shannon each.
   - D6.Chao1.diversity for CDI vs ND: p-value = 0.3362 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis, so we assume that the true difference in means for D6.Chao1.diversity CDI and ND are equal to 0; they are similar.

```
        Two Sample t-test

data:  mydata[mydata$Outcome == "CDI", ]$D6.Chao1.diversity and mydata[mydata$Outcome == "ND", ]$D6.Chao1.diversity
t = -0.96323, df = 311, p-value = 0.3362
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -100.1675   34.3268
sample estimates:
mean of x mean of y
 145.4545  178.3749
```

- D6.Shannon.diversity for CDI vs ND: p-value = 0.7164 which is bigger than α(0.05), thus, we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and the true difference in means for D6.Shannon.diversity CDI and ND are equal to 0; they are similar.

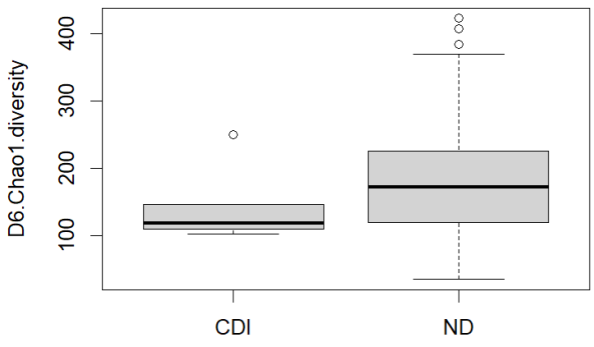
```
        Two Sample t-test

data:  mydata[mydata$Outcome == "CDI", ]$D6.Shannon.diversity and mydata[mydata$Outcome == "ND", ]$D6.Shannon.diversity
t = 0.3636, df = 311, p-value = 0.7164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6437360  0.9355816
sample estimates:
mean of x mean of y
 3.049826  2.903903
```

6.2 Assess whether the previous test assumptions have been met for the test:

1. D6.Chao1.diversity for CDI vs ND: -
   - Test normality: The p-values are significant; thus, we have enough evidence to reject the null hypothesis and the data is not normal.

```
> shapiro.test(mydata[mydata$Outcome=="CDI",]$D6.Chao1.diversity)

        Shapiro-Wilk normality test

data:  mydata[mydata$Outcome == "CDI", ]$D6.Chao1.diversity
W = 0.7679, p-value = 0.04323

> shapiro.test(mydata[mydata$Outcome=="ND",]$D6.Chao1.diversity)

        Shapiro-Wilk normality test

data:  mydata[mydata$Outcome == "ND", ]$D6.Chao1.diversity
W = 0.97345, p-value = 1.831e-05
```

   - Test homoscedasticity: From the boxplot we can see that they don't have the same variance, thus heteroscedasticity.



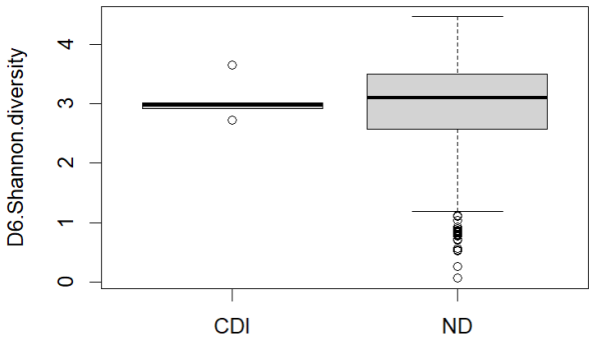
   - Since the assumptions have not been met for D6.Chao1.diversity, so we used Wilcoxon rank sum test. The p-value = 0.2786 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and D6.Chao1.diversity in CDI and ND are similar.

```
> # Than we use Wilcoxon rank sum test
> wilcox.test(mydata[mydata$Outcome=="CDI",]$D6.Chao1.diversity,
+             mydata[mydata$Outcome=="ND",]$D6.Chao1.diversity, paired= F)

        Wilcoxon rank sum test with continuity correction

data:  mydata[mydata$Outcome == "CDI", ]$D6.Chao1.diversity and mydata[mydata$Outcome == "ND", ]$D6.Chao1.diversity
W = 552, p-value = 0.2786
alternative hypothesis: true location shift is not equal to 0
```

2. D6.Shannon.diversity for CDI vs ND: -
   - Test normality: The p-values one is significant and the other is not; thus, we have enough evidence to reject the null hypothesis and the data is <u>not normal</u>.

```
> shapiro.test(mydata[mydata$Outcome=="CDI",]$D6.Shannon.diversity)

        Shapiro-Wilk normality test

data:  mydata[mydata$Outcome == "CDI", ]$D6.Shannon.diversity
W = 0.83065, p-value = 0.1407

> shapiro.test(mydata[mydata$Outcome=="ND",]$D6.Shannon.diversity)

        Shapiro-Wilk normality test

data:  mydata[mydata$Outcome == "ND", ]$D6.Shannon.diversity
W = 0.91799, p-value = 6.086e-12
```

   - Test homoscedasticity: From the boxplot we can see that they don't have the same variance, thus <u>heteroscedasticity</u>.



   - Since the assumptions have not been met for D6.Shannon.diversity, so we used Wilcoxon rank sum test. The p-value = 0.8014 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and D6.Shannon.diversityin CDI and ND are similar.

```
> # Than we use Wilcoxon rank sum test
> wilcox.test(mydata[mydata$Outcome=="CDI",]$D6.Shannon.diversity,
+             mydata[mydata$Outcome=="ND",]$D6.Shannon.diversity, paired= F)

        Wilcoxon rank sum test with continuity correction

data:  mydata[mydata$Outcome == "CDI", ]$D6.Shannon.diversity and mydata[mydata$Outcome == "ND", ]$D6.Shannon.diversity
W = 719, p-value = 0.8014
alternative hypothesis: true location shift is not equal to 0
```

6.3 We hypothesis that Jacard distance "different" in the group receiving BOL Antibiotics compared to the FQ antibiotics B:

1. H_0: Jacard distance in the group receiving OBL compared to the FQN is similar.
   H_A: Jacard distance in the group receiving OBL compared to the FQN is different.
2. Assuming heteroscedasticity, we used Two-sample Welch t-test. The p-value = 0.4538 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and D1.D6.Jaccard.distance is similar between OBL and FQN Antibiotics.

```
> t.test(mydata[mydata$Antibiotic.class=="FQN",]$D1.D6.Jaccard.distance,mydata[mydata$Antibiotic.class=="OBL",]$D1.D6.Jaccard.distance, var.equal = F, paired = F)

        Welch Two Sample t-test

data:  mydata[mydata$Antibiotic.class == "FQN", ]$D1.D6.Jaccard.distance and mydata[mydata$Antibiotic.class == "OBL", ]$D1.D6.Jaccard.distance
t = -0.75196, df = 102.69, p-value = 0.4538
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06663772  0.02999888
sample estimates:
mean of x mean of y
0.6539343 0.6722537
```

6.4 Assess the previous test assumption:

1. Test normality: The p-values are not significant; thus, we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis, so we assume normality.

```
> shapiro.test(mydata[mydata$Antibiotic.class=="FQN",]$D1.D6.Jaccard.distance)

        Shapiro-Wilk normality test

data:  mydata[mydata$Antibiotic.class == "FQN", ]$D1.D6.Jaccard.distance
W = 0.98216, p-value = 0.573

> shapiro.test(mydata[mydata$Antibiotic.class=="OBL",]$D1.D6.Jaccard.distance)

        Shapiro-Wilk normality test

data:  mydata[mydata$Antibiotic.class == "OBL", ]$D1.D6.Jaccard.distance
W = 0.97742, p-value = 0.05651
```

2. Test heteroscedasticity: The p-value = 0.4596 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and D1.D6.Jaccard.distance has equal variance between OBL and FQN Antibiotics.

```
> var.test(mydata[mydata$Antibiotic.class=="FQN",]$D1.D6.Jaccard.distance,
+          mydata[mydata$Antibiotic.class=="OBL",]$D1.D6.Jaccard.distance)

        F test to compare two variances

data:  mydata[mydata$Antibiotic.class == "FQN", ]$D1.D6.Jaccard.distance and mydata[mydata$Antibiotic.class == "OBL", ]$D1.D6.Jaccard.distance
F = 1.1802, num df = 55, denom df = 110, p-value = 0.4596
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7573297 1.9065161
sample estimates:
ratio of variances
          1.180246
```

3. The assumptions have not been met and we used Two-sample t-test. The p-value = 0.4407 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and D1.D6.Jaccard.distance is similar between OBL and FQN Antibiotics.

```
> t.test(mydata[mydata$Antibiotic.class=="FQN",]$D1.D6.Jaccard.distance,mydata[mydata$Antibiotic.class=="OBL",]$D1.D6.Jaccard.distance, paired = F, var.equal = T)

        Two Sample t-test

data:  mydata[mydata$Antibiotic.class == "FQN", ]$D1.D6.Jaccard.distance and mydata[mydata$Antibiotic.class == "OBL", ]$D1.D6.Jaccard.distance
t = -0.77285, df = 165, p-value = 0.4407
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06512087  0.02848203
sample estimates:
mean of x mean of y
0.6539343 0.6722537
```

6.5 We hypothesis that Jacard distance is different between the different Antibiotics. Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity):

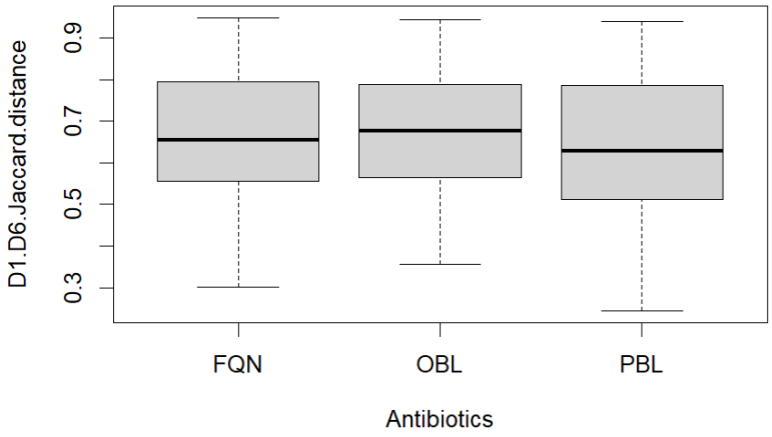
1. H_0: There is no significant difference in Jaccard distances between the different Antibiotics.
   H_A: There is a significant difference in Jaccard distances between at least two of the different Antibiotics.

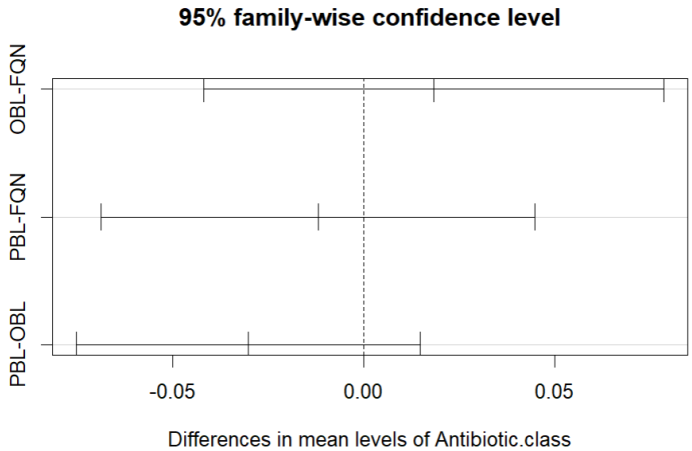
2. Assuming normality and homoscedasticity, we used ANOVA test. The p-value = 0.287 which is bigger than α(0.05), thus we don't have enough evidence to reject the null hypothesis in support of alternative hypothesis and D1.D6.Jaccard.distance is similar between the different Antibiotics.

```
> summary(AnovaModel)
                  Df Sum Sq Mean Sq F value Pr(>F)
Antibiotic.class    2  0.061 0.03056   1.253  0.287
Residuals         332  8.098 0.02439
```

3. To determine which pairs of means are significantly different from each other we used Tukey multiple comparisons of means. And the result of the p-values shows that there's no significant difference between any one of the Antibiotics. We can also see that in the boxplot and the Tukey result plotted.

```
> TukeyHSD(AnovaModel)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = D1.D6.Jaccard.distance ~ Antibiotic.class, data = mydata)

$Antibiotic.class
              diff         lwr        upr     p adj
OBL-FQN  0.01831942 -0.04194691 0.07858574 0.7544150
PBL-FQN -0.01192043 -0.06865495 0.04481408 0.8739116
PBL-OBL -0.03023985 -0.07521348 0.01473378 0.2544052
```

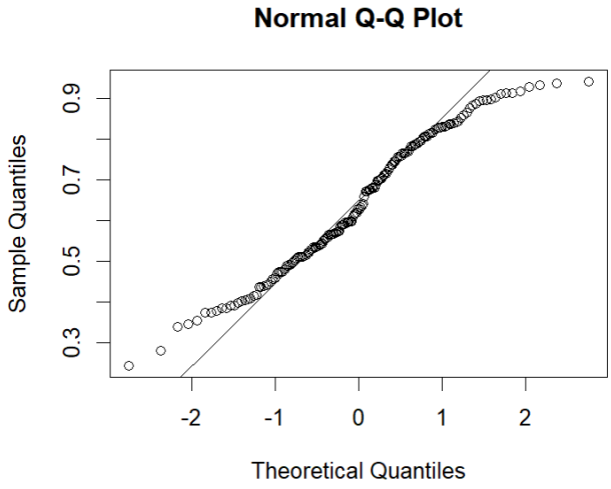


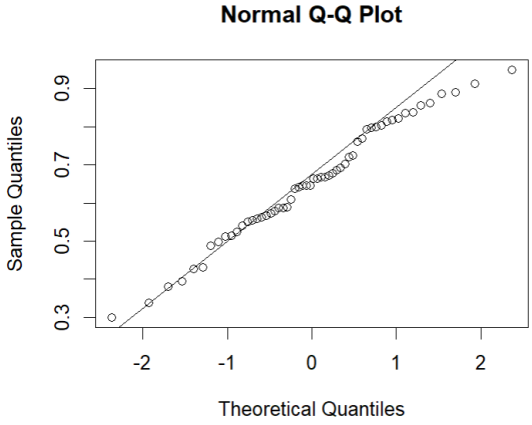


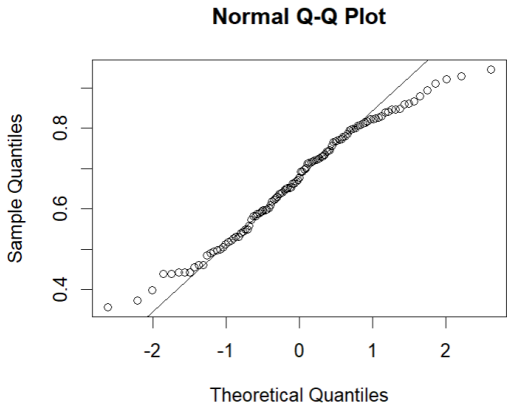
4. Assess assumption of normality by Shapiro and Q-Q plot. Based on Shapiro test, p-value = 0.573 for FQN, p-value = 0.05651 for OBL, p-value= 0.001216 for PBL and since all of them

are bigger than α(0.05), except for PBL, thus it is significant; we have enough evidence to reject the null hypothesis and the data is not normal. Then we used Q-Q plot for certainty and the all the plots showed non-normality as well.

```
mydata$Antibiotic.class: FQN

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98216, p-value = 0.573

------------------------------------
mydata$Antibiotic.class: OBL

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97742, p-value = 0.05651

------------------------------------
mydata$Antibiotic.class: PBL

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97054, p-value = 0.001216
```







5. The right test to be used in this case is Kruskal Wallis rank based test for non-normal data and Dunn test.

```
> kruskal.test(D1.D6.Jaccard.distance~Antibiotic.class, AAD)

        Kruskal-Wallis rank sum test

data:  D1.D6.Jaccard.distance by Antibiotic.class
Kruskal-Wallis chi-squared = 2.2031, df = 2, p-value = 0.3324
```

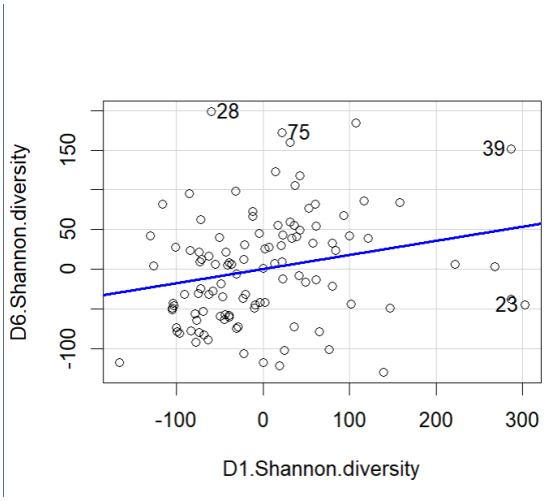
```
> dunn_test = dunn.test(AAD$D1.D6.Jaccard.distance, AAD$Antibiotic.class, method = "b
onferroni")
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 2.2031, df = 2, p-value = 0.33


                           Comparison of x by group
                                 (Bonferroni)
Col Mean-|
Row Mean |        FQN        OBL
---------+----------------------
     OBL |  -0.652742
         |     0.7709
         |
     PBL |   0.483142   1.484185
         |     0.9435     0.2066

alpha = 0.05
Reject Ho if p <= alpha/2
```

7. Linear Regression

7.1 Fit a linear regression to the data and interpret the regression coefficient.

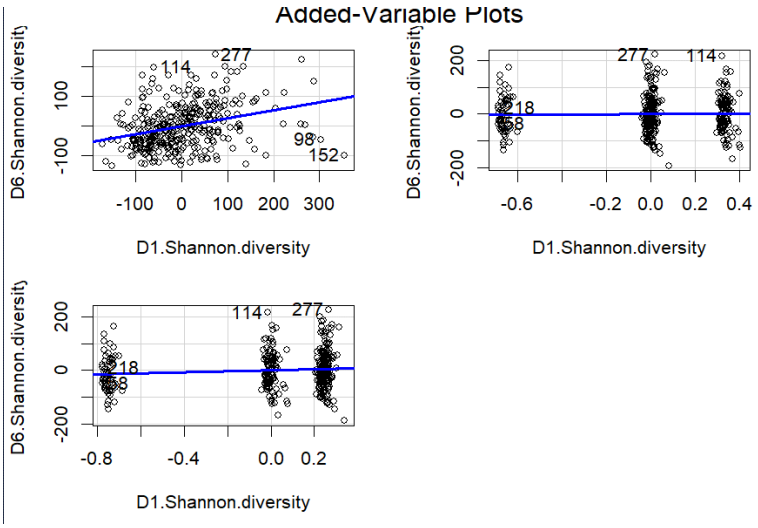
1. We take D1.Chao1.diversity as our regressor and D6.Chao1.diversity as the outcome according to the "OBL" Antibiotic. After using the linear model and do the summary, the results show the p-value = 0.0145 which is bigger than α(0.05), thus we have enough evidence to reject the null hypothesis and D1.Chao1.diversity can predict D6.Chao1.diversity.

```
Residuals:
     Min       1Q   Median       3Q      Max
-154.556  -49.832   -0.889   40.502  209.269

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   132.00757   15.69084   8.413 1.72e-13 ***
D1.Chao1.diversity[AAD$Antibiotic.class == "OBL"]   0.17925    0.07217   2.484   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 67.82 on 109 degrees of freedom
Multiple R-squared:  0.05357,   Adjusted R-squared:  0.04488
F-statistic: 6.169 on 1 and 109 DF,  p-value: 0.01452
```



2. Another model with multiple linear regression with the same axis but for all Antibiotics. The Intercept represents the value of the response variable when all the predictor variables are equal to zero. The residual standard error is 71.84, which means that the typical distance between the observed and predicted values of the response variable is around 71.84 units. The multiple R-squared is 0.1018, which means that the predictor variables explain only 10.18% of the variance in the response variable. The F-statistic is 12.5, with a p-value of 9.148e-08, which is highly significant. It suggests that the predictor variables collectively have a significant effect on the response variable.

```
Residuals:
     Min       1Q   Median       3Q      Max
-194.395  -48.435   -6.777   42.272  221.307

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        109.34963   13.79541   7.927 3.48e-14 ***
D1.Chao1.diversity   0.27316    0.04654   5.870 1.06e-08 ***
Antibiotic.classOBL  4.03880   11.79444   0.342    0.732
Antibiotic.classPBL 17.69613   11.10497   1.594    0.112
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.84 on 331 degrees of freedom
Multiple R-squared:  0.1018,    Adjusted R-squared:  0.09366
F-statistic:  12.5 on 3 and 331 DF,  p-value: 9.148e-08
```



7.2 Calculate and interpret a 95% confidence interval of the regression.

The output of confint() is a table that provides the 95% confidence intervals for each coefficient in the linear regression model.

8. Contribution

Name	Task
Aya	Statistical Inference, Hypothesis
Dina	Normality, Hypothesis, Linear
Marvy	Graphics, Hypothesis
Hazem	Outliers, Descriptive, Linear