ANTIBIOTICS ASSOCIATED DIARRHEA I

BIOMEDICAL STATISTICS – BMD407

Dr. Mohamed Mysara & Eng. Mariam Oweda

Mayar Tarek	202002151
Fatema Gamal	202002674
Pola Girgis	202002248
Raghad Mohamed	202000571

Table of Contents

INTRODUCTION

Antibiotics-associated diarrhea (AAD) is a prevalent and economically significant disease worldwide. It occurs as a result of disruptions in the normal commensal bacteria population, which can lead to the colonization of pathogenic bacteria such as Clostridioides difficile (C. difficile), causing C. difficile infection (CDI), or diarrhea itself, referred to as AAD. Some individuals may not experience diarrhea (referred to as ND) despite being exposed to antibiotics.

To address the challenges posed by AAD and CDI, a prospective clinical trial was conducted across six European countries: France, the Netherlands, Spain, Romania, Germany, and Greece. The study enrolled approximately 1000 patients who met specific inclusion criteria, including testing negative for CDI and not having taken antibiotics prior to enrollment. These patients were divided into three groups, each receiving one of the three classes of antibiotics: PBL (penicillin + beta-lactamase inhibitor), OBL (other beta-lactamase antibiotics), and FQN (fluoroquinolones). Stool samples were collected from the patients on Day 1 (prior to antibiotic administration) and again on Day 6. The patients were then monitored for a period of 90 days, with additional stool samples collected in case of diarrhea.

The microbial composition of the stool samples was analyzed using 16S microbial profiling. The study focused on assessing microbial richness (Chao index), microbial evenness (Shannon index), and beta-diversity distance (Jaccard distance) between Day 1 and Day 6 samples (referred to as D1 and D6, respectively).

The objectives of the study include investigating the following outcomes:

- Changes in microbial composition between baseline samples across different countries, gender, and age groups.
- Temporal changes in microbial composition over time (D1, D6, and stool samples) for each antibiotic class.
- Comparisons of microbial composition changes across different antibiotic classes over time.

By examining these outcomes, the study aims to enhance our understanding of the impact of antibiotics on the gut microbiota and to identify potential factors that contribute to AAD and CDI. The findings may provide valuable insights for the development of preventive and therapeutic strategies to mitigate the adverse effects of antibiotics on the gut microbiome and reduce the incidence of AAD and CDI.

1. DESCRIPTIVE STATISTICS

1.1. Data reading

The first thing we do is read our AAD data, which stands for antibiotic-associated diarrhoea I data by uploading the file in our project.

1. 2. Descriptive statistics

- We apply some descriptive statistics to the data after reading it.
- To view the descriptive information for each column, we used summary function.

```
> summary(AAD)
  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
```

- For each variable in our data, we calculate the mean, median, minimum, maximum, first and third quartiles.

```
> mean(AAD$D1.Shannon.diversity,na.rm=TRUE)
[1] 3.249321
> median(AAD$D1.Shannon.diversity,na.rm=TRUE)
[1] 3.392265
> min(AAD$D1.Shannon.diversity,na.rm=TRUE)
[1] 0.127635
> max(AAD$D1.Shannon.diversity,na.rm=TRUE)
[1] 4.465318
> quantile(AAD$D1.Shannon.diversity,na.rm=TRUE,c(0.25,0.75))
     25%      75%
2.898378 3.765255
```

Results of D1. Shannon.diversity

```
# calculate the following: mean, median, minimum, maximum,
# D1.Shannon.diversity variable
mean(AAD$D1.Shannon.diversity,na.rm=TRUE)
median(AAD$D1.Shannon.diversity,na.rm=TRUE)
min(AAD$D1.Shannon.diversity,na.rm=TRUE)
max(AAD$D1.Shannon.diversity,na.rm=TRUE)
quantile(AAD$D1.Shannon.diversity,na.rm=TRUE,c(0.25,0.75))

# D6.Shannon.diversity variable
mean(AAD$D6.Shannon.diversity,na.rm=TRUE)
median(AAD$D6.Shannon.diversity,na.rm=TRUE)
min(AAD$D6.Shannon.diversity,na.rm=TRUE)
max(AAD$D6.Shannon.diversity,na.rm=TRUE)
quantile(AAD$D6.Shannon.diversity,na.rm=TRUE,c(0.25,0.75))

# D1.Chao1.diversity variable
mean(AAD$D1.Chao1.diversity,na.rm=TRUE)
median(AAD$D1.Chao1.diversity,na.rm=TRUE)
min(AAD$D1.Chao1.diversity,na.rm=TRUE)
max(AAD$D1.Chao1.diversity,na.rm=TRUE)
quantile(AAD$D1.Chao1.diversity,na.rm=TRUE,c(0.25,0.75))

# D6.Chao1.diversity variable
mean(AAD$D6.Chao1.diversity,na.rm=TRUE)
median(AAD$D6.Chao1.diversity,na.rm=TRUE)
min(AAD$D6.Chao1.diversity,na.rm=TRUE)
max(AAD$D6.Chao1.diversity,na.rm=TRUE)
quantile(AAD$D6.Chao1.diversity,na.rm=TRUE,c(0.25,0.75))

# D1.D6.Jaccard.distance variable
mean(AAD$D1.D6.Jaccard.distance,na.rm=TRUE)
median(AAD$D1.D6.Jaccard.distance,na.rm=TRUE)
min(AAD$D1.D6.Jaccard.distance,na.rm=TRUE)
max(AAD$D1.D6.Jaccard.distance,na.rm=TRUE)
quantile(AAD$D1.D6.Jaccard.distance,na.rm=TRUE,c(0.25,0.75)
```

- For the categorical variables, Antibiotic.class and Outcome, we calculate the frequency table. The table function gets the number of each categorical variable in our data like the outcome and antibiotics. The results in figure below.

```
> table(AAD$Antibiotic.class)

FQN OBL PBL
 56 111 168
> table(AAD$Outcome)

AAD CDI  ND
 22   5 308
```

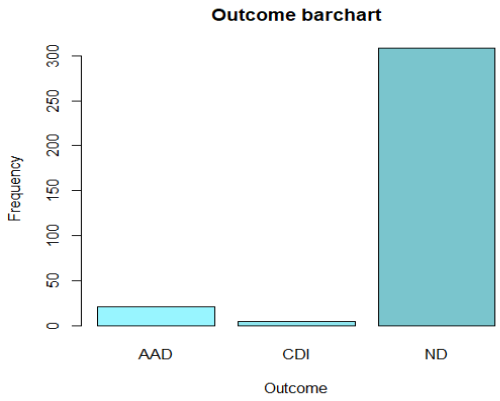
- We perform two calculations of the correlation coefficient. The first one included D1 and D6 Shannon, and the second one included D1 and D6 Chao.

```
> cor(AAD$D1.Shannon.diversity,AAD$D6.Shannon.diversity, use="complete.obs")
[1] 0.2208003
> cor(AAD$D1.Chao1.diversity,AAD$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)

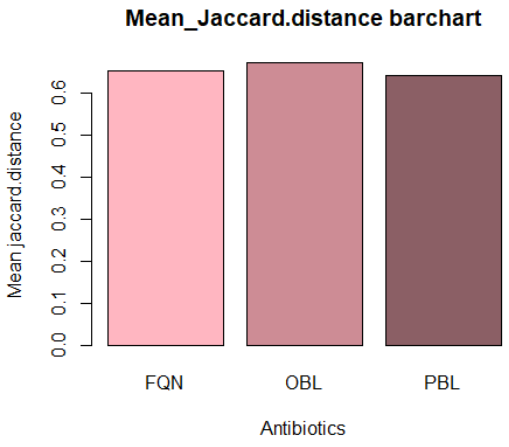
```
#1st: Barchart
barplot(table(AAD$Outcome), xlab="Outcome",ylab="Frequency",
        col = c("cadetblue1","cadetblue2","cadetblue3"),main="Outcome barchart")
```



- We used a table on variable outcomes to be divided into outcome categories to see the frequency of each category. And used the col argument to distinguish between them.
- The output of this code is a bar plot with the x-axis representing the different outcomes, the y-axis representing the frequency counts, and bars representing the frequencies of each outcome.
- From the plot, we can conclude that ND is the most frequent outcome in the dataset. which experienced no diarrhea (ND).

2.2. Generate a bar chart graph with mean Jaccard distance in BOL, FQ, OBL

```
#2nd: Barchart
barplot(tapply(AAD$D1.D6.Jaccard.distance,
        list(Antibiotic=AAD$Antibiotic.class),mean,na.rm=T),
        xlab="Antibiotics",ylab="Mean jaccard.distance",
        main="Mean_Jaccard.distance barchart", col = c("lightpink","lightpink3","#8B5F65"))
```

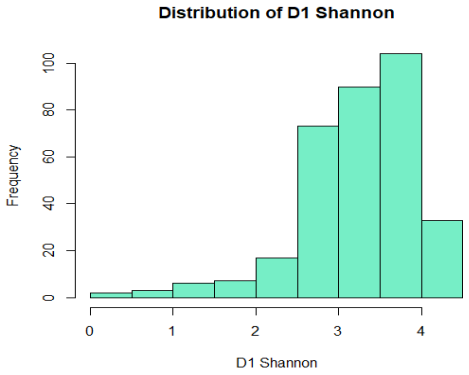


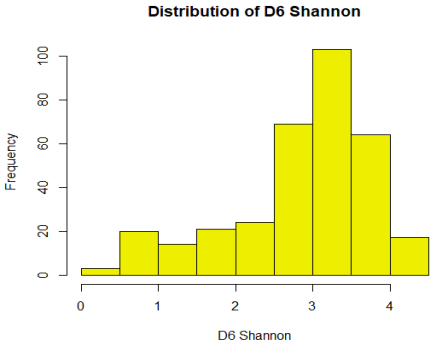
- The output of this code is a bar plot where each bar represents the mean Jaccard distance for a specific antibiotic. The x-axis represents the different antibiotics, the y-axis represents the mean Jaccard distance, and the bars represent the mean values.
- From the plot, we can conclude that the most affecting antibiotic across the days is OBL

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

```
#3rd: Histogram
hist(AAD$D1.Shannon.diversity,xlab="D1 Shannon",main="Distribution of D1 Shannon",col = "aquamarine2")
hist(AAD$D6.Shannon.diversity,xlab="D6 Shannon",main="Distribution of D6 Shannon",col = "yellow2")
```

- We used the hist function to plot the histogram and display the frequencies of D1 and D6 Shannon.
- Both plots are left-skewed not normally distributed, so they have a smaller mean.

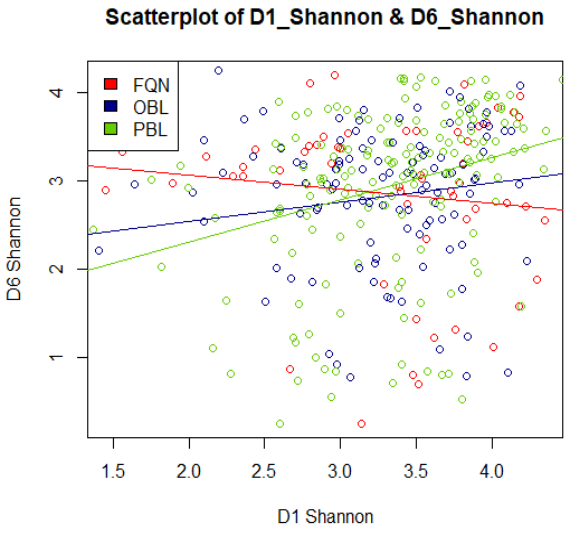




2.4. Make a scatterplot of 2 continuous variables D1 Shannon and D6 Shannon, and add the regression lines for each antibiotic

```
#4th: Scattarplot
plot(D6.Shannon.diversity[Antibiotic.class=="FQN"]~D1.Shannon.diversity[Antibiotic.class=="FQN"],data = AAD,xlab="D1 Shannon",ylab="D6 Shannon",col="red",main=
points(D6.Shannon.diversity[Antibiotic.class=="OBL"]~ D1.Shannon.diversity[Antibiotic.class=="OBL"],data = AAD,xlab="D1 Shannon",ylab="D6 Shannon",col="darkblu
points(D6.Shannon.diversity[Antibiotic.class=="PBL"]~ D1.Shannon.diversity[Antibiotic.class=="PBL"],data = AAD,xlab="D1 Shannon",ylab="D6 Shannon",col="chartre
legend("topleft", legend = c("FQN","OBL","PBL"), fill=c("red","darkblue","chartreuse3"))
#regression lines
abline(lm(AAD$D6.Shannon.diversity[AAD$Antibiotic.class=="FQN"]~AAD$D1.Shannon.diversity[AAD$Antibiotic.class=="FQN"]),col="red")
abline(lm(AAD$D6.Shannon.diversity[AAD$Antibiotic.class=="OBL"]~AAD$D1.Shannon.diversity[AAD$Antibiotic.class=="OBL"]),col="darkblue")
abline(lm(AAD$D6.Shannon.diversity[AAD$Antibiotic.class=="PBL"]~AAD$D1.Shannon.diversity[AAD$Antibiotic.class=="PBL"]),col="chartreuse3")
```

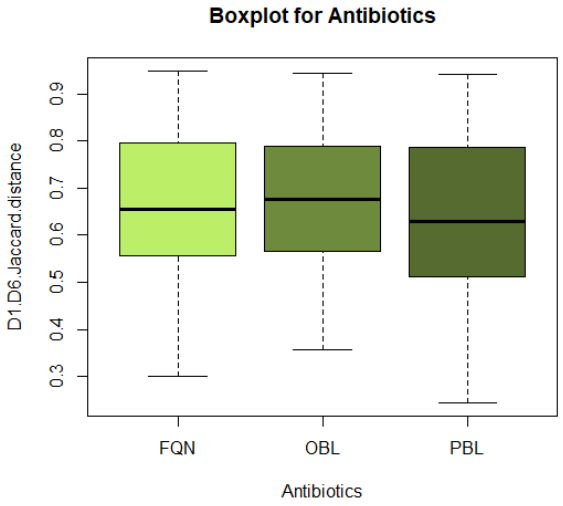
- we used plot and points function to make scatterplot between the three antibiotics to know the relation between them.
- Code generates a scatterplot of the "D1 Shannon" diversity values on the x-axis against the "D6 Shannon" diversity values on the y-axis, for three different categories of the "Antibiotic.class" variable.
- From the plot we can see that there is no correlation between the three antibiotics over days.
- For the regression lines, the relationship is nonlinear between them, and the three have weak relationships as there are a lot of outliers. So, we need linear regression for this dataset.



2.5. Make a boxplot of Jaccard distance and a separate boxplots per Antibiotics.

```
#5th: Box plot
boxplot(D1.D6.Jaccard.distance~Antibiotic.class,main="Boxplot for Antibiotics"
        ,col = c("darkolivegreen2","darkolivegreen4","darkolivegreen"),xlab="Antibiotics",data = AAD)
```

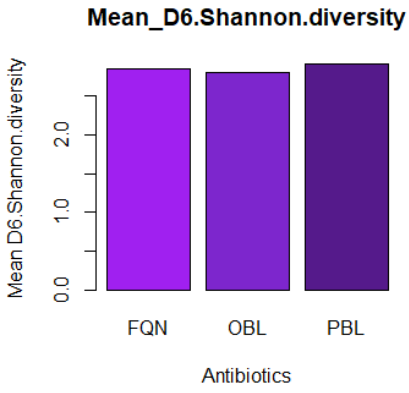
- From the plot, there is no significant difference between FQN & OBL, while PBL has a significant difference, so we must check by making a var.test or lavene test.



2.6. Generate a bar chart graph with mean D6.Shannon in BOL, FQ, OBL

```
barplot(tapply(AAD$D6.Shannon.diversity, list(Antibiotic=AAD$Antibiotic.class),mean,na.rm=T), xlab="Antibiotics",ylab="Mean D6.Shann
        main="Mean_D6.Shannon.diversity", col = c("purple","purple3","purple4"))
```

- The output of this code is a bar plot where each bar represents the mean D6. Shannon for a specific antibiotic. The x-axis represents the different antibiotics, the y-axis represents the mean D6. Shannon, and the bars represent the mean values.
- From the plot, we can conclude that the most affecting antibiotic in day 6 is PBL.

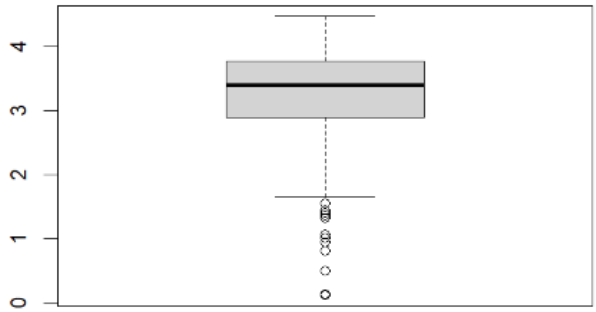


3. OUTLIERS

Plotting the Outliers in all the Numerical Data that we have (Upon analyzing the numerical columns in the dataset, we have identified outliers that need to be acknowledged in order to prevent potential errors).

3.1. D1 Shannon diversity

```
boxplot(AAD$D1.Shannon.diversity)
boxplot(AAD$D1.Shannon.diversity)$out
```



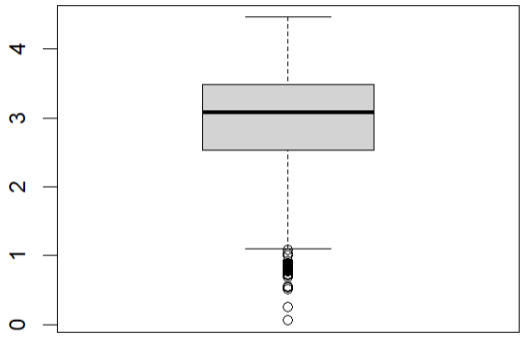
Outliers:

```
[1] 1.010512 1.364614 0.127635 0.507176 1.063788 1.561207 0.816340 1.448330 1.331024 0.9505
[11] 1.404700 0.143722
```

3.2. D6 Shannon diversity

From the results obvious that has greatest no. of outliers

```
boxplot(AAD$D6.Shannon.diversity)
boxplot(AAD$D6.Shannon.diversity)$out
```

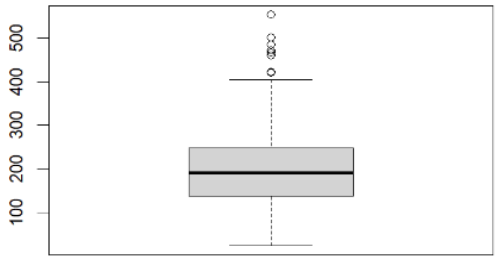


Outliers:

```
[1] 0.537671 1.011686 0.805769 0.070407 0.519615 0.931239 0.878636 1.092522 0.822535 0.847067
[11] 0.927137 0.698893 0.874226 0.901139 0.738428 0.821437 0.854022 0.789490 1.039348 0.259205
[21] 0.262554 0.716793 0.777983 0.836249 0.558723 0.804012
```

3.3. D1 Chao diversity

```
boxplot(AAD$D1.Chao1.diversity)
boxplot(AAD$D1.Chao1.diversity)$out
```

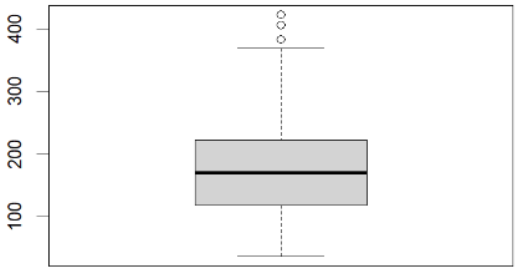


Outliers:

```
459.4412 422.8971 500.5446 484.5753 552.9333 485.3214 420.5882 472.1224 466.5200
```

3.4. D6 Chao diversity

```
boxplot(AAD$D6.Chao1.diversity)
boxplot(AAD$D6.Chao1.diversity)$out
```

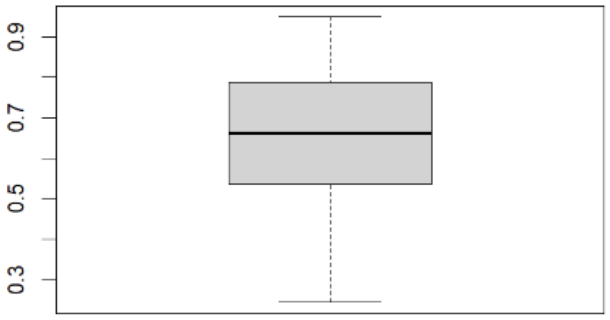


Outliers:

```
407.0286 383.6596 422.7500
```

3.5. D1.D6 Jaccard distance

```
boxplot(AAD$D1.D6.Jaccard.distance)
boxplot(AAD$D1.D6.Jaccard.distance)$out      #Has No Outliers
```

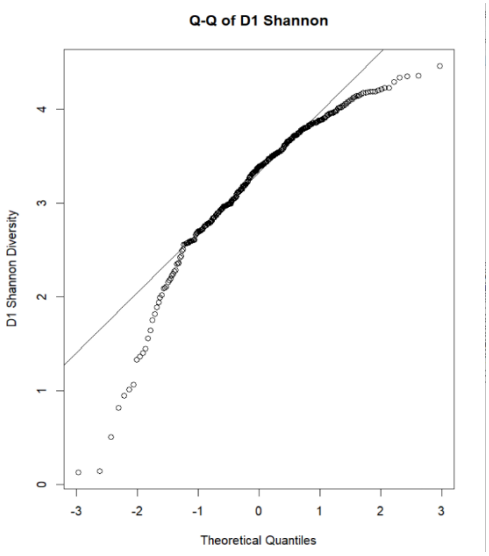


4. NORMALITY/ HOMOSCEDASTICITY

4.1. Checking the normality on all the numeric datasets:

4. D1 Shannon diversity

```
#1- D1.Shannon.diversity
qqnorm(AAD$D1.Shannon.diversity, ylab="D1 Shannon Diversity")
qqline(AAD$D1.Shannon.diversity,ylab="D1 Shannon Diversity")
#from the qq plot we can see the d1 shannon diversity data is not r

shapiro.test(AAD$D1.Shannon.diversity) #p-value = 4.134e-13 < 0.05
#from the shapiro test, we can see the p-value is less than 0.05 s
```



First, we check the normality of D1 Shannon by graphical representation using a QQ plot to plot our data against a theoretical distribution to compare to and added a reference line using (qqline)

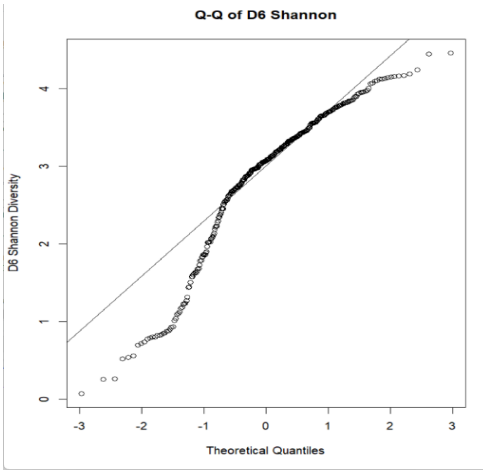
From our qq plot and reference line we can see that not all points in our data Fall on the reference line, suggesting that the data isn't normally distributed. We can make sure of this using Shapiro-wilk test.

The null hypothesis of Shapiro-wilk test is that the data is normally distributed. The p-value for D1 Shannon diversity was 4.134e-13 which is less than the significance level (0.05) so we can reject the null and assume that D1 Shannon data is not normally distributed.

5. D6 Shannon diversity

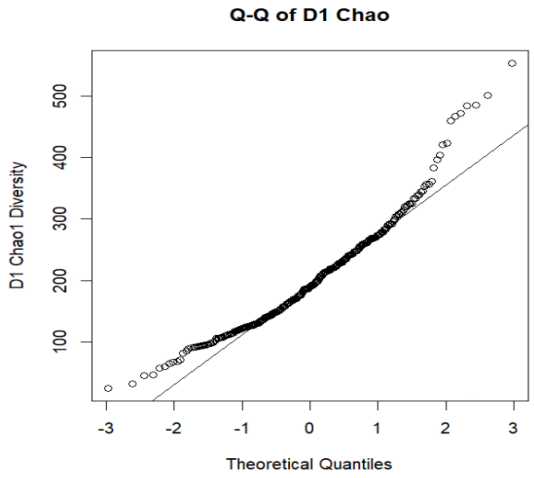
```
#2- D6.Shannon.diversity
qqnorm(AAD$D6.Shannon.diversity,ylab="D6 Shannon Diversity", main="Q-Q of D6 Shannon")
qqline(AAD$D6.Shannon.diversity, ylab="D6 Shannon Diversity")    #from the qq plot we can

shapiro.test(AAD$D6.Shannon.diversity) #(p-value = 8.795e-13) < 0.05
#from shapiro the p value is less than 0.05 so we can reject the assumption of normality
```

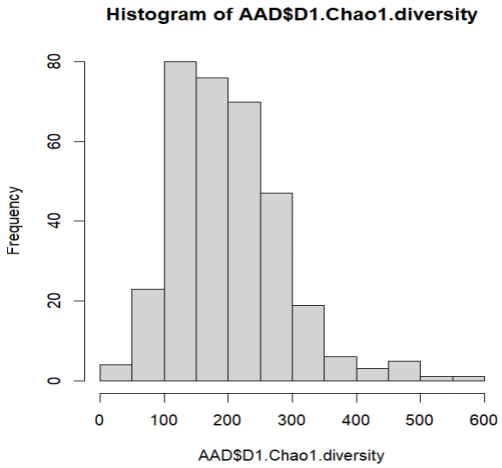
Using QQ plot and then Shapiro-wilk test (p-value = 8.795e-13) < 0.05 so we can reject the null and assume non-normality of D6 Shannon



6. D1 Chao diversity

```
#3- D1.Chao1.diversity
hist(AAD$D1.Chao1.diversity)  #histogram is skewed not normal
qqnorm(AAD$D1.Chao1.diversity,ylab="D1 Chao1 Diversity",main="Q-Q of D1 Chao")
qqline(AAD$D1.Chao1.diversity)

shapiro.test(AAD$D1.Chao1.diversity) #(p-value = 2.489e-09) < 0.05
#from shapiro the p value < 0.05 so we can reject the assumption of normality for
```





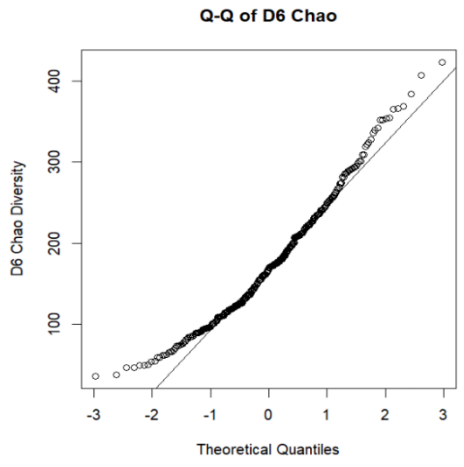
Here, we added another way of checking normality graphically using a histogram to plot the data. The data appears to be right-skewed and not all the points follow the reference line in the QQ plot

To check with Shapiro-wilk the p-value is 2.489e-09 < 0.05 so we can assume non-normality.

7. D6 Chao diversity

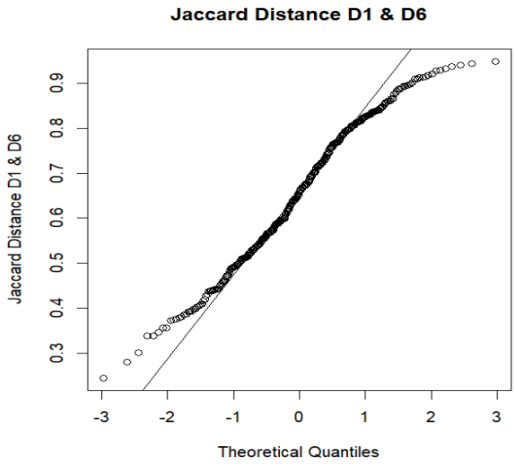
```
#4- D6.Chao1.diversity
qqnorm(AAD$D6.Chao1.diversity,ylab="D6 Chao Diversity",main="Q-Q of D6 Chao")
qqline(AAD$D6.Chao1.diversity)                              #from the qq plot we can s

shapiro.test(AAD$D6.Chao1.diversity)  #(p-value = 1.776e-06) < 0.05
#from shapiro the p value < 0.05 so we can reject assumption of normality
```

Using QQ plot and then Shapiro-wilk test (p-value = 1.776e-06) < 0.05 so we can reject the null and assume non-normality of D6 Shannon



8. D1.D6 Jaccard distance

```
#5- D1.D6.Jaccard.distance
qqnorm(AAD$D1.D6.Jaccard.distance, ylab="Jaccard Distance D1 & D6",main="Jaccard Distance D1 & D6")
qqline(AAD$D1.D6.Jaccard.distance)
shapiro.test(AAD$D1.D6.Jaccard.distance) #(p-value = 8.022e-05) < 0.05
#from shapiro the p value < 0.05 so we can reject assumption of normality
```



Using QQ plot and then Shapiro-wilk test (p-value = 8.022e-05) < 0.05 so we can reject the null and assume non-normality of D6 Shannon

Conclusion: After performing two tests for each numerical variable, Using graphical and Shapiro test none of the variables appeared to be normally distributed.

4.2. Checking the homoscedasticity on all the numeric datasets:

To test homoscedasticity we used two methods, Levene test as its robust against any types of non-normality and Bartlett's test.
The null hypothesis for both Levene's test and Bartlett's test is that the variances of the groups or samples are equal (homoscedastic)

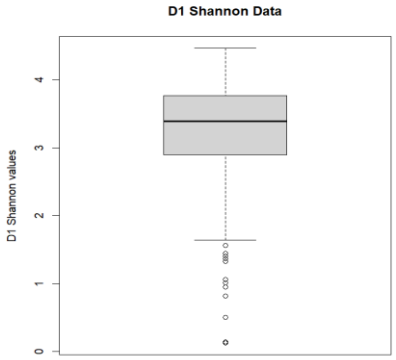
```
#####################################
#4.2.testing for homoscedasticity
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#Graphical visualizing using box plots (more reliable since our data has outliers)
#method 1: levene test & method 2: barlett test , we applied normality test on all numeric variables in g
library(car)
#Note: levene test as is robust to non-normality

#1- D1.Shannon.diversity
boxplot(AAD$D1.Shannon.diversity,ylab="D1 Shannon values",main="D1 Shannon Data") #

leveneTest(D1.Shannon.diversity ~ Antibiotic.class, data=AAD)  #p value 0.857 > 0.05 (homoscedastic)

bartlett.test(D1.Shannon.diversity ~ Antibiotic.class, data = AAD)  #p-value = 0.9746  > 0.05 so assume h
#From 2 methods D1.Shannon.diversity so we have enough evidence to accept assumption that it is homosceda
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#2- D6.Shannon.diversity
boxplot(AAD$D6.Shannon.diversity,ylab="D6 Shannon values",main="D6 Shannon Data")

leveneTest(D6.Shannon.diversity~Antibiotic.class,data=AAD) #p value= 0.4289 > 0.05 assume homoscedasticit

bartlett.test(D6.Shannon.diversity ~ Antibiotic.class, data = AAD)  #p-value = 0.1757 > 0.05 so we have e
#From 3 methods D6.Shannon.diversity so we have enough evidence to accept assumption that it is homosceda
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#3- D1.Chao1.diversity
boxplot(AAD$D1.Chao1.diversity,ylab="D1 Chao values",main="D1 Chao Data")

leveneTest(D1.Chao1.diversity~Antibiotic.class,data=AAD) # p value 0.9278 > 0.05 assume homoscedasticity

bartlett.test(D1.Chao1.diversity ~ Antibiotic.class, data = AAD) #p-value = 0.5621 > 0.05
#since p-value is greater than 0.05 so we fail to reject the assumption of homoscedasticity
#From 2 methods D1.Chao1.diversity so we have enough evidence to accept assumption that it is homoscedast
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#4- D6.Chao1.diversity

boxplot(AAD$D6.Chao1.diversity,ylab="D6 Chao values",main="D6 Chao Data")

leveneTest(D6.Chao1.diversity~Antibiotic.class,data=AAD) # p value 0.295 > 0.05 assume homoscedasticity
```

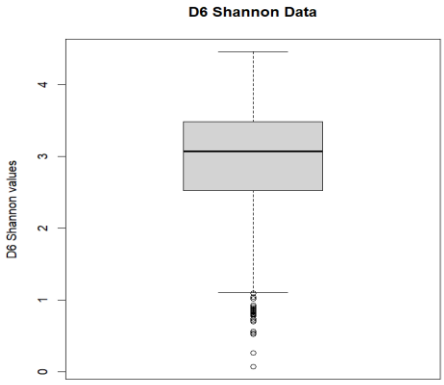
1. D1 Shannon diversity

Levene's test: p value 0.857 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)
Bartlett's test: p-value = 0.9746 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)

2. D6 Shannon diversity

Levene's test: p value= 0.4289 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)
Bartlett's test: p-value = 0.1757 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)





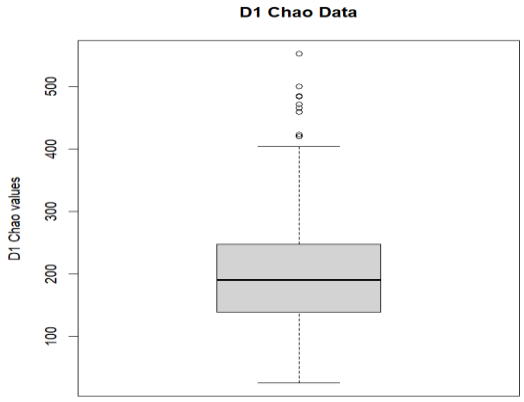
Since Levene's test and Bartlet are sensitive to outliers, we use boxplot to assume heteroscedasticity.

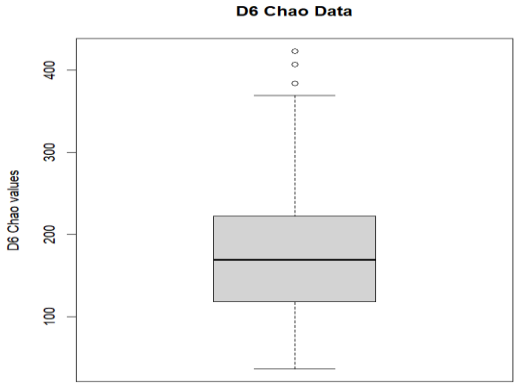
3. D1 Chao diversity

Levene's test: p value =0.9278 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)
Bartlett's test: p-value = 0.5621 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)

4. D6 Chao diversity

Levene's test: p value 0.295 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)
Bartlett's test: p-value = 0.206 > 0.05 (we fail to reject the null hypothesis, assuming homoscedasticity)

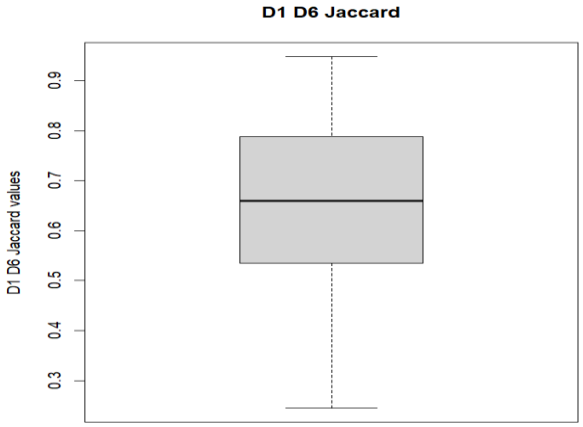




Since Levene's test and Bartlet are sensitive to outliers, we use boxplot to assume heteroscedasticity.

5. D1.D6 Jaccard Distance

Levene's test: p value 0.02891 < 0.05 (we can reject the null and assume heteroscedasticty)
Bartlett's test: p-value = 0.1448 > 0.05
From the boxplot and from levene's test we can assume heteroscedasticity as levene is more robust against non-normality than bartlett



Conclusion: All numeric data are assumed to be heteroscedastic.

5. STATISTICAL INFERENCE

To calculate the 90%, 95%, and 99% confidence intervals for the means of Jaccard distance for each antibiotic, we used the confint method using the below formula.

From the results, it is obvious that as the level increased from 0.90 to 0.99, the confidence intervals got much greater. So, the higher the confidence level, the wider the interval becomes, reflecting a higher level of confidence in capturing the true value.

```
antibiotics = unique(AAD$Antibiotic.class)
for (antibiotic in antibiotics){
  level_90 = confint(lm(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class == antibiotic] ~ 1, AAD), level=0.90)
  print(antibiotic)
  cat("level 0.90 of ", antibiotic, level_90,"\n")
}

for (antibiotic in antibiotics){
  level_95 = confint(lm(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class == antibiotic] ~ 1, AAD), level=0.95)
  print(antibiotic)
  cat("level 0.95 of ", antibiotic, level_95,"\n")
}

for (antibiotic in antibiotics){
  level_99 = confint(lm(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class == antibiotic] ~ 1, AAD), level=0.99)
  print(antibiotic)
  cat("level 0.99 of ", antibiotic, level_99,"\n")
}
```

- In 90% confidence interval

```
[1] "OBL"
level 0.90 of  OBL 0.650139 0.6943684
[1] "FQN"
level 0.90 of  FQN 0.6198198 0.6880488
[1] "PBL"
level 0.90 of  PBL 0.6207265 0.6633013
```

- In 95% confidence interval

```
[1] "OBL"
level 0.95 of  OBL 0.6458337 0.6986737
[1] "FQN"
level 0.95 of  FQN 0.6130703 0.6947983
[1] "PBL"
level 0.95 of  PBL 0.616605 0.6674228
```

```
[1] "OBL"
level 0.99 of  OBL 0.6373082 0.7071992
[1] "FQN"
level 0.99 of  FQN 0.5995273 0.7083413
[1] "PBL"
level 0.99 of  PBL 0.6084799 0.6755479
```

- **In terms of inferences** and observations related to interval width when requesting a higher confidence level:

Increased certainty: A higher confidence level provides a greater degree of certainty that the true parameter value falls within the calculated confidence interval. This is because a wider interval allows for a larger range of potential values to be captured with a higher level of confidence.

Reduced precision: However, the increased certainty associated with a higher confidence level comes at the cost of reduced precision. The wider interval encompasses a larger range of values, leading to a less precise estimate of the true parameter value. As a result, the interval width tends to be larger when requesting a higher confidence level.

Trade-off between precision and certainty: It's important to note that the choice of confidence level involves a trade-off between precision and certainty. A higher confidence level provides a higher level of confidence in capturing the true parameter value but sacrifices precision by yielding wider intervals. Conversely, a lower confidence level (e.g., 90%) results in narrower intervals, providing a more precise estimate, but with a lower level of confidence in capturing the true value.

6. HYPOTHESIS TESTING

6.1. We hypothesis that Chao/Shannon on day 6 is different between CDI vs ND. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework.

H0: There is no difference in Chao/Shannon day 6 between CDI and ND
HA: The Chao/Shannon day 6 is different between CDI and ND

```
#We hypothesis that Chao/Shannon at day 6 different between CDI vs ND

#null hypothesis: there is no difference between D6.chao/shannon between CDI and ND
#alternative: there is a difference between D6.chao/shannon between CDI and ND

#Assuming normality and homoscedasticity we use t-test
t.test(AAD$D6.Chao1.diversity[AAD$Outcome=="CDI"],AAD$D6.Chao1.diversity[AAD$Outcome=="ND"], var.equal= TRUE)
# p-value = 0.3362 (not significant so we fail to reject the null ) so we can say there is no difference
t.test(AAD$D6.Shannon.diversity[AAD$Outcome=="CDI"],AAD$D6.Shannon.diversity[AAD$Outcome=="ND"], var.equal= TRUE)
# p-value = 0.7164 (not significant so we fail to reject the null ) so we can say there is no difference


# assessing normality assumption
shapiro.test(AAD$D6.Chao1.diversity[AAD$Outcome=="CDI"])   #p-value = 0.04323 assumption of normality is rejected
shapiro.test(AAD$D6.Chao1.diversity[AAD$Outcome=="ND"])    #p-value = 1.831e-05 assumption of normality is rejected

shapiro.test(AAD$D6.Shannon.diversity[AAD$Outcome=="CDI"]) #p-value = 0.1407
shapiro.test(AAD$D6.Shannon.diversity[AAD$Outcome=="ND"])  #p-value = 6.086e-12


#assessing homoscedasticity
boxplot(AAD$D6.Chao1.diversity[AAD$Outcome=="CDI"],AAD$D6.Chao1.diversity[AAD$Outcome=="ND"],names = c("Chao6 CDI", "Chao6 ND"))     #variances differ
boxplot(AAD$D6.Shannon.diversity[AAD$Outcome=="CDI"],AAD$D6.Shannon.diversity[AAD$Outcome=="ND"],names = c("Shannon6 CDI", "Shannon6 ND")) #variances d


# since the two assumptions are not met we can use a non-parametric test
wilcox.test(AAD$D6.Chao1.diversity[AAD$Outcome=="CDI"],AAD$D6.Chao1.diversity[AAD$Outcome=="ND"])
```

Since we are assuming normality and homoscedasticity, we can use two sample t-test and set the var.equal to true:

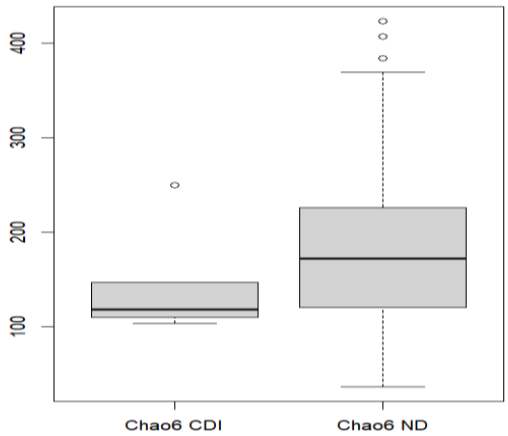
- The p-value of the t-test between Chao day6 CDI and ND groups = 0.3362 > 0.05 so we fail to reject the null and assume there's no difference between Chao day 6 CDI and ND groups.

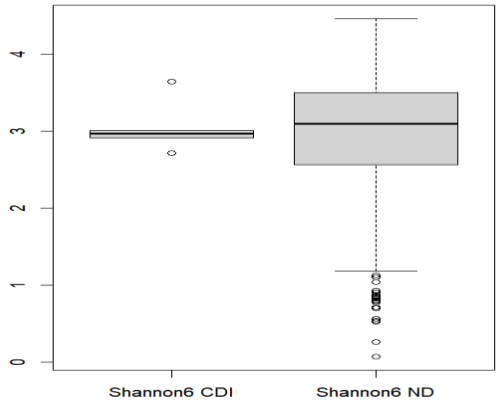
- The p-value of the t-test between Shannon day6 CDI and ND groups = 0.7164 > 0.05 so we fail to reject the null and assume there's no difference between Shannon day 6 CDI and ND groups.

When assessing the two previous assumptions:

- All the parameters are not normal except Shannon day6 group CDI. And the variances of the parameters are different as seen in the boxplot.

Since the two assumptions are not met, we can use a non-parametric test as Wilcox test.





6.2. We hypothesis that Jaccard distance "different" in the group receiving OBL Antibiotics compared to the FQN antibiotics. Can you test this hypothesis assuming heteroscedasticity.

H0: There is no difference in Jaccard distance in OBL and FQN antibiotics groups.
HA: The Jaccard distance differs in OBL and FQN antibiotics groups.

```
#2nd point: We hypothesis that Jacard distance "different" in the group receiving OBL Antibiotics compared to the FQN antibiotics

#We hypothesis that Jacard distance "different" in the group receiving OBL Antibiotics  compared to the FQN
#null hypothesis: there is no difference in the jaccard distance groups receiving obl and fqn
#alternative there is difference between the two groups

#using welch's t test assuming heteroscedasticity
t.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"],AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"],var.equal=FALSE )
#p-value = 0.4538 (greater than 0.05 so not significant)
#so we fail to reject the null

#assessing the assumption
var.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"],AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"])
# p-value = 0.4596 (not significant) so data is homoscedastic and assumption is not met

shapiro.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"]) #p-value = 0.05651
shapiro.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"]) #p-value = 0.573
```

Since we assume heteroscedasticity, we can use welch's t-test:

- The p-value is 0.4538>0.05 so we fail to reject the null hypothesis and assume that there is no difference between the two groups.

Assessing the assumption:

- Using var.test to check the variance, p value is 0.4596 > 0.05 so we assume homoscedasticity (equal variance), And previous assumption is not met.

6.3. We hypothesis that Jaccard 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).

H0: There is no difference in Jaccard Distance between the different antibiotics.
HA: There is atleast one difference between one group and another in antibiotics of Jaccard distance.

```
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#3rd point: We hypothesis that Jacard distance is different between the different Antibiotics

#H0: there is no difference in jaccard distance between different antibiotics
#HA: there is at least one antibiotic different than one other antibiotic


#assuming normality and homoscedasticity
ANOVAModel = aov(D1.D6.Jaccard.distance ~ Antibiotic.class,data=AAD)
summary(ANOVAModel)
coef(ANOVAModel)
#p value=0.287 not signifcant (so we fail to reject the null)

TukeyHSD(ANOVAModel) #there is no significant differenences between the groups
plot(TukeyHSD(ANOVAModel))

#assessing normality and homoscedasticity
shapiro.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"]) #p-value = 0.05651 not significant assume normality
qqnorm(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"])
qqline(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"])
hist(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="OBL"])
# we assume normality

shapiro.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="PBL"]) #p-value = 0.001216 we can't assume normality
qqnorm(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="PBL"])
qqline(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="PBL"])
hist(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="PBL"])

shapiro.test(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"]) #p-value = 0.573 assume normality
qqnorm(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"])
qqline(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"])
hist(AAD$D1.D6.Jaccard.distance[AAD$Antibiotic.class=="FQN"])
# since one group is not normal we can use a non-parametric test to be more sure
```

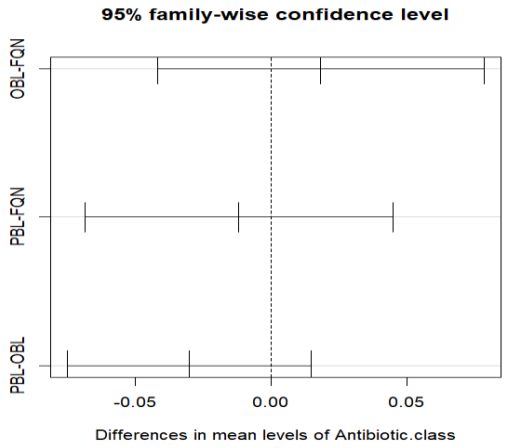
Assuming normality and homoscedasticity, we use parametric test for 3 samples ANOVA.

- The p-value=0.287 > 0.05 so we fail to reject the null hypothesis.
- Then we use post-hoc test Tukey to determine which specific groups are significantly different from each other but there is no significant difference between any of the groups.

Then we assess the assumptions:

- Two groups are normal (OBL and FQN), and one isn't normal so we can use a non-parametric test to make sure of our hypothesis using Kruskal test where its p-value = 0.3324.

```
> 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
> coef(ANOVAModel)
       (Intercept) Antibiotic.classOBL Antibiotic.classPBL
         0.65393430          0.01831942         -0.01192043
> TukeyHSD(ANOVAModel) #there is no significant differenences between the groups
  Tukey multiple comparisons of means
    95% family-wise confidence level

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

$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
```



7. LINEAR MODEL

In our Project We Did two types of linear Regression:

7.1. Simple linear Regression:

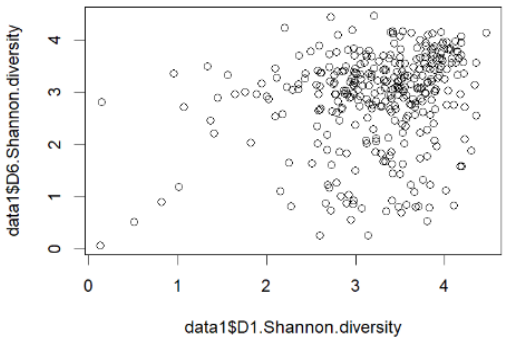
Were we plotted the data Between Day 1 Shannon as the independent variable on the X-Axis and Day 6 Shannon as the dependent variable on the Y-Axis and this is first model then We applied the same steps also when we plotted Day 1 Chao Against Day 6 Chao this is second model and compared between them to see which model is better for this data according to the value of R squared, pvalue, st.error and median.

```
# STEP 1: Draw a graph of the data to make sure the relationship between D1 & D6 Shannon.diversity make sense
plot(AAD$D1.Shannon.diversity, AAD$D6.Shannon.diversity)

# STEP 2: Do the regression
simple.regression1 <- lm(AAD$D6.Shannon.diversity ~ AAD$D1.Shannon.diversity , data=AAD)

# STEP 3: Look at the R^2, F-value and p-value
summary(simple.regression1)
```

Steps of Linear regression:

1. Plot the data to make sure that the relation we chose make sens

```
# STEP 1: Draw a graph of the data to make sure the relationship between D1 & D6 Shannon.diversi
plot(AAD$D1.Shannon.diversity, AAD$D6.Shannon.diversity)

# STEP 2: Do the regression
simple.regression1 <- lm(AAD$D6.Shannon.diversity ~ AAD$D1.Shannon.diversity , data=AAD)
```



- The plot scattered in random shape not linear relationship

2. We do the regression function then we look at the R-Squared F- Value and P-Value.

- Using summary method

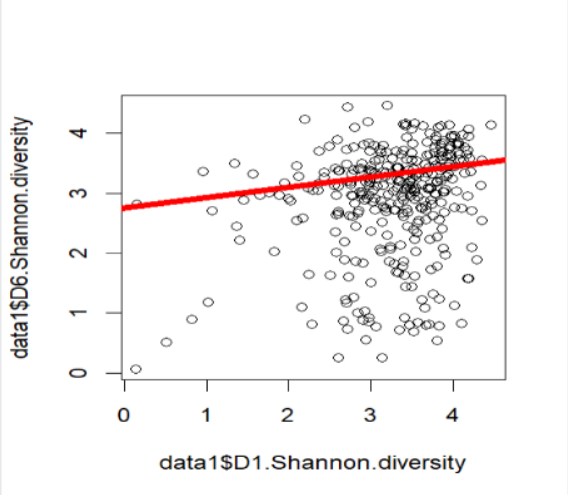
```
Residuals:
    Min      1Q  Median      3Q     Max
-3.0974 -0.3459  0.1170  0.4867  1.2125

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.76133    0.12403  22.264  < 2e-16 ***
data1$D6.Shannon.diversity 0.17040    0.04125   4.131 4.57e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.692 on 333 degrees of freedom
Multiple R-squared:  0.04875,   Adjusted R-squared:  0.0459
F-statistic: 17.07 on 1 and 333 DF,  p-value: 4.569e-05
```

- From the results the R^2 only 4.875% of the variance in D1.Shannon.diversity is explained by D6.Shannon.diversity. also the associated p-value is 4.569e-05 (very small). This indicates that the regression model as a whole is statistically significant, suggesting that D6.Shannon.diversity is a significant predictor of D1.Shannon.diversity.

- The reason that we look at these specific values as the R-squared value helps assess the goodness of fit of the model, while the F-value and P-value determine the overall significance and statistical significance of the model, respectively. These metrics are essential for evaluating the reliability and validity of the linear regression analysis and its results.

3. We plot the best fit Regression line for our graphs

```
plot(AAD$D1.Shannon.diversity, AAD$D6.Shannon.diversity)
abline(simple.regression1, lwd=5, col="red")
```



4. Interpret the regression coefficient

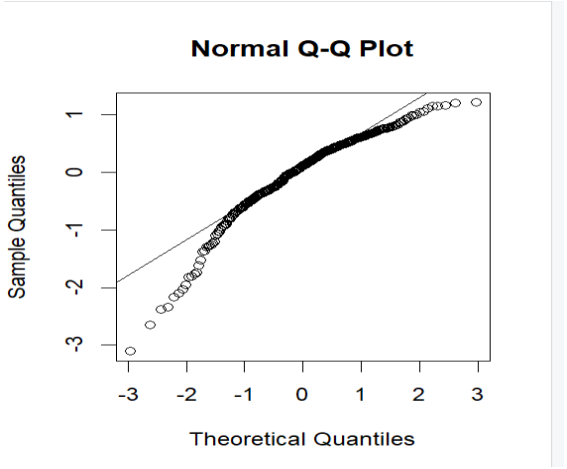
```
#Interpret the regression coefficient
intercept <- coef(simple.regression1)[1]
coefficient <- coef(simple.regression1)[2]

cat("Intercept:", intercept, "\n")
cat("Coefficient:", coefficient, "\n")
```

```
Intercept: 141.6345
> cat("Coefficient:", coefficient, "\n")
Coefficient: 0.339413
```

result

- We interpret the regression coefficients and assess the assumptions of linear regression, particularly the normality assumption. Then we use QQ plot helps visualize if the residuals deviate significantly from a normal distribution.

```
qqnorm(simple.regression1 $residuals)
qqline(simple.regression1 $residuals)
```



- From QQ plot the residuals deviate significantly from a normal distribution.

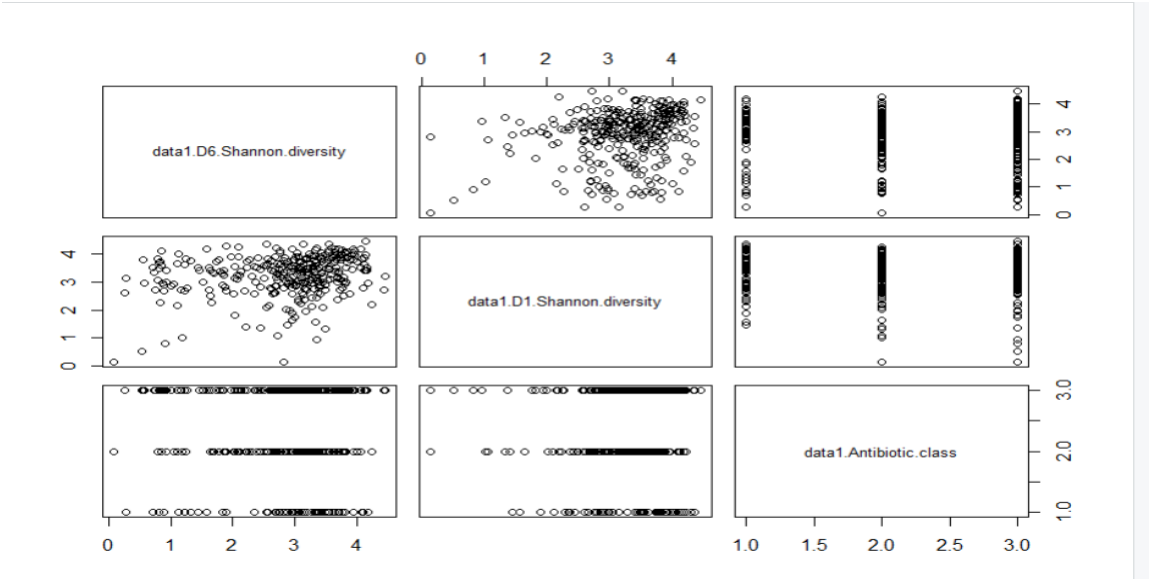
7.2.Second Multiple Linear Regression:

In the Multiple linear Regression, we plot Day 6 Shannon as the dependent variable on the Y-Axis and the multiple independent variables are the Antibiotics Class and Day 1 Shannon on the X-Axis

```
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#7.2. MULTIPLE regression: we used D6.Shannon.diversity with D1 & Antibiotic.class

plot(data.frame(AAD$D6.Shannon.diversity, AAD$D1.Shannon.diversity, AAD$Antibiotic.class))
multiple.regression <- lm(D6.Shannon.diversity ~ D1.Shannon.diversity + Antibiotic.class, data=AAD)
summary(multiple.regression)

if(!require(car)){
  install.packages("car")
}
# Load car package
library(car)

# Produce added variable plots
avPlots(multiple.regression)
```

Steps of Multiple linear regression:

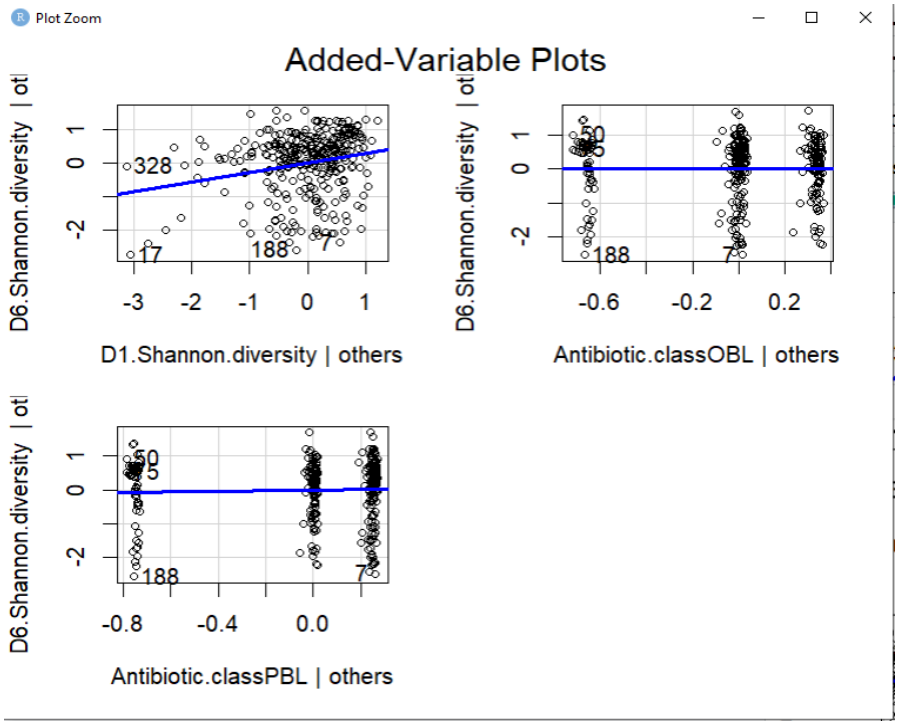
1- Plot the data to make sure that the relation we chose make sense.



2- We do the regression function and we look at the R-Squared F- Value and P-Value.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.5309 -0.3643  0.2218  0.6235  1.7193

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.898713   0.261053   7.273 2.56e-12 ***
D1.Shannon.diversity 0.285092   0.069556   4.099 5.23e-05 ***
Antibiotic.classOBL -0.003358   0.147593  -0.023    0.982
Antibiotic.classPBL  0.079302   0.138722   0.572    0.568
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8984 on 331 degrees of freedom
Multiple R-squared:  0.05073,   Adjusted R-squared:  0.04213
F-statistic: 5.897 on 3 and 331 DF,  p-value: 0.0006251
```

3-We plot the best fit Regression line for our graphs



7.3.Bonus

```
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ BONUS
#1st
confint(simple.regression1, level=0.95)
confint(simple.regression2, level=0.95)

#2nd
Estimate <- predict(lm(AAD$D1.D6.Jaccard.distance ~ AAD$Antibiotic.class), AAD)
Change <- Estimate[2] - Estimate[1]
```

- To get 95% confidence interval of the regression slope, we used confint function

  - In summary, the output provides the estimated values for the intercept and coefficient along with their corresponding 95% confidence intervals. These values help you understand the average response and the range of likely values for the intercept and coefficient in the regression model.

```
> confint(simple.regression1, level=0.95)
                                2.5 %    97.5 %
(Intercept)              1.4810576 2.3871115
AAD$D1.Shannon.diversity 0.1498715 0.4223336
> confint(simple.regression2, level=0.95)
                                2.5 %      97.5 %
(Intercept)            119.7385175 163.5304151
AAD$D6.Chao1.diversity   0.2241708   0.4546551
```

- To Estimate the average Jacard distance change for with changing the Antibiotics, we used the predict function to estimate value and get the change value by difference of estimate.

  - Since the value is negative, it indicates a decrease in the average Jacard distance when the Antibiotics value is changed from the first level to the second level.

```
> Change
          2
-0.01831942
>
```