**Family Project Data Analysis**

**BIO 199**

**By: Mildred Monsivais**

Table of Contents

[Project Overview 4](#_Toc58497878)

[Tools Used 4](#_Toc58497879)

[Python Libraries Used 4](#_Toc58497880)

[Goal 4](#_Toc58497881)

[Data Process Flow Chart 5](#_Toc58497882)

[Heatmap Flow Chart 5](#_Toc58497883)

[PCA Flow Chart 5](#_Toc58497884)

[Shannon H Flow Chart 6](#_Toc58497885)

[Heatmaps 7](#_Toc58497886)

[Level: All Families 7](#_Toc58497887)

[0 Values in the Data 7](#_Toc58497888)

[Level: Families 8](#_Toc58497889)

[Level: Family 1 Individuals 9](#_Toc58497890)

[Level: Family 2 Individuals 10](#_Toc58497891)

[Level: Family 3 Individuals 11](#_Toc58497892)

[Level: Family 4 Individuals 12](#_Toc58497893)

[Principal Component Analysis 13](#_Toc58497894)

[Why use PCA 13](#_Toc58497895)

[What are we looking for in the PCA Charts 13](#_Toc58497896)

[PCA Plot Interpretation 13](#_Toc58497897)

[Level: All Families 13](#_Toc58497898)

[Level: Families 14](#_Toc58497899)

[Family 1 14](#_Toc58497900)

[Family 2 14](#_Toc58497901)

[Family 1 Values: k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_ 14](#_Toc58497902)

[Family 2 Values: k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_ 14](#_Toc58497903)

[Family 3 15](#_Toc58497904)

[Family 4 15](#_Toc58497905)

[Family 3 Values: k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus 15](#_Toc58497906)

[Family 4 Values: k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus 15](#_Toc58497907)

[Important Indexes Plot Based on PCA Outliers 16](#_Toc58497908)

[Plotted Values 16](#_Toc58497909)

[Shannon H Values 17](#_Toc58497910)

# **Project Overview**

## **Tools Used**

Python (Analysis)

Draw.io (Flow Charts)

Microsoft Word (Report)

## **Python Libraries Used**

1. pandas == 1.1.3
2. ecopy == 0.1.2.2
3. plotly.graph\_objects == 4.13.0
4. scipy == 1.2.1

## **Goal**

To interpret the family data set by applying exploratory data analysis to summarize interesting features using Python in the dataset.

# **Data Process Flow Chart**

## **Heatmap Flow Chart**

**Diagram

Description automatically generated**

## **PCA Flow Chart**

**Diagram

Description automatically generated**

## **Shannon H Flow Chart**

**Diagram

Description automatically generated**

# **Heatmaps**

## **Level: All Families**

A screenshot of a computer

Description automatically generated

## **0 Values in the Data**

In this heatmap we can see that there is a lot of values that are blue and on the color scale blue equates to 0 values. With this in mind, I counted how many individual values are 0’s and I found that 5576 out of 9504 values are 0. This is something that should be looked at because 58.67%(5576/99504) of the values in the data are 0’s. I believe this should be looked at more closely because this is something that is odd and could be an error with how the data was preprocessed.

## **Level: Families**

A picture containing text, screenshot, electronics, display

Description automatically generatedA picture containing text, electronics, display, screenshot

Description automatically generated

A picture containing text, electronics, screenshot

Description automatically generatedA picture containing text, electronics, screenshot, display

Description automatically generated

## **Level: Family 1 Individuals**

**Chart

Description automatically generatedGraphical user interface, chart

Description automatically generated**

**Chart, bar chart

Description automatically generatedGraphical user interface

Description automatically generated**

## **Level: Family 2 Individuals**

**A picture containing text, screenshot, electronics, display

Description automatically generatedGraphical user interface, chart

Description automatically generated**

**Chart

Description automatically generatedGraphical user interface, chart

Description automatically generated**

## **Level: Family 3 Individuals**

**Graphical user interface, chart

Description automatically generatedGraphical user interface, chart

Description automatically generated**

**Chart, bar chart

Description automatically generatedGraphical user interface, chart

Description automatically generated**

## **Level: Family 4 Individuals**

**Graphical user interface, chart

Description automatically generatedChart

Description automatically generated**

**A picture containing text, screenshot, electronics, display

Description automatically generatedA picture containing text, screenshot, electronics, display

Description automatically generated**

# **Principal Component Analysis**

**Why use PCA**  
I have 96 columns of data, and you can't visualize the interaction between the 96 variables at once. We can do this in pairs; however we will not see how all variables interact. Therefore, we will use PCA to reduce the 96 dimensions down to 2 dimensions of Principal Components measuring variability captured between the 2 components.

## **What are we looking for in the PCA Charts**

We are looking for any values that look like outliers. The outlier values that are away from the cluster of values (In this case names of the Taxon) in the PCA charts are the most important values, due to their distances. With this in mind, I decided to look at the outliers in each PCA graph and saw that the outlier values for the Family 1 chart is the same as the Family 2 chart. Also, the Family 3 chart and Family 4 chart have the same outlier values that I chose. These values are highlighted in the PCA charts.

## **PCA Plot Interpretation**

The values that are highlighted (with the ovals) in the following plots are considered outliers in the PCA plots, therefore these outliers mean that they of high significance. So I looked at the index(row) of these outliers and found that Family 1 and Family 2 have similarities with the k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_ taxon and Family 3 and Family 4 have similarities with the k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus taxon.

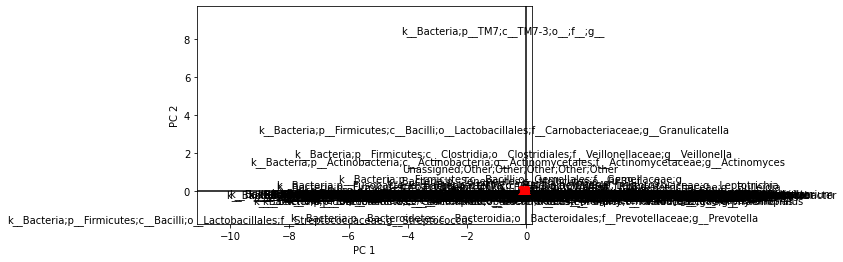
## **Level: All Families**

**Graphical user interface, application, Teams

Description automatically generated**

## **Level: Families**

## **Family 1**

****

## **Family 2**

**Graphical user interface, application

Description automatically generated**

## **Family 1 Values: k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_**

****

## **Family 2 Values: k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_**

****

## **Family 3**

**Graphical user interface, application, Teams

Description automatically generated**

## **Family 4**

**Graphical user interface, application

Description automatically generated**

## **Family 3 Values: k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus**

****

## **Family 4 Values: k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus**

****

## **Important Indexes Plot Based on PCA Outliers**

**Chart, histogram

Description automatically generated**

## **Plotted Values**

Family 1 PCA Outlier (k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_)

Family 2 PCA Outlier (k\_\_Bacteria;p\_\_TM7;c\_\_TM7-3;o\_\_;f\_\_;g\_\_)

Family 3 (k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus)

Family 4 (k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Streptococcus)

Again, these values are the outliers we found in the PCA plots above. As we can see, the values between Family 1and Family 2 are very close and the values between Family 3 and Family 4 are very close.

# **Shannon H Values**

|  |  |
| --- | --- |
| **Category** | **Value** |
| D Alpha | 5.945649222183524 |
| D Beta | 1.3877187447376873 |
| D Gamma | 8.250888875259127 |