Step 1: Extracted Meta data and the cluster of Species from “EMPmetadata_animal.csv” file.

Step 2: Data Clean:

1. Some cells were out of data. I filled them with value 0.0.
2. I separated the qualitative Meta data and transformed them into Quantitative data. E.g. For Biomass, I replaced the data [large, medium, small] with [1,2,3].

Step 3: I wanted to find out which Meta Data are prominent for predicting the cluster of the Species. For that, I used the Model : “Anova (Analysis of Variance)” . This model calculates the “F Score” and based on that we can select the best Meta Data for the cluster analysis. Higher the F score, more statistically significant to reject Null Hypothesis. So, we can use those Meta Data which have higher F score for clustering. For more details about ANOVA and F score (F value), you can visit this site: https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476

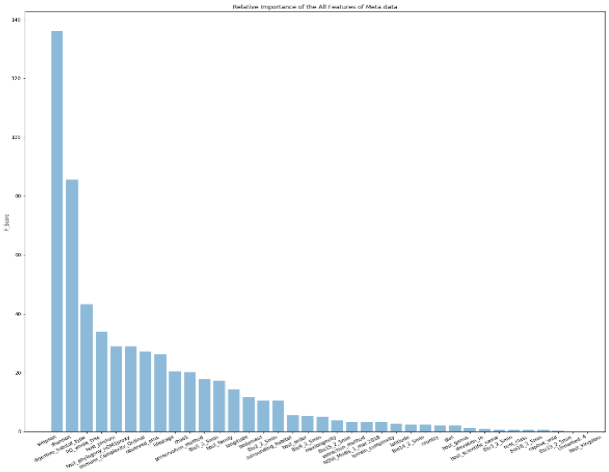
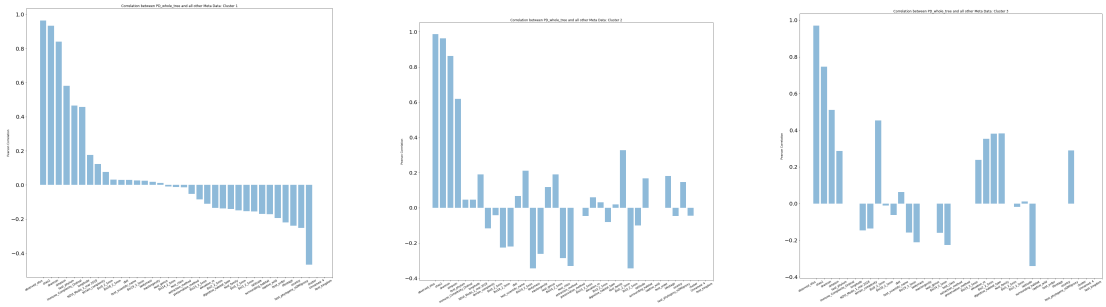


Figure 1: F score of all Meta Data

Step 4: I wanted to view the correlation between the “PD_Whole_Tree” and the rest of the Meta Data across the Clusters. This can give us insight about how different those clusters are from each other. It also gives us ideas for each cluster, which Meta Data can best describe the alpha diversity of “PD_whole_tree”. I used Pearson Correlation Model.



(a)Cluster 1 (b) Cluster 2 (c) Cluster 3

Figure 2: The Correlation between “PD_whole_tree” and rest of the Meta Data across the Clusters