Mini-Project 2 Checkpoint 2

ECE/CS 498DS
Spring 2020

Anirudh Sharma (ashar29), Anunay (anunays2), Badri (br17)

Task 1 - Question 0

1. **Why do biologists need multiple samples to identify microbes with significantly altered abundance?**

We encourage the process of taking multiple sample sizes so that we can say with more confidence that these results confirm to the total population (Like mean and variance of the population).
Drawing multiple sample sizes give more reliable results about the abundance of microbes with greater precision and confidence.

This is the statistical reason why biologists need multiple samples to identify microbes with significantly al tered abundance

2. **Number of samples analyzed:** 764

3. **Number of microbes identified**: 149

Task 1 – Question 1

A. Factorization of joint probability distribution:

- P(S, C, CM, L, Q) = P(Q|L,C) x P(L) x P(C|CM,S) x P(S) X P(CM)

B. Number of parameters needed to define conditional probability distribution: 11

C. Conditional probability tables:

S No.	Contamination	Lab Time	Quality	Count	Probability
1	High	Long	Bad	57	0.97
2	High	Short	Bad	16	0.06
3	Low	Long	Bad	78	0.08
4	Low	Short	Bad	160	0.04
5	High	Long	Good	2	0.03
6	High	Short	Good	233	0.94
7	Low	Long	Good	885	0.92
8	Low	Short	Good	3569	0.96

S No.	Storage Temperature	Collection Method	Contamination	Count	Probability
1	Cold	Nurse	High	178	0.04
2	Cold	Patient	High	34	0.08
3	Cool	Nurse	High	39	0.09
4	Cool	Patient	High	57	0.84
5	Cold	Nurse	Low	3869	0.96
6	Cold	Patient	Low	410	0.92
7	Cool	Nurse	Low	402	0.91
8	Cool	Patient	Low	11	0.16

Task 1 – Question 1(continued)

S No.	Storage Temperature	Count	Probability
1	Cold	4491	0.898
2	Cool	509	0.102

S No.	Coll	Count	Probability
1	Nurse	4488	0.898
2	Patient	512	0.102

S No.	Labtime	Count	Probability
1	Long	1022	0.204
2	Short	3978	0.796

Task 1 – Question 1 (continued)

D. Table of P(Quality | Storage Temp, Collection Method, Lab Time):

S No.	Storage Temperature	Collection Method	Lab Time	P(Bad)	P(Good)
1	Cold	Nurse	Long	0.11	0.89
2	Cold	Nurse	Short	0.04	0.96
3	Cold	Patient	Long	0.14	0.86
4	Cold	Patient	Short	0.06	0.94
5	Cool	Nurse	Long	0.18	0.82
6	Cool	Nurse	Short	0.03	0.97
7	Cool	Patient	Long	0.88	0.12
8	Cool	Patient	Short	0.04	0.96

E) Total number of samples dropped: 65 samples were dropped. (Due to the combination in the 7th row)

Initial sample count: 764, Final sample count: 699

Note : The red one highlighted in the table are the bad quality samples

Task 1 – Question 2

1. The data provided is accurate as the sum of relative abundance of microbes in each sample is 1 (With an error rate of 10^{-10} as mentioned in piazza)

2. **What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?**

Benefits

- Relative abundance data helps us to understand which microbe is present in highest/lowest proportion in the stool. Since we collected 764 samples, we can be quite confident about it. It also helps to compare and analyze the proportion of different microbes across various samples.
- It can also let us know if the microbes are present in almost same proportions across different samples (for different patients)

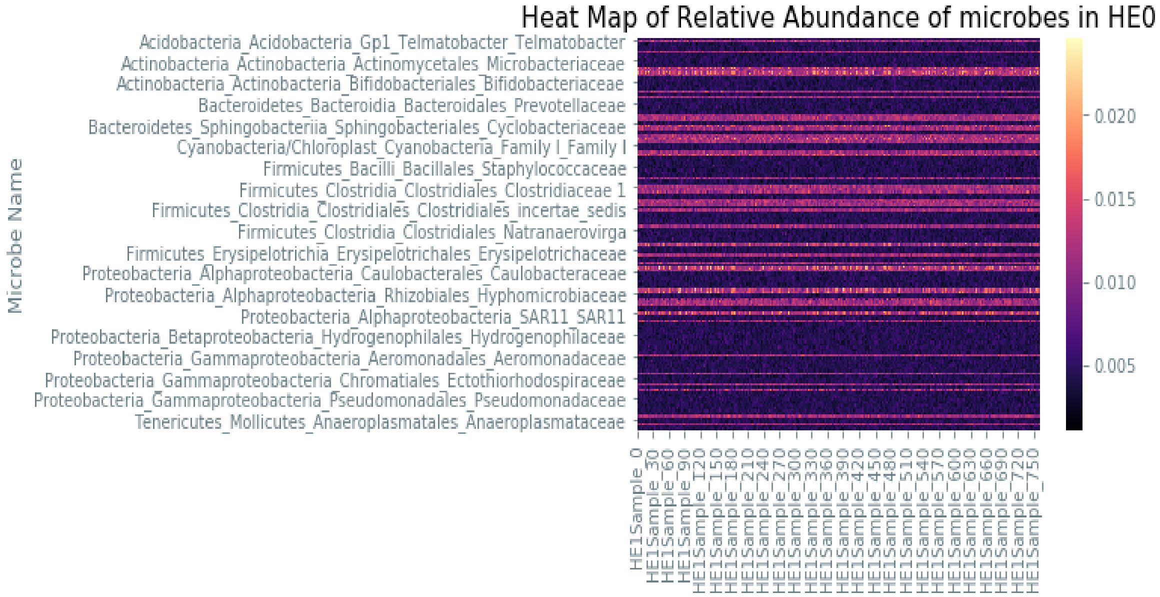
Drawbacks

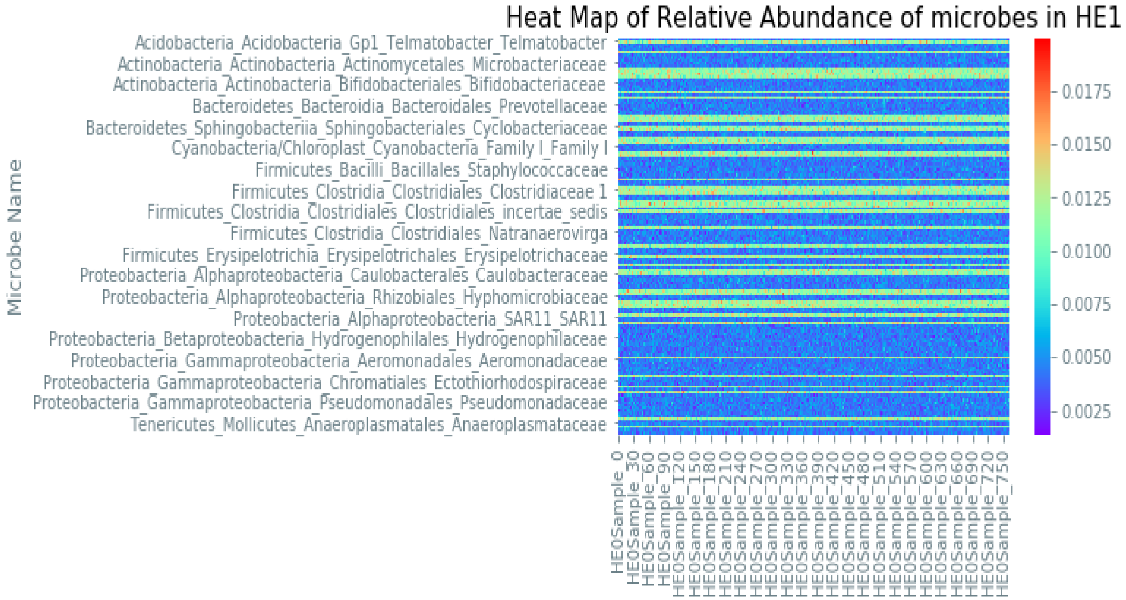
- We wouldn't be aware of the magnitude/amount of microbes present as that can vary across samples.

No, we won't lose any information as the range and scale of concentration of different microbes in the data is similar.

Task 1 – Question 3

- **Heatmaps (HE0 on left HE1 on right):**





Summarize your observations

- The concentration of microbes across different samples seems to be more or less consistent.
- The relative abundance of some samples like Hyphomicrobeaceae shows slight variation (pink and yellow) across different samples and their abundance is higher. While the samples like Tenericutes(Last line) have very low abundance variation throughout all the samples.

Task 1 – Question 3 (continued)

Q. Which aspects of the data are the heatmaps good at highlighting? What types of things are heatmaps less suitable for?

Heatmaps are less suitable for data which shows no variation across different categories because they leverage the idea of color coding to show distinct elements.

Consider the revenue of a supermarket giant which has a fixed number of customers visiting everyday. Displaying this in a heatmap serves no purpose as it doesn't highlight any change without any variation.

Task 2 – Question 1

B. What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.

Null Hypothesis: The given microbe is unaltered in both the population (HE0 and HE1) (Both the samples of microbes come from the same distribution)

Alternate Hypothesis: The given microbe is altered in both the population (The samples of microbes doesn't come from the same distribution)

For example if we consider the microbe -Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter it has a p-value of 0.18 (approximately) which is greater than the significance level, so we fail to reject the null hypothesis.

Conclusion : Microbes abundance is not altered across this sample

C. Count the number of microbes with significantly altered expression at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Summarize your answers in a table below:

S No.	Alpha	Altered Count
0	0.100	48
1	0.050	36
2	0.010	27
3	0.005	24
4	0.001	20

Task 2 – Question 2

A. What does a p-value of 0.05 represent in our context?

The p-value is the probability of obtaining results as extreme as the observed results of a hypothesis test, assuming that the null hypothesis is correct. Typically a p-value of less than or equal to .05 indicates strong evidence against the null hypothesis, thus we can reject the null hypothesis. In the given problem's context, p-value of 0.05 would mean that the Null Hypothesis: The given microbe is unaltered in both the population (HE0 and HE1) is not true. There is a 5% chance that the microbe content will be unaltered in both the population.

Alternatively, we can also say that the microbe content is altered in both the populations with 95% confidence.

B. If the null hypothesis is true, what distribution will the p-values follow?

When the null hypothesis is true and the underlying random variable is continuous, then the probability distribution of the p-value is uniform in the interval [0,1]

Under the null hypothesis, your test statistic T has the distribution $F(t)$ (e.g., standard normal). We show that the p-value $P=F(T)$ has a probability distribution $Pr(P<p)=Pr(F-1(P)<F-1(p))=Pr(T<t)\equiv p$; in other words, P is distributed uniformly. This holds so long as $F(\cdot)$ is invertible, a necessary condition of which is that T is not a discrete random variable.

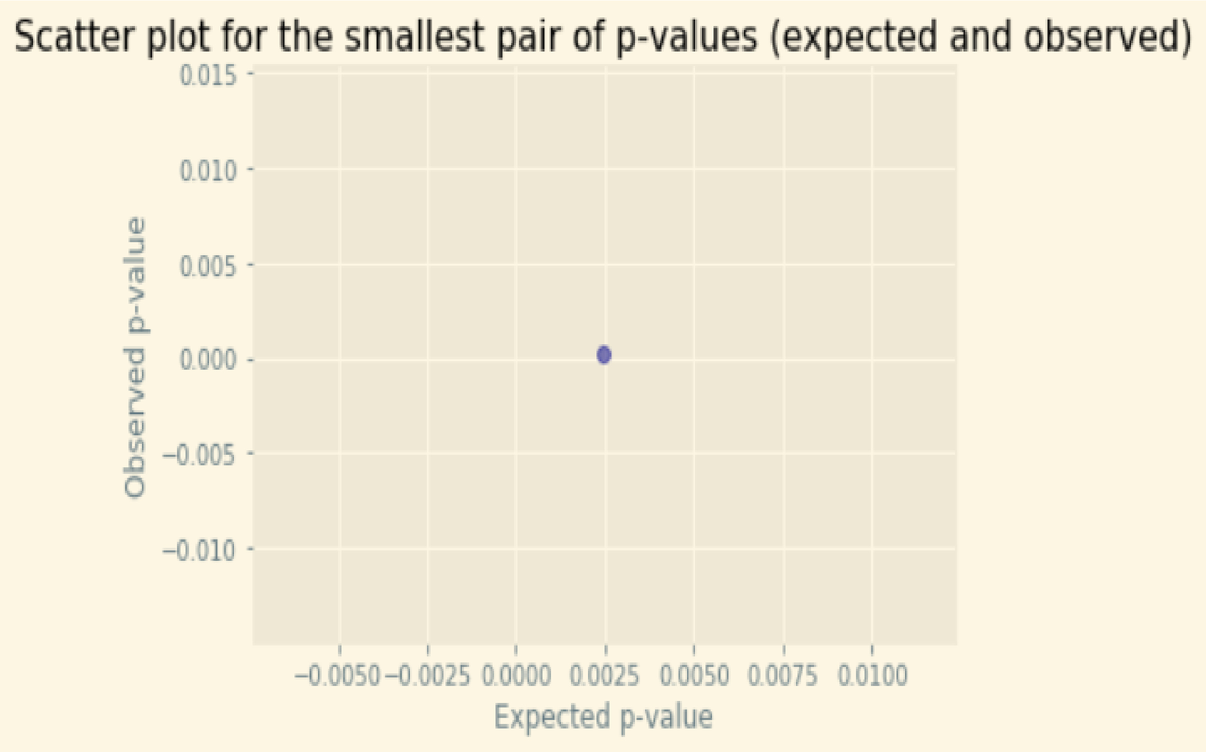
C. If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:

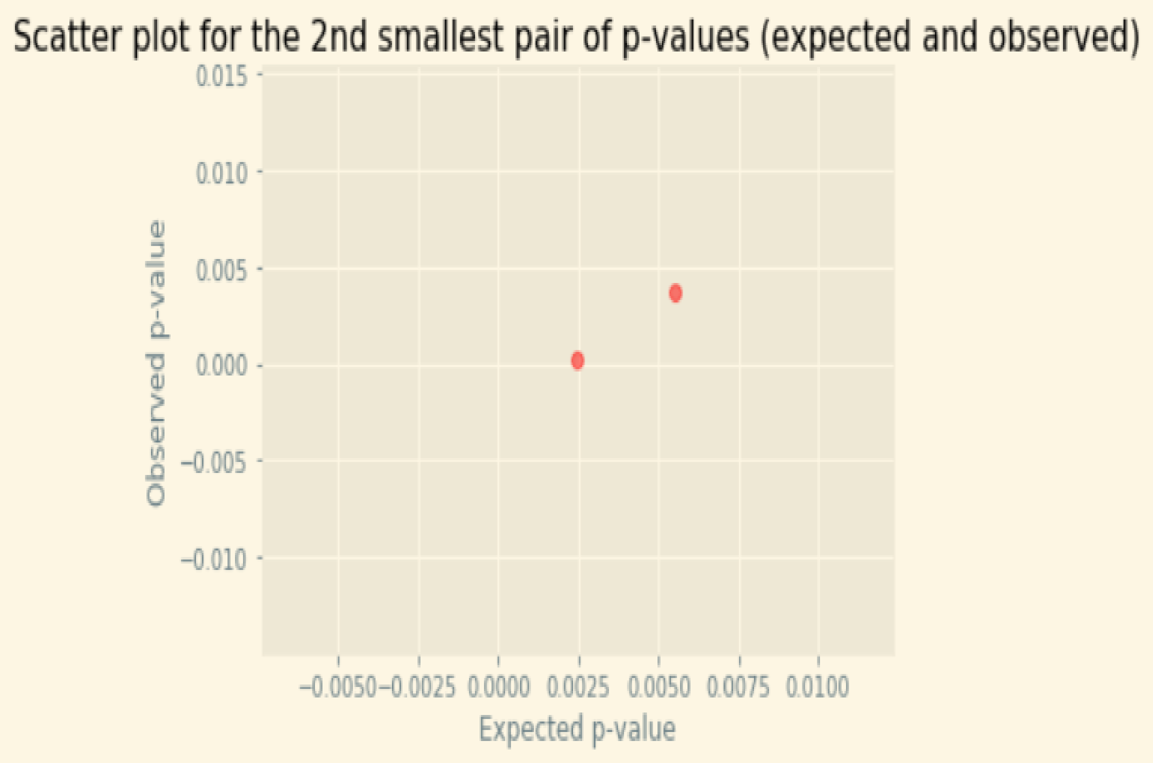
S No.	Alpha (α)	Expected Altered Count (Rounded value)	Observed Altered Count
1	0.100	15	48
2	0.050	7	36
3	0.010	1	27
4	0.005	1	24
5	0.001	0	20

The expected altered count would be a threshold value with observed altered count generally being equal to or greater than the expected count

Task 2 – Question 2 (continued)

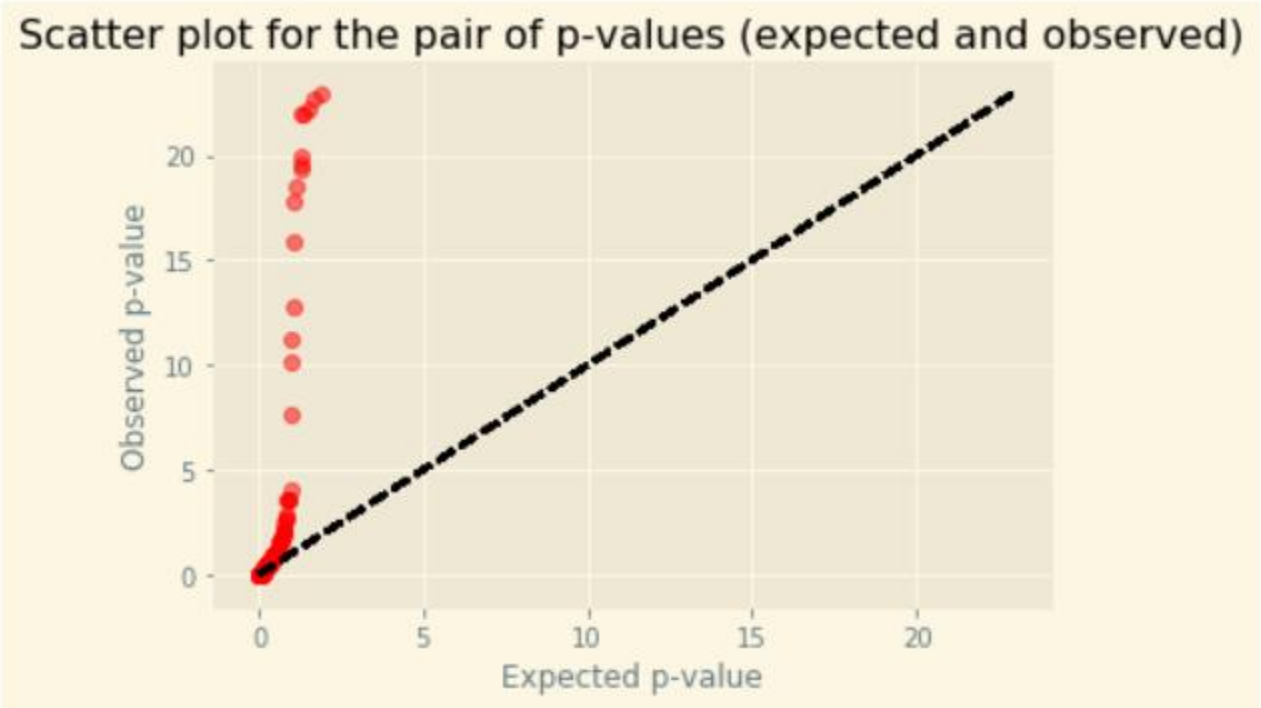
D. Q-Q plot:





Task 2 – Question 2 (continued)

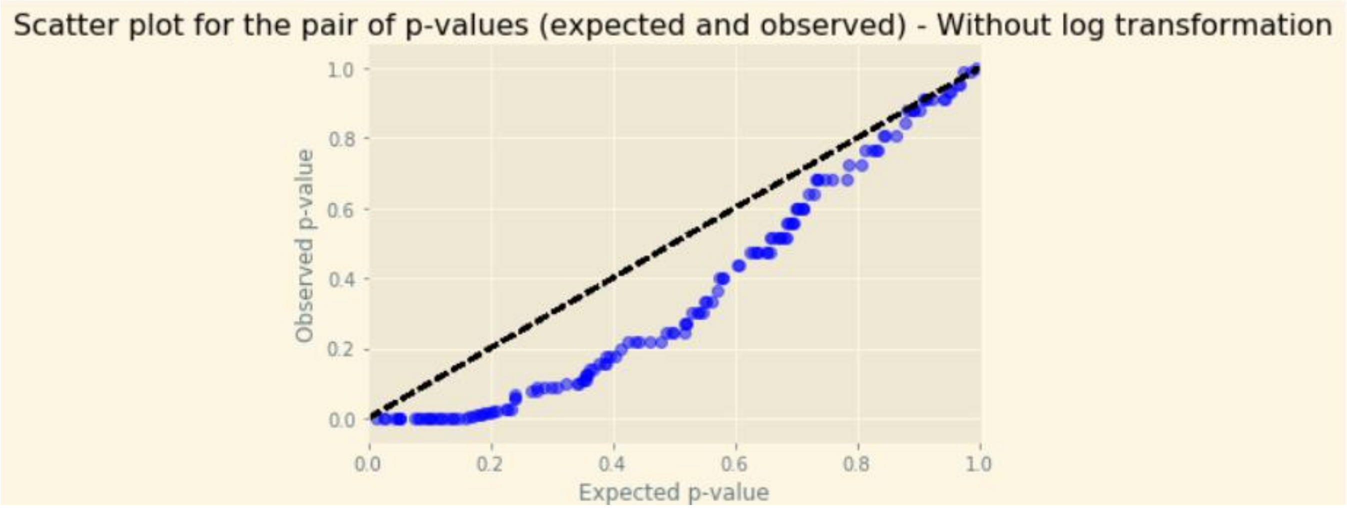
D. Q-Q plot:



Task 2 – Question 2 (continued)

E.(i). **How does taking the -log10() of the p-values help you visualize the p-value distribution?**

- Taking –log10() of the p-values will help normalize the data and helps us to expand the scale of the values a bit so that it is easier to draw insights from the visual representation (in this case a Q-Q plot)

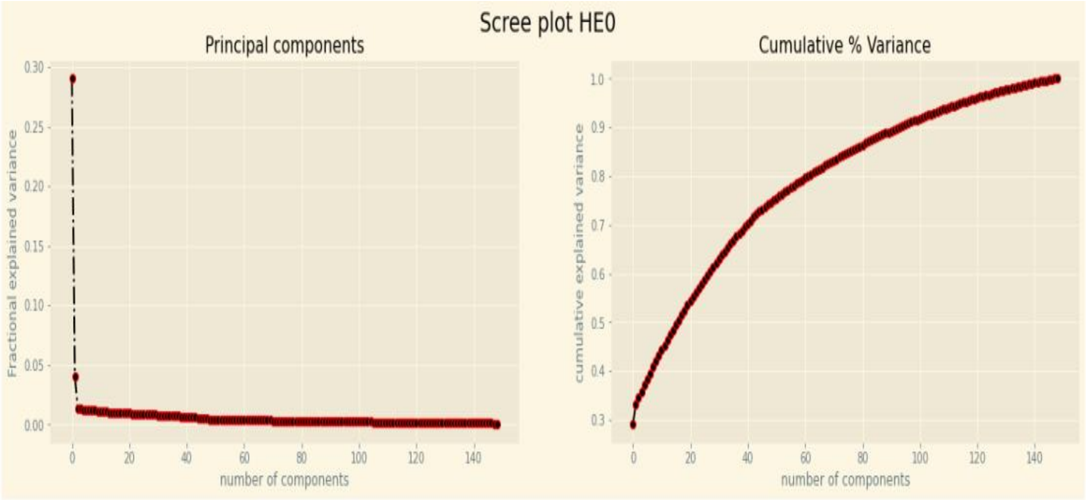


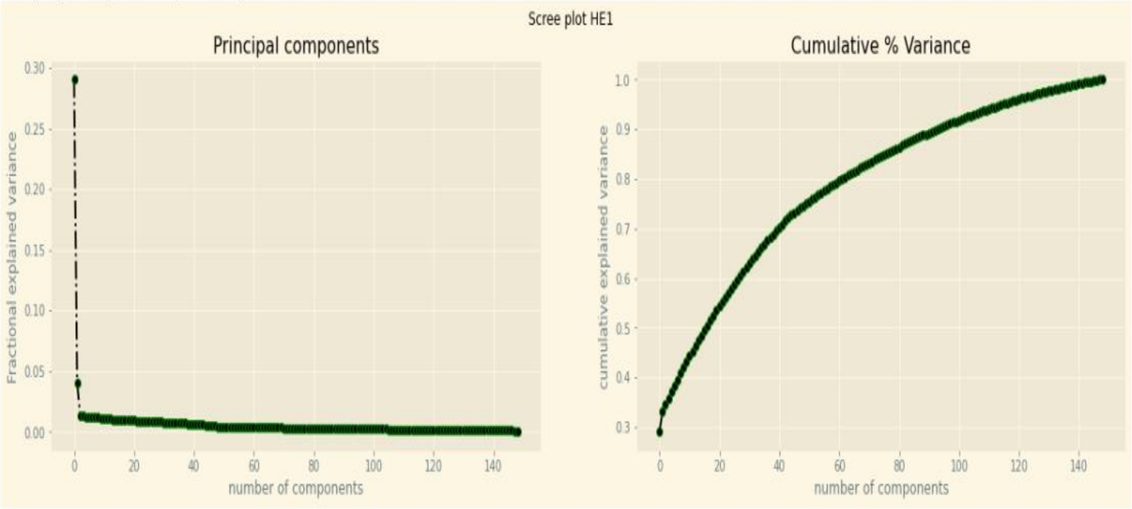
E.(ii). **What can you conclude from the Q-Q plot?**

The QQ plot shows that the points do not align along the x=y line. It indicates that the data sets come from different distributions. In the conte xt of the problem, the null hypothesis does not hold true and the plot suggests that the microbe content is altered in both the populations

Task 3 – Question 1

- b. Scree plots:



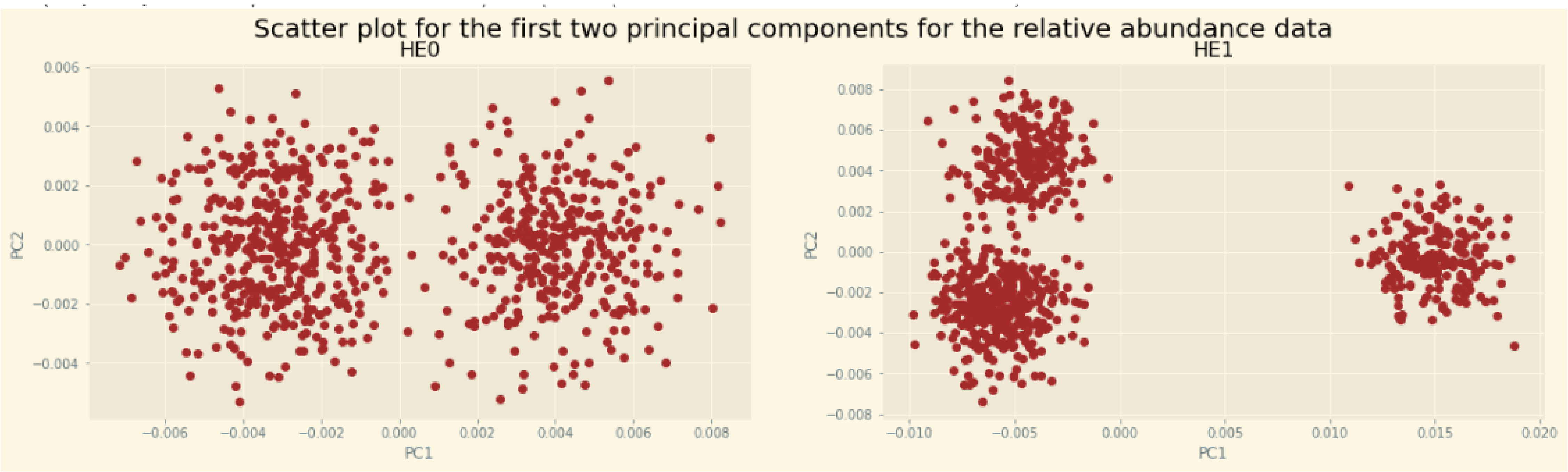


- Number of principal components needed to explain 30% of the total variance (HE0 and HE1):

  HE0 : The number of components to explain 30% of the variance are **16**

  HE1 **:** The number of components to explain 30% of the variance are **2**

Task 3 – Question 1 (continued)

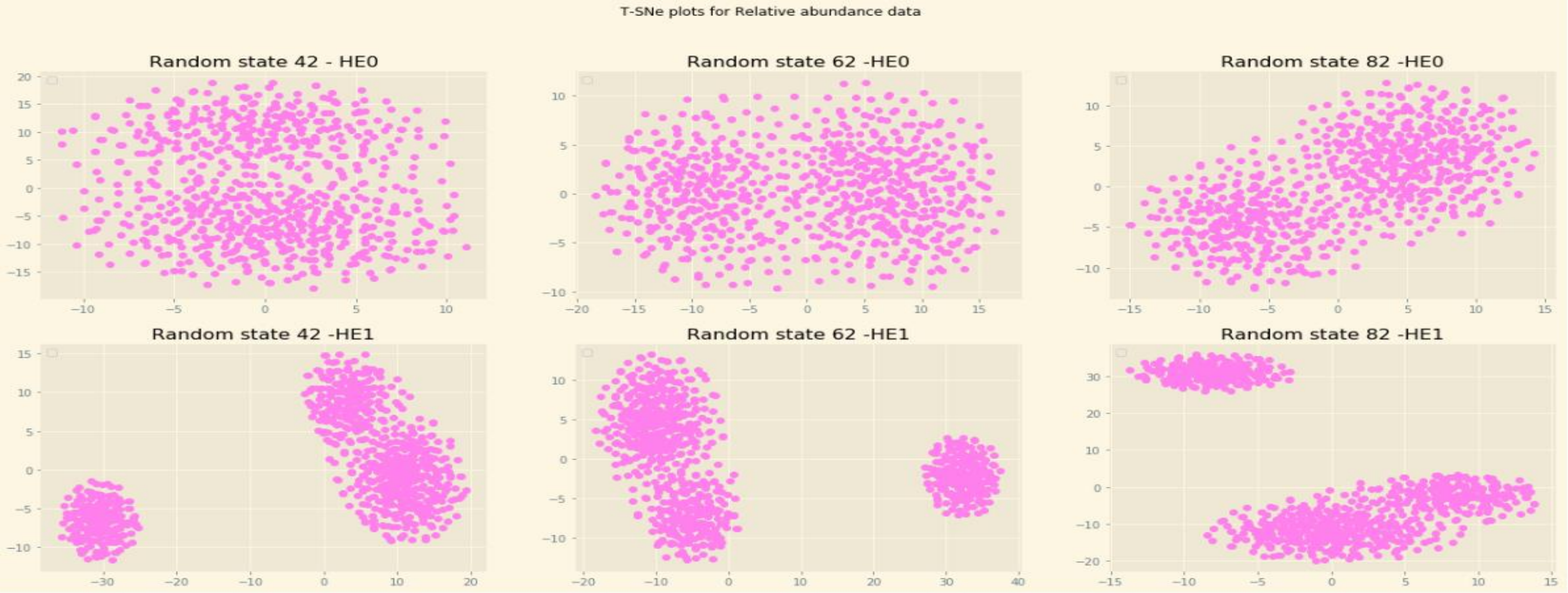


- Observations:

From the plot, we could clearly see that there exists 2 distinct clusters in RH0 and 3 distinct clusters in RH1. The 2 principal components have been able to capture around **9%** and **33%** of variance. That is why we are able to see more clear clusters in the RH1 data.

The intra-cluster distance between the samples in HE0 is more and the inter-cluster is less. Conversely, for HE1 samples, inter-cluster distance is more and intra-cluster distance between samples is less.

Task 3 – Question 2



Observations:

- 1) For the same population, the orientation of the clusters depend on the initial random seed. For different seed value, we can see different positioning of the clusters in the d-dimensional space.
- 2) Clusters are clearly distinguishable (far apart) in HE1, whereas the cluster boundary is not well defined for HE0 population.

Task 3 – Question 2 (continued)

- d. Discussion of similarities and differences between PCA and t-SNE results:

Similarity :

1) Both the algorithms are used for dimensionality reduction of bigger size datasets

2) Both are unsupervised learning techniques which can be used for data exploration and visualization.

Differences :

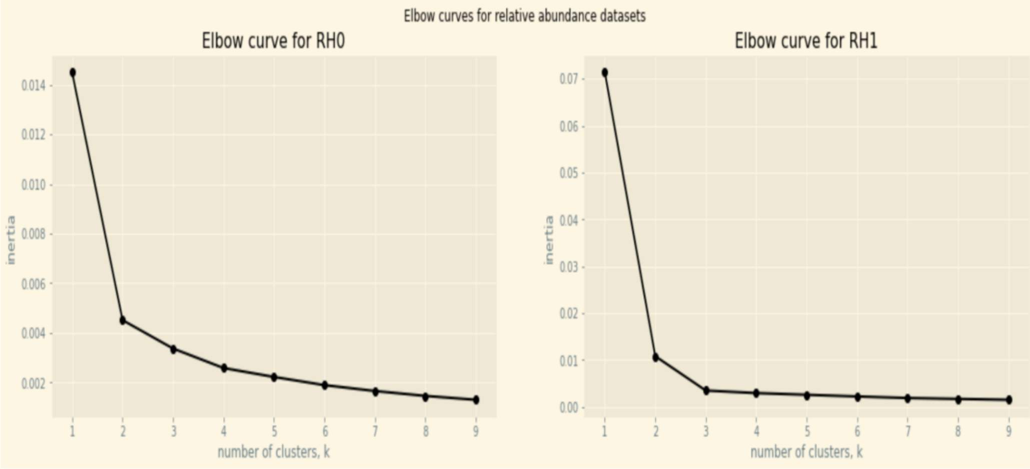
1) PCA uses linear feature extraction technique to find the components with the largest variance

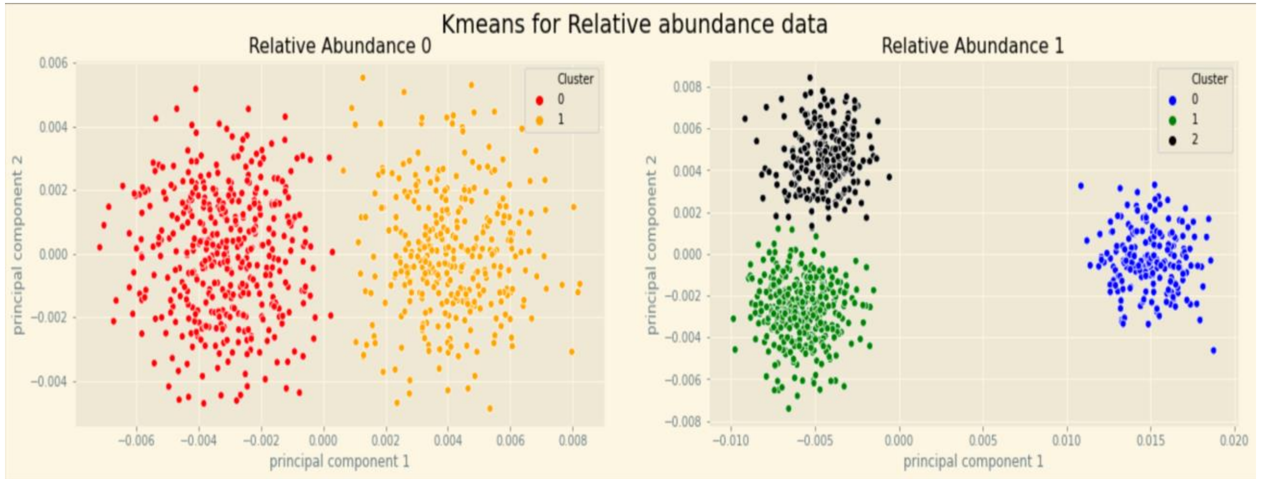
2)t-SNE uses non linear feature reduction from high dimensional space to low dimensional space

3) PCA is a mathematical approach and tries to separate points as far as possible based on highest variance while TSNE is a probabilistic approach and tries to group points as close as possible based on probability that two close points came from the same population distribution

Task 3 – Question 3

- a. K-means:





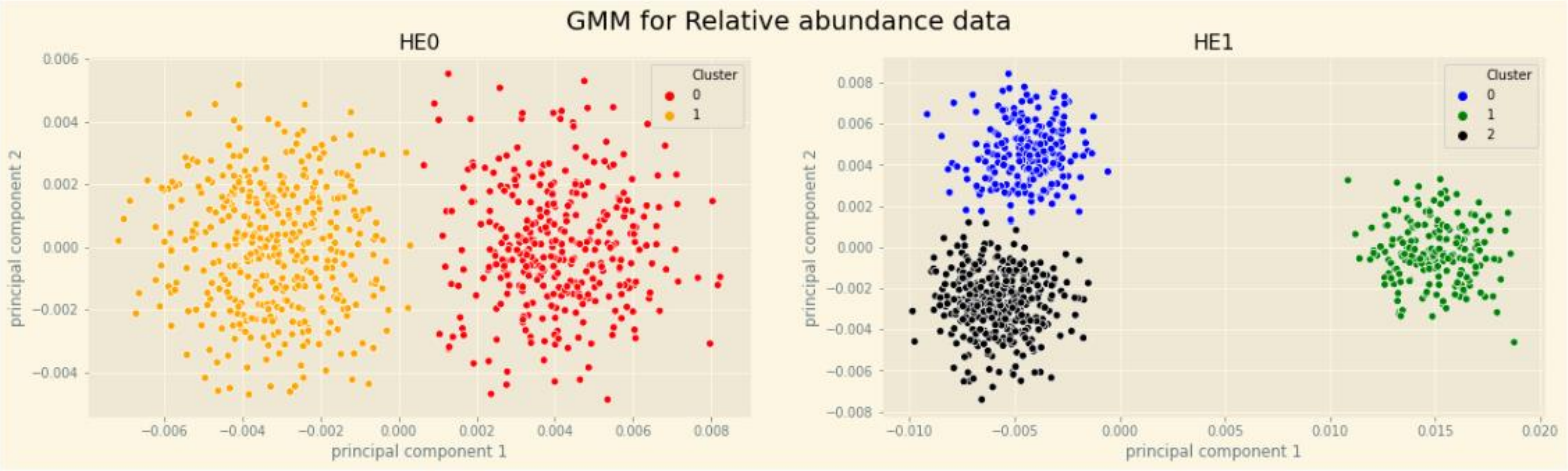
Visually , we saw that relative abundance 0 had 2 clusters and relative abundance 1 had 3 clusters .

Verification of the visual inspection for clusters can be seen through the elbow curve on the left :

The elbow curve which tells us the Within sum of squares (WSS) for different values of K. The point at which the curve shows a sudden decrease in the WSS is the elbow point which can be considered as the optimum number of K From above we could see that K=2 and 3 respectively.

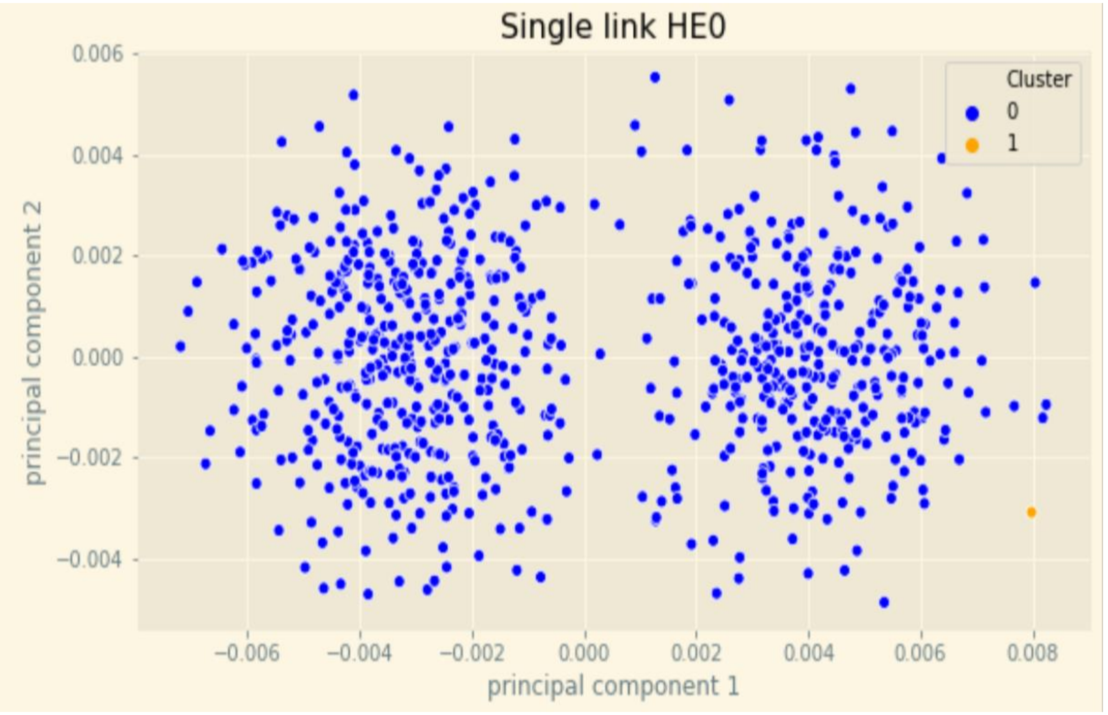
Task 3 – Question 3 (continued)

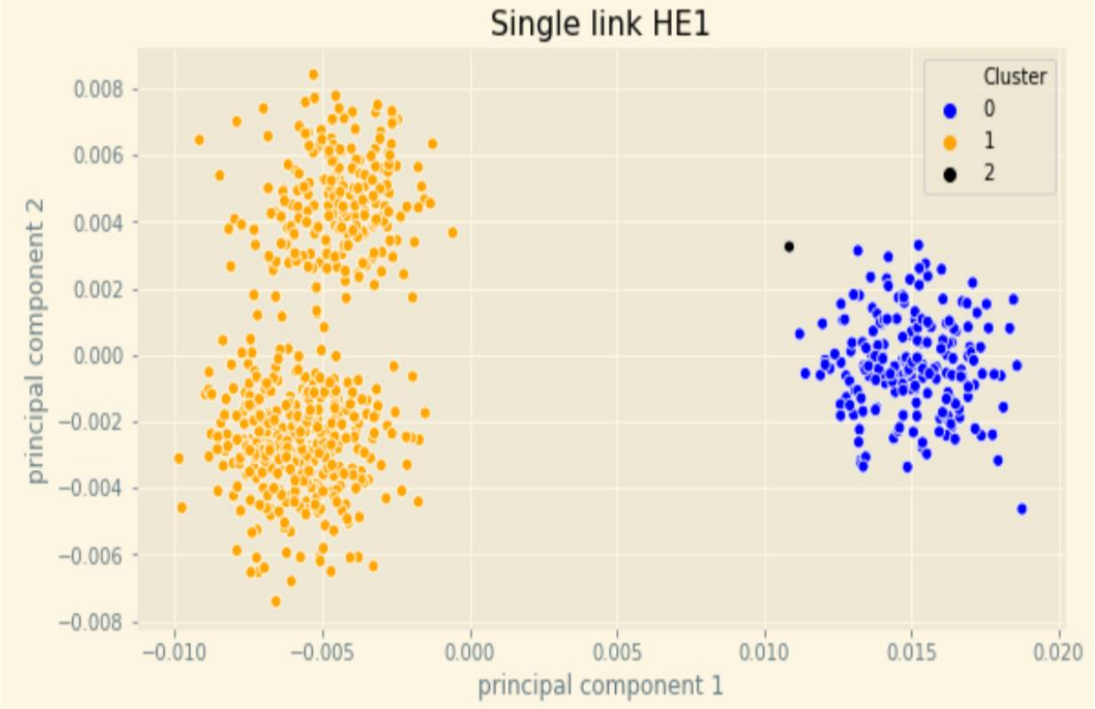
- b. Gaussian mixture model:



Task 3 – Question 3 (continued)

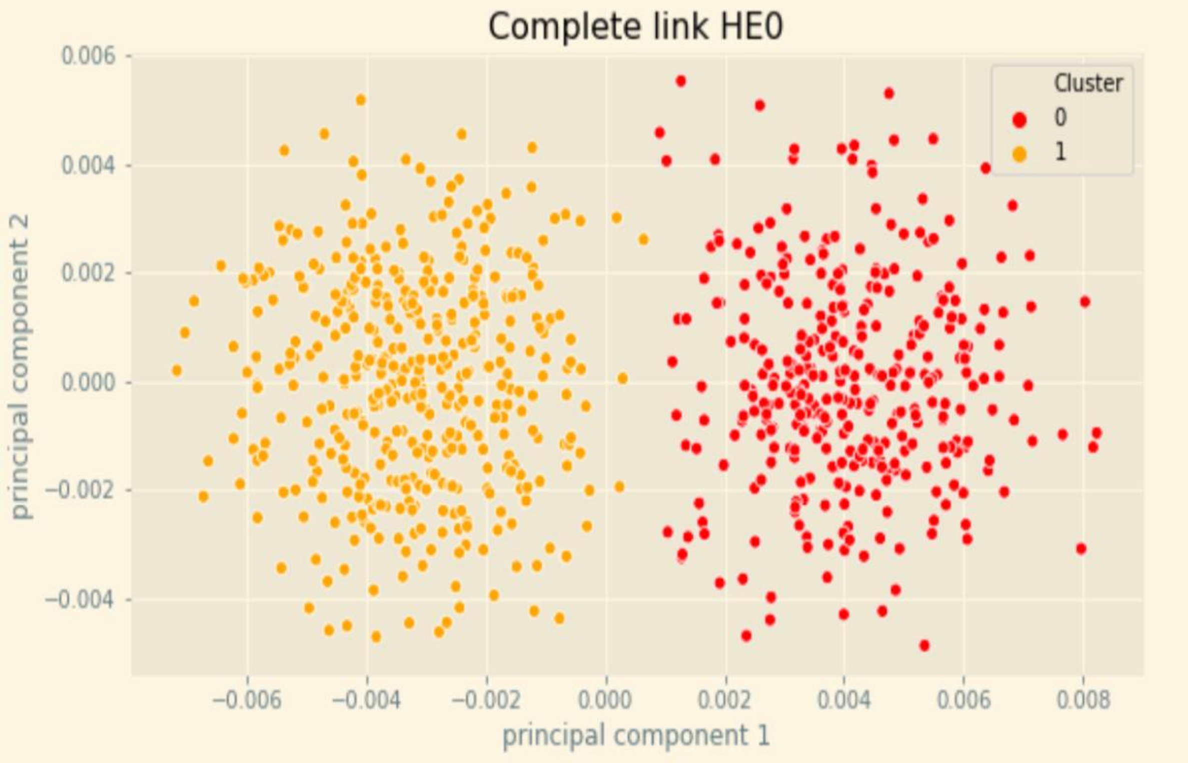
- c. single linkage hierarchical:

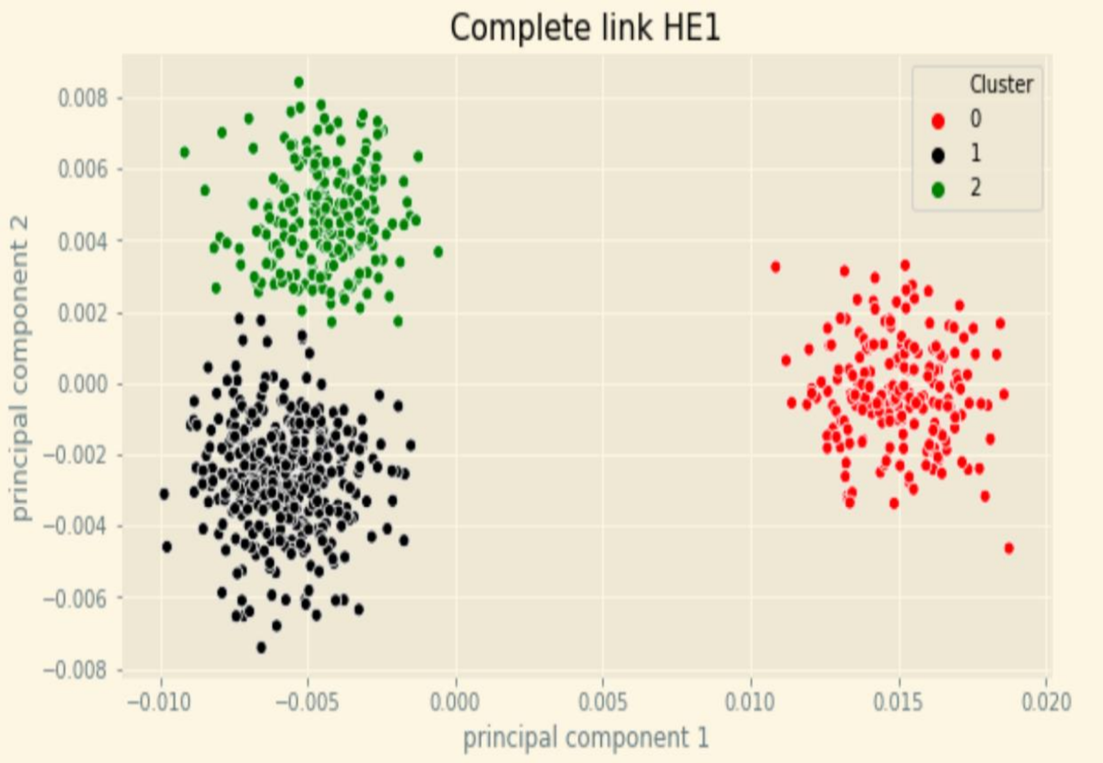




Task 3 – Question 3 (continued)

- c. complete linkage hierarchical:





Task 3 – Question 3 (continued)

d. Discussion on single vs. complete linkage hierarchical methods:

- Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.
- Unlike single link, Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics

Yes. We could see a difference in the clusters being formed in single and complete linkage clustering. There is one single data point which is located little further away from the clusters (Right side in the last 2 picture) in single linking which is forming its own cluster

Yes. The data contains some outliers which are quite distant to be included as a part of any of the inherent clusters in the dataset. This is also evident in the single link clusters formed, as single link clustering is more sensitive to noise and outliers and can find irregular cluster shapes, which is present in this case.

e. Interpretation and comparison of the different methods:

Based on the clusters formed , it can be seen that K-means and GMM produce almost identical clusters for both HE0 and HE1 clusters . One reason for this may be due to the initial seed selection. This is very evident from the centroid of clusters whose values are very close in Kmeans and GMM (Can be seen in the code). It’s not always necessary that these method produce similar results.

Hierarchal clustering produce similar results when the complete linkage distance is used while when single linkage distance is used we see that there is one single data point which is located little further away from the clusters in single linking which is forming its own cluster.

We have finally used cluster results from K-means as the basis for our analysis.

Task 3 – Question 3 (continued)

f. In context, what do the clusters you have found represent? What are some factors which could account for this type of clustering pattern?

- The clusters found through different algorithm represent the subpopulation of samples which shows certain attribute as compared to another subpopulation. For example, the first cluster might have higher relative abundance of certain specific type microbes as compared to the other clusters.
- The clusters represent samples having similar relative abundance of all microbes. Different clusters may/may not have different samples (sample number)
- The factors that could account for these type of clustering pattern would be the difference in the relative abundance of different microbes in different samples

g. Based on your process for deciding the number of clusters to partition the data into, what situations or factors might result in your decision being inaccurate?

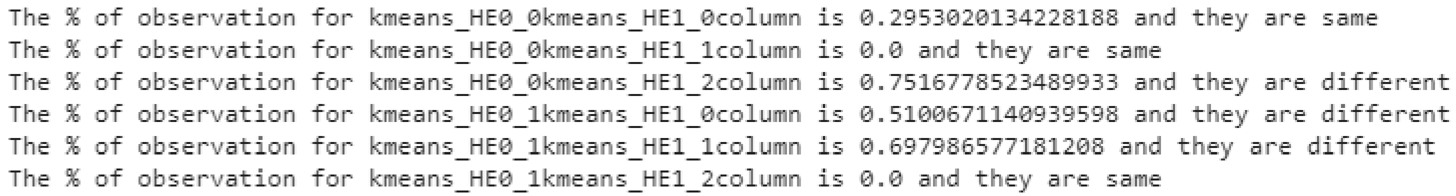
- The visual inspection approach based on the clusters visible in PCA can go wrong if the difference between the relative abundance of microbes is small.
- Conceptually, the clusters might not be distinctly visible (absence of distinct decision boundary) and can overlap, resulting in incorrect number of clusters.
- A better approach will be to use different metrics like elbow-curve, BIC etc. to determine the optimum number of clusters.

Task 4 – Question 1

- a. Determining which HE1 subpopulations had a significantly different microbiome than the HE0 samples. Explain your decision process and provide evidence supporting your conclusions.

For the 3 subpopulations of HE1,
1) kmeans_HE1_2 is statistically different from kmeans_HE0_0
2) kmeans_HE1_0 is statistically different from kmeans_HE0_1
3) kmeans_HE1_1 is statistically different from kmeans_HE0_1

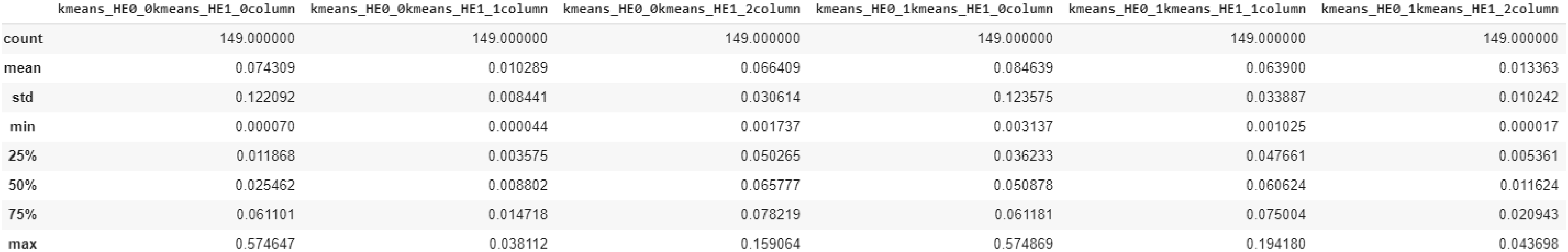
The screenshot below shows the % of microbes having % difference in abundance more than 5%



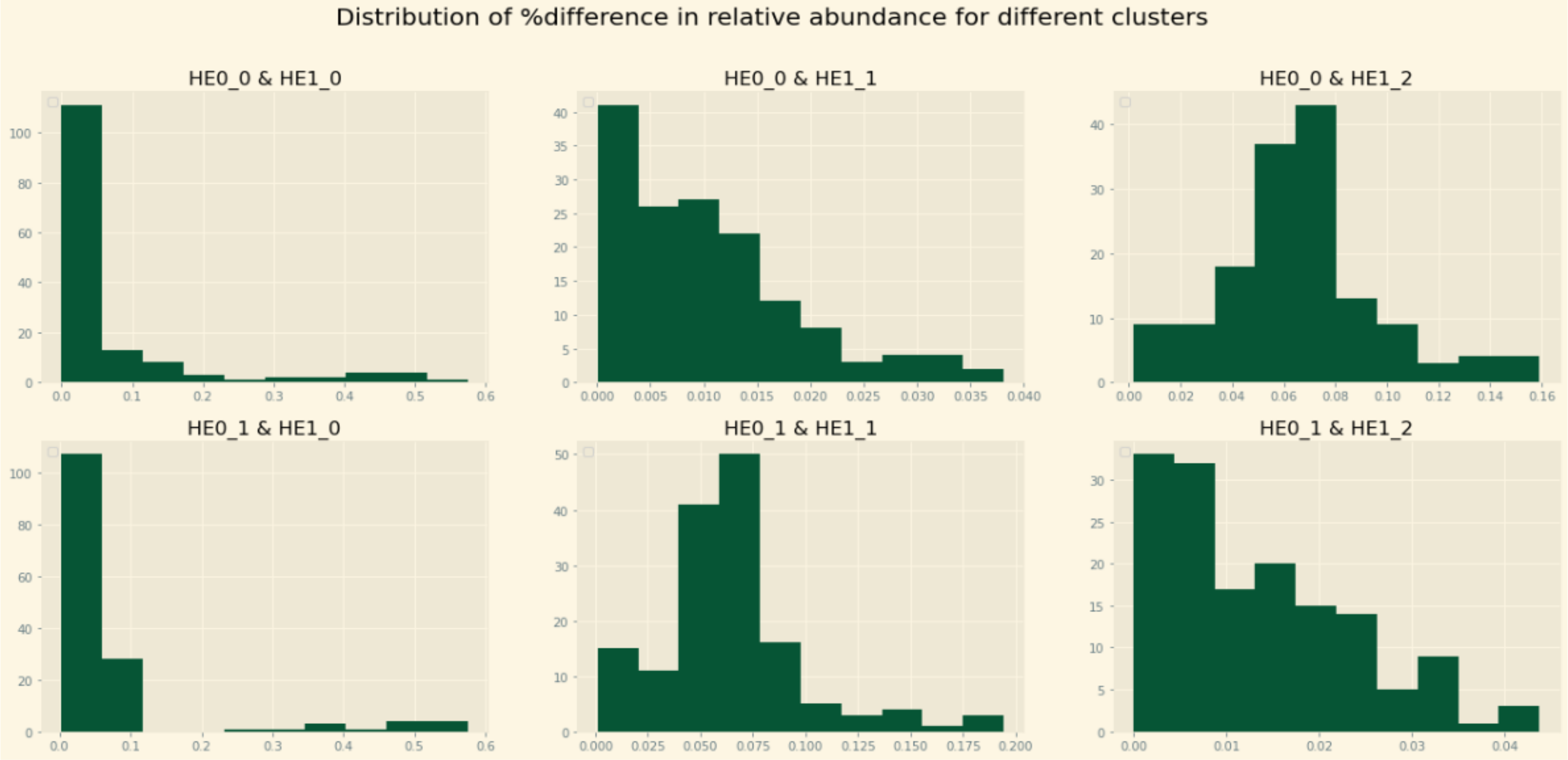
```
The % of observation for kmeans_HE0_0kmeans_HE1_0column is 0.2953020134228188 and they are same
The % of observation for kmeans_HE0_0kmeans_HE1_1column is 0.0 and they are same
The % of observation for kmeans_HE0_0kmeans_HE1_2column is 0.7516778523489933 and they are different
The % of observation for kmeans_HE0_1kmeans_HE1_0column is 0.5100671140939598 and they are different
The % of observation for kmeans_HE0_1kmeans_HE1_1column is 0.697986577181208 and they are different
The % of observation for kmeans_HE0_1kmeans_HE1_2column is 0.0 and they are same
```

- Decision process – Steps:
  - There are different cluster data at a sample level for all the 149 columns.
  - We take the average relative abundance of microbes in each of these clusters.
  - Then we compare cluster 0,1 of HE0 with cluster 0,1,2 of HE1 by taking the %difference of average relative abundance of each microbes between clusters.
  - There exist 6 pairwise comparisons in thus scenario
  - We extract the descriptive statistics of % difference of microbe for different pairwise cluster comparison as shown in the next slide.
  - Based on our mean % difference of different cluster comparison, we could see that it the %age ranges from 1% - 9% (table in the next slide). Taking the average of the range we choose the % cutoff allowed to be 5%.
  - If the total #microbes having more 5% change in relative abundance change constitutes 30% of the total, then we say that the subpopulations are different

Below table is the descriptive statistics of the %different of average abundance of different cluster comparisons



	kmeans_HE0_0kmeans_HE1_0column	kmeans_HE0_0kmeans_HE1_1column	kmeans_HE0_0kmeans_HE1_2column	kmeans_HE0_1kmeans_HE1_0column	kmeans_HE0_1kmeans_HE1_1column	kmeans_HE0_1kmeans_HE1_2column
count	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000
mean	0.074309	0.010289	0.066409	0.084639	0.063900	0.013363
std	0.122092	0.008441	0.030614	0.123575	0.033887	0.010242
min	0.000070	0.000044	0.001737	0.003137	0.001025	0.000017
25%	0.011868	0.003575	0.050265	0.036233	0.047661	0.005361
50%	0.025462	0.008802	0.065777	0.050878	0.060624	0.011624
75%	0.061101	0.014718	0.078219	0.061181	0.075004	0.020943
max	0.574647	0.038112	0.159064	0.574869	0.194180	0.043698

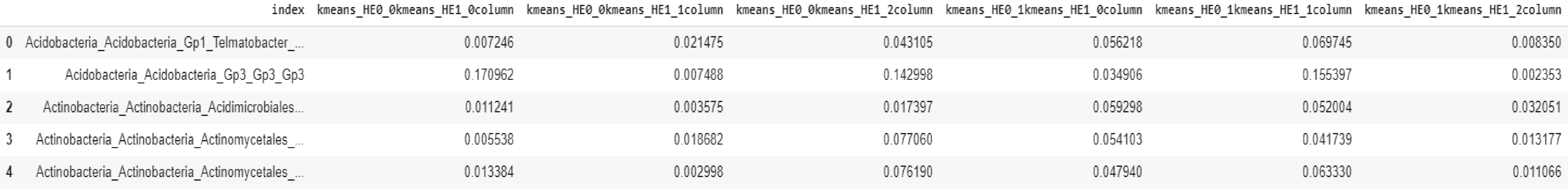


The diagram shows the distribution of % difference of relative abundance of microbes between each cluster

Note: HE0_0 means the cluster 0 from HE0. Similarly for other

Task 4 – Question 1 (continued)

Below table is the descriptive statistics of the %different of average abundance of different cluster comparisons



	index	kmeans_HE0_0kmeans_HE1_0column	kmeans_HE0_0kmeans_HE1_1column	kmeans_HE0_0kmeans_HE1_2column	kmeans_HE0_1kmeans_HE1_0column	kmeans_HE0_1kmeans_HE1_1column	kmeans_HE0_1kmeans_HE1_2column
0	Acidobacteria_Acidobacteria_Gp1_Telmatobacter_...	0.007246	0.021475	0.043105	0.056218	0.069745	0.008350
1	Acidobacteria_Acidobacteria_Gp3_Gp3_Gp3	0.170962	0.007488	0.142998	0.034906	0.155397	0.002353
2	Actinobacteria_Actinobacteria_Acidimicrobiales...	0.011241	0.003575	0.017397	0.059298	0.052004	0.032051
3	Actinobacteria_Actinobacteria_Actinomycetales_...	0.005538	0.018682	0.077060	0.054103	0.041739	0.013177
4	Actinobacteria_Actinobacteria_Actinomycetales_...	0.013384	0.002998	0.076190	0.047940	0.063330	0.011066

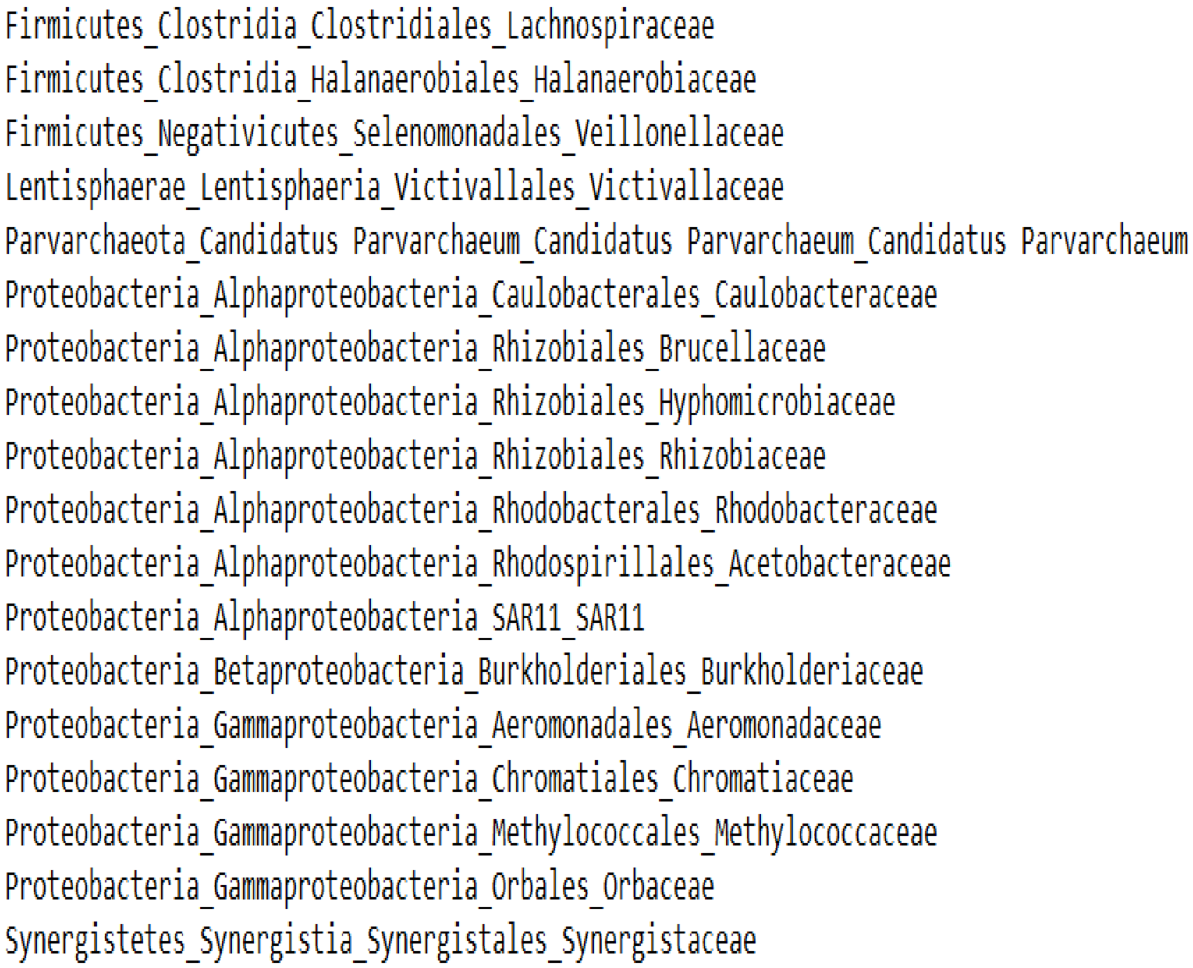
Task 4 – Question 1 (continued)

- b. Determining the HE0subpopulation most similar to each HE1 subpopulation with a significantly different microbiome. Explain the decision process and provide evidence to support your conclusions.
- Decision process – Steps:
  - There are different cluster data at a sample level for all the 149 columns
  - We take the average relative abundance of microbes in each of these clusters
  - We compare cluster 0,1 of HE0 with cluster 0,1,2 of HE1 by taking the %difference of average relative abundance of each microbes between clusters
  - There exist 6 pairwise comparisons in thus scenario

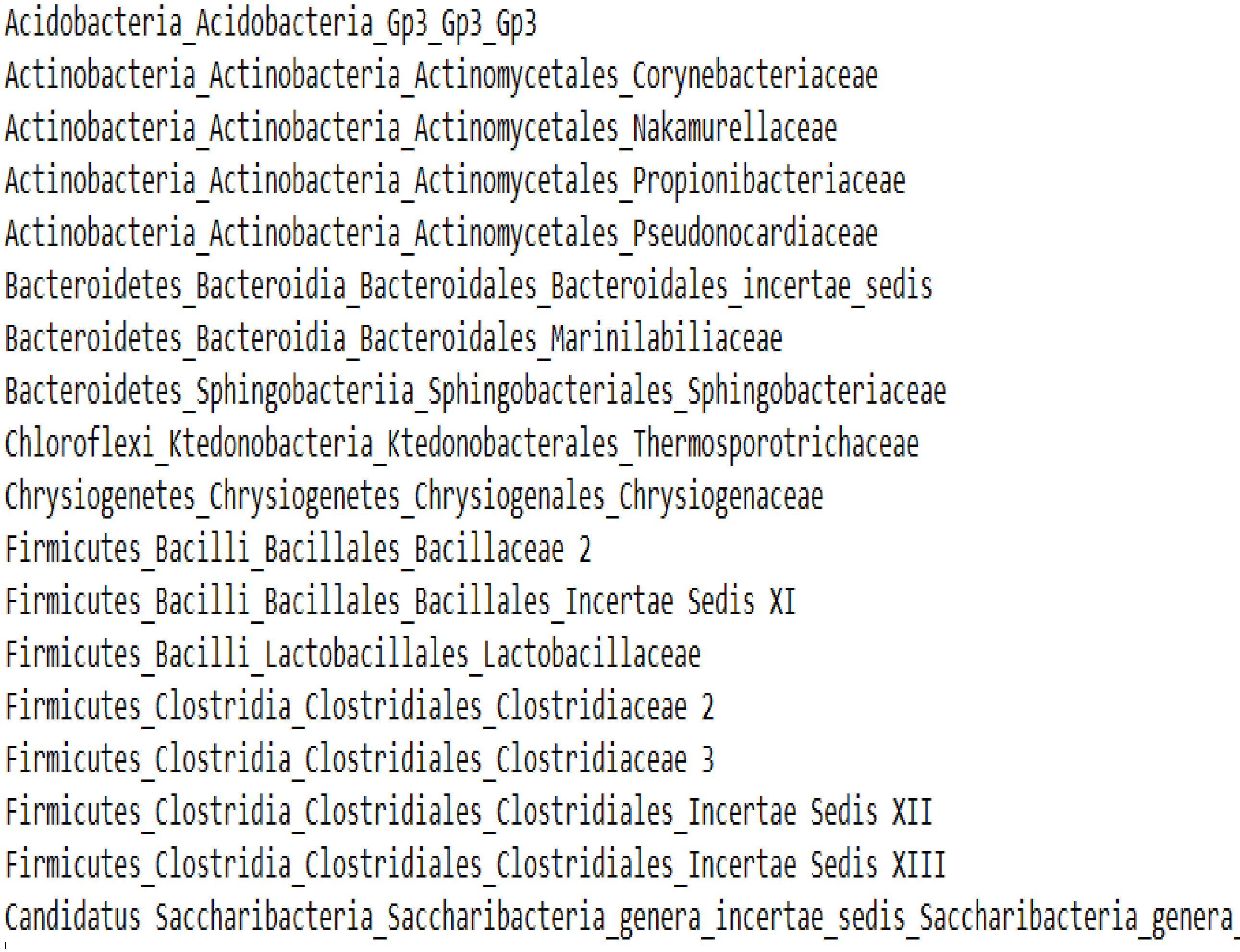
  - We extract the descriptive statistics of %diff of microbe for different pairwise cluster comparison as shown in the next slide
  - Based on our mean % difference of different cluster comparison, we could see that it the %age ranges from 1% - 9%. Hence we take the %cutoff allowed to be 5%
  - If the total #microbes having more 5% change in relative abundance change constitute 30% of the total, then we say that the subpopulations are different
  - Similarly if #microbes is less than 30% of the total, then we say the two sub population are similar

Task 4 – Question 1 (continued)

c. Microbes with significantly altered abundance based on KS test:



```
Firmicutes_Clostridia_Clostridiales_Lachnospiraceae
Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae
Firmicutes_Negativicutes_Selenomonadales_Veillonellaceae
Lentisphaerae_Lentisphaeria_Victivallales_Victivallaceae
Parvarchaeota_Candidatus Parvarchaeum_Candidatus Parvarchaeum_Candidatus Parvarchaeum
Proteobacteria_Alphaproteobacteria_Caulobacterales_Caulobacteraceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae
Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae
Proteobacteria_Alphaproteobacteria_Rhodospirillales_Acetobacteraceae
Proteobacteria_Alphaproteobacteria_SAR11_SAR11
Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae
Proteobacteria_Gammaproteobacteria_Aeromonadales_Aeromonadaceae
Proteobacteria_Gammaproteobacteria_Chromatiales_Chromatiaceae
Proteobacteria_Gammaproteobacteria_Methylococcales_Methylococcaceae
Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae
Synergistetes_Synergistia_Synergistales_Synergistaceae
```



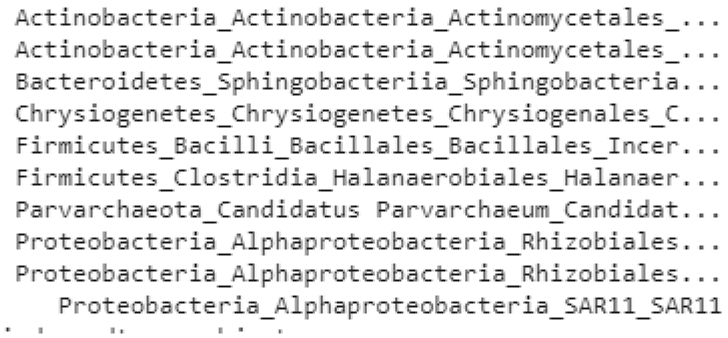
```
Acidobacteria_Acidobacteria_Gp3_Gp3_Gp3
Actinobacteria_Actinobacteria_Actinomycetales_Corynebacteriaceae
Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae
Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae
Actinobacteria_Actinobacteria_Actinomycetales_Pseudonocardiaceae
Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidales_incertae_sedis
Bacteroidetes_Bacteroidia_Bacteroidales_Marinilabiliaceae
Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae
Chloroflexi_Ktedonobacteria_Ktedonobacterales_Thermosporotrichaceae
Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae
Firmicutes_Bacilli_Bacillales_Bacillaceae 2
Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI
Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae
Firmicutes_Clostridia_Clostridiales_Clostridiaceae 2
Firmicutes_Clostridia_Clostridiales_Clostridiaceae 3
Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XII
Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XIII
Candidatus Saccharibacteria_Saccharibacteria_genera_incertae_sedis_Saccharibacteria_genera_
```

Task 4 – Question 2

- a. Which of the microbes that you identified show an increase of relative abundance in the HE1 sample? Do any show a decrease?

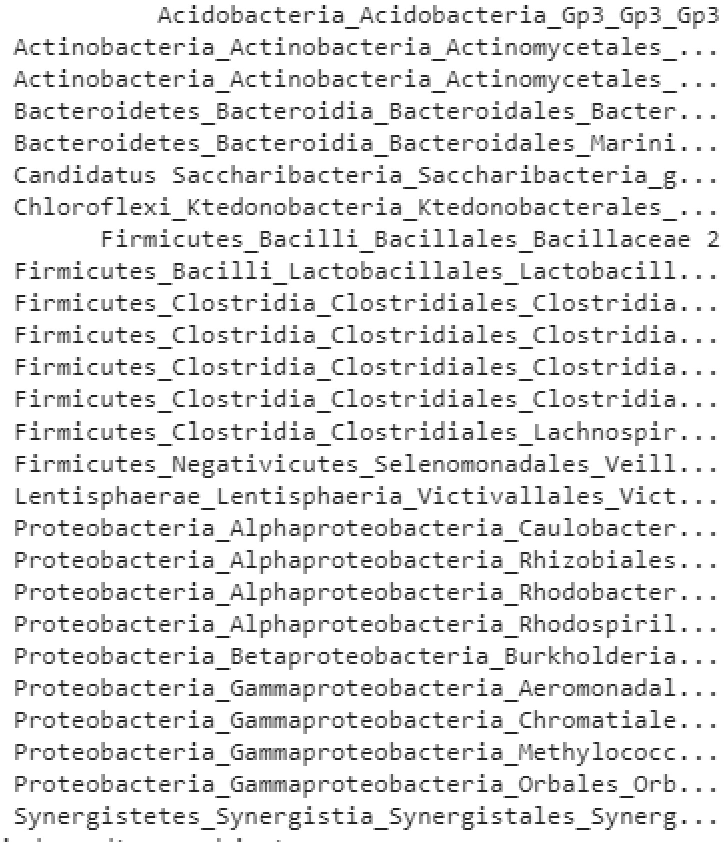
There are a total of 10 microbes which has shown an increase of relative abundance in HE1 sample and 26 shows decrease

Increase



```
Actinobacteria_Actinobacteria_Actinomycetales_...
Actinobacteria_Actinobacteria_Actinomycetales_...
Bacteroidetes_Sphingobacteriia_Sphingobacteria...
Chrysiogenetes_Chrysiogenetes_Chrysiogenales_C...
Firmicutes_Bacilli_Bacillales_Bacillales_Incer...
Firmicutes_Clostridia_Halanaerobiales_Halanaer...
Parvarchaeota_Candidatus Parvarchaeum_Candidat...
Proteobacteria_Alphaproteobacteria_Rhizobiales...
Proteobacteria_Alphaproteobacteria_Rhizobiales...
   Proteobacteria_Alphaproteobacteria_SAR11_SAR11
```

Decrease



```
          Acidobacteria_Acidobacteria_Gp3_Gp3_Gp3
Actinobacteria_Actinobacteria_Actinomycetales_...
Actinobacteria_Actinobacteria_Actinomycetales_...
Bacteroidetes_Bacteroidia_Bacteroidales_Bacter...
Bacteroidetes_Bacteroidia_Bacteroidales_Marini...
Candidatus Saccharibacteria_Saccharibacteria_g...
Chloroflexi_Ktedonobacteria_Ktedonobacterales_...
      Firmicutes_Bacilli_Bacillales_Bacillaceae 2
Firmicutes_Bacilli_Lactobacillales_Lactobacill...
Firmicutes_Clostridia_Clostridiales_Clostridia...
Firmicutes_Clostridia_Clostridiales_Clostridia...
Firmicutes_Clostridia_Clostridiales_Clostridia...
Firmicutes_Clostridia_Clostridiales_Clostridia...
Firmicutes_Clostridia_Clostridiales_Lachnospir...
Firmicutes_Negativicutes_Selenomonadales_Veill...
Lentisphaerae_Lentisphaeria_Victivallales_Vict...
Proteobacteria_Alphaproteobacteria_Caulobacter...
Proteobacteria_Alphaproteobacteria_Rhizobiales...
Proteobacteria_Alphaproteobacteria_Rhodobacter...
Proteobacteria_Alphaproteobacteria_Rhodospiril...
Proteobacteria_Betaproteobacteria_Burkholderia...
Proteobacteria_Gammaproteobacteria_Aeromonadal...
Proteobacteria_Gammaproteobacteria_Chromatiale...
Proteobacteria_Gammaproteobacteria_Methylococc...
Proteobacteria_Gammaproteobacteria_Orbales_Orb...
Synergistetes_Synergistia_Synergistales_Synerg...
```

Question 4.b

- Taxonomical relationships and groups among microbes with altered abundance:
- Yes, the altered microbes does have a parent or phylum (in scientific terms) which is an umbrella of group of microbes. There are a total of 11 groupings for these altered microbes

They are the following:

Acidobacteria
Actinobacteria
Bacteroidetes
Candidatus Saccharibacteria
Chloroflexi
Chrysiogenetes
Firmicutes
Lentisphaerae
Parvarchaeota
Proteobacteria
Synergistetes