Visual Analysis of the Ocean Microbiome

This is the submission table to be used for the summative assessment in CSC8636 – Complex Data Visualization. Fill in your comments and answers in the table below for each of the assessment parts, as lined out in the Summative Assessment description. Add relevant screenshots as appendix in this document, and refer to them in your table comments. You should also include a list of references to sources you have used, and cite them appropriately in your answers. Submit this document in pdf format together with your Python code and any datasets that are loaded by the code, as a single zip file in NESS. ***The submission deadline is 16:30 on Friday 19th February***

Part 1 – OTU table (30%)	**Your comments and answers**
Describe your choice of method for dealing with the high-dimensionality of the data	There are various methods for dealing with the high-dimensionality data during analysis process. Out of the methods, I chose to aggregate the dataset based on the clustering mechanism. This clustering method is used based on the different hierarchy levels representing the Ocean microbiome dataset. Based on the hierarchy levels, the abundance values were aggregated by the summation within the obtained samples and calculated the average across the SampleId's.
Reflect on your choice of method for dealing with dimensionality, in context of alternative approaches (e.g. "I chose *method_A*, because it does *abc*, another alternative could have been *method_B*, which does *cde*, but I believe *method_A* is a better choice because …")	I choose the clustering method for reducing dimensions in this high-dimensionality data because, there won't be any loss in the details that has been stored in all the variables. The clustering method aggregates the data at any required hierarchical level which is easy to use for visualization without losing the data. One of the well-known alternative method of dimensionality reduction method would be identifying the dimensions based on the Principal Component analysis method. Where the dimensions which are of less important are removed for the analysis process. Since if we use the Principal component analysis method we might lose some valuable information from the removed variables and keeping in mind of the loss of data, Clustering method is approached.
Describe your choice of visualization method to use	I choose to represent the abundance pattern of the ocean microbiome in a scatter plot. The scatter plot visualization is generated by aggregating the abundance value using the Phylum and displayed it by the color combination of different Domains. The abundance scale is represented by taking the logarithmic value as apart from the top 5 to 6 most abundant phylums, all the other phylums are below the range of 0.
Motivate/justify your choice of visualization method, and explain how you decided on	Representing the patterns in scatterplot is easy for users to identify the outliers as it helps in focusing on which Phylum should be taken in account more than considering all the Phylums. Based on the color combination and logarithmic

number of dimensions to display in the visualization	scale it is easy for viewers to identify the most abundant Phylums and which Domain are they representing. I have limited the dimensions to 3 which consists of Domains, Phylums and Abundance as it would be easy to interpret the details for the users. This visual also helps in providing interactivity to view the abundance percentage by hovering over the scatterplot.
Reflect on your choice of visualization method, in context of alternative approaches	The Scatterplot is better for displaying the outliers or to identify patterns in a dataset. Apart from this visualization, alternative approaches can be made using Pie charts or Bar charts or box plot by reducing the dimensionality based on the Principal Component Analysis method.
Give some examples of possibly interesting data patterns that you can find through your visualization	**Possible data patterns:** • Most of the Phylums belong to the Bacteria domain. • The Proteobacteria phylum has the highest abundance of nearing at the range of 60%. • The Phylum which belongs to Archaea domain seems to be the second most abundant domain from the data. • Apart from the Proteobacteria Phylum all the other phylum’s are below the range of 50%.
Part 2 – Sample mapping (20%)	
Describe your choice of visualization approach(es) for representing sample categories	For representing the sample categories, I choose to use the Parallel Categories visualization with the reduction of dimensionality based on clustering method. The visualization is made by finding the topmost abundant Phylums with respect to the categories that they belong to while observing the microbiome data.
Reflect on and justify your choice of visualization approach(es), in context of alternative approaches	The alternative approach for visualizing this can be done using the Scatter-Plot matrix in a 3D visual approach to understand the pattern with respect to the categories. But comparing to Parallel Categories, the Scatter-plot matrix will not be able to showcase more than 3 or 4 variables and hence Parallel Categories is used in-order to give the user the clear idea of how the most abundant Phylums is taken from the different categories. The color combination used in the visualization is based on ranking with respect to the top 3 most abundant Phylums.
Give some examples of data patterns that can be found through your visualization	**Possible data patterns:** • The plot is represented and color combination is used based on the order of most abundant Phylums. • We could find that most abundant phylum is colored in Black followed by Purple and LightBlue. • The Phylum Proteobacteria is the most abundant and they are obtained from all the three regions of the

	categories and are equally obtained from all the ocean regions. • Cyanobacteria and Thaumarchaeota Phylums are the second most abundant. • All the top 3 abundant microbiomes are mostly from the SRF layer.
Part 3 – Multiple levels (35%)	
Describe your choice of visualization methods for visualizing the three levels of detail. Provide detail on which level you are visualizing in each visualization view.	For visualizing different hierarchical levels in detail again Parallel categories visualization is used which gives a clear information of the origin of the microbiomes based on different levels. The first visualization is made at Domain, Class and Order levels of the Top 3 most abundant microbiomes. Each visualization is made by linking it with the Levels of origin category to give a clear understanding.
Reflect on your choice of visualization methods for the three levels of detail, in context of alternative methods	I choose Parallel categories to represent the visuals as the hierarchical levels can be interlinked to each other to identify the pattern towards the data. An alternative approach is to make visualization using stacked bar charts which can represent the most abundant microbiomes in a stacked manner whereas through Parallel Categories it is easy to interlink with the furthermore levels which is easy for reproducing the visualization.
Describe the approaches you have used to provide consistency across the multiple views	To provide the consistency across the multiple visuals, the number of samples displayed are same and the details of most abundant microbiomes are displayed for each level. Further to this all the visuals are represented based on the color combination of Levels of origin category and the hierarchical levels related to that levels is included.
Describe how you have used visualization theory and guidelines when designing the multiple views visualization, using appropriate references	• Color combination used in the visuals is maintained with the respect to the vision impaired users. • The number of levels, categories and samples represented are limited which makes user to understand the visualization easily. • The interactivity in the visuals is made by hovering over the links to identify the counts of samples which are linked to the next levels and categories.
Give some examples of data patterns that can be found through your visualization	**Possible data patterns:** • From the visuals we can identify that the most abundant microbiomes belong to the Domain Bacteria. • While displaying the visual based on level Class, we can identify that Alphaprocteobacteria has the highest abundant microbiomes which belongs to the Phylum Procteobacteria and this is displayed in the visuals.

	• The third level of visual is represented based on the Order level where we can view few microbiomes which belong to Archaea and Undef domain has the most abundant microbiomes. • The color combination is represented based on the category Levels of origin to identify from which category the most abundant microbiomes are created.
Part 4 - Uncertainty (10%)	
Describe potential sources of uncertainty that may exist in the data	The sources of uncertainty are present when we move onto visualize the data based on granular level of the taxonomy dataset. Where we could find some of the undefined values in the Order, Family and Genus levels. The undefined microbiome is even observed in the Domain level. Some of the class and order of the microbiomes are recorded as Dates which is also a potential source of uncertainty. This uncertainty exists in the data as for some of the microbiome there might not be an accurate classification, or it might be the error made while recording the observations.
Describe how you could visualize this uncertainty, based on your visualization in part 1, 2 or 3	• The uncertainty is visualized using the scatter plot in Part 1 to represent the pattern based on the Domain and Phylum level. • Apart from that all the other parts are focused on displaying the most abundant microbiomes and the class level which contains dates seems to be less abundant as it is not displayed in the visuals. • Even though the Undef is represented at the domain level, since it has 5% of abundant microbiome it can be visualized in all the visuals.

Part 1- Abundance Pattern:

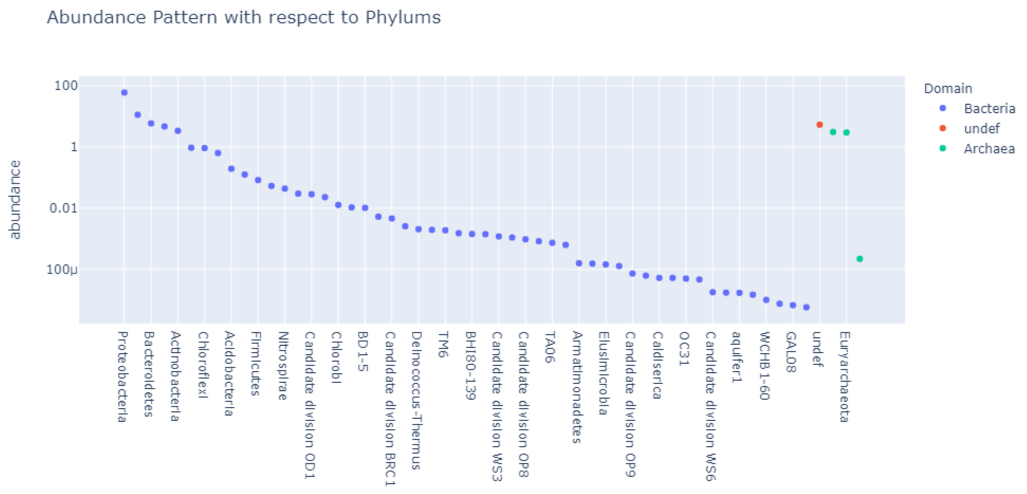


Fig.1

Part – 2: Sample Categorical visualization:

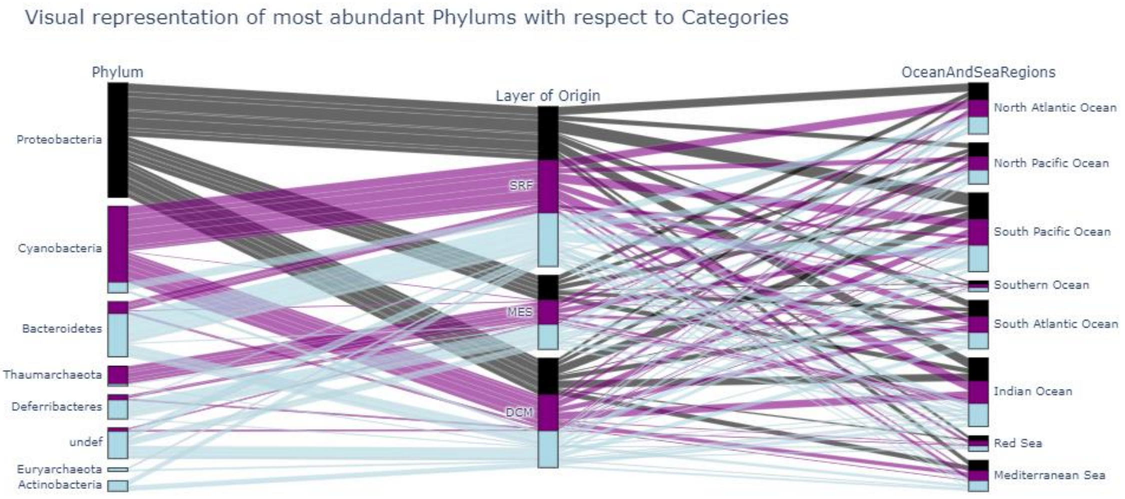

Visual representation of most abundant Phylums with respect to Categories

Fig.2

Part – 3: Visualizations at different hierarchical levels:

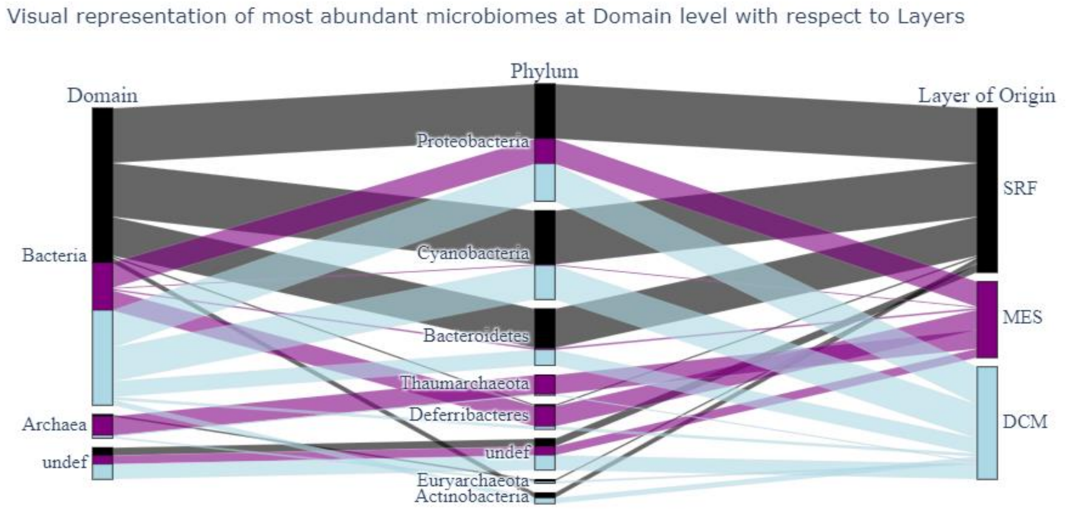
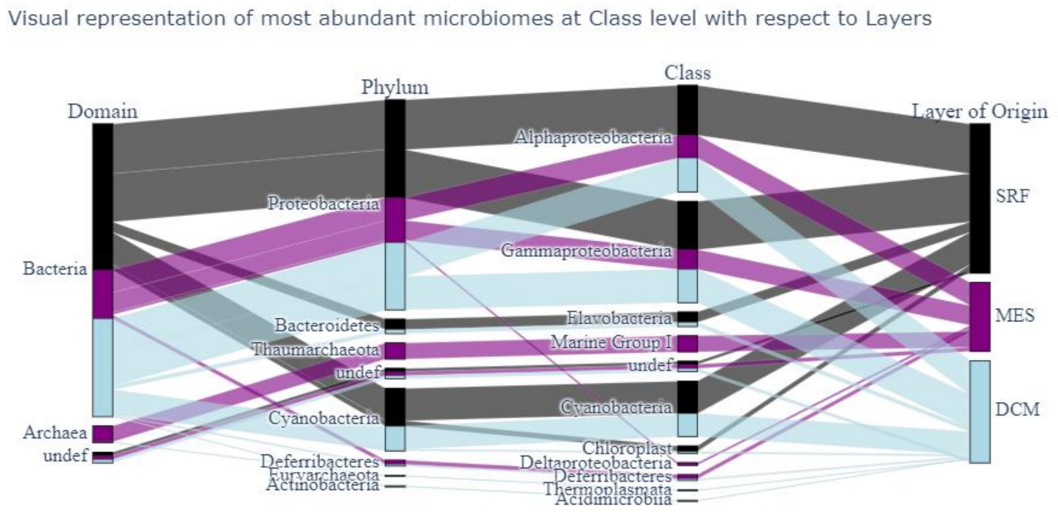
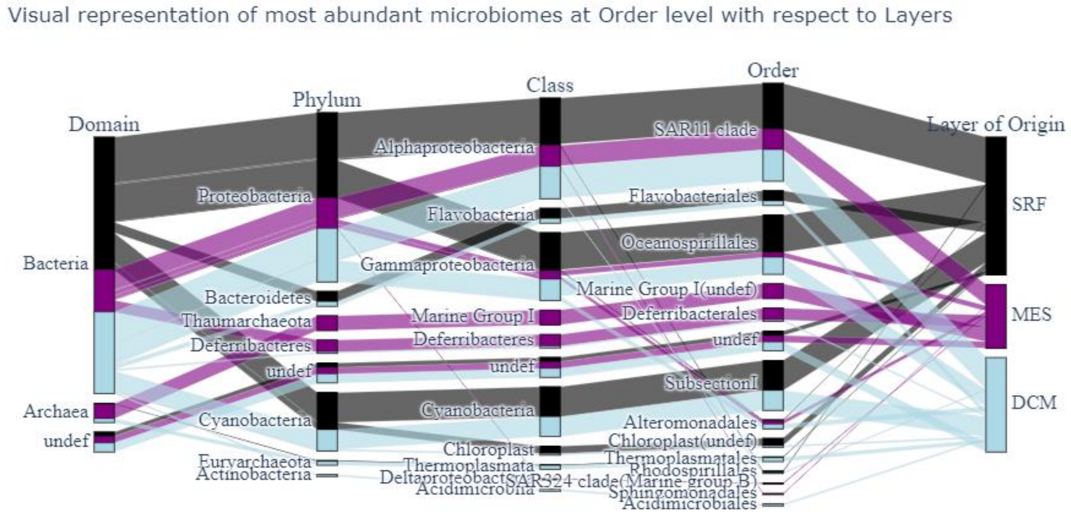

Visual representation of most abundant microbiomes at Domain level with respect to Layers

Fig.3.1



Visual representation of most abundant microbiomes at Class level with respect to Layers

Fig.3.2



Visual representation of most abundant microbiomes at Order level with respect to Layers

Fig.3.3

Appendix:

- All the pre-processing steps of merging and creating a master data is done using the Python.
- The interactive visuals are created using some of the packages in python.
- The packages that are used are "Pandas" for creating the master data.
- "Plotly" package is used for creating the interactive visuals with the ocean microbiome dataset.