Visual Analysis of the Ocean Microbiome

This is the submission table to be used as report 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. You should 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 visualization as html page, Python code and any datasets that are loaded by the code, as a single zip file in Canvas. ***The submission deadline is 16:30 on Thursday 22nd February.***

Student name: Vidadi Nasibov **Student ID: 230302626**

Part 1 – Interactive visualization using multiple coordinated views (60%)	**Your comments and answers**
Describe and justify how your visualization meet the aim of understanding overall abundance patterns and diversity (aim 1), and reflect on alternative approaches. Explain how you dealt with the high dimensionality.	The visualization achieves Aim 1 by employing bar charts and scatter plots to showcase the most abundant microbiomes across different oceanic regions. This approach simplifies complex multi-dimensional data, enabling users to discern patterns of microbial distribution and relative abundance with ease. To manage high dimensionality, the data was aggregated, averaged, and top-20 filtering was applied, ensuring clarity and focus on significant trends without overwhelming the user. Alternative approaches could involve the use of dimensionality reduction techniques such as PCA (Principal Component Analysis) or t-SNE (t-distributed Stochastic Neighbor Embedding) to visualize high-dimensional data in a low-dimensional space, but these methods might abstract the data too much for non-expert users.
Describe and justify how your visualization meet the aim of understanding abundance patterns in context of sample meta data (aim 2), and reflect on alternative approaches. Explain how you dealt with the heterogeneity of the data?	Visualization 3, a scatter plot, correlates the abundance of specific genera with sample metadata such as depth and biome type, providing insight into environmental influences on microbial populations. This visualization uses colour coding to distinguish between phyla and integrates depth, which is a piece of sample metadata, into the analysis. This meets Aim 2 by contextualizing abundance data within the environment from which it was sampled. The heterogeneity of the data was addressed by categorizing it into discrete, color-coded groups, facilitating easier comparison across categories. Additionally, visualization 4 is a scatter plot that compares microbial diversity across phyla and depths. An alternative could be to use multidimensional scaling or similar methods to display the data in a way that reflects the similarity of samples based on their metadata profiles,

	which could reveal additional patterns or clusters within the data.
Describe and justify how your visualization meet the aim of investigating both overview and detail (aim 3), and reflect on alternative approaches.	The visualizations employ a multi-faceted approach to meet aim 3. Visualization 5 is a parallel coordinates plot that compares different samples across multiple dimensions, including depth and geographic location, which helps to provide both an overview and the ability to investigate detailed patterns. Sankey diagrams provide a macro overview of microbial abundance flows at different taxonomic levels. Alternative approaches could include using tree maps for hierarchical data or linked brushing for exploring high-dimensional data spaces. These methods can provide a simultaneous overview and detail but may not be as intuitive for all user groups.
Describe and justify your use of visualization theory and design principles, including e.g. the use of visual channels, colour, gestalt theory, and design guidelines.	The design principles are grounded in visualization theory. Colour is employed as a primary visual channel to differentiate between taxa, adhering to the gestalt principle of similarity. The principle of proximity is observed in the grouped bars of the bar charts, facilitating pattern recognition. The layouts follow established guidelines for clarity and simplicity, ensuring that the visualizations are easily interpretable and aesthetically pleasing.
Describe and justify your use of interactivity, referring to theory where relevant.	Interactivity is leveraged through dropdown menus and tooltips, enhancing user engagement and data exploration. This approach is supported by the information-seeking mantra: "overview first, zoom and filter, then details-on-demand". It provides users with the autonomy to navigate through the data at their discretion, which is essential for effective data analysis.
Describe and justify your approach to coordinating the multiple views, referring to theory where relevant.	Multiple views are coordinated through interactive elements that enable users to control the display of data, which is consistent with the coordinated multiple views (CMV) theory. This theory suggests that different views provide different perspectives of the same data, and user interactions in one view should reflect in others, enriching the data exploration experience.
Give examples of possibly interesting data patterns that you can find through your visualization, you are welcome to include screenshots of the patterns below this table.	One interesting pattern observable is the predominance of certain microbial genera in specific ocean regions, hinting at ecological preferences. Another pattern is the variation in microbial diversity with depth, which can be crucial for understanding marine ecosystems' health and function.
Part 2 - Uncertainty (10%)	
Describe potential sources of uncertainty that may exist in the data	Uncertainty in the data may arise from measurement errors, incomplete taxonomic classification, sampling bias, and environmental variability. These uncertainties can

	affect the accuracy and reliability of the inferred microbial abundance and diversity patterns.
Describe how you could visualize this uncertainty, based on your visualization in part 1.	Uncertainty could be visualized using error bars in bar charts, confidence intervals in scatter plots, or varying the opacity of lines in Sankey diagrams. These visual cues would provide users with an understanding of the reliability of the data and caution in interpreting the results. For example, in scatter plots, a gradient can denote the concentration of data points, indicating less certainty in areas with sparse data. In Sankey diagrams, the thickness of the links can vary to represent the degree of uncertainty in the abundance flow between taxa.
Part 3 – Heuristics evaluation (20%)	**Describe how your visualization in part 1 meet the following heuristic criteria, or how it could be changed to meet these criteria.**
The visualization facilitates answering questions about the data.	The visualizations are structured to guide users in querying the dataset effectively, offering multiple angles and levels of detail to answer a variety of questions about microbial abundance and diversity.
The visualization provides a new or better understanding of the data.	By transforming raw data into visual formats, the visualizations unveil patterns and trends that may not be immediately evident in tabular data, offering new insights.
The visualization provides opportunities for serendipitous discoveries.	Interactive features allow users to stumble upon unexpected patterns, fostering discoveries beyond the initial scope of exploration.
The visualization affords rapid parallel comprehension for efficient browsing.	The use of coordinated views enables users to quickly compare and contrast data across different visualizations, enhancing the browsing experience.
The visualization provides mechanisms for quickly seeking specific information.	Dropdown menus and tooltips facilitate swift navigation to specific data points or trends, streamlining information retrieval.
The visualization provides a big picture perspective of the data.	Sankey diagrams and bar charts give a comprehensive view of the overall data structure and the main abundance patterns.
The visualization provides an understanding of the data beyond individual data cases.	The visualizations contextualize individual data points within larger ecological and environmental frameworks, offering a broader understanding.
The visualization helps avoid making incorrect inferences.	The careful design and implementation of the visualizations aim to minimize misinterpretations by providing clear, accurate representations of the data.
The visualization facilitates learning more broadly about the domain of the data.	The visualizations serve as educational tools, enhancing the user's knowledge of marine microbiology through interactive learning.
The visualization helps understand data quality.	By providing detailed views and the ability to drill down into the data, the visualizations help assess the quality and robustness of the data.

List of references:

1. Grimaudo, L., & Concas, A. (2021). Visual Analytics in High-Dimensional Data Applied to Dimensionality Reduction Techniques for Biological Data Sets. *Frontiers in Bioinformatics*, 1. https://www.frontiersin.org/articles/10.3389/fbinf.2021.774631/full
2. Roberts, J. C. (2007). State of the Art: Coordinated & Multiple Views in Exploratory Visualization. *University of Kent*. https://www.cs.kent.ac.uk/pubs/2007/2559/content.pdf
3. Plotly. (n.d.). *Plotly Python Graphing Library*. https://plotly.com/python/
4. The Carpentries Incubator. (n.d.). Introduction to High-Dimensional Data and Dimension Reduction. *High-Dimensional Statistics with R*. https://carpentries-incubator.github.io/high-dimensional-stats-r/01-introduction-to-high-dimensional-data/index.html
5. Lê, S., Josse, J., & Husson, F. (2022). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10320150/
6. Analytics Vidhya. (2021). Interactive Plots in Python with Plotly: A Complete Guide. https://www.analyticsvidhya.com/blog/2021/10/interactive-plots-in-python-with-plotly-a-complete-guide/
7. Towards Data Science. (n.d.). The What, Why, and How of Sankey Diagrams. https://towardsdatascience.com/the-what-why-and-how-of-sankey-diagrams-430cbd4980b5