Final Project - Microbial Composition of Freshwater Lakes and Bogs

This final project ties together all aspects of the course and requires tidying and filtering data, string cleaning, and plotting, as well as fitting a linear model. There are 3 data files necessary to complete this assignment:

1. An OTU (Operational Taxonomic Unit) table. Operational Taxonomic Units (OTUs) are groups of organisms defined by a specified level of DNA sequence similarity at a marker gene (e.g. 97% similarity at the V4 hypervariable region of the 16S rRNA gene). OTUs are a measure of microbial abundance.
2. A metadata table with the sample name, depth, dissolved oxygen (DO), Temperature, pH and Secchi.Depth.
3. A taxonomy table of Kingdom, Phylum, Class, Order, Lineage, Clade and Tribe for each OTU number and a percent identity value in brackets.

Note: Each sample name is composed of: An abbreviated lake or bog name. Whether the sample is from the Hypolimnia (H - lower layer), Epilimnia (E - upper layer), or an unknown (U) Layer of the lake or bog, the date of sampling, and possible replicate information (i.e. R1, R2).

Data is from the paper:

_ Linz et al., "Bacterial Community Composition and Dynamics Spanning Five Years in Freshwater Bog Lakes". (mSphere, 2017)_

and was retrieved from the 'OTUtable' R package.

Functions from the tidyverse are to be used to complete whenever possible. Any plots generated should have appropriate axis labels and titles.

The merged dataset will become quite large. Make sure to think about the subset of data you will need to answer each question.

Questions:

1. Read in the OTU table and reshape it into tidy format. Filter out all instances of OTUs less than 2 to reduce the size of the data set. Familiarize yourself with the data set and perform any other data cleaning necessary. (3 marks)
2. Collapse replicates by taking the mean of each OTU_num for any replicates. Make sure you do not have redundant information in your data frame after this operation. (2 marks)
3. Which site and Layer has the greatest total number (sum) of OTUs? Include the number of OTUs at this site and layer. (2 marks)
4. Read in your metadata file and combine it with your otu table. Why is the number of rows so large after the join with this particular dataset? (2 marks)
5. Make a plot showing dissolved oxygen vs pH. Describe what you see. Remove the outlier and plot the data again making any necessary adjustments. Based on the metadata available to you, can you find any relationships explaining either of these variables? Can you show any additional information to support your claim? (6 marks)
6. Which Site has the greatest range in Temperature? Use descriptive statistics to illustrate your answer. (2 marks)
7. Read in the taxonomy file. Subset to include only the Phylum, Order and Class information for each OTU number. Remove the prefix (ie. p___), the percent identity, and any round or square brackets from the taxonomy information. (3 marks)
8. Combine the taxonomy information with your joined metadata and otu table. Remove Secchi.Depth. (2 marks)

9. Is the difference in mean OTUs significant for Proteobacteria between the epilimnia and hypolimnia layers? Make an exploratory boxplot to visualize the data. Make a prediction. Then use a t.test with a 99% confidence interval to obtain your final answer. (5 marks)
10. Given a model of Dissolved Oxygen (DO) predicted from Sites only, what would you conclude after multiple hypothesis testing with the bonferroni correction? For this question, use only the subset of data for which there is no missing data. (5 marks)
11. Given the varibles Site, Layer, Temperature, and Depth - what would the best linear model be to predict Dissolved Oxygen (DO)? Use your knowlege of model fitting, model comparison, and statistical interpretation to build a model from scratch. State the linear model type as part of your answer. For this question, use only the subset of data with the relevant data for which there is no missing data and no duplicated data. (8 marks)

5 marks will be given according to the following rubric:

- 5.0 - code is well-documented and concise
- 2.5 - code is either well-documented or concise, but not both
- 0 - no attempt was made to document code, extra variables are created, code is difficult to read

Total: 45 marks

Submission: Each student will upload a .R file to Quercus. Please include your first and last name, the date of submission, and the assignment number.

Due date: 11:59pm February 27th, 2019

The late penalty for the final project is 5% for each day the assignment is late up to a maximum of 5 days, after which a mark of 0 will be given.