hw8

Hello and welcome

This homework will be an opportunity to practice common dplyr functions for working with and analyzing data frames. It will take the form of several analysis questions - with specific requirements that will test your knowledge of R and dplyr.

Important notes before you begin:

1. Your final output will be a PDF or html file (you pick) that documents both your code and analysis of this data frame.
2. For this hw, you will only need to import the `dplyr` package only - *NO* other packages are needed.
3. We will be working with the same `hw8_data.csv` file from the in-class workshop (from the `hw8` directory).
4. Use Markdown to explain certain outputs and document your code with comments too.
5. Bonus points if you style your tables (for better readability) in your output html/PDF files.
6. Keep in mind that some of these questions are worded in a such way to mirror questions that non-bioinformaticians may ask you. Try to apply the `dplyr` verbs to help you answer the question.
7. Save your code on Github. Practice committing and pushing your code to Github.
8. Good luck :)

Questions

1. How many tissue samples are there by donor and what is their disease status?
    - Display your answer as a data frame.
    - Disease status should be ordered as follows: Normal, Potential, Affected.
2. What is the average AND standard deviation values for `CD45+` and `PBMCs (%CD45)` when grouped by disease status and tissue?
    - Display your answer as a data frame.
    - Disease status should be ordered as follows: Normal, Potential, Affected.
3. What is the mean difference between the Eosinophil populations between the different disease statuses, irrespective of tissue?
    - Display your answer as a data frame.
    - Disease status should be ordered as follows: Normal, Potential, Affected.
4. When stratified by tissue and disease status, which donors have the highest `CD8 Tem CD69+` proportions?

   - Display your answer as a data frame.
   - Disease status should be ordered as follows: Normal, Potential, Affected.

5. What are the average AND standard deviation values for all columns that start with “CD4 Tem” when grouped by disease status and tissue?

   - Display your answer as a data frame.
   - Disease status should be ordered as follows: Normal, Potential, Affected.
   - There are 26 columns that start with “CD4 Tem”. You should not be manually specifying “mean(‘CD4 Tem blah blah‘)” for each of those 26 columns. `dplyr` has additional ways to solve this.