Assignment 3

Please note that due dates can be found in the Syllabus; submission instructions can be found on the Assignment Instructions page. In this assignment, you can submit a Google Doc (or other text editor, pictures, etc.) but also your `R` code via Google Drive. Aki will go over the submission process in the lab.

You might consider (but it is not mandatory) using R Markdown to write your answers.

50 total marks.

This assignment focuses on wrangling the NCBI Taxonomy database into R and stored as a tibble. I would recommend that you start by becoming familiar with the database and how it is organized here.

You will need to show your code and output for each step throughout the following parts. Sometimes screen shots can be used if you prefer.

Part 1 [2 marks] Download `taxdmp` directly to RStudio Cloud and uncompress it if necessary. Follow good practices and store this information in a folder `raw` as done in the lecture.

Part 2. [2 marks] The `readme.txt` file describes each file in the download of `taxdmp`. We do not need to consider some of these files. We will just focus on

`nodes`, `names`, `division`. Show how to read each of these into their own tibble.

Part 3. [2 marks] Some variables of these three tibbles are no longer necessary because we don't consider here `gencode`, `merged`, `delnodes`, `citations`. Remove these columns from the three tibbles. In addition remove any column with the term `flag` in it and `mitochondrial genetic code id`. We don't need these.

Part 4. [3 marks] For each tibble, is it tidy? Why or why not?

Part 5. [3 marks] For each tibble, are the classes of the individual columns appropriate? Why or why not? If they are not, show how to correct this.

(There is no single right or wrong answer for Parts 4 and 5. It depends on your design choices. You need to justify in point form your answer.)

Part 6. [4 marks] Show how to join the `nodes` and `names` tibbles. What is your primary and foreign keys? Which join function did you use and why?

Part 7. [6 marks] Show how to join the tibbles for Part 6 with the `division` tibble. What is your primary and foreign keys? Which join function did you use and why? Why or why isn't joining the `division` tibble with the tibble from Part 6 a good idea?

Part 8. [6 marks] Write a function that accepts as arguments the tibble from question 6 and the name of a taxon. It should return the `tax_id`.

Part 9. [3 marks] Write a function that accepts as arguments the tibble from question 6 and a `tax_id` for a target taxon. The function returns a tibble consisting of all the direct children of target.

Part 10. [6 marks] Which domain (Eukaryota, Bacteria, Archae) has the most species, and how many of them are there for each? What about genera?

Part 11. [6 marks] Write a function that accepts as arguments the tibble from question 6 and a `division`. The function shoudl return a tibble consisting of taxa in that division.

Part 12. [7 marks] Write a function called `path_to_root` that accepts as arguments the tibble from question 6 and a `tax_id` for a target taxon. The function returns a string vector whose order is the path of that taxon to the root of the tree of life. For example,

```
#print( path_to_root( tax_id=632,
         taxonomy=part6_tibble) )
#    root.cellular organisms, Bacteria,
         Proteobacteria, Gammaproteobacteria,
         Enterobacterales, Yersiniaceae,
         Yersinia, NA, Yersinia pestis
```

Note that some nodes along the path from the taxon to the root may not have a name. Your function should print out `NA`

Good luck!