Assessment 3 - Applied Data Science

Instructions:

This assignment explores a student academic performance data set. The performance metrics come from a learning management system (similar to our moodle). You can read more about the data set *here* where you will also find documentation about the 16 variables in the data set. If you scroll down the page to the data you can download it, or you can download it from moodle. Once downloaded you will need to move it to to your working directory, or somewhere you know the file path to before you can load it.

The documentation of the dataset can also be found in the text file "Edu-Data Documentation" in the folder for assignment 3.

I suggest that you utilise the RStudio cheat-sheets and R4DS textbook significantly.

When you are using data frames to determine values you must show the code to create the data frame and leave a comment below the code with your answer, and what your rationale was to get there if the code/data frame does not show it explicitly (comment using "#" then type).

Using pipes is preferred, but creating objects in the environment is acceptable.

The weight of each question is shown next to it, and the whole assignment is worth 25% of the final grade.

Due: 22/06/23 2330 NZT

Tasks:

1) (12.5%) Download and import the data set, call it *education_data_raw*

2) (25%) Test if there are any missing values in the data set, if you find any, remove them. Then, tidy the data so that: * "PlaceofBirth" is called "place_of_birth" and only the first letter of each entry in this column is capitalised * Repeat the previous step for "StageID" (call it stage_id, and change the capitalisation)
   * Change the names of "NationalITy" and "VisITedResources" to snake_case
   * Remove any duplicate columns resulting from the previous changes so that only the improved columns remain
   * Change gender to a factor
   * Call this tidied data set "education_data"

3) (12.5%) Group the data by grade and topic then determine which group has the highest total hand raises. You will first need to determine the total hand raises in each group.

4) (12.5) Filter the data to only include 8th graders and then explore how subject choice relates to average discussion participation. What is the average discussion for each topic in this grade, AND how many students are in the group?

5) (12.5%) Create a linear model to determine how raisedhands impacts discussion group participation. Then pipe the model to a BaseR plot it using "plot()".

6) (25%) New Zealand classification
   * Write a csv called "nz_classifications" which converts the grades in the dataset to how we classify schools in New Zealand (Primary, Intermediate and High)
   * Add/join this to the existing "education_data" * Create a publication quality graph using ggplot to show the most popular topics in each NZ school type

   - Facet the graph
   - Apply appropriate theme
   - Title
   - Labels
   - Caption with the reference for the data