Midterm Project 1

MTH 3270 Data Science

Due Wed., Mar. 18

Rules: You must do your own work, and you're only allowed to speak about this project with the instructor (Grevstad) and your partner from the class (if you will be working with one).

If you will be working with a partner, you must notify the instructor (Grevstad) via email *prior* to beginning work together. The two of you will only submit one project (with both your names on it).

The projects are **due** in **hard copy** (paper) format no later than **Mar. 18, 2020** at **6:00 PM**.

Instructions: Check your **email** immediately for the **data sets** and also regularly during the week in case there are important announcements about this project.

The project will use the following data sets (in your **MSU Denver email**):

API_4_DS2_en_csv_v2_820954.csv
Metadata_Country_API_4_DS2_en_csv_v2_820954.csv
Metadata_Indicator_API_4_DS2_en_csv_v2_820954.csv

The first one is the main data set, and contains data on education for each of 264 of the world's countries for the years 1960-2019. The second contains information about each country. The third contains information about the variables recorded for each country.

The data were obtained by going here:

`https://data.worldbank.org/topic/education`

and then clicking on "CSV" under "Download" on the right.

You will need to either manually **delete** the **first four rows** of **API_4_DS2_en_csv_v2_820954.csv** *or* use the `skip = 4` argument in `read.csv()` before reading the data into R (and don't forget `header = TRUE` and `stringsAsFactors = FALSE` in `read.csv()`).

You will need to do some **data wrangling** and **tidying** (which *may* involve combining data sets, selecting columns, adding new columns, filtering rows, grouping by a categorical variable, converting between wide and narrow formats, etc.). All **data wrangling** and **tidying** must be done in R (except by permission of the instructor).

Your **tasks** are:

1. Pick one of the **years** from 1960 to 2019 (`X1960`, `X1961`, ..., `X2019`) and one of the **variables** (`Indicator.Names`). Then either:

   - Summarize that variable (in that year) **by Income Group** to compare values of the variable across different **Income Groups** of countries,

     or

   - Summarize that variable (in that year) **by Region** to compare values of the variable across different **Regions** of countries.

2. Create one or more visualizations (graphical displays) for comparing the variable you chose (in the year you chose) for the different **Income Groups** *or* **Regions** of countries.

3. Pick one of the **variables** (`Indicator.Names`). Summarize that variable for each of the **years** from 1960 to 2019 (`X1960`, `X1961`, ..., `X2019`) to compare the values of the variable across different **years**, i.e. to look for any **trend** in that variable. The yearly summary you use in the trend analysis may be based on *all* countries or just a particular **Income Group** or **Region** of countries.

4. Create one or more visualizations (statistical graphs) of the yearly summary values for the variable you chose, i.e. to visualize whether there's any **trend** in that variable.

What to turn in: A **write-up** (perhaps 2-5 pages including graphs) consisting of:

1. A **brief description** (e.g. 1-3 paragraphs) of any data wrangling and tidying you had to do in order to answer questions **1-4** above.

2. For questions **1-2** above, indicate **which year** you chose to focus on, and **which variable** you chose to compare across the **Income Groups**, and then state the results of your comparisons (including the graphs).

3. For questions **3-4** above, indicate **which variable** you chose to compare across the **years** (i.e. to investigate for a trend), and then state the results of your comparisons/trend investigation (including the graphs).

4. Your **R code** with **comments** (use `#`) indicating **what** each chunk of code does and **why** it does it. The instructor (Grevstad) may request an electronic copy of your **R code** in a **.R file** (as produced by RStudio's editor), so please hold on to it.

Suggestions:

1. Try using one of the `*_join()` functions (from `"tidyr"`) to combine the **Income Group** and **Region** information from **Metadata_Country_API_4_DS2_en_csv_v2_820954.csv** with the data in **API_4_DS2_en_csv_v2_820954.csv**.

2. You can remove unneeded columns (variables) using `select()` (from the `"dplyr"` package).

Grading: Your grade will be based on:

1. Your attainment of **tasks 1-4** above.

2. Your **write-up**, including the graphs (as described above).

3. The inclusion of and correctness of your **R code**.