Homework 2: Introduction to Statistical Computing Using R

STAT 42100: Modern Statistical Modeling Using R and SAS
STAT 52100: Statistical Computing

Instruction: Collaboration on homework assignments is acceptable, but all write-ups must be done independently and clearly indicate the submitter's understanding of the material. Unclear or disorganized homework may have points removed, even if the content is correct. **No hardcopy of the typed report needs to be turned in. The pdf file should be submitted electronically on canvas.** An assignment handed in after the deadline is late, and may or may not be accepted. My solutions to the assignment questions will be available when everyone has handed in their assignment.

Edit the program(s) and output together into a single document, showing the lines of code and relevant output produced by SAS/R. Your answers must be easy for the grader to find. A simple structure is, for each part of each question in order, to put these three things:

- Your code
- Your output
- Your answers and explanation

For SAS, the code and output are naturally separate, so this order is good; for R, the code and output can be intermingled, and that is OK. If your assignment is disorganized or otherwise difficult for the grader to deal with, you can expect to lose marks.

You are reminded that work handed in with your name on it must be entirely your own work. It is as if you have signed your name under it. If it was done wholly or partly by someone else, you have committed an academic offence, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The graders will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you will do badly on the exams.

Problem This problem is based on Hans Rosling talks New Insights on Poverty and The Best Stats You've Ever Seen. The assignment uses data visualization to gain insights on global health and economics. We will use the following datasets: `children_mortality.xlsx`, `life_expectancy.xlsx`, `fertility.xlsx`, `population.xlsx`, `continent-info.tsv` and `GDP.xlsx`.

1. Create five `tbl_df` table objects, one for each of the tables provided in the above files. Hints: Use the `read_excel` function.

2. Write a function called `my_func` that takes a table as an argument and returns the column name. For each of the five tables, what is the name of the column containing the country names? Print out the tables or look at them with View to determine the column. And assign a common name to this column in the various tables.

3. Notice that in these tables, years are represented by columns. Create a tidy dataset in which each row is a unit or observation and our 5 values of interest, including the year for that unit, are in the columns. Hints: Use the `gather` function from the `tidyr` package and `full_join` from the `dplyr` package.

4. Add a column to the consolidated table containing the continent for each country. Hint: Use the file `continent-info.tsv` that maps countries to continents here. Hint: Learn to use the `left_join` function.

5. Report the child mortalilty rate in 2015 for these 5 pairs:

   ```
   Sri Lanka or Turkey
   Poland or South Korea
   Malaysia or Russia
   Pakistan or Vietnam
   Thailand or South Africa
   ```

6. Use `ggplot2` to create a plot of life expectancy versus fertility for 1962 for Africa, Asia, Europe, and the Americas. Use color to denote continent and point size to denote population size.

7. Learn about OECD and OPEC. Add a couple of columns to your consolidated tables containing a logical vector that tells if a country is OECD and OPEC respectively. We use the membership on 2015:

   ```
   oecd <- c("Australia","Austria","Belgium","Canada","Chile",
             "Country","Czech Republic","Denmark","Estonia",
             "Finland","France","Germany","Greece","Hungary",
             "Iceland","Ireland","Israel","Italy","Japan",
             "Korea","Luxembourg","Mexico","Netherlands",
             "New Zealand","Norway","Poland","Portugal",
             "Slovak Republic","Slovenia","Spain","Sweden",
             "Switzerland","Turkey","United Kingdom","United States")
   ```

```
opec <- c("Algeria", "Angola", "Ecuador", "Iran", "Iraq",
          "Kuwait", "Libya","Nigeria", "Qatar", "Saudi Arabia",
          "United Arab Emirates", "Venezuela")
```

8. Make the same plot as in Problem above, but this time use color to annotate the OECD countries and OPEC countries. For countries that are not part of these two organization annotate if they are from Africa, Asia, or the Americas.

9. Explore how this figure changes across time. Show 4 figures that demonstrate how this figure changes through time.

10. Let's compare France and its former colony Tunisia. Make a plot of fertility versus year with color denoting the country. Do the same for life expecancy. How would you compare Tunisia's improvement compared to France's in the past 60 years? Hint: use `geom_line`.