Graded Lab Assignment 3: Programming in R

Please refer to the following ebook which is available in the public domain:

https://www.cs.upc.edu/~robert/teaching/estadistica/rprogramming.pdf

Please read through and understand the chapter on the case-study titled **"Data Analysis Case Study: Changes in Fine Particle Air Pollution in the U.S."** on Page 131. This chapter essentially discusses the changes in fine particle outdoor air pollution in the US during 1999-2012.

This graded lab assignment is designed to be a capstone mini-project for you to use your learning from the non-graded assignment on R towards reading, manipulating and performing visualizations on **real-world datasets** using R.

Consider the following URL as your data source:
https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html
(Citation: US Environmental Protection Agency. Air Quality System Data Mart [internet database] available at http://www.epa.gov/ttn/airs/aqsdatamart. Accessed March, 2017.)

In the above URL, please click on the **File Download** link.
http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

In the above webpage, click on the link **"Table of Annual Summary Data"** and you will see annual summary data for years between 1990 and 2016. Now click on annual_all_2016.zip (2016 annual data), download and unzip the file and you will see an Excel file with a large number of columns. Note that you are NOT required to have any prior domain knowledge about all the attributes (columns) in this Excel file. Carefully decide the attributes that you want to consider in your analysis (consider only those attributes, which you have some idea about e.g., arithmetic mean, state, county, city names etc.).

Required:

1) Read in the 2016 annual data from the Excel file and after performing analysis on the data, create an effective visualization to depict air pollution across different states. Within each state, you can go more fine-grained such as county/city level. Your visualization should be clickable as much as possible. For example, a user should be able to zoom into a given state and look at the air pollution across different counties in that state; then for a given county, one should be able to zoom into the county and observe the air pollution at different cities.

2) Repeat the above for multiple years (at least 5 different years; ideally, space out the years e.g., 1990, 2000, 2005, 2010, 2016 etc.)

3) Create effective visualizations for changes in air-pollution across time for different states. Here also, you should additionally create a clickable interface for counties and cities as described in step 1.

Deliverables:

- You need to upload your completed codes on Blackboard by **April 7, 2017 at 3 pm**. This is a hard deadline and will NOT be extended.
- You are also required to provide a demo of your program.
- You need to write a document (12 point Times New Roman font, 1.5 line spacing, maximum of 3 pages) to explain the results of your visualizations in a concise manner. *In this document, you can also discuss which regions can be clustered based on air pollution and if you found any*

outliers and/or discovered any interesting patterns in the data. ***In essence, you should use this document to summarize your interpretation of the results.*** You should also submit this document (in PDF format only) via Blackboard.

Please note the following points:

1. This lab assignment will contribute to **10% of your grades** for the course.

2. Please write the Names and Roll Numbers of your group members in the Comments section at the beginning of your program file. All programs should be submitted on Blackboard on or before the stated deadline. Please note that your programs may be subjected to plagiarism checks.

3. The **deadline** for assignment submission is **April 7, 2017 at 3 pm**.

4. This is a **HARD deadline** and no points will be awarded for the assignment if you submit after the deadline, unless there are extenuating circumstances.

5. The grading criteria for this assignment will be based on effort, adherence to learning points from your previous non-graded assignment, code quality, visualization, results and scalability.

6. Any act of plagiarism will result in a zero for the entire assignment. Hence, please avoid any form of plagiarism.

7. This is a group assignment, hence please do NOT collaborate with your fellow students in other groups towards the completion of this assignment. However, you are obviously required to collaborate effectively with members of your own group to ensure that you are able to function as a team player.

8. If you encounter any problems with downloading the datasets, please contact your TA (Saurabh Mishra).

This assignment carries 10 points.