DAT 301 Project 1

For Project 1 you need to choose a topic that requires either a data manipulation or a simple statistical or machine learning technique that should be equivalent in load to about 1.5 - 2 labs or 1.5 - 2 homework assignments.

- Don't forget to submit all the files you used, including source code (.R), Rmarkdown that produces pdf and/or slides, data files, possible images, etc.
- Do **NOT** zip the files into a single folder. Submit all the files unzipped.

Steps and Requirements

- Choose a data set, define a problem (or problems) you want to solve
- Perform some exploratory data analysis and if appropriate, apply some statistical procedures to answer your questions. In your report (see below), you will need to provide the code, explain what you have done, and what you have found or what your conclusion is.
- Use Rmarkdown to create either a pdf report or html ioslides. You can interweave the (textual) discussion with code chunks and their output. Explain which data set you used (what the resource is and what the variables are). Formulate the problem(s) and explain how you tried to answer your questions and what the conclusion is. Examples of projects made by students in earlier semesters are provided in Canvas course shell.
- Give a presentation. A part of the lab is practicing presentation of your analysis. **Record a short video** (≈ 10 min) with a brief introduction to the dataset (**including its source**) and your problem, as well as what analysis you have done and some findings or conclusion. You can use Zoom or any other way to record your voice (sound) and the screen on which you can show your report or slides and go over them. Do NOT pay attention to the quality of the video or editing - that's not important. Just make sure that your voice could be heard and understood, and the text on the screen could be read. The video should be in mp4 or avi, or some other common format. Submit it in Canvas, with other files. If you need more than 10 minutes, it's okay to make a bit longer video.
- You **MUST** submit all the files you used (including source code (.R), Rmarkdown that produces slides and/or pdf, data files, possible images, etc.)
- Do **NOT** zip the files into a single folder. Submit all the files unzipped.

Where to Look for Datasets

If you used some dataset in your HW3 or Lab3, it is okay to make this project as an extension of the previous assignment, as long as you meet the aforementioned steps and requirements.

Otherwise, you can look for a dataset wherever you want. The file `opendatasites.csv` with the list of links where you can search for datasets is provided in Canvas. Geographical notions in this file to which the links are related include regions, states, countries, cities. In addition, here are two popular sites with a lot of datasets:

- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.php

Grading Policy

You will be assessed on the quality of the written report as well as the presentation in the following way:

1. Background and problem definition (5pts)
2. Data wrangling, munging and cleaning (20pts)
3. Exploratory Data Analysis (30pts)
4. Data Visualization (15pts)
5. Final paper/slides and presentation (30pts)

IMPORTANT: IF YOU CHOOSE A TOPIC OTHER THAN SUGGESTED ABOVE, YOU MUST BE ORIGINAL! COPYING PROJECTS FOUND ONLINE IS PLAGIARISM AND ACADEMIC DISHONESTY (CHEATING!) WHICH VIOLATES ACADEMIC INTEGRITY, AND WILL RESULT IN FINAL GRADE OF E!

For information on academic integrity, including the policy and appeal procedures, please visit http://provost.asu.edu/academicintegrity

Reminder

- Don't forget to submit all the files you used, including source code (.R), Rmarkdown that produces pdf and/or slides, data files, possible images, etc.
- Do **NOT** zip the files into a single folder. Submit all the files unzipped.