Final Project
STAT / BIST 5225 - Spring 2021

Due on Sunday, May 2, 2021

The data:
For this project use the "*Mortality Multiple Cause Files*" from the CDC:
https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Mortality_Multiple
Use the U.S. Data, which is provided as zip files and is available for the years 1968-2019.
Start by reading the User's Guide:
https://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm

The objective:
To write a research paper on mortality in the U.S., which utilizes the dataset and the tools taught in the class.

Instructions:

- Use either R or SAS to analyze the U.S. mortality causes dataset. You may use SQL to store the data. Do not use languages not covered in the class.
- It is prohibited to consult with other people – you must complete this project on your own.
- Using solutions found on the Internet is not allowed. You can use documentation and relevant tutorials and examples, but do not copy existing code.
- Any source that you use (tutorials, blogs, articles, etc.) must be cited a References section in your paper.
- Your code must work and must be well documented. If any packages are needed, your documentation must state clearly what is needed, and how to install external components.
- Submit your paper in Word or pdf format. Your paper has to include a description of what you did in terms of data management and your observations and results. Your paper should not include code. Only verbal explanations, plots, tables, and references.
- Submit your code and paper separately. The file names should contain your last and first name and the type of file. For example, BarHaimSource.R would contain the source code in R, and BarHaimPaper.docx or BarHaimPaper.pdf will contain the paper.
- Do not submit large data sets via HuskyCT! You may provide access to your data via github. The dataset which you create for this project must also be reproducible with your code.
- The maximum number of pages in the final paper is 15, including a cover page, graphs, tables, references. In Appendix A, include a clear description of your dataset. Additional figures, tables, and results may be provided in a separate appendix.
- Your paper's cover page must include a title, author name, date, and a short abstract which states the main results of your analysis.

- Figures and tables must be in high-quality. Do not paste screenshots.
- Proof-read your paper and make sure it contains no typos or grammatical errors.

The research objective of the project is open-ended. The dataset is very detailed, so there are many possible research questions. Choose questions which you think are interesting and be sure to explain why investigating these questions is important.

You do not have to use all the data (from 1968). You should create a subset of the data which is sufficient for your analysis. The steps needed to obtain your subset should be well documented and reproducible, so that if others follow them precisely, they will get the exact same dataset. You may augment the data with information from other sources (e.g., geographic data, demographic data, etc.)

Deliverables and due dates:

1. **Project plan** – a one-page description of your proposed research questions and your work plan. This is **due by March 14**, via HuskyCT. Do not start working on your actual project before you get my approval for your plan!
2. **Final project** – **due by May 2**, 2021 via HuskyCT. Submit one docx/pdf file with your paper, and an R and/or SAS file with your code
3. **Optional** – class presentation. If you want to present your paper in class, please let me know by April 10. I will reserve time slots for presentations on April 20, 22, and 27 (during normal class hours). Presentations should be up to 15 minutes long.