BSTA001: Population Health Data Science - I Final

The Goal

For your final project you will propose your own hypothesis, collect an open source dataset, analyze the data, summarize the results of your analysis, and discuss how the results either support or disprove your hypothesis.

I expect a brief review of some literature or background information on the topic and, most important, I expect you to draw on what we've learned in class. At minimum, your project should include Python3 code that runs an exploratory data analysis, studies the data in the context of random variables, and reports (informally, formally, or both) the results of statistical computations.

I also ask you focus on interpreting your results for a lay audience who may not necessarily understand statistics.You are welcome to work in groups of 4. If you decide to work in groups please email me who you are working with.

Deliverables for the project are a **single** Jupyter notebook.

Data and your topic:

The topic you pick is up to you! Pick a topic you are passionate about, or a topic you are interested in learning about.

Progress Dates (Suggested):

2021-04-22 - Please send 2-3 topics you are interested to study and potential datasets that you can use for your project.

2021-04-30 - Please send a brief email about your progress. You should have imported your dataand begun at least one visualization to support or refute your hypothesis.

2020-05-07 - Please send a second brief email about your progress. You should have run a statistical hypothesis test (z-test, t-test, chi-square, etc) to support or refute your hypothesis.

2020-05-14 - Please send a second brief email about your progress. You should have the methods and results written, and be in progress with writing the introduction and discussion.

Due Dates (Not Suggested):

2021-05-17 - Final Project due.

Guidelines for the primary analysis

Your final project write-up should be written using a jupyter notebook that combines text together with code written in Python3. Any code that is written should be described in the code block itself, using #comments, or before the code block.
Please follow the structure below for your write-up:

Title: The title of your project and a list of student's names.

Summary: an introduction to your project, brief description of statistical concepts you used to study the problem, and a summary of the results and conclusions (About a 1/2 page).
Data: a brief summary of key features of the dataset. You should define each variable that will be used, how the data was collected, your target population and observations, and whether the sample is representative and ways it could be biased (About 1 page).

Methods: A description of the statistical tools you used for analysis. At minimum a description of any summary statistics (mean, median, variance, etc), a description of one exploratory data analysis, and a description of the statistical hypothesis test used to support or refute your hypothesis. (about 1/2 page)

Results: A presentation of your results. Including at minimum on exploratory data analysis plot (histogram, scatterplot, boxplot, barplot, etc) and at leaste one formal analysis of a random variables (hypothesis test) that supports or refutes your hypothesis. (About 1 page)

Discussion: Summarize your work, its limitations, and possible future steps/improvements.

References: Cite all sources in a standard format.

You should use 11 or 12 point font and no less than 1 inch margins all around.

Proper Citations

Any direct quotes from other sources should be minimal (a sentence or two, not multiple paragraphs!), should be between quotation marks, and should be cited at the point of quotation. It's ok to include a figure from another source, but this should be cited in the figure caption.