General Instructions

The purpose of this group project is to analyze a data set, using Python and/or SQLite to process the data and to generate graphs, figures, and visualizations as necessary, with the ultimate goal of producing a single written report and presentation which presents your results and explains their significance.

You may work in groups of two or three for this project. You should submit the code that you wrote (there may be multiple .py files), a written report of your results including the graphs that your code generated, and a small statement of each group member’s contribution to the project. The deadline for turning in all files is 11:55pm on the due date.

The report should include the following:

1. **Overview**: A description of what this section of the analysis is about. Include information about the dataset(s) and where they were obtained, and any special preparation needed in order to use/analyze the data.

2. **Methodology**: How you performed the analysis.

3. **Results**: The results that your analysis produced.

4. **Discussion**: Interpreting the results.

The above is a general outline of what you should present, but you are not required to “number” or outline your report using these terms. Just be sure that it is clear that you are providing the required parts.

Do not be concerned with trying to figure out a single “right” answer. You are not being graded on whether or not you come up with a specific analysis, but rather how correct your methodology is and on the quality of execution.

For all written work (i.e., the reports, not the .py files), you may use any word processor that you like, but please make sure to export the report as a PDF, so that there will not be any software compatibility issues when grading.
Overall Goals

Finding Interesting Data

For this project, you will need to choose one or more datasets to use. While you may source your data from any reputable source, some suggested sources are:

- [https://toolbox.google.com/datasetsearch](https://toolbox.google.com/datasetsearch)
- Also see the comments here: [https://news.ycombinator.com/item?id=17919297](https://news.ycombinator.com/item?id=17919297)

Your goals are as follows:

1. Identify an appropriate dataset (one or more) that can be analyzed for non-trivial information.

2. Prepare visualizations of **five** or more types of observations. You may have multiple visualizations for a single observation (if it makes understanding the data clearer). For clarification, if you have two graphs, one for the number of male smokers and the other for female smokers, then that only counts as one type of observation (data on smokers). It’s like double-spacing your paper. It takes up more room, but you didn’t provide anything of substance.

3. See if you can make your observations follow a theme of some sort or have some other relationship. Ultimately, you are trying to build an informed, nuanced understanding of the data.

4. Do not misrepresent the data. Do not sensationalize the data (*e.g.*, no “clickbait” claims). Hopefully you will find something interesting, but that is not guaranteed.

5. Try to combine data in interesting ways. This may be joins across related tables, or combining data from multiple datasets.

**Written Report: Due May 1, 2019**

Provide the written report as described in the General Instructions section. Although it is stated to have **five** types of observations, only a minimum of **three** are required at this point. The additional observations may be added for the final presentation.
Group Presentation: Due May 6, 2019

The final exam time for this class has been scheduled by the Registrar to be Monday, May 6th, from 4:15pm to 6:15pm.

At this time, each group will give a short, 5-minute slide presentation to the class about their dataset and their results. This does not need to be limited to the written report. That is, visualizations may be improved and new analysis, if discovered, may be added.

A copy of the slides and any updated materials, if any, should be submitted online.