MidTerm Project

About This Project

The mid-semester project is a core component used to assess your master of the course content in CS 181 / DA 210. This is an individual project. In this project, you'll compile three to four data sets from the web and attempt to use these data sets, along with the tools and techniques we are developing in this course, to answer a central question of your choosing. **The topic, central question, and data sets you choose are entirely up to you.**

Objectives

The objectives of this project include:

- working with a real-world dataset, and one that has greater volume and scale than we have seen;
- going beyond the "building-block" mentality that comes with the small data and limited scope of homework sets;
- building a larger whole, synthesized from many different skills learned so far over the semester;
- thinking about the data sets themselves, and the information the study of those data can provide;
- effectively communication what was learned.

Requirements

The methods and tools you use to perform the analysis towards answering your central question must satisfy the following requirements:

1. Requirement #1: Your data sets should be tabular data, likely .csv files similar to those we're using in class.
2. Requirement #2: **One of your data sets should be read in and used to construct a DoL, and another must be read in and used to construct an LoL. These data structures can then be used to construct pandas DataFrames. Note that all remaining tables can be directly converted into DataFrames.**
3. Requirement #3: You must convert all DataFrames to be tidy, and clearly state in your writeup the mappings of independent to dependent variables.
4. Requirement #4: You must use pandas and DataFrames in some meaningful way.

Data

You should identify multiple data sources (at least three or four) that can be expressed in the tabular model. These data sources may be .csv files, files on the web, or any other data source that you learn how to access. These data sets should be picked with the goal of answering your central question, which should only be able to be answered by combining data from the various data tables.

Project Deliverables

This project will be broken into three deliverables.

Deliverable #1: Proposal (Due 9/19)

A short mid-semester project proposal is due **by 11:59pm on Monday, September 19th**. The proposal should do the following in a single document (.pdf file):

- Identify the central question you want to try and answer.
- Specify all the data sets that you have found (that can be read in tabular form) and which you intend to use for this project.
- Give an outline of how you intend to use these datasets to answer your central question. This should include a list of functions that you plan to write. If you do a good job of breaking down the task into smaller functions, you should be able to describe in a short sentence what each function will accomplish. I will give you feedback about the complexity of your chosen project and whether you will need to make it simpler or more complex. If you are working with a partner, you need only submit one proposal.

Deliverable #2: Data Parsing (Due 10/3)

For the first half of the project work itself, you will parse the data into a single Google Colab notebook, completing Requirement #2. This part of the project, due **by 11:59pm on Wednesday, October 3rd**, should include the following:

- Code to read in one dataset and use it to build a DoL, which should then be used to construct a pandas DataFrame.
- Code to read in another dataset and use it to build an LoL, which should then be used to construct a pandas DataFrame.
- Code to read in any remaining datasets directly as pandas DataFrames.

Your notebook should be self-documenting, with lots of Markdown (text) cells explaining the datasets you have identified, and how the functions you write enable you to parse the data. For both this deliverable and the next, your notebook should have all cells runnable; **errors should not occur during processing**.

Deliverable #3: Data Cleaning and Processing (Due 10/12)

In the final part of the project, you will clean up your data and use to answer your central question. This part is due **by 11:59pm on Wednesday, October 12th**, and should include the following:

- A markdown cell describing the mapping of independent to dependent variables.
- Code to make all data tidy.
- Markdown cells explaining the central question, and your answer(s) to that question, in the form of output values, graphs, tables, or any other format that is appropriate.

Your notebook should now include Markdown cells discussing how you have made the data tidy and the process by which you have answered the central question you posed.

Project Assessment

Each deliverable in your project will be graded separately, with the overall project score becoming a combination of those for the deliverables.

Deliverable #1	25%
Deliverable #2	40%
Deliverable #3	35%