Mid-Semester Project

DA 210-02 / CS 181-02: Data Systems
Spring 2023
Instructor: Dr. Tanya Amert

About This Project

The mid-semester project is a core component used to assess your master of the course content in DA 210 / CS 181. You must complete the project with a partner.

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; and
- effectively communication what was learned.

A secondary objective is to give a "preliminary run" for a synthesis project like this, so as to be better prepared for the final project at the end of the semester.

Requirements

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

- **Requirement #1:** Your data sets should be tabular data, likely .csv files similar to those we're using in class.
- **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.
- **Requirement #3:** You must convert all DataFrames to be tidy, and clearly state in your writeup the mappings of independent to dependent variables.
- **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

A short mid-semester project proposal is due by **11:59pm** on **Wednesday, February 8th**. The proposal should do the following in a single document (e.g., .docx or .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.

You and your partner will submit one proposal together.

Deliverable #2: Data Parsing

For the first half of the project work itself, you will parse the data into a single Jupyter notebook, completing Requirement #2. This part of the project, due by **11:59pm** on **Wednesday, February 15th**, 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 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.

Despite working with a partner, *you must submit separate lab reports*. The code in the two reports can be identical, but the writing and exposition should be your own.

Deliverable #3: Data Cleaning and Processing

In the final part of the project, you will clean up your data and use to answer your central question. This part is due at **11:59pm** on **Wednesday, March 1st**, 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. **To receive a score of "E" on the overall project, you must earn a grade of at least "M" on the first two deliverables, and an "E" on the third deliverable.**

Just to emphasize: for this mid-semester project (as well as the final project), you will be evaluated both on how you do your work and on how well you communicate your results. A project that just hacks together some monolithic or ill-structured code to "get an answer," or does a poor job expressing how the output of the code addresses your central question, will not receive a good grade.

Deliverable #1: Proposal

You can earn a grade of "M" on the proposal by meeting the following criteria:

- The central question is identified.
- The data files are included with the proposal document, and are properly cited in the document itself.
- Each data set is described in the proposal document, including the column names and data format, and the number of rows.
- The steps that you will use to answer the central question are clearly laid out.

You can earn a grade of "E" on the proposal by meeting the following additional criteria:

- You provide function signatures and descriptions (e.g., "`computeAverageAge(ageTable) -> float` computes the average age using the 'Age' column in the given table, and returns that value as a float) for several functions you will write to use in answering your central question.
- Your list of steps that you will use to answer the central question are clear enough that, along with a clear description of the data files, another student could reasonably complete your project based on your proposal alone.

You will earn a grade of "R" on the proposal if you do not meet the criteria for an "M" but have shown a good-faith effort in completing the proposal document. Failure to submit a document, or a document that is lacking almost all criteria for an "M", will result in a grade of "N".

Deliverable #2: Data Parsing

You can earn a grade of "M" on the second deliverable (a single Jupyter notebook) by meeting the following criteria:

- You parse one of your data files as a DoL before converting it to a pandas DataFrame.
- You parse another of your data files as an LoL before converting it to a pandas DataFrame.
- Any remaining data files are read directly in as pandas DataFrames.
- Your notebook executes without any errors.

- Your notebook should contain Markdown cells introducing the datasets.
  - Provide citation(s) for the data, and a brief description for each.
  - Explain your central question. (You do not need to discuss how you'll answer it.)
- If you did not submit your data with the proposal, you must include it with this deliverable. (If you did submit your data already, you needn't include it again.)

You can earn a grade of "E" on the second deliverable by meeting the following additional criteria:

- Your data files are all in a folder labeled "data". This folder should be a sibling of your notebook in a tree representing the file system layout.
  - If you submitted your data with the proposal, you should not submit it now. (This is unnecessary space and time spent uploading/downloading what could be large files.)
- Provide additional information about your data sources:
  - Give context for the problem – why is this data interesting?
  - For each of the data sources, display part of the table, and explain the columns (you need only describe the columns you plan to use).
- If you encountered any issues while processing the data, explain them in Markdown cells (e.g., values that contain commas, cells that should be numbers but can't be parsed, etc.).

Despite working with a partner, *you must submit separate lab reports*. The code in the two reports can be identical, but the writing and exposition should be your own.

You will earn a grade of "R" on the second deliverable if you do not meet the criteria for an "M" but have shown a good-faith effort in parsing the data. Failure to submit a Jupyter notebook, or a notebook that is lacking almost all criteria for an "M", will result in a grade of "N".

Deliverable #3: Data Cleaning and Processing

You can earn a grade of "M" on the third and final deliverable (a single Jupyter notebook) by meeting the following criteria:

- You list the mappings of independent to dependent variables for your data in a Markdown cell.
- You successfully transform one of your datasets to be tidy, and make an attempt to transform all others to be tidy.
- Your notebook executes without any errors.
- You answer your central question(s) using output values, graphs, tables, or any other format that is appropriate.
- You include Markdown cells discussing how you made the data tidy and the process by which you answer the question.
- If you made any changes to your data during processing and did not resubmit the data files, they must be included with this deliverable.

You can earn a grade of "E" on the second deliverable by meeting the following additional criteria:

- Your identification of variables and classifications of independent and dependent is correct.
- All of your datasets are made tidy.
- Your use of pandas and DataFrames to answer your question is meaningful (i.e., more than just finding a cell in a table, or finding the min/mean/max of a single column).

Despite working with a partner, *you must submit separate lab reports*. The code in the two reports can be identical, but the writing and exposition should be your own.

You will earn a grade of “R” on the third deliverable if you do not meet the criteria for an “M” but have shown a good-faith effort in tidying the data and answering your question. Failure to submit a Jupyter notebook, or a notebook that is lacking almost all criteria for an “M”, will result in a grade of “N”.

Overall Project Grade

Your overall project grade is dependent upon all three deliverables.

- To earn a score of “E”, you must earn a grade of at least “M” on the first two deliverables, and an “E” on the third deliverable.
- To earn a score of “M”, you must earn a grade of at least “M” on one of the first two deliverables, and an “M” on the third deliverable.
- To earn a score of “R”, you must earn a grade of at least “R” on all three deliverables.