Final Project

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

About This Project

The final project is another core component that will be used to assess your mastery of the course content in DA 210 / CS 181. As with the previous project, you will work with one partner to complete your project.

In this project, you'll combine the major pieces of each unit from this semester—webscraping, JSON/HTML/XML parsing, database design and query access, and tabular data processing. The final product of this project is to produce two notebooks that comprise a single data story. As with the mid-semester project, you'll have to make two creative choices: (1) what data sources you will use, and (2) what central question you will seek to answer (i.e., what story you will try to tell with the data sources you choose).

Objectives

As with the mid-semester project, the objectives of this project include:

- working with a real-world dataset, and one that has greater volume and scale than we have seen in homework and software labs;
- 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 throughout the semester;
- thinking about the data itself, and the information the study of that data can provide; and
- effectively communicating what you've learned.

Requirements

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

- **Requirement #1:** You must combine data from at least two different sources. At least one of these must be an HTML website. Both sources must be hierarchical data.
- **Requirement #2:** At least one of your sources must involve some degree of web scraping. You should access the direct HTML file, or a JSON/XML format from a website. You must include the downloaded HTML file with your submitted notebook.
- **Requirement #3:** You must store your data in a database. You will have to design and create the database, including tables, columns, and how you'll populate rows.
- **Requirement #4:** You must query a subset of the data from the DB in Requirement #3.
- **Requirement #5:** You must use a `pandas DataFrame` in some way to process the data.
- **Requirement #6:** You must use markdown cells in your Jupyter notebook to balance your code with the appropriate description. For example:
    - What web sources did you use? What format was the data in? How did you access and parse the data?

    - Explain the database you designed. How does it adhere to the principles of good DB design?
    - Explain briefly the DB queries you used to access and the data frame(s) you used to process the data.
- **Requirement #7:** You must include at least one graph that depicts something related to your data story and central question, and a final markdown cell summarizing the results of your project.

Data

You should identify multiple data sources, at least one of which requires some degree of web scraping (e.g., a table on Wikipedia). These data sets should be picked with the goal of answering your central question, which should only be able to be answered by combining from the various data tables.

Project Deliverables

This project will be broken into three deliverables, with code spread across two Jupyter notebooks.

Deliverable #1: Proposal

A short final project proposal is due by **11:59pm** on **Wednesday, March 29th**. 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 and which you intend to use for this project. You must include link to these dataset in your proposal, but you do not need to download anything yet.
- 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/use to answer the question. 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/more complex.

You and your partner will submit one proposal together.

Deliverable #2: Data Acquisition

For the second deliverable, you'll create the first of your two Jupyter notebooks. This part of the project is due by **11:59pm** on **Wednesday, April 5th**. In a file called `data_acquisition.ipynb`, you should have the following:

- Markdown cells describing your central question and citing your data sources (Requirement #6).
- Web scraping for at least one of your data sources (Requirement #2).
- Markdown cells describing the process of accessing the data and parsing/scraping it (Requirement #6).
- Markdown cells explaining how you plan to use the data to answer your question.

Your notebook should be self-documenting, with lots of Markdown cells explaining the datasets you have identified, and how the functions you write will 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. ***Submission of reports with too-similar Markdown cells will be considered an academic integrity violation.***

Deliverable #3: Data Storage and Analysis

In the final part of the project, you will store your data in a SQL database that you create, and then use SQL queries of that data to answer your central question. This final part is due at **11:59pm** on **Wednesday, April 19th**.

You will modify `data_acquisition.ipynb` to add the following:

- Markdown cells describing your database design, including how it adheres to principles of good DB design (Requirement #6).
- SQL statements to create tables in your database (Requirement #3).
- SQL statements to populate the tables in your database using the data parsed from your data sources (Requirement #3).
- Markdown cells describing any challenges faced when storing your data in your SQL database (Requirement #6).

Additionally, you'll create a new notebook, `data_analysis.ipynb`, that includes the following:

- Markdown cells to summarize how the data is stored in the DB and the queries you will use to access that data (Requirement #6).
- SQL statements to query a subset of the data from the DB (Requirement #4).
- Code cells using a `pandas DataFrame` in some way to process the data (Requirement #5).
- Markdown cells explaining how you use the data frame(s) to process the data. (Requirement #6).
- A final Markdown cell explaining the answer(s) to your central question (Requirement #7).

Answer(s) to your question(s) can be in the form of output values, graphs, tables, or any other format that is appropriate. The notebook should conclude with a markdown cell with all the results. Your notebook should have all cells runnable.

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 earn 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 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 earn 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 sources are properly cited in the document.
- Each data set is described in the proposal document, including the column names and data format, and approximate data size (e.g., how many 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 discuss the “entities” in your datasets (approximately corresponding to the independent variable set for each table in tidy data).
- Your list of steps that you will use to answer your central question are clear enough that, along with a clear description of your data, another student could reasonably complete your project based on your proposal alone.
- You explain how you know that the data sources are eligible for scraping (this likely involves a discussion of the Terms of Service or a similar document for each data source).

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 in almost all criteria for an “M”, will result in a grade of “N”.

You and your partner need only submit one proposal document.

Deliverable #2: Data Acquisition

You can earn a grade of “M” on the second deliverable by meeting the following criteria:

- You save your Jupyter notebook as `data_acquisition.ipynb`.
- You include your 2+ HTML/XML/JSON data files with your notebook in a single .zip file.
- You use web scraping to extract one table of data before converting it to a pandas DataFrame.
- You acquire data from the other dataset and convert it to a pandas DataFrame.
- Your notebook includes Markdown cells introducing the datasets.
    - Provide citation(s) for the data, and a brief description for each.
    - Explain your central question, and give a brief explanation of how you plan to use the datasets to answer it (e.g., a single paragraph).
    - Explain the process by which you have accessed the data and parsed/scraped it.
- Your notebook executes without any errors.

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

- Provide additional information about your data sources:
    - Give context for the problem – why is this data interesting?
    - Explain how you know that the data sources are eligible for scraping (this likely involves a discussion of the Terms of Service or a similar document for each data source).
    - 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 scraping or otherwise processing the data, explain them in Markdown cells.
- You store your HTML/XML data files in a folder named `data` that is a sibling of your Jupyter notebook.

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. ***Submission of reports with too-similar Markdown cells will be considered an academic integrity violation.***

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 scraping and parsing the data. Failure to submit a Jupyter notebook, or a notebook that is lacking in almost all criteria for an "M", will result in a grade of "N".

Deliverable #3: Data Storage and Analysis

You can earn a grade of "M" on the third and final deliverable (two Jupyter notebooks and a SQL database) by meeting the following criteria:

- Your Jupyter notebooks are named `data_acquisition.ipynb` and `data_analysis.ipynb`.
- You submit your notebooks, your data, and your SQL database in a single .zip file.
- You store your data in a SQL database.
- You query the data from your SQL database into `pandas DataFrames`.
- You answer your central question using at least one graph.
- Your notebook executes without any errors.
- You include Markdown cells to explain:
  - your database design, including diagrams showing the fields in each table, the primary key(s) for each table, and any relationships (e.g., foreign key constraints) between tables.
  - the process by which you answer your question.

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

- All data storage (e.g., parsing HTML/XML/JSON and writing to SQL) occurs in your `data_acquisition.ipynb` notebook, and all data analysis occurs in `data_analysis.ipynb`.
- Explain in `data_acquisition.ipynb` how you know that the data sources are eligible for scraping (this likely involves a discussion of the Terms of Service or a similar document for each data source).
- Your data goes from HTML/XML/JSON -> `pandas DataFrames` -> SQL -> `pandas DataFrames` before you answer your question, rather than using the original `DataFrames` directly.
- You explain in Markdown cells:
  - how your database design adheres to principles of good DB design.
  - any challenges faced when storing your data in your SQL database.
- The problem solving needed to scrape the data and/or process the data was complex.
- Your explanations in Markdown cells are clear and well written.

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. ***Submission of reports with too-similar Markdown cells will be considered an academic integrity violation.***

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 using SQL to store and access the data and answer your question. Failure to submit this deliverable, or notebooks that are lacking in 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.