**Links to Data from Arizona Dept of Education (ADE)** <https://www.azed.gov/accountability-research/data>is the actual website

* [Cohort 2023 Four Year Graduation Rate Data](https://www.azed.gov/sites/default/files/2024/01/4Year_Grad_Rate_Cohort2023_publish%5B1%5D.xlsx)
* [Cohort 2022 Four Year Graduation Rate Data](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.azed.gov%2Fsites%2Fdefault%2Ffiles%2F2023%2F05%2F4Ygraduation_rate_cohort_2022_redacted.xlsx&wdOrigin=BROWSELINK)
* [Cohort 2021 Four Year Graduation Rate Data](https://www.azed.gov/sites/default/files/2022/03/Cohort%202021%204%20Year%20Graduation%20Rate.xlsx)
* [Cohort 2020 Four Year Graduation Rate Data](https://www.azed.gov/sites/default/files/2021/06/Cohort%202020%204%20Year%20Graduation%20Rate%20Final.xlsx)
* [Cohort 2019 Four Year Graduation Rate Data](https://www.azed.gov/sites/default/files/2020/09/20194YearGradRateFinal.xls)

**Data Engineering Track Requirements (75 points)**

1. **xlsx -> csv ---Javier**
2. **csv into python, clean the data---Javier**
3. **break up the data into 3rd normal form ----Priscilla**
4. **Create an ERD----Priscilla**
5. **Use SQLAlchemy to add your tables into a database (SQLite)---Jonathan**
6. **Create a Flask API that queries your database and returns your data---Drew**

**Database Design (40 points)**

* The project uses ETL workflows to ingest data into the database. (10 points) Javier to save excel file into CSV, and narrow the data to <1000 records, use SQLAlchemy to upload data into a database, use SQLite as the database instead of Postgres because SQLite is a portable database file.
* The original dataset(s) are transformed prior to storing it in the database. (5 points) Javier
* A database is used to house the data (SQL, MongoDB, SQLite, etc.). (5 points)
* The database has at least two tables (SQL) or collections (NoSQL). (5 points)
* The project documents the choice of the database used and why. (5 points)
* The project includes documentation of the ETL workflow with diagrams or ERD. (10 points) Priscilla create ERDs before putting into sqlite. Break out the csv into smaller tables by school LEA id

**Data and Delivery (35 points)**

* The database contains at least 100 unique records. (5 points)
* The project uses one additional library not covered in class related to data engineering. (10 points)
* The project includes a method for reading data from the database and displaying it for future use, such as: (10 points)
  + Pandas DataFrame
  + Flask API with JSON output
  + SQLAlchemy Jonathan
* The GitHub repo has a README.md that includes the following: (10 points)
  + An overview of the project and its purpose
  + Instructions on how to use and interact with the project
  + At least one paragraph summarizing efforts for ethical considerations made in the project
  + References for the data source(s)
  + References for any code used that is not your own

**Both Track Requirements**

**Group Presentation (25 points)**

* All group members speak during the presentation. (5 points)
* The content is relevant to the project. (5 points)
* The presentation maintains audience interest. (5 points)
* Content, transitions, and conclusions flow smoothly within any time restrictions. (10 points)

**Project Guidelines**

The following project guidelines focus on teamwork, your project proposal, data sources, and data cleanup and analysis.

**Collaborating with Your Team**

Remember that these projects are a group effort. The experience of close collaboration will create better project outcomes and help you in your future careers. Specifically, you’ll learn collaborative workflows that will enable you to approach and solve complex problems. Working in groups allows you to work smart and dream big. Take advantage!

**Project Proposal**

Before you start writing any code, your group should outline the scope and purpose of your project. This will help provide direction and safeguard against **scope creep** (the tendency for projects to become more complex after work begins).

The proposal is essentially a brief summary of your interests and intent. Be sure to include the following details:

* The kind of data you’d like to work with and the field you’re interested in (finance, healthcare surveys, etc.)
* The questions you’ll ask of the data
* Possible source for the data

Use the following example for guidance:

"The aim of our project is to uncover patterns in credit card fraud. We’ll examine relationships between transaction types and location, purchase prices and times of day, purchase trends over the course of a year, and other related relationships derived from the data."

**Finding Data**

Once your group has written a proposal, it’s time to start searching for data. We recommend the following curated sources of high-quality data:

* [data.worldLinks to an external site.](https://www.data.world/)
* [KaggleLinks to an external site.](https://www.kaggle.com/)
* [Data.govLinks to an external site.](https://www.data.gov/)
* [Awesome Public DatasetsLinks to an external site.](https://github.com/awesomedata/awesome-public-datasets)
* [Public-APIsLinks to an external site.](https://github.com/n0shake/Public-APIs)
* [Awesome APILinks to an external site.](https://github.com/Kikobeats/awesome-api)
* [Medium API ListLinks to an external site.](https://benjamin-libor.medium.com/a-curated-collection-of-over-150-apis-to-build-great-products-fdcfa0f361bc)

**important**

Whenever you use a dataset or create a new dataset based on other sources (such as existing datasets or information scraped from websites), make sure to use the following guidelines:

1. Check for copyright protections, and make sure that the way you plan to use this dataset is within the bounds of fair use.
2. Document how you intend to use this dataset now and in the future. Find any licenses or terms of use associated with the dataset, and review them to confirm that your intended use is in compliance.
3. Investigate how the dataset was collected. Identify any indicators that the data was obtained from a source that the compilers were not authorized to access.

You’ll likely have to adjust your project plan as you explore the available data. That’s okay! This is all part of the process. Just make sure that everyone in the group is aligned on the project’s goals as you make changes.

Make sure that your datasets are not too large for your personal computer. Big datasets are difficult to manage locally, so consider using data subsets or different datasets altogether.

**Data Cleanup and Analysis**

Now that you’ve picked your data, it’s time to tackle development and analysis. This is where the fun starts!

The analysis process can be broken into two broad phases: (1) exploration and cleanup, and (2) analysis.

As you’ve learned, you’ll need to explore, clean, and reformat your data before you can begin answering your research questions. We recommend keeping track of these exploration and cleanup steps in a dedicated Jupyter notebook to keep you organized and make it easier to present your work later.

After you’ve cleaned your data and are ready to start crunching numbers, you should track your work in a Jupyter notebook dedicated specifically to analysis. We recommend focusing your analysis on multiple techniques, such as aggregation, correlation, comparison, summary statistics, sentiment analysis, and time-series analysis. Don’t forget to include plots during both the exploration and analysis phases. Creating plots along the way can reveal insights and interesting trends in the data that you might not notice if you wait until you’re preparing for your presentation. Presentation requirements will be further explained in the next module.