Project #2 Requirements Sneak Peek

Note: Project #2 will start on Week13, by Feb 22nd, and you don't have to prepare anything before this date.

Project Week - Week 13 – starts Feb 22nd

Day 1	Form groups
	Identify datasets
	Perform ETL on the data
Day 2	Develop database
Day 3	Complete final report

Project Proposals

Team effort

- Due to the brief timeline, teamwork will be crucial to your success!
- Work closely with your team through all phases of the project to ensure that there are no surprises at the end of the week.
- Working in a group enables you to address more difficult problems than you'd be able to manage on your own.
- Take advantage by working smart and dreaming big!

Project proposal

- Before you write any code, remember that you only have one week to complete this project.
- Think of this project like a typical work assignment. Imagine that a bunch of data came in, and you and your team have been tasked with migrating it to a production database.
- Try to take advantage of instructor and TA support during office hours and in-class project time.

Project 2: ETL

Data Cleanup and Analysis Requirements

Teams will be responsible for:

- Citing the data sources
- Extracting the data from those sources
- Transforming the data (cleaning, joining, filtering, aggregating, etc.)
- Loading the data into a database (relational or non-relational)

Report Requirements

You will also prepare a report to address the following points:

Extract

Your original data sources and how the data were formatted (CSV, JSON, pgAdmin 4, etc.)

Transform

What data cleaning or transformation was required

Load

The final database, tables/collections, and why this was chosen.

Project Rubric

Rubric Summary

Grading Categories

Project proposal	(20 points)
Technical report	(20 points)
GitHub repository	(20 points)

Data Suggestions

Data Suggestions

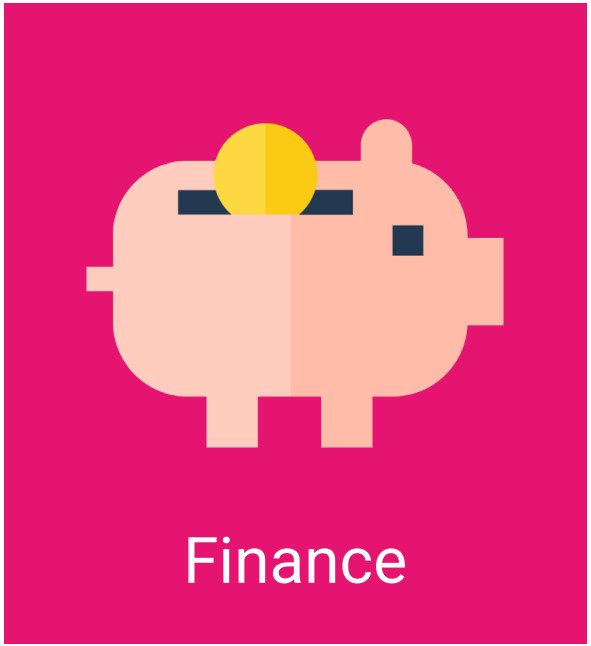
Feel free to ask us for input, but our general advice is to use data sources that:

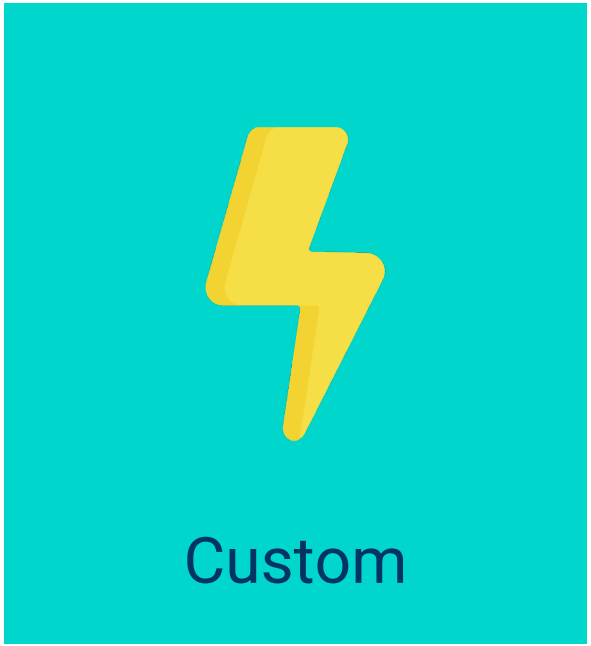
- Are sufficiently large
- Have a consistent format
- Ideally, contain more data than we need
- Are well documented

Choosing a Project Track

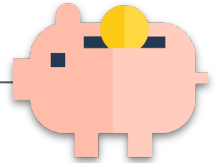
Choosing a Project Track

For this project, you can focus your efforts on a specific industry, including the following specializations:





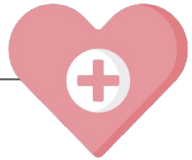
ETL and Finance



When to use ETL in finance

- Current treasury benchmarks are at an all-time low, and a financial analyst has decided to study the last 30 years worth of rates.
- After pulling historical data, the analyst cleans and explores the data to perform their analysis, with the intent of predicting future benchmark trends.
- Once the historical data have been collected, processed, and loaded into a database, the financial analyst turns their attention to present-day data. Using an API, they pull the most up-to-date information so it can be added to their established database.
- They've already extracted the new data, but before loading them into the existing database, they need to ensure that they have the correct format. Once the data are transformed, they can load them to the database and continue with the analysis.

ETL and Healthcare



When to use ETL in healthcare

An analyst working at a major hospital is tasked with reviewing policies regarding the upcoming flu season. The analyst is keeping the following questions in mind:

- How many patients does the hospital expect this year?
- How severe will flu season be this year?
- Will there be regional differences? Similarities?

The analyst wants to collect and analyze data from different sources so they can make predictions about the upcoming flu season.

Before combining the hospital's own data with regional data acquired externally, the analyst will need to extract the new data, transform them to match the existing data, and then load them into the database.

ETL in the Wild

Customize it!

Several other industries use the ETL process, as well.

In marketing, analysts may acquire data from competitors to see how their products measure up. Multiple data sources would need to be extracted, transformed, and loaded into a common database prior to analysis.

An analyst working for a large retail chain is in charge of moving a legacy database into a cloud-based data warehouse.

An entrepreneur has a big business idea but wants to get a feel for their product idea. hey use web scraping and APIs to pull data from a variety of social media platforms with the intent of analyzing consumer reactions.