

Term Project

Term Project

Throughout the term, you will be working on a project that has 5 milestones. The goal of the project is to build on the materials covered during each 2-week period.

Click on the milestones below to learn more about what is due when.
















Milestone 1



Identify Datasets

The first milestone of this project will be to select the data you want to work with. You will need to select 3 different data sources that have different file types of information – and the data will need to have a relationship between them. If one doesn't exist, you will have to create one. It is likely you will need to manipulate the data to create a relationship. Finding the data, you want to work with for this project, will likely be the hardest part of the project. You must have one of each of the following types of datasets – and you need a minimum of 1000 rows across all datasets. You need a total of 30 columns across the 3 datasets you select.

- CSV/Excel/PDF or another flat file source.
- Website you want to pull data from--you will want to identify a website that has data stored in a table, similar to the screenshot below.

Rank ↕	Country (or dependent territory) ↕	Population ↕	% of world population ↕	Date ↕	Source
1	 China ^[b]	1,403,496,680	18.0%	12 Jul 2020	National population clock ^[3]
2	 India ^[c]	1,364,603,167	17.5%	12 Jul 2020	National population clock ^[4]
3	 United States ^[d]	329,940,508	4.23%	12 Jul 2020	National population clock ^[5]
4	 Indonesia	269,603,400	3.46%	1 Jul 2020	National annual projection ^[6]
5	 Pakistan ^[e]	220,892,331	2.83%	1 Jul 2020	UN Projection ^[2]
6	 Brazil	211,782,426	2.72%	12 Jul 2020	National population clock ^[7]
7	 Nigeria	206,139,587	2.64%	1 Jul 2020	UN Projection ^[2]
8	 Bangladesh	168,940,146	2.17%	12 Jul 2020	National population clock ^[8]
9	 Russia ^[f]	146,748,590	1.88%	1 Jan 2020	National estimate ^[9]
10	 Mexico	127,792,286	1.64%	1 Jul 2020	National annual projection ^[10]
11	 Japan	125,930,000	1.61%	1 Jun 2020	Monthly provisional estimate ^[11]
12	 Philippines	108,881,966	1.40%	12 Jul 2020	National population clock ^[12]
13	 Egypt	100,608,449	1.29%	12 Jul 2020	National population clock ^[13]
14	 Ethiopia	98,665,000	1.27%	1 Jul 2019	National annual projection ^[14]
15	 Vietnam	96,208,984	1.23%	1 Apr 2019	2019 census result ^[15]

- API you will pull data from.

Some places you can find datasets are listed below:

- [Tableau Community](#)
- [Kaggle Datasets](#)
- [Data.Gov](#)
- [Science.Gov](#)
- [Data.Gov.UK](#)
- [NORC](#)
- [European Social Survey](#)
- [API List](#)
- [PrommableWeb](#)
- [Public APIs](#)
- [OpenWeatherMap](#)

Wikipedia is a good source to find data that is in a table - and the structure of the HTML is usually very similar.

There are no restrictions on what dataset you use, other than you cannot use the specific datasets used in the book(s).

For the first milestone, you need to submit the following:

- Project Subject Area: Describe your project in 1-2 sentences
- Data Sources:
 - Flat File:
 - Description
 - Link or Flat File uploaded
 - API:
 - Description
 - Link

- Website:
 - Description
 - Link
- Relationships
 - Describe how the data from each source is connected (see example below).
 - If there isn't an obvious relationship, explain how you will make one
- 250 Words describing how you plan to tackle the project, what the data means, ethical implications of your project scenario/topic, and what challenges you might face.

Submit via a PDF to the assignment link.

Example of Relationships:

In case you are confused what is meant by a relationship between the data sources here is an example (this is a very simple example and I would expect your datasets to have more variables)

CSV File: Contains a list of stores by store ID and other metadata about the stores

Website: Contains a list of store locations, by location ID and store ID and the various departments each store has by department ID.

API: Contains the transactions at each store – contains a transaction ID and store ID.

All 3 of these data sources are related by Store ID. The CSV file has a 1 to many relationship with the Website by StoreID and has a one to many relationship with the API data by StoreID as well.

Milestone 2



Cleaning/Formatting Flat File Source

Perform at least 5 data transformation and/or cleansing steps to your flat file data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone.

Examples:

- Replace Headers
- Format data into a more readable format
- Identify outliers and bad data
- Find duplicates
- Fix casing or inconsistent values
- Conduct Fuzzy Matching

Make sure you clearly label each transformation step (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences.

You must submit the following:

- Jupyter Notebook File or PDF of your code with Milestone # listed.
- Each transformation should be labeled with description or what it is doing.
- Human readable dataset after all transformations should be printed at the end of your notebook.
- 1 paragraph of the ethical implications of data wrangling specific to your datasource and the steps you completed.

You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

Milestone 3



Cleaning/Formatting Website Data

Perform at least 5 data transformation and/or cleansing steps to your website data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone.

Examples:

- Replace Headers
- Format data into a more readable format
- Identify outliers and bad data
- Find duplicates
- Fix casing or inconsistent values
- Conduct Fuzzy Matching

Make sure you clearly label each transformation step (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences.

You must submit the following:

- Jupyter Notebook File or PDF of your code with Milestone # listed.
- Each transformation should be labeled with description or what it is doing.
- Human readable dataset after all transformations should be printed at the end of your notebook.
- 1 paragraph of the ethical implications of data wrangling specific to your datasource and the steps you completed.

You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

Milestone 4



Connecting to an API/Pulling in the Data and Cleaning/Formatting

Perform at least 5 data transformation and/or cleansing steps to your API data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone.

Examples:

- Replace Headers
- Format data into a more readable format
- Identify outliers and bad data
- Find duplicates
- Fix casing or inconsistent values
- Conduct Fuzzy Matching

Make sure you clearly label each transformation step (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences.

You must submit the following:

- Jupyter Notebook File or PDF of your code with Milestone # listed.
- Each transformation should be labeled with description or what it is doing.
- Human readable dataset after all transformations should be printed at the end of your notebook.
- 1 paragraph of the ethical implications of data wrangling specific to your datasource and the steps you completed.

You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.



Merging the Data and Storing in a Database/Visualizing Data

Now that you have cleaned and transformed your 3 datasets, you need to load them into a database. You can choose what kind of database (SQLite or MySQL, Postgre SQL are all free options). You will want to load each dataset into SQL Lite as an individual table and then you must join the datasets together in Python into 1 dataset.

Once all the data is merged together in your database, create 5 visualizations that demonstrate the data you have cleansed. You should have at least 2 visualizations that have data from more than one source (meaning, if you have 3 tables, you must have visualizations that span across 2 of the tables – you are also welcome to use your consolidated dataset that you created in the previous step, if you do that, you have met this requirement).

For the visualization portion of the project, you are welcome to use a python library like Matplotlib, Seaborn, or an R package ggPlot2, Plotly, or Tableau/PowerBI.

PowerBI is a free tool that could be used – Tableau only has a free web author. If you use Tableau/PowerBI you need to submit a PDF with your assignment vs the Tableau/PowerBI file.

Clearly label each visualization. Submit your code for merging and storing in the database, with your code for the visualizations along with a 250-500-word summary of what you learned and had to do to complete the project. In your write-up, make sure to address the ethical implications of cleansing data and your project topic. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

You must submit the following:

- Jupyter Notebook File or PDF of your code with Milestone # listed.
- Each transformation and visualization should be clearly labeled.
- Human readable dataset after all transformations should be printed at the end of your notebook.
- 250-500 word summary of what you learned and a summary of the ethical implications.

Remember – your GitHub repository can act as a portfolio for potential employers! I would highly suggest using this to submit your work, so you can fill it with good content that demonstrates the projects you are working on!

