Project 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, meaning one dataset can have 5 columns, another can have 10 and the final could have 15. A total of 1000 rows is needed across all datasets, not each. Also - you will likely not have that many rows of data at the end of the project after you transform & blend the data together.

- 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. You will not be able to export the data to CSV for this step.
- API you will pull data from. You will not be able to export data to CSV for this step.

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 regardless of topic.

There are no restrictions on what dataset you use.

For the first milestone, you need to submit the following as a Word document:

- Project Subject Area: Describe your project in 1-2 sentences
- Data Sources:
    - Flat File:
        - Description of data source
        - Link or Flat File uploaded
    - API:
        - Description of data source
        - Link
    - Website:
        - Description of data source
        - 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
- 250-500 Words covering the following:
    - Project approach/plan
    - What concerns/challenges you think you will face with the data/project topic
    - Ethical Implications of your project topic

Submit via a Word document 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 1 is due Sunday, by Midnight of Week 4. Refer to the rubric for more grading detail.

Project 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 number and 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 answering the following questions:
    - What changes were made to the data?
    - Are there any legal or regulatory guidelines for your data or project topic?
    - What risks could be created based on the transformations done?
    - Did you make any assumptions in cleaning/transforming the data?
    - How was your data sourced / verified for credibility?
    - Was your data acquired in an ethical way?
    - How would you mitigate any of the ethical implications you have identified?

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 2 is due Sunday, by Midnight of Week 6. Refer to the rubric for more grading detail.

Project 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. As a reminder - you cannot export your website data to CSV to work with it, you must do all the work directly against the HTML source.

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 number and 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 answering the following questions:
    - What changes were made to the data?
    - Are there any legal or regulatory guidelines for your data or project topic?
    - What risks could be created based on the transformations done?
    - Did you make any assumptions in cleaning/transforming the data?
    - How was your data sourced / verified for credibility?
    - Was your data acquired in an ethical way?
    - How would you mitigate any of the ethical implications you have identified?

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 is due Sunday, by Midnight of Week 8. Refer to the rubric for more grading detail.

Project 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. As a reminder - you cannot export your API data to CSV to work with it, you must do all the work directly against the API/JSON source.

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 number and 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 answering the following questions:
    - What changes were made to the data?
    - Are there any legal or regulatory guidelines for your data or project topic?
    - What risks could be created based on the transformations done?
    - Did you make any assumptions in cleaning/transforming the data?
    - How was your data sourced / verified for credibility?
    - Was your data acquired in an ethical way?
    - How would you mitigate any of the ethical implications you have identified?

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 is due Sunday, by Midnight of Week 10. Refer to the rubric for more grading detail.

Project Milestone 5

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 (SQLLite 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 step 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 answering the following questions:

- What changes were made to the data?
- Are there any legal or regulatory guidelines for your data or project topic?
- What risks could be created based on the transformations done?
- Did you make any assumptions in cleaning/transforming the data?
- How was your data sourced / verified for credibility?
- Was your data acquired in an ethical way?
- How would you mitigate any of the ethical implications you have identified?

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!

Milestone 5 is due **Saturday**, by Midnight of Week 12. Refer to the rubric for more grading detail.