

Project 2

[Start Assignment](#)

Due May 1 by 11:59pm **Points** 50 **Submitting** a file upload
File Types html and ipynb

Project 2

For this project, you will be conducting your own data science analysis of a dataset of your choosing. You are empowered to find a dataset that interests you. The final product of your analysis will be a Jupyter Notebook with some explanations and the results of your computation alongside the code.

This project assesses the following learning objectives:

- Explore a question that can be answered through a Data Science investigation
- Appropriately access, cite, and share data
- Generate and interpret a data visualization to effectively convey information to your audience
- Develop a compelling narrative using data

Finding a Dataset

You should find a file to conduct your analysis on. There are many public repositories of data, such as the **CORGIS collection** (<https://think.cs.vt.edu/corgis/csv/>), which has a large number of .csv files from a variety of sources. You could analyze historical data about diseases, battle logs from a video game, weather records in your home state, or whatever you please.

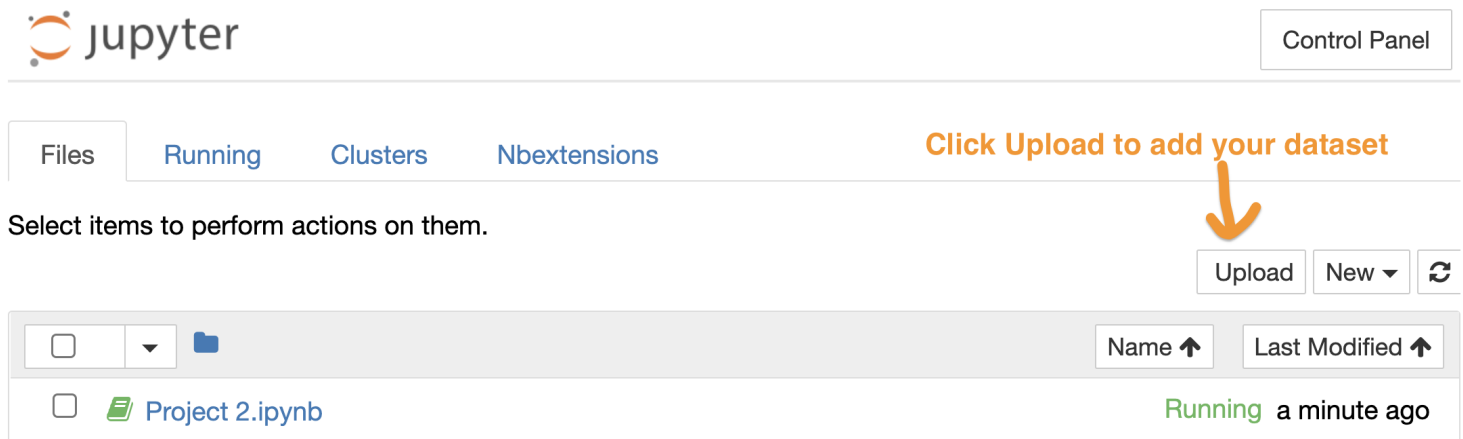
Although there is great flexibility in the shape and nature of your data, you must ensure that there is sufficient numeric data in the dataset to conduct your analysis.

Regardless of what dataset you choose, you need to clearly identify where the data come from, appropriately **site the source of the data** (<https://libguides.lib.msu.edu/citedata>), and make an objective argument for the importance of the data. It is not enough to just say that you find it personally interesting - you must provide a justification that a neutral third party will find believable. Any kind of analysis can be justified: consider arguments from different perspectives such as economic factors, expert testimonials, etc.

Once you have chosen your dataset, please complete this quiz (you are free to change your dataset anytime during the project, but please remember to update your answer to this quiz): **Project 2: Dataset**

Loading the Dataset

Workspace has been created for your project in [Vocareum](#). The assignment is called Project 2 and contains a starter Project 2.ipynb file with the usual imports already added. You will need to load the dataset to have access to it within your Jupyter Notebook. To do so, use the upload button before opening your Jupyter Notebook.



Once the data is loaded, you can perform any preprocessing or cleaning that is necessary to use the data in the subsequent steps.

Histogram Analysis

First, you are required to generate and explain a Histogram of some numeric data (with at least 30 data points). Although this could mean processing a list of numbers found directly in the data, you are also free to do analysis of the data that leads to numeric data. For example, you could analyze text data to compute some numbers and find their distribution. The only requirement is that the histogram you produce must have at least 30 data points represented.

After you have generated and shown the Histogram, you must interpret its meaning. What does the distribution say about the nature of the data?

Secondary Analysis

Second, you must then do a further analysis that interests you, such as any of the following:

- A line plot showing trends
- A scatter plot comparing related values
- A **bar graph** (<https://pythonspot.com/matplotlib-bar-chart/>) showing values across categories
- Descriptive statistics such as mean, median, sums, etc.
- **Inferential statistics** (<https://docs.scipy.org/doc/scipy/reference/stats.html>) such as regression
- A more advanced **regression** (<https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>) type of analyses

Some of the above could be done without any special libraries, but some may require the use of the Scipy, Scikit-learn, Matplotlib modules (or other interesting data science tools). It is up to you to decide how much extra analysis you want to do, but you should make sure that your analysis is a reasonable choice for the data.

Be sure to clearly explain what kind of secondary analysis you did and interpret the results.

Stakeholders

You will need to identify two stakeholders who would be interested in your analyses. These stakeholders should be distinctive from each other. For example, for weather data, a *non-distinctive* pair of stakeholders would be "Weathermen" and "Forecasters". A much better *distinctive* pair would be "Farmers planning their watering schedule" and "Scientists studying climate change". Try to think of general classes of people in different parts of society.

For each stakeholder, you should clearly explain what the stakeholders should learn from your analysis. This could be in the form of recommendations, or a description of how the results are relevant.

Report

You should combine your code and the results of running that code into a Jupyter Notebook. In general, here is a recommended outline:

1. Title
2. Your name
3. Explain your dataset and its origin
4. Load your dataset using the JSON or Requests module, then clean and preprocess your dataset in preparation for visualization
5. Create a histogram of your data and interpret its results
6. Conduct a secondary analysis of your data and interpret the results
7. Identify two distinctive stakeholders and contextualize your results for the stakeholders.
8. The honor code

Grading

You will be graded on the following components:

- 5 points for clearly identifying and correctly citing the source of the dataset
- 5 points for objectively explaining the importance of the data
- 10 points for loading and potentially cleaning the dataset with good code organization
- 10 points for a properly labelled histogram with at least 30 data points and a clear interpretation of your histogram

- 10 points for an additional statistical, visual, or other Data Science analysis and a clear explanation of the results of your secondary analysis
- 10 points for clearly identifying at least 2 stakeholders and what they should learn from the results

Refer to the rubric at the bottom of the page for more in-depth explanations of the exact criteria used to grade you. Notice that each element can receive one of the following marks:

Submission

You **must** submit two files:

1. The Jupyter Notebook (.ipynb) file source code of the notebook.
2. The HTML (.html) file of your Jupyter Notebook

You **must** include your name and the following honor code in your report: *"I have neither given nor received unauthorized assistance on this assignment."*

You **must** fill out this survey to clearly indicate what dataset you chose: [Project 2: Dataset](#)

Failure to follow the above submission requirements will result in a rejection of your submission and an automatic zero.

Rubric

Criteria	Ratings		
Appropriately access, cite, and share data	4 pts	3 pts	0 pts
To achieve full marks, the report must:	5 pts	Adequate	Inadequate
1. Provide enough information for a reader to find the original dataset	Full Marks	One of the criteria is not met	Two of the criteria is not met.
2. Load data file and preprocess/clean data		None of the criteria was met	

Criteria**Ratings****Explained Source of Dataset**

To achieve full marks, the report must:

1. Clearly explain the source of the data such that the reader can understand the nature of the data
2. Objectively explain the importance of the dataset with a coherent argument.

	4 pts	3 pts	0 pts
5 pts	Adequate	Inadequate	Missing
Full Marks	One of the criteria is not met	Two of the criteria is not met.	None of the criteria was met

Loaded and Processed the Dataset using Good Style and Organization

To achieve full marks, the report must:

1. Show code that loads the dataset in a reproducible way
2. Performs any preprocessing necessary to analyze the data as a histogram of the secondary analysis
3. Uses good code organization that makes it easy to read and follow the code (e.g., good variable names, documentation for complex code, no excessively nested code)

	8 pts	6 pts	0 pts
10 pts	Adequate	Inadequate	Missing
Full Marks	One of the criteria is not met.	Two of the criteria are not met.	None of the criteria was met

Labelled Histogram with 30+ Data Points with a Clear Interpretation

To achieve full marks, the report must:

1. Show the code to create and label the histogram
2. Show the histogram with proper labels
3. Have at least 30 data points
4. Clearly interpret the meaning of the histogram

	8 pts	6 pts	0 pts
10 pts	Adequate	Inadequate	Missing
Full Marks	One of the criteria is not met.	Two of the criteria are not met.	None of the criteria was met

Criteria**Ratings****Secondary Analysis with Clear Explanation**

To achieve full marks, the report must:

1. Show the code of a secondary analysis that answers a new question.
2. Concisely explains the kind of analysis that was performed.
3. Clearly interprets the meaning of the result of the analysis

	8 pts	6 pts	0 pts
10 pts	Adequate	Inadequate	Missing
Full Marks	One of the criteria was not met.	Two of the criteria was not met.	None of the criteria was met

Identified Stakeholders and Contextualized the Results for Them

To achieve full marks, the report must:

1. Clearly identify a stakeholder who would be interested in the results of the analyses.
2. Clearly identify a second, distinctive stakeholder who would be interested in the results from a different perspective
3. Contextualize the results for each stakeholder to clearly explain what that stakeholder should learn from the analyses.

	8 pts	6 pts	0 pts
10 pts	Adequate	Inadequate	Missing
Full Marks	One of the criteria was not met	Two of the criteria was not met	None of the criteria was met