

Data 200, Fall 2025 Project 1

Due: 11/3, 11:59 p.m.

With a tabular dataset of your choosing (locate one online):

1. Import the data into a Jupyter notebook and print the first 5 rows.
2. Retrieve general information about the data using info.
3. Create a set of descriptive statistics for the data.
4. Determine whether the tabular data is TidyData.
5. If it is TidyData, explain how it meets each of the 3 traits. If it is not, apply transformations to make it TidyData.
6. Within your dataset, group the data based upon a variable and calculate 3 different aggregations. If your data are not suitable to grouping, just calculate the 3 aggregations for the whole dataset. What do these aggregations suggest about the data?
7. Create two different types of visualizations with your dataset. These must be different kinds of visualizations (ex. scatterplot and histogram, not two histograms). Label them appropriately. Outline what they mean.
8. Examining your data and determine if a relationship exists between two of the variables. If your data are not suited to that task, determine if a general trend or pattern is present. Write up your findings (250 words).
9. Evaluate based upon the current methods how confident you can be with this hypothesis and map out some next steps of how you might go about more credibly (150 words).
10. Write a brief statement on what additional data you may need in your analysis. Where could you possibly locate it? (100 words)

When all parts are completed submit your Jupyter notebook, 500 words (these can be in the notebook or a separate file), and dataset in Moodle.