Project-1 Description

The Project-1 for this class is the analysis of a dataset of your own choosing (or your own dataset). You can choose the data based on your interests. The goal of this project is for you to demonstrate proficiency in the techniques we have covered in this class (and beyond, if you like) and apply them to a dataset in a meaningful way.

Choosing Dataset

You can pick a dataset from a repository like UCI, Kaggle, etc. It is important that you use a manageable dataset. This means that the source data should be accessible, and the final dataset should be large enough that multiple relationships can be explored. As such, your dataset must have **at least 50 observations** and **between 10 to 20 variables**. The dataset's variables should include **categorical** variables, **discrete** numerical variables, and **continuous** numerical variables.

Data fetching (downloading), data preprocessing shall be included into your Jupyter notebook `analysis.ipynb`. Optionally, you may provide your dataset together with your submission if it does not exceed the Moodle upload limit.

Building a Custom Dataset (optional, 5 bonus points)

Instead of using an available dataset you may wish to collect your own that reflects *your* interests. You need to write a Python script `get_data.py` that will

1. connect to a website (or multiple websites)
2. get the data by parsing the HTML source(s); in addition, you can parse sources in other formats (such as XML, JSON, etc.)
3. clean/process the data (or this can be deferred to the analysis part)
4. save the data in a file (or multiple files)

IMPORTANT: You cannot prepare a dataset, upload it to a server, and then fetch it in the `get_data.py`.

In order for you to have the greatest chance of success with this project it is important that you prepare a manageable dataset. This means that the source data should be accessible and the final dataset should be large enough that multiple relationships can be explored. As such, your dataset must have **at least 50 observations** and **between 10 to 20 variables**. The dataset's variables should include **categorical** variables, **discrete** numerical variables, and **continuous** numerical variables.

You may provide your final dataset together with your submission, but in principle I should be able to get it (possibly with variations) by running your `get_data.py`.

Analysis and Write up (required, 25 points)

All analyses and writeup must be done in a Jupyter notebook `analysis.ipynb`. If you prepare a dataset in a format that we haven't encountered in class (e.g. json), make sure that you are able to load it into Jupyter as this can be tricky depending on the source.

Your **introduction** should introduce your general topic and research questions[1] (I expect 3—5 interrelated questions) and your data (where it came from, how it was collected, what is the sampling design, what are the variables, etc.).

After providing the description of your dataset and research questions in the introduction use the remainder of your write up to showcase how you have arrived at answers to your questions using any techniques we have learned in this class (and some beyond, if you're feeling adventerous). The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather let me know that you are proficient at asking meaningful questions and answering them with results of data analysis, that you are proficient in using Python, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research questions. You do not have to apply every procedure we learned. Also pay attention to your writing. Neatness, coherency, and clarity will count.

Your write up must also include a 1—2 page **conclusion** and discussion. This will require a summary of what you have learned about your research question along with arguments supporting your conclusions. Also critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the analysis should be discussed here. A paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project should also be included.

The project is very open ended. You should create some kind of compelling visualization(s) of this data in Python. There is no limit on what tools or packages you may use. You do not need to visualize all of the data at once. A single high quality visualization will receive a much higher grade than a large number of poor quality visualizations.

You can add sections as you see fit. Make sure you have a section called Introduction at the beginning and a section called Conclusion at the end. The rest is up to you!

[1] Here you can learn what a **good** research question is.

Deliverables

Your submission should be a ZIP-file `Lastname.zip`, that contains:

- (optionally) `get_data.py`
- (optionally) dataset file(s) if they do not exceed Moodle limits
- `analysis.ipynb`