Mid-Semester Project

INFO 1998 | Fall 2020

I. Key Dates

Assigned: Oct 16, 2020
Due: Oct 30, 2020 (11:59pm)
Groups: 1-3 students

II. Problem

In the first section of the course, you have learned some fundamentals of Python and acquired some data manipulation and visualization skills. You also now have detailed tutorials (lecture notes) as a resource going forward. As we move towards the machine learning component of the course, it's important to ensure that you're confident with these fundamentals - and this project is the way to do that. **For this project, you will analyze a dataset of your choice and share your code and inferences through a well-documented Jupyter Notebook.**

1. For this project, we recommend that you find a dataset of your choice online - Kaggle and Data.gov are some incredible resources but feel free to explore.
2. After choosing a dataset, come up with a question that you want to answer. For example, questions relevant to the Titanic Dataset could be 'What was the distribution of deaths of men to women?' or 'What were the most common age groups on board?'
3. Clean and manipulate the data, and come up with visualizations (at least 2) based on your questions.

III. Rubric and Submission

- Preprocessing and Manipulation (30): Any necessary cleaning and manipulation of the dataset
- Visualizations (2 x 20 = 40): At least two visualizations of different types (i.e. you can't have merely two bar charts, for instance). Visualizations are clearly visible, clean, well-labeled, and serve a clear purpose for your question(s).
- Write-Up (20): The question, hypothesis (if any), and inferences are explained in detail. We recommend that you use 'Markdown' in contrast to the 'Code' on the Notebook.
- Presentation (10): Code and write up is written and structured sensibly

Note that you can work in groups of up to 3 students, but you may also choose to work individually for this project.

Please submit your Jupyter Notebook (that includes the link to the dataset source) and dataset in a zip file through CMS. If you are submitting as a group, you will have to first form a group on CMS and then submit just once.

IV. Other Notes

- Start early - finding a suitable dataset and cleaning it takes more time than you'd expect. Additionally, you may sometimes preprocess it only to find that the dataset does not reveal sufficient insights and will have to find another dataset.

- Seek help - The instructors would be happy to help you with dataset selection and/or any other components of the assignment. Feel free to come and work at office hours and ask questions as they arise.

- Be authentic - the datasets you find online would likely have multiple other projects stemming from them that you'd easily find online. We encourage you to skim through these for some inspiration but also warn against copying those or conducting identical analyses. We'll cross-check all the submissions and this has led to academic integrity violations in the past.