
Project Phase 2 - Data Cleaning and EDA

In the final project for this course, you will apply the techniques learned in this class to analyze datasets of interest to you. Your goal should be to create an original project that you would be proud to show off to a potential employer. The project will require you to collect and obtain several data-sets that can be correlated in some ways, clean, integrate and apply EDA on the data, perform analysis using ML techniques, and then present your findings by building a dashboard and visualizations for the intended audience.

The project is divided into several phases. In the SECOND phase (Data Cleaning and EDA), we ask you to examine the collected data and perform data cleaning and EDA. Each project is different, so it's hard to dictate what each team should do, but your notebook should include code that examines the dataset and performs data cleaning. This can include filling-in missing values, standardizing fields, removing irrelevant columns, binning data into categories because data appears sparse, etc. In doing data cleaning, you are performing EDA, but include additional EDA to visualize your data. You must include a histogram, at least two scatter plots or a pair-wise scatter plot, a parallel coordinates plot, and a box-plot. If a plot is not applicable in your case, then note that and include a different plot. Note, you should perform as much EDA as necessary and this is just minimum requirements.

This is a really good example notebook on visualization, so check it out <https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57>.

To get started, accept the assignment link here <https://classroom.github.com/g/hf69ksqp>

This submission MUST include a README that :

- identifies and describes the datasets that were used;
- Describe the EDA process and include comments on what you are trying to accomplish with each code snippet. It should also include text that summarizes your findings. (For example, you can comment that the data is skewed left or the data seems to be non-uniform);
- Contain information about how to run your code.

0.1 Submission

Submit your jupyter notebook (and other code) to github. Note, your notebook file should be named **teamname_proj_phase1**. You should include comments at the beginning of the notebook that includes your name and student id. Submissions will not be accepted via email; you must turn your assignment via Github.