EDA and visualization using Python

Stats 242 Project Proposal

Zhewen Hu and Clark Fitzgerald

In this project we plan to write a Python module centered primarily around data visualization. The focus will be on understanding and explaining statistical concepts as well as exploratory data analysis.

We have several high level objectives. First we would like to craft a quality piece of software that can be shown to potential employers. Second we'd like to build tools that simplify and accelerate common types of statistical data visualization. This would be personally useful to us and statistics learners.

Ideally this project would contribute functionality that is missing from the Python data analysis ecosystem. Python has several full featured visualization libraries including matplotlib, Seaborn, and bokeh. We have no intention of duplicating what's already been done. The goal is rather to use these as base platforms, and extend them to have more statistical capabilities. Hence this project will be more specialized than the general purpose libraries. If we are successful, and the code is general enough to have broad application then we will consider submitting patches to these projects.

Areas of concentration

There are several aspects of this project where we may delve more deeply.

- EDA : What are the fastest and most informative plots for exploratory data analysis? How can we quickly learn if the necessary assumptions for statistical modeling hold? For example, by binning and plotting boxplots along with a scatterplot one can better judge the linear model assumption of equal variance.
- Interactivity : Create html widgets / apps which require no technical skills to use. An example would be a tool that allows a user to vary the parameters of the Beta distribution and then plots the resulting shape.
- Maps : Statistical visualizations on maps offer rich opportunity as well as challenge. For example, one can easily draw the Leaflet map or Choropleth map.

Learning objectives

There are two aspects of data visualization: knowing what to plot, and knowing how to do it. In this project we hope to enhance our skills and experience in both aspects.