Stats 242 Revised Project Proposal

EDA and visualization of large data sets using Python

Zhewen Hu and Clark Fitzgerald

(3rd group member dropped the class)

This project will focus on visualizing data sets containing between 10 thousand and 100 million data points. It is too many points for most conventional EDA plots, and yet it still can fit easily in memory. There are two main challenges with data sets of this size- the time needed to plot, and overplotting.

The Bokeh project is actively developing this set of capabilities within Python. Downsampling in particular is under active development. By downsampling we can use fewer data points to achieve the same effect as a plot with more data. We will determine if this works on a local machine as well as on a server.

In this project we plan to push the limits of what Bokeh can do. We'll start with some statistical simulations of data up to the limits of memory. For a more advanced application we will examine climate data collected by NASA. This data consists of several GB of data collected daily and stored in HDF5 format. It's made publicly available through an FTP server.

The focus will be on rapid exploratory data analysis at scale. More precisely, when given a data set on a local laptop where n > 10,000, we'd like to create informative plots. If we're feeling really ambitious we'll try to create an application that uses the climate data to plot and animate maps.

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. The technologies we'll be working with are Python, Bokeh, XML, and HDF5.