**Capstone Project Proposal**

The purpose of this project is to analyze data on US housing costs. In particular, the aim of this study is to view trends in housing costs as a function of year, region, median income, and other relevant factors. Some interesting questions would be:

1. Determination of whether housing has gone up, down, or stabilized.
2. What specific factors have led changes in cost.
3. It would be of interest to measure any lingering impacts of the 2008 financial crisis.

The finance industry and public institutions could use this analysis as an indicator of the overall health of the economy. Government agencies may find this analysis useful to allocate better resources to households in distress. Lending agencies may be able to better gauge market conditions and the health of local markets based on this analysis.

The data is provided by the U.S. Department of Housing and Urban Development (HUD) from 1985 to 2013. The data is available in <https://www.huduser.gov/portal/datasets/hads/hads.html>, accessible for download to the general public. The data consists of separate csv text files for each year that the data was compiled.

The first approach is to understand the data available. Fortunately, the data fields are well documented and explained. The second step is to clean up the data (data wrangling) if necessary: Look for missing data, perform initial statistical analysis, look for outliers, and decide how to deal with missing data and outliers. The next step would be to plot correlations among the various data fields. This step should provide a first glimpse into any significant correlations among the data points. In here, the data in these correlations could be labelled according to the value of another field (e.g. color coding according to region or owner/renter status). In parallel, a hypothesis would be made on what are the important factors (data fields) driving housing costs. This hypothesis could then be tested using a linear (or nonlinear) regression model. Further analysis using machine learning techniques would then be applied to the data.

The project deliverables would include the Python code used in the statistical analysis and modeling, a final paper on trends discovered in the data (correlations, factors driving costs), and a presentation deck of slides showcasing the important results of this project.