Thomas Martinez

Jacob Fiola

Katherine Best

**Project Proposal**

**Data Set Link**

<https://web.archive.org/web/20170428050251/https://bouldercolorado.gov/open-data/city-of-boulder-public-trees/> (The dataset itself is the csv file on this page)

**Question of Interest**

The city of Boulder is home to hundreds of trees of varying species. As with any population, some of these trees are in better condition than others. We aim to investigate what factors most affect the health of Boulder's trees. This has the potential to yield some interesting results what trees do best or worst in Boulder. For example, are certain types of trees the healthiest? Or are trees that grow in certain locations the healthiest? Or does some combination of these factors yield the healthiest trees?

**Dataset**

We will use the 2015 version of the Boulder Public Trees dataset. This dataset contains information about the location and type of trees in the city of Boulder. We are using this version instead of the most recent version because it also contains information about the condition of the trees. The condition takes on one of six values (dead, very poor, poor, fair, good, and excellent) which we will represent as a scale from 1 to 6.

**Methods To Use**

There are a large number of attributes (39) some of which we believe are redundant. For example, we have multiple attributes describing the street that the tree can be found on. We will use a dimension reduction technique (PCA) to remove redundant features and simplify the computational analysis of the data. We will then use multivariate linear regression with the axes obtained from PCA to try to predict the condition of the trees. We will split the data into a training set and a testing set and use the training set to build the model. We will then test the model on the testing set to develop a sense of how accurate the model will be for unseen data.

We expect to find that the best indicator of a tree’s condition will be the species of the tree.