# What’s been done so far?

### Data Processing

* There is one column in the data file (Trees\_Public.csv) in which someone used a comma in the notes section. This was causing problems reading the file in so the comma was removed and replaced with a semicolon. The modified data file is Trees\_public\_Modified.csv.
* Some columns have been dropped for various reasons
  + Columns that contain strings that do not contain interesting / important information and that don’t have a limit on the number of possible entries have been dropped
  + Columns that contain all one value or almost all one value have been dropped if the entries were not boolean valued
  + Some other columns were dropped if there were not interesting and would have been difficult to deal with
* Note that columns with numerical values that are not linear have not been dealt with yet.
* For columns that have string values that do not correspond to numerical rankings we need to create indicator columns for each possible value.
  + I wrote a function to do this and it currently produces a warning which should probably be fixed.

#### Centering and Scaling

* Currently using a built in function from sklearn to center and standardize all columns.

### PCA

* Currently using a function from sklearn to do this so we can get an idea of how this will turn out.
* Note that the condition is still in with the data