Clustering

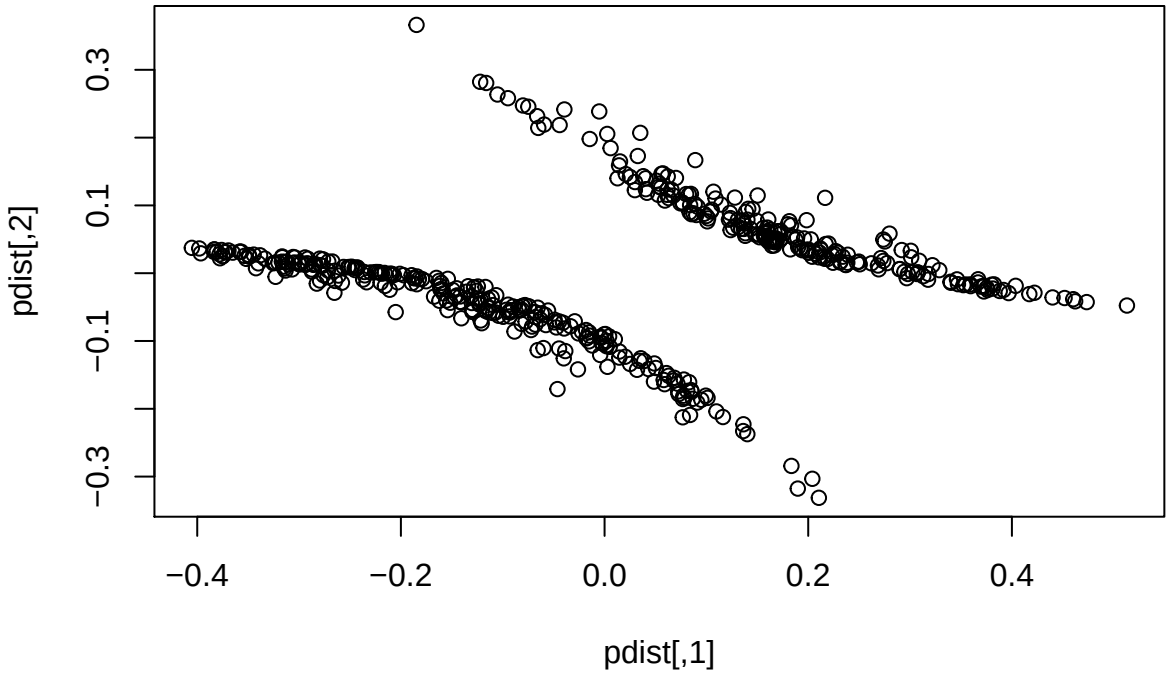
Jeff B

April 4, 2019

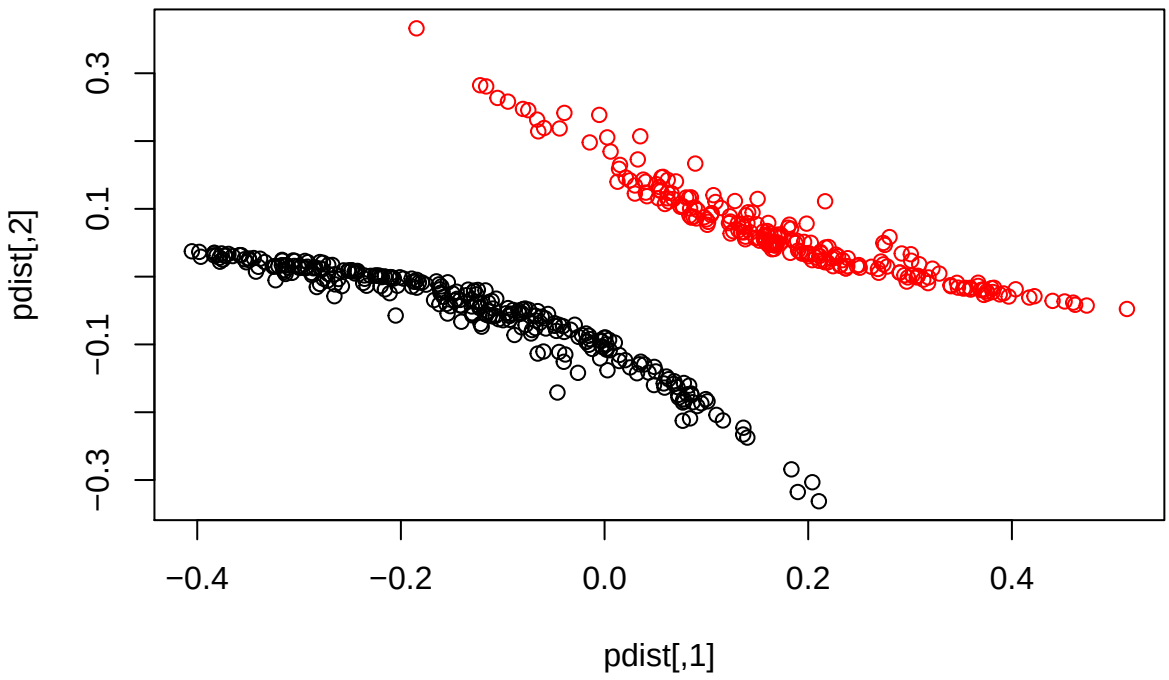
Clustering

We begin by computing the respective pairwise distances in our data, and plotting the output.

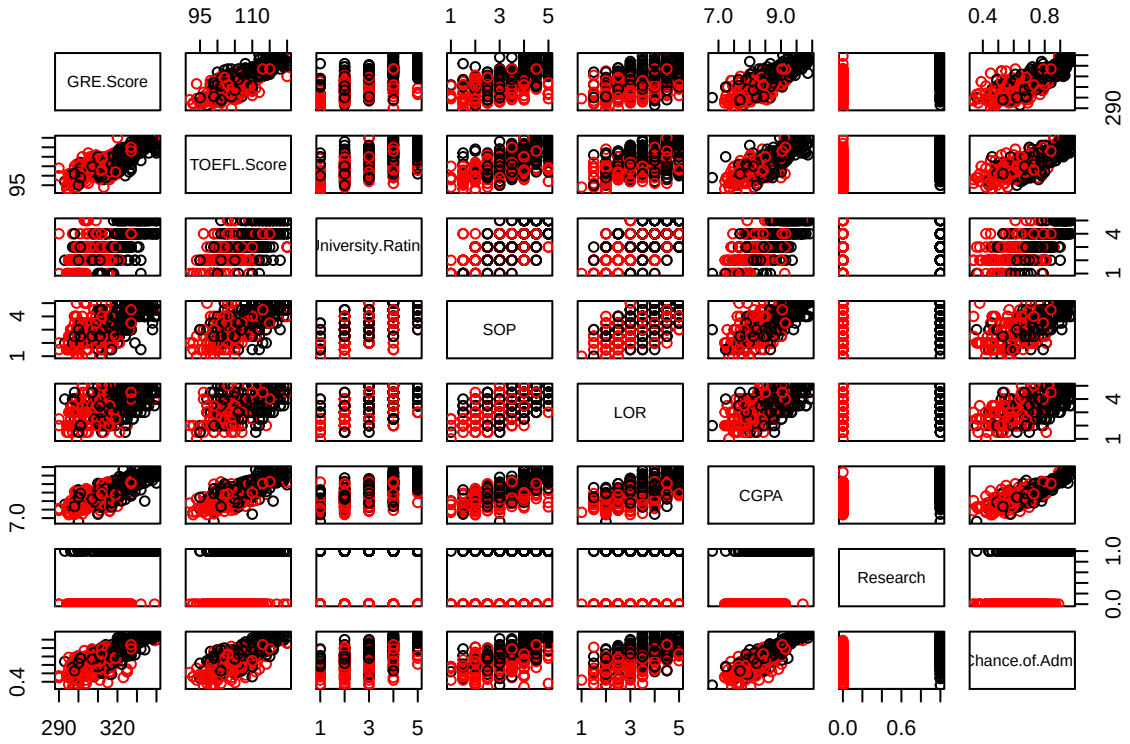
```
## Warning in daisy(admissionsData[, -1], metric = "gower"): binary
## variable(s) 7 treated as interval scaled
```



We quickly see that two clear groups appear. We can isolate these two groups using hierarchical clustering with single-linkage chaining.



We can then use scatterplots to show the entirety of the data, while still keeping the groups intact, to see if we can determine which predictors most affect these clusters.



We notice that, using the single linkage chaining from above, we can predict whether or not a student performs research almost perfectly.

So, by applying Gower's Distance on all predictors and using single-linkage chaining, we have two clear clusters directly coinciding with the presence of a research variable. This tells us that we should use Research as a response variable in models, in addition to Chance of Admit.

We can now perform analyses on the data to attempt to predict a candidate's Chance of Admission, as well as the presence of Research Experience.