ST 599 Project 1

Clustering

Dataset: All Sky Automated Survey

Overview

1. Data Cleaning
2. PCA
3. K-Means
4. Hierarchical

Data Cleaning

439 of the 1539 training stars with NA in their covariates (Color!)

- Used stars with complete sets of covariates
- Imputation for Hierarchical analysis

PCA

Goal: Dimensional Reduction

Challenge: Interpreting output is hard.
p-dimensional vectors (“loadout” coefficients)

Our solution: Top Variables

PCA

1. Separate dataset into 25 classes
2. Perform PCA on subsets
3. Which covariates most explain variation class-wide?
4. Top *p** most recurring covariates

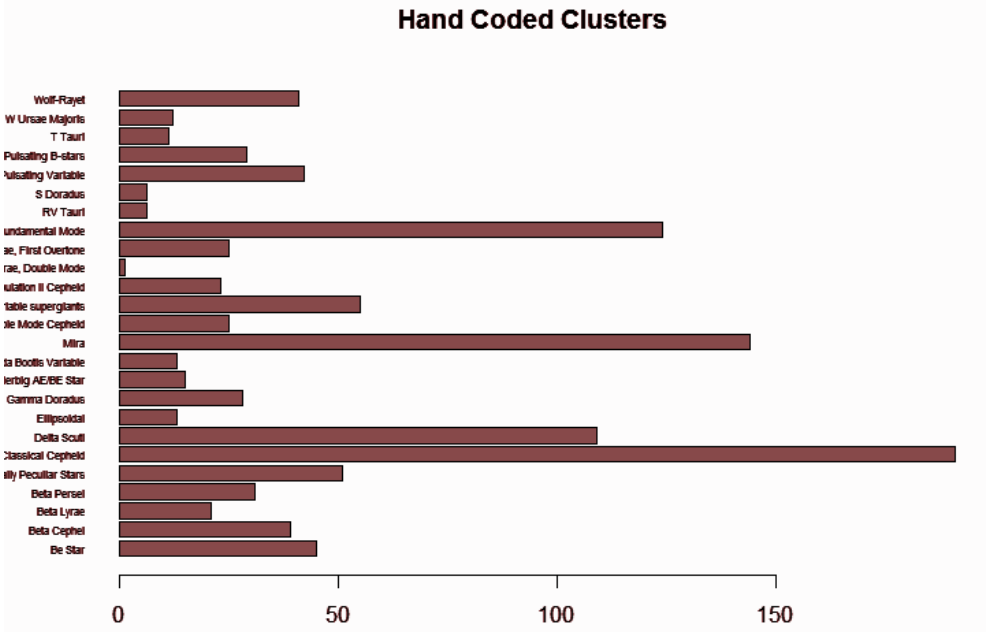
PCA

Top Variable Selection:

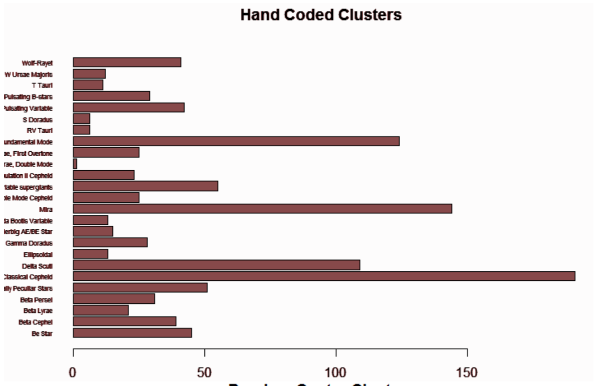
- Top five combination vectors from each PCA (explained roughly 50% of variation)
- Top ten contributors (by amplitude)
- Tally covariates contributions globally

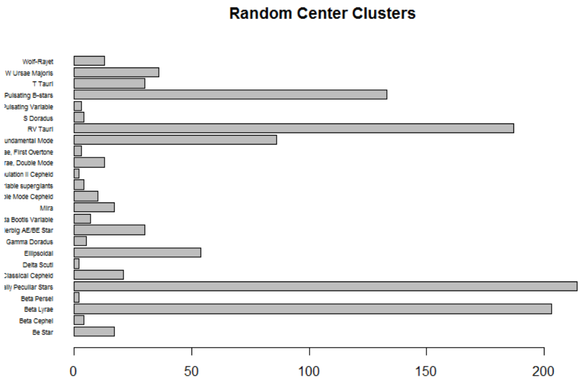
K-means

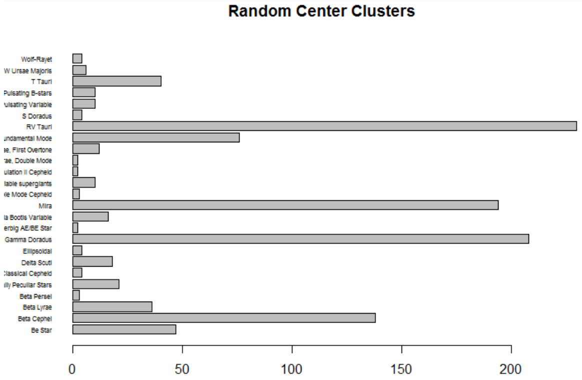
The “real” clusters



K-means Random Selection of centers (default)





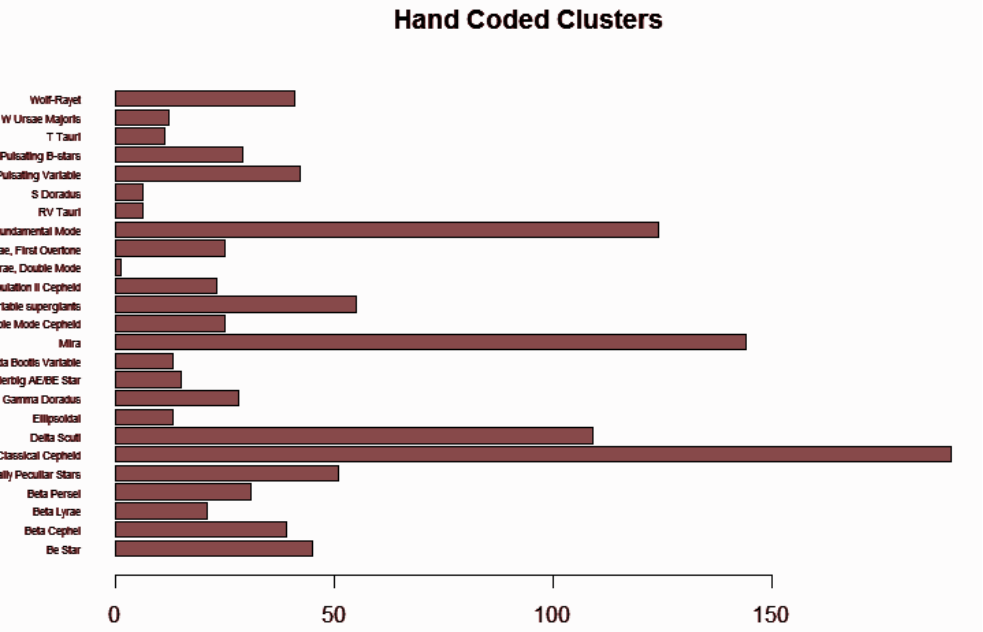


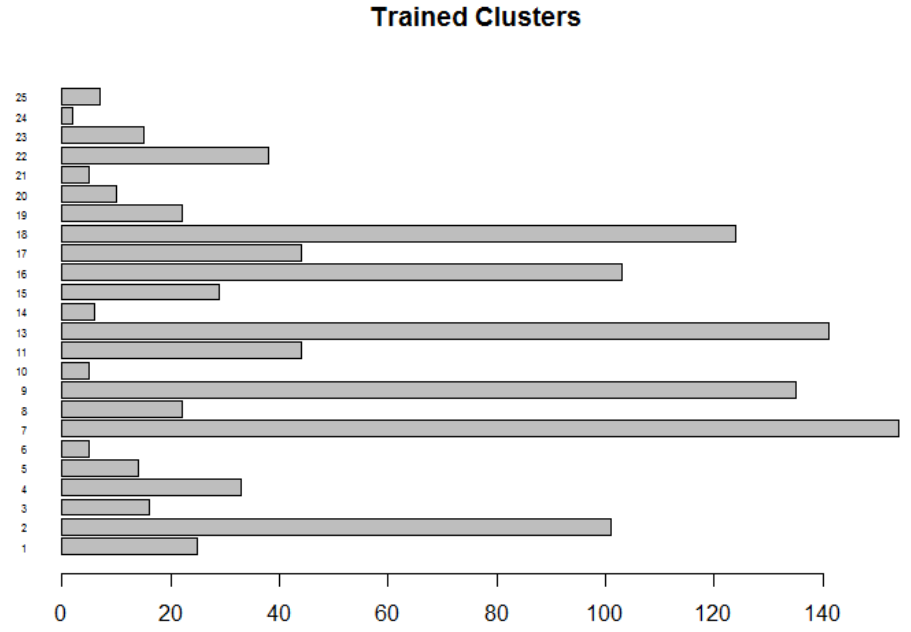
K-means

Custom Centers

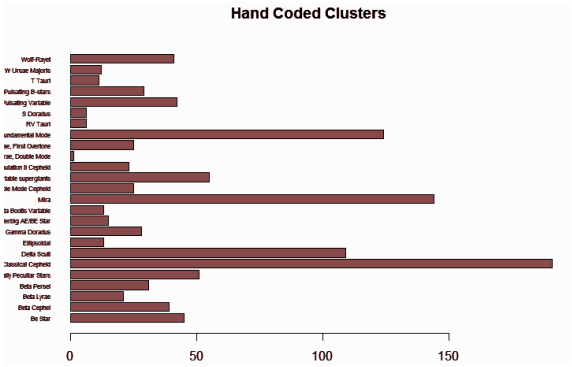
- with hand classified data, took the mean of each variable, for each star type
- Clusters of size 0
- Hartigan-Wong vs Lloyd algorithm

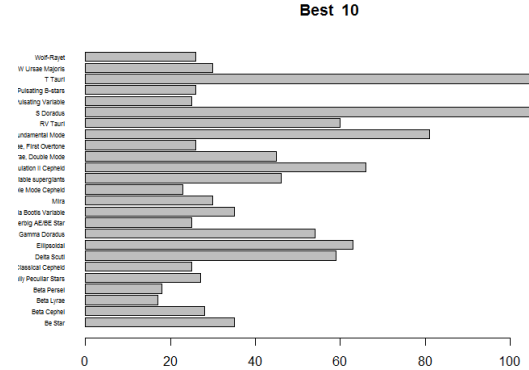
K-means custom centers

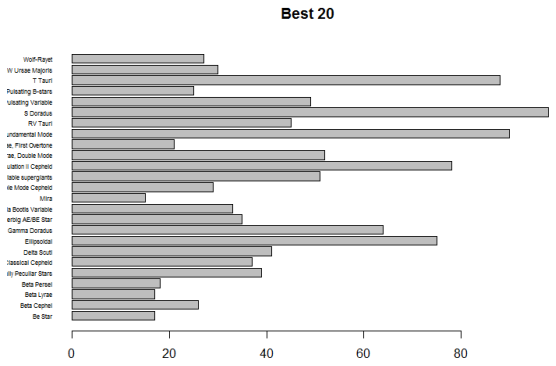


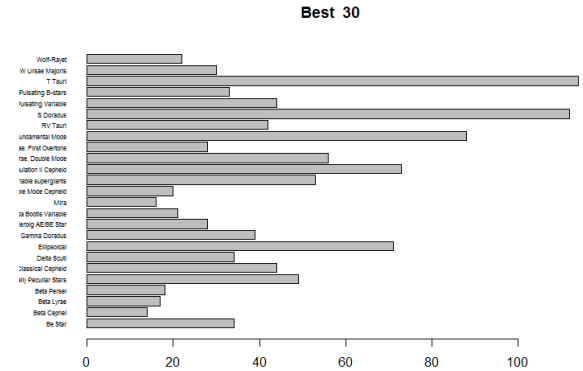


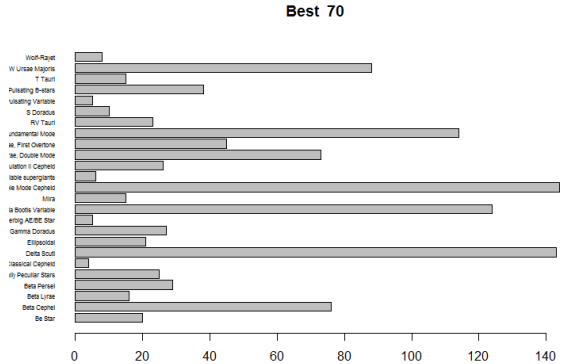
K-means with top *n* variables







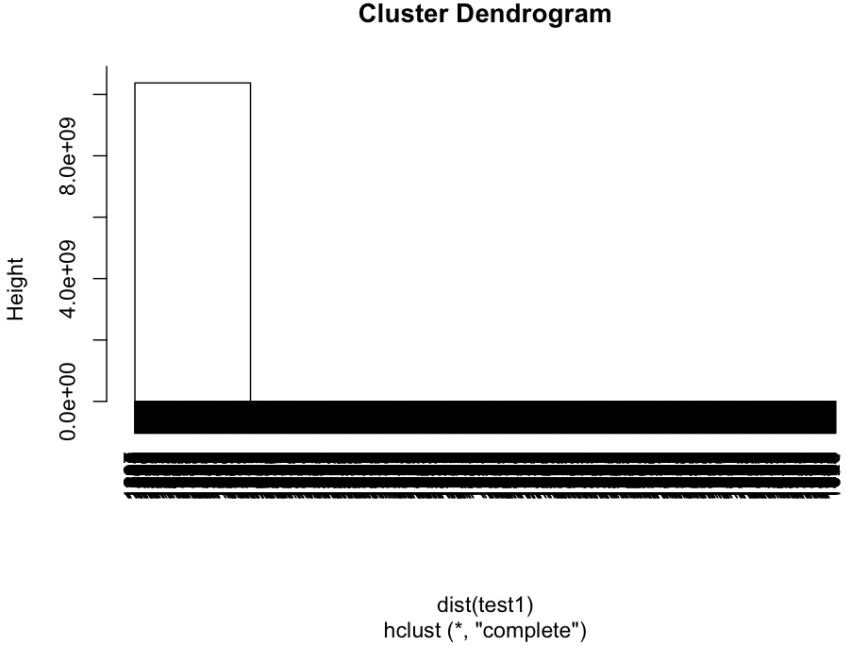


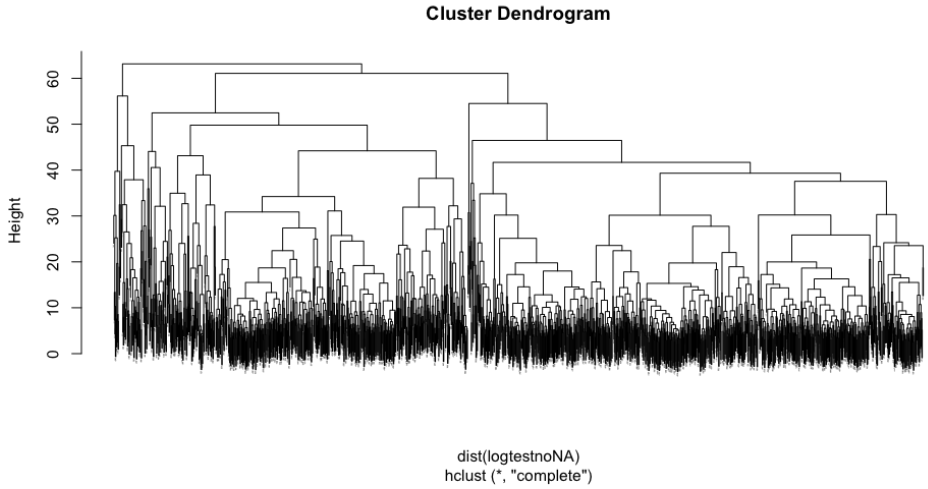


Hierarchical

- NA Imputation
- Log Transformation
- Single vs Complete

Hierarchical

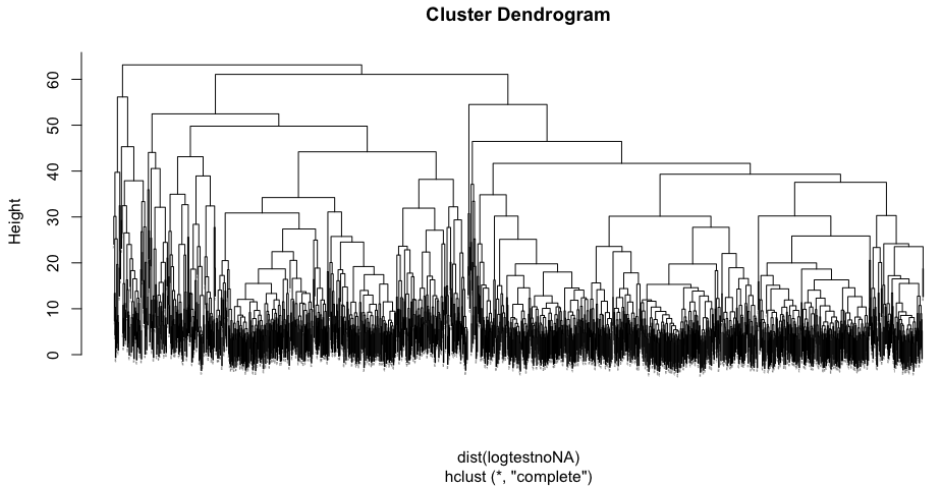




Future Work

- Better NA solution
- Where to Cut the Dendogram
- Star frequency explanation

  i.e. How to detect Mira?



Thanks!

Questions?