Introduction to Data Science

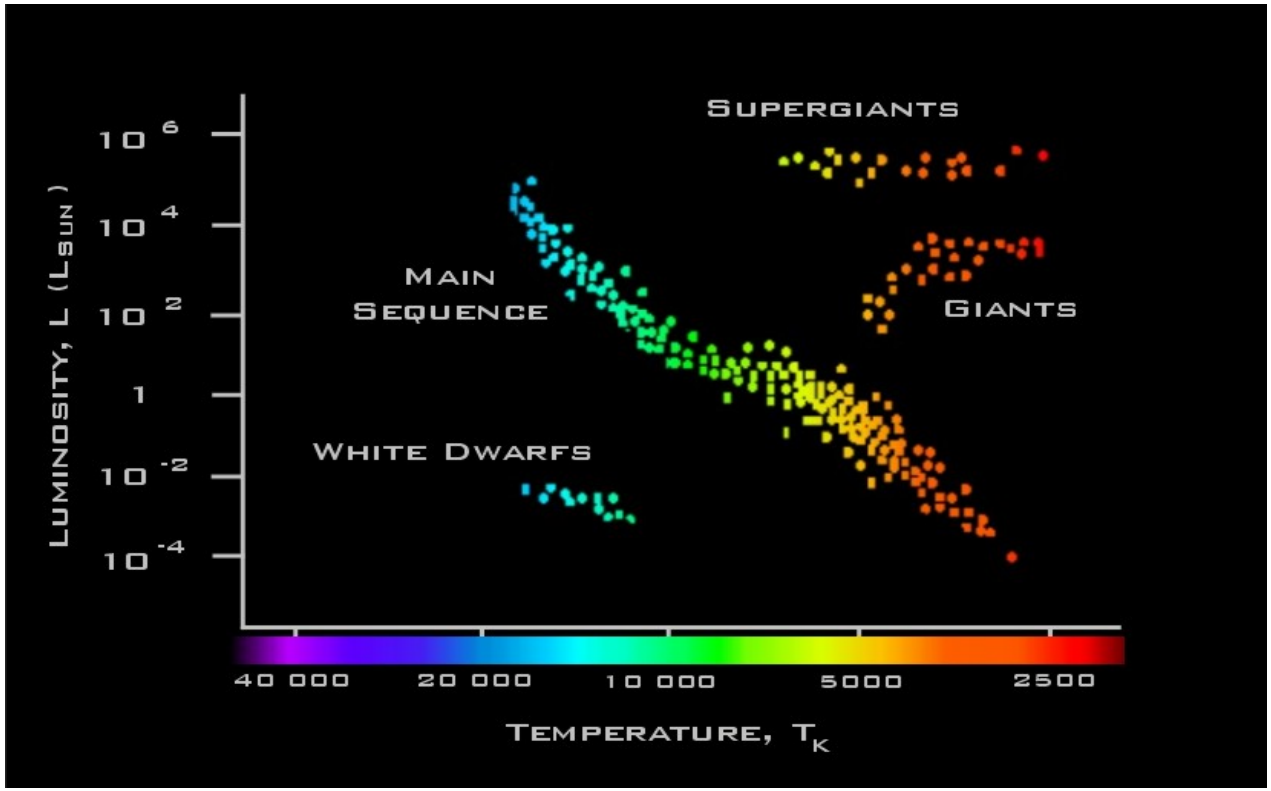
CLUSTERING

BRIAN D'ALESSANDRO

BASIC CLUSTERING GOAL

How do I find distinct groupings of similar objects?

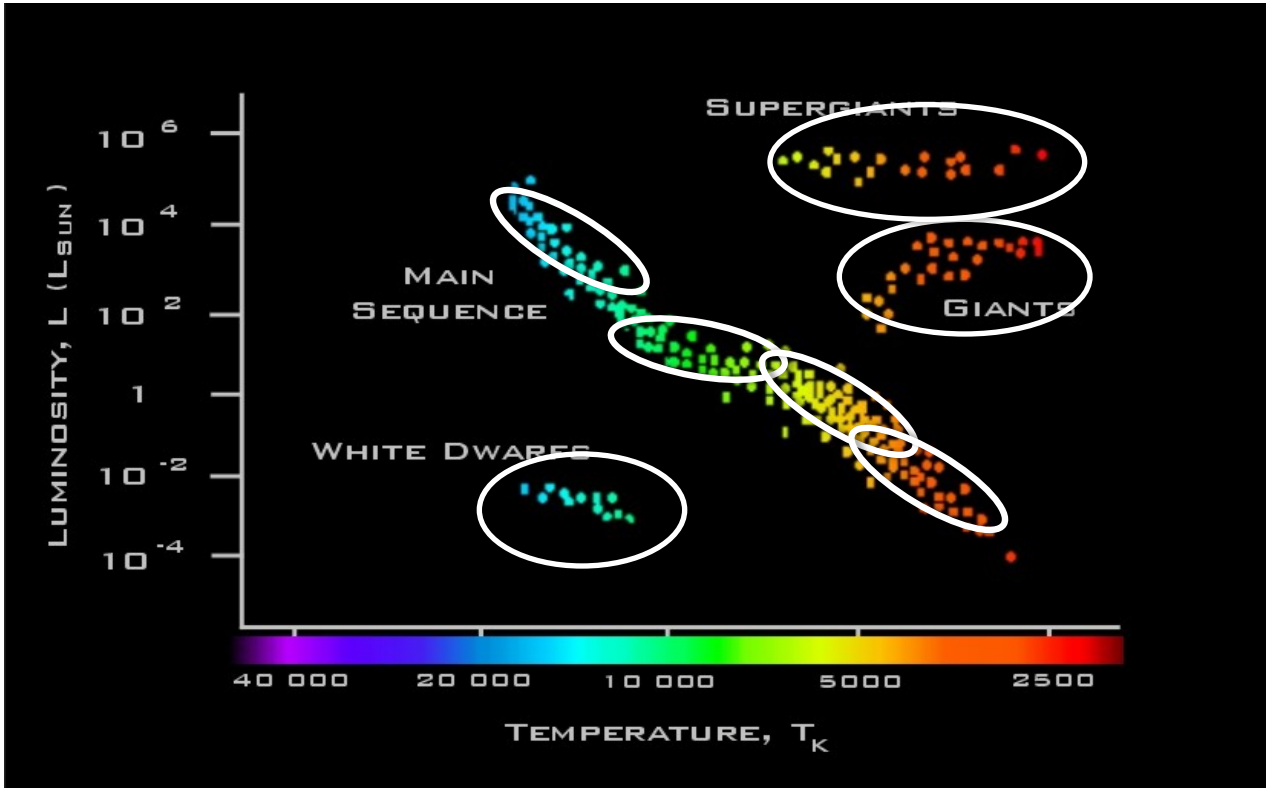
An object is an row in a data matrix X with a feature vector X_i



BASIC CLUSTERING GOAL

Compute similarity/distance between objects

Create distinct groups that minimize intra-group distance and maximize inter-group distances



2 DOMINANT TECHNIQUES

K-Means (partitioning)

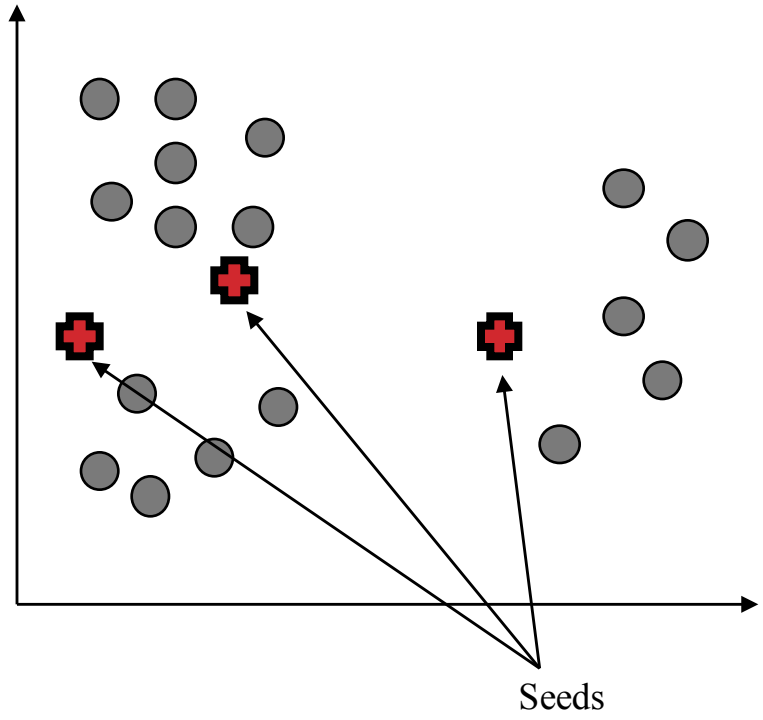
- Clusters are defined by a center point
- # groupings chosen in advance
- Each object belongs to cluster in which it has minimum distance to cluster center
- Generally cheaper, but not stable

Hierarchical

- Clusters are arranged in a nested taxonomy
- K can be chosen after the fact
- Stable but computationally expensive

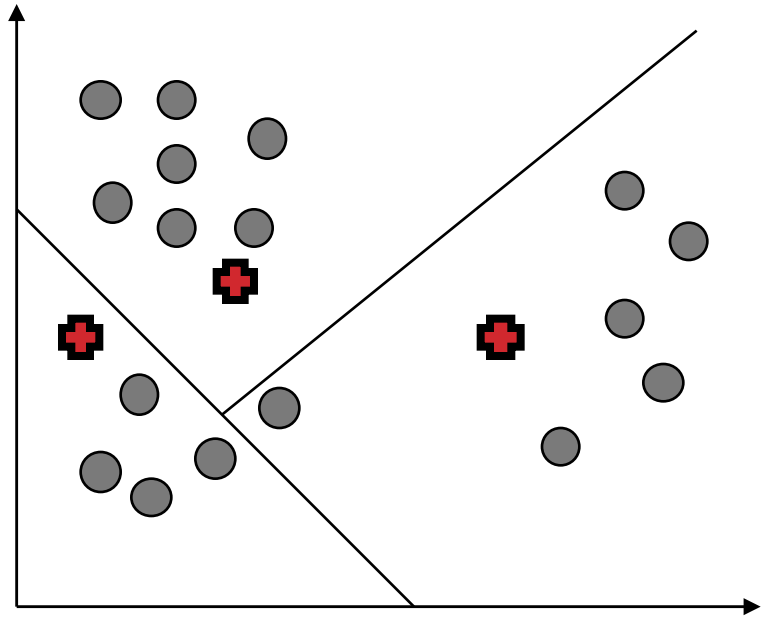
K-MEANS

1. Start with data and a predefined number of clusters, k.
2. Choose k seeds randomly.



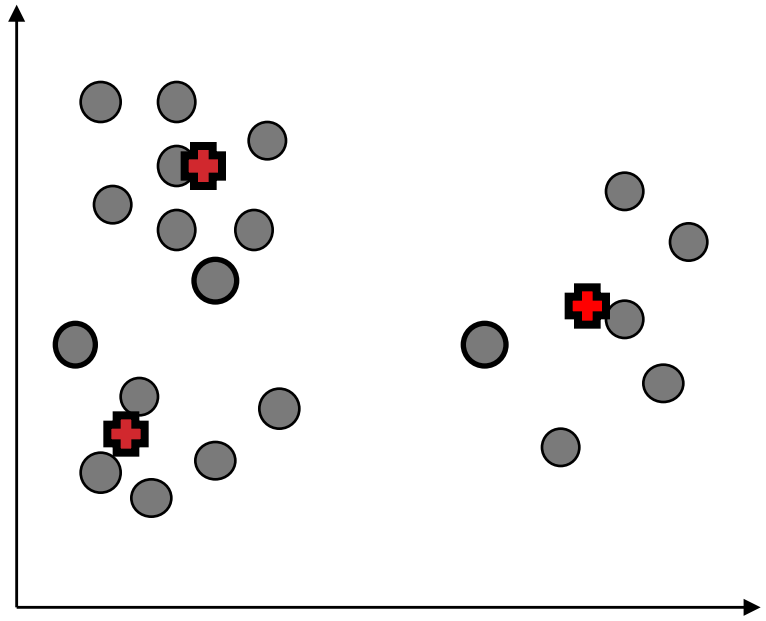
ASSIGN INSTANCES TO CLUSTERS

3. Assign each instance to the cluster for which it is closest to the cluster center.



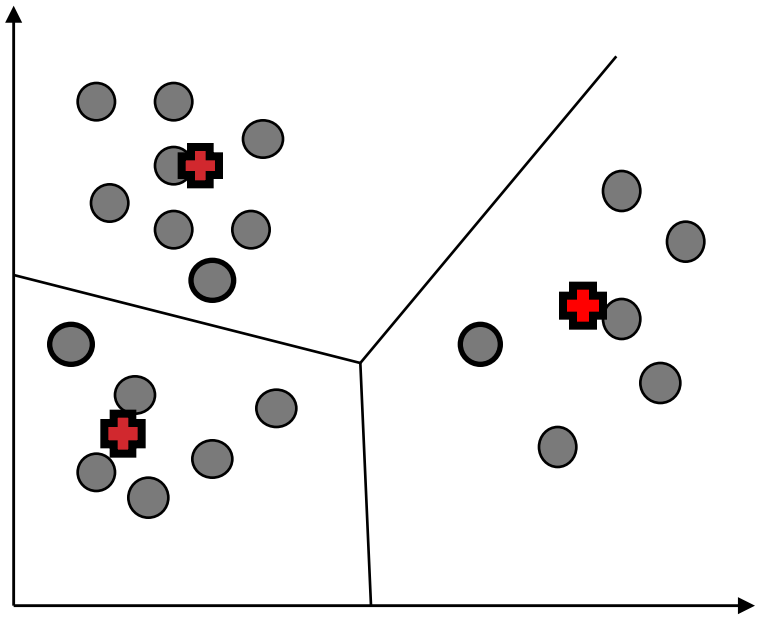
FIND NEW CENTROIDS

4. Compute new centroids per cluster, which is the avg coordinate for all objects within the cluster.



DEFINE NEW CLUSTER

5. Repeat steps 3-4 until cluster centroids converge



Demo:
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

KMEANS DISCUSSION

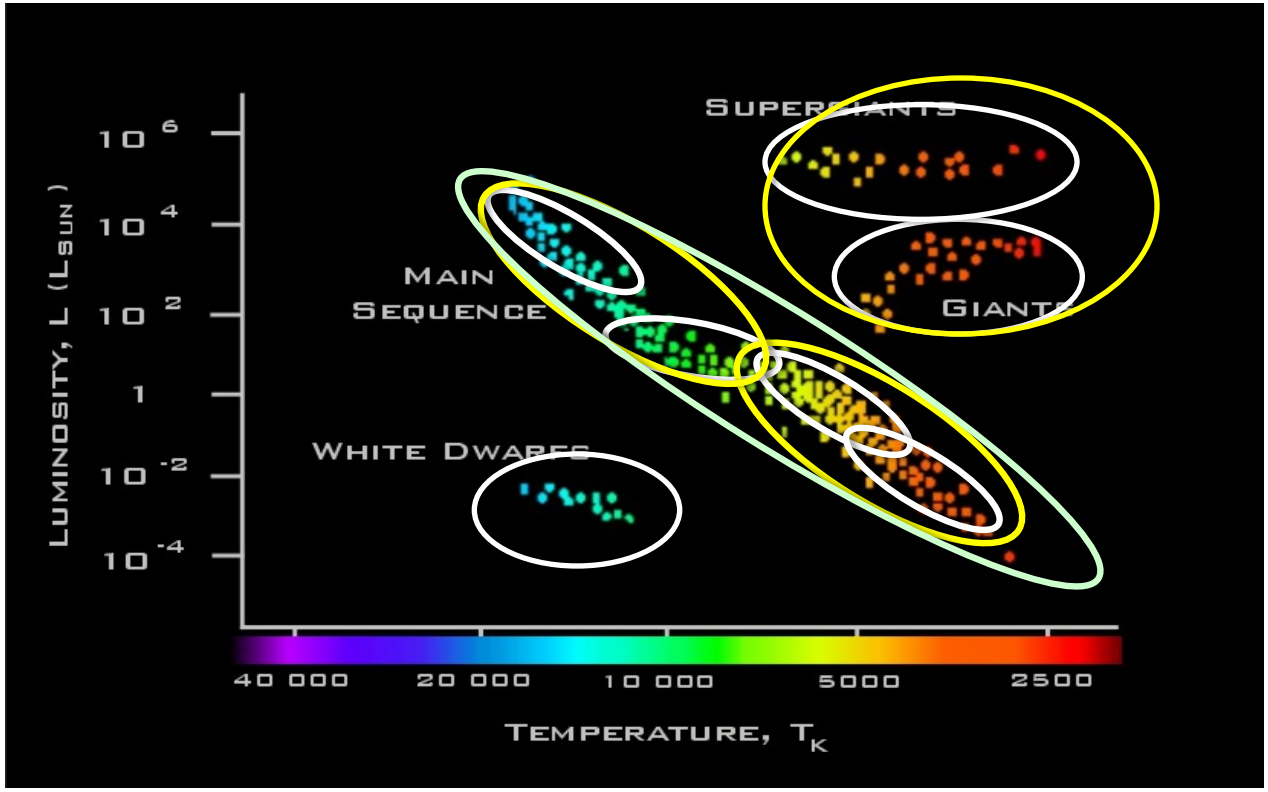
- Most common distance metric is Euclidean distance,
- but others are appropriate

- Many rules from k-NN apply here
  - Scaling of features matters
  - Can up-weight important features
  - Features need to be numeric
  - Curse of dimensionality

- Assessment
  - What is best k?
  - Is this a good fit?
  - What do the clusters mean?

CLUSTER HIERARCHY

Clusters can be embedded in other clusters.
Hierarchical clustering attempts to uncover this embedding.



GENERIC ALGORITHM

This is the generic algorithm for agglomerative clustering methods.

- Compute all pairwise similarities
- Place each instance into its own cluster
- Merge the two most similar clusters into one
  - Replace two clusters into the new cluster
  - Recompute intercluster similarity scores

- Repeat until there are only k clusters left

EXAMPLE

In our first homework all students filled out a DS profile self-assessment

6. In each box, give a ranking of 1-10 on well you think you are in the category listed. A 1 should mean you are a complete novice and a 10 means you are pretty much an expert.

Data Visualization	
Computer Science	
Mathematics	
Statistics	
Machine Learning	
Business Strategy	
Communication	

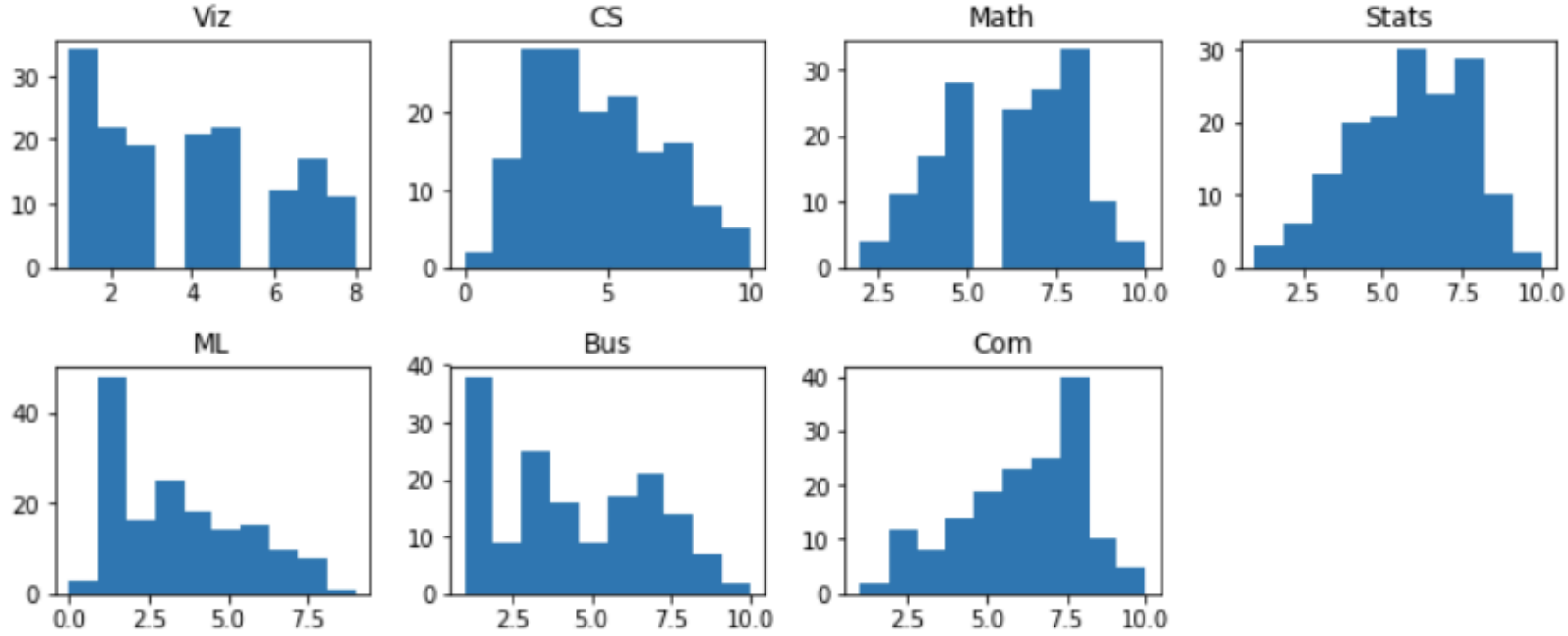
APPLICATION

Goal :
assign students into study groups of size 4, where each study group is maximally diverse with respect to skill distribution

We'll use cluster analysis to segments students into roughly equal sized groups. We'll also attempt to qualify each group.

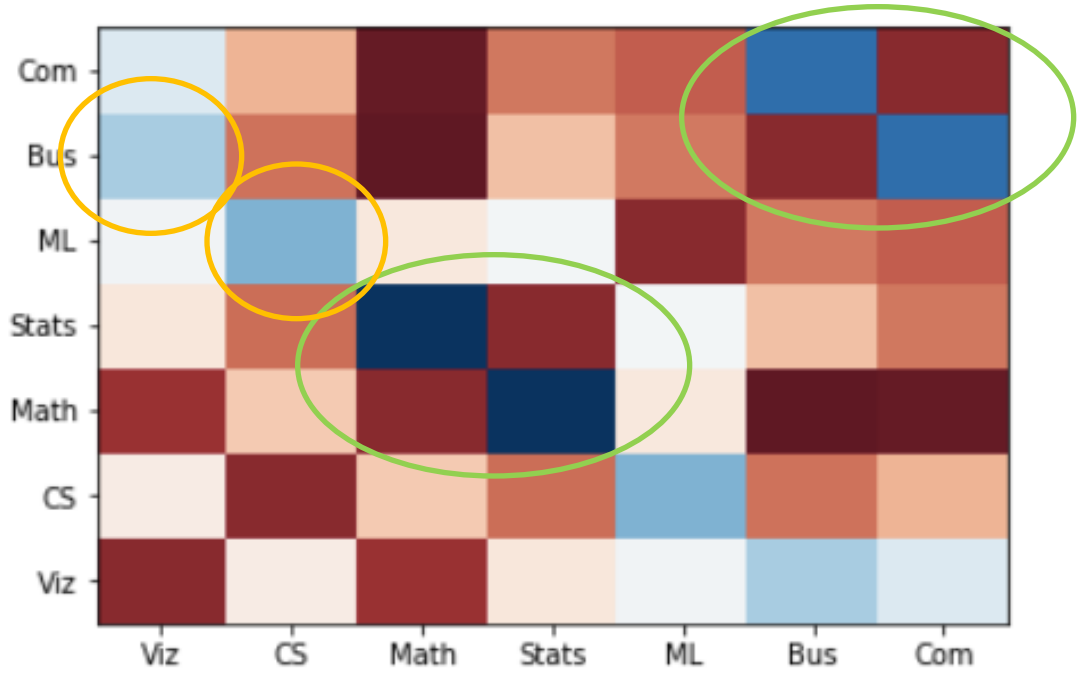
FIRST LETS UNDERSTAND OUT DATA - DISTRIBUTIONS

Simple histograms of fields in question will give us a sense of the data distributions



FIRST LETS UNDERSTAND OUR DATA - COVARIANCE

Visualizing the covariance matrix will give us a sense of which variables are more correlated to each other

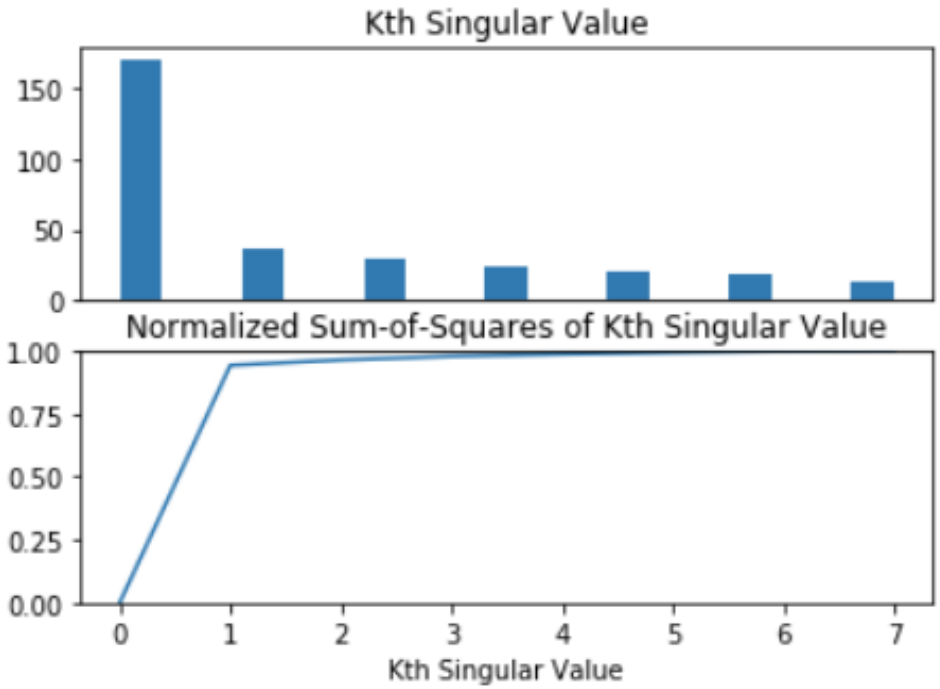


The covariance matrix shows us two pairs of variables with very high correlation: {Business & Communication skills} and {Math and Statistics skills}

We also find slightly less but significant correlation between two other pair: {Business & Visualization skills} and {CS and ML skills}

PREPPING DATA FOR CLUSTERING – SVD DIM REDUCTION

Our covariance analysis reveals strong redundancy amongst the features. We may be better off reducing the dimensions with SVD



The skew-ness of the singular value distribution confirms our hypothesis of data redundancy

This data set is a good candidate for SVD dim reduction prior to clustering.

ONCE WE BEGIN CLUSTERING, WE NEED TO THINK ABOUT EVALUATION

Goodness of fit

Use metrics like “Inertia” and “Silhouette coefficient” to determine which value of k to commit to.

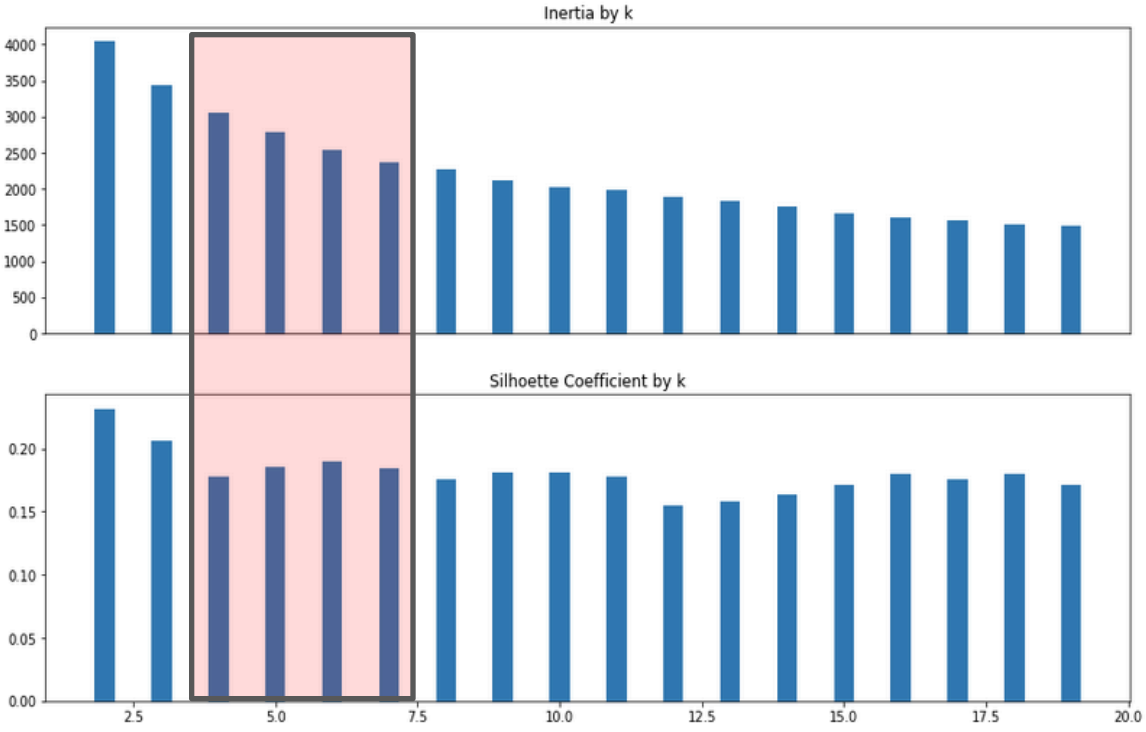
Cluster Distribution

Do the clusters have practical distributions across them.

Interpretation

Do the clusters have meaningful and useful interpretations.

WE CAN QUANTIFY GOODNESS OF FIT TO IDENTIFY THE RIGHT K



This region looks like the area that has the best fit

The *inertia* measures the within cluster sum of squares distance from the mean.

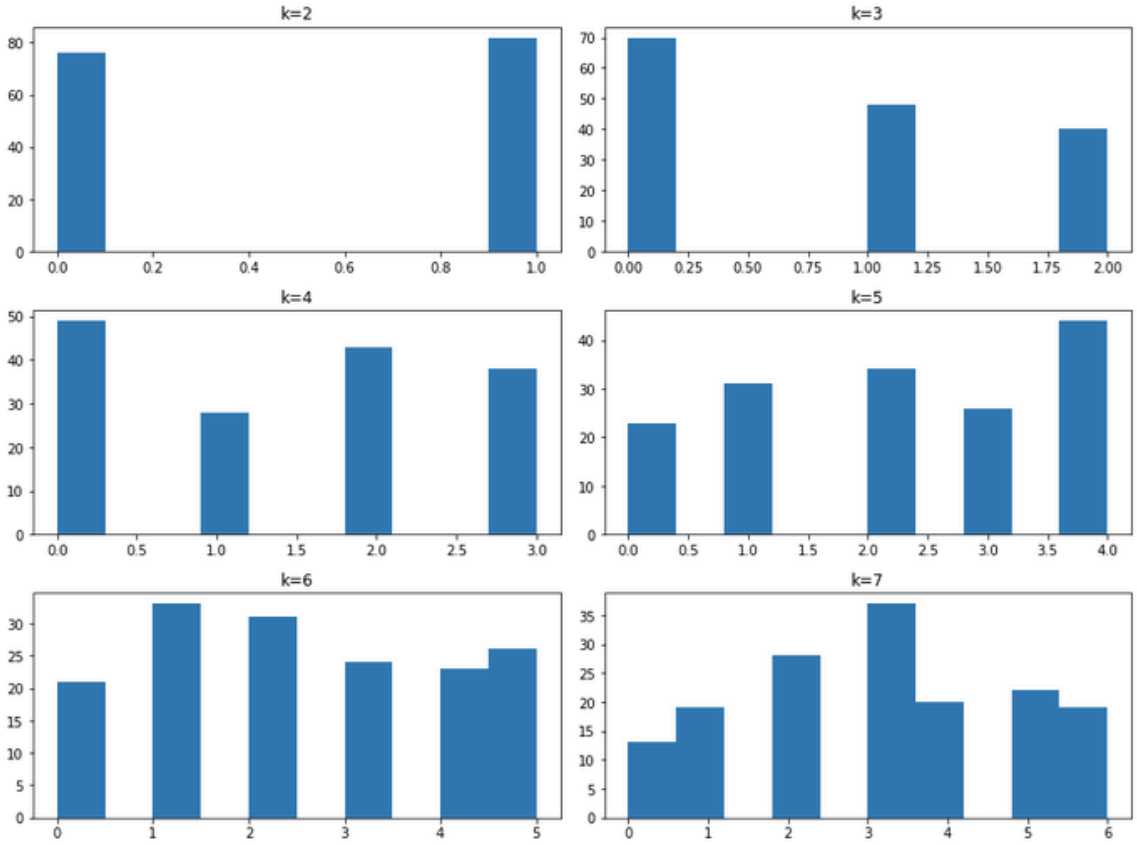
$$inertia = \sum_{j=1}^{k} \sum_{x_i \in C_j} |x_i - \mu_j|^2$$

a = the mean distance between a sample and all other points in the same cluster

b = the mean distance between a sample and all other points in the nearest cluster

$$s = \frac{b - a}{max(a, b)}$$

CLUSTER DISTRIBUTIONS ARE ANOTHER MEANS TO CHOOSE K

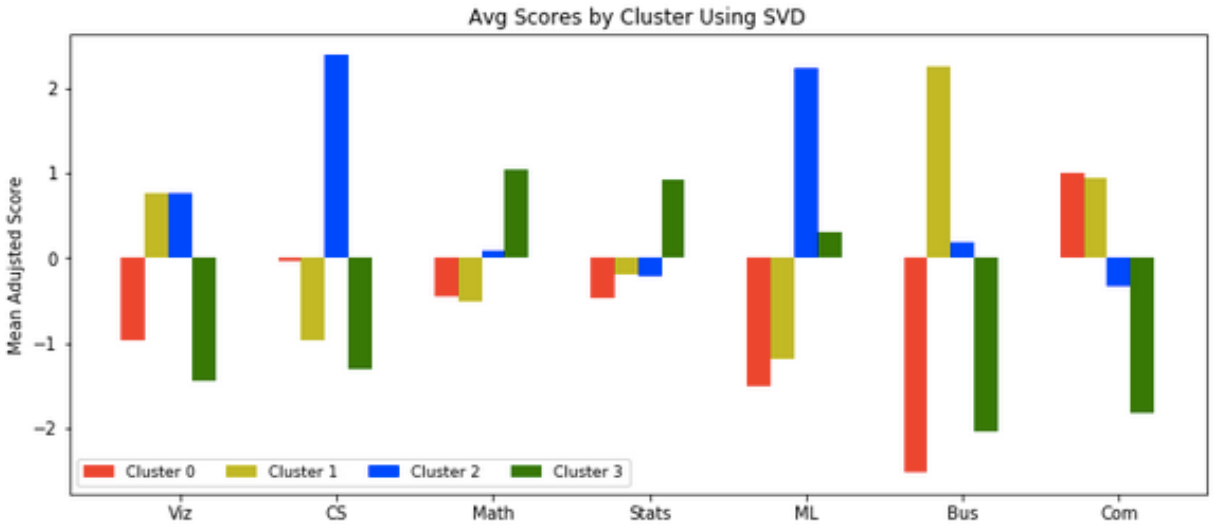


Depending on our expected application of the clusters, we may want our clusters to have a certain distribution.

In many cases, we want to avoid cluster distributions that are too skewed.

EXPLAINING CLUSTERS

Often the best way to explain clusters is to look at the values of the global mean adjusted cluster centroid.



Cluster 0 is generally below avg on all points except Com (perhaps these are people that are more conservative with self-ratings)

Cluster 1 is the business & communications group

Cluster 2 is the group dominant in CS and ML

Cluster 3 is the group dominant in math and stats