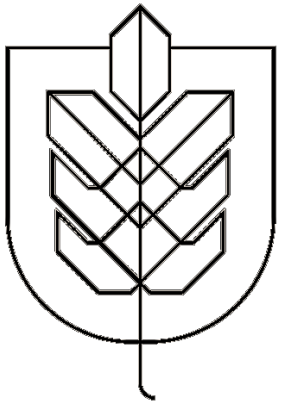
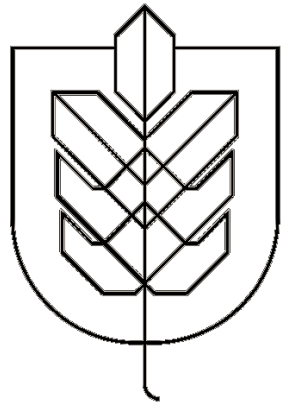


Distance- and Density-Based Clustering

Unsupervised Learning

Reminder!



Distance-based unsupervised learning techniques require variance scaling!

Modeling in a linear space requires standardized distance

All variables *should* be continuous and on the same scale

Can we still use categorical variables?

11 famous products that were originally intended for a completely different purpose

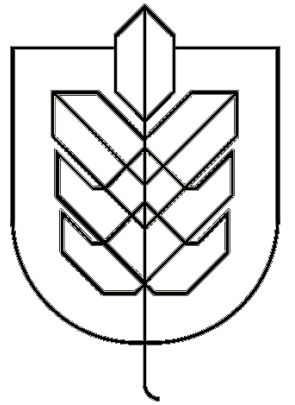
Will Heilpern Apr. 1, 2016, 10:50 AM

Some of the best discoveries happen by accident. As a result, many of the world's most famous brands and products started out doing something completely different to that what they are known for today.

Bubble wrap was initially used for something completely different. Getty

They range from soft drinks that were originally laced with powerful mind-altering drugs to, medicines with unexpected, but profitable, side-effects.

The Rules Were Meant to Be…



Machine learning models rely on assumptions.

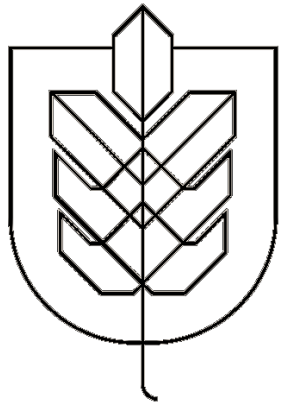
Violations can be measured mathematically

… but that is beyond the scope of this course.

Knowing a model's assumptions is the second most valuable aspect of the machine learning process

…right behind domain knowledge

Fundamentals of Clustering

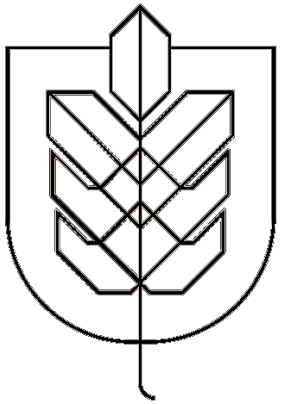


Objective:

Divide the data into groups (clusters).

+ Observations in the same cluster should be similar.
+ Observations in another cluster should be different.

What does it mean to be similar or different?



Agglomerative Clustering

Different ways to be different

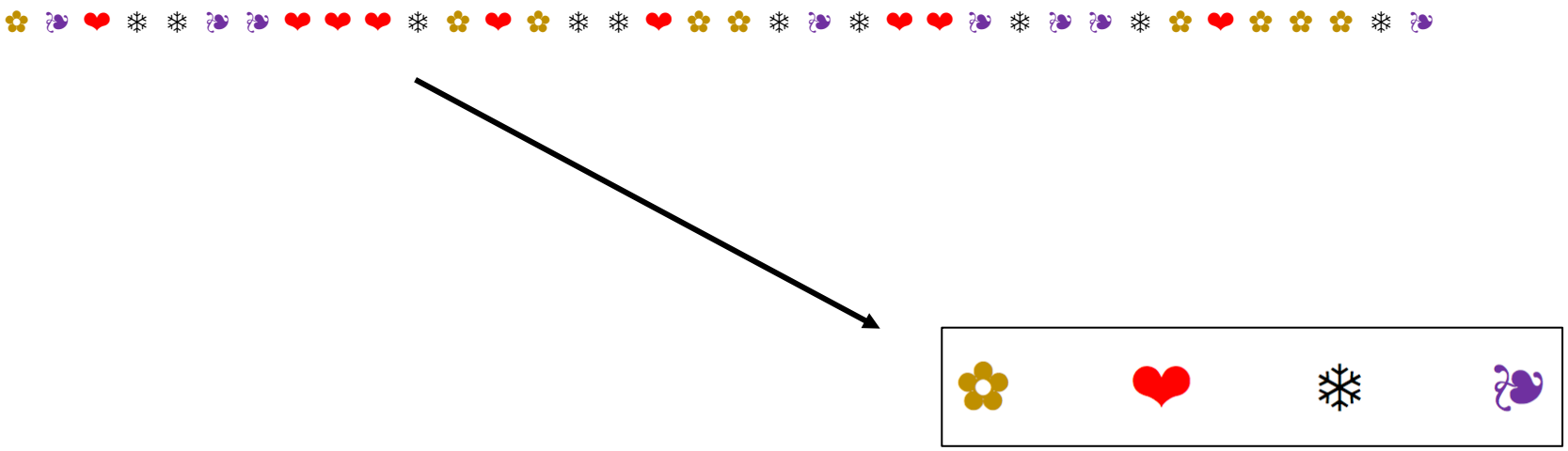
Agglomerative Clustering

Starts with the assumption that each observation is unique.

36 data points, each represented as its own cluster.

Looking Deeper into the Data

Then looks for similarities in the data.



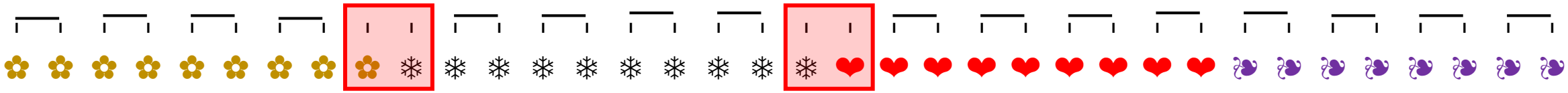
Looking Deeper into the Data

Then looks for similarities in the data.

Looking Deeper into the Data

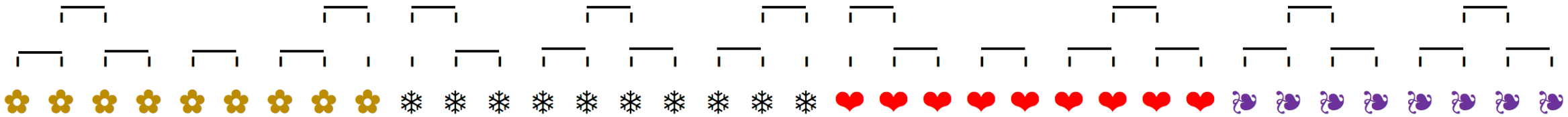
And groups observations accordingly.

Some observations are still
by themselves.



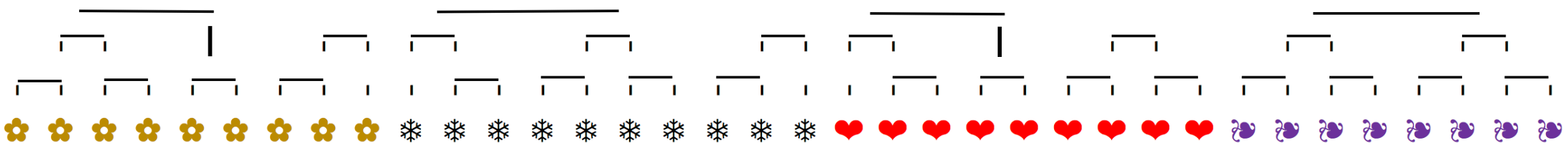
Looking Deeper into the Data

Clustering continues…



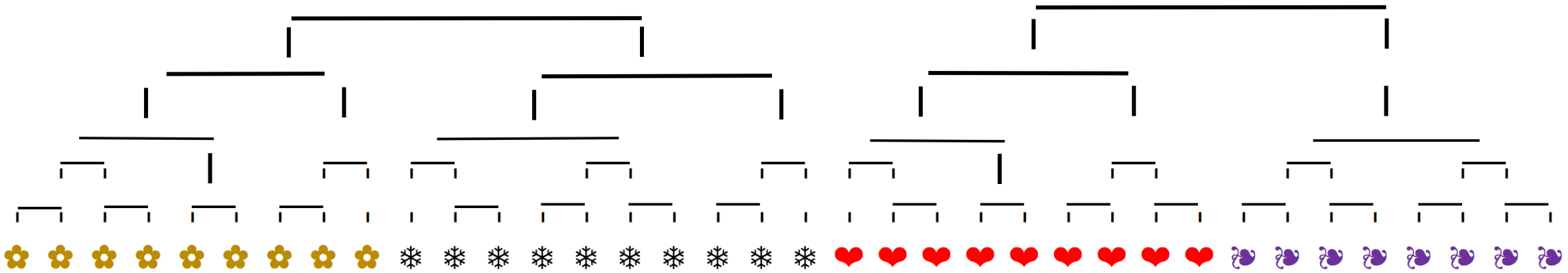
Looking Deeper into the Data

Clustering continues…



Looking Deeper into the Data

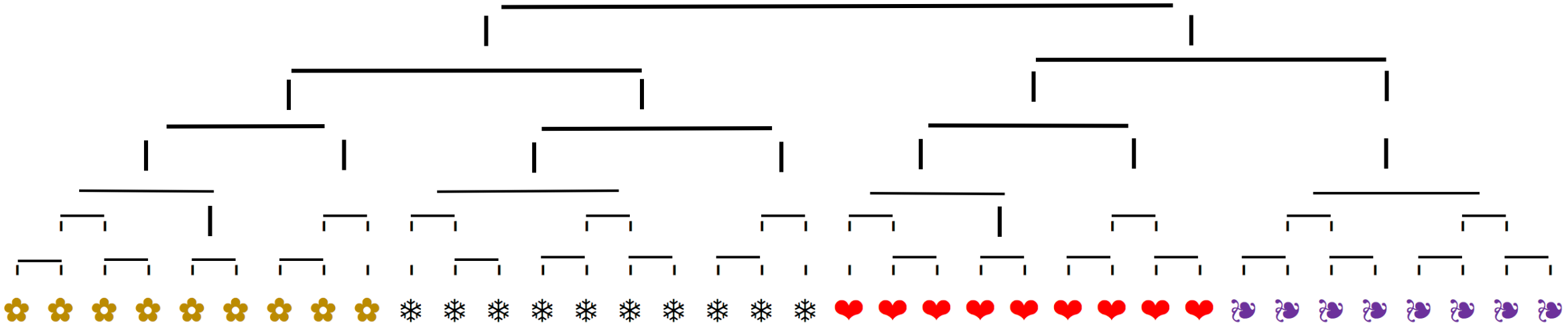
Clustering continues…



Looking Deeper into the Data

Clustering continues…

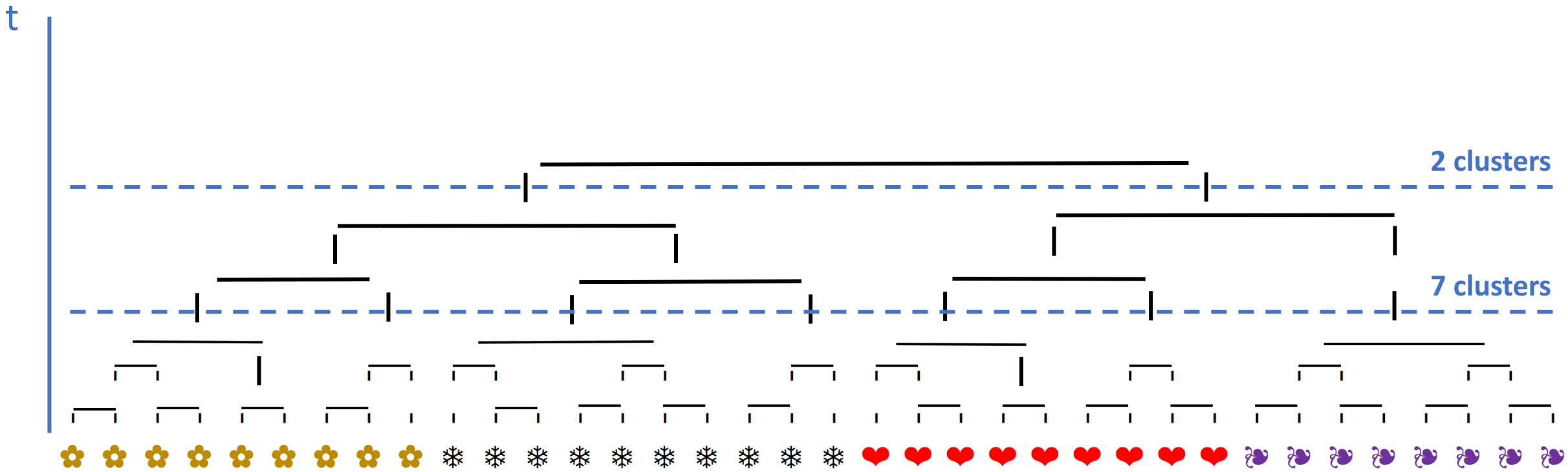
…until all of the data is in one cluster

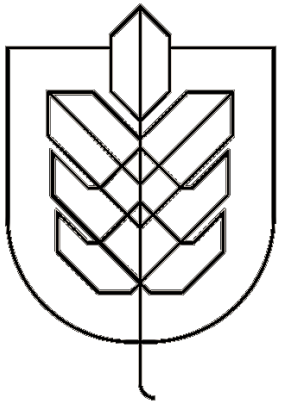


Looking Deeper into the Data

…or until we specify a stoppage parameter.

In K-Means, this would be a number of clusters.

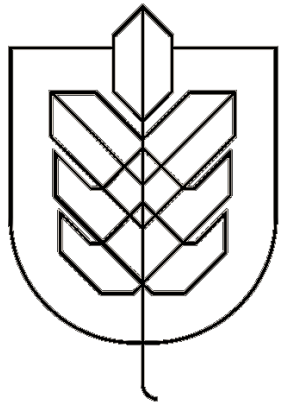




K-Means Clustering

Expanding on measures of center

K-Means Clustering



Iterative algorithm for when we know how many clusters to create

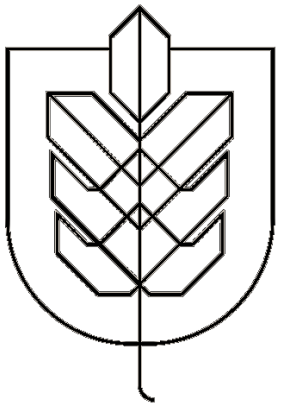
Attempts to find optimal points to act as cluster centers

→ means for each feature

Key Advantage

Can predict on new data

K-Means Clustering - Python

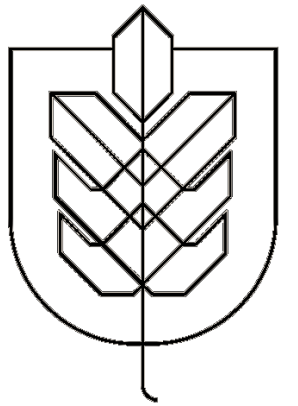


```
138
139 from sklearn.cluster import KMeans # k-means clustering
140
141 # Creating a model with 3 clusters
142 customers_k3 = KMeans(n_clusters = 3,
143                       random_state = 508)
144
145
146 # Fit model to points
147 customers_k3.fit(df)
148
```

Instantiating a KMeans model with three clusters

Fitting the model instance to the data

K-Means Clustering - Python



```
150
151 # Checking to see if we got the same clusters as when using fcluster
152 kmeans_clusters = pd.DataFrame({'cluster': customers_k3.labels_})
153
154 print(kmeans_clusters.iloc[: , 0].value_counts())
155
156 centroids = customers_k3.cluster_centers_
157
158 centroids_df = pd.DataFrame(centroids)
159
```

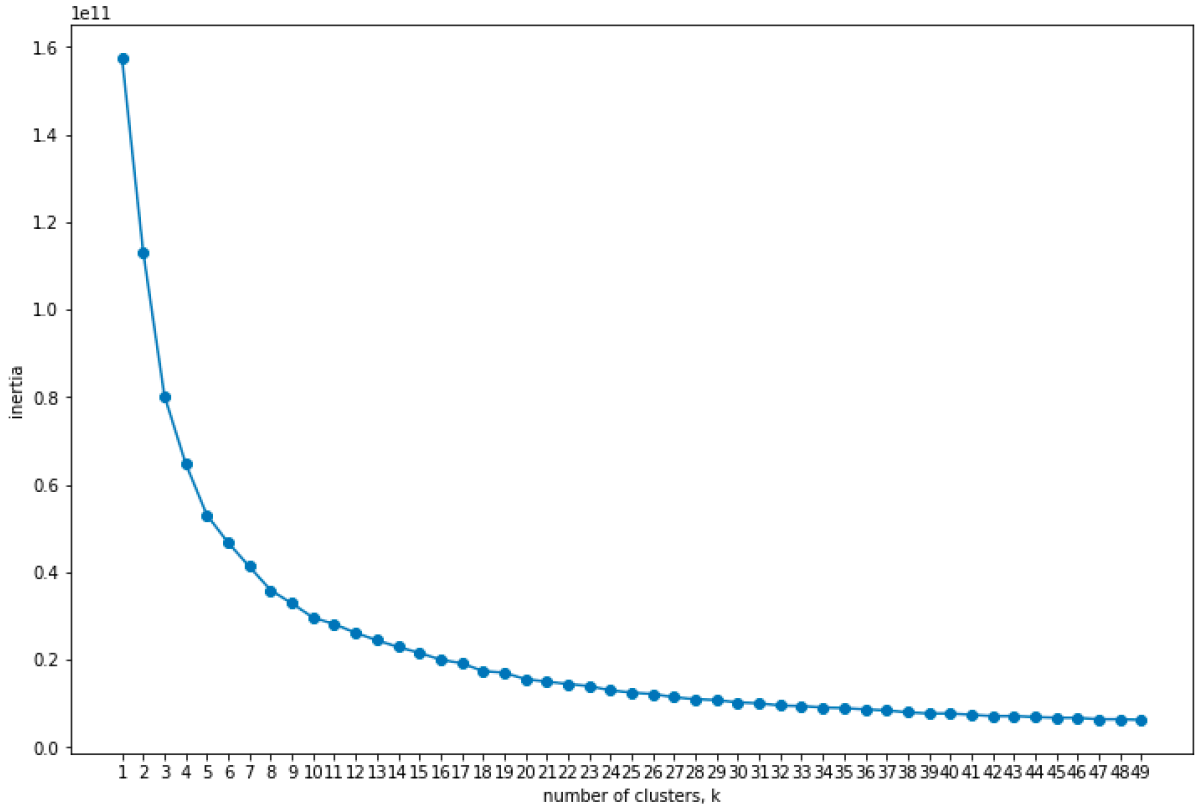
Extracting cluster labels (cluster number for each observation)

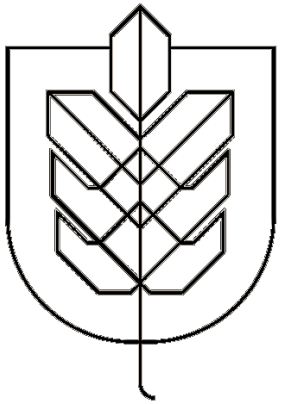
Extracting cluster centers (means of each feature)

How Many Clusters?

Inertia - within cluster sum of squares

Lower inertia is better





Practice in Python