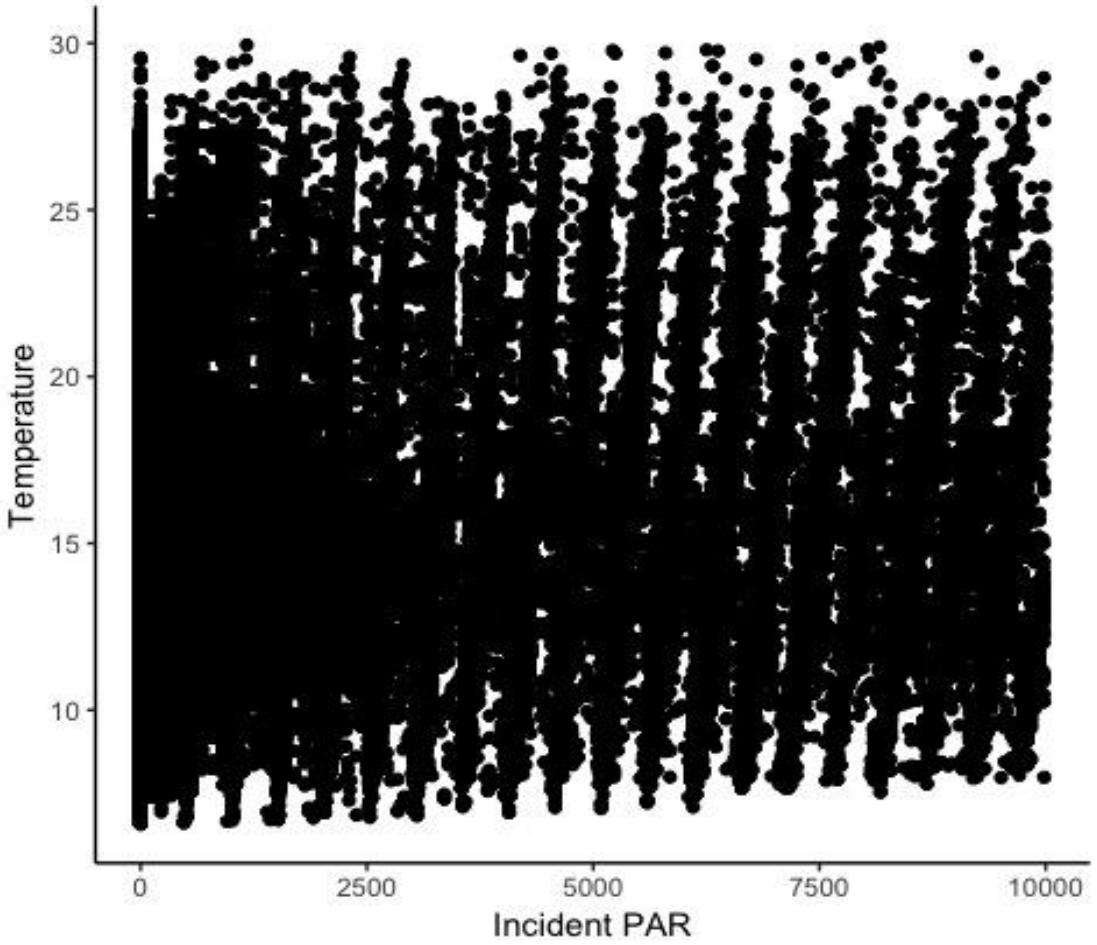
Stat 215A - Week 5

Clustering

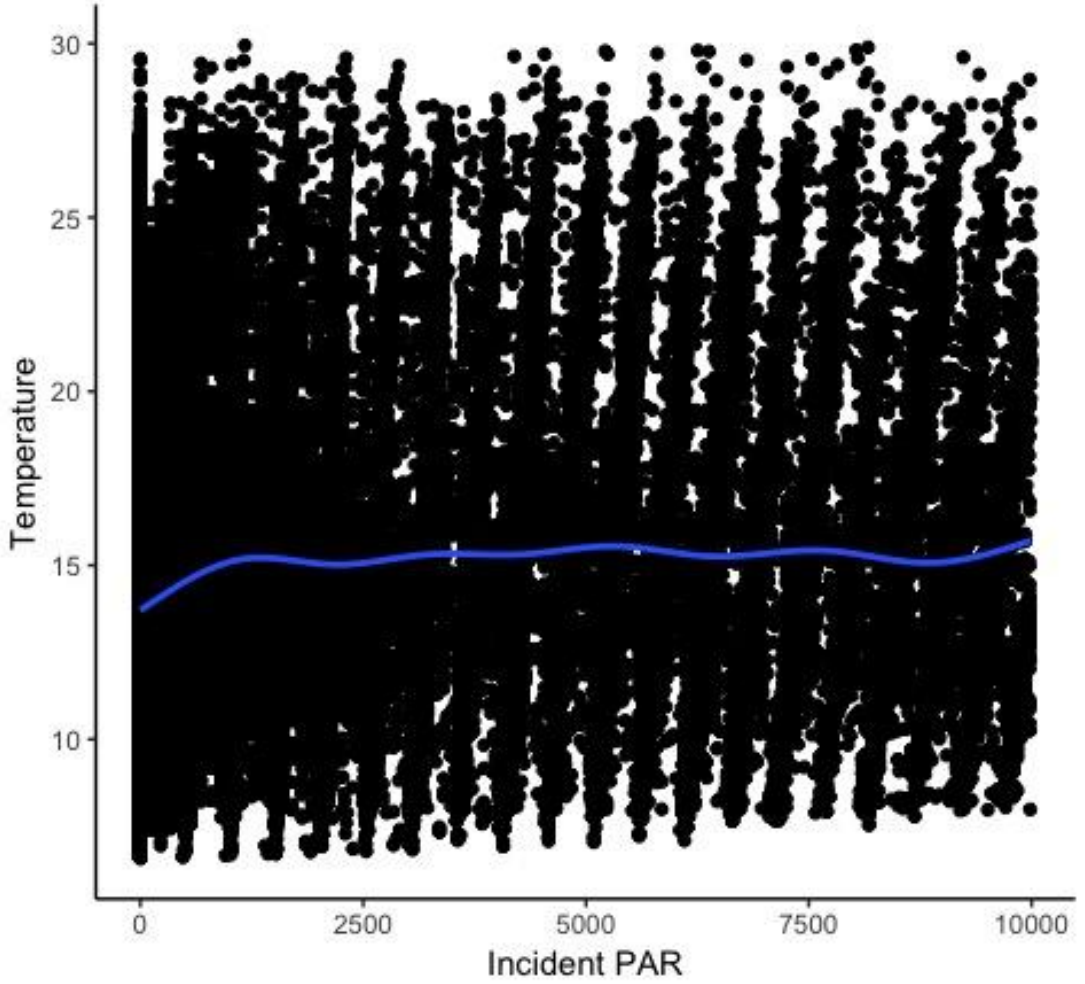
Lab 1 Discussion

Overplotting



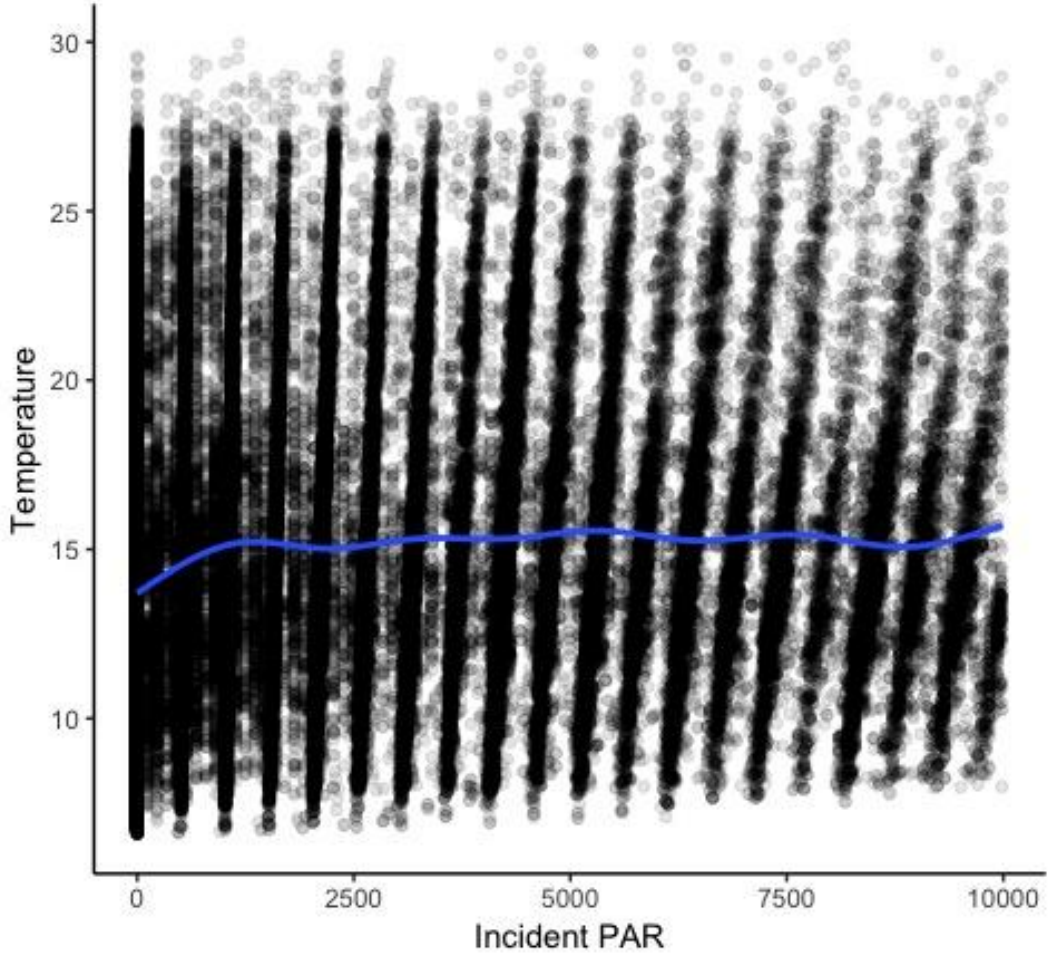
Overplotting

Add trendline



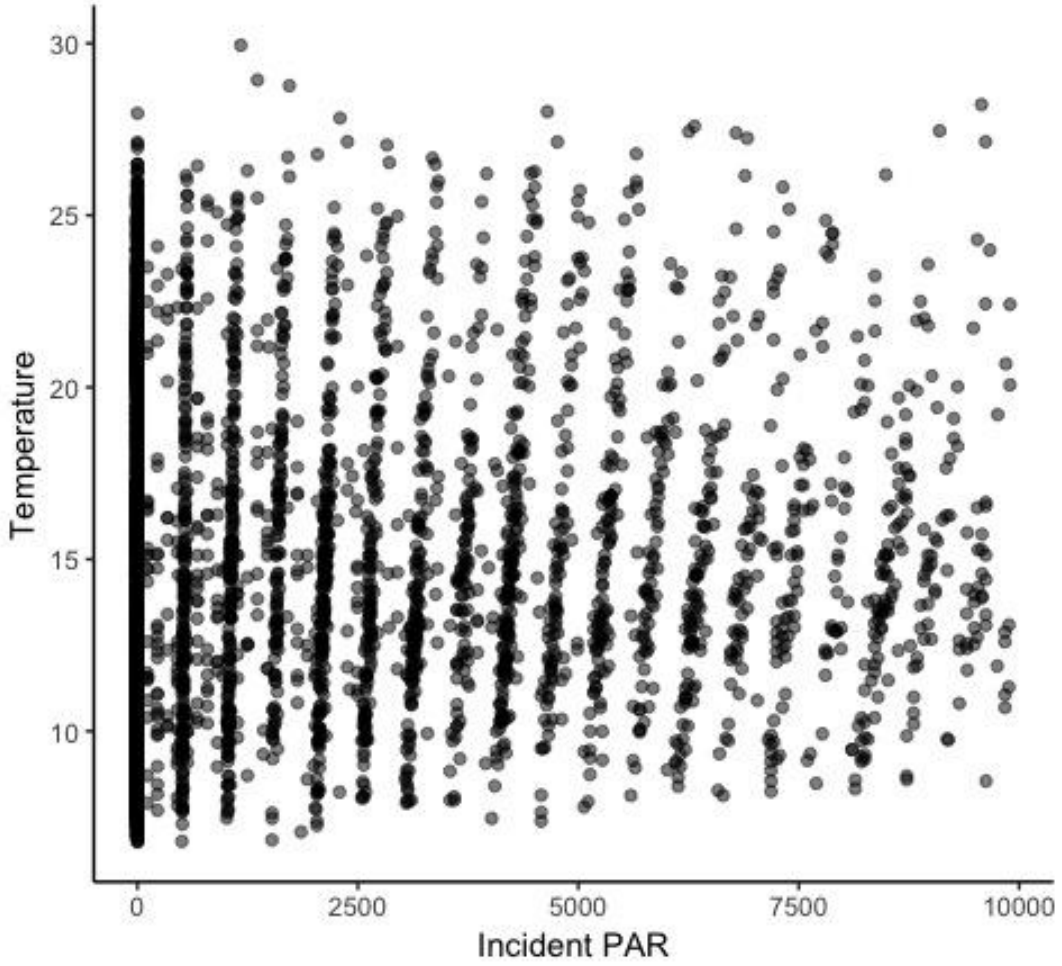
Overplotting

Use transparency



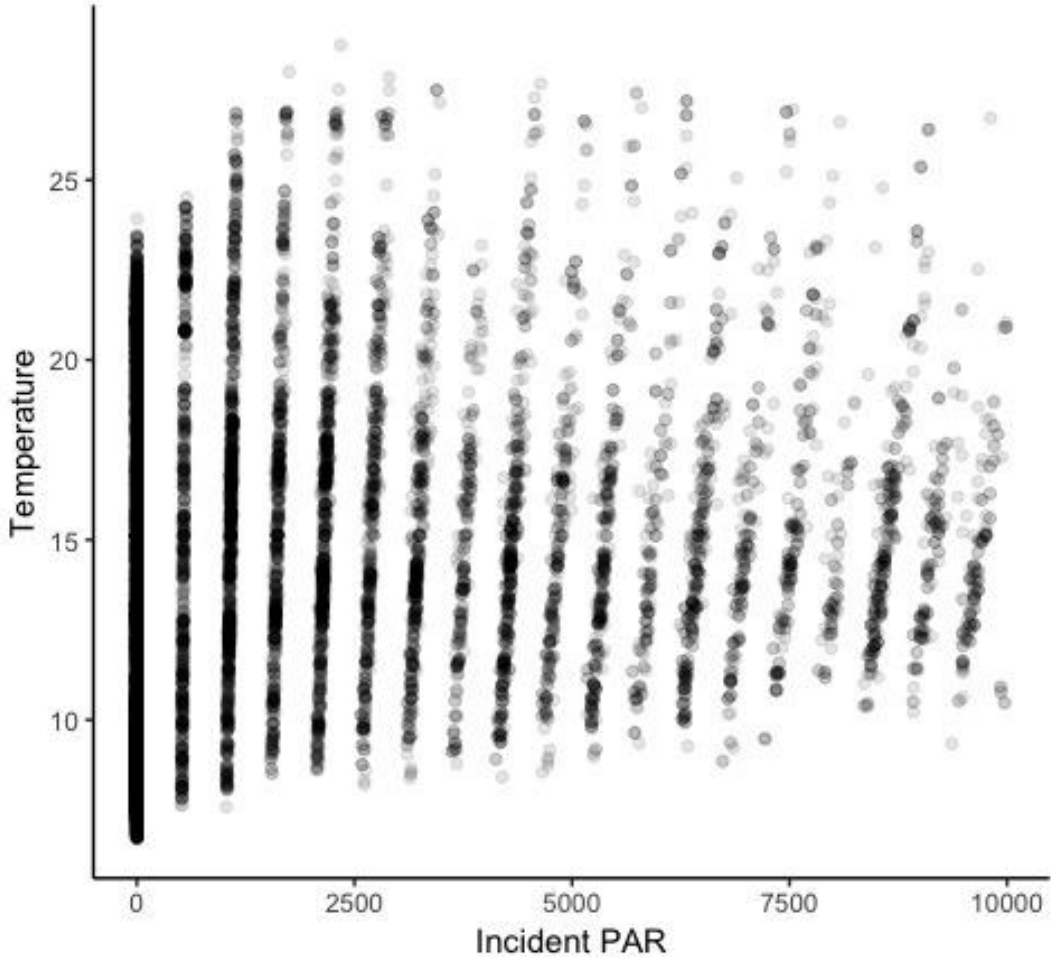
Overplotting

Subsample 10,000 points

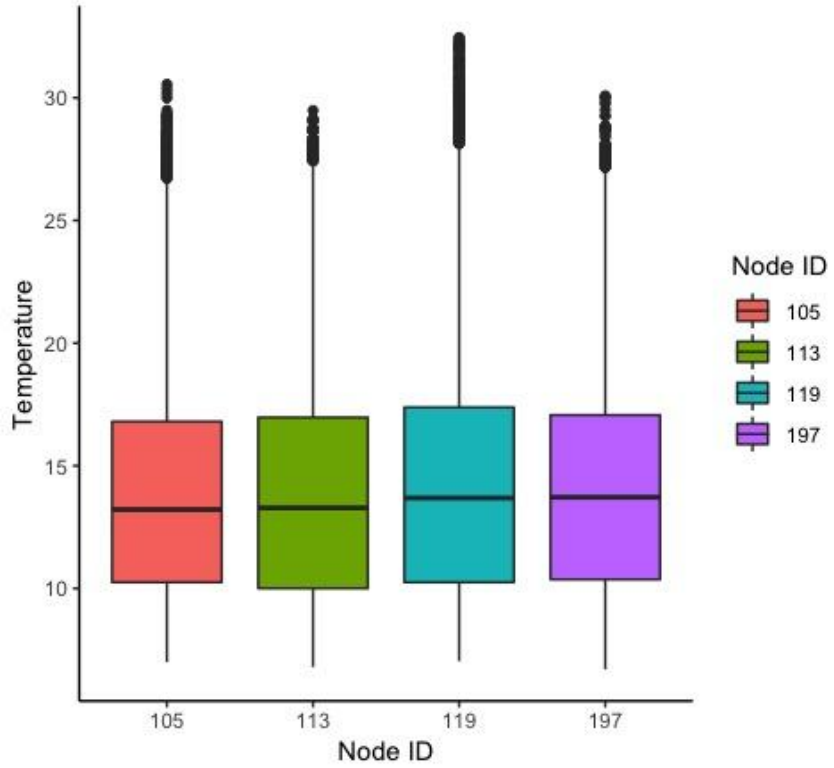


Overplotting

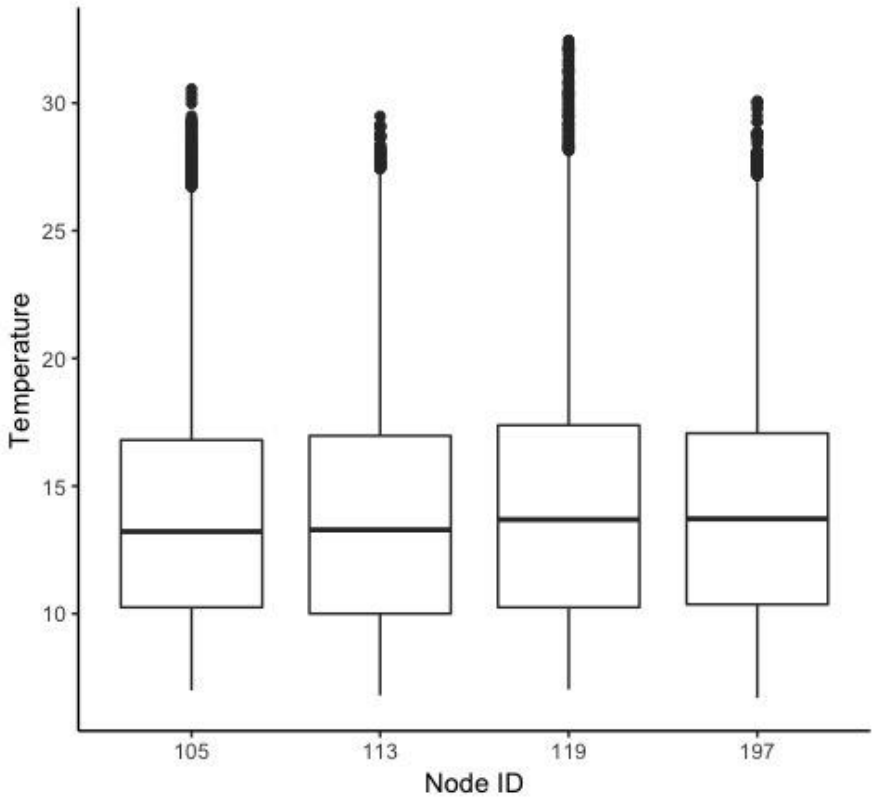
Subsample a single node



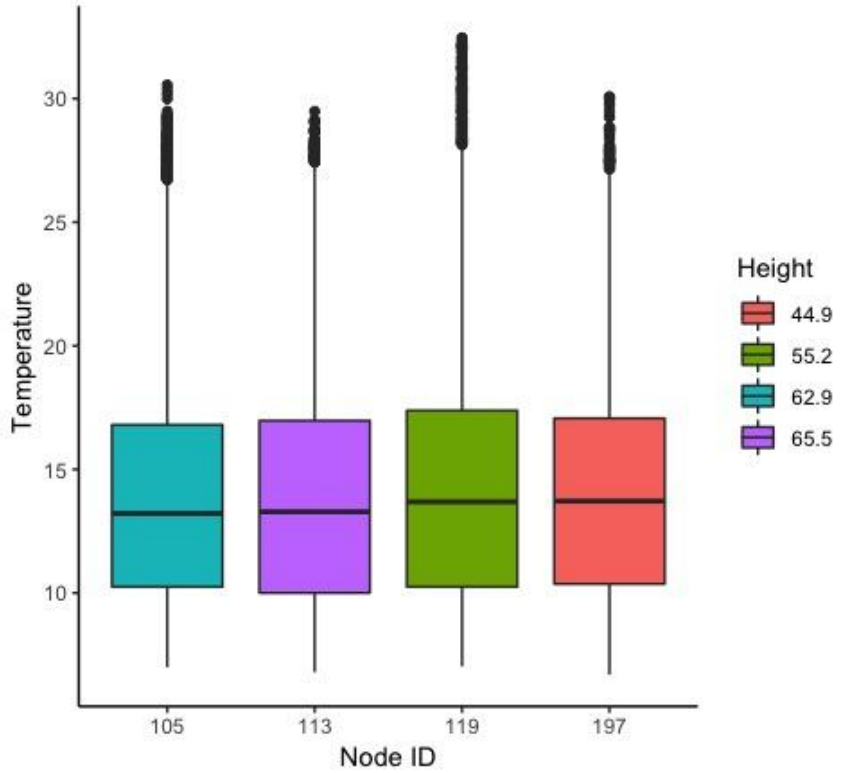
Using color appropriately



Using color appropriately

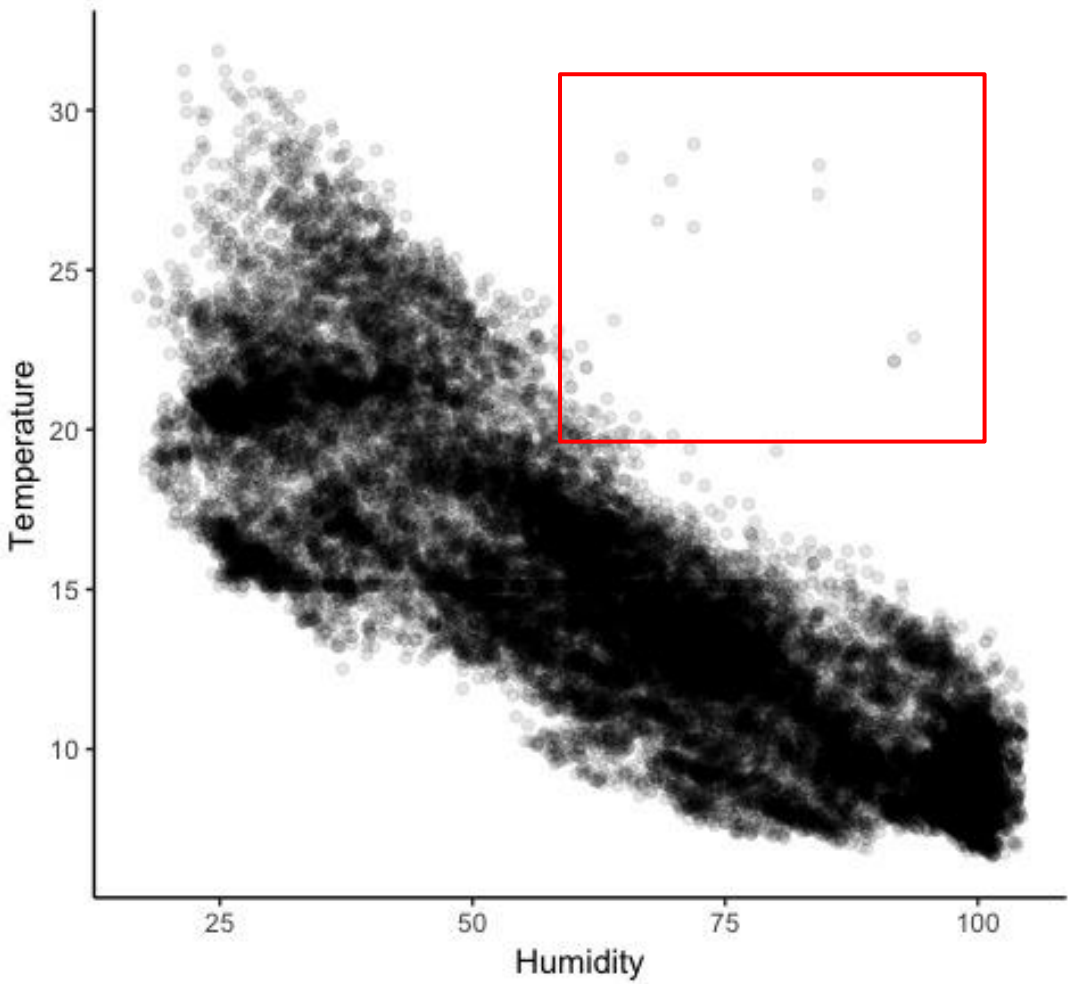


Using color appropriately



Data cleaning

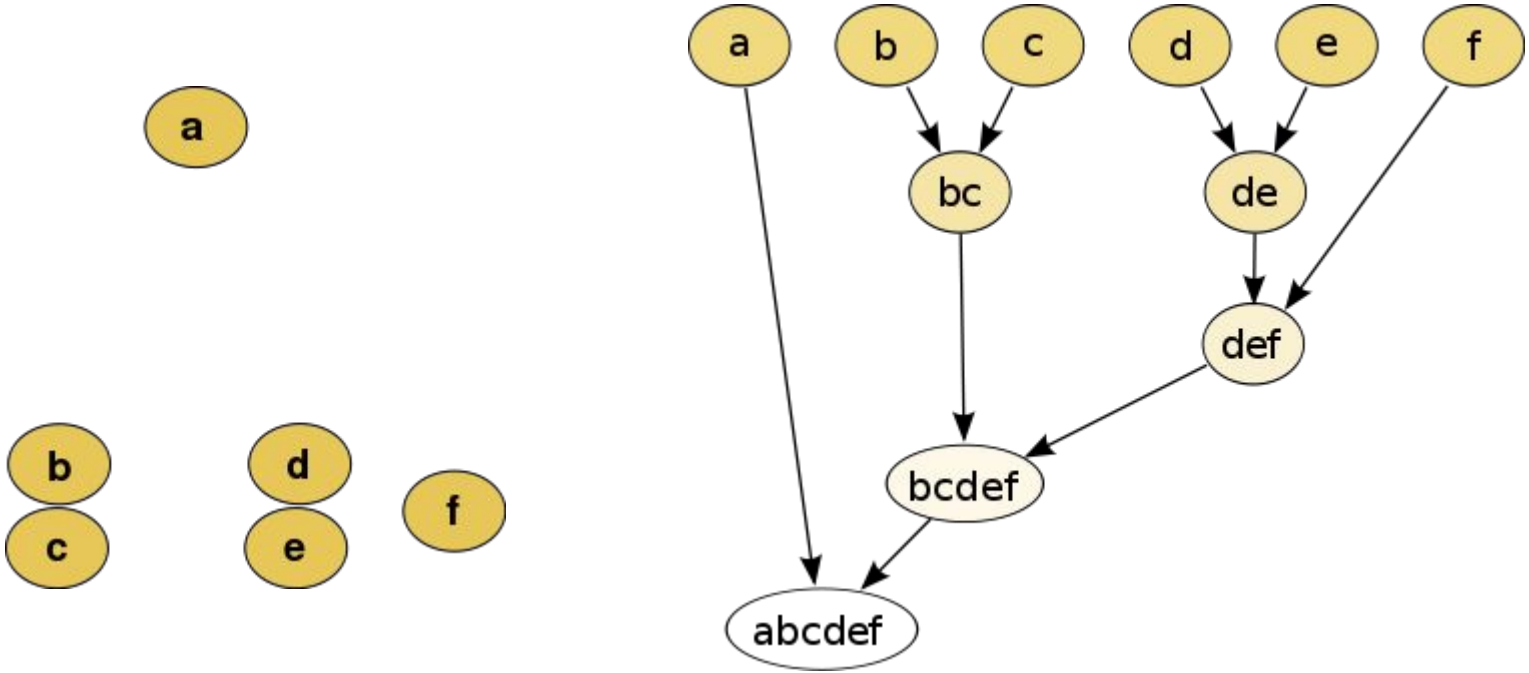
Problematic to only clean by apply bounds to the variables.



Clustering

K-means, spectral clustering, hierarchical clustering

Hierarchical clustering



https://en.wikipedia.org/wiki/Hierarchical_clustering

Hierarchical clustering

Pros:

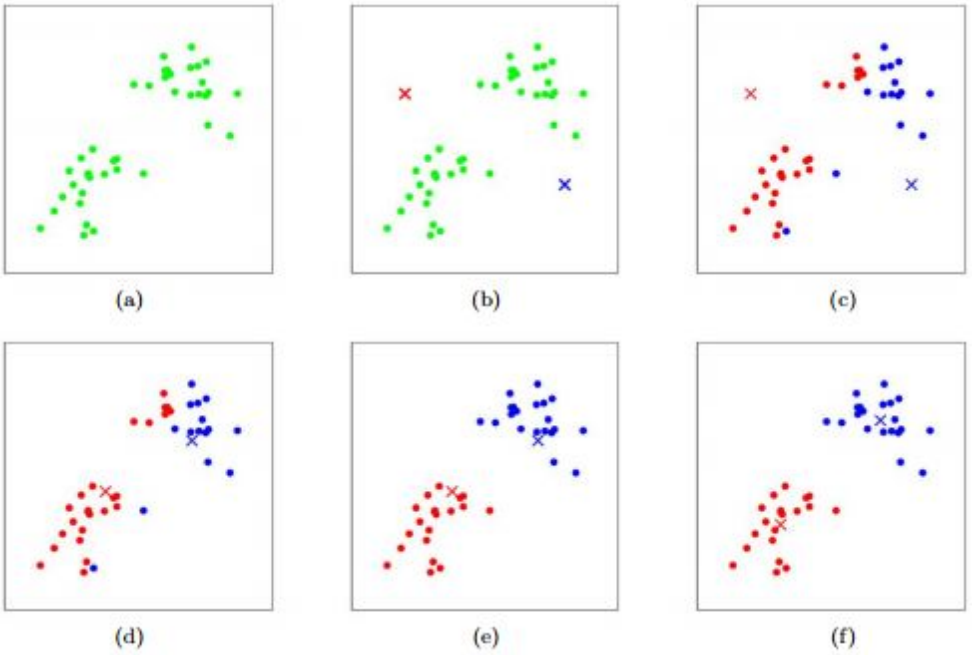
- Does not require choosing number of clusters
- Not sensitive to choice of distance metric
- Good when underlying data has a hierarchical structure

Cons:

- Slower than k-means

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

K - means



http://stanford.edu/~cpiech/cs221/handouts/kmeans.html

K-means: pros and cons

Pros:

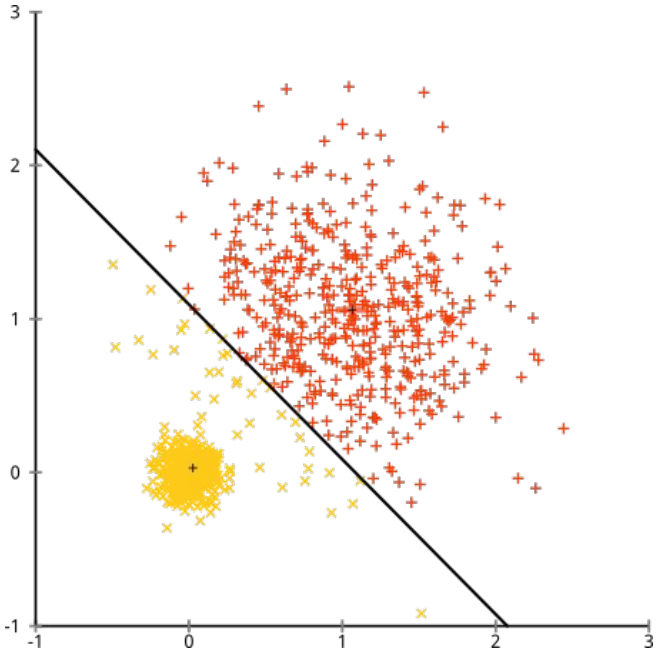
- Linear complexity: *O(n)*

Cons:

- Have to manually choose the number of clusters
- Randomly choosing starting points for clusters
- Assumes the variance of each variable is spherical
- Assumes all variables have the same variance

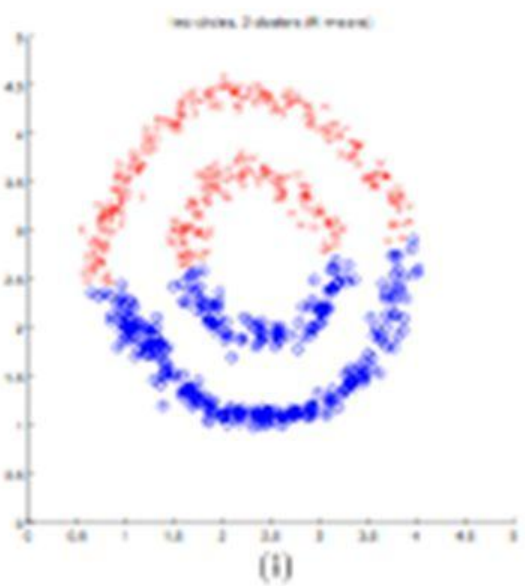
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

K-means: different variances

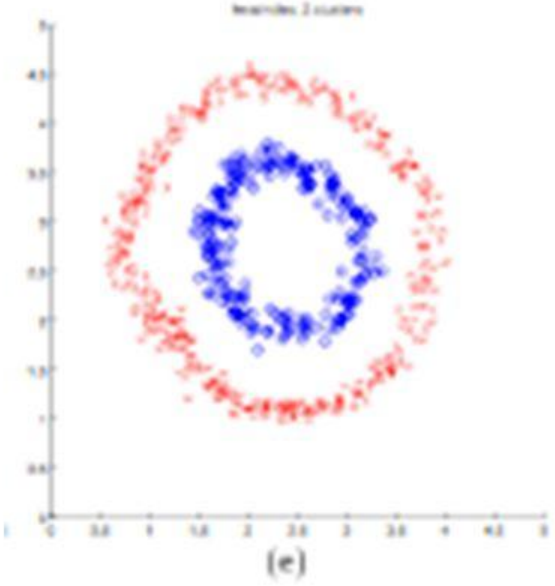


https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Spectral clustering



K-means



Spectral Clustering

https://slideplayer.com/slide/4855874/

Spectral clustering

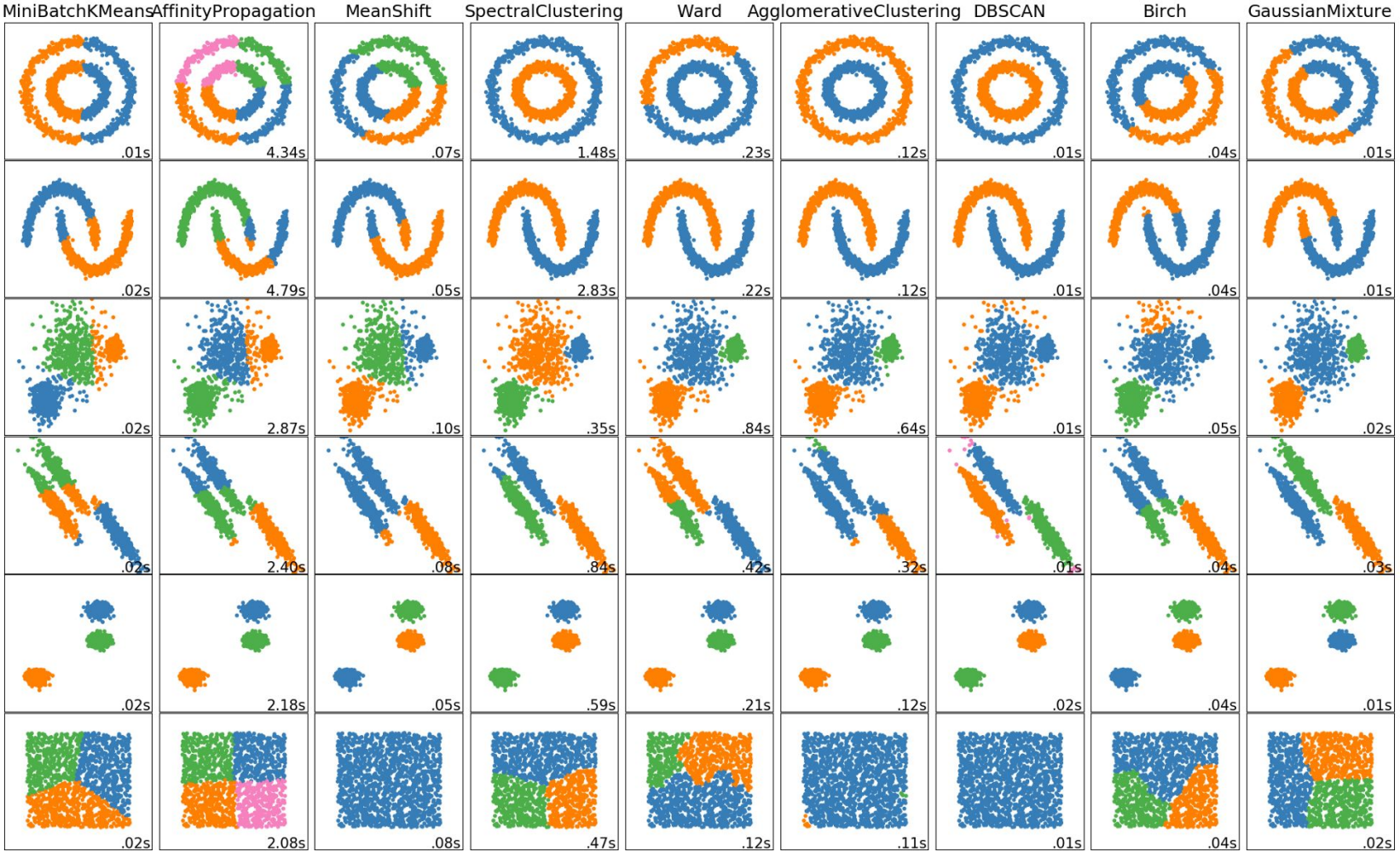
Pros:

- Good for clustering data with more complex shapes
- Reasonably fast if the data is sparse
- Produces good results

Cons:

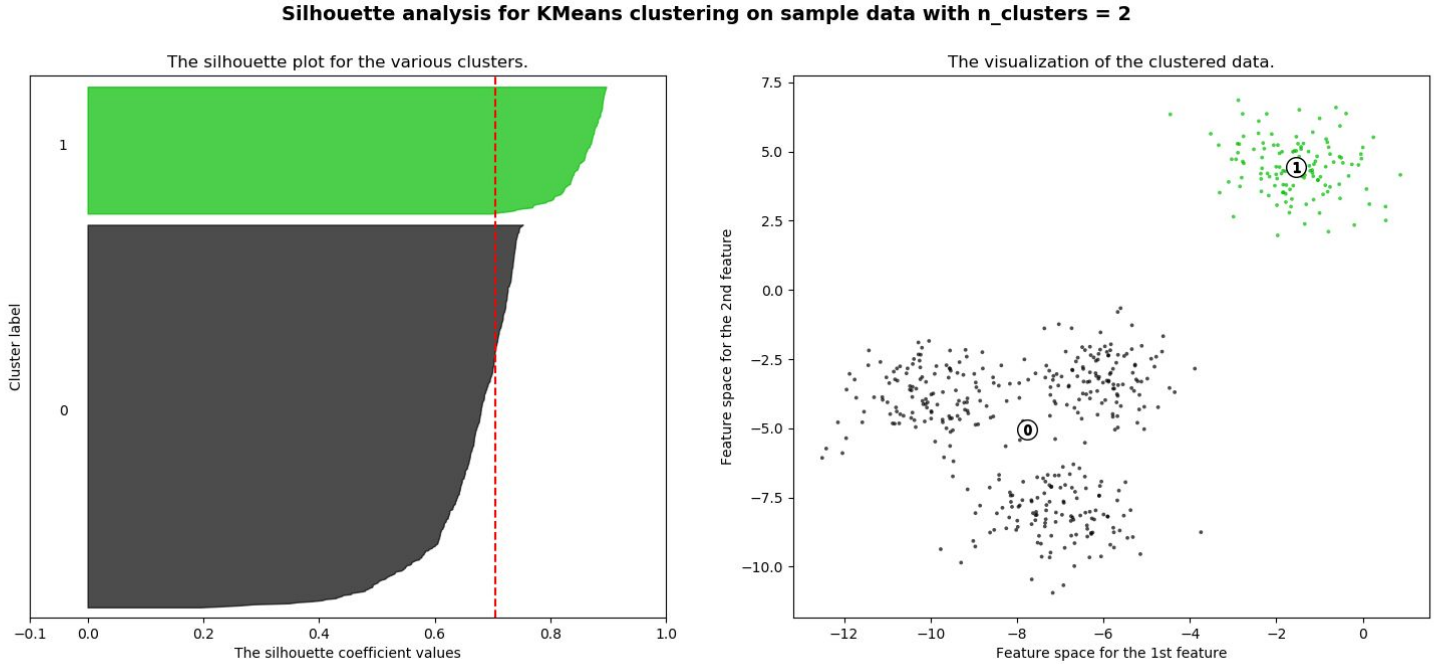
- Have to manually choose the number of clusters
- Slow in large datasets

http://www.cvl.isy.liu.se:82/education/graduate/spectral-clustering/SC_course_part1.pdf


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s 4.34s .07s 1.48s .23s .12s .01s .04s .01s
.02s 4.79s .05s 2.83s .22s .12s .01s .04s .01s
.02s 2.87s .10s .35s .84s .64s .01s .05s .02s
.02s 2.40s .08s .84s .42s .32s .01s .04s .03s
.02s 2.18s .05s .59s .21s .12s .02s .04s .01s
.02s 2.08s .08s .47s .12s .11s .01s .04s .02s

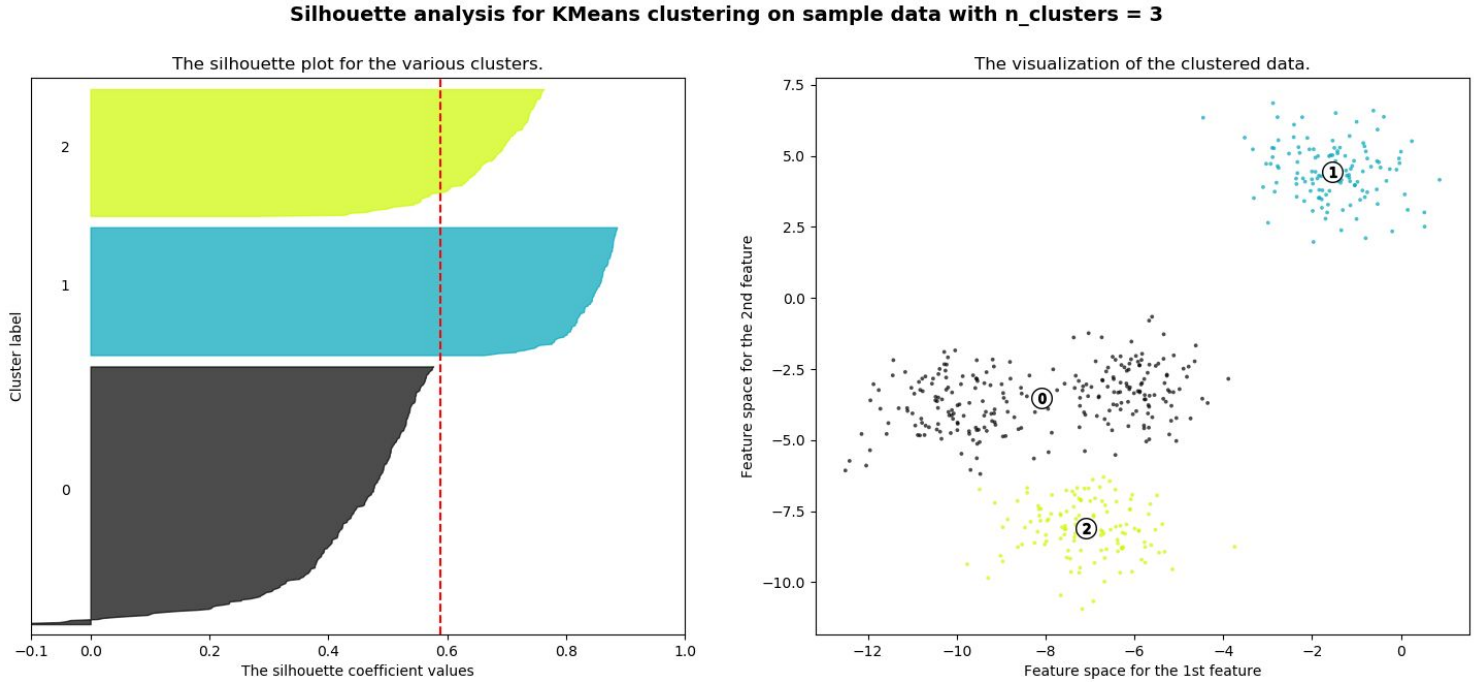
Silhouette plots

Silhouette plots



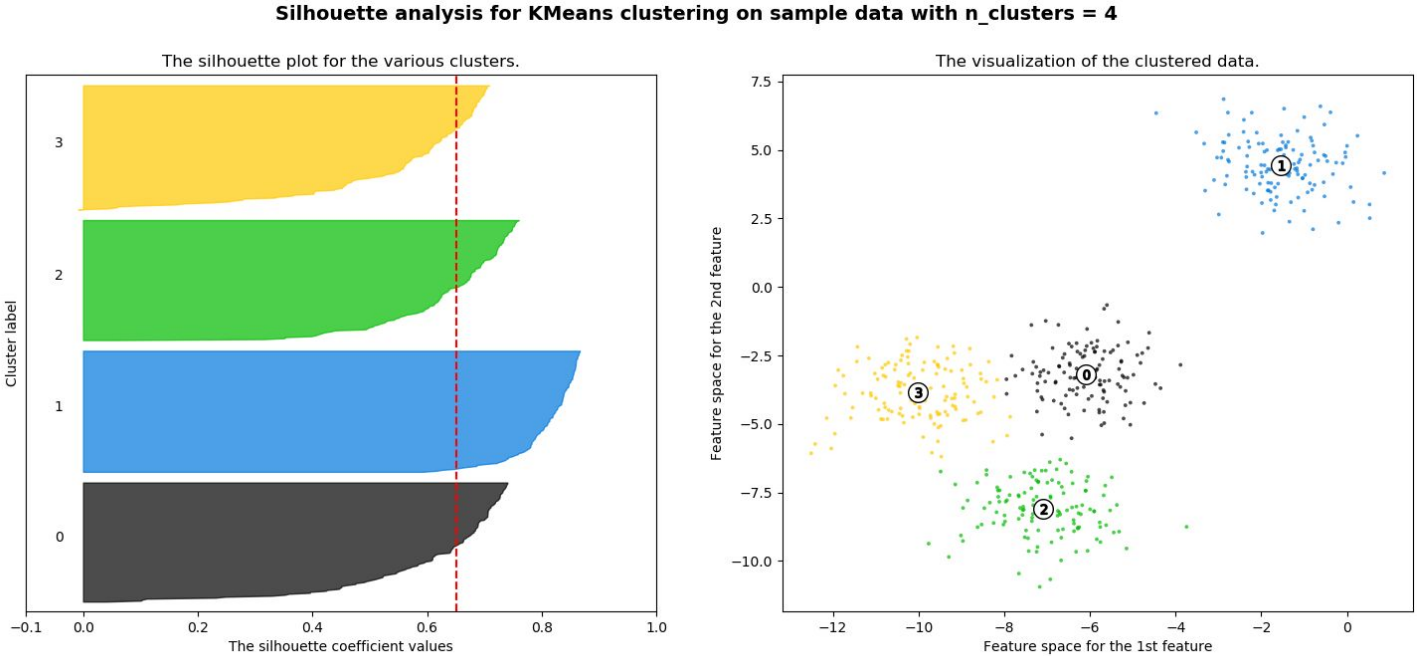
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots



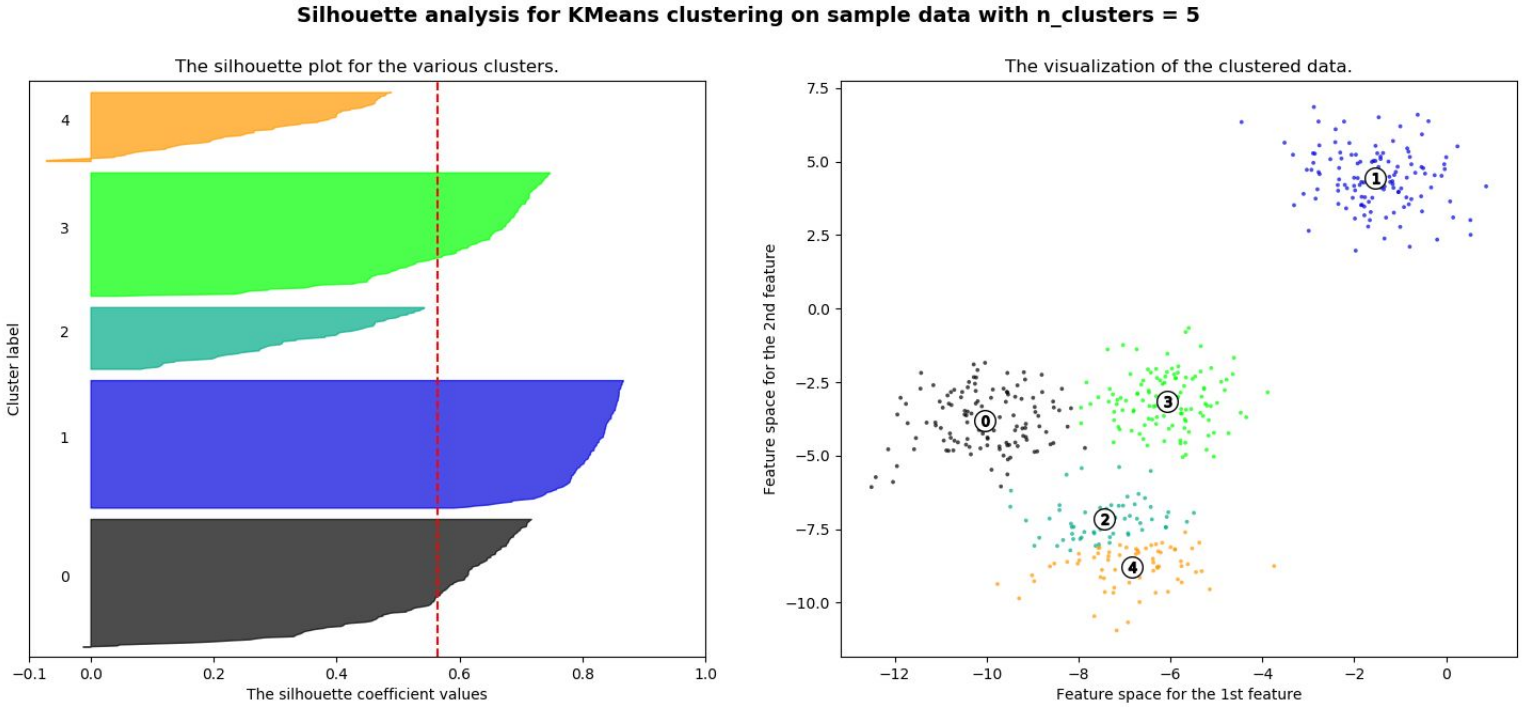
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots



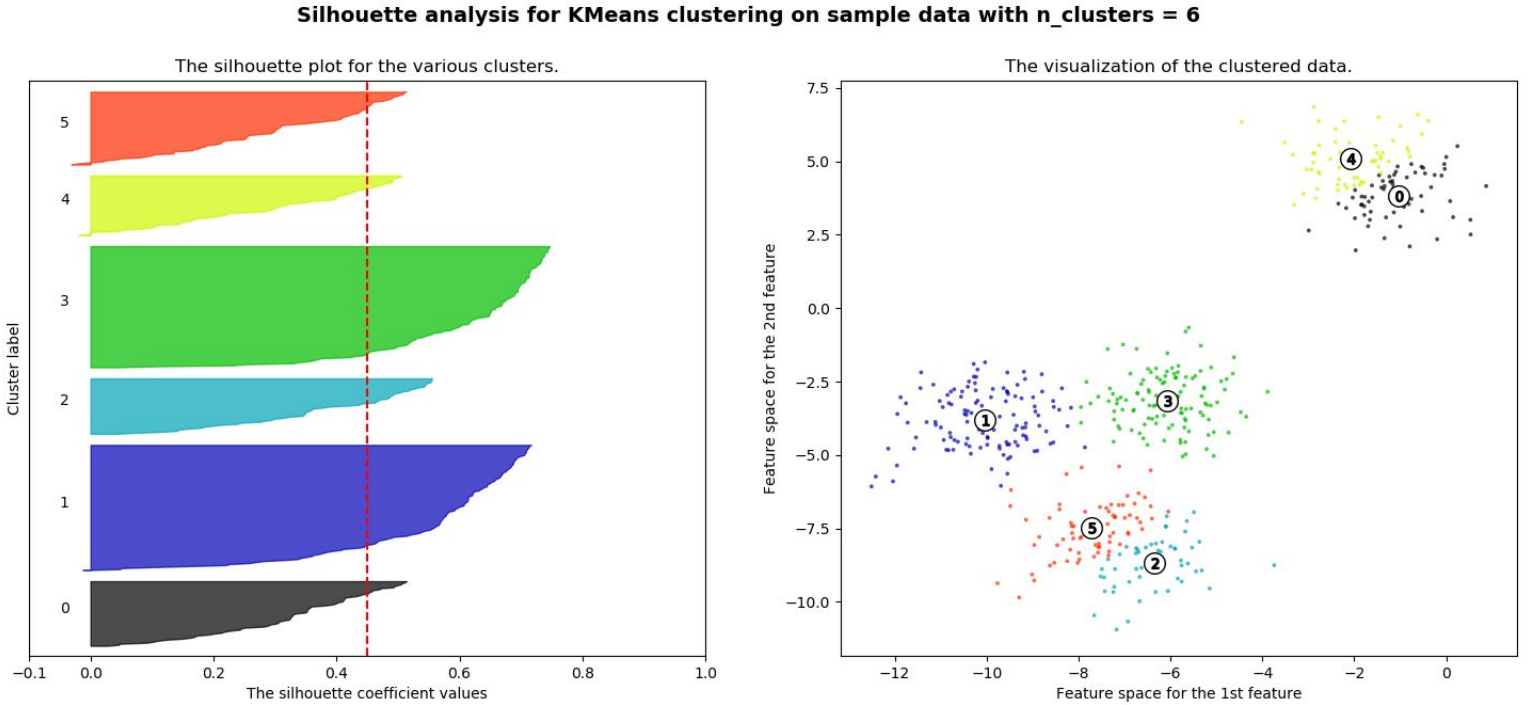
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots



http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots



http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots (pick k = 2)

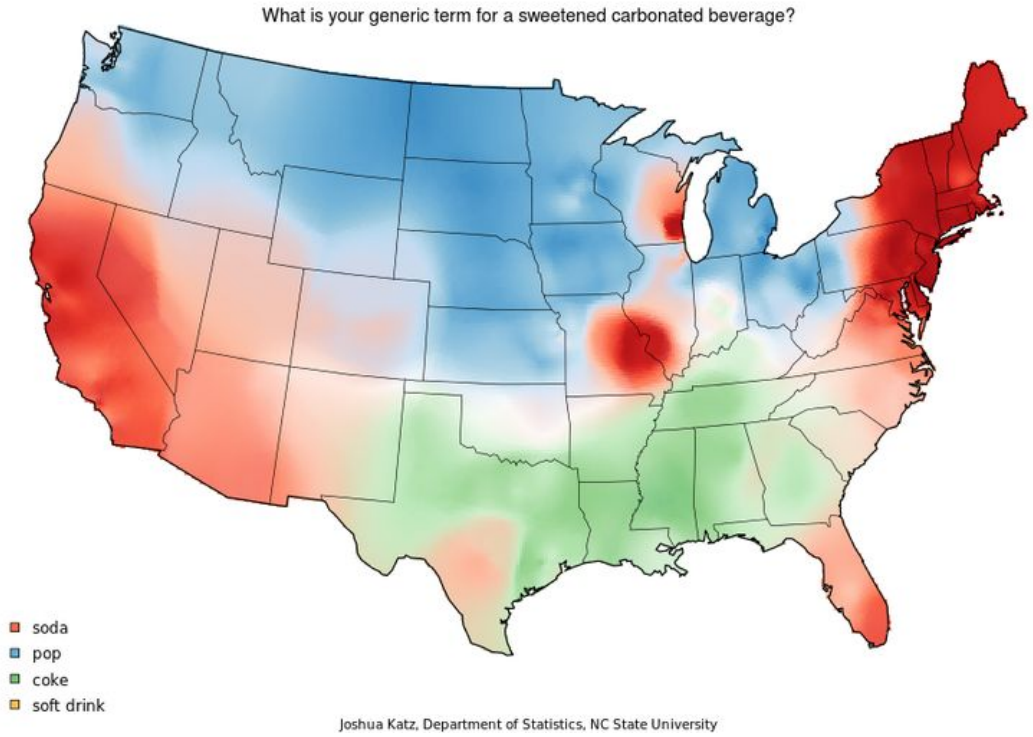
```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083261
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

Clustering exercises

1. Load wine.csv (14 characteristics of 178 wines from three different cultivars)
2. Plot the wines in the space defined by the first two principal components. Color each wine by cultivar
3. Run k-means with 3 cluster centers using all variables (except cultivar). Color each point in step 2 by cluster
4. Run k-means using the first two principal components only. Color each point in 2 by cluster. Compare the clustering results from (3) and (4)
5. Re-run steps 3 and 4 each four times. Do the results change?
6. Re-run steps 3-5 with 10 clusters. Compare silhouette plots.

Lab 2 Introduction

Lab 2



https://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6#ok-this-one-is-crazy-everyone-pronounces-pecan-pie-differently-10