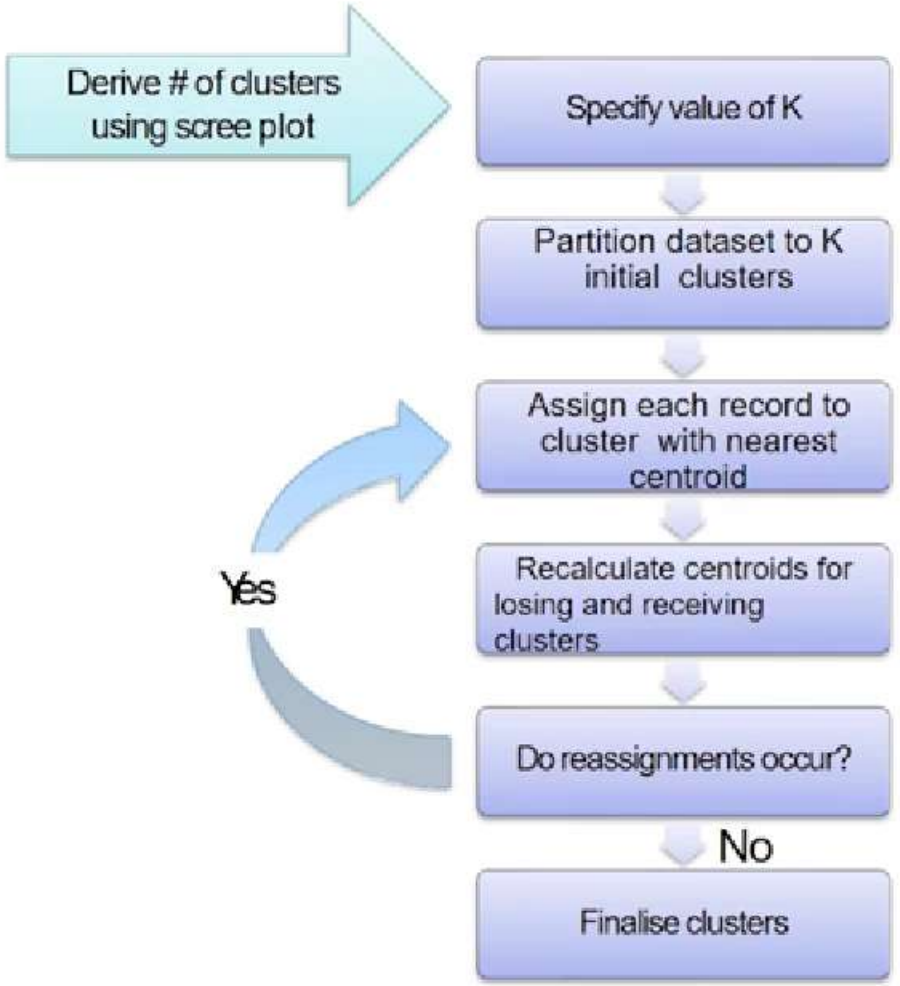
K-means Clustering

- A non-hierarchical approach to forming good clusters is to pre-specify a desired number of clusters, *k*
- *A*ssign each record to one of the *k* clusters, according to their distance from each cluster
- So as to minimize a measure of dispersion within the clusters
- *The 'means'* in the K-means refers to averaging of the data; that is, finding the centroid
- *K*-means clustering is widely used in large dataset applications

How does k-means clustering work?



Scaling – Z scaling & Min-max scaling

Z Scaling

- features will be rescaled
- have the properties of a standard normal distribution
- μ=0 and σ=1

$$z = \frac{x - \mu}{\sigma}$$

Min Max scaling

- the data is scaled to a fixed range - 0 to 1.
- The cost of having this bounded range - smaller standard deviations, which can suppress the effect of outliers

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where is scaling used?

k-nearest neighbors

- k-means

perceptrons, neural networks

- principal component analysis

Validating Clusters

- The resulting clusters should be valid to generate insights
- Cluster interpretability
- Cluster stability
- Cluster separation
- Number of clusters