1. Feature Selection

In the case of unsupervised learning, output is not defined, the goal is to submit the data set and let the algorithms find the clusters and correlations between the datasets. But in this case, the issue could be dataset could contain a lot of unrelated features, this could lead to algorithms providing results which are not of business value. This is where PCA would come into the picture which will help in reducing the dimensions of the data set and helps in identifying the optimal number of features/dimensions which cover maximum variance across various features.

1. Scaling units

Feature scaling is another important factor that needs to be factored in before running any unsupervised learning algorithms especially K-mean or PCA. Suppose the dataset contains 2 features which are measure in different units e.g. weight of a person, average salary, in this case, the weight would be Kilograms or pounds and salary in dollars. Obviously, weight comparison with average salary is not apple to apple comparison, in this case, features with higher weightage which is salary will start playing a critical role in distance calculation (specifically k-mean or PCA) than feature with lower magnitude. This is where feature scaling comes into the picture and puts both features on the same footing. There are 4 common scaling features

* 1. Standardization Normalization – This replaces the value with z-sore
  2. Mean Normalization – This replaces value between 1 and -1
  3. Min-Max Scaling – Value ranges from 0 to 1
  4. Unit Vector– Value ranges from 0 to 1

1. Identifying Clusters.

Identifying the right k-value is difficult to conclude especially when there is a lot of overlapping data across different clusters. In this case have to identify k-value with different methods like the elbow method, Average silhouette width method, Gap-statistic method. These different methods could give identical results or different results. So, in case if 2 methods give identical results then go with that value. In some cases, the mean of different methods can be considered to identify the correct k-value.

1. Data size

Unsupervised learning algorithms tend to be data-hungry compared to that of the supervised learning algorithms. Performance increases as the sample size increases, this holds good till a certain limit. The sample size is dictated by the amount of variance that you have across the features in the

1. Feature engineering (Encoding)

This is the most important step for both supervised and unsupervised learning. As part of this step, missing values are addressed and categorical variables are encoded using the appropriate technique. There would be categorical variables that will be one hot encoded or target encoded, depending on the choice the dimensions of the dataset may significantly grow depending upon the choice. If the dimensions grow significantly then it could impact the output generated by the unsupervised models. So the right choice has to be made for the encoding technique.

All about Feature Scaling

https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35