Q-1. Movie Recommendation systems are an example of: i) Classification ii) Clustering iii) Regression Ans- d) 2 and 3

Q-2. Sentiment Analysis is an example of: i) Regression ii) Classification iii) Clustering iv) Reinforcement

Ans- d) 1, 2 and 4

Q-3. Can decision trees be used for performing clustering?

Ans- a) True

Q-4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points: i) Capping and flooring of variables ii) Removal of outliers

Ans- a) 1 only

Q-5. What is the minimum no. of variables/ features required to perform clustering?

Ans- b) 1

Q-6. For two runs of K-Mean clustering is it expected to get same clustering results?

Ans- b) No

Q-7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

Ans- a) Yes

Q-8. Which of the following can act as possible termination conditions in K-Means? i) For a fixed number of iterations. ii) Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum. iii) Centroids do not change between successive iterations. iv) Terminate when RSS falls below a threshold.

Ans- d) All of the above

Q-9. Which of the following algorithms is most sensitive to outliers?

Ans- a) K-means clustering algorithm

Q-10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning): i) Creating different models for different cluster groups. ii) Creating an input feature for cluster ids as an ordinal variable. iii) Creating an input feature for cluster centroids as a continuous variable. iv) Creating an input feature for cluster size as a continuous variable.

Ans- d) All of the above

Q-11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

Ans- d) All of the above

Q-12. Is K sensitive to outliers?

Ans- The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. The group of points in the right form a cluster, while the rightmost point is an outlier. K-means can be quite sensitive to outliers. So if you think you need to remove them, I would rather remove them first, or use an algorithm that is more robust to noise. For example, k medians is more robust and very similar to k-means, or you use DBSCAN.

DBSCAN Full From is – Density Based Scan

Density-Based Spatial Clustering of Applications with Noise

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, a opposed to k-means which assumes that clusters are convex shaped. The central component to the DBSCAN is the concept of core samples, which are samples that are in areas of high density

DBSCAN uses 2 parameters to find clusters in the data sets - cluster radius(epsilon)and minimum number of points needed to form a cluster(n). So, we scan each point in the data and check whether it has n or more number of points around it within the radius epsilon. And if a point meets the criteria, it becomes a core point and the points surrounding it become border points and a cluster is created and so on.

Q-13. Why is K means better?

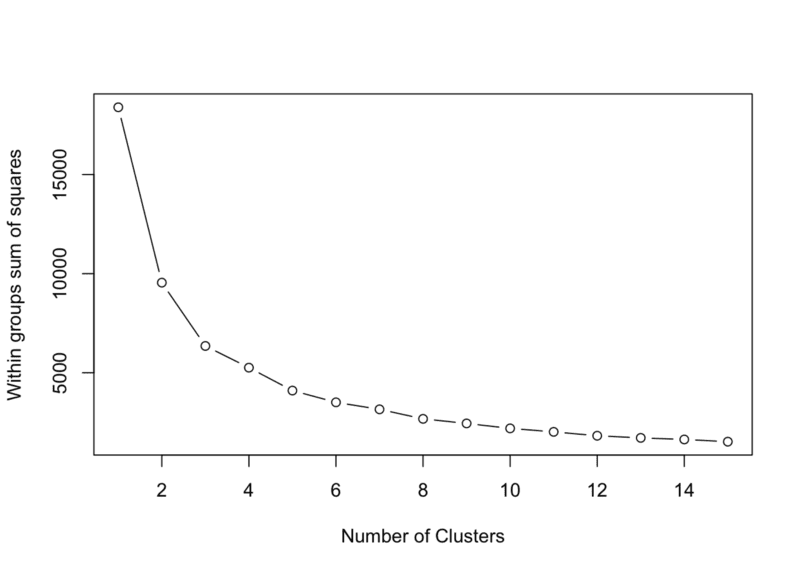
Ans- Unsupervised learning has emerged as the most effective technique for discovering patterns in data. It is also being used to develop labels on top of the supervised models. This is one of the most widely used techniques for market or customer segmentation wherein the company’s data can be segregated into clusters and used to identify certain patterns which leads to a more customised approach. This technique comprises machine learning algorithms through which data analysts can draw inferences from datasets without labelled responses. Cluster analysis is also widely used for exploratory data analysis to find hidden patterns or grouping in data.

Unsupervised Learning Algorithms Can Be Divided Into Two Wide Categories:

Clustering: A clustering problem is where one can find the inherent groupings in the data, such as grouping customers by purchasing behaviour.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y. Some of the common clustering algorithms are hierarchical clustering, Gaussian mixture models and K-means clustering. The last one is considered one of the simplest unsupervised learning algorithms, wherein data is split into k distinct clusters based on distance to the centroid of a cluster.

Why K-Means Clustering Is So Popular

[](https://149695847.v2.pressablecdn.com/wp-content/uploads/2018/10/1_t2CuFhihU_vKNE6u1wI6Og.png)

K-Means for Clustering is one of the popular algorithms for this approach. Where K means the number of clustering and means implies the statistics mean a problem. It is used to calculate code-vectors (the centroids of different clusters). According to a[tutorial](https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html), for any word/value/key that needs to be ‘vector quantized’, it is by calculating the distance from all the code vectors and assign the index of the code vector with the minimum distance to this value. For example, clustering can be applied to MP3 files, cellular phones are the general areas that use this technique.

Is K-Means Really Used In Production?

K-means has been around since the 1970s and fares better than other clustering algorithms like density-based, expectation-maximisation. It is one of the most robust methods, especially for image segmentation and image annotation projects. According to some users, K-means is very simple and easy to implement. However, it is unlikely to be the state-of-the-art, but for straightforward clustering, it is also a part of a larger data-processing pipeline, K-means is a reasonable default choice, at least until you figure out that the clustering step is your bottleneck in terms of overall performance.

Strengths Of K-Means Clustering Algorithm

According to[this paper](https://www-cs.stanford.edu/~acoates/papers/coatesng_nntot2012.pdf), (Learning Feature Representations With K-means) K-means is used to learn feature representations for images (use k-means to cluster small patches of pixels from natural images, then represent images in the basis of cluster centres; repeat this several times to form a “deep” network of feature representations) gives image classification results that are competitive with much more complex / intimidating deep neural network models. In fact, a lot of k-means applications are now done using support vector machines.

* It gives good results
* It is already implemented in the software
* Number of clusters has to be fixed before
* Dependent of the initialisation parameters and the chosen distance

Weakness

* The results given are usually dependent on the initial values for the means.
* And the way to initialise the means is not specified, one can start by randomly choosing K of the samples.

How K-Means Algorithm Functions:

The algorithm clusters into k groups and here k is the input parameter. In this procedure, a dataset is classified through a certain number of clusters, commonly known as k clusters and the main idea is to define k centres, one for each cluster. These centers should be placed in a way since different location causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. However, the main disadvantage is one has to specify the number of clusters as an input in the algorithm.

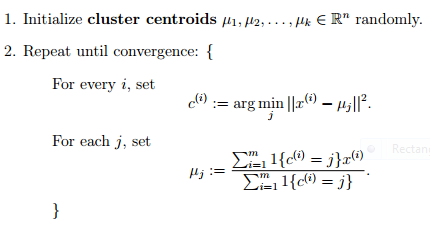
Conclusion

Many argue that in the field of data science, one should primarily use simple, self-learning algorithms. And clustering algorithm, the most commonly used unsupervised learning algorithm is self-improving and one doesn’t need to set parameters. In fact, most data science teams rely on simple algorithms like regression and completely because they solved all normal business problems with simple algorithms like XG Boost. Another key upside of K-means, the standard data mining tool is that as opposed to conventional statistical methods, the clustering algorithms do not depend on statistical distributions of data and can be used with little prior knowledge exists, a paper on data mining using [K-means discussed](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0087-2).

Q-14. Is K means a deterministic algorithm?

Ans- The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.

K-means, that you used as example, starts with randomly chosen cluster centroids so to find optimal ones. Besides the initialization, the algorithm is totally deterministic, as you can make sure [looking at it's pseudocode](http://stanford.edu/~cpiech/cs221/handouts/kmeans.html):

[](https://i.stack.imgur.com/rElZ5.png)

Nothing prohibits you from starting with non-random centroids. We use random centroids so to make sure that badly chosen starting points would not lead us to poor results. The same with other "random" algorithms: you can use them in "deterministic" fashion, but in most cases, this is not a wise thing to do. In case of k-means the algorithm deterministically minimizes the within-cluster sum of squares to find the optimal clustering solution. Unfortunately, it is sensitive to how the algorithm was initialized. Clustering problems in most cases do not have clear-cut solutions, because of that we often want to use randomized procedures to robustify them. Imagine that you used some deterministic [hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering) algorithm. Imagine that it goes through your data sequentially, starting from the first observation. What would happen if the first case was an outlier? On the other hand, if you initialized it several times at random points, the procedure would be less prone to problems with data. Moreover, if you run non-"deterministic" algorithm multiple times and then use majority vote to choose for each case the class that appeared most commonly among the results, then the final output will also by highly deterministic.