Q 1. Movie Recommendation systems are an example of:

i) Classification
ii) Clustering
iii) Regression

Ans - b) 1 and 2

Q 2. Sentiment Analysis is an example of:

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) ii) Removal of outliers Options:

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 witha 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 - *k*-Means is a well studied clustering problem that finds applications in many fields related to unsupervised learning. *k*-means clustering is "highly sensitive" to the isolated points. These isolated points called outliers.

For example -
is the average of the salaries of the following people:

50k, 20k, 35k, 65k and 1MM

The average ends up being (50k + 20k + 35k + 65k + 1MM) / 5 = 1170k / 5 = 234k.

If we did not have the $1MM outlier, the average would have been (50k + 20k + 35k + 65k) / 4 = 170k / 4 = 42.5k.

Note that the two average results are wildly different from one another.

So we can say k-mean algorithm is not making sense to use.

Also we can say k-means clustering is Highly sensitive to outliers.

Q 13. Why is K means better?

Ans –

1) Relatively simple to implement.

2) Scales to large data sets.

3) Guarantees convergence.

4) Can warm-start the positions of centroids.

5) Easily adapts to new examples.

6) Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Q 14. Is K means a deterministic algorithm?

Ans – No ,Nature of K-Means is due to its random selection of data points as initial centroids making "non- deterministic algorithm ".

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.