Q1. ANS - b)4

Q2.ANS- d) 1,2 and 4

Q3.ANS – d) formulating the clustering problem

Q.4ANS – a) Euclidean distance

Q.5ANS- b) Divisive clustering

Q.6ANS- d) All answers are correct

Q.7ANS- a) Divide the data points into groups

Q.8ANS- b) Unsupervised learning

Q.9ANS- d) All of the above

Q.10ANS- a) K-means clustering algorithm

Q.11ANS- d) All of the above

Q.12ANS- a) Labeled data

**Q13. How is cluster analysis calculated?**

ANS – STEPS

**Step 1 – Hypothesis building**

This is the most crucial step of the target segmentation(customer segmentation).Try to identify all possible variables that can help segment the portfolio regardless of its availability. Let’s take an example of X bank that wants to understand the profile of its customer base to build targeted campaigns.

a. Customer balance with bank X

b. Number of transaction done in last 1/3/6/12 months

c. Balance change in last 1/3/6/12 months

d. Demographics of the customer

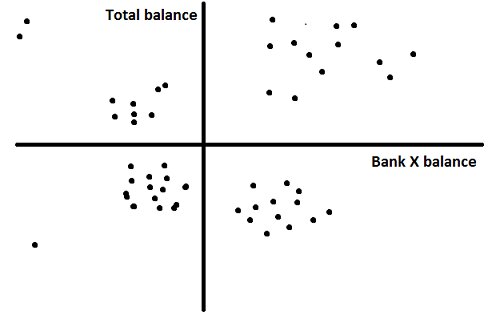
e. Customer total balance with all Indian banks

**Step 2 – Initial shortlist of variable**

Once we have all possible variable, start selecting variable as per the data availability. Let’s say, for the current example we have only data for Customer balance with bank X and Customer total balance with all Indian banks (total balance)

**Step 3 – Visualize the data**

It is very important to know the population spread across the selected variable before starting any analysis. For the current scenario, the exercise becomes simpler as the number of selected variables is only 2. Following is a scatter plot between total balance and Bank X balance (origin taken as mean of both the variables):



This visualization helps us to identify clusters which we can expect after the final analysis. Here, we can see there are four clear clusters in four quadrants. We can expect the same result in the final solution.

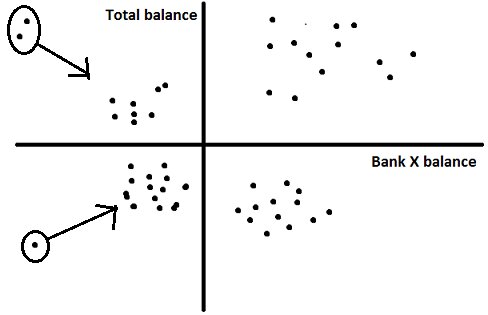
**Step 4 – Data cleaning**

Cluster analysis is very sensitive to outliers. It is very important to clean data on all variables taken into consideration. There are two industry standard ways to do this exercise:

1. Remove the outliers: (Not recommended in case the total data-points are low in number) We remove the data-points beyond mean +/- 3\*standard deviation.

2. Capping and flouring of variables: (Recommended approach) We cap and flour all data-points at 1 and 99 percentile.

Let’s use the second approach for this case.

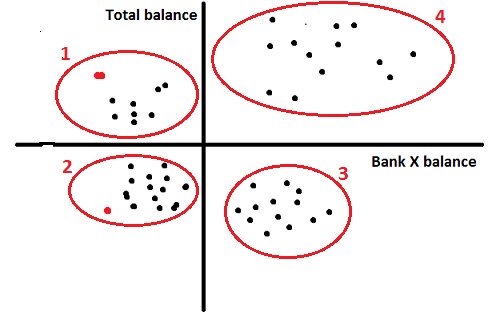


**Step 5 – Variable clustering**

This step is performed to cluster variables capturing similar attributes in data. And choosing only one variable from each variable cluster will not drop the sepration drastically compared to considering all variables. Remember, the idea is to take minimum number of variables to justify the seperation to make the analysis easier and less time consuming.

**Step 6 – Clustering**

We can use any of the technique i.e Hierarchical clustering, k-mean clustering etc depending on the number of observation. k-means is used for a bigger samples. Run a proc fastclus with k=4 (which is apparent from the visualization).

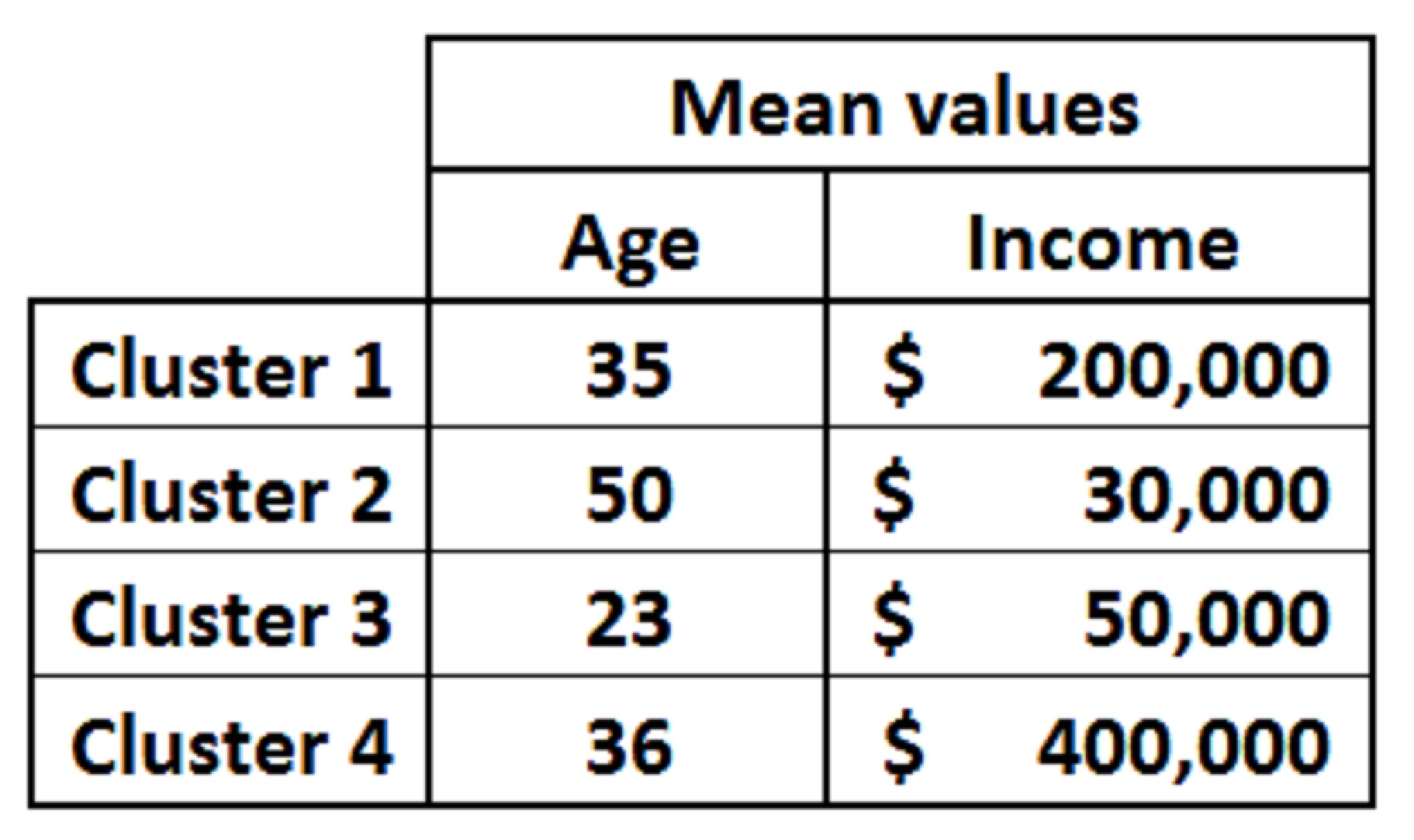


**Step 7 – Convergence of clusters**

A good cluster analysis has all clusters with population between 5-30% of the overall base. Say, our total number of customer for bank X is 10000. The minimum and maximum size of any cluster should be 500 and 3000. If any of the cluster is beyond the limit than repeat the procedure with additional number of variables..

**Step 8 – Profiling of the clusters**

After validating the convergence of cluster analysis, we need to identify behavior of each cluster. Lets say we map age and income to each of the four clusters and get following results.



Now is the time to build story around each cluster. Lets take any two cluster and analyze.

Cluster 1 : (High Potential Low balance customer) These customers do have high balance in aggregate but low balance with bank X. Hence, they are high potential customer with low current balance. Also the average salary is on a higher side which validates our hypothesis of customer being high potential.

Cluster 3 : (High Potential high balance customers) Even though the salary and total balance in aggregate is on a lower side, we see a lower average age. This indicates that the customer has a high potential to increase their balance with bank X.

**14. How is cluster quality measured?**

ANS -Assessing the quality of your model is one of the most important considerations when deploying any [**machine learning**](https://opendatascience.com/?s=machine+learning) algorithm.

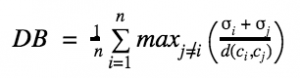
For supervised learning problems, this is easy. There are already labels for every example, so the practitioner can test the model’s performance on a reserved evaluation set.

We don’t have that luxury when we’re dealing with unlabeled data in unsupervised learning contexts. There’s nothing to even test since there’s no ground truth — the idea of testing in this arena is a flawed premise.

That doesn’t mean assessing the model is a lost cause. Numerous metrics examine the quality of clustering results when labeled data is unavailable. These metrics can give the practitioner insight into how the clusters might change depending on the algorithm’s selection and the natural tendency of the data to group together.

**Davies-Bouldin Index –**

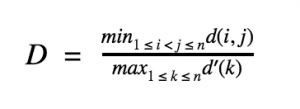
The DB Index is calculated by the following formula:



where n is the number of clusters and σi is the average distance of all points in cluster i from the cluster centroid ci.The DB index captures the intuition that clusters that are (1) well-spaced from each other and (2) themselves very dense are likely a ‘good’ clustering. This is because the measure’s ‘max’ statement repeatedly selects the values where the average point is farthest away from its centroid, and where the centroids are closest together. As the DB index shrinks, the clustering is considered ‘better’.

## Dunn Index –

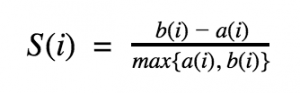
The formula for the Dunn Index is as follows:



where i, j and k are each indices for clusters, d measures the inter-cluster distance and d’ measures the intra-cluster difference.The Dunn Index captures the same idea as the DB Index: it gets better when clusters are well-spaced and dense. But the Dunn Index increases as performance improves.What differs is the way this problem is approached. While the DB index considers the dispersion and separation of all clusters, the Dunn Index only considers the worst cases in the clustering: the clusters that are closest together and the single most dispersed cluster. Depending on our application, the change in objective may introduce unexpected problems.

## Silhouette Coefficient –

The Silhouette Coefficient is measured like so:



where a(i) is the average distance of point i from all other points in its cluster and b(i) is the smallest average distance of i to all points in any other cluster. To clarify, b(i) is found by measuring the average distance of i from every point in cluster A, the average distance of i from every point in cluster B, and taking the smallest resulting value.

The Silhouette Coefficient tells us how well-assigned each individual point is. If S(i) is close to 0, it is right at the inflection point between two clusters. If it is closer to -1, then we would have been better off assigning it to the other cluster. If S(i) is close to 1, then the point is well-assigned and can be interpreted as belonging to an ‘appropriate’ cluster.

The Silhouette Coefficient is a very intuitive and sophisticated measure of distance. Its downfall is that it can be extremely expensive to compute on all n points. This is because we must compute the distance of i from all other n — 1 points for each i, which leads to a complexity of O(n2).

Many practitioners will balk at that cautious assessment and shrug, saying it’s less than NP. However, for very large datasets these time complexities can become unmanageable.

These are just three of the most popular methods for assessing cluster quality. There are a variety of other techniques, but They’ll certainly give us more insight about our model’s accuracy than blind trust.

**15. What is cluster analysis and its types?**

ANS - Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets.

Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal.

In other words, the clusters are regions where the density of similar data points is high. It is generally used for the analysis of the data set, to find insightful data among huge data sets and draw inferences from it. Generally, the clusters are seen in a spherical shape, but it is not necessary as the clusters can be of any shape.

It depends on the type of algorithm we use which decides how the clusters will be created. The inferences that need to be drawn from the data sets also depend upon the user as there is no criterion for good clustering.

Types of Clustering –

**Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, each customer is put into one group out of the 10 groups.

**Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, each costumer is assigned a probability to be in either of 10 clusters of the retail store.

Types of Clustering Algorithms –

**Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

**Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

**Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

**Density Models:**These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.