**CUSTOMER MARKET SEGMENTATION**

1. **INTRODUCTION**

Market segmentation is a great way to identify marketing groups from a larger population. This is done by identifying target customers belonging to a specific type based on features/categories like age, income, gender, etc. so that ads or targeting marketing can be carried out to increase the customer base of a company or boost their sales i.e., grow their business.

The target here is to divide the mall target market into approachable groups. This is done by creating subsets of a market based on demographics/behavioral data of the customers.

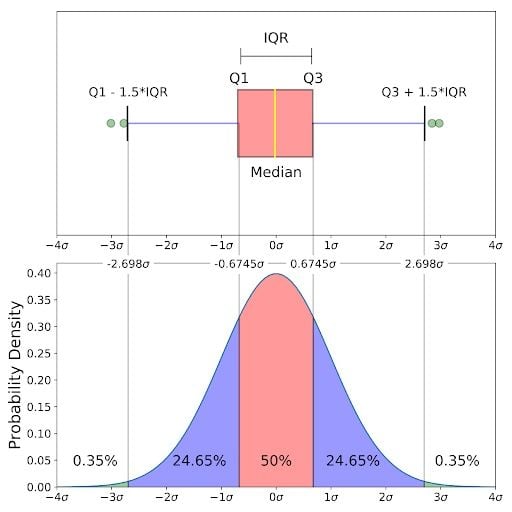
1. **DATASET**

The dataset being used is Mall

1. **DATA DESCRIPTION**
2. **PRE-PROCESSING**
3. **DATA EXPLORATION**

Boxplot:

* A boxplot is a graph that gives you a good indication of how the values in the data are spread out.
* It also tells you about the outliers and their values.
* Gives information about the variability or dispersion of the data.



If the notches in the boxplot (notch plot) do not overlap we can say that with 95 % confidence, the true medians differ between the categories.

1. **MODEL TRAINING**

For customer segmentation, we employ the K-Means algorithm. What K-Means does is simply divide a set of N sample (our dataset) into K disjoint clusters C, where each cluster is defined by the mean of the samples present in the cluster (their centroid). This can also be interpreted such that K-means tries to separate samples in n groups of equal variance, minimizing the inertia or within-cluster sum-of-squares. This algorithm scales well to larger samples as well.

Even the simplicity and advantages of K-means are pretty evident, a few disadvantages of the same can be said:

* The “Inertia” which we aim to minimize, assumes that clusters are convex and isotropic, which may not always be the case in real-life.
* It responds poorly to elongated clusters, or manifolds with irregular shapes.
* The other factor is that inertia is not a normalized metric – as Euclidean distances become inflated in higher dimensions; Hence, dimension reduction is a good approach to distill this problem and speed up computations.

1. **RESULTS**

The clustering or segmentation was evaluated based on the following:

* *Inertia*:
* *Silhouette coefficient*: It is defined for each sample and has 2 scores; “a” being the score between a sample and all other points in the same cluster. “b” being the mean distance between a sample and all other points in the next nearest cluster.

Silhouette Coefficient

So, the silhouette coefficient for a set of samples is given as the mean of the Silhouette coefficient for each sample. It ranges from -1 (incorrect clustering) to +1 (highly dense clustering and well separated). A zero silhouette coefficient can indicate overlapping clusters.

* *Dunn’s Index*: This is the minimum inter-cluster distance divided by the maximum cluster size. Simply put, a higher DI means better clustering. But, this also assumes that better clustering means that clusters are compact and well-separated from other clusters.

We use the elbow plot to determine the best

1. **CONCLUSION**
2. **REFERENCES**