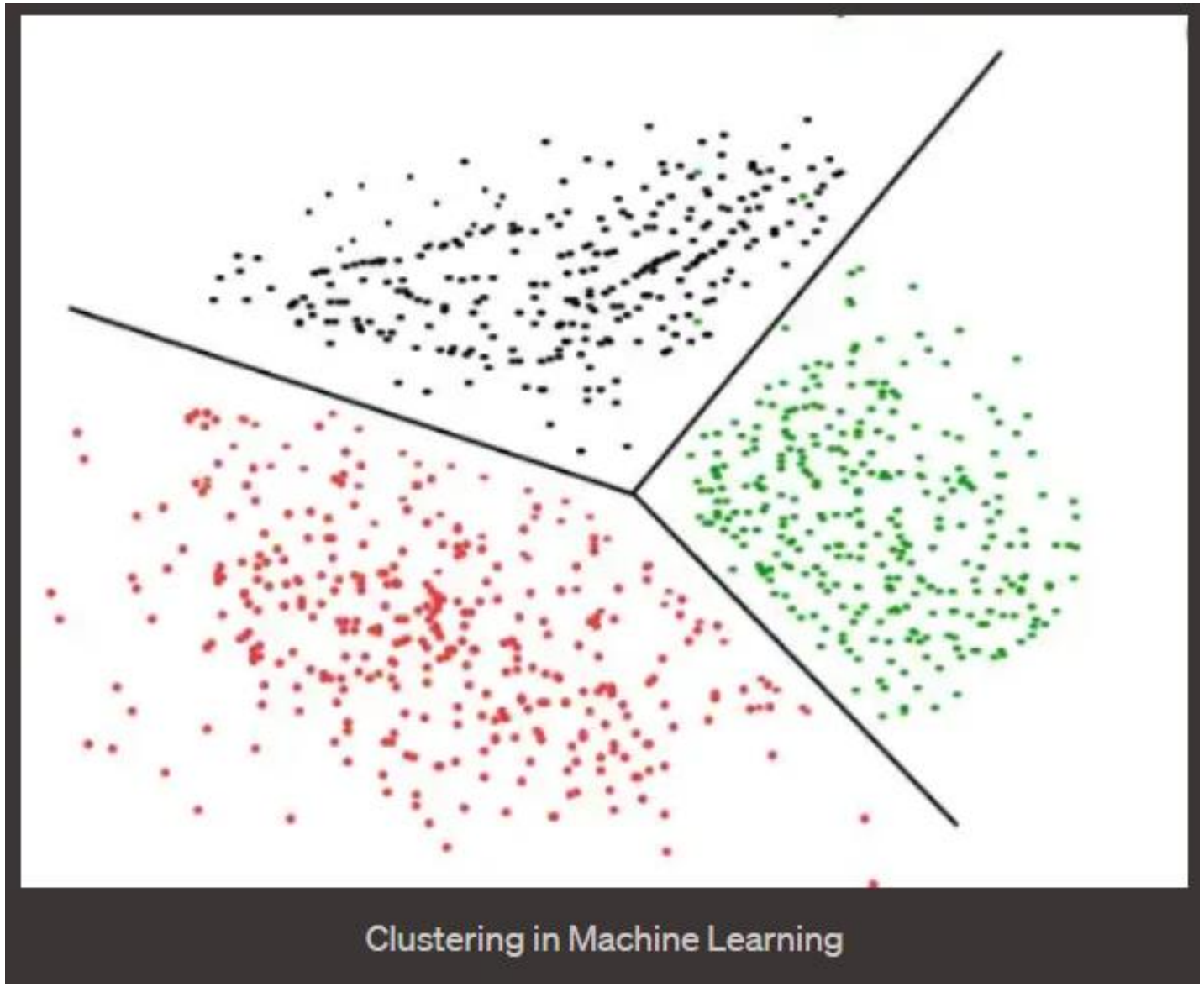
K Mode Clustering

UNSUPERVISED MACHINE LEARNING ALGORITHM

Why K mode clustering

- In the real world, the data might be having different data types, such as numerical and categorical data. To perform a certain analysis, for instance, clustering analysis, we should consider the data type in the data we have. The clustering algorithm commonly used in clustering techniques and efficiently used for large data is k-Means. But, it only works for the numerical data. It's actually not suitable for the data that contains the categorical data type. So, Huang proposed an algorithm called k-Modes which is created in order to handle clustering algorithms with the categorical data type.



Clustering in Machine Learning

Problem Statement

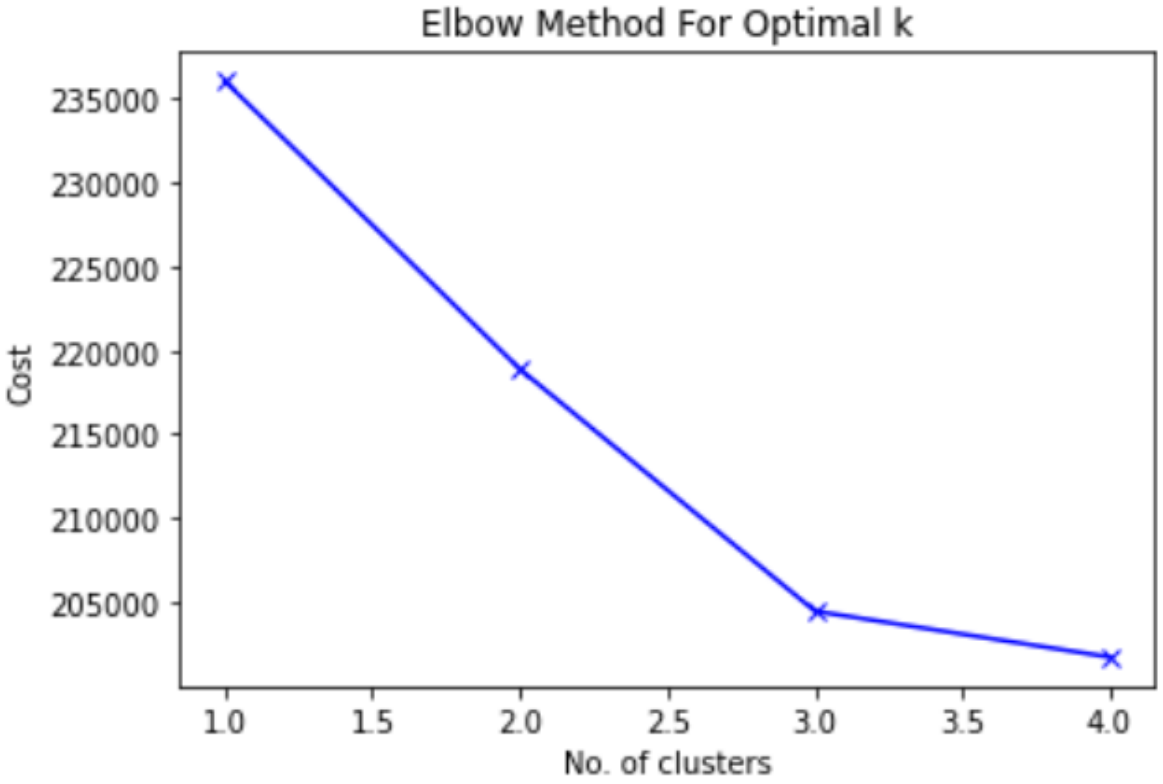
- K Mode machine learning clustering algorithm on Hotel mapping data to remove duplicate hotel inventory across multiple suppliers and map content with the highest level of accuracy and zero duplicates

Overall Approach

- 1. Understanding of problem statement and Business purpose
- 2. Understanding of Data
- 3. DATA CLEANING –
- 3.1Fixing rows and columns
- 3.2 Finding out missing or null values and treating them
- 4. Prepare data for model building-
- 4.1 Check the DATATYPE of column and Data Binning /astype()
- 4.2 converting the labels into a numeric form so as to convert them into the machine-readable form.(Label Encoding)

Building Machine Learning Model

- 5. Finding optimal k for clustering(Elbow curve/Cao/Huang initialisation methos
- 6. Building model with no k clusters
- 7. MERGING of Datasets
- 8. CONCLUSION



Elbow Curve to initializing k-The curve is bended at 3 points so build the model using k=3(clusters)

FINAL MODEL SUMMARY: Merge actual dataset and predicted clusters.

In [48]:
```
VTech7.insert(0, "Cluster", clusters, True)
VTech7.head()
```

Out[48]:

	Cluster	VervotechId	HotelName	Provider	ProviderHotelId	ProviderRoomCode	ProviderRoomName	ProviderBedInfo	RoomInSquareFeet
0	0	1285	73	0	581	2096	630	262	23
1	2	1285	73	1	228	5921	630	261	24
2	1	1285	73	2	2747	24671	630	217	24
3	2	1285	73	1	228	8410	10471	41	321
4	0	1285	73	0	581	2102	631	68	23

Conclusion

- Dataset has divided into k=3 clusters(Cluster0,Cluster1,Cluster2)
- Inference from the model predictions:

  row 0,4 are merged as a cluster 0;

  row 2,6 are merged as a cluster 1;

  row 1,3 are merged as a cluster 2.