**AMA Assignment**

**Applying PCA & Clustering on Airbnb Dataset**

Airbnb Distribution Project when Airbnb decides to expand to a new city, it is crucial to gain insights into the Airbnb Distribution Pattern to perfect budgeting and advertising strategies in that city. Therefore, our team aims to analyse relevant factors that impact Airbnb’s spatial distribution. We picked data of year 2019 of New York city for our investigation .

**Introduction to the dataset:**

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generatedSnapshot of data:**

**Data preparation:**

There are no duplicates in the data.

There are missing values in the dataset for certain variables.

A screenshot of a computer

Description automatically generated

Handling Missing values:

* Host\_name and name are not that much of null values, so we are good to fill those with some substitutes in both the columns first.
* We fill name with value ‘unknown’ and fill host\_name with value ‘no\_name’.
* last\_review and reviews\_per\_month have total 10052 null values each.
* last\_review column is not required for our analysis as compared to number\_of\_reviews & reviews\_per\_month.
* So, removing last\_review column beacause of not that much important.
* We fill null values of reviews\_per\_month with 0 as there can be zero reviews for that listing.
* As the price column consist of 0's(11 obs) so we remove them as they do not add any value in our analysis.

**Handling Outliers:**

We use IQR technique for dealing with the outliers in data for price variable as the price data is an important variable for our analysis.

Dealing outliers for price variable:

IQR is 106.0

lower\_range is -90.0

upper\_range is 334.0

Outliers are 6.08% of total data

So we remove outliers for the price variable.

We will use scaling for all numeric variables before applying clustering on the data which will reduce the impact of outliers.

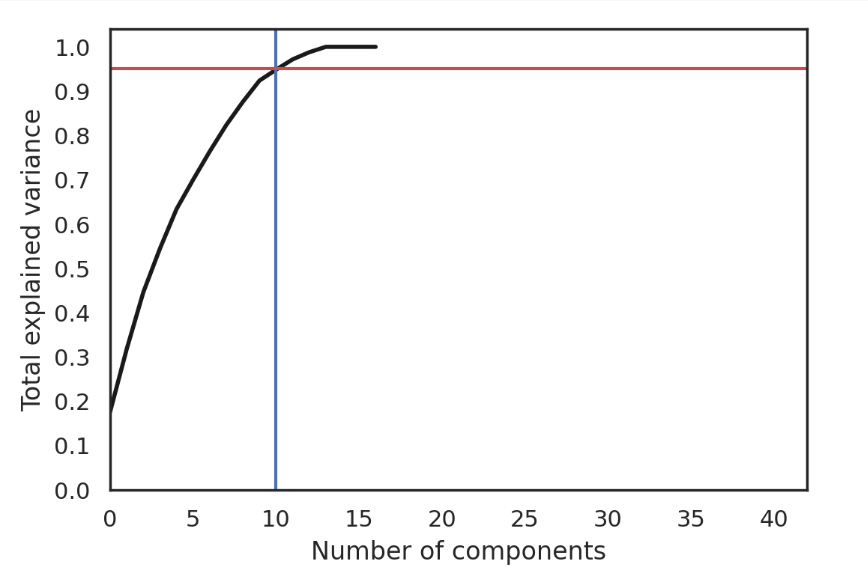
Also apply One-Hot Encoding to the categorical variables to convert them to numerical data.

**Unsupervised Machine Learning:**

**Apply PCA & K-means Clustering:**

It is a dimension reduction method that is preferably used for an Unsupervised Learning Problem.

It is utterly important to carefully standardize your data because PCA works under the assumption that the data is normally distributed and is very sensitive to variance of variables.

****

* 4 components can represent about 50% information of X
* 10 components can represent about 95% information of X

**K-Means:**

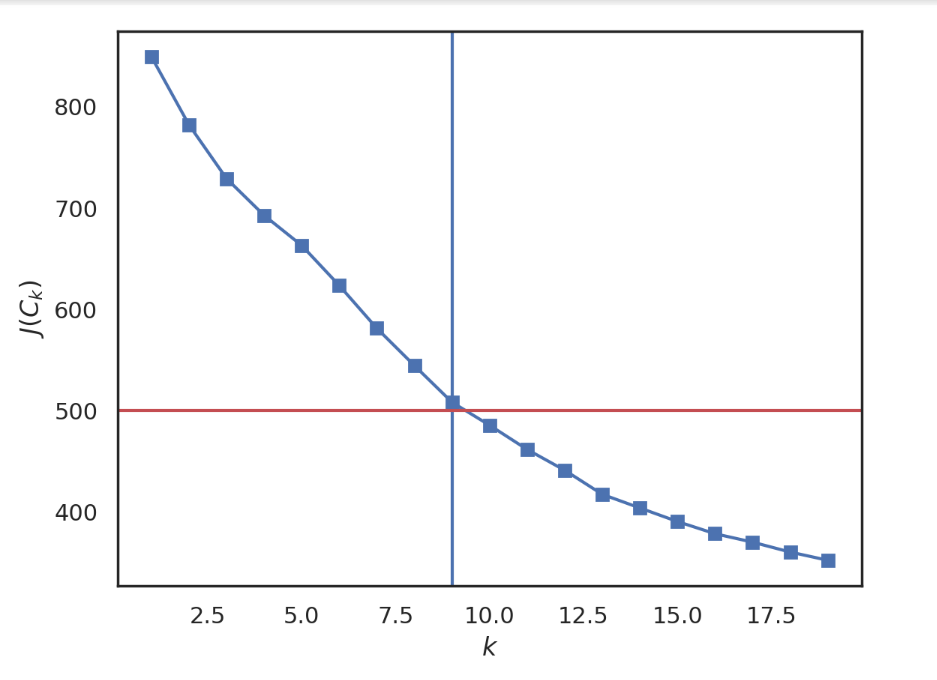
**K-Means Clustering** is a distance based algorithm that is used for unsupervised learning that **does not support categorical features**.

Method:

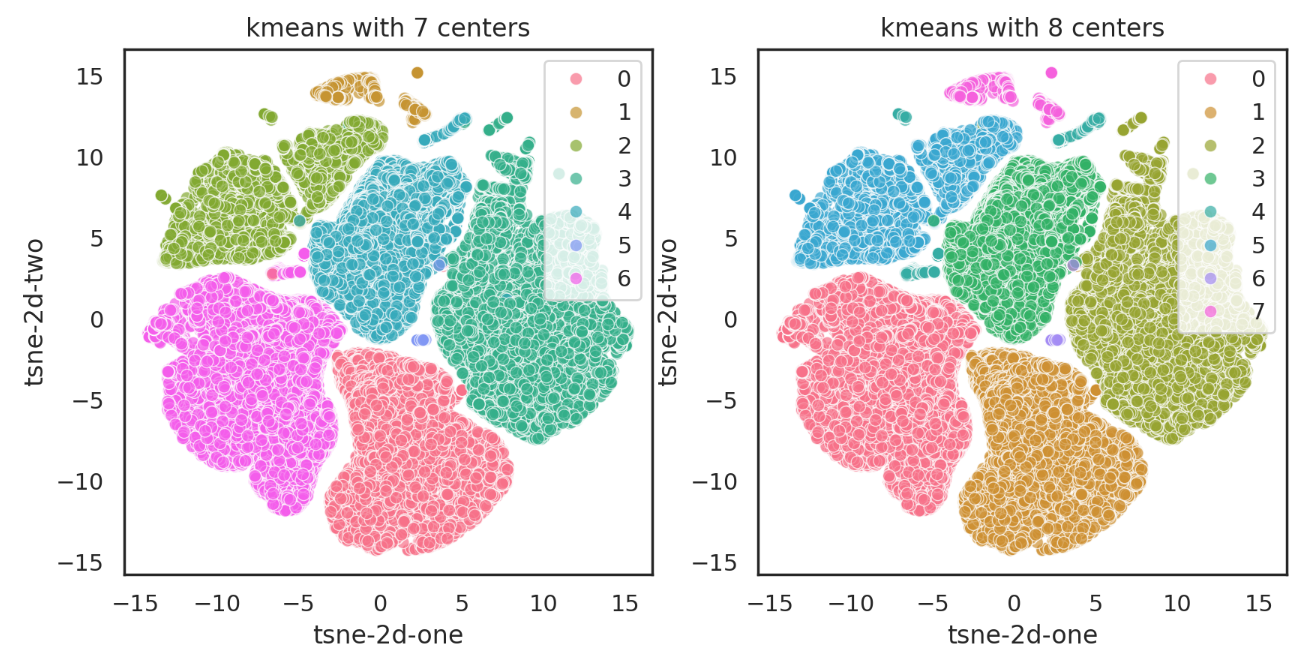
* Selects 2 (assume k = 2) centroids randomly.
* Distance between the centroids and all the data points are calculated.
* If the data point is closer to one of the centroids, then it gets labelled as that centroid and vice-versa.
* For the 2 clusters formed, average value of the data points, grouped with either of the centroid, is calculated and these average values are then defined as the new clusters.
* This process repeats itself until both centroids converge to fixed points.

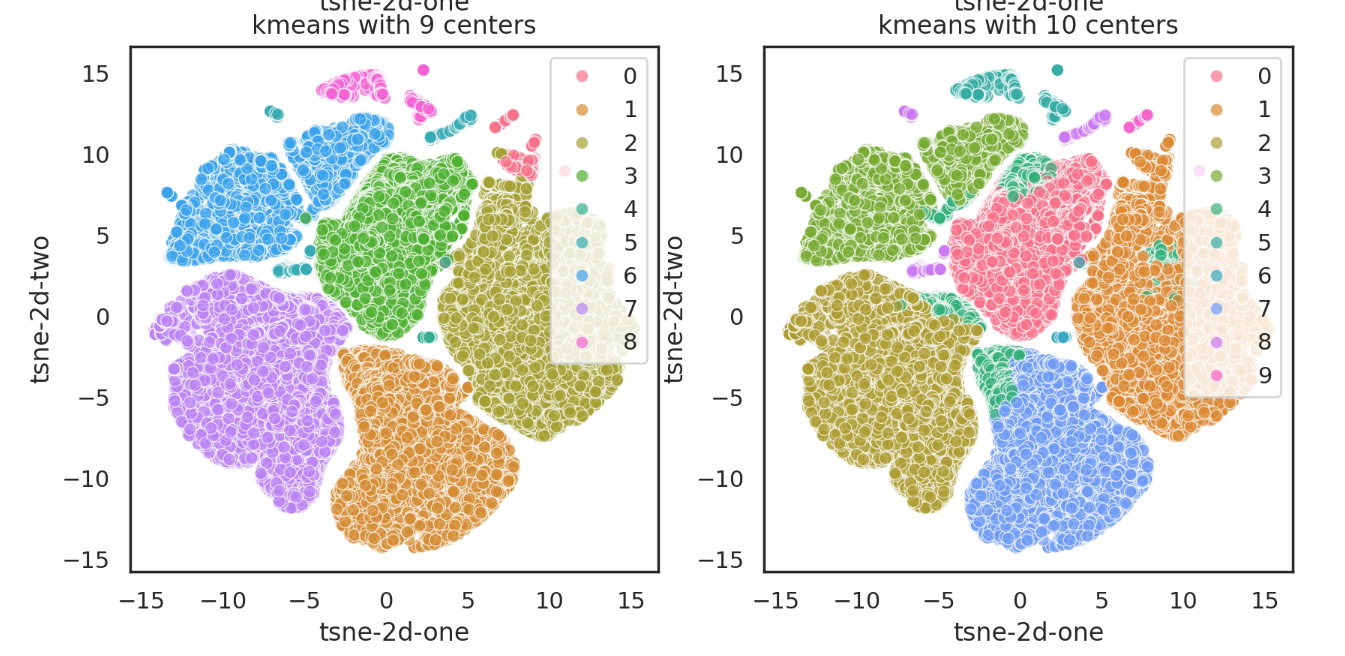
How to select k for clusters ?

**Elbow Method** : It is a method that plots the sum of squared error for a range of values of **k**. If this plot looks like an arm, then **k** is the value that resembles an **elbow** is selected. From this **elbow** value, sum of squared values (inertia) starts decreasing in a linear fashion and thus is considered as an optimal value.

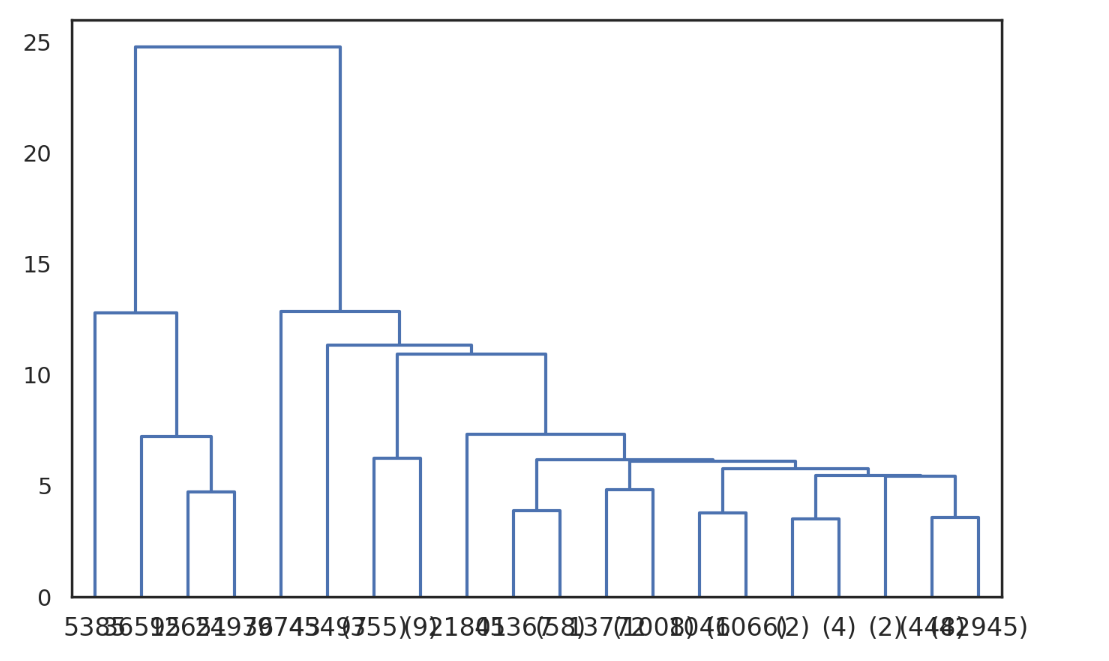


In this plot we can observe that for the values of **k > 9** the inertia tends to layoff so we can get clusters within the range of 7 to 11 and then select the best cluster with k centers.  
**Let's plot the clusters with 7 , 8 , 9 , 10 centers**





**The kmeans cluster with 8 centers has the optimal clustering performance.**

**Hierarchy clustering**

Choosing optimum number

of clusters using

Dendogram.

* In hierarchy clustering, to decide clusters with dendograms is difficult in this case.  
  Having too many clusters can make it difficult to understand the underlying structure of the data.
* As you move up the dendrogram, the distance between merged clusters gets larger.
* Choosing 8 clusters might represent a point where the clusters are still reasonably distinct while still offering some separation between them.
* Choosing 8 could be a balance between interpretability, maintaining within-cluster similarity, and maximizing between-cluster difference.

**Airbnb Cluster analysis**  
Understanding the behaviour of clusters using box\_plots for each variable.

First we add a column in our dataset as Kmeans\_pred which Is the prediction of all data observations into different clusters based on their behaviour by using Kmeans model with 8 clusters.

