**Applied Machine Learning**

**Lab Report 10**

**Hafiz Ahmad**

**19l-1316**

**Section-8A**

**INTRODUCTION:**

Unsupervised learning is a branch of machine learning that concentrates on analyzing data without predefined labels or target variables. Within unsupervised learning, clustering analysis is a widely utilized technique that aims to group similar data points based on their intrinsic characteristics. Two popular clustering algorithms are K-means and K-medoids. These algorithms find applications in diverse fields including data mining, pattern recognition, customer segmentation, and image analysis. In this experiment, the focus is on exploring the concept of clustering analysis using K-means and K-medoids algorithms, evaluating their efficacy, and investigating their potential applications.

**OBJECTIVES:**

The main objective of this experiment is to utilize the K-means and K-medoids algorithms on a given dataset, examine their clustering behavior, compare their performances, and acquire insights into the influence of algorithm parameters. By conducting this experiment, we aim to understand how these clustering algorithms operate, assess their effectiveness in grouping data points, and explore how different parameter settings can impact their performance.

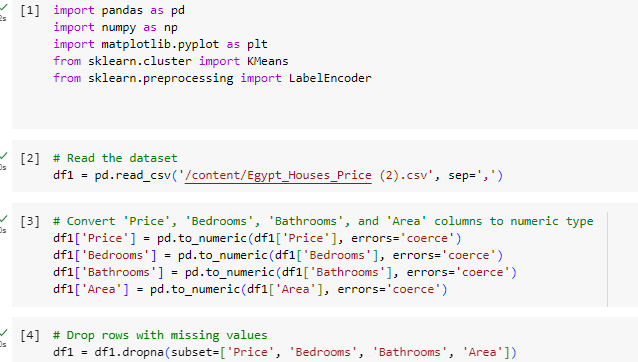
**Procedure:**

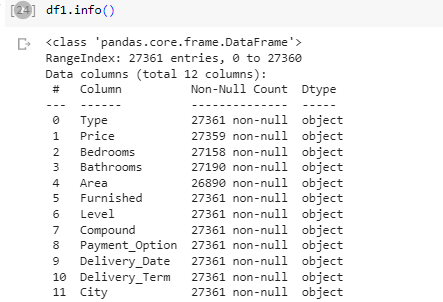
Data Preprocessing: The dataset must be preprocessed by converting relevant columns to numerical values before K-means clustering can be applied. The "Price," "Bedrooms," "Bathrooms," and "Area" columns are transformed into numerical types in this instance. Furthermore, any missing qualities in these segments should be tended to.

Process of Conversion: Label encoders are used to convert the categorical columns into numerical values.

Clustering by K-means: The elbow method is used to figure out how many clusters the data needs. The 'Price' and 'Area' columns are then selected as clustering numerical features. To see the results, the first few rows with the type, price, and assigned cluster are printed.

Clustering of K-medoids: K-medoids begins by perusing the dataset and switching the necessary segments over completely to numeric sorts. The missing values in rows are removed, and categorical columns are numerically encoded. The 'Restrooms' and 'Region' segments are chosen as the mathematical highlights for grouping. K-medoids bunching is applied with a predetermined number of groups. Medoids and cluster assignments are obtained as a result. A dissipate plot is created, where information focuses are plotted in view of their 'Restrooms' and 'Region' values and hued by their doled out groups. The plot provides a visual representation of the K-medoids algorithm's cluster formation.



**Before conversion:**

**After conversion:**A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated with medium confidence

A picture containing line, plot, diagram, screenshot

Description automatically generated

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with low confidence

A screen shot of a graph

Description automatically generated with low confidence



A picture containing text, font, screenshot

Description automatically generated

A screen shot of a computer program

Description automatically generated with low confidence

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

**Application:**

Grouping examination involving K-means and K-medoids calculations tracks down applications in different fields, including:

Identifying Anomalies: Outliers or abnormal patterns that deviate from expected behavior can be identified with the aid of clustering.

Analyses of Social Networks: Clustering provides insights into social structures and relationships by assisting in the identification of communities or groups within a social network.

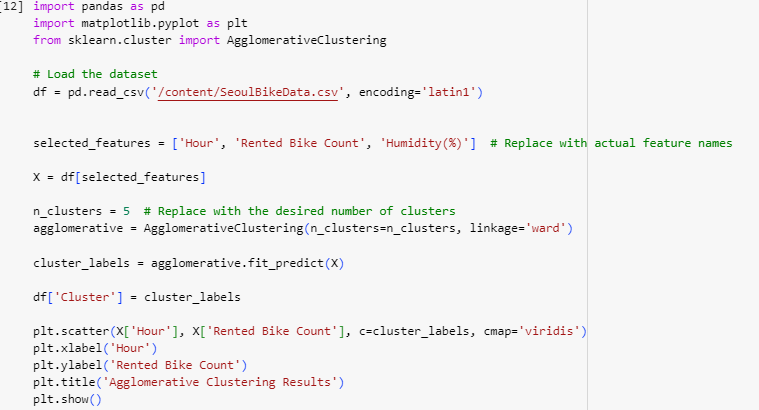
**Issues:**

During the process of clustering analysis, there are various challenges that can arise. One common challenge is determining the optimal number of clusters for the given data. It is important to strike a balance between oversimplifying the clustering solution by using too few clusters and overcomplicating it by using too many clusters. If the number of clusters is too low, the resulting clusters may be too broad and fail to capture meaningful distinctions within the data. This can lead to a loss of information and insights. On the other hand, if the number of clusters is too high, it may result in overfitting, where the model is excessively tailored to the training data and does not generalize well to new data. To address this challenge, various methods can be employed, such as the elbow method, silhouette analysis, or the gap statistic. These techniques help determine the optimal number of clusters by assessing the compactness and separation of the clusters.

**Conclusion:**

We investigated the capabilities, limitations, and wide range of applications of the K-means and K-medoids clustering algorithms in this experiment, which included customer segmentation, image compression, anomaly detection, and social network analysis. The bits of knowledge acquired can be important for dynamic cycles.

**Post lab: On Seoul bike data set:**



A picture containing text, screenshot, diagram, plot

Description automatically generated