****

**Final Project of Data Science**

**Report on K-Means Clustering Algorithm**

**Section: S8**

**Submitted To: Dr. Abdullah**

|  |  |
| --- | --- |
| **Name** | **Roll no.** |
| **Sara Arshad** | **L1F20BSSE0226** |

Contents

[**Introduction:** 3](#_Toc157801902)

[**Uses:** 3](#_Toc157801903)

[**Advantages:** 3](#_Toc157801904)

[**Disadvantages:** 3](#_Toc157801905)

[**Code:** 4](#_Toc157801906)

[**Results:** 5](#_Toc157801907)

**K-Means Clustering**

## **Introduction:**

K-Means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping subsets (clusters). The algorithm aims to minimize the variance within each cluster and maximize the variance between different clusters.

## **Uses:**

1. **Image Segmentation:**

K-Means clustering can be used to segment images by grouping similar pixels together.

1. **Customer Segmentation:**

Businesses use K-Means to segment customers based on their behavior, helping in targeted marketing.

1. **Anomaly Detection:**

Detecting unusual patterns or outliers in data by identifying clusters with significantly fewer data points.

1. **Document Clustering:**

Grouping similar documents together, aiding in tasks like topic modeling.

1. **Genetic Data Analysis:**

Identifying patterns in genetic data for understanding genetic similarities and differences.

## **Advantages:**

1. **Simplicity:**

Easy to implement and understand, making it a go-to choice for initial exploratory data analysis.

1. **Efficiency:**

Computationally efficient, especially with a large number of variables.

1. **Scalability:**

Can handle large datasets and is relatively scalable.

1. **Versatility:**

Applicable to various types of data, not limited to a specific domain.

## **Disadvantages:**

1. **Sensitive to Initial Centroids:**

The choice of initial centroids can influence the final clustering result.

1. **Dependence on K:**

The algorithm requires specifying the number of clusters (K) beforehand, which might not always be known.

1. **Spherical Clusters Assumption:**

Assumes that clusters are spherical and equally sized, which may not be true for all datasets.

1. **Sensitive to Outliers:**

Outliers can significantly impact the clustering result.

1. **Hard Assignment:**

Each data point is assigned to exactly one cluster, which might not reflect the true underlying structure of complex datasets.

## **Code:**

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

dataframe = pd.read\_csv('train.csv')

print(dataframe.info())

non\_numeric\_columns = dataframe.select\_dtypes(exclude=['float64', 'int64']).columns

dataframe\_numeric = dataframe.drop(columns=non\_numeric\_columns)

print(dataframe\_numeric.describe())

scaler = StandardScaler()

dataframe\_scaled = scaler.fit\_transform(dataframe\_numeric)

inertia = []

for i in range(1, 11):

    kmeans = KMeans(n\_clusters=i, random\_state=42)

    kmeans.fit(dataframe\_scaled)

    inertia.append(kmeans.inertia\_)

optimal\_clusters = 3

kmeans = KMeans(n\_clusters=optimal\_clusters, random\_state=42)

dataframe['Cluster'] = kmeans.fit\_predict(dataframe\_scaled)

print(dataframe['Cluster'].value\_counts())

outliers = dataframe[dataframe['Cluster'] == -1]

print('Outliers:')

print(outliers)

plt.figure(figsize=(8, 6))

plt.plot(range(1, 11), inertia, marker='o')

plt.title('Elbow Method')

plt.xlabel('Number of Clusters')

plt.ylabel('Inertia')

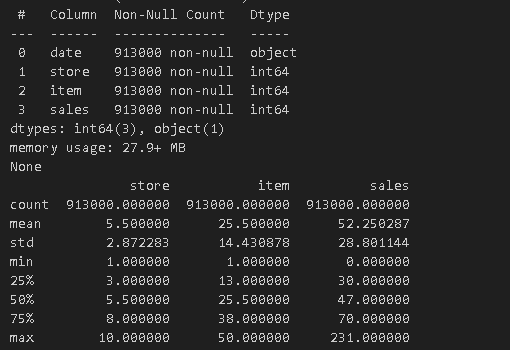
plt.show()

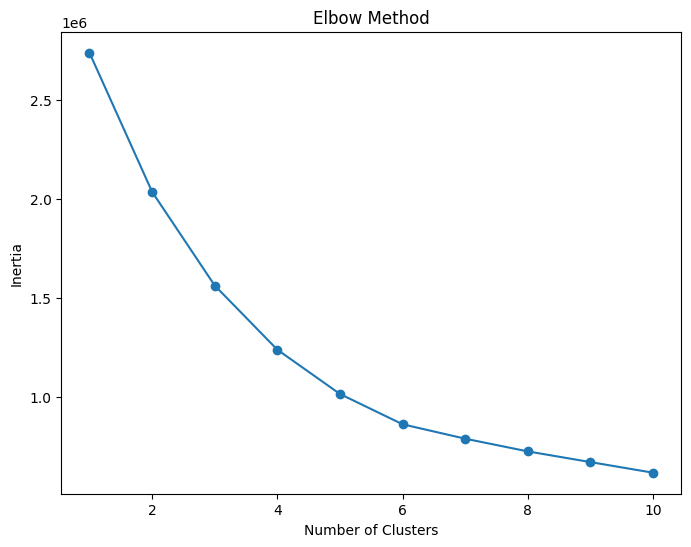
pd.plotting.scatter\_matrix(dataframe, c=dataframe['Cluster'], figsize=(12, 8), marker='o', hist\_kwds={'bins': 20}, alpha=0.8)

plt.suptitle('Scatter Matrix with Clusters')

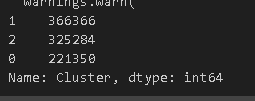
plt.show()

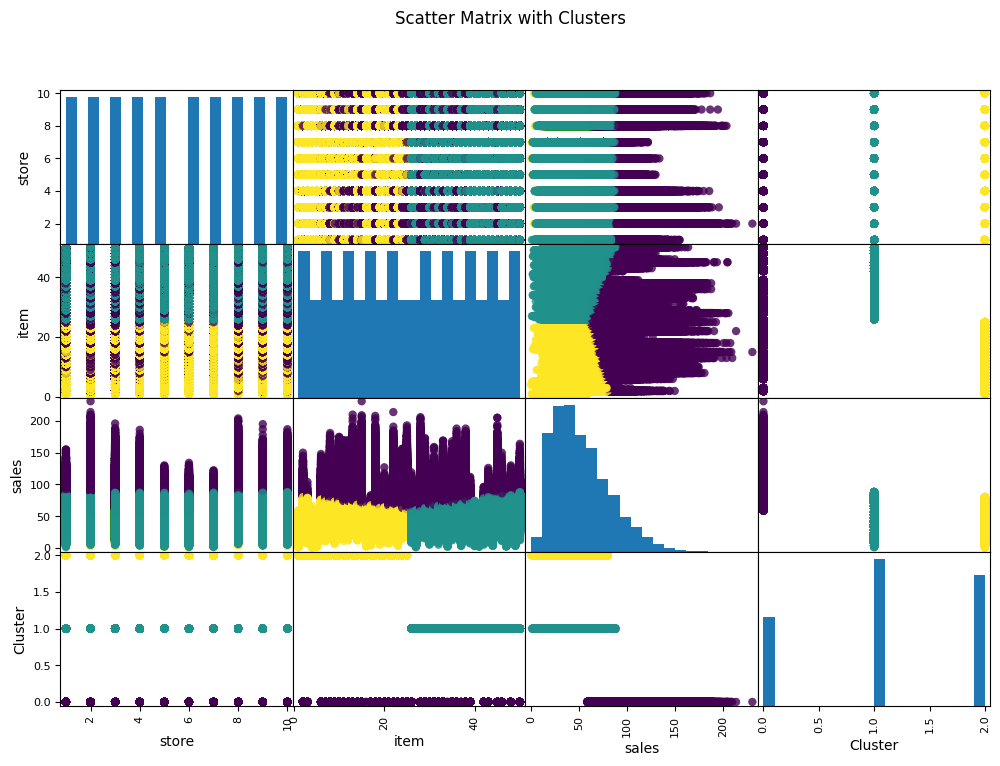
## **Results:**





* The graph shows the result of applying the Elbow Method to determine the optimal number of clusters for the K-means clustering algorithm. The Elbow Method plots the inertia (or within-cluster sum of squared distances) against the number of clusters and looks for the point where the inertia curve starts to bend or form an elbow shape. This point indicates the optimal number of clusters that minimizes the inertia while avoiding overfitting.
* The graph has the title "Elbow Method" and two axes: the x-axis is labeled "Number of Clusters" and the y-axis is labeled "Inertia". The x-axis ranges from 0 to 10 and the y-axis has a scale indicated by '1e6' at the top, suggesting it is multiplied by a factor of one million. The graph has blue dots representing data points for each even number on the x-axis from 2 to 10. The line connecting these points starts high at (2, ~2.5e6) and decreases sharply until around point (4, ~1.5e6), after which it continues decreasing but at a much slower rate.
* The graph suggests that the optimal number of clusters is somewhere between 2 and 4, as the inertia curve starts to bend at that range. However, the exact number of clusters is not very clear, as there is no distinct elbow point. This might indicate that the data is not well suited for clustering or that the K-means algorithm is not the best choice for this data. Alternatively, it might require further analysis or additional methods to confirm the optimal number of clusters.





* The image shows a scatter matrix with clusters that visualizes the data relationships between store, item, sales, and cluster. Each cell in the matrix represents a scatter plot or histogram comparing two variables. Different colors (blue, yellow, green) are used to represent different clusters within the data.
* The image suggests that the data has been clustered using the K-means algorithm with four clusters as the optimal number. The cluster labels are shown on the bottom right cell of the matrix.
* The histograms on the diagonal show the distribution of each variable across the data. The store variable has a uniform distribution, meaning that each store has an equal number of data points. The item variable has a bimodal distribution, meaning that there are two peaks or modes in the data. The sales variable has a skewed distribution, meaning that most of the data points are concentrated on the lower end of the scale. The cluster variable has a balanced distribution, meaning that each cluster has an equal number of data points.
* The scatter plots on the off-diagonal cells show the relationship between two different variables. Some of the plots show a clear clustering pattern, especially between sales and item. This indicates that there is a strong correlation between these two variables and that they are good predictors of the cluster membership. Other plots show a weak or no clustering pattern, such as between store and item. This indicates that there is little or no correlation between these two variables and that they are not good predictors of the cluster membership.
* The image provides some useful insights into the data and the clustering results. However, it also has some limitations and areas for further investigation. For example:
* The image does not provide any context or numerical data to support the analysis. It is not clear what the data represents, what the units are, or what the source is. It is also not clear how the data was preprocessed, scaled, or normalized before applying the clustering algorithm.
* The image does not provide any evaluation or validation of the clustering results. It is not clear how the optimal number of clusters was determined, what the quality or performance of the clustering algorithm was, or how the clusters were interpreted or labeled. It is also not clear how the clustering results compare to other methods or models, such as supervised learning or regression.
* The image does not provide any recommendations or implications based on the clustering results. It is not clear what the purpose or goal of the clustering analysis was, what the benefits or drawbacks of the clustering results are, or what actions or decisions can be made based on the clustering results.

**Dataset File:**

Dataset is present in: **train.csv**