REPORT

Steps in KMeans –

1. **Loaded the Data using Pandas. pd.read_csv('/path/file.csv')**

2. **Selected the features that will be included in the KMeans model.**

   For Example, in Mall data analysis original data had following features

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)

   In this case CustomerID is not required for forming cluster in KMeans.

3. **The Gender feature is a categorical feature. It must be converted to numerical to feed it to KMeans model.**

4. **After all these steps found all the necessary features and in required format to feed it to the KMeans model.**

	Age	Annual Income (k$)	Spending Score (1-100)	Female	Male
0	19	15	39	0.0	1.0
1	21	15	81	0.0	1.0
2	20	16	6	1.0	0.0
3	23	16	77	1.0	0.0
4	31	17	40	1.0	0.0

5. **Performing standardization for scaling the values the values. Being a distance-based algorithm KMeans works best with features being in same scale.**

REPORT

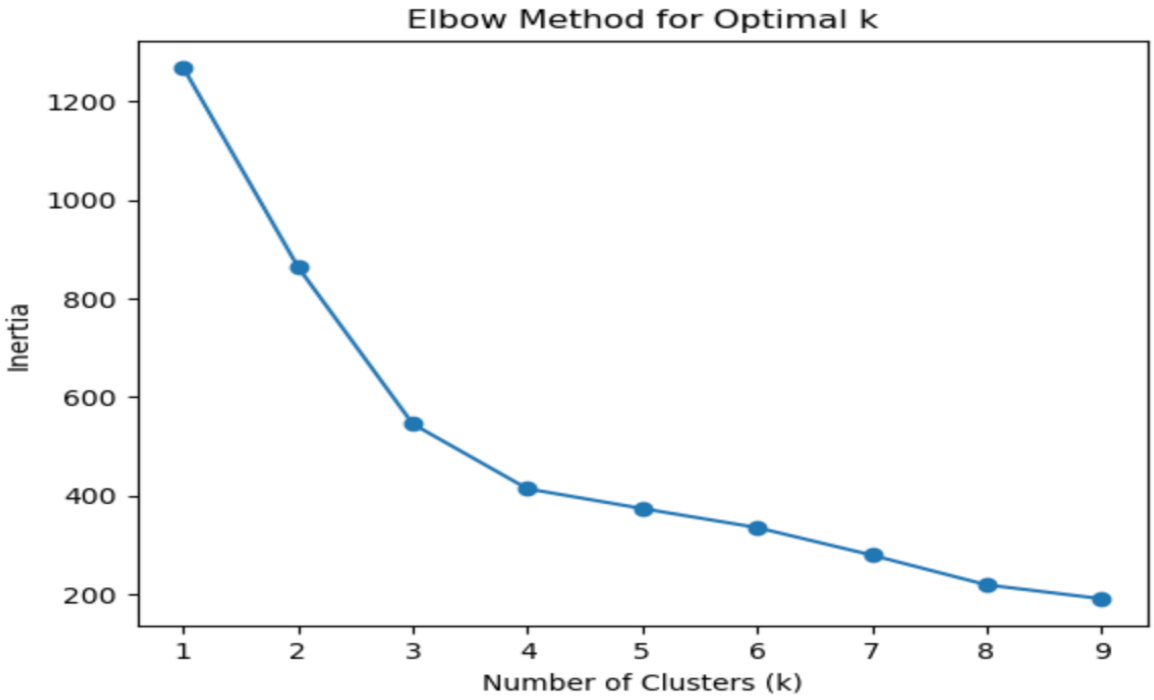
```
array([[-1.42456879, -1.73899919, -0.43480148, -1.12815215,  1.12815215],
       [-1.28103541, -1.73899919,  1.19570407, -1.12815215,  1.12815215],
       [-1.3528021 , -1.70082976, -1.71591298,  0.88640526, -0.88640526],
       [-1.13750203, -1.70082976,  1.04041783,  0.88640526, -0.88640526],
       [-0.56336851, -1.66266033, -0.39597992,  0.88640526, -0.88640526],
       [-1.20926872, -1.66266033,  1.00159627,  0.88640526, -0.88640526],
       [-0.27630176, -1.62449091, -1.71591298,  0.88640526, -0.88640526],
       [-1.13750203, -1.62449091,  1.70038436,  0.88640526, -0.88640526],
       [ 1.80493225, -1.58632148, -1.83237767, -1.12815215,  1.12815215],
```

6. **Now preforming KMeans but an initial estimate of 4 clusters .**

   kmeans = KMeans(n_clusters= 4, random_state=20) . Using scaled values.

	Age	Annual Income (k$)	Spending Score (1-100)	Female	Male	labels
0	-1.424569	-1.738999	-0.434801	-1.128152	1.128152	2
1	-1.281035	-1.738999	1.195704	-1.128152	1.128152	2
2	-1.352802	-1.700830	-1.715913	0.886405	-0.886405	3
3	-1.137502	-1.700830	1.040418	0.886405	-0.886405	3

7. **For estimating the optimum number of clusters implementing Elbow Method.**

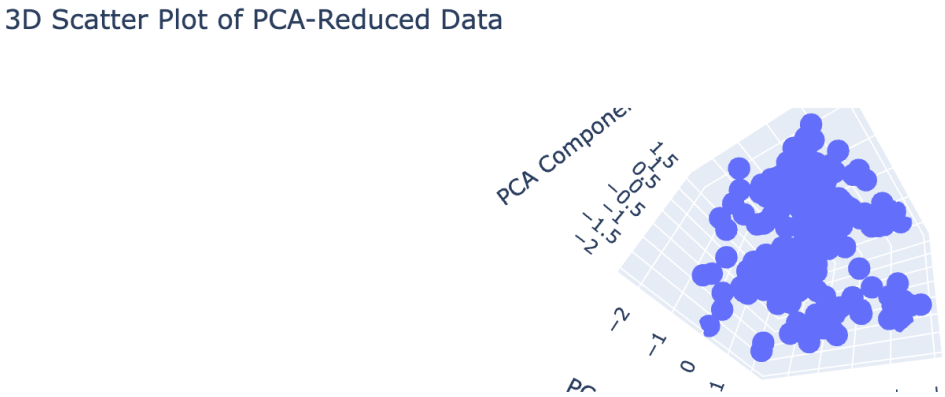


REPORT

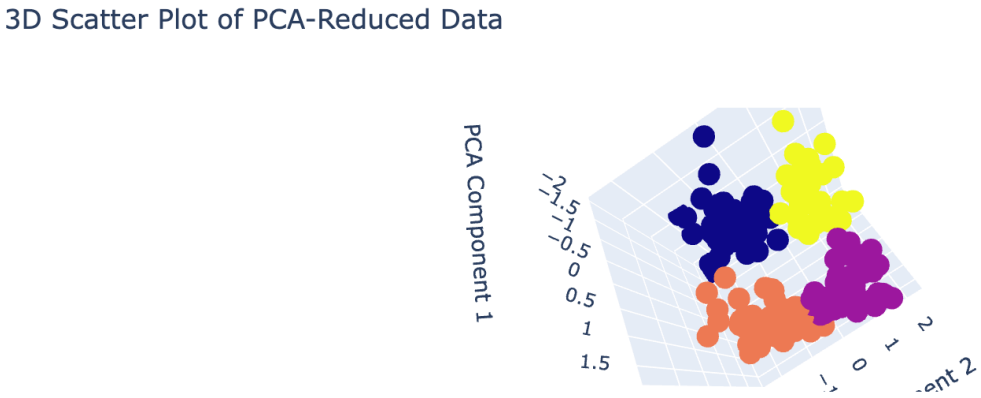
Observation -

At k=4 there a stabilization in inertia values. Hence 4 clusters are optimum.

8. **Now using the scaled value applying PCA dimensionality reduction technique to reduce the number of features to 3 principal components.**

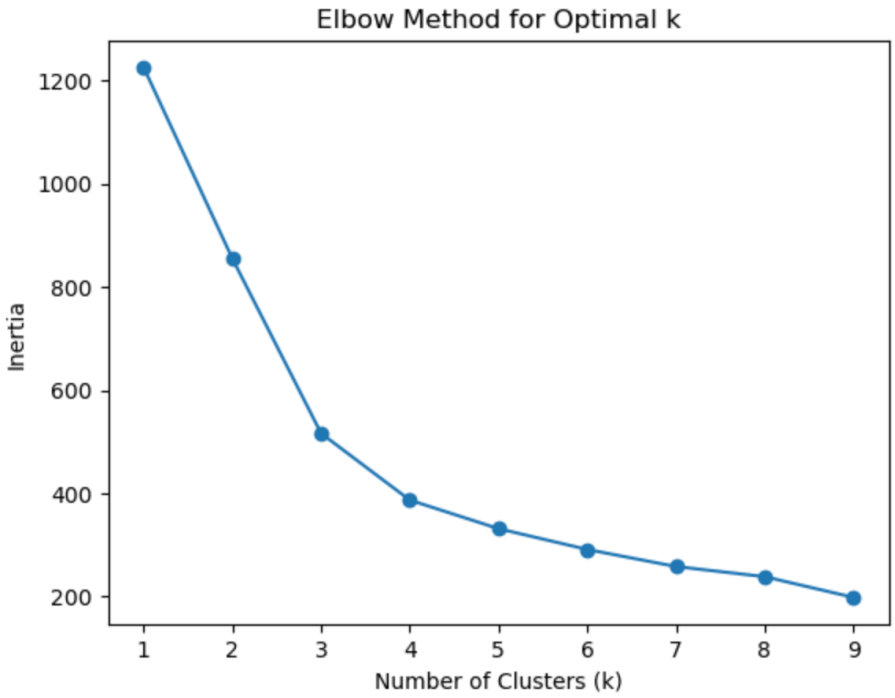


9. **Now implementing kMeans on PCA with n_components = 3.**



10. Again, using Elbow method.

REPORT



Observations for the Clusters(Centroid below) –

[[0.34458885 0.49853901 0.10811216 0.88640526 -0.88640526]
[0.75982983 0.07086791 -0.81492926 -1.12815215 1.12815215]
[-0.76072691 0.05496398 0.83369302 -1.12815215 1.12815215]
[-0.62577433 -0.83703899 -0.02970693 0.88640526 -0.88640526]]

Analyzing Each Cluster:

1. **Cluster 0** (Centroid: [0.34458885 0.49853901 0.10811216 0.88640526 -0.88640526]):
    - **Age**: The value 0.34458885 suggests that the average customer in this cluster is slightly above the mean age.
    - **Income**: The value 0.49853901 indicates that this group has a slightly above-average income.
    - **Spending Score**: The value 0.10811216 means that the spending score is close to average for this cluster.
    - **Gender**: Since the values for Female (0.886) and Male (-0.886) are opposites, this suggests that this cluster is predominantly female.

Interpretation: This cluster likely represents **younger, relatively higher-income female customers** with an average spending score.

2. **Cluster 1** (Centroid: [0.75982983 0.07086791 -0.81492926 -1.12815215 1.12815215]):

REPORT

- **Age**: The value 0.75982983 indicates that this group is older than average.
- **Income**: The value 0.07086791 suggests that this group has close to average income.
- **Spending Score**: The value -0.81492926 indicates that this group has a **low spending score**, suggesting they spend less.
- **Gender**: Since the values for Female (-1.12815215) and Male (1.12815215) are opposites, this indicates that this cluster is predominantly male.

Interpretation: This cluster could represent **older male customers** with average income and lower spending behavior.

3. **Cluster 2** (Centroid: [-0.76072691 0.05496398 0.83369302 -1.12815215 1.12815215]):
   - **Age**: The value -0.76072691 indicates that this group is younger than average.
   - **Income**: The value 0.05496398 suggests that this group has almost average income.
   - **Spending Score**: The value 0.83369302 indicates that this group has a **high spending score**.
   - **Gender**: Again, the gender values (-1.12815215 for female, 1.12815215 for male) suggest this is a predominantly male cluster.

Interpretation: This cluster represents **young, high-spending male customers** with average income.

4. **Cluster 3** (Centroid: [-0.62577433 -0.83703899 -0.02970693 0.88640526 -0.88640526]):
   - **Age**: The value -0.62577433 suggests this group is younger than average.
   - **Income**: The value -0.83703899 indicates this group has lower income than average.
   - **Spending Score**: The value -0.02970693 suggests an average spending score, close to the mean.
   - **Gender**: The values for Female (0.88640526) and Male (-0.88640526) indicate this is a predominantly **female cluster**.

Interpretation: This cluster could represent **younger, lower-income female customers** with an average spending score.

Key Takeaways –

Cluster 0 (Young, higher-income females): You might want to target this group with premium products or services, as they have relatively high income.
Cluster 1 (Older, lower-income males): This group might respond better to budget-friendly or value-focused products.
Cluster 2 (Young, high-spending males): You could target this group with high-end, trendy products since they have high spending potential.

REPORT

Cluster 3 (Young, low-income females): This group might be ideal for budget-friendly offerings, especially in fashion or affordable luxury items.

Customer Segmentation in Retail and E-commerce(Observations from my results)

Application: Clustering is widely used in customer segmentation, allowing businesses to categorize their customers into distinct groups based on shared characteristics, such as purchasing behavior, demographic attributes (age, gender, income), or spending patterns.

Use Case:

- **Targeted Marketing**: Based on the clustering results, you can personalize marketing campaigns. For example, if Cluster 0 represents **young, higher-income females** with an average spending score, businesses can target them with high-end products or exclusive offers.
- **Customer Engagement**: A retailer can use the clusters to design specific promotions for **Cluster 1** (older, lower-income males), such as discounts or loyalty programs, that appeal to this group's price sensitivity.
- **Product Recommendations**: For **Cluster 2** (young, high-spending males), businesses can recommend trendy or premium products as this group is likely to spend more.

Benefit: Increases customer satisfaction and conversion rates by offering personalized and relevant products and services to each group.

2. Anomaly Detection in Financial Services

Application: Clustering can help in anomaly detection, especially in the financial industry. It can be used to identify unusual customer behavior, such as abnormal spending patterns or fraud detection.

3. Inventory Management in Retail and Supply Chain

Application: In inventory management, clustering helps identify products with similar demand patterns, which can lead to more efficient stock management and forecasting.

4. Healthcare Industry: Patient Segmentation

Application: Clustering can be used to segment patients based on their medical history, symptoms, or risk factors, which can lead to better diagnosis, treatment planning, and personalized care.

And many more....