Task 3: Customer Segmentation / Clustering

Report on my clustering results

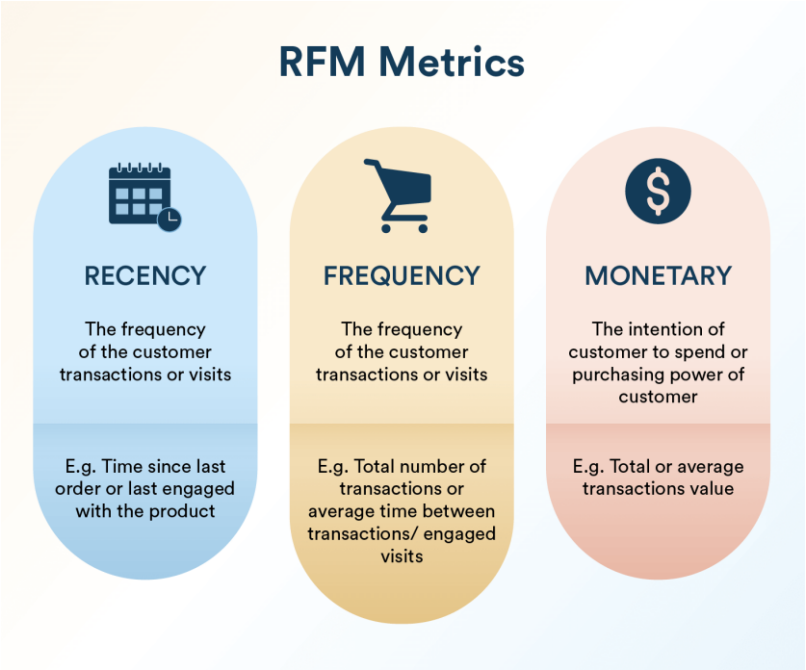
Here, for the customer segmentation task, I have used **K-Means Clustering** algorithm.

- **Data Preprocessing**

Initially, the data was pre-processed by

- Adding RFM features and Selecting the required features for clustering

  The Recency, Frequency and Monetary (RFM) features were extracted from the 'transactions' data and merged with the customers profile data.



  Then the customer's tenure i.e. days since they have signed up was calculated and appended. The following features were selected:

  ['Region', 'Tenure', 'Recency', 'Frequency', 'Monetary']

  One-hot encoding of the Region feature was done since it was categorical.

- Handling Missing Values

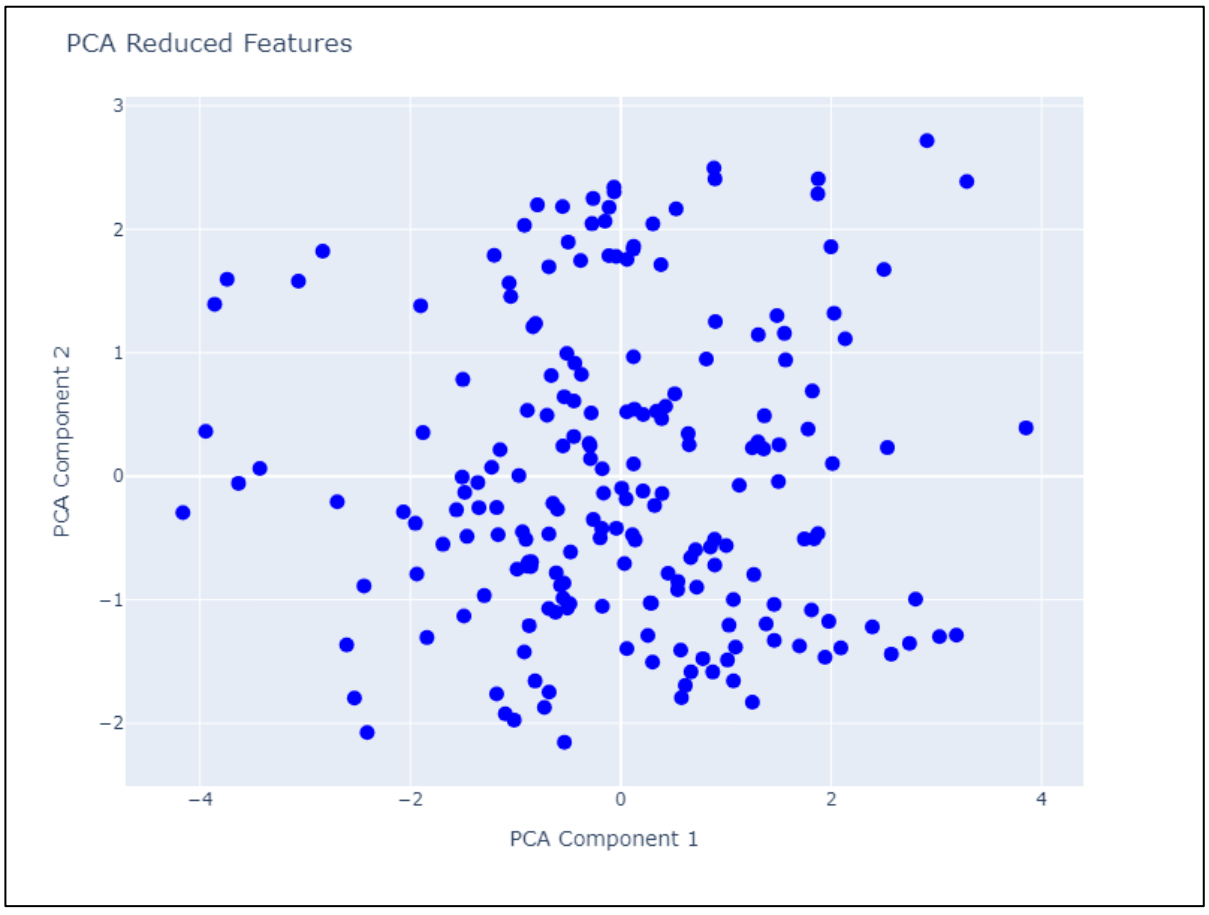
  Upon investigation, C0180 – Amy Carpenter hadn't made any transaction, hence the RFM values were imputed as '0' to avoid the case of having a NaN.

- Standardizing the data

  The customer features were standardized. Standardization is a common preprocessing step in machine learning that involves scaling the features of the dataset so that they have a mean of 0 and a standard deviation of 1. This process can improve the performance of many machine learning algorithms by ensuring that each feature contributes equally to the model.

- Dimensionality Reduction

  The data's dimensions were reduced with Principal Component Analysis (PCA) for easier computation and visualisation of the clusters. The number of components it was reduced was 2.



- **Performing Cluster Evaluation with K-Means using the following metrics.**

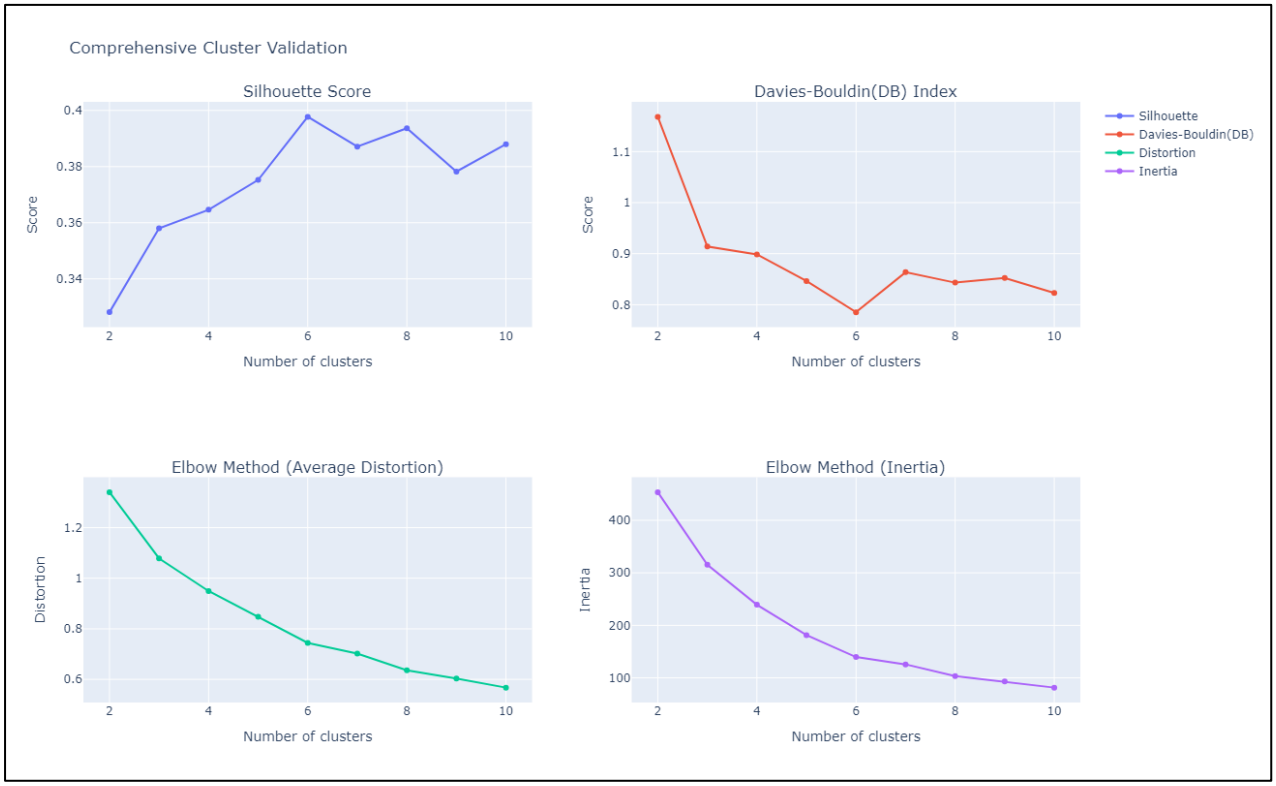
For finding the optimal number of clusters, I have performed cluster validation with the following metrics:

- Silhouette Scores
  It measures how similar an object is to its own cluster compared to other clusters, with values ranging from -1 to 1. A higher score indicates well-defined and distinct clusters, while a lower score suggests overlapping or poorly separated clusters.

- **Davies-Bouldin(DB) Index**
  It evaluates clustering performance by calculating the average similarity between each cluster and its most similar cluster. Lower DB index values indicate better clustering, as they signify more compact and well-separated clusters.

- Elbow Method (Average Distortion)
  This by using average distortion measures the sum of squared distances between data points and their corresponding cluster centroids. The optimal number of clusters is chosen at the "elbow" point, where adding more clusters no longer significantly reduces distortion.

- Elbow Method (Inertia)
  This by using inertia calculates the total within-cluster sum of squared distances from data points to their centroids. The optimal number of clusters is identified at the "elbow" point, where the inertia reduction starts to diminish as more clusters are added.



```
Recommendations based on different metrics:
----------------------------------------
Silhouette Score (higher is better): Optimal k = 6
Davies-Bouldin(DB) Index (lower is better): Optimal k = 6
Elbow Method (Distortion): Suggested k = 4
Elbow Method (Inertia): Suggested k = 4
```

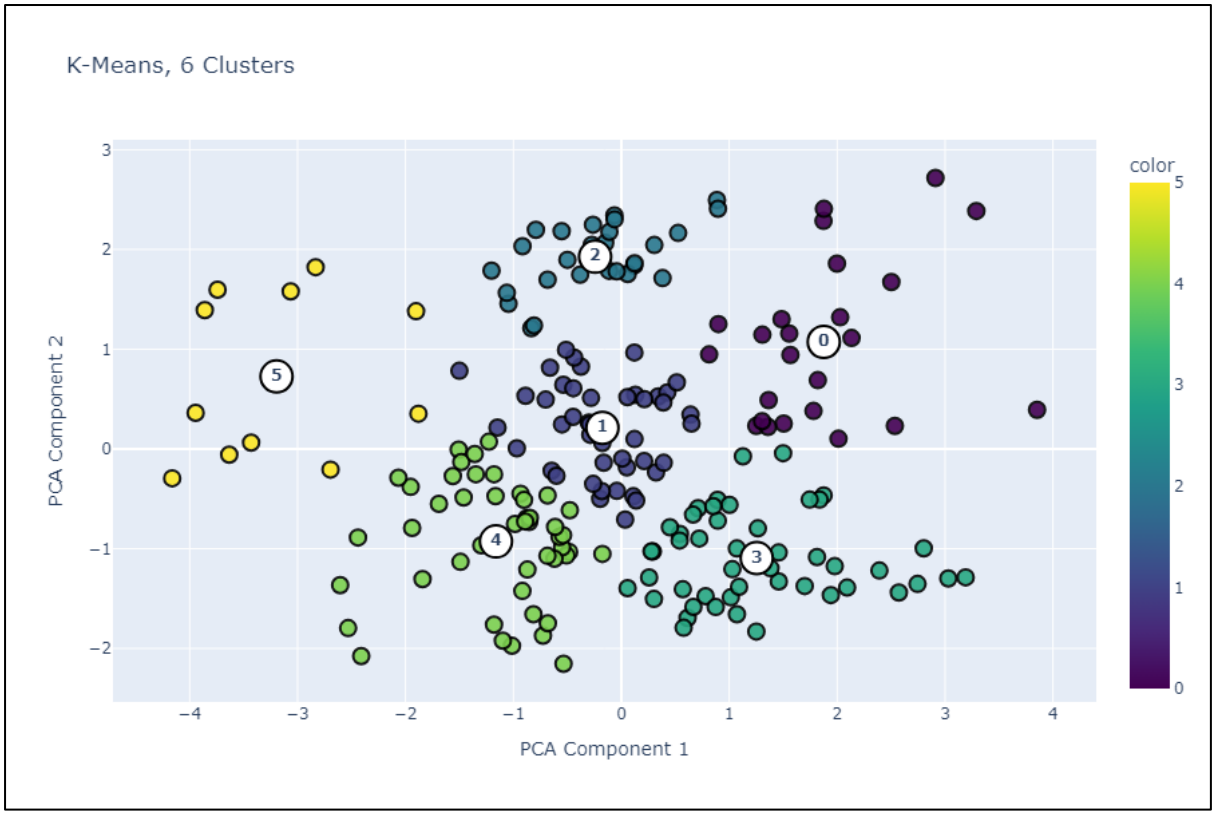
The recommended optimal number of clusters were:

- **6**, as per Silhouette Score and DB Index
- **4**, as per Elbow Method

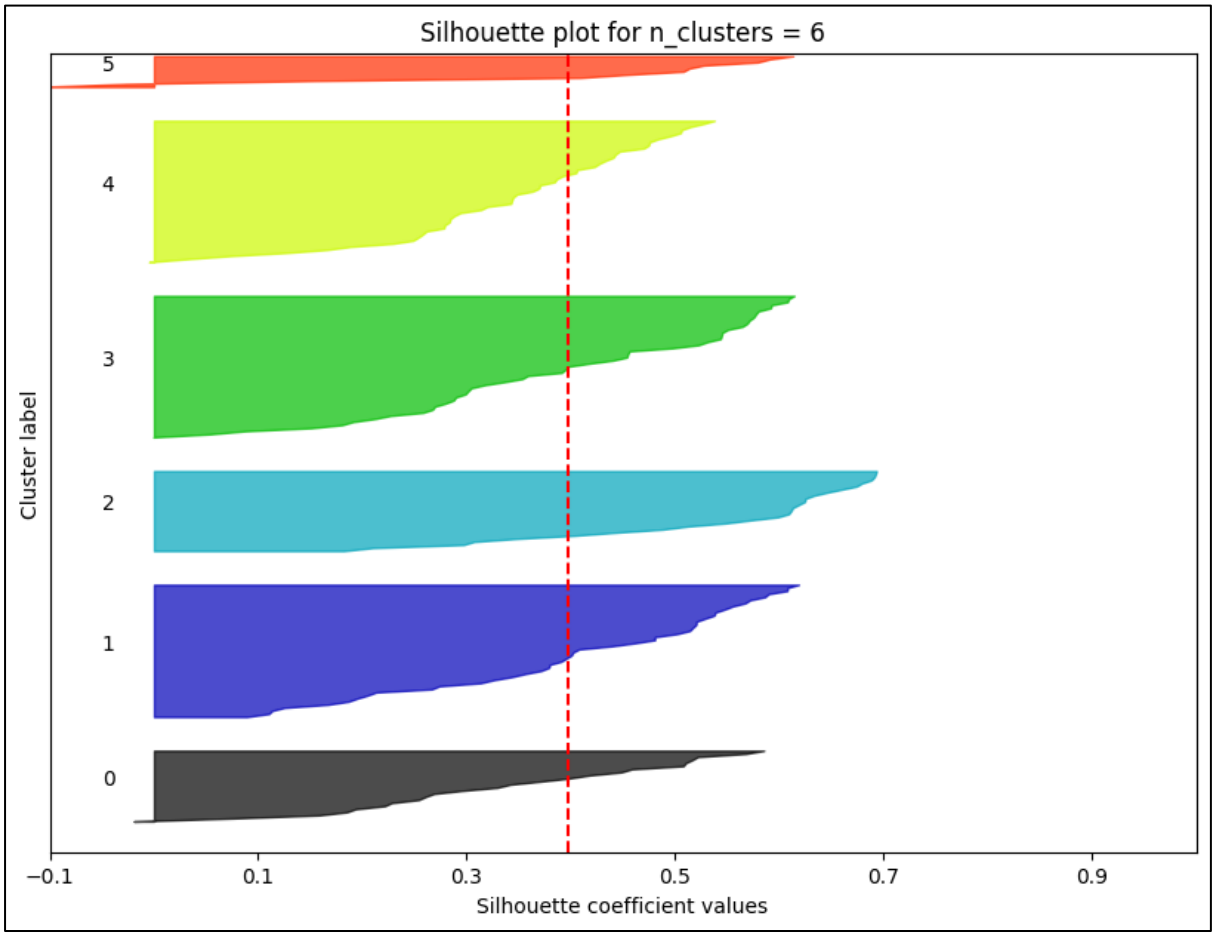
- **Clustering for Optimal k = 6 (As per Silhouette Score and DB Index)**

```
Silhouette Coefficient for n_clusters=6: 0.3977
Davies-Bouldin Index for n_clusters=6: 0.7853
```

Clusters visualisation



Plotting the Silhouette Coefficient values



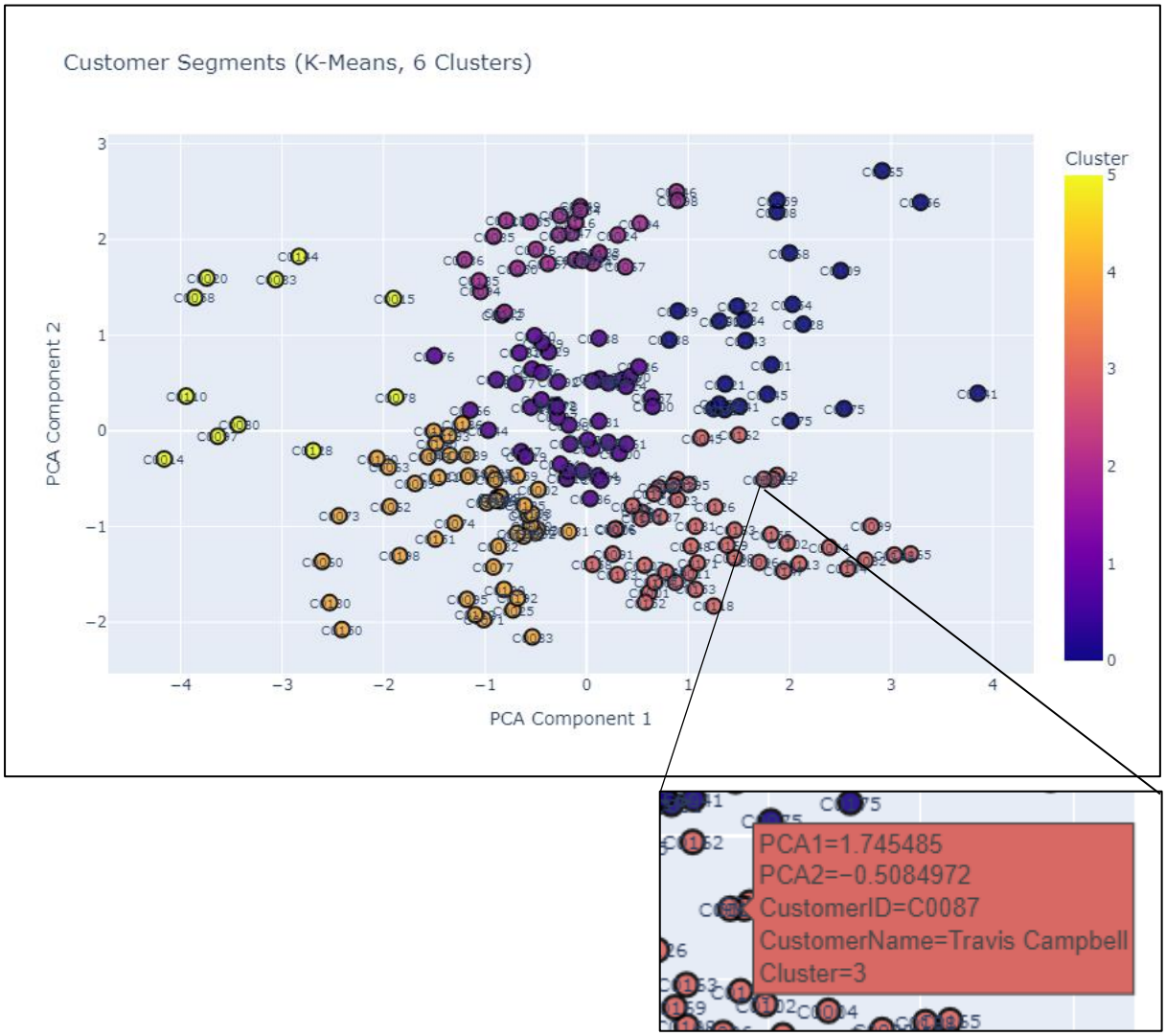
Mapping the 6 Clusters back to Customers

	CustomerID	CustomerName	Tenure	Recency	Frequency	Monetary	Region_Asia	Region_Europe	Region_North America	Region_South America	Cluster	PCA1	PCA2
0	C0001	Lawrence Carroll	933	87.0	5.0	3354.52	False	False	False	True	3	0.613997	-1.694224
1	C0002	Elizabeth Lutz	1080	56.0	4.0	1862.74	True	False	False	False	4	-0.477287	-0.613687
2	C0003	Michael Rivera	327	157.0	4.0	2725.38	False	False	False	True	4	-0.483765	-1.032006
3	C0004	Kathleen Rodriguez	842	36.0	8.0	5354.88	False	False	False	True	3	2.391476	-1.219433
4	C0005	Laura Weber	897	85.0	3.0	2034.24	True	False	False	False	4	-0.933873	-0.448706

Customers count for each cluster

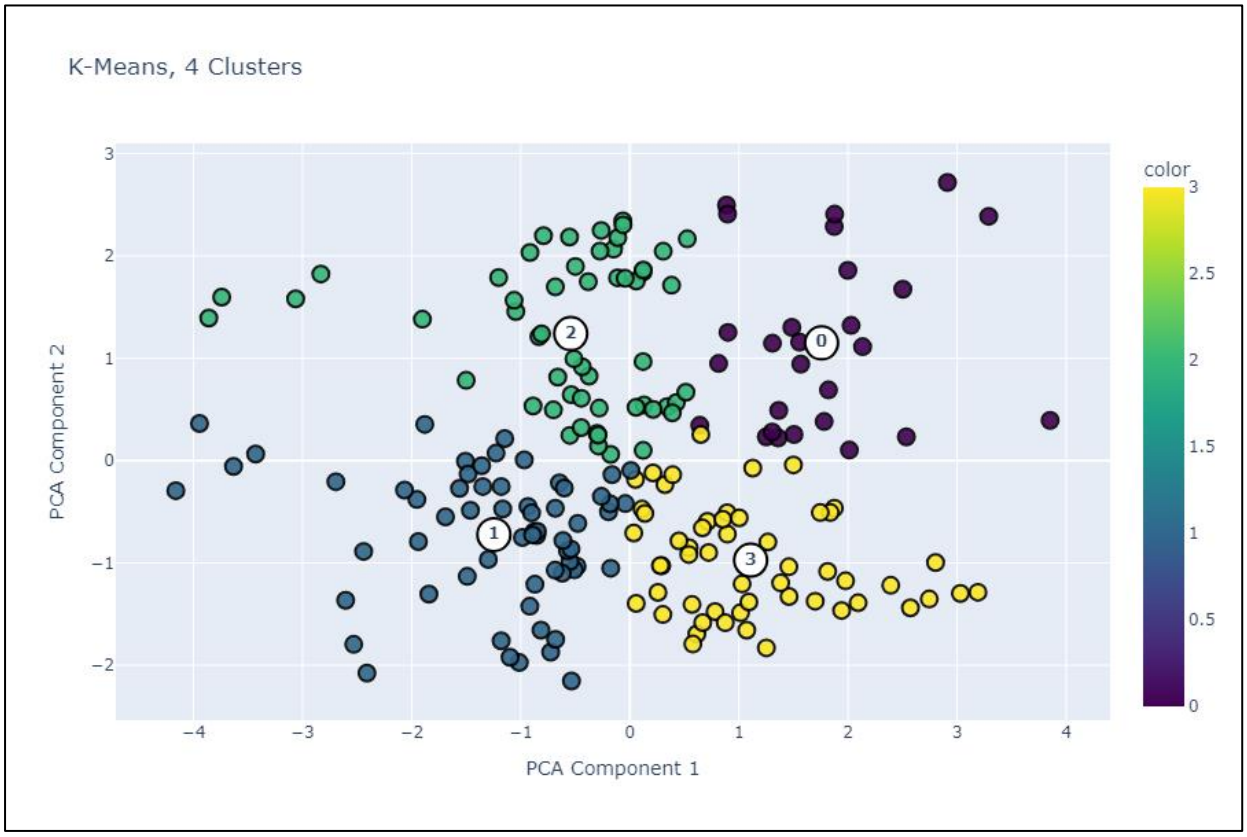
```
Cluster
0    24
1    44
2    27
3    47
4    47
5    11
Name: count, dtype: int64
```

Customers segments



- **Clustering for Optimal k = 4 (As per Elbow Method)**

Clusters visualisation



Mapping the 4 Clusters back to Customers

	CustomerID	CustomerName	Tenure	Recency	Frequency	Monetary	Region_Asia	Region_Europe	Region_North America	Region_South America	Cluster	PCA1	PCA2
0	C0001	Lawrence Carroll	933	87.0	5.0	3354.52	False	False	False	True	3	0.613997	-1.694224
1	C0002	Elizabeth Lutz	1080	56.0	4.0	1862.74	True	False	False	False	1	-0.477287	-0.613687
2	C0003	Michael Rivera	327	157.0	4.0	2725.38	False	False	False	True	1	-0.483765	-1.032006
3	C0004	Kathleen Rodriguez	842	36.0	8.0	5354.88	False	False	False	True	3	2.391476	-1.219433
4	C0005	Laura Weber	897	85.0	3.0	2034.24	True	False	False	False	1	-0.933873	-0.448706

Customers count for each cluster

```
Cluster
0    27
1    63
2    55
3    55
Name: count, dtype: int64
```

Customers segments