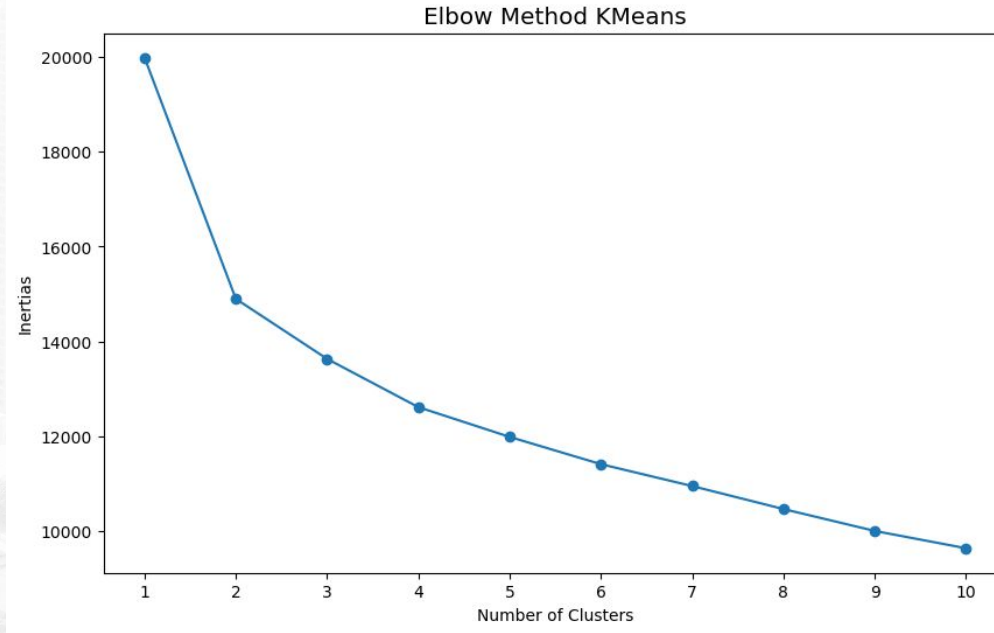


- The **Elbow Method** using **K-Means Clustering** is performed to get better visualisation

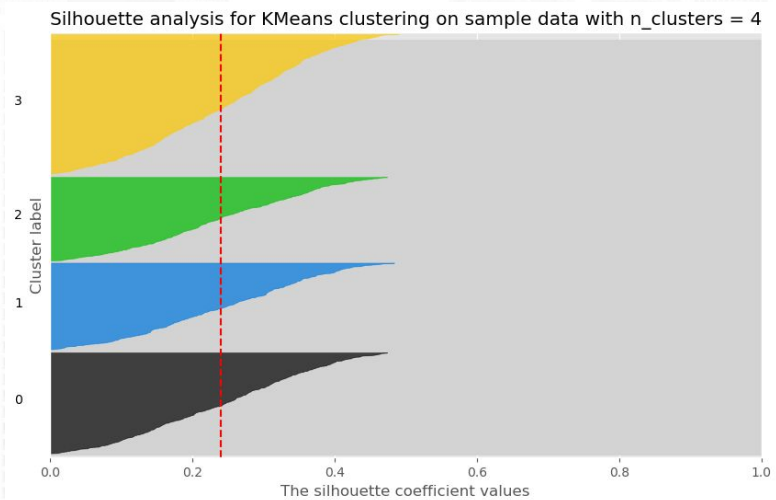
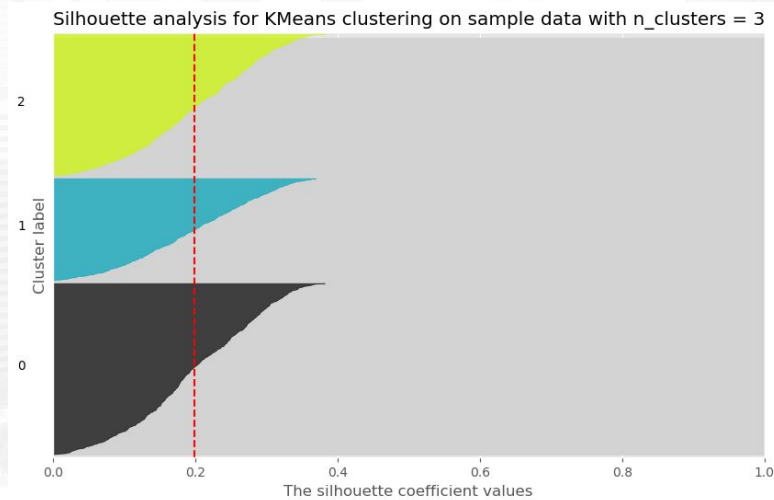


```
0    5085.083127
1    1265.713722
2    1017.344492
3     627.259852
4     575.144725
5     461.160959
6     484.931421
7     461.497115
8     364.992917
9                NaN
dtype: float64
```

- From the elbow method, the best cluster is 3. Although some confusion with 4 clusters, the difference between 3 to 4 is smaller rather than 4 to 5. However, keeping in mind 4 clusters to be check for modelling is cautious.

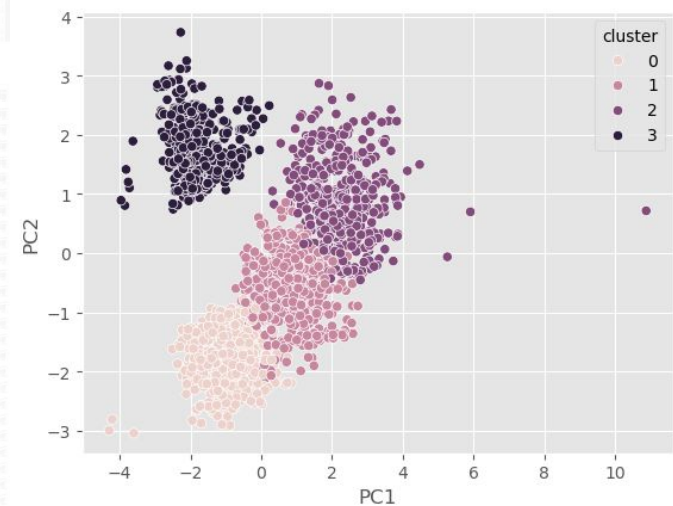
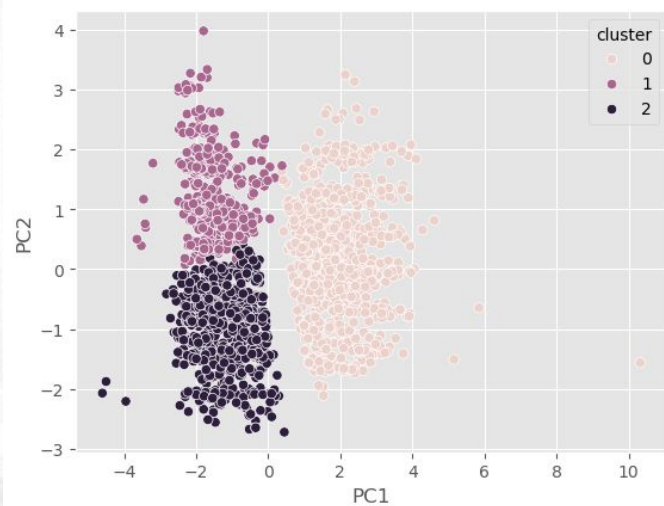
For more information, see jupyter notebook [here](#)

- Evaluation using **Silhouette Score** can be seen below between 3 and 4 clusters.



- From the graph, it can be seen that 1 and 0 in  $n\_cluster=3$  is being separated into three clusters in  $n\_cluster=4$  with more uniform shape.
- Silhouette coefficient is also higher for  $n\_cluster=4$  than  $n\_cluster=3$ .

- Evaluation using **PCA** can be seen below between 3 and 4 clusters.



- From the graph, it can be seen that a slight difference in clustering with `n_cluster=3` seems to be more spread than `n_cluster=4`
- In `n_cluster=4`, although cluster 2 and 3 seems to be overlap, there's a clear distinct as such cluster 2 is high values on PC2 but low values on PC1. While the opposite is for cluster 3.

In conclusion, the 4-cluster solution appears to be the better choice because:

- It has a slightly better silhouette score, indicating better-defined clusters.
- The PCA visualization strongly supports the existence of 4 distinct groups.
- It likely provides a more detailed and accurate representation of the data's underlying structure.