This report is a written documentation of the findings from the .ipynb file. Graphs covered in this report can also be found in the .ipynb file, there are added explanations and reasoning present in this documentation, and it is to be used as a supplementary material to the .ipynb file.

Choosing the optimal k-value

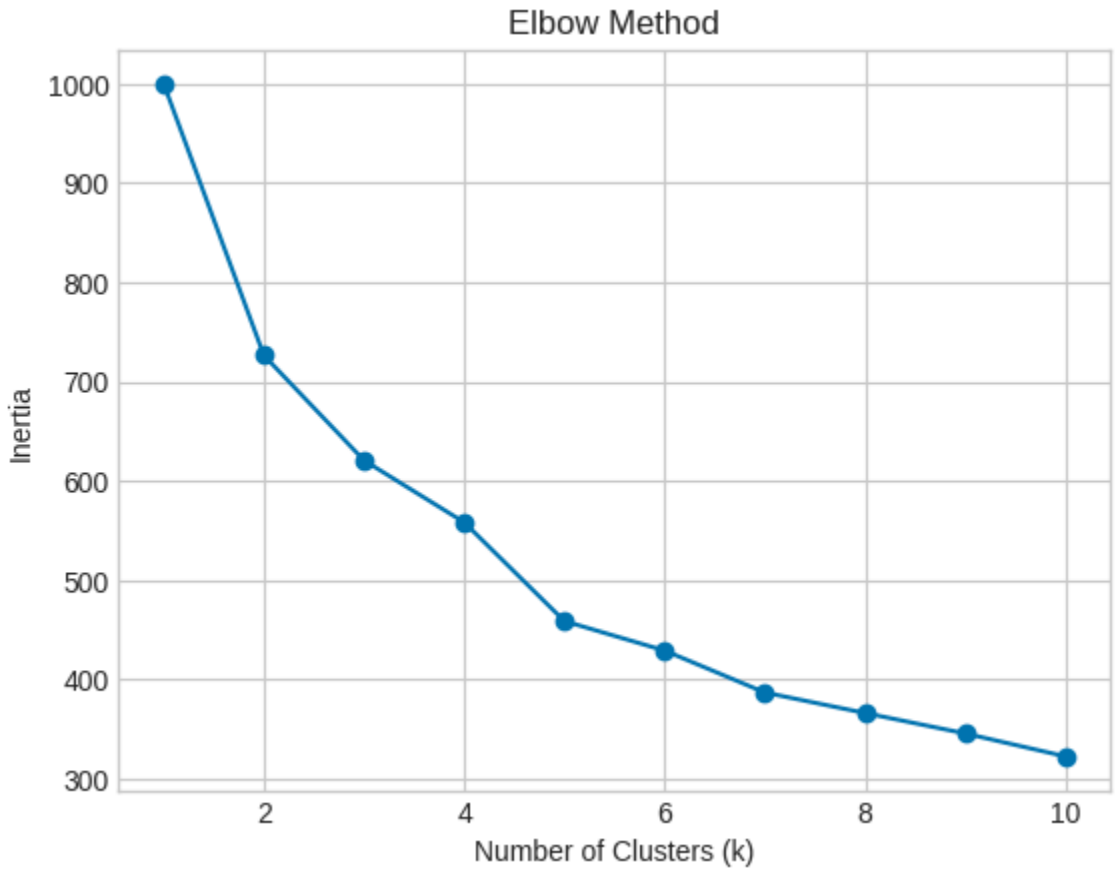

Fig 1. Graph of elbow method implementation for different k values

From the plotted elbow method graph, we can observe that there are 2 potential k values, that is k=2 or k=5. Both these points show a significant drop in inertia as compared to their respective previous k values (k=1 and k=4). They also show a smaller drop in inertia compared to their respective subsequent k value (k=3 and k=6).

But it is observed that k=5 features a gentler decrease to its subsequent k value (k=6), potentially signifying that k=5 is a more optimal choice than k=2 based on this graph.

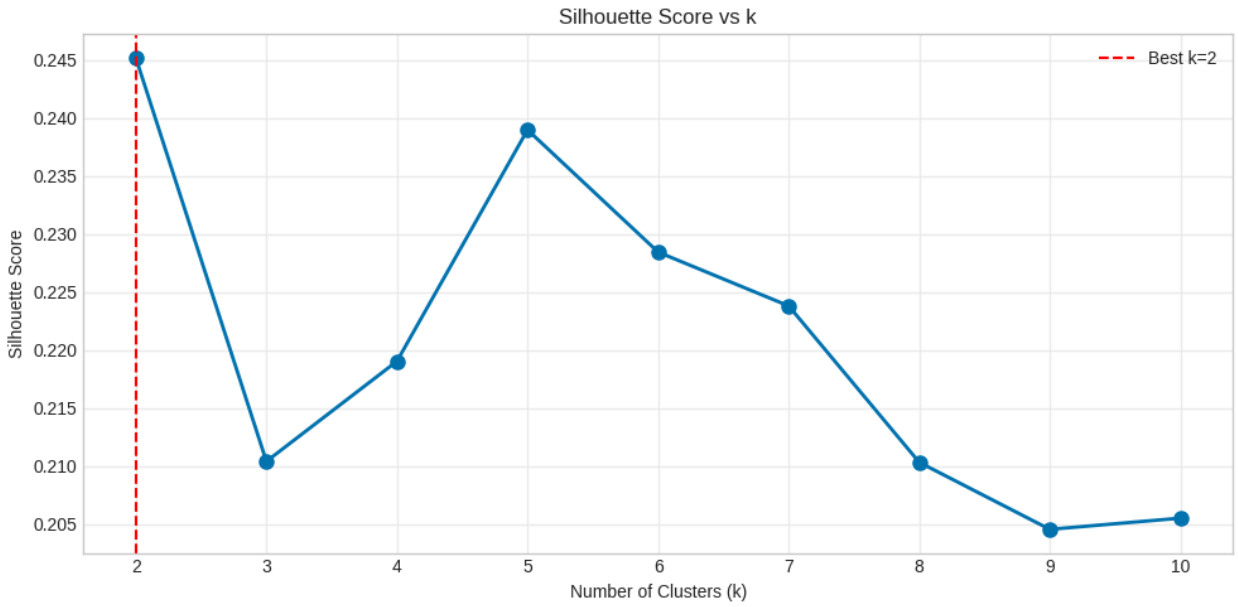

Fig 2. Graph of silhouette score for different k values

In addition, we can observe from this silhouette analysis graph, that k=2 is chosen as the optimal k value, with a score of 0.245. However, k=5 offers a good score of 0.2390, and does not compromise too much on silhouette score. k=5 score also is significantly higher than the other k values.

Ultimately, the choice of k-value for the model comes down to k=2 or k=5.

k=5 was chosen as it is beneficial in this context to further segment the market down to more granular sub groups, allowing for greater marketing and promotional targeting of these subgroups.

K-means model findings

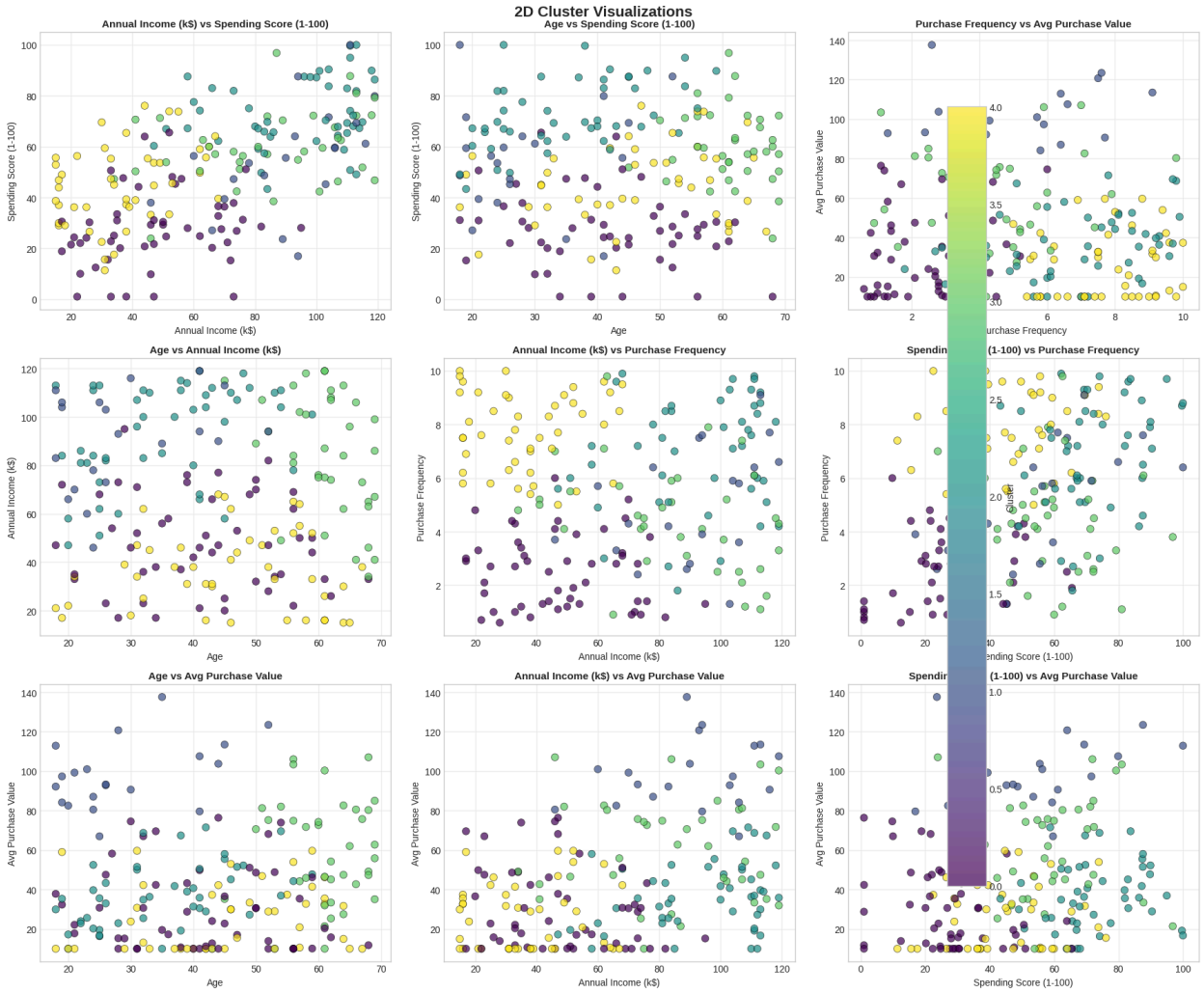

Fig 3. 2D plots of different feature combinations

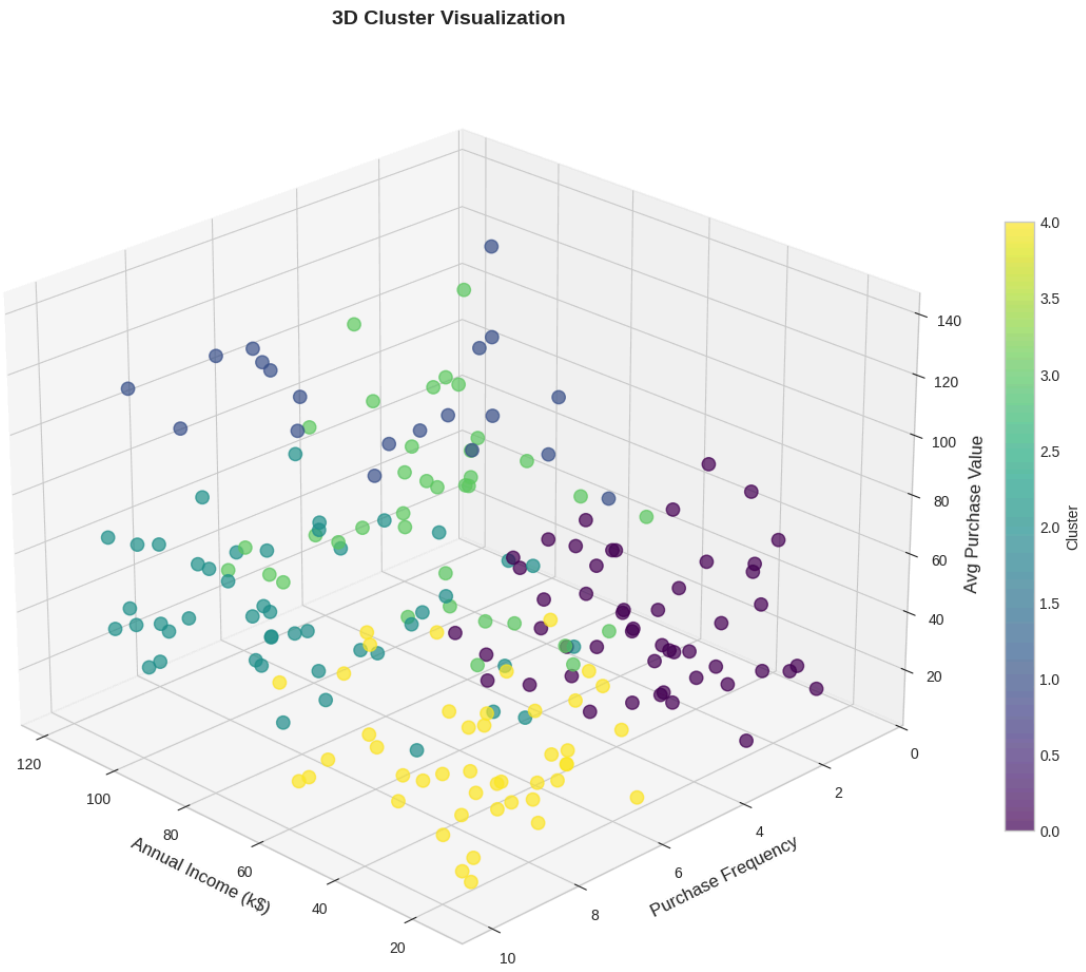

Fig 4. 3D plots of key features - Annual Income, Purchase Frequency, Avg Purchase Value

While it can be said that the graphs present statistical imperfection, with clusters overlapping each other, choosing k=5 provides better segmentation and capturing of different sub groups.

From the graphs, there are 5 clusters that could be interpreted:

1. Lower income customers with frequent low spendings (Yellow)
2. Lower income customers with infrequent medium spendings (Purple)
3. Middle income customers with average spendings and frequency, likely representing the average customer (Green)
4. High income customers who purchase frequently but with low spendings per transactions (Turquoise)
5. High income customers who purchase less frequently but are high spenders per transaction, likely representing VIP customers (Dark Blue)

The marketing strategy for this retail business can correctly identify and strategise marketing campaigns for future customers based on the cluster they are categorised into.