****

**VIRGINIA COMMONWEALTH UNIVERSITY**

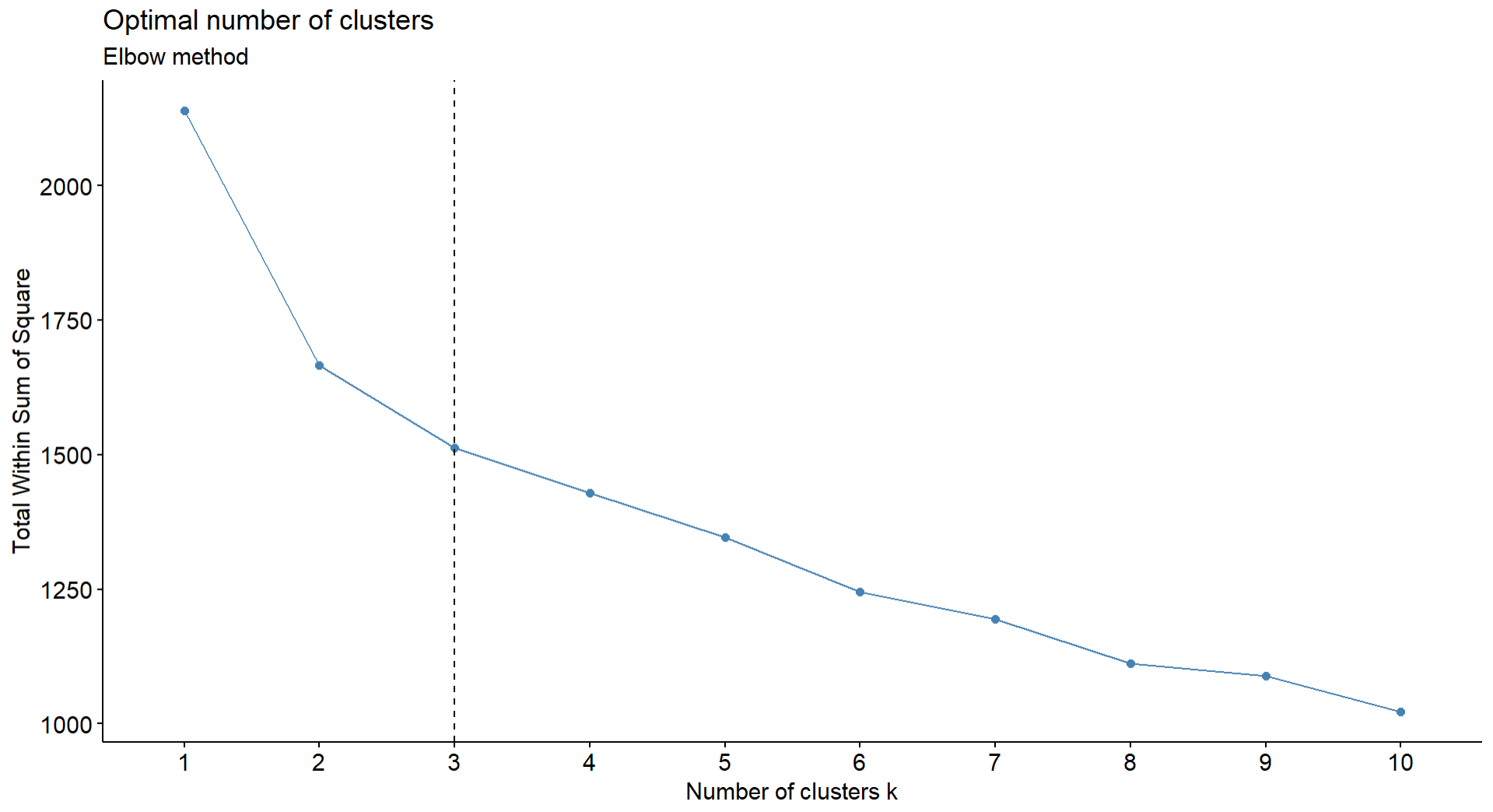
**Statistical analysis and modelling (SCMA 632)**

**A4b- Conduct Cluster Analysis to characterize respondents based on background variables**

**Rohan Arpit Dungdung**

**V01108247**

**Date of Submission: 15-07-2024**



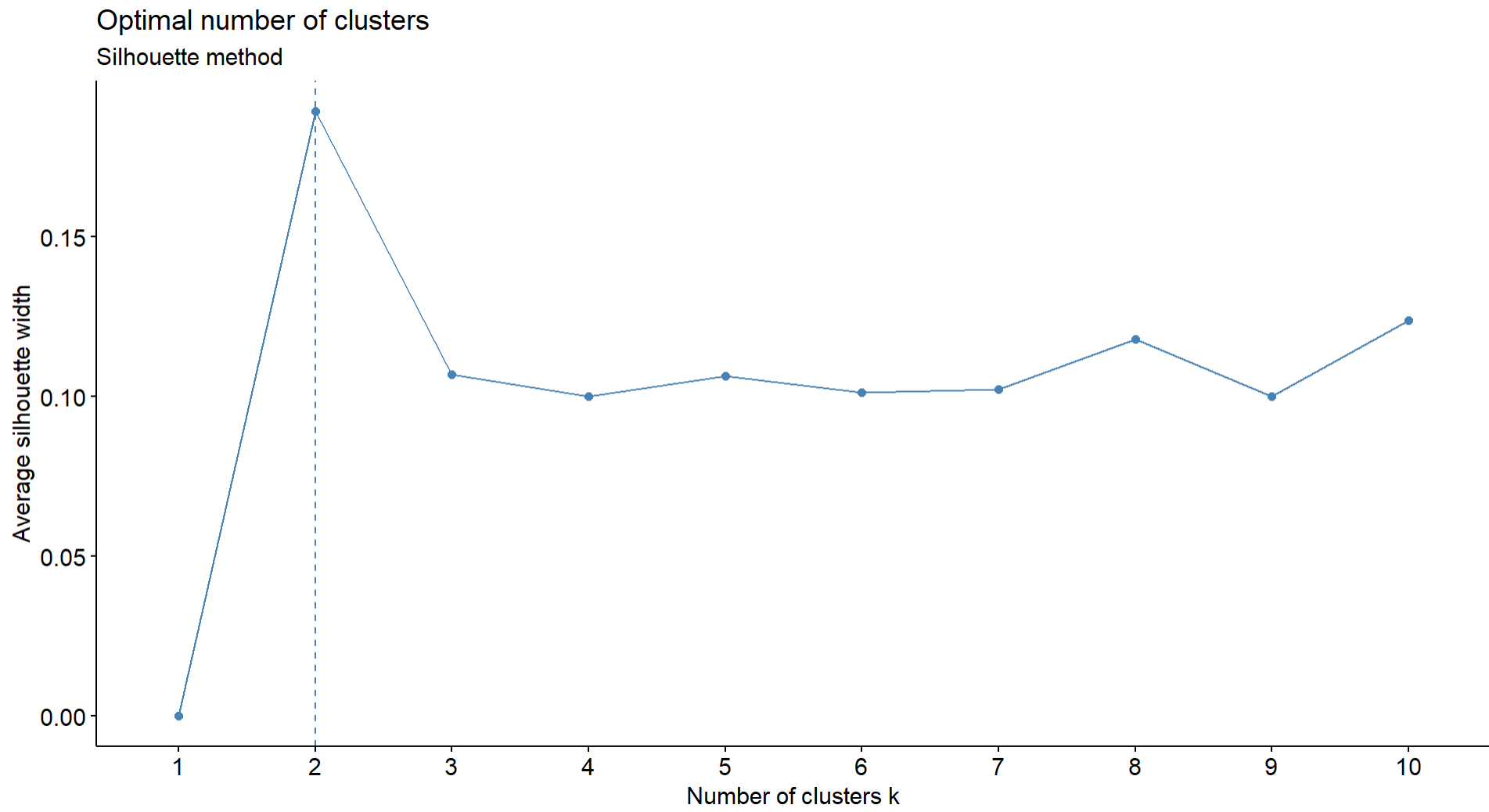
Interpretation

The WSS measures the compactness of the clusters. It is calculated as the sum of squared distances between each point and the centroid of its cluster. A lower WSS indicates more compact clusters.

The "elbow" in the plot represents the point where the WSS begins to level off. This point suggests the optimal number of clusters, as adding more clusters beyond this point results in diminishing returns in reducing WSS.

In this plot, the elbow appears around k = 3, where the curve starts to flatten. This indicates that three clusters are optimal for this dataset, balancing compactness (low WSS) with simplicity (fewer clusters).

The number of clusters chosen is critical in ensuring that the clusters are distinct enough to provide meaningful groupings while avoiding overfitting or overly complex models.



Interpretation

The average silhouette width measures the cohesion within clusters and separation between clusters. A value close to 1 indicates that data points are well matched to their own cluster and poorly matched to neighboring clusters. A value near 0 indicates overlapping clusters, and negative values indicate that points may be assigned to the wrong cluster.

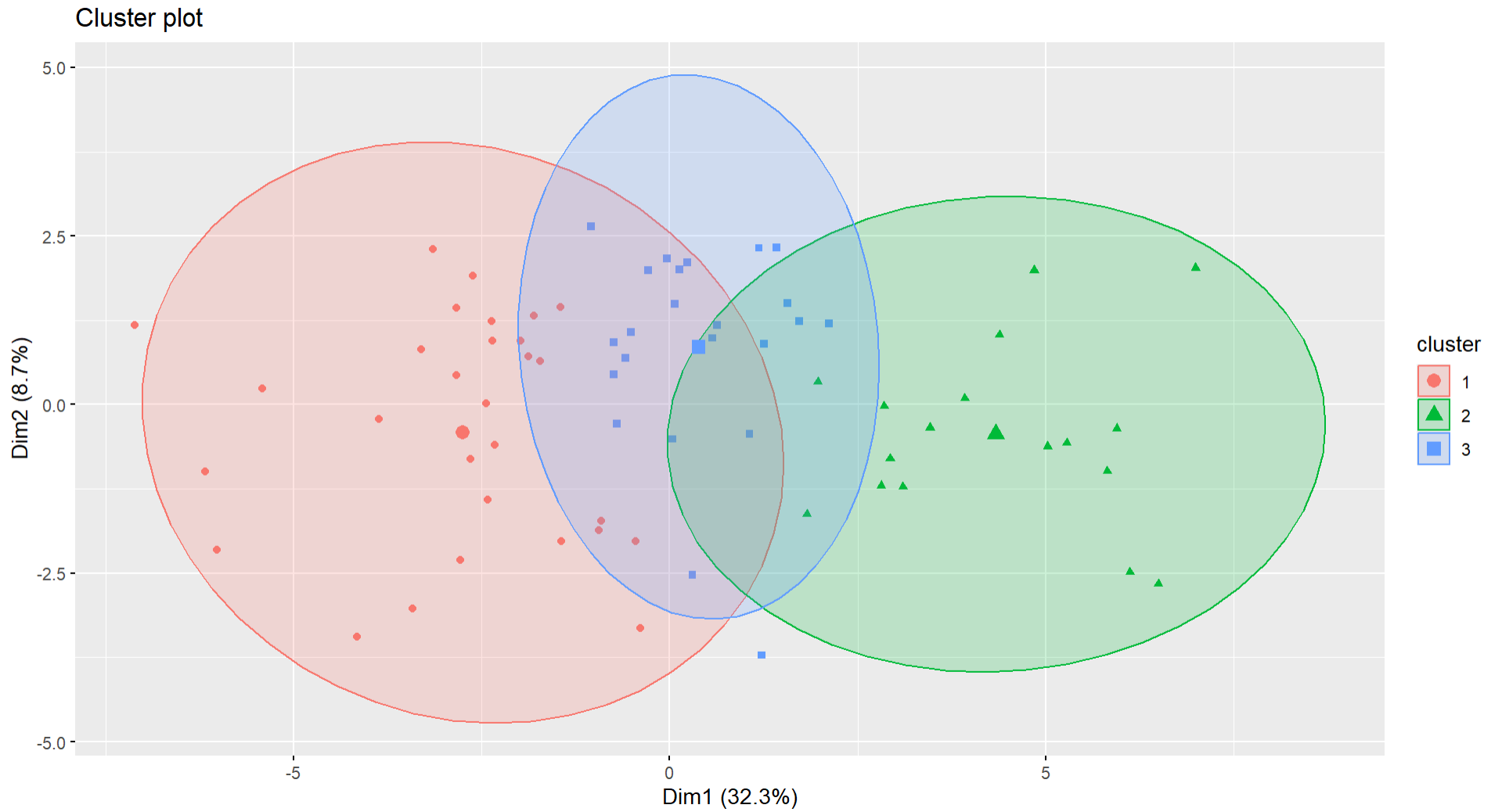
The plot shows the average silhouette width for different numbers of clusters (k). The optimal number of clusters is typically where the silhouette width is highest, as this indicates the best separation and cohesion.

In this case, the highest silhouette width occurs at k = 2, suggesting that two clusters provide the best-defined grouping for this dataset.

If the silhouette width for k = 2 is substantially higher than for other values of k, it may indicate that the dataset naturally divides into two distinct groups. These groups could correspond to fundamentally different respondent characteristics, preferences, or behaviors.

Choosing k = 2 could lead to more cohesive clusters, which are easier to interpret and analyze. However, the decision should also consider the context and the specific needs of the analysis.

The choice between k = 2 and k = 3 (or more) clusters depends on the balance between interpretability, the natural grouping of the data, and the specific goals of the analysis. While the Silhouette Method provides a useful metric for assessing clustering quality, practical considerations and domain knowledge should also play a role in the final decision.



Interpretation

The three clusters are represented by different symbols and colors:

Cluster 1 (red circles): Represented on the left side of the plot.

Cluster 2 (green triangles): Located mostly on the right side.

Cluster 3 (blue squares): Positioned between Clusters 1 and 2.

The plot shows that Cluster 2 is relatively well-separated from the other two clusters, indicating a distinct group of observations. Clusters 1 and 3 have some overlap, which may suggest that they share similar characteristics or that the data points in these clusters are not as distinctly different from one another as they are from Cluster 2.

Dim1 (32.3%): This axis captures 32.3% of the variance in the data. The distribution along this dimension helps distinguish Cluster 2 from the other clusters, as Cluster 2 primarily occupies the positive side of Dim1.

Dim2 (8.7%): This axis captures 8.7% of the variance. There is less clear separation along Dim2 compared to Dim1, but it still helps differentiate the clusters, particularly Cluster 3 from Cluster 1.