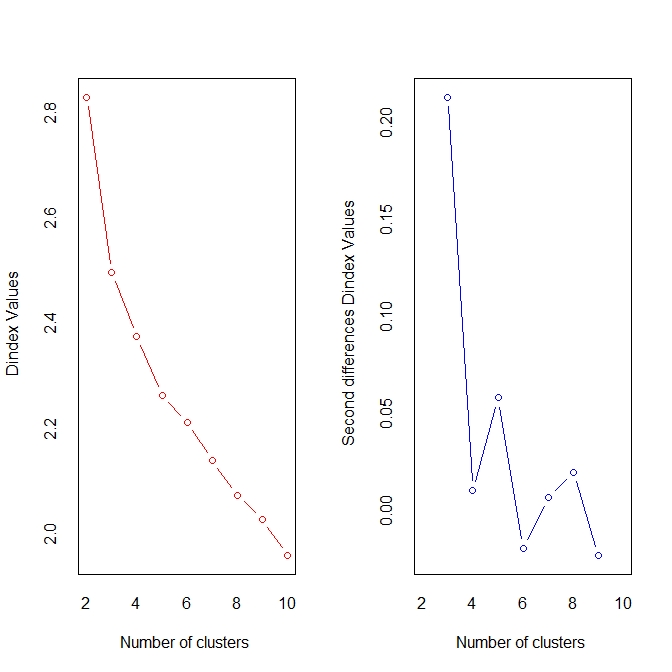
**Introduction**

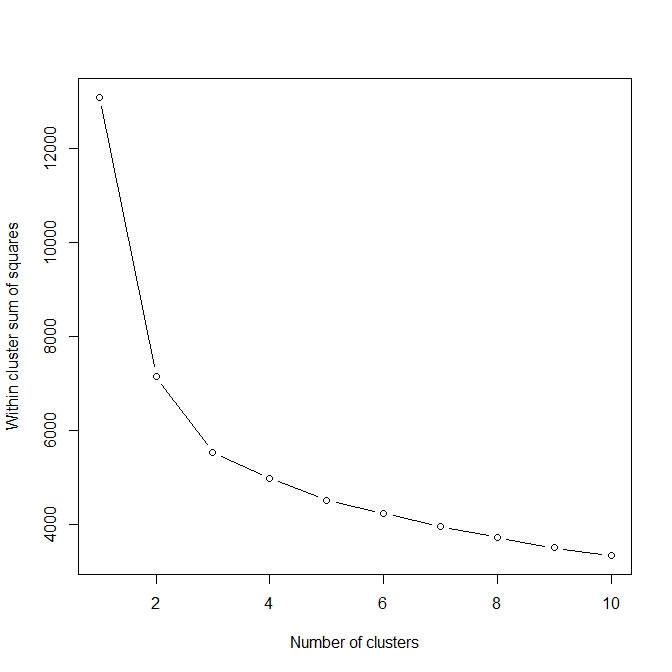
Clustering being an unsupervised machine learning technique used for exploring the structure of datasets by grouping similar data points. On the project, I analyzed the vehicle dataset provided using k-means clustering, with and without principal component analysis, with the aim of understanding the effects of dimensionality reduction on clustering analysis. The dataset had information of different cars that were used to classify the vehicle type without using the class variable.

**1st Subtask**

On the first subtask, I performed data cleaning followed by identifying and removing of the outliers on the dataset. That was followed by determining the optimal number of clusters using the specified four “automated” tools that included NBclust, Elbow, Gap statistics and silhouette methods.



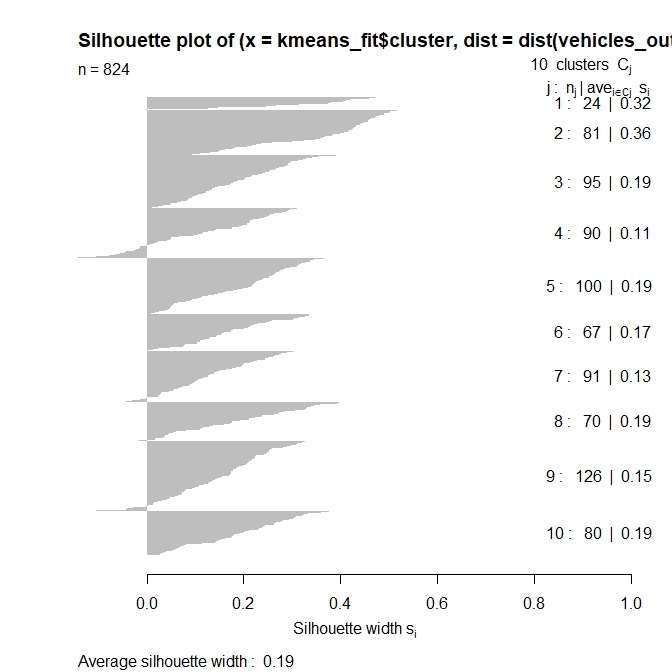
As shown by the visualization above, the NbClust results showed significant drop between 2 and 4 showing that the best number of cluster was three.



That is also evident when I used the elbow method as seen on the elbow plot above, the best number of clusters is between two and four clusters.

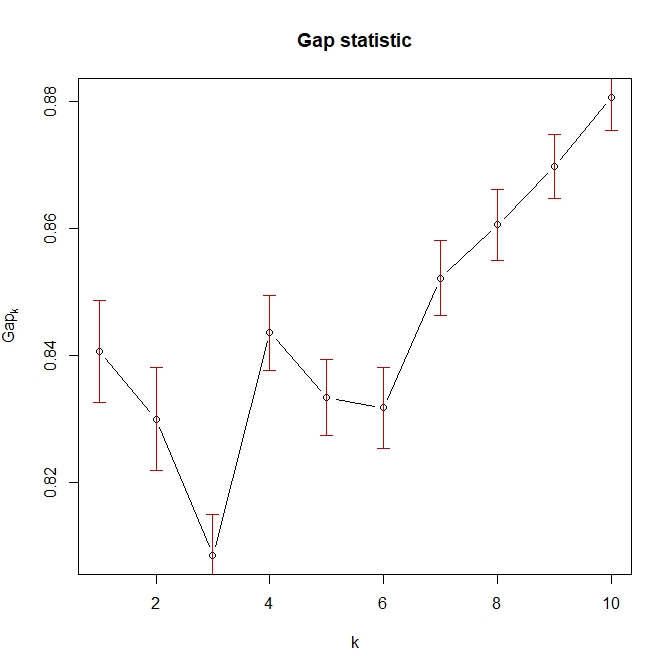
That resulted to me performing K-means clustering with k=3 on the cleaned dataset. The clustering results showed that the observations were separated into three distinct clusters. I then evaluated the clustering using internal evaluation metrics such as the between-cluster sum of squares (BSS) and within-cluster sum of squares (WSS) indices that displayed that the clustering was relatively good (Saji, 2021).

The between-cluster sum of squares value of 57.77922 for the k-means clustering analysis, while the within-cluster sum of squares value was 3338.338 and the total sum of squares value was 3396.117. The BSS/TSS ratio was 0.01701332 that was relatively low showing the cluster analysis was not optimal and there was some overlap between the clusters. Furthermore the average silhouette width was 0.1867381 showing the quality of the obtained clusters was not very high. The silhouette plot below shows that some of the clusters have negative silhouette scores, thus suggesting that those clusters are not well-separated and the data points within those clusters may belong to other clusters (Shutaywi & Kachouie, 2021).



**2nd Subtask**

On the second subtask, I performed PCA on the cleaned dataset to reduce the input dimensionality. The PCA analysis showed that the first four principal components (PCs) explained 92.78% of the variance in the data. Having selected the first four PCs and created a new transformed dataset. Next, I applied the same four "automated" tools as before to the new pca-based dataset to determine the optimal number of clusters. The NbClust and elbow method highlighted that k=2 was still the best choice but the gap statistic and silhouette method suggested k=3. I decided to use k=3 for the final k-means clustering analysis.



After performing the k-means clustering on the new pca-based dataset with k=3 the results showed that the observations were again separated into three distinct clusters with sizes 548, 8 and 290. It was that evident there were 2 clusters that had highly negative or highly positive values for certain Principal Components, hence indicated that there were some features in the data that strongly separate the two groups from the rest of the data (Google Developers, 2019). The BSS/TSS ratio was 0.518457 that indicated only a small proportion of the data was not explained by the principal components I retained.