SDS 3786 Lab 7: Le regroupement

Professor: Patrick Boily

Anthony Le, John Bute, Kelly Gao

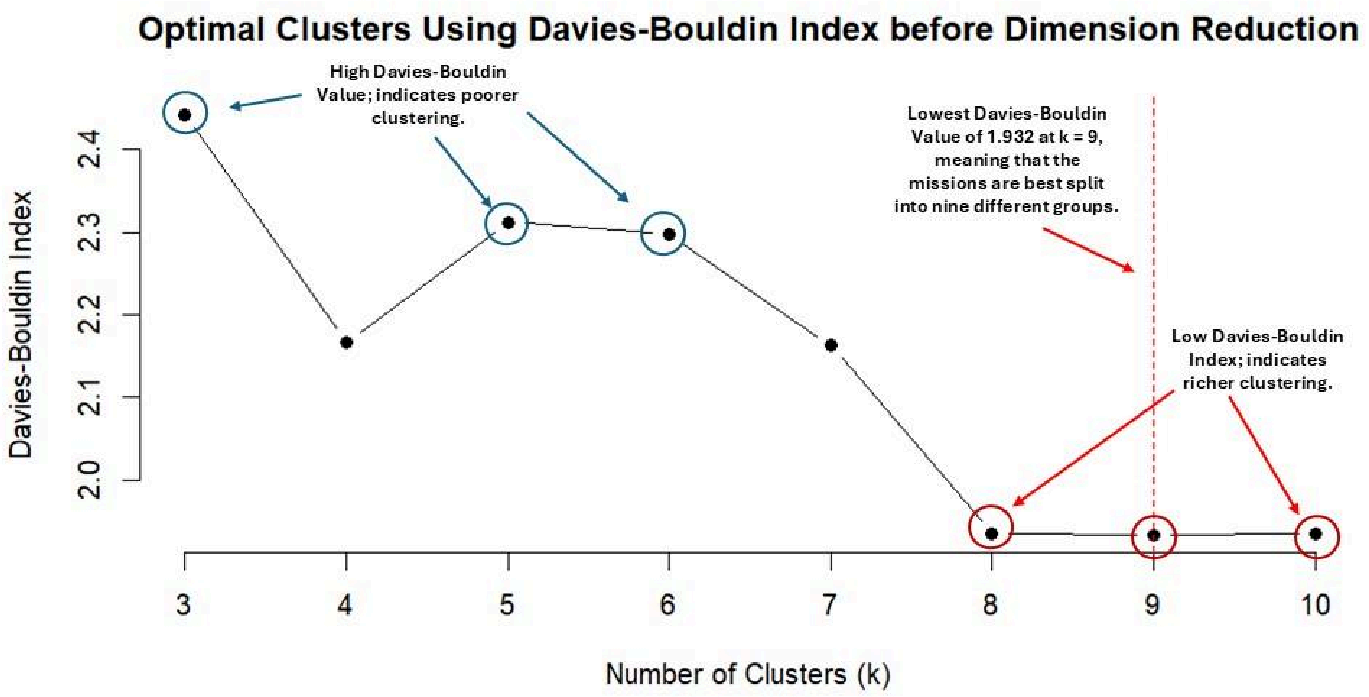
Introduction

This report provides a comprehensive analysis of the similarities between each embassy through the use of methods of unsupervised machine learning. Through the use of clustering algorithms, several indexes as well as various visualizations on an aggregated dataset based on each mission, we have identified several missions with strong similarities, which can allow us to posit the reason for these similarities, whether it is geographical or political.

Results

Through our observations, we were able to identify numerous embassies that have a perfect similarity score among each region. Following our investigation, we have determined that a collection of missions are true pairs among each other, meaning that each region has both a similar intake and activity to one another. Though this may be due to certain calculation or processing errors, we will nonetheless reveal the list of embassies that are identified as true pairs among each other.

Missions identified as true pairs among each other
Ammertuma, Atarillo, Bigowon, Bonaira, Borgen, Bosphorus, Cebu, Chach, Charles Town, Dun Eideann, Freetown, Garonne, Groniet, Huidobro, Kasim, Keupenhavn, Malake, Malkajgiri, Masqat, Pressburg, Puranupakorn, Qudaa, San Martin, Sao Manuelm Schduagert, Sitka, Tartu, Usumbara, Wilna, Yamai, eGoli.



Optimal Clusters Using Davies-Bouldin Index before Dimension Reduction

High Davies-Bouldin Value; indicates poorer clustering.

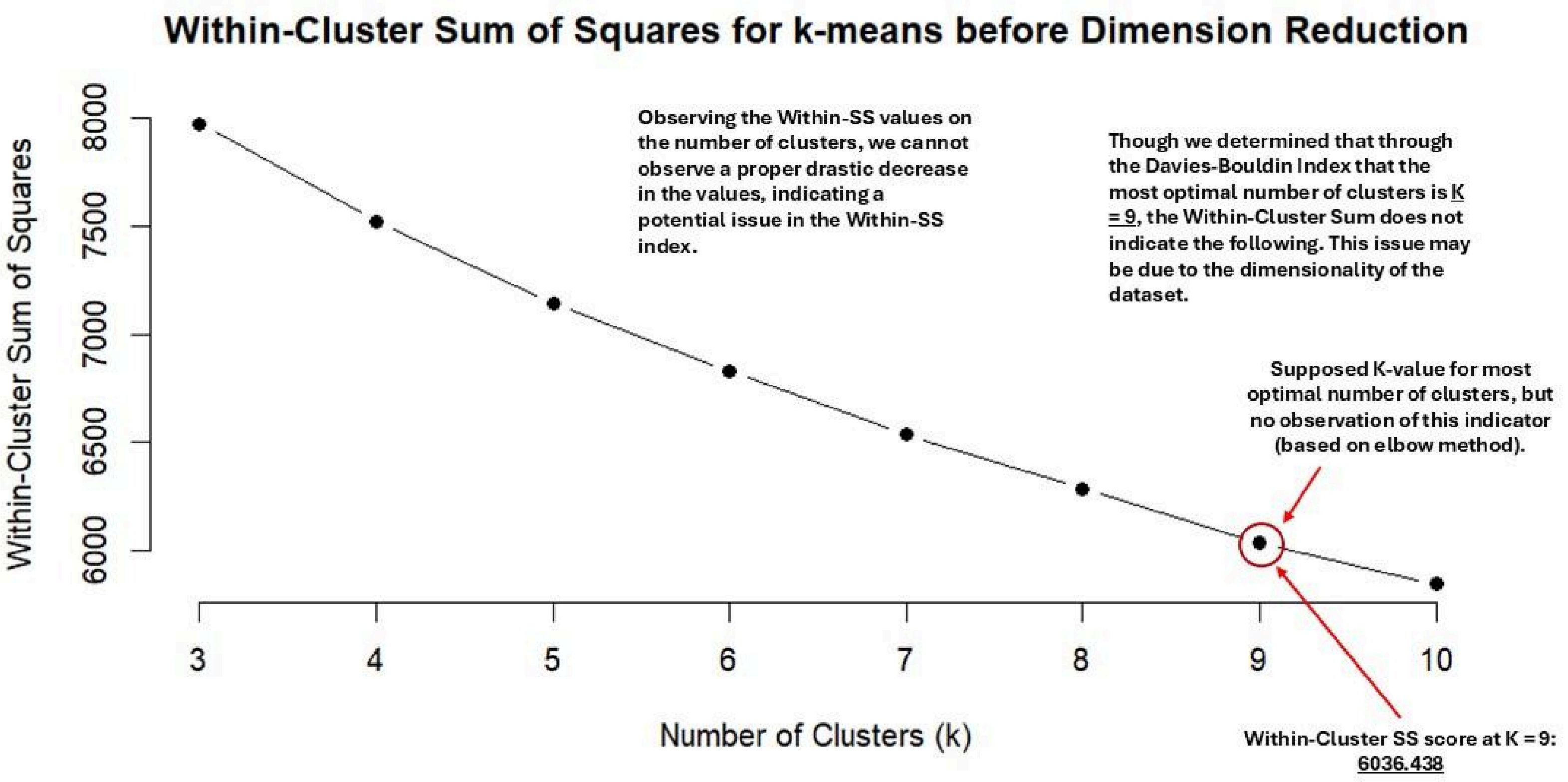
Lowest Davies-Bouldin Value of 1.932 at k = 9, meaning that the missions are best split into nine different groups.

Low Davies-Bouldin Index; indicates richer clustering.

Observing the graph which depicts the Davies-Bouldin index against the number of clusters, the initial k-values in the graph are seen to have a higher value with less clusters. This indicates that the lower number of clusters would provide us with a poorer clustering, as they are less compact and would result in a higher overlap. This also tells us that the missions within the dataset are not entirely like one another and are more unique overall.

As the number of clusters increase, we can see that the Index begins to decrease similarly, indicating an increase in richness among the clusters. This also tells us that the missions within the dataset are more unique and diversified in terms of their overall intake and spent time.

Finally at K = 9, we can observe the lowest index value of 1.932, positing that using nine clusters would provide us with the accurate clustering possible.

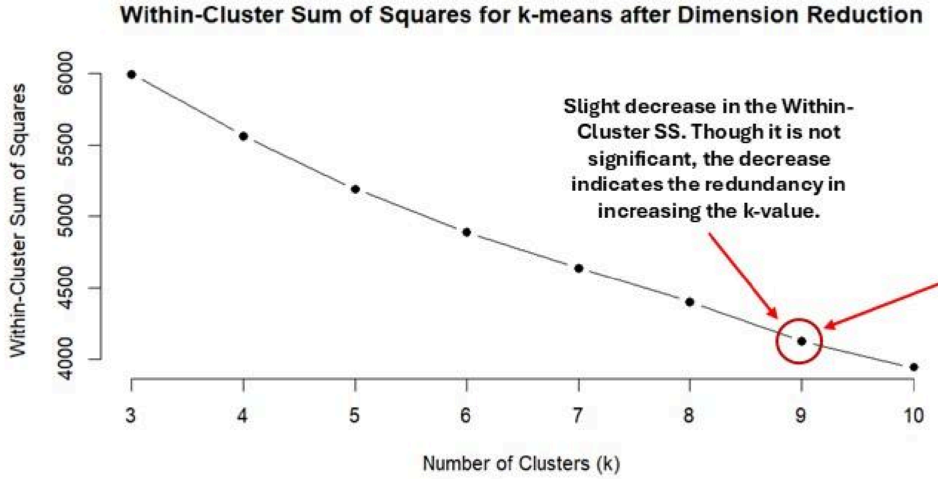


Observing the Within-SS values on the number of clusters, we cannot observe a proper drastic decrease in the values, indicating a potential issue in the Within-SS index.

Though we determined that through the Davies-Bouldin Index that the most optimal number of clusters is <u>K = 9</u>, the Within-Cluster Sum does not indicate the following. This issue may be due to the dimensionality of the dataset.

Supposed K-value for most optimal number of clusters, but no observation of this indicator (based on elbow method).

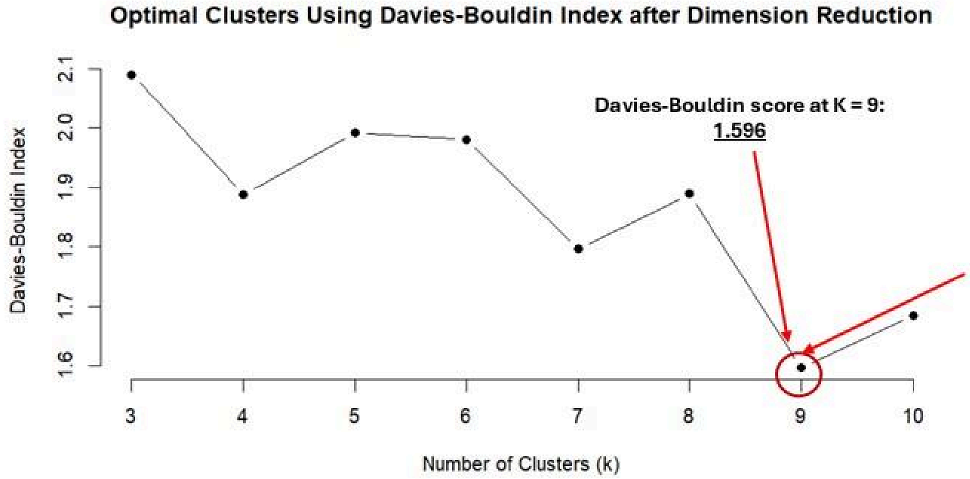
Within-Cluster SS score at K = 9: <u>6036.438</u>



Slight decrease in the Within-Cluster SS. Though it is not significant, the decrease indicates the redundancy in increasing the k-value.

Within-Cluster SS score at K = 9: 4130.369

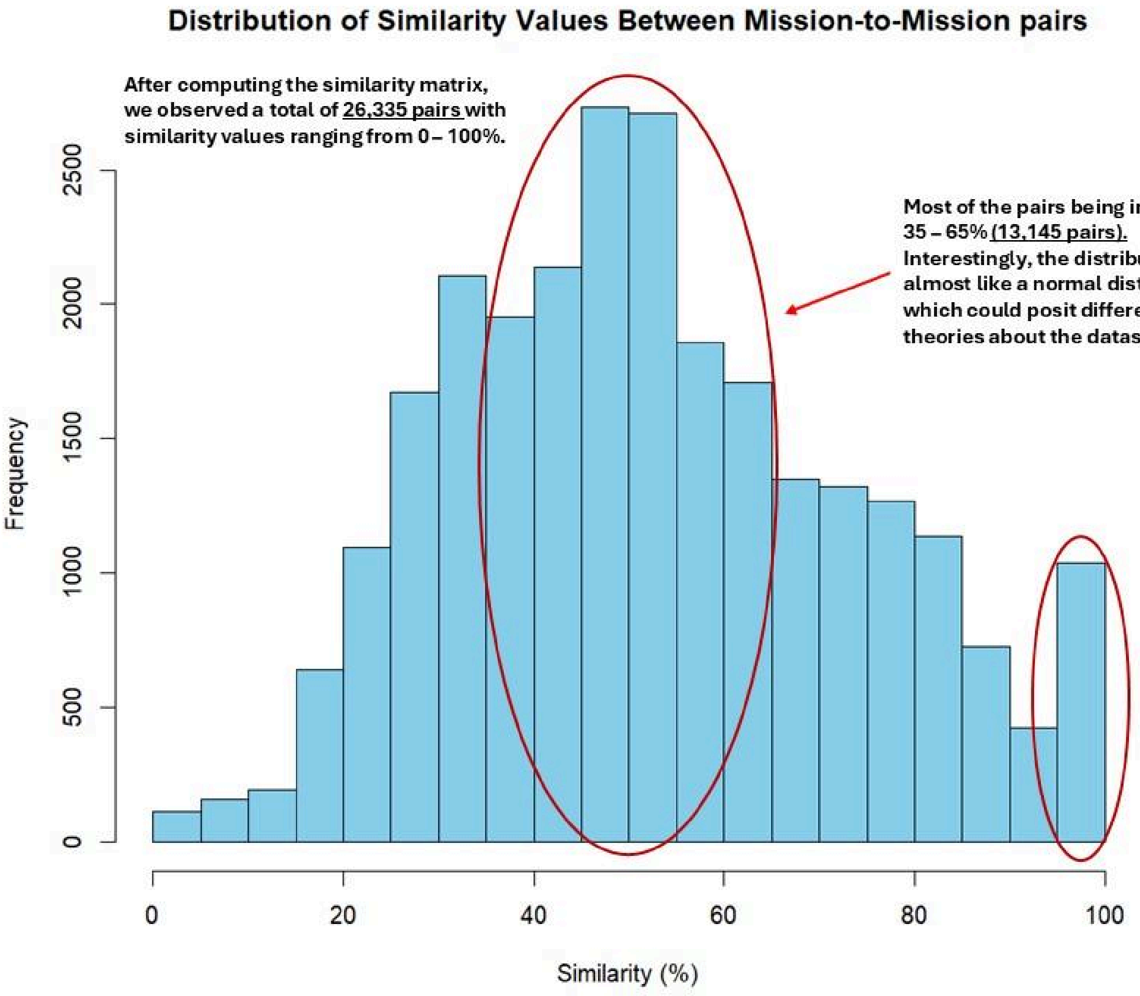
After reducing the dimensions of the dataset by performing a Principal Component Analysis, we viewed stronger evidence with the Davies-Bouldin and Within-SS indexes that concluded the optimal number of clusters for clustering to be k = 9. This was done after retaining principal components that explained for 80% of the data's variability.



Davies-Bouldin score at K = 9: 1.596

Massive drop in Davies-Bouldin index seen at k = 9. Following the dimension reduction, this enforces our previous belief that K = 9 is the optimal number of clusters for clustering.

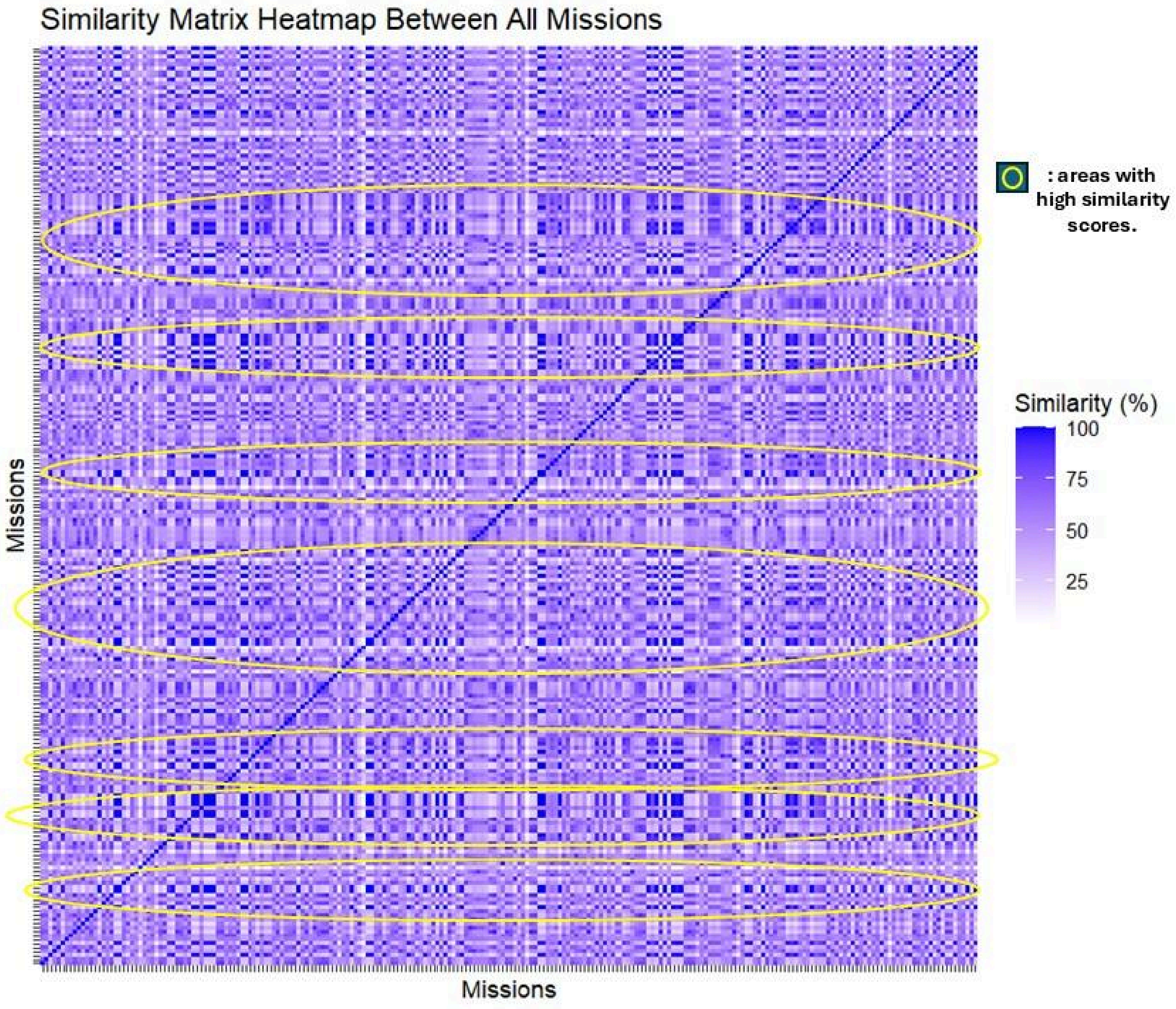
After developing an algorithm to select a set of features and a random clustering algorithm, we additionally implemented a similarity matrix to account for the pairs of missions within the same cluster. This similarity matrix was computed through 500 runs of the original algorithm.



After computing the similarity matrix, we observed a total of 26,335 pairs with similarity values ranging from 0 – 100%.

Most of the pairs being in between 35 – 65% (13,145 pairs). Interestingly, the distribution is almost like a normal distribution, which could posit different theories about the dataset overall.

More Importantly, we can observe a strong total number of pairs that have a similarity value between 95 – 100%. After calculating the similarity matrix, we determined that over 1,047 mission-pairs have a similarity value that is equivalent or very close to being true pairs. This observation of true pairs indicates several missions that have an almost similar or exact number of intake and spent time on activities, which could be due to region or political similarities.



(Unfortunately, due to number of missions within the dataset, this heatmap has omitted all mission names)

After creating the similarity matrix depicting the similarity values between all missions, we had developed a heatmap that showed several numerous groupings of strong similarity scores, indicating the presence of similar activity among the missions.

Though there is not a specific spot where the similarity values are grouped, we can observe that the areas with strong similarity scores are placed on the same rows and columns as each other. Additionally, we can observe that a lot of the strong similarity scores are almost replicated in other rows, which could indicate the existence of missions that appear to be like each other to more than 95%.