Clustering VC Funds based on Investments

Merging

The datasets **investmets.csv** and **funding_round.csv** were merged with on their unique_id with the SQL like pandas join to improve clarity of the clustering and to deepen the conjecture of the clusters. The merged dataset in the notebook is **inv_funrnd**

Pre-Processing

The **inv_funrnd** dataset had a lot of quality issues like duplicated columns, inept id's and impractical text data.

A list containing the columns with data assessment issues was created and used to generate a new standard merged dataset without data assessment problems. [**std_mrg**]

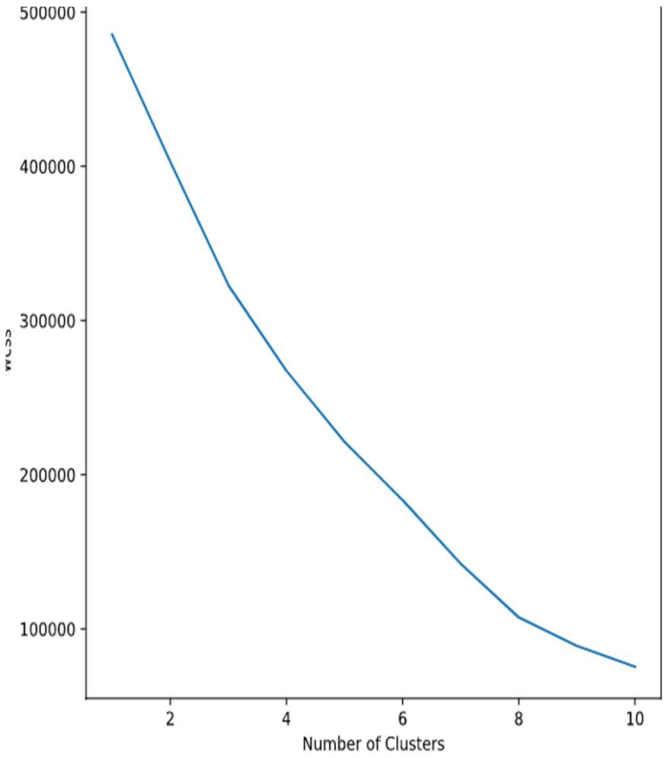
Since the clustering algorithm can only work with numerical data, the same was extracted from the [**std_mrg**] to implement the clustering on.

The resulting dataset [**c_std_mrg**] was scaled using sklearn standard scaler to improve the clustering algorithm performance.

Elbow Method for Optimal K || Kmeans Clustering (no-pca)

The elbow method is used to mathematically determine the optimum number of clusters

We graph the relationship between the number of clusters and Within Cluster Sum of Squares (WCSS) then we select the number of clusters where the change in WCSS begins to level off (elbow method).



From the image above, the optimum number of clusters for the data is 8.

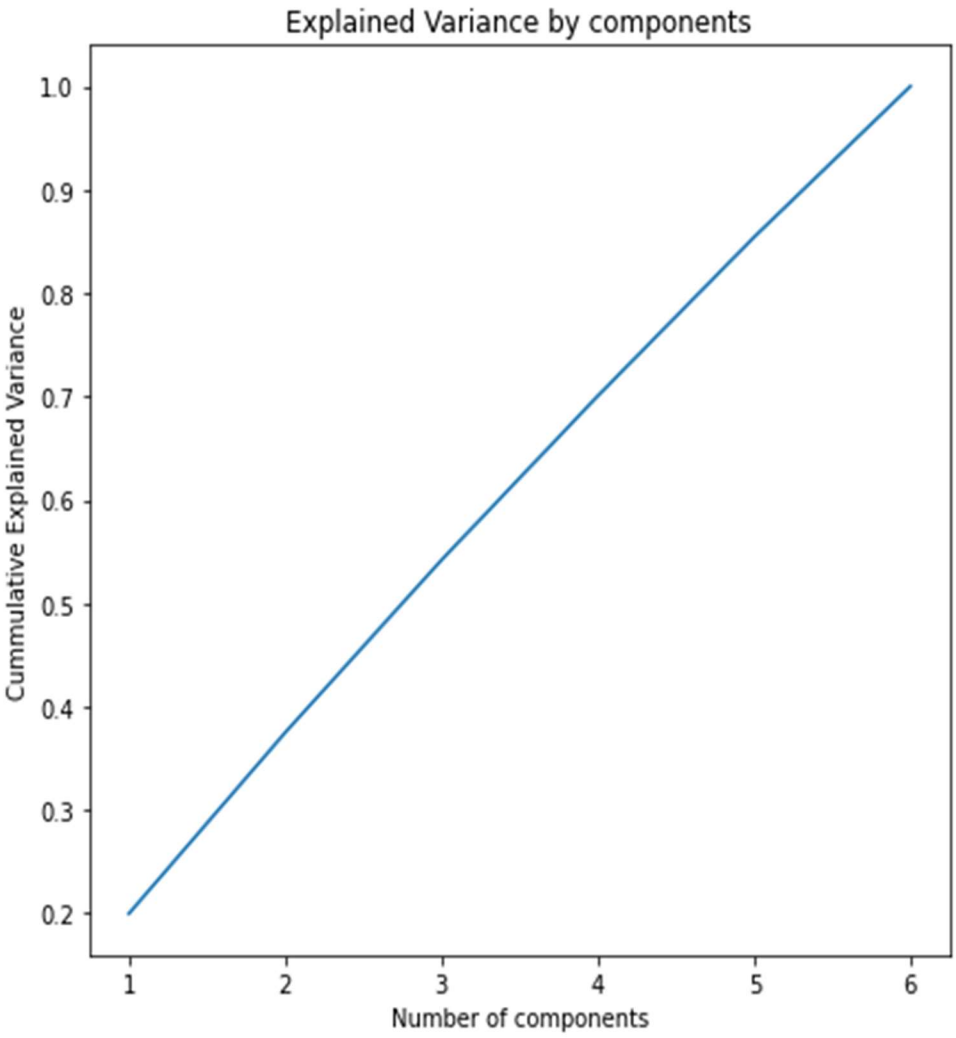
The K means clustering algorithm is read to be fitted to the dataset using the optimum number of clusters. A new column containing the cluster each data row belongs to is now added to the standard merged dataset [*std_mrg_kmeans*] for further segmentation analysis which would be helpful for new and existing startups to gauge their progress and success and make strategic moves to improve their chances of success. This would be more strategically useful than a supervised classification model.

Modest analysis was done on the clustered dataset and the following was inferred;

- Startups in the cluster **6** that were in their first funding rounds significantly higher raised funds than those in other clusters; the particular data points could be easily extracted for exploration.
- All the startups in cluster 5 and 6 when compared through their unique id in the **objects.csv** dataset all had excellent statuses; either done their **ipo, acquired or operating.**

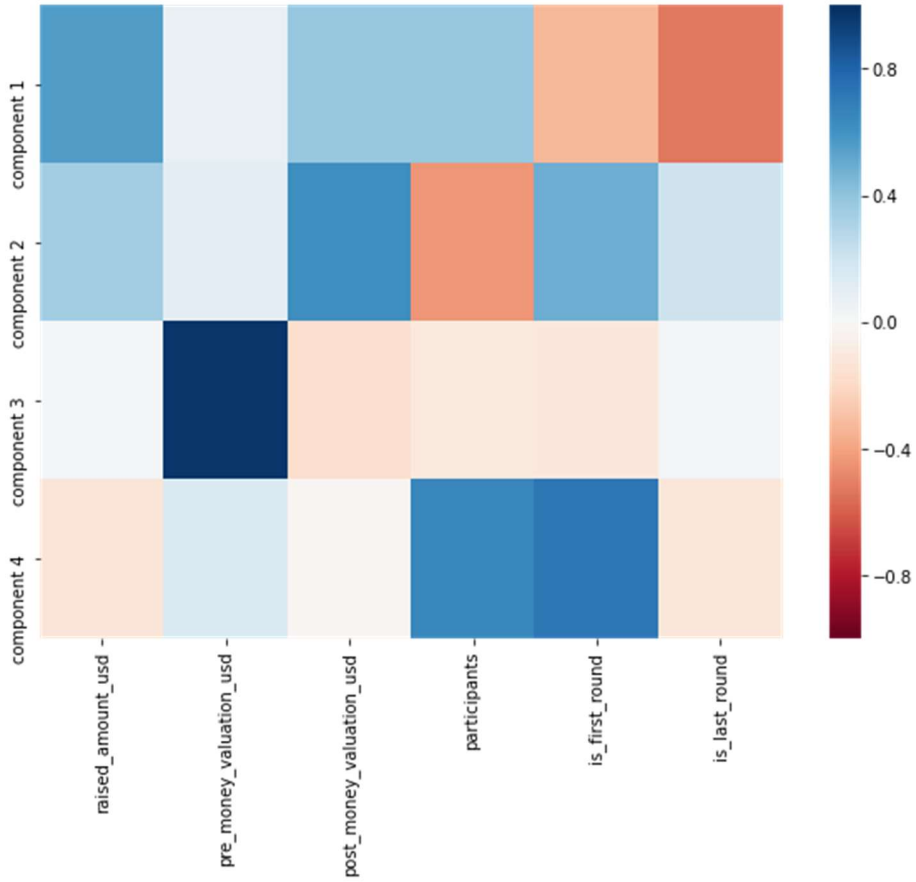
Kmeans Clustering (pca)

PCA (Principal Component Analysis), albeit complex in subtleties is simply used to reduce the dimensionality and to create a better milieu for visualizations.



The graph above [**explained variance graph in images folder**]; simply shows the explained variance of the components possible. E.g. (1component = 20% explained variance etc.)

In context with this particular project, the number of components was reduced to 4(over 70% explained variance), and to overcome the curse of interpretability, the visual heatmap below shows the correlation between the new components and the old.



The same process of clustering was followed with the first clustering with the same amount of optimal clusters and very similar results, hence any new segment visualization would be best understood with the help of the heatmap or rather be done on the first clustering results.

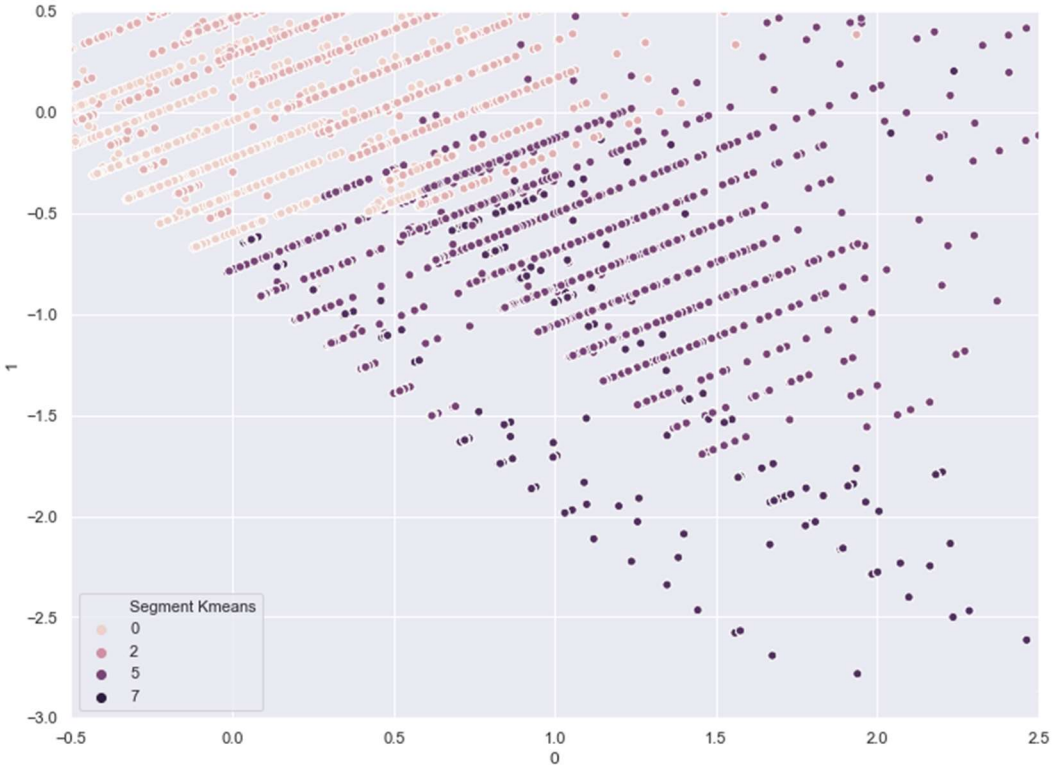
Segment Visuals

Theses visuals highlight the arrangement of the clusters.

Refer to heatmap to interpret components.

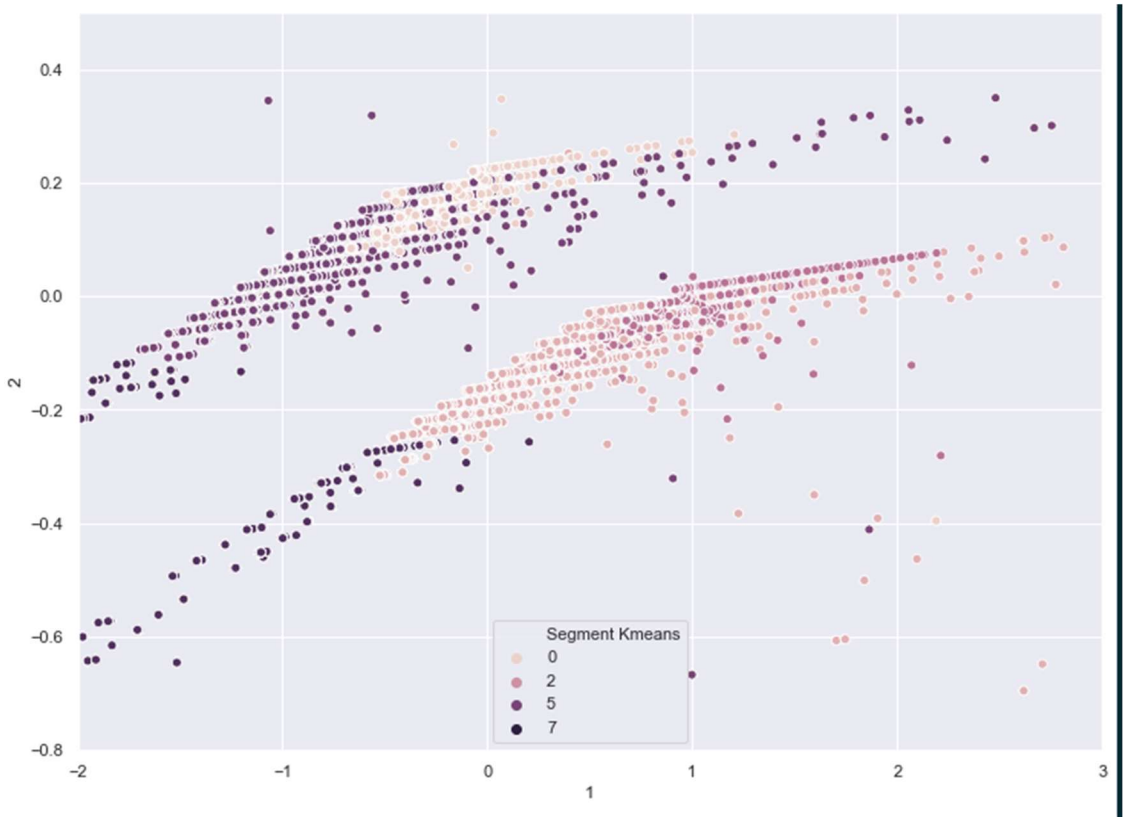
(Y-axis: component 2||x-axis: component 4)

Refer to heatmap to interpret components.



(Y-axis: component 3||x-axis: component 4)

Refer to heatmap to interpret components.



(Y-axis: component 2||x-axis: component 3)