KMeans Clustering

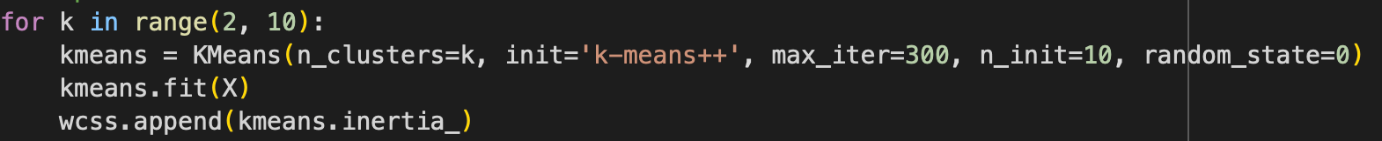
After all the data has been processed, we intend to use the whole dataset for clustering to get a better insight of the data. However, before we move on to the clustering method. We decided to use Elbow method to visualize the optimal n number of clusters:

**What is Elbow method?**

The elbow method is a technique used in clustering problems to determine the optimal number of k clusters in a dataset. It is mostly common use for KMeans clustering.

**In order to implement Elbow method:**

1. **Run the K-means algorithm** on the dataset for a range of values for k. For our implementation, we tried to run from 2 to 10 number of clusters

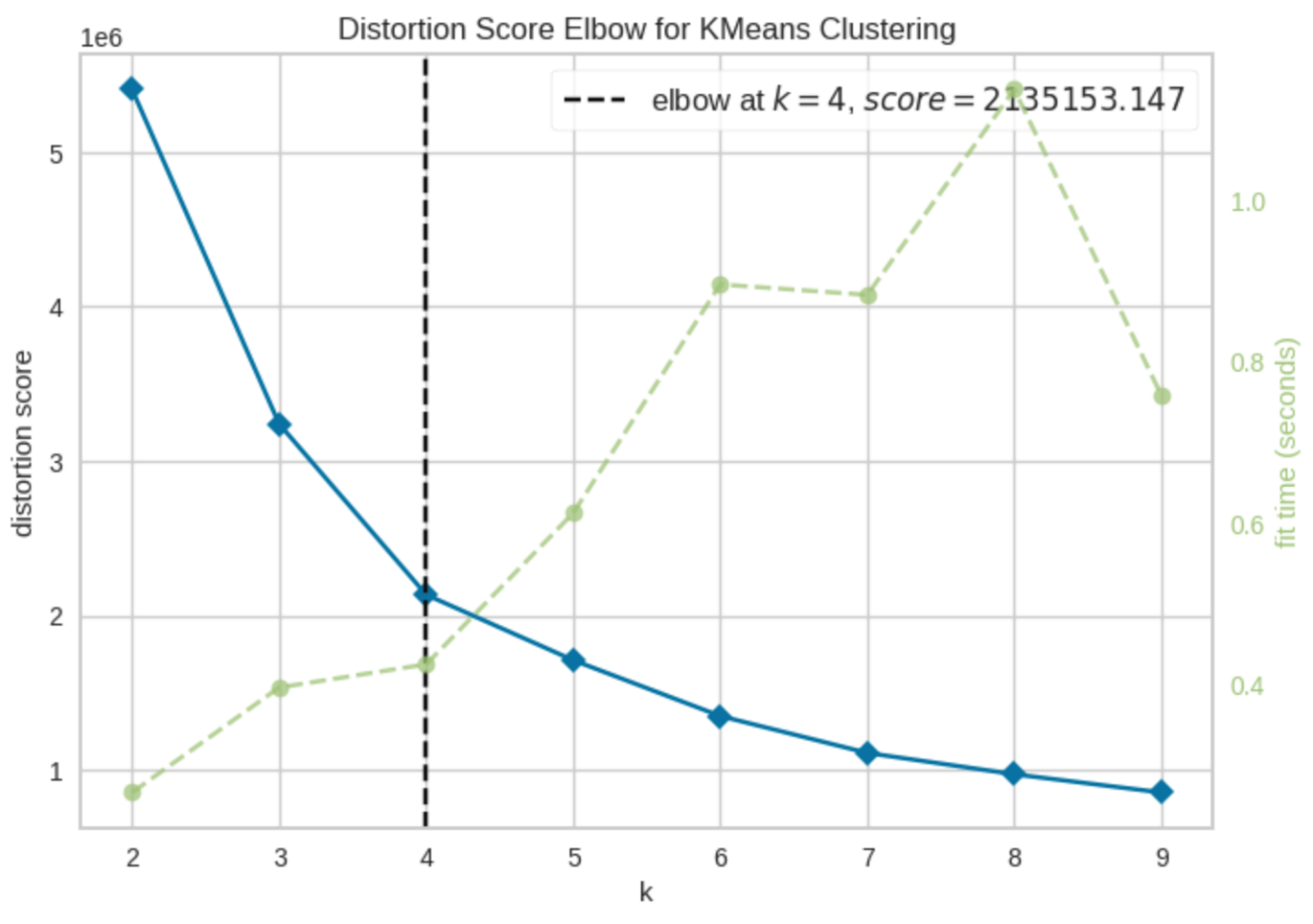


1. For each k we test, we **calculate the Within-Cluster Sum of Squares (WCSS)**, that is the metrics to measures the sum of the squared distances between each point in a cluster with its belonging centroid. WCSS will decrease as the number of clusters increase, since all data point are more tightly to each group it has been clustered.
2. **Identify the "elbow" point**: The optimal number of clusters is often identified as the point where the curve starts to flatten out. This point is known as the "elbow" because it looks like the bend in an arm.

**Why the Elbow Method?**

* Before the elbow, adding more clusters significantly reduces WCSS, indicating that the clusters are becoming more compact.
* After the elbow, the reduction in WCSS diminishes, suggesting that adding more clusters doesn't significantly improve the model. Thus, the elbow point balances model complexity and fit quality

**Final result**



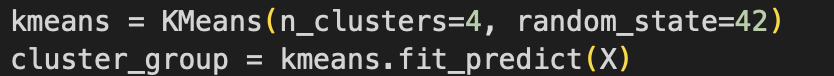
**Kmeans**

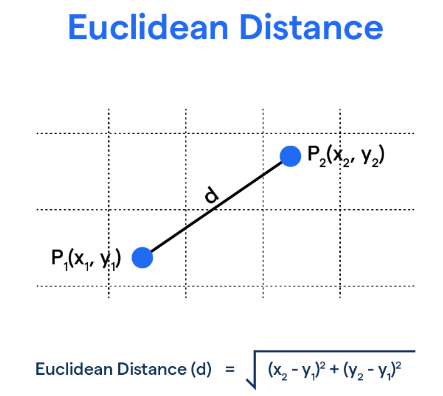
Now the optimal cluster has been identified which is 4. It is time for us to perform clustering with Kmeans.

K-means is a straightforward clustering algorithm used in unsupervised machine learning to predefined number of groupings k

**How K-means Works:**

1. **Choose the Number of Clusters (k)** which is 4



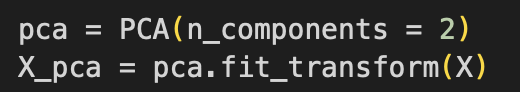
1. **Initialize Centroids in random**
2. **Assign Points to the Nearest Centroid**:
   * Each data point is assigned to the nearest centroid using the Euclidean distance for our implementation (default parameter)
3. **Recalculate Centroids**:
   * Once all points are assigned to clusters, the algorithm recalculates the centroids as the mean of all points in each cluster.
4. **Repeat Assignment and Recalculation** until the centroids no longer move significantly, indicating that the algorithm has converged.

Visualizing the output

Since our dataset has up to 8 features, can be consider as the eight dimensional data which is no tool or any other method for us to visualize the 8th dimensional data.

Luckily there is one technique called dimensional reduction, we use that to reduce 8th dimensional into 2 dimensional data for better visualize. The method we use is Principal Component Analysis(PCA), mostly use in data science and machine learning. Its main purposes are:

1. **Reducing Dimensionality**
2. **Feature Extraction**: to reduce noise and improve model performance, as well as computing time
3. **Data Visualization**: for our visualization purposes
4. **Dealing with Multicollinearity**



After determining that 4 clusters were optimal, K-Means clustering was performed using 4 clusters. To visualize the clusters, Principal Component Analysis (PCA) was used to reduce the dimensionality of the dataset to 2D, making it easier to plot and interpret.



In this, we are running Kmeans clustering using 4 clusters as the result for the Elbow method determined the number of clusters to be 4.

PCA is used to reduce the dimensions of data to 2D representation.

X\_pca = pca.fit\_transform(X) reduces the data to 2 dimensions for visualization.

plt.scatter(X\_pca[:, 0], X\_pca[:, 1], c=cluster\_group, cmap='viridis',): Plots the PCA-transformed data with colours corresponding to cluster labels.

plt.scatter(centers\_pca[:, 0], centers\_pca[:, 1], s=300, c='red', marker='x', label='Centroids'): Adds cluster centroids to the plot and determines the colour of the visualisation.

plt.xlabel and ylabel along with title indicates the labels and title respectively.



**Explanation of Visualization:**

* **PCA Transformation:** The dataset is transformed into two principal components using PCA, which simplifies the high-dimensional data into two dimensions while retaining most of the variance.
* **Cluster Representation:** Each point in the scatter plot represents a cluster. The scatter plot shows distinct groups of data points, each color-coded according to the cluster assigned by the K-Means algorithm.
* **Interpretation:** The visualization confirms the separation between clusters, indicating that the K-Means algorithm successfully grouped similar data points. If the clusters are well-separated, it implies that the features used were effective in differentiating the customer segments.
* **Insights:** The clustering revealed distinct segments within the data, which could be further analyzed for actionable insights, such as identifying key customer segments or predicting customer behavior.
* By labeling and analyzing the clusters, several important insights were derived. These include the identification of key customer segments and potential strategies for targeted marketing or customer retention.

**INSIGHTS FOR OUR STAKEHOLDERS:**

Clustering in a telecommunication company can help a company separate and determine the trends and insights of their customer base.

Clusters can help separate customers based on titles such as High value customers who decided to stay with the company whereas, it can also help to determine customers who are likely to churn based on the prices or high costs. The clusters can be compared with various fields.

It can also help streamline target strategies. By delving deeper into the clusters, strategies such as offering discounts to customers who are showing churn due to price or offering new types of subscriptions to customers to retain more customers.

**Future tasks:**

In the upcoming part of the project, churn rate for each cluster can be calculated, which can help a company decide the most influential characters of churn and appropriate strategies can be implemented. Various marketing strategies can be implemented such as offering low prices for yearly subscription