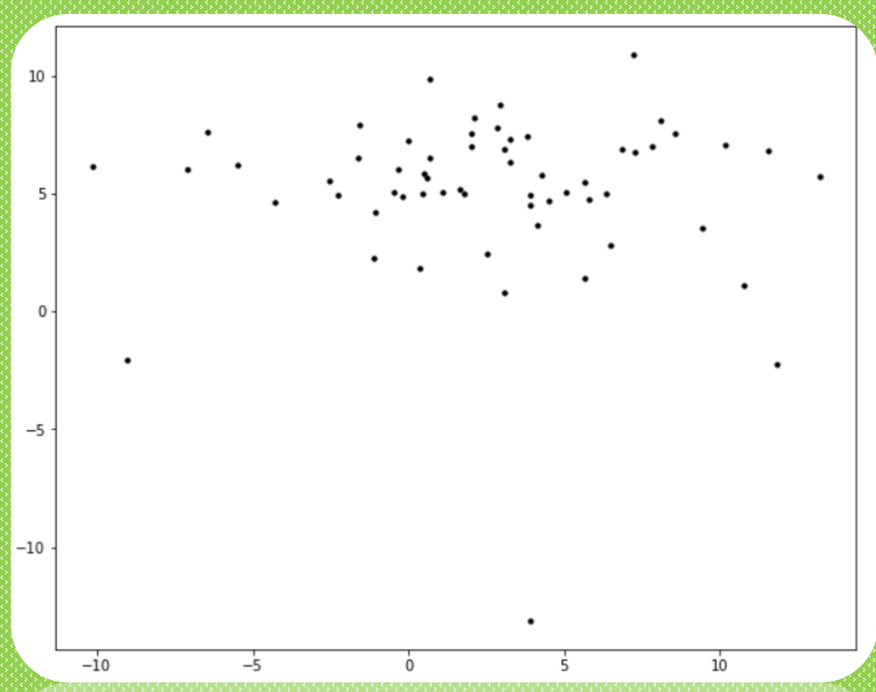
GithubLink
https://github.com/Arslanilyas1/Assignment3-datascience

DATA SCIENCE ASSIGNMENT 03

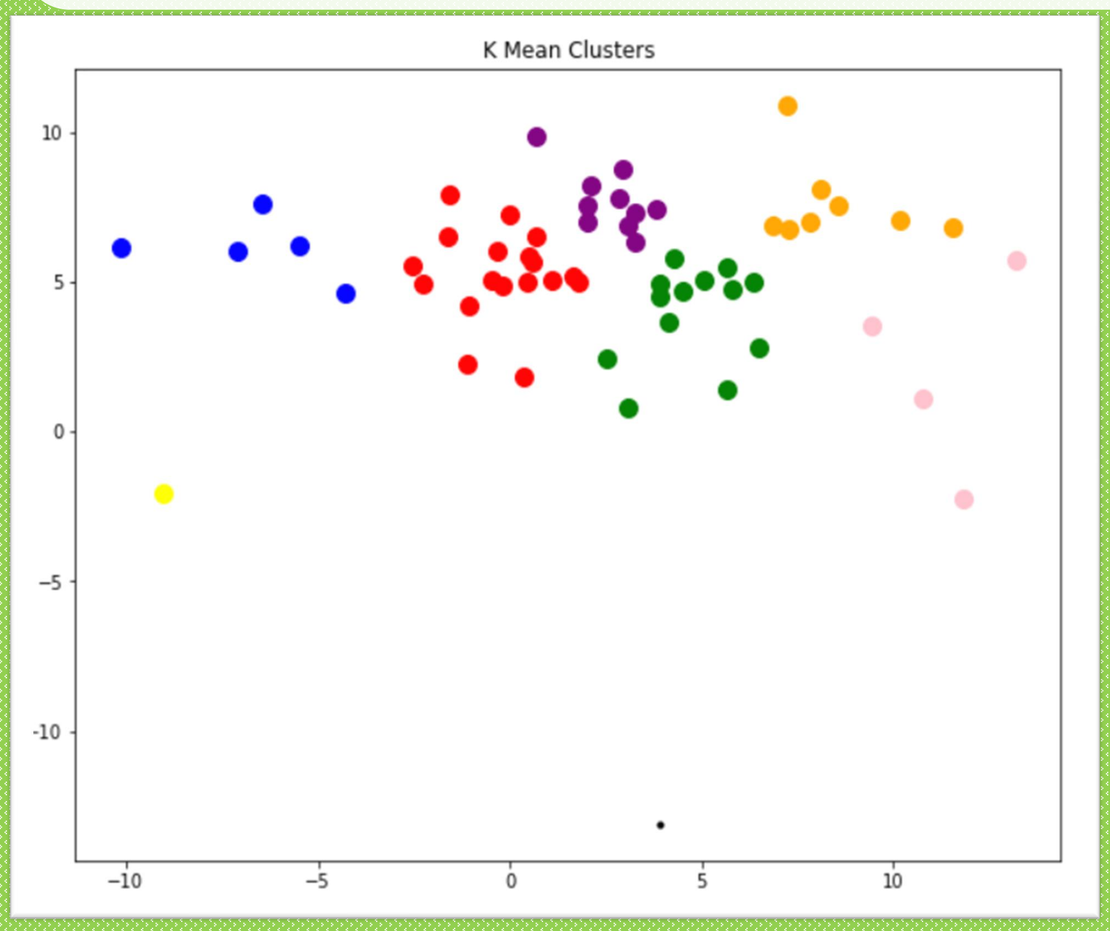
Introduction.

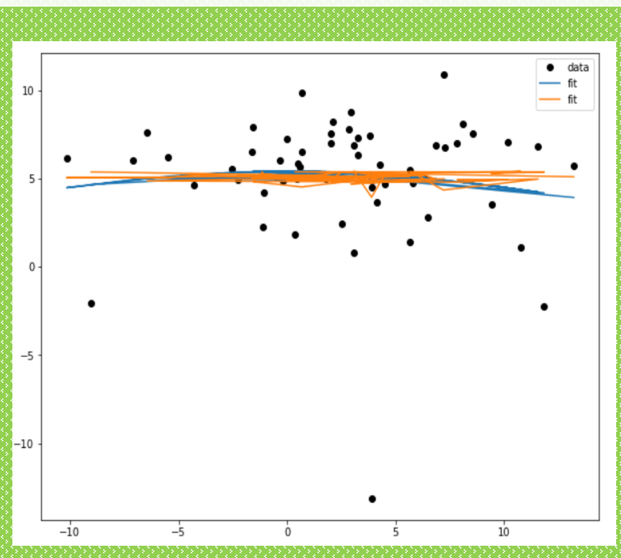
This project is about the clustering technique, which is highly useful in data science when you need to group data in a data set. We use data from the World Bank in this assignment that is relevant to GDP growth in Greece and indonasia. This time, we introduce several new libraries, such as the Python third, which is built in for data science-related words, comparable to SK. Learn a few things from that library, such as preprocessing for data normalization and Science Pi, where we import curve fit technique and cluster where we utilize k mean cluster from the same library as K learn. You can also use some of the built-in libraries we utilized in previous assignments . Can we go over the findings one by one and share them with you in this poster? We not only received the results, but we got them with the help of data visualization, which we used to build a data visualization for more precise learning. We'll now see how many clusters we can create with just this clustering algorithm from this dataset. On the left side, we explain data in the same way. This data is for 2019 and 2020, including an explanation of the entire dataset.



Normalization.

The process of turning real-valued numeric properties into a 0 to 1 range is known as normalization. In machine learning, data normalization is used to make model training less sensitive to feature scale. As a result, our model will be able to converge to more accurate weights. The Python preprocessing package includes the normalize function for data normalization. It accepts an array of numbers as input and converts them to a 0 to 1 range. We divide the result by the range after subtracting the minimal value from each item. The difference between the maximum and least value is known as range.





Curve fit method

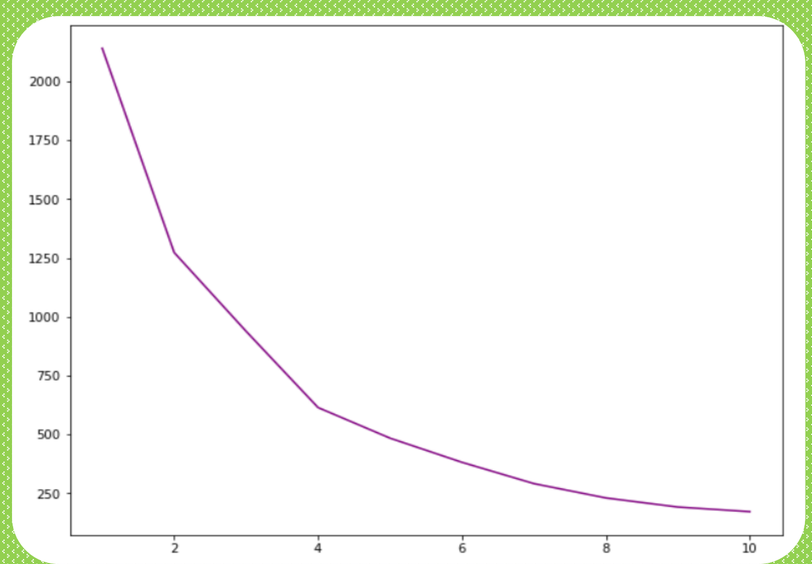
Curve fitting is a type of optimization in which the optimal set of parameters for a certain function that best fits a set of data is determined. Unlike supervised learning, curve fitting requires the specification of the function that translates input examples to outputs. We also use built-in libraries to fit the dataset using the curve fit method.

Clustering.

Clustering is a set of techniques for organizing data into groups. Clusters are groups of data objects that are more similar to one another than they are to data objects from other clusters. To locate groupings of data objects in a dataset, the unsupervised machine learning technique k-means clustering is utilized. There are several clustering methods to choose from, but k-means is one of the most popular and accessible.

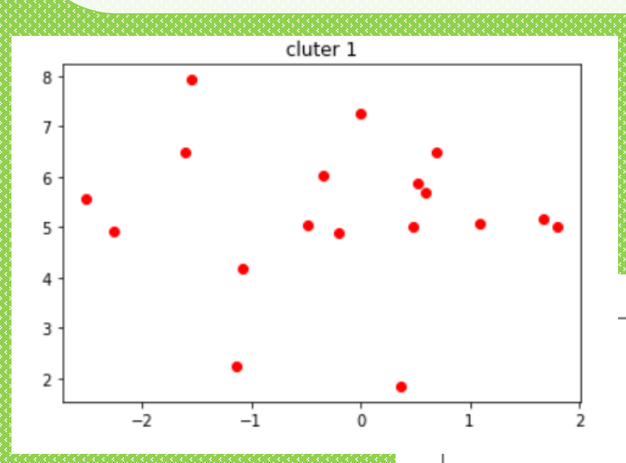
K mean cluster

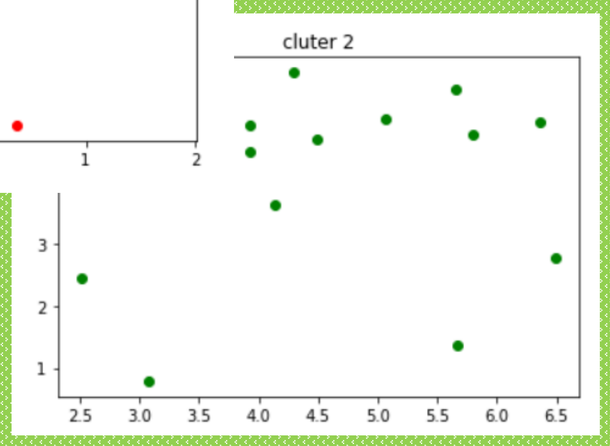
This is the k-means cluster, in which we clearly observed three different clusters in the scattered form of matplotlib graph according to elbow curve vi set number of cluster equals to 8 and we observe three different clusters of k mean algorithm now we examine all three plasters separately and check which clusters have more elements on the left top figure is about the entire dataset, then curve fit, and finally elbow chart. Finally, we make all of the clusters couple together and visualise using Matplotlib scatter plotting all of the clusters in a single plot.

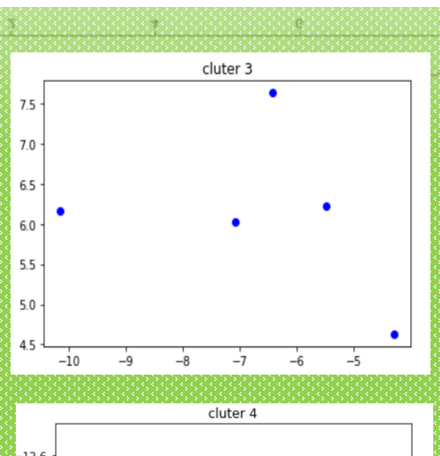


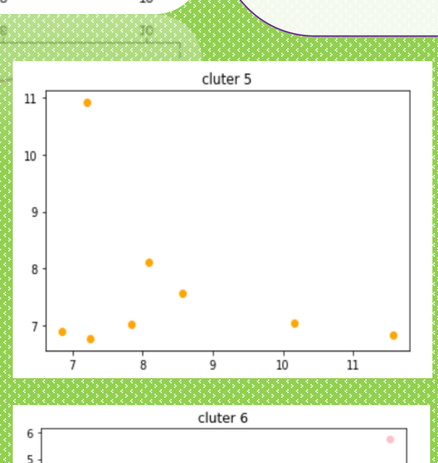
Elbow Curve

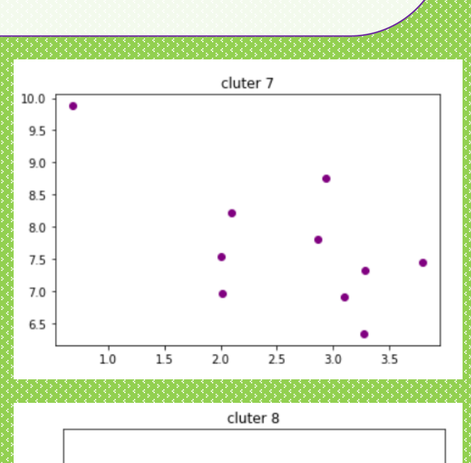
Before clustering, it's critical to determine how many clusters are possible in a particular data set by applying the elbow curve approach to the library SK. Learn that this is a method of inertia by default and append it to the list, then plot the entire list data using matplotlib and claim that where the line becomes straight until that point, note the value of x and assume that number of clusters.

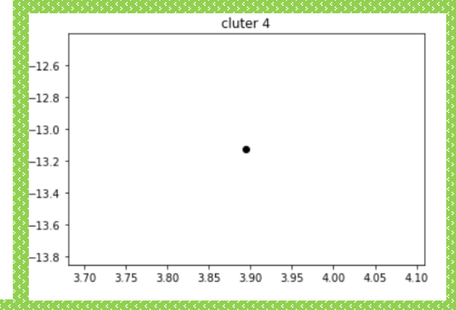


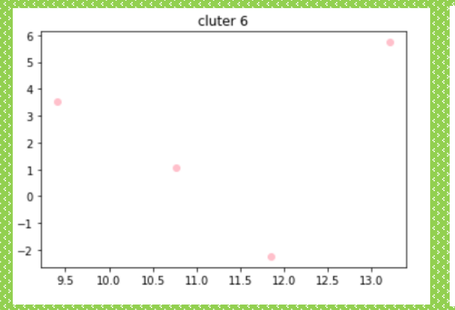


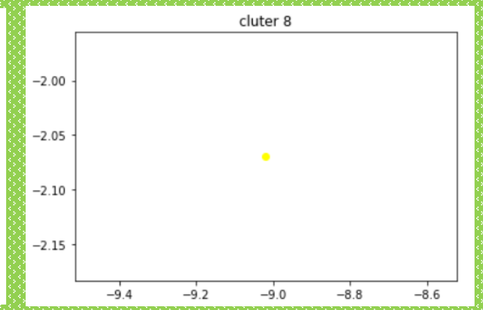












Interpretation of Results.

according to above prediction we make different visualization to check which cluster contains more elements
all of the 8 cluster is in front of you and we visualize all cluster separately