**Choose a dataset suitable for a clustering problem that includes at least 1,000 samples and 7 features. It can be any real-world or benchmark dataset that you have permission to access and use.**

* **Explain the dataset, conduct exploratory data analysis (EDA) and preprocess the data as needed to prepare it for modelling, discuss ethical/social issues.**
* **Apply two clustering algorithms on the selected dataset of your choice using Python.**
* **Provide detailed explanations of your data mining processes, including the rationale behind the choice of algorithms and parameters.**
* **Present the results, compare the performance of the two algorithms, and critically evaluate the clustering output**
* **Discuss how it will benefit the related business or help solve the problem at hand. Explain how the insights gained from the model could influence decision-making, or improve processes.**
* **Suggest ways the business or organization can leverage these insights to achieve specific goals or improve performance (Provide actionable recommendations)**

Answer:-

# Explain the dataset, conduct exploratory data analysis (EDA) and preprocess the data as needed to prepare it for modelling, discuss ethical/social issues.

Dataset:-

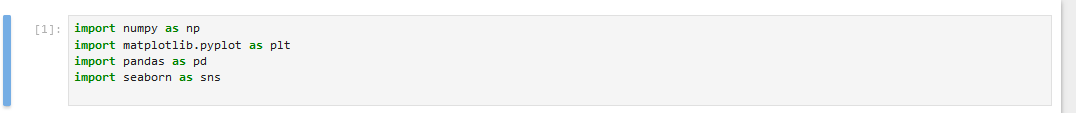
The dataset I have chosen is from civil engineering department where the different components adds and make the concrete compressive strenghts. In this dataset there are 1035 samples and 11 components that adds up at different ratio’s and make concrete compressive strengths accordingly.

# Reference:-

[https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength](https://archive.ics.uci.edu/ml/datasets/Concrete%2BCompressive%2BStrength)

EDA:-

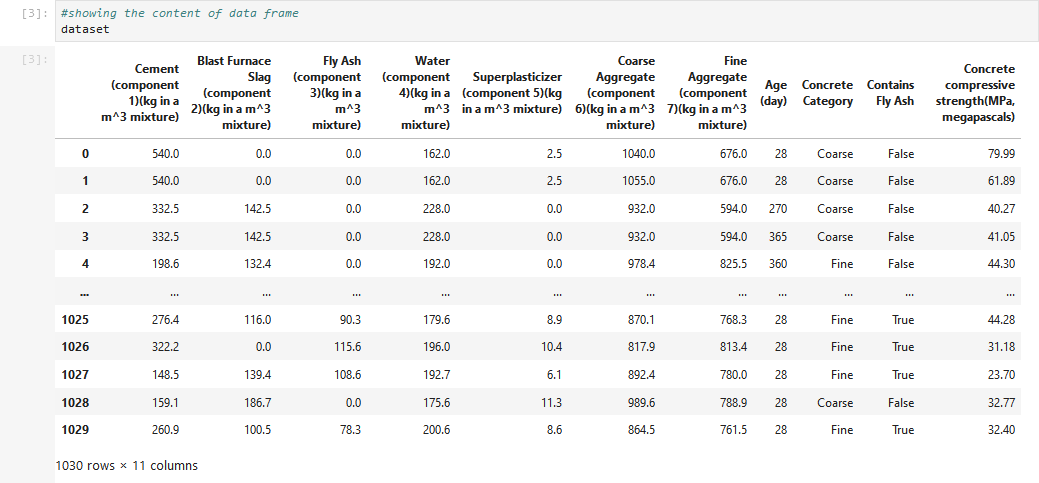
Now I am going to explore the data set. First of all I import the necessary libraries as

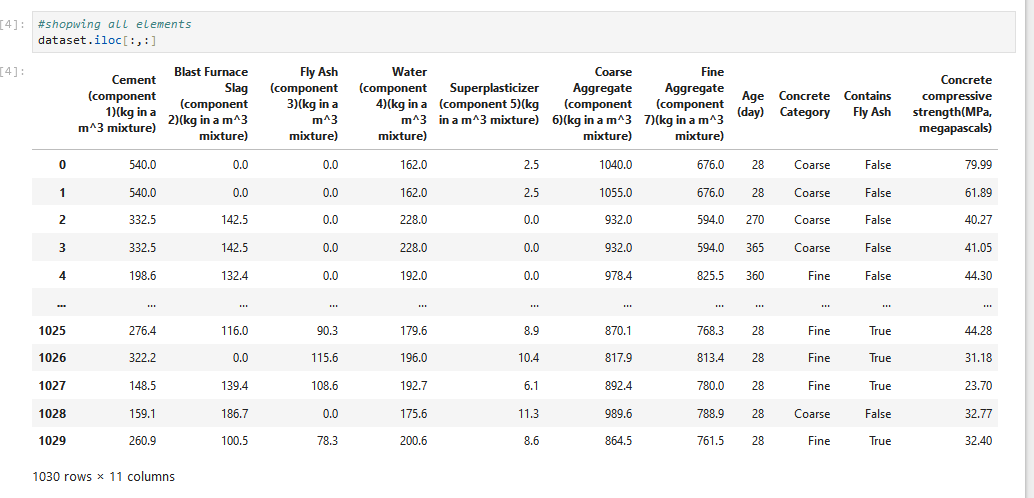


Then I load my data set file into pandas as follow

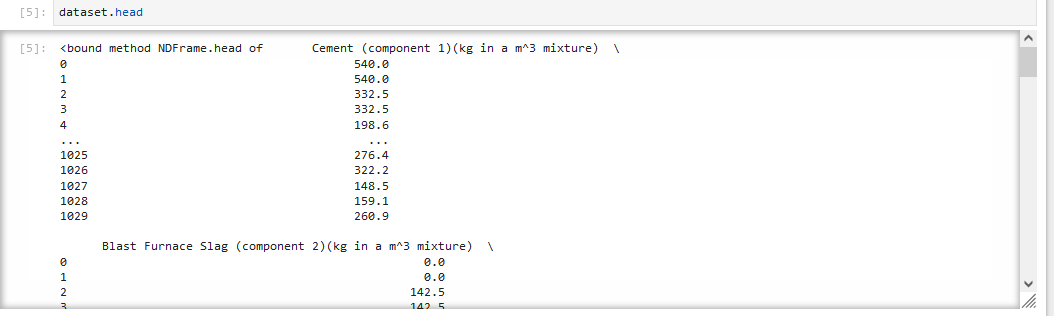


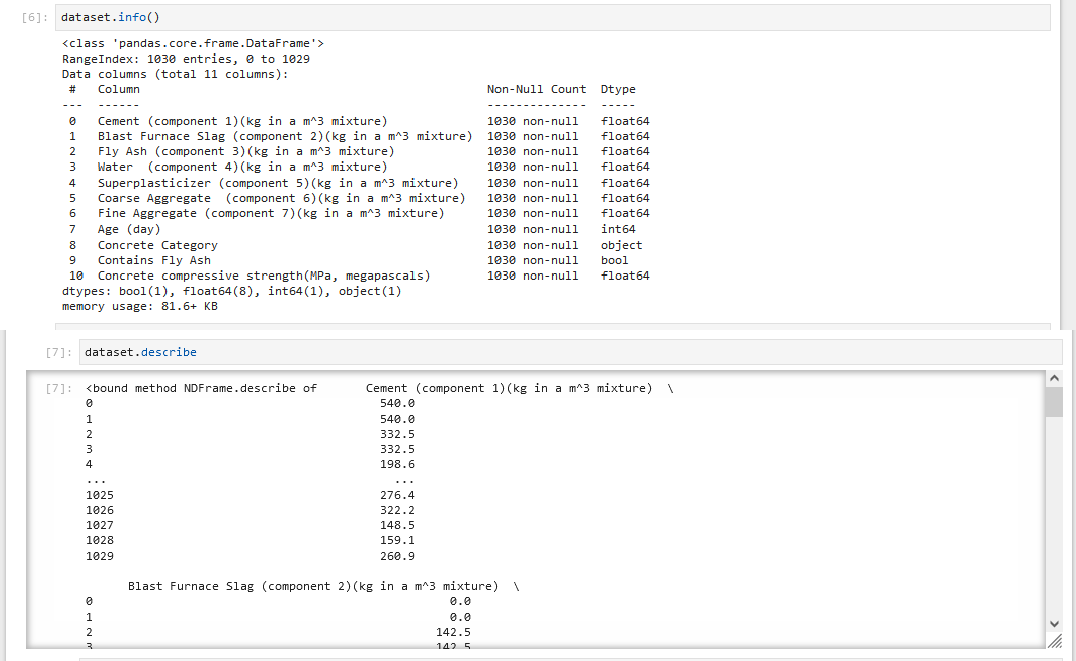
Then I write a command to show the data set that I just uploaded as



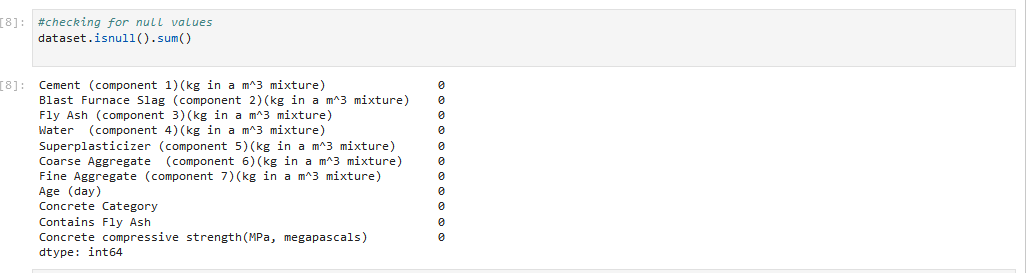
Now showing all the elements using iloc command as follow

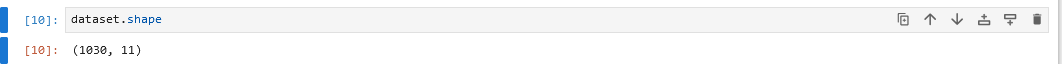
Now I explor the data set using head(), info() and describe() to get the better understanding of the dataset as follow





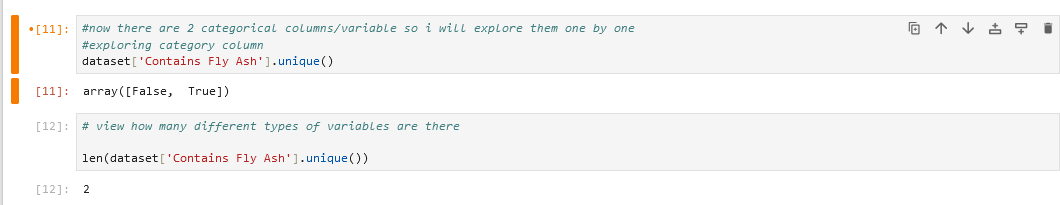
Then I write a code to see the missing value in my data set as follow



From the output we can see that my data set does not have any missing value. Then I write a code to see the shape of my dataset.

From the output we can see that it has 1030 samples and 11 features.

Now in my dataset there are 2 categorical variables so I am going to explore these categories one by one.



From above we can see that contain fly ash column has 2 categories.

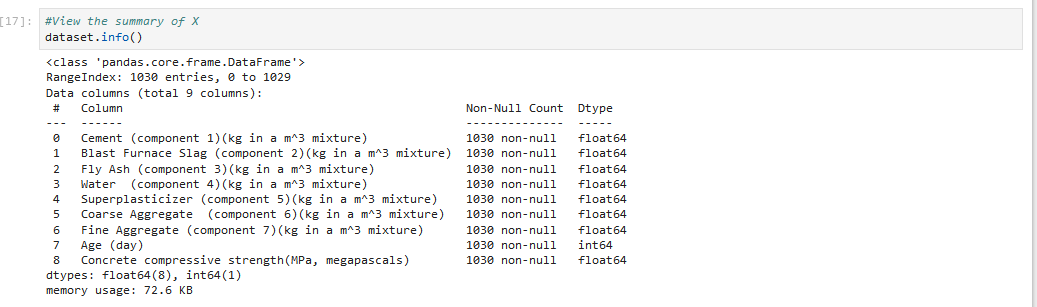


From above we can see that concrete category has also 2 categories.

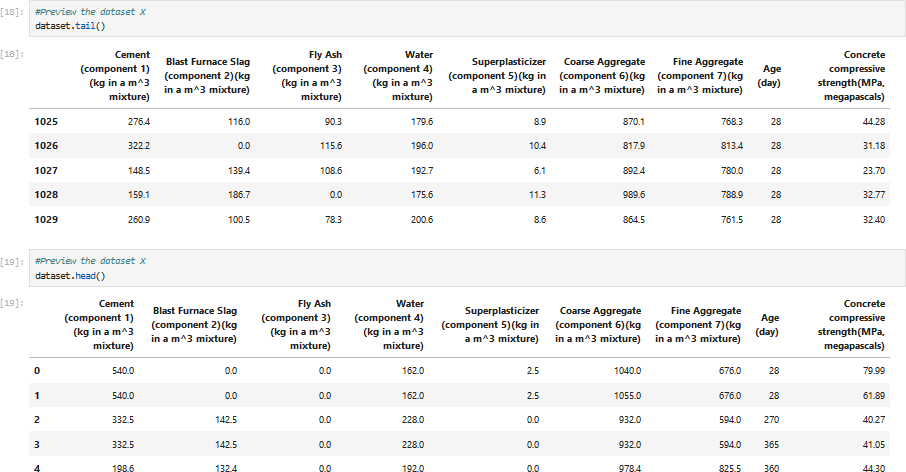
As our aim is to do clustering on this dataset So these categorical columns are not useful for us. Now I am going to remove these columns from my data set as



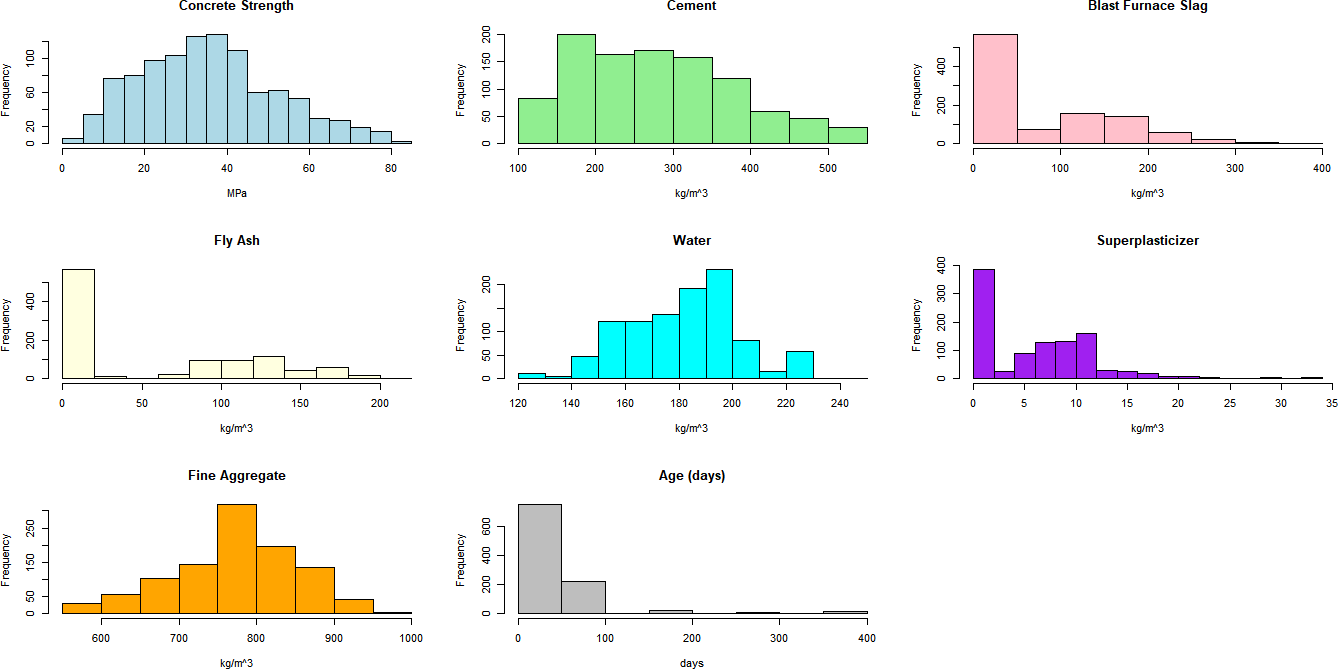
After removing these columns my dataset becomes as



And these are the head and tail of my transformed data set as follow



And this is the graphical representation of my data



Now from above graphical distribution we can see that the data is not standardize like values for cement is much more than that of age(days) and we can see also see difference among every graph as well. So we need to standardize the data before applying clustering.

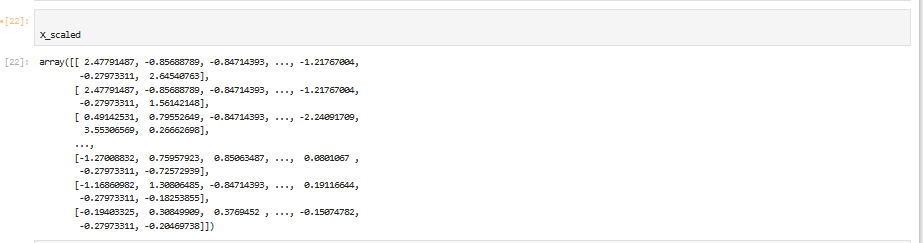
First I declare my dataset as X



I initialize the standard scaler and then I fit and transform the data as follow

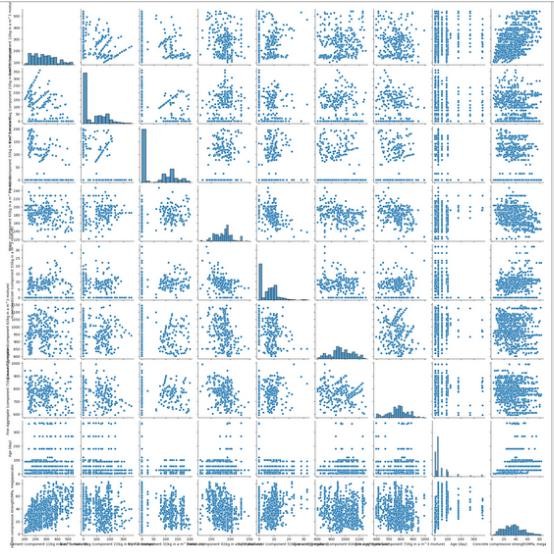


Now you can view the transformed data as



Now to see roughly how many clusters are there in the data set I am going to plot the scatter plot using seaborn as follow





From above scatter plots we can see that the data set is very compact and there are no visual clusters. So we have to apply different techniques to visualize the clustering in the data set.

# Social ethical issues:-

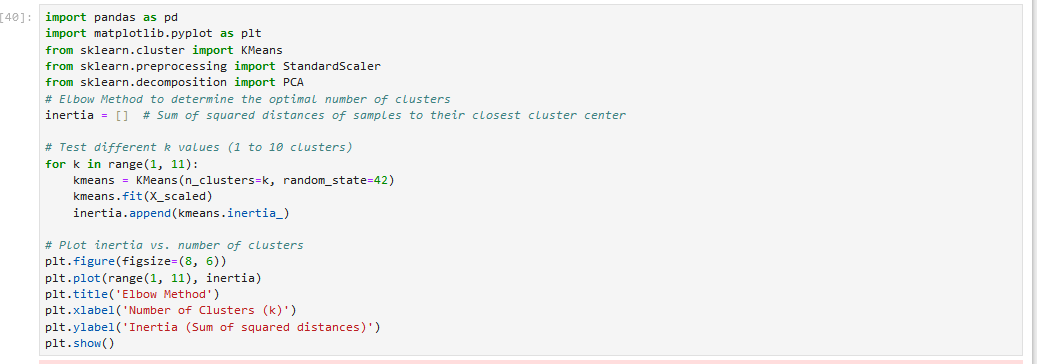
Now for the purpose of clustering I will make sure that the grouping of concrete mixtures is not misleading. The data is properly cleaned to avoid mistakes that could affect safety. I will assure that the clustering groups should be easy to understand and the result should help use material efficiently to protect the environment.

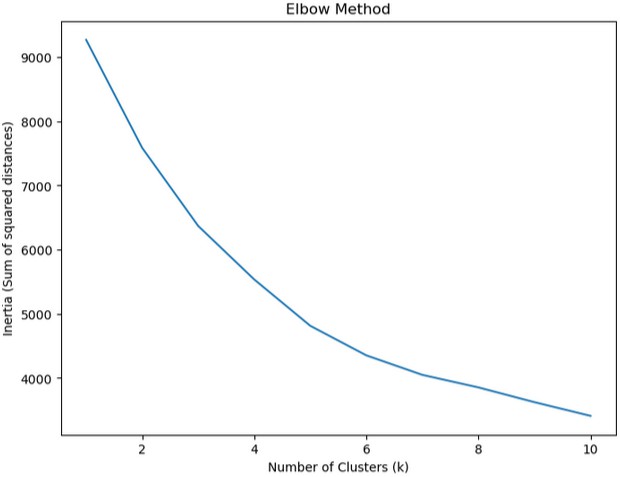
# Apply two clustering algorithms on the selected dataset of your choice using Python.

K-means clustering:-

I choose k-means clustering because it is very efficient for large set data set and easy to implement and understand.

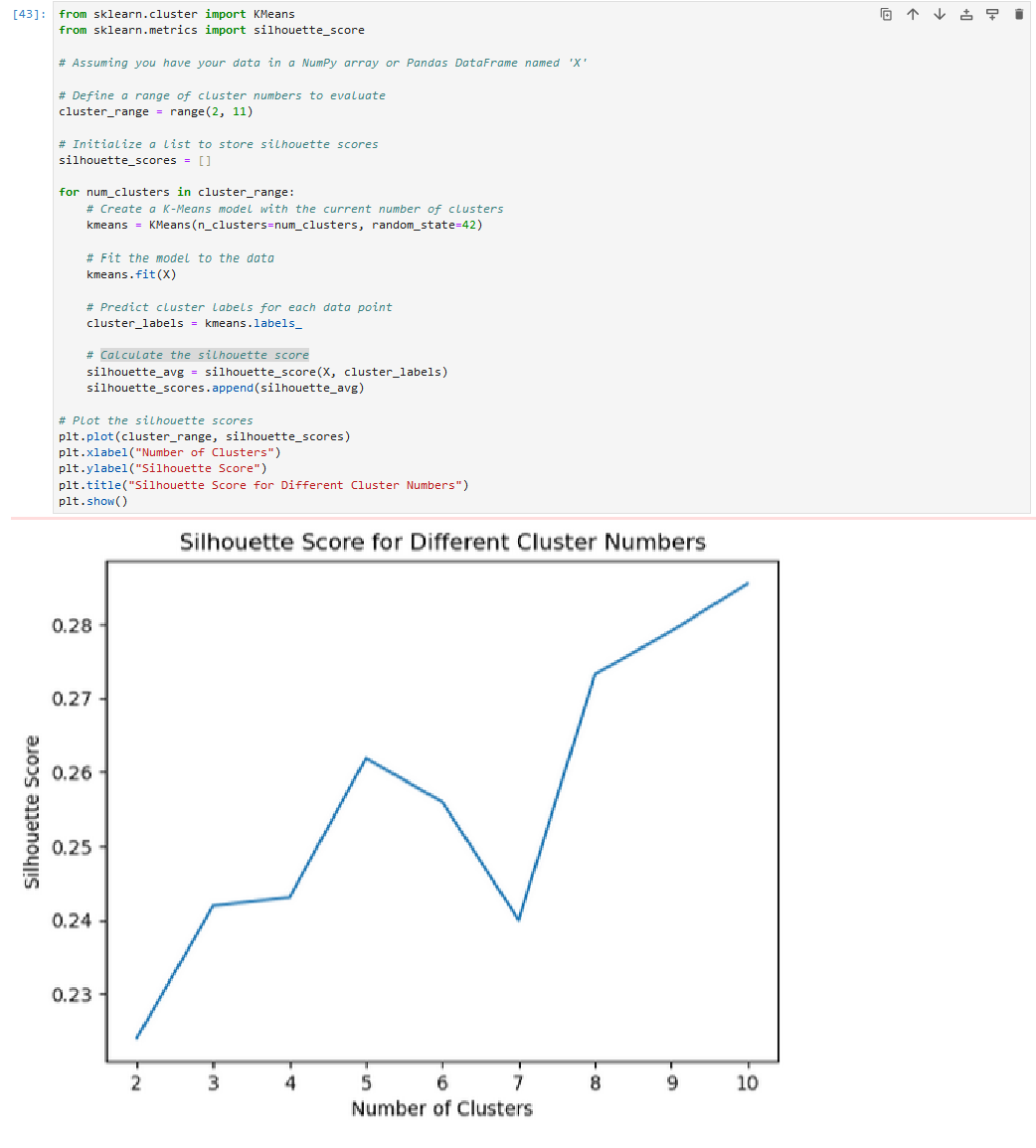
Now I am going to implement the K-mean clustering algorithm, for this purpose first of all I will apply elbow method. The elbow method is used to determine the number of clusters in the data set.



The above code uses Scitkit Learn to perform K-means clustering with values of *k* from 1 to 10 and saves the resulting wcss to an array. We then plot this using matplotlib. I setted the value of random variable equal to 42 so that every time I run this code with the same data set it will give me same output.

In above graph we can see that the elbow is at 5, so its mean that the number of clusters is 5.

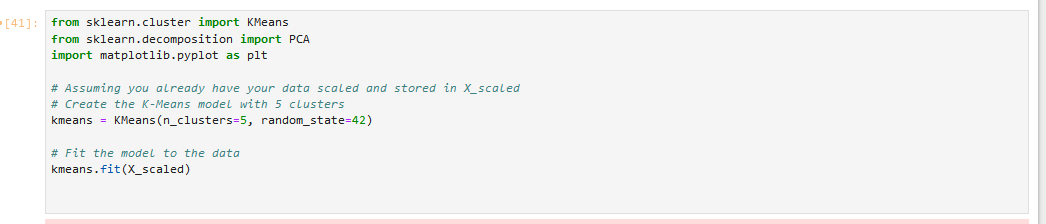
But for more clarification of the number of clusters I am going to calculate and plot average silhouette score for different number of cluster.



From above graph we can see that there is a peak at 5th cluster and after 5th cluster silhouette score goes downward. So its mean that the total number of clusters in this dataset is 5.

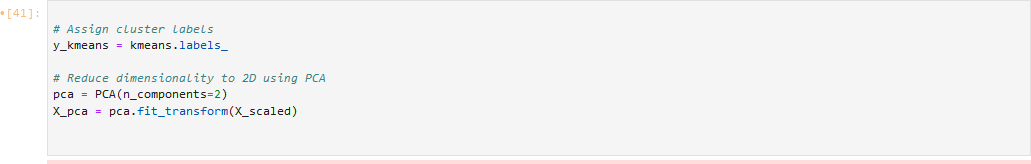
Implementation:-

Now I am going to implement the K-mean clustering algorithm using k-means.fit() method as



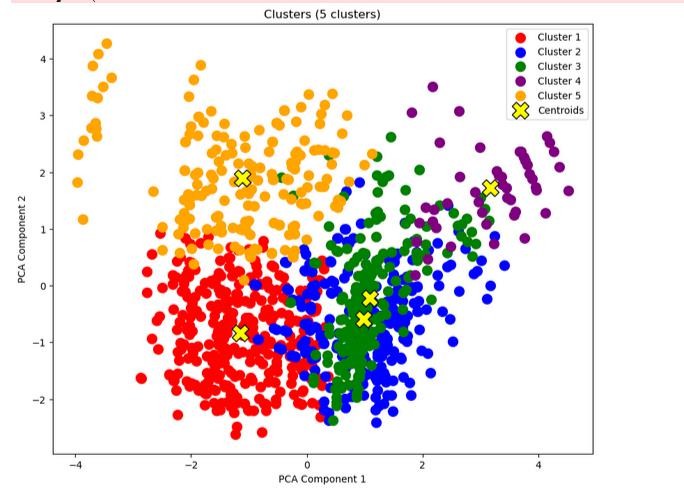
In this code I labeled number of cluster =5 which we have determined earlier and random state =42 so that it will give same output after every run.

Then I assigned the cluster labels and apply PCA to reduce the dimensionality to 2D as follow



Then I plot the clusters and as well as their centers so that we can visualize that how far the cluster is from one another.

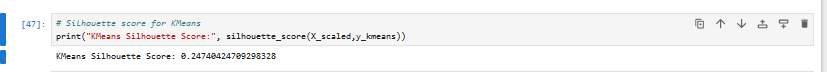




From above we can see that there are 5 clusters. The dataset is so compact that we cannot completely separate each clusters from another cluster. But still by applying different colors and centeriod for each cluster we are somehow able to see the grouping in the dataset.

# Result:-

I will write a code to calculate the silhouette score to determine the correctness of our k- mean clustering algorithm.



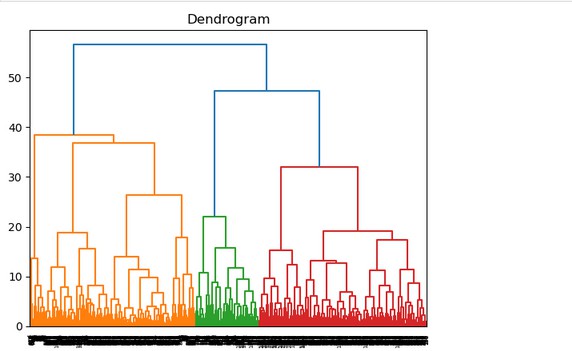
The silhouette score near to 1 or 1 mean very good cluster and 0 or below zero means very bad clustering. But in our case the dataset was too much compact and there are several mixtures with similar compositions and compressive strength. so that we got silhouette score of 0.247 it mean not very bad clustering and not very good clustering. It’s a moderate result.

# Hierarchical Clustering:-

I choose hierarichical clustering because it provides a hierarchical structure of clusters: allowing for more granular analysis.

Now I am going to apply hierarchical clustering on the same data set as I describe above. For hierarchical clustering I will use aglomerative method. Now for applying

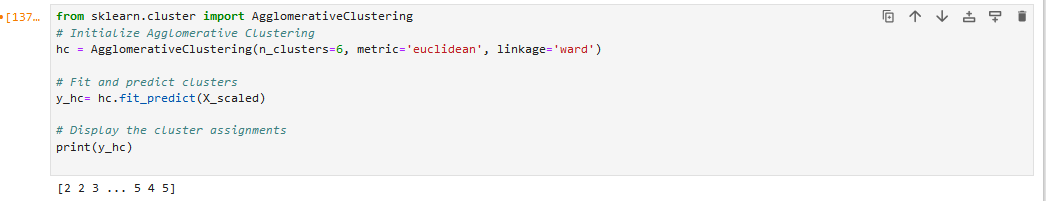
aglomerative method we first to applydendrogram which shows the point at which each cluster merges with another.



From above graph if we divide the y-axis into halve than we can see that it is giving 6 clusters.

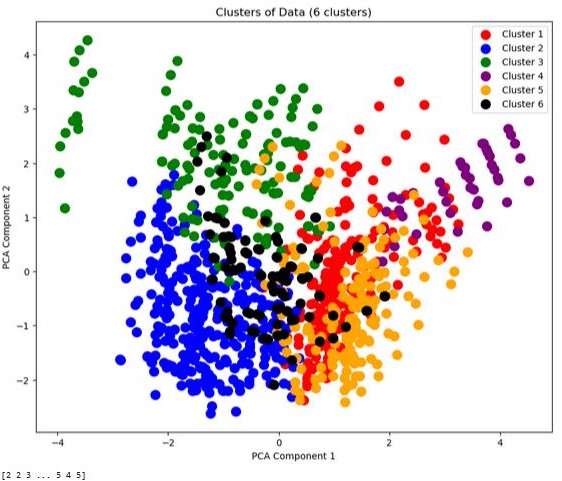
# Agglomerative Clustering:-

Now I am going to apply agglomerative clustering using fit and transform method and 6 number of clusters on scaled data set as follow



Then I reduced the dimensionality of data to 2D using PCA and then plot the clusters.





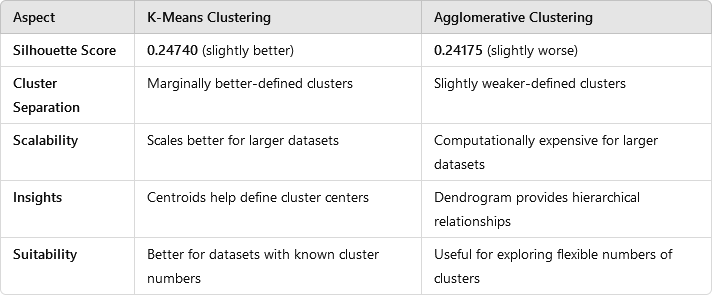
Here you can see the 6 number of clusters.

Now I am going to calculate the Silhouette score for Agglomerative Clustering as



From above results we can see that the Silhouette score for Agglomerative Clustering is 0.2417 because the data set was too much compact and there are several components with similar compositions and compressive strength. I can not make clear clusters but this score is also acceptable mean it is not considered as very bad clustering but you can say that it’s a moderate clustering.

# Comparison:-

****

**Business Benefits and Problem Solving:-**

* If we talk about construction businesses, they can use this clustering results to identify clusters that represent high strength concrete for critical structures like bridges or skyscrapers and lower-strength concrete for less demanding projects like pavements or residential buildings.
* A construction business can also use clustering results to identify underperforming concrete mixtures in terms of strength or cost-effectiveness.
* In this clustering model each cluster can serve as a "recipe" for specific applications which business can use to uplift their business.

# Actionable Recommendations:-

* A construction business can use this clustering results to develop concrete categories, such as "economy,""standard," and "high-performance" concrete.
* A business can use this clustering results to educate their customers about cluster-based concrete solutions so that customers can buy and apply these concrete mixtures according to their requirement.
* A business can focus on mixtures which have high concrete strength and low cement component because manufacturing of cement is not environment friendly.
* A research team of a construction business can use this clustering results to focus on clusters with higher demands.