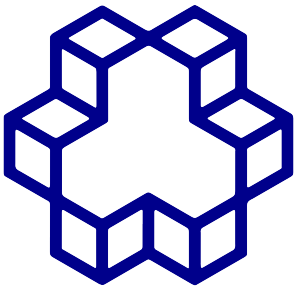
CODING ASSIGNMENT 2

Kmeans



دانشگاه صنعتی خواجه نصیرالدین طوسی

Artificial Intelligence

Instructor: Dr.Pishgoo

TAs: Amirreza Mehrzadian, Alireza Saeednia

Deadline: 1403/02/27

1 EDA

It is always better to get some intuition from your data before building any model. The overall objective of exploratory data analysis is to obtain vital insights and hence usually includes the following sub-objectives:

- Identifying and removing data outliers
- Identifying trends in time and space
- Uncover patterns related to the target
- Creating hypotheses and testing them through experiments
- Identifying new sources of data

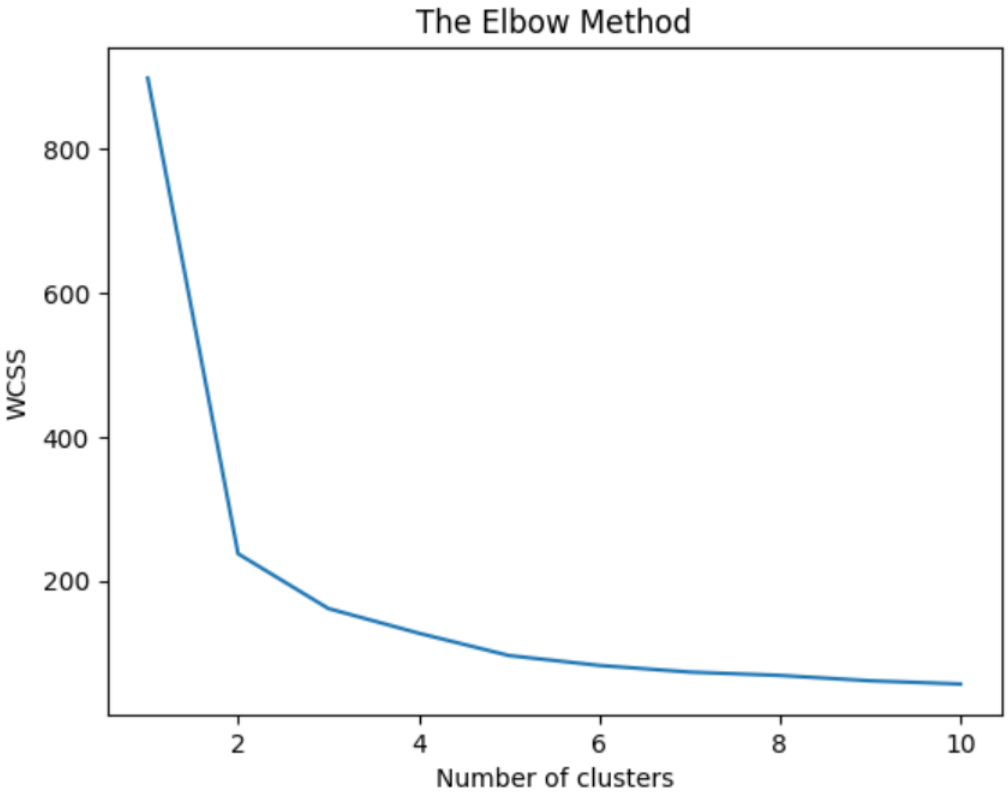
The key components in an EDA are the main steps undertaken to perform the EDA. These are as follows:

1. Data Collection
2. Finding all Variables and Understanding Them
3. Cleaning the Dataset
4. Identify Correlated Variables
5. Choosing the Right Statistical Methods
6. Visualizing and Analyzing Results

TODO: your dataset is almost clean but has some null columns, 3 categorical variables, and 1 non-numeric variable. handle them. At the end Scale the features and get them into a similar range.

2 KMeans

perform K-means clustering. Use the elbow method to determine how many clusters there should be. This method involves running the K-means algorithm for a range of values of K (number of clusters), and for each value of K, calculate the WCSS. Plot these scores on a graph. The point where the decrease in WCSS becomes less abrupt (forming an "elbow" shape) can be used as an indicator of the optimal number of clusters. kmeans.inertia_ gives you Within-Cluster Sum of Squares (WCSS): This is the sum of the squared distance between each member of the cluster and its centroid. The idea is to minimize WCSS. If the WCSS is low, that means the clusters are compact, and if it's high, the clusters are spread out.


The Elbow Method
WCSS
800
600
400
200
2
4
6
8
10
Number of clusters