Homework Part 2:

Technology used: Python, IDE: Jupyter Notebook

Report submitted by: Menaka Katapadi Kamath,

email: [mxk165230@utdallas.edu](mailto:mxk165230@utdallas.edu)

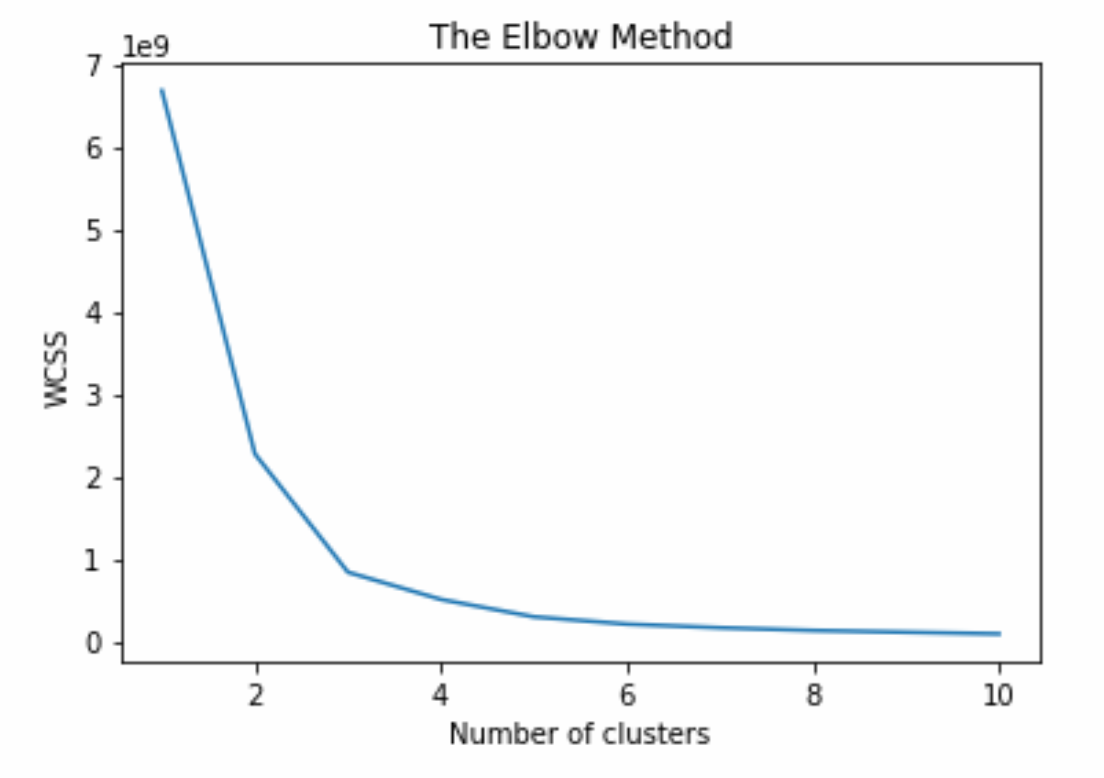
Dataset used: FileType.csv

Number of clusters formed: 3

Cluster Labels: 0, 1, 2

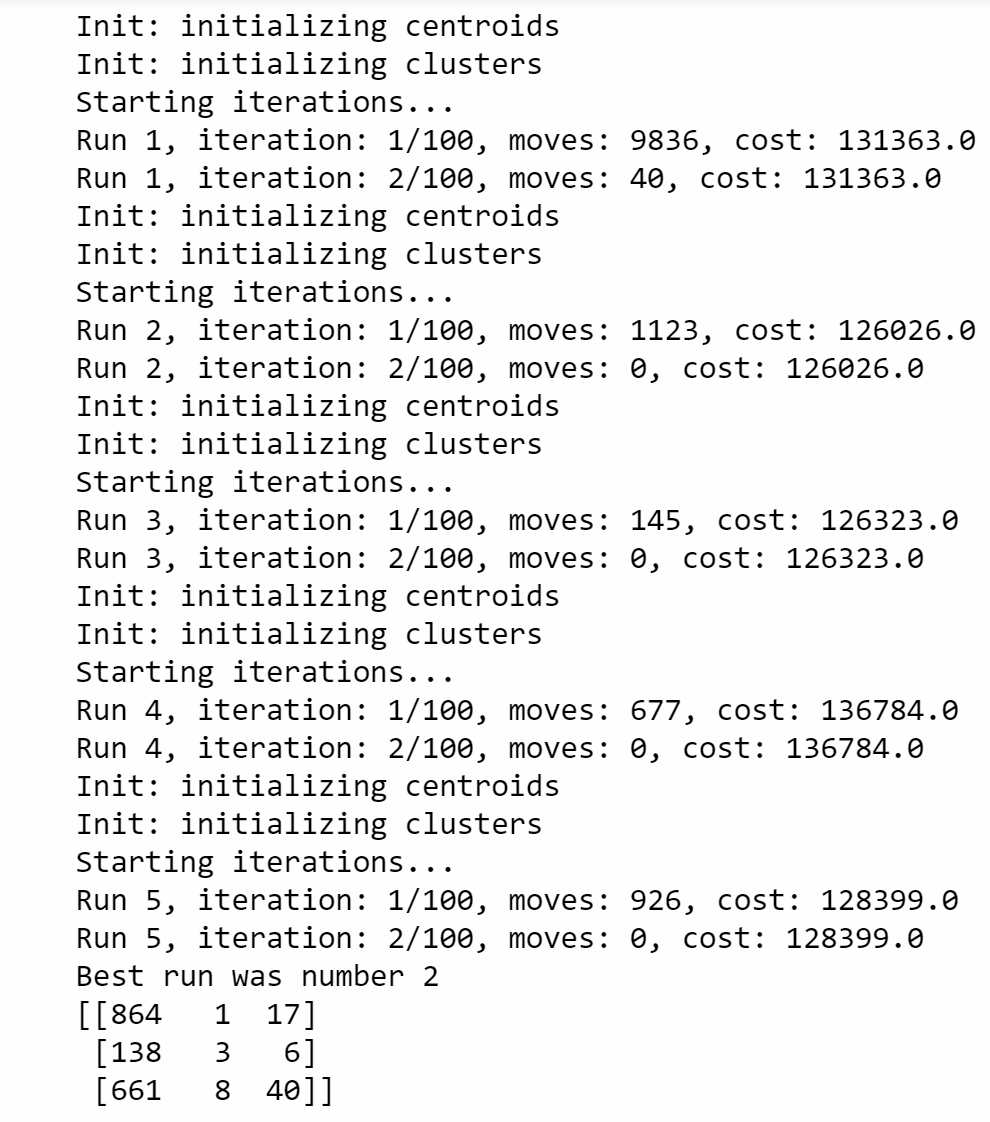
Subset of features used: File Type, InfraId(indicating the repository), Date.

I used K-means clustering to find the optimal number of clusters using Elbow Method.



Upon dropping the elbow point towards the X axis, we get optimal number of clusters for dataset containing subset of features(File Type, InfraId(indicating the repository), Date).

Since all the three features are categorical, I used Kmodes algorithm to assign the cluster labels to every observation. Kmodes algorithm works very well for categorical data.

Below is the clustering analysis output:

More details can be understood from the code.