

# Travel Review Ratings

Rajaraman Ganesan  
Master Engineering  
Electrical & Computer Dept.  
Western University, London  
[rganesa@uwo.ca](mailto:rganesa@uwo.ca)

Vatsal Shah  
Master Engineering  
Electrical & Computer Dept.  
Western University, London  
[vshah56@uwo.ca](mailto:vshah56@uwo.ca)

**Abstract--** Advancement in technology has fundamentally changed how information produced and consumed by all users involved in travel. Travelers can now access different sources of information, and they can generate their own content and share their views and experiences. Reviews shared through online has become a very influential information source that affects travel in terms of both reputation and performance. However, the volume of data on the Internet has reached a level that makes manual processing almost impossible, demanding new analytical approaches. Clustering technique considered as a possible solution to limit the volume of data. In travel review, based on users rating on social media say After pre-processing the dataset, the problem reduced to Clustering problem. Here, we make use of clustering algorithm techniques like the k-mean, k-medoids and CLARA algorithm using Python and R. We conclude by evaluating results and compared with custom models and available libraries in python such as Sklearn. In addition, we evaluate results of algorithm in R language with available libraries and visualized the results. Moreover, explored about Amazon SageMaker and deployed python code.

**Keywords--** Clustering models, k-means, k-medoids, Clara, Cluster evaluation, tourism, Sklearn, python, R, silhouette plot, Amazon SageMaker

## I. INTRODUCTION

In the age of e-commerce, every industry is involved in online sales, and the hospitality and tourism industry are no exception. The participatory nature of the Internet in recent years has led to an explosive growth of travel-related user-generated content. Travel planning has become one of important commercial use. Sharing on the web has become a major tool in expressing customer thoughts about a product or Service. Many tourists look for some places like fun malls, restaurants or vacation spots, etc. online in recent times [1]. After consumption customers give feedback/rating, online so online reviews have become increasingly important. They are fast, updated and available everywhere and have become the word-of-mouth of the digital age. Thus, online review plays a critical role in the tourism industry, which mainly offers services and focuses on customer satisfaction. This is the main reason

people spend time online reading the review/rating backing their decision-making.

In this study, user ratings are captured from Google reviews across the Europe region and average rating ranges from 1 to 5. The dataset contains information on 25 variables, obtained from the UCI Machine Learning Repository. With these reviews, we can make a good decision about the places about to visit, nature of the user.

### A. Background & Motivation

The reason for choosing the topic is to find the best places the people can visit. We process the data provided analyze and clusters the range of rating provided by the tourist. The goal of this project is to resolve this problem by building and comparing various techniques using unsupervised learning algorithm. Moreover, to encounter the difference in the process and issues by applying the custom model and pre-built libraries.

### B. Aim & Objective

The objective of the problem is to cluster the range of ratings provided by various consumers in various places they have visited. This project is helpful for solving problems using various clustering techniques such as K-means, K-medoids and CLARA algorithm and find the average ratings given by users on different places. We implement by applying various methods, custom model and pre-built libraries in python by understanding the process, compare and evaluate results. At the end, export more about cloud technology in Machine learning like Amazon SageMaker.

The steps followed to manage these goals:

1. Understand the selected dataset
2. Display some graphical information and visualize the features
3. Data pre-processing
4. Apply clustering algorithms on the selected dataset using a custom model of k-means in Python
5. Apply k-means algorithm using pre-built libraries in python
6. Apply algorithms like k-means, k-medoids, Clara in R
7. Evaluate the model in python & R
8. Compare the model and find the optimal one
9. Explore cloud technology in Machine Learning on Amazon SageMaker

### C. High Level Overview

The major purpose of this is to observe the ratings provided by users in various places. Clustering algorithms implemented to analyze this model. Many methods of clustering including soft clustering been used to develop this model of users review.

The remainder of the paper organized as follows. Section 2 summarizes the basic properties of applied models, section 3 explores the methodology with data preprocessing. Section 4 comprises the evaluation process and section 5 presents a summary.

## II. BACKGROUND & LITERATURE SURVEY

The research work of [2] presents a comparison of three different datasets gathered from travel and tourism domain. The first dataset has 249 user records with 6 attributes, seconds dataset has 980 user records with 10 attributes and the third dataset has 5456 user records with 24 rating attributes. They applied various clustering techniques such as k-means, k-medoids, and CLARA and Fuzzy c-means using R packages. In the end, they concluded the k-means algorithm performed better than other clustering algorithms.

There is much research on review rating because every people believe the rating provided online and plan his or her visit to those selected places. Many statistical methods have been applied to develop a travel review rating model, using, K-means, soft clustering algorithms.

### A. Clustering:

Clustering is a set of observations into subset in the same cluster are similar in some sense. In the world of machine learning, it is an unsupervised approach. Unsupervised learning applied while there is input data, but there are no corresponding output variables associated with it. The objective of clustering is to find different groups within elements in the data. Clustering algorithms find the structure in the data so that elements of the same cluster are more like each other than to those from different clusters. It has manifold usage in many fields such as machine learning, pattern recognition, image analysis, information retrieval, bio-informatics, data compression, and computer graphics [2].

### B. K-means Algorithm:

K-means is one of the simplest algorithms uses unsupervised learning method to solve clustering issues. It is mainly used when u have unlabeled data. The goal of this algorithm is to find groups in data, with the number of groups represented by the variable.

$$WCV(C_k) = \sum_{x_i \in \mu_k} (x_i - \mu_k)^2 \quad [3]$$

Where,

$x_i$  = entity belong to cluster  $C_k$

$C_k$  &  $\mu_k$  = mean of all entities forming the cluster.

The following steps followed to implement the k-means algorithm:

1. Determine the total no: of clusters to be formed.
2. Start by identifying centroids initially and random sampling done within the dataset to find initial points.
3. Calculate the distance between the entity and each of centroids and assign the entity to the cluster close to the centroids.
4. Revise the cluster centroid by calculating mean values of all entities. Repeat this for each cluster.
5. Reduce the total within-cluster variation. In addition, repeat the above step 3 and 4 until max number of iterations reached.

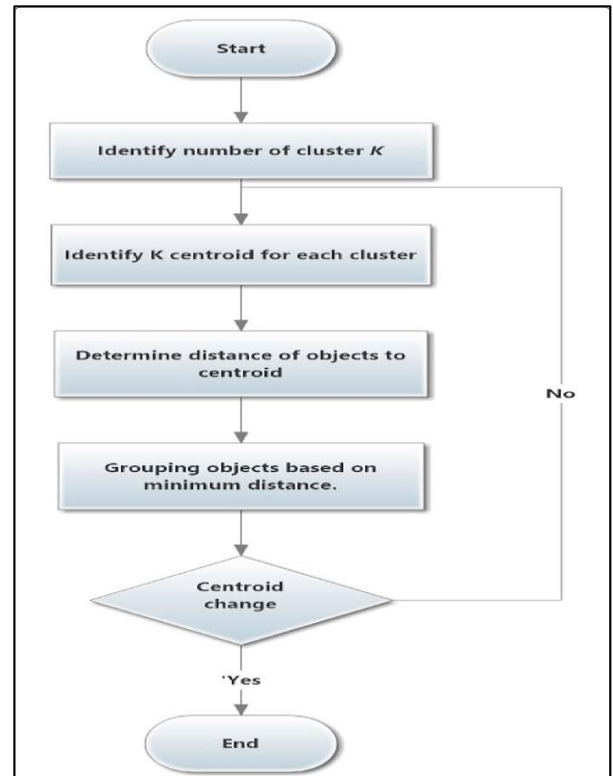


Fig (2.1) Process for k-means algorithm

### C. K-medoids Algorithm:

The k-medoids algorithm works similar to k-means algorithm, but less sensitive to the presence of noises and outliers in the data set. The key difference is in fitting the centroids. While in k-means algorithm the centroids need not be the actual entities from the dataset, k-medoids algorithm chose an entity from the dataset whose average dissimilarity to all the other entities is minimum as the medoid.

The steps to perform k-medoids algorithm is given below:

1. Determine k, the total number of clusters to be populated
2. Identify k initial medoids to start with. Medoids are always selected from the members of the dataset.

3. For each entity in the dataset
  - a. Calculate the distance between the entity and each of the k medoids
  - b. Allocate the entity to a cluster whose medoid is the nearest
4. For each of the k clusters
  - a. Scan for an entity that can reduce the average dissimilarity within the cluster.
  - b. If an entity is found that can reduce the objective function, swap the medoid and the entity in consideration.
5. If there is any change to at least one medoid in step 4, go to step 3 else end.

#### D. CLARA Algorithm:

CLARA (Clustering for Large Applications) algorithm is formulated as an extension to k-medoids algorithm to deal with datasets containing very large number of entities (when it ranges in several thousands). This is to overcome the scalability challenges of k-medoids algorithm like the needs for high computing time and large memory requirements. This is achieved by means of sampling approach.

The below given are the steps to perform Clara algorithm:

1. Split the dataset randomly into multiple subsets having a fixed size.
2. Apply PAM algorithm on each subset and choose the corresponding k medoids.
3. For each entity in the complete dataset
  - a. Calculate the distance between the entity and each of the k medoids.
  - b. Allocate the entity to a cluster whose medoid is the nearest.
4. For each subset
  - a. Calculate the average dissimilarities of the entities to their closest medoid (quality or goodness of the clustering).
5. Select sub-dataset with the minimum average dissimilarity.

### III. METHODOLOGY

In methodology, data description, independent variable, and dependent variable described with the scale of variables. Moreover, in the process data preprocessing and feature engineering described below.

#### A. Data Description & Preparation:

This dataset consists of 5456 total instances and 25 features including:

TABLE I  
ATTRIBUTES OF THE DATASET

Attribute No	Description
1	Unique user ID
2	Average ratings on churches
3	Average ratings on resorts

4	Average ratings on beaches
5	Average ratings on parks
6	Average ratings on theatres
7	Average ratings on museums
8	Average ratings on malls
9	Average ratings on zoo
10	Average ratings on restaurants
11	Average ratings on pubs/bars
12	Average ratings on local services
13	Average ratings on burger/pizza shops
14	Average ratings on hotels/other lodgings
15	Average ratings on juice bars
16	Average ratings on art galleries
17	Average ratings on dance clubs
18	Average ratings on swimming pools
19	Average ratings on gyms
20	Average ratings on bakeries
21	Average ratings on beauty & spas
22	Average ratings on cafes
23	Average ratings on viewpoints
24	Average ratings on monuments
25	Average ratings on gardens

There are 24 categories reviewed by users. It is shown in the below rating ranger from 1 to 5. Each categories has various number of users rating. To get better understanding of the data, visualization with hist and plot diagram as below:

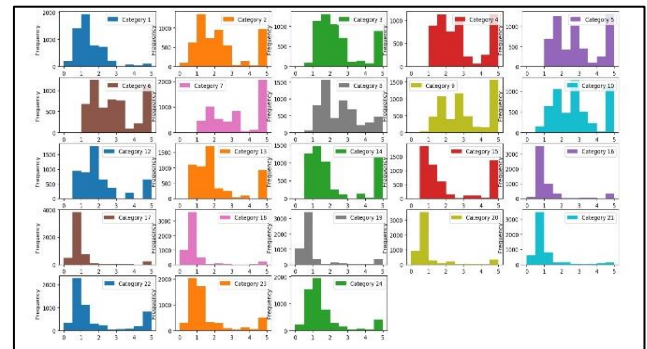


Fig (3.1) Visualization of data with hist diagram

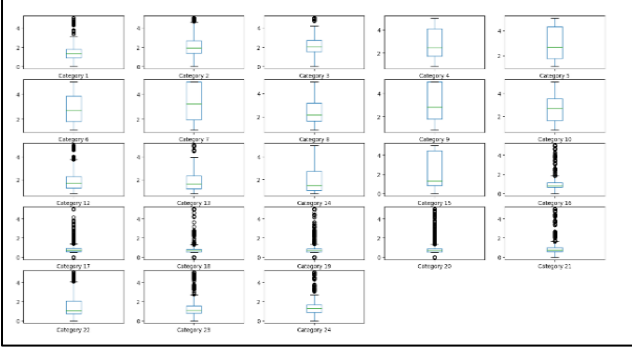


Fig (3.2) Visualization of data with box-plot diagram

In this dataset, to understand the importance of features with compare to each other and find the best features to evaluate model, used heatmap as shows below:

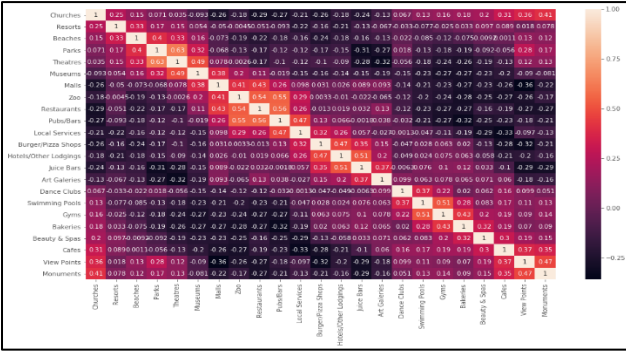


Fig (3.3) Visualization of data with heatmap diagram

### B. Data Clensing & Normalizing:

Mostly, data gathered from various sources, which have missing values and noises. Because of this, data cleaning is an important step to perform before applying an algorithm. There are various approaches for data cleaning. In Python, `isna()` used to identify null values of all attributes and `isna().sum()` gives the summery of all attributes with the number of null records. It can be replaced with mean value. In python, it can be implemented as `fillna()` function which replaces all missing values with the mean value.

Scaling or normalizing of the attribute is a practice [2] can be performed in clustering problem. By performing normalization all, the attributes are in the same range.

## IV. EVALUATION

The objective of a project is to identify the optimal number of clusters along with the best approach to apply KMeans algorithms. The objective of a clustering algorithm is to accomplish minimum inter-cluster similarity among clusters and maximum intra-cluster similarity within each cluster [4][5].

### A. Elbow Method:

A method of interpretation and validation of consistency within cluster analysis designed to help to find the appropriate number of clusters in a dataset [6][7].

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

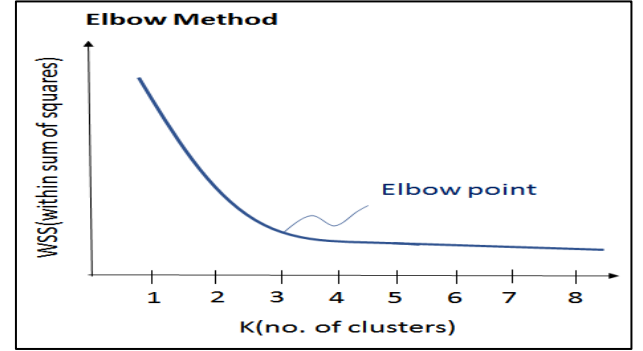


Fig (4.1) Elbow method graph represents WSS vs No. of clusters

### B. Silhouette Plot:

Silhouette is a method of interpret and validate consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance. [8]

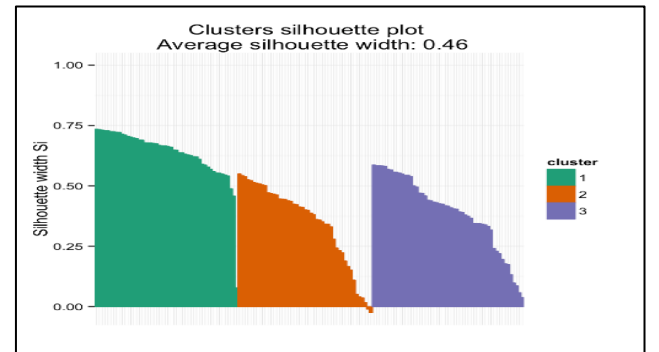


Fig (4.2) Represents the Silhouette plot

### C. Euclidean distance:

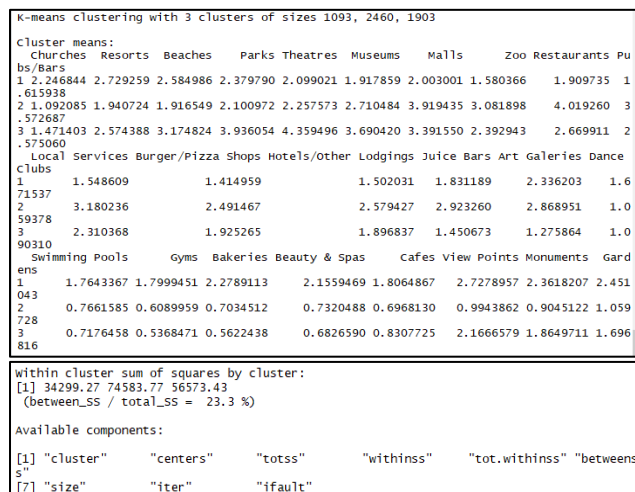
The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points. It is the most obvious way of representing distance between two points.

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad [9]$$

Where x and y are two vectors of length n.

Kmeans () function returns a list including:

- cluster: A vector of integers (from 1:k) indicating the cluster to which each point is allocated
- centers: A matrix of cluster centers (cluster means)
- totss: The total sum of squares (TSS), TSS measures the total variance in the data.
- withinss: Vector of within-cluster sum of squares, one component per cluster
- tot.withinss: Total within-cluster sum of squares,
- betweenss: The between-cluster sum of squares
- size: The number of observations in each cluster



Fig(4.3) Represents K-means function return list

#### D. Results in R

The graphical representation shown below to have a better understanding of algorithm we implemented using R language. Clustering is performed using k-means, k-medoids (portioning around medoids), and clustering for large applications (CLARA), approaches and the resultant clusters are plotted using fviz\_cluster() function.

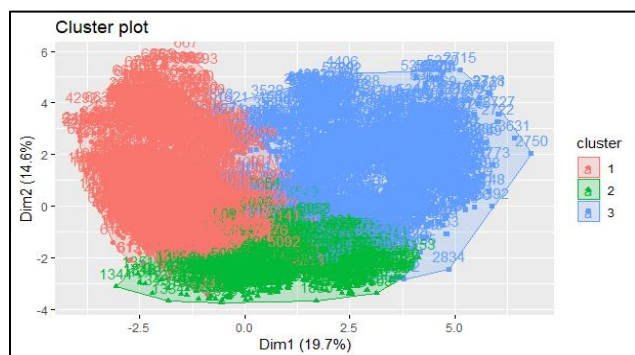
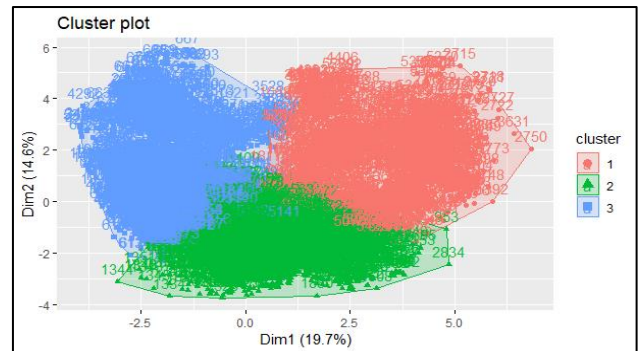


Fig (4.4) Cluster plot using k-means algorithm



Fig(4.5) Cluster plot using k-medoids algorithm

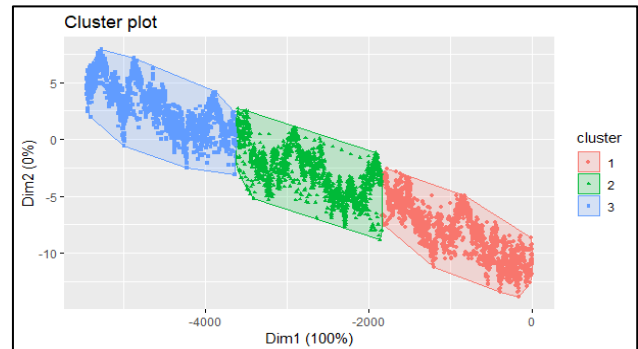


Fig (4.6) Cluster plot using CLARA algorithm

Silhouette Plot technique provides a graphical representation of how well each object has been classified. Each vertical line corresponds to an element.

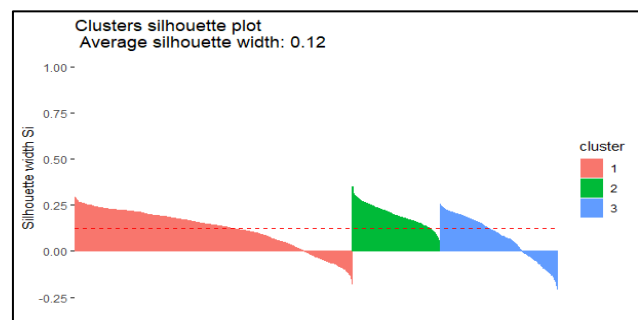
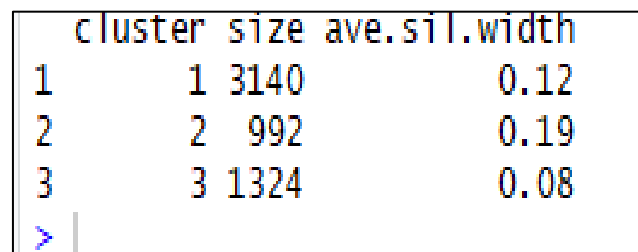


Fig (4.7) Silhouette Plot for various clusters



Fig(4.8) Represents Average Silhouette width

The below given is the elbow plot, is a method to validate optimal number of cluster. Also, help finding the appropriate number of clusters in a dataset.



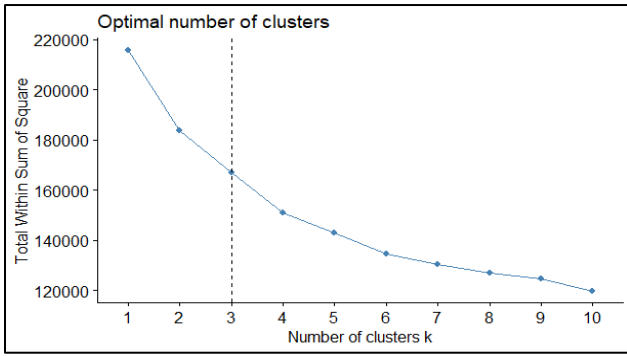


Fig (4.9) Optimal no: of cluster for elbow plot in R

#### E. Results in Python

When implemented the code with python programming language we have a graphical representation of plotting of centroid given below. It is apply random centroid to the data and using eculidean distance calculator assign each value to its closest cluster. In the process, everytime we finds the distance from the cluster and assign to nearest one. Moreover, this process continues untill error become zero and no more changes in centroid.

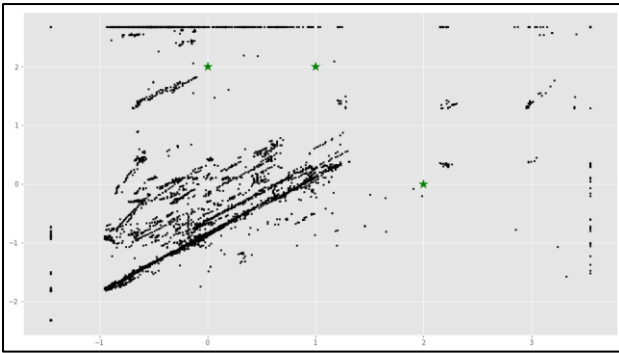


Fig (4.10) Represents plotting of centroid randomly

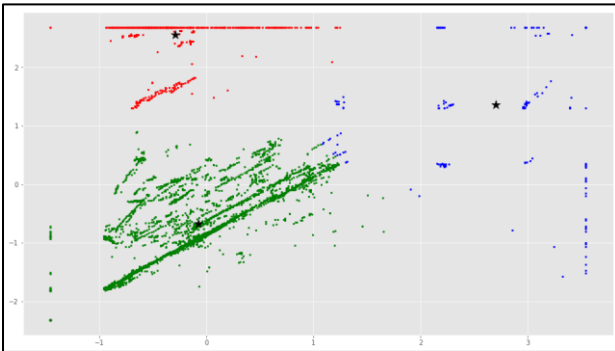


Fig (4.11) Represents assignment of new centroid

After creating separation of data in 3 various cluster, below graph represents data in 3D frame. It's showing separation of data in 3 different clusters. Centroid are represents as start in the center of the data and datapoints shows in red, green and blue color.

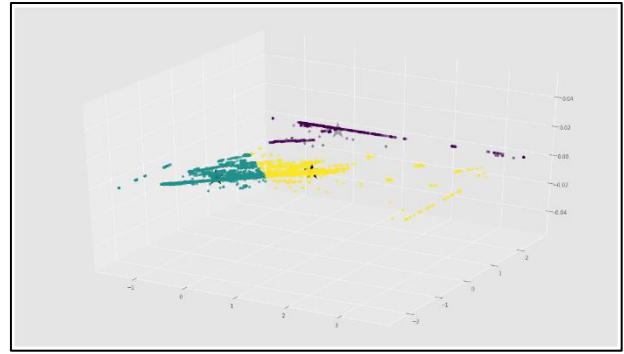


Fig (4.12) Represents clustering in 3D

Elbow graph representation in python and option value for K is 3.

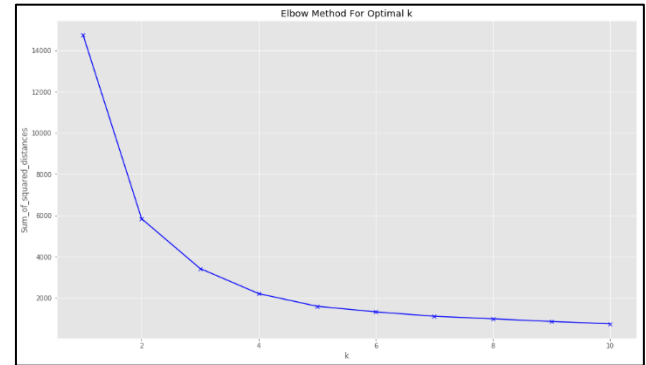


Fig (4.13) Optimal no: of cluster for elbow plot in Python

#### F. Time-complexity:

Time-complexity in python and R includes time for fitting an algorithm, predicting an algorithm, total time to run a code.

TABLE II  
TIME-COMPLEXITY IN PYTHON

In Python	Time (sec)
Time for custom model - Assigning each value to its closest cluster	1.728 s
Time for fitting a model	0.102 s
Time for predicting a model	0.001 s
<b>Total time to run a code</b>	<b>16.59 s</b>

TABLE III  
TIME-COMPLEXITY IN R

In R	Time (sec)
Time for k-means	0.84 s
Time for k-medoids	8.50 s
Time for CLARA	14.2 s
<b>Total time to run a code</b>	<b>23.9 s</b>

#### G. Tools & Libraries used for implementation:

Python: To implement the code we have made use of Python 3.6.1 version. We had made use of numpy library for multidimensional array used to store of same datatype. Pandas library provide high performance and used for data analysis tools. We make use of matplotlib.pyplot for comprehensive 2D/3D plotting and displaying in understandable manner. Seaborn is a visualization library based on matplotlib.

R: The code implemented in R 3.5.2 version. We made use of cluster library to find data in groups, ggplot2 to create decorative visualization and factoextra library to visualize the clusters.

## V. SUMMARY & FUTURE WORK

We studied the data, checking for data unbalancing, preparation, visualizing features and understanding relationships between various features. We are building the model with K-means, K-medoids and CLARA algorithm from R libraries, K-means algorithm with python libraries and custom model. Then selected the best cluster value k, compared and checked with elbow graph. Finally, conclude the results and ended with best one. Moreover, we worked on Amazon SageMaker to run python code and explore scope in cloud technology for Machine learning.

In the future work, we would like to explore more on unsupervised learning approach in Machine learning and select unlabeled dataset of Gigabyte, save on Amazon S3, run a model on Amazon Sage Maker, save and use a model for real time data.

## VI. REFERENCES

- [1]. Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). "A comparative study of efficient initialization methods for the k-means clustering algorithm". Expert Systems with Applications.
- [2]. Renjith, Shini, and C. Anjali. "A personalized mobile travel recommender system using hybrid algorithm." In Computational Systems and Communications (ICCSC), 2014 First International Conference on, pp. 12-17. IEEE, 2014
- [3]. Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.
- [4]. Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices." IEEE Transactions on Pattern Analysis and Machine Intelligence 24, no. 12 (2002): 1650-1654.
- [5]. Kovács, Ferenc, Csaba Legány, and Attila Babos. "Cluster validity measurement techniques." In a 6th International symposium of hungarian researchers on computational intelligence. 2005.
- [6]. David J. Ketchen, Jr; Christopher L. Shook (1996). "The application of cluster analysis in Strategic Management Research: An analysis and critique". Strategic Management Journal.
- [7]. Tupti M. Kodinariya, Dr. Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering". In International Journal of Advance Research in Computer Science and Management Studies. Volume 1, Issue 6, November 2013
- [8]. Selecting the number of clusters with silhouette analysis on KMeans clustering , Sklearn ([Link](#))
- [9]. Partitioning cluster analysis: Quick start guide - Unsupervised Machine Learning ([Link](#))

## CONTRIBUTIONS

**Rajaraman Ganesan:** He worked on data preparation and understanding of clustering algorithms. He applies algorithm like k-means, k-medoids and CLARA in R. His role is to find best value of cluster, compare and check with elbow graph in R. At the end, evaluate the best results after applying libraries. He also found the time-complexity to run each algorithm and total time to run a code in R.

**Vatsal Shah:** He worked on dataset understanding, data pre-processing. His role is to understand and apply k-means algorithms with custom models. It includes create equations, find distance and select best value of cluster K number and evaluate results. He created output.txt file to write all results of the process in python. Moreover, found the time-complexity for kmeans in fitting and predicting a mode, total time to run a code. Lastly, he used Amazon SageMaker to explore cloud technology in Machine learning and run python code to find the results.