ABSTRACT

Clustering technique is critically important step in segmentation. It is a multivariate procedure quite suitable for segmentation applications in the market forecasting and planning research. This is a comprehensive report of k-means clustering technique .

A machine learning model determining the Customer Segmentation or classifying customers in different levels and to Explore the general distribution of the data to get a sense of Male vs. Female customers and how their income, age, and spending scores are similar or different followed by which gender has a higher income and which gender has a higher average spending score.

The model developed was an intelligent tool which upon receiving inputs from sales data records visualizes the data and divides the data into five main income groups. There are 5 rough clusters obtained by the data Low income, low spending score Low income, high spending score Mid income, medium spending score High income, low spending score and High income, high spending score

The model was successfully implemented and tested over a period of one month. A total of n = 200, customer, were tested for observations which were then divided into k = 5 distinct groups. The classification was based on nearest mean. Results were quite encouraging and had shown high accuracy.

Index Terms - Cluster analysis, customer segmentation, K means.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Clustering is a statistical technique much similar to classification. It sorts raw data into meaningful clusters and groups of relatively homogeneous observations. The objects of a particular cluster have similar characteristics and properties but differ with those of other clusters. The grouping is accomplished by finding similarities among data according to characteristics found in raw data. The main objective was to find optimum number of clusters. There are two basic types of clustering methods, hierarchical and non-hierarchical. Clustering process is not one time task but is continuous and an iterative process of knowledge discovery from huge quantities of raw and unorganized data. For a particular classification problem, an appropriate clustering algorithm and parameters must be selected for obtaining optimum results. Clustering is a type of explorative data mining used in many application oriented areas such as machine learning, classification and pattern recognition. In recent times, data mining is gaining much faster momentum for knowledge based services such as distributed and grid computing. For clustering method, the most important property is that a tuple of particular cluster is more likely to be similar to the other tuples within the same cluster than the tuples of other clusters.

2. EXISTING METHOD

Finding the right market for customers in major industries has attracted a huge attention for several years. Customer Segmentation processes have been performed using one of these methodologies.

- Hierarchical Clustering
- CRM

DISADVANTAGES

The main problem with these algorithms was that they performed grouping only once which caused to get less accuracy as the data has been incorrectly grouped.

3. PROPOSED METHOD WITH ARCHITECTURE

Identifying right customer and providing right service at right time and treating different types of customers differently is the key to success in business. So, a predictive model will be used to segregate customers into different groups based on their transactional data. Once the customers are segregated then their associative buying pattern are identified to enhance the profit for the organization future coming customer.

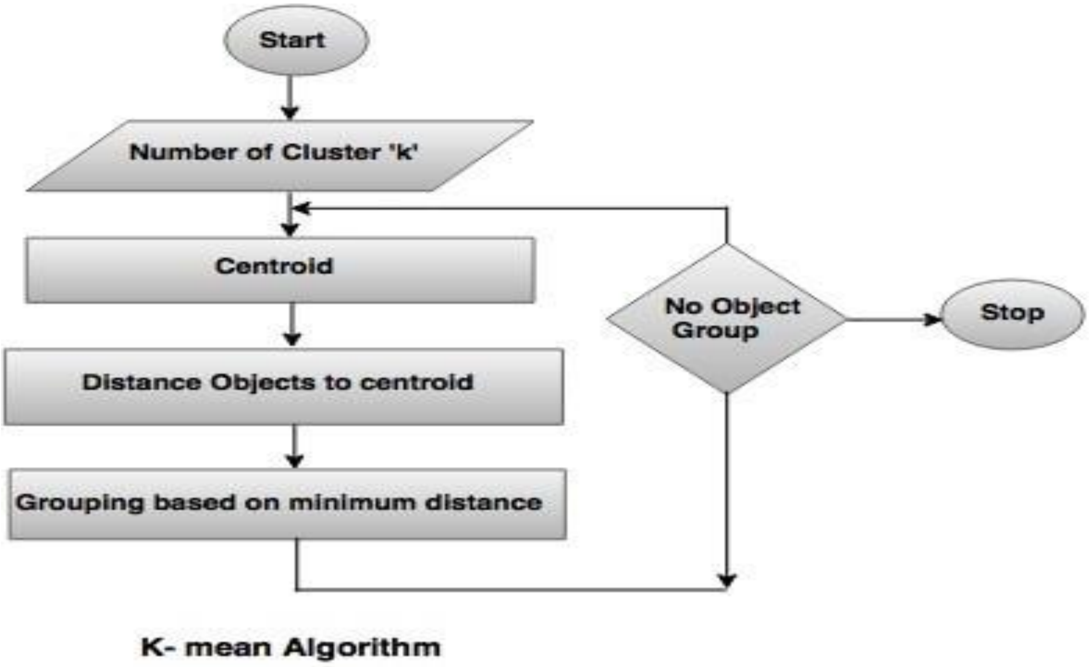


Fig1.1 K-means Architecture

ADVANTAGES

- Proposed method groups data into a specific number of clusters as it uses the elbow method to find the optimum number of clusters.
- It is much accurate because the final output is selected based on the elbow method instead of randomly selecting like in the previous cases.

4. METHODOLOGY

In order to identify the target customers Clustering technique can be used for cluster analysis. Clustering is defined as to group data in clusters/segments so that data within segment are similar while data across the segments are dissimilar. Various techniques can be used for clustering like k means, hierarchical, grid based model based technique. In this paper we proposed to use K-means technique for customer segmentation due its following advantages:

- This technique suits for the data with numeric features and often terminates at local optimum.
- It is highly scalable and efficient for large data sets.
- It is fast in modelling and its result is more understandable.

K means Algorithms is discussed as follows

The k-means clustering algorithm divides the n records into k segments of records called clusters where $k \leq n$, so as to minimize the distances between records within a particular cluster.

Step 1: Choose K points at random as segment centres.

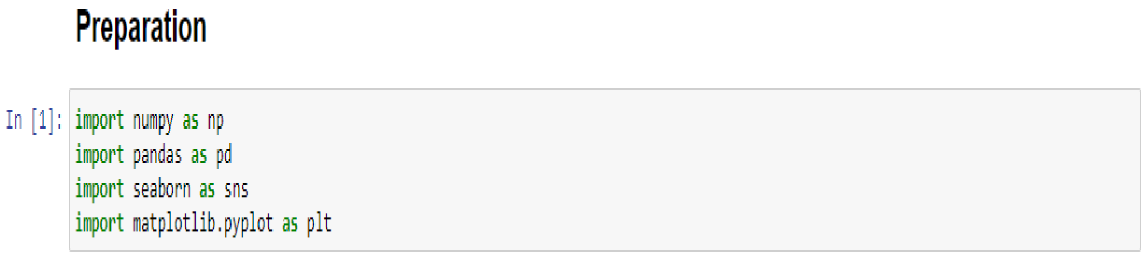
Step 2:Assign each record to its closest segment centre using certain distance measure (usually Euclidean or Manhattan).

Step 3: Calculate the centroid of each segment, use it as the new segment centre (one measure of centroid is mean).

Step 4: Go back to Step 2, stop when segment centres do not change any more.

5. IMPLEMENTATION

It is implemented in python using jupyter notebook. The following libraries are need and imported accordingly. These are further used for the visualization of the data.



Preparation

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
```

Fig 1.2 Installing Libraries

The gender distribution from the dataset is represented and from that it is clear that there are more women compared to men.

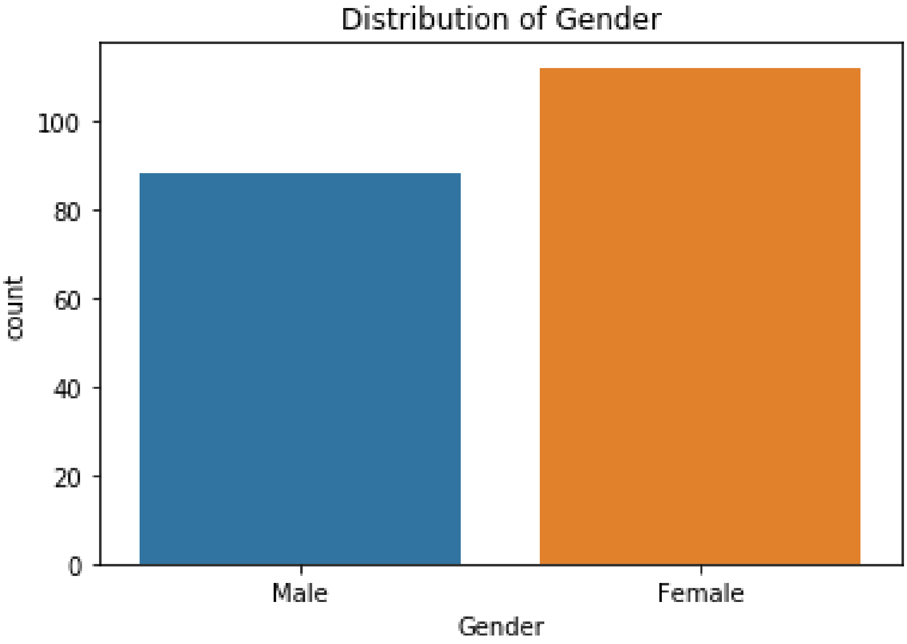


Fig 1.3 Gender Visualization

The annual income distributions are represented using matplotlib.

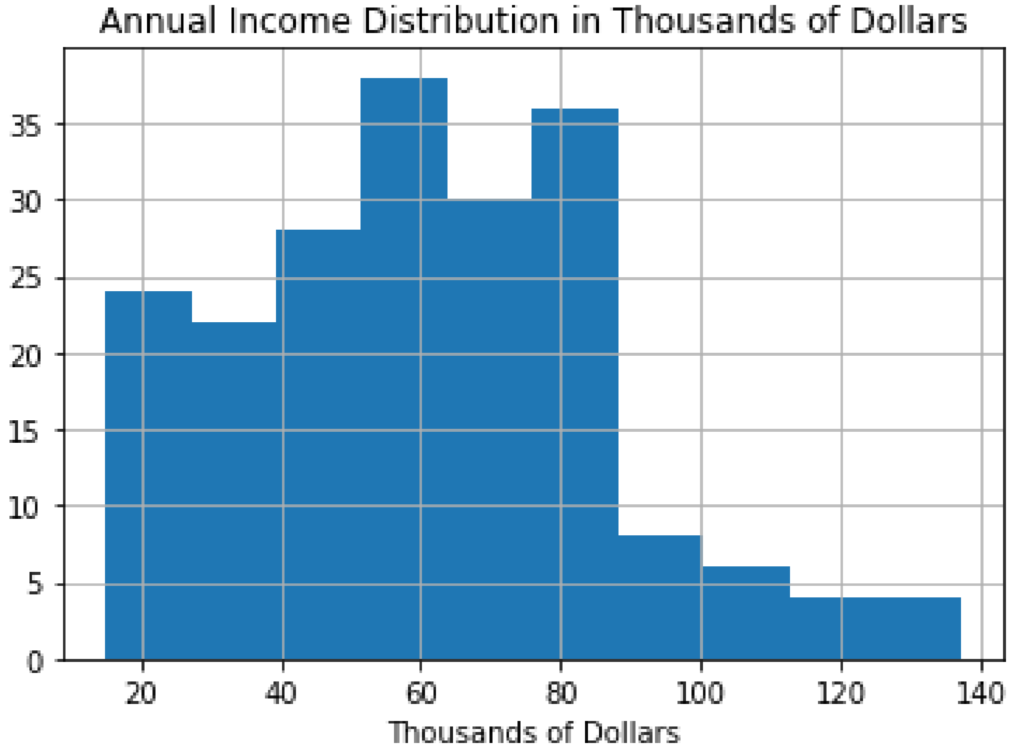


Fig 1.4 Annual Income Distribution

The distribution of age is represented in the bar plot as follows

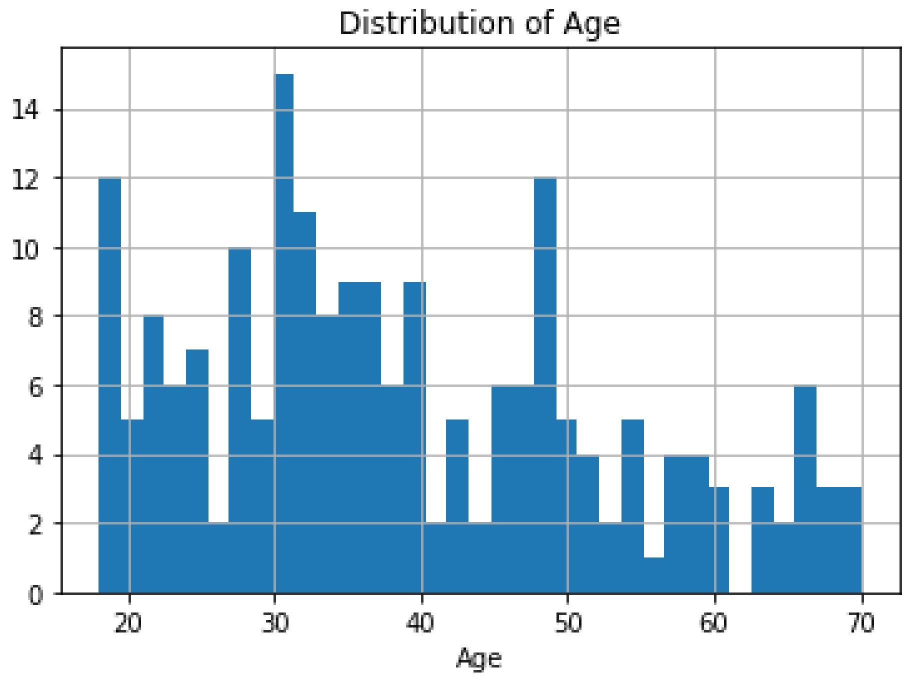


Fig 1.5 Age Visualization

Bar plot on Income distribution by gender

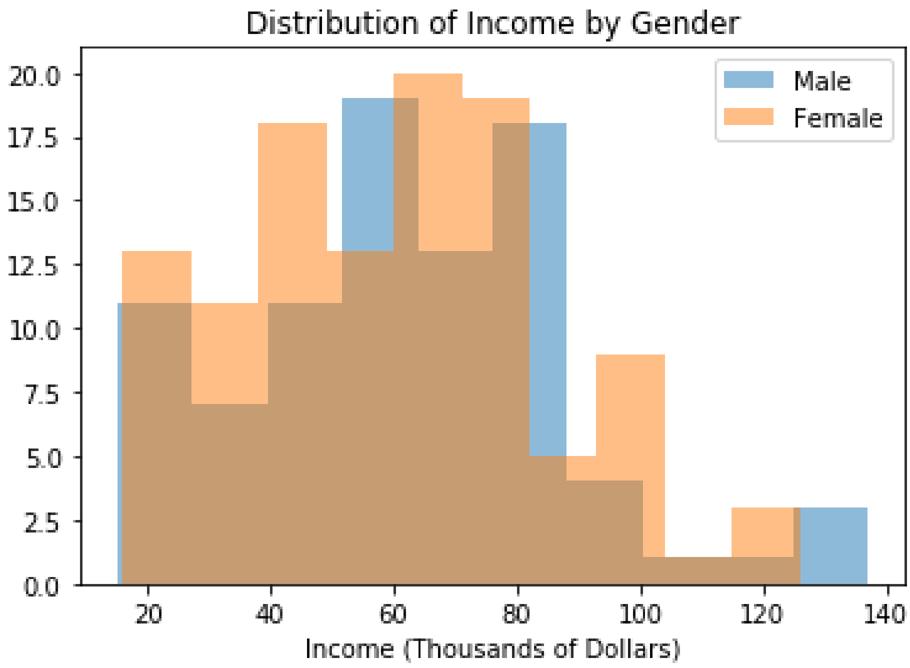


Fig 1.6 Distribution of income by Gender

The optimum number of clusters are found by the elbow method. It is represented as follows.

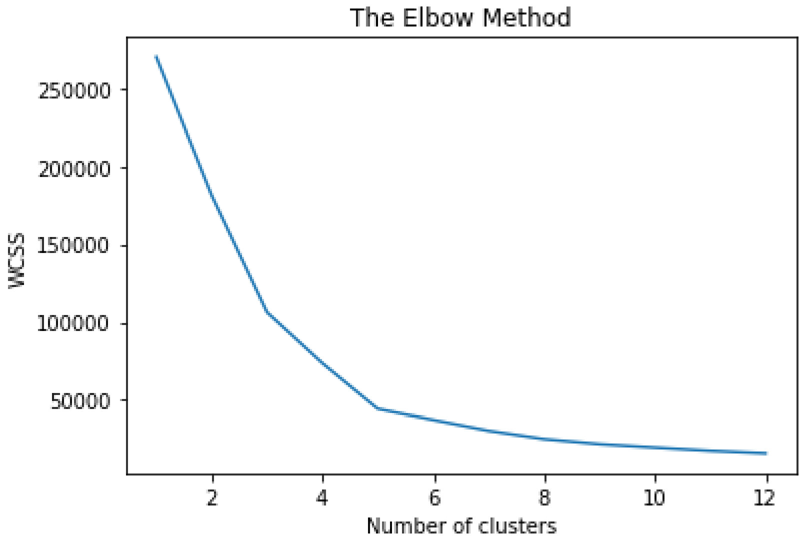


Fig 1.7 Finding the number of clusters

Final result of the segregated customers is shown by the scatterplot.

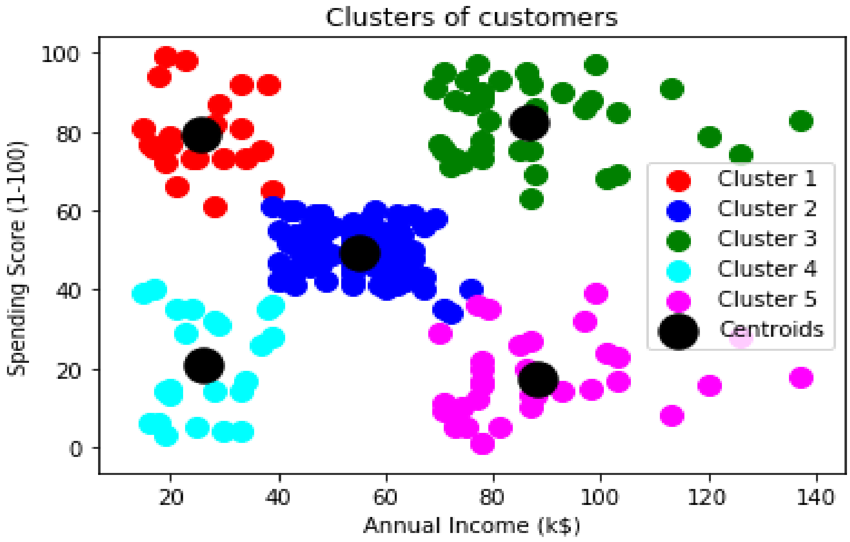


Fig 1.8 Final Result

6. CONCLUSION

Customer segmentation is a way to improve communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built. Customer Segmentation needed to get potential customers used to increase profits. Potential customer data can be used to provide service the characteristics of customer including ecommerce services as a media buying and selling online.

1. There are 5 rough clusters obtained by the data
   - Low income, low spending score
   - Low income, high spending score
   - Mid income, medium spending score
   - High income, low spending score
   - High income, high spending score
2. Based on these data, the following hypotheses could be tested:
   - Marketing cheaper items to women to see if they purchase more frequently or more volume.
   - Marketing more to younger women because their spending score tends to be higher.
   - Thinking up new ways to target advertising, pricing, branding, etc. to the older women (older than early 40s) who have lower spending scores.
   - Figure out a way to gather more data to build a data set that has more features. The more features, the better understanding of what determines Spending Score. Once Spending Score is better understood, we can understand what factors will lead to increasing Spending Score, thus lead to greater profits.