MALL CUSTOMERS

Introduction to big data and data analytics

(20B12CS333)

Project Report

Submitted by:

Anish Khare (9919103195)

Rana Aishwarya Pratap Singh (9919103198)

Piyush Gupta (9919103203)

Submitted to:

Dr. Neeraj Jain

Department of CSE/IT

Jaypee Institute of Information Technology University ,Noida

Problem statement:

To segment customers into groups with different income age and spending scores in a mall and derive an understanding of it from various charts and exploratory analysis.

The clustering of customers will provide valuable information on which group has higher spending score and which group mall should target to increase profits. Model is trained using the K-Means algorithm by which we can segment customers into n different groups.

Introduction:

In this project, we have used a dataset about the customers available in the mall. Identifying potential target customers is essential for any business. This project shows how big data analytics can be used to segment customers according to income age and spending scores at a mall. We have plotted various graphs to visualize the data and derive an understanding of it. By clustering customers, we can determine which groups have higher spending scores, and which groups malls should set their sights on to increase profits.

This project analyzes the relationship between customers' age, income, and expenditures at a mall in order to understand the perfect consumer base to target. This project uses an iterative clustering algorithm called K-Means to cluster data.K-Means algorithm is used to segment customers into n different groups.

Proposed work with tools and datasets used:

Proposed work:

1) Find an appropriate dataset according to our requirement
2) Clean data and be left with only the data which is relevant for our study
3) Plot distribution of annual income, age and spending scores among the customers
4) Plot gender distribution of customers in the mall
5) Plot a 3d scatter plot with age, annual income and spending score as x, y and z axes.
6) Use elbow method and silhouette method to find the number of clusters
7) Apply K-mean clustering to the data
8) Derive a conclusion from our analysis

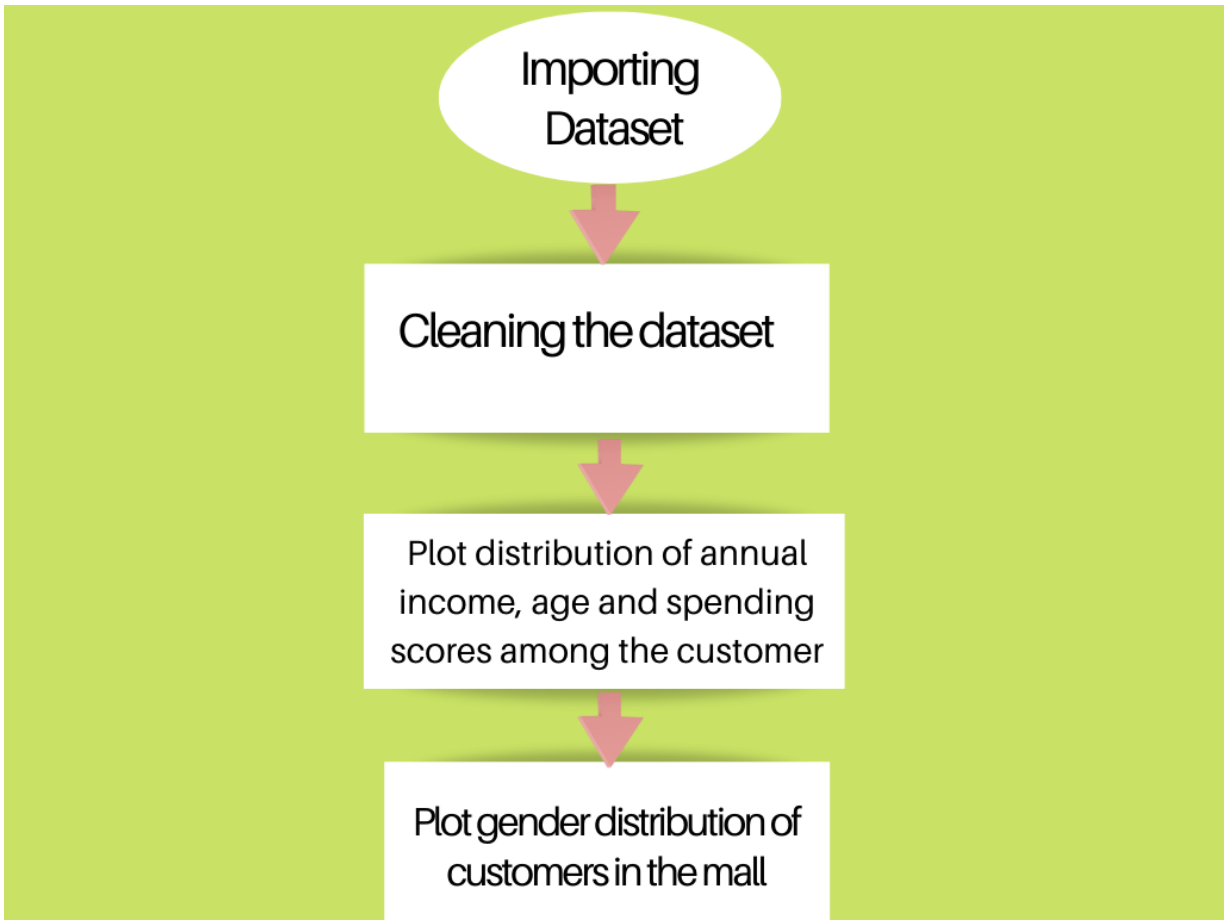
Tools:

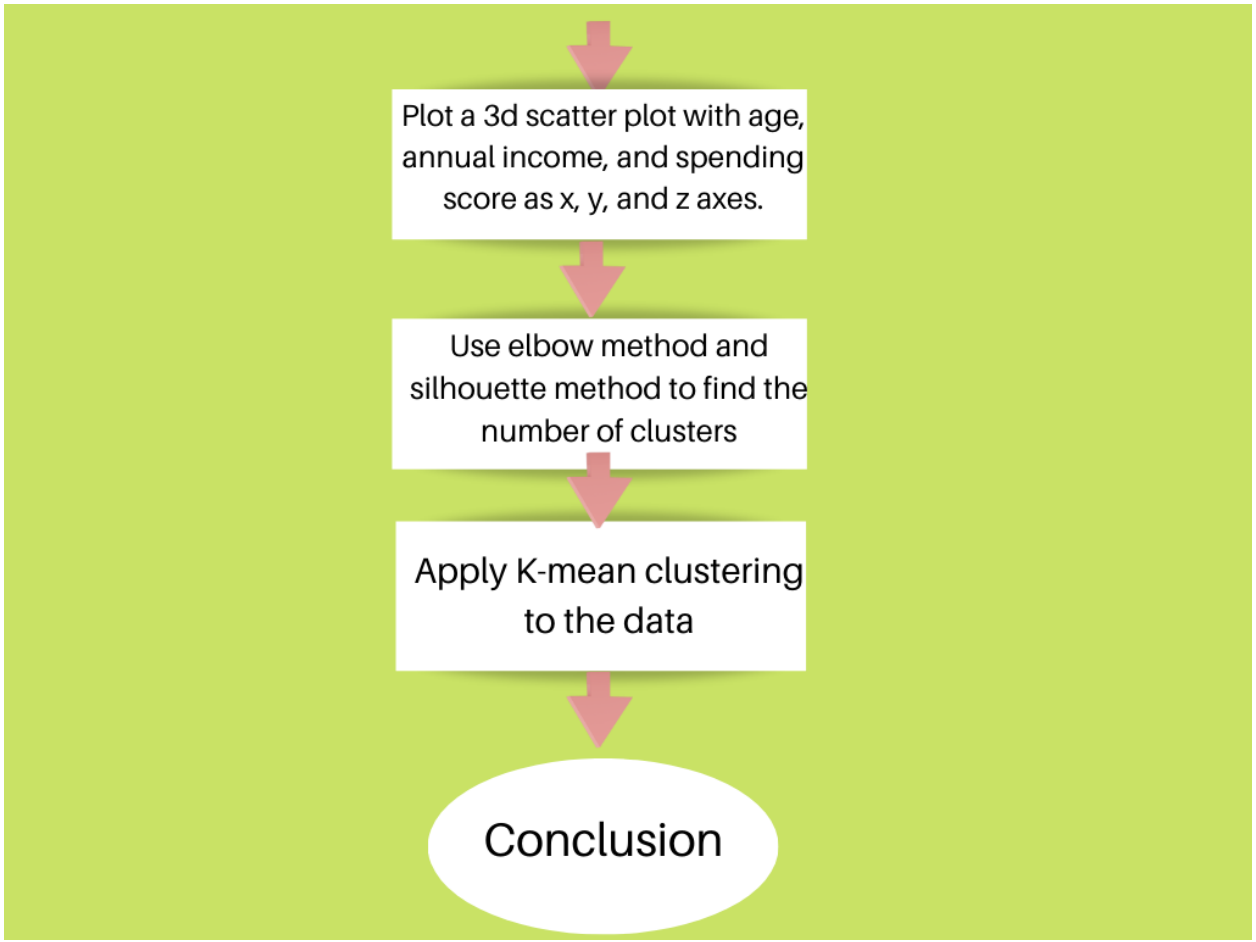
1) Jupyter Notebook and google colab for compiling code and visualizing data
2) Numpy for linear algebra
3) Pandas for data preparation
4) Matplotlib, seaborn, plotly for data visualization
5) Sklearn to apply k-mean algorithm and silhouette method

Dataset

We have used the mall_customers dataset from Kaggle.
The dataset is based on small customers, and it contains 200 rows and 5 columns. This dataset will be used in the data analysis and machine learning project.

Workflow Diagram of Proposed Work

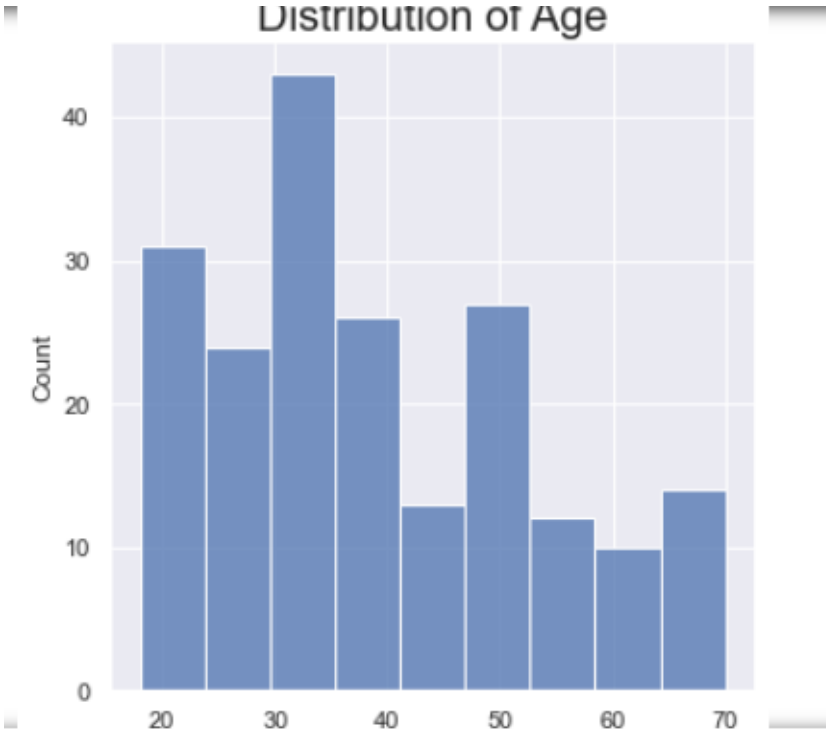




Results:

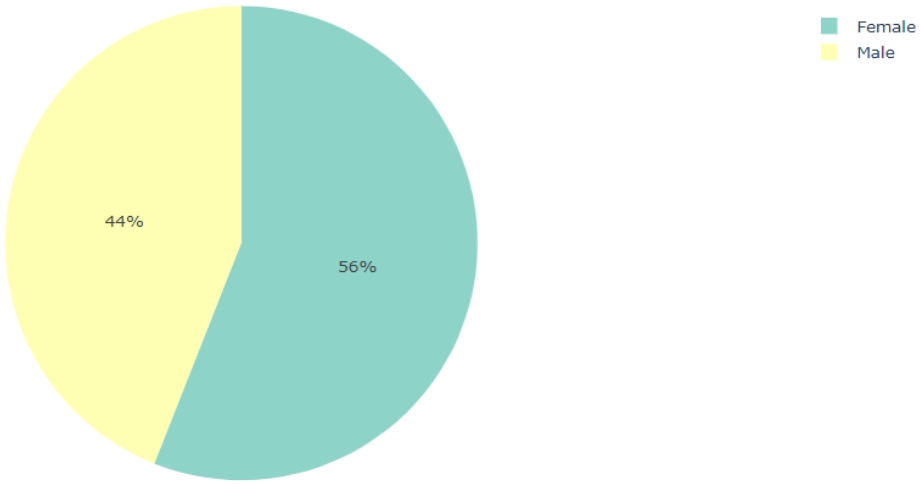
1) We found out that people with income in the middle bracket were the most common type of customers
2) We found out that most of the customers belong to relatively younger age groups
3) We found out that most customers belonged to the middle of the spending score scale
4) We found out that 44% of the customers were male and 66% were female
5) We applied the k-mean algorithm successfully and divided the dataset into 5 clusters. The yellow and purple clusters in the above graph are clearly higher spending clusters compared to others. Hence it would be beneficial if the mall tries to target these customers. Also orange and pink clusters spend less than other clusters and the mall can try new methods for attracting customers from these clusters.

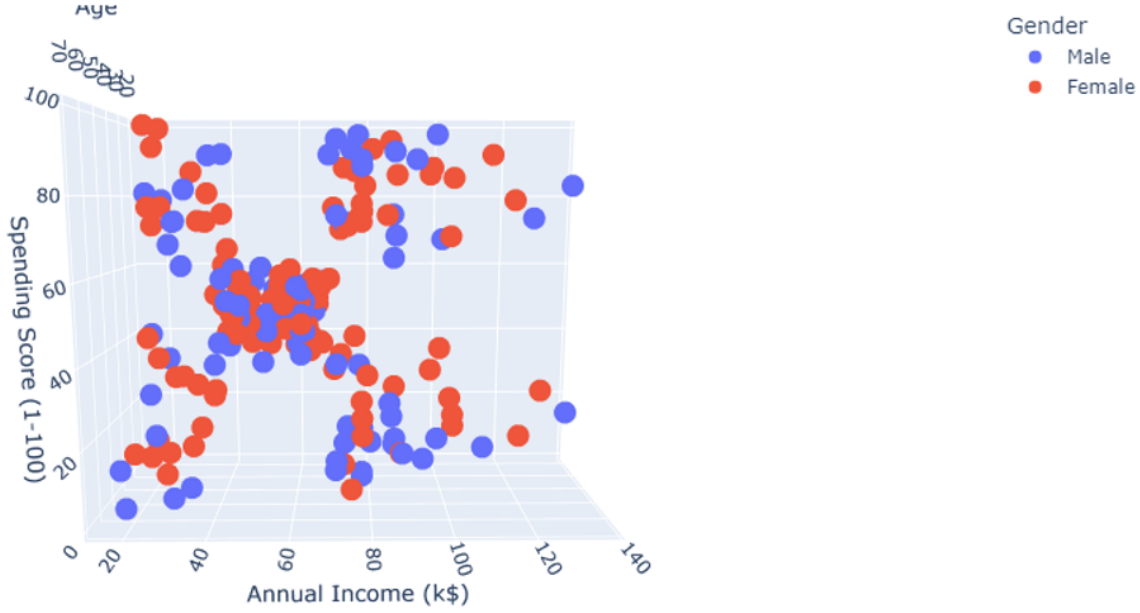
Snapshots of results:

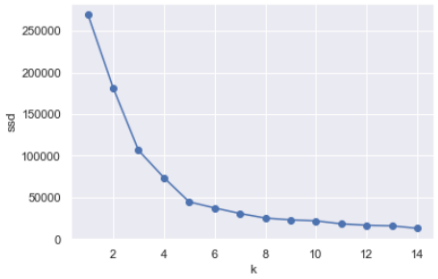

Distribution of Age
Count
0
10
20
30
40
20
30
40
50
60
70


Distribution of Spending Score (1-100)
Count
0
10
20
30
40
0
20
40
60
80
100
Range of Spending Score (1-100)

Distribution of Gender on dataset

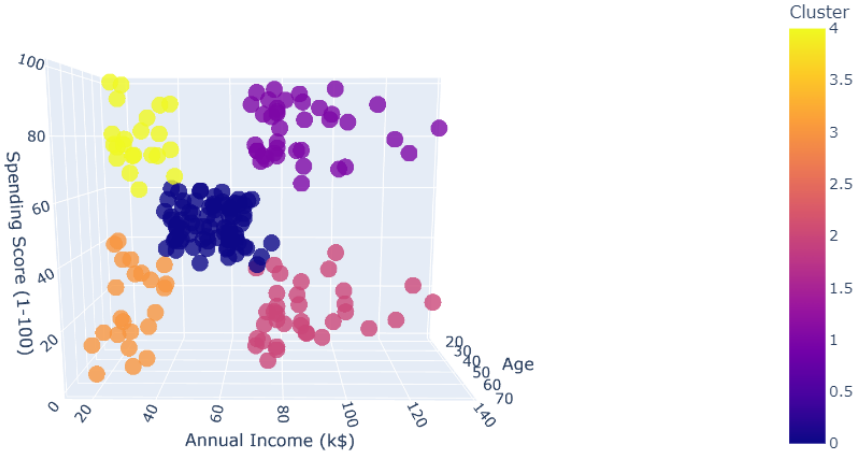






In [39]:
```
for n_cluster in range(2, 14):
    kmeans = KMeans(n_clusters=n_cluster).fit(X)
    label = kmeans.labels_
    sil_coeff = silhouette_score(X, label, metric='euclidean')
    print("cluster={}, The silhouette Coeff = {}".format(n_cluster, sil_coeff))
```

```
cluster=2, The silhouette Coeff = 0.2968969162503008
cluster=3, The silhouette Coeff = 0.46761358158775435
cluster=4, The silhouette Coeff = 0.4931963109249047
cluster=5, The silhouette Coeff = 0.553931997444648
cluster=6, The silhouette Coeff = 0.53976103063432
cluster=7, The silhouette Coeff = 0.5270287298101395
cluster=8, The silhouette Coeff = 0.4564394045323282
cluster=9, The silhouette Coeff = 0.4565077334305076
cluster=10, The silhouette Coeff = 0.44791983625403836
cluster=11, The silhouette Coeff = 0.4489710248005492
cluster=12, The silhouette Coeff = 0.4374829582024546
cluster=13, The silhouette Coeff = 0.3995045991135986
```



In [41]:
```
finaldata.head()
```

Out[41]:

	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Cluster
0	Male	19	15	39	3
1	Male	21	15	81	4
2	Female	20	16	6	3
3	Female	23	16	77	4
4	Female	31	17	40	3

Conclusion:

In this project the K-Means algorithm has been implemented successfully and the dataset was clustered into 5 groups successfully. This grouping will help the mall to devise different methods for targeting high spending customers and also attracting low spending customers for increasing profits. We also plotted various graphs and charts and accomplished all of the work we proposed to derive an understanding of the customers that go to the mall.

References:

- Kaggle: Your Home for Data Science
- https://numpy.org/
- Stack Overflow - Where Developers Learn, Share, & Build Careers
- https://matplotlib.org/
- https://plotly.com/
- scikit-learn: machine learning in Python — scikit-learn 1.0.1 documentation
- https://seaborn.pydata.org/
- pandas - Python Data Analysis Library (pydata.org)