SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF NETWORKING AND COMMUNICATIONS
18CSP109L / 18CSP110L- MAJOR PROJECT / INTERNSHIP

Online customer profiling and segmentation using machine learning algorithms.

AAKRITI NAG : RA1811029010038
ROHIT RANJAN : RA1811029010039
Batch ID: NWC039038
Guide : Dr. C.N.S. Vinoth Kumar

Table of contents

- Abstract
- Introduction - Problem Statement/Objectives
- Literature survey
- Proposed block diagram / Working of each module
- UML diagrams / ER diagrams/Use case diagram/Activity diagram
- Algorithms Used
- Hardware and software used
- Implementation and Output
- References

Abstract

Identifying the right market for products is important for targeting and improving business . In this project, we use unsupervised machine learning algorithms to implement customer segmentation in Python. We will see the descriptive analysis of our data and then implement several versions of the K-means algorithm. Furthermore, through the data collected, we can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, we can strategize the marketing techniques more efficiently and minimize the possibility of risk to the investment.

INTRODUCTION

The practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. Customer Segmentation is the process of division of the customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

Objectives

- Identify the potential customer base for selling the product using segmentation.
- Implement Clustering Algorithms to group the customer base.
- Use the elbow method to find optimal results.

Literature Survey

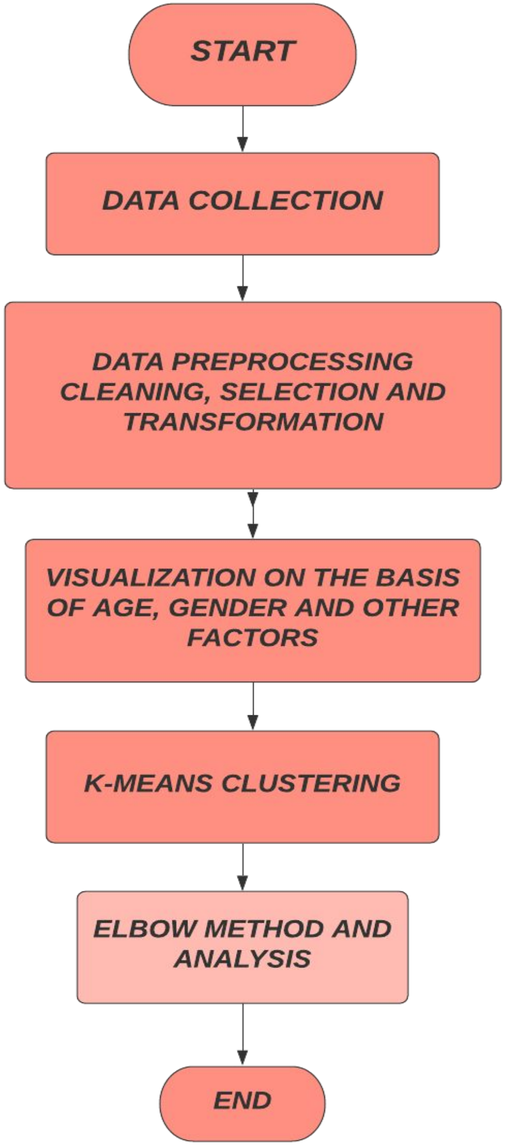
Sl. No.	Journal	Contribution	Inference
1.	Mediana Aryuni; Evaristus Didik Madyatmadja; Eka Miranda Published in: 2018 International Conference on Information Management and Technology (ICIMTech), “Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering ” in IEEE.	Built the cluster model and calculated the number of clusters using K-means clustering algorithm. The clusters were made based on similar characteristics between them, and the resultant clusters can be used to group customers based on their buying behaviour.	This research built clustering models on customer profile data based on their usage of Internet Banking for customer segmentation in XYZ bank using K-Means method and K-Medoids method. The performance of clustering result was measured and compared.

SI. No	Journal	Contribution	Inference
2.	Chinedu Pascal Ezenkwu , Simeon Ozuomba , Constance Kalu on “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”	A MATLAB program of the k-Means algorithm was developed (available in the appendix) and the program is trained using a z-score normalised two-feature dataset of 100 training patterns acquired from a retail business. The features are the average amount of goods purchased by customer per month and the average number of customer visits per month.	The ability of any business to understand each of its customers’ needs will earn it greater leverage in providing targeted customer services and developing customised marketing programs for the customers. This understanding can be possible through systematic customer segmentation.

Sl. No.	Journal	Contribution	Inference
3.	Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury. Published in: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), “Customer Segmentation using K-means Clustering” in IEEE.	3 different clustering algorithms (k-Means, Agglomerative, and Mean shift) are been implemented to segment the customers and finally compare the results of clusters obtained from the algorithms, opted for internal clustering validation rather than external clustering validation, which depends on some external data like labels. Internal cluster validation can be used for choosing clustering algorithm which best suits the dataset and can correctly cluster data into its opposite cluster.	The silhouette score for the three algorithms applied in this paper, the graph shows there is not much significant difference in K-means and Agglomerative clustering. Hence, these two algorithms were able to cluster our data well than Mean shift algorithm as displayed by the low value of silhouette score.

SI. No.	Journal	Contribution	Inference
4.	Kishana R. Kashwan, Member, IACSIT, and C. M. Velu. Published in : International Journal of Computer Theory and Engineering, Vol. 5, No. 6, December 2013, "Customer Segmentation Using Clustering and Data Mining Techniques".	This research paper is a comprehensive report of k-means clustering technique and SPSS Tool to develop a real time and online system for a particular super market to predict sales in various annual seasonal cycles. The model developed was an intelligent tool which received inputs directly from sales data records and automatically updated segmentation statistics at the end of day's business.	The analyses at the end provided further illustrations of using cluster method for market segmentation for forecasting. Computing based system developed was an intelligent and it automatically presented results to the managers to infer for quick and fast decision making process.

Block Diagram



START

DATA COLLECTION

DATA PREPROCESSING CLEANING, SELECTION AND TRANSFORMATION

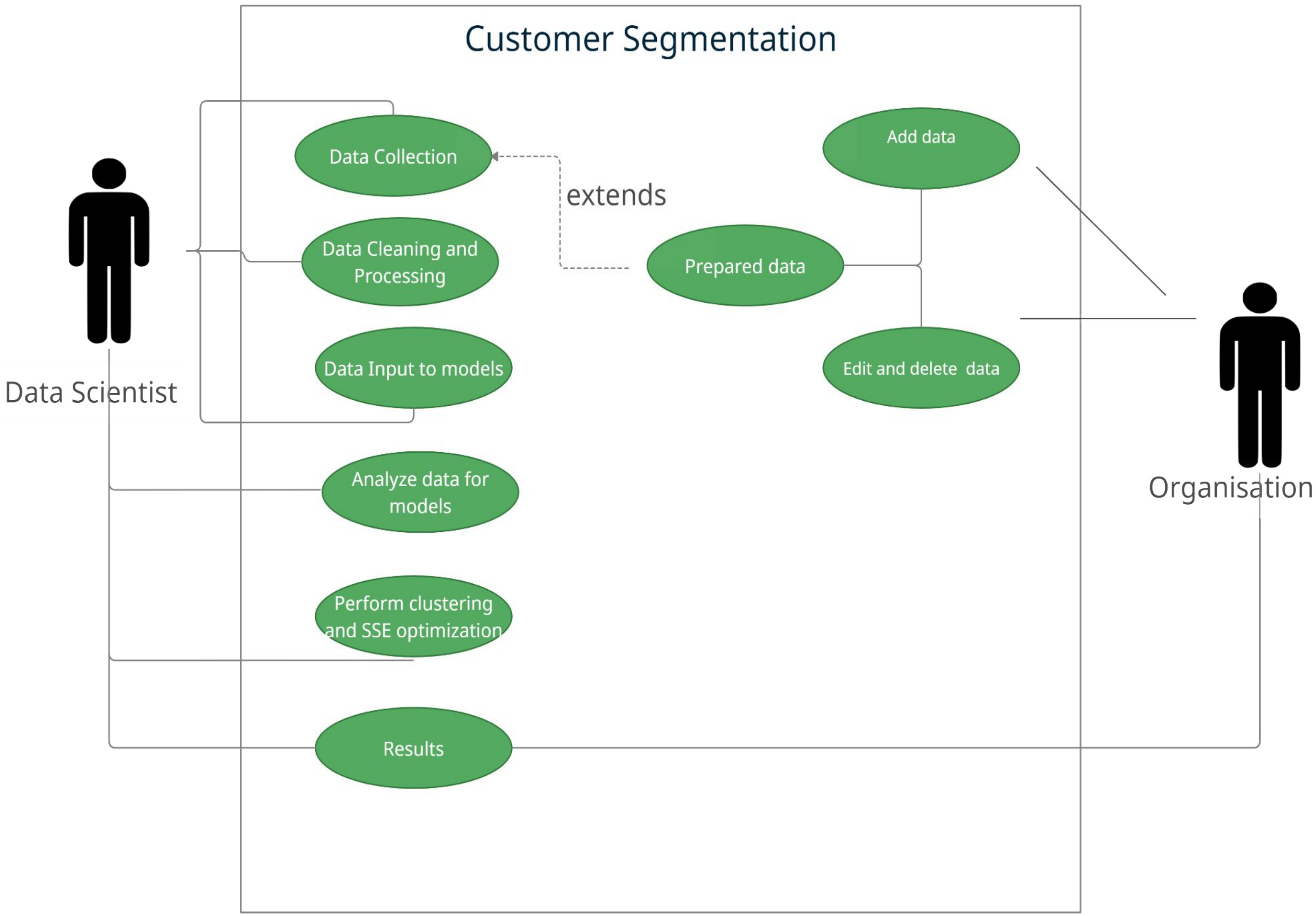
VISUALIZATION ON THE BASIS OF AGE, GENDER AND OTHER FACTORS

K-MEANS CLUSTERING

ELBOW METHOD AND ANALYSIS

END

Use-case Diagram

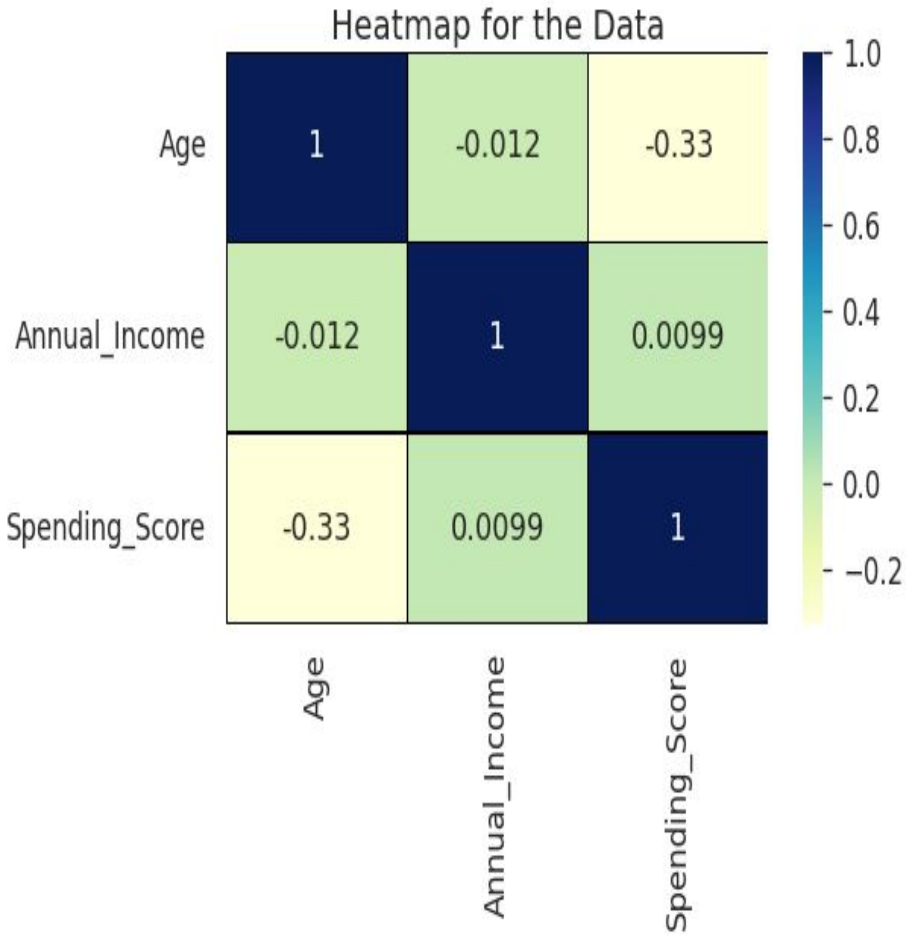


Algorithms Used

- Correlation
- K-means clustering algorithm
- SSE Optimization

Correlation

Correlation analysis in research is **a statistical method used to measure the strength of the linear relationship between two variables and compute their association**. A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.



K-means Clustering

K-Means Clustering is an unsupervised machine algorithm, which groups the unlabeled dataset into different clusters. It is a clustering algorithm that aims to partition n observations into k clusters.

An iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties.

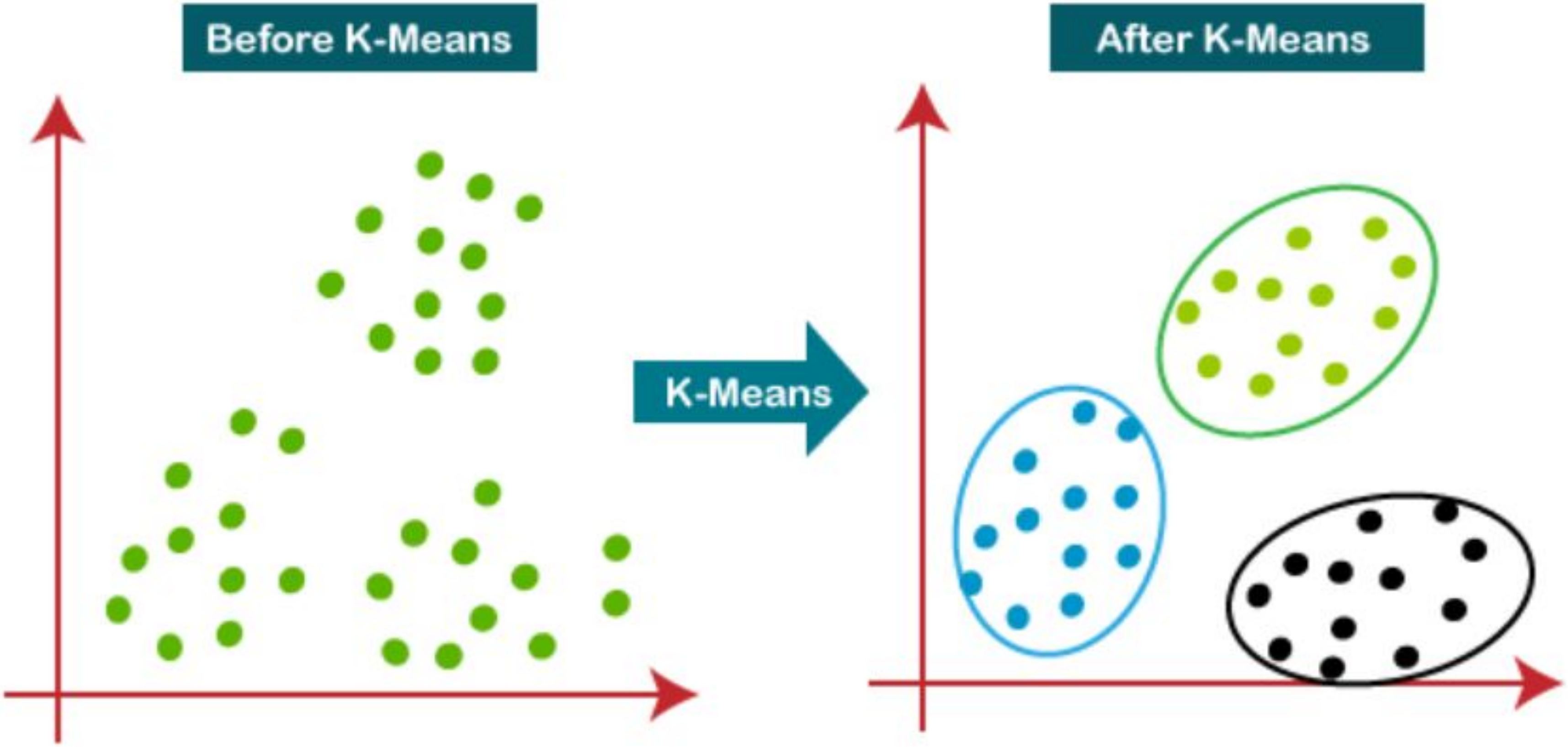
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

K-means clustering

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

K-means clustering



SSE Optimization

The Elbow method is a visual method to test the consistency of the best number of clusters by comparing the difference of the sum of square error (SSE) of each cluster, the most extreme difference forming the angle of the elbow shows the best cluster number.

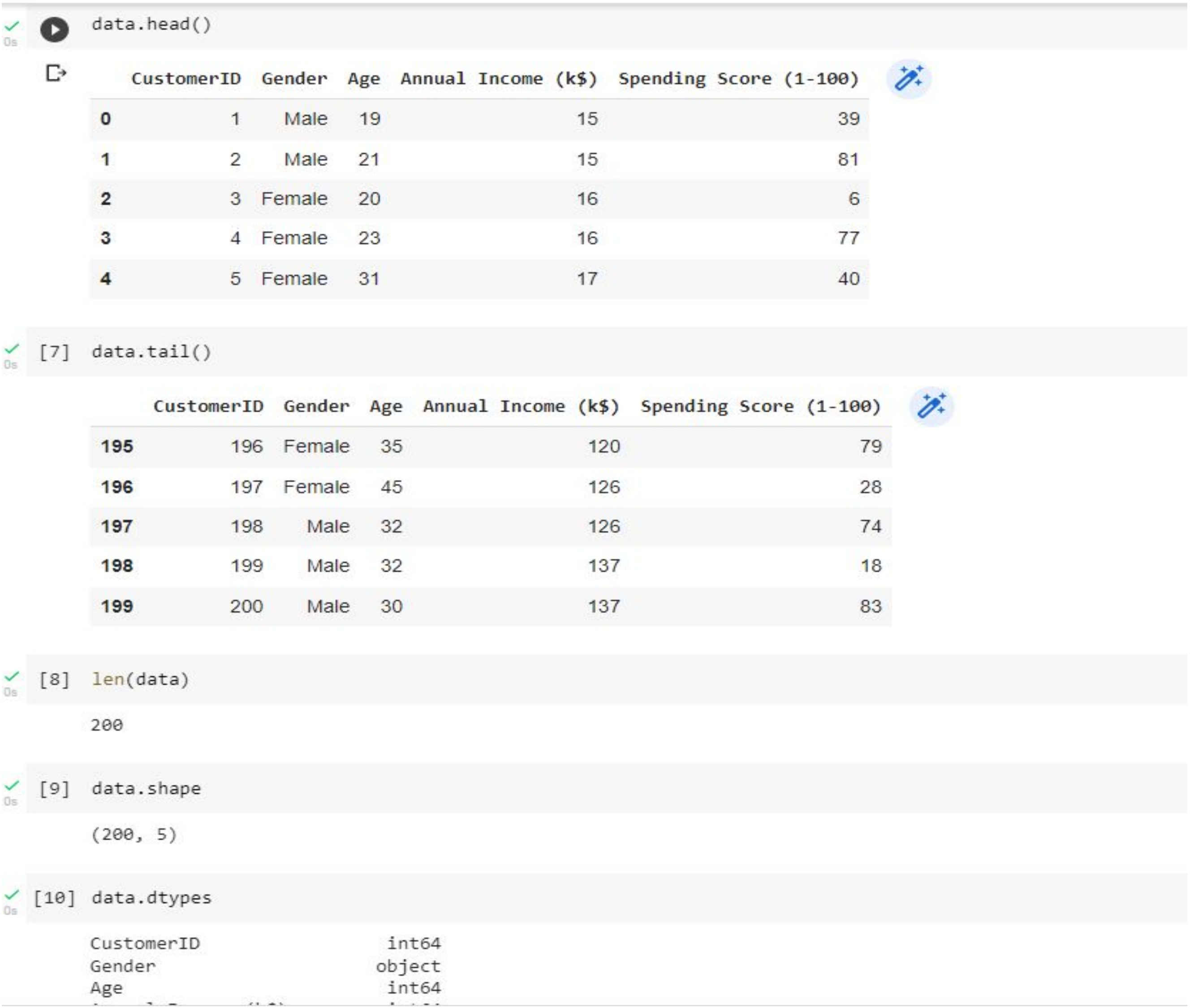
In cluster analysis, the elbow method is **a heuristic used in determining the number of clusters in a data set**. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

Software Used

- Jupyter Notebook
- Pandas
- Numpy
- Matplotlib
- Scikit Learn
- Seaborn
- Google Colab

Implementation & Output

Data Exploration



```
data.head()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

```
[7] data.tail()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

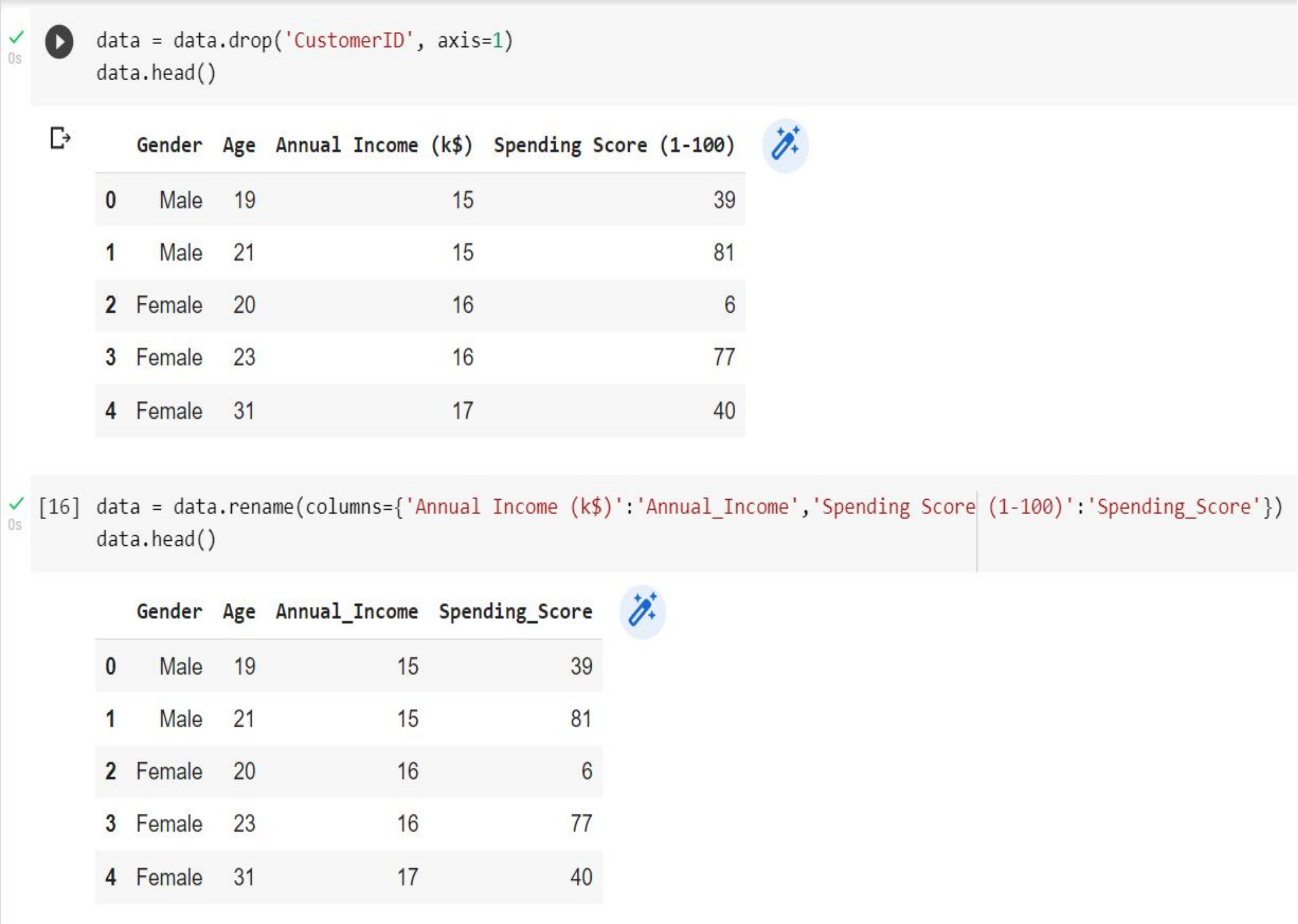
```
[8] len(data)

200
```

```
[9] data.shape

(200, 5)
```

```
[10] data.dtypes

CustomerID                 int64
Gender                    object
Age                        int64
```

Implementation and Output

Data Cleaning



```
data = data.drop('CustomerID', axis=1)
data.head()
```

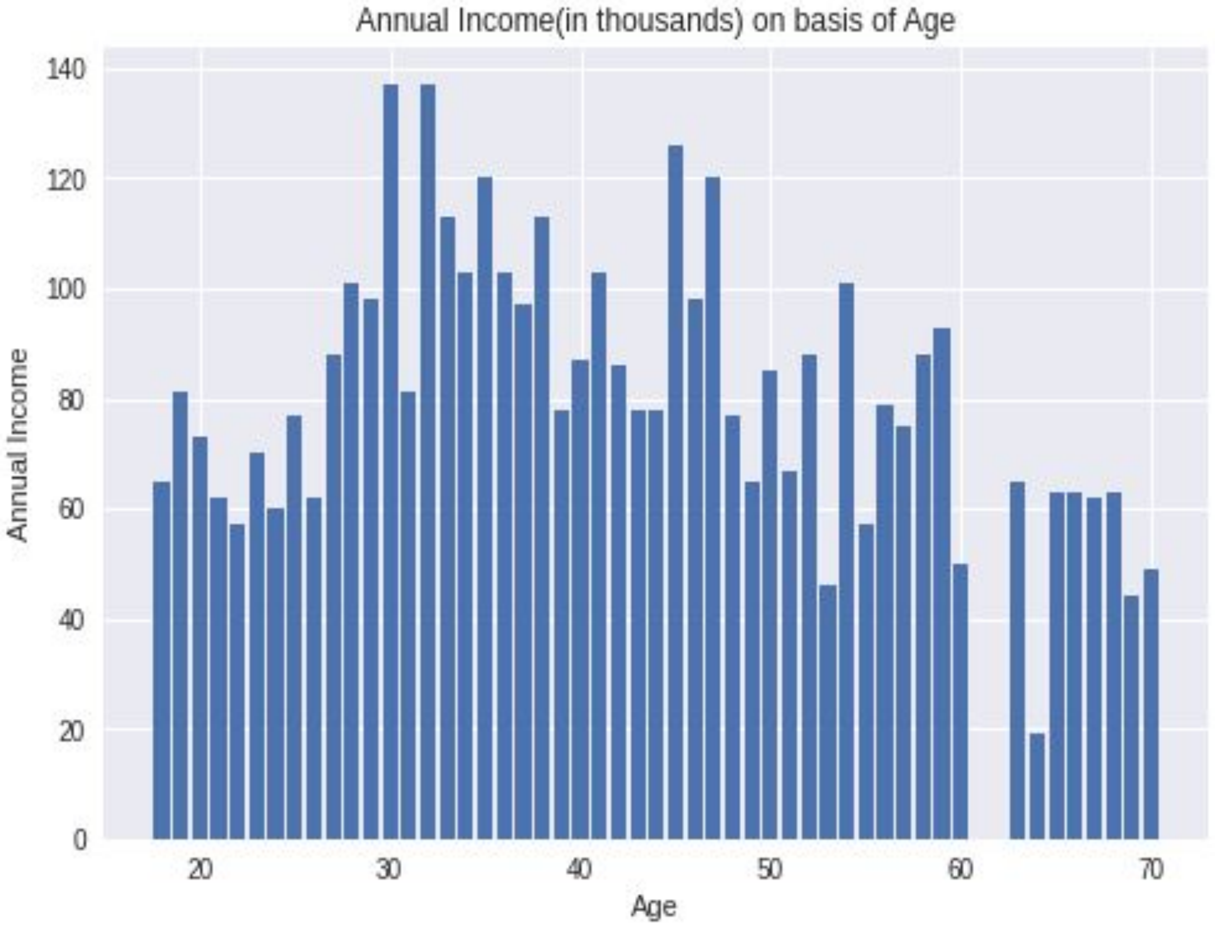
	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

```
data = data.rename(columns={'Annual Income (k$)':'Annual_Income','Spending Score (1-100)':'Spending_Score'})
data.head()
```

	Gender	Age	Annual_Income	Spending_Score
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

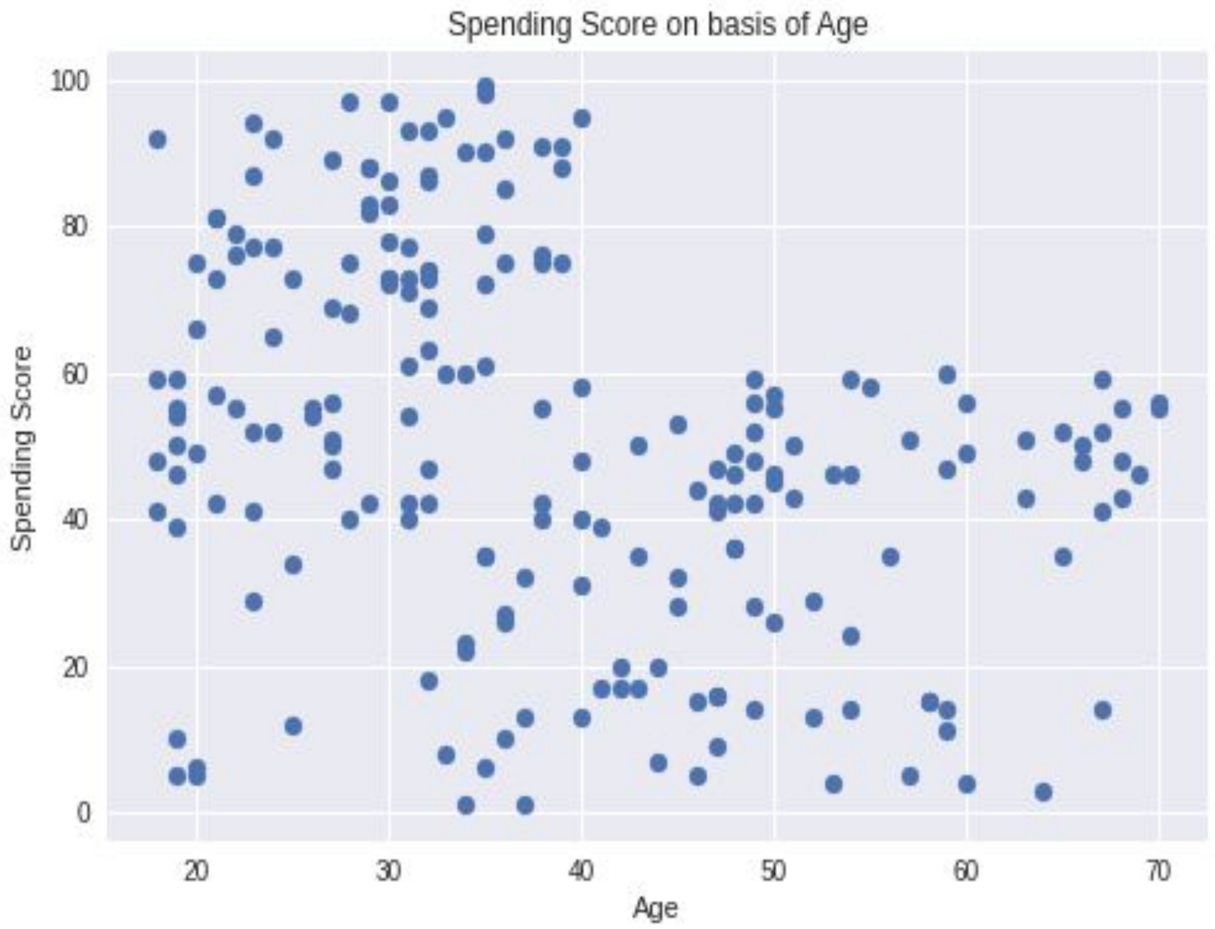
Implementation and Output

Data Preprocessing

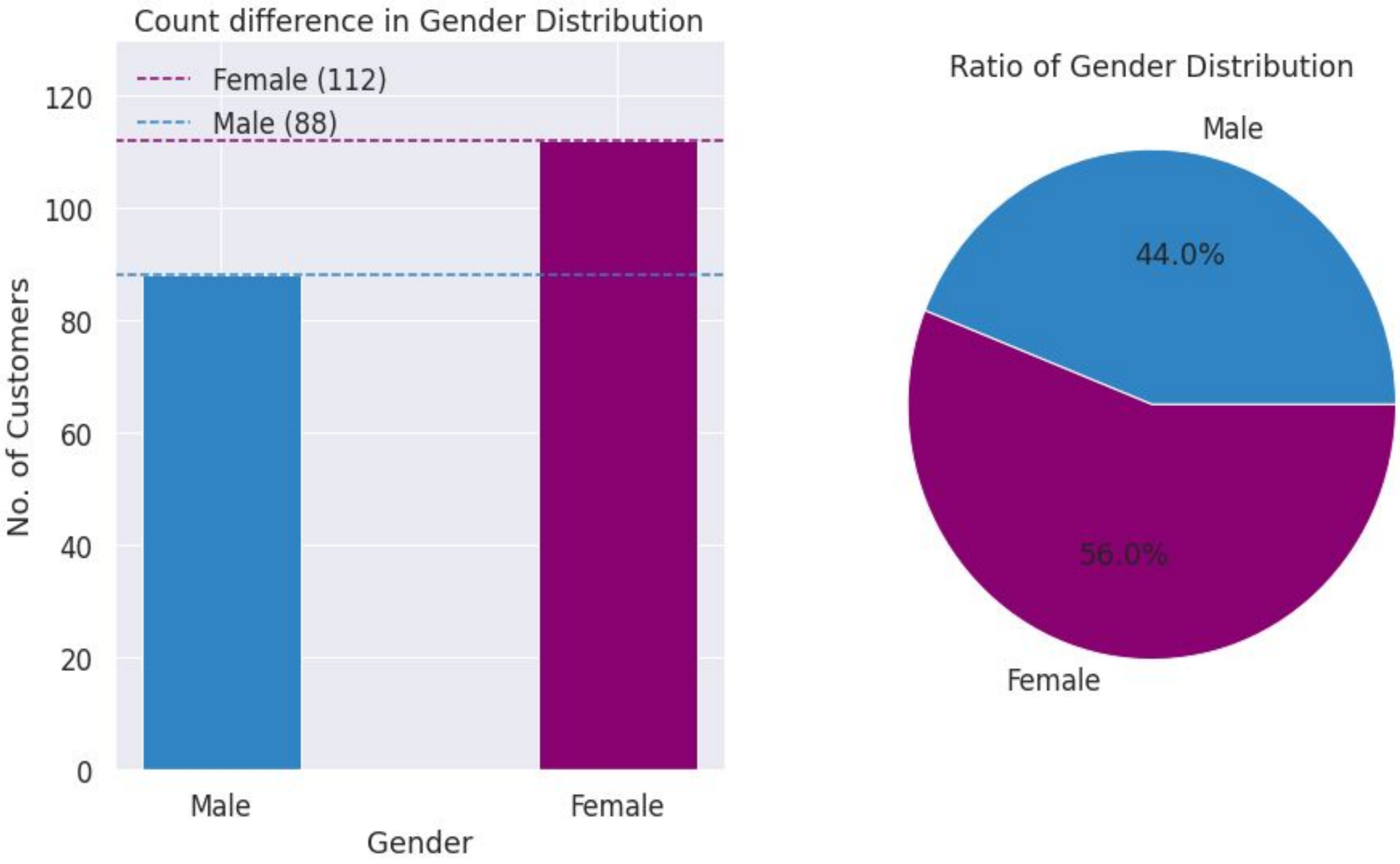


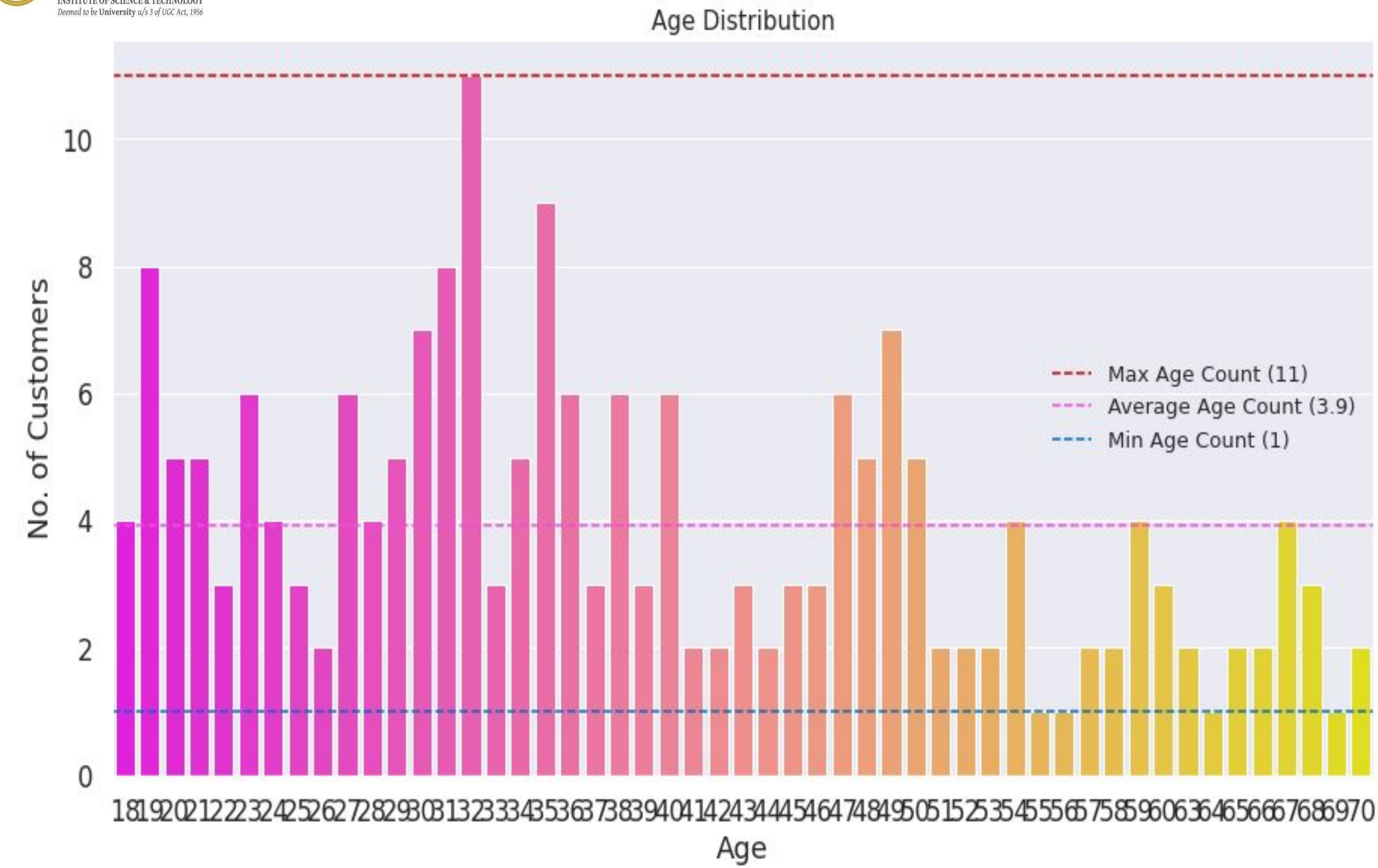
Implementation and Output

Data Preprocessing



Gender Distribution




Age Distribution
No. of Customers
0
2
4
6
8
10
Max Age Count (11)
Average Age Count (3.9)
Min Age Count (1)
18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 63 64 65 66 67 68 69 70
Age

References

1. https://ieeexplore.ieee.org/abstract/document/8528086
2. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.736.3182
3. https://ieeexplore.ieee.org/abstract/document/8769171
4. https://www.researchgate.net/profile/Kr-Kashwan/publication/271302240_Customer_Segmentation_Using_Clustering_and_Data_Mining_Techniques/links/57093e7908ae2eb9421e2d86/Customer-Segmentation-Using-Clustering-and-Data-Mining-Techniques.pdf
5. https://ieeexplore.ieee.org/document/5453745

THANK YOU!