Mall Customer Segmentation Analysis - Clustering

PROJECT REPORT

18CSC479T - STATISTICAL MACHINE LEARNING

(2018 Regulation)

III Year / V Semester

Academic Year: 2022 -2023

By

Varun Khachane (RA2011026010072)

Sahil Satasiya (RA2011026010110)

Under the guidance of

Dr. K Suresh

Assistant Professor

Department of Computational Intelligence

FACULTY OF ENGINEERING AND TECHNOLOGY

SCHOOL OF COMPUTING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

NOVEMBER 2022

BONAFIDE

This is to certify that **18CSC479T - STATISTICAL MACHINE LEARING project report** titled "**Mall Customer Segmentation Analysis - Clustering"** is the bonafide work of **Varun Khachane (RA2011026010072)**
Sahil Satasiya (RA2011026010110)
who undertook the task of completing the project within the allotted time.

Signature of the Guide
Dr. K Suresh
Assistant Professor
Department of CINTEL,
SRM Institute of Science and Technology

Signature of the HoD
Dr. Annie Uthra
Professor and Head
Department of CINTEL
SRM Institute of Science and Technology

About the course:-

18CSC479T - Statistical Machine Learning is three credit course with **L T P C as 3-0-0-3**

Objectives:

The student should be made to:

- Learn the basics of statistical machine learning techniques.
- Learn the basics of build model based on logistic regression and random forest techniques
- Learn basic idea of ideas of probability and work on probabilistic approaches like Naïve Bayes, Bayes Theorem
- Be familiar with knowledge of Kernel functions in practical applications
- Be familiar with knowledge of K-means clustering on real world examples
- Learn PCA and SVD with Scikit-learn

Course Learning Rationale (CLR): The purpose of learning this course is to:

Course Learning Rationale (CLR):	The purpose of learning this course is to:
CLR-1 :	Understand the Fuzzy Logic Basics
CLR-2 :	Gain knowledge on the Machine learning concepts
CLR-3 :	Gain knowledge on Fuzzy based clustering concepts
CLR-4 :	Acquire knowledge on Fuzzy Integrated classification
CLR-5 :	Understanding Neuro-Fuzzy Modeling concepts
CLR-5 :	Acquiring better understanding on Fuzzy logic usage
CLR-6	Understanding the fuzzylogics in Machine learning

Course Learning Outcomes (CLO): At the end of this course, learners will be able to:

Course Learning Outcomes (CLO):	At the end of this course, learners will be able to:
CLO-1 :	Acquire the knowledge on statistical machine learning techniques.
CLO-2 :	Acquire the ability to build model based on logistic regression and random forest techniques
CLO-3 :	Understand the basic ideas of probability and work on probabilistic approaches like Naïve Bayes, Bayes Theorem
CLO-4 :	Apply the knowledge of Kernel functions in practical applications
CLO-5 :	Apply the knowledge of K-means clustering on real world examples
CLO-6 :	Acquire the knowledge on using PCA and SVD with Scikit-learn

Table 1: Internal Mark Split-up:- As per Curriculum

Learning Assessment											
	Bloom's Level of Thinking	Continuous Learning Assessment (50%								Final Examination (50% weightage)	
		CLA – 1 (10%)		CLA – 2 (15%)		CLA – 3 (15%)		CLA – 4 (10%)#			
		Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice	Theory	Practice
Level 1	Remember	*40 %*	-	*30 %*	-	*30 %*	-	*30 %*	-	*30%*	-
	Understand										
Level 2	Apply	*40 %*	-	*40 %*	-	*40 %*	-	*40 %*	-	*40%*	-
	Analyze										
Level 3	Evaluate	*20 %*	-	*30 %*	-	*30 %*	-	*30 %*	-	*30%*	-
	Create										
	Total	100 %		100 %		100 %		100 %		100 %	

CLA – 4 can be from any combination of these: Assignments, Seminars, Tech Talks, **Mini-Projects**, Case-Studies, Self-Study, MOOCs, Certifications, Conf. Paper etc.,

ABSTRACT

Management and maintain of customer relationship have always played a vital role to provide business intelligence to organizations to build, manage and develop valuable long term customer relationships. The importance of treating customers as an organizations main asset is increasing in value in present day and era. Organizations have an interest to invest in the development of customer acquisition, maintenance and development strategies. The business intelligence has a vital role to play in allowing companies to use technical expertise to gain better customer knowledge and Programs for outreach. Customer segmentation helps the marketing team to recognize and expose different customer segments that think differently and follow different purchasing strategies. Customer segmentation helps in figuring out the customers who vary in terms of preferences, expectations, desires and attributes. The main purpose of performing customer segmentation is to group people, who have similar interest so that the marketing team can converge in an effective marketing plan. Clustering is an iterative process of knowledge discovery from vast amounts of raw and unorganized data.

Customer segmentation is a separation of a market into multiple distinct groups of consumers who share the similar characteristics. Segmentation of market is an effective way to define and meet customer needs. Unsupervised Machine Learning Techniques, K-Means Clustering Algorithm, Minibatch K-Means and Hierarchical Clustering are used to perform Market Basket Analysis. Market Basket Analysis is carried out to predict the target customers who can be easily converged, among all the customers. In order to allow the marketing team to plan the strategy to market the new products to the target customers which are similar to their interests. Key words: Target Customers, Clusters, Unsupervised Learning, K-Means, Minibatch K-Means, Hierarchical Clustering Segmentation, Market Basket Analysis

Dataset Description

You are owing a supermarket mall and through membership cards, you have some basic data about your customers. Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29
20	Female	35	23	98
21	Male	35	24	35
22	Male	25	24	73
23	Female	46	25	5
24	Male	31	25	73
25	Female	54	28	14
26	Male	29	28	82
27	Female	45	28	32
28	Male	35	28	61
29	Female	40	29	31

Similarly in total data of 200 customer was collected to get more accurate and precise output and simultaneously train the model more accurately.

Modules Description (Architecture diagram, Algorithms used etc.)

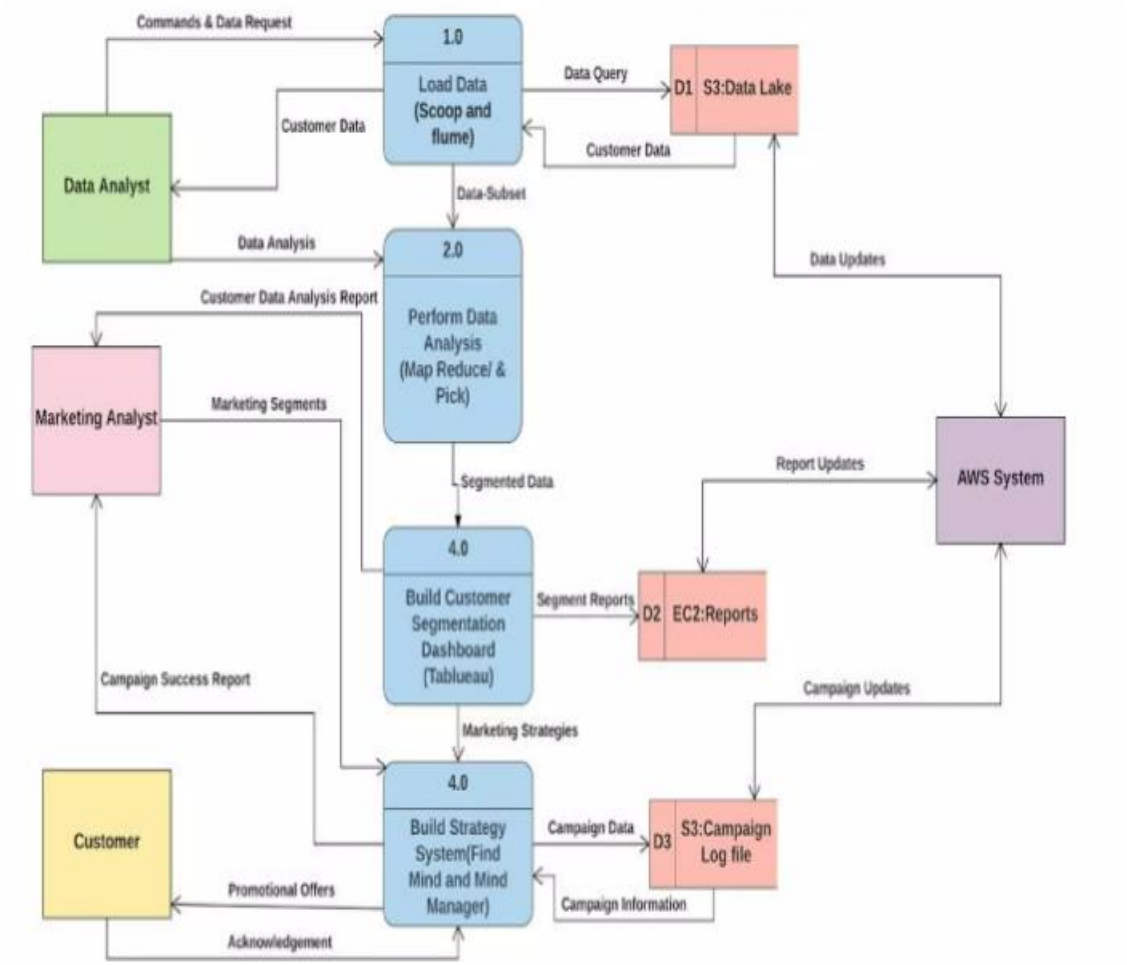
Libraries Used:

- Pandas
- Matplotlib
- Seaborn
- Scikit-learn

Attributes:

- Customer ID
- Age
- Gender
- Annual Income
- Spending Score

Architecture Diagram:



Algorithm: (K-mean Clustring)

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

```python
# cluster on 3 features
df2 = df[['Annual Income (k$)', 'Spending Score (1-100)', 'Age']]
df2.head()
```

Python

```python
errors = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(df2)
    errors.append(kmeans.inertia_)
```

Python

```python
# plot the results for elbow method
plt.figure(figsize=(13,6))
plt.plot(range(1,11), errors)
plt.plot(range(1,11), errors, linewidth=3, color='red', marker='8')
plt.xlabel('No. of clusters')
plt.ylabel('WCSS')
plt.xticks(np.arange(1,11,1))
plt.show()
```

Python

```python
km = KMeans(n_clusters=5)
km.fit(df2)
y = km.predict(df2)
df2['Label'] = y
df2.head()
```

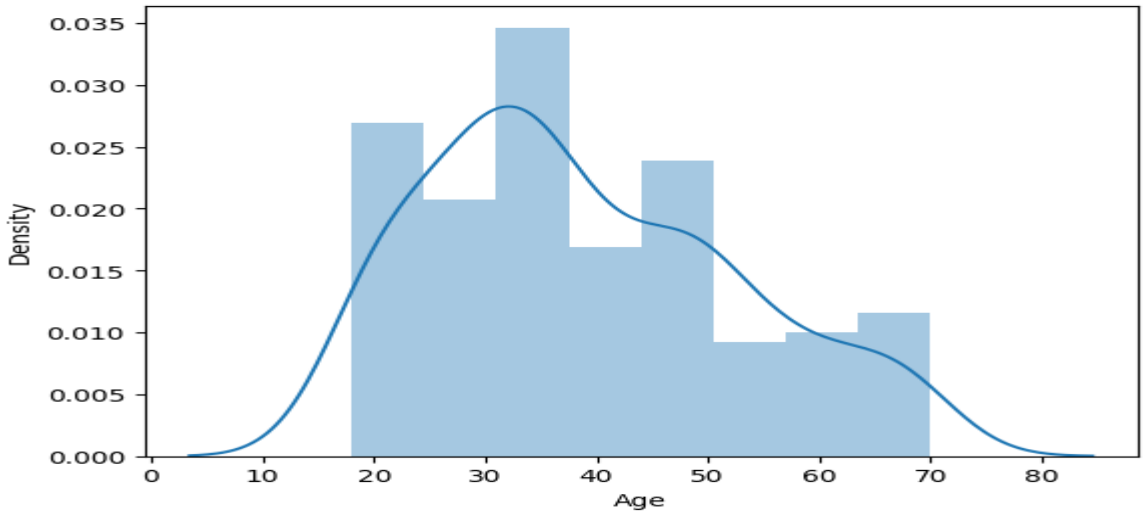
Python

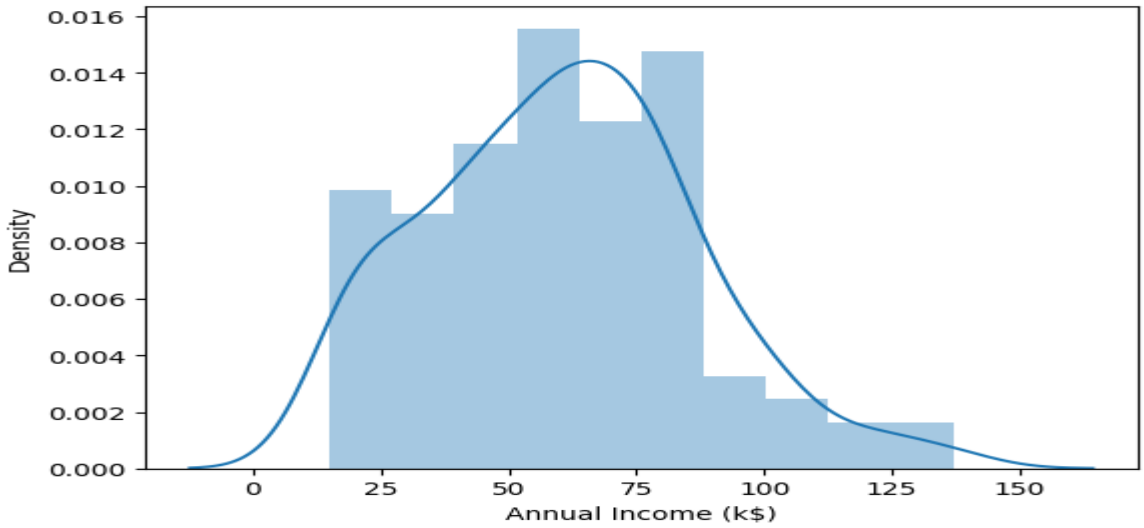
```python
# 3d scatter plot
fig = plt.figure(figsize=(20,15))
ax = fig.add_subplot(111, projection='3d')

ax.scatter(df2['Age'][df2['Label']==0], df2['Annual Income (k$)'][df2['Label']==0], df2['Spending Score (1-100)'][df2['Label']==0], c='red', s=50)
ax.scatter(df2['Age'][df2['Label']==1], df2['Annual Income (k$)'][df2['Label']==1], df2['Spending Score (1-100)'][df2['Label']==1], c='green', s=50)
ax.scatter(df2['Age'][df2['Label']==2], df2['Annual Income (k$)'][df2['Label']==2], df2['Spending Score (1-100)'][df2['Label']==2], c='blue', s=50)
ax.scatter(df2['Age'][df2['Label']==3], df2['Annual Income (k$)'][df2['Label']==3], df2['Spending Score (1-100)'][df2['Label']==3], c='brown', s=50)
ax.scatter(df2['Age'][df2['Label']==4], df2['Annual Income (k$)'][df2['Label']==4], df2['Spending Score (1-100)'][df2['Label']==4], c='orange', s=5
ax.view_init(30, 190)
ax.set_xlabel('Age')
ax.set_ylabel('Annual Income $')
ax.set_zlabel('Spending Score')
plt.show()
```

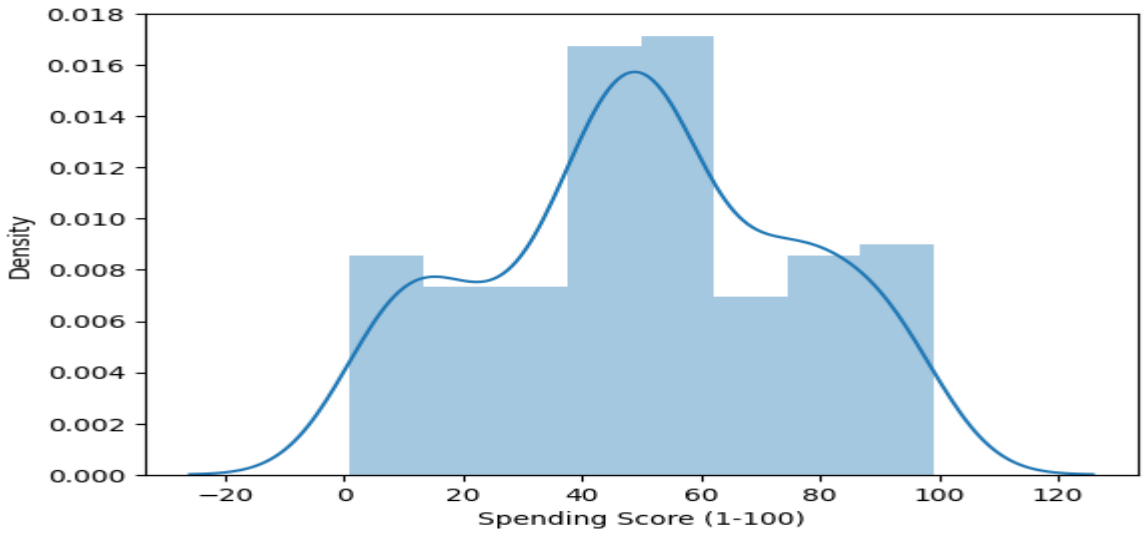
Python

Results and Discussion (Confusion matrix, output image/values, evaluation parameters table, graphs etc.)

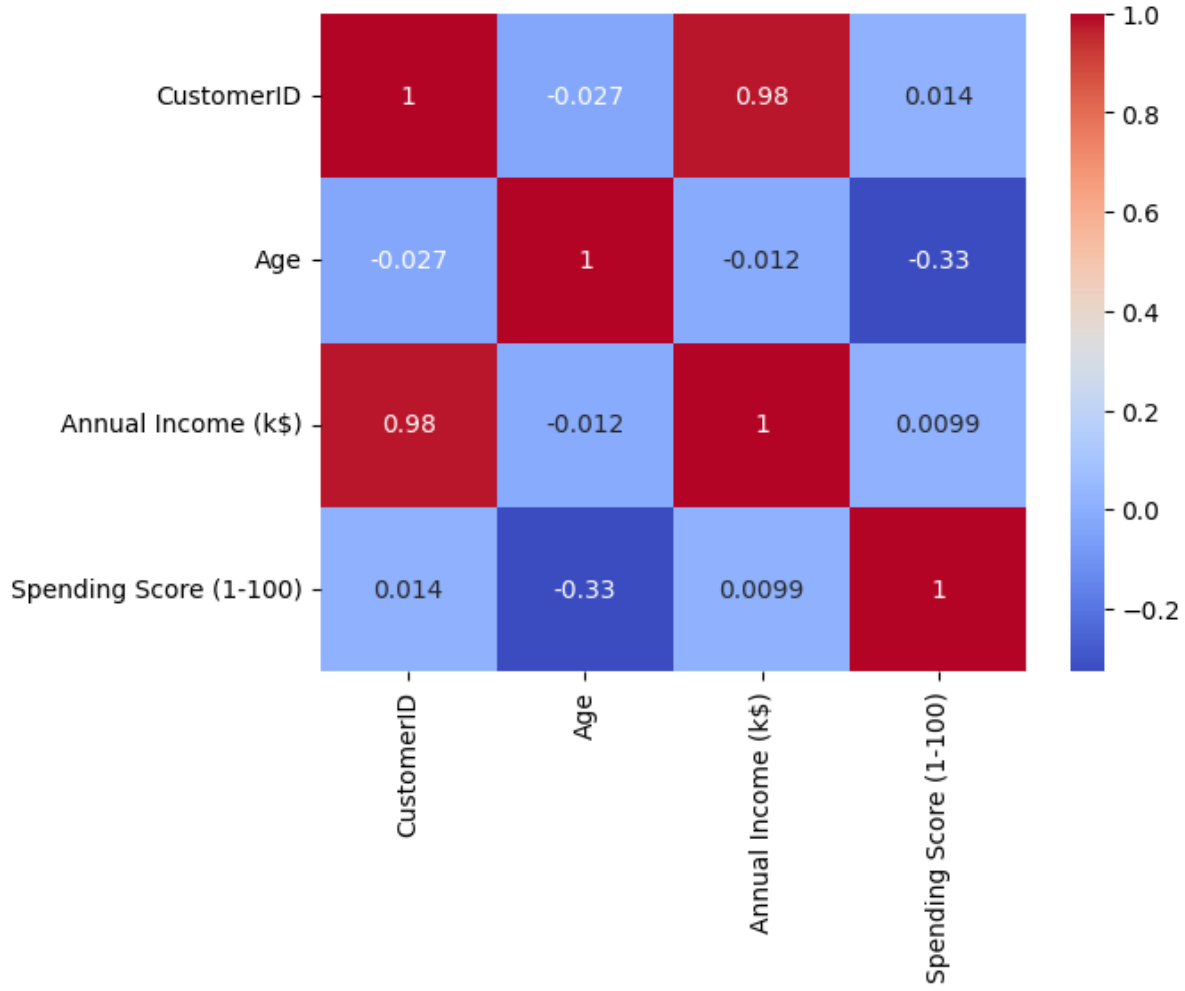
Displot:



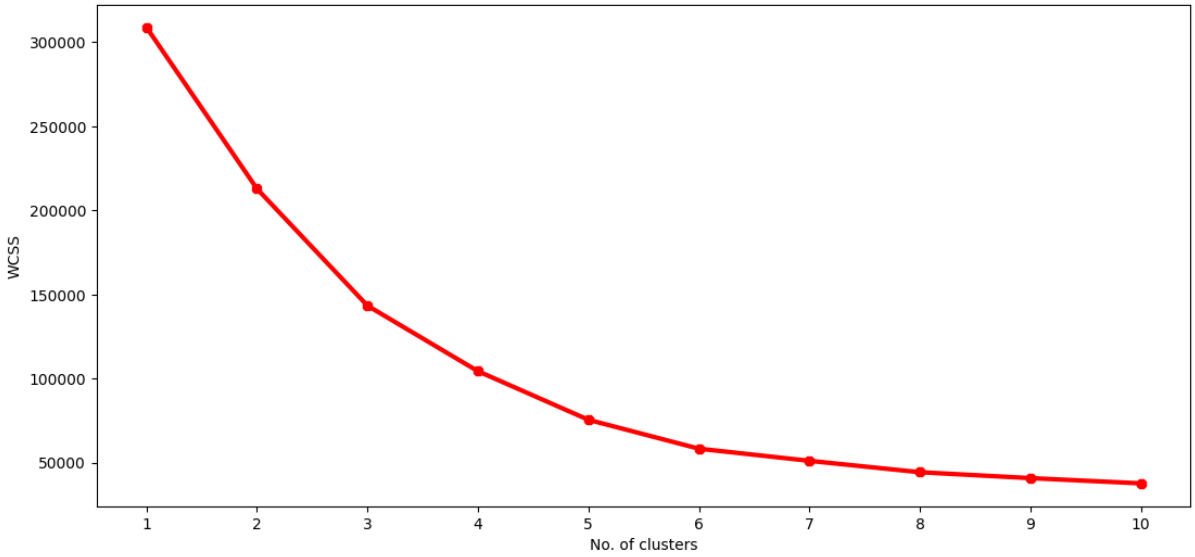




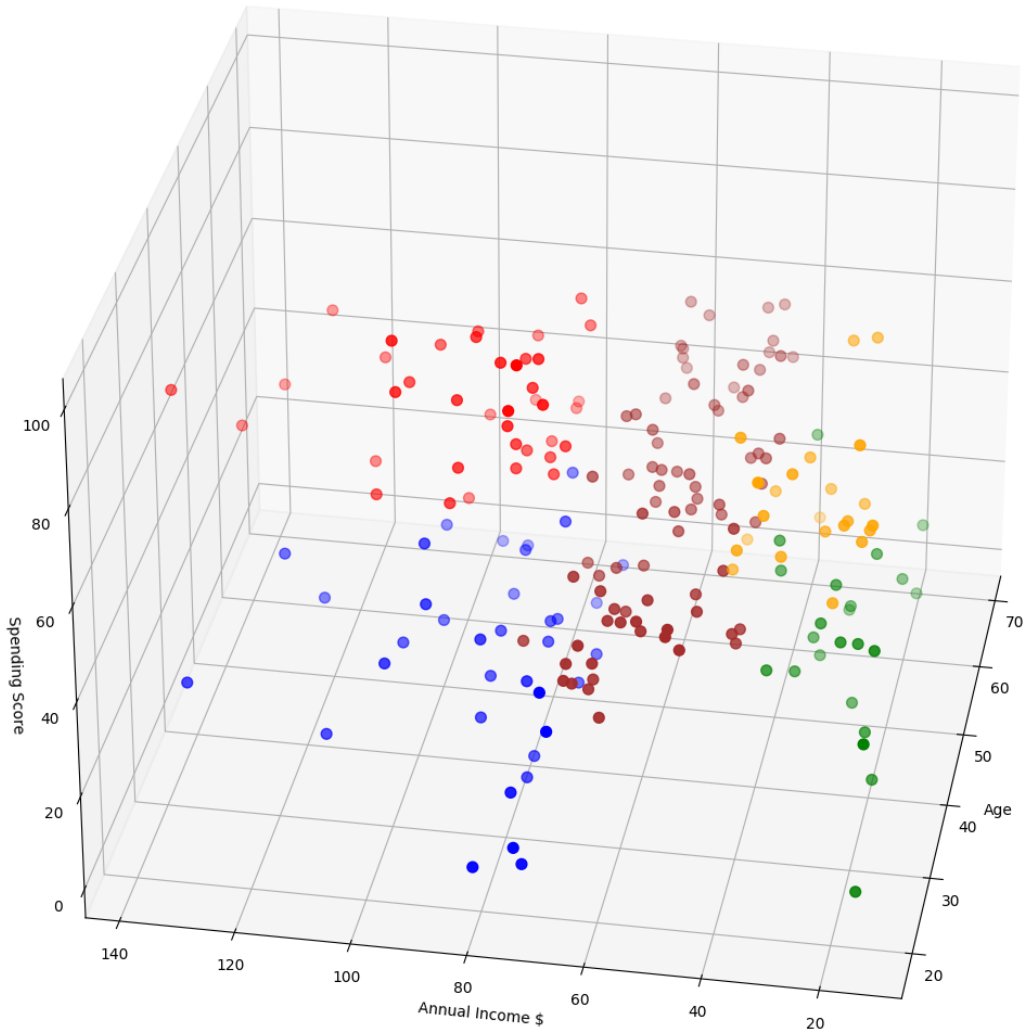
Co-relation Matrix:

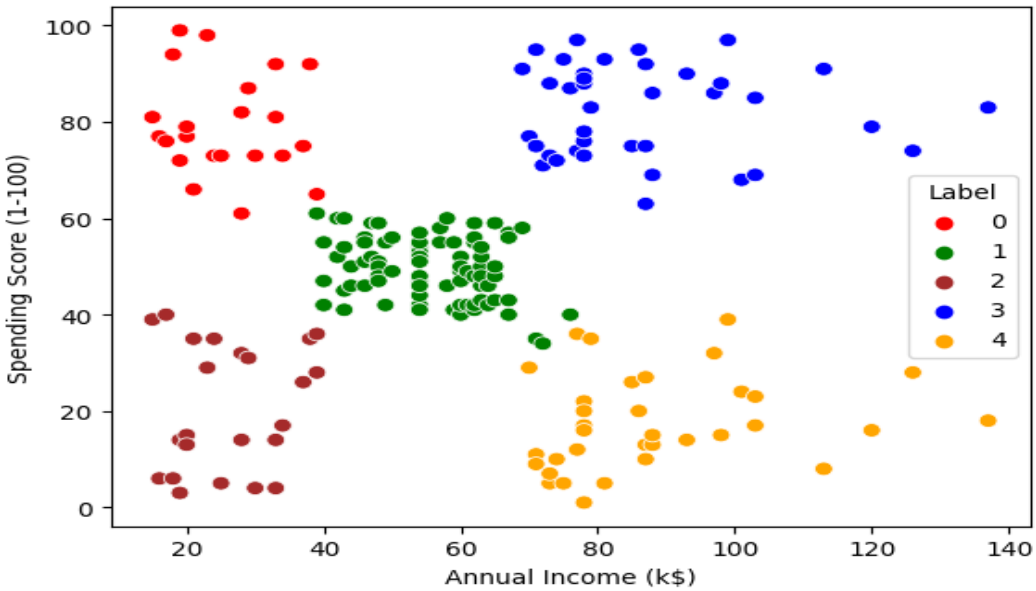


Elbow Diagram



Final Output:





Conclusion

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company

Reference

https://copyassignment.com/8-steps-to-build-a-machine-learning-model/

https://www.kaggle.com/code/fabiendaniel/customer-segmentation

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning