Assignment - 1

Tomson George (C0857730)

Praveen Mahaulpatha (C0860583)

R. B. C. M. W. Thulana Vimukthi Abeywardana

Artificial Intelligent and Machine Learning

2022S AML 1413 2 - Introduction to Artificial Intelligence

Debashish Roy

July 15, 2022

Part 2 - Unsupervised Learning

1. Understanding the Dataset

The dataset includes 200 instances and 5 attributes about customers of a shopping mall. The data set was retrieved from Kaggle.com . Attributes of the dataset as follows:

Features

- Customer ID
- Age
- Annual Income
- Spending Score (1-100)

2. Exploratory Data Analysis

2.1. Loading Dataset

```
1 df = pd.read_csv('Mall_Customers.csv')
```

2.2. Dataset dimensions

```
1 df.shape
```

```
(200, 5)
```

2.3. Descriptive Statistics

```
1 df.describe()
```

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

2.4. Data types

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
```

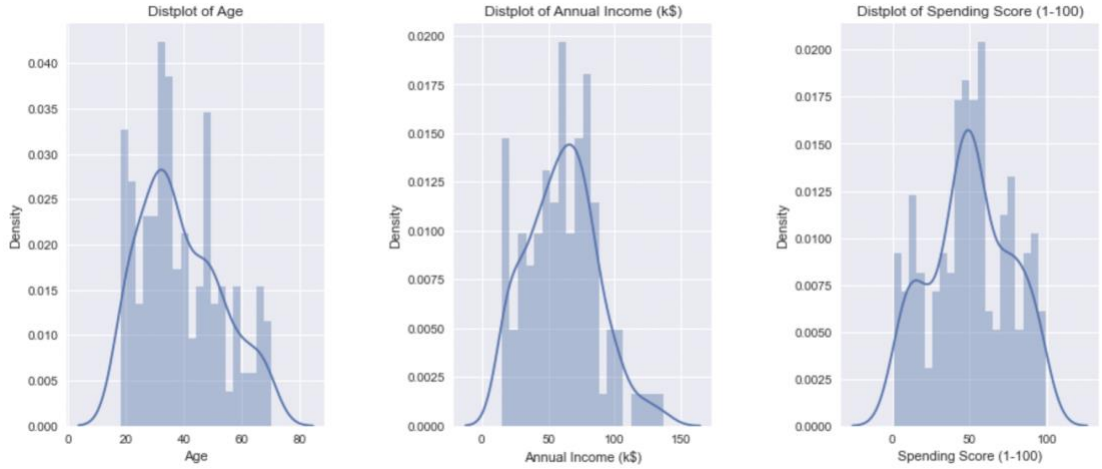
2.5 Checking for null values

```
1 df.isnull().sum()
```

```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

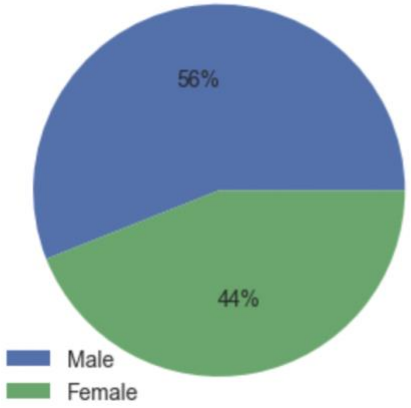
No null values exists in the dataset.

2.6. Plotting Histograms



Data seems to be having a normally distributed but slight skewness can be identified. The ranges of axis among the labels are having similar values hence normalization will not be necessary.

2.7. Plotting Pie Chart



Data seems to be more biased towards Male population hence we can draw a better sample in the next phase of the project.

3. Fitting the Model

In this project, we will be using a K means clustering algorithm to find out hidden clusters among customers. We will define a target number k, which refers to the number of centroids needed. A centroid is an imaginary or real location representing the center of the cluster. Then, the algorithm allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The key is to findout the best value for variable k, inorder to yield the most efficient results.

3.1. Finding out the Inertia

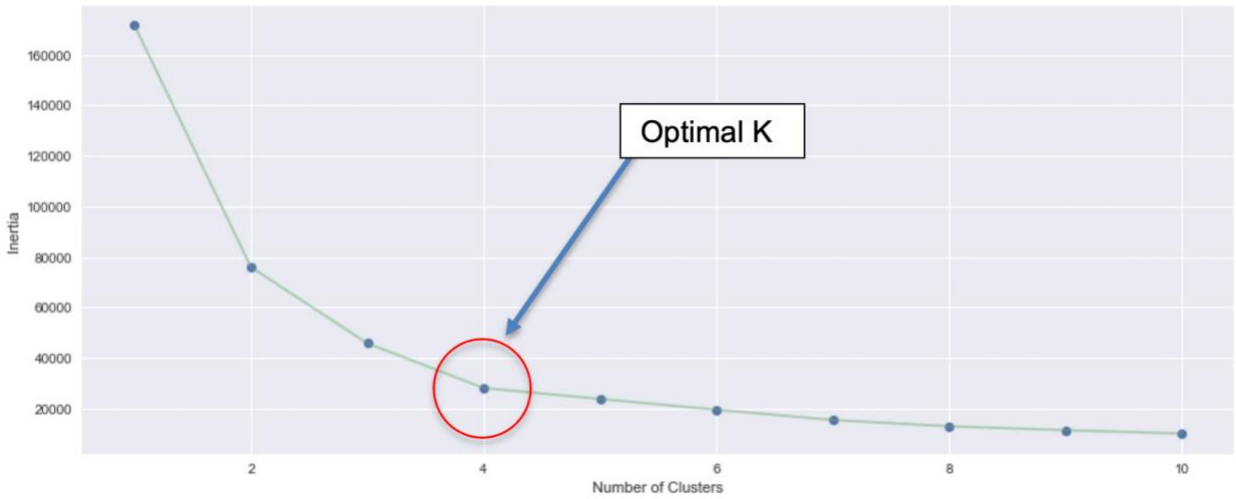
Inertia measures how well a dataset was clustered by K-Means. It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.

We have given k, values from 1 to 10 and found out the algorithm's inertia and appended it to a list.

```
'''Age and spending Score'''
X1 = df[['Age' , 'Spending Score (1-100)']].iloc[: , :].values
inertia = []
for n in range(1 , 11):
    algorithm = (KMeans(n_clusters = n ,init='k-means++', n_init = 10 ,max_iter=300,
                        tol=0.0001,  random_state= 111  , algorithm='elkan') )
    algorithm.fit(X1)
    inertia.append(algorithm.inertia_)
```

3.2. Finding the optimal K value

To find the optimal K value, we can use the commonly used elbow method. Where we plot the inertia value against the number of k and consider the k value where the decrease in inertia begins to slow as the optimal k value.



3.3. Fitting K means

Once the model is fit for k = 4, we can see clear clusters in the relationship between Age and spending score attributes shown below. Similarly we can perform such segmentation for Annual Income and Spending Score in the next phase of the project.