Data Science Specialization

Academic Year 2019 - 2023

Python for Data Science Project

Customer Segmentation

Detailed Analysis and Visualization on Customer Data Attributes
K-means Clustering to Segment the Customers

Group: 04
Members:
Preethi G - S20190020241
Subash J - S20190020253
Vamsi Chittoor - S20190020208
Kajal - S20190020215

1 Project Description

Owning any business involves some basic data about customers like name, age, customer ID, amount spent etc. Understanding and planning a strategy on the customers and targeting the marketing towards the buyers not only helps the business grow faster but also leads to customer satisfaction.

2 Description of the Data

The data set includes basic customer data such as Customer ID, Age, Gender, Annual Income and Spending Score. Spending Score is a score (out of 100) given to a customer by the business analysts, based on the amount spent on purchases and the behavior of the customer.

2.1 Data Set Information:

The data set contains a total of 5 columns (Customer ID, Age, Gender, Annual Income and Spending Score) with 200 data points. (There were no missing values in the data set). All attributes except gender are numerical attributes.

Attribute Information:

No.	Attribute	Description
1	CustomerID	Numbers range from 1 to 200 (unique integers)
2	Gender	Gender of the customer (Male and Female).
3	Age	Age of the customers range from 18 to 70
4	Annual Income (k$)	Annual Income of customers measured in thousands($)
5	Spending Score (1-100)	Score of a customer based on money spent & behaviour

3 Methodology

3.1 Visualizing and Pre-processing the Data set

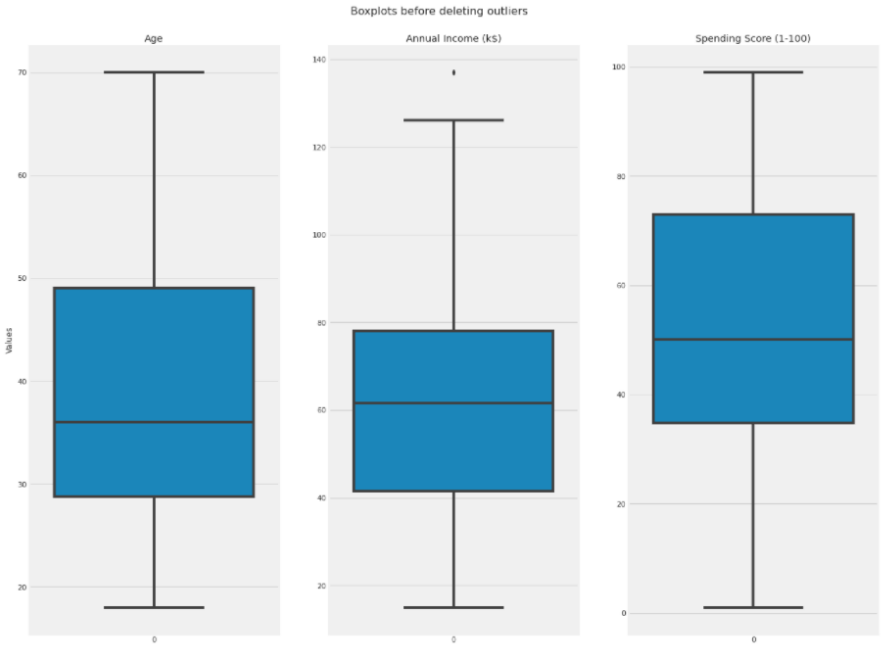
- The Data set is read and the descriptive statistics of the numerical attributes of the data set is calculated using describe().

	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

- The descriptive statics of the categorical attributes is also observed using describe(include='object').

	Gender
count	200
unique	2
top	Female
freq	112

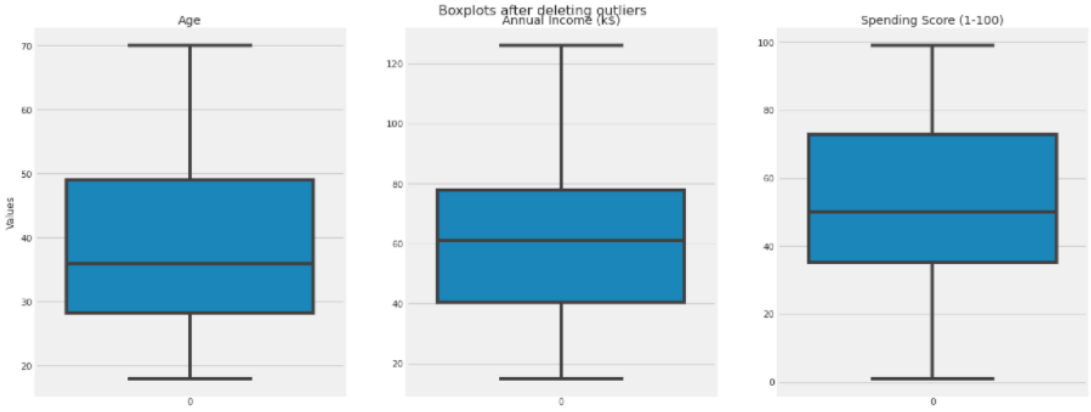
- Next the Exploratory Data Analysis (EDA) on the data set is done. The box plot of the data is plotted for each numerical columns (only Age, Annual Income (k$) and Spending Score (1-100)). The CustomerID is not taken since its a unique number given to every customer which is equivalent to label encoding.



- Now the Interquartile Range of each column is calculated and the outliers are removed from the data.

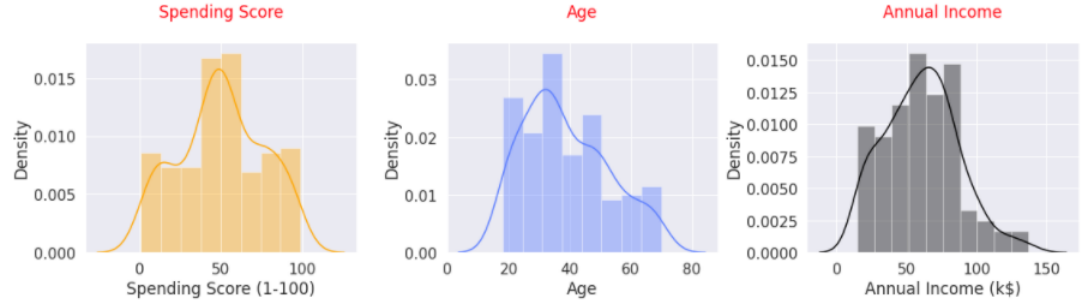
```
Here we will get IQR for each column
 Age                     20.25
Annual Income (k$)       36.50
Spending Score (1-100)   38.25
dtype: float64

(200, 5)

(198, 5)
```

- After the removal, the boxplot is plotted again to make sure that all the outlier values are removed and the data is free of outliers.

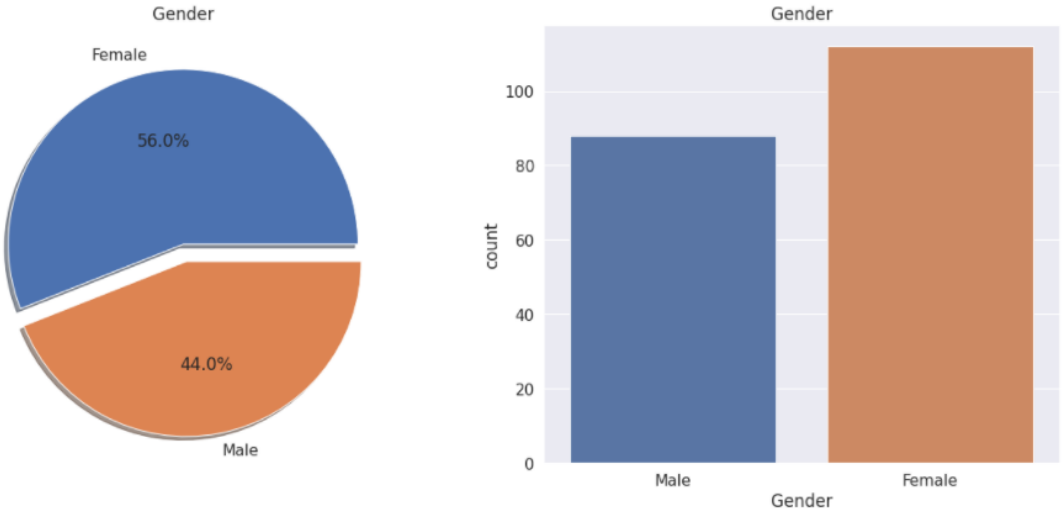


- The correlation for the Data Frame is calculated and the heatmap is plotted.

- Followed by the heatmaps the distribution plots are plotted using distplot() to understand about the distribution of each attributes better.



- A pie plot is plotted to understand the gender percentage and count plot is plotted to observe the count of each gender



- In the data pre-processing step the CustomerID column is dropped since it is not a redundant attribute. The data is checked for missing values. (No missing values).

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

  Since Gender is a categorical column but still might play a role in the data, one hot encoding is performed.

3.2 Model Building

Since there is no labels for the classification of data, an unsupervised learning technique must be used. K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in data science.

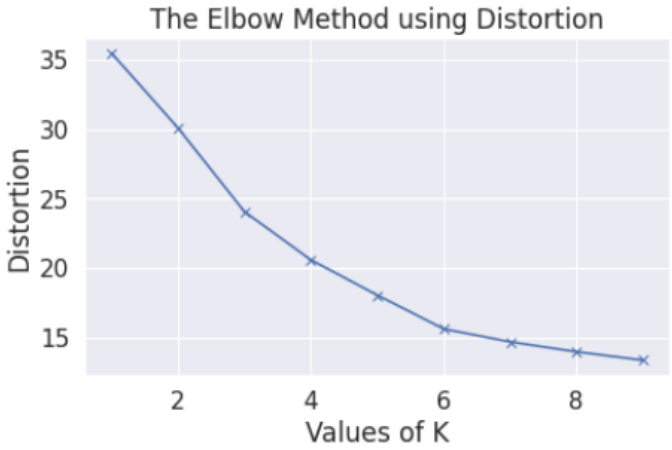
K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

- The data is split into Training set and Testing set using train_test_split with a test_set of 0.05 (implies 190 data points for training and 10 data points for testing).

- For an unsupervised algorithm, it is required to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k using calculations of distortion, inertia etc.

- Distortion is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.

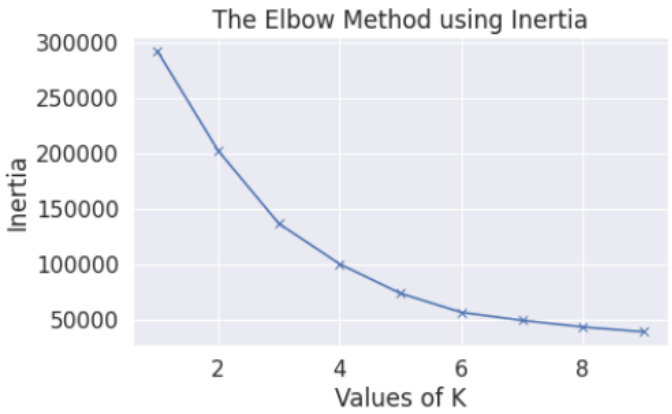
```
1 : 35.4815677834514
2 : 30.094868981688894
3 : 24.070753045711776
4 : 20.593552685291694
5 : 18.057790612746228
6 : 15.623759349995327
7 : 14.689156666444633
8 : 13.975282926441825
9 : 13.368839628807743
```



It can be observed from the above plot that the optimum number of clusters k is 6.

- Inertia is the sum of squared distances of samples to their closest cluster center.

```
1 : 292384.35263157886
2 : 202491.06046616373
3 : 136531.14039460564
4 : 100105.9779989059
5 : 73585.46065885868
6 : 56499.32887899534
7 : 49209.14606971975
8 : 43257.705320790614
9 : 38956.92703480627
```



The Elbow method using inertia also verifies that the optimum number of clusters k is around 6.

- To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e. the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 6.

- So, K-Means is found with KMeans(n_cluster=6). The algorithm labels are given as follows

```
array([5, 2, 4, 4, 0, 3, 3, 4, 3, 0, 2, 2, 3, 2, 0, 2, 2, 2, 5, 1, 1, 0,
       4, 5, 1, 2, 2, 1, 0, 0, 4, 0, 1, 5, 4, 0, 1, 5, 0, 2, 0, 5, 4, 5,
       2, 1, 1, 4, 0, 0, 2, 0, 2, 3, 2, 0, 0, 2, 5, 0, 2, 0, 2, 0, 4, 0,
       0, 1, 3, 2, 2, 0, 4, 2, 0, 0, 1, 0, 4, 2, 5, 1, 4, 0, 4, 4, 2, 5,
       5, 5, 4, 2, 0, 1, 5, 0, 5, 0, 2, 3, 5, 5, 5, 0, 3, 2, 3, 4, 2, 1,
       3, 5, 0, 5, 5, 0, 3, 0, 4, 4, 1, 5, 5, 5, 2, 2, 4, 5, 2, 3, 0, 0,
       3, 0, 5, 3, 5, 4, 1, 5, 5, 2, 3, 1, 2, 0, 2, 4, 1, 1, 5, 2, 3, 4,
       2, 4, 1, 5, 2, 2, 3, 5, 0, 0, 5, 4, 2, 0, 5, 3, 2, 5, 4, 4, 4, 3,
       5, 4, 5, 0, 5, 1, 3, 3, 5, 0, 4, 0, 5, 4], dtype=int32)
```

  The cluster centers are given below and the number of iterations made is 6

```
array([[ 0.4047619 , 56.5952381 , 53.30952381, 49.64285714],
       [ 0.4       , 43.25      , 26.2       , 20.1       ],
       [ 0.44736842, 32.84210526, 86.5       , 82.47368421],
       [ 0.42857143, 25.42857143, 26.14285714, 79.52380952],
       [ 0.53333333, 41.26666667, 89.23333333, 16.2       ],
       [ 0.35897436, 27.41025641, 57.02564103, 48.76923077]])
```

3.3 Principal Component Analysis

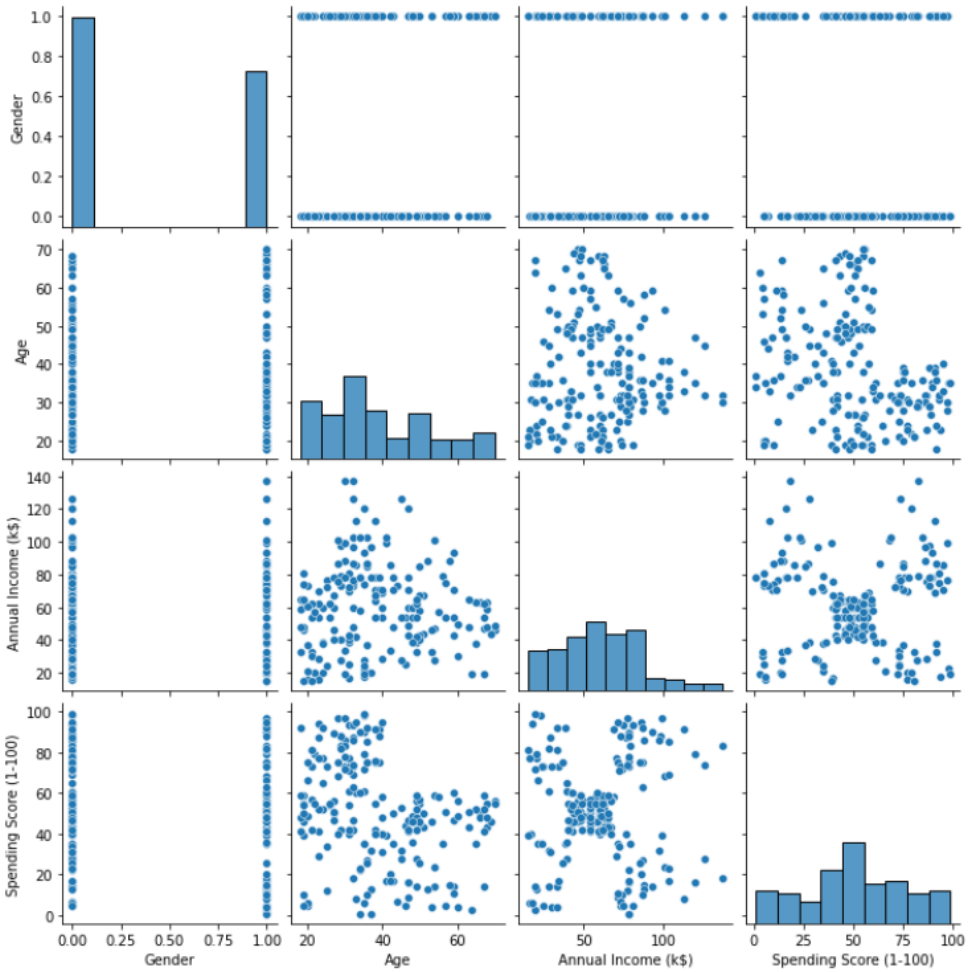
- PCA is an algorithm that is used for dimensionality reduction - meaning, informally, that it can take in a DataFrame with many columns and return a DataFrame with a reduced number of columns that still retains much of the information from the columns of the original DataFrame.

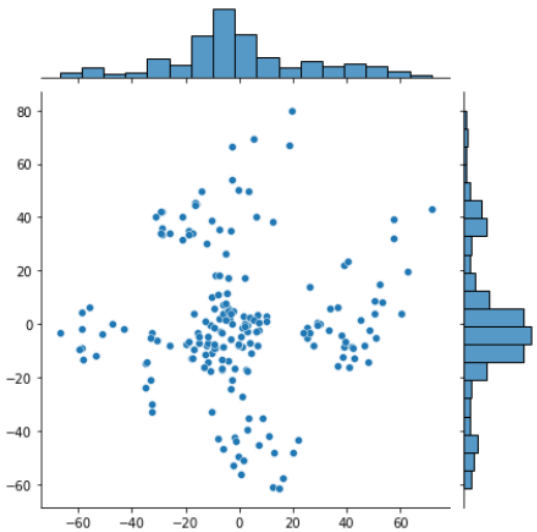
- The columns of the DataFrame produced from the PCA procedure are called Principal Components.

- We will use these principal components to help us visualize our clusters in 1-D, 2-D, and 3-D space, since we cannot easily visualize the data we have in higher dimensions.

- For example, we can use two principal components to visualize the clusters in 2-D space, or three principal components to visualize the clusters in 3-D space.

The pair plots and joint plots are plotted to understand the relation between the attributes better.



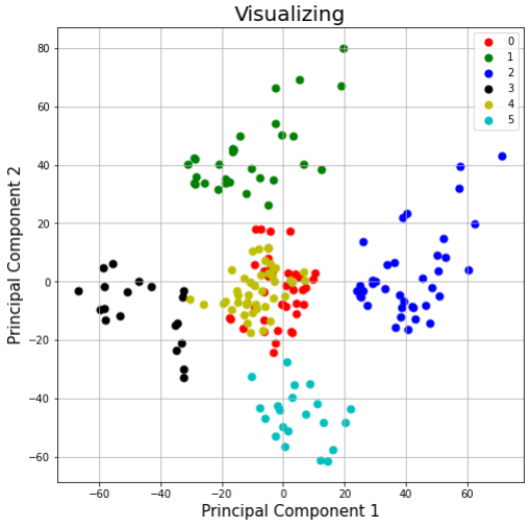


The PCA components and the explained variance ratio is shown below respectively.

```
array([[-3.90649915e-04, -1.79208314e-01,  5.82038045e-01,
         7.93168293e-01],
       [ 1.03373976e-03,  1.28376580e-01,  8.13161319e-01,
        -5.67703316e-01]])
```

```
array([0.45453298, 0.43379727])
```

Visualizing the clusters using Principle Component 1 and Principle Component 2



The over lapping clusters doesn't indicate that the clusters are improperly separated. The overlap is due to the reduction in dimensions by PCA. In higher dimensions the clusters are well separated.

3.4 Prediction

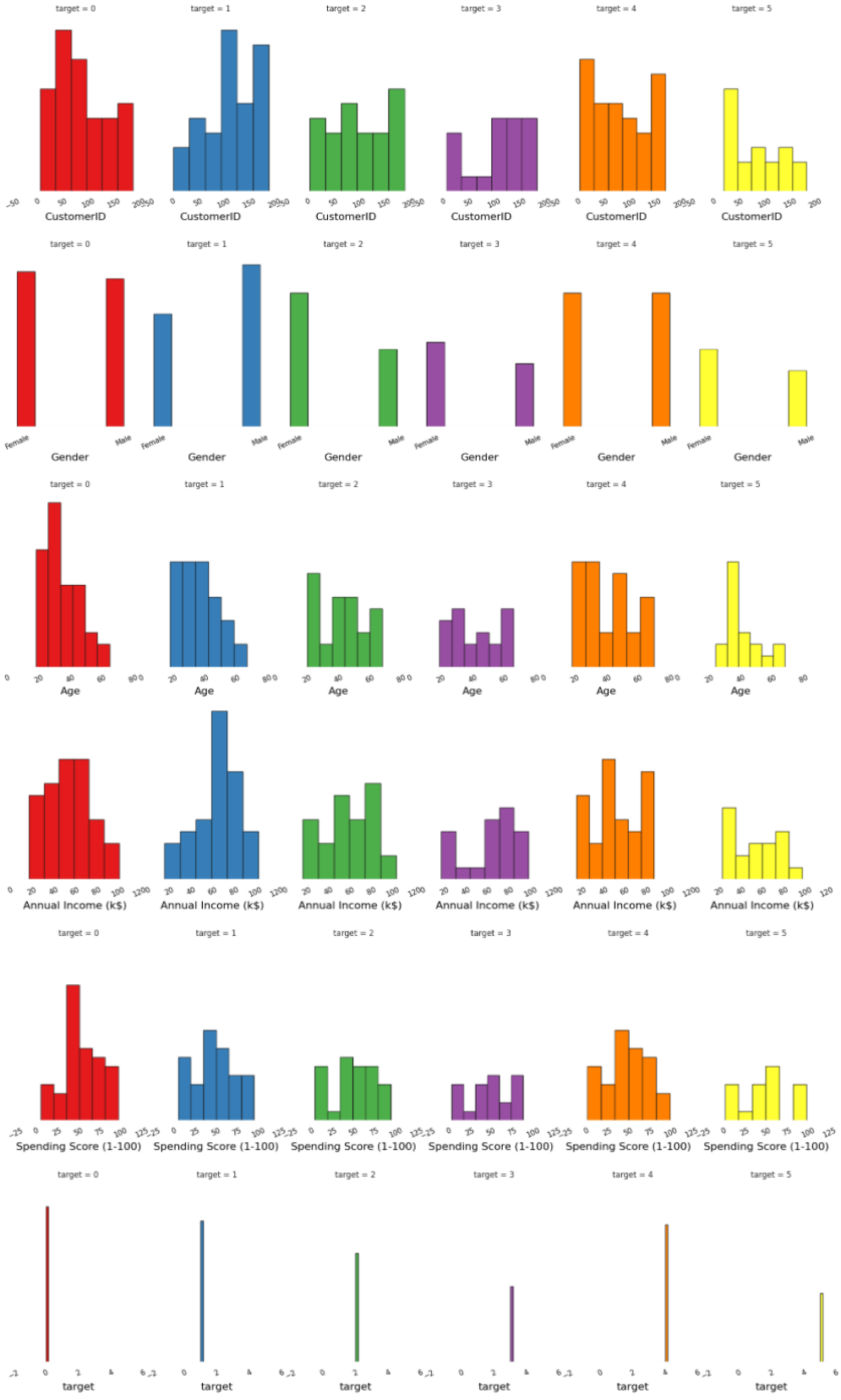
Prediction on the test data is given below

```
Input data
     Gender  Age  Annual Income (k$)  Spending Score (1-100)
18        1   52                  23                      29
170       1   40                  87                      13
107       1   54                  63                      46
98        1   48                  61                      42
177       1   27                  88                      69
182       1   46                  98                      15
5         0   22                  17                      76
146       1   48                  77                      36
12        0   58                  20                      15
152       0   44                  78                      20
Predicted cluster
 [3 1 4 4 2 1 5 1 3 1]
```

4 Visualization and Results

The count plot of distribution of the clusters clusters is shown in the figure below

The Cluster Interpretation



5 References

1. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

2. https://towardsdatascience.com/understanding-k-means-clustering