Big Data Science (CSCI 5952)

Mahsa Rahimian

University of Colorado Denver

04/26/2021

https://github.com/Mahsa7915/Mahsa/projects/1

https://github.com/Mahsa7915?tab=repositories

Contents

Problem Statement and Background 4

Methods 4

Results..6

Tools 10

Lesson Learned 12

Problem Statement and Background

Customer segmentation is one of the crucial applications in unsupervised learning. Companies can determine several segments of customers by using clustering techniques which can allow them to target the protentional user base.

Customer segmentation can divide customers into several groups which share a similarity in different ways that is based on gender, interests, miscellaneous spending habits, and age. This project can help organization to solve a big problem which is target the potential customer for a particular product. The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

Methods

While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as “cluster assignment”. When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data.

Another important methods that we should use for determining optimal clusters are:

- Elbow method

- Silhouette method
- Gap statistic

Elbow Method: The main goal behind cluster partitioning methods like k-means is to define the clusters such that the intra-cluster variation stays minimum.

Silhouette Method: “With the help of the average silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the average silhouette over significant values for k clusters”.

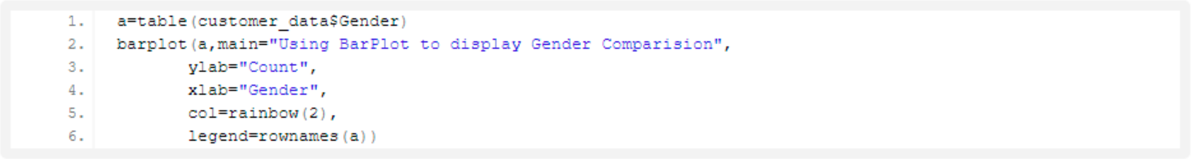
Gap Statistic Method: “In 2001, researchers at Stanford University – R. Tibshirani, G.Walther and T. Hastie published the Gap Statistic Method. We can use this method to any of the clustering method like K-means, hierarchical clustering etc. Using the gap statistic, one can compare the total intracluster variation for different values of k along with their expected values under the null reference distribution of data. With the help of Monte Carlo simulations, one can produce the sample dataset. For each variable in the dataset, we can calculate the range between min(xi) and max (xj) through which we can produce values uniformly from interval lower bound to upper bound.For computing the gap statistics method we can utilize the clusGap function for providing gap statistic as well as standard error for a given output”.

Results

Customer Gender Visualization

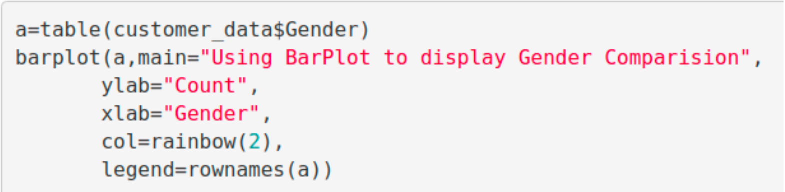
In this, we will create a barplot and a piechart to show the gender distribution across our customer_data dataset.

Code:



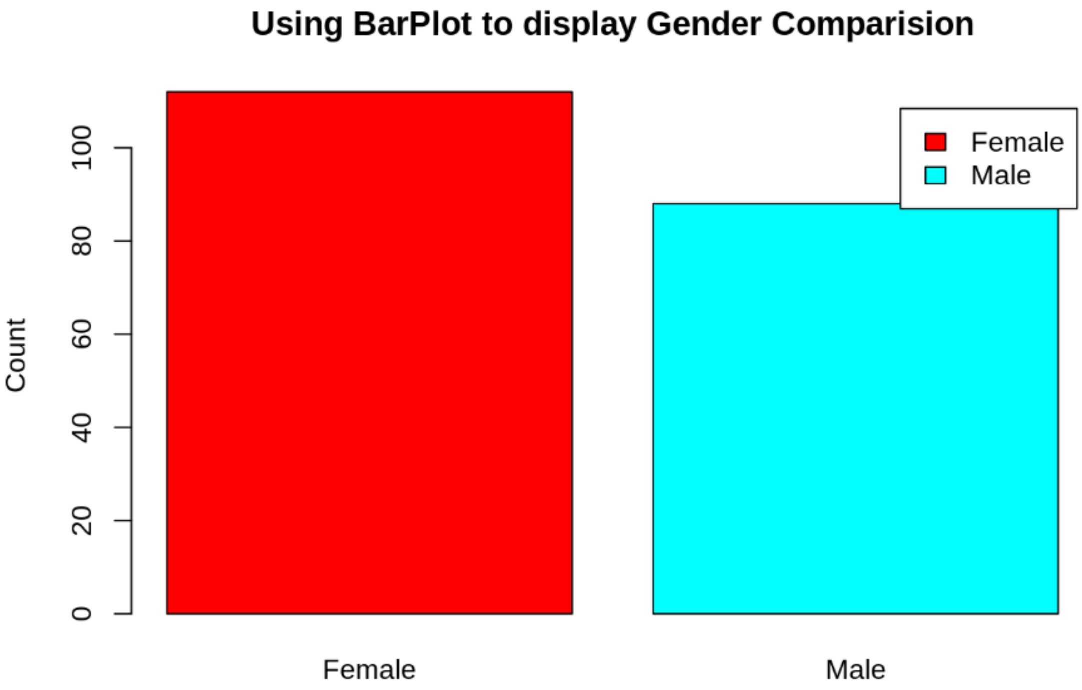
```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

Screenshot:



```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

Output:

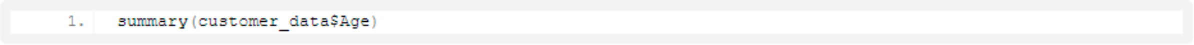


By analyzing above barplot we can see the number of females is higher than males.

Visualization of Age Distribution

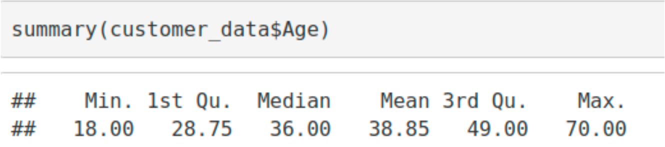
Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.

Code:



```
summary(customer_data$Age)
```

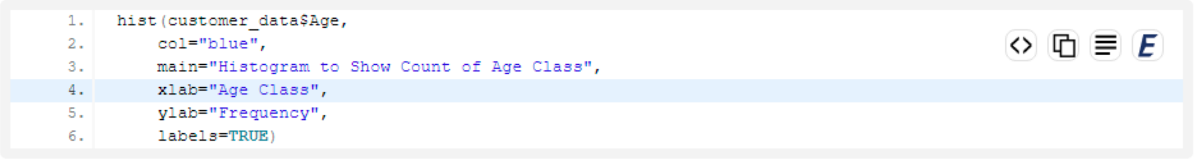
Output Screenshot:



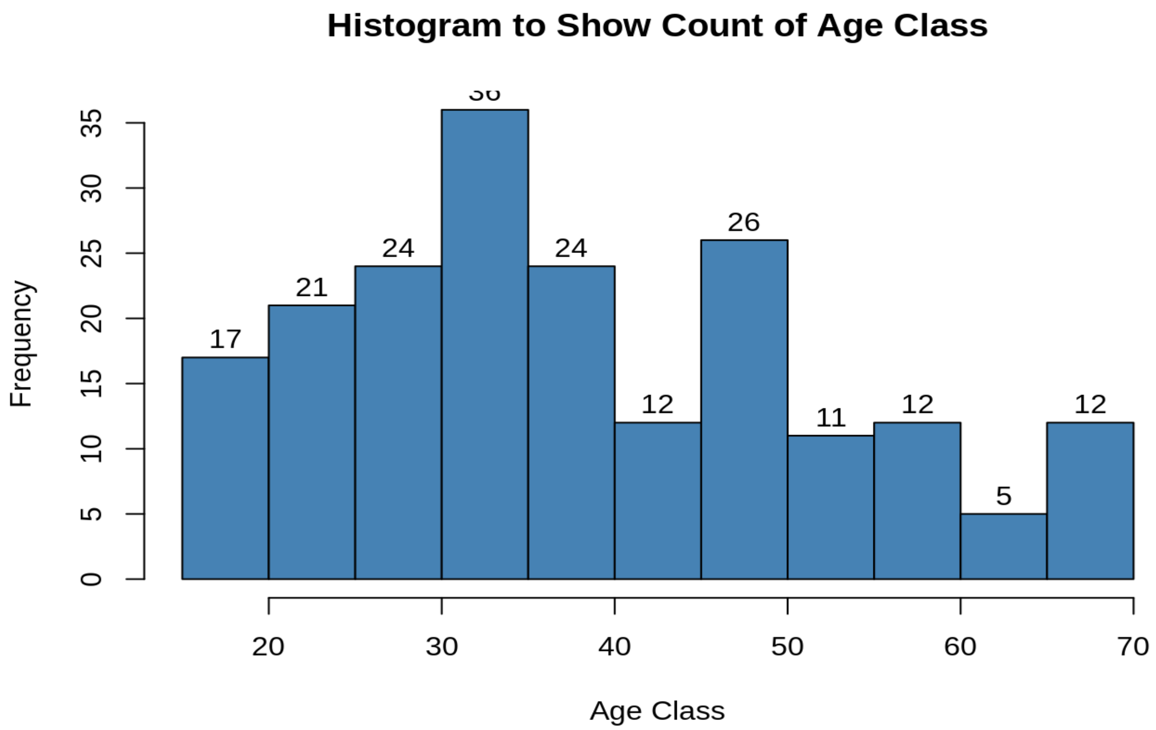
```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   28.75   36.00   38.85   49.00   70.00
```

Code:



```
hist(customer_data$Age,
    col="blue",
    main="Histogram to Show Count of Age Class",
    xlab="Age Class",
    ylab="Frequency",
    labels=TRUE)
```



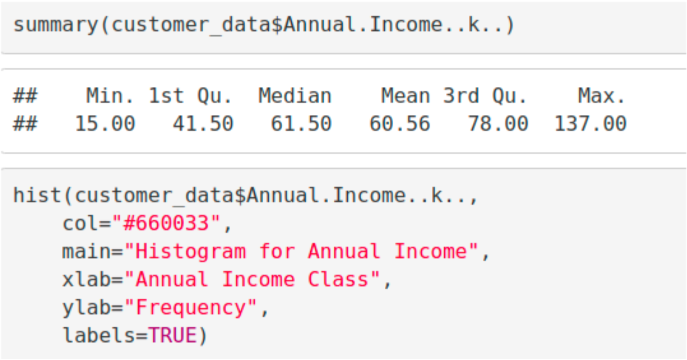
Analysis of the Annual Income of the Customers

In this section of the R project, we will create visualizations to analyze the annual income of the customers. We will plot a histogram and then we will proceed to examine this data using a density plot.

Code:

```
summary(customer_data$Annual.Income..k..)
hist(customer_data$Annual.Income..k..,
  col="#660033",
  main="Histogram for Annual Income",
  xlab="Annual Income Class",
  ylab="Frequency",
  labels=TRUE)
```

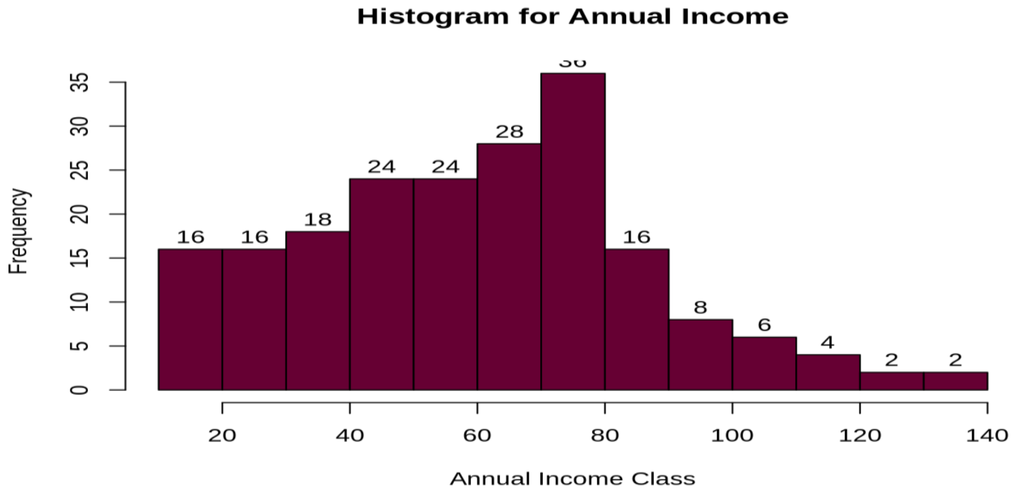
Screenshot:



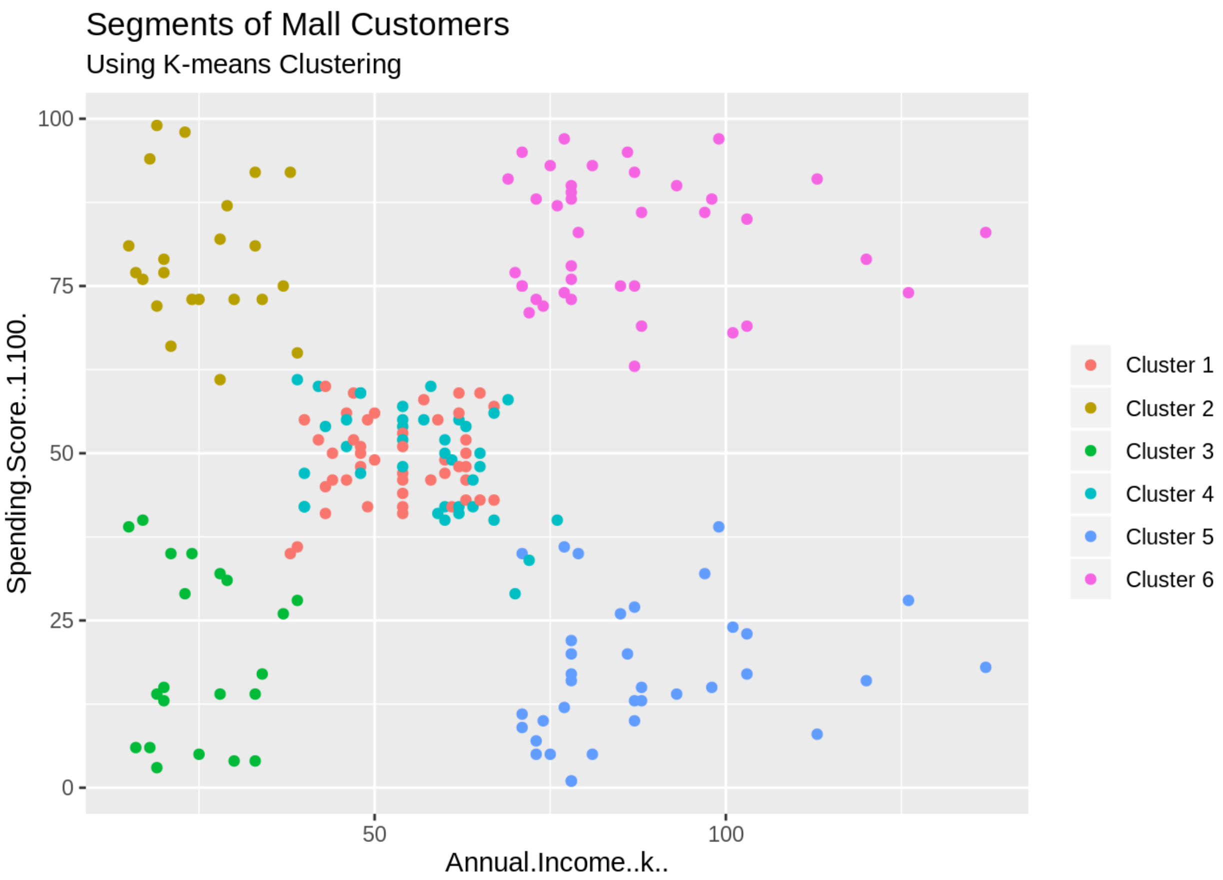
```
summary(customer_data$Annual.Income..k..)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00
```

```
hist(customer_data$Annual.Income..k..,
    col="#660033",
    main="Histogram for Annual Income",
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)
```



Visualizing the Clustering Results using the First Two Principle Components:



Cluster 6 and 4 : "These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary".

Cluster 1: "This cluster represents the customer_data having a high annual income as well as a high annual spend".

Cluster 3 : " This cluster denotes the customer_data with low annual income as well as low yearly spend of income".

Cluster 2 : " This cluster denotes a high annual income and low yearly spend".

Cluster 5: " This cluster represents a low annual income but its high yearly expenditure".

Tools

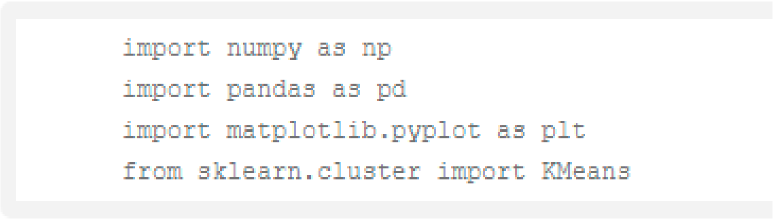
One of the most popular Machine Learning algorithms is K-means clustering. It is an unsupervised learning algorithm, meaning that it is used for unlabeled datasets. Imagine that you have several points spread over an n-dimensional space.

We can use K-means over random data using Python libraries.

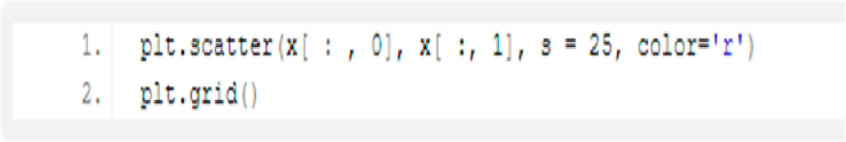
1. First, we import the essential Python Libraries required for implementing our k-means algorithm

2. We then randomly generate 200 values divided in two clusters of 100 data points each.

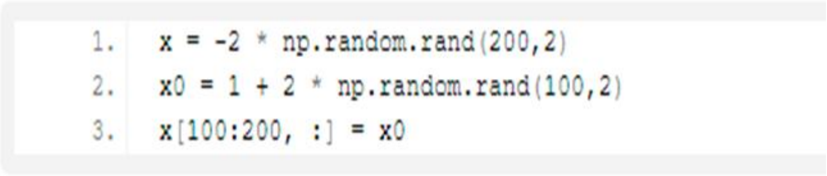
3. We proceed to plot our generated random values and obtain the following graph.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```



```
plt.scatter(x[ : , 0], x[ :, 1], s = 25, color='r')
plt.grid()
```



```
x = -2 * np.random.rand(200,2)
x0 = 1 + 2 * np.random.rand(100,2)
x[100:200, :] = x0
```

The first step of using k-means clustering is to determine the number of clusters (K) that we are going to produce the final output. This algorithm needs to select k objects from dataset randomly that it plays role as the initial centers for our clusters. These selected clusters called means or centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as “cluster assignment”. When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration.

Summing up the K-means clustering:

- “We specify the number of clusters that we need to create.
- The algorithm selects k objects at random from the dataset. This object is the initial cluster or mean.
- The closest centroid obtains the assignment of a new observation. We base this assignment on the Euclidean Distance between object and the centroid.

- k clusters in the data points update the centroid through calculation of the new mean values present in all the data points of the cluster. The kth cluster's centroid has a length of p that contains means of all variables for observations in the k-th cluster. We denote the number of variables with p.
- Iterative minimization of the total within the sum of squares. Then through the iterative minimization of the total sum of the square, the assignment stop wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations".

Lesson Learned

In this data science project, as you know we went through customer segmentation model. We used a class of machine learning called unsupervised learning. As I mentioned in this report, we used a clustering algorithm called k-means clustering. In this project we learned how effective is this K-means clustering in this project. We analyzed and visualized the data and then proceeded to implement our algorithm.