**Customer Segmentation**

**(using K-Means)**

**Project Report**

Submitted

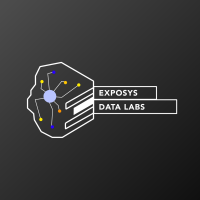
**By**

Name - Mayur Ramchandra Machhi

Email - [mayurmachhi19@gmail.com](mailto:mayurmachhi19@gmail.com)

Contact ­– 8378023120 / 7972843446d

**For**



EXPOSYS DATA LABS

Bengaluru-India

**ABSTRACT**

In our project, we will perform one of the most essential applications of machine learning Customer Segmentation by using K-Means Clustering Algorithm. In this project, we will implement customer segmentation in Python. Whenever you need to find your best customer, customer segmentation is the ideal methodology. Then we will explore the data upon which we will be building our segmentation model. Also, in this data science project, 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**

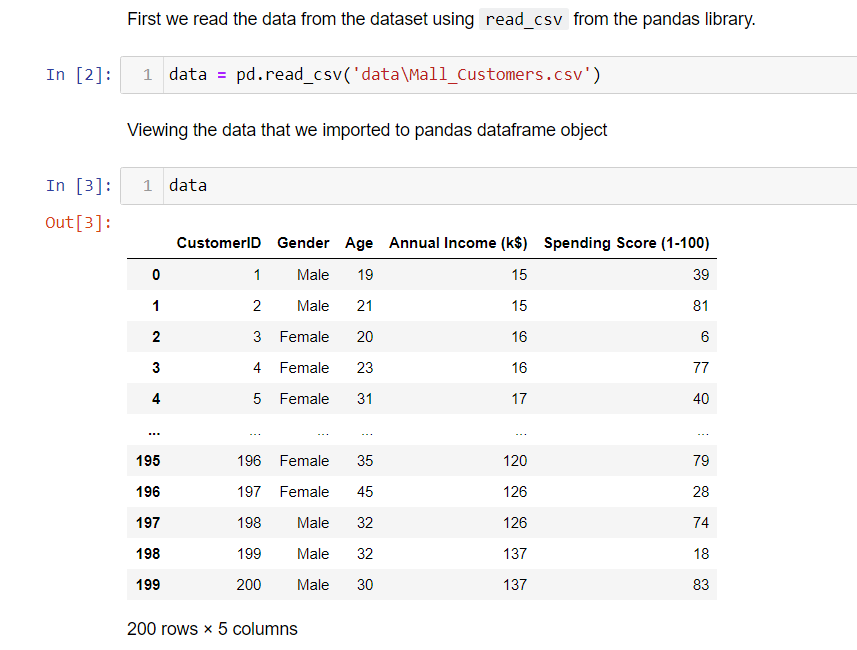
Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.

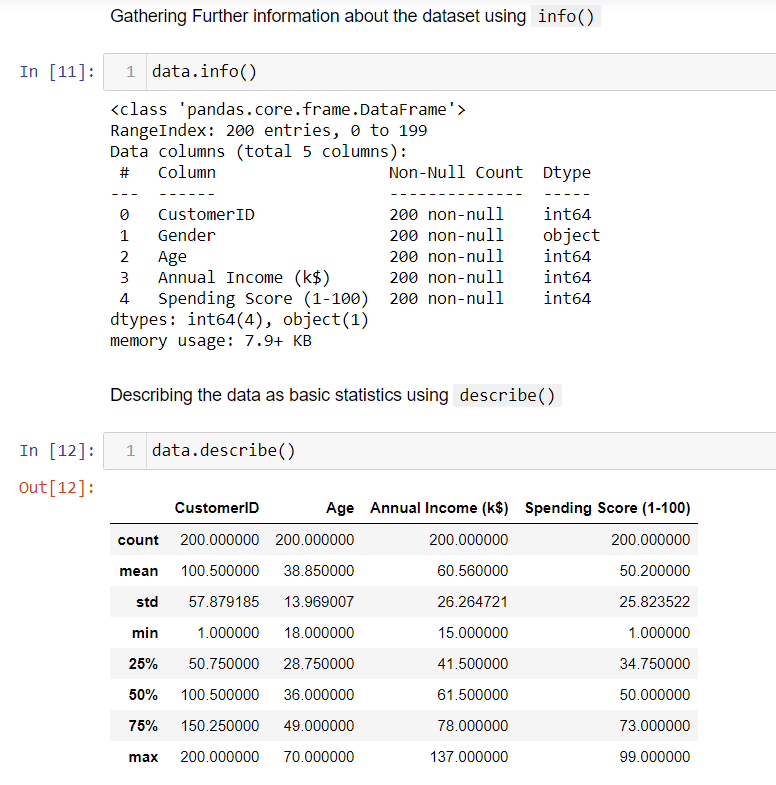
**Dataset Description:**

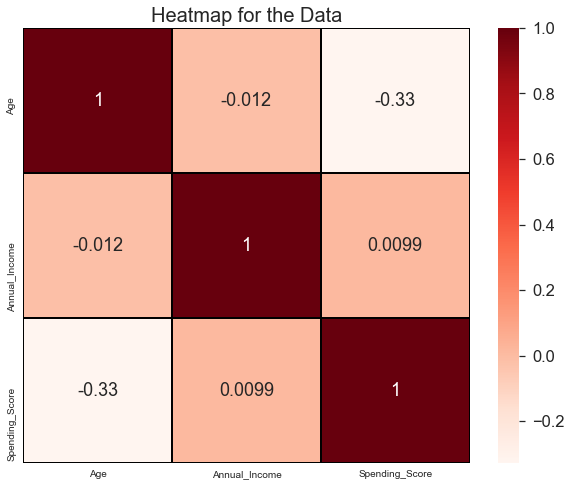
Mall Customer Segmentation Data:

The data is given by Exposys Data Labs. It has individual unique customer IDs, A categorical variable in the form of Gender and three columns of Age, Annual Income and Spending Score which will be our main targets to identify the patterns in the customers shopping and spending spree.

Data URL – drive.google.com/file/d/19BOhwz52NUY3dg8XErVYglctpr5sjTy4

****

****

****

**Purpose:**

To find the best customer, using customer segmentation methodology. To explore the data upon which building a segmentation model. Also, in this project, we will see the descriptive analysis of our data and then implement the K-means algorithm.

**Objectives:**

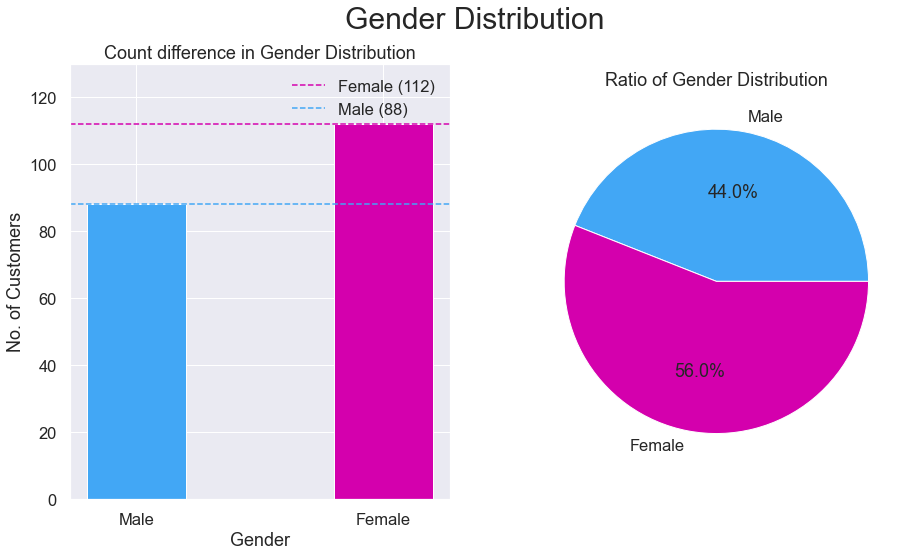
The objective of the project are as follows:

• Identify the potential customer base for selling the product.

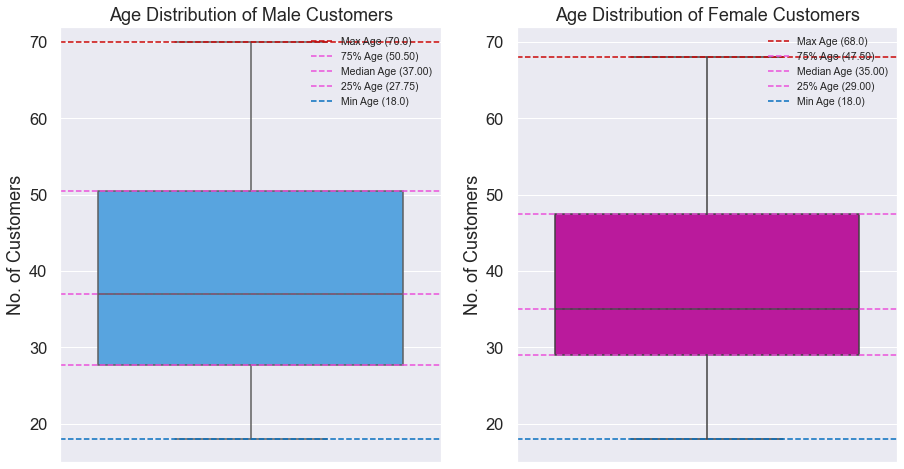
• Implement Clustering Algorithms to group the customer base.

**Exploratory Data Analysis:**

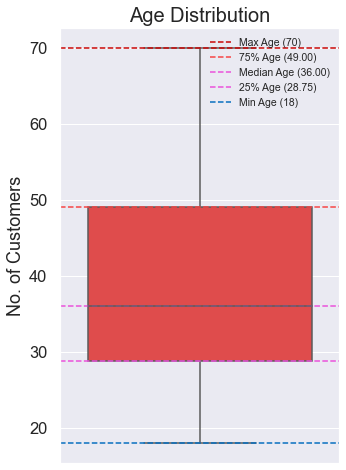
Visualization of Distribution of Males and Females:



From the above graphs, we observe that the number of females(112) is higher than the males(88). The Ratio of Gender population is 56% Females and 44% Males. By this we can say that majority of the customers that visit the mall are Females.

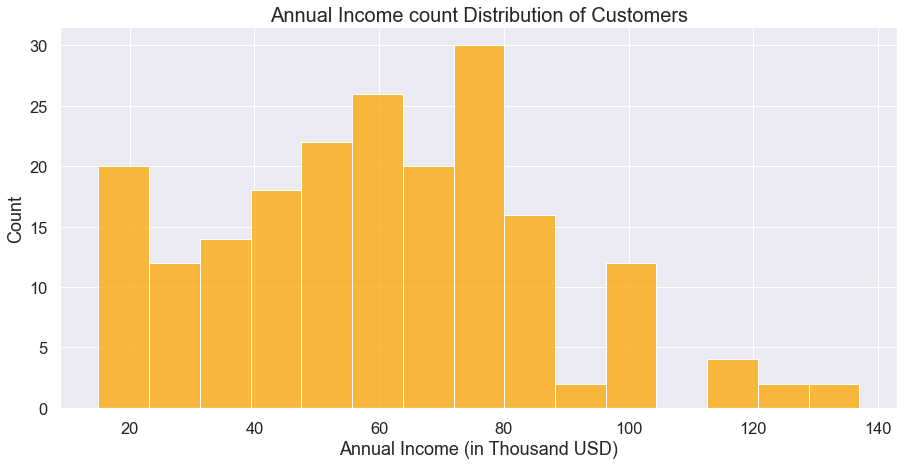


Age Analysis of Customers:

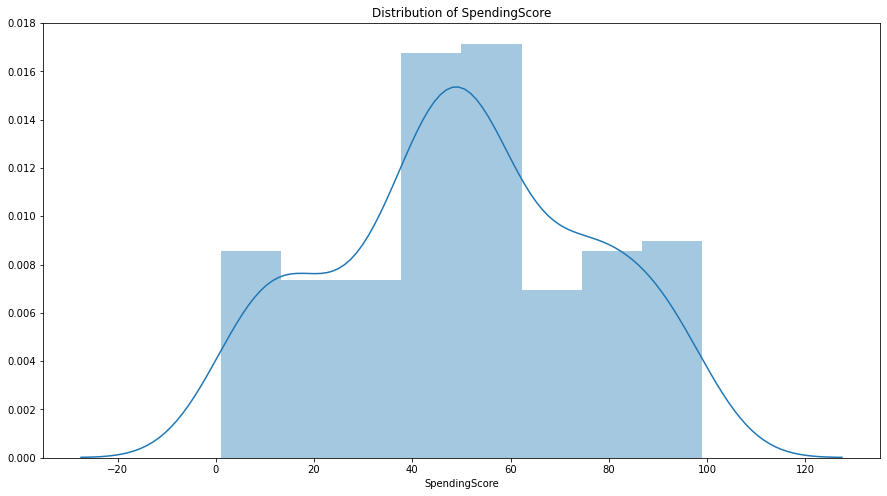


From the above boxplot, we can conclude that a large amount of ages are between 30 and 35. Min Age is 18, Max Age is 70. By comparing the age distribution of the customers, we can conclude that most of the customers were within the band between 30 to 50, where the mean is around 35 years old.

Annual Income and Spending Score Analysis:



The distribution of Annal Income and Spending Score exhibited an approximation of normal distribution, with highest density around the mean of the variables. The maximum and minimum of Annual Income are 137 and 15 respectively, with the mean at 60.56. From the plot, we can see that the peak of the distribution fell in the region of 60 to 75.

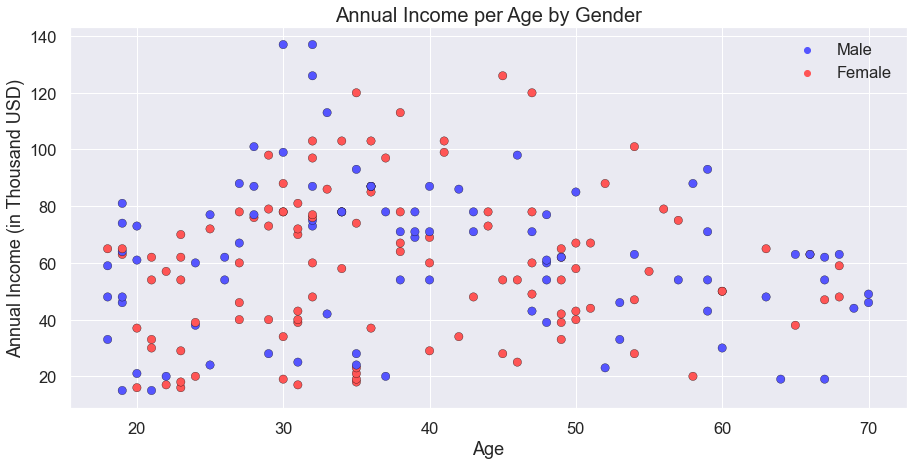


For the Spending score, the maximum and minimum are 99 and 1, while the histplot indicated that the highest number of customers have the spending score ranging from 40 to 60.

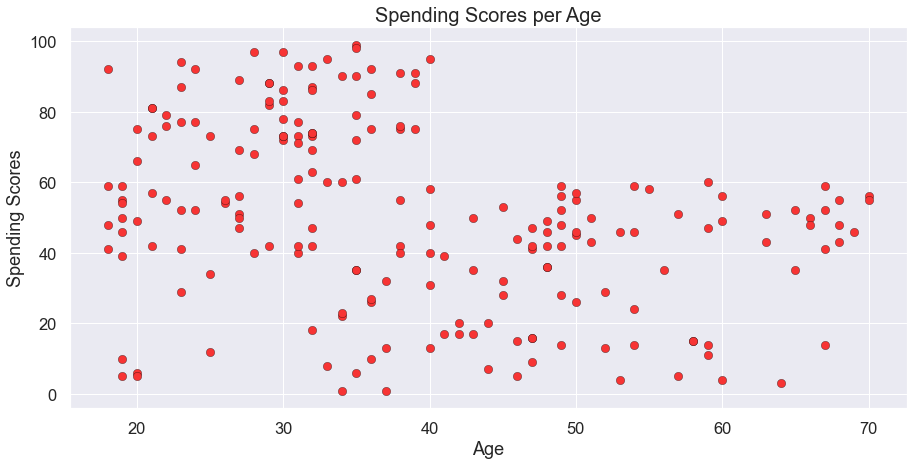
**Characteristic relations:**

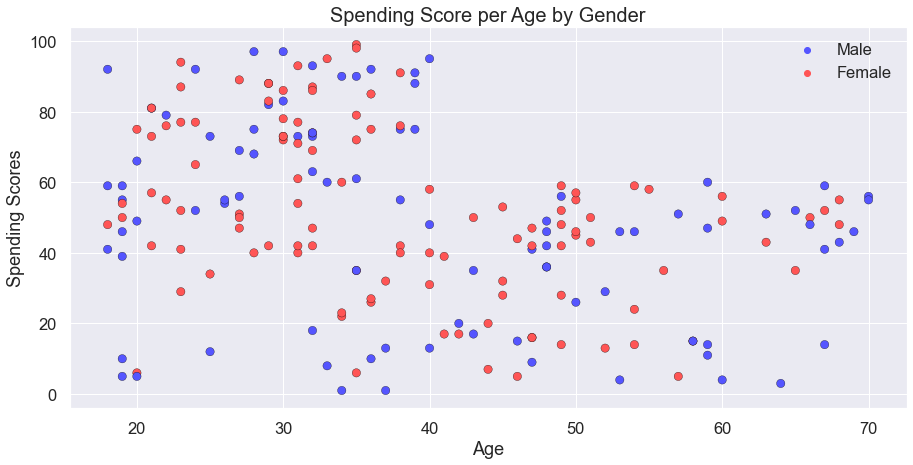
Annual Income vs Age analysis:





Spending Score vs Age analysis:





**METHODOLOGY**

In our project we used following packages:

* Pandas
* Numpy
* Matplotlib
* Scikit Learn
* Seaborn

**Pandas:**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.

**Numpy:**

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

**Matplotlib:**

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

**Scikit Learn:**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**Seaborn:**

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

**IMPLEMENTATION**

**What is Clustering?**

Imagine that you have a group of chocolates and liquorice candies. You are required to separate the two eatables. Intuitively, you are able to separate them based on their appearances. The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group.

Clustering is used in various fields like image recognition, pattern analysis, medical informatics, genomics, data compression etc. It is part of the unsupervised learning algorithm in machine learning. This is because the data-points present are not labelled and there is no explicit mapping of input and outputs. As such, based on the patterns present inside, clustering takes place.

**K-Means Clustering:**

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. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

We then proceeded to perform K-means Clustering which will create different clusters to group similar spending activity based on their age and annual income. K-Means Clustering selects random values from the data and forms clusters assigned. The closest values from the centre of each cluster were taken to update the cluster and reshape the plot (just like k-NN). The closest values are based on Euclidean Distance.