CUSTOMER SEGMENTATION

(USING K-MEANS)

PROJECT REPORT

SUBMITTED

BY

NAME- ANSHU RAJ

[EMAIL-2018UMT1807@MNIT.AC.IN](mailto:EMAIL-2018UMT1807@MNIT.AC.IN)

CONTACT-7281081395/6204744265

ABSTRACT

Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively. This illustrates how firms can construct novel and inventive approaches that provide great value.

In this project we will perform Customer Segmentation using K-Means Clustering Algorithms. For this, we will use Python as our programming language and we will be using several of its libraries for visualization and analysis of our data. 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.

Table Of Contents

1. Introduction
   1. Dataset Description
   2. Purpose
   3. Objectives
2. Existing and Proposed method with architecture
   1. What is Clustering?
   2. Various Clustering Algorithms
   3. K-means Clustering
3. Methodology
   1. Jupyter Notebook
   2. Packages/ Libraries Used
   3. Exploratory Data Analysis
4. Implementation
5. Result and Conclusion

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:

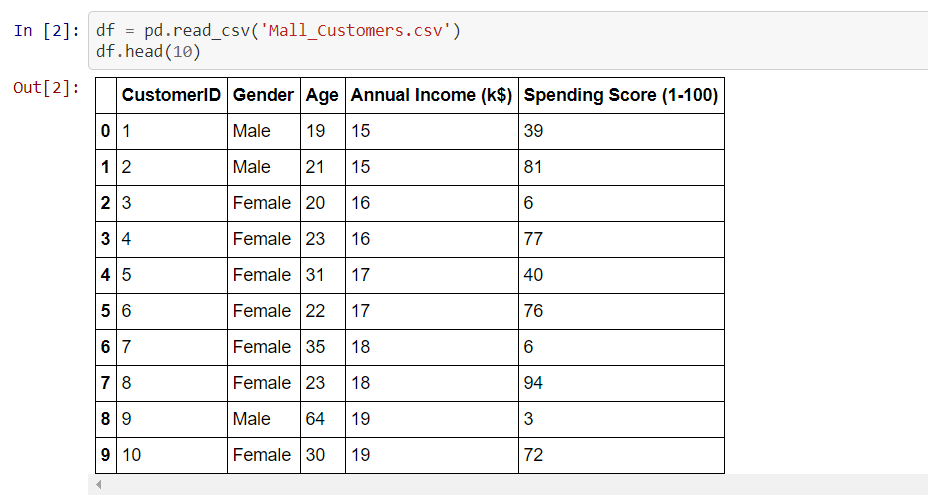
The data was given by Exposys Data Labs. The columns in dataset are: CustomerId, Gender, Age, Annual Income (k$), Spending Score (1-100).

The data can be found here :

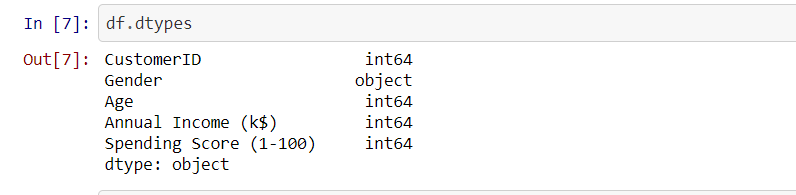
URL: <https://drive.google.com/file/d/19BOhwz52NUY3dg8XErVYglctpr5sjTy4>

First we will import the csv file into a pandas dataframe.

After this the dataset looks like this.



The ‘Gender’ column is a categorical variable while the rest of the columns are of integer value.



The dataset contains 5 columns and 200 rows.



**Purpose:**

The purpose of this project is to build a clustering model for customer segmentation of mall customers so that a proper business model can be adopted to maximise sale in the mall.

**Objectives:**

The objectives of this course are as follows:

* Implement Clustering Algorithms to group the customers of a mall in customer dataset.
* Identify the potential customer base for selling the product.

Existing Methods and Proposed Method with Architecture

**What is Clustering?**

Clustering means finding clusters in a dataset. A cluster is a group of data points or objects in a dataset that are like other objects in the group, and dissimilar to datapoints in other clusters. Clustering is a form of unsupervised machine learning. This is because the data points present are not labelled and there is no explicit mapping of input and output. As such, based on the pattern inside, clustering takes place.

**Various Clustering Algorithms**

1. K-Means Clustering
2. Mean shift Clustering
3. Agglomerative Hierarchical Clustering
4. Density Based Spatial Clustering of Applications with Noise (DBSCAN)

E.T.C

Among the above types of clustering algorithms, for customer segmentation K-Means Clustering is used.

**Proposed Method:** K-Means Clustering.

**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.

Create clusters based on smallest distance.

Objects move to clusters.

Recalculate the new cluster centres.

Put object to closest cluster centres.

Set initial cluster centre randomly.

Select the number of cluster centres.

Methodology

**Jupyter Notebook** was used in this project. It is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

**Following packages were used in this project:**

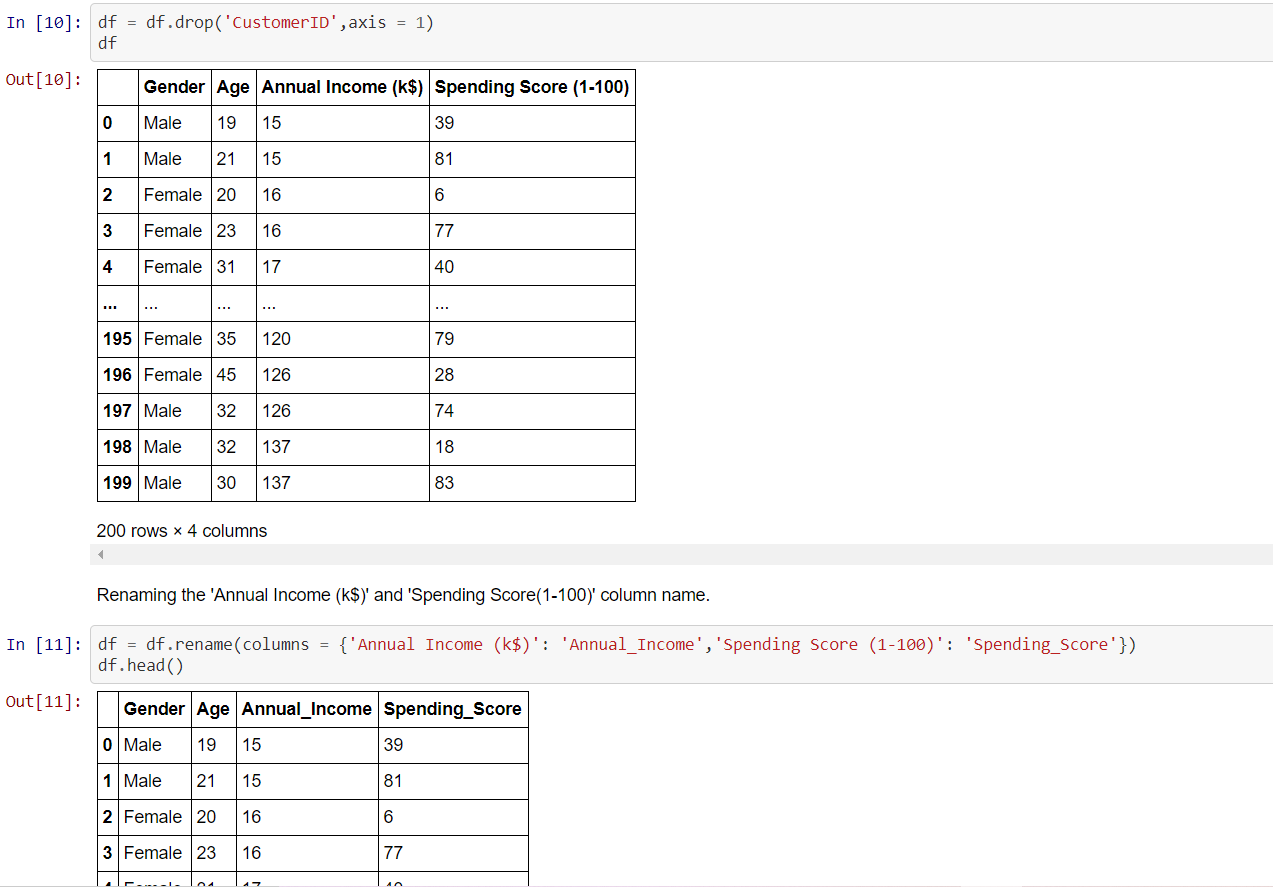
1. Pandas (version-1.2.4)
2. NumPy (version-1.20.3)
3. Matplotlib (version-3.4.2)
4. Scikit-Learn (version-0.24.2)
5. Seaborn (version-0.11.1)

**Exploratory Data Analysis:**

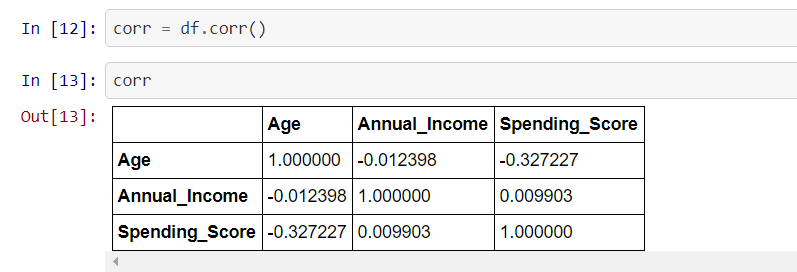
Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

First the dataset was cleaned a little bit.

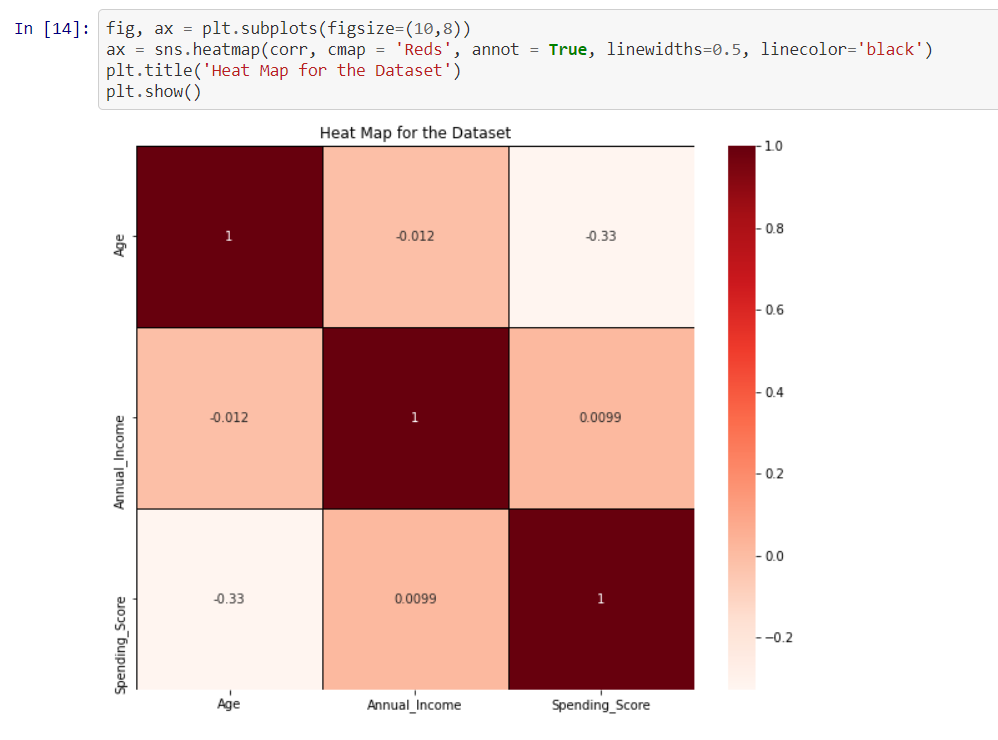
We dropped the ‘Customer Id’ column as this was redundant. We also changed the column names from ‘Annual Income (k$)’ and ‘Spending Score (1-100) to ‘Annual\_Income’ and ‘Spending\_Score’ respectively for our convenience.



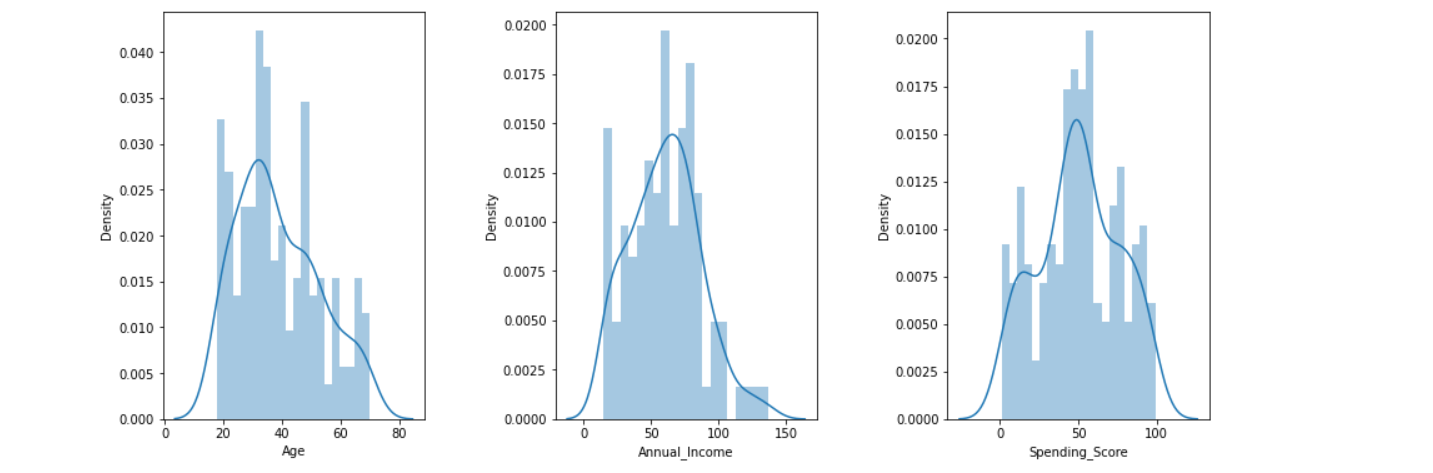
Then the corelation between the features was checked.



A heat map of the corelation was plotted.

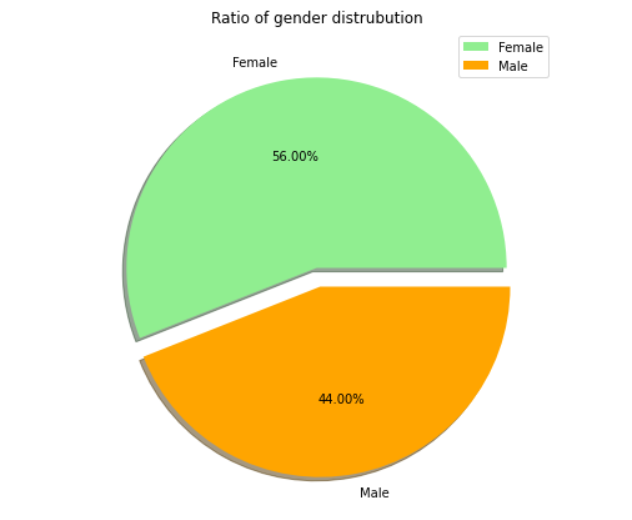


Now we visualized the dataset through Distplot to get the distribution of the data.



In these histograms we can observe that the distribution of these values resembles a Gaussian distribution, where the vast majority of the values lay in the middle with some exceptions in the extremes.

Gender Distribution:

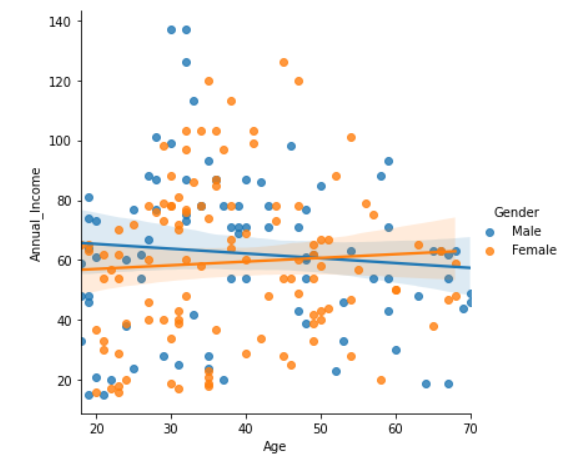


It is clear from above chart that there are a greater number of females than males.

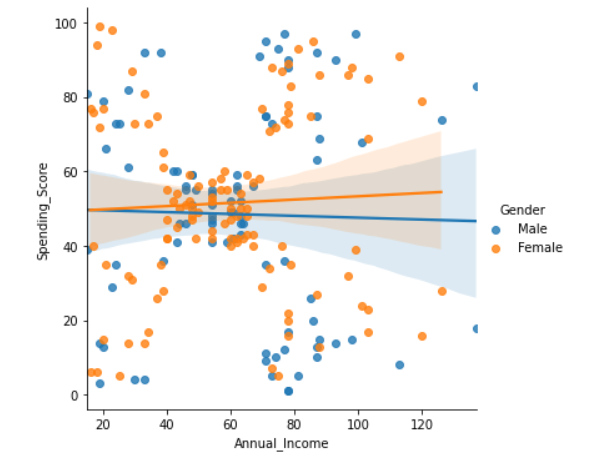
Now relation between various features were visualized.

Age and Annual Income.

In the below plot we can clearly see how people in their thirties, forties and fifties tend to earn more money annually than the ones younger than thirty or older than fifty years old. We can also see how males tend to earn a little bit more money than females, at least until fifty years old.

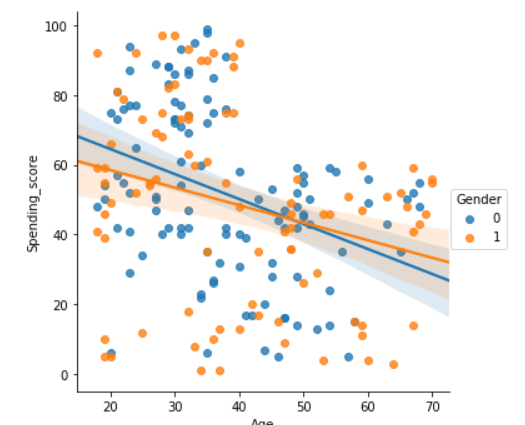


Spending Score and Annual Income



Here we can observe how a better annual income leads to having a higher spending score, specially for women.

Age and Spending Score

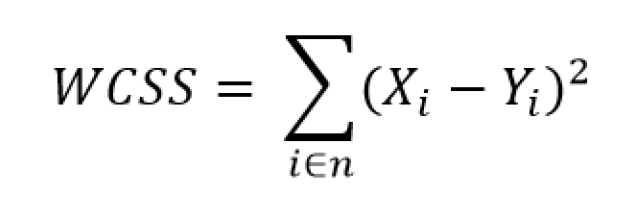


We could predict, young people tend to spend way more than older people from the above plot.

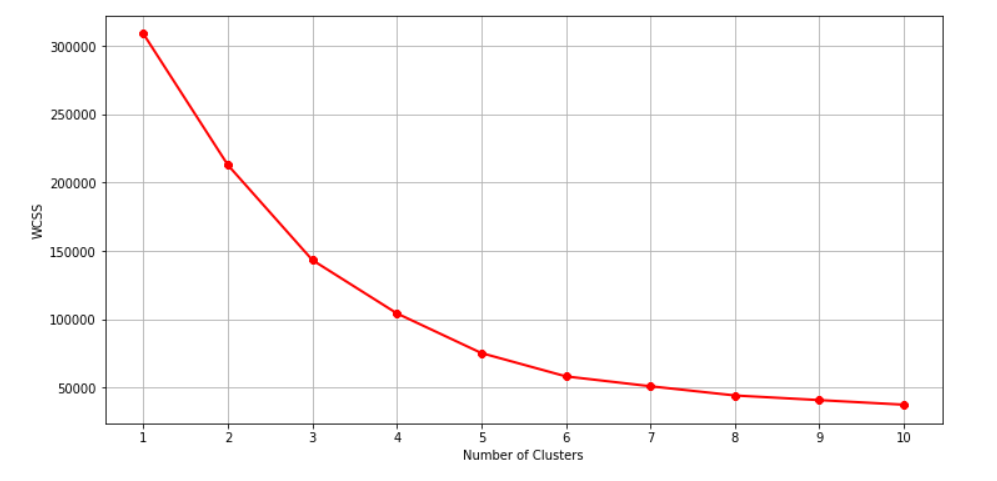
Implementation

Before building the k-means model we determined the number of clusters we need to make. For this we used the very popular Elbow Method.

We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.



where Yi is centroid for observation Xi. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.



If we consider the above plot, the elbow can be found, approximately, where the number of clusters is equal to 5. Therefore, we are selecting **5** as the number of clusters to divide our data in.

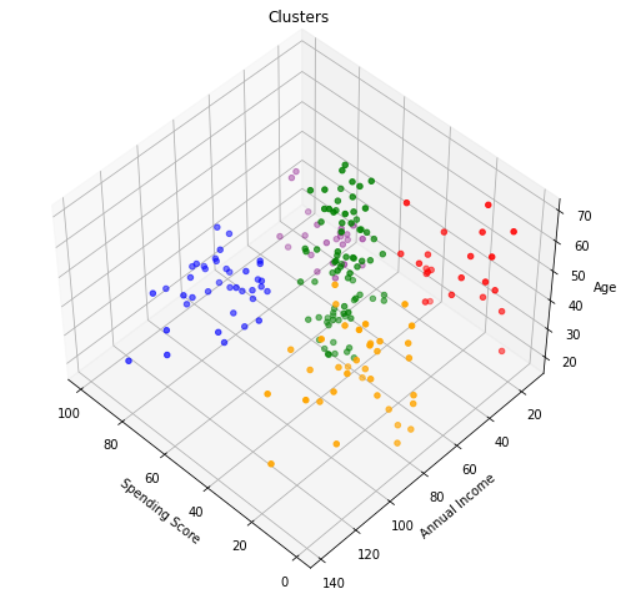
**Clustering:**

In the process of clustering, we did not consider the gender factor anymore. The first main reason of why we did take this approach is because the difference between male and female in this data is not particularly high and making a gender differentiation won't provide any further information.

Result and Conclusion

**Result:**

The following clusters were generated.



1. Purple Cluster: Groups young people with moderate to low annual income who spend a lot.
2. Blue Cluster: Groups reasonably young people with pretty decent salaries who spend a lot.
3. Green Cluster: Groups people of all ages whose salary isn’t pretty high and their spending score is moderate.
4. Orange Cluster: Groups people who actually have pretty decent salary but barely spends money.
5. Red Cluster: Groups people whose salary is pretty low and don’t spend much money. They are people of all ages.

**Conclusions:**

After developing the solution for the model we have come to following conclusions:

* K-Means Clustering is a powerful technique in order to achieve a decent customer segmentation.
* Customer segmentation is a good way to understand the behaviour of different customers and plan a good marketing strategy accordingly.
* There is not a much difference between the spending score of men and women.
* Observing the clustering plot, it can be clearly observed that the ones who spend more money in malls are young people(purple and blue cluster). They are the main target when it comes to marketing.
* Although young people seem to be the ones spending the most, we can't forget there are more people we have to consider, like people who belong to the pink cluster, they are what we would commonly name after "middle class" and it seems to be the biggest cluster.