**EXPOSYS DATA LABS DATA SCIENCE INTERNSHIP**

**PROJECT REPORT**

**Problem Statement**: 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. Use K-means clustering and also visualize the gender and age distributions. Then analyze their annual incomes and spending scores.

***What Is Customer Segmentation?***

Customer segmentation, also known as market segmentation, is the practice of dividing consumers into segments that can be focused on. Data, for example, a clients’ socioeconomics, topography, psychographic and behavioural preferences are considered while deciding customer segmentation.

Segmenting a market according to customers is sound practice. It empowers you to build up a more profound comprehension of your clients and find what influences them to tick. When you’re conveying a message, it will be more compelling if the beneficiary of the message thinks that its applicable.

Division is essentially a method for orchestrating your clients into littler gatherings as indicated by write. These unmistakable sub-gatherings or fragments ought to be described by specific traits. Presently you can target particular, pertinent advertising messages at each gathering.

Furthermore, it’s not just about what you say. How you convey is additionally essential, and division frequently requires a precisely organized advertising blend. That is on the grounds that a few clients may favour the immediate approach, for example, phone promoting, while others react better to a neighbourhood publicising effort.

The most common ways in which businesses segment their customer base are:

1. **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.
2. **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer’s city, state, or even country of residence.
3. **Psychographics**, such as social class, lifestyle, and personality traits.
4. **Behavioral data**, such as spending and consumption habits, product/service usage, and desired benefits.

### ****Customer segmentation procedures****

Customer segmentation, also called consumer segmentation or client segmentation, procedures include:

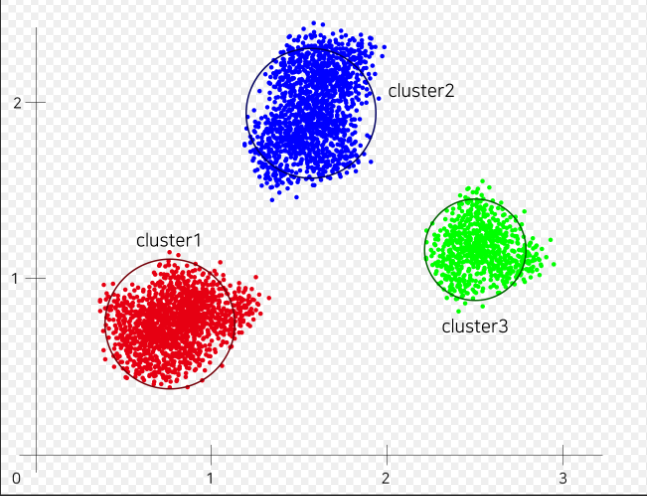
* Deciding what data will be collected and how it will be gathered
* Collecting data and integrating data from various sources
* Developing methods of [data analysis](https://searchsqlserver.techtarget.com/definition/data-mining) for segmentation
* Establishing effective communication among relevant business units (such as marketing and customer service) about the segmentation
* Implementing applications to effectively deal with the data and respond to the information it provides

# Advantages of Customer Segmentation

1. Determine appropriate product pricing.
2. Develop customized marketing campaigns.
3. Design an optimal distribution strategy.
4. Choose specific product features for deployment.
5. Prioritize new product development efforts.

# K Means Clustering Algorithm

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.



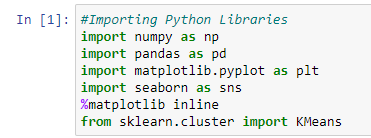
K Means Clustering where K=3

# Environment and tools used

* Python 3.8.2
* Jupyter Notebook
* numpy 1.19.1
* pandas 1.1.0
* matplotlib 3.3.0
* seaborn 0.10.1
* scikit-learn 0.23.2

CODE & DATA ANALYSIS:

**First importing the necessary Python Libraries-**



*Numpy for linear algebra*

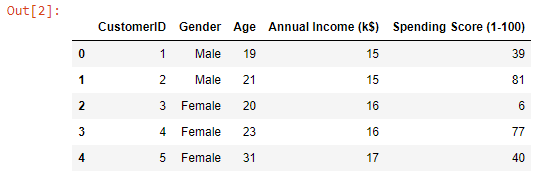
*Pandas for data processing, CSV file I/O*

*Matplotlib for data visualization*

*Seaborn for data visualization*

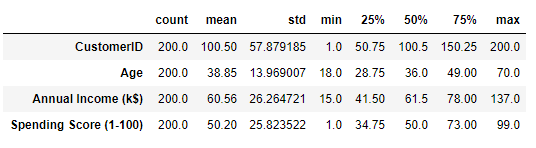
*Scikit-learn for predictive data analysis*

**Importing the dataset and viewing the first five rows of the data-**

*So we can see the columns in the dataset are CustomerId, Gender, Age, Annual Income and Spending Score.*

**To understand more about the data data.describe() is used**

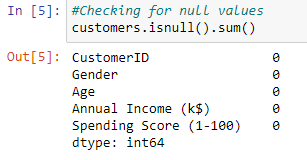
We can see that-

* *The count is 200 means we have records of 200 customers with us.*
* *The minimum age of customer in our data is 18 yrs and maximum age is 70.*

*The mean here is 38 and median is 36.Here Mean>Median means our data has high outliers that is more of youngsters prefer to go malls.*

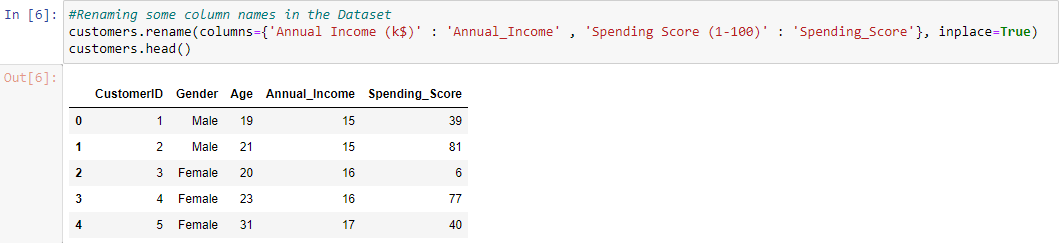
* *The minimum annual income of customer is 15k$ and maximum is 137k$. The mean and median here is 60k$ and 61k$ respectively.*
* *Spending Score is something you assign to the customer based on your defined parameters like customer behaviour and purchasing data. Here the minimum spending score assigned is 1 and maximum ranges till 99.Both mean and median is 50.*

**now checking whether the data has null value or not –**



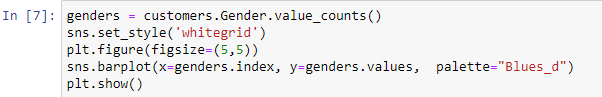
*So, there is no null values in the dataset.*

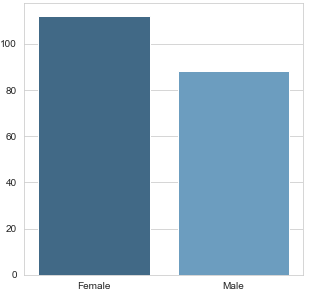
**Renaming some columns for the ease of use –**



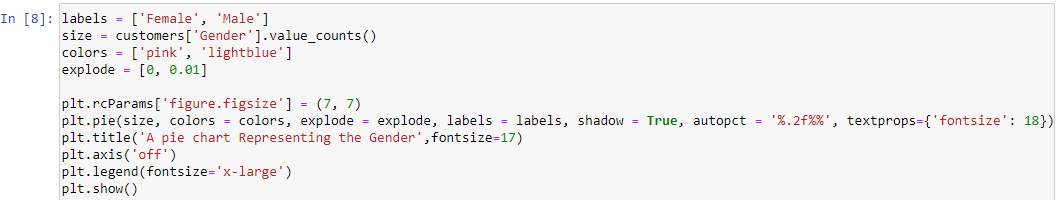
**Checking the distribution of male and female population in the dataset. –**

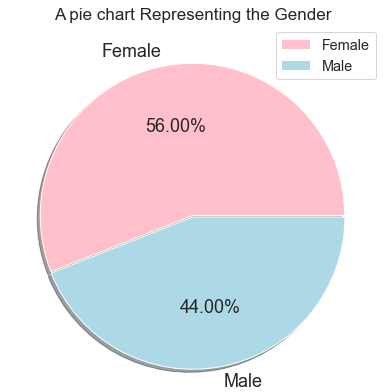
* Bar Plot :





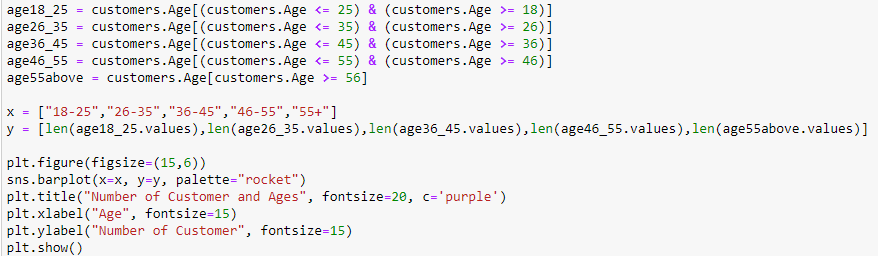
* Pie Chart :

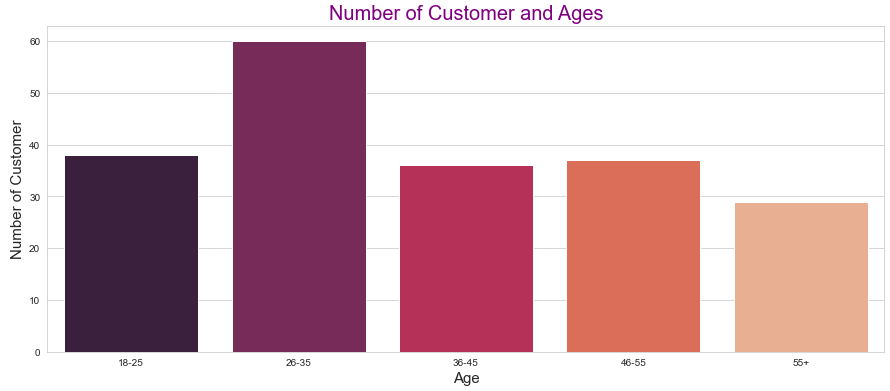




*Interestingly, The Females are in the lead with a share of 56% whereas the Males have a share of 44%, that’s a huge gap specially when the population of Males is comparatively higher than Females.*

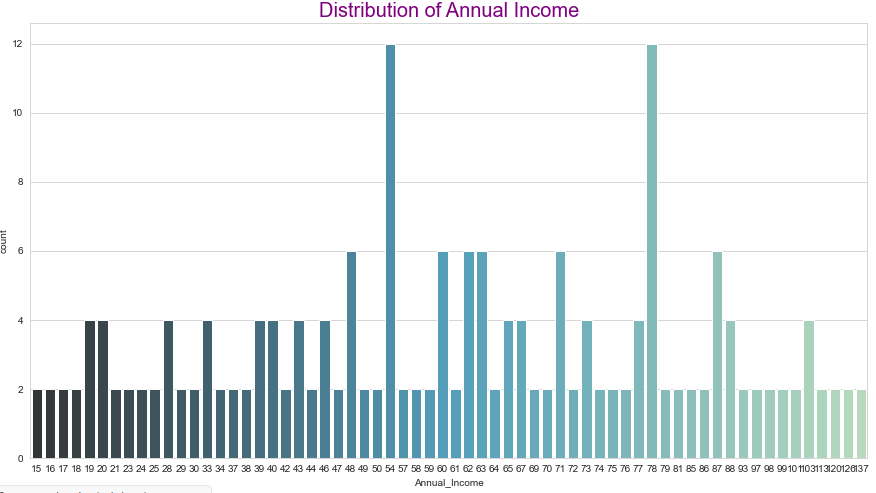
## Checking for the Distribution of number of customers in each age group





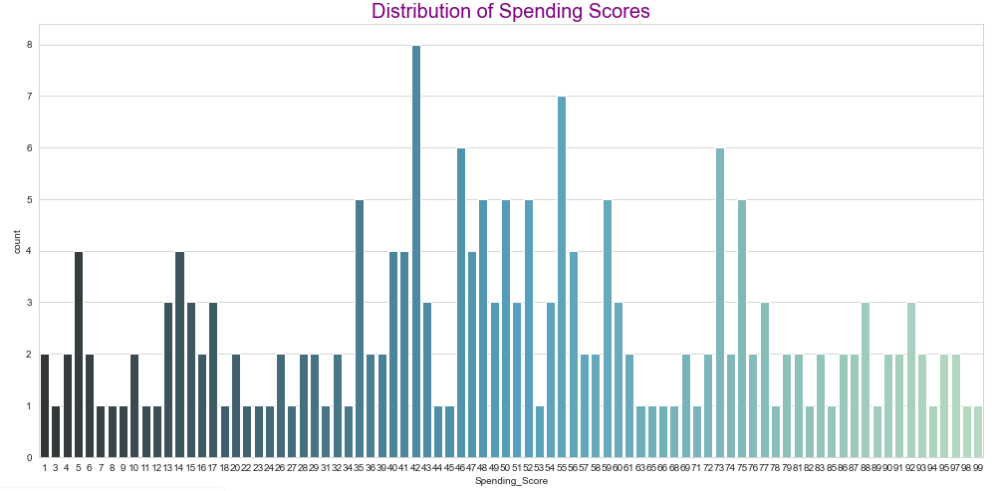
*So most of the customers falls under the age group of 26 to 35 so they are the most frequent visitors in the mall where the people of ages 55 and above are very less frequent in the mall.*

**Now I made a count plot to see the distribution of Annual Income –**

****

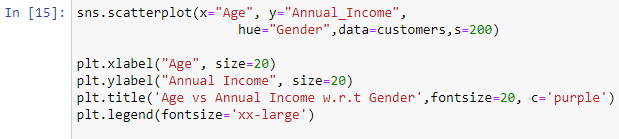
*Interestingly there are customers in the mall with a very much comparable frequency with their Annual Income ranging from 15 US Dollars to 137K US Dollars. There are more Customers in the Mall who have their Annual Income as 54k US Dollars or 78 US Dollars.*

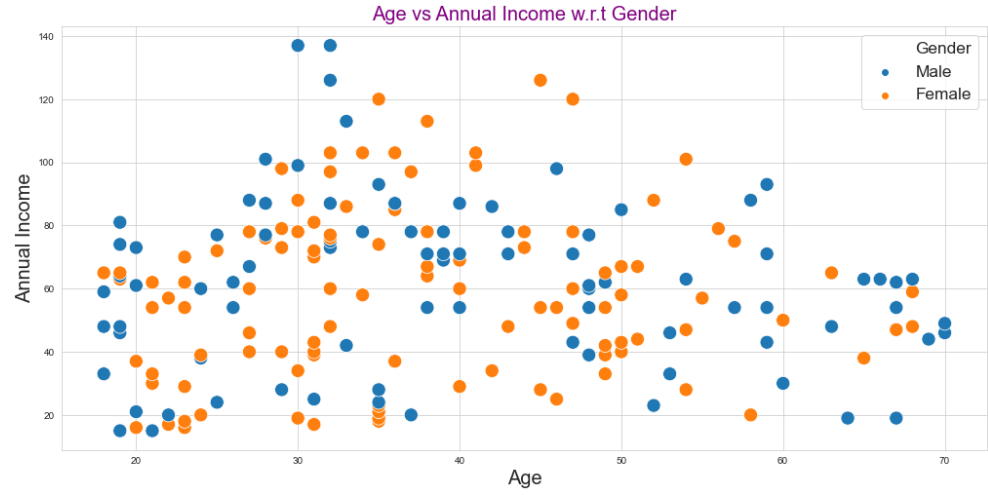
**Similarly I made a count plot to see the distribution of Spending Score –**

****

*On a general level, we may conclude that most of the Customers have their Spending Score in the range of 35–60. Interesting there are customers having 1 spending score also, and 99 Spending score also, Which shows that the mall cares to the variety of Customers with Varying needs and requirements available in the Mall.*

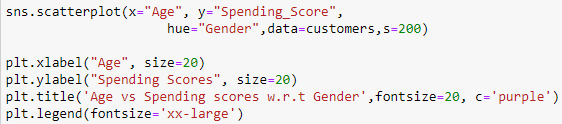
**Age vs Annual Income –**

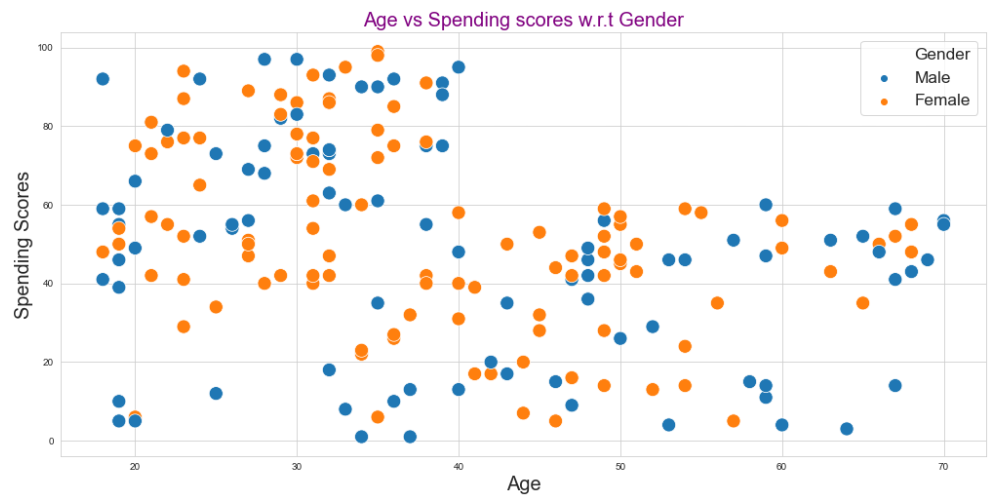




*We can see in scatter plot that most of the people annual income is more when there age is in between 20 to 50.*

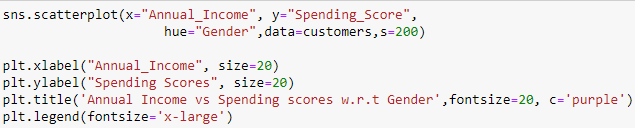
**Age vs Spending scores :**

****



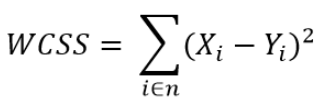
*We can see in the scatter plot that people of ages between 20 and 40 spend more*

**Annual Income vs Spending Scores –**

****

*\*

Next I plotted 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.

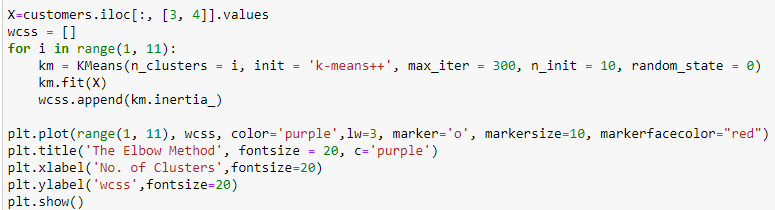
# The Elbow Method

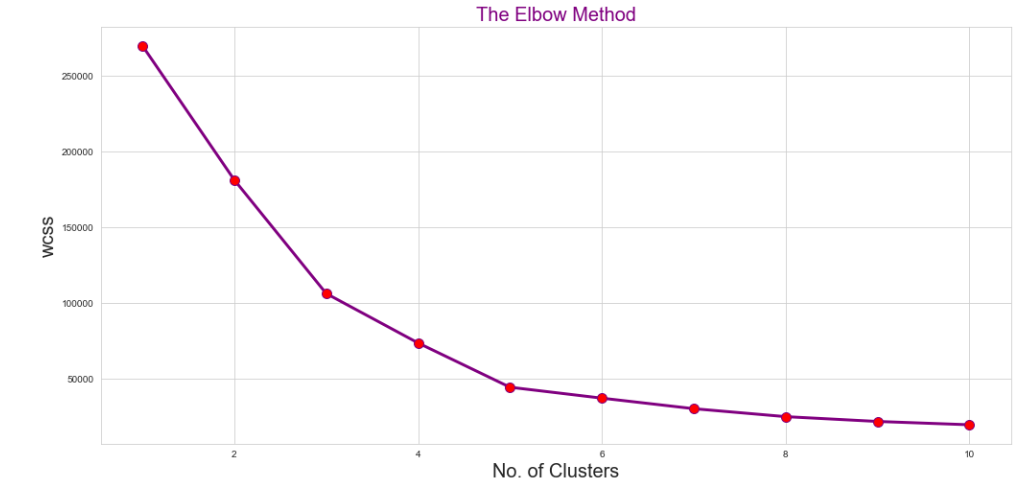
Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

The steps can be summarized in the below steps:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WCSS).
3. Plot the curve of WCSS vs the number of clusters K.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

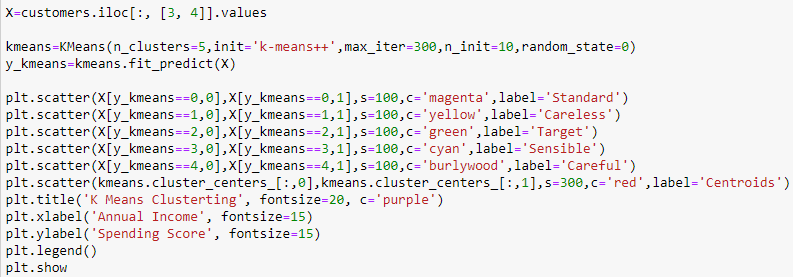
**choosing the optimal no of clusters by plotting elbow plot.**

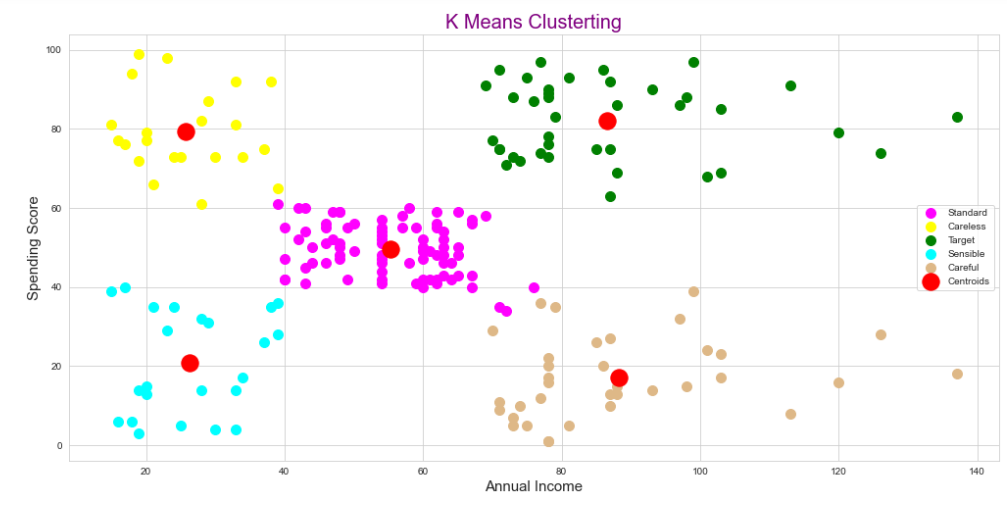
****



*So we found the optimal value of k is 5 using elbow method.*

**Visualizing the clusters**

**



*Cluster 1- High income low spending =Careful*

*Cluster 2- Medium income medium spending =Standard*

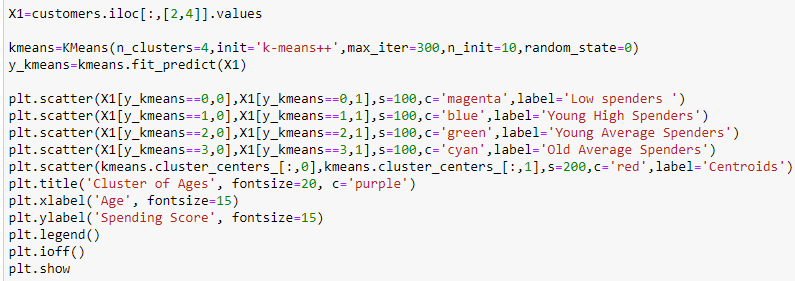
*Cluster 3- High Income and high spending =Target*

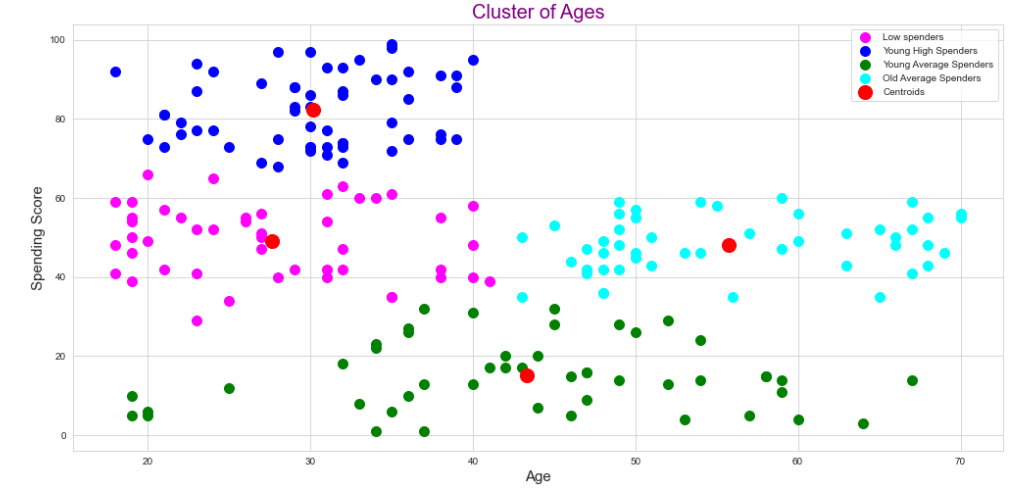
*Cluster 4- Low Income and high spending =Careless*

*Cluster 5- Low Income and low spending =Sensible*

*This Clustering Analysis gives us a very clear insight about the different segments of the customers in the Mall. There are clearly Five segments of Customers namely Careful, Standard, Target, Careless, Sensible based on their Annual Income and Spending Score which are reportedly the best factors/attributes to determine the segments of a customer in a Mall.*

**Finding the relation between Age and spending score**

**



*So we can cluster the data into four groups.*

*1.Low spenders*

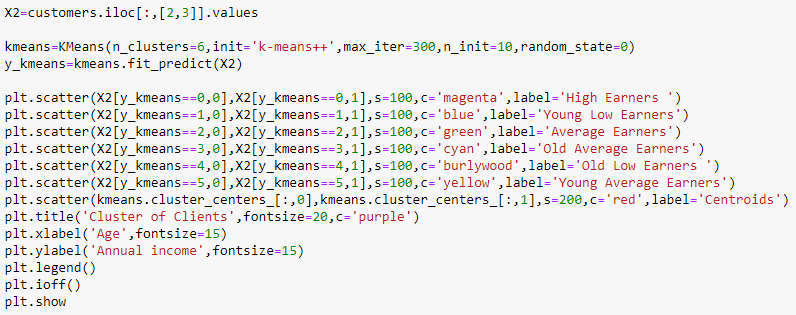
*2.Young High Spenders*

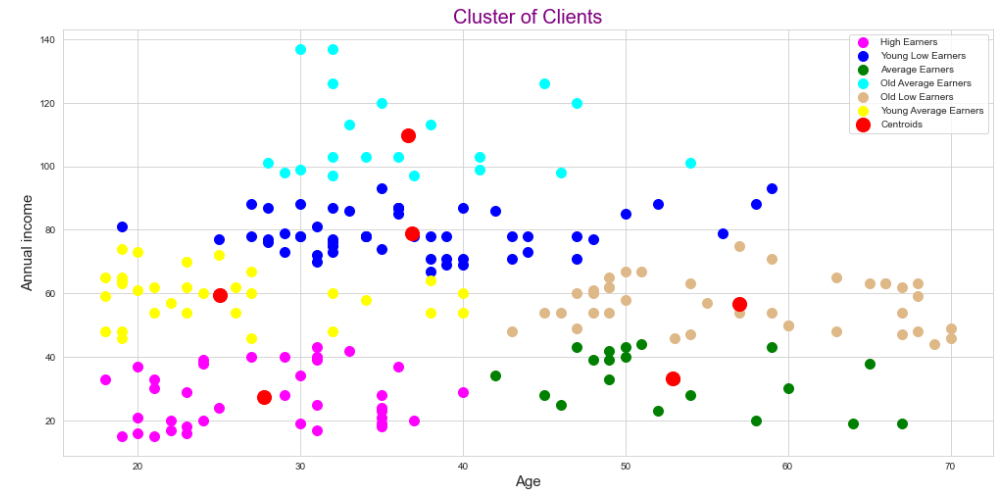
*3.Young Average Spenders*

*4.Old Average spenders*

*We can clearly see that Only young people(18-40 age group) are involved in High Spending.As age increases people fall into average or Low spending catogery.*

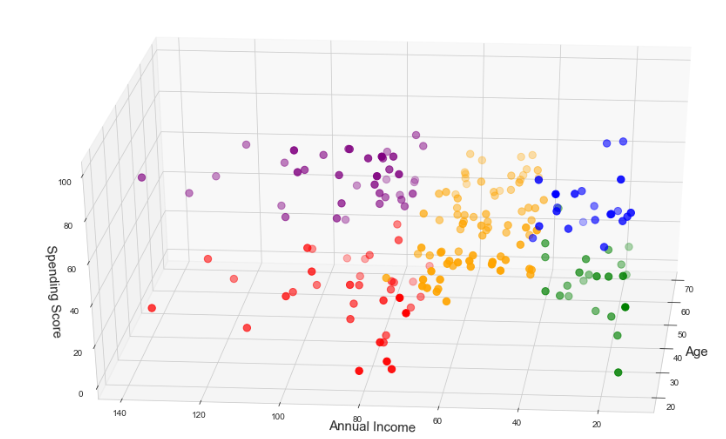
**Finding out the relation between Age and Income**





*We can see people in age group 0f 30-40 have high number of high income people*

**Finding out the relation between Age,Annual Income and spending score**

******

*It is a 3D plot to visualize the spending score of the customers with their annual income. The data points are separated into 5 classes which are represented in different colours as shown in the 3D plot.*

**Conclusions**

*Now we have 4 different clusters grouped by Age and Spending Score.*

*The analysis shows there is low score concentration in male gender (between 0 and 25 score points). In female gender, we have high concentration in ranges between 75 and 100 compared to male gender. In general, women have higher Spending Score than men.*

*In other hand, the Annual Income distribution shows that in general, men have higher annual income than women. These two analysi together could give good insights for mall administrators.*

*Senior Spending Scores concentrates in low and medium values; In high score valuation, adults have the highest levels; In gender comparison, young and senior women have higher Spending Score values than young and senior men.*