Table of Contents

ABSTRACT

Mall Customer data is an interesting dataset that has hypothetical customer data. It puts you in the shoes of the owner of a supermarket. You have customer data, and on this basis of the data, you have to divide the customers into various groups. Malls do customer profiling to target the potential customers with irresistible offers whenever they visit the mall. In this tutorial, you will learn several EDA and customer profiling techniques with the help of unsupervised K-Means Clustering algorithm. Segmentation of market is an effective way to define and meet customer needs. Unsupervised Machine Learning Techniques, K- Means Clustering Algorithm, Mini batch K-Means and Hierarchical Clustering are used to perform Market Basket Analysis. Clustering is an iterative process of knowledge discovery from vast amounts of raw and unorganised data. The purpose of this analysis is to uncover underlying patterns in the customer base, and to groups of customers accordingly, often known as market segmentation. In doing so, the marketing team can have a more targeted approach to reach consumers, and the mall can make more informed strategic decisions to increase profits.

Introduction

Management and maintain of customer relationship have always played a vital role to provide business intelligence to organizations to build, manage and develop valuable long term customer relationships. The importance of treating customers as an organizations main asset is increasing in value in present day and era. Organizations have an interest to invest in the development of customer acquisition, maintenance and development strategies. The business intelligence has a vital role to play in allowing companies to use technical expertise to gain better customer knowledge and Programs for outreach. By using clustering techniques like k-means, customers with similar means are clustered together. Customer segmentation helps the marketing team to recognise and expose different customer segments that think differently and follow different purchasing strategies.

Customer segmentation helps in figuring out the customers who vary in terms of preferences, expectations, desires and attributes. The main purpose of performing customer segmentation is to group people, who have similar interest so that the marketing team can converge in an effective marketing plan. Clustering is an iterative process of knowledge discovery from vast amounts of

raw and unorganised data. Clustering is a type of exploratory data mining that is used in many applications, such as machine learning, classification and pattern recognition.

Role of the student in internship:

The role is to analyse the the type of people entering the mall . The age group and the type of ads that needs to be displayed to enable maximum profit to the organisation

Schedule of Internship:

1st September 2021 to 30th September 2022

DATA ACQUISITION AND CLEANING

The data includes the following features:

1. CustomerID : Unique ID assigned to the customer .
2. Gender : Gender of the customer .
3. Age : Age of the customer .
4. Annual Income (k$) : Annual Income of the customer .
5. Spending Score (1-100) : Score assigned by the mall based on customer behaviour and spending nature .

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 ie 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.

Load the downloaded .csv file into pandas dataframe:

```
In [7]: # Load the downloaded .csv file into pandas dataframe:
        data = pd.read_csv('Mall_Customers.csv')
        data.head()
```

Out[7]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Check the detailed summary of the dataset by calling the method:

```
In [16]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   CustomerID      200 non-null    int64 
 1   Gender          200 non-null    object
 2   Age             200 non-null    int64 
 3   annual_income   200 non-null    int64 
 4   spending_score  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Changing the name of some columns:

```
In [14]: #Changing the name of some columns
data = data.rename(columns={'Annual Income (k$)': 'annual_income', 'Spending Score (1-100)': 'spending_score'})
data.head()
```

Out[14]:

	CustomerID	Gender	Age	annual_income	spending_score
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Data Cleaning:

Checking for the null values:

```
In [18]: data.isnull().sum()
         # there is no null value

Out[18]: CustomerID        0
         Gender            0
         Age               0
         annual_income     0
         spending_score    0
         dtype: int64
```

Replacing categorical values with numerics:

```
In [20]: data.Gender.replace('Male', 0, inplace=True)
         data.Gender.replace('Female', 1, inplace=True)
         data
```

Out[20]:

	CustomerID	Gender	Age	annual_income	spending_score
0	1	0	19	15	39
1	2	0	21	15	81
2	3	1	20	16	6
3	4	1	23	16	77
4	5	1	31	17	40
...	...	...	...	...	...
195	196	1	35	120	79
196	197	1	45	126	28
197	198	0	32	126	74
198	199	0	32	137	18
199	200	0	30	137	83

200 rows × 5 columns

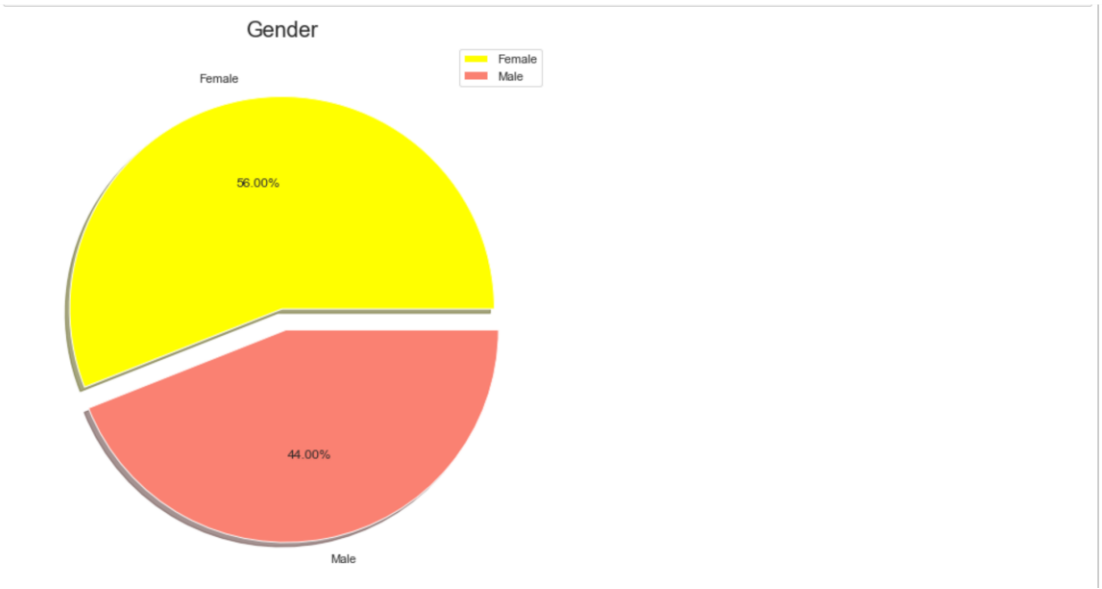
DATA VISUALISATION

- To explore and understand the dataset, we generate the various plots. We begin by plotting the gender distribution.

Gender Distribution

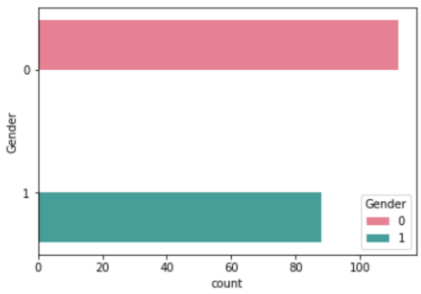
We will check the distribution of male/female customers in our dataset by creating a pie chart. Certainly we would like to know if the female shoppers outnumber the males? We do this plot using the following code:

```
In [32]: labels = ['Female', 'Male']
         size = data['Gender'].value_counts()
         colors = ['yellow', 'salmon']
         explode = [0, 0.1]

         plt.rcParams['figure.figsize'] = (9, 9)
         plt.pie(size, colors = colors, explode = explode, labels = labels, shadow = True, autopct = '%.2f%%')
         plt.title('Gender', fontsize = 20)
         plt.axis('off')
         plt.legend()
         plt.show()
```



Total Number of male and female:

```
In [37]: #Count and plot gender
         sns.countplot(y = 'Gender', data = data, palette="husl", hue = "Gender")
         data["Gender"].value_counts()

Out[10]: 0    112
         1     88
         Name: Gender, dtype: int64
```

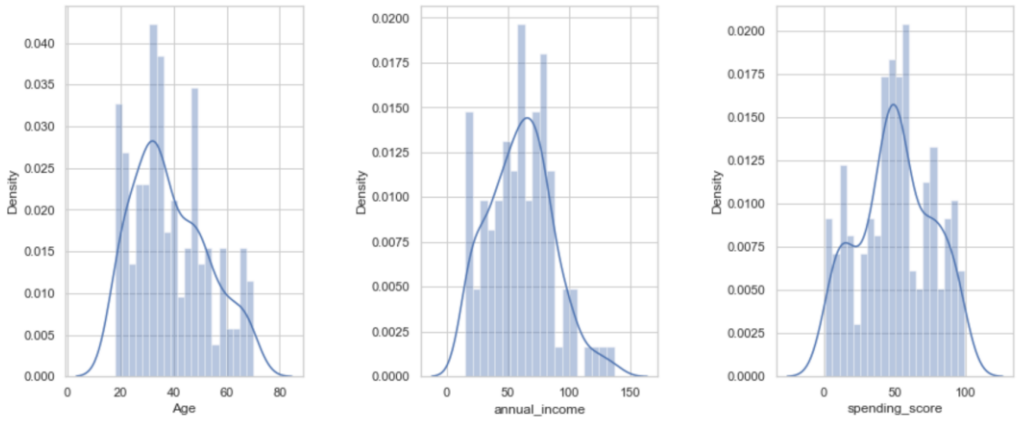


Distribution of Age ,Income,Spending score:

This is also a chart to better explain the Distribution of Each Income level, Interesting 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 .It can be seen that the Ages from 27 to 40 are very much frequently coming to malls and purchasing.People of Age 55, 56, 69, 64 are very less frequent in the Malls. People at Age 32 are the Most Frequent Visitors in the Mall.

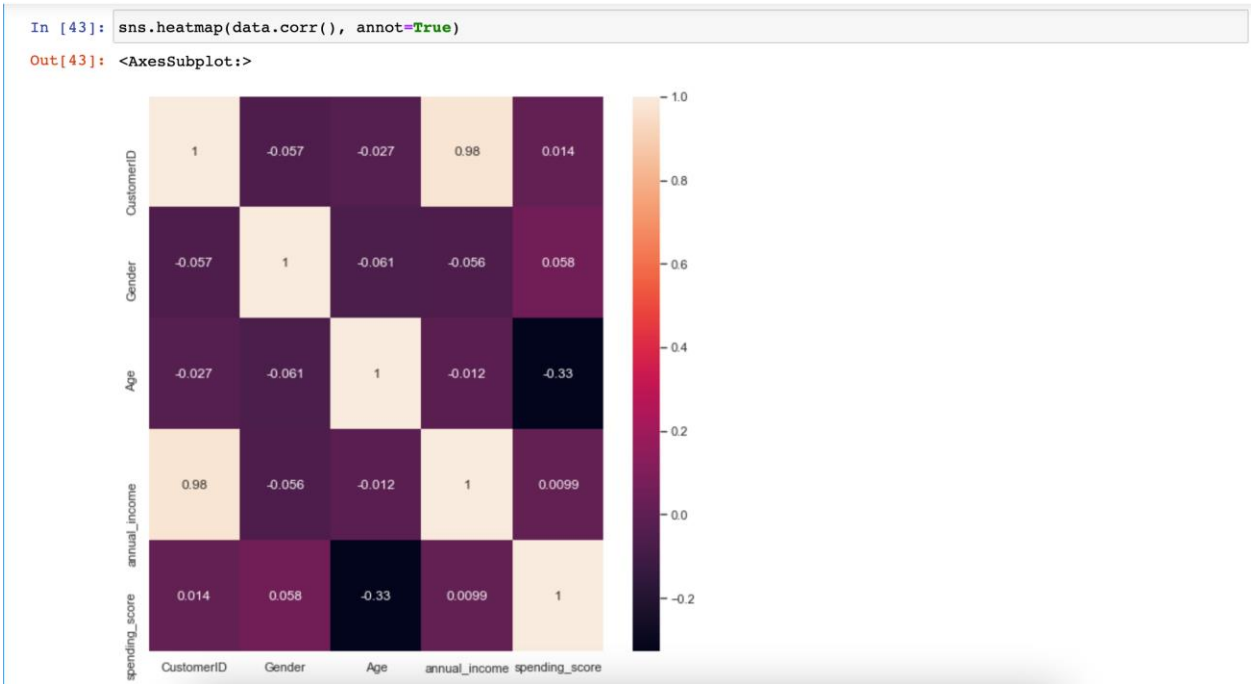
```
In [41]: plt.figure(1 , figsize = (15 , 6))
         feature_list = ['Age','annual_income', "spending_score"]
         feature_listt = ['Age','annual_income', "spending_score"]
         pos = 1
         for i in feature_list:
             plt.subplot(1 , 3 , pos)
             plt.subplots_adjust(hspace = 0.5 , wspace = 0.5)
             sns.distplot(data[i], bins=20, kde = True)
             pos = pos + 1
         plt.show()
```

In [30]:
```
plt.subplot(1, 3, 1)
sns.set(style = 'whitegrid')
sns.distplot(data['annual_income'])
plt.title('Distribution of Annual Income', fontsize = 20)
plt.xlabel('Range of Annual Income')
plt.ylabel('Count')
plt.show()


plt.subplot(1, 3, 2)
sns.set(style = 'whitegrid')
sns.distplot(data['Age'], color = 'red')
plt.title('Distribution of Age', fontsize = 20)
plt.xlabel('Range of Age')
plt.ylabel('Count')
plt.show()

plt.subplot(1, 3, 3)
sns.set(style = 'whitegrid')
sns.distplot(data['spending_score'],color='green')
plt.title('Distribution of Spending Score', fontsize = 20)
plt.xlabel('Range of Spending Score ')
plt.ylabel('Count')
plt.show()
```



Heat Map for the given data

The Above Graph for Showing the correlation between the different attributes of the Mall Customer Segmentation Dataset, This Heat map reflects the most correlated features with Orange Colour and least correlated features with yellow colour.



Variable Correlations

In this pairplot graphic we can clearly observe the relation between the different variables we have in our dataset and we will explain this relation better in the following section. Having said that it's interesting to take a look at the diagonal of this pairplot where we can see the same histograms we've studied before, but this time making a differentiation between women and man, for each value of each variable we can now see the proportion between men and women.

Age and Gender

We know the population at large is female skewed, and that the most represented age group is 30-39, but what is the distribution of gender by age.

Income and Spending Score

It is difficult to identify a clear relationship between income and spending score, however, clusters do appear to form within the data. It is difficult to interpret the relationship btween gender and spending score with this plot, so more testing should be done.

CLUSTERING / DATA MODELLING

Standardising data is a good practice when clustering, as the range of values within each feature will influence how the cluster is formed, which is not usually desirable. K-Means clustering uses Euclidean distance to measure the similarity between objects, so if a feature has a range much larger than another feature, it will dominate the other features in the clustering process.

The model was initiated on the standardised data to cluster based on age, income, and spending score.

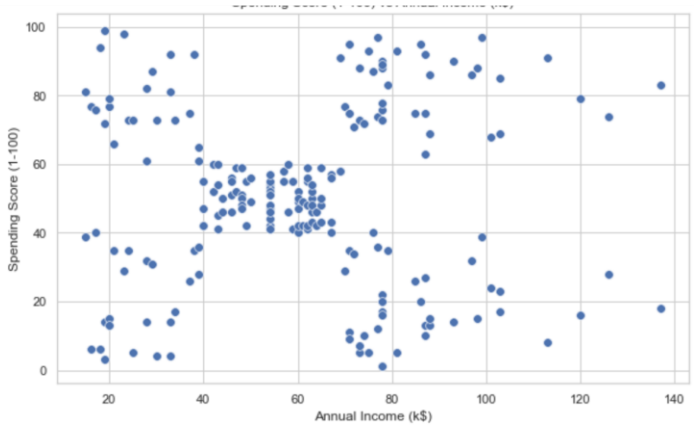
The relationship between spending score and income for both men and women. From the output, you can see that the customers having an annual income between 50 to 60 have reasonable spending habits, while some customers having either low or high income are better spenders.

```
In [51]: df1=data[["CustomerID","Gender","Age","annual_income","spending_score"]]
         X=df1[["annual_income","spending_score"]]
         X.head()
```

Out[51]:

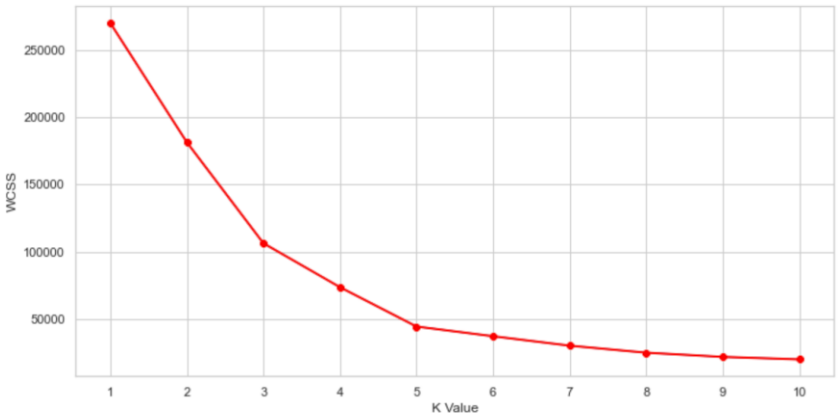
	annual_income	spending_score
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

```
In [53]: #Scatterplot of the input data
         plt.figure(figsize=(10,6))
         sns.scatterplot(x = 'annual_income',y = 'spending_score',  data = X  ,s = 60 )
         plt.xlabel('Annual Income (k$)')
         plt.ylabel('Spending Score (1-100)')
         plt.title('Spending Score (1-100) vs Annual Income (k$)')
         plt.show()
```



Now we calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k. Next, we choose the k for which WSS first starts to diminish. This value of K gives us the best number of clusters to make from the raw data.

In [57]:
```
#Importing KMeans from sklearn
from sklearn.cluster import KMeans
wcss=[]
for i in range(1,11):
    km=KMeans(n_clusters=i)
    km.fit(X)
    wcss.append(km.inertia_)
#The elbow curve
plt.figure(figsize=(12,6))
plt.plot(range(1,11),wcss)
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show()
```



This is known as the elbow graph, the x-axis being the number of clusters, the number of clusters is taken at the elbow joint point. This point is the point where making clusters is most relevant as here the value of WCSS suddenly stops decreasing. Here in the graph, after 5 the drop is minimal, so we take 5 to be the number of clusters.

In [59]:
```
#Taking 5 clusters
km1=KMeans(n_clusters=5)
#Fitting the input data
km1.fit(X)
#predicting the labels of the input data
y=km1.predict(X)
#adding the labels to a column named label
df1["label"] = y
#The new dataframe with the clustering done
df1.head()
```

Out[59]:

	CustomerID	Gender	Age	annual_income	spending_score	label
0	1	0	19	15	39	3
1	2	0	21	15	81	0
2	3	1	20	16	6	3
3	4	1	23	16	77	0
4	5	1	31	17	40	3

```
In [60]: #Scatterplot of the clusters
plt.figure(figsize=(10,6))
sns.scatterplot(x = 'annual_income',y = 'spending_score',hue="label",
                palette=['green','orange','brown','dodgerblue','red'], legend='full',data = df1  ,s = 60 )
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.title('Spending Score (1-100) vs Annual Income (k$)')
plt.show()
```

5 different clusters have been formed from the data. The red cluster is the customers with the least income and least spending score, similarly, the blue cluster is the customers with the most income and most spending score.

3-D Visualisation of Clusters

3.Segmentation using Age , Annual Income and Spending Score

The above plot displays firstly what a K-Means algorithm would yield using five clusters. We only fed three features to train our cluster model. This gives us enough data to put these features on a 3-D scale. Here, x-axis represents , y- axis represents and z-axis represents in the above graph.

After plotting the results got by K-Means on this 3-D graphic, it’s our job now to identify and describe the five clusters that have been created:

Yellow Cluster - The yellow cluster groups young people with moderate to low annual income who actually spend a lot

Purple Cluster - The purple cluster groups reasonably young people with pretty decent salaries who spend a lot

Pink Cluster - The pink cluster basically groups people of all ages whose salary isn't pretty high and their spending score is moderate

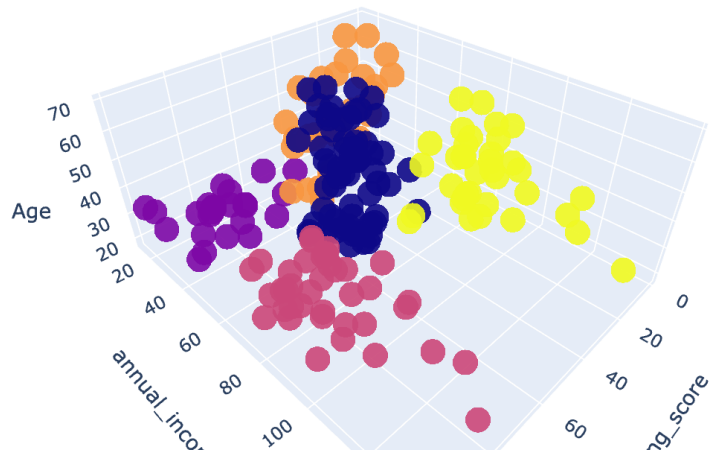
Orange Cluster - The orange cluster groups people who actually have pretty good salaries and barely spend money, their age usually lays between thirty and sixty years

Blue Cluster - The blue cluster groups whose salary is pretty low and spend a little money in stores, they are people of all ages.

```
In [78]: means_k = KMeans(n_clusters=5, random_state=0)
         means_k.fit(data)
         labels = means_k.labels_
         centroids = means_k.cluster_centers_
```


```
In [78]: means_k = KMeans(n_clusters=5, random_state=0)
         means_k.fit(data)
         labels = means_k.labels_
         centroids = means_k.cluster_centers_
```


```
In [79]: import plotly as py
         import plotly.graph_objs as go
         from sklearn.cluster import KMeans
         import warnings
         trace1 = go.Scatter3d(
            x= data['spending_score'],
            y= data['annual_income'],
            z= data['Age'],
            mode='markers',
             marker=dict(
                color = labels,
                size= 10,
                line=dict(
                    color= labels,
                ),
                opacity = 0.9
             )
         )
         layout = go.Layout(
            title= 'Clusters',
            scene = dict(
                    xaxis = dict(title  = 'spending_score'),
                    yaxis = dict(title  = 'annual_income'),
                    zaxis = dict(title  = 'Age')
                )
         )
         fig = go.Figure(data=trace1, layout=layout)
         py.offline.iplot(fig)
```



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.

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 isn't much difference between the spending score of women and men, which leads us to think that our behaviour when it comes to shopping is pretty similar.
- Observing the clustering graphic, it can be clearly observed that the ones who spend more money in malls are young people. That is to say they are the main target when it comes to marketing, so doing deeper studies about what they are interested in may lead to higher profits.
- Although younglings 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.