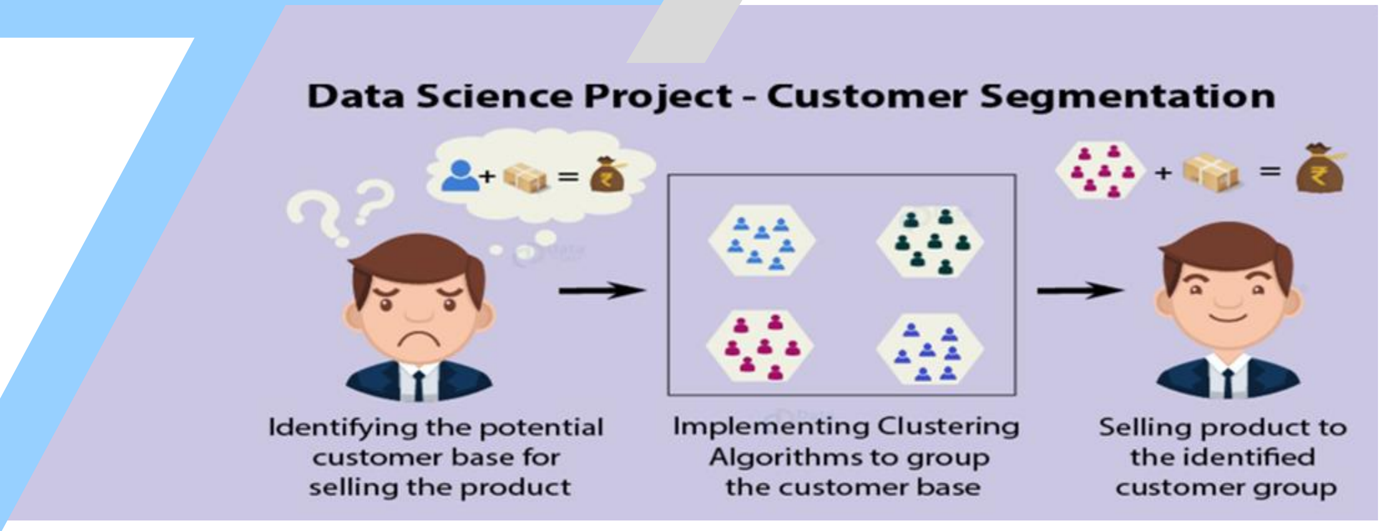
Exposys Data Labs

Internship Project


Data Science Project - Customer Segmentation
Identifying the potential customer base for selling the product
Implementing Clustering Algorithms to group the customer base
Selling product to the identified customer group

Submitted By:

Twinkle Paul

Data Lab Intern

Abstract

Customer segmentation is the practice of dividing a customer base into groups of individuals, such that each group will have distinct characteristics and need that will remain constant for members within the group. In this post - digital era, customer segmentation using data analytics and machine learning is of prime importance. Companies such as Netflix, Amazon are already leading their business towards profit by exploiting this benefit of data science to cater to the needs of the customer.

This project performs demographic and behavioural customer segmentation on a dataset of 200 customers with 5 attributes. The segmentation is carried out using the K – means clustering algorithm. Several other algorithms like Elbow Method, Silhouette Method and Principal Component Analysis have also been used, in addition to the K – means, to ensure an efficient outcome.

The project generates 6 customer segments, each with their own marketing need. Based on the analysis provided in this report, the mall may choose to cater to the needs of customers with high budget or low budget, or may even identify the group of prime customers who generate the most revenue. The report generated in this project has been generalized to specify the characteristics of all 6 segments. The mall may use the analysis to target any or all of the segments to increase its share of profits.

CONTENTS

Customer Segmentation

Introduction:

Customer Segmentation – What Does It Mean?

Customer segmentation (also known as market segmentation) is the division of potential customers in a given market into discrete groups. The division is based on customers having similar enough:

- Needs, i.e., so that a single whole product can satisfy them.
- Buying characteristics, i.e., responses to messaging, marketing channels, and sales channels, that a single go-to-market approach can be used to sell to them competitively and economically.

Who Needs It and Why?

Not all customers are the same. Each person has unique characteristics and requirements that may not be found in any other customer. Segmentation gives a company a greater ability to better satisfy the needs of its customers.

According to a 2017 Mailchimp survey:

- Segmented campaigns have a 14 percent higher open rate on average than non-segmented campaigns
- Campaigns segmented on customer interest have a 74 percent higher click rate on average than non-segmented campaigns.

Thus, if the marketing campaigns are targeted to very specific subsets of customers, companies should be able to obtain a better response rate when compared to a broad marketing campaign that advertises to the masses.

Hence, companies that want to gain an advantage over their competitors need to understand their customers and their unique requirements. By servicing their customers at a higher level than their customers, businesses are able to maintain a competitive advantage and target new customers.

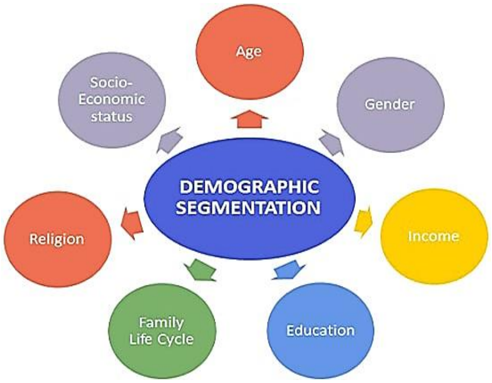
Existing Methods

Companies must determine what makes up a customer segment to identify the m. There are several criteria that can be used such as accessibility, homogenous, differentiable and measurable. Good market segmentation will result in a segment where customers are as similar as possible within the segment, and as different as possible between segments.

While there are dozens of subcategories and traits that can be used to identify different markets, there are only five types of market segmentation viz*. Demographic, Geographic, Psychographic, Behavioural* and *Industry-Based* Segmentations. The dozens of separate subcategories are placed under each of these five types. The purpose of breaking the market up into five types of segments is that businesses can more accurately achieve similarity in each of the segment.

1. Demographic Segmentation

The demographic market segmentation is focused entirely on who the customer is. In a B2B company, the traits that would likely be included in this segment extend to industry type, company size, time in position, and role within the company. On the other hand, a B2C company would include such demographic traits as age, education, gender, occupation, family status, and income.



Demographic Segmentation is a very common segmentation type that's used within market research to determine what a company's main target audience is. This information is also easy to obtain as they can be pulled from any trusted published data repository (Ex: Census).

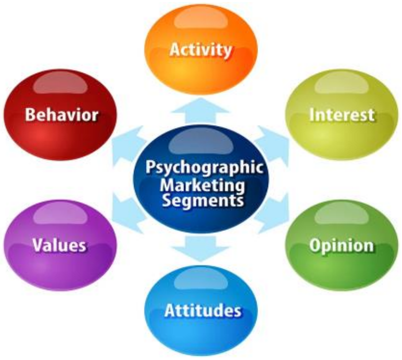
2. Geographic Segmentation

The geographic market segmentation allows to effectively split the entire audience based on their location. The location of the customers plays a part in their overall purchase decision as most customers are influenced at least partly based on where they live. The core traits and segments that can be used with the geographic segmentation include region, continent, country, city, and district.

This is an exceedingly popular type of segmentation. This form of segmentation is considered to be ideal for international companies. Customers who live in different countries will have different wants and needs, which can be precisely targeted in a marketing campaign. It's also highly beneficial for small businesses with a limited budget.

3. Psychographic Segmentation

The psychographic market segmentation is aimed at separating the audience based on their personalities. The different traits within this segmentation include lifestyle, attitudes, interests, and values.



However, extensive research will be necessary with this form of segmentation since identifying demographics based on personality is relatively subjective. If the company belongs to an industry, whose main audience values quality and energy-efficiency above all else, then the marketing platform can be altered to account for these core values.

4. Behavioural Segmentation

The behavioral market segmentation divides the whole audience based on the previous behavior that they've exhibited with the brand. Some of the main traits within this segmentation type include product knowledge, purchase patterns, previous purchases, awareness of your business, and product rating.

5. Industry - Based Segmentation

Unlike retail consumers, industrial consumers can be segmented on characteristics such as location, company type and buying characteristics, etc. When segmenting industrial customers, the location of a customer can be used to define a segment. This may be important for shipping and deliveries. Customers within a certain geographical region may have similar requirements.

Customers can be segmented by the type of company. For example, segments can be created based on company size, type of industry or purchasing criteria. The buying characteristics of customers can also define a segment. Characteristics such as purchasing volume or purchasing history.

Proposed Method

Strategy Chosen: What & Why?

Proper customer segmentation requires proper segmentation strategy. When we are considering small-scale B2C (Business to Consumer) industries (like, the shopping mall) with not much concern over the geographical spread, we can eliminate the choices of Industry-Based and Geographic segmentation method.

On the other hand, combining ***Demographic*** and the ***Behavioural*** market Segmentation has its own benefits in understanding the needs, requirements and buying habits of the consumers, which are mentioned in the section below. Also, as a bonus, when these two methods are applied in a combination, they don't require the extensive research of Psychographic Segmentation, for yielding a result of near-equivalent value.

Demographic Segmentation

Demographic segmentation is the process of dividing your market into segments based on things like ethnicity, age, gender, income, family makeup, and education.

Behavioural Segmentation

Behavioural segmentation is the process of sorting and grouping customers based on the behaviours they exhibit. These behaviours include the types of products and content they consume, and the cadence of their interactions with an app, website, or business.

Advantages of the Combination of Both:

1. ***Easily Available Data:***

   It's easy to acquire the required data through some trusted published dataset repository (Ex: census data, analytics software, consumer insights, and more).

2. ***Cost – Efficiency***

   It's also considered by many businesses to be the most cost-efficient way to divide a target market.

3. ***Build long-lasting customer relationships***

   Reaching customers on a more human level with demographic-based personalized marketing creates deeper customer loyalty.

7. Improve products and services

When companies have a deeper understanding of target audience, they can put themselves in their shoes to better serve them.

8. Optimize marketing strategies

This segmentation allows companies to get more specific with their marketing strategies. It helps to clarify vision, have more direction with future advertising plans, and optimize resources, time, and budget.

9. Budget allocation

Since the company knows the spending habit of the customers, they can better allocate their efforts to target them.

10. Forecasting

Looking at each segment's patterns, the companies can identify trends and more effectively plan for the future.

Machine Learning – the Ticket to Effective Customer Segmentation

Machine Learning – Understanding its Role in Segmentation

The use of machine learning can be seen almost everywhere around us, be it Facebook recognizing you or your friends, or YouTube recommending a video or two based on the user history. Machine Learning is broadly categorized as Supervised and Unsupervised Learning.

Supervised Machine Learning is one in which we teach the machine by providing both independent and dependent variables, for example, Classifying or predicting values.

Unsupervised Machine Learning mainly deals with identifying the structure or pattern of the data. In this type of algorithms, we do not have labeled data (or the dependent variable is absent), for example, clustering data, recommendation systems, etc.

Clustering Algorithms of Unsupervised Machine Learning, are highly effective in the scenarios of Market Segmentation. They provide amazing results as one can deduce many hidden relations between different attributes or features.

Understanding Clustering Algorithms

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present.

There are several types of clustering algorithms present. However, in case of market segmentation, the most obvious and efficient choice of a clustering algorithm is the K-Means Algorithm which is a partitioning method based on the centroid model.

K – Means Clustering: The Strategic Key

K - Means is one of the simplest unsupervised learning algorithms that can solve the customer segmentation problem. It is an iterative clustering 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 (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

This algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

where,

‘$||x_i$ - $v_j||$’ is the Euclidean distance between x_i and v_j.

‘c_i’ is the number of data points in i^{th} cluster.

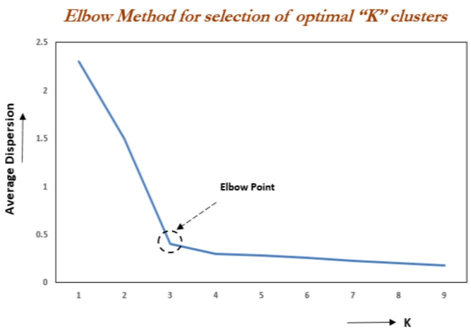
‘c’ is the number of cluster centres.

Obtaining the Mysterious Value of ‘k’

In K – Means Clustering Algorithm, the value **‘k’ s**tands for the optimal no. of clusters A fundamental step for any clustering algorithm is to determine the optimal number of clusters into which the data may be clustered, before starting the its implementation on dataset.

‘k’ has been obtained using the following algorithms:

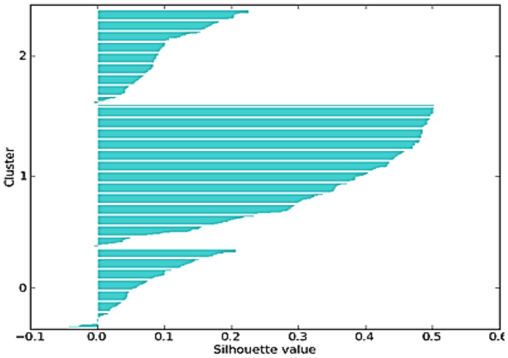
a. *Elbow Method:*



The elbow method plots the value of the cost function produced by different values of *k*. If *k* increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as *k* increases. The value of *k* at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

b. Silhouette Analysis:

S.A. is a way to measure how close each point in a cluster is to the points in its neighbouring clusters. Silhouette values lies in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighbouring cluster and very close to the cluster its assigned. Similarly, value of -1 indicates that the point is close to its neighbouring cluster than to the cluster its assigned. And, a value of 0 means it’s at the boundary of the distance between the two cluster. Hence, higher the value better is the cluster configuration.



Using the above combination of algorithms, the optimal value of ‘k’ obtained in this project was 6. Hence, K – Means Clustering Algorithm was used on the dataset of Mall Customers, to efficiently segment the customers into 6 segments as per the attributes of the dataset as shown later.

However, the obtained segments were multi – dimensional and difficult to be visualized. Hence, dimensionality reduction algorithm was applied on them to enable their visualization in a 2-D graphical plot.

Better Visualization with Dimensionality Reduction

Dimensionality Reduction Using Principal Component Analysis (PCA)

Principal Component Analysis (PCA) in conjunction with k-means is a powerful method for visualizing high dimensional data. PCA is used in exploratory data analysis and for making predictive models.

It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

The first principal component of a set of data points in a multidimensional space is the direction of a line that best fits the data, in that it minimizes the variance of the projected data or minimizes the sum of squared distances from points to the line. Each subsequent principal component ‘i’ is a direction of a line that minimizes the sum of squared distances and is orthogonal to the first ‘i-1’ principal components. Principal component analysis or PCA is the process of finding or using such components.

Using Principal Component Analysis in the project enabled to visualize the 3 – dimensional clusters on a 2 – dimensional space without the loss of any information.

Implementation

Steps Taken Towards Market Segmentation:

1. Collect Dataset & Develop an Understanding of the Variables

The dataset for this project has been collected from the following link:
https://drive.google.com/file/d/19BOhwz52NUY3dg8XErVYglctpr5sjTy4/view

	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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

The collected dataset has 200 entries and it has 5 columns in it. or features in it, listing the Customer ID, Gender, Age, Annual Income and Spending Habits.

The Variables that this project takes under consideration are:

1. Age

Age is the most basic variable of them all, albeit the most important because consumer preferences continually change with age. Almost all marketing campaigns target age-specific audiences. Besides, it categorizes the customers according to their shopping habits. Usually, middle-aged people have more spending tendency, whereas, the ripe-aged people have a saving tendency.

2. *Gender*

Men and women generally have different likes, dislikes, needs, and thought processes, making it an important feature.

3. *Annual Income*

If people can't afford your product or service, there is no point in targeting them. Income targeting lets you measure the buying power of your audience. Knowing the income range of consumers, allows to find data to support how people spend money on both the higher and lower end of the spectrum.

4. *Spending Score:*

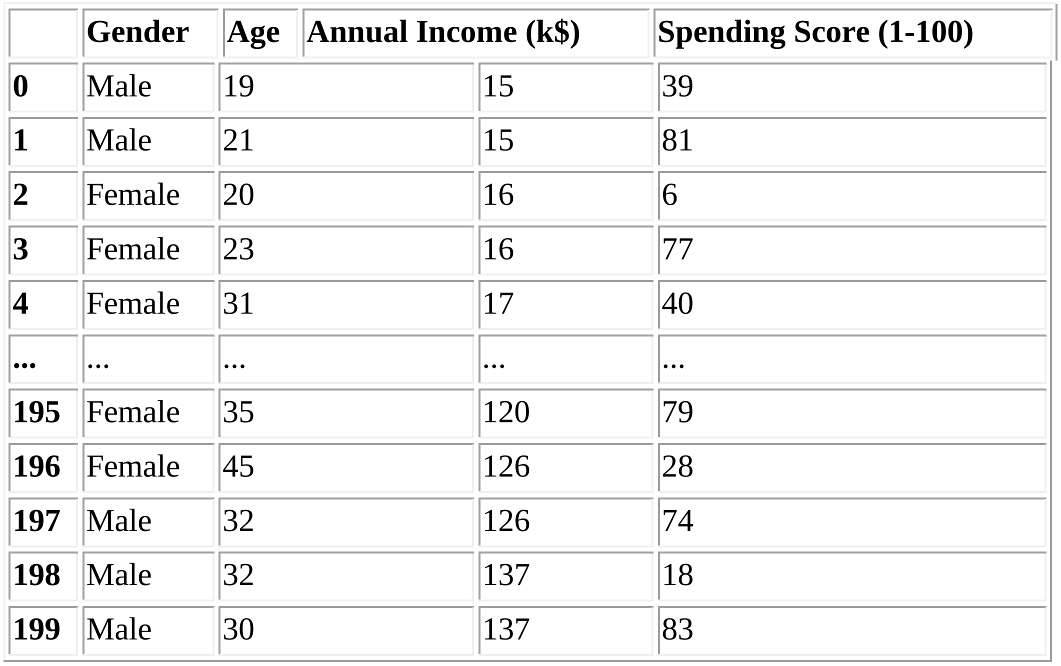
It is the score (out of 100) given to a customer by the mall authorities, based on the purchasing data and the behaviour of the customer. Based on this value, we can understand the marketing tendency of the customer and hence, is an important behavioural feature for marketing campaign.

Hence, in the next sections, the project will be trying to establish a relation between these 4 attributes to divide the customer database into segments.

2. Cleaning the Dataset

The dataset obtained previously needs to be cleaned of any unrequired data before further operation. Hence, it requires the removal of the Customer ID, which is nothing but just a unique number assigned to distinguish the customers in the database. It has no relevance in the segmentation procedure.

The cleaned dataset can be listed as follows:

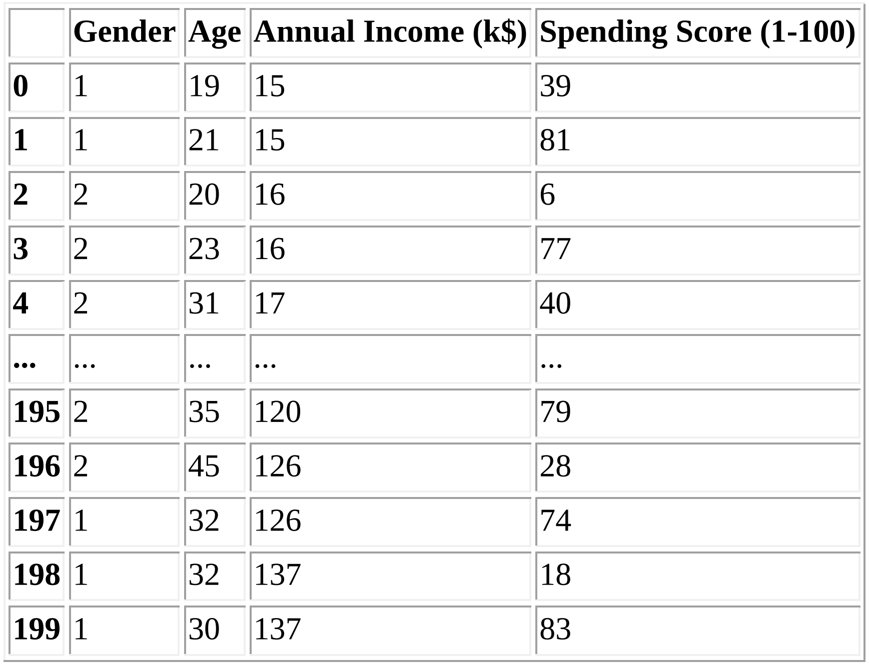


	Gender	**Age**	**Annual Income (k$)**	**Spending Score (1-100)**
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40
...	...	...	...	...
195	Female	35	120	79
196	Female	45	126	28
197	Male	32	126	74
198	Male	32	137	18
199	Male	30	137	83

3. Pre - Processing the Dataset

The cleaned dataset needs to be pre-processed before further use. In other words, the dataset thus obtained, is comprised of both numeric and non-numeric data types. In order to ensure that the data gets correctly represented, the non-numeric data type needs to be converted to numeric type.

The following table shows a dataset, where the males have been represented by 1's and females by 2's:



	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	19	15	39
1	1	21	15	81
2	2	20	16	6
3	2	23	16	77
4	2	31	17	40
...	...	...	...	...
195	2	35	120	79
196	2	45	126	28
197	1	32	126	74
198	1	32	137	18
199	1	30	137	83

4. Visualizing the Dataset

Once, the dataset has been prepared for evaluation, it needs to be evaluated, and the co-relation (if any) between the variables needs to be studied along with their individual influence on the data. This helps to analyse the segmented data in a later phase of the project.

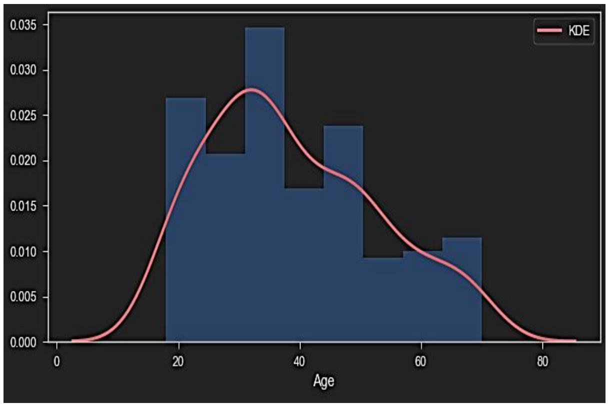
- ***Using the Probability Distribution Plot***

Studying the distribution plot gives information on the reach of each of the attributes, and how they influence the customer.

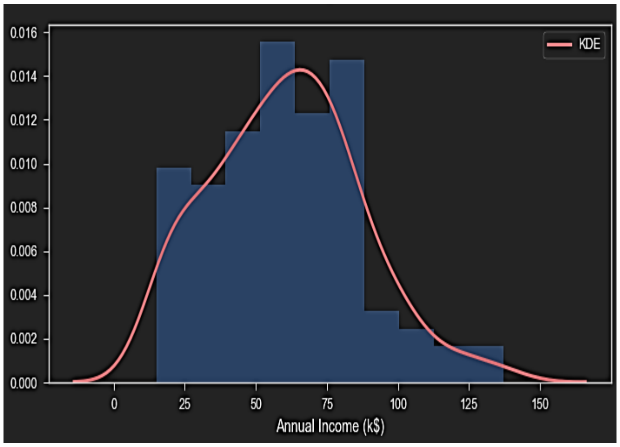
The evaluation of distribution plot led to the following assumptions:

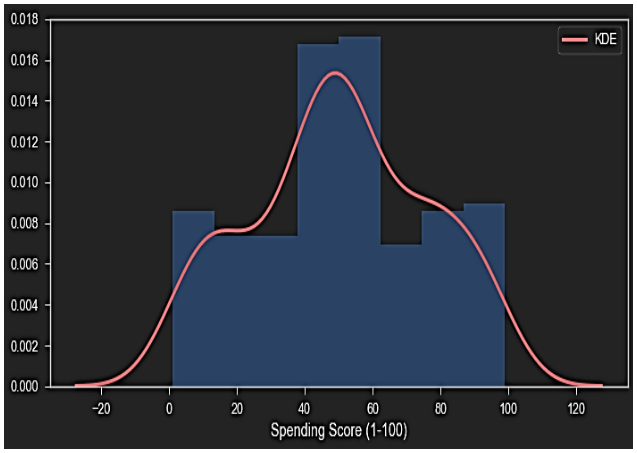
1. Most of the customers are around the age of 38.

2. The Probability Density Curve confirms that there are more young customers than old.



3. Among the middle-aged customers, there's an increased no. of customers at the age of approximately, 38 and 45.





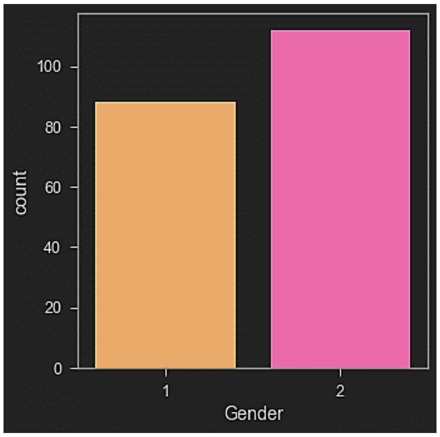
4. The Annual Income curve shows, most customers have an annual income of less than 75k dollars.

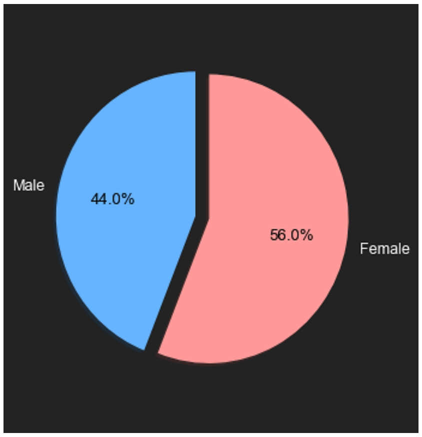
5. The mean Annual Income of the customers can be considered around 60k dollars.

6. Only a few customers have very low (below 20), or, very high (above 80) spending score.

7. Most customers have a spending score of around 40 - 60 %.

- ## *Using the Count Plot & Pie Chart to Visualize Gender Ratio*

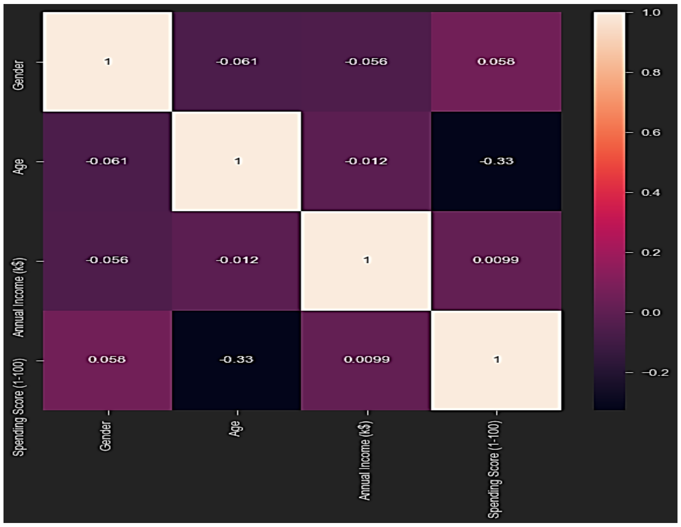




The above graphs depict that there are more female customers in the mall than male.

• Checking for Co – Relation among the Data

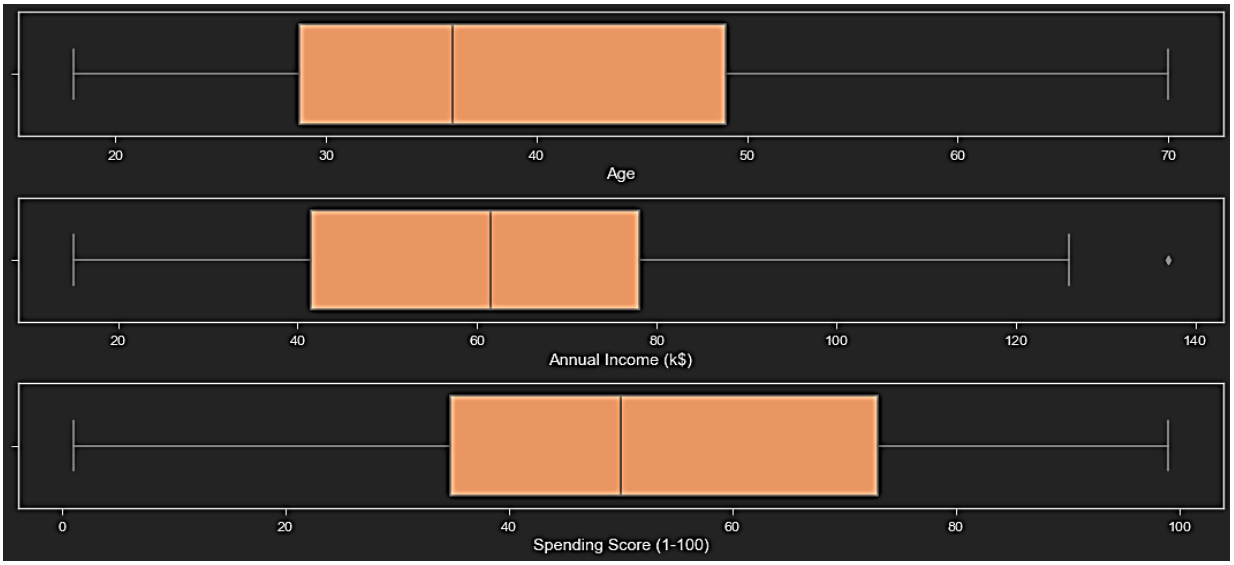
Checking for Co – relation helps to analyse if any, then how does one variable influence other, and to what degree.



However, in the dataset considered in this project the 3 demographic variables (i.e. Age, Gender, Annual Income) and the behavioural variable (i.e. Spending Habit) are all independent of each other.

• Checking for Skewness

Even though the dataset has been cleansed, it's quite possible that some erroneous data might have made its way into the dataset. Visualizing the skewness, would give us an insight into the presence of such erroneous data, if any.



The above plot ensures, that the data hasn't been skewed by any mis – represented data or error.

5. Start Processing the Dataset

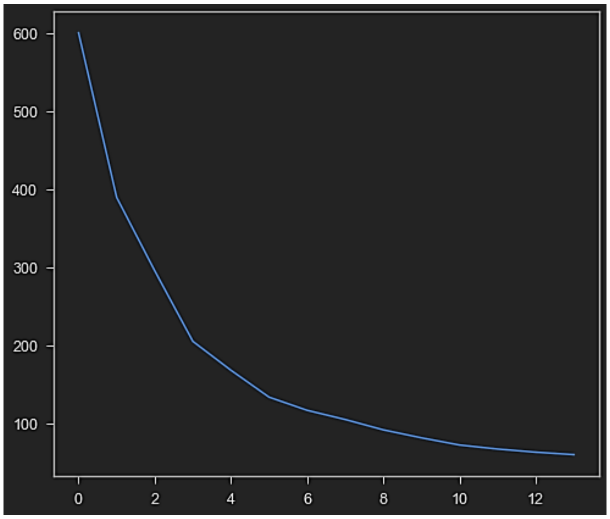
- ### *Scale the Data*

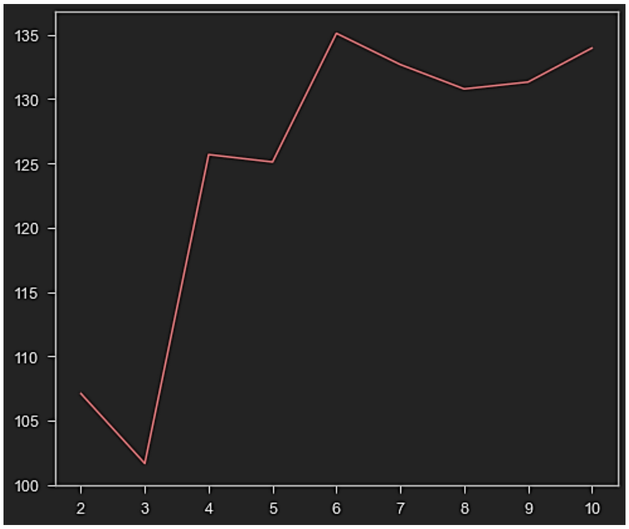
The data needs to be scaled, so that all the variables have uniform influence in the following algorithms and do not bias the outcome.

- ### *Checking for the Optimal No. of Clusters*

The scaled dataset is ready to be applied to a training algorithm. However, as mentioned in the Methodology Section, before applying k – means algorithm, it's important to determine the optimal number of clusters i.e. 'k'.

As per the methodology explained earlier, the elbow Plot and Silhouette method are applied to the dataset.





The Elbow Plot places the value of 'k' as 5, whereas, the more efficient Silhouette Analysis places the value at approximately 6.

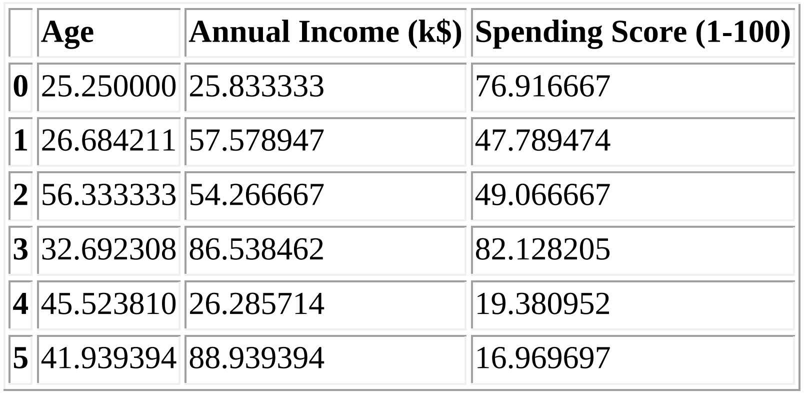
In regards to this project, the value of 'k' has been chosen as 6.

6. Applying K – Means to Segment the data

- ### *Training the Model*

The scaled dataset is used to train a clustering model using K – means algorithm. The trained dataset results in 6 cluster, each with their own centroid value per attribute.

The clusters are listed as:



	Age	Annual Income (k$)	Spending Score (1-100)
0	25.250000	25.833333	76.916667
1	26.684211	57.578947	47.789474
2	56.333333	54.266667	49.066667
3	32.692308	86.538462	82.128205
4	45.523810	26.285714	19.380952
5	41.939394	88.939394	16.969697

The above clusters can be classified as:

- **Young Moderate Spenders:** Customers belonging to this group are mostly the young individuals (mean around 25). Their annual income and spending score are close to the mean annual income (60) and the mean spending score (50) that was observed earlier with respect to the non - clustered set of data. This group needs further observations for any other remark.

- **Middle-aged Moderate Spenders:** Customers belonging to this group are mostly the people in their middle adulthood or older (above 55-year-old) individuals. Their annual income (near 60) and spending score (near 50) needs further speculation as they are close to the mean values for the entire dataset.

- **Young Spendthrifts:** Customers belonging to this group are once again the young individuals (around 25). However, their annual income is distinctly low (below 30k) and spending score is distinctly high (above 70k). These group of customers can be designated as frequent buyers, but they can't be expected to shop much in the high price range.

- **Prime Customers:** Customers belonging to this group belong to their middle-age or early adulthood (30 - 40 years old). Both their annual income (above 80k) and spending score (above 80) are distinctly high. These are the prime customers of the mall. They are the main revenue generators.

- **Low Budget Methodical Customers:** Customers belonging to this group belong to their middle-age (around 45 years old). Both their annual income (below 30k) and spending score (below 40) are distinctly low. These customers are neither frequent shoppers, nor can they afford to pay much. They are definitely, the thoughtful buyers, who use their experience and better judgement for every single penny spent.

- **Middle-aged, High Earner, Low Spender:** Customers belonging to this group also belong to their early middle-age (around 40s). Their annual income is distinctively high (above 80k). However, the spending score is quite low (below 30). They probably prefer some other mall. These customers can be converted into the Prime Customers with proper marketing campaign.

- ***Fitting the Model***

Once the model has been trained, and the clusters have been generated, the entire dataset is applied to the trained model and each data point is associated with the corresponding cluster.

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

The above list shows the first five data points or consumers, along with their respective clusters or segments.

7. Applying PCA for Dimensionality Reduction

As the clusters have 3 variables (viz. Age, Annual Income, Spending Score) associated with them, they can’t be visualized in a 2 – dimensional space. Hence, the application of Principal Component Analysis is important to reduce the no. of dimensions without losing any relevant information.

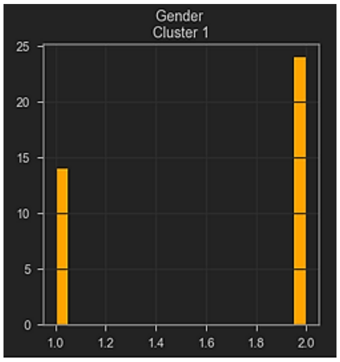
	pca1	pca2	Cluster
0	-0.615720	-1.763481	0
1	-1.665793	-1.820747	0
2	0.337862	-1.674799	4
3	-1.456573	-1.772430	0
4	-0.038465	-1.662740	4

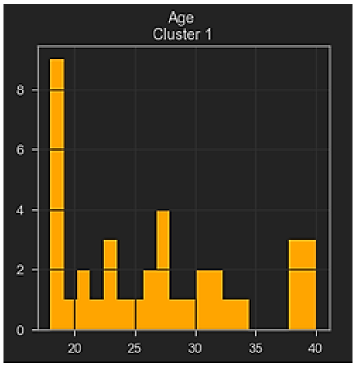
The above list shows the first five data points or consumers as a function of 2 principal components (viz. pca1 and pca2), instead of the combined function of the dataset attribute.

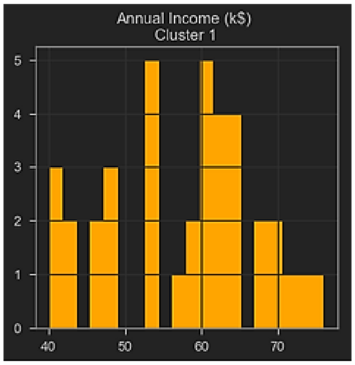
Since, the clusters associated with the datapoints in the list with variables is same as that with the principal components, its verified that the information is preserved in dimensionality reduction method.

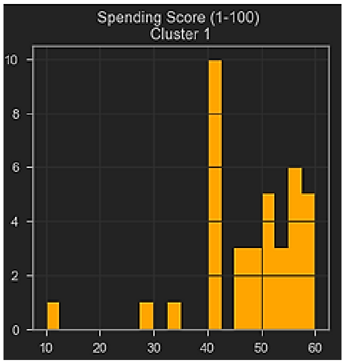
Analysing Customer Segments vs Variables

1. Young Moderate Spenders:





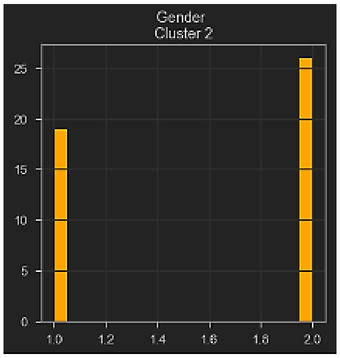


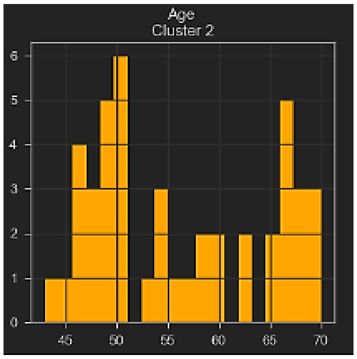


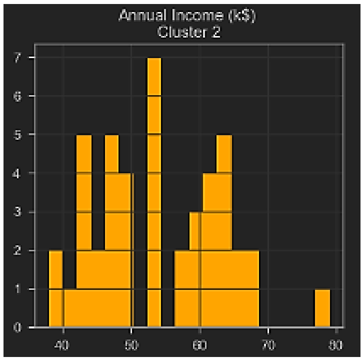
- These customers are mostly the young individuals (around 20 - 40), and includes a lot of teenagers.
- Though there's a significant group of male customers, the no. of female shoppers is more as expected (approx. 12 units more than male).
- Their annual income, is quite diverse, however, the graph indicates they range from 40k to beyond 70k, with majority of them having an income within 55-70k.
- Their spending score, however, is quite low, with majority customers lying in the zone of 40 - 60.

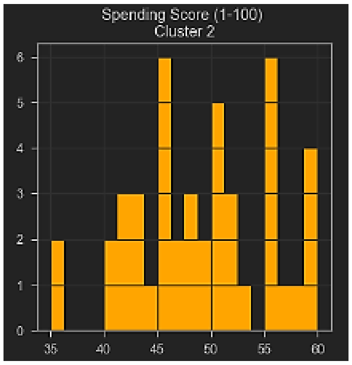
Clearly, these customers can be made to shop more. Keeping in mind their low age, they can be targeted towards impulse buying of products needed by the young generation.

2. Middle-aged Moderate Spenders:





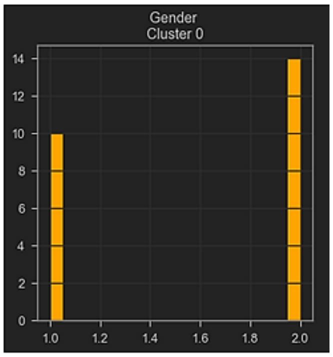


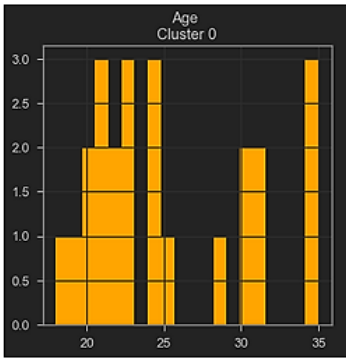


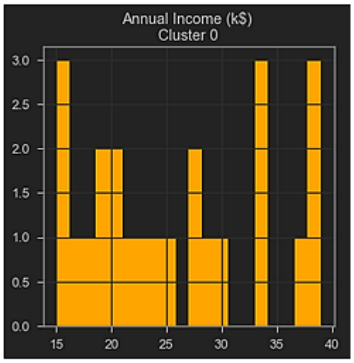
- These customers are mostly around mid-40s to early 70s.
- Though there's a significant group of male customers, the no. of female shoppers is more as expected (approx. 9 units more than male).
- Their annual income lies in the range of approximately 40-70k with a spectacular high frequency in the region of 55k and some exceptions. They can be safely put in the category of individuals with stable income.
- They have a mediocre spending score, with notable highs in the range of 45 - 60.

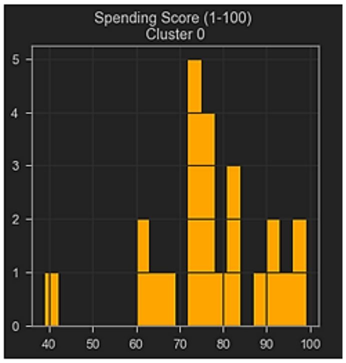
These group of customers are careful spenders, and are probably people with family.

3. Young Spendthrifts:





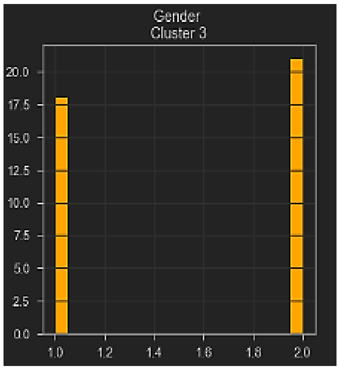


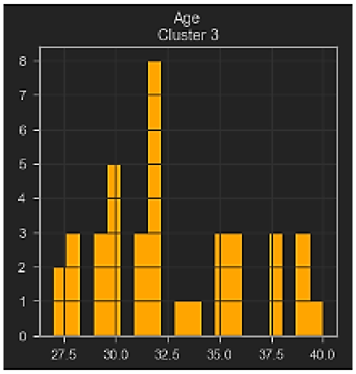


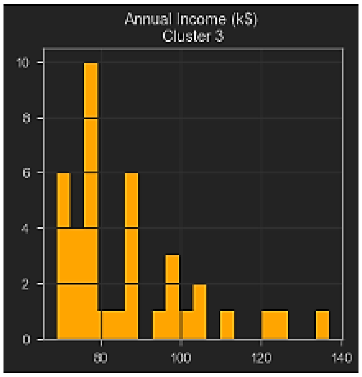
- These customers are mostly in their early adulthood (around 20 -35). A significant amount of them are teenagers.
- There isn't a lot of difference between the number of male and female customers, though the female customers are definitely, a bit more (about 2.5 units).
- Their annual income is below the average annual income of the dataset (below 40k).
- Their spending score, the other hand, is quite high (mostly, above 50).

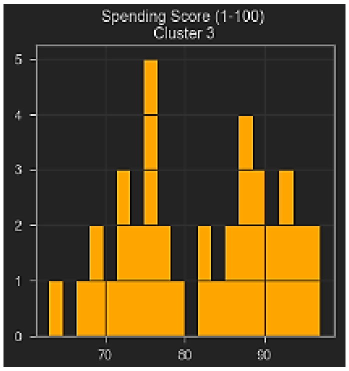
Hence, it can be presumed that these group of people are probably college goers, or, people with new jobs, or promotions. Their annual income vs spending score ratio, definitely, puts them in the list of impulse buyers. The mall must target this group to bring in more customers. Also, they are an asset to the mall who can be expected to buy costlier products in the future.

4. Prime Customers:





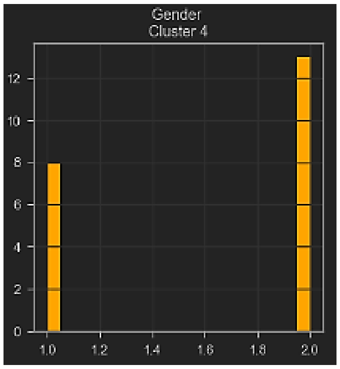


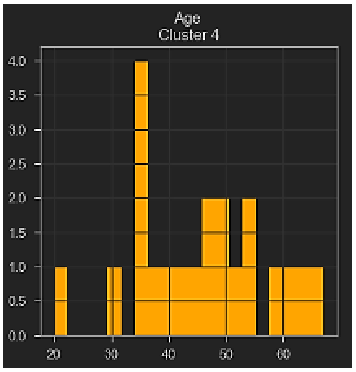


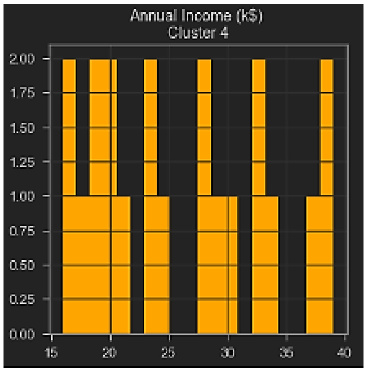
- These customers are in mid-age i.e. late twenties to early forties.
- They are almost evenly distributed gender-wise (only, a negligible 2.5 units more male than female).
- Their annual income is significantly higher (mostly above 70k).
- Their spending scores are also, distinctly high (above 60).

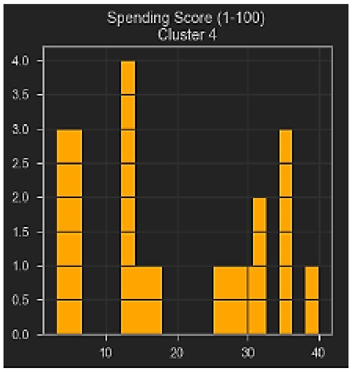
Hence, these are the prime customers of the mall that bring in most of the revenue.

5. Low Budget Methodical Customers:





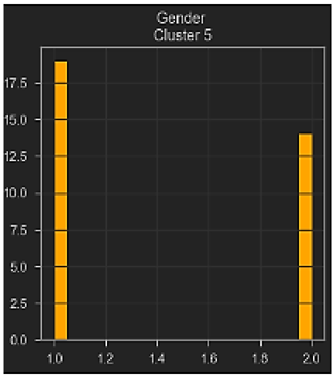


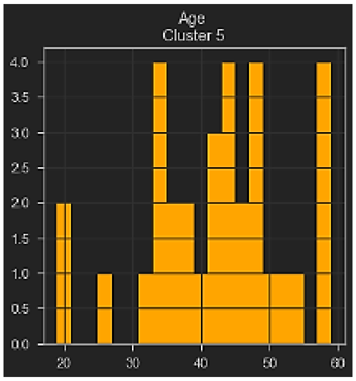


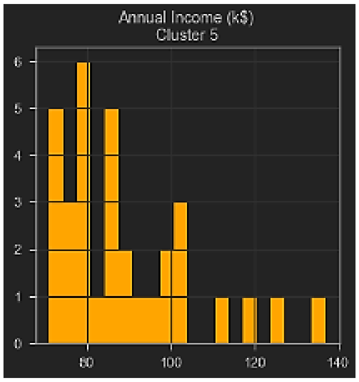
- These customers belong to diverse age, though, a notable high can be noted in the region of mid-thirties.
- There are more females than male customers (about 5 units).
- These customers have low annual income (below 40k)
- Their spending scores are also low spending score (below 40).

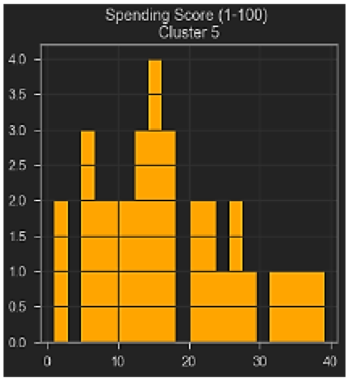
Hence, these are the customers with low - budget and methodical purchasing behaviour.

6. Middle-aged, High Earner, Low Spenders:





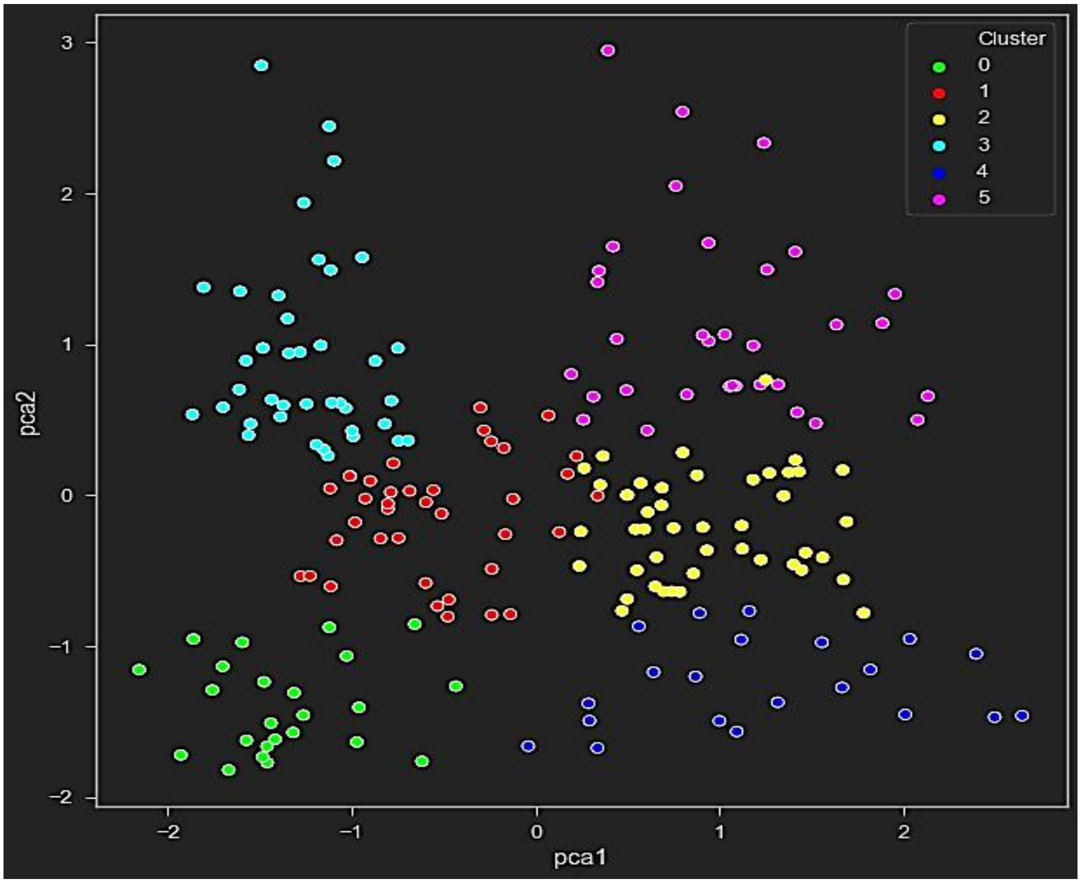




- These customers are in the diverse age-group. However, most of them are in their late thirties to late fifties.
- This is the only group with more male (3.5 units more) than female customers.
- Their annual income is the highest (above 70k) among all other groups.
- Their spending score is quite low (below 40).

They probably prefer some other mall or are new customers. These customers are definitely the worthy target of a marketing campaign.

Analysing Customer Segments in 2D - space



Using the graphical plot above, we can further analyse the identified clusters *(listed features may not be in order with the data in graph)*:

- **Young Moderate Spenders:** The customers with both pca1 and pca2 value in range of -1 to +1. These were the young customers with moderate buying tendency, and the target for impulse buying of products needed by the young generation.

- **Middle-aged Moderate Spenders:** The customers with positive pca1 and pca2 value in range of -1 to +1. These were the middle-aged customers with stable or high income, and a mediocre spending score. They are the cautious buyers.

- **Young Spendthrifts:** The customers with negative pca1 and negative pca2 value. These were the customers in their early adulthood, with below average annual income and high spending score. They are the impulse buyers.

- **Prime Customers:** The customers with negative pca1 and positive pca2 value. These were the middle-aged customers with significantly high annual income as well as spending scores. These were the group of prime customers of the mall who bring in most of the revenue.

- **Low Budget Methodical Customers:** The customers with positive pca1 and negative pca2 value. These were the diverse-aged with low annual income as well as low spending score. They are careful spenders.

- **Middle-aged, High Earner, Low Spender:** The customers with positive pca1 and positive pca2 value. These were also the diverse-aged individuals. However, it was the only male majority group with the highest annual income among all other groups but low spending scores. They aren't frequent buyers, but are definitely, a good target for the marketing campaign.

Thus, using K-Means Clustering on a dataset containing demographic and behavioural variables, yields us 6 different segments of customers, each with their own characteristics, and with their own distinct marketing needs.

The Mall can use the trained model generated in this project, to allocate unknown customers or data points to the relevant segment. This would enable to understand the customer’s specific needs and target them with a marketing technique that has been tailored to that individual. This would in turn, increase the mall’s share of profits.

Thank
you