Exposys Data Labs

Internship Report

On

CUSTOMER SEGMENTATION

Submitted in partial fulfilment of
the requirements of the
internship

in

DATA SCIENCE

by

ARUN GOVIND

DECLARATION

I, Arun Govind, declare that this internship report on “Customer Segmentation” prepared for Exposys Labs, Bangalore represents my ideas in my own words, adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented any idea or data in my submission. I understand that disciplinary action by the organization may be taken for any violation in the rules and any action may be taken if I falsified any data or rules being broken.

ABSTRACT

The emergence of many business competitors has engendered severe rivalries among competing businesses in gaining new customers and retaining old ones. Due to the preceding, the need for exceptional customer services becomes pertinent, notwithstanding the size of the business. Furthermore, the ability of any business to understand each of its customers' needs will earn it greater leverage in providing targeted customer services and developing customised marketing programs for the customers. This understanding can be possible through systematic customer segmentation. Each segment comprises customers who share similar market characteristics. The ideas of big data and machine learning have fuelled a terrific adoption of an automated approach to customer segmentation in preference to traditional market analyses that are often inefficient especially when the number of customers is too large. In this report, the K-Means clustering algorithm is applied for this purpose.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

In the present generation, it is important to comprehend client behaviour and sort clients based on their demography and purchasing behaviour. In the Retail sector, the various chains of hypermarkets and malls generate an exceptionally large amount of customer data. This data is generated daily across these stores and outlet malls. This extensive database of customers transactions needs to be analysed for designing profitable strategies.

All customers have different kinds of needs based on their age and income. With the increase in customer base and transactions, it is hard to understand the requirements of the market. To attain a huge success over their market, they must boost their strategy to identify, understand and target the profitable or valuable customers from other non-profitable customers. Identifying potential customers can improve the marketing campaign, which ultimately increases the sales.

Segmentation can play a better role in grouping those customers into various segments. The critical aspect of customer segmentation is that it allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies. Marketing messages sent to well-thought-out customer segments have seen 200% greater conversions than those sent to a general audience. If you can create well-defined customer segments based on common attributes, you can get better results from marketing. For example, segmentation based on the age of a customer can help you send the right messages that matter the most to them. If you create customer segments based on the social channels they frequent, you can engage with them at the right place and time. Tools such as Google Analytics can help you create customer segments effectively. Effective customer segmentation can help you stay relevant and valuable to your audience and ahead of your competition as well.

EXISTING WORK

There are various methods used by marketers after collecting enough data about their customers to divide their customers and prospects into segments. Some types of customer segregation used to segregate customers used by marketers are:

1. Demographic Segmentation
2. Geographic Segmentation
3. Behavioural Segmentation
4. Psychographic Segmentation

Some advanced schemas and solutions that retailers might leverage to help in the segmentation based on demographic, geographic, behavioural and psychographic characteristics involve:

- CHAID/Decision Tree Analysis

CHAID (chi-square automatic interaction detection) or Tree Analysis starts with partitioning each object (customer) into one of two outcomes (this is the dependent variable). Examples include response to a marketing promotion (0=No, 1=Yes), purchasing a specific product or making a repeat purchase. This is the starting point for the tree. Other characteristics (the independent variables) are used to further partition customers. Each partition creates a new branch of the tree.

CHAID is a tool used to discover the relationship between variables. CHAID analysis builds a predictive model, or tree, to help determine how variables best merge to explain the outcome in the given dependent variable. CHAID creates all possible cross tabulations for each categorical predictor until the best outcome is achieved and no further splitting can be performed.

In the CHAID technique, we can visually see the relationships between the split variables and the associated related factor within the tree. The development of the decision, or classification tree, starts with identifying the target variable or dependent variable; which would be considered the root. CHAID analysis splits the target into two or more categories that are called the initial, or parent nodes, and then the nodes are split using statistical algorithms into child nodes. Unlike in regression analysis, the CHAID technique does not require the data to be normally distributed.

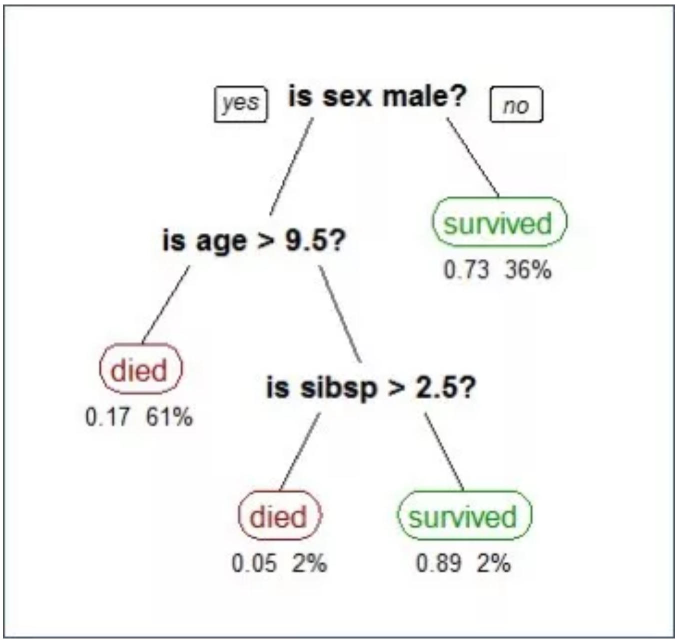


Figure 1: Tree Analysis of Titanic Survivors Example

As an example, below is a tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the nodes show the probability of survival and the percentage of observations in the branch. The tree analysis shows that your chances of survival were good if you were part of segment (1) a female or segment (2) a male older than 9.5 years with three or more siblings.

<u>- Neural Networks</u>

A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. The concept of neural networks for segmentation is derived from an imitation of neural networks in the human brain. In short, it's a sophisticated version of connect-the-dots, where the algorithm deduces certain outcomes based on multiple layers of input data.

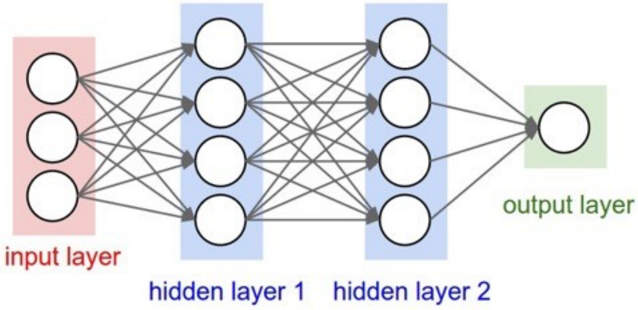


Figure 2: Neural Network example

A neural network must be trained by being presented with numerous examples of the behaviour to be understood. Much like the human brain considers a multitude of inputs to conclude or decide, the neural network algorithm connects all the available data to recognize patterns in the information that can be most associated with an expected outcome.

Some applications of neural networks include:

1. Creating buyer segments from the total population of consumers
2. Developing customized content for specific segments
3. Making predictions regarding certain customer preferences/behaviours

Neural networks have also gained widespread adoption in business applications such as forecasting and marketing research solutions, fraud detection and risk assessment.

A neural network evaluates price data and unearths opportunities for making trade decisions based on the data analysis. The networks can distinguish subtle nonlinear interdependencies and patterns other methods of technical analysis cannot.

PROPOSED WORK

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment to maximize the value of each customer to the business. In this project we will be using K-Means clustering method of unsupervised learning to visualize the gender and age distributions to analyse their annual incomes and spending scores using Python.

Clustering has proven efficient in discovering subtle but tactical patterns or relationships buried within a repository of unlabelled datasets. Clustering algorithms include K-Means algorithm, k-Nearest Neighbour algorithm, Self-Organising Map (SOM) and so on. These algorithms, without any knowledge of the dataset beforehand, can identify clusters therein by repeated comparisons of the input patterns until the stable clusters in the training examples are achieved based on the clustering criterion or criteria. Each cluster contains data points that have very close similarities but differ considerably from data points of other clusters. Clustering has got immense applications in pattern recognition, image analysis, bioinformatics and so on. In this paper, the k-Means clustering algorithm has been applied in customer segmentation.

The main idea of cluster analysis is to identify groupings of objects (customers) that are similar with respect to a collection of characteristics and are as dissimilar as possible from an adjacent grouping of objects. The number of clusters is dependent upon the type of algorithm used to identify the clusters. The characteristics can be a set of data elements that describe each customer.

Within a cluster analysis, one widely used methodology for datamining is K-Means Clustering, where K represents the number of clusters to be created. Each cluster is centred around a point called a centroid. Think of the clusters as planets in multi-dimensional space, with the number of dimensions equivalent to the number of data elements and where each object has a centre point. The centre point is made up from the average values of the data elements making up the clusters.

In the end, the K-Means solution might yield five manageable segments. A customer is assigned to a segment by calculating a distance measurement to each of the centroids. The score is commonly calculated from the Euclidean distance. The centroid with the closest distance becomes the customer's home segment.

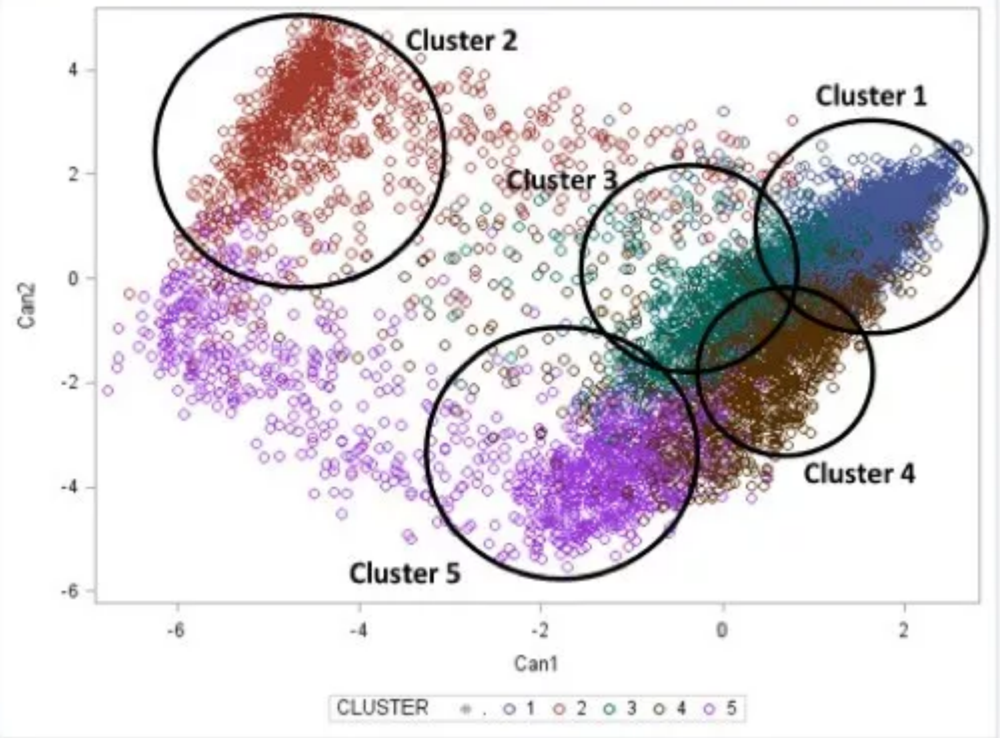


Figure 3: Proposed K-Means Clustering Analysis Diagram

METHODOLOGY

The data used in this project was given by Exposys Data Labs inform of a drive link which consisted of an excel file with mall customers dataset. The dataset consists of 5 features representing 200 selected customers. The five features include the customer ID, gender of the customers, age of the customers, annual income of a customer and their spending score respectively. In this project, four steps were adopted in realising an accurate result using K-Means algorithm.

Generic K-Means clustering algorithm:

- STEP 1: Feature Normalisation
- STEP 2: Centroids Initialization
- STEP 3: Assignment Stage
- STEP 4: Updating Stage
- STEP 5: Repeat steps 3 and 4 until the changes in positions of centroids are zero.

Step 1: Feature Normalisation

The first step in k-means is to pick the number of clusters, k. This is a data preparation stage. Feature normalisation helps to adjust all the data elements to a common scale to improve the performance of the clustering algorithm.

Step 2: Centroids Initialisation

One of the important steps in K-Means Clustering is to determine the optimal no. of clusters we need to give as an input. This can be done by iterating it through several n values and then finding the optimal n value, it gives us the number of clusters hence the total number of centroids. We find this optimal n, using the Elbow Method.

The elbow method runs k-means clustering on the dataset for a range of values for k and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned centre.

Step 3: Assignment Stage

In the assignment stage, each data point is assigned to the cluster whose centroid yields the least within cluster sum of squares compared with other clusters. That is, the square Euclidean norms of each data point from the current centroids are computed. Thereafter, the data points are assigned membership of the cluster that gives the minimum square Euclidean norm.

Step 4: Updating Stage

In this stage we update the centroid of each cluster using the data points therein. After each iteration, new centroid is computed for each cluster as the mean of all the data points present in the cluster.

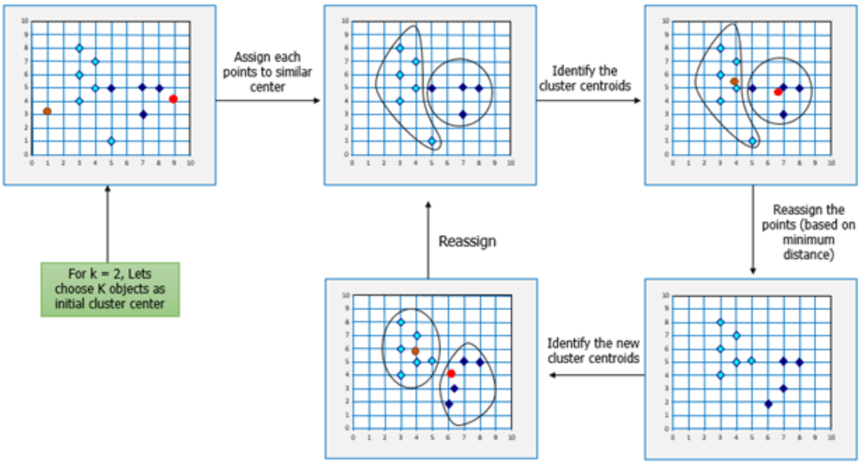

Assign each points to similar center
Identify the cluster centroids
Reassign the points (based on minimum distance)
Identify the new cluster centroids
Reassign
For k = 2, Lets choose K objects as initial cluster center

Figure 4: Example for Step by Step method for K-Means method

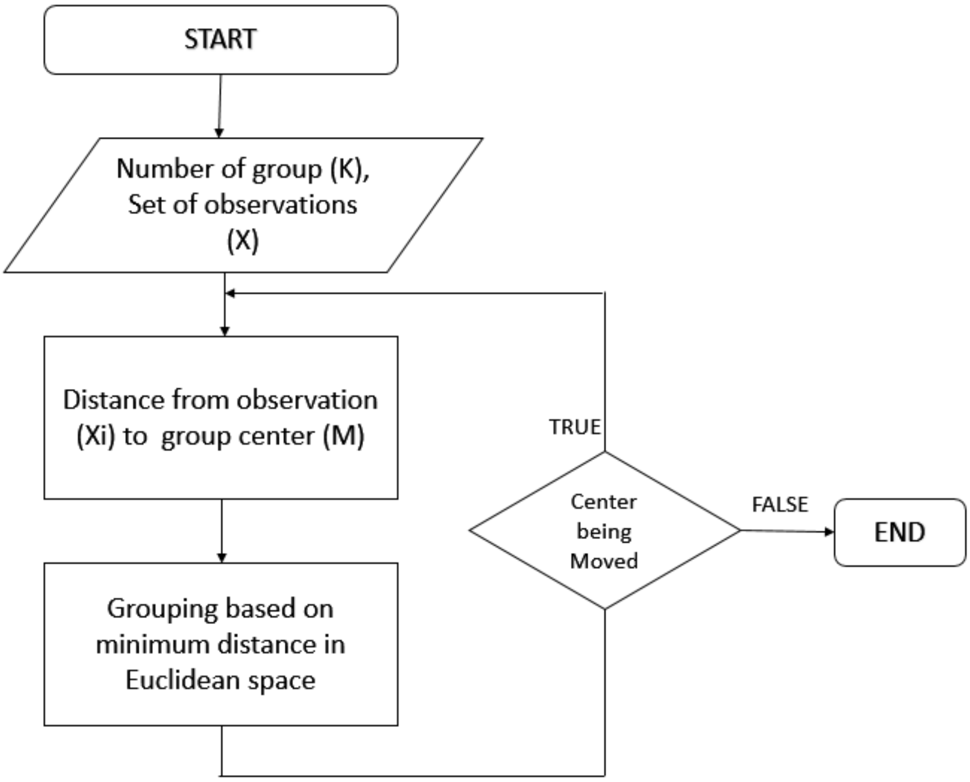

START
Number of group (K), Set of observations (X)
Distance from observation (Xi) to group center (M)
Grouping based on minimum distance in Euclidean space
TRUE
Center being Moved
FALSE
END

Figure 5: Flowchart for K-Means method of clustering

IMPLEMENTATION

We used various tools to implement the provided data and use K-Means method of clustering.

1. Python

Python is a popular programming language. It was created by Guido van Rossum and released in 1991. Python was designed for readability and has some similarities to the English language with influence from mathematics. It uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

It is used for:

- web development (server-side),
- software development,
- mathematics,
- system scripting.

Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc). Python has a simple syntax like the English language. It has syntax that allows developers to write programs with fewer lines than some other programming languages. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.

Data analysis of this project is on Python.

2. Pandas

Pandas has been one of the most popular and favourite data science tools used in Python programming language for data wrangling and analysis. Data is unavoidably messy in real world. And Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analysing data.

Here are just a few of the things that pandas does well:

- Easy handling of missing data in floating point as well as non-floating-point data
- Size mutability: columns can be inserted and deleted from Data Frame and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, Data Frame, etc. automatically align the data for you in computations
- Powerful, flexible group by functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data
- Make it easy to convert ragged, differently indexed data in other Python and NumPy data structures into Data Frame objects.

We used pandas for the data exploratory and extracting data information.

3. Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. It is used in Python scripts, web application servers, and various graphical user interface toolkits.

In our project we use Matplotlib to visualise gender distribution and implement the Elbow method to find the optimal number of clusters required for K-Means clustering.

4. Seaborn

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

Here is some of the functionality that seaborn offers:

- A dataset-oriented API for examining relationships between multiple variables
- Specialized support for using categorical variables to show observations or aggregate statistics
- Automatic estimation and plotting of linear regression models for different kinds dependent variables
- Concise control over matplotlib figure styling with several built-in themes
- Tools for choosing colour palettes that faithfully reveal patterns in your data

Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented 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.

Seaborn was used to implement distribution graphs to depict the age, annual income and spending score in terms of Regression plots, histograms and lmplots.

5. Scikit-learn

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction. It is an open source Python library that implements a range of machine learning, pre-processing, cross-validation and visualization algorithms using a unified interface. Its name stems from the notion that it is a "SciKit" (SciPy Toolkit), a separately developed and distributed third-party extension to SciPy. The original codebase was later rewritten by other developers.

Important features of scikit-learn:

- Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, etc.
- Accessible to everybody and reusable in various contexts.
- Built on the top of NumPy, SciPy, and matplotlib.

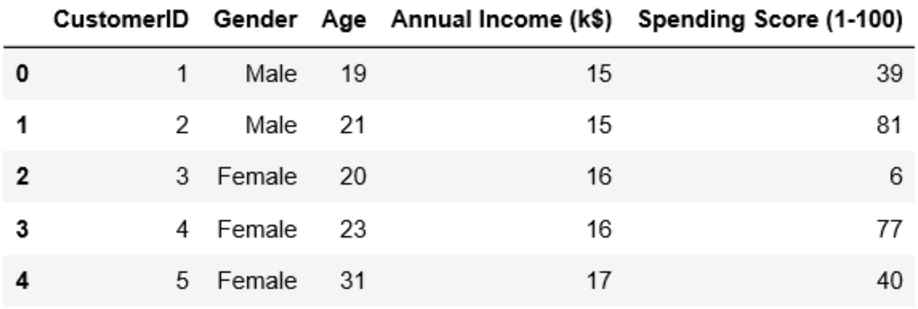
We implemented scikit-learn in our project to process our data through the means of K-Means clustering to classify the data into their respective clusters giving valuable information about their age, annual income and spending scores respectively.

INFERENCE

1. Data Information

The dataset consists of 5 features representing 200 selected customers. The five features include the customer ID, gender of the customers, age of the customers, annual income of a customer and their spending score respectively.

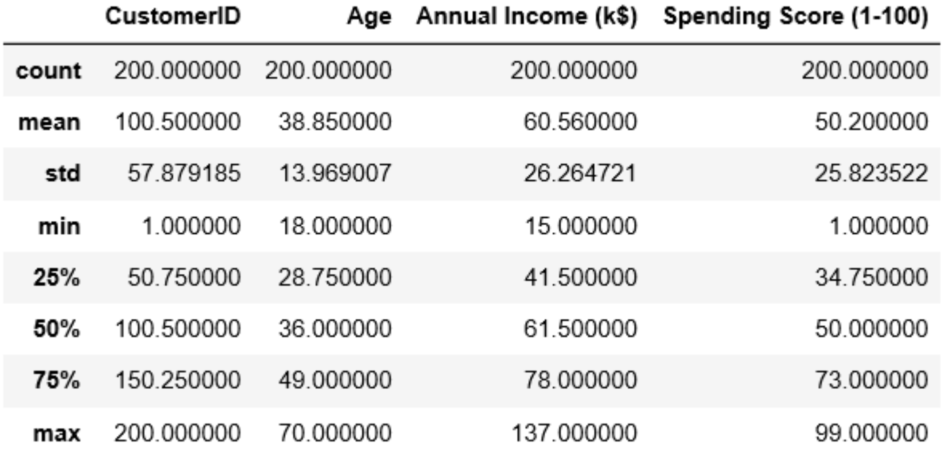
To visualise and have an idea of how our data looks like, below is the head of the data of the first 5 customers:



	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

Figure 6: Mall Customers data of first 5 customers

To understand and have a better overlook about our data, below is what the describe() function tells us about the Mall Customers data:



	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Figure 7: Describing the mathematical factors for all features

- Count of the customers gives the meaning of having 200 customers in the data.
- The minimum age of the customers from the data is 18yrs and maximum age is 70yrs. The mean age is 38yrs and the median among age is 36yrs.
- The minimum annual income of the customer data is $15,000 and maximum among the data is $137,000. The mean of annual income is $60,000 and the median is $61,000 respectively.
- Spending score ranges from 1-99 for minimum and maximum while the mean and median is 50 for the score. Spending score is assigned based on customers behaviour and purchasing data.

2. Data Exploration

Data exploration utilizes both manual data examination (regularly thought to be one of the most repetitive and tedious errands in data science) and automated tools that remove data into initial reports that incorporate data visualizations and graphs. This procedure empowers further data examination as examples and patterns are distinguished. Data exploration makes a clearer perspective on datasets instead of pouring more than many figures in unstructured data.

- Missing Data

Missing data can be a huge problem for any data scientist. Missing data becomes a huge barrier unless it is handled properly. Handling of missing data involves:

- Remove rows with missing data
- Remove rows for specific values
- Drop variables with missing data
- Impute missing data
- Predict missing data using ML

In our project we will be using heatmaps to find out the presence of any missing values. A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colours. The seaborn python package allows the creation of annotated heatmaps which can be tweaked using Matplotlib tools as per the creator's requirement.

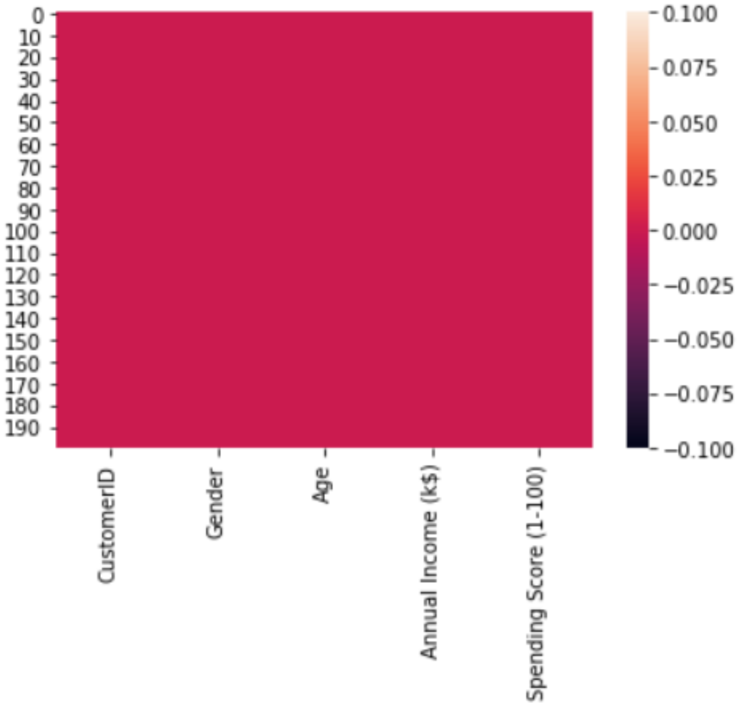


Figure 8: Heatmap of Mall Customers Data

Here we have used the heat map to check for null values and as we can see, we have zero null values in any columns.

3. Data Visualisation

- Distribution Charts

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions. It looks very much like a bar chart, but there are important differences between them. It shows the frequency of values in data by grouping it into equal-sized intervals or classes (so-called bins). In such a way, it gives you an idea about the approximate probability distribution of your quantitative data.

- Distribution of Gender in the Mall

Creating a pie chart of the distribution of gender from the Mall Customers data given to us we can see that Females are in lead with a share of 56% from the total customers whereas the Males have a share of 44%. Compared to the world population and distribution of men vs women, it’s a huge gap with the ratio of female to men who visit the mall more often.

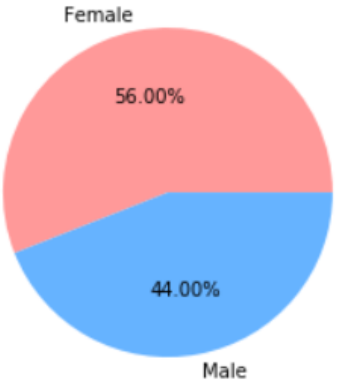


Figure 9: Pie Chart of Gender

- Distribution of Age

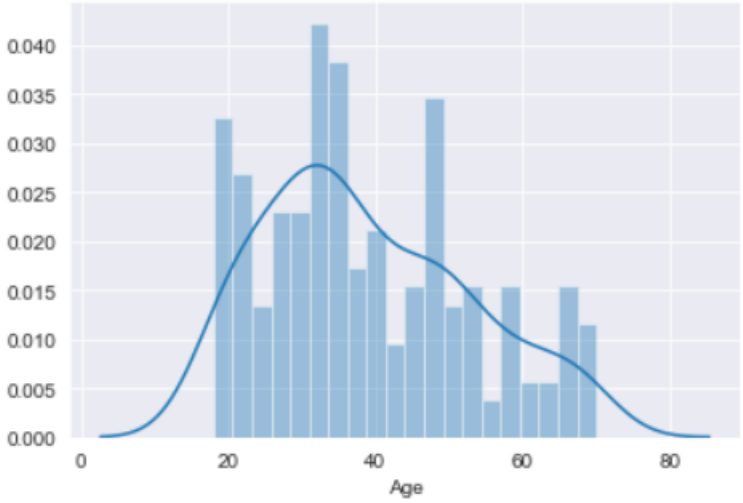


Figure 10: Distribution of Age

It tends to be pictured that the age group of 24-40 are as often as possible appearing at the malls and include in purchases. Individuals of the age 55,56,64 and 69 are less continuous to visit malls. Individuals of age 32 are the most regular visitors in the shopping centre.

- Distribution of Annual Income

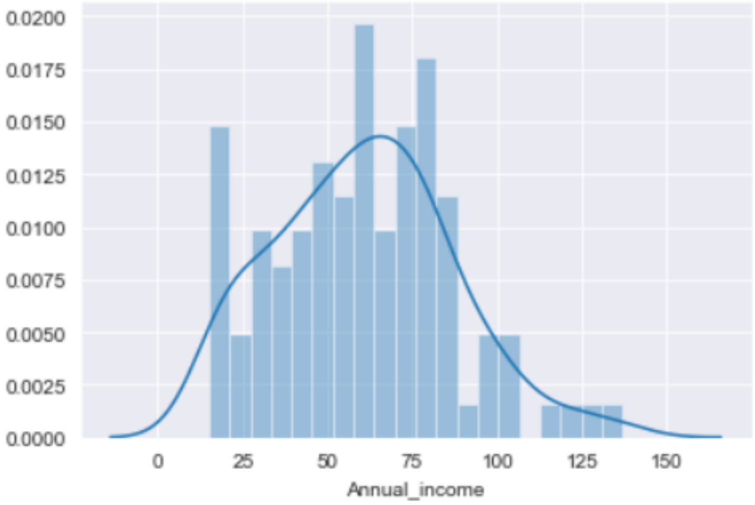


Figure 11: Distribution of Annual Income

It is pleasing to see that the customers in the mall with a very much comparable frequency with their annual income ranging from $15,000-$137,000. There are more customers in the mall who have their annual income as $54,000-$78,000.

- Distribution of Spending Score

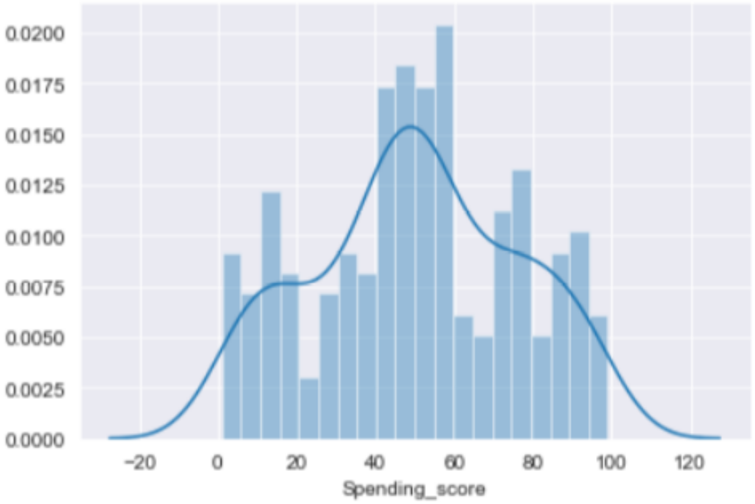


Figure 12: Distribution of Spending Score

This chart is the most important chart in the perspective of the shopping mall because it is necessary to have a prediction of an idea about the spending score of the customers visiting the mall. We can conclude that most of the customers have their spending score in the range of 35-60. It is amusing to know that there are customers with such a variety of spending scores ranging from 1-99 which shows that the shopping mall caters to the variety of customers with varying needs and requirements available in the mall.

4. Clustering

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

The optimal K value usually found is the square root of N, where N is the total number of samples. Use an error plot or accuracy plot to find the most favourable K value. KNN performs well with multi-label classes, but you must be aware of the outliers.

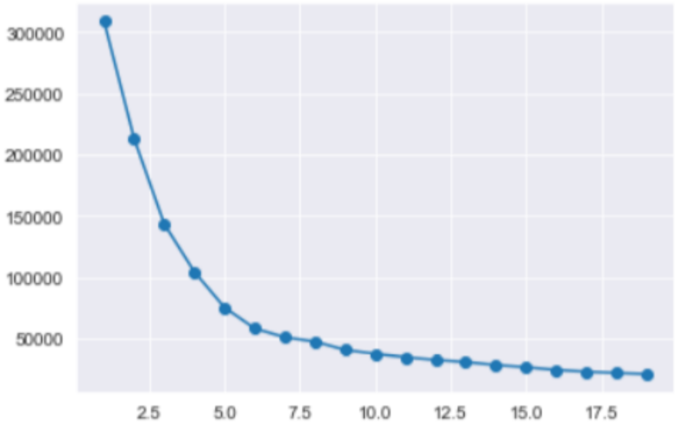


Figure 13: Elbow method for Mall Customers

After using the elbow method, we find out that n=5 is the optimal number of clusters best suited to implement the K-Means clustering. We choose the n value based on if the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

RESULTS

This clustering analysis gives us an away from about the various fragments of the customers in the mall. There are obviously five sections of customers to be specific Miser, General, Target, Spendthrift and Careful dependent on their yearly pay and spending score which are supposedly the best factors, credits to decide the fragments of a client in a mall.

I have segregated them into four distinct classifications to be specific Usual Customers, Priority Customers, Senior Citizen Target Customers and Young Target Customers and in the wake of getting the outcomes we can accordingly make diverse promoting procedures and arrangements to enhance the spending scores of the client in the mall.

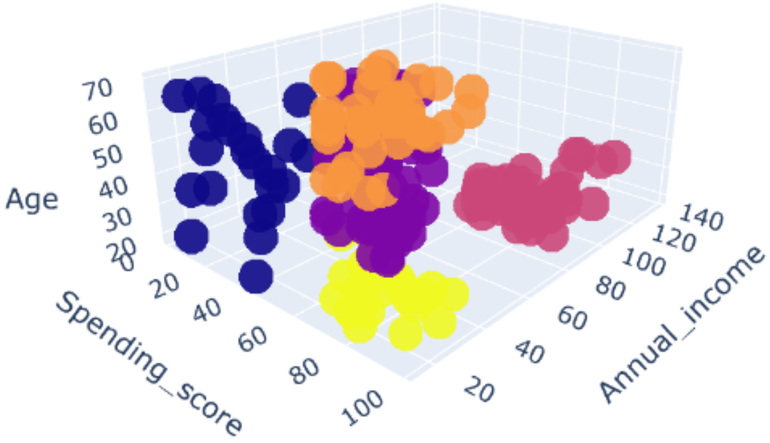


Figure 14: 3D model of the clusters formed after K-Means clustering

- In cluster 1(red coloured), we see that individuals have high income and high spending scores, this is the ideal case for the shopping centre/shops as these individuals are the prime wellsprings of benefit. These individuals may be ordinary clients of the shopping centre and are persuaded by the shopping centre's offices.

- In cluster 2(yellow coloured), we can see that the individuals have low income however high spending scores, these are the individuals who love to purchase items more regularly even though they have a low income. Possibly this is claiming these individuals are happy with the shopping centre administrations. The shopping centres probably won't focus on these individuals viably yet won't have any desire to lose them.

- In cluster 3(orange coloured), we see that the individuals have high income however low spending scores. These are the individuals who aren't happy with the shopping centre administrations. These help as prime focuses of the shopping centre as they can possibly go through cash. Thus, the shopping centre specialists will attempt to include new offices, so they can draw in these individuals and can address their issues.

- In cluster 4(blue coloured), we see that the individuals have low yearly income and low spending scores. This is very sensible as individuals having low compensations want to purchase less these are the savvy individuals who realize how to go through and set aside cash. The shops will be least keen on individuals having a place with this cluster.

- In cluster 5(purple coloured) we see that individuals have normal income and a normal spending score; these individuals won't be the objective of the shopping centres however they will be thought of and other information investigation procedures are utilized to build their spending score.

CONCLUSION

- The analysis shows there is a low score concentration in male gender (between 0 and 25 score points). In female gender, we have a high concentration in ranges between 75 and 100 compared to the male gender. In general, women have higher spending scores than men.
- The annual income distribution shows that in general men have higher annual income than women. These can be analysed together to give great insights for mall administrators.
- Senior spending scores amasses in low and medium qualities; in high score valuation, grown-ups have the most significant levels; in gender examination, youthful and senior ladies have higher spending score esteems than youthful and senior men.