**ABSTRACT**

Looking around and finding that most companies are now data-driven. They make strategic decisions based on data analysis, enabling them to examine and organize their data for better service. There has always been a lot of competition in the market as to who can provide the best customer experience, attract new customers based on their needs, and satisfy their demands, enhancing their profit and growth. However, this is not very easy and calls for various data mining techniques and algorithms. Machine learning can help them target potential customers. The algorithms deep dive into the data pool to extract hidden treasures and patterns that can bring wonderous profits to many organizations and better decision making. Customer segmentation is one such beautiful concept. Customer segmentation finds its use in many sectors. For example, in Netflix, it can be used as a recommendation system to find a group of similar users and use it for filtering, categorizing, or recommending movies. Banks or insurance companies use it for fraud detection or to evaluate certain insurance risks to segmented customers. Will be using Customer Segmentation in the retail industry, a Mall, to segment customers into various groups and target potential. The industry can then work towards attractive ideas to sell products and services inclined towards these specific

**COMPANY PROFILE**

History of the Organization Exposys Data Labs is a world leader in Robotics, Universe Intelligence (UI), Artificial Intelligence (AI) research and its applications that directly impact Planet Earth and human life.

**Objectives**

Exposys Data Labs aims to Solve real world business problems like Automation, Big Data and data Science. our core team of experts in various technologies help businesses to identify issues,oppurtunities and prototype solutions using trending technologies like AI, ML, Deep Learning and Data Science. we follow a human-focussed and not technology driven approach to achieve success in our clients endeavours.

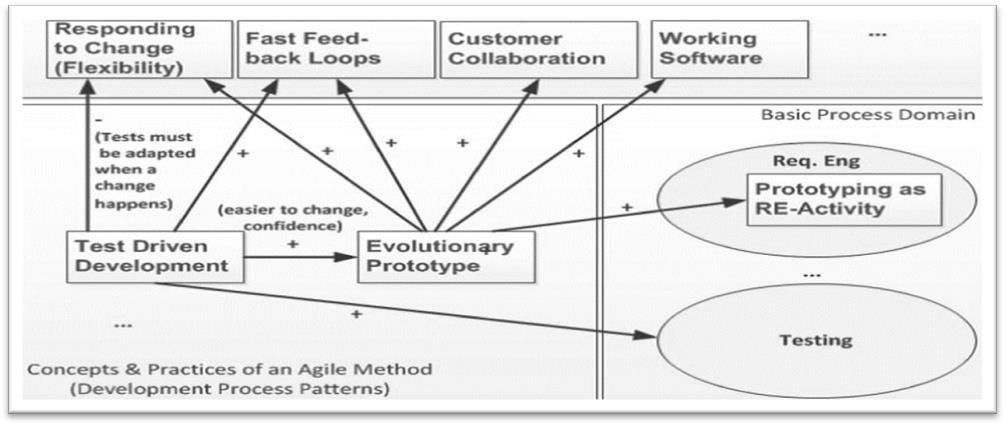
**Operation of the Organization**

We are based in Bengaluru India. Exposys Data Labs aims to Solve real world business problems like Automation, Big Data and data Science.

**Major Milestones**

* Realistic,Scalable Marketing Strategy
* Profitable Business Model
* Hire and Train a Solid Team
* Gain Authority in Your Industry

#### Structure of the Department



#### Roles and Responsibilities of Individuals

Since the internship was remotely conducted by the company, to ensure easy onboarding of interns, the company had individuals who took care of the smooth run of online training.

* Operation and Strategy Head- Ensured there were no difficulties for interns while onboarding. Best of mentors and doubt clarifying sessions were arranged too.
* Technical Lead- Ensured the technicalities of online training to be smooth. Bestplatforms were arranged for our meetings and trainings.
* Mentors- They have helped us to understand the concepts, gave us tasks to get practical take a way and clarified doubts to the best.
* Interns- Worked through the tasks given either individually or in a group.

#### Training Program

The internship is a platform where the trainees are assigned with the specific task. In the initial days of the internship, I was trained on the following:

* Python Programming
* Machine Learning Algorithms

**DATA SET:**

For this project we have used Mall Customer Dataset, our main objective is to divide customers into groups according to common characteristics.

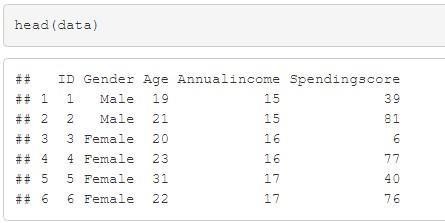


Table: Fields in Mall Customer CSV File

**DATASET EXTRACTION AND TRANSFORMATION:**

We imported the dataset and using Elbow Method and K-means. we identified the segments of customers to target the potential user base. By observing these customers are divided into groups according to common characteristics.

**Elbow Method:**

In [cluster](https://en.wikipedia.org/wiki/Cluster_analysis) [analysis,](https://en.wikipedia.org/wiki/Cluster_analysis) [t](https://en.wikipedia.org/wiki/Cluster_analysis)he elbow method is a [heuristic](https://en.wikipedia.org/wiki/Heuristic) [u](https://en.wikipedia.org/wiki/Heuristic)sed in [determining](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [the](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [number](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [of](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [clusters in](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [a](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [data](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) [set.](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) The method consists of plotting the [explained](https://en.wikipedia.org/wiki/Explained_variation) [variation](https://en.wikipedia.org/wiki/Explained_variation) [a](https://en.wikipedia.org/wiki/Explained_variation)s a function of the number of clusters, and picking the [elbow](https://en.wikipedia.org/wiki/Elbow_of_the_curve) [of](https://en.wikipedia.org/wiki/Elbow_of_the_curve) [the](https://en.wikipedia.org/wiki/Elbow_of_the_curve) [curve](https://en.wikipedia.org/wiki/Elbow_of_the_curve) as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of [principal](https://en.wikipedia.org/wiki/Principal_component) [components](https://en.wikipedia.org/wiki/Principal_component) to describe a data set.

**K-Means:**

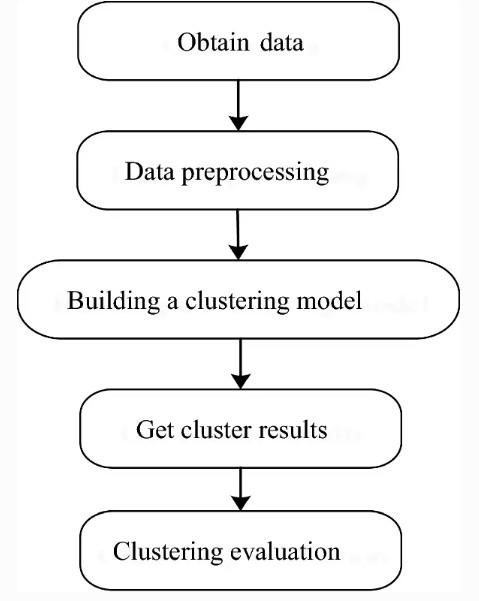
K-Means clustering is a method of [vector](https://en.wikipedia.org/wiki/Vector_quantization) [quantization,](https://en.wikipedia.org/wiki/Vector_quantization) originally from [signal](https://en.wikipedia.org/wiki/Signal_processing) [processing,](https://en.wikipedia.org/wiki/Signal_processing) that aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) n observations into k clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) [w](https://en.wikipedia.org/wiki/Cluster_(statistics))ith the nearest [mean](https://en.wikipedia.org/wiki/Mean) [(](https://en.wikipedia.org/wiki/Mean)cluster centers or cluster [centroid)](https://en.wikipedia.org/wiki/Centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into [Voronoi](https://en.wikipedia.org/wiki/Voronoi_cell) [cells.](https://en.wikipedia.org/wiki/Voronoi_cell) k-means clustering minimizes within-cluster variances ([squared](https://en.wikipedia.org/wiki/Squared_Euclidean_distance) [Euclidean](https://en.wikipedia.org/wiki/Squared_Euclidean_distance) [distances)](https://en.wikipedia.org/wiki/Squared_Euclidean_distance), but not regular Euclidean distances, which would be the more difficult [Weber](https://en.wikipedia.org/wiki/Weber_problem) [problem:](https://en.wikipedia.org/wiki/Weber_problem) the mean optimizes squared errors, whereas only the [geometric](https://en.wikipedia.org/wiki/Geometric_median) [median](https://en.wikipedia.org/wiki/Geometric_median) [m](https://en.wikipedia.org/wiki/Geometric_median)inimizes Euclidean distances. For instance, better Euclidean solutions can be found using [k-medians](https://en.wikipedia.org/wiki/K-medians_clustering) [a](https://en.wikipedia.org/wiki/K-medians_clustering)nd [k-medoids.](https://en.wikipedia.org/wiki/K-medoids)

The problem is computationally difficult ([NP-hard)](https://en.wikipedia.org/wiki/NP-hardness); however, efficient [heuristic algorithms](https://en.wikipedia.org/wiki/Heuristic_algorithm) [c](https://en.wikipedia.org/wiki/Heuristic_algorithm)onverge quickly to a [local](https://en.wikipedia.org/wiki/Local_optimum) [optimum.](https://en.wikipedia.org/wiki/Local_optimum) [T](https://en.wikipedia.org/wiki/Local_optimum)hese are usually similar to the [expectationmaximization](https://en.wikipedia.org/wiki/Expectation-maximization_algorithm) [algorithm](https://en.wikipedia.org/wiki/Expectation-maximization_algorithm) for [mixtures](https://en.wikipedia.org/wiki/Mixture_model) [o](https://en.wikipedia.org/wiki/Mixture_model)f [Gaussian](https://en.wikipedia.org/wiki/Gaussian_distribution) [distributions](https://en.wikipedia.org/wiki/Gaussian_distribution) via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the Gaussian mixture model allows clusters to have different shapes. The unsupervised k-means algorithm has a loose relationship to the [k-nearest](https://en.wikipedia.org/wiki/K-nearest_neighbor) [neighbor classifier,](https://en.wikipedia.org/wiki/K-nearest_neighbor) a popular supervised [machine](https://en.wikipedia.org/wiki/Machine_learning) [learning](https://en.wikipedia.org/wiki/Machine_learning) technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as [nearest](https://en.wikipedia.org/wiki/Nearest_centroid_classifier) [centroid](https://en.wikipedia.org/wiki/Nearest_centroid_classifier) [classifier](https://en.wikipedia.org/wiki/Nearest_centroid_classifier) [o](https://en.wikipedia.org/wiki/Nearest_centroid_classifier)r [Rocchio](https://en.wikipedia.org/wiki/Rocchio_algorithm) [algorithm.](https://en.wikipedia.org/wiki/Rocchio_algorithm)

#### Experience

As per our experience during the internship, Exposys Data Labs follows a good work culture and it has friendly employees, starting from the staff level to the management level. The trainers are well versed in their fields and they treat everyone equally. There is no distinguishing between fresher graduates and corporates and everyone is respected equally. There is a lot of teamwork followed in every task, be it hard or easy and there is a very calm and friendly atmosphere maintained at all times. There is a lot of scope for self-improvement due to the great communication and support that can be found. Interns have been treated and taught well and all our doubts and concerns regarding the training or the companies have been properly answered. All in all, Knowledge Solutions India was a great place for a fresher to start career and also for a corporate to boost his/her career.

#### Data Flow Diagram



#### 5 Implementation

##### Customer segmentation

Customer segmentation is partitioning a customer database into group of people with similar characteristics. It is an application of unsupervised learning. It is a business strategy that allows targeting a specific group of customers and effectively allocate marketing resources. For such large datasets, we need an analytical approach like clustering to make customer segments.

There are four major ways of segmentation, i.e., geographical, economic, demographic, and behavioural patterns.

In this project, we divide a Mall customer's dataset based on gender, age, income, spending habits, etc. We also visualize gender and age distributions and analyse their annual incomes and spending scores to target the potential user base. The method used is K-means clustering.

Language: Python

##### Clustering

Clustering can group data unsupervised solely based on similarities to each other. It will partition customers into mutually exclusive groups aka clusters. Having the result would help understand and predict customer preferences and differences, thus making the company deliver personalised experiences for each group of customers.

Types of Clustering:

Partition-based clustering is a group of clustering algorithms that produces sphere-like clusters, such as; K-Means, K-Medians or Fuzzy c-Means. These algorithms are relatively efficient and are used for medium to large- sized databases.

Hierarchical clustering algorithms produce trees of clusters, such as agglomerative and divisive algorithms. This group of algorithms are very intuitive and are generally suitable for use with small-size datasets.

Density-based clustering algorithms produce arbitrary-shaped clusters. They are outstanding when dealing with spatial clusters or noise in the data set, for example, the DB scan algorithm.

#### Screen Shots

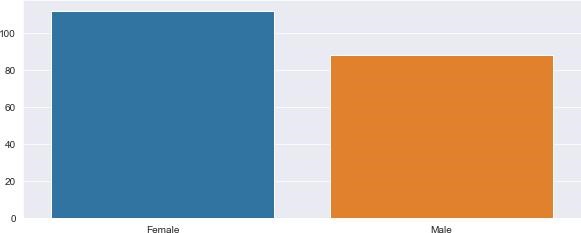


Figure Gender Distribution

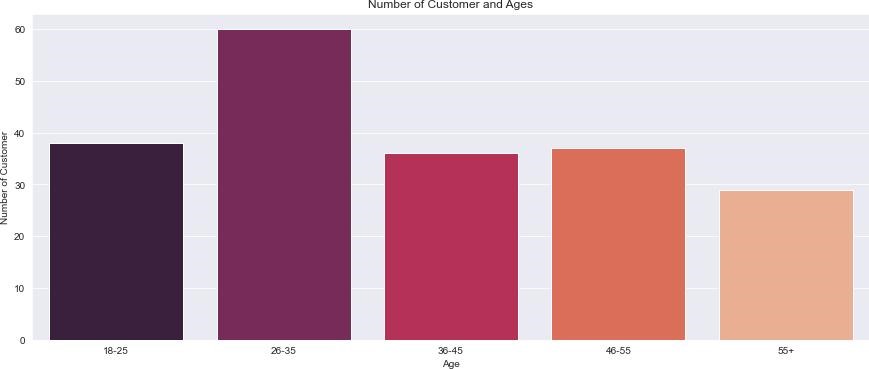


Figure Age Distribution

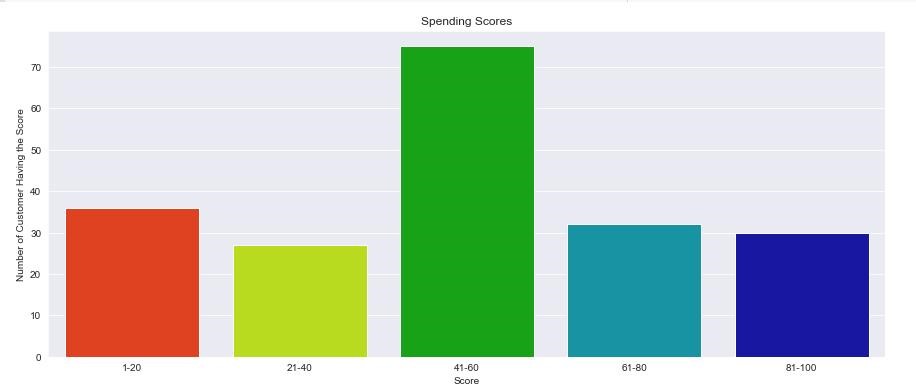


Figure Annual Income Distribution

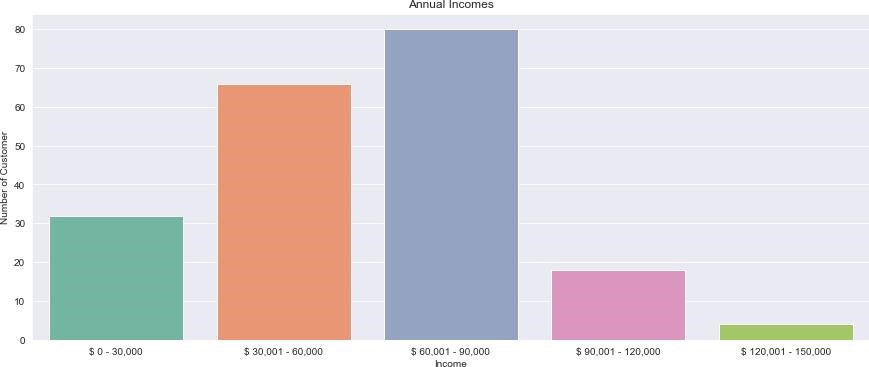


Figure Spending score Distribution

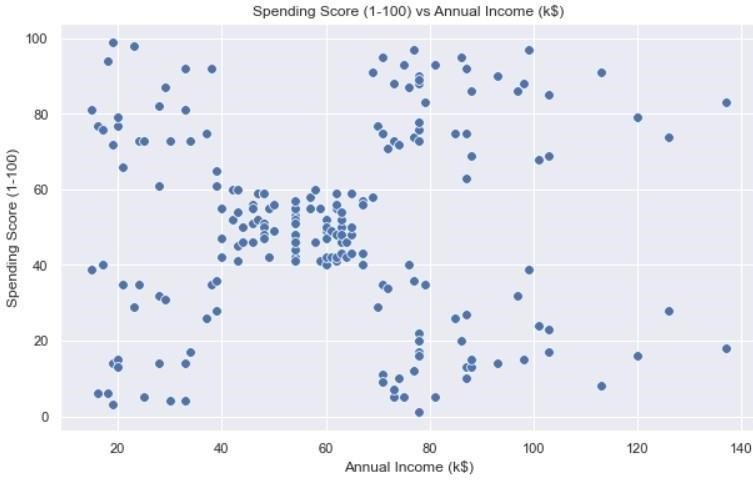


Figure Annual Income vs Spending Score

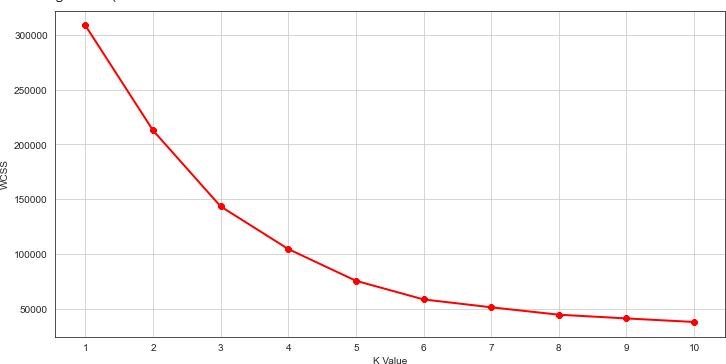


Figure Elbow Graph

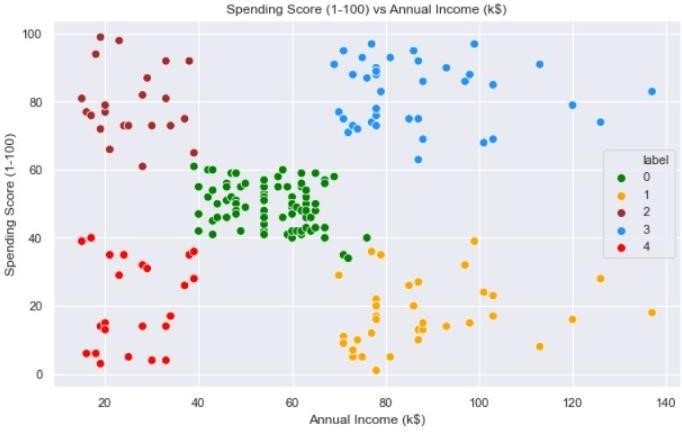
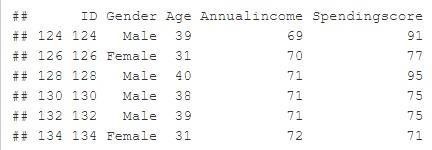


Figure Graph after applying K-Means

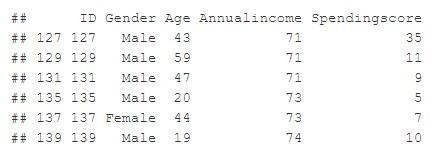
### CONCLUSION

The goal of K means is to group data points into distinct non-overlapping subgroups.

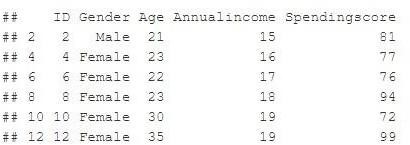
Cluster 3: high spending scores and high-income; alert them with new arrivals as they are potential customer for increase in revenue.



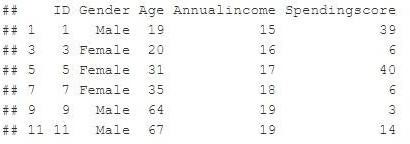
Cluster 1: high income and low spending score; ask them for feedback and advertise them with new products that might attract them, they have the potential to convert into cluster 4.



Cluster 2: low income and high spending scores; can help them by providing new deals and sales offers so that despite low income they still remain loyal.



Cluster 4: low income and low spending score; it won’t be beneficial to both the parties to target these customers.



Rest are average and the company can use them according to market conditions.