Ed-Tech
Market Segmentation

Contributors

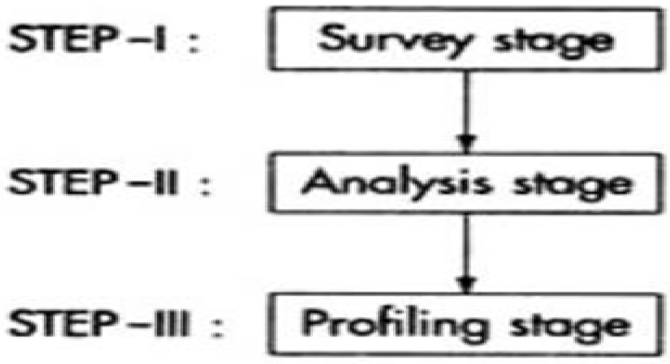
NISHANTH SINHA
TANUSHRI KUMAR
SAHASI KORRAPATI

Batch 12 - team F

Introduction

Market segmentation is founded on the presumption of winning large by specializing onethemall. Big thinking isn't always the greatest strategy. Targeting tiny market niches in emerging economies, even some extremely huge markets like India and China, can be a more effective strategy.

What is Market Segmentation?
Market segmentation is a marketing phrase that refers to grouping prospective buyers based on their demands and responses to marketingactions.One definition of market segmentation is the split of a market into subsets of customers to simplify target branding and marketing methods. It's simpler to approach someone when you know who you're attempting to attract and what they're interested in.



What Is Edtech?

Edtech is the practice of introducing IT tools into the classroom to create a more engaging, inclusive and individualized learning experience.

Today's classrooms have moved beyond the clunky desktop computers that were once the norm and are now tech-infused with tablets, interactive online courses and even robots that can take notes and record lectures for students who are ill.

This influx of edtech tools are changing classrooms in a variety of ways: edtech robots are making it easy for students to stay engaged through fun forms of learning; IoT devices are being hailed for their ability to create digital classrooms for students, whether they're physically in school, on the bus or at home; even machine learning and blockchain tools are assisting teachers with grading tests and holding students accountable for homework.

The potential for scalable individualized learning has played an important role in edtech's ascendance. The way we learn, how we interact with classmates and teachers, and our overall enthusiasm for the same subjects is not a one-size-fits-all situation. Everyone learns at their own pace and in their own style. Edtech tools make it easier for teachers to create individualized lesson plans and learning experiences that foster a sense of inclusivity and boost the learning capabilities of all students, no matter their age or learning abilities.

And it looks like technology in the classroom is here to stay — 92% of teachers believe tech is going to have a major impact on the way they educate in the near future. For that reason, it's vital to understand the benefits edtech brings in the form of increased communication, collaboration and overall quality of education.

Fermi Estimation

"Better to be approximately right than precisely wrong"
What is it? The process of formulating a solution to a problem based on a set of logical assumptions. The outcome will be an order-of-magnitude solution.
How are we going to do it? Questions of this sort are frequently severe in nature, and thus cannot typically be answered using ordinary mathematical or scientific knowledge, such as How many pizzas do students in a hostel consume on a daily basis? It's tempting to just say "probably a million" and move on with your life, but this is a problem that can be. Here's how we might be able to tackle the problem.

1. This year, how many students are enrolled in educational institutions in India? 250 million is a good guess.
2. How many students can afford and are interested in taking online learning assistance ? roughly 50%.

So, if we do the arithmetic, out of the entire students in India 80 million students are interested in taking online learning assistance.

What's more, Fermi demonstrated that this strategy for predicting values is very precise. Your personal estimates are sometimes exaggerated. They might be too low at times. But, in the end, everything balances out.

Problem Statement

Understanding the elements that influence the Ed-Tech market, such as location, facilities,quality and consumer behavior, is the first step in estimating Ed-Tech market infographics.

We are going to work for a launch of a Learning Assistant app that has major roles in online tutoring and simplifying the concepts through animation.

We are going to analyze the Ed-Tech market in India using segmentation analysis and come up with a feasible strategy to enter the market , targeting the segments most likely to use their product in terms of geographic , demographic, psychographic and behavioral data of consumers.

Data sources

- 2015_16_Districtwise (csv file)

- 2015_16_Statewise_Elementary (csv file)
- 2015_16_Statewise_Secondary (csv file)

  We have chosen data sets from which we could find a solution for the problem statement, taking into geographic and demographic and physical characteristics of the students enrolled in schools. The data we chose will largely determine the sort of conclusion we obtain from the market segmentation. Data plays a significant part in this and we gathered data from reliable sources.

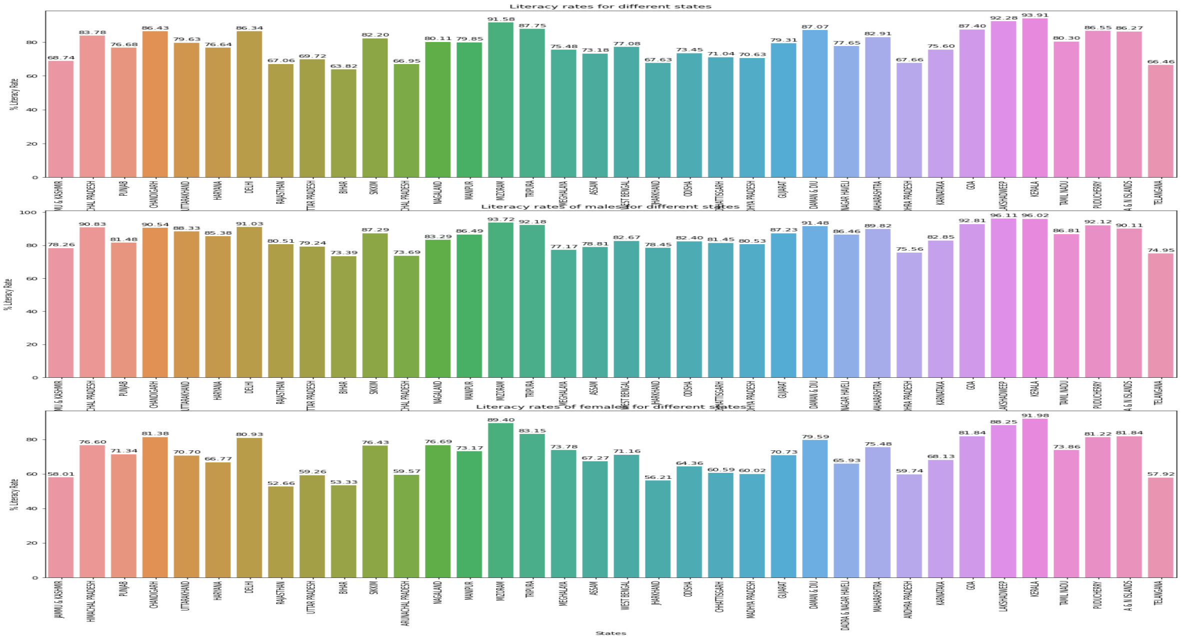
Exploratory Data Analysis

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

We have briefly performed EDA for getting a better perspective of the collected data and to get initial-kick start insights for market segmentation. The major python libraries we used are numpy,pandas,matplotlib.pyplot,seaborn.
In EDA

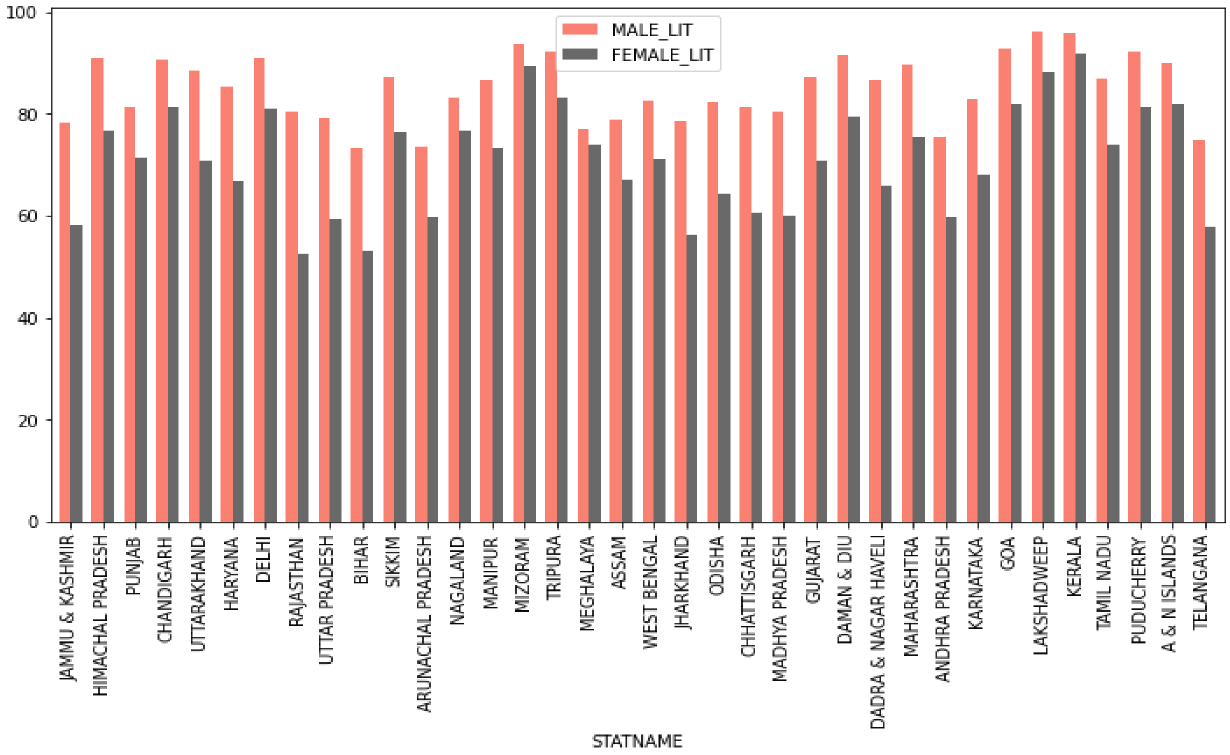
- We have checked the first five rows of all three data sets using head().
- We got the shape of all three data sets that are : (680,819),(36,816),(36,630)
- We got the unique states and districts from all the data sets.

- We have visualized literacy rates of different states,literacy rates of males for different states,literacy rates of females for different states.



States with highest and lowest literacy rates secondary and elementary have same literacy rates stats

- Found differences in literacy rates between males and females in different states.



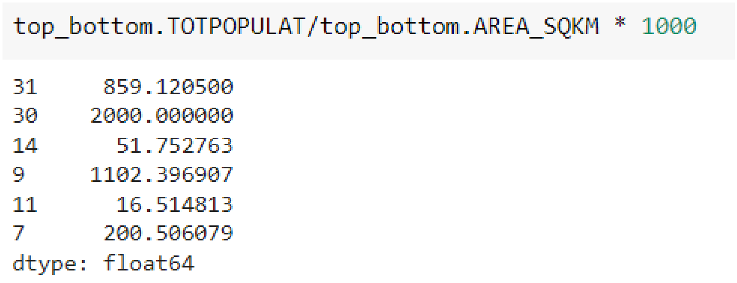
The states with the highest overall literacy rates are : Kerala, Lakshadweep, Mizoram, Tripura and Goa. The states with the lowest overall literacy rates are : Bihar, Telangana, Arunachal Pradesh, Rajasthan and Jharkhand.

Analysis between the 3 top most literacy rate states and 3 bottom most literacy rate states. Creating a data frame top_bottom containing only top3 and bottom 3 states with respect to literacy rates.

6 states taken are : Kerala,Lakshadweep,Mizoram,Bihar,Arunachal Pradesh,Rajasthan

Exploring features that might affect literacy rates.

1.total population

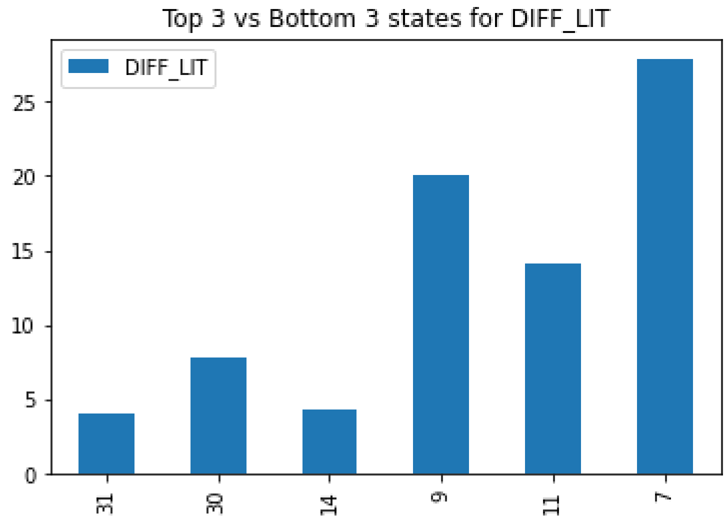


```
top_bottom.TOTPOPULAT/top_bottom.AREA_SQKM * 1000
```

```
31     859.120500
30    2000.000000
14      51.752763
9     1102.396907
11      16.514813
7      200.506079
dtype: float64
```

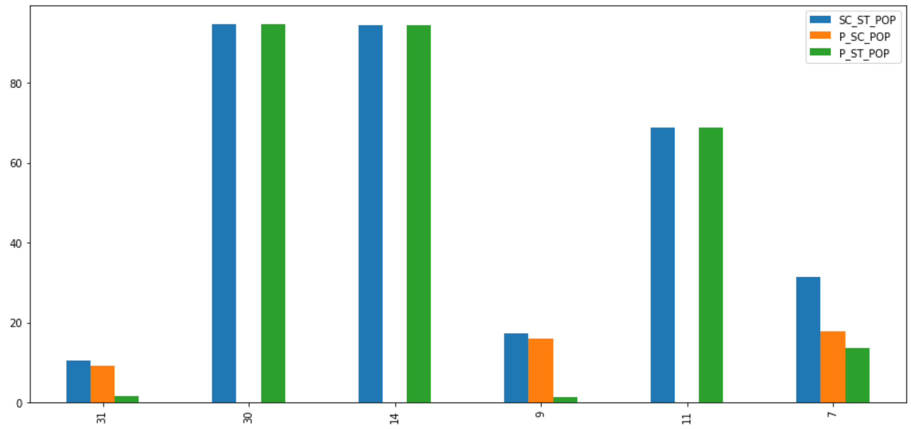
It was evident that population density doesn't affect the literacy rates . there doesn't seem to be any particular trend between the top 3 most and bottom 3 most states.

2.using feature differences in literacy rates of male and female.



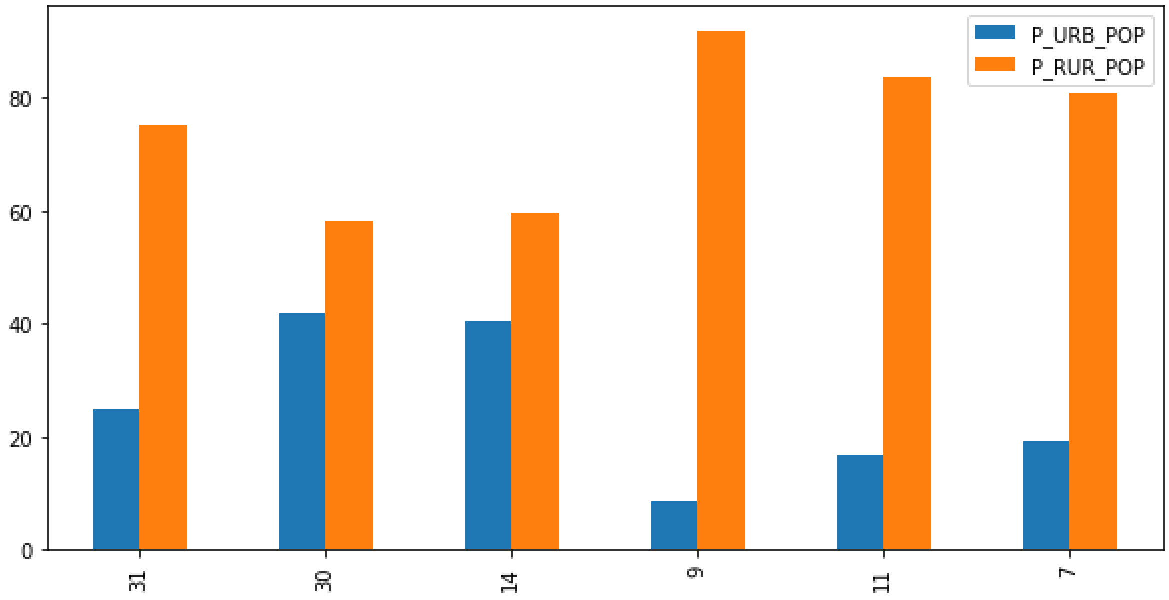
DIFF_LIT says a lot. The differences are really high in states with low overall literacy rates. So even if the bottom most states have good male literacy rates, female literacy rates are really low and that takes their overall literacy rate down. Thus these states really need to work on educating their females and increasing their literacy rate.

3. using features P_SC_POP and P_ST_POP.



The overall literacy rate doesn't depend on the SC and ST population proportions because these are very different for all the above states.

4. P_RUR_POP is created which is 100 - urban population proportion. This will help us in comparing the rural and urban population proportions.



The difference between rural and urban population is much much bigger in the bottom 3.The rural population percentage in the bottom 3 states is much more than the rural population percentage in the top 3. That's an important factor to note. People living in rural areas lead a very different life compared to the people living in urban areas.Not many children go to school as there isn't much awareness among rural people.Children mostly take up their parents work.

Data Preprocessing

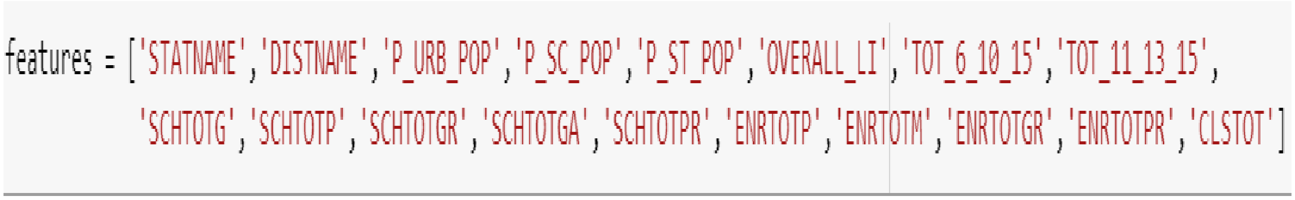
Data preparation is critical in any data mining process since it directly affects the project's success rate. If there are missing attributes, attribute values, noise or outliers, and duplicate or incorrect data, the data is considered to be unclean. If any of these are present, the quality of the findings will suffer. The major python libraries we used for data pre-processing are numpy,pandas,matplotlib.pyplot,seaborn and sci-kit learn for PCA.

- Data labeling -data labeling is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels to provide context so that a machine learning model can learn from it.

- Label encoding -We have converted the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning

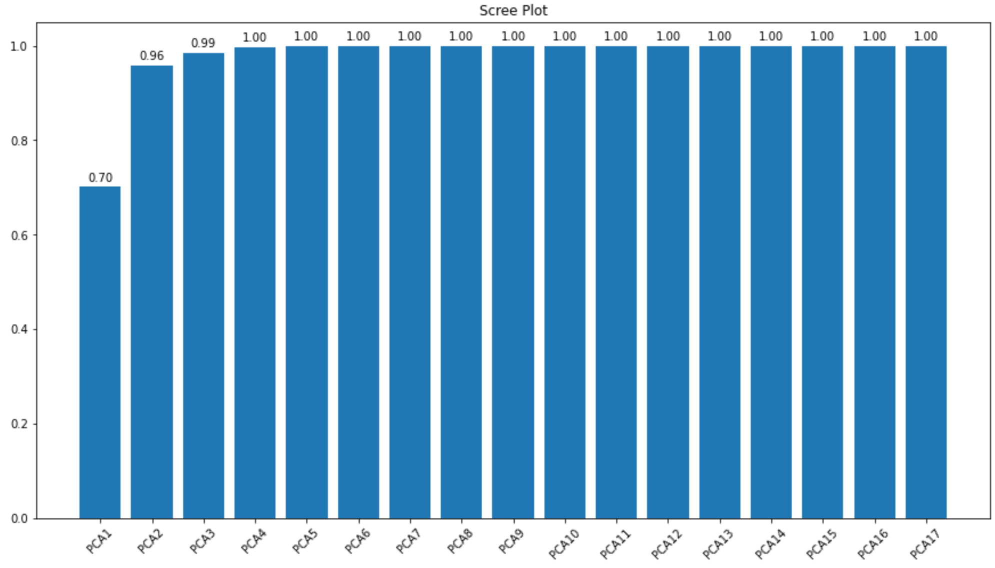
- Dropping Null values -The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

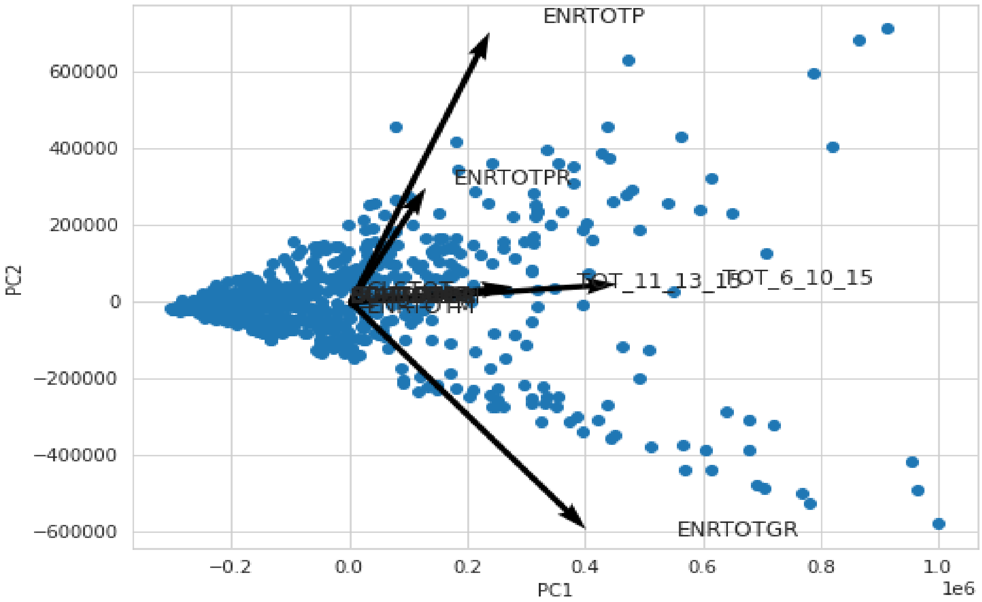
- We have extracted relevant features



```
features = ['STATNAME','DISTNAME','P_URB_POP','P_SC_POP','P_ST_POP','OVERALL_LI','TOT_6_10_15','TOT_11_13_15',
            'SCHTOTG','SCHTOTP','SCHTOTGR','SCHTOTGA','SCHTOTPR','ENRTOTP','ENRTOTM','ENRTOTGR','ENRTOTPR','CLSTOT']
```

- We have performed PCA-principal to represent a multivariate data table as a smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables and to adjust the dimensionality of the data and to proceed with segment extraction.





Visualization of PCA

Segment Extraction

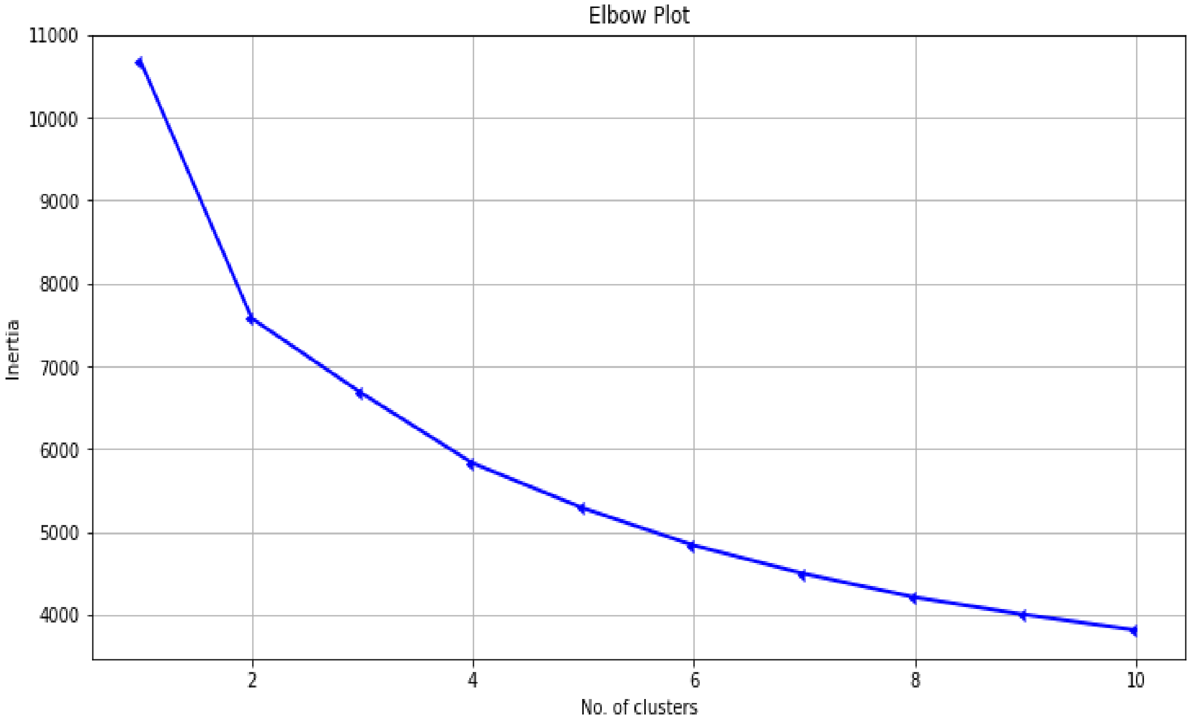
For extracting appropriate segments we have used different ML techniques for clustering like k-means , hierarchical clustering, DBSCAN (density-based spatial clustering of applications with noise).

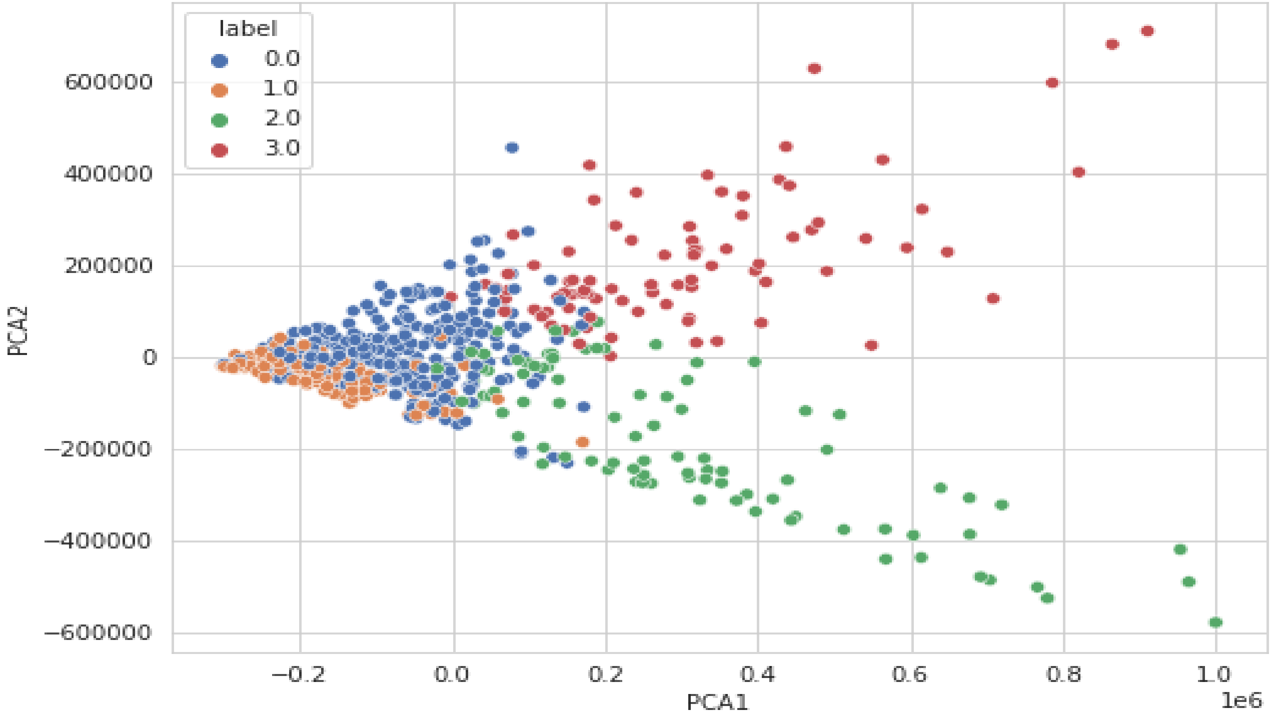
K Means Clustering Algorithm:

The k-means clustering algorithm is an iterative process of moving cluster centers or centroids to the mean position of their constituent points and reassigning instances to their closest clusters until there is no significant change in the number of cluster centers possible or a number of iterations is reached. The k-means clustering technique primarily accomplishes two goals: • 'Uses an iterative procedure to find the optimal value for K center points or centroids, then assigns each data point to the nearest kcenter.

A cluster is formed by data points that are close to a specific k-center. One of the most prominent methods for determining the ideal number of clusters is the Elbow approach. This approach makes use of the WCSS value notion. It's the sum of the squares of the distances between each data point and its cluster's centroid1, with the other two terms being the same. We can use any approach, such as Euclidean distance or Manhattan distance, to calculate the distance between data points and the centroid. The elbow technique uses the procedures below to get the best cluster value:

- It performs K-means clustering on a dataset for various K values (ranges from 1-10).
- Calculate the WCSS value for each value of K.
- Draws a line between the WCSS values computed and the number of clusters K.
- If a sharp point of bend or a plot point resembles an arm, that point is regarded as the optimal K value.





Hierarchical clustering Algorithm:

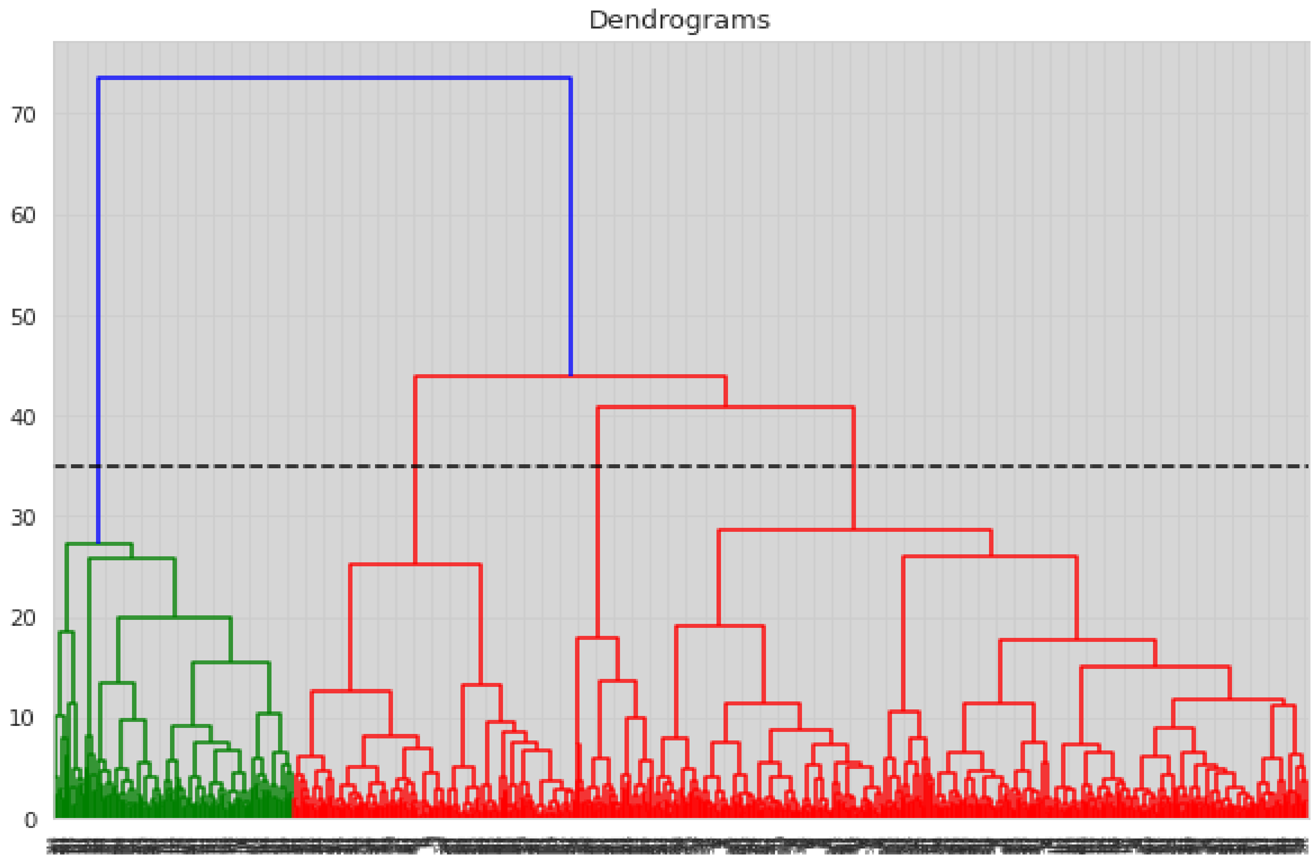
In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).

Agglomerative clustering:

Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method). At first every data set is considered as an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.

Algorithm for Agglomerative Hierarchical Clustering is:

- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
- Consider every data point as a individual cluster
- Merge the clusters which are highly similar or close to each other.
- Recalculate the proximity matrix for each cluster
- Repeat Step 3 and 4 until only a single cluster remains.



-

DBSCAN algorithm:

Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

The DBSCAN algorithm uses two parameters:

> minPts**:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
> eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point.

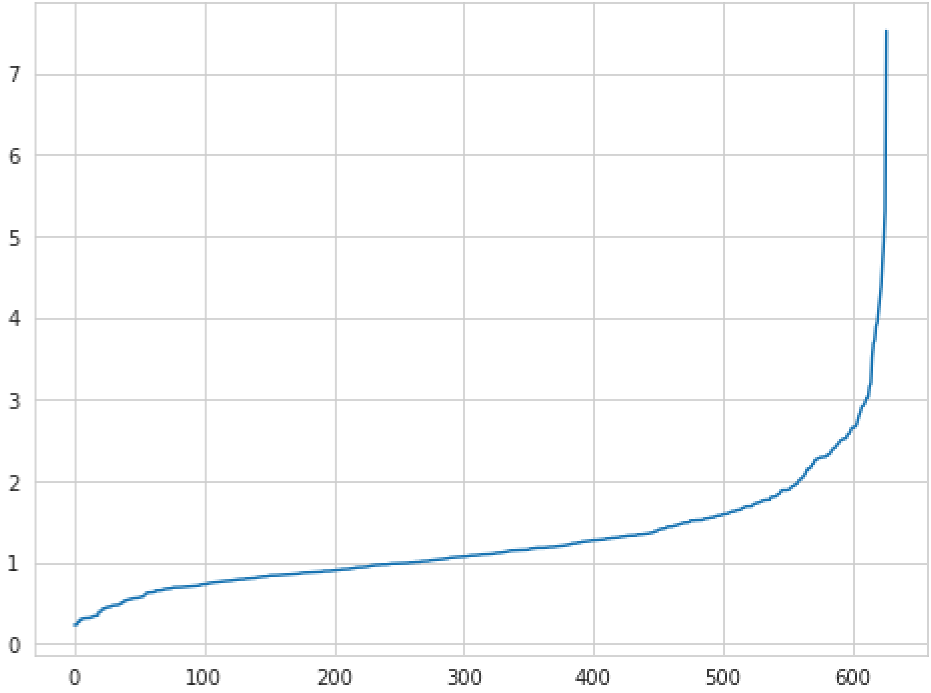
These parameters can be understood if we explore two concepts called Density Reachability and Density Connectivity.

Reachability in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it.

Connectivity, on the other hand, involves a transitivity based chaining-approach to determine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighborhood of a.

Algorithmic steps for DBSCAN clustering

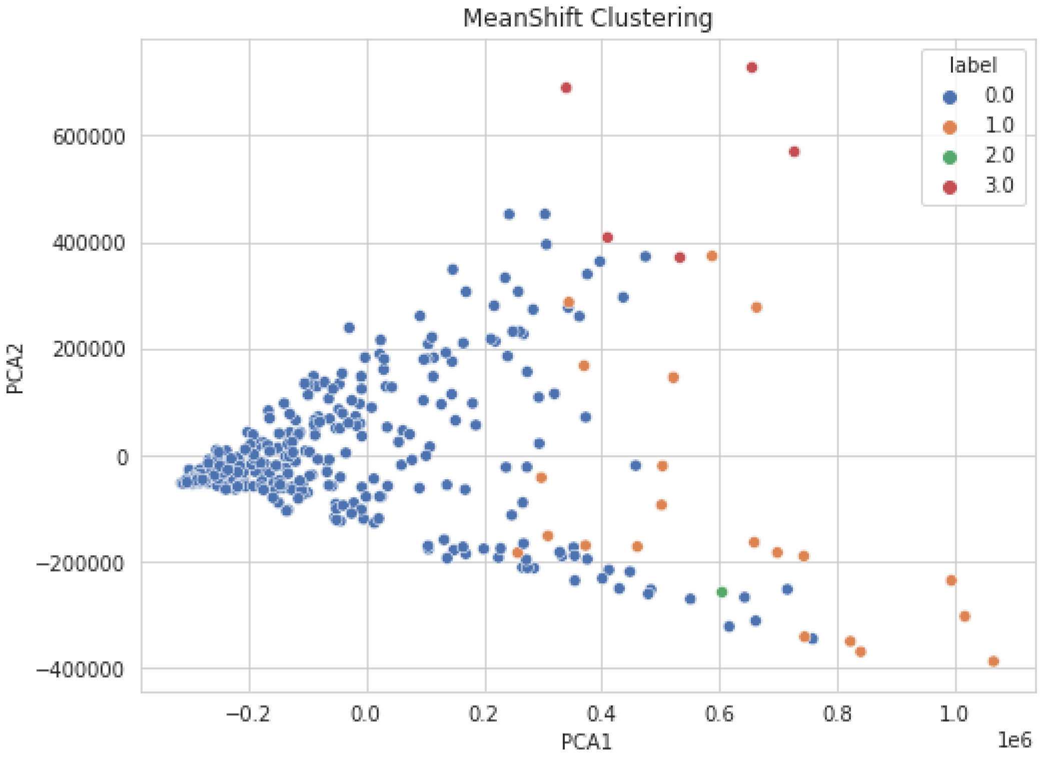
- The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited).
- If there are at least ‘minPoint’ points within a radius of ‘ε’ to the point then we consider all these points to be part of the same cluster.
- The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring points



DBSCAN algorithm is not working properly on the dataset as it is a density based algorithm and after visualization it can be seen that our data will not work well for density based algorithms

Mean shift algorithm:

Mean shift is a non-parametric feature-space mathematical analysis technique for locating the maxima of a density function, a so-called mode-seeking algorithm.[Application domains include cluster analysis in computer vision and image processing.



Analyzing the market segments

Numerous bases can be used to split the market into various segments. Market segmentation bases may be roughly classified into two types. Consumer traits(consumer segmentation) and reactions (product related segmentation). We'll be diving deeper into the former.

Geographic Segmentation:

The most common and popular foundation for market segmentation is geographic geography. It's a marketing technique that focuses on selling items to individuals who reside or shop in a certain area. This method is especially beneficial if you offer items that are affected by regional variances in culture, climate, or population. The size of a city can influence a buyer's desire. Even within the same city or suburb, there can be variances in customer preferences. Any population separated by geography that also separates customers into groups with similar demands might be useful to a marketer. Geographically, the All-India dataset may be segmented.

Demographic Segmentation:

Demographic segmentation is the process of categorizing a target market based on factors such as age, education, and gender. It is a sort of market segmentation that aids firms in better understanding their customers and successfully meeting their demands.

Psychographic Segmentation:

Psychographic segmentation is defined as a market segmentation technique where groups are formed according to psychological traits that influence consumption habits drawn from people's lifestyle and preferences. It is mainly conducted on the basis of "how" people think and "what" do they aspire their life to be.

Behavioral Segmentation:

Behavioral segmentation refers to a process in marketing which divides customers into segments depending on their behavior patterns when interacting with a particular business or website.

Volume Segmentation:

The division of a market into segments on the basis of the varying volume of demand for the product by individuals, groups or types of customers; typically, the segments are ranked to denote heavy usage, medium usage or light usage.

Product Space Segmentation:

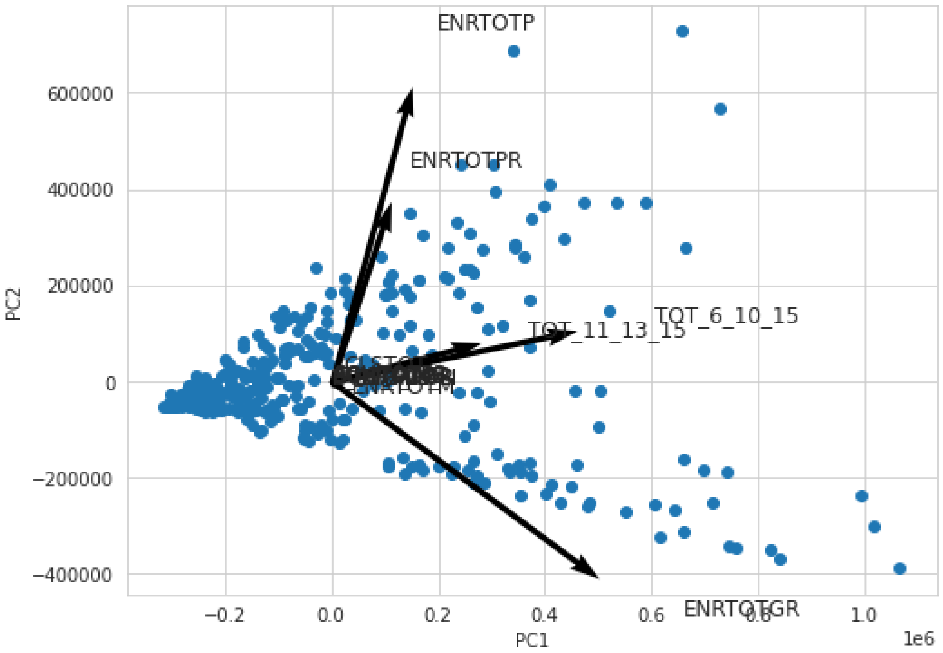
Product segmentation is when a company modifies its product into several different products in order to attract different kinds of customers or target different markets.

Market segmentation simply modifies the marketing strategy in an effort to do the same. In real estate housing this segmentation may not prove that much beneficial.

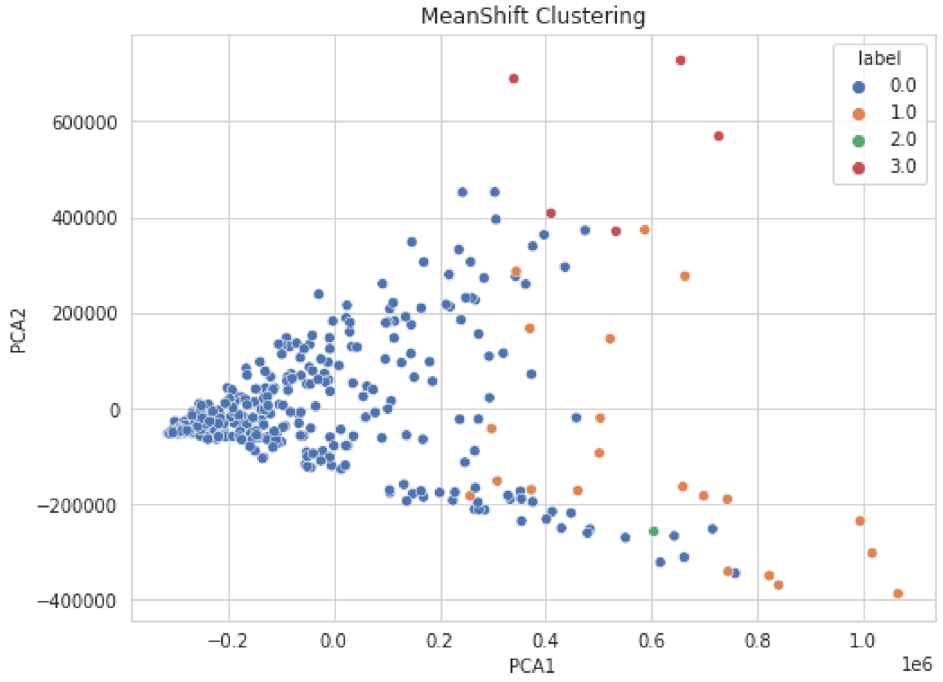
Benefit Segmentation:

Benefit segmentation is a method of market segmentation that involves segmenting your market based on the perceived value or advantages that consumers believe they will receive from your product. This can involve categorizing consumers based on perceived benefits such as quality, features, customer service, etc.

Selection of target segment



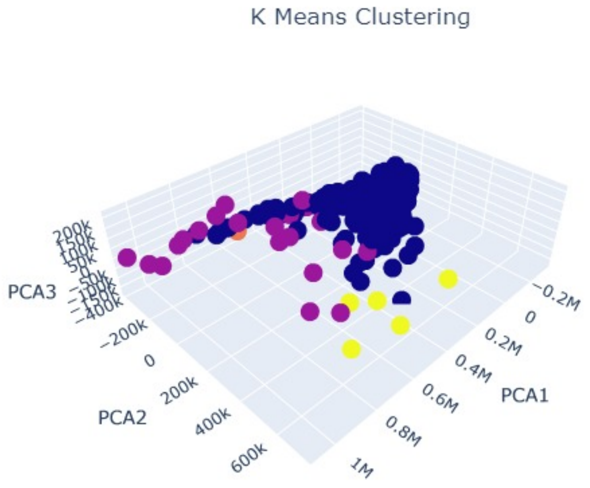
The PCA transformation has majority of these features namely ENRTOTP , ENRTOTPR,TOT_6_10_15,TOT_11_13_15. The alignment of the data points along the direction of the PCA plot signifies that increase of that particular feature.

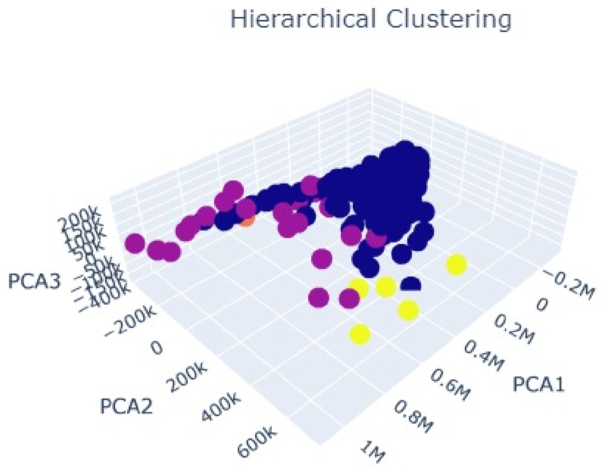


The segment 0 and segment 2 consists of populations who are having low enrollment rate in total, government and rural.

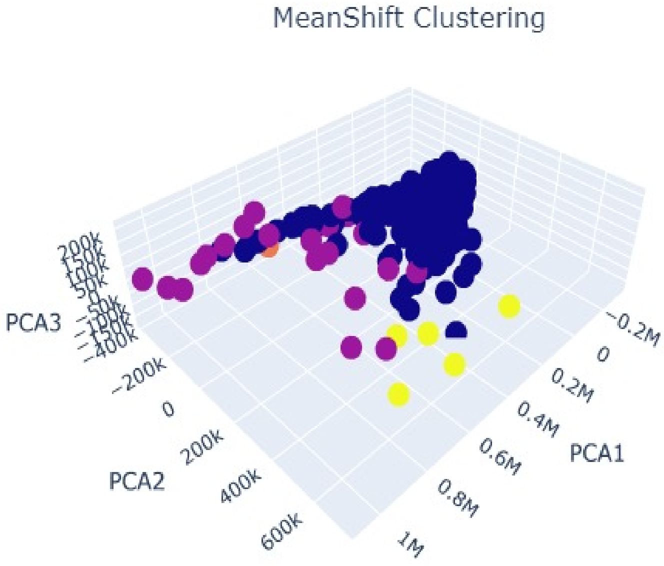
The segment 1 and segment 3 consists of populations who are having high enrollment rates in total, government and rural.

THE 3D VISUALIZATIONS OF SEGMENTS





Hierarchical Clustering



MeanShift Clustering

So we have observed from the analysis we have earlier done in the EDA, Data Pre-Processing and Segment Extraction. We have seen that urban populations had high literacy rates since they can be more interested in availing the latest services/courses that our learning assistant app would be providing. The targeted customers should be primarily students enrolled in any educational institutions.

Customizing the Marketing Mix

A marketing mix includes multiple areas of focus as part of a comprehensive marketing plan. The term often refers to a common classification that began as the four Ps: product, price, placement, and promotion.

Effective Marketing touches on a broad range of areas as opposed to fixating on one message. Doing so helps reach a wider audience, and by keeping the four Ps in mind, marketing professionals are better able to maintain focus on the things that really matter. Focusing on a marketing mix helps organizations make strategic decisions when launching new products.

Product
This represents an item or service designed to satisfy customer needs and wants. To effectively market a product or service, it's important to identify what differentiates it from competing products or services. It's also important to determine if other products or services can be marketed in conjunction with it. The product that is being developed here is a LEARNING ASSISTANT APP. The traditional learning system will be tackled through this. This learning app will increase the ease of learning through animations and graphics and hands-on examples.

Price

The sale price of the product reflects what consumers are willing to pay for it. Marketing professionals need to consider costs related to research and development, manufacturing, marketing, and distribution—otherwise known as cost-based pricing. The survey will be conducted in different regions then we can come to the conclusion of pricing the courses that can be provided and the beta version of the app will be released to check how much customers are satisfied with the proposed APP.

Placement

The type of product created is important to consider when determining areas of distribution.
The APP would work best in metropolitan cities. However providing services to small cities and towns of the country would prove beneficial as there would be high demand.

Promotion

Joint marketing campaigns also are called a promotional mix. Activities might include advertising, sales promotion, personal selling, and public relations. A key consideration should be for the budget assigned to the marketing mix. Marketing professionals carefully construct a message that often incorporates details from the other three Ps when trying to reach their target audience. Determination of the best mediums to communicate the message and decisions about the frequency of the communication also are important. Customer service businesses often will take a consumer-centric approach that incorporates additional elements to address their unique needs.

Three additional Ps tied to this type of marketing mix might include people, process, and physical evidence. People refer to employees who represent a company as they interact with clients or customers. Process represents the method or flow of providing service to the clients and often incorporates monitoring service performance for customer satisfaction.

Potential Customer Base:

- Quality of Education:
- Customized learning for students from all walks: AI-enabled platforms can help in identifying knowledge gaps, understanding how students learn and tailoring online courses as per the most effective learning pathway for each student.
- Lifelong learning and professional development: India is expected to have a population of 140 million college-going people by 2030. A surge is expected in demand for learning new and advanced skills around digital, analytics, automation and cyber security from students and professionals across industries.
- Cost of Education: EdTech platforms have the potential to offer flexible learning models at different price points that are suited for a student’s socioeconomic background and specific needs.
- Contribution Economy: As per KPMG, online certification courses and test preparation will continue to account for a large proportion of the online education market in India which was estimated to touch the $1.96 billion mark in 2021.
- Employment Generation: "The sector witnessed larger demand in the backroom, especially for software developers, support function, teacher, content writer, marketing professionals. Experience bracket which is larger in demand is 3-5 years and 6-10 years," Yadav said.

These factors will most certainly attract customers and the potential customers can be identified from the identified target segment.

The code and data sets are here -

https://github.com/sahasi24/Ed-Tech_market_segmentation