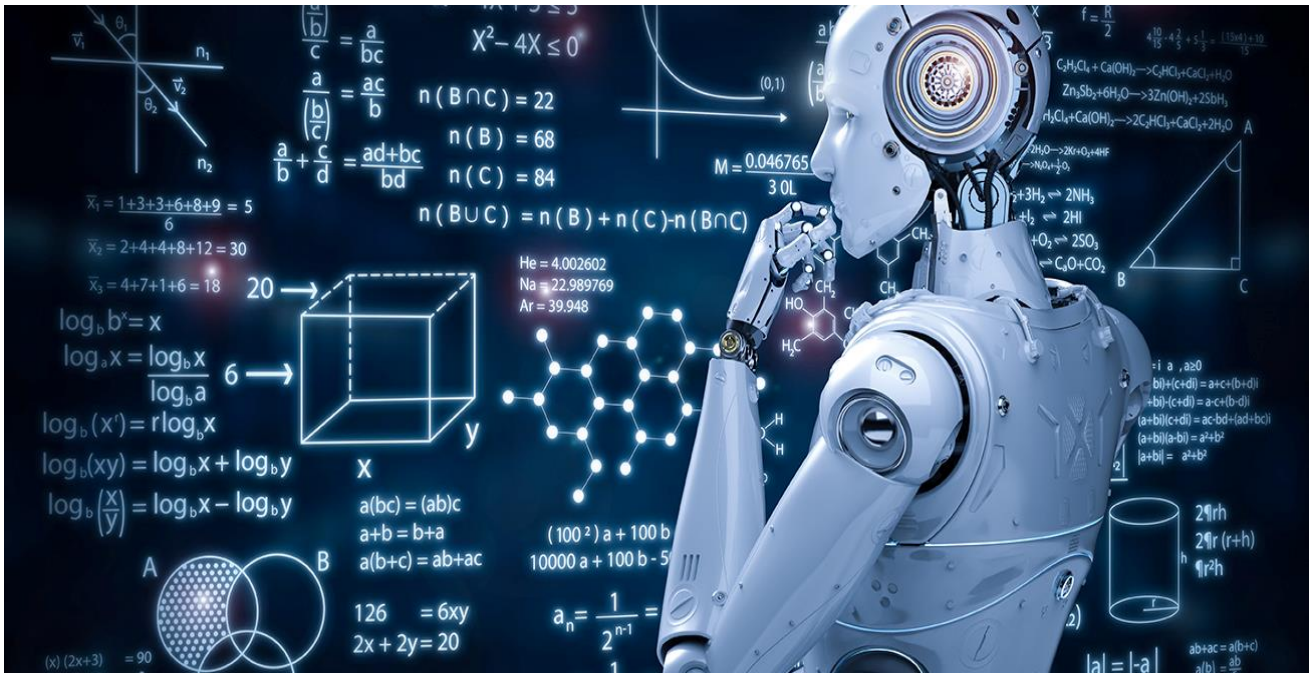
Market Segmentation Analysis of Machine Learning Job market in India



By

Team ‘E’

Santosh kumar

Sunku Sai Nisvas Sankarsh

Debang Mandal

Qamar Alam

Raunak Ghosh

Feynn Labs Machine Learning Internship

Dated-15/03/22

Table of Content

1. Problem Statement

Machine Learning is an Application of Artificial Intelligence (AI) it gives devices the ability to learn from their experiences and improve their self without doing any coding. For Example, when you shop from any website it's shows related search like:- People who bought also saw this.

Finding Companies most probable to hire an ML Engineer/Data Analyst Applicant in respect to his/her skillset.

=>Data Collection/Scraping based on

1. Geography,

2. Company's field of work,

3. Company size,

4. Upcoming vacancies in respect to company's growth (IPO/Funding etc.)

5. Machine Learning/Data Analysis Skills currently most demanded in the market in respect to i) Experience required, ii) Time required to acquire the skill, iii) Vacancies open iv) Salary etc.

We have to analyze Machine Learning Job Market in India with respect to the given problem statement using Segmentation analysis and outline the segments most optimal to apply or prepare for Machine Learning Jobs.

2. Data Sources

Data collection is defined as the procedure of collecting, measuring, and analysing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information. Regardless of the field of study or preference for defining data (quantitative or qualitative), accurate data collection is essential to maintain research integrity. The selection of appropriate data collection instruments (existing, modified, or newly developed) and delineated instructions for their correct use reduce the likelihood of errors. A formal data collection process is necessary as it ensures that the data gathered are both defined and accurate. This way, subsequent decisions based on arguments embodied in the findings are made using valid data. The process provides both a baseline from which to measure and in certain cases an indication of what to improve.

According to the problem statement, the dataset should mainly contain Job title, Experience, Skills, Location, etc. Different dataset sources are as follow: -

https://www.kaggle.com/lekuid/data-scientist-job-listings-in-india

https://www.kaggle.com/ankitkalauni/predict-the-data-scientists-salary-in-india

https://www.kaggle.com/andrewmvd/data-analyst-jobs

https://www.kaggle.com/halhuynh/it-jobs-dataset

3. Data Pre-processing

Data pre-processing is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. Using interactive dashboards and point-and-click data exploration, users can better understand the bigger picture and get to insights faster. This process makes deeper analysis easier because it can help target future searches and begin the process of excluding irrelevant data points and search paths that may turn up no results. More importantly, it helps build a familiarity with the existing information that makes finding better answers much simpler. Many times, data pre-processing uses visualization because it creates a more straightforward view of data sets than simply examining thousands of individual numbers or names. In any data pre-processing, the manual and automated aspects also look at different sides of the same coin. Manual analysis helps users familiarize themselves with information and can point to broad trends.

Exploring Machine Learning Jobs Dataset

```
In [12]: df.head()
```

Out[12]:

	Job Title	Company Name	Exp	Location	Skills
0	Data Engineer: Machine Learning	IBM	4-8 Yrs	Bangalore/Bengaluru	deep learning|Interpersonal skills|Time manage...
1	Data Engineer: Machine Learning	IBM	4-6 Yrs	Bengaluru/Bangalore	deep learning|Interpersonal skills|Time manage...
2	Manager - Machine Learning Engineer (Data Sci...	Pylon Management Consulting Pvt Ltd	7-9 Yrs	Remote	Data Science|Machine Learning|Deep Learning|IT...
3	Data Scientist-Python Machine Learning	Jubna	3-5 Yrs	Noida, NCR	mapping|ML algorithms|analyses data|machine le...
4	Senior/Lead Data Scientist - Machine Learning/...	Squareroot Consulting Pvt Ltd.	1-6 Yrs	Bangalore/Bengaluru	Visualization|Exploratory Testing|Machine Lear...

Checking for null values in the dataset

```
In [7]: # checking for null values
        df.isnull().sum()
```

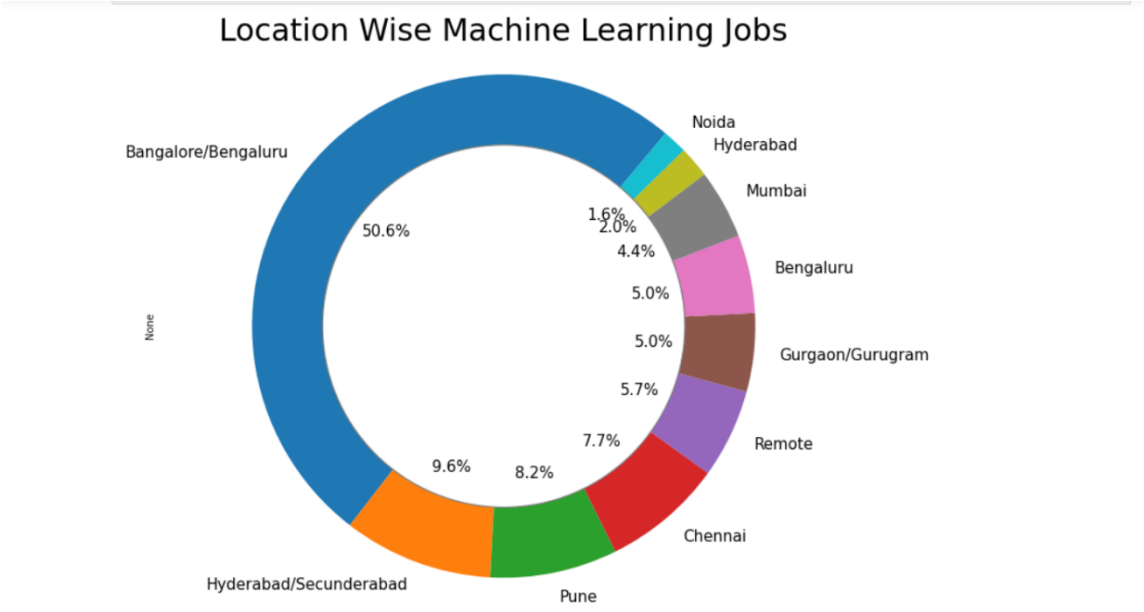
```
Out[7]: Job Title         0
        Job URL           0
        Company Name      0
        Company URL       0
        Exp               0
        Salary            0
        Location          0
        Skills            0
        Posted          368
        dtype: int64
```

Columns of the dataset

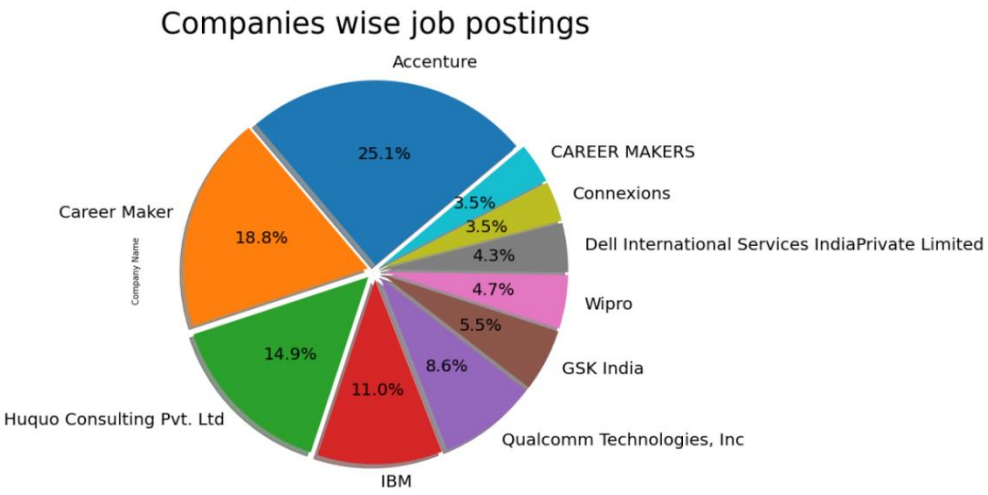
```
In [8]: # checking columns of our dataset
        df.columns
```

```
Out[8]: Index(['Job Title', 'Job URL', 'Company Name', 'Company URL', 'Exp', 'Salary',
               'Location', 'Skills', 'Posted'],
              dtype='object')
```

Location-wise Machine Learning Jobs

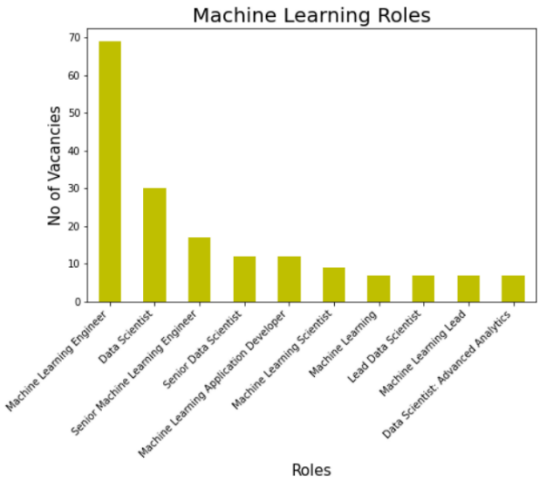


Company-wise Job Postings:

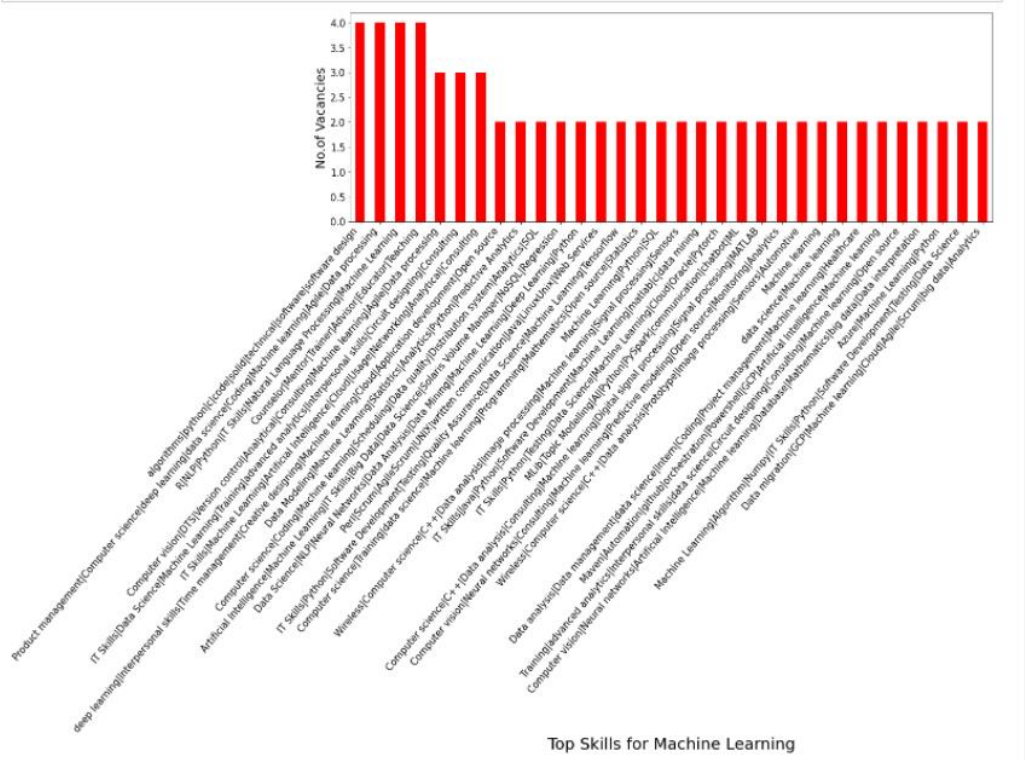


Experience wise No. of Vacancies:

Machine Learning Roles in the Market:



Skills-wise plot of Machine Learning Jobs in the Market:



Info of the dataset:

```
In [10]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 994 entries, 0 to 993
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Job Title     994 non-null    object
 1   Job URL       994 non-null    object
 2   Company Name  994 non-null    object
 3   Company URL   994 non-null    object
 4   Exp           994 non-null    object
 5   Salary        994 non-null    object
 6   Location      994 non-null    object
 7   Skills        994 non-null    object
 8   Posted        626 non-null    object
dtypes: object(9)
memory usage: 70.0+ KB
```

4 Segments Extraction

Methodologies that can be used:

1 Distance-Based Methods

The distance-based models sequester the sequence data into pairwise distances. This step loses some information, but sets up the platform for direct tree reconstruction. The two steps of this method are hereby discussed in detail.

Most of the time, data available for market segmentation is in the form of tables, where each column represents features about customers and rows represents customers. Numerous approaches to measuring the distance between two vectors exist, several are used routinely in cluster analysis and market segmentation.

1.1 Distance Measures

The most common distance measures used in market segmentation analysis are

a) Euclidean distance:

b) Manhattan or absolute distance: Both Euclidean and Manhattan distance treat all dimensions of the data equally. The way this kind of algorithms works is, they find the distance of each vector with all other vectors and then group those which are closer to each other. Usually, these algorithms are used when the data available is small.

1.2 Hierarchical Methods:

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations (consumers) into k groups (segments). Market segmentation analysis occurs between those two extremes. First method is Divisive hierarchical clustering methods start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment, Second method is Agglomerative hierarchical clustering approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters). Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment. Underlying both divisive and agglomerative clustering is a measure of distance between groups of observations (segments). This measure is determined by specifying

(1) a distance measure d(x, y) between observations (consumers) x and y, and **(2)** a linkage method. The linkage method generalises how, given a distance between pairs of observations, distances between groups of observations are obtained. Single linkage: distance between the two closest observations of the two sets. Complete linkage: distance between the two observations of the two sets that are farthest away from each other. Average linkage: mean distance between observations of the two sets. Clustering in general, and hierarchical clustering in specific, are exploratory techniques. Different combinations can reveal different features of the data. The result of hierarchical clustering is typically presented as a 9 dendrogram. A dendrogram is a tree diagram. The root of the tree represents the one-cluster solution where one market segment contains all consumers. The leaves of the tree are the single observations (consumers), and branches in between correspond to the hierarchy of market segments formed at each step of the procedure. The height o the branches corresponds to the distance between the clusters. Higher branches point to more distinct market segments. Dendrograms are often recommended as a guide to select the number of market segments. however, dendrograms rarely provide guidance of this nature because the data sets underlying the analysis are not well structured enough.

1.3 Partitioning Methods

Hierarchical clustering methods are particularly well suited for the analysis of small data sets with up to a few hundred observations. For larger data sets, dendrograms

are hard to read, and the matrix of pairwise distances usually does not fit into com-

puter memory.

This means that – instead of computing all distances between all pairs of observations in the data set at the beginning of a hierarchical partitioning,

cluster analysis using a standard implementation – only distances between each consumer in the data set and the center of the segments are computed.

A partitioning clustering algorithm aiming to extract five market segments, in contrast, would only have to calculate between 5 and 5000 distances at each step of the iterative or stepwise process.

1.3.1 k-Means and k-Centroid Clustering

The most popular partitioning method is k-means clustering. Let X = {x1,..., Xn} be a set of observations (consumers) in a data set. Partitioning clustering methods divide these consumers into subsets (market segments) such that consumers assigned to the same market segment are as similar to one another as possible, while consumers belonging to different market segments are as dissimilar as possible In addition, the algorithm requires the specification of the number of segments In fact, the choice of the distance measure typically has a bigger impact on the nature of the resulting market segmentation solution than the choice of algorithm

2.Model-Based Methods:

model-based methods have been proposed as an alternative of Distance-based methods, model-based segment extraction methods are based on the assumption that the true market segmentation solution –

which is unknown – has the following two general properties:

(1) each market segment has a certain size

(2) if a consumer belongs to market segment A, that consumer will have characteristics which are specific to members of market segment A. Model Based methods use the empirical data to find those values for segment sizes and segment-specific characteristics that best reflect the data

The values that need to be estimated are called parameters. Different statistical frameworks are available for estimating the parameters of the finite mixture model

Maximum likelihood estimation is commonly used. Maximum likelihood estimation aims at determining the parameter values for which the observed data is most likely to occur. An alternative statistical inference approach is to use the Bayesian framework for estimation If a Bayesian approach is pursued, mixture models are usually fitted using Markov chain Monte Carlo methods. consumers in the empirical data set can be assigned to segments using the following

approach. First, the probability of each consumer to be a member of each segment is determined. This is based on the information available for the consumer, which consists of y, the potentially available x, and the estimated parameter values of the finite mixture model: The consumers are then assigned to segments using these probabilities by selecting the segment with the highest probability. A standard strategy to select a good number of market segments is to extract finite mixture models with a varying number of segments and compare them. Selecting the correct number of segments is as problematic in model-based methods as it is to select the correct number of clusters when using partitioning methods. At first glance, finite mixture models may appear unnecessarily complicated. The advantage of using such models is that they can capture very complex segment characteristics, and can be extended in many different ways.

2.1 Finite Mixtures of Distributions:

The simplest case of model-based clustering has no independent variables x, and simply fits a

distribution to y. The statistical distribution function f () depends on the measurement level or scale of the segmentation variables y.

2.3 Extensions and Variations:

Finite mixture models are more complicated than distance-based methods. The additional complexity makes finite mixture models very flexible. It allows using any statistical model to describe a market segment. As a consequence, finite mixture models can accommodate a wide range of different data characteristics: for metric data we can use mixtures of normal distributions, for binary

data we can use mixtures of binary distributions. For nominal variables, we can use mixtures of multinomial distributions or multinomial logit models**.**

Ordinal variables are tricky because they are susceptible to containing response styles. To address this problem, we can use mixture models disentangling response style effects from content-specific

responses while extracting market segments. If the data set contains repeated observations over time, mixture models can cluster the time series, and extract groups of similar consumers. Alternatively, segments can be extracted on the basis of switching behavior of consumers between groups over time using Markov chains. This family of models is also referred to as dynamic latent change models, and can be used to track changes in brand choice and buying decisions over time. Mixture models also allow to simultaneously include segmentation and descriptor variables.

Segmentation variables are used for grouping, and are included in the segment-specific model as usual**.**

3. Algorithms with Integrated Variable Selection

Most algorithms focus only on extracting segments from data. These algorithms

assume that each of the segmentation variables makes a contribution to determining the segmentation solution. But this is not always the case. Sometimes, segmentation variables were not carefully selected, and contain redundant or noisy variables. Pre-processing methods can identify them.

When the segmentation variables are binary, and redundant or noisy variables

cannot be identified and removed during data pre-processing

3.1 Biclustering Algorithms

Biclustering simultaneously clusters both consumers and variables. Biclustering algorithms exist for

any kind of data, including metric and binary. Several popular biclustering algorithms exist; in particular they differ in how a bicluster is defined. In the simplest case, a bicluster is defined for 13 binary data as a set of observations with values of 1 for a subset of variables. Each row corresponds to a consumer, each column to a segmentation variable. Biclustering is particularly useful in market segmentation applications with many segmentation variables. Standard market segmentation techniques risk arriving at suboptimal groupings of consumers in such situations.

1. No data transformation. 2. Ability to capture niche markets. Biclustering methods, however, do not group all consumers. Rather, they select groups of similar consumers, and leave ungrouped consumers who do not fit into any of the groups.)

3.2 Variable Selection Procedure for Clustering Binary Data:

This method is based on the k-means algorithm as clustering method, and assumes that not all variables available are relevant to obtain a good clustering solution. In particular, the method assumes the presence of

masking variables. They need to be identified and removed from the set of segmentation variables.

3.3 Variable Reduction: Factor-Cluster Analysis

The term factor-cluster analysis refers to a two-step procedure of data-driven market segmentation analysis.

1. Factor analyzing data leads to a substantial loss of information.

2. Factors-cluster results are more difficult to interpret.

4 Data Structure Analysis:

Extracting market segments is inherently exploratory, irrespective of the extraction algorithm used. Validation in the traditional sense, where a clear optimality criterion is targeted, is therefore not possible. Ideally, validation would mean calculating different segmentation solutions, choosing different segments, targeting them, and then comparing which leads to the most profit, or most success in mission achievement. This is clearly not possible in reality because one organization cannot run multiple segmentation strategies simultaneously just for the sake of determining which performs best.

Data structure analysis provides valuable insights into the properties of the

data. These insights guide subsequent methodological decisions. Most importantly, stability-based data structure analysis provides an indication of whether natural, distinct, and well-separated market segments exist in the data or not. If they do, they can be revealed easily. If they do not, users and

data analysts need to explore a large number of alternative solutions to identify the most useful segment(s) for the organization.

data structure analysis can also help to choose a suitable number of segments

to extract.

4.1 Cluster Indices

Data analysts need guidance to make some of the most critical decisions, such as selecting the number of market segments to extract. cluster indices represent the most common approach to obtaining such guidance.

Generally, two groups of cluster indices are distinguished: internal cluster indices and external cluster indices.

Internal cluster indices are calculated on the basis of one single market seg-

mentation solution

example for an internal cluster index is the sum of all distances between pairs of segment members. The lower this number, the more similar members of the same segment are. Segments containing similar members are attractive to users.

External cluster indices cannot be computed on the basis of one single market

segmentation solution only. Rather, they require another segmentation as additional input. The external cluster index measures the similarity between two segmentation solutions. If the correct market segmentation is known, the correct assignment of members to segments serves as the additional

input. The correct segment memberships, however, are only known when artificially generated data is being segmented.

4.2 Gorge Plots:

A simple method to assess how well segments are separated, is to look at the distances of each consumer to all segment representatives. Similarity values can be visualized using gorge plots.

Each gorge plot contains histograms of the similarity values separately for each segment. The x-axis plots similarity values. The y-axis plots the frequency with which each similarity value occurs high

similarity values indicate that a consumer is very close to the centroid (the segment representative) of the market segment. Low similarity values indicate that the consumer is far away from the centroid.

If natural, well-separated market segments are present in the data, we expect the gorge plot to contain many very low and many very high values. This is why this plot is referred to as gorge plot Producing and inspecting a large number of gorge plots is a tedious process, and has the disadvantage of not accounting for randomness in the sample used. These disadvantages are overcome by stability analysis, which can be conducted at the global or segment level.

4.3 Global Stability Analysis:

Global stability analysis acknowledges that both the sample of consumers, and the algorithm used in data-driven segmentation introduce randomness into the analysis.

Bootstrapping generates a number of new data sets by drawing observations with replacement from the original data. These new data sets can then be used to compute replicate segmentation solutions for different numbers of segments.

4.4 Segment Level Stability Analysis:

Target market represents a group of individuals who have similar needs, perceptions and interests. They show inclination towards similar brands and respond equally to market fluctuations. Individuals who think on the same lines and have similar preferences form the target audience.

Target market includes individuals who have almost similar expectations from the organizations or marketers.

To select a target market, it is essential for the organizations to study the following factors:

- Understand the lifestyle of the consumers
- Age group of the individuals
- Income of the consumers
- Spending capacity of the consumers
- Education and Profession of the people
- Gender
- Mentality and thought process of the consumers
- Social Status
- Kind of environment individuals are exposed to

5. Profiling and describing potential segments

First identifying Key Market Segment Characteristics. The goal of the pro- filing phase is to learn more about the market categories that have been identified. When the extraction is in progress only when data-driven market segmentation is employed does this become possible. Profiling is now required. In common sense, the segment profiles are predefined segmentation. If age is used as a segmentation variable in commonsense, for example,

The resulting segments will very certainly be age groups as a result of seg- mentation. As a result, while you're using the picture is radically different in the event of data-driven segmentation: segmentation Users of the service may have decided to extract categories based on the consumer benefits they were looking for.

However, until the market categories are defined, the differentiating characteristics of the resulting market categories are unknown.

The information has been analyzed. The purpose of profiling is to find these market segment distinguishing traits in relation to segmentation variables. The process of profiling requires describing each individual. Market segment on its own and in comparison, to other market segments When pressed, when asked about their vacation plans, the majority of Austrian winter visitors indicate they are going skiing.

Skiing in the mountains. Alpine skiing may define a market sector, but it does not necessarily distinguish it.

from competitors in other markets

Visualizations for Segment Profiling Although utilizing visuals to visualize data is an important part of statistical data analysis, neither the highly simple nor the very detailed tabular representations often employed to visualize data are ideal.

In exploratory statistical analysis (such as cluster analysis), graphics are very useful because they reveal the complicated relationships between variables. Furthermore, in an age of enormous data, visualization provides a straight forward approach to track changes over time.

In the data-driven market segmentation process, visualizations are important for inspecting one or more segments in detail for each segmentation solution. The comprehension of segment profiles is made easier by statistical graphs. They also make evaluating the utility of a market segmentation approach more easier. When it comes to data segmentation, there are always a lot of different options. The choice of one of the various options is a crucial on

This task is made easier for the data analyst and user by using visualizations of solutions.

1. Identifying Defining Characteristics of Market Segments
2. Assessing Segment Separation

6. Selection of target segment

We all need to find the optimal solution for which company can take a group of students who have specialized skills that are required for the company

For this we use clustering, In Clustering we use k-means clustering

We will find the suitable K value for our K-means clustering where we will get the result that every company will get the required candidates that are good with required skills.

we have initialized the for loop for the iteration on a different value of k ranging from 1 to 10; since for loop in Python, exclude the outbound limit, so it is taken as 11 to include 10^{th} value.

we have got the number of clusters, so we can now train the model on the dataset. To train the model, we will use the below lines of code we will use k as 4, as we know there are 4 clusters that need to be formed. The code is given below:

kmeans=KMeans(n_clusters=4,init='k-means++',random_state=42)

y_kmeans=kmeans.fit_predict(X)

7. Customising the Marketing Mix

7.1 Implications for Marketing Mix Decisions

- Marketing was originally seen as a toolbox to assist in selling products, with marketers mixing the ingredients of the toolbox to achieve the best possible sales results.
- The most commonly marketing mix is understood as consisting of the *4Ps*: Product, Price, Promotion and Place.
- The segmentation process is frequently seen as part of what is referred to as the *segmentation-targeting-positioning* (STP) approach.
- The segmentation-targeting-positioning approach postulates a sequential process. The process starts with *market segmentation* (The extraction, profiling and description of segments), followed by

targeting (the assessment of segments and selection of a target segment), and finally *positioning* (the measures an organisation can take to ensure that their product is perceived as distinctly different from competing products, and in line with segment needs).

Fig. 11.1 How the target segment decision affects marketing mix development

- For reasons of simplicity, the traditional 4Ps model of the marketing mix including Product, Price, Place and Promotion serves as the basis of this discussion.
- The selection of one or more specific target segments may require the design of new, or the modification or re-branding of existing products (Product), changes to prices or discount structures (Price), the selection of suitable distribution channels (Place), and the development of new communication messages and promotion strategies that are attractive to the target segment (Promotion).

7.2 Product

- One of the key decisions an organisation needs to make when developing the product dimension of the marketing mix, is to specify the product in view of customer needs.
- Often this does not imply designing an entirely new product, but rather modifying an existing one.

- Other marketing mix decisions that fall under the product dimension are: naming the product, packaging it, offering or not offering warranties, and after sales support services.

7.3 Price

- Typical decisions an organisation needs to make when developing the price dimension of the marketing mix include setting the price for a product, and deciding on discounts to be offered.
- Figure 11.2 shows the expenditures of segment 3 members on the right, and those of all other consumers on the left.

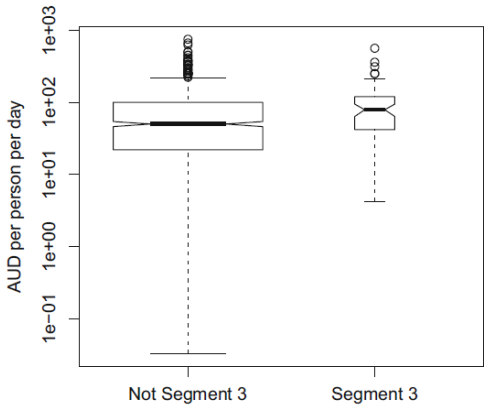


Fig. 11.2 Total expenditures in Australian dollars (AUD) for the last domestic holiday for tourists in segment 3 and all other tourists

- But the information contained in Fig. 11.2 is still valuable. It illustrates how the price dimension can be used to best possibly harvest the targeted marketing approach.

7.4 Place

- The key decision relating to the place dimension of the marketing mix is how to distribute the product to the customers.
- This includes answering questions such as: should the product be made available for purchase online or offline only or both; should the manufacturer sell directly to customers; or should a wholesaler or a retailer or both be used.

7.5 Promotion

- Typical promotion decisions that need to be made when designing a marketing mix include: developing an advertising message that will resonate with the target market, and identifying the most effective way of communicating this message.
- Other tools in the promotion category of the marketing mix include public relations, personal selling, and sponsorship.

8. Link to GitHub profile with codes and dataset well documented.

Dataset and Code Implementation