Market Segmentation

Analysing the respective market in India using Segmentation analysis for Online

Travel Hostel Chain Start Up *by*

Team 3

Shubham Urmaliya

Overview

The Indian outbound tourist market has evolved considerably in recent years. Rising incomes, multiple income households, exposure to international lifestyles, easier financial credits and an upbeat economy have enhanced the travel aspirations and consumption of the Indian consumer. With 15% growth recorded in this market over the last few years, many tourism boards have opened up offices in India to capture a share of this lucrative market.

Understanding visitors and the factors that pull them to visit a destination is imperative for successful branding as visitor characteristics and salient attributes are used to differentiate and position. This study identifies important attributes and segments of international tourists visiting India

Historically the tourism industry has largely used demographics including age, nationality, income, gender, and country of origin to define segments.
In addition to demographics, however, travel characteristics such as purpose of visit, length of stay, and visitation levels have also been used as grounds for segmentation.

What is Market Segmentation?

Market Segmentation

Market segmentation is nothing but dividing the total consumer market into groups to be able to communicate with them and provide their specific needs.

Smith (1956) introduced the concept of market segmentation as a strategic tool. He stated that “Market segmentation (…) can be viewed as a heterogeneous market (one characterized by divergent demand) as a number of smaller homogeneous markets”.

Why Segment the Tourism Market?

Every tourist being different, the tourism industry possibly is not capable of satisfying every individual’s need. This is the foundation of segmenting the total market.

While all tourists are different, some of them are similar to each other. Marketing force of a tourism business group the tourists into various segments that categorize the similar as well as distinct members. Market segmentation can be applicable to any of the tourism supply components and provides benefits as given below −

It helps to understand specific demands of the consumers.

It helps to allocate marketing expenses efficiently.

It helps to create effective marketing strategies to target specific market segment.

Tourism Market Segmentation

The tourism market segmentation can be broadly divided into the following types −

Geographic

Geographic market segmentation is done considering the factors such as tourists' place of origin. This factor is important as the tourists belonging to different places are brought up with different cultures and show different traits of behavior. It is the most basic type of segmentation.

Demographic

This segmentation is done by considering the tourist's gender, age, marital status, ethnicity, occupation, religion, income, education, and family members.

Psychographic

The marketing people do this segmentation by taking into account the psyche of the tourists. They gather information about the tourists' interests, attitudes, their way of living life, opinions, and overall personality.

Classes of Tourists

Depending upon the motives and the way of touring, there are various classes of tourists −

Tourists Travelling with Families

The tourists who visit places with their first and extended families, or families of relatives. One person, generally the head of the tourist family is the decision maker. The families generally travel for holidays and leisure and tend to expend sparingly.

Single Tourists

They travel alone and are independent. They are alone but not lonely; as tourism is what they pursue as a hobby. The gap year travelers, unmarried persons, widows/widowers, backpackers, and solitary tourists travel single. They decide for themselves and tend to expend more. They tend to carry less stuff on the journey. They tend to behave balanced if any challenging situation occurs and are rational towards tour schedules.

Groups of Tourists

Students from schools and universities as members of educational tours, fellows of various fraternities with common interests, groups of newly-weds, or senior citizens.

Tourists Visiting Friends and Relatives

These tourists travel to meet friends or relatives, or to attend a celebration or gatheringl. These tourists generally plan their tours in the breaks such as Diwali holidays, Christmas holidays, or any kind of long break when most of the people have break.

Business Tourists

They are the professional tourists on the business trips. They decide for themselves but do not spend much money. For example, a sales or a marketing person travels to another city to attend a business fair, and business manager travels to another country for business deals.

Incentive Tourists

They tour for consuming the reward they received in the form of a few days' family holiday package at some hotel or resort. Such rewards are generally distributed if an employee performs outstanding to achieve the goals.

Health Tourists

These tourists travels to places with the agenda of health on their mind. They travel to avail some special medical treatment, operation, surgery, medication, or inexpensive aesthetic surgeries available in different country. Some tourists in this category also travel if they are receiving some illness from the climate at their residence such as Asthma.

Place

The place is where the tourists visit and stay. The potential of a tourist destination lies in its attractiveness or aesthetic value, accessibility, and the facilities it provides to the tourists. The tourists also seek a place highly for the activities it offers, the amenities and skilled workforce it provides, and its location

Analysis and Approaches used for Segmentation

Clustering

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.
Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples.

K Means Algorithm

K Means algorithm is an iterative algorithm that tries to partition the dataset into pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

The approach k-means follows to solve the problem is **expectation maximization** The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster.

Below is a break down of how we can solve it mathematically

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

And M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}}$$

Applications

K means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

1. Get a meaningful intuition of the structure of the data we're dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups.

Implementation

In the Beginning We've analysed Domestic & Foreign Tourist Arrivals in 2016-17 & 17-18 in a structured manner.

1. **First and foremost, we Extracted State with most Tourist Arrival.**
2. **We Extracted Most Visited Cities or Circles of Extracted States.**
3. **We calculated the Distance of the Monuments from Airport and Railway Station.**
4. **We Analysed the GoIbibo Hotel Dataset and Extracted the relevant Hotels in particular State or Circle.**

Data Sources

We have gathered some datasets from Kaggle and OGD (Open Government Data) which are somehow related to the case.

Packages/ Tools used:

1. Numpy : To calculate various calculations related to arrays.
2. Pandas : To read or load the datasets.

We have considered two datasets for the analysis.

1. Tourism Dataset of 2016-18 (Foreign & Domestic)
2. Goibibo Hotel Dataset

Exploratory Data Analysis

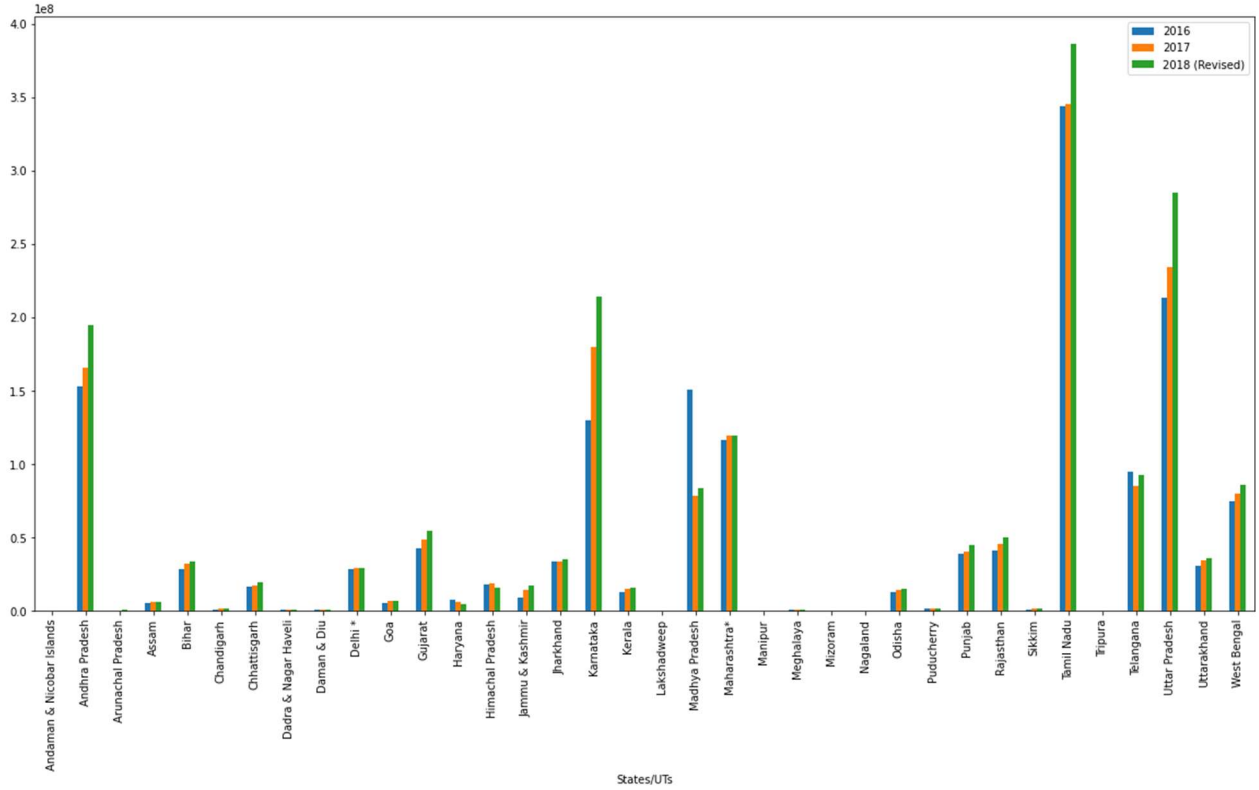
Behavioural Analysis

Behavioural segmentation is possibly the most useful of all for e-commerce businesses. As with psychographic segmentation, it requires a little data to be truly effective – but much of this can be gathered via your website itself. Here we group customers with regards to their:

- Spending habits
- Purchasing habits
- Browsing habits
- Interactions with the brand
- Loyalty to brand
- Previous product feedback

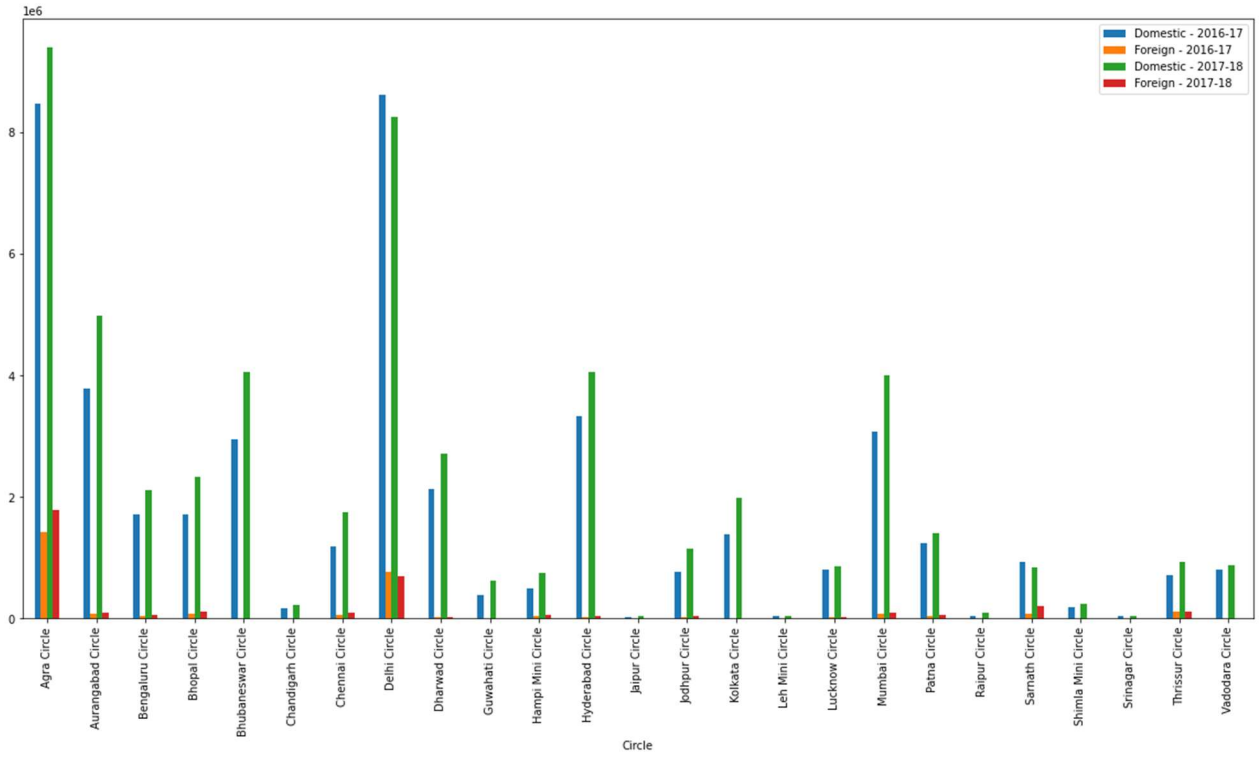
We've Plotted Multiple Bar Graph of No of Visitors vs States In 2016-18.

State Wise Tourism Distribution (Fig 1.1)

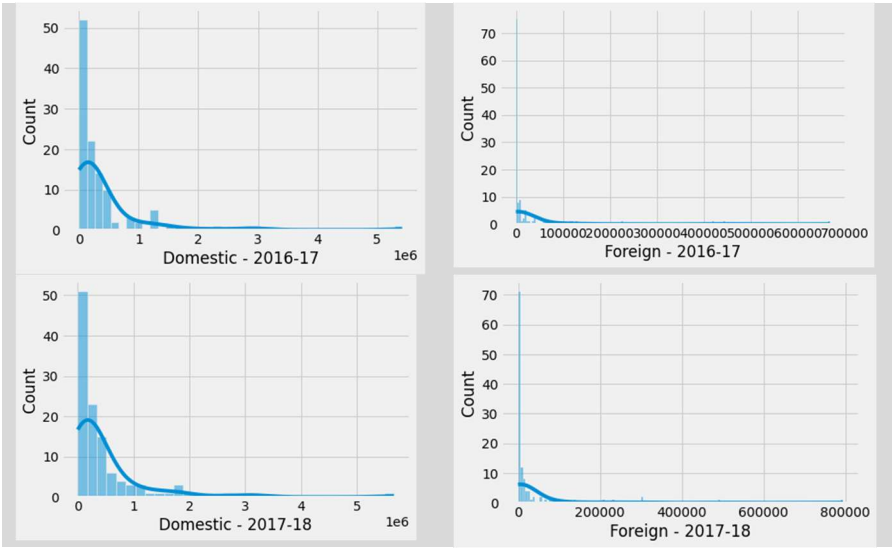


As Shown in (Fig 1.1) Maximum No of Tourist Arrival is in Tamil Nadu, Uttar Pradesh, Karnataka & Maharashtra.

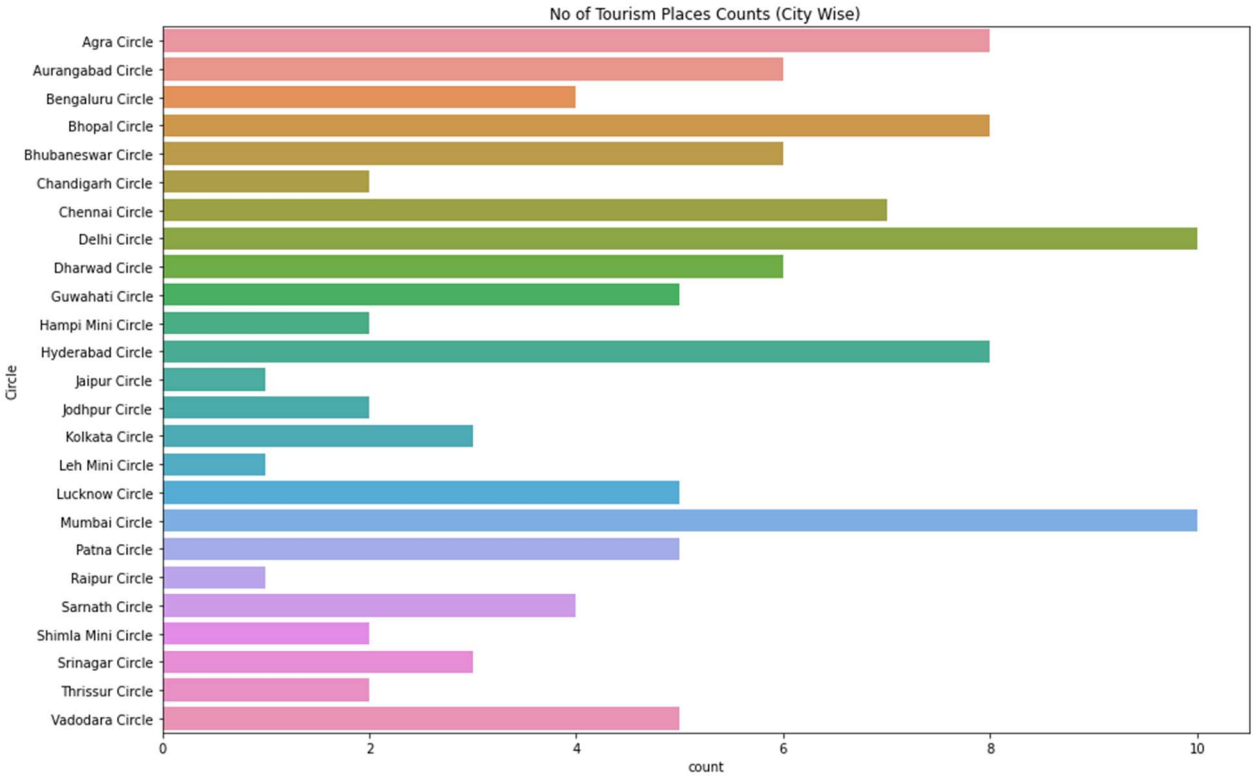
Circle Wise Tourism Distribution (Fig 1.2)



As Shown in (Fig 1.2) Maximum No of Tourist Arrival is in Agra, Delhi, Hyderabad, Mumbai, Aurangabad, Bhubaneshwar, Dharwad, Bhopal, Bangalore, Kolkata, Chennai, Patna.



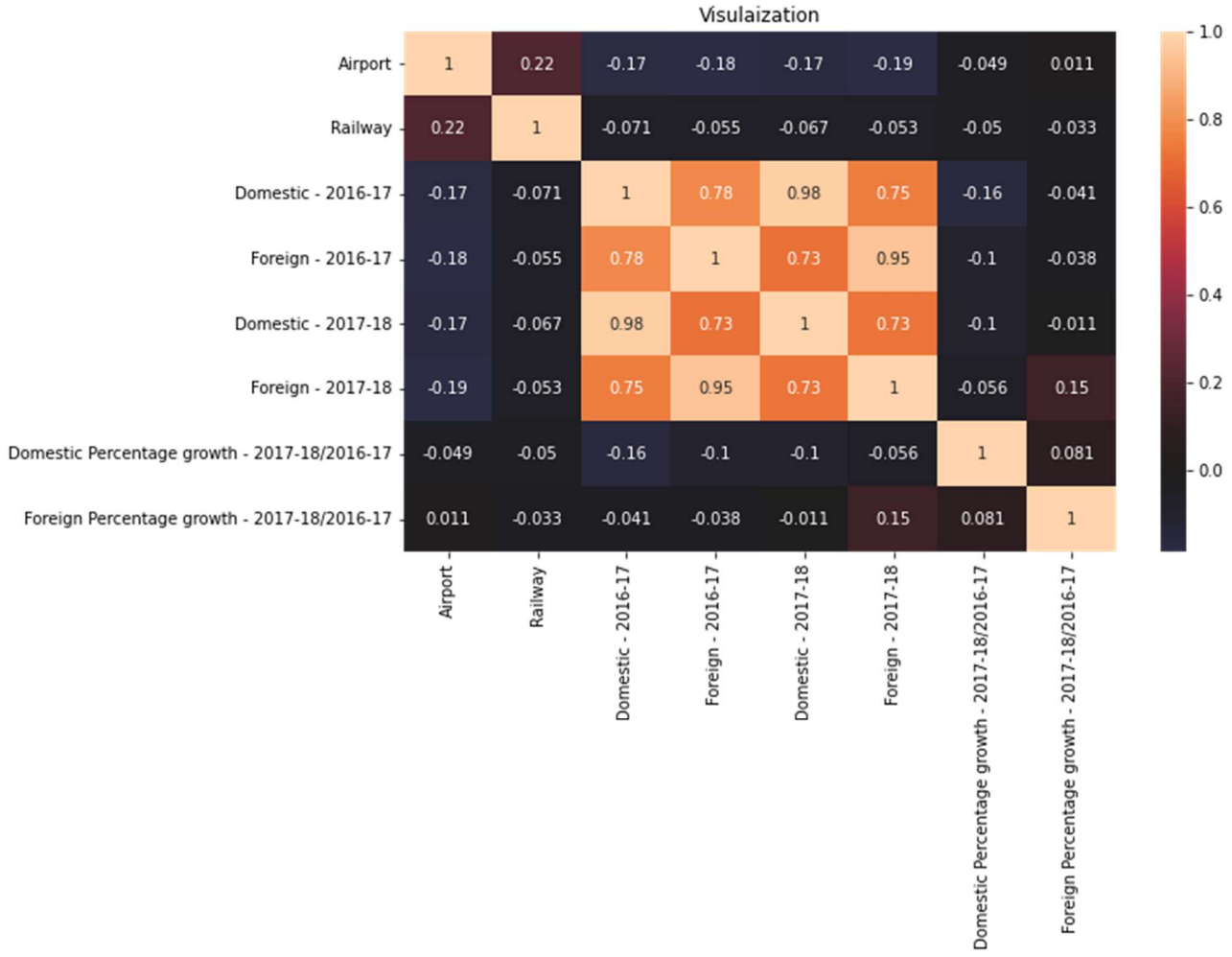
Tourist Places Count Circle Wise (Fig 1.3)



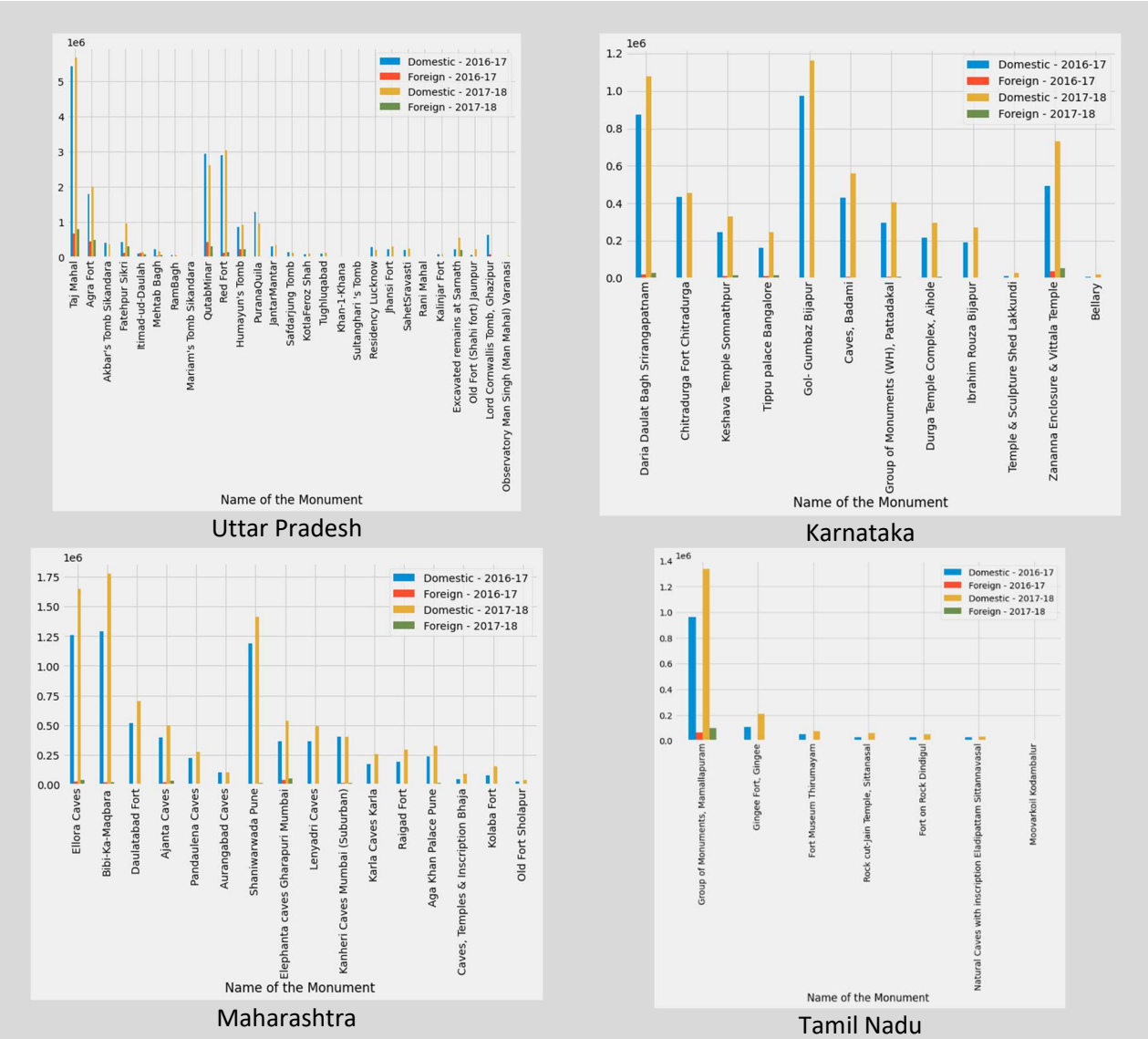
As Shown in (Fig 1.3) Maximum No of Tourist Places is in Agra, Delhi, Hyderabad, Mumbai, Aurangabad, Bhubaneshwar, Dharwad, Bhopal, Bangalore, Kolkata, Chennai, Patna.

Correlation between Features (Fig 1.4)

```
fig, ax = plt.subplots(figsize=(10,6))
sns.heatmap(df.corr(), center=0, annot=True )
ax.set_title('Visulaization')
```



Tourist Place Wise Tourism Distribution (Fig 1.4)



In (Fig 1.4) We accessed the tourist places that are Visited More by people in particular State which we have extracted in (Fig 1.1)

Geographic Analysis

Geographic market segmentation is done considering the factors such as tourists' place of origin. This factor is important as the tourists belonging to different places are brought up with different cultures and show different traits of behaviour. It is the most basic type of segmentation.

Here Weve taken GoIbibo Dataset and Extracted necessary features.

```
df
```

	city	country	property_name	state
0	Manali	India	Baragarh Regency	Himachal Pradesh
1	Gurgaon	India	Asian Suites A- 585	Haryana
2	Goa	India	Bevvan Resort	Goa
3	Manali	India	Apple Inn Cottage	Himachal Pradesh
4	Delhi	India	Anmol Hotel Pvt.Ltd	Delhi
...	...	...	...	...
3995	Ujjain	India	Hotel Shreenath Palace	Madhya Pradesh
3996	Ahmedabad	India	Hotel Sarvottam	Gujarat
3997	Mumbai	India	Hotel Silver Inn	Maharashtra
3998	Deoghar	India	Hotel Shivam International	Jharkhand
3999	Sambalpur	India	Hotel Silver Moon	Orissa

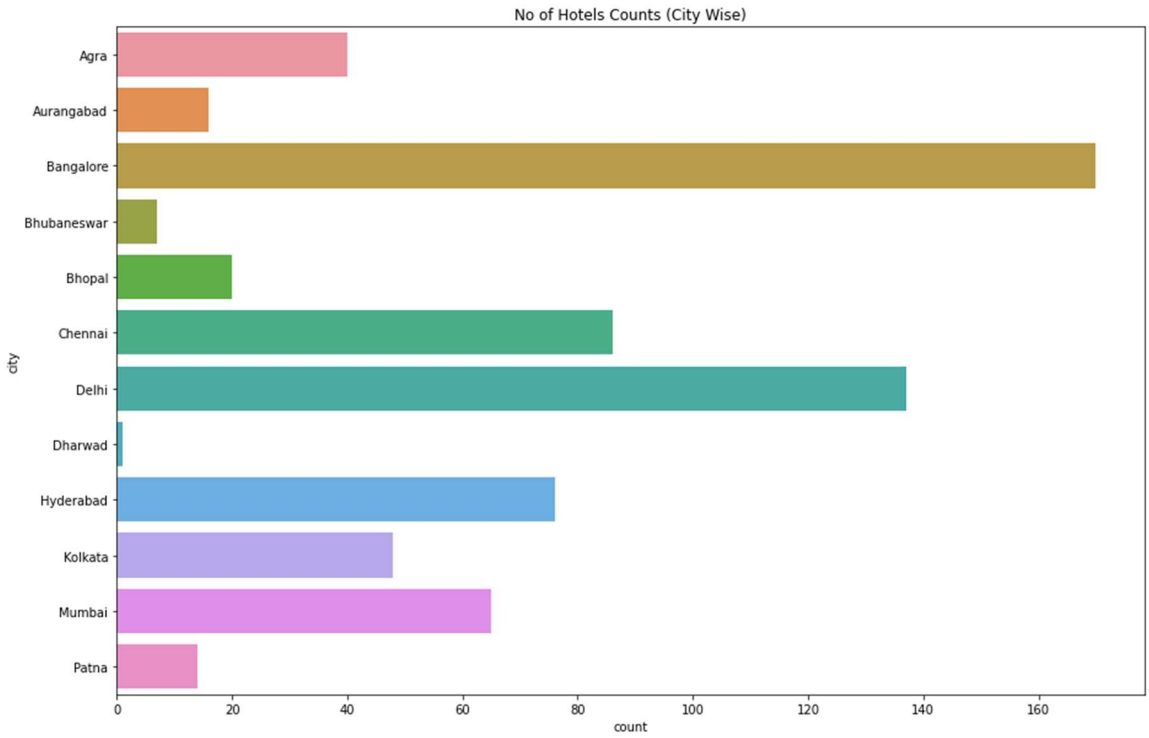
4000 rows × 4 columns

We have already extracted the cities which are most visited by tourists, So rather than calculating the hotel count of all the circles we used only extracted circles to find hotel count.

```
hotels_count = hotels_df.groupby('city')[['value_count']].sum().reset_index()
hotels_count
```

	city	value_count
0	Agra	40
1	Aurangabad	16
2	Bangalore	170
3	Bhopal	20
4	Bhubaneswar	7
5	Chennai	86
6	Delhi	137
7	Dharwad	1
8	Hyderabad	76
9	Kolkata	48
10	Mumbai	65
11	Patna	14

Hotel Count Distribution (Fig 2.1)



So by getting the insight of plot (Fig 2.1) we can distinguish range of hotels according to particular Circles.

Psychographic Analysis

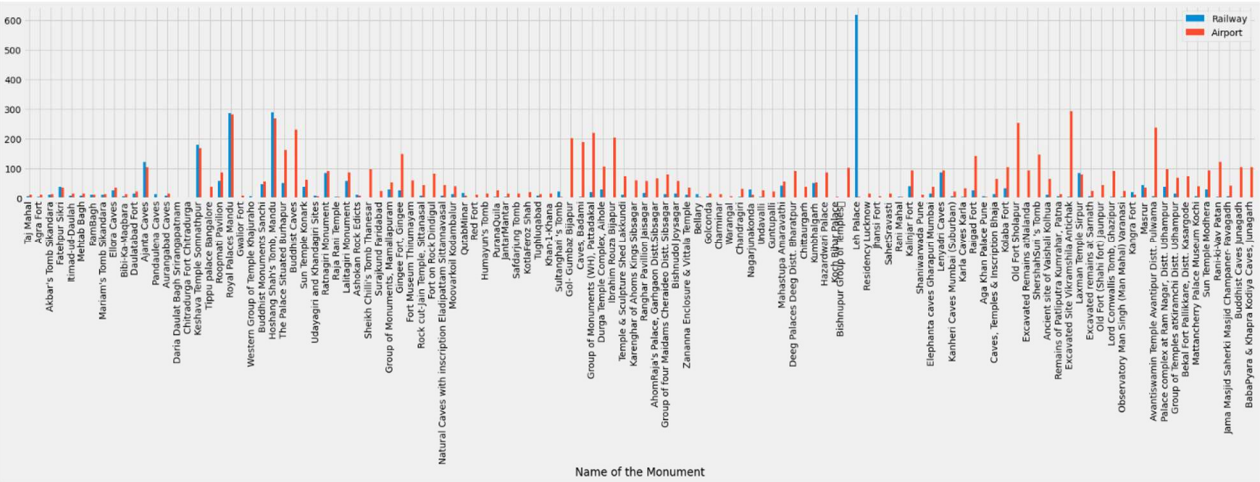
Psychographic segmentation has been done by dividing consumers into subgroups based on shared psychological characteristics, including subconscious or conscious beliefs, motivation, and priorities to explain and predict consumer behaviour.

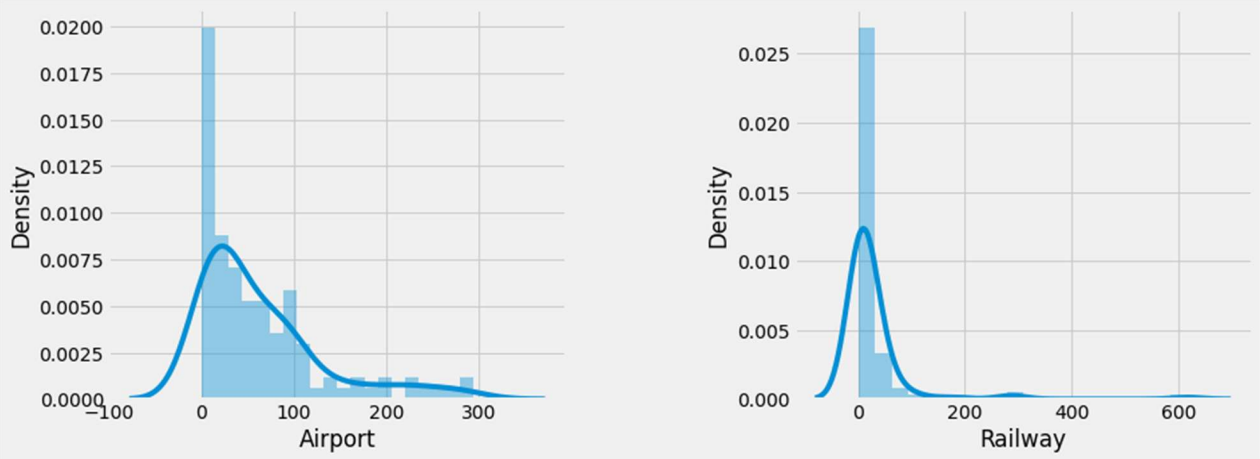
Railway & Airport Distribution (Fig 3.1)

```
df.plot(x="Name of the Monument", y=['Railway','Airport'], kind="bar",figsize=(30,5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7621619810>
```

```
/usr/local/lib/python3.7/dist-packages/matplotlib/backends/backend_agg.py:214: RuntimeWarning: Glyph 13 missing from current font.
  font.set_text(s, 0.0, flags=flags)
/usr/local/lib/python3.7/dist-packages/matplotlib/backends/backend_agg.py:183: RuntimeWarning: Glyph 13 missing from current font.
  font.set_text(s, 0, flags=flags)
```





When looking at Airport and Railway features (Fig 3.1), we can infer that either the place of visit is very near to the structure i.e., less than a kilometre, or the place of visit like “LEH circle” has no station in the near range of 500km

Segmentation

Using K Means

At present, the application of cluster analysis in transformer data processing can
improve the processing of unbalanced data sets, improve the evaluation level of transformer state, and more accurately reflect the real operation state of transformer.
We have tried to improve the processing effect with Power Transformer on unbalanced data sets.

Power Transformer (Fig 4.1)

```
from sklearn.preprocessing import StandardScaler,MinMaxScaler,PowerTransformer,RobustScaler

pt=PowerTransformer()


features = ["Domestic - 2016-17", "Foreign - 2016-17"]

X_subset = df[features]
scaler = PowerTransformer().fit(X_subset)
X = scaler.transform(X_subset)
pd.DataFrame(X, columns=X_subset.columns).describe()
```

Segmentation of 2016-17 Data

Next step is to select the right number of clusters.

Elbow Point Method

The maximum possible number of clusters will be equal to the number of observations in the dataset.

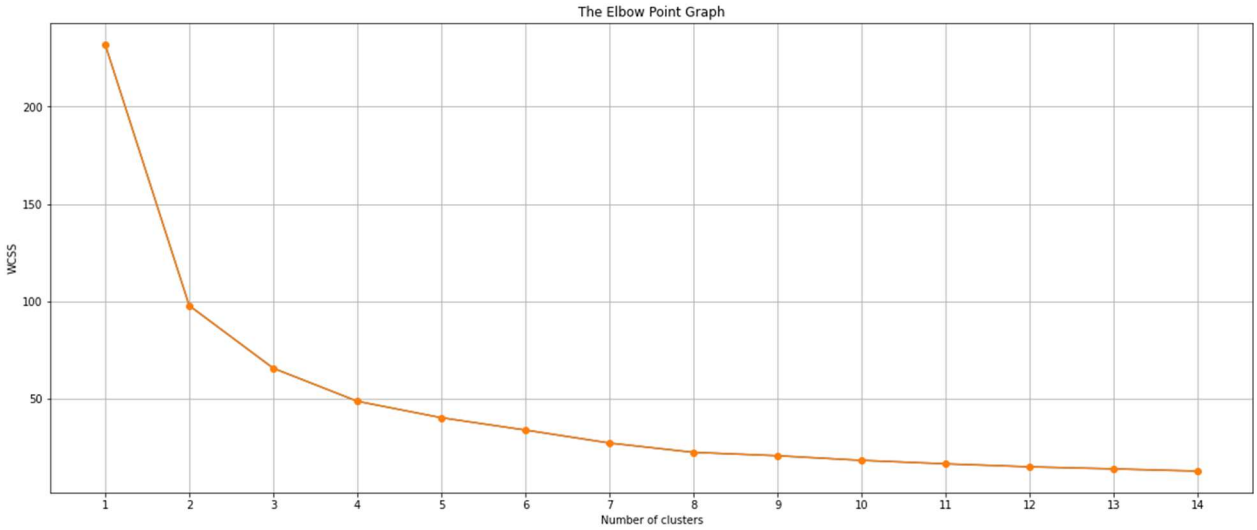
But then how can we decide the optimum number of clusters? One thing we can do is plot a graph, also known as an elbow curve, where the x-axis will represent the number of clusters and the y-axis will be an evaluation metric.

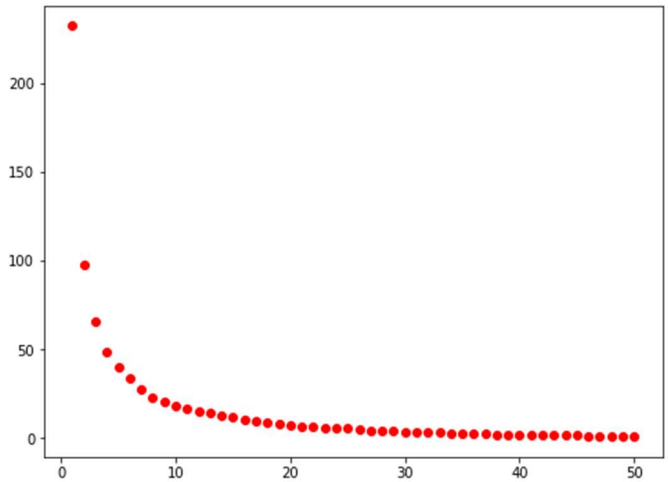
the cluster value where this decrease in inertia value becomes constant can be chosen as the right cluster value for our data.

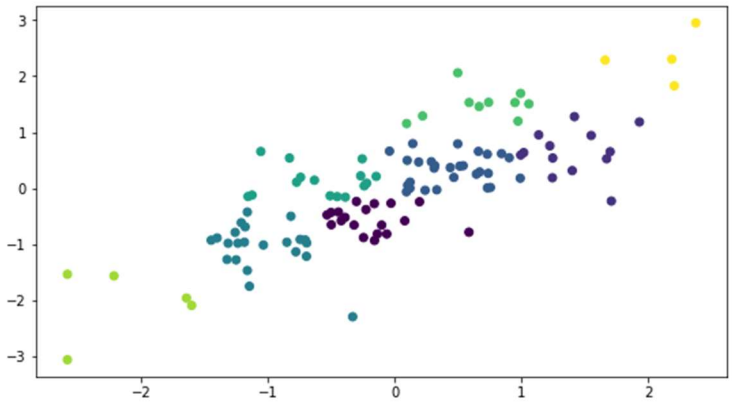
```
from sklearn.cluster import KMeans
wcss = []  # Within-Cluster-Sum-of-Squares
for i in range(1, 15):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=10)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(20, 8))
plt.plot(range(1, 15), wcss)
plt.title('The Elbow Point Graph')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.plot(range(1, 15), wcss, "-o")
plt.xticks(range(1, 15))
plt.grid(True)
plt.show()
```

Elbow Point Method (Fig 4.2)







We can't find optimum number of clusters with above plot so we performed Silhouette Analysis to find Silhouette coefficient for clusters.

Silhouette Analysis

Let's start with the equation for calculating the silhouette coefficient for a particular data point:

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}$$

where,

- s(o) is the silhouette coefficient of the data point o

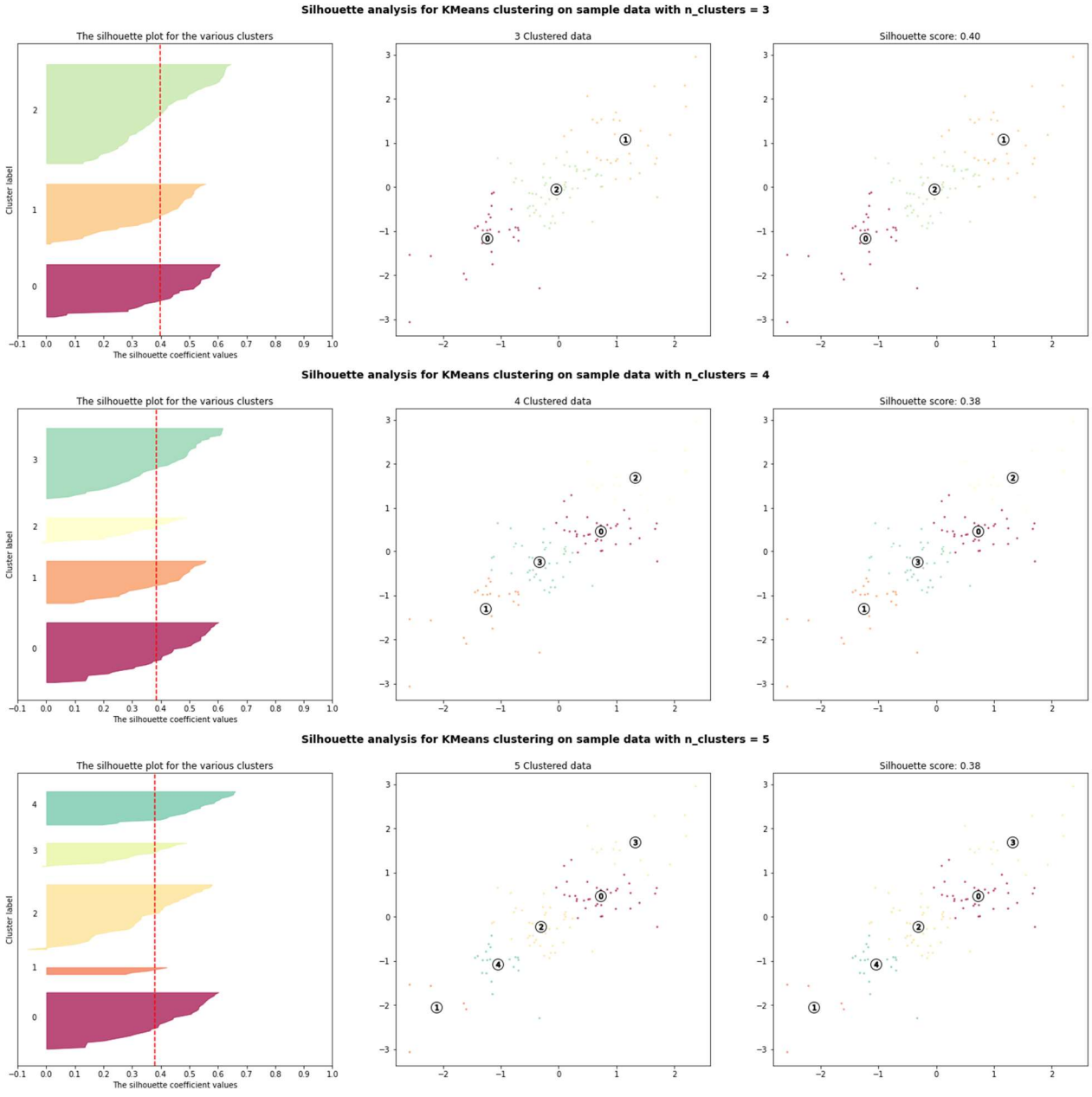
- a(o) is the average distance between o and all the other data points in the cluster to which o belongs

- b(o) is the minimum average distance from o to all clusters to which o does not belong

```
   for 3 clusters the silhouette score is 0.399
            Centers of each cluster:
      Domestic - 2016-17  Foreign - 2016-17
   0        15987.772147          50.174217
   1       765044.575855       31723.741933
   2       148444.343790        1371.398608
-------------------------------------------------
   for 5 clusters the silhouette score is 0.380
            Centers of each cluster:
      Domestic - 2016-17  Foreign - 2016-17
   0        3.325696e+05       10530.724564
   1        1.024494e+03           2.916819
   2        1.089550e+05         657.777064
   3        1.539438e+06       48382.004520
   4        1.956705e+04          84.687637
-------------------------------------------------
```

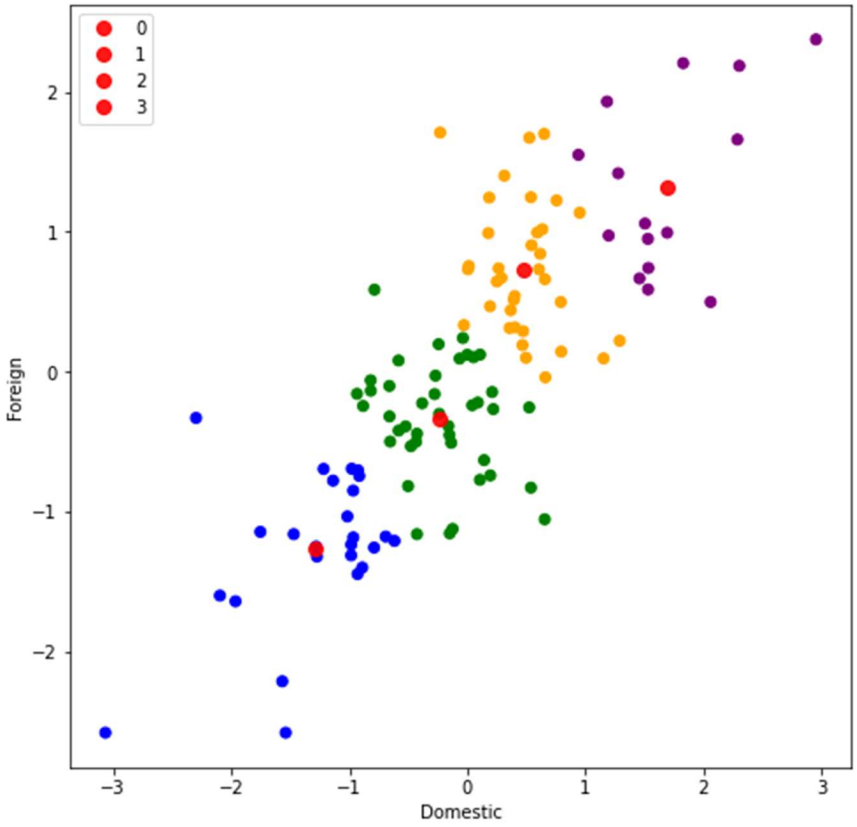
The value of the silhouette coefficient is between [-1, 1]. A score of 1 denotes the best meaning that the data point o is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.

Silhouette Analysis (Fig 4.3)



Below is the segmentation plot for above found clusters.

Segmentation Plot of 4 Clusters (Fig 4.4)



The Red encoded circles are centroids of their particular segment.

Cluster 0 –

Circle_Bhopal Circle	0.142857
Circle_Agra Circle	0.114286
Circle_Dharwad Circle	0.114286
Circle_Bengaluru Circle	0.085714
Circle_Mumbai Circle	0.085714
Circle_Lucknow Circle	0.057143
Circle_Aurangabad Circle	0.057143
Circle_Jodhpur Circle	0.057143
Circle_Vadodara Circle	0.057143
Circle_Patna Circle	0.057143

Cluster 1 –

Circle_Chennai Circle	0.16
Circle_Srinagar Circle	0.12
Circle_Hyderabad Circle	0.08
Circle_Bhubaneswar Circle	0.08
Circle_Delhi Circle	0.08
Circle_Guwahati Circle	0.08
Circle_Vadodara Circle	0.04
Circle_Bhopal Circle	0.04
Circle_Chandigarh Circle	0.04
Circle_Dharwad Circle	0.04

Cluster 2 –

Circle_Delhi Circle	0.266667
Circle_Agra Circle	0.133333
Circle_Aurangabad Circle	0.133333
Circle_Hyderabad Circle	0.133333
Circle_Bengaluru Circle	0.066667
Circle_Sarnath Circle	0.066667
Circle_Bhubaneswar Circle	0.066667
Circle_Mumbai Circle	0.066667
Circle_Chennai Circle	0.066667
Circle_Hampi Mini Circle	0.000000

Cluster 3 –

Circle_Mumbai Circle	0.121951
Circle_Hyderabad Circle	0.097561
Circle_Guwahati Circle	0.073171
Circle_Patna Circle	0.073171
Circle_Delhi Circle	0.073171
Circle_Vadodara Circle	0.048780
Circle_Aurangabad Circle	0.048780
Circle_Bhopal Circle	0.048780
Circle_Sarnath Circle	0.048780
Circle_Bhubaneswar Circle	0.048780

Segmentation of 2017-18 Data

Repeating same steps as performed above in 2017-18 data we evaluated –

Cluster 0 -

```
Circle_Bhopal Circle        0.142857
Circle_Agra Circle          0.114286
Circle_Dharwad Circle       0.114286
Circle_Bengaluru Circle     0.085714
Circle_Mumbai Circle        0.085714
Circle_Lucknow Circle       0.057143
Circle_Aurangabad Circle    0.057143
Circle_Jodhpur Circle       0.057143
Circle_Vadodara Circle      0.057143
Circle_Patna Circle         0.057143
```

Cluster 1 -

```
Circle_Chennai Circle        0.16
Circle_Srinagar Circle       0.12
Circle_Hyderabad Circle      0.08
Circle_Bhubaneswar Circle    0.08
Circle_Delhi Circle          0.08
Circle_Guwahati Circle       0.08
Circle_Vadodara Circle       0.04
Circle_Bhopal Circle         0.04
Circle_Chandigarh Circle     0.04
Circle_Dharwad Circle        0.04
```

Cluster 2 -

```
Circle_Delhi Circle          0.266667
Circle_Agra Circle           0.133333
Circle_Aurangabad Circle     0.133333
Circle_Hyderabad Circle      0.133333
Circle_Bengaluru Circle      0.066667
Circle_Sarnath Circle        0.066667
Circle_Bhubaneswar Circle    0.066667
Circle_Mumbai Circle         0.066667
Circle_Chennai Circle        0.066667
Circle_Hampi Mini Circle     0.000000
```

Cluster 3 -

```
Circle_Mumbai Circle         0.121951
Circle_Hyderabad Circle      0.097561
Circle_Guwahati Circle       0.073171
Circle_Patna Circle          0.073171
Circle_Delhi Circle          0.073171
Circle_Vadodara Circle       0.048780
Circle_Aurangabad Circle     0.048780
Circle_Bhopal Circle         0.048780
Circle_Sarnath Circle        0.048780
Circle_Bhubaneswar Circle    0.048780
dtype: float64
```

Result

In above analysis we Focused on

1. High tourism/most visited places in India
2. High concentration of hotels
3. Low concentration of hotels
4. Estimation of Railway and Airport Distance to determine shared hostel's locations based on previous/recent census data.

And we got these following Results

1. Maximum No of Tourist Arrival –
   1.1. In State is Tamil Nadu, Uttar Pradesh, Karnataka & Maharashtra.
   1.2. In Circle is Agra, Delhi, Hyderabad, Mumbai, Aurangabad, Bhubaneshwar, Dharwad, Bhopal, Bangalore, Kolkata, Chennai, Patna.
2. High Concentration of Hotels are in Bangalore, Delhi, Chennai, Hyderabad, Mumbai.
3. Low Concentration of Hotels are in Agra, Dharwad, Patna, Bhubaneshwar, Bhopal, Aurangabad, Kolkata.
4. As shown in Fig (3.1) the place of visit like "LEH circle" has no station in the near range of 500km so we can't open a shared hostel in particular location.

From Cluster analysis and above results we concluded that these are the optimum location to open a shared hostel chain-

- ✓ Dharwad
- ✓ Patna
- ✓ Bhubaneshwar
- ✓ Bhopal
- ✓ Agra

Marketing Mix

Setting prices for our products is both an art and a science. Most importantly, you must know and understand your cost of production. From there you can adjust based on product characteristics, a specific pricing strategy, customer price sensitivity, customer values, and other factors. Price contributes to the perception of your product, that is, when consumers see a product price it sends signals to them about quality, match with the market outlet, expectations for assistance, etc. Keeping accurate and complete records accounting for all steps – production, packaging, storage, promotion, transportation/distribution, and sales – will assist you in setting a price and making adjustments as necessary.

4Ps of Marketing Mix

The 4Ps helps companies to review and define key issues that affect the marketing of its products and services and is often now referred to as the 7Ps framework for the digital marketing mix.

Marketing as a whole relies on all seven Ps.

It is essential to consider them as a whole, and not in isolation. Customers must experience a coherent view of your company and your product, and that can only come from viewing the customer experience from end-to-end across all seven Ps.

Importance of Marketing Mix

It helps understand what our product or service can offer to our customers and helps plan a successful product offering. Helps with planning, developing and executing effective marketing strategies. Help determine whether your product or service is suitable for your customers.

Product: The product i.e., hostel chains will definitely sustain in the market, in compliance with the accommodation services provided to the customers.

Price:- As we have accommodation services, the prices may vary according to the demands, as well as the number of availability of tourists.

Place:- Through the analysis, we can see that Agra, Bhopal, and Dharwad are the best cities for starting a hostel chain, among all the other cities.

Promotion: Promotion can be based upon the analysis. More offers and promotions can be given to the segments that are more valuable to the company.

Codes

All the codes used in this project can be found on
tripathishubham1/Travel-Hostel-Startup---Market-Segmentation (github.com)

References

Datasets that have been used in this project are taken from

Kaggle –
Indian Hotels on Goibibo | Kaggle

OGD –
https://data.gov.in/resources/domestic-tourists-visits-india-2017-2019-ministry-tourism

Tutorials Point –
Market Segmentation (tutorialspoint.com)

Analytics Vidya –
https://medium.com/@cmukesh8688/k-means-clustering-in-machine-learning-252130c85e23