Tourism Market in India

Analyzing of Tourism Market in India for a Travel Hostel Startup

By Team – 1

Harsh

Divyansh Bobade

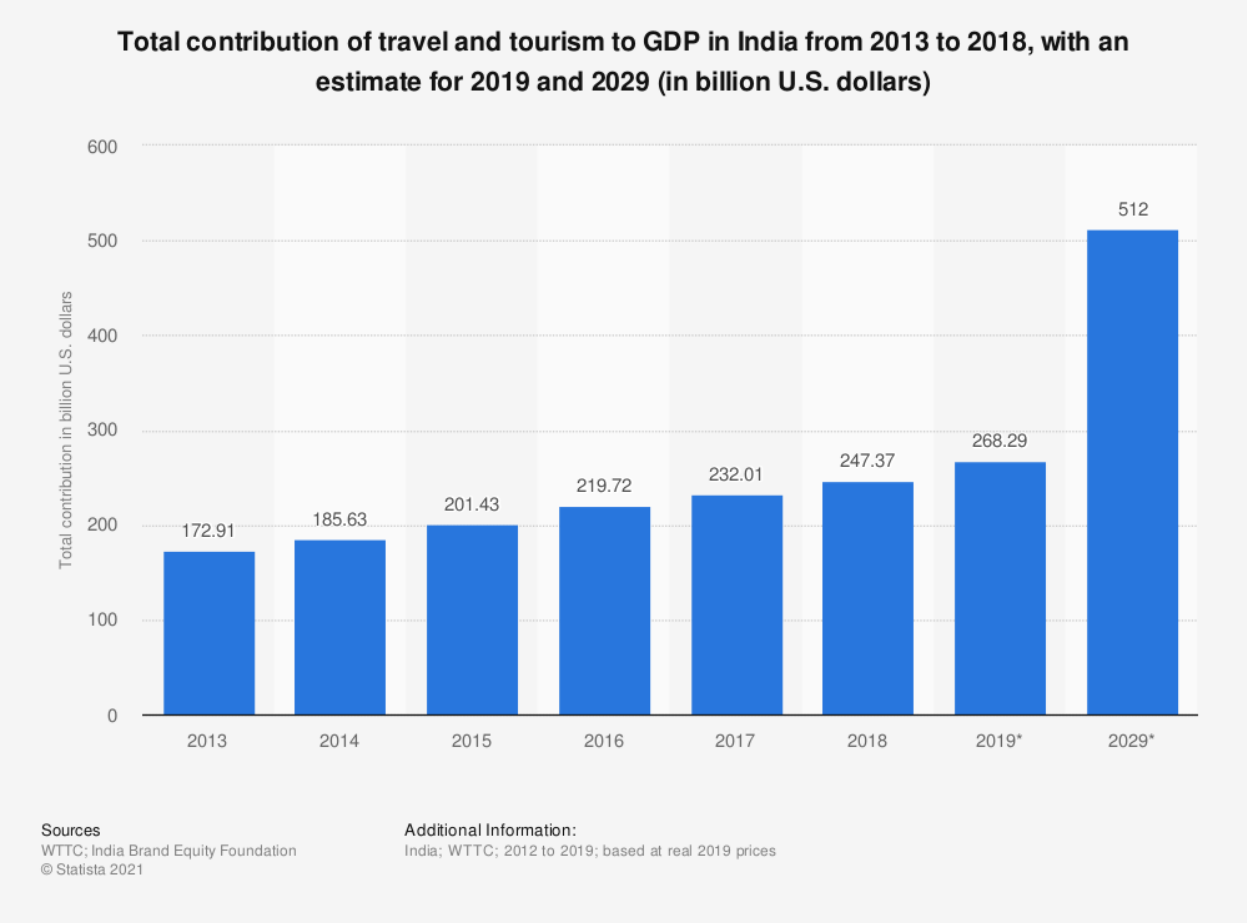
Aanangsha Borah

Sweety Biswas

Introduction:

One of the oldest civilizations in the world, India is a mosaic of multicultural experiences. With a rich heritage and myriad attractions, the country is among the most popular tourist destinations in the world. It covers an area of 32, 87,263 sq. km, extending from the snow-covered Himalayan heights to the tropical rain forests of the south. As the 7th largest country in the world, India stands apart from the rest of Asia, marked off as it is by mountains and the sea, which give the country a distinct geographical entity.

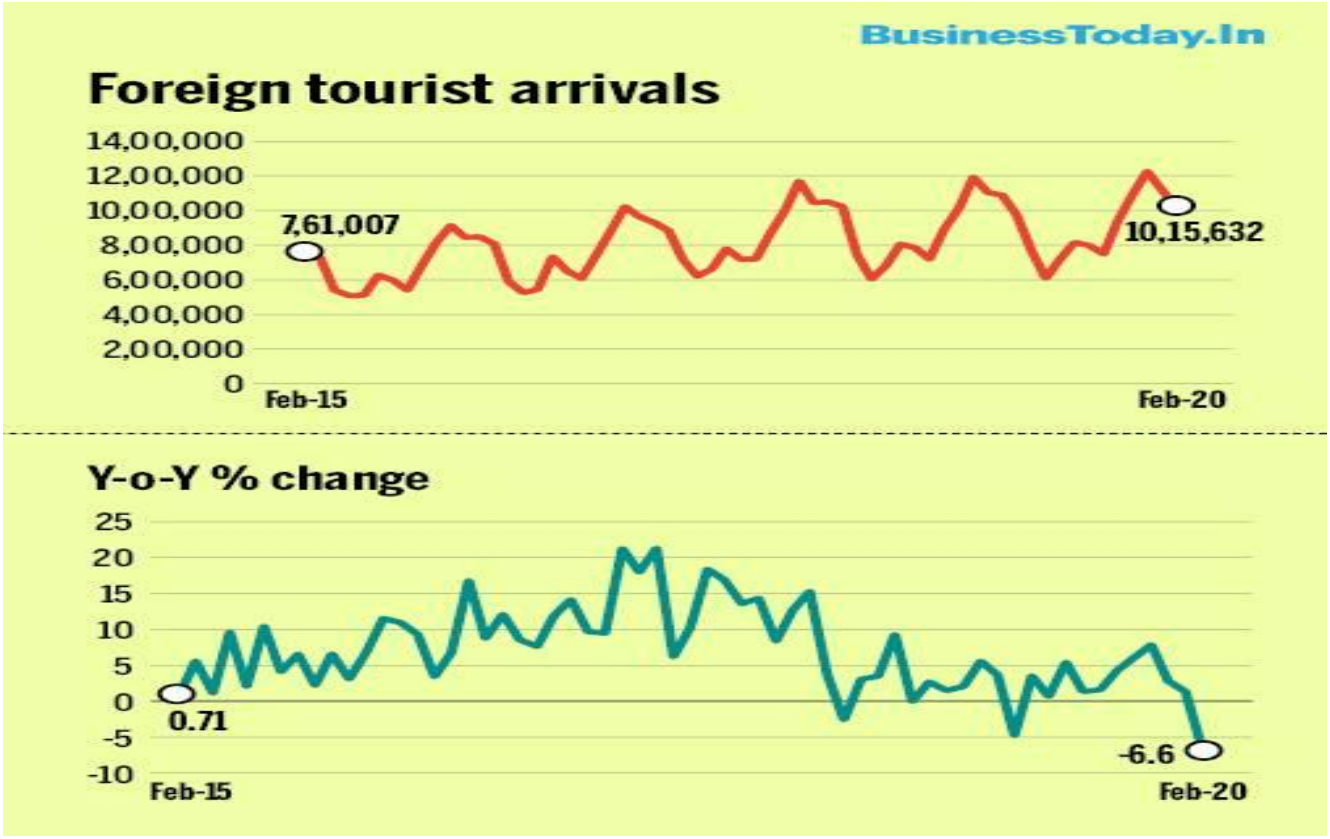
The Travel and Tourism Competitiveness Report 2019 ranked India 34th out of 140 countries overall. In FY20, the tourism sector in India accounted for 39 million jobs, which was 8.0% of the total employment in the country. By 2029, it is expected to account for about 53 million jobs.



Travel and Tourism was the largest service industry in India and it was worth $234 bn in 2018. But over the past year, the pandemic had a worsening effect on the tourism sector in India.

India has seen a major decline in incoming tourists, both on a domestic and international level, which has cost us a lot of GDP growth. The following graph shows incoming tourist arrivals in

India.



This market segmentation of the tourism market in India will help us to get an idea of what are the various factors that attract tourists, what sectors need improvement and what strategy should we implement to bring tourists, which will not only help us in growing India's GDP, but will only help many small scale business which depends upon tourists for their living, who had lost their means of income during this pandemic.

Exploration of Datasets:

Here we have a dataset of mainly number of travelers incoming to different parts of India both domestic and foreign from 2016 to 2018 with 115 rows (values) and 10 columns. Also we have the information regarding the places like connectivity, the tourist spot, growth rate in travelers etc.

Let us have a look into the dataset.

```
travellerwise=pd.read_csv("Tourism.csv")
travellerwise.head(10)
```

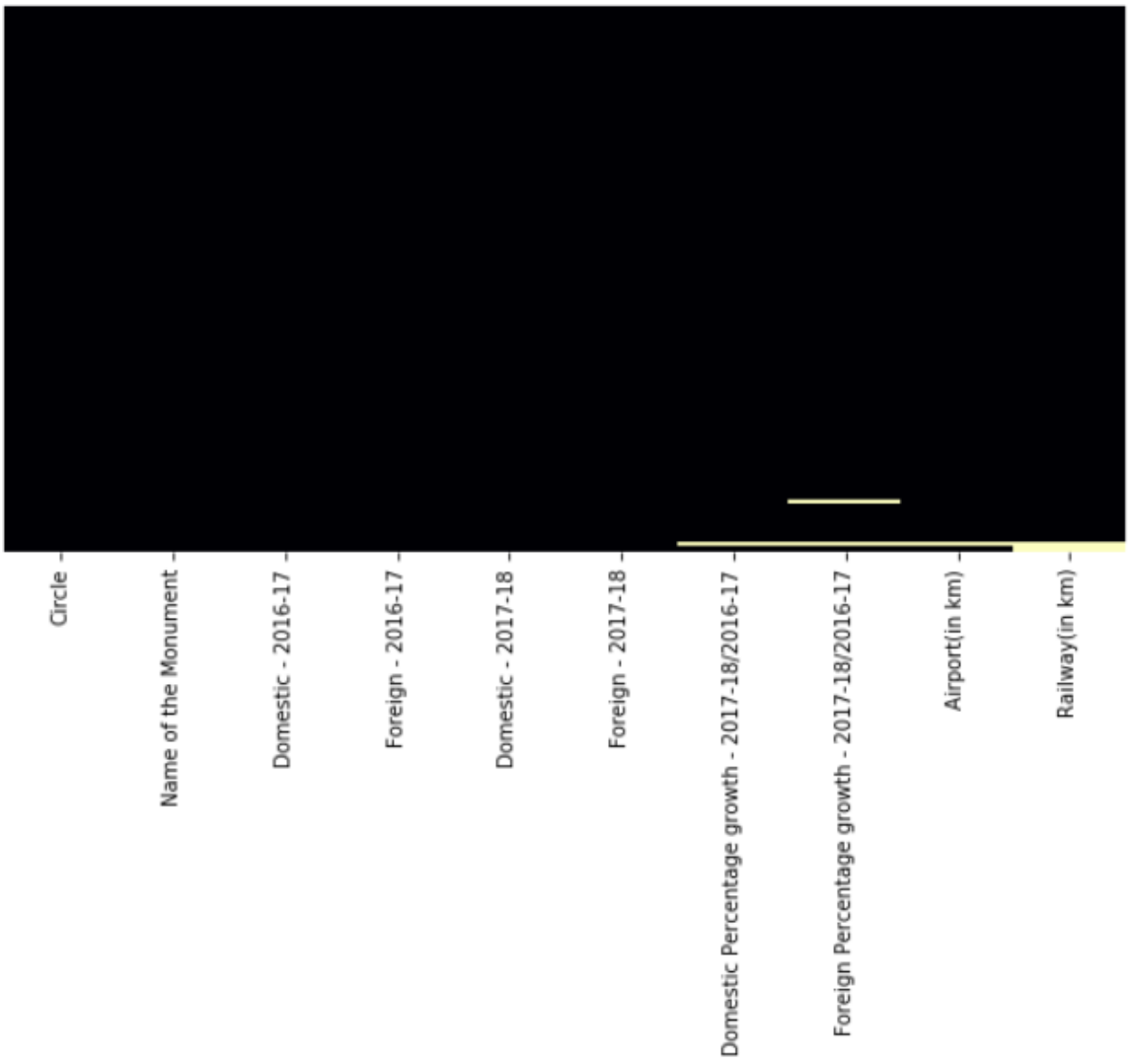
	Circle	Name of the Monument	Domestic - 2016-17	Foreign - 2016-17	Domestic - 2017-18	Foreign - 2017-18	Domestic Percentage growth - 2017-18/2016-17	Foreign Percentage growth - 2017-18/2016-17	Airport(in km)	Railway(in km)
0	Agra Circle	Taj Mahal	5419608	668403	5663136	794556	4.5	18.9	12.0	6.9
1	Agra Circle	Agra Fort	1799953	441326	2008208	489822	11.6	11.0	11.0	5.0
2	Agra Circle	Akbar's Tomb Sikandara	418502	22389	362070	23638	-13.5	5.6	13.0	12.0
3	Agra Circle	Fatehpur Sikri	427854	126114	962069	301181	124.9	138.8	35.0	37.0
4	Agra Circle	Itimad-ud-Daulah	108392	129177	138524	76572	27.8	-40.7	15.0	9.0
5	Agra Circle	Mehtab Bagh	215845	39969	172790	68897	-19.9	72.4	16.0	10.0
6	Agra Circle	RamBagh	56238	1911	67541	13493	20.1	606.1	12.0	11.0
7	Agra Circle	Mariam's Tomb Sikandara	28026	201	29880	12776	6.6	6256.2	14.0	12.0
8	Aurangabad Circle	Ellora Caves	1255537	24866	1645484	40405	31.1	62.5	36.0	28.0
9	Aurangabad Circle	Bibi-Ka-Maqbara	1291040	18756	1773100	20290	37.3	8.2	13.0	6.5

Let’s check for null values and if it has if there is any pattern to it. As null or empty values in the dataset might create problem so it is always better to look into it before use.

```
travellerwise.isnull().sum()
```

```
Circle                                          0
Name of the Monument                            0
Domestic - 2016-17                              0
Foreign - 2016-17                               0
Domestic - 2017-18                              0
Foreign - 2017-18                               0
Domestic Percentage growth - 2017-18/2016-17    1
Foreign Percentage growth - 2017-18/2016-17     2
Airport(in km)                                  1
Railway(in km)                                  2
dtype: int64
```

```
plt.figure(figsize=(10,5))
sns.heatmap(travellerwise.isnull(),cmap='magma',yticklabels=False,cbar=False)
plt.show()
```



There are a very few null values in our datasets. It seems from the above plot that null values are not following any such significant pattern. Only some values like distance from railway and airport are missing and it is a pretty general case as some places might not be well connected. They can be removed because we already have sufficient amount of dataset to work with.

After cleaning the datasets now we are left with a dataset of 112 rows and 10 columns.

Let us look into the structure of the datasets and some statistical measurements of the datasets.

```
df.describe()
```

	Domestic - 2016-17	Foreign - 2016-17	Domestic - 2017-18	Foreign - 2017-18	Domestic Percentage growth - 2017-18/2016-17	Foreign Percentage growth - 2017-18/2016-17	Airport(in km)	Railway(in km)
count	1.120000e+02	112.000000	1.120000e+02	112.000000	112.000000	112.000000	112.000000	112.000000
mean	3.991450e+05	26542.321429	4.759064e+05	32318.883929	48.383036	1297.487500	62.323214	18.895625
std	7.165584e+05	88792.877717	7.810398e+05	100170.832913	101.089579	9978.534996	64.898156	36.450659
min	3.430000e+02	2.000000	6.400000e+01	0.000000	-97.000000	-100.000000	0.300000	0.500000
25%	4.869550e+04	229.500000	7.132100e+04	331.750000	12.950000	19.800000	14.750000	3.000000
50%	1.730935e+05	1405.500000	2.057655e+05	2736.000000	34.550000	42.550000	39.250000	6.700000
75%	3.960450e+05	10658.250000	4.936770e+05	15126.500000	51.400000	68.725000	91.175000	17.075000
max	5.419608e+06	668403.000000	5.663136e+06	794556.000000	678.100000	101485.200000	294.000000	288.000000

```
print(travellerwise['Circle'].value_counts())
```

```
Delhi Circle          10
Mumbai Circle         10
Agra Circle            8
Hyderabad Circle       8
Chennai Circle         7
Bhopal Circle          7
Dharwad Circle         6
Vadodara Circle        6
Aurangabad Circle      6
Bhubaneswar Circle     5
Lucknow Circle         5
Guwahati Circle        5
Patna Circle           5
Sarnath Circle         4
Bengaluru Circle       4
Srinagar Circle        3
Kolkata Circle         3
Chandigarh Circle      2
Thrissur Circle        2
Hampi Mini Circle      2
Shimla Mini Circle     2
Jodhpur Circle         2
Jaipur Circle          1
Raipur Circle          1
Leh Mini Circle        1
Name: Circle, dtype: int64
```

We have maximum number of tourist spot in Delhi circle and Mumbai circle followed by Agra and Hyderabad circle.

Data Visualization

Let's look into the flow of tourist grouped by circles.

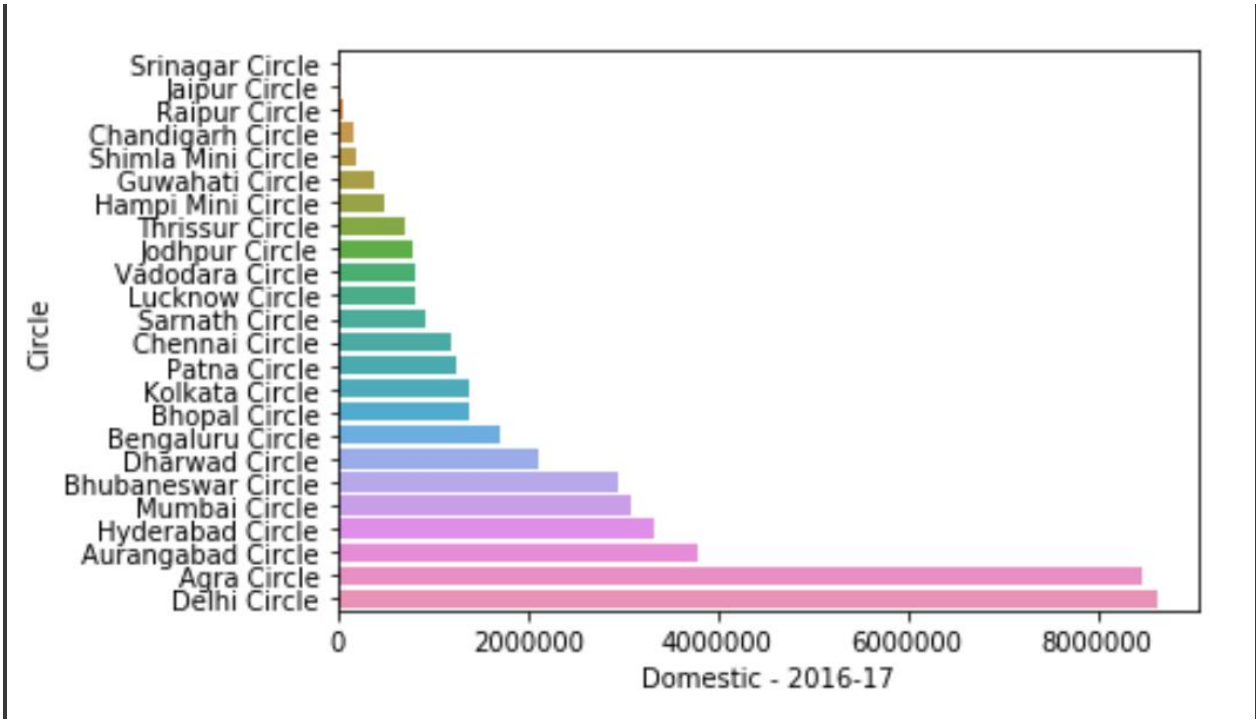


Fig: flow of domestic tourist in the year 2016-17

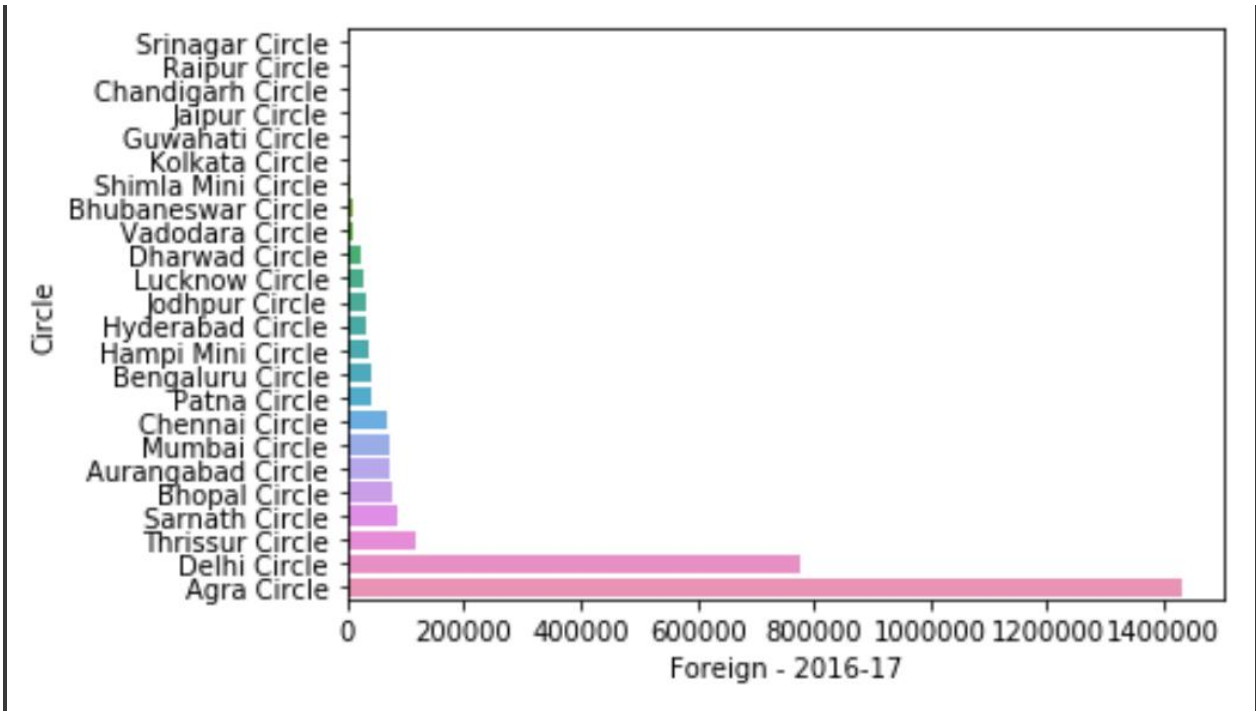


Fig: flow of foreign tourist in the year 2016-17

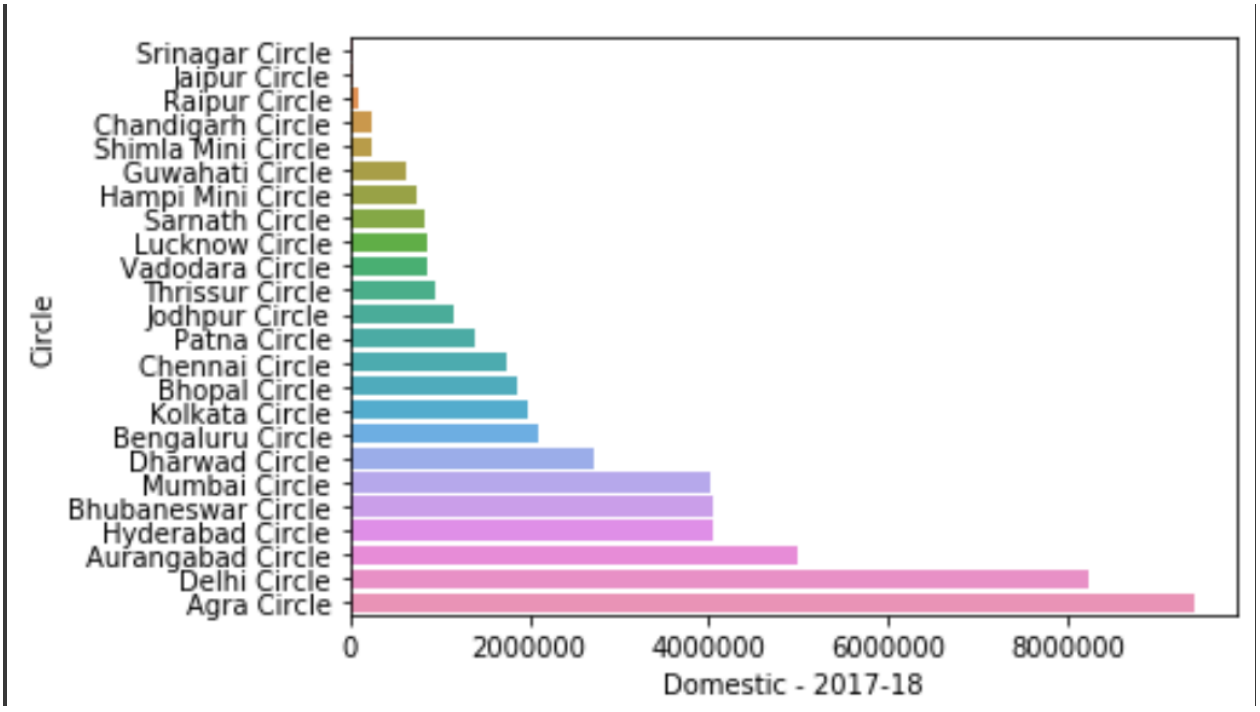


Fig: flow of domestic tourist in the year 2017-18

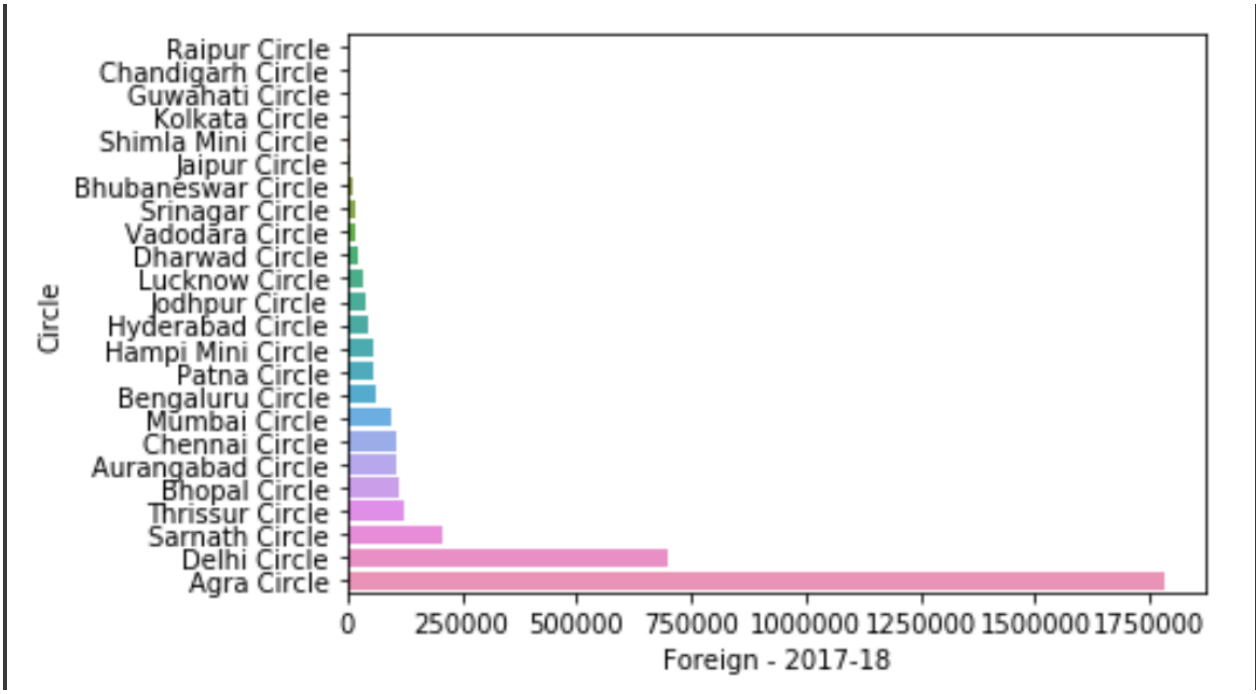
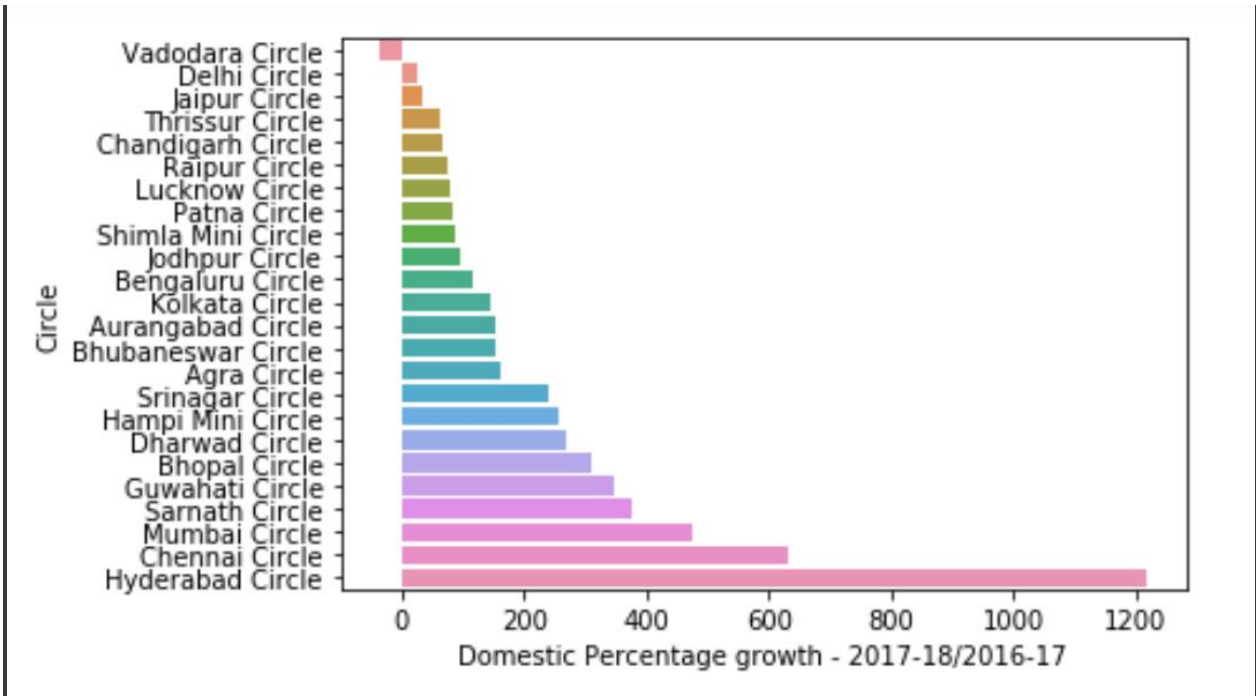


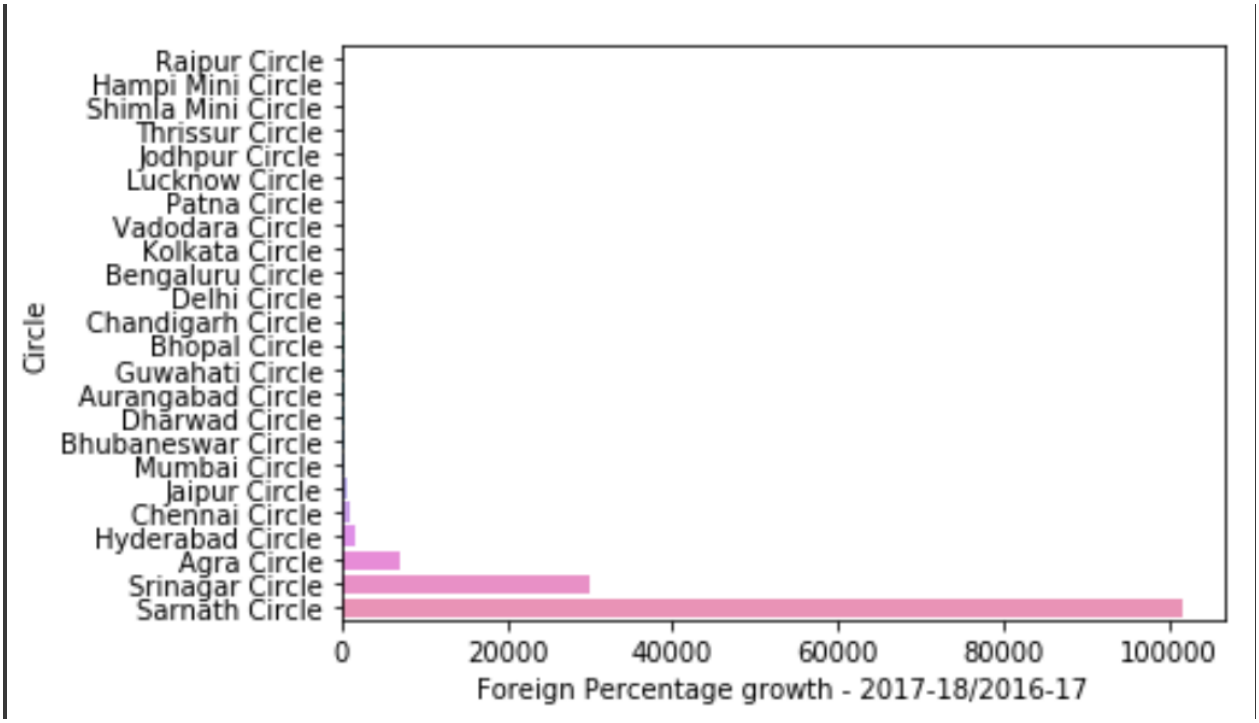
Fig: flow of foreign tourist in the year 2017-18.

The above exploration implies that Agra circle is holding largest number of intake for both domestic and foreign tourists except for domestic tourist in 2016-17 but with a very little margin. Their amount of tourist incoming is comparatively very larger than other circles.

Lets us look into the circle wise growth of tourist from the year 2016 to 2018.



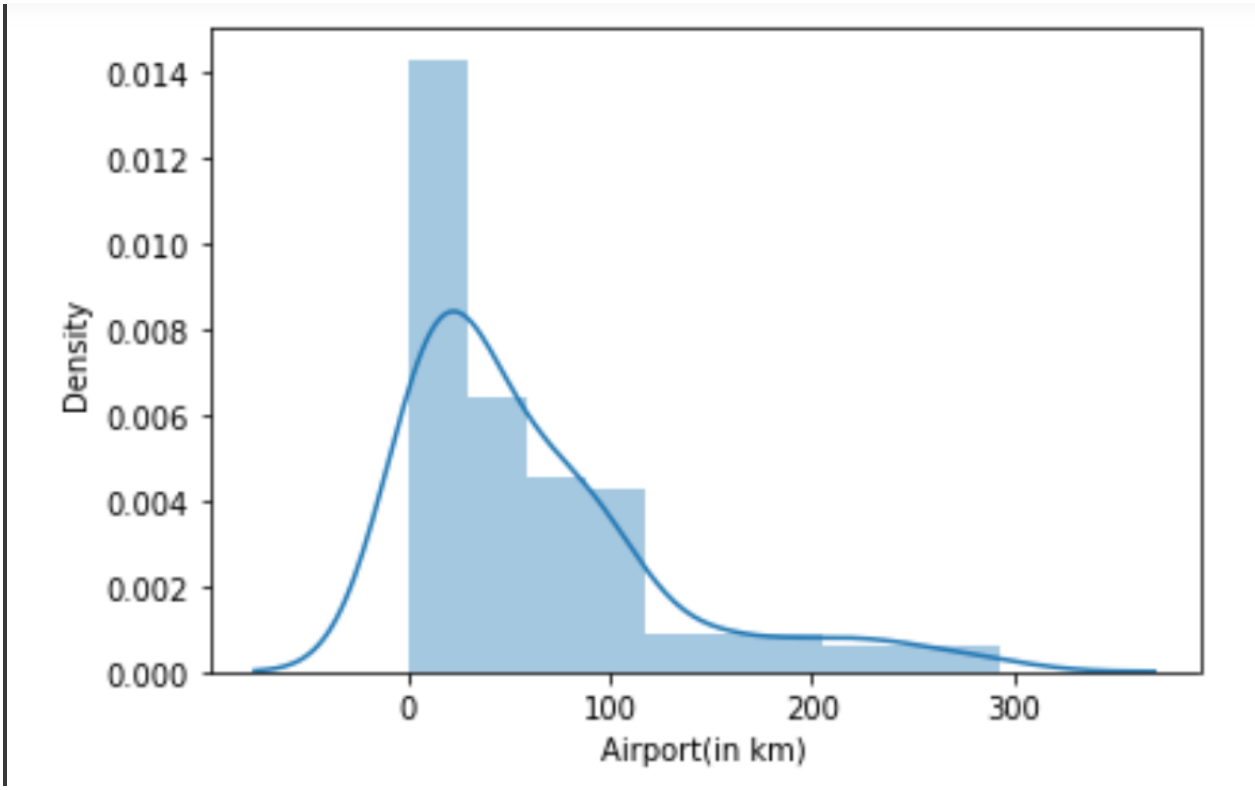
Hydrabad circle has highest increase in the growth of tourist flow followed by Chennai circle. Vadodara circle has shown decrease in tourist flow growth. Also most of the circles have an increased growth rate.

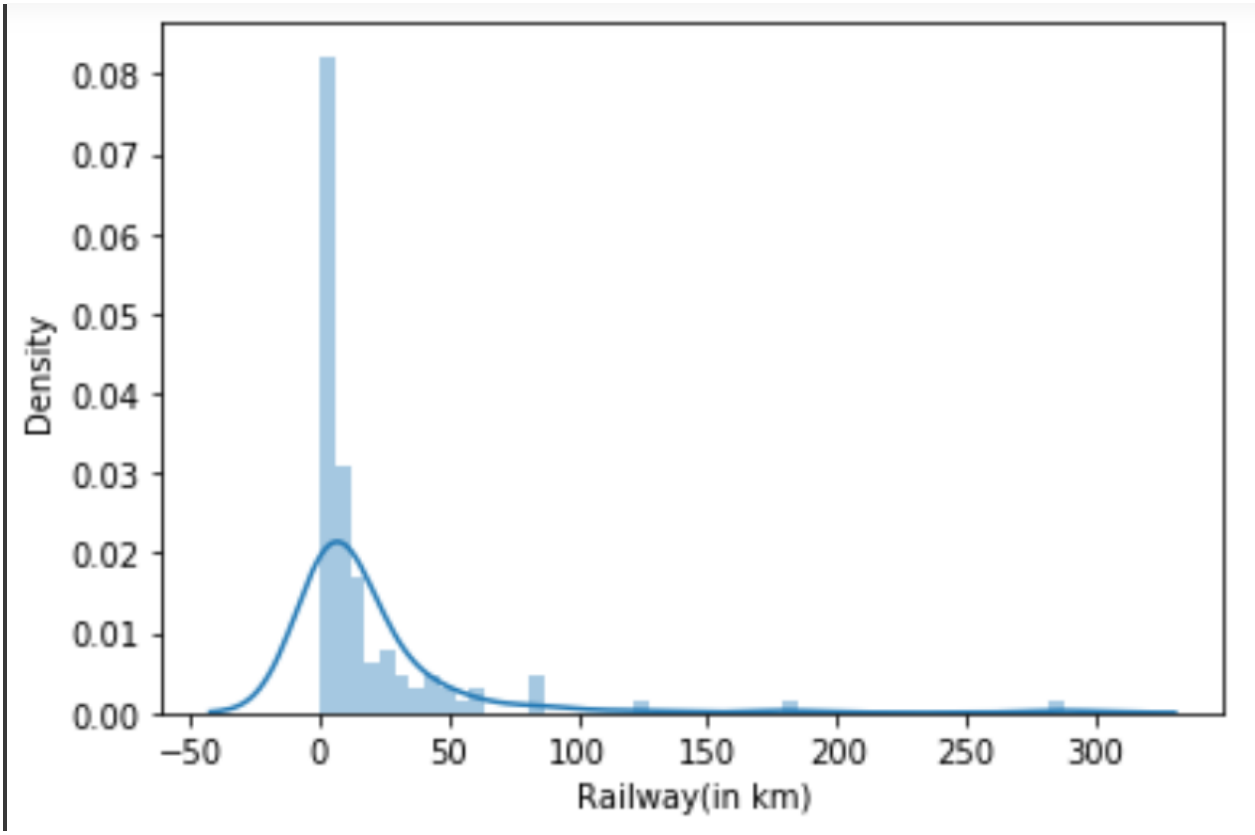


It is seen that most of the circles have very constant flow of tourist in the past two years except for Sarnath Circle which an tremendous increase in the Foreign tourist flow.
Below graph represents density of tourist spot according to its distance from railway station and airport.

Fig: It's very obvious that the connectivity of a place plays an important role in the tourism sector. The distance of a particular travel destination from the major transportation facilities is a major factor to consider.

Therefore let us look into how the distance is affecting the number of tourist incoming.





As expected more the distance of a place from Airport and railway less is the number of tourist. Also it is clear from the above graph that more tourist spots have grown which are closer to railway and airport. Connectivity is an important issue to consider.

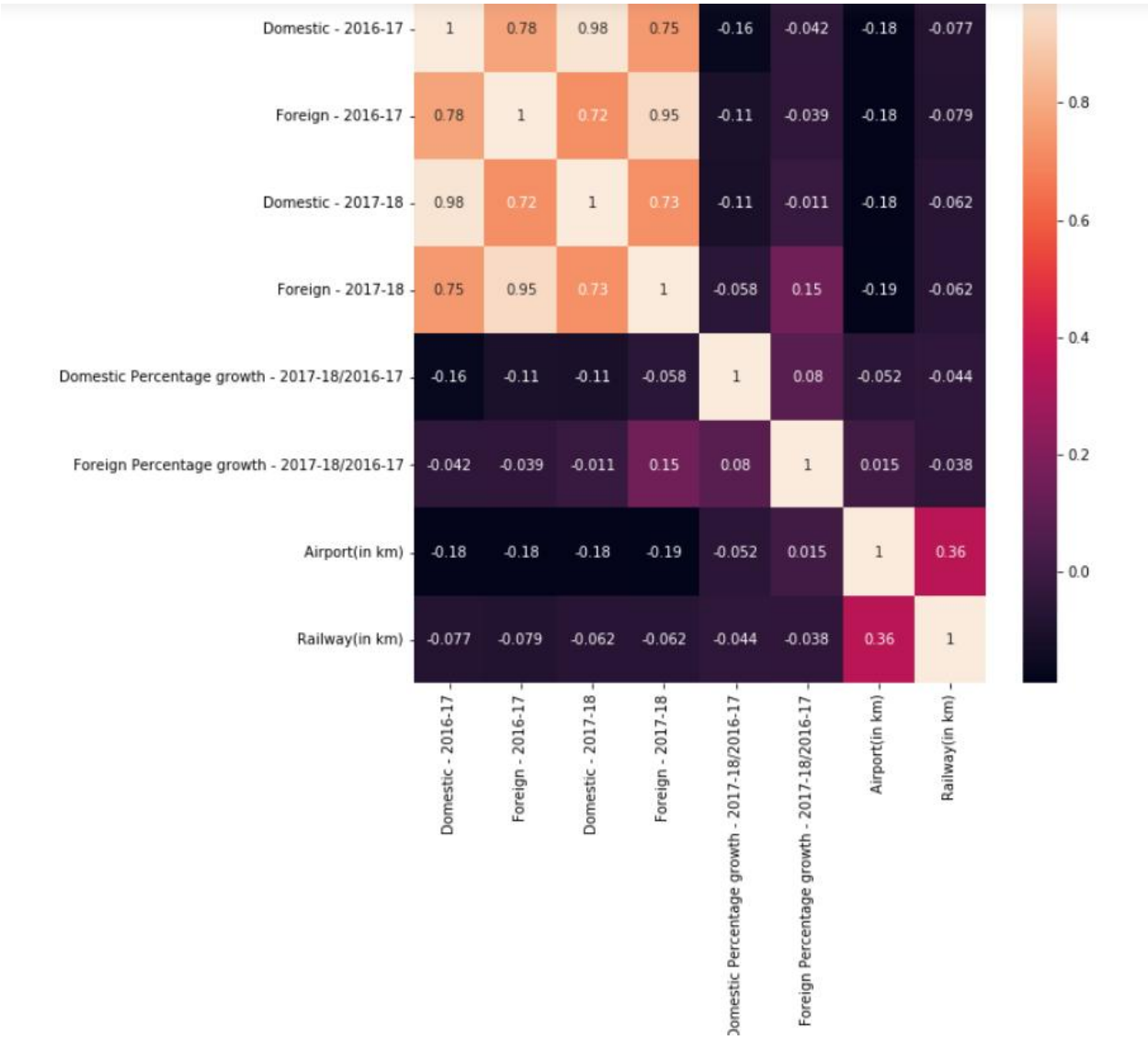
```
num_data = df.select_dtypes(exclude='object').columns.tolist()
cat_data = df.select_dtypes(include='object').columns.tolist()

print(num_data)
print(cat_data)
```

```
['Domestic - 2016-17', 'Foreign - 2016-17', 'Domestic - 2017-18', 'Foreign - 20
17-18', 'Domestic Percentage growth - 2017-18/2016-17', 'Foreign Percentage gro
wth - 2017-18/2016-17', 'Airport(in km)', 'Railway(in km)']
['Circle', 'Name of the Monument']
```

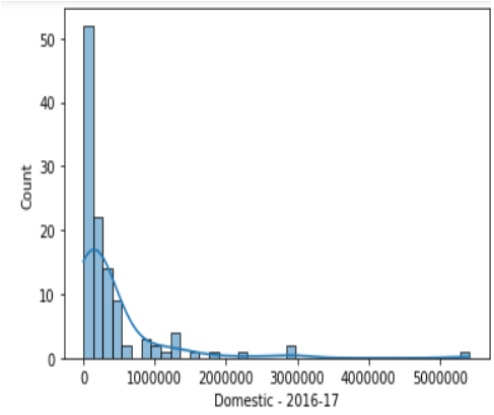
By taking a look at some of our features we could say that the features are highly skewed and have Positive Skewness which suggests that there are outliers in our dataset. Large outliers need to be transformed or removed for clustering purposes.

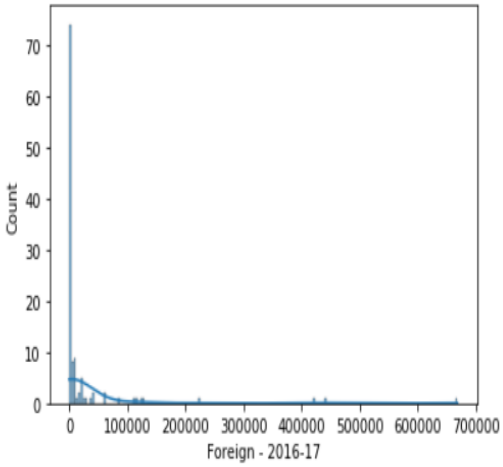
```
plt.figure(figsize=(9,9))
sns.heatmap(df.corr(),annot=True)
```

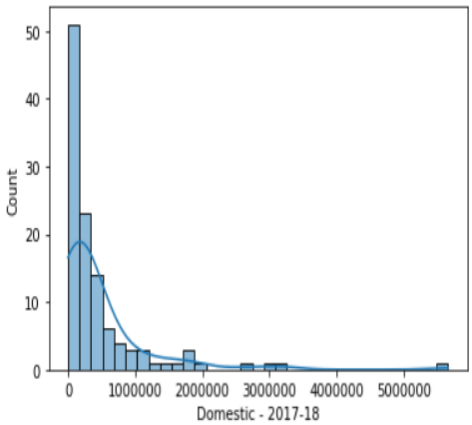


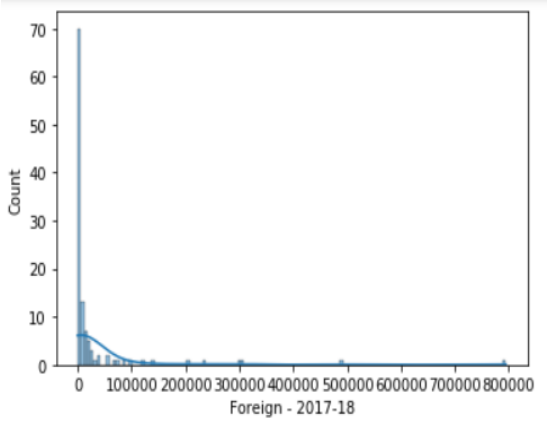
From the Heat map we see that our two features have a high correlation in between them i.e., between these features :- ["Domestic - 2016-17" and "Domestic - 2017-18"] and ["Foreign - 2017-18" and "Foreign - 2016-17"] So, we will be choosing only one of the features from the

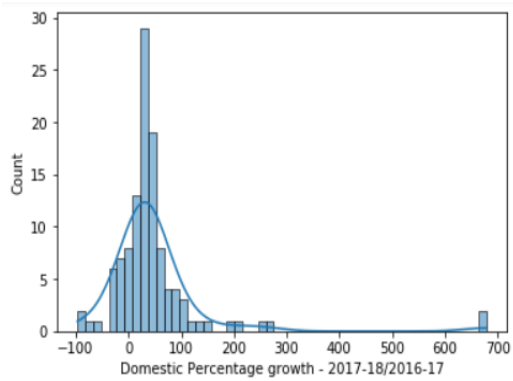
above two i.e., ["Domestic - 2017-18", "Foreign - 2017-18"] So, we preprocessed/transformed our features using Power Transformer which made the data have mean=0 and variance=1.

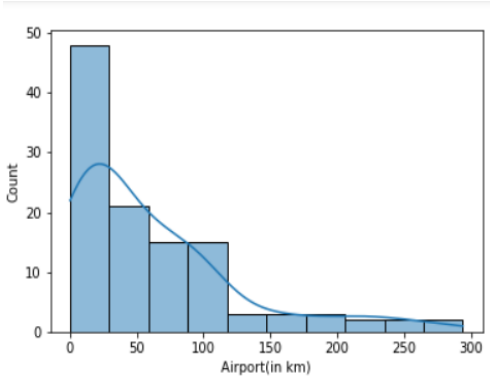


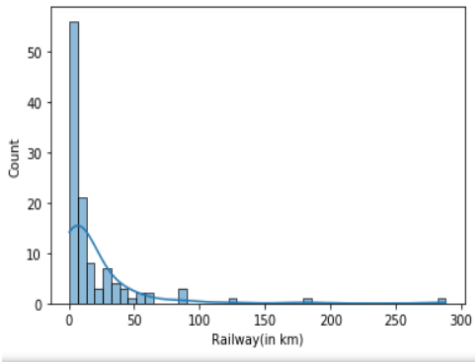





Count
Foreign - 2017-18


Count
Domestic Percentage growth - 2017-18/2016-17


Count
Airport(in km)


Count
Railway(in km)

Extracting Segments

Distance Based Methods

1) Hierarchical Clustering

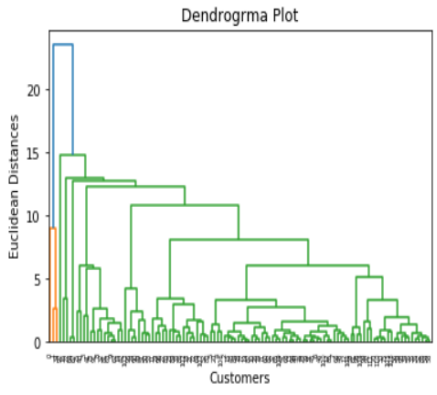
Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations (consumers) into k groups (segments). If the aim is to have one large market segment (k = 1), the only possible solution is one big market segment containing all consumers in data X. At the other extreme, if the aim is to have as many market segments as there are consumers in the data set (k = n), the number of market segments has to be n, with each segment containing exactly one consumer. Each consumer represents their own cluster. Market segmentation analysis occurs between those two extremes.

In this we will use Agglomerative hierarchical clustering which approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters).

```
In [29]: import scipy.cluster.hierarchy as shc
         from sklearn.preprocessing import MinMaxScaler,StandardScaler
         from sklearn.cluster import AgglomerativeClustering
```

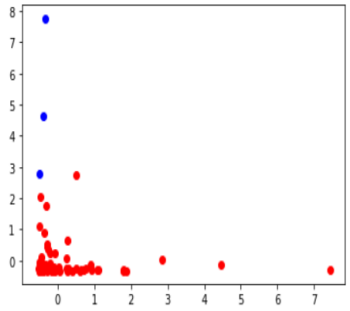
Complete linkage hierarchical cluster analysis of the tourist data set

```
In [42]: dendro = shc.dendrogram(shc.linkage(x, method="ward"))
         plt.title("Dendrogrma Plot")
         plt.ylabel("Euclidean Distances")
         plt.xlabel("Customers")
         plt.show()
```



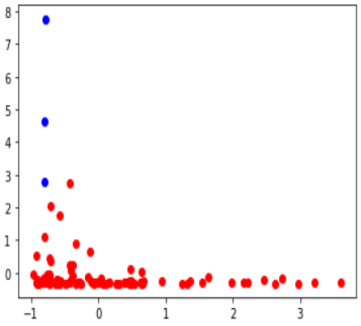
```
In [49]: plt.scatter(df1['Railway(in km)'],df1['Foreign - 2017-18'],c='red')
         plt.scatter(df2['Railway(in km)'],df2['Foreign - 2017-18'],c='blue')

Out[49]: <matplotlib.collections.PathCollection at 0x23e09b5c4c8>
```



```
In [50]: plt.scatter(df1['Airport(in km)'],df1['Foreign - 2017-18'],c='red')
         plt.scatter(df2['Airport(in km)'],df2['Foreign - 2017-18'],c='blue')

Out[50]: <matplotlib.collections.PathCollection at 0x23e0764d208>
```



Partitioning Methods

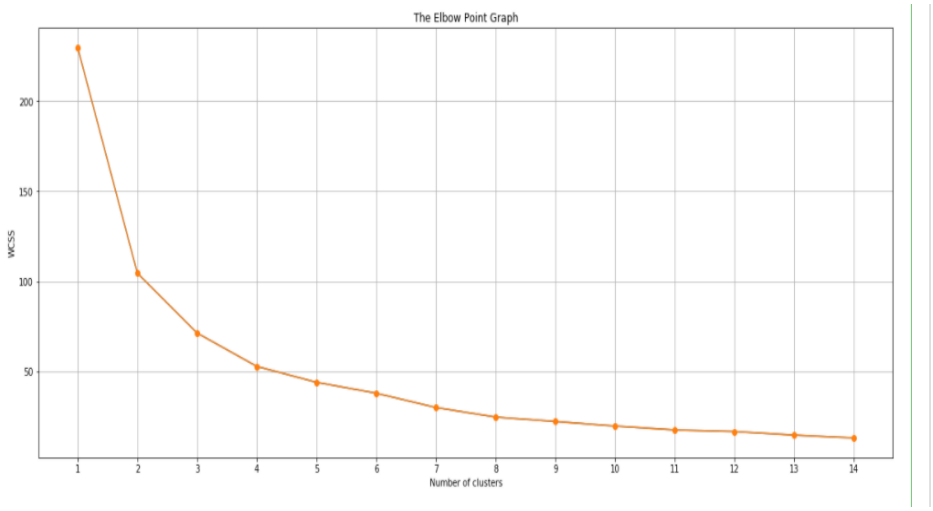
- **K-means Clustering**

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

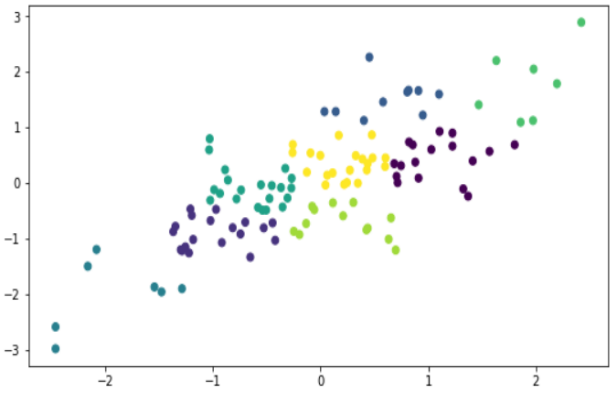
Step 1: Choose the number of clusters *k*
Step 2: Select k random points from the data as centroids
Step 3: Assign all the points to the closest cluster centroid
Step 4: Recompute the centroids of newly formed clusters
Step 5: Repeat steps 3 and 4

```
In [56]: 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()
```

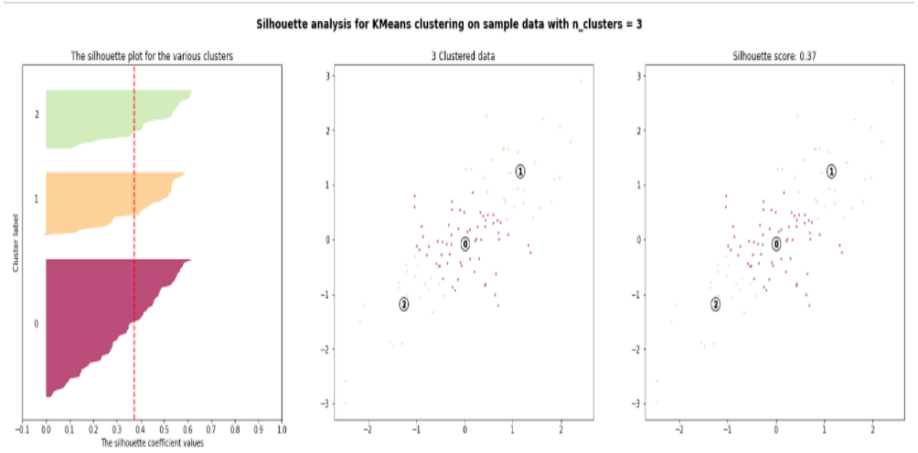


Out[58]: <matplotlib.collections.PathCollection at 0x23e0805c448>

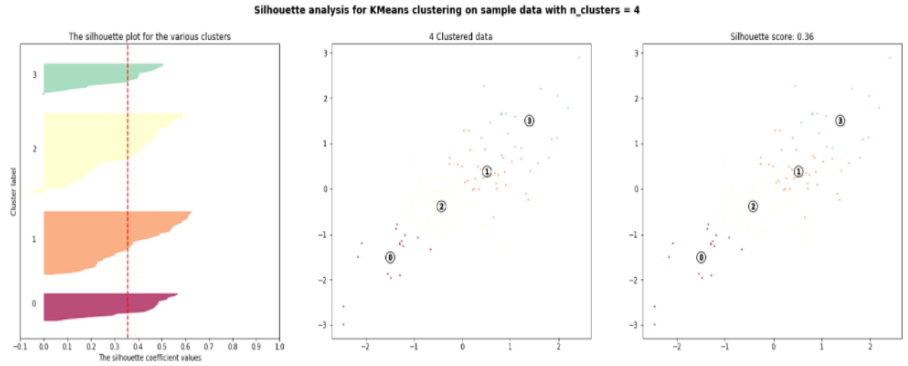


K equal to 3 -> observed low rates of gain in the decay of the distortions with the decrease of K reaching the limit of 10 % with the K equal to 9. Silhouette analysis on K-Means clustering

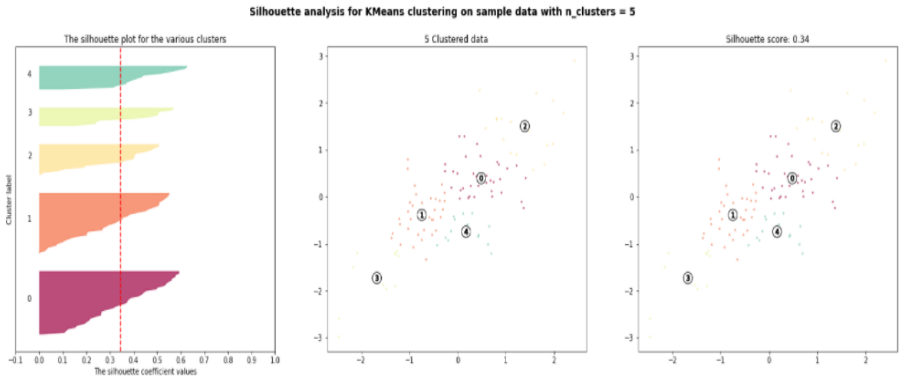
For Cluster 3



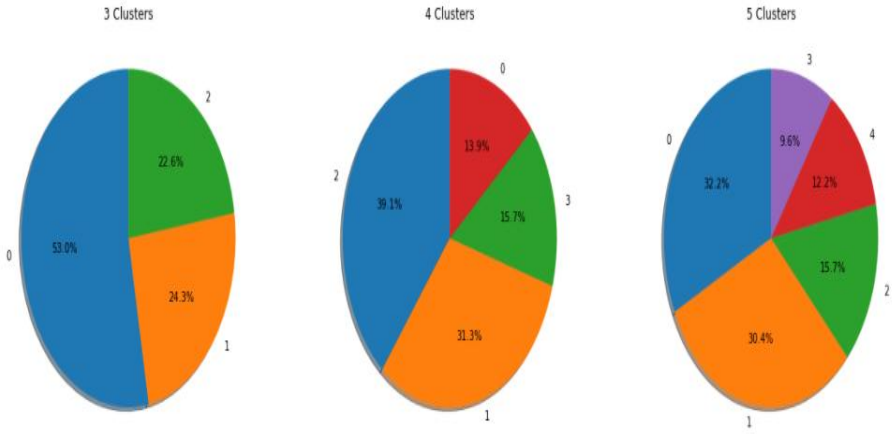
For Cluster 4



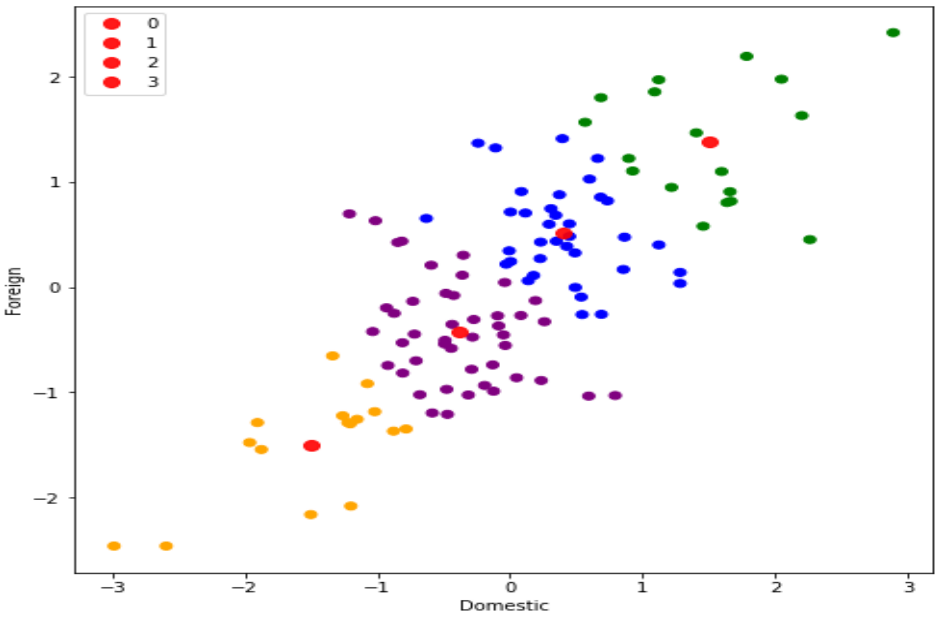
For Cluster 5



The Silhouette score if greater than 0 tells us that our clusters are separated well. The Closer to 1 the better the clusters are separated and closer to -1 the worse the separation.



From the Silhouette plot and Coefficient and the number of data points associated with our cluster, we pick 4 as our optimum number of clusters. Segmenting our feature with the K-Means algorithm with the number of clusters is equal to 4.



K-means plot for 4 clusters

Now, We One hot encoded our Circle features and created a data frame table for our Circle, encoded feature and the prediction made for 4 clusters. For each cluster, we tried to find the circle that has a great impact on each of our cluster values.

- For Cluster value 0:

```
: c0 = clust_prod[clust_prod['cluster'] == 0].drop('cluster', axis=1).mean()
  c0.sort_values(ascending=False)[0:10]
```

```
C:\Users\Dell\anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarni

Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None'
ise TypeError.  Select only valid columns before calling the reduction.
```

```
: Circle_Vadodara Circle      0.1250
  Circle_Srinagar Circle      0.1250
  Circle_Chennai Circle       0.1250
  Circle_Delhi Circle         0.1250
  Circle_Hyderabad Circle     0.1250
  Circle_Lucknow Circle       0.0625
  Circle_Bhopal Circle        0.0625
  Circle_Sarnath Circle       0.0625
  Circle_Guwahati Circle      0.0625
  Circle_Hampi Mini Circle    0.0625
  dtype: float64
```

- **For Cluster value 1 :**

```
c1 = clust_prod[clust_prod['cluster'] == 1].drop('cluster', axis=1).mean()
c1.sort_values(ascending=False)[0:10]
```

```
C:\Users\Dell\anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarni

Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None'
ise TypeError.  Select only valid columns before calling the reduction.
```

```
Circle_Dharwad Circle       0.138889
Circle_Agra Circle          0.111111
Circle_Bhopal Circle        0.111111
Circle_Mumbai Circle        0.111111
Circle_Bengaluru Circle     0.083333
Circle_Lucknow Circle       0.055556
Circle_Jodhpur Circle       0.055556
Circle_Aurangabad Circle    0.055556
Circle_Vadodara Circle      0.055556
Circle_Delhi Circle         0.055556
dtype: float64
```

- **For Cluster value 2** :

```
c2 = clust_prod[clust_prod['cluster'] == 2].drop('cluster', axis=1).mean()
c2.sort_values(ascending=False)[0:10]
```

```
C:\Users\Dell\anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWar

Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=Nor
ise TypeError.  Select only valid columns before calling the reduction.
```

```
Circle_Hyderabad Circle      0.088889
Circle_Mumbai Circle         0.088889
Circle_Guwahati Circle       0.088889
Circle_Patna Circle          0.066667
Circle_Bhubaneswar Circle    0.066667
Circle_Chennai Circle        0.066667
Circle_Delhi Circle          0.066667
Circle_Vadodara Circle       0.044444
Circle_Kolkata Circle        0.044444
Circle_Aurangabad Circle     0.044444
dtype: float64
```

- **For Cluster value 3**:

```
c3 = clust_prod[clust_prod['cluster'] == 3].drop('cluster', axis=1).mean()
c3.sort_values(ascending=False)[0:10]
```

```
C:\Users\Dell\anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarn

Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None
ise TypeError.  Select only valid columns before calling the reduction.
```

```
Circle_Agra Circle           0.166667
Circle_Delhi Circle          0.166667
Circle_Hyderabad Circle      0.111111
Circle_Aurangabad Circle     0.111111
Circle_Bengaluru Circle      0.055556
Circle_Thrissur Circle       0.055556
Circle_Bhubaneswar Circle    0.055556
Circle_Chennai Circle        0.055556
Circle_Sarnath Circle        0.055556
Circle_Hampi Mini Circle     0.055556
dtype: float64
```

Result :

From K-means plot for 4 clusters we could infer that as the Domestic and Foreign tourists increase cluster 1 gives us a better result. We could say that Circle Delhi is our best choice for establishing the tourism market. But Cluster 1 doesn't satisfy one of our knock-out criteria i.e., The Segment must be large enough. Cluster 2 accounts for 30.1% of total value while cluster 1 accounts for 13.3%. So, cluster 1 fails on one criterion. Let's take a look at another Knock Out Criteria i.e., The Segment must be homogenous. So, we found out the mean average distance between the data points in a cluster.

```
from scipy.spatial import distance_matrix

dist_mat = pd.DataFrame(distance_matrix(X, centroids))

dist_mat.groupby(c_preds).mean()
```

	0	1	2	3
0	0.678943	2.783230	1.579199	4.175453
1	2.803146	0.510383	1.309192	1.492366
2	1.625812	1.341509	0.568743	2.676501
3	4.199277	1.520784	2.678040	0.736140

Distance Matrix for our cluster

From the above figure, we see that the mean average distance in cluster 1 is 0.510383 while that of cluster 2 is 0.568743. We know customers in the same segment prefer the same product quality preferences and characteristics that affect their purchasing decisions. So, we could say that Cluster 2 passes the Knock Out Criteria and Cluster 1 again fails on this criterion. So, we fail to reject our Cluster 2. Hence Cluster 2 is the perfect Cluster for further analysis. Figure for cluster value, 2 says that the most appropriate destinations for establishing our tourism market are: Agra Circle, Dharwad Circle, and Bhopal Circle. Again, from one of Our Knock Criteria i.e., Segment must be Large Enough and from one of the figures in the market overview section, we can see that Agra circle definitely has a large number of tourists compared to Bhopal and Dharwad Circle.

Target Segment :

After doing an analysis of our data and applying different models, we need to select a market segment that satisfies the knock-out criteria and is more profitable compared to others. We came to the conclusion that cluster 2 would be the best fit as it satisfies the knock out criteria and gives certain advantages, such as:

1. Cluster 2 is homogenous compared to others, as the average distance b/w data points in a cluster are least in 2.

2. Cluster 2 is well separated as the in-between distance of cluster 2 from other clusters is distinguishable

3. Cluster 2 contain a large number of data points

4. The members of cluster 2 have a high tourist population compared to others.

The major circles that comprise cluster 2 are from:-

1. Agra Circle

2. Bhopal Circle

3. Dharwad Circle

- Therefore for the Travel Startup, it would be best to open hotel chains first in these regions as they offer distinctive advantages and are highly profitable.

Marketing Mix :

In order to reach a wider audience, we need to focus on multiple areas, for a comprehensive marketing plan. Effective marketing is generally met by looking into broad areas, rather than fixating into one thing. The 4Ps help marketing professionals maintain focus on things that really matter. It gives power to organizations to make strategic decisions when launching new products or revising existing ones.

The 4 P's of Marketing Mix

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

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

3. 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.

4. 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.

GitHub: https://github.com/iamharsh1312/Tourism-Market-Segmentation-in-India