National Institute of Technology, Kurukshetra

Master of Computer Application

A Project Report On

Data Analytics Semester Long Assignment

Submitted By

Vikas Verma 52221204

Samiran Saha 52212212

Mahesh Patidar 52222106

Under the guidance of

Dr. Kapil Gupta

Assistant Professor, MCA Department,NIT KKR

April 2024

Department of Computer Application

Acknowledgement

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to **National Institute of Technology Kurukshetra,Haryana** for providing us with a great opportunity to pursue our Master's degree in this institution.

It is a matter of immense pleasure to express our sincere thanks to **Dr.Kapil Gupta**,Assistant Professor &Department of Computer Application, NIT Kurukshetra, for his constant encouragement and expert advice & providing right academic guidance that made our task possible.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

Vikas Verma 52221204

Samiran Saha 52212212

Mahesh Patidar 52222106

Abstract

This dataset presents a comprehensive exploration guide highlighting must-visit destinations across various regions of India. It serves as a valuable resource for travelers and enthusiasts, providing detailed insights into the unique characteristics of each location. The dataset encompasses a diverse range of attractions, including historical landmarks, religious shrines, and natural wonders, offering a glimpse into India's rich cultural heritage. Key features of the dataset include geographical categorization by zones (e.g., Northern, Southern), state-wise location details, city/town information, and the names of tourist spots or landmarks. Each destination is classified based on its type (e.g., Temple, War Memorial, Natural Park) and includes the establishment year, estimated time needed for a thorough visit, and average Google review rating. Additionally, the dataset provides practical information such as entrance fees in Indian Rupees, the presence of airports within a 50-kilometer radius for accessibility, and the weekly off day for each destination. The significance of each place, whether historical, religious, or environmental, is also outlined, along with permissions for DSLR camera use. Moreover, the dataset includes the total number of Google reviews received by each destination, expressed in lakhs (hundreds of thousands), and suggests the best time of day to visit for an optimal experience. This abstract serves as a concise summary of the dataset's contents, facilitating its use for research, analysis, and travel planning purposes.

Contents

1 Introduction **6**

1.1 Title of dataset 6

1.2 Dataset Link 6

1.3 Dataset Source 6

1.4 Dataset Description 6

1.5 Dataset Attributes 6

1.5.1 Zone 6

1.5.2 State 6

1.5.3 City Name 6

1.5.4 Name 6

1.5.5 Type 6

1.5.6 Establishment Year 6

1.5.7 Time needed to visit(hrs) 7

1.5.8 Google Review Rating 7

1.5.9 Entrance Fee(IN INR) 7

1.5.10 Airport with 50km Radius 7

1.5.11 Weekly Off 7

1.5.12 Significance 7

1.5.13 DSLR Allowed 7

1.5.14 Number of Google Review Rating (in lakhs) 7

1.5.15 Best time to visit 7

2 Problem Statement **8**

2.1 Question no 1 8

2.2 Question no 2 8

2.3 Question no 4 8

2.4 Question no 5 8

2.5 Question no 6 8

2.6 Question no 9 8

3 Methodology **9**

3.1 Data Collection 9

3.2 Data Pre-processing 9

3.2.1 Exploratory data analysis(EDA) 9

3.3 Clustering 21

3.3.1 Elbow method to find the value of k 21

3.3.2 K-means Clustering 21

3.4 Model Implementation 22

3.4.1 Regression problem 22

3.4.2 Classification 22

3.5 Principal Component Analysis (PCA) 23

4 Conclusion **24**

5 References **25**

1 Introduction

1.1 Title of dataset

Travel Dataset: Guide to India's Must See Places

1.2 Dataset Link

https://www.kaggle.com/datasets/saketk511/travel-dataset-guide-to-indias-must-see-places/data

1.3 Dataset Source

Kaggle

1.4 Dataset Description

The "Guide to India's Must-See Places" dataset provides comprehensive information about various tourist destinations across India. It serves as a valuable resource for traveler's seeking to explore India. The dataset includes essential details such as location, type of attraction, establishment year, visitation duration, Google review ratings, entrance fees, nearby airports, weekly off days, significance, DSLR permissions, number of Google reviews, and the best time to visit.

1.5 Dataset Attributes

1.5.1 Zone

The geographical region or zone where the destination is located (e.g., Northern,Southern, Eastern, Western, Central)

1.5.2 State

The state within India where the destination is situated (e.g., Rajasthan, Kerala, Uttar Pradesh, West Bengal etc).

1.5.3 City Name

The name of the city or town where the attraction is located (e.g., Konark, Purulia etc).

1.5.4 Name

The name of the tourist destination or landmark (e.g., India Gate, Marine Drive etc).

1.5.5 Type

Classification based on the nature of the attraction (e.g., Park, Market, Temple etc)

1.5.6 Establishment Year

The year in which the attraction was established or built.

1.5.7 Time needed to visit(hrs)

Approximate duration required to explore the attraction fully.

1.5.8 Google Review Rating

The average rating given by visitors on Google Reviews, representing overall satisfaction.

1.5.9 Entrance Fee(IN INR)

The cost of entry or admission fee for visiting the attraction, denominated in Indian Rupees (INR).

1.5.10 Airport with 50km Radius

The nearest airport located within a 50-kilometer radius of the destination.

1.5.11 Weekly Off

The day(s) of the week when the attraction is closed or has limited operating hours.

1.5.12 Significance

A brief description highlighting the historical, cultural, or natural significance of the destination.

1.5.13 DSLR Allowed

Indicates whether visitors are permitted to use DSLR cameras for photography at the attraction.

1.5.14 Number of Google Review Rating (in lakhs)

The total count of Google reviews received by the attraction, denominated in lakhs (one lakh equals 100,000).

1.5.15 Best time to visit

The optimal season or time of year recommended for visiting the destination e.g., Evening, Day etc).

2 Problem Statement

2.1 Question no 1

Demonstrate various techniques for feature engineering, such as encoding categorical variables, handling missing values, creating interaction terms, and deriving new features from existing ones.

2.2 Question no 2

Perform exploratory data analysis, visualizing the distributions of key variablesand identifying outliers. Investigate the relationship between different variables ina dataset using rank, , correlation matrices and scatter plots.

2.3 Question no 4

Implement two or more clustering algorithms to group similar data points together. Visualize the resulting clusters using a scatter plot.

2.4 Question no 5

Compare and contrast different evaluation metrics for classification (e.g., accuracy, precision, recall, F1-score) and regression (e.g., mean squared error, R-squared). Discuss the strengths and limitations of each metric and when to use them.

2.5 Question no 6

Use principal component analysis (PCA)/Singular Value Decomposition (SVD) to reduce the dimensionality of a dataset and visualize the principal components.

2.6 Question no 9

Compare and contrast different types of data visualizations (e.g., histograms, scatter plots, heatmaps, boxplots etc.)

3 Methodology

3.1 Data Collection

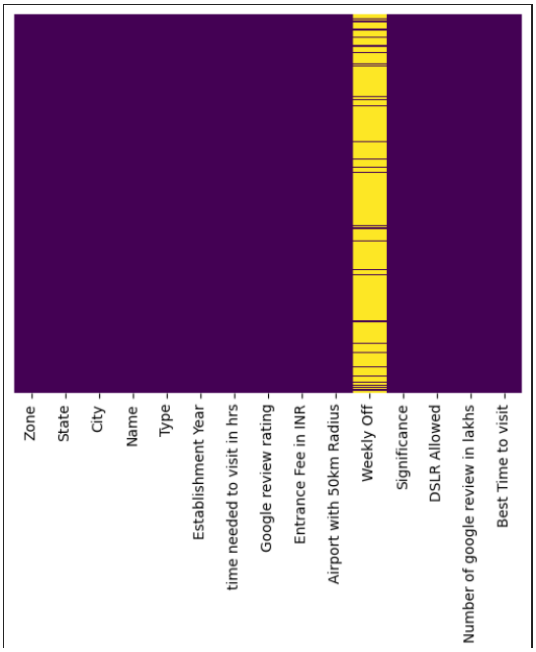
We collect data from Kaggle and do the process like loading the dataset. Convert the 'unknown' values in 'Establishment Year' to NaN and impute NaN values with median as median is less sensitive to outliers.

3.2 Data Pre-processing

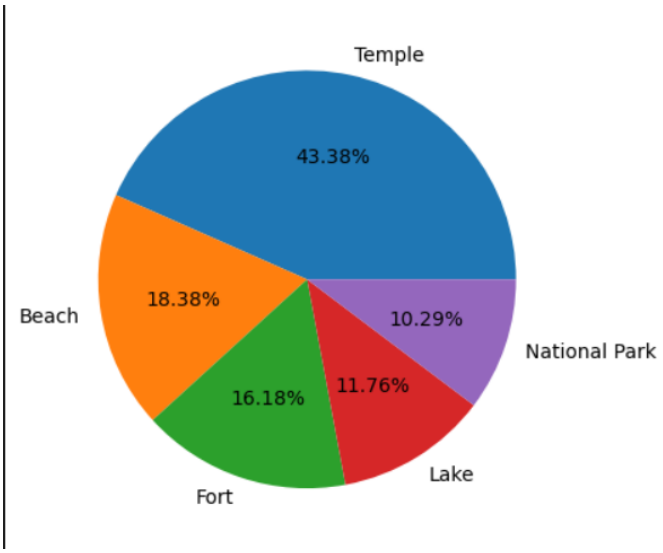
- Covert unknown values in establishment year (NaN) and impute NaN values with median.
- Used level Encoding for our cateogrical attribute like Zone, state, DSLR allowed etc
- To visualize the correlation matrix among all the attribute.

3.2.1 Exploratory data analysis(EDA)

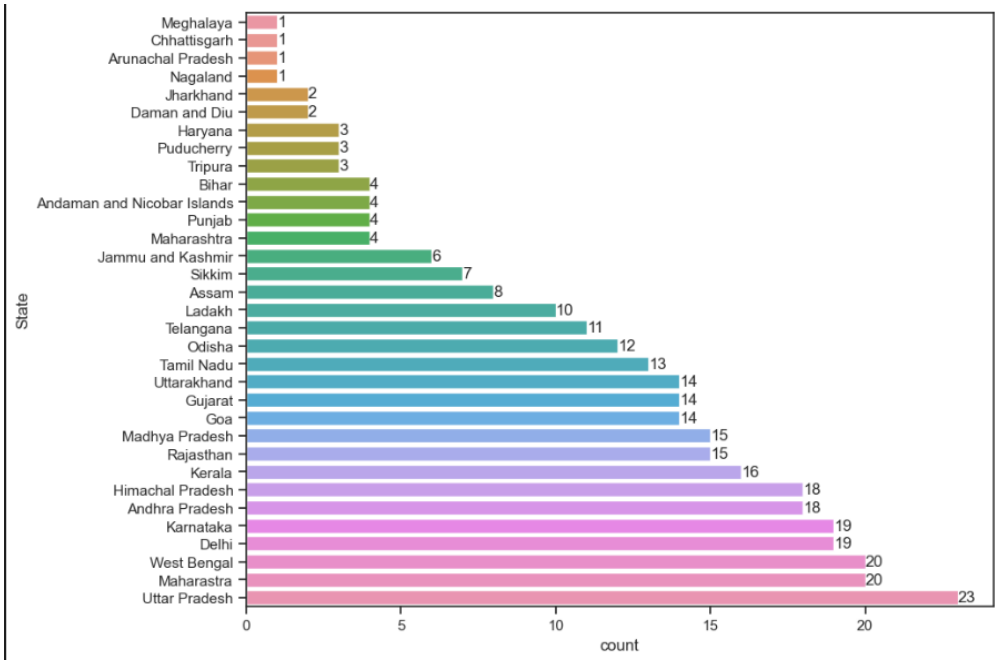
Exploratory Data Analysis (EDA): EDA is a crucial step to understand the structure and characteristics of the dataset. Techniques such as summary statistics, data visualization (e.g., histograms, scatter plots, and heatmaps), and correlation analysis can be used to explore relationships and patterns within the data.



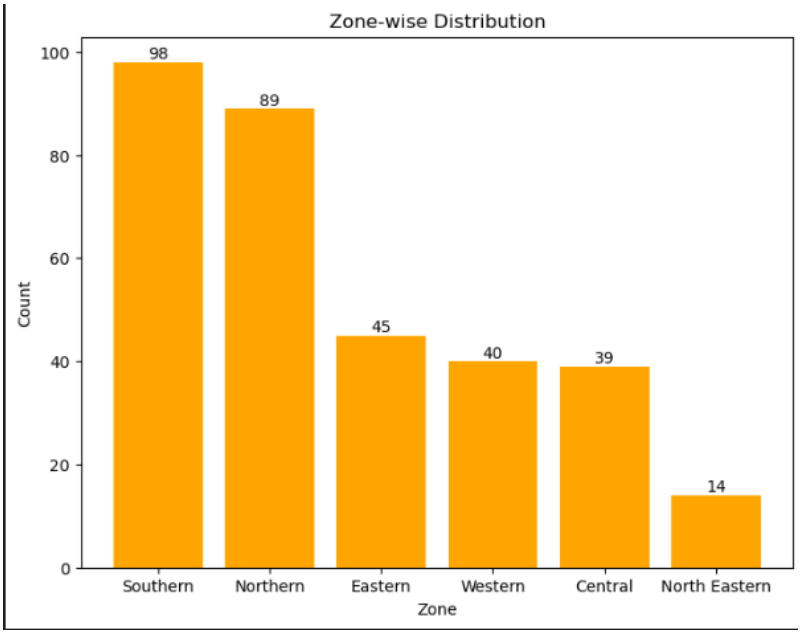
- What are the top 5 visiting places type.



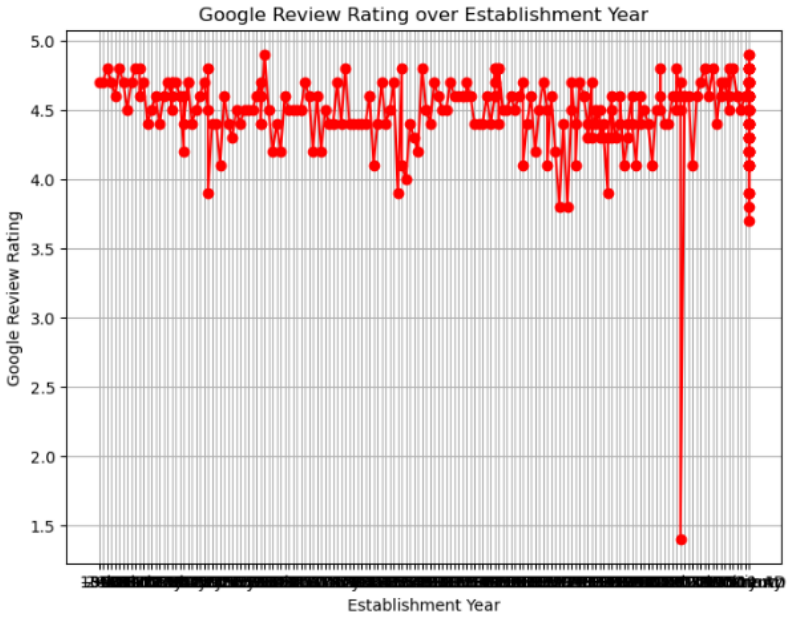
- Which is the state with most visiting places.



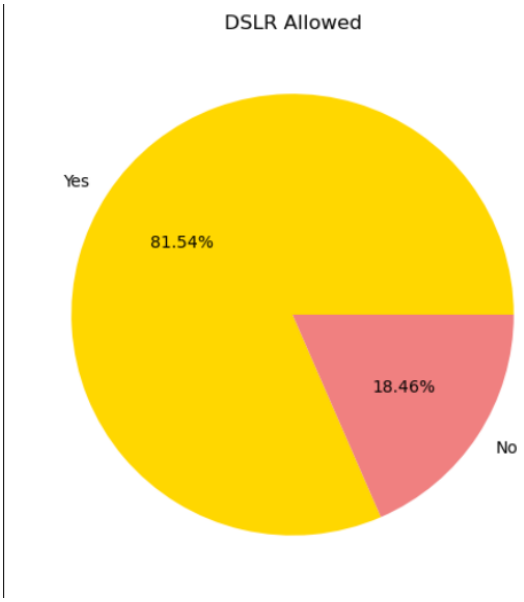
- Zone-wise distribution (bar chart).



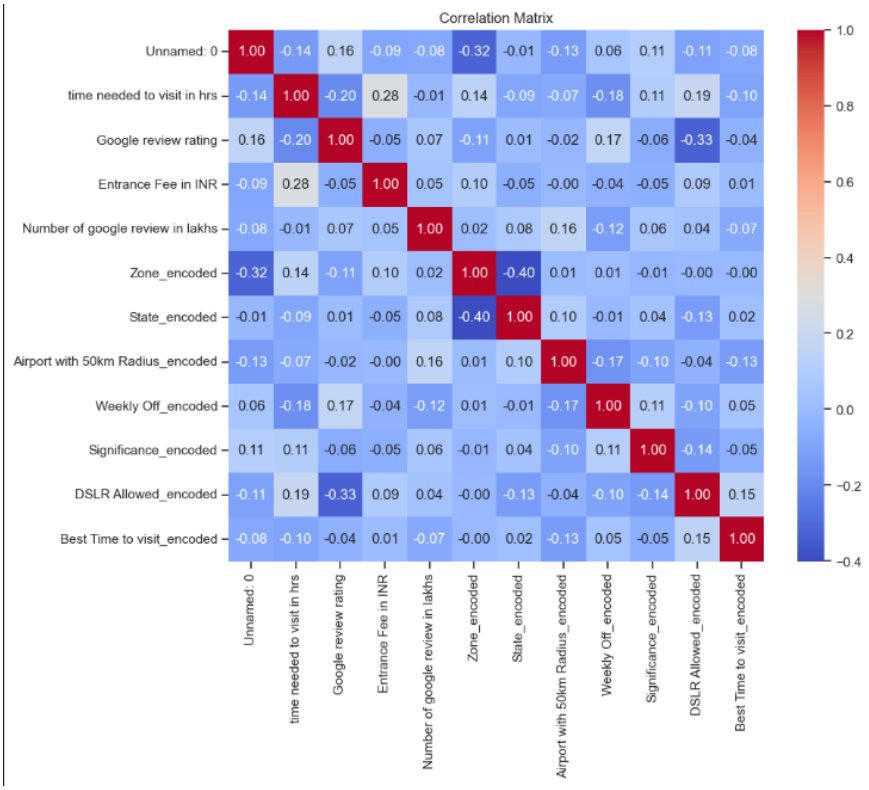
- Line chart Google Review Rating over Establishment Year.



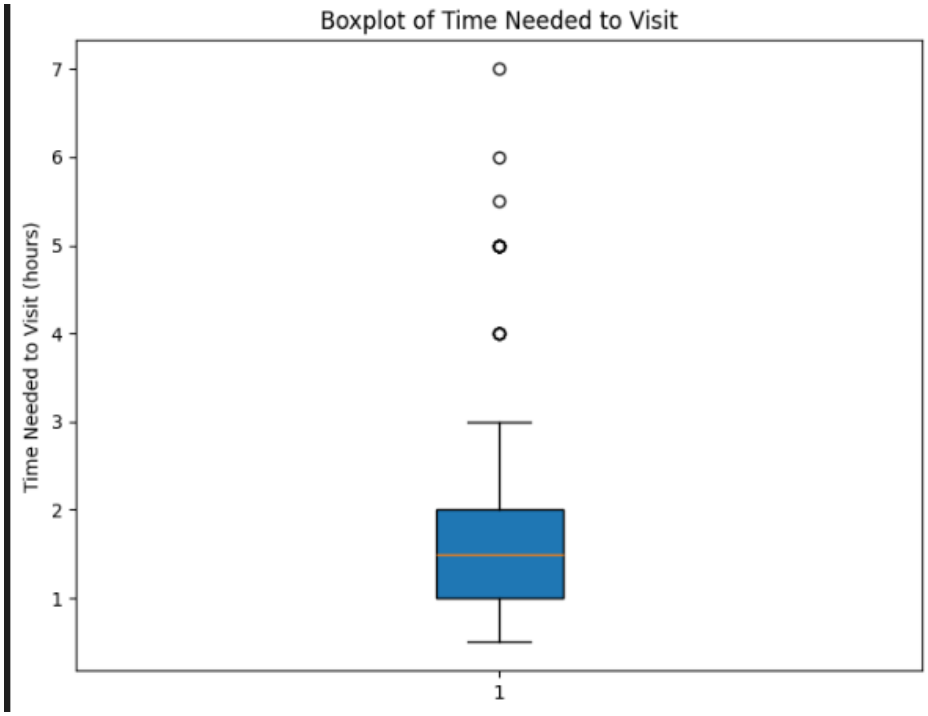
- Percentage of DSLR allowed.



- Heatmap of correlation



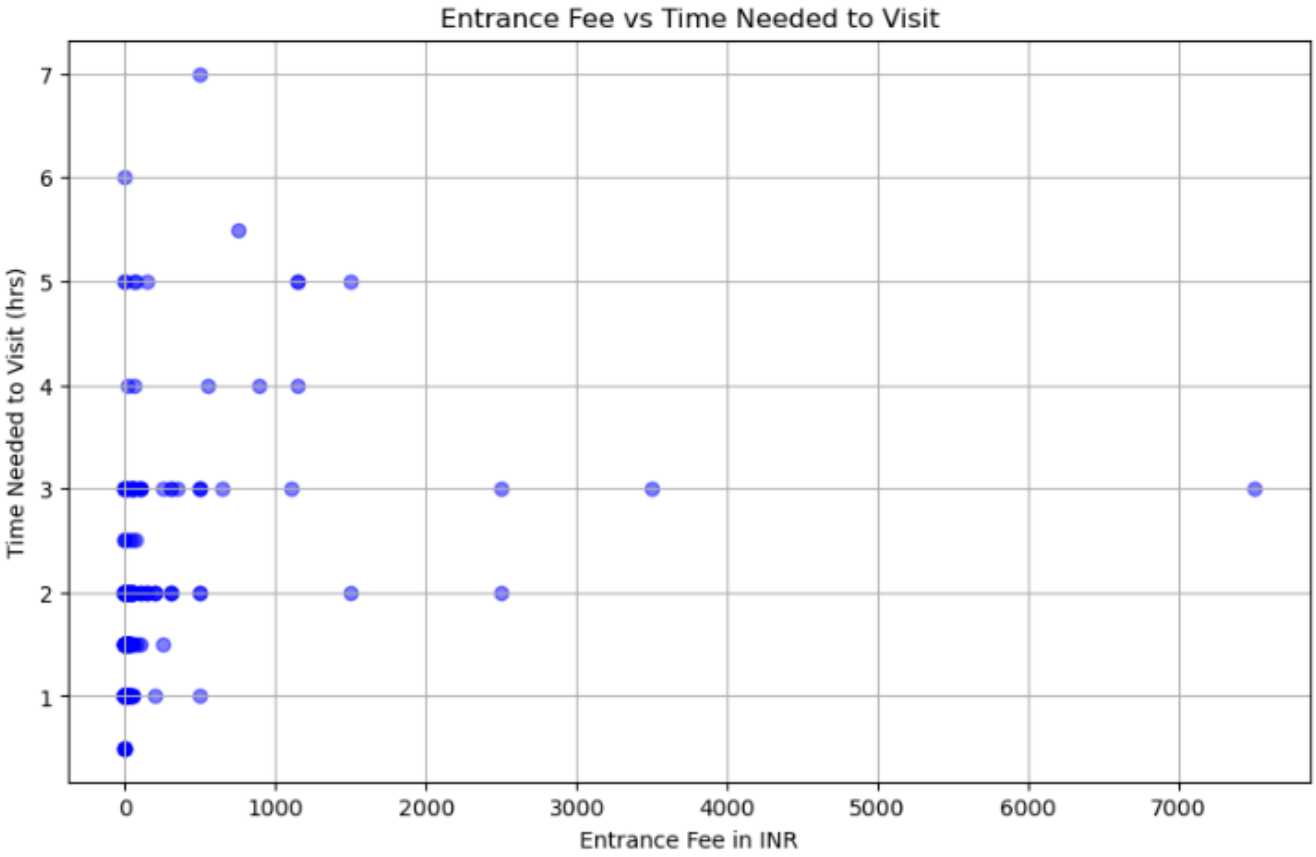
- Box plot for time needed to visit(hrs)



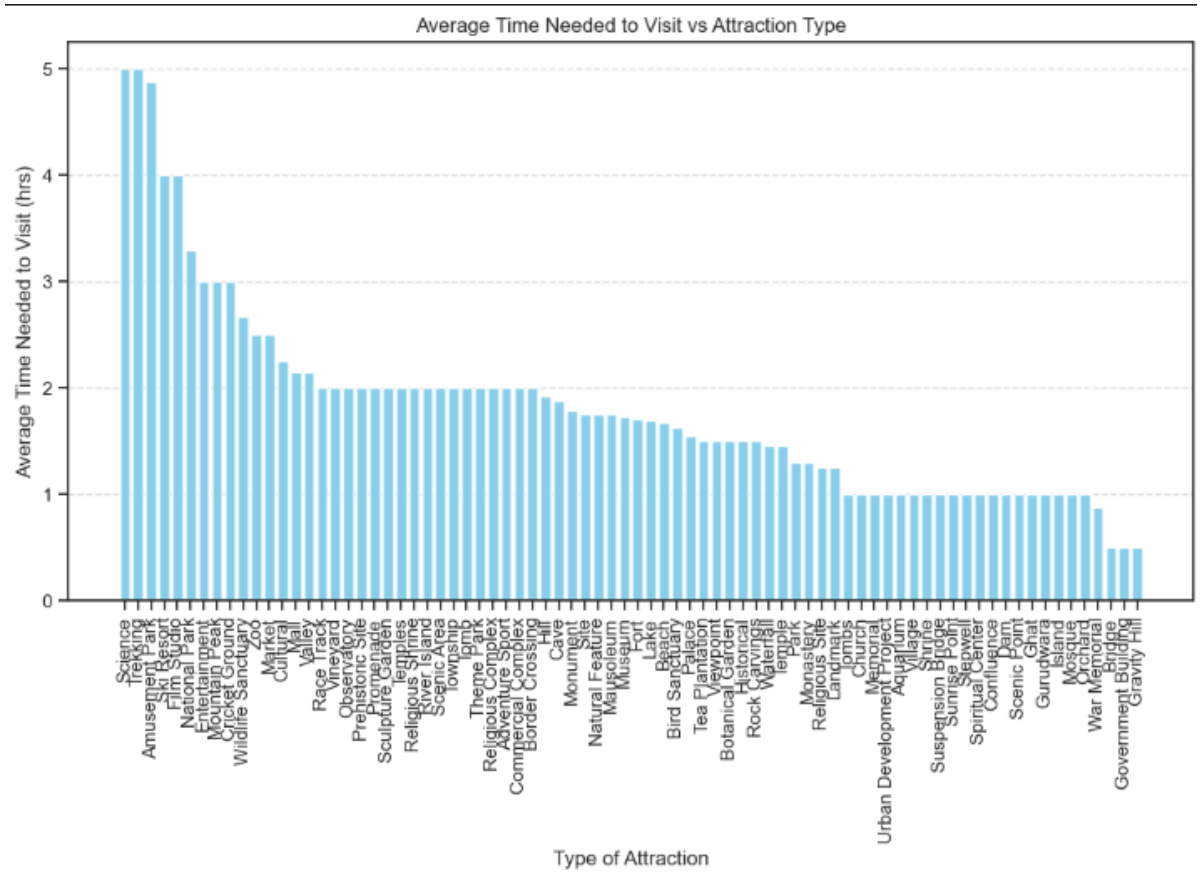
	Unnamed: 0	Zone	State	City	Name	Type	Establishment Year	time needed to visit in hrs	Google review rating	Entrance Fee in INR	...	DSLR Allowed	Number of google review in lakhs	Best Time to visit	Zone_encoded	State_encoded	Airport with 50km Radius_encoded	Weekly Off_encoded
2	2	Northern	Delhi	Delhi	Akshardham Temple	Temple	2005	5.0	4.6	60	...	No	0.40	Afternoon	3	7	1	5
15	15	Northern	Delhi	Delhi	National Science Centre	Science	1992	5.0	4.4	70	...	Yes	0.23	All	3	7	1	5
24	24	Western	Maharastra	Mumbai	Essel World	Amusement Park	1986	5.0	4.3	1149	...	Yes	0.27	All	5	19	1	5
25	25	Western	Maharastra	Mumbai	Elephanta Caves	Monument	1987	4.0	4.3	550	...	Yes	0.35	All	5	19	1	5
26	26	Western	Maharastra	Lonavala	Imagicaa	Amusement Park	2013	5.0	1.4	1149	...	Yes	0.95	All	5	19	0	1

5 rows × 23 columns

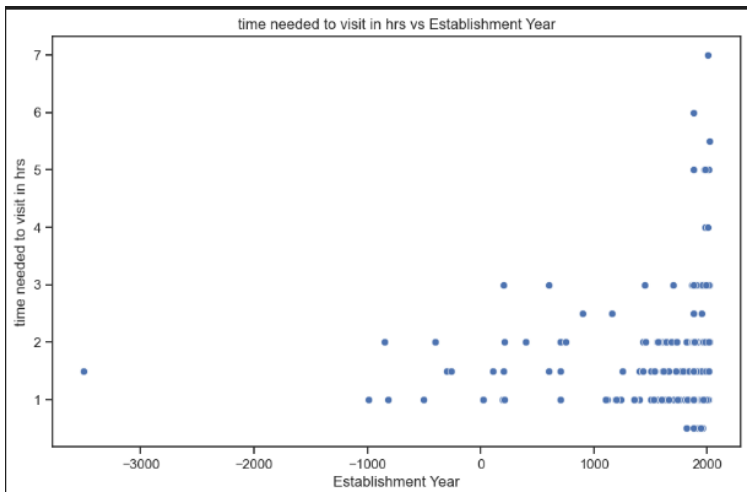
- REALTION BETWEEN ENTRANCE FEE VS TIME NEEDED TO VISIT



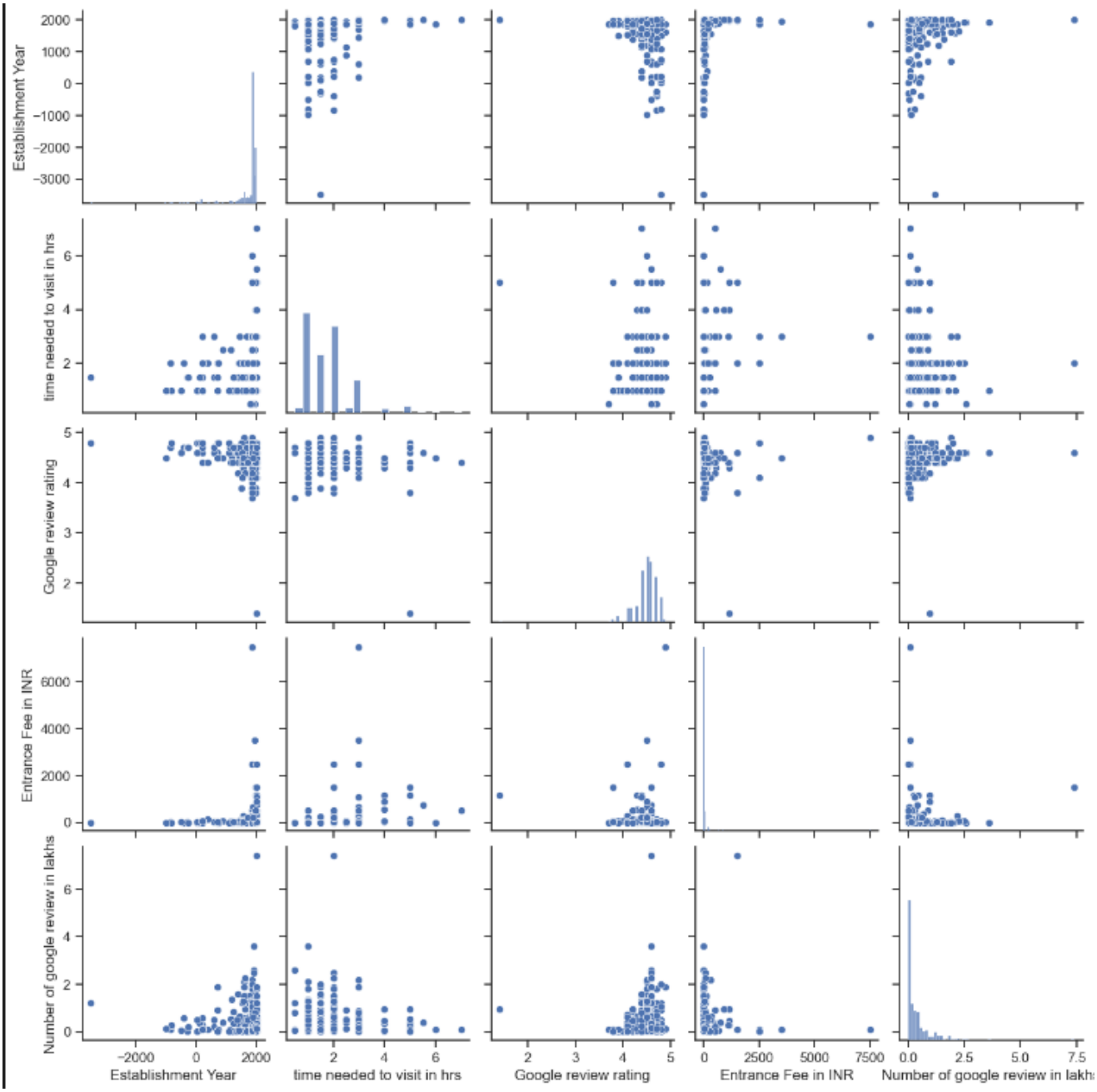
- Average Time Needed to Visit vs Attraction Type



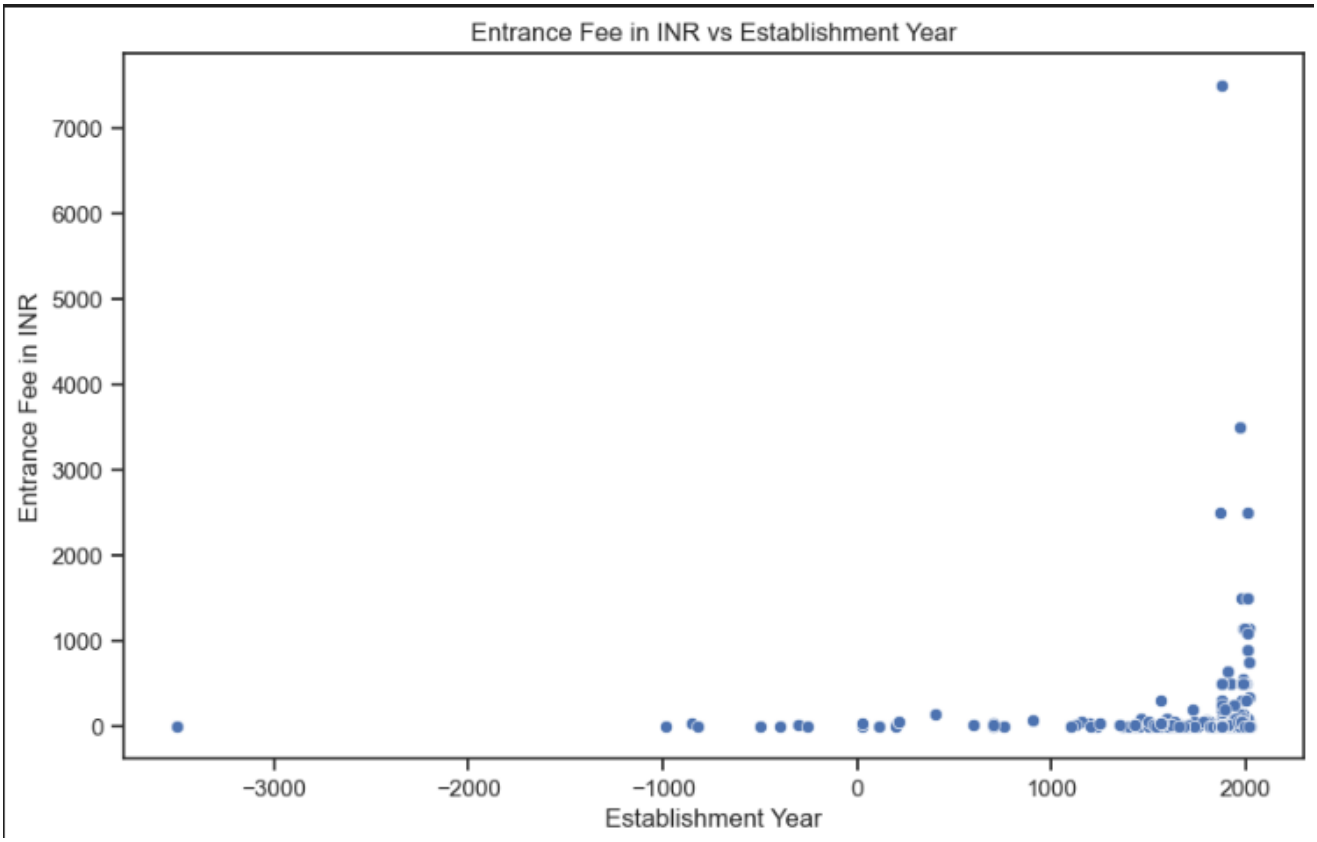
- Time need to visit vs Establishment year



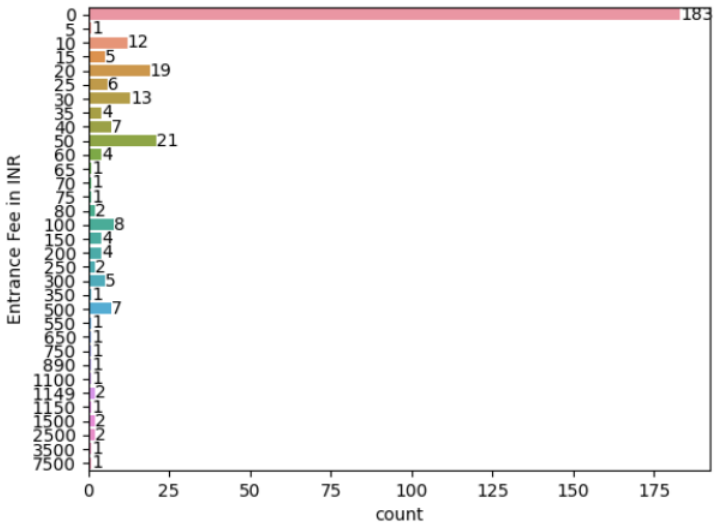
- Pair plot


Establishment Year
time needed to visit in hrs
Google review rating
Entrance Fee in INR
Number of google review in lakhs
Establishment Year
time needed to visit in hrs
Google review rating
Entrance Fee in INR
Number of google review in lakhs

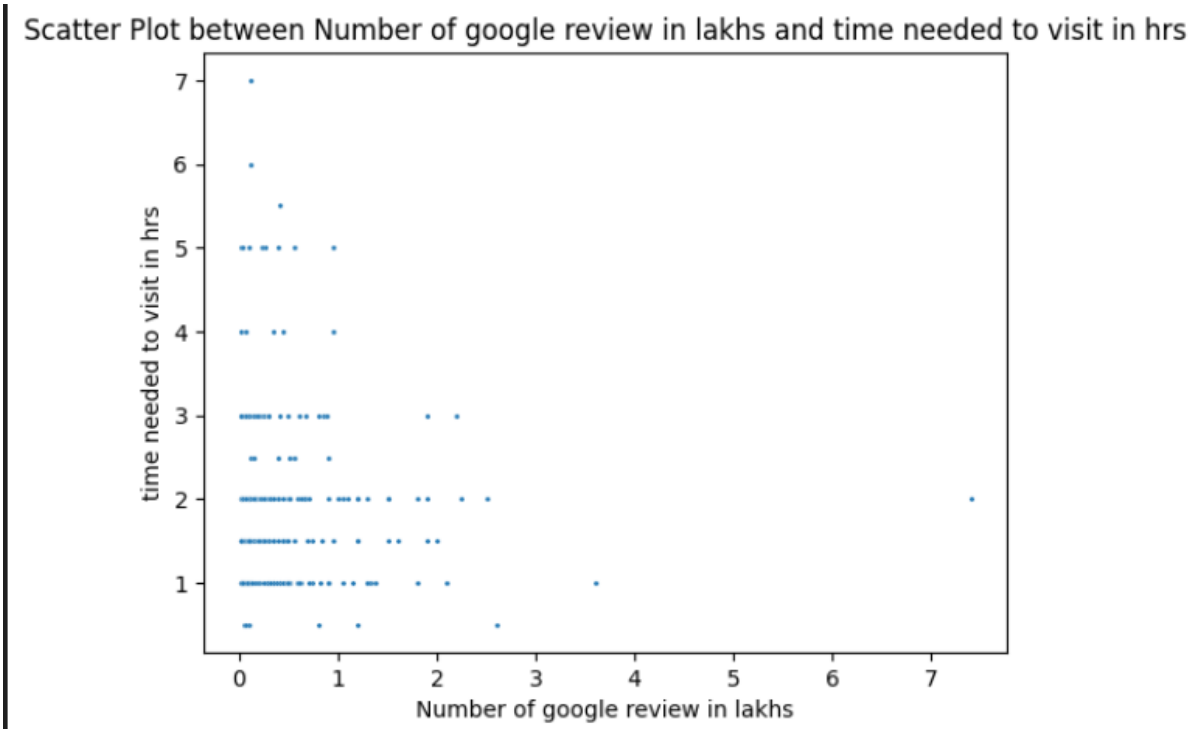
- Entrance fee in INR vs Establishment year



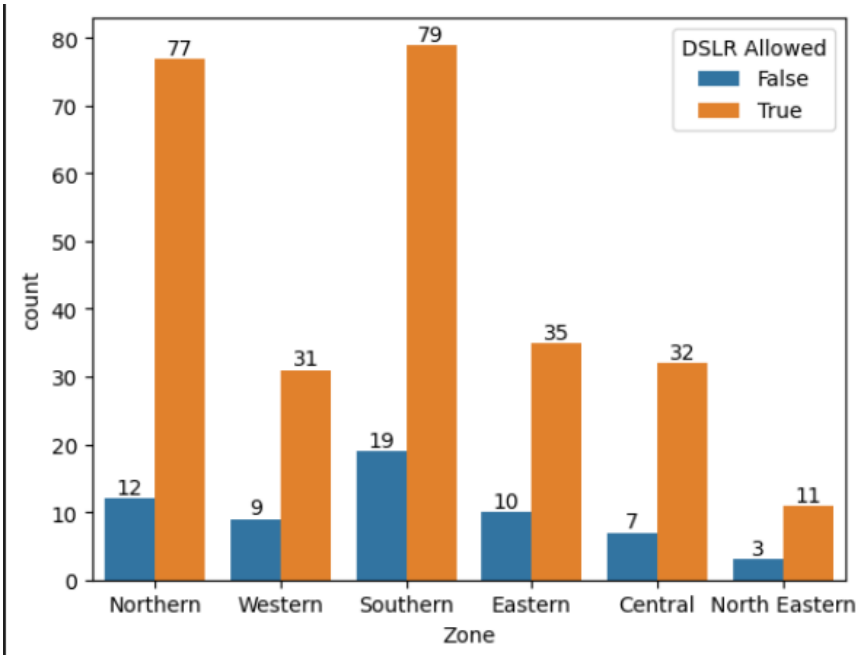
- Entrance fee in INR



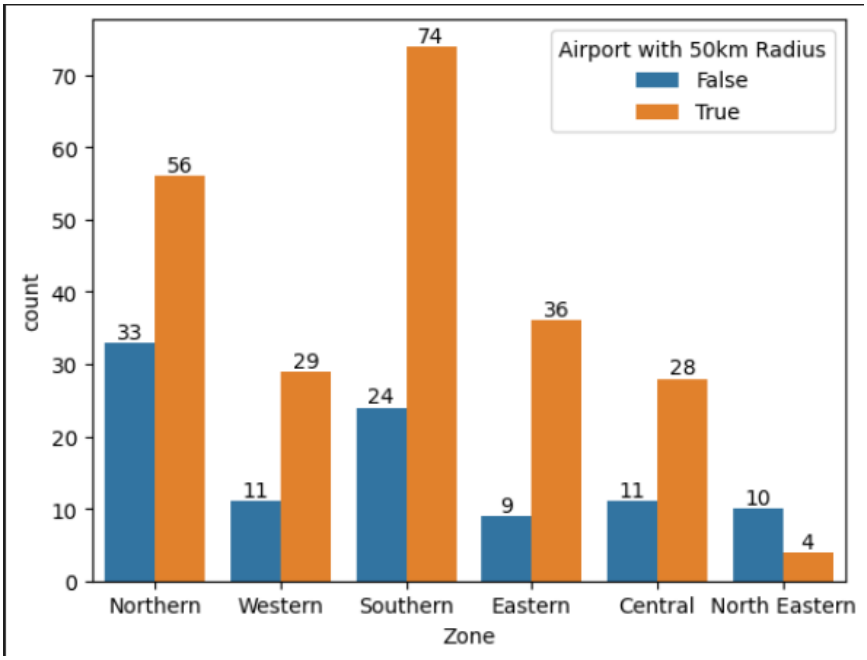
- Number of google review in lakhs vs time needed to visit in hrs



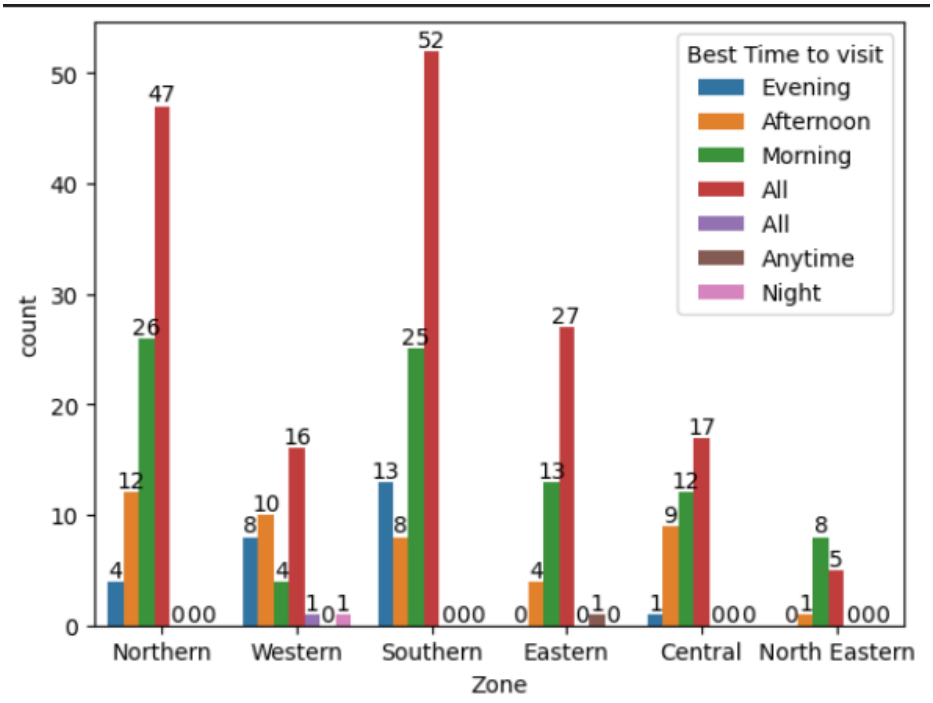
- DSLR allowed zone wise



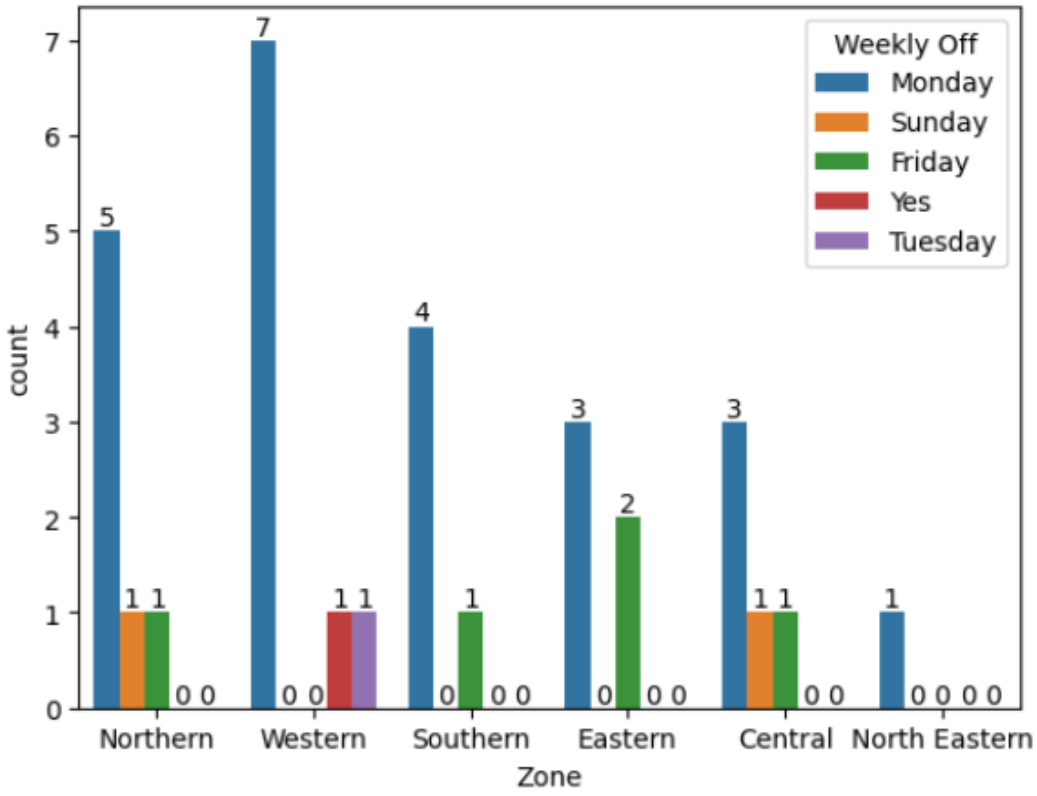
- Airport with 50km Radius Zone wise



- Best Time to visit Zone wise



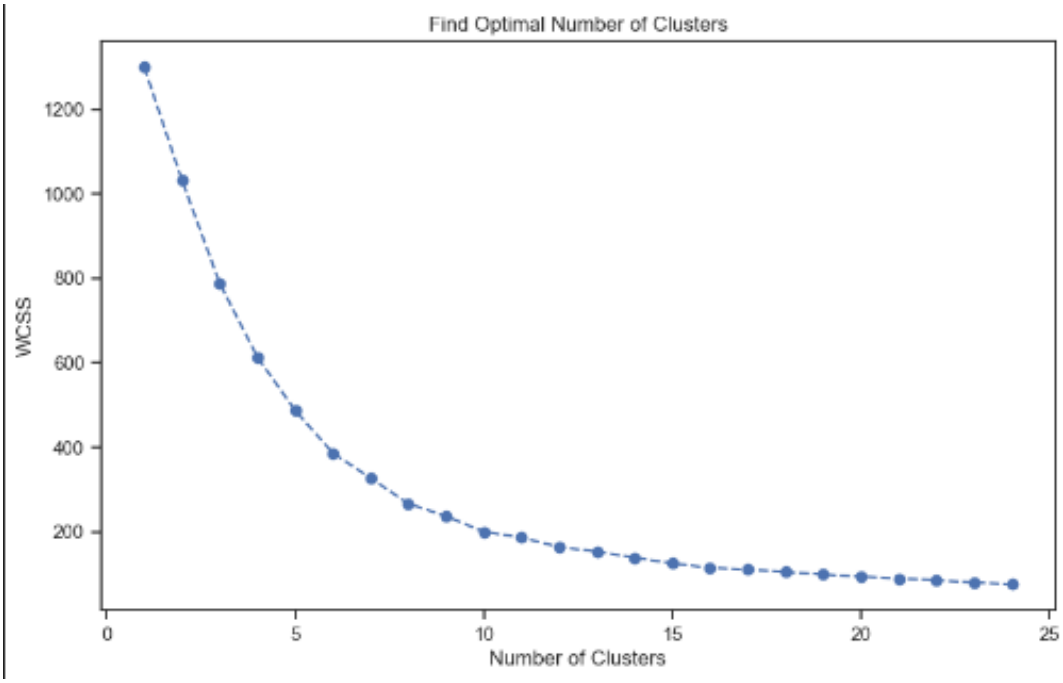
- Weekly off zone wise



3.3 Clustering

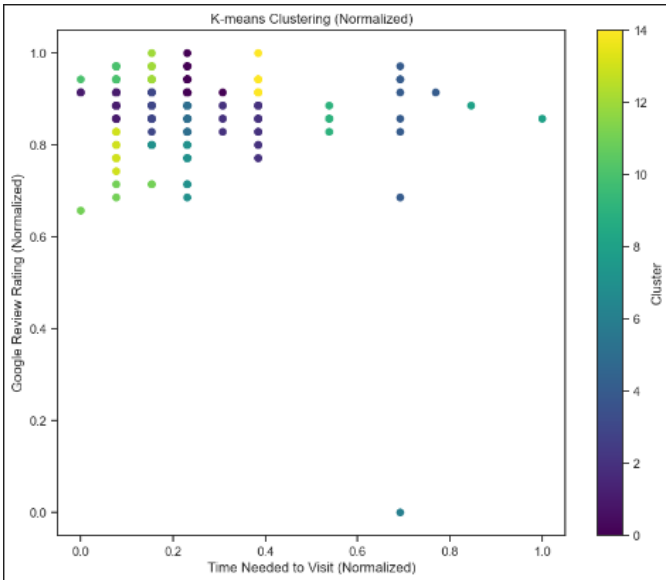
3.3.1 Elbow method to find the value of k

The elbow method is a graphical representation of finding the optimal 'K' in a K-means clustering. It works by finding WCSS (Within-Cluster Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.



3.3.2 K-means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.



3.4 Model Implementation

In our major project, a total of 2 classification models were implemented for better understanding of the dataset and the domain

3.4.1 Regression problem

- Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It is one of the simplest and most commonly used techniques in predictive modeling and data analysis.

Input: y, y_{pred} (predicted values)
Output: R^2 (coefficient of determination)
Function `CalculateR2(`y, y_{pred}`):`

```
// Calculate Mean Squared Error (MSE)
n ← length(y)
mse ← (1/n) Σ_{i=1}^{n} (y_i − y_{pred,i})^2
// Calculate Mean Squared Deviation (MSD)
ȳ ← (1/n) Σ_{i=1}^{n} y_i
msd ← (1/n) Σ_{i=1}^{n} (y_i − ȳ)^2
// Calculate R^2
r^2 ← 1 − mse/msd
return r^2
```

Algorithm 1: Calculation of R^2

After applying : Mean Squared Error: 0.41493979820207255

Mean Squared Data: 0.4168933715976331

R-squared Value: 0.004686026520580144

3.4.2 Classification

- Random Forest is a popular ensemble learning method used in both classification and regression tasks in machine learning. It belongs to the class of tree-based models and is known for its high accuracy, robustness, and ability to handle large datasets with high-dimensional feature spaces

After applying : Random forest classifier:

Accuracy: 0.6615384615384615

Precision: 0.7575757575757576

Recall: 0.6410256410256411

F1-score: 0.6944444444444444

- A decision tree is a popular and intuitive predictive modeling technique used in machine learning for both classification and regression tasks. It's a tree-like structure where each internal node represents a decision based on the value of a feature, each branch represents the outcome of the decision, and each leaf node represents the final decision or outcome.

  After applying: Decision tree:

  Accuracy: 0.5846153846153846

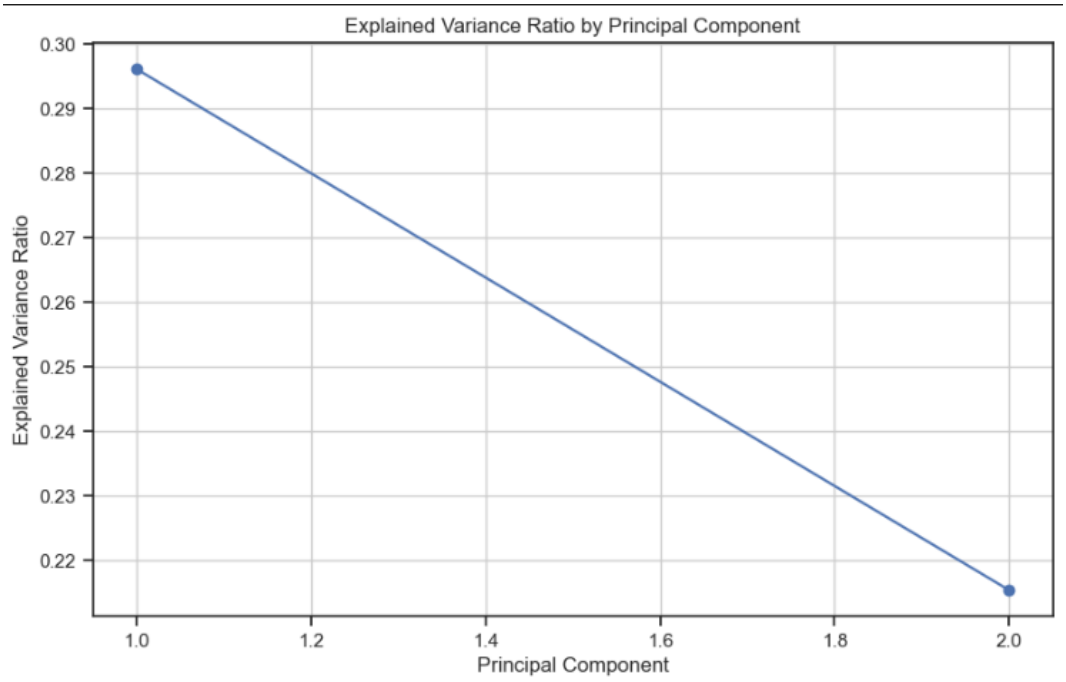
  Precision: 0.7

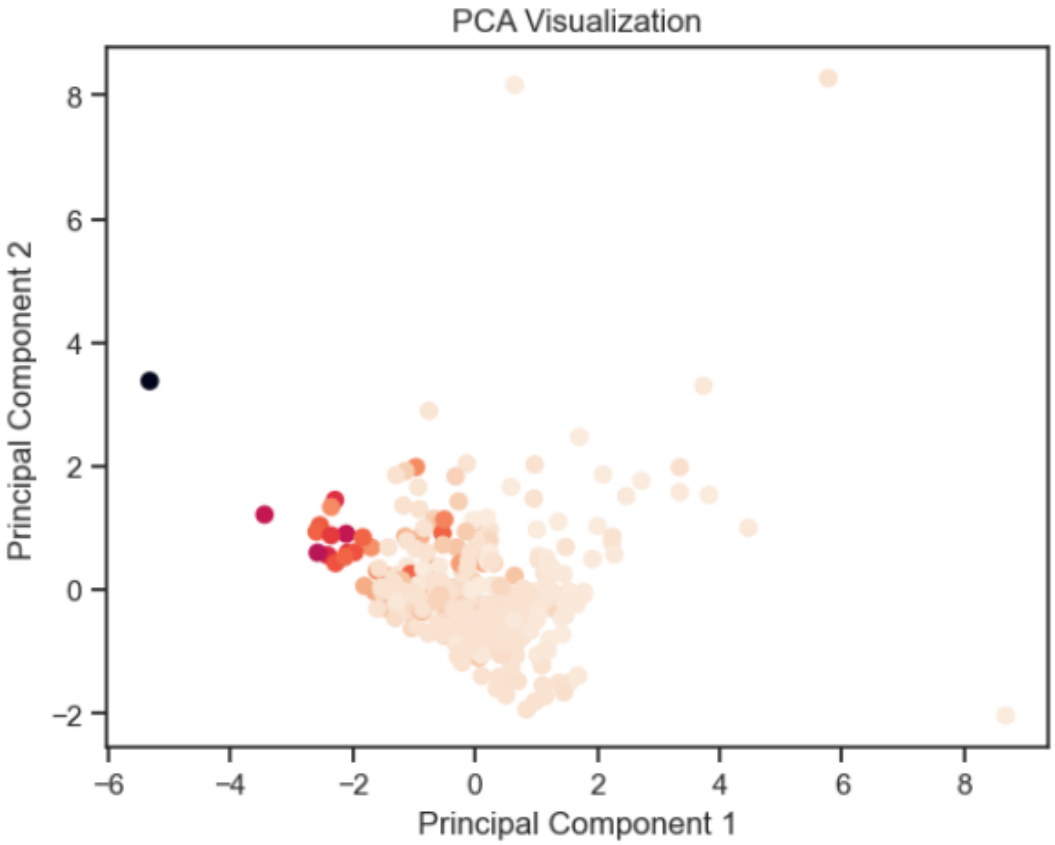
  Recall: 0.5384615384615384

  F1-score: 0.608695652173913

3.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables.PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, Principal Component Analysis (PCA) is an unsupervised learning algorithm technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.





4 Conclusion

- We found temple,beaches,fort,lake ,national park,respectively are the top 5 visiting places type
- Uttarpradesh is the state with most visiting places.
- Southern zone has the most number of visting places type
- There is nothing drop in the Google Review Rating over Establishment Year.But the year where the Google review rating drops is 2013.
- 18.46 percent of the places where dslr is not allowed and the type is temple mostly.
- Most time needed to visit is the type Science.
- In southern zone most no airport within 50km of radius
- There is no better cluster we got because of the quality of data. We visualize through PCA components to visualize those components

5 References

- https://www.kaggle.com/
- https://towardsdatascience.com/
- https://www.geeksforgeeks.org/principal-component-analysis-pca/
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://youtu.be/H99JRtDDnvk?si=$_Y dXcDY6kf2NdHGA$