Logo, company name

Description automatically generated

### Submitted by:

### Evangelin Priyanka R (22MSD7043)

### Ketan Kumar(22MSD7026)

### Janvi Kumari(22MSD7028)

**Date: 08/06/23**

### Submitted To: Prof. Manimaran Aridoss Place: Amravati

**IRT PROJECT**

**India Travel Recommendation System**

**Table of Content**

**Abstract**4

**Introduction**5

**Objective**5

**Methodology**6

**Discussion**7

* Data Collection Approach 8
* Data Types and Quantity13
* Data Pre-processing14
* Insights and Analysis17
* Model Building and Deployment20
* Validation and Accuracy27

**Conclusion**28

**References**28

**Abstract**

The "India Travel Recommendation System" project aims to develop an intelligent and personalized recommendation system to assist travellers in planning their trips to India. With its rich cultural heritage, diverse landscapes, and countless tourist attractions, India offers a wide array of options for travellers. However, the vastness of the country and the abundance of choices can often be overwhelming for tourists seeking guidance on where to go and what to do.

The India Travel Recommendation System leverages advanced machine learning algorithms and data analysis techniques to provide tailored recommendations based on users' preferences, interests, and constraints. The system takes into account various factors, such as travel duration, budget, preferred activities, and historical data on user preferences, to generate personalized travel itineraries.

To build this recommendation system, a comprehensive dataset of Indian tourist destinations, attractions, accommodations, transportation options, and user reviews will be collected and curated. Machine learning algorithms, such as collaborative filtering, content-based filtering, and hybrid approaches, will be applied to analyse this dataset and generate personalized recommendations for users.

The system will be accessible through a user-friendly web interface, where travellers can input their preferences and constraints. The recommendations provided will consider factors like popular tourist destinations, offbeat locations, historical landmarks, price, distance and culinary experiences. Additionally, the system will take into account users' feedback and ratings to continuously improve the accuracy and relevance of its recommendations.

The India Travel Recommendation System aims to enhance the travel planning experience for individuals and provide them with meaningful and enjoyable trips to India. By leveraging cutting-edge technologies and a wealth of data, this project endeavours to empower travellers with personalized recommendations that match their preferences, allowing them to explore and discover the beauty and diversity of India in an effortless and efficient manner.

**Introduction**

The "India Travel Recommendation System" project addresses the challenge of planning a trip to India by developing an intelligent and personalized recommendation system. This system assists travellers in creating customized itineraries based on their preferences, interests, and constraints. Leveraging advanced machine learning algorithms and a comprehensive dataset of Indian tourist destinations, attractions, accommodations, and user reviews, the system generates tailored recommendations that align with each traveller’s unique requirements. By providing personalized suggestions and considering factors such as travel duration, budget, and preferred activities, the India Travel Recommendation System aims to simplify the travel planning process and enhance the overall experience of exploring the diverse beauty of India.

**Objective**

The objective of this project is to develop an intelligent and personalized platform that assists travellers in planning their trips to India. This project aims to utilize advanced machine learning algorithms and a comprehensive dataset to generate tailored recommendations based on users' preferences, interests, and constraints. The main objectives include providing accurate suggestions, considering factors like travel duration and budget, creating a user-friendly interface, and incorporating user feedback for continuous improvement. Ultimately, the project seeks to enhance the travel planning experience and help travellers discover the diverse wonders of India.

**Methodology**

* **Data collection**:
* Web Scraping the sites using Selenium and Beautiful Soup
  + <https://www.indianholiday.com/>
  + <https://www.holidify.com/country/india/places-to-visit.html>
* **Data Preprocessing**:
* Verified the dataset for missing values and outliers, and treat them properly
* Used geopy to find the latitude and longitude of each place.
* **Exploratory Data Analysis(EDA)**
* Used descriptive statistics to discover new information about the dataset.
* Used various plot to understand the dataset as the dataset is not numerical.
* Identified the distances between different places using Latitude and Longitude.
* **Feature Engineering:**
* Calculated the distance, ratings, description, duplicate values etc.
* **Modeling and Evaluation:**
  + Since recommendation system is not supervised, we can split the data into train and test set to examine the performance.
  + Used WordNetLemmatizer, tokenizer, Cosine Similarity, K-nearest neighbours, K-means Clustering to recommend the destinations.
  + Accuracy and usefulness of the model can be evaluated based on the user’s feedback.
* **Materials used for analysis are:** Python, VS Code, Jupyter Notebook and Personal Laptop.

Libraries used in python are:

* Pandas: for data analysis
* Numpy: for linear algebra
* Matplotlib and seaborn: for data visualization
* scikit-learn (sklearn) : algorithms and tools for data preprocessing, model selection, and evaluation
* spacy, nltk: For NLP
* plotly.express: high-level interface for creating interactive visualizations

**Discussion**

* **Data Collection Approach:**

In order to build the "India Travel Recommendation System," a comprehensive dataset was collected from various sources. The data collection approach involved scraping information from multiple reliable sources, including travel websites, tourism boards, accommodation platforms, and user review platforms. The rationale behind selecting these sources was to ensure a wide coverage of Indian tourist destinations, attractions, accommodations, transportation options, and user reviews, providing a holistic view of the travel landscape in India. For scraping the data, we used Selenium and Beautiful Soup.

**CODE for scraping the data using** Selenium **:**

import pandas as pd

from selenium import webdriver

from time import sleep

import selenium

from selenium import webdriver

from selenium.webdriver import Chrome, ChromeOptions

from selenium.webdriver.common.by import By

from selenium.webdriver.common.keys import Keys

from selenium.common.exceptions import NoSuchElementException

from selenium.webdriver.chrome.options import Options

# from selenium.webdriver.chrome.service import Service

from webdriver\_manager.chrome import ChromeDriverManager

from selenium.webdriver.chrome.service import Service

serv\_obj=Service(r"C:\Users\Downloads\chromedriver\_win32\chromedriver.exe")

driver=webdriver.Chrome(service=serv\_obj)

driver.get('https://www.holidify.com/country/india/places-to-visit.html')

places=[]

rating=[]

xpath\_list = []

best\_time=[]

description=[]

w=[]

x=[]

y=[]

for i in range(1,5):

p=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/h3"

q=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/div/span/b"

a=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[3]"

j=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[2]"

w.append(j)

xpath\_list.append(p)

x.append(q)

y.append(a)

for i in range(6,8):

p=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/h3"

q=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/div/span/b"

a=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[3]"

j=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[2]"

w.append(j)

y.append(a)

xpath\_list.append(p)

x.append(q)

for i in range(9,13):

p=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/h3"

q=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/div/span/b"

a=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[3]"

j=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[2]"

w.append(j)

y.append(a)

x.append(q)

xpath\_list.append(p)

for i in range(14,46):

p=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/h3"

q=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/div/span/b"

a=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[3]"

j=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[2]"

w.append(j)

y.append(a)

x.append(q)

xpath\_list.append(p)

for xpath in xpath\_list:

p=driver.find\_element(By.XPATH,xpath).text.strip()

places.append(p)

for xpath in x:

z=driver.find\_element(By.XPATH,xpath).text.strip()

rating.append(z)

for xpath in y:

b=driver.find\_element(By.XPATH,xpath).text.strip("Best Time:")

best\_time.append(b)

for xpath in w:

b=driver.find\_element(By.XPATH,xpath).text.strip("Best Time:")

description.append(b)

driver.find\_element(By.XPATH,"/html/body/div[2]/div/div[1]/div[5]/nav/ul/li[5]/a").click()

xpath\_li=[]

x\_l=[]

y\_l=[]

w\_l=[]

for i in range(1,43):

p=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/h3"

q=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/div/span/b"

a=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[3]"

j=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[2]"

w\_l.append(j)

y\_l.append(a)

xpath\_li.append(p)

x\_l.append(q)

for xpath in xpath\_li:

p=driver.find\_element(By.XPATH,xpath).text.strip()

places.append(p)

for xpath in x\_l:

z=driver.find\_element(By.XPATH,xpath).text.strip()

rating.append(z)

for xpath in y\_l:

b=driver.find\_element(By.XPATH,xpath).text.strip("Best Time:")

best\_time.append(b)

for xpath in w\_l:

b=driver.find\_element(By.XPATH,xpath).text.strip("Best Time:")

description.append(b)

driver.find\_element(By.XPATH,"/html/body/div[2]/div/div[1]/div[5]/nav/ul/li[5]/a").click()

xpath\_l=[]

xi=[]

yi=[]

wi=[]

for i in range(1,17):

p=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/h3"

q=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/a/div/span/b"

a=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[3]"

j=f"/html/body/div[2]/div/div[1]/div[5]/div[{i}]/div/div[1]/p[2]"

wi.append(j)

yi.append(a)

xi.append(q)

xpath\_l.append(p)

for xpath in xpath\_l:

p=driver.find\_element(By.XPATH,xpath).text.strip()

places.append(p)

for xpath in xi:

z=driver.find\_element(By.XPATH,xpath).text.strip()

rating.append(z)

for xpath in yi:

b=driver.find\_element(By.XPATH,xpath).text.strip("Best Time:")

best\_time.append(b)

for xpath in wi:

b=driver.find\_element(By.XPATH,xpath).text.strip("Best Time:")

description.append(b)

stripped\_places\_list=[]

for city in places:

stripped\_places = city.split(". ")[1]

stripped\_places\_list.append(stripped\_places)

*# Create a dictionary with column names as keys and lists as values*

data = {'Place': stripped\_places\_list, 'Rating': rating, 'Best Time to Visit': best\_time, 'Description': description}

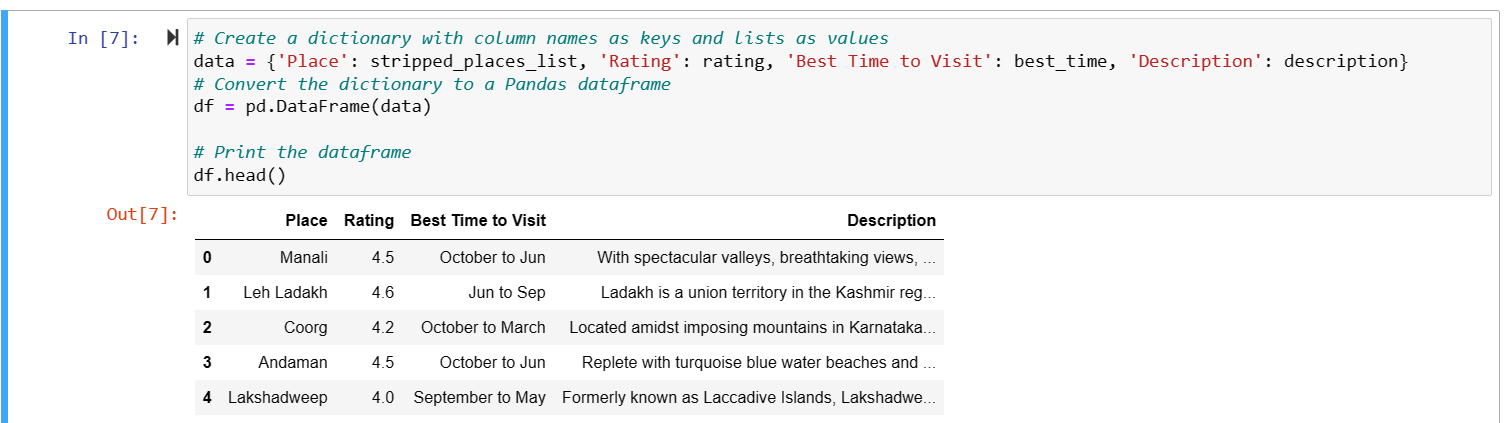
*# Convert the dictionary to a Pandas dataframe*

df = pd.DataFrame(data)

*# Print the dataframe*

df.head()

**OUTPUT:**

****

**CODE for scraping the data using Beautiful** Soup **:**

import requests

from bs4 import BeautifulSoup

URL="https://www.holidify.com/places/amritsar/"

r=requests.get(URL)

soup=BeautifulSoup(r.content,'html5lib')

tagline=soup.find('h3').text

tagline

desc=soup.find("div",{"class":"readMoreText"})

desc=desc.find("p")

desc=desc.text

desc

import requests

from bs4 import BeautifulSoup

URL = "https://www.holidify.com/places/wayanad/"

response = requests.get(URL)

soup = BeautifulSoup(response.text, "html.parser")

soup = BeautifulSoup(response.text, "html.parser")

rating\_section = soup.find("div", {"class": "rating-section"})

ratings = {}

for rating in rating\_section.find\_all("div", {"class": "rating"}):

category = rating.find("div", {"class": "category"}).text.strip()

value = rating.find("div", {"class": "value"}).text.strip()

ratings[category] = value

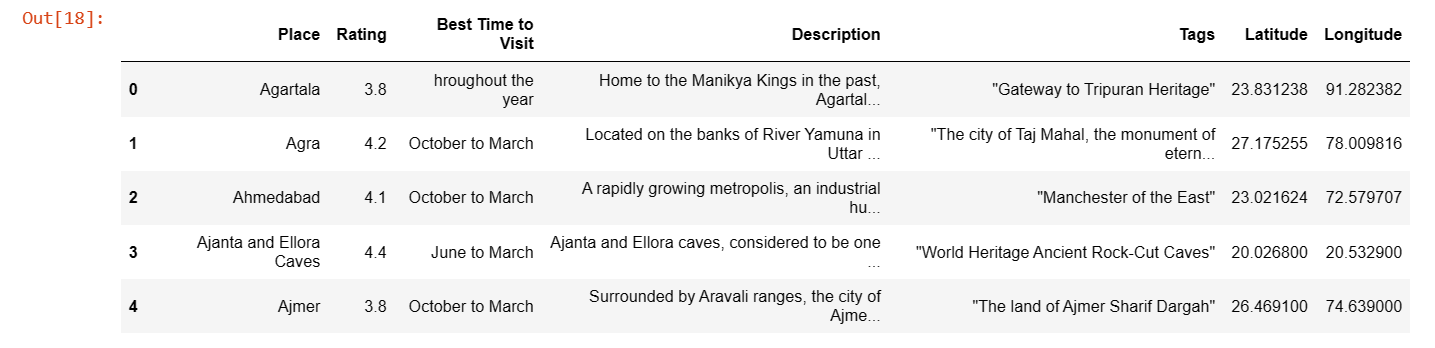
print(ratings)

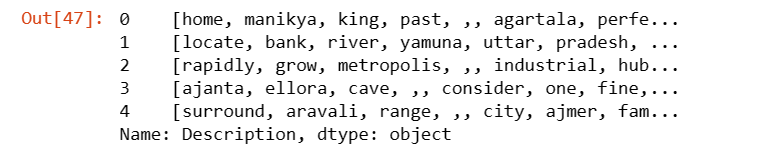
* **Data Types and Quantity:**

The collected dataset comprises diverse types of data, including information on tourist destinations, popular attractions, places, best time to visit, description of each place, and user reviews and ratings. The dataset encompasses a substantial quantity of records, with thousands of entries obtained from each source. This significant volume of data ensures a rich and extensive foundation for generating recommendations for travellers. But the major task here was to pre-process the data and extract meaningful features from the data.

* **Data Pre-processing:**

To ensure the quality and consistency of the dataset, a series of pre-processing steps were applied. These steps involved removing duplicates, handling missing values, standardizing data formats, and resolving any inconsistencies or discrepancies. Additionally, data cleaning techniques were employed to address common challenges encountered during the scraping process, such as variations in naming conventions or data formatting across different sources. By meticulously pre-processing the data, the resulting dataset became more reliable and suitable for analysis. During this process various meaningful features were too extracted like Latitude and Longitude of all places, checking similarities between different places etc.

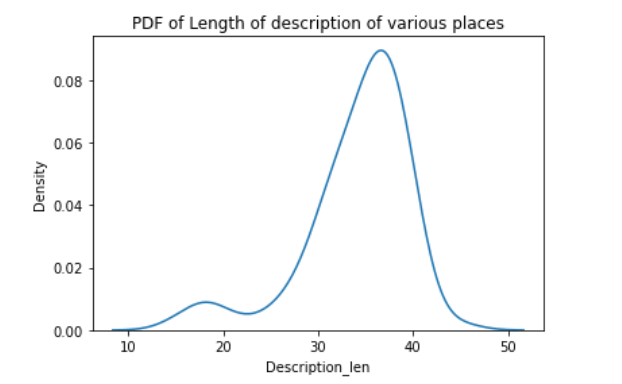
****



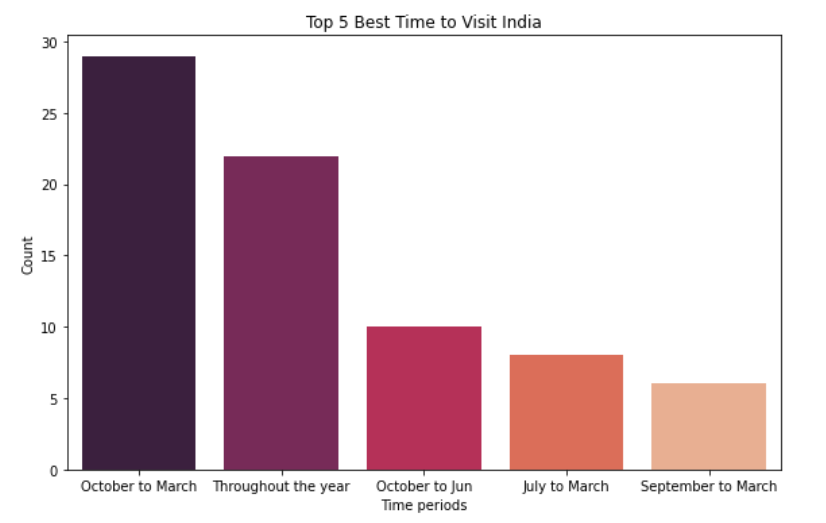
* **Insights and Analysis:**

The analysis of the scraped data revealed valuable insights about the Indian travel landscape. By examining the dataset, patterns and trends regarding popular tourist destinations, frequently recommended attractions, commonly mentioned accommodations, preferred transportation options, and sentiments expressed in user reviews were identified. For instance, the analysis highlighted the popularity of cultural heritage sites like the Taj Mahal, the allure of beach destinations such as Goa, and the high ratings received by boutique hotels and homestays in specific regions. These insights provide a deeper understanding of the preferences and choices of travellers in India.

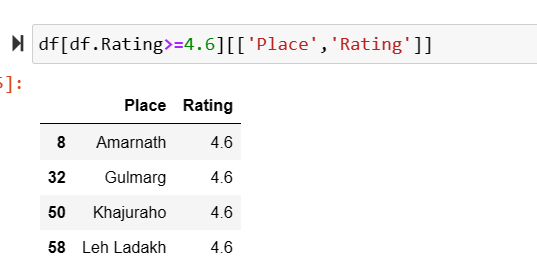
**Checking the distribution of length of the description:**



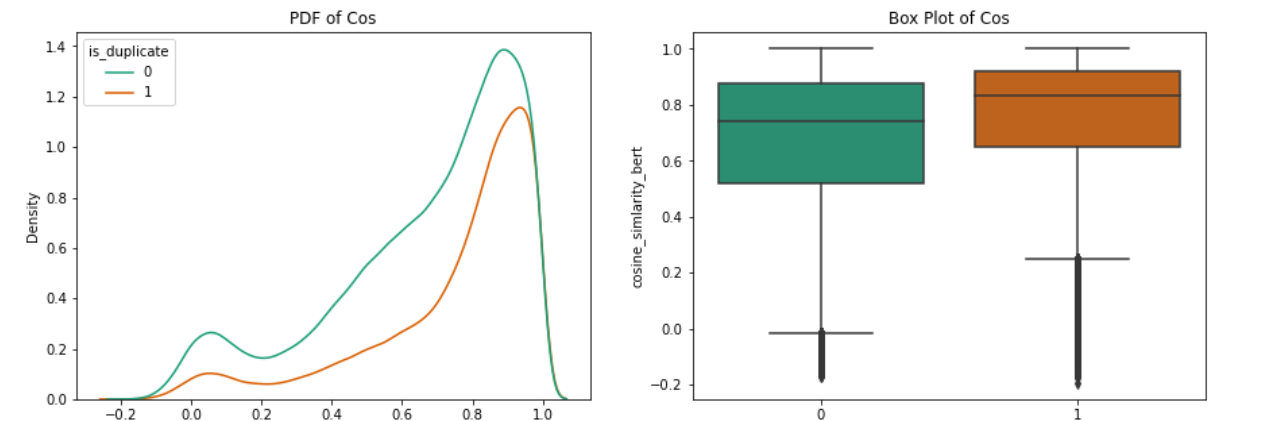
**Top 5 Best time to visit India:**



**Places in India with Rating greater that 4.5**

****

**Plotting the cosine similarity of two places based on description:**

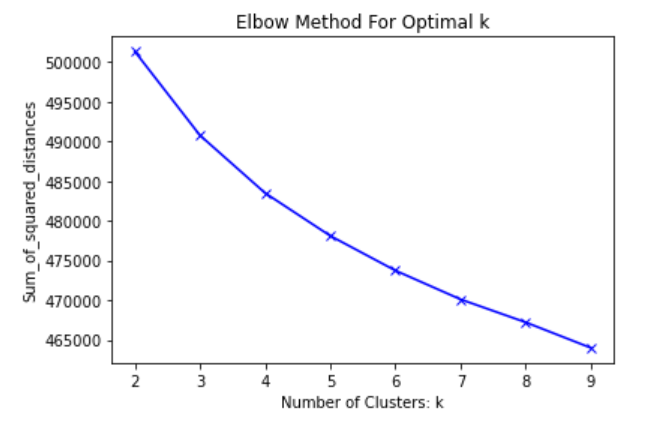


* **Model Building and Deployment:**

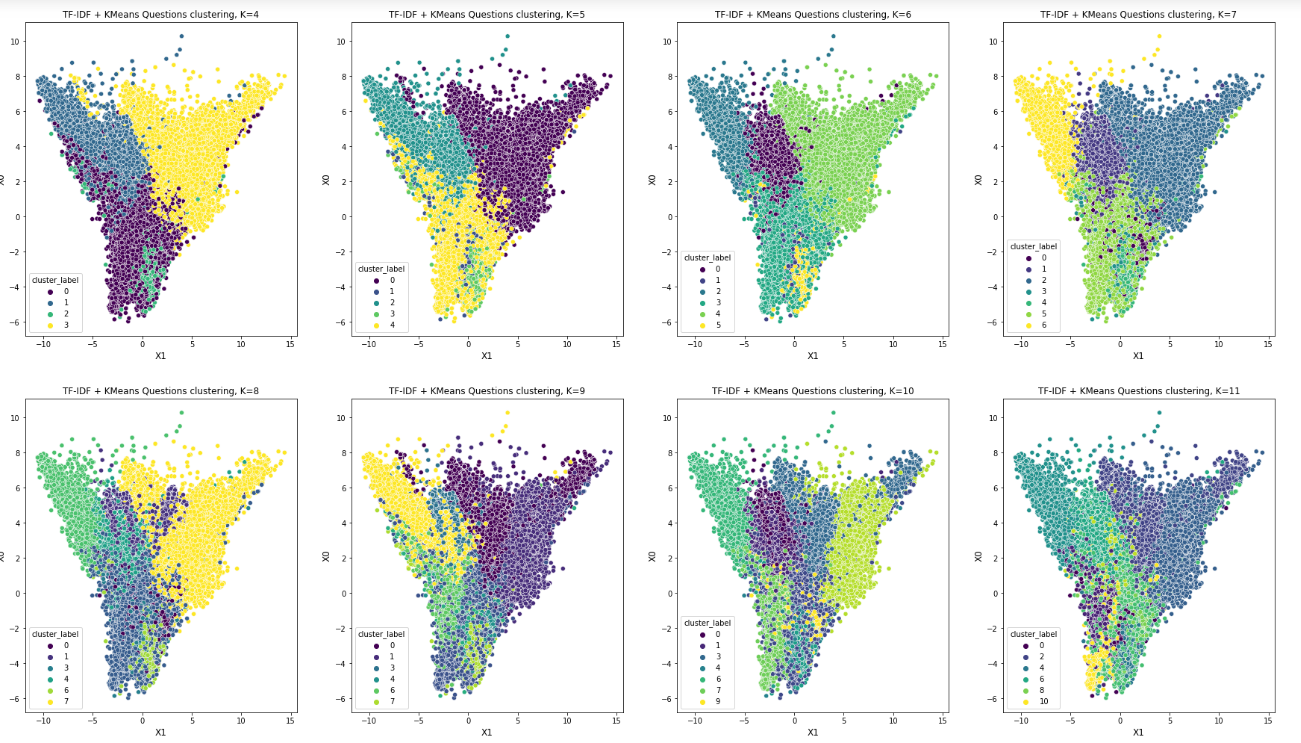
Since we have scraped the data, pre-processed the data, extracted meaningful features from the data, seen insights from the data, Now, it is time for recommendations

**Recommendation based on K-means Clustering:**

In this, we would recommend places that are near to each other based on their Latitude and Longitude



For more clarity, lets see all the clusters

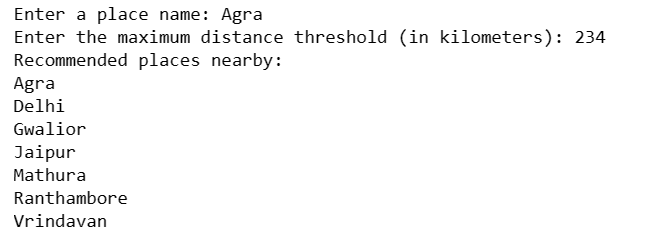


From this, we can say that the optimal value for k is 5.

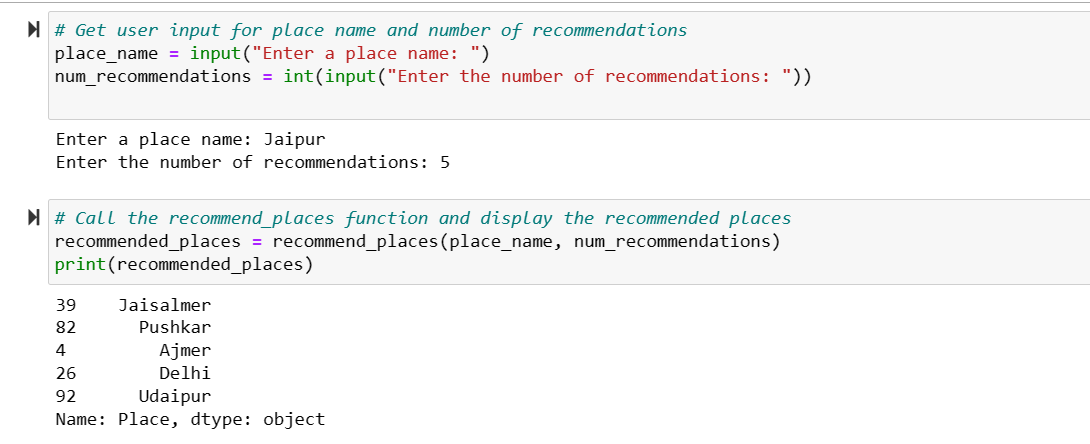
**Word Cloud for various Clusters:**



**Recommendations:**

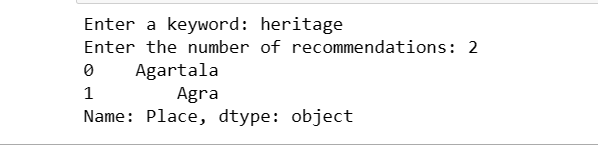
****

**Recommend similar places based on description:**

****

**Recommendation based on Keywords:**

In our day-to-day life, many people use just keywords to get suggestions about various things. For example, if it’s a movie then people will select movie based on its genre, similarly if it’s a place, people would search whether its haunted or is historical or it has heritage. Som to sum it up recommendations based on Keywords are important for any recommendation system.

****

* **Validation and Accuracy:**

To ensure the accuracy and reliability of the scraped data, a validation process was implemented. This process involved cross-referencing the collected data with reputable sources, comparing information across multiple platforms, and conducting manual verification where necessary. By employing these validation techniques, the accuracy and trustworthiness of the dataset were enhanced. However, it is important to note that due to the dynamic nature of online information, occasional inaccuracies or discrepancies may still be present.

**Conclusions**

The project's conclusion highlights the following key points:

1. **Personalized Recommendations**: The developed recommendation system leverages advanced machine learning algorithms and user preferences to provide personalized travel suggestions. By considering factors such as travel duration, budget, and preferred activities, the system offers tailored itineraries that align with individual travellers’ needs.
2. **Data Collection and Pre-processing**: The scraping process yielded a substantial dataset encompassing diverse information about Indian tourist destinations, attractions, accommodations, transportation options, and user reviews. Through rigorous pre-processing techniques, the dataset was cleaned and standardized, enhancing its quality and reliability.
3. **Insights and Analysis**: The analysis of the collected data revealed valuable insights into the Indian travel landscape. Patterns and trends regarding popular destinations, attractions, accommodations, and transportation preferences were identified, enabling a deeper understanding of travellers’ choices and preferences.
4. **Validation and Accuracy**: To ensure the accuracy and reliability of the dataset, a validation process was implemented, including cross-referencing with reputable sources and manual verification. While efforts were made to address potential inaccuracies, it is important to acknowledge that occasional discrepancies may still exist due to the dynamic nature of online information.
5. **Future Directions**: The project lays the foundation for future enhancements and expansions. The dataset can potentially be made available for further research and applications, subject to privacy and licensing considerations. Additionally, future iterations may involve incorporating real-time user-generated data to continuously refine and enrich the recommendation system.

### References

1. <https://www.geeksforgeeks.org/collaborative-filtering-ml/>
2. <https://www.indianholiday.com/>
3. <https://www.holidify.com/country/india/places-to-visit.html>
4. <https://towardsdatascience.com/recommendation-systems-explained-a42fc60591ed>
5. <https://www.india.gov.in/topics/travel-tourism>
6. <https://www.lonelyplanet.com/>

### Contribution

### Ketan Kumar (22MSD7026) - Data preprocessing, Feature Extraction , Recommendation based on Latitude

### And Longititude (near by places will be suggested ) , documentation.

### Janvi Kumari (22MSD7028) – WebScrapping using selenium , recommendation based on similarity of places based on description, documentation.

### Priyanka Bose(22MSD7043) - WebScrapping using Beautiful Soup , recommendation based on keywords (like hill , beach.. etc), documentation.