Capstone Project - Travel Insight

Akrant Varshney

12/17/2024

Contents

1 *Introduction*

This project helps in analyzing seasonal visitation trends in Indian tourist destinations. The study seeks to understand which places in India are more frequently visited during specific seasons. The project combines data analysis techniques with web scraping to create a comprehensive tourism recommendation system. By leveraging various data sources, including existing databases and travel websites the study aims to provide valuable insights for both tourists and the tourism industry. The project's outcomes are expected to contribute to the field of data science and have practical implications for tourism planning and management

2 *Objectives and Methodology*

The primary objective of this study is to develop a robust system that assists tourists in selecting the optimal month to visit specific Indian states. This system aims to analyze and interpret vast amounts of tourism data to identify seasonal visitation patterns across various destinations in India. By understanding these trends, the project seeks to provide personalized recommendations to travelers, enhancing their trip planning experience. Additionally, the study aims to investigate the factors influencing these seasonal tourism patterns, such as weather conditions, local festivals, and cultural events, to offer a comprehensive view of each destination's appeal throughout the year.

The methodology employed in this project revolves around extensive data collection and analysis. The primary sources of data are Google Maps and official Indian tourism websites. Google Maps provided a wealth of information through user reviews, ratings, and check-ins, offering insights into visitor preferences and peak seasons for various locations. The Indian tourism websites, both at the national and state levels, contributed official statistics, event calendars, and promotional information that helped in understanding the tourism landscape of each region. This data was extracted using web scraping techniques, ensuring a large and diverse dataset. The collected information was then processed, cleaned, and analyzed using various data science tools and statistical methods to identify patterns and trends in seasonal tourism across different Indian states.

3 Data Analysis

Data collection for this project is extensive and draws from multiple carefully selected sources to ensure a comprehensive understanding of tourism patterns in India. The primary sources of data include the official Indian tourism website, Google Maps, and Wikipedia. These platforms were chosen for their reliability, breadth of information, and user-generated content, which provide both official statistics and real-world experiences of travelers.

The official Indian tourism website serves as a rich repository of information, offering detailed insights into various destinations, seasonal events, and visitor statistics. This source provides authoritative data on tourist attractions, cultural festivals, and recommended travel periods for different regions across India. The information gathered from this platform forms the backbone of our understanding of the official tourism landscape and promotional efforts by the Indian government.

Google Maps has been instrumental in providing location-based data and user reviews. The platform's extensive database of points of interest, coupled with its review and rating system, offers valuable insights into visitor experiences and preferences. By analyzing the temporal patterns of reviews and ratings, we can discern seasonal trends in visitor satisfaction and popularity for various destinations. The geospatial data from Google Maps also allows for a more nuanced understanding of regional tourism patterns and the relationship between different attractions.

Wikipedia serves as a supplementary source of information, providing historical context, geographical details, and general information about various tourist destinations in India. While not used for primary data collection, Wikipedia helps in corroborating information and filling in gaps in our understanding of certain locations or cultural events that may influence tourism patterns.

The data analysis process involves a combination of quantitative and qualitative methods. Quantitative analysis focuses on visitor numbers, ratings, and seasonal fluctuations, while qualitative analysis examines the content of reviews and descriptions to understand the factors that contribute to a destination's appeal during specific seasons. This dual approach allows for a nuanced understanding of not just when people visit certain places, but also why they choose those times.

By synthesizing data from these diverse sources, the project aims to create a comprehensive picture of seasonal tourism patterns in India. This approach allows for the identification of trends that might not be apparent from any single data source, providing a robust foundation for the development of a tourism recommendation system.

4 Implementation

4.1 *Data Preprocessing*

The data preprocessing phase involved several key steps to extract meaningful information from the collected data. First, the Google reviews data was processed to determine the number of comments for each month across different places. This involved parsing the review dates and aggregating them by month for each location.

Next, seasons were assigned to different months to facilitate seasonal analysis. This step was crucial for identifying patterns in visitation trends across different times of the year. The preprocessing pipeline was designed to handle large volumes of data efficiently, ensuring that the analysis could be scaled to cover a wide range of locations across India.

To address the issue of the first maximum month being skewed due to nationwide festival holidays, the system was adjusted to consider the second maximum month as the recommended time to visit. This approach provides a more balanced recommendation that takes into account location-specific peak seasons rather than just national holiday patterns.

```python
def processReviews(data, outputFileName='processed_reviews.csv'):
    processedData = (data
        .sort_values(by='published_at_date')
        .assign(published_at_date=lambda x: pd.to_datetime(x['published_at_date']))
        .assign(month=lambda x: x['published_at_date'].dt.strftime('%B'))
        .groupby('month')
        .size()
        .reset_index(name='reviewCount')
    )

    # Create a DataFrame with all months
    allMonths = pd.DataFrame({'month': list(month_name)[1:]})  # Excludes the empty string at index 0

    # Merge all months with the processed data
    processedData = allMonths.merge(processedData, on='month', how='left').fillna(0)
    processedData['reviewCount'] = processedData['reviewCount'].astype(int)

    # Sort the DataFrame by month order
    monthOrder = {month: index for index, month in enumerate(month_name[1:])}
    processedData['monthOrder'] = processedData['month'].map(monthOrder)
    processedData = processedData.sort_values('monthOrder').drop('monthOrder', axis=1)
    # Define seasons
    seasons = {
        "January": "Winter", "February": "Spring", "March": "Spring",
        "April": "Spring", "May": "Summer", "June": "Summer",
        "July": "Summer", "August": "Monsoon", "September": "Monsoon",
        "October": "Monsoon", "November": "Winter", "December": "Winter"
    }
    # Add season information
    processedData['season'] = processedData['month'].map(seasons)

    # Create the "testData" folder if it doesn't exist
    if not os.path.exists('Data/processedData'):
        os.makedirs('Data/processedData')

    # Determine the output filename
    if outputFileName is None:
        frame = inspect.currentframe().f_back
        calling_var = [var for var, val in frame.f_locals.items() if val is data][0]
        outputFileName = f"{calling_var}.csv"
    else:
        outputFileName = os.path.basename(outputFileName)

    # Save the processed data to a CSV file in the "testData" folder
    outputPath = os.path.join('Data/processedData', outputFileName)
    processedData.to_csv(outputPath, index=False)

    print(f"CSV file saved as: {outputPath}")

    return processedData
```

Figure 1: Function written to process the data (dataProcessing.py)

	column 1	column 2	column 3
1	month	reviewCount	season
2	January	4	Winter
3	February	12	Spring
4	March	4	Spring
5	April	5	Spring
6	May	3	Summer
7	June	8	Summer
8	July	7	Summer
9	August	12	Monsoon
10	September	8	Monsoon
11	October	289	Monsoon
12	November	3	Winter
13	December	5	Winter

Figure 2: Sample of the processed data

4.2 *UI Development*

The next step was the UI development. The user interface was developed using Streamlit, a powerful framework for creating web applications with Python. The UI incorporates several features to enhance user experience and provide comprehensive information:

1. **Search Functionality:** A search field was added to allow users to directly access data for specific places by entering the place name. This feature improves the ease of use and allows for quick information retrieval.
2. **Dropdown Menus:** Various dropdown menus were implemented to offer users different options for data visualization and analysis. This includes options to view the top 5 safest states or the top 5 most dangerous states based on crime data.
3. **Crime Ranking Display:** The application now shows the crime ranking of the state once a user searches for a place. This provides valuable safety information to potential tourists.
4. **Dynamic Bar Plot:** A visually appealing bar plot was integrated into the UI. This plot not only displays the crime ranking of various states using a heat map but also highlights the state of the currently searched location. This feature offers a quick visual comparison of safety across different states.
5. **Map Integration:** The UI includes a map showing the location of the searched place, providing geographical context to the user.
6. **Seasonal Information:** The application displays the most visited season for the selected location, along with other relevant tourism information gathered from the preprocessed data.
7. **Sources Page:** The application displays the most visited season for the selected location, along with other relevant tourism information gathered from the preprocessed data.

These UI elements work together to create a comprehensive and user-friendly interface that will allow tourists to make informed decisions about their travel plans based on both tourism trends and safety considerations.

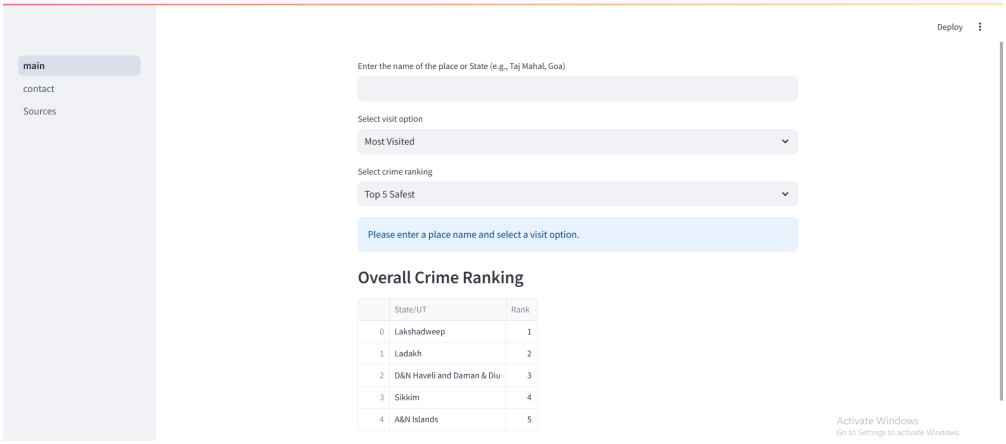

Figure 3: Web-app with empty search field

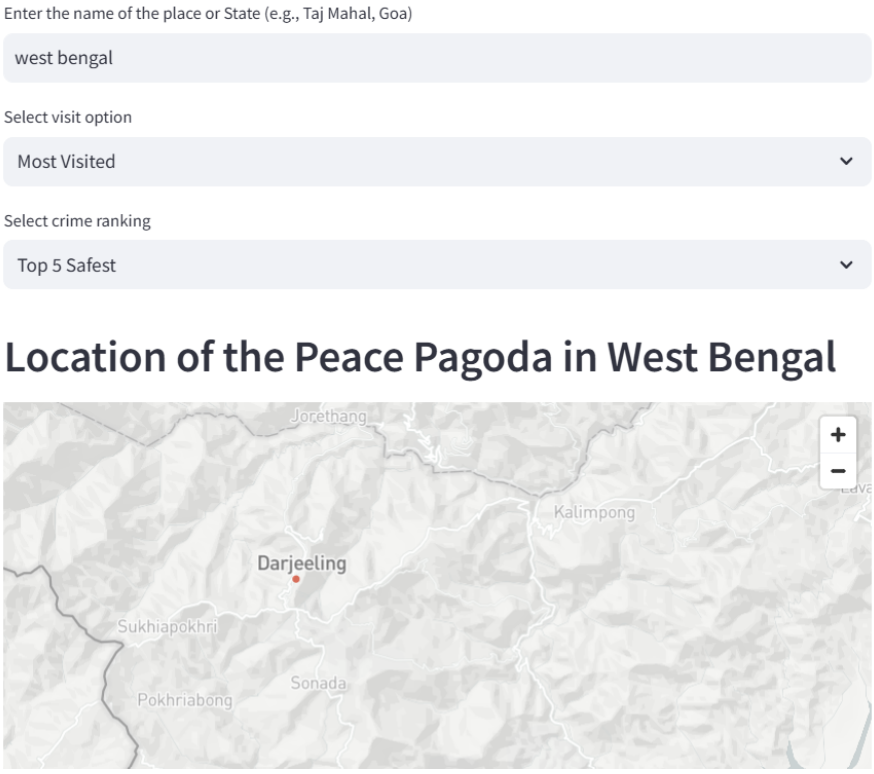

Figure 4: Web-app after searching for a specific state

West Bengal - Most Visited Time of the year

Month	Season	Review Count
January	Winter	17

Go to reviews

Crime Ranking

Crime Rank of West Bengal: 25

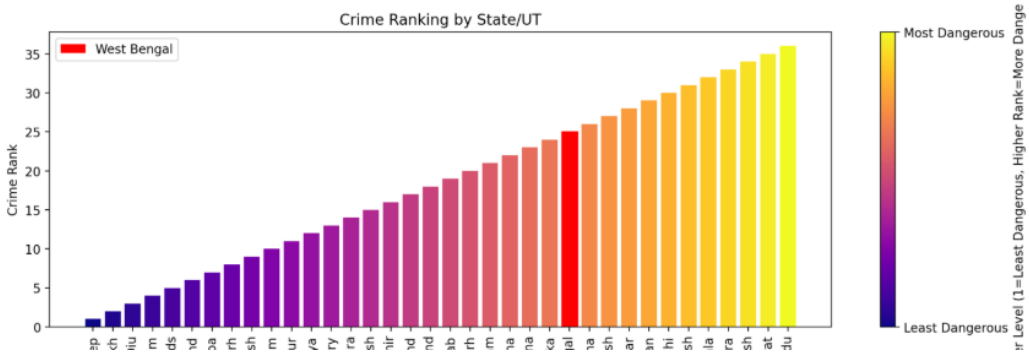

Figure 5: More details about the same state shown when scrolled down

5 Conclusion

This project has successfully developed a comprehensive tourism recommendation system for India, leveraging data science techniques to analyze seasonal visitation trends and safety considerations. By preprocessing Google review data, assigning seasons to different months, and integrating crime statistics, the system provides valuable insights for tourists planning their trips. The implementation of a user-friendly Streamlit interface, complete with search functionality, dropdown menus, and dynamic visualizations, enhances the accessibility and utility of the gathered data. The project's ability to combine tourism patterns with safety rankings demonstrates the power of data-driven decision-making in the travel industry. By offering recommendations based on the second most popular month rather than nationwide holiday patterns, the system provides more nuanced and location-specific advice. This approach not only aids tourists in making informed decisions but also contributes to the broader field of recommendation systems, showcasing the practical applications of data science in enhancing user experiences and industry practices.

6 Future Scope

The project has significant potential for expansion and refinement in the future. One key area for development is increasing the granularity of data from state-level to city-level analysis. This would allow for more precise recommendations, taking into account the unique characteristics and seasonal patterns of individual cities within each state. Additionally, the system could be enhanced by incorporating more tourist spots and attractions, further diversifying the range of recommendations available to users. A cross-analysis feature could be implemented to compare multiple destinations simultaneously, helping users make more informed decisions when choosing between different locations. The project could also benefit from continuous data updates to ensure the recommendations remain current and relevant. Furthermore, expanding the scope to include global tourist locations would transform the system into a powerful international travel planning tool. As the tourism industry evolves, particularly in response to changing travel patterns and safety concerns, this system could be adapted to include real-time data on factors such as crowd levels, health advisories, and environmental conditions, making it an even more comprehensive and valuable resource for travelers worldwide.

7 Important Links

GitHub repository - Travel-Insight

Web-app Link - travelinsight.streamlit.app/