Tourist Experience & Travel Risk Intelligence System

(A Big Data Project)

Satyaprakash Behera (24MBMB14)

School of Management studies.

1. Introduction

Tourism is one of the fastest-growing sectors in India and acts as a major contributor to cultural exchange, employment generation, and national revenue. The diversity of landscapes, heritage monuments, hill stations, beaches, and spiritual destinations attracts millions of travellers every year. However, the experience of tourists can significantly vary due to factors like crowd density, weather conditions, safety concerns, transportation availability, hygiene, travel scams, and seasonality.

With the rise of digital platforms, travellers increasingly share reviews, opinions, and personal experiences through online travel websites and social media channels. These reviews contain **rich experiential insights**, but they exist mostly in **unstructured text form**, making them difficult to analyse manually. Many travellers still struggle to identify:

- Which destinations offer the best experience?
- What is the right season to visit a location?
- Which cities are safer and more comfortable for tourists?
- What complaints do travellers frequently report?
- Which places match their travel preferences (family trip, solo trip, friend group travel, etc.)?

This project addresses these gaps by developing a **Tourist Experience and Travel Risk Intelligence System** that applies **Big Data Analytics and Natural Language Processing (NLP)** to large-scale Indian tourist review data. The system analyses ratings, sentiments, travel patterns, complaints, and contextual attributes to provide **data-driven tourism insights** and **recommendations**.

The project has been implemented in **Databricks**, leveraging the scalability of **Apache Spark (PySpark)** and the structured data lake foundation provided by **Delta Lake Medallion Architecture (Bronze → Silver → Gold)**.

2. Objectives of the Project

The main objectives of the project are:

1. To collect and organize large-scale Indian tourist experience data.
2. To preprocess the dataset by cleaning, enriching, and standardizing review information.
3. To perform **sentiment analysis** on tourist review text using NLP techniques.
4. To design a **city-level travel risk scoring model** that evaluates safety perception.
5. To evaluate **seasonal travel suitability** using experience scores.
6. To profile traveller behaviour based on traveller type (Solo, Family, Couple, Friends, Business).
7. To identify **top-rated attractions** and **frequent complaint themes**.
8. To create dashboards and visual insights for interpretation and decision-making.

3. Dataset Description

The dataset used in this study is named **tourist_experience_india.csv** and contains **around 1500 tourist review records** collected from Kaggle.

Attribute	**Description**
city	Tourist destination city visited
state	The Indian state where the city is located
attraction	Name of the tourist attraction
category	Type of place (heritage, nature, temple, theme park, etc.)
visit date	Date when the tourist visited
season	Seasonal time of the visit (Winter, Summer, etc.)

Attribute	Description
traveller type	Type of traveller (Solo, Friends, Family, Couple, Business)
duration days	Number of days spent
rating	Tourist's rating (1–5 scale)
sentiment label	Sentiment output (Positive / Neutral / Negative)
sentiment score	Sentiment polarity score (-1 to +1)
cleanliness score	Perception about cleanliness
safety score	Perceived travel safety
crowd level	Crowd conditions (Low / Medium / High)
cost bucket	Overall expense level of travel
review text	Detailed written review by the tourist

This dataset is rich because it contains both **quantitative fields** (rating, duration, safety score) and **qualitative text feedback**, making it suitable for **hybrid analytics**.

4. Methodology

To ensure structured data analysis and scalable computation, the project follows the **Medallion Data Architecture**, widely used in data Lakehouse systems.

4.1 Bronze Layer: Raw Data Ingestion

- The dataset was uploaded into the Databricks environment.
- Stored in **Delta Lake** format without applying any modifications.
- Serves as a **single source of truth** for the project.

4.2 Silver Layer: Data Cleaning and Enrichment

Data transformations applied:

- Handling missing or inconsistent values.

- Converting **visit_date** into proper Date format and extracting month and season.
- Standardizing traveller_type and attraction names.
- Applying **NLP Sentiment Analysis** to compute sentiment_score and sentiment_label.
- Creating a combined **experience_score** using:

$$\text{experience\_score} = (0.7 \times \text{rating}) + (0.3 \times \text{sentiment\_score})$$

4.3 Gold Layer: Feature Aggregation & Analytics

Analytical gold tables were created for different use cases:

Gold View/Table	Purpose
gold_city_rating_analysis	Identify best and lowest-rated tourist cities
gold_sentiment_analysis	Overall emotional experience trends
gold_city_risk_score	Safety risk classification of cities
gold_season_recommendation	Best months and seasons for travel
gold_traveller_profile	Preference analysis for traveller types
gold_top_attractions	Most loved tourist attractions
gold_complaint_patterns	NLP-based frequent complaint extraction

This layered design ensures **clean separation of processing stages**, enhances **reusability**, and improves **scalability**.

Models and Analytical Techniques Used

This project uses a combination of **statistical analysis**, **rule-based modelling**, and **Natural Language Processing (NLP)** techniques. Each use case is supported by a relevant analytical model described below:

Use Case	Model / Technique Used	Why This Model Was Used
Tourist Rating Analysis by City	**Descriptive Statistics (Mean Aggregation Model)**	To rank cities based on average tourist ratings, which reflects overall satisfaction.
Sentiment Analysis of Tourist Reviews	**VADER Sentiment Analysis (Rule-Based NLP Model)**	VADER is optimized for short social reviews and performs well without needing training data.
Travel Safety & Risk Score	**Weighted Risk Scoring Model (Custom Formula)**	Combines sentiment and rating behaviour to measure *perceived safety* more accurately.
Best Travel Season Recommendation	**Experience Score Model + Quartile Classification**	Provides an unbiased recommendation framework based on statistical distribution of experience scores.
Traveler Type Preference Profiling	**Group-Based Rating Aggregation (Segmentation Model)**	Highlights patterns within traveller groups to understand their destination preferences.
Top Loved Attractions	**Experience Score Ranking Model**	Ranks attractions by combining emotional tone and satisfaction score.
Complaint & Pain Point Analysis	**Keyword Frequency & Category Mapping (Text Mining Model)**	Extracts recurring complaint themes to identify tourism service gaps.

Explanation of Key Models

1. VADER Sentiment Analysis Model

- **Type:** Rule-Based Lexicon NLP Model
- **Why Used:**
    - Works extremely well for **short, informal travel reviews**

- No training required
- Computes a **compound score** used to classify sentiment into:
  - Positive
  - Neutral
  - Negative

2. Experience Score Model

$$\text{experience\_score} = 0.7 \times \text{rating} + 0.3 \times \text{sentiment\_score}$$

- Places higher weight on **rating** but still incorporates **emotion**
- Helps compare cities, attractions, and seasons more holistically

3. Travel Risk Scoring Model

$$\text{risk\_score} = (\text{Negative Review Ratio} \times 8) + (5 - \text{avg\_rating}) \times 0.2$$

- Higher negative review ratio = higher risk
- Helps classify cities into:

Risk Score	Safety Level
< 1.8	Safe
1.8 – 3.0	Moderate Risk
≥ 3.0	High Risk

4. Traveler Profiling Model

- Uses **segmentation logic**:
  Group reviews by traveller type → compute mean satisfaction → identify preference patterns.

5. Complaint Mining Model

- Uses:
    - Tokenization
    - Stop word Removal
    - Frequency Count
    - Category Mapping (Safety, Cost, Crowd, Cleanliness)

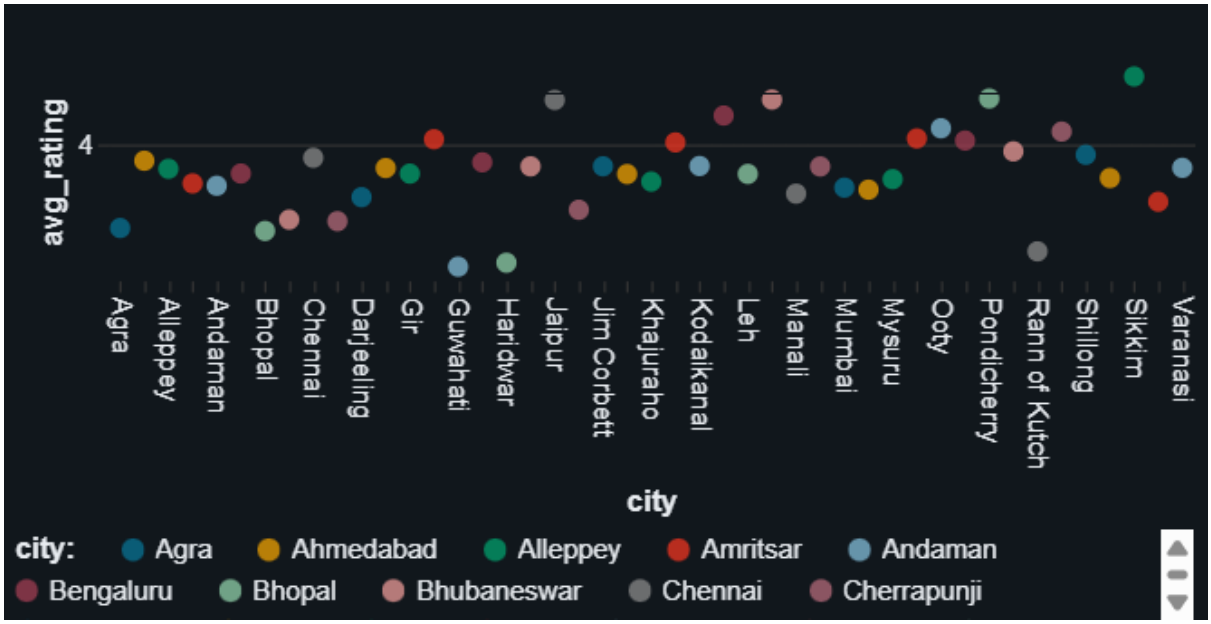
5. Use Case Analysis

5.1 Use Case 1: Tourist Rating Analysis by City

Cities were ranked based on average tourist ratings. This helps identify cities that consistently deliver good visitor satisfaction.

Insights:

- Cities like **Ooty, Jaipur, Pondicherry, Hyderabad, and Rishikesh** scored high.
- Congested metro areas show slightly lower satisfaction due to traffic and crowd.

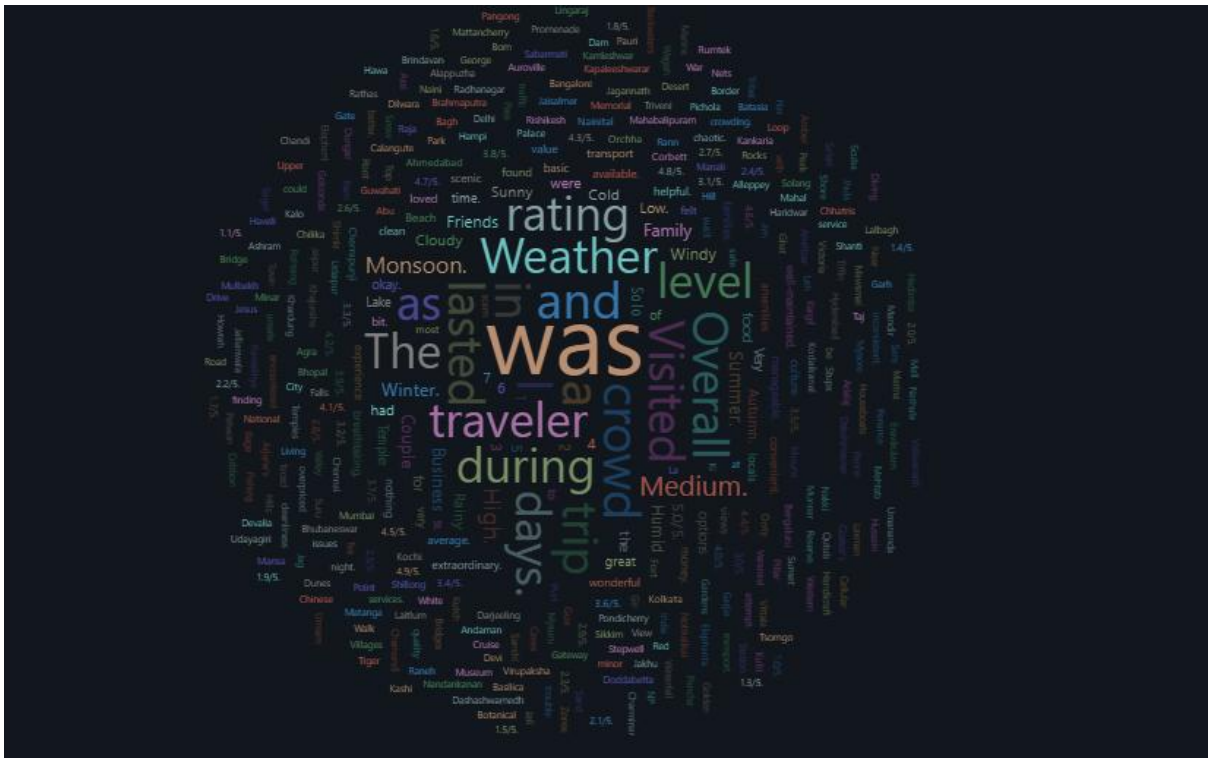


5.2 Use Case 2: Sentiment Analysis of Tourist Review

The VADER sentiment analyser was applied to text reviews.

Sentiment Type	Interpretation
Positive	Tourist was satisfied
Neutral	No strong emotional expression
Negative	Tourist faced issues or discomfort

Observation:
Heritage and nature destinations displayed **higher emotional positivity**, while urban areas showed **mixed sentiment** due to congestion.



5.3 Use Case 3: City Safety & Travel Risk Scoring Model

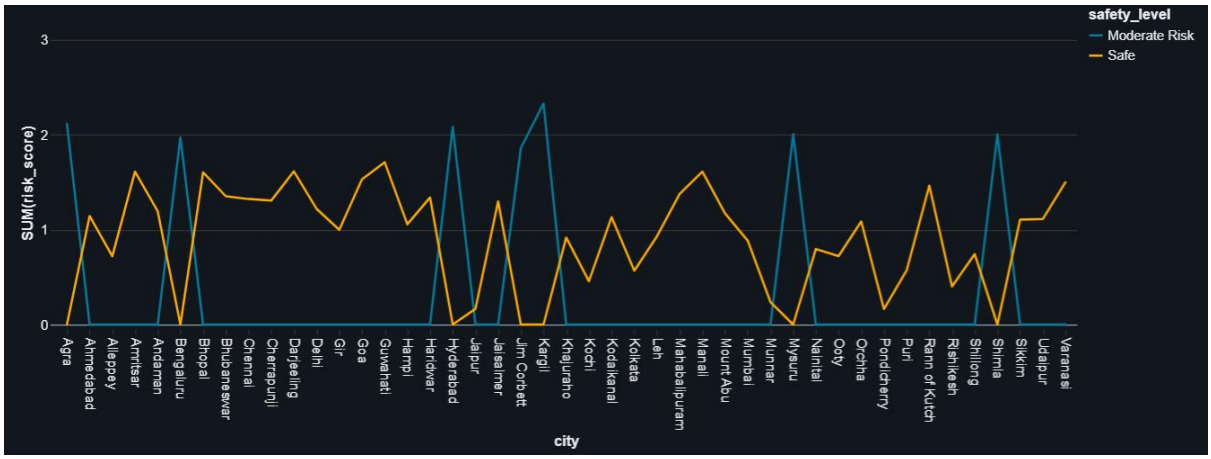
A **risk scoring model** was created:

$$\text{risk\_score} = (\text{Negative Review Ratio} \times 8) + (5 - \text{avg\_rating}) \times 0.2$$

Risk Score	Safety Level
< 1.8	Safe
1.8 – 3.0	Moderate Risk
≥ 3.0	High Risk

Insights:

- Highly crowded tourist regions showed **moderate or high travel risk**.
- Scenic and spiritual destinations remained **consistently safe**.

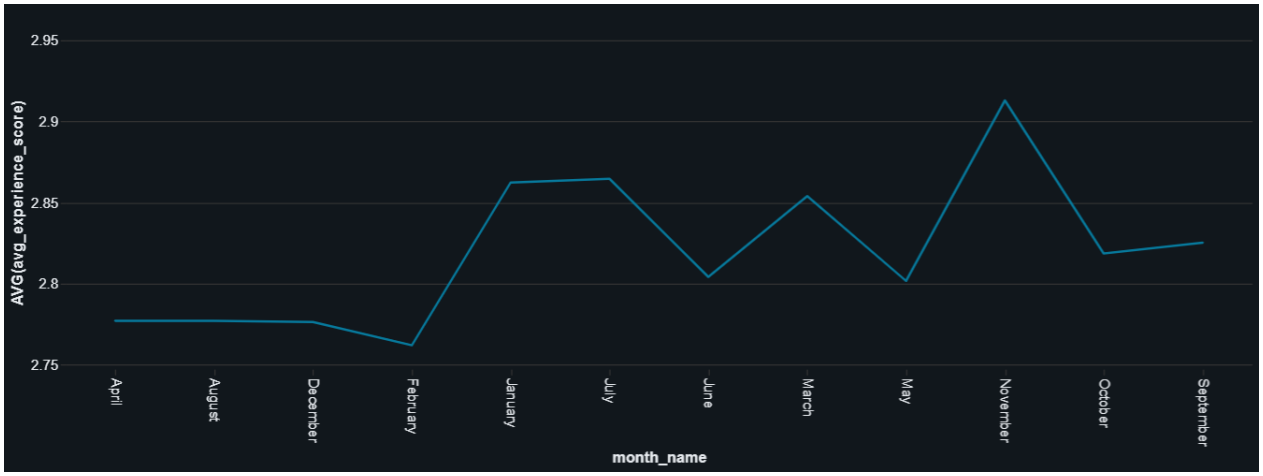


5.4 Use Case 4: Best Travel Season Recommendation

Experience scores were computed monthly to determine ideal travel seasons.

Patterns Found:

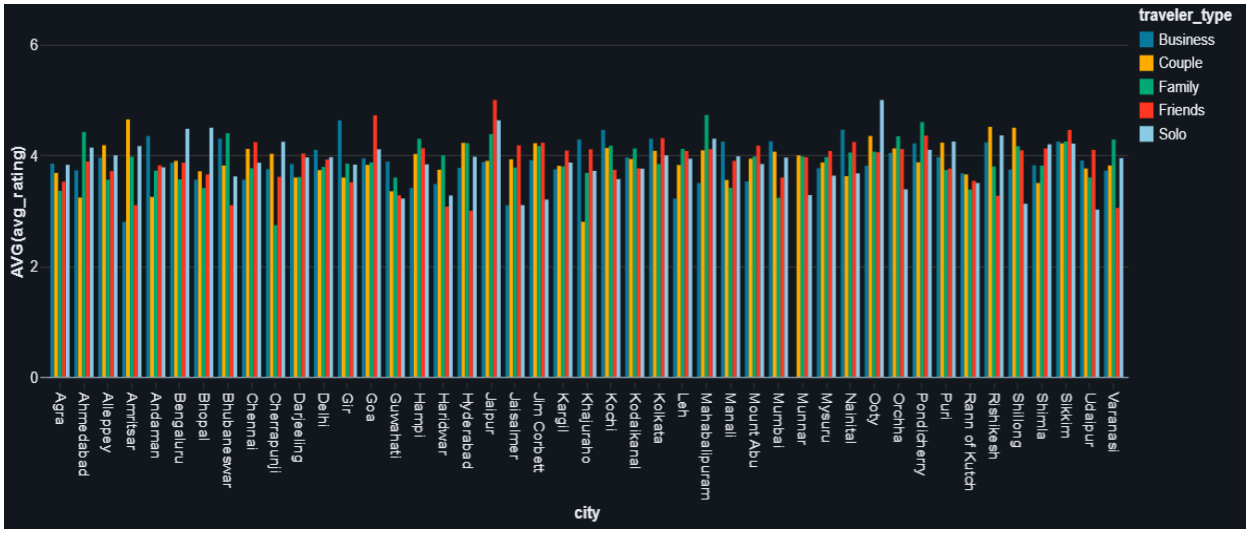
Season	Recommendation Trend
Winter	Highly Recommended
Summer	Travel with caution (heat & crowd)
Monsoon	Mixed experience
Autumn	Mostly good to travel



5.5 Use Case 5: Traveller-Type Preference Profiling

Different traveller groups prefer different types of destinations:

Traveller Type	Prefers	Reason
Couples	Peaceful & scenic hill stations	Privacy & scenic beauty
Friends	Adventure & nightlife	Shared group enjoyment
Families	Clean & safe cultural destinations	Comfort & convenience
Solo Travellers	Spiritual & low-budget cities	Personal comfort & independence
Business Travellers	Metro cities	Work travel convenience

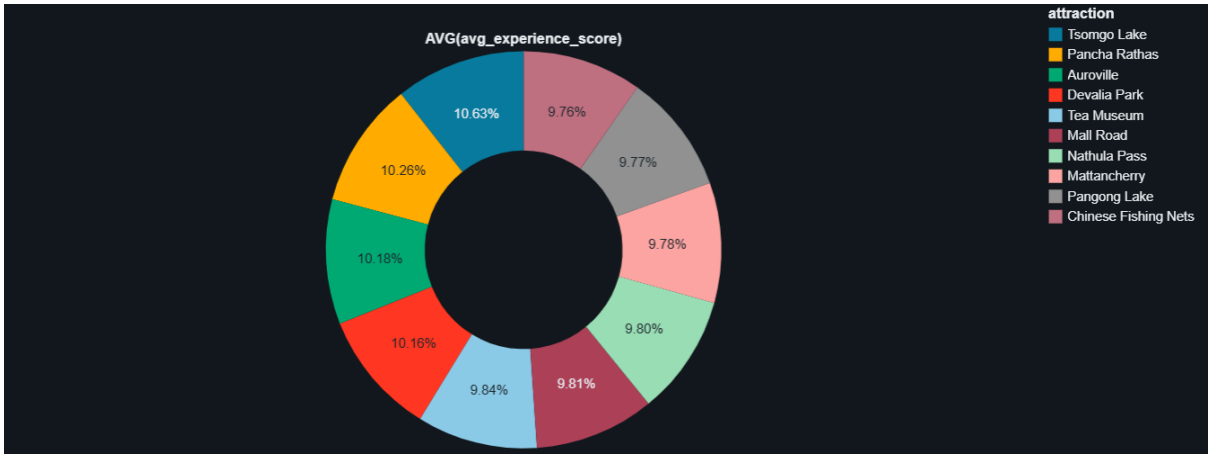


5.6 Use Case 6: Top Attractions Analysis

Top attractions were ranked based on experience score. Examples include:

- City Palace (Jaipur)
- Ooty Botanical Garden
- Golden Temple (Amritsar)

These reflect **well-maintained environments and positive visitor emotions**.

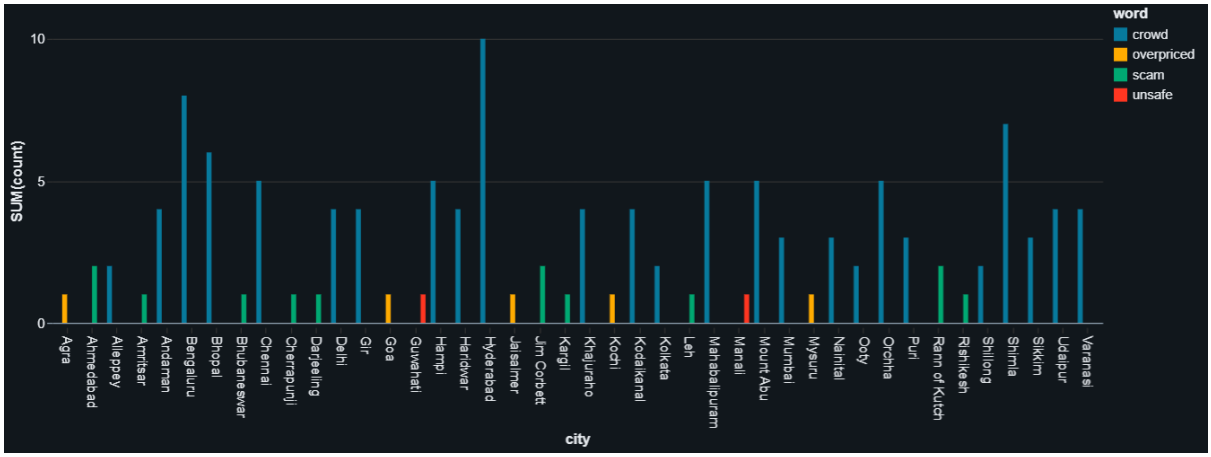


5.7 Use Case 7: Tourist Complaint & Pain Point Analysis (NLP)

Using keyword frequency clustering, common issues identified:

Complaint Category	Cities Often Mentioned
Crowd Congestion	Delhi, Manali, Shimla
Scams & Overcharging	Agra, Jaipur Markets
Cleanliness Issues	Beaches & transport hubs
High Cost	Goa, Mumbai Metro Zones

These findings help authorities improve **tourism infrastructure and management**.



6. Key Findings

1. **Season greatly affects tourist experience** — winter and post-monsoon are best for most destinations.
2. **Safety perception varies by crowd level** — crowded cities tend to have more negative sentiment.
3. **Traveller preferences are strongly role-based** (solo, couple, family, etc.).
4. **Cultural & heritage destinations** consistently rank high in satisfaction.

7. Conclusion

The project successfully demonstrates the use of **big data analytics, sentiment analysis, and risk modelling** to evaluate tourist experiences in India. Using Databricks and PySpark enabled scalable data processing, structured pipeline development, and repeatable analytics workflows. The insights generated can help tourists make informed travel decisions and support tourism departments in **improving visitor experience quality**.

8. Future Scope

- Develop a **real-time tourist advisory dashboard**
- Implement **AI-based personalized trip recommendations**
- Integrate **geo-spatial heatmaps for live crowd monitoring**
- Deploy as a **mobile travel planning application**