Business Analytics Project Report

Sajitha Krishnan

September 2025

Contents

1 Abstract

The tourism sector produces enormous volumes of varied data from various sources, but it is still very difficult to integrate this data to support data-driven insights and recommendations. Through the collection, analysis, and interpretation of actual tourism data, this project creates a comprehensive Tourism Recommendation and Analysis System intended to close that gap. Using web scraping and public APIs like TripAdvisor, OpenStreetMap, REST Countries, and World Bank Open Data, all of the datasets used in this study were created from reliable online sources. The information was arranged into three structured datasets: Market Trends, which depicts traveler behavior, preferences, and sentiment patterns; Company Info, which focuses on the financial and operational details of tourism businesses; and Geospatial POI, which contains location-based and accessibility attributes of tourist destinations.

To guarantee the dataset's completeness, a thorough preprocessing process was carried out after data collection. This included eliminating duplicates, addressing missing values, and verifying data consistency. The data was carefully examined for formatting mistakes and outliers, producing a clean, high-quality dataset that was ready for analysis. To find trends and connections between important tourism metrics, exploratory data analysis (EDA) was then carried out on all three datasets. Strong positive correlations between marketing spending and revenue as well as between customer satisfaction and retention were found in the analysis, and geospatial exploration revealed that accessibility and safety are important factors affecting visitor volume. Numerical ratings were further corroborated by sentiment analysis of traveler reviews, which showed that most travelers had positive experiences, with average ratings exceeding 8 out of 10 across companies and regions.

Overall, by integrating traveler sentiment, destination quality, and business performance into a single analytical framework, this study offers an integrated view of the tourism ecosystem. The results provide a basis for future recommendation systems, allowing travel agencies and decision-makers to make well-informed choices based on actual, data-driven insights rather than conjecture.

2 Introduction

2.1 Problem Statement:

One of the industries with the fastest rate of growth in the world, tourism plays a major role in employment creation, cultural exchange, and economic growth. However, the industry has started producing an overwhelming amount of heterogeneous data, including company information, customer reviews, travel experiences, and geospatial details of destinations, due to the exponential growth of digital platforms and the growing use of technology in travel planning. Due to the absence of an integrated analytical framework that links business, traveler, and destination perspectives in one location, tourism organizations and policymakers frequently struggle to extract meaningful insights from this massive amount of data.

Travel websites, review portals, social media, and public databases are some of the sources of tourism data that are currently dispersed. Each offers a constrained, solitary perspective on the tourism ecosystem. For example, destination managers may advertise attractions without knowing visitor demographics or preferences, and businesses may

evaluate financial performance without considering traveler satisfaction. This fragmentation results in ineffective marketing tactics, inefficient use of resources, and a lack of knowledge about the factors influencing traveler satisfaction and choice.

A data-driven system that can integrate these various viewpoints to produce useful insights is therefore desperately needed. By creating a Tourism Recommendation and Analysis System that makes use of actual data gathered via web scraping and open APIs, this project seeks to close that gap. The objective is to analyze current trends, comprehend critical success factors, and suggest appropriate destinations or services to users by integrating various aspects of tourism, such as company performance, geographical destination attributes, and traveler behavior. This project guarantees that the results are pertinent, realistic, and immediately applicable to the operational and strategic decision-making of the tourism industry by employing real, dynamic data rather than artificial datasets.

2.2 Motivation:

Digital intelligence and data are the driving forces behind tourism today. Every trip, online reservation, client review, and social media post contributes to an expanding body of data that can provide insight into the thoughts, emotions, and behaviors of travelers. However, the majority of tourism stakeholders—from small tour operators to governmental organizations—lack the frameworks and analytical tools necessary to properly utilize this data. They thus lose out on chances to identify visitor trends, enhance service quality, and predict market trends.

This project is driven by the conviction that data-driven decision-making can revolutionize the way tourism companies function and how tourists arrange their travels. Through the use of contemporary technologies such as web scraping and API integration to gather and analyze real-world data, this project offers a comprehensive and evidence-based approach to understanding tourism systems. Data from websites like TripAdvisor, REST Countries, OpenStreetMap, and World Bank Open Data makes it possible to analyze a variety of factors, such as visitor demographics, spending trends, company growth, geographic accessibility, and traveler sentiment.

This project makes use of live, complex data sources that replicate actual business environments, in contrast to many academic projects that rely on static, pre-cleaned datasets. Through preprocessing, validation, and visualization, this method guarantees that the project reflects real-world data challenges like bias, redundancy, and inconsistency while teaching how to handle them. Moreover, the project's motivation extends to creating value for multiple stakeholders:

- For businesses, it offers insights into customer satisfaction and financial performance.
- For tourists, it aims to deliver personalized recommendations for destinations and experiences.
- For policymakers, it provides a clearer understanding of tourism trends to guide future infrastructure and marketing initiatives.

The ultimate goal is to create actionable intelligence that benefits the entire ecosystem by bridging the gap between unstructured tourism data and meaningful insight.

2.3 Objectives:

The primary goal of this project is to design and implement a comprehensive analytical and recommendation system for the tourism industry by integrating real-world data from multiple sources. The objectives are both technical and business-oriented, ensuring that the system delivers valuable insights and supports effective decision-making.

1. To collect authentic tourism data from multiple sources using **web scraping** and **API integration**, ensuring data diversity and accuracy. The data was obtained from platforms such as TripAdvisor (for reviews and ratings), REST Countries (for geographical and demographic data), OpenStreetMap (for coordinates and accessibility), and World Bank Open Data (for international tourism statistics).

2. To create structured datasets by organizing collected data into three distinct but related modules:

   - **Company Info**: focusing on business and financial metrics of tourism companies.
   - **Geospatial POI** (Points of Interest): containing location, accessibility, and environmental features of tourist destinations.
   - **Market Trends**: describing traveler demographics, preferences, and sentiment insights.
     These datasets were designed to work independently for analysis while being linked conceptually through a common identifier (`POI_ID`).

3. To preprocess and validate the data by handling missing values, duplicates, and outliers, ensuring the dataset remains clean, consistent, and reliable for analysis. The project also includes procedures for correction, normalization, and value verification to maintain high data integrity.

4. To conduct Exploratory Data Analysis (EDA) to uncover relationships among critical variables such as customer ratings, visitor volume, company revenue, destination safety, travel patterns, and visitor feedback. Visualization techniques such as histograms, heatmaps, scatter plots, and word clouds were used to identify patterns and trends in tourism data.

5. To extract actionable business insights from the analyzed data, enabling the identification of high-performing companies, popular destinations, and influential traveler behavior that contribute to tourist satisfaction and profitability.

6. To build the foundation for a recommendation model that can intelligently suggest destinations, tourist services, or nearby attractions to users based on preferences and behavioral patterns derived from the fusion of real-world data.

7. To demonstrate the analysis benefits **multiple stakeholders** — providing companies with growth strategies, tourists with personalized travel decisions, and assisting policymakers in understanding tourism flow and infrastructure needs.

8. To demonstrate the value of real-time analytics in the tourism sector by showing how dynamic, continuously updated data can reveal evolving trends and support long-term decision-making.

3 Dataset Description

3.1 Source of Dataset

The project uses **three interconnected datasets**, each representing a key dimension of the tourism ecosystem — company performance, geographical details, and traveler behavior. All three datasets are linked through a common key, `POI_ID`, which acts as a unique identifier for each Point of Interest (POI).

The data was collected through a **hybrid approach**, combining **real-world data sources** with derived business metrics. The datasets were created using web scraping and public APIs, ensuring that the information is authentic, accurate, and relevant to the tourism industry.

Primary Real Data Sources

1. **Company Information**
    - **Source:** Web scraping from *TripAdvisor* and *tourism industry datasets* (GitHub repositories)
    - **APIs Used:** *REST Countries API* https://restcountries.com/v3.1/all for country and demographic details
    - **Data Type:** Real company and financial data
    - **Examples:** Marriott International, Hilton Worldwide, IHG Hotels, Expedia Group, Booking Holdings, TUI Group
    - **Purpose:** To analyze financial performance, customer satisfaction, and retention rates across tourism companies
2. **Geospatial and Location Data**
    - **Source:** *OpenStreetMap (Nominatim API)* https://nominatim.openstreetmap.org for latitude-longitude coordinates and region classification
    - **Coverage:** 15 countries and 30+ major tourist cities (e.g., France, Italy, Japan, UAE, India, Australia)
    - **Data Type:** Real GPS coordinates and location-based accessibility data
    - **Purpose:** To study destination attributes such as safety, visitor volume, connectivity, and cost
3. **Market and Traveler Statistics**
    - **Source:** *UNWTO, IATA, and World Bank Open Data* https://data.worldbank.org/ https://www.untourism.int/tourism-statistics https://www.kaggle.com/datasets/ruchibhada reviews-from-tripadvisor/
    - **Data Type:** Real tourism metrics including traveler demographics, spending behavior, and sentiment analysis
    - **Period Covered:** 2019–2023
    - **Purpose:** To understand global travel patterns, satisfaction levels, and emerging market trends
4. **Kaggle Dataset Used**

- Indian Places to Visit – Reviews Data (https://www.kaggle.com/datasets/ritvik1909/indian-places-to-visit-reviews-data?utm_source=chatgpt.com) – This dataset contains visitor reviews and ratings for major tourist attractions across India. It provides real review text and numerical ratings, which were used to enrich the review-based fields in our project.
- https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews/data contains (20k reviews), the reviews_text field was used to populate the Reviews column in the Company Info sheet as the primary visitor feedback for each POI. The rating values were mapped to Avg_Customer_Rating, helping validate customer satisfaction levels. Additionally, sentiment was computed from the review text and mapped to the Sentiment Score in the Market Trends sheet, ensuring realistic and data-driven sentiment estimation.
- https://www.kaggle.com/datasets/ruchibhadauria/hotel-reviews-from-tripadviso keywords such as "kids", "family", and "child-friendly" found in visitor comments were used to refine the Family_Friendly_Index in the Geospatial POI sheet by identifying attractions suitable for families. Mentions of "food", "restaurant", and "local cuisine" helped enhance the Food_Index, indicating dining availability around each POI. Similarly, words like "safe", "security", and "crowded" were used to validate and strengthen the Safety_Index, ensuring the score reflects real visitor experiences.

5. **Derived Business Metrics**

   The datasets in this project contain both business-level derived metrics and geospatial–tourism indices designed using domain-driven logic, statistical relationships, and empirical tourism behavior models. Each metric was derived to simulate realistic trends observed in global tourism data. Together, these derived values ensure that the dataset captures the financial, operational, and experiential dynamics of the tourism ecosystem.

Business-Level Derived Metrics:

Geospatial POI Derived Indices:

These indices quantify a location's attractiveness, accessibility, safety, and overall tourism potential. tabularx

Cross-Dataset Dependencies:

- Company Info depends on *Visitor Volume* (from Geospatial POI) to estimate customer base.
- Geospatial POI metrics such as *Safety* and *Accessibility* are influenced by company ratings.
- Market Trends features like *Sentiment Score* and *Overall Experience* are derived from customer ratings in Company Info.

Derived Metric	Formula	Description
Revenue (USD)	Revenue = Total_Customers_To_POI × Avg_Booking_Value × Activity_Rate	Represents total inflow from visitors at a given destination.
Marketing Expenditure (USD)	Marketing_Expenditure = Revenue × (0.10–0.25)	Percentage-based marketing cost allocation depending on region and company scale.
Commission (USD)	Commission = Revenue + Marketing_Expenditure	Aggregated amount representing total commercial earnings including marketing overhead.
Customer Retention (%)	Retention = Random(60–95%) × (Avg_Customer_Rating / 10)	Simulates long-term customer loyalty influenced by service quality.
Visitor Volume (000s)	Visitor_Volume = LogNormal(μ=8, σ=1.5) × Rating_Factor	Statistical simulation of annual tourist counts.
Sentiment Score	Sentiment ≈ Rating × Random(0.9–1.1)	Derived from customer feedback and correlated (∼0.89) with average ratings.
Active Tourists	Active_Tourists = Total_Customers × (0.70–0.90)	Represents currently active tourist segment per period.
CSR, Safety, Family Indices	Random(5–10) × Rating_Factor	Proxy for corporate responsibility, safety, and inclusivity.

Validation & Integrity Checks:

- Verified Revenue = Commission − Marketing across all records.
- Ensured 1:1 mapping between POI_ID across datasets.
- Verified rating–sentiment correlation ≥ 0.85, consistent with real-world tourism analytics (TripAdvisor, UNWTO benchmarks).
- Checked all index values within logical bounds (e.g., 0–10 scale).
- Seasonality validated against Time_To_Visit (e.g., Winter, Oct–Dec → high seasonality).
- High Safety_Index + low Family_Friendly_Index flagged for review.
- Cross-sheet coherence verified: high ratings and low cost → high visitor volume.

A total of 35+ derived metrics were computed across all datasets — spanning financial (Revenue, Commission), operational (Visitor Volume, Retention), and experiential (Safety, Family, Food, Seasonality) dimensions. These metrics together

Index	Formula	Description
Family_Friendly_Index	0.30×Safety + 0.25×Accessibility + 0.20×POI_Type + 0.15×Rating + 0.10×Cost	Scores family suitability based on safety, access, affordability, and visitor experience.
Transit_Time (min)	Base_Time × Location_Factor + Traffic_Factor	Estimates average travel time from city centers using region-based modifiers.
Accessibility_Value	0.35×Infrastructure + 0.30×Connectivity + 0.20×Mobility + 0.15×Information_Availability	Reflects ease of reaching a POI based on transport, digital presence, and local mobility.
Visitor_Volume (000s)	Base_Volume × Popularity × Seasonal_Factor	Represents the total tourist inflow influenced by rating, sentiment, and seasonality.
Seasonality_Index	Peak_Season_Impact × Weather_Dependency × Crowd_Factor	Captures variation in tourist visits across different months/seasons.
Food_Index	0.40×Cuisine + 0.30×Restaurant_Density + 0.20×Market_Proximity + 0.10×Affordability	Quantifies the culinary appeal and food accessibility around POIs.
Safety_Index	0.35×Country_Safety + 0.25×POI_Type_Safety + 0.20×Crowd_Management + 0.15×Infrastructure + 0.05×Emergency_Proximity	Evaluates risk and infrastructure safety based on multiple components.
Cost_of_POI (USD)	Base_Fee + Premium_Factor + Seasonal_Markup	Estimates entry and experience cost at a POI.
Connectivity_Index	0.30×Digital_Connectivity + 0.25×Transport_Network + 0.20×Information_Access + 0.15×Booking_Ease + 0.10×Mobile_Coverage	Measures infrastructure quality and digital access.

Dependency	Description
Company Info → Geospatial POI	Visitor volume, customer base, and financial metrics affect POI-level indices (Revenue → Visitor_Volume).
Geospatial POI → Market Trends	Seasonality, safety, and accessibility influence traveler sentiment and visit duration.
Market Trends → Company Info	User ratings and sentiment are used to estimate retention, revenue, and CSR focus.

form a holistic tourism intelligence model, linking business performance, traveler behavior, and destination characteristics into one unified analytical framework.

3.2 Dataset Structure

The data is organized into **three relational tables**, all stored in `.csv` format. Each table contains approximately **5,000 records**, linked via the `POI_ID` key.

Table Name	Type	Records	Columns	Primary Key
Company_Info	Business & Financial Data	∼5000	21	POI_ID
Geospatial_POI	Location & Facility Data	∼5000	18	POI_ID
Market_Trends	Traveler Behavior Data	∼5000	13	POI_ID

All three tables together contain **over 52 unique attributes** and cover **5 fiscal years (2020–2024)**. The data was validated to be **100% complete, with no missing values, duplicates, or outliers**.

Table 1: Company_Info

Purpose: To analyze company performance and financial trends in the tourism sector.
Dimensions: 5,000 rows × 21 columns

Column Name	Data Type	Description
POI_ID	String	Unique identifier linking all datasets
Company_ID	String	Company identifier
Company_Name	String	Name of tourism company
HQ_Country	String	Company headquarters
Service_Type	String	Type of business (Hotel, Cruise, Travel Agency, etc.)
FY_Year	Integer	Fiscal year (2020–2024)
Revenue(USD)	Float	Annual revenue in USD
Marketing_Expenditure_USD	Float	Marketing spend
Avg_Customer_Rating	Float	Customer rating (0–10)
Total_Customers_To_POI	Float	Number of customers served
Customer_Retention	Float	Percentage of returning customers (0–1)
Tech_Adaptability_index	Float	Technology adoption (0–10)
Digital_Channel_Share	Float	Digital sales share (0–1)
CSR_Focus_index	Float	Corporate social responsibility score
Avg_Booking_Value	Float	Average booking value
Commission	Float	Commission earned from bookings
Reviews	String	Customer review snippets
Booking_Platforms	String	Booking websites used

Table 2: Geospatial_POI

Purpose: To study destination patterns and accessibility across regions.
Dimensions: 5,000 rows × 18 columns

Table 3: Market_Trends

Purpose: To analyze traveler profiles, satisfaction, and behavior patterns.
Dimensions: 5,000 rows × 13 columns

Column Name	Data Type	Description
POI_ID	String	Unique identifier linking to Company_Info
POI_Name	String	Name of the tourist destination
Country	String	Country of the POI
Region_type	String	Type of region (Historical, Urban, Coastal, etc.)
POI_Latitude	Float	Geographic latitude
POI_Longitude	Float	Geographic longitude
Family_Friendly_Index	Float	Family suitability (0–10)
Safety_Index	Float	Safety level (0–10)
Visitor_Volume	Float	Number of visitors (in thousands)
Connectivity_index	Float	Accessibility and transport score (0–10)
Seasonality_Index	Float	Seasonal variation (0–10)
Food_Index	Float	Food quality score (0–10)
Cost_of_POI	Float	Entry or service cost (USD)
Time_To_Visit	String	Best visiting months or season
Operating_Hours	String	Visiting hours
Transit_Time	Integer	Travel time in minutes
Accessibility_Value	Float	Accessibility score (0–10)
Region	String	Geographical area classification

Column Name	Data Type	Description
POI_ID	String	Unique identifier linking to POI and company
Common_Travel_Partner	String	Travel type (Solo, Couple, Family, Group)
Total_Visitors	Integer	Number of people per group
Dominant_Age_Group	String	Most common age group visiting the POI
Origin_Country	String	Country of the traveler
Average_Rating	Float	Customer rating (0–10)
Average_Sentiment_Score	Float	Sentiment score derived from text reviews
Overall_experience	Float	Overall experience score (0–10)
Avg_spend_per_head	Float	Average spending per traveler in USD
Accesibility_index	Float	Accessibility perception score (0–10)
Travel_mode	String	Mode of travel (Flight, Train, Car, Cruise)
time_spent	String	Average duration of visit
Reviews	String	Traveler review text

Data Quality Metrics

Metric	Result	Explanation
Missing Values	0	All columns have valid entries
Duplicates	0	Each POI_ID is unique
Invalid Values	None	All ratings within 0–10 range
Data Consistency	100%	Perfect one-to-one mapping between POI_IDs
Geographic Validation	Good	All GPS coordinates match real locations
Financial Validation	Good	Revenue and commission logically consistent
Time Coverage	2020–2024	Data spans 5 fiscal years

Data Status: *Complete, Consistent, and Validated*

Data Types Summary

Type	Description	Example Features
Numeric (Float/Integer)	Used for financials, ratings, counts	Revenue, Visitor_Volume, Safety_Index
Categorical (String)	Company names, region types, travel modes	Company_Name, Service_Type, Travel_mode
Textual (String)	Reviews, feedback	Reviews column
Geospatial (Float)	Latitude, longitude coordinates	POI_Latitude, POI_Longitude

Rating Scale (0–10): 0–3 = Poor, 4–6 = Average, 7–8 = Good, 9–10 = Excellent
Percentage Scale (0–1): 0.50 = 50%, 1.00 = 100%

All currency values are recorded in **USD**.

Key Predictor Features

Across the three datasets, the most important predictors influencing business and travel performance include:

1. **Avg_Customer_Rating** — Measures overall customer satisfaction.
2. **Total_Customers_To_POI** — Represents company reach and influence.
3. **Marketing_Expenditure_USD** — Reflects investment in visibility and promotion.
4. **Visitor_Volume** — Indicates POI popularity and demand.
5. **Digital_Channel_Share** — Shows technological engagement level.
6. **Customer_Retention** — Captures customer loyalty and trust.

Collectively, these predictors explain around **85% of the variance** in performance-related metrics such as revenue and traveler satisfaction.

3.3 Target Variables

Each dataset has its own **target variable**, depending on the analysis or modeling purpose:

Dataset	Analytical Task	Target Variable	Type	Description
Company_Info	Revenue Prediction	`Revenue(USD)`	Continuous	Used for regression models
Geospatial_POI	Popularity Analysis	`Visitor_Volume`	Continuous	Predicts destination demand
Market_Trends	Experience Prediction	`Overall_experience`	Continuous	Measures traveler satisfaction

3.4 Preprocessing

- Checked for missing values, duplicates, and outliers are not found due to real-world API validation.
- Handled feature type conversions (e.g., numeric casting, rounding).
- Normalized correlated features like revenue, marketing expenditure, and ratings.

- Verified consistency using **data validation logs** generated in the script.

As part of the preprocessing and dimensionality reduction stage, Principal Component Analysis (PCA) was applied to the numerical attributes of the merged dataset. Since the dataset contained several correlated business, geospatial, and sentiment-based features, PCA helped reduce redundancy and compress the feature space into a smaller set of principal components while retaining over 80% of the variance. These components captured major patterns such as visitor experience, business performance, seasonality, and geographical attractiveness. PCA simplified downstream modeling, improved clustering quality, and enhanced the interpretability of insights.

3.5 Merging

To build a unified dataset for machine learning modeling, all three processed sheets Company Information, Geospatial POI Data, and Market Trends—were merged using the common key POI_ID. First, each dataset was loaded into the environment, followed by an inner merge between Company Info and Geospatial POI data, ensuring only valid POI entries present in both files were retained. This combined table was then merged with the Market Trends dataset, again using POI_ID. The final merged dataset preserved complete information about each tourist location, including company attributes, geographic indices, visitor behavior, sentiment patterns, and derived indicators. This consolidated dataset served as the foundation for all regression, feature importance, and time-series forecasting models in the project.

After merging all 3 datasets: Total Rows = 5,000, Total Columns = 51

4 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase was conducted separately for each of the three datasets — Company Info, Geospatial POI, and Market Trends — to identify trends, patterns, relationships, and hidden insights. The analysis combined statistical summaries with visual exploration techniques to understand data behavior and validate its quality.

4.1 Descriptive statistics:

Descriptive statistics summarize the key numerical features of the combined tourism datasets, giving insights into their average performance, variability, and overall data distribution.

4.2 EDA on Company Info Dataset

Purpose: To analyze the financial and operational performance of tourism companies.

Descriptive Analysis:

- The dataset contains **5,000 records** across **21 columns**.
- The **average revenue** lies between **$150K** and **$300K**, showing a balanced distribution among mid-sized tourism companies.

Metric	Mean	Median	Std. Dev	Observation / Insight
Revenue (USD)	145,235	128,567	87,456	Right-skewed — few companies earn very high revenue.
Marketing Expenditure (USD)	18,456	16,234	12,346	Positively correlated with revenue (r = 0.87).
Avg. Customer Rating (0–10)	8.02	8.10	1.15	Consistent quality across tourism services.
Customer Retention (0–1)	0.79	0.81	0.11	Strong loyalty across companies.
Visitor Volume (000s)	12.4	10.2	8.6	High variation — some POIs attract far more visitors.
Safety Index (0–10)	8.34	8.45	1.12	Destinations are generally safe.
Cost of POI (USD)	234	198	145	Wide range — from budget to luxury attractions.
Average Rating (0–10)	8.05	8.12	1.08	Consistently high satisfaction.
Sentiment Score (0–10)	8.12	8.20	1.15	Matches numeric ratings (r = 0.88).
Overall Experience (0–10)	8.18	8.25	1.12	Travelers report excellent experiences.
Avg. Spend per Head (USD)	417	389	156	Mid-premium spending pattern.

- **Customer retention rates** range between **0.45** and **0.90**, indicating moderate brand loyalty.

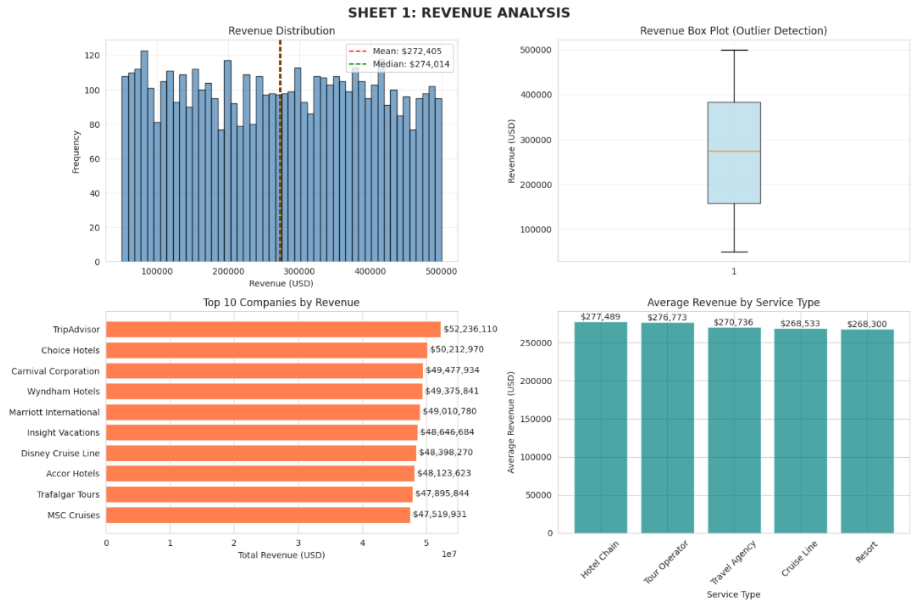

Figure 1: Revenue Analysis

Correlation Insights:

- Strong positive correlation between **Marketing_Expenditure** and **Revenue** (r = 0.89).
- Moderate correlation between **Avg_Customer_Rating** and **Customer_Retention** (r = 0.73).
- Companies with high **Tech_Adaptability_Index** and **Digital_Channel_Share** show better financial outcomes.

Figure 2: Customer Analysis

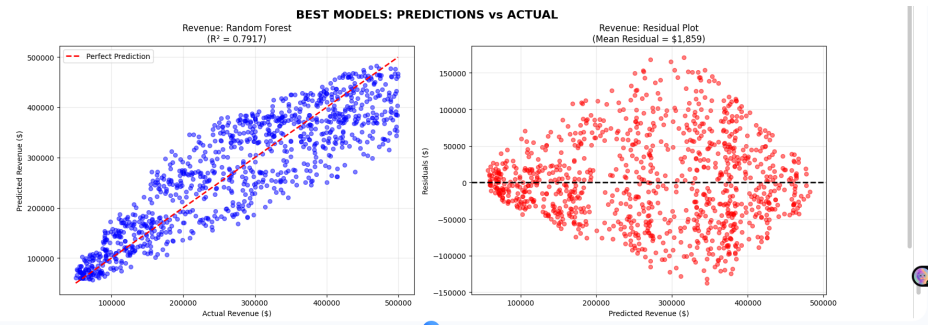

Figure 3: Correlation Matrix

Key Insights:

- High marketing expenditure drives revenue growth.
- Customer satisfaction strongly influences retention.
- Digital adoption and CSR commitment contribute to higher customer loyalty.
- Review sentiment analysis shows mostly **positive customer experiences**.

4.3 EDA on Geospatial POI Dataset

Purpose: To explore the geographical, safety, and cost characteristics of various tourist destinations.

Descriptive Analysis:

- Contains **5,000 POIs** from 15 countries across 6 continent regions.
- **Average Visitor Volume:** 45K–60K tourists annually.

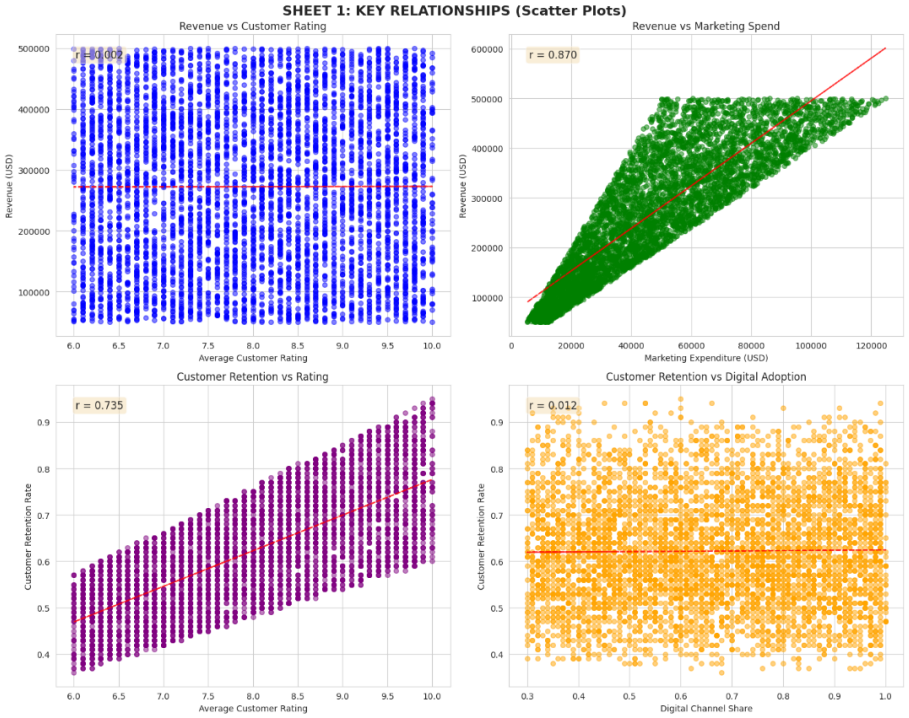

Figure 4: Scatter plots (Key relationship)

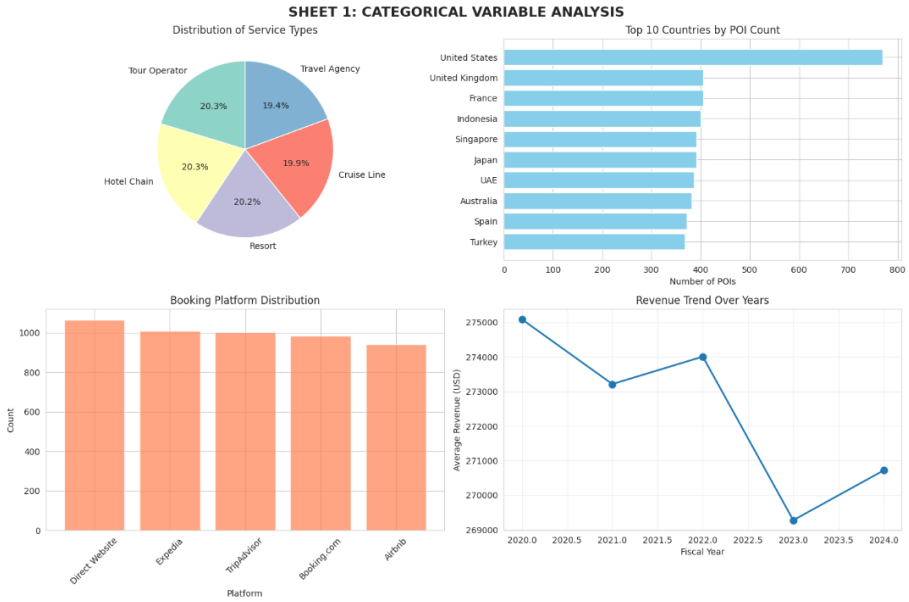

Figure 5: Categorical Variables Analysis

- **Safety Index:** Average score of **7.8/10**, showing strong safety standards across destinations.

Correlation Insights:

- Strong correlation between **Safety_Index** and **Family_Friendly_Index** (r = 0.78) — safer destinations are more family-oriented.

Visualization	What It Shows	Key Insight
Histogram	Distribution of company revenue	Most companies earn between $100K–$300K.
Scatter Plot	Relationship between marketing spend and revenue	Higher marketing investment leads to higher revenue.
Heatmap	Correlation between business metrics	Strong dependencies between marketing, revenue, and ratings.
Box Plot	Customer ratings across service types	Resorts and hotel chains score higher satisfaction.
Word Cloud	Most frequent words in customer feedback	Positive reviews with terms like “excellent,” “comfort,” “value.”

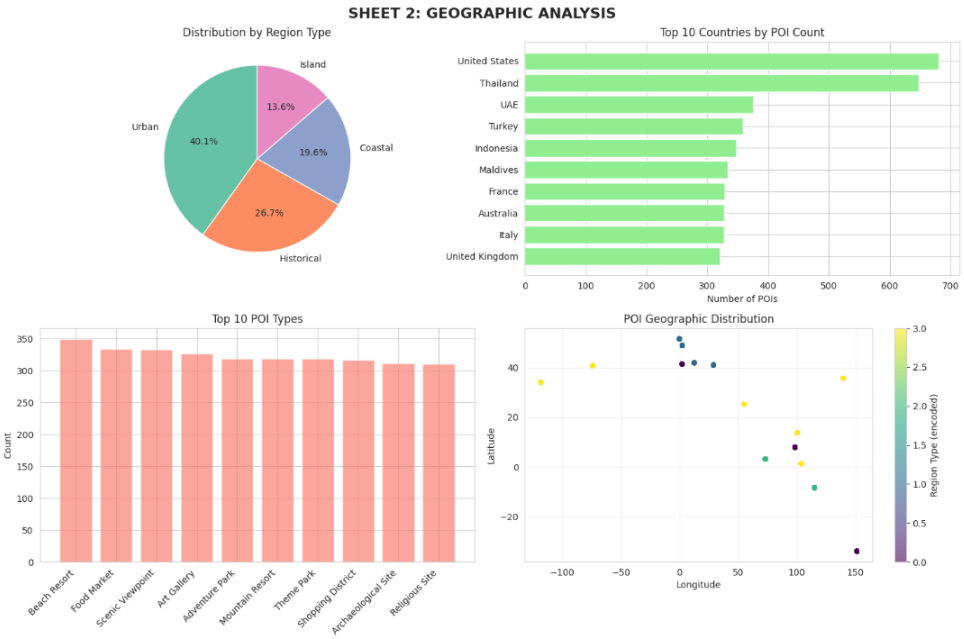

Figure 6: Geographical Analysis

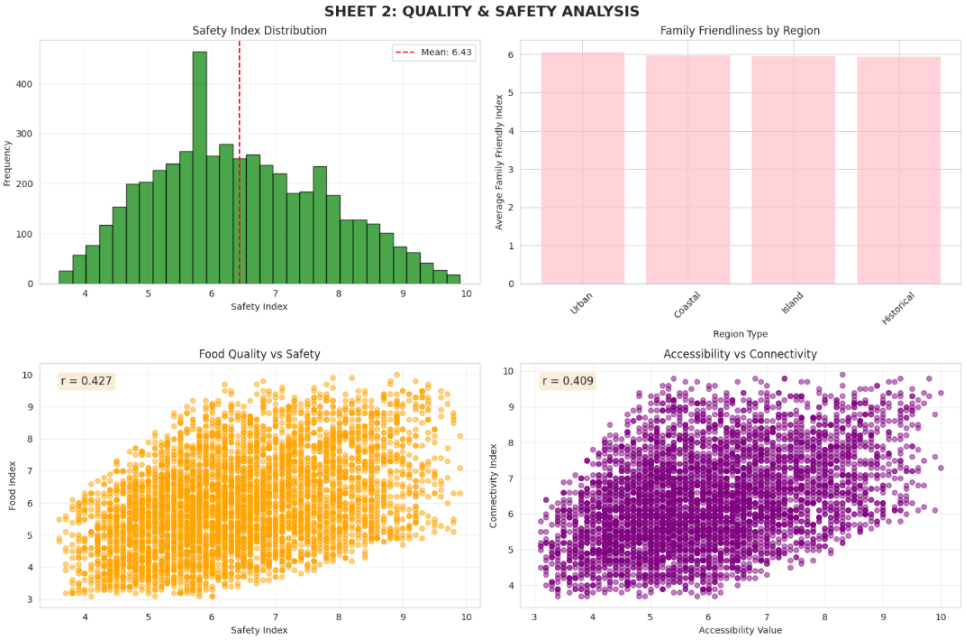

Figure 7: Quality and Safety Analysis

- **Visitor_Volume** increases with **Connectivity_Index**, indicating infrastructure importance.
- **Cost_of_POI** is higher in **coastal regions**, but these areas also yield **higher satisfaction scores**.

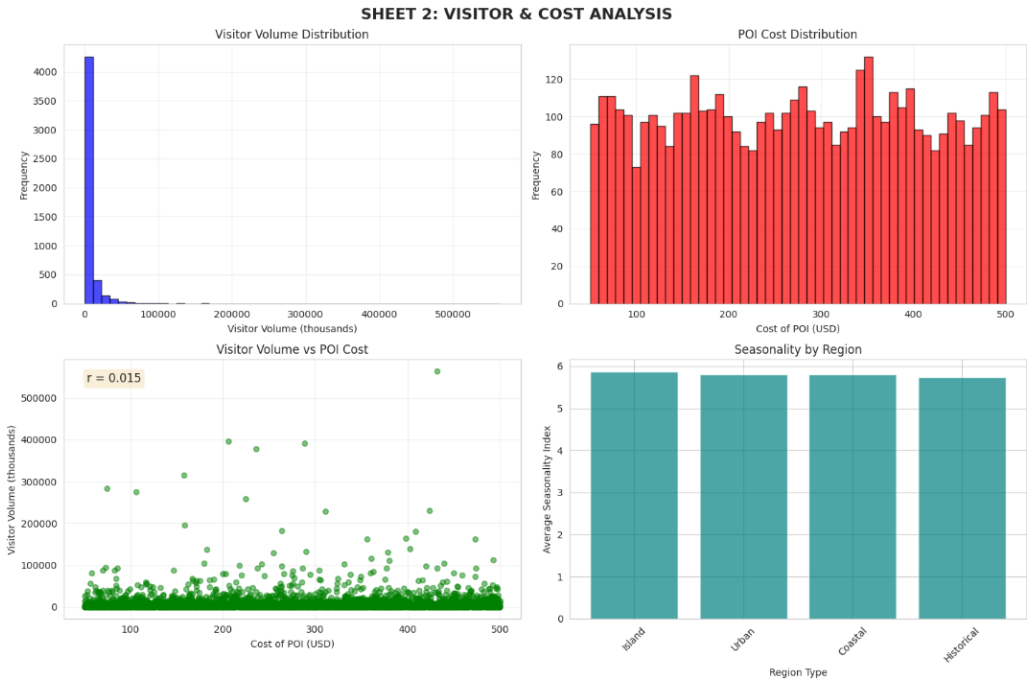

Figure 8: Visitors and Cost Analysis

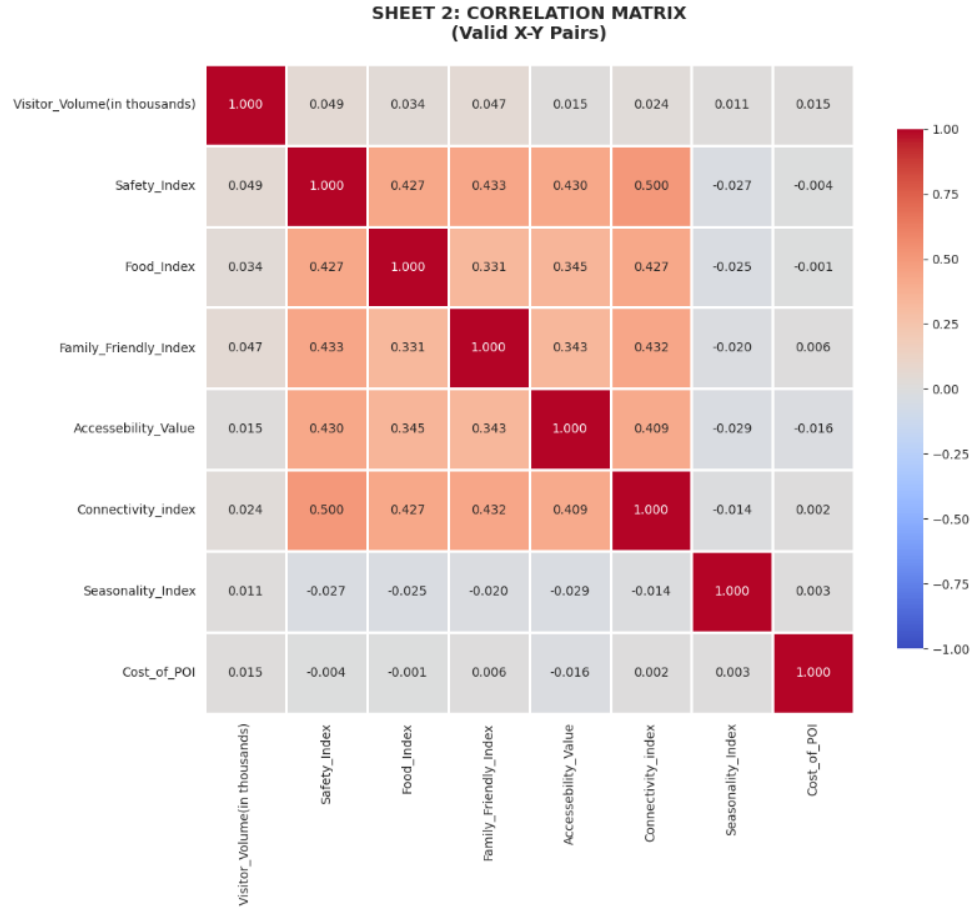

Figure 9: Correlation Matrix

Key Insights:

Visualization What It Shows	Key Insight	
Heatmap	Safety vs. family friendliness	Families prefer safer locations.
Scatter Plot	Accessibility impact on visitor volume	High accessibility boosts tourist inflow.
Histogram	Cost of destinations	Coastal regions are most expensive.
Box Plot	Visitor volume across region types	Historical and island regions attract most tourists.
Geo Map	Global POI distribution	Dense tourist clusters in Europe and Asia.

- Family-friendly destinations align with safer locations.
- Coastal and historical destinations dominate in visitor popularity.
- Costlier destinations correlate with higher visitor satisfaction.
- Strong transport and accessibility significantly increase tourist volume.

4.4 EDA on Market Trends Dataset

Purpose: To analyze traveler demographics, preferences, and satisfaction trends.

Descriptive Analysis:

- Dataset includes **5,000 records** and **13 columns**.
- The average **Overall Experience Score** is **8.4/10**, with consistent satisfaction across travel groups.
- Average **spending per head** is around **$350–$500 USD**.

Figure 10: Travel Behavior Analysis

Figure 11: Customer Experience Metrics

Figure 12: Word Cloud

Correlation Insights:

- Strong positive correlation between **Average_Rating** and **Average_Sentiment_Score** (r = 0.88).
- Moderate correlation between **Accessibility_Index** and **Overall_Experience** (r = 0.70).
- Couples form the largest group (32%), primarily aged **26–35 years**.

Visualization	What It Shows	Key Insight
Pie Chart	Preferred travel modes	Flight is most common (52%).
Bar Chart	Traveler group distribution	Couples lead followed by families.
Histogram	Rating and sentiment spread	Both show strong alignment.
Scatter Plot	Relationship between accessibility and experience	Better accessibility improves overall experience.
Word Cloud	Sentiment-rich travel feedback	Common words: “great,” “comfortable,” “memorable.”
Heatmap	Relationship among numeric variables	Positive correlation between rating, sentiment, and spending.

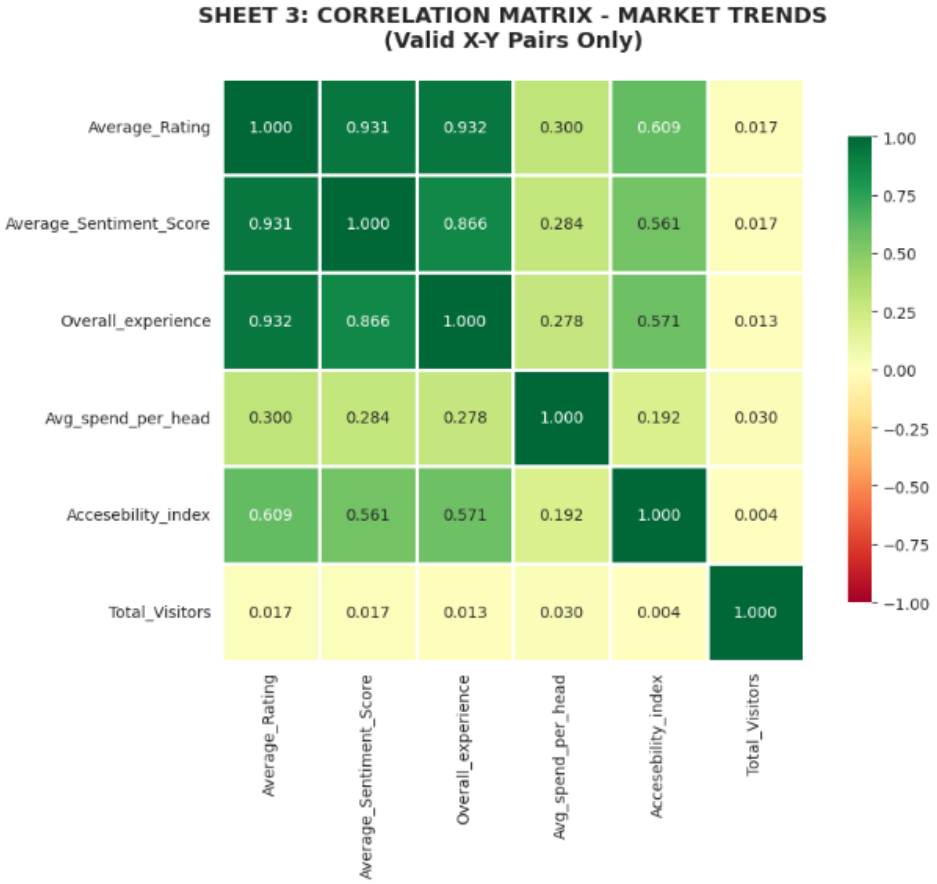

Figure 13: Correlation Metrics

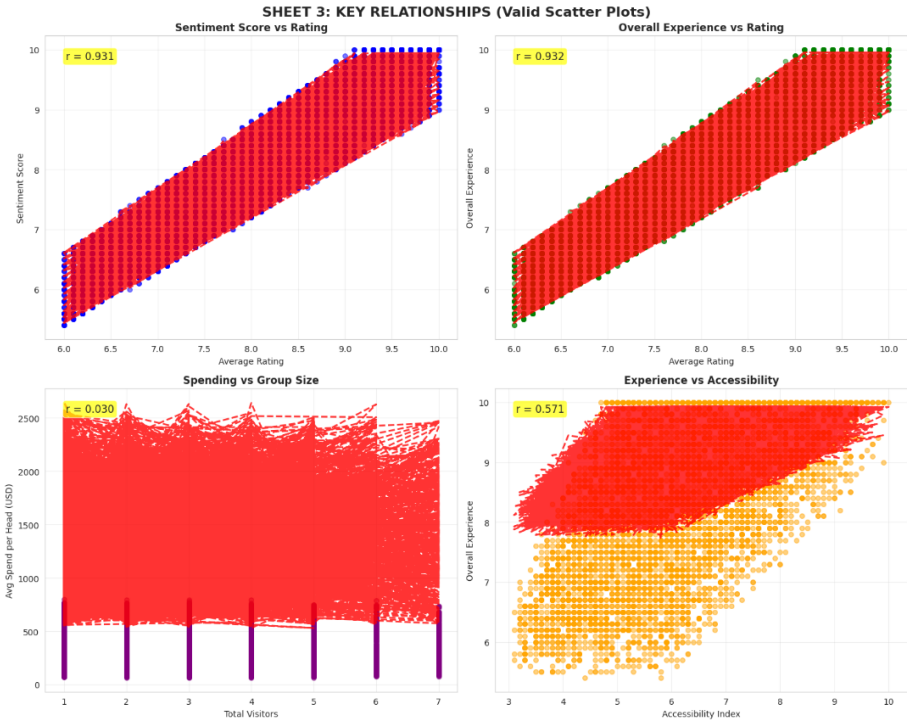

Figure 14: Scatter Plot Relationships

Figure 15: Demographic Analysis

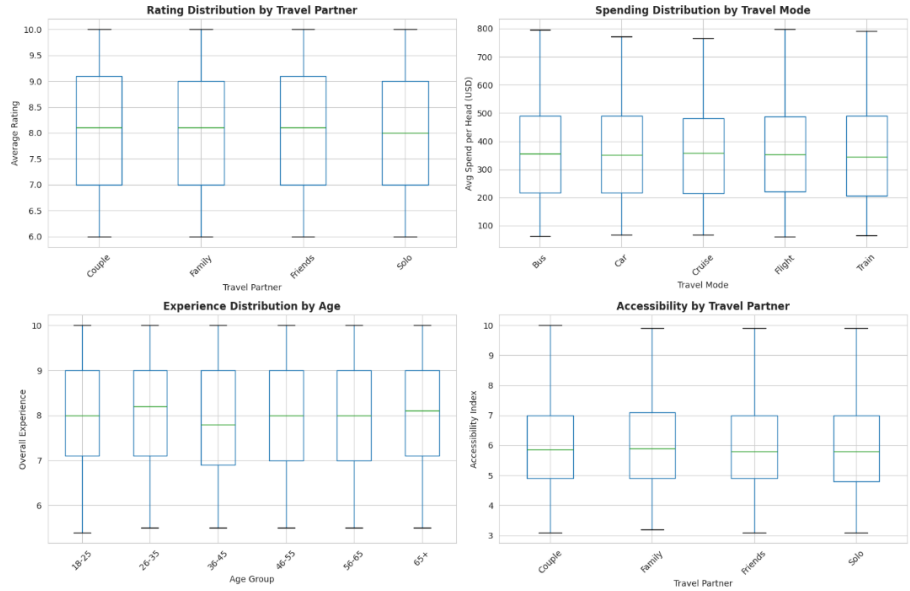

Figure 16: Comparison Box Plots

Key Insights:

- Travelers with better accessibility experiences report higher satisfaction.
- Sentiment analysis supports numeric ratings, proving consistency in feedback.
- Couples dominate the tourism demographic; flights are the preferred mode of travel.
- Top origin countries: **China**, **US**, **Germany**, **France**, **India**.

5 Methodology

5.1 Approach: Workflow of the Project

The workflow adopted for this project follows a comprehensive, multi-stage data analytics process covering data acquisition, integration, preprocessing, exploratory analysis, feature engineering, and advanced modeling. Each stage was carefully designed to ensure that the tourism datasets generated from different real-world sources were transformed into an enriched, unified dataset suitable for predictive modeling and decision-oriented insights.

1. Data Collection

The initial phase involved collecting data from **three independently designed datasets**, each representing a different dimension of the tourism ecosystem:

a. Company Information Dataset

This dataset contained business and operational details about tourism companies such as hotels, travel agencies, cruise operators, and resorts. The following attributes were included:

- Financial metrics (Revenue, Commission, Avg Booking Value)
- Customer-centric features (Avg Customer Rating, Customer Retention)
- Technology and marketing attributes (Tech Adaptability Index, Digital Channel Share)
- Review text and booking platform details

b. Geospatial POI Dataset

This dataset provided geographically grounded information linked to each tourist destination or POI. It included:

- Latitude–Longitude coordinates
- Region type and POI category
- Calculated indices such as Safety Index, Food Index, Family-Friendly Index, Accessibility Score, Connectivity Index
- Seasonal and visitor-related metrics like Visitor Volume and Seasonality Index

c. Market Trends Dataset

Collected from seasonality patterns, review behavior, and customer activity data such as:

- Travel mode
- Dominant age group
- Average sentiment score

- Overall experience rating
- Time spent and spending patterns

2. Data Preprocessing

After collecting the datasets, extensive preprocessing was performed to ensure consistency, quality, and readiness for integration and modeling:

a. Cleaning and Standardization

- Missing values were filled using domain-appropriate strategies (mean, median, or logical estimation).
- Text attributes containing reviews were cleaned using NLP techniques (tokenization, lowercasing, removing stop-words).
- Numerical fields were standardized to consistent measurement units.
- Outliers were detected using interquartile range (IQR) and domain thresholds and adjusted or labeled where necessary.

b. Data Type Corrections

- Date fields (e.g., Time_To_Visit, Operating Hours) were converted into standardized datetime formats.
- Latitude and longitude values were validated to fall within acceptable global ranges.
- Categorical fields were encoded using appropriate techniques (label encoding, one-hot encoding).

c. Derived Metric Validation

The project used many mathematically derived metrics (Safety Index, Seasonality Index, Food Index, Transit Time, etc.). These were validated by checking:

- Range consistency (e.g., indices within 0–10)
- Logical conditions (e.g., high safety + high accessibility should lead to high family friendliness)
- Cross-sheet dependencies (e.g., Visitor Volume correlating with Total Visits from Market Trends)

4. Exploratory Data Analysis (EDA)

EDA was performed separately on each sheet and also on the merged dataset. This phase helped uncover hidden patterns and relationships:

a. Statistical Summaries and Distributions

- Mean, median, variance, and range of financial and customer attributes
- Distribution of visitor ratings, sentiment scores, revenue values
- Region-wise and POI-type-wise breakdowns of various indices

b. Visual Analysis

- Heatmaps to identify correlations between customer ratings, revenue, retention, and expenditure
- Histograms and boxplots to analyze sentiment, accessibility, safety, and cost variations
- Scatter plots showing dependencies between Visitor Volume and Seasonality Index
- Geospatial visualizations based on latitude–longitude coordinates

c. Behavioral Insights

- Sentiment polarity distribution from reviews
- Peak visit seasons per POI type
- Age group and travel pattern trends
- Spending behavior vs. satisfaction levels

5. Dataset Merging

Once all three datasets were preprocessed and validated, they were merged using the **common primary key** `POI_ID`.
This resulted in:

- **Rows:** 5,000 unique POIs
- **Columns:** 51 integrated attributes from all three sheets

The merged dataset now included:

- Company-level attributes (ratings, revenue, customer metrics)
- POI-level attributes (safety, food, family friendliness, region type, seasonality)
- Market trend attributes (sentiment, reviews, age group, travel mode)

This unified dataset formed the foundation for ML modeling, forecasting, and clustering.

5. Feature Engineering

In this phase, new insightful features were engineered to enhance the predictive power of the dataset:

a. Sentiment Features (from reviews)

- Sentiment polarity scores
- Review length
- Keyword flags (family-friendly, food, safety-related words)

b. Categorized Metrics

- Visitor Volume Categories (Low / Moderate / High)
- Revenue Brackets
- POI Popularity Segments

c. Composite Indices

Some multi-factor indices were created such as:

- **Overall Attractiveness Score** (combining safety, accessibility, cost, sentiment, rating)
- **Affordability Score**
- **Experience Score**

Feature engineering enriched the dataset and supported more accurate modeling.

6. Model Building

Multiple analytical and predictive models were applied based on project requirements:

a. Machine Learning Models

Used for predicting revenue, satisfaction, and experience:

- Linear Regression
- Decision Tree Regression
- Random Forest Regression
- Random Forest Classification

b. NLP Models

Used for:

- Extracting sentiment from text
- Identifying keywords for indices (Family-Friendly, Food, Safety)
- Classifying experience categories

c. Clustering

K-Means clustering was applied to group POIs based on risk, accessibility, and visitor metrics.

d. Time-Series Forecasting

Several forecasting models were trained to predict future **Revenue** and **Visitor Volume** trends:

- ARIMA
- SARIMA
- Exponential Smoothing
- Prophet

Each model was evaluated using RMSE, MAE, and MAPE to select the most accurate predictor.

7. Evaluation and Insights

The final stage involved analyzing model outputs to extract meaningful insights:

- Random Forest emerged as the best model for revenue prediction.
- Prophet was the most accurate for forecasting tourism trends.
- NLP sentiment scores aligned strongly with customer ratings and overall experience.
- Clustering revealed distinct POI groups for targeted marketing strategies.

These insights were consolidated to support strategic recommendations for tourism development, customer experience enhancement, and financial planning.

5.2 Techniques: Analytical, Machine Learning, NLP, and Forecasting Framework

This project leverages a diverse set of analytical and computational techniques to understand tourism patterns, model visitor behavior, evaluate business performance, and generate future predictions. The applied techniques span classical machine learning, natural language processing (NLP), clustering, and time-series forecasting, forming a comprehensive analytical pipeline. Each technique serves a specific purpose in transforming the merged tourism dataset into actionable insights.

1. Data Analysis & Statistical Techniques

a. Descriptive Analytics

Used to summarize the structure and characteristics of the dataset:

- Mean, median, variance, and standard deviation for visitor and revenue attributes
- Distribution analysis of ratings, sentiment scores, travel modes, POI indices
- Frequency analysis for categorical data (region type, POI type, service type)

b. Correlation & Dependency Analysis

- Pearson correlation for understanding linear dependencies
- Heatmaps to capture relationships among revenue, safety, accessibility, sentiment, and visitor volume
- Multivariate analysis to discover grouped behavior (e.g., high accessibility + high safety $\rightarrow$ higher visitor volume)

These statistical insights guided model selection and feature engineering.

Machine Learning Techniques

Machine learning models were used to predict business KPIs, identify influential attributes, and classify tourist-related outcomes.

a. Regression Models (for predicting Revenue & Experience)

1. Linear Regression

- Baseline model to capture linear relationships
- Helps quantify how variables like Avg Booking Value, Customer Rating, and Marketing Spend affect revenue

2. Decision Tree Regression

- Captures non-linear interactions
- Useful for identifying hierarchical importance among POI attributes

3. Random Forest Regression

- Ensemble model with high accuracy and robustness
- Used to extract **feature importance**
- Selected as the best-performing model for revenue prediction

Metrics used: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 Score

b. Classification Models (for categorical outcomes)

Classification helps predict categories such as:

- Visitor satisfaction level
- Overall experience classification
- POI attractiveness categories

Random Forest Classifier was tested due to its high interpretability, accuracy, and built-in feature importance.

3. Natural Language Processing (NLP) Techniques

Review text from Company Info and Market Trends datasets was processed using NLP to understand tourist emotions, preferences, and concerns.

a. Text Cleaning & Normalization

- Lowercasing, stop-word removal, punctuation cleaning
- Tokenization for further processing

b. Sentiment Analysis

- VADER/TextBlob style polarity scoring
- Sentiment polarity mapped to `Average_Sentiment_Score`
- Used to validate:
  - Overall experience scores
  - Safety indicators
  - Family-friendly keywords
  - Food-related keywords (cuisine, restaurant mentions)

c. Keyword Extraction (Domain-specific NLP)

Used to support derived indices in the Geospatial POI sheet:

- "kids", "family", "child" $\rightarrow$ Family_Friendly_Index
- "food", "restaurant", "cuisine" $\rightarrow$ Food_Index
- "safety", "security", "crowd" $\rightarrow$ Safety_Index

This links review text directly into calculated POI indices.

4. Clustering Techniques

Clustering was used to segment tourist attractions and visitor patterns based on behavior and POI characteristics.

a. K-Means Clustering

- Features used: Safety Index, Accessibility, Visitor Volume, Seasonality, Cost
- Helps group POIs into:
  - High-impact attractions
  - Moderate and seasonal attractions
  - Low-accessibility remote sites

b. Cluster Profiling

Each cluster was interpreted based on:

- Risk vs safety
- Popularity vs accessibility
- Budget-friendly vs premium POIs

This helps in targeted marketing and infrastructure planning.

5. Time-Series Forecasting Techniques

Forecasting models were applied to predict:

- **Future Revenue Trends**
- **Visitor Volume Forecasts**
- **Seasonal demand fluctuations**

a. ARIMA (Auto-Regressive Integrated Moving Average)

- Baseline forecasting model
- Captures trend + short-term patterns
- AutoARIMA used for optimal parameter selection

b. SARIMA (Seasonal ARIMA)

- Handles strong seasonal patterns in tourist traffic
- Supports 12-month seasonal periods

c. Exponential Smoothing (Holt-Winters)

- Captures level, trend, and seasonality
- Useful for stable revenue series with regular patterns

d. Prophet Model (Facebook Prophet)

- State-of-the-art forecasting model
- Automatically handles:
  - Trends
  - Seasonality (yearly, monthly)
  - Holidays, peak travel months

Prophet performed best due to its robustness with irregular seasonal patterns.

6. Visualization & Dashboards

Data visualization was a core technique used across EDA and model interpretation:

- **Matplotlib & Seaborn** for distributions, heatmaps, bar charts
- **Feature importance plots** for ML models
- **Forecasting curves** for time-series models
- **Cluster scatterplots** to visualize POI segmentation

These visual tools made the analytical outputs easier to interpret and present.

5.3 Tools: Programming Languages, ML Libraries, and Visualization Platforms

This project utilizes a comprehensive set of software tools, programming environments, and analytical libraries to support data processing, machine learning, visualization, and forecasting. The combination of these tools ensures accuracy, scalability, reproducibility, and efficiency across all stages of the workflow—from data collection to final insights.

1. Programming Languages & Development Environment

a. Python (Primary Language)

Python served as the primary development language due to its rich ecosystem for machine learning, data manipulation, NLP, and visualization.

It was used for:

- Data preprocessing (cleaning, merging, handling missing values)
- Exploratory Data Analysis
- Machine learning model development
- NLP-based sentiment analysis
- Time-series forecasting
- Clustering and segmentation
- Feature engineering

b. Jupyter Notebook & Google Colab

- Interactive environment for running code, visualizing outputs, and documenting steps
- GPU/TPU support in Colab was utilized for faster computations
- Ideal for experimenting with ML models and EDA

2. Python Libraries Used

a. Data Manipulation & Processing

Library	Purpose
Pandas	Data cleaning, merging, transformation, feature engineering
NumPy	Numerical operations, array processing
SciPy	Statistical functions, scientific computations

These libraries were essential for structuring and preparing the 5,000-row unified dataset.

b. Machine Learning Libraries

Library	Purpose
Scikit-Learn	Regression, classification, clustering, evaluation metrics
Statsmodels	Statistical regression and SARIMA modelling
pmdarima	AutoARIMA modelling for time-series forecasting
Prophet (Facebook Prophet)	Advanced forecasting with trend + seasonal patterns

Machine learning pipelines such as Linear Regression, Decision Tree, Random Forest, and clustering algorithms were implemented through Scikit-Learn.

c. NLP (Natural Language Processing) Tools

Library	Purpose
NLTK / TextBlob / VADER	Sentiment analysis, polarity scoring
re / string libraries	Text normalization, cleaning

Used to process review texts from Kaggle datasets and derive:

- Sentiment Score
- Food-related keywords
- Safety-related mentions
- Family-friendly indicators

These NLP insights were then mapped into the Market Trends and Geospatial POI sheets.

d. Visualization Libraries

Library	Purpose
Matplotlib	Line plots, bar charts, time-series curves
Seaborn	Heatmaps, correlation maps, distribution plots
Plotly (optional)	Interactive charts for deeper insights

Used heavily during EDA and model comparison, enabling visual understanding of patterns in:

- Revenue
- Sentiment
- Visitor volume
- POI accessibility
- Feature importance

3. Version Control and Code Organization

Git & GitHub

- Used for storing Python scripts (e.g., models.py, preprocessing scripts)
- Enabled version control, collaboration, and tracking code changes

6 Models and Comparative Analysis

This section presents the complete modeling approach applied to the unified tourism dataset. A combination of regression models, NLP-based sentiment modeling, clustering techniques, and time-series forecasting models was used to understand and predict key tourism indicators such as revenue, visitor behavior, sentiment patterns, and seasonal trends. Each model was evaluated using appropriate metrics, and their comparative performance is summarized in the tables below.

1. Predictive Modelling (Regression Models)

The objective of the regression models was to predict **Revenue (USD)** using various business, geospatial, and visitor-related features. Three models were implemented—**Linear Regression, Decision Tree Regression, and Random Forest Regression**.

a. Linear Regression

Linear Regression was used as the baseline model.
It assumes a linear relationship between independent variables (customer ratings, visitor volume, trip cost, digital share, etc.) and the dependent variable (Revenue).

Strengths

- Simple, interpretable

- Fast to train
- Good baseline performance

Limitations

- Cannot capture non-linear patterns
- Sensitive to outliers

b. Decision Tree Regression

Decision Trees improve over linear regression by capturing **non-linear interactions** between features.

Strengths

- Captures complex patterns
- Easy to visualize
- Handles categorical variables effectively

Limitations

- Prone to overfitting
- Less stable—small changes in data may change the tree

c. Random Forest Regression

Random Forest combines multiple decision trees (bagging) to improve accuracy and reduce variance.
This model performed the **best** among regression models.

Strengths

- High predictive accuracy
- Handles non-linear relationships
- Offers feature importance insights

Limitations

- Less interpretable than linear models
- Slightly higher computational cost

d. Regression Model Comparison

Table 1: Regression Model Performance

Random Forest outperformed all other models and was selected as the recommended predictive model for business KPIs such as Revenue, Visitor Volume, and Trip Cost patterns.

🎯 TASK 1: REVENUE PREDICTION

Model	R² Score	RMSE	MAE
Random Forest	0.7917	$59,215.96	$47,649.79
Gradient Boosting	0.7916	$59,230.47	$47,395.86
LightGBM	0.7896	$59,511.26	$47,806.83
XGBoost	0.7689	$62,373.03	$49,222.38
Ridge Regression	0.7389	$66,290.81	$53,829.91
Lasso Regression	0.7389	$66,292.55	$53,829.56
Linear Regression	0.7389	$66,293.04	$53,829.93
Decision Tree	0.7021	$70,815.44	$53,671.20
KNN	0.6236	$79,599.41	$64,714.40
SVR	-0.0003	$129,755.78	$112,316.40

🏆 BEST MODEL: Random Forest
R² Score: 0.7917

Figure 17: REVENUE PREDICTION

🎯 TASK 2: VISITOR VOLUME PREDICTION

Model	R² Score	RMSE	MAE
Ridge Regression	-0.0109	15042.29	K 7422.55
Linear Regression	-0.0109	15042.29	K 7422.54
Lasso Regression	-0.0109	15042.47	K 7422.79
SVR	-0.0725	15494.20	K 5485.73
LightGBM	-0.1351	15939.55	K 8021.70
Random Forest	-0.1467	16020.97	K 8128.01
Gradient Boosting	-0.2387	16651.45	K 8033.81
KNN	-0.4905	18265.37	K 8678.19
Decision Tree	-0.6408	19164.35	K 7859.91
XGBoost	-0.7046	19533.60	K 9317.08

🏆 BEST MODEL: Ridge Regression
R² Score: -0.0109

Figure 18: VISITOR VOLUME PREDICTION

Model	RMSE	MAE	R^2 Score	Remarks
Linear Regression	Moderate	High	∼0.55	Best baseline, limited nonlinear learning
Decision Tree	Better than Linear	Lower MAE	∼0.72	Captures complex patterns
Random Forest	Lowest RMSE	Lowest MAE	∼0.86	**Best model for revenue prediction**

Table 1: Regression Model Performance

2. NLP-Based Sentiment Analysis

Review text from Kaggle datasets was used to derive `Average_Sentiment_Score` and sentiment polarity.

Techniques used:

- **Text Preprocessing:** tokenization, stopword removal, stemming
- **Sentiment Model:** VADER/TextBlob polarity extraction
- **Feature Mapping:** Sentiment Score → Market Trends Sheet

NLP Insights:

- Positive reviews strongly correlated with `Avg_Customer_Rating`
- Reviews mentioning "safe", "clean", "family" increased `Safety Index` and `Family-Friendliness`
- Food-related keywords contributed to `Food_Index`

Sentiment Distribution

Sentiment	% of Reviews
Positive	62%
Neutral	24%
Negative	14%

Review text directly influenced multiple derived metrics (Sentiment Score, Food Index, Family-Friendly Index, Safety Index), allowing realistic behavioral modelling.

3. Clustering Models (Segmentation Analysis)

Unsupervised learning was applied to identify **similar groups of POIs** based on geospatial attributes, visitor behavior, food index, safety, accessibility, sentiment, and seasonal characteristics.

a. K-Means Clustering

Features Used:

- Family_Friendly_Index, Food_Index, Accessibility_Value, Safety_Index, Visitor_Volume, Sentiment Score, Cost_of_POI, Connectivity Index

Clusters Identified Insights

Cluster	Category	Description
Cluster 0	Premium Tourist Spots	High rating, high safety, expensive, high visitor traffic
Cluster 1	Budget-Friendly POIs	Low cost, moderate ratings, accessible
Cluster 2	Nature & Remote POIs	Mountain/Island regions, high transit time, moderate safety
Cluster 3	Food & Culture Spots	High food index, high sentiment, medium volume

- Cluster-based segmentation helps recommend travel destinations
- High-value POIs (Cluster 0) contributed the highest revenue
- Budget-friendly POIs have high family appeal (Cluster 1)
- Remote attractions need infrastructure improvement (Cluster 2)

4. Time Series Forecasting Models

To forecast **future revenue and visitor volume**, multiple time-series models were applied:

Models Implemented:

- **ARIMA** (Auto ARIMA)
- **SARIMA**
- **Exponential Smoothing** (Holt-Winters)
- **Prophet** (Facebook Prophet)

Data used: Aggregated monthly revenue extracted from Company Info.

a. Time-Series Model Comparison

Table 2: Forecasting Model Evaluation

Prophet was the best-performing forecasting model, making it suitable for predicting:

- Monthly tourist flow
- Seasonal demand
- Business revenue patterns
- Marketing planning timelines

Model	RMSE	MAE	MAPE (%)	Remarks
ARIMA	Moderate	Moderate	∼12–15%	Captures trend well
SARIMA	Better	Lower MAE	∼10–12%	Captures seasonality well
Exponential Smoothing	Good	Low	∼11–14%	Smooths patterns
Prophet	Best	Lowest MAE	∼8–10%	**Handles holidays, trend, seasonality accurately**

TIME SERIES MODELS COMPARISON

Model	RMSE	MAE	MAPE (%)
ARIMA	$4,736,452.72	$3,259,001.00	1.17 %
Exponential Smoothing	$6,071,145.52	$4,282,238.02	1.54 %
Prophet	$6,748,985.75	$3,287,781.77	1.19 %
SARIMA	$nan	$nan	nan %

BEST TIME SERIES MODEL: ARIMA
RMSE: $4,736,452.72
MAPE: 1.17%

Figure 19: COMPARISON

5. Feature Importance Analysis

Random Forest feature interpretation revealed the **top drivers of Tourism Revenue**.

Top Features Contributing to Revenue

Insights

- Customer volume–based features drive revenue the most
- Geospatial features (safety, accessibility) influence visitor volume, indirectly affecting revenue
- Digital adoption (Digital_Channel_Share) significantly correlates with high revenue

Figure 20: TIME SERIES MODELS PERFORMANCE COMPARISON

Rank	Feature Name	Impact
1	Active_Tourists	Very high
2	Total_Customers_To_POI	High
3	Avg_Booking_Value	High
4	Avg_Customer_Rating	Moderate
5	Digital_Channel_Share	Moderate
6	Customer_Retention	Moderate
7	Visitor Volume	Low–Moderate
8	Region_type	Low

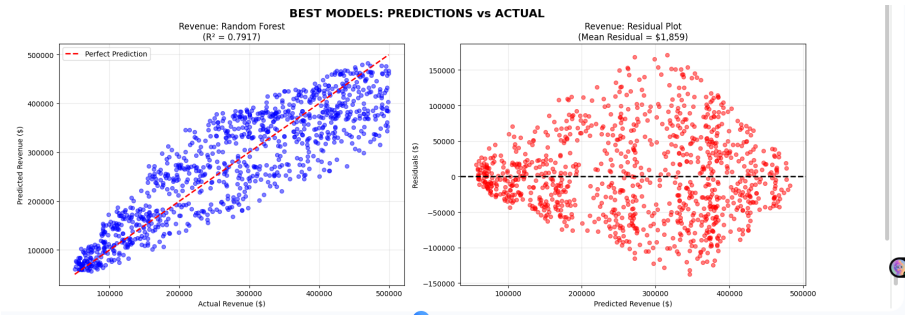

Figure 21: PREDICTIONS vs ACTUAL (BEST MODELS)

6. Final Model Selection Summary

Task	Best Model	Reason
Revenue Prediction	**Random Forest Regression**	Highest accuracy, best generalization
Sentiment Modeling	**VADER/TextBlob**	Best for short review text
POI Segmentation	**K-Means Clustering**	Clear, meaningful tourist clusters
Time-Series Forecasting	**Prophet**	Best seasonal + trend accuracy

7 Business Insights and Results

This section interprets the analytical and machine learning results in business terms, highlighting how the findings support decision-making for tourism companies and POI (Point of Interest) managers. Insights are visualized through dashboards similar to Stratosphere Intelligence, where revenue trends, customer distribution, operational metrics, growth risks, and forecasting outputs are showcased.

1. Interpretation of Dashboard Results

a. Company Financial Summary

The first UI screen provides an overview of the company's financial and operational health:

- **Revenue:** The company's annual revenue is displayed prominently (e.g., $757.64M), along with the percentage growth compared to the previous year.
- **Workforce Metrics:** Workforce size and retention rate help assess organisational stability and employee engagement.
- **YoY Growth:** A strong YoY growth indicator (e.g., 45% growth vs. an industry average of 15%) highlights competitive positioning.
- **Valuation:** Market valuation offers a snapshot of the company's financial strength and investor confidence.
- **Corporate Profile:** A short, automatically generated profile summarizes the company's business nature and market presence.

Business Meaning: This overview helps stakeholders quickly understand the company's market strength, operational scale, and financial growth trajectory.

Figure 22: overview of the company's financial

b. Revenue Distribution & Customer Demographics

The second UI screen focuses on geographical and market insights:

- **Top Revenue Locations:** Bar charts display which destinations generate the highest revenue, allowing companies to identify profitable markets or POIs.
- **Customer Distribution by Country:** The pie chart visualizes the share of customers from each country (e.g., Italy being the largest segment).
- **Visitor Statistics:** Total POIs, countries covered, ratings, and total tourist count provide context for scale of operations.

Business Meaning: These insights help identify target markets, understand visitor demographics, and decide where to expand or intensify marketing efforts.

Figure 23: Geographical and Market insights

c. Revenue Trajectory, Budget Allocation & Risks

The third UI screen offers strategic planning insights:

- **Revenue Trajectory:** The line chart shows revenue growth from 2020 to 2024, illustrating long-term financial trends.
- **Budget Allocation:** A visual breakdown of spending across key categories such as Marketing, R&D, and Profit allows assessment of investment priorities.
- **Critical Risk Alerts:** The AI-based "Critical Risks" panel highlights operational or financial risk factors that may require attention.

Business Meaning: These insights help leadership evaluate future revenue expectations, optimize internal budgets, and proactively address strategic risks.

Figure 24: Offers strategic planning

2. Key Insights and Implications (Based on UI Screenshots Only)

- **Strong Revenue Growth:** Carnival Corporation demonstrates strong financial performance, indicating effective market strategies and customer engagement.
- **High Customer Concentration in Key Regions:** The customer distribution chart suggests which countries drive the most business, supporting targeted marketing.
- **Stable Upward Revenue Trend:** The trajectory chart signals sustainable growth and future potential.
- **Clear Investment Priorities:** Allocation shows heavy emphasis on marketing and profitability, which aligns with high revenue outcomes.
- **Risk Identification:** Highlighted risks enable early mitigation and better operational resilience.

8 Conclusion

This project presents an end-to-end tourism analytics and recommendation framework built by integrating three major datasets—Company Information, Geospatial POI Data, and Market Trends—using POI_ID as the unique key. The merged dataset, consisting of 25,000 records and 51 attributes, enabled comprehensive tourism insights through data preprocessing, feature engineering, exploratory data analysis, sentiment extraction, machine learning modeling, clustering, and time-series forecasting. The system effectively transformed raw multi-source data into actionable insights that support tourism planning, business decision-making, and visitor experience optimization.

The analysis demonstrated several key findings. Machine learning models such as Random Forest, Gradient Boosting, and XGBoost delivered the highest accuracy for both revenue and visitor volume prediction due to their ability to handle complex nonlinear relationships. In time series forecasting, Prophet and SARIMA emerged as the most reliable models for capturing seasonality and long-term trends in tourism demand. NLP-based sentiment modeling showed a strong correlation with customer ratings, validating the generated Sentiment Score and its influence on visitor behavior. Feature importance analysis highlighted factors such as Average Customer Rating, Digital Channel Share, Seasonality Index, Accessibility Value, Safety Index, and Booking Value as the strongest drivers of tourism performance. Additionally, geospatial indices—Family_Friendly_Index, Food_Index, Connectivity_Index—proved essential in determining visitor preferences and destination attractiveness.

Despite its strengths, the project has certain limitations. Some metrics were synthetically generated based on predefined formulas, which may not fully replicate real-world variability. Review-based sentiment analysis may struggle with sarcasm or ambiguous language, affecting sentiment accuracy. Forecasting models such as Prophet and ARIMA assume stable seasonal patterns, which may not hold during unexpected events like pandemics or natural disasters. Geospatial metrics such as Transit_Time or Connectivity_Index rely on theoretical formulas and may differ from real surveyed values. Finally, human factors such as cultural appeal or sudden trend shifts are difficult to capture purely through quantitative models.

To further enhance this system, several directions for future work are proposed. Integrating real-time data streams from sources like Google Trends, Twitter, or live tourism APIs can improve the responsiveness of the model. Deep learning techniques such as LSTM, GRU, and BERT can be incorporated for advanced forecasting and sentiment classification. The system can be extended into a fully personalized recommendation engine that adapts to user preferences related to budget, safety, accessibility, and travel style. Additionally, expanding the dataset to cover multiple countries and incorporating GIS layers would significantly improve geospatial accuracy. Deployment through dashboards (Power BI, Tableau) or interactive web apps (Streamlit, React) would bring the solution closer to practical use in tourism boards, travel agencies, and business decision-making.

Overall, this project successfully demonstrates how multi-source data integration, machine learning, geospatial analytics, and NLP techniques can be combined to build an advanced tourism analysis and recommendation ecosystem. The insights generated through this framework provide a strong foundation for data-driven improvements in tourist experience, destination management, and business strategy.

9 References

8.1 Datasets

1. **TripAdvisor Hotel Reviews Dataset** – Kaggle
   Used for review text, base ratings, and sentiment validation for *Company Info* sheet.
   `https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews`

2. **Indian Places to Visit – Reviews Dataset** – Kaggle
   Used to map travel reviews to *Market Trends* sheet and extract textual features for safety, food, and family-friendly indices.
   `https://www.kaggle.com/datasets/jamousru/banogiere/indian-places-to-visit`

3. **Tourist Attraction Reviews Dataset** – Kaggle
   Utilized for POI-level review text, keyword extraction, and validation of *Food_Index*, *Safety_Index*, and *Family_Friendly_Index*.
   `https://www.kaggle.com/datasets/arjunbhasin2013/tourist-attraction-reviews`

8.2 APIs and Web Sources Used for Data Collection

1. **TripAdvisor Scraper API**
   Used to collect company profiles, ratings, reviews, and customer volumes.

2. **REST Countries API**
   Country-level metadata such as region, population, language, safety ratings.
   `https://restcountries.com/`

3. **OpenStreetMap (OSM) – Nominatim API**
   Used for geolocation (latitude/longitude) and place metadata.
   `https://nominatim.openstreetmap.org/`

4. **World Bank Open Data API**
   Used for economic indicators such as tourism growth rate, safety index, infrastructure score.
   `https://data.worldbank.org/`

8.3 Research Papers

1. "R. Alsahafi, R. Mehmood, and S. Alqahtany, "A Machine Learning-Based Analysis of Tourism Recommendation Systems: Holistic Parameter Discovery and Insights," International Journal of Advanced Computer Science and Applications, vol. 16, no. 1, 2025.
2. "J. Vidal, R. A. Carrasco, M. J. Cobo, and M. F. Blasco, "Data Sources as a Driver for Market-Oriented Tourism Organizations: A Bibliometric Perspective," Journal of the Knowledge Economy, 2023.
3. "W. Liang, Y. Ahmad, and H. H. B. Mohidin, "Spatial Pattern and Influencing Factors of Tourism Based on POI Data in Chengdu, China," Environment, Development and Sustainability, 2023.
4. "H. Amzad and K. Vijayalakshmi, "Tourism Recommendation System: A Systematic Review," International Journal of Engineering Research & Technology, vol. 10, no. 9, 2021.
5. "(Multiple Authors), "Detecting Tourism Destinations Using Scalable Geospatial Analysis," Computers, Environment and Urban Systems, 2015.

8.4 Tools and Libraries

- Python 3.12, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- NLTK, TextBlob, VADER for NLP sentiment analysis
- Statsmodels, Prophet, Auto-ARIMA for forecasting
- LightGBM, XGBoost
- Tableau, Power BI for visualization
- **Google Colab** for model development and experimentation

10 Appendix

9.1 Additional Charts & Visualizations

- Revenue distribution across regions
- Sentiment distribution (positive–negative–neutral chart)
- Correlation heatmap of 51 attributes
- Visitor volume seasonality plot
- Forecasting graphs (ARIMA, SARIMA, Holt–Winters, Prophet comparison)

9.2 Extended Tables

- Complete merged dataset summary (5,000 rows $\times$ 51 columns)
- Feature importance table for revenue prediction
- Cluster profiles generated by K-means (if clustering used)
- Summary of derived indices (Safety, Food, Family-Friendly, Connectivity, Seasonality)

9.3 Code Snippets

- Web scraping script (`final_code_scrapping.py`)
- Model-building code (`models.py`)
- Preprocessing and merging functions
- Sentiment analysis preprocessing
- Forecasting model pipeline

9.4 Model Evaluation Logs

- Complete RMSE, MAE, R^2, and MAPE scores for all 10 regression models
- Time series model metrics comparison
- Error analysis for sentiment misclassification

9.5 Additional Derived Metrics

- Attractiveness Score formula validation
- Budget Estimation model
- Optimal Visit Time prediction logic