# Unit 2: Tourism Data Collection, Preprocessing and Exploratory Data Analysis

Course: AI and Tourism – MIT-AI @ Gandaki University

Bidur Devkota, PhD
GCES Pokhara

December 17, 2025



## Unit 2 Overview

**Duration:** 8 Hours

**Learning Objectives:**

- Understand tourism data sources and collection methods
- Learn to preprocess and transform raw tourism data
- Apply descriptive statistics and visualization techniques
- Conduct exploratory analysis to identify trends and patterns

**Structure:**

1. Data Sources and APIs
2. Data Preprocessing
3. Descriptive Statistics & Visualization
4. Case Study Applications



## 2.1 Data Sources and APIs

**Modern Tourism Data Landscape**

**Traditional Sources**

- Government surveys
- Airport arrival records
- Hotel registration data
- Tourism board reports

**Digital Sources**

- Online booking platforms
- Social media posts
- Review websites
- Mobile app data

**Key Technologies**

VGI (Volunteered Geographic Information) + APIs = Real-time Tourism Insights



## Volunteered Geographic Information (VGI)

**Definition:** Geographic data created by non-experts via digital platforms

**Tourism VGI Examples:**

- TripAdvisor reviews/ratings
- Google Maps location reviews
- Instagram geotagged posts
- Foursquare check-ins
- Travel blog posts

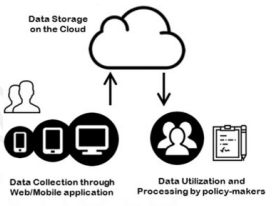



Figure: VGI Data Flow

https://link.springer.com/article/10.1007/s13132-017-0504-y

**Impact**

Provides real-time, localized insights impossible with traditional surveys

## APIs for Tourism Data Collection

| API | Data Provided |
|---|---|
| Google Places API | Location details, reviews, popularity times, photos |
| TripAdvisor API | Hotel/attraction reviews, ratings, rankings |
| Booking.com API | Availability, prices, property features |
| OpenWeather API | Weather conditions, forecasts |
| Skyscanner API | Flight prices, availability |

Table: Major Tourism Data APIs



## API Considerations

**Key Considerations:**

- API rate limits
- Data usage terms
- Authentication
- Data pagination
- Error handling



## 2.2 Data Preprocessing

**Why Preprocess Tourism Data?**

**Common Issues**

- Missing values
- Inconsistent formats
- Duplicate entries
- Outliers
- Unstructured text

**Preprocessing Steps**

1. Data cleaning
2. Handling missing data
3. Text processing
4. Feature engineering
5. Normalization

**Golden Rule**

Garbage In = Garbage Out
Quality preprocessing = Reliable insights



## Processing Unstructured Tourism Data

**Text Processing Pipeline:**

1. Tokenization
2. Lowercasing
3. Stopword removal
4. Stemming/Lemmatization
5. Sentiment analysis
6. Entity recognition

**Example Review:**

*"Amazing hotel! Great location near the beach. Service was exceptional but rooms were small."*

**Extracted Features:**

- Sentiment: Positive (0.7)
- Keywords: hotel, beach, service, rooms
- Aspects: location, service, room size

## Tourism Feature Engineering

| Raw Feature | Engineered Features |
|---|---|
| Arrival Date | Season, Day of Week, Holiday Flag, Peak Season Flag |
| Hotel Price | Price Category, Relative Price Index, Discount Percentage |
| Review Text | Sentiment Score, Topic Categories, Review Length |
| Location Coordinates | Distance to City Center, Proximity to Attractions |
| Booking Date | Advance Booking Days, Lead Time Category |

Table: Feature Engineering Examples



## 2.3 Descriptive Statistics & Visualization

**Descriptive Statistics for Tourism**

**Central Tendency**
- Average stay duration
- Mean tourist spending
- Median hotel price
- Mode travel package

**Dispersion**
- Range of tour prices
- Variance in monthly arrivals
- Standard deviation of ratings
- IQR of spending patterns

**Application Example**

Calculating average daily rate (ADR) and revenue per available room (RevPAR) for hotel performance analysis



## Tourism Data Visualization Techniques

**Temporal Patterns**
- Line charts (arrivals over time)
- Heatmaps (seasonal demand)
- Calendar plots (daily occupancy)

**Spatial Patterns**
- Choropleth maps
- Point density maps
- Flow maps (tourist movement)

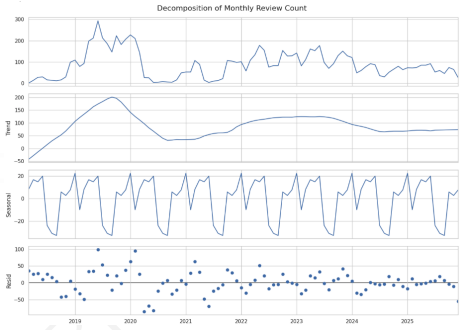
**Comparative Analysis**
- Bar charts (spending by category)
- Pie charts (visitor demographics)
- Radar charts (destination attributes)

**Relationship Analysis**
- Scatter plots (price vs. rating)
- Correlation matrices
- Bubble charts



## Visualization Examples

**Time Series Decomposition**

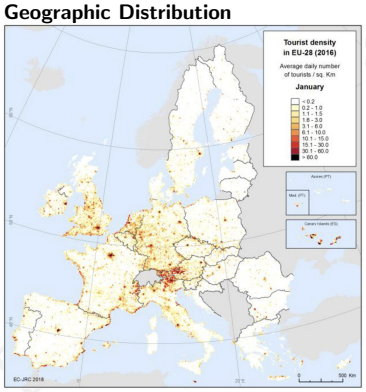



timeseries decomposition(Trend + Seasonality + Residuals) of reviews of Bindabasini Temple

**Tools:** matplotlib, seaborn, plotly, Tableau, Power BI

## Visualization Examples

**Geographic Distribution**




Tourist density by region (choropleth map)

https://www.reddit.com/r/MapPorn/comments/gibz5q/tourist_density_in_eu_january_vs_august/



## 2.4 Case Study: Nepal Tourism Analysis

**Dataset:** Nepal Tourism Statistics 2023 (2014-2023)

**Key Data Points from Report:**

- Total tourists in 2023: **1,014,882** (65.05% growth from 2022)
- Top nationalities: India (31.52%), USA (9.88%), China (5.99%), UK (5.20%)
- Average length of stay: 13.2 days
- Foreign exchange earnings: $548.2 million (68% increase)
- Peak months: October (117,306), March (99,426), April (98,773)

**Objectives:**

1. Analyze post-COVID recovery patterns
2. Identify changing nationality trends
3. Assess seasonality in Himalayan tourism
4. Evaluate adventure tourism growth (trekking/expeditions)
5. Analyze pilgrimage tourism patterns



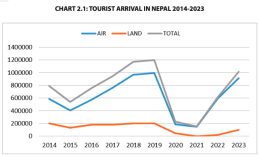
## Nepal Tourism Trend Analysis (2014-2023)

**Key Trends from Data:**

- **Pre-COVID Peak:** 1,197,191 tourists in 2019
- **COVID Impact:** 80.7% drop in 2020 (230,085 tourists)
- **Recovery:** Gradual increase 2021-2023
- **2023 Recovery:** Back to 85% of pre-COVID levels
- **Growth Drivers:** Indian tourists (31.5%), Return of Chinese tourists

**Notable Shifts:**

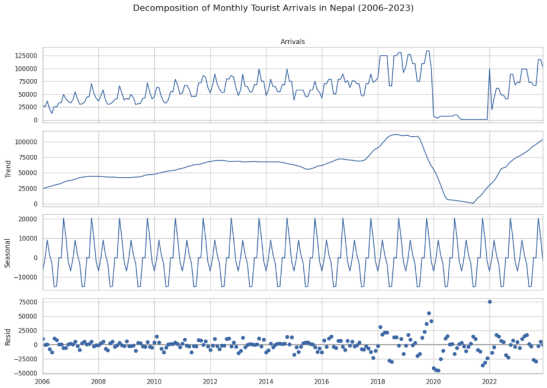
- **Mode of Entry:** 90.09% by air, 9.91% by land
- **Purpose:** Holiday (62.5%), Trekking (15.2%), Pilgrimage (13.1%)




Tourist Arrivals 2014-2023



## Nepal Tourism Seasonality Analysis

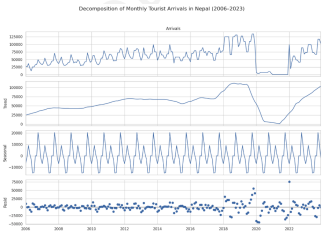



Residual: Deviations from trend (e.g., 2015 earthquake, 2020–2021 COVID drop)

## Nepal Tourism Seasonality Analysis

**Observed:**

- raw monthly data
- Trend — Long-term growth (rise until 2019, collapse in 2020–2021 due to COVID, strong recovery in 2022–2023).
- Seasonal — Consistent yearly cycle: peaks in October–November (post-monsoon/autumn trekking) and March (spring), lows in June–August (monsoon) and January.
- Residual — Irregular variations/noise, with large negative residuals in 2020–2021



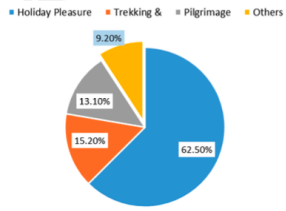

Residual: Deviations from trend (e.g., 2015 earthquake, 2020–2021 COVID drop)



## Nepal Tourism Market Segmentation

**By Purpose of Visit (2023):**

- **Holiday/Pleasure:** 634,301 (62.5%)
- **Trekking & Mountaineering:** 154,262 (15.2%)
  - Expedition teams: 2,253
  - Climbers: 9,398
  - Royalty: Rs. 1,013.6 million
- **Pilgrimage:** 132,949 (13.1%)
  - Lumbini visitors: 903,883 total
  - Pashupatinath: 118,531 foreign visitors
- **Other:** 93,370 (9.2%)




Purpose of Visit Distribution (2023)

**Adventure Tourism Highlights:**

- Popular peaks: Everest, Ama Dablam, Manaslu
- TIMS trekkers: 28,691 in 2023



## Geographic Distribution & Entry Points

**Top Entry Points (2023):**

- **By Air:** 914,270 (90.09%) - TIA Kathmandu, Gautam Buddha Airport
- **By Land:** 100,612 (9.91%)
  - Belahiya (India border): 81,338 (81% of land arrivals)
  - Rasuwagadi (China border): 10,447
  - Kakarbhitta (India): 3,125
  - Birgunj (India): 2,006

**Top Source Markets Analysis:**

| Country | 2023 Visitors | Change from 2022 |
|---|---|---|
| India | 319,936 | +52.8% |
| USA | 100,355 | +30.2% |
| China | 60,878 | +534.1% |
| UK | 52,865 | +17.2% |
| Australia | 38,798 | +44.2% |
| Bangladesh | 36,483 | +43.8% |



## Economic Impact Analysis

**Foreign Exchange Earnings:**

- **2023 Earnings:** $548.2 million (Rs. 72.46 billion)
- **Growth:** 68% increase from 2022
- **Daily Spending:** $41 per tourist per day
- **GDP Contribution:** 1.98% of GDP (hotel & restaurant sector)

**Tourism Infrastructure (2023):**

- **Star Hotels:** 182
- **Non-star Hotels:** 1,416
- **Total Beds:** 54,370
- **Travel Agencies:** 4,845
- **Trekking Agencies:** 3,191
- **Tourist Guides:** 5,123
- **Trekking Guides:** 26,292

**Key Insight**

Tourism recovery significantly contributes to foreign exchange reserves and employment generation

## Hands-on Exercise: Nepal Tourism Data Analysis

**Dataset:** Nepal Tourism Statistics 2023 (provided PDF)

**Tasks:**

1. Clean and structure tourist arrival data (2014-2023)
2. Analyze seasonality patterns for different nationalities
3. Calculate recovery rates post-COVID by market segment
4. Visualize trekking and expedition trends
5. Identify emerging markets and declining trends
6. Forecast 2024 arrivals based on historical patterns

**Expected Deliverables**

- Time-series analysis of tourist arrivals
- Market segmentation dashboard
- Seasonality heatmaps
- Recovery trend analysis
- Strategic recommendations for tourism board



## Recommended Tools & Libraries

**Python Ecosystem:**

- `pandas` - Data manipulation
- `numpy` - Numerical computing
- `matplotlib/seaborn` - Visualization
- `scikit-learn` - Machine learning
- `statsmodels` - Statistical analysis
- `geopandas` - Spatial analysis
- `nltk/spaCy` - Text processing

**Other Tools:**

- R with tidyverse
- Tableau/Power BI
- Google Data Studio
- QGIS (spatial analysis)
- SQL databases

**Data Sources:**

- UNWTO datasets
- National tourism boards
- Kaggle tourism datasets
- API collections



## Key Takeaways

1. **Quality preprocessing** is essential for reliable analysis
2. **Visualization** reveals patterns not obvious in raw data
3. **Seasonality and trends** drive strategic decisions
4. **Outlier detection** helps identify opportunities and risks

**Professional Competency**

Ability to transform raw tourism data into actionable business intelligence



## References & Further Reading


- Smith, J. (2022) *Tourism Analytics: Data Driven Decision Making*
- Chen, L., & Wang, Y. (2021) *Big Data in Tourism*
- UNWTO (2023) *Global Tourism Data Standards*
- Google (2023) *Travel Insights with Google*
- Kaggle Tourism Datasets `https://www.kaggle.com/tags/tourism`
- GitHub Tourism Analytics Projects `https://github.com/topics/tourism-analytics`
- Nepal Tourism Statistic 2023 `https://trade.ntb.gov.np/wp-content/uploads/2024/06/Nepal-Tourism-Statistic_2023-final.pdf`


**Thank You**

Questions?