Foreign Tourist Arrivals and their Economic Impact on India

Analysis by Group 9

Dishant Patel - 202201260
Shravan Kakadiya - 202201333
Madhav Kanjilimadom - 202203018

Problem Statement

How can we analyze global trends in foreign tourist arrivals (FTAs) to India and their impact on foreign exchange earnings (FEE) to identify high-growth regions, optimize seasonal and demographic strategies, and enhance India’s overall tourism competitiveness?

Objectives

1. Regional Insights - Identify the regions contributing the most to FTAs and FEE and assess their growth trends over time.
2. Forecasting Tourism Trends - Predict continent-wise FTAs for future years to create data-driven planning and decision-making.
3. Demographic and Seasonal Analysis - How age groups and seasonal trends affect tourism patterns.
4. Maximize Economic Impact: Use data and results from our analysis to propose strategies to enhance tourism revenue.

Understanding Our Data

Data Available at data.gov.in:

Total FTAs in India: 2001 - 2023

Continent & Country-wise FTAs in India: 2001 - 2021

Quarterly Arrival of FTAs in India: 2001 - 2019

Age Group-wise FTAs in India: 2001 - 2019

FEE due to Tourism: 2001 - 2023

India World Tourism Rank: 2001 - 2021

Two Datasets - Geographic vs Demographic Analysis

Dataset 1:

- Continent and Country-wise FTAs for 21st century.
- 2 Categorical features: Continent & Country.
- 21 Numerical Features: FTAs for each year.

- Goal: Predictive analysis of FTA for future years.

Dataset 2:

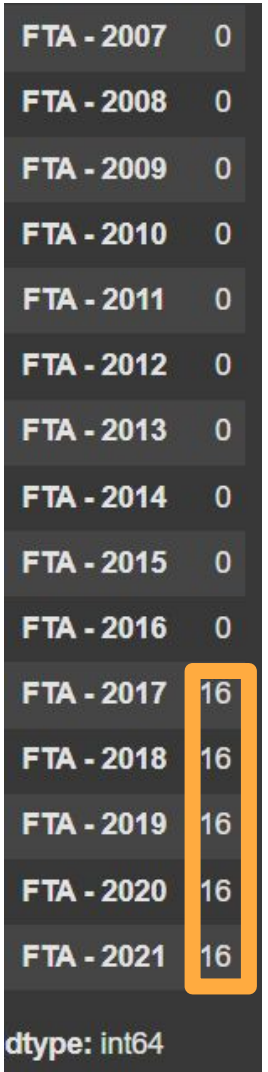
- Year-wise over FTAs in India with demographic features.
- All numerical data:
- Growth in FTAs and FEE per year,
- Age-group wise distribution of FTAs,
- Quarterly distribution of FTAs,
- Tourism share percentage in India,
- Rank of India in world tourism.
- Goal: Understanding seasonal patterns to suggest optimizing strategies.

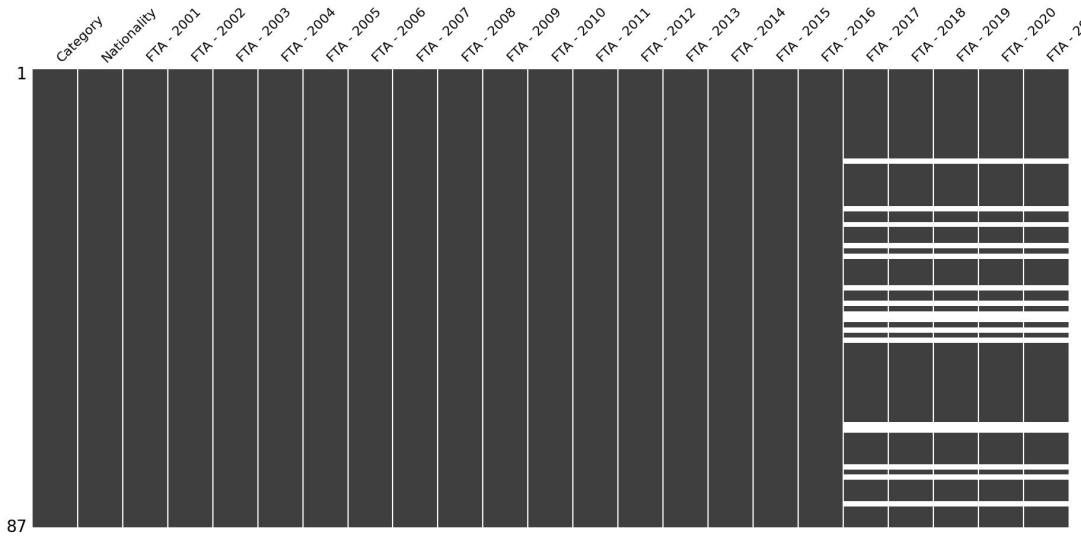
Data Cleaning

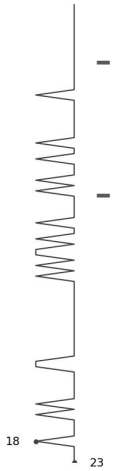
- Dropping irrelevant columns

- Missing Data Analysis - Identify, Analyse, and Handle

- Renaming columns for easier interpretation



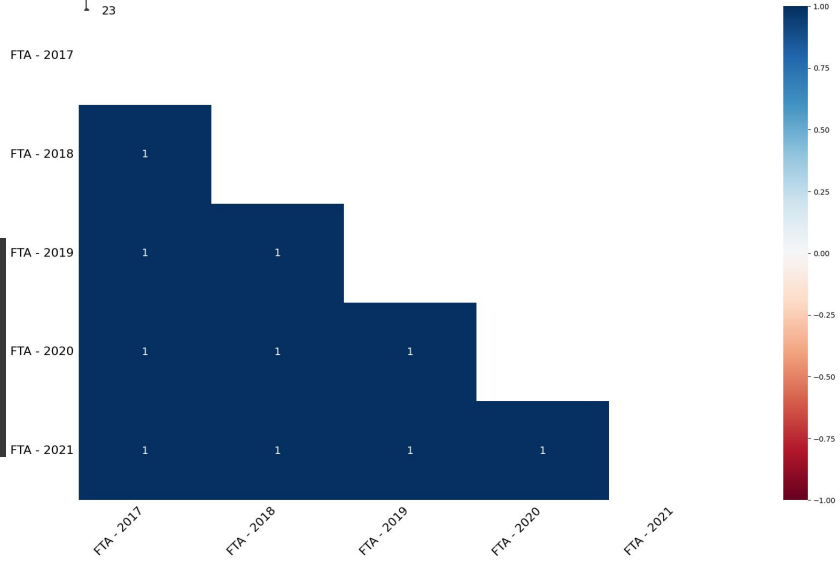




- Relationship between missing data.
- MCAR? MAR? MNAR?

- Little's MCAR Test is used to identify MCAR data.

```
Chi-Square Test Result:
P-Value: 1.0
The missing data is likely MCAR.
```



Missing Values

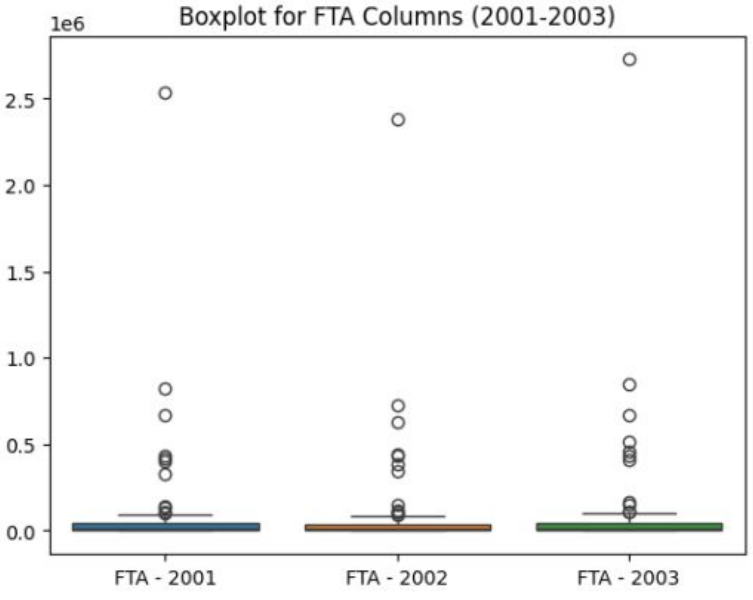
- Our dataset shows consistency across all rows. Each value follows a proportion with respect to its value in the previous row. Each value follows a proportion of sorts.
- Can use a variation of mean imputation. Simple to use and does not affect the accuracy of our data.

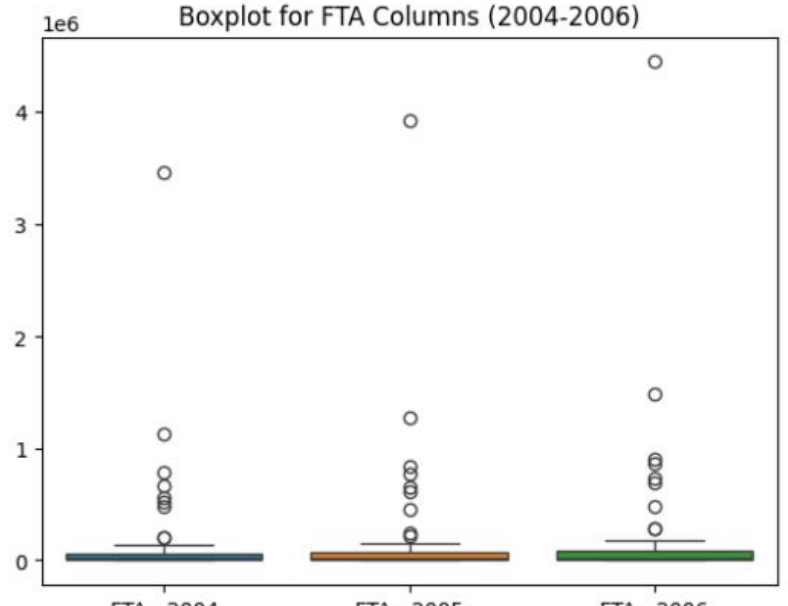
Data Transformation

- Using the data available in our second dataset and the above method, we proportionally create two more rows (2022 and 2023) to match the completeness of our data.

Outlier Analysis

- Observations significantly far from the distribution. Maybe genuine or incorrect.
- Identified using IQR, visualised using boxplots. Some of the box plots from both our datasets are shown below.





Outlier Analysis

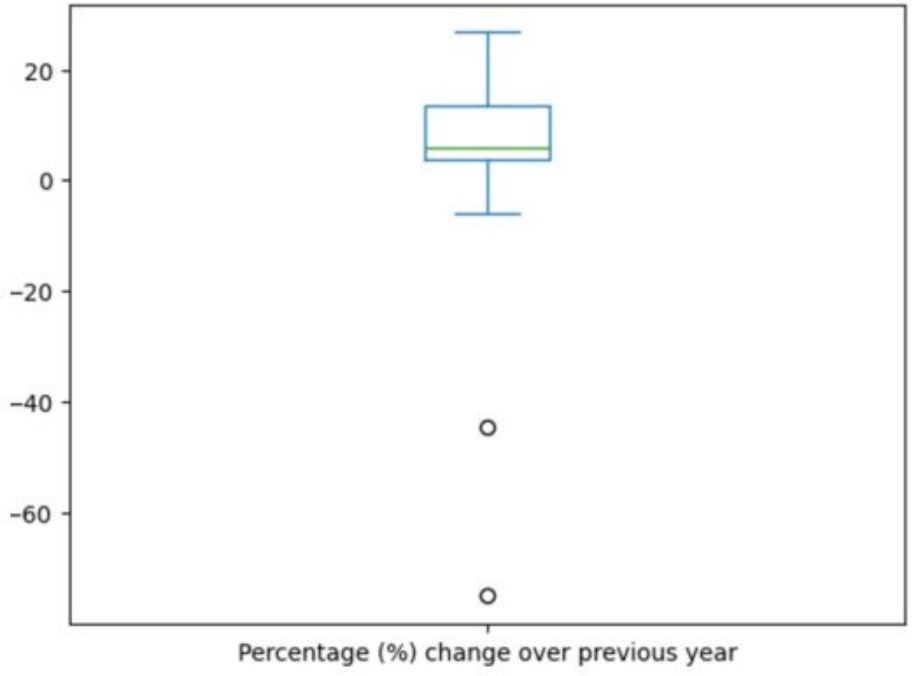
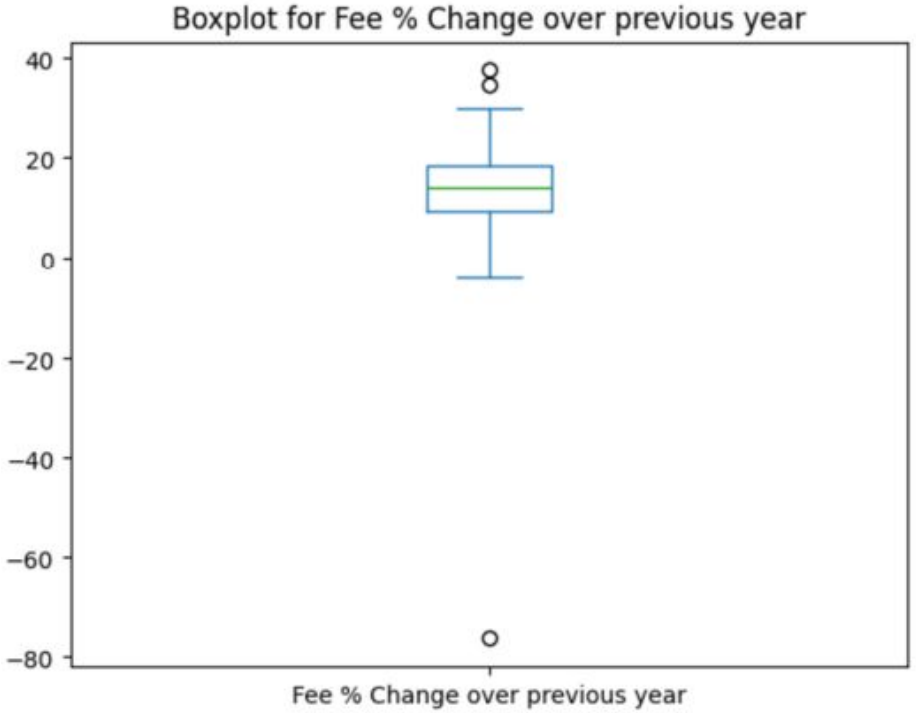


Figure 4.9: Outliers in FTAs



Insights using Data Visualization

- We have used two types of Data Visualization:
  1. Univariate Analysis
  2. Bivariate Analysis
- Univariate Analysis used to summarize and understand individual variables
- Histograms, pie charts, box plots are used to explore data distribution, detect outliers, and assess data quality.
- Useful for understanding key metrics like FTA in India per quarter over years
- Bivariate Analysis used to examine relationships between two variables, identifying trends, dependencies or correlations.
- Scatter plots, correlation matrices, line charts are employed to evaluate how two features impact each other.
- Used to understand the relationship between FTA/FEE and Tourism Share /Rank of India

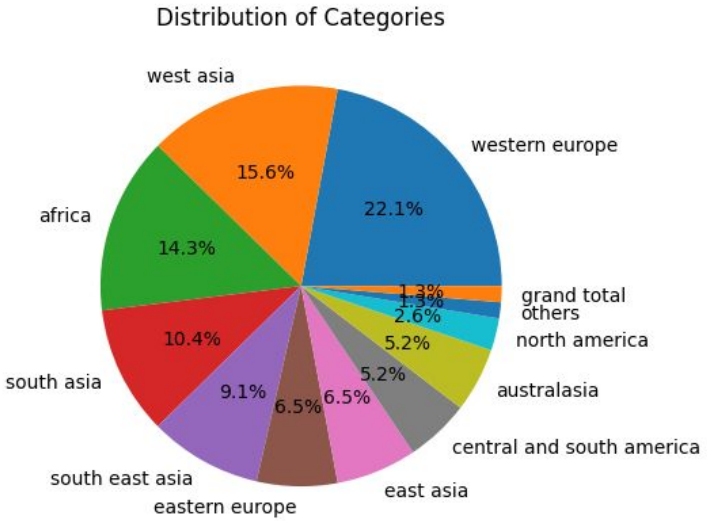
Univariate & Bivariate Analysis

We have used:

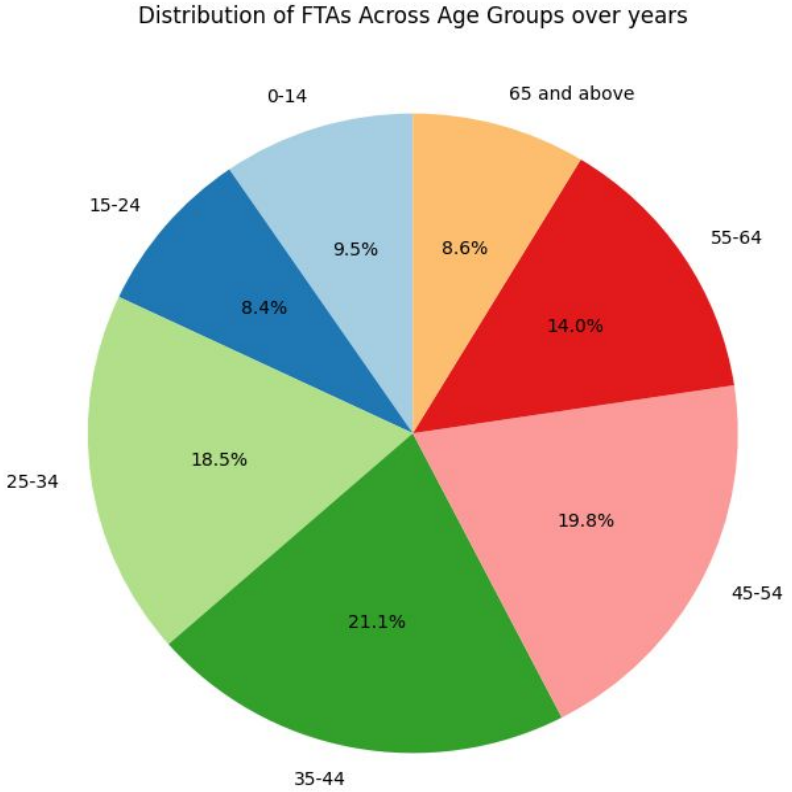
1. Box plots - Outlier detection for FTAs, FEEs and percentage change of these variables over previous year
2. Histograms - Quarterly (of a year) distribution of FTAs over years
3. Pie plots - Distribution of regions of origins for FTAs, age group wise and quarterly total FTAs in India over years
4. Line graph - Trend of FTAs over categories

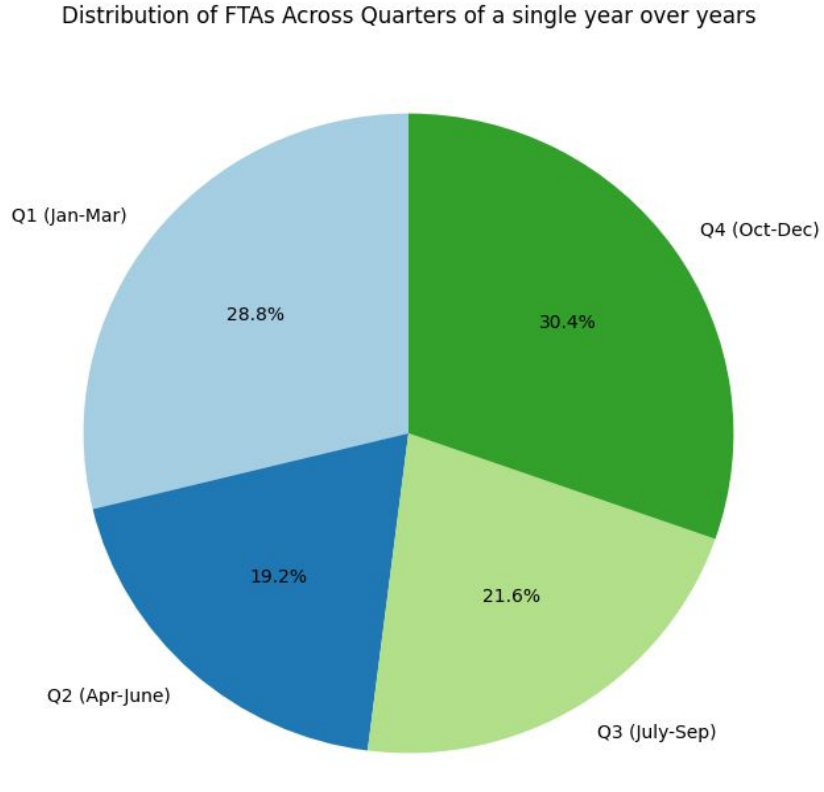
We have used:

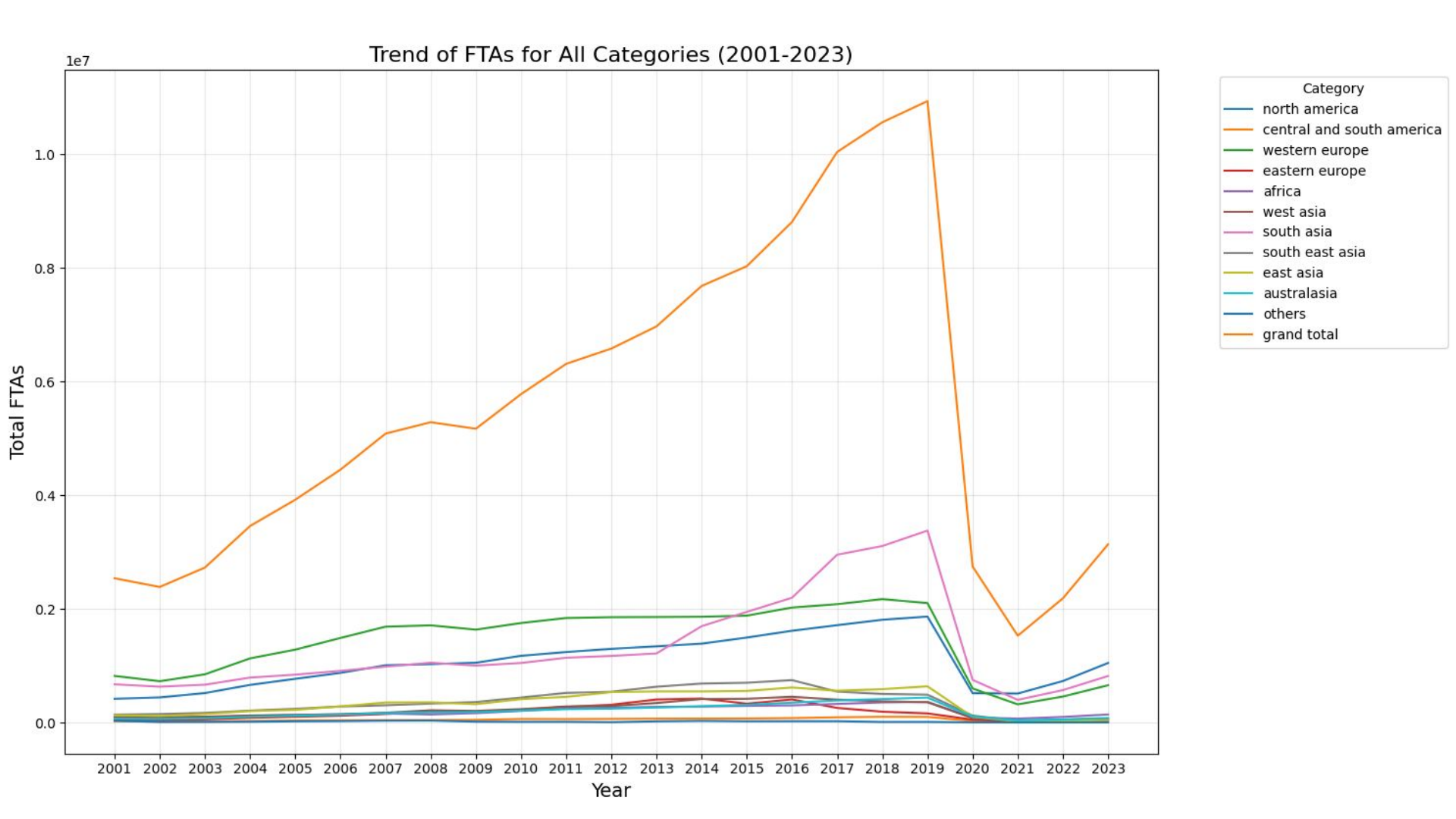
1. Heatmap - Correlation heatmap for FTA data
2. Scatterplots - shows the relation between FTA/FEE and Tourism Share/ Rank of India
3. World Map Choropleth showing region-based density of FTAs into India



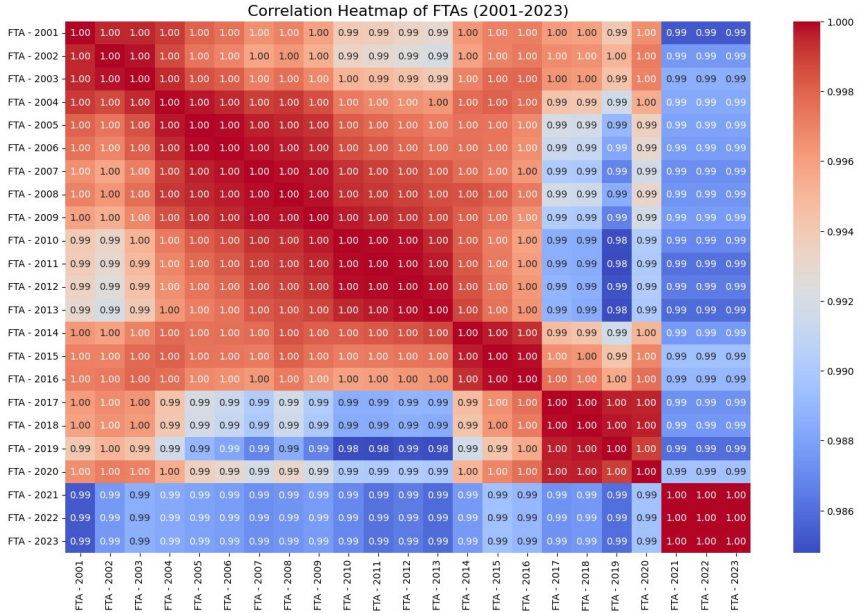
Demographic and Seasonal distribution of FTAs



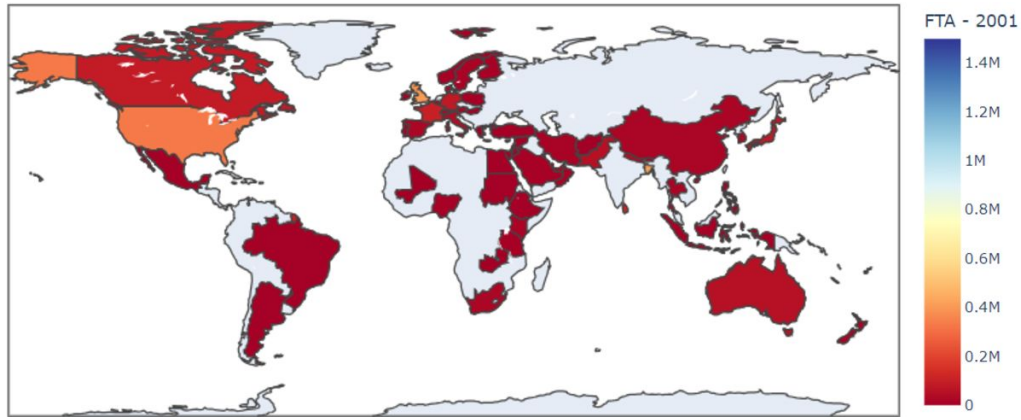



Trend of FTAs for All Categories (2001-2023)
1e7
Total FTAs
Year
Category
north america
central and south america
western europe
eastern europe
africa
west asia
south asia
south east asia
east asia
australasia
others
grand total

Heatmap - Correlation heatmap for FTA data

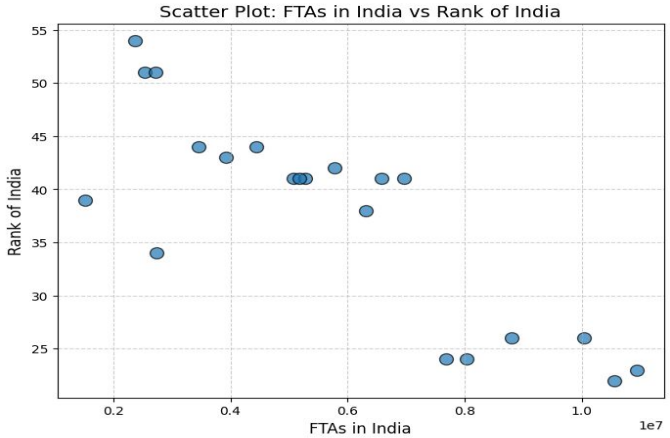


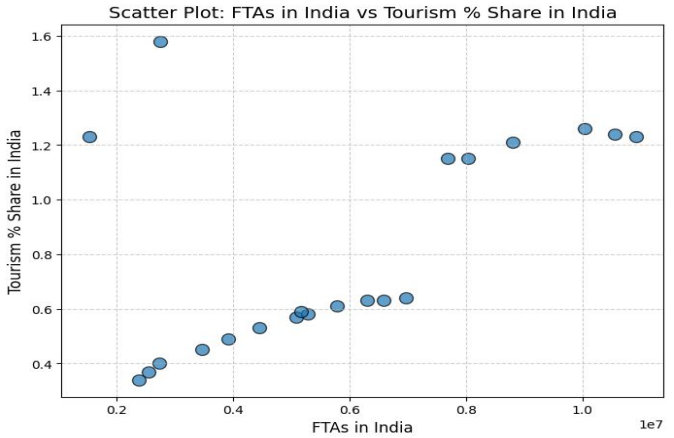
World map Choropleth showing region-based density

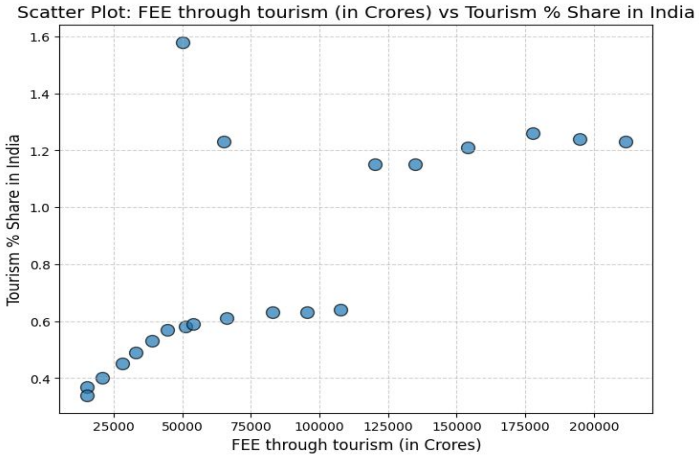


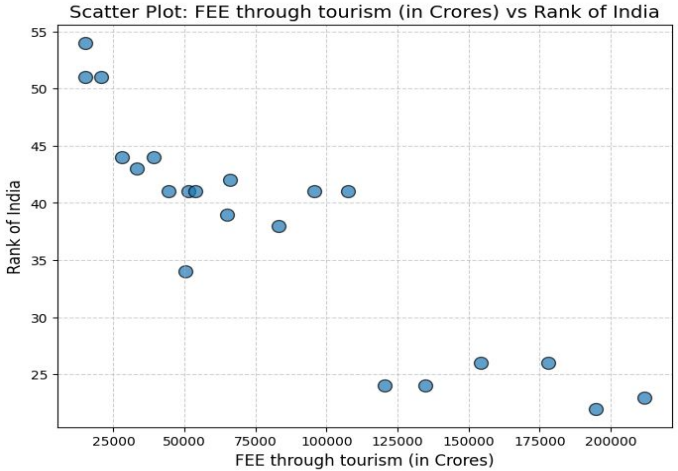
*Note that the data of many countries is unavailable. This is because this data was not recorded and unavailable in our dataset

Scatterplots - Gives an idea about possible correlation among features









Feature Engineering

- Given the year-wise FTA values, it is clear that the dataset is a time-series data.
- To prepare the data for time-series analysis and prediction, the dataframe was reshaped by converting year values into rows.
- We have created some additional input features (data transformation) that will enhance the model predictions.
- We will use the ‘Year’ feature along with additional features as selected features for our model.

Feature Engineering

Year Squared: Adds a quadratic term for the Year column. Helps capture non-linear trends in time-series data.

Lag feature: Captures the influence of past values on the current value, which is common in time-series data (autocorrelation).

Rolling Mean(3-Year): Smoothens the data to reduce short-term fluctuations and highlights longer-term trends.

Implementation: These features are calculated using groupby to ensure they are computed separately for each category/nationality.

Model Fitting

Three regression models are implemented to predict FTA values:

1. Linear Regression:
2. Random Forest Regressor:
3. Gradient Boosting Regressor:

Model Training and Evaluation:

- Train Test Split
- Feature Scaling
- Model Fitting
- Performance metrics - MSE, MAE, R2 Score

Performance Metrics Comparisons for Models

For each Model we measure MAE, MSE and R2 score. Below is the reason and values of R2 score for all the models we used for predictions.

Time-Series Relevance: Since your data is time-series, R^2 can highlight how well temporal and trend-based features (e.g., Year and Rolling Mean) explain FTA variations.

Models: R2 score

LinearRegression: 'R2': **0.916**

RandomForestRegressor: 'R2': **0.931**

GradientBoostingRegressor: 'R2': **0.937**

Conclusion and Results

- Age Group Distribution: Middle-aged adults (25-54 years) emerged as the predominant contributors to FTAs, reflecting their higher travel frequency and disposable incomes.
- Quarterly Trends: Quarterly analysis underscored seasonal peaks in FTAs, particularly winter months, indicating the impact of climate on tourist inflows.
- Regional Analysis: Western Europe and West Asia dominated as primary sources of foreign tourists.
- Impact of COVID-19: The pandemic significantly disrupted the tourism sector, with sharp declines in FTAs and FEE during 2020-2021. This anomaly underscored the sector’s vulnerability to global crises.


Thank You!