## **Predicting the Most Visited Countries**

Name: Ghaith Salameen

Supervised by: Maysa Khalil

# 1. Introduction

## Background

The rise of data science and machine learning (ML) has enabled deeper insights into global trends, including international tourism. Predicting tourist arrivals using historical and economic indicators allows stakeholders to make informed decisions. This project uses ML techniques to analyze and predict the most visited countries in 2024.

## Problem Statement

The tourism industry lacks predictive tools that estimate future tourist arrivals with high accuracy. This project addresses the challenge of forecasting international tourist arrivals to identify which countries are expected to receive the most visitors in 2024.

## Objectives

- Analyze historical data of tourist arrivals.  
- Handle missing data using intelligent imputation techniques.  
- Compare arrival trends across years.  
- Build predictive models to forecast 2024 tourist arrivals.

## Scope

This project focuses on using publicly available tourism data from 2022 and 2023, supplemented with World Bank statistics, to predict 2024 arrival figures. Deep learning is not utilized; instead, the focus is on classical ML regression models.

# 2. Literature Review

## Previous Work

Previous works in tourism forecasting have used time series analysis and statistical methods. These studies often lacked advanced imputation techniques for missing data and did not integrate machine learning models for prediction.

## Comparison

Many past methods relied solely on complete datasets and basic regression, ignoring the potential of models like Random Forests or SVR and data enrichment strategies.

## Contribution

This project fills data gaps using multi-stage imputation (year-over-year, World Bank data, KNN), then builds and compares ML models, offering a more robust prediction framework.

# 3. Methodology

## Data Collection

Data sourced from a CSV file titled 'most-visited-countries-2024.csv', includes 203 countries with data on arrivals in 2022, 2023, predictive values for 2024, and World Bank stats.

## Data Preprocessing

- Initial analysis showed 70–90% missing data in some columns.  
- Missing data filled in three stages:  
 1. Using previous year values.  
 2. Filling from World Bank data.  
 3. Applying KNN Imputer.

## Model Selection

Linear Regression, Random Forest Regressor, and Support Vector Regressor (SVR) were used for modeling.

## Justification

These models are well-suited for regression tasks with numerical features and offer different strengths (interpretability, robustness, flexibility).

## Training Strategy

- Split data into train/test sets.  
- Evaluation using Mean Squared Error (MSE) and R² score.

## Validation Strategy

Used train-test split and manual inspection of trends across years.

# 4. Experiments

## Experimental Setup

- Environment: Google Colab.  
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn.

## Hyperparameter Tuning

- Models used default parameters for baseline.  
- Future iterations could explore GridSearchCV.

## Baseline Models

- Initial predictions with Linear Regression serve as the baseline.  
- Random Forest and SVR models are used for performance comparison.

# 5. Results and Discussion

## Performance Metrics

(Metrics not listed in the PDF; should be added if available, e.g., MSE, R²)

## Visualizations

- Histograms and box plots to understand distribution and outliers.  
- Line plots comparing 2022 and 2023 arrivals per country.  
- Pairplot to examine feature relationships.

## Comparative Analysis

Initial plots show visible trends in increased arrivals from 2022 to 2023, justifying the forecasting strategy.

## Error Analysis

Model errors may arise due to high initial missingness and generalization limits of the selected models.

# 6. Conclusion and Future Work

## Summary of Findings

- Effective imputation strategies greatly improved data usability.  
- Predictive models were successfully trained to estimate 2024 arrivals.  
- Some countries consistently ranked high in tourist numbers.

## Limitations

- Heavy reliance on imputation may reduce real-world accuracy.  
- Limited feature set (only arrival counts, no socio-economic indicators).

## Future Directions

- Incorporate additional features (GDP, travel restrictions, events).  
- Apply deep learning models for time series forecasting.  
- Use cross-validation and automated hyperparameter tuning.

# 7. References

- World Bank Open Data  
- Scikit-learn documentation  
- Matplotlib, Seaborn documentation  
- Original CSV dataset used in Colab

# 8. Appendices

- Figures of boxplots, histograms, and time-series charts.  
- Code for imputation and model training.