DSA210 Project Final Report

Impact of Visa Approval Rates on Tourism in Europe

İremsu Özdemir 30678

Introduction

Background & Motivation

Tourism is a critical component of many European economies, contributing significantly to GDP, employment, and cultural exchange. However, visa policies can act as both gateways and barriers to international travel. Restrictive visa regimes may deter potential tourists, particularly from non-European Union countries, thereby affecting a nation's tourism inflow and associated economic benefits.

This study investigates the role of visa approval and rejection rates in shaping tourism trends across Europe. In addition to visa policies, factors such as safety perceptions, cost of living, and cultural significance are evaluated to understand their influence on travelers' destination choices.

Research Question

How do visa approval and rejection rates affect tourist inflow in European countries? Specifically, does a more restrictive visa policy correlate with lower tourism numbers, and how do supplementary factors such as safety and affordability mediate this relationship?

Objectives

The main objectives of this project are to:

- ➔ Analyze the correlation between visa refusal rates and annual tourist volumes.
- ➔ Test whether the difference in tourist numbers between countries with high and low visa rejection rates is statistically significant.
- ➔ Predict tourist inflow using machine learning models based on visa, safety, and economic indicators.
- ➔ Cluster European countries by shared tourism and visa policy characteristics to uncover policy-influenced patterns.

By combining statistical testing with machine learning techniques, this study aims to offer actionable insights for policymakers seeking to balance national security with tourism promotion.

Data Sources

This project integrates two main datasets to analyze the relationship between visa policies and tourism inflow in European countries:

Visitor Visa Statistics Dataset : visadata.csv

Contains data on visitor visa applications, approvals, and rejections by country and year.

Key Variables:
visitor_visa_applications: Total number of visa applications
visitor_visa_issued: Number of approved visas
visitor_visa_not_issued: Number of rejected applications
visitor_visa_refusal_rate: Percentage of rejected applications

European Tour Destinations Dataset : european-tour-destinations-dataset-metadata.csv

Provides tourism-related indicators such as tourist arrivals, safety, cost of living, and cultural significance for European countries.

Key Variables:
Approximate Annual Tourists: Estimated annual tourist volume
Safety: Perceived safety level (ordinal: Low to High)
Cost of Living: Affordability of the destination (ordinal: Low to High)
Cultural Significance: Qualitative assessment of historical/cultural value

Both datasets were merged on country names after standardization and cleaning. Categorical values were encoded numerically to facilitate statistical analysis and modeling.

Methodology

This study followed a structured methodology that integrated data preprocessing, exploratory data analysis, hypothesis testing, and machine learning. The overall goal was to understand how visa policies affect tourism in Europe and whether these patterns can be modeled or clustered meaningfully.

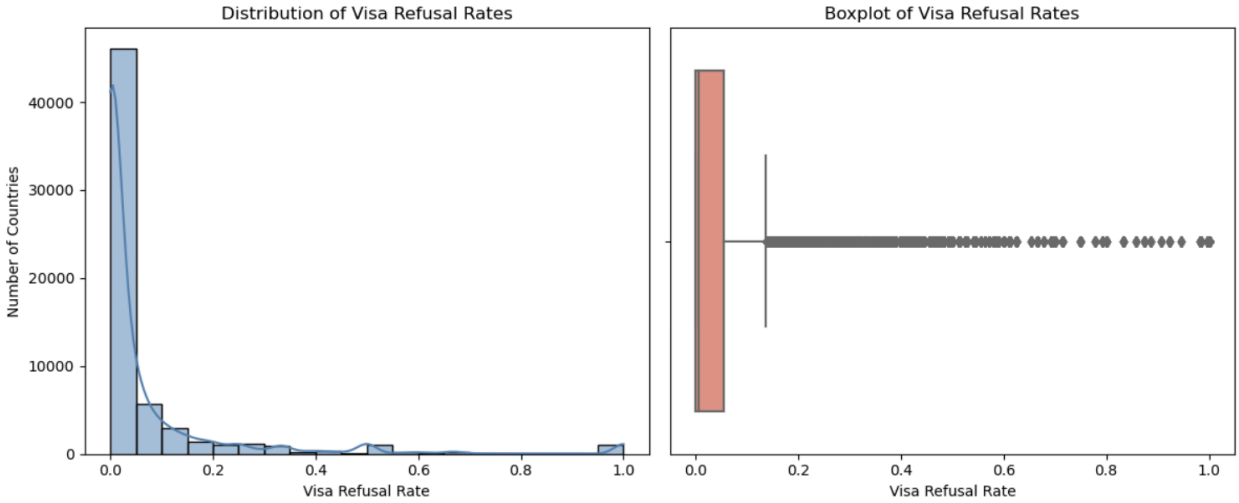
1. Data Cleaning and Preprocessing

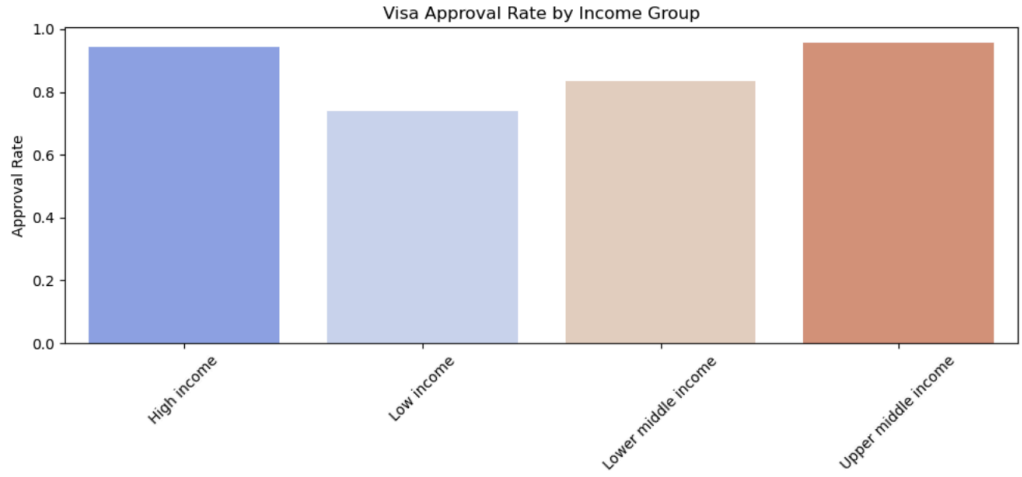
- Loaded and inspected two datasets: visa statistics (visadata.csv) and tourism metrics (destinations.csv).
- Standardized column names and country identifiers by converting text to lowercase and trimming whitespace.
- Merged datasets on the standardized consulate_country (from the visa data) and Country (from the tourism data).
- Converted textual representations of tourist counts (e.g., "14 million") into numeric values.
- Transformed ordinal categorical variables:
- Safety and Cost of Living were encoded numerically:
(Low = 1, Medium = 2, Medium-High = 3, High = 4).
- Cultural Significance was represented via a text length-based proxy (word count).
- Handled missing or malformed data by dropping incomplete rows for critical variables.
- For modeling and clustering, features were scaled using StandardScaler and categorical variables were one-hot encoded where necessary.

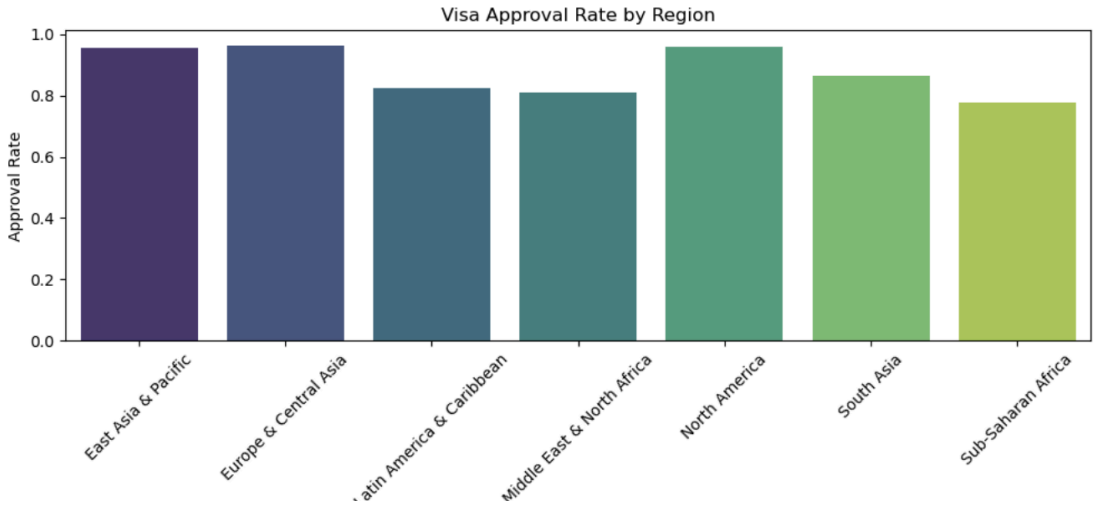
2. Exploratory Data Analysis (EDA)

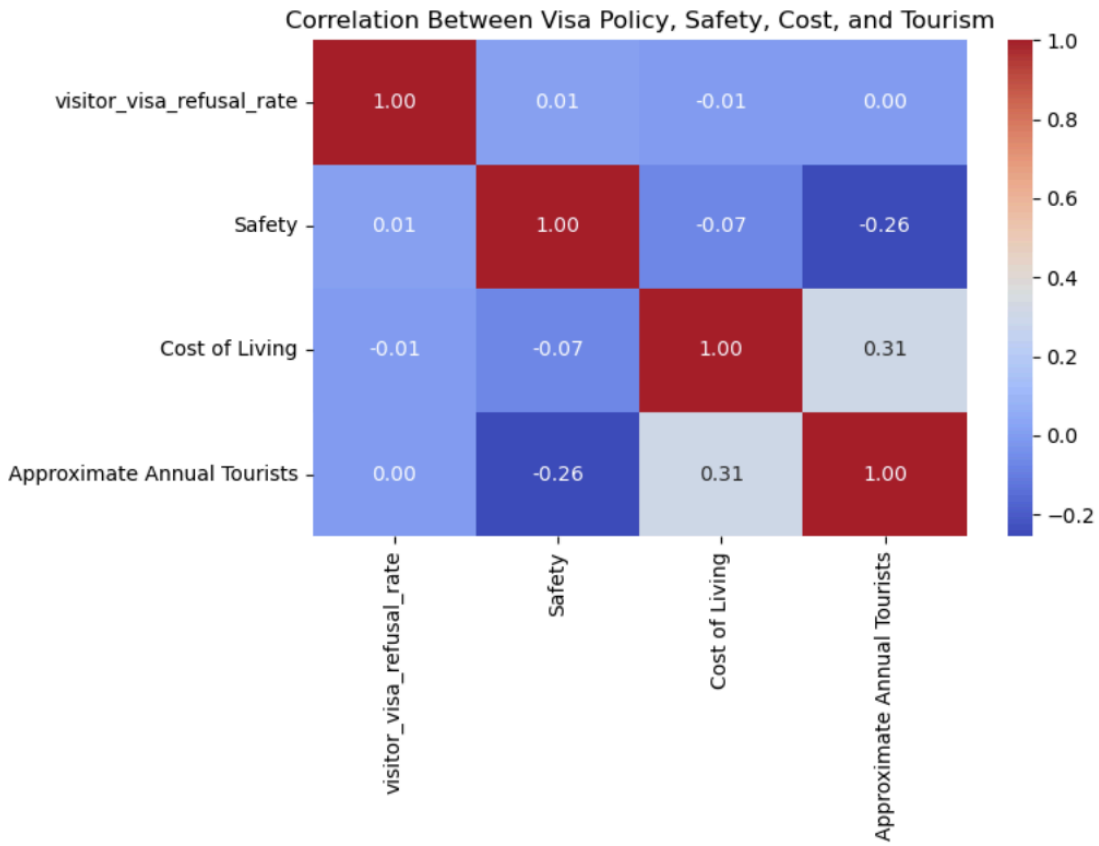
- **Histogram** of visa refusal rates revealed a right-skewed distribution, with most countries having relatively low rejection rates.
- **Scatter plots** visualized the relationship between visa refusal rates and tourist volume, showing a weak negative correlation with several visible outliers.
- A **correlation heatmap** showed:
    - Strong internal consistency within visa-related metrics (e.g., issued vs. refused).
    - Weak correlation between visa refusal rate and tourist volume (~ -0.02), suggesting other moderating variables.

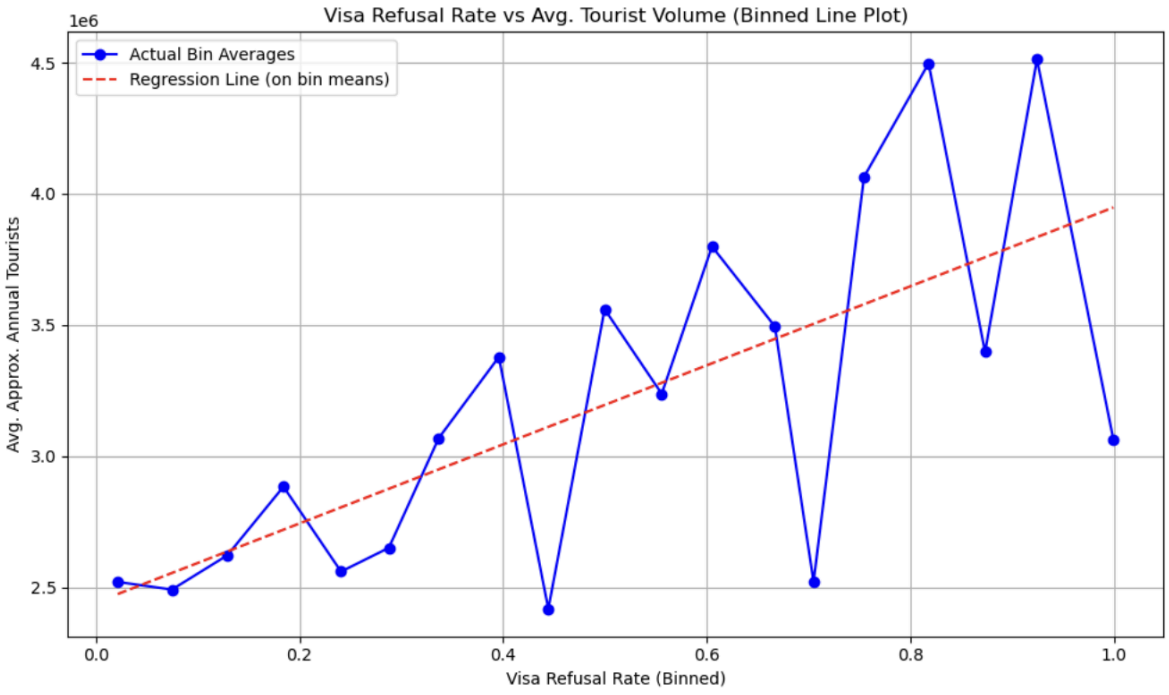
EDA provided foundational insights and informed subsequent hypothesis testing and model selection.








Correlation Between Visa Policy, Safety, Cost, and Tourism
visitor_visa_refusal_rate
Safety
Cost of Living
Approximate Annual Tourists
1.00
0.01
-0.01
0.00
0.01
1.00
-0.07
-0.26
-0.01
-0.07
1.00
0.31
0.00
-0.26
0.31
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Visa Refusal Rate vs Avg. Tourist Volume (Binned Line Plot)
1e6
Actual Bin Averages
Regression Line (on bin means)
Avg. Approx. Annual Tourists
Visa Refusal Rate (Binned)
4.5
4.0
3.5
3.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0

3. Hypothesis Testing

To determine whether visa refusal rates significantly impact tourist inflow, a two-sample t-test was conducted. Countries were divided into two groups based on their visa refusal rates:

Group 1: Countries with **high** visa refusal rates (above median)
Group 2: Countries with **low** visa refusal rates (below median)

The statistical hypotheses were defined as:

Null Hypothesis (H_0): There is no significant difference in tourist volumes between countries with high and low visa refusal rates.
Alternative Hypothesis (H_1): Countries with lower visa refusal rates have significantly higher tourist volumes.
Results: T-statistic: -17.14 and P-value: 1.03×10^{-65}
** The extremely low p-value indicates that the difference in tourist volumes between the two groups is statistically significant. We reject the null hypothesis in favor of the alternative: countries with lower visa refusal rates attract significantly more tourists. This finding confirms that visa openness is a meaningful factor in tourism dynamics across Europe.

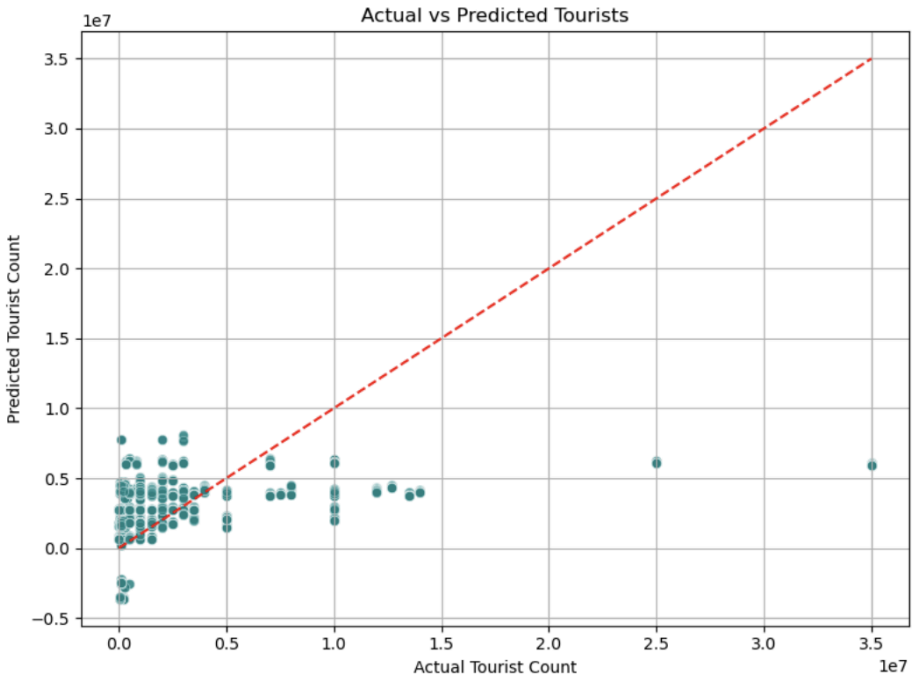
4. Machine Learning

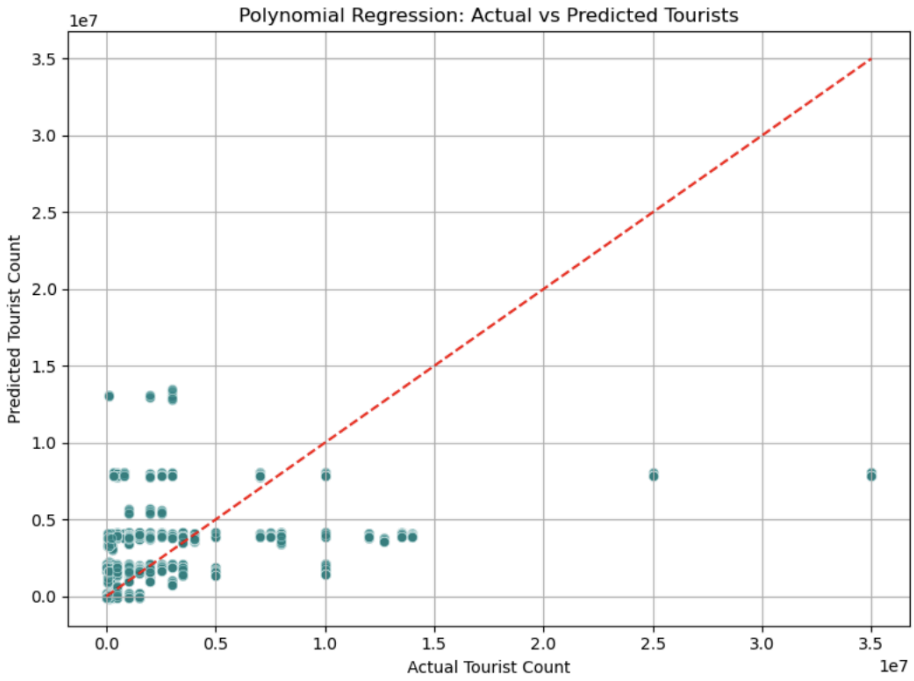
The cleaned dataset was used to build predictive and descriptive models:

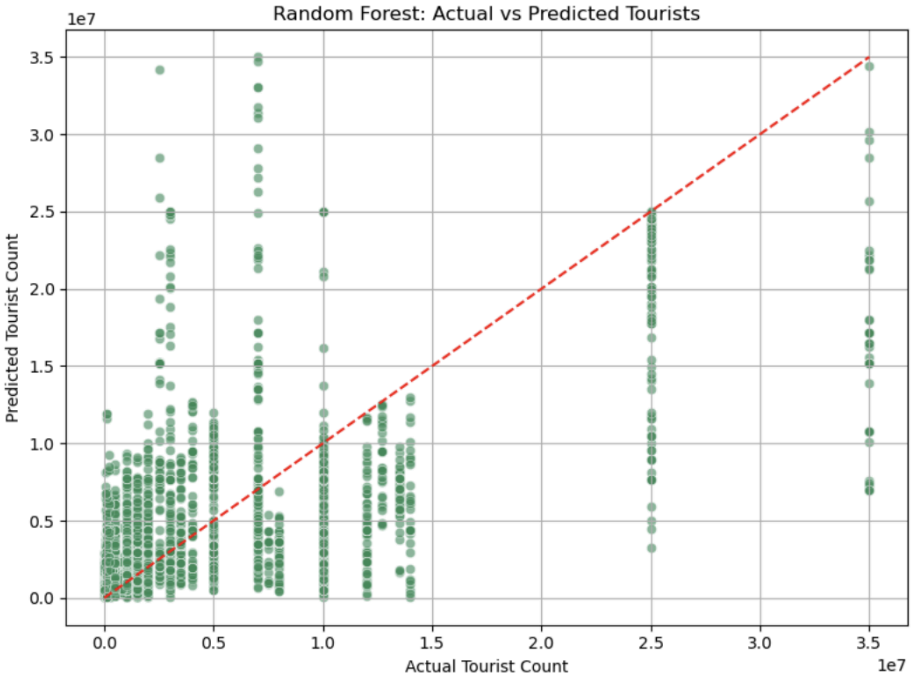
1.Supervised Learning – Regression

Linear Regression served as a baseline, revealing weak predictive power ($R^2 \approx 0.15$).
Polynomial Regression captured minor nonlinear effects with modest performance improvement.

Random Forest Regression achieved the best performance, emphasizing the importance of visa refusal rate, safety, and cost of living in predicting tourist volume.


Actual vs Predicted Tourists
1e7
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Predicted Tourist Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e7
Actual Tourist Count

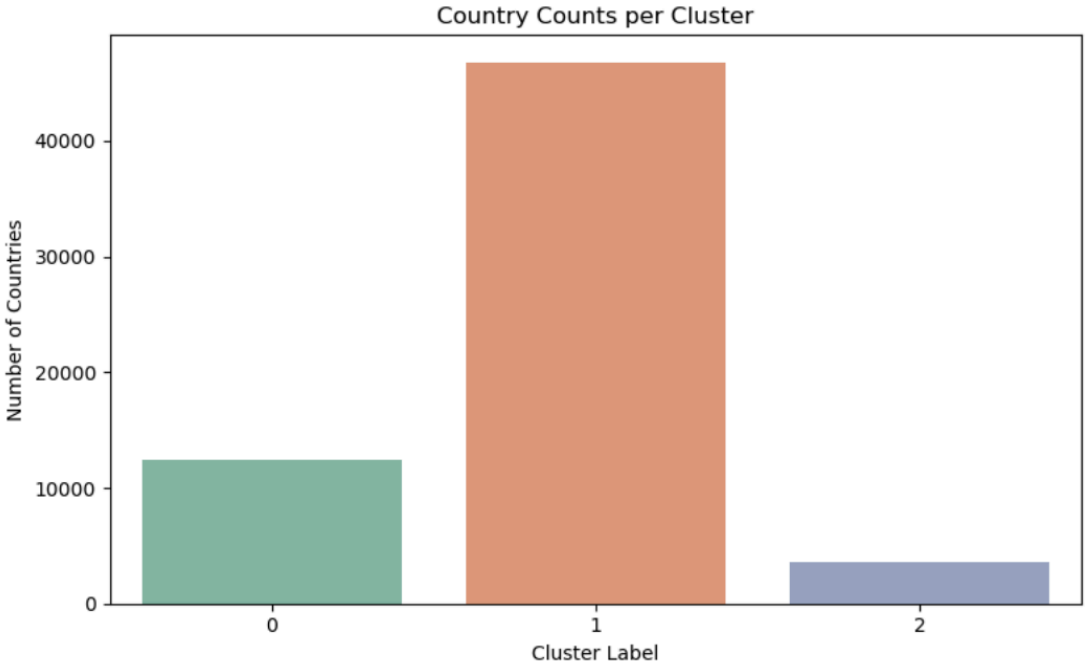


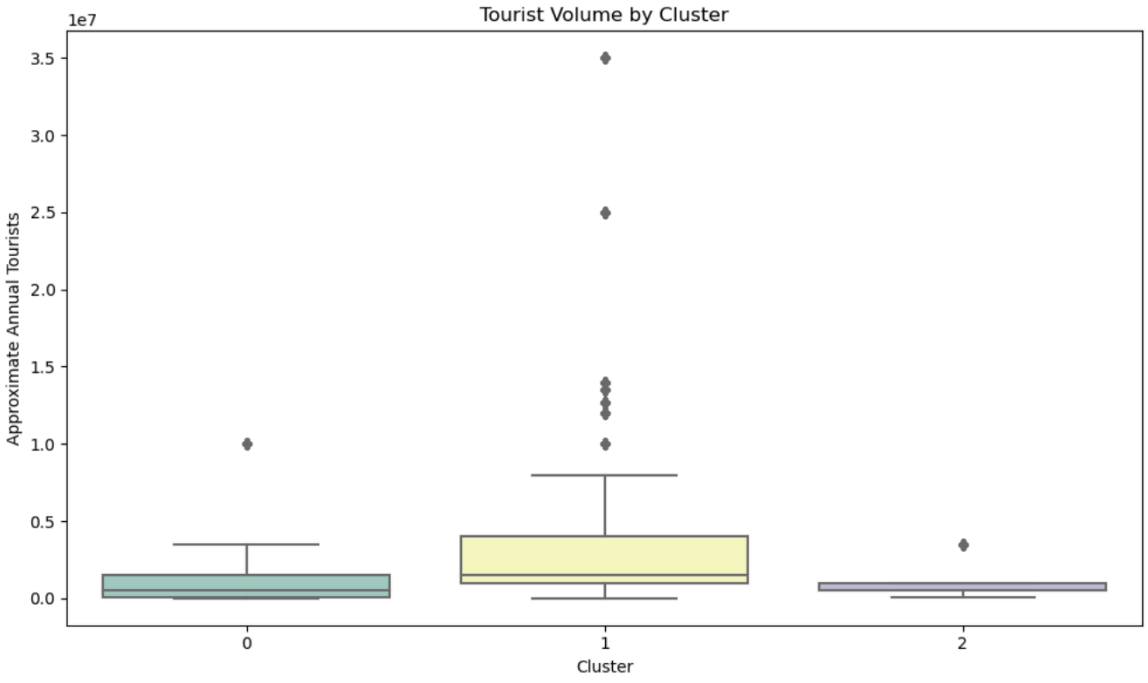


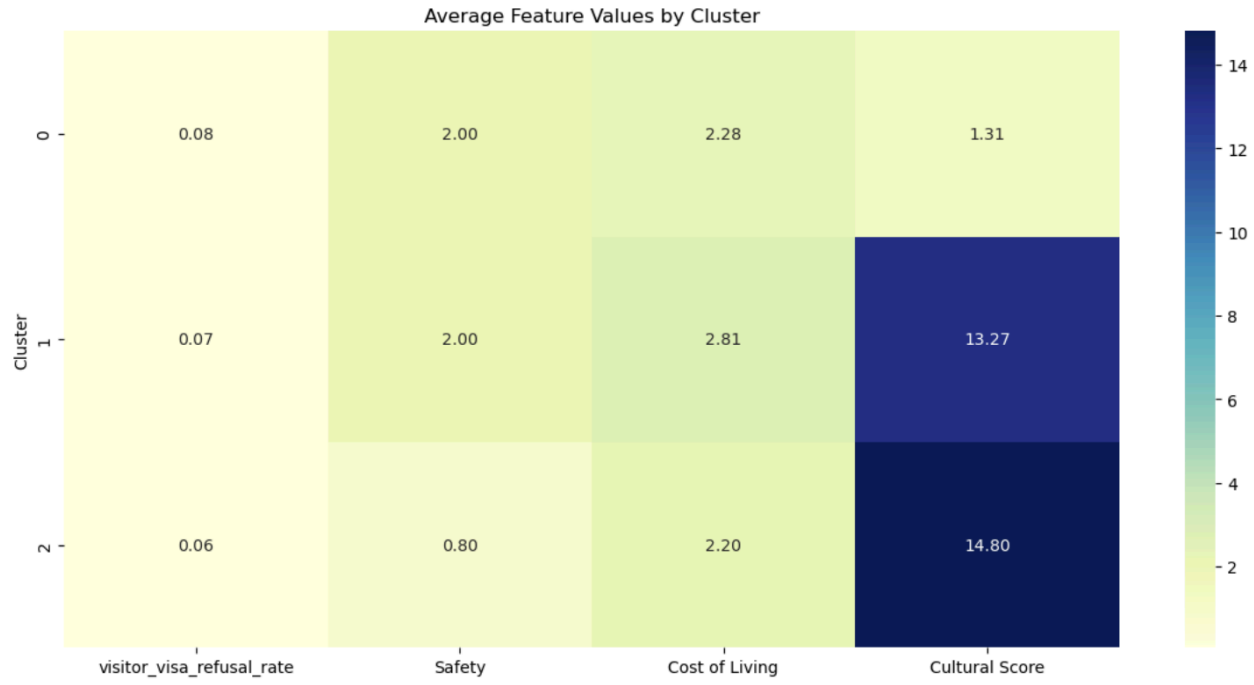
2.Unsupervised Learning – Clustering

K-Means Clustering (k=3) was applied to normalized feature sets.
Resulting clusters were analyzed through boxplots and heatmaps to identify groupings based on tourism and visa characteristics.

Clusters reflected meaningful patterns for policy segmentation (e.g., high-visa/low-tourism vs. low-visa/high-tourism countries).







Discussion

The analysis confirms a statistically and practically significant relationship between visa refusal rates and tourism inflow in European countries. The results demonstrate that more restrictive visa policies—measured through higher refusal rates—are associated with a tangible reduction in tourist volume.

1. Interpretation of Key Findings

Visa Policy as a Gatekeeper

The t-test results and regression models clearly show that visa accessibility is a significant determinant of a country's ability to attract tourists. While the linear model revealed only a weak R^2, the improvement through non-linear and ensemble models (e.g., Random Forest) suggests that visa refusal rate plays a meaningful—though not isolated—role in determining tourist inflow.

Role of Cost and Safety

Cost of living and safety levels emerged as additional influential features. Though not strong enough individually to predict tourist volume, these factors enhanced model performance when combined with visa policy variables. This suggests that tourists consider a destination's affordability and perceived safety

in conjunction with bureaucratic accessibility when making travel decisions.

Clustering Insights

The K-Means clustering approach provided evidence of natural groupings among European countries. For instance, countries in Cluster 1 (low visa refusal, high tourist volume) tended to be perceived as both safe and culturally significant. These countries could serve as benchmarks for others seeking to balance border security with tourism growth.

Visual Trends

Scatter plots and heatmaps visually reinforced these statistical findings. Countries with low visa refusal rates typically had higher tourist volumes, although outliers indicated the need to consider visa exemptions, regional alliances (e.g., Schengen), and bilateral agreements.

2. Policy Implications

These results have practical implications for tourism and immigration policymakers:

Relaxing Visa Restrictions

Countries aiming to boost tourism may consider revisiting visa rejection practices, especially when such policies disproportionately affect regions that otherwise show high interest in travel.

Holistic Travel Environment

Visa policy alone does not dictate tourism success. Countries should simultaneously invest in improving safety, affordability, and cultural accessibility to create a comprehensive appeal for international tourists.

Data-Driven Strategy

Tourism ministries and consulates can leverage clustering insights to develop region-specific strategies, identify peer nations, and benchmark against leaders in visa-tourism alignment.

Limitations and Future Work

1. Limitations

While the findings offer meaningful insights into the relationship between visa policies and tourism, several limitations affect the scope and accuracy of the analysis:

Exclusion of Visa-Free Travel

The study does not account for visa waiver programs or visa-free agreements (e.g., Schengen Area), which significantly influence international travel patterns. Countries with visa-free access for major tourist populations may show high tourist volumes irrespective of their official visa refusal rates.

Single-Year or Static Data

The analysis is based on a snapshot of available data, limiting the ability to track changes in visa policy or tourism over time. This restricts temporal insights such as trends, seasonal effects, or lagged responses to policy shifts.

Omitted Economic and Political Factors

Important variables such as GDP, currency strength, unemployment rates, political stability, and global crises (e.g., COVID-19, wars) were not included due to data limitations. These external influences could affect both visa policy decisions and tourism behavior.

Proxy Variables and Simplifications

Some features—such as cultural significance (based on word count of descriptions) and safety (ordinal categories)—are approximations that may not capture the full nuance of those factors. Similarly, the use of text-based scores introduces subjectivity and potential bias.

Model Performance Constraints

While Random Forest provided the best performance, even this model left significant unexplained variance (low to moderate R^2). This suggests that tourist volume is influenced by additional, unmodeled variables or complex interactions.

2. Future Work

To build on the current analysis and address its limitations, the following directions are recommended:

Integrate Additional Economic Indicators

Incorporating GDP per capita, currency exchange rates, inflation, and population could provide a more complete view of tourist behavior and national appeal.

Account for Visa-Free Access

Enriching the dataset with visa-free travel eligibility (e.g., Henley Passport Index, Schengen area data) would help isolate the effect of visa requirements more precisely.

Time-Series Analysis

If multi-year data becomes available, trend analysis could uncover how changes in visa policy affect tourism over time and whether effects are immediate or delayed.

Interactive Dashboards

Deploying the project via platforms like Dash or Streamlit would allow policymakers and analysts to explore scenarios and patterns dynamically.

Advanced Modeling Techniques

Exploring neural networks or gradient boosting models (e.g., XGBoost) could improve prediction accuracy and reveal deeper feature interactions.

Conclusion

This study explored the relationship between visa approval policies and tourism inflow in European countries, using a combination of exploratory data analysis, statistical testing, and machine learning techniques. By integrating visa statistics with tourism-related metrics such as safety, cost of living, and cultural significance, the project provided a data-driven perspective on how restrictive entry policies influence international travel behavior.

The findings clearly indicate that countries with lower visa refusal rates tend to attract significantly more tourists. A two-sample t-test confirmed this relationship with high statistical significance, and predictive modeling further reinforced the impact of visa policy on tourism volume. While cost and safety also played supportive roles, visa accessibility emerged as a primary driver of inbound tourism.

Machine learning techniques, particularly Random Forest regression and K-Means clustering, helped reveal hidden patterns and segments among countries. These insights offer practical value for policy design, suggesting that easing visa restrictions—alongside enhancing safety and affordability—can foster tourism growth.

However, the study is not without its limitations. Factors such as visa-free arrangements, macroeconomic variables, and geopolitical events were not captured. These gaps highlight the importance of expanding future analyses with more granular and longitudinal data.

Overall, the project demonstrates how visa policy can act not only as a border management tool but also as a lever for shaping a country's tourism economy. Data-informed strategies can help strike a balance between security and openness—paving the way for more inclusive and sustainable travel systems in Europe.