Tanzania Tourism Prediction

Data Analysis and Machine Learning Project

By:

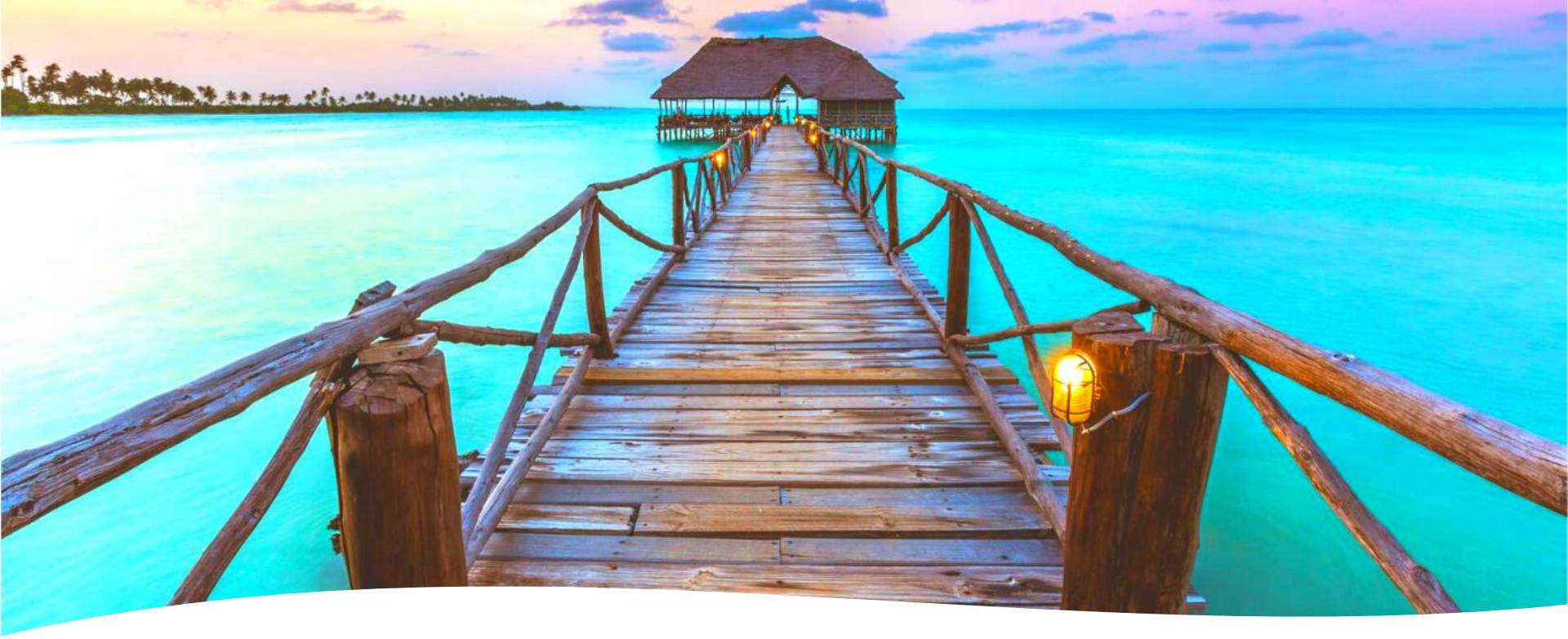
- ***M. RadWan***

What Are We Trying to Achieve?

Goal: Predict tourist expenditure when visiting Tanzania.

Dataset: 6,476 tourists surveyed by National Bureau of Statistics (NBS)

Deliverable: Build a model to forecast & help tourism stakeholders understand spending patterns to optimize offerings & plan better.



Why It Matters

Accurate forecasts empower stakeholders to:

- **Optimize resource** allocation (e.g., staffing, infrastructure).
- **Boost tourism** revenue by targeting peak seasons.
- **Support sustainable** growth in Tanzania's economy.
- Our project **bridges data** and **real-world decisions**.

Project Steps

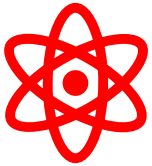
1. Setup and Initialization

2. Load Data

3. Check Data Quality

4. Exploratory Data Analysis

5. Data Preprocessing



6. Feature Engineering

7. Model Selection

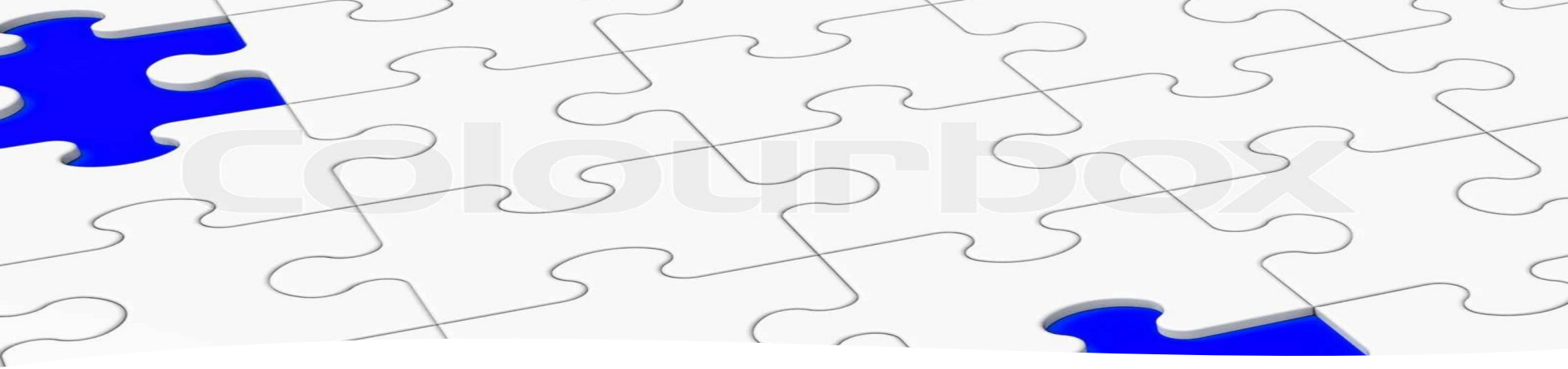
8. Model Training and Evaluation

9. Model Optimization

10. Feature Importance Analysis

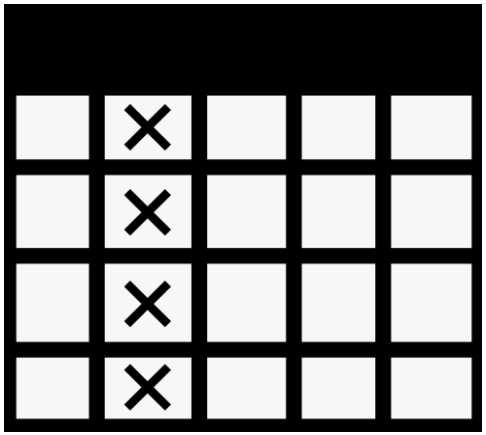
11. Data Product Concept

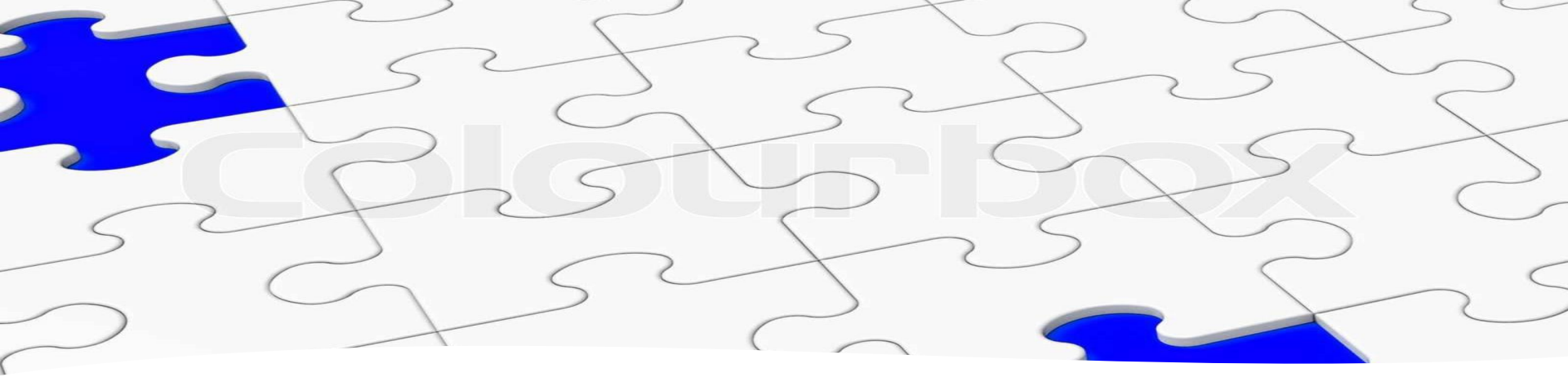
12. Conclusion and Recommendations



The Data Puzzle – What We Started With

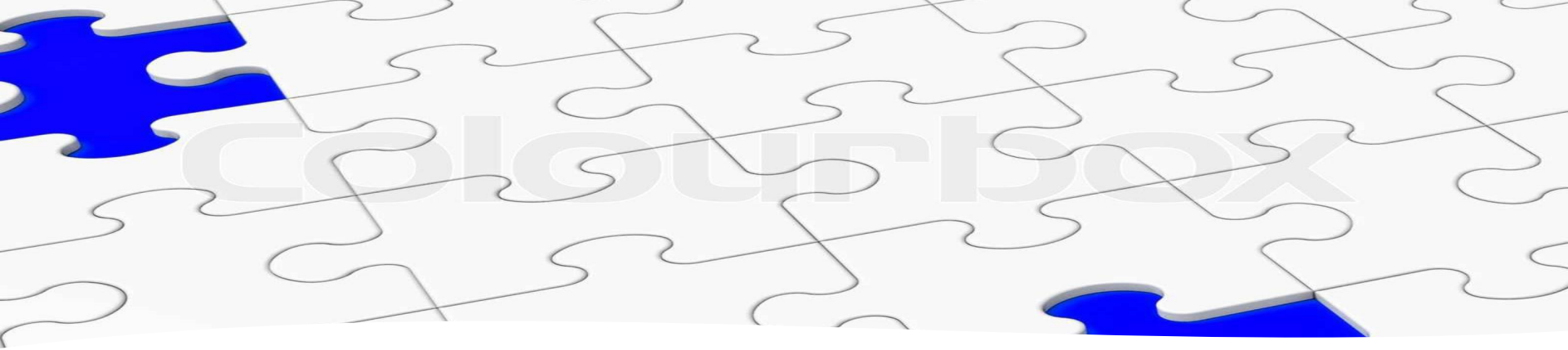
- Our data was like a jigsaw puzzle with missing pieces.
- 4,809 tourist records, 23 details each—like country, travel purpose, and spending.
- But not all pieces fit perfectly from the start!





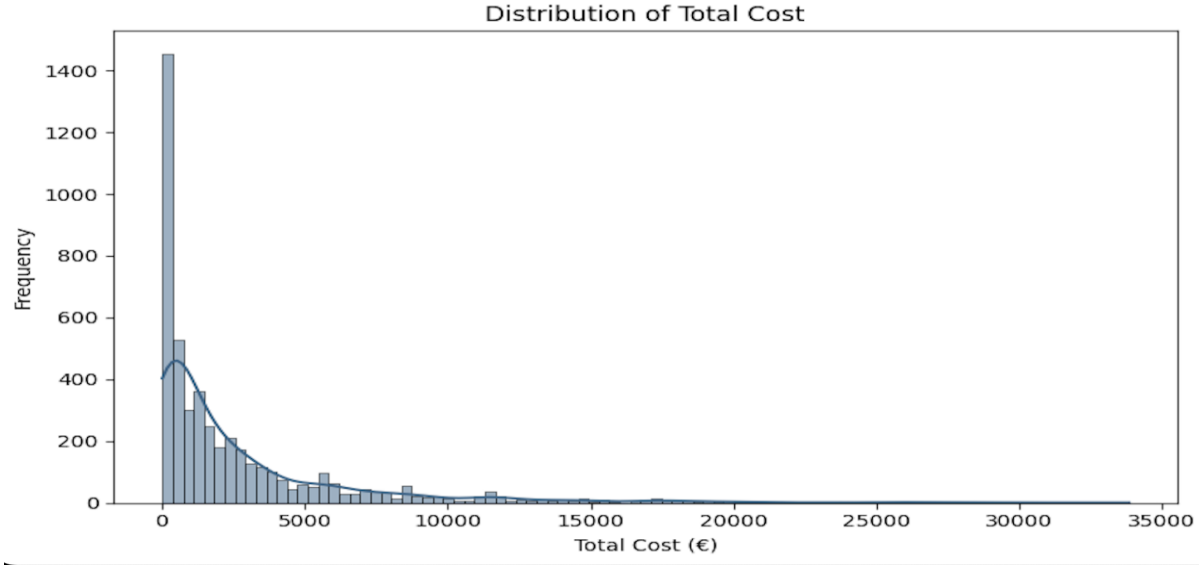
Missing Information – Gaps in the Story

- Over 23% of tourists didn't say who they traveled with.
- 6.5% didn't share what impressed them most.
- Small gaps in male/female counts too (0.1%).
- Challenge: How do we predict spending without the full picture?



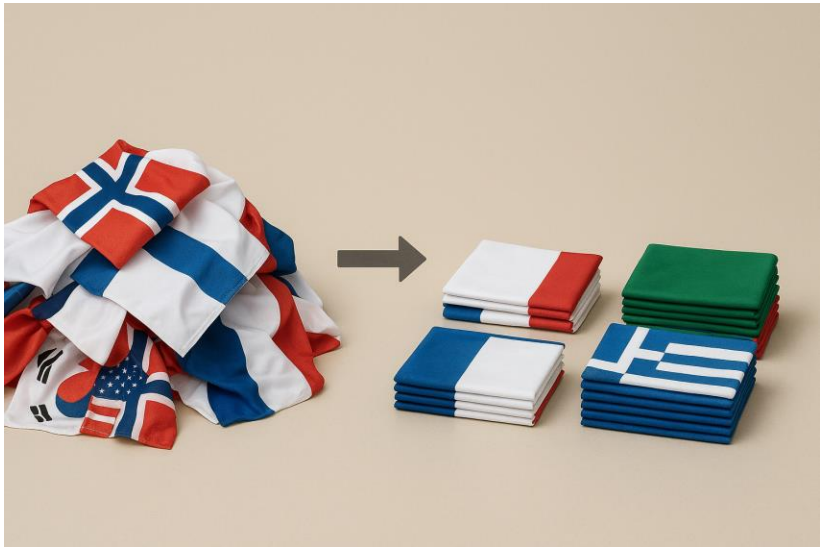
Uneven Data – Not Everyone Spends the Same

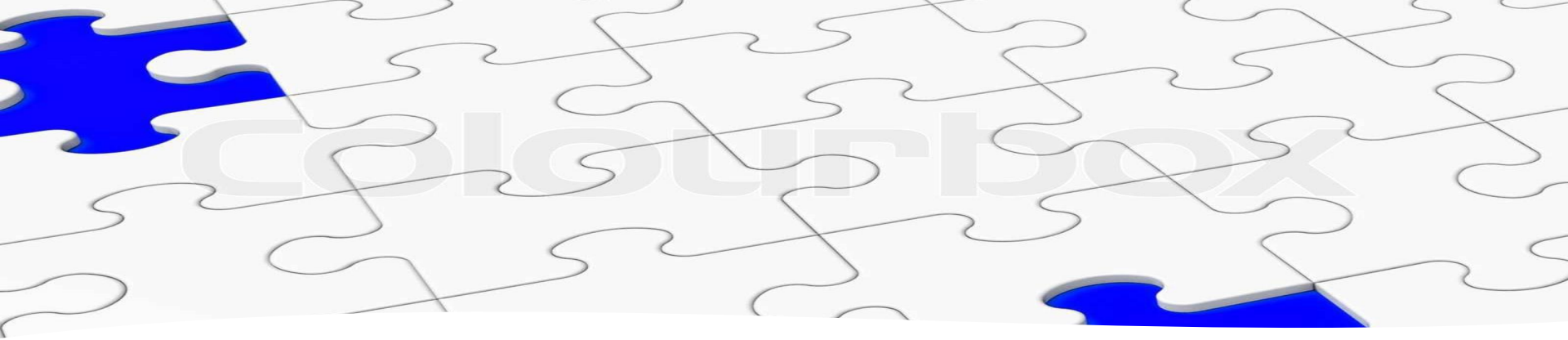
- Spending ranged from 49,000 TZS to 99 million TZS!!!!
- Most tourists spent less, but a few spent a LOT.
- Challenge: Our predictions could focus too much on big spenders.



Too Many Choices – Sorting Through the Noise

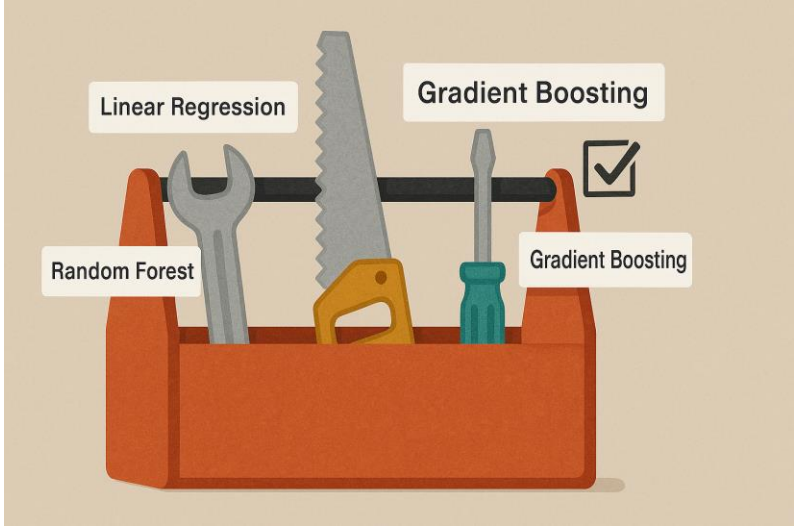
- Data had lots of categories: countries, activities, payment types, and more.
- Example: 100+ countries turned into continents (e.g., Europe, Asia).
- Challenge: Too much variety confused the model.

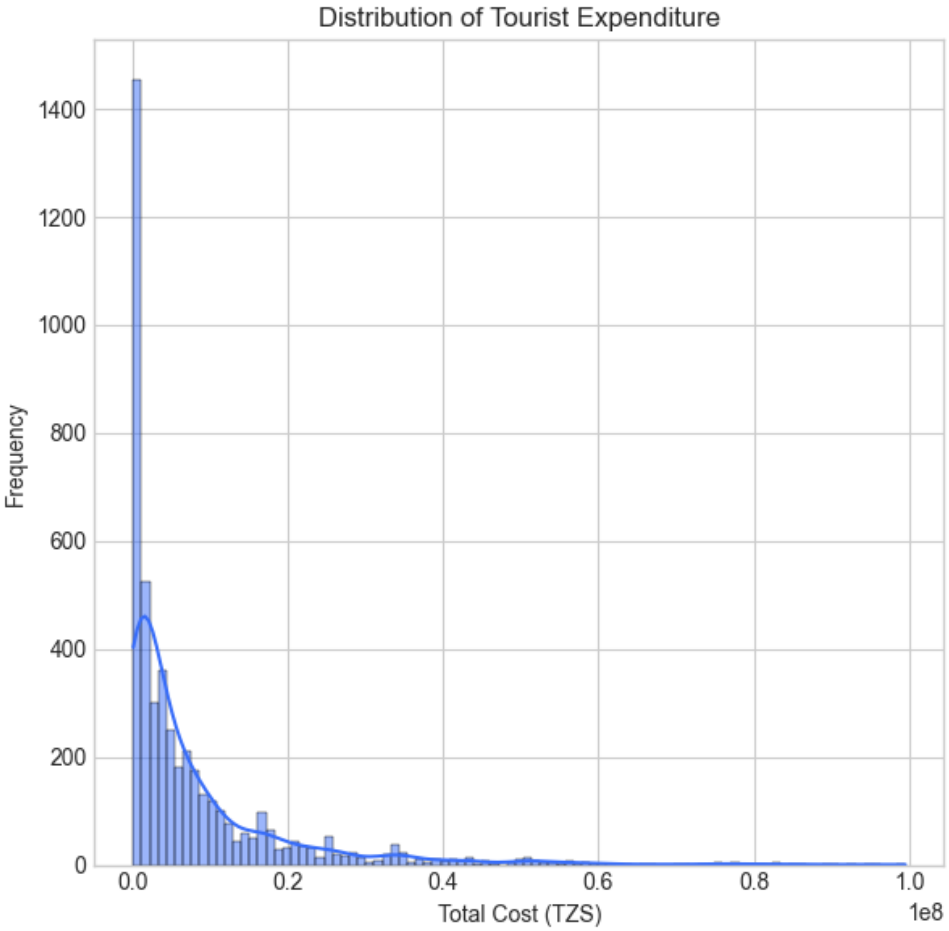


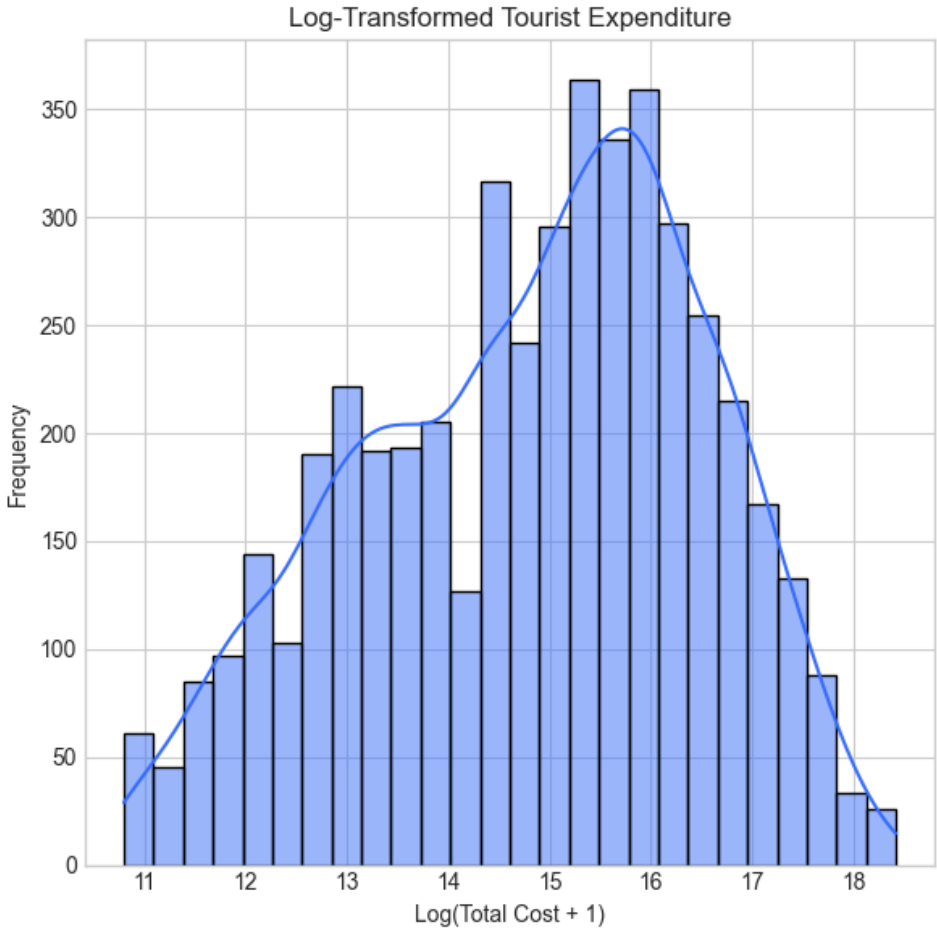


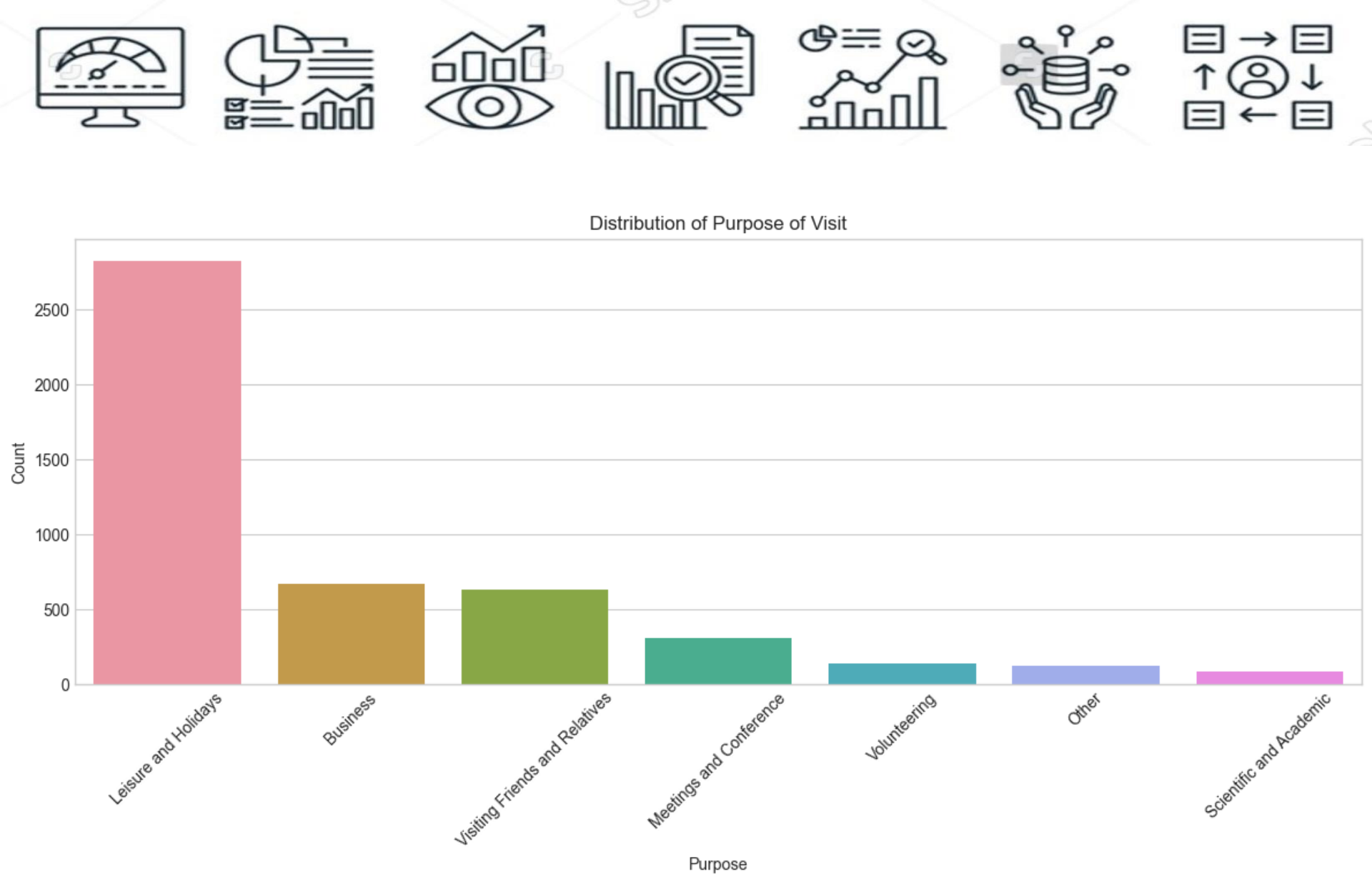
Finding the Right Fit – Testing Our Tools

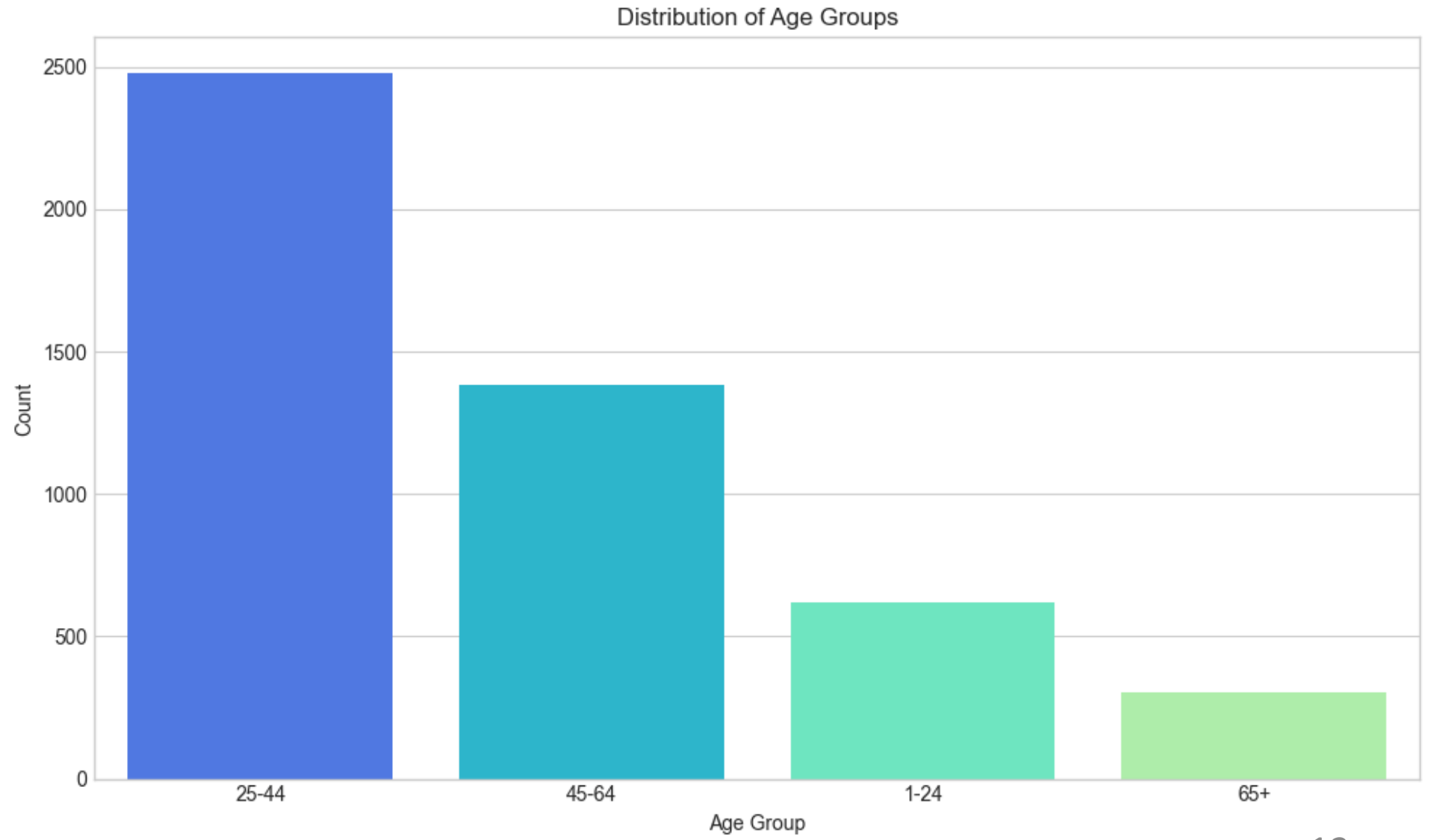
- We tried multiple tools (models) to predict spending.
- Some worked better than others—Gradient Boosting won!
- Challenge: Picking the best tool took time and testing.

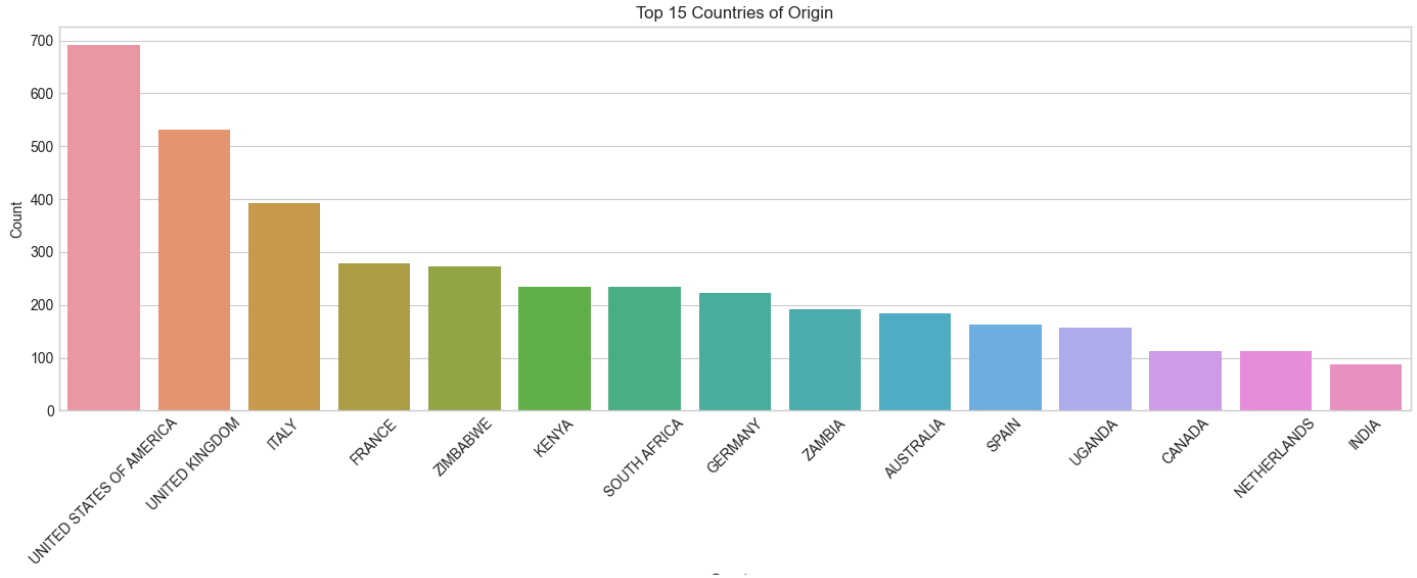


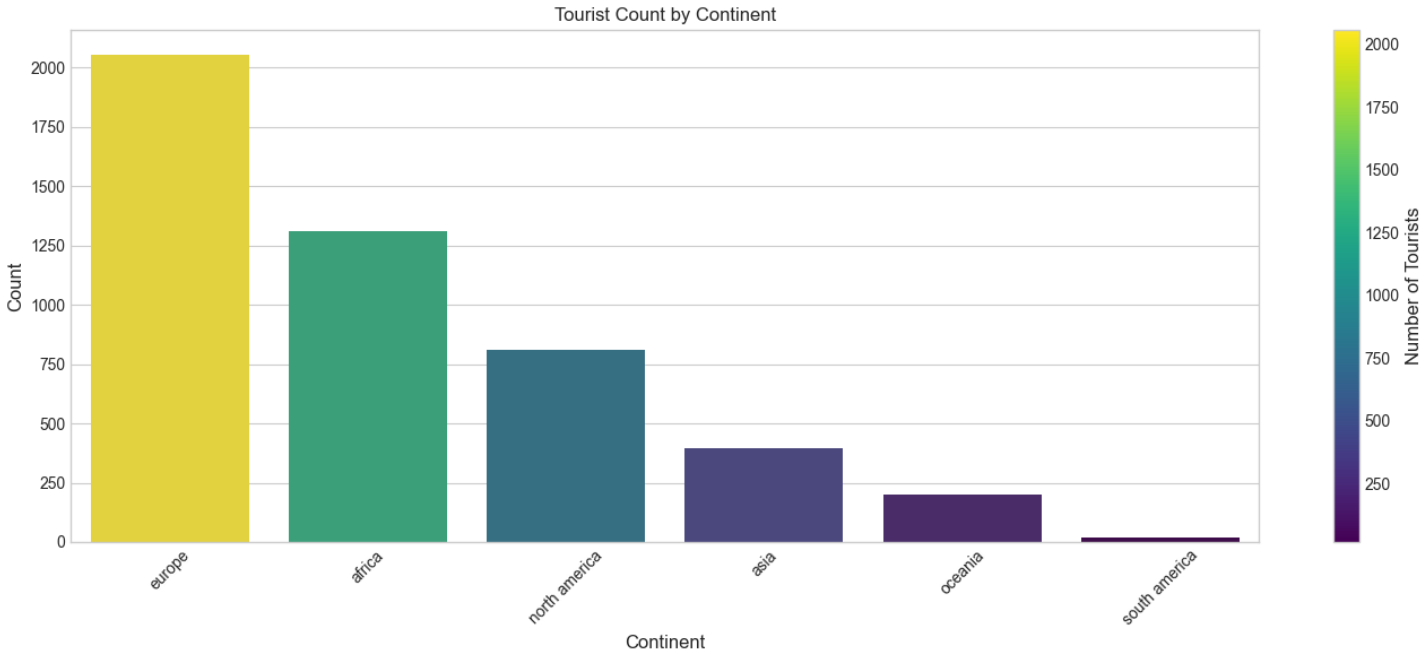

Distribution of Tourist Expenditure
Frequency
Total Cost (TZS)
1e8
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0

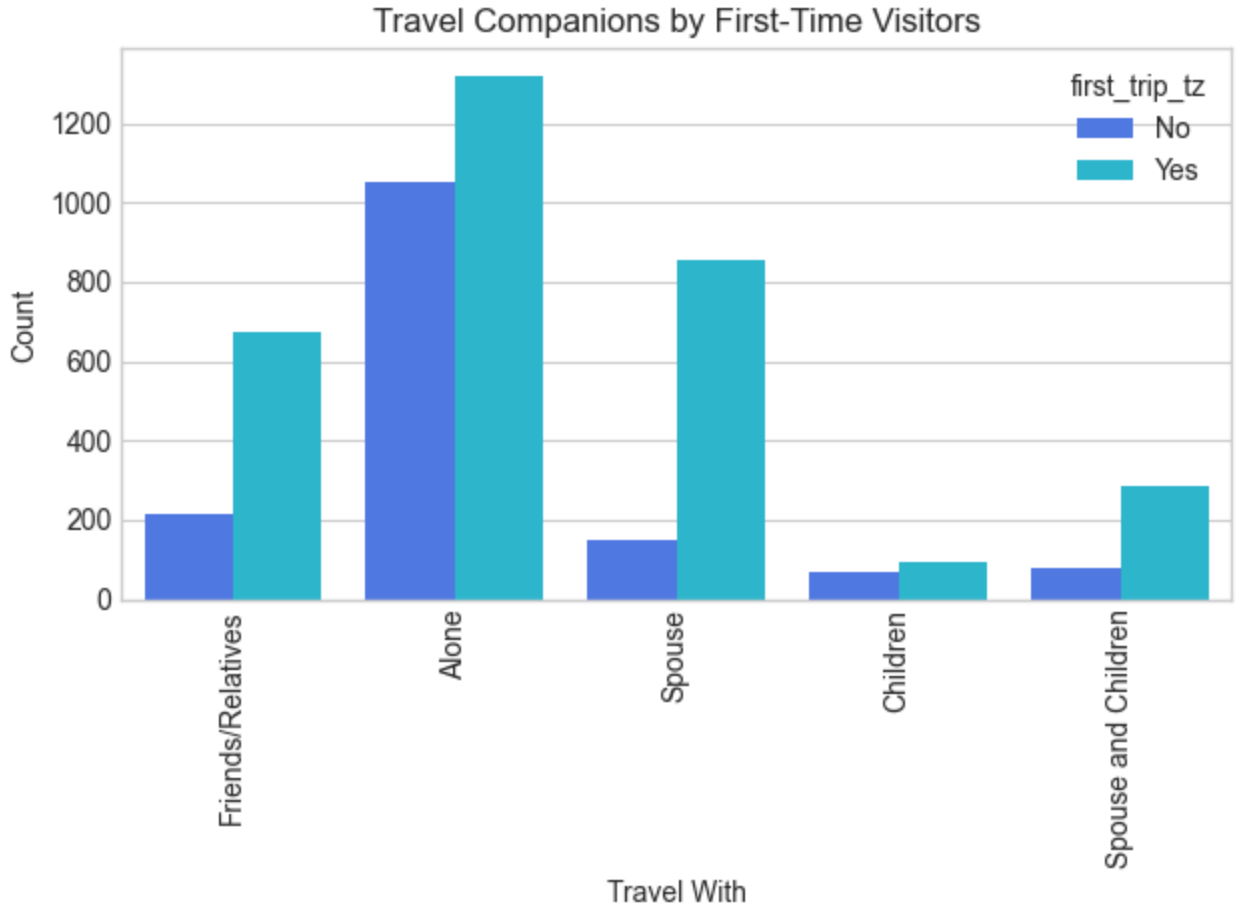

Log-Transformed Tourist Expenditure
Frequency
Log(Total Cost + 1)
0
50
100
150
200
250
300
350
11
12
13
14
15
16
17
18

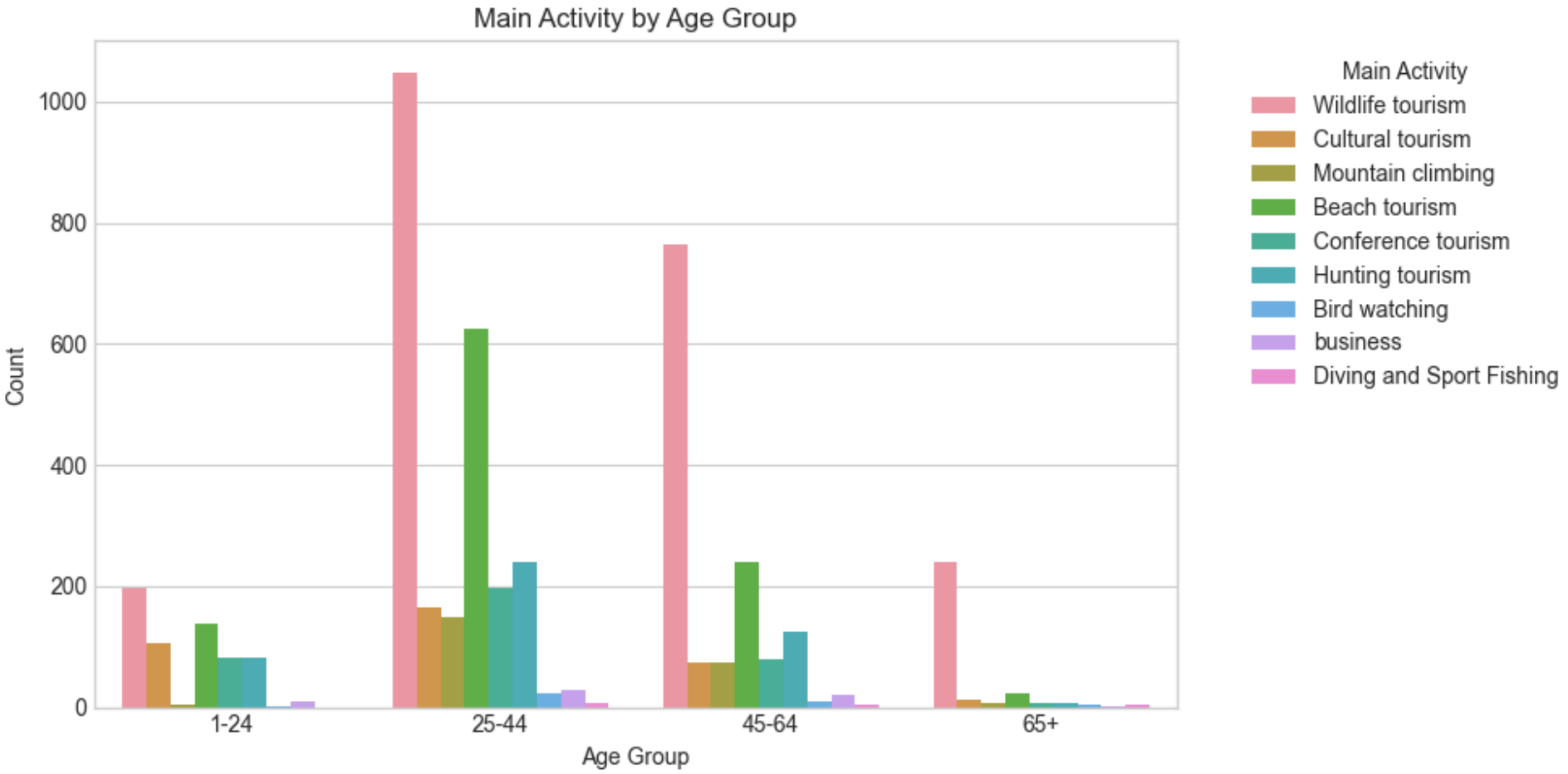

Distribution of Purpose of Visit
Count
2500
2000
1500
1000
500
0
Leisure and Holidays
Business
Visiting Friends and Relatives
Meetings and Conference
Volunteering
Other
Scientific and Academic
Purpose


Distribution of Age Groups
Count
2500
2000
1500
1000
500
0
25-44
45-64
1-24
65+
Age Group

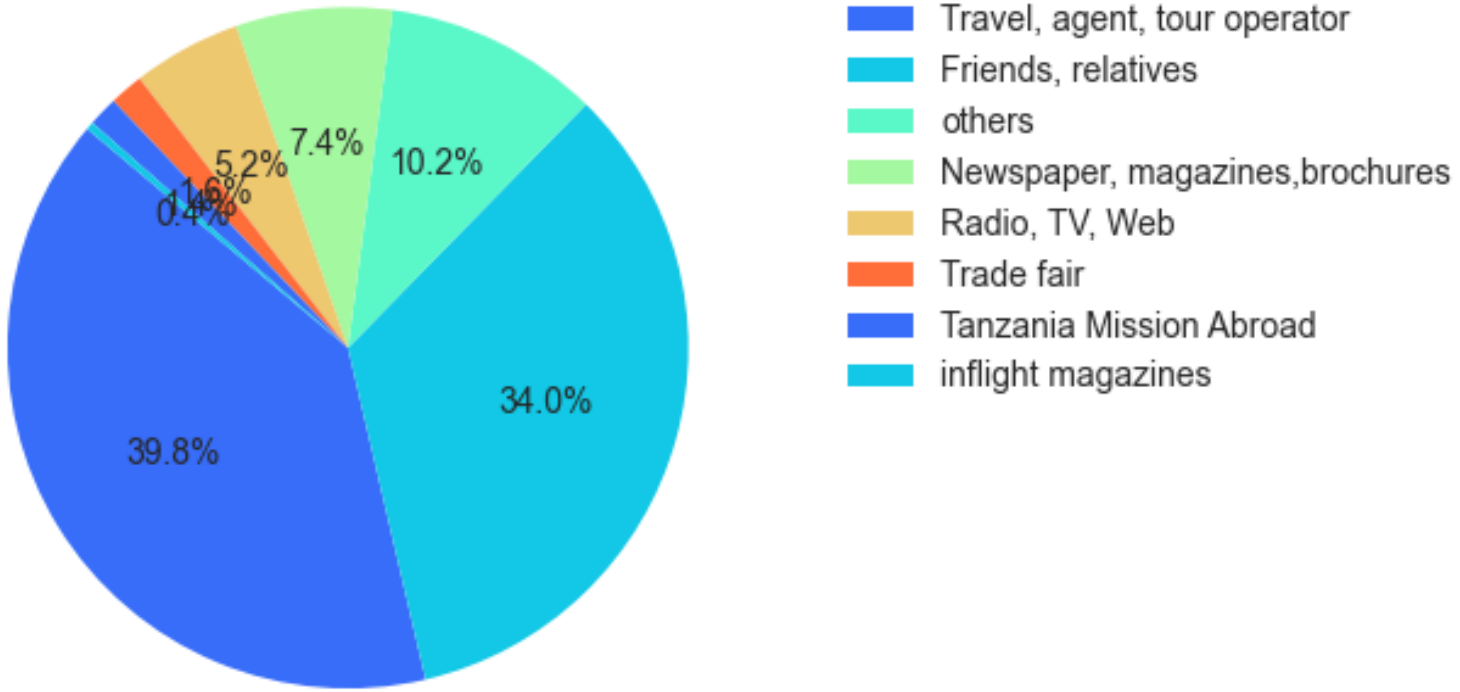

Top 15 Countries of Origin
Count
0
100
200
300
400
500
600
700
UNITED STATES OF AMERICA
UNITED KINGDOM
ITALY
FRANCE
ZIMBABWE
KENYA
SOUTH AFRICA
GERMANY
ZAMBIA
AUSTRALIA
SPAIN
UGANDA
CANADA
NETHERLANDS
INDIA

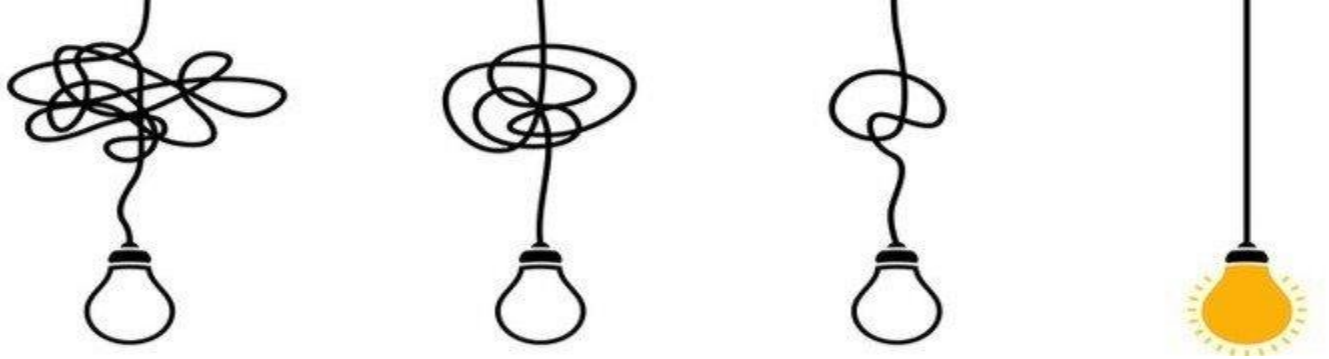

Tourist Count by Continent
Count
0
250
500
750
1000
1250
1500
1750
2000
europe
africa
north america
asia
oceania
south america
Continent
Number of Tourists
250
500
750
1000
1250
1500
1750
2000


Travel Companions by First-Time Visitors
first_trip_tz
No
Yes
Count
0
200
400
600
800
1000
1200
Friends/Relatives
Alone
Spouse
Children
Spouse and Children
Travel With


Main Activity by Age Group
Count
1000
800
600
400
200
0
1-24
25-44
45-64
65+
Age Group
Main Activity
Wildlife tourism
Cultural tourism
Mountain climbing
Beach tourism
Conference tourism
Hunting tourism
Bird watching
business
Diving and Sport Fishing

Where did you get the information to Tanzania

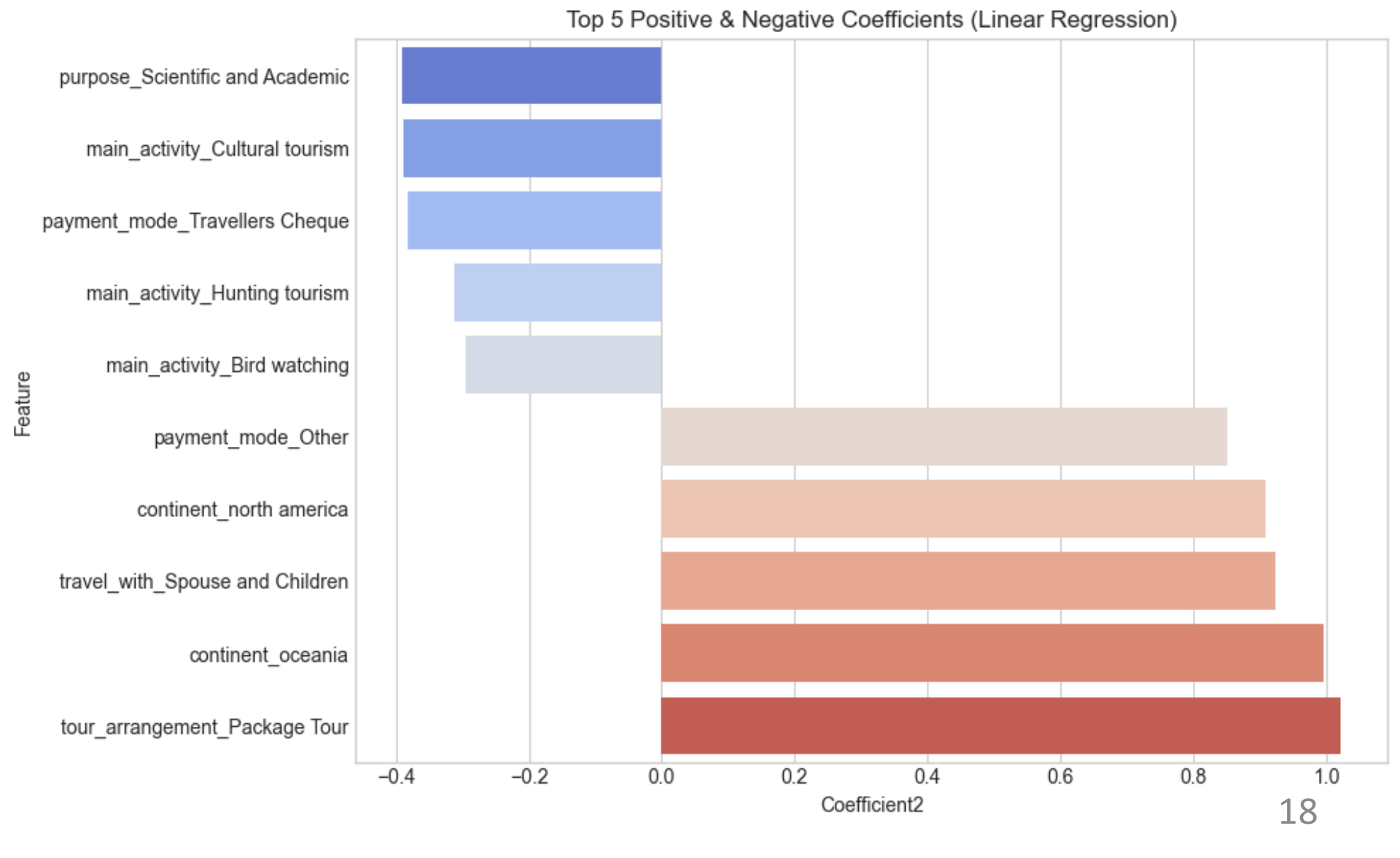


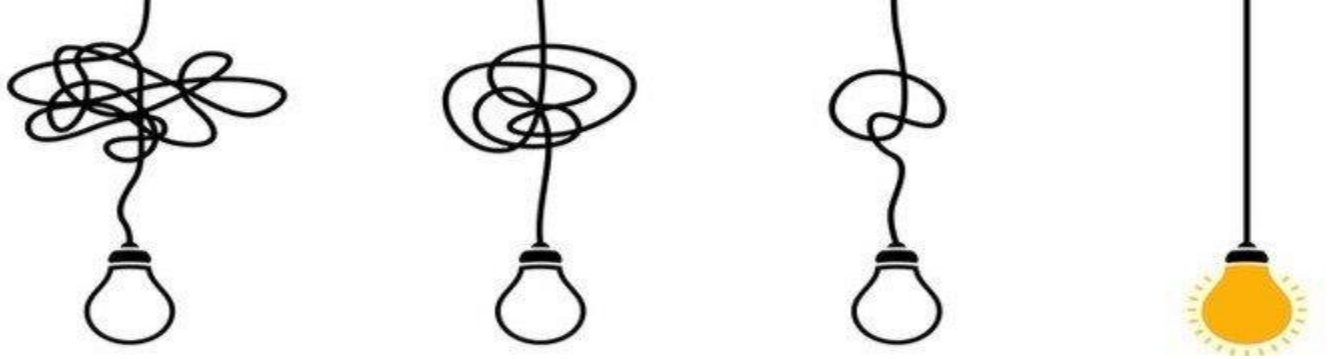


Solution

Base-Line

- **Objective**: Established a baseline to predict tourist expenditure using simple models.
- **Data Prep**: Used preprocessed data with encoded categorical variables (e.g., OneHotEncoder for "country", "purpose").
- **Model Selection**: Implemented Linear Regression as the initial baseline model.
- **Evaluation Matric:** Mean Absolute Error (MAE), R^2.
- **Performance**: Achieved a baseline R^2 of ~0.52 and MAE of ~0.88, indicating moderate fit.
- **Insights**: Length of stay and package tours showed significant impact on expenditure.
- **Next Steps**: Identified need for advanced models (e.g., Random Forest, Gradient Boosting) to improve accuracy.


Top 5 Positive & Negative Coefficients (Linear Regression)
purpose_Scientific and Academic
main_activity_Cultural tourism
payment_mode_Travellers Cheque
main_activity_Hunting tourism
main_activity_Bird watching
payment_mode_Other
continent_north america
travel_with_Spouse and Children
continent_oceania
tour_arrangement_Package Tour
Feature
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Coefficient2



Solution

Comparison of Models

Machine Learning Models Used

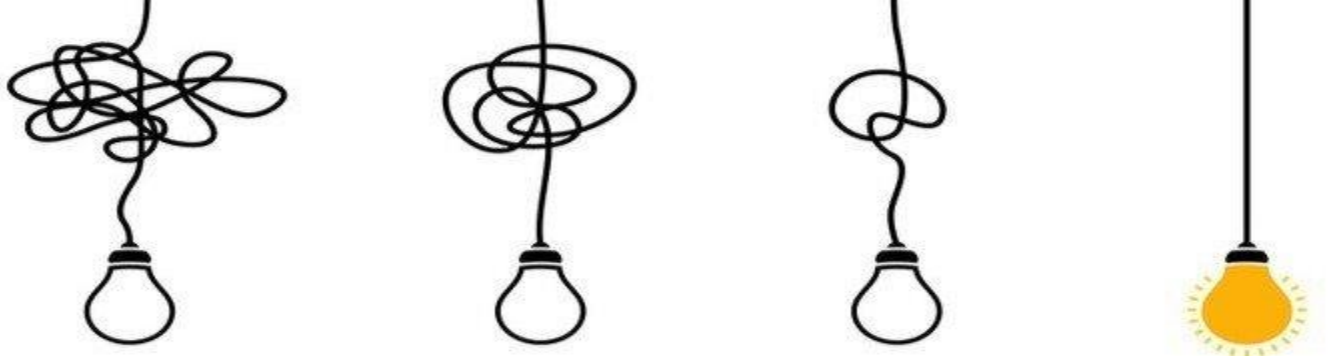
- **Polynomial Features, Ridge Regression**
- **Polynomial Features, Elastic Net**

- **Ridge Regression**
  - Purpose: A regularized linear regression model to handle potential multicollinearity in the data.

- **Lasso Regression**
  - Purpose: Another regularized linear regression model, useful for feature selection by shrinking less important coefficients to zero.

- **Elastic Net**
  - Purpose: Combines L1 (Lasso) and L2 (Ridge) regularization to balance feature selection and regularization.

- **Random Forest Regressor**
  - Purpose: An ensemble model using multiple decision trees, likely tested for its ability to capture non-linear relationships in the data.



Solution

Comparison of Models

Machine Learning Models Used

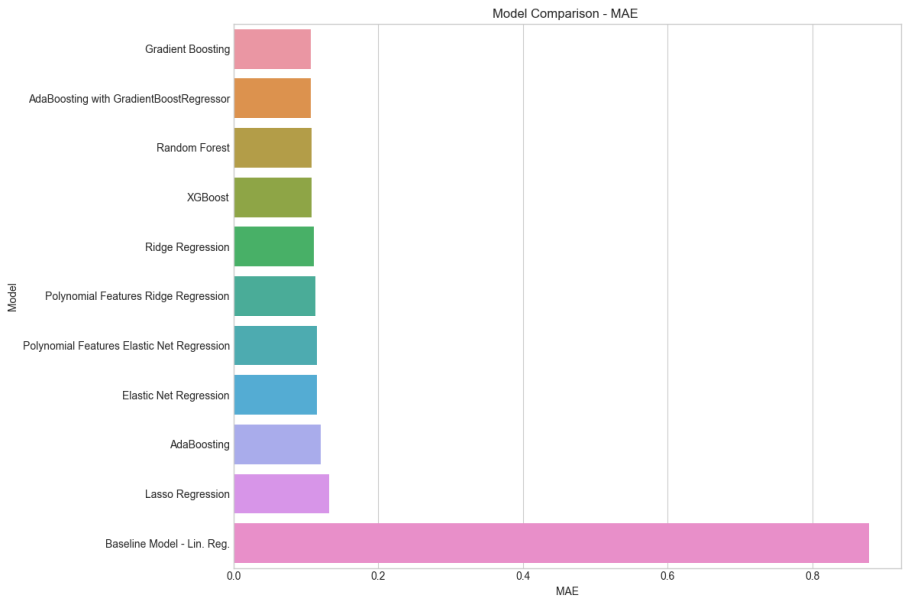
- **Gradient Boosting Regressor**
  - Purpose: An ensemble boosting model that builds trees sequentially to minimize errors. Noted as the best-performing model with MAE of 0.1060 and R^2 of 0.5544 in the conclusions.

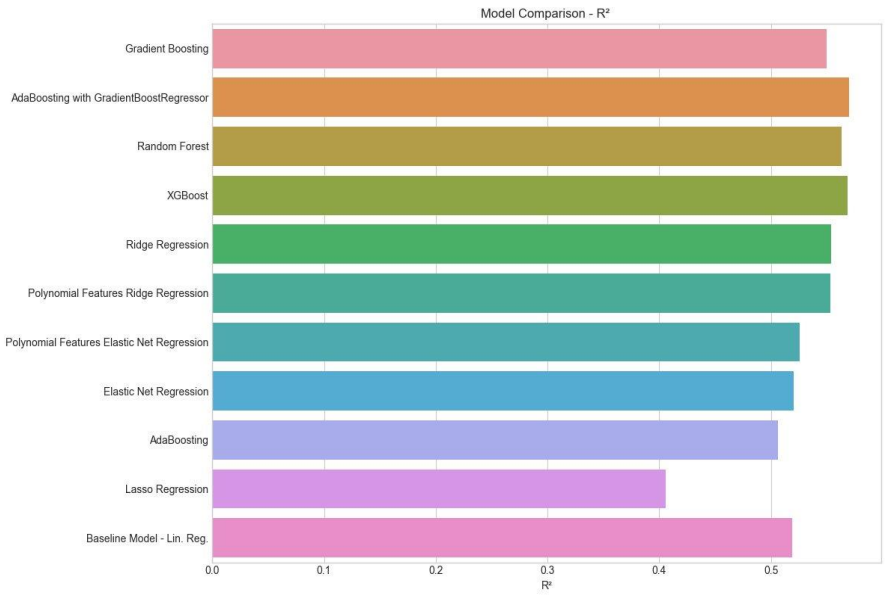
- **AdaBoost Regressor**
  - Purpose: An ensemble boosting model that adjusts weights of incorrectly predicted instances, potentially tested for robustness.

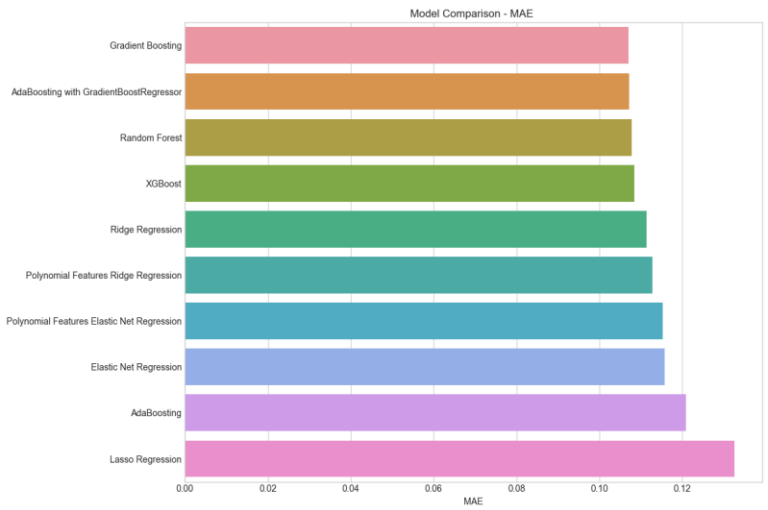
- **AdaBoost with Gradient Boosting Regressor**

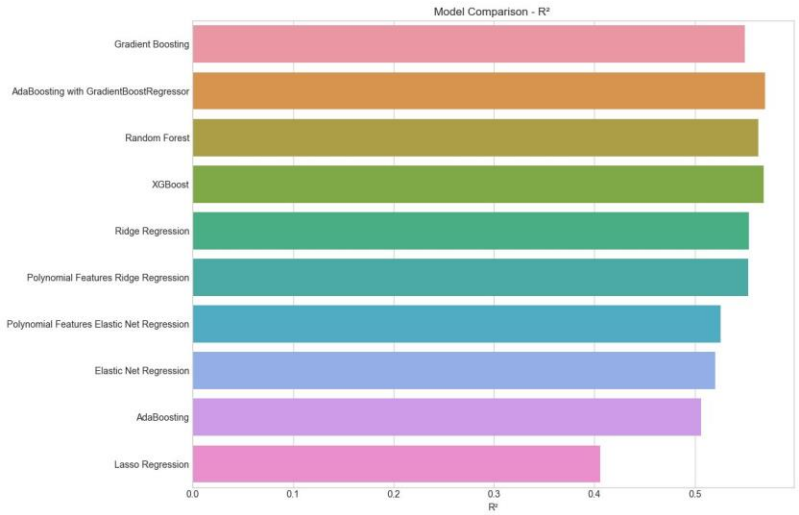
- **XGBoost Regressor**
  - Purpose: An optimized gradient boosting model known for high performance and scalability, likely evaluated for its predictive power.

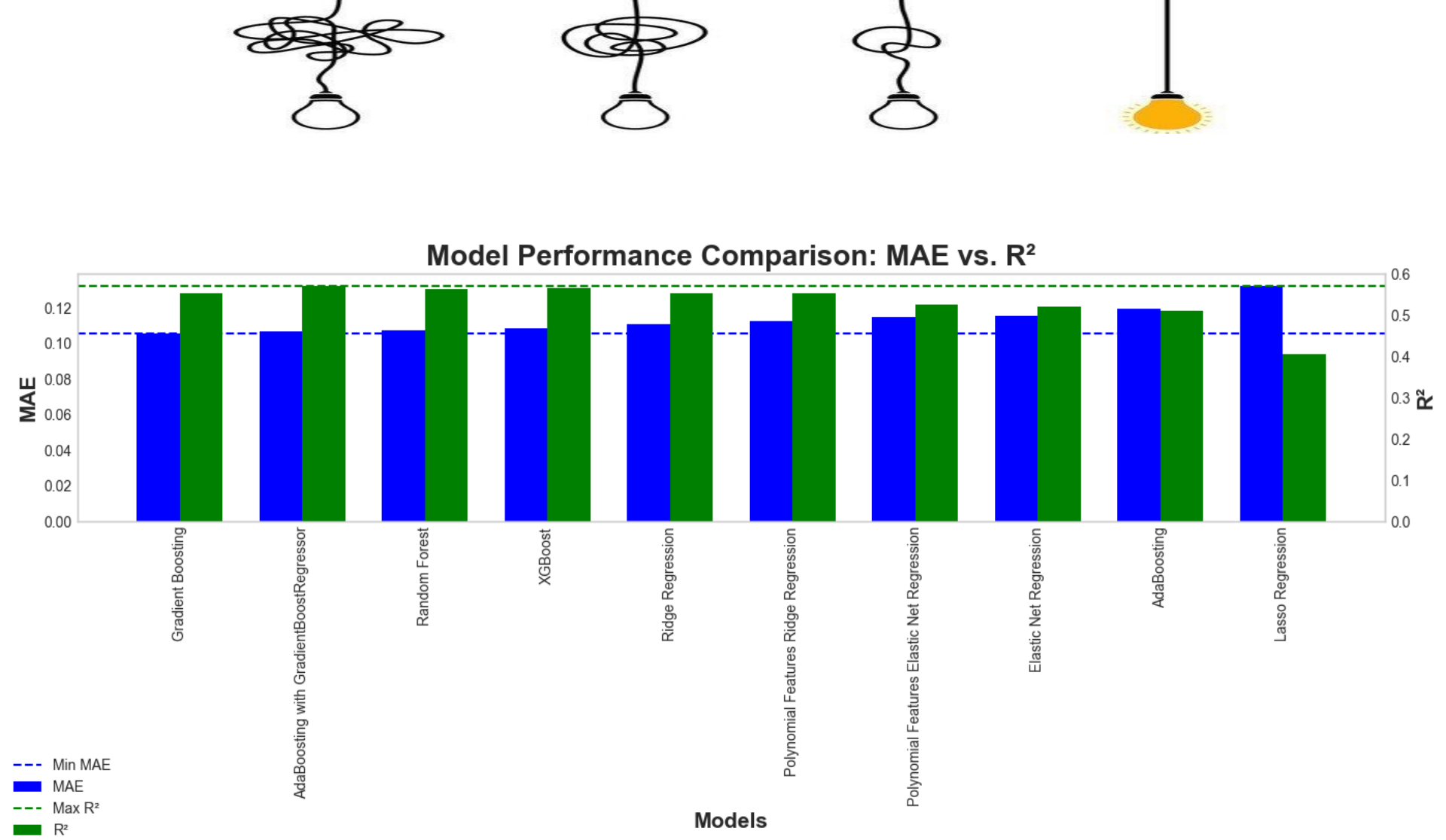
- **Decision Tree Regressor**
  - Purpose: A single decision tree model, possibly used as a baseline or component in ensemble methods.

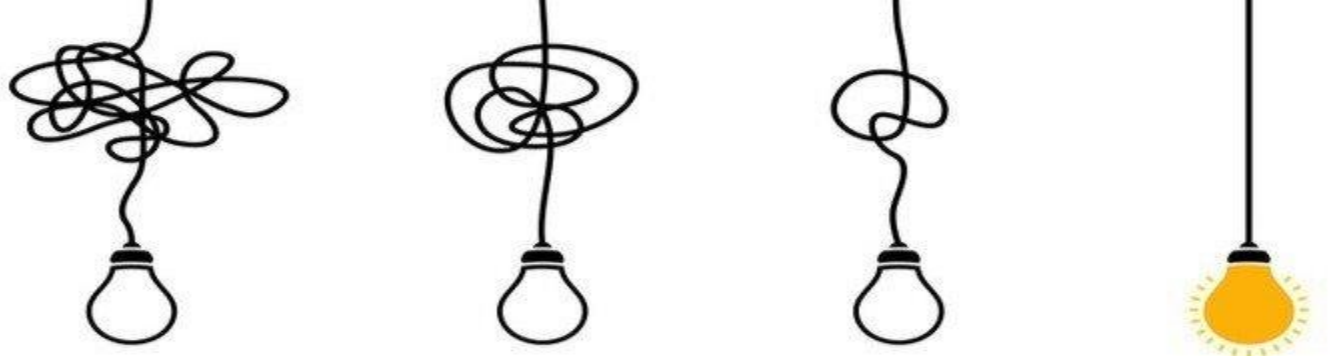

Model Comparison - MAE
Gradient Boosting
AdaBoosting with GradientBoostRegressor
Random Forest
XGBoost
Ridge Regression
Polynomial Features Ridge Regression
Polynomial Features Elastic Net Regression
Elastic Net Regression
AdaBoosting
Lasso Regression
Baseline Model - Lin. Reg.
Model
0.0
0.2
0.4
0.6
0.8
MAE


Model Comparison - R²
Gradient Boosting
AdaBoosting with GradientBoostRegressor
Random Forest
XGBoost
Ridge Regression
Polynomial Features Ridge Regression
Polynomial Features Elastic Net Regression
Elastic Net Regression
AdaBoosting
Lasso Regression
Baseline Model - Lin. Reg.
0.0
0.1
0.2
0.3
0.4
0.5
R²


Model Comparison - MAE
Gradient Boosting
AdaBoosting with GradientBoostRegressor
Random Forest
XGBoost
Ridge Regression
Polynomial Features Ridge Regression
Polynomial Features Elastic Net Regression
Elastic Net Regression
AdaBoosting
Lasso Regression
0.00
0.02
0.04
0.06
0.08
0.10
0.12
MAE


Model Comparison - R²
Gradient Boosting
AdaBoosting with GradientBoostRegressor
Random Forest
XGBoost
Ridge Regression
Polynomial Features Ridge Regression
Polynomial Features Elastic Net Regression
Elastic Net Regression
AdaBoosting
Lasso Regression
0.0
0.1
0.2
0.3
0.4
0.5
R²

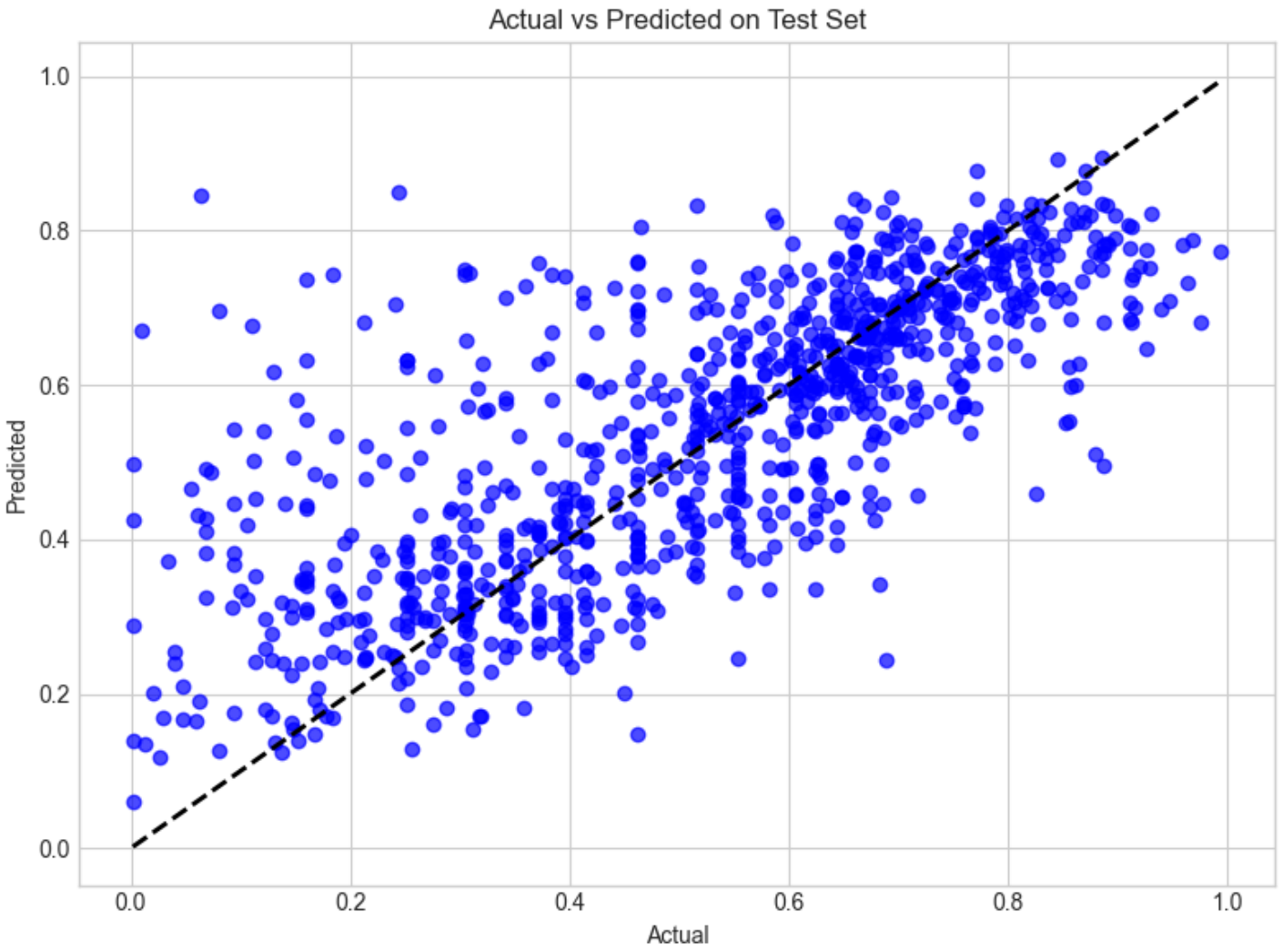

Model Performance Comparison: MAE vs. R²
MAE
R²
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Gradient Boosting
AdaBoosting with GradientBoostRegressor
Random Forest
XGBoost
Ridge Regression
Polynomial Features Ridge Regression
Polynomial Features Elastic Net Regression
Elastic Net Regression
AdaBoosting
Lasso Regression
Min MAE
MAE
Max R²
R²
Models

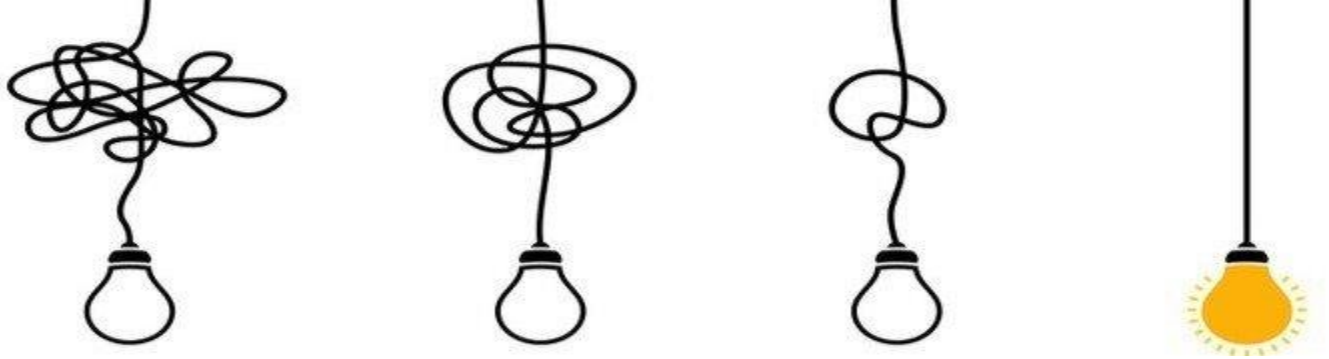


Solution

Analyzing Gradient Boosting

- **Model Performance**:
  - Gradient Boosting Regressor emerged as the best-performing model.
  - Metrics: MAE = 0.1060, R^2 = 0.5544.
- **Why It Excelled**:
  - Handles non-linear relationships and interactions between features effectively.
  - Key features: package tours, length of stay, group size, purpose (leisure), accommodation.
- **Impact**:
  - Predicts tourist expenditure (total_cost in TZS) with moderate accuracy (R^2 ~ 0.55).
  - Supports actionable insights: promote packages, extend stays, target groups.
- **Takeaway**:
  - Gradient Boosting balances complexity and interpretability, making it ideal for tourism revenue optimization.


Actual vs Predicted on Test Set
Predicted
Actual
0.0
0.2
0.4
0.6
0.8
1.0

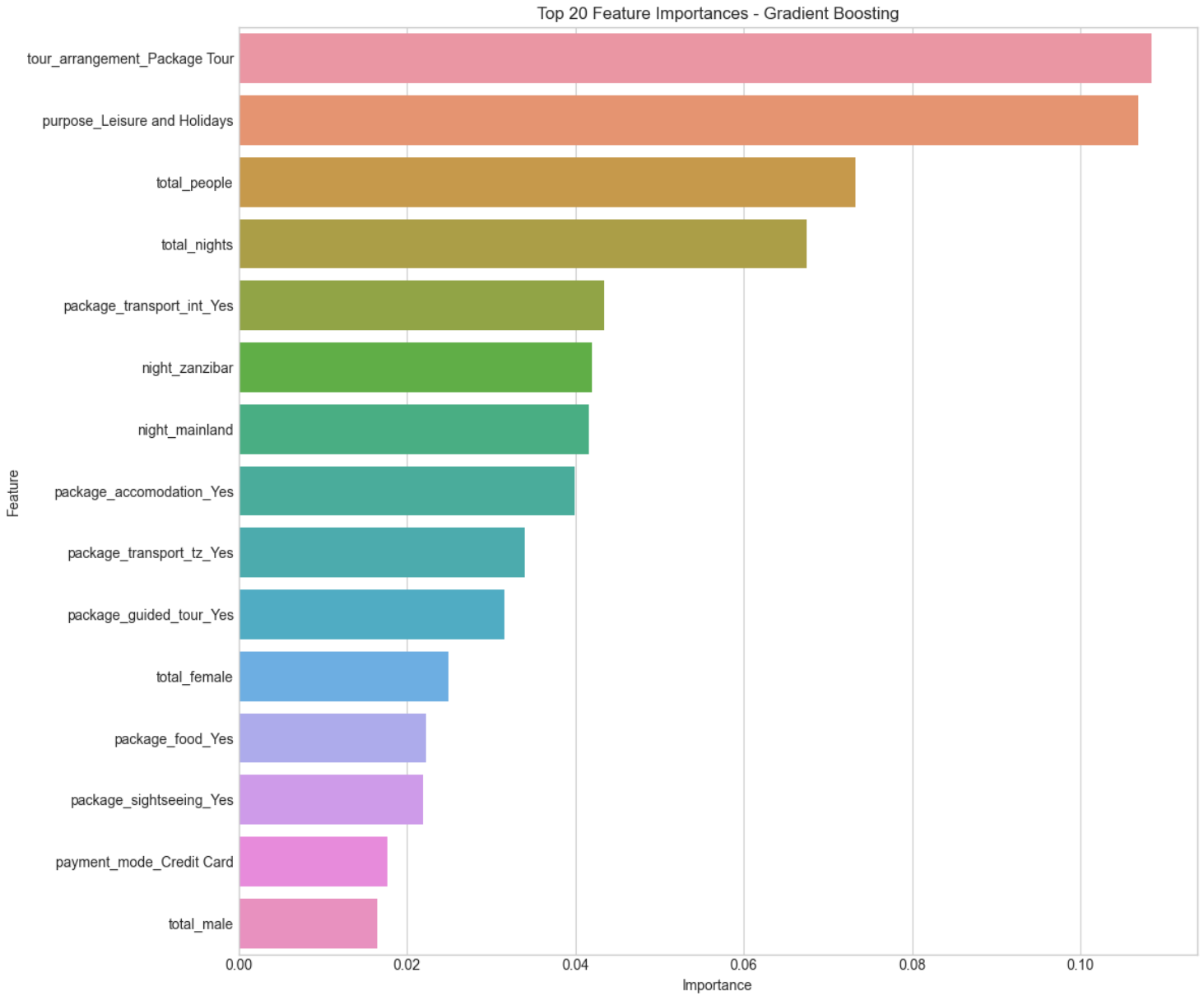


Solution

Gradient Boosting Features

- The Gradient Boosting model leverages a mix of categorical (encoded) and numerical (scaled) features, with the top influencers being (package tours, travel purpose, length of stay, group size, and accommodation packages).

- Additional features from the dataset enhance its predictive power, capturing diverse aspects of tourist behavior.


Top 20 Feature Importances - Gradient Boosting
tour_arrangement_Package Tour
purpose_Leisure and Holidays
total_people
total_nights
package_transport_int_Yes
night_zanzibar
night_mainland
package_accomodation_Yes
package_transport_tz_Yes
package_guided_tour_Yes
total_female
package_food_Yes
package_sightseeing_Yes
payment_mode_Credit Card
total_male
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance

Optimizing Revenue through Predictive Insights

- **Key Objective**: Predict tourist expenditure (total_cost in TZS) using tourist data.

- **Best Model**: Gradient Boosting (MAE: 0.1060, R^2: 0.5544).

- **Key Drivers**:
  - Package tours vs. independent travel
  - Leisure & holiday purpose
  - Length of stay (nights)
  - Group size (people)
  - Accommodation packages

- **Data Product Opportunity**:
  - Tool for stakeholders to forecast revenue based on tourist profiles.
  - Suggests optimizations (e.g., extend stays, add packages).

- **Impact**: Enhances tourism planning, targets high-spending segments (e.g., leisure groups).

Top 5 Factors Influencing Tourist Expenditure

Package Tours (41.2%): Tourists on package tours spend significantly more

Leisure Travel (14.7%): Leisure travelers outspend other visitor types

Length of Stay (14.1%): Longer stays correlate with higher spending

Group Size (7.8%): Larger travel groups spend more in total

Accommodation Packages (2.4%): Tourists with accommodation included spend more

Recommendations for Tourism Stakeholders

1. Develop Comprehensive Package Tours

- Include accommodation, transportation, and guided experiences
- Focus on high-quality, all-inclusive offerings

2. Target Leisure Travelers

- Allocate more marketing resources to leisure travel segment
- Develop specialized offerings for holiday travelers

3. Encourage Extended Stays

- Create multi-day itineraries and tour packages
- Offer incentives for longer bookings

4. Cater to Group Travel

- Develop group-friendly accommodation and activities
- Create special rates for larger travel parties


TRAVEL
BUDGET
• FLIGHT
• HOTEL
• CAR
• RESTAURANT
• FOOD
• FUEL
http://localhost:8501/

Tanzania Tourism Expenditure Predictor

A tool for tourism stakeholders to predict and optimize revenue

Input Parameters:

- Tourist Origin: United States
- Age Group: 25-44
- Travel Type: Package Tour
- Length of Stay: 10 nights
- Group Size: 2 people
- Purpose: Leisure and Holidays
- Main Activity: Wildlife Tourism

Predicted Expenditure:

$3,250 USD

(7,475,000 TZS)

Optimization Suggestions:

- Add guided tour package (+15%)
- Extend stay to 14 nights (+25%)
- Include Zanzibar (+18%)


Thank
you