Tanzania Tourism Prediction

hh-ds-24-9

by Marina & Felicitas

Reason and Challenge

1. Tourism sector plays a significant role in the Tanzanian economy, contributing about 17% to the country's GDP and 25% of all foreign exchange revenues

2. Tanzania generated 2.4 billion dollars in 2018

3. Tanzania is the **only country in the world** which has allocated more than 25% of its total area for wildlife, national parks, and protected areas.

Challenge:

Goal is to accurately predict tourist expenditure when visiting Tanzania for tourist agencies to improve their offer.

Summary of Data

- Total observations: **4809**
- Total features: **23**

- Target Feature: Total Cost; The total tourist expenditure in TZS (Tanzanian Shillings*)

- Type of data:
  - Demographic
    - Country
    - Age group (1-24, ...)

  - Type of trip
    - People travelling (Alone, Couple, Family, Group...)
    - Purpose of the Trip (Wildlife Exploring, Sport, Business, Beach,,,)
    - Tour arrangements (including transport, insurance, guides...)

  - Destination
    - Zanzibar
    - Mainland

* 1€ = 2662,98 TZS

Data Challenges

- **Predominantly categorical features**

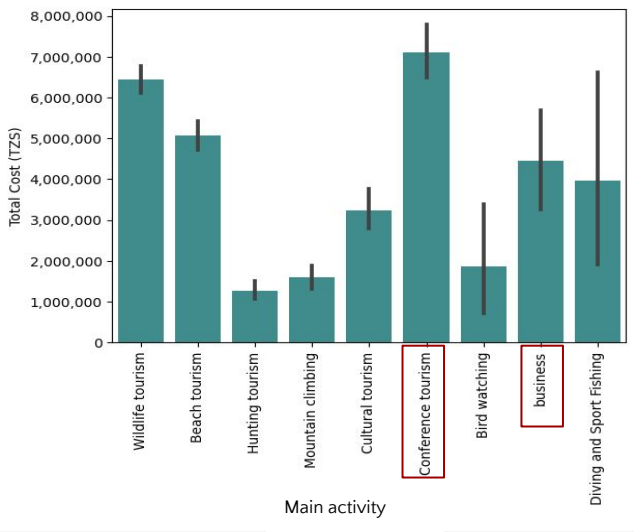
Data Challenges

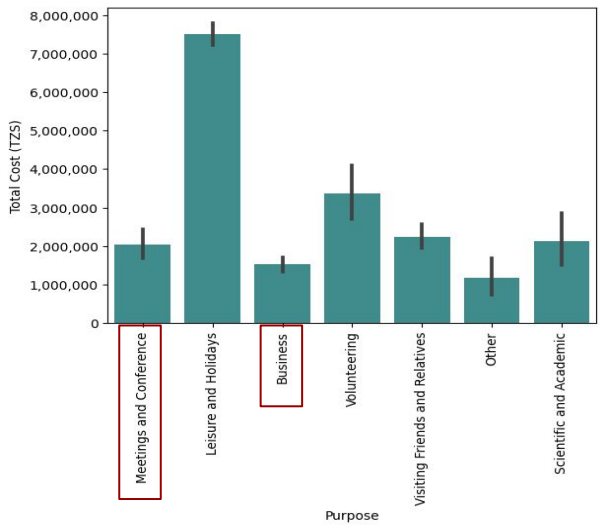
- **Predominantly categorical features**

- **Lack of useful information for the business case**: The dataset could include additional features that would have significantly improved insights and model performance.

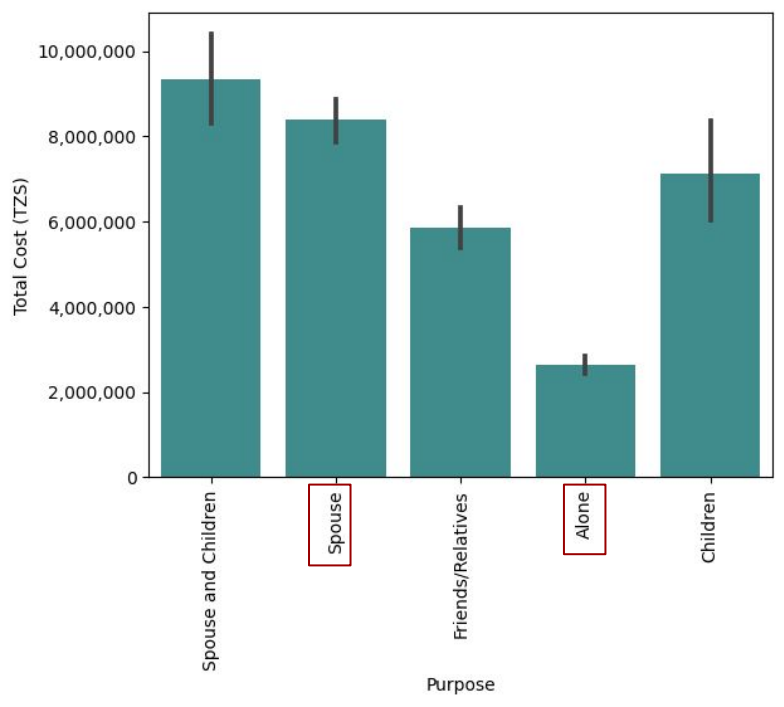
Data Challenges

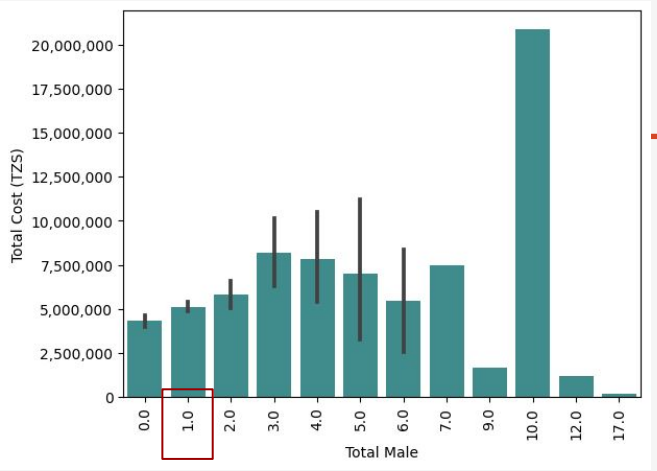
- **Predominantly categorical features**
- **Lack of useful information for the business case**: The dataset could include additional features that would have significantly improved insights and model performance.
- **Redundant survey questions**: Many survey questions exhibited strong correlations, leading to potential multicollinearity issues and redundant information.

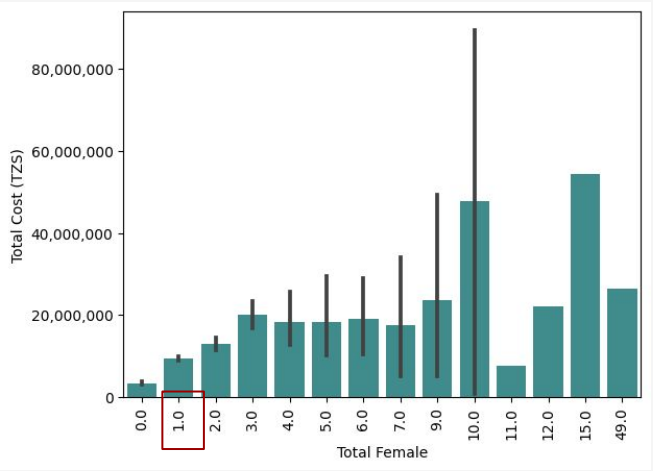




Data Challenges

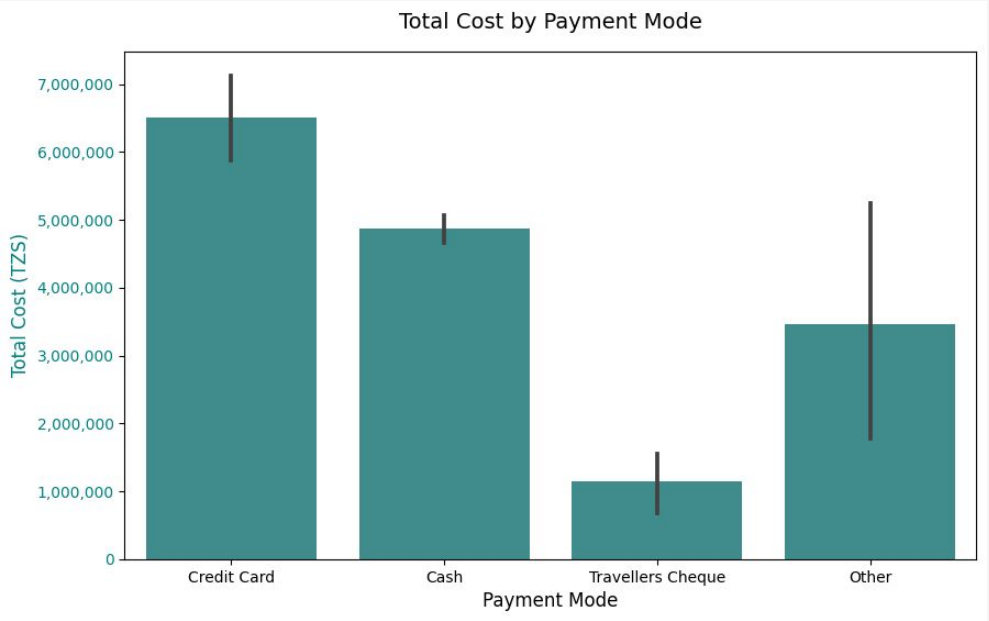


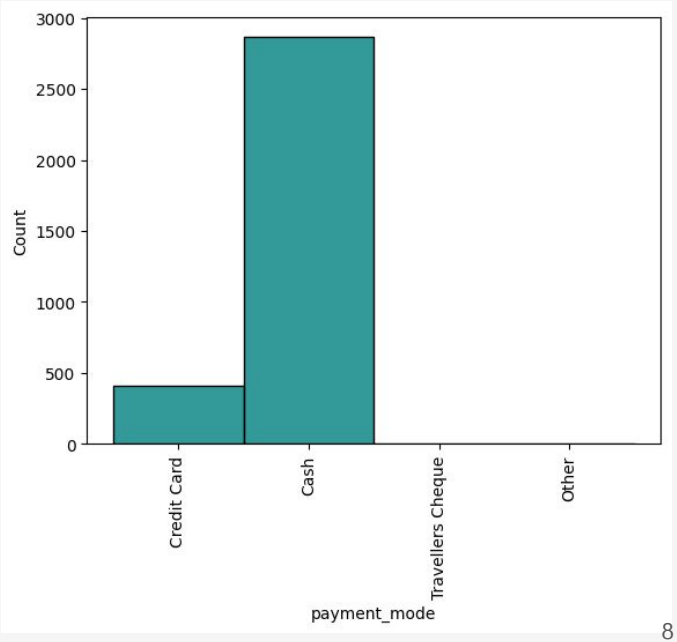




Data Challenges

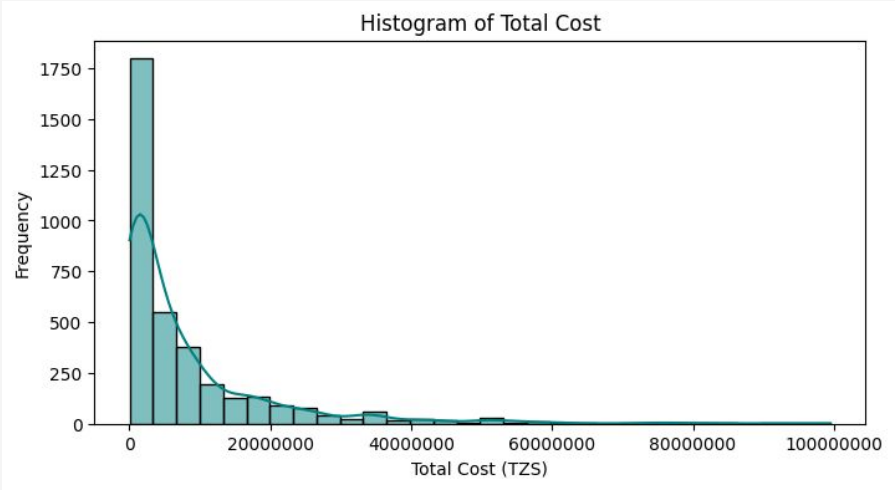
- **Highly skewed data**: Both the features and the target variable were imbalanced, complicating modeling and performance evaluation.

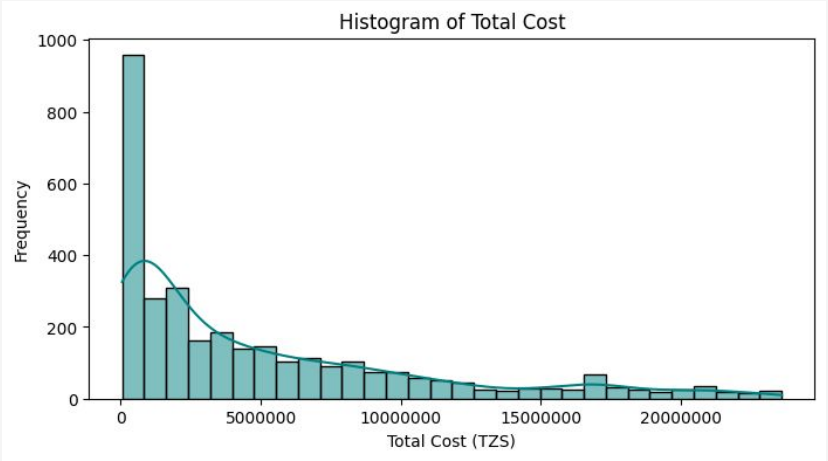




Data Challenges

- **Highly skewed data**: Both the features and the target variable were imbalanced, complicating modeling and performance evaluation.





Removed outliers from train set

Baseline Model

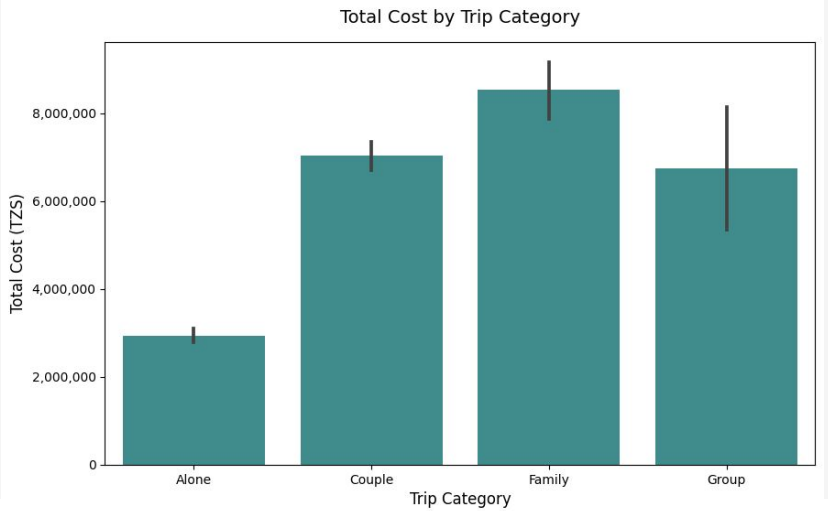
- Used all features (except of a few that had many NA)
- Type of problem: Regression
- Evaluation Metrics: RMSE, MAE, R^2
- Baseline Model: Decision Tree Regressor

Model	Root Mean Squared Error (RMSE):	Mean Absolute Error (MAE):	R-squared (R^2):	R-squared (R^2, %):
Decision Tree (Test Data)	10,979,865 TZS ≈ 4,139.89 €	5,770,815 TZS ≈ 2,175.89 €	0.170	**17%**
Decision Tree (Train Data)	1,058,342 TZS ≈ 398.89 €	263,110 TZS ≈ 99.18 €	0.965	**96.5%**

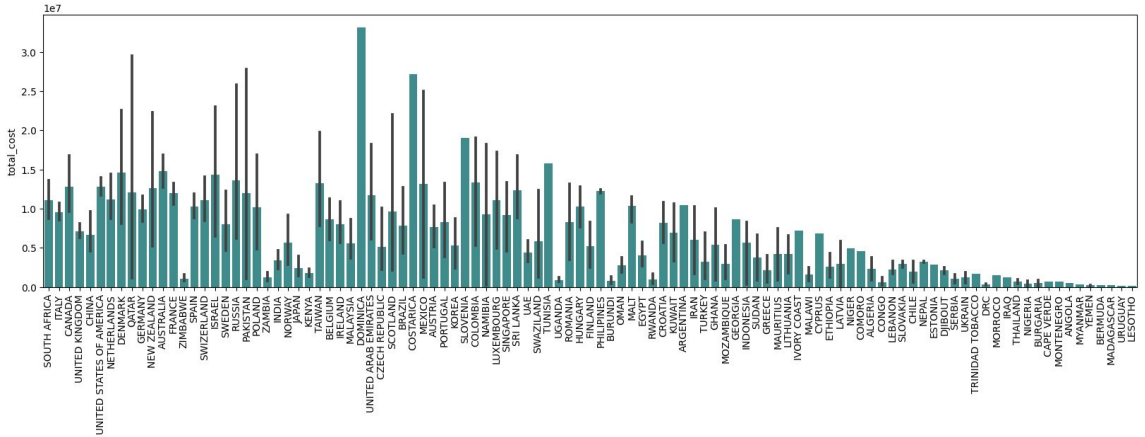
Feature Engineering and iterations on the model

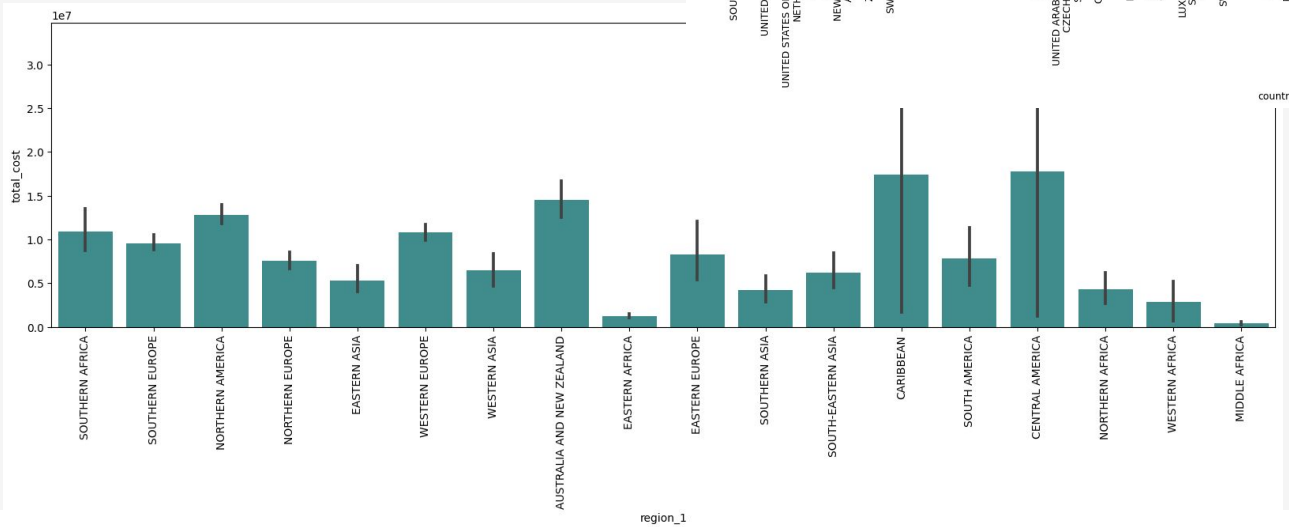
- Iterations on the model
  - Tried different parameters. Tried different models. Grid Search

- Iterations on the data set
  - Dropped some features
  - Created new ones (Trip category, World region)
  - Encoding + Scaling



Feature Engineering and iterations on the model





Summary

Model	Root Mean Squared Error (RMSE):	Mean Absolute Error (MAE):	R-squared (R^2):	R-squared (R^2, %):
Decision Tree (Test Data)	10,979,865 TZS ≈ 4,139.89 €	5,770,815 TZS ≈ 2,175.89 €	0.170	**17%**
Decision Tree (Train Data)	1,058,342 TZS ≈ 398.89 €	263,110 TZS ≈ 99.18 €	0.965	**96.5%**
RandomForestRegressor (Test)	10,808,097 TZS ≈ 4,073.45 €	5,553,628 TZS ≈ 2,094.71 €	0.192	**19.2%**
RandomForestRegressor (Train)	3,970,504 TZS ≈ 1,496.90 €	2,741,397 TZS ≈ 1,032.71 €	0.514	**51.4%**
RandomForestRegressor (Test)	9,495,002 TZS ≈ 3,578.93 €	5,368,076 TZS ≈ 2,022.12 €	0.26	**26%**
RandomForestRegressor (Train)	10,298,890 TZS ≈ 3,876.69 €	5,845,418 TZS ≈ 2,201.08 €	0.402	**40.2%**

Conclusions

- **Prediction**: Total Expenses = Adjusted based on trip features using a Random Forest Regression model.
- On average, our current model's predictions are off by around **2,022.12 €** (Mean Absolute Error), suggesting there is **significant room for improvement** in model accuracy.
- **R^2 Value**: The model explains **only 26.71%** of the variance in total expenses, indicating that while it captures some underlying patterns in the data, it still fails to account for most of the variability in the expenses.

Next Steps

1. **Understand and clean better the data**
2. **Continue iterating on the model**
   a. Feature engineering
   b. Try other models
   c. Hyperparameter tuning

3. [*if possible*] **Improve the survey to have more and better data – aligned with the use case**

Thank you

Conclusions