INTRODUCTION

Tanzania's tourism makes up about **17%** of the country's GDP and **25%** of all foreign exchange revenues. This is a significant proportion of Tanzania's income.

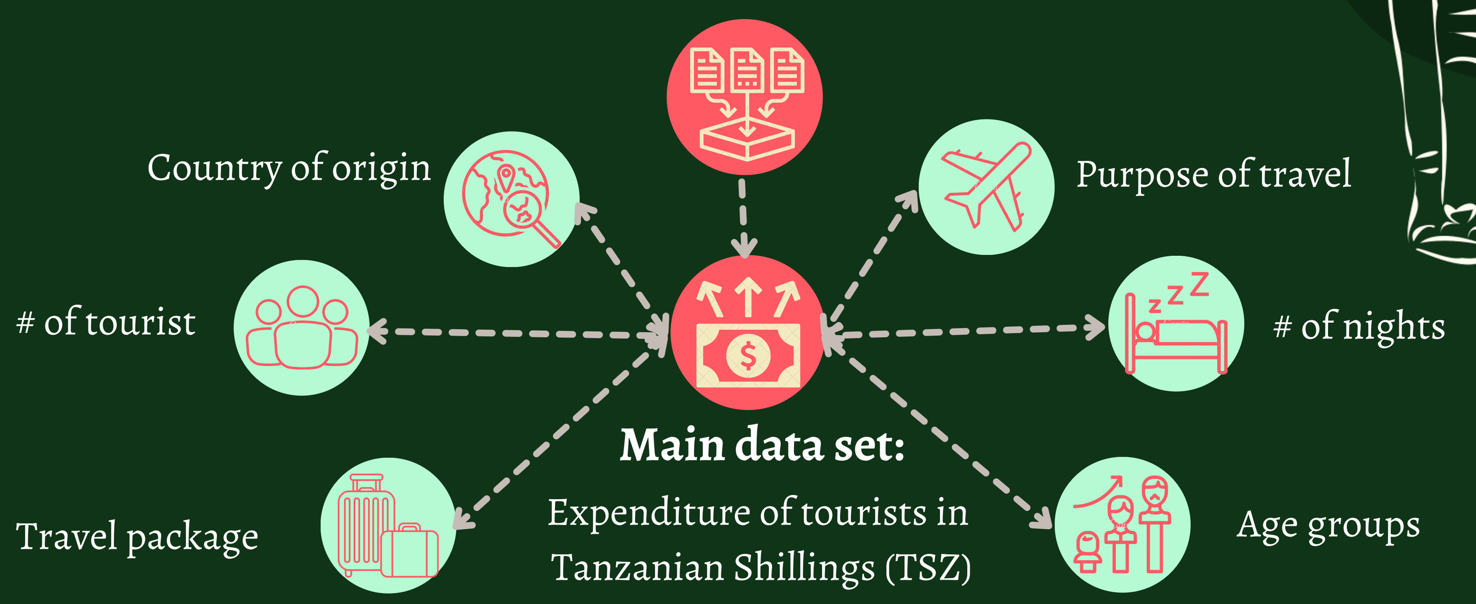
Correspondingly, **25%** of Tanzania's land makes up the main attraction including wildlife, national parks and protected areas.

This Shiny app can help users including the Tanzania Tourism Board and tour operators **invest** their marketing budget wisely by **analyzing factors** that contribute to tourism income.

METHODOLOGY

SOURCE OF DATA

Zindi & Natual Bureau of Statistics of Tansania



SHINY APP USAGE

COMPONENTS OF SHINY APP

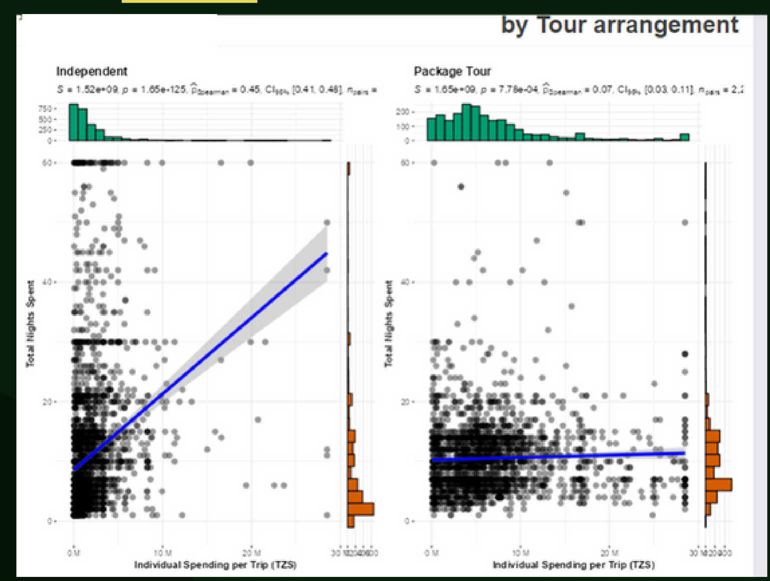
1. **DASHBOARD**
   Overview of largest tourist contribution by count, expenditure, spent/night etc
2. **HYPOTHESIS TESTING**
   Using different parts of data analysis to discover statistically significant factors from spending behaviour of groups, tourist region of origin and specific drivers
3. **CLUSTERING**
   Based on categorical factors, users can cluster the tourist groups using Latent Class Analysis
4. **REGRESSION MODEL**
   Predict tourist expenditure based factors. The options to use decision tree and random forest are presented as well, with the ability to tune the models further.

PROGRAMMING LANGUAGE

R programming was used for the data processing, statistical analyses and building models. Shiny Dashboard is used to build the web application. Packages used include Shiny, shinydashboard, shinyWidgets, shinyjs, tidyverse, ggstatsplot, plotly, DT, caret, rpart, sparkline, visNetwork, ranger, poLCA, sf, tmap, and ExPanDaR.

ANALYSIS

Correlation of # nights and $/trip by categories

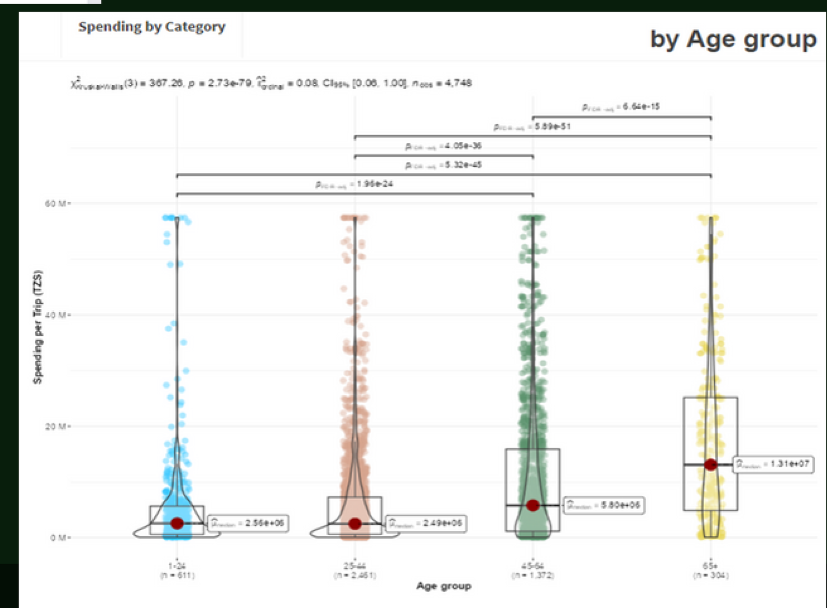


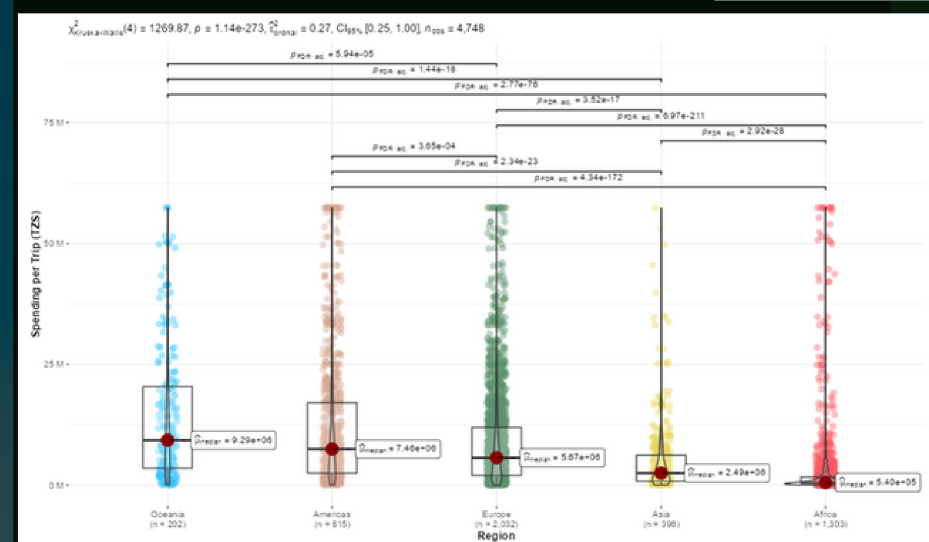
Correlation scores suggest significant differences in spending based on tour arrangement. This is also reflected in other variables relating to package tour arrangement (accommodation/ food/ transportation/ tour/ sightseeing).

Independent travel is highly correlated with more nights and increased spending.

Spending by Categories

Mean spending is significantly different across age groups; **older tourists** tend to **spend more**, likely due to higher spending power.

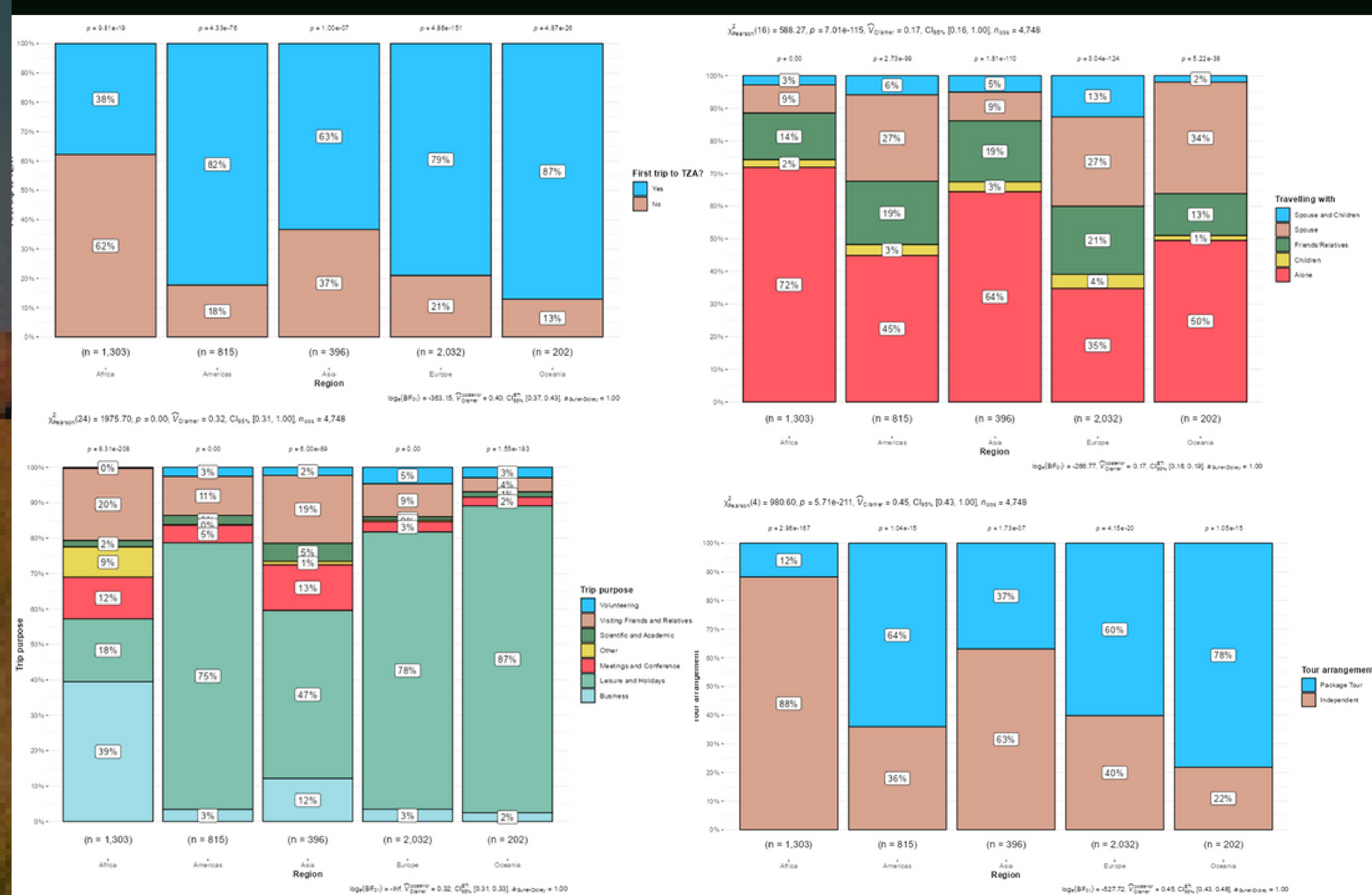




Spending by Region

Mean spending differs significantly across regions. Tourists from **Africa** spend **significantly lower** than other regions, while tourists from **Oceania** spend the **most**.

Demographics by Region



Why do African tourist spend so little? Their demographics show them being:

Highest % in:
Business travel
Travel Alone
Independent (no tour)
Not first timers

Lowest % in:
Leisure travel

Why do Oceania tourist spend the most? Their demographics show them being:

Lowest % in:
Business travel
Independent (no tour)
Not first timers

Highest % in:
Leisure travel

Clustering insights

When the number of classes is 5 and below, repetitions have minimal impact on diagnostic statistics; **BIC score** is visibly impacted **by repetition** when there are **>5 classes**.
Trend for AIC score and Likelihood Ratio mirrors that of BIC score but not Entropy i.e. model with the **lowest BIC** score **does not** always give **best Entropy score**.

- Literature review showed that BIC is most widely reported and considered the most reliable model fit indicator. In contrast, an Entropy value close to 1 is deemed ideal, which is the case for all our solutions. Hence, we will not rely solely on Entropy to determine the final solution.
- **Class = 8** and **repetition = 6** gives the best class solution.

Number of Classes	BIC	AIC	Likelihood	Entropy	BIC	AIC	Likelihood	Entropy
	Number of Repetitions = 1				Number of Repetitions = 6			
2	104,060	104,439	33,502	0.999	104,060	104,439	33,502	0.999
3	100,014	100,587	29,396	0.999	100,014	100,587	29,396	0.999
4	96,170	96,937	25,493	1	96,170	96,937	25,493	1
5	94,539	95,498	23,801	0.991	94,539	95,498	23,801	0.991
6	93,848	95,000	23,050	0.986	93,493	94,646	22,695	0.983
7	93,438	94,784	22,580	0.981	92,803	94,149	21,945	0.978
8	93,015	94,554	22,098	0.985	92,473	94,012	21,555	0.974

Literature also suggest reviewing the class size when identifying the best solution, but there are no guidelines on what is a good size. In our selected solution, the smallest class has **at least 5%** of the sample. Thus, our **selected** solution should be taken as the **final class solution**.

Regression insights

Random Forest will always produce a **better baseline model** (higher RMSE/MAE) than Decision Trees, as it aggregates many Decision Trees to limit overfitting and error due to bias.

Among the 3 resampling techniques for Random Forest models, **kfold Cross Validation outperforms** Bootstrap Resampling across all diagnostic statistics when there are **fewer trees (50)**, but there is **marginal difference** as the number of trees **increases (200/500)**.

- Literature review suggests that with increased iterations, both methods will produce a similar error estimate as once the OOB error stabilizes, it will converge to the cross-validation error. However, **Bootstrap Resampling** has the advantage of requiring **less computation**.

Random Forest	RMSE	MAE	Rsquared	RMSE	MAE	Rsquared	RMSE	MAE	Rsquared
	Number of Trees = 50			Number of Trees = 200			Number of Trees = 500		
Bootstrap Resampling	8,478	5,082	0.42	8,375	4,998	0.434	8,346	4,967	0.438
kfold (k=10) Cross Validation	8,406	5,050	0.43	8,338	4,977	0.439	8,134	4,968	0.442
Repeated (x5) kfold (k=10) Cross Validation	8,309	5,030	0.443	8,345	4,984	0.448	8,347	4,980	0.438

As the number of trees increase in the forest (200 vs. 500), the increase in accuracy becomes marginal. Literature review suggests that Random Forest models generally hit a performance plateau with increased trees. As such, selecting a **smaller number of trees within the "plateau"** provides a good balance of **better diagnostic statistics and faster processing**.

FUTURE WORK

High- spending visitors are currently from:
Europe
USA

With developing countries becoming more affluent, there may be new key tourist sources.

An area for future work would be to allow for more updated datasets collected by Tanzania NBS to be uploaded into the app.

Current prediction models have mediore results

Other predictive models like XGBoost could be explored for improved outcomes.

The team also hopes to inspire **other countries' tourism agencies** to develop similar apps that democratize data analyses of their tourism data, which could in turn better optimise marketing spend.

SMU SINGAPORE MANAGEMENT UNIVERSITY
ARIANA TAN RUI MIN | MAGDALENE CHAN JING WEI | MICHAEL KEVIN DJOHAN