

Tanzania Tourism Expenditure

**Presented by: Ricarda Albers, Jaad
Bishti, Angelina Neunzig,
Qurratulain Khaleeq**



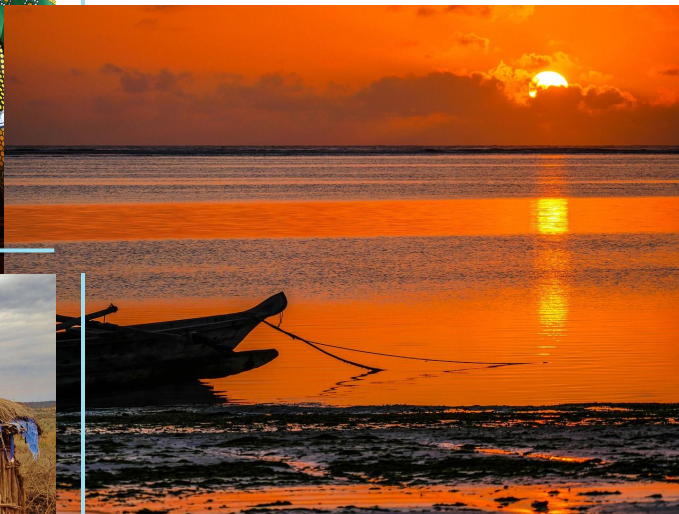
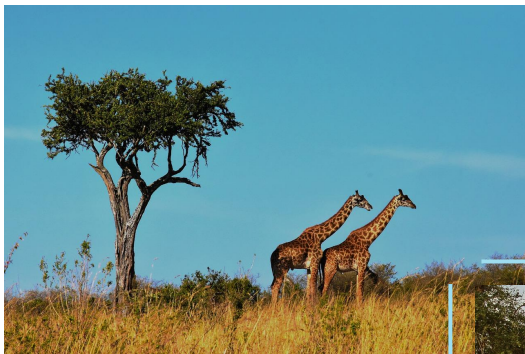


TABLE OF CONTENTS

01

**Business
understanding**

02

**Exploratory Data
Analysis**

03

**Predictive model of
expenditure for
tourists in Tanzania**

04

**Conclusion and
recommendation**

05

Future work




01

Business Understanding





Business Case

- The Tanzanian Tourism Board would like to establish a service on its website that tourists can use to calculate their expenses on site before they start their trip.
 - On the one hand, this enables customer acquisition/retention and on the other hand, it generates relevant data about customers.
 - Future data could be used to cluster target groups for advertising or improving services for tourists in Tanzania.
-
- 

Business Goal

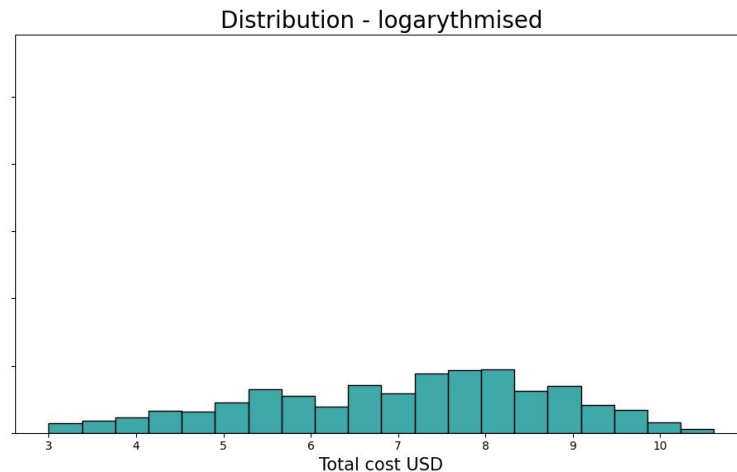
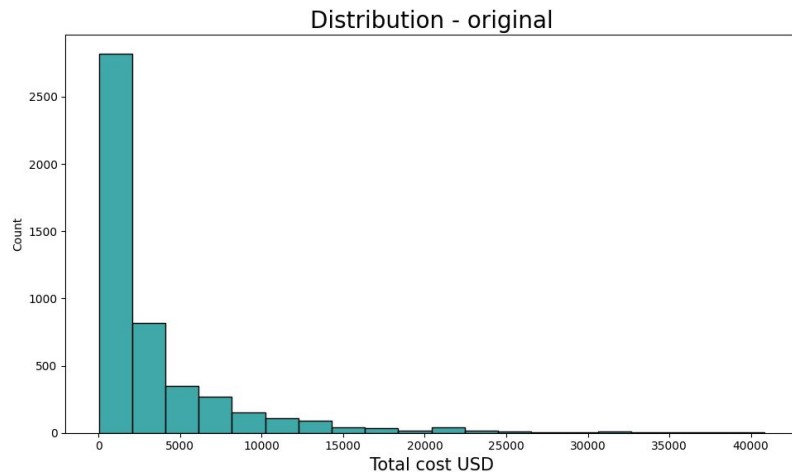
- Using machine learning and previous data on tourist spending in Tanzania to create a predictive model for estimating a tourist's expenses accurately.
- The model considers various factors like trip duration, the number of travelers, and activities to predict total spending. This helps tourists plan their budget effectively before their Tanzania visit.
- This predictive tool benefits both tourists and the tourism industry by providing personalized expenditure estimates and improving the overall travel experience in Tanzania

02

Exploratory Data Analysis



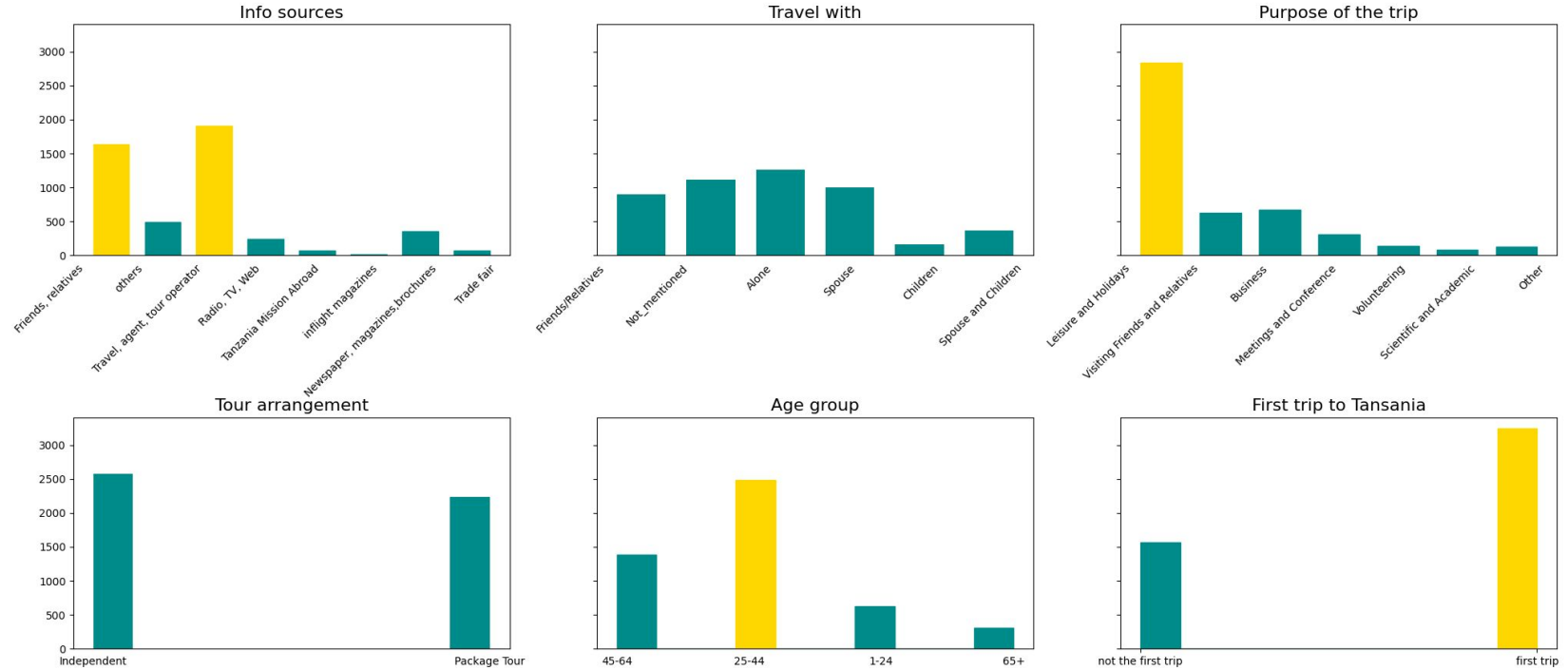
Frequency distribution of the target



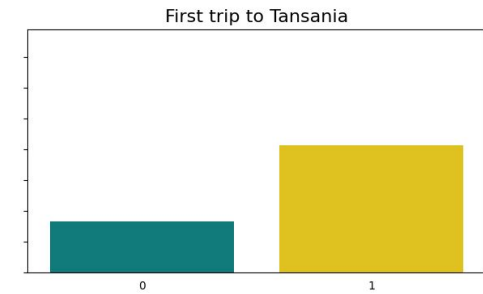
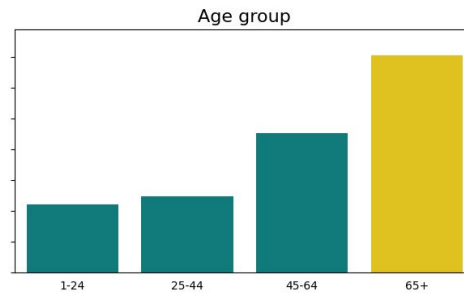
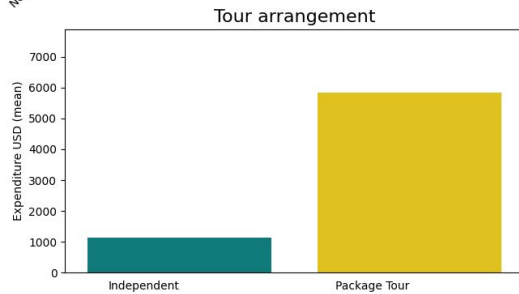
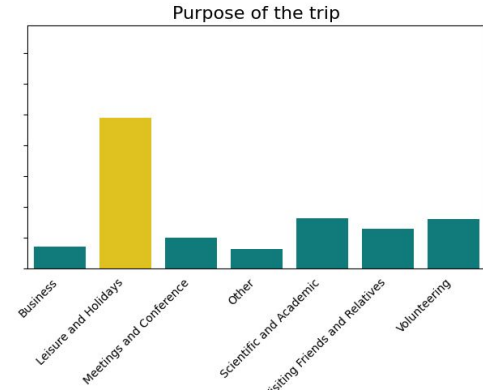
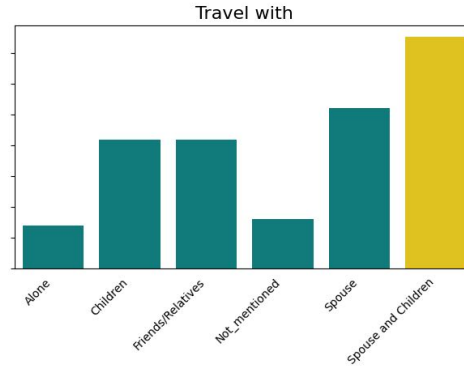
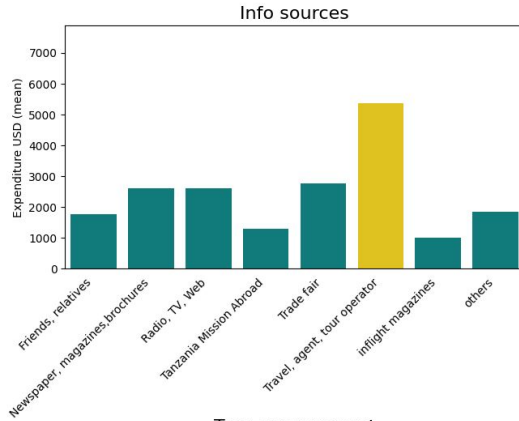
- For the target (total cost in USD), the data set shows a clear right skew.
- In addition, it can be seen that there are very few high values.
- For this reason, the data for the target were logarithmized. Now a distribution close to a normal distribution is shown, which is more suitable for modelling. Finally, this calculation is reversed to allow our users to calculate their local travel costs in advance.



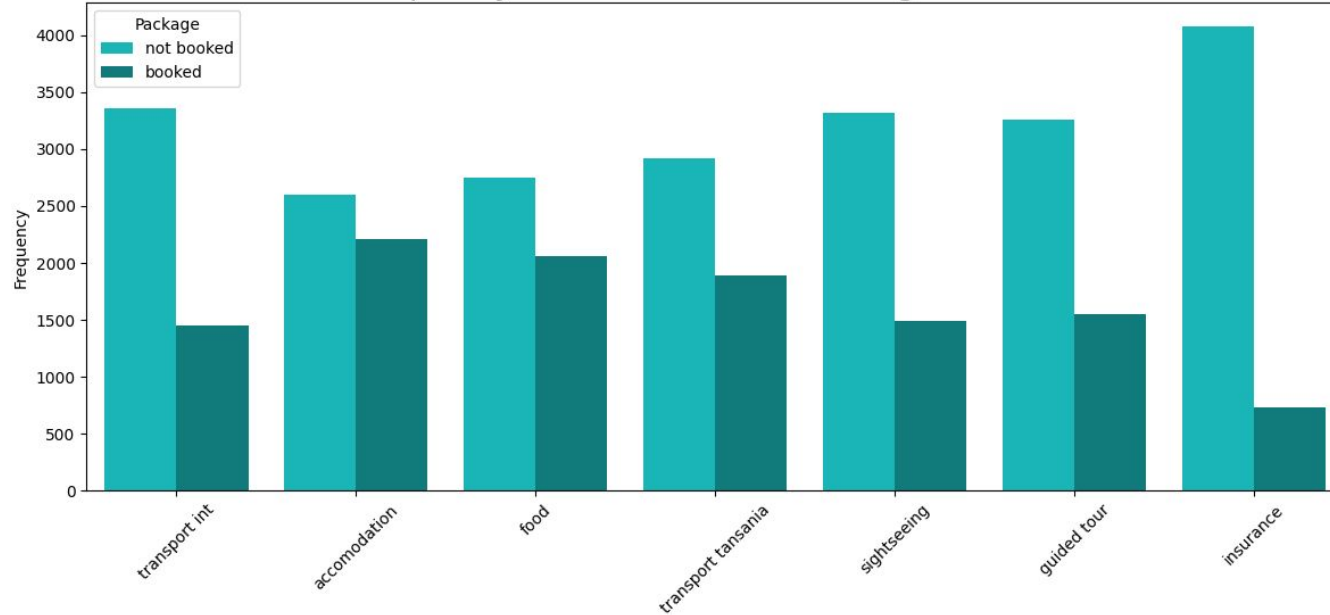
Distribution of Categorical Features



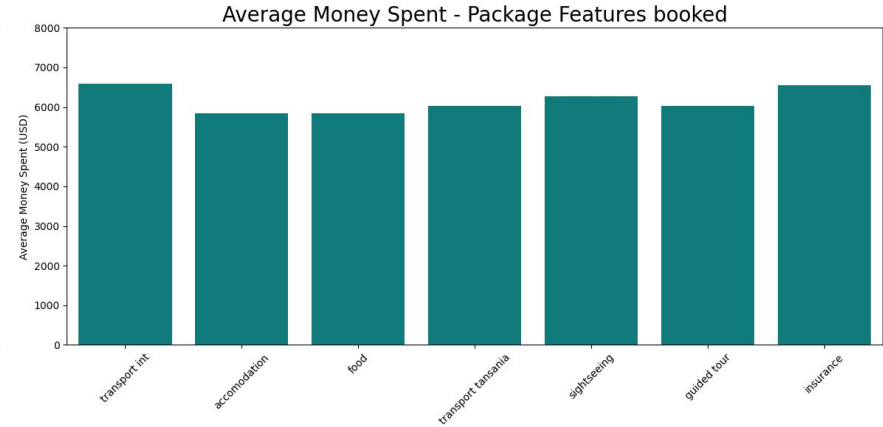
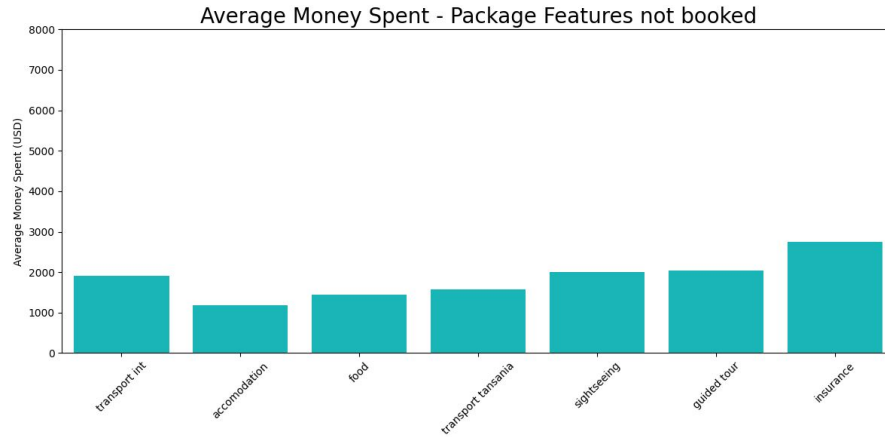
Mean expenditure Categorical Features



Frequency Distribution of all Package Features



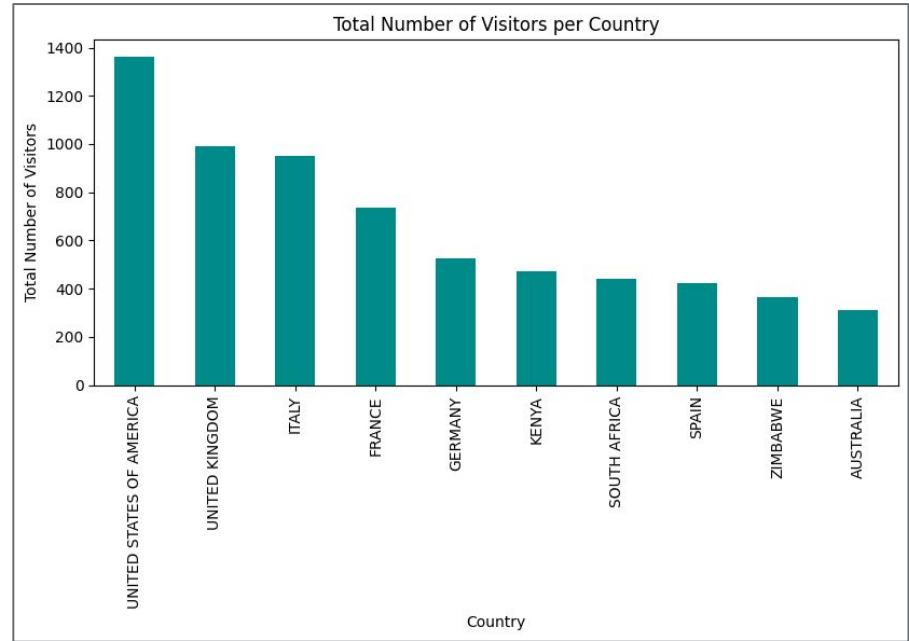
- The frequency distribution of the added packages shows for each package that the majority of travellers do not book the respective package.
- Across all packages, about one third of travellers books packaged tours.



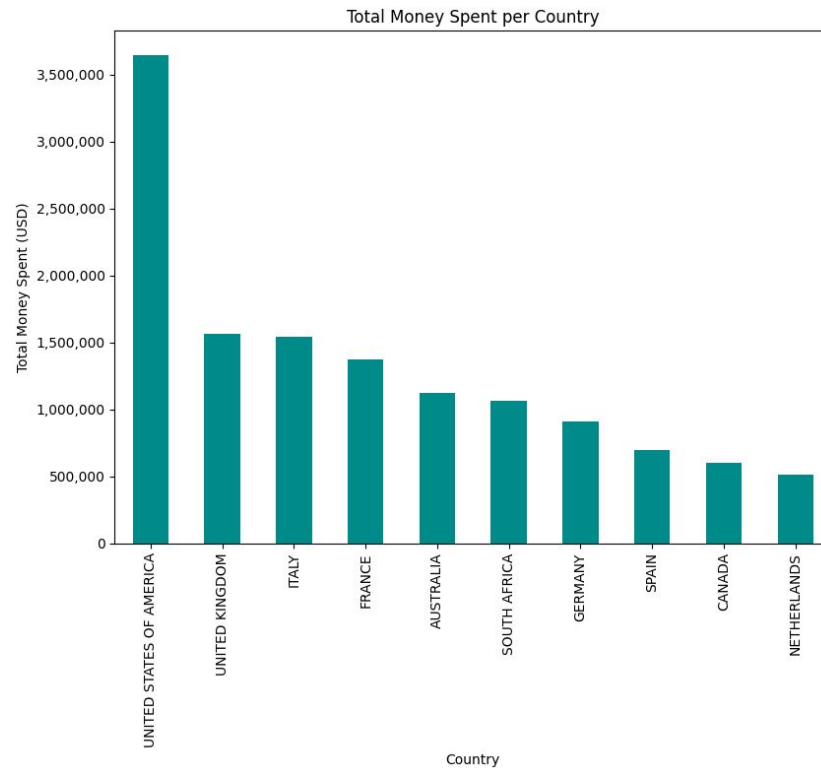
- When calculating the average (mean) USD spent per package booked or not booked, the clear trend is that travellers who book packages spend more money on holidays to Tanzania.



The majority of visitors from Tanzania are from the United States of America (1300 visitors) followed by the UK and Italy

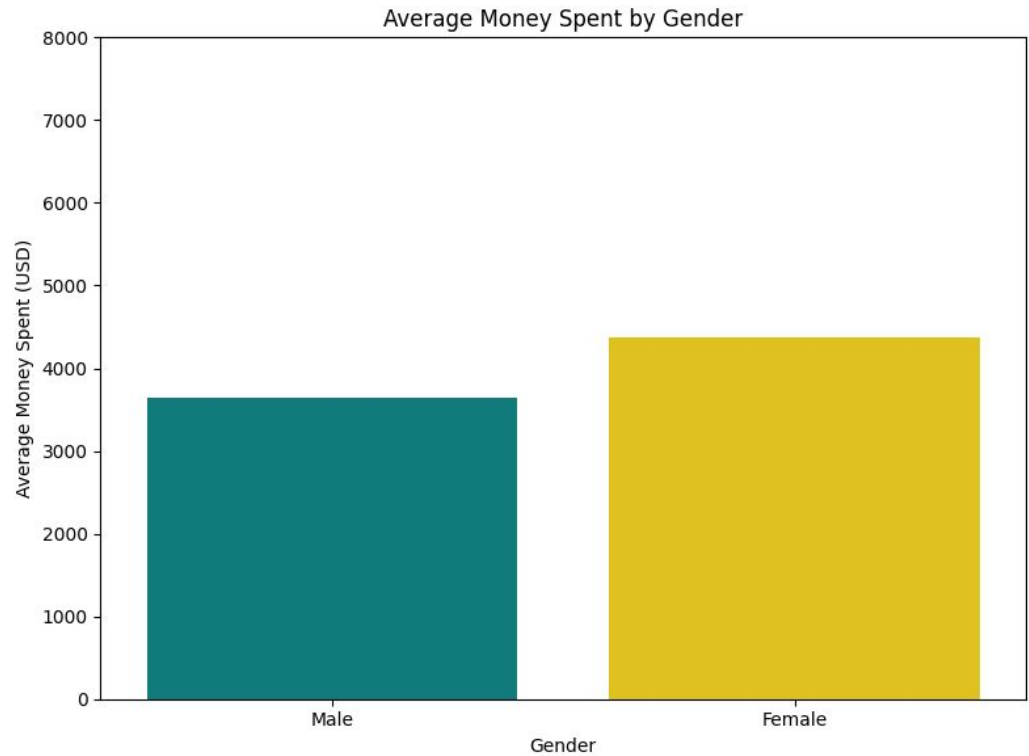


Tourists from the United States of America are the overall highest spenders in Tanzania, spending a total of 3.6 Million USD



More females spend money when travelling in Tanzania than males

- Females: average of 4500 USD
- Males: average of 3600 USD



03

Predictive model of expenditure for tourists in Tanzania



Target variable, machine learning model

- Target variable: amount of money in US Dollars, a tourist will need for his/her trip to Tanzania
- Model: linear regression model

Evaluation Metric

- RMSE (root mean square error): RMSE punishes outliers, it is more sensitive to extreme outliers (than MAE for example); RMSE has the benefit of penalizing large errors more, so because our tourists want to have an accurate predicted amount of money, which they need for their trip to Tanzania

Baseline model

- First, simple model which shall predict the expenditure only with the variables “age groups” (0–24, 25–44, 45–64, 65+)
- Model: linear regression model

	train data set	test data set
RMSE	5,11	4,95
mean expenditure	1130,96 USD	1115,51 USD
slope	age group 25–44: 1,132 age group 45–64: 2,204 Age group 65+: 4,786	
intercept	761,80	

Final model

- Linear regression model with all features of the data set and no additional transformation (23 features resp. 52 after creating dummies)
- Model: linear regression model

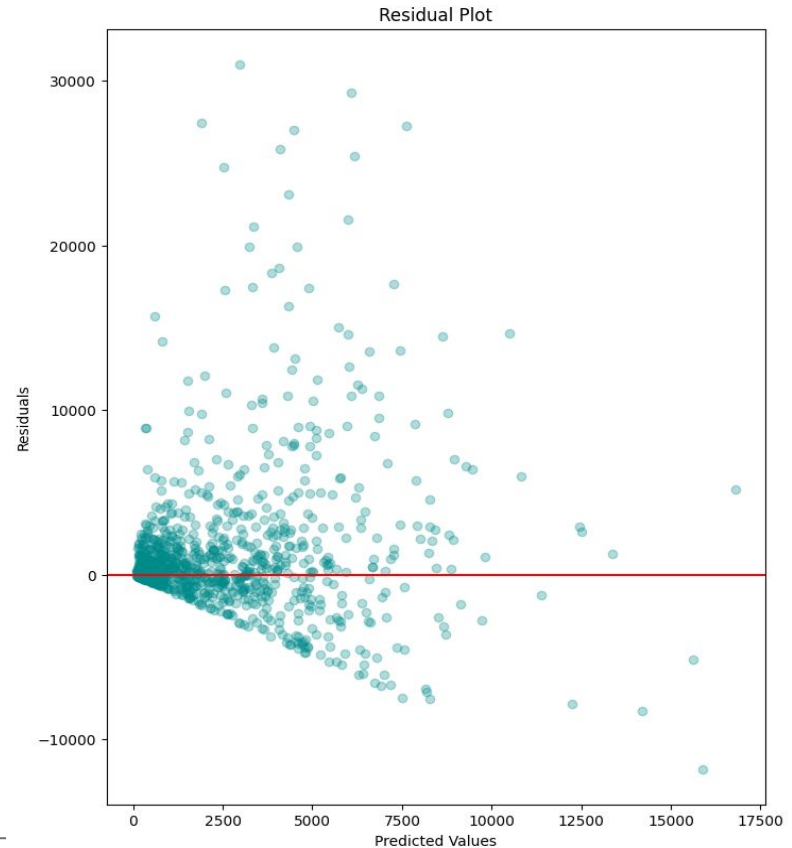
	train data set	test data set
RMSE final	3,14	3,28
RMSE baseline	5,11	4,95
mean expenditure	1130,96 USD	1115,51 USD

Modelling and transforming work in between

- Logarithmic transformation of the target variable
- Data scaling
- Cross validation to evaluate our model
- Regularization by Elastic Net model (cv with ElasticNetCV)
- Decision tree/regression tree as a different model
- Grid search and cross validation on decision tree model

→ nothing of this improved our results compared to a simple linear regression model with all the features of the data set in it

- until the amount of around 5.000 USD the residuals is predicting well
- to interpret the error values which lie above, we would have needed more time and have a closer look at it
- the target variable has a huge spread, which can be one of the reasons, (removing outliers of the target variable didn't change anything)




04

Conclusion and Recommendations






Conclusion

- With our machine learning model the Tanzanian Tourism Corporation gets a model which they could use to establish a service on their website that tourists can use to calculate their expenses before they start their trip.
 - On the one hand, this enables customer acquisition/retention and on the other hand, it generates relevant data about customers.
 - This predictive tool has the potential to benefit both tourists and the tourism industry by providing a personalized estimate of expenditure and enhancing the overall travel experience in Tanzania.
-
- 



Recommendations

- Focus on the age group that is frequently travelling (25–44)
 - People spend more with packages so create interesting affordable packages for the said age group
 - Most people get information from travel agents, create more advertisement specially to target younger travelers
 - Focus on the first time travellers, they are spending more on average
-
- 

05

Future work



Future work

- look more closely to the values which aren't predicted correctly and find out the reasons for that
- more feature engineering i.e. combining expenditure and days or gender
- remove some features and try, if it improves the model



Thank you for your time and attention!
