Kimberly Adams
DSC 680 Spring 2024



# Predicting the Monetary Benefits of Ecotourism

Ecotourism unites regular tourism with nature, local communities, and environmental awareness and conservation. The idea is that nature has intrinsic value and can also be a vital part of a local economy through its preservation. The easiest example of this is the concept of a safari. Visitors go to African countries to see the charismatic wildlife (e.g., lions, elephants, cheetahs). If this wildlife, and the habitats these and many other species depend on, is not preserved, then the local economy loses out on revenue when tourists no longer come to see the sights.

Ecotourism promotes the protection of wildlife and wild places by providing alternative sources of revenue for locals and the countries that support it. In another simple example, a poacher can kill an elephant for its tusks and get a one-time payout, but this is not sustainable as the supply of adult, tusk-bearing elephants quickly dwindles. However, allowing the elephant to survive in its natural habitat can provide income for the local guides and other businesses supporting tourists coming to see the elephant for many years, especially since an elephant can live on average around 56 years in the wild.

So now that we have explored some of the benefits of ecotourism, is there a way to predict the value in direct monetary terms? Also, are there certain characteristics of the tourists themselves that correlate with increased revenue for the recipient country? Being able to predict these things would give countries greater incentive to provide organization and possibly infrastructure to facilitate visitors coming to the country.

## Data

To explore these ideas, this project uses a dataset on Tanzanian tourism from Kaggle (https://www.kaggle.com/datasets/alfredkondoro/tanzania-tourism-prediction-zindi-africa?select=Train.csv). The dataset has information about 4,809 different tours originating from various countries. It includes numerous descriptive variables about the tour including: tourist composition of the tour (i.e., age, gender, count), main purpose, length, how it was planned, components of the tour package (i.e., food, lodging, transportation, insurance, guides, etc.), payment method, and most importantly, the total cost of the tour.

One limitation in the dataset is that the Kaggle files include both train and test subsets, but the test set does not include the tour cost values. This means that the training data will need to be split into the actual train and test datasets (ideally at an 80-20% split). This will shrink the size of the data going into the model, so it will not have as much to train on or as much to validate the predictions.

Also, because many variables are categorical and not numerical, dummy variables will be needed to be able to include some of them in the model. This also makes the model a little more challenging especially when some of the variables are not straight binary this or that but rather include several options.

## Methods

The first part of the project looked at the makeup of the data through exploratory data analysis. The distributions of the different variables were graphed by themselves to see if there were any lopsided values. Null values were dropped and a few extra variables were constructed including the total travelers on the tour (males + females), the total tour length (nights on the mainland + nights in Zanzibar), and the total number of tour components that

where included in the tour price (international transport + accommodations + food + Tanzania transport + sightseeing + guided tour + insurance). A correlation heatmap was also constructed to see which variables were most correlated with each other and with the total cost.

The second part of the project attempted to build different models to try to predict the total tour cost. Categorical variables were one-hot encoded so they could be included in the models. The first and second models utilized the XGBRegressor model with the first using the default parameters and the second using optimized parameters. A third, multivariate linear regression model was also created to see if the four most highly correlated variables could predict the total tour cost.

## Analysis

*Exploratory Analysis*

The exploratory analysis showed that most tours are originating from the United States followed by the United Kingdom, Italy, France, Zimbabwe, and Germany (figure 1). This shows that tours are coming from all over the world including from within other parts of Africa. The United States is by far the greatest with over 500 tours visiting. This could be due to increased marketing in the US, or greater interest in which case increased marketing there could lead to even more tours.

When looking at the tour age group breakdown, 25-44 year old group was the greatest proportion of the tours followed by the 45-64 year old group (figure 2). This would suggest that marketing to the other age groups (1-24 and 65+) is probably a waste of time as these other groups are unlikely to have the interest, funds, or the health to be able to go.

Although there were multiple purposes for going on a tour, the number one reason was for leisure and holidays (figure 3). This was followed by business and vising friends and relatives. This makes intuitive sense as most folks are unlikely to book a tour for other purposes as often.

The main activity of the tours was predominantly wildlife tourism which fits with our ecotourism theme (figure 4). This was followed by beach tourism which also fits. All other purposes made up much smaller percentages of the tours.

Most tours got their information from a travel agent or tour operator followed by friends and relatives (figure 5). This suggests that if countries can foster partnerships with travel agencies, they are most likely to reach their target audience with relevant information.

There is very little difference between the number tours booked through packages and those booked independently (figure 6). This suggests that despite the previous exploration showing that many tours got information from travel agents, they might not actually book their tours through them. That shouldn't prohibit countries from making information available through them though.

When looking at the package contents of the tours, most tours did not include many components (figures 7-13). The only component that was in more tours than not was accommodations and that was only by a small margin. Food was a close second, but still had more tours without it than with. Travel insurance was the component least included in tour packages.

The length of most tours was on the shorter side, usually three weeks and under (figures 14 and 15). Most did not visit Zanzibar at all or if they did, spent little time on the mainland to

compensate. There were a range of longer tours as long as 145 days on the mainland and up to 61 days on Zanzibar, but thee tours were much rarer.

Most tours were paid for in cash, with a few paid for by credit card (figure 16). Other forms of payment were negligible. This could possibly be due to limited infrastructure and acceptance of any form of payment outside of cash locally. Most tours cost less than 20,000,000 TZS with some reaching almost up to 100,000,000 TZS (figure 18). These high values in the local currency suggest high inflation rates within Tanzania.

Most tours were the participants first visits to Tanzania, though a surprising chunk were repeat visitors (figure 17). This could indicate great satisfaction with the tour and great word of mouth for future tours from friends and neighbors of those visitors.

*Modeling*

Unfortunately none of the models attempted were very good at predicting the total cost of the tour. The default XGBRegressor model had a mean absolute error of 196,843,672 (figure 20). The optimized version got that down to 24,863,447 (figure 21) so significantly better than the default, but still pretty high.

The multivariate linear regression utilized the number of tour components included in the tour package, total number of travelers, total females on the tour, and if this was the tour's first visit to Tanzania as inputs based on the higher correlation shown in the correlation heat map (figure 19). The highest correlation was between the number of tour components and total cost with a correlation score of 0.47. All other correlations were low at around 0.27. The linear regression model had a mean absolute error of 6,562,337 (figure 22) which was better than even the optimized XGBRegressor model, but still high compared to the mean cost of the tours in the database (9,732,343).

## Conclusion

Overall, this project aimed to help local communities by showing that there are employment opportunities and economic benefits gained by protecting natural resources. Although this project's models were not optimal at predicting the potential costs of incoming tours, it still provided insight into the characteristics of the tours and perhaps some opportunity to focus marketing efforts. The models, particularly the linear regression, might also have potential for further refinement to make them more accurate, such as perhaps focusing on tours from a single country or purpose. As it stands, it currently can provide some very conservative estimates for tour values.

The potential monetary infusion from tourism provides further leverage to support conservation initiatives, local environmental education, and responsible infrastructure improvements. Although the particular dataset in this study is focused on an African country, the same benefits of responsible tourism apply even to your local natural areas. It's one of the reasons our National Parks are so popular here in the United States. Not only do they inspire visitors with their often breathtaking views, but they also provide an economic boost to the local communities surrounding them.

## Ethical Considerations

The biggest ethical consideration I can see with this kind of study is that a country might decide to bulk up infrastructure to support tourism in such a way that it negatively impacts the environment the ecotourism is targeting. For example, if a country decides to build a large

airport, the placement of the airport needs to be carefully considered to minimize the effects that habitat loss, increased human presence, and even strange visuals and sounds have on the local wildlife. The idea is to coexist in harmony with the wildlife, not supersede it.

# Appendix

## Figures:

*Figure 1. Distribution of countries the tours originated from.*

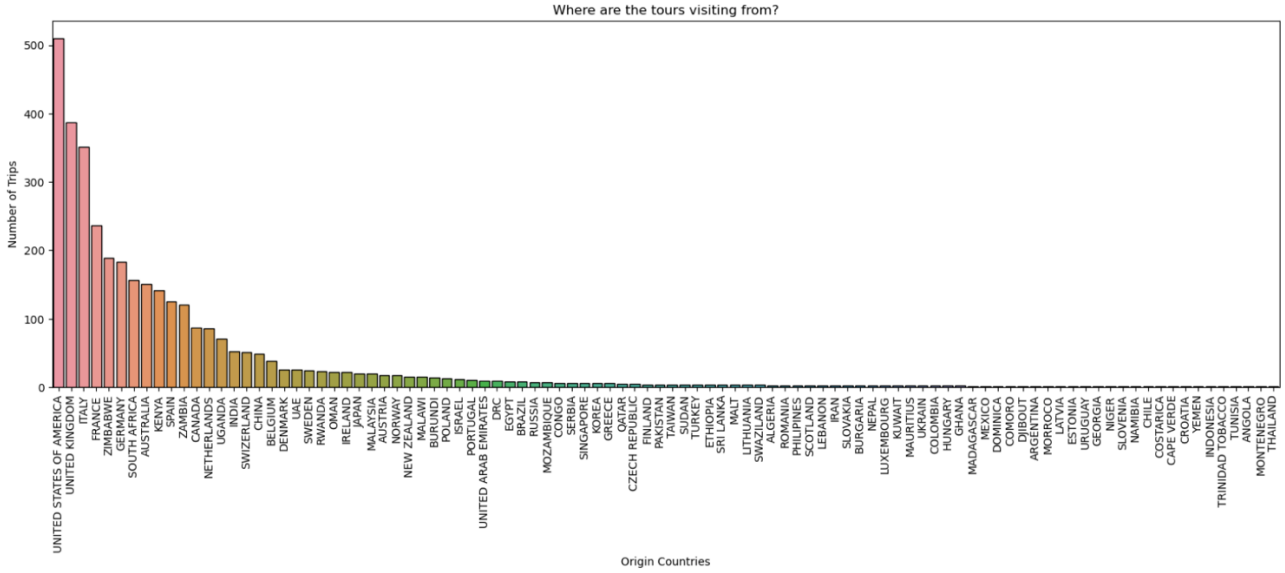



*Figure 2. Distribution of tour age groups.*

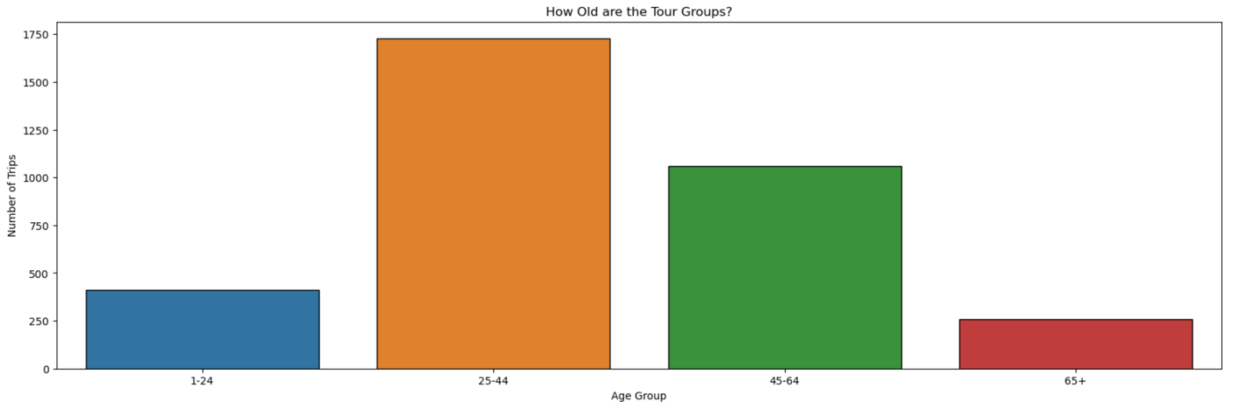


*Figure 3. Distribution of main tour purposes.*

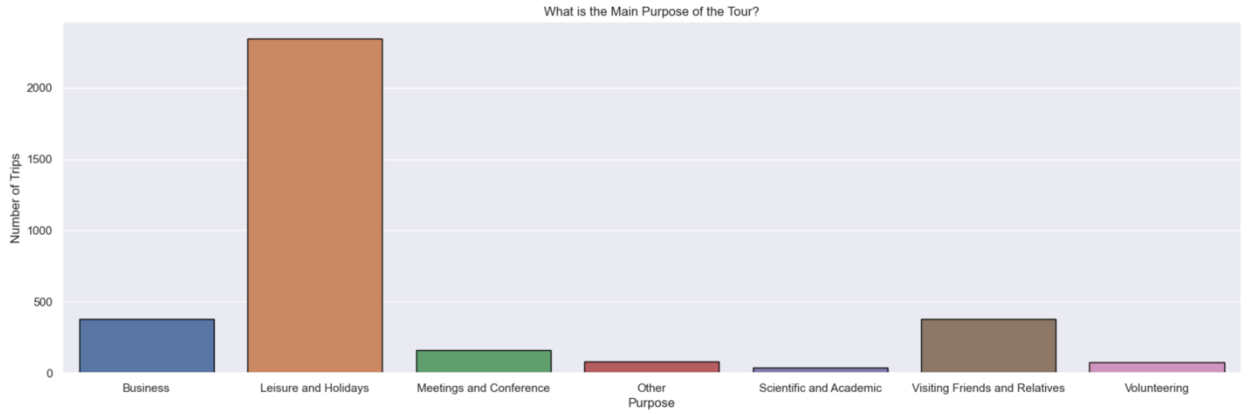



*Figure 4. Distribution main activity of the tour.*

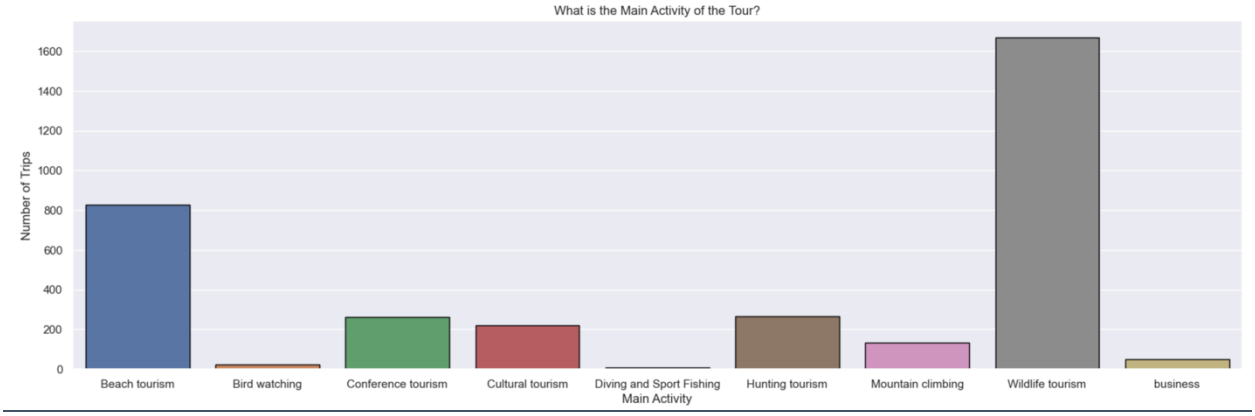



*Figure 5. Distribution tour sources of information.*

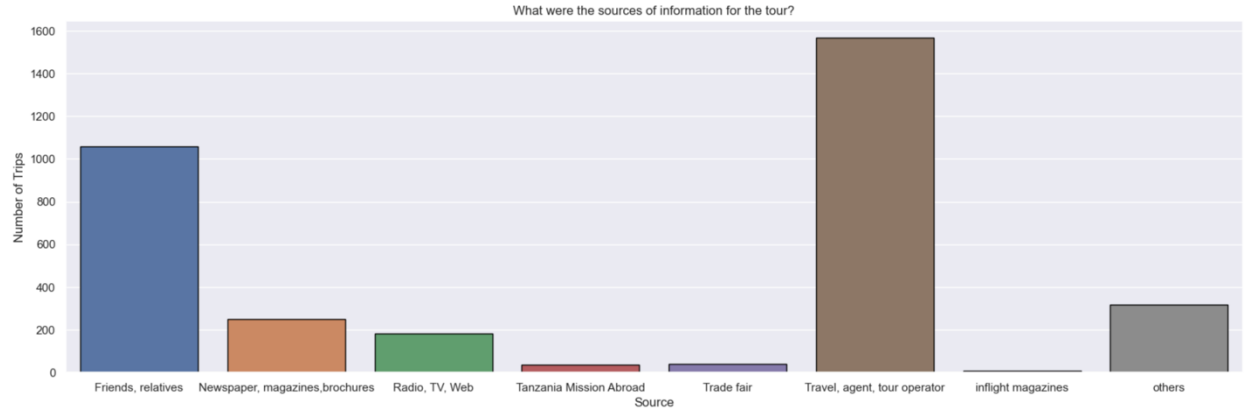


*Figure 6. Distribution of how the tour was arranged.*

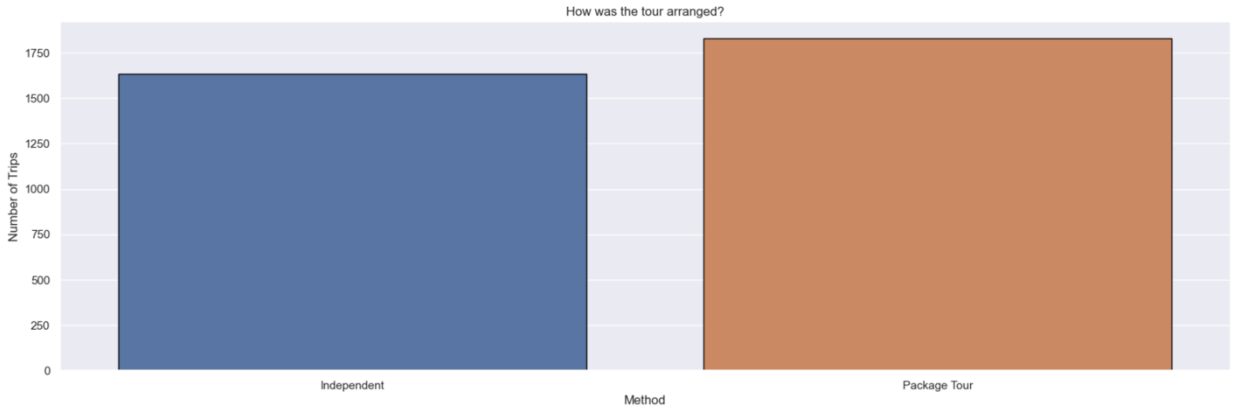



*Figure 7. Was international transport included?*

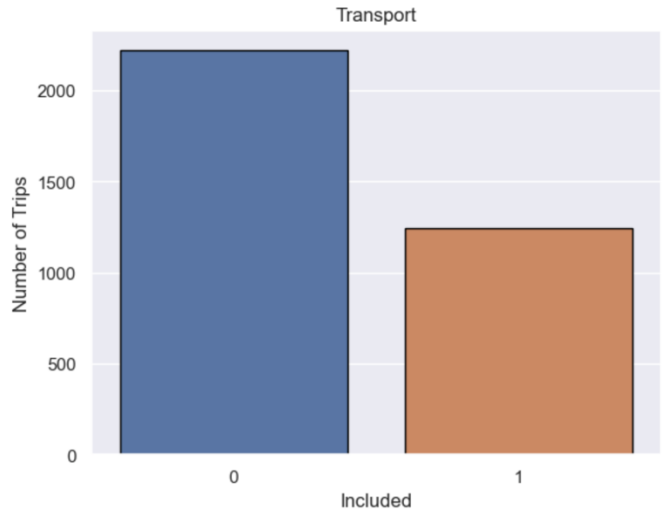



*Figure 8. Was accommodation included?*

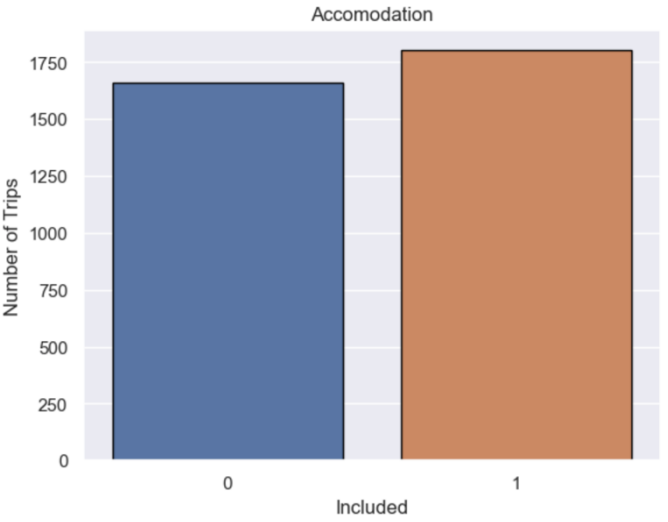


*Figure 9. Was food included?*

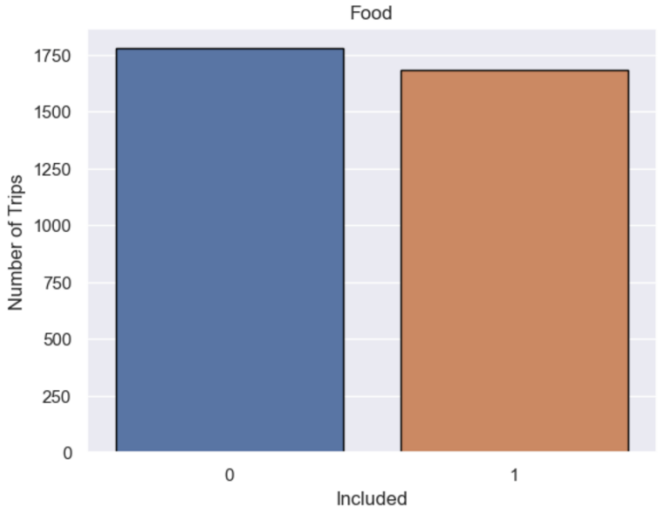



*Figure 10. Was travel within Tanzania included?*

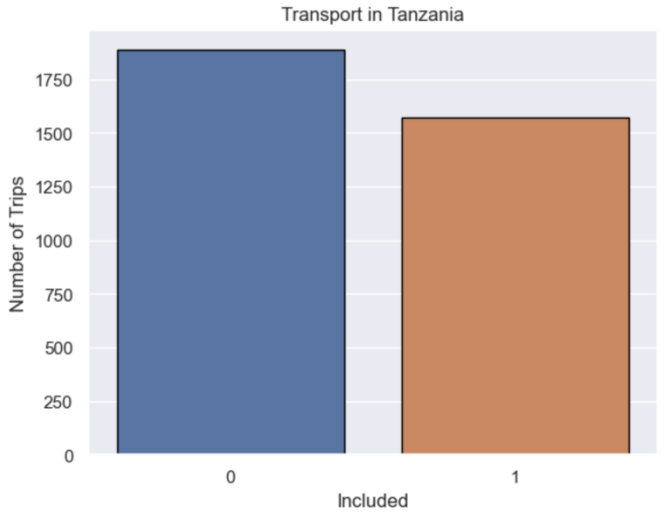


*Figure 11. Was sightseeing included?*

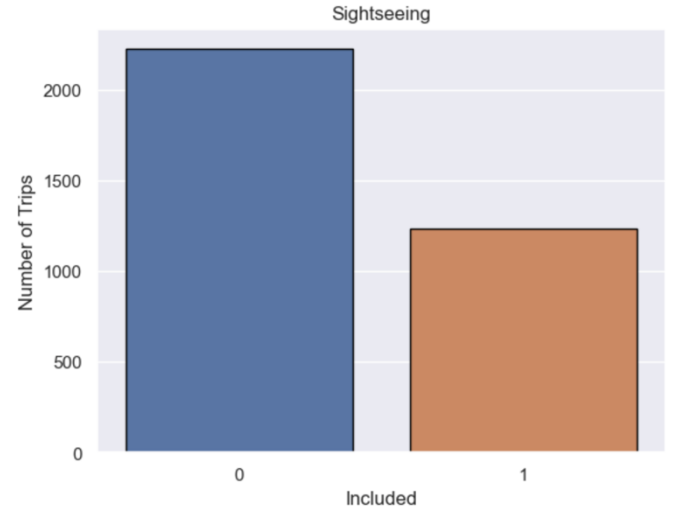



*Figure 12. Was a guided tour included?*

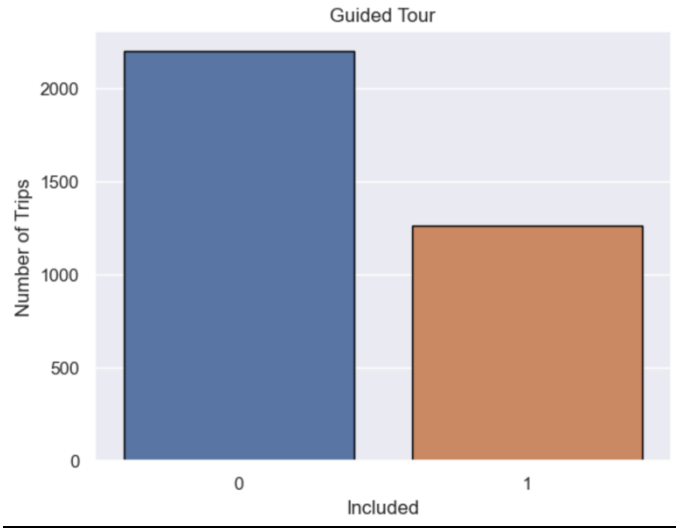


*Figure 13. Was travel insurance included?*

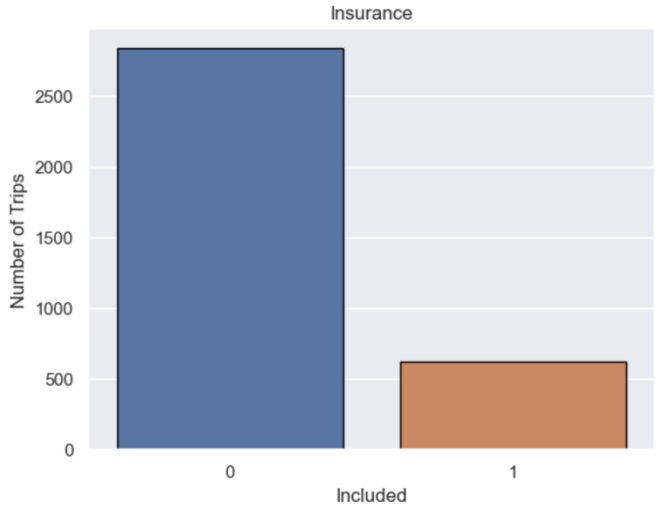



*Figure 14. Distribution of nights the tour had on the mainland.*

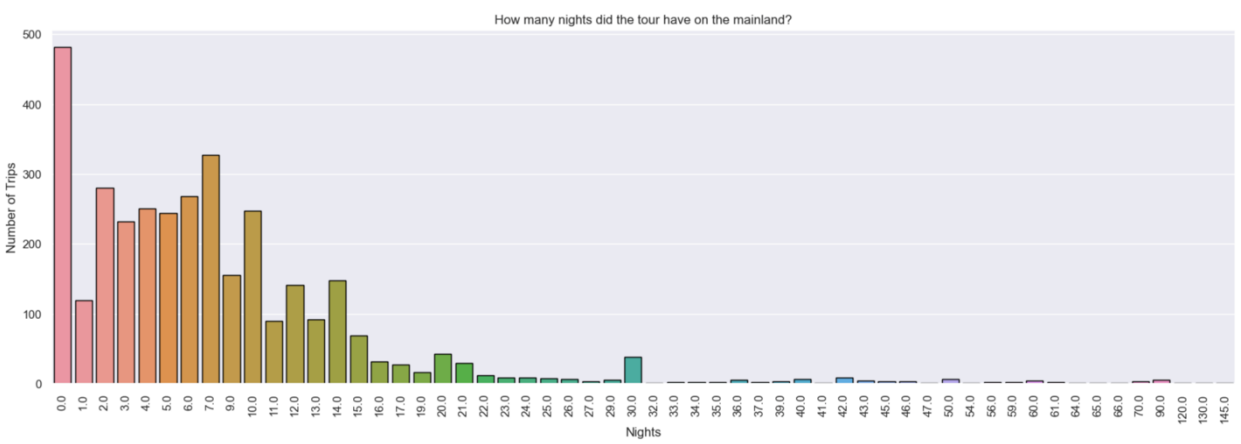



*Figure 15. Distribution of nights the tour had on Zanzibar.*

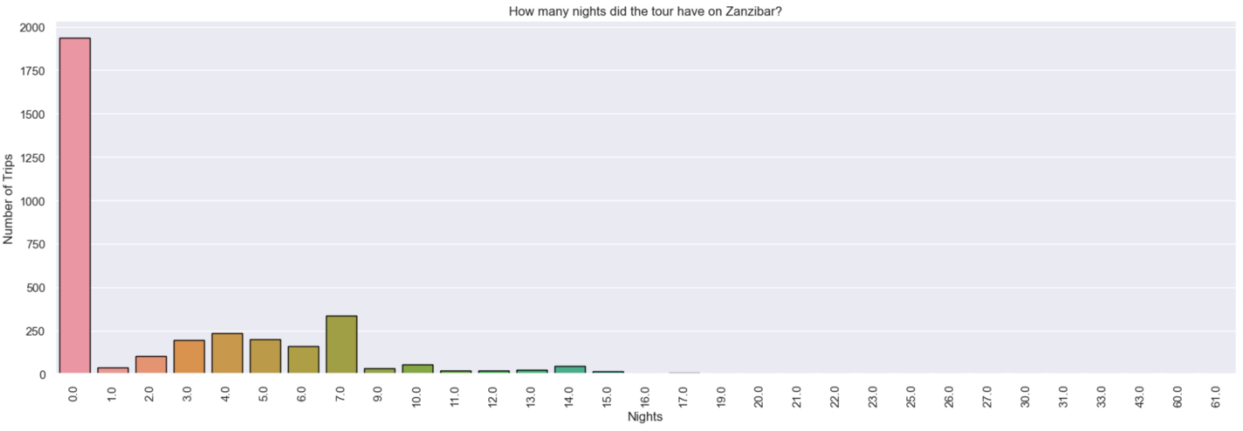


*Figure 16. Distribution payment methods.*

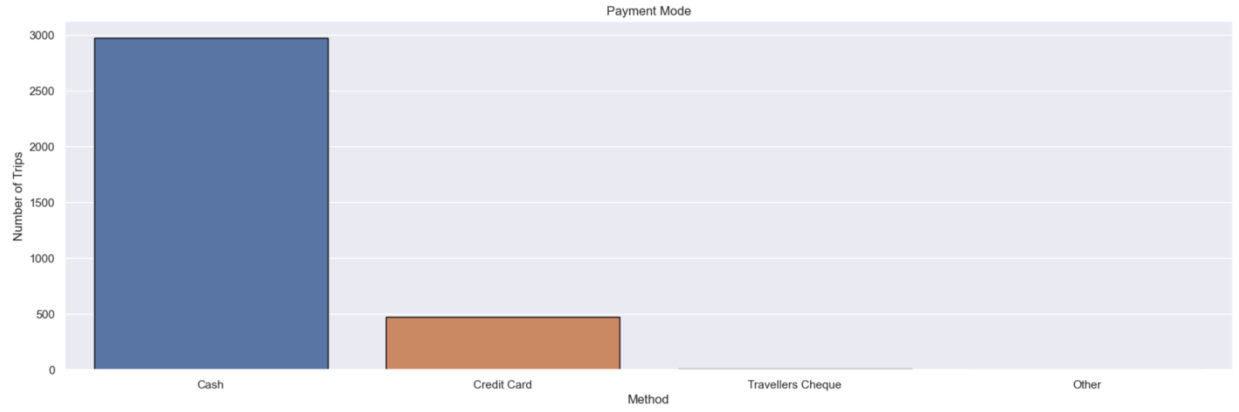



*Figure 17. Distribution of first trippers.*

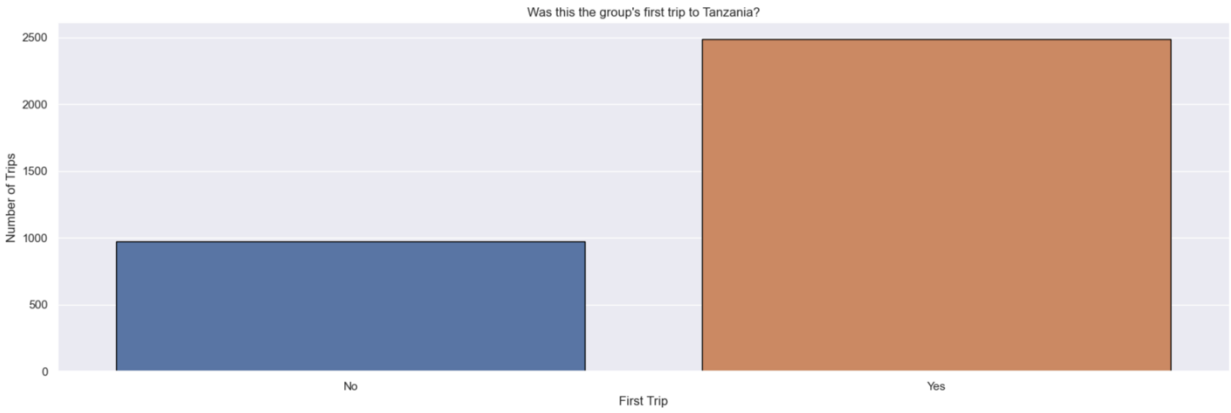



*Figure 18. Distribution of total tour costs.*

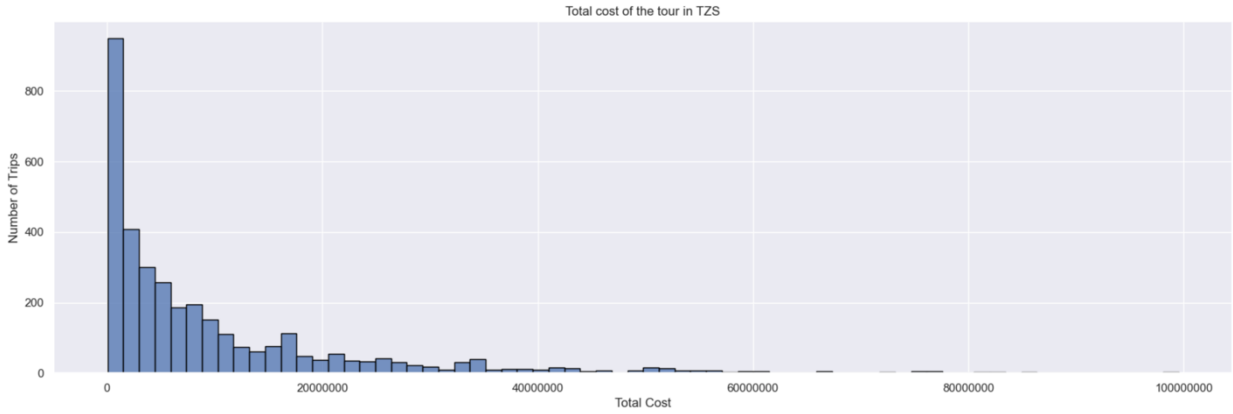


*Figure 19. Correlation heat map.*

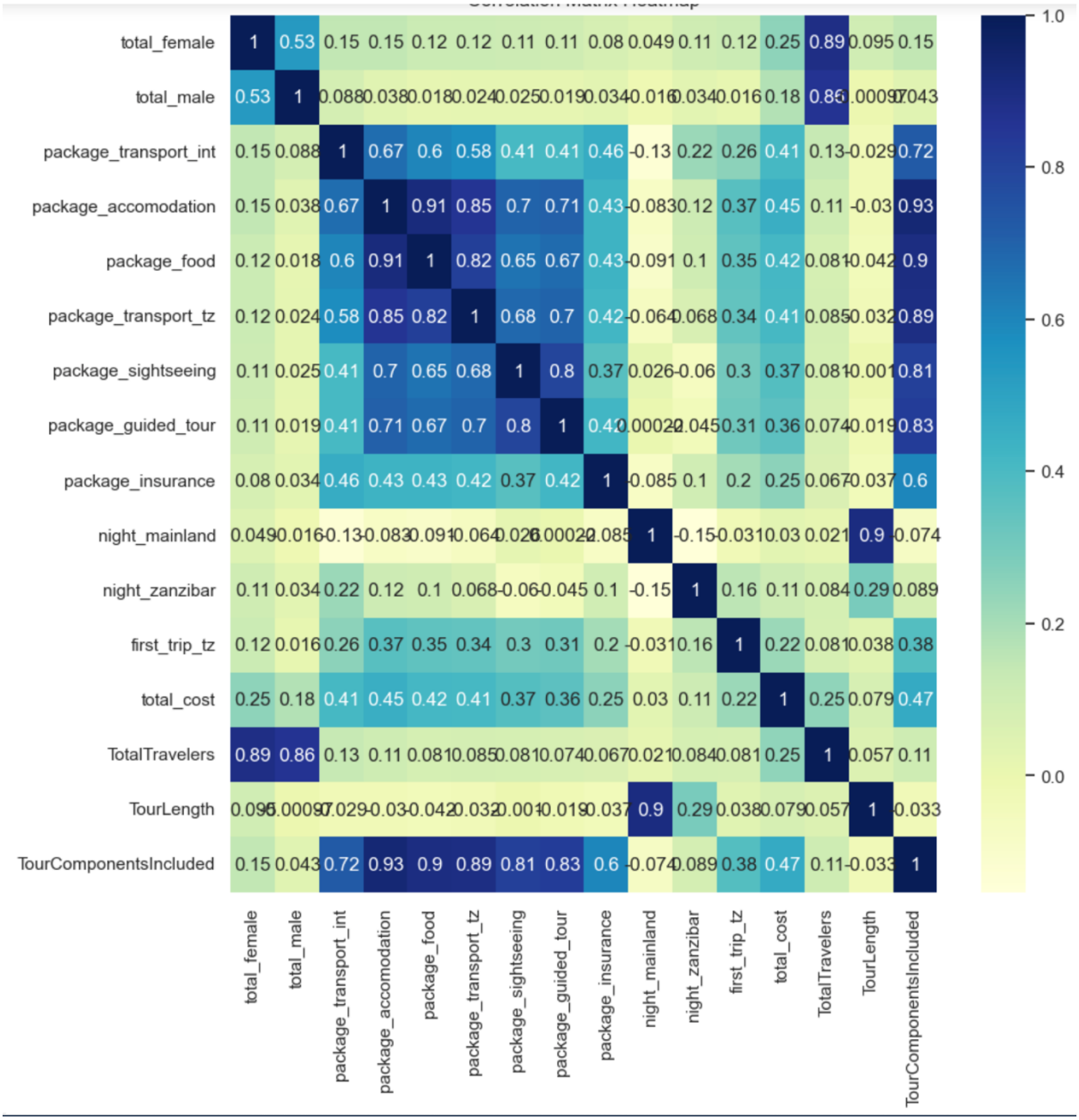


*Figure 20. Actual vs Predicted total tour costs with default XGBRegressor model.*

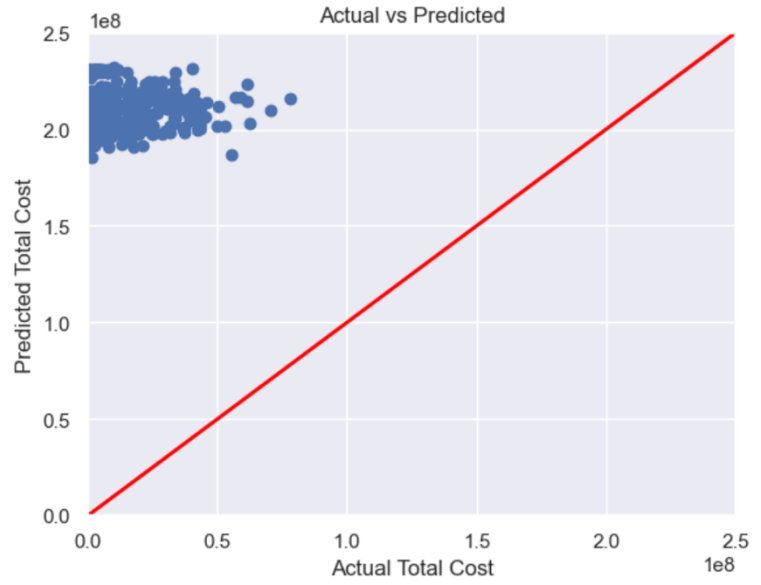



*Figure 21. Actual vs Predicted total tour costs with optimized XGBRegressor model.*

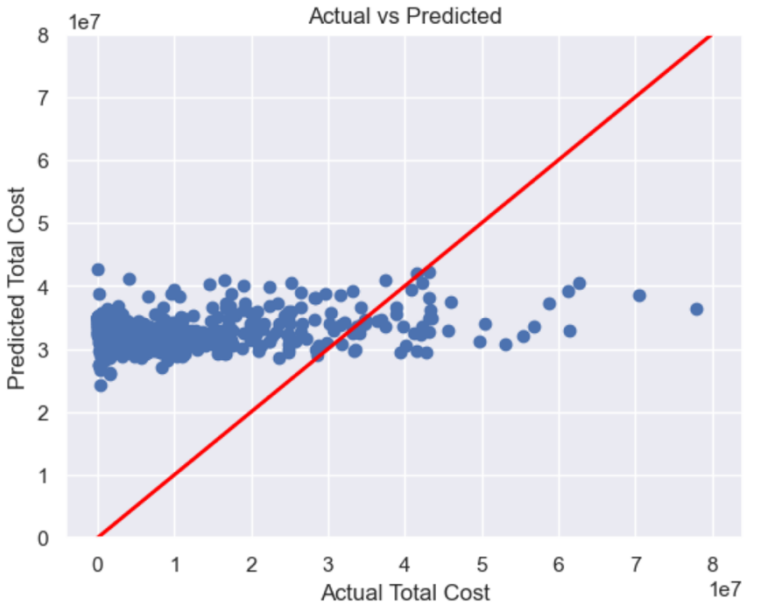
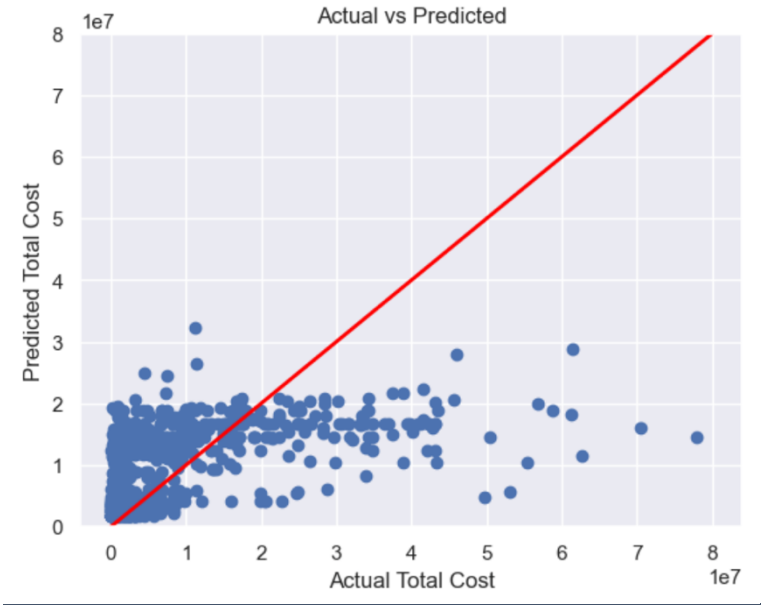


*Figure 22. Actual vs Predicted total tour costs with multivariate linear regression model.*



## Data Dictionary:

| Column Name | Definition |
|---|---|
| **id** | Unique identifier for each tourist |
| **country** | The country a tourist coming from. |
| **age_group** | The age group of a tourist. |
| **travel_with** | The relation of people a tourist travel with to Tanzania |
| **total_female** | Total number of females |
| **total_male** | Total number of males |
| **purpose** | The purpose of visiting Tanzania |
| **main_activity** | The main activity of tourism in Tanzania |
| **infor_source** | The source of information about tourism in Tanzania |
| **tour_arrangement** | The arrangement of visiting Tanzania |
| **package_transport_int** | If the tour package include international transportation service |
| **package_accomodation** | If the tour package include accommodation service |
| **package_food** | If the tour package include food service |
| **package_transport_tz** | If the tour package include transport service within Tanzania |
| **package_sightseeing** | If the tour package include sightseeing service |
| **package_guided_tour** | If the tour package include tour guide |
| **package_insurance** | if the tour package include insurance service |
| **night_mainland** | Number of nights a tourist spent in Tanzania mainland |
| **night_zanzibar** | Number of nights a tourist spent in Zanzibar |
| **payment_mode** | The mode of payment for tourism service |
| **first_trip_tz** | If it was a first trip to Tanzania |
| **most_impressing** | what impressed a toursit in Tanzania |
| **total_cost** | The total tourist expenditure in TZS(currency) |

**References:**

*Elephants live longer in the wild, study shows*. (2008, December 12). African Wildlife Foundation. https://www.awf.org/news/elephants-live-longer-wild-study-shows

Kazmi, H. (2024, June 1). *Regression using XGBoost in Python*. Educative. https://www.educative.io/answers/regression-using-xgboost-in-python#

*Tanzania Tourism Prediction - Zindi*. (2024, February 15). https://www.kaggle.com/datasets/alfredkondoro/tanzania-tourism-prediction-zindi-africa?select=Train.csv

*What is ecotourism - the International Ecotourism Society*. (2019, January 11). The International Ecotourism Society. https://ecotourism.org/what-is-ecotourism/