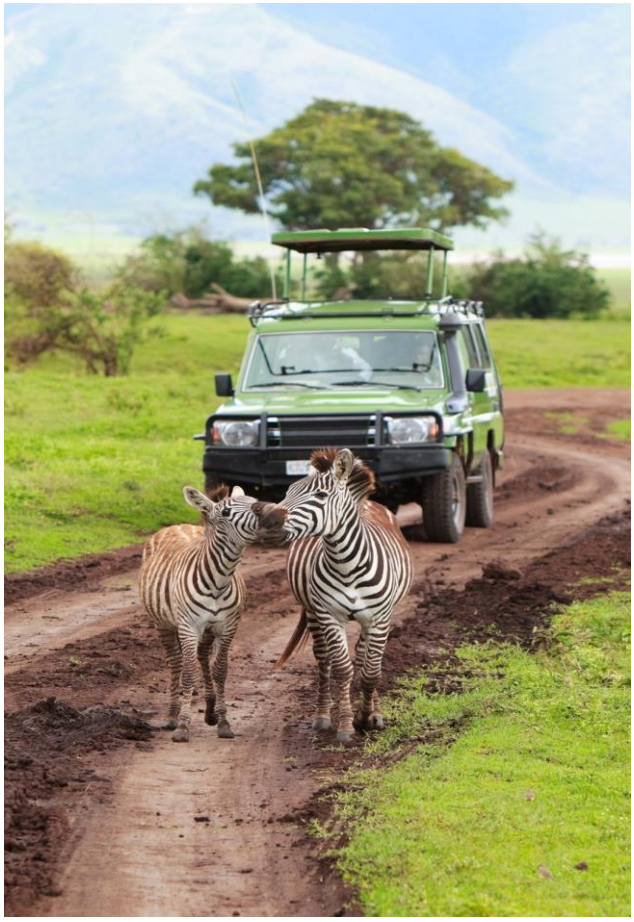


TANZNIA TOURISM SECTOR ANALYSIS

ABSTRACT

Tanzania's tourist industry is important to the country's economy, accounting for around 17 percent of GDP and 25 percent of all foreign exchange earnings. This analysis is aimed at discovering trends within the tourism sector and developing a model to estimate tour prices

King Oshobugie Eshiebor

Dataline Analytics

INTRODUCTION

Tanzania has almost 38% of its land reserved as protected areas, one of the world's highest percentages. Tanzania boasts 16 national parks and is home to a large variety of animal life. Among the large mammals include the Big five; cheetahs, wildebeest, giraffes, hippopotamuses and various antelopes. Tanzania's most well-known wildlife attractions are located in the northern part of the country and include the Serengeti National Park, Tarangire National Park and Lake Manyara National Park. The Serengeti National Park encompasses the world-famous great migrations of animals. The Serengeti National Park is the most popular park in the country and had the chance to host more than 330,000 visitors in 2012. (Wikipedia, 2022).

Business Overview of the Problem

With the amount of attention focused on the tourism sector in Tanzania, there is always the potential for more. In order to do this, more tourists have to be attracted, more revenue has to be gotten and the attractive sites have to be maintained. In order to achieve the afore mentioned, analysis is required.

From figuring out what makes tourist the most attracted, to what they are most impressed with, to how much they are likely to spend, analysis unveils ways of improving the sector even further.

SOLUTION APPROACH

This study is based on the Tanzania tourism sector dataset gotten from Dataline Analytics and looks at the observable trends and relationships between variables in order to discover how changing parameters for one variable can help improve overall performance of the sector.

This study also aims to build a regression model to help predict how much potential tourists would spend on a tourism trip to Tanzania. This helps tourists plan and prepare for their trips.

Data gotten from Dataline Analytics contains the following variables and their definitions:

Column Name	Definition
id	Unique identifier for each tourist
country	The country a tourist is coming from.
age_group	The age group of a tourist.
travel_with	The relation of people a tourist travels 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_arrangment	The arrangement of visiting Tanzania

package_transport_int	If the tour package includes international transportation service
package_accomodation	If the tour package includes accommodation service
package_food	If the tour package includes food service
package_transport_tz	If the tour package includes transport service within Tanzania
package_sightseeing	If the tour package includes sightseeing service
package_guided_tour	If the tour package includes tour guide
package_insurance	if the tour package includes 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 tourist the most in Tanzania
total_cost	The total tourist expenditure in TZS (currency)

KEY FINDINGS AND INSIGHTS

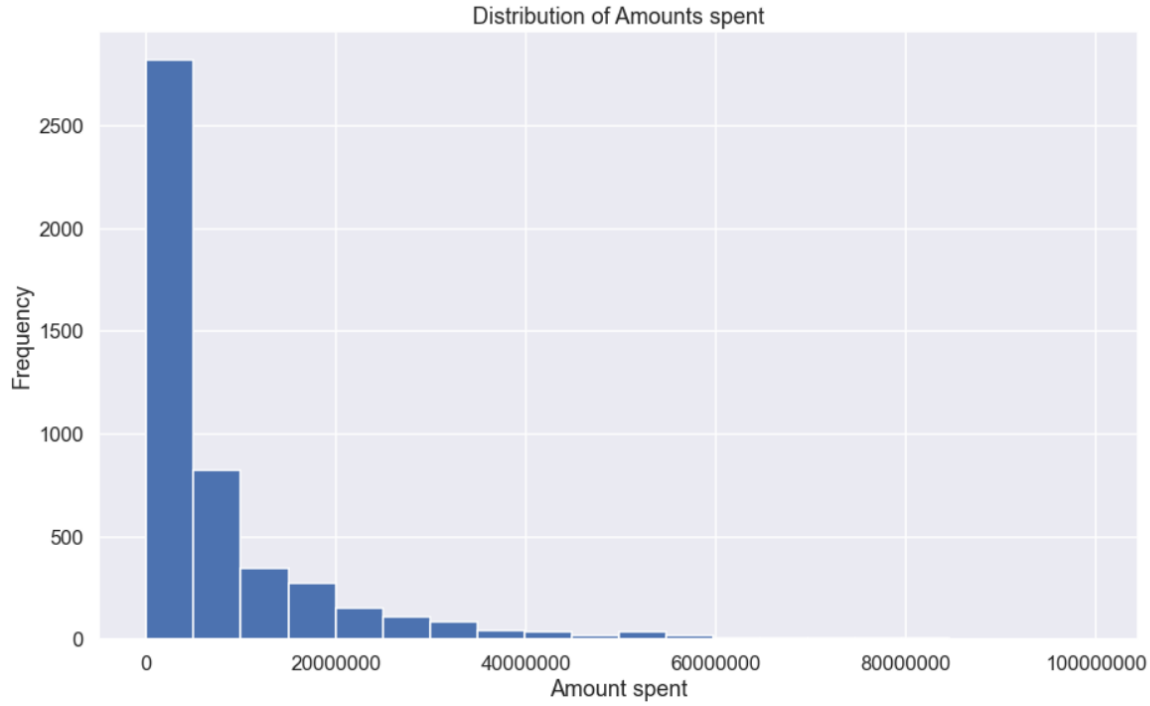
Univariate Exploration

These results have to do with exploring one variable at a time; without observing its relationship to other variables. It is concerned mostly with the count or distribution of values under that particular column.

Beginning with obviously the most interesting variable, total_cost, we see the distribution of values below.

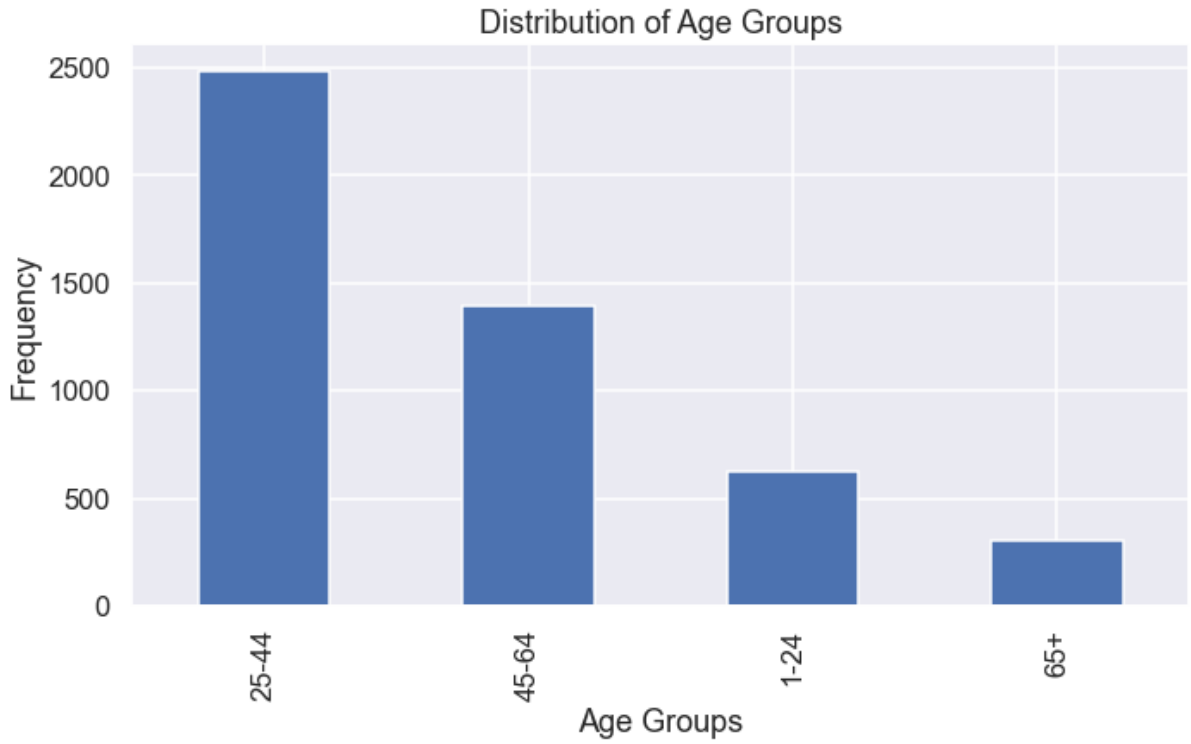
Total Amount Spent

```
The mean amount spent on tours is 8114388.777617801
```



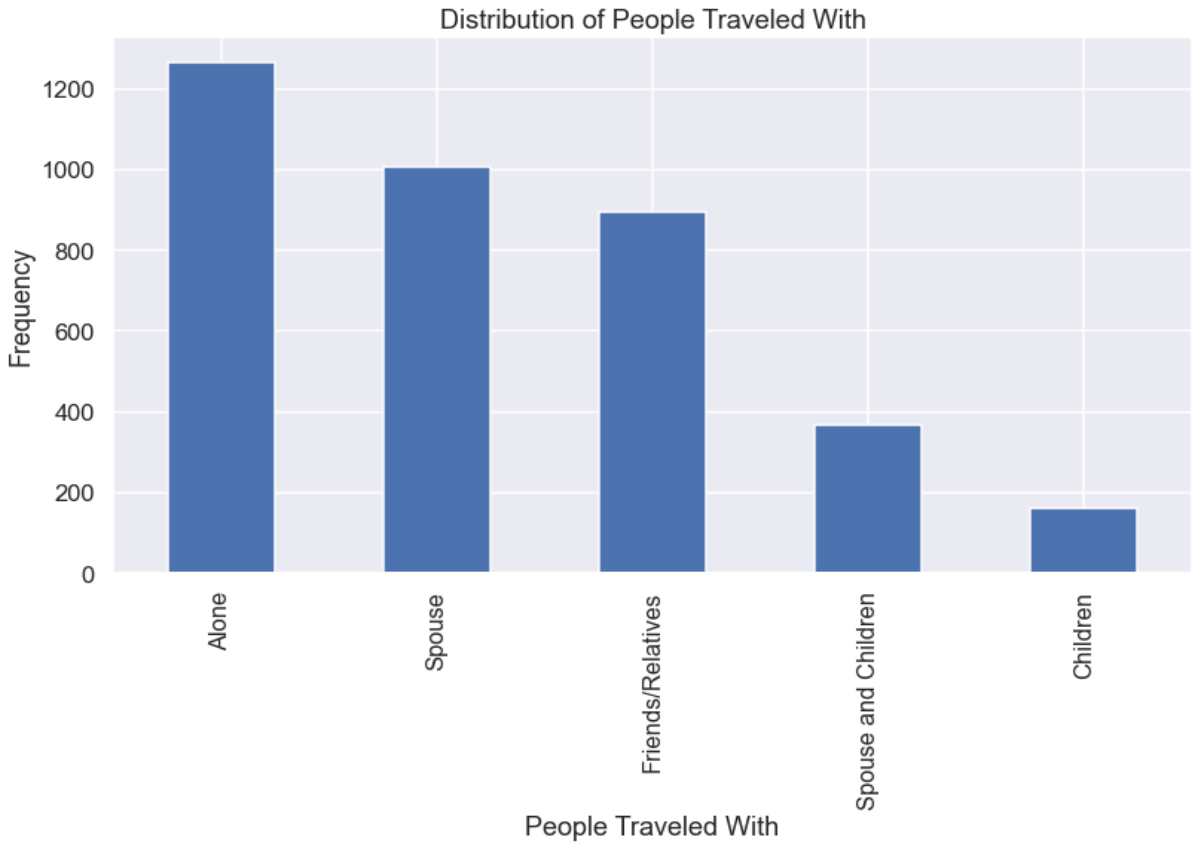
From the plot above, we see that the distribution of total amounts spent (total_cost) is severely right skewed, with the average being just over TSh8.1 Million although this is most likely affected by outliers.

Age Groups



The bar graph above illustrates that those between the ages of 25 and 44 go on the most tours, while those 65 and older go on the fewest. In terms of frequency, those aged 1 to 24 are second to last. This might be due to the fact that most people in this age group do not have a lot of money to spend on tours. However, there isn't enough data in this dataset to make that determination.

People the Most Traveled With



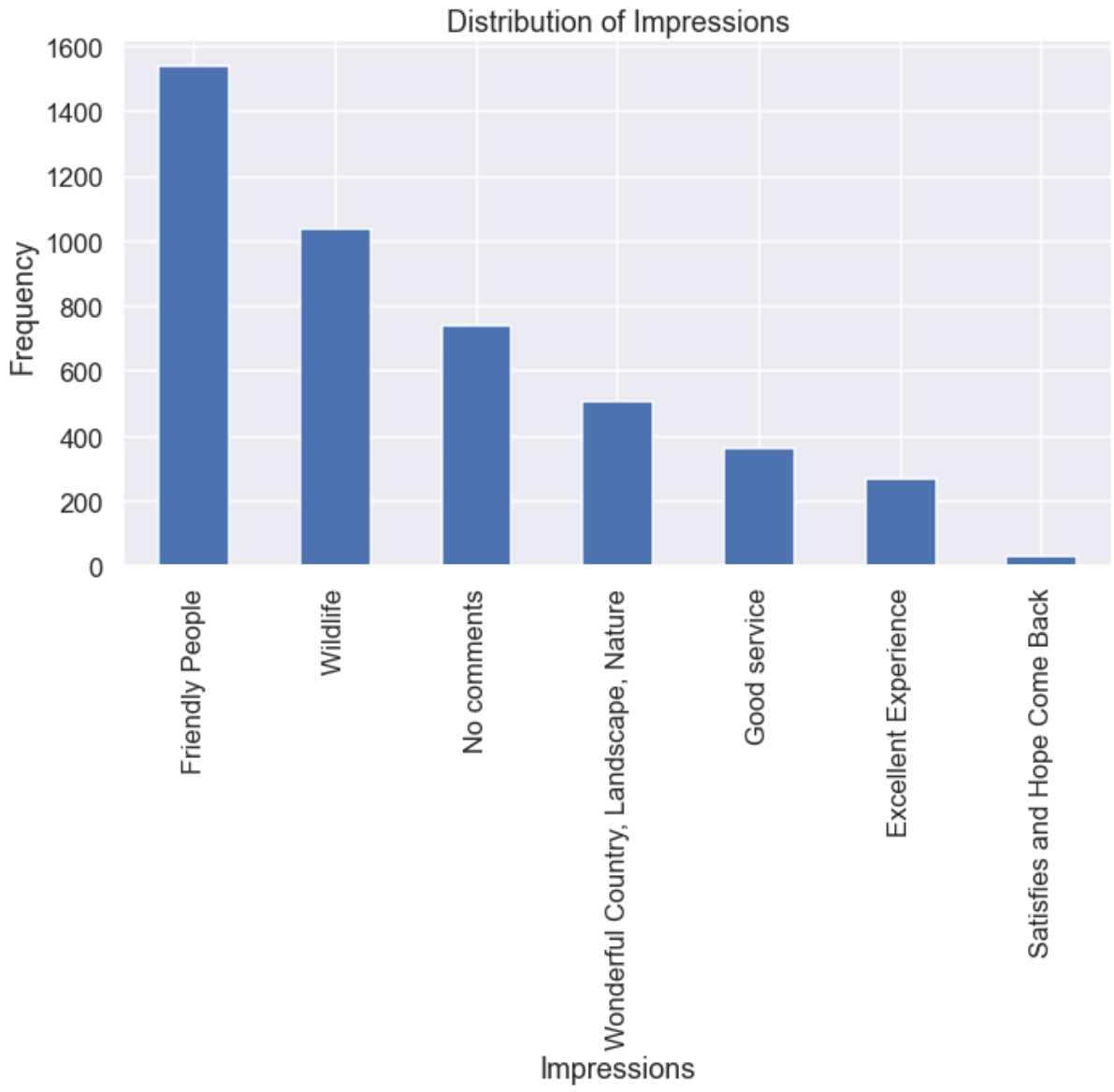
The majority of visitors either go alone or with their spouses on trips. According to the distribution, a single parent is the least likely to bring their children.

Information Sources

The use of travel agencies and tour operators, followed by word of mouth from friends and family, appears to be quite efficient in recruiting travelers. In-flight magazines appear to be the least successful strategy, maybe because visitors travelling into the nation are already aware of the tourist attractions through other methods or have other plans.

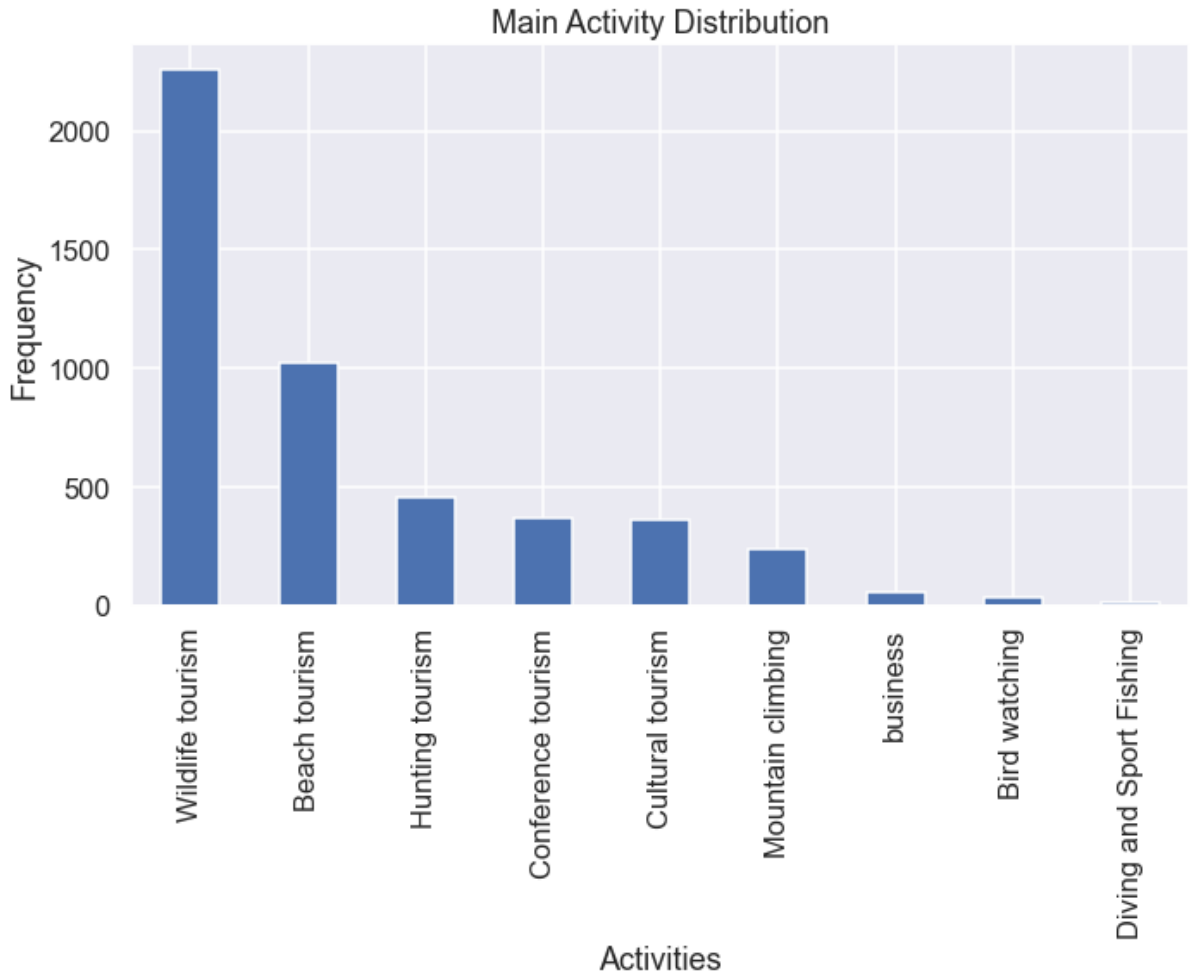
Increased advertising rates on television and, especially, the internet would be a good idea, given these media are used by many more people than newspapers, magazines, and brochures. As a result, it has the ability to attract more individuals.

Most Impressive Aspects of Tours



Tourists are impressed by how nice and inviting Tanzanians appear to be to visitors, which is a fantastic thing. People are also drawn to wildlife, which accounts for a significant portion of the tourist attractions.

Main Activity of Tourism



The graph above supports the previous statement that wildlife attractions are by far the most popular tourist attraction. Beach tourism comes in second, with fewer than half the number of visitors as wildlife tourism. However, diving and sport fishing only draw a tiny number of visitors.

Bivariate Exploration

This entails the observation of relationships between two variables. The main variable of interest is considered to be total_cost. This is because it is what is going to be predicted and it also seems the most significant.

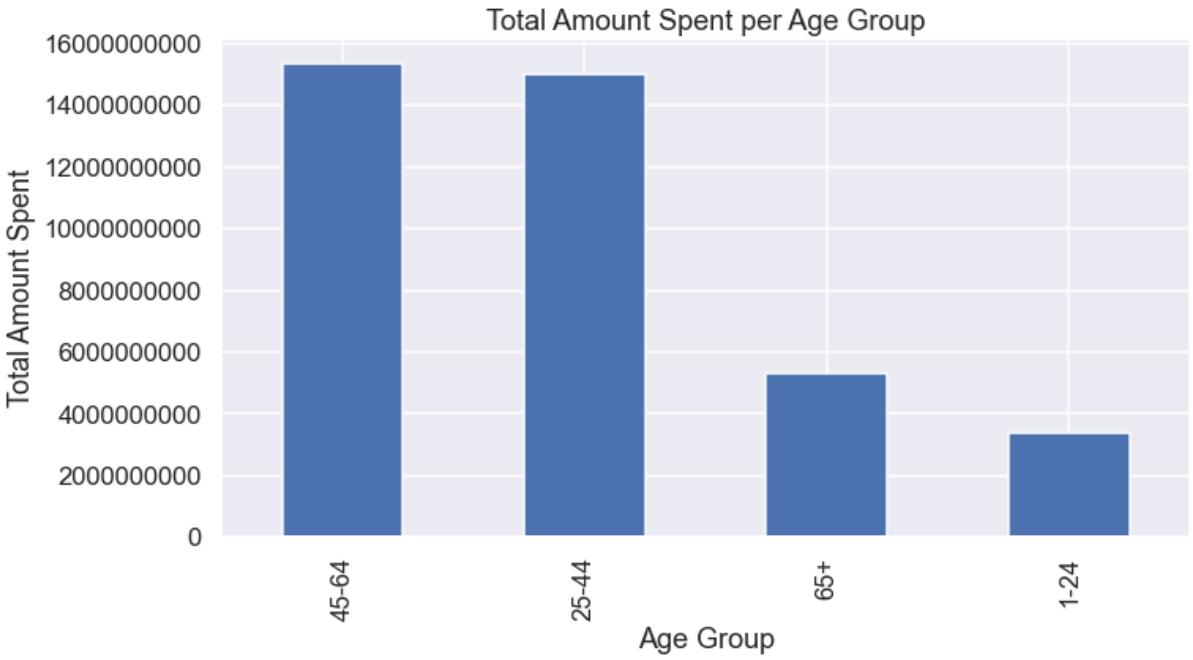
Country vs Total Amount Spent

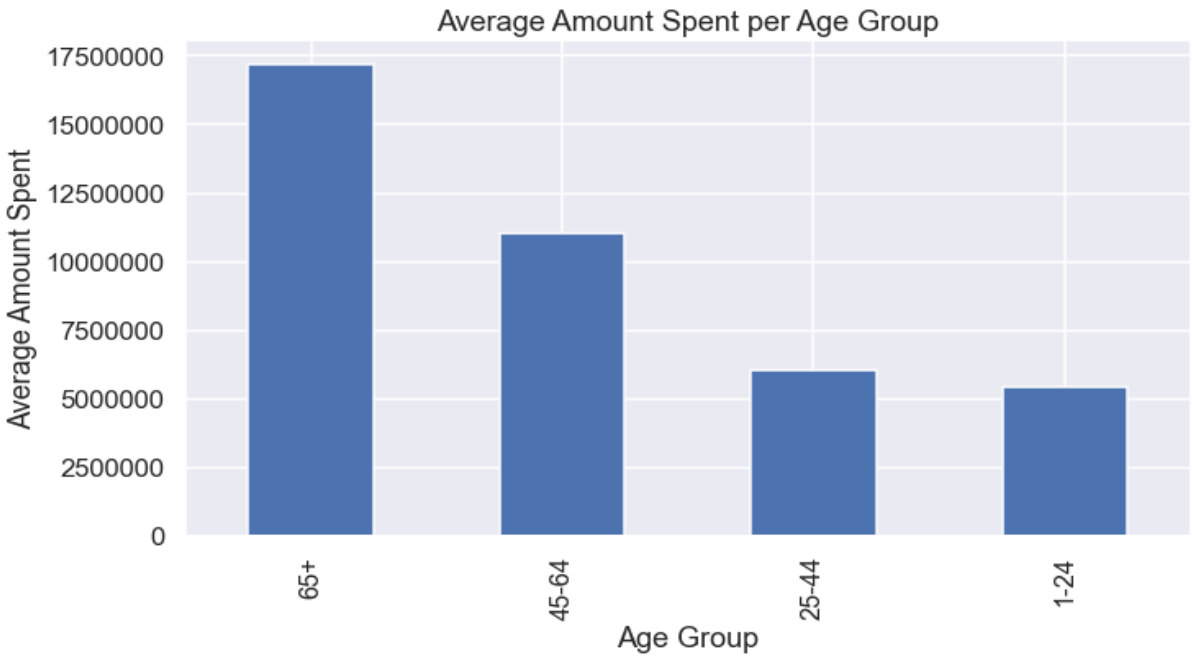
country	total_cost
UNITED STATES OF AMERICA	8.890832e+09
UNITED KINGDOM	3.808383e+09
ITALY	3.762160e+09
FRANCE	3.344496e+09
AUSTRALIA	2.743132e+09

country	total_cost
DOMINICA	3.315000e+07
COSTARICA	2.718300e+07
SLOVENIA	1.906237e+07
TUNISIA	1.574625e+07
AUSTRALIA	1.474802e+07

The table by the left shows the top 5 countries with tourists who have spent **the most on tours in total.** The table on the right shows the countries which spend **the most on average in tours.** Tourists from the United States, the United Kingdom, Italy, France, and Australia have spent the most money in Tanzania, in that order. People from the United States have spent over TSh8.8 billion, while Australians have spent over 2.7 billion. When we look at the greatest average amount spent by countries, Dominica, Costa Rica, Slovenia, Tunisia, and Australia come in first, second, and third, respectively. This means that visitors from the top five nations outnumber those from the last five. Dominicans have spent an average of TSh33.3 million.

Amount Spent vs Age Group

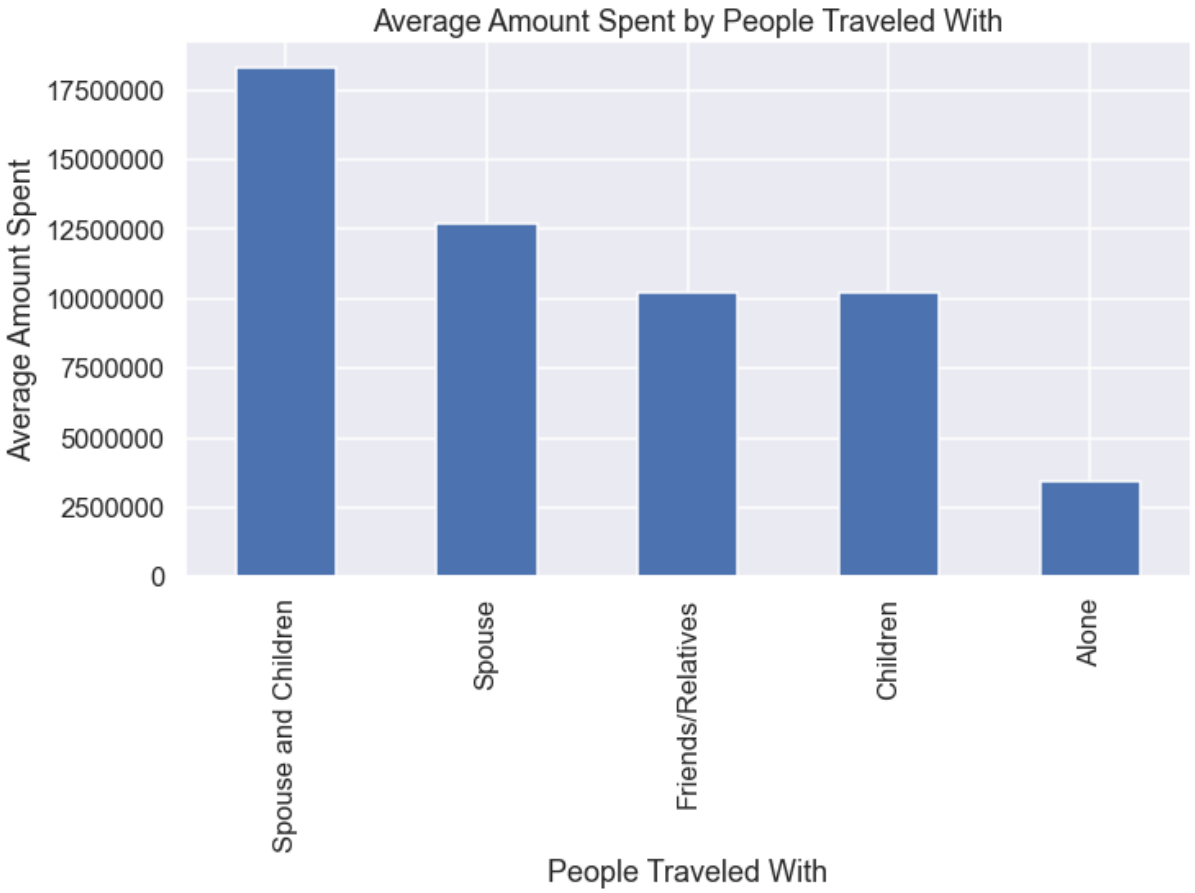




When total cost is compared to nation, the two plots above reveal a gap comparable to when total cost was compared to country. When the total amount spent on excursions by each age group is examined, it can be observed that those aged 25-44 and 45-64 have spent the most. This is simply a repeat of a previous plot that showed that the most common visitors are between the ages of 25 and 44 and 45 and 64.

The second graph demonstrates this by demonstrating that people aged 65 and up spend more money each tour on average. The second figure likewise shows that there is a positive relationship between age range and total amount spent each tour, with the total amount spent increasing as the age range grows.

Amount Spent vs People Traveled With



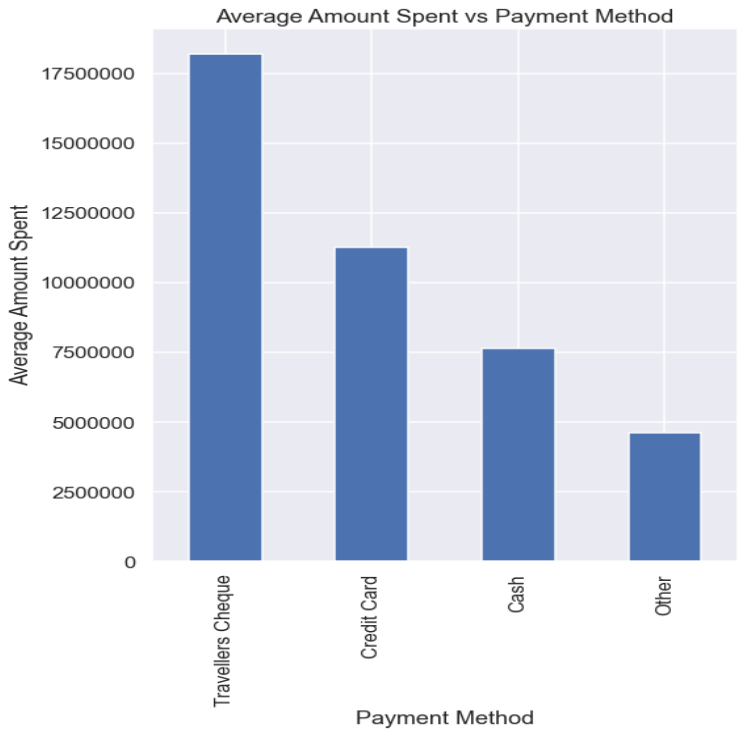
People who come with their husbands and children are most likely to spend the most money, according to the graph above. Individuals who travel alone, on the other hand, tend to spend the least amount of money. This is fascinating because the bar chart depicting the distribution of travel with shows that more people arrive alone than with a companion. As a result, a large number of visitors makes up the group that spends the least, which does not transfer well into revenue.

One option would be to increase advertising with topics such as:

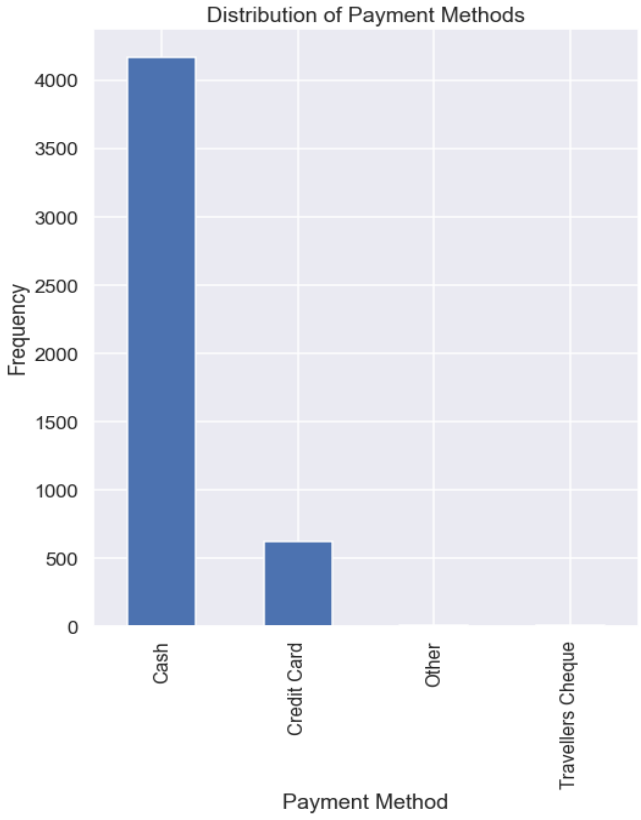
- A romantic getaway
- Family vacations

- Get-togethers and reunions

Amount Spent vs Payment Method

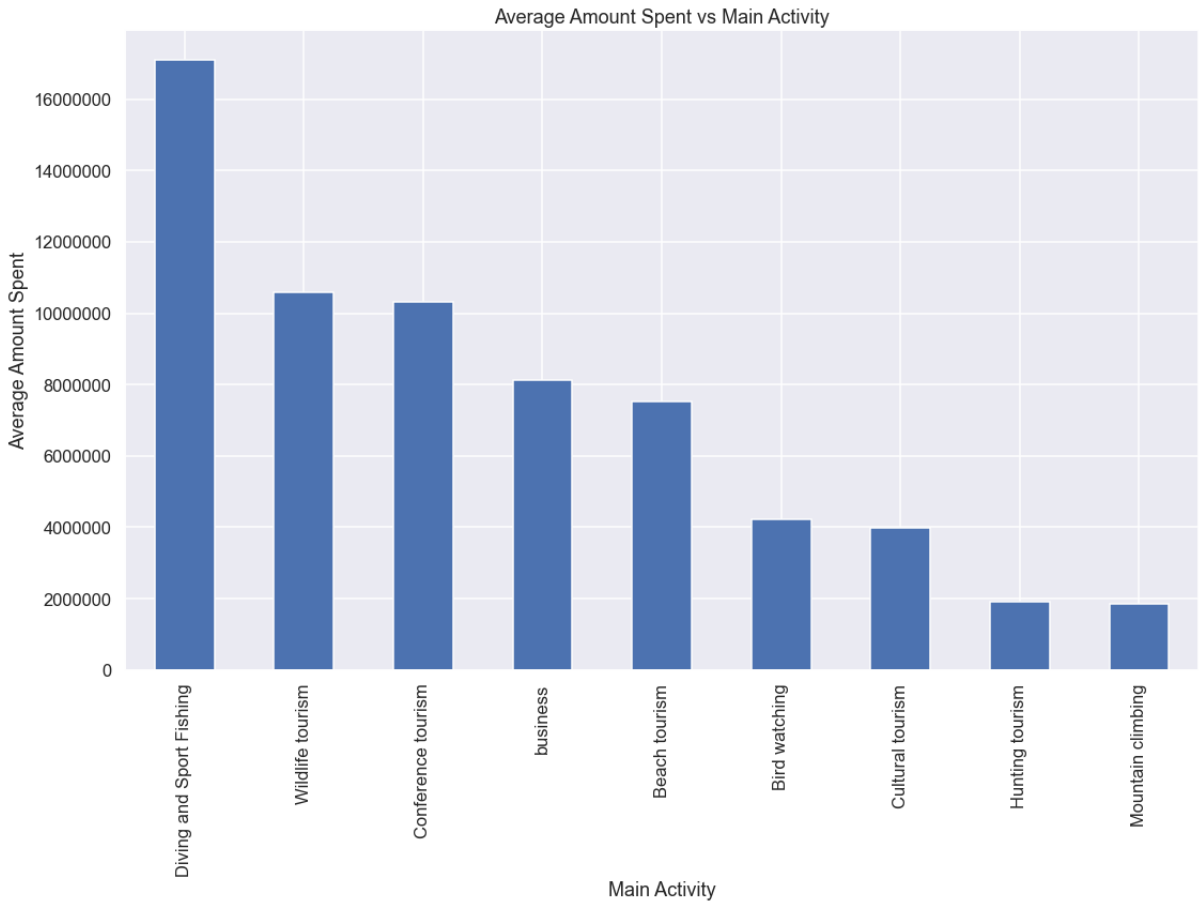


It seems tourists who use travellers' cheques spend the most amount of money. People may find them more comfortable to pay with seeing as they are safer to carry around and can be replaced if misplaced. This is however not the case as can be seen on further analysis.



Most people pay with cash; The traveller's cheque payment may be an outlier. This was confirmed by a really high value under the total_cost column in one or more rows, and a few rows for payments made using Travellers Cheques.

Main Activity vs Amount Spent

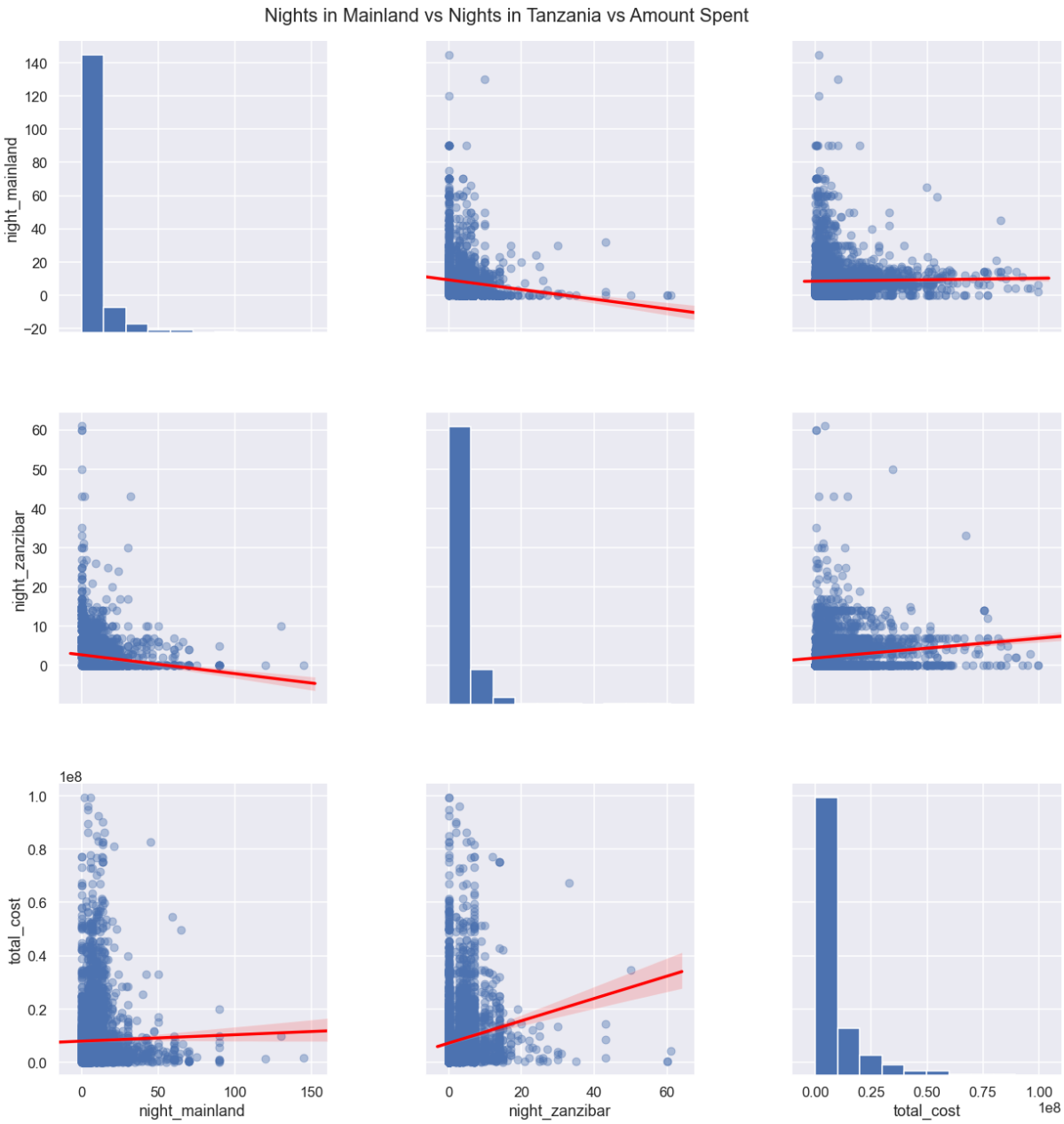


Diving and Sport Fishing attracts the most money on tours, despite it being the least popular tourist activity as seen in the univariate plot of main activities earlier. This is likely due to a higher cost related to the attraction. Wildlife tourism is the next money fetcher, and also the most popular activity on tour.

Multivariate Exploration

This entails studying the relationship between more than two variables. Once again, the variable total_cost will be prioritized.

Comparison of Nights Spent on the Mainland, Nights Spent in Zanzibar and Amount Spent



The regplot matrix proves a positive correlation between total cost against nights spent on the mainland and total cost against night spent in zanzibar. Although the scatter points seem to depict higher amounts spent in lesser nights, the regression line shows there is a weak positive correlation between night_mainland and total_cost, and also a weak correlation between night_zanzibar and total_cost. As expected, there is a negative correlation between night_mainland and night_zanzibar.

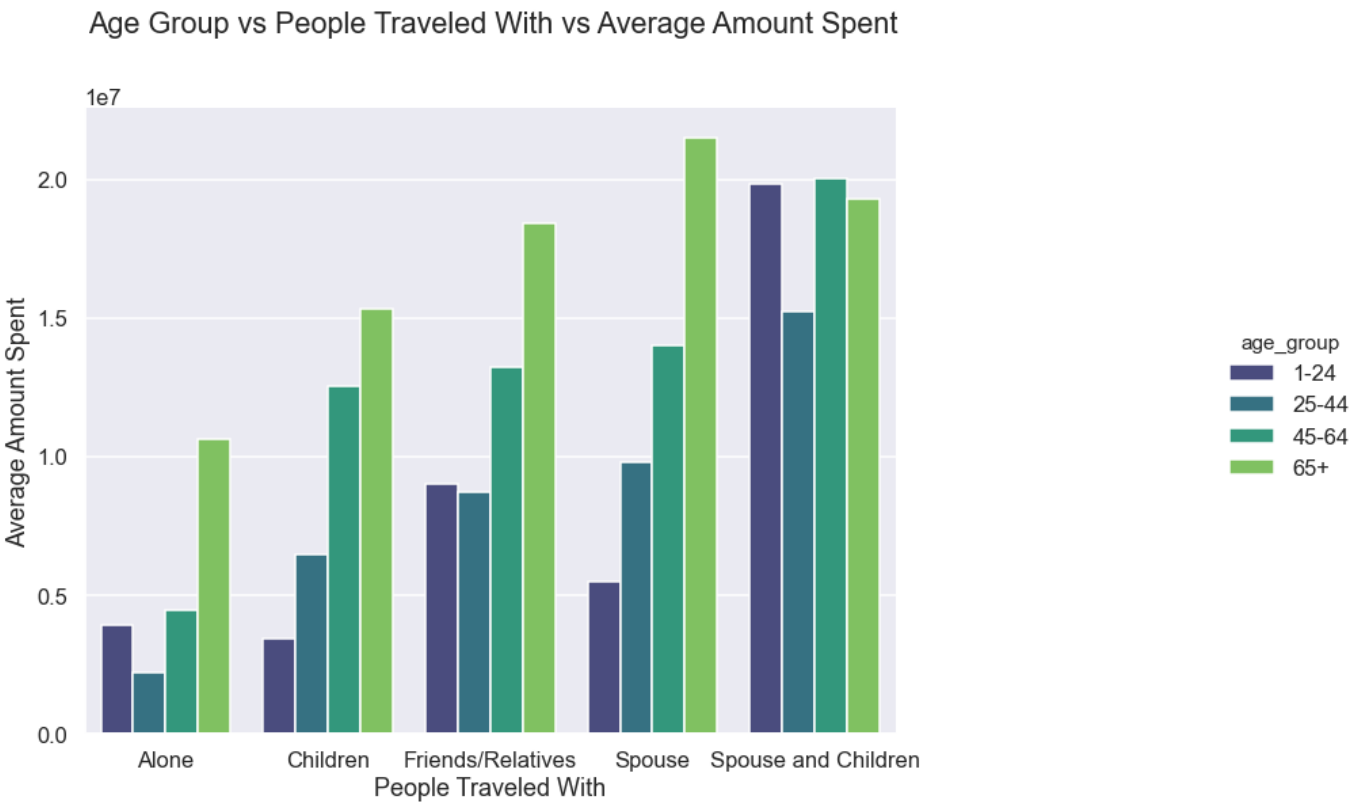
The actual correlations can be seen below:

- Correlation coefficient between night_mainland and total_cost is 0.0204731102797182
- Correlation coefficient between night_zanzibar and total_cost is 0.14513910874128153

- Correlation coefficient between night_mainland and night_zanzibar is 0.11815514846039063

All weak correlations.

Age Group vs People Traveled with vs Amount Spent



When visiting with their spouses and children, couples as young as 1-24 appear to spend the most on tours, even more than persons 65 and older, who appear to dominate the average spendings. We can observe that the average amount spent by visitors visiting with simply their spouse increases as they become older. It is clear from this narrative that encouraging couples to come on tours with their children, regardless of age, is one approach to increase income.

THE MODEL

The model was built using linear regression. Linear regression is a fundamental and widely utilized sort of predictive analysis. The goal of regression is to look at two things:

1. Is it possible to forecast an outcome (dependent) variable using a set of predictor variables?
2. Which factors in particular are significant predictors of the outcome variable, and how do they influence the outcome variable?

These regression estimations are used to illustrate how one dependent variable interacts with one or more independent variables. The simplest version of the regression equation with one dependent and one independent variable is y = bx + c, where y represents the estimated

dependent variable score, c represents the constant, b represents the regression coefficient, and x represents the independent variable score.

Model Overview

For this dataset, a multiple linear regression model was built. This is because a lot of factors seem to contribute to the response but one factor alone cannot suffice in explaining the response variable (total_cost).

The model was built and refined by removing variables with the highest p-values and highest Variance Inflation Factors (VIFs) one by one. The VIF is a measure of collinearity; high VIFs indicate a correlation between predictor variables. Ideally, we'd want high correlation between the predictor and response variables but not between the different predictor variables. This refinement was carried out until all p-values were less than 0.05 and all VIFs were less than 5

Performance Summary

The summary of the model is seen below:

OLS Regression Results

Dep. Variable:	total_cost	R-squared:	0.394
Model:	OLS	Adj. R-squared:	0.392
Method:	Least Squares	F-statistic:	210.2
Date:	Mon, 06 Jun 2022	Prob (F-statistic):	0.00
Time:	16:25:31	Log-Likelihood:	-77920.
No. Observations:	4542	AIC:	1.559e+05
Df Residuals:	4527	BIC:	1.560e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.486e+06	3.88e+05	-8.992	0.000	-4.25e+06	-2.73e+06
package_insurance	1.32e+06	3.22e+05	4.094	0.000	6.88e+05	1.95e+06
night_mainland	1.159e+05	1.57e+04	7.387	0.000	8.51e+04	1.47e+05
night_zanzibar	2.475e+05	3.67e+04	6.749	0.000	1.76e+05	3.19e+05
first_trip_tz	6.156e+05	2.54e+05	2.428	0.015	1.19e+05	1.11e+06
25-44	1.765e+06	3.33e+05	5.303	0.000	1.11e+06	2.42e+06
45-64	3.796e+06	3.65e+05	10.405	0.000	3.08e+06	4.51e+06
65+	7.129e+06	5.27e+05	13.524	0.000	6.1e+06	8.16e+06
Children	2.601e+06	5.06e+05	5.137	0.000	1.61e+06	3.59e+06
Friends/Relatives	1.872e+06	2.75e+05	6.807	0.000	1.33e+06	2.41e+06
Spouse	1.806e+06	2.89e+05	6.245	0.000	1.24e+06	2.37e+06
Spouse and Children	4.97e+06	3.97e+05	12.532	0.000	4.19e+06	5.75e+06
Leisure and Holidays	1.968e+06	2.87e+05	6.849	0.000	1.4e+06	2.53e+06
Wildlife tourism	1.4e+06	2.18e+05	6.430	0.000	9.73e+05	1.83e+06
Package Tour	5.453e+06	2.78e+05	19.632	0.000	4.91e+06	6e+06

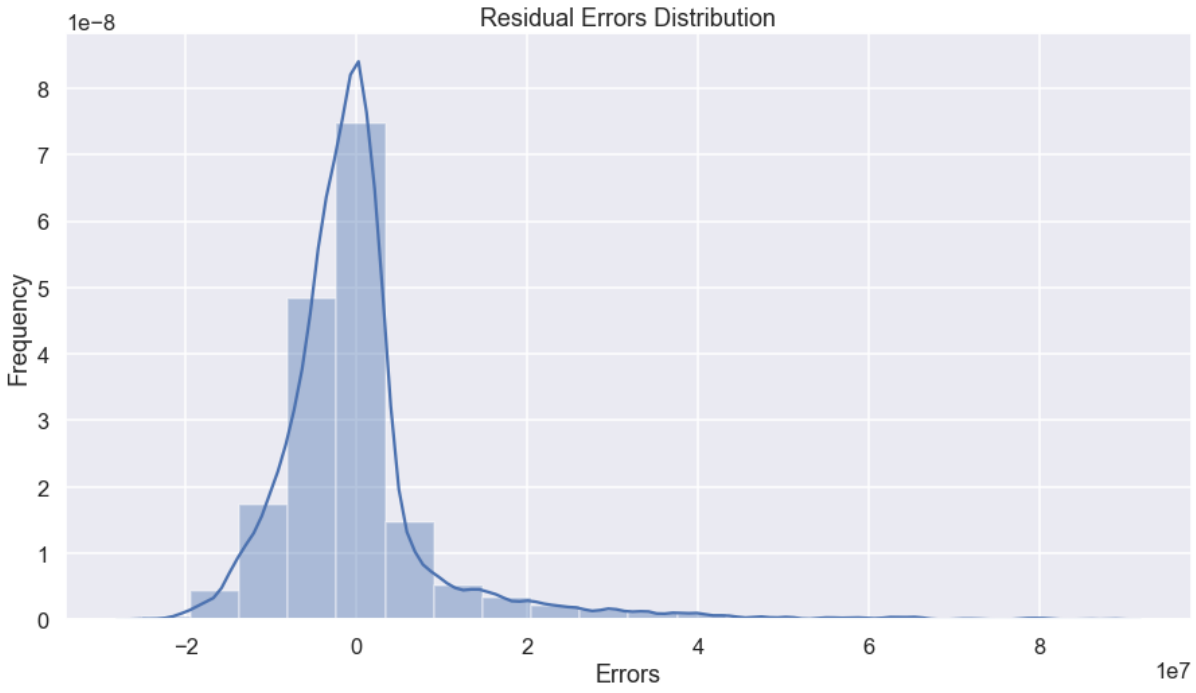
Omnibus:	1405.740	Durbin-Watson:	2.039
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5193.392
Skew:	1.513	Prob(JB):	0.00
Kurtosis:	7.276	Cond. No.	68.9

After removing variables and reviewing the model 13 times, the 14th model met the conditions of <0.05 p-value and <5 VIF. An R-squared value of 0.394 shows that the model has

only a 39.4% fit. This however is not necessarily a bad thing. In order to assess the linear model's validity, let us plot the histogram of the error terms to see if they are distributed normally (which is one of the fundamental assumptions of linear regression).

Residual Analysis

Residual analysis is used to assess the appropriateness of a linear regression model by defining residuals and examining the residual plot graphs. Residual(*e*) refers to the difference between observed value(*y*) vs predicted value (*y*^). Every data point has one residual (tutorialspoint, 2022). We will be plotting a histogram of the residual points. For a good regression model, the histogram should be distributed normally.



The error terms, as can be seen, closely match a normal distribution. As a result, we may use the model in the test dataset to generate predictions.

CONCLUSIONS

Using the Tanzania tourist dataset, an analysis was carried out and noticeable trends were pointed out. Further analysis was done with respect to the total amount spent on tours.

Actionable Insights and Recommendations

Major ways to improve tourist influx include:

- Create publicity aimed more at the middle-aged by appealing to things they would find more relatable such as sightseeing and romantic events.
- Way too many tourists visit alone. More tourists should be encouraged to come with family; especially their spouses and children. This can be pitched as family fun time tourism or a romantic getaway.

- The internet and media should be used to create more awareness of the tourist locations. Newspapers and Magazines providing more information sources than the internet should not be the case in this age
- Conference tourism should be popularized
- The model should be used to enable tourists better plan for tours before embarking.

In the end, a simple multiple linear regression model was built using machine learning and tuned manually. A multiple linear regression model was used because the target variable is always dependent on more than one variable.

REFERENCES

i. *Tourism in Tanzania, Wikipedia article. Retrieved 05/06/2022 from https://en.wikipedia.org/wiki/Tourism_in_Tanzania*

ii. *Statistics - Residual analysis, Tutorials Point. Retrieved 05/06/2022 from https://www.tutorialspoint.com/statistics/residual_analysis.htm#*