Tanzania Tourist Prediction

Prepared by:
-Sorabh Vasudeva
-Mustapha
-Sohee

Objective

- Our goal is to predict how much money a tourist will spend when visiting Tanzania?

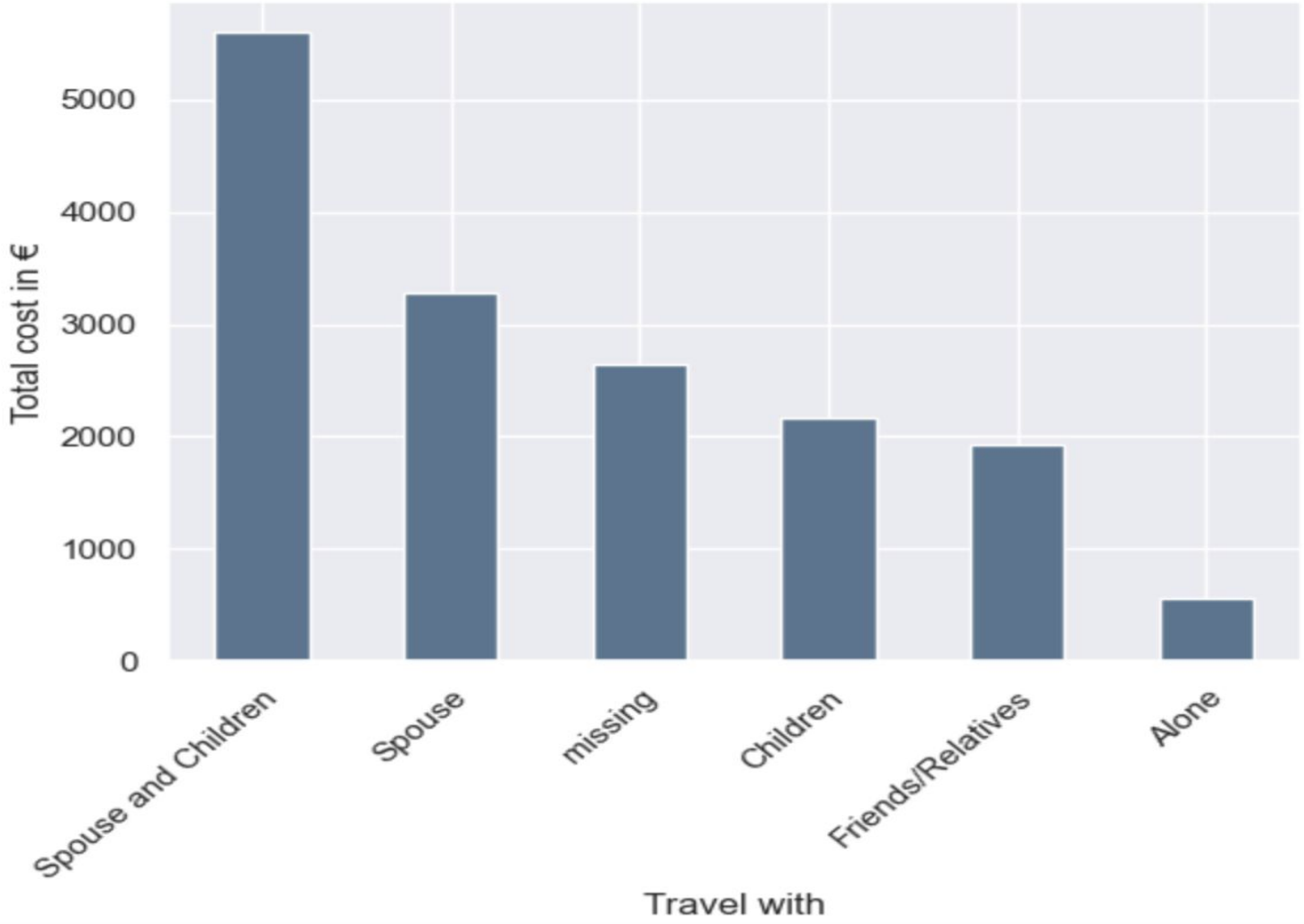
Dataset Overview

- Dataset includes over 6476 tourist records with 23 columns:

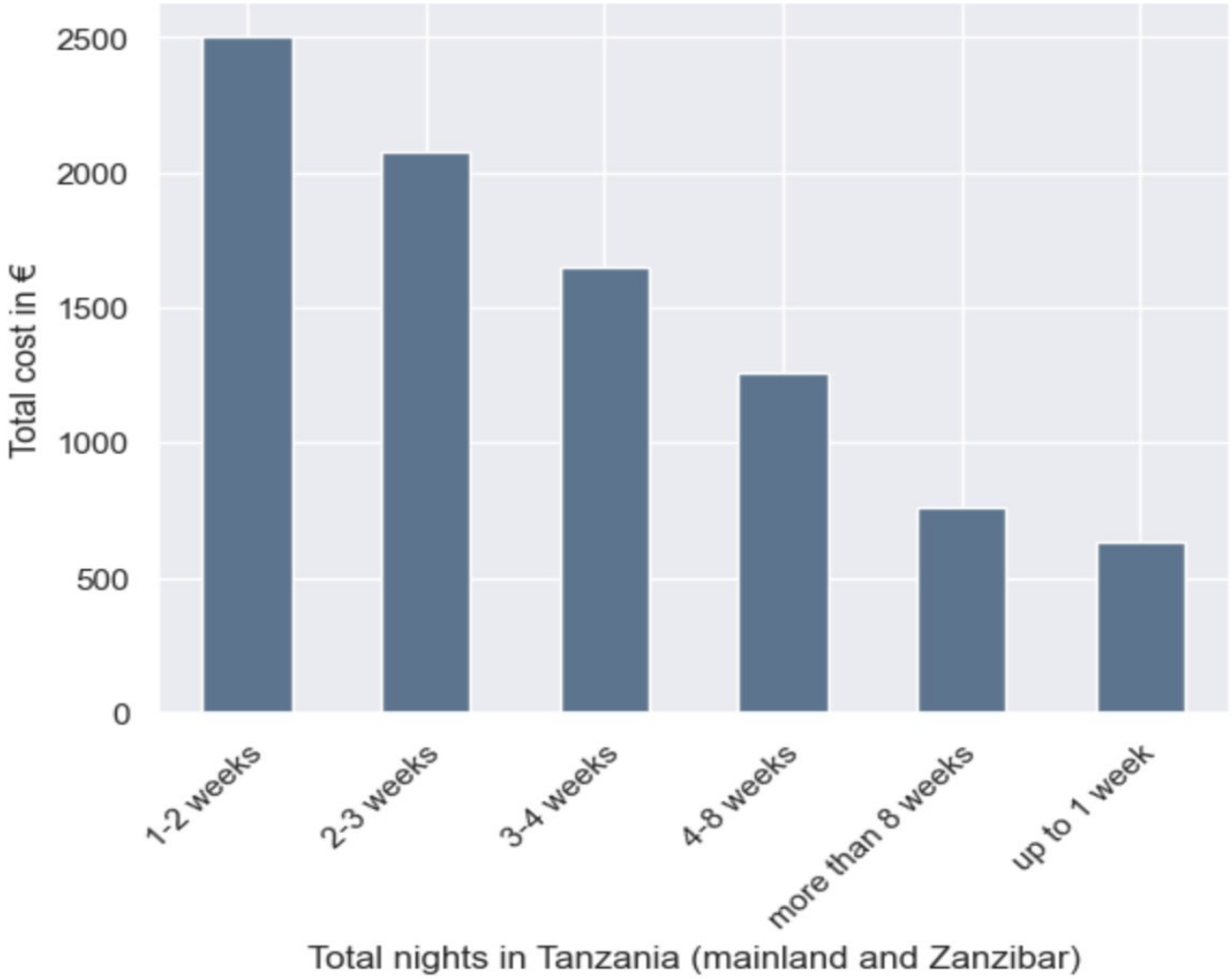
	Column Name	Definition
0	id	Unique identifier for each tourist
1	country	The country a tourist coming from.
2	age_group	The age group of a tourist.
3	travel_with	The relation of people a tourist travel with t...
4	total_female	Total number of females
5	total_male	Total number of males
6	purpose	The purpose of visiting Tanzania
7	main_activity	The main activity of tourism in Tanzania
8	infor_source	The source of information about tourism in Tan...
9	tour_arrangment	The arrangment of visiting Tanzania
10	package_transport_int	If the tour package include international tran...
11	package_accomodation	If the tour package include accommodation service
12	package_food	If the tour package include food service
13	package_transport_tz	If the tour package include transport service ...
14	package_sightseeing	If the tour package include sightseeing service
15	package_guided_tour	If the tour package include tour guide
16	package_insurance	if the tour package include insurance service
17	night_mainland	Number of nights a tourist spent in Tanzania m...
18	night_zanzibar	Number of nights a tourist spent in Zanzibar
19	payment_mode	The mode of payment for tourism service
20	first_trip_tz	If it was a first trip to Tanzania
21	most_impressing	what impressed a toursit in Tanzania
22	total_cost	The total tourist expenditure in TZS(currency)

Random Insights/ Hypothesis

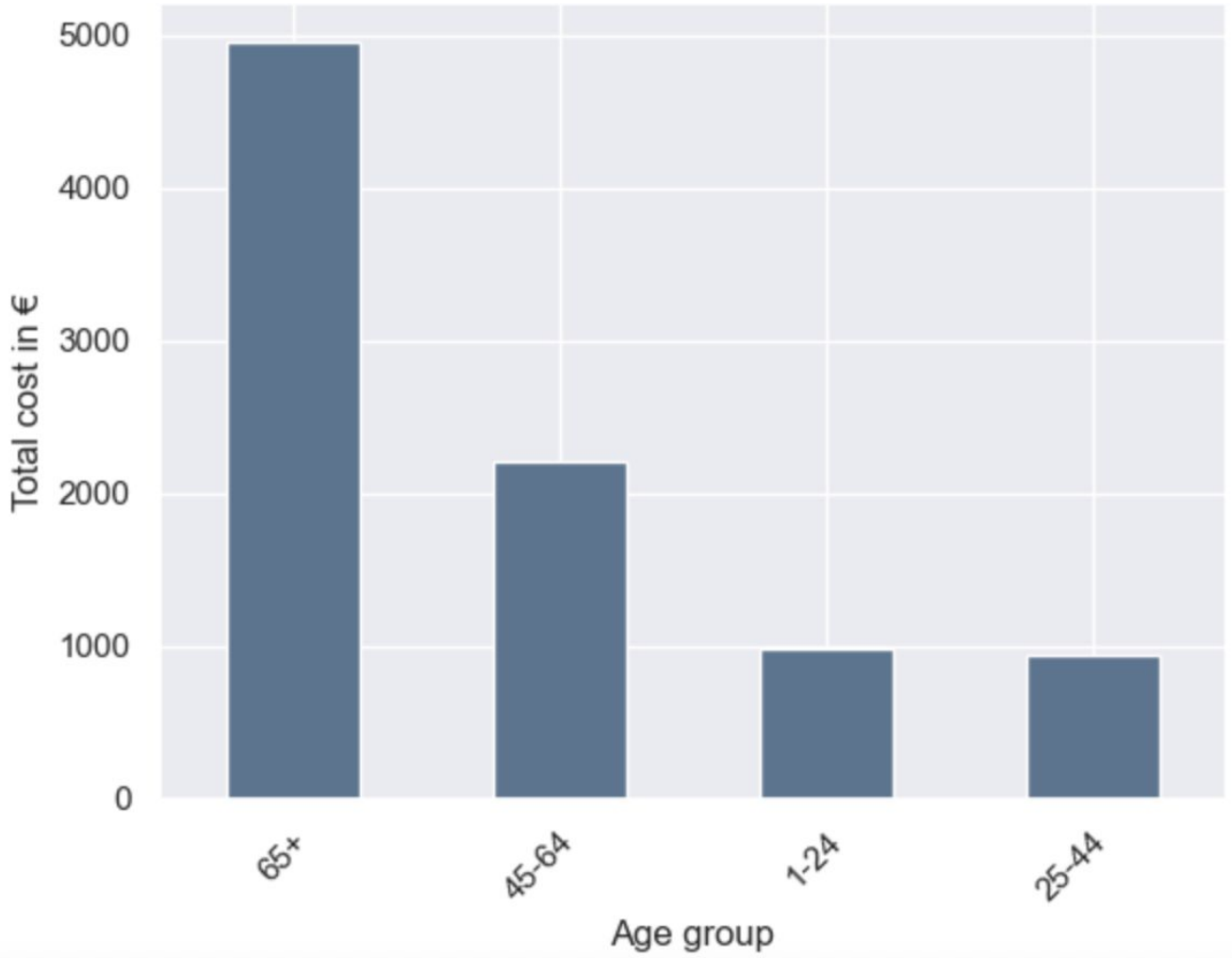
The bigger a group is, the more money they spend.



The longer a group stays, the more money they spend.



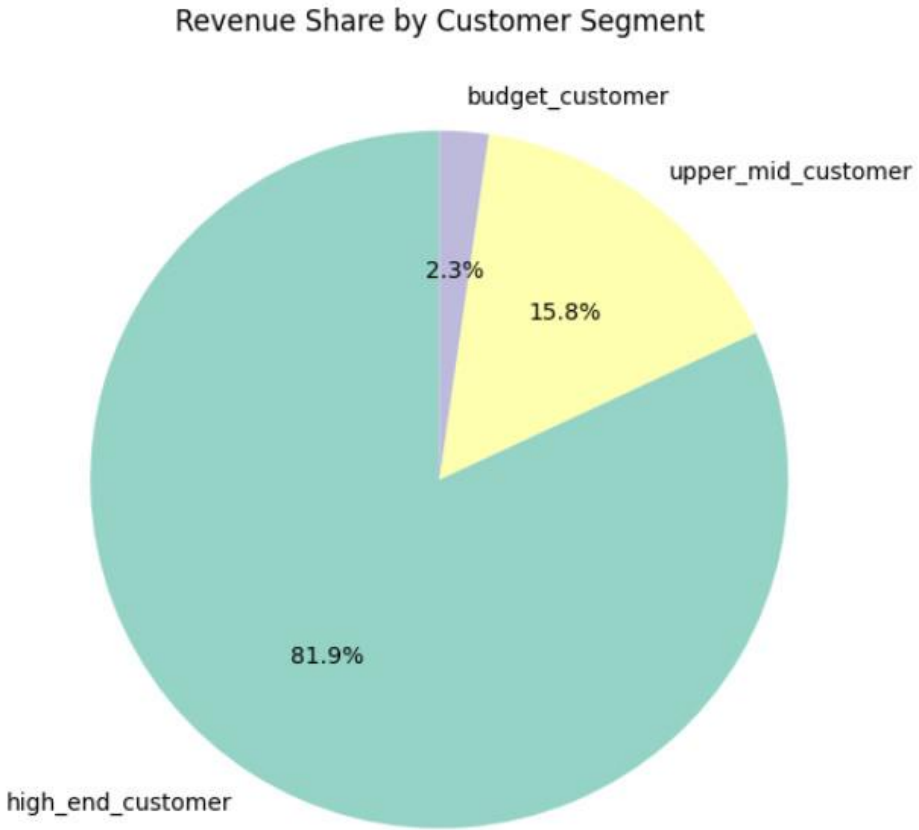
Older tourists spend more than younger tourists.

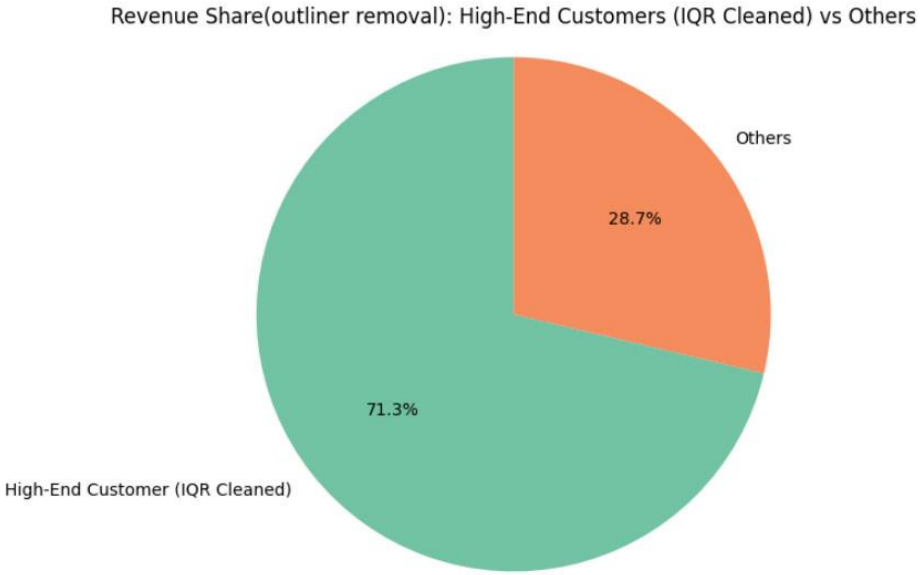


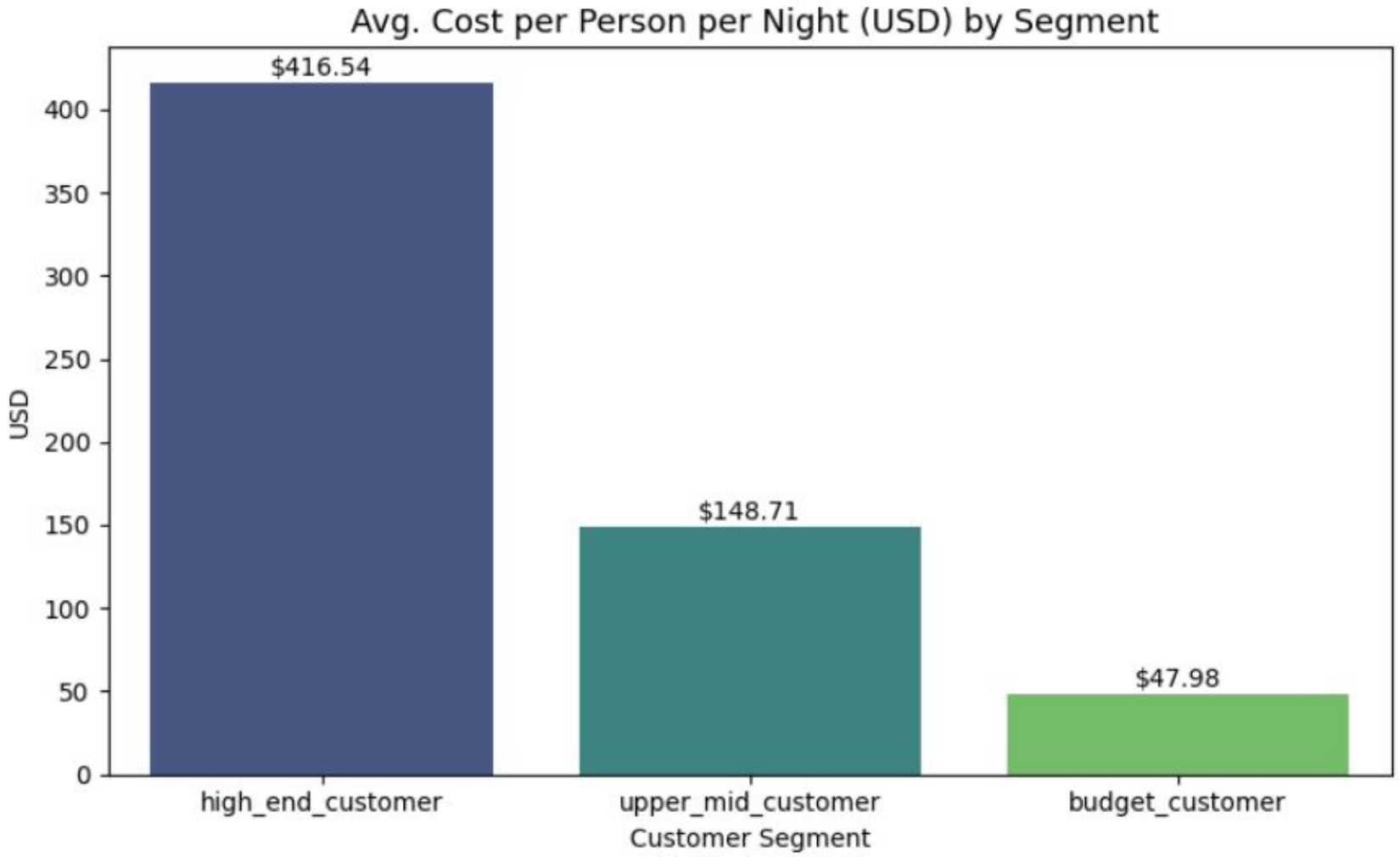
Customer Segment

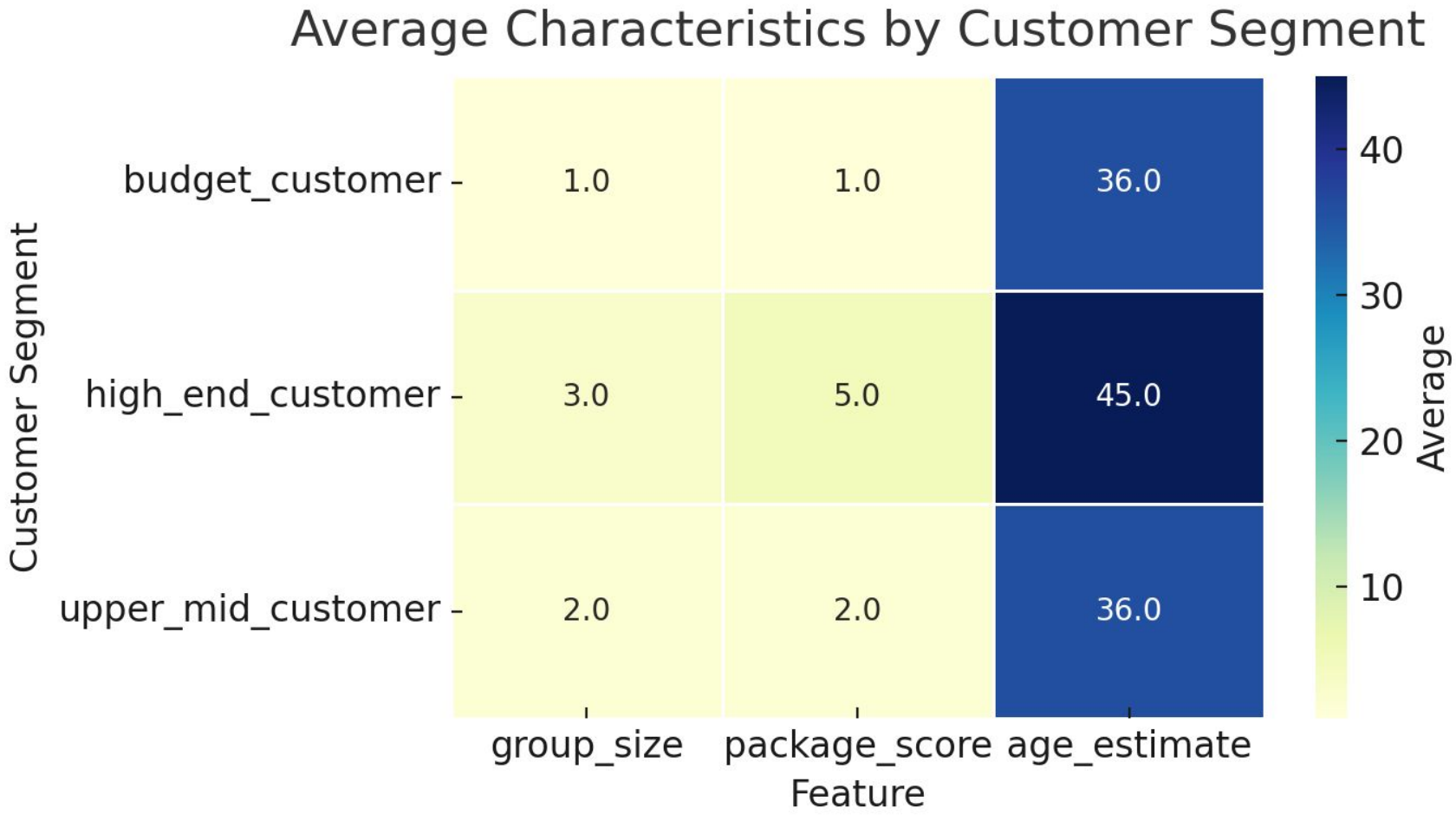
Segment	Count	Percentage
High-end customer	1610	34.00%
Budget customer	1563	33.00%
Upper-mid customer	1562	32.98%
Total	**4735**	**100**

Revenue Share by Customer Segment

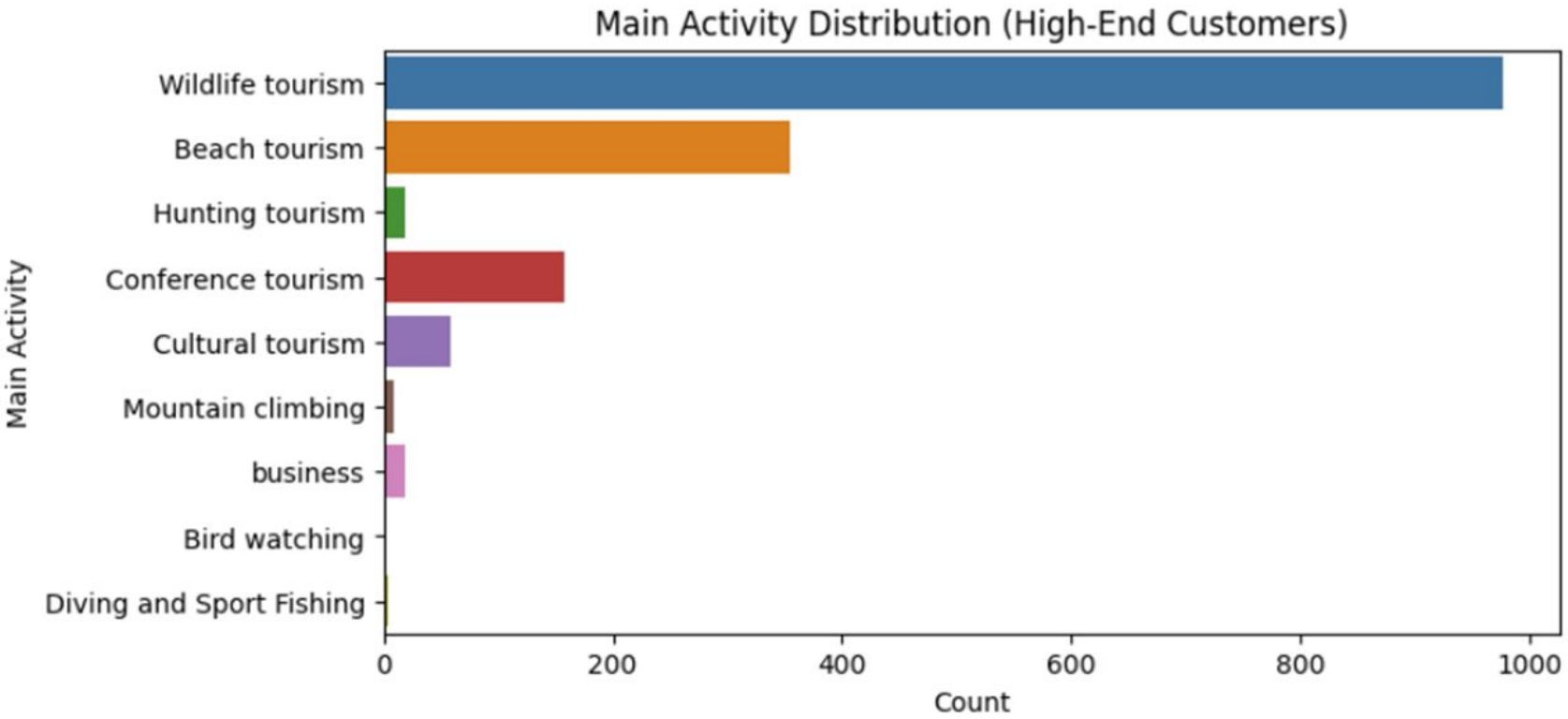





Avg. Cost per Person per Night (USD) by Segment
$416.54
$148.71
$47.98
400
350
300
250
200
150
100
50
0
USD
high_end_customer
upper_mid_customer
budget_customer
Customer Segment



- Package Score: How many packages are used?


Main Activity Distribution (High-End Customers)
Wildlife tourism
Beach tourism
Hunting tourism
Conference tourism
Cultural tourism
Mountain climbing
business
Bird watching
Diving and Sport Fishing
Main Activity
0
200
400
600
800
1000
Count

Summary of High-End Customer Characteristics Analysis

Feature	Analysis Result
group_size	Mostly 2 or more people (Median: 2, Upper 25%: 3+)
package_score (number of package)	Very high (Mean: 5.2, Median: 6)
age_estimate	Mainly 34–54 years old (Median: 34, Upper quartile: 54)
age_group	25-44: 43%, 45-64: 38%
travel_with	Spouse (36%), friends/relatives + alone (each 22%)
payment_mode	Mostly cash (83%) → Credit card payment 17%
main_activity	Focused on wildlife tourism (59%), beach tourism (23%)

Insight-Driven Recommendation Strategy

Element	Example Strategy
Package Composition	Include 5+ items such as accommodation, meals, guide, insurance, transport
Target Age Group	Average age 44 → Tailored for middle-aged and older adults
Travel Companion	Spouse / alone / friends → Emphasize privacy and luxury
Message	“Luxury Wildlife + Beach Twin Tour Just for You”, “Private All-Inclusive Package for Couples”

Predictions using ML model

Modeling

•**Cleaned the Data for Accuracy**
We removed extreme values and adjusted the data so the model could learn more consistently and accurately.

•**Added Smart Features**
We created helpful new data points — like group size, payment method, and estimated age — to give the model better insight into customer behavior.

•**Used a Powerful Model (XGBoost)**
This model is known for high accuracy and works well even with messy or complex data.

•**Focused on High-Spending Customers**
We trained the model to pay extra attention to top spenders, making it more useful for business decision-making.

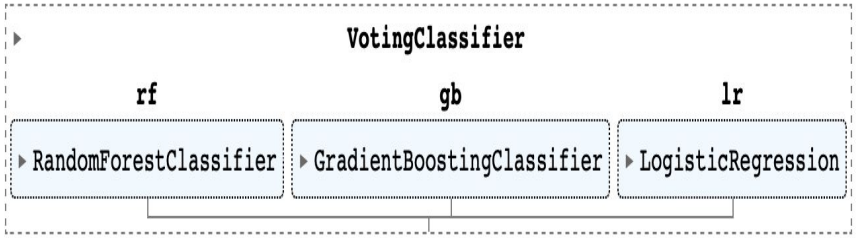
•**Optimized the Model for Better Results**
We tested many settings automatically to improve performance — and saw about a **5% increase in accuracy**.

Revenue Prediction Model Performance Summar

Metric	Business Interpretation
MAE($1,335)	On average, the prediction is off by ±$1,335 per customer
RMSE($2,304)	Some customers may have significantly inaccurate predictions
SMAPE(54%)	Predictions differ from actuals by about ±54% on average
$R^2$0(73%)	The model explains 73% of the variability in revenue

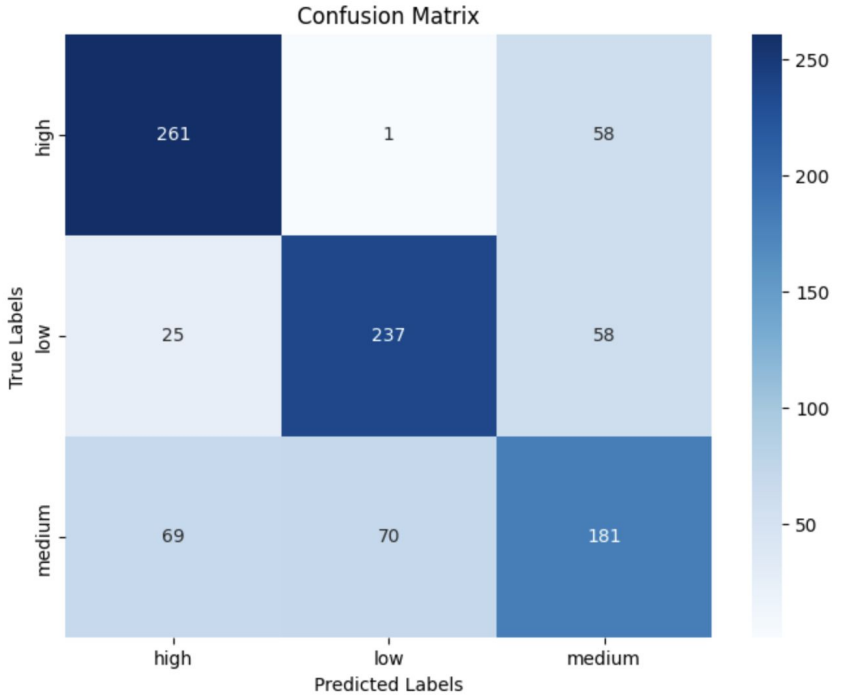
Total Actual Revenue: Approximately $35,400,000 (based on TZS 81,420,000,0000)
Average Revenue Per Customer: Approximately $7,200 (based on average TZS 16,560,000)

Voting Methods

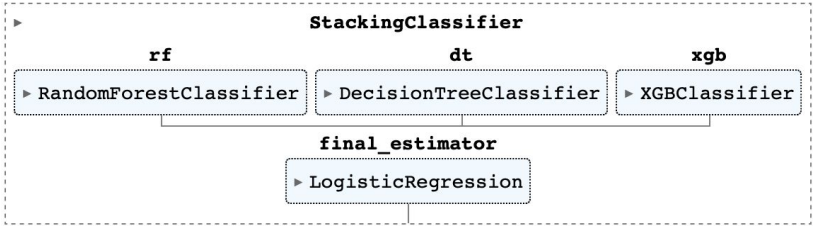


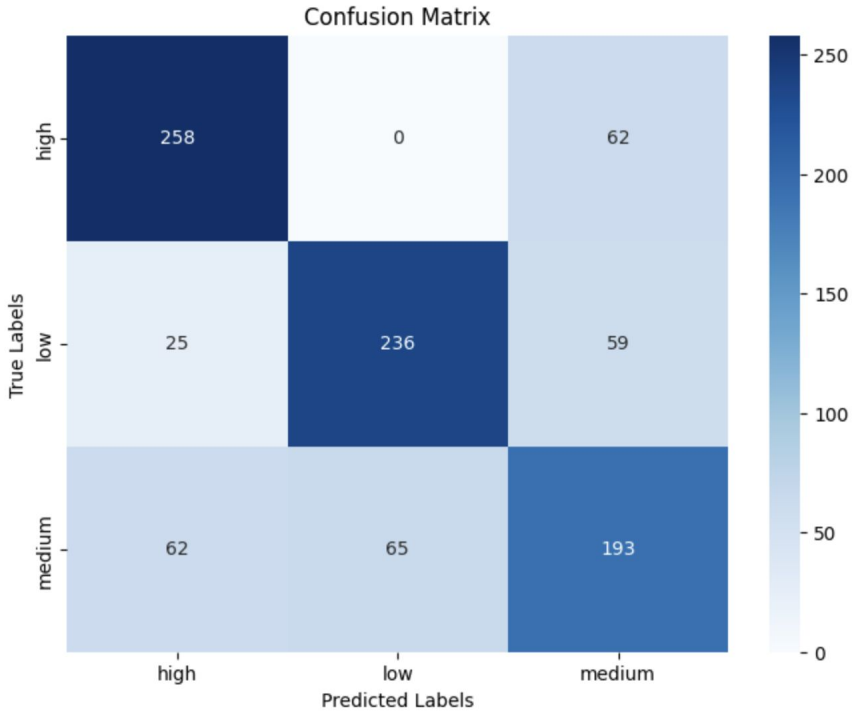
- **The diagonal values represent correct predictions**
- **Off-diagonal values show misclassifications:**
- 69 true “medium” were predicted as “high”, and 70 as “low”

- **ACCURACY 71%**



Stacking Classifier





- ACCURACY 72%
- **Both models often confuses “medium” class with both “low” and “high”.**

Data Product Idea

‘Data Product: test predictions can be used in practice’

- **Personalized Customer Recommendations**
  Test data simulates real customers.
  → Use predictions to group customers, suggest products, and automate marketing.
- **Can Be Used in Real-Time**
  Connect the model to a system.
  → When customer info is entered, spending and suggestions are shown instantly. (ex: Flask, FastAPI)
- **Supports Tourism Strategy**
  → See which customer types spend more (by nationality, age, purpose)
  → Plan better campaigns, services, and investments

Future Work

- **Use More Helpful Data**
  (e.g. season, weather, currency rates, holidays — *external data*)
- **Group Similar Tourists**
  (divide customers into types — *clustering, e.g. KMeans*)
- **Try Smarter Model Tools**
  (combine models or let a system choose the best — *AutoML, model ensembles, Catboost*)
- **Build a Live Prediction Tool**
  (predict in real-time when user info is entered — *API with Flask or FastAPI*)
- **Make the Model Even More Accurate**
  (improve step by step — *prediction optimization*)

Q & A