Empower Tanzania

Eliaskouloures.com

EliasKouloures.com

EliasKouloures.com

EliasKouloures.com


🧠 Create Insights for Tanzania


Use Dataset of Statistics Bureau

EliasKouloures.com

🔧 Build accurate ML Model

EliasKouloures.com

EliasKoutoures.com


💪 Master Data Quality Issues


💰 Generate Action-Items
EliasKouloures.com

Data Cleaning

a) Handle missing values

b) Check for outliers

c) Encode categorical variables

d) Normalize numerical features

```
missing_values
```

[2] ✓ 0.0s

```
ID                          0
country                     0
age_group                   0
travel_with              1114
total_female                3
total_male                  5
purpose                     0
main_activity               0
info_source                 0
tour_arrangement            0
package_transport_int       0
package_accomodation        0
package_food                0
package_transport_tz        0
package_sightseeing         0
package_guided_tour         0
package_insurance           0
night_mainland              0
night_zanzibar              0
payment_mode                0
first_trip_tz               0
most_impressing           313
total_cost                  0
dtype: int64
```

EliasKouloures.com

```python
# Step 1a: Handle missing values based on the strategies discussed

# Impute 'travel_with' and 'most_impressing' with 'Unknown'
df['travel_with'].fillna('Unknown', inplace=True)
df['most_impressing'].fillna('Unknown', inplace=True)

# Impute 'total_female' and 'total_male' with their respective medians
df['total_female'].fillna(df['total_female'].median(), inplace=True)
df['total_male'].fillna(df['total_male'].median(), inplace=True)

# Verify if all missing values are handled
df.isnull().sum().sum()
```

✓ 0.0s

```
0
```

```python
# Step 1b: Identify and remove duplicate rows
duplicates = df.duplicated().sum()

# Remove duplicates if any
if duplicates > 0:
    df.drop_duplicates(inplace=True)

duplicates
```

```
0
```

```python
# Step 1c: Check for invalid/inconsistent values

# Check for negative numbers in 'total_cost'
negative_cost_count = (df['total_cost'] < 0).sum()

# Summary of checks
invalid_values_summary = {
    'Negative Costs': negative_cost_count,
}

invalid_values_summary
```

✓ 0.0s

```
{'Negative Costs': 0}
```

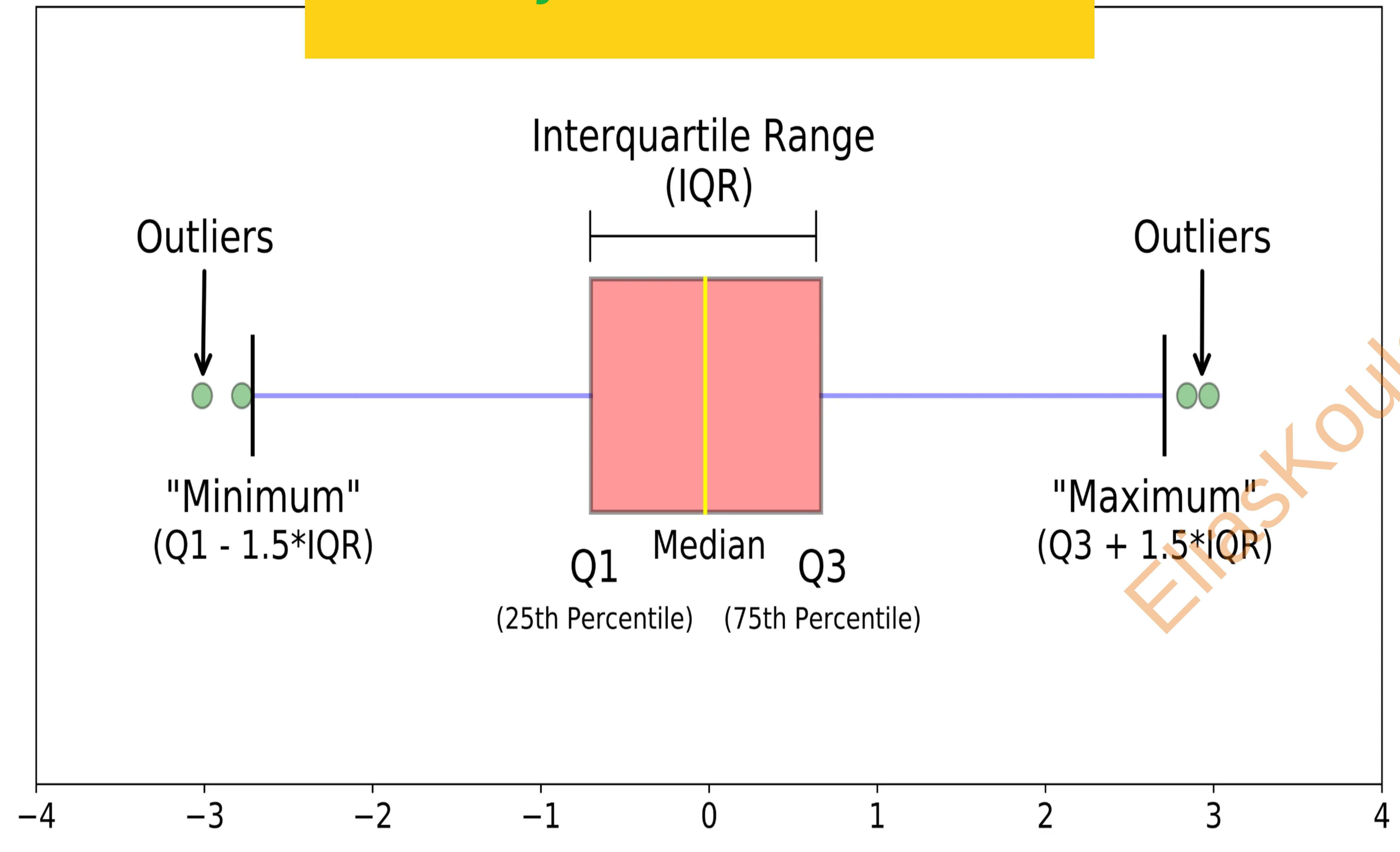
✓ 1.3s

	ID	total_female	total_male	package_transport_int	package_accomodation	package_food	package_transport_tz	package_sightseeing	pack
0	tour_0	1.0	1.0	No	No	No	No	No	
1	tour_10	1.0	0.0	No	No	No	No	No	
2	tour_1000	0.0	1.0	No	No	No	No	No	
3	tour_1002	1.0	1.0	No	Yes	Yes	Yes	Yes	
4	tour_1004	1.0	0.0	No	No	No	No	No	

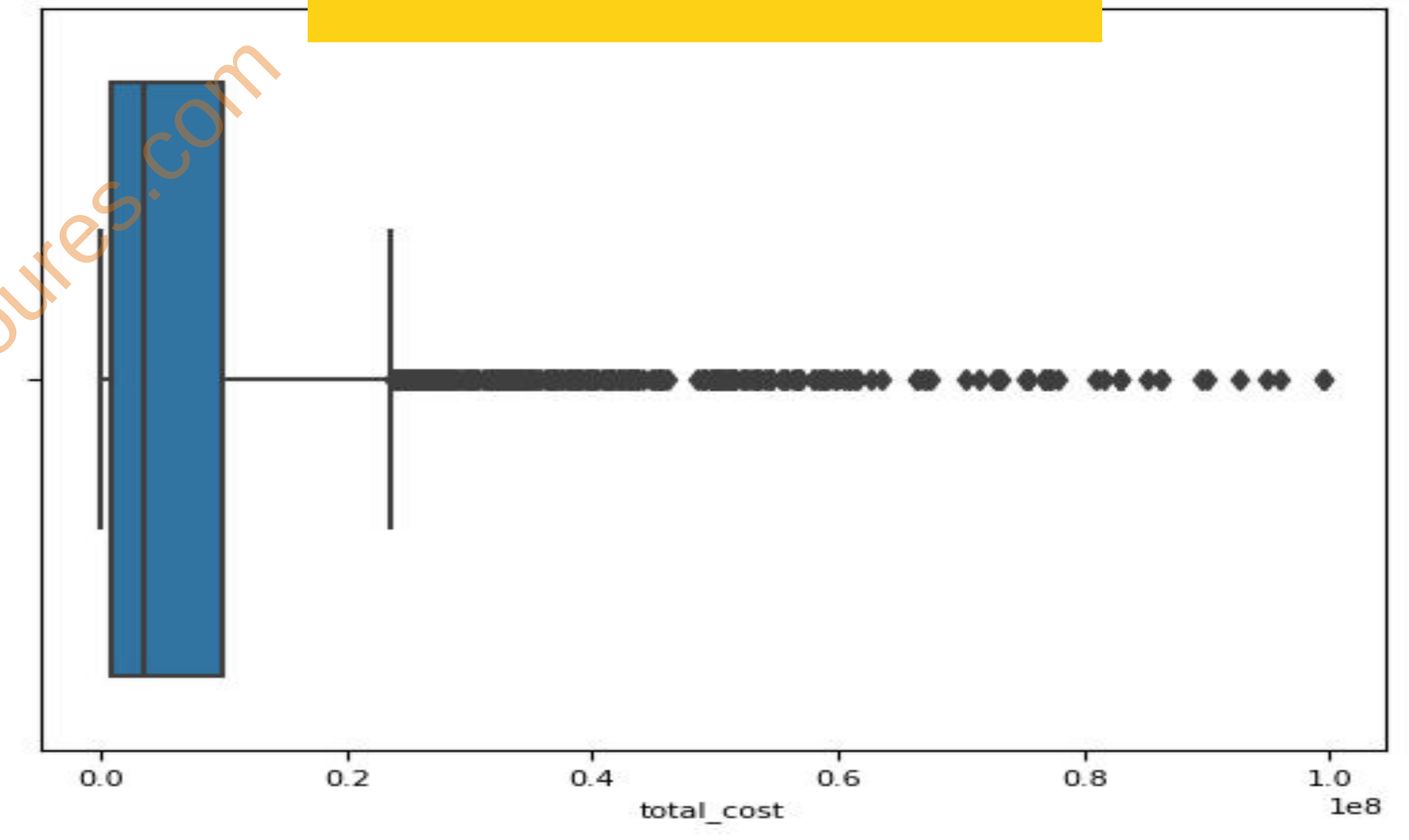
5 rows × 156 columns

Univariate Analysis heavily right-skewed

Evenly distributed data

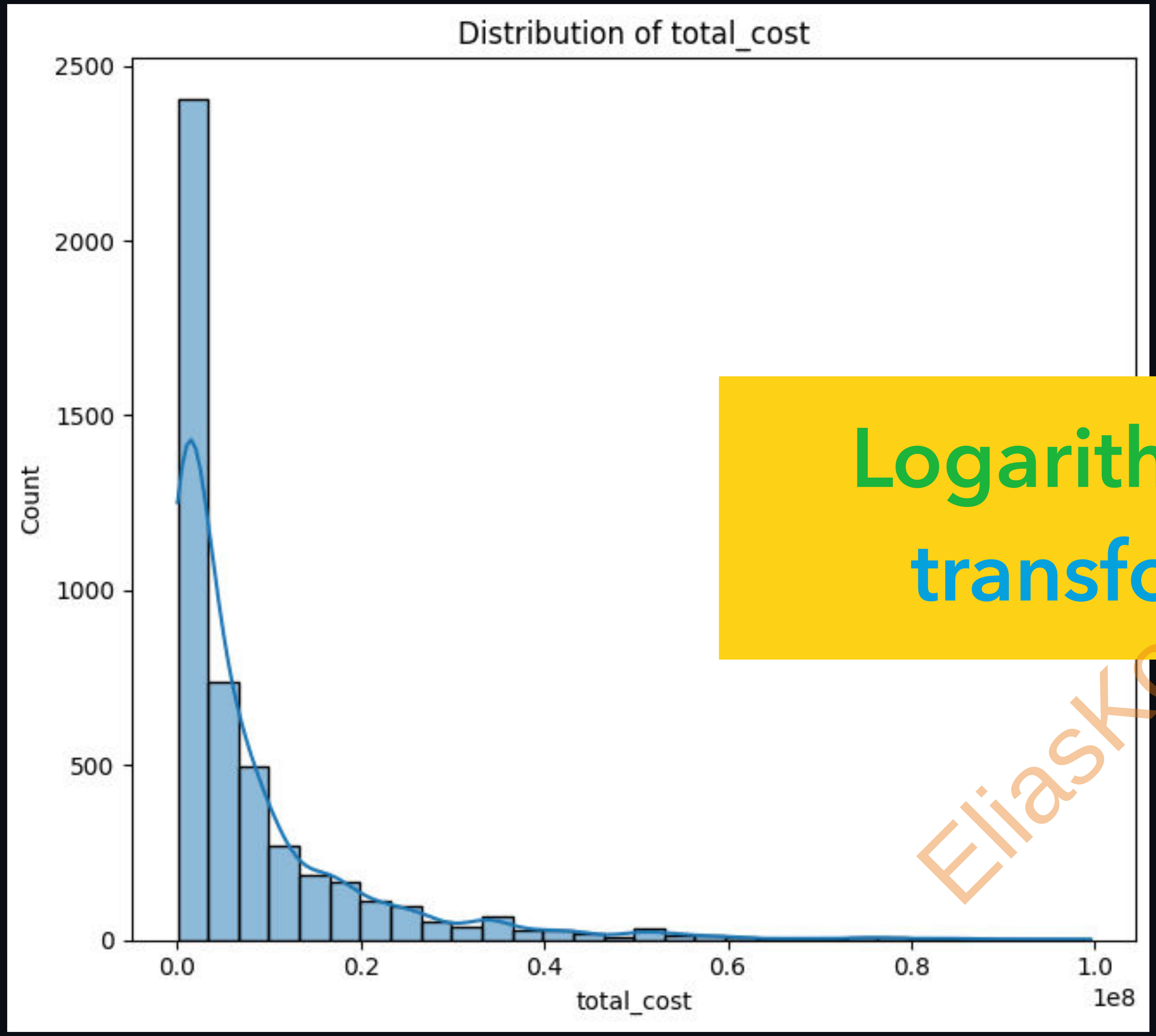


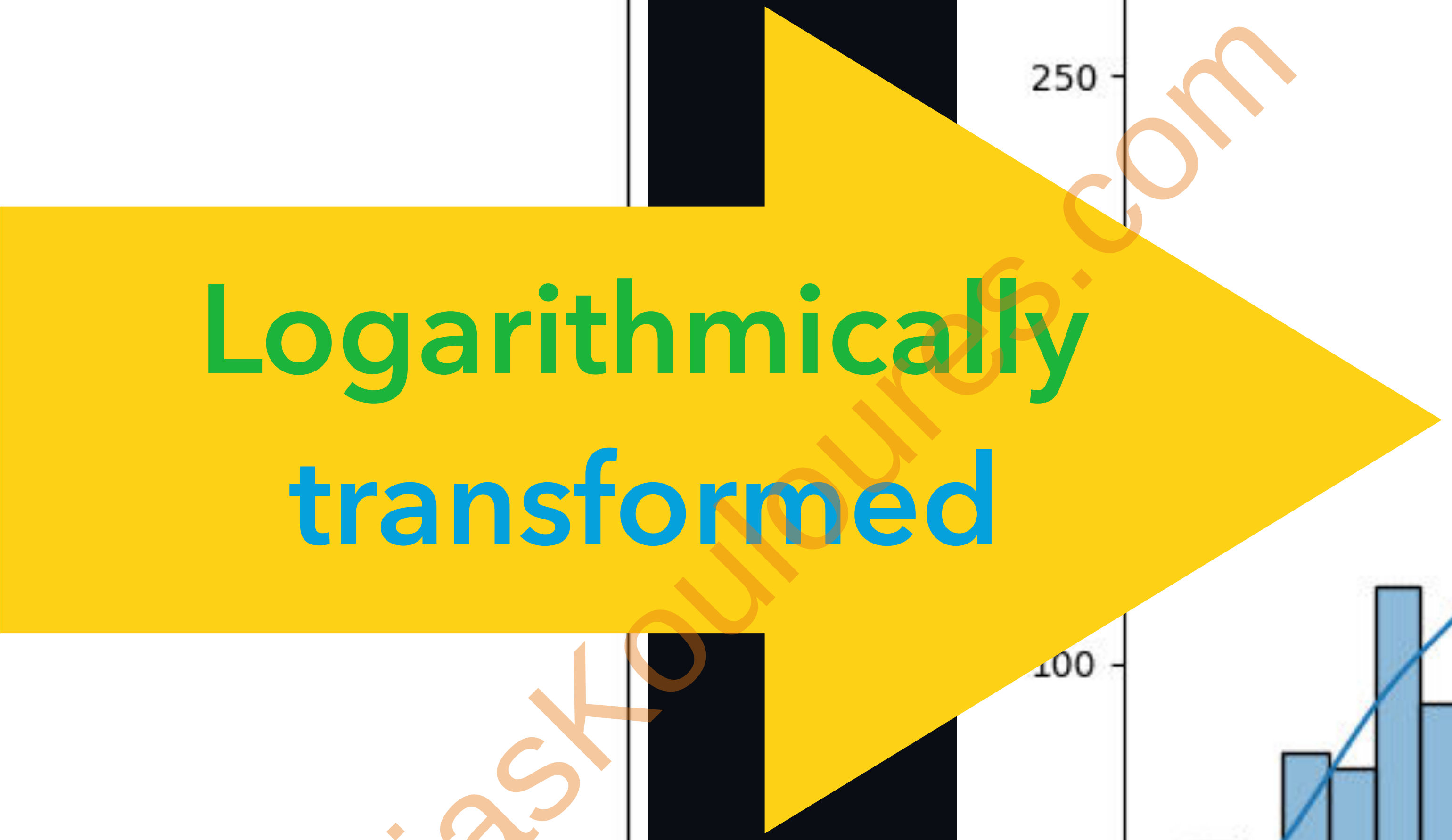
Tanzania Tourism data

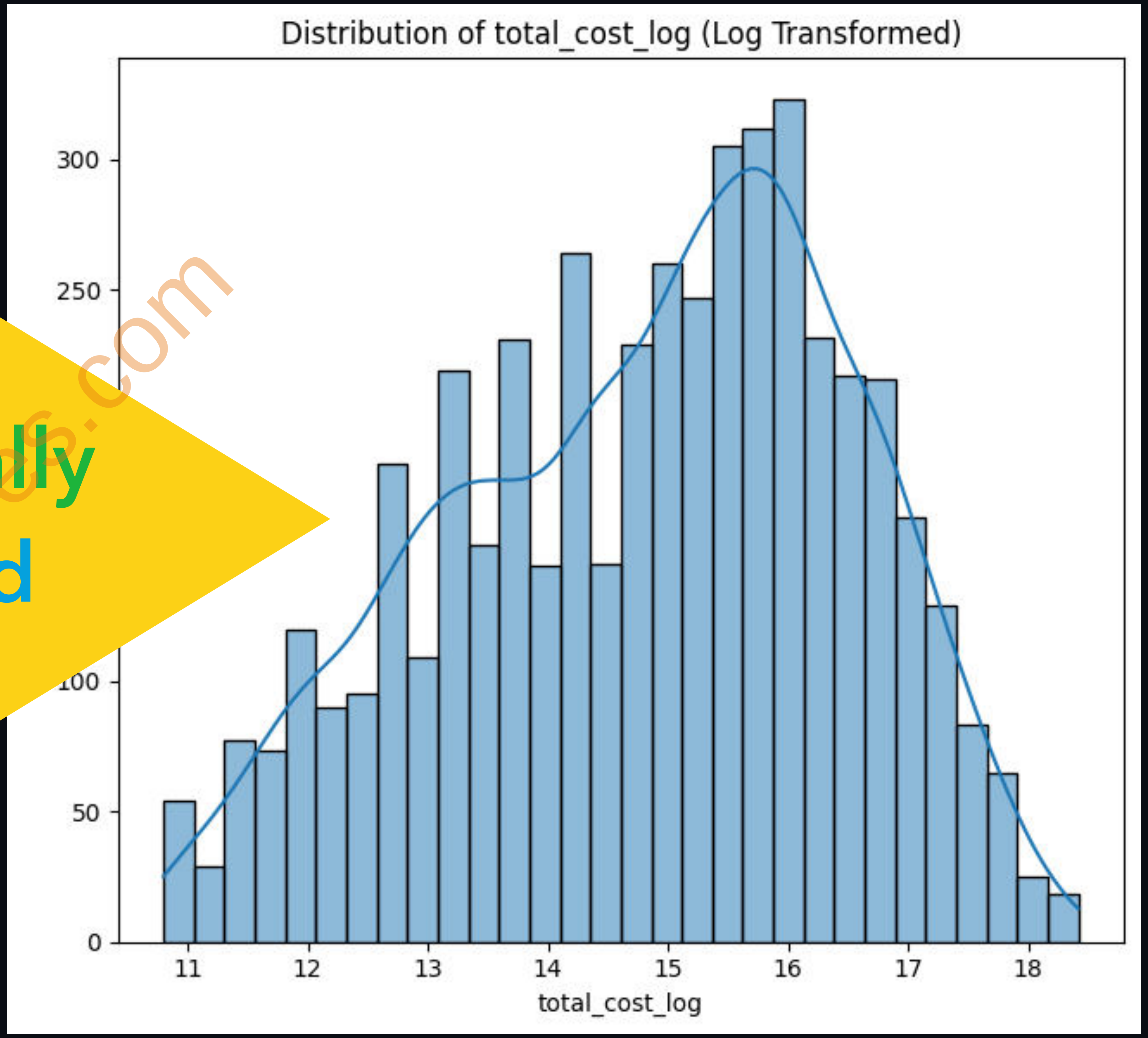


EliasKouloures.com

Visitor Spendings heavily right-skewed







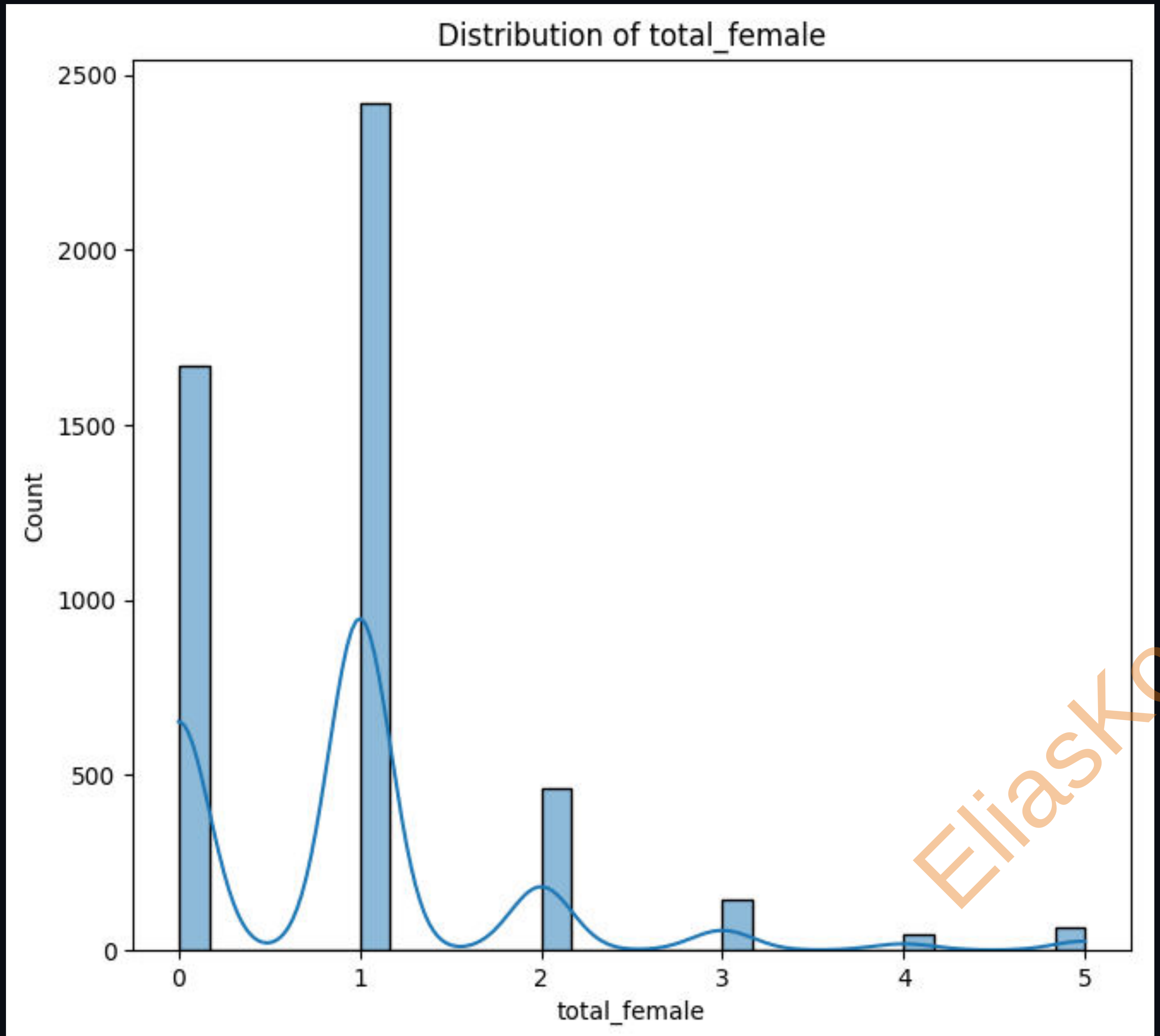
EliasKoufoures.com

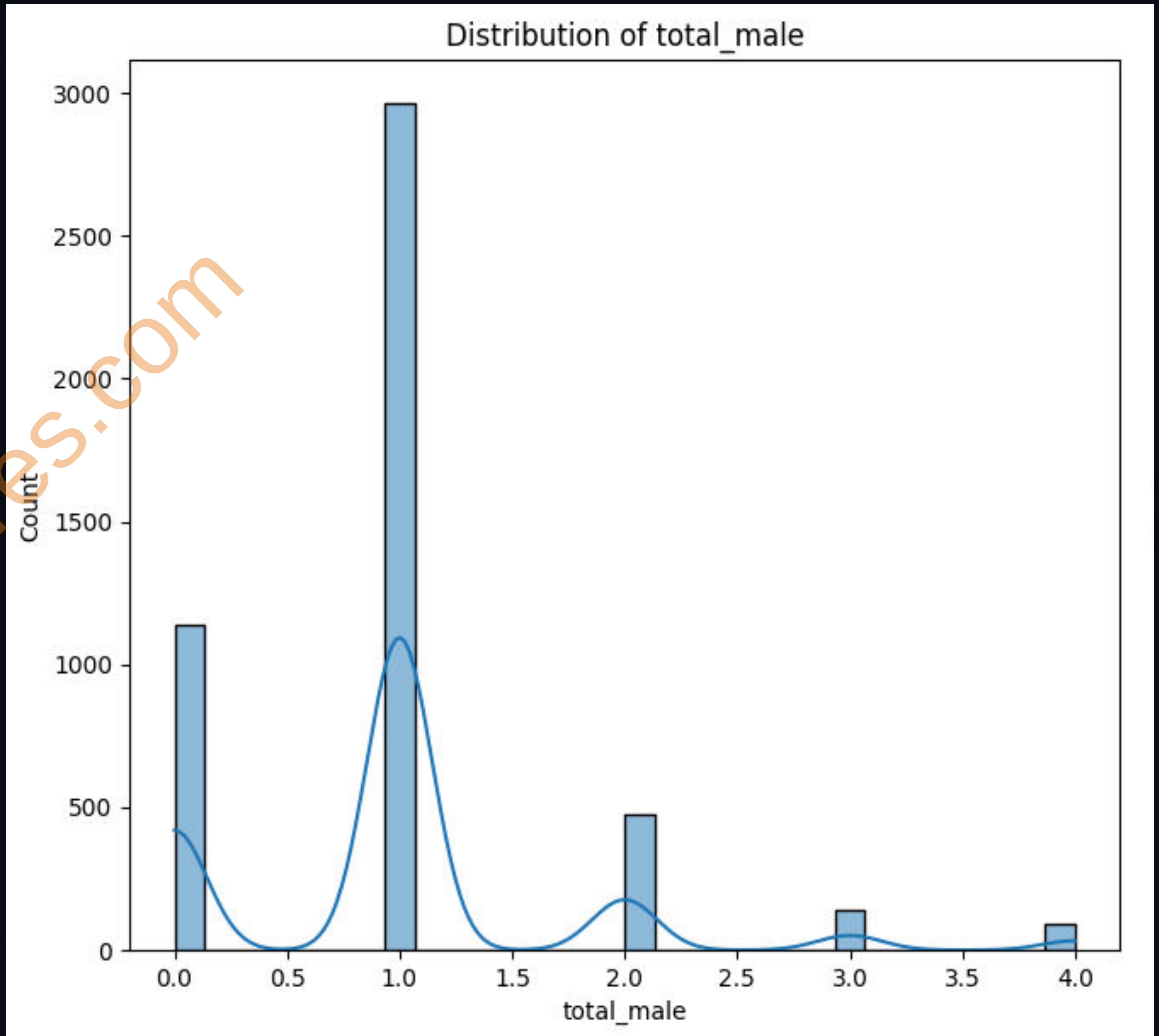
Observation: Wide distribution ➡️ high standard deviation.

Pitfall: Skewed target variable ➡️ issues with linear models.

Solution: Log transform to normalize distribution.

Number of Visitor clustered around 1





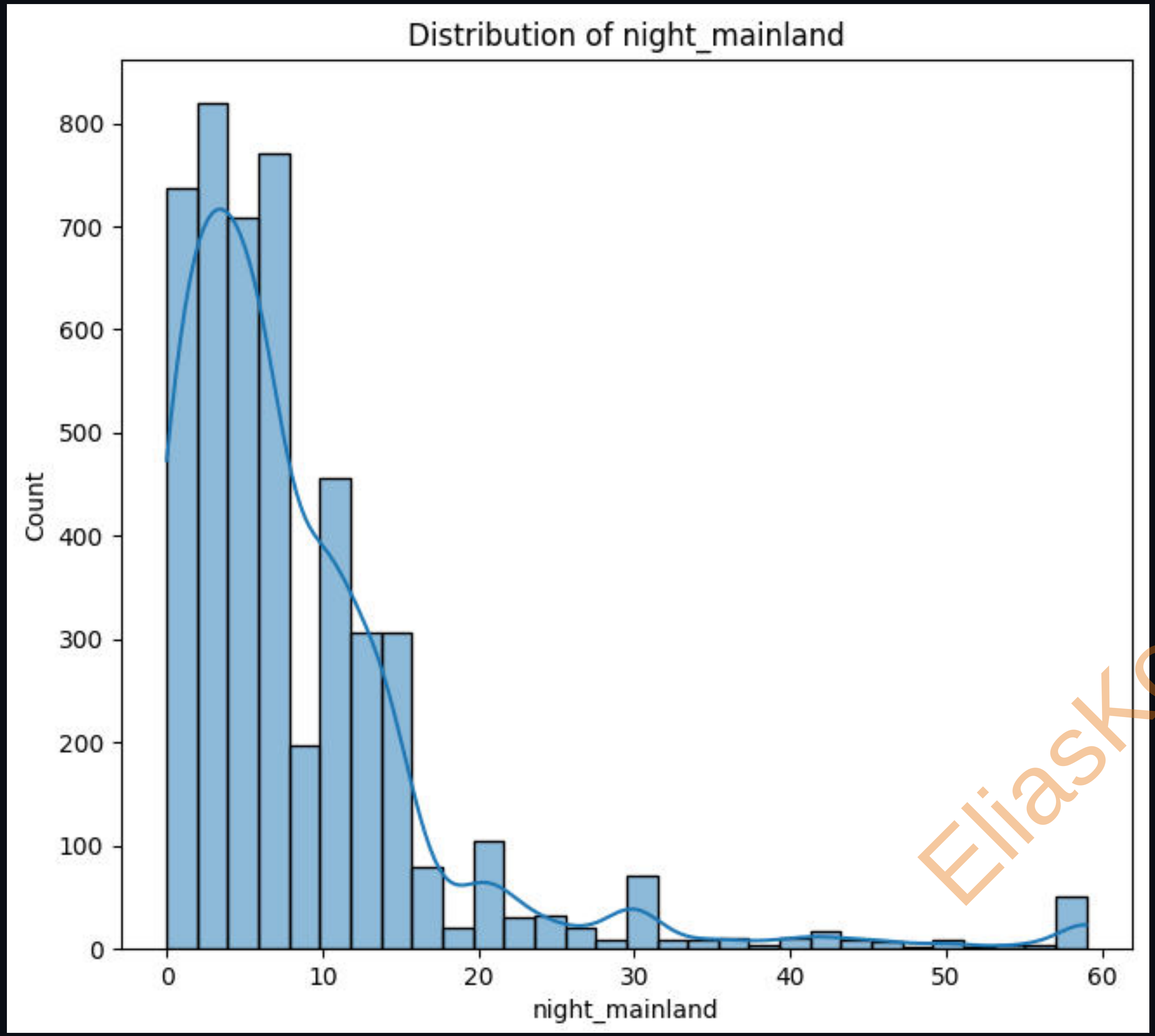
EliasKouloures.com

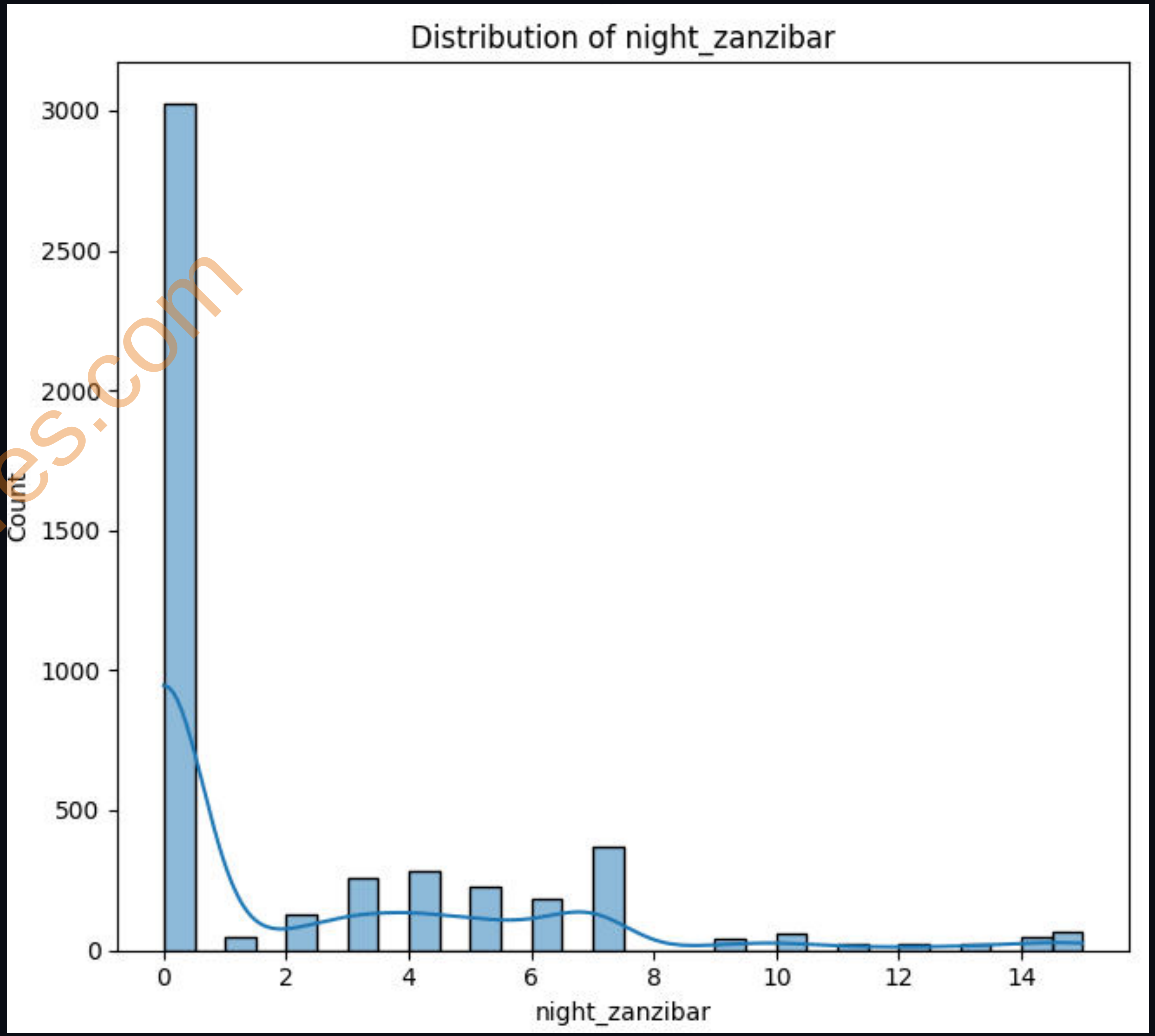
Observation: Most tourists travel alone ➡️ outliers up to 49 & 44.

Pitfall: Outliers affect linear models ➡️ sensitive to data range.

Solution: Robust scaling & outlier capping ➡️ females 0–5 & males 0–4 ➡️ reduces outlier impact.

Most stay <8 Nights on Mainland & 2 Zanzibar





EliasKouloures.com

Observation: Most tourists stay few nights ➡️ outliers up to 145 & 61.

Pitfall: Large values skew understanding of spending by night.

Solution: Robust scaling ➡️ nights Mainland capped max. 59 & Zanzibar 15 ➡️ reduces skewness.

Most visitors 25-44 years of age

➡️ followed by 45-64

Most travel alone

➡️ or with a spouse

Leisure & holidays

➡️ most common visit reasons

Wildlife tourism dominates

➡️ natural reserves & parks

Most rely on friends & relatives for info

➡️ Word-of-Mouth & SoMe biggest leverage

Majority prefers package tours

➡️ big spending predictor

Most pay cash

➡️ encourage digital payments (better tracking & security)

1st Tanzanian Insights

EliasKouloures.com

Majority of tourists from few countries

Crucial for predicting spending.

Potential source of bias if not handled carefully.

Eliaskouloures.com

Bivariate Analysis – aka – Heatmap

Usually provide insights how 2 variables relate to a 3rd, e.g. spending. ➡️ If linear data.

EliasKouloures.com

```
from
from
from
impor

# Ste
X = d
y = d

X_tra

# Ste
linea
linea

# Ste
y_pre

# RMS
rmse

# R^2
r2 =

# MAP
mape

rmse,
```

✓ 1.6s

RMSE = Root Mean Square Error: 7,286,064.30 TZS

2.768 €

2,911 $

Average magnitude of errors between predicted & observed values.

R^2 = R-Squared: 0.624 = ✋🪙

62.4% prediction variability of model. ➡️ room for improvement.

MAPE = Mean Absolute Percentage Error: 710.69%

High error ➡️ model not accurate ➡️ predictions 710.69% off on average.

Interpretation of Coefficients

High RMSE & MAPE ➡️ linear model not best fit.

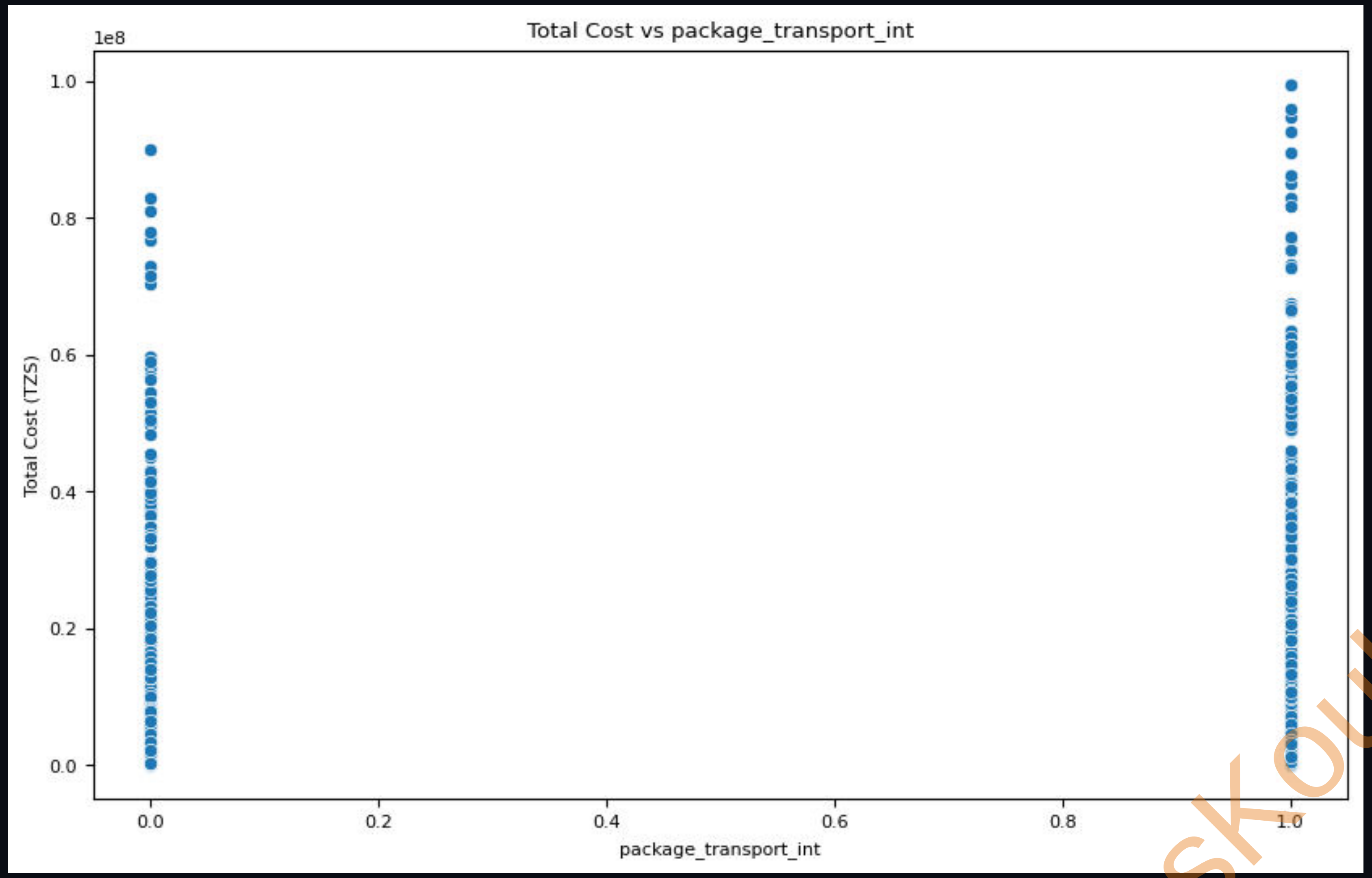
Check scatter plots, if relationships are strongly linear.

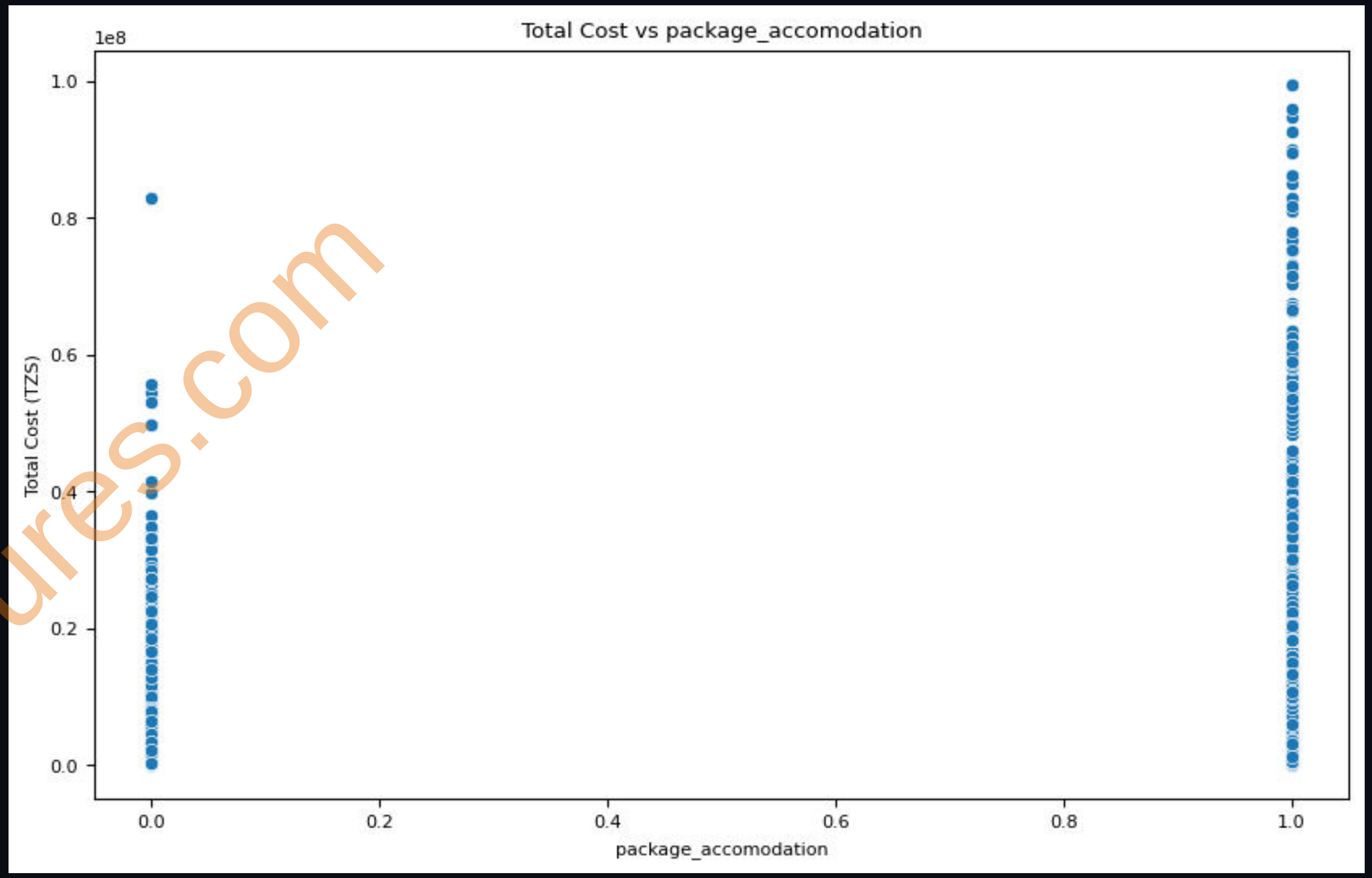
Eliaskouloures.com

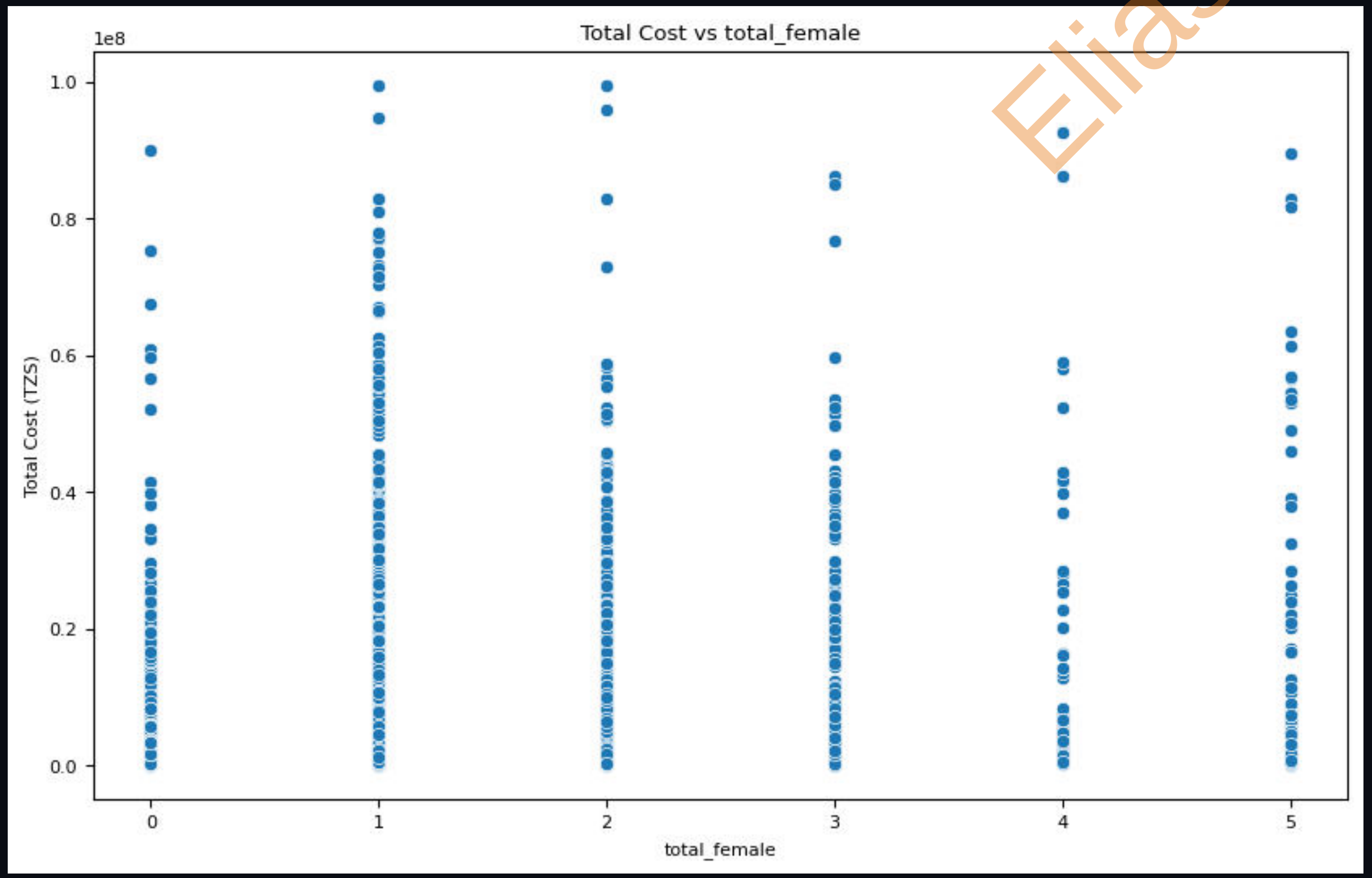
```
(7286064.300077519, 0.6240310522633432, 710.6852970446578)
```

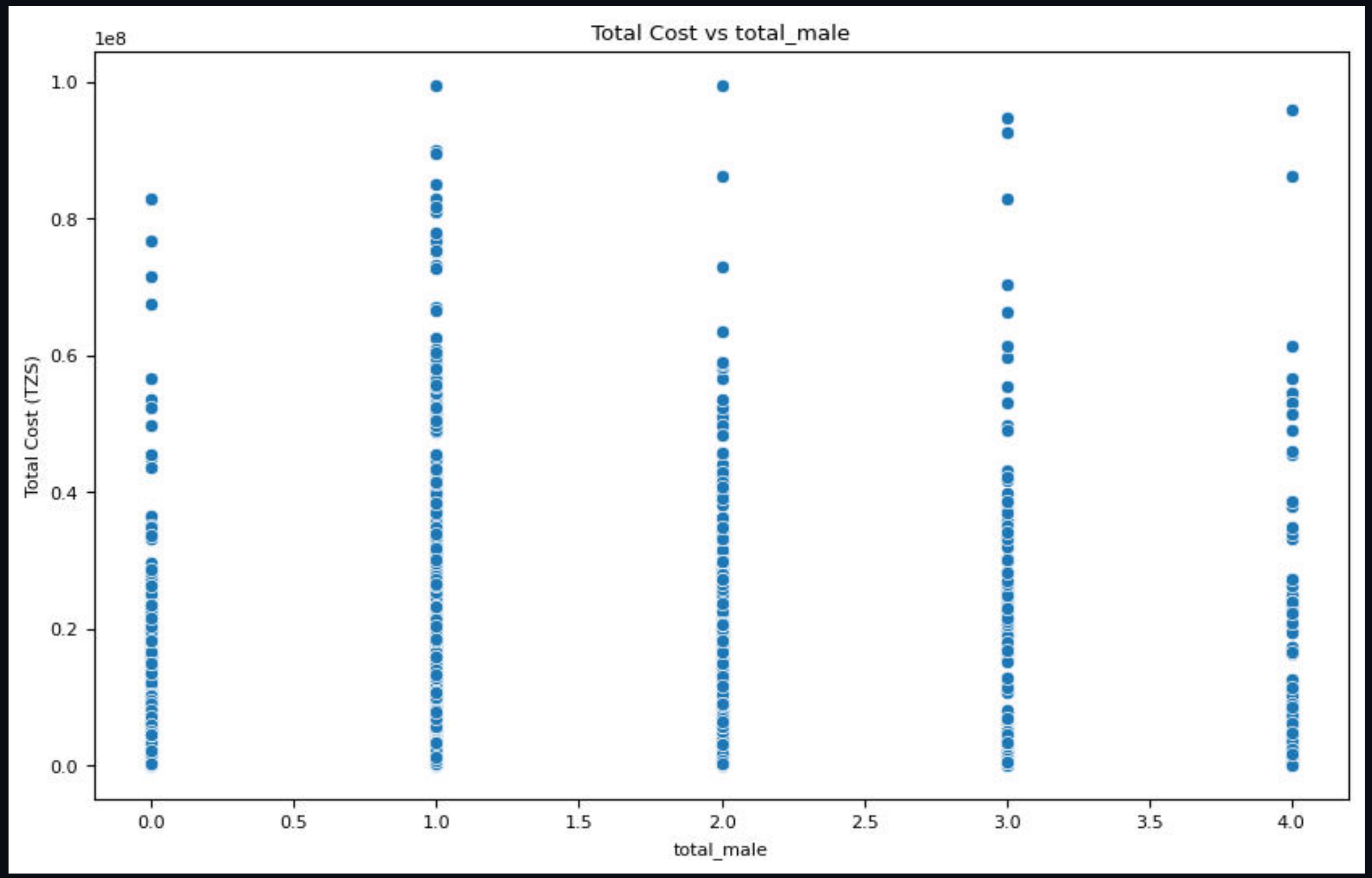
Relationships not strongly linear

Nonlinear models perform better. ➡️ Try RF, SVM & GB.









EliasKouloures.com

Advanced Models & Scientific Analysis

```python
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.svm import SVR
from sklearn.model_selection import cross_val_score

# Initialize models
rf_model = RandomForestRegressor(random_state=42)
svm_model = SVR()
gbm_model = GradientBoostingRegressor(random_state=42)

# Create
models =

# Initia
results =

# Perfor
for mode
    scor
    rmse
    avg_
    r2_s
    avg_
    results.append((name, avg_rmse, avg_r2))

# Convert results to DataFrame for easier interpretation
results_df = pd.DataFrame(results, columns=['Model', 'Average RMSE', 'Average R2'])

results_df
```

	Model	Average RMSE	Average R2
0	Random Forest	1.745866e+05	0.999756
1	Support Vector Machine	1.317062e+07	-0.153609
2	Gradient Boosting	1.220810e+05	0.999895

✓ 34.6s

Python

EliasKouloures.com

Hyperparameter Tuning

Max. Depth of Trees:
- Limits # of splits
- Lower value ➡️ underfitting
- High value ➡️ overfitting

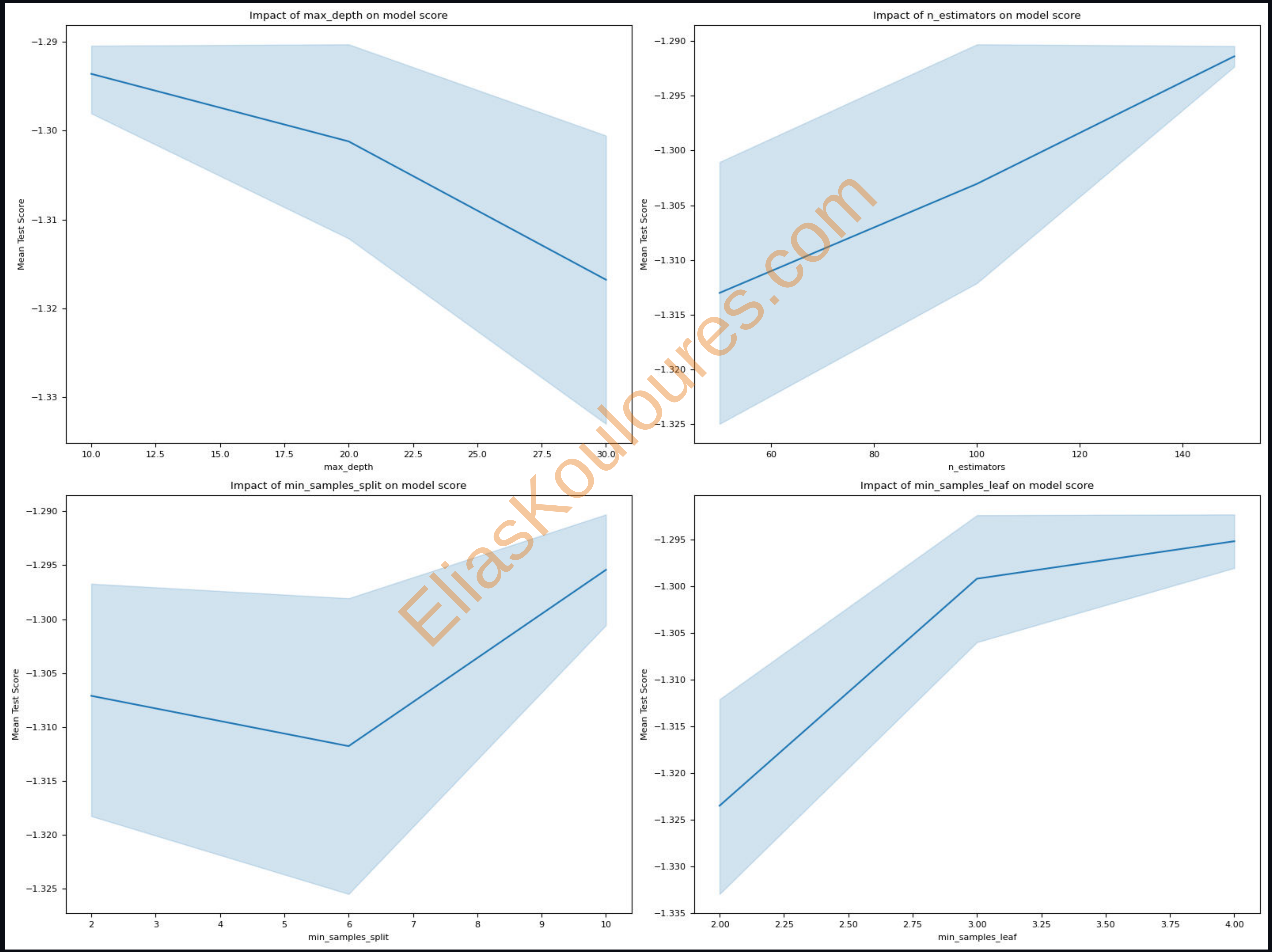
Observations:
- As depth increases ➡️ score decreases
- Implies better performance (RMSE or log loss)
- Uncertainty widens ➡️ depth increases
- Indicating greater variance ➡️ higher depths.

Implications:
- Greater depths fit data more closely
- Increasing variance ➡️ indicator of overfitting
- Score stabilizes after certain depth
- Diminishing returns beyond a certain depth

Min_samples_split:
- Minimum sample number to split a node
- Higher values prevent from learning fine-grained patterns (potentially noise) in the training data.

Observations:
- Model improves ➡️ score decreases
- Up to a certain value and then levels off
- Increasing split ➡️ uncertainty decreases

Implications:
- Increasing split ➡️ prevents overfitting by not splitting nodes with very few samples
- After certain point ➡️ model might become too generalized or underfitted ➡️ score plateau
- Reducing variance suggests that higher values of min_samples_split lead to more stable models



EliasKoulOures.com

N_estimators:
- # of trees in Random Forest
- Higher value ➡️ more trees
- More robust model ➡️ increases computation

Observations:
- Increasing # ➡️ better score ➡️ decreasing number
- Uncertainty remains consistent

Implications:
- Adding trees increases robustness of model
- Certain point ➡️ diminishing returns
- Consistent variance ➡️ stable model performance

Min_samples_leaf:
- Minimum sample number required in leaf node
- Higher values ➡️ deter overfitting

Observations:
- Score decreases ➡️ model improves
- Plateau after a certain point
- Uncertainty relatively stable throughout

Implications:
- Increasing samples ➡️ robuster model
- Ensures leaves are meaningful data representation
- Too high values ➡️ can underfit
- Consistent variance ➡️ parameter impact consistent

1. Best Model

Random Forest & Gradient Boosting. Support Vector Machine performs poorly ➡️ discard for this analysis.

2. Overfitting

High R-Squared for RF & GB indicates overfitting. ➡️ Validate model on unseen data. ✅

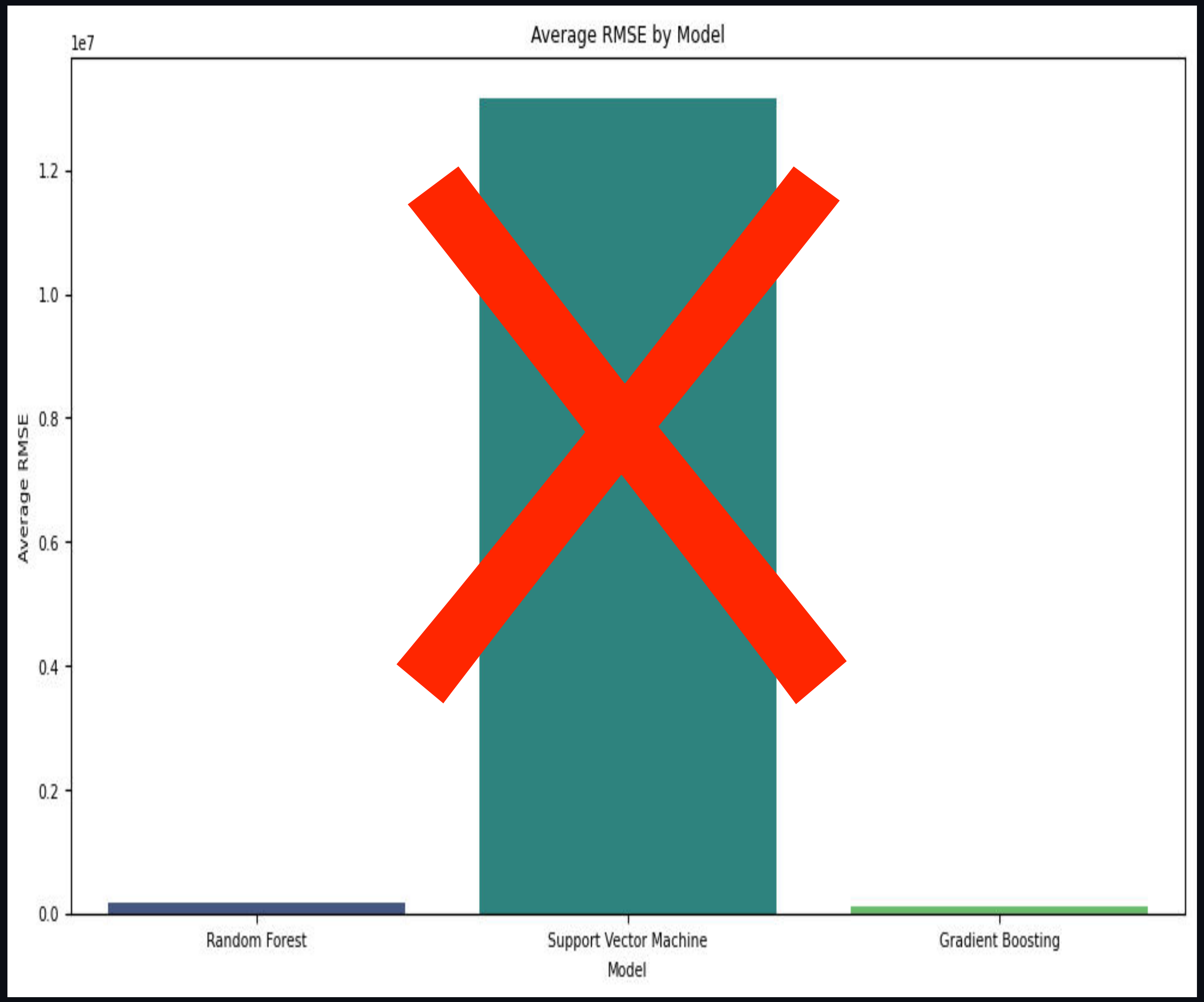
3. Complexity

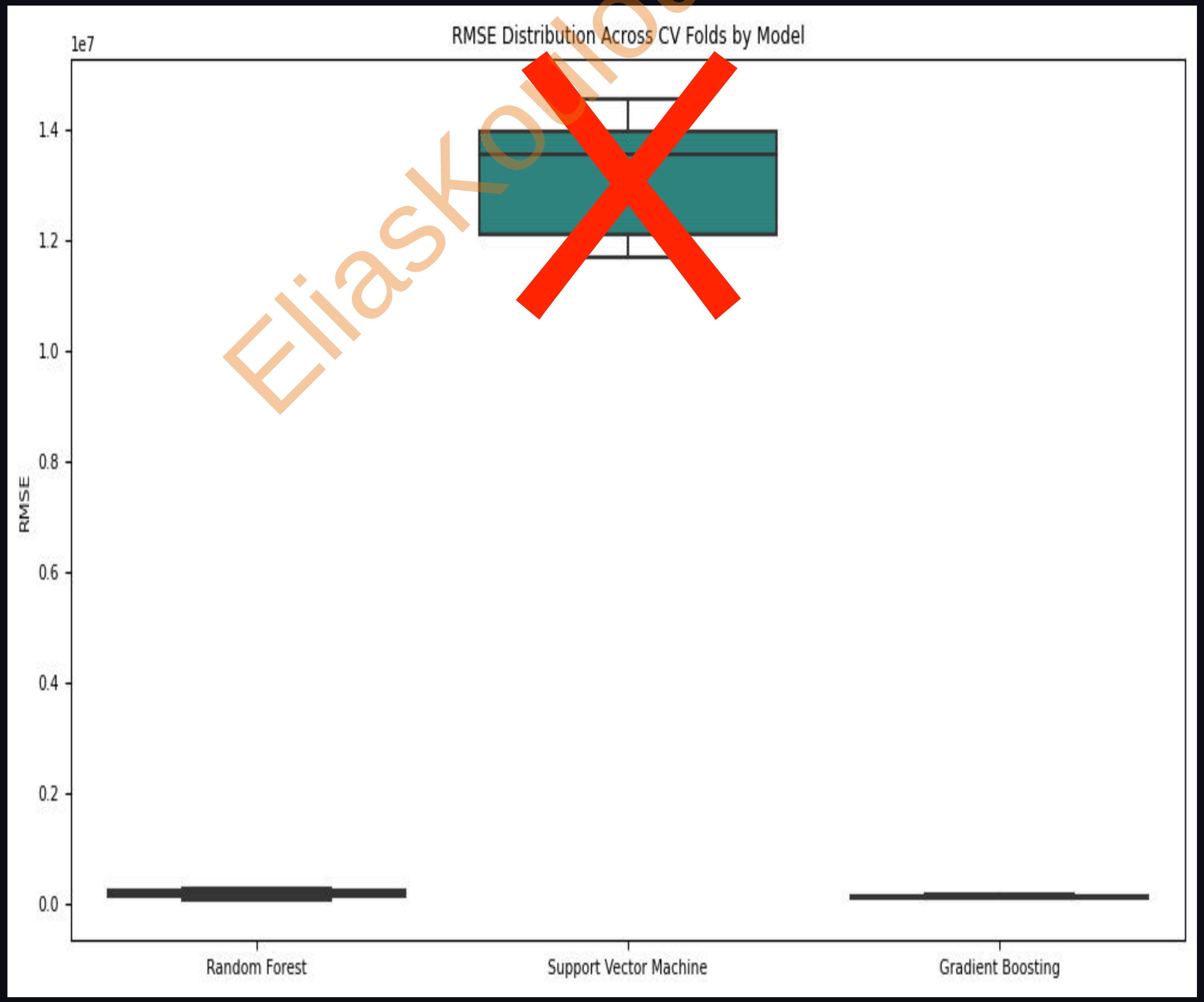
RF & GB are ensemble models ➡️ more complex than linear ➡️ more computation effort ➡️ perform better with non-linear data.

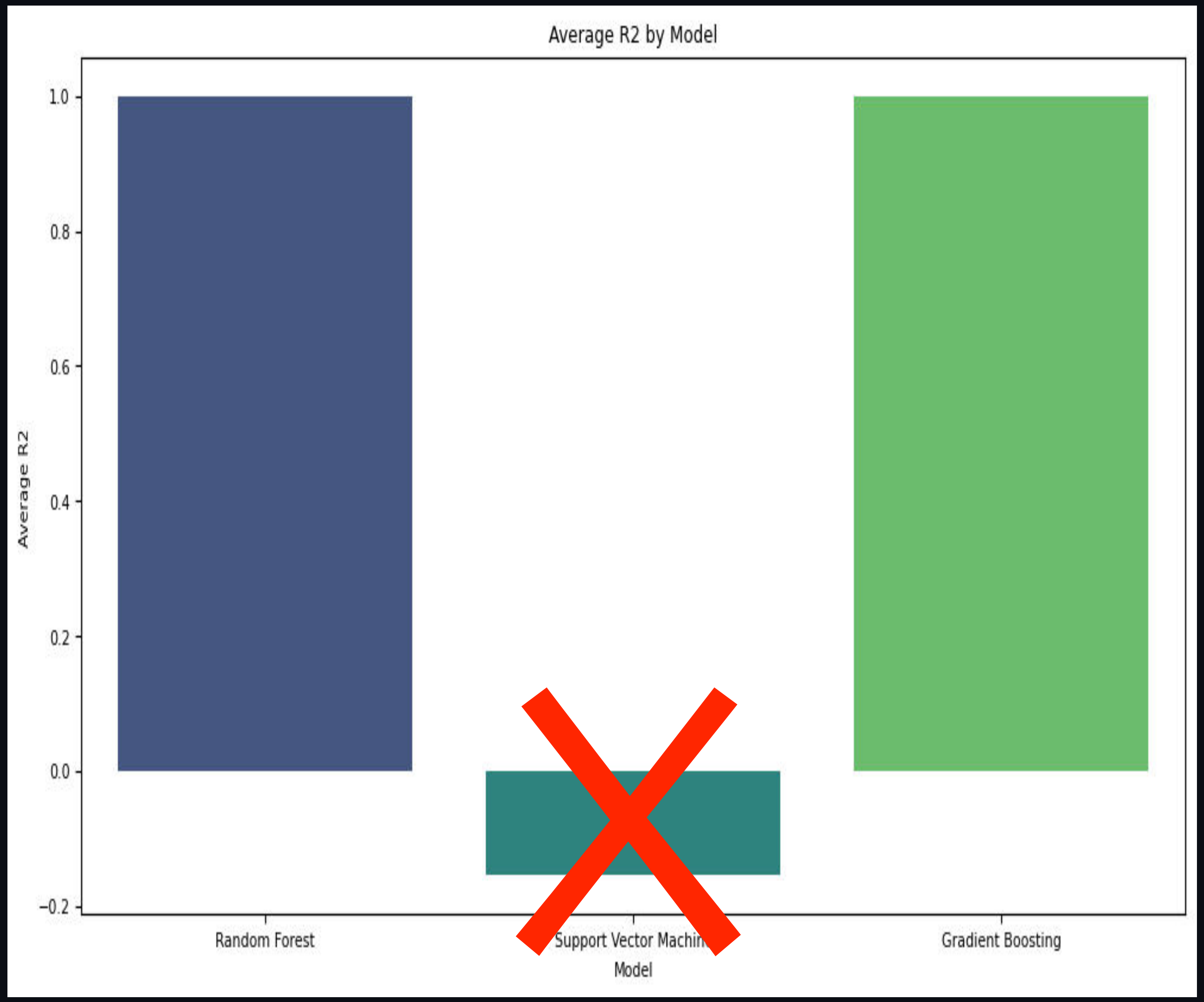
4. Insights

Before ML deployment – esp. with near-perfect metrics ➡️ critical to understand feature importances & ensure objectives alignment.

EliasKoulloures.com





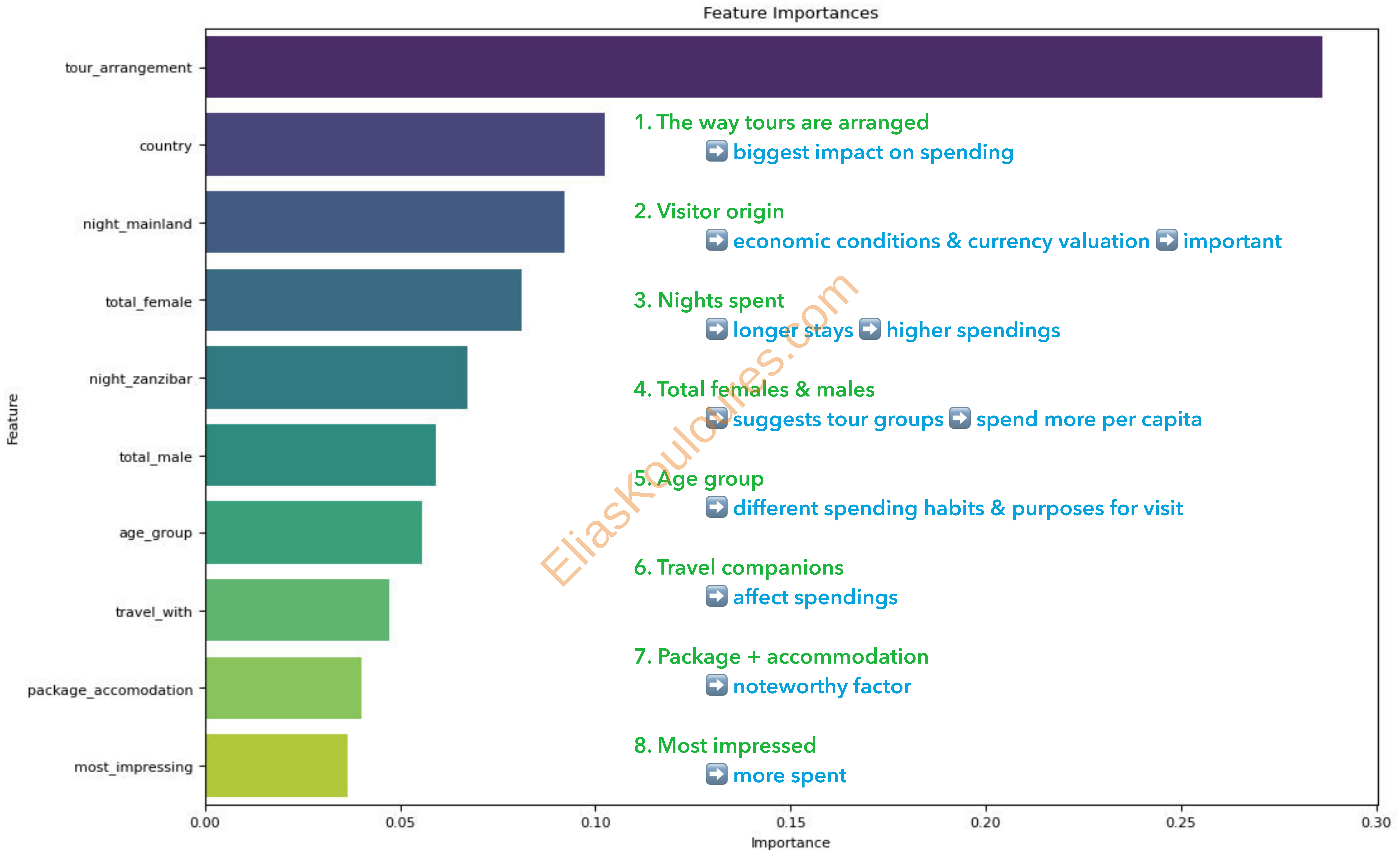


```python
# Calculate the performance metrics for the best Random Forest model
best_rf_model = RandomForestRegressor(**best_rf_params, random_state=42)
best_rf_scores = cross_val_score(best_rf_model, X_train, y_tra
best_rf_rmse_scores = np.sqrt(-best_rf_scores)
best_rf_avg_rmse = np.mean(best_rf_rmse_scores)
best_rf_r2_scores = cross_val_score(best_rf_model, X_train, y_
best_rf_avg_r2 = np.mean(best_rf_r2_scores)

# Update the results DataFrame with the new Random Forest mode
results_df.loc[results_df['Model'] == 'Random Forest', 'Averag
results_df.loc[results_df['Model'] == 'Random Forest', 'Averag

# Create a bar plot for average RMSE and R2 metrics
plt.figure(figsize=(12, 6))
sns.barplot(x='Model', y='Average RMSE', data=results_df, pale
plt.title('Average RMSE by Model')
plt.show()

plt.figure(figsize=(12, 6))
sns.barplot(x='Model', y='Average R2', data=results_df, palett
plt.title('Average R2 by Model')
plt.show()

# Create box plots to show distribution of RMSE scores across
rf_cv_scores = cross_val_score(best_rf_model, X_train, y_train
svm_cv_scores = cross_val_score(svm_model, X_train, y_train,
gbm_cv_scores = cross_val_score(gbm_model, X_train, y_train,

cv_scores_df = pd.DataFrame({
    'Random Forest': np.sqrt(-rf_cv_scores),
    'Support Vector Machine': np.sqrt(-svm_cv_scores),
    'Gradient Boosting': np.sqrt(-gbm_cv_scores)
})

plt.figure(figsize=(12, 6))
sns.boxplot(data=cv_scores_df, palette='viridis')
plt.title('RMSE Distribution Across CV Folds by Model')
plt.ylabel('RMSE')
plt.show()
```

✓ 21.0s

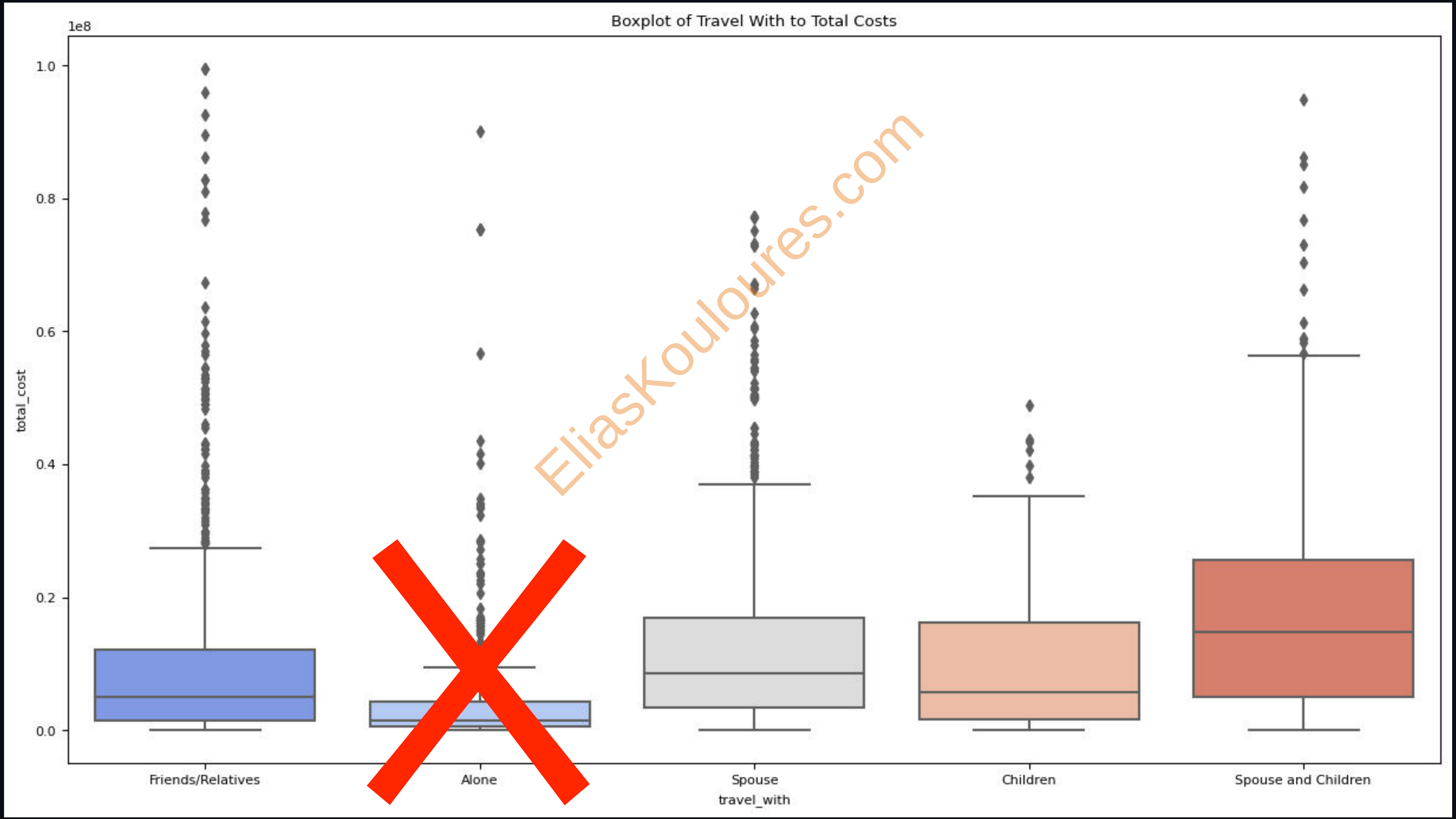
	Feature	Importance
8	tour_arrangement	0.286102
0	country	0.102374
16	night_mainland	0.092218
3	total_female	0.081190
17	night_zanzibar	0.067167
4	total_male	0.059006
1	age_group	0.055510
2	travel_with	0.047351
10	package_accomodation	0.039959
20	most_impressing	0.036648

EliasKouloures.com


Feature Importances
tour_arrangement
country
night_mainland
total_female
night_zanzibar
total_male
age_group
travel_with
package_accomodation
most_impressing
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance
1. The way tours are arranged
➡️ biggest impact on spending
2. Visitor origin
➡️ economic conditions & currency valuation ➡️ important
3. Nights spent
➡️ longer stays ➡️ higher spendings
4. Total females & males
➡️ suggests tour groups ➡️ spend more per capita
5. Age group
➡️ different spending habits & purposes for visit
6. Travel companions
➡️ affect spendings
7. Package + accommodation
➡️ noteworthy factor
8. Most impressed
➡️ more spent
EliasKouloures.com

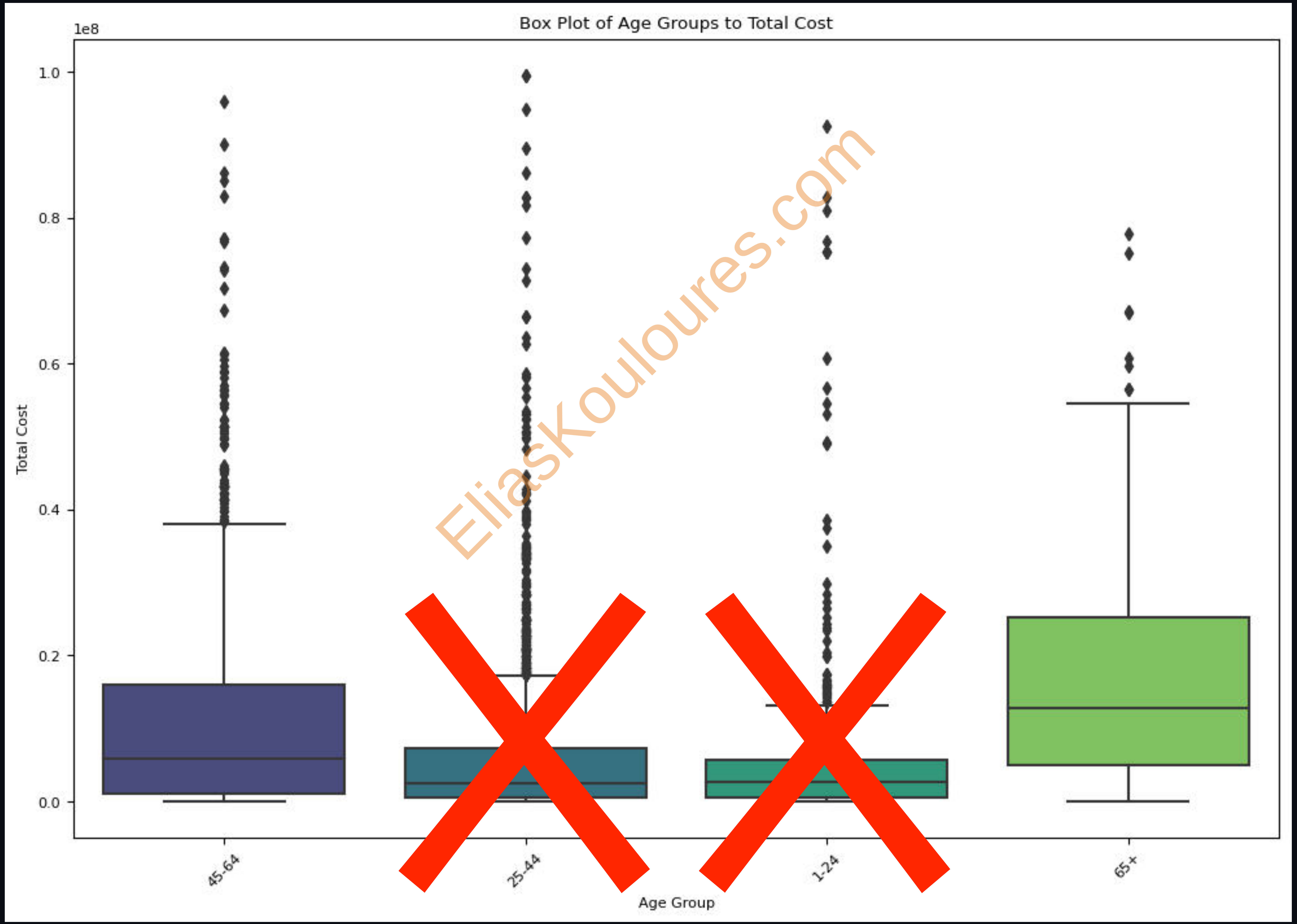
Spendings by Travel Companions

Focus on families, couples & single parents ➡️ maybe friends ➡️ ignore lone travellers.



Spending Distribution for Age Groups

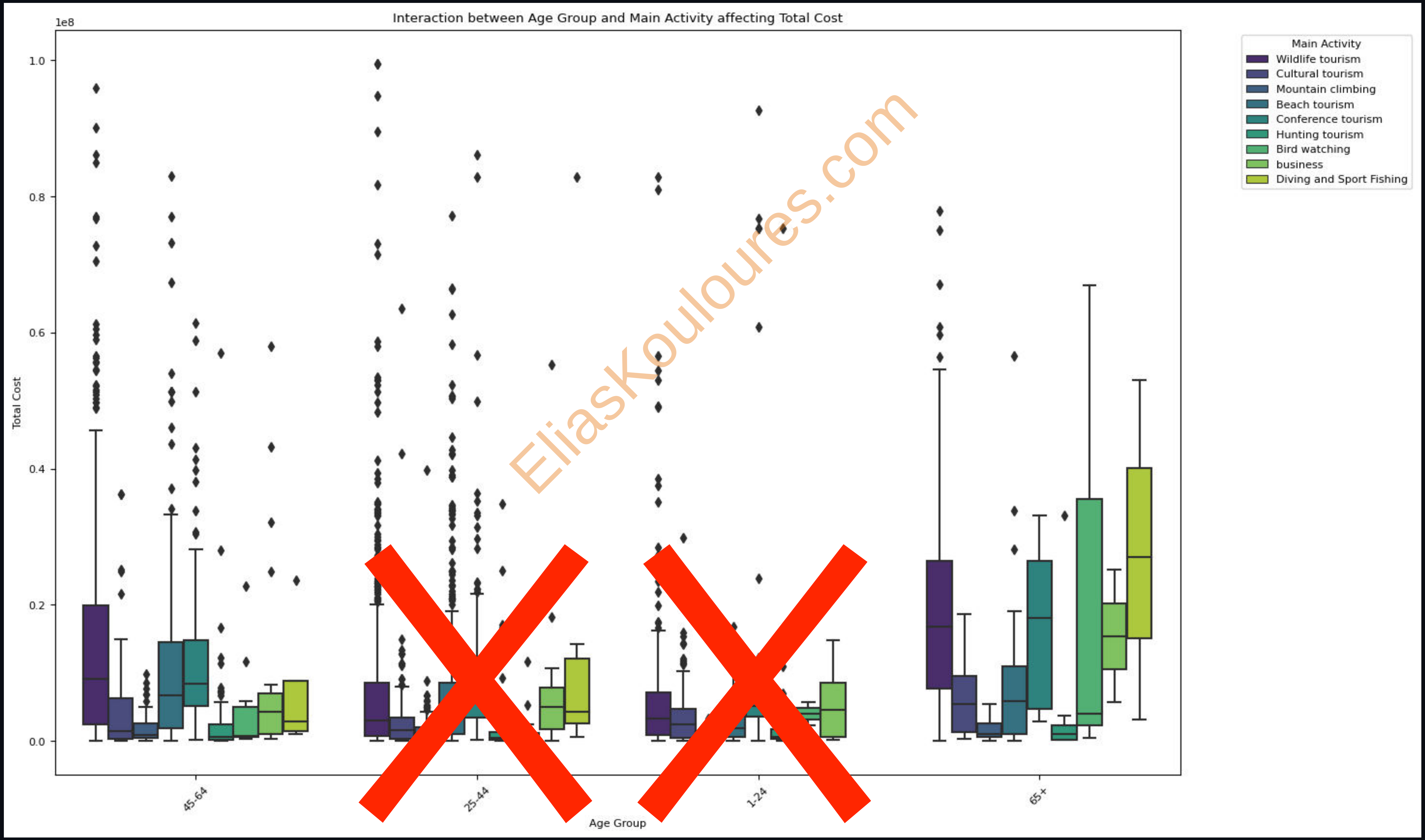
The older the tourists ➡️ the higher AND more predictable their spendings.



Activities & Spendings for Age Groups

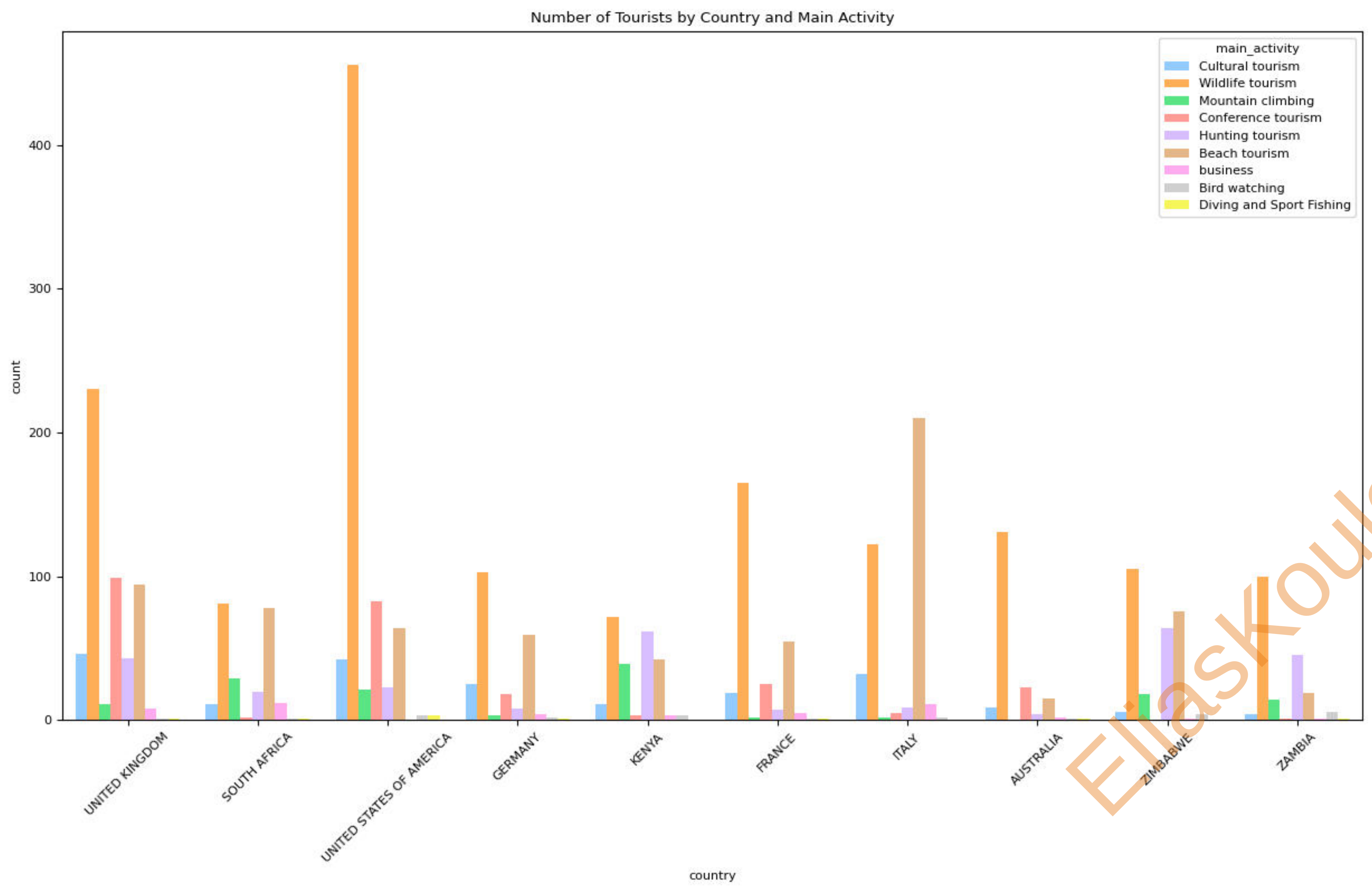
45-64 ➡️ wildlife, beach, business & conferences.

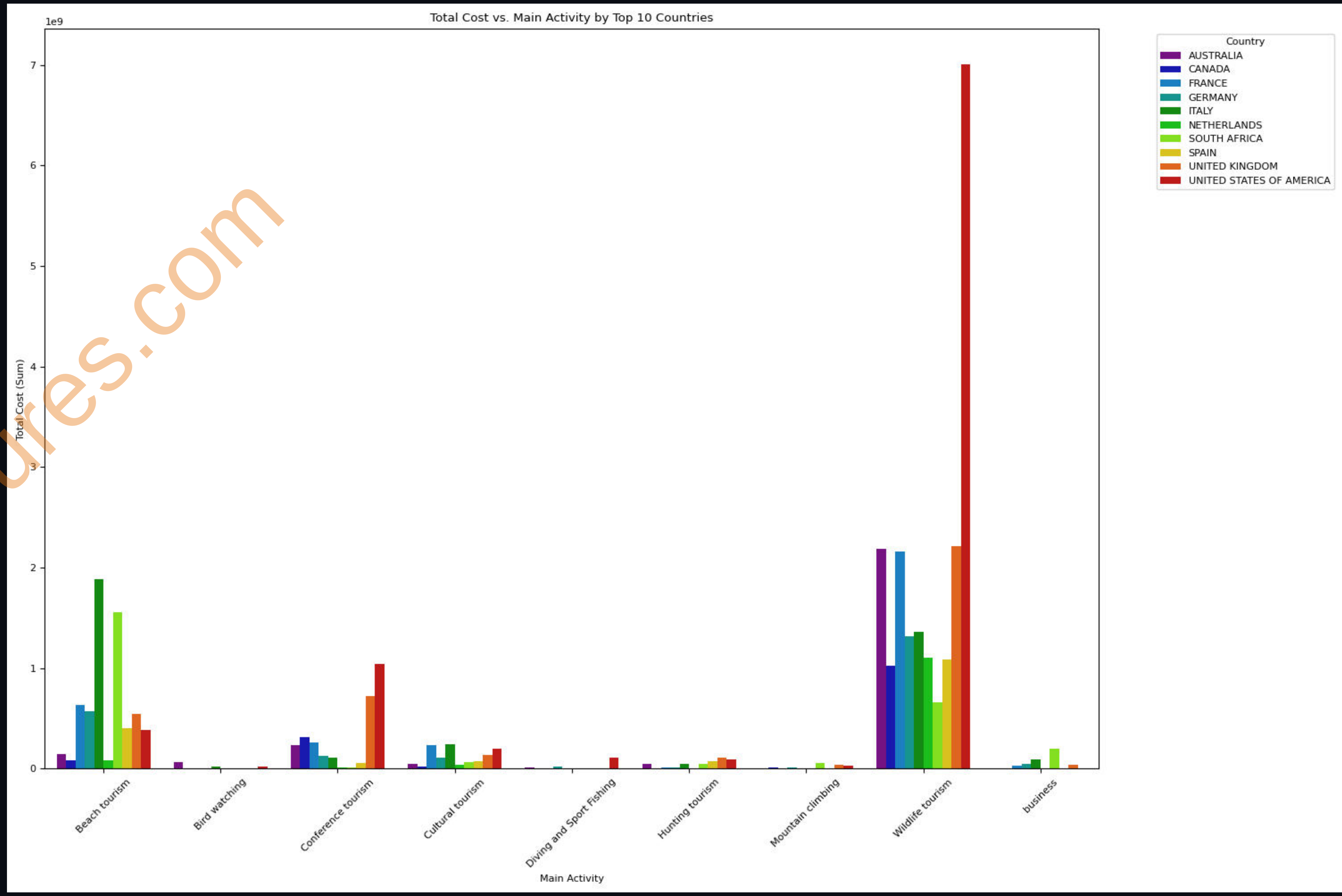
65+ ➡️ wildlife, bird watching, diving, sports fishing, business & conferences.



Top 10 Activity by Origin: Doing vs. Spending

What tourists did. ↔️ How much they spent on it.





Eliaskouloures.com

USA: Most overall + most Wildlife + most $$$ wildlife & conferences ➡️ upsell all non-wildlife activities + advertise on Conferences!

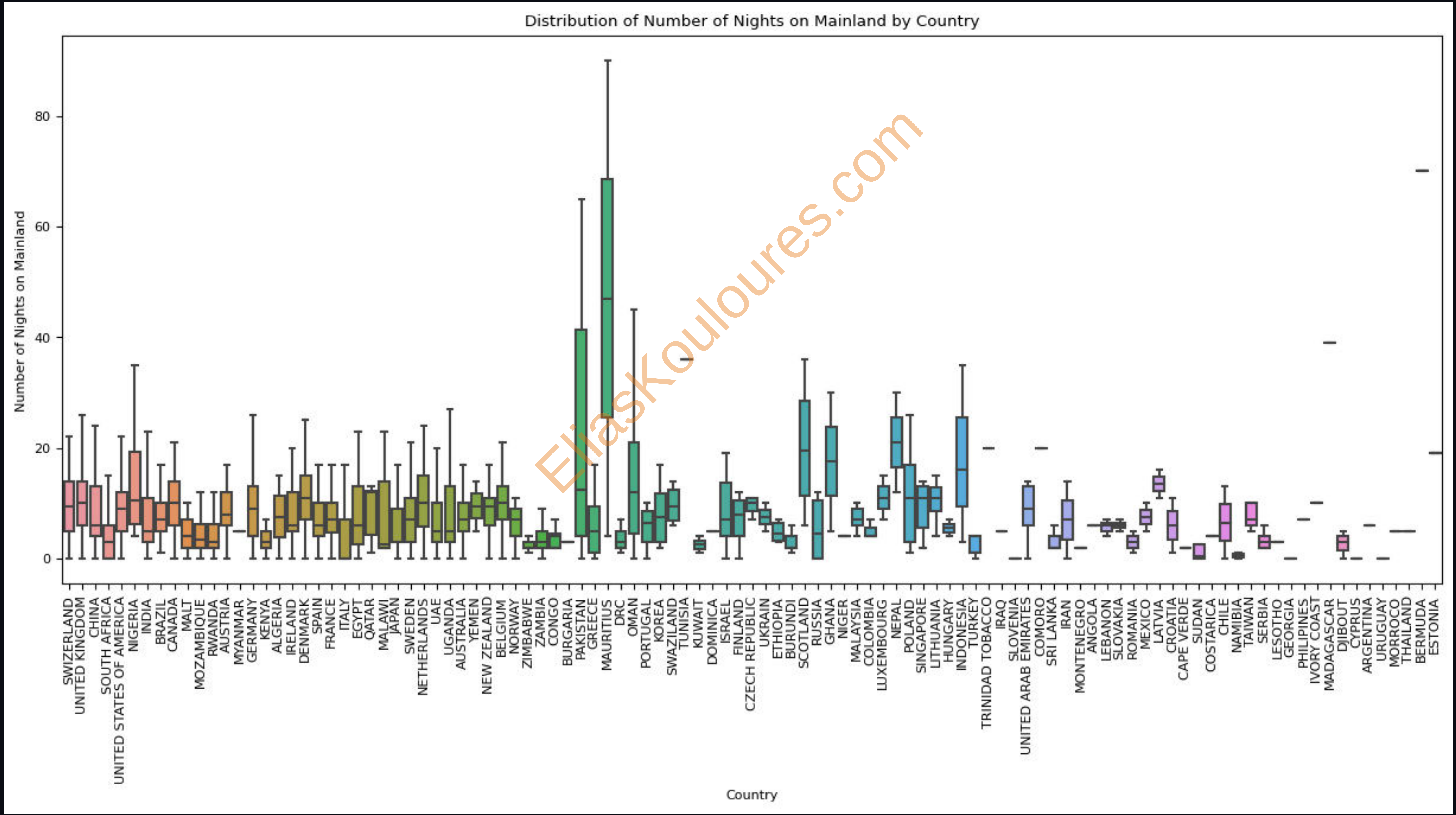
UK: 2nd overall + 2nd wildlife ➡️ below average $$$ wildlife + Beach ➡️ encourage spending + upsell other activities!

Italiens: 3rd overall + most Beach visitors + $$$ + 5th biggest wildlife $$$ + non existent otherwise ➡️ encourage other activities!

South Africans: 2nd biggest Beach group + $$$ + biggest conference $$$ ➡️ create packages & promos + advertise conferences!

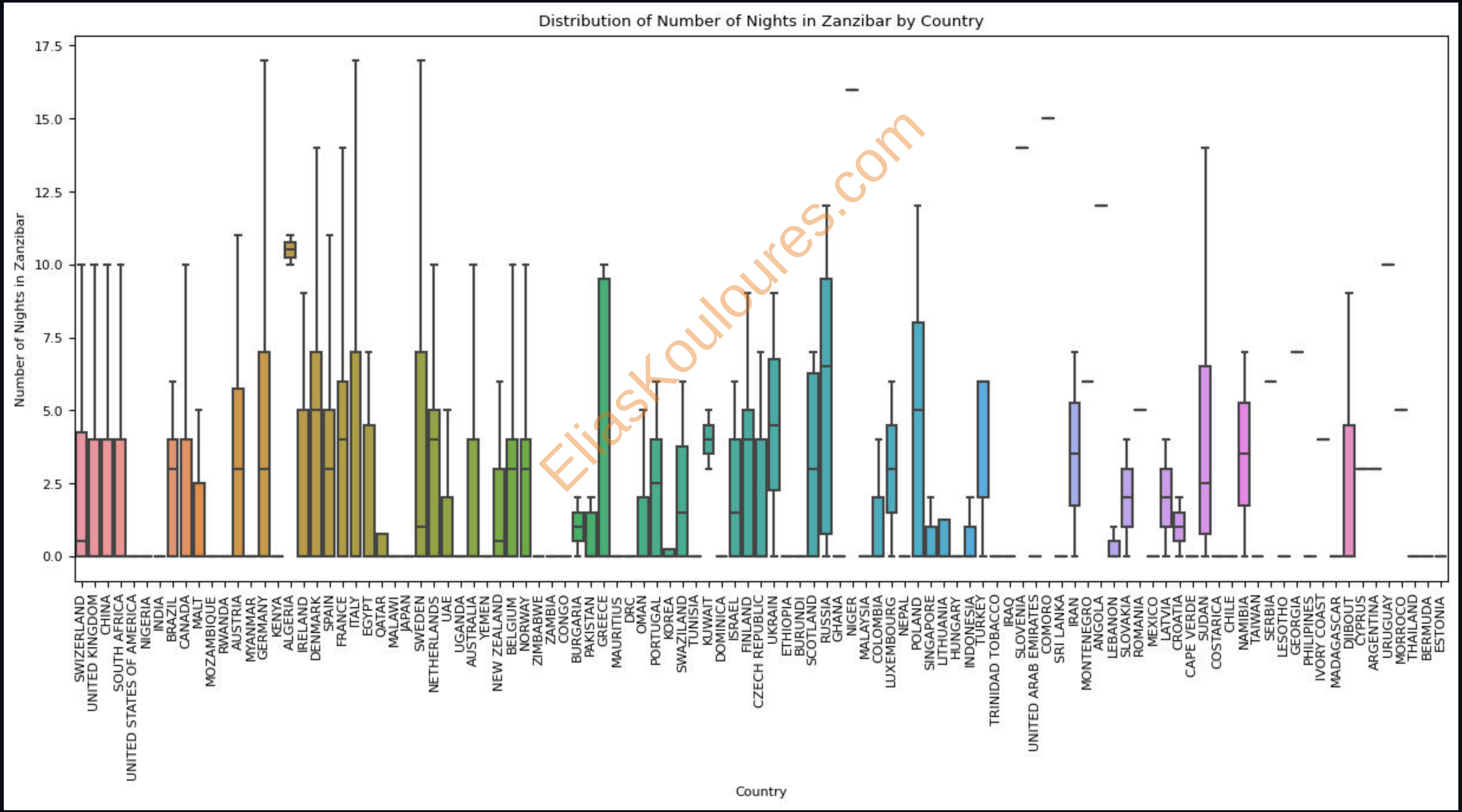
Nights spent on mainland – by Country

Longest: Mauritius, Oman, Pakistan, Scotland, Indonesia, Ghana & Nepal ➡️ Encourage spending.



Nights spent on Zanzibar – by Country

Longest: Greece, Poland, Sweden, Italy, Denmark & Germany ➡️ Advertise in & encourage SoMe posts.



1st Action-Item
Child-friendly Tanzania

Encourage single parent & family visits:

- **Offer child-care services,** e.g. kids clubs, family rooms, babysitter & special activities
- **Create Social Media campaigns,** e.g. show (single) parents with kids enjoying Tanzania
- **Develop new services,** e.g. family & single parent vacation packages, kid discounts, and interactive trip planning tools for kids

EliasKoulourès.com

2nd Action-Item Business People Promo

Allure business visitors to book vacations:

- Create VR experiences, e.g. promos at Tanzanian business congresses, and online via business websites & LinkedIn
- Create tailored offers – packages to come back with family, kid or & partner
- Offer „Alpha Animal Adventures" – Premium Safaris incl. Big-5 sightings & Networking events with local businesspeople

EliasKouloures.com

3rd Action-Item
Deluxe Lifestyle Web Series

Tourism data shows many spending outliers visiting Tanzania – aka – rich people.

We create a web series showcasing deluxe events in premium locations, e.g. Zanzibar.

We invite VIPs & SoMe influencers from high GDP countries to meet Tanzanian stars, artists, musicians, etc.

EliasKouloures.com

4th Action-Item
Beyond Big-5 Offers

Encourage more bookings & upsell tourists:

- **Highlight underutilised activities & sights,** e.g. scuba diving, bird-watching, sports fishing, etc.
- **Create new offers,** e.g. Hot Air Ballooning in Serengeti or at Mount Kilimanjaro, Cultural Immersion Workshops, Artisan Market visits, private dinners in Ngorongoro Crater, etc.

EliasKouloures.com

5th Action-Item Target Profitable Travellers

Optimise marketing ROI & increase profits by focussing on:

- **Most lucrative tourists,** e.g. couples & groups over 45 years of age, single parents & families.
- **Most lucrative origins,** e.g. USA, UK, Europe, Poland, Australia, Canada, Switzerland & Japan.

EliasKouloures.com

6th Action-Item
Boost Travel Storytelling

Amplify Word-of-Mouth & SoMe sharing by:

- **Create WOW experiences**, e.g. draw paintings/postcards with baby elephants, cage dive with crocodiles, etc.
- **Install „Wi-Fi in the Wild"** – to enable kids without roaming to live-stream on SoMe
- **Launch Social Media Challenges**, e.g. Best Sunset Photo, Dance with Locals, etc.

EliasKouloures.com

Thank you for your time

Any Questions?

EliasKouloures.com

Elias Kouloures

Creative Data Scientist, Prompt Engineer & GenAI Expert

📞 +491602448800

💌 elias.kouloures@gmail.com

👀 EliasKouloures.com