USER GUIDE TO HABARI TANZANIA

Habari : /hɑˈbɑ.ri/

(n class, plural habari)

1. the casual way of asking - what is the news?
2. Habari Tanzania - literally what is the news of Tanzania; The local way of asking - Hello Tanzania, how are you?

1 Dashboard

An overview of key figures. Users can explore tourists and spending from a map view and gather key statistics of the top countries.

2 Data Analysis

Spending Behaviour

a Compare correlation between individual spending and nights spent. Option to group by:

Categories

b Compare tourist spending across categories.

Identify high-rollers

Regional Overview

Users can compare between regions or top countries of each region. Options include:

a Spending

b Demographics

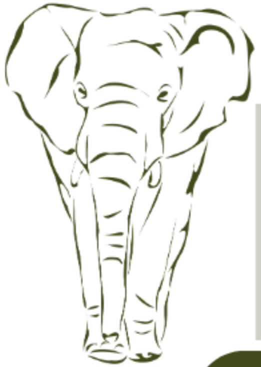
Country Comparison

Users can compare drivers of spending between any 2 countries. Options to filter for:

a Demographics

b Behaviour traits

3 Clustering



Users can adjust the # of clusters and # of repetitions to find the best cluster to group the behavior of tourists

*Note: If #of repetitions is greater than 1, a global search was done to obtain the lowest Bayesian Information Criterion (BIC) score.

4 Regression Model

Users can select:

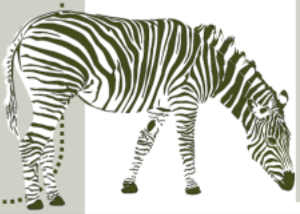
a Preferred variables to improve prediction results

b Train-test partition for inputs to build the model

Random Forest

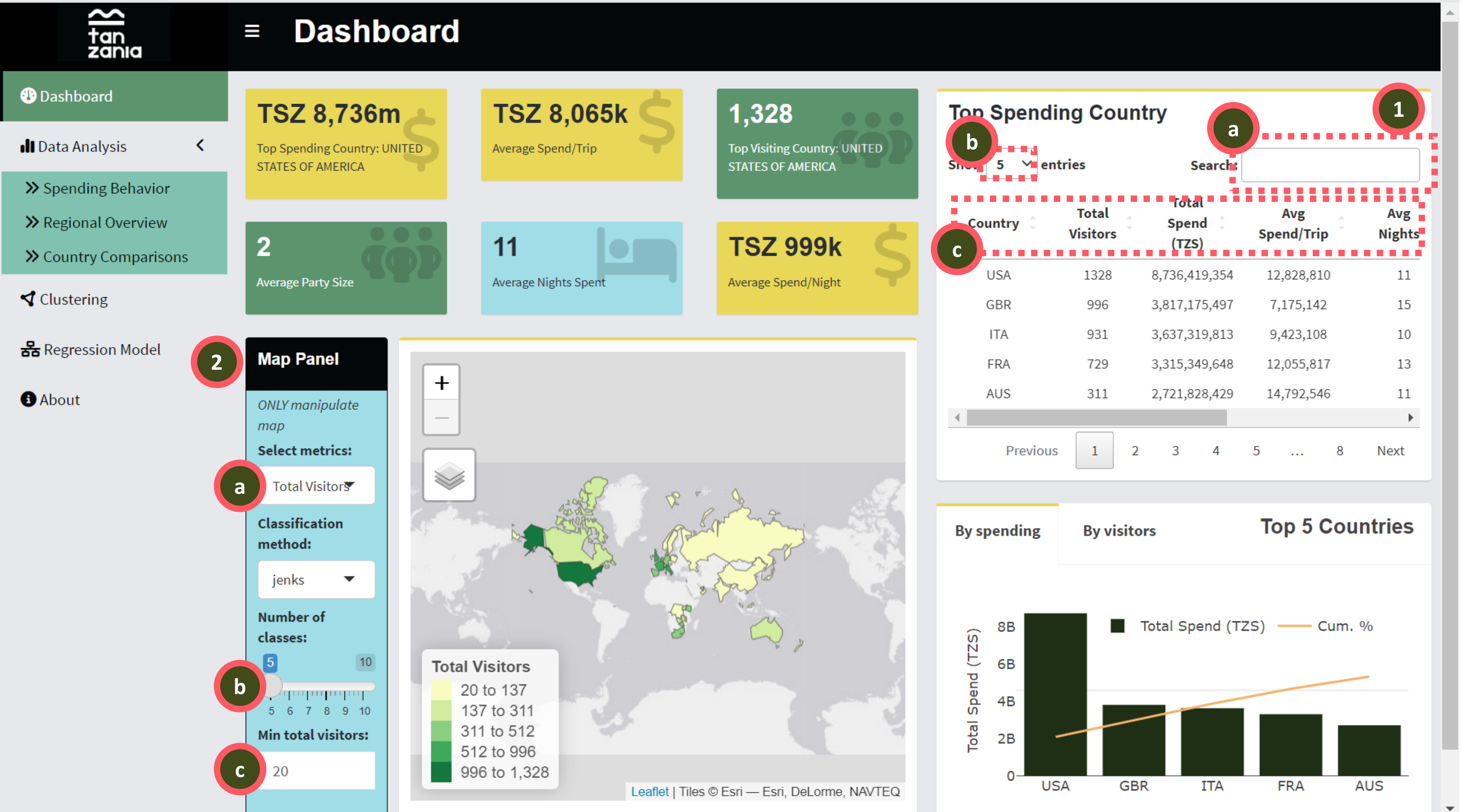
Users can select:

1. Resampling technique preferred
2. # of trees
3. Variable importance mode
4. Split Rule



DASHBOARD

On the landing page is a Dashboard providing an overview of key figures. Users can explore the distribution of tourists and spending from a map view and gather key statistics of the top spending countries.



TSZ 8,736m
Top Spending Country: UNITED STATES OF AMERICA

TSZ 8,065k
Average Spend/Trip

1,328
Top Visiting Country: UNITED STATES OF AMERICA

2
Average Party Size

11
Average Nights Spent

TSZ 999k
Average Spend/Night

Top Spending Country

Show 5 entries

Search:

Country	Total Visitors	Total Spend (TZS)	Avg Spend/Trip	Avg Nights
USA	1328	8,736,419,354	12,828,810	11
GBR	996	3,817,175,497	7,175,142	15
ITA	931	3,637,319,813	9,423,108	10
FRA	729	3,315,349,648	12,055,817	13
AUS	311	2,721,828,429	14,792,546	11

Previous 1 2 3 4 5 … 8 Next

Map Panel

ONLY manipulate map

Select metrics: Total Visitors

Classification method: jenks

Number of classes: 5 (5 6 7 8 9 10)

Min total visitors: 20

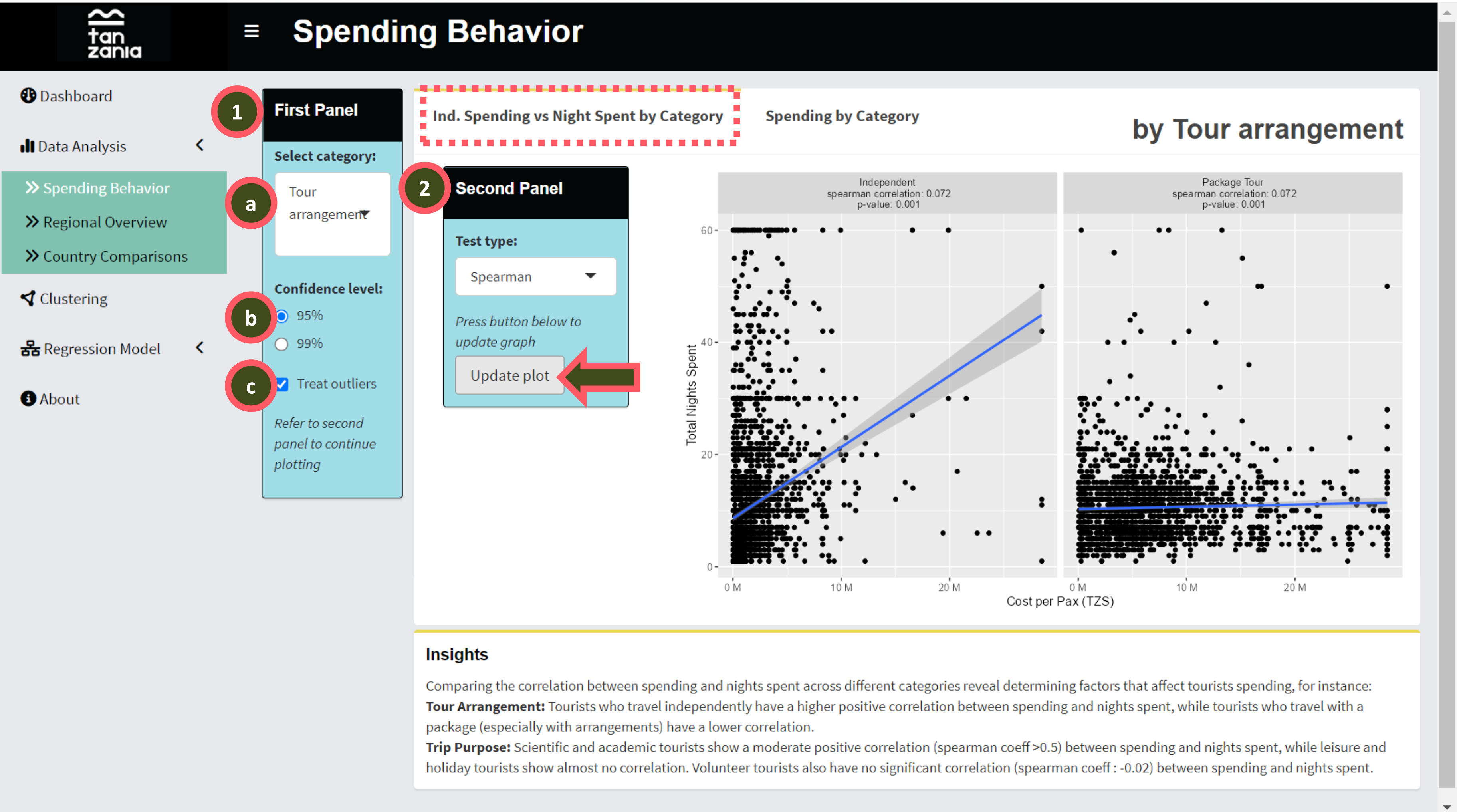
Total Visitors
- 20 to 137
- 137 to 311
- 311 to 512
- 512 to 996
- 996 to 1,328

1. **Top Spending Country:** Users can scroll through the table on Top Spending Countries. Each parameter takes in only one input.
   a. Users can search for a particular country through the search box.
   b. Users can use the dropdown box to choose whether to display 5 or 10 entries.
   c. User can click on each variable to sort details by ascending or descending order of that variable.

2. **Map Panel:** Users can visualise continuous metrics on the map. Each parameter takes in only one input.
   a. Users can use the dropdown boxes to select the continuous metric and the classification methodology for visualisation.
   b. Users can use the sliding bar to decide the number of classes.
   c. Users can input a minimum number for classification.

EXPLORE: Data Analysis – Spending Behaviour

This exploration page allows users to compare the correlation between individual spending and nights spends in Tanzania. Users are given the option to group by demographics or behavioural trait categories such as tour arrangement, trip purpose, source of information etc. to identify which are the factors that are more likely to affect tourist spending.



First Panel

Select category: Tour arrangement

Confidence level:
● 95%
○ 99%

☑ Treat outliers

Refer to second panel to continue plotting

Second Panel

Test type: Spearman

Press button below to update graph

Update plot

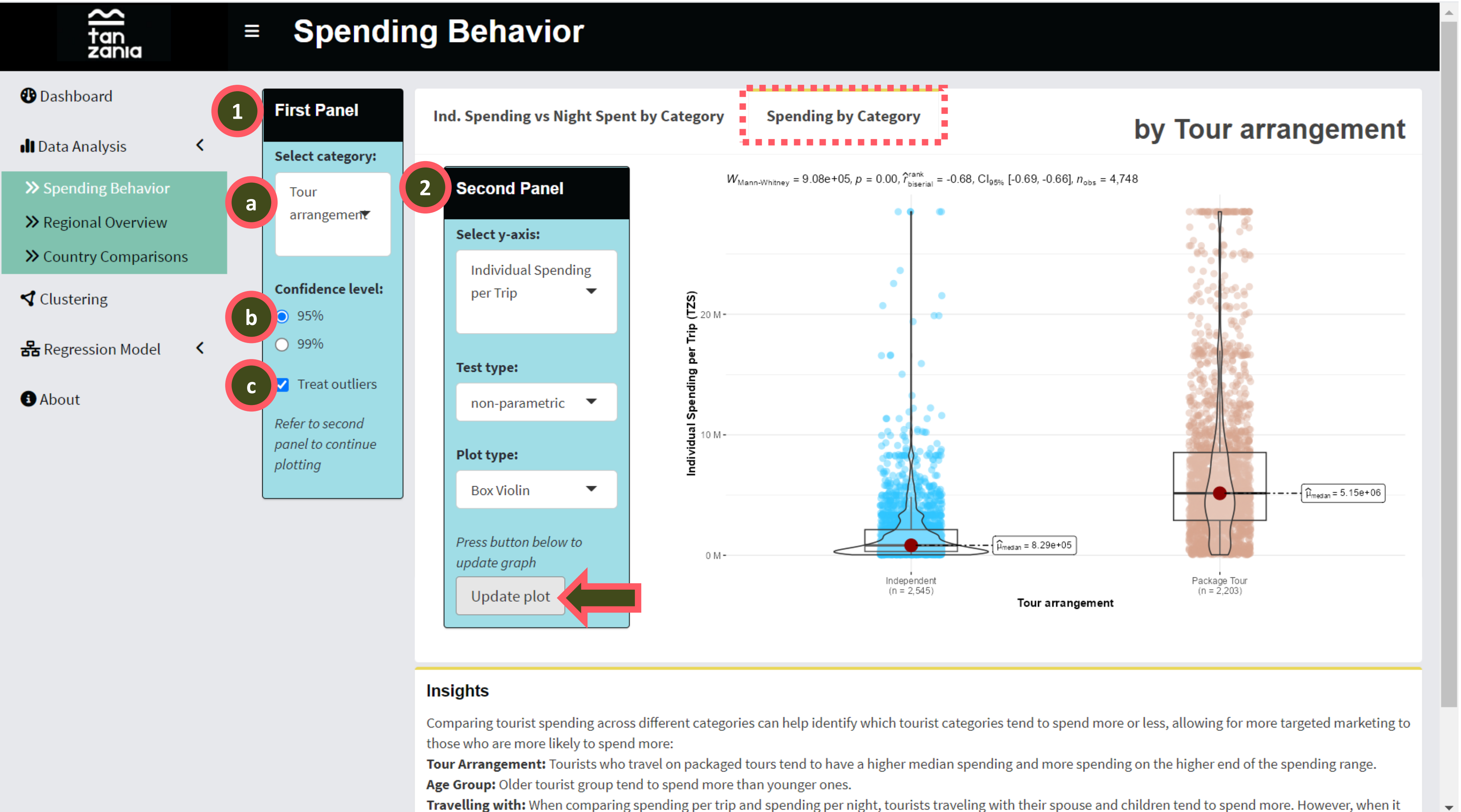
Insights

Comparing the correlation between spending and nights spent across different categories reveal determining factors that affect tourists spending, for instance:
Tour Arrangement: Tourists who travel independently have a higher positive correlation between spending and nights spent, while tourists who travel with a package (especially with arrangements) have a lower correlation.
Trip Purpose: Scientific and academic tourists show a moderate positive correlation (spearman coeff >0.5) between spending and nights spent, while leisure and holiday tourists show almost no correlation. Volunteer tourists also have no significant correlation (spearman coeff : -0.02) between spending and nights spent.

1. **First Panel:** Users can perform confirmatory data analysis on the correlation between individual spending per trip vs. total nights spent. Each parameter takes in only one input.
   a. Users can use the dropdown boxes to select the demographics or behavioural trait category to analyze.
   b. Users can use the radio buttons to select the confidence level.
   c. Users can select the check box “Treat outliers” if they would like the app will use data which has been treated for outliers.
2. **Second Panel:** Users can use the dropdown boxes to select the type of correlation test to perform. Each parameter takes in only one input. The charts will appear after users click on the “Update plot” button.

EXPLORE: Data Analysis – Spending Behaviour

This exploration page allows users to compare tourist spending across various demographics or behavioural trait categories. Such comparisons can help users identify which tourist groups / demographics tend to spend more, allowing for more targeted marketing.



First Panel

Select category: Tour arrangement

Confidence level:
● 95%
○ 99%

☑ Treat outliers

Refer to second panel to continue plotting

Second Panel

Select y-axis: Individual Spending per Trip

Test type: non-parametric

Plot type: Box Violin

Press button below to update graph

Update plot

$W_{\text{Mann-Whitney}}$ = 9.08e+05, p = 0.00, $\hat{r}^{\text{rank}}_{\text{biserial}}$ = -0.68, $\text{CI}_{95\%}$ [-0.69, -0.66], n_{obs} = 4,748

Insights

Comparing tourist spending across different categories can help identify which tourist categories tend to spend more or less, allowing for more targeted marketing to those who are more likely to spend more:

Tour Arrangement: Tourists who travel on packaged tours tend to have a higher median spending and more spending on the higher end of the spending range.

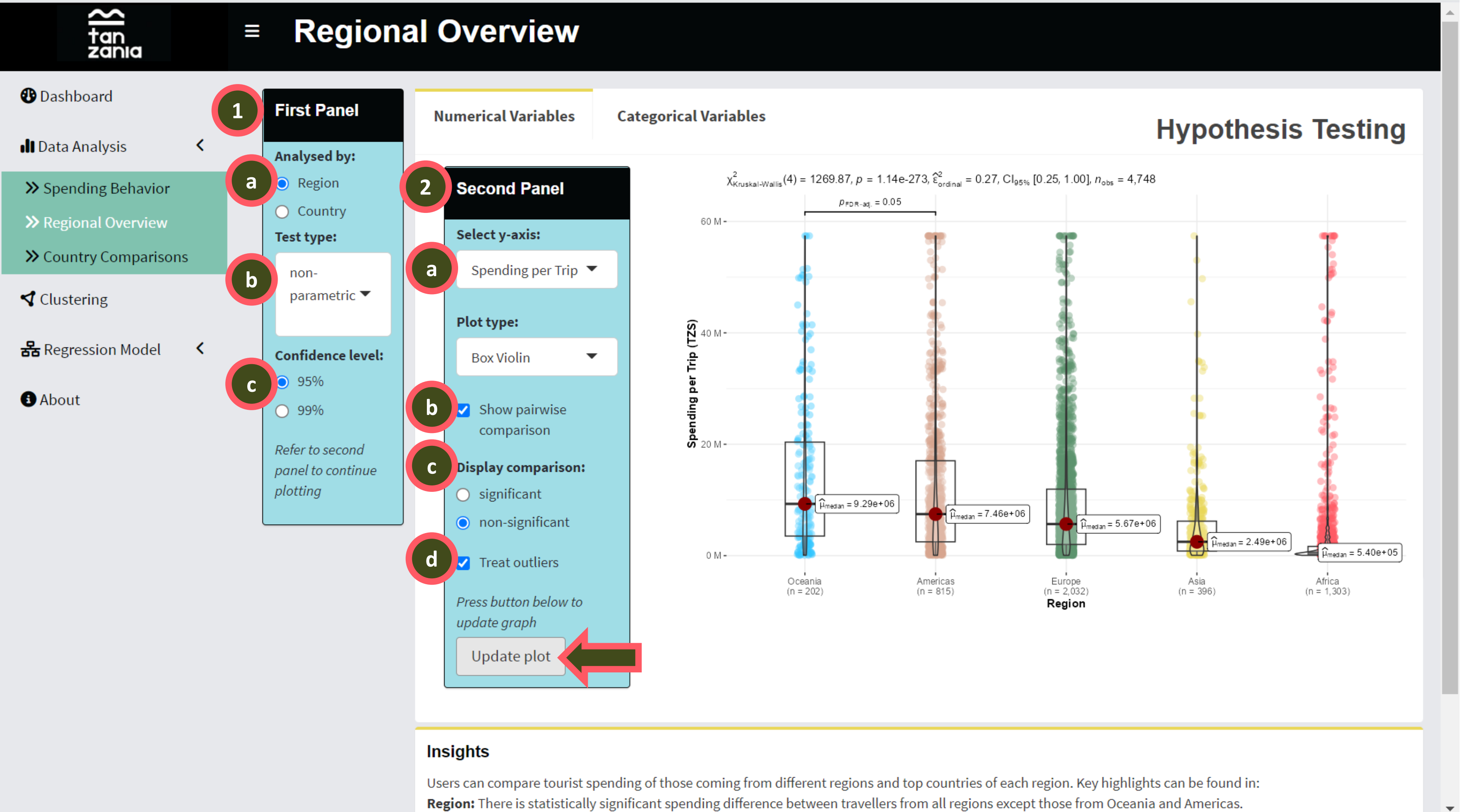
Age Group: Older tourist group tend to spend more than younger ones.

Travelling with: When comparing spending per trip and spending per night, tourists traveling with their spouse and children tend to spend more. However, when it

1. **First Panel:** Users can perform confirmatory data analysis on tourist spending. Each parameter takes in only one input.
   a. Users can use the dropdown boxes to select the demographics or behavioural trait category to analyze.
   b. Users can use the radio buttons to select the confidence level.
   c. Users can select the check box “Treat outliers” if they would like to use data which has been treated for outliers.
2. **Second Panel:** Users can select the spending metric to plot (y-axis), correlation test type and plot type. Each parameter takes in only one input. The charts will appear after users click on the “Update plot” button.

EXPLORE: Data Analysis – Regional Overview

This exploration page allows users to compare trends and perform confirmatory data analysis on tourist spending between regions or top countries in each region.



First Panel

Analysed by:
● Region
○ Country

Test type: non-parametric

Confidence level:
● 95%
○ 99%

Refer to second panel to continue plotting

Second Panel

Select y-axis: Spending per Trip

Plot type: Box Violin

☑ Show pairwise comparison

Display comparison:
○ significant
● non-significant

☑ Treat outliers

Press button below to update graph

$\chi^2_{\text{Kruskal-Wallis}}(4) = 1269.87$, $p = 1.14\text{e-}273$, $\hat{\varepsilon}^2_{\text{ordinal}} = 0.27$, $\text{CI}_{95\%}$ [0.25, 1.00], $n_{\text{obs}} = 4{,}748$

$p_{\text{FDR-adj.}} = 0.05$

Insights

Users can compare tourist spending of those coming from different regions and top countries of each region. Key highlights can be found in:

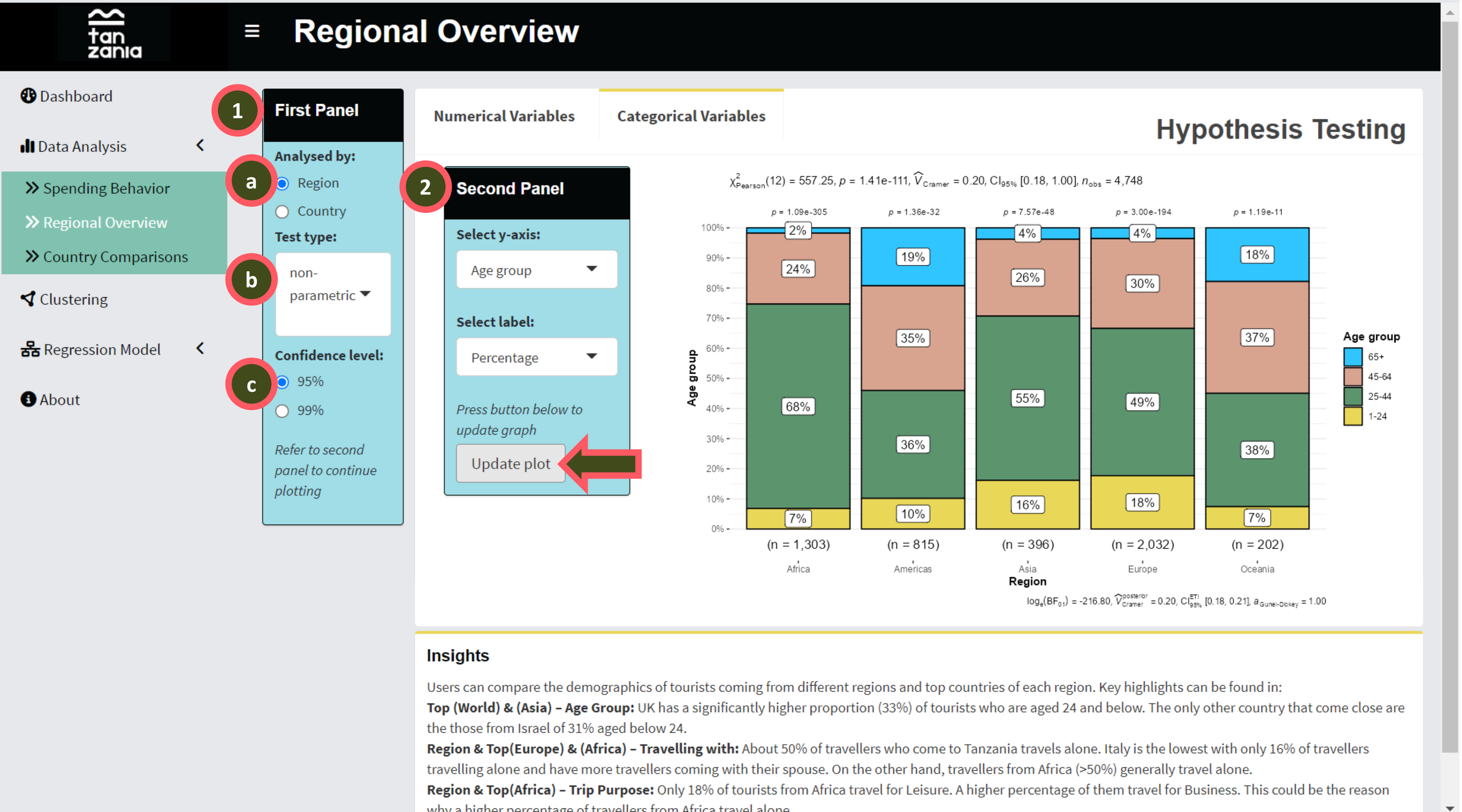
Region: There is statistically significant spending difference between travellers from all regions except those from Oceania and Americas.

1. **First Panel:** Users can perform confirmatory data analysis on tourist spending. Each parameter takes in only one input.
   a. Users can use the radio buttons to select either region or country comparisons.
   b. Users can use the dropdown boxes to select the type of correlation test.
   c. Users can use the radio buttons to select the confidence level.

2. **Second Panel:** Each parameter takes in only one input. The charts will appear after users click on the "Update plot" button.
   a. Users can select the spending metric to plot (y-axis) and the plot type.
   b. Users can select the check box "Show pairwise comparison" if they would like to see this in the charts.
   c. Users can use the radio buttons to choose whether to display significant or non-significant comparisons in the chart.
   d. Users can select the check box "Treat outliers" if they would like to use data which has been treated for outliers.

EXPLORE: Data Analysis – Regional Overview

This exploration page allows users to compare trends and perform confirmatory data analysis on tourist demographics between regions or top countries in each region.



First Panel

Analysed by:

● Region
○ Country

Test type: non-parametric

Confidence level:

● 95%
○ 99%

Refer to second panel to continue plotting

Second Panel

Select y-axis: Age group

Select label: Percentage

Press button below to update graph

$\chi^2_{\text{Pearson}}(12) = 557.25$, $p = 1.41\text{e-}111$, $\widehat{V}_{\text{Cramer}} = 0.20$, $\text{CI}_{95\%}\ [0.18, 1.00]$, $n_{\text{obs}} = 4{,}748$

$\log_e(\text{BF}_{01}) = -216.80$, $\widehat{V}^{\text{posterior}}_{\text{Cramer}} = 0.20$, $\text{CI}^{\text{ETI}}_{95\%}\ [0.18, 0.21]$, $a_{\text{Gunel-Dickey}} = 1.00$

Insights

Users can compare the demographics of tourists coming from different regions and top countries of each region. Key highlights can be found in:

Top (World) & (Asia) – Age Group: UK has a significantly higher proportion (33%) of tourists who are aged 24 and below. The only other country that come close are the those from Israel of 31% aged below 24.

Region & Top(Europe) & (Africa) – Travelling with: About 50% of travellers who come to Tanzania travels alone. Italy is the lowest with only 16% of travellers travelling alone and have more travellers coming with their spouse. On the other hand, travellers from Africa (>50%) generally travel alone.

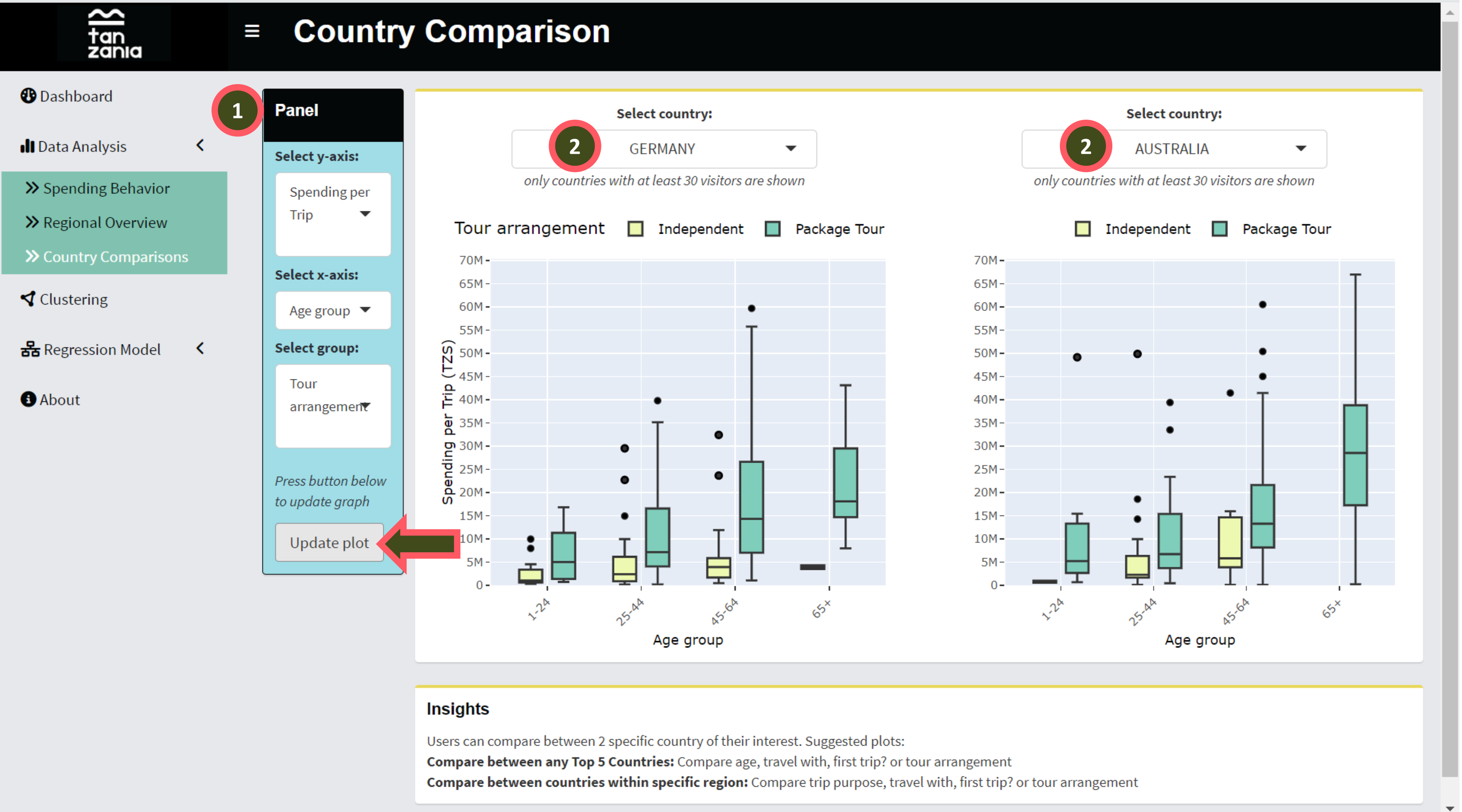
Region & Top(Africa) – Trip Purpose: Only 18% of tourists from Africa travel for Leisure. A higher percentage of them travel for Business. This could be the reason why a higher percentage of travellers from Africa travel alone.

1. **First Panel:** Users can perform confirmatory data analysis on categorical variables. Each parameter takes in only one input.
   a. Users can use the radio buttons to select either region or country comparisons.
   b. Users can use the dropdown boxes to select the type of correlation test to perform.
   c. Users can use the radio buttons to select the confidence level.

2. **Second Panel:** Users can use the dropdown boxes to select the demographics metric to plot (y-axis) and whether the labels should show percentages or absolute counts. Each parameter takes in only one input. The charts will appear after users click on the “Update plot” button.

EXPLORE: Data Analysis – Country Comparison

This exploration page allows users to compare drivers of spending between any two countries in terms of demographics and behavioural traits.



Select y-axis: Spending per Trip

Select x-axis: Age group

Select group: Tour arrangement

Press button below to update graph

Select country: GERMANY

only countries with at least 30 visitors are shown

Select country: AUSTRALIA

only countries with at least 30 visitors are shown

Insights

Users can compare between 2 specific country of their interest. Suggested plots:

Compare between any Top 5 Countries: Compare age, travel with, first trip? or tour arrangement

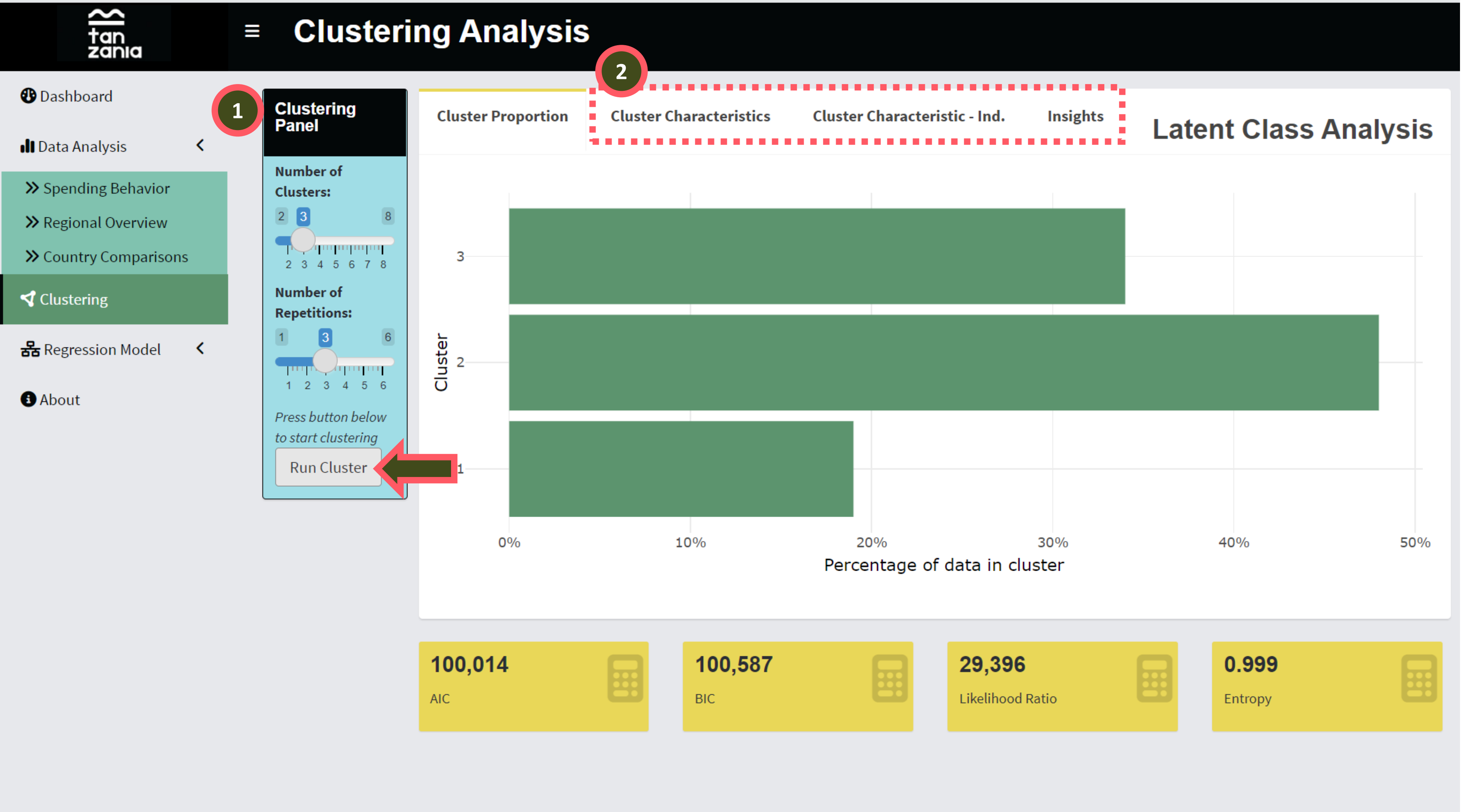
Compare between countries within specific region: Compare trip purpose, travel with, first trip? or tour arrangement

1. **Panel:** Users can use the dropdown boxes to select the spending metric (y-axis) to plot against a selected demographics or behavioural trait category (x-axis). Users can also select which demographics or behavioural trait category (group) they would like to compare the data on. Each parameter takes in only one input. The charts will appear after users click on the "Update plot" button.

2. **Select Country:** Users can use the dropdown boxes to select the two countries to compare between. Each parameter takes in only one input. The charts can be refreshed after users click on the "Update plot" button in the **Panel**.

ANALYSIS: Clustering

This analysis page allows users to perform Clustering analysis to group tourist behaviour. Users can adjust the number of clusters and number of repetitions to find the best cluster based on the AIC, BIC, Likelihood Ratio and Entropy values. Specifically, this tab allows users to review the proportion of sample that fall within each class when trying to identify the best solution.



Clustering Analysis

Clustering Panel

Number of Clusters: 3 (2–8)

Number of Repetitions: 3 (1–6)

Press button below to start clustering

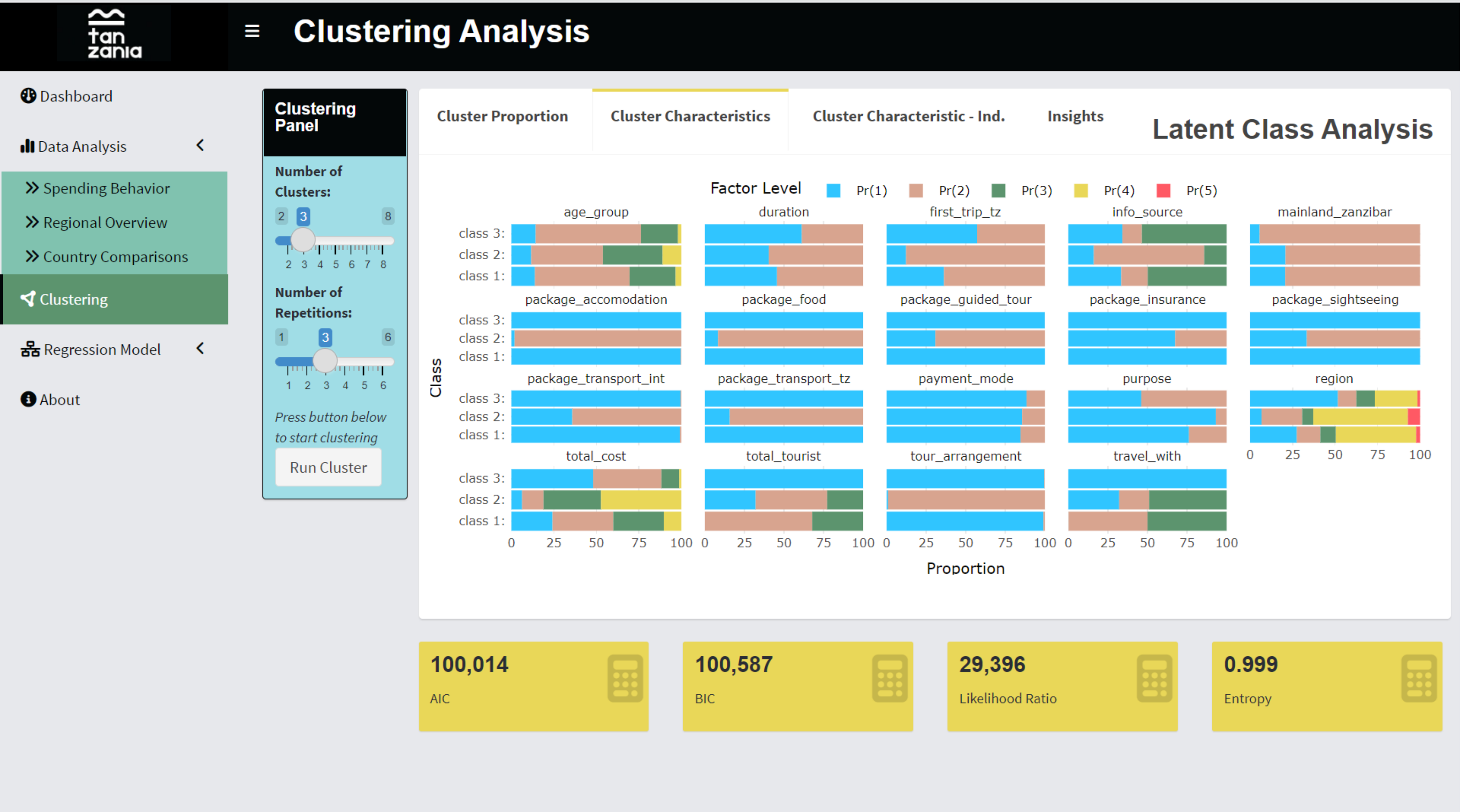
100,014	100,587	29,396	0.999
AIC	BIC	Likelihood Ratio	Entropy

1. **Clustering Panel:** Users can use the sliding bar to decide on the number of clusters and repetitions to build the clustering model. Each parameter takes in only one input. The model will run after users click on the “Run Cluster” button.

2. **Tabs:** Once the model has been run, users can click on the other tabs to view the relevant charts without having to rerun the model.

ANALYSIS: Clustering

This analysis page allows users to perform Clustering analysis to group tourist behaviour. Users can adjust the number of clusters and number of repetitions to find the best cluster based on the AIC, BIC, Likelihood Ratio and Entropy values. Specifically, this tab allows users to review the cluster characteristics at one glance.



Number of Clusters: 3

Number of Repetitions: 3

Press button below to start clustering

100,014 AIC

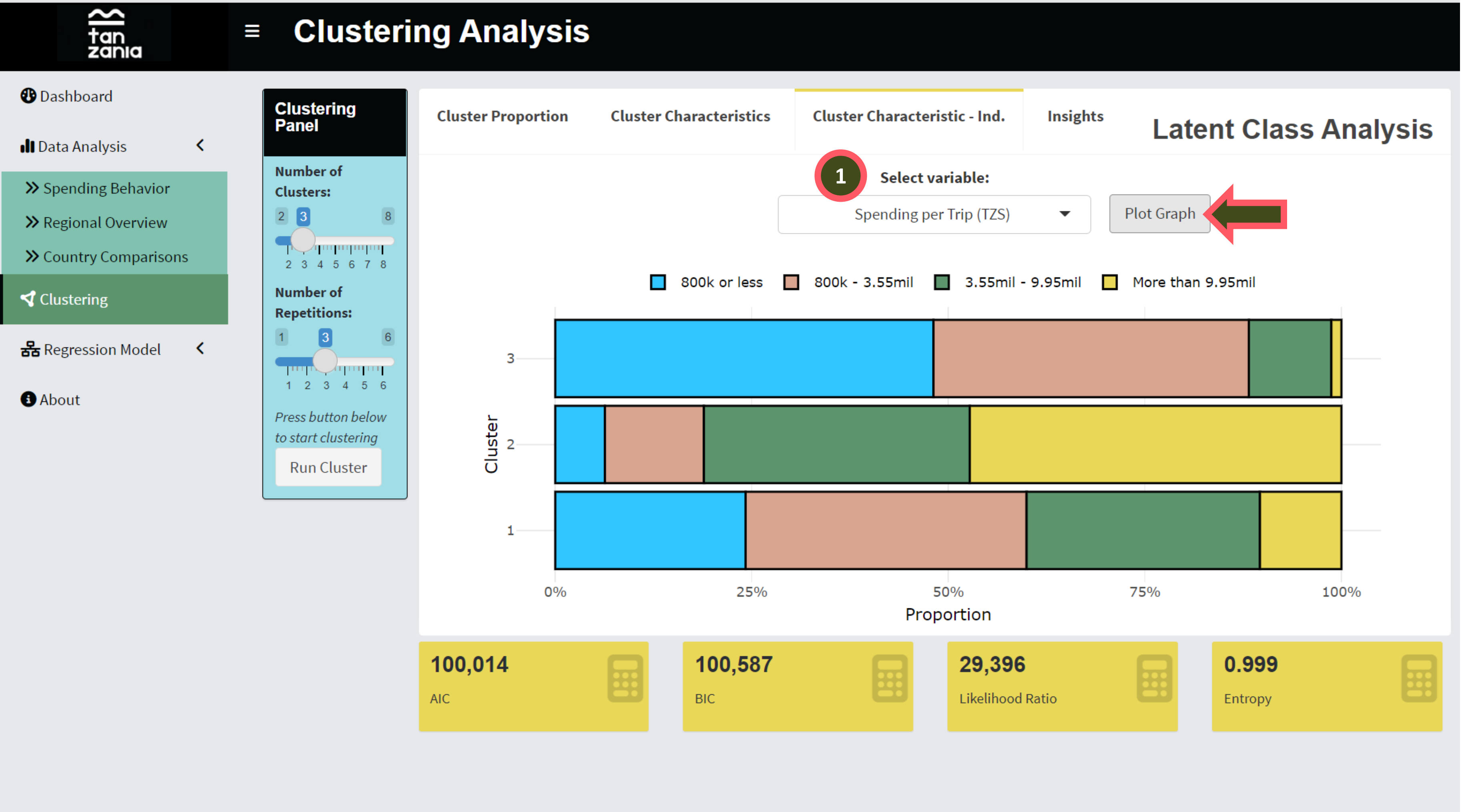
100,587 BIC

29,396 Likelihood Ratio

0.999 Entropy

ANALYSIS: Clustering

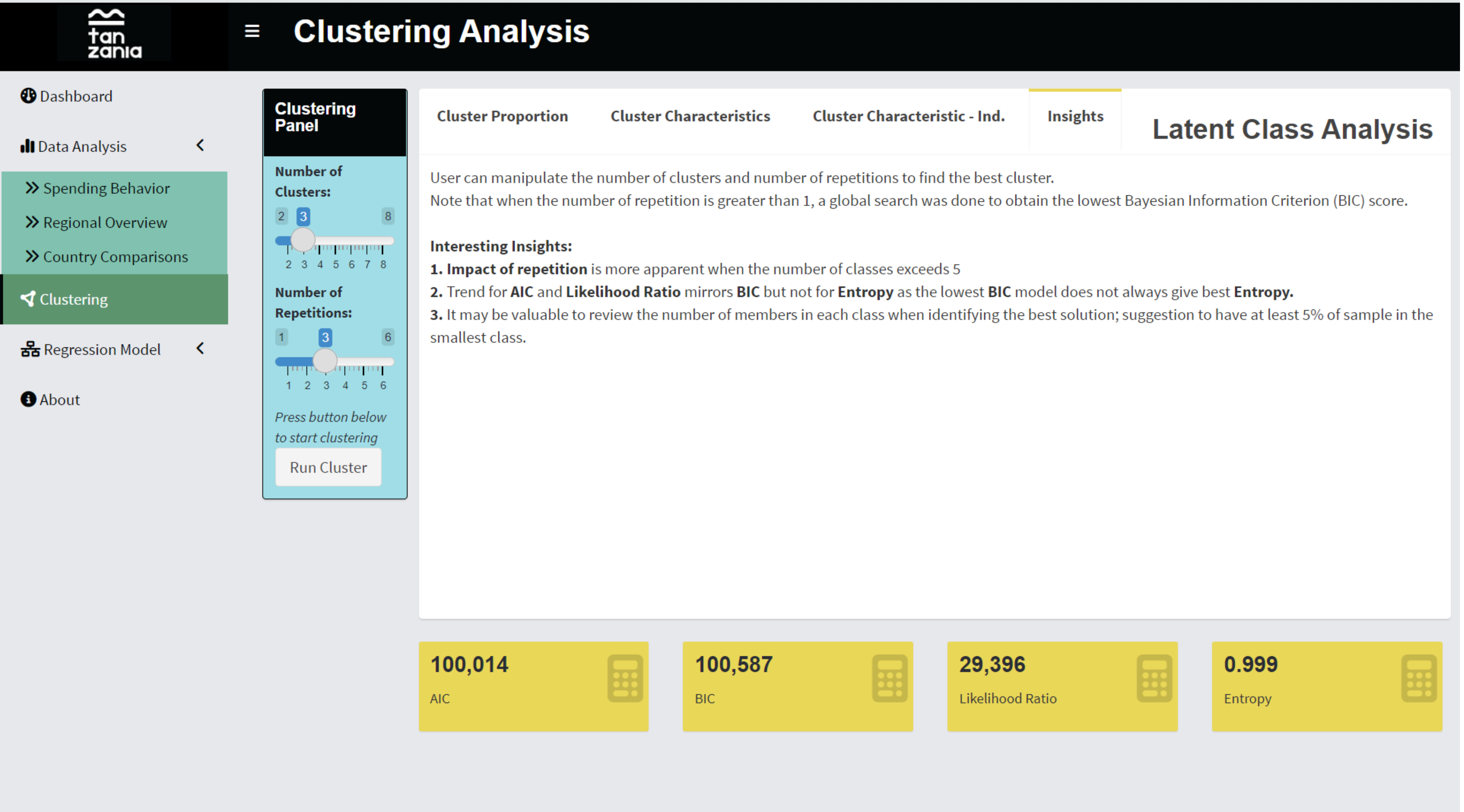
This analysis page allows users to perform Clustering analysis to group tourist behaviour. Users can adjust the number of clusters and number of repetitions to find the best cluster based on the AIC, BIC, Likelihood Ratio and Entropy values. Specifically, this tab allows users to review in detail the cluster characteristics for each variable.



1. **Select Variable:** Once the model has been run, in the Cluster Characteristic – Ind. tab, users can use the dropdown box to select the variable that they would like to review in detail. Each parameter takes in only one input. The charts will appear after users click on the “Plot Graph” button.

ANALYSIS: Clustering

This analysis page allows users to perform Clustering analysis to group tourist behaviour. Users can adjust the number of clusters and number of repetitions to find the best cluster based on the AIC, BIC, Likelihood Ratio and Entropy values. Specifically, this tab provides users with suggestions on how to identify the best clustering solution.



Clustering Analysis

Clustering Panel

Number of Clusters: 3

Number of Repetitions: 3

Press button below to start clustering

Latent Class Analysis

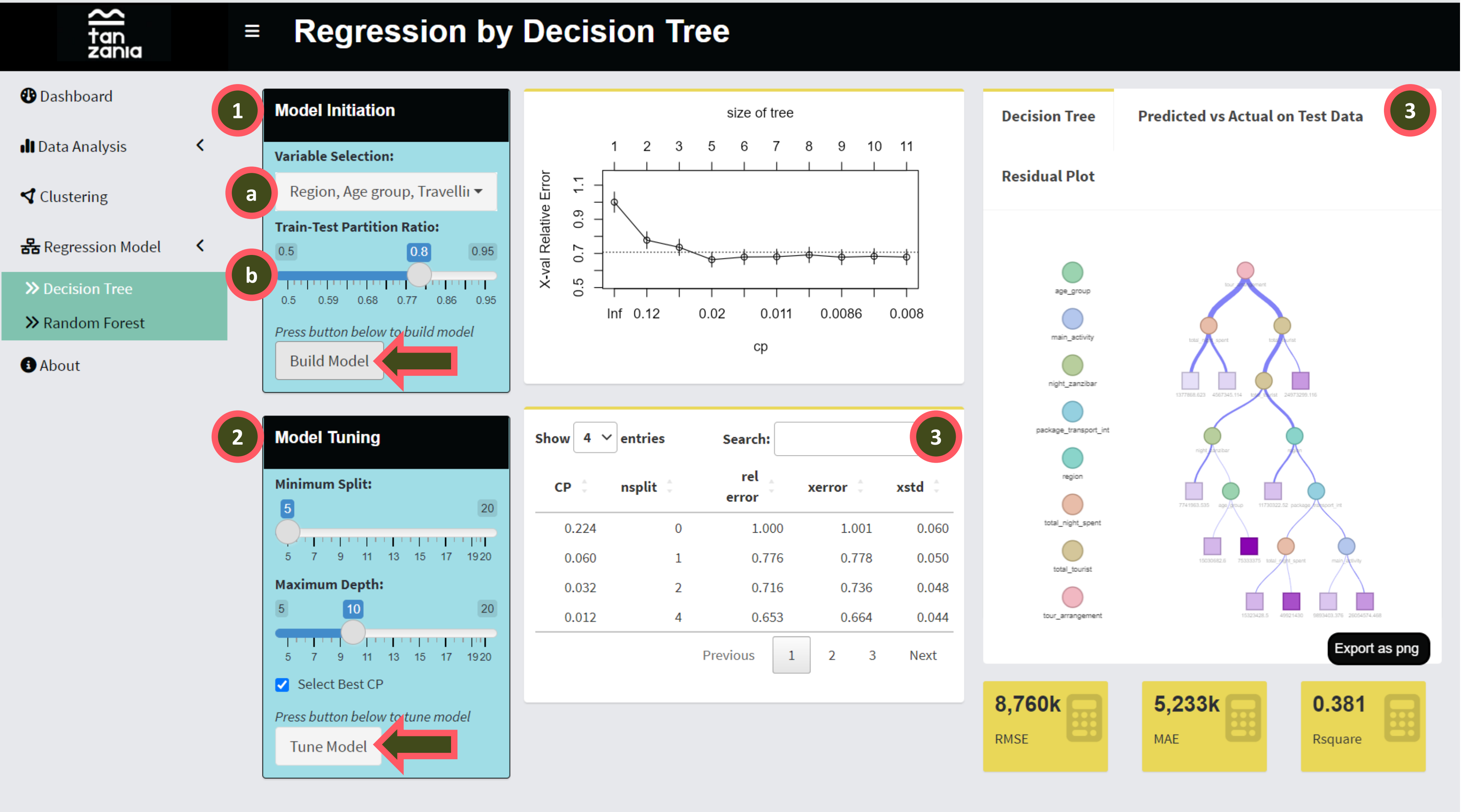
User can manipulate the number of clusters and number of repetitions to find the best cluster.
Note that when the number of repetition is greater than 1, a global search was done to obtain the lowest Bayesian Information Criterion (BIC) score.

Interesting Insights:
1. Impact of repetition is more apparent when the number of classes exceeds 5
2. Trend for **AIC** and **Likelihood Ratio** mirrors **BIC** but not for **Entropy** as the lowest **BIC** model does not always give best **Entropy.**
3. It may be valuable to review the number of members in each class when identifying the best solution; suggestion to have at least 5% of sample in the smallest class.

100,014 AIC	**100,587** BIC	**29,396** Likelihood Ratio	**0.999** Entropy

ANALYSIS: Regression Model – Decision Tree

This analysis page allows users to perform Regression analysis to predict tourist spending. Specifically, this tab allows users to build a Regression Tree model to find the best model based on the RMSE, MAE and R-squared values. Users can adjust variables to be included in the model, the train-test partition ratio and perform hyperparameter tuning (minimum split, maximum depth and complexity parameter).



Regression by Decision Tree

Model Initiation

Variable Selection: Region, Age group, Travellii

Train-Test Partition Ratio: 0.5 – 0.8 – 0.95

Press button below to build model

Build Model

Model Tuning

Minimum Split: 5 (range 5–20)

Maximum Depth: 10 (range 5–20)

☑ Select Best CP

Press button below to tune model

Tune Model

Show 4 entries Search:

CP	nsplit	rel error	xerror	xstd
0.224	0	1.000	1.001	0.060
0.060	1	0.776	0.778	0.050
0.032	2	0.716	0.736	0.048
0.012	4	0.653	0.664	0.044

Previous 1 2 3 Next

8,760k RMSE

5,233k MAE

0.381 Rsquare

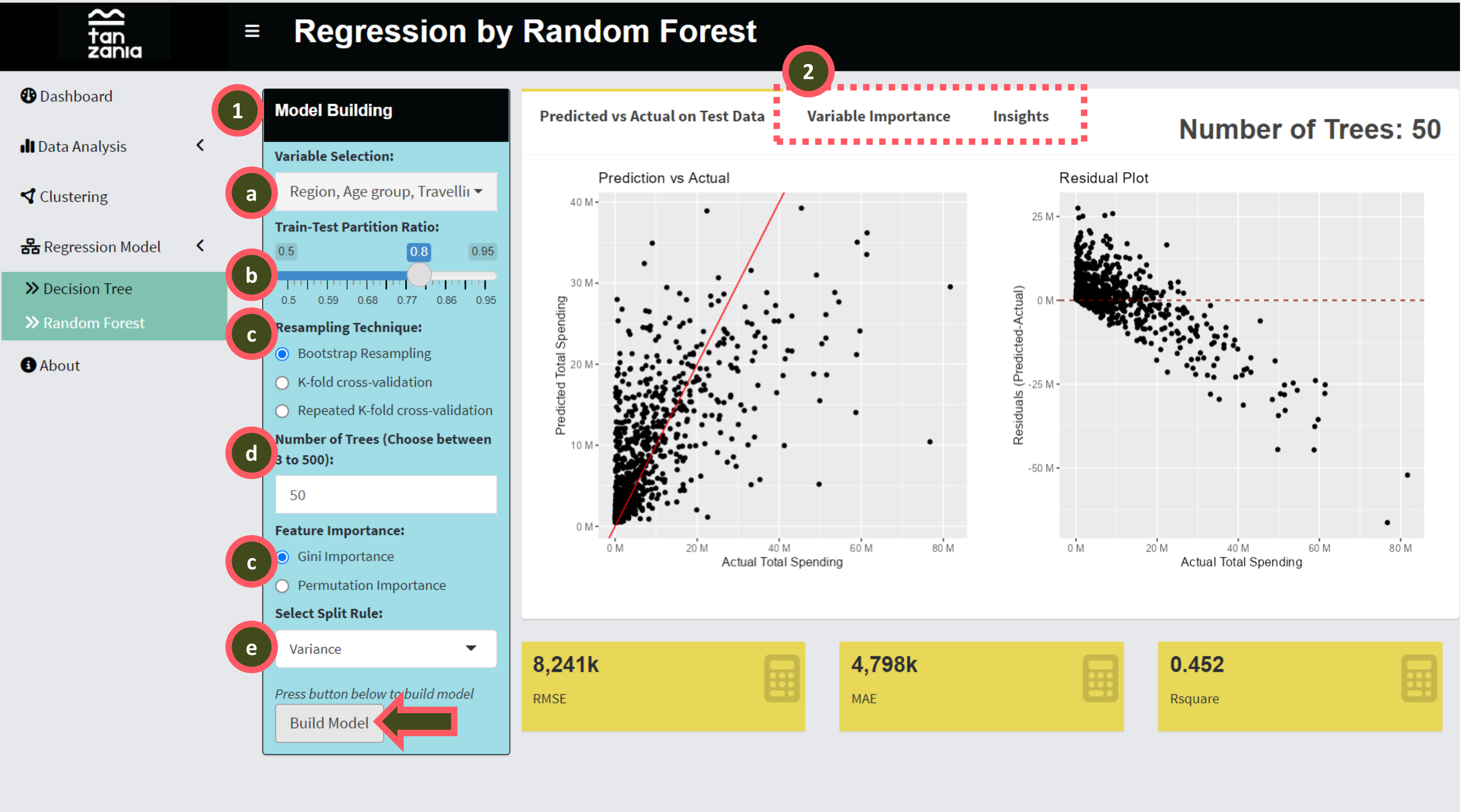
1. **Model Initiation:** The model will run after users click on “Build Model”.
   a. Users can use the dropdown box to select the variables to be used to build the regression tree model. This parameter takes in multiple inputs.
   b. Users can use the sliding bar to select the train-test partition ratio. This parameter takes in only one input.

2. **Model Tuning:** This section appears after the model has been run. Each parameter takes in only one input. The model will be modified after users click on “Tune Model”.
   a. Users can use the sliding bars to select the minimum split and maximum depth to tune the model.
   b. If users select the check box “Select Best CP”, no further action is needed, else users must input the preferred complexity parameter (CP).

3. **Charts / Table:** Users can interact with the charts / table once the model has been built.

ANALYSIS: Regression Model – Random Forest

This analysis page allows users to perform Regression analysis to predict tourist spending. Specifically, this tab allows users to build a Random Forest model to find the best cluster based on the RMSE, MAE and R-squared values. Users can adjust variables to be included in the model, the train-test partition ratio, resampling techniques, number of trees in the forest, variable importance mode and split rule.



Regression by Random Forest

Model Building

Variable Selection: Region, Age group, Travellii

Train-Test Partition Ratio: 0.8

Resampling Technique:
● Bootstrap Resampling
○ K-fold cross-validation
○ Repeated K-fold cross-validation

Number of Trees (Choose between 3 to 500): 50

Feature Importance:
● Gini Importance
○ Permutation Importance

Select Split Rule: Variance

Press button below to build model

Number of Trees: 50

8,241k RMSE

4,798k MAE

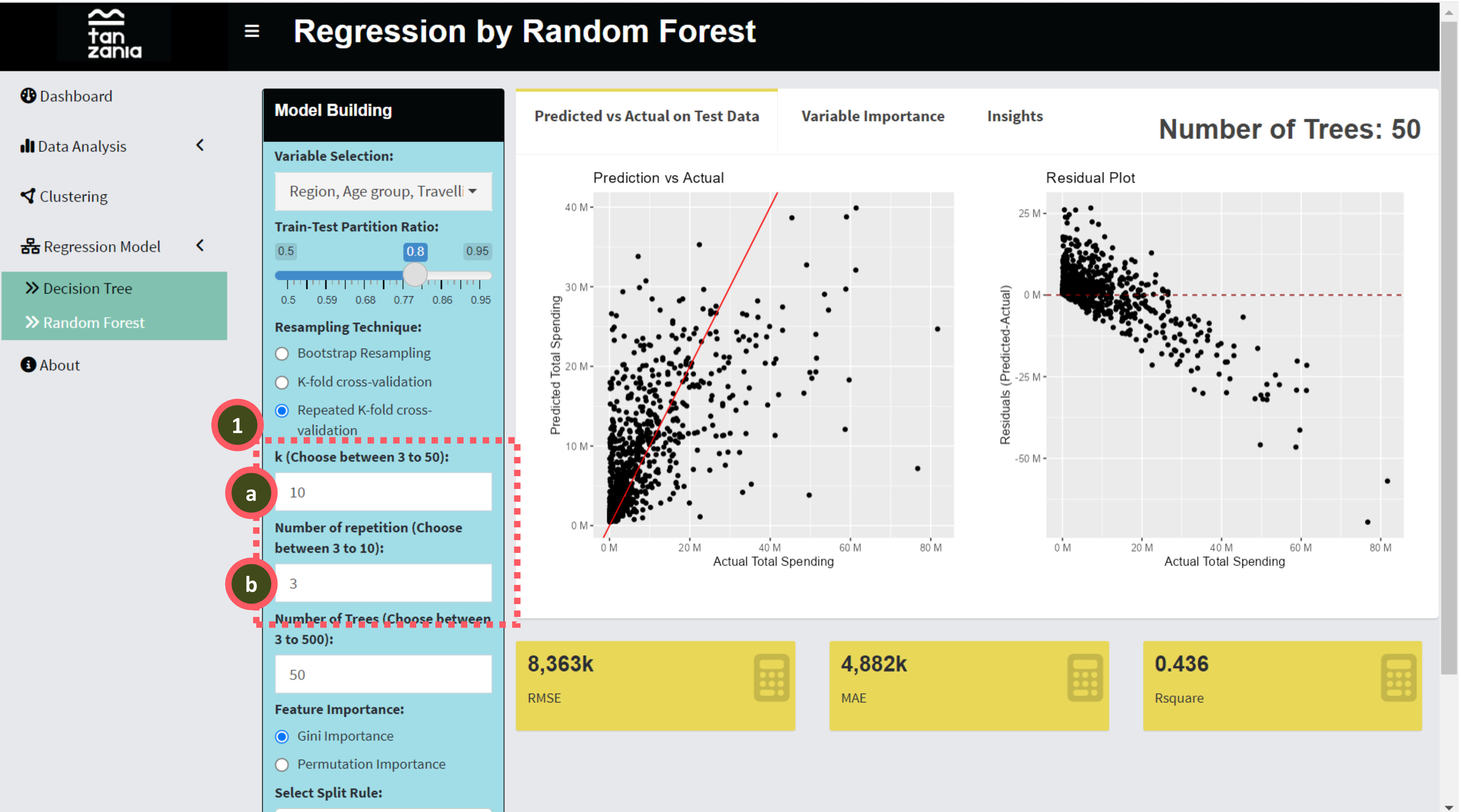
0.452 Rsquare

1. **Model Initiation:** Except for part 1a, each parameter takes in only one input. The model will run after users click on "Build Model".
   a. Users can use the dropdown box to select the variables to be used to build the random forest model. This parameter takes in multiple inputs.
   b. Users can use the sliding bar to select the train-test partition ratio. This parameter takes in only one input. The model will run after users click on "Build Model".
   c. Users can use the radio buttons to select a resampling technique and feature importance mode.
   d. Users can input a preferred number of trees to build the random forest model on.
   e. Users can use the dropdown box to select the split rule.

2. **Tabs:** Once the model has been run, users can click on the other tabs to view the relevant charts without having to rerun the model.

ANALYSIS: Regression Model – Random Forest

This analysis page allows users to perform Regression analysis to predict tourist spending. Specifically, this tab allows users to build a Random Forest model to visualize the top 20 most important variables based on the model built in the previous page.



Regression by Random Forest

Model Building

Variable Selection: Region, Age group, Travelli

Train-Test Partition Ratio: 0.8 (0.5 – 0.95)

Resampling Technique:

- ○ Bootstrap Resampling
- ○ K-fold cross-validation
- ● Repeated K-fold cross-validation

k (Choose between 3 to 50): 10

Number of repetition (Choose between 3 to 10): 3

Number of Trees (Choose between 3 to 500): 50

Feature Importance:

- ● Gini Importance
- ○ Permutation Importance

Select Split Rule:

Number of Trees: 50

8,363k RMSE

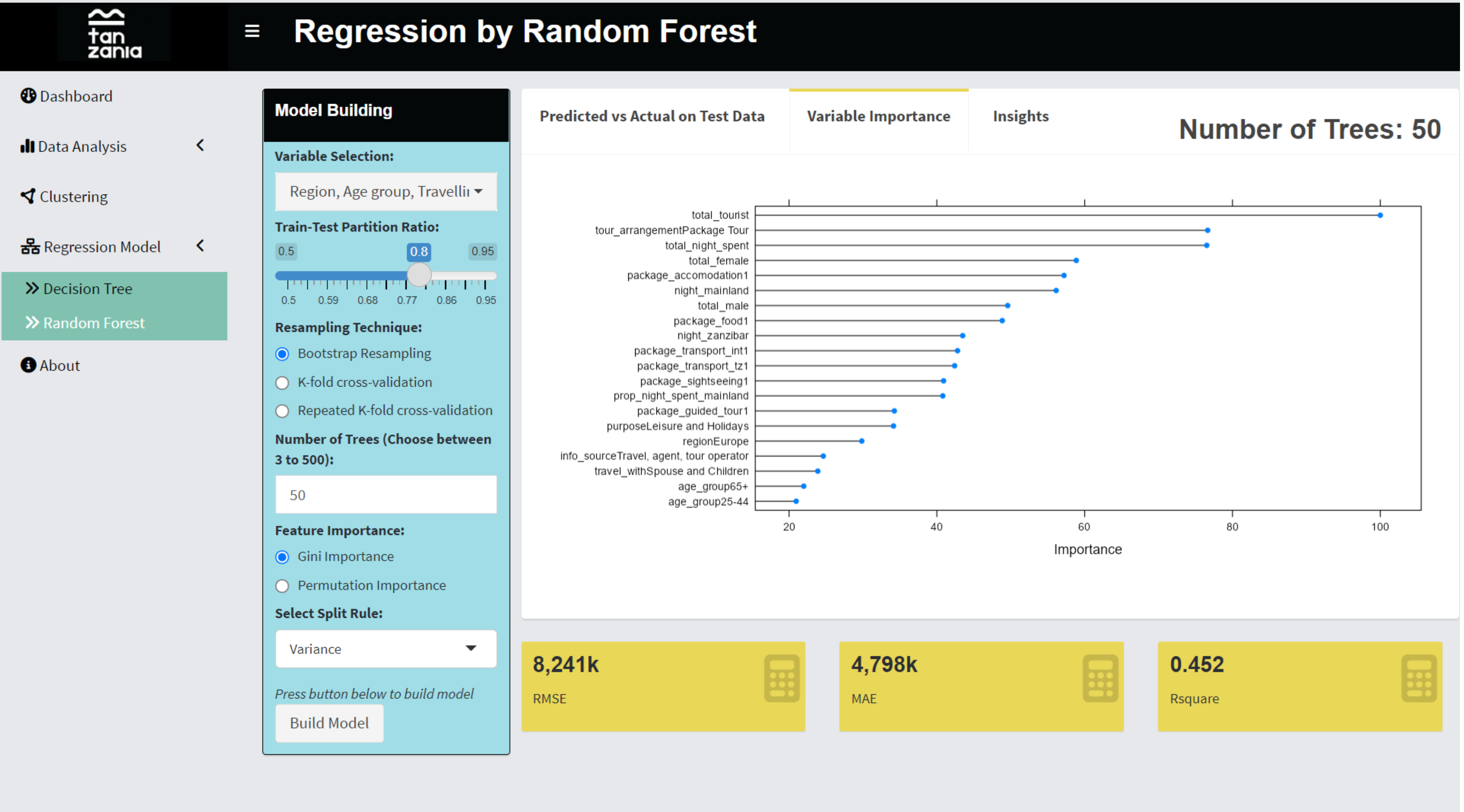
4,882k MAE

0.436 Rsquare

1. **(Repeated) k-fold cross-validation:** This section appears if either of the k-fold cross-validation resampling options are selected. Each parameter takes in only one input.
   a. Users can input a preferred number for k.
   b. If repeated k-fold cross-validation option is selected, users can input a preferred number of repetition.

ANALYSIS: Regression Model – Random Forest

This analysis page allows users to perform Regression analysis to predict tourist spending. Specifically, this tab allows users to build a Random Forest model to visualize the top 20 most important variables based on the model built in the previous page.



Regression by Random Forest

Model Building

Variable Selection: Region, Age group, Travellii

Train-Test Partition Ratio: 0.5 – 0.8 – 0.95

Resampling Technique:

- ● Bootstrap Resampling
- ○ K-fold cross-validation
- ○ Repeated K-fold cross-validation

Number of Trees (Choose between 3 to 500): 50

Feature Importance:

- ● Gini Importance
- ○ Permutation Importance

Select Split Rule: Variance

Press button below to build model

Build Model

Number of Trees: 50

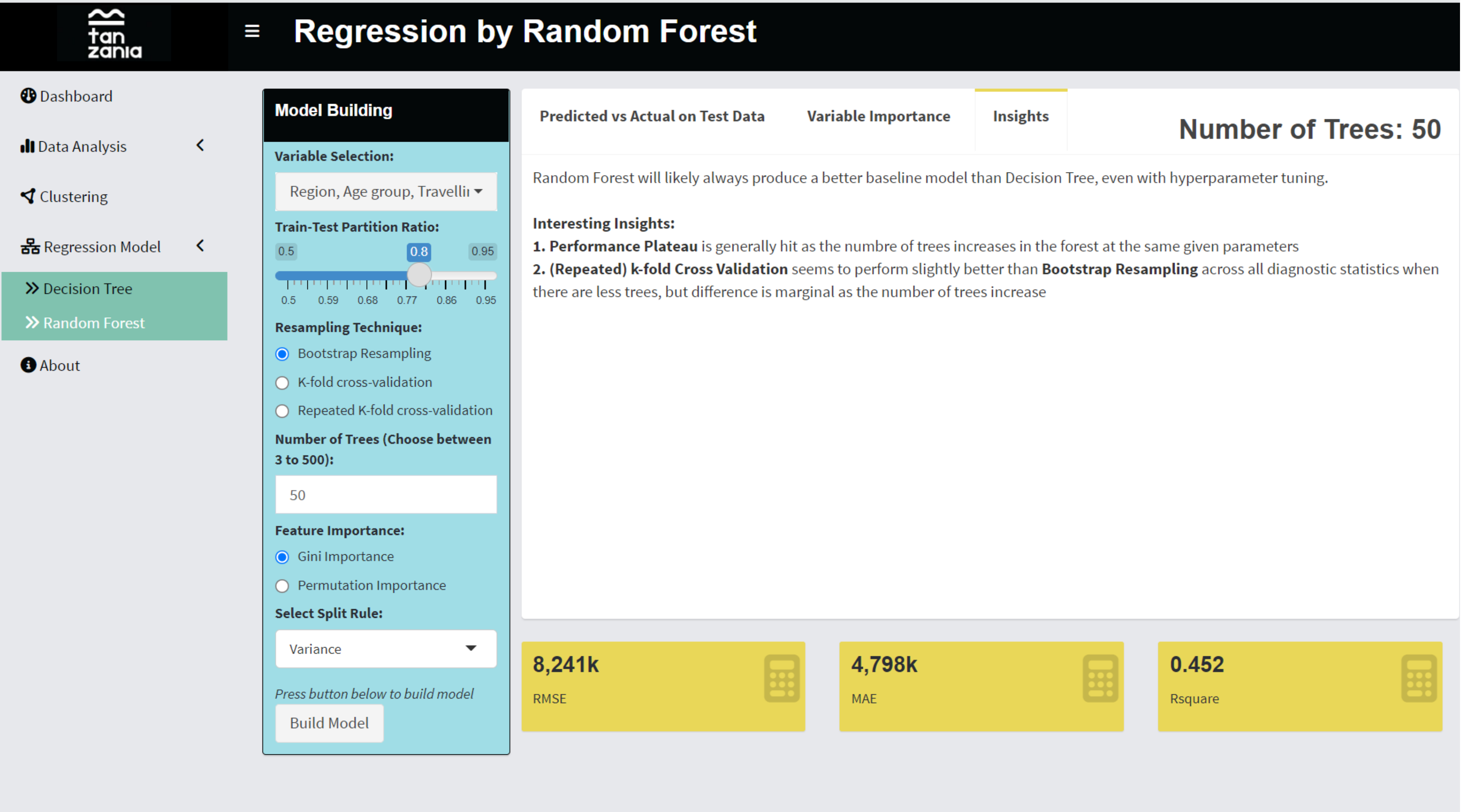
8,241k RMSE

4,798k MAE

0.452 Rsquare

ANALYSIS: Regression Model – Random Forest

This analysis page allows users to perform Regression analysis to predict tourist spending. Specifically, this tab provides users with suggestions on how to identify the best regression model solution.



Regression by Random Forest

Model Building

Variable Selection: Region, Age group, Travellii

Train-Test Partition Ratio: 0.8 (0.5 – 0.95)

Resampling Technique:

- ● Bootstrap Resampling
- ○ K-fold cross-validation
- ○ Repeated K-fold cross-validation

Number of Trees (Choose between 3 to 500): 50

Feature Importance:

- ● Gini Importance
- ○ Permutation Importance

Select Split Rule: Variance

Press button below to build model

Number of Trees: 50

Random Forest will likely always produce a better baseline model than Decision Tree, even with hyperparameter tuning.

Interesting Insights:

1. Performance Plateau is generally hit as the numbre of trees increases in the forest at the same given parameters

2. (Repeated) k-fold Cross Validation seems to perform slightly better than **Bootstrap Resampling** across all diagnostic statistics when there are less trees, but difference is marginal as the number of trees increase

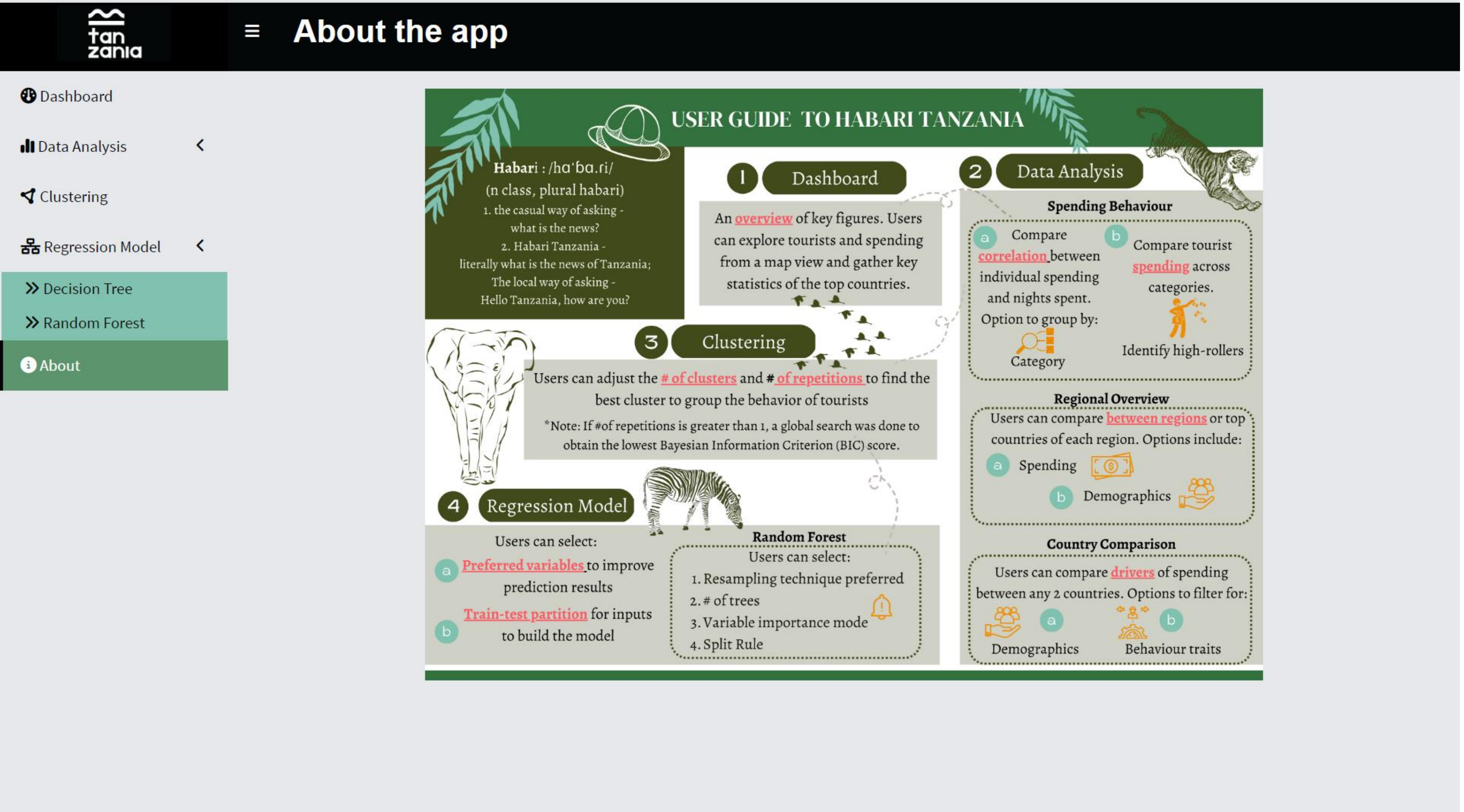
8,241k
RMSE

4,798k
MAE

0.452
Rsquare

ABOUT

This about page serves as a quick user guide to using this Shiny app. Upon clicking the image, users will be redirected to this detailed user guide.



USER GUIDE TO HABARI TANZANIA

Habari : /hɑˈbɑ.ri/
(n class, plural habari)
1. the casual way of asking - what is the news?
2. Habari Tanzania - literally what is the news of Tanzania;
The local way of asking - Hello Tanzania, how are you?

1 Dashboard

An overview of key figures. Users can explore tourists and spending from a map view and gather key statistics of the top countries.

2 Data Analysis

Spending Behaviour

a Compare correlation between individual spending and nights spent. Option to group by: Category

b Compare tourist spending across categories. Identify high-rollers

Regional Overview

Users can compare between regions or top countries of each region. Options include:

a Spending

b Demographics

Country Comparison

Users can compare drivers of spending between any 2 countries. Options to filter for:

a Demographics

b Behaviour traits

3 Clustering

Users can adjust the # of clusters and # of repetitions to find the best cluster to group the behavior of tourists

*Note: If #of repetitions is greater than 1, a global search was done to obtain the lowest Bayesian Information Criterion (BIC) score.

4 Regression Model

Users can select:

a Preferred variables to improve prediction results

b Train-test partition for inputs to build the model

Random Forest

Users can select:

1. Resampling technique preferred
2. # of trees
3. Variable importance mode
4. Split Rule