### **Context**

The Tanzanian tourism sector plays a significant role in the Tanzanian economy, contributing about 17% to the country’s GDP and 25% of all foreign exchange revenues. The sector, which provides direct employment for more than 600,000 people and up to 2 million people indirectly, generated approximately $2.4 billion in 2018 according to government statistics. Tanzania received a record 1.1 million international visitor arrivals in 2014, mostly from Europe, the US and Africa.

Tanzania is the only country in the world which has allocated more than 25% of its total area for wildlife, national parks, and protected areas.There are 16 national parks in Tanzania, 28 game reserves, 44 game-controlled areas, two marine parks and one conservation area.

### **Objective**

The dataset describes 6476 rows of up-to-date information on tourist expenditure collected by the National Bureau of Statistics (NBS) in Tanzania.The dataset was collected to gain a better understanding of the status of the tourism sector and provide an instrument that will enable sector growth. Your goal is to accurately predict tourist expenditure when visiting Tanzania.

The objective of this Challenge is to develop a machine learning model to predict what a tourist will spend when visiting Tanzania.The model can be used by different tour operators and the Tanzania Tourism Board to automatically help tourists across the world estimate their expenditure before visiting Tanzania.

***More details on the metric are available in the template notebook provided.***

**Data Description**

* **Train.csv** - To be used for training and tuning of models.
* **Test.csv** - To be used only for testing the performance of the final best model.
* **variableDescribtion -** csv file carrying the variables describtion

#### **Submission Guidelines**

1. There are two parts to the submission:
   1. A well commented Jupyter notebook [format - .html]
   2. A presentation as you would present to the top management/business leaders [format - .pdf]
2. Any found copied/ plagiarized with other competitors will not be graded and is an automatic disqualification
3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation.
4. Submission will not be evaluated if
   1. it is submitted post-deadline, or,
   2. more than 2 files are submitted.

#### **Best Practices for Notebook**

* The final notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* It is important to remove all warnings and errors before submission.
* The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb).
* A template notebook with steps to follow and a few questions to answer has been provided for reference. It will help you approach the analysis in the right manner and generate insights from the data. It is not mandatory to use it for submission.
* It is recommended that you read the problem statement and go through the criteria and description mentioned in the rubric before starting the project..

#### **Best Practices for Presentation**

* The presentation should be made keeping in mind that the audience will be a business leader like CMO, COO, CFO, or CEO.
* The key points in the presentation should be the following:
  + Business overview of the problem and solution approach
  + Key findings and insights which can drive business decisions
  + Model overview and performance summary
  + Conclusions
* Focus on explaining the takeaways in an easy-to-understand manner.
* The inclusion of the potential benefits of implementing the solution will give you the edge.
* Copying and pasting from the notebook is not a good idea, and it is better to avoid showing codes unless they are the focal point of your presentation.
* The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file.
* A presentation template has been provided for reference. It is not mandatory to use it for submission.

##### **Scoring guide (Rubric)**

| **Criteria** | **Points** |
| --- | --- |
| **Exploratory Data Analysis and Insights** - Overview of the data - Univariate analysis  - Bivariate analysis  -Multivariate analysis |  |
| **Data pre-processing** - Prepare the data for analysis - Missing value Treatment - Ensure no data leakage |  |
| **Model building** - Build a Linear Regression model |  |
| **Model Performances** - evaluate the model’s performance |  |
| **Productionize the model** - Productionize the model using a pipeline |  |
| **Business Insights & Conclusions** - Business insights and Conclusions. |  |
| **Notebook - Overall quality** - Structure and flow. - Well commented code. |  |
| **Presentation - Overall quality** - Structure and flow - Crispness - All key insights and recommendations covered |  |
|  |  |