

## Homestays Data analysis and price prediction

**Step 1: Please find the dataset required for the following tasks at this link:** [Homestays\\_Data.zip](https://cerinaco-my.sharepoint.com/:u:/g/personal/omkar_cerina_co/ETARwr3mqEtAoZFmWmYKnOIBtPhRtBczH9svOO1glGqxUg?e=3NwzrD) [[https://cerinaco-my.sharepoint.com/:u:/g/personal/omkar\\_cerina\\_co/ETARwr3mqEtAoZFmWmYKnOIBtPhRtBczH9svOO1glGqxUg?e=3NwzrD](https://cerinaco-my.sharepoint.com/:u:/g/personal/omkar_cerina_co/ETARwr3mqEtAoZFmWmYKnOIBtPhRtBczH9svOO1glGqxUg?e=3NwzrD)]

**Download the “Homestays\_Data.zip” file and find the “Homestays\_Data.csv” which contains the dataset.**

*Note: The entire dataset has multiple columns and has more than 70,000 rows. Processing the following tasks on a Jupyter notebook on your local environment (your personal computer) might be time consuming due to limited processing power. We encourage you to use cloud-based tools such as Google Colab or AWS Sagemaker or anything that you prefer.*

---

### Step 2: Complete the following tasks.

Objective: Build a robust predictive model to estimate the `log\_price` of homestay listings based on comprehensive analysis of their characteristics, amenities, and host information.

First make sure that the entire dataset is clean and ready to be used.

#### 1. Feature Engineering:

Task: Enhance the dataset by creating actionable and insightful features. Calculate `Host\_Tenure` by determining the number of years from `host\_since` to the current date, providing a measure of host experience. Generate `Amenities\_Count` by counting the items listed in the `amenities` array to quantify property offerings. Determine `Days\_Since\_Last\_Review` by calculating the days between `last\_review` and today to assess listing activity and relevance.

#### 2. Exploratory Data Analysis (EDA):

Task: Conduct a deep dive into the dataset to uncover underlying patterns and relationships. Analyze how pricing (`log\_price`) correlates with both categorical (such as `room\_type` and `property\_type`) and numerical features (like `accommodates` and `number\_of\_reviews`). Utilize statistical tools and visualizations such as correlation matrices, histograms for distribution analysis, and scatter plots to explore relationships between variables.

#### 3. Geospatial Analysis:

Task: Investigate the geographical data to understand regional pricing trends. Plot listings on a map using `latitude` and `longitude` data to visually assess price distribution. Examine if certain neighbourhoods or proximity to city centres influence pricing, providing a spatial perspective to the pricing strategy.

#### 4. Sentiment Analysis on Textual Data:

Task: Apply advanced natural language processing techniques to the `description` texts to extract sentiment scores. Use sentiment analysis tools to determine whether positive or negative descriptions influence listing prices, incorporating these findings into the predictive model being trained as a feature.

## 5. Amenities Analysis:

Task: Thoroughly parse and analyse the `amenities` provided in the listings. Identify which amenities are most associated with higher or lower prices by applying statistical tests to determine correlations, thereby informing both pricing strategy and model inputs.

## 6. Categorical Data Encoding:

Task: Convert categorical data into a format suitable for machine learning analysis. Apply one-hot encoding to variables like `room\_type`, `city`, and `property\_type`, ensuring that the model can interpret these as distinct features without any ordinal implication.

## 7. Model Development and Training:

Task: Design and train predictive models to estimate `log\_price`. Begin with a simple linear regression to establish a baseline, then explore more complex models such as RandomForest and GradientBoosting to better capture non-linear relationships and interactions between features. Document (briefly within Jupyter notebook itself) the model-building process, specifying the choice of algorithms and rationale.

## 8. Model Optimization and Validation:

Task: Systematically optimize the models to achieve the best performance. Employ techniques like grid search to experiment with different hyperparameters settings. Validate model choices through techniques like k-fold cross-validation, ensuring the model generalizes well to unseen data.

## 9. Feature Importance and Model Insights:

Task: Analyze the trained models to identify which features most significantly impact `log\_price`. Utilize model-specific methods like feature importance scores for tree-based models and SHAP values for an in-depth understanding of feature contributions.

## 10. Predictive Performance Assessment:

Task: Critically evaluate the performance of the final model on a reserved test set. Use metrics such as Root Mean Squared Error (RMSE) and R-squared to assess accuracy and goodness of fit. Provide a detailed analysis of the residuals to check for any patterns that might suggest model biases or misfit.

Final Deliverables:

**Comprehensive Jupyter notebook (PDF File/ PDF Export):** Contains all codes, analytical steps, visualizations, and detailed commentary explaining each action and decision.