Problem Statement and Data

You have been given access to data from the data warehouse of an online aggregator for properties that tourists can rent out for short durations. The current sample is from all the listings in Antwerp Belgium.
The task is to generate an ML based solution that can be used to suggest appropriate listing prices to the property owner when they try to list a property out for rent.
The dataset was pulled from a data warehouse and has the following tables:

1. **Calendar (Data on listings in Chronological Order)**

    - listing_id (Unique id for each listing)
    - date (Date on which listing was made)
    - available (1->Available, 0->Rented out)
    - price (Price in USD)
    - adjusted_price (Adjusted Price USD, use price column instead of this for any analysis)
    - minimum_nights (Minimum number of nights a guest can book)
    - maximum_nights (Maximum number of nights a guest can book)

2. **Listings (Data on details of a particular listing)**

    - listing_id (Unique id for each listing)
    - listing_url (Unique url to each listing)
    - name (Name of the listing)
    - description (Description of the listing)
    - lat (latitude)
    - long (longitude)
    - property_type (Type of property)
    - room_type (Type of room)
    - accommodates (Number of guests that can be accommodated)
    - bathrooms (Number of bathrooms)
    - bedrooms (Number of bedrooms)
    - beds (Number of beds)
    - amenities (List of amenities)
    - host_id (Unique Id of the host)

3. **Hosts (Data on hosts who've posted their listing)**

    - host_id (Unique id of host)
    - host_name (Name of the host)
    - host_since (Timestamp of host registration on the platform)
    - host_location (Location of host)
    - host_about (Self reported about section of each host)

4. **Reviews (Data on Reviews)**

    - listing_id (Unique id for each listing)
    - review_id (Unique review id)
    - date (Date of posting the review)
    - reviewer_id (Id of reviewer)
    - reviewer_name (Name of the reviewer)
    - comments (Review comment)

Tasks

Data Understanding and feature creation (Task 1):

- Look at the table Calendar how many rows and unique listing ids are present? Are there any implications when it comes to having more rows and less unique listing ids?
- Look at the price column in Calendar table. What transformations you will need to perform so that you can create a column that can be used as a target/response variable?
- Look at the tables Listings, Hosts and Reviews to come up with a list of potential transformations needed in order to have predictors that can be used to predict the listing price.
- Create an aggregated view of data spread across different tables, containing the target as well as predictor variables.

Data Quality and checks (Task 1):

- Once the aggregated dataset has been created, do a data audit. Create a data quality report which has the following basic structure:
    - Continuous Variables: (#unique values, percentage_missing_values, min, max, average, 25th percentile, 75th percentile, 90th percentile, 95th percentile)
    - Categorical Variables: (#Unique values, percentage_missing_values)
- Highlight any data anomaly that you find and fix it.

Variable profiling and checking relationships between variables (Task 2):

- Assess the relationship between target and predictor variables. You can compute correlations, plot bivariate relationships.

- Based on the above analysis summarize your findings and list down the transformations you will do on different predictors, remove the variables from further analysis.

Modelling and insights (Task 2):

- Explain your approach on creating train/test/validation splits.
- Create a comparison matrix to compare different regression models you've run
- Experiment with Linear Regression, Regression Trees, Random Forest Regressor and GBM. Not compulsory but you can also experiment with Xgboost, Lightgbm
- Explain which model you've finalized and why you finalize the model.
- Explain what the top 5 most important predictors are and also explain the direction of impact of these top 5 predictors on the response variable.

Deliverables

- A well-designed deck outlining the conclusions and the analysis (.ppt)
- A well-structured code pushed on github (Write an informative README, well-structured code/notebooks)
- *Optional*: A blog post on medium/personal blog/blogger/linkedin