# MINI PROJECT CANVAS Title (preliminary): Airbnb Price Prediction System. Group members: Ari Kauppi, Yu Wang, Muhammad Rafay Workshop # : 4 .

| MOTIVATION 🎯*Which is the target group of our mini-project? Who is the end-user?  What are their objectives? What needs do we need to address with our work?  How will they benefit from this proposed solution?* Airbnb hosts will be our end users. The objective is to provide customers with compatible options to find the best Airbnb prices. To achieve this, we decided to create a price prediction model that predicts prices based on user needs, allowing customers to quickly find the best options at competitive prices.  This will enable Airbnb hosts to easily set a competitive price for their listings based on its features. | DATA COLLECTION 🧩*Which data sources are we planning to use?  Mention database tables, API methods, websites to scrape, etc.  Which is the data management plan?* We are collecting data from InsideAirbnb.com, which is available in CSV format at the following link:<https://insideairbnb.com/get-the-data/>.  We will not store any personal or sensitive data. Only publicly available data regarding Airbnb listings will be used, and no end-user inputs will be stored.  We are targeting Airbnb listings in Boston, Massachusetts, United States. | PREPROCESSING 🛠*What are the goals of the preprocessing pipeline?*Clean and prepare the data for effective training of the model. Including handling missing values, encoding categorical variables, normalizing numerical data. *Give some examples of data preprocessing steps.*Dropping multiple different measures of the same variable. For example, a measure of *availability*, or how many nights are available to be booked in the next 30, 60, or 90 days. Only one will be kept. *What are some possible data cleaning/wrangling methods you’re planning to use?* Handling missing values by either filling or removing them. Encoding categorical variables into numbers to aid modeling. Normalizing and scaling. Removing outliers. *What are some possible data transformations that could be useful?*Data transformations may not be necessary. We can read the data from CSV. Possibly reading the data into an SQL database could be useful, but we’ve not decided whether it’s necessary yet. *Any feature engineering necessary?* Yes, we will turn some of the features into boolean features (for example, amenities data), normalize numerical data, and encode categorical variables. | EXPLORATORY DATA ANALYSIS (EDA) 🔎*Look at the data!  What steps are you planning to take towards exploring and understanding better the data you have?  What properties would be meaningful to summarize/visualize in this step?* Following are the steps we will take in order to explore our airbnb listing data.  First we will understand the dataset structures of each and every column in the data like shape and datatypes then we will move to, handling missing of different columns in the data, construct histogram to understand skewness in the data, outlier detection and featuring engineering are the final steps  Meaningful properties to visualize:  In this step, it would be meaningful to use a **correlation map** to explore the relationship between various features and price. **Box plots** can help detect outliers and assess variability in the data. Additionally, **scatter plots** will clarify the relationship between price and other key features, providing deeper insights for the model. | VISUALIZATIONS 📊*List any meaningful visualizations you are planning to produce that will be useful to the end user?  Are you planning to produce any interactive visualizations?  If so, which types of interactivity might be useful to the end user?* The Most Interactive Visualization are   * Bar Charts * Line Charts * Heat Maps * Pie Charts * TreeMaps   The following types of interactivity might be useful to the end user:   * Filters * Dynamic Charts * Dropdown Menus   **Price vs. Occupancy Scatter Plots**: Illustrate how pricing affects booking rates, guiding users to balance between higher prices and occupancy levels.  **Heatmaps**: Map out local price ranges, enabling users to set competitive prices based on neighborhood trends.  **Feature-Based Boxplots**: Display how different features (e.g., amenities, property type) impact prices, helping users decide which features to highlight or invest in. |
| --- | --- | --- | --- | --- |
| LEARNING TASK 🐭 (focus on problem definition) *Define the problem setting.  Is this supervised / unsupervised / other…?  Classification / regression / other…?  What are we planning to learn? E.g. What is the target variable / learning outcome?  What variables are we using as input?* This is a supervised learning problem where we are targeting the regression area.  Outcome Learning: Our main aim is to predict Airbnb prices in target areas so users can find the most compatible prices based on their demand.  Currently, we haven't decided on the input variables yet as we will determine them during the EDA steps. We have a total of 60 input variables in the dataset.  Here are a few variables mention below::   * listing\_url * scrape\_id * last\_scraped * source * name * description * neighborhood\_overview * picture\_url * host\_id * host\_url * host\_name * price | LEARNING APPROACH 🪤 (focus on solution implementation) *Which ML/statistical methods seem more relevant for the defined problem setting and why?**Which evaluation metrics could be relevant?**Is any special treatment relevant regarding how we choose to split the data or how we cross-validate?* We will test out our data on different machine learning algorithms like xgboost, randomforest, linear regression and we will select the best one which works well in terms of accuracy. The most relevant methods for predicting Airbnb prices include regression. Our approach is to use a trial-and-error method to determine the best regression model for this data.  XGBoost, for example, should be well suited for price prediction, because real estate pricing frequently has non-linear dependencies, which this model is able to capture. XGBoost also provides insights into important metrics.. Random forest, on the other hand is also suitable as it’s robust to non-linear relationships between features and provides feature importance insights.  **Evaluation methods will be**:   * MSE (Mean Squared Error) * RMSE (Root Mean Squared Error) * MAE (Mean Absolute Error)   We will apply the **k-fold cross-validation** method to split the data. |  | COMMUNICATION OF RESULTS 📢 *Which type of deliverable will benefit most the end-user? Do we choose to write a blog post, create a website, an app, or other..?  How do we communicate best our results to the predefined target group?  Short description of your interface/workflow (if applicable).* We will create a website to introduce our work.  The Airbnb hosts can visit our website to get the price prediction.  We may use some front-end technologies like React to create the website.  The Interface will be simple and clear consist of different inputs columns where the user fill the columns and get the price prediction on the data. | DATA PRIVACY AND ETHICAL CONSIDERATIONS 🔐 (if applicable)*Are there any fairness constraints that apply to our proposed pipeline?  Is there a need to ask for consent during the data collection process?  Is there a need for data pseudonymization/anonymization?  Any other privacy considerations that come to mind?* We will do some work to ensure the fairness constraints in our proposed pipeline. We will consider several aspects like data fairness, algorithmic fairness, model interpretability and outcome fairness.  This data is licensed under a Creative Commons Attribution 4.0 International License. The data can be copied and redistributed for any purpose, even commercially.  Because the data could be accessed on Airbnb's website, there is no need for data anonymization. We may order some links to meet the end users’ needs.  No other privacy considerations that come to mind. |
|  | ADDED VALUE 🎁*Is there a possibility for added value from the data we’re planning to use?**What is the added value?**How are predictions turned into added value for the end-user?* | We provide Airbnb hosts with insights on optimal pricing for their accommodations by analyzing comparable listings in the dataset.  The trained model offers accurate price predictions, which are accessible to the end-user through an easy-to-use website. |  | LEGENDWEEK 1: Data collection/preprocessing  WEEK 2: EDA & visualizations   WEEKS 3-4: Machine/deep learning  WEEK 5: Fairness & data privacy |