Data Collection and Preprocessing Phase

Date	18 July 2024
Team ID	740111
Project Title	Unvieling Airbnb Price Patterns : Machine Learning Models For Forecasting
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Reports

Data Collection Plan: To forecast Airbnb prices using machine learning, the data collection plan involves gathering detailed listing information, pricing, booking data, and user reviews from sources such as the Airbnb API, web scraping, Inside Airbnb datasets, and Kaggle.

Section	Description
Project Overview	The project aims to develop a machine learning model to forecast Airbnb prices by analyzing various factors that influence pricing patterns. The process begins with a comprehensive data collection plan, which involves obtaining detailed information on Airbnb listings, including property features, pricing, and booking history, from multiple sources such as the Airbnb API, web scraping, Inside Airbnb datasets, and Kaggle. Additionally, the project incorporates external data, including local events and economic indicators, to account for factors affecting pricing trends. This approach ensures a robust dataset for training the model, which is crucial for accurately predicting future prices. By analyzing and integrating diverse data sources, the project seeks to uncover patterns and insights that enhance the forecasting accuracy and provide valuable information for both property owners and potential guests.
Data Collection Plan	To forecast Airbnb prices using machine learning, the data collection plan involves gathering comprehensive data from multiple sources. This includes extracting detailed listing information, such as location, property type, and amenities, through the Airbnb API or web scraping. Pricing data, including base rates and seasonal variations, will also be collected via these methods. Booking data, encompassing occupancy rates and booking patterns, along with user reviews from Airbnb or datasets like Inside Airbnb and Kaggle, will be integrated. Additionally, external factors such as local events and economic indicators will be sourced from public databases and local government websites to enrich the dataset and enhance forecasting accuracy.
Raw Data Sources Identified	The raw data sources for this project include datasets obtained from Kaggle & UCI, the popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as gender,

Additionally, integrating local event and economic data will enrich the model by incorporating external factors influencing price trends.

	marital status, income, and loan-related details for machine learning analysis.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Per
Kaggle Dataset	The dataset comprises (location,booking data, hotel review,property type) outcomes.	https://www.kaggle.com/datasets/stevezhenghp/airbnb-price-prediction	CSV	32.77MB	Pub