High Level Design (HLD)

House Price Data Analysis

Revision Number: 1.0
Last date of revision: 16/09/2024

Mayank Yadav
Nitish Kr Dash

Document Version Control

Date Issued	Version	Description	Author
16th Sept 2024	1.0	First Version of Complete HLD	Mayank Yadav

Contents

Abstract

The rise of Airbnb has transformed the way people travel, offering a wide variety of accommodation options, from shared rooms to entire homes. This project focuses on analyzing a dataset of Airbnb listings in **Amsterdam**, exploring key attributes such as pricing, neighborhood influence, room types, reviews, and overall satisfaction. By delving into the patterns of bookings, amenities, and reviews, we aim to provide insights into pricing strategies, neighborhood desirability, and host performance.

This creative data analysis will reveal underlying trends that can help hosts optimize their pricing and improve guest satisfaction while assisting potential guests in making informed decisions.

1 Introduction

1.1 Why this High-Level Design Document?

This High-Level Design (HLD) document provides a comprehensive overview of the Airbnb Data Analysis project. It is intended to guide the data analysis process, focusing on how various attributes such as price, location, reviews, and accommodation features influence booking decisions. The document outlines the key design aspects, tools, and performance metrics, ensuring a structured approach to uncovering valuable insights.

This document will:

- Outline the data structure and provide an overall project design.
- Define the functional and non-functional requirements.
- Offer guidance on how various analysis components interact to drive key insights.

1.2 Scope

The primary goal of this project is to analyze Airbnb listing data from **Amsterdam** to uncover significant factors influencing prices and bookings. The dataset includes details on room types, neighborhoods, reviews, satisfaction levels, and amenities. Key analyses will focus on relationships between these variables and pricing strategies. Using modern data visualization techniques, the findings will assist stakeholders in understanding market trends and making data-driven decisions.

2 General Description

2.1 Product Perspective & Problem Statement

The Airbnb dataset presents a unique opportunity to examine the pricing structure and booking patterns of listings in Amsterdam. The core objective is to explore how factors such as room type, neighborhood, reviews, and accommodations influence price and booking rates.

Key research questions include:

- **Host Insights**: Who are the top earners among hosts? Is there a relationship between price and earnings?
- **Neighborhood Dynamics**: Which neighborhoods attract the most bookings? How does location impact pricing?
- **Reviews and Satisfaction**: What is the relationship between reviews and prices? Do higher review scores correlate with higher prices?
- **Amenities**: How do amenities offered affect pricing?

By addressing these questions, we aim to provide actionable insights for Airbnb hosts and guests alike.

2.2 Tools used

The following tools and technologies will be used to perform the analysis:

- **Python (Pandas, NumPy)**: For data cleaning, analysis, and manipulation.
- **Tableau/Power BI**: For interactive data visualization and dashboards.
- **Jupyter Notebooks**: For coding and reporting results.
- **Excel**: For preliminary data exploration and processing.

3 Design Details

3.1 Functional Architecture

The functional architecture of this Airbnb Data Analysis project can be broken down into five key steps:

a. **Data Acquisition**: Load and explore the Airbnb dataset.
b. **Data Cleaning**: Handle missing values, standardize formats, and filter outliers (e.g., properties with zero reviews or missing location data).
c. **Data Analysis**: Use Python to perform statistical analysis on room types, reviews, pricing, and neighborhoods.
d. **Visualization**: Create compelling visuals using Tableau/Power BI to represent trends and relationships.
e. **Reporting**: Summarize findings and generate interactive dashboards.

3.2 Optimization

To optimize the data analysis process, we will adopt the following strategies:

- **Minimize Redundant Fields: Only the necessary fields (price, reviews, satisfaction, room type, etc.) will be retained to reduce memory usage.**
- **Efficient Queries: Apply efficient filtering and grouping techniques in Python to speed up analysis.**
- **Optimized Visualizations: Simplify charts and reduce the number of filters in Tableau/Power BI for faster load times and smoother interactivity.**
- **Avoid Complex Calculations: Minimize nested calculations during data processing to improve performance.**

4 KPIs

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the disease.

As and when, the system starts to capture the historical/periodic data for a user, the dashboards will be included to display charts over time with progress on various indicators or factors.

To measure and track the progress and insights derived from the dataset, the following KPIs will be used:

1. **Top Earning Hosts**: Identify which hosts earn the most and why.
2. **Price vs. Neighborhood**: Analyze how prices vary across different neighborhoods in Amsterdam.
3. **Review Impact**: Assess the influence of reviews and satisfaction scores on pricing.
4. **Booking Trends**: Determine which room types and neighborhoods receive the most bookings.
5. **Price vs. Amenities**: Explore how different amenities affect pricing decisions.

5 Deployment

Once the analysis is complete, the findings will be deployed through interactive dashboards and reports:

- **Tableau/Power BI Dashboards**: These dashboards will allow stakeholders to filter data by neighborhood, room type, price, and more, providing an intuitive exploration of key trends.
- **Jupyter Notebooks**: Detailed analysis and insights will be shared in notebook format, offering code and visualizations.
- **Excel Reports**: For those preferring spreadsheet formats, key metrics and trends will be summarized in Excel.

The results will empower hosts to optimize their listings and potential guests to make well-informed booking decisions.