CAPSTONE PROJECT

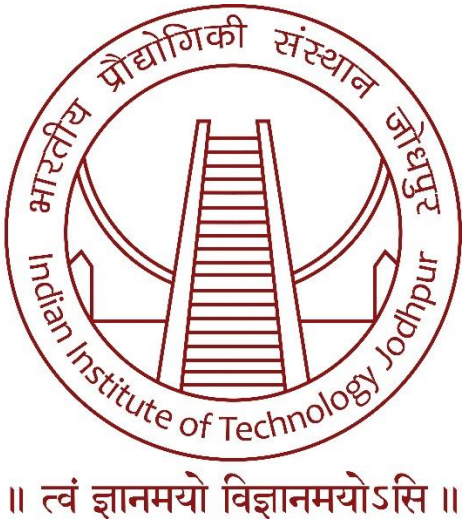
AIRBNB DATA CLEANING, PROCESSING, AND ADVANCED ANALYSIS

A POST GADUATION DIPLOMA

PROJECT PROPOSAL REPORT

SUBMITTED BY

NIRAJ BHAGCHANDANI [G23AI2087]
SHUBHAM RAJ [G23AI2082]
BHAVESH ARORA [G23AI2126]
PARAS PANDA [G23AI2117]
JATIN SHRIVAS [G23AI2094]
JAI SINGH KHUSHVAH [G23AI2018]



SUBMISSION DATE: *20TH OCTOBER, 2024*

DEPARTMENT OF ARTIFICIAL INTELLIGENCE DATA ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY - JODHPUR

Table of Contents

Introduction

Overview

This project focuses on cleaning, processing, and performing advanced analysis on Airbnb datasets, which include details about listings, calendars, and reviews. The objective is to improve the overall data quality and extract meaningful insights that can help in decision-making. We will work through various phases, starting from understanding the dataset to implementing advanced machine learning models.

Purpose and Objectives

The purpose of this project is to ensure that the raw Airbnb data is transformed into a format that can be easily analyzed and understood. The key objectives include:

- Identifying and fixing data quality issues such as missing values and inconsistencies.
- Integrating multiple datasets (listings, reviews, calendars) for a holistic view.
- Using advanced techniques to predict trends, user behavior, and potential revenue optimization.

Problem Statement

Airbnb's dataset is quite large and diverse, making it prone to several issues like missing entries, outliers, and unstructured formats. These problems affect the ability to derive accurate insights and make data-driven decisions. The challenge lies in cleaning and transforming the data, then developing machine learning models to predict useful outcomes, such as revenue trends and booking behaviors.

Project Scope

Data Cleaning

The first step is to ensure that the data is free from errors, such as missing values, duplicated entries, or inconsistent formats. We will identify these issues using exploratory data analysis (EDA) techniques and fix them by using Python libraries like Pandas and Scikit-learn. This step will improve the overall integrity and usability of the dataset.

Data Processing

After cleaning, the next step is processing the data to prepare it for analysis. This involves transforming raw data into a structured format, consolidating multiple datasets into one, and ensuring that the data is standardized. We will also augment the data using techniques such as synthetic data generation if needed.

Advanced Analysis

Once the data is cleaned and processed, we can move towards advanced analysis. This involves developing machine learning models to predict key business outcomes. Examples include predicting occupancy rates, estimating potential revenue, or analyzing user reviews to determine satisfaction levels. We will use algorithms like regression and classification, depending on the specific business problem.

Methodology

Phase 1: Exploratory Data Analysis (EDA)

In this phase, we will perform an initial analysis of the Airbnb dataset to understand its structure and identify any data quality issues. EDA helps us visualize the data, discover trends, and detect outliers. We'll use visual tools like heatmaps, histograms, and box plots to explore the data.

Phase 2: Data Cleaning and Augmentation

This phase involves addressing the issues identified during EDA. We'll clean the dataset by filling missing values, removing duplicates, and handling outliers. Additionally, we may need to augment the data by generating synthetic data points, especially where key information is missing or insufficient.

Phase 3: Data Integration and Optimization

Once the data is cleaned, we will integrate the various Airbnb datasets (listings, calendars, reviews) into one comprehensive dataset. This step will involve merging and resolving inconsistencies. We will also optimize the dataset for efficient querying and analysis, which may involve techniques like indexing and partitioning.

Phase 4: Advanced Analytics and Reporting

In the final phase, we will perform advanced analytics using machine learning models and data visualization tools. Predictive models will be developed to provide insights such as future booking trends or price optimizations. The results will be presented using interactive dashboards (e.g., in Power BI or Tableau) to facilitate business decision-making.

Task Breakdown

Phase-by-phase Division of Work and Time Allocation

The project is divided into four main phases, each focusing on a specific aspect of the data analysis process:

- **Phase 1 (Week 1-3): Exploratory Data Analysis (EDA)**

During this period, the team will conduct a deep exploration of the dataset. This includes identifying missing values, outliers, and patterns using tools like Python and Pandas. The goal here is to understand the data thoroughly before diving into cleaning.

- **Phase 2 (Week 4-5): Data Cleaning and Augmentation**
  In this phase, the team will address all issues identified during EDA. This involves handling missing values, eliminating duplicate records, and normalizing the data for further analysis. Additionally, synthetic data generation may be applied where necessary to fill in gaps.
- **Phase 3 (Weeks 6-7): Data Integration and Optimization**
  Once the data is cleaned, the next step is to merge the various datasets (listings, reviews, and calendar data) into a single comprehensive dataset. During this phase, we will also focus on optimizing the dataset for storage and processing efficiency.
- **Phase 4 (Weeks 8-9): Advanced Analytics and Reporting**
  The final phase will involve applying machine learning models to analyze the data and generate actionable insights. This phase also includes creating visual reports and dashboards to present findings clearly.

Timeline

Phase	Timeline	Milestones
Phase 1: Exploratory Data Analysis	Weeks 1-3	Completion of initial EDA, data profiling report
Phase 2: Data Cleaning & Augmentation	Weeks 4-5	Cleaned dataset, handling of outliers and missing values
Phase 3: Data Integration & Optimization	Weeks 6-7	Integrated datasets, optimized data pipeline
Phase 4: Advanced Analytics & Reporting	Weeks 8-9	Machine learning models, BI dashboards, final report

Roles and Responsibilities

Specific Tasks for Each Team Member

- **Niraj Bhagchandani** (Team Leader)
  Responsible for managing the overall project timeline, ensuring smooth communication among team members, and leading the integration of different datasets.
- **Shubham Raj**
  Primarily focused on conducting exploratory data analysis and handling the initial data cleaning processes.
- **Bhavesh Arora**

In charge of developing machine learning models, training and testing the models to ensure predictive accuracy.

- **Paras Panda**
  Will handle data visualization and reporting, developing dashboards and clear visual representations of insights.
- **Jatin Shrivas**
  Focuses on cloud infrastructure setup and optimizing data storage, using tools like AWS or Google Big Query.
- **Jai Singh Khushvah**
  Responsible for feature engineering, including creating new features from existing datasets and optimizing the machine learning models.

Expected Outcomes

By the end of this project, the team expects to deliver:

- A thoroughly cleaned and well-structured Airbnb dataset, ready for analysis.
- Predictive models capable of estimating trends such as booking behaviors and potential revenue.
- Comprehensive insights into Airbnb listings, user reviews, and occupancy trends.
- Interactive business intelligence (BI) dashboards to allow stakeholders to easily explore key metrics and results.
- A detailed final report summarizing the methodology, findings, and actionable recommendations for Airbnb.

Conclusion

This project aims to address critical issues in the Airbnb dataset through a structured approach involving data cleaning, augmentation, and advanced analysis. By following a phased methodology, the team will be able to deliver clean data, powerful machine learning models, and meaningful insights. The expected outcomes will not only improve data usability but also provide Airbnb with valuable insights for decision-making and revenue optimization.