**Air BnB Data Pipeline**

**Abstract**

This project is focused on designing and implementing a comprehensive data engineering pipeline to analyze Airbnb listings using publicly available datasets provided by Inside Airbnb. The primary objective is to extract valuable insights from large and unstructured datasets by transforming them into a clean, organized, and easily analysable format. The raw datasets are obtained in CSV format and contain information about thousands of Airbnb listings across multiple cities, including details such as listing prices, availability, property types, number of reviews, host information, and geographic location. Since the raw data can contain inconsistencies, missing values, and formatting issues, the project applies extensive preprocessing and data cleaning using Python and pandas. This involves detecting and handling missing or incorrect entries, converting data types to appropriate formats, standardizing categorical and numerical features, and structuring the data into a well-organized schema suitable for storage and analysis. By performing these steps, the project ensures that the data is reliable, consistent, and ready for further analytical processes.

After cleaning and structuring, the processed data is stored in a SQLite database, which provides a lightweight and efficient solution for data storage and retrieval. This database allows for easy querying of specific subsets of data and supports subsequent analysis tasks. Using the structured dataset, the project investigates trends and patterns within Airbnb listings, focusing on metrics such as pricing variations, availability rates, and property type distribution across different locations. By analyzing these patterns, the project provides insights into factors that influence pricing, occupancy, and listing popularity, enabling data-driven understanding of the short-term rental market. Additionally, this work highlights practical applications of data engineering techniques, including data ingestion, transformation, storage, and analysis, using a straightforward yet powerful tech stack consisting of Python, pandas, and SQLite.

Overall, this project demonstrates how real-world data can be transformed from raw, unorganized formats into meaningful information that can support decision-making and market analysis. The approach emphasizes the importance of clean, structured data and efficient storage mechanisms for deriving actionable insights. By implementing this data engineering pipeline, the project not only showcases technical skills in data processing but also provides a framework for future studies and applications, such as predictive modelling, pricing optimization, and trend forecasting in the short-term rental industry. The methodologies applied here can be extended to similar large-scale datasets in other domains, illustrating the relevance and scalability of the techniques used. In conclusion, this work serves as a practical example of applying data engineering principles to real-world datasets, delivering both technical and analytical value while providing a foundation for informed decision-making and deeper understanding of market dynamics in the Airbnb ecosystem.

Y Rushita – 2320030091

Mrinalini B – 2320030309

A Bala Aditya - 2320030043