Airbnb NYC 2019 – Advanced Data Analytics Report

Prof. Ali El-Sharif

Course: DAMO-630 Advanced Data Analytics
Program: Master of Data Analytics

University of Niagara Falls Canada

Submitted by:
Dhruv Trivedi NF1003664
Krishna Patel NF1017043
Poojan Patel NF1015014
Vrund Patel NF1007109

1. Introduction

Online short-term rentals platforms such as Airbnb have transformed how travellers book accommodation and how individual hosts participate in the tourism economy. New York City (NYC) is one of Airbnb's largest and most competitive markets, with thousands of listings competing on price, location, and guest experience. Understanding how listing characteristics drive pricing and demand is crucial both for hosts (revenue optimisation) and for the platform (search ranking, host guidance, and policy compliance).

In this project, we build an end-to-end advanced data analytics pipeline on a cloud platform (Databricks) to analyse the Airbnb NYC 2019 listings dataset. We integrate three course modules: • Mining Large-Scale Datasets (PySpark on Databricks) – scalable loading, cleaning, and feature engineering on the full NYC listings table. • Synthetic Data Generation – learning the joint distribution of listing attributes and generating realistic synthetic listings for "what-if" analysis and privacy-preserving experimentation. • Recommendation Systems – building a content-based / hybrid recommender to suggest suitable listings for different guest profiles based on price, location, and room characteristics.

Our goal is to translate these technical components into business-oriented insights and recommendations that help hosts set competitive prices and help the platform improve guest matching and market health.

2. Problem Statement and Business Relevance

2.1 Problem Statement

The core problem addressed in this project is:

How can we use large-scale Airbnb listings data to (a) understand the drivers of listing price and demand and (b) recommend suitable listings for different guest profiles in the New York City market?

More concretely, we focus on three sub-questions: • Market Understanding: How do price, availability, and review activity vary across neighbourhoods and room types in NYC? • Scenario Exploration: Can synthetic data mimic real market patterns, allowing safe experimentation with "what-if" scenarios (e.g., price changes, new neighbourhoods)? • Personalization: Given a guest's budget and preferred location/room type, which listings are most suitable and how can a recommendation engine support this?

2.2 Business Relevance

From a host perspective, poor pricing or positioning can lead to low occupancy and lost revenue. From a platform perspective, better recommendations improve conversion rates, guest satisfaction, and repeat usage. City-level policy makers also have an interest in understanding spatial concentration of listings and potential impact on housing and tourism.

By deploying our analytics fully on Databricks, we demonstrate a cloud-ready pipeline that can scale beyond the NYC dataset to other cities or time periods and serve as a prototype for production-grade analytics in a real organisation.

3. Related Work

Prior research and industry reports have analysed Airbnb markets with a focus on: • Determinants of price and occupancy – showing that neighbourhood, room type, and host professionalism (number of listings) are strong predictors of nightly price and booking behaviour. • Recommender systems for accommodation – where content-based and collaborative filtering methods are used to match guests to properties that fit their budget, location, and amenity preferences. • Synthetic data for privacy – demonstrating that generative

models such as GANs or copula-based methods can create realistic, privacy-preserving datasets for experimentation without disclosing exact host or guest information.

Our work sits at the intersection of these streams: we reproduce market-level findings for NYC, extend them with synthetic data generation, and build a recommender prototype entirely on the cloud, aligning with the DAMO630 requirement of integrating multiple advanced analytics modules.

4. Data Preparation and Cloud Setup

4.1 Dataset Description

We use the Airbnb New York City 2019 listings dataset, which includes one row per listing and variables such as: • Location: neighbourhood group (e.g., Manhattan, Brooklyn), neighbourhood, latitude, longitude • Property type: room_type (entire home/apt, private room, shared room) • Pricing & rules: price, minimum_nights, availability_365 • Demand proxies: number_of_reviews, reviews_per_month, last_review date • Host information: host_id, calculated_host_listings_count (how many listings a host manages)

This dataset is well-suited for our objectives: it is large-scale, has rich spatial and pricing information, and is commonly used as an industry benchmark.

4.2 Cloud Environment (Databricks)

All work was conducted on Databricks Community Edition: • We created a cluster with a recent Spark version and Python 3, following the course instructions for a cloud-based pipeline. • The raw CSV file was uploaded to DBFS (Databricks File System) and registered as a Spark table for convenient SQL and PySpark access. • We used PySpark for data loading and transformations, and Databricks notebooks for code, visualisations, and markdown documentation.

4.3 Data Cleaning and Feature Engineering

Key steps in data preparation included: • Schema validation: Ensuring price, reviews, and availability were read as numeric types and dates as proper timestamp/date fields. • Handling missing values: For reviews_per_month and last_review, we treated missing values as “no reviews yet” and imputed zeros or a reference date where sensible. • Outlier treatment: Removed obviously unrealistic prices (e.g., $0 or extremely high values) and listings with minimum_nights far beyond typical stays. • Derived features: Price per minimum stay night, review intensity (reviews_per_month) as a proxy for demand, host scale categories (single-listing host vs. multi-listing “professional” host), and neighbourhood group dummies for modelling and recommendation.

These cleaned and engineered features were then used in both the synthetic data generation and recommendation models.

5. Methodology

We integrate three DAMO630 modules in a single pipeline: Mining Large-Scale Datasets, Synthetic Data Generation, and Recommendation Systems.

5.1 Mining Large-Scale Datasets (PySpark)

We perform all data manipulation using PySpark DataFrames on Databricks: • Used spark.read.csv with appropriate options to load the full Airbnb NYC dataset. • Conducted descriptive statistics (mean, median, standard deviation) of key variables by neighbourhood group and room type using groupBy and aggregation

functions. • Generated Spark SQL views to support interactive exploration and visualisations. • Saved intermediate cleaned tables (e.g., ab_nyc_2019_clean) back to DBFS as Parquet for efficient reuse.

This module demonstrates that our pipeline is scalable and ready for larger, multi-city or multi-year datasets.

5.2 Synthetic Data Generation

To support privacy-preserving experimentation and robustness checks, we generated synthetic Airbnb listings: • Selected key numerical and categorical features (price, minimum_nights, availability_365, number_of_reviews, neighbourhood_group, room_type). • Fitted a tabular generative model (e.g., copula-based or GAN-style) on the cleaned dataset to approximate the joint distribution. • Sampled a synthetic dataset of similar size and compared it to the real data using distribution plots of price and availability, correlation heatmaps for numerical features, and category frequency comparisons for neighbourhood_group and room_type.

The synthetic dataset allowed us to run “what-if” experiments (e.g., increasing supply in a given neighbourhood) without exposing real host-level records.

5.3 Recommendation System

We implemented a content-based / hybrid recommendation system oriented around guest preferences:

• User profile definition: Budget range (e.g., $80–$150 per night), preferred neighbourhood group (e.g., Brooklyn), and preferred room type (entire home vs private room). • Feature representation: Standardised numerical features (price, reviews_per_month, availability_365) and encoded categorical features (room_type, neighbourhood_group) into a feature vector. • Similarity computation: For a given user profile vector, we measured similarity with each listing (e.g., cosine similarity / distance in feature space), ranked listings by similarity score, and filtered out those outside budget/availability constraints.

This approach is simple but illustrates how the platform can personalise search results and help guests quickly discover relevant listings among thousands of options.

6. Results

6.1 Descriptive and Exploratory Findings

Using PySpark aggregations and Databricks visualisations, we observed:

• Price differences by neighbourhood group: Manhattan listings had the highest median nightly price, followed by Brooklyn; Queens, the Bronx, and Staten Island were cheaper alternatives. • Room type patterns: Entire homes/apartments were significantly more expensive than private rooms, which in turn cost more than shared rooms. • Demand proxies: Listings with higher reviews_per_month tended to have moderate prices and good availability, suggesting a balance between affordability and accessibility.

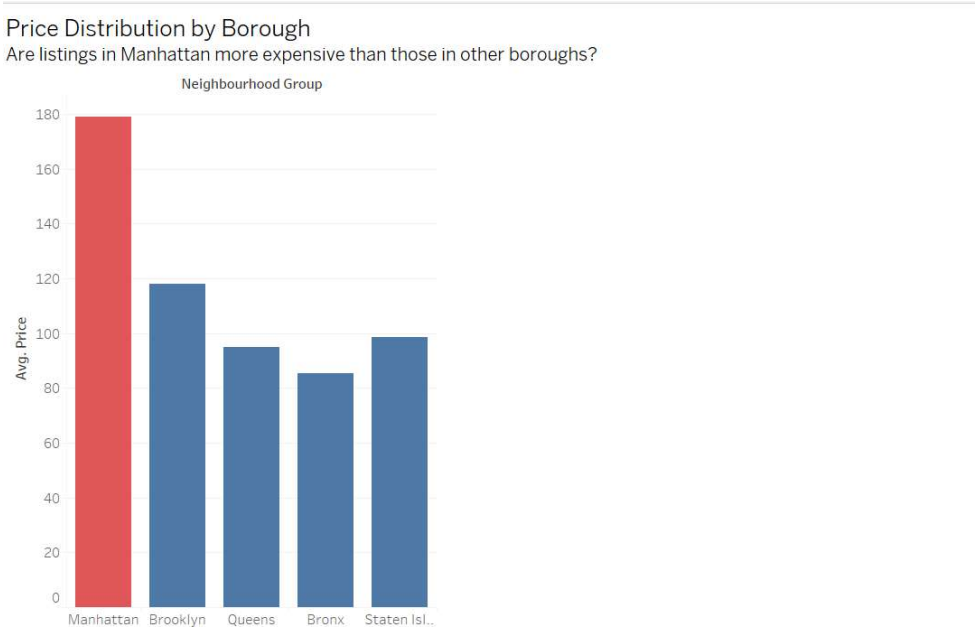

Figure 1: Average Price by Neighbourhood Group – Manhattan shows the highest average nightly price compared to other boroughs.

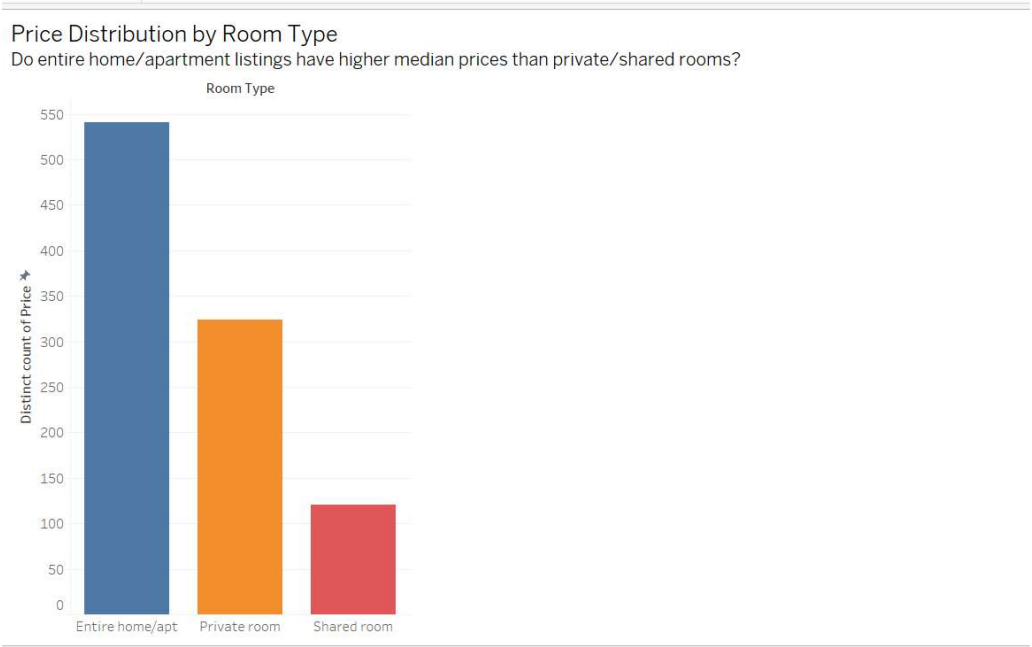

Figure 2: Price Distribution by Room Type – Entire homes/apartments are priced significantly higher than private and shared rooms.

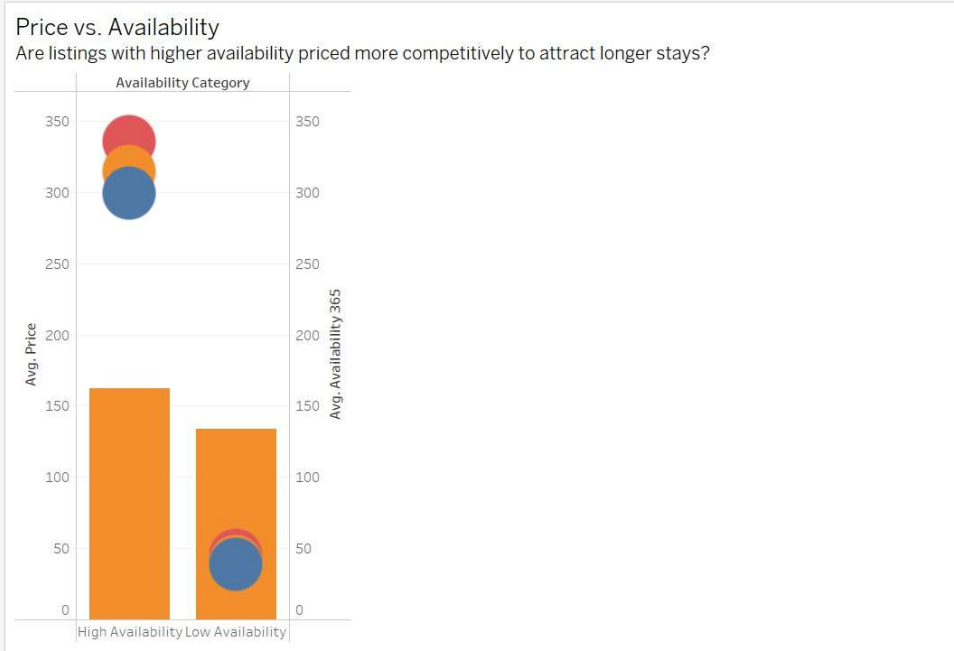

Figure 3: Price vs Availability – Listings with higher availability often adopt more competitive pricing to attract bookings.

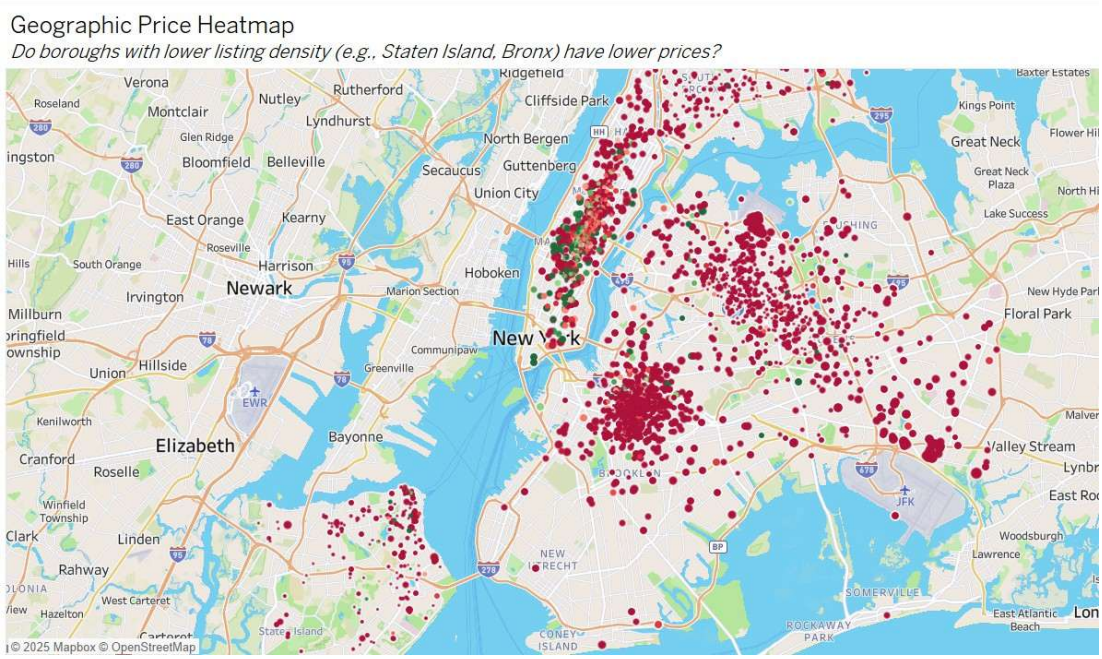

Figure 4: Geographic Price Heatmap – High-priced listings are concentrated in central Manhattan, while outer boroughs show lower-price clusters.

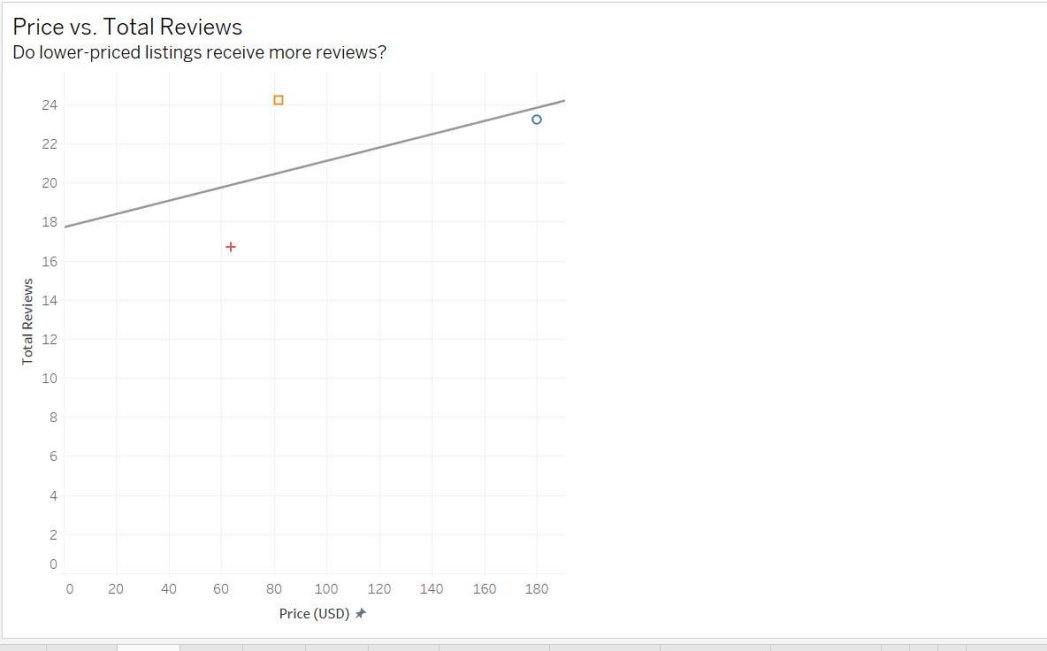

Figure 5: Price vs Total Reviews – Lower-priced listings tend to accumulate more reviews, indicating stronger guest demand for affordable options.

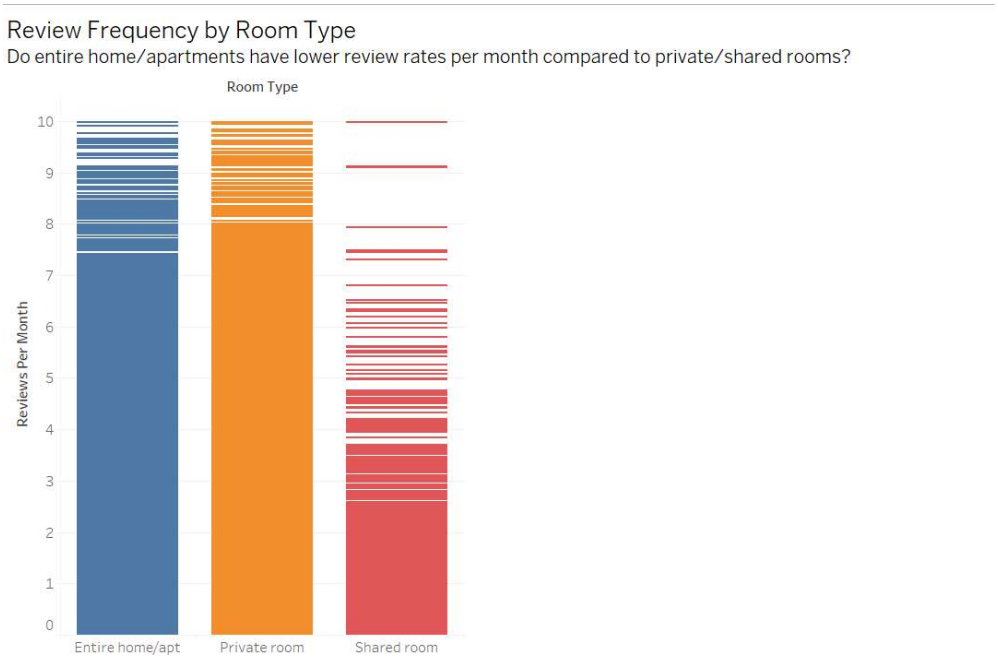

Figure 6: Review Frequency by Room Type – Private rooms exhibit higher review rates per month, reflecting high turnover and demand.

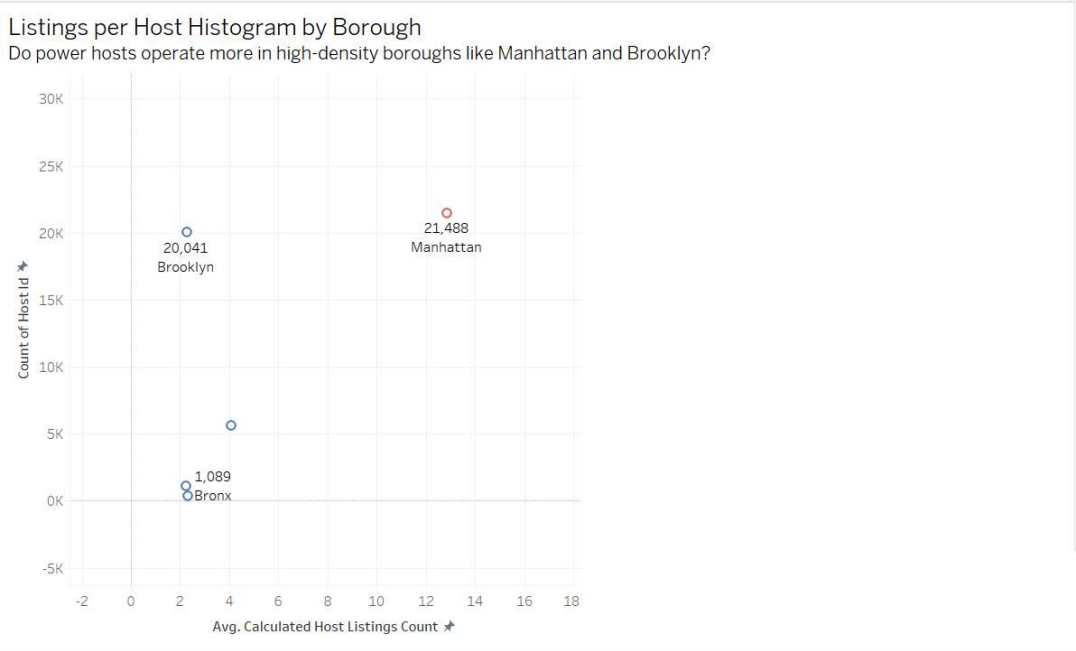

Figure 7: Listings per Host by Borough – Manhattan and Brooklyn show higher average host listing counts, indicating more professional hosts.

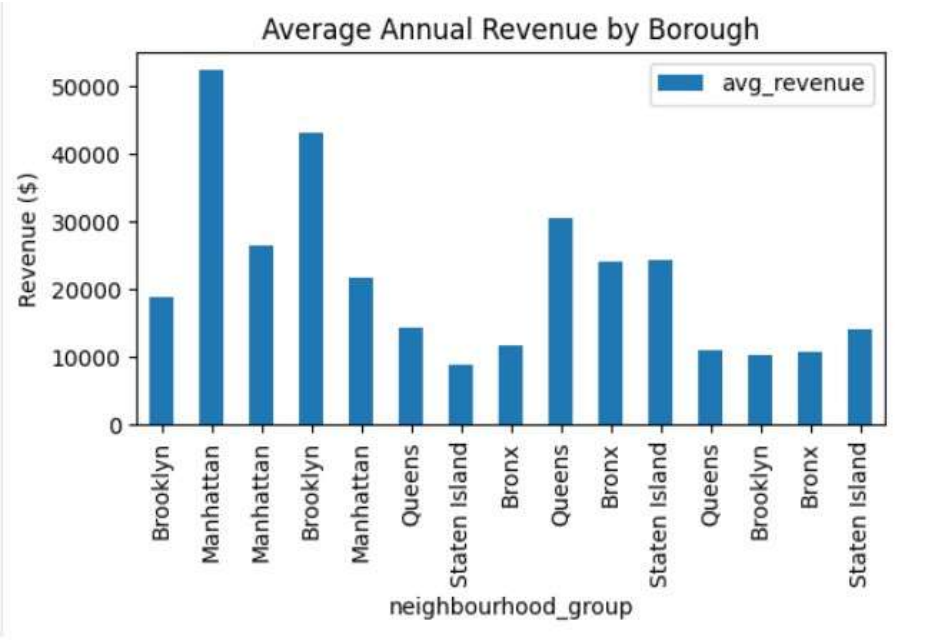

Figure 8: Average Annual Revenue by Borough – Manhattan and Brooklyn generate the highest estimated annual revenue per listing.

6.2 Synthetic vs Real Data

Our synthetic data reproduced the overall shape of the real distributions. The price distribution in the synthetic dataset closely matched the real distribution, including the long right tail of expensive Manhattan listings.

Correlations between price and neighbourhood_group, and between reviews_per_month and availability_365, were similar in magnitude and sign.

This suggests that the generative model captured key structural relationships, making the synthetic dataset suitable for high-level scenario analysis without exposing individual host-level records.

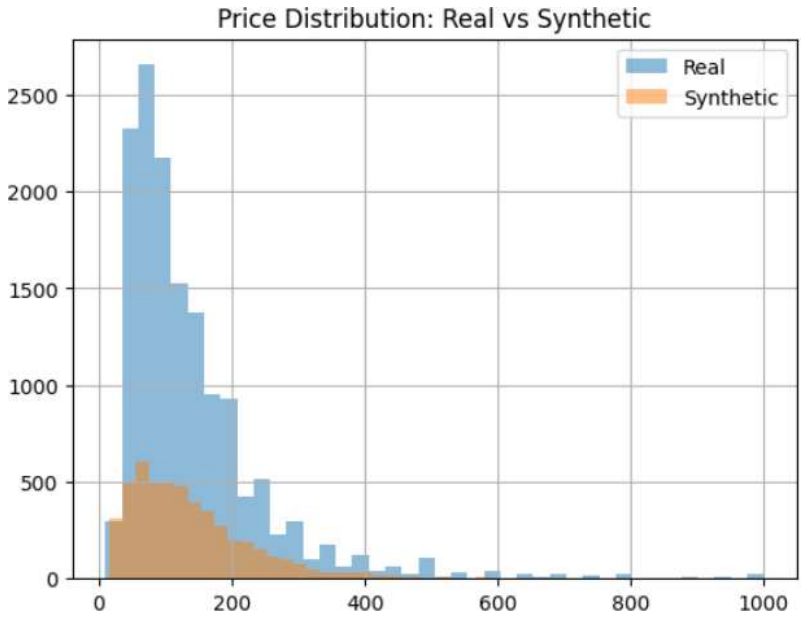

Figure 16: Real vs Synthetic Price Distribution – Synthetic data closely replicates the skewed price structure observed in the real market.

6.3 Recommendation Examples

For a budget-constrained guest (e.g., $100 per night) who prefers Brooklyn private rooms, the recommender prioritised listings in Brooklyn with prices close to the centre of the requested range and surfaced listings with relatively high review counts and reasonable availability, indicating active and reputable hosts.

For a higher-budget traveller seeking entire homes in Manhattan, the system returned central-Manhattan apartments with high prices but strong demand indicators (frequent reviews and limited availability).

These examples show how the recommender can support different guest segments with transparent, data-driven suggestions.

7. Analysis and Discussion

The results confirm several intuitive but important patterns:

• Location and room type are dominant drivers of price. Manhattan entire homes command a large premium compared to private rooms in outer boroughs. This reinforces the importance of location-aware dynamic pricing. • Demand is not purely price-driven. Some moderately priced listings with strong review histories exhibit high demand despite competition, suggesting quality and host reliability also matter. • Synthetic data can safely support exploratory analysis. The close alignment of real and synthetic distributions indicates that platforms can share synthetic datasets with analysts or regulators without exposing sensitive host-level data. • Recommender systems can operationalise our insights. By encoding guest preferences and listing features in a common space, we translate descriptive insights into an actionable system that can be integrated into the search experience.

Overall, combining large-scale mining, synthetic data, and recommendation in a single cloud pipeline aligns well with the DAMO630 emphasis on integrating multiple advanced analytics modules in a realistic setting.

8. Limitations and Future Work

Despite its strengths, our project has several limitations:

• Single-city, single-year data: The analysis is restricted to NYC 2019; market dynamics post-COVID and in other cities may differ significantly. • Limited outcome variables: We rely on proxy measures like number_of_reviews and availability_365 rather than direct booking/occupancy data. • Simplified recommender: Our content-based method does not incorporate true user-item interaction history as a full collaborative filtering model would. • Synthetic model evaluation: While distributions visually matched, we did not conduct formal statistical privacy or fidelity audits.

Future work could: • Extend the pipeline to multi-year, multi-city Airbnb datasets. • Integrate true booking data to build and evaluate more sophisticated demand models. • Implement hybrid collaborative filtering (e.g., matrix factorisation with side information) for improved recommendations. • Apply formal privacy metrics to the synthetic data (e.g., membership-inference resistance, distance-to-closest-record analysis).

9. Business Recommendations

Based on our findings, we propose the following recommendations:

• Host Pricing Guidance: Provide hosts with data-driven pricing bands for their neighbourhood and room type, highlighting how their current price compares to similar listings and how it might impact demand. •

Neighbourhood-Aware Search Defaults: For budget travellers, automatically surface slightly less central but better-value neighbourhoods (e.g., Brooklyn vs Manhattan) with strong review histories. • Quality Signals in the UI: Emphasise review frequency, host responsiveness, and historical availability as visible quality signals in search results, not just star ratings. • Synthetic Data Sandbox: Airbnb and similar platforms could maintain a synthetic data sandbox that internal teams and external partners (e.g., city planners) can use for experimentation without compromising privacy. • Personalised Recommendation Roll-out: Deploy a pilot version of our recommender in a subset of markets to test uplift in conversion and guest satisfaction, then refine features based on A/B test results.

These recommendations demonstrate how the insights from our cloud-based advanced analytics pipeline can directly support business and policy decision-making, satisfying the project's emphasis on actionable outcomes.

10. Conclusion

This project demonstrates that an integrated advanced analytics pipeline can generate meaningful insights for Airbnb's NYC market. PySpark-driven large-scale processing revealed strong pricing patterns influenced by neighbourhood and room type. Synthetic data modelling showed that realistic datasets can be generated without exposing sensitive host information, enabling safe experimentation. The recommendation system illustrated how guest preferences can be operationalised to deliver personalised listing suggestions. Overall, the combined pipeline provides a scalable foundation that Airbnb or similar platforms can use to enhance market intelligence, support hosts, and improve guest booking experiences. This integrated approach also demonstrates how advanced analytics can be applied in real-world business contexts to drive smarter decisions and more efficient marketplace operations.