Data Workflow & Market Analysis

Marco Flavio Delgado Martinez

Udacity Institute of AI and Technology

Capstone Project 1

February 2026

ABSTRACT

In this project, I established a reproducible data engineering pipeline to analyze the New York City Airbnb market using 2019 listing data. My primary objective was to determine how listing prices vary across boroughs and whether high market saturation correlates with reduced pricing power. The workflow involved cleaning raw data, imputing missing review metrics to preserve inventory, and removing pricing outliers to ensure statistical validity. My analysis reveals that while Manhattan and Brooklyn account for over 85% of the market saturation, Manhattan retains significantly higher pricing power compared to other boroughs. This report details the data engineering decisions, ethical considerations regarding missing data, and the strategic implications of the findings.

Data Workflow & Market Analysis

Overview

The ability to transform raw data into actionable intelligence is a core competency in modern AI and marketing strategy. In this capstone project, I utilized the New York City Airbnb Open Data dataset to build a complete data engineering workflow. My goal was to move beyond simple descriptive statistics and uncover the relationship between supply (listing volume) and value (price) across New York City's five boroughs. By standardizing the data ingestion and cleaning process, I ensured that the resulting insights are both reproducible and statistically robust.

Dataset Description

The dataset comprises approximately 49,000 unique listings from the year 2019. It provides a snapshot of the rental market, including variables such as price, location (latitude/longitude and neighborhood), room type, and host activity metrics. For this analysis, I focused specifically on the price variable as the primary target for revenue analysis and neighbourhood_group (borough) as the key segmentation variable. I also utilized reviews_per_month and availability_365 as proxies for listing activity and inventory pressure.

Workflow and Methodology

My workflow followed a modular design to ensure scalability.

- **Ingestion:** I loaded the raw CSV data using the Pandas library (McKinney, 2013), performing immediate validation checks to confirm the schema and row count.
- **Cleaning and Governance:** I implemented functions to address data quality issues. Specifically, I identified that null values in the review columns represented zero activity rather than missing error data.
- **Exploratory Analysis:** I aggregated the clean data to calculate key performance indicators, focusing on the density of listings per borough.
- **Visualization:** I employed a custom color palette and specific statistical plot types (such as violin plots) using the Seaborn library (Waskom, 2021). Furthermore, I utilized the Matplotlib framework (Hunter, 2007) to generate the foundational geographic scatter maps required for the spatial analysis.

Key Decisions and Assumptions

A critical governance decision involved the handling of missing values in the reviews_per_month column, which affected approximately 20% of the dataset. I chose to impute these values with zero rather than dropping the rows. My reasoning was that dropping these records would eliminate new market entrants who have not yet received a review, thereby creating "survivorship bias" where only established hosts are analyzed (McKinney, 2013). Preserving these rows ensures a more accurate reflection of the total market supply. Additionally, I removed listings with a price of $0 (errors) and capped prices at the 99th percentile to prevent luxury outliers from skewing the average price metrics.

Results and Interpretation

The analysis demonstrated a clear bifurcation in the market. Manhattan and Brooklyn are heavily saturated, containing the vast majority of active listings. Despite this competition, Manhattan maintains the highest pricing power, characterized by a high median price and a significant long-tail distribution of high-value properties. In contrast, Queens and the Bronx display high price elasticity, with rates clustered tightly around the lower end of the spectrum. This suggests that while Manhattan listings compete on features and location, listings in the outer boroughs compete primarily on price.

Responsible Practice and Bias

I identified potential bias in the availability_365 metric. A value of zero could indicate a fully booked property (success) or an abandoned listing (churn). By treating all zero values as valid inventory, my analysis may overestimate the active supply. Furthermore, the decision to remove the top 1% of pricing outliers means that my findings apply to the general consumer market but do not accurately reflect the dynamics of the ultra-luxury segment.

Reproducibility

To ensure this project can be replicated by other data engineers, I have included a requirements.txt file specifying the exact versions of Pandas, Matplotlib, and Seaborn used. The analysis is contained within a Jupyter Notebook (data_workflow.ipynb) that executes linearly from data ingestion to visualization. Version control was managed via Git, with specific feature branches used to isolate the development of exploratory analysis functions.

References

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90-95. 10.1109/MCSE.2007.55

McKinney, W. (2013). *Python for Data Analysis*. O'Reilly Media, Incorporated.

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 1-4. 10.21105/joss.03021