Hüseyin Soykök – 3804308 – Group 24

Milestone 1 Report: A Geographical and Economic Analysis of the New York City Airbnb Market

A. Title & Source

Title: New York City Airbnb Open Data (2019)

Primary URL: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Publisher/Author: Dgomonov (Kaggle user)

Publication or Last-Update Date: The data is a snapshot from 2019. It was uploaded to Kaggle approximately 5 years ago.

License: CC0 1.0 (Public Domain Dedication). This license permits the free use, modification, and distribution of the dataset for any purpose, including academic use, without restriction.

B. Motivation

As an avid traveller who frequently utilizes Airbnb for accommodation, the rich data offered by this platform holds significant personal relevance. Unlike the often-uniform pricing of hotels, the flexible, budget-diverse options on Airbnb are what I find most compelling as a traveller.

My primary motivation for this project is to uncover the "price-performance" balance a key concern for any traveller using the power of data visualization. My goal is to transform the intuitive, and often subjective, search for the 'best' accommodation into a data-driven, visual analysis. The raw latitude and longitude coordinates within the dataset provide a unique opportunity to create rich, interactive maps that reveal the very "shape" of New York City's pricing, popularity, and rental-type landscape.

The findings from this analysis will serve as a personal guide for my future travel decisions while also fulfilling my objective of understanding the geographical dynamics of a rental market.

I plan to explore two concrete, high-potential questions through visualization:

1. How is listing price geographically distributed across New York City? Which neighbourhoods (neighbourhood) constitute the most expensive "red zones," and how does this price map change dramatically when filtered by room_type ('Entire home/apt' vs. 'Private room')?

2. Where are the true "value" or "price-performance" hotspots? That is, are listings that are both affordable (e.g., price < $100) and popular (e.g., number_of_reviews > 100) geographically clustered in specific neighbourhoods (e.g., in Queens or the deeper parts of Brooklyn)?

C. Scope & Granularity

- **Size:** The dataset consists of 48,895 items (rows/listings) and 16 attributes (columns). This size is substantial enough for in-depth analysis yet manageable without requiring heavy preprocessing.
- **Unit of Analysis:** Each item (row) represents a **single, unique Airbnb listing** in New York City as of 2019.
- **Keys:** The id attribute is a primary key, uniquely identifying each listing. The host_id identifies the host; this key is not unique per row, as one host can have multiple listings.

D. Schema (Types & Ranges)

Attribute	**Role**	**Data Type &Semantics**	**Unit**	**Range / Domain**	**Example**
id	ID/Key	Quantitative	Unique ID	48,895 unique integers	2539, 2595
name	Attribute	Categorical	Text	~47,900+ unique strings	Skylit Midtown Castle
host_id	ID/Key	Quantitative	Unique ID	37,457 unique hosts	2787, 2845
host_name	Attribute	Categorical	Text	~11,400+ unique strings	John, Jennifer
neighbourhood_group	Attribute	Categorical (Spatial)		5 Categories: Manhattan (21.6k), Brooklyn (20.1k), Queens (5.6k), Bronx (1.0k), Staten Island (0.3k)	Manhattan
neighbourhood	Attribute	Categorical (Spatial)		221 Categories	Midtown, Harlem
latitude	Attribute	Quantitative (Spatial)	Degree (Lat)	[40.499, 40.913] (NYC Bounds)	40.75362
longitude	Attribute	Quantitative (Spatial)	Degree (Lon)	[-74.244, -73.712] (NYC Bounds)	-73.98377
room_type	Attribute	Categorical		3 Categories: Entire home/apt (25.4k), Private room (22.3k), Shared room (1.1k)	Entire home/apt
price	Attribute	Quantitative	USD ($)	[0, 10000] (Contains outliers)	149, 225, 89
minimum_nights	Attribute	Quantitative	Nights	[1, 1250] (Contains outliers)	1, 3, 10
number_of_reviews	Attribute	Quantitative	Count	[0, 629]	9, 45, 270
last_review	Attribute	Quantitative (Temporal)	Date	[2011-03-28, 2019-07-08]	2018-10-19
reviews_per_month	Attribute	Quantitative	Count/Month	[0.01, 58.5]	0.21, 4.64
calculated_host_listings_count	Attribute	Quantitative	Count	[1, 327] (Total listings by host)	6, 2, 1
availability_365	Attribute	Quantitative	Days	[0, 365] (Days available in next year)	365, 194, 0

E. Quality & Limitations

This dataset exhibits typical characteristics of a real-world dataset, making it a realistic project target.

1. Outliers: This is the most critical issue. The price attribute contains extreme values, such as $0 and $10,000. These values distort simple statistics (like the mean) and will break the scale of most visualizations. Similarly, minimum_nights has irrational outliers (max 1250).

- **Anticipated Cleaning:** Prior to any price visualization, the price attribute must be filtered to a rational range (e.g., $10 - $1,000). minimum_nights will be similarly filtered (e.g., < 31 days).

2. Missing Values: The ~10,052 missing values in last_review and reviews_per_month (approx. 20% of the data) are not a data error. They are an implicit property of the data, corresponding directly to listings where number_of_reviews is 0. This is "inapplicable" data, not "missing" data. These items must be filtered out when analyzing review-based popularity.

3. Temporal Scope: The data is a static snapshot from 2019. It does not reflect market changes from the COVID-19 pandemic or the subsequent rise of remote work. Findings must be interpreted within this 2019 context.

F. Suitability

This dataset strikes an ideal balance of difficulty for a data visualization project. It possesses a manageable complexity, as it is provided in a single CSV file and does not require advanced operations like complex joins or NLP. However, this is not an easy dataset; it presents realistic challenges, such as significant outliers in the price column, which necessitate mandatory and thoughtful data cleaning steps. Its strongest feature is its high-impact visualization potential; the latitude and longitude attributes provide a direct invitation to create geographical dot maps and heatmaps, moving beyond simple bar charts. This rich mix of categorical and quantitative attributes fully supports the creation of a multi-faceted dashboard to tell a cohesive story.