ALAN SAMPEDRO
DATA IMMERSION
TASK 6.1

NYC AIRBNB MARKET ANALYSIS

DATASET DESCRIPTION

SOURCE

The dataset was obtained from *Inside Airbnb*, a project dedicated to enhancing the accessibility of Airbnb data. It is important to note that this website is independent and not affiliated with Airbnb or any of its competitors. The dataset is licensed under a Creative Commons Attribution 4.0 International License.

COLLECTION

The data was collected directly from the official Airbnb website and pre-processed for distribution purposes. For more detailed information, please refer to the Data Assumptions page on the website.

CONTENTS

The dataset comprises various attributes of individual Airbnb property listings in New York City. These attributes include price, room type, availability, and other relevant information.

LIMITATIONS

- Listings in the same building are individually anonymized by Airbnb, which may result in their scattered appearance around the actual address.
- Listings can be deleted from the Airbnb platform, so the data presented here represents a snapshot of available listings at a specific time.
- The neighborhood names associated with each listing were determined by comparing geographic coordinates with the city's defined neighborhoods. Airbnb's neighborhood names were not used due to their inaccuracies.
- Some hosts might not keep their calendar updated or have it highly available even though they live in the entire home/apartment.

ETHICS

- No private or personally identifiable information was utilized. The Airbnb site publicly displays names, photographs, listings, and review details.
- This site asserts the "fair use" of compiled information to produce a non-commercial derivation, enabling public analysis, discussion, and community benefit.

- To ensure privacy, Airbnb anonymized the location information for each listing, guaranteeing that the provided location falls within a range of 0-450ft (150m) from the actual address.
- All copyright and registered trademarks remain the property of their respective owners.

DATA PROFILE

SUMMARY

- 43566 records.
- 18 variables.

VARIABLE	DESCRIPTION	QUALITATIVE/QUANTITATIVE	TIME VARIANT/INVARIANT	NOMINAL/ORDINAL DISCRETE/CONTINUOUS
id	Airbnb's unique identifier for the listing.	Qualitative	Time-invariant	Nominal
name	Listing title.	Qualitative	Time-invariant	Nominal
host_id	Airbnb's unique identifier for the host.	Qualitative	Time-invariant	Nominal
host_name	First name of host.	Qualitative	Time-invariant	Nominal
neighbourhood_group	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	Qualitative	Time-invariant	Nominal
neighbourhood	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	Qualitative	Time-invariant	Nominal
latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	Qualitative	Time-invariant	Nominal
longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	Qualitative	Time-invariant	Nominal
room_type	Property category of listing.	Qualitative	Time-invariant	Nominal
price	Daily price in local currency.	Quantitative	Time-variant	Continuous
minimum_nights	Minimum number of night stay for the listing.	Quantitative	Time-variant	Discrete
number_of_reviews	The number of reviews the listing has.	Quantitative	Time-variant	Continuous
last_review	The date of the last/newest review.	Qualitative	Time-variant	Discrete
calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.	Quantitative	Time-variant	Discrete
availability_365	The availability of the listing 365 days in the future as determined by the calendar. Note a listing may be available because it has been booked by a guest or blocked by the host.	Quantitative	Time-variant	Discrete
number_of_reviews_ltm	The number of reviews the listing has (in the last 12 months)	Quantitative	Time-variant	Discrete
license	Unknown.	NA	NA	NA

DATA QUALITY

INTEGRITY ISSUES

- The majority of quantitative variables, except for 'availability_365', exhibit a significant proportion of mild outliers (1.5 * IQR) as well as extreme outliers (3 * IQR).
- The ‘price’ column has a minimum value of zero. This suggests that one or more listings are either “free of charge” or have unavailable pricing information.
- High proportion of missing values in ‘last_review’ and ‘reviews_per_month’ columns.

MODIFICATIONS

- The columns ‘license’, ‘host_name’, and ‘name’ were removed from the analysis as they were deemed irrelevant for the intended purpose.

- To enhance conciseness and readability, certain columns in the DataFrame were renamed.
    - 'neighbourhood_group' → 'nbhd_group'
    - 'neighbourhood' → 'nbhd'
    - 'minimum_nights' → 'min_nights'
    - 'number_of_reviews' → 'reviews_count'
    - 'reviews_per_month' → 'monthly_reviews'
    - 'calculated_host_listings_count' → 'host_list_count'
    - 'number_of_reviews_ltm' → 'reviews_count_ltm'
- Records where ‘price’ is greater than 50,000 were removed from the dataset since they are not plausible and do not align with the expected range of prices.
- There are 16 records with a price value of 0. Among them, 14 follow a consistent pattern, representing hotel rooms located in Manhattan. The dataset was filtered using the identified pattern, and the median value of ‘price’ in this subset was calculated and used as the replacement value. The other 2 records were removed from the dataset since they didn’t exhibit any noticeable patterns.
- Records where ‘min_nights’ is greater than 800 were removed from the dataset since it is highly unreasonable to expect hosts to require customers to book a listing for 2-3 years.
- For the 'last_review' column, we will substitute missing values with "Unavailable".
- For the 'monthly_reviews' column, we will replace missing values with "0".

QUALITATIVE VARIABLES

VARIABLE	MODE	COUNT
id	2595	43557
host_id	107434423	43557
nbhd_group	Manhattan	43557
nbhd	Bedford-Stuyvesant	43557
latitude	40.76153	43557
longitude	-73.99878	43557
room_type	Entire home/apt	43557
last_review	nan	43557

QUANTITATIVE VARIABLES

VARIABLE	MIN	MEDIAN	MAX	IQR	MEAN	STD
price	10	136	20500	149	218.86	442.46
min_nights	1	15	500	28	18.57	25.37
reviews_count	0	5	2024	23	26.37	57.74
monthly_reviews	0.01	0.55	63.95	1.67	1.22	1.78
host_list_count	1	1	569	4	39.94	99.12
availability_365	0	89	365	278	137	137.34
reviews_count_ltm	0	1	1128	8	7.88	18.25

RESEARCH QUESTIONS

GEOSPATIAL ANALYSIS

1. How does the distribution of Airbnb listings vary across different neighborhoods in New York City?
2. Are there any spatial patterns or clusters of high-priced listings in specific areas of the city?
3. How does the spatial distribution of Airbnb listings in New York City correlate with the proximity to public transportation hubs, such as subway stations or bus stops?
4. Is there a correlation between the proximity to popular tourist attractions and the price of Airbnb listings?
5. Are there any spatial variations in the average number of reviews for listings across different neighborhoods in the city?

REGRESSION ANALYSIS

1. What factors influence the price of Airbnb listings in New York City? (e.g., room type, neighborhood, number of reviews, availability)
2. Can we predict the price of a listing based on its attributes, such as room type, neighborhood, and number of reviews?
3. How does the minimum number of nights required for a stay affect the price of listings?
4. Can we identify any interaction effects between room type and neighborhood on the price of Airbnb listings in New York City?
5. How does the availability of listings throughout the year impact their price, taking into account the seasonality factor?

CLUSTER ANALYSIS

1. Can we identify distinct clusters of listings based on their attributes (e.g., price, room type, availability)?
2. Are there specific neighborhoods that have similar characteristics in terms of Airbnb listings?
3. Do clusters of listings exhibit different patterns in terms of price trends over time?
4. Are there any distinct clusters of listings that cater specifically to different types of travelers, such as budget-conscious, luxury-seeking, or family-oriented?
5. Do the clusters of listings exhibit different patterns in terms of the minimum number of nights required for a stay?

TIME SERIES ANALYSIS

1. How has the availability of Airbnb listings in New York City changed over time?
2. Are there seasonal patterns in the number of reviews received by listings?
3. Is there a relationship between the availability of listings and the number of reviews in the last 12 months?
4. Is there a long-term trend in the average price of Airbnb listings in New York City, and how does it compare to the overall inflation rate in the region?
5. Are there any temporal patterns in the number of bookings made through Airbnb, and do they coincide with major events or tourist seasons in the city?

APPENDIX

DETAILED DATA PROFILE

SCRIPT