Machine Learning Group Project: Analytical Report

Word Count: 996

“What are the primary factors that affect the listing price of properties on Airbnb, and how can these insights help hosts achieve the most value from their listings?”

This report utilises the New York City Airbnb Open Data (Dgomonov, 2019) to explore factors that influence listing price. It examines the impact of listing names and descriptions, the room type, minimum night requirements, and location. The report concludes that room type, and location are the most significant factors affecting price, but other elements influence search visibility and utilisation.   This will help hosts to optimise pricing strategies for their listings.

1. Data Overview

The dataset includes 48,895 listings with 16 features, including price, neighbourhood and geographical information, room type, and availability. Missing data is limited, primarily occurring in 'last_review' and 'reviews_per_month', likely aligning to un-booked properties (Figure 1). '

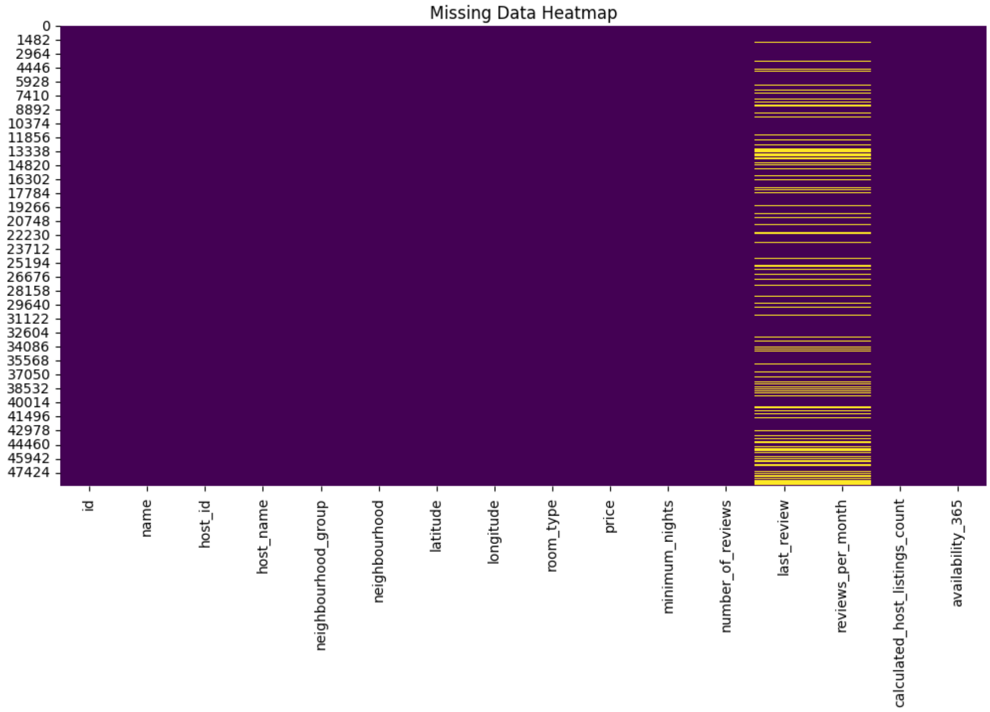


Figure 1: Heatmap showing missing data

Some key features are highly skewed. Price is extremely skewed, with a long tail of extreme values. Minimum nights are highly skewed, suggesting potential data errors. Variables related to reviews are highly skewed, reflecting a smaller number of highly active listings. After application of a simple IQR outlier removal, the key numerical variables still display a high skew (Figure 2).

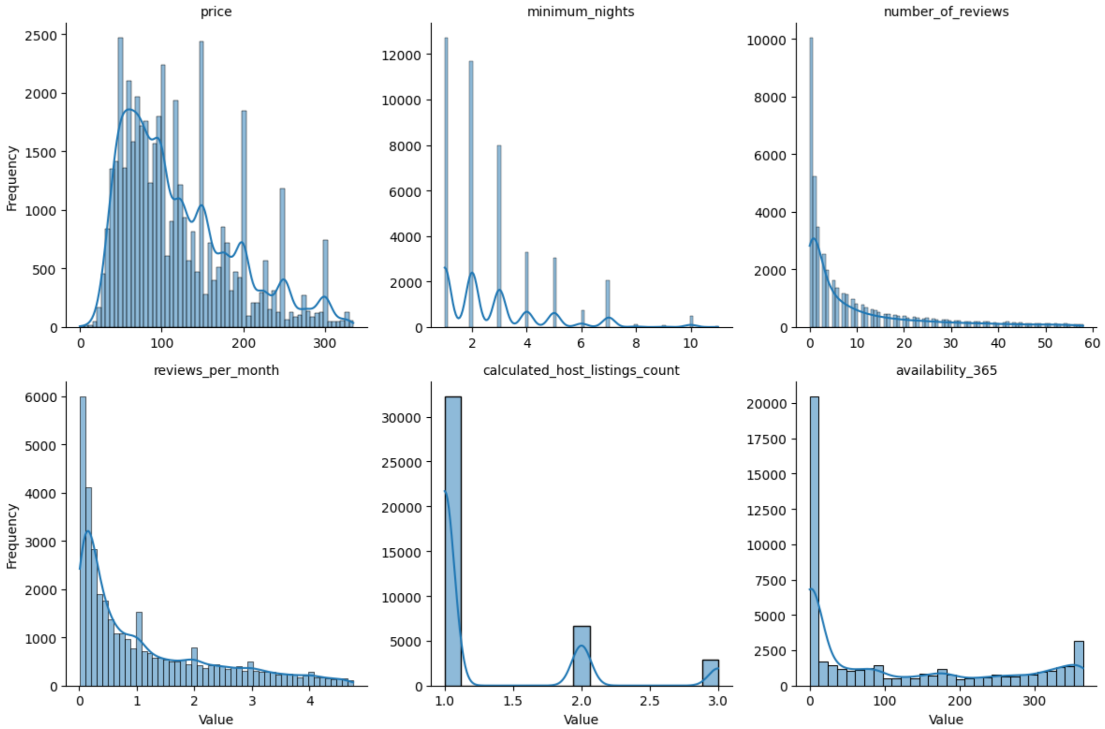


Figure 2: Histogram excluding IQR outliers

Most hosts manage one property, with a few hosts managing a large number of properties, signalling professionalisation (Figure 3). Correlation analysis shows weak relationships between price and other numerical variables, with only ‘number_of_reviews’ and ‘reviews_per_month’ showing moderate positive correlation (Figure 4).

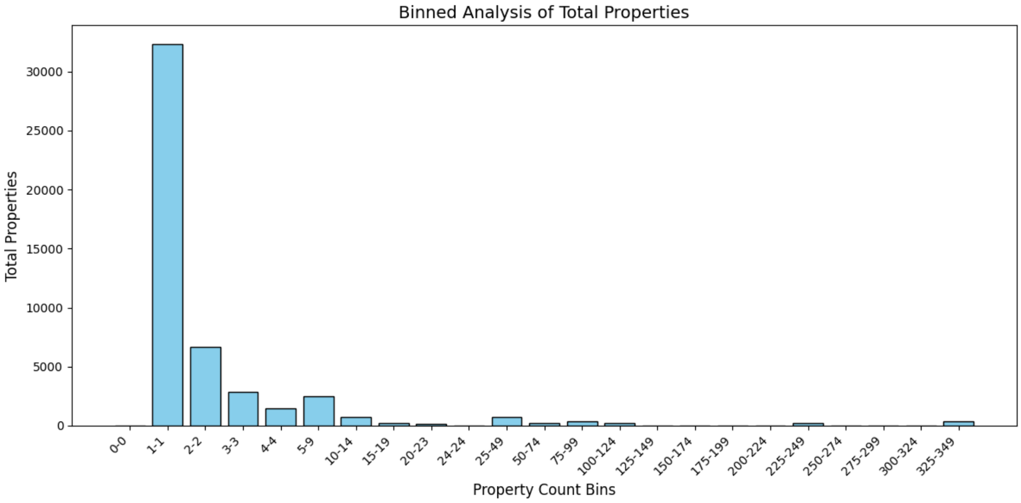


Figure 3: Binned properties per host

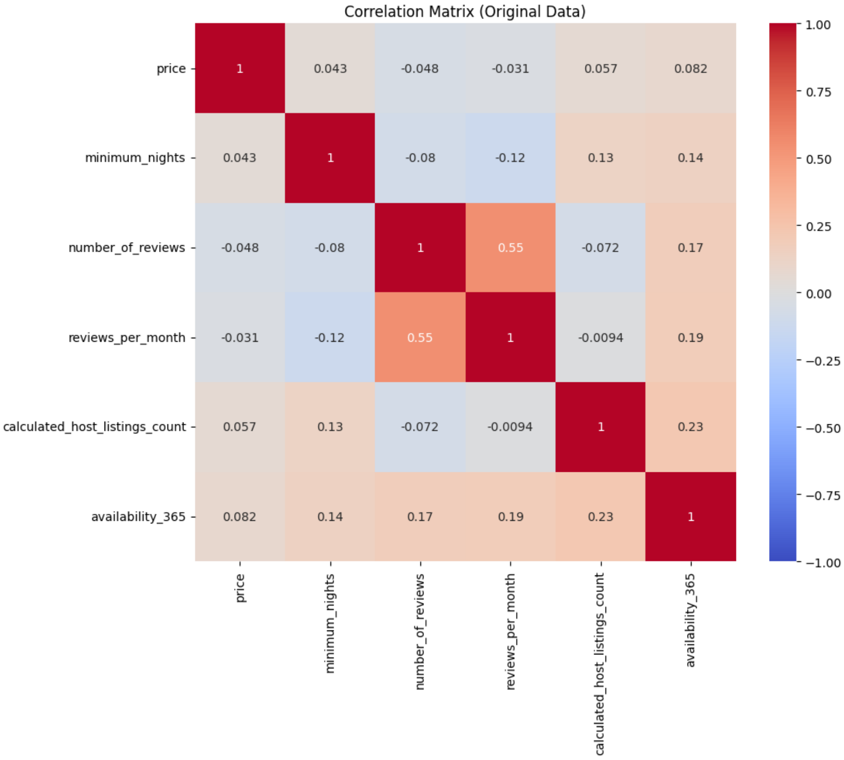


Figure 4: Correlation Matrix

Categorical analysis indicates most listings are in Manhattan and Brooklyn, with “Entire home/apartment” being the most popular type of listing (Figure 5).

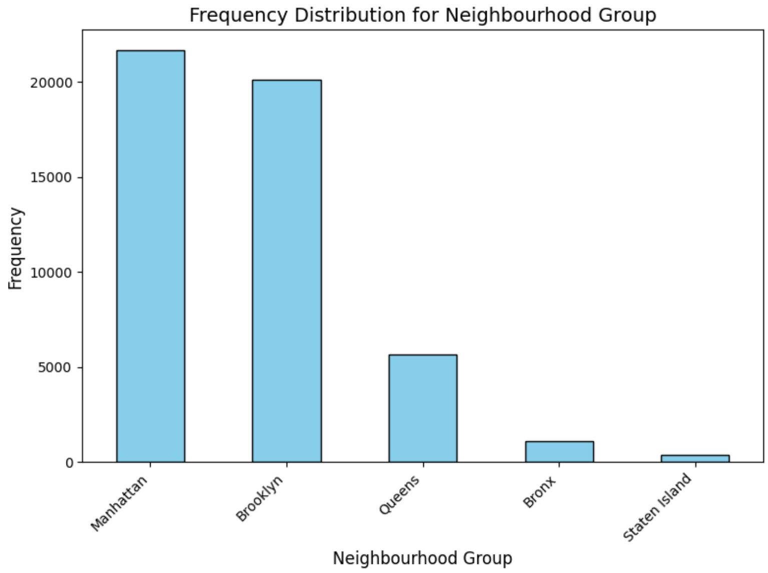


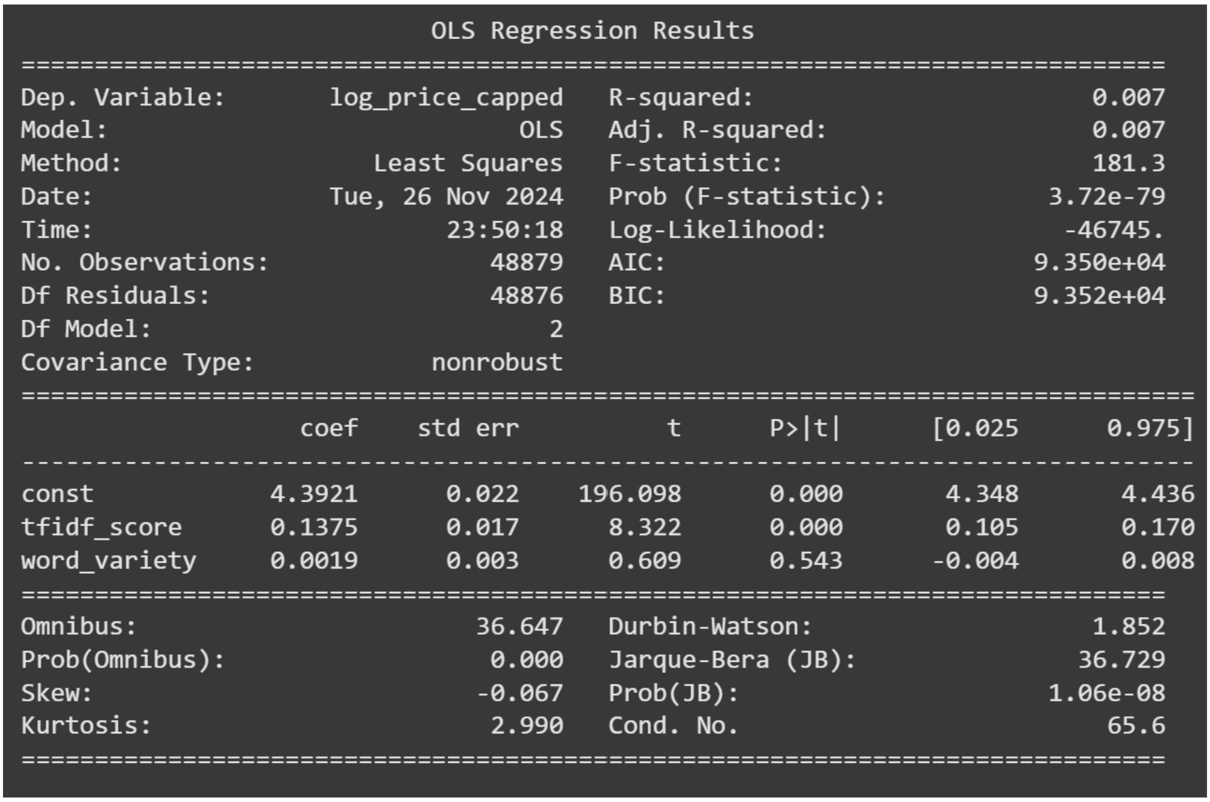
Figure 5: Count by Neighbourhood Group

2. Influence of Listing Names on Airbnb Pricing

This section investigates whether Airbnb listing names influence pricing by exploring textual features including word count, uniqueness, and keywords, using TF-IDF vectorisation (Aizawa, 2003) and regression modelling (Alharbi, 2023). While unique, descriptive names showed slight tendencies toward higher prices, most results were statistically insignificant.

Regression models demonstrated limited evidence of a meaningful relationship between listing names and pricing. Minor patterns emerged, suggesting that optimising listing titles with relevant keywords could enhance visibility but has minimal impact on pricing. Instead, other factors such as location or amenities likely play a more significant role (Yang et al., 2021).

These findings suggest that attractive titles may improve a listing's appeal but is unlikely to have a substantial impact on pricing.



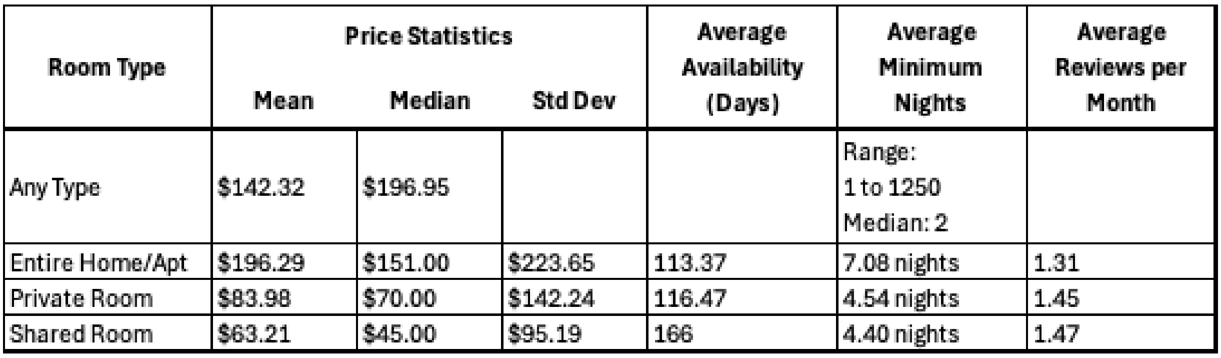
```
                            OLS Regression Results
==============================================================================
Dep. Variable:       log_price_capped   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     181.3
Date:                Tue, 26 Nov 2024   Prob (F-statistic):           3.72e-79
Time:                        23:50:18   Log-Likelihood:                -46745.
No. Observations:               48879   AIC:                         9.350e+04
Df Residuals:                   48876   BIC:                         9.352e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            4.3921      0.022    196.098      0.000       4.348       4.436
tfidf_score      0.1375      0.017      8.322      0.000       0.105       0.170
word_variety     0.0019      0.003      0.609      0.543      -0.004       0.008
==============================================================================
Omnibus:                       36.647   Durbin-Watson:                   1.852
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.729
Skew:                          -0.067   Prob(JB):                     1.06e-08
Kurtosis:                       2.990   Cond. No.                         65.6
==============================================================================
```

Figure 6: Regression Results

3. Influence of Minimum Night Stays on Airbnb Pricing

This section explores the relationship between pricing and minimum night requirements. The price shows significant variability with an average price of $142.32 and a standard deviation of $196.95. The median minimum stay is 2 nights, with a maximum of 1,250 nights. These variables possess a very low correlation coefficient of 0.026 which indicates that the number of minimum nights have a minimal impact on pricing (Di Persio & Lalmi, 2024). The analysis indicates that room type has a more significant impact on price - entire Home/Apt listings have the highest prices, followed by Private and Shared Rooms (Hong & Yoo, 2020) – this is addressed more in Section 4.

Room type and location are critical in determining competitive prices. While minimum night stays can affect occupancy rates, they have a minimal effect on the variability of prices. Figure 7 summarises key findings regarding price, availability, minimum nights, and guest engagement across room types.



Room Type	Price Statistics			Average Availability (Days)	Average Minimum Nights	Average Reviews per Month
	Mean	Median	Std Dev			
Any Type	$142.32	$196.95			Range: 1 to 1250 Median: 2	
Entire Home/Apt	$196.29	$151.00	$223.65	113.37	7.08 nights	1.31
Private Room	$83.98	$70.00	$142.24	116.47	4.54 nights	1.45
Shared Room	$63.21	$45.00	$95.19	166	4.40 nights	1.47

Figure 7: Price, availability, minimum nights and review by room type

4. Room Type and Price

Room type significantly impacts listing prices. Entire homes or apartments have the highest prices (Figure 8), with fewer extreme outliers (Figure 9). Other room types, such as shared and private rooms, tend to have lower prices (Figure 10). This is likely because guests prefer the privacy and exclusive use of facilities offered by entire homes, whereas sharing spaces like kitchens and bathrooms may deter renters.

Shared rooms typically have the lowest price and are generally less attractive due to privacy and safety concerns. Research suggests that there is a potential link between shared rooms and higher crimes rates (Lanfear & Kirk, 2024).

Private rooms are in the mid-price range, offering a balance between affordability and privacy. When pricing listings, hosts need to take the room type into account, ensuring that shared spaces are competitively priced to attract guests, without competing with more private listings.

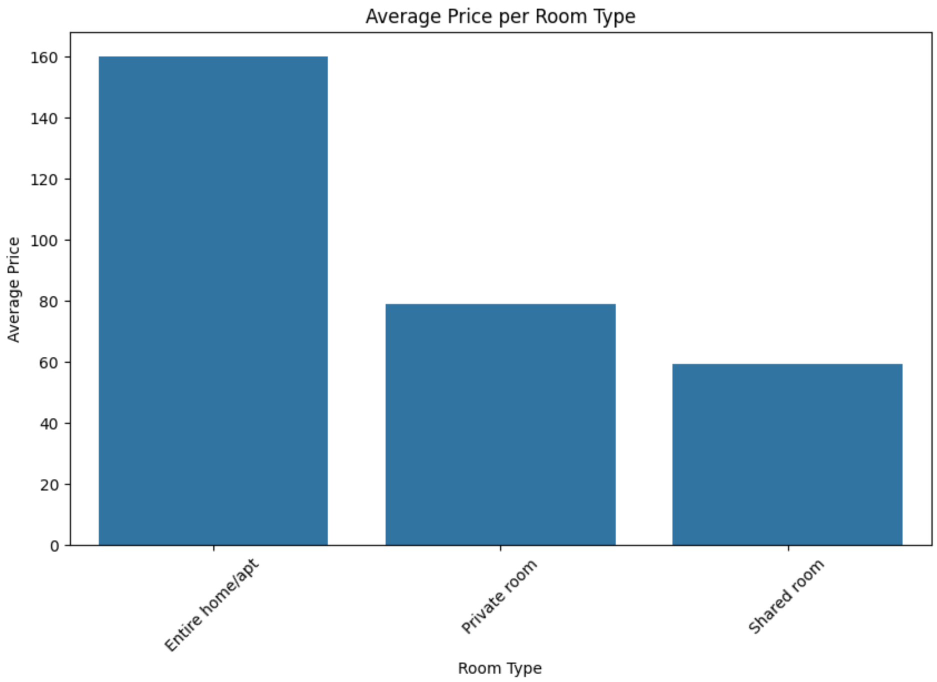


Figure 8: Average price per room type.

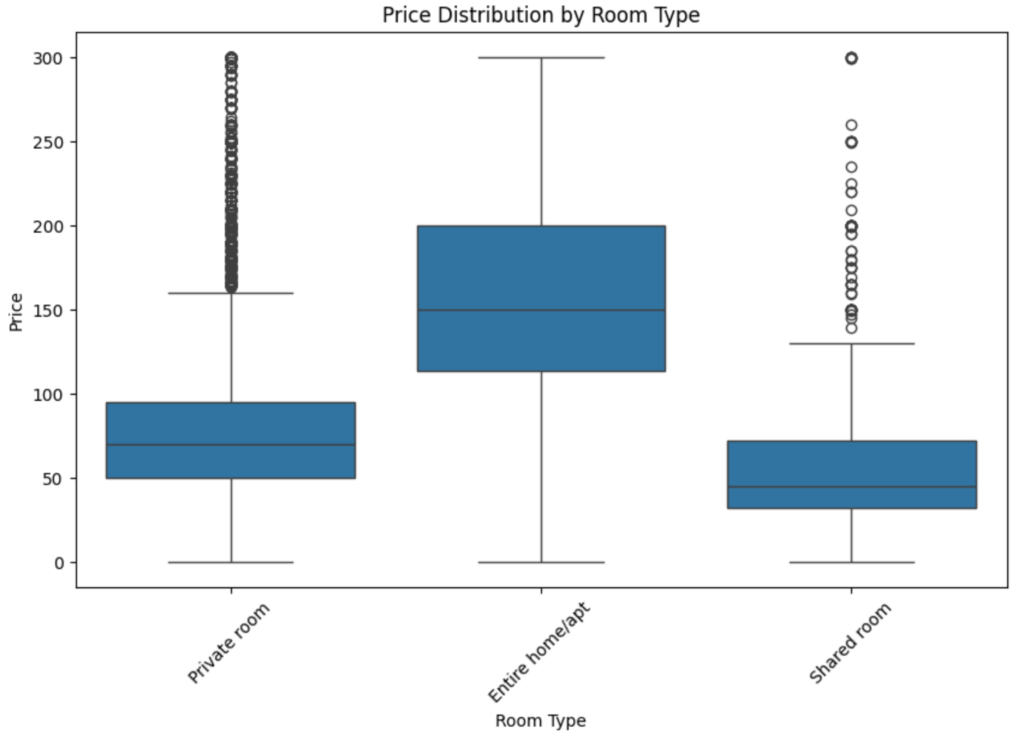


Figure 9: Price distribution boxplot.

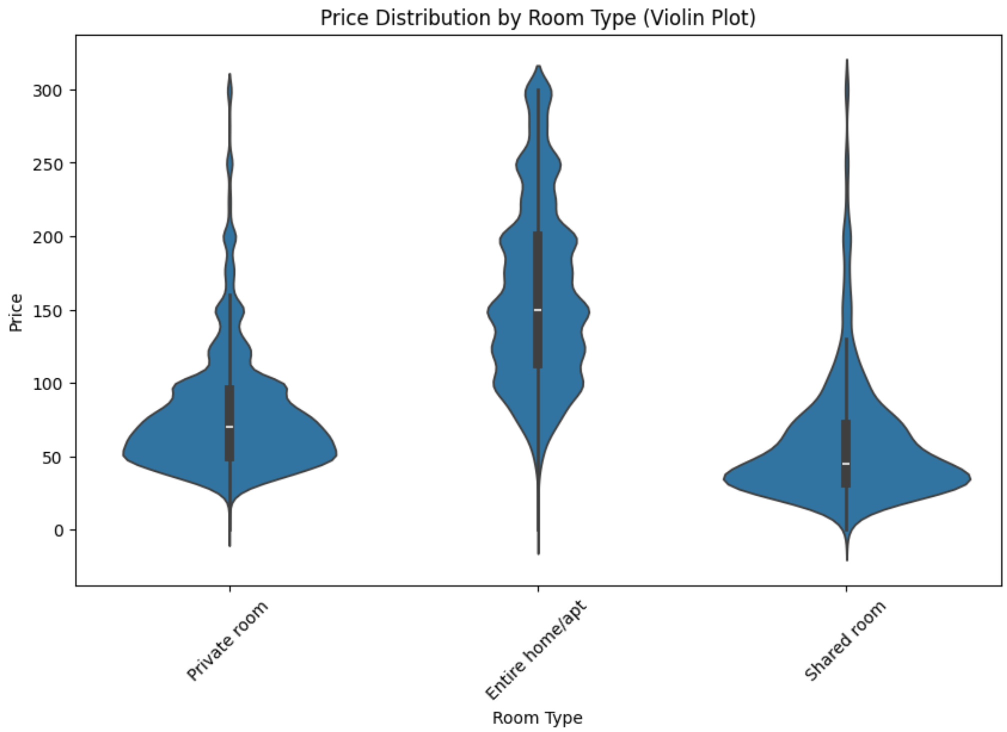


Figure 10: Price distribution violin graph.

5. Geographical Analysis

This section analyses the relationship between geographic location and Airbnb listing prices in New York City, following similar approaches as Gyódi & Nawaro (2021) and Toader et al. (2022). Neighbourhood group, longitude, and latitude have no missing values. Outliers are retained to capture high-priced special locations, and only listings with positive reviews are included to ensure bookings. Using a 5% significance level, we conduct two linear regressions. Results show that listings further north and west are associated with higher prices. Manhattan, Brooklyn, and Queens exhibit significantly higher prices, while Staten Island shows no significant effect.

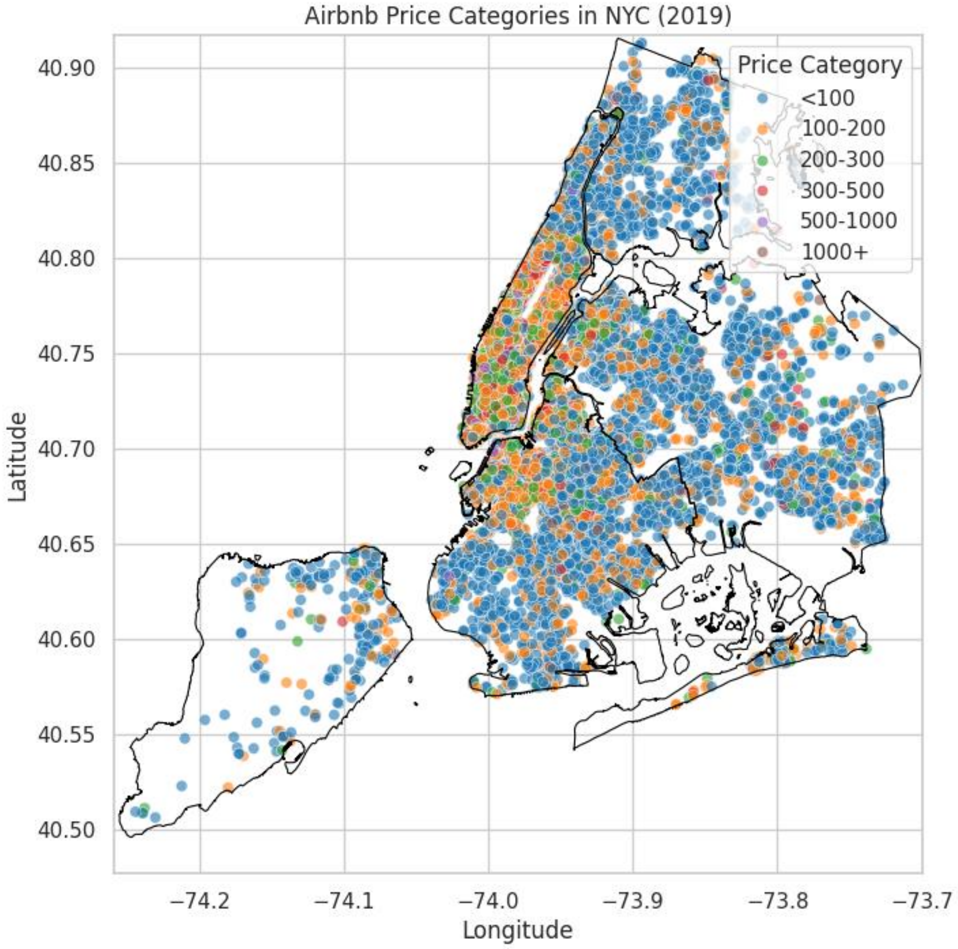


Figure 11: Prices by geographic region.

Figure 11 illustrates NYC price distributions, with Manhattan and Brooklyn showing above-average prices, moderate increases in Queens, and lower prices in Staten Island. Staten Island also has fewer listings, reflecting lower tourist demand. This suggests a potential low-price strategy for Staten Island to attract budget-conscious tourists unable to afford pricier areas.

6. Conclusion

Room type and location are the most influential factors affecting Airbnb listing prices in this dataset. Entire home and apartments typically achieve the highest price, whilst shared rooms achieve the lowest. Hosts may optimise their pricing by taking into account the level of privacy and proximity to popular destinations, such as Manhattan or Brooklyn. Whilst other factors such as listing names and minimum night requirements have minimal direct impact on price, they may affect visibility in search results and ultimately affect property utilisation.

Reference List:

Aizawa, A. (2003) An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39(1): 45–65. DOI: https://doi.org/10.1016/S0306-4573(02)00021-3.

Alharbi, Z.H. (2023) A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis. *Sustainability* 15(17): 13159. DOI: https://doi.org/10.3390/su151713159.

Dgomonov (2019) Airbnb listings and metrics in NYC, NY, USA. Available from: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data [Accessed 18/11/2024].

Di Persio, L. & Lalmi, E. (2024) Maximizing Profitability and Occupancy: An Optimal Pricing Strategy for Airbnb Hosts Using Regression Techniques and Natural Language Processing. *Journal of Risk and Financial Management* 17(9): 414. DOI: https://doi.org/10.3390/jrfm17090414.

Gyódi, K. & Nawaro, Ł. (2021) Determinants of Airbnb prices in European cities: A spatial econometrics approach. *Tourism Management* 86: 104319. DOI: https://doi.org/10.1016/j.tourman.2021.104319.

Hong, I. & Yoo, C. (2020) Analyzing Spatial Variance of Airbnb Pricing Determinants Using Multiscale GWR Approach. *Sustainability* 12(11): 4710. DOI: https://doi.org/10.3390/su12114710.

Lanfear, C.C. & Kirk, D.S. (2024) The promise and perils of the sharing economy: The impact of Airbnb lettings on crime. *Criminology* 1745-9125: 12383. DOI: https://doi.org/10.1111/1745-9125.12383.

Toader, V., Negrușa, A.L., Bode, O.R. & Rus, R.V. (2022) Analysis of price determinants in the case of Airbnb listings. *Economic Research-Ekonomska Istraživanja* 35(1): 2493–2509. DOI: https://doi.org/10.1080/1331677X.2021.1962380.

Yang, Y. (Yvonne), Kim, S.I., Kim, J. & Koh, Y. (2021) How Airbnb Titles Influence Guests' Decision Making: Linguistic and Spatial Analysis Approach. *International Journal of Hospitality & Tourism Administration* 25(2): 382–405. DOI: https://doi.org/10.1080/15256480.2022.2114973.

Appendix A – Code

File 1: Notebook PDF Attached (Final_Version_Group_Project_Group_1.pdf)

File 2: Notebook File (Final_Version_Group_Project_Group_1.ipynb)