Development Team Project:

Project Report

Nikos Archontas, Fabian Narel, Matilda Nilsson, João Torres and Maryum Rasool

# Introduction

The rise of online platforms like Airbnb has transformed the hotel industry by giving customers more control over pricing, reviews, and location research (Sharma & Gupta, 2021). Airbnb has become a top accommodation option worldwide, especially in New York City (Jiao & Bai, 2019). This report analyzes Airbnb listings in NYC, focusing on pricing, availability, location, reviews, and property types. Understanding these factors can help Airbnb optimize pricing and improve guest engagement.

## Proposed Business Analytic Question:

**Which NYC neighborhood had the highest average Airbnb listing price in 2019, and what factors contributed to its high pricing? Additionally, is there a statistical relationship between Airbnb prices and the number of reviews a listing receives?**

# Data Analysis

## Data Preprocessing

The dataset initially contained 48,895 rows and 16 columns. We removed records with zero reviews, zero prices, and extreme outliers. After preprocessing, the dataset was reduced to 38,833 entries.

## Exploratory Data Analysis (EDA)

The number of Airbnb listings by neighborhood groups showed that Manhattan had the highest number of listings, closely followed by Brooklyn. Staten Island and the Bronx had considerably fewer listings (Figure 1).

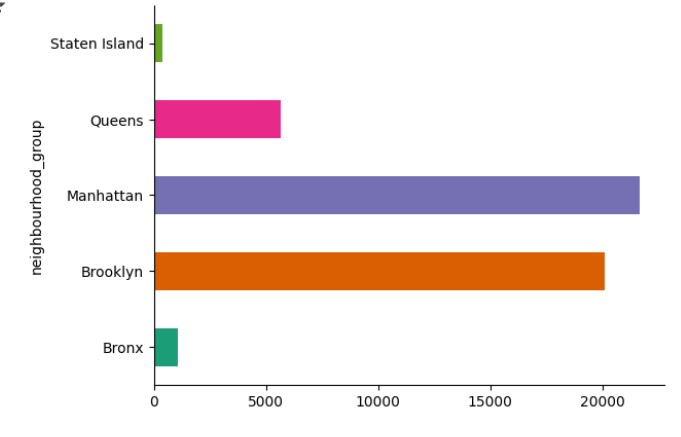


Figure 1: Visualisation of the number of Airbnb listings per neighborhood group

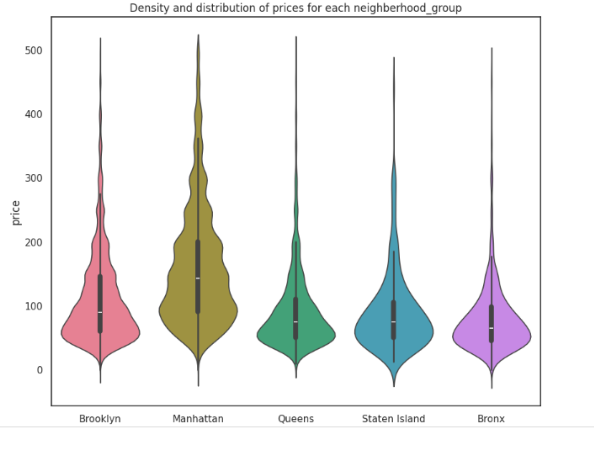


Figure 2: Violin Plot depicting the density and distribution of prices for airbnb listings for each New York neighborhood group in 2019

Manhattan had the highest median price at $150 per night, while the Bronx had the lowest at $65 per night. Summary statistics revealed the overall median price was $142.33, with a standard deviation of $196.99, indicating significant price differences across listings (Figure 2, Figure 3).

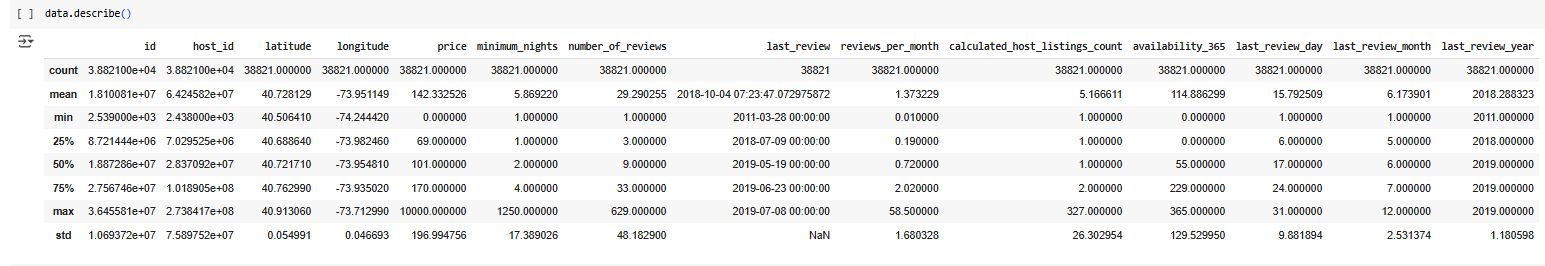


Figure 3: Descriptive statistics on the dataset, achieved through the data.describe command.

In Manhattan and Brooklyn, room type had a significant impact on pricing. Entire homes and apartments had a moderate positive correlation with price (r=0.476), while private rooms had a negative correlation (r=-0.450), supporting the preference for privacy in these neighborhoods (Perez-Sanchez et al., 2019). Longitude also showed a negative correlation with price, reflecting Manhattan's property value decline from expensive areas like SoHo to more affordable locations like East Harlem (Coles et al., 2017) (Figure 4). The number of reviews had a weak correlation with price.

| Count | 38821 |
| --- | --- |
| Mean | $142.33 |
| Standard Deviation | 196.99 |
| 25th Percentile | $69 |
| 50th Percentile | $101 |
| 75th Percentile | $170 |
| Max | $10000 |

Table 1: Descriptive statistics on Airbnb listing prices.

Box plots confirmed that Entire Home/Apartment listings had more than double the median price of Private and Shared rooms. The number of reviews had no significant impact on price. Minimum nights had a weak correlation with price, showing lower prices for properties that allowed single-night stays (Figure 5).

## 

Figure 4: Correlation Matrix for Manhattan and Brooklyn data.

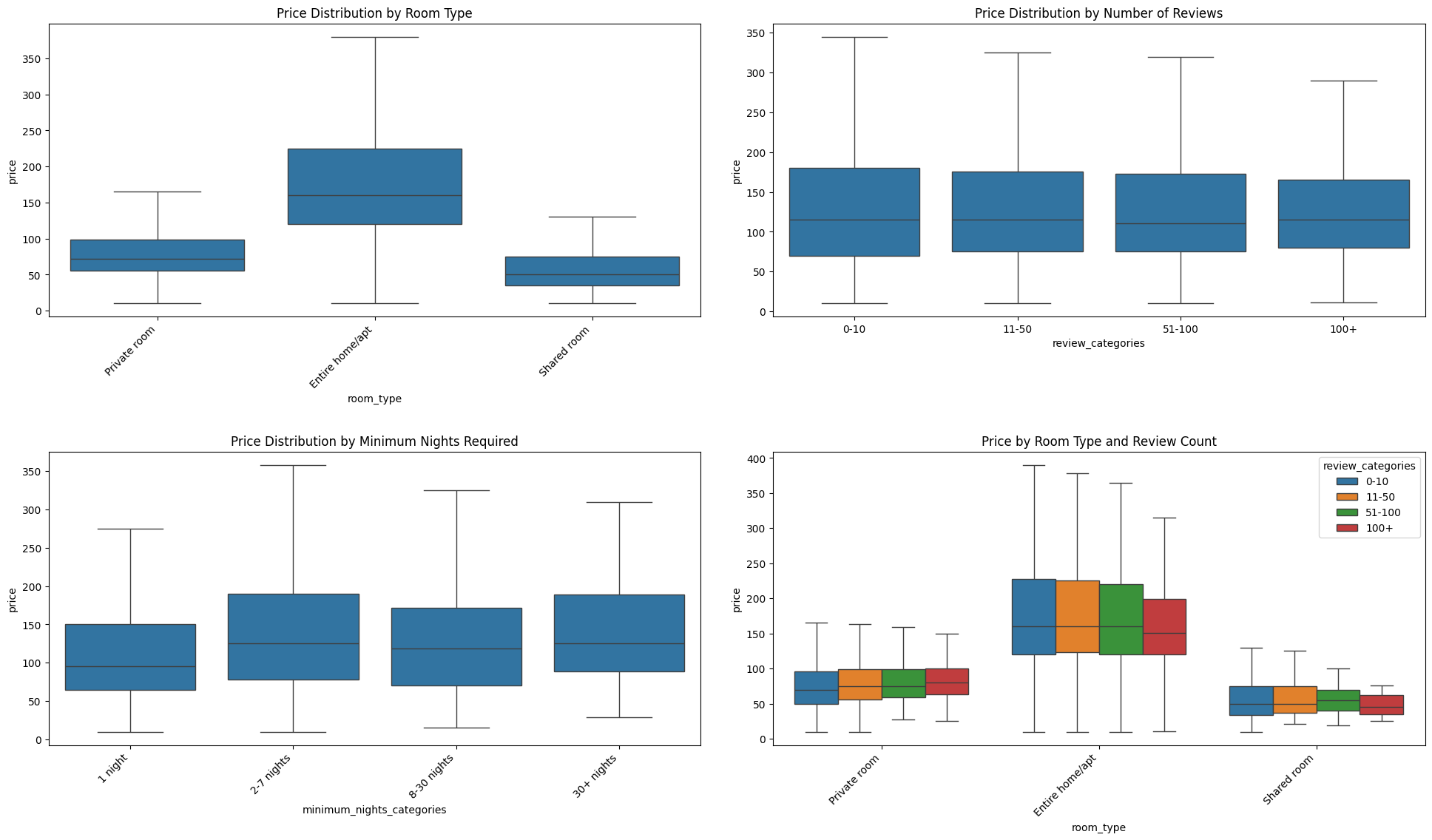


Figure 5: Box plots comparing Price Distribution by Room Type, Number of Reviews, Minimum Nights Required and Room Type combined with Number of Reviews

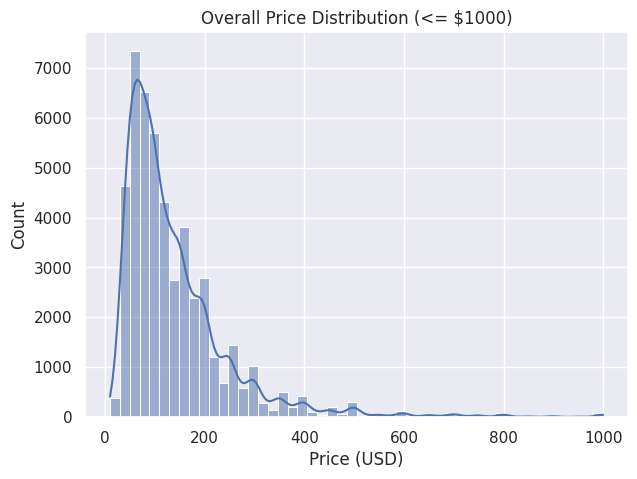


Figure 6: Histogram of nightly prices for NYC Airbnb listings (prices capped at $1000).

The distribution of NYC Airbnb prices was highly skewed, with most listings priced under $1,000, though a few luxury properties exceeded $10,000. The median price was inflated by high-end properties. The market reflected a mix of budget and premium options, with 75% of listings allowing stays of five nights or fewer (Figure 6).

## Correlation Analysis

Expensive listings tend to have fewer reviews, suggesting lower booking frequency. In the 2019 NYC data, the weak negative correlation (r ≈ -0.058, p ≈ 2.6×10⁻³⁷) indicates that budget-friendly listings attract more guests and reviews, while high-priced ones cater to a niche market (Martinez et al., 2017). This aligns with research on tourism demand dynamics, where lower prices can boost occupancy, but higher prices reduce the guest pool (Toader et al., 2022).



Figure 7: Visualisation of scatter plot of price vs. number of reviews

The scatter plot (Figure 7) supports this, showing a dense cluster of low-priced, well-reviewed listings and scattered high-priced listings with few reviews. Many expensive listings had 0–5 reviews, reflecting rare bookings. This reinforces the price-occupancy trade-off, where lower prices attract more bookings, while higher prices target fewer guests. Price showed minimal correlation with other features, aside from the slight negative link with reviews.

## Machine Learning Analysis

To further explore patterns at the neighborhood level, we performed K-Means clustering on neighborhoods based on average price and average number of reviews per listing as well as linear regression to examine the relationship between prices and various predictors, including neighborhood group, number of reviews, and host listing count, with the idea to determine whether these factors impact price (Schroeder et al., 2017).

## Linear Regression

The linear regression model tests whether neighborhood group, reviews, and host listing count influence price. While statistically significant (F = 205.263, p < 0.001), it explains only 9.3% of price variance (R² = 0.093), indicating many unaccounted factors (Chatterjee & Hadi, 2006). The high RMSE (112.457) reflects substantial predictive error. We reject the null hypothesis, confirming that at least one predictor impacts price (Field, 2024).

| **Predictor** | **Coefficient** | **P value** | **Interpretation** |
| --- | --- | --- | --- |
| Number of reviews | -0.116 | <0.001 | Every additional review results in the price dropping by $0.12 on average.  P value suggests significant impact. |
| Reviews per month | -0.073 | 0.917 | Each additional review decreases the price by $0.07.    Not statistically significant. |
| Neighbourhood group (Brooklyn) | 41.529 | <0.001 | $41.53 more expensive than the baseline. |
| Neighbourhood group (Manhattan) | 99.871 | <0.001 | Listings are $99.87 more expensive that the baseline reference |
| Neighbourhood group (Queens) | 16.440 | 16.440 | Queens listing are 16.44$ more expensive    But have a weaker impact |
| Neighbourhood group (Staten Island) | 10.920 | 0.392 | Does not significantly impact price  (p > 0.05) |

Table 2: Regression table highlighting the effects individuals have on price

The analysis confirms significant price differences among neighborhood groups, with Manhattan and Brooklyn commanding higher prices than the Bronx. Listings with more reviews tend to have lower prices, indicating budget-friendly properties attract more guests (Perez-Sanchez et al., 2018; Coles et al., 2017), while "reviews per month" had no significant impact.

Linear regression identifies neighborhood groups as a key price predictor, with Manhattan and Brooklyn as premium markets. However, its low explanatory power (R² = 9.3%) suggests future models should include factors like room type, amenities, seasonality, and demand for better accuracy (Chatterjee & Hadi, 2006; Schroeder et al., 2017).

## K-Means Clustering

Performing K-Means clustering on NYC neighborhoods based on average price and number of reviews to segment the market into three distinct groups:

* **Cluster 1 – "Luxury Markets"**: High price, low reviews. Representing about 1% of neighborhoods, these areas have ultra-expensive listings with minimal bookings, catering to an exclusive clientele (e.g., Fort Wadsworth, Staten Island).
* **Cluster 2 – "Mid-Range Markets"**: Moderate price, moderate reviews. The largest group (79%) includes areas like Manhattan and Brooklyn, offering moderately priced listings with steady guest traffic (e.g., popular parts of Brooklyn and Lower Manhattan).
* **Cluster 3 – "Budget-Friendly/High-Occupancy"**: Lower price, high reviews. Comprising 20% of neighborhoods, these areas attract budget travelers and high occupancy, with listings priced under $100 and many reviews (e.g., East Elmhurst, Queens).

These clusters represent different market segments, from high-end to budget-friendly options, and illustrate varying pricing strategies and guest demands across the city.

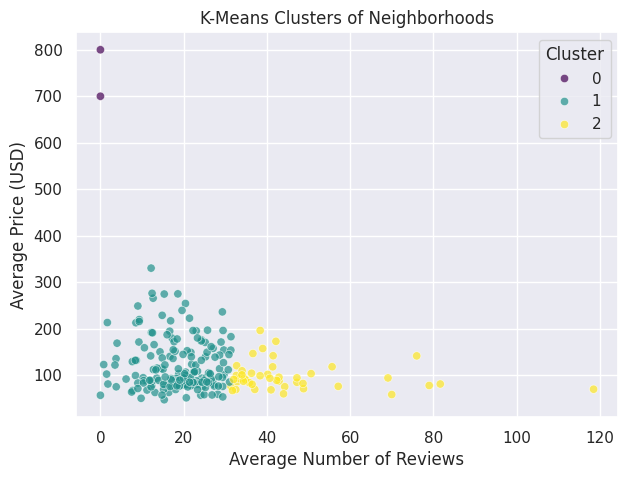


Figure 8: Visualisation of K-means clustering of neighborhoods

The cluster visualization (Figure 8) shows three distinct groups: budget (low-price, high-review), mid-range (moderate price and reviews), and luxury (high-price, low-reviews). This segmentation offers strategic insights for Airbnb hosts and stakeholders. Hosts in mid-range areas should differentiate their listings to stand out, while budget hosts should focus on high occupancy and positive reviews. Luxury hosts should emphasize exclusive amenities and experiences. Airbnb can tailor marketing and pricing strategies based on these segments, targeting budget travelers with promotions and luxury segments through niche channels.

# Findings and Recommendations

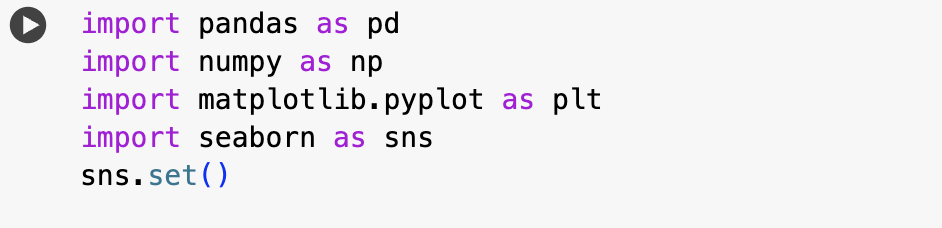
Airbnb has significant growth potential in New York City. The dataset reveals high demand near major landmarks, which could be leveraged to increase profits by targeting accommodations in these areas. Additionally, offering a mix of premium housing in high-demand sectors and smaller rooms in less popular areas could help Airbnb remain competitive through dynamic pricing. By tracking location-specific trends and local events, Airbnb can better plan promotional offers tailored to neighborhoods and hot spots.

# Conclusion

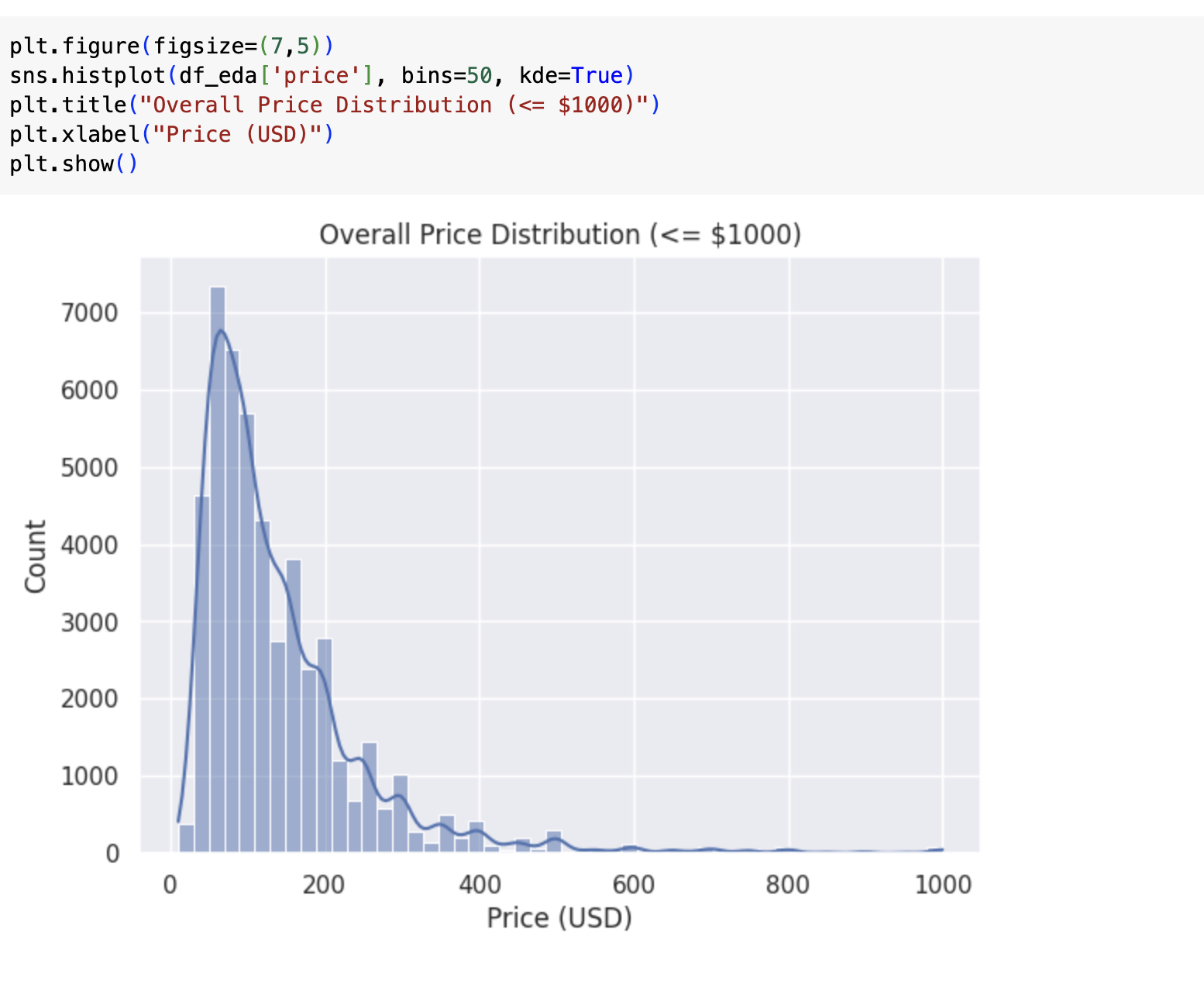
This report identifies key business opportunities in New York City for Airbnb, including targeted marketing strategies and revenue optimization. By focusing on specific areas, Airbnb can enhance customer satisfaction and increase revenue. The report also demonstrates the use of data analysis and machine learning for strategic decision-making and business growth.

# Appendix

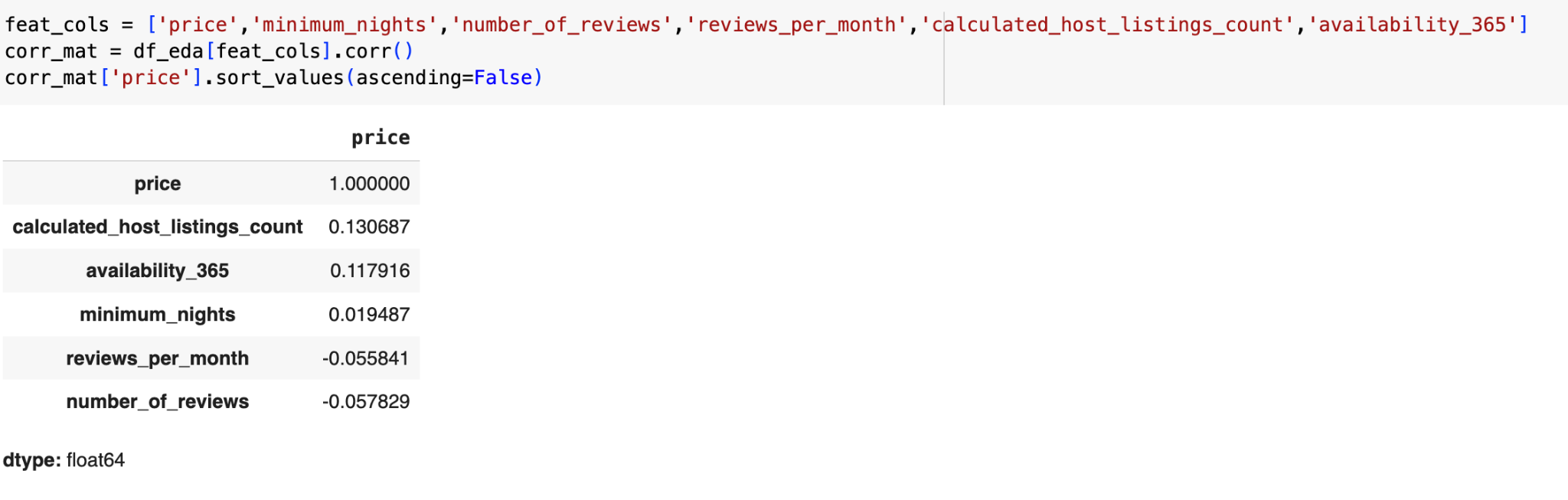
Correlation and Machine Learning Analysis

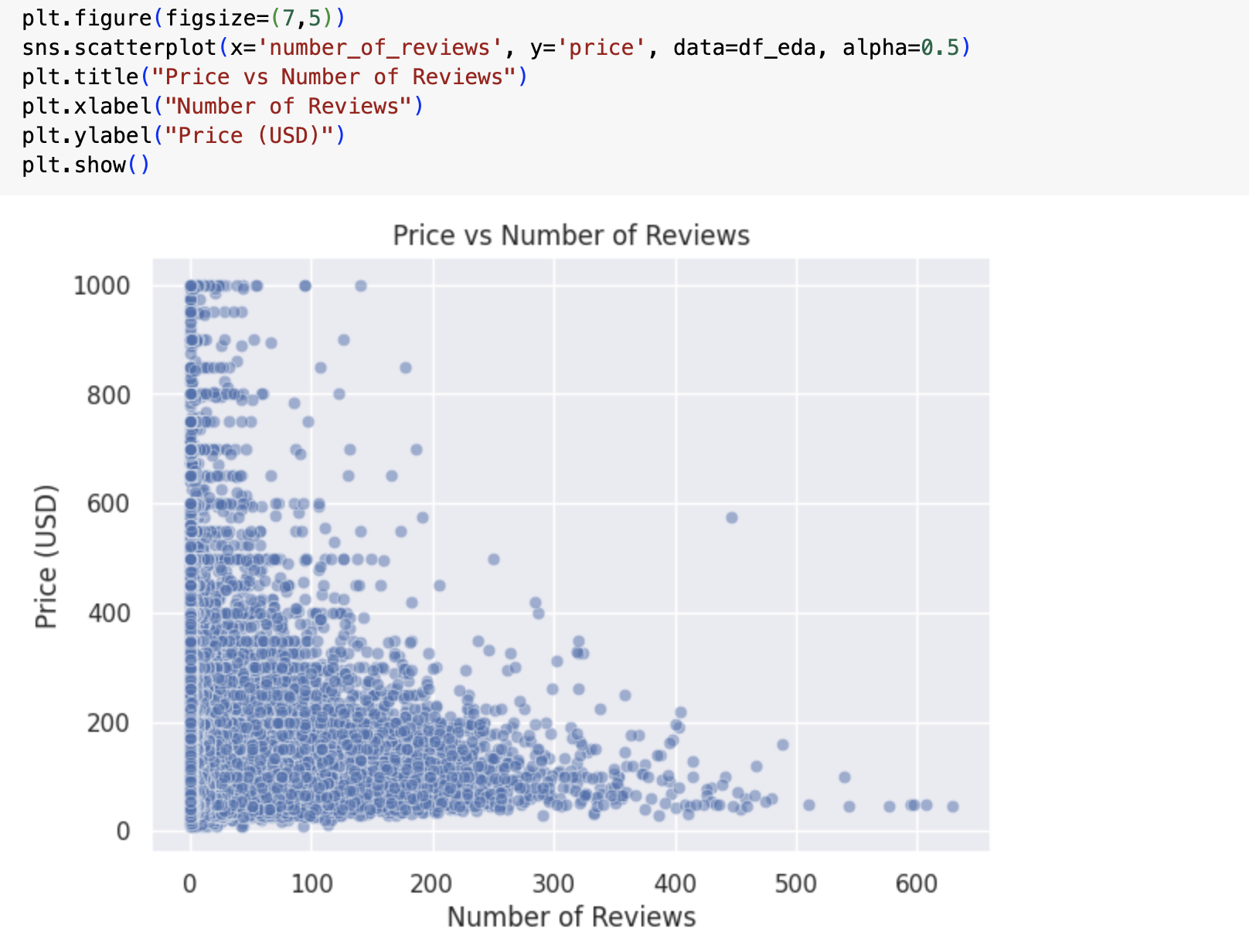


A histogram for price distribution (<= 1000)

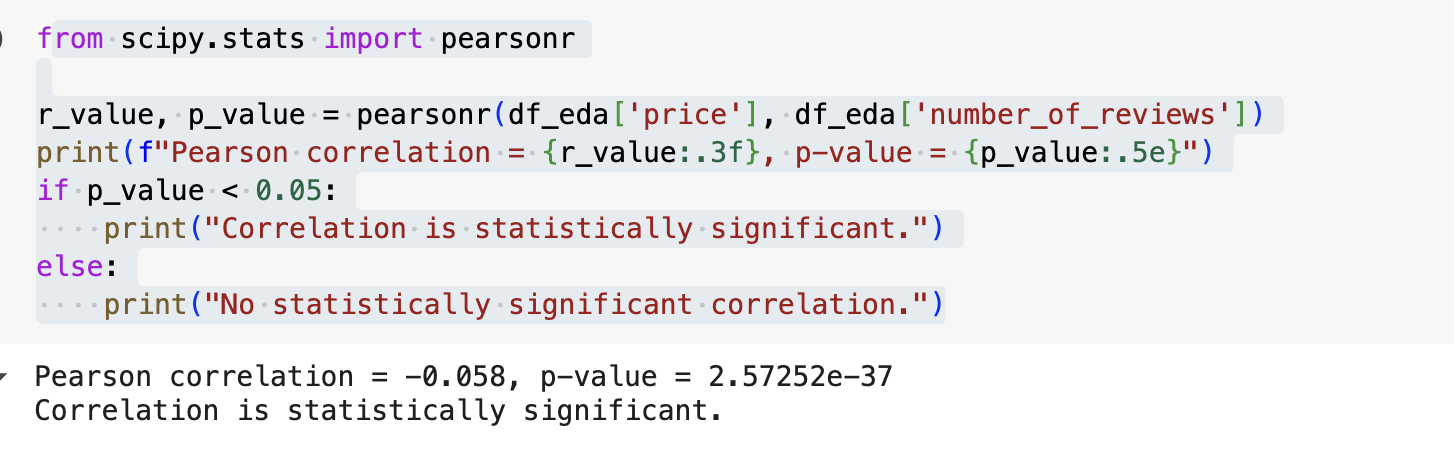


Correlation analysis between price and numeric features number\_of\_reviews, minimum\_nights, availability\_365, etc.

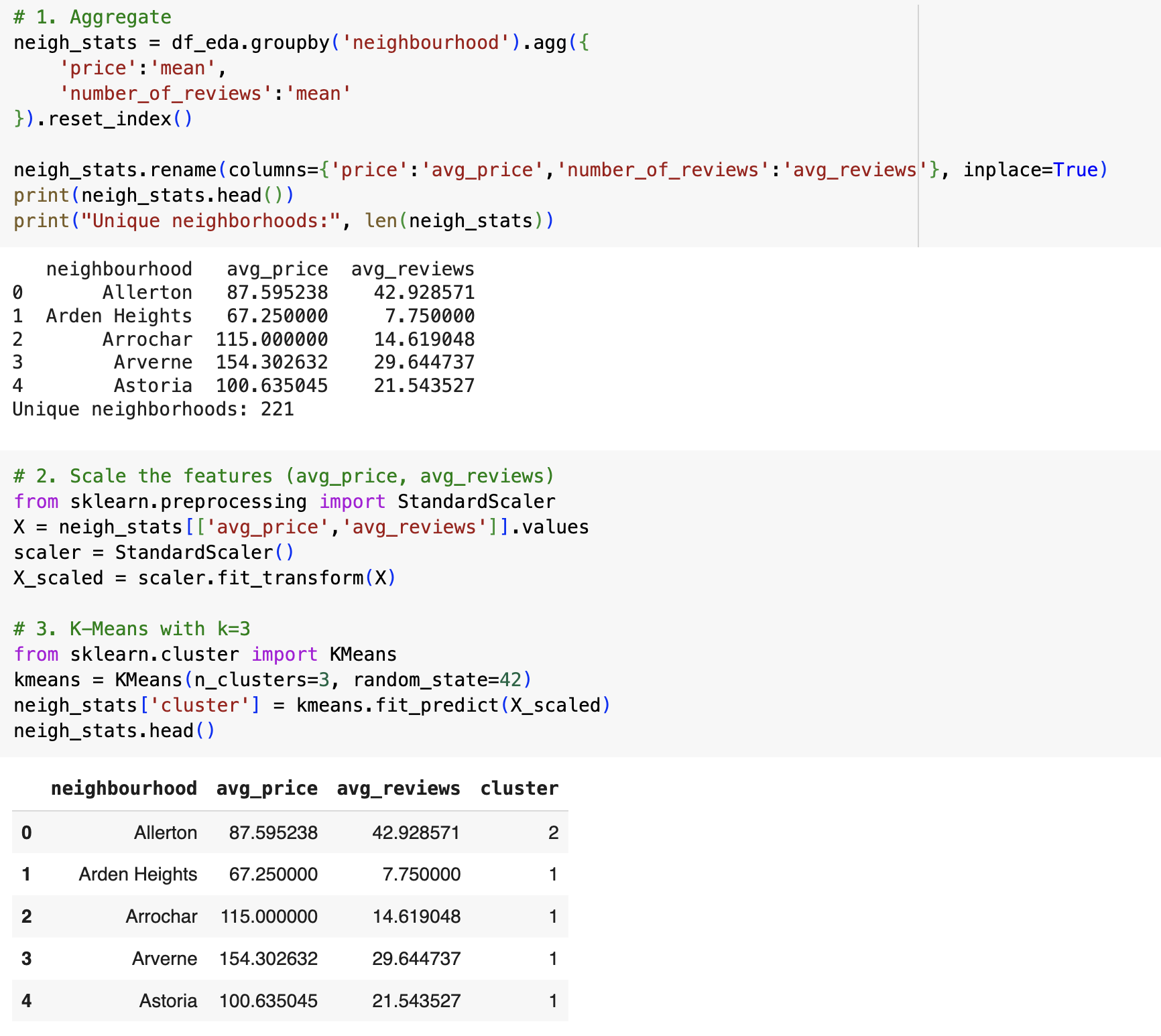


And plotting

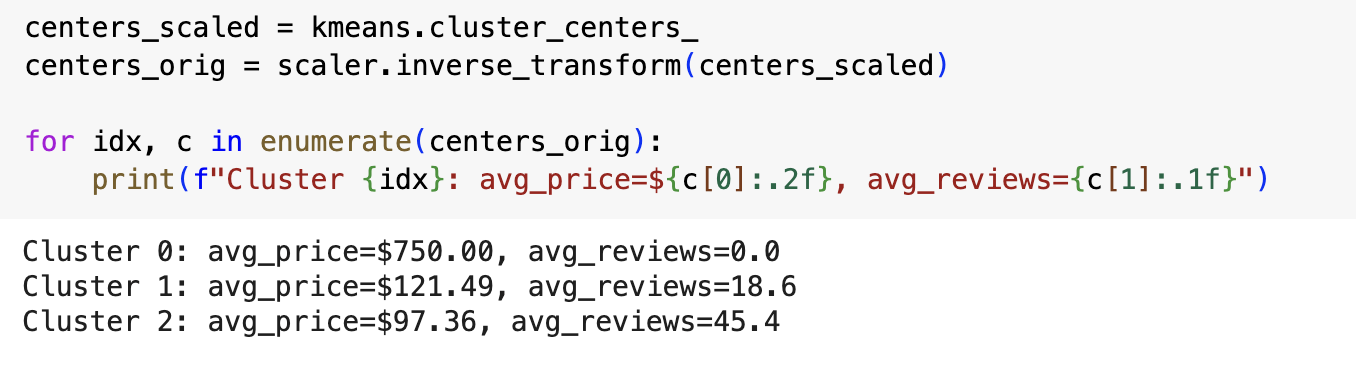
Pearson correlation for Price vs Numbers



K-means clustering by aggregating data by neighbourhood to compute mean price and mean reviews. Feature scaling and clustering.



Cluster profiling.

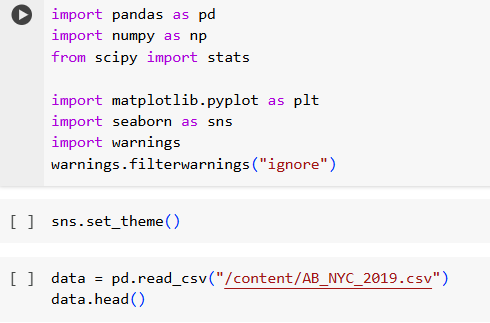


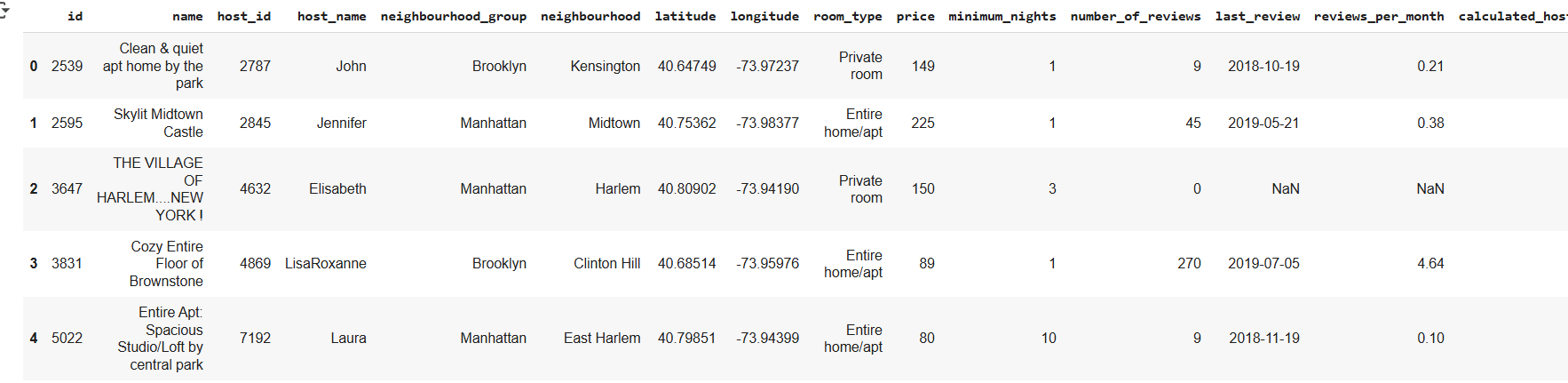
Plotting:



## Data Pre-Processing

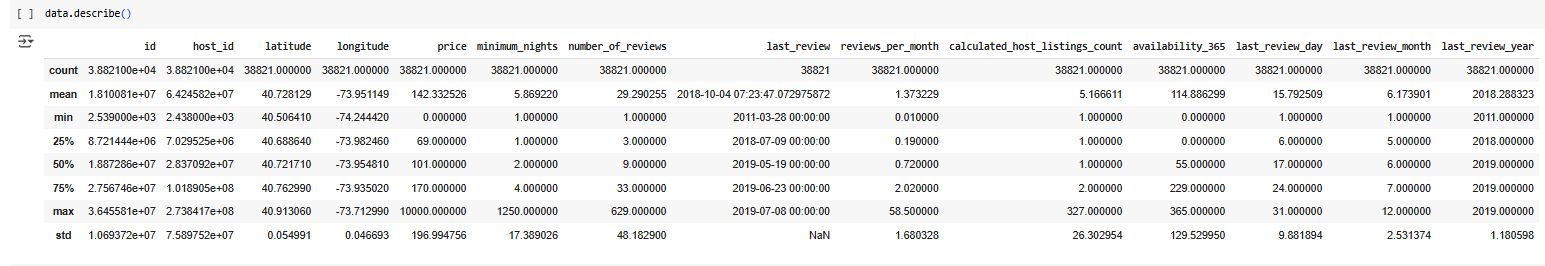
Libraries used, Data import and initial analysis:





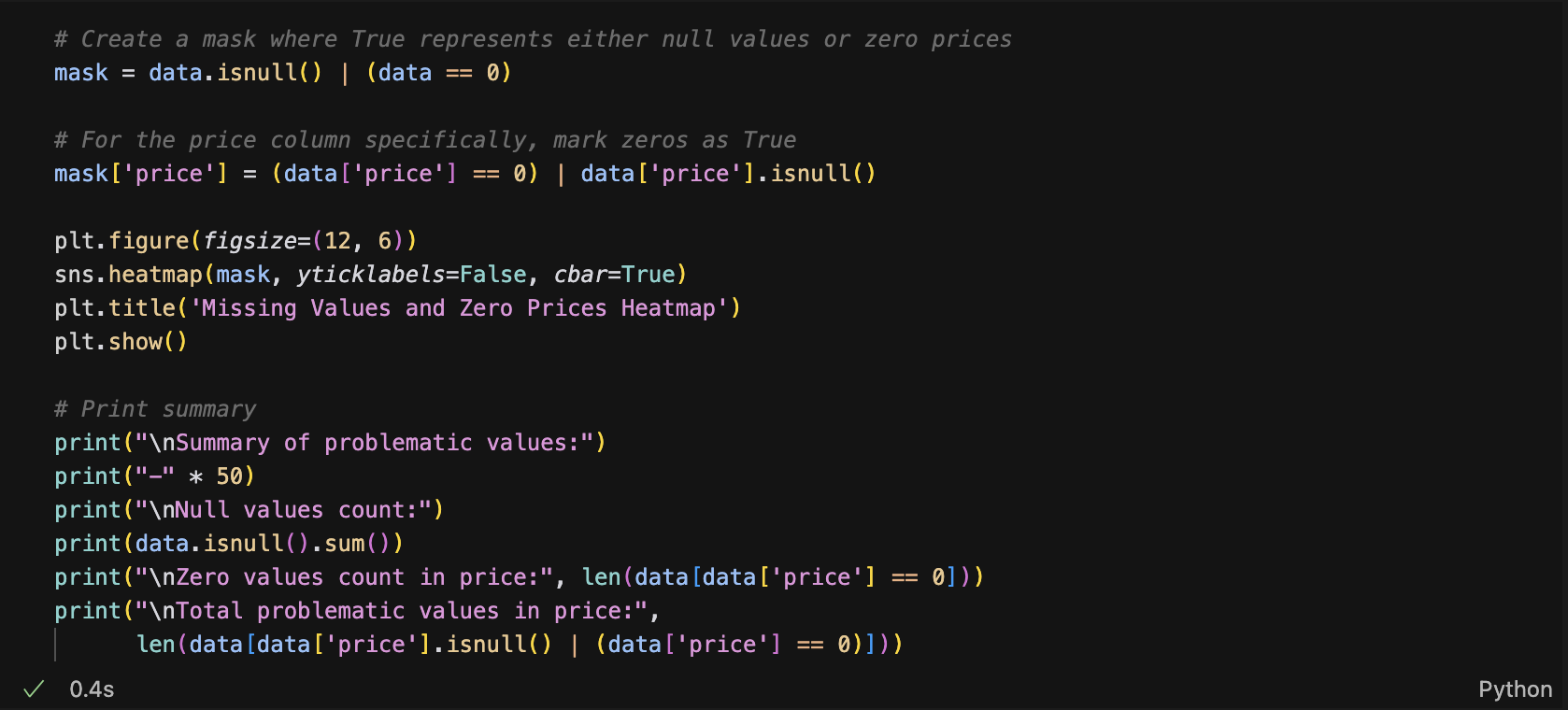


Descriptive statistics of the data set:

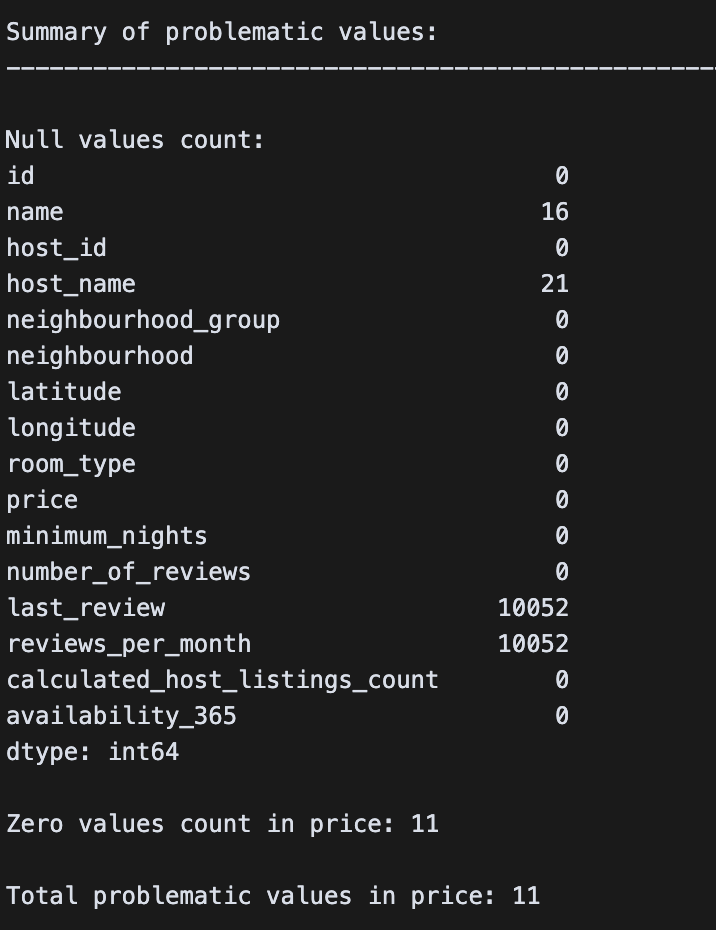
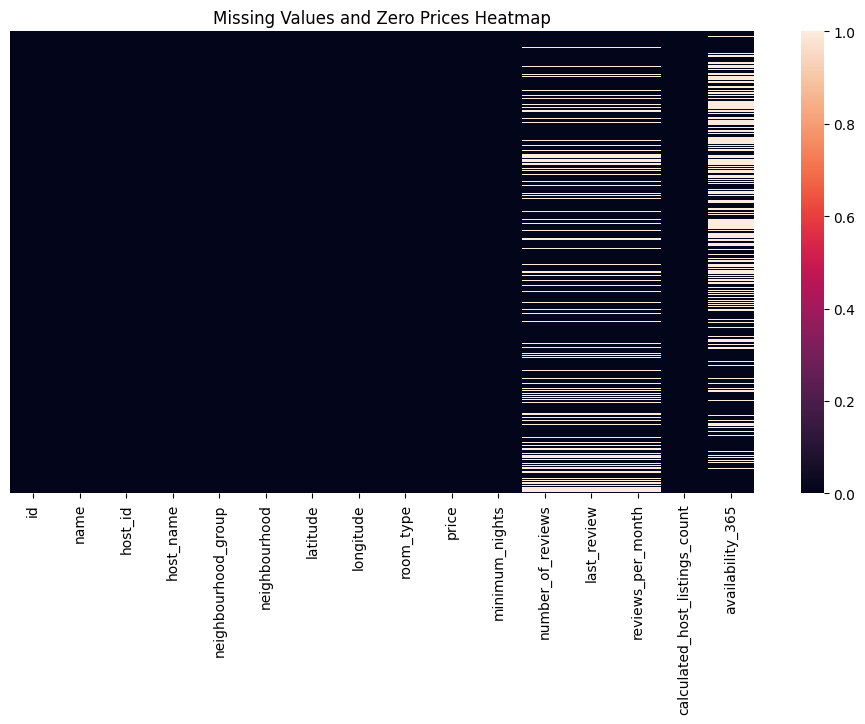




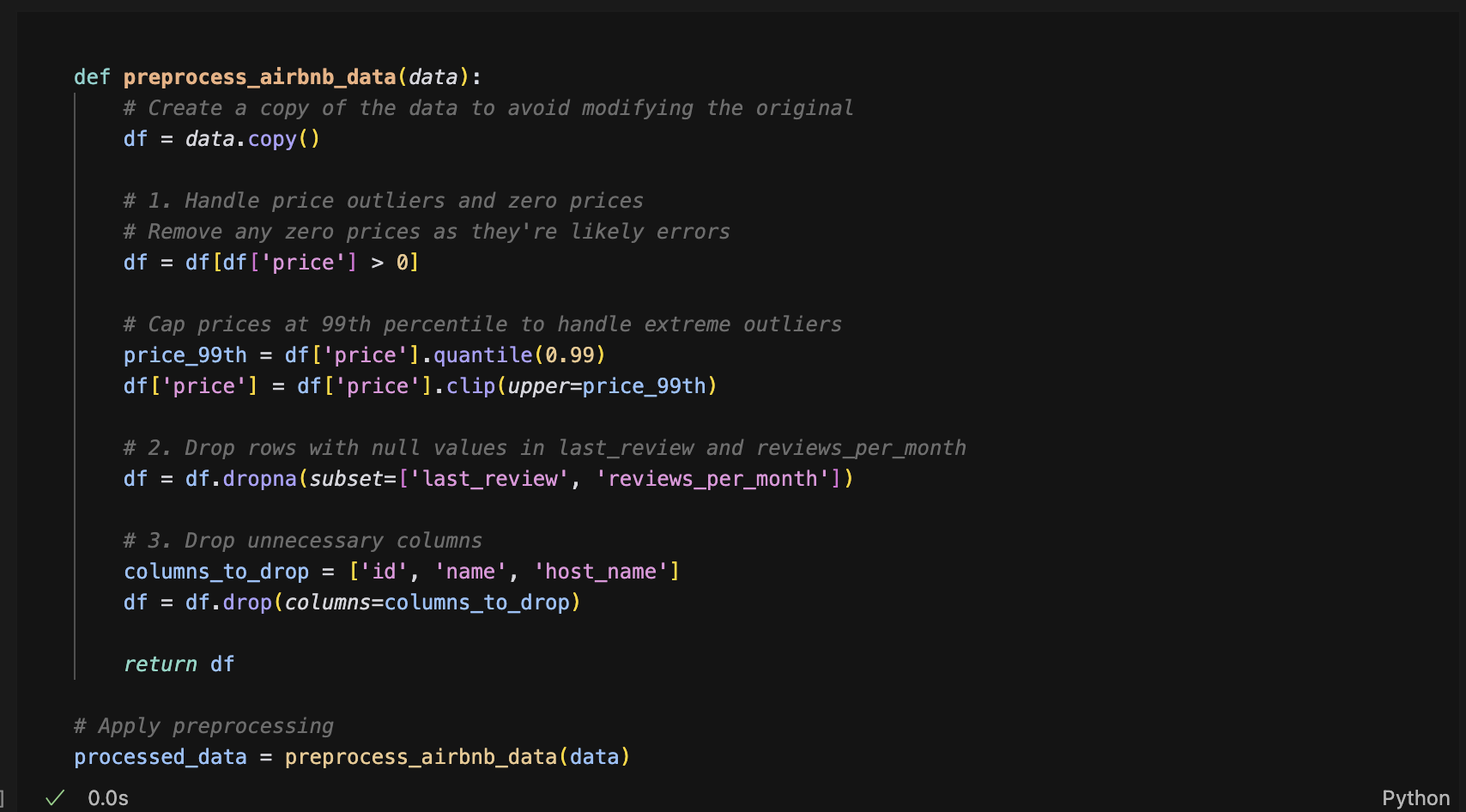
Code to create missing zeroes heatmap:



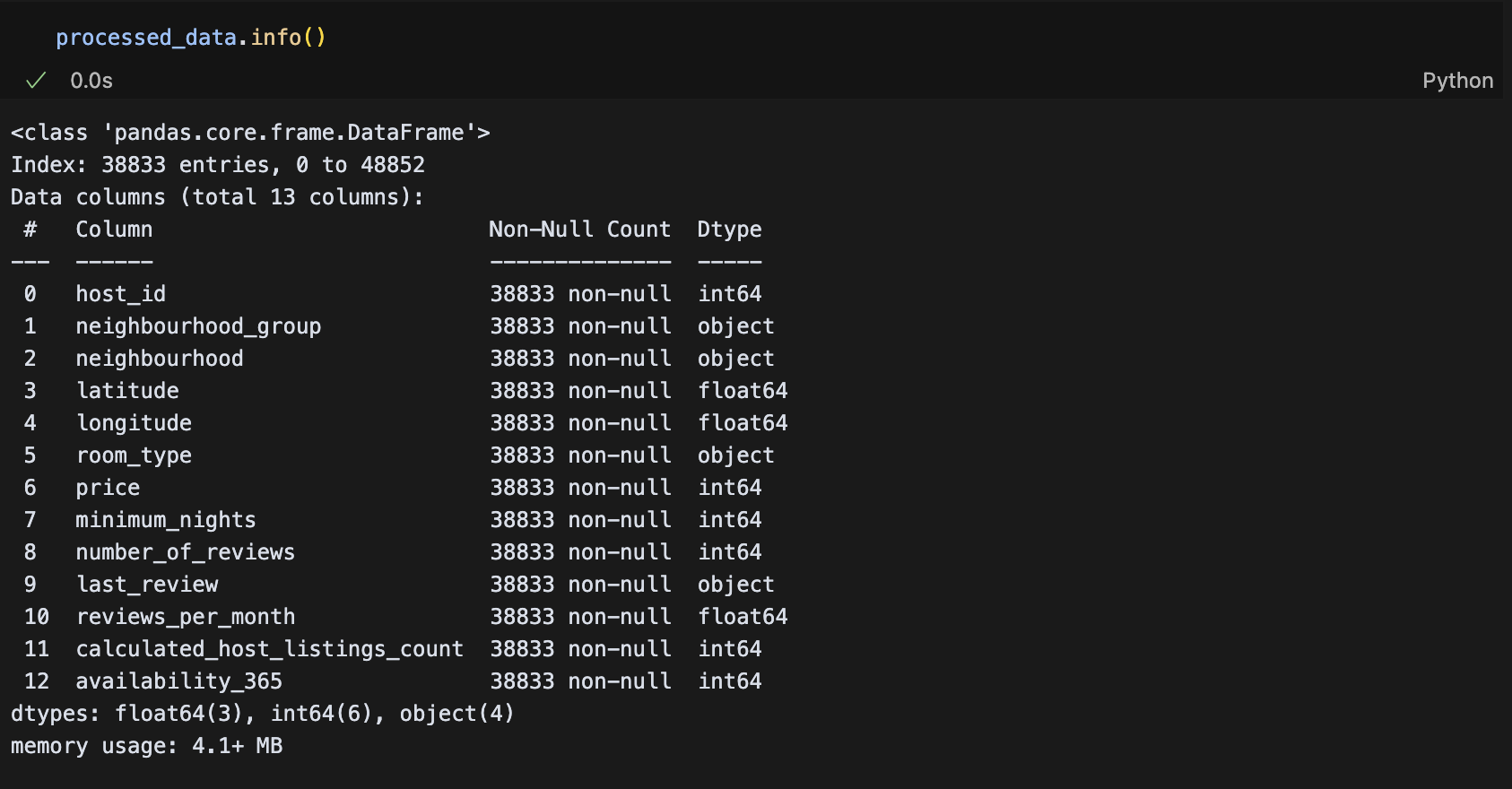
Visualization:



Pre-processing function:



Info on the pre-processed data:

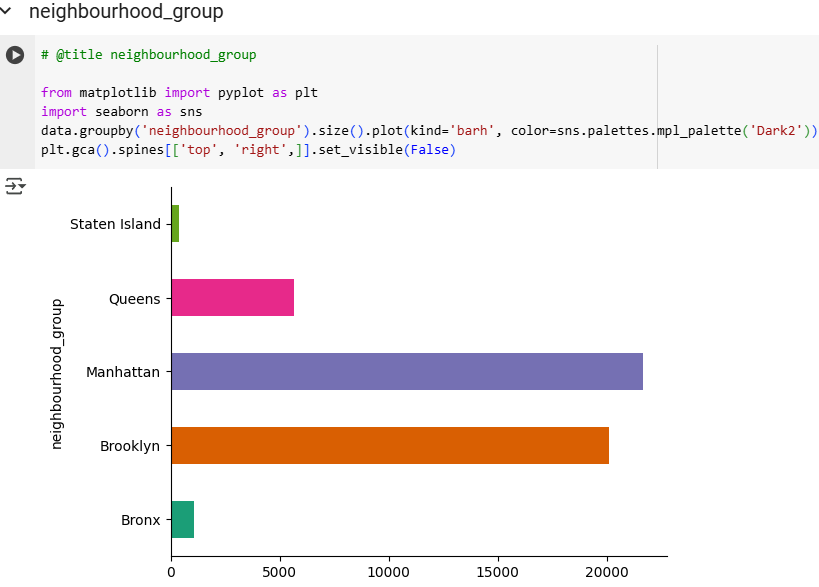
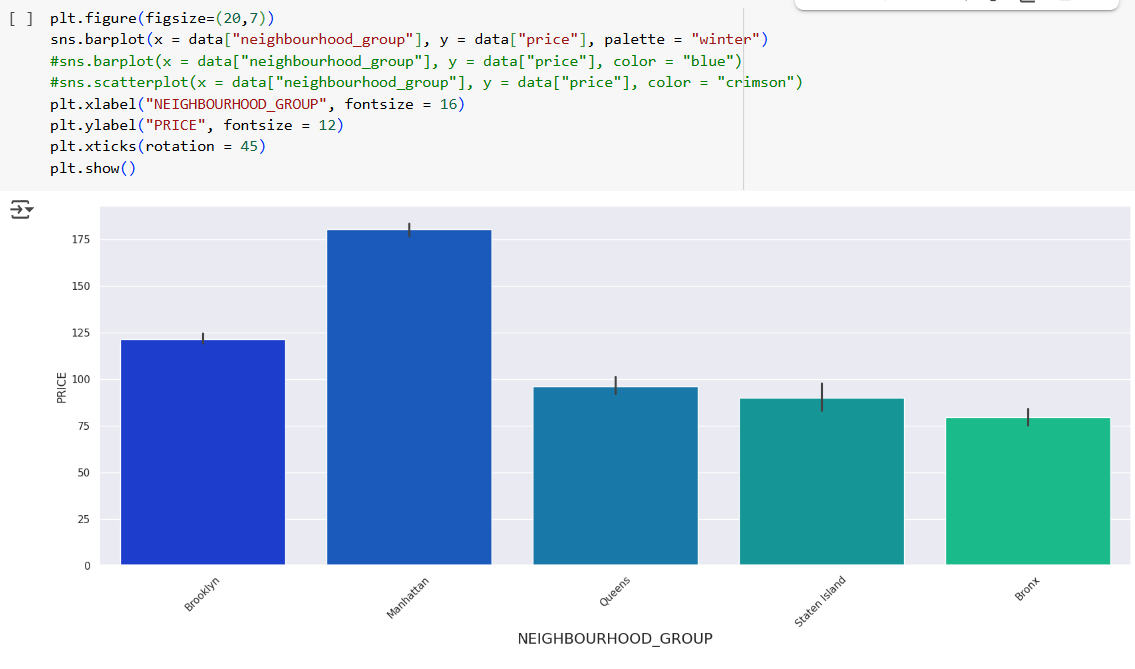


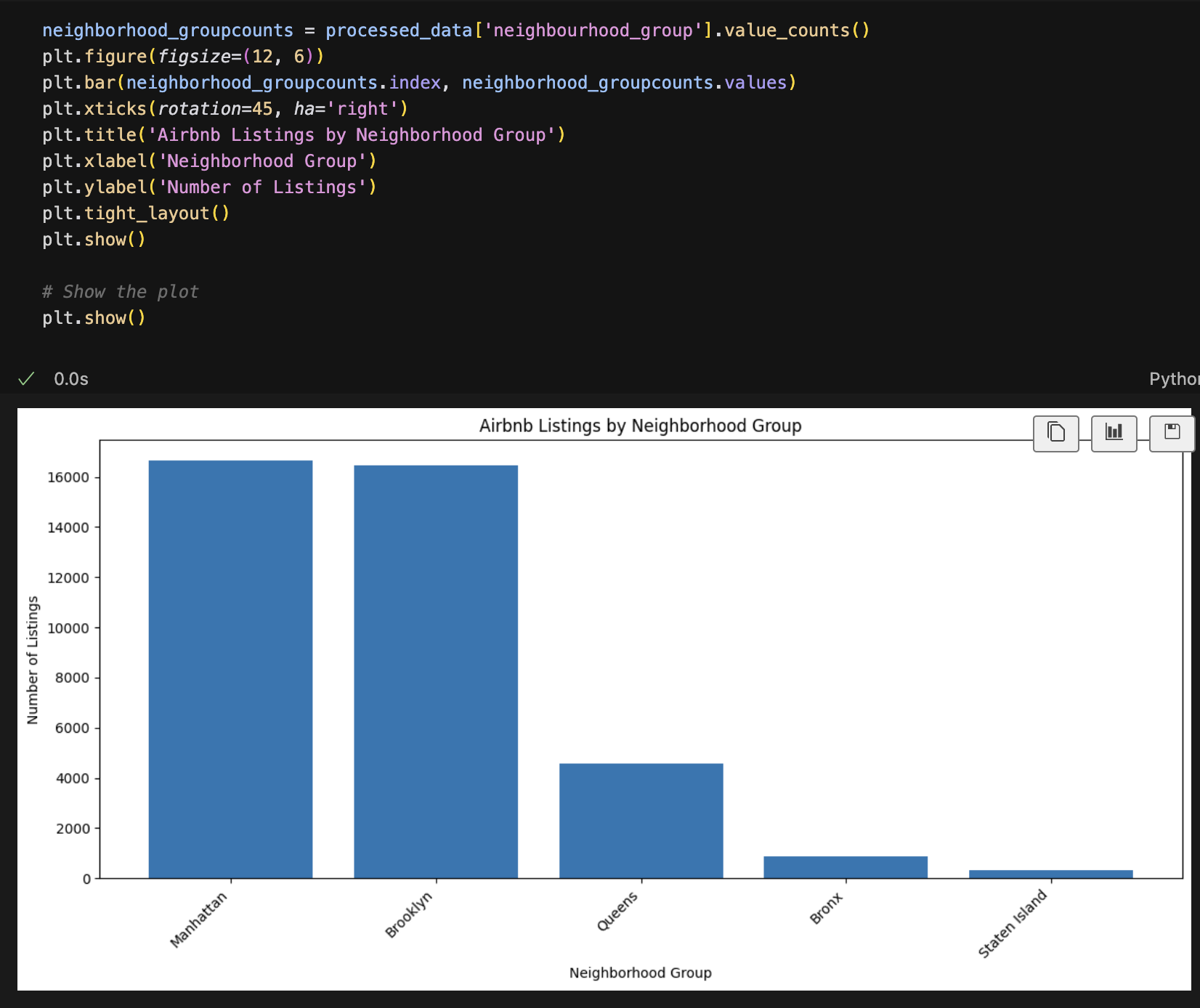
## EDA

Code and visualization of top 20 NYC Neighborhoods by number of listings.

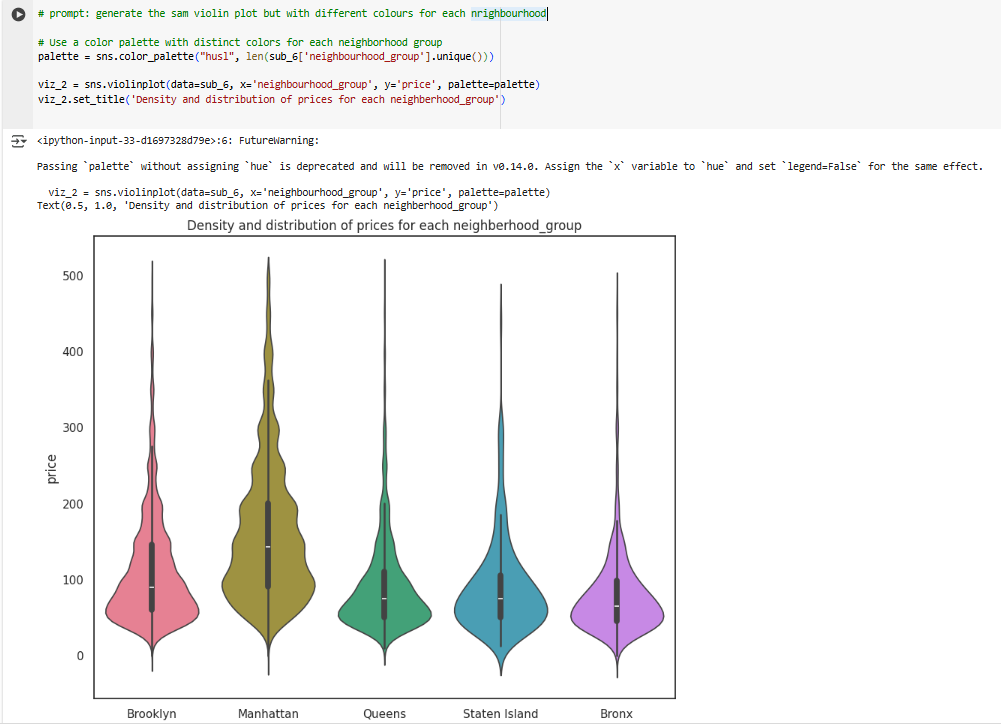
## 

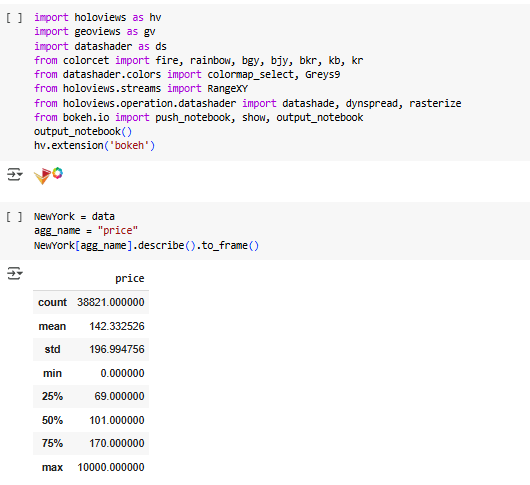
Code and visualization of Neighborhood groups by number of listings.

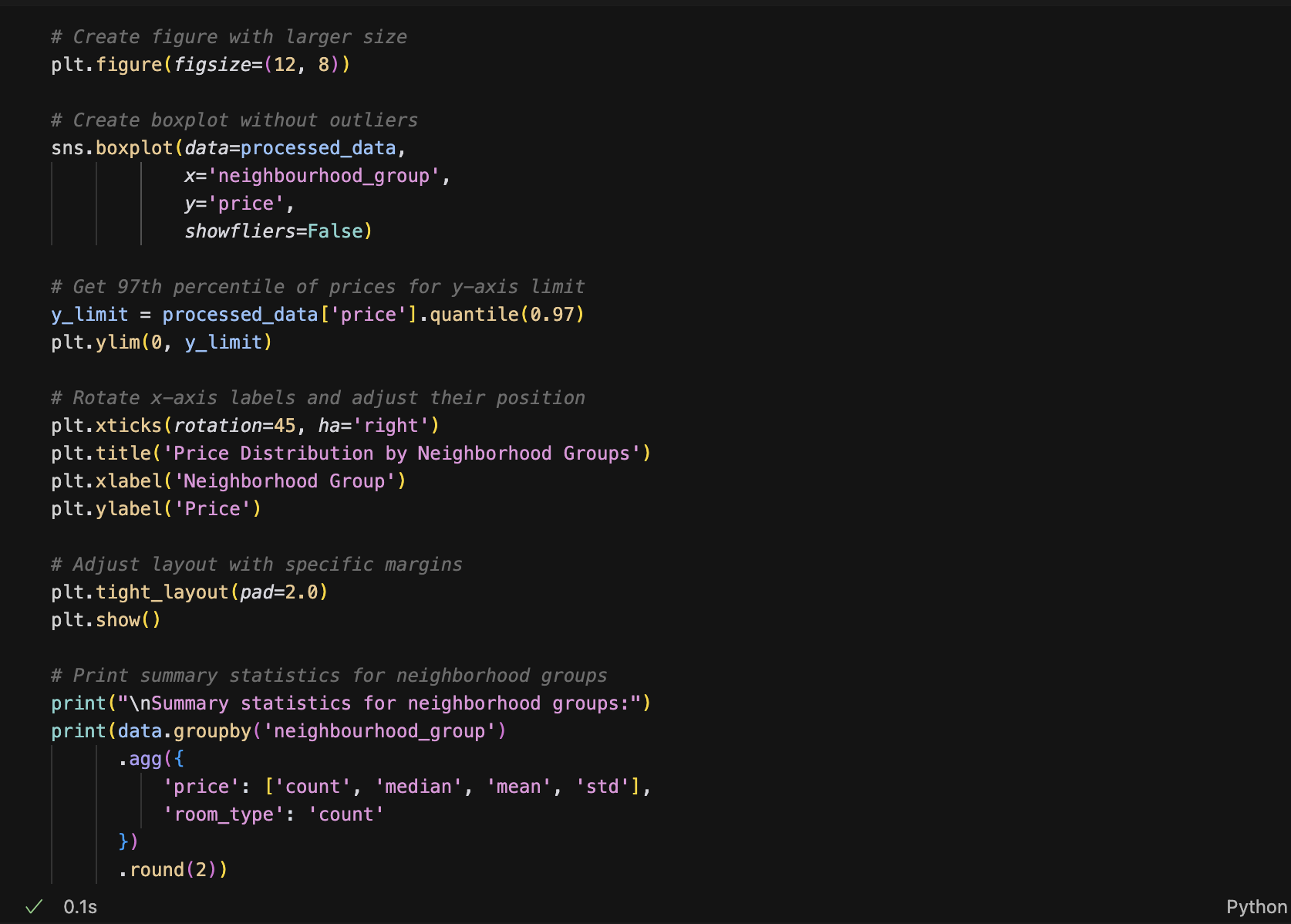


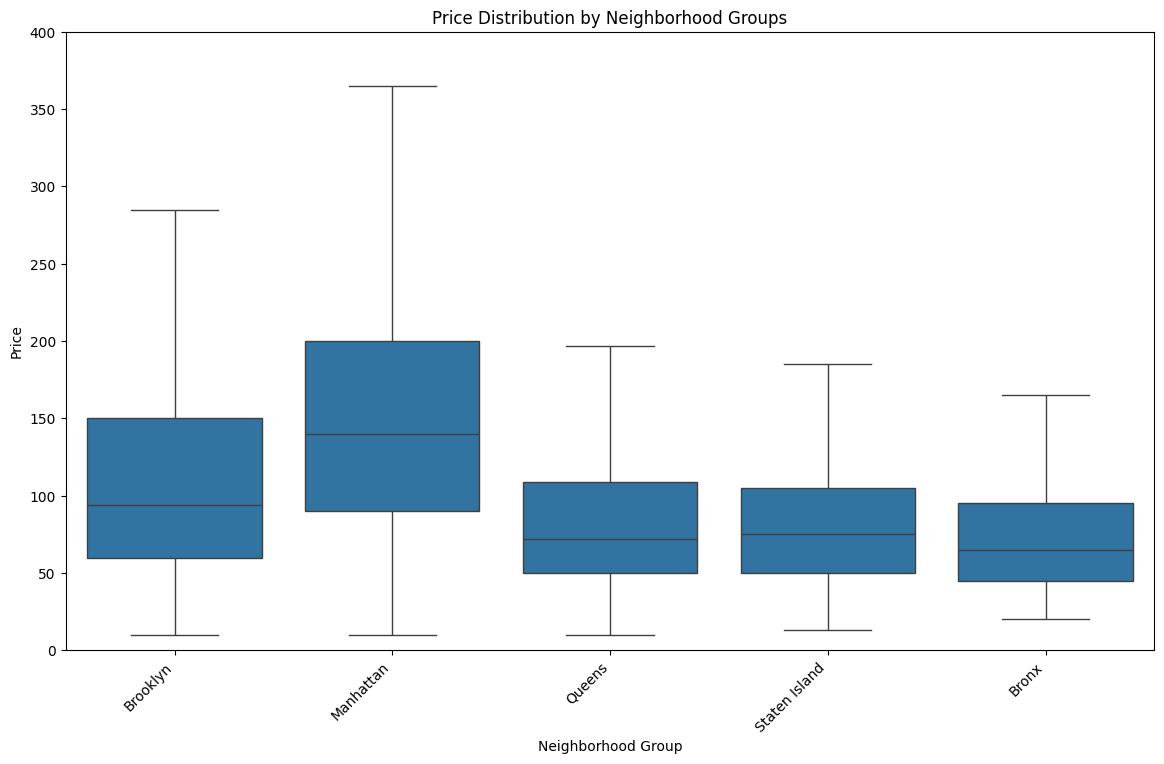


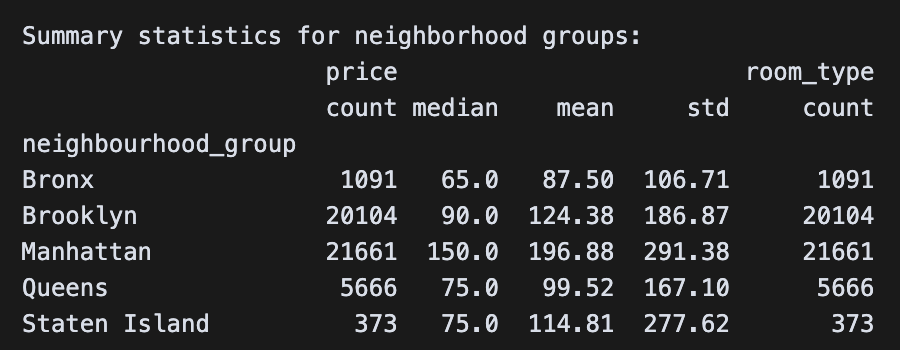
Code and visualization of price distribution per neighborhood group:



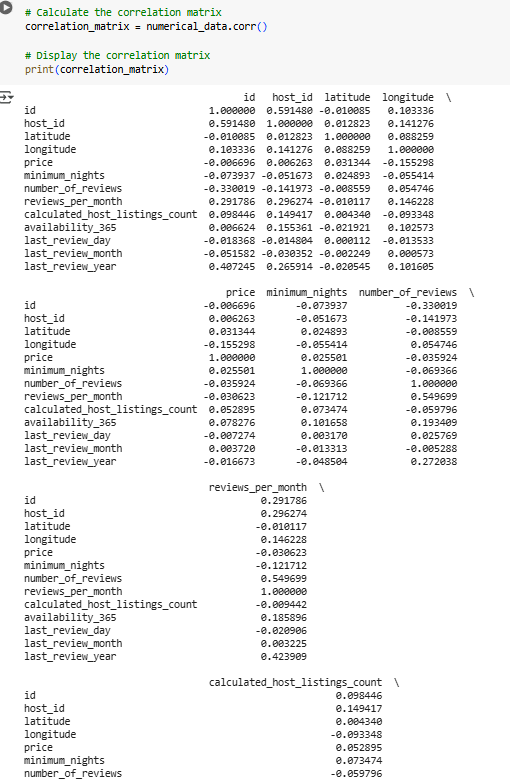


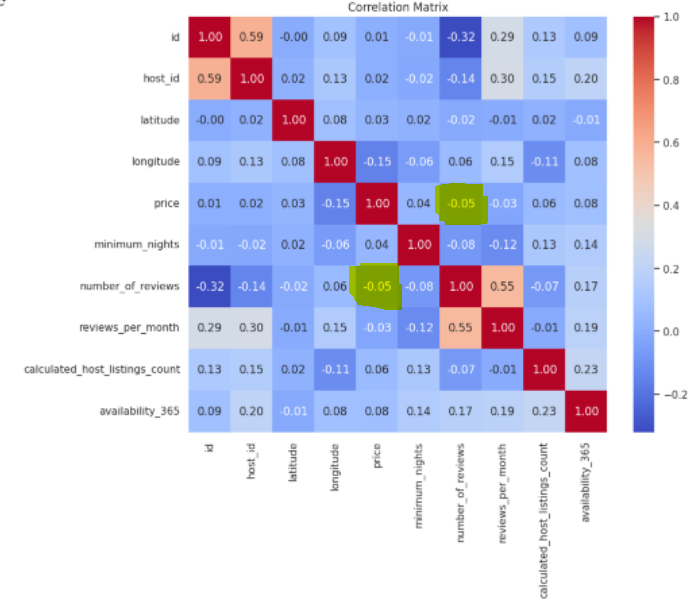
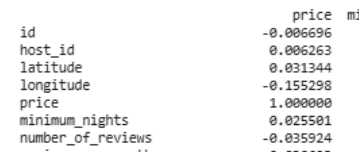




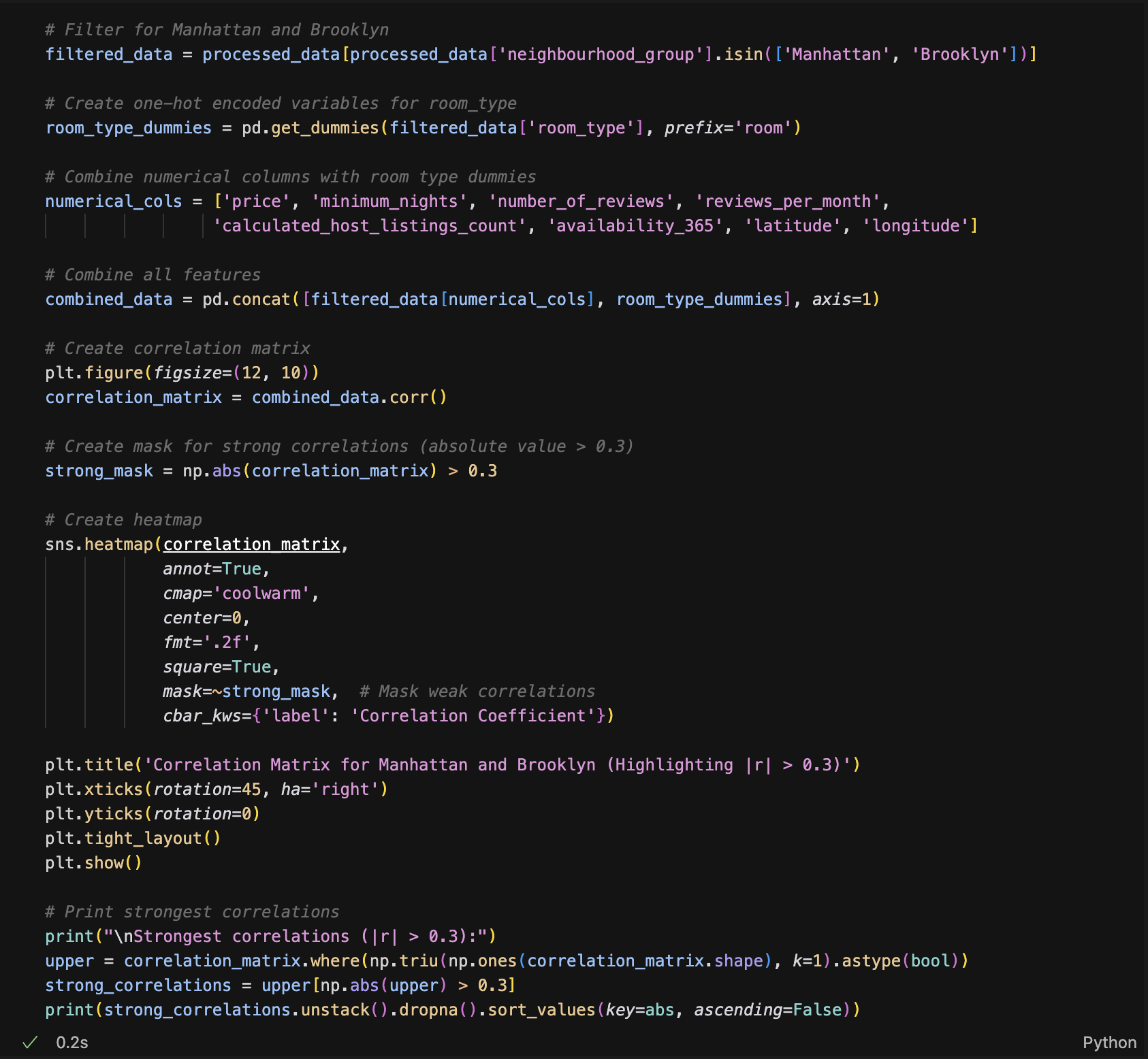


Calculation of the correlation matrix:

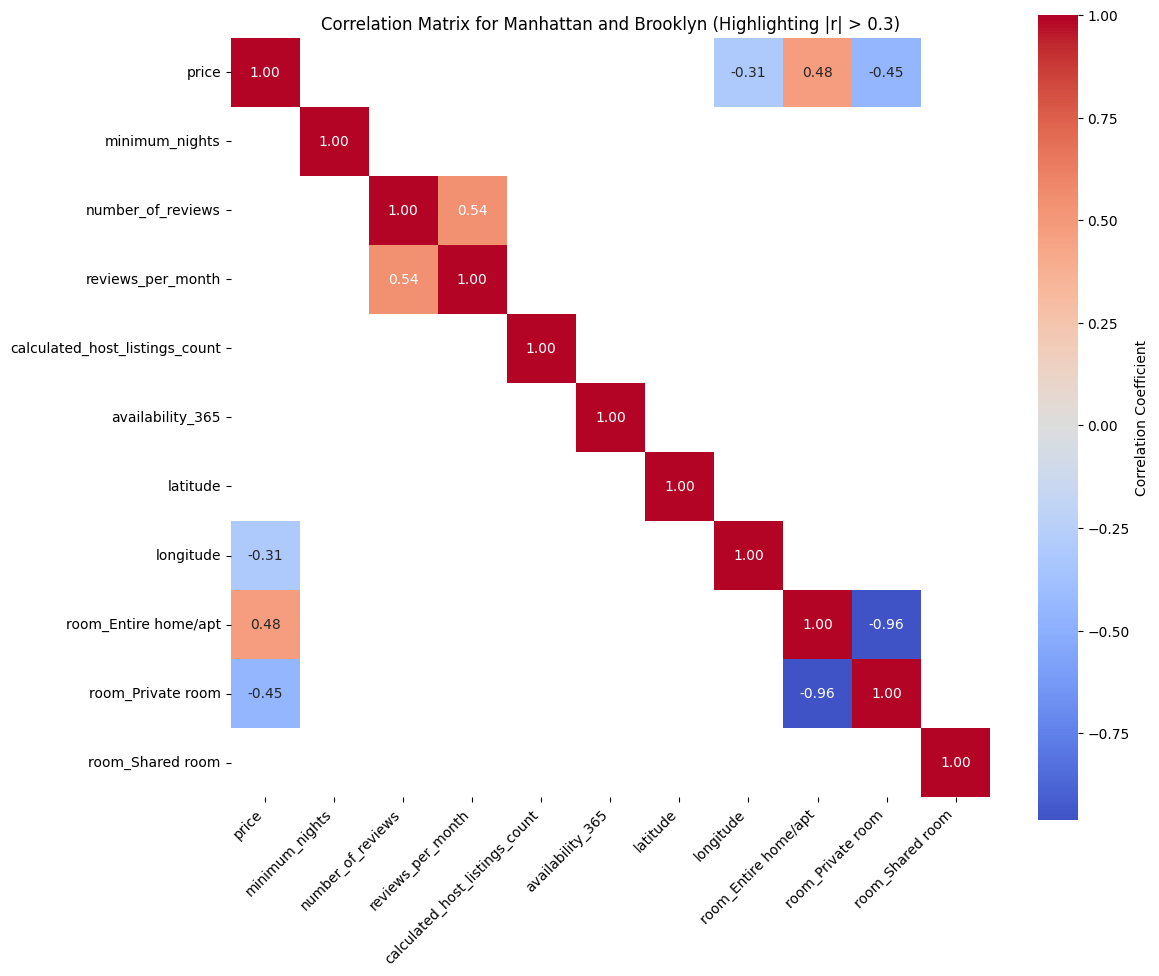


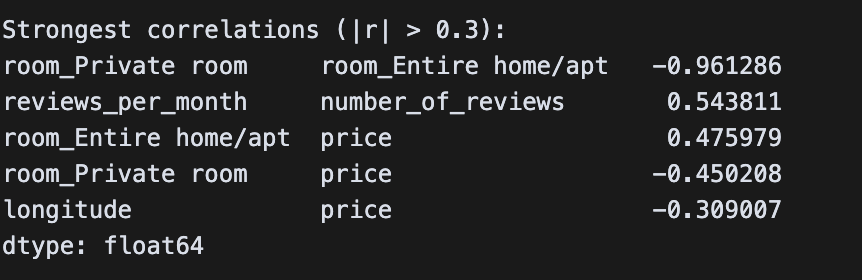


Code of the Correlation matrix used for only for Brooklyn and Manhattan

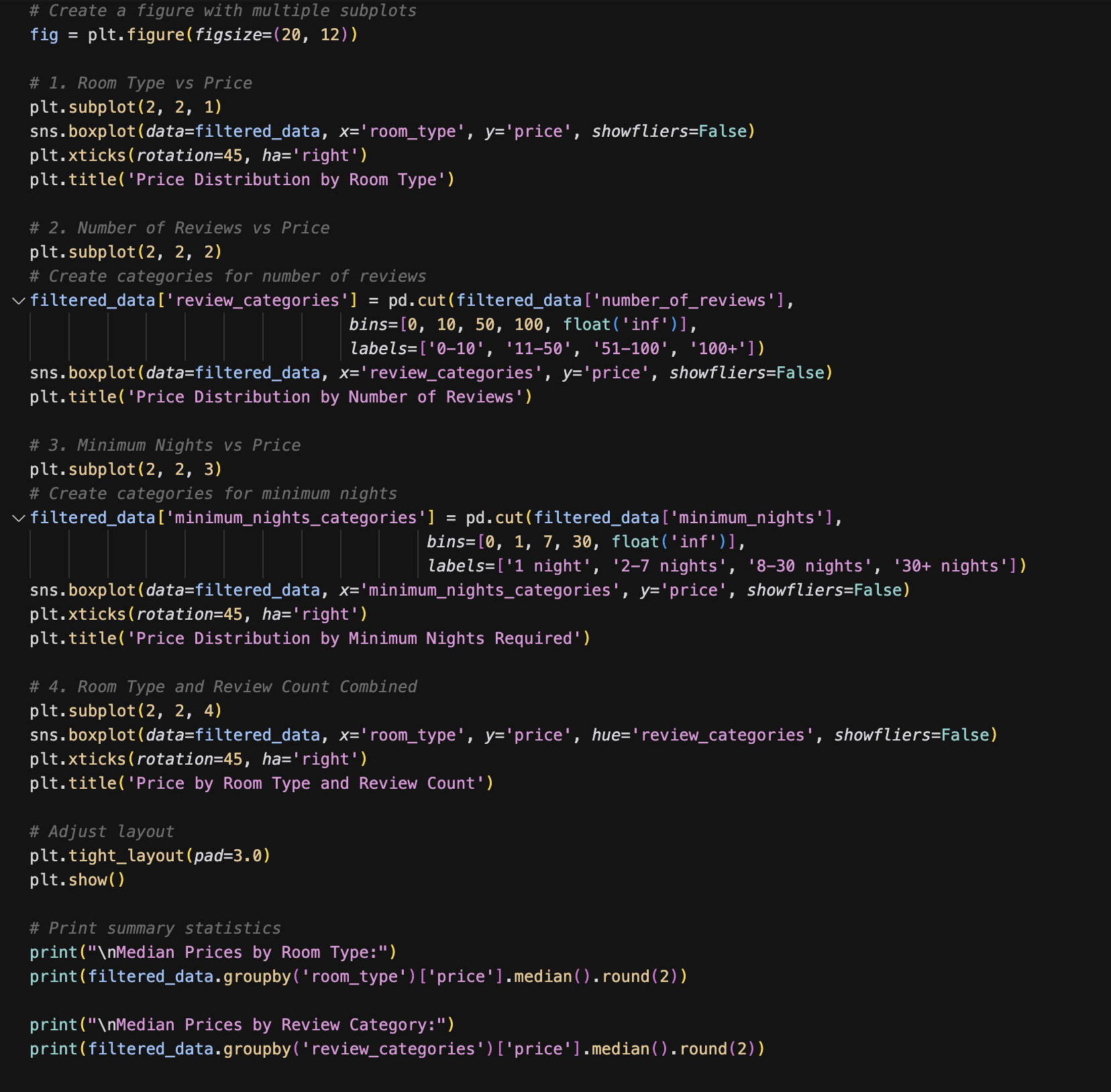


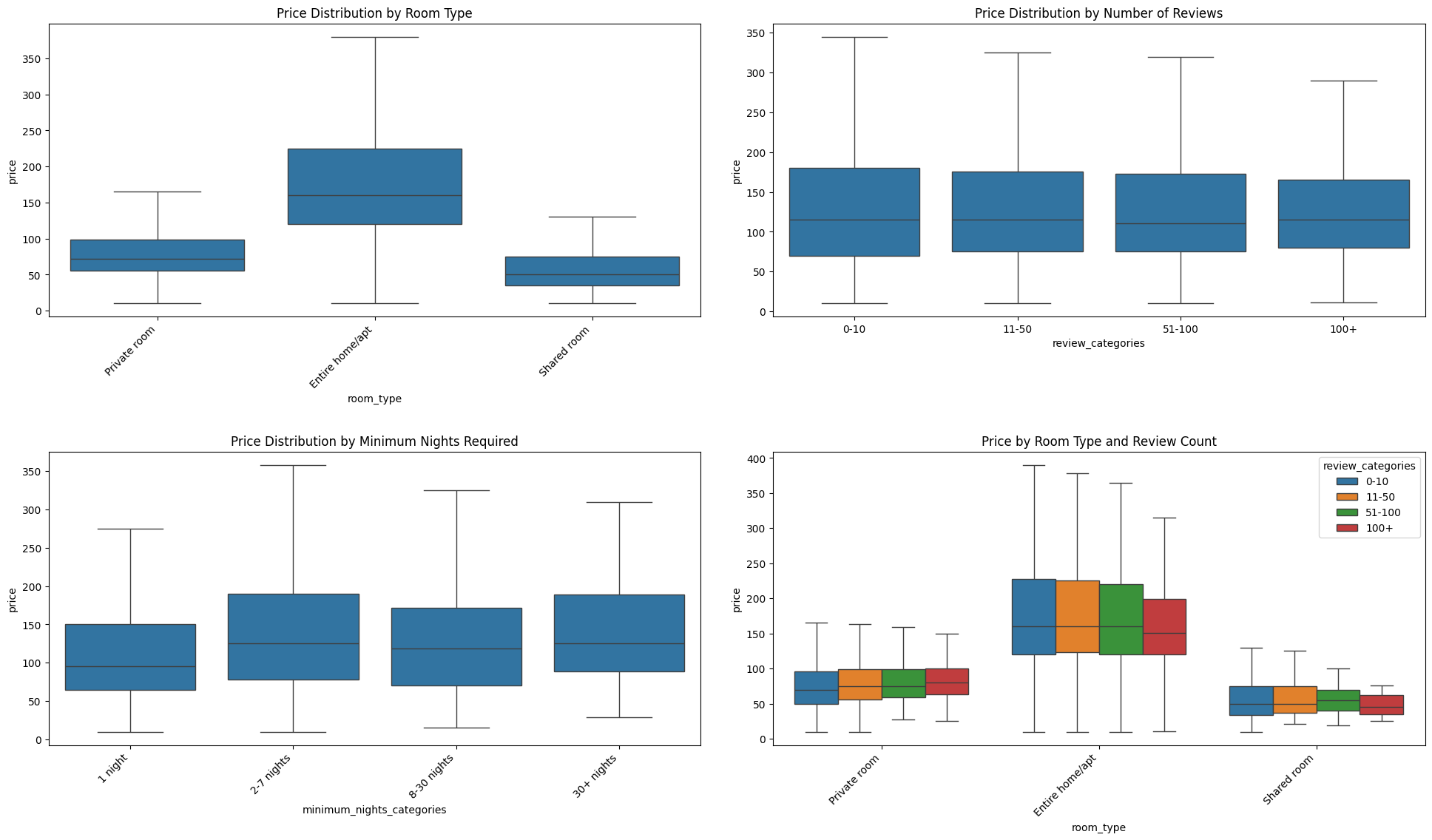
Visualisation:

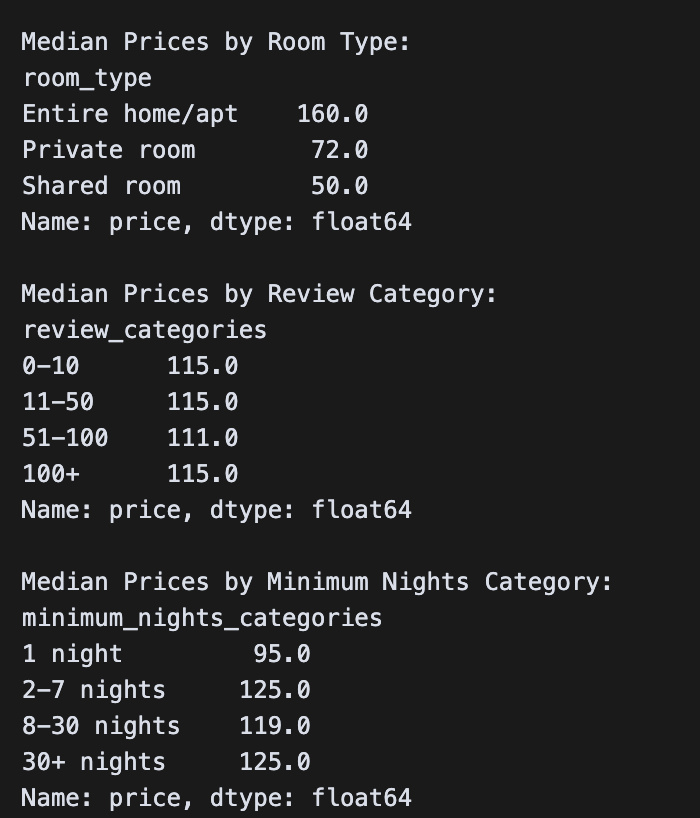


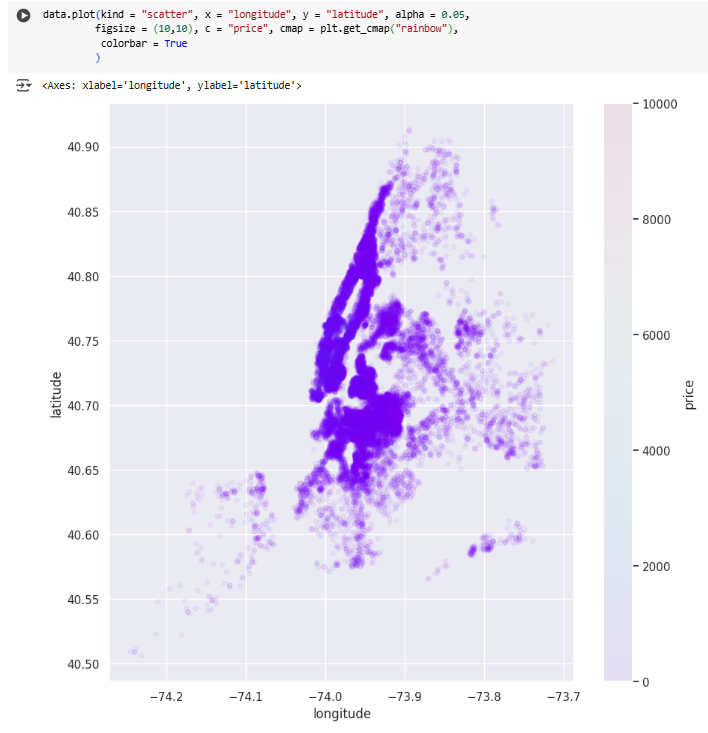


Code and visualization for Box plots comparing Price Distribution by Room Type, Number of Reviews, Minimum Nights Required and Room Type combined with Number of Reviews

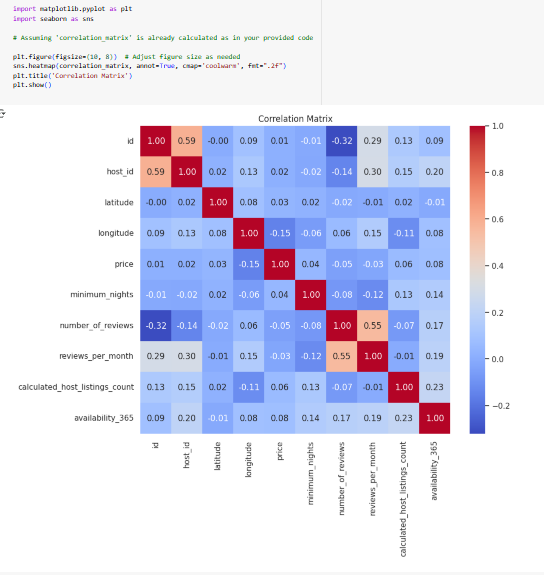


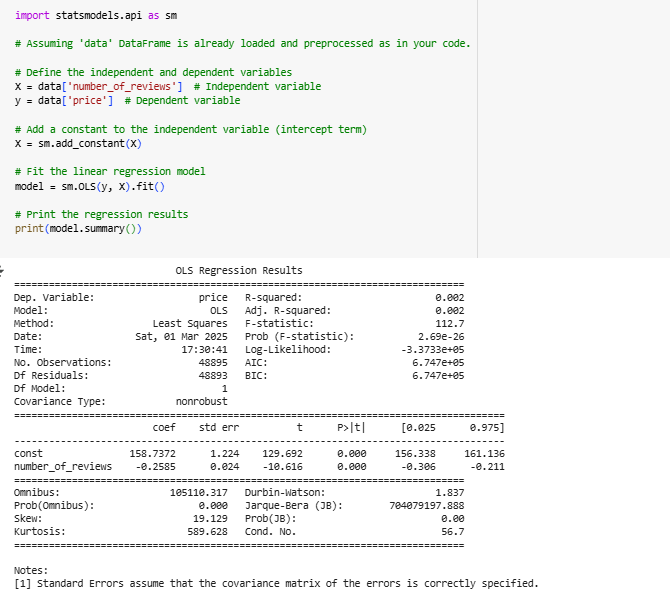




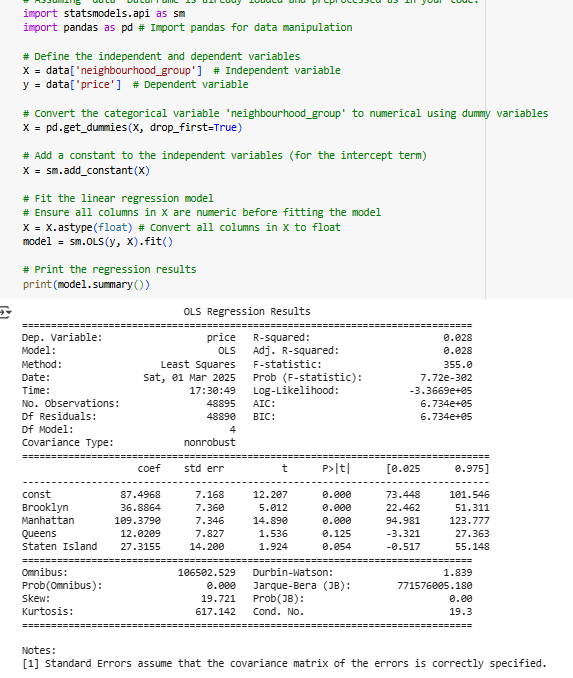


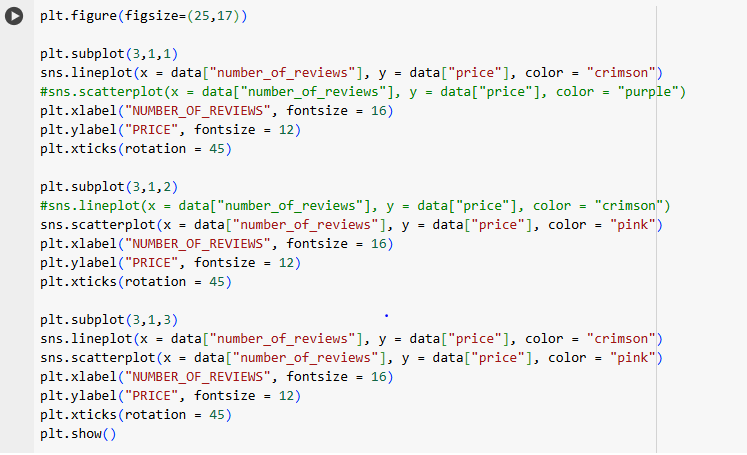
## Data Analysis

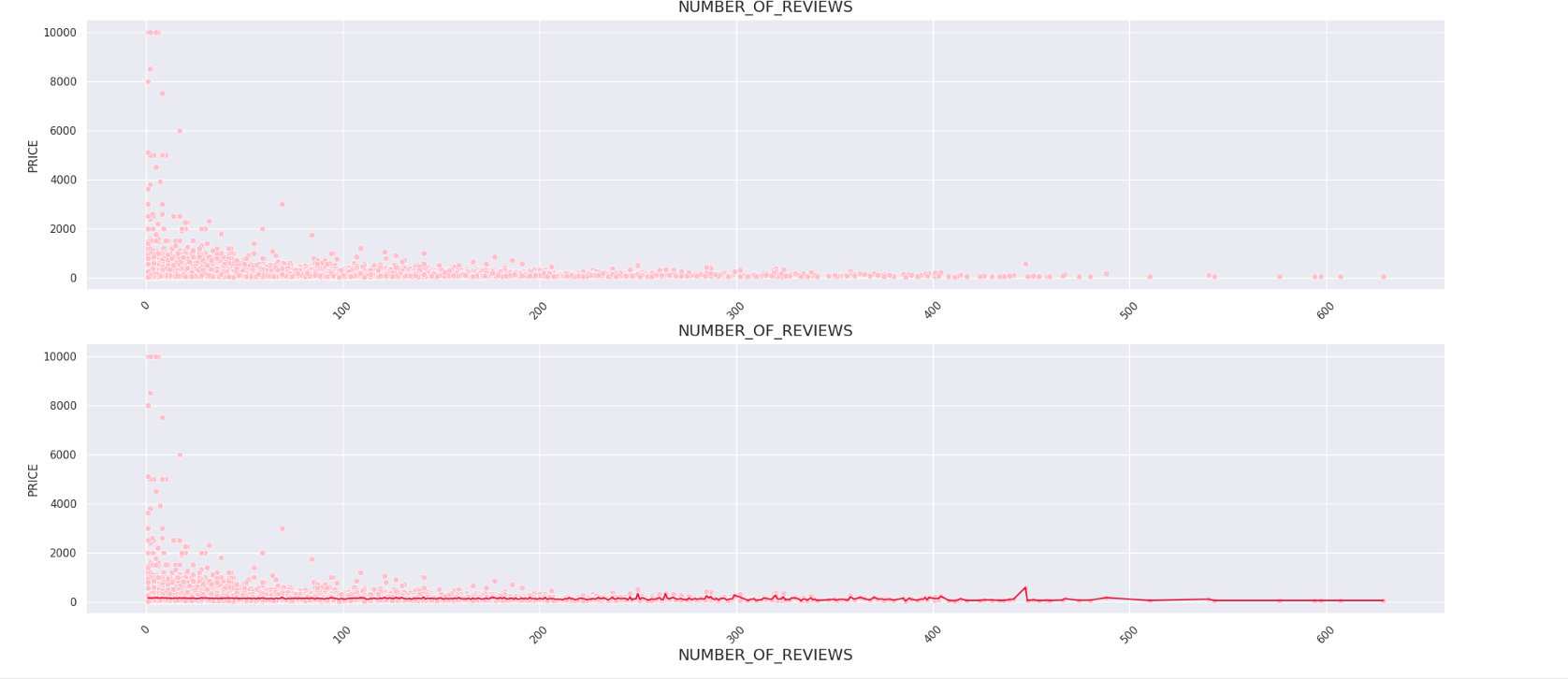




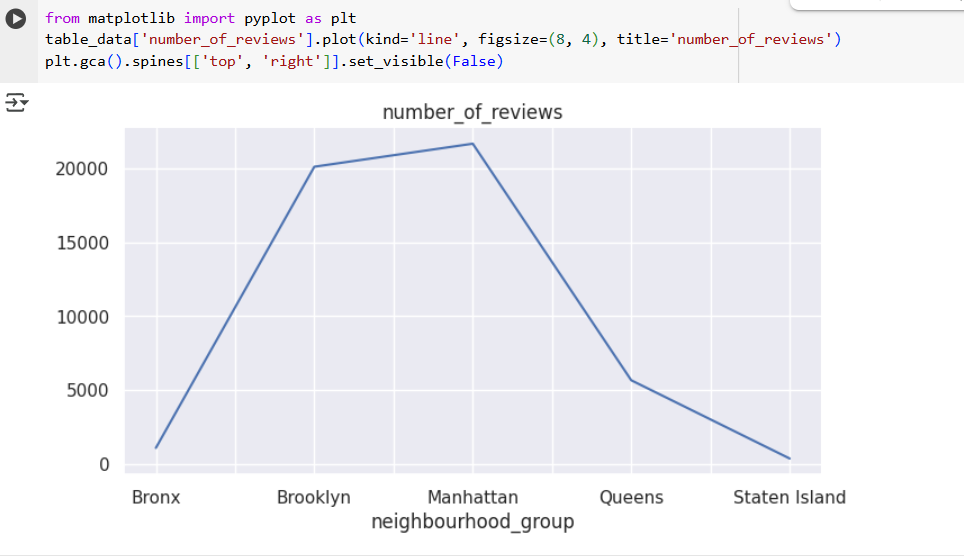
Model Fit (R-squared = 0.002, Adjusted R-squared = 0.002)

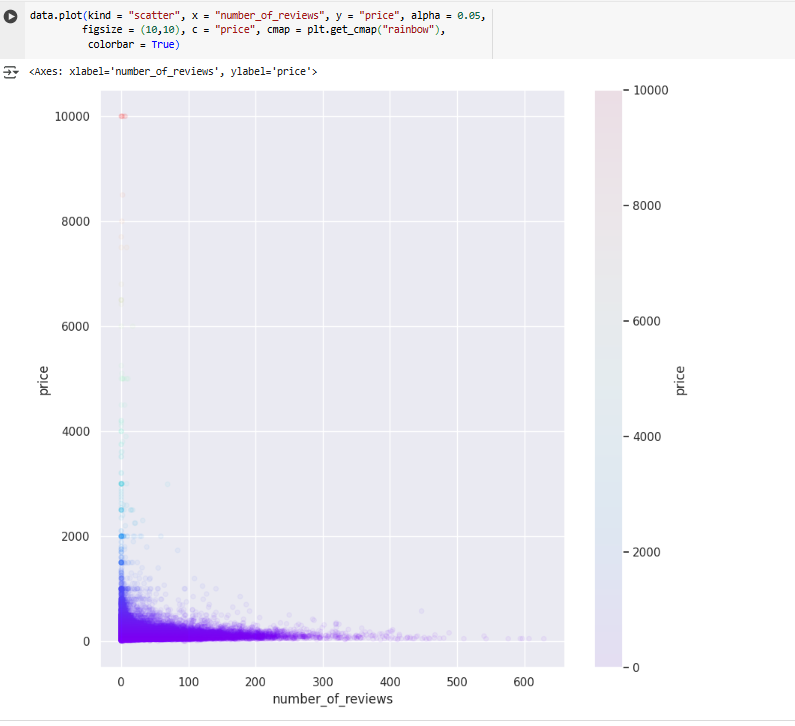






# 

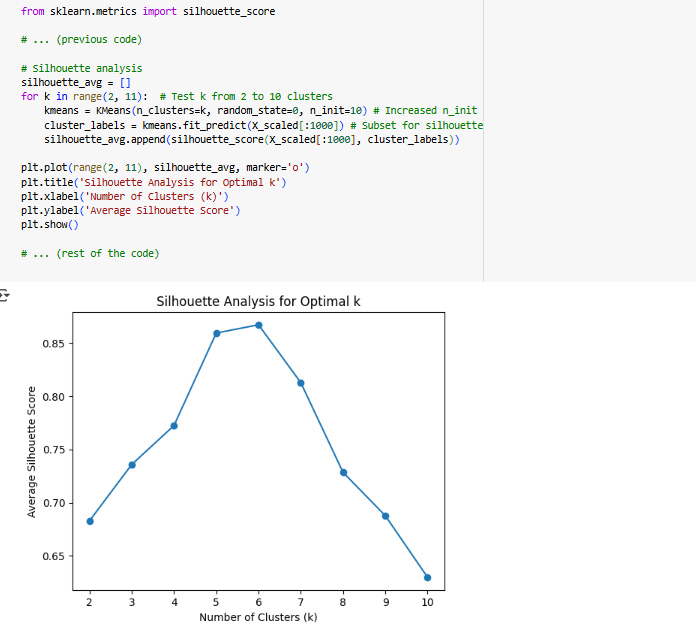


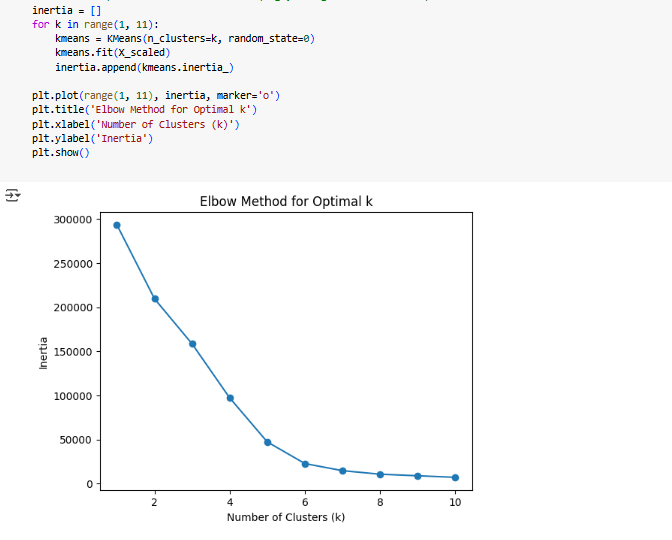


## 

## 

## Unsupervised learning-Clustering

Silhouette Analysis

Elbow Method

# 

# 

# 

# References:

Carvache-Franco, M., Regalado-Pezúa, O., Sirkis, G., Carvache-Franco, O., &

Carvache-Franco, W. (2023). Market segmentation in urban tourism: A study in Latin America. *PLOS ONE, 18*(5), e0285138.<https://doi.org/10.1371/journal.pone.0285138> (Accessed: 7 March 2025).

Chatterjee, S., & Hadi, A. S. (2006). *Regression Analysis by Example* (4th ed.).

Hoboken, NJ: Wiley-Interscience.

Coles, P. A., Egesdal, M., Ellen, I. G., & Sundararajan, A. (2017). Airbnb usage across

New York City neighborhoods: Geographic patterns and regulatory implications. *Cambridge Handbook on the Law of the Sharing Economy*.<https://doi.org/10.2139/ssrn.3048397>

Field, A. P. (2024). *Discovering statistics using IBM SPSS statistics* (6th ed.). Los

Angeles: Sage.

Jaureguizar Cervera, D., Pérez-Bustamante Yábar, D.C., & de Esteban Curiel, J.

(2022). Factors affecting short-term rental first price: A revenue management model. *Frontiers in Psychology, 13*, article 994910.<https://doi.org/10.3389/fpsyg.2022.994910> (Accessed: 7 March 2025).

Jiao, J., & Bai, S. (2019). Cities reshaped by Airbnb: A case study in New York City,

Chicago, and Los Angeles. *Environment and Planning A: Economy and Space, 52*(1), 10–13.<https://doi.org/10.1177/0308518x19853275>

Martinez, R.D., Carrington, A., Kuo, T., Tarhuni, L., & Abdel-Motaal, N.A.Z. (2017). The

impact of an Airbnb host’s listing description “sentiment” and length on occupancy rates. *arXiv preprint arXiv:1711.09196* [cs.CY].<https://arxiv.org/abs/1711.09196> (Accessed: 7 March 2025).

Nivón, T. (2019). Data Analysis on Airbnb in NYC. NYC Data Science Academy Blog,

Dec 2, 2019. (Analysis of Inside Airbnb NYC September 2019 data). Available at: nycdatascience.com.

Perez-Sanchez, V.R., Serrano-Estrada, L., Marti, P., & Mora-Garcia, R-T. (2018). The

what, where, and why of Airbnb price determinants. *Sustainability, 10*(12), 4596.<https://doi.org/10.3390/su10124596>

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (2017). *Understanding regression*

*analysis: An introductory guide* (2nd ed.). Los Angeles, CA: SAGE Publications, Inc.

Sharma, U., & Gupta, D. (2021). Analyzing the applications of the internet of things in

the hotel industry. *Journal of Physics: Conference Series*.<https://doi.org/10.1088/1742-6596%2F1969%2F1%2F012041>

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), 2487–2509.

Wedel, M., & Kamakura, W.A. (2000). *Market segmentation: Conceptual and*

*methodological foundations* (2nd ed.). Boston: Kluwer Academic Publishers.