draft

**Introduction**  
The growth of online platforms such as Airbnb has reshaped the hotel industry by allowing travelers greater influence over pricing, reviews, and location choices (Sharma & Gupta, 2021). Today, Airbnb stands as one of the leading accommodation options globally, with a particularly strong presence in New York City (Jiao & Bai, 2019). This report examines Airbnb listings in NYC, with emphasis on pricing, availability, location, guest reviews, and property types. Gaining insights into these elements can support Airbnb in refining its pricing strategies and enhancing guest experiences.

**1.Proposed Business Analytic Question**

Which neighborhoods in New York City recorded the highest average listing prices, and what combination of factors contributed to these elevated costs? Furthermore, is there a statistically significant relationship between listing prices and the number of reviews a property receives, and what insights can this relationship provide about pricing strategies? Finally, how can understanding these patterns help Airbnb hosts optimize pricing, and improve overall guest engagement in a highly competitive urban market?

**2. Data Analysis**

**Data Preprocessing**  
The original dataset consisted of 48,895 rows and 16 columns. During preprocessing, we removed listings with zero reviews, zero prices, and extreme outliers to ensure data quality and reliability. Following these cleaning steps, the dataset was reduced to 36,753 valid entries, forming the basis for subsequent analysis.

**Exploratory Data Analysis (EDA)**  
An examination of Airbnb listings across different neighborhood groups revealed that Manhattan had the largest concentration of listings, with Brooklyn trailing closely behind. In contrast, Staten Island and the Bronx had significantly fewer listings, indicating a notable disparity in Airbnb presence across NYC boroughs (Figure 1).

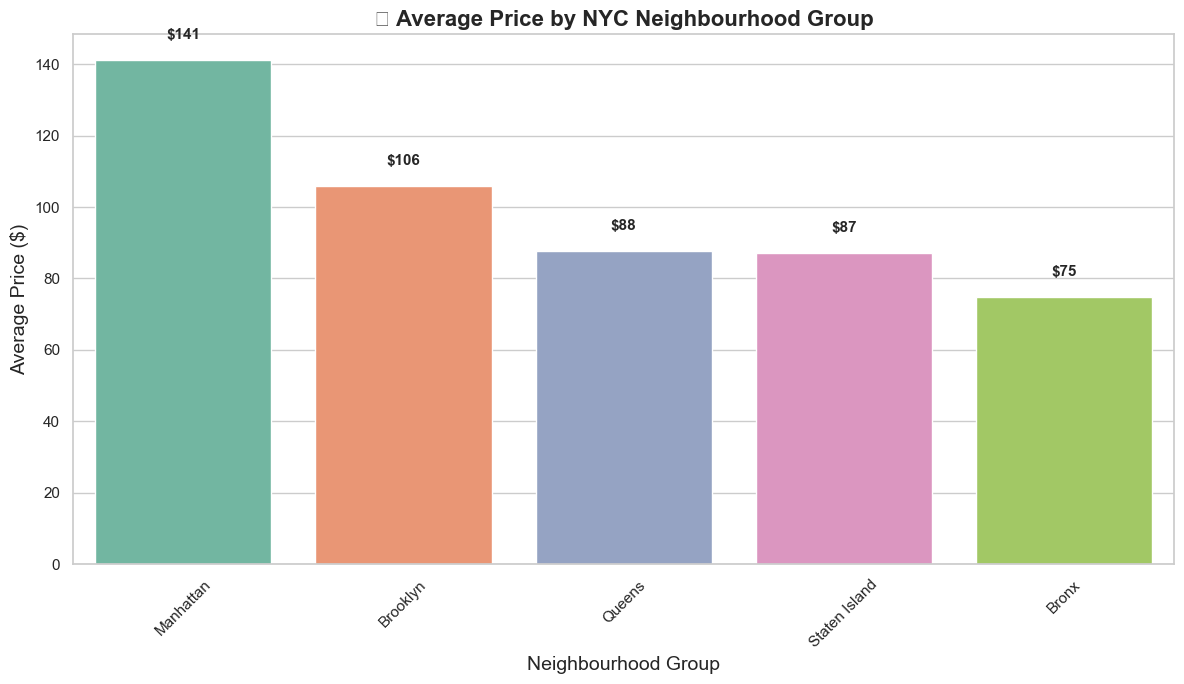
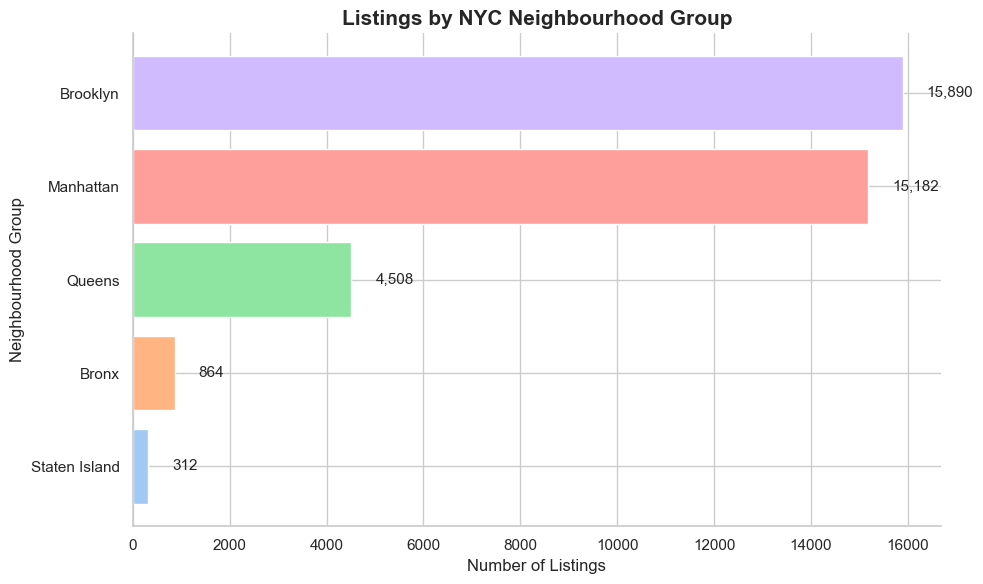


Figure 1

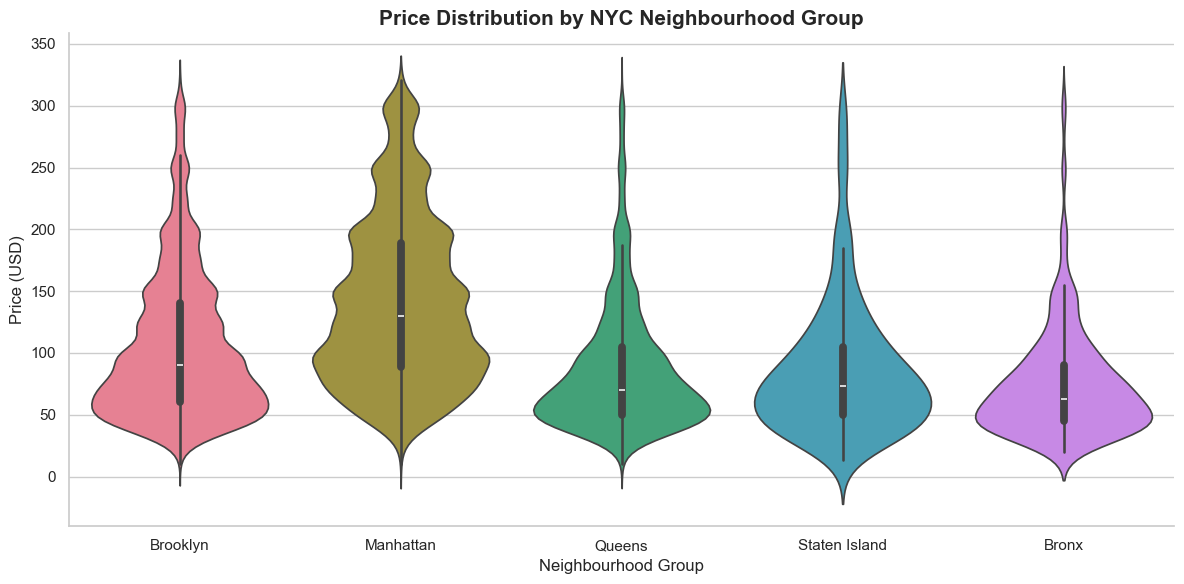


Figure 2

Brooklyn and Manhattan showed the widest variation in listing prices, with Manhattan generally having the highest concentration of expensive listings. Staten Island displayed the broadest spread of prices, including some extreme outliers, while the Bronx and Queens had relatively lower and more compact price distributions. (Figure 2).

The correlation analysis revealed several notable relationships among Airbnb listing variables. Longitude showed a moderate negative correlation with price (r = -0.31), indicating that listings located further east or west may be slightly cheaper, possibly reflecting geographic differences in property value. Reviews per month and number of reviews were strongly positively correlated (r = 0.56), suggesting that frequently reviewed listings consistently attract monthly attention, highlighting host engagement and guest activity.

Availability over the year (availability\_365) showed a small positive correlation with number of reviews (r = 0.20), suggesting that listings available year-round may accumulate more reviews over time.

Price had a weak negative correlation with longitude (r = -0.31), suggesting that location still plays a role in pricing, though other factors like room type or amenities may have stronger effects. Other variables, including latitude, minimum nights, and calculated host listings count, showed no strong correlation with price, implying limited linear association in this dataset (Figure 3).

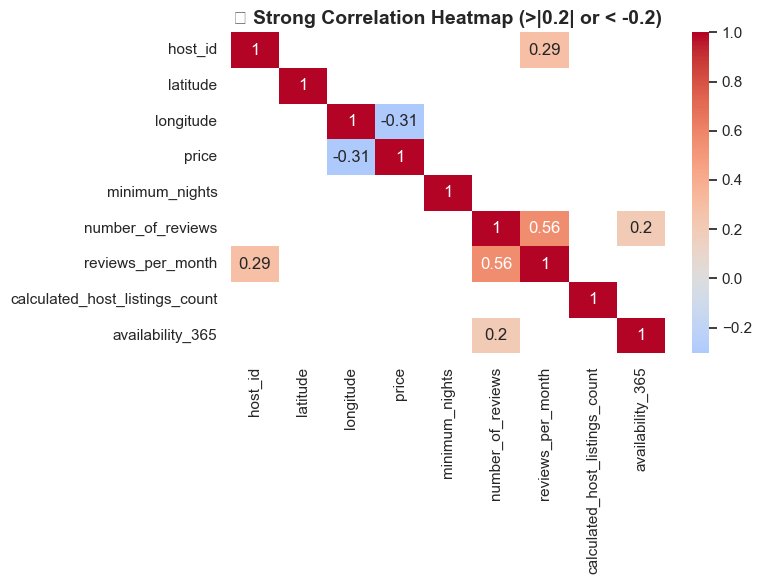


Figure 3

The box plots reveal that **entire homes/apartments** are the most expensive, followed by **private rooms** and **shared rooms**, which are the least expensive. **Review count** has little impact on pricing, and **minimum nights required** shows that listings with longer stays (8-30 nights) tend to be more expensive. Longer-term stays (30+ nights) may offer slightly lower prices.(figure 4)

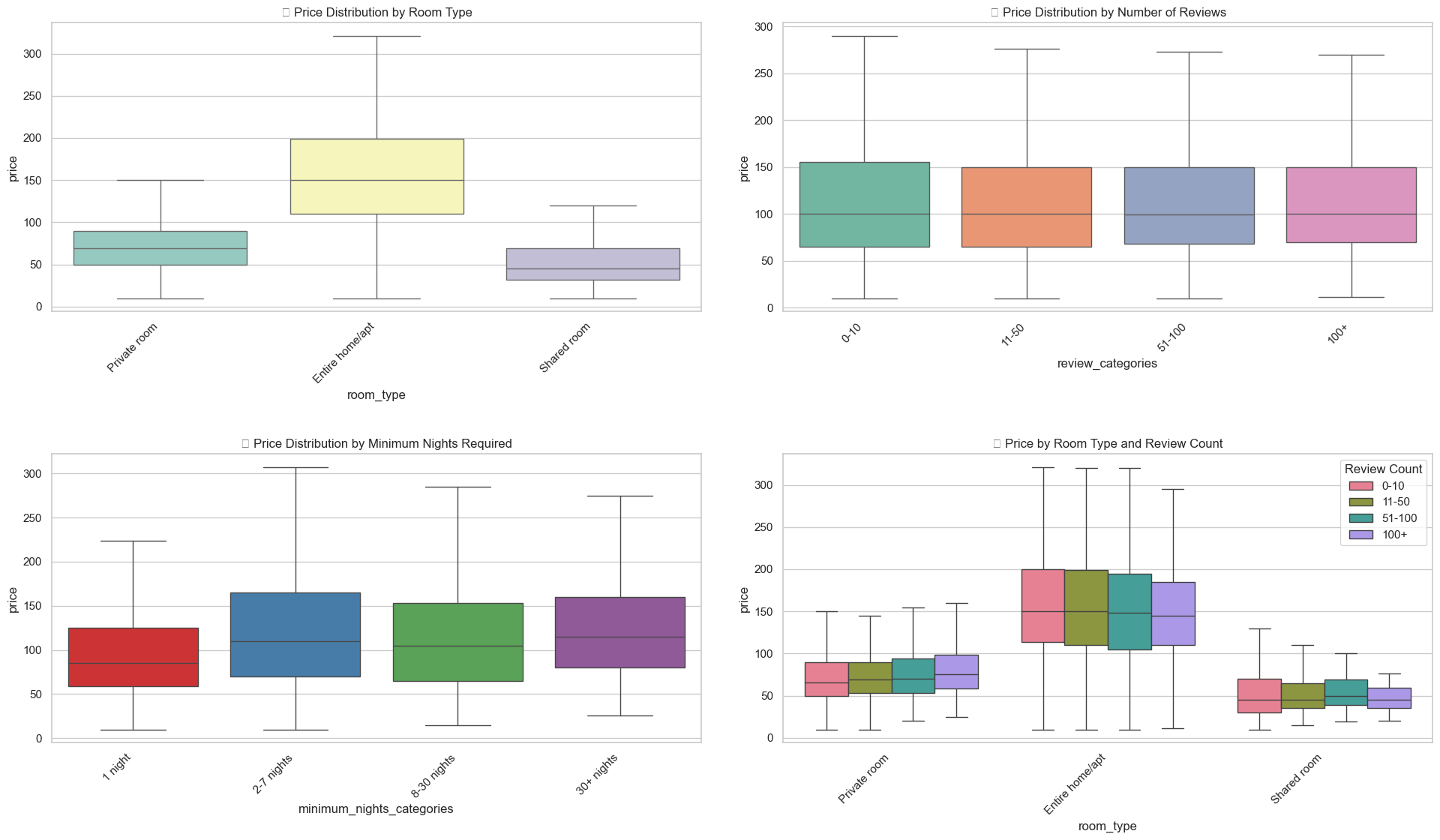


Figure 4

The price distribution reveals a clear concentration of listings priced between $50 and $100, with the highest frequency occurring around the $60-$80 range. This suggests that most options available are affordable, making up the bulk of the dataset. The distribution is right-skewed, meaning that as the price increases, the number of listings decreases. Although most properties are on the lower end of the price spectrum, there is a noticeable tail extending towards higher prices, particularly in the $200-$300 range. These higher-priced listings are less frequent and likely represent more luxurious or larger properties, indicating that while affordable options are more common, there is a smaller, but significant, portion of premium listings available as well.

(figure 5)

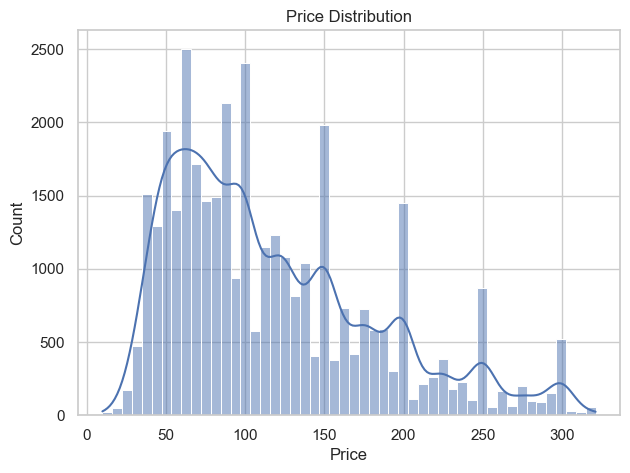


Figure 5

Corellation

Expensive listings tend to have fewer reviews, suggesting they receive less frequent bookings. The weak negative correlation (r ≈ -0.0188, p ≈ 3.1871e-04) indicates that more affordable listings attract more guests and reviews, while high-priced listings cater to a smaller, niche market. This aligns with findings in tourism research, where lower prices typically lead to higher occupancy rates, but higher prices target fewer customers.

The scatter plot supports this, showing a concentration of low-priced listings with more reviews, while high-priced listings are scattered with fewer reviews. Many expensive listings have 0-5 reviews, indicating infrequent bookings. This observation underscores the price-occupancy trade-off, where lower prices drive more bookings, while higher prices attract fewer guests. Price shows minimal correlation with other factors, aside from the slight negative relationship with reviews. (Figure 6)



Figure 6

**Machine Learning Analysis**  
To investigate neighborhood-level trends, we applied K-Means clustering using the average price and average number of reviews per listing. Additionally, we conducted linear regression to assess how different factors—such as neighborhood group, number of reviews, and host listing count—affect prices, aiming to identify which variables have a significant impact (Schroeder et al., 2017).

Liner Regression  
  
The regression analysis table provides valuable insights into factors influencing the pricing of listings, which could be relevant for platforms like Airbnb. The constant term indicates that when all predictors are zero, the baseline price is approximately 119 units.

The **minimum nights** required for booking a listing has a negative impact on price, with each additional night of minimum stay reducing the price by 0.22 units. This suggests that listings with more stringent stay requirements are generally priced lower.

The **number of reviews** also has a slight negative effect on price, with each additional review leading to a decrease of 0.03 units in the listing's price. This could indicate that popular listings, which typically have more reviews, are slightly cheaper, possibly due to their higher availability or greater competition.

**Reviews per month** does not show a statistically significant impact on the price, suggesting that it doesn't have a meaningful effect on the pricing.

The **calculated host listings count** shows a positive effect, with hosts having more listings pricing their offerings 0.11 units higher. This could reflect the increased pricing flexibility or resources of hosts who are more experienced or manage multiple properties.

**Availability** throughout the year has a small but positive impact, with listings available 365 days being priced 0.05 units higher, reflecting a premium on properties that are always available.

Regarding the **room types**, both private rooms and shared rooms are priced lower than entire homes. Private rooms are about 77 units less expensive, while shared rooms are 103 units cheaper, underlining the higher price associated with renting entire homes.

**Neighborhood group** shows the significant impact of location on price. Listings in **Brooklyn** are priced 23 units higher than the baseline area, whereas listings in **Manhattan** are much more expensive, with a price difference of 52 units. On the other hand, **Queens** listings are only slightly more expensive, with a 11-unit increase compared to the baseline.

Overall, the analysis suggests that listing prices are influenced by factors like the minimum nights required, host activity, room type, and location, with Manhattan and Brooklyn commanding the highest price premiums.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Coefficient** | **P-value** | **Interpretation** |
| **const** | 119.55 | 0.000 | When all predictors are 0, the average price is ~119 units. |
| **minimum\_nights** | -0.22 | 0.000 | More minimum nights required → lower price. |
| **number\_of\_reviews** | -0.03 | 0.000 | More reviews → slightly lower price (popular listings tend to be cheaper). |
| **reviews\_per\_month** | +0.03 | 0.859 | Not statistically significant → no meaningful effect on price. |
| **calculated\_host\_listings\_count** | +0.11 | 0.000 | Hosts with more listings tend to charge slightly higher prices. |
| **availability\_365** | +0.05 | 0.000 | Listings available year-round are priced slightly higher. |
| **room\_type\_Private room** | -76.89 | 0.000 | Private rooms are ~77 units cheaper than entire homes. |
| **room\_type\_Shared room** | -102.92 | 0.000 | Shared rooms are ~103 units cheaper than entire homes. |
| **neighbourhood\_group\_Brooklyn** | +23.37 | 0.000 | Listings in Brooklyn are ~23 units more expensive than the baseline area. |
| **neighbourhood\_group\_Manhattan** | +51.89 | 0.000 | Listings in Manhattan are ~52 units more expensive than the baseline area. |
| **neighbourhood\_group\_Queens** | 11.29 | 0 | Listings in Queens are ~11 units more expensive than the baseline area. |

Table 1

K-means

**1. Budget-Friendly Listings (Cluster 0)**

These listings have **lower prices** (50–100 units) and **moderate reviews** (20–50). They typically cater to budget-conscious travelers and may be newer or in less popular areas. The fewer reviews suggest these listings are less established but still offer affordable options.

**2. Standard Listings (Cluster 1)**

With **moderate pricing** (100–150 units) and **higher reviews** (40–70), these listings appeal to a broader audience. They are likely in popular areas, well-established, and offer a reliable balance between cost and quality.

**3. Luxury Listings (Cluster 2)**

These listings are **premium-priced** (150–250 units) with **high review counts** (60+). They target guests seeking high-end experiences and are located in high-demand areas. The strong reputation and higher pricing reflect exceptional quality and service.

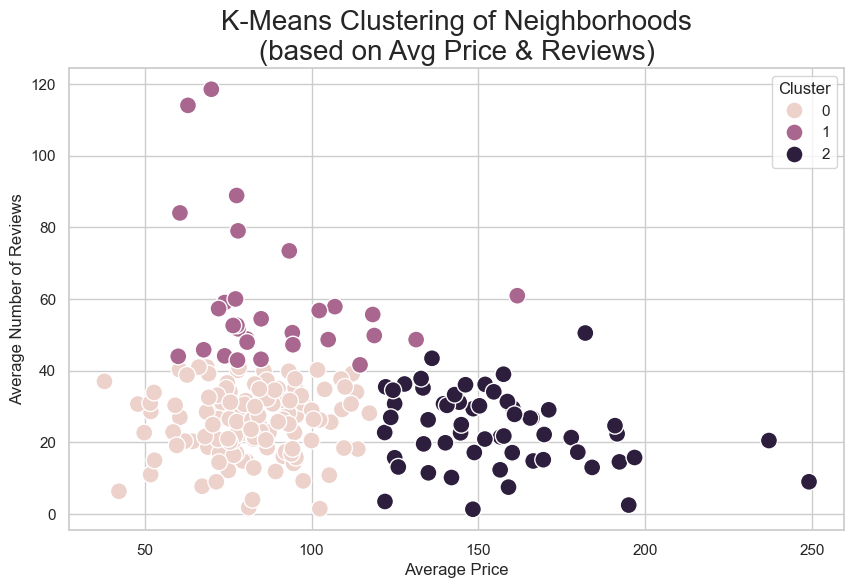


Figure 7

The cluster visualization (Figure 8) reveals three distinct groups: budget-friendly (low-price, moderate reviews), standard (moderate price and reviews), and luxury (high-price, high-reviews). This segmentation provides valuable insights for Airbnb hosts and stakeholders. Hosts in the standard segment should focus on maintaining competitive pricing and high guest satisfaction to stand out, while budget-friendly hosts can prioritize increasing occupancy and generating positive reviews. Luxury hosts should highlight unique experiences and high-end amenities to justify premium pricing. Airbnb can leverage these insights to optimize marketing strategies, targeting budget travelers with special offers and luxury segments through tailored, exclusive channels.

**3. Findings and Recommendations**

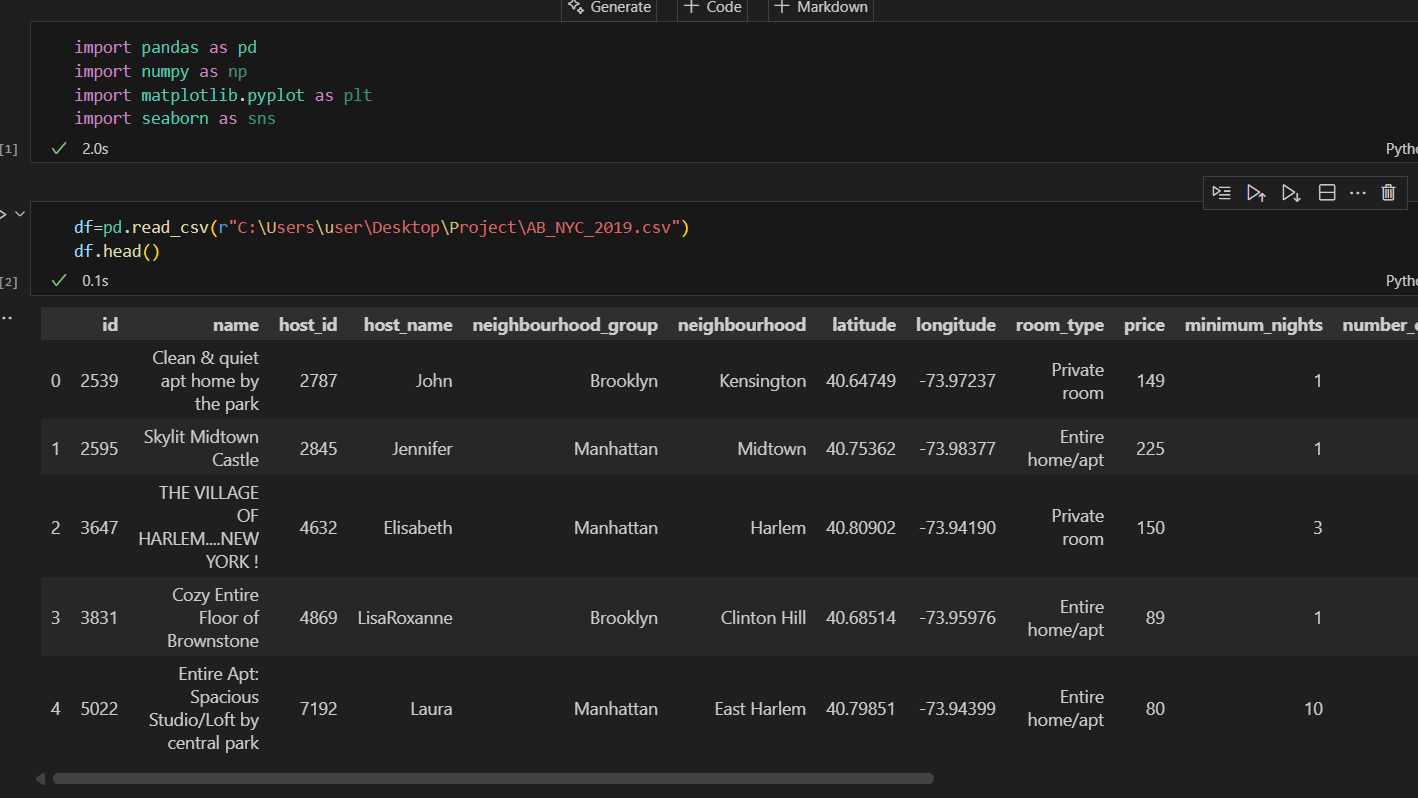
Airbnb has substantial growth opportunities in New York City. The analysis of the dataset highlights strong demand around key landmarks, presenting an opportunity to maximize profits by focusing on accommodations in these prime locations. Furthermore, a strategic approach involving a combination of upscale properties in high-demand areas and budget-friendly rooms in less sought-after neighborhoods can help Airbnb stay competitive by utilizing dynamic pricing strategies. By closely monitoring location-specific trends and local events, Airbnb can create targeted promotional campaigns that align with the unique characteristics of different neighborhoods and attractions, enhancing its market positioning and overall customer engagement.

**4. Conclusion**

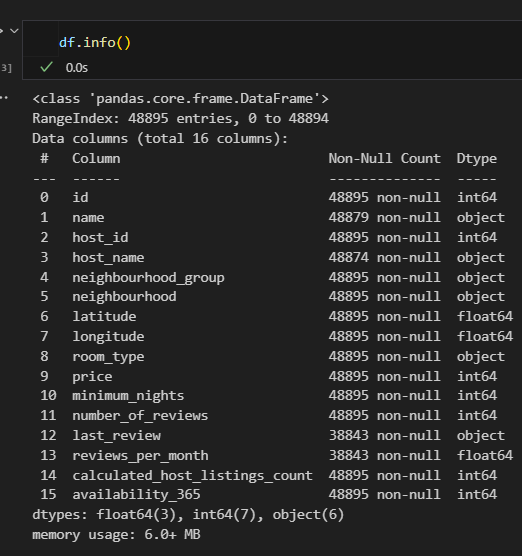
This report highlights crucial business opportunities for Airbnb in New York City, emphasizing targeted marketing tactics and strategies for maximizing revenue. By concentrating on specific neighborhoods and adapting offerings accordingly, Airbnb can improve customer satisfaction and drive increased earnings. Additionally, the report showcases how data analysis and machine learning can be leveraged for informed decision-making, ultimately contributing to Airbnb’s strategic growth and long-term success.

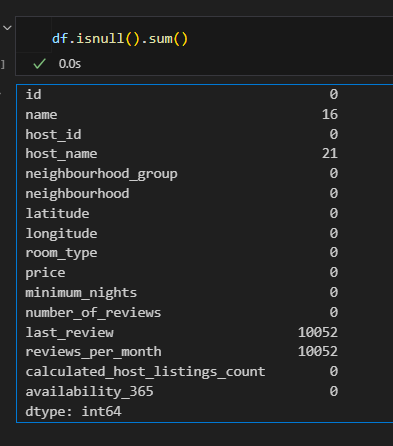
Appendix

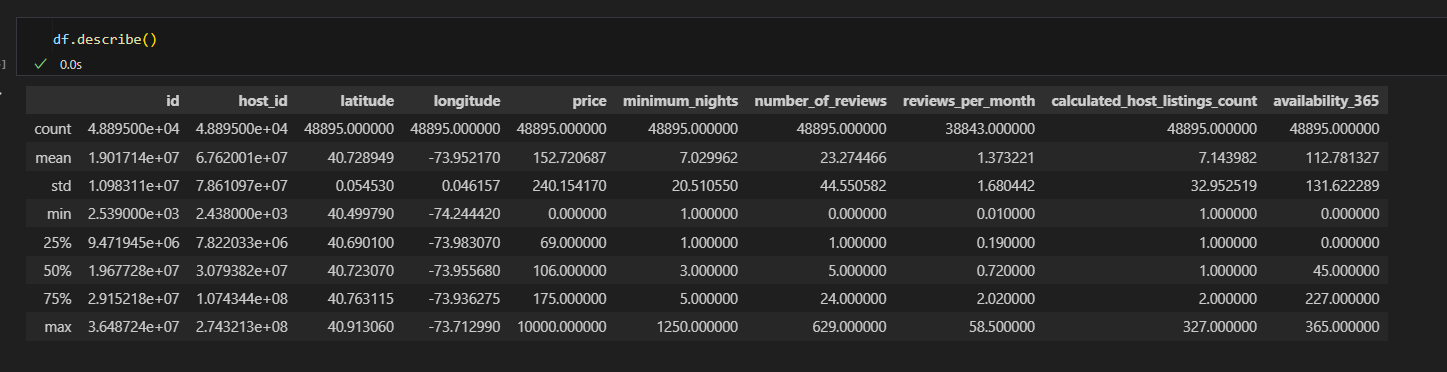
importing



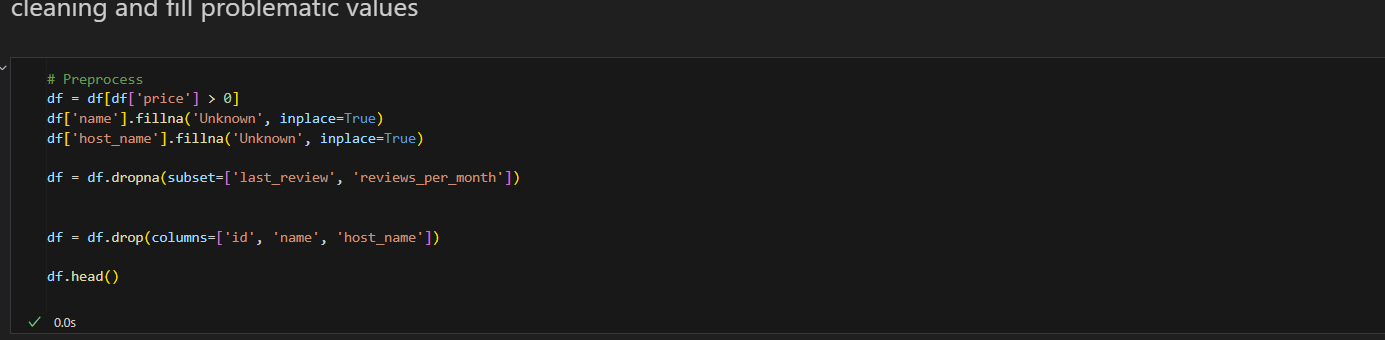
Preprocessing

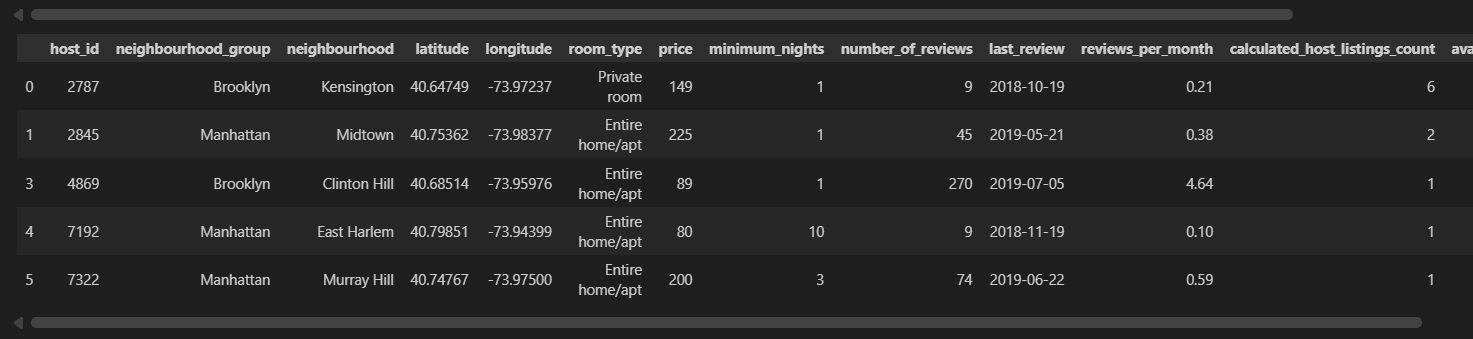




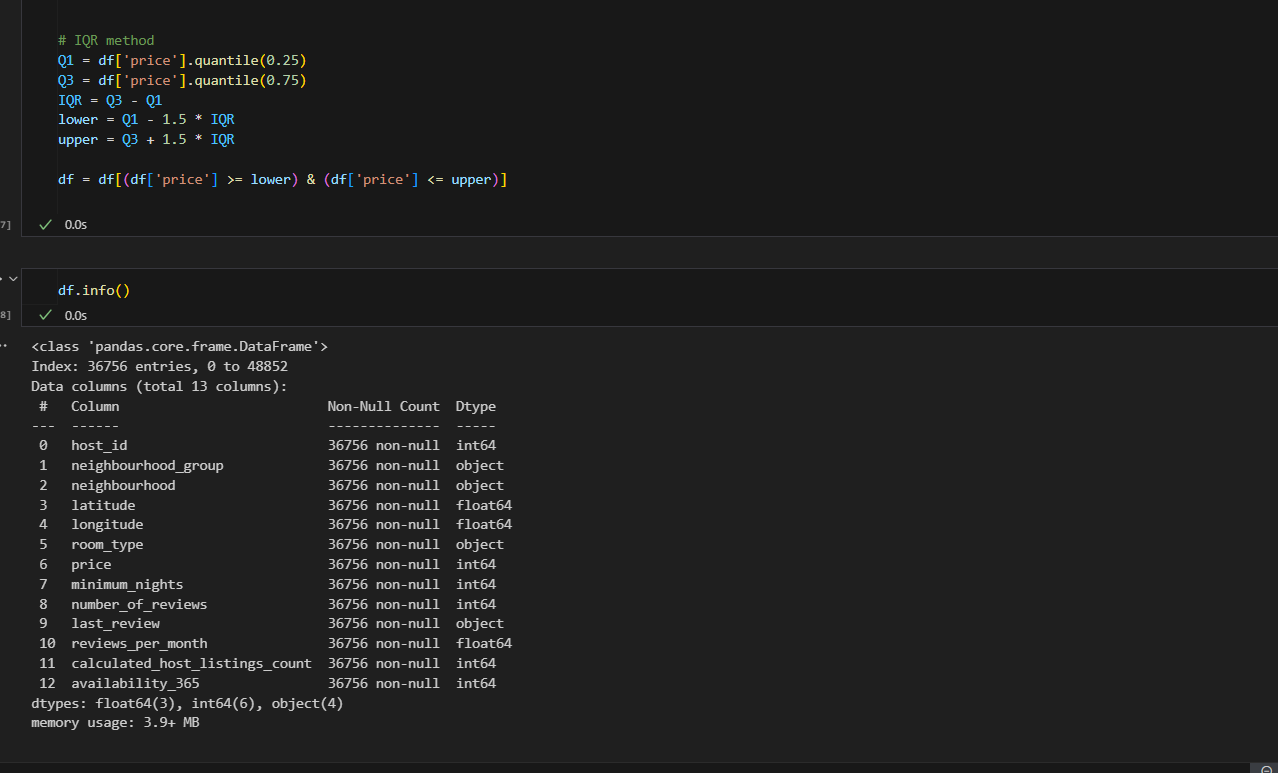


Cleaning



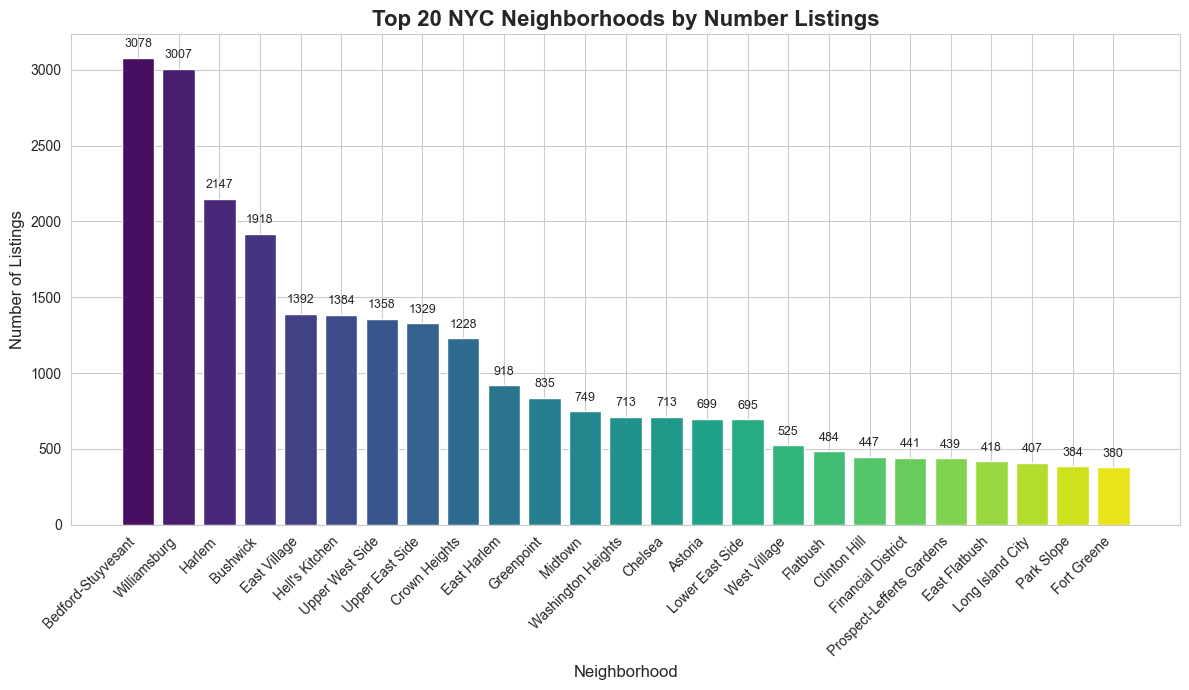


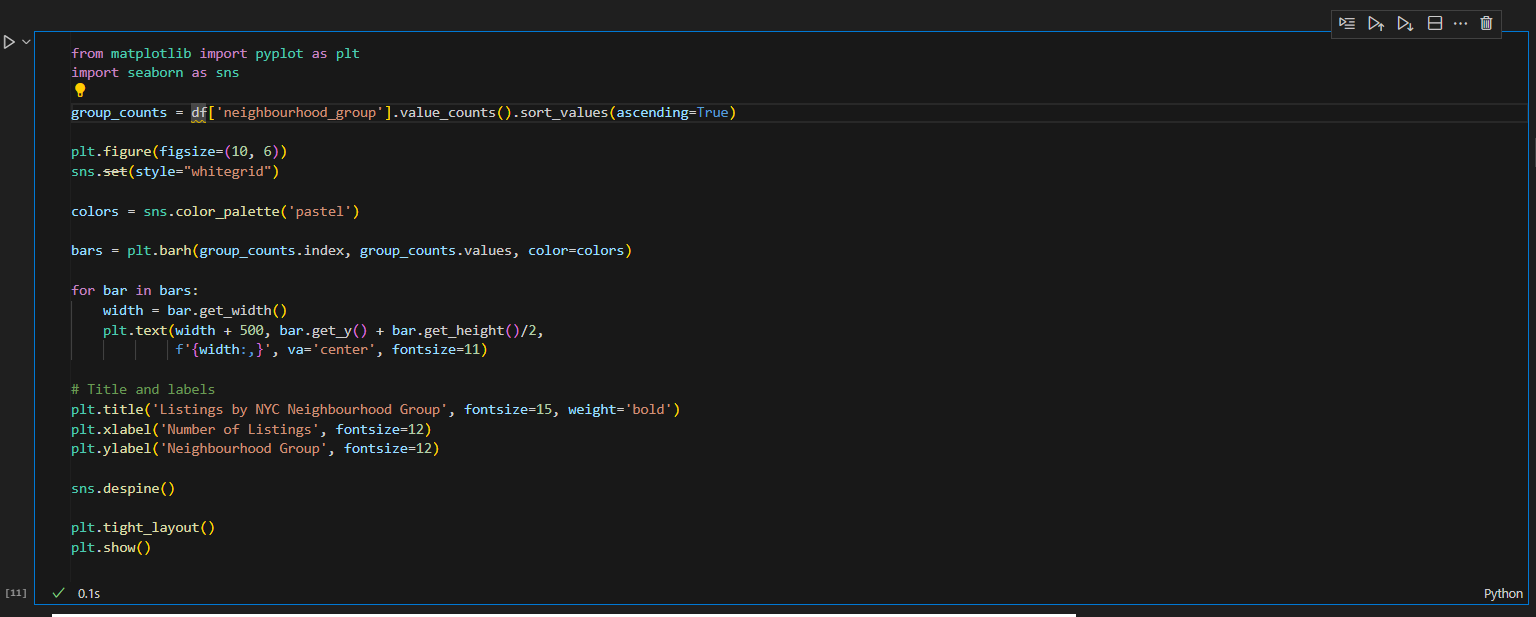
Detect and remove oulier

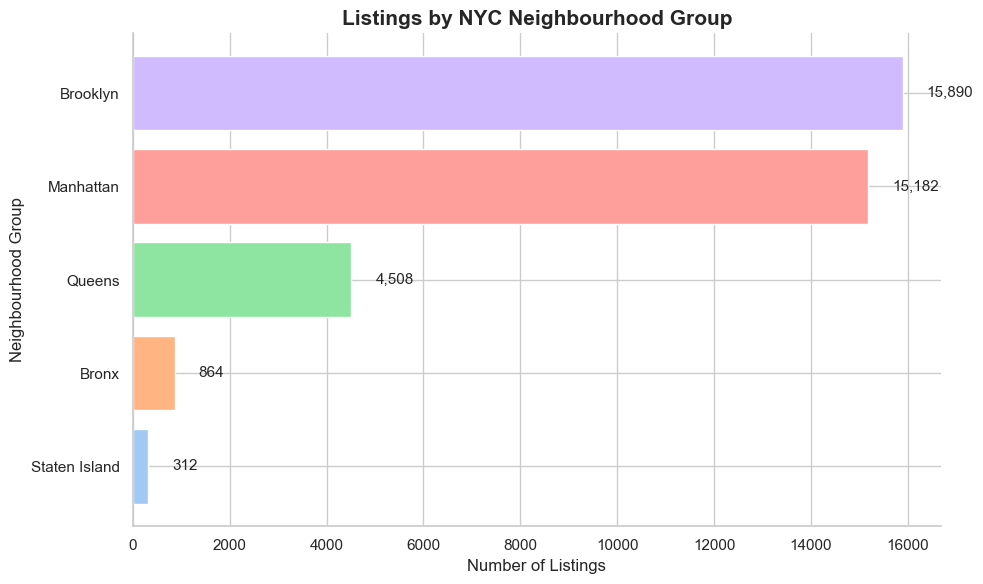


EDA



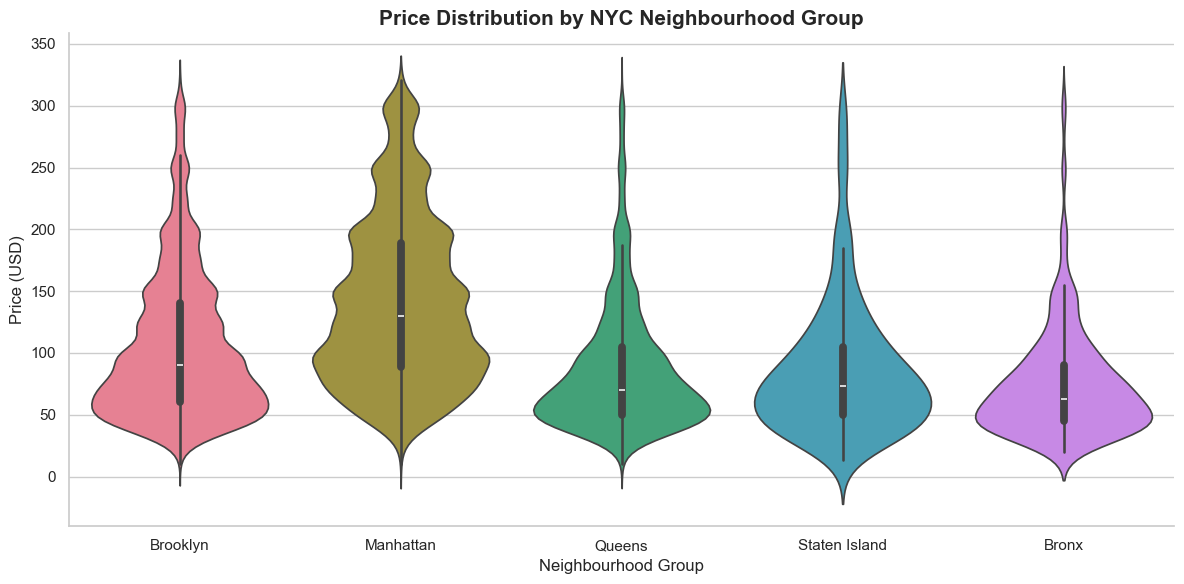


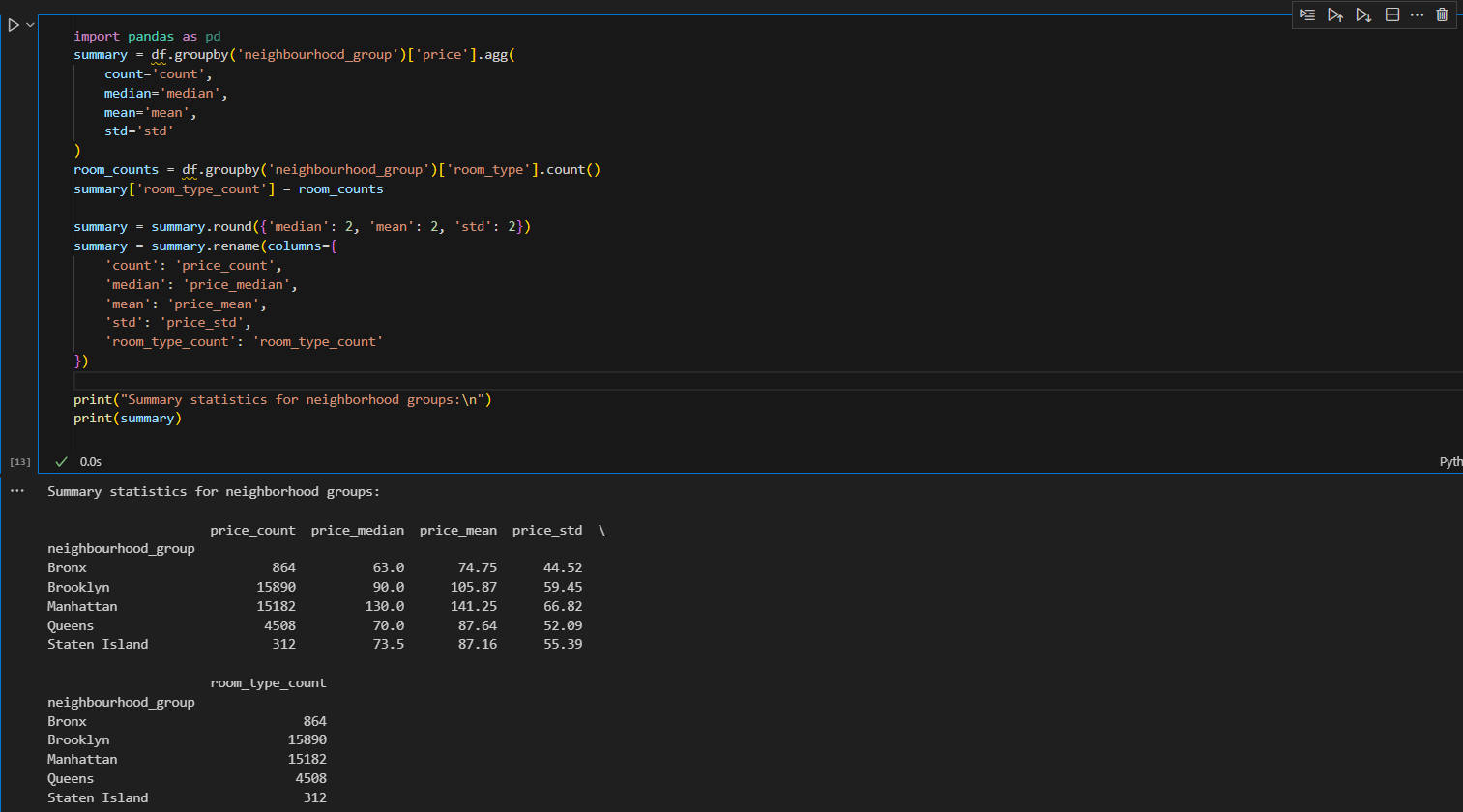




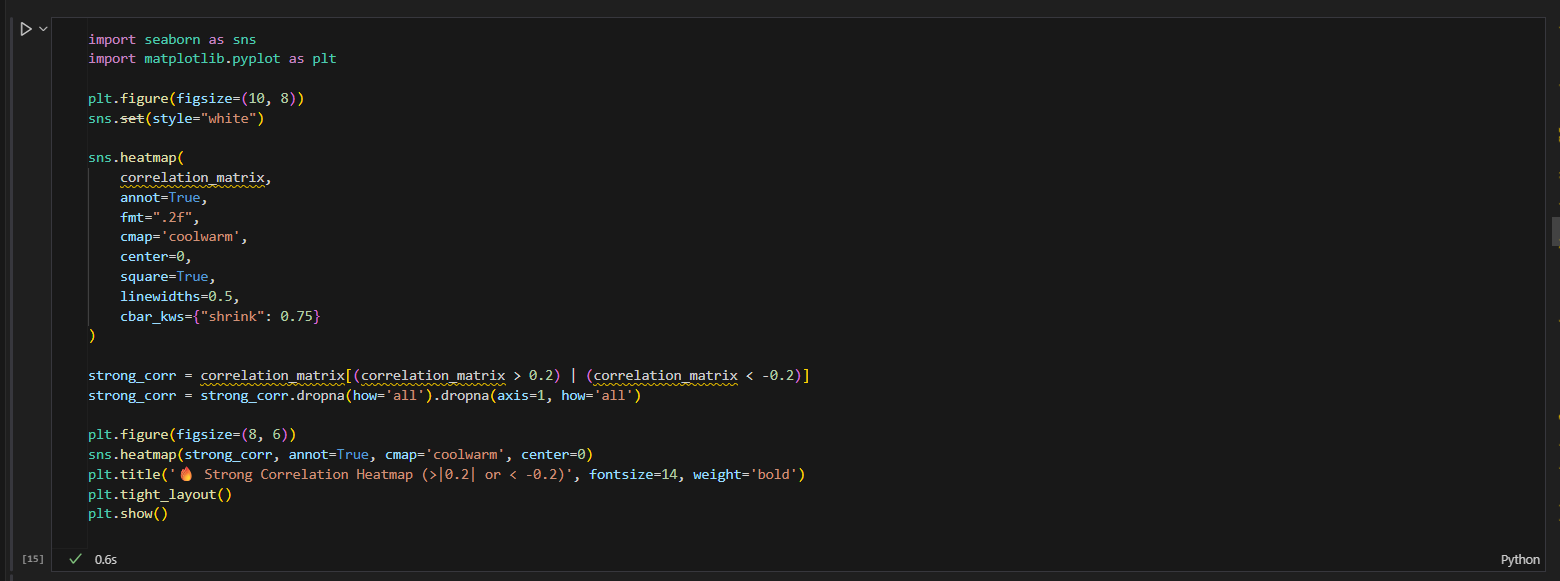
Violit chart

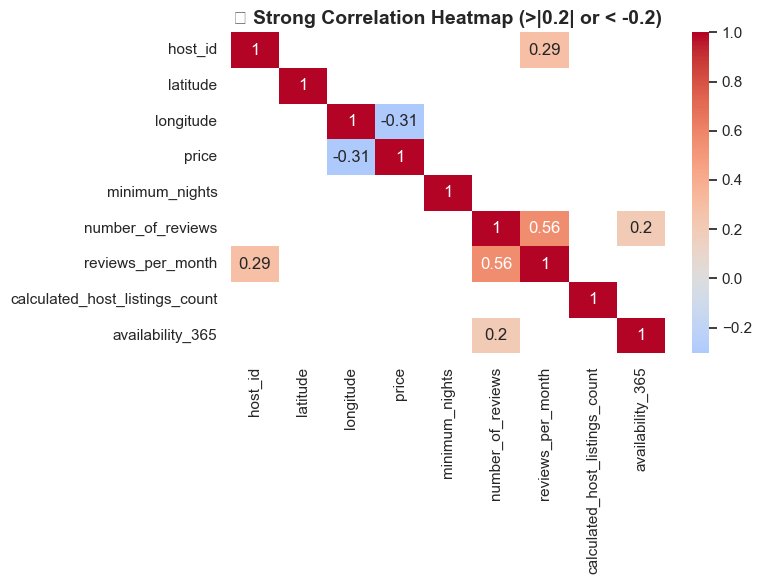
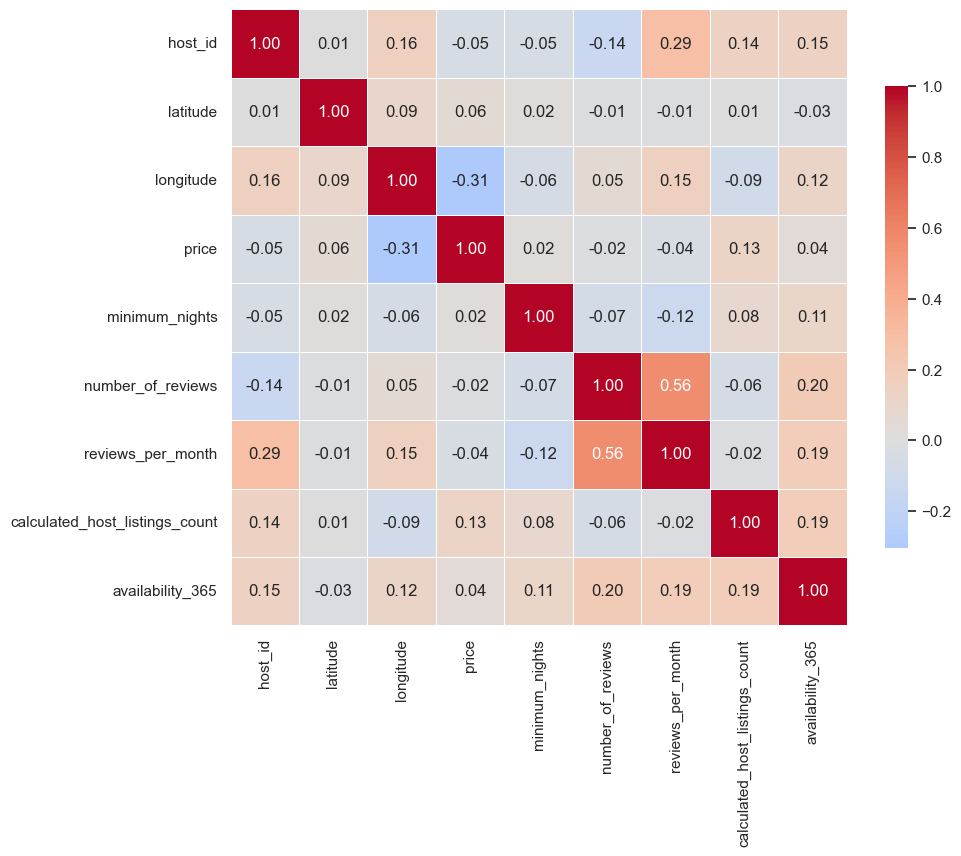


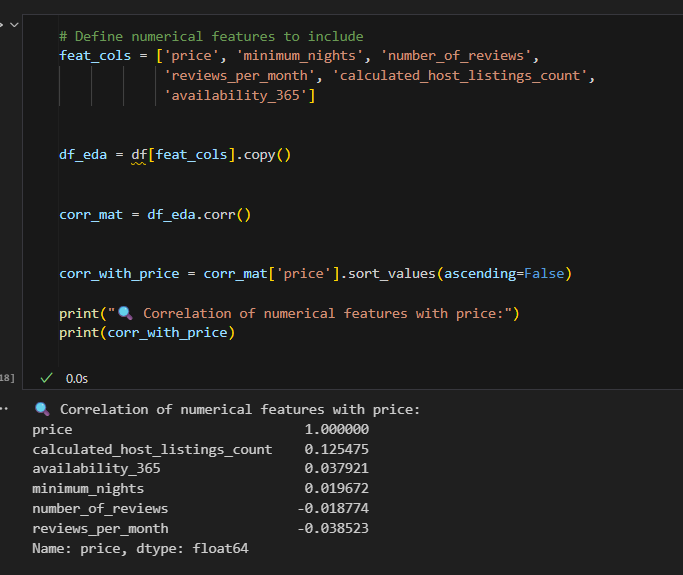




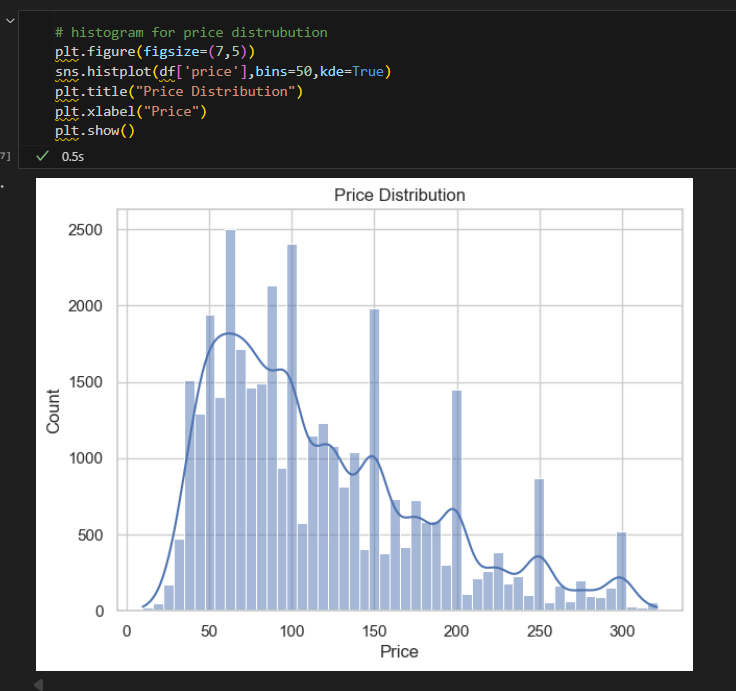
Correlation matrix



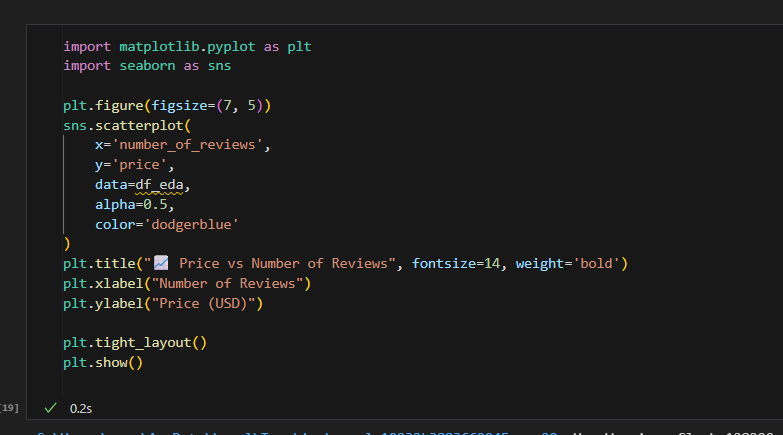




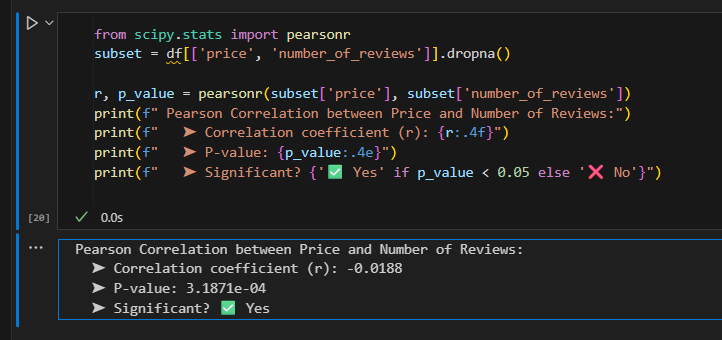
Histogram



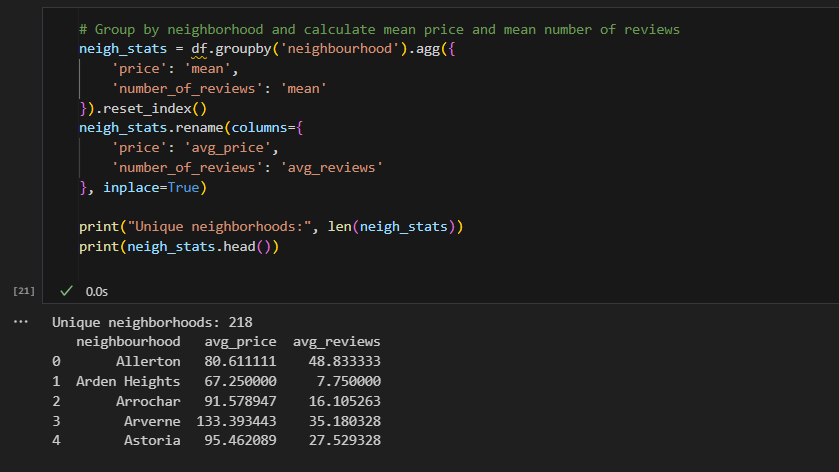
Pearson correlation

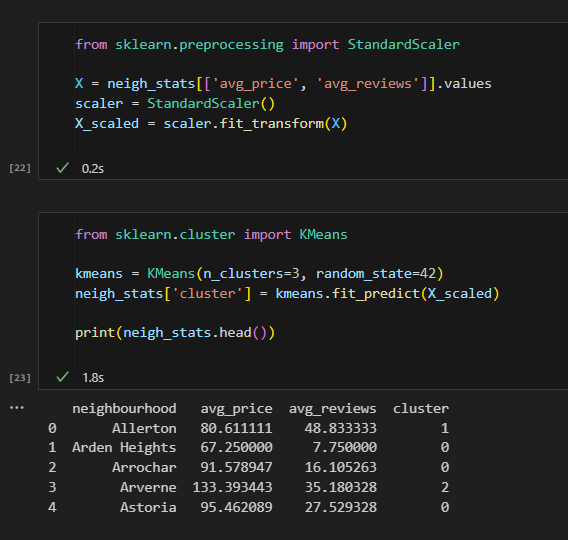


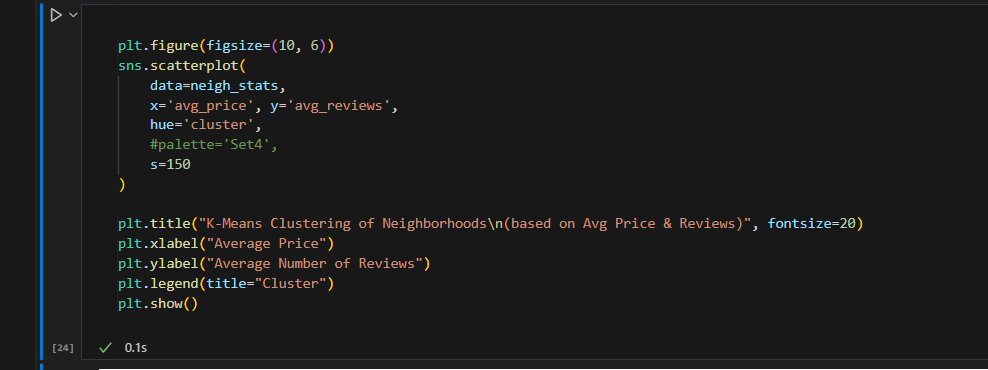


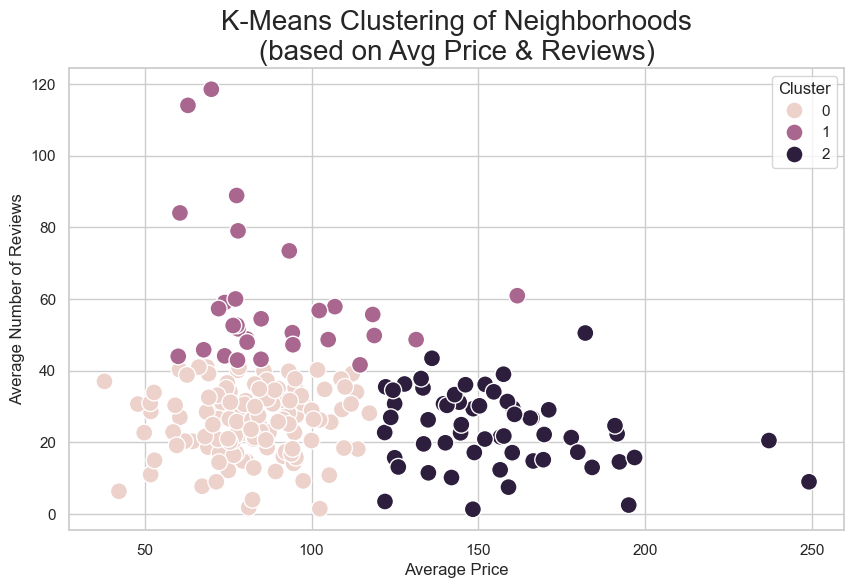


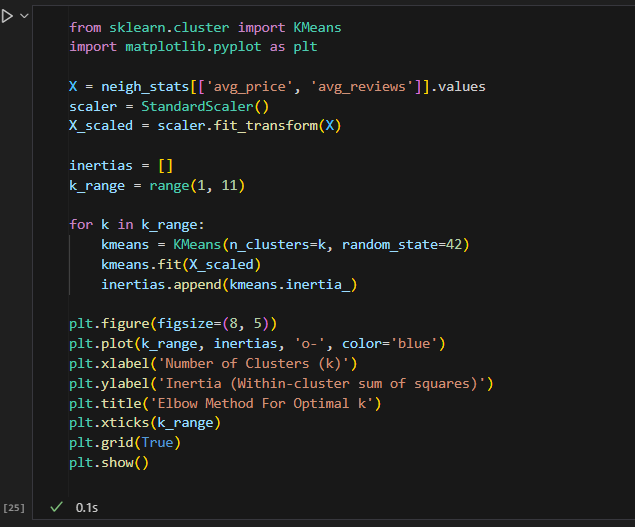
Grouping

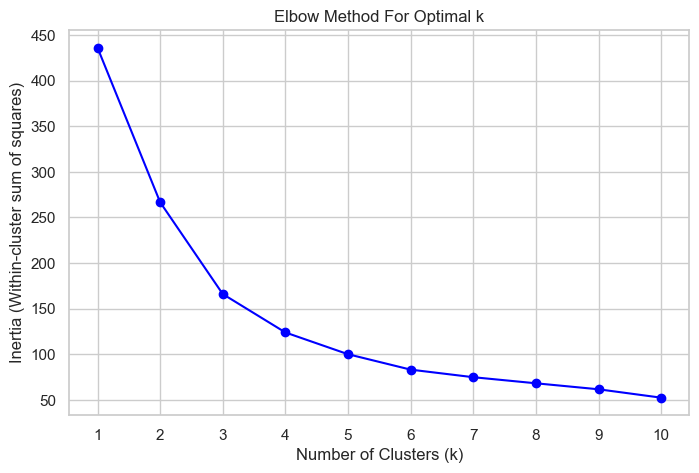


Clustering with K-means:  
  




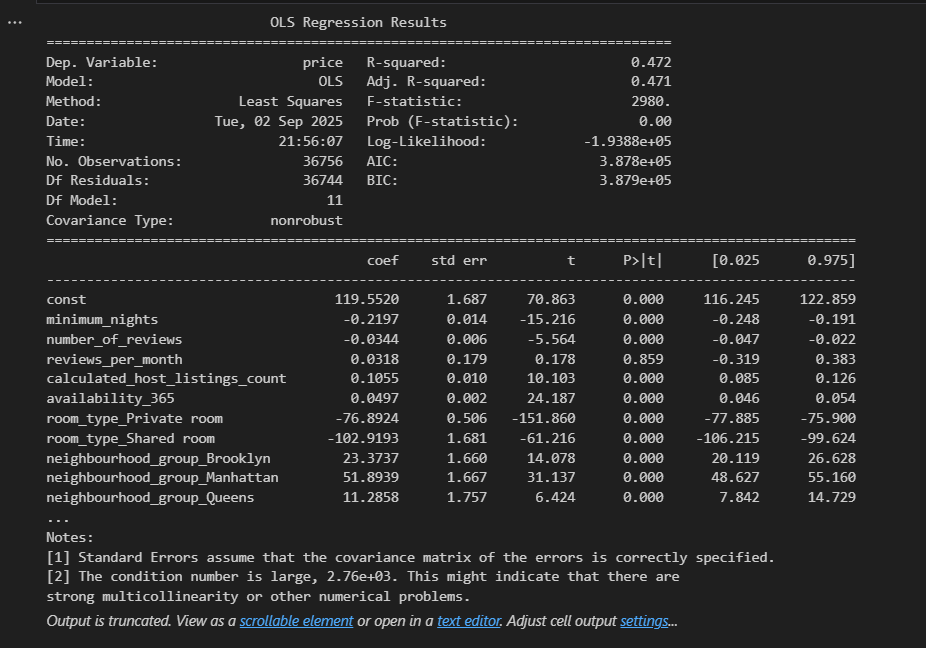


Controlling Optimal clustering number with elbow method:  




Linear regression





Reference

1. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.
2. **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).
3. **Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (2017).** *Understanding Regression Analysis: An Introductory Guide* (2nd ed.). Los Angeles, CA: SAGE Publications, Inc.
4. **Tufte, E. R. (2001).** *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press.
5. **Field, A. P. (2024).** *Discovering Statistics Using IBM SPSS Statistics* (6th ed.). Los Angeles: Sage.
6. **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).
7. **Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013).** *Bayesian Data Analysis* (3rd ed.). Boca Raton: CRC Press.
8. **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>
9. **Bishop, C. M. (2006).** *Pattern Recognition and Machine Learning.* New York: Springer.
10. **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.
11. **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>
12. **Provost, F., & Fawcett, T. (2013).** *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* Sebastopol, CA: O'Reilly Media.
13. **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>
14. **Chawla, N. V., & Bowyer, K. W. (2002).** *Data Mining for Imbalanced Datasets: An Overview.* In *Proceedings of the International Conference on Data Mining*. IEEE.
15. **Hocking, R. R. (2013).** *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley-Interscience.
16. **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).
17. **Wedel, M., & Kamakura, W.A. (2000).** *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). Boston: Kluwer Academic Publishers.
18. **McKinney, W. (2018).** *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). Sebastopol, CA: O'Reilly Media.
19. **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).
20. **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>
21. **Tufte, E. R. (2001).** *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press.
22. **Hinton, G. E. (2012).** *A Practical Guide to Neural Networks*. Cambridge: MIT Press.
23. **Chatterjee, S., & Hadi, A. S. (2006).** *Regression Analysis by Example* (4th ed.). Hoboken, NJ: Wiley-Interscience.
24. **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.