# Analyzing NYC Airbnb Open Data

Data Science 2024 Bootcamp →

Shihui Feng, Dong Li, Sofia Celorio , Antong Lei, KaiYuan Ma, Tujie Guo

# Table of Contents



# 01

# Introduction

# Overview of the Project




The main idea of this project is to use the **NYC Airbnb Open Data** to conduct a comprehensive **analysis** that will yield actionable insights for both hosts and potential renters.

By delving into Airbnb's extensive dataset, we hope to uncover patterns, trends, and factors **influencing rental prices** and **demand** in New York City's dynamic accommodation market.

# Company Background

- Online platform that allows people to **rent out** their **homes** or **spare rooms** to guests looking for **short-term lodging**.
- Founded in **2008**, has rapidly transformed the hospitality industry by offering an alternative to traditional hotels.
- The platform connects **hosts** and **guests**, providing a variety of accommodation options such as apartments, houses, and unique stays.

***Importance of Airbnb in NYC:***

- Economic impact
- Flexibility and affordability
- Accommodation diversity
- Neighborhood revitalization
  - Encouraging tourism beyond traditional tourist areas, benefiting local businesses and communities.
- Cultural exchange

# 02

# Goal & Hypothesis

# Current Goal

**Goal**

Predicting rental prices based on relevant features.

Learn how to use area (locations) to predict apartment pricings

# Hypothesis

**The size of the location will increase with season –> Affecting Price**

**Price of Location –> Depends on Availability**

**More Review –> Less Availability**

1. **Holiday** season rents will be **more expensive** due to high demand
2. The **larger** the apts size, the more **pricey** it is
3. On average **Manhattan** apts price will be **more expensive** than other boroughs
4. The **central** the neighborhood, the **higher** the apts price
5. The neighborhood with **better public transportations** have a **higher price**
6. The more **reviews** the apts has, the **less availability** apt will be
7. The **cheaper** the apt is for the **location**, the **less availability** apt has
8. The more **listings**(calculated_host_listings_count) the host has, the **less availability** it would be (due to experience)
9. The **longer characters** the airbnb name has, the **more availability** there is
10. **Entire home/apt renting** has **less availability** than **private room**

# 03

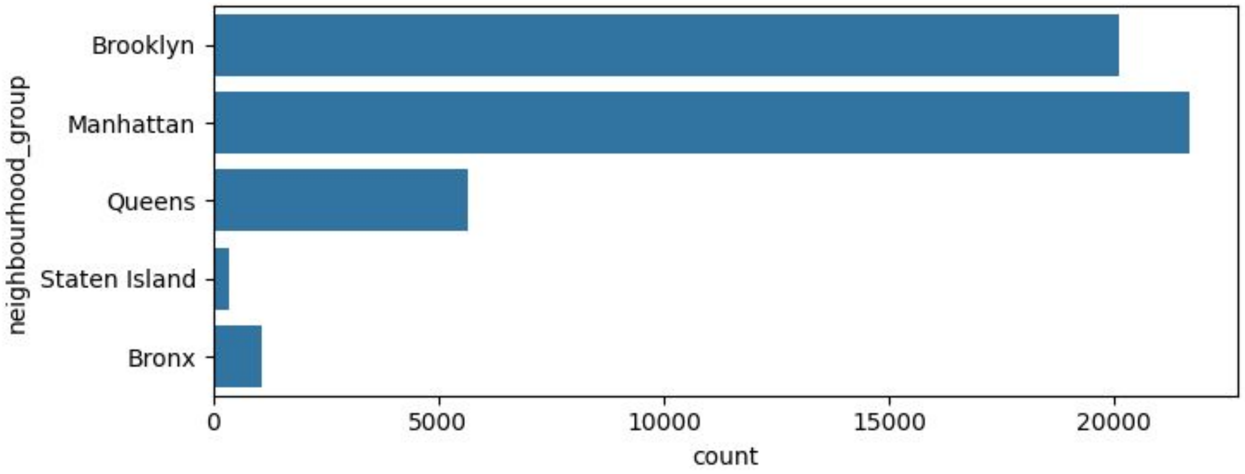
# Our Process

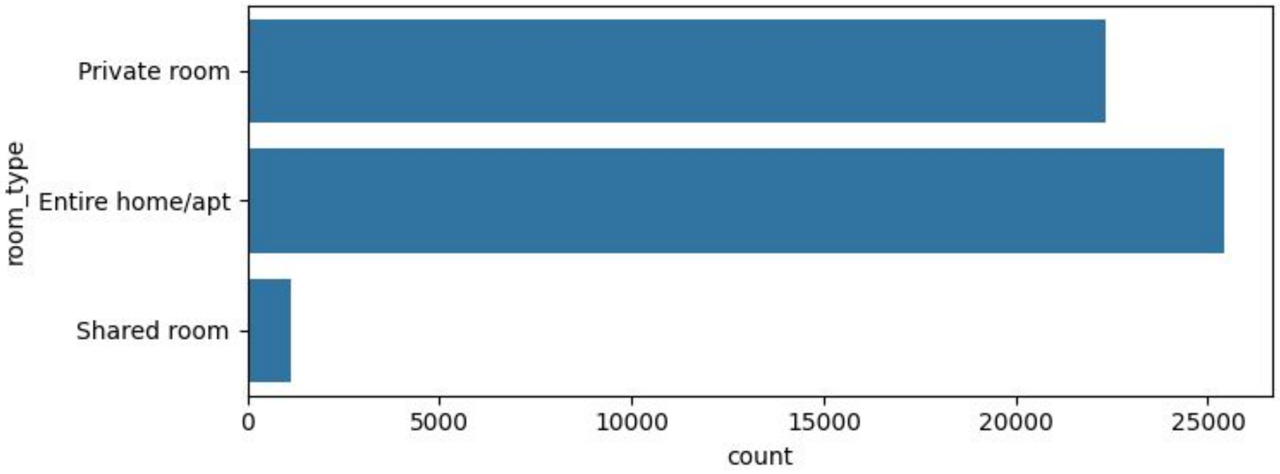
```
#data informtion
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

```
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/AB_NYC_2019.csv')
df.head()
```

| | id | name | host_id | host_name | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2539 | Clean & quiet apt home by the park | 2787 | John | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 | |
| **1** | 2595 | Skylit Midtown Castle | 2845 | Jennifer | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 | |
| **2** | 3647 | THE VILLAGE OF HARLEM....NEW YORK ! | 4632 | Elisabeth | Manhattan | Harlem | 40.80902 | -73.94190 | Private room | 150 | |
| **3** | 3831 | Cozy Entire Floor of Brownstone | 4869 | LisaRoxanne | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | Entire home/apt | 89 | |
| **4** | 5022 | Entire Apt: Spacious Studio/Loft by central park | 7192 | Laura | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | 80 | |



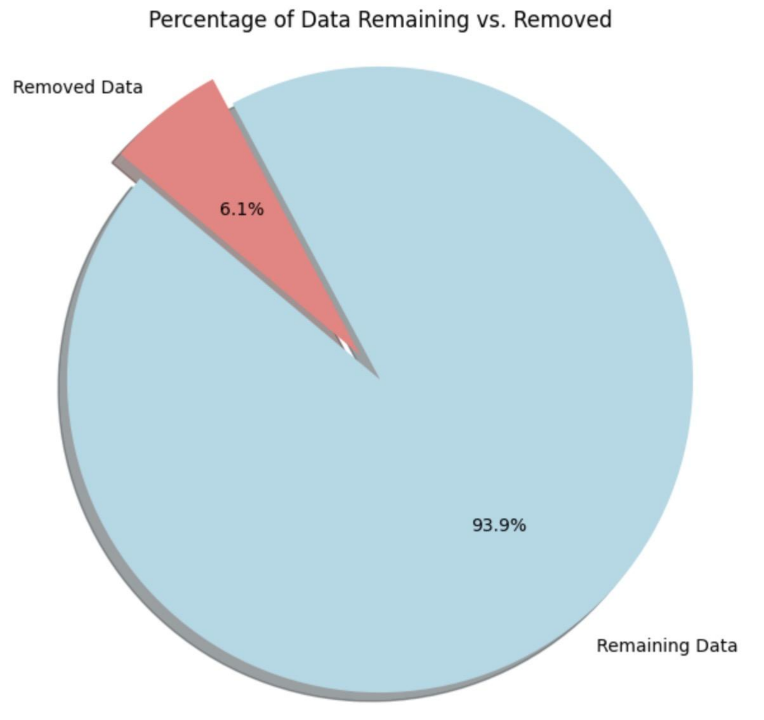
# Overview of NYC Airbnb Dataset (Data Description)

- **Data Information**
  - Acquired the NYC Airbnb dataset containing 48,895 entries and 16 columns.
  - Checked the shape and info of the DataFrame to understand its structure and dimensions.
  - Data begin at **2019**
- **Delete Missing Data**:
  - Identified missing data by generating a boolean DataFrame.
  - Summarized missing data by column.
  - Found rows with missing data and specifically targeted columns ('name' and 'host_name').
    - Replaced missing values:
    - Filled missing values in 'name' and 'host_name' with 'unknown'.
    - Imputed missing values in 'reviews_per_month' with 0.
    - Forward-filled missing values in 'last_review' column.

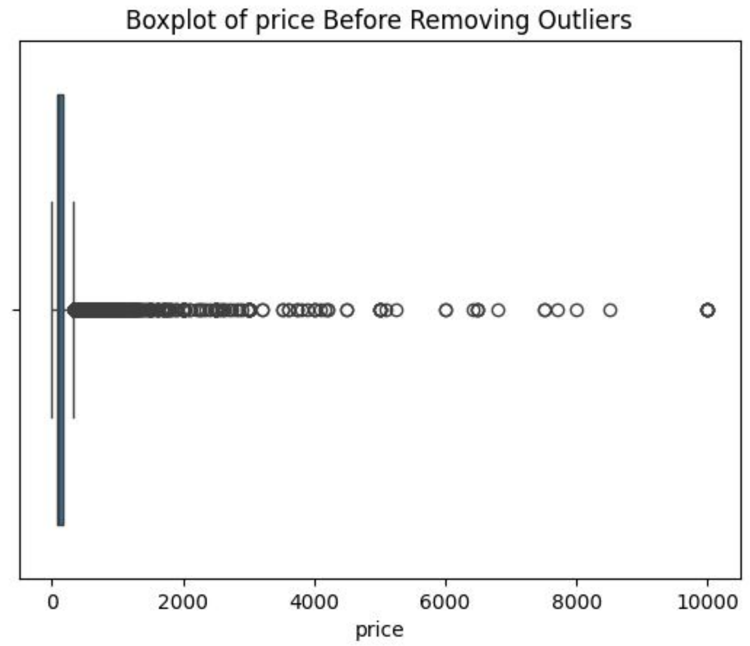
```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

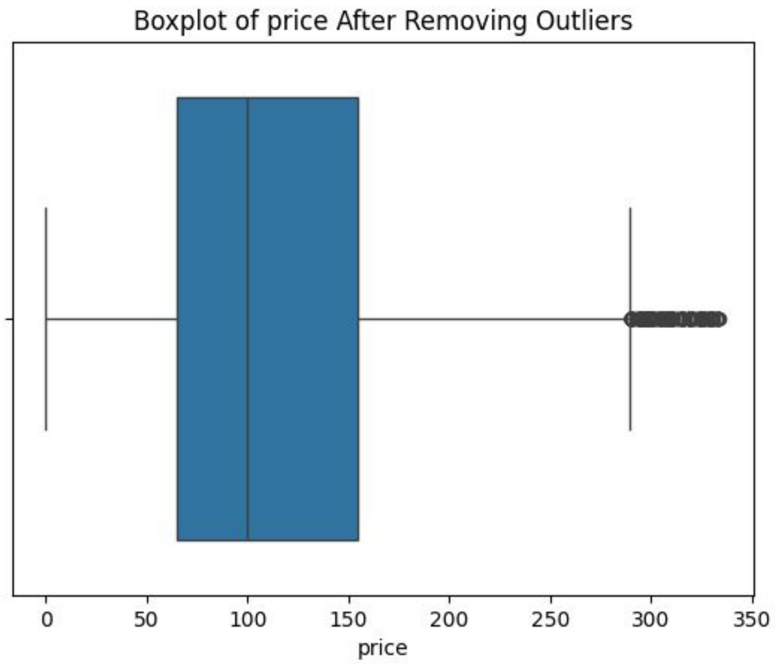
```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

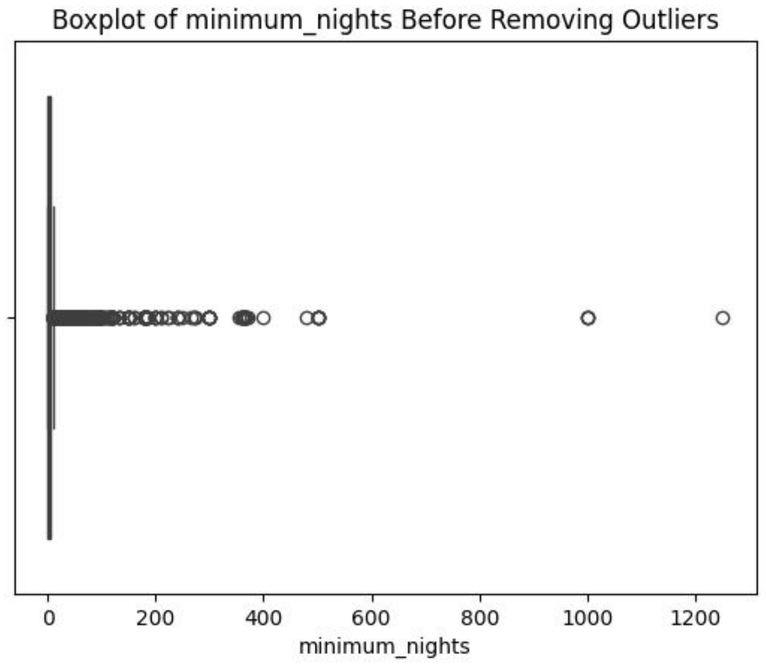
# Distribution of features and identification of patterns

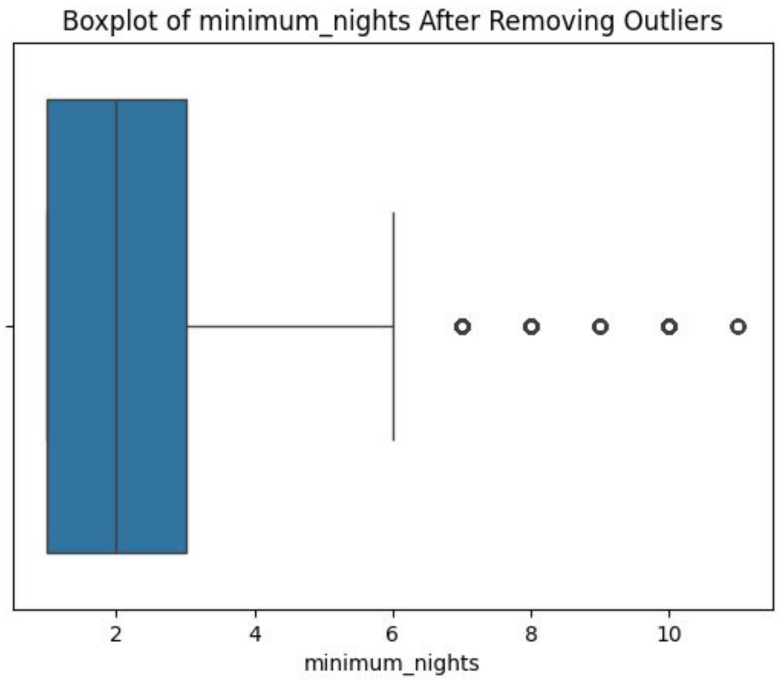



- **Removal**:
  - Detected outliers in the 'price' column using the Interquartile Range (IQR) method.
  - Calculated lower and upper bounds for outlier detection.
  - Removed outliers falling outside the defined bounds.
  - Compared the original data size (48,895 rows) with the size after outlier removal.
- **Visualization**:
  - Plotted a pie chart to illustrate the percentage of data remaining after outlier removal compared to the removed data.
  - Remaining Data- 93.9%
  - Removed Data - 6.1 %


Boxplot of price Before Removing Outliers
0
2000
4000
6000
8000
10000
price


Boxplot of price After Removing Outliers
0
50
100
150
200
250
300
350
price


Boxplot of minimum_nights Before Removing Outliers
0
200
400
600
800
1000
1200
minimum_nights


Boxplot of minimum_nights After Removing Outliers
2
4
6
8
10
minimum_nights

# 04

# Data Visualization

# Find the Median price for each neighborhood

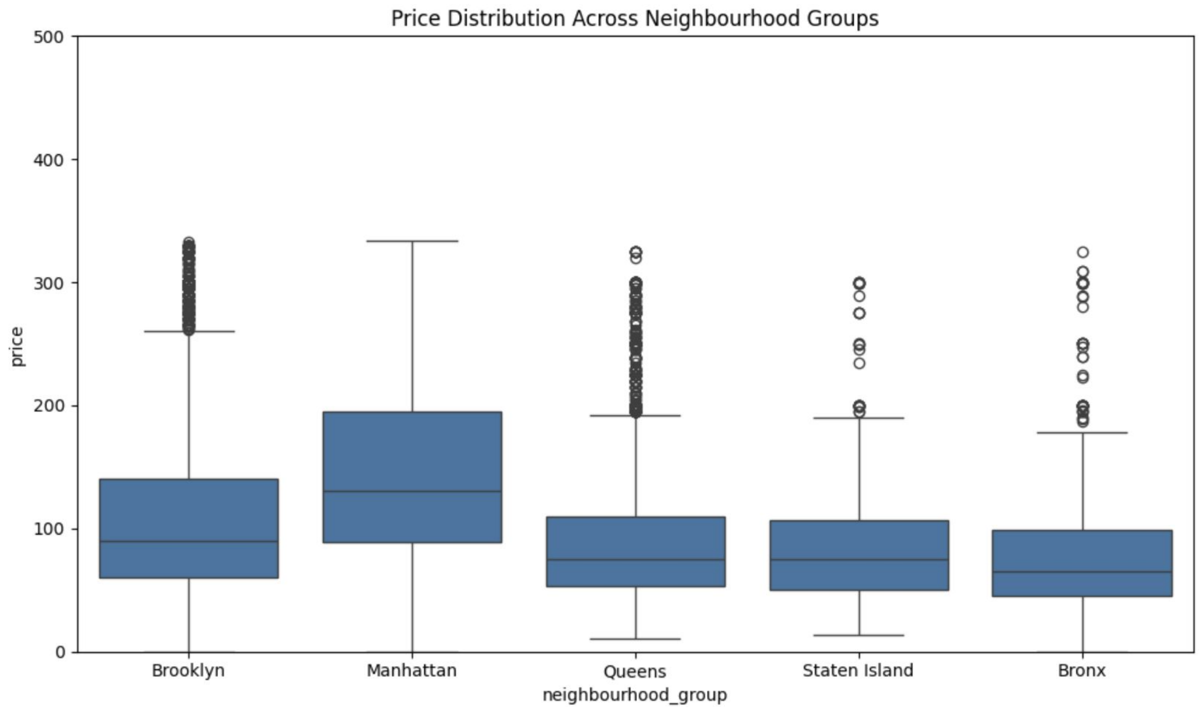
**Median Price Analysis**:

a. Get the median price per neighborhood group for every column.
b. Utilized **groupby** to calculate and display median prices per neighborhood group, both with and without outliers.

| neighbourhood_group | mean | min | max |
|---|---|---|---|
| Bronx | 87.496792 | 0 | 2500 |
| Brooklyn | 124.383207 | 0 | 10000 |
| Manhattan | 196.875814 | 0 | 10000 |
| Queens | 99.517649 | 10 | 10000 |
| Staten Island | 114.812332 | 13 | 5000 |

| neighbourhood_group | mean | min | max |
|---|---|---|---|
| Bronx | 77.365421 | 0 | 325 |
| Brooklyn | 105.699614 | 0 | 333 |
| Manhattan | 145.952835 | 0 | 334 |
| Queens | 88.904437 | 10 | 325 |
| Staten Island | 89.235616 | 13 | 300 |

# Groupby Neighborhood box plot




- Price distribution across neighborhood groups using box plots.
- Displayed the number of listings in each neighborhood group using count plots.
- Plotted bar graphs showing the median price in each neighborhood group.
- correlation relationships between 'price', 'reviews_per_month', 'calculated_host_listings_count', 'availability_365', and 'minimum_nights'.

```
df_cleaned = df.dropna()
correlation_matrix_all_data = df_cleaned[['price', 'reviews_per_month', 'availability_365','minimum_nights','calculated_host_listings_count']].corr()
correlation_matrix_all_data
```

| | price | reviews_per_month | availability_365 | minimum_nights | calculated_host_listings_count |
|---|---|---|---|---|---|
| **price** | 1.000000 | -0.043289 | 0.027122 | 0.059901 | 0.088708 |
| **reviews_per_month** | -0.043289 | 1.000000 | 0.254287 | -0.231585 | 0.036455 |
| **availability_365** | 0.027122 | 0.254287 | 1.000000 | -0.097462 | 0.129584 |
| **minimum_nights** | 0.059901 | -0.231585 | -0.097462 | 1.000000 | -0.031173 |
| **calculated_host_listings_count** | 0.088708 | 0.036455 | 0.129584 | -0.031173 | 1.000000 |

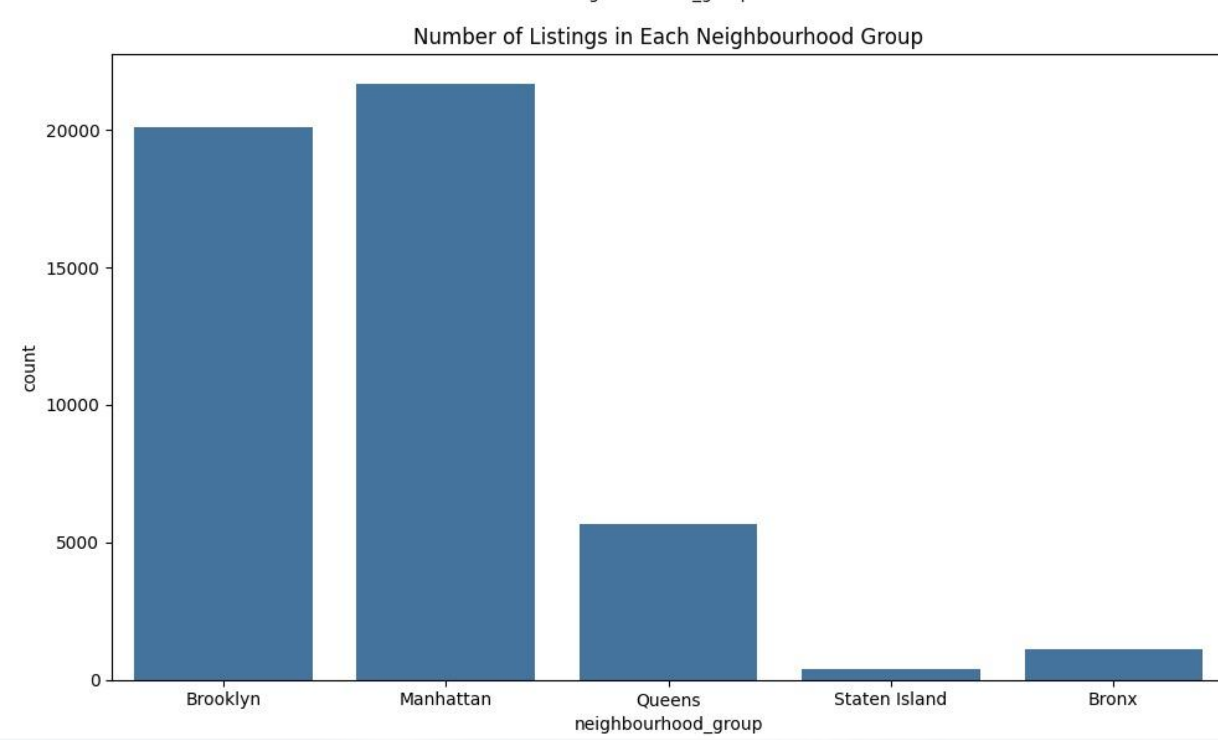
# Median Price + Housing Numbers

**From the Number of Housing :**

- Manhattan & Brooklyn have the most housing availability and listing at NYC
- Queens, Bronx & Staten Island are less

**From the Median Price:**

- Manhattan has the Highest Median Price
- The other parts (Bronx, Queens, Staten Island & Brooklyn) are mostly the same

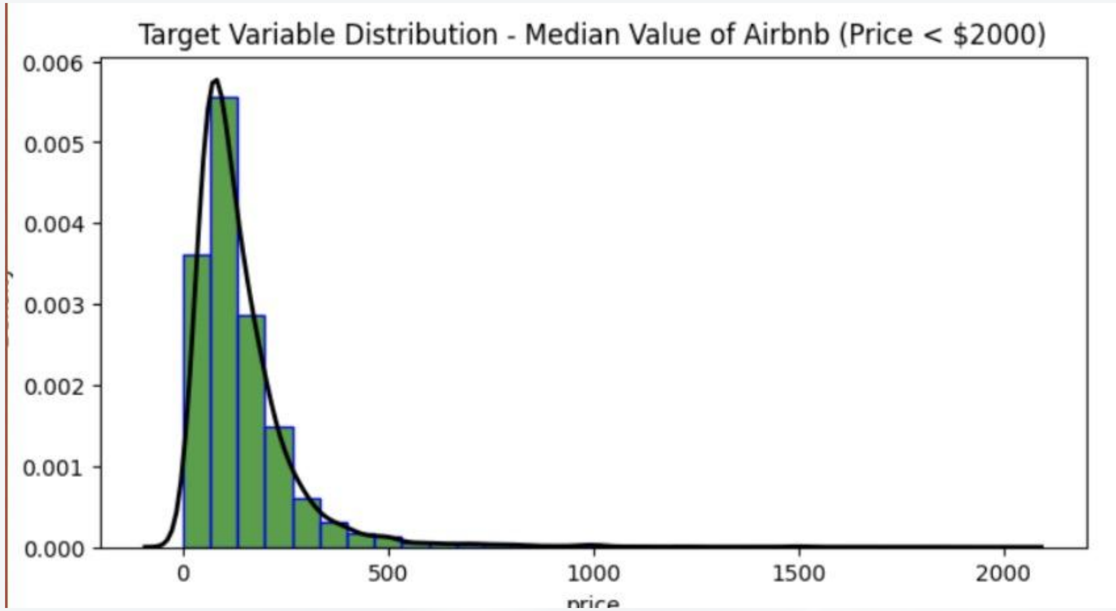


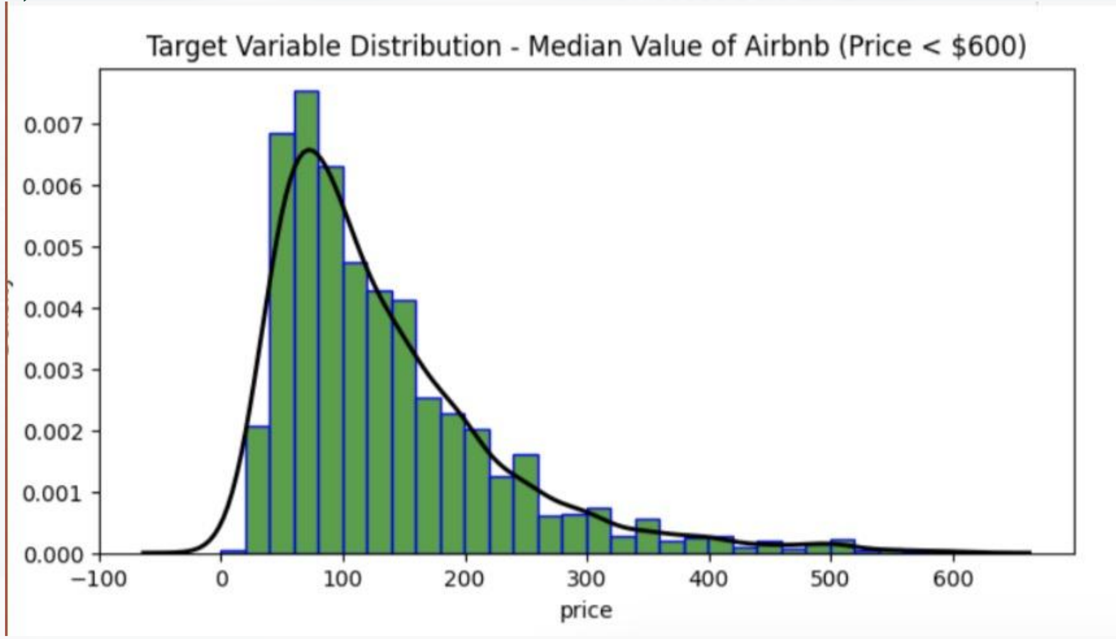

# Median Value of Airbnb

**Target Variable Distribution**:

- Visualized the distribution of the target variable 'price' using histograms with Kernel Density Estimation (KDE).
- Plotted histograms and KDE plots for the entire dataset and subsets filtered for prices less than $600

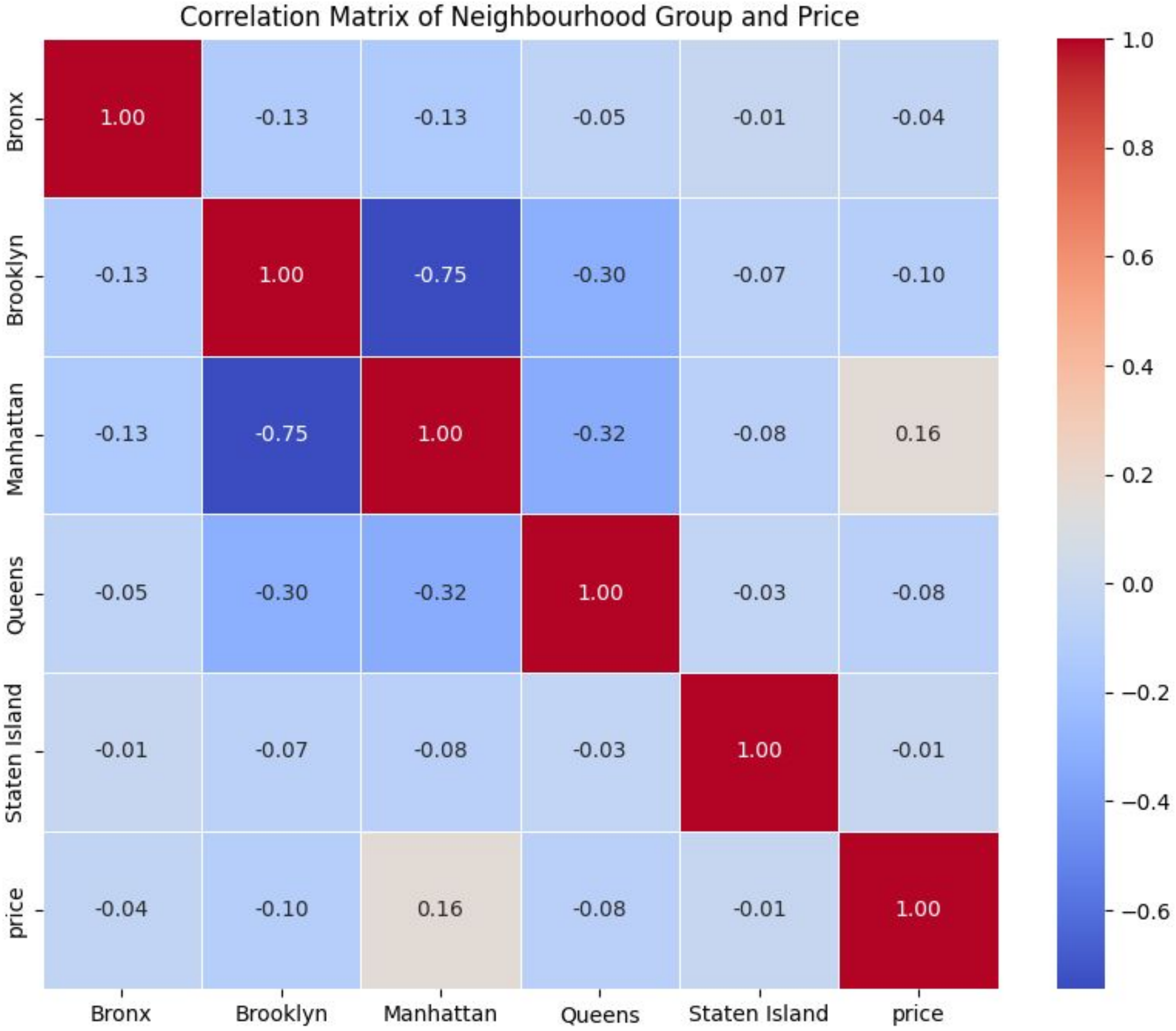
The Median Value of NYC will be at the range (0<x<200)

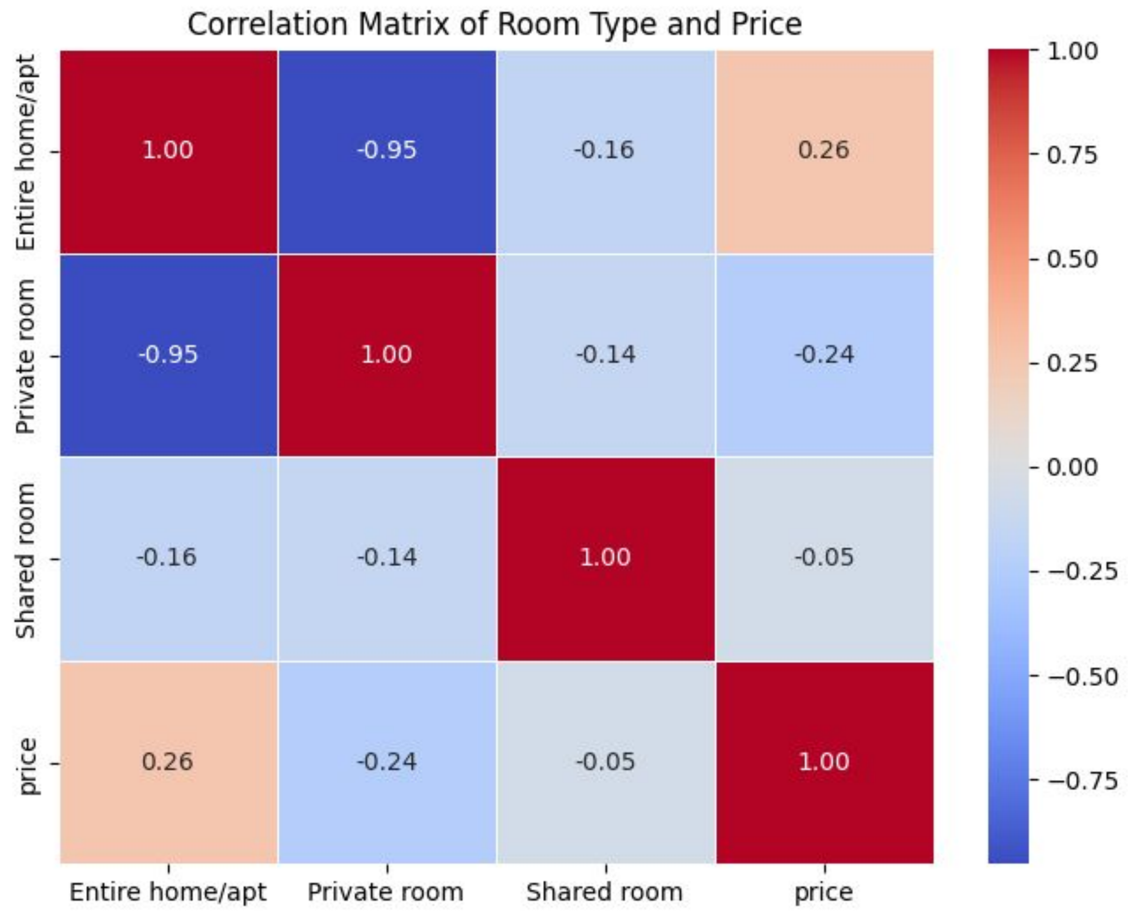


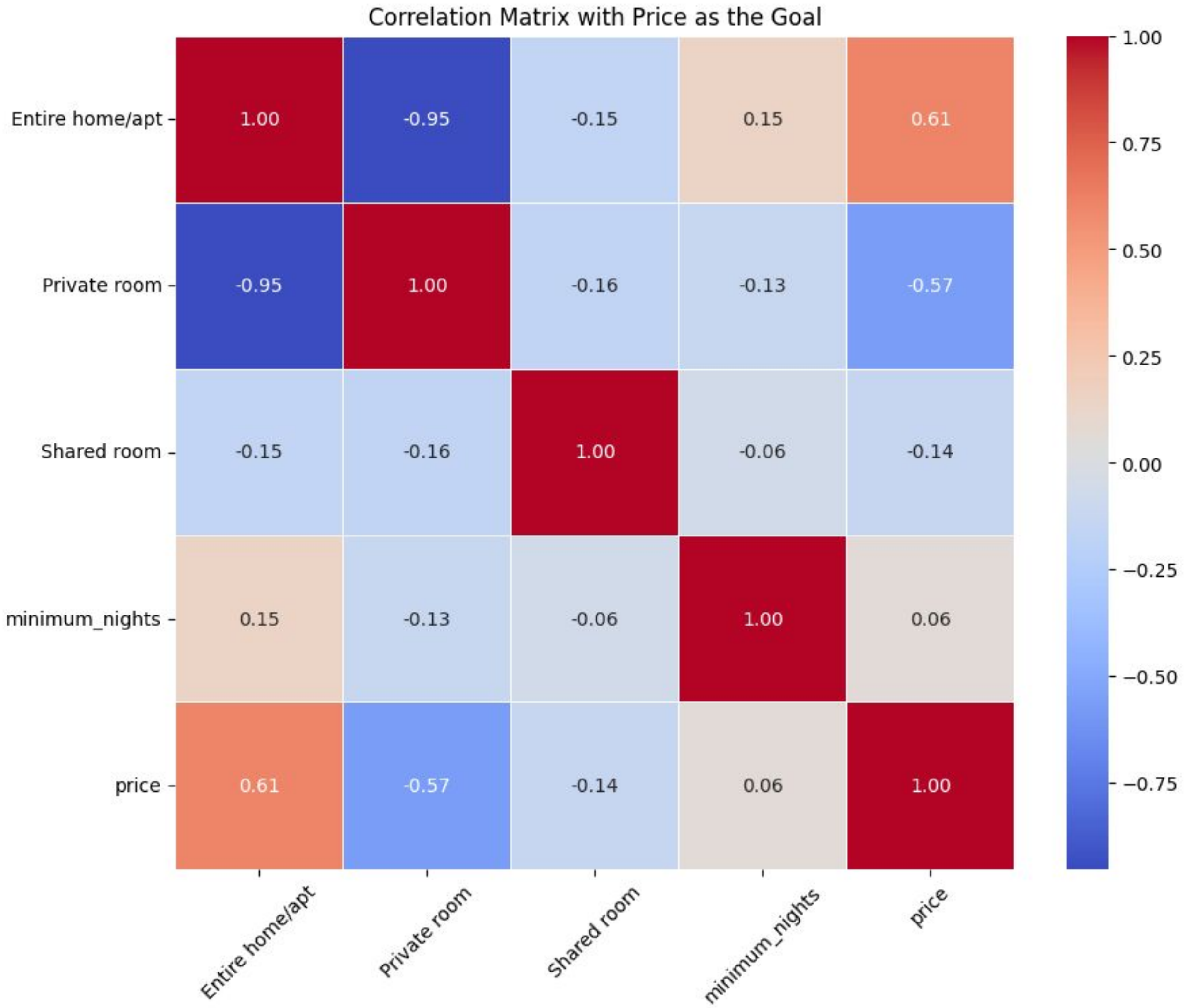



# Data Transformation

- **<u>Data Transformation</u>:**
  1. Created the list named room_type containing the possible values for the 'room_type' column: **'Entire home/apt', 'Private room', and 'Shared room'**.
- **<u>Encoding to Dummy Variables</u>:**
  1. Room type and neighborhood group are both strings, and they both have to be dummy variables before you can see how they relate to price.
  2. Used pd.get_dummies() function to encode the 'room_type' column into dummy variables.
  3. Specified the columns parameter as ['room_type'] to indicate the column to be encoded.
  4. Set the dummy_na parameter to False to avoid creating dummy variables for any potential missing values.
  5. Assigned the resulting DataFrame to df_encoded.
- **<u>Output Display</u>:**
  1. Printed the DataFrame df_encoded to show the encoded representation of the 'room_type' column using dummy variables.


Correlation Matrix of Neighbourhood Group and Price
Bronx
Brooklyn
Manhattan
Queens
Staten Island
price
1.00 -0.13 -0.13 -0.05 -0.01 -0.04
-0.13 1.00 -0.75 -0.30 -0.07 -0.10
-0.13 -0.75 1.00 -0.32 -0.08 0.16
-0.05 -0.30 -0.32 1.00 -0.03 -0.08
-0.01 -0.07 -0.08 -0.03 1.00 -0.01
-0.04 -0.10 0.16 -0.08 -0.01 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Correlation Matrix of Room Type and Price
Entire home/apt
Private room
Shared room
price
1.00
-0.95
-0.16
0.26
-0.95
1.00
-0.14
-0.24
-0.16
-0.14
1.00
-0.05
0.26
-0.24
-0.05
1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Correlation Matrix with Price as the Goal
Entire home/apt
Private room
Shared room
minimum_nights
price
1.00
-0.95
-0.15
0.15
0.61
-0.95
1.00
-0.16
-0.13
-0.57
-0.15
-0.16
1.00
-0.06
-0.14
0.15
-0.13
-0.06
1.00
0.06
0.61
-0.57
-0.14
0.06
1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

05

# Results and Conclusion

# Assumptions

## 01

1. Room_type has influence on price
2. Neighbourhood_group has influence on price
3. Availability has influence on price
4. Minimum_night has influence on price

# Next Step

**Prediction Model :**

- Linear Regression
- Decision Model
  - Linear regression models are usually easier to explain because they provide direct coefficients to describe the relationship between features and targets. Gradient boosting regression trees are more difficult to explain, as they are ensemble models based on a large number of decision trees.

& Learn From Peers & Guest Speaker‘s Recommendation

# Thanks for listening!


Data Science 2024 Bootcamp →

Shihui Feng, Dong Li, Sofia Celorio , Antong Lei, KaiYuan Ma, Tujie Guo

# Works Cited

Aydin, Rebecca. "How 3 Guys Turned Renting Air Mattresses in Their Apartment into a $31 Billion Company, Airbnb." *Business Insider*, www.businessinsider.com/how-airbnb-was-founded-a-visual-history-2016-2. Accessed 30 Mar. 2024.

# Hypothesis Testing

# Description of Hypothesis Testing Methodology

01

Describe the methodology used for hypothesis testing.

# Visualizations in Supporting Findings

01

# Table of Contents

# Consideration of Host, Listing and Geographical factors

01

Discuss the consideration of host characteristics, listing details, and geographical factors in feature selection.

# Potential Areas of Improvement

Provide recommendations for potential areas of improvement.

# Closing Remarks + Implications

01

Conclude with closing remarks and discuss the implications of the findings.

# Challenges Faced during Project

01

Discuss the challenges faced during the project.