AirBnB NYC

# Data Analysis

## Target Audience:

## DATA ANALYSTS

By

Tarriq Ferrose Khan

Batch May 2024 DSC 68

tarriqferrosekhan@gmail.com

# Approach Note

## Data Understanding

- Data is analyzed to understand the variables, data types and data distribution.
- Formatted Data for any like formatting inconsistencies

## Exploratory Data Analysis (EDA)

- Handled Null Value in columns and replaced with appropriate values.
- Checked for Outliers in numerical columns and cleaned those values.
- Created Additional Columns to understand data better.
- Correlations among numeric variables observed to baseline variables before model building

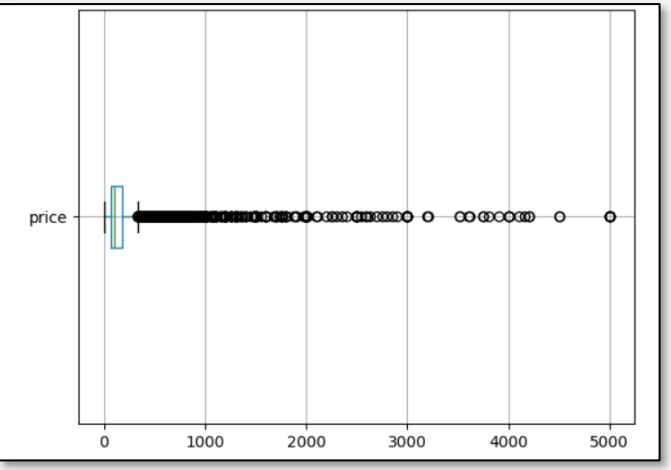
## Model Building

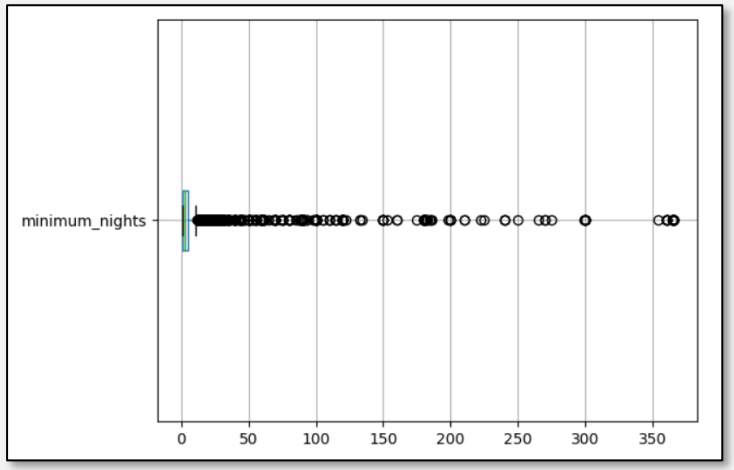
- Used Linear Regression to build a model to understand the price impacting features.

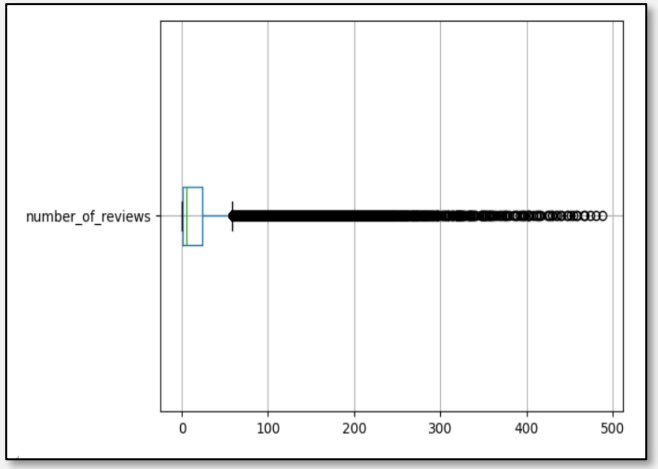
# Data Analysis, processing and Observation Details.

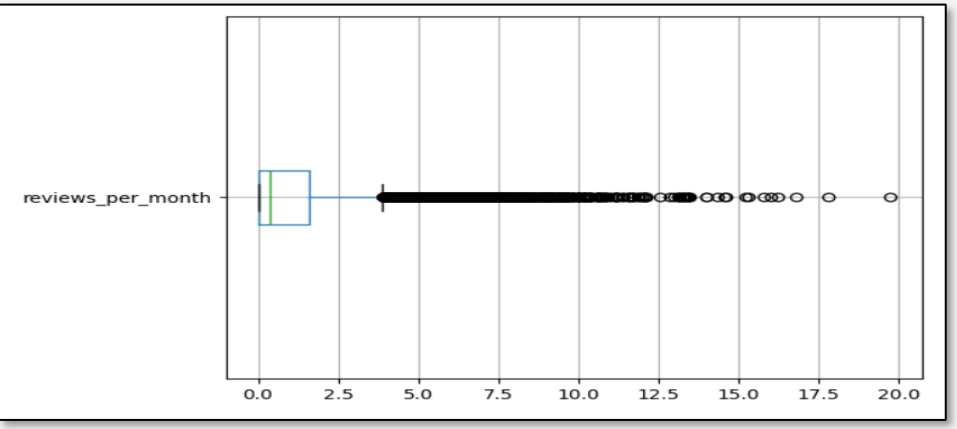
**Exploratory Data Analysis**

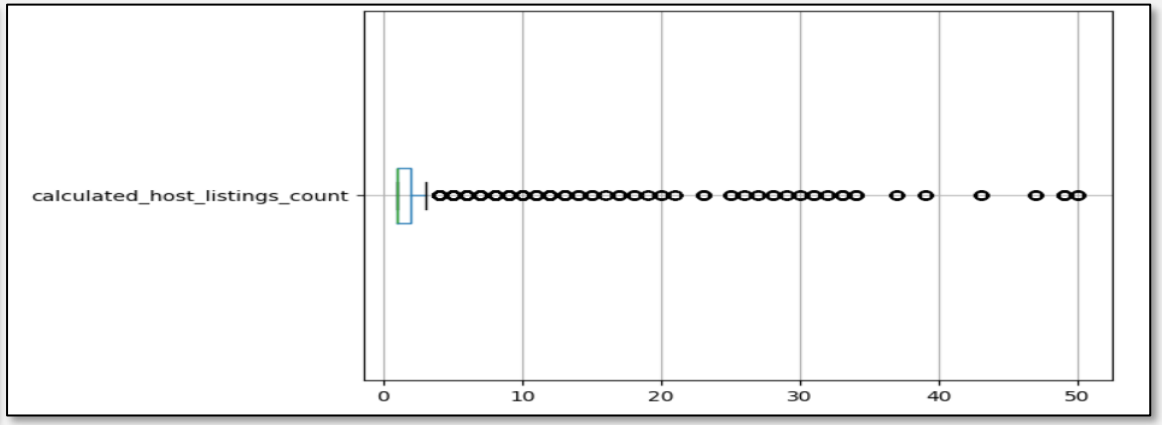
# EDA: Outliers Handled



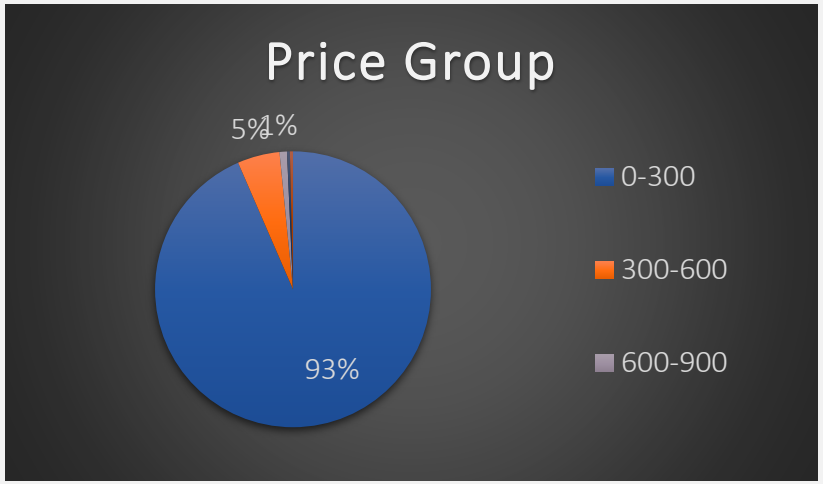



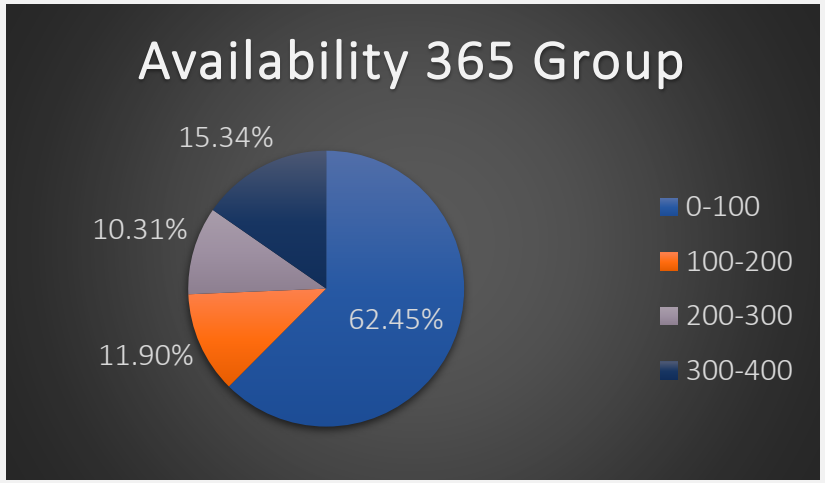



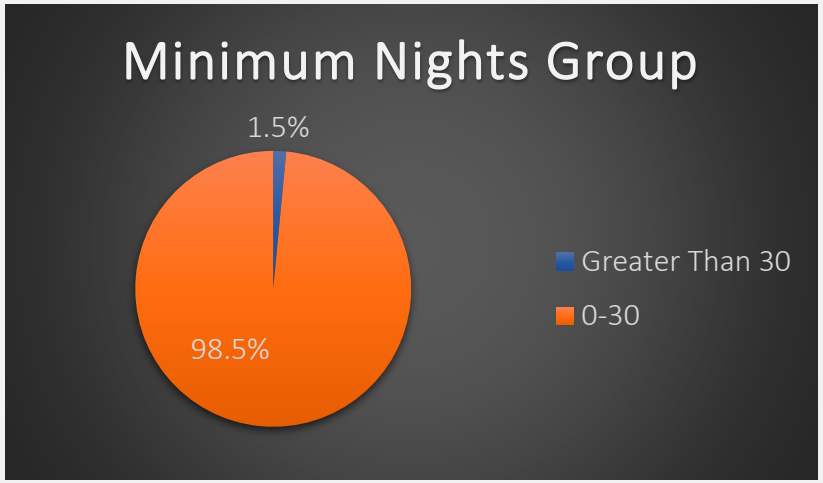



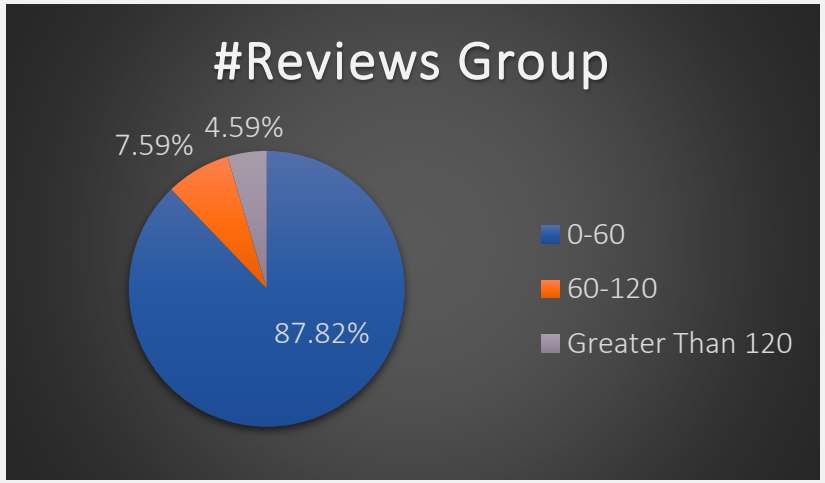

Analysed various numeric Variables for Outliers and removed.

# EDA : Create Additional columns for binning & Distributions (value counts)




Created additional columns through user defined function for binning certain numerical variables:

- PriceGroup - To understand the Price Range
- Availabilty365Group – To understand what's the range of availability
- MinimumNightsGroup- To understand the #listings that falls in the range of Minimum nights.
- ReviewsGroup- To understand the #listings that falls in the range of #Reviews

# EDA: Create additional columns - Text Processing on Listing Name (Regex)

| Gender_Spec | #Listings |
|---|---|
| No | 99.56% |
| Yes | 0.44% |
| **Grand Total** | **100.00%** |

| Event_Spec | #Listings |
|---|---|
| No | 99.92% |
| Yes | 0.08% |
| **Grand Total** | **100.00%** |

| Landmark_Spec | #Listings |
|---|---|
| No | 89.89% |
| Yes | 10.11% |
| **Grand Total** | **100.00%** |

| Bedroom_Spec | #Listings |
|---|---|
| No | 77.51% |
| Yes | 22.49% |
| **Grand Total** | **100.00%** |

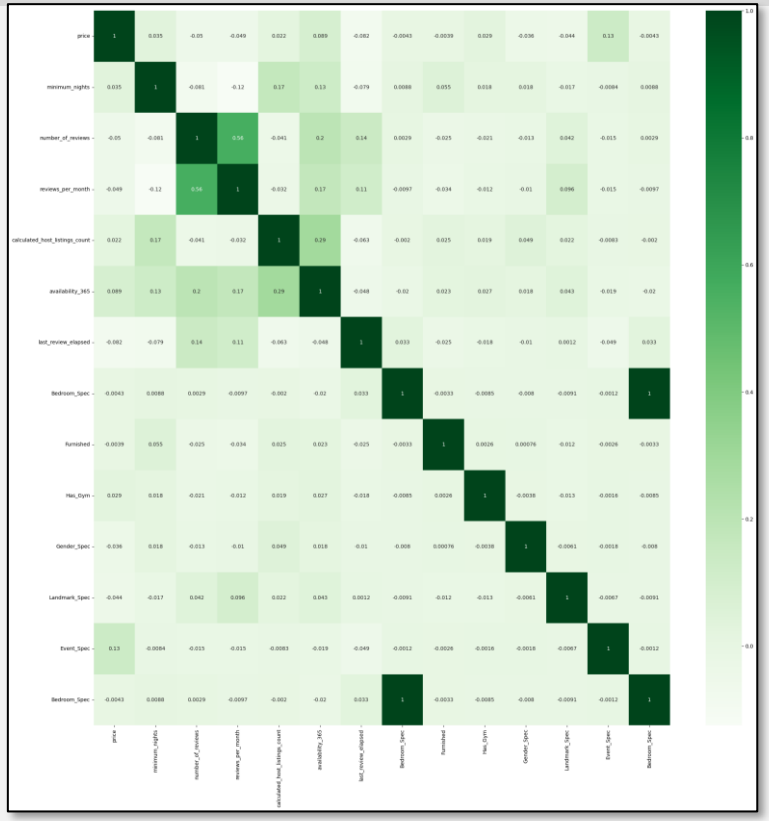
| Furnished | #Listings |
|---|---|
| No | 99.14% |
| Yes | 0.86% |
| **Grand Total** | **100.00%** |

Created additional columns by checking for below using the Regex Processing and Text parsing:

- Bedroom_Spec : Implemented Regex to search for specific Keywords where Bedrooms are specified.
- Furniture_Spec: Checked if Furniture specified.
- Gender_Spec: Checked if any Gender Specifications are mentioned in the property
- Event_Spec: Event Like Super Bowl Highlighted

# EDA: Understanding the correlations

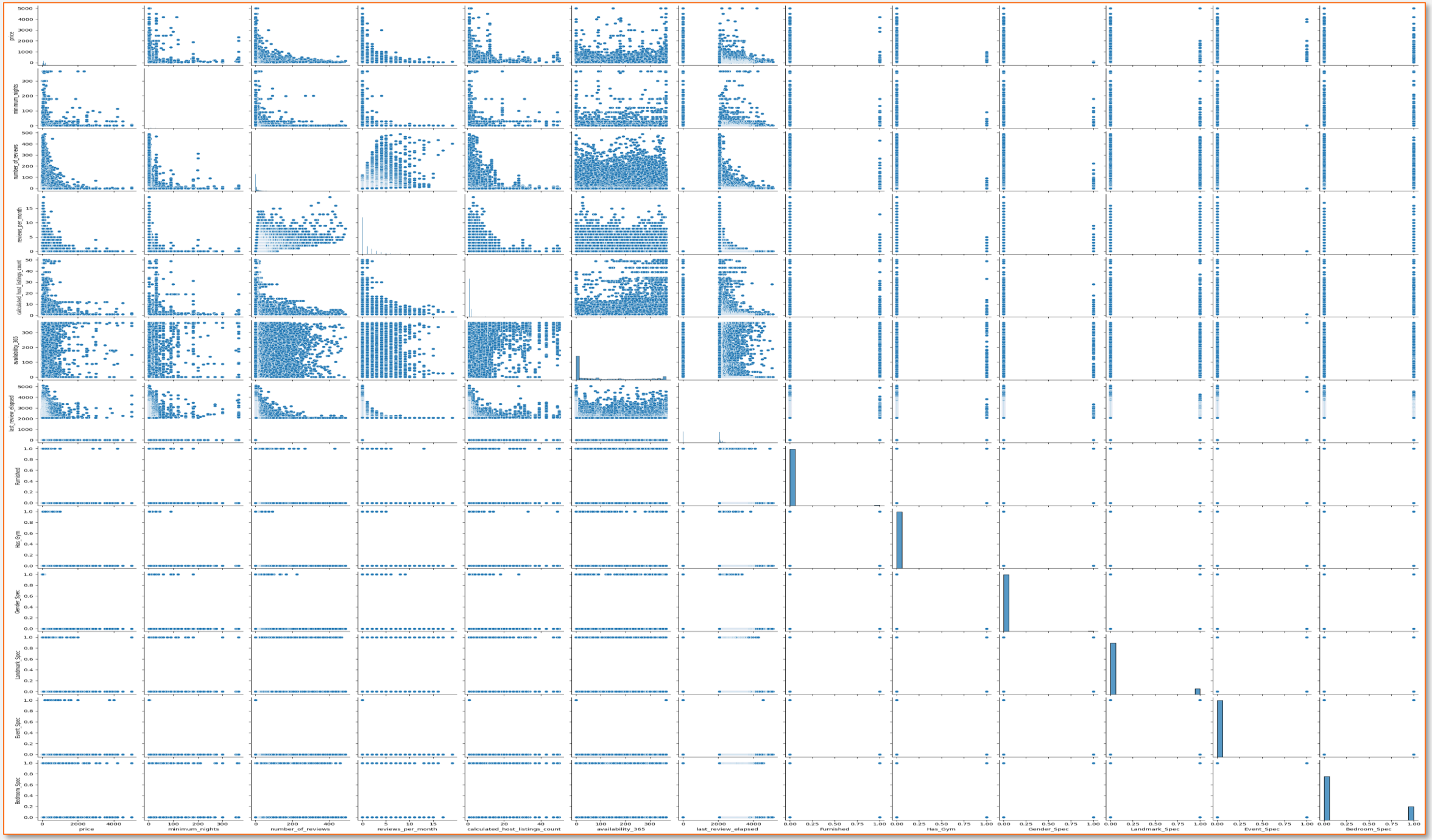
| | price | minimum_nights | number_of_reviews | reviews_per_month | calculated_host_listings_count | availability_365 | last_review_elapsed | Bedroom_Spec | Furnished | Has_Gym | Gender_Spec | Landmark_Spec | Event_Spec | Bedroom_Spec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| price | 1 | 0.035407 | -0.049904 | -0.048924 | 0.022131 | 0.089382 | -0.082341 | -0.00429 | -0.003885 | 0.028879 | -0.035502 | -0.044172 | 0.134253 | -0.00429 |
| minimum_nights | 0.035407 | 1 | -0.081397 | -0.124911 | 0.169042 | 0.129283 | -0.078554 | 0.00881 | 0.055147 | 0.017748 | 0.017893 | -0.017329 | -0.008418 | 0.00881 |
| number_of_reviews | -0.049904 | -0.081397 | 1 | 0.56088 | -0.04104 | 0.195981 | 0.142963 | 0.002924 | -0.025256 | -0.020548 | -0.013353 | 0.042468 | -0.014917 | 0.002924 |
| reviews_per_month | -0.048924 | -0.124911 | 0.56088 | 1 | -0.032106 | 0.170373 | 0.114897 | -0.00973 | -0.033543 | -0.011944 | -0.010404 | 0.096303 | -0.015436 | -0.00973 |
| calculated_host_listings_count | 0.022131 | 0.169042 | -0.04104 | -0.032106 | 1 | 0.288522 | -0.062834 | -0.001957 | 0.025145 | 0.019197 | 0.048624 | 0.022486 | -0.008271 | -0.001957 |
| availability_365 | 0.089382 | 0.129283 | 0.195981 | 0.170373 | 0.288522 | 1 | -0.048015 | -0.019762 | 0.023172 | 0.027343 | 0.018081 | 0.04348 | -0.019026 | -0.019762 |
| last_review_elapsed | -0.082341 | -0.078554 | 0.142963 | 0.114897 | -0.062834 | -0.048015 | 1 | 0.03299 | -0.024891 | -0.018278 | -0.009953 | 0.001242 | -0.049008 | 0.03299 |
| Bedroom_Spec | -0.00429 | 0.00881 | 0.002924 | -0.00973 | -0.001957 | -0.019762 | 0.03299 | 1 | -0.00326 | -0.00854 | -0.007965 | -0.009072 | -0.001155 | 1 |
| Furnished | -0.003885 | 0.055147 | -0.025256 | -0.033543 | 0.025145 | 0.023172 | -0.024891 | -0.00326 | 1 | 0.002563 | 0.000764 | -0.012103 | -0.002577 | -0.00326 |
| Has_Gym | 0.028879 | 0.017748 | -0.020548 | -0.011944 | 0.019197 | 0.027343 | -0.018278 | -0.00854 | 0.002563 | 1 | -0.003785 | -0.013072 | -0.001593 | -0.00854 |
| Gender_Spec | -0.035502 | 0.017893 | -0.013353 | -0.010404 | 0.048624 | 0.018081 | -0.009953 | -0.007965 | 0.000764 | -0.003785 | 1 | -0.006111 | -0.001829 | -0.007965 |
| Landmark_Spec | -0.044172 | -0.017329 | 0.042468 | 0.096303 | 0.022486 | 0.04348 | 0.001242 | -0.009072 | -0.012103 | -0.013072 | -0.006111 | 1 | -0.006745 | -0.009072 |
| Event_Spec | 0.134253 | -0.008418 | -0.014917 | -0.015436 | -0.008271 | -0.019026 | -0.049008 | -0.001155 | -0.002577 | -0.001593 | -0.001829 | -0.006745 | 1 | -0.001155 |
| Bedroom_Spec | -0.00429 | 0.00881 | 0.002924 | -0.00973 | -0.001957 | -0.019762 | 0.03299 | 1 | -0.00326 | -0.00854 | -0.007965 | -0.009072 | -0.001155 | 1 |



Including the newly created Feature columns correlated with Price and observed these were not having any effect on the rest of the numeric columns. please refer to the Pair plot in the next slide

Columns considered for Correlation Analysis:

```
'price',
 'minimum_nights',
 'number_of_reviews',
 'reviews_per_month',
 'calculated_host_listings_count',
 'availability_365',
 'last_review_elapsed',
 'Bedroom_Spec',
 'Furnished',
 'Has_Gym'
```


price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365
last_review_elapsed
Furnished
Has_Gym
Gender_Spec
Landmark_Spec
Event_Spec
Bedroom_Spec

# Understanding Price impacting Factors

**Model Building**

# Model Building – Price impacting Factors

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.833 |
| Model: | OLS | Adj. R-squared: | 0.833 |
| Method: | Least Squares | F-statistic: | 1.046e+04 |
| Date: | Sun, 09 Mar 2025 | Prob (F-statistic): | 0.00 |
| Time: | 18:55:57 | Log-Likelihood: | 93415. |
| No. Observations: | 33656 | AIC: | -1.868e+05 |
| Df Residuals: | 33639 | BIC: | -1.867e+05 |
| Df Model: | 16 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0158 | 0.000 | 34.547 | 0.000 | 0.015 | 0.017 |
| price_300-600 | 0.0511 | 0.000 | 129.815 | 0.000 | 0.050 | 0.052 |
| price_600-900 | 0.1169 | 0.001 | 130.122 | 0.000 | 0.115 | 0.119 |
| price_900-1200 | 0.1741 | 0.001 | 117.794 | 0.000 | 0.171 | 0.177 |
| price_Greater Than 1200 | 0.4225 | 0.001 | 311.872 | 0.000 | 0.420 | 0.425 |
| Private room | -0.0142 | 0.000 | -82.674 | 0.000 | -0.015 | -0.014 |
| Shared room | -0.0185 | 0.001 | -33.929 | 0.000 | -0.020 | -0.017 |
| Brooklyn | 0.0123 | 0.000 | 26.376 | 0.000 | 0.011 | 0.013 |
| Manhattan | 0.0178 | 0.000 | 38.261 | 0.000 | 0.017 | 0.019 |
| Queens | 0.0100 | 0.001 | 19.697 | 0.000 | 0.009 | 0.011 |
| Staten Island | 0.0090 | 0.001 | 8.686 | 0.000 | 0.007 | 0.011 |
| AVL_365_100-200 | 0.0021 | 0.000 | 7.823 | 0.000 | 0.002 | 0.003 |
| AVL_365_200-300 | 0.0020 | 0.000 | 7.064 | 0.000 | 0.001 | 0.003 |
| AVL_365_300-400 | 0.0037 | 0.000 | 15.378 | 0.000 | 0.003 | 0.004 |
| RVWSPERMONTH_Greater Than 8 | -0.0031 | 0.002 | -2.048 | 0.041 | -0.006 | -0.000 |
| NUM_RVW_60-120 | -0.0007 | 0.000 | -2.299 | 0.022 | -0.001 | -0.000 |
| NUM_RVW_Greater Than 120 | -0.0012 | 0.000 | -2.842 | 0.004 | -0.002 | -0.000 |

```
from sklearn.metrics import r2_score
r2_score(y_test_new, y_pred_cnt_new)
```

✓ 0.0s

```
0.8059996807702011
```

**Equation from the predicted Model**

const x 0.0158 + price_300-600 x 0.0511 + price_600-900 x 0.1169 + price_900-1200 x 0.1741 + price_Greater Than 1200 x 0.4225 + Private room x -0.0142 + Shared room x -0.0185 + Brooklyn x 0.0123 + Manhattan x 0.0178 + Queens x 0.0100 + Staten Island x 0.0090 + AVL_365_100-200 x 0.0021 + AVL_365_200-300 x 0.0020 + AVL_365_300-400 x 0.0037 + RVWSPERMONTH_Greater Than 8 x -0.0031 + NUM_RVW_60-120 x -0.0007 + NUM_RVW_Greater Than 120 x -0.0012

# Summary

## Exploratory Data Analysis

Cleaned up data, created additional columns to understand the patterns of distribution and correlations.

## Model

Regression Model was built to understand the Price impacting Features.

# Conclusion

Data Analysis Reveals 0-300$ is preferred Price Range and Manhattan is the most preferred neighborhood.

Availability_365 – the number of days the property is available in a year as significant correlation.

Gender and other specifications were not much influencing the prices.

Thank you,

Tarriq Ferrose Khan

tarriqferrosekhan@gmail.com

# Appendix

Part-1_Methodology.pdf