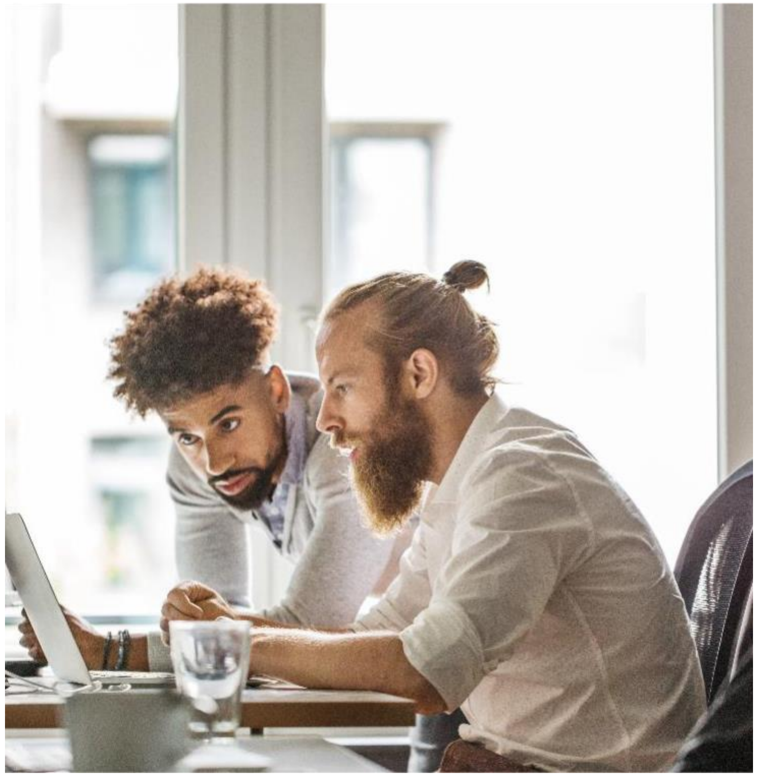
Strategic Data Insights for Revenue Growth: A Focus on Airbnb New York Listings

Sidharth Jayanth Kini
Sukanya Debnath
Sathwik Bidayanth
August 2024

Agenda



- Objective
- Data Lifecycle Management
- Data Overview and Cleaning
- Data Analysis and Insight generation
- Conclusion
- Appendix

Objective

- Study the given Dataset thoroughly and Analyze key attributes of given Dataset to identify strategies for increasing Airbnb's revenue.
- Ask the effective questions that can lead to data insights
- Provide insights on pricing ranges, property types, and adjustments needed to attract more customers.
- Share findings by data visualization and statistical techniques.

Data Lifecycle Management

- **Data Acquisition and Loading**
  - •Data is captured from various sources and loaded into different environments.
- **Data Cleaning and Preparation**
  - •The data undergoes cleaning and preprocessing.
  - •Exploratory Data Analysis (EDA) is performed.
  - •New features are engineered to enhance the dataset.
- **Analysis and Insight Generation**
  - •Meaningful insights are derived using various analytical methods and techniques.
  - •These insights inform strategic decision-making and business processes.

Data Overview

- The given dataset contains information about different Airbnb listings in New York City along with their hosts, locations, prices and other attributes.

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Detailed Dataset Description

```
Categorical Variables:
    - room_type
    - neighbourhood_group
    - neighbourhood

Continous Variables(Numerical):
    - Price
    - minimum_nights
    - number_of_reviews
    - reviews_per_month
    - calculated_host_listings_count
    - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
    - latitude
    - longitude

Time Varibale:
    - last_review
```

Variable Categories

Data Cleaning

- **Categorization** : Continuous Variables were binned into groups for deeper understanding of relationships and connections between elements.
    - Categorization was done for Price, minimum_night, availability_356,number_of_reviews.
- **Fixing Columns and Data types**: Updating the columns with correct Data type to combat inconsistencies
- **Missing Value Analysis :** To review the Missing Values in give data

```python
def price_categorization(row):

    if row <= 50:
        return 'very Low'
    elif row <= 100:
        return 'Low'
    elif row <= 200:
        return 'Medium'
    elif (row <= 250):
        return 'High'
    else:
        return 'very High'
```

Categorization

Note: we observed that the data type of last_review column is object.

This needs to be converted into datetime format for our analysis

```
In [29]: airbnb.last_review = pd.to_datetime(airbnb.last_review)
         airbnb.last_review

Out[29]: 0        2018-10-19
         1        2019-05-21
         2               NaT
         3        2019-05-07
         4        2018-11-19
                     ...
         48890           NaT
         48891           NaT
         48892           NaT
         48893           NaT
         48894           NaT
         Name: last_review, Length: 48895, dtype: datetime64[ns]
```

Fixing Columns and Data Types

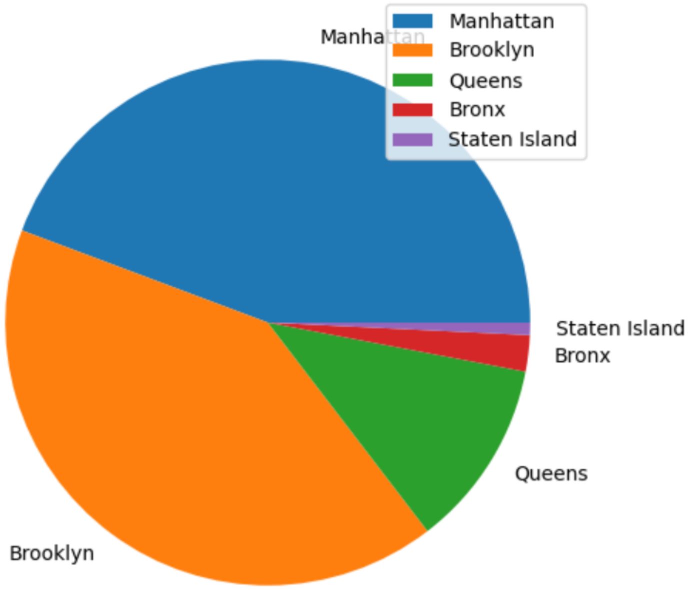
```
In [39]: missing_perc=round(airbnb.isna().sum()/len(airbnb)*100,2)
         missing_perc.sort_values(ascending=False)

Out[39]: reviews_per_month                  20.56
         last_review                        20.56
         host_name                           0.04
         name                                0.03
         id                                  0.00
         number_of_reviews                   0.00
         number_of_reviews_categories        0.00
         minimum_night_categories            0.00
         price_categories                    0.00
         availability_365                    0.00
         calculated_host_listings_count      0.00
         minimum_nights                      0.00
         price                               0.00
         room_type                           0.00
         longitude                           0.00
         latitude                            0.00
         neighbourhood                       0.00
         neighbourhood_group                 0.00
         host_id                             0.00
         availability_365_categories         0.00
         dtype: float64
```

Missing Value Analysis

Note: It is imperative to highlight that we will **neither drop nor impute any columns**, as our primary focus is on analyzing the dataset rather than constructing a model. Additionally, the majority of the features hold significant importance for our analysis

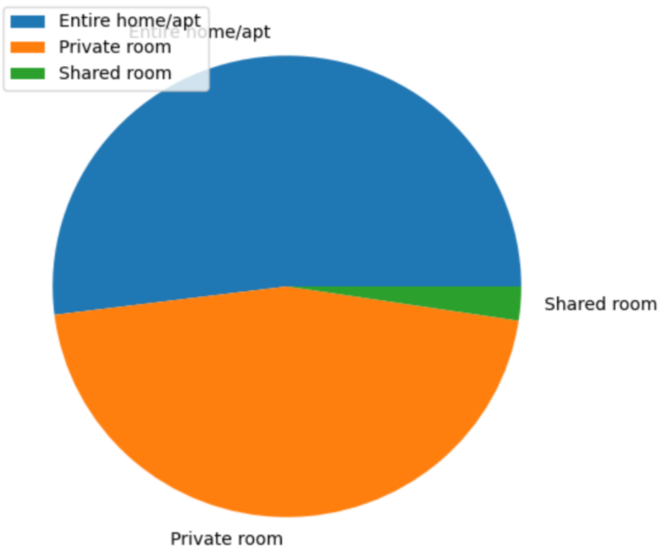
Target Neighbourhood Groups



Manhattan	44.301053
Brooklyn	41.116679
Queens	11.588097
Bronx	2.231312
Staten Island	0.762859

- Manhattan and Brooklyn Neighbourhood groups have highest contributing towards the Airbnb listings. They have a combined contribution of **85%**.
- Staten Island has lowest contribution towards the Airbnb listings.

Room Type Analysis



Contribution of Room Type for Listing

number_of_reviews_categories	High	Low	Medium	very High	very Low
room_type					
Entire home/apt	4281	5177	3015	5306	7630
Private room	3758	4213	2290	4850	7215
Shared room	197	207	125	180	451

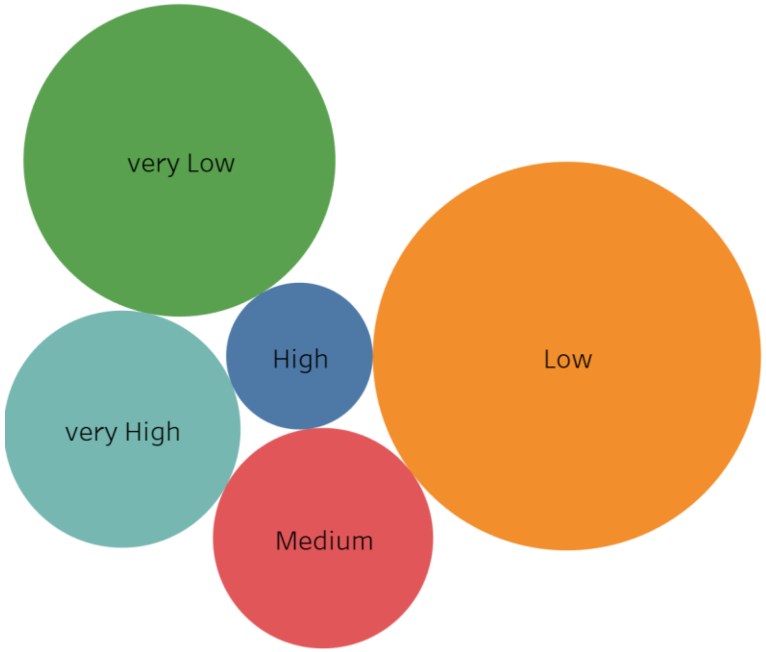
```
room_type
Entire home/apt    22.842418
Private room       24.112962
Shared room        16.600000
dtype: float64
```

Number of Reviews vs Room Type

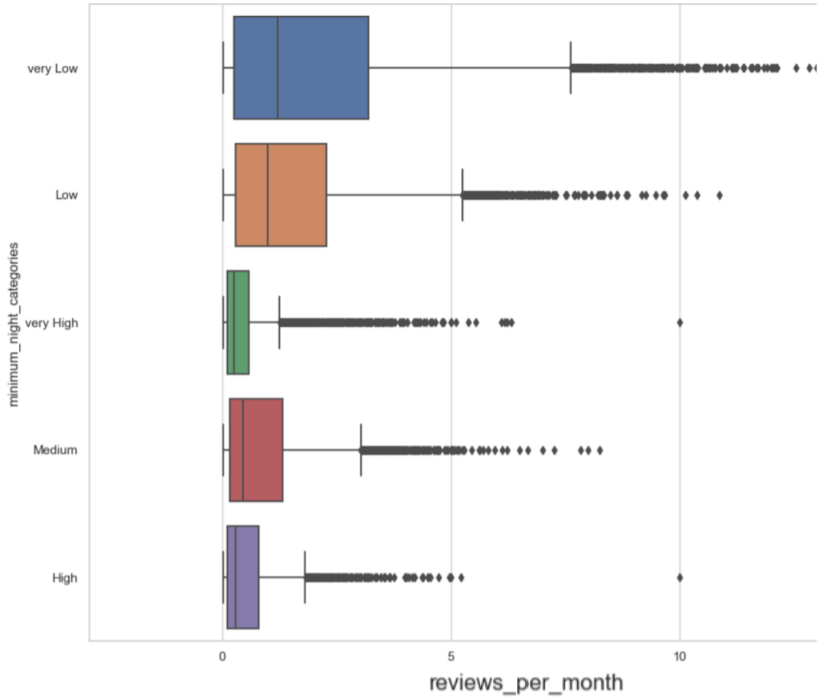
- Entire home/apt (51.9%) and Private rooms (45.6%) forms Majority of Listing space type.
- Shared rooms forms only 2.3% of listing space type.

- Shared Room type has the lowest numbers across all review categories
- This suggests shared rooms are less popular compared to entire homes or private rooms.

Minimum Nights Categories Analysis



Distribution of Minimum Nights for Airbnb Listing

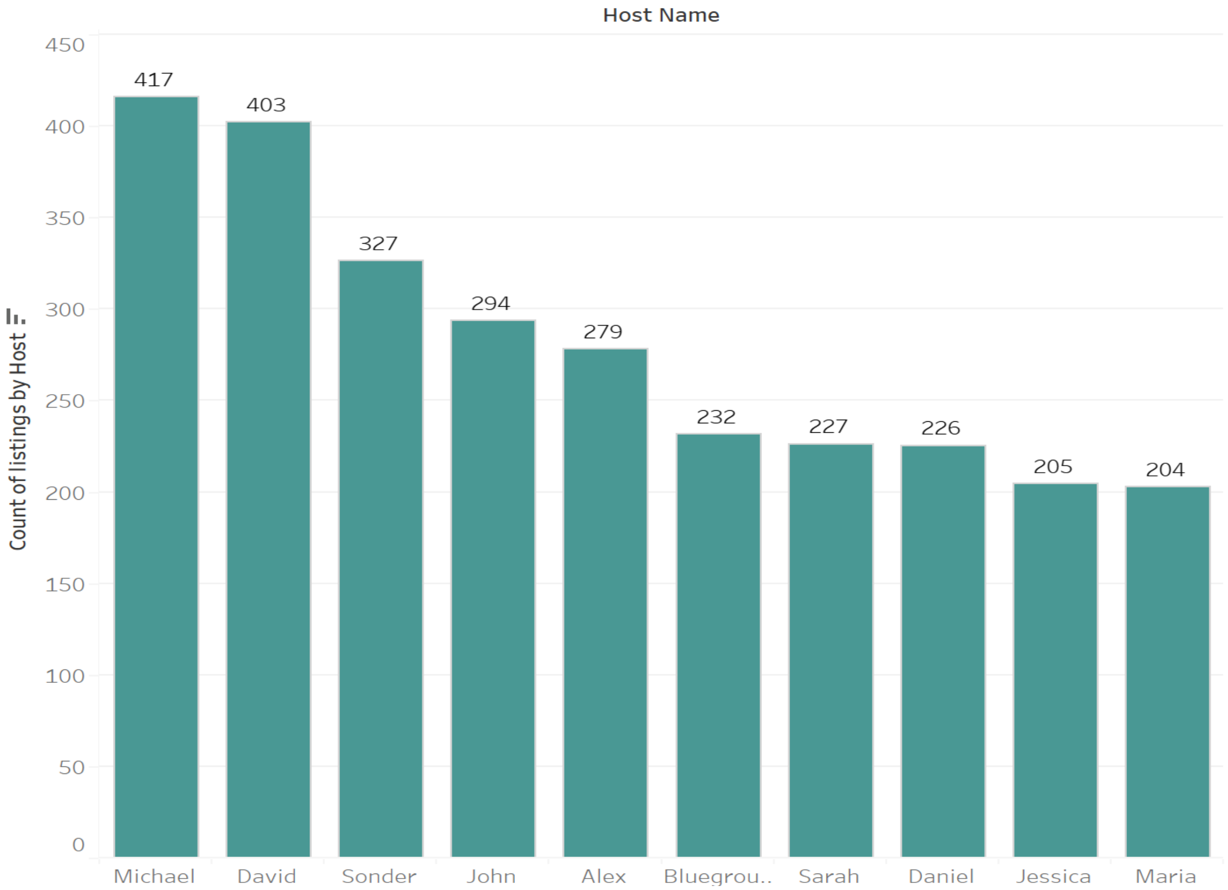


Effects on Minimum Nights on monthly review

- The majority of Airbnb listings in NYC require only a few minimum nights for booking.

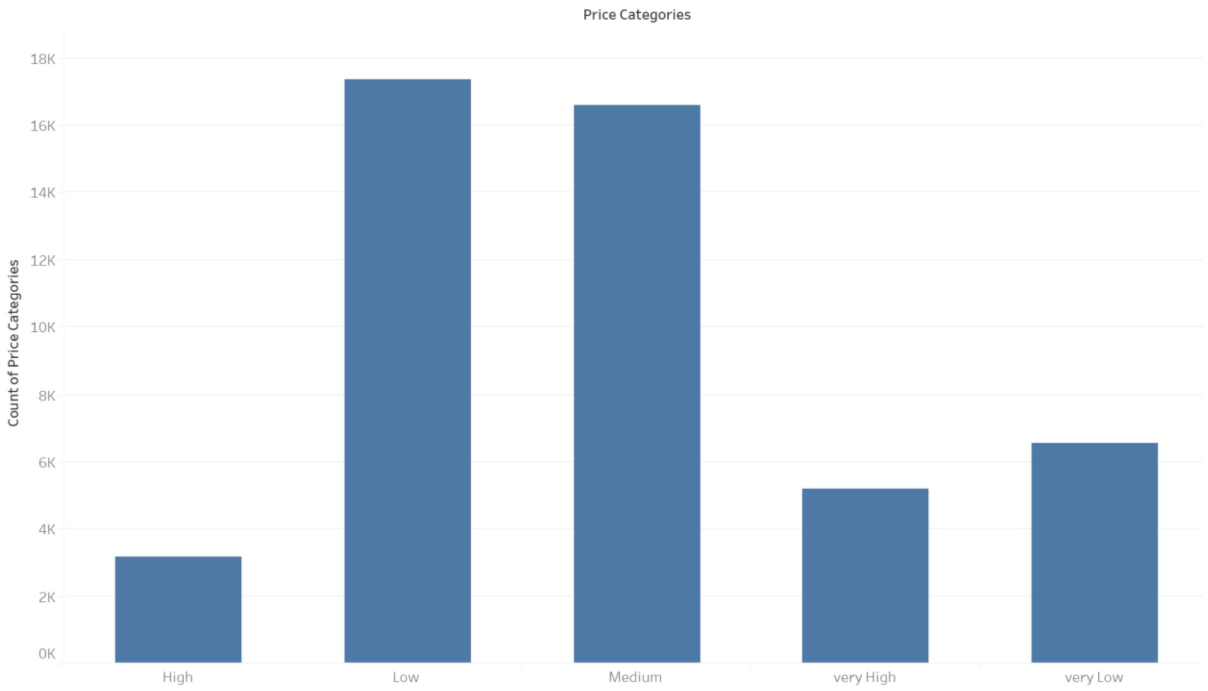
- Listings with “Very low” and “low” number of minimum nights categories tend to receive high reviews on a monthly basis

Host Analysis



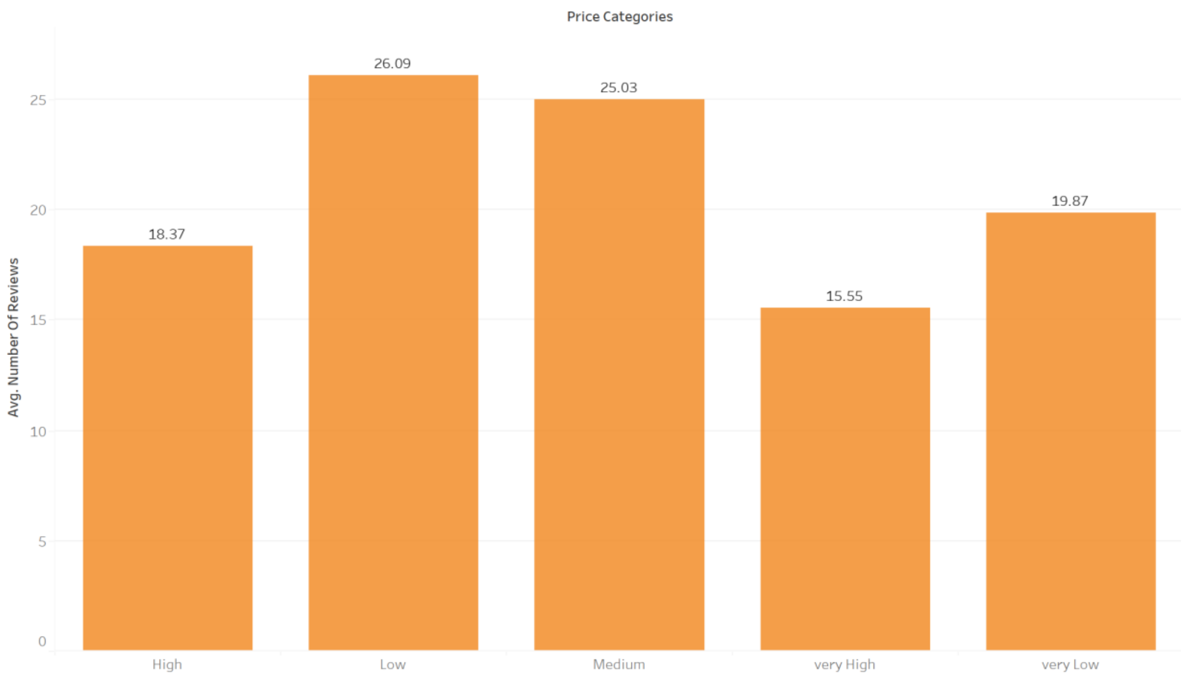
- ➢ Top 10 Hosts with highest number of Airbnb property listings in NYC.
- ➢ Micheal Owns Highest listings summing up to 417 Airbnb properties

Price Categories Analysis



Price category distribution for Airbnb Listing

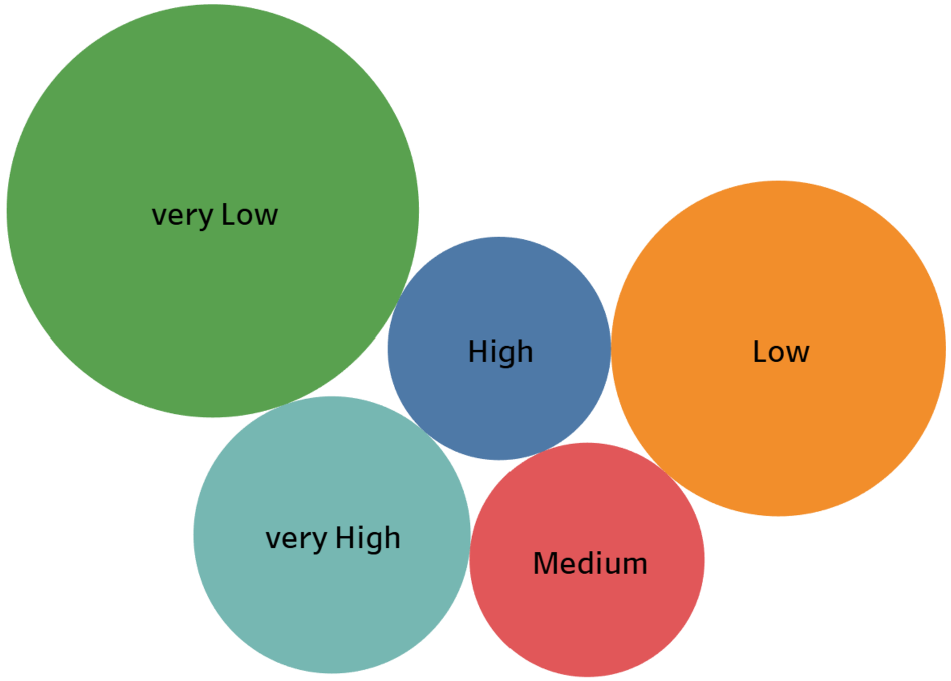
- The majority of Airbnb listings in NYC falls under Low and Medium Price Categories



Effect of listing price on reviews

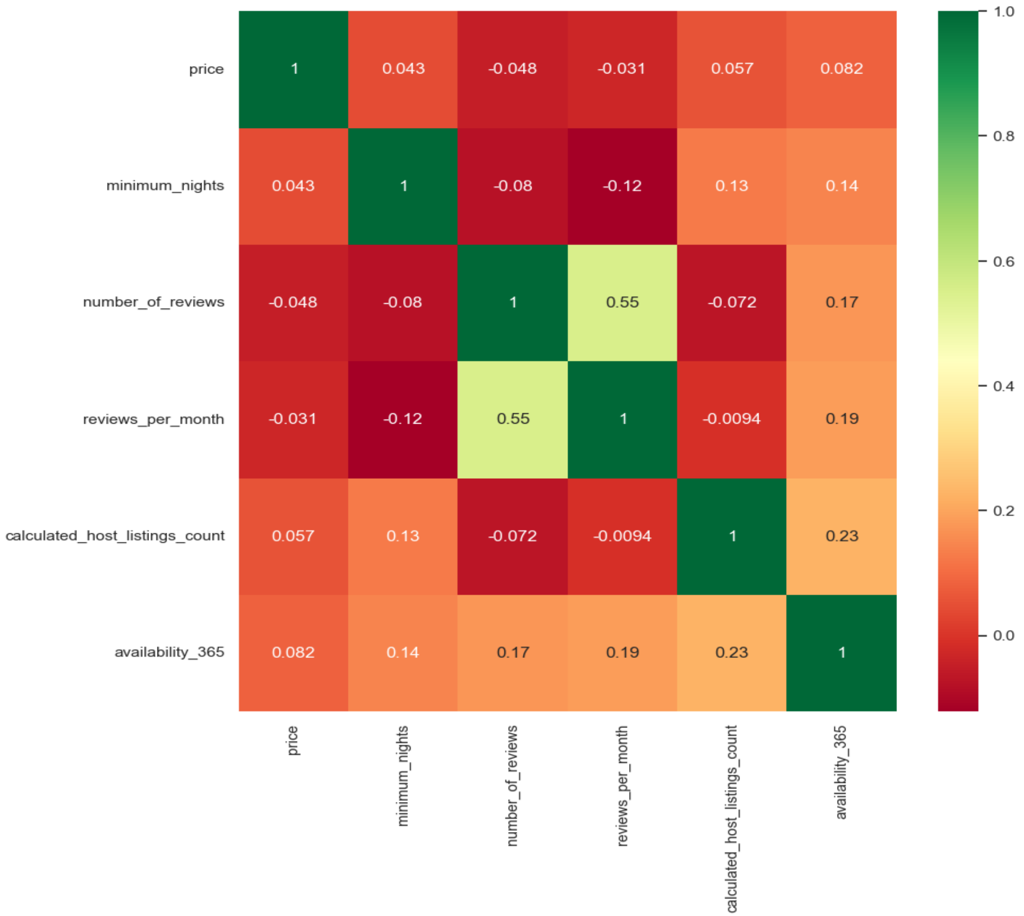
- Listings with Low and Medium price categories tend to receive high reviews on an average

Availability Analysis



- The availability of majority of the Airbnb listing in NYC fall between “Very Low” and “Low” availability_365 category.
- Only Few Airbnb listings are available throughout the year for booking.

Correlation Analysis



Heat Map

Top correlation are given below

```
sol[1:8]
```

```
calculated_host_listings_count  availability_365                  0.225701
reviews_per_month               availability_365                  0.185791
number_of_reviews               availability_365                  0.172028
minimum_nights                  availability_365                  0.144303
                                calculated_host_listings_count    0.127960
                                reviews_per_month                 0.121702
price                           availability_365                  0.081829
dtype: float64
```

- To understand the correlation between the Variables Heat was plotted.
- Top meaningful correlation between the variables were also derived. (refer the above table)

Conclusion

- Significant insights have been obtained based on various attributes in the dataset.
- A substantial amount and variety of visuals have been utilized in the presentations for stakeholders.
- The data collection team should gather data on review scores to enhance future analyses.
- A machine learning model can be developed to identify groups of similar objects in datasets with multiple variables.

Appendix –Data Methodology

- Performed a comprehensive analysis of the New York Airbnbs dataset.
- Cleaned the dataset using Python programming.
- Extracted the essential features.
- Applied group aggregation, pivot tables and other statistical techniques.
- Developed charts and visualizations using Tableau.