Storytelling Case Study: Airbnb, NYC

Debanik, Pratyusha and Abhishek

CONTENTS

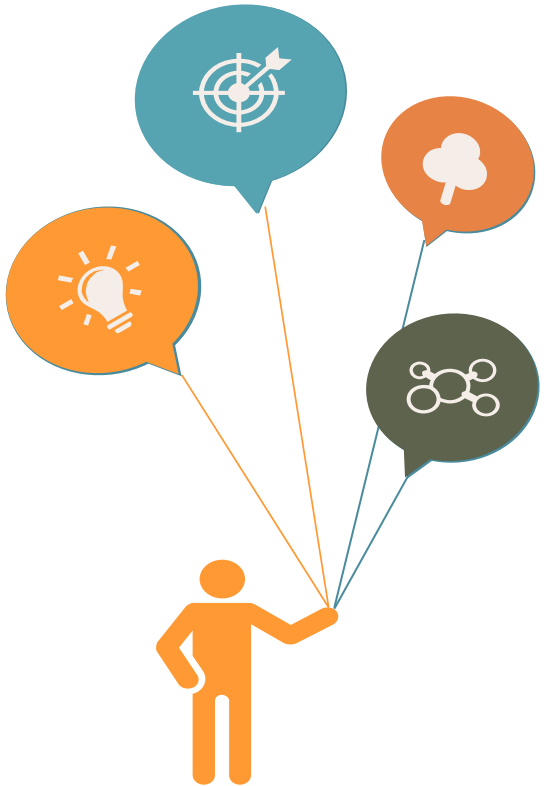
OBJECTIVE

Data Analysis

1. Get a better understanding about Airbnb listings with respect to various parameters.
2. Understand the customer preferences.
3. Understand the customer booking trend.

Exploratory Data Analysis

To understand some important insights, we have explored the following questions:
1. How are the Airbnb listings spread out in NYC?
2. What type of rooms do customers prefer?
3. What could be the ideal number of minimum nights to increase customer bookings?



Data Presentation

Based on customer review:
1. Most preferred neighborhood
2. Most preferred room type
3. Who are the Hosts who have the highest listings w.r.t Neighborhood?

Methodology

-> The data was analyzed through univariate and bivariate analysis.
-> The analysis and visualizations were done using Tableau considering various parameters.
-> The main parameters that have been taken into account for analysis are –
1. Geography based bookings
2. Bookings based on room type
3. Number of reviews
4. Minimum number of nights

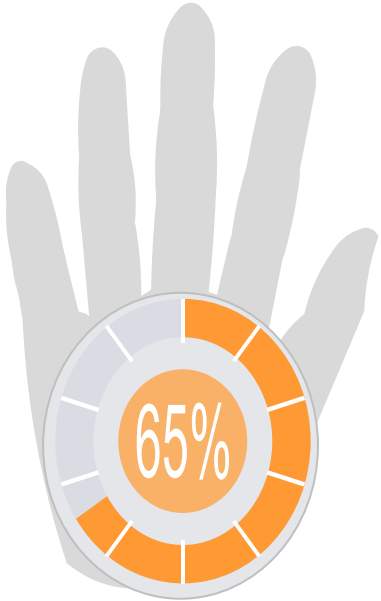
DATA ANALYSIS

Read Data

Data is read using python and tableau.

EDA

Analyze and clean data and create bins for continuous variables



Recommendations

Perform univariate, bivariate analysis and correlations to capture insights

Visualizations

Create visualizations using Tableau for further insights

Read & Review Data

In [2]:
```
# Read the dataset
airbnb_df = pd.read_csv('AB_NYC_2019.csv')
```

In [3]:
```
# View the data
airbnb_df.head(10)
```

Out[3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM...NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	
5	5099	Large Cozy 1 BR Apartment In	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire	200	3	

In [4]:
```
airbnb_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
```

In [5]:
```
airbnb_df.shape
```

Out[5]:
```
(48895, 16)
```

In [6]:
```
airbnb_df.describe()
```

Out[6]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.

Binning Continuous Variables to Categories

```
In [14]: airbnb_df['price_categories'].value_counts()
```

```
Out[14]: price_categories
         very High    24967
         High         17367
         Medium        6474
         Low             59
         very Low        28
         Name: count, dtype: int64
```

```
In [19]: airbnb_df['minimum_night_categories'].value_counts()
```

```
Out[19]: minimum_night_categories
         Low          19695
         very Low     12720
         very High     6640
         Medium        6337
         High          3503
         Name: count, dtype: int64
```

```
In [24]: airbnb_df['number_of_reviews_categories'].value_counts()
```

```
Out[24]: number_of_reviews_categories
         very Low     15296
         Low           9597
         Medium        8612
         High          8431
         very High     6959
         Name: count, dtype: int64
```

```
In [29]: airbnb_df['reviews_per_month_categories'].value_counts()
```

```
Out[29]: reviews_per_month_categories
         Medium       25765
         very High    11519
         High          8298
         Low           2352
         very Low       961
         Name: count, dtype: int64
```

This is done to create relationships between variables and get more insights into the data

Data Classification

4.1 Categorical Columns

In [42]:
```
# Categorical columns
cat_cols = airbnb_df.columns[[0,1,3,4,5,8,16,17,18,19,20,21]]
cat_cols
```

Out[42]:
```
Index(['id', 'name', 'host_name', 'neighbourhood_group', 'neighbourhood',
       'room_type', 'price_categories', 'minimum_night_categories',
       'number_of_reviews_categories', 'reviews_per_month_categories',
       'calculated_host_listings_count_categories',
       'availability_365_categories'],
      dtype='object')
```

In [43]:
```
airbnb_df[cat_cols].head()
```

Out[43]:

	id	name	host_name	neighbourhood_group	neighbourhood	room_type	price_categories	minimum_night_categories	number_of_reviews_categ
0	2539	Clean & quiet apt home by the park	John	Brooklyn	Kensington	Private room	very High	very Low	Me

4.2 Numerical Columns

In [44]:
```
num_cols = airbnb_df.columns[[9,10,11,13,14,15]]
num_cols
```

Out[44]:
```
Index(['price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')
```

In [45]:
```
airbnb_df[num_cols].head()
```

Out[45]:

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
0	149	1	9	0.21	6	365

4.3 Location & Time Variables

```
loc = airbnb_df.columns[[5,6,12]]
airbnb_df[loc]
```

	neighbourhood	latitude	last_review
0	Kensington	40.64749	19-10-2018
1	Midtown	40.75362	21-05-2019
2	Harlem	40.80902	NaN
3	Clinton Hill	40.68514	05-07-2019
4	East Harlem	40.79851	19-11-2018
...	...	...	...

This is done to help perform univariate analysis

Missing Value Analysis

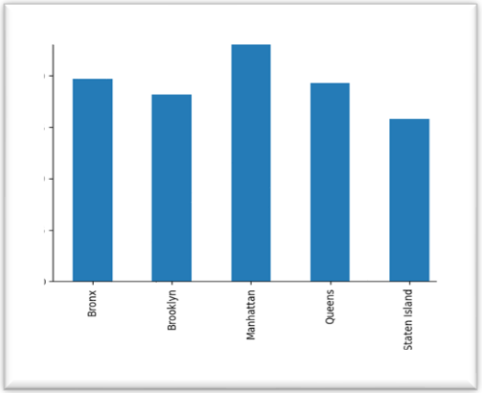
```
# Percentage of missing values
round((airbnb_df.isnull().sum()/len(airbnb_df))*100,2)
```

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

1. Two columns (last_review , reviews_per_month) has around 20.56% missing values. name and host_name has 0.3% and 0.4 % missing values
2. We need to see if the values are missing at random or not.
3. Most of the features are important for analysis and we are not making a model, only analysing hence we are not dropping any columns.

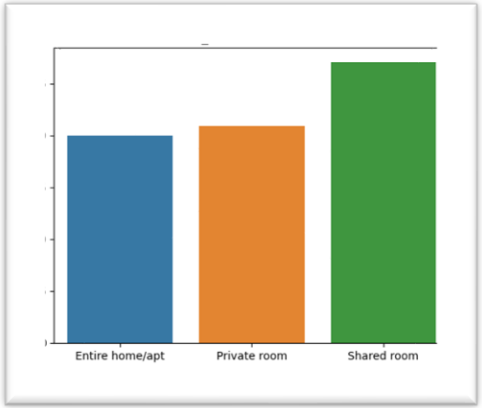
Conclusion:

1. The pricing is higher when 'last_review' is missing.
2. Lesser reviews are given for shared rooms.
3. Higher room prices have more reviews.
4. Missing values are not at random



Observation:

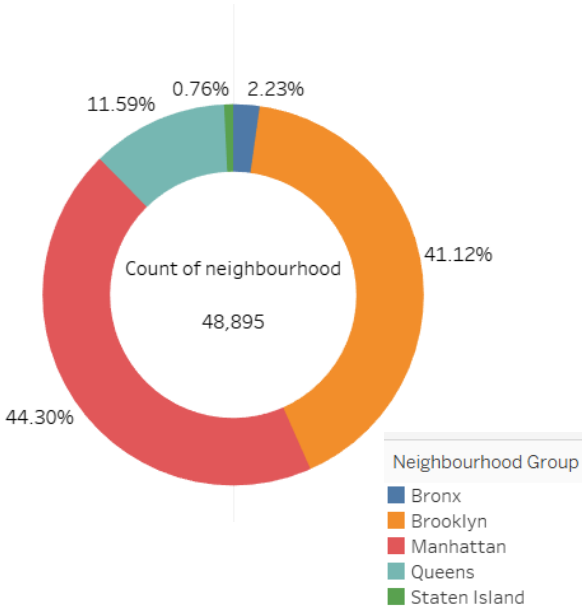
Each neighbourhood_group has about 19 % missing values in 'last_review' feature.



Observation:

'Shared room' has the highest missing value percentage (27 %) while other room types have only about 20 %.

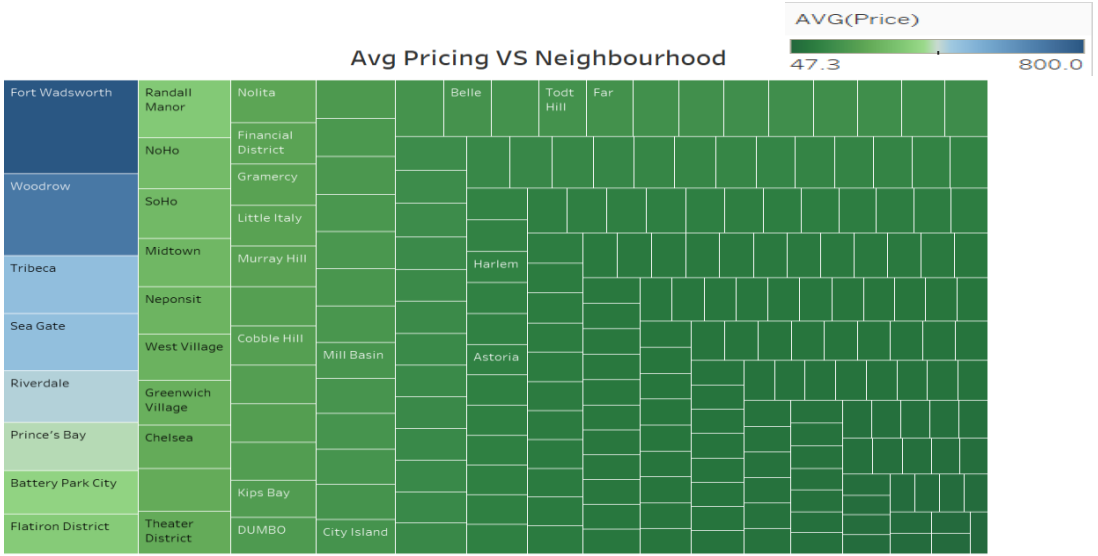
Analysis – Neighborhood



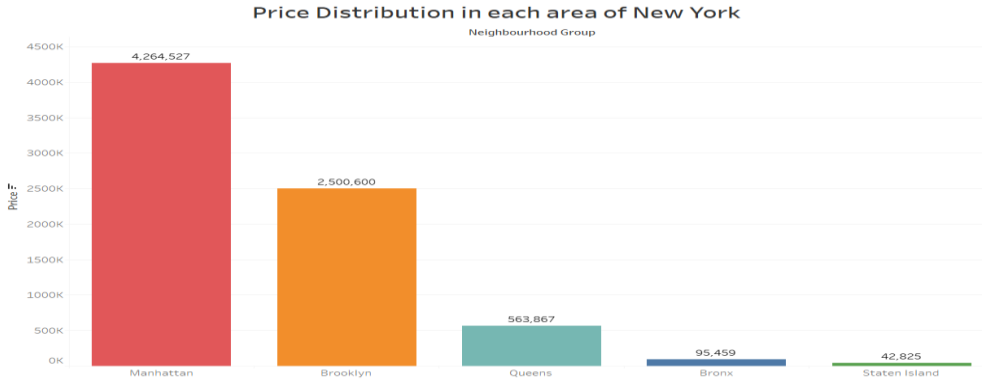
neighbourhood_group	
Manhattan	44.301053
Brooklyn	41.116679
Queens	11.588097
Bronx	2.231312
Staten Island	0.762859

Observation:

85% of the listing are Manhattan and Brooklyn neighbourhood_group

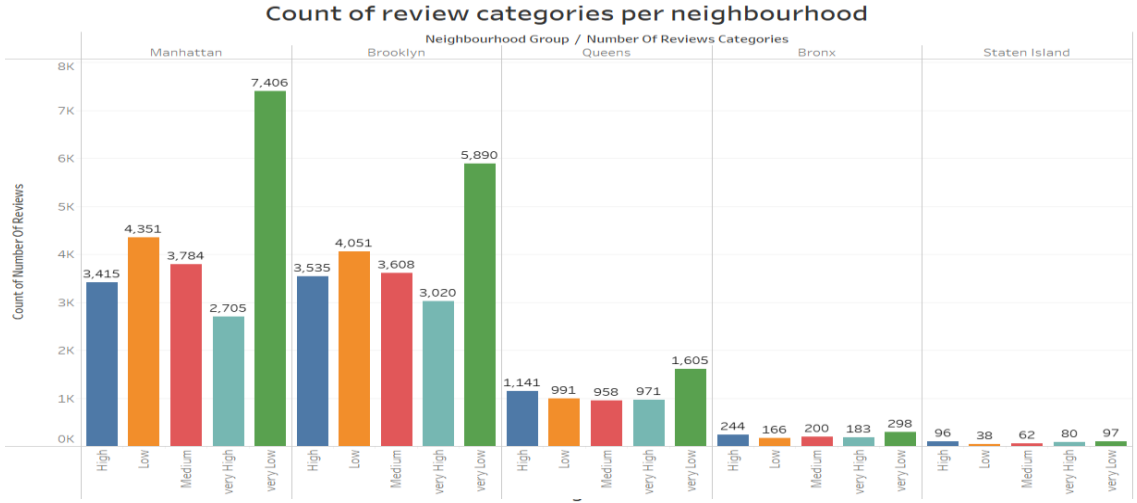


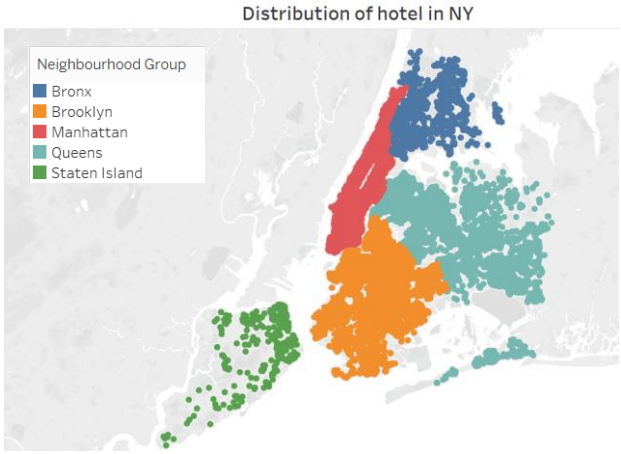
Average pricing is higher than average in only 6 neighborhoods

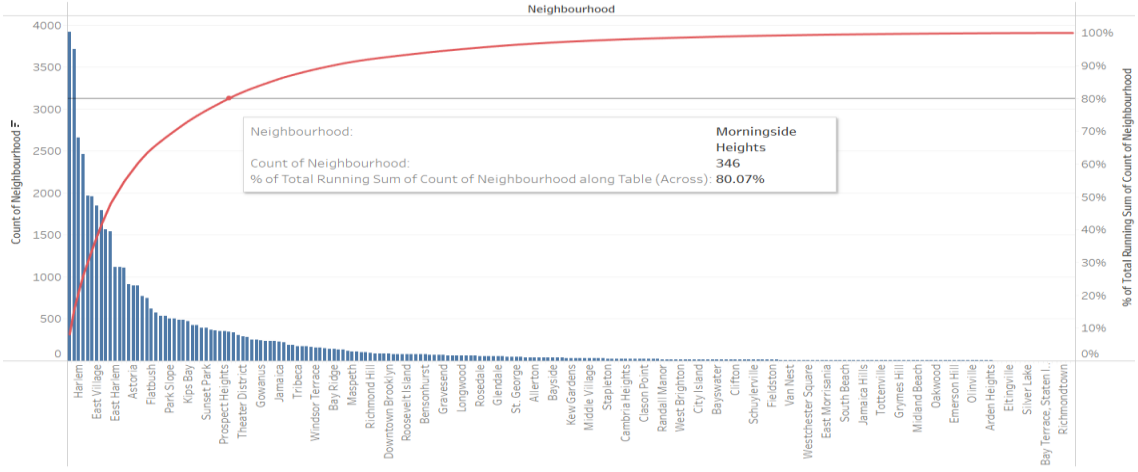


Manhattan and Brooklyn are more expensive than other neighborhood groups

Analysis – Neighborhood Contd.

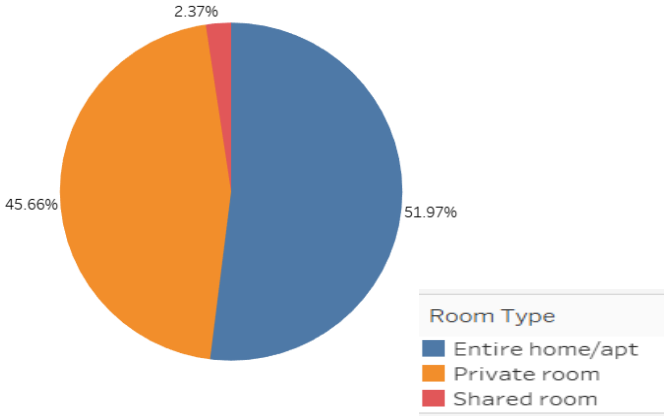


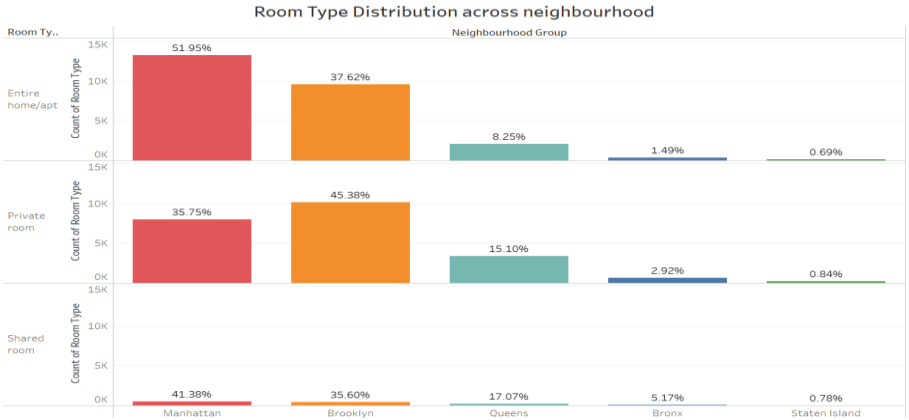


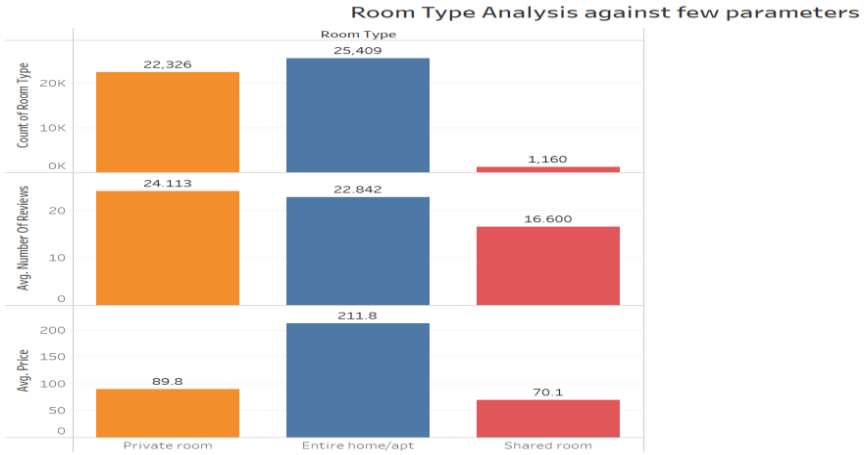


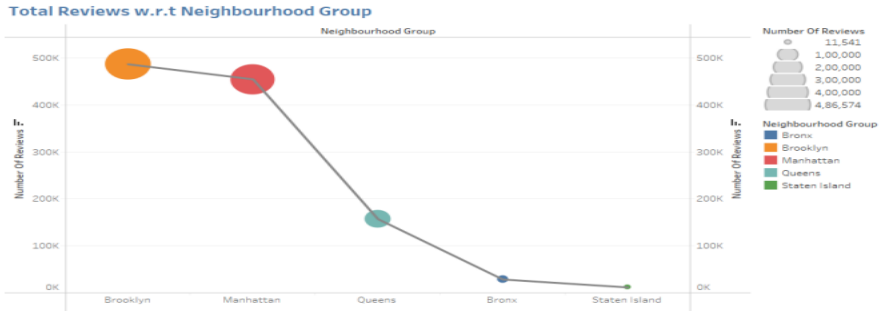
- 346 neighborhoods make up 80% of the entire neighborhoods
- Lower reviews are observed across Manhattan and Brooklyn
- Airbnb's footprint appears dominant in Manhattan, Brooklyn, and Queens. These boroughs boast the highest number of listings, likely due to their high population density and central roles as NYC's financial and tourism hubs.
- Conversely, Staten Island sees minimal Airbnb presence. With its lower population density and fewer tourist attractions, listings dwindle to around 1%, reflecting a less vibrant market for short-term rentals.

Analysis – Room Types



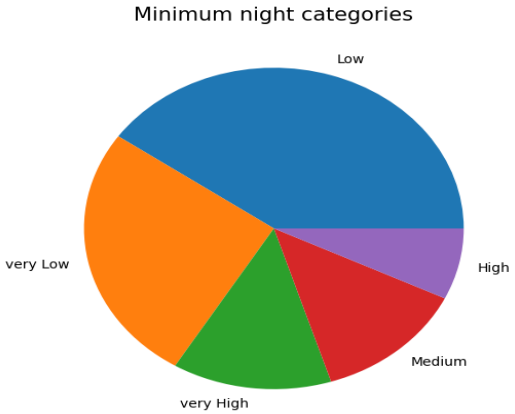




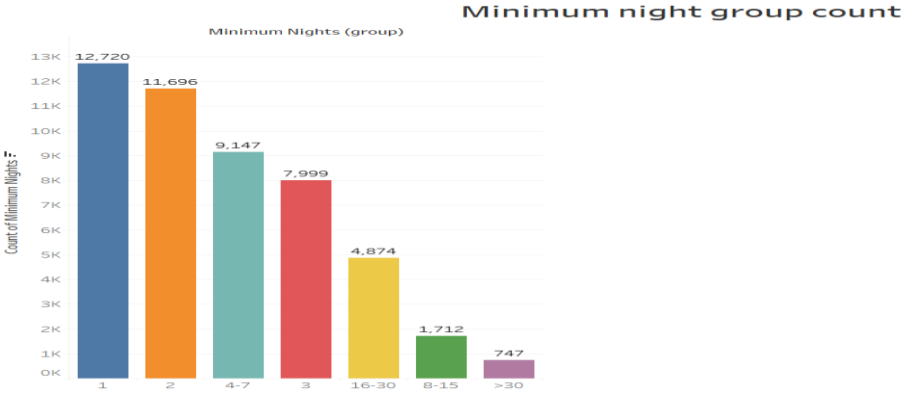


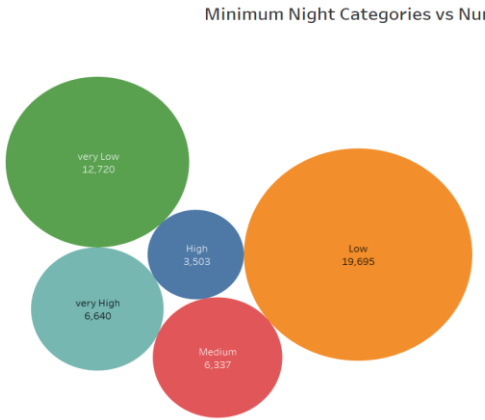
- Private rooms and Entire home/apt take up 97% of the room types
- Manhattan and Brooklyn have more Pvt rooms and Entire home/apt
- Pvt rooms and Entire home/apt have more rooms, reviews and are pricier than shared rooms
- The popularity of Manhattan and Brooklyn for bookings is reinforced by positive customer reviews, highlighting their appeal.

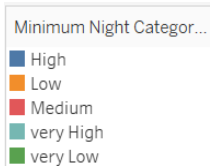
Analysis – Minimum Night Categories

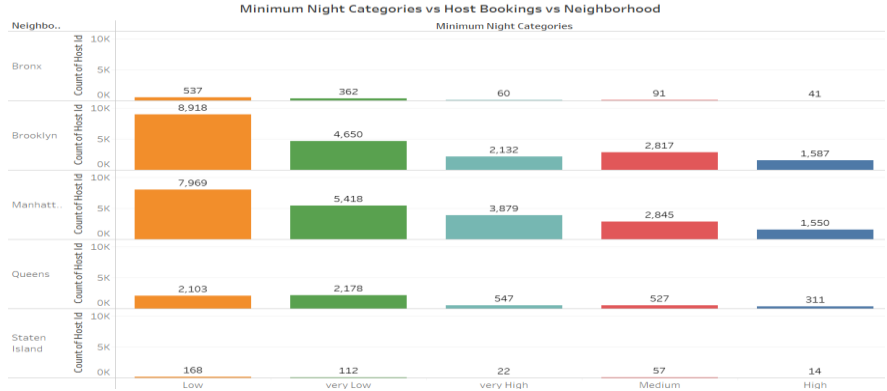


minimum_night_categories	
Low	40.280192
very Low	26.014930
very High	13.580121
Medium	12.960425
High	7.164332



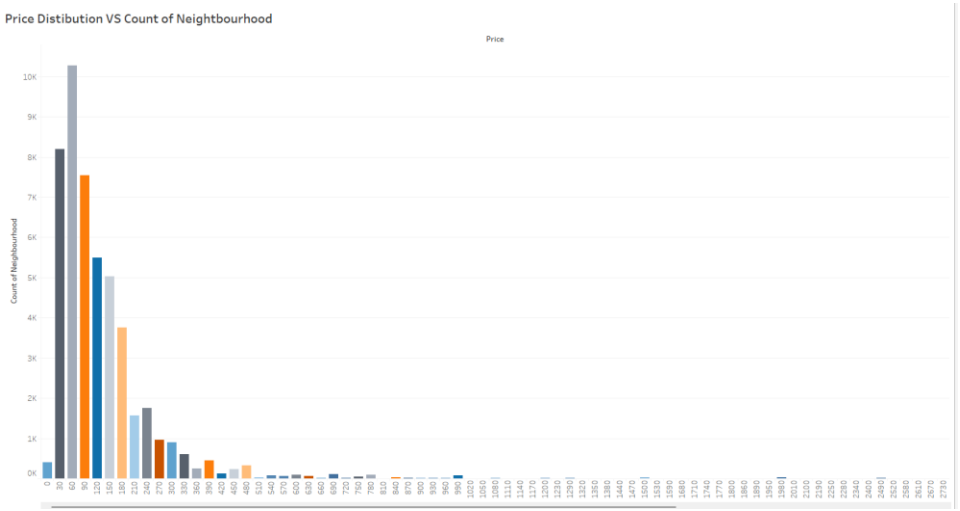


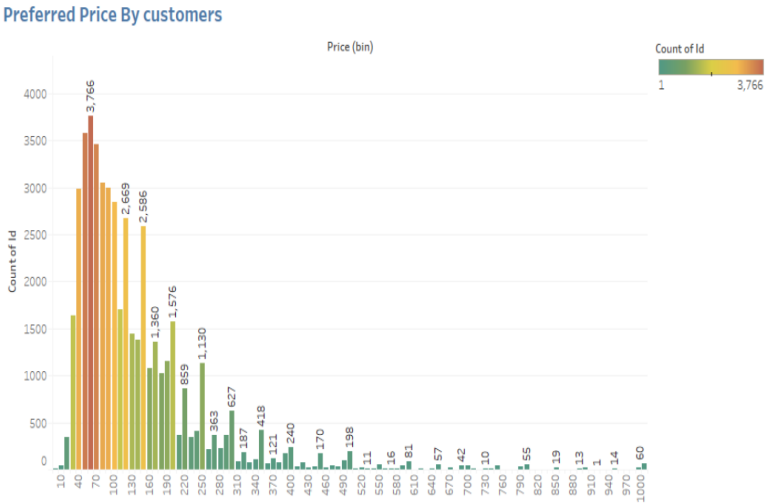


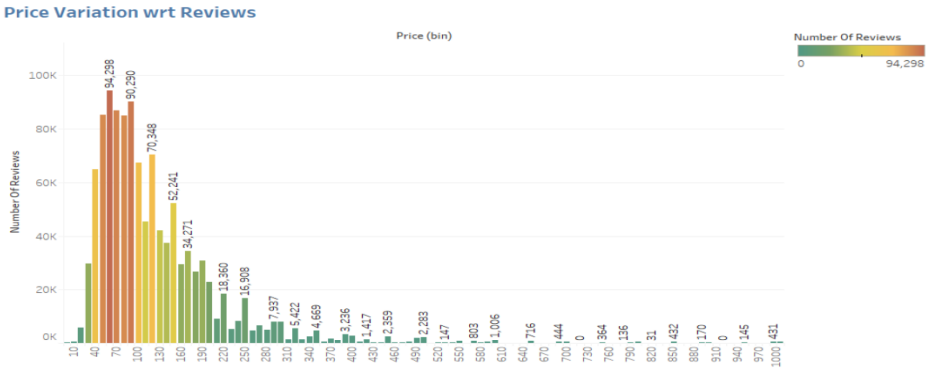


- Lower category (1-6 days) minimum nights take up 66% of the bookings, , suggesting a preference among customers for shorter stays.
- Reviews for lower category minimum nights is more
- There is a noticeable surge in bookings at the 30-day mark (13% for very high category), indicating a trend of customers renting on a monthly basis.
- Manhattan and Brooklyn stand out with a higher number of 30-day bookings compared to other areas. Possible explanations include tourists opting for extended stays or mid-level employees choosing budget-friendly options for company visits.

Analysis – Price







Pricing Preference:

The analysis considers two parameters: volume of bookings in a price range and the number of reviews in that range.

Both graphs indicate that the most favorable price range for customers is $40 - $190. This range aligns with the majority of bookings and also receives a significant number of positive reviews, indicating high customer satisfaction.

Recommendation:

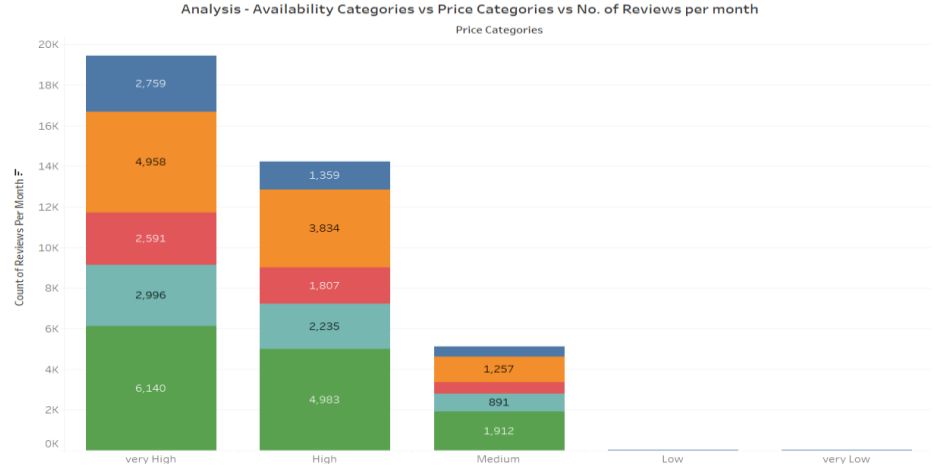
Expansion and Acquisition Strategy:

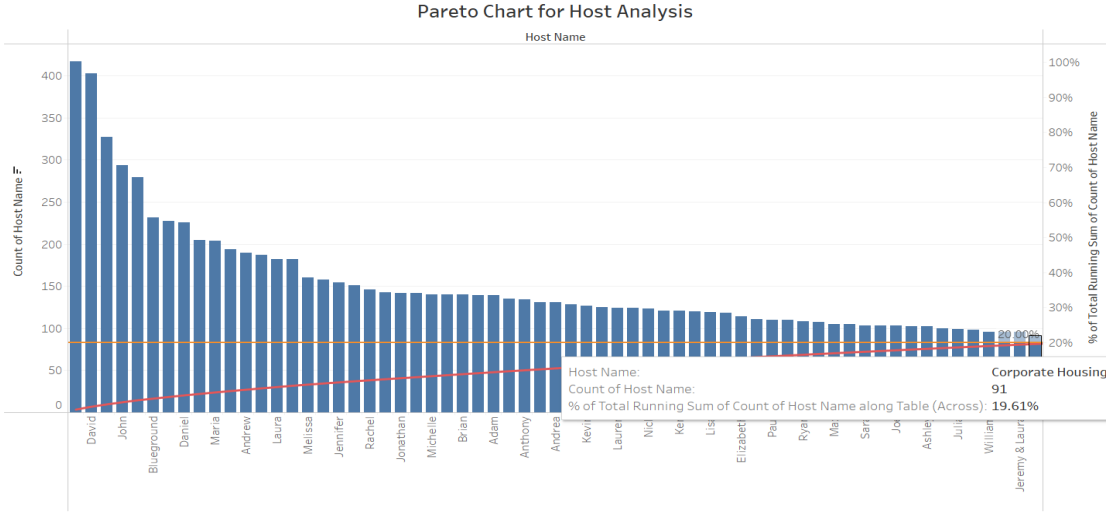
It is recommended to focus on new acquisitions and expansion within the price range of $40 - $190.

This range demonstrates a balance between attracting a substantial volume of customer traffic and ensuring high customer satisfaction.

Investing in this price range is likely to yield positive results in terms of both booking volume and customer experience, contributing to overall business success.

Analysis – Remaining Features





- If the combination of availability and price is very high, reviews per month will be low on average.
- Very high availability and very low price are likely to get more reviews.
- Lower categories for availability and price will get lesser reviews.
- Top 91 hosts make up only 20% of the bookings

RECOMMENDATIONS

Neighborhood

- 85% of the listings are in Manhattan and Brooklyn
- Average pricing is higher than average in only 6 neighborhoods
- Manhattan and Brooklyn have more expensive offerings
- 346 neighborhoods make up 80% of the entire neighborhoods

Room Types

- Private rooms and Entire home/apt take up 97% of the room types
- Manhattan and Brooklyn have more Pvt rooms and Entire home/apt
- Initiate targeted promotional campaigns and offer discounts to boost bookings for shared rooms by highlighting affordability.
- Prioritize acquiring more hosts and listings offering monthly rental durations, especially in Manhattan and Brooklyn where there is a higher demand for such extended stays.
- Tap into the market of customers requiring short-term rentals for quarantine purposes. Consider offering weekly or bi-weekly rentals to accommodate individuals stranded in NYC

Minimum Night Categories

- Lower category minimum nights take up 66% of the bookings
- Reviews for lower category minimum nights is more
- AirBnB should promote lowering of minimum nights bookings to increase its revenue

Additional Observations

- Top 91 hosts make up only 20% of the bookings so all guests should be given equal importance
- Plan new acquisitions in the price range of $40 - $190. This range, identified through analysis of both booking volume and customer reviews, signifies a sweet spot that attracts substantial customer traffic while ensuring high satisfaction levels.
- Consider acquiring more 'private rooms' in Manhattan and Brooklyn, and 'entire homes' in Bronx and Queens. This targeted approach aims to meet the specific preferences and demands of customers in different regions.
- With an average price of $124 and potential for growth, prioritize expansion efforts in Brooklyn. This recommendation takes into account the existing saturation in Manhattan and identifies Brooklyn as a viable location for acquiring new listings and accommodating diverse customer preferences.
- Increase acquisitions in coastal regions to diversify the property portfolio and attract customers seeking accommodations with scenic views or proximity to waterfronts.

APPENDIX – Data Methodology

Data Analysis

Data loading and analysis using Python

Binning of continuous variables

Categorization of numerical columns for EDA

EDA

Univariate, Bivariate Analysis and Correlations identified using Python

Missing Value Analysis

Missing value analysis using Python

Data Visualizations

Used Tableau for building graphs for analysis and identifying relationships and correlations between variables

APPENDIX – Data Methodology Contd.

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

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

THANK YOU