Airbnb Case Study NYC

By Prithibi Mondal

Objective:

1 Goal

To conduct a thorough analysis of Airbnb dataset

2 Revenue Loss

Airbnb experienced significant revenue losses during the COVID-19 pandemic.

3 Recovering Business

With travel resuming, Airbnb is now focused on reviving its business and is prepared to offer services to its customers once more.

Data Preparation

1

Data Cleaning

Cleaned data to remove any missing values and duplicates.

2

Outlier Identification

Identified outliers

3

Data Visualization

Visualize the data through Tableu and Python

Importing Libraries and Reading the Data:

[1]:
```
#importing necessary libraries
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

[2]:
```
df = pd.read_csv("AB_NYC_2019.csv")
df.head(5)
```

[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	19-1
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	21-0
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	05-0
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	19-1

Data Types and Missing Values:

```
[3]: df.columns

[3]: Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')

[4]: df.shape

[4]: (48895, 16)

[5]: df.isnull().sum()

[5]: 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
```

Handling the Missing Values:

```
[6]: df.drop(['id','name','last_review'], axis = 1, inplace = True)
```

```
[7]: df.head(5)
```

[7]:

	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calcula
0	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	0.21	
1	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	0.38	
2	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	
3	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	4.64	
4	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	0.10	

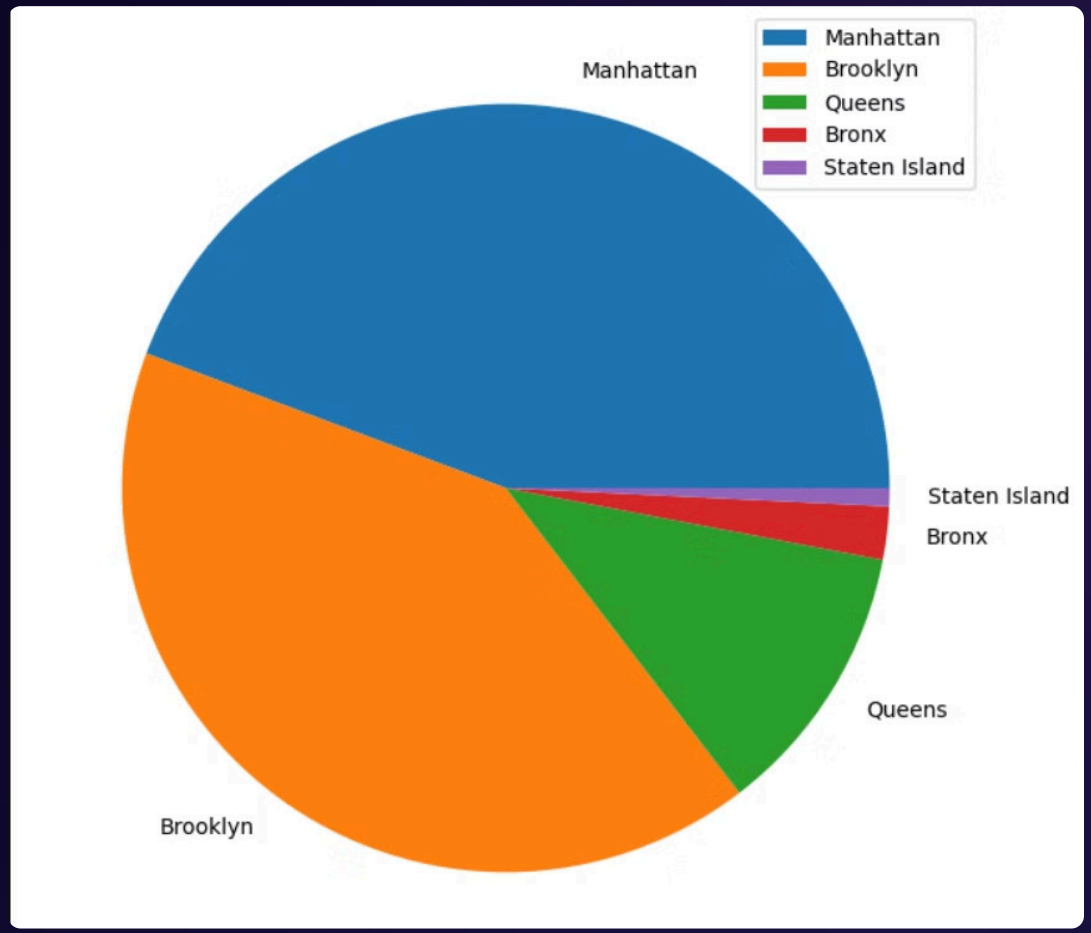
```
[8]: df.fillna({'reviews_per_month':0},inplace=True)
```

```
[9]: df.room_type.unique()
```

```
[9]: array(['Private room', 'Entire home/apt', 'Shared room'], dtype=object)
```

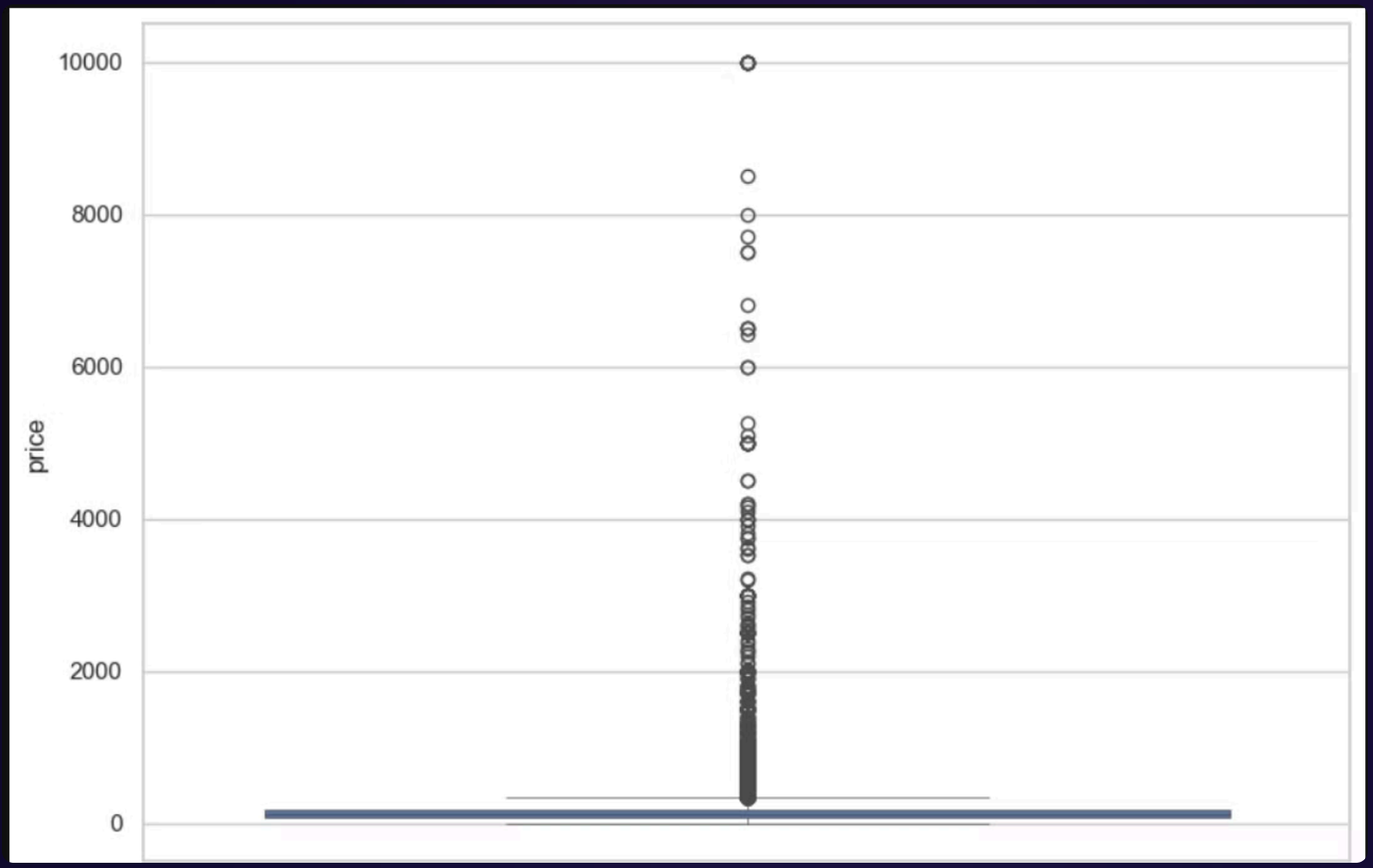
Neighbourhood

```
plt.figure(figsize=(8,8))
plt.pie(x = df2.neighbourhood_group.value_counts(normalize= True) * 100,labels = df2.neighbourhood_group.value_counts(normalize= True).index)
plt.legend()
plt.show()
```



Price Analysis:

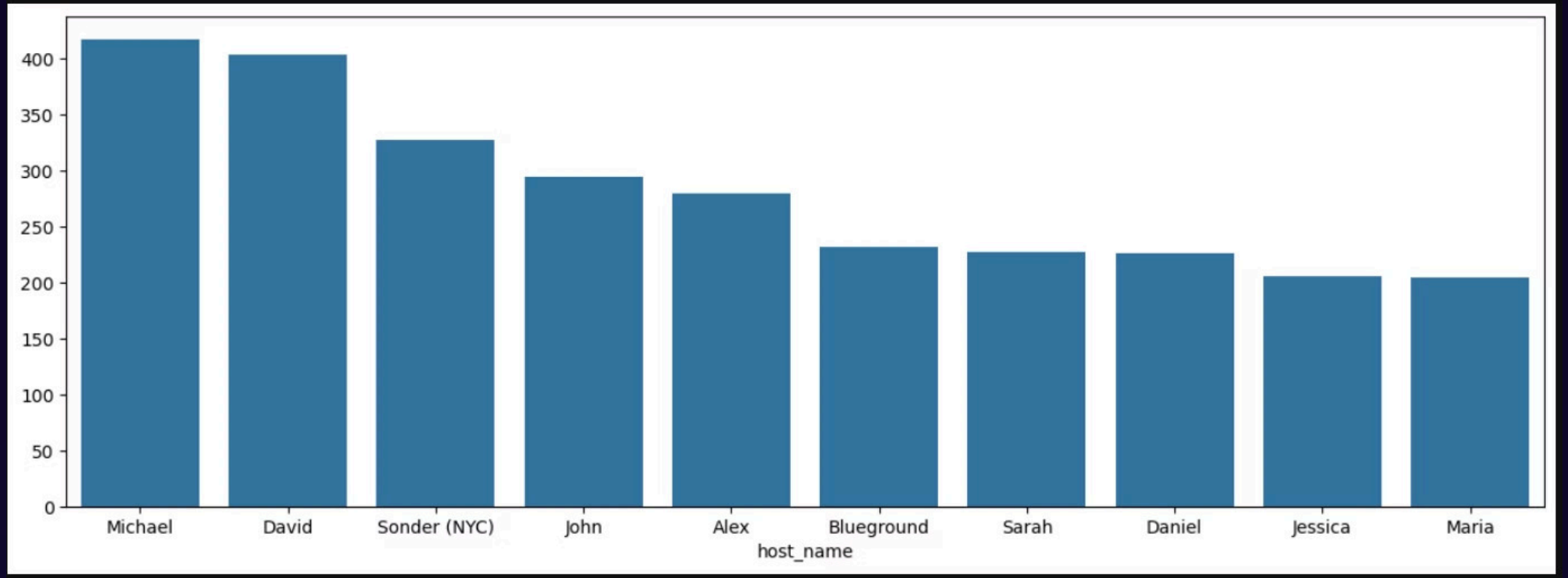
```
plt.figure(figsize=(10,7))
sns.set_theme(style="whitegrid")
#tips = sns.load_dataset("tips")
sns.boxplot(y = df2.price,width=0.8,
    dodge=True,
    fliersize=6,
    linewidth=.5,
    whis=1.5,
    color=None)
plt.show()
```



•Most of the outliers in Price column are for Brooklyn and Manhattan.

Top Host:

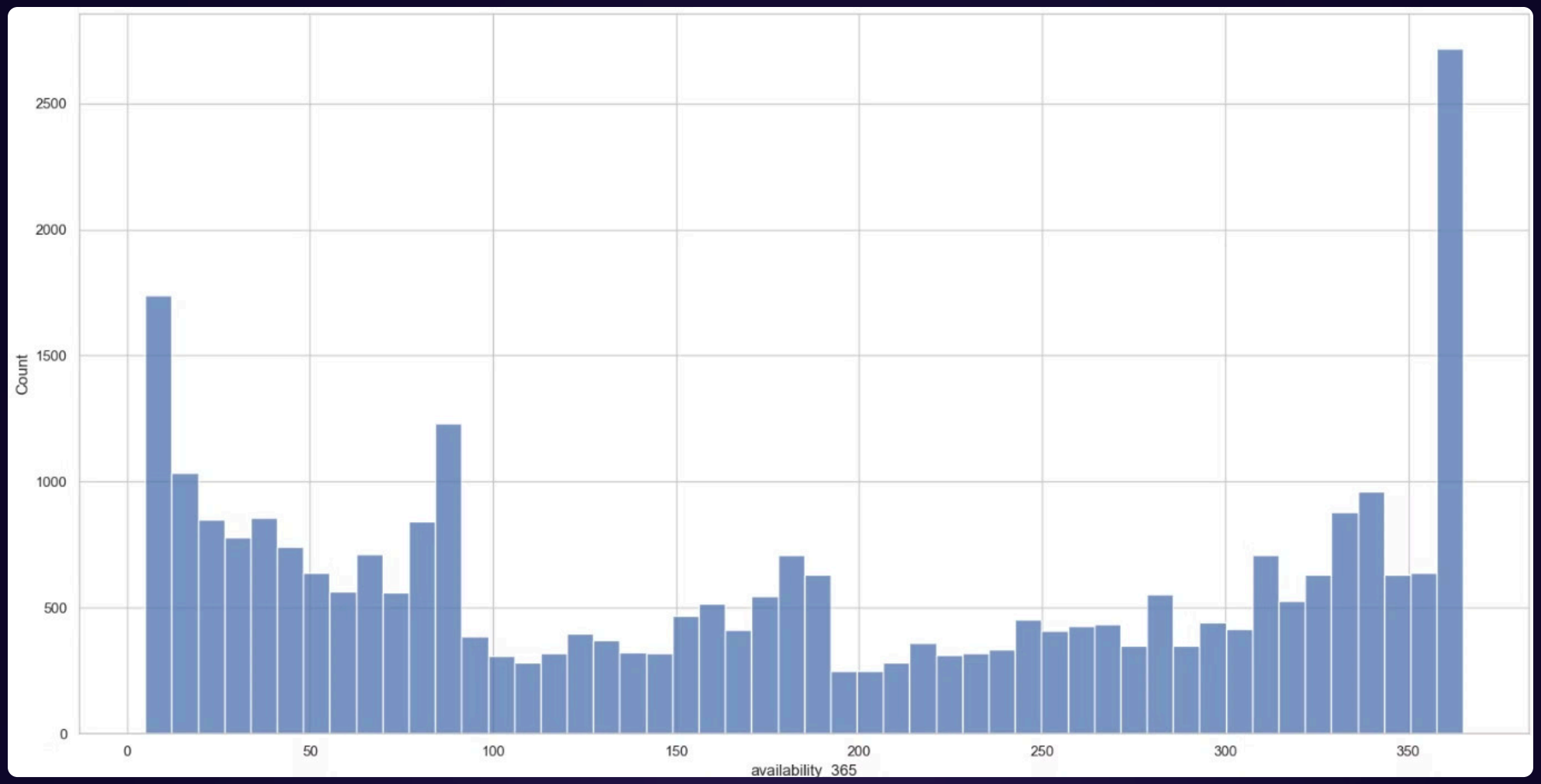
```
# Top 10 host's
plt.figure(figsize=(15,5))
sns.barplot(x = df2.host_name.value_counts().index[:10] , y = df2.host_name.value_counts().values[:10])
plt.show()
```



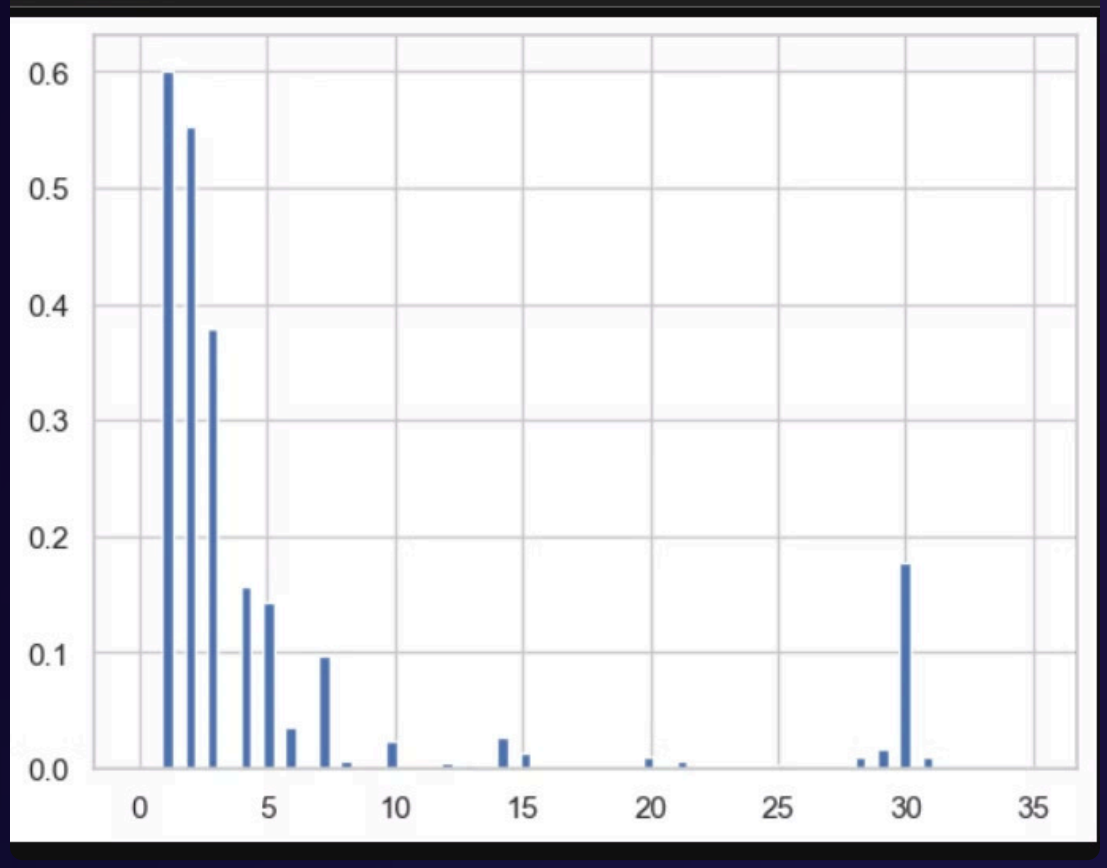
Review Analysis:

```
plt.figure(figsize=(70,8))
sns.boxplot(data = df2, y = 'room_type' ,x = 'reviews_per_month')
plt.xticks(np.arange(0,100,.5))
plt.show()
```

Availability and Minimum Nights:




```
plt.hist(data = df2, x = 'minimum_nights',bins=80,range=(0,35),density=True)
plt.show()
```



Conclusion

1 Revenue Impact

Airbnb suffered major financial losses during the COVID-19 pandemic as global travel came to a halt, drastically reducing demand for accommodations.

2 Business Objective

Airbnb's primary goal is to provide a platform for individuals to rent out unused properties, offering travelers unique and flexible lodging options.

3 Recovery Efforts

As travel resumes, Airbnb is actively working to regain its market presence and is prepared to offer services to meet the increasing demand for accommodations.

Appendix- Data Sources

The columns in the dataset are self-explanatory. You can refer to the diagram given below to get a better idea of what each column signifies.

Note: The price column contains the price/night.

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

Variable Categories:

Variables can be classified into four main types: **categorical, numeric, location, and time**. Choosing the right plot depends on the variable type. **Categorical** variables are best visualized with bar or pie charts, while **numeric** variables suit histograms, box plots, and scatter plots. **Location** variables work well with maps and geospatial plots, and **time** variables are effectively represented with line charts or time series plots. Understanding these distinctions helps in selecting the most appropriate visualization for the data.

```
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 Methodology:

- **Data Collection**: Gather Airbnb data
- **Data Cleaning**: Clean the data using Python (pandas), handling missing values, duplicates, and converting data types (e.g., dates, prices).
- **Exploratory Data Analysis (EDA)**: Use Python (matplotlib, seaborn) to analyze trends in pricing, occupancy rates, reviews, and booking patterns.
- **Tableau Visualization**: Import the cleaned data into Tableau to create interactive dashboards visualizing booking trends, revenue changes, and customer behavior.
- **Insights & Recommendations**: Use the visualizations to highlight key findings, such as Airbnb's recovery trends, pricing strategies, and customer preferences, and provide actionable recommendations.