International Institute of Information Technology Bangalore

upGrad

STORYTELLING CASE STUDY : AIRBNB NYC

NAME : PRITKUMAR PARMAR

AGENDA

- OBJECTIVE
- DATA LIFE-CYCLE
- ANALYSIS METHODS
- CONCLUSION and RECOMMANDATIONS
- APPENDIX

OBJECTIVE

- To handle an analysis of NewYork AirBNB dataset.
- Detect meaningful insights from dataset
- Prepare data to do Data Visulisation and extract important insights.

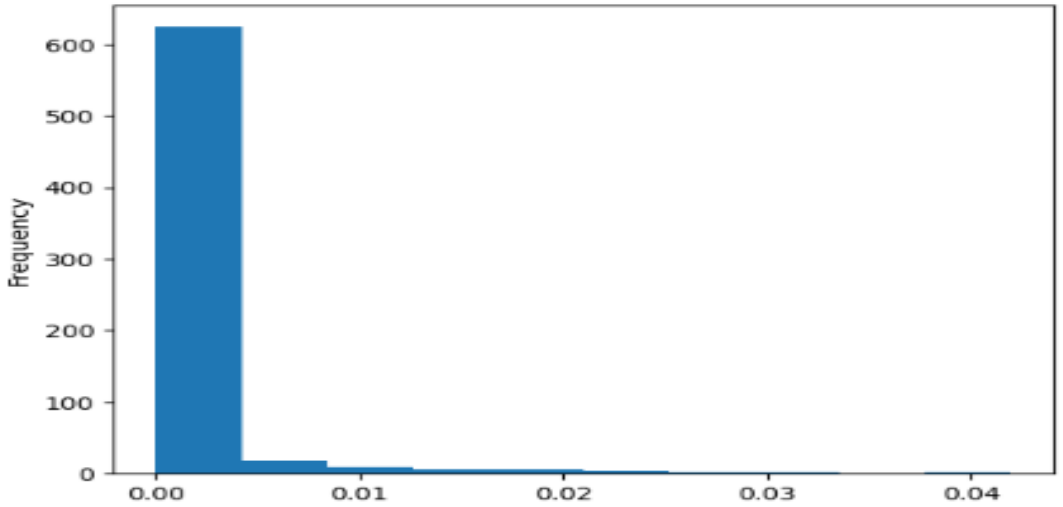
BACKGROUND STEPS

- In first stage, load the data and import the necessary libraries
- In second stage, Clean the data and do the needful opeartions for null values
- In the third stage, Prepare the data for Data Visulisation/EDA
- In fourth stage, extract meaningful insights from data

UNIVARIATE ANALYSIS

6.7 Price

In [510]:
```
inp0.price.value_counts(normalize=True).plot.hist()
plt.show()
```



In [511]:
```
sns.distplot(inp0.price,kde=True)
plt.show()
```

```
C:\Users\Administrator\AppData\Local\Temp\ipykernel_15496\4203382358.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(inp0.price,kde=True)
```

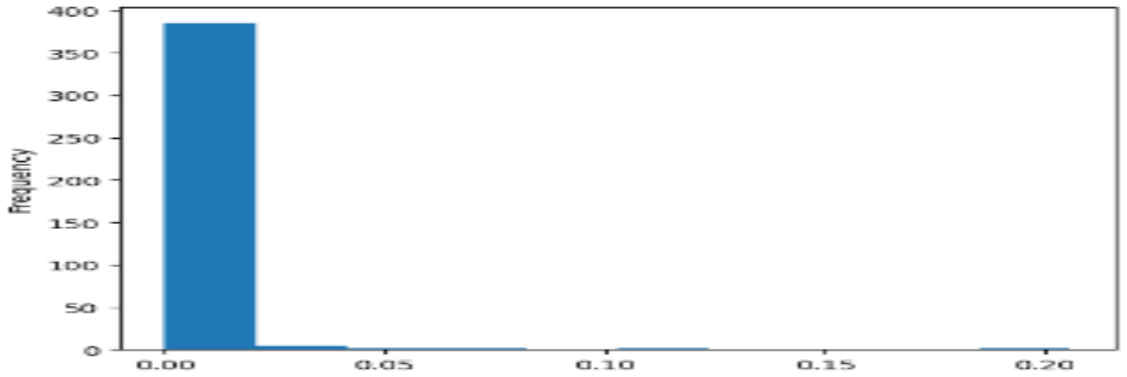
ANALYSIS OF NO. OF REVIEW FEATURE

6.9 NUMBER OF REVIEWS

In [S16]:
```
inp0.number_of_reviews.describe()
```

Out[S16]:
```
count     48895.000000
mean         23.274466
std          44.550582
min           0.000000
25%           1.000000
50%           5.000000
75%          24.000000
max         629.000000
Name: number_of_reviews, dtype: float64
```

In [S17]:
```
inp0.number_of_reviews.value_counts(normalize=True).plot.hist()
plt.show()
```



In [S18]:
```
sns.distplot(inp0.number_of_reviews)
plt.show()
```

```
C:\Users\Administrator\AppData\Local\Temp\ipykernel_15496\4258575927.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(inp0.number_of_reviews)
```

ANALYSIS OF ROOM TYPE FEATURE

6.6 ROOM TYPE

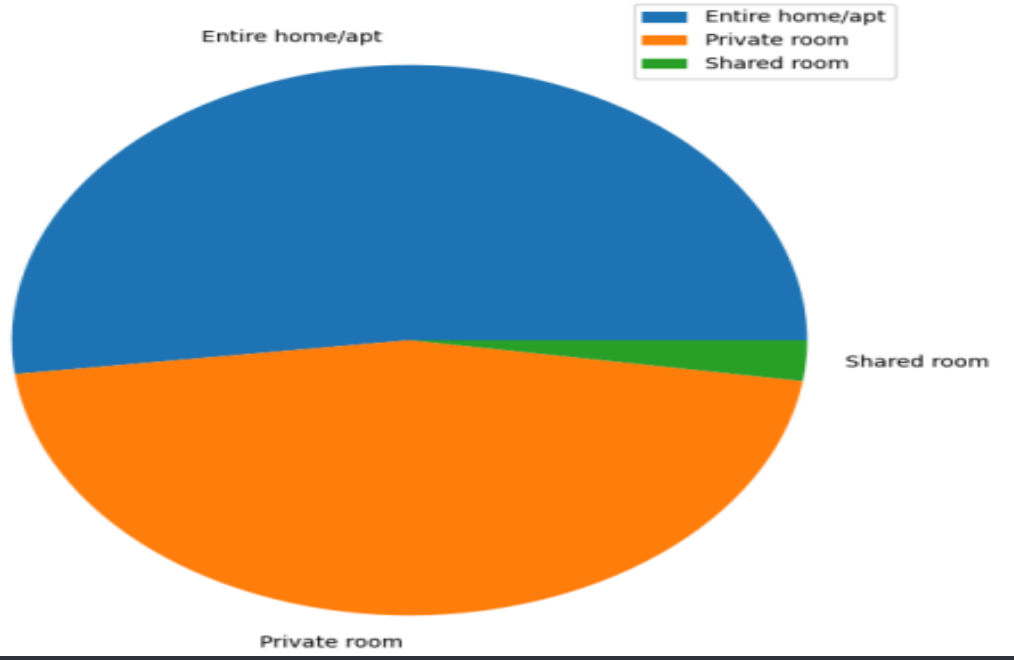
In [507…
```
inp0.room_type.value_counts()
```

Out[507]:
```
Entire home/apt     25409
Private room        22326
Shared room          1160
Name: room_type, dtype: int64
```

In [508…
```
inp0.room_type.value_counts(normalize=True)*100
```

Out[508]:
```
Entire home/apt     51.966459
Private room        45.661111
Shared room          2.372431
Name: room_type, dtype: float64
```

In [509…
```
plt.figure(figsize=(8,8))
plt.pie(x = inp0.room_type.value_counts(normalize= True) * 100,labels = inp0.room_type.value_
plt.legend()
plt.show()
```

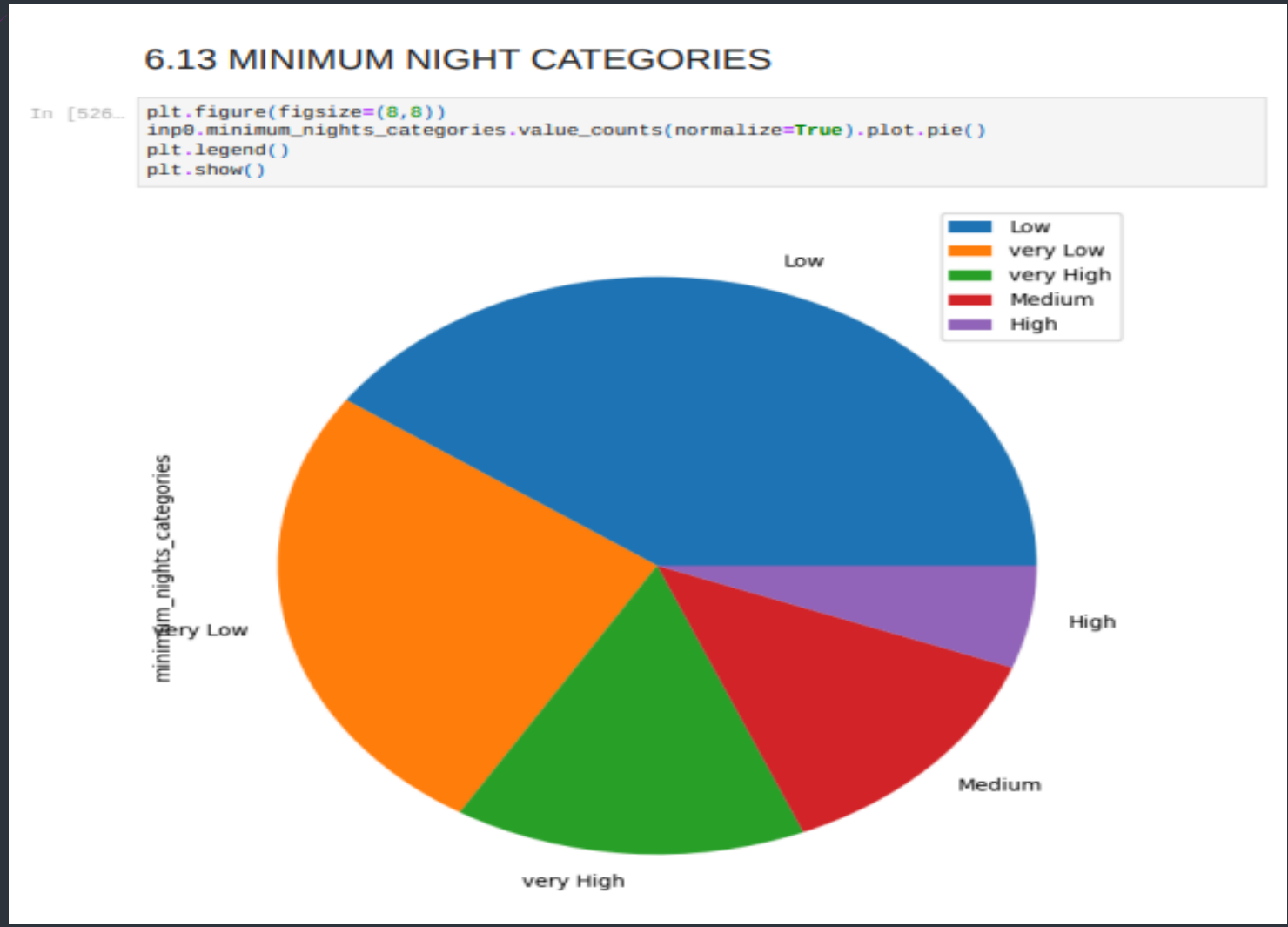


ANALYSIS OF MINIMUM NIGHTS SPENT FEATURE

6.13 MINIMUM NIGHT CATEGORIES

In [526…

```
plt.figure(figsize=(8,8))
inp0.minimum_nights_categories.value_counts(normalize=True).plot.pie()
plt.legend()
plt.show()
```



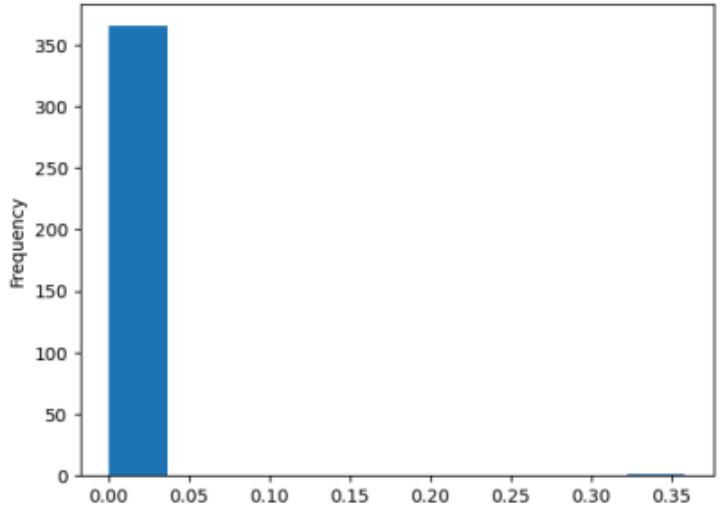
ANALYSIS OF AVAILABILUTY_365 FEATURE

6.11 AVAILABILITY 365

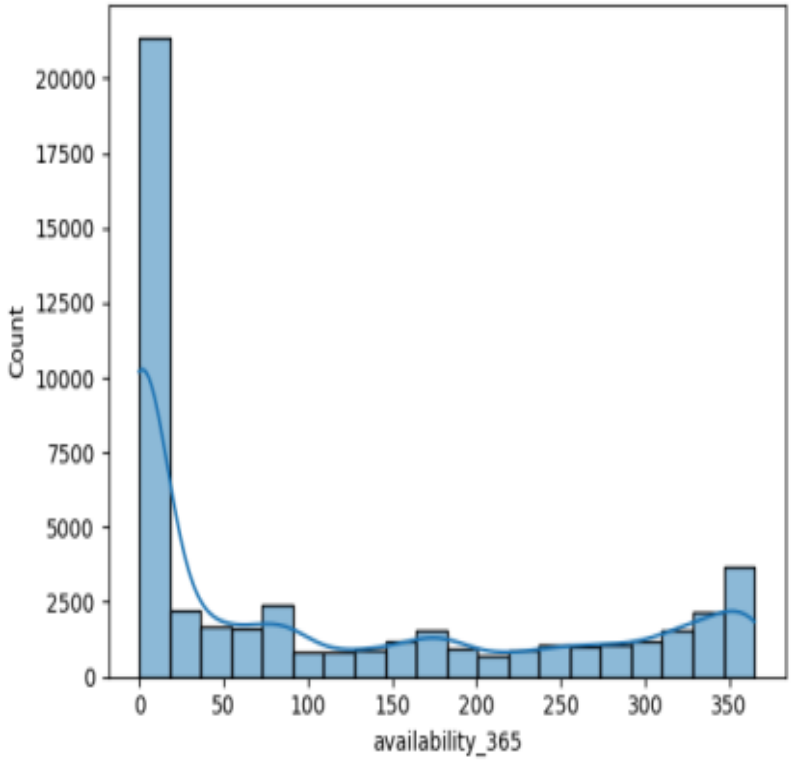
In [522…
```
inp0.availability_365.describe()
```

Out[522]:
```
count    48895.000000
mean       112.781327
std        131.622289
min          0.000000
25%          0.000000
50%         45.000000
75%        227.000000
max        365.000000
Name: availability_365, dtype: float64
```

In [523…
```
inp0.availability_365.value_counts(normalize=True).plot.hist()
plt.show()
```



In [524…
```
sns.histplot(inp0.availability_365,bins = 20,kde = True)
plt.show()
```



BIVARIATE/MULTIVARIATE ANALYSIS

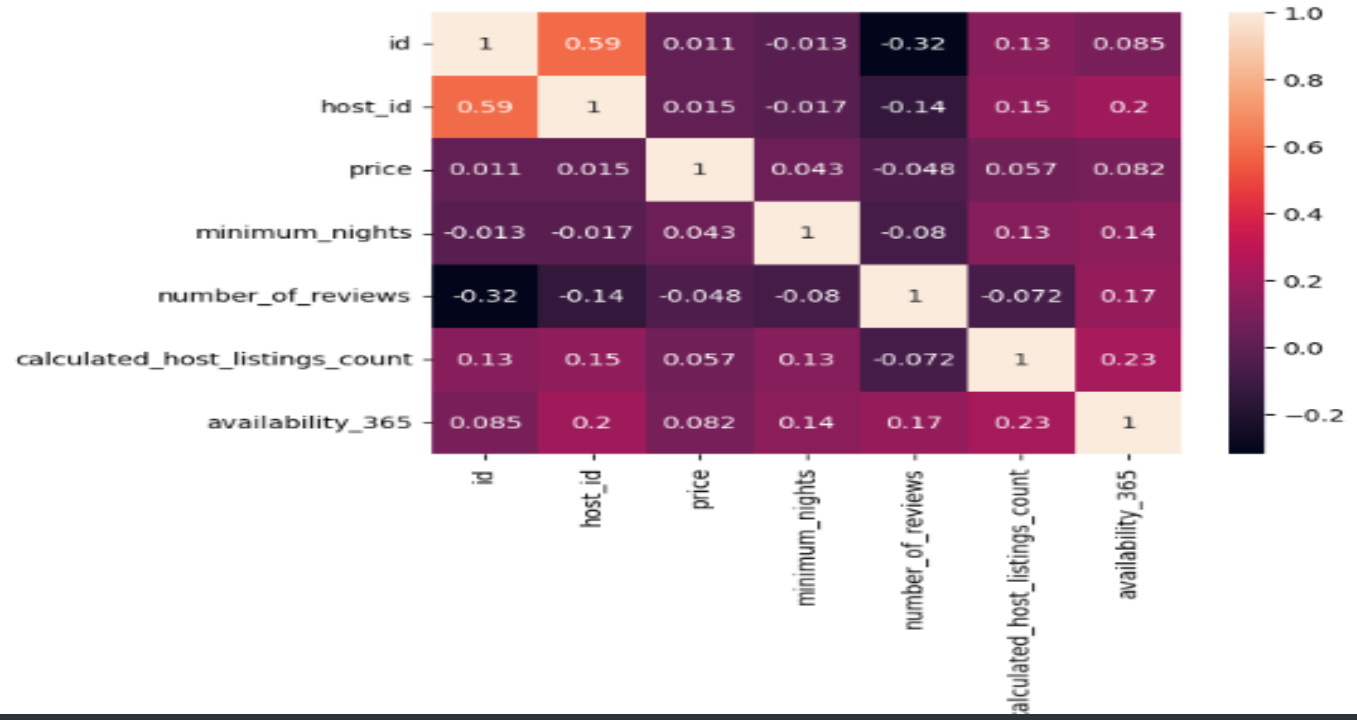
7. BIVARIATE AND MULTIVARIATE ANALYSIS

In [528]:
```
inp0[numerical_columns].head()
```

Out[528]:

	id	host_id	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
0	2539	2787	149	1	9	6	365
1	2595	2845	225	1	45	2	355
2	3647	4632	150	3	0	1	365
3	3831	4869	89	1	270	1	194
4	5022	7192	80	10	9	1	0

In [529]:
```
res = inp0[numerical_columns].corr()
sns.heatmap(res, annot=True)
plt.show()
```



CONCLUSION AND RECOMMANDATION

- Strong significant insights are derived based on various attributes in the dataset
- Data collection team should collect data about review scores so that it can strengthen the later analysis.
- Ample amount and variety of visuals have can used in the presentations for the stake-holders.
- A clustering machine learning model to identify groups of similar objects in datasets with two or more variable quantities can be made.

APPENDIX : DATA SOURCE

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

APPENDIX : FEATURE DATATYPE

```
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
```


Thank You !