Methodology

EDA :-

1. First the data set was imported in excel , the row and columns formatting was done.
2. Then for basic data cleaning , missing values, outliers treatment was done.
3. Name – 16, host_name – 21, last_review and review_month – 10052 values were mssing , deleting certain rows were names were missing as its very low compared to total rows in data set
4. The date columns were imputed with mode method.
5. 5. Then outliers fixing was done price, min_nights , no of reviews, availability, calculated hostings has certain values which are extremely high.
6. 6. Applied quartile cut off methods and set a cut off < 98% percentile and deleted the values to make it a normal distribution.
7. Did descriptive analysis
8. 8. Certain data table which are there in visuals are extracted from Excel pivot.
9. 9 . All the visual were made on tableau desktop.
10. 10 . The python script link in provided for better reference.

Python codes used to cleaning and EDA:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings("ignore")
```

```
airbnb = pd.read_csv("AB_NYC_2019.csv")

airbnb.info()
```

missing values :-

```
missing_values = []

for i in airbnb.columns:

    missing_values.append(len(airbnb[airbnb[i].isnull()]))



list2 = missing_values

## there are last_review and reviews per month has max missing

for i , j in zip(list1, list2):

    print(i,"=",j,"\n")
```

```
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
```

```
airbnb["last_review"].value_counts(normalize = True)*100
```

```
airbnb.describe(percentiles = [0.80,0.90,0.95,0.99])
```

Outliers test

```
plt.figure(figsize=(8,20))

plt.boxplot(airbnb.price)
```

```
list1 = ["minimum_nights" , "number_of_reviews" , "reviews_per_month" , "calculated_host_listings_count"]
```

```
[plt.boxplot(i) for i in airbnb(list1)]
```

```
plt.figure(figsize=(20,8))

sns.boxplot(airbnb["minimum_nights"])
```

removing qutliers by percentile method :

```
max_thres = airbnb["minimum_nights"].quantile(0.99)


airbnb = airbnb[~(airbnb["minimum_nights"]>max_thres)]


airbnb = airbnb[~(airbnb["reviews_per_month"]>max_thres3)]


airbnb.groupby("host_name")["calculated_host_listings_count"].sum().head(30)
```

Then excel pivot was implemented