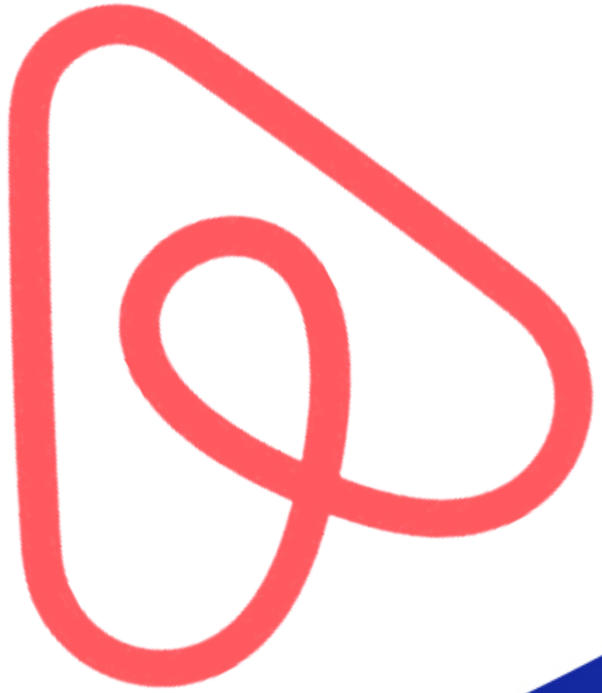
SAMSUNG

Airbnb price prediction



Supervised by :
Dr. Doaa Mahmoud & Eng. Haneen Eldaly

NOTEBOOK

Together for Tomorrow!
Enabling People
Education for Future Generations

Abdelrahman Raslan

Agenda

01 About Dataset

02 Problem statement

03 Data Preprocessing

04 EDA

05 Modelling and Data fitting

06 Business Solution

07 Deployment

01

About Dataset

Know more about data

What is Airbnb ?

Airbnb began in 2008 when two designers who had space to share hosted three travelers looking for a place to stay. Now, millions of hosts and travelers choose to create a free Airbnb account so they can list their space and book unique accommodations anywhere in the world. And Airbnb experience hosts share their passions and interests with both travelers and locals.

What is the dataset ?

Context

Since 2008, guests and hosts have used Airbnb to travel in a more unique, personalized way. As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in New York City

Content

The following Airbnb activity is included in this New York dataset:

Listings, including full descriptions and average review score Reviews, including unique id for each reviewer and detailed comments Calendar, including listing id and the price and availability for that day

Tables	Description
NAME	Title
host_identity_verifed	Hosts that have been verified by Airbnb
host_name	Name of the host/house owner
neighbourhood group	Boroughs
neighbourhood	Neighbourhood of the borough
lat	Latitude
long	Longitude
instant_bookable	Wheter you can book inmediately
cancellation_policy	Kind of cancelalation policy
room_type	Kind of property
Construction year	In which year it was built?
Price	Rental price
service_fee	Airbnb profit
Minimum nights	Minimum amount of stay
Number of reviews	How many people have qualified the property?
last review	Last time that has been qualified
reviews per month	Average number of reviews per month
review rate	Total average of reviews
calculated host listings count	Amount of guests
availability 365	number of days the property is available in the year.

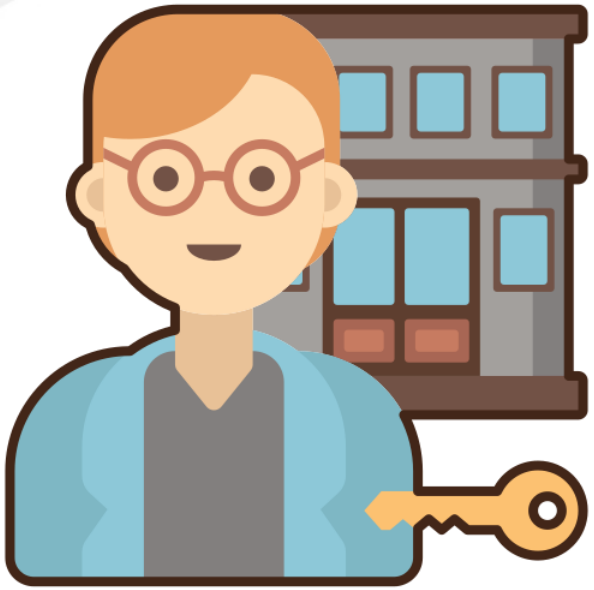
https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

02

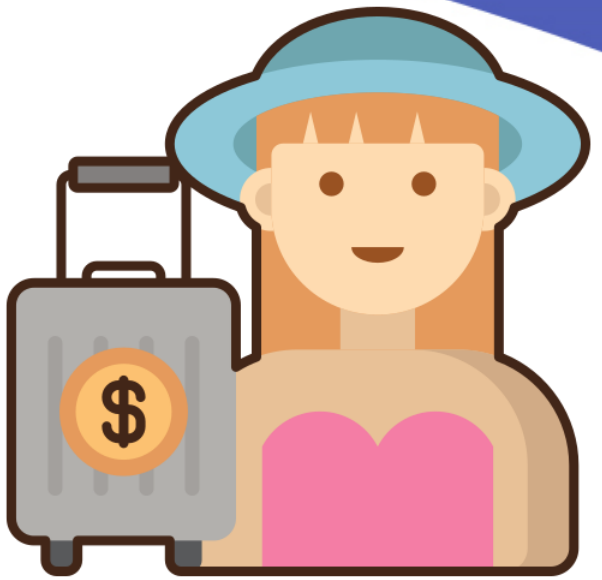
Problem Statements

The problems that society suffers from, and this data can solve them

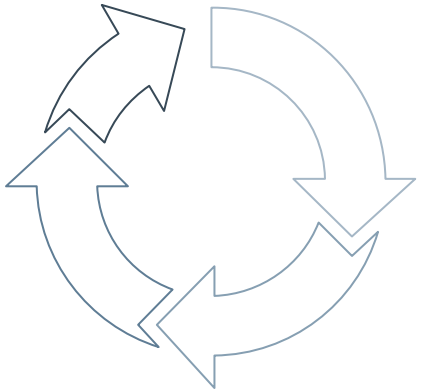
For Host & Guest



Pricing an Airbnb listing is still a challenging task for the host as there is a need to consider several features and amenities considering the amount of competition in the market.



Not all guests can search for this number of features, and the price is difficult to know, and what is important to the user is the price first



03

Data Preprocessing

Clean the data and make it suitable for the rest of the operations

Operation done on the Dataset

1. Remove duplicated columns and rows

2. Change in values type

- Convert 'last review' column to datetime type
- Convert price and services fee to numerical values

3. Removed nulls values from some columns

- 247 row deleted from price
- 250 row deleted from name
- 273 row deleted from host identity verified
- 29 row deleted from neighbourhood group
- 76 row deleted from cancellation policy
- 105 row deleted from instant bookable
- 182 row deleted from Construction year
- 390 row deleted from host name
- 14 row deleted from neighbourhood
- 7 row deleted from lat & long
- Total rows deleted --> 1,442
- remove availability 365 that less than 0 and more than 360
- remove minimum nights that less than 0

4. Deal with nulls value by fill with mean and median

- reviews per month → mean
- minimum nights → mean
- availability 365 → mean
- calculated host listings count → mean
- number of reviews → median
- review rate number→ median

After prepossessing

```
In [83]: airbnb.isnull().sum().sort_values()
```

```
Out[83]: id                                0
         calculated host listings count    0
         review rate number                0
         reviews per month                 0
         last review                       0
         number of reviews                 0
         minimum nights                    0
         service fee                       0
         price                             0
         Construction year                 0
         availability 365                  0
         room type                         0
         instant_bookable                  0
         long                              0
         lat                               0
         neighbourhood                     0
         neighbourhood_group               0
         host name                         0
         host_identity_verified            0
         host id                           0
         NAME                              0
         cancellation_policy               0
         availability_grp                  0
         dtype: int64
```


```
[10]: airbnb.shape
```

```
[10]: (102599, 26)
```

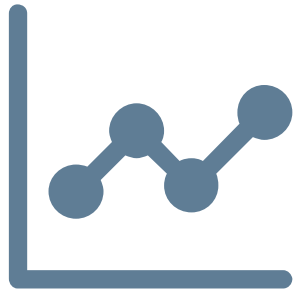
```
[94]: airbnb.shape
```

```
[94]: (100594, 23)
```

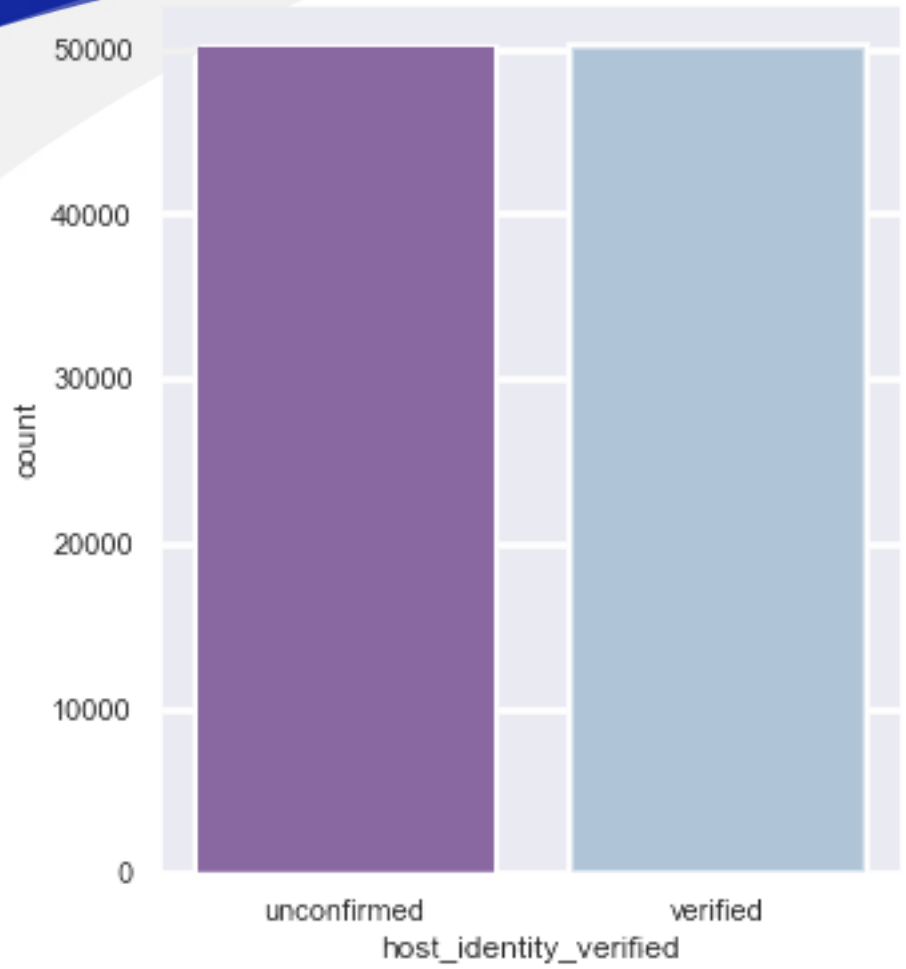
04

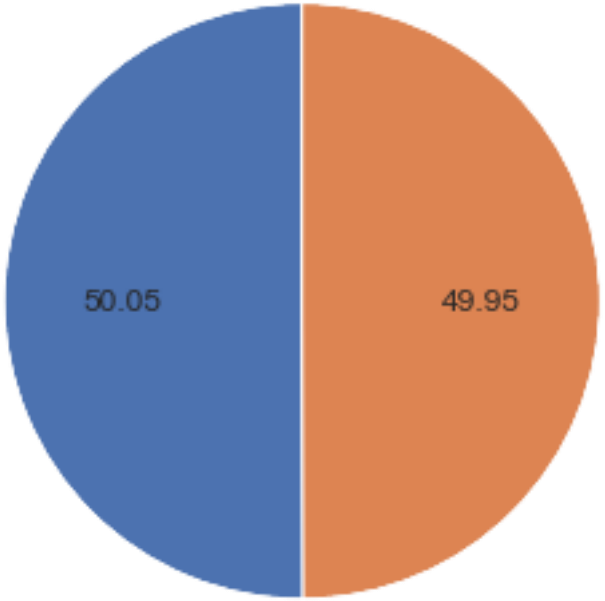
EDA

Exploratory Data Analysis



Number of verification and non-verification of the identity of the host

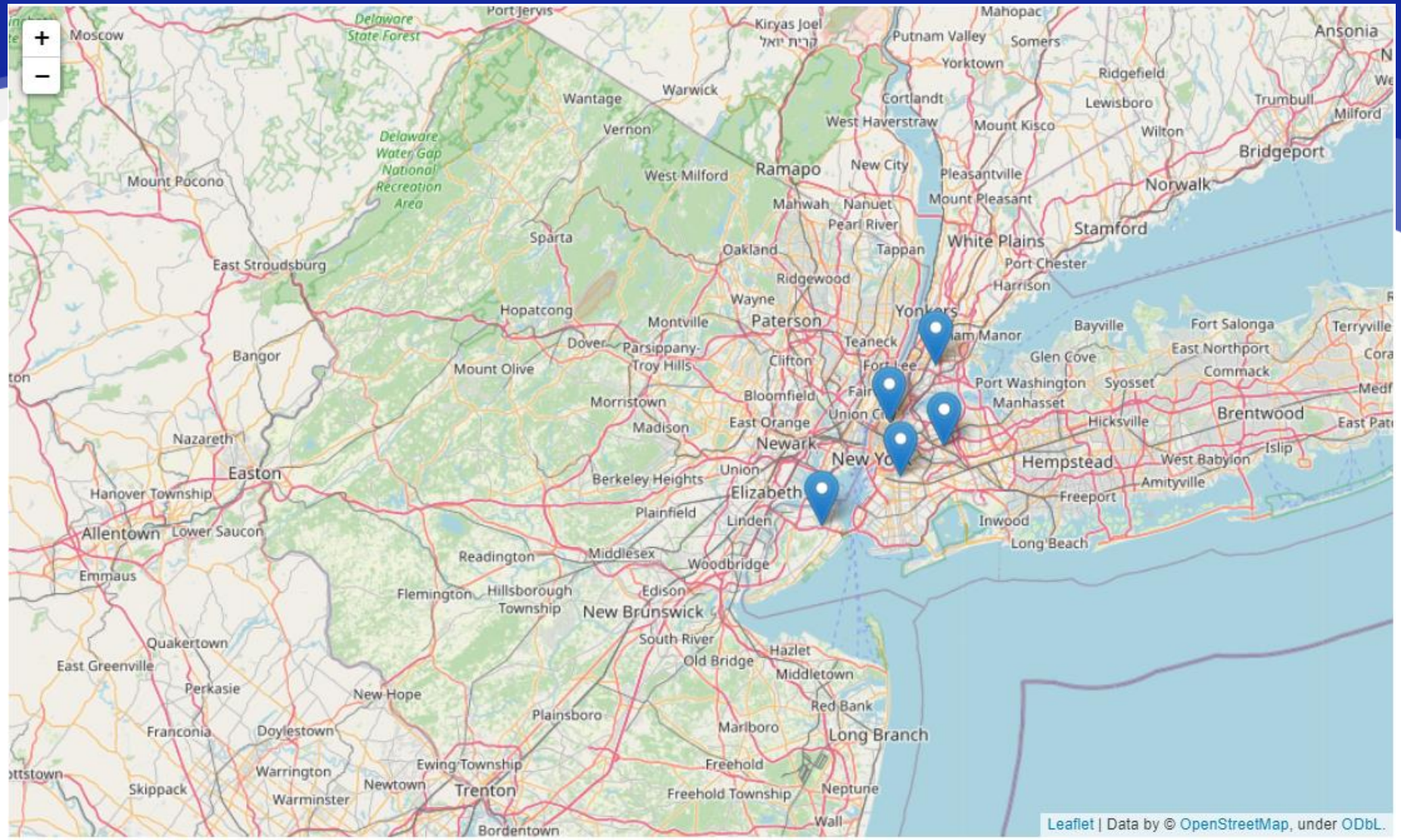




Conclusion :

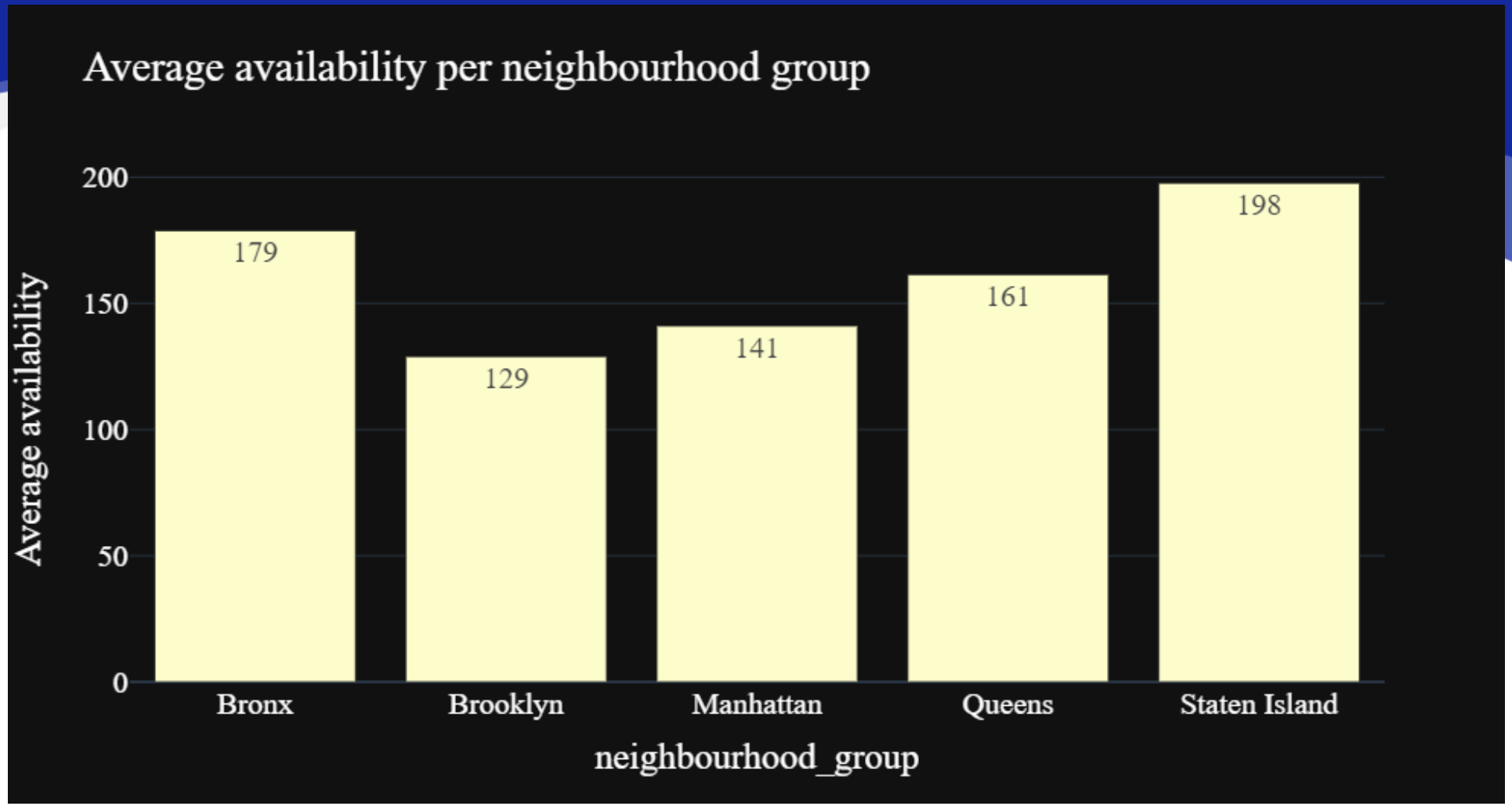
From the shown plot , the number of verification and non-verification is very close, which causes insecurity for the guest, and we can fix this and ensure a more level of safety for the guest by working to increase the number of verification

Geographical location on the world map



Conclusion : All in New York City

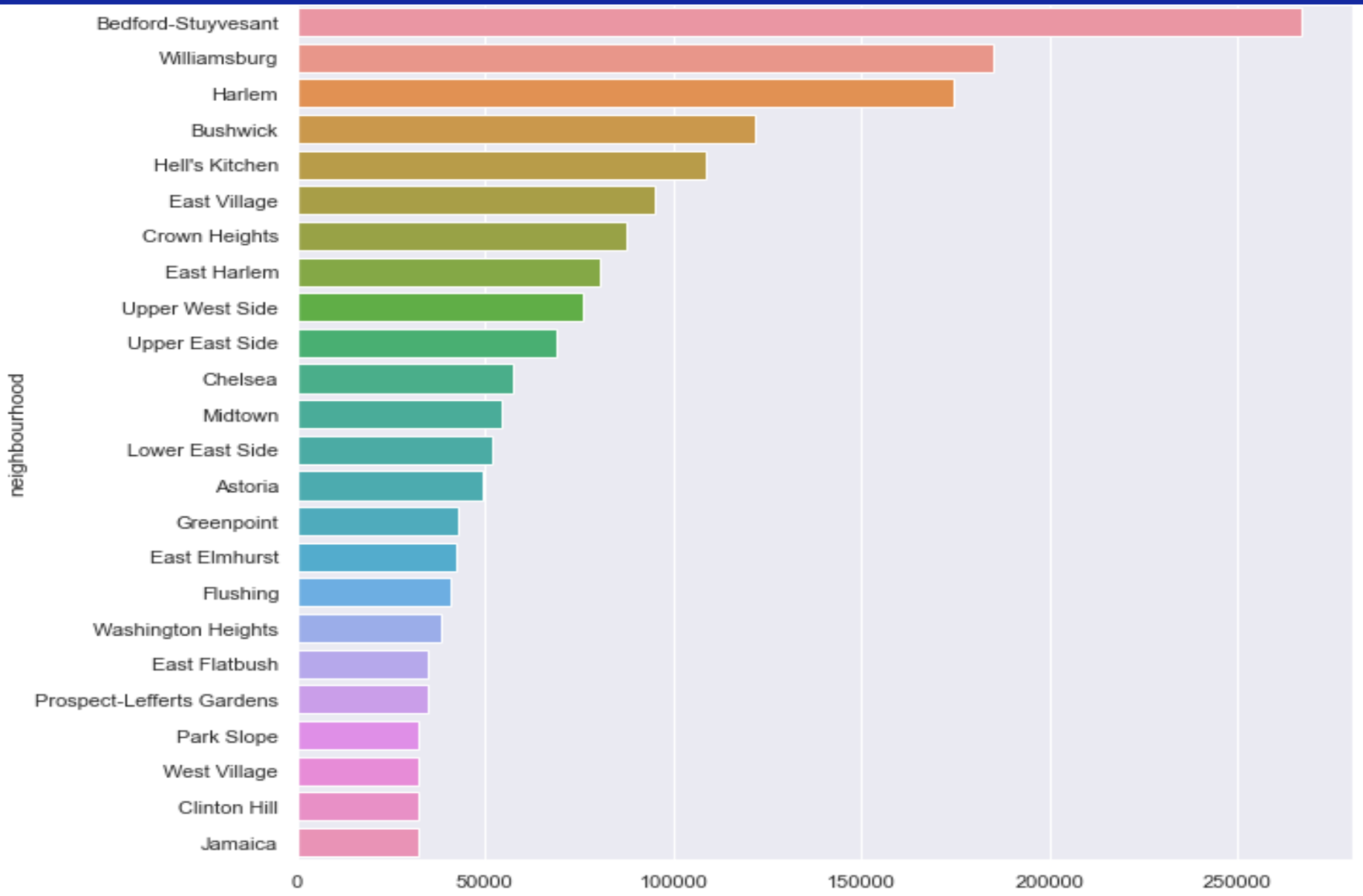
Average availability for each neighborhood group



Conclusion :

Brooklyn is the most popular listing, followed by Manhattan. It seems that Staten Island are the least popular listings. . So that we can make offers on it

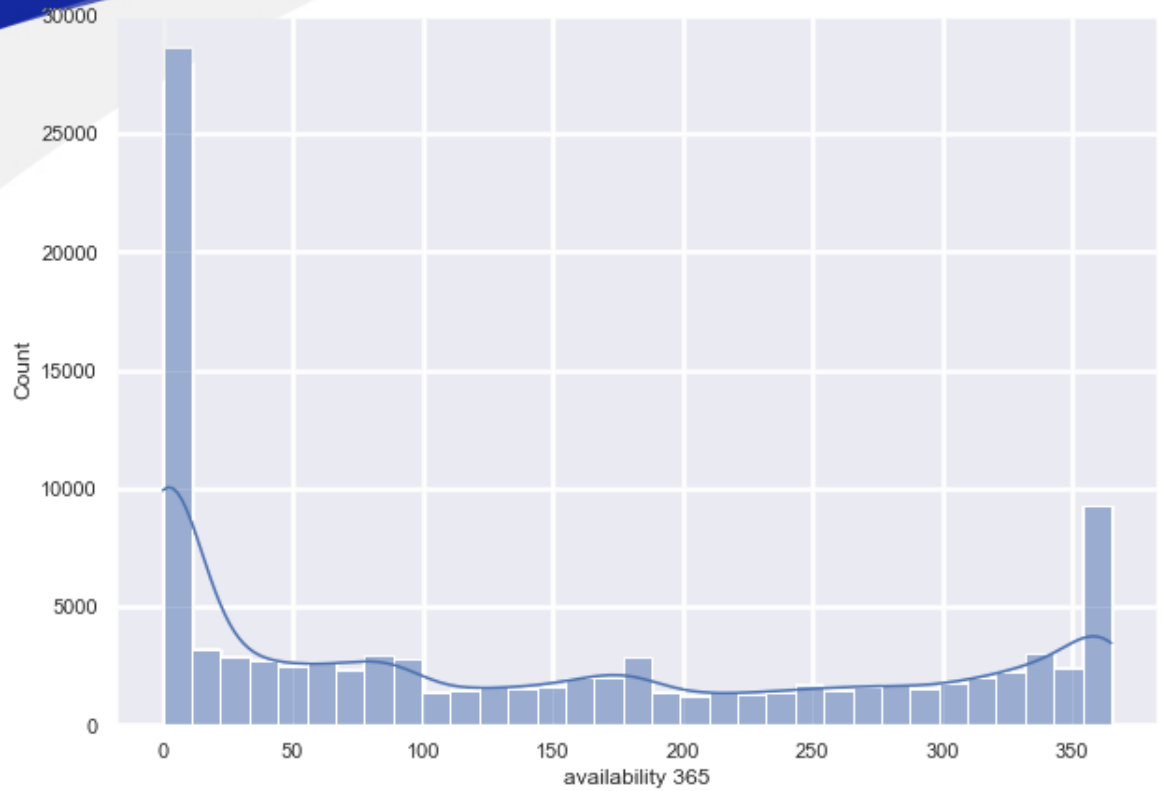
Top 25 most reviewed neighborhoods

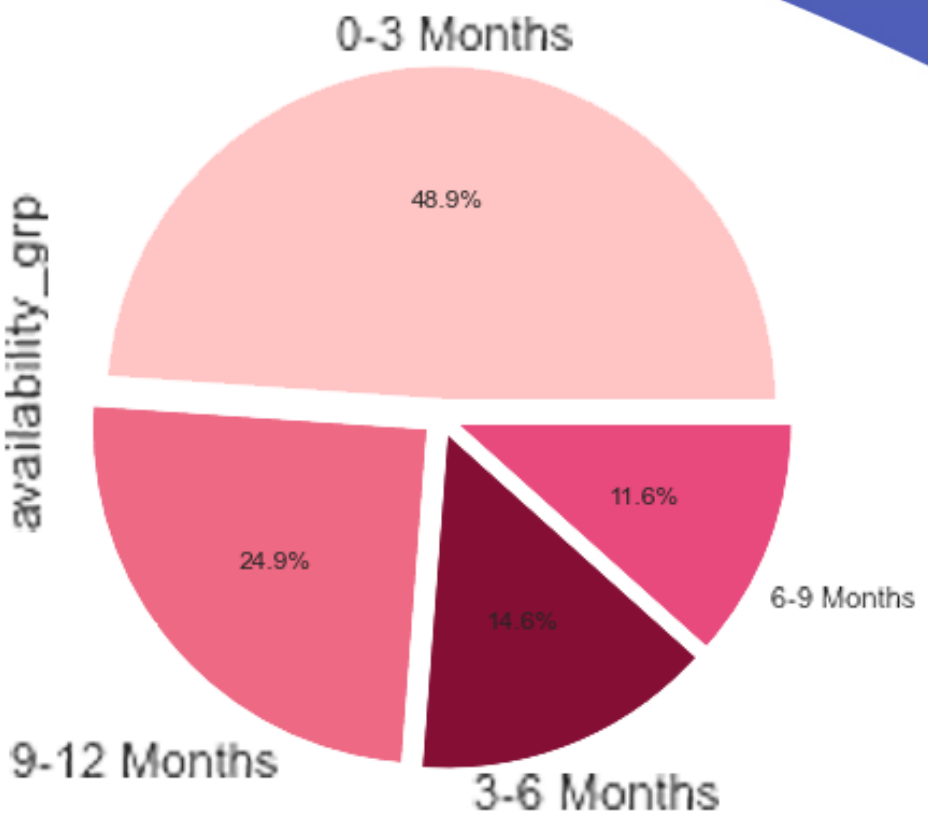


Conclusion :

Bedford – Stuyvesant is top 1

The most available set of dates the house receive the guest

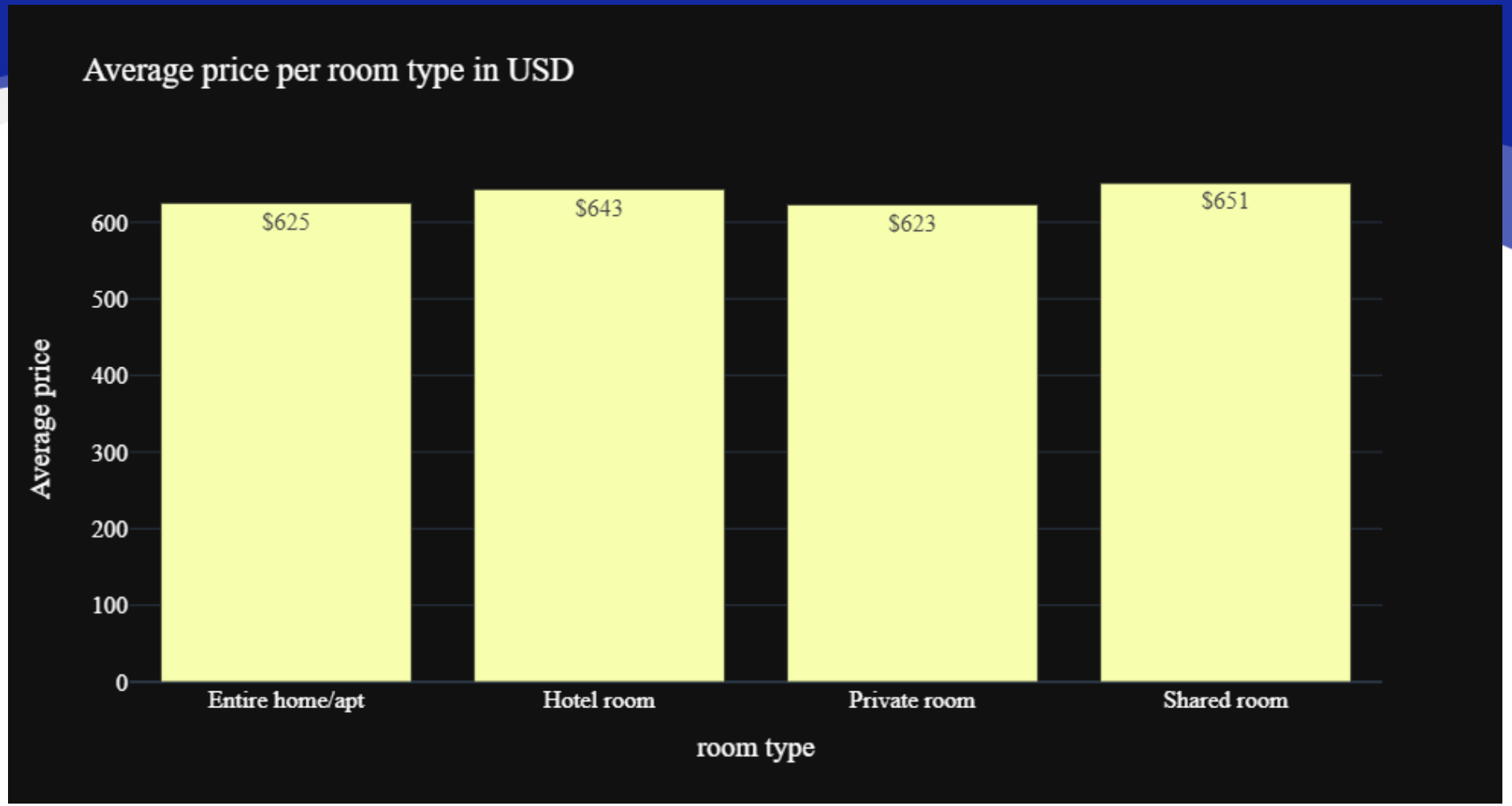




Conclusion :

- from 0-3 months is the most.
  - 0 and 365 day is the most

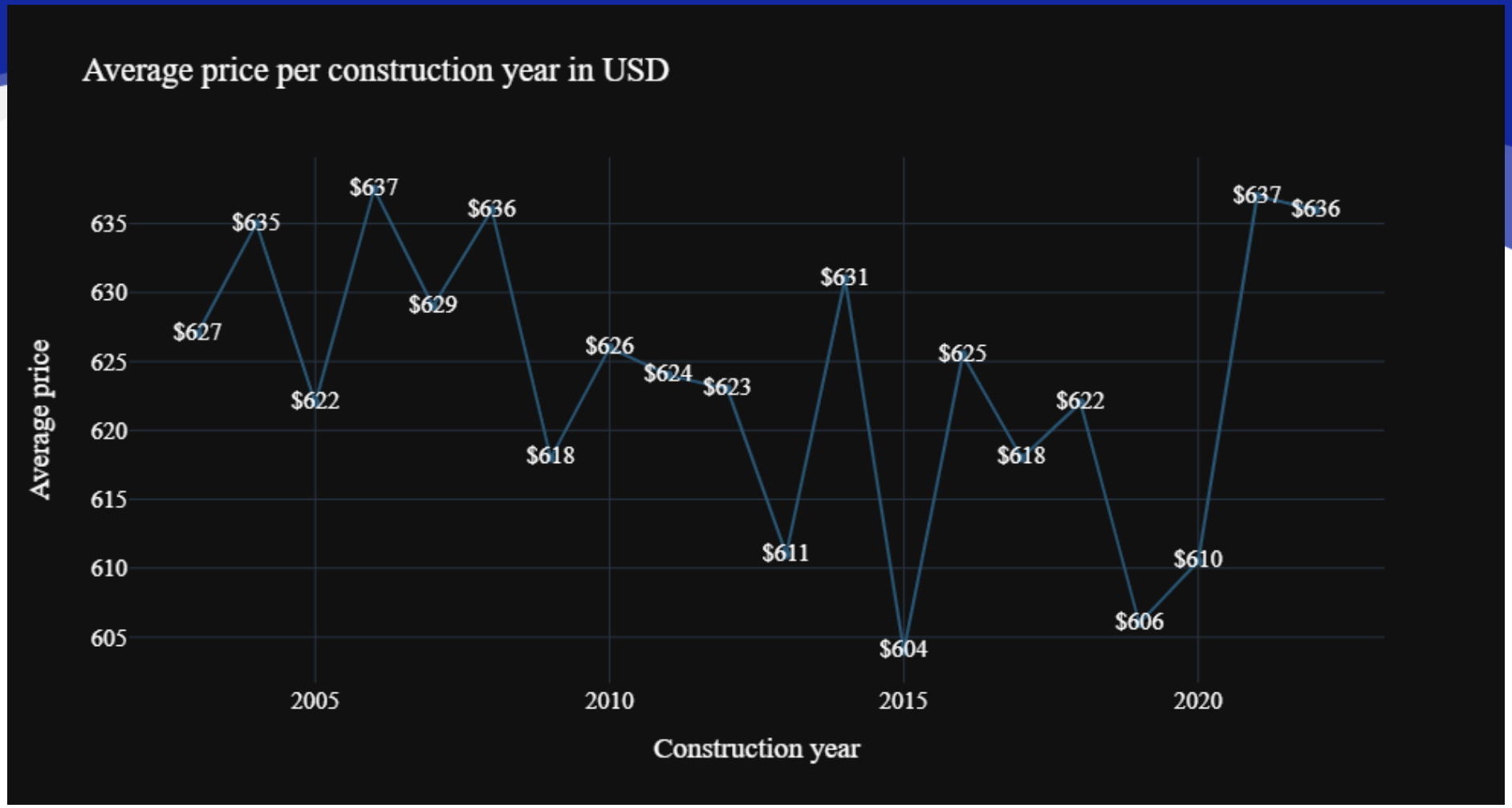
The average price for each room type



Conclusion :

The average price per night is $600 for all types

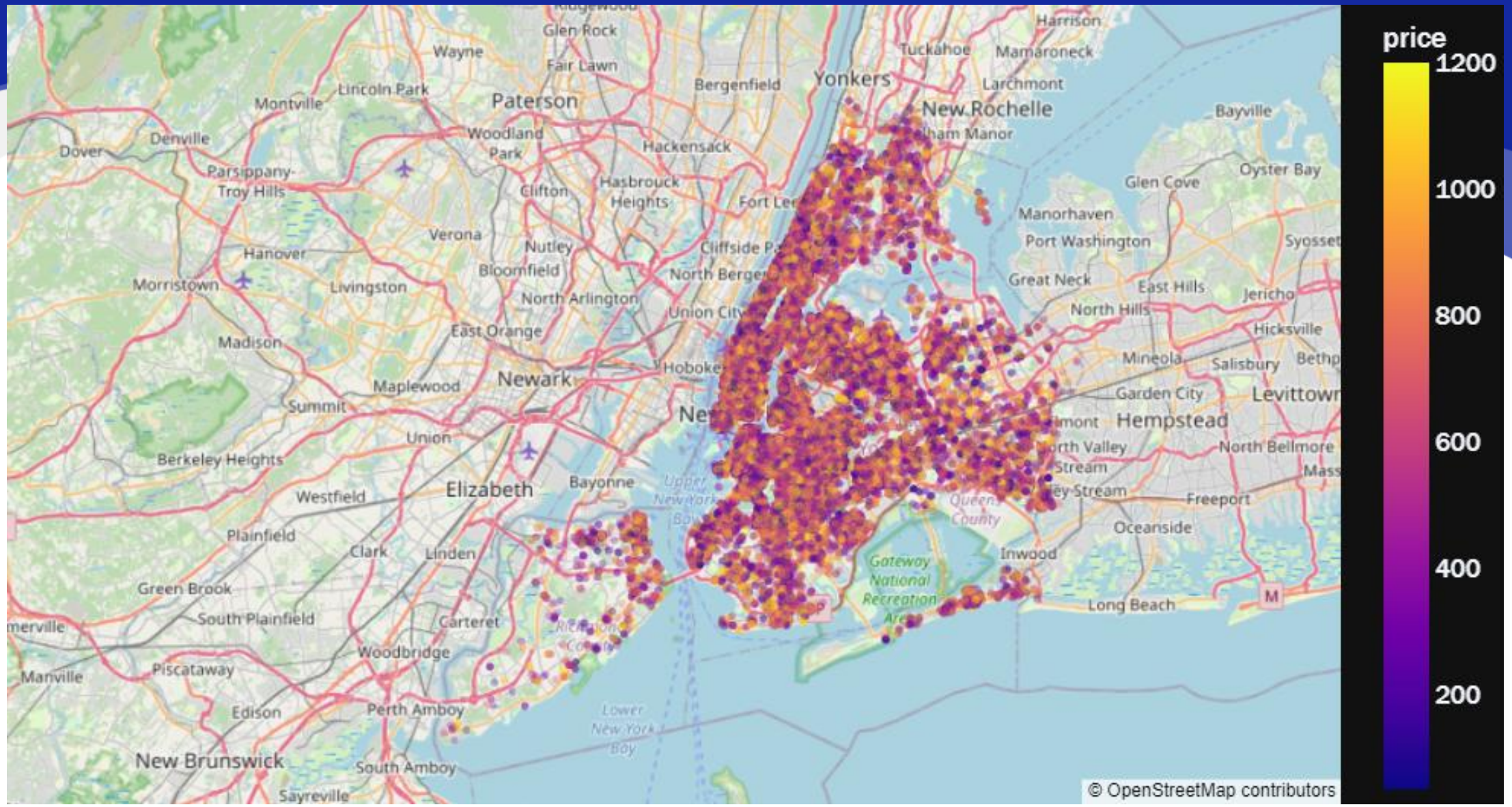
The later construction year correlated with the higher price



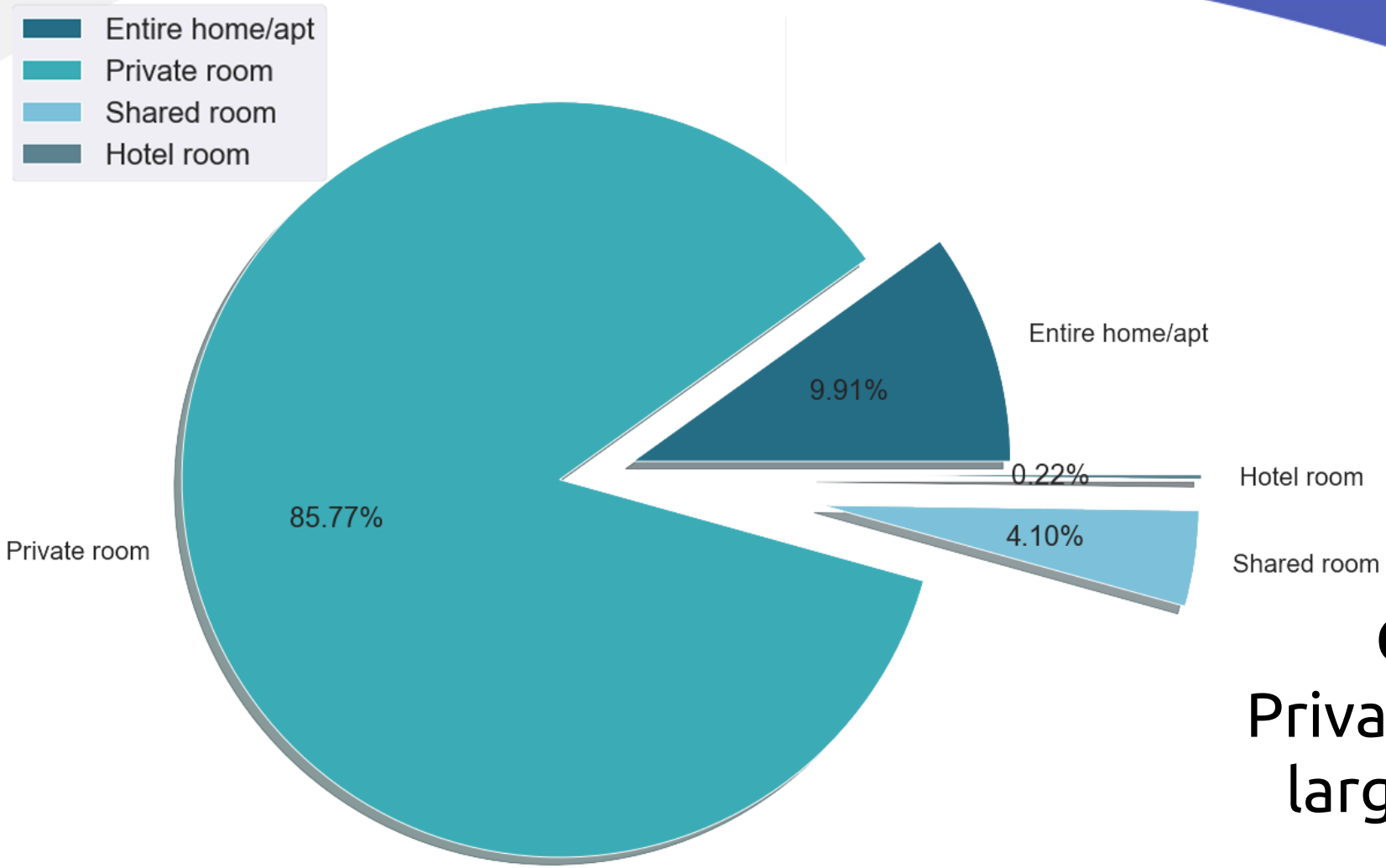
Conclusion :

The average price managed to stay between 600 and 620 per night, but there are some outliers that depends on more luxury

Prices depends on neighborhood group by map

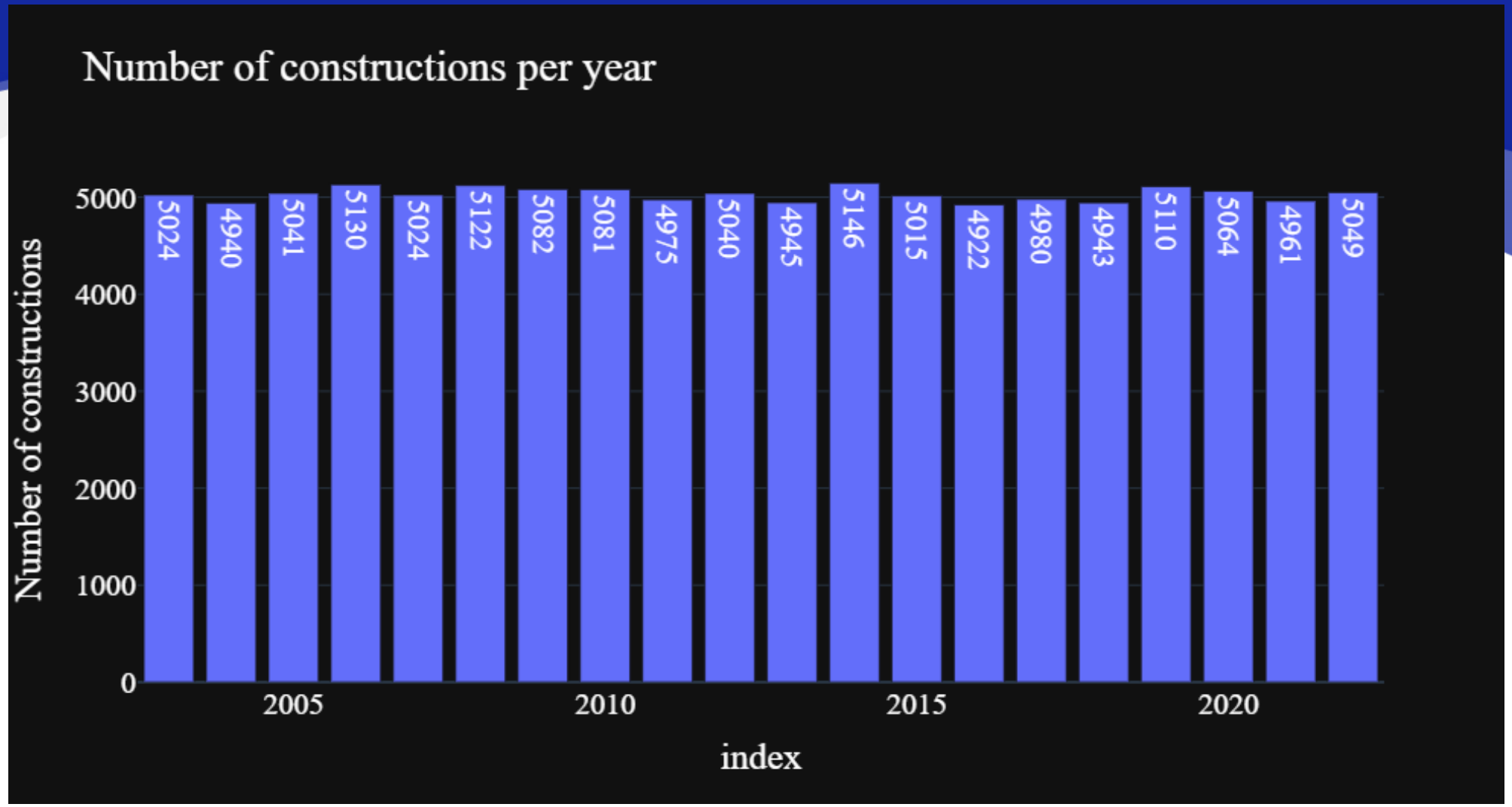


The proportions of room types?



Conclusion :
Private rooms get the largest percentage

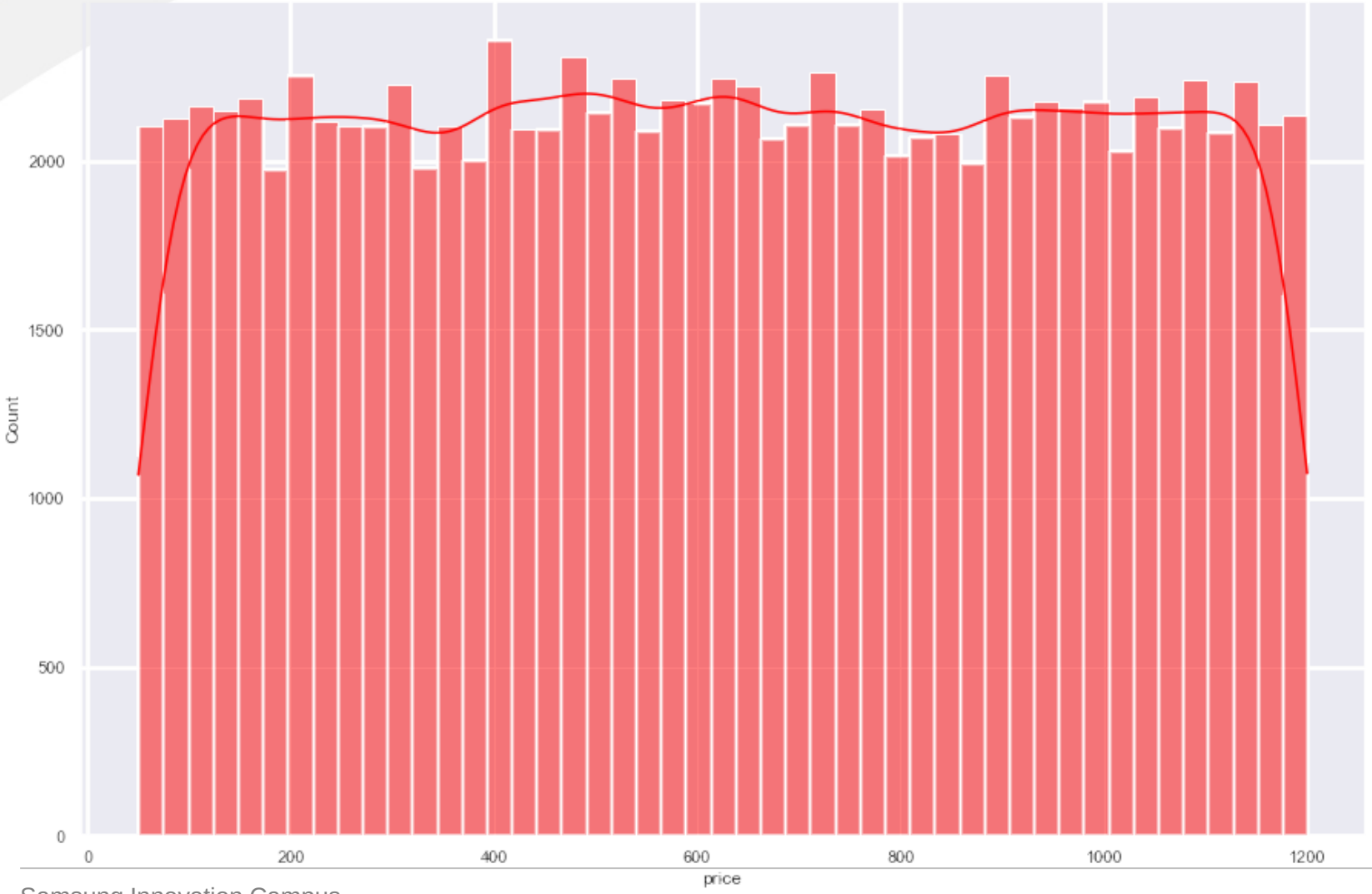
Number of constructions per year



Conclusion :

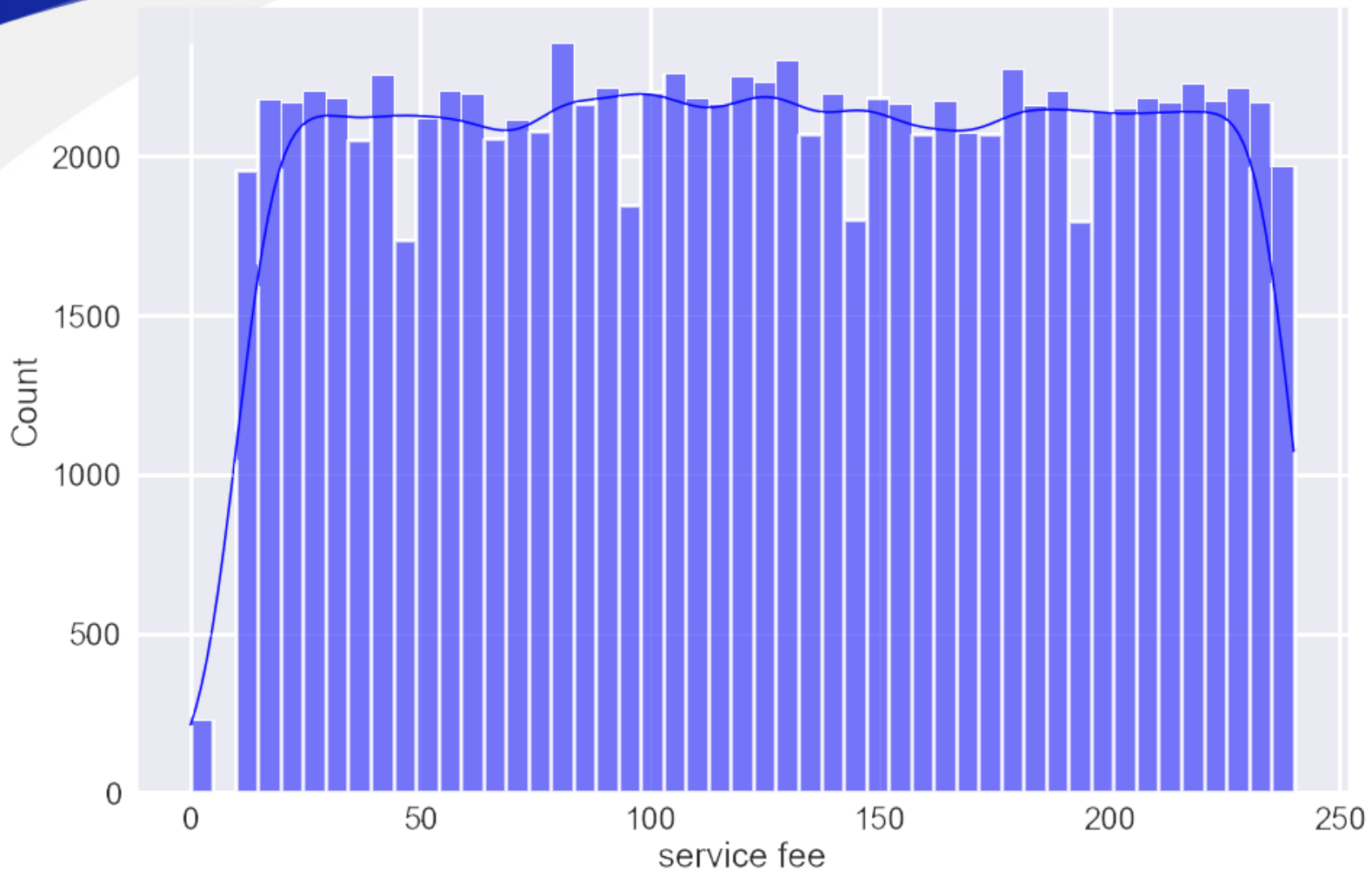
It looks like there are about ~4,000 constructions per year on the dataset

The range of prices



Conclusion :
Prices range from 50 to 1200 dollars per night

Service fee



Conclusion :
Service fee range from 0 to 240 dollars

Host name

Conclusion :

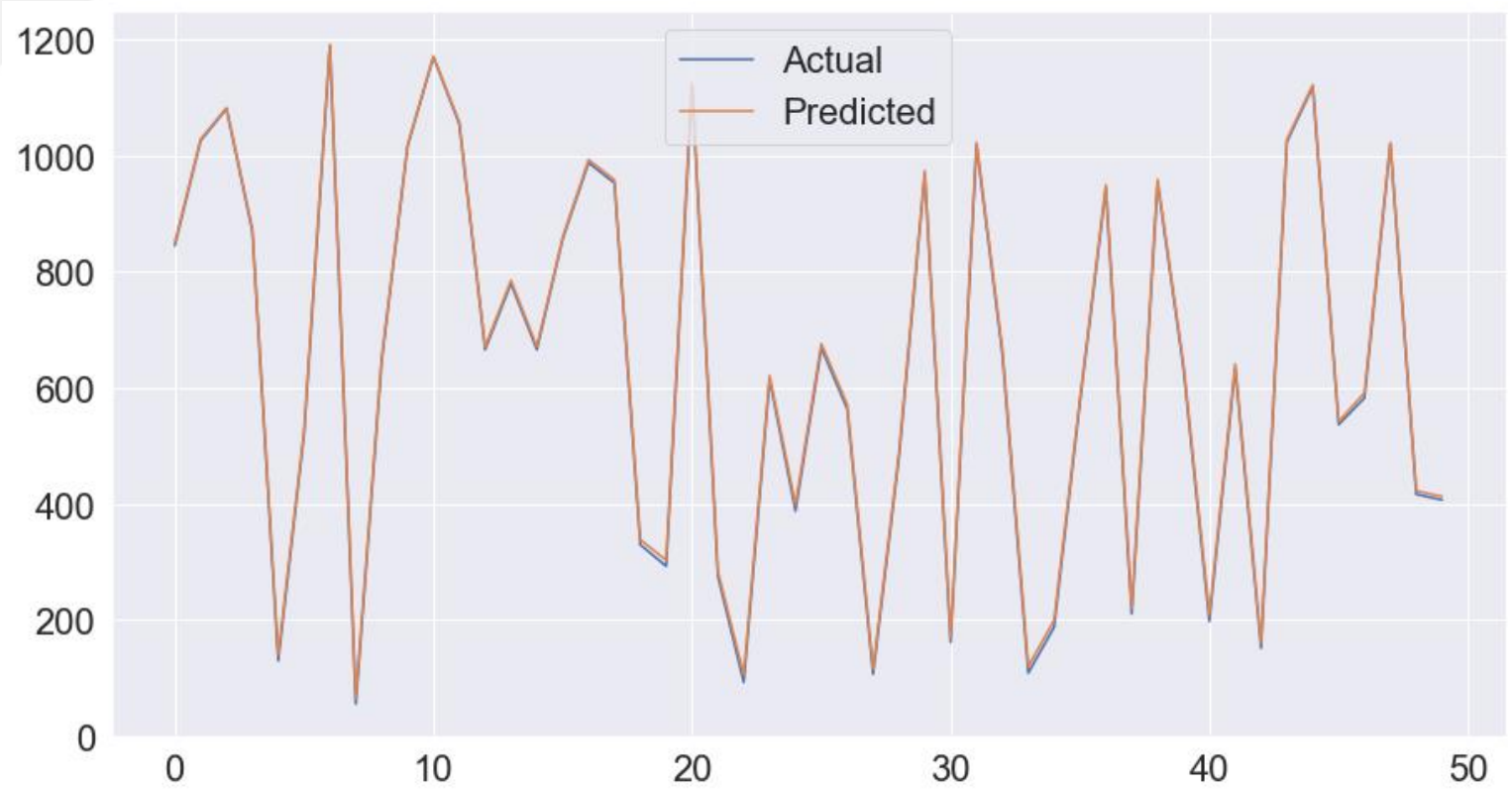
Michael, David, and John are the most popular hosts names

05

Modelling and Data fitting

Apply some important models in machine learning

Linear Regression



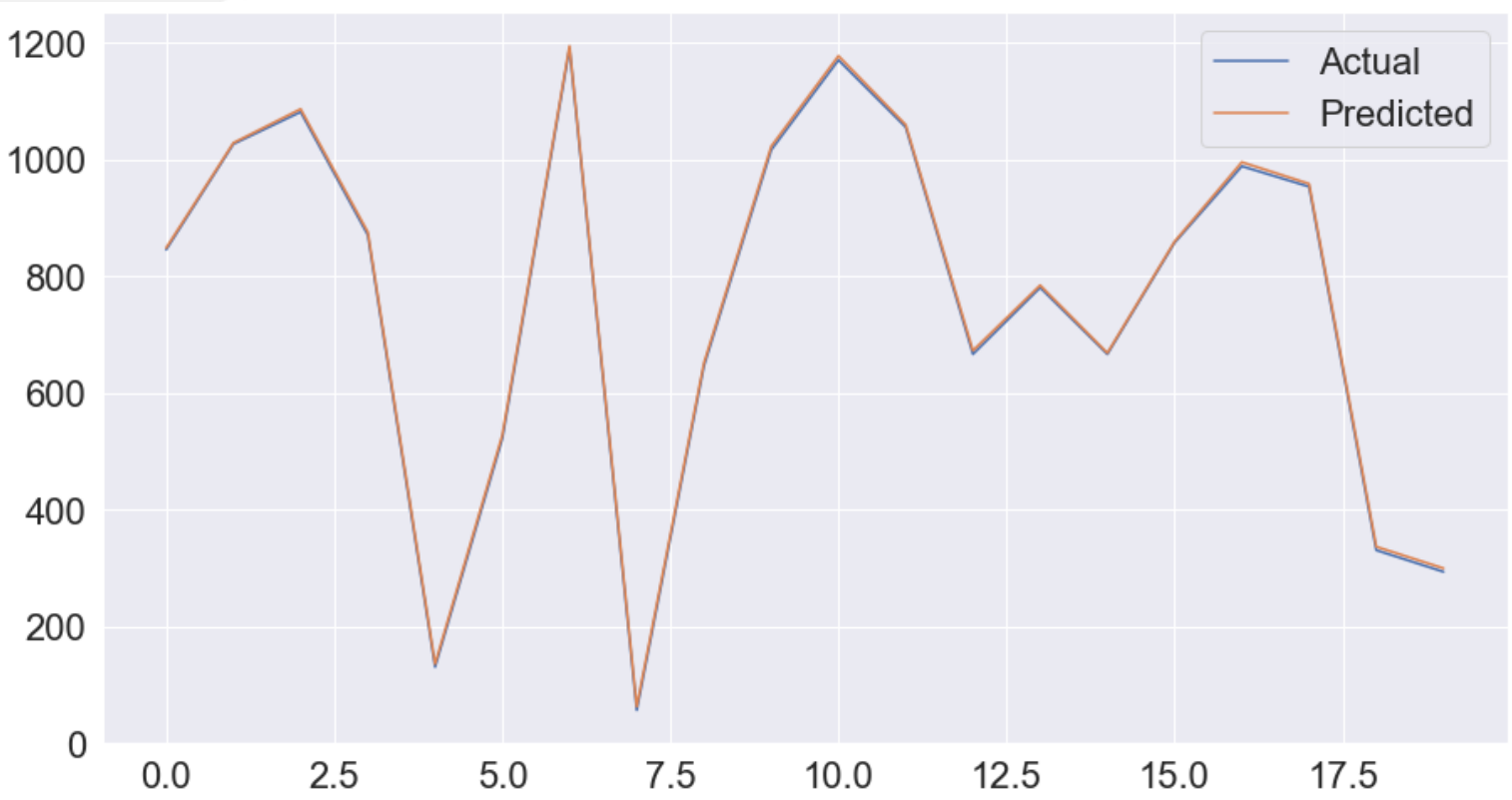
Evaluation of Linear Regression

```
In [115]: reg_score = r2_score(y_test , y_pred)
          reg_score
Out[115]: 0.9877797759896411

In [116]: p = len(x_train[0])
          n = len(y_train)
          adj_R2 = 1-(1-reg_score)*(n-1)/(n-p-1)
          adj_R2
Out[116]: 0.9877776699203574

In [117]: adj_R2 < reg_score
Out[117]: True
```

```
[102]: reg = linear_model.LinearRegression()
       reg.fit(x_train,y_train)
       regv =reg.score(x_train,y_train)
       regv
[102… 0.9886161767574628
```

```
[103]: reg.score(x_test,y_test)
[103… 0.9877797759896411
```

Decision Tree



Evaluation of Decision Tree

```
In [139]: tdt_score = r2_score(y_test , y_pred)
          tdt_score
Out[139]: 0.9949502872387704

In [140]: p = len(x_train[0])
          n = len(y_train)
          adj_R2 = 1-(1-tdt_score)*(n-1)/(n-p-1)
          adj_R2
Out[140]: 0.9949494169564476

In [141]: adj_R2 < tdt_score
Out[141]: True
```

iii. Decision Tree

```
129]: tdt = DecisionTreeRegressor().fit(x_train, y_train)
      tdtv = tdt.score(x_train,y_train)
      tdtv
129... 1.0
```

Comparison of all models in Accuracy

By R^2

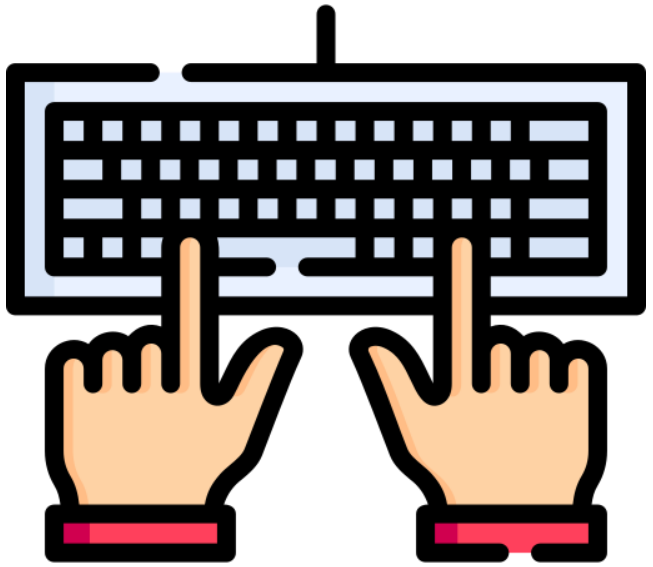
06

Business Solution

Solve the problems that were presented at the beginning through the data

Make it easier for host and guests to know the prices

The problem will be solved by creating a site that takes from the user, whether it is a host or a guest, some additional information that is not available to search for on the site, so that all parties benefit through that the host knows how much the product is offered at a reasonable price, and the guest knows in one way or another the average cost of the room with the specifications required

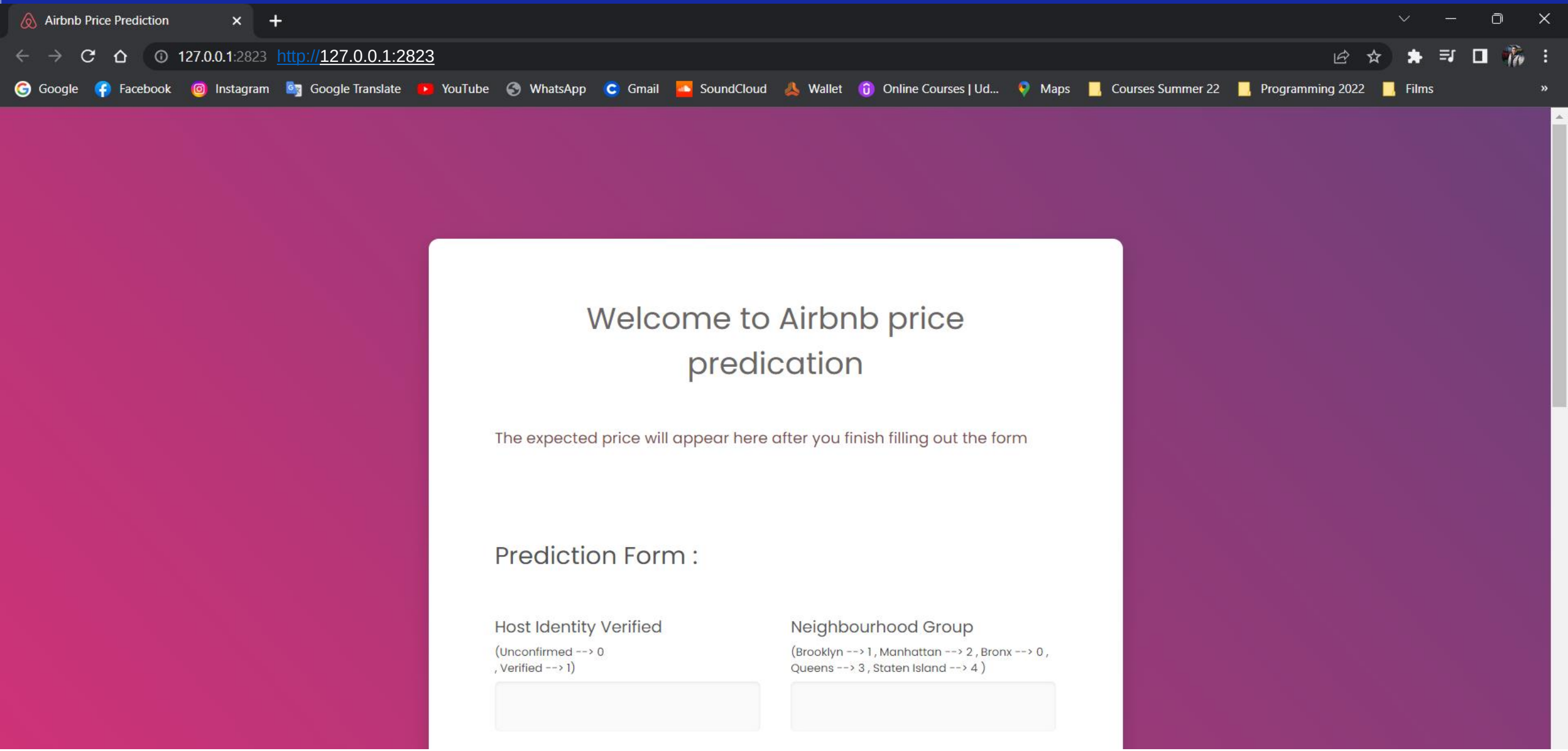


07

Deployment

Price Prediction Website

The site's interface



It is required to enter numbers to predict

Host Identity Verified
(Unconfirmed --> 0
, Verified --> 1)

Neighbourhood Group
(Brooklyn --> 1 , Manhattan --> 2 , Bronx --> 0 , Queens --> 3 , Staten Island --> 4)

Instant Bookable
(False --> 0 , True --> 1)

Cancellation Policy
(Moderate --> 1 , Strict --> 2 , Flexible --> 0)

Room Type
(Hotel Room --> 1 , Private Room --> 2 , Entire Home/Apt --> 0 , Shared Room --> 3)

Construction Year
(In Which Year It
Was Built)

Minimum Nights
(Minimum Number Of Night Stay For The Listing)

Number Of Reviews
(The Number Of Reviews
The Listing Has)

Reviews Per Month
(Average Number Of Reviews Per Month)

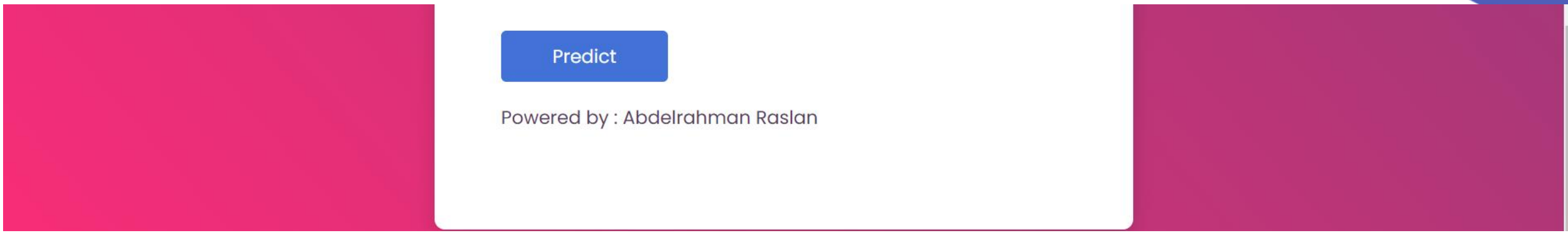
Review Rate Number
(Total Average Number Of Reviews)

Calculated Host Listings Count
(Amount Of
Guests)

Availability 365
(Number Of Days That Room Available In The Year)

Service Fee
(Airbnb Profit)

Button Action to predict



Example

Prediction Form :

Host Identity Verified
(Unconfirmed --> 0 , Verified --> 1)
1

Neighbourhood Group
(Brooklyn --> 1 , Manhattan --> 2 , Bronx --> 0 , Queens --> 3 , Staten Island --> 4)
2

Instant Bookable
(False --> 0 , True --> 1)
1

Cancellation Policy
(Moderate --> 1 , Strict --> 2 , Flexible --> 0)
2

Room Type
(Hotel Room --> 1 , Private Room --> 2 , Entire Home/Apt --> 0 , Shared Room --> 3)
3

Construction Year
(In Which Year It Was Built)
2013

Minimum Nights
(Minimum Number Of Night Stay For The Listing)
15

Number Of Reviews
(The Number Of Reviews The Listing Has)
222

Reviews Per Month
(Average Number Of Reviews Per Month)
5

Review Rate Number
(Total Average Number Of Reviews)
101

Calculated Host Listings Count
(Amount Of Guests)
28

Availability 365
(Number Of Days That Room Available In The Year)
37

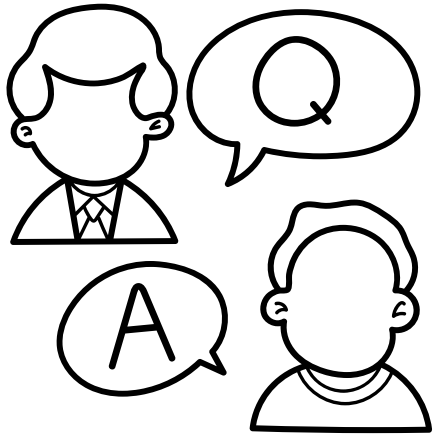
Service Fee
(Airbnb Profit)
0

Predict

Welcome to Airbnb price predication

The expected price will appear here after you finish filling out the form

The price is : [1077.] $ per night



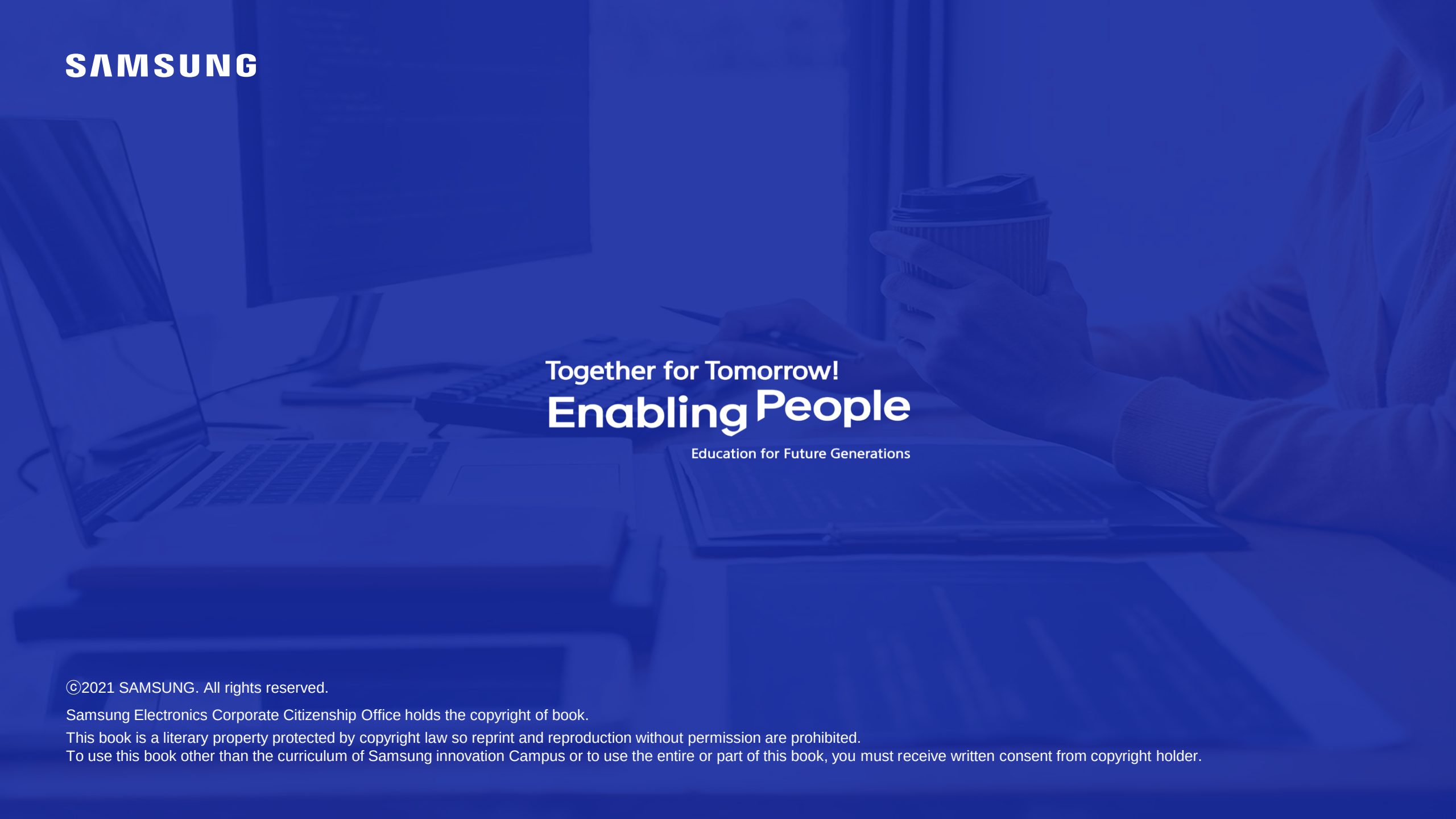
Any question

I hope everyone enjoys and takes advantage of the presentation


Morris Charts
Line Chart
Area Chart
Bar Chart
Donut Chart
25%
Donut Chart
Sparkline Charts
Pie Chart
Easy Pie Charts
25%
50%
75%

Thank You

SAMSUNG



ⓒ2021 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of book.

This book is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.
To use this book other than the curriculum of Samsung innovation Campus or to use the entire or part of this book, you must receive written consent from copyright holder.