

# WHERE TO STAY IN NYC?

## 1. Introduction:

### 1.1 Background

As the center of globe commerce and culture, New York has a huge population and attract a large amount of travellers every year. This project is to help travelers to find a best way to stay during their travelling based on their locations and interests.

Therefore, this analysis aims to gather basic understanding of the different boroughs in NYC, using NYC house market data available, in conjunction with supporting geographic information such as crime rate, and education data.

People who come to NYC to stay for a while, or who are keen to buy a property in NYC but with limited experience may find the analysis helpful.

## 2. Data

### 2.1 Data Source

A list of NYC boroughs has been sources from Wikipedia, to help narrow down the target areas for analysis.

### 2.2 Data Cleansing

Python is utilized for the data analysis; sources data are mainly stored in csv or xlsx format that easily loaded into Python for data cleansing.

For the timeliness of the information, I decided to only use last 24-month data for house price, crime rate, and education information.

It is noted that the spelling of certain Boroughs is slightly different which lead to null values been produced when different data sources are joined together using Borough as a join key. Further data cleansing was performed to align the spelling of Boroughs, such as: "Westminster" is now changed to "City of Westminster", etc.

## 3. Methodology

The analysis employed Python as programming platform, and used various analysis packages such as Pandas, NumPy, and Matplotlib etc, for different features.

To gain a basic understanding of the regions, firstly we run a few tests to describe the data, using Pandas and demonstrate the result using data virilization technics.

After that, we asked Python to performed unsupervised Machine Learning technicism, to cluster data in to different cluster.

### 3.1 Descriptive Statistics

Firstly, we chose to look at the prices of 3 house types (ONE, TWO, THREE BEDROOM) in NYC over last 12 months for each Borough.

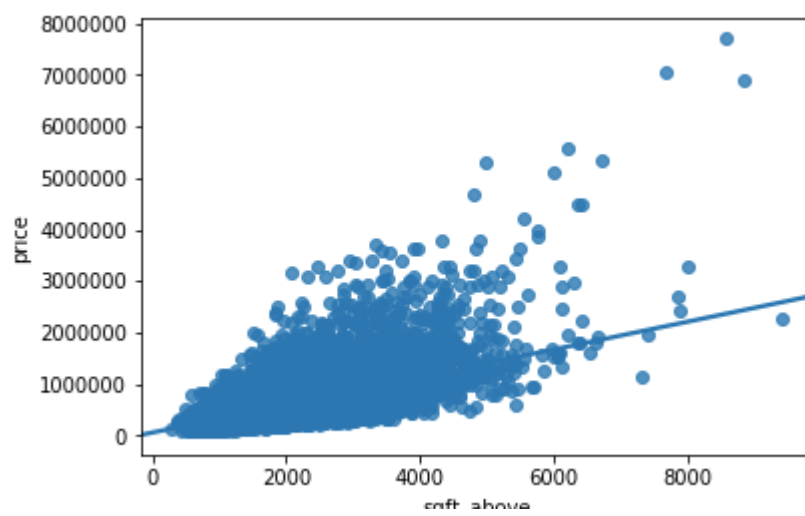
This chart provided readers with a basic understanding of house housing price to fit individual's budget consideration.

```
]:
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
count	2.161300e+04	21600.000000	21603.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000
mean	5.400881e+05	3.372870	2.115736	2079.899736	1.510697e+04	1.494309	0.007542	0.234303
std	3.671272e+05	0.926657	0.768996	918.440897	4.142051e+04	0.539989	0.086517	0.766318
min	7.500000e+04	1.000000	0.500000	290.000000	5.200000e+02	1.000000	0.000000	0.000000
25%	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000
75%	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000

We then look at the value of information BETWEEN PRICE AND ROOMTYPES.

It can be observed that a relatively clear positive correlation between average house price and number of rooms.



## 4. Results

Post K-means clustering analysis, it is noticed that the outsider Brough's of NYC are largely grouped together (blue dots). As K-means is a way of unsupervised machine learning technique, we are not expecting the result to provide any opinion on where to buy or to live. It is a way to differentiate each region based on the features we feed into model. For example, a red dot from west NYC (Hounslow), share similar characteristics from a red dot from east NYC (Romford).

## 5. Discussion

Although data clustering provided an un-biased view of the grouping of regions, there are a few limitations of the analysis that worth to call out:

There are still room to enhance the quality of the data feed. If we compare an area with high crime rate but mainly robbery, to another area with lower water rate but mainly murder, relying on pure volume will not provide sufficient insight.

## 6. Conclusion

To conclude, the analysis is helpful for any reader who have no or limited knowledge of living in NYC. People who might be interested in this projects are travelers who plan to choose airbnb in NYC.