Exploratory Data Analysis OF Lagos House Prices

- by Inyama Chiedoziem

Objective

To determine the best place to live in lagos based on the following parameters

- Security
- Budget
- Other Utilities
- House Density
- How Landlords can Increase their Prices

Table of Content

- Data Wrangling
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Introduction

Lagos is state in the south western part of Nigeria, that boasts of the highest population density of any city in Africa. Here the the cost of renting houses are quite high, so I decided analyze the cost of owning a house in the city. We are hoping to look into the pricing patterns of lagos houses draw observation and render conclusion. This dataset was provided by quantum analytics

```python
In [2]:    # Importing Libraries
           import numpy as np
           import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sns
```

Data Wrangling

In [4]:
```
# read file into pandas dataframe
lagos_houses = pd.read_csv(r'C:\Users\Mhizfair\Desktop\Quantum analytics\Python\lagos_house_prices_raw.csv')
```

In [5]:
```
lagos_houses.head()
```

Out[5]:

	location	bed	bath	toilet	price	Property_Type	Parking_Space	Security	Electricity	Furnished	Security_Doors	CCTV	Pool	Gym	BQ
0	yaba	1	1	2	700000.0	Mini flat	0	0	0	0	0	0	0	0	0
1	yaba	1	1	2	700000.0	Mini flat	0	0	0	0	0	0	0	0	0
2	yaba	1	1	2	650000.0	Mini flat	0	0	0	0	0	0	0	0	0
3	yaba	1	1	1	450000.0	Mini flat	0	0	0	0	0	0	0	0	0
4	yaba	3	3	4	800000.0	Detached duplex	0	1	0	0	0	0	0	0	0

In [6]:
```
lagos_houses.shape
```

Out[6]:
```
(5336, 15)
```

In [7]:
```
lagos_houses.dtypes
```

Out[7]:
```
location            object
bed                  int64
bath                 int64
toilet               int64
price              float64
Property_Type       object
Parking_Space        int64
Security             int64
Electricity          int64
Furnished            int64
Security_Doors       int64
CCTV                 int64
Pool                 int64
Gym                  int64
BQ                   int64
dtype: object
```

In [8]:
```
lagos_houses.columns
```

Out[8]:
```
Index(['location', 'bed', 'bath', 'toilet', 'price', 'Property_Type',
       'Parking_Space', 'Security', 'Electricity', 'Furnished',
       'Security_Doors', 'CCTV', 'Pool', 'Gym', 'BQ'],
      dtype='object')
```

In [9]:
```
# Checking if data is duplicated
sum(lagos_houses.duplicated())
```

Out[9]:
```
3434
```

In [21]:
```
# Viewing duplicate File
lagos_houses[lagos_houses.duplicated()]
```

Out[21]:

	location	bed	bath	toilet	price	Property_Type	Parking_Space	Security	Electricity	Furnished	Security_Doors	CCTV	Pool	Gym	BQ
1	yaba	1	1	2	700000.0	Mini flat	0	0	0	0	0	0	0	0	0
8	yaba	1	1	2	700000.0	Mini flat	0	0	0	0	0	0	0	0	0
9	yaba	1	1	2	700000.0	Mini flat	0	0	0	0	0	0	0	0	0
22	yaba	1	1	1	350000.0	Self contain	0	0	0	0	0	0	0	0	0
23	yaba	1	1	2	700000.0	Mini flat	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5327	ajah	1	1	1	600000.0	Mini flat	0	0	0	0	0	0	0	0	0
5329	ajah	1	1	2	500000.0	Mini flat	1	0	0	0	0	0	0	0	0
5330	ajah	1	1	2	550000.0	Mini flat	0	0	0	0	0	0	0	0	0
5333	ajah	4	4	5	1700000.0	Semi detached duplex	1	0	0	0	0	0	0	0	0
5334	ajah	1	1	2	500000.0	Mini flat	0	0	0	0	0	0	0	0	0

3434 rows × 15 columns

In [13]:
```python
# View Locations In Lagos
lagos_houses.location.unique()
```

Out[13]:
```
array(['yaba', 'surulere', 'gbagada', 'lekki phase 1', 'ikorodu', 'ikeja',
       'ajah'], dtype=object)
```

In [18]:
```python
# Price Of Most Expensive House
lagos_houses.price.max()
```

Out[18]:
```
2450000.0
```

In [22]:
```python
lagos_houses.price.min()
```

Out[22]:
```
150.0
```

In [25]:
```python
columns = ['Parking_Space', 'Security', 'Furnished', 'Electricity', 'Security_Doors', 'CCTV', 'Pool', 'Gym', 'BQ' ]
for col in columns:
```

```
        lagos_houses[col].replace({0: 'no', 1: 'yes'}, inplace = True)
```

In [27]: `lagos_houses.head()`

Out[27]:

	location	bed	bath	toilet	price	Property_Type	Parking_Space	Security	Electricity	Furnished	Security_Doors	CCTV	Pool	Gym	BQ
0	yaba	1	1	2	700000.0	Mini flat	no	no	no	no	no	no	no	no	no
1	yaba	1	1	2	700000.0	Mini flat	no	no	no	no	no	no	no	no	no
2	yaba	1	1	2	650000.0	Mini flat	no	no	no	no	no	no	no	no	no
3	yaba	1	1	1	450000.0	Mini flat	no	no	no	no	no	no	no	no	no
4	yaba	3	3	4	800000.0	Detached duplex	no	yes	no	no	no	no	no	no	no

In [26]: `lagos_houses.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5336 entries, 0 to 5335
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   location        5336 non-null   object
 1   bed             5336 non-null   int64
 2   bath            5336 non-null   int64
 3   toilet          5336 non-null   int64
 4   price           5336 non-null   float64
 5   Property_Type   5336 non-null   object
 6   Parking_Space   5336 non-null   object
 7   Security        5336 non-null   object
 8   Electricity     5336 non-null   object
 9   Furnished       5336 non-null   object
 10  Security_Doors  5336 non-null   object
 11  CCTV            5336 non-null   object
 12  Pool            5336 non-null   object
 13  Gym             5336 non-null   object
 14  BQ              5336 non-null   object
dtypes: float64(1), int64(3), object(11)
memory usage: 625.4+ KB
```

What is the structure of your dataset? There were 5336 records of data and 15 features (location,bed,bath,toilet,price, Property_Type,Parking_Space,Security,Electricity,Furnished,Security_Doors,CCTV,Pool,Gym,BQ). The variables can be grouped as numeric and

categorical.

(1=Yes) ——> (0=No) What is/are the main feature(s) of interest in your dataset? I'm interested in figuring out what features are best for predicting the prices of rent for Houses in lagos.

What features in the dataset do you think will help support your investigation into your feature(s) of interest? I expect that blood pressure will have the strongest effect on cvd: the higher the blood pressure, the greater risk of cvd. I also think that age, weight, height, cholesterol levels would have contributing effects.

Univariate Analysis:

This helps to get an overview of each variable in the data.

In [33]:
```python
# Lets see how each location listed their houses
count_listing = lagos_houses['location'].value_counts()
count_listing
```

Out[33]:
```
yaba               1460
ajah                900
surulere            728
gbagada             692
ikorodu             578
lekki phase 1       521
ikeja               457
Name: location, dtype: int64
```

In [38]:
```python
count_listing.plot.bar(title= 'Distribution Of Listing By Location', ylabel= 'Frequency', xlabel= 'Location')

plt.show()
```

## Distribution Of Listing By Location
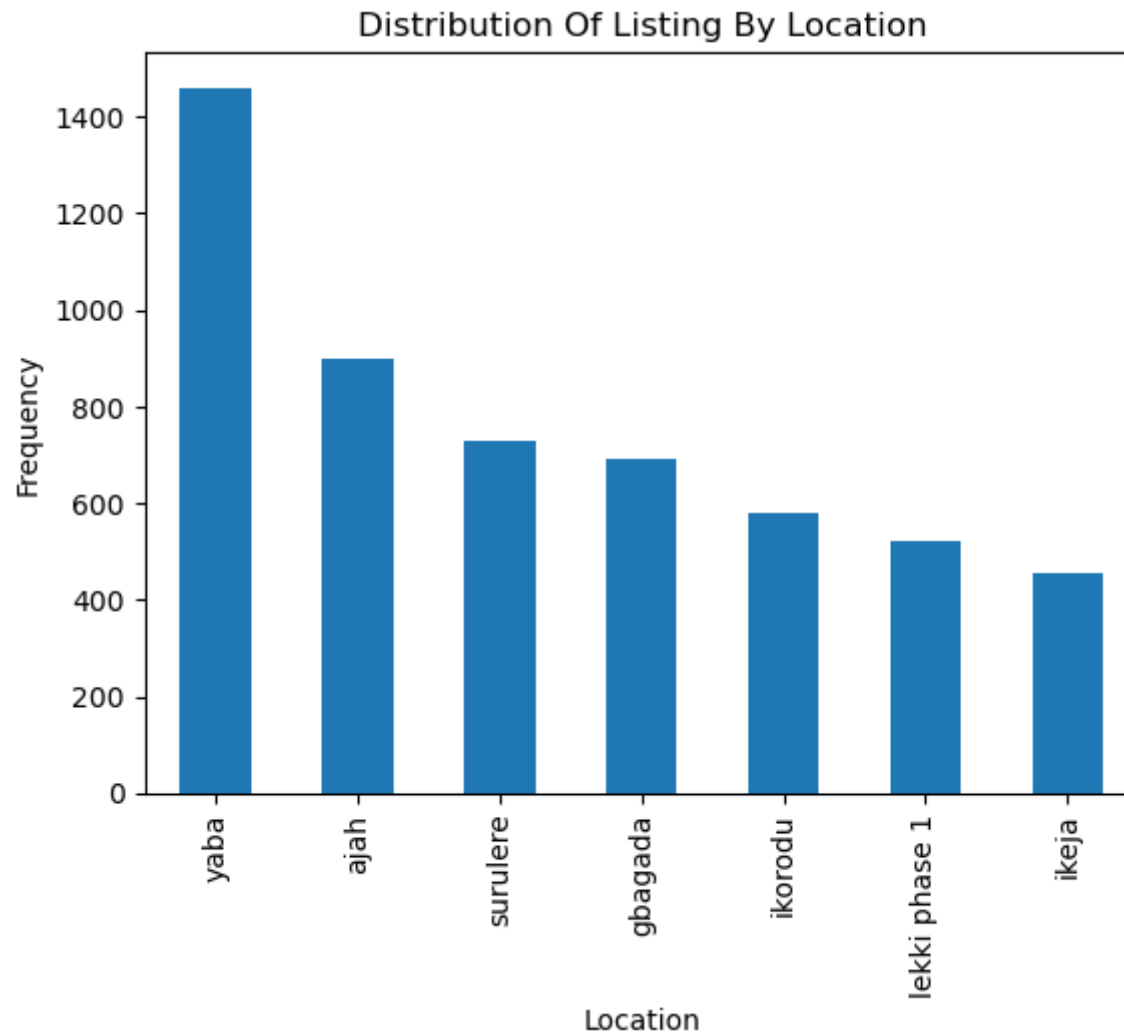
Observation

From the above barchart we can see that yaba had the highest amount of listing. Later we would investigate if the prices of the houses in yaba is responsible for the high amount of listing there also to note is that Ikeja had the least amount of listing despite being the state capital

```
In [60]:   # Can we deduce the Location that has CCTV?
           ctv= lagos_houses.groupby('location')['CCTV'].apply(lambda x: (x== 'yes').sum()).sort_values()
```

```
cctv
```

Out[60]:
```
location
lekki phase 1      0
gbagada            1
ajah               3
ikeja              5
yaba               5
ikorodu            8
surulere          12
Name: CCTV, dtype: int64
```

In [61]:
```python
cctv.plot.barh(title= 'No of CCTV By Location', ylabel= 'Frequency', xlabel= 'Location')

plt.show()
```
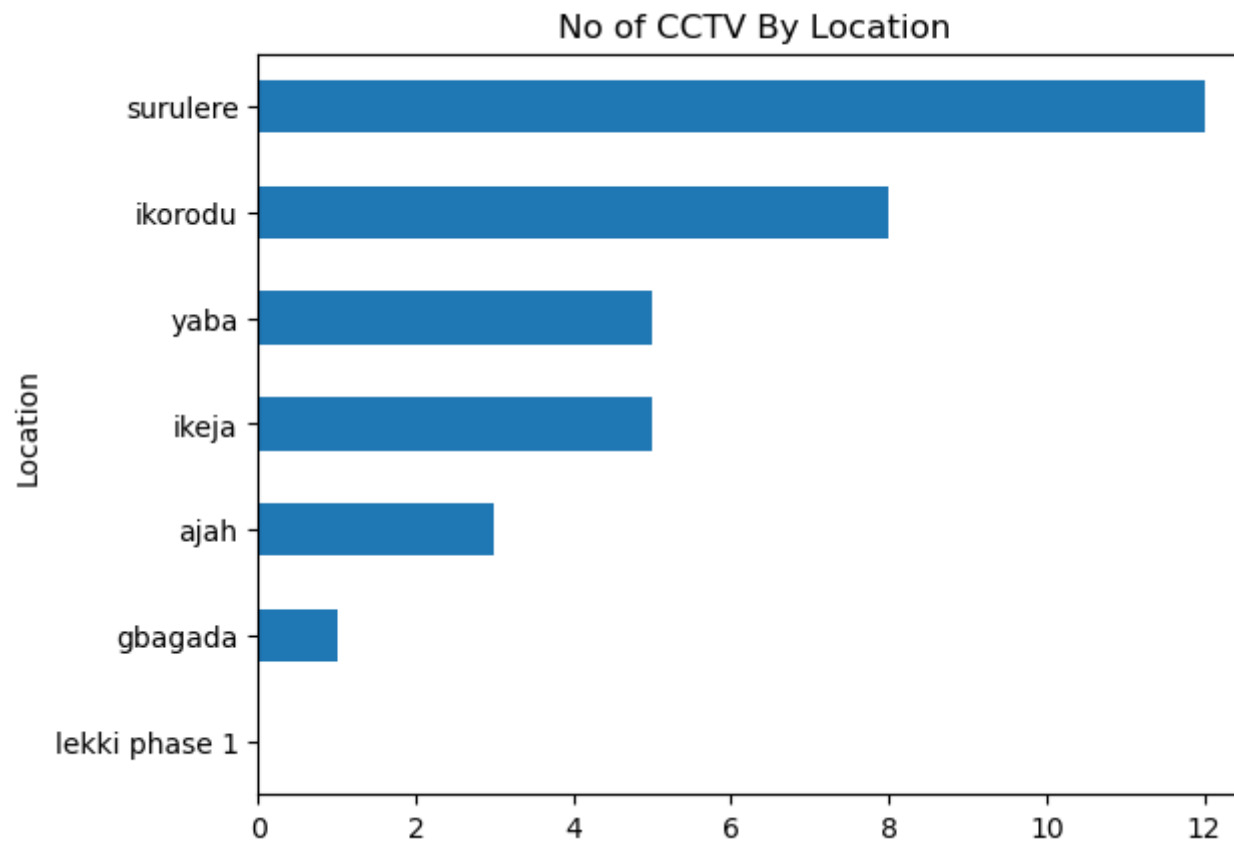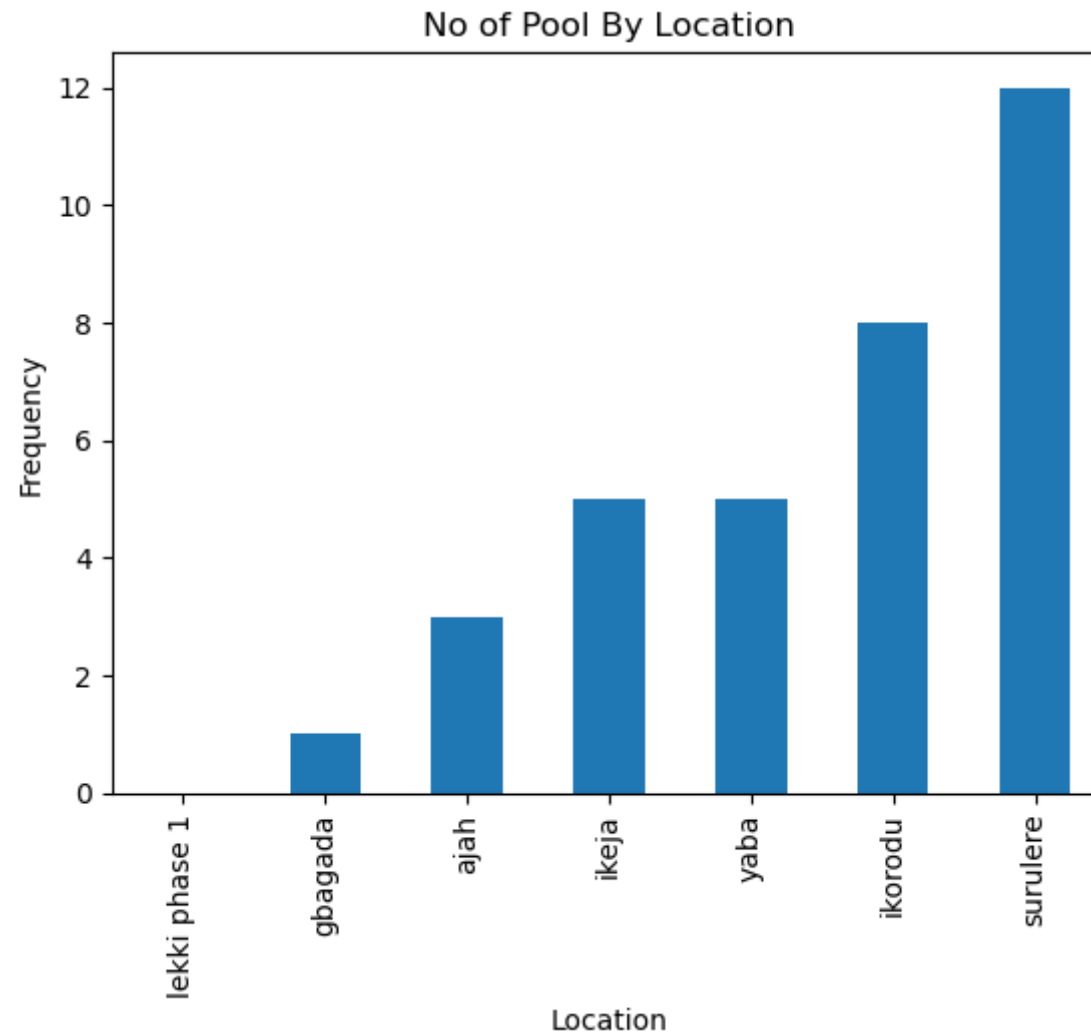
In [59]:
```python
# No. Of locations with Pools
pool= lagos_houses.groupby('location')['Pool'].apply(lambda x: (x== 'yes').sum())
pool
```

Out[59]:
```
location
ajah               1
gbagada            7
ikeja              5
ikorodu            0
lekki phase 1     15
surulere           3
yaba               1
Name: Pool, dtype: int64
```

In [63]:
```python
cctv.plot.bar(title= 'No of Pool By Location', ylabel= 'Frequency', xlabel= 'Location')

plt.show()
```

No of Pool By Location


```
In [68]:  # Average Price of Houses by Location
          Avg_price= lagos_houses['price'].groupby(lagos_houses.location).mean().astype('int').sort_values()
          Avg_price
```

Out[68]:
```
location
ikorodu            155095
yaba               550986
surulere           589189
ajah               693335
gbagada            742290
ikeja              772700
lekki phase 1     1211013
Name: price, dtype: int32
```

In [82]:
```python
Avg_price.plot(kind='line', x='location', y='price', title='Average price By Location')
plt.show()
```
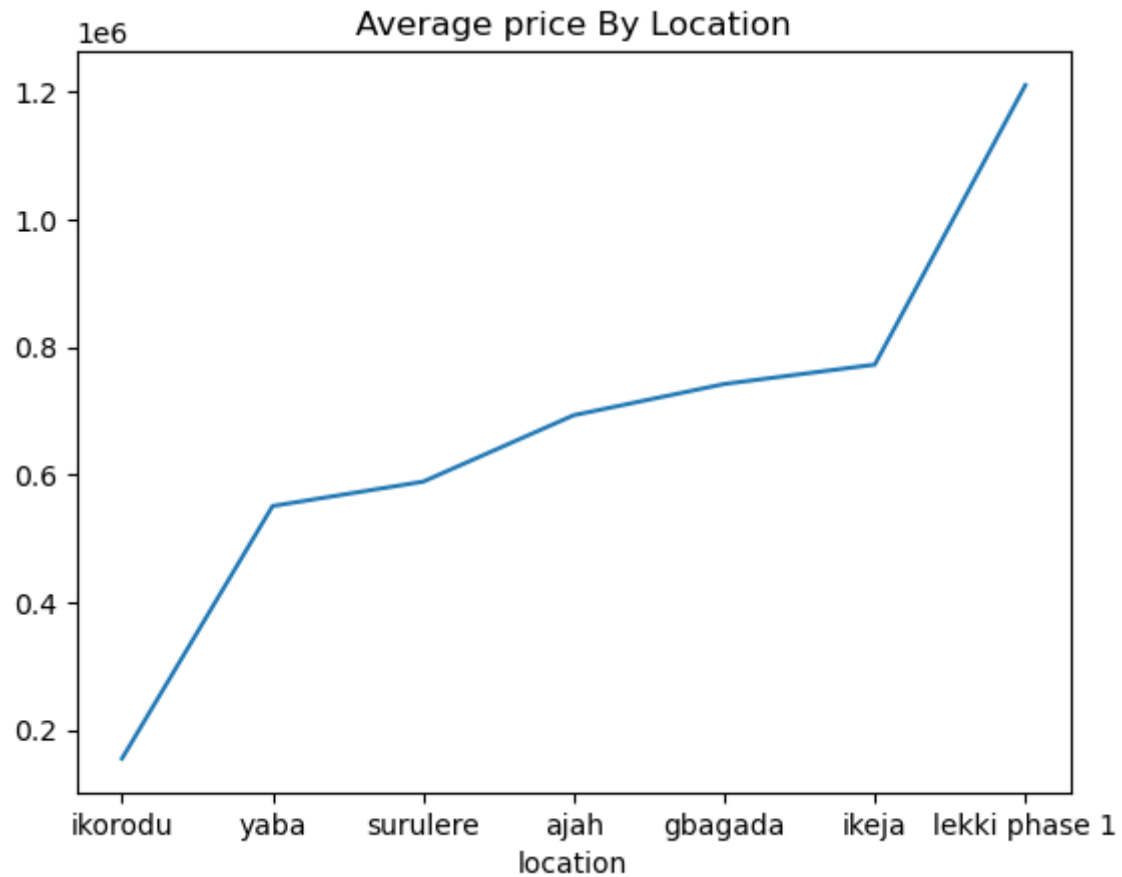

Observation

So far we can deduce that Yaba has the highest amount of listing, Surulere has the highest amount of houses that has CCTV in them as much as 12 listings from Surulere boast of CCTV its rather unclear why lekki phase 1 has no CCTV or pool despite being the highest the most expensive when we look at the average price by location. I think so far it is safe to say that Swimming pool and CCTV has little or no impact on the prices of houses as Surulere who has the highest amount in those have just a little below average when we look at the average of house prices based on location. As we investigate further lets see if we can find relationships that we could link to the bizzare high cost of prices in Lekki Phase 1

Bivarate Analysis

-Lets investige to see if we can link anything in the data to the high cost of houses in lekki phase 1.

```
In [84]:   # Before we go deep into it lets look at the type of Property that has the highest amount of average price
           Prop_type= lagos_houses['price'].groupby(lagos_houses.Property_Type).mean().astype('int').sort_values()
           Prop_type
```

```
Out[84]:   Property_Type
           Self contain              447286
           Mini flat                 590483
           Flat / apartment          977908
           Blocks of flats          1108333
           Penthouse flat           1160714
           Semi detached bungalow   1200000
           Detached bungalow        1313888
           Terraced duplex          1632894
           Detached duplex          1695396
           Semi detached duplex     1747878
           Massionette house        2200000
           Name: price, dtype: int32
```

There is no surprise here as the average prices of each type of property tend to increase as the property type grew bigger. Here the most expensive type of proprty was unsurprisingly the Massionette type of houses while the least expensive was the Self contain appartments

```
In [124…   ### Now lets see if we can see the type of properties listed per location
           counts=lagos_houses.groupby(['location', 'Property_Type']).size().reset_index(name= 'prop')
           table= counts.pivot(index= 'location', columns= 'Property_Type', values= 'prop')
           table=table.fillna(0)
           table.head(7)
```

Out[124]:

Property_Type	Blocks of flats	Detached bungalow	Detached duplex	Flat / apartment	Massionette house	Mini flat	Penthouse flat	Self contain	Semi detached bungalow	Semi detached duplex	Terraced duplex
location											
ajah	15.0	10.0	29.0	89.0	0.0	483.0	5.0	191.0	3.0	47.0	28.0
gbagada	6.0	1.0	4.0	186.0	0.0	387.0	5.0	90.0	0.0	5.0	8.0
ikeja	6.0	1.0	8.0	49.0	0.0	316.0	0.0	70.0	0.0	6.0	1.0
ikorodu	0.0	1.0	0.0	40.0	0.0	296.0	0.0	240.0	0.0	1.0	0.0
lekki phase 1	0.0	1.0	2.0	9.0	1.0	336.0	1.0	167.0	0.0	3.0	1.0
surulere	4.0	4.0	3.0	73.0	0.0	444.0	1.0	196.0	0.0	3.0	0.0
yaba	11.0	0.0	7.0	240.0	0.0	808.0	2.0	391.0	0.0	1.0	0.0

In [139…

```python
fig, ax= plt.subplots(figsize=(10, 6))
plt.rcParams.update({'font.size': 10})
im= ax.imshow(table, cmap='YlOrRd', interpolation='nearest')
ax.set_xticks(range(len(table.columns)))
ax.set_yticks(range(len(table.index)))
ax.set_xticklabels(table.columns, rotation=90)
ax.set_yticklabels(table.index)
ax.set_xlabel('Property_Type')
ax.set_ylabel('Location')
fig.colorbar(im)
plt.show()
```
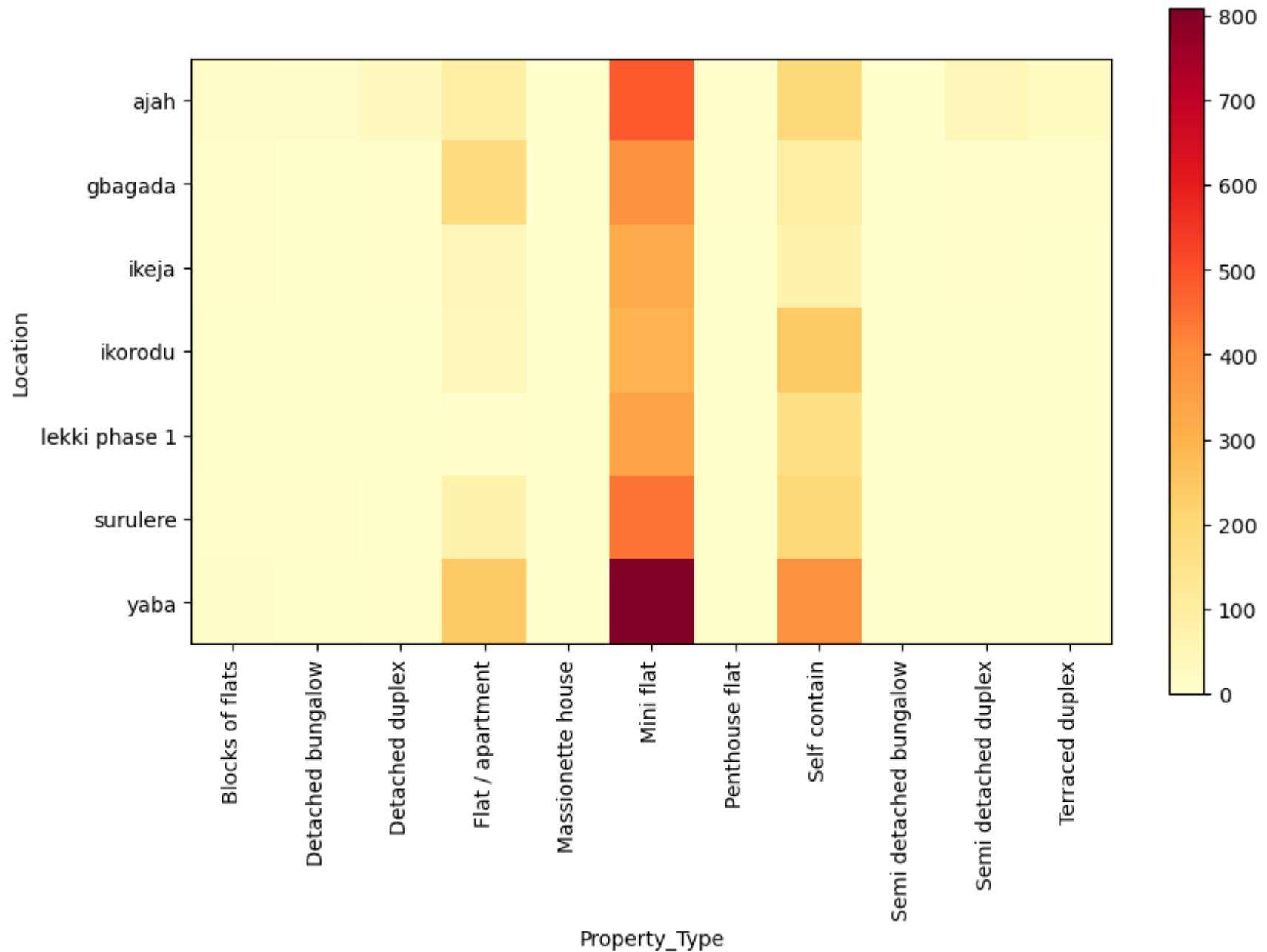
If you look at the table closely you'd discover that we have an aswer to why lekki phase 1 has a very high average price compared to others despite not having swimming pool and cctv cameras like surulere. This is because lekki Phase 1 is the only location that boasts of a massion and

from the property type analysis we did we learnt that that property is outrageously expensive. Ajah seems to be the location that has a perfect distribution for different property Types. Lets dig deeper and see if security has any effect on price

In [147...
```python
S= lagos_houses.groupby('location')['Security'].apply(lambda x: (x== 'yes').sum())
S_Doors=pool= lagos_houses.groupby('location')['Security_Doors'].apply(lambda x: (x== 'yes').sum())
print(S)
print(S_Doors)
```

```
location
ajah               83
gbagada            43
ikeja              37
ikorodu            93
lekki phase 1      45
surulere           39
yaba              218
Name: Security, dtype: int64
location
ajah               5
gbagada            0
ikeja              0
ikorodu            0
lekki phase 1      0
surulere           0
yaba               0
Name: Security_Doors, dtype: int64
```

Conclusion

From the above exploratory analysis I was able to deduce the following to answer the Analysis Objective

- Yaba is the mos densely concentrated amount of listing with over 700+ house listing
- Lekki Phase 1 is the most expensive neighborhood acoording to the dataset as it has the highest amount of average price(though this is largely due to the fact that it is the only location that has a moansion and mansion is the most expensive kind of property.
- Mini Flat is the most comon type of property in lagos and every single location has atleast 200 mini flats while Mansion is the least popular kind of propert with just one mansion in the whole of lagos.
- Ajah is the most security conscious neighborhood with a good number of security outfits and the land lords here are very conscious of installing security doors as landlords in other areas did not install any security door.

- CCTV, Swimming pools and other utilities has little or no effect on the prices of property.

Recommendations

After making the above conclusions I have the following recommendations to make to the landlords

- Landlords should decongest Ajah as the amount of listing in Ajah is too much and excess demand in a location seems to reduce the price of houses in that area
- While building Landlords could concentrate more on security doors rather than utilities like swimming pools and CCTV as the laters has little or no impact on the price of property.
- Priorities should be put in building other types of propertyies sucha as pent houses because houses like mini flats are already over saturated
- Landlords in Lekki Phase one should prioritize installing CCTV cameras and swimming pools into their properties as this would help the price appreciation of their properties

In [ ]:

In [ ]:

In [ ]: