In [3]:

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
import category_encoders as ce
from catboost import CatBoostRegressor
from sklearn.model_selection import train_test_split,KFold, StratifiedKFold
from sklearn.metrics import mean_squared_error
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
import matplotlib.pyplot as plt
from lightgbm import LGBMRegressor
#from sklearn.impute import KNNImputer

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

```
C:\Users\Adekunle\anaconda3\lib\site-packages\xgboost\compat.py:36: Futur
eWarning: pandas.Int64Index is deprecated and will be removed from pandas
in a future version. Use pandas.Index with the appropriate dtype instead.
  from pandas import MultiIndex, Int64Index
```

In [4]:

```
df= pd.read_csv('Housing_dataset_train.csv')
test = pd.read_csv('Housing_dataset_test.csv')
df.head()
```

Out[4]:

	ID	loc	title	bedroom	bathroom	parking_space	price
0	3583	Katsina	Semi-detached duplex	2.0	2.0	1.0	1149999.565
1	2748	Ondo	Apartment	NaN	2.0	4.0	1672416.689
2	9261	Ekiti	NaN	7.0	5.0	NaN	3364799.814
3	2224	Anambra	Detached duplex	5.0	2.0	4.0	2410306.756
4	10300	Kogi	Terrace duplex	NaN	5.0	6.0	2600700.898

In [5]:

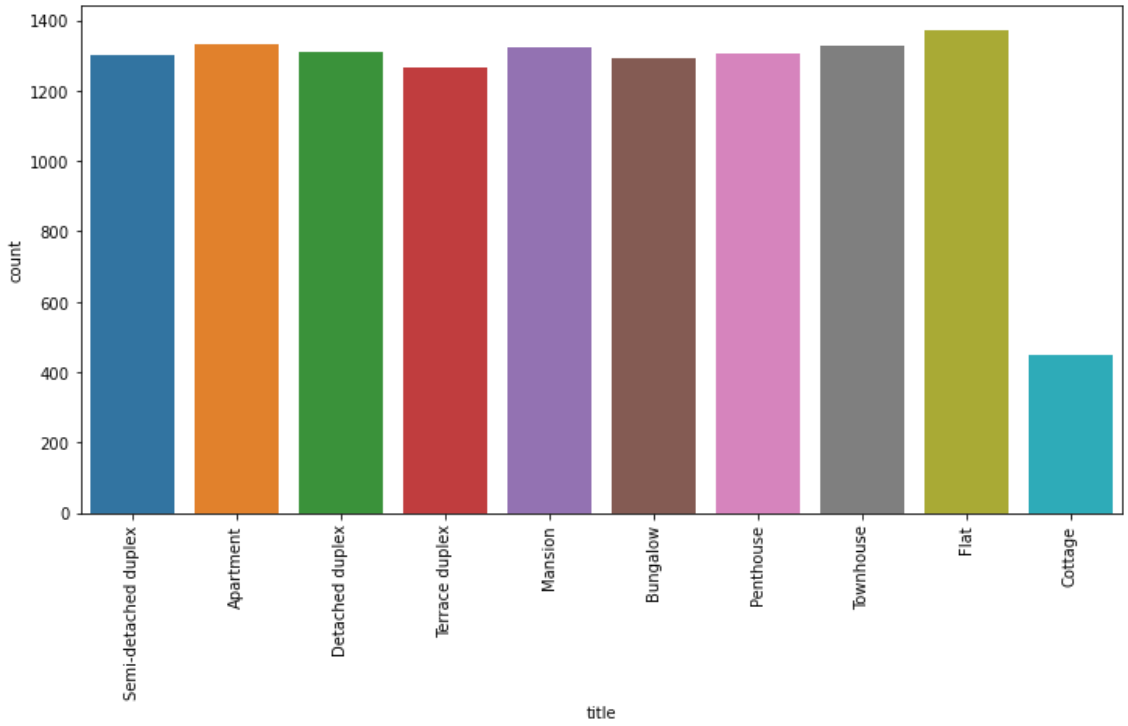
```
sub = pd.read_csv('Sample_submission.csv')
sub.head()
```

Out[5]:

	ID
0	845
1	1924
2	10718
3	12076
4	12254

In [6]:

```
#What is the distribution of house types in the dataset?
plt.figure(figsize=(12,6))
sns.countplot(x='title', data=df)
plt.xticks(rotation=90)
plt.show()
```



What is the distribution of house types in the dataset?

Cottage House types is low compared to others

In [7]:

```
#Which state has the highest number of houses in the dataset?

df['loc'].value_counts()
```

Out[7]:

```
Kaduna         370
Anambra        363
Benue          355
Yobe           353
Borno          351
Kano           351
Nasarawa       349
Cross River    349
Zamfara        348
Imo            348
Ebonyi         346
Kebbi          346
Katsina        345
Ogun           345
Ondo           344
Gombe          343
Bauchi         342
Oyo            341
Adamawa        341
Bayelsa        340
Plateau        338
Osun           338
Jigawa         337
Ekiti          336
Kwara          333
Niger          330
Akwa Ibom      329
Lagos          328
Sokoto         326
Delta          325
Enugu          324
Rivers         323
Kogi           321
Taraba         315
Abia           312
Edo            302
Name: loc, dtype: int64
```

Which state has the highest number of houses in the dataset?

Kaduna has the highest number of houses in the dataset

In [8]:

```
#What is the average house price in each state?

df.groupby('loc').mean()[['price']]
```

Out[8]:

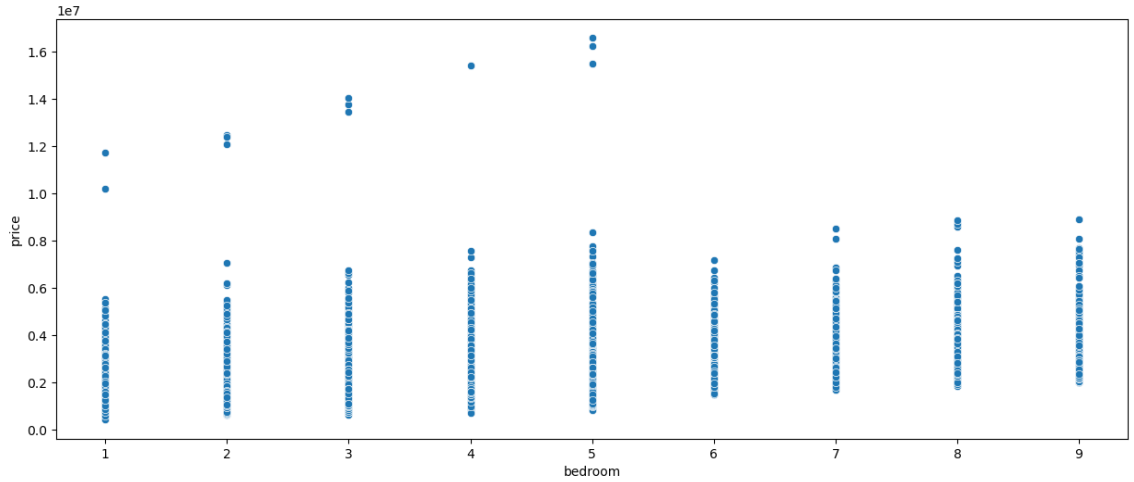
loc	price
Abia	1.717083e+06
Adamawa	1.905287e+06
Akwa Ibom	2.725454e+06
Anambra	2.337230e+06
Bauchi	1.772961e+06
Bayelsa	3.112322e+06
Benue	1.920461e+06
Borno	1.735704e+06
Cross River	2.507765e+06
Delta	2.712493e+06
Ebonyi	1.635850e+06
Edo	2.310452e+06
Ekiti	2.109220e+06
Enugu	2.272887e+06
Gombe	1.860851e+06
Imo	2.067489e+06
Jigawa	1.735867e+06
Kaduna	1.846993e+06
Kano	2.081931e+06
Katsina	1.947589e+06
Kebbi	1.616372e+06
Kogi	1.763416e+06
Kwara	1.903424e+06
Lagos	4.210546e+06
Nasarawa	2.061764e+06
Niger	1.885325e+06
Ogun	2.564020e+06
Ondo	2.277494e+06
Osun	2.180570e+06
Oyo	2.293159e+06
Plateau	1.942316e+06
Rivers	2.957098e+06
Sokoto	1.681016e+06
Taraba	1.855306e+06
Yobe	1.747938e+06
Zamfara	1.689541e+06

In [13]:

```
#Is there any relationship between the number of bedrooms and the house price?
plt.figure(figsize=(15,6),dpi=100)
sns.scatterplot(x='bedroom', y='price', data=df)
```

Out[13]:

```
<AxesSubplot:xlabel='bedroom', ylabel='price'>
```



Is there any relationship between the number of bedrooms and the house price?

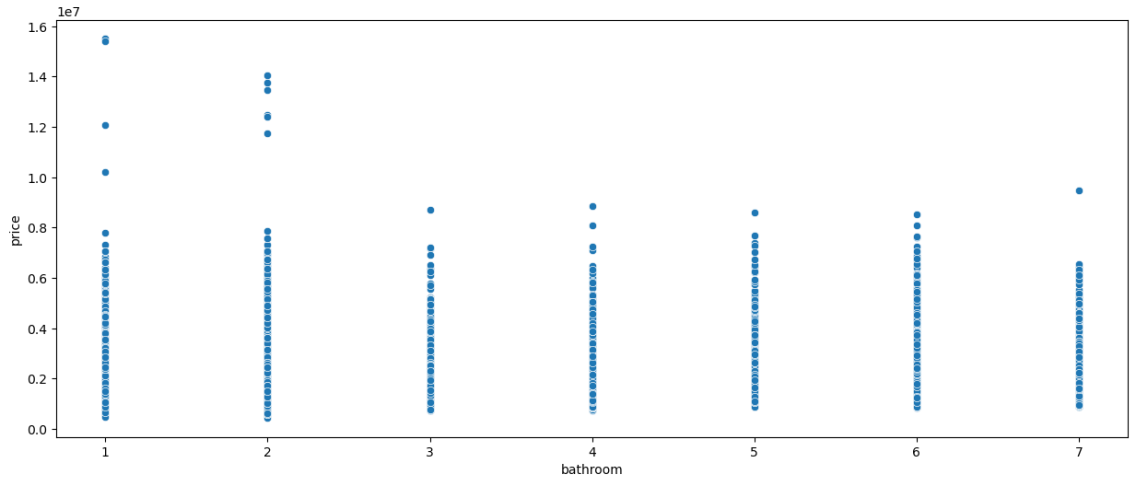
From the above Plot I can see that houses that have 5 bedrooms have the highest price

In [20]:

```
#Is there any relationship between the number of bathrooms and the house price?
plt.figure(figsize=(15,6),dpi=100)
sns.scatterplot(x='bathroom', y='price', data=df)
```

Out[20]:

```
<AxesSubplot:xlabel='bathroom', ylabel='price'>
```



Is there any relationship between the number of bathrooms and the house price?

From the above Plot I can see that houses that have 1 bathroom have the highest price

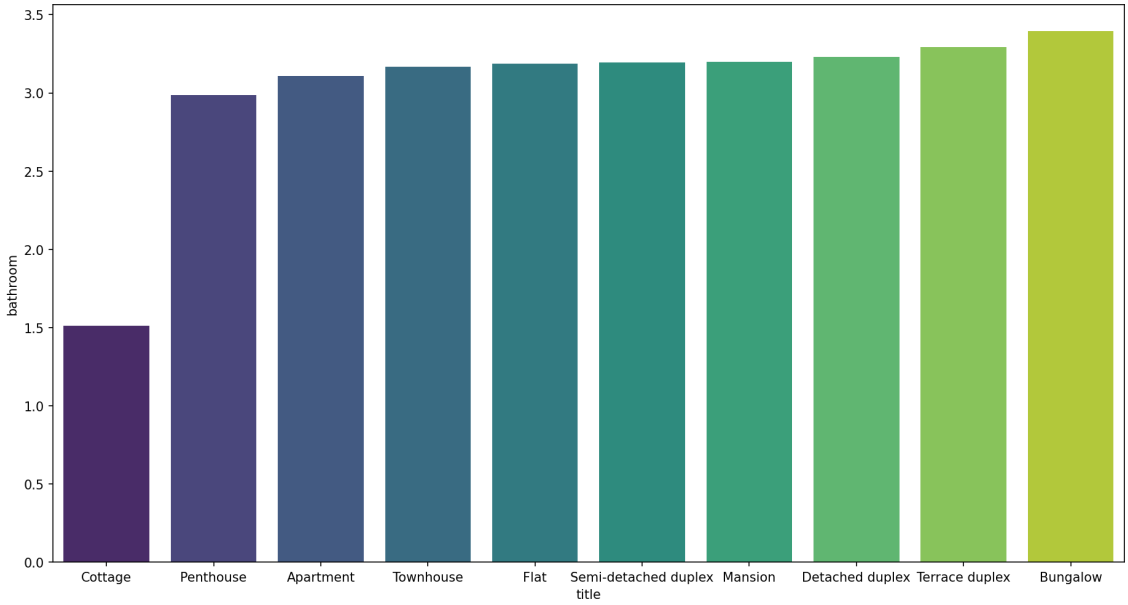
In [19]:

```
#Which house type has the highest average number of bathrooms?


bat = df.groupby('title').mean()[['bathroom']].sort_values('bathroom').reset_index()
plt.figure(figsize=(15,8),dpi=150)
sns.barplot(x='title', y='bathroom', data=bat, palette='viridis')
```

Out[19]:

```
<AxesSubplot:xlabel='title', ylabel='bathroom'>
```

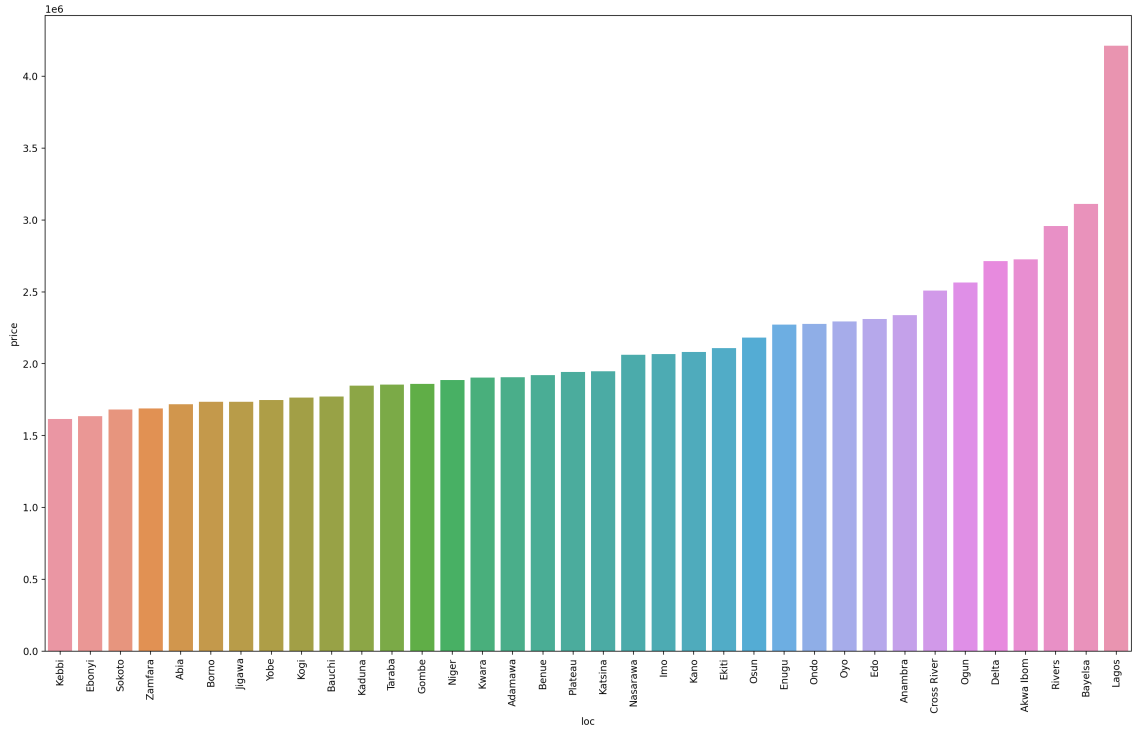


Insight

Bungalow house type has the highest number of bathroom at an average while cottage house type has the lowest

In [27]:

```
#Is there a significant difference in house prices between different states?
lag = df.groupby('loc').mean()[['price']].sort_values('price').reset_index()
plt.figure(figsize=(20,12),dpi=200)
sns.barplot(x='loc', y='price', data=lag)
plt.xticks(rotation=90);
```



Insight

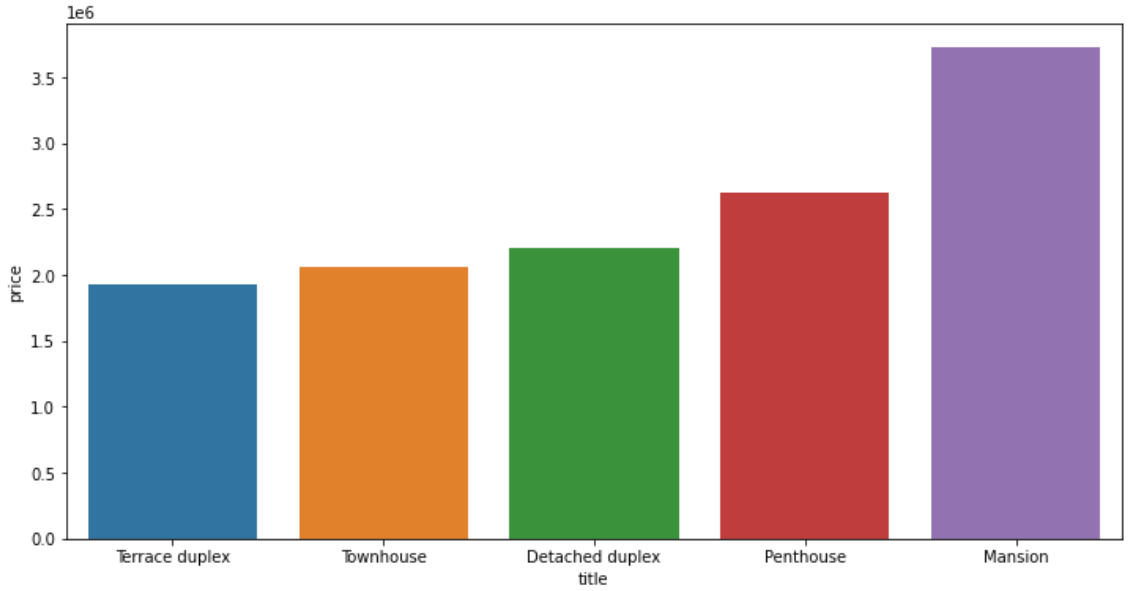
Lagos State has the highest number of Price

In [37]:

```
#What are the top 5 most expensive house types in the dataset?
title = df.groupby('title').mean()[['price']].sort_values('price').tail().reset_index()
plt.figure(figsize=(12,6))
sns.barplot(x='title', y='price', data=title)
```

Out[37]:

```
<AxesSubplot:xlabel='title', ylabel='price'>
```



Insight

Mansion is the most expensive house type in the dataset

In [40]:

```
plt.figure(figsize=(12,6))
sns.heatmap(df.corr(), annot=True, cmap='viridis')
```

Out[40]:

<AxesSubplot:>

Insight

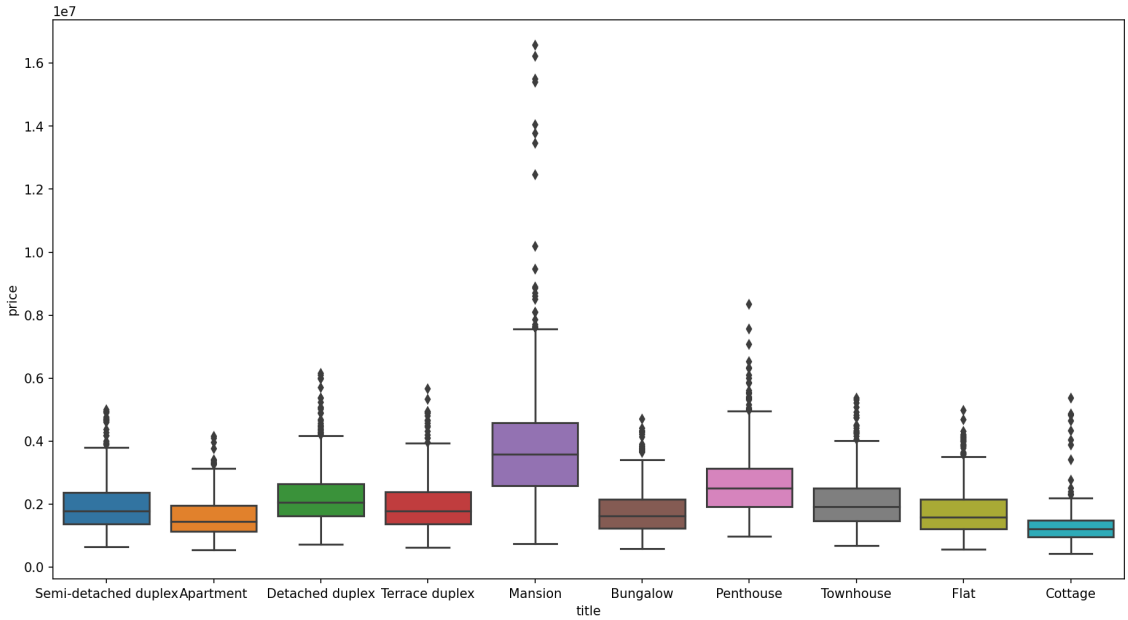
The number of bedrooms and bathrooms seem to be the most correlated of the numerical features

In [41]:

```
#How does the house price vary with different house types?
plt.figure(figsize=(15,8),dpi=150)
sns.boxplot(x='title', y='price' , data=df)
```

Out[41]:

```
<AxesSubplot:xlabel='title', ylabel='price'>
```



In [43]:

```
#What is the average number of bedrooms for each house type?
df.groupby('title').mean()[['bedroom']]
```

Out[43]:

title	bedroom
Apartment	4.344219
Bungalow	4.402852
Cottage	2.905512
Detached duplex	4.327840
Flat	4.378877
Mansion	4.333929
Penthouse	4.342982
Semi-detached duplex	4.414903
Terrace duplex	4.340639
Townhouse	4.298759

In [ ]:

```
```