australia-housing-prices

May 17, 2023

[39]:
```
# import the dataset
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('RealEstateAU_1000_Samples.csv')
df
```

[39]:

	index	TID	breadcrumb	category_name	property_type	building_size	land_size	preferred_size	open_date
0	0	1350988	Buy>NT>DARWIN CITY	Real Estate & Property for sale in DARWIN CITY…	House	NaN	NaN	NaN	Added 2 hours ago
1	1	1350989	Buy>NT>DARWIN CITY	Real Estate & Property for sale in DARWIN CITY…	Apartment	171m²	NaN	171m²	Added 7 hours ago
2	2	1350990	Buy>NT>DARWIN CITY	Real Estate & Property for sale in DARWIN CITY…	Unit	NaN	NaN	NaN	Added 22 hours ago
3	3	1350991	Buy>NT>DARWIN CITY	Real Estate & Property for sale in DARWIN CITY…	House	NaN	NaN	NaN	Added yesterday
4	4	1350992	Buy>NT>DARWIN CITY	Real Estate & Property for sale in DARWIN CITY…	Unit				
..	…	…	…	…	…				
995	995	1351983	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT …	House				
996	996	1351984	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT …	House				
997	997	1351985	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT …	House				
998	998	1351986	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT …	House				
999	999	1351987	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT …	Unit				

```
4           201m²       NaN          201m²     Added yesterday
..            …         …             …              …
995           NaN    9.17ha         9.17ha         Under offer
996         203m²     600m²          600m²                NaN
997       209.6m²     800m²          800m²                NaN
998         180m²     450m²          450m²                NaN
999         120m²       NaN          120m²                NaN

                                    listing_agency  … state  zip_code  \
0                        Professionals - DARWIN CITY  …    NT       800
1    Nick Mousellis Real Estate - Eview Group Member  …    NT       800
2                  Habitat Real Estate - THE GARDENS  …    NT       800
3                              Ray White - NIGHTCLIFF  …    NT       800
4                Carol Need Real Estate - Fannie Bay  …    NT       800
..                                                  …  …   …        …
995                         United Realty NT - Parap  …    NT       834
996            Kassiou Constructions - HOWARD SPRINGS  …    NT       836
997            Kassiou Constructions - HOWARD SPRINGS  …    NT       836
998            Kassiou Constructions - HOWARD SPRINGS  …    NT       810
999                              Home Zone NT - DARWIN  …    NT       820

            phone latitude longitude product_depth bedroom_count  \
0    08 8941 8289      NaN       NaN      premiere           2.0
1       411724000      NaN       NaN      premiere           3.0
2    08 8981 0080      NaN       NaN      premiere           2.0
3    08 8982 2403      NaN       NaN      premiere           1.0
4       418885966      NaN       NaN      premiere           3.0
..            …       …         …             …             …
995  08 8981 2666      NaN       NaN       feature           4.0
996   08 89834326      NaN       NaN      standard           4.0
997   08 89834326      NaN       NaN      standard           4.0
998   08 89834326      NaN       NaN      standard           4.0
999  0418 895 345      NaN       NaN       feature           2.0

    bathroom_count  parking_count           RunDate
0              1.0            1.0  27-05-2022 15:54
1              2.0            2.0  27-05-2022 15:54
2              1.0            1.0  27-05-2022 15:54
3              1.0            0.0  27-05-2022 15:54
4              2.0            2.0  27-05-2022 15:54
..             …              …              …
995            3.0            6.0  27-05-2022 15:54
996            2.0            2.0  27-05-2022 15:54
997            2.0            2.0  27-05-2022 15:54
998            2.0            3.0  27-05-2022 15:54
999            2.0            2.0  27-05-2022 15:54
```

```
[1000 rows x 27 columns]
```

[40]:
```
df.columns
```

[40]:
```
Index(['index', 'TID', 'breadcrumb', 'category_name', 'property_type',
       'building_size', 'land_size', 'preferred_size', 'open_date',
       'listing_agency', 'price', 'location_number', 'location_type',
       'location_name', 'address', 'address_1', 'city', 'state', 'zip_code',
       'phone', 'latitude', 'longitude', 'product_depth', 'bedroom_count',
       'bathroom_count', 'parking_count', 'RunDate'],
      dtype='object')
```

[41]:
```
df.shape
```

[41]:
```
(1000, 27)
```

[42]:
```
df.isnull().sum()
```

[42]:
```
index              0
TID                0
breadcrumb         0
category_name      0
property_type      0
building_size    720
land_size        467
preferred_size   391
open_date        698
listing_agency     0
price              0
location_number    0
location_type      0
location_name      0
address           12
address_1         12
city               0
state              0
zip_code           0
phone              0
latitude        1000
longitude       1000
product_depth      0
bedroom_count     33
bathroom_count    33
parking_count     33
RunDate            0
dtype: int64
```

[43]:
```
df.describe()
```

[43]:

	index	TID	location_number	zip_code	latitude	\
count	1000.000000	1.000000e+03	1.000000e+03	1000.00000	0.0	
mean	499.500000	1.351488e+06	1.474125e+08	816.64600	NaN	
std	288.819436	2.888194e+02	6.121381e+07	13.22057	NaN	
min	0.000000	1.350988e+06	1.085305e+08	800.00000	NaN	
25%	249.750000	1.351238e+06	1.386598e+08	800.00000	NaN	
50%	499.500000	1.351488e+06	1.390458e+08	820.00000	NaN	
75%	749.250000	1.351737e+06	1.393042e+08	830.00000	NaN	
max	999.000000	1.351987e+06	7.001996e+08	839.00000	NaN	

	longitude	bedroom_count	bathroom_count	parking_count
count	0.0	967.000000	967.000000	967.000000
mean	NaN	2.866598	1.739400	2.152017
std	NaN	1.151914	0.635663	1.514818
min	NaN	0.000000	1.000000	0.000000
25%	NaN	2.000000	1.000000	1.000000
50%	NaN	3.000000	2.000000	2.000000
75%	NaN	4.000000	2.000000	2.000000
max	NaN	9.000000	5.000000	12.000000

[44]:
```
df = df.drop( columns = ['TID', 'breadcrumb','category_name', 'preferred_size',␣
 ↪'open_date', 'zip_code', 'phone', 'latitude', 'longitude' ,'product_depth',␣
 ↪'RunDate', 'address', 'address_1'])
df.describe()
```

[44]:

	index	location_number	bedroom_count	bathroom_count	\
count	1000.000000	1.000000e+03	967.000000	967.000000	
mean	499.500000	1.474125e+08	2.866598	1.739400	
std	288.819436	6.121381e+07	1.151914	0.635663	
min	0.000000	1.085305e+08	0.000000	1.000000	
25%	249.750000	1.386598e+08	2.000000	1.000000	
50%	499.500000	1.390458e+08	3.000000	2.000000	
75%	749.250000	1.393042e+08	4.000000	2.000000	
max	999.000000	7.001996e+08	9.000000	5.000000	

	parking_count
count	967.000000
mean	2.152017
std	1.514818
min	0.000000
25%	1.000000
50%	2.000000
75%	2.000000
max	12.000000

[45]:
```
df.columns
```

[45]:
```
Index(['index', 'property_type', 'building_size', 'land_size',
       'listing_agency', 'price', 'location_number', 'location_type',
       'location_name', 'city', 'state', 'bedroom_count', 'bathroom_count',
       'parking_count'],
      dtype='object')
```

[46]:
```
df['price'] = df['price'].astype(str)
df['price'] = df['price'].str.extract('(\d+)', expand=False)
df['price'] = pd.to_numeric(df['price'])

df['building_size'] = df['building_size'].astype(str)
df['building_size'] = df['building_size'].str.replace(r'\D', '', regex=True)
df['building_size'] = pd.to_numeric(df['building_size'])

df['land_size'] = df['land_size'].astype(str)
df['land_size'] = df['land_size'].str.replace(r'\D', '', regex=True)
df['land_size'] = pd.to_numeric(df['land_size'])

df.head()
```

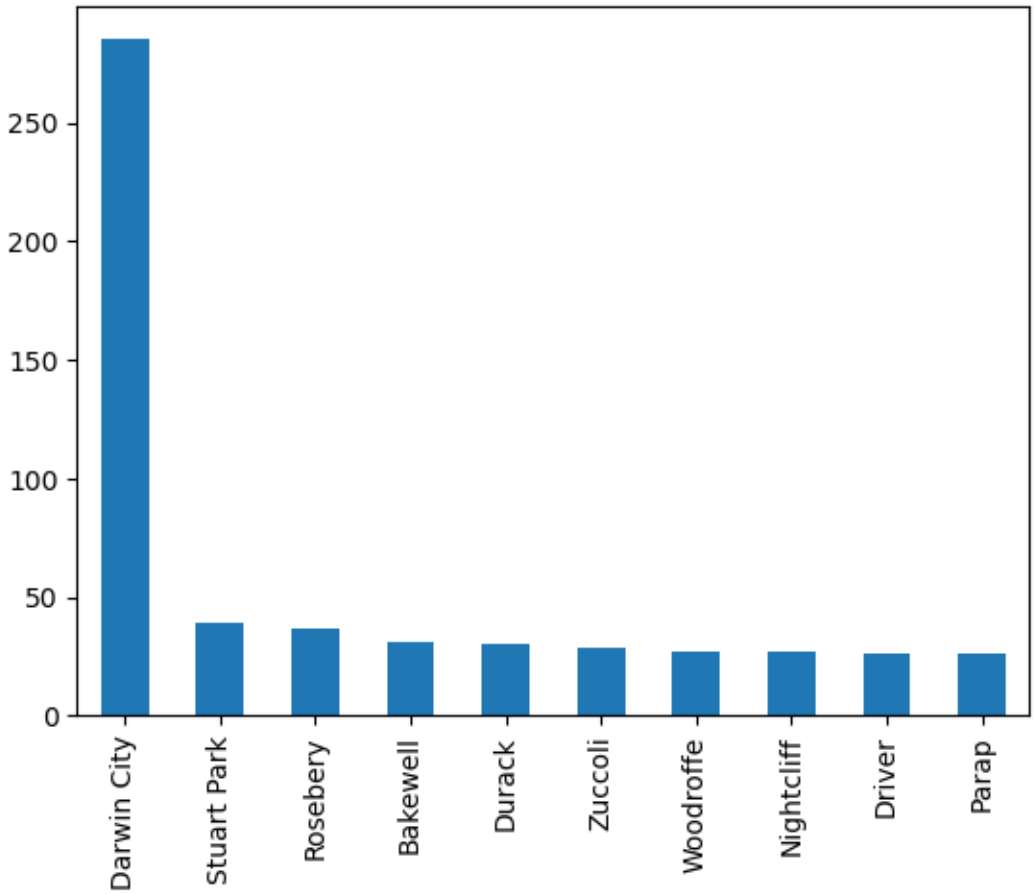
[46]:

	index	property_type	building_size	land_size	listing_agency	price	location_number	location_type	location_name	city	state	bedroom_count	bathroom_count	parking_count
0	0	House	NaN	NaN	Professionals - DARWIN CITY	435.0	139468611	Buy	$435,000	Darwin City	NT	2.0	1.0	1.0
1	1	Apartment	171.0	NaN	Nick Mousellis Real Estate - Eview Group Member	320.0	139463755	Buy	Offers Over $320,000	Darwin City	NT	3.0	2.0	2.0
2	2	Unit	NaN	NaN	Habitat Real Estate - THE GARDENS	310.0	139462495	Buy	$310,000	Darwin City	NT	2.0	1.0	1.0
3	3	House	NaN	NaN	Ray White - NIGHTCLIFF	259.0	139451679	Buy	$259,000	Darwin City	NT	1.0		
4	4	Unit	201.0	NaN	Carol Need Real Estate - Fannie Bay	439.0	139433803	Buy	$439,000	Darwin City	NT	3.0		

```
3              1.0            0.0
4              2.0            2.0
```

[47]:
```
city =  df['city'].value_counts()
city.head(10).plot.bar()

plt.show()
```

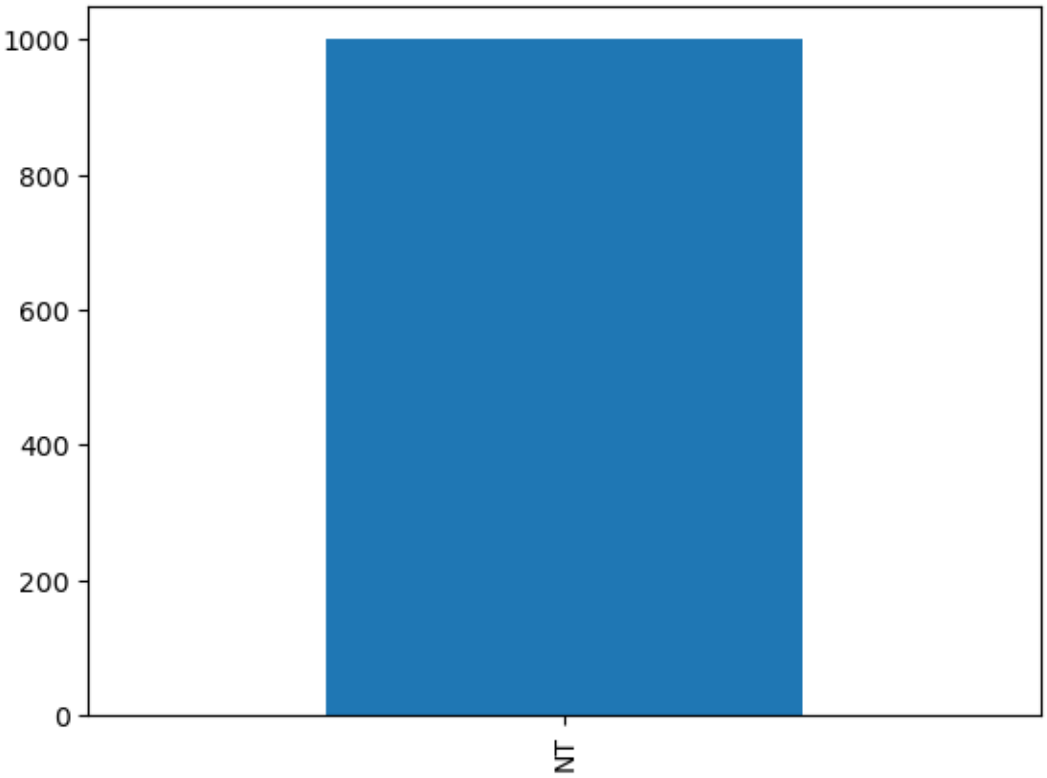


[48]:
```
df.columns
```

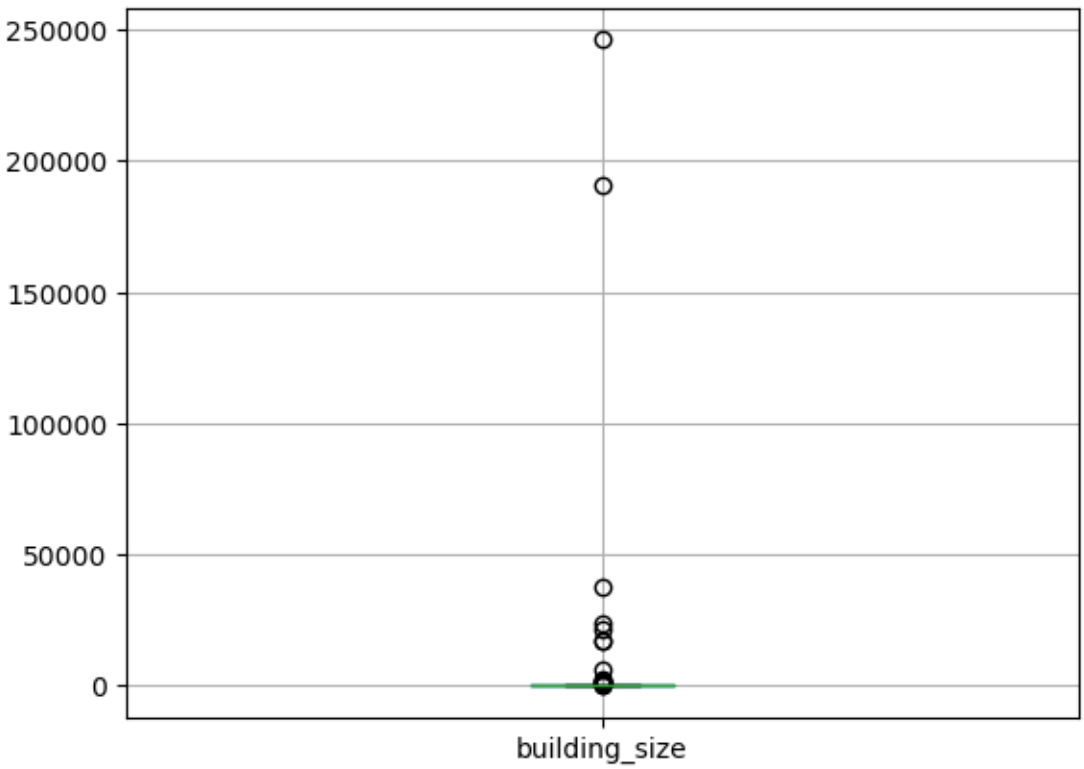
[48]:
```
Index(['index', 'property_type', 'building_size', 'land_size',
       'listing_agency', 'price', 'location_number', 'location_type',
       'location_name', 'city', 'state', 'bedroom_count', 'bathroom_count',
       'parking_count'],
      dtype='object')
```

[49]:
```
state =  df['state'].value_counts()
state.head().plot.bar()
```

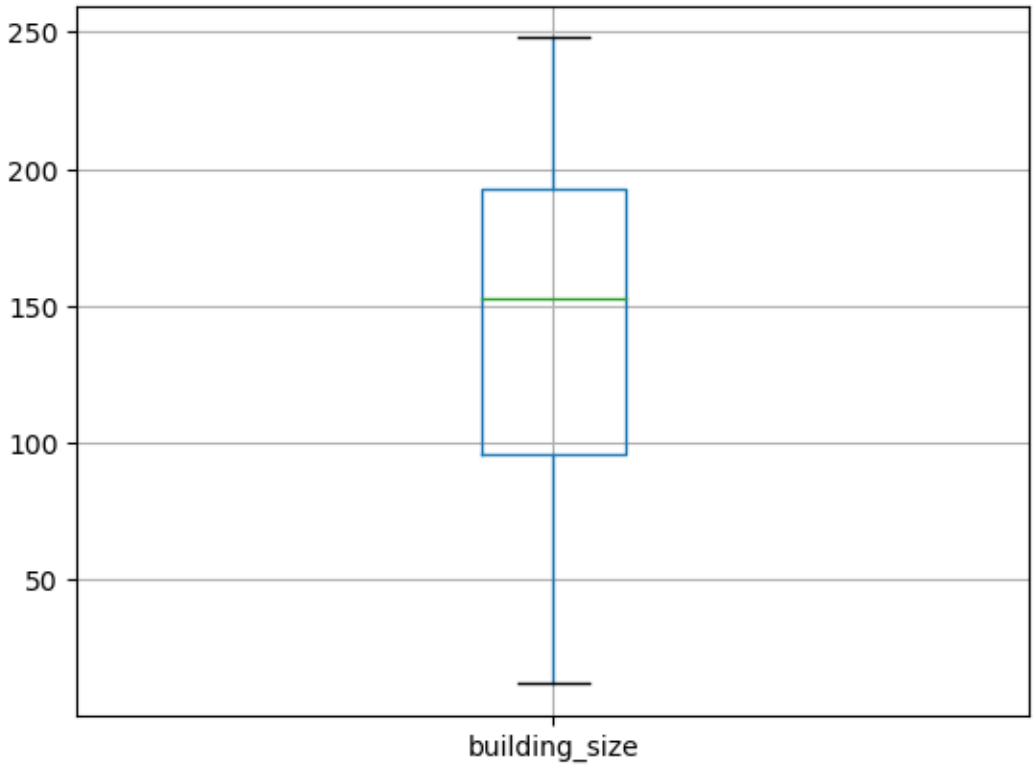
```
plt.show()
```



[50]:
```
boxplot = df.boxplot(column=['building_size'])
```



[51]:
```
df = df[df['building_size'] < 250]
boxplot = df.boxplot(column=['building_size'])
```



[52]:
```
df = df.sort_values(by=['price'], ascending=False)
df = df[df['price'] < 400]
df.dtypes
df.head(10)
```

[52]:

	index	property_type	building_size	land_size	\
655	655	Apartment	130.0	NaN	
999	999	Unit	120.0	NaN	
49	49	Apartment	130.0	NaN	
912	912	Townhouse	103.0	212.0	
307	307	Apartment	184.0	NaN	
492	492	House	138.0	815.0	
735	735	Apartment	138.0	NaN	
58	58	Apartment	138.0	NaN	
141	141	Apartment	99.0	NaN	
236	236	House	180.0	819.0	

	listing_agency	price	location_number	\
655	Call2View Real Estate - Palmerston	399.0	139024171	
999	Home Zone NT - DARWIN	399.0	138330946	
49	Call2View Real Estate - Palmerston	399.0	139024171	
912	@realty - National Head Office Australia	390.0	138557035	

```
307                          Smith Real Estate NT  385.0        139372067
492    Real Estate NT by George Pikos - FANNIE BAY  380.0        139202223
735                        Raine & Horne - Darwin  375.0        138917719
58                         Raine & Horne - Darwin  375.0        138917719
141                         LJ Hooker Darwin -     370.0        136546906
236  Real Value Properties NT - Northern Territory  370.0        139430295

    location_type         location_name         city state  bedroom_count  \
655           Buy              $399,000  Darwin City    NT            2.0
999           Buy              $399,000  Stuart Park    NT            2.0
49            Buy              $399,000  Darwin City    NT            2.0
912           Buy             $390,000+     Rosebery    NT            3.0
307           Buy              $385,000  Stuart Park    NT            2.0
492           Buy              $380,000    Woodroffe    NT            3.0
735           Buy  Offers Over $375,000  Darwin City    NT            2.0
58            Buy  Offers Over $375,000  Darwin City    NT            2.0
141           Buy              $370,000  Darwin City    NT            2.0
236           Buy  OFFERS OVER $370,000      Moulden    NT            3.0

     bathroom_count  parking_count
655             2.0            2.0
999             2.0            2.0
49              2.0            2.0
912             2.0            2.0
307             2.0            2.0
492             1.0            1.0
735             2.0            2.0
58              2.0            2.0
141             2.0            0.0
236             1.0            4.0
```

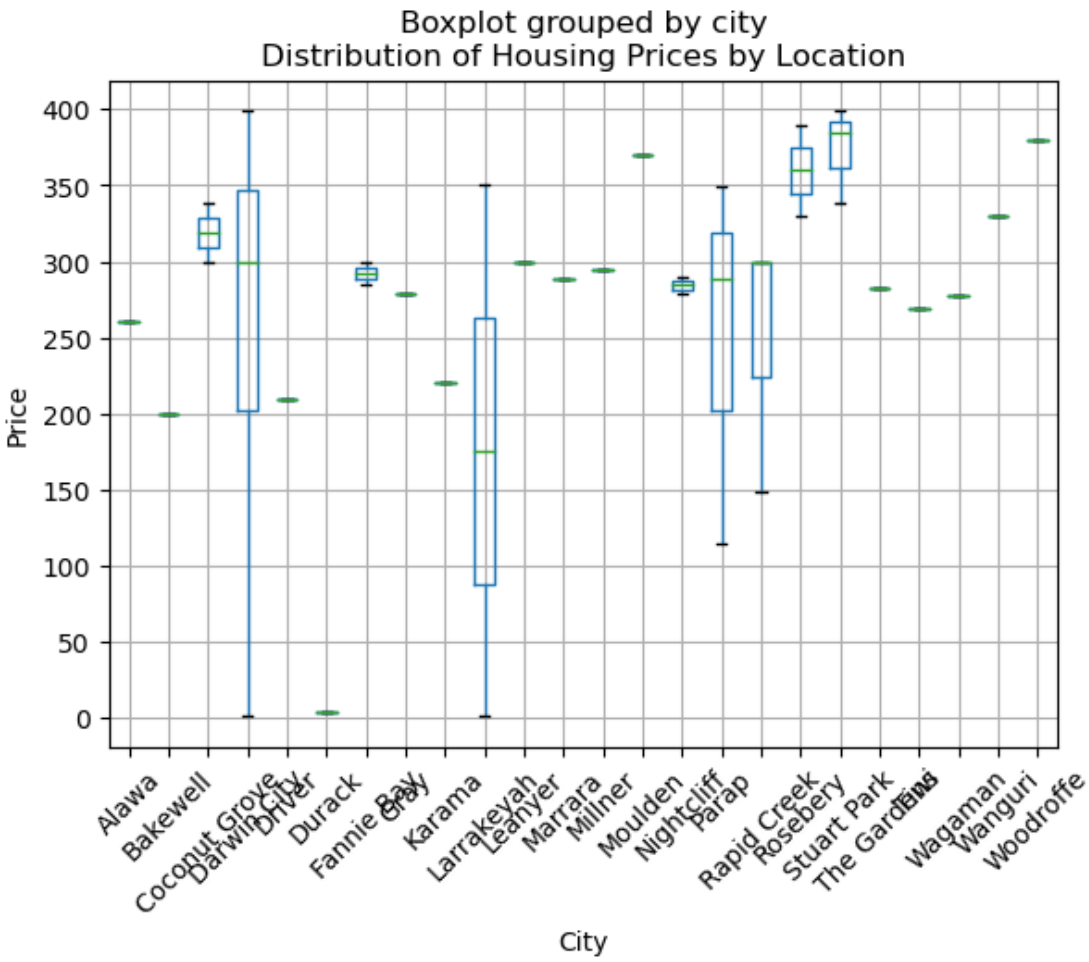
[53]:
```
# Relationship between housing prices and location
city_price = df[['price', 'city']].copy()
city_price.groupby('city').mean().sort_values(by=['price'], ascending=False)
```

[53]:
```
                   price
city
Woodroffe      380.000000
Stuart Park    374.333333
Moulden        370.000000
Rosebery       360.000000
Wanguri        330.000000
Coconut Grove  319.000000
Leanyer        300.000000
Millner        295.000000
Fannie Bay     292.000000
Marrara        289.000000
```

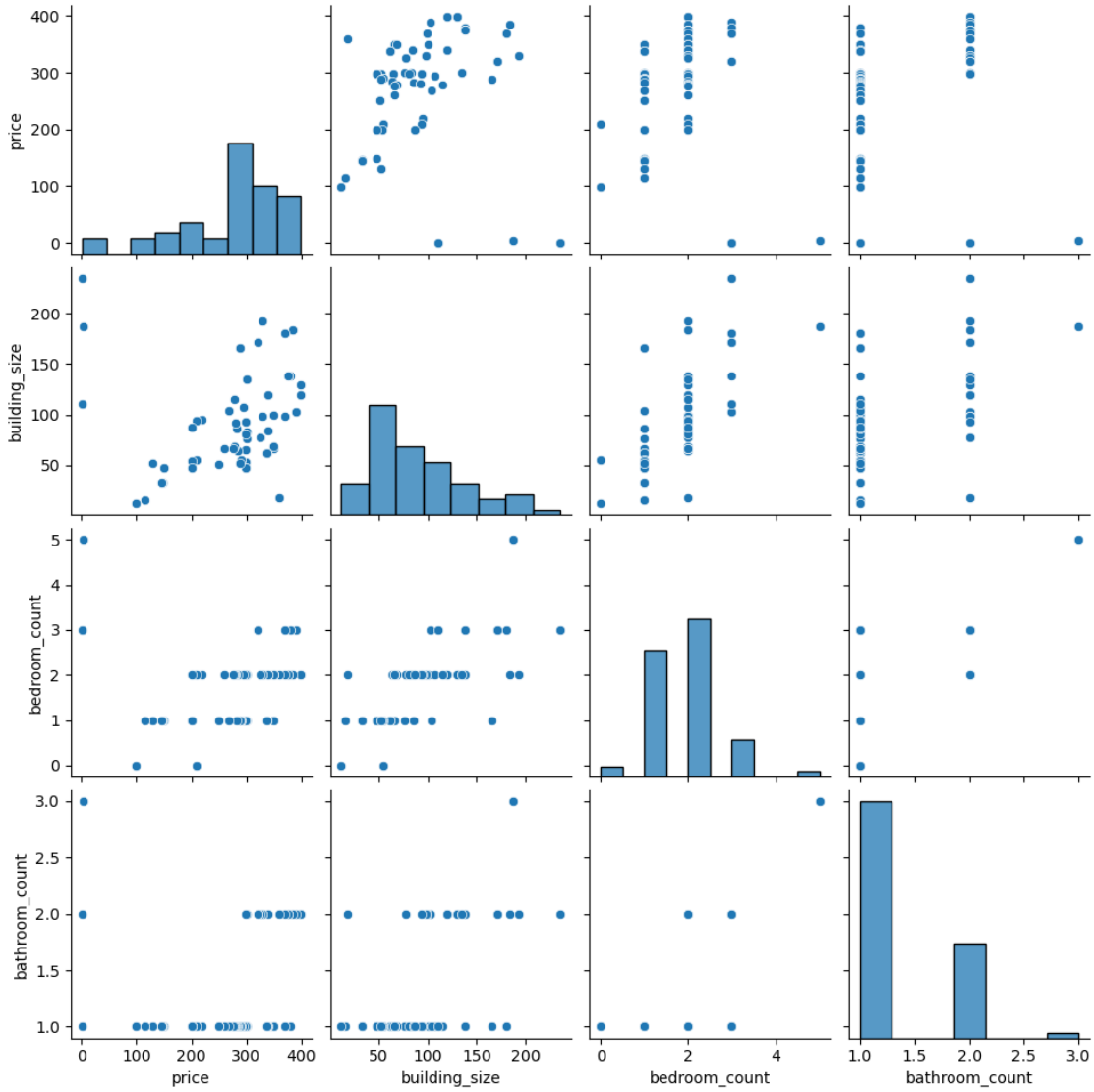
```
Nightcliff      284.500000
The Gardens     283.000000
Gray            279.000000
Wagaman         277.000000
Darwin City     272.333333
Tiwi            269.000000
Alawa           260.000000
Parap           251.000000
Rapid Creek     249.333333
Karama          220.000000
Driver          209.000000
Bakewell        200.000000
Larrakeyah      175.500000
Durack            4.000000
```

[54]:
```
plt.figure(figsize=(10, 6))
df.boxplot(column='price', by='city')
plt.xlabel('City')
plt.ylabel('Price')
plt.title('Distribution of Housing Prices by Location')
plt.xticks(rotation=45)
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



[55]:
```
# Relationship between housing prices and features such as size, number of␣
↪bedrooms, and number of bathrooms
import seaborn as sns
sns.pairplot(df[['price', 'building_size', 'bedroom_count', 'bathroom_count']])
plt.show()
```


price
building_size
bedroom_count
bathroom_count
0
100
200
300
400
50
100
150
200
0
1
2
3
4
5
1.0
1.5
2.0
2.5
3.0