In [34]:
```
import numpy as np # linear algebra
import pandas as pd
from pandas import read_csv
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression
import sklearn.metrics as metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score

pd.set_option('display.float_format', lambda x: '%.3f' % x)
```

In [ ]:
```
```

In [35]:
```
filename = 'RealEstateAU_1000_Samples.csv'
df = read_csv(filename)

df.sample(5)
```

Out[35]:

	index	TID	breadcrumb	category_name	property_type	building_size	land_size	pr
133	133	1351121	Buy>NT>DARWIN CITY	Real Estate & Property for sale in DARWIN CITY...	Apartment	NaN	NaN	
749	749	1351737	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT ...	House	NaN	868m²	
320	320	1351308	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT ...	Residential Land	NaN	450m²	
901	901	1351889	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT ...	House	NaN	782m²	
376	376	1351364	Buy>NT>DARWIN	Real Estate & Property for sale in DARWIN, NT ...	House	141m²	2.02ha	

5 rows × 27 columns

In [36]:
```
df = df.drop( columns = ['TID', 'breadcrumb','category_name', 'preferred_size', 'o
df.describe()
```

Out[36]:

	index	location_number	bedroom_count	bathroom_count	parking_count
count	1000.000	1000.000	967.000	967.000	967.000
mean	499.500	147412522.866	2.867	1.739	2.152
std	288.819	61213814.676	1.152	0.636	1.515
min	0.000	108530451.000	0.000	1.000	0.000
25%	249.750	138659781.000	2.000	1.000	1.000
50%	499.500	139045835.000	3.000	2.000	2.000
75%	749.250	139304193.000	4.000	2.000	2.000
max	999.000	700199623.000	9.000	5.000	12.000

In [37]:
```
df.head()
```

Out[37]:

	index	property_type	building_size	land_size	listing_agency	price	location_number	locati
0	0	House	NaN	NaN	Professionals - DARWIN CITY	$435,000	139468611	
1	1	Apartment	171m²	NaN	Nick Mousellis Real Estate - Eview Group Member	Offers Over $320,000	139463755	
2	2	Unit	NaN	NaN	Habitat Real Estate - THE GARDENS	$310,000	139462495	
3	3	House	NaN	NaN	Ray White - NIGHTCLIFF	$259,000	139451679	
4	4	Unit	201m²	NaN	Carol Need Real Estate - Fannie Bay	$439,000	139433803	

gotta fix the price column

In [38]:
```
df = df[pd.notnull(df['bedroom_count'])]
df = df[pd.notnull(df['building_size'])]
df = df[pd.notnull(df['land_size'])]
df = df[pd.notnull(df['price'])]


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

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

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


df.head()
```

```
C:\Users\Mahi\AppData\Local\Temp\ipykernel_3440\1830399713.py:10: FutureWarning: T
he default value of regex will change from True to False in a future version.
  df['building_size'] = df['building_size'].str.replace(r'\D', '')
C:\Users\Mahi\AppData\Local\Temp\ipykernel_3440\1830399713.py:13: FutureWarning: T
he default value of regex will change from True to False in a future version.
  df['land_size'] = df['land_size'].str.replace(r'\D', '')
```

Out[38]:

	index	property_type	building_size	land_size	listing_agency	price	location_number	loca
26	26	Unit	340	340	Colliers International - Darwin	795.000	139277147	
42	42	House	81	81	Ray White City (NT) -	450.000	139095611	
88	88	Unit	80	92	Nick Mousellis Real Estate - Eview Group Member	580.000	138568447	
153	153	Unit	92	92	No Agent Property - BRIGHTON EAST	280.000	134958294	
162	162	Unit	54	54	Raine & Horne - Darwin	200.000	130692394	

In [39]:

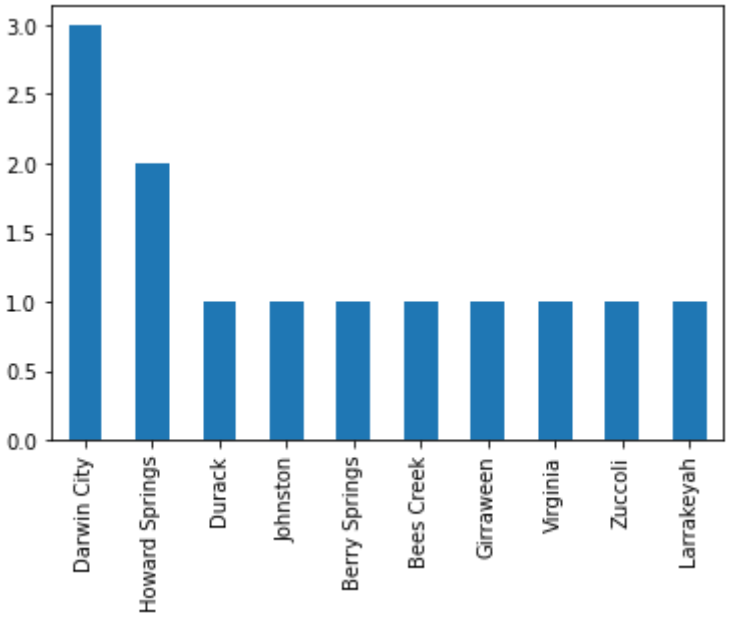
```
df = df.sort_values(by=['price'], ascending=False)
df = df[df['price'] < 400]
df.dtypes
df.head(10)
```

Out[39]:

	index	property_type	building_size	land_size	listing_agency	price	location_number	loca
912	912	Townhouse	103	212	@realty - National Head Office Australia	390.000	138557035	
492	492	House	138	815	Real Estate NT by George Pikos - FANNIE BAY	380.000	139202223	
236	236	House	180	819	Real Value Properties NT - Northern Territory	370.000	139430295	
424	424	Unit	120	120	LJ Hooker Darwin -	339.000	139272035	
268	268	Unit	65	65	Real Estate Central - DARWIN CITY	299.000	139403051	
797	797	Unit	52	80	For Sale By Owner ...	289.000	138792207	
263	263	Unit	64	84	@realty - National Head Office Australia	285.000	139411639	
153	153	Unit	92	92	No Agent Property - BRIGHTON EAST	280.000	134958294	
683	683	Unit	115	115	Call2View Real Estate - Palmerston	279.000	138979211	
887	887	Unit	66	115	Nick Mousellis Real Estate - Eview Group Member	277.000	138610195	

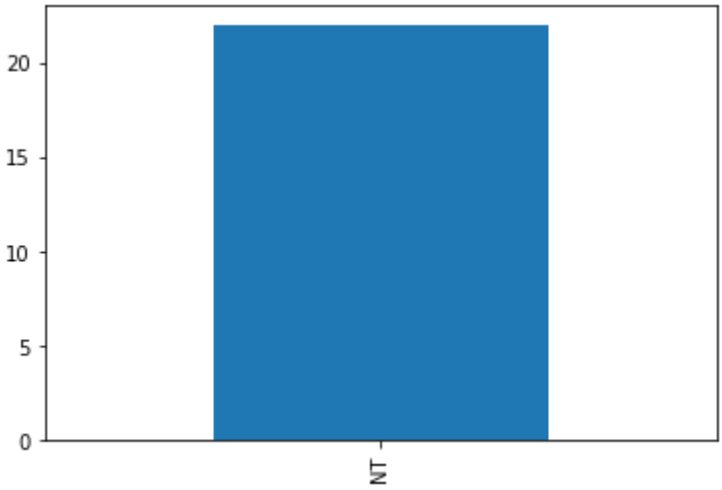
Now let's see where the most properties are located

In [40]:
```
city =  df['city'].value_counts()
city.head(10).plot.bar()

plt.show()
```



Next let's look at states

In [41]:
```
state =  df['state'].value_counts()
state.head(5).plot.bar()

plt.show()
```

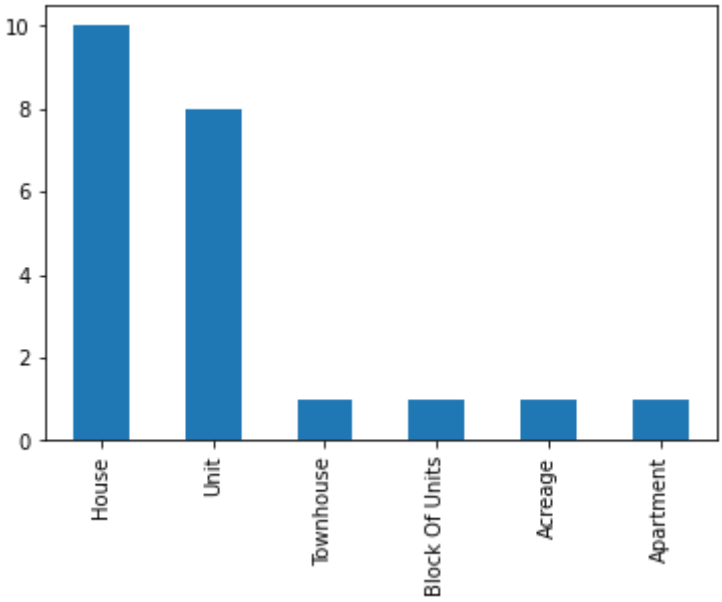


In [42]:
```
df.state.unique()
```

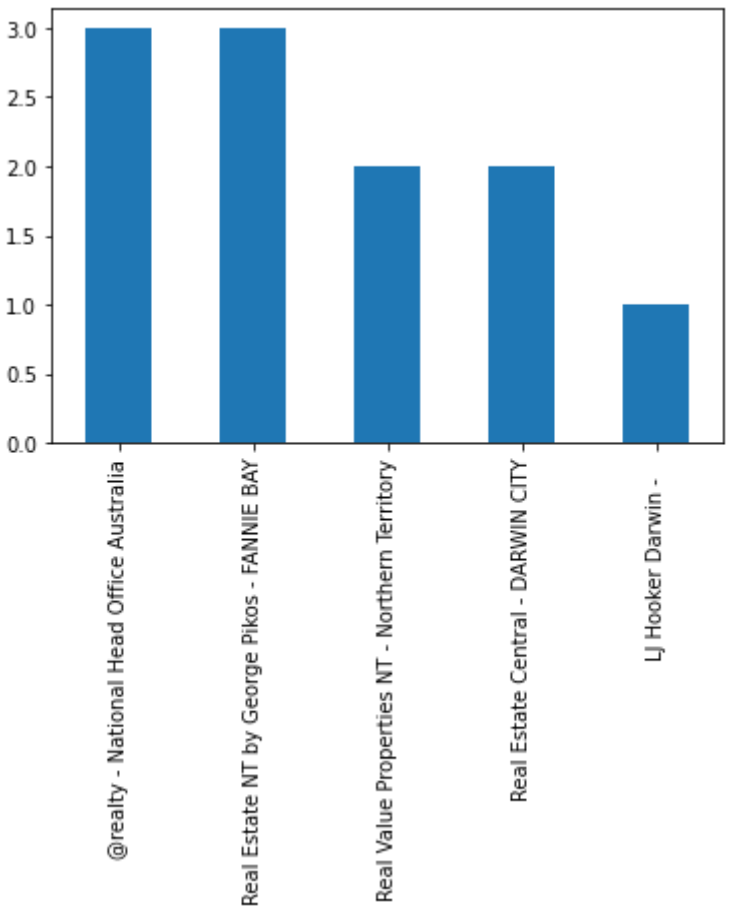
Out[42]:
```
array(['NT'], dtype=object)
```

so all the data come from the same state. now let's look at property types

In [43]:
```
protype =  df['property_type'].value_counts()
protype.head(10).plot.bar()

plt.show()
```



In [44]:
```
l_agenc =  df['listing_agency'].value_counts()
l_agenc.head(5).plot.bar()

plt.show()
```



now let's see where the mean price of each company

In [45]:
```
agen_price = df[['price', 'listing_agency']].copy()
agen_price.groupby('listing_agency').mean()
agen_price.sort_values(by=['price'], ascending=False)
```

Out[45]:

	price	listing_agency
912	390.000	@realty - National Head Office Australia
492	380.000	Real Estate NT by George Pikos - FANNIE BAY
236	370.000	Real Value Properties NT - Northern Territory
424	339.000	LJ Hooker Darwin -
268	299.000	Real Estate Central - DARWIN CITY
797	289.000	For Sale By Owner ...
263	285.000	@realty - National Head Office Australia
153	280.000	No Agent Property - BRIGHTON EAST
683	279.000	Call2View Real Estate - Palmerston
887	277.000	Nick Mousellis Real Estate - Eview Group Member
162	200.000	Raine & Horne - Darwin
339	4.000	Elders Real Estate - Palmerston
661	1.000	Elders Real Estate - Darwin
788	1.000	Kassiou Constructions - HOWARD SPRINGS
260	1.000	Piening Property Sales - COOLALINGA
948	1.000	Real Value Properties NT - Northern Territory
954	1.000	Real Estate NT by George Pikos - FANNIE BAY
969	1.000	@realty - National Head Office Australia
667	1.000	Real Estate NT by George Pikos - FANNIE BAY
318	1.000	Real Estate Central - DARWIN CITY
671	1.000	Renee's Realty NT - DURACK
183	1.000	David Booth Real Estate Pty Ltd - Darwin

lets also look at the most expensive cities

In [46]:
```
city_price = df[['price', 'city']].copy()
city_price.groupby('city').mean().sort_values(by=['price'], ascending=False)
```

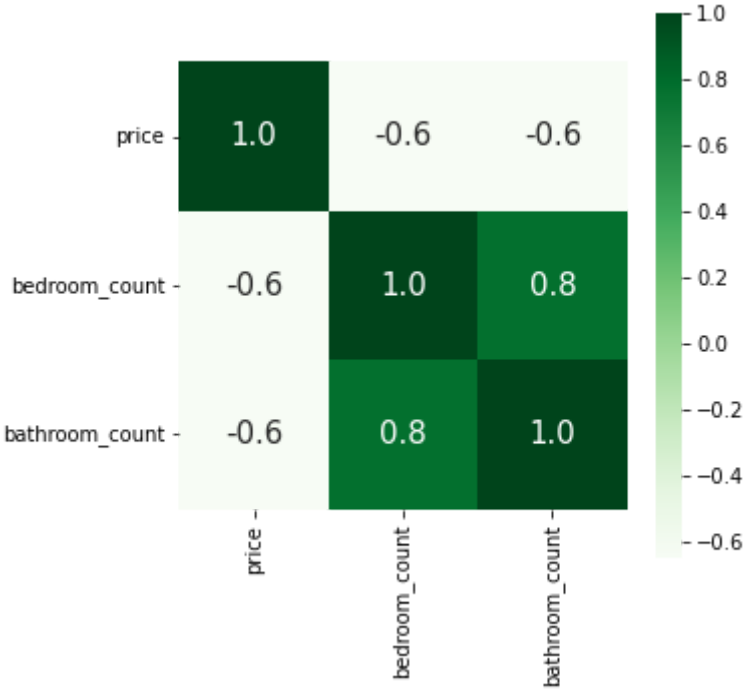
Out[46]:

city	price
Rosebery	390.000
Woodroffe	380.000
Moulden	370.000
Coconut Grove	339.000
Rapid Creek	299.000
Parap	289.000
Fannie Bay	285.000
Gray	279.000
Wagaman	277.000
Darwin City	160.333
Durack	4.000
Bees Creek	1.000
Virginia	1.000
Johnston	1.000
Larrakeyah	1.000
Berry Springs	1.000
Howard Springs	1.000
Girraween	1.000
Zuccoli	1.000

In [47]:
```
newdf = df[['price', 'bedroom_count', 'bathroom_count']].copy()
newdf = newdf.dropna()
```

In [48]:
```
corr = newdf.corr()
corr.shape
# Plotting the heatmap of correlation between features
plt.figure(figsize=(5,5))

sns.heatmap(corr, cbar=True, square= True, fmt='.1f', annot=True, annot_kws={'size
```

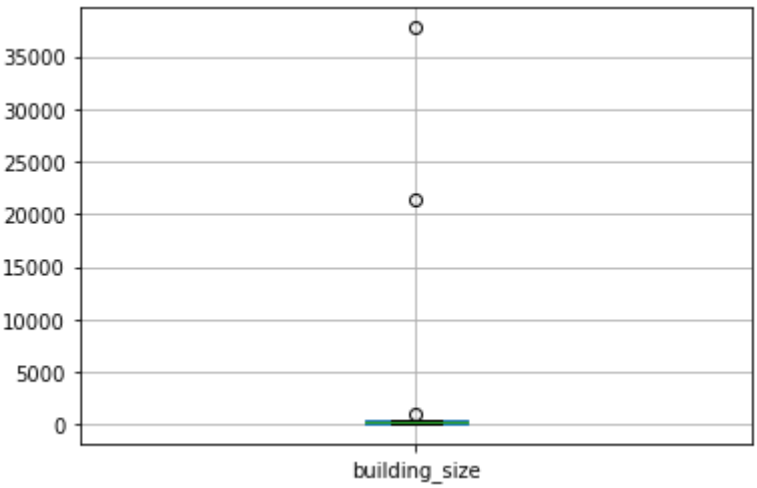
Out[48]:
```
<AxesSubplot:>
```



let's look at the size too

ok that's not right. our data set is corrupt

In [49]:
```
boxplot = df.boxplot(column=['building_size'])
```

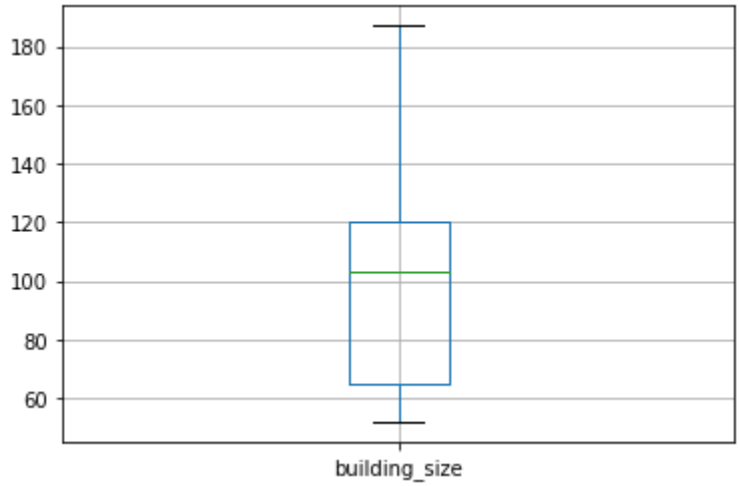


apparently there is a building size 250000, which is bigger than the entire United Kingdom.

In [50]:
```
print(df.size)
```

```
308
```

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



Done