eng22cs0130

October 8, 2024

[173]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

[174]:
```
data = pd.read_csv(r"C:\Users\DSU-CSE513-16\Downloads\Housing (2).csv")
```

[175]:
```
data
```

[175]:
```
        price  area  bedrooms  bathrooms  stories mainroad guestroom basement  \
0    13300000  7420         4          2        3      yes        no       no
1    12250000  8960         4          4        4      yes        no       no
2    12250000  9960         3          2        2      yes        no      yes
3    12215000  7500         4          2        2      yes        no      yes
4    11410000  7420         4          1        2      yes       yes      yes
..        …     …         …          …        …        …         …
561   1820000  3000         2          1        1      yes        no      yes
562   1767150  2400         3          1        1       no        no       no
563   1750000  3620         2          1        1      yes        no       no
564   1750000  2910         3          1        1       no        no       no
565   1750000  3850         3          1        2      yes        no       no

    hotwaterheating airconditioning  parking prefarea furnishingstatus
0                no             yes        2      yes        furnished
1                no             yes        3       no        furnished
2                no              no        2      yes   semi-furnished
3                no             yes        3      yes        furnished
4                no             yes        2       no        furnished
..              …               …      …        …                …
561              no              no        2       no      unfurnished
562              no              no        0       no   semi-furnished
563              no              no        0       no      unfurnished
564              no              no        0       no        furnished
565              no              no        0       no      unfurnished

[566 rows x 13 columns]
```

[176]:
```
print("\n Sample Data:")
print(data.head(10))
```

```
 Sample Data:
      price   area  bedrooms  bathrooms  stories mainroad guestroom basement  \
0  13300000   7420         4          2        3      yes        no       no
1  12250000   8960         4          4        4      yes        no       no
2  12250000   9960         3          2        2      yes        no      yes
3  12215000   7500         4          2        2      yes        no      yes
4  11410000   7420         4          1        2      yes       yes      yes
5  10850000   7500         3          3        1      yes        no      yes
6  10150000   8580         4          3        4      yes        no       no
7  10150000  16200         5          3        2      yes        no       no
8   9870000   8100         4          1        2      yes       yes      yes
9   9800000   5750         3          2        4      yes       yes       no

  hotwaterheating airconditioning  parking prefarea furnishingstatus
0              no             yes        2      yes        furnished
1              no             yes        3       no        furnished
2              no              no        2      yes   semi-furnished
3              no             yes        3      yes        furnished
4              no             yes        2       no        furnished
5              no             yes        2      yes   semi-furnished
6              no             yes        2      yes   semi-furnished
7              no              no        0       no      unfurnished
8              no             yes        2      yes        furnished
9              no             yes        1      yes      unfurnished
```

[177]:
```
print(data.tail(10))
```

```
       price  area  bedrooms  bathrooms  stories mainroad guestroom basement  \
556  2100000  3360         2          1        1      yes        no       no
557  1960000  3420         5          1        2       no        no       no
558  1890000  1700         3          1        2      yes        no       no
559  1890000  3649         2          1        1      yes        no       no
560  1855000  2990         2          1        1       no        no       no
561  1820000  3000         2          1        1      yes        no      yes
562  1767150  2400         3          1        1       no        no       no
563  1750000  3620         2          1        1      yes        no       no
564  1750000  2910         3          1        1       no        no       no
565  1750000  3850         3          1        2      yes        no       no

    hotwaterheating airconditioning  parking prefarea furnishingstatus
556              no              no        1       no      unfurnished
557              no              no        0       no      unfurnished
558              no              no        0       no      unfurnished
559              no              no        0       no      unfurnished
```

```
560            no            no       1       no      unfurnished
561            no            no       2       no      unfurnished
562            no            no       0       no   semi-furnished
563            no            no       0       no      unfurnished
564            no            no       0       no        furnished
565            no            no       0       no      unfurnished
```

[178]:

```python
print("basic Information:")
print(data.info())
```

```
basic Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 566 entries, 0 to 565
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price             566 non-null    int64
 1   area              566 non-null    int64
 2   bedrooms          566 non-null    int64
 3   bathrooms         566 non-null    int64
 4   stories           566 non-null    int64
 5   mainroad          566 non-null    object
 6   guestroom         566 non-null    object
 7   basement          566 non-null    object
 8   hotwaterheating   566 non-null    object
 9   airconditioning   566 non-null    object
 10  parking           566 non-null    int64
 11  prefarea          566 non-null    object
 12  furnishingstatus  566 non-null    object
dtypes: int64(6), object(7)
memory usage: 57.6+ KB
None
```

[179]:

```python
data.dtypes
```

[179]:

```
price                int64
area                 int64
bedrooms             int64
bathrooms            int64
stories              int64
mainroad            object
guestroom           object
basement            object
hotwaterheating     object
airconditioning     object
parking              int64
prefarea            object
furnishingstatus    object
```

```
dtype: object
```

[180]:
```
data.columns
```

[180]:
```
Index(['price', 'area', 'bedrooms', 'bathrooms', 'stories', 'mainroad',
       'guestroom', 'basement', 'hotwaterheating', 'airconditioning',
       'parking', 'prefarea', 'furnishingstatus'],
      dtype='object')
```

[181]:
```
data.shape
```

[181]:
```
(566, 13)
```

[182]:
```
print("\n Summary Statistics:")
print(data.describe())
```

```

 Summary Statistics:
              price          area    bedrooms   bathrooms     stories  \
count  5.660000e+02    566.000000  566.000000  566.000000  566.000000
mean   4.666197e+06   5076.773852    2.950530    1.275618    1.786219
std    1.906052e+06   2168.049072    0.746217    0.496008    0.861294
min    1.750000e+06   1650.000000    1.000000    1.000000    1.000000
25%    3.360000e+06   3514.000000    2.000000    1.000000    1.000000
50%    4.270000e+06   4500.000000    3.000000    1.000000    2.000000
75%    5.639375e+06   6357.500000    3.000000    2.000000    2.000000
max    1.330000e+07  16200.000000    6.000000    4.000000    4.000000

          parking
count  566.000000
mean     0.674912
std      0.856194
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      3.000000
```

[183]:
```
data.isnull()
```

[183]:
```
     price   area  bedrooms  bathrooms  stories  mainroad  guestroom  \
0    False  False     False      False    False     False      False
1    False  False     False      False    False     False      False
2    False  False     False      False    False     False      False
3    False  False     False      False    False     False      False
4    False  False     False      False    False     False      False
..     …      …       …         …        …         …          …
561  False  False     False      False    False     False      False
```

```
562  False  False     False      False    False     False      False
563  False  False     False      False    False     False      False
564  False  False     False      False    False     False      False
565  False  False     False      False    False     False      False

     basement  hotwaterheating  airconditioning  parking  prefarea  \
0       False            False            False    False     False
1       False            False            False    False     False
2       False            False            False    False     False
3       False            False            False    False     False
4       False            False            False    False     False
..        …               …              …      …       …
561     False            False            False    False     False
562     False            False            False    False     False
563     False            False            False    False     False
564     False            False            False    False     False
565     False            False            False    False     False

     furnishingstatus
0               False
1               False
2               False
3               False
4               False
..                 …
561             False
562             False
563             False
564             False
565             False

[566 rows x 13 columns]
```

[184]:

```
data.isnull().sum()
```

[184]:

```
price               0
area                0
bedrooms            0
bathrooms           0
stories             0
mainroad            0
guestroom           0
basement            0
hotwaterheating     0
airconditioning     0
parking             0
prefarea            0
```

```
furnishingstatus    0
dtype: int64
```

[185]:

```
data.dropna(inplace=True)
```

[186]:

```
data.count()
```

[186]:

```
price               566
area                566
bedrooms            566
bathrooms           566
stories             566
mainroad            566
guestroom           566
basement            566
hotwaterheating     566
airconditioning     566
parking             566
prefarea            566
furnishingstatus    566
dtype: int64
```

[187]:

```
duplicate_rows_df=data[data.duplicated()]
print("number of duplicate row:",duplicate_rows_df.shape)
```

```
number of duplicate row: (21, 13)
```

[188]:

```
data.count()
```

[188]:

```
price               566
area                566
bedrooms            566
bathrooms           566
stories             566
mainroad            566
guestroom           566
basement            566
hotwaterheating     566
airconditioning     566
parking             566
prefarea            566
furnishingstatus    566
dtype: int64
```

[189]:

```
data=data.drop_duplicates()
```

[190]:

```
data.count()
```

```
[190]: price               545
       area                545
       bedrooms            545
       bathrooms           545
       stories             545
       mainroad            545
       guestroom           545
       basement            545
       hotwaterheating     545
       airconditioning     545
       parking             545
       prefarea            545
       furnishingstatus    545
       dtype: int64
```

```
[191]: features=data.columns
       features
```

```
[191]: Index(['price', 'area', 'bedrooms', 'bathrooms', 'stories', 'mainroad',
              'guestroom', 'basement', 'hotwaterheating', 'airconditioning',
              'parking', 'prefarea', 'furnishingstatus'],
             dtype='object')
```

```
[192]: zero_val_cols=(data[features]==0).sum()
       zero_val_cols
```

```
[192]: price                 0
       area                  0
       bedrooms              0
       bathrooms             0
       stories               0
       mainroad              0
       guestroom             0
       basement              0
       hotwaterheating       0
       airconditioning       0
       parking             299
       prefarea              0
       furnishingstatus      0
       dtype: int64
```

```
[193]: data.isnull().sum()/len(data)*100
```

```
[193]: price               0.0
       area                0.0
       bedrooms            0.0
       bathrooms           0.0
       stories             0.0
```

```
mainroad            0.0
guestroom           0.0
basement            0.0
hotwaterheating     0.0
airconditioning     0.0
parking             0.0
prefarea            0.0
furnishingstatus    0.0
dtype: float64
```

[194]:

```
#data[['price','area']]=data[['price','area']].replace(0,np.NaN)
# before removing 0s we have to convert 0s to NaN
data.loc[:,['price','area']]=data[['price','area']].replace(0,np.NaN)
```

[195]:

```
#filling null values with median of that column
#data.area.fillna(data.area.median(),inplace=True)
data.loc['area']=data['area'].fillna(data.area.median())
```

```
C:\Users\DSU-CSE513-16\AppData\Local\Temp\ipykernel_6136\492706799.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data.loc['area']=data['area'].fillna(data.area.median())
```

[196]:

```
data
```

[196]:

```
           price    area  bedrooms  bathrooms  stories mainroad guestroom  \
0     13300000.0  7420.0       4.0        2.0      3.0      yes        no
1     12250000.0  8960.0       4.0        4.0      4.0      yes        no
2     12250000.0  9960.0       3.0        2.0      2.0      yes        no
3     12215000.0  7500.0       4.0        2.0      2.0      yes        no
4     11410000.0  7420.0       4.0        1.0      2.0      yes       yes
…            …       …         …          …        …        …
541    1767150.0  2400.0       3.0        1.0      1.0       no        no
542    1750000.0  3620.0       2.0        1.0      1.0      yes        no
543    1750000.0  2910.0       3.0        1.0      1.0       no        no
544    1750000.0  3850.0       3.0        1.0      2.0      yes        no
area         NaN     NaN       NaN        NaN      NaN      NaN       NaN

     basement hotwaterheating airconditioning  parking prefarea  \
0          no              no             yes      2.0      yes
1          no              no             yes      3.0       no
2         yes              no              no      2.0      yes
3         yes              no             yes      3.0      yes
4         yes              no             yes      2.0       no
…       …             …              …             …        …
```

```
541        no              no              no      0.0       no
542        no              no              no      0.0       no
543        no              no              no      0.0       no
544        no              no              no      0.0       no
area      NaN             NaN             NaN      NaN      NaN

     furnishingstatus
0           furnished
1           furnished
2      semi-furnished
3           furnished
4           furnished
…                …
541    semi-furnished
542       unfurnished
543         furnished
544       unfurnished
area              NaN

[546 rows x 13 columns]
```

[197]:

```
     #one-hot encoding
one_hot_encoded = pd.get_dummies(data,columns=['mainroad'],prefix=['mainroad'])
print("one-hot encoded data:")
print(one_hot_encoded)
```

```
one-hot encoded data:
           price    area  bedrooms  bathrooms  stories guestroom basement  \
0     13300000.0  7420.0       4.0        2.0      3.0        no       no
1     12250000.0  8960.0       4.0        4.0      4.0        no       no
2     12250000.0  9960.0       3.0        2.0      2.0        no      yes
3     12215000.0  7500.0       4.0        2.0      2.0        no      yes
4     11410000.0  7420.0       4.0        1.0      2.0       yes      yes
…          …     …       …        …      …       …       …
541    1767150.0  2400.0       3.0        1.0      1.0        no       no
542    1750000.0  3620.0       2.0        1.0      1.0        no       no
543    1750000.0  2910.0       3.0        1.0      1.0        no       no
544    1750000.0  3850.0       3.0        1.0      2.0        no       no
area         NaN     NaN       NaN        NaN      NaN       NaN      NaN

     hotwaterheating airconditioning  parking prefarea furnishingstatus  \
0                 no             yes      2.0      yes        furnished
1                 no             yes      3.0       no        furnished
2                 no              no      2.0      yes   semi-furnished
3                 no             yes      3.0      yes        furnished
4                 no             yes      2.0       no        furnished
…              …             …      …      …              …
541               no              no      0.0       no   semi-furnished
```

```
542                no              no      0.0       no      unfurnished
543                no              no      0.0       no        furnished
544                no              no      0.0       no      unfurnished
area              NaN             NaN      NaN      NaN              NaN

      mainroad_no  mainroad_yes
0           False          True
1           False          True
2           False          True
3           False          True
4           False          True
…           …           …
541          True         False
542         False          True
543          True         False
544         False          True
area        False         False

[546 rows x 14 columns]
```

[198]:

```python
from sklearn.preprocessing import LabelEncoder
```

[199]:

```python
label_encoder = LabelEncoder()
data['Guestroom_LabelEncoded']=label_encoder.fit_transform(data['guestroom'])
print("\n Label Encoded Data:")
print(data)
```

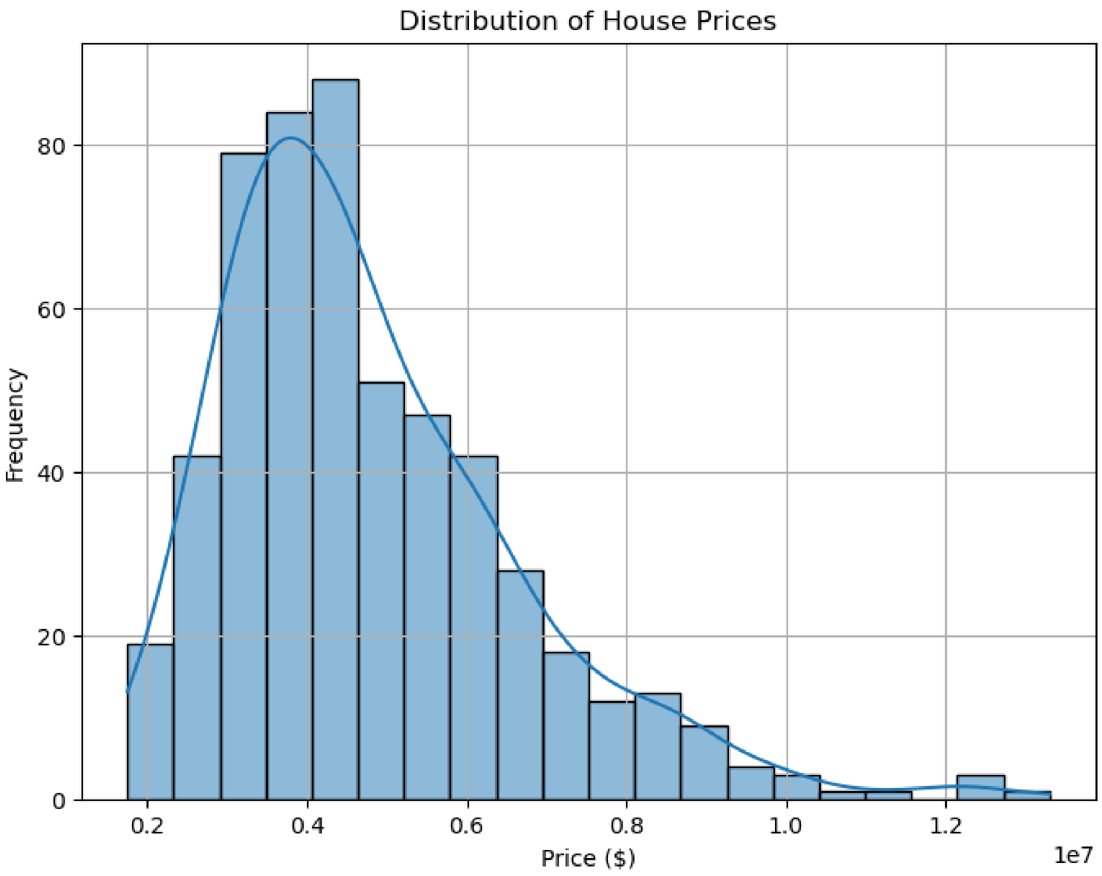
```
 Label Encoded Data:
           price    area  bedrooms  bathrooms  stories mainroad guestroom  \
0     13300000.0  7420.0       4.0        2.0      3.0      yes        no
1     12250000.0  8960.0       4.0        4.0      4.0      yes        no
2     12250000.0  9960.0       3.0        2.0      2.0      yes        no
3     12215000.0  7500.0       4.0        2.0      2.0      yes        no
4     11410000.0  7420.0       4.0        1.0      2.0      yes       yes
…          …     …       …        …      …      …      …
541    1767150.0  2400.0       3.0        1.0      1.0       no        no
542    1750000.0  3620.0       2.0        1.0      1.0      yes        no
543    1750000.0  2910.0       3.0        1.0      1.0       no        no
544    1750000.0  3850.0       3.0        1.0      2.0      yes        no
area         NaN     NaN       NaN        NaN      NaN      NaN       NaN

     basement hotwaterheating airconditioning  parking prefarea  \
0          no              no             yes      2.0      yes
1          no              no             yes      3.0       no
2         yes              no              no      2.0      yes
3         yes              no             yes      3.0      yes
4         yes              no             yes      2.0       no
```

```
…       …             …             …      …      …
541        no              no              no      0.0       no
542        no              no              no      0.0       no
543        no              no              no      0.0       no
544        no              no              no      0.0       no
area      NaN             NaN             NaN      NaN      NaN

     furnishingstatus  Guestroom_LabelEncoded
0           furnished                       0
1           furnished                       0
2      semi-furnished                       0
3           furnished                       0
4           furnished                       1
…                …                     …
541    semi-furnished                       0
542       unfurnished                       0
543         furnished                       0
544       unfurnished                       0
area              NaN                       2

[546 rows x 14 columns]
```

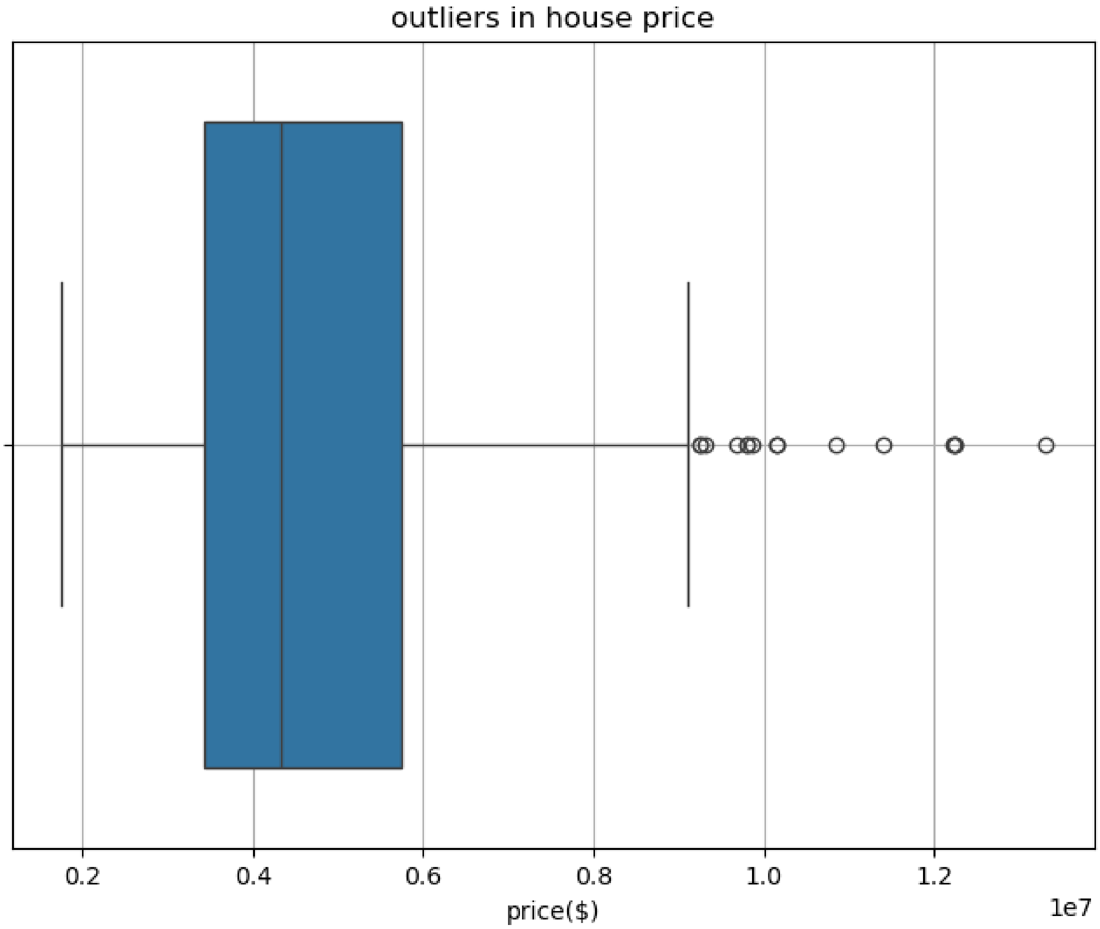
```
C:\Users\DSU-CSE513-16\AppData\Local\Temp\ipykernel_6136\2421631340.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data['Guestroom_LabelEncoded']=label_encoder.fit_transform(data['guestroom'])
```

[200]:

```python
#visualize the distribution of house prices
plt.figure(figsize=(8,6))
sns.histplot(data['price'],bins=20,kde=True)
plt.title("Distribution of House Prices")
plt.xlabel("Price ($)")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()
```



```
[201]: #visualize outliers in house price
       plt.figure(figsize=(8,6))
       sns.boxplot(data=data,x='price')
       plt.title("outliers in house price")
       plt.xlabel("price($)")
       plt.grid(True)
       plt.show()
```



```
[202]: data = data[data['price']<9000000]
```

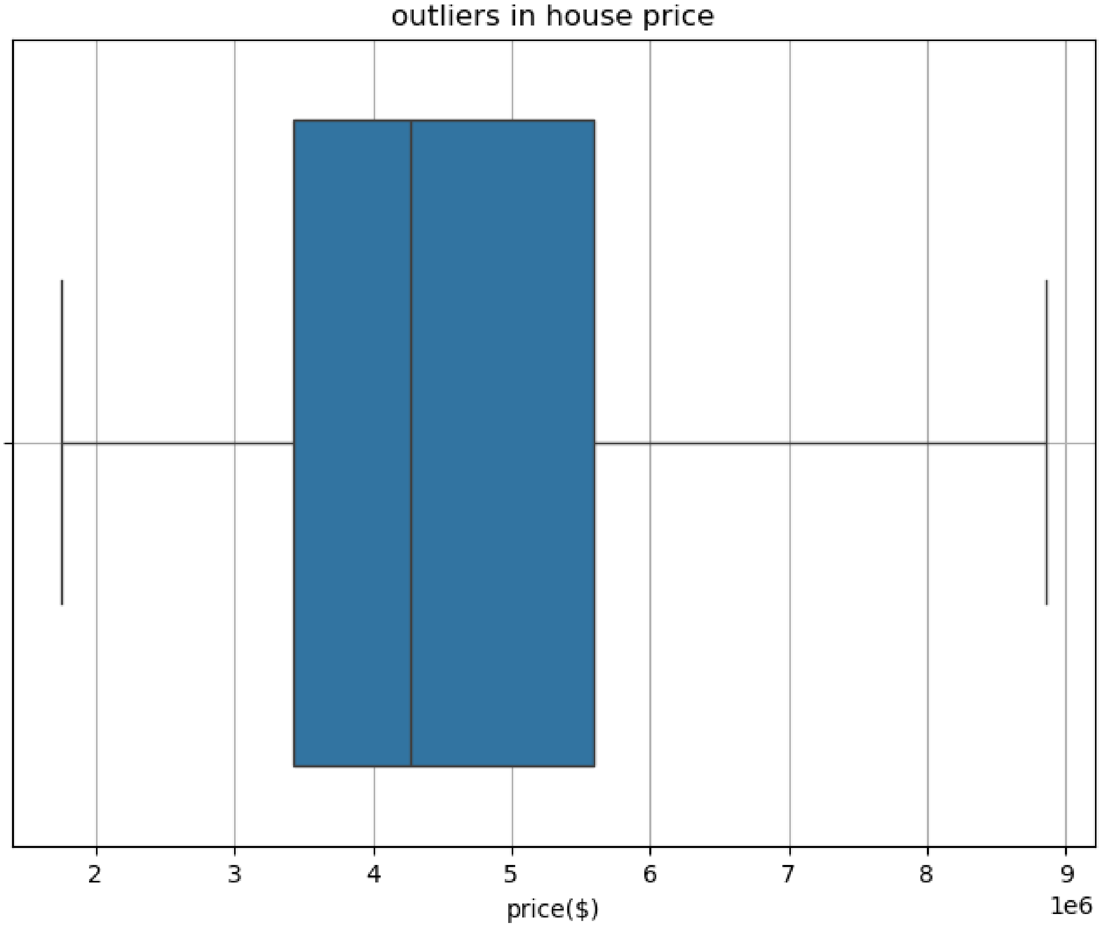
```
[203]: def find_boundaries(variable):
           q1 = data[variable].quantile(0.25)
           q3 = data[variable].quantile(0.75)
           iqr = q3 - q1
           lower_boundary = q1 - 1.5 * iqr
           upper_boundary = q3 + 1.5 * iqr
           return lower_boundary, upper_boundary

       lower_price, upper_price = find_boundaries('price')
       data.price = np.where(data.price > upper_price, upper_price, data.price)
       data.price = np.where(data.price < lower_price, lower_price, data.price)
```

```
C:\Users\DSU-CSE513-16\AppData\Local\Temp\ipykernel_6136\2133715101.py:10:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data.price = np.where(data.price > upper_price, upper_price, data.price)
C:\Users\DSU-CSE513-16\AppData\Local\Temp\ipykernel_6136\2133715101.py:11:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data.price = np.where(data.price < lower_price, lower_price, data.price)
```

[204]:

```
plt.figure(figsize=(8,6))
sns.boxplot(data=data,x='price')
plt.title("outliers in house price")
plt.xlabel("price($)")
plt.grid(True)
plt.show()
```



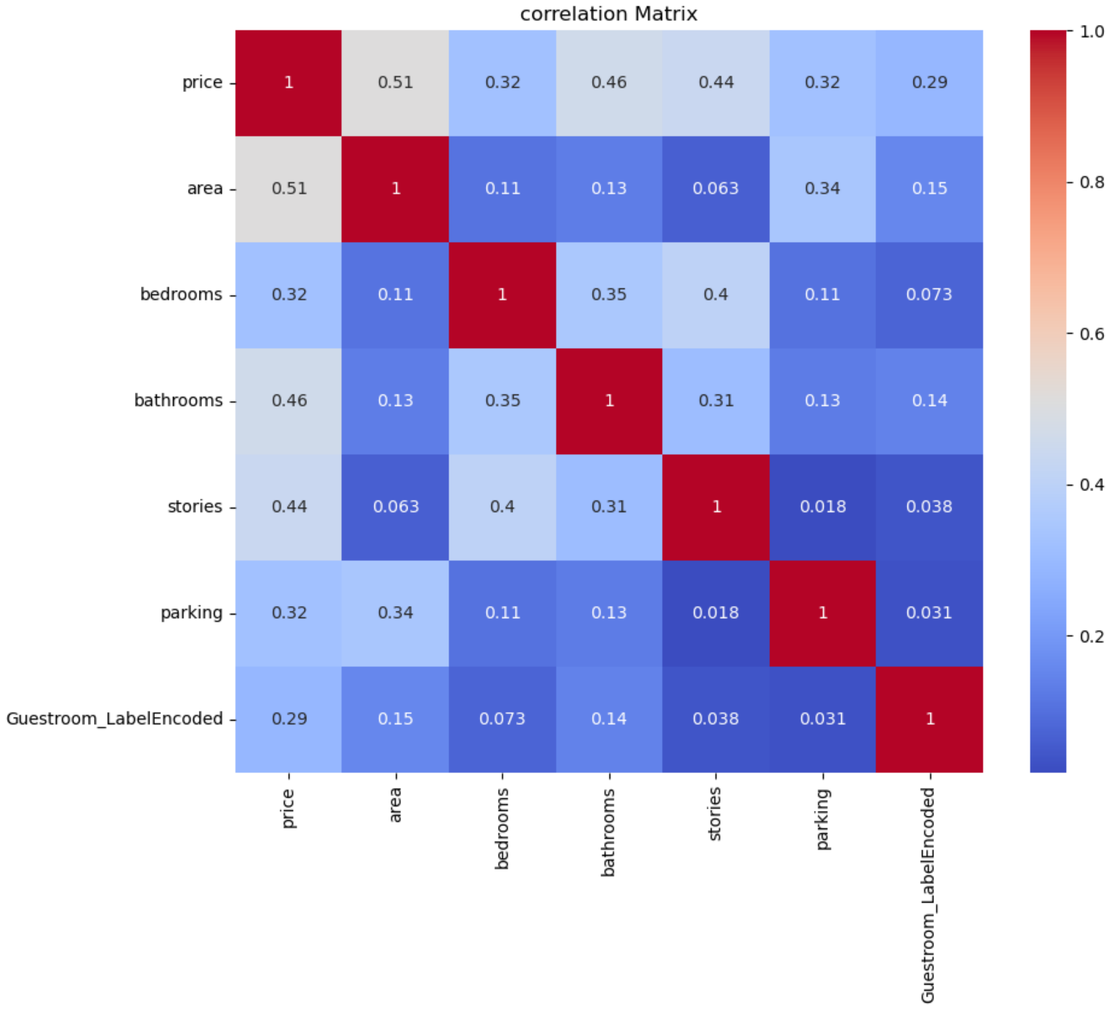
[205]:

```
#calculate correlation matrix
numeric_data =data.select_dtypes(include=[np.number])
correlation_matrix=numeric_data.corr()
print(correlation_matrix)
```

```
                        price      area  bedrooms  bathrooms   stories  \
price                1.000000  0.511642  0.323128   0.460610  0.436452
area                 0.511642  1.000000  0.109293   0.132166  0.063436
bedrooms             0.323128  0.109293  1.000000   0.349523  0.404938
bathrooms            0.460610  0.132166  0.349523   1.000000  0.308414
stories              0.436452  0.063436  0.404938   0.308414  1.000000
parking              0.323307  0.343992  0.105479   0.128327  0.018348
Guestroom_LabelEncoded  0.286845  0.153728  0.072505   0.141416  0.037742

                      parking  Guestroom_LabelEncoded
price                0.323307                0.286845
area                 0.343992                0.153728
bedrooms             0.105479                0.072505
bathrooms            0.128327                0.141416
stories              0.018348                0.037742
parking              1.000000                0.030774
Guestroom_LabelEncoded  0.030774                1.000000
```

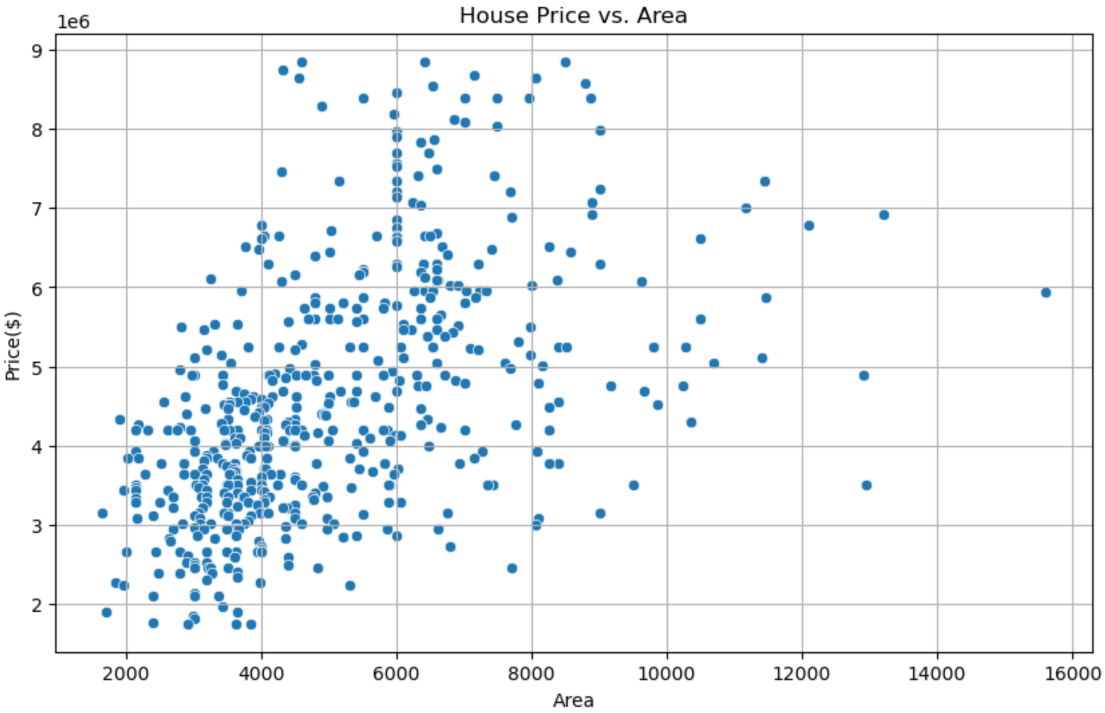
[207]:

```
#visualize correlation matrix
plt.figure(figsize=(10,8))
sns.heatmap(correlation_matrix,annot=True,cmap="coolwarm")
plt.title("correlation Matrix")
plt.show()
```



[209]:

```
plt.figure(figsize=(10,6))
sns.scatterplot(data=data,x='area',y='price')
plt.title("House Price vs. Area")
plt.xlabel("Area")
plt.ylabel("Price($)")
plt.grid(True)
plt.show()
```


House Price vs. Area
1e6
9
8
7
6
5
4
3
2
Price($)
2000
4000
6000
8000
10000
12000
14000
16000
Area

[ ]: