dmv4

November 3, 2024

1 Data Wrangling Problem Statement: Data Wrangling on Real Estate Market The goal is to perform data wrangling to gain insights into the factors influencing housing prices and prepare the dataset for further analysis or modeling.

[1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
```

[2]:

```
df=pd.read_csv(r"C:\Users\dell\Desktop\DMV and ML\DMV Datasets\Real-Estate␣
 ↪dataset.csv")
```

[3]:

```
df.head()
```

[3]:

```
      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

  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
```

[4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
```

```
 0   price             545 non-null    int64
 1   area              545 non-null    int64
 2   bedrooms          545 non-null    int64
 3   bathrooms         545 non-null    int64
 4   stories           545 non-null    int64
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
memory usage: 55.5+ KB
```

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

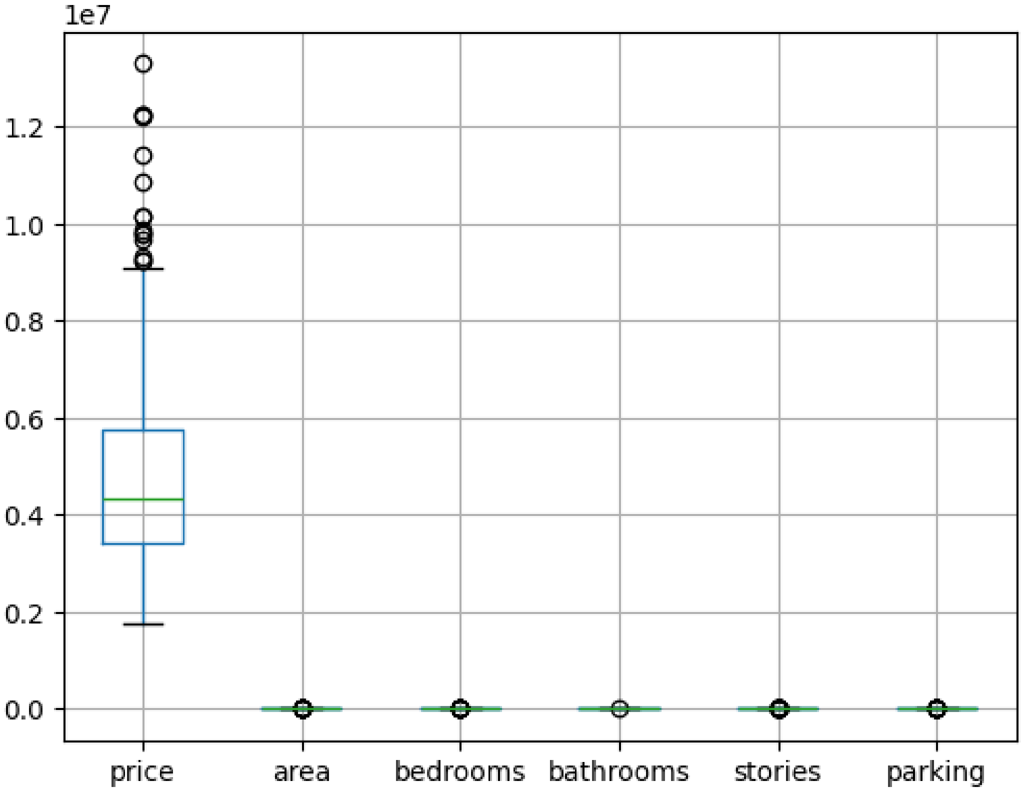
```
[5]:              price          area    bedrooms   bathrooms     stories  \
count  5.450000e+02    545.000000  545.000000  545.000000  545.000000
mean   4.766729e+06   5150.541284    2.965138    1.286239    1.805505
std    1.870440e+06   2170.141023    0.738064    0.502470    0.867492
min    1.750000e+06   1650.000000    1.000000    1.000000    1.000000
25%    3.430000e+06   3600.000000    2.000000    1.000000    1.000000
50%    4.340000e+06   4600.000000    3.000000    1.000000    2.000000
75%    5.740000e+06   6360.000000    3.000000    2.000000    2.000000
max    1.330000e+07  16200.000000    6.000000    4.000000    4.000000

          parking
count  545.000000
mean     0.693578
std      0.861586
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      3.000000
```

```
[6]: df.isna().sum()
```

```
[6]: 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
```

[7]:
```
df.boxplot()
plt.show()
```

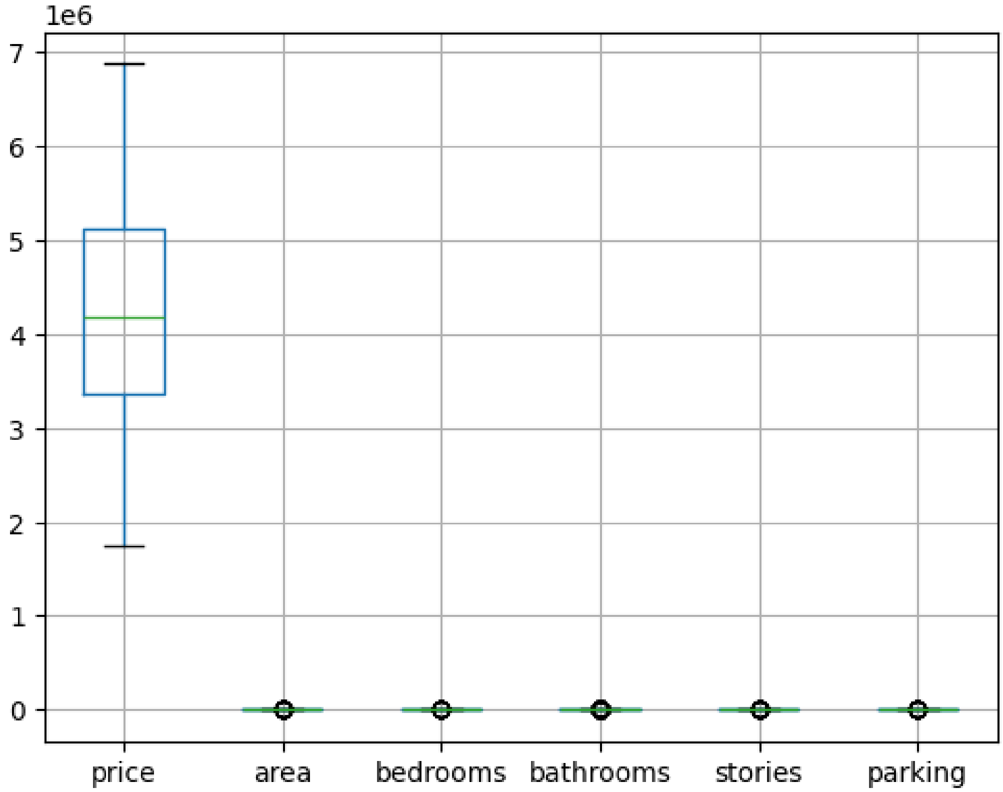


[8]:
```
Q1 = df["price"].quantile(0.25)
Q3 = df["price"].quantile(0.75)
iqr= Q3 - Q1
minm = Q1 - (1.5 * iqr)
maxm = Q3 + (1.5 * iqr)
df = df[(df["price"]>minm) & (df["price"]<maxm)]


# another method for same results
upper_limit = df['price'].quantile(0.9)
```

```
df = df[df['price'] < upper_limit]
```

[9]:

```
df.boxplot()
plt.show()
```



[10]:

```
df.dtypes
```

[10]:

```
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
```

[11]:

```
df.columns
```

[11]:

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

[12]:

```
le =LabelEncoder()
df['mainroad'] = le.fit_transform(df['mainroad'])
df['guestroom'] = le.fit_transform(df['guestroom'])
df['basement'] = le.fit_transform(df['basement'])
df['hotwaterheating'] = le.fit_transform(df['hotwaterheating'])
df['airconditioning'] = le.fit_transform(df['airconditioning'])
df['prefarea'] = le.fit_transform(df['prefarea'])
df['furnishingstatus'] = le.fit_transform(df['furnishingstatus'])
```

[13]:

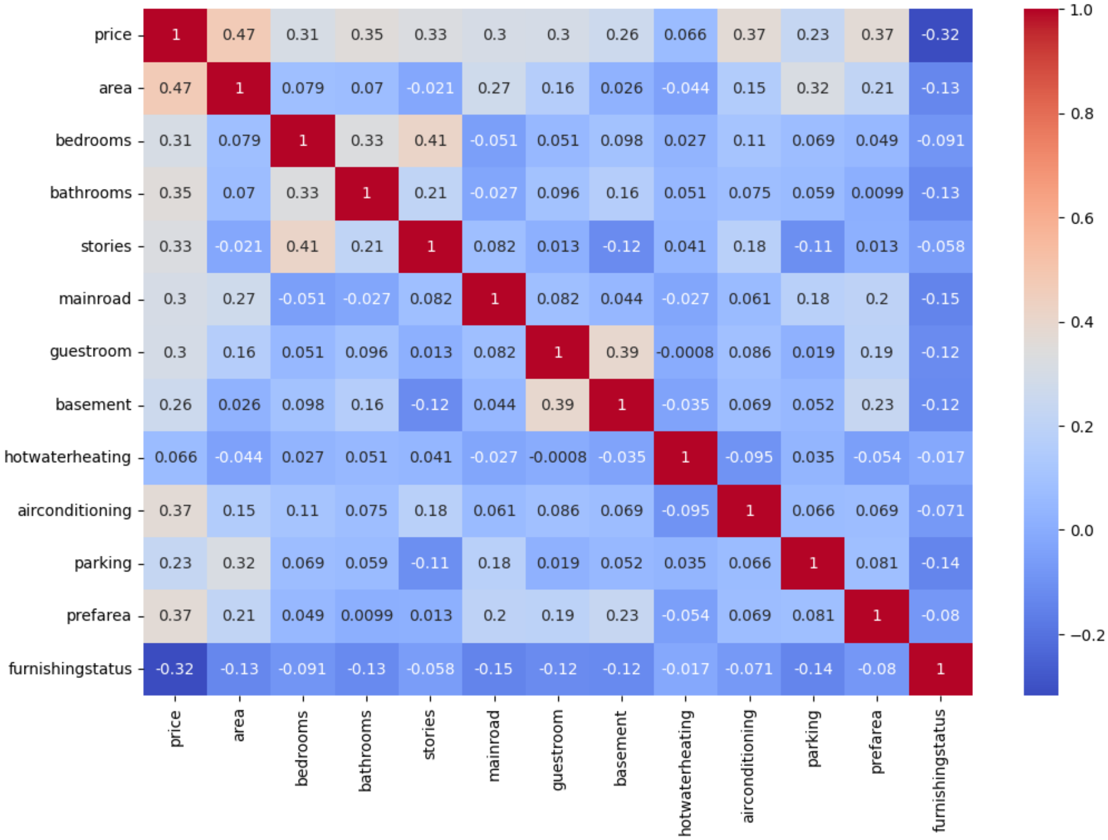
```
df.head()
```

[13]:

```
      price   area  bedrooms  bathrooms  stories  mainroad  guestroom  \
68  6860000   6000         3          1        1         1          0
69  6790000  12090         4          2        2         1          0
70  6790000   4000         3          2        2         1          0
71  6755000   6000         4          2        4         1          0
72  6720000   5020         3          1        4         1          0

    basement  hotwaterheating  airconditioning  parking  prefarea  \
68         0                0                1        1         0
69         0                0                0        2         1
70         1                0                1        0         1
71         0                0                1        0         0
72         0                0                1        0         1

    furnishingstatus
68                 0
69                 0
70                 1
71                 2
72                 2
```

[14]:

```
plt.figure(figsize=(12, 8))
sns.heatmap(data = df.corr(), cmap = "coolwarm", annot=True)
plt.show()
```



	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
price	1	0.47	0.31	0.35	0.33	0.3	0.3	0.26	0.066	0.37	0.23	0.37	-0.32
area	0.47	1	0.079	0.07	-0.021	0.27	0.16	0.026	-0.044	0.15	0.32	0.21	-0.13
bedrooms	0.31	0.079	1	0.33	0.41	-0.051	0.051	0.098	0.027	0.11	0.069	0.049	-0.091
bathrooms	0.35	0.07	0.33	1	0.21	-0.027	0.096	0.16	0.051	0.075	0.059	0.0099	-0.13
stories	0.33	-0.021	0.41	0.21	1	0.082	0.013	-0.12	0.041	0.18	-0.11	0.013	-0.058
mainroad	0.3	0.27	-0.051	-0.027	0.082	1	0.082	0.044	-0.027	0.061	0.18	0.2	-0.15
guestroom	0.3	0.16	0.051	0.096	0.013	0.082	1	0.39	-0.0008	0.086	0.019	0.19	-0.12
basement	0.26	0.026	0.098	0.16	-0.12	0.044	0.39	1	-0.035	0.069	0.052	0.23	-0.12
hotwaterheating	0.066	-0.044	0.027	0.051	0.041	-0.027	-0.0008	-0.035	1	-0.095	0.035	-0.054	-0.017
airconditioning	0.37	0.15	0.11	0.075	0.18	0.061	0.086	0.069	-0.095	1	0.066	0.069	-0.071
parking	0.23	0.32	0.069	0.059	-0.11	0.18	0.019	0.052	0.035	0.066	1	0.081	-0.14
prefarea	0.37	0.21	0.049	0.0099	0.013	0.2	0.19	0.23	-0.054	0.069	0.081	1	-0.08
furnishingstatus	-0.32	-0.13	-0.091	-0.13	-0.058	-0.15	-0.12	-0.12	-0.017	-0.071	-0.14	-0.08	1

[ ]: