In [20]:

```
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt
#df = pd.read_csv("C:\Code\D214\move.csv")
df = pd.read_csv("C:\\Code\\D214\\rent_apartment_moscow.csv")
```

In [2]:

```
#viewing size of dataset
print('Length of dataset: ', len(df.index))
print('Number of columns: ', len(df.columns))
```

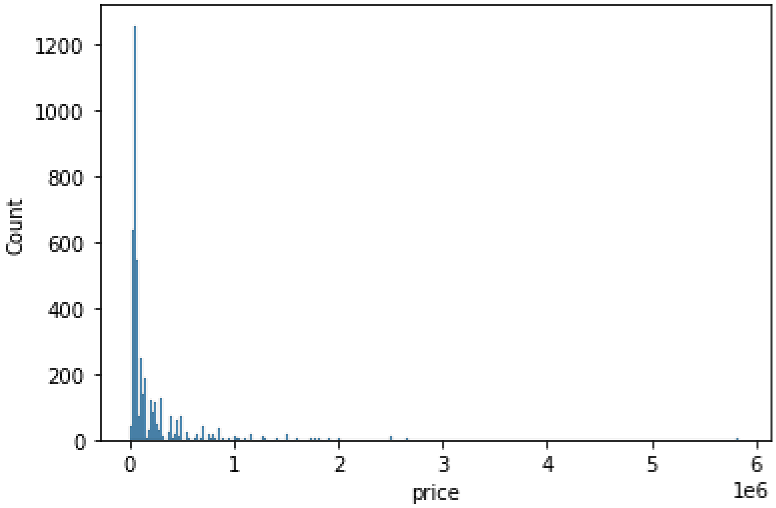
```
Length of dataset:  4241
Number of columns:  14
```

In [3]:

```
#viewing price distribution
sns.histplot(df['price'])
```

Out[3]:

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



In [4]:

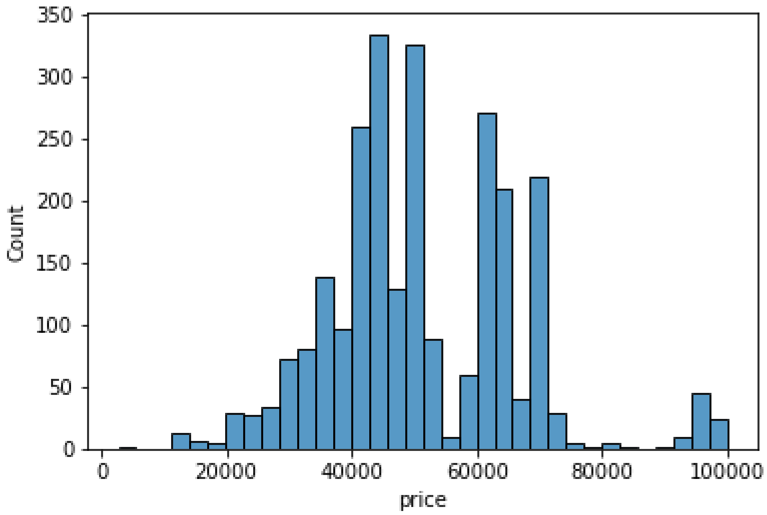
```
#remove price outliers
df = df[df['price'] < 100000]
```

In [5]:

```
#viewing price distribution again
sns.histplot(df['price'])
```

Out[5]:

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



In [6]:

```
#creating variable for sale price
price = df['price']
price.describe()
```

Out[6]:

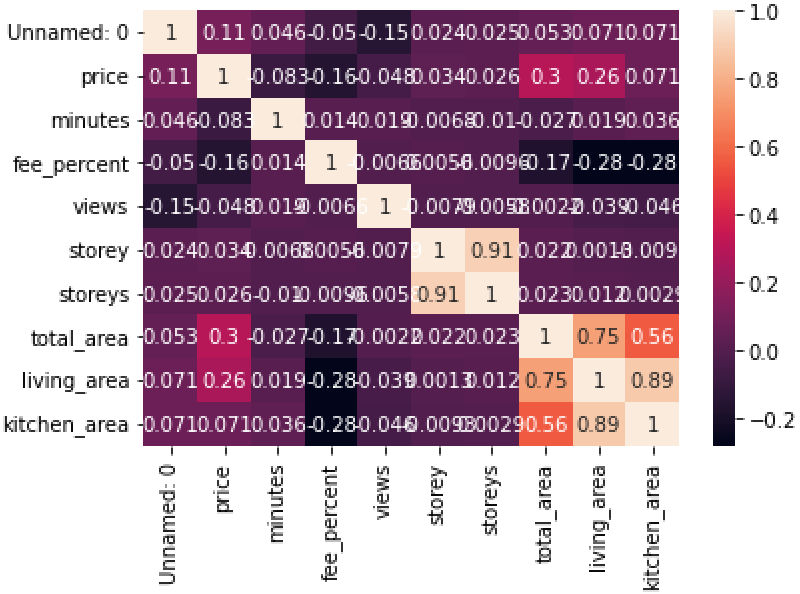
```
count     2569.000000
mean     50987.761386
std      15154.275901
min       2700.000000
25%      40000.000000
50%      50000.000000
75%      60000.000000
max      99999.000000
Name: price, dtype: float64
```

In [7]:

```
#viewing correlation
cor = df.corr()
sns.heatmap(cor, annot = True)
```

Out[7]:

```
<AxesSubplot:>
```



In [8]:

```
x = df['total_area']
y = df['price']
```

In [9]:

```
x = np.array(x)
```

```
y = np.array(y)
```

In [11]:

```
x = x.reshape(-1,1)
y = y.reshape(-1,1)
```

In [10]:

```
print(x)
```

```
[50 38 55 ... 50 64 64]
```

In [12]:

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2)
```

In [13]:

```
model = DecisionTreeRegressor()
model = model.fit(x_train,y_train)
```

In [14]:

```
predict = model.predict(x_test)
```

In [15]:

```
r2_score(y_test, predict)
```

Out[15]:

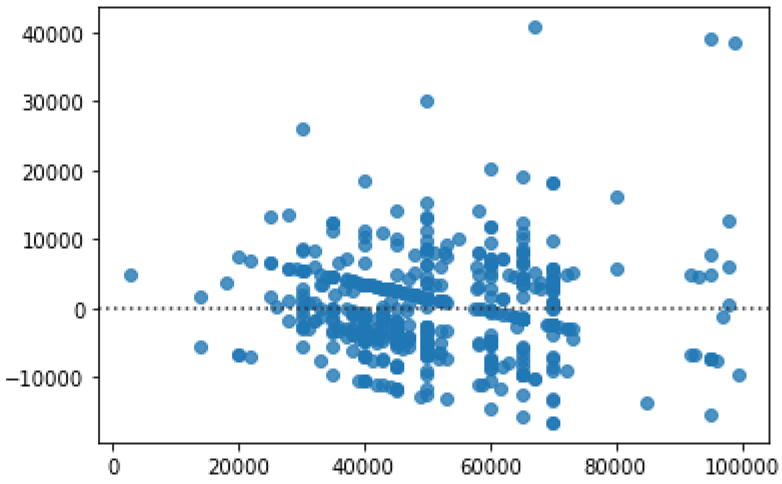
```
0.1439700610069341
```

In [27]:

```
sns.residplot(x = y_test, y = predict)
```

Out[27]:

```
<AxesSubplot:>
```



In [ ]:

```
```