In [1]:
```
import pandas as pd
import numpy  as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
rent_df = pd.read_csv('https://raw.githubusercontent.com/DSNote/fas
rent_df
```

Out[2]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City
0	2022-05-18	2.0	10000	1100.0	Ground out of 2	Super Area	Bandel	Kolkata
1	2022-05-13	2.0	20000	800.0	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata
2	2022-05-16	2.0	17000	1000.0	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata
3	2022-07-04	NaN	10000	800.0	1 out of 2	Super Area	Dumdum Park	Kolkata
4	2022-05-09	2.0	7500	850.0	1 out of 2	Carpet Area	South Dum Dum	Kolkata
...	...	...	...	...	...	...	...	...
4741	2022-05-18	2.0	15000	1000.0	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad
4742	2022-05-15	3.0	29000	2000.0	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad
4743	2022-07-10	3.0	35000	1750.0	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad
4744	2022-07-06	3.0	45000	1500.0	23 out of 34	Carpet Area	Gachibowli	Hyderabad
4745	2022-05-04	2.0	15000	1000.0	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad

4746 rows × 12 columns

In [3]:
```
rent_df.tail()
```

Out[3]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	
4741	2022-05-18	2.0	15000	1000.0	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	
4742	2022-05-15	3.0	29000	2000.0	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	
4743	2022-07-10	3.0	35000	1750.0	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	
4744	2022-07-06	3.0	45000	1500.0	23 out of 34	Carpet Area	Gachibowli	Hyderabad	
4745	2022-05-04	2.0	15000	1000.0	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	U

In [4]:
```
rent_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Posted On          4746 non-null   object 
 1   BHK                4743 non-null   float64
 2   Rent               4746 non-null   int64  
 3   Size               4741 non-null   float64
 4   Floor              4746 non-null   object 
 5   Area Type          4746 non-null   object 
 6   Area Locality      4746 non-null   object 
 7   City               4746 non-null   object 
 8   Furnishing Status  4746 non-null   object 
 9   Tenant Preferred   4746 non-null   object 
 10  Bathroom           4746 non-null   int64  
 11  Point of Contact   4746 non-null   object 
dtypes: float64(2), int64(2), object(8)
memory usage: 445.1+ KB
```

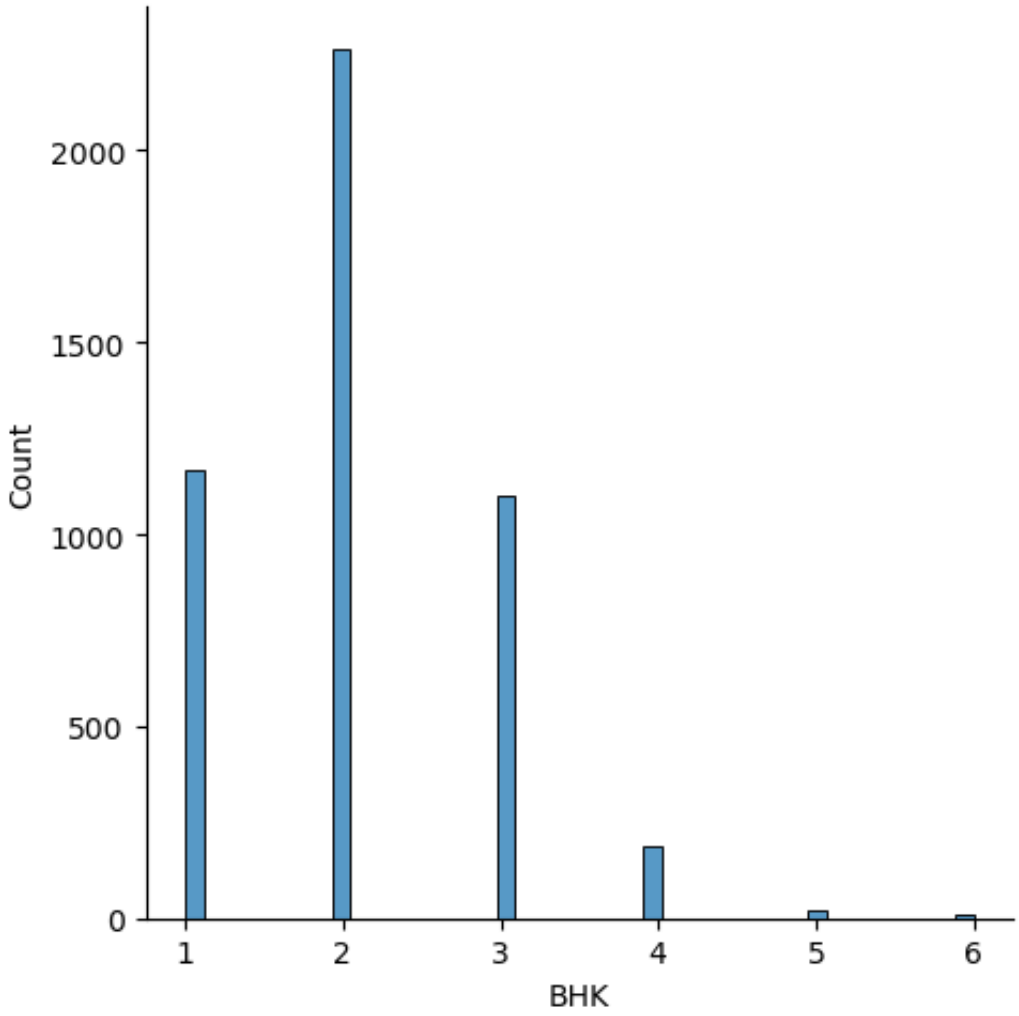
In [5]:
```
round(rent_df.describe(),2)
```

Out[5]:

	BHK	Rent	Size	Bathroom
count	4743.00	4746.00	4741.00	4746.00
mean	2.08	34993.45	967.48	1.97
std	0.83	78106.41	634.53	0.88
min	1.00	1200.00	10.00	1.00
25%	2.00	10000.00	550.00	1.00
50%	2.00	16000.00	850.00	2.00
75%	3.00	33000.00	1200.00	2.00
max	6.00	3500000.00	8000.00	10.00

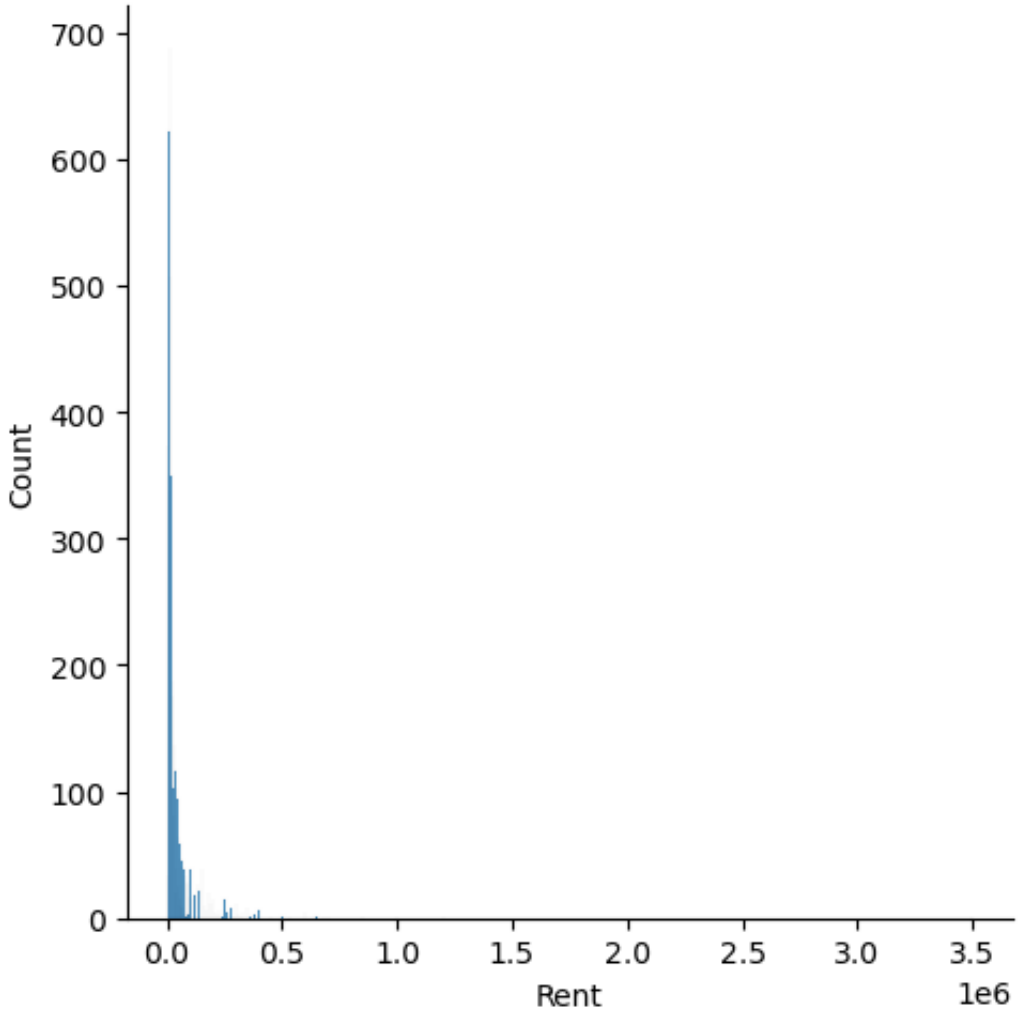
In [6]:
```
sns.displot(rent_df['BHK'])
```

Out[6]:
```
<seaborn.axisgrid.FacetGrid at 0x1256f27b0>
```



In [7]:
```
sns.displot(rent_df['Rent'])
```

Out[7]:
```
<seaborn.axisgrid.FacetGrid at 0x12599d810>
```



In [8]:
```
rent_df['Rent'].sort_values() #오름차순 정렬
```

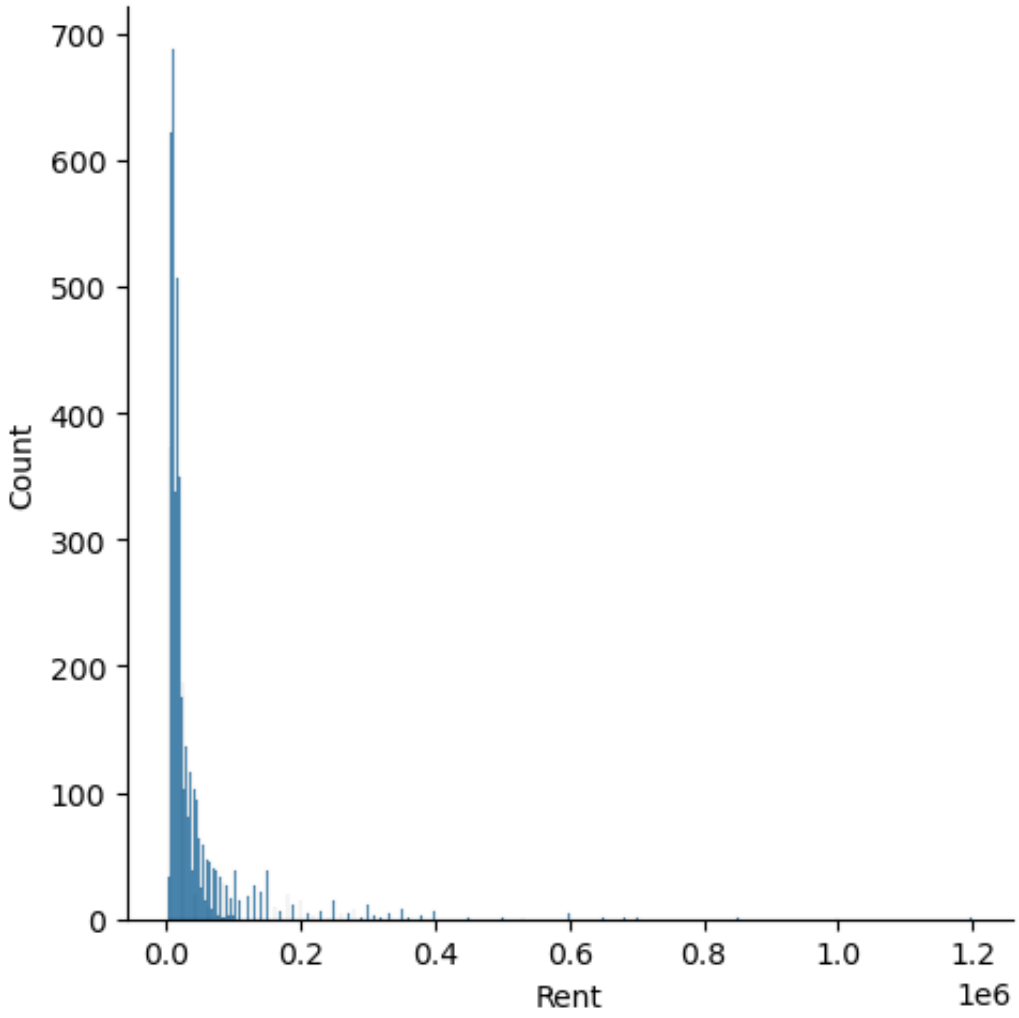
Out[8]:
```
4076       1200
285        1500
471        1800
2475       2000
146        2200
          ...
1459     700000
1329     850000
827     1000000
1001    1200000
1837    3500000
Name: Rent, Length: 4746, dtype: int64
```

In [9]:
```
rent_df.drop(1837)['Rent'].sort_values()
```

Out[9]:
```
4076       1200
285        1500
471        1800
2475       2000
146        2200
          ...
1484     680000
1459     700000
1329     850000
827     1000000
1001    1200000
Name: Rent, Length: 4745, dtype: int64
```

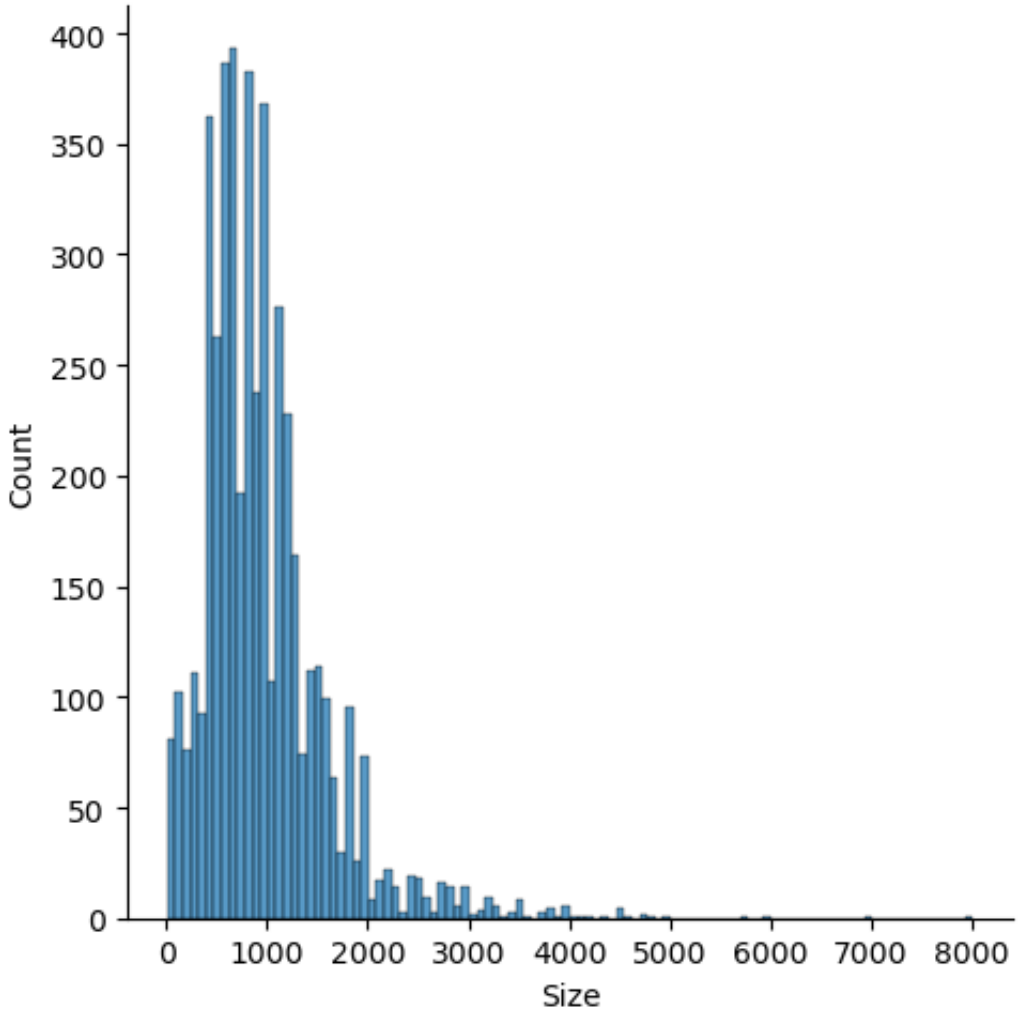
In [10]:
```
sns.displot(rent_df.drop(1837)['Rent'])
```

Out[10]:
```
<seaborn.axisgrid.FacetGrid at 0x126427250>
```



In [11]:
```
sns.displot(rent_df['Size'])
```

Out[11]:
```
<seaborn.axisgrid.FacetGrid at 0x1264efc50>
```

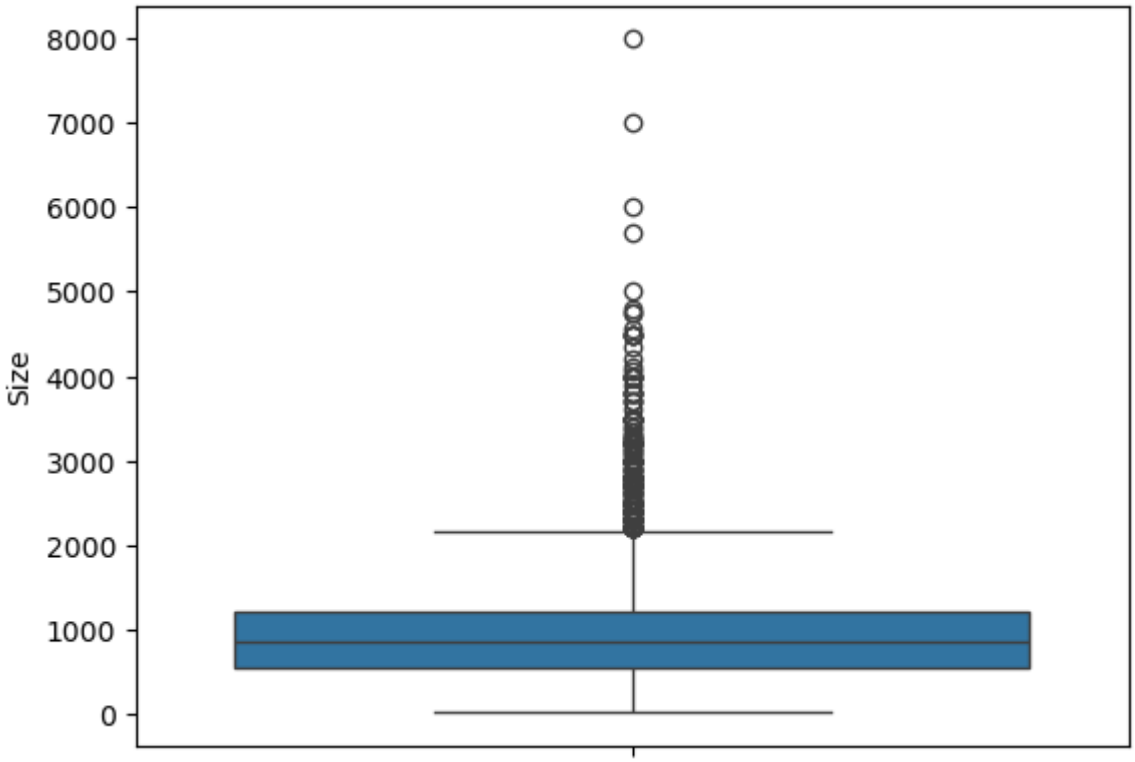


In [ ]:

In [12]:
```
#(Q3 - Q1) = IQR
#IQR*1.5 이상 떨어진 선에서 가장 가까운 값부터 아웃라이어
```

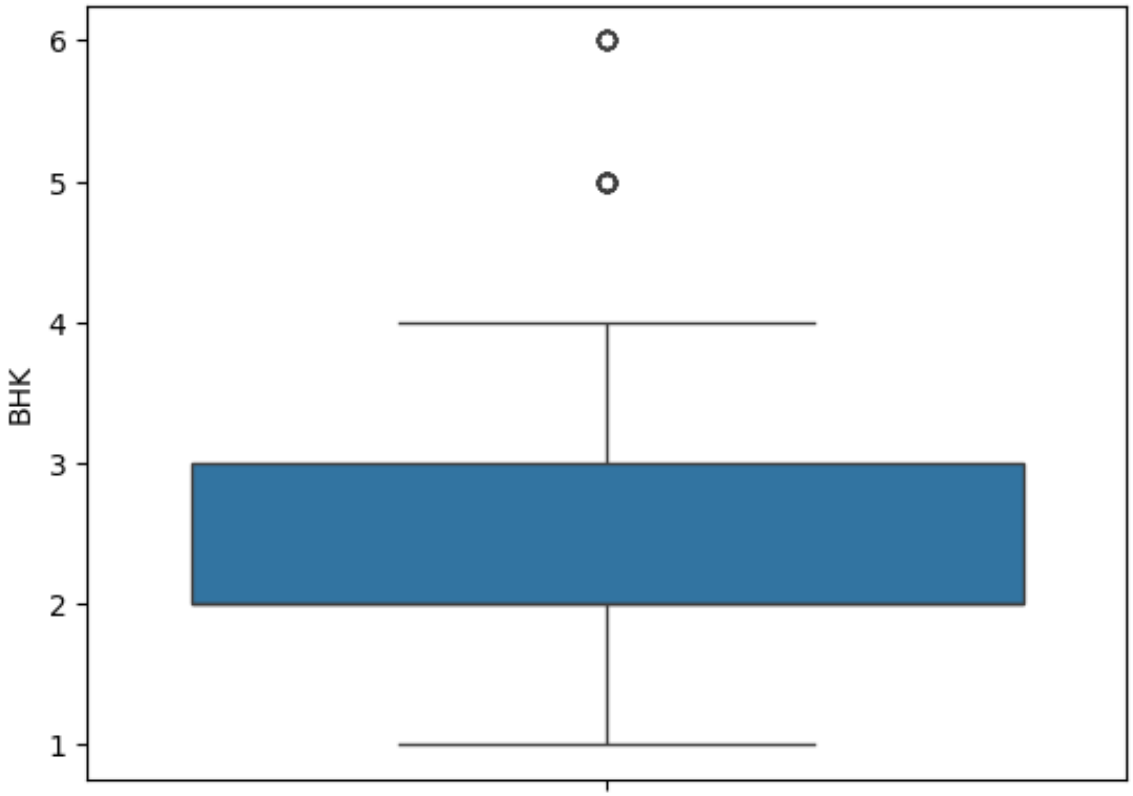
In [13]:
```
sns.boxplot(y=rent_df['Size'])
```

Out[13]:
```
<Axes: ylabel='Size'>
```



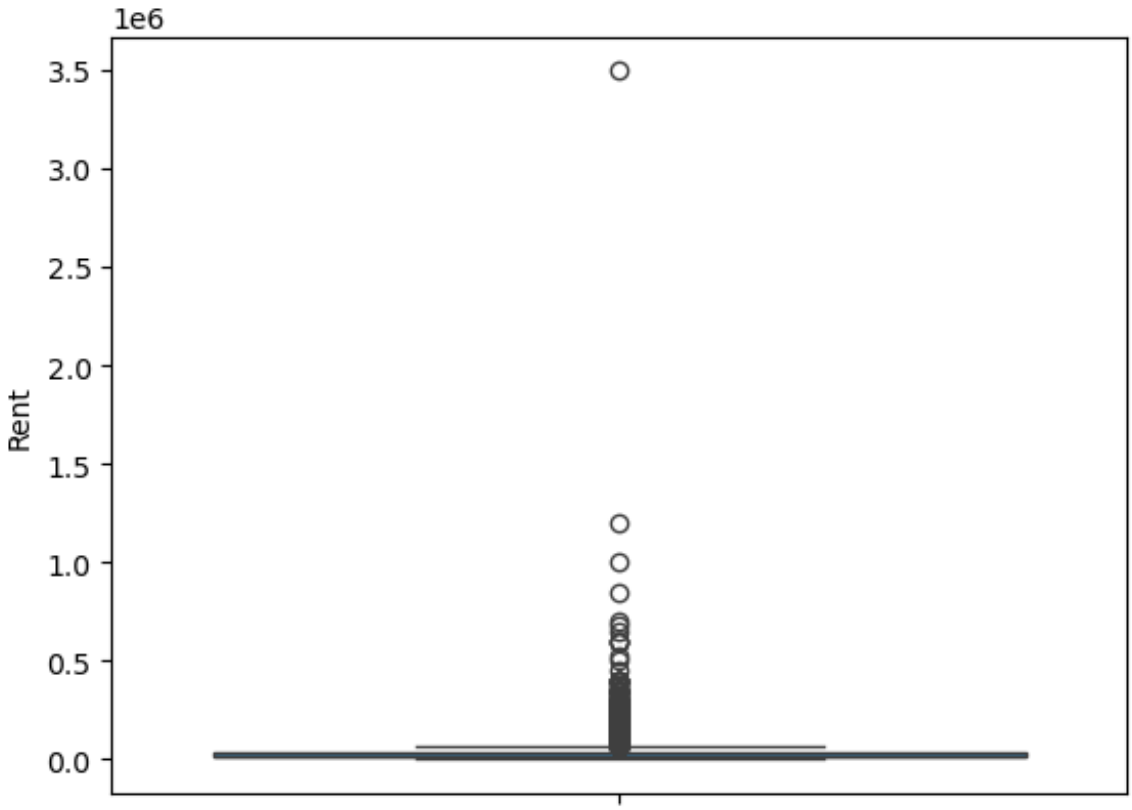
In [14]:
```
sns.boxplot(y=rent_df['BHK'])
```

Out[14]:
```
<Axes: ylabel='BHK'>
```



In [15]:
```
sns.boxplot(y=rent_df['Rent'])
```

Out[15]:
```
<Axes: ylabel='Rent'>
```



In [16]:
```
#결측치 처리
rent_df.isna().sum()
```

Out[16]:
```
Posted On            0
BHK                  3
Rent                 0
Size                 5
Floor                0
Area Type            0
Area Locality        0
City                 0
Furnishing Status    0
Tenant Preferred     0
Bathroom             0
Point of Contact     0
dtype: int64
```

In [17]:
```
rent_df.isna().mean() #결측치비율
```

Out[17]:
```
Posted On            0.000000
BHK                  0.000632
Rent                 0.000000
Size                 0.001054
Floor                0.000000
Area Type            0.000000
Area Locality        0.000000
City                 0.000000
Furnishing Status    0.000000
Tenant Preferred     0.000000
Bathroom             0.000000
Point of Contact     0.000000
dtype: float64
```

In [18]:
```
rent_df.dropna(subset=['Size'])
```

Out[18]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City
0	2022-05-18	2.0	10000	1100.0	Ground out of 2	Super Area	Bandel	Kolkata
1	2022-05-13	2.0	20000	800.0	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata
2	2022-05-16	2.0	17000	1000.0	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata
3	2022-07-04	NaN	10000	800.0	1 out of 2	Super Area	Dumdum Park	Kolkata
4	2022-05-09	2.0	7500	850.0	1 out of 2	Carpet Area	South Dum Dum	Kolkata
...	...	...	...	...	...	...	...	...
4741	2022-05-18	2.0	15000	1000.0	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad
4742	2022-05-15	3.0	29000	2000.0	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad
4743	2022-07-10	3.0	35000	1750.0	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad
4744	2022-07-06	3.0	45000	1500.0	23 out of 34	Carpet Area	Gachibowli	Hyderabad
4745	2022-05-04	2.0	15000	1000.0	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad

4741 rows × 12 columns

In [19]:
```
rent_df.dropna(axis = 1)
```

Out[19]:

	Posted On	Rent	Floor	Area Type	Area Locality	City	Furnishing Status
0	2022-05-18	10000	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished
1	2022-05-13	20000	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished
2	2022-05-16	17000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished
3	2022-07-04	10000	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished
4	2022-05-09	7500	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished
...	...	...	...	...	...	...	...
4741	2022-05-18	15000	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	Semi-Furnished
4742	2022-05-15	29000	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	Semi-Furnished
4743	2022-07-10	35000	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	Semi-Furnished
4744	2022-07-06	45000	23 out of 34	Carpet Area	Gachibowli	Hyderabad	Semi-Furnished
4745	2022-05-04	15000	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	Unfurnished

4746 rows × 10 columns

In [20]:
```
rent_df.drop(['BHK','Size'],axis=1)
```

Out[20]:

	Posted On	Rent	Floor	Area Type	Area Locality	City	Furnishing Status
0	2022-05-18	10000	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished
1	2022-05-13	20000	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished
2	2022-05-16	17000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished
3	2022-07-04	10000	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished
4	2022-05-09	7500	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished
...	...	...	...	...	...	...	...
4741	2022-05-18	15000	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	Semi-Furnished
4742	2022-05-15	29000	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	Semi-Furnished
4743	2022-07-10	35000	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	Semi-Furnished
4744	2022-07-06	45000	23 out of 34	Carpet Area	Gachibowli	Hyderabad	Semi-Furnished
4745	2022-05-04	15000	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	Unfurnished

4746 rows × 10 columns

In [21]:
```
na_index = rent_df[rent_df.isna().any(axis=1)].index
na_index
```

Out[21]:
```
Index([3, 53, 89, 425, 430, 4703, 4731, 4732], dtype='int64')
```

In [22]:
```
ㄱ둣
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[22], line 1
----> 1 ㄱ둣

NameError: name 'ㄱ둣' is not defined
```

In [23]:
```
rent_df.loc[na_index]
```

Out[23]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	Cit
3	2022-07-04	NaN	10000	800.0	1 out of 2	Super Area	Dumdum Park	Kolka
53	2022-04-23	NaN	15000	1000.0	Ground out of 2	Super Area	Bansdroni	Kolka
89	2022-05-31	NaN	8500	550.0	2 out of 3	Carpet Area	Kasba -East	Kolka
425	2022-05-22	2.0	9000	NaN	2 out of 3	Super Area	Airport Area Behala	Kolka
430	2022-05-08	2.0	8500	NaN	Ground out of 1	Carpet Area	Nayabad	Kolka
4703	2022-07-06	2.0	12000	NaN	4 out of 4	Super Area	Anandbagh, Secunderabad, Moula Ali Road	Hyderaba
4731	2022-06-24	2.0	13000	NaN	2 out of 2	Super Area	Manikonda, Outer Ring Road	Hyderaba
4732	2022-07-08	2.0	7000	NaN	Ground out of 2	Super Area	Vinayaka Nagar	Hyderaba

In [24]:
```
rent_df.fillna(rent_df[['BHK','Size']].median()).loc[na_index]
```

Out[24]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	Cit
3	2022-07-04	2.0	10000	800.0	1 out of 2	Super Area	Dumdum Park	Kolka
53	2022-04-23	2.0	15000	1000.0	Ground out of 2	Super Area	Bansdroni	Kolka
89	2022-05-31	2.0	8500	550.0	2 out of 3	Carpet Area	Kasba -East	Kolka
425	2022-05-22	2.0	9000	850.0	2 out of 3	Super Area	Airport Area Behala	Kolka
430	2022-05-08	2.0	8500	850.0	Ground out of 1	Carpet Area	Nayabad	Kolka
4703	2022-07-06	2.0	12000	850.0	4 out of 4	Super Area	Anandbagh, Secunderabad, Moula Ali Road	Hyderaba
4731	2022-06-24	2.0	13000	850.0	2 out of 2	Super Area	Manikonda, Outer Ring Road	Hyderaba
4732	2022-07-08	2.0	7000	850.0	Ground out of 2	Super Area	Vinayaka Nagar	Hyderaba

In [25]:
```
rent_df['BHK'] = rent_df['BHK'].fillna(rent_df['BHK'].median())
```

In [26]:
```
rent_df = rent_df.fillna(rent_df[['BHK','Size']].median())
```

In [27]:
```
rent_df.isna().mean()
```

Out[27]:
```
Posted On           0.0
BHK                 0.0
Rent                0.0
Size                0.0
Floor               0.0
Area Type           0.0
Area Locality       0.0
City                0.0
Furnishing Status   0.0
Tenant Preferred    0.0
Bathroom            0.0
Point of Contact    0.0
dtype: float64
```

In [28]:
```
rent_df.tail()
```

Out[28]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	
4741	2022-05-18	2.0	15000	1000.0	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	
4742	2022-05-15	3.0	29000	2000.0	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	
4743	2022-07-10	3.0	35000	1750.0	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	
4744	2022-07-06	3.0	45000	1500.0	23 out of 34	Carpet Area	Gachibowli	Hyderabad	
4745	2022-05-04	2.0	15000	1000.0	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	U

In [29]:
```
rent_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Posted On          4746 non-null   object 
 1   BHK                4746 non-null   float64
 2   Rent               4746 non-null   int64  
 3   Size               4746 non-null   float64
 4   Floor              4746 non-null   object 
 5   Area Type          4746 non-null   object 
 6   Area Locality      4746 non-null   object 
 7   City               4746 non-null   object 
 8   Furnishing Status  4746 non-null   object 
 9   Tenant Preferred   4746 non-null   object 
 10  Bathroom           4746 non-null   int64  
 11  Point of Contact   4746 non-null   object 
dtypes: float64(2), int64(2), object(8)
memory usage: 445.1+ KB
```

In [30]:
```
rent_df['Area Type'].unique()
```

Out[30]:
```
array(['Super Area', 'Carpet Area', 'Built Area'], dtype=object)
```

In [31]:
```
rent_df['Area Type'].nunique()
```

Out[31]:
```
3
```

In [32]:
```
rent_df['Area Type'].value_counts()
```

Out[32]:
```
Area Type
Super Area     2446
Carpet Area    2298
Built Area        2
Name: count, dtype: int64
```

In [33]:
```
index = ['Area Type','Area Locality','City','Furnishing Status','Te
```

In [34]:
```
for i in index:
    print(i, rent_df[i].nunique())
```

```
Area Type 3
Area Locality 2235
City 6
Furnishing Status 3
Tenant Preferred 3
Point of Contact 3
```

In [35]:
```
rent_df.drop(['Posted On','Floor','Area Locality'], axis=1, inplace
```

In [36]:
```
rent_df = pd.get_dummies(rent_df, columns = ['Area Type', 'City','F
```

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

Out[37]:

	BHK	Rent	Size	Bathroom	Area Type_Built Area	Area Type_Carpet Area	Area Type_Super Area	C
0	2.0	10000	1100.0	2	False	False	True	
1	2.0	20000	800.0	1	False	False	True	
2	2.0	17000	1000.0	1	False	False	True	
3	2.0	10000	800.0	1	False	False	True	
4	2.0	7500	850.0	1	False	True	False	

5 rows × 22 columns

In [38]:
```
X = rent_df.drop('Rent', axis=1)
y = rent_df['Rent']
```

In [39]:
```
from sklearn.model_selection import train_test_split
```

In [40]:
```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0
#random_state 랜덤 돌린거 고정시키려고 아무값으로 정함
```

In [41]:
```
from sklearn.linear_model import LinearRegression
```

In [42]:
```
lr = LinearRegression()
```

In [43]:
```
lr.fit(X_train, y_train)
```

Out[43]:
```
LinearRegression()
```

In [44]:
```
pred = lr.predict(X_test)
```

In [45]:
```
y_test
```

Out[45]:
```
4039     19800
81        8200
3399     13000
2893      9000
4371     13000
         ...
876     250000
2099     26000
3089     11000
4430      8000
280       4500
Name: Rent, Length: 1424, dtype: int64
```

In [46]:
```
from sklearn.metrics import mean_absolute_error, mean_squared_error
```

In [47]:
```
mean_absolute_error(y_test, pred)
```

Out[47]:
```
23006.355884193235
```

In [48]:
```
mean_squared_error(y_test, pred)**0.5
```

Out[48]:
```
38927.79016246277
```

In [49]:
```
y_train.loc[1837]
```

Out[49]:
```
np.int64(3500000)
```

In [50]:
```
X_train.loc[1837]
```

Out[50]:
```
BHK                                      3.0
Size                                  2500.0
Bathroom                                 3
Area Type_Built Area                 False
Area Type_Carpet Area                 True
Area Type_Super Area                 False
City_Bangalore                        True
City_Chennai                         False
City_Delhi                           False
City_Hyderabad                       False
City_Kolkata                         False
City_Mumbai                          False
Furnishing Status_Furnished          False
Furnishing Status_Semi-Furnished      True
Furnishing Status_Unfurnished        False
Tenant Preferred_Bachelors            True
Tenant Preferred_Bachelors/Family    False
Tenant Preferred_Family              False
Point of Contact_Contact Agent        True
Point of Contact_Contact Builder     False
Point of Contact_Contact Owner       False
Name: 1837, dtype: object
```

In [51]:
```
X_train.drop(1837, inplace = True)
y_train.drop(1837, inplace = True)
```

In [52]:
```
lr.fit(X_train, y_train)
```

Out[52]:
```
LinearRegression()
```

In [54]:
```
new_pred = lr.predict(X_test)
```

In [58]:
```
mean_squared_error(y_test, new_pred)**0.5
```

Out[58]:
```
38550.251728440446
```

In [59]:
```
y_train_log = np.log(y_train)
```

In [60]:
```
lr.fit(X_train, y_train_log)
```

Out[60]:
```
LinearRegression()
```

In [61]:
```
newnew_pred = lr.predict(X_test)
```

In [62]:
```
pred_exp = np.exp(newnew_pred)
```

In [64]:
```
mean_squared_error(y_test, pred_exp)**0.5
```

Out[64]: 32632.780523342746

In [65]:
```
print("회귀 계수:", lr.coef_)
print("절편:", lr.intercept_)
```

```
회귀 계수: [ 2.28264168e-01  4.01251785e-04  1.64383148e-01 -6.6006973
2e-02
  5.06875920e-02  1.53193812e-02 -1.05068454e-01 -1.38082649e-01
  5.21658765e-02 -2.48671017e-01 -4.25448948e-01  8.65105191e-01
  1.40222527e-01 -1.36294128e-02 -1.26593115e-01  6.87854278e-02
  4.71370052e-03 -7.34991283e-02  2.84418221e-01 -1.92718765e-01
 -9.16994554e-02]
절편: 8.607305156394444
```

In [ ]: