House_Price_Prediction

November 9, 2024

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

[2]:
```
data= pd.read_csv("House Price India.csv")
```

[3]:
```
data
```

[3]:

	id	Date	number of bedrooms	number of bathrooms	\
0	6762810145	42491	5	2.50	
1	6762810635	42491	4	2.50	
2	6762810998	42491	5	2.75	
3	6762812605	42491	4	2.50	
4	6762812919	42491	3	2.00	
…	…	…	…	…	
14615	6762830250	42734	2	1.50	
14616	6762830339	42734	3	2.00	
14617	6762830618	42734	2	1.00	
14618	6762830709	42734	4	1.00	
14619	6762831463	42734	3	1.00	

	living area	lot area	number of floors	waterfront present	\
0	3650	9050	2.0	0	
1	2920	4000	1.5	0	
2	2910	9480	1.5	0	
3	3310	42998	2.0	0	
4	2710	4500	1.5	0	
…	…	…	…	…	
14615	1556	20000	1.0	0	
14616	1680	7000	1.5	0	
14617	1070	6120	1.0	0	
14618	1030	6621	1.0	0	
14619	900	4770	1.0	0	

	number of views	condition of the house	…	Built Year	\
0	4	5	…	1921	
1	0	5	…	1909	

```
2                    0                       3  …        1939
3                    0                       3  …        2001
4                    0                       4  …        1929
…                  …                       …  …         …
14615                0                       4  …        1957
14616                0                       4  …        1968
14617                0                       3  …        1962
14618                0                       4  …        1955
14619                0                       3  …        1969

       Renovation Year  Postal Code  Lattitude  Longitude  living_area_renov  \
0                    0       122003    52.8645   -114.557               2880
1                    0       122004    52.8878   -114.470               2470
2                    0       122004    52.8852   -114.468               2940
3                    0       122005    52.9532   -114.321               3350
4                    0       122006    52.9047   -114.485               2060
…                  …           …         …         …                  …
14615                0       122066    52.6191   -114.472               2250
14616                0       122072    52.5075   -114.393               1540
14617                0       122056    52.7289   -114.507               1130
14618                0       122042    52.7157   -114.411               1420
14619             2009       122018    52.5338   -114.552                900

       lot_area_renov  Number of schools nearby  Distance from the airport  \
0                5400                         2                         58
1                4000                         2                         51
2                6600                         1                         53
3               42847                         3                         76
4                4500                         1                         51
…                …                          …                          …
14615           17286                         3                         76
14616            7480                         3                         59
14617            6120                         2                         64
14618            6631                         3                         54
14619            3480                         2                         55

         Price
0      2380000
1      1400000
2      1200000
3       838000
4       805000
…         …
14615   221700
14616   219200
14617   209000
14618   205000
```

```
14619   146000

[14620 rows x 23 columns]
```

[4]:
```
data.info()
```

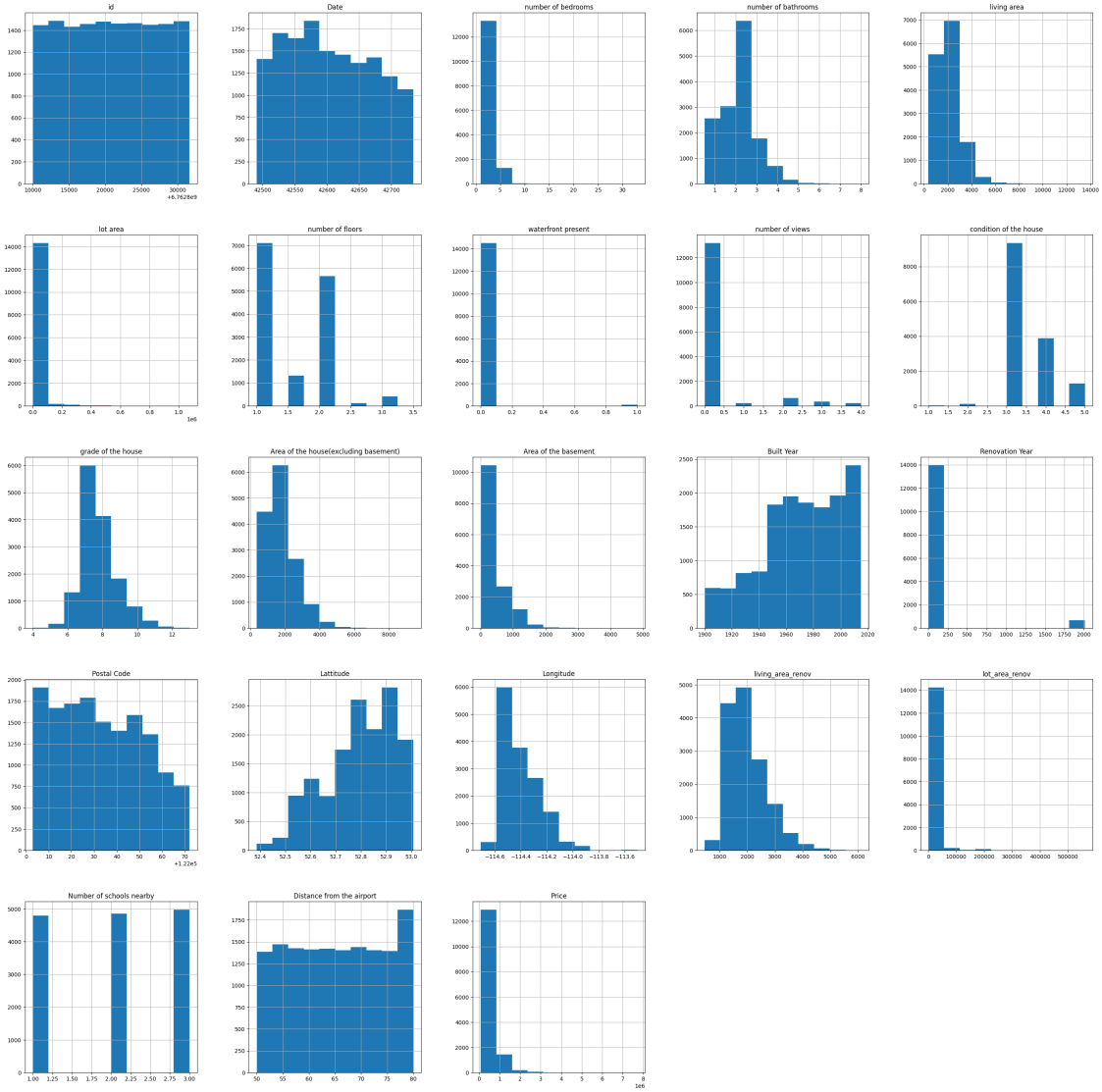
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14620 entries, 0 to 14619
Data columns (total 23 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   id                                     14620 non-null  int64
 1   Date                                   14620 non-null  int64
 2   number of bedrooms                     14620 non-null  int64
 3   number of bathrooms                    14620 non-null  float64
 4   living area                            14620 non-null  int64
 5   lot area                               14620 non-null  int64
 6   number of floors                       14620 non-null  float64
 7   waterfront present                     14620 non-null  int64
 8   number of views                        14620 non-null  int64
 9   condition of the house                 14620 non-null  int64
 10  grade of the house                     14620 non-null  int64
 11  Area of the house(excluding basement)  14620 non-null  int64
 12  Area of the basement                   14620 non-null  int64
 13  Built Year                             14620 non-null  int64
 14  Renovation Year                        14620 non-null  int64
 15  Postal Code                            14620 non-null  int64
 16  Lattitude                              14620 non-null  float64
 17  Longitude                              14620 non-null  float64
 18  living_area_renov                      14620 non-null  int64
 19  lot_area_renov                         14620 non-null  int64
 20  Number of schools nearby               14620 non-null  int64
 21  Distance from the airport              14620 non-null  int64
 22  Price                                  14620 non-null  int64
dtypes: float64(4), int64(19)
memory usage: 2.6 MB
```

[5]:
```
data.dropna(inplace=True)
```

[6]:
```
data.hist(figsize=(35, 35))
```

[6]:
```
array([[<Axes: title={'center': 'id'}>, <Axes: title={'center': 'Date'}>,
        <Axes: title={'center': 'number of bedrooms'}>,
        <Axes: title={'center': 'number of bathrooms'}>,
        <Axes: title={'center': 'living area'}>],
       [<Axes: title={'center': 'lot area'}>,
        <Axes: title={'center': 'number of floors'}>,
        <Axes: title={'center': 'waterfront present'}>,
```

```
  <Axes: title={'center': 'number of views'}>,
  <Axes: title={'center': 'condition of the house'}>],
 [<Axes: title={'center': 'grade of the house'}>,
  <Axes: title={'center': 'Area of the house(excluding basement)'}>,
  <Axes: title={'center': 'Area of the basement'}>,
  <Axes: title={'center': 'Built Year'}>,
  <Axes: title={'center': 'Renovation Year'}>],
 [<Axes: title={'center': 'Postal Code'}>,
  <Axes: title={'center': 'Lattitude'}>,
  <Axes: title={'center': 'Longitude'}>,
  <Axes: title={'center': 'living_area_renov'}>,
  <Axes: title={'center': 'lot_area_renov'}>],
 [<Axes: title={'center': 'Number of schools nearby'}>,
  <Axes: title={'center': 'Distance from the airport'}>,
  <Axes: title={'center': 'Price'}>, <Axes: >, <Axes: >]],
dtype=object)
```



[7]:
```
data.corr()
```

[7]:

	id	Date	number of bedrooms	\
id	1.000000	0.045966	-0.329034	
Date	0.045966	1.000000	-0.015663	
number of bedrooms	-0.329034	-0.015663	1.000000	
number of bathrooms	-0.516909	-0.026485	0.509784	
living area	-0.648127	-0.021958	0.570526	
lot area	-0.100269	0.004392	0.034416	
number of floors	-0.312305	-0.010335	0.177294	
waterfront present	-0.112937	0.012006	-0.006257	
number of views	-0.293004	-0.004782	0.078665	
condition of the house	-0.045061	-0.027402	0.026597	

grade of the house	-0.673448	-0.033097	0.352945
Area of the house(excluding basement)	-0.565116	-0.015994	0.473599
Area of the basement	-0.290806	-0.015711	0.300332
Built Year	-0.068645	-0.005869	0.152954
Renovation Year	-0.109155	-0.011636	0.016132
Postal Code	0.294709	0.018243	-0.044156
Lattitude	-0.479334	-0.023327	-0.013163
Longitude	-0.070841	-0.018231	0.135712
living_area_renov	-0.599900	-0.032495	0.389855
lot_area_renov	-0.089604	-0.000050	0.029400
Number of schools nearby	-0.004821	-0.004071	0.003397
Distance from the airport	-0.004542	0.011457	-0.006157
Price	-0.773114	-0.027919	0.308460

	number of bathrooms	living area	\
id	-0.516909	-0.648127	
Date	-0.026485	-0.021958	
number of bedrooms	0.509784	0.570526	
number of bathrooms	1.000000	0.753517	
living area	0.753517	1.000000	
lot area	0.080806	0.174420	
number of floors	0.502924	0.354743	
waterfront present	0.060104	0.105837	
number of views	0.183789	0.287728	
condition of the house	-0.128232	-0.063358	
grade of the house	0.663054	0.761835	
Area of the house(excluding basement)	0.684391	0.875793	
Area of the basement	0.287190	0.441491	
Built Year	0.498127	0.309602	
Renovation Year	0.049669	0.059400	
Postal Code	-0.105546	-0.080303	
Lattitude	0.031156	0.054518	
Longitude	0.223904	0.240208	
living_area_renov	0.570530	0.757571	
lot_area_renov	0.078627	0.180312	
Number of schools nearby	0.002180	0.002370	
Distance from the airport	0.009206	0.002511	
Price	0.531735	0.712169	

	lot area	number of floors	\
id	-0.100269	-0.312305	
Date	0.004392	-0.010335	
number of bedrooms	0.034416	0.177294	
number of bathrooms	0.080806	0.502924	
living area	0.174420	0.354743	
lot area	1.000000	-0.004138	
number of floors	-0.004138	1.000000	

waterfront present	0.026282	0.016316
number of views	0.078308	0.020153
condition of the house	-0.008548	-0.269928
grade of the house	0.110546	0.463082
Area of the house(excluding basement)	0.183553	0.525643
Area of the basement	0.019755	-0.242976
Built Year	0.051615	0.481565
Renovation Year	0.006848	0.006705
Postal Code	0.070131	-0.129788
Lattitude	-0.090983	0.050731
Longitude	0.221432	0.127550
living_area_renov	0.149744	0.285093
lot_area_renov	0.706812	-0.010120
Number of schools nearby	-0.012671	-0.007579
Distance from the airport	0.003291	0.016567
Price	0.081992	0.262732

	waterfront present	number of views	\
id	-0.112937	-0.293004	
Date	0.012006	-0.004782	
number of bedrooms	-0.006257	0.078665	
number of bathrooms	0.060104	0.183789	
living area	0.105837	0.287728	
lot area	0.026282	0.078308	
number of floors	0.016316	0.020153	
waterfront present	1.000000	0.400206	
number of views	0.400206	1.000000	
condition of the house	0.018644	0.052533	
grade of the house	0.079831	0.254532	
Area of the house(excluding basement)	0.071865	0.162672	
Area of the basement	0.085441	0.293062	
Built Year	-0.024226	-0.055357	
Renovation Year	0.085865	0.102944	
Postal Code	0.038318	0.039268	
Lattitude	-0.021795	-0.004555	
Longitude	-0.047791	-0.079706	
living_area_renov	0.085743	0.281452	
lot_area_renov	0.032055	0.072300	
Number of schools nearby	0.001563	0.008004	
Distance from the airport	0.001448	-0.001657	
Price	0.263687	0.395973	

	condition of the house	…	\
id	-0.045061	…	
Date	-0.027402	…	
number of bedrooms	0.026597	…	
number of bathrooms	-0.128232	…	

```
living area                                -0.063358  …
lot area                                   -0.008548  …
number of floors                           -0.269928  …
waterfront present                          0.018644  …
number of views                             0.052533  …
condition of the house                      1.000000  …
grade of the house                         -0.152530  …
Area of the house(excluding basement)      -0.167695  …
Area of the basement                        0.180609  …
Built Year                                 -0.381718  …
Renovation Year                            -0.062126  …
Postal Code                                 0.045334  …
Lattitude                                  -0.002998  …
Longitude                                  -0.121189  …
living_area_renov                          -0.099743  …
lot_area_renov                             -0.004748  …
Number of schools nearby                   -0.006939  …
Distance from the airport                  -0.002136  …
Price                                       0.041376  …

                                        Built Year  Renovation Year  \
id                                       -0.068645        -0.109155
Date                                     -0.005869        -0.011636
number of bedrooms                        0.152954         0.016132
number of bathrooms                       0.498127         0.049669
living area                               0.309602         0.059400
lot area                                  0.051615         0.006848
number of floors                          0.481565         0.006705
waterfront present                       -0.024226         0.085865
number of views                          -0.055357         0.102944
condition of the house                   -0.381718        -0.062126
grade of the house                        0.440358         0.014501
Area of the house(excluding basement)     0.419369         0.025727
Area of the basement                     -0.138843         0.075104
Built Year                                1.000000        -0.233683
Renovation Year                          -0.233683         1.000000
Postal Code                              -0.062349         0.018006
Lattitude                                -0.143153         0.028908
Longitude                                 0.414591        -0.080050
living_area_renov                         0.328625        -0.002601
lot_area_renov                            0.072874         0.005869
Number of schools nearby                 -0.001631        -0.000826
Distance from the airport                -0.003968         0.005342
Price                                     0.050307         0.133173

                                        Postal Code  Lattitude  Longitude  \
id                                         0.294709  -0.479334  -0.070841
```

Date	0.018243	-0.023327	-0.018231
number of bedrooms	-0.044156	-0.013163	0.135712
number of bathrooms	-0.105546	0.031156	0.223904
living area	-0.080303	0.054518	0.240208
lot area	0.070131	-0.090983	0.221432
number of floors	-0.129788	0.050731	0.127550
waterfront present	0.038318	-0.021795	-0.047791
number of views	0.039268	-0.004555	-0.079706
condition of the house	0.045334	-0.002998	-0.121189
grade of the house	-0.146342	0.115256	0.203754
Area of the house(excluding basement)	-0.083730	-0.000088	0.345899
Area of the basement	-0.010542	0.112989	-0.145879
Built Year	-0.062349	-0.143153	0.414591
Renovation Year	0.018006	0.028908	-0.080050
Postal Code	1.000000	-0.310172	-0.099003
Lattitude	-0.310172	1.000000	-0.131472
Longitude	-0.099003	-0.131472	1.000000
living_area_renov	-0.108454	0.046148	0.341221
lot_area_renov	0.077483	-0.091622	0.258066
Number of schools nearby	0.010605	0.014949	-0.010163
Distance from the airport	0.011528	0.007193	-0.003100
Price	-0.115908	0.297490	0.024414

	living_area_renov	lot_area_renov	\
id	-0.599900	-0.089604	
Date	-0.032495	-0.000050	
number of bedrooms	0.389855	0.029400	
number of bathrooms	0.570530	0.078627	
living area	0.757571	0.180312	
lot area	0.149744	0.706812	
number of floors	0.285093	-0.010120	
waterfront present	0.085743	0.032055	
number of views	0.281452	0.072300	
condition of the house	-0.099743	-0.004748	
grade of the house	0.720019	0.116725	
Area of the house(excluding basement)	0.737744	0.194670	
Area of the basement	0.196403	0.011283	
Built Year	0.328625	0.072874	
Renovation Year	-0.002601	0.005869	
Postal Code	-0.108454	0.077483	
Lattitude	0.046148	-0.091622	
Longitude	0.341221	0.258066	
living_area_renov	1.000000	0.189225	
lot_area_renov	0.189225	1.000000	
Number of schools nearby	-0.001203	-0.025014	
Distance from the airport	-0.005673	-0.014587	
Price	0.584924	0.075535	

	Number of schools nearby
id	-0.004821
Date	-0.004071
number of bedrooms	0.003397
number of bathrooms	0.002180
living area	0.002370
lot area	-0.012671
number of floors	-0.007579
waterfront present	0.001563
number of views	0.008004
condition of the house	-0.006939
grade of the house	0.000986
Area of the house(excluding basement)	-0.002894
Area of the basement	0.010284
Built Year	-0.001631
Renovation Year	-0.000826
Postal Code	0.010605
Lattitude	0.014949
Longitude	-0.010163
living_area_renov	-0.001203
lot_area_renov	-0.025014
Number of schools nearby	1.000000
Distance from the airport	0.004035
Price	0.009890

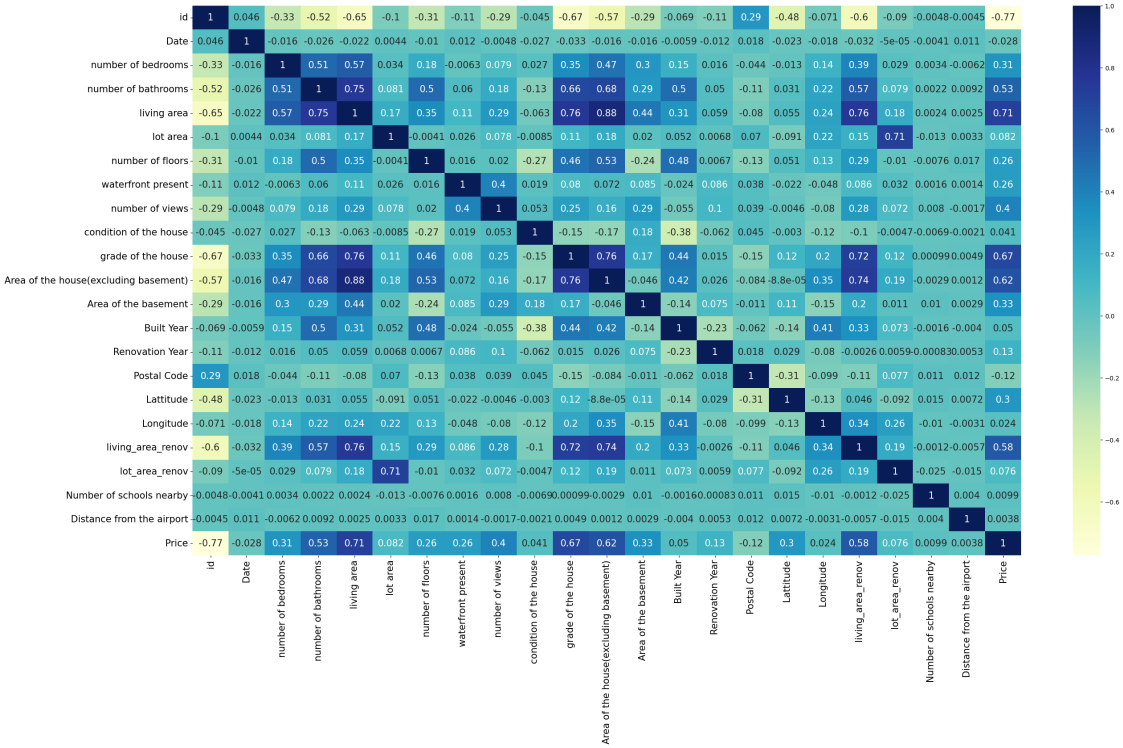
	Distance from the airport	Price
id	-0.004542	-0.773114
Date	0.011457	-0.027919
number of bedrooms	-0.006157	0.308460
number of bathrooms	0.009206	0.531735
living area	0.002511	0.712169
lot area	0.003291	0.081992
number of floors	0.016567	0.262732
waterfront present	0.001448	0.263687
number of views	-0.001657	0.395973
condition of the house	-0.002136	0.041376
grade of the house	0.004940	0.671814
Area of the house(excluding basement)	0.001222	0.615220
Area of the basement	0.002926	0.330202
Built Year	-0.003968	0.050307
Renovation Year	0.005342	0.133173
Postal Code	0.011528	-0.115908
Lattitude	0.007193	0.297490
Longitude	-0.003100	0.024414
living_area_renov	-0.005673	0.584924
lot_area_renov	-0.014587	0.075535

```
Number of schools nearby                                0.004035  0.009890
Distance from the airport                               1.000000  0.003804
Price                                                   0.003804  1.000000

[23 rows x 23 columns]
```

[8]:
```
plt.figure(figsize=(30,16))
sns.heatmap(data.corr(),annot=True,cmap="YlGnBu",annot_kws={"size": 15},)
plt.xticks(rotation=90, fontsize=15)
plt.yticks(rotation=0, fontsize=15)
```

[8]:
```
(array([ 0.5,  1.5,  2.5,  3.5,  4.5,  5.5,  6.5,  7.5,  8.5,  9.5, 10.5,
        11.5, 12.5, 13.5, 14.5, 15.5, 16.5, 17.5, 18.5, 19.5, 20.5, 21.5,
        22.5]),
 [Text(0, 0.5, 'id'),
  Text(0, 1.5, 'Date'),
  Text(0, 2.5, 'number of bedrooms'),
  Text(0, 3.5, 'number of bathrooms'),
  Text(0, 4.5, 'living area'),
  Text(0, 5.5, 'lot area'),
  Text(0, 6.5, 'number of floors'),
  Text(0, 7.5, 'waterfront present'),
  Text(0, 8.5, 'number of views'),
  Text(0, 9.5, 'condition of the house'),
  Text(0, 10.5, 'grade of the house'),
  Text(0, 11.5, 'Area of the house(excluding basement)'),
  Text(0, 12.5, 'Area of the basement'),
  Text(0, 13.5, 'Built Year'),
  Text(0, 14.5, 'Renovation Year'),
  Text(0, 15.5, 'Postal Code'),
  Text(0, 16.5, 'Lattitude'),
  Text(0, 17.5, 'Longitude'),
  Text(0, 18.5, 'living_area_renov'),
  Text(0, 19.5, 'lot_area_renov'),
  Text(0, 20.5, 'Number of schools nearby'),
  Text(0, 21.5, 'Distance from the airport'),
  Text(0, 22.5, 'Price')])
```



[9]:
```
Q1 = data['Price'].quantile(0.25)
Q3 = data['Price'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

data= data[(data['Price'] >= lower_bound) & (data['Price'] <= upper_bound)]
```

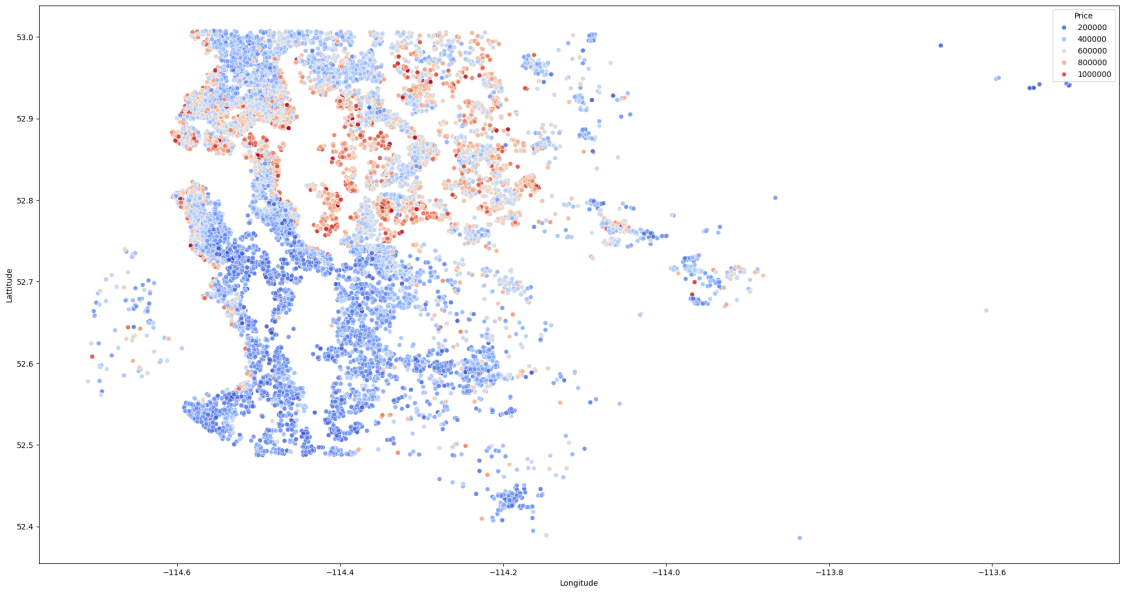
[10]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13859 entries, 3 to 14619
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              13859 non-null  int64
 1   Date                            13859 non-null  int64
 2   number of bedrooms              13859 non-null  int64
 3   number of bathrooms             13859 non-null  float64
 4   living area                     13859 non-null  int64
 5   lot area                        13859 non-null  int64
 6   number of floors                13859 non-null  float64
 7   waterfront present              13859 non-null  int64
```

```
 8   number of views                          13859 non-null  int64
 9   condition of the house                   13859 non-null  int64
 10  grade of the house                       13859 non-null  int64
 11  Area of the house(excluding basement)    13859 non-null  int64
 12  Area of the basement                     13859 non-null  int64
 13  Built Year                               13859 non-null  int64
 14  Renovation Year                          13859 non-null  int64
 15  Postal Code                              13859 non-null  int64
 16  Lattitude                                13859 non-null  float64
 17  Longitude                                13859 non-null  float64
 18  living_area_renov                        13859 non-null  int64
 19  lot_area_renov                           13859 non-null  int64
 20  Number of schools nearby                 13859 non-null  int64
 21  Distance from the airport                13859 non-null  int64
 22  Price                                    13859 non-null  int64
dtypes: float64(4), int64(19)
memory usage: 2.5 MB
```

[11]:
```
plt.figure(figsize=(25,13))
sns.
 ↪scatterplot(x='Longitude',y='Lattitude',data=data,hue='Price',palette='coolwarm')
```

[11]: <Axes: xlabel='Longitude', ylabel='Lattitude'>



[12]:
```
from sklearn.model_selection import train_test_split
```

[13]:
```
x = data.drop(['Price'], axis=1)  # Drop the 'Price' column from the dataframe␣
 ↪to use the rest as features
```

```
y = data['Price']
```

[14]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2)
```

[15]:
```
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline

pipe=Pipeline([
    ('scale',StandardScaler()),
    ('model',LinearRegression())
])
#mod = LinearRegression()

pipe.fit(x_train, y_train)
```

[15]:
```
Pipeline(steps=[('scale', StandardScaler()), ('model', LinearRegression())])
```

[16]:
```
pipe.predict(x_test)  # Predicting the target values
```

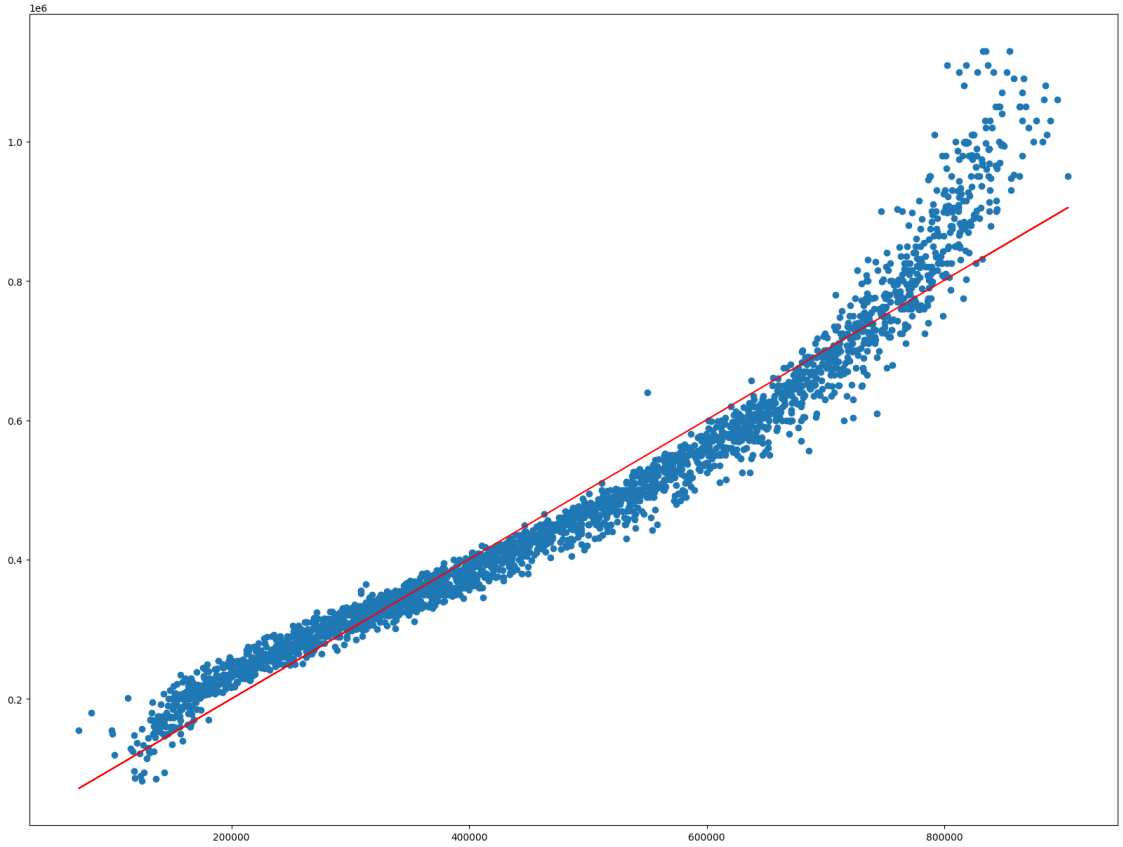
[16]:
```
array([311739.00709258, 163411.19180275, 627288.59441258, …,
       504981.34290785, 820194.03195988, 823320.49272666])
```

[17]:
```
pred = pipe.predict(x_test)  # Predicting the target values
pipe.score(x_test, y_test)  # Evaluating the model with test data
```

[17]:
```
0.9426109186432817
```

[18]:
```
plt.figure(figsize=(20,15))
plt.scatter(pred,y_test)
slope,intercept=np.polyfit(pred,y_test,1)
regression_line = slope * pred + intercept
plt.plot(pred, regression_line, color='red', label='Regression Line')
```

[18]:
```
[<matplotlib.lines.Line2D at 0x248b7d35d10>]
```



[19]:
```
 from sklearn.neighbors import KNeighborsRegressor

knn = Pipeline([
    ('scale', StandardScaler()),  # Feature scaling
    ('mod', KNeighborsRegressor())  # KNN model
])

# Fit the pipeline with the training data
knn.fit(x_train, y_train)
```

[19]:
```
Pipeline(steps=[('scale', StandardScaler()), ('mod', KNeighborsRegressor())])
```

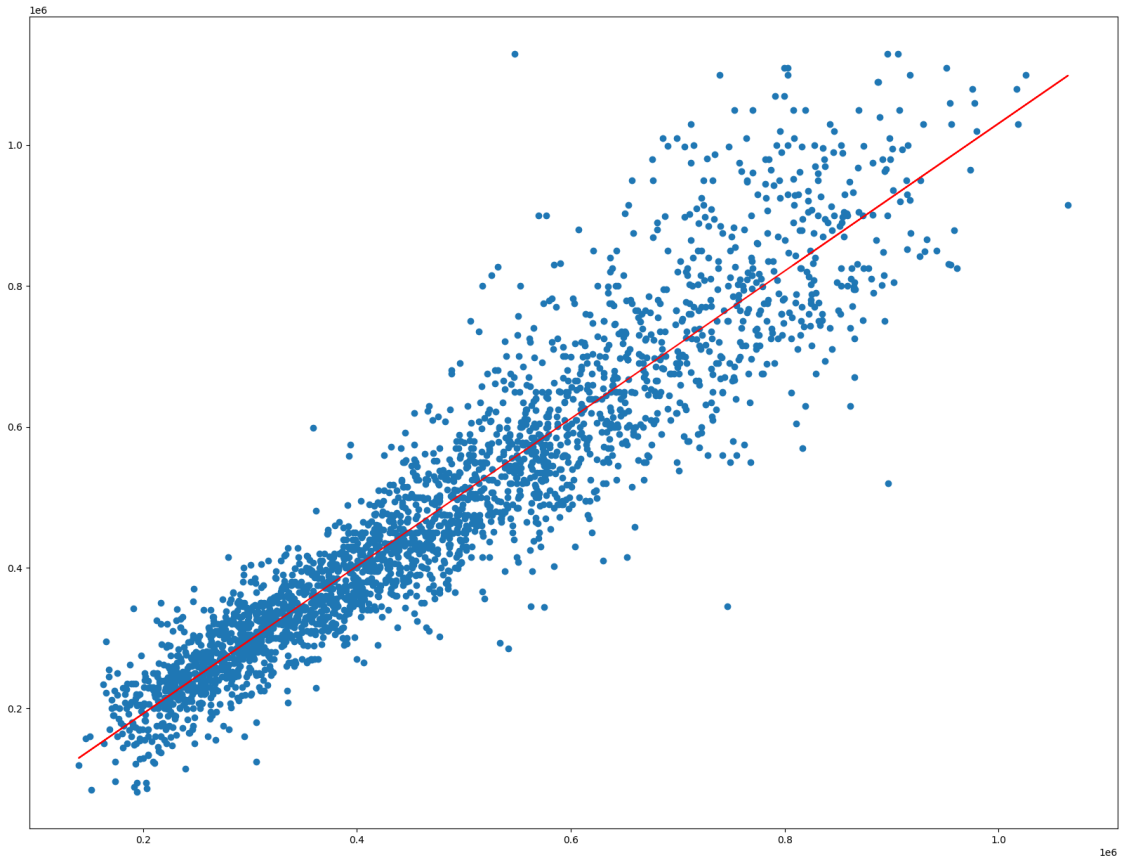
[20]:
```
Knnpred = knn.predict(x_test)  # Predicting the target values
knn.score(x_test, y_test)
```

[20]:
```
0.8709536432983814
```

[ ]:
```
plt.figure(figsize=(20,15))
plt.scatter(Knnpred,y_test)
slope,intercept=np.polyfit(Knnpred,y_test,1)
```

```
regression_line = slope * Knnpred + intercept
plt.plot(Knnpred, regression_line, color='red', label='Regression Line')
```

[ ]:
```
[<matplotlib.lines.Line2D at 0x248ca333750>]
```



[22]:
```
knn.get_params()
```

[22]:
```
{'memory': None,
 'steps': [('scale', StandardScaler()), ('mod', KNeighborsRegressor())],
 'verbose': False,
 'scale': StandardScaler(),
 'mod': KNeighborsRegressor(),
 'scale__copy': True,
 'scale__with_mean': True,
 'scale__with_std': True,
 'mod__algorithm': 'auto',
 'mod__leaf_size': 30,
 'mod__metric': 'minkowski',
 'mod__metric_params': None,
 'mod__n_jobs': None,
```

```
 'mod__n_neighbors': 5,
 'mod__p': 2,
 'mod__weights': 'uniform'}
```

[23]:
```
from sklearn.model_selection import GridSearchCV
gridknn= GridSearchCV(estimator= knn,
                      param_grid={'mod__n_neighbors':[1,2,3,4,5,6,7,8,9,10]},
                      cv=3)
```

[24]:
```
gridknn.fit(x_train,y_train)
best_score = gridknn.best_score_

print(f"Best k value: {best_score}")
```

```
Best k value: 0.8623147023439707
```

[ ]:
```
gridknn.score(x_test,y_test)
```

[ ]:
```
0.8728762890558124
```

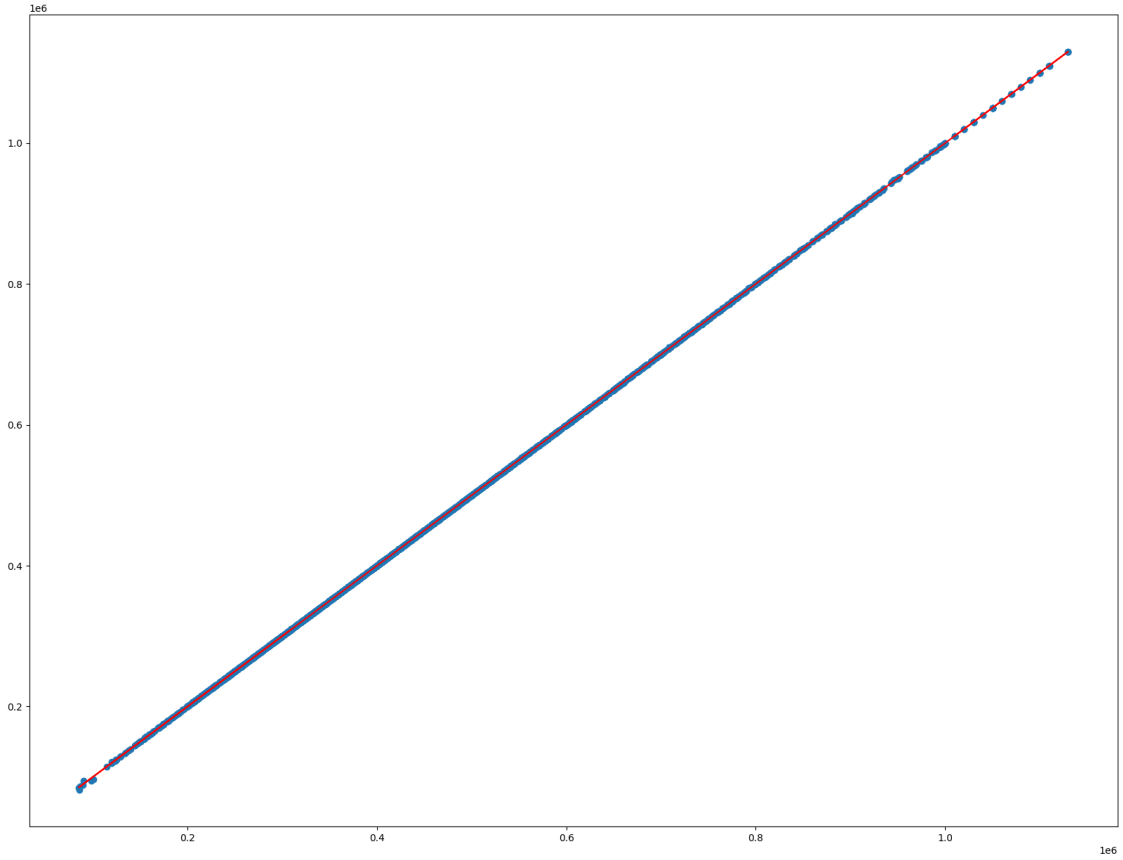
[26]:
```
from sklearn.ensemble import RandomForestRegressor
rf = Pipeline([
    ('scale', StandardScaler()),
    ('rf', RandomForestRegressor())
])

rf.fit(x_train, y_train)
rfpred=rf.predict(x_test)
rf.score(x_test, y_test)
```

[26]:
```
0.9999988927382535
```

[27]:
```
plt.figure(figsize=(20,15))
plt.scatter(rfpred,y_test)
slope,intercept=np.polyfit(rfpred,y_test,1)
regression_line = slope * rfpred + intercept
plt.plot(rfpred, regression_line, color='red', label='Regression Line')
```

[27]:
```
[<matplotlib.lines.Line2D at 0x248b886fd90>]
```


1e6
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
1e6

[ ]: