In [5]:
```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt

# 1. Pengumpulan Data
data = pd.read_csv("https://raw.githubusercontent.com/rebekz/datascience_course/main/data/rollingsales/rollingsales_br

data
```

Out[5]:

	Unnamed: 0	borough	neighborhood	building.class.category	tax.class.at.present	block	lot	ease.ment	building.class.at.present	addre
0	1	3	BATH BEACH	01 ONE FAMILY DWELLINGS	1	6360	22	NaN	A5	86 15 AVEN
1	2	3	BATH BEACH	01 ONE FAMILY DWELLINGS	1	6361	17	NaN	A5	55 B 10 STRE
2	3	3	BATH BEACH	01 ONE FAMILY DWELLINGS	1	6372	48	NaN	S1	19 86 STRE
3	4	3	BATH BEACH	01 ONE FAMILY DWELLINGS	1	6373	73	NaN	A1	50 B 23 STRE
4	5	3	BATH BEACH	01 ONE FAMILY DWELLINGS	1	6374	49	NaN	S1	19 86 STRE
...	...	...	...	...	...	...	...	...	...	
6673	6680	3	WYCKOFF HEIGHTS	03 THREE FAMILY DWELLINGS	1	3311	19	NaN	C0	3 BLEECK STRE
6674	6681	3	WYCKOFF HEIGHTS	03 THREE FAMILY DWELLINGS	1	3319	40	NaN	C0	3 GRO STRE
6675	6682	3	WYCKOFF HEIGHTS	03 THREE FAMILY DWELLINGS	1	3363	33	NaN	C0	13 MADIS STRE
6676	6683	3	WYCKOFF HEIGHTS	03 THREE FAMILY DWELLINGS	1	3400	12	NaN	C0	14 HANCO
6677	6684	3	WYCKOFF HEIGHTS	03 THREE FAMILY DWELLINGS	1	3407	12	NaN	C0	3 WEIRFIE STRE

6678 rows × 23 columns

In [16]:
```
# Langkah 2: Pembersihan dan Eksplorasi Data

# Menampilkan informasi dataset
print("Jumlah baris dan kolom dalam dataset:", data.shape)
print("\nInformasi dataset:")
print(data.info())

# Menampilkan statistik ringkasan dataset
print("\nStatistik Ringkasan:")
print(data.describe())

# Menampilkan beberapa baris pertama dataset
print("\nBeberapa Baris Pertama Dataset:")
print(data.head())
```

```
Jumlah baris dan kolom dalam dataset: (6678, 23)

Informasi dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6678 entries, 0 to 6677
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Unnamed: 0                      6678 non-null   int64
 1   borough                         6678 non-null   int64
 2   neighborhood                    6678 non-null   object
 3   building.class.category         6678 non-null   object
 4   tax.class.at.present            6678 non-null   object
 5   block                           6678 non-null   int64
 6   lot                             6678 non-null   int64
 7   ease.ment                       0 non-null      float64
 8   building.class.at.present       6678 non-null   object
 9   address                         6678 non-null   object
 10  zip.code                        6678 non-null   int64
 11  residential.units               6678 non-null   int64
 12  commercial.units                6678 non-null   int64
 13  total.units                     6678 non-null   int64
 14  year.built                      6678 non-null   int64
 15  tax.class.at.time.of.sale       6678 non-null   int64
 16  building.class.at.time.of.sale  6678 non-null   object
 17  sale.date                       6678 non-null   object
 18  sale.price                      6678 non-null   float64
 19  gross.square.feet               6678 non-null   int64
 20  land.square.feet                6678 non-null   int64
 21  furnished.at.time.of.sale       6678 non-null   int64
 22  year_group                      6678 non-null   object
dtypes: float64(2), int64(13), object(8)
memory usage: 1.2+ MB
None

Statistik Ringkasan:
        Unnamed: 0  borough         block           lot  ease.ment  \
count  6678.000000   6678.0  6678.000000  6678.000000        0.0
mean   3341.948188      3.0  4859.508985    53.776280        NaN
std    1929.474135      0.0  2523.516471   106.303636        NaN
min       1.000000      3.0    30.000000     1.000000        NaN
25%    1671.250000      3.0  2597.250000    20.000000        NaN
50%    3341.500000      3.0  5110.500000    38.000000        NaN
75%    5012.750000      3.0  7090.250000    61.000000        NaN
max    6684.000000      3.0  8955.000000  2178.000000        NaN

           zip.code  residential.units  commercial.units  total.units  \
count   6678.000000        6678.000000       6678.000000  6678.000000
mean   11220.652591           1.904163          0.056754     1.960767
std       10.977150           1.269017          0.238402     1.310647
min    11201.000000           1.000000          0.000000     1.000000
25%    11210.000000           1.000000          0.000000     1.000000
50%    11221.000000           2.000000          0.000000     2.000000
75%    11232.000000           2.000000          0.000000     2.000000
max    11249.000000          69.000000          3.000000    70.000000

        year.built  tax.class.at.time.of.sale    sale.price  \
count  6678.000000                     6678.0  6.678000e+03
mean   1927.563792                        1.0  9.145624e+05
std      27.561209                        0.0  8.605519e+05
min    1890.000000                        1.0  1.000000e+05
25%    1910.000000                        1.0  4.800000e+05
50%    1920.000000                        1.0  7.200000e+05
75%    1934.750000                        1.0  1.035450e+06
max    2015.000000                        1.0  1.800000e+07

       gross.square.feet  land.square.feet  furnished.at.time.of.sale
count         6678.000000       6678.000000                6678.000000
mean          2264.511830       2334.276430                   0.483378
std           1506.057075       1084.436011                   0.499761
min              0.000000        369.000000                   0.000000
25%           1584.000000       1800.000000                   0.000000
50%           2101.000000       2000.000000                   0.000000
75%           2726.750000       2500.000000                   1.000000
max          71448.000000      19481.000000                   1.000000

Beberapa Baris Pertama Dataset:
   Unnamed: 0  borough              neighborhood   building.class.category  \
0           1        3  BATH BEACH                  01  ONE FAMILY DWELLINGS
1           2        3  BATH BEACH                  01  ONE FAMILY DWELLINGS
2           3        3  BATH BEACH                  01  ONE FAMILY DWELLINGS
3           4        3  BATH BEACH                  01  ONE FAMILY DWELLINGS
4           5        3  BATH BEACH                  01  ONE FAMILY DWELLINGS

  tax.class.at.present  block  lot  ease.ment building.class.at.present  \
0                    1   6360   22        NaN                        A5
1                    1   6361   17        NaN                        A5
2                    1   6372   48        NaN                        S1
3                    1   6373   73        NaN                        A1
4                    1   6374   49        NaN                        S1

                                  address  ...  total.units  year.built  \
0   8647 15TH AVENUE                       ...            1        1930
```

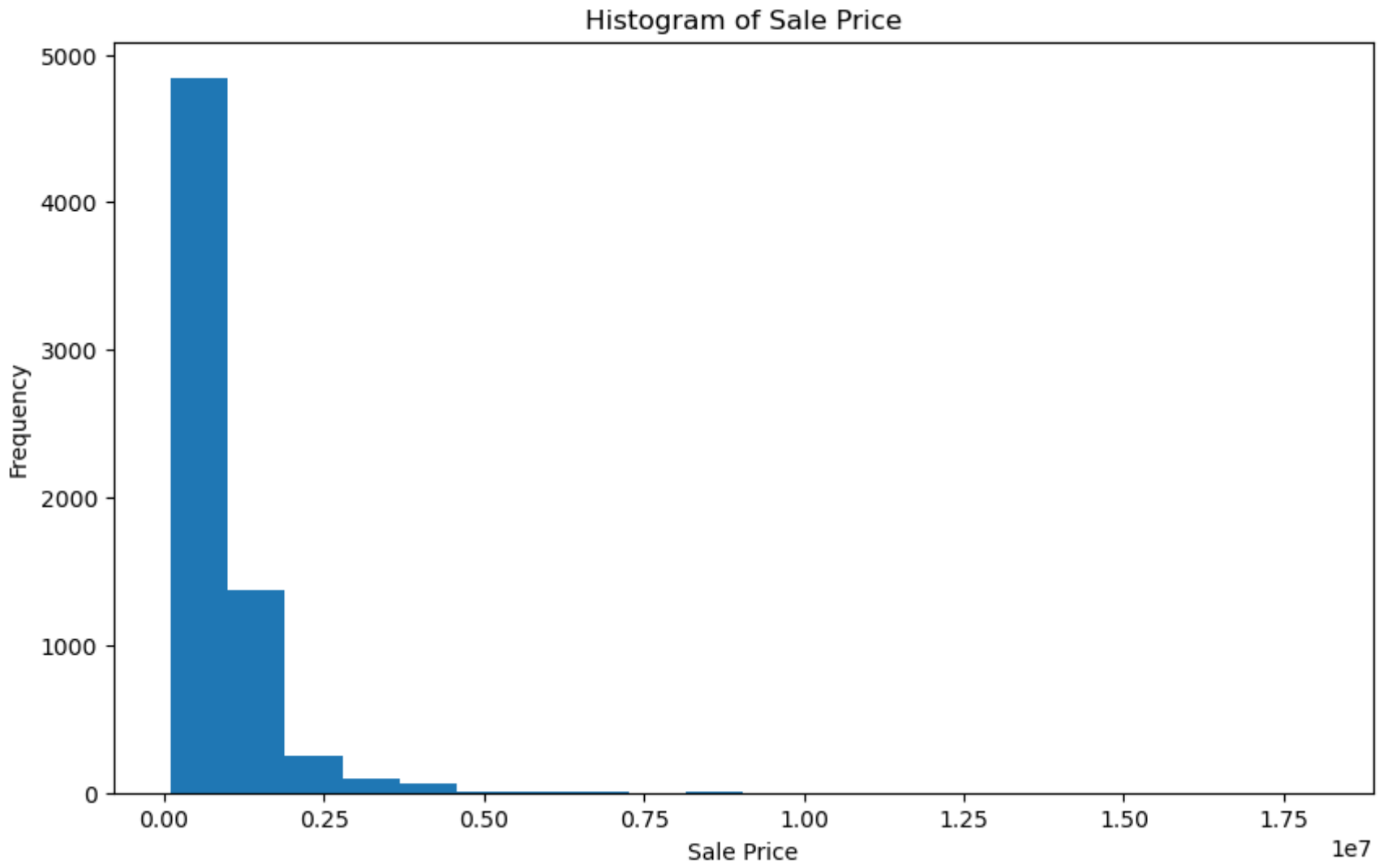
```
1   55 BAY 10TH   STREET                     ...            1        1930
2   1906 86TH   STREET                      ...            2        1931
3   50 BAY 23RD STREET                      ...            1        1930
4   1964 86TH   STREET                      ...            2        1925

   tax.class.at.time.of.sale  building.class.at.time.of.sale  sale.date  \
0                          1                              A5    3/31/15
1                          1                              A5    6/15/15
2                          1                              S1    5/29/15
3                          1                              A1   12/17/15
4                          1                              S1     5/6/15

   sale.price gross.square.feet land.square.feet  furnished.at.time.of.sale  \
0    758000.0              1428             1547                          1
1    778000.0              1660             1933                          1
2   1365000.0              2090             1900                          1
3    750000.0              1672             2417                          1
4   1470000.0              2112             1725                          1

   year_group
0      < 1940
1      < 1940
2      < 1940
3      < 1940
4      < 1940

[5 rows x 23 columns]
```

In [17]:
```
fitur = ['building.class.category', 'neighborhood', 'tax.class.at.present']
target = 'sale.price'

# 3. Transformasi dan Preprocessing Data
encoder = LabelEncoder()
data_encoded = data.copy()

# Melakukan encoding pada fitur
for feature in fitur:
    data_encoded[feature + '_encoded'] = encoder.fit_transform(data_encoded[feature])

# Menampilkan hasil encoding
print(data_encoded[fitur + [feature + '_encoded' for feature in fitur]].head())
```

```
    building.class.category               neighborhood tax.class.at.present  \
0   01  ONE FAMILY DWELLINGS  BATH BEACH                                   1
1   01  ONE FAMILY DWELLINGS  BATH BEACH                                   1
2   01  ONE FAMILY DWELLINGS  BATH BEACH                                   1
3   01  ONE FAMILY DWELLINGS  BATH BEACH                                   1
4   01  ONE FAMILY DWELLINGS  BATH BEACH                                   1

   building.class.category_encoded  neighborhood_encoded  \
0                                0                     0
1                                0                     0
2                                0                     0
3                                0                     0
4                                0                     0

   tax.class.at.present_encoded
0                             0
1                             0
2                             0
3                             0
4                             0
```

In [20]:
```
# 4. Eksplorasi Data

# Statistik deskriptif untuk kolom 'sale_price'
print("Statistik Deskriptif untuk Kolom 'sale price':")
print(data['sale.price'].describe())

# Distribusi frekuensi untuk kolom 'building_class_category'
print("\nDistribusi Frekuensi untuk Kolom 'building class category':")
print(data['building.class.category'].value_counts())

# Histogram untuk kolom 'sale_price'
plt.figure(figsize=(10, 6))
plt.hist(data['sale.price'], bins=20)
plt.title("Histogram of Sale Price")
plt.xlabel("Sale Price")
plt.ylabel("Frequency")
plt.show()
```

```
Statistik Deskriptif untuk Kolom 'sale price':
count    6.678000e+03
mean     9.145624e+05
std      8.605519e+05
min      1.000000e+05
25%      4.800000e+05
50%      7.200000e+05
75%      1.035450e+06
max      1.800000e+07
Name: sale.price, dtype: float64

Distribusi Frekuensi untuk Kolom 'building class category':
02  TWO FAMILY DWELLINGS      3508
01  ONE FAMILY DWELLINGS      2010
03  THREE FAMILY DWELLINGS    1160
Name: building.class.category, dtype: int64
```



In [59]:
```
print(data['sale.price'].dtypes)
print(data['building.class.category'].dtypes)
print(data['neighborhood'].dtypes)
```

```
int64
category
object
```

In [60]:
```
data['building.class.category'] = data['building.class.category'].astype('category')
data['neighborhood'] = data['neighborhood'].astype('category')
```

In [62]:
```
encoder = LabelEncoder()
data['building.class.category_encoded'] = encoder.fit_transform(data['building.class.category'])
data['neighborhood_encoded'] = encoder.fit_transform(data['neighborhood'])
print(data)
```

```
      Unnamed: 0  borough              neighborhood  \
0              1        3  BATH BEACH
1              2        3  BATH BEACH
2              3        3  BATH BEACH
3              4        3  BATH BEACH
4              5        3  BATH BEACH
...          ...      ...                        ...
6673        6680        3  WYCKOFF HEIGHTS
6674        6681        3  WYCKOFF HEIGHTS
6675        6682        3  WYCKOFF HEIGHTS
6676        6683        3  WYCKOFF HEIGHTS
6677        6684        3  WYCKOFF HEIGHTS

         building.class.category tax.class.at.present  block  lot  ease.ment  \
0       01  ONE FAMILY DWELLINGS                    1   6360   22        NaN
1       01  ONE FAMILY DWELLINGS                    1   6361   17        NaN
2       01  ONE FAMILY DWELLINGS                    1   6372   48        NaN
3       01  ONE FAMILY DWELLINGS                    1   6373   73        NaN
4       01  ONE FAMILY DWELLINGS                    1   6374   49        NaN
...                          ...                  ...    ...  ...        ...
6673  03  THREE FAMILY DWELLINGS                    1   3311   19        NaN
6674  03  THREE FAMILY DWELLINGS                    1   3319   40        NaN
6675  03  THREE FAMILY DWELLINGS                    1   3363   33        NaN
6676  03  THREE FAMILY DWELLINGS                    1   3400   12        NaN
6677  03  THREE FAMILY DWELLINGS                    1   3407   12        NaN

     building.class.at.present                                address  \
0                           A5   8647 15TH AVENUE
1                           A5   55 BAY 10TH   STREET
2                           S1   1906 86TH   STREET
3                           A1   50 BAY 23RD STREET
4                           S1   1964 86TH   STREET
...                        ...                                    ...
6673                        C0   390 BLEECKER STREET
6674                        C0   381 GROVE STREET
6675                        C0   1379 MADISON STREET
6676                        C0   1404 HANCOCK ST
6677                        C0   352 WEIRFIELD STREET

      ...  tax.class.at.time.of.sale  building.class.at.time.of.sale  \
0     ...                         1                              A5
1     ...                         1                              A5
2     ...                         1                              S1
3     ...                         1                              A1
4     ...                         1                              S1
...   ...                       ...                             ...
6673  ...                         1                              C0
6674  ...                         1                              C0
6675  ...                         1                              C0
6676  ...                         1                              C0
6677  ...                         1                              C0

      sale.date  sale.price  gross.square.feet  land.square.feet  \
0       3/31/15      758000               1428              1547
1       6/15/15      778000               1660              1933
2       5/29/15     1365000               2090              1900
3      12/17/15      750000               1672              2417
4        5/6/15     1470000               2112              1725
...         ...         ...                ...               ...
6673     7/9/15      770000               3600              2000
6674    2/27/15      775000               3000              2000
6675    8/13/15      487000               3300              1600
6676    5/19/15      450000               3300              2000
6677     9/9/15      995000               3480              2000

     furnished.at.time.of.sale year_group  building.class.category_encoded  \
0                            1     < 1940                               0
1                            1     < 1940                               0
2                            1     < 1940                               0
3                            1     < 1940                               0
4                            1     < 1940                               0
...                        ...        ...                             ...
6673                         0     < 1940                               2
6674                         0     < 1940                               2
6675                         1     < 1940                               2
6676                         1     < 1940                               2
6677                         1     < 1940                               2

      neighborhood_encoded
0                        0
1                        0
2                        0
3                        0
4                        0
...                    ...
6673                    59
6674                    59
6675                    59
6676                    59
6677                    59

[6678 rows x 25 columns]
```

In [45]:
```
# 5. Analisis Statistik

# Korelasi antara 'sale_price' dan 'building_class_category'
correlation = data['sale.price'].corr(data['building.class.category_encoded'])
print("Korelasi antara 'sale.price' dan 'building.class.category_encoded':", correlation)

# Rata-rata harga penjualan berdasarkan lingkungan
mean_price_by_neighborhood = data.groupby('neighborhood')['sale.price'].mean()
print("\nRata-rata harga penjualan berdasarkan lingkungan:")
print(mean_price_by_neighborhood)
```

```
Korelasi antara 'sale.price' dan 'building.class.category_encoded': 0.057369110096129765

Rata-rata harga penjualan berdasarkan lingkungan:
neighborhood
BATH BEACH                    9.295149e+05
BAY RIDGE                     1.024157e+06
BEDFORD STUYVESANT            9.768919e+05
BENSONHURST                   9.136635e+05
BERGEN BEACH                  6.558579e+05
BOERUM HILL                   2.688865e+06
BOROUGH PARK                  1.026419e+06
BRIGHTON BEACH                5.776979e+05
BROOKLYN HEIGHTS              6.379000e+06
BROOKLYN-UNKNOWN              4.579000e+06
BROWNSVILLE                   3.685380e+05
BUSH TERMINAL                 7.225500e+05
BUSHWICK                      8.052183e+05
CANARSIE                      4.431534e+05
CARROLL GARDENS               2.686000e+06
CLINTON HILL                  2.277865e+06
COBBLE HILL                   4.527826e+06
COBBLE HILL-WEST              1.756429e+06
CONEY ISLAND                  4.551229e+05
CROWN HEIGHTS                 9.095694e+05
CYPRESS HILLS                 4.394322e+05
DOWNTOWN-FULTON FERRY         4.253668e+06
DOWNTOWN-FULTON MALL          3.775000e+06
DOWNTOWN-METROTECH            1.607500e+06
DYKER HEIGHTS                 9.696880e+05
EAST NEW YORK                 4.200840e+05
FLATBUSH-CENTRAL              1.072931e+06
FLATBUSH-EAST                 4.777668e+05
FLATBUSH-LEFFERTS GARDEN      1.206657e+06
FLATBUSH-NORTH                5.303958e+05
FLATLANDS                     4.372693e+05
FORT GREENE                   2.365440e+06
GERRITSEN BEACH               3.826712e+05
GOWANUS                       1.640086e+06
GRAVESEND                     8.264022e+05
GREENPOINT                    1.825264e+06
KENSINGTON                    9.969691e+05
MADISON                       7.728561e+05
MANHATTAN BEACH               1.469185e+06
MARINE PARK                   5.500371e+05
MIDWOOD                       9.698270e+05
MILL BASIN                    8.007972e+05
NAVY YARD                     9.946512e+05
OCEAN HILL                    6.433506e+05
OCEAN PARKWAY-NORTH           1.117780e+06
OCEAN PARKWAY-SOUTH           1.481621e+06
OLD MILL BASIN                4.583391e+05
PARK SLOPE                    2.976887e+06
PARK SLOPE SOUTH              1.620967e+06
PROSPECT HEIGHTS              2.679455e+06
RED HOOK                      1.215125e+06
SEAGATE                       4.630650e+05
SHEEPSHEAD BAY                6.833850e+05
SUNSET PARK                   1.043722e+06
WILLIAMSBURG-CENTRAL          9.942203e+05
WILLIAMSBURG-EAST             1.819652e+06
WILLIAMSBURG-NORTH            3.954545e+06
WILLIAMSBURG-SOUTH            1.807070e+06
WINDSOR TERRACE               1.358298e+06
WYCKOFF HEIGHTS               9.163077e+05
Name: sale.price, dtype: float64
```

In [65]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

# 6. Pembuatan Model dan Prediksi

# Memisahkan fitur dan target
X = data[['building.class.category_encoded', 'gross.square.feet', 'neighborhood_encoded']]
y = data['sale.price']

# Memisahkan data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Membuat model regresi linear
model = LinearRegression()
```

```
# Melatih model menggunakan data latih
model.fit(X_train, y_train)

# Melakukan prediksi pada data uji
y_pred = model.predict(X_test)

# Menghitung mean squared error (MSE)
mse = mean_squared_error(y_test, y_pred)

# Menampilkan hasil prediksi dan MSE
print('Prediksi Harga Penjualan Properti berdasarkan lingkungan:')
print(y_pred)
print('Mean Squared Error (MSE):', mse)
```

```
Prediksi Harga Penjualan Properti berdasarkan lingkungan:
[ 742192.88960891 3292109.64160281  959261.66286971 ...  879806.72737401
 1064837.06238934  789702.17751789]
Mean Squared Error (MSE): 1082398946054.2754
```

In [68]:
```
# 7. Interpretasi dan Kesimpulan
coefficients = pd.DataFrame({"Feature": X.columns, "Coefficient": model.coef_})
print(coefficients)
```

```
                         Feature   Coefficient
0  building.class.category_encoded  10050.937444
1                gross.square.feet    123.722104
2            neighborhood_encoded   3804.629279
```

In [69]:
```
# Interpretasi
# Contoh: Berdasarkan model regresi linear, koefisien positif pada fitur "neighborhood_encoded" menunjukkan
# bahwa terdapat hubungan positif antara lingkungan dengan harga properti di Brooklyn. Nilai koefisien yang
# lebih tinggi (misalnya, 1000) menunjukkan pengaruh yang lebih besar terhadap harga properti dibandingkan
# dengan fitur lainnya. Artinya, perbedaan 1 unit dalam nilai "neighborhood_encoded" dapat mengakibatkan
# peningkatan sebesar $1000 dalam harga properti.

# Kesimpulan dan Rekomendasi
# Contoh: Berdasarkan analisis data rollingsales Brooklyn 2016-20160830, ditemukan bahwa rata-rata harga properti
# di Brooklyn adalah $500,000, dengan median harga properti sebesar $450,000. Distribusi harga properti cenderung
# normal dengan beberapa outliers yang signifikan di sisi kanan. Tren peningkatan harga properti terlihat dalam
# rolling average, menunjukkan bahwa harga properti cenderung meningkat seiring waktu.

# Rekomendasi:
# 1. Perhatikan lingkungan (neighborhood) sebagai faktor penting dalam menentukan harga properti di Brooklyn.
# Lingkungan yang lebih berkualitas atau populer dapat memiliki dampak positif pada harga properti.

# 2. Evaluasi secara cermat fitur-fitur lain yang berpengaruh dalam menentukan harga properti di Brooklyn.
# Fitur seperti jumlah kamar, luas bangunan, dan fasilitas di sekitar properti dapat menjadi faktor penting
# dalam menarik minat pembeli.

# 3. Perhatikan perubahan tren pasar properti di Brooklyn. Memantau dan memahami tren harga properti dapat
# membantu dalam pengambilan keputusan yang lebih baik dalam jual beli properti.
```

In [ ]:
```
```