```
# Mount Google Drive to access the dataset
from google.colab import drive
drive.mount('/content/drive')

# File path to the dataset
file_path = '/content/drive/MyDrive/House Price Prediction & Feature Impact Analysis/Real es

# Check if the file exists
import os
print("File exists:", os.path.exists(file_path))

# Load the dataset into a Pandas DataFrame
import pandas as pd
data = pd.read_csv(file_path)

# Display the first few rows and dataset shape
print(data.head())
print("Dataset shape:", data.shape)
```

```
Mounted at /content/drive
File exists: True
   No  X1 transaction date  X2 house age  \
0   1             2012.917          32.0
1   2             2012.917          19.5
2   3             2013.583          13.3
3   4             2013.500          13.3
4   5             2012.833           5.0

   X3 distance to the nearest MRT station  X4 number of convenience stores  \
0                                84.87882                                10
1                               306.59470                                 9
2                               561.98450                                 5
3                               561.98450                                 5
4                               390.56840                                 5

   X5 latitude  X6 longitude  Y house price of unit area
0     24.98298     121.54024                        37.9
1     24.98034     121.53951                        42.2
2     24.98746     121.54391                        47.3
3     24.98746     121.54391                        54.8
4     24.97937     121.54245                        43.1
Dataset shape: (414, 8)
```

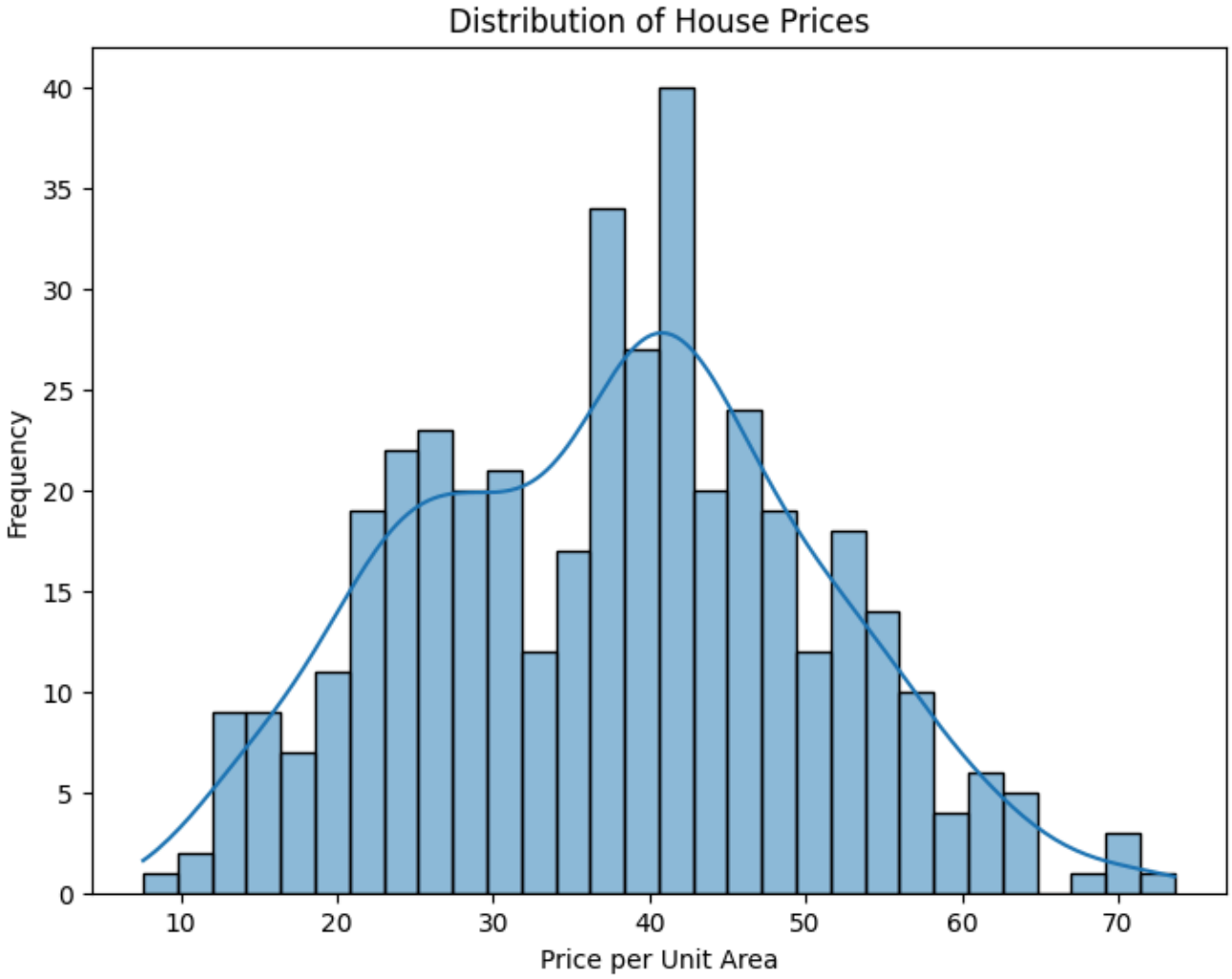
```
# Rename columns for better readability
data.columns = [
    "ID", "Transaction_Date", "House_Age",
    "Distance_to_MRT", "Convenience_Stores",
    "Latitude", "Longitude", "Price_per_Unit_Area"
]
```

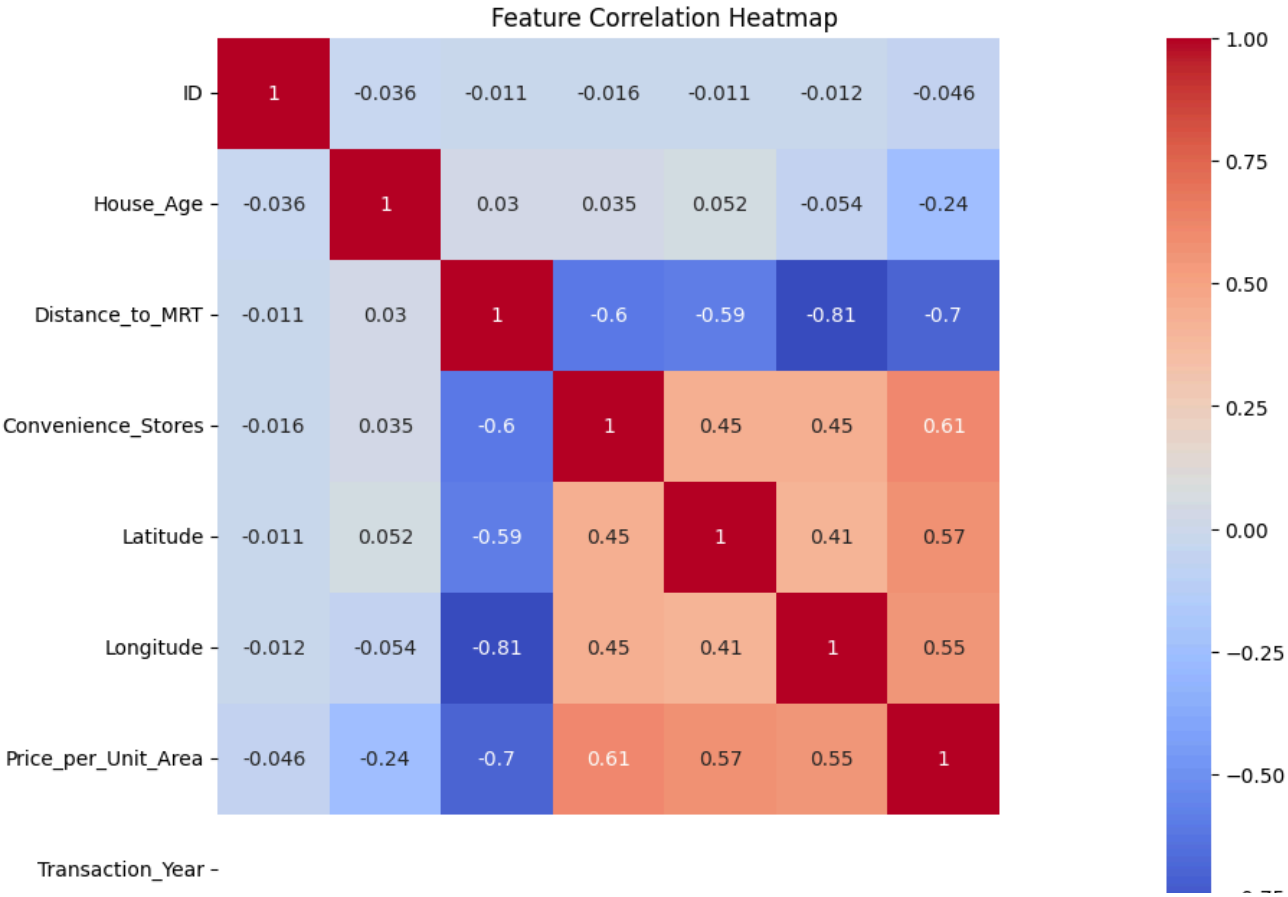
```
# Convert Transaction_Date to a datetime format and extract year
data['Transaction_Date'] = pd.to_datetime(data['Transaction_Date'], format='%Y.%f', errors='
data['Transaction_Year'] = data['Transaction_Date'].dt.year

# Remove outliers from the Price_per_Unit_Area column using the IQR method
def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

data = remove_outliers(data, 'Price_per_Unit_Area')

# Create distance categories
import numpy as np
data['Distance_Category'] = pd.cut(
    data['Distance_to_MRT'],
    bins=[0, 500, 1000, 3000, 5000, np.inf],
    labels=['Very Close', 'Close', 'Moderate', 'Far', 'Very Far']
)

# Save the cleaned dataset
processed_file_path = '/content/drive/MyDrive/House Price Prediction & Feature Impact Analys
data.to_csv(processed_file_path, index=False)
print("Processed data saved at:", processed_file_path)
```

```
Processed data saved at: /content/drive/MyDrive/House Price Prediction & Feature Impact
<ipython-input-2-833b4a2ec6b2>:25: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
  data['Distance_Category'] = pd.cut(
```

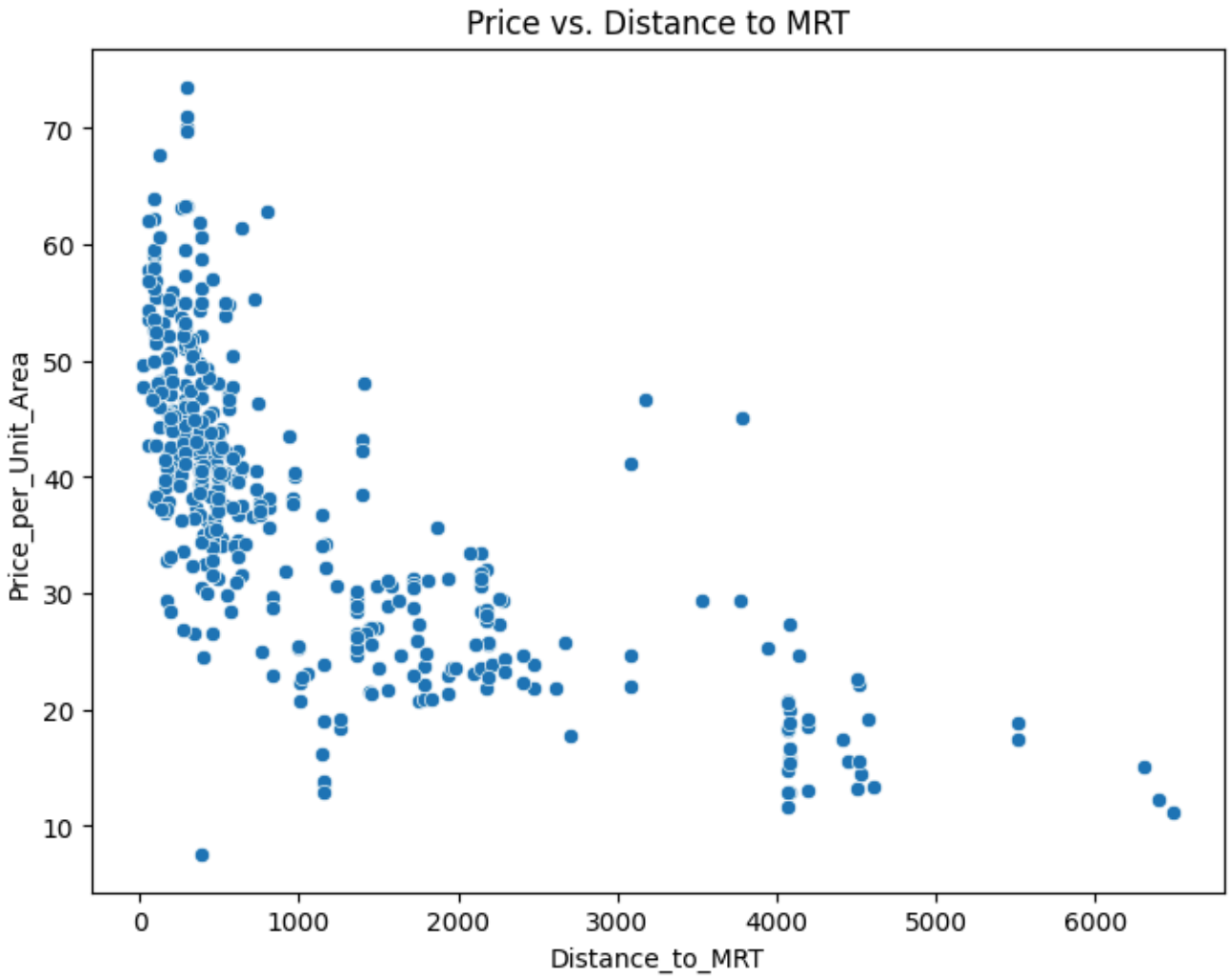
```
import seaborn as sns
import matplotlib.pyplot as plt

# Plot house price distribution
plt.figure(figsize=(8, 6))
sns.histplot(data['Price_per_Unit_Area'], kde=True, bins=30)
plt.title('Distribution of House Prices')
plt.xlabel('Price per Unit Area')
plt.ylabel('Frequency')
plt.show()

# Plot feature correlations
plt.figure(figsize=(10, 8))
```

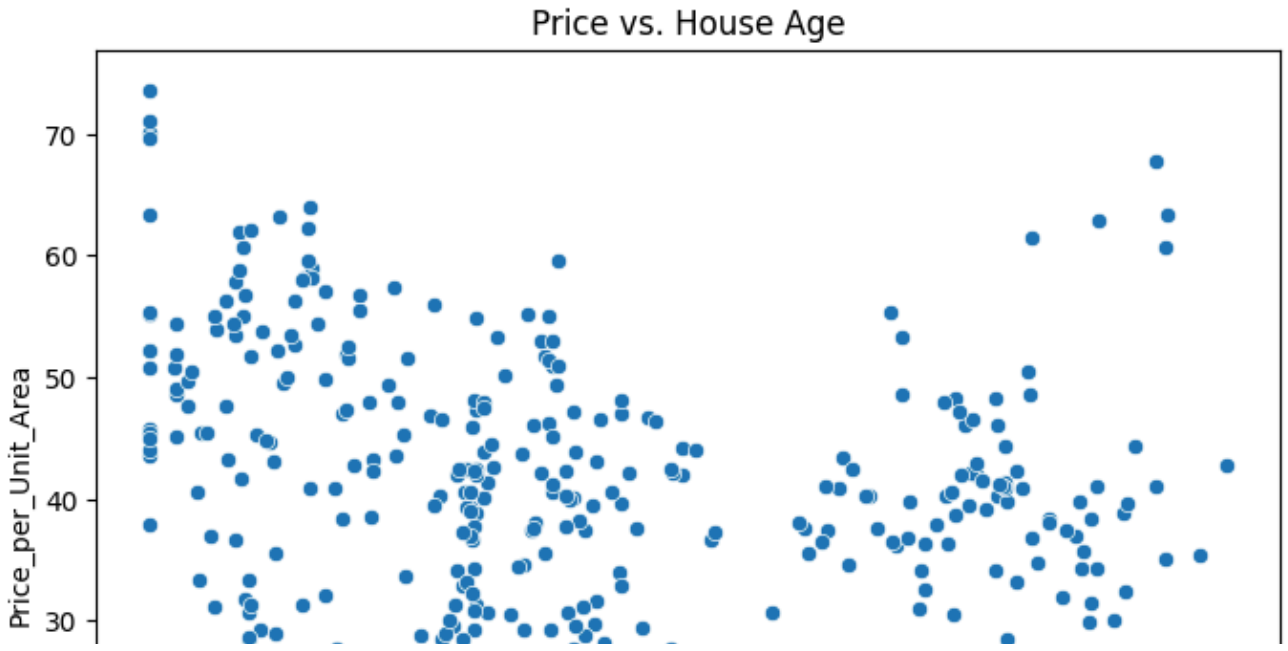
```
sns.heatmap(data.select_dtypes(include=['float64', 'int64']).corr(), annot=True, cmap='coolw
plt.title('Feature Correlation Heatmap')
plt.show()

# Scatter plots for key relationships
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Distance_to_MRT', y='Price_per_Unit_Area', data=data)
plt.title('Price vs. Distance to MRT')
plt.show()

plt.figure(figsize=(8, 6))
sns.scatterplot(x='House_Age', y='Price_per_Unit_Area', data=data)
plt.title('Price vs. House Age')
plt.show()
```


Distribution of House Prices
Frequency
Price per Unit Area
0
5
10
15
20
25
30
35
40
10
20
30
40
50
60
70


Feature Correlation Heatmap
ID
House_Age
Distance_to_MRT
Convenience_Stores
Latitude
Longitude
Price_per_Unit_Area
Transaction_Year
1 -0.036 -0.011 -0.016 -0.011 -0.012 -0.046
-0.036 1 0.03 0.035 0.052 -0.054 -0.24
-0.011 0.03 1 -0.6 -0.59 -0.81 -0.7
-0.016 0.035 -0.6 1 0.45 0.45 0.61
-0.011 0.052 -0.59 0.45 1 0.41 0.57
-0.012 -0.054 -0.81 0.45 0.41 1 0.55
-0.046 -0.24 -0.7 0.61 0.57 0.55 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50


ID
House_Age
Distance_to_MRT
Convenience_Stores
Latitude
Longitude
Price_per_Unit_Area
Transaction_Year
−0.75


Price vs. Distance to MRT
Price_per_Unit_Area
Distance_to_MRT
10
20
30
40
50
60
70
0
1000
2000
3000
4000
5000
6000


Price vs. House Age
Price_per_Unit_Area
30
40
50
60
70