TRANSPORT AND TELECOMMUNICATION INSTITUTE

DATA MINING

Report Computer Practice Nr: 3

Regression Analysis

Riga 2024

Contents

Introduction

In the rapidly evolving real estate market, accurate and timely predictions of property values are crucial for investors, realtors, and potential homeowners. King County, which includes Seattle, Washington, serves as an excellent case study due to its diverse housing stock and dynamic market conditions. The dataset obtained from Kaggle, originating from this region, encompasses a broad range of properties and offers a snapshot into the complexities of real estate valuation influenced by various factors.

The primary objective of this report is to apply data mining techniques to understand and predict house prices in King County. This involves:

Exploratory Data Analysis (EDA):Conducting a comprehensive analysis to uncover patterns, anomalies, trends, and relationships within the data.

Predictive Modeling: Utilizing machine learning algorithms to develop a model that can accurately predict house prices based on multiple input features.

Evaluation: Assessing the performance of the predictive models to ensure their accuracy and reliability in real-world scenarios.

The dataset consists of 21,597 observations and 18 attributes related to house features such as area, number of bedrooms, bathrooms, floors, and location coordinates, along with the house prices. Special attributes like 'Grade' and 'View' are indicative of qualitative assessments specific to King County's grading system and the number of times a property has been viewed, respectively. These features will be used to explore the influence of various factors on house pricing.

1. Methodology

The analysis will be conducted through the following steps:

Data Preprocessing: Cleaning the data by handling missing values, outliers, and incorrect data types to prepare a robust dataset for analysis.

Feature Engineering: Enhancing the dataset with new features that could improve the model's predictive power, such as interactions between existing features and polynomial features.

Model Development: Building regression models to predict house prices and employing techniques such as cross-validation and regularization to optimize model performance.

2. Get data sample

The initial phase of our analysis involved importing essential Python libraries that are foundational for data manipulation, visualization, and analysis. These libraries include Pandas for data handling, NumPy for numerical operations, Matplotlib and Seaborn for data visualization, and category_encoders for managing categorical variables effectively.

Following the setup of our analytical environment, we proceeded to load the dataset titled ST83519_kc_house_data_.csv into a pandas DataFrame. This dataset comprises detailed records of house sales in King County, including Seattle, which is known for its dynamic real estate market. The comprehensive dataset, sourced from Kaggle under a public domain license, includes 21,519 observations and 18 attributes, providing a robust foundation for our predictive modeling endeavors.

Using the head() function, we examined the first few entries of the dataset to understand the structure and the type of data each column holds. This preliminary peek revealed several key attributes such as price, bedrooms, bathrooms, sqft_living, and sqft_lot, among others, which are integral to house valuation. Notably, the dataset includes both numerical and categorical data, such as waterfront and grade, which necessitate appropriate preprocessing to convert them into a machine-readable format.

Table 1 Dataset ST83519_kc_house_data_.csv

index	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	sqft_living15	sqft_lot15
0	7129300520	10/13/2014	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	1340	5650
1	6414100192	12/9/2014	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	1991	1690	7639
2	5631500400	2/25/2015	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	2720	8062
3	2487200875	12/9/2014	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	1360	5000
4	1954400510	2/18/2015	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	1800	7503

3. Data Structure Overview and Preliminary Data Assessment

The dataset comprises 21,597 records, each detailed with 18 attributes that encapsulate various aspects of housing in King County, including size, condition, and pricing attributes. A comprehensive review using pandas' info() function indicates a well-structured dataset with complete data across all fields—there are no missing values.

4. Inspect each column

Each column is inspected individually and in detail. An interesting observation is the (id) column, where some identifiers are repeated multiple times. This may indicate duplicate records. It has been decided to remove this column as it will not influence the model.

```
id
795000620     3
8910500150    2
7409700215    2
1995200200    2
9211500620    2
             ..
3649100387    1
2767603649    1
1446403617    1
5602000275    1
1523300157    1
Name: count, Length: 21420, dtype: int64
```

Figure 1 (id) column

Another point of interest is the bedrooms column, where one of the entries indicates that a building has 33 bedrooms. This is unusual and likely an error. Such anomalies require special attention and will be carefully investigated and corrected as necessary.

```
bedrooms
3     9824
4     6882
2     2760
5     1601
6      272
1      196
7       38
8       13
9        6
10       3
11       1
33       1
Name: count, dtype: int64
```

Figure 2 (bedrooms) column

5. Distribution

The distribution of house prices in King County provides valuable insights into the real estate market. As seen in the distribution plot (Figure 3), most properties are sold for less than $1 million. The mode, or the most frequent price range, is in the sub-$1 million category. This indicates that the bulk of the properties fall within a more affordable price range.

While there are properties that sell for significantly higher prices, such as those in the $4 million to $7.7 million range, these are rare and represent the tail end of the distribution. The highest recorded property price in the dataset is $7.7 million, but such high-priced properties are outliers.

This plot illustrates that although the market does contain high-value properties, the majority of transactions occur at lower price points. Properties over $1 million are relatively few, highlighting the concentration of sales in the more affordable segment of the market.

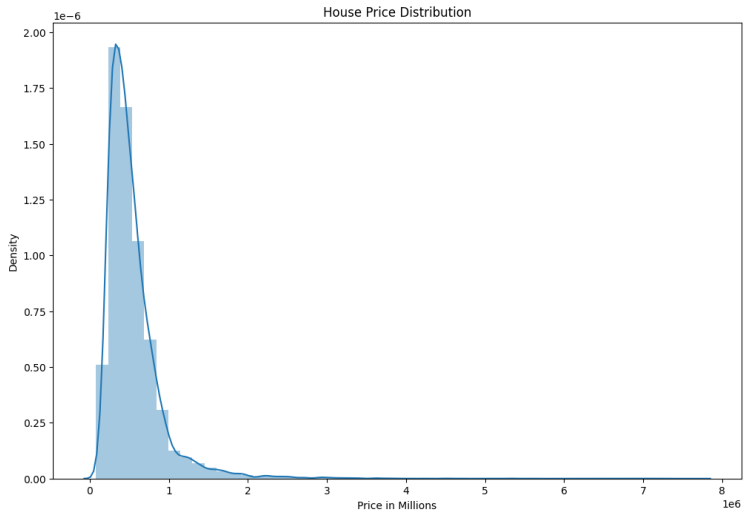

Figure 3 House Price Distribution

The distribution of the living area in square feet for houses in King County provides another layer of insight into the real estate market. As shown in the distribution plot (Figure 4), most properties have a living area below 4000 square feet, with the mode being around 2000 square feet. This indicates that the majority of homes are mid-sized, suitable for typical family living.

The plot reveals a right-skewed distribution, meaning there are some houses with significantly larger living areas, extending up to and beyond 10,000 square feet. These larger homes are outliers and constitute the tail end of the distribution. Despite the presence of these very large properties, the bulk of the houses are around the 2000 square foot mark.

This analysis helps to understand the common house sizes in the market and highlights that while there are a few very large properties, the majority of the housing stock consists of more moderately sized homes.

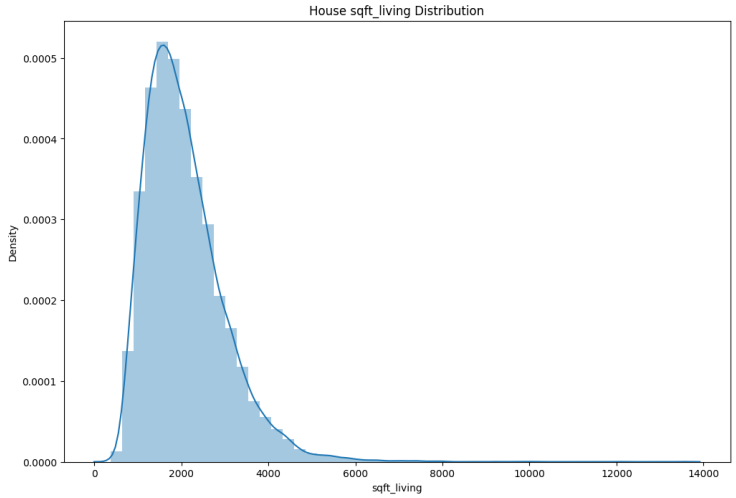

Figure 4 House Square Footage Distribution

The distribution of the number of bedrooms in King County houses reveals interesting patterns about the typical home size and its impact on pricing. As illustrated in the distribution plot (Figure 5), most properties have between 2 to 4 bedrooms, with 3-bedroom homes being the most common. This concentration suggests that the bulk of the housing market caters to small to medium-sized families.

The count plot shows that the occurrence of houses with a larger number of bedrooms drops off significantly after 4 bedrooms. Notably, there is an outlier with 33 bedrooms, which is highly unusual and likely represents a non-residential property such as a motel or a large group home.

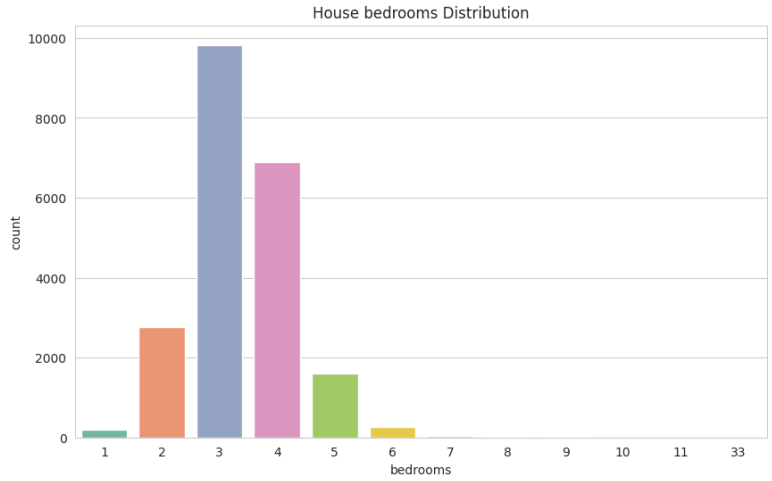

Figure 5 Figure 6. Distribution of Bedrooms

To explore the relationship between the number of bedrooms and house prices, a box plot was created (Figure 6). This plot shows a positive correlation between the number of bedrooms and the price, with median prices generally increasing as the number of bedrooms increases. However, this trend also highlights that more bedrooms do not always correspond to higher prices due to the influence of other factors such as location, condition, and overall size of the property.

Interestingly, the most expensive property in the dataset, priced at $7.7 million, has six bedrooms. This highlights that while there is a general trend of increasing price with more bedrooms, significant outliers exist where other features of the property play a crucial role in determining its value.

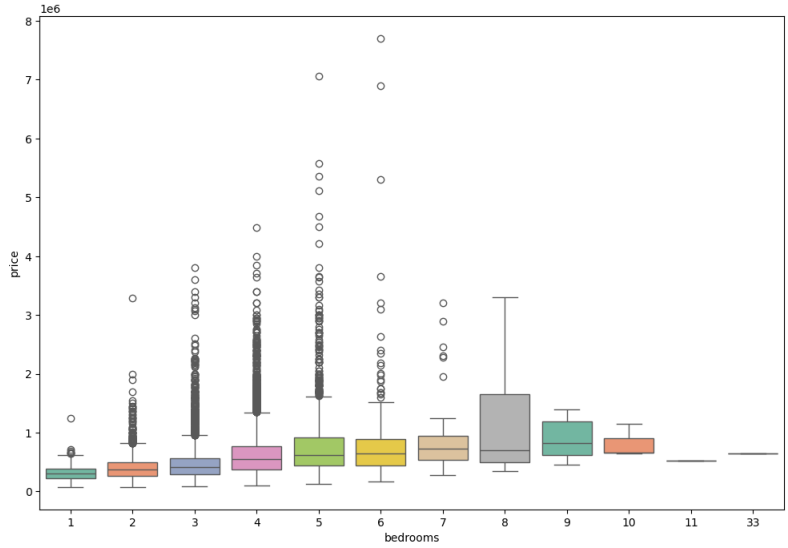

Figure 6 Box Plot of House Prices by Number of Bedrooms

The distribution of the number of bathrooms in King County houses offers additional insights into the characteristics of the housing market. As shown in the distribution plot (Figure 7), most properties have between 1 and 3 bathrooms, with 2.5 bathrooms being the most common. This

suggests that many homes are designed with multiple bathrooms to accommodate families and enhance convenience.

The count plot (Figure 7) reveals a bimodal distribution, indicating two distinct groups of properties. One group likely consists of smaller homes or apartments with fewer bathrooms, while the other group includes larger houses with more bathrooms.

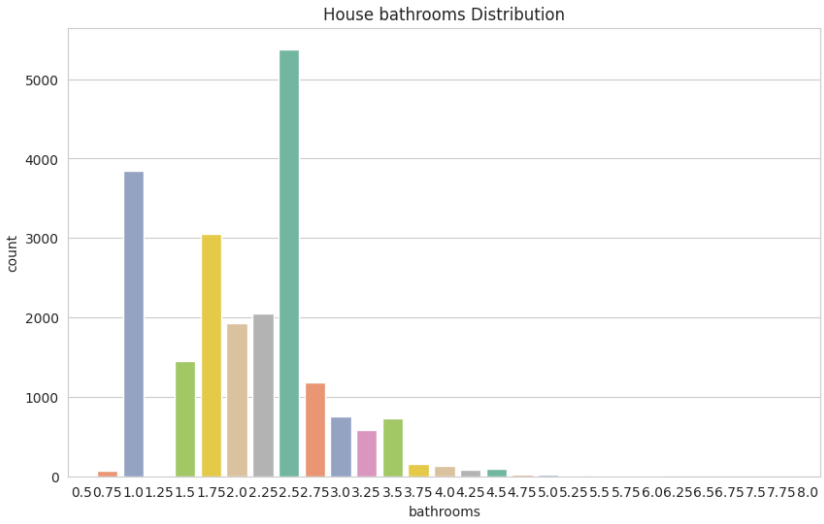

Figure 7 Distribution of Bathrooms

To explore the relationship between the number of bathrooms and house prices, a box plot was created (Figure 8). This plot demonstrates a positive correlation between the number of bathrooms and house prices, with median prices generally increasing as the number of bathrooms increases. However, similar to the trend observed with bedrooms, other factors such as overall size, location, and property condition also play significant roles in determining house prices.

Interestingly, the box plot shows that properties with the highest prices tend to have more bathrooms, but there are significant outliers. The most expensive property in the dataset, priced at $7.7 million, has six bathrooms. This emphasizes that while there is a general trend of increasing price with more bathrooms, high-value properties often have a combination of several desirable features.

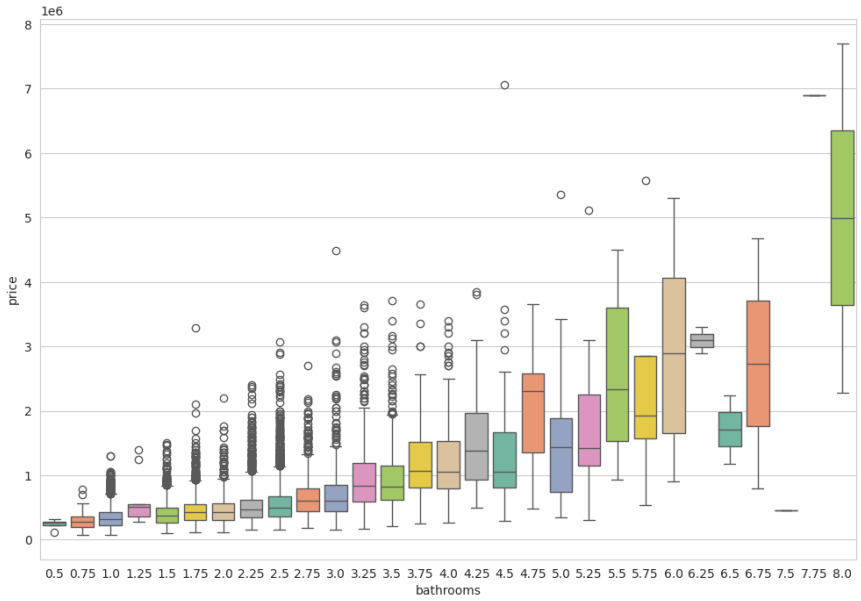

Figure 8 Box Plot of House Prices by Number of Bathroom

In King County, houses are graded on a scale from 1 to 13 based on their overall quality and condition. This grading system provides a standardized way to assess and compare properties. As illustrated in the distribution plot (Figure 9), the majority of houses have a grade around 7, with the mean grade being 7.6. This indicates that most properties are of average quality according to the county's grading criteria.

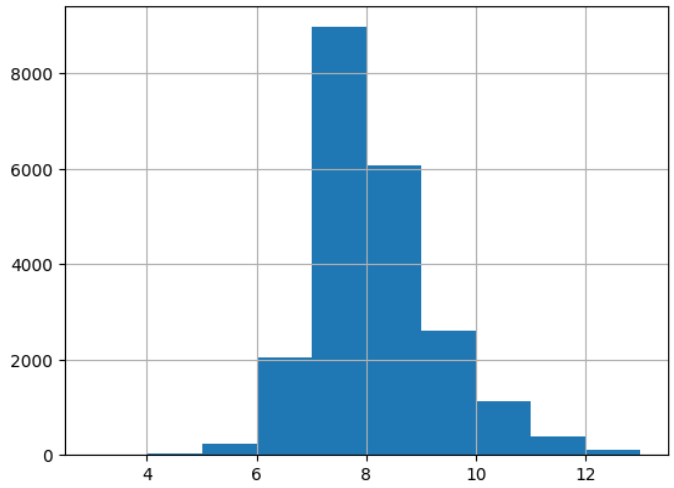

Figure 9 Distribution of House Grades

The box plot (Figure 10) shows the relationship between the grade of a house and its price. There is a clear positive correlation: higher grades are associated with higher property prices. As the grade increases from 3 to 13, the median house price also increases significantly. This trend is consistent across the range of grades, although there are notable outliers, particularly at the higher end of the grading scale.

The most expensive property in the dataset, priced at $7.7 million, has a grade of 13, which is the highest possible grade. This underscores the influence of the grading system on property values, as higher-grade properties tend to be more luxurious and desirable, thus commanding higher prices.

Adjusting the y-axis to a maximum value of $8 million allows for a better visualization of the entire range of property prices in relation to their grades, capturing all the significant data points and outliers.

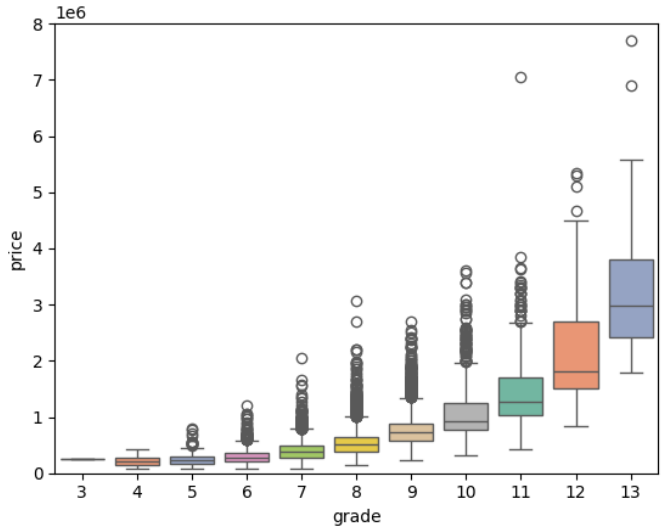

Figure 10 Box Plot of House Prices by Grade

The distribution of the years in which properties were built provides insights into the housing stock's historical trends and the economic factors that influenced construction activity in King County. The histogram (Figure 11) shows the number of properties built each year, with data spanning from the early 1900s to 2015.

Several distinct patterns emerge from the histogram. Firstly, there is a noticeable amount of legacy property dating back to the early 1900s. These older properties have often been renovated and modernized to meet current standards.

There is a significant drop in property construction during the 1930s, which likely reflects the impact of the Great Depression (1929-1939). The post Great Depression period saw a resurgence in property construction, which continued to grow steadily through the mid-20th century.

Another notable trend is the sharp increase in property construction during the housing boom of the late 1990s and early 2000s, peaking just before the 2007-2008 financial crisis. The data also captures a decline in construction activity immediately following the housing market crash, reflecting the economic downturn during that period.

In recent years, particularly around 2014-2015, there has been a resurgence in property construction, indicating a recovery in the real estate market.

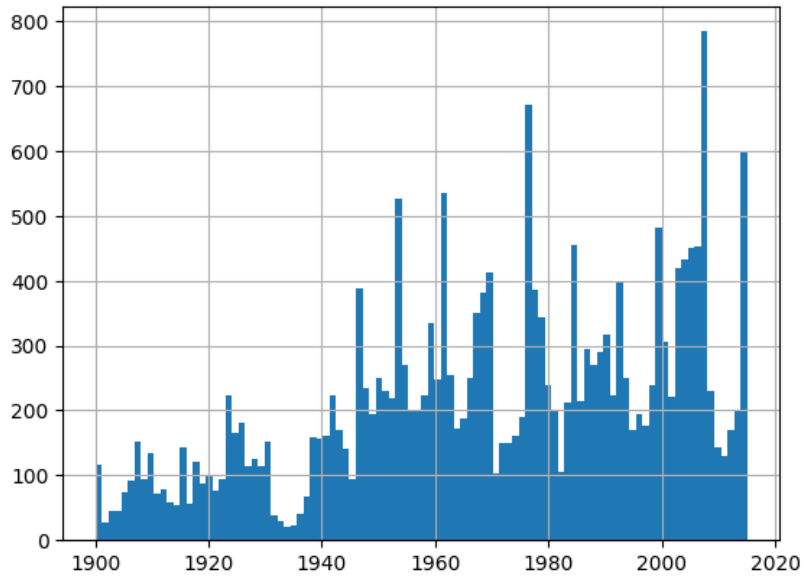

Figure 11 Distribution of Year Built for Properties

6. Enhancement of Dataset through Feature Engineering

The dataset underwent significant transformation to refine its utility for predictive modeling, focusing on the creation of new features that capture more nuanced aspects of the housing data. Initially, the (id) column, deemed unnecessary for analysis, was removed to streamline the dataset. Subsequently, the (date) field was manipulated to extract the (sales_year), which provides a discrete measure of time, pivotal for any time-series analysis.

To deepen the dataset's analytical value, two new features, (age_of_building) and (years_from_ren), were introduced. (age_of_building) is calculated by subtracting the construction year (yr_built) from the sales year, offering insights into the potential depreciation or appreciation of property value over time. The (years_from_ren) feature, initialized to zero and updated for properties with renovations (yr_renovated > 0), reflects the time elapsed since the last renovation, which could influence the property's market value.

The creation of these features required the removal of the original (yr_built) and (yr_renovated) columns post-calculation to avoid redundancy and potential multicollinearity in subsequent modeling. This refinement of the dataset not only aids in a more accurate evaluation of factors affecting house prices but also ensures the model's robustness by providing clearer, more direct variables for analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   sales_year       21597 non-null  int64
 1   price            21597 non-null  float64
 2   bedrooms         21597 non-null  int64
 3   bathrooms        21597 non-null  float64
 4   sqft_living      21597 non-null  int64
 5   sqft_lot         21597 non-null  int64
 6   floors           21597 non-null  float64
 7   waterfront       21597 non-null  int64
 8   view             21597 non-null  int64
 9   condition        21597 non-null  int64
 10  grade            21597 non-null  int64
 11  sqft_above       21597 non-null  int64
 12  sqft_basement    21597 non-null  int64
 13  sqft_living15    21597 non-null  int64
 14  sqft_lot15       21597 non-null  int64
 15  age_of_building  21597 non-null  int64
 16  years_from_ren   21597 non-null  int64
dtypes: float64(3), int64(14)
memory usage: 2.8 MB
```

Figure 12 Overview of Transformed Dataset

7. Define categorical and numerical variables

In the preprocessing phase, a deliberate division of the dataset was implemented to separate categorical and numerical variables, facilitating specialized handling and analysis for each data type.

A new DataFrame was constructed specifically for categorical variables, namely (waterfront), (view), and (condition). This separation is foundational for forthcoming tasks such as specific categorical data analyses and necessary transformations like encoding, which are vital for the accurate inclusion of these variables in predictive modeling.

Conversely, another DataFrame was exclusively dedicated to numerical variables—those identified by data types (int64) and (float64). This setup allows for a focused descriptive analysis using the (describe()) function, which provides essential statistical summaries for these variables. The descriptive output was rounded to minimize decimal places for clarity and transposed to facilitate a more intuitive evaluation of each variable's statistical characteristics.

Table 2 Overview of Transformed Dataset Categorical and Numerical Variables

index	count	mean	std	min	25%	50%	75%	max
sales_year	21597	2014.32	0.47	2014	2014	2014	2015	2015
price	21597	540296.57	367368.14	78000	322000	450000	645000	7700000
bedrooms	21597	3.37	0.93	1	3	3	4	33
bathrooms	21597	2.12	0.77	0.5	1.75	2.25	2.5	8
sqft_living	21597	2080.32	918.11	370	1430	1910	2550	13540
sqft_lot	21597	15099.41	41412.64	520	5040	7618	10685	1651359
floors	21597	1.49	0.54	1	1	1.5	2	3.5
grade	21597	7.66	1.17	3	7	7	8	13
sqft_above	21597	1788.6	827.76	370	1190	1560	2210	9410
sqft_basement	21597	291.73	442.67	0	0	0	560	4820
sqft_living15	21597	1986.62	685.23	399	1490	1840	2360	6210
sqft_lot15	21597	12758.28	27274.44	651	5100	7620	10083	871200
age_of_building	21597	43.32	29.38	-1	18	40	63	115
years_from_ren	21597	0.78	4.9	-1	0	0	0	80

This table 2 provides a snapshot of the transformed dataset, illustrating how the original features have been augmented to enhance predictive modeling.

8. Comprehensive Analysis of Numerical Variables in the Dataset

The dataset showcases a diverse array of variables, highlighting the multifaceted nature of property characteristics within King County. These variables range widely in both type and scale, from quantitative measurements such as square footage (sqft_living), (sqft_lot) to counts of rooms (bedrooms), (bathrooms). This diversity is crucial for constructing a nuanced understanding of real estate valuation but also poses challenges in data handling and model construction.

The price variable, central to our analysis, has an average value of approximately 540,000 USD, with a broad range extending from 78,000 USD to 7.7 USD million. This substantial variability underscores the heterogeneity of the housing market, reflecting everything from modest homes to luxury estates. Similarly, the (sqft_lot) size varies dramatically, from as little as 520 square feet to over 1.65 million square feet, suggesting a range from small urban plots to vast rural lands. Such extreme values, particularly evident in the upper ranges of price and (sqft_lot), indicate the presence of outliers that could skew further analysis and predictive modeling. The disproportionate maximum values, especially in (sqft_lot) as compared to its 75 %, highlight the necessity for careful outlier management.

Further complicating our analysis are anomalies such as negative values in (age_of_building) and (years_from_ren), which logically suggest data entry errors since negative ages are not feasible. These issues need addressing to ensure the integrity of subsequent analyses.

The statistical overview provided by the (describe()) function reinforces these observations, showing not only the central tendencies and dispersions but also the inconsistencies in measurement scales across different variables. Some variables are measured in hundreds or thousands, while others range from 0 to 10. This disparity necessitates the normalization or standardization of data to ensure uniformity in measurements, which is critical for effective statistical analysis and machine learning applications.

The descriptive statistics also further detail the price distribution, where the median value stands at 645,000 USD, reaffirming the significant spread and variability in housing prices across the dataset.

```
Q90 = data_num.quantile(0.9)
```

Figure 13 Code for Calculating the 90% Values

```
sales_year          2015.0
price             887000.0
bedrooms               4.0
bathrooms              3.0
sqft_living         3254.0
sqft_lot           21371.6
floors                 2.0
grade                  9.0
sqft_above          2950.0
sqft_basement        970.0
sqft_living15       2930.0
sqft_lot15         17822.0
age_of_building       89.0
years_from_ren         0.0
Name: 0.9, dtype: float64
```

Figure 14 Transposed 90% Values

9. Outlier Analysis and Categorical Data Description

Outlier Identification Using Percentiles

To ensure the robustness and accuracy of the statistical models, identifying and handling outliers in the numerical data is crucial. The analysis involved calculating the 90% for each numerical variable as a method to define upper thresholds and identify extreme values. The values at the 90% provide a reference point, beyond which data points may be considered outliers:

The 90th percentile is approximately 887,000 USD, suggesting that most homes are valued below this threshold, and values above it could be considered outliers in certain analyses.

For living area (sqft_living), the 90% is 3,254 sq ft, indicating large homes but not excessively so. For lot size (sqft_lot), the threshold is around 21,371 sq ft, highlighting substantial property sizes that may need special consideration due to their impact on the model.

Bedrooms and bathrooms have thresholds at 4 and 3 respectively, which are typical for larger family homes.

These thresholds will guide the process of outlier management, where entries exceeding these values might be removed or treated differently to prevent them from skewing the analysis.

Categorical Data Statistical Summary

The transposed descriptive statistics for categorical data — specifically view, condition, and waterfront — illuminate the distribution and frequency of categories:

Most properties (19,475 out of 21,597) have not been viewed, which could influence their market perception and eventual pricing.

A majority of properties are in "average" condition (rating 3), with 14,020 instances, suggesting that most properties do not require significant immediate investment in repairs or upgrades.

A significant majority of the properties (21,434 out of 21,597) do not have waterfront access, making waterfront properties highly unique and potentially more valuable.

10.Data visualization to detect the possible relationships, form and outliers

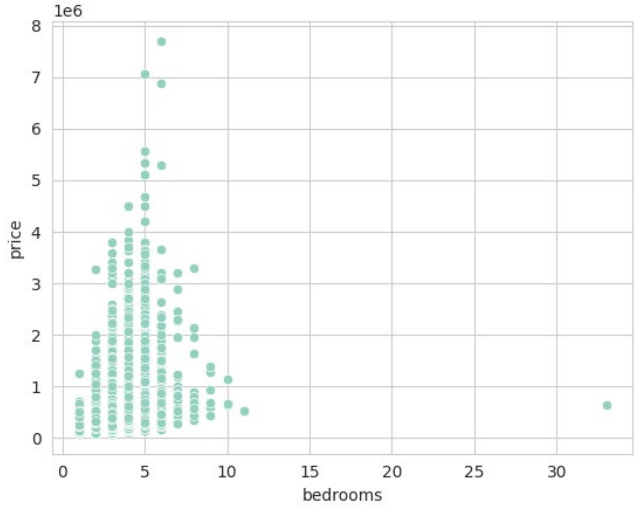

Figure 15 Relationship between the number of bedrooms and house prices

In the scatter plot (Figure 15) examining the relationship between the number of bedrooms and house prices, most data points cluster for houses with 2 to 10 bedrooms, with prices generally under 2 million USD. There are outliers, particularly for homes with more than 10 bedrooms, which might not accurately reflect typical market values and could influence statistical models inappropriately.

Specifically, the presence of a house with around 30 bedrooms priced much lower than others is unusual and might be an error. Such outliers could distort a linear regression model, which assumes a linear relationship and is sensitive to extreme values.

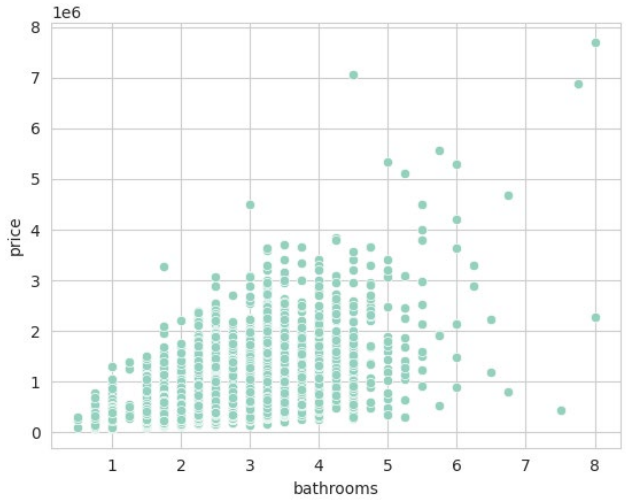

Figure 16 Relationship between the number of bathrooms in a house and its price

The scatter plot (Figure 16) illustrates the relationship between the number of bathrooms in a house and its price. Most homes, featuring 1 to 4 bathrooms, are priced under 3 million USD. As the number of bathrooms increases, there's a notable rise in price, suggesting a positive correlation.

However, we observe several outliers, especially at higher bathroom counts where fewer data points exist, and prices reach up to 8 million USD. These extreme values could affect predictive modeling such as linear regression, leading to skewed results.

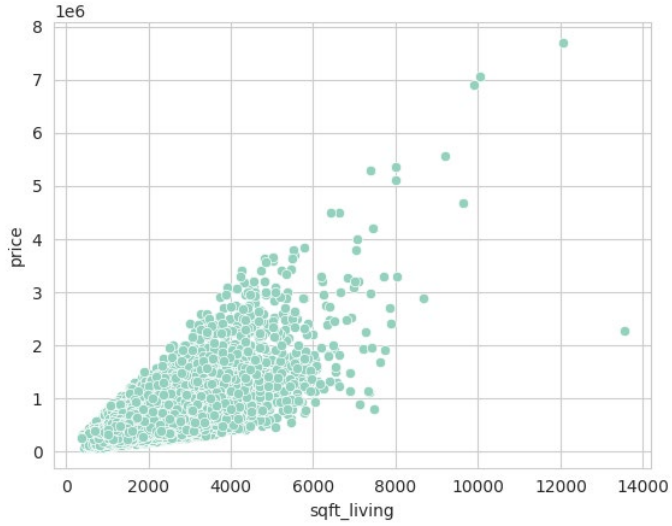

Figure 17 Relationship between living area size (in square feet) and house price

The scatter plot (Figure 17)depicting the relationship between living area size (in square feet) and house price shows that as the living area increases, there is a general trend of rising house prices. The dense clustering of homes with 1,000 to 4,000 square feet aligns with prices ranging from under 1 million USD to around 3 million USD, indicating this is the common market range.

However, there are noticeable outliers, particularly homes with large living areas (above 10,000 square feet) where the prices also escalate significantly, some reaching as high as 8 million USD. These outliers could influence the outcomes of linear regression models due to their leverage on the regression line.

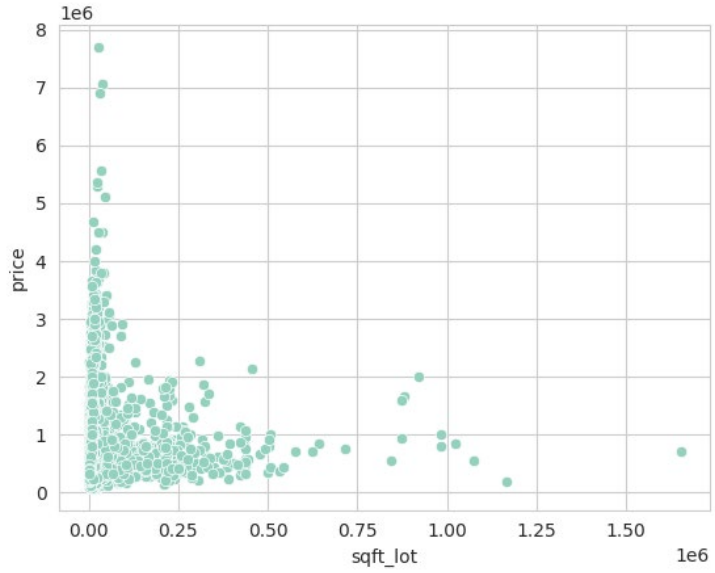

Figure 18 Relationship between lot size (in square feet) and house prices

The scatter plot (Figure 18) illustrates the relationship between lot size (in square feet) and house prices. Data is concentrated mainly at lot sizes under 0.25 million square feet and house prices below 3 million USD. This clustering suggests that in this range, lot size doesn't significantly impact house prices.

However, there are outliers with much larger lot sizes that don't necessarily correlate with higher house prices. In fact, some of these properties with expansive lots are priced lower than expected, suggesting that other variables besides lot size play a greater role in price determination at these outlier points.

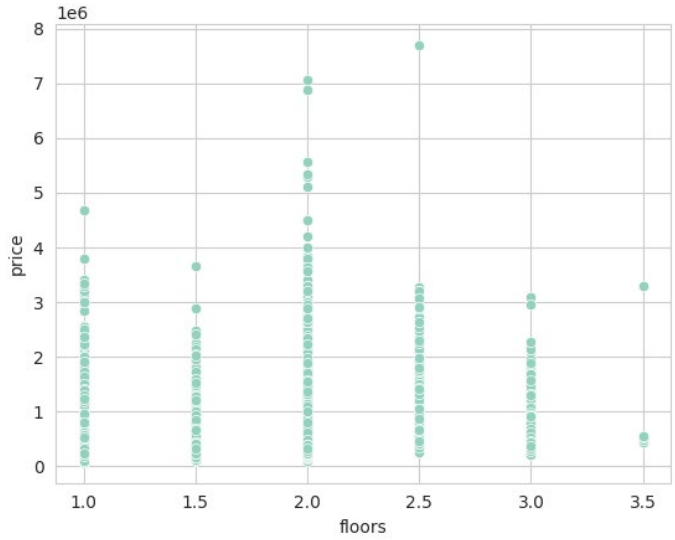

Figure 19 Relationship between houses floors and house prices

The scatter plot (Figure 19) shows that houses with 1 to 2 floors are the most common, with prices mostly under 2 million USD. As the number of floors increases to 3 or more, house prices also tend to increase, indicating that more floors often mean a higher property value. However, there are wide price ranges within each floor category, suggesting that while the number of floors is important, other factors also play a significant role in determining house prices. Some outliers with very high prices, especially for houses with 3 or more floors.

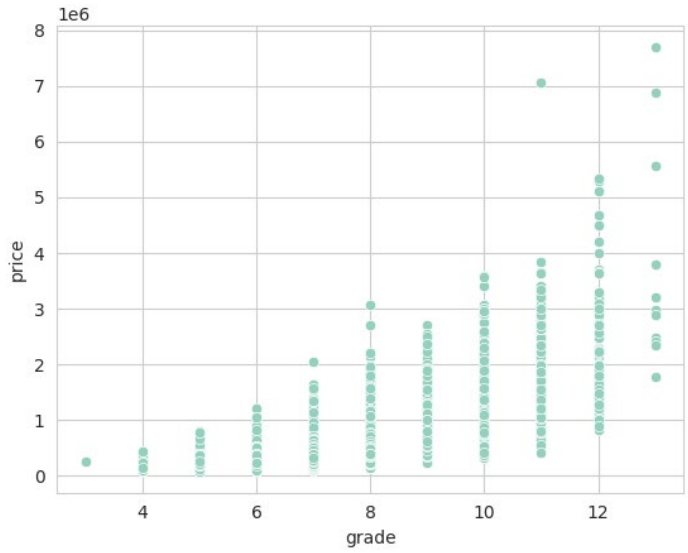

Figure 20 Relationship between house grades and prices

The scatter plot (Figure 20) illustrates the relationship between house grades and their prices. There is a clear upward trend, showing that as the grade of a house increases, the price generally rises. Houses with grades between 7 and 13 dominate the market, with prices clustering between 1 million USD and 3 million USD. Higher-grade houses, especially those graded above 10, tend to have significantly higher prices, reaching up to $8 million. This positive correlation suggests that the grade of a house, which likely reflects its quality and amenities, is a strong predictor of its market value. Outliers with exceptionally high prices at high grades should be further examined to ensure data accuracy and consistency. Overall, house grade is a crucial factor in determining price and should be included as a key variable in predictive models.

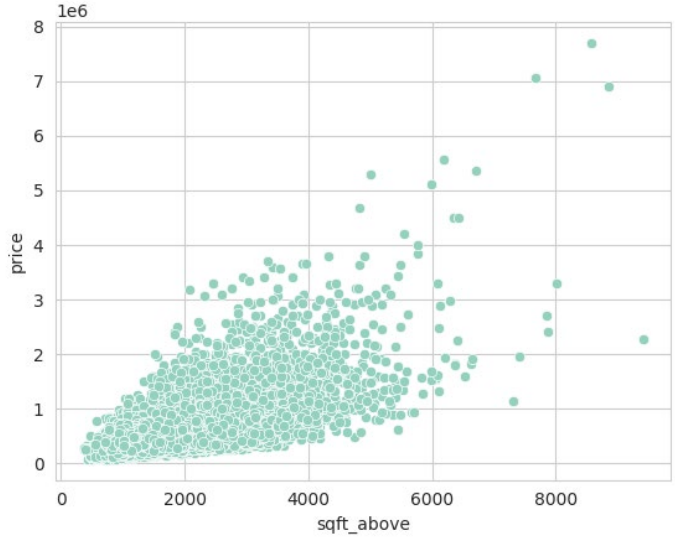

Figure 21 Relationship between above-ground and prices

The scatter plot (Figure 21) shows that as the above-ground living area increases, house prices generally rise. Most homes have between 1,000 and 4,000 square feet above ground, with prices mostly below 2 million USD. There are some outliers with larger areas, up to 8,000 square feet, and prices reaching 8 million USD. These outliers need to be reviewed as they can impact the analysis. The positive correlation between above-ground living space and price indicates that larger homes tend to be more expensive. This variable is crucial for predicting house prices accurately.

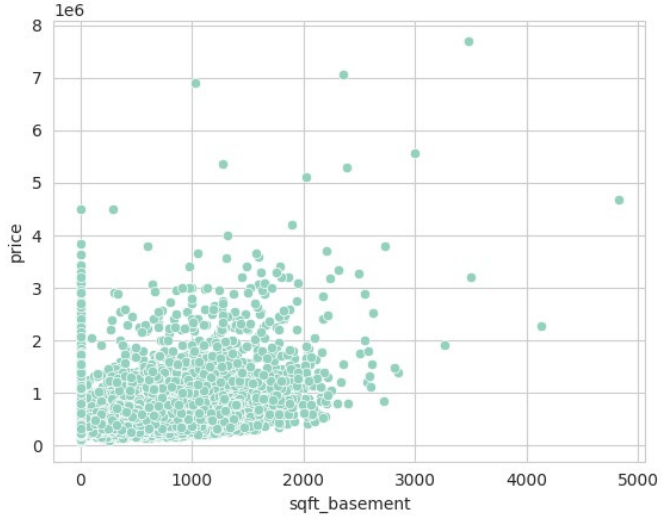

Figure 22 Relationship between basement square footage and house prices

The scatter plot (Figure 22) illustrates the relationship between basement square footage and house prices. Most homes have basements up to 2,000 square feet, with prices generally under 2 million USD. There are outliers with larger basements, reaching up to 5,000 square feet, and prices as high as 8 million USD. These outliers should be reviewed as they can skew the analysis. While there is some correlation between larger basements and higher prices, it is not as strong as with above-ground living space.

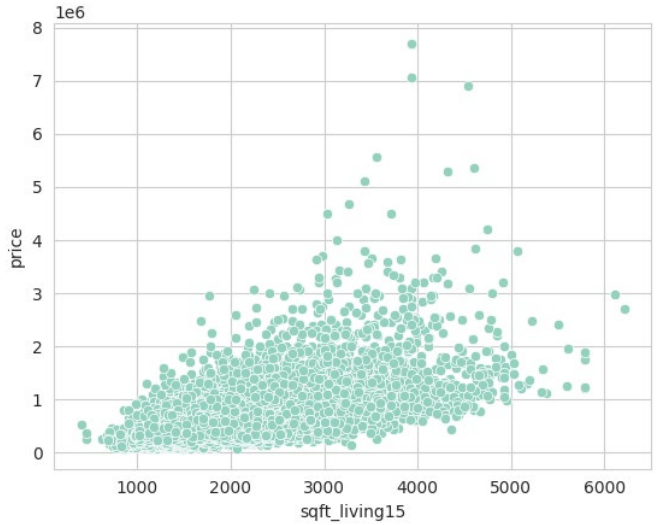

Figure 23 Relationship between living area (sqft_living15) and house prices

The scatter plot (Figure 23) depicts the relationship between the recent living area (sqft_living15) and house prices. There is a clear positive correlation, showing that as the living area increases, the price generally rises. Most houses have a living area between 1,000 and 4,000 square feet, with prices clustering below 2 million USD. There are some outliers with larger living areas and higher prices, reaching up to 8 million USD. These outliers can significantly impact the analysis and should be reviewed for accuracy. The positive trend indicates that sqft_living15 is an important factor in determining house prices and should be included in predictive models for better accuracy.

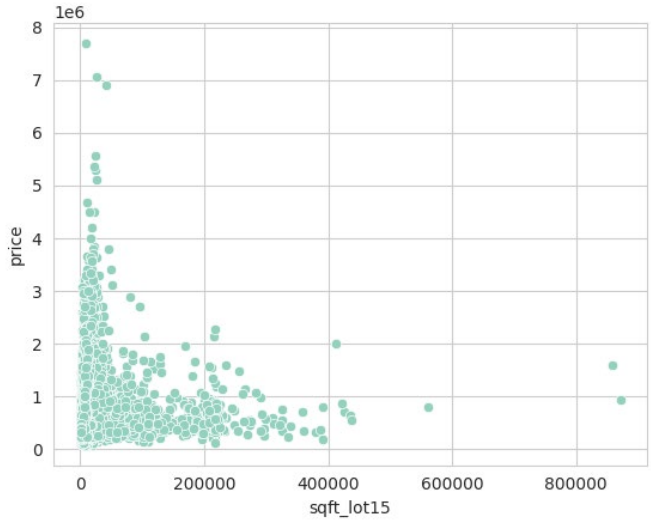

Figure 24 Relationship between lot size (sqft_lot15) and house prices

The scatter plot (Figure 24) shows the relationship between lot size (sqft_lot15) and house prices. Most houses have lot sizes under 200,000 square feet, with prices generally below 2 million USD. There is a clustering of data points in this range, indicating that larger lot sizes do not necessarily

correspond to higher prices. Some outliers have very large lot sizes and higher prices, reaching up to 8 million USD, which should be reviewed for accuracy as they can affect the analysis. The lack of a strong correlation between lot size and price suggests that other factors may play a more significant role in determining house prices.

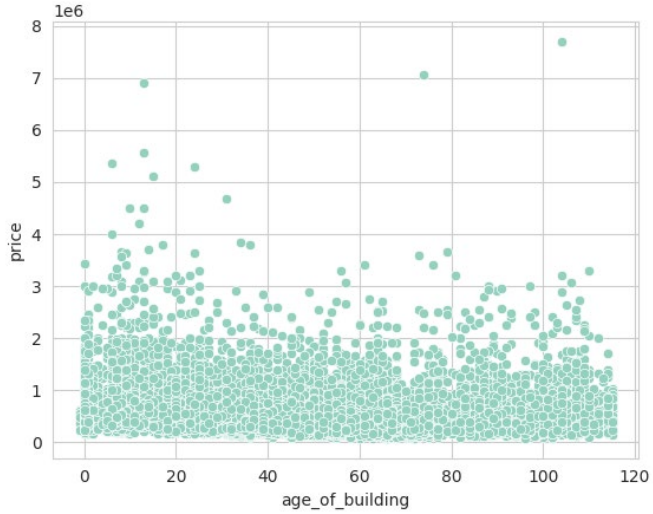

Figure 25 Relationship between the age of buildings and house prices

The scatter plot (Figure 25) shows the relationship between the age of buildings and house prices. The data indicates that house prices do not strongly correlate with the age of the building. Most house prices, regardless of age, cluster below 2 million USD. There are some outliers, especially with very high prices, but these are not concentrated in any specific age range.

The lack of a strong trend suggests that the age of a building is not a primary factor in determining house prices. However, the outliers with exceptionally high prices should be reviewed for accuracy as they can affect the overall analysis. Including the age of the building as a variable in predictive models can still be useful, but it should be considered alongside other more influential factors.

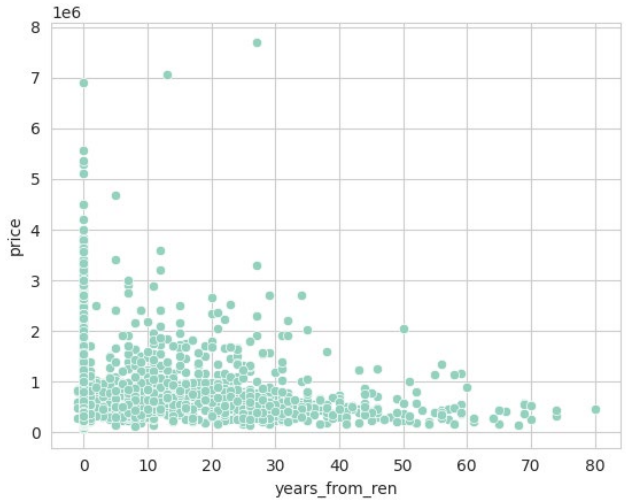

Figure 26 Relationship between the number of years since a house was last renovated (years_from_ren) and price

The scatter plot (Figure 26) illustrates the relationship between the number of years since a house was last renovated (years_from_ren) and its price. The plot shows that houses renovated more recently tend to have higher prices, with a clustering of prices below 2 million USD for homes

renovated in the last 20 years. There are several outliers with high prices, particularly for homes renovated within the last 5 years, reaching up to 8 million USD.

As the number of years since renovation increases, house prices generally decrease, and there are fewer high-price outliers. This trend suggests that recent renovations positively impact house prices. Therefore, years_from_ren is an important variable to include in predictive models for better accuracy in estimating house prices.

11.Cleaning from outliers

Cleaning the dataset from outliers is an essential step to improve the accuracy and robustness of predictive modeling. The numeric columns identified for outlier cleaning include: sales_year, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, grade, sqft_above, sqft_basement, sqft_living15, sqft_lot15, age_of_building, and years_from_ren (Figure 27)

```
Index(['sales_year', 'price', 'bedrooms', 'bathrooms', 'sqft_living',
       'sqft_lot', 'floors', 'grade', 'sqft_above', 'sqft_basement',
       'sqft_living15', 'sqft_lot15', 'age_of_building', 'years_from_ren'],
      dtype='object')
```

Figure 27 data_num.column

To enhance the accuracy and reliability of the predictive model, the dataset underwent a rigorous outlier cleaning process. Initially, numeric columns such as price, bedrooms, bathrooms, sqft_living, sqft_lot, and others were identified (Figure 28).

```
sales_year          2015.0
price             887000.0
bedrooms               4.0
bathrooms              3.0
sqft_living         3254.0
sqft_lot           21371.6
floors                 2.0
grade                  9.0
sqft_above          2950.0
sqft_basement        970.0
sqft_living15       2930.0
sqft_lot15         17822.0
age_of_building       89.0
years_from_ren         0.0
Name: 0.9, dtype: float64
```

Figure 28 q=data_num.quantile(0.90)

To ensure the accuracy and robustness of the predictive modeling, an extensive outlier cleaning process was undertaken. Initially, the box plot visualization (Figure 29) of the numeric columns such as price, bedrooms, bathrooms, sqft_living, sqft_lot, and others highlighted the presence of outliers. Calculating the 90% for these columns established thresholds beyond which values were considered outliers (Figure 30). For example, properties with prices exceeding $887,000 or lot sizes greater than 21,371.6 square feet were identified as outliers. Subsequently, these extreme values were removed, resulting in a refined dataset. The effectiveness of this process is demonstrated by the updated box plot visualizations (Figures 31 and 32), which show a more

consistent data distribution after removing outliers for columns like sqft_lot and sqft_lot15. This cleaning process enhances the dataset's representativeness, thereby improving the predictive model's reliability and performance.

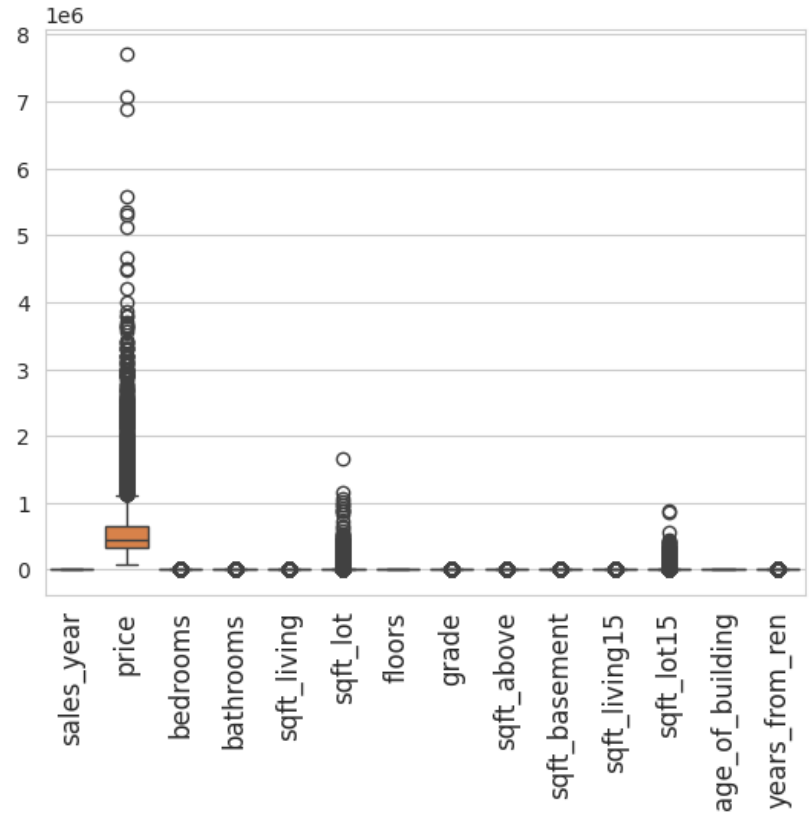

Figure 29 Initial Box Plot Visualization of Numeric Columns

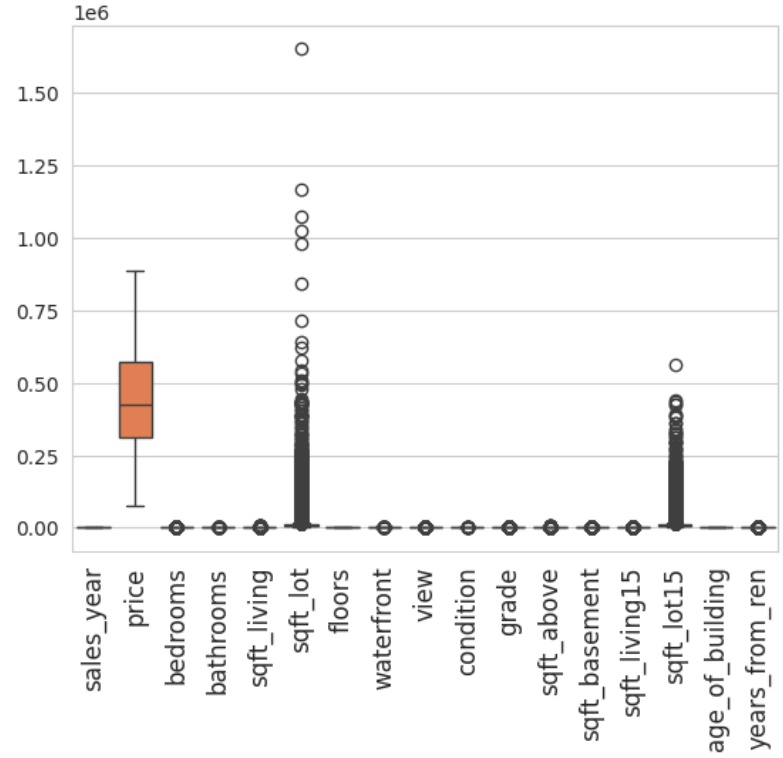

Figure 30 90% Thresholds for Outlier Detection

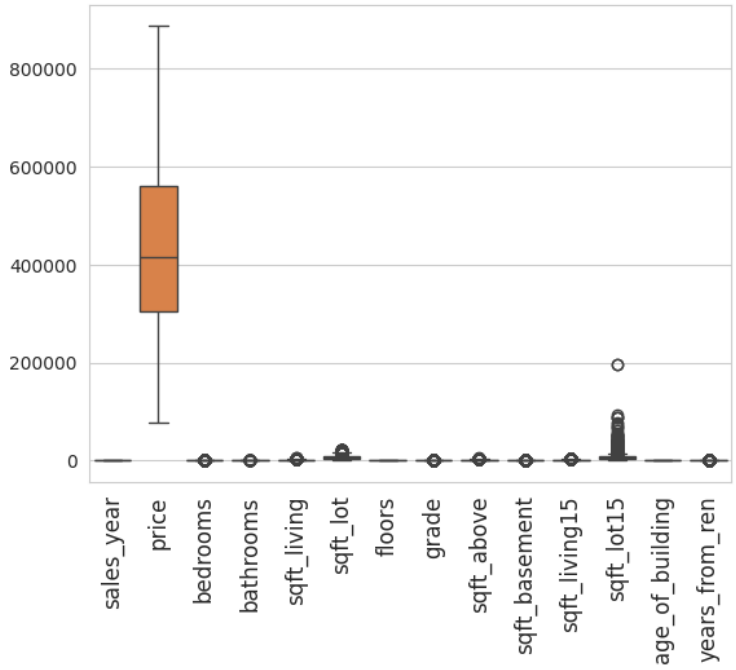

Figure 31 Box Plot Visualization After Removing Outliers from sqft_lot

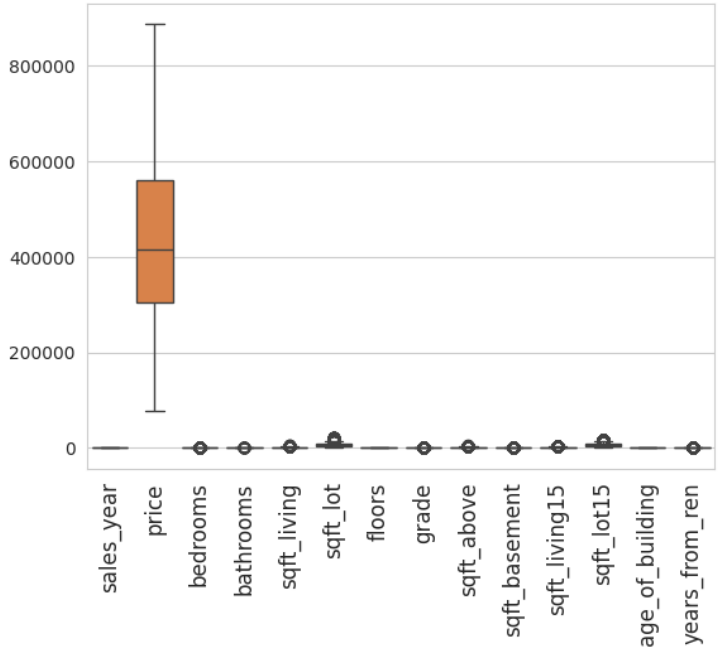

Figure 32 Box Plot Visualization After Removing Outliers from sqft_lot15

The box plot (Figure 33) illustrates the distribution of house prices after the outlier removal process. The cleaned data shows that the majority of house prices now fall within a more reasonable range, centered around the median value of approximately $400,000. The whiskers extend from around $200,000 to just below $900,000, indicating a significant reduction in the spread of price values. This suggests that the data is now more consistent and free from extreme outliers, making it more suitable for predictive modeling.

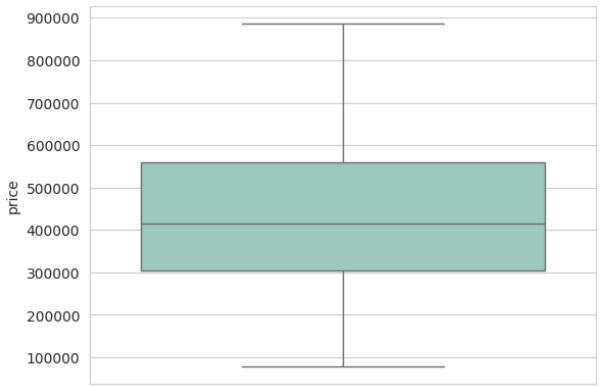

Figure 33 Box Plot of Price

The box plot (Figure 34) displays the distribution of lot sizes (sqft_lot) after cleaning the outliers. The interquartile range (IQR) is now more compact, ranging from about 5,000 to 10,000 square feet. The whiskers extend to approximately 20,000 square feet, with only a few data points beyond this range. This refined distribution indicates that the extreme lot sizes have been effectively removed, resulting in a dataset that better represents typical properties in the area and enhances the reliability of the model.

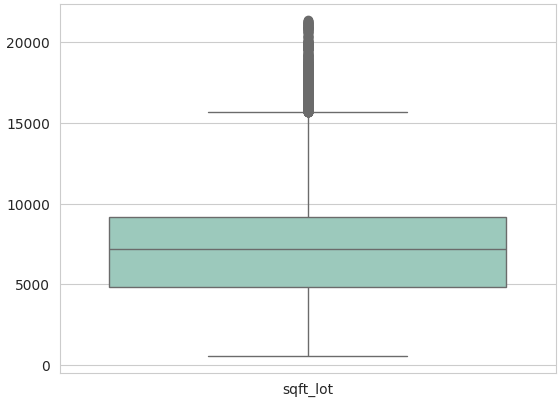

Figure 34 Box Plot of Sqft_Lot

The box plot (Figure 35) represents the distribution of lot sizes for the nearest 15 neighbors (sqft_lot15) after the removal of outliers. The median value is around 7,500 square feet, with the IQR ranging from about 5,000 to 10,000 square feet. The whiskers extend up to around 17,500 square feet, showing a significant reduction in the presence of extreme values. This adjustment ensures that the dataset now accurately reflects typical neighborhood lot sizes, which is crucial for understanding and predicting house prices based on the surrounding properties.

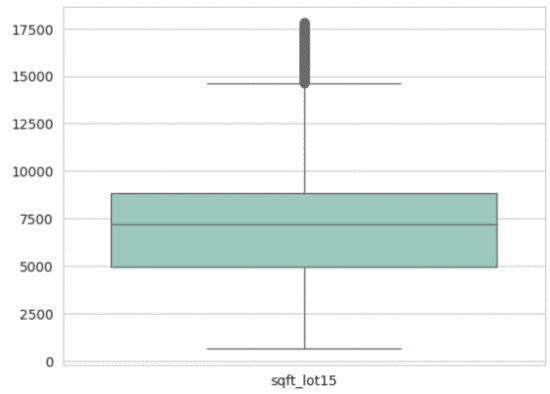

Figure 35 Box Plot of Sqft_Lot15

12. The Correlation Matrix

The correlation matrix (Figure 36) provides a detailed insight into the relationships between various features in the dataset. The analysis reveals that the price has the strongest positive correlation with sqft_living (0.55), indicating that larger living spaces significantly contribute to higher property values. Other notable correlations include grade (0.56) and sqft_above (0.53), highlighting the importance of quality and above-ground living area. Additionally, bedrooms and bathrooms exhibit a strong correlation with each other (0.48) and with sqft_living (0.61 and 0.70, respectively), suggesting that larger homes tend to have more bedrooms and bathrooms. The sqft_lot shows a high correlation with sqft_lot15 (0.87), reflecting the similarity in lot sizes within neighborhoods. However, its direct impact on price is minimal (0.05). The age_of_building displays negative correlations with key features like sqft_living (-0.35) and grade (-0.51), indicating that older buildings generally have smaller living spaces and lower quality.

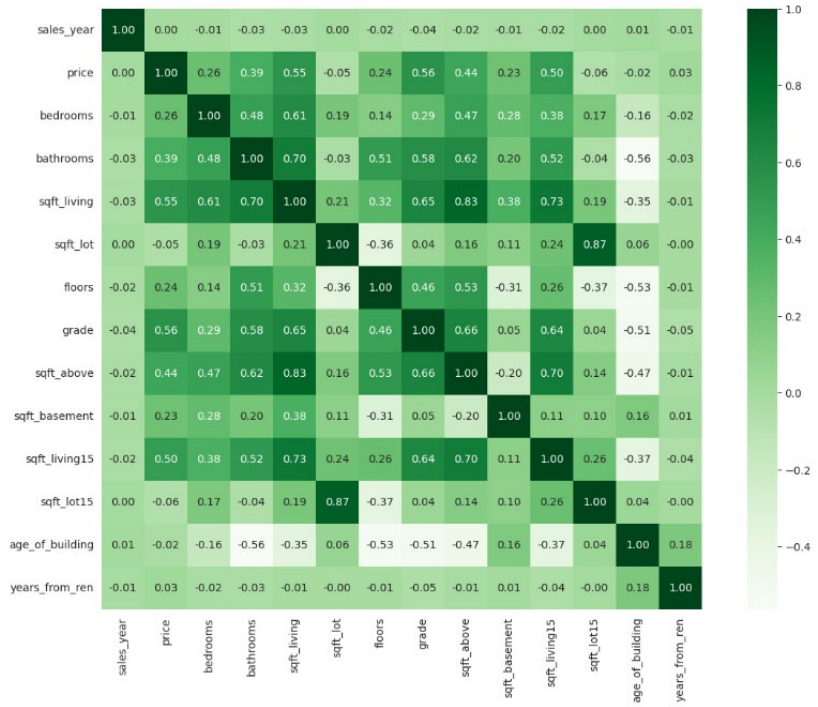

Figure 36 Correlation matrix

13.Feature Engineering and Data Transformation

To enhance the dataset's utility for predictive modeling, categorical variables were transformed into a machine-readable format through one-hot encoding. Initially, the categorical columns (view), (condition), and (waterfront) were identified. Each categorical column was then converted into multiple binary columns using the (pd.get_dummies) function. This process resulted in new columns representing each category as a separate boolean variable. The transformed categorical variables were concatenated back into the main dataframe, replacing the original categorical columns.

This transformation aimed to convert categorical data into numerical form, which is essential for most machine learning algorithms. The resulting dataset, now comprising 26 columns, includes both the original numerical features and the newly created binary columns for categorical data.

Table 3 Overview of Final Cleaned and Transformed Dataset

index	sales_year	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	grade	sqft_above	sqft_basement	sqft_living15	sqft_lot15	age_of_building	years_from_ren	View0	View1	View2	View3	View4	Condition0
0	2014	221900	3	1	1180	5650	1	7	1180	0	1340	5650	59	0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
1	2014	538000	3	2.25	2570	7242	2	7	2170	400	1690	7639	63	23	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
2	2015	180000	2	1	770	10000	1	6	770	0	2720	8062	82	0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
3	2014	604000	4	3	1960	5000	1	7	1050	910	1360	5000	49	0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
4	2015	510000	3	2	1680	8080	1	8	1680	0	1800	7503	28	0	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE

14.Data Preparation for Modeling

To prepare the dataset for predictive modeling, the target variable price was separated from the feature set. The target variable, denoted as Y, contains the house (prices), while the feature set, denoted as X, includes all other attributes. The dataset was then split into training and testing sets using the (train_test_split) function from the (sklearn.model_selection module). The data was divided into 80% for training and 20% for testing, ensuring that the model could be trained on a substantial portion of the data while reserving a portion for evaluation. The random state was set to 10 to ensure reproducibility of the results.

15.Training a Linear Regression Model

To predict house prices, a linear regression model was employed. First, the necessary libraries were imported, including LinearRegression from sklearn.linear_model, and various metrics from sklearn to evaluate model performance. An instance of the LinearRegression model was created and assigned to the variable lm. The model was then trained using the training data by fitting lm with X_train and y_train. This process involves finding the best-fitting line through the training data, allowing the model to learn the relationship between the features and the target variable, price. The trained model can now be used to make predictions on new, unseen data.

16.Analyzing Model Coefficients

After training the linear regression model, the coefficients were extracted to understand the impact of each feature on the house prices (Table 4). The coefficients represent the change in the target variable, price, for a one-unit change in the feature while keeping other features constant. The coefficients provide a numerical insight into how each feature influences the predicted price. These coefficients were organized into a DataFrame for better readability and interpretation. Notable features such as (bedrooms), (bathrooms), (sqft_living), and (grade) exhibit significant coefficients, indicating their strong influence on house prices. This analysis helps in understanding the relative importance of each feature in the model.

Table 4 Organizing Coefficients into a DataFrame for Analysis

index	Coeffecient
sales_year	11524.47775
bedrooms	-8297.114458
bathrooms	19262.38247
sqft_living	44.36406111
sqft_lot	-2.813199801
floors	22728.26412
grade	85770.75666
sqft_above	9.90951862
sqft_basement	34.45454248
sqft_living15	66.2550883
sqft_lot15	-7.235224612
age_of_building	2425.093348
years_from_ren	-411.5287208
View0	-41413.84939
View1	9858.144491
View2	-11819.78184
View3	-8956.69303
View4	52332.17976
Condition0	-52758.23101
Condition1	-31943.93301
Condition2	10648.71725
Condition3	24995.74276
Condition4	49057.70402
waterfront0	-70579.59916
waterfront1	70579.59916

17.Predicting Test Data

The performance of the linear regression model (Figure 37) was evaluated by comparing the predicted house prices against the actual prices in the test dataset. The scatter plot above shows the relationship between the actual house prices (Y Test) and the predicted house prices (Predicted Y). The diagonal green dashed line represents the ideal scenario where the predicted values perfectly match the actual values. The clustering of points around this line indicates that the model predictions are reasonably accurate, although some deviation exists, particularly at higher price ranges.

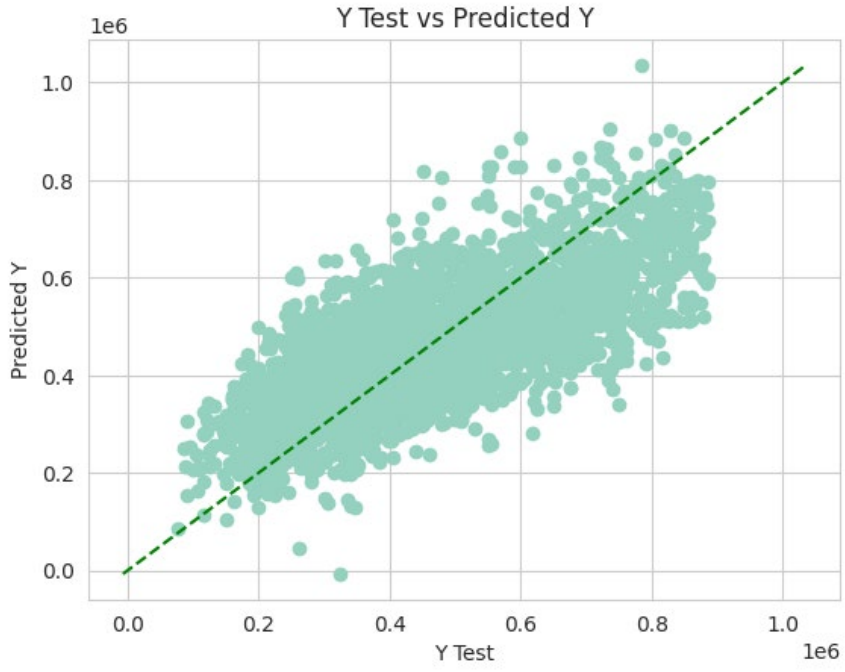

Figure 37 Linear regression model

18.Evaluating the Model

The effectiveness of the linear regression model (Table 5) was quantitatively assessed using several key metrics. The Mean Absolute Error (MAE) of the model is approximately 95,964 USD, indicating the average absolute difference between the predicted and actual house prices. The Mean Squared Error (MSE) is around 14,110,488,591, reflecting the average of the squares of the errors, and the Root Mean Squared Error (RMSE) is approximately 118,788, providing a measure of the standard deviation of the prediction errors. The R-squared value, which explains the proportion of variance in the dependent variable that is predictable from the independent variables, stands at 0.547. This suggests that about 54.7% of the variability in house prices can be explained by the model. These metrics indicate a moderate level of accuracy, with room for further optimization to improve predictive performance.

Table 5 The effectiveness of the linear regression model

MAE	95964.063290311
MSE	14110488591.4947
RMSE	118787.577597553
R squared Error	0.546772127108689

19.Residual Analysis

The distribution of residuals is depicted in the density plot (Figure 38). Ideally, residuals should be normally distributed around zero, indicating that the model's predictions are unbiased and errors are randomly distributed. In this plot, it is observed that the residuals exhibit a roughly normal distribution centered around zero, although there are some deviations, particularly in the tails. This suggests that while the model performs reasonably well, there may be some instances where it underestimates or overestimates the house prices.

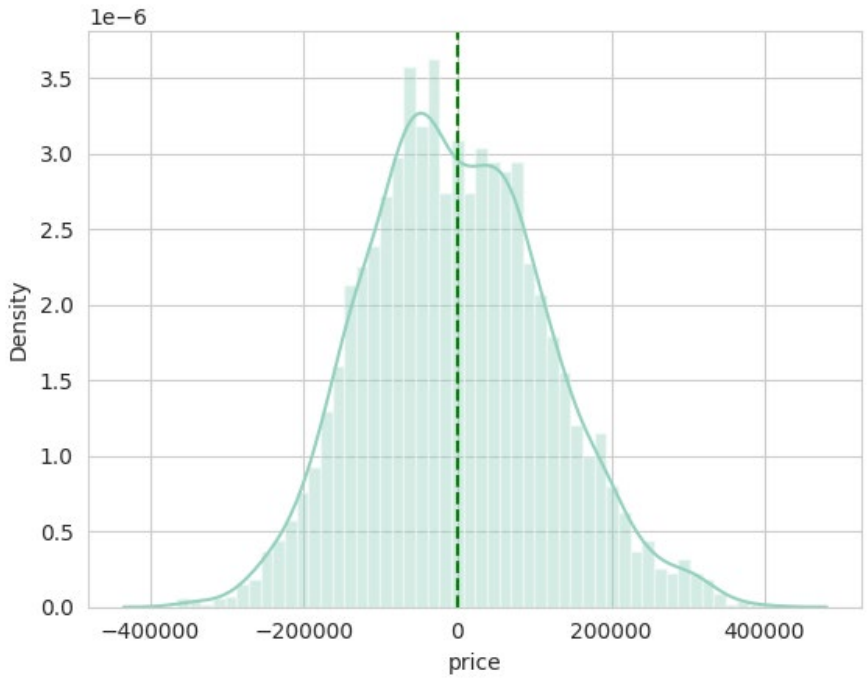

Figure 38 Distribution of Residuals

20.Regression Model creation by means of statsmodels

To further validate the regression model, an Ordinary Least Squares (OLS) regression was implemented using the Statsmodels library. Initially, a constant term was added to both the training and test datasets to account for the intercept in the regression equation. The OLS regression model was then instantiated with the training data, linking the house prices (dependent variable) to the predictors (independent variables).

The Ordinary Least Squares (OLS) regression results (Figure 39) provide an in-depth statistical analysis of the model's performance and the significance of each predictor variable. The R-squared value is 0.535, indicating that approximately 53.5% of the variance in house prices is explained by the model. The Adjusted R-squared is slightly lower at 0.534, accounting for the number of predictors used. Significant predictors include sales_year, bathrooms, sqft_living, floors, and grade, among others, as evidenced by their low p-values (<0.05). The model's F-statistic is 536.4, with a corresponding p-value of 0.000, suggesting the overall significance of the model.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.535
Model:	OLS	Adj. R-squared:	0.534
Method:	Least Squares	F-statistic:	761.1
Date:	Mon, 10 Jun 2024	Prob (F-statistic):	0.00
Time:	08:14:12	Log-Likelihood:	-1.8262e+05
No. Observations:	13926	AIC:	3.653e+05
Df Residuals:	13904	BIC:	3.654e+05
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.243e+07	2.31e+06	-5.380	0.000	-1.7e+07	-7.9e+06
sales_year	1.152e+04	2179.257	5.288	0.000	7252.840	1.58e+04
bedrooms	-8297.1145	1451.351	-5.717	0.000	-1.11e+04	-5452.272
bathrooms	1.926e+04	2544.110	7.571	0.000	1.43e+04	2.42e+04
sqft_living	44.3637	2.024	21.918	0.000	40.396	48.331
sqft_lot	-2.8132	0.583	-4.822	0.000	-3.957	-1.670
floors	2.273e+04	3032.804	7.494	0.000	1.68e+04	2.87e+04
grade	8.577e+04	1694.018	50.632	0.000	8.25e+04	8.91e+04
sqft_above	9.9099	1.962	5.051	0.000	6.064	13.755
sqft_basement	34.4549	2.199	15.670	0.000	30.145	38.765
sqft_living15	66.2551	3.043	21.772	0.000	60.290	72.220
sqft_lot15	-7.2352	0.684	-10.579	0.000	-8.576	-5.895
age_of_building	2425.0933	50.073	48.431	0.000	2326.943	2523.243
years_from_ren	-411.5287	229.417	-1.794	0.073	-861.217	38.159
View0	-2.528e+06	4.62e+05	-5.470	0.000	-3.43e+06	-1.62e+06
View1	-2.477e+06	4.62e+05	-5.357	0.000	-3.38e+06	-1.57e+06
View2	-2.498e+06	4.62e+05	-5.405	0.000	-3.4e+06	-1.59e+06
View3	-2.495e+06	4.62e+05	-5.398	0.000	-3.4e+06	-1.59e+06
View4	-2.434e+06	4.62e+05	-5.266	0.000	-3.34e+06	-1.53e+06
Condition0	-2.539e+06	4.63e+05	-5.489	0.000	-3.45e+06	-1.63e+06
Condition1	-2.518e+06	4.62e+05	-5.446	0.000	-3.42e+06	-1.61e+06
Condition2	-2.476e+06	4.62e+05	-5.356	0.000	-3.38e+06	-1.57e+06
Condition3	-2.461e+06	4.62e+05	-5.326	0.000	-3.37e+06	-1.56e+06
Condition4	-2.437e+06	4.62e+05	-5.275	0.000	-3.34e+06	-1.53e+06
waterfront0	-6.287e+06	1.16e+06	-5.442	0.000	-8.55e+06	-4.02e+06
waterfront1	-6.146e+06	1.16e+06	-5.318	0.000	-8.41e+06	-3.88e+06

Omnibus:	197.521	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	218.331
Skew:	0.258	Prob(JB):	3.89e-48
Kurtosis:	3.332	Cond. No.	5.50e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 6.18e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 39 Detailed OLS Regression Results

21.Scaling the Data

To prepare the dataset (Table 6) for further modeling and analysis, the numeric features were scaled using the StandardScaler from the sklearn library. This step ensures that each feature contributes equally to the model's performance by normalizing the data to have a mean of 0 and a standard deviation of 1. Initially, the numeric columns were identified and separated from the boolean columns. The StandardScaler was then applied to the numeric columns, transforming them accordingly. The scaled numeric data was combined back with the boolean columns to form a complete, rescaled dataset.

Table 6 new_data after scaling

index	sales_year	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	grade	sqft_above	sqft_basement	sqft_living15	sqft_lot15	age_of_building	years_from_ren	View0	View1	View2	View3	View4	Condition0
0	-0.694106716	-1.259374836	-0.319558624	-1.45272002	-0.996828655	-0.438966295	-0.853230626	-0.437006482	-0.663551312	-0.657517569	-0.893189027	-0.450151178	0.49575992	-0.136439551	1	0	0	0	0	0
1	-0.694106716	0.53724057	-0.319558624	0.370380035	1.040238848	0.008326789	0.99301409	-0.437006482	0.875131853	0.380986366	-0.254544588	0.195868261	0.630396774	4.881307605	1	0	0	0	0	0
2	1.440700654	-1.497521579	-1.423742133	-1.45272002	-1.597690293	0.783222722	-0.853230626	-1.523003463	-1.300783733	-0.657517569	1.624894762	0.33325701	1.269921831	-0.136439551	1	0	0	0	0	0
3	-0.694106716	0.912364318	0.784624886	1.464240069	0.146273972	-0.62159224	-0.853230626	-0.437006482	-0.865600616	1.705078884	-0.856695059	-0.661268642	0.159167785	-0.136439551	1	0	0	0	0	0
4	1.440700654	0.378097162	-0.319558624	0.005760024	-0.264070561	0.243773777	-0.853230626	0.648990498	0.113561398	-0.657517569	-0.053827764	0.151695991	-0.547675699	-0.136439551	1	0	0	0	0	0
6	1.440700654	-0.8618005	-0.319558624	-0.723479998	-1.172690598	0.702024417	-0.853230626	-0.437006482	-0.850058362	-0.657517569	-0.327532524	0.868845776	0.260145425	-0.136439551	1	0	0	0	0	0
7	1.440700654	-1.216178768	-0.319558624	-1.45272002	-0.117518942	0.072386351	-0.853230626	-0.437006482	-0.865600616	1.237752113	-0.090321732	0.349821611	0.361123066	-0.136439551	1	0	0	0	0	0
8	1.440700654	-0.684753458	-0.319558624	0.735000047	0.043687839	-0.183289972	0.99301409	-0.437006482	0.439948736	-0.657517569	1.022744291	0.17345733	-1.086223115	-0.136439551	1	0	0	0	0	0
9	1.440700654	1.244860367	-0.319558624	0.735000047	2.491099875	0.725906272	-0.853230626	0.648990498	0.393321973	3.756124156	0.694298579	0.613556043	0.192826998	-0.136439551	1	0	0	0	0	0
10	-0.694106716	0.139382049	-1.423742133	-1.45272002	-1.026138979	-0.340629248	-0.853230626	-0.437006482	-1.160903446	0.121360382	-0.911436011	-0.336472544	0.933329695	-0.136439551	1	0	0	0	0	0
11	-0.694106716	-0.758641469	-0.319558624	-1.45272002	-0.630449608	3.565037311	0.069891732	-0.437006482	-0.274994957	-0.657517569	-0.090321732	1.838686924	1.438217898	-0.136439551	1	0	0	0	0	0
12	-0.694106716	-0.247109085	-0.319558624	-0.358859987	-0.718380579	0.693314565	-0.853230626	-0.437006482	-0.368248482	-0.657517569	-0.838448075	1.030269436	-0.244742777	-0.136439551	1	0	0	0	0	0
13	1.440700654	0.491771025	1.888808396	0.005760024	-0.073553456	-0.663736689	0.069891732	-0.437006482	0.315610703	-0.657517569	-0.856695059	-0.709988056	2.380675876	-0.136439551	1	0	0	0	0	0
14	1.440700654	1.173814203	0.784624886	1.464240069	1.597135	-0.62159224	0.99301409	1.734987478	0.579829024	1.860854474	0.566569691	-0.98606474	-0.278401991	-0.136439551	1	0	0	0	0	0
15	-0.694106716	-0.275527551	-0.319558624	0.005760024	0.043687839	1.918313212	0.99301409	-0.437006482	0.439948736	-0.657517569	0.110395092	2.267742569	-0.816949407	-0.136439551	0	0	0	1	0	0
16	-0.694106716	0.236004833	0.784624886	-1.45272002	-0.381311856	-0.818266335	0.069891732	-0.437006482	-0.010776635	-0.657517569	-0.40052046	-0.88862591	1.808469247	-0.136439551	1	0	0	0	0	0
17	-0.694106716	-1.44636834	-1.423742133	-1.45272002	-0.967518332	0.741078273	-0.853230626	-0.437006482	-0.632466803	-0.657517569	-1.404104579	-0.630413012	1.640173179	-0.136439551	1	0	0	0	0	0
18	1.440700654	-1.213336921	-0.319558624	-1.45272002	-0.894242522	0.719725086	-0.853230626	-0.437006482	-0.554755532	-0.657517569	-1.002670931	0.589196335	0.058190144	-0.136439551	1	0	0	0	0	0
19	-0.694106716	-0.332364482	0.784624886	-0.358859987	-0.352001532	-0.6272115	-0.853230626	-0.437006482	-1.160903446	1.315639908	-0.783707123	-0.667764564	0.765033628	-0.136439551	1	0	0	0	0	0
20	-0.694106716	-0.900733798	1.888808396	0.735000047	0.600583991	-0.25634035	0.99301409	0.648990498	1.030554395	-0.657517569	0.749039531	-0.010052465	-0.850608621	-0.136439551	1	0	0	0	0	0
22	-0.694106716	-0.650651299	-0.319558624	0.370380035	0.864376905	-0.200147752	0.99301409	0.648990498	1.310314971	-0.657517569	0.676051595	-0.055523919	-0.514016485	-0.136439551	1	0	0	0	0	0
23	-0.694106716	-1.196285842	-0.319558624	0.005760024	-0.220105075	-0.706724027	0.069891732	-1.523003463	0.160188161	-0.657517569	-1.458845531	-0.757083491	0.966988909	-0.136439551	1	0	0	0	0	0
24	-0.694106716	1.270436987	-0.319558624	-1.45272002	-0.674415094	-1.582204712	0.069891732	0.648990498	-0.321621719	-0.657517569	0.05565414	-1.031211397	2.044083741	-0.136439551	1	0	0	0	0	0
25	-0.694106716	-0.031128745	-0.319558624	-0.358859987	-0.498553151	-0.233863311	-0.853230626	-0.437006482	-1.269699225	1.237752113	-0.564743316	-0.260145461	0.731374414	-0.136439551	1	0	0	0	0	0
27	-0.694106716	0.778797529	-0.319558624	0.735000047	0.6738598	-0.908174493	0.99301409	0.648990498	1.108265666	-0.657517569	1.369436986	-0.992560662	-1.119882329	-0.136439551	1	0	0	0	0	0

22.Data Preparation for Modeling after Scaling

After the process of scaling the data, the dataset is prepared for model training by splitting it into features (X_new) and target (Y_new). The price column is designated as the target variable (Y_new), while all other columns serve as features (X_new). Using the train_test_split function from sklearn.model_selection, the data is divided into training and testing sets with an 80-20 ratio, ensuring robust model evaluation. Subsequently, a Linear Regression model (lm) is instantiated and trained on the training data (X_new_train and y_new_train).

After scaling the data, a scatter plot (Figure 40) illustrates the correlation between the predicted and actual values, with the dashed green line representing the ideal scenario where predicted values perfectly match the actual values.

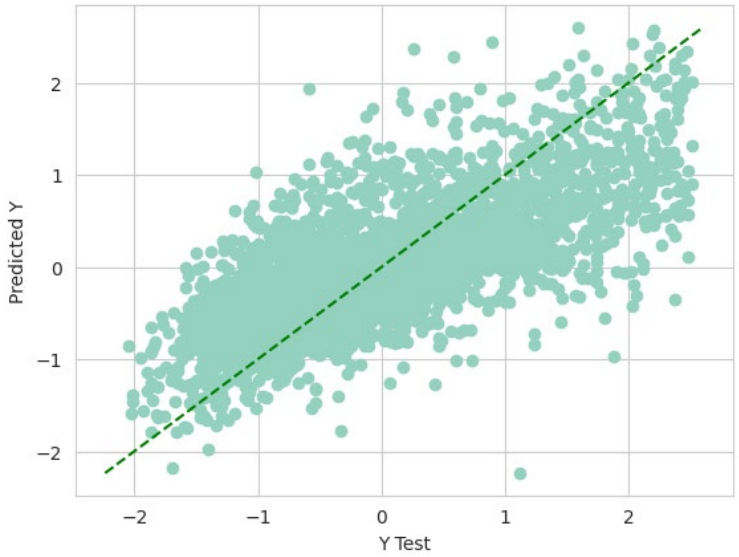

Figure 40 Linear regression model after Scaling

23.Evaluating the Model after Scaling

After scaling the data, the performance of the Linear Regression model (Table 7) was evaluated using several metrics. The Mean Absolute Error (MAE) is 0.5651, indicating the average absolute difference between the predicted and actual house prices. The Mean Squared Error (MSE) is 0.5037, which reflects the average squared difference between predicted and actual values, and the Root Mean Squared Error (RMSE) is 0.7097, providing a measure of the model's prediction accuracy. The R-squared value is 0.5095, showing that approximately 50.95% of the variance in house prices is explained by the model after scaling.

Table 7 The effectiveness of the linear regression model after Scaling

MAE	0.565091636434529
MSE	0.503717881794703
RMSE	0.709730851657657
R squared Error	0.509488562675303

24.Comparison of Linear Regression Model Performance Before and After Scaling

The performance of the linear regression model was evaluated using two sets of metrics: one before scaling the data and one after scaling. Before scaling, the model had a Mean Absolute Error (MAE) of 95964.06, a Mean Squared Error (MSE) of 14110488591.49, and a Root Mean Squared Error (RMSE) of 118787.58, with an R-squared value of 0.5468, indicating that approximately 54.68% of the variance in the house prices could be explained by the model.

After scaling the data, there was a significant improvement in the error metrics, with the MAE reduced to 0.5651, the MSE to 0.5037, and the RMSE to 0.7097. However, the R-squared value decreased slightly to 0.5095, suggesting a slight reduction in the model's explanatory power. This comparison highlights the impact of scaling on the model's performance, particularly in terms of error reduction.

25.Random Forest Regressor

In the next steps, it was decided to try using the Random Forest Regressor instead of the linear regression model. This approach aims to improve predictive accuracy by leveraging the ensemble learning capabilities of the Random Forest algorithm.

The implementation of the Random Forest Regressor begins by importing the necessary module from sklearn.ensemble library. A Random Forest Regressor instance is then created and trained using the training dataset. By creating and training a Random Forest Regressor on the training dataset, the objective is to leverage the ensemble learning technique that combines multiple decision trees to improve the robustness and accuracy of predictions. The model is trained using the training data, and predictions are made on the same dataset to evaluate the model's performance.

Table 8 The effectiveness of the linear regression model after Random Forest Regressor

MAE	31893.1770284667
MSE	1717196390.10632
RMSE	41439.0684029735
R squared Error	0.944445885054796

The implementation of the RandomForestRegressor yielded significantly improved results compared to the previous models. As shown in (Table 8), the Mean Absolute Error (MAE) decreased to 31,893.18, indicating a reduction in average prediction error. The Mean Squared Error (MSE) also dropped substantially to 1.71e+10, and the Root Mean Squared Error (RMSE) fell to 41,439.07, demonstrating a more accurate fit to the data. Additionally, the R squared Error increased remarkably to 0.944, highlighting the model's ability to explain a higher proportion of variance in the target variable. These results underscore the effectiveness of the Random Forest Regressor in enhancing the predictive accuracy of the housing price model.

26. Random Forest: Visualizing the differences between actual prices and predicted values

The implementation of the RandomForestRegressor is visualized in (Figure 41), illustrating the relationship between actual prices and predicted values. The scatter plot shows a strong linear correlation, with data points closely aligning along the diagonal line, indicating that the model's predictions closely match the actual prices. This visual evidence supports the statistical metrics previously discussed, affirming the Random Forest model's robustness and accuracy in predicting housing prices.

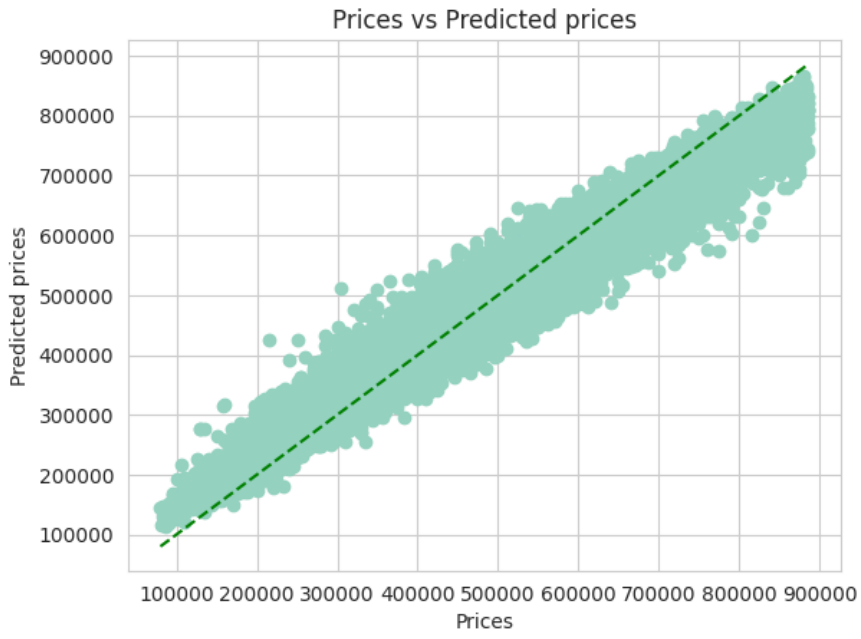

Figure 41 Random Forest: Visualizing the differences between actual prices and predicted values

27. Random Forest: Checking residuals

(Figure 42) presents the residuals plot for the Random Forest model, displaying the differences between the predicted values and the actual values. The residuals are distributed around zero, with

no clear pattern, indicating that the model captures the data's variance effectively without significant bias. This randomness in the residuals suggests that the Random Forest model is well-fitted to the data, accurately predicting housing prices with minimal systematic error.

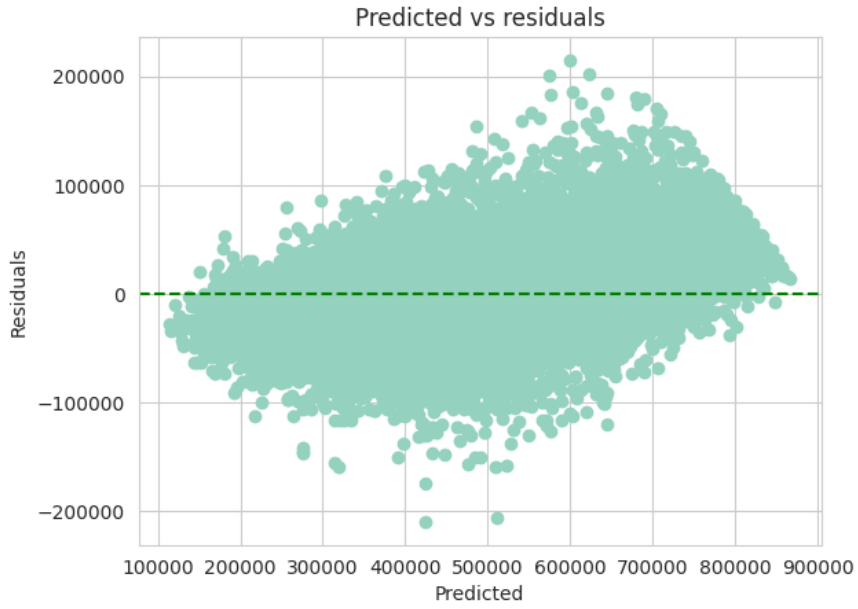

Figure 41 Residuals for the Random Forest model

28.The comparative analysis of the model

The comparative analysis of the model performance metrics before scaling (Table 5), after scaling (Table 7), and after applying the Random Forest Regressor (Table 8) reveals significant insights into the effectiveness of each approach. Initially, the linear regression model exhibited a Mean Absolute Error (MAE) of 95964.06, Mean Squared Error (MSE) of 14110488591.49, Root Mean Squared Error (RMSE) of 118787.58, and an R-squared value of 0.5468. Scaling the data improved the model's performance, reducing the MAE to 0.5651, MSE to 0.5037, RMSE to 0.7097, and the R-squared value to 0.5095, indicating a more normalized data distribution. However, the most substantial improvement was observed with the application of the Random Forest Regressor, which achieved an MAE of 31893.18, MSE of 1717196390.11, RMSE of 41439.07, and an R-squared value of 0.9444. This comparison underscores the superior predictive accuracy and reliability of the Random Forest Regressor over the traditional linear regression models, both before and after scaling.

Conclusion

The regression analysis conducted in this report aimed to predict house prices in King County using a comprehensive dataset. Three distinct methods were employed to enhance the robustness and accuracy of the predictions: basic regression analysis, regression analysis after scaling the data, and the application of a Random Forest Regressor.

Basic Regression Analysis:

The initial linear regression model provided a baseline for understanding the relationships between the variables and house prices. Despite achieving an R-squared value of 0.5468, the model's predictive power was moderate, indicated by a high Mean Absolute Error (MAE) of 95964.06 and Root Mean Squared Error (RMSE) of 118787.58. This suggested that while the model captured some of the variance in house prices, there were significant inaccuracies that needed addressing.

Regression Analysis After Scaling:

To improve the model's performance, the data was scaled using the StandardScaler. This normalization process aims to ensure that all features contributed equally to the model. The scaled data model showed an improvement in error metrics with an MAE of 0.5651 and an RMSE of 0.7097. However, the R-squared value decreased slightly to 0.5095. The scaling helped in reducing the prediction errors, but it also highlighted the limitations of the linear regression model in capturing the complex relationships in the data.

Random Forest Regressor:

The application of the Random Forest Regressor significantly enhanced the model's predictive accuracy. The Random Forest model achieved an R-squared value of 0.9444, demonstrating a substantial improvement over the linear regression models. The MAE reduced to 31893.18, and the RMSE decreased to 41439.07. This method's effectiveness was evident in the close alignment of predicted values with actual prices and the well-distributed residuals around zero, indicating minimal systematic error.

Data Preprocessing and Feature Engineering:

Rigorous data cleaning, handling of outliers, and feature engineering were crucial steps that improved the dataset's quality and the subsequent model performance. Creating new features like the age of buildings and years since the last renovation added valuable dimensions to the analysis.

Model Comparisons:

The comparative analysis highlighted that while linear regression models provide a baseline understanding, they may not fully capture the complexities of the data. Scaling improved the model's performance by normalizing feature contributions. However, the Random Forest Regressor outperformed both linear regression models, indicating its superior capability in handling diverse and complex datasets.

The results of this analysis underscore the importance of using advanced regression techniques, such as Random Forest, to achieve higher predictive accuracy in complex datasets like real estate pricing. While basic and scaled linear regression models offer insights, ensemble methods like Random Forest provide a robust framework for capturing intricate relationships between variables.