House Price Prediction

INFO 6105 Data Science Engineering Methods and Tools

Sisi Tian 002060372

Ming-Hsiang Lee 002308390

Yun Yang 002050051

Yiyi Wang 002302810

Nov. 2024

Table of Contents

Project Overview

Motivation

Our project aims to develop a predictive model to estimate house sale prices in King County using data science and machine learning techniques. The focus will be on accurately predicting prices and analyzing key property characteristics that influence valuation.

Data Source and Timeframe

The dataset covers home sales in King County from May 2014 to May 2015, featuring variables such as property features, location details, and sale prices.

Evaluation Metric

The model's accuracy is evaluated using Root Mean Square Error (RMSE) on the logarithmic scale of predicted and observed sale prices. This approach ensures that errors are normalized across high- and low-priced properties.

Data Summary and Inital Analysis

Quantitative Variables:

Continuous

- **price** – Sale price of the home
- **sqft_living** – Interior living space square footage
- **sqft_lot** – Land space square footage
- **sqft_above** – Square footage above ground level
- **sqft_basement** – Square footage below ground level
- **lat** – Latitude
- **long** – Longitude
- **sqft_living15** – Living space square footage of the nearest 15 neighbors
- **sqft_lot15** – Land lot square footage of the nearest 15 neighbors

Discrete

- **floors** - Number of floors
- **bedrooms** – Number of bedrooms
- **bathrooms** – Number of bathrooms (can include fractional values like 0.5)

Categorical Variables

Binary

- **waterfront** – Dummy variable (1 if the home overlooks waterfront, 0 if not)

Nominal

- **date** – Date of the home sale
- **zipcode** – Zip code of the house location

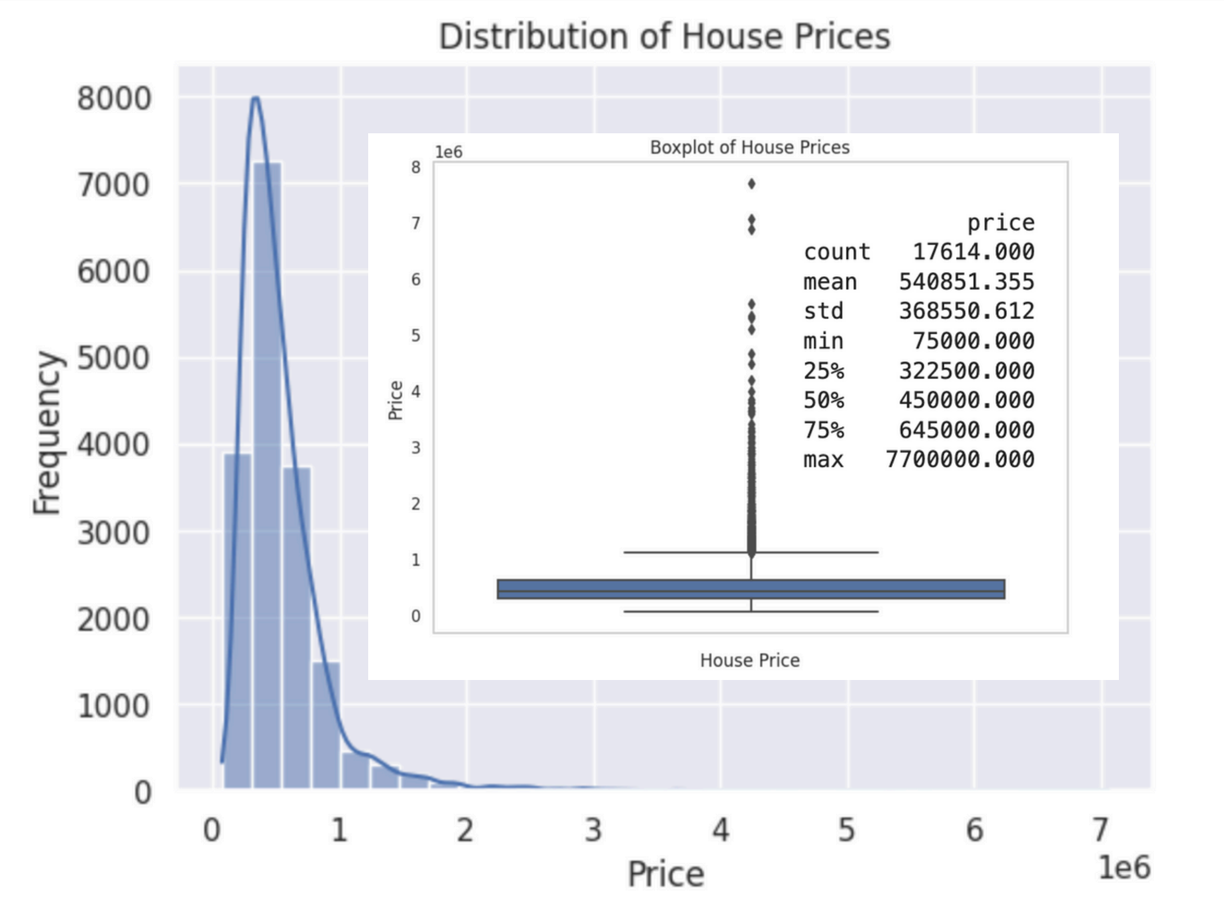
Ordinal

- **view** – View index (0 to 4)
- **condition** – Condition index (1 to 5)
- **grade** – Grade index (1 to 13)

---

yr_built: Year the house was built
yr_renovated: Year of last renovation

Data Summary and Inital Analysis



Observations on House Price Distribution

- Right-Skewed Distribution

Majority of homes are priced below $1M.

A small number of very high-priced homes (above $3M) skew the average upwards.

- Potential Outliers

High-end properties (>$3M) could distort model predictions.

Consider removing or treating these outliers.

- Price Normalization Needed

Log transformation of prices may help address the skewness, improving model performance on both low- and high-priced homes.

Data Summary and Inital Analysis

Property Characteristics

	sqft_living	sqft_above	sqft_basement
count	17603.0000	17603.0000	17603.0000
mean	2077.9761	1787.4277	290.5484
std	911.1863	824.4028	439.8333
min	290.0000	290.0000	0.0000
25%	1420.0000	1190.0000	0.0000
50%	1910.0000	1560.0000	0.0000
75%	2550.0000	2210.0000	560.0000
max	10040.0000	7880.0000	3000.0000

- There is a **positive correlation** between square footage of the living area and price, suggesting larger homes tend to be more expensive.
- sqft_living = sqft_above + sqft_basement

Data Summary and Inital Analysis

Property Characteristics

	bedrooms	bathrooms	floors
count	17614.000	17614.000	17614.000
mean	3.372	2.113	1.496
std	0.936	0.774	0.540
min	0.000	0.000	1.000
25%	3.000	1.500	1.000
50%	3.000	2.250	1.500
75%	4.000	2.500	2.000
max	33.000	8.000	3.500

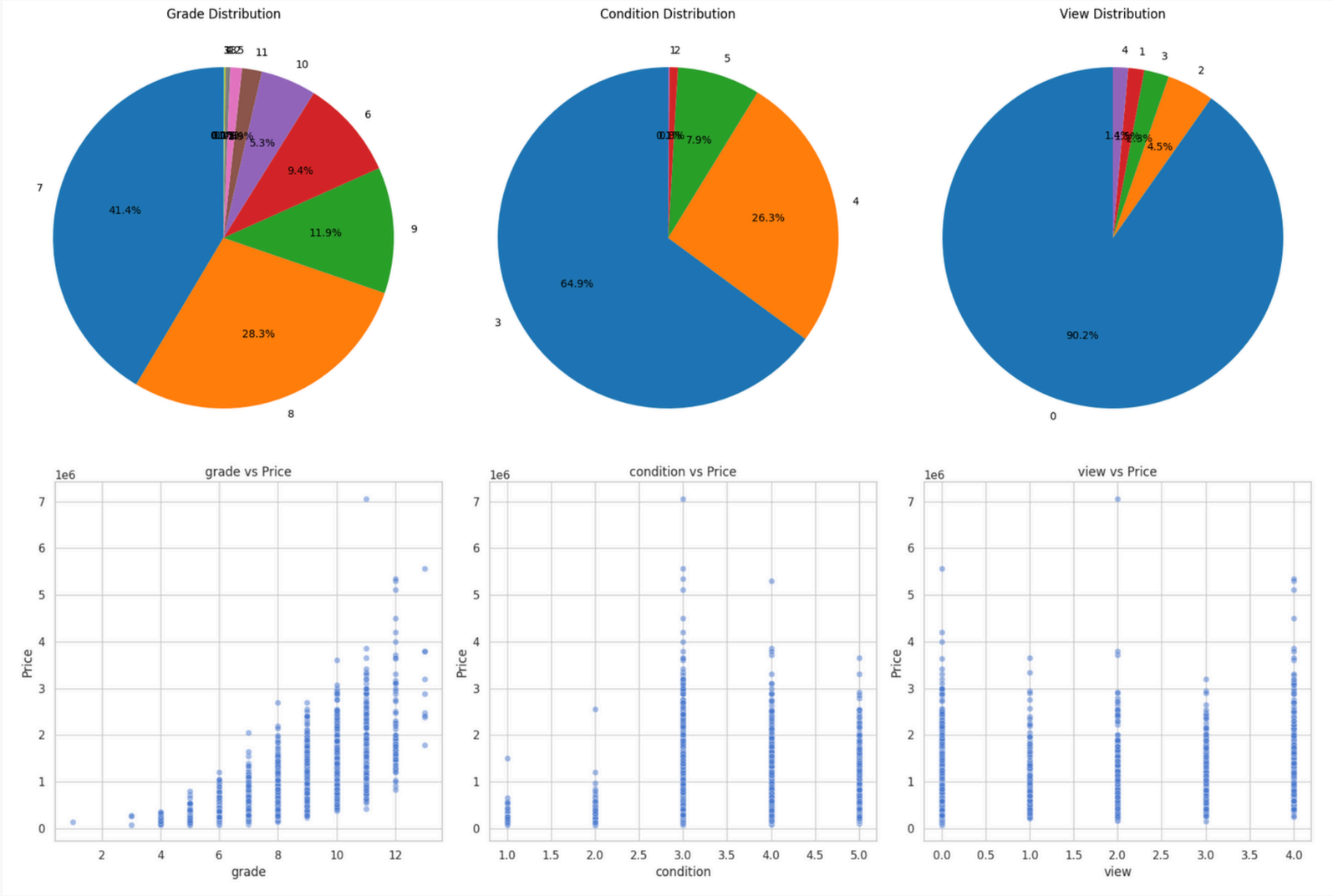
Bathrooms and Bedrooms: The distribution of bathrooms and bedrooms shows **potential skewness**, especially with outliers. Consider log transformation or removing outliers.

Data Summary and Inital Analysis

```
Summary of grade:
grade
7      7296
8      4983
9      2101
6      1658
10      928
11      329
5       196
12       77
4        22
13        9
3         3
1         1
Name: count, dtype: int64
```

```
Summary of condition:
condition
3    11426
4     4632
5     1383
2      140
1       22
Name: count, dtype: int64
```

```
Summary of view:
view
0    15882
2      786
3      413
1      268
4      254
Name: count, dtype: int64
```



Higher grades, higher prices. Condition alone does not strongly influence price.

Data Summary and Inital Analysis

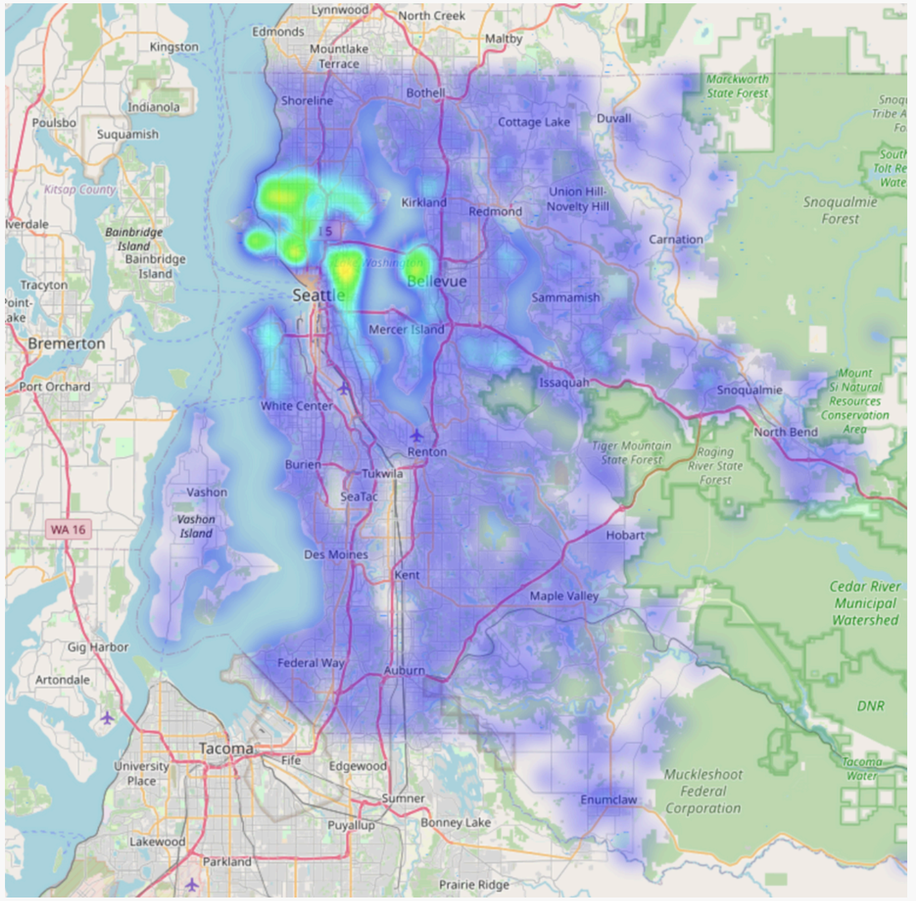
Location-Based Variables

```
Summary of zipcode:
zipcode
370.5582    487
353.4907    474
173.8543    474
279.9155    452
362.7332    450
           ...
427.6166     84
208.1152     77
260.0724     72
183.3251     45
570.8198     36
Name: count, Length: 70, dtype: int64
```

```
                lat          long
count  17603.0000  17603.0000
mean      47.5605   -122.2142
std        0.1384      0.1401
min       47.1559   -122.5190
25%       47.4716   -122.3280
50%       47.5726   -122.2300
75%       47.6780   -122.1250
max       47.7776   -121.3150

Summary of waterfront:
waterfront
0    17479
1      124
Name: count, dtype: int64
```

Zipcode exhibits **high cardinality**, with 70 unique values



Price heatmap based on lat and long

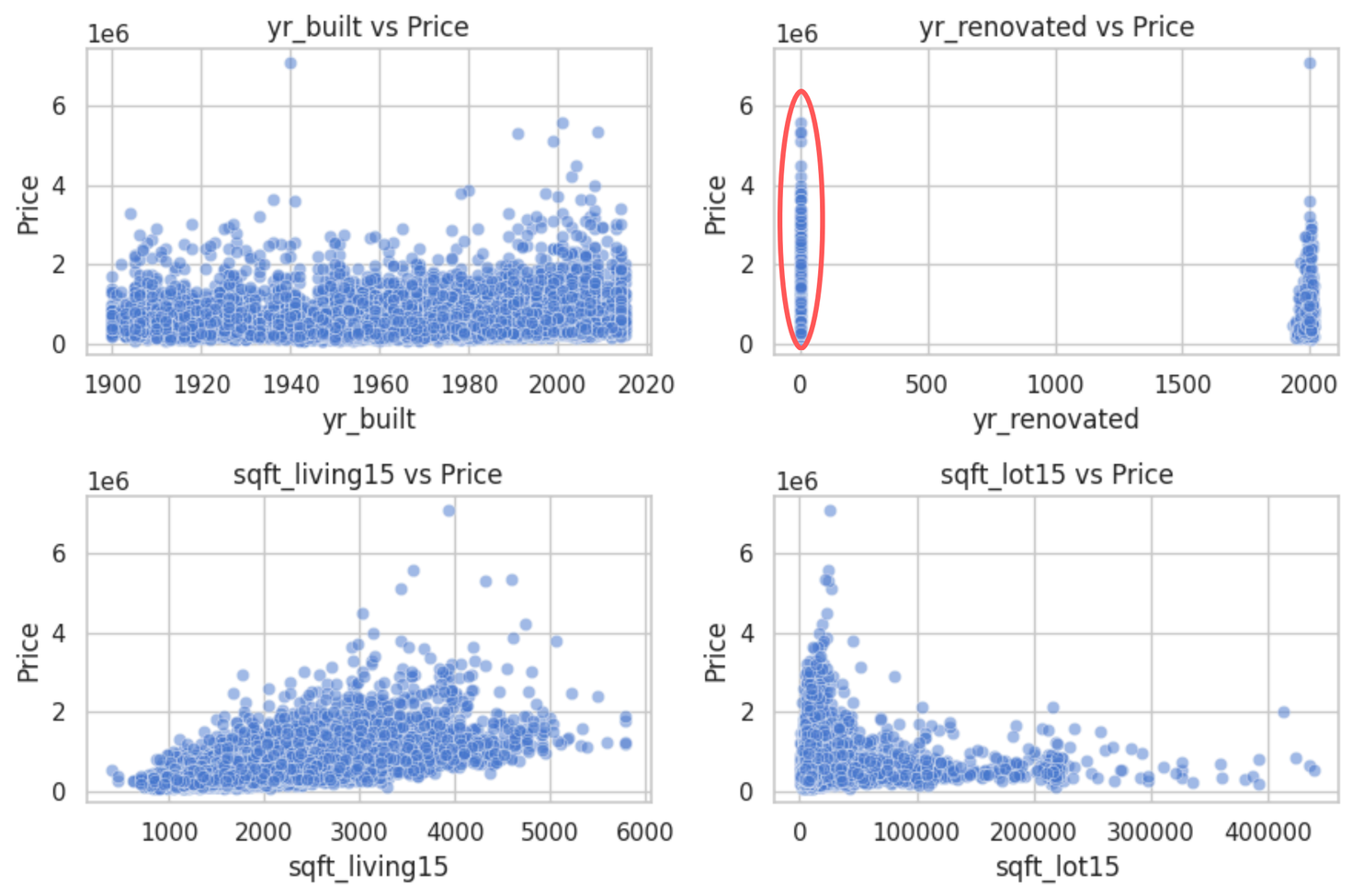
Data Summary and Inital Analysis

Age and Renovations

	yr_built	yr_renovated
count	17603.0000	17603.0000
mean	1970.9141	80.5979
std	29.4231	392.8758
min	1900.0000	0.0000
25%	1951.0000	0.0000
50%	1975.0000	0.0000
75%	1997.0000	0.0000
max	2015.0000	2015.0000

Neighborhood Information:

	sqft_living15	sqft_lot15
count	17603.0000	17603.0000
mean	1984.7411	12559.6566
std	683.8203	25349.4798
min	399.0000	659.0000
25%	1480.0000	5100.0000
50%	1840.0000	7600.0000
75%	2360.0000	10050.0000
max	5790.0000	438213.0000

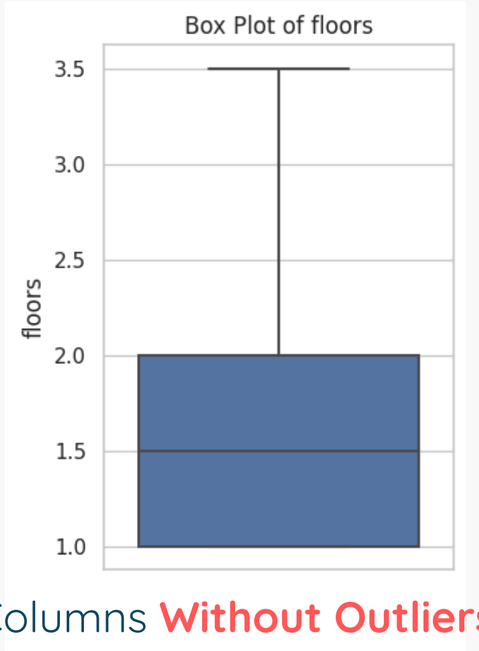


Data Preprocessing and Cleaning

Identifying Outliers

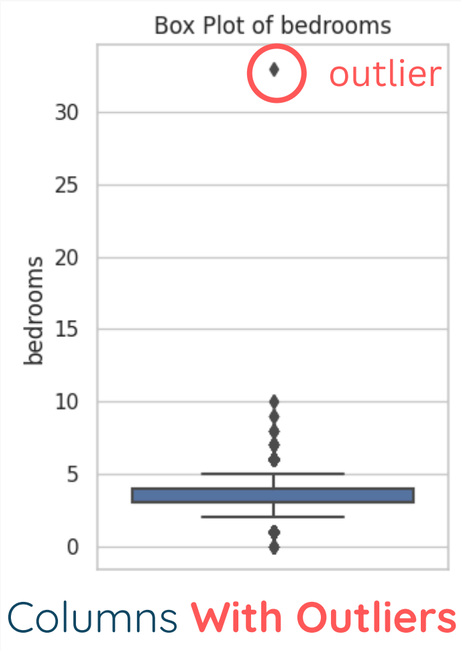
- **Method:**

Use box plots to visualize the relationship between numeric features and house prices.



Columns **Without Outliers**

feature “floor” as an example



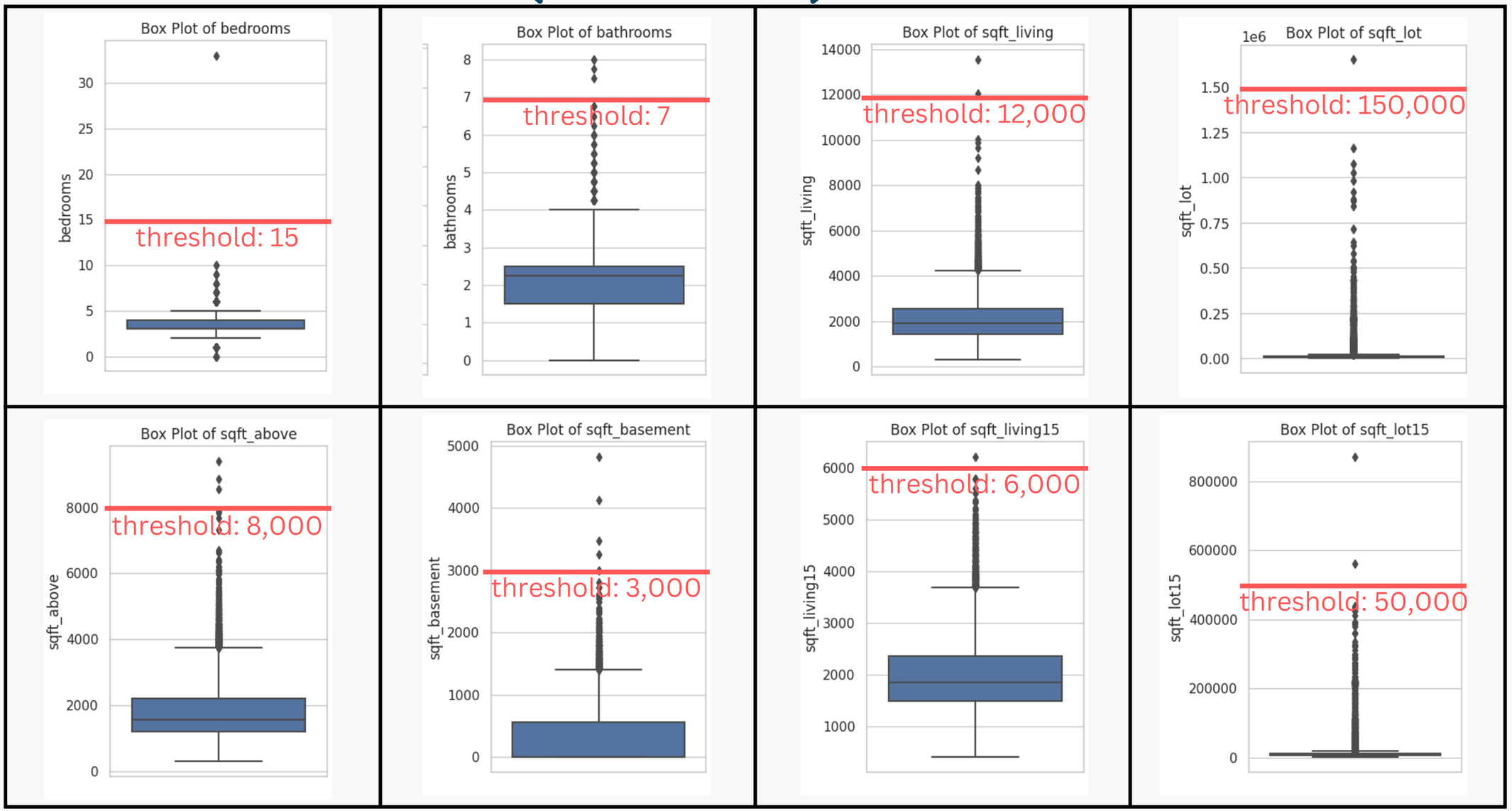
Columns **With Outliers**

feature “bedrooms” as an example

- **Comment:** 8 features contain at least one outlier and those outliers need to be excluded from the dataset.

Columns Without Outliers (10 features)	Columns With Outliers (8 features)
floors	bedrooms
long	bathrooms
waterfront	sqft_living
view	sqft_lot
condition	sqft_above
grade	sqft_basement
yr_built	sqft_living15
yr_renovated	sqft_lot15
zipcode	
lat	

Columns With Outliers (8 features)



- Comment: Determined thresholds for these 8 features based on visual differences.

Exclude 11 rows of data in total.

Data Preprocessing and Cleaning

Handling Missing Value - null values

id	0
date	0
bedrooms	0
bathrooms	0
sqft_living	0
sqft_lot	0
floors	0
waterfront	0
view	0
condition	0
grade	0
sqft_above	0
sqft_basement	0
yr_built	0
yr_renovated	0
zipcode	0
lat	0
long	0
sqft_living15	0
sqft_lot15	0

- **Method:**

1. Check for missing values using the **isnull().sum()** function, except for **price** column

```
missing_values = all_data.filter(regex='^(?!price$)').isnull().sum()
print(missing_values)
```

- **Comment:**

All features show zero missing values, indicating our dataset is complete.

Data Preprocessing and Cleaning

Handling Missing Value - 0 values

	yr_renovated
count	17603.0000
mean	80.5979
std	392.8758
min	0.0000
25%	0.0000
50%	0.0000
75%	0.0000
max	2015.0000

- **Method:**

2. Treat **0** in **yr_renovated** as missing, replacing it with **yr_built** to ensure completeness.
(affecting 16,901 values in training dataset, and 3798 values in test dataset)

```
all_data['yr_renovated'] = np.where(all_data['yr_renovated'] == 0,
                                    all_data['yr_built'],
                                    all_data['yr_renovated'])
```

	bedrooms	bathrooms
count	17603.0000	17603.0000
mean	3.3697	2.1116
std	0.9069	0.7677
min	0.0000	0.0000
25%	3.0000	1.5000
50%	3.0000	2.2500
75%	4.0000	2.5000
max	10.0000	6.7500

3. Remove rows where **bedrooms** or **bathrooms** had a value of **0**.
(removing 15 rows from training dataset, and 1 row in test dataset)

```
train_df = train_df[(train_df['bedrooms'] != 0) & (train_df['bathrooms'] != 0)]
test_df = test_df[(test_df['bedrooms'] != 0) & (test_df['bathrooms'] != 0)]
```

- **Comment:**

Replaced **0** values in **yr_renovated** with **yr_built** to accurately reflect the renovation year.
Rows with **0** values for **bedrooms** or **bathrooms** were removed to maintain realistic values

Feature Engineering

Classification type features

date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
20140901T000000	450000	3	2.250	1780	9969	1.000
20140930T000000	430000	2	2.250	1040	1516	2.000
20140731T000000	230000	4	1.500	1520	8800	1.000
20140715T000000	275000	3	1.500	1060	6954	1.000
20141031T000000	400000	3	2.000	1350	7216	1.000

view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat
0	...	8	1450	330	1985	0	98034	47.
0	...	8	1040	0	2008	0	98122	47.
0	...	7	1520	0	1960	0	98002	47.
0	...	6	1060	0	1983	0	98106	47.
0	...	7	1350	0	1964	0	98177	47.

SPECIAL FEATURE PROCESSING

zipcode

- This results in too many feature dimensions and increases the complexity of the model.
- Zipcode is essentially a label, without continuity or proximity
- Potential risk of data leakage
- The lack of geographic information

date

- The original date feature has no order meaning
- This date format contains the year, month, day and specific time.

Distribution of Target Value

The distribution of the target variable, price, shows significant skewness, indicating that it is not normally distributed. This lack of normality can negatively affect the performance of models that assume a normal distribution for optimal predictions.

Feature Engineering

SPECIAL FEATURE PROCESSING

ZIPCODE

Zip codes often reflect differences in local property markets (e.g., expensive vs. affordable neighborhoods).By using the average price per square foot, we can quantifying these local trends more accurately.Homes in the same area will likely share similar price-to-size ratios, helping the model capture relevant neighborhood effects.

	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	\
0	590	1942	1942	362.733	47.700	-122.371	
1	840	1931	1931	433.223	47.642	-122.351	
2	220	1936	1936	230.524	47.570	-122.360	
3	0	1986	1986	177.188	47.426	-122.163	
4	0	1988	2000	287.663	47.608	-122.147	
...	...	...	...	...	...	...	
3994	660	1918	1918	230.524	47.533	-122.347	
3995	0	1985	1985	154.940	47.368	-122.182	
3996	0	1992	1992	147.370	47.292	-122.375	
3997	0	1997	1997	477.553	47.644	-122.185	
3998	1580	2006	2006	221.566	47.570	-122.296	

	sqft_living15	sqft_lot15
0	1450	6800
1	1790	3000
2	1920	5000
3	2030	8183
4	1880	3350
...	...	...
3994	1190	4200
3995	1780	7210
3996	2001	7547

Feature Engineering

DATE

Date and time information often carries important trends and patterns that affect housing prices.

Processing date features can help the model capture important factors such as time-related trends, periodicity, policy impacts, etc., thereby improving the accuracy and reliability of housing price predictions.

SPECIAL FEATURE PROCESSING

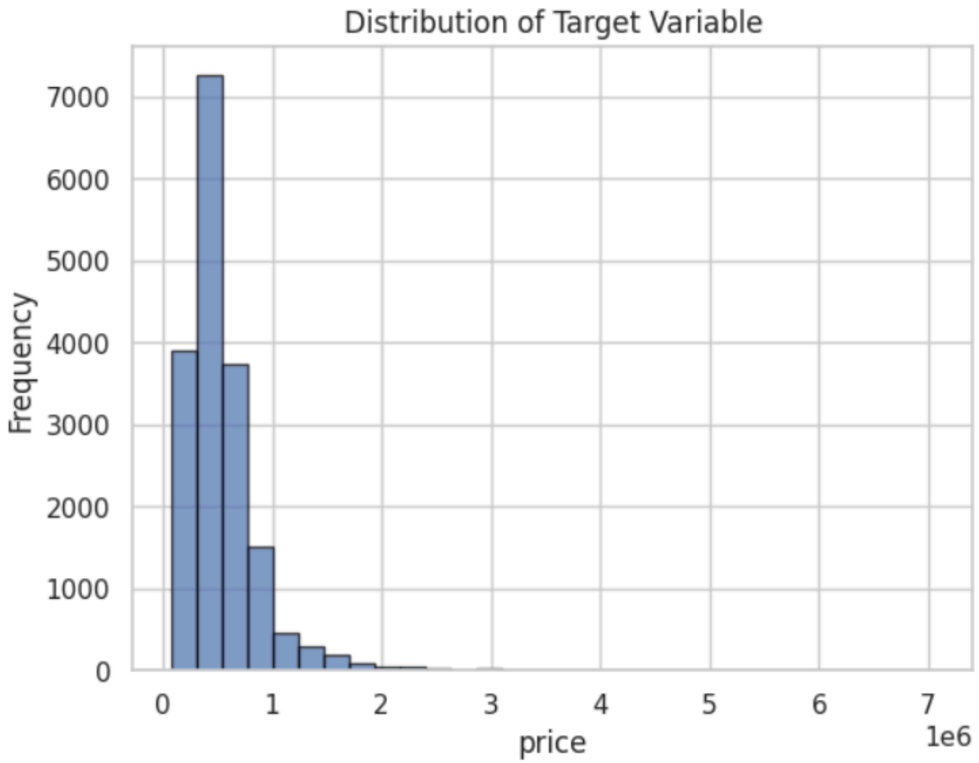
id	date	bedrooms	bathrooms	sqft_li
6825100015	20140604T000000	2	1.75	
1698900195	20140902T000000	3	2.00	
2848700585	20150424T000000	1	1.00	
2597690050	20150409T000000	4	1.75	
8944600200	20140623T000000	3	2.50	

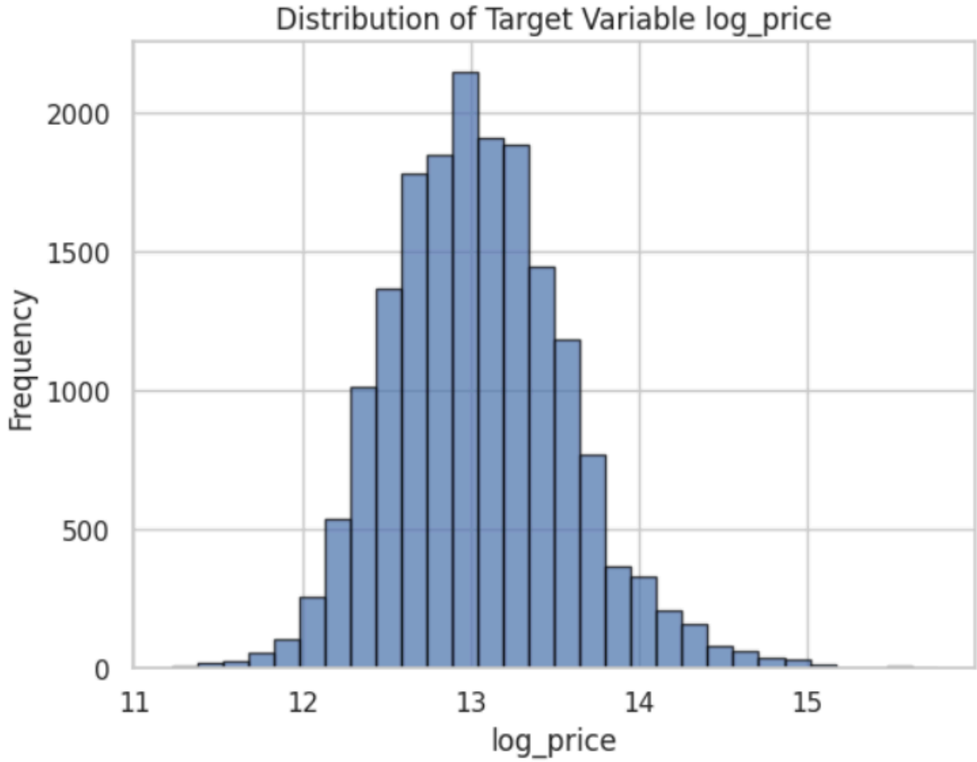
year_sold	month_sold
2014	9
2014	9
2014	7
2014	7
2014	10

Feature Engineering

PRICE

FEATURE TRANSFORMATION





Feature and Model Selection

Model 1: Linear Regression with All Features

```
Cross-validated RMSE scores: [0.18477929 0.18683658 0.18621464 0.18557236 0.18833121]
Mean RMSE: 0.18634681658319716
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.874
Method:                 Least Squares   F-statistic:                     6440.
Date:                Sun, 27 Oct 2024   Prob (F-statistic):               0.00
Time:                        07:42:41   Log-Likelihood:                 4624.5
No. Observations:               17588   AIC:                            -9209.
Df Residuals:                   17568   BIC:                            -9053.
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -116.3922      9.899    -11.757      0.000    -135.796     -96.988
bedrooms       0.0042      0.002      2.083      0.037       0.000       0.008
bathrooms      0.0402      0.003     11.939      0.000       0.034       0.047
sqft_living    0.0001   2.37e-06     46.192      0.000       0.000       0.000
sqft_lot    5.386e-07   5.03e-08     10.717      0.000     4.4e-07    6.37e-07
floors        -0.0297      0.004     -7.872      0.000      -0.037      -0.022
waterfront     0.4370      0.018     23.891      0.000       0.401       0.473
view           0.0560      0.002     25.717      0.000       0.052       0.060
condition      0.0605      0.002     25.063      0.000       0.056       0.065
grade          0.0958      0.002     41.783      0.000       0.091       0.100
sqft_above  8.936e-05   2.37e-06     37.771      0.000    8.47e-05    9.4e-05
sqft_basement  1.997e-05   2.75e-06      7.276      0.000    1.46e-05    2.54e-05
yr_built      -0.0016      0.000    -13.063      0.000      -0.002      -0.001
yr_renovated   0.0013      0.000     10.497      0.000       0.001       0.002
zipcode        0.0032   2.59e-05    121.820      0.000       0.003       0.003
lat            0.4070      0.013     30.616      0.000       0.381       0.433
long           0.1998      0.013     15.969      0.000       0.175       0.224
sqft_living15  8.968e-05   3.56e-06     25.197      0.000    8.27e-05    9.67e-05
sqft_lot15    -5.321e-08   7.89e-08     -0.674      0.500   -2.08e-07    1.01e-07
year_sold      0.0658      0.005     13.611      0.000       0.056       0.075
month_sold     0.0029      0.001      3.937      0.000       0.001       0.004
==============================================================================
Omnibus:                     1523.253   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5873.273
Skew:                          -0.377   Prob(JB):                         0.00
Kurtosis:                       5.729   Cond. No.                     3.25e+17
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.83e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

- use **k-fold cross-validation** to calculate RMSE
- prevent overfitting and provide a more reliable performance assessment

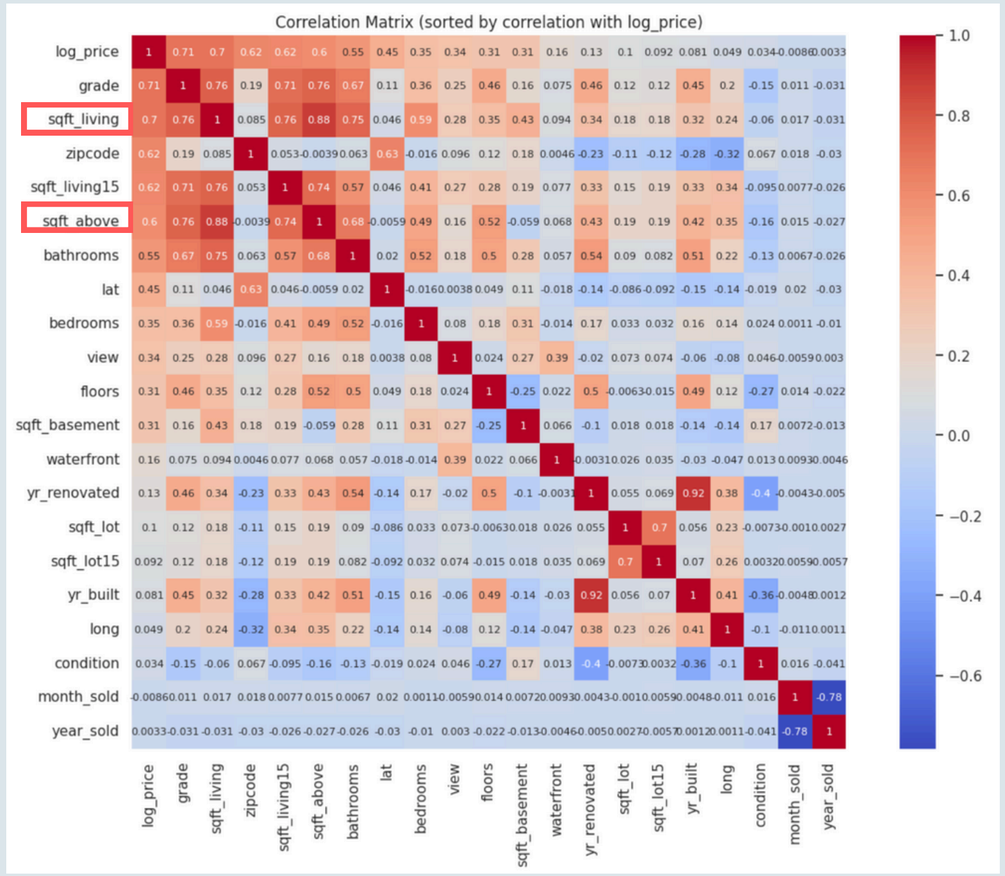
RMSE = 0.1863

R^2 = 0.874

- high R^2 does not indicate the quality of the model.
- ignore **multicollinearity**
- when some independent variables are highly correlated, their coefficients can become unstable, reducing the model's interpretability

Feature and Model Selection

Model 2: Linear Regression with Multicollinearity Handling



	feature	VIF
0	price	21.687
1	bedrooms	26.097
2	bathrooms	29.355
3	sqft_living	inf
4	sqft_lot	2.289
5	floors	18.623
6	waterfront	1.289
7	view	1.607
8	condition	36.658
9	grade	176.869
10	sqft_above	inf
11	sqft_basement	inf
12	yr_built	30904.570
13	yr_renovated	32109.885
14	zipcode	49.843
15	lat	218687.746
16	long	1180621.325
17	sqft_living15	29.340
18	sqft_lot15	2.525
19	year_sold	1493364.536
20	month_sold	5.783
21	log_price	6662.798

Features with VIF greater than 10

	feature	VIF
0	price	21.687
1	bedrooms	26.097
2	bathrooms	29.355
3	sqft_living	inf
5	floors	18.623
8	condition	36.658
9	grade	176.869
10	sqft_above	inf
11	sqft_basement	inf
12	yr_built	30904.570
13	yr_renovated	32109.885
14	zipcode	49.843
15	lat	218687.746
16	long	1180621.325
17	sqft_living15	29.340
19	year_sold	1493364.536
21	log_price	6662.798

```
Cross-validated RMSE scores: [0.18944384 0.1921918  0.19206282 0.19202521 0.19406627]
Mean RMSE: 0.19195798791002786
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.867
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     7142.
Date:                Sun, 27 Oct 2024   Prob (F-statistic):               0.00
Time:                        07:42:42   Log-Likelihood:                 4098.7
No. Observations:               17588   AIC:                            -8163.
Df Residuals:                   17571   BIC:                            -8031.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -97.8874     10.182     -9.614      0.000    -117.844     -77.930
bedrooms       0.0067      0.002      3.255      0.001       0.003       0.011
sqft_living    0.0002   3.02e-06     73.680      0.000       0.000       0.000
sqft_lot    4.719e-07   5.16e-08      9.141      0.000    3.71e-07    5.73e-07
floors        -0.0027      0.003     -0.810      0.418      -0.009       0.004
waterfront     0.4361      0.019     23.170      0.000       0.399       0.473
view           0.0589      0.002     26.880      0.000       0.055       0.063
condition      0.0564      0.002     22.869      0.000       0.052       0.061
grade          0.1221      0.002     55.255      0.000       0.118       0.126
yr_built      -0.0014      0.000    -10.909      0.000      -0.002      -0.001
yr_renovated   0.0012      0.000      9.343      0.000       0.001       0.001
zipcode        0.0031   2.62e-05    118.362      0.000       0.003       0.003
lat            0.4088      0.014     29.860      0.000       0.382       0.436
long           0.3127      0.012     25.455      0.000       0.289       0.337
sqft_lot15  7.207e-08   8.11e-08      0.889      0.374   -8.69e-08    2.31e-07
year_sold      0.0634      0.005     12.714      0.000       0.054       0.073
month_sold     0.0025      0.001      3.351      0.001       0.001       0.004
==============================================================================
Omnibus:                     1450.446   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5301.460
Skew:                          -0.370   Prob(JB):                         0.00
Kurtosis:                       5.586   Cond. No.                     3.37e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.37e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

RMSE = 0.1920 R^2 = 0.867

Feature and Model Selection

Model 3: Linear Regression with Multicollinearity Handling and Removal of Insignificant Variables

Model 2

```
Cross-validated RMSE scores: [0.18944384 0.1921918  0.19206282 0.19202521 0.19406627]
Mean RMSE: 0.19195798791002786
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.867
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     7142.
Date:                Sun, 27 Oct 2024   Prob (F-statistic):               0.00
Time:                        07:42:42   Log-Likelihood:                 4098.7
No. Observations:               17588   AIC:                            -8163.
Df Residuals:                   17571   BIC:                            -8031.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -97.8874     10.182     -9.614      0.000    -117.844     -77.930
bedrooms       0.0067      0.002      3.255      0.001       0.003       0.011
sqft_living    0.0002   3.02e-06     73.680      0.000       0.000       0.000
sqft_lot    4.719e-07   5.16e-08      9.141      0.000    3.71e-07    5.73e-07
floors        -0.0027      0.003     -0.810      0.418      -0.009       0.004
waterfront     0.4361      0.019     23.170      0.000       0.399       0.473
view           0.0589      0.002     26.880      0.000       0.055       0.063
condition      0.0564      0.002     22.869      0.000       0.052       0.061
grade          0.1221      0.002     55.255      0.000       0.118       0.126
yr_built      -0.0014      0.000    -10.909      0.000      -0.002      -0.001
yr_renovated   0.0012      0.000      9.343      0.000       0.001       0.001
zipcode        0.0031   2.62e-05    118.362      0.000       0.003       0.003
lat            0.4088      0.014     29.860      0.000       0.382       0.436
long           0.3127      0.012     25.455      0.000       0.289       0.337
sqft_lot15  7.207e-08   8.11e-08      0.889      0.374   -8.69e-08    2.31e-07
year_sold      0.0634      0.005     12.714      0.000       0.054       0.073
month_sold     0.0025      0.001      3.351      0.001       0.001       0.004
==============================================================================
Omnibus:                     1450.446   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5301.460
Skew:                          -0.370   Prob(JB):                         0.00
Kurtosis:                       5.586   Cond. No.                     3.37e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.37e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

RMSE = 0.1920 R^2 = 0.867

Model 3

```
Cross-validated RMSE scores: [0.18935838 0.19206642 0.1920824  0.19204003 0.19405078]
Mean RMSE: 0.1919196025030181
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.867
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     8162.
Date:                Sun, 27 Oct 2024   Prob (F-statistic):               0.00
Time:                        07:42:42   Log-Likelihood:                 4098.0
No. Observations:               17588   AIC:                            -8166.
Df Residuals:                   17573   BIC:                            -8049.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -97.6766     10.179     -9.596      0.000    -117.629     -77.724
bedrooms       0.0066      0.002      3.199      0.001       0.003       0.011
sqft_living    0.0002   3.01e-06     73.904      0.000       0.000       0.000
sqft_lot    5.036e-07   3.86e-08     13.046      0.000    4.28e-07    5.79e-07
waterfront     0.4358      0.019     23.167      0.000       0.399       0.473
view           0.0591      0.002     26.983      0.000       0.055       0.063
condition      0.0567      0.002     23.082      0.000       0.052       0.062
grade          0.1219      0.002     55.541      0.000       0.118       0.126
yr_built      -0.0014      0.000    -11.076      0.000      -0.002      -0.001
yr_renovated   0.0012      0.000      9.309      0.000       0.001       0.001
zipcode        0.0031   2.57e-05    120.473      0.000       0.003       0.003
lat            0.4093      0.014     29.961      0.000       0.383       0.436
long           0.3145      0.012     25.822      0.000       0.291       0.338
year_sold      0.0634      0.005     12.718      0.000       0.054       0.073
month_sold     0.0025      0.001      3.353      0.001       0.001       0.004
==============================================================================
Omnibus:                     1453.437   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5312.598
Skew:                          -0.371   Prob(JB):                         0.00
Kurtosis:                       5.588   Cond. No.                     2.97e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.97e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

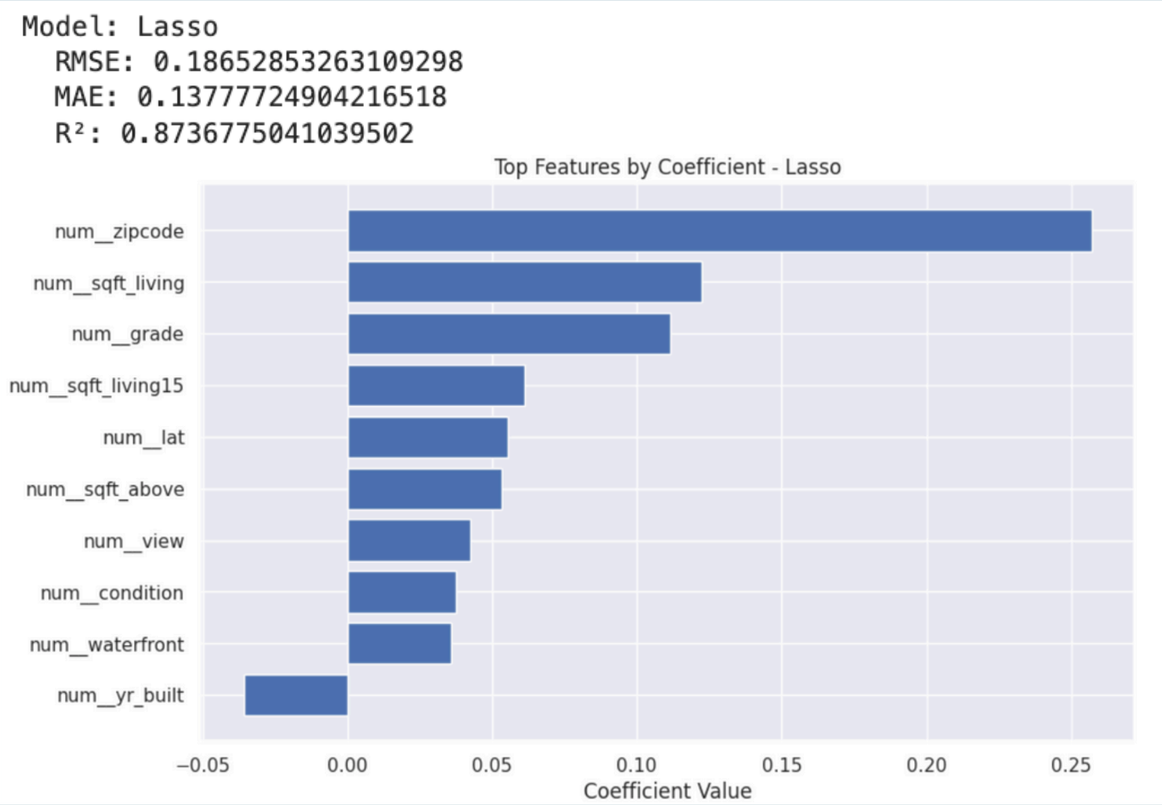
RMSE = 0.1919 R^2 = 0.867

simplify the model

Feature and Model Selection

Model 4: Lasso Regression

```
Model: Lasso
  RMSE: 0.18652853263109298
  MAE: 0.13777724904216518
  R²: 0.8736775041039502
```



- apply a penalty term to the model parameters
- shrink unimportant feature coefficients to zero
- automatically perform feature selection
- select fewer features
- avoid overfitting

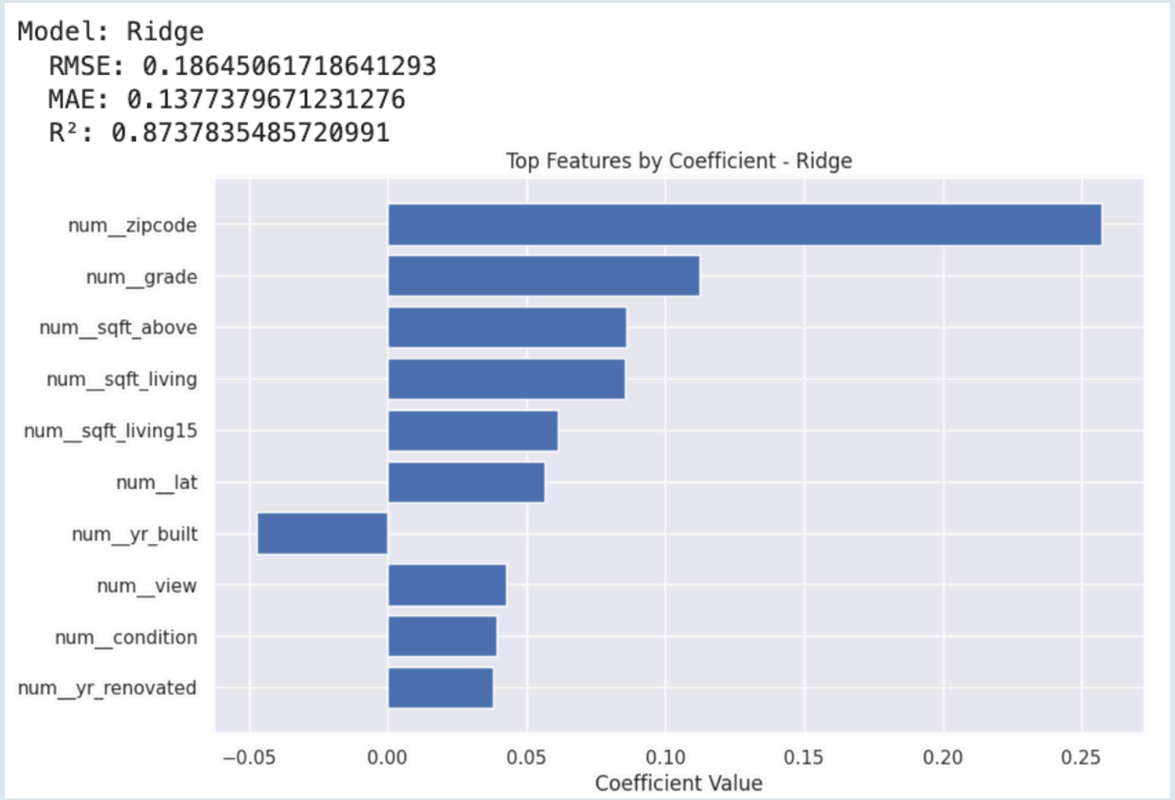
RMSE = 0.1865

R^2 = 0.8737

- zipcode
- sqft_living
- grade
- sqft_living15
- lat

Feature and Model Selection

Model 5: Ridge Regression



- does not shrink coefficients to zero
- can effectively address multicollinearity issues. When there is a high correlation between variables, Ridge reduces the coefficient values, lowering the model's reliance on specific features, which makes the model more stable.

RMSE = 0.1865

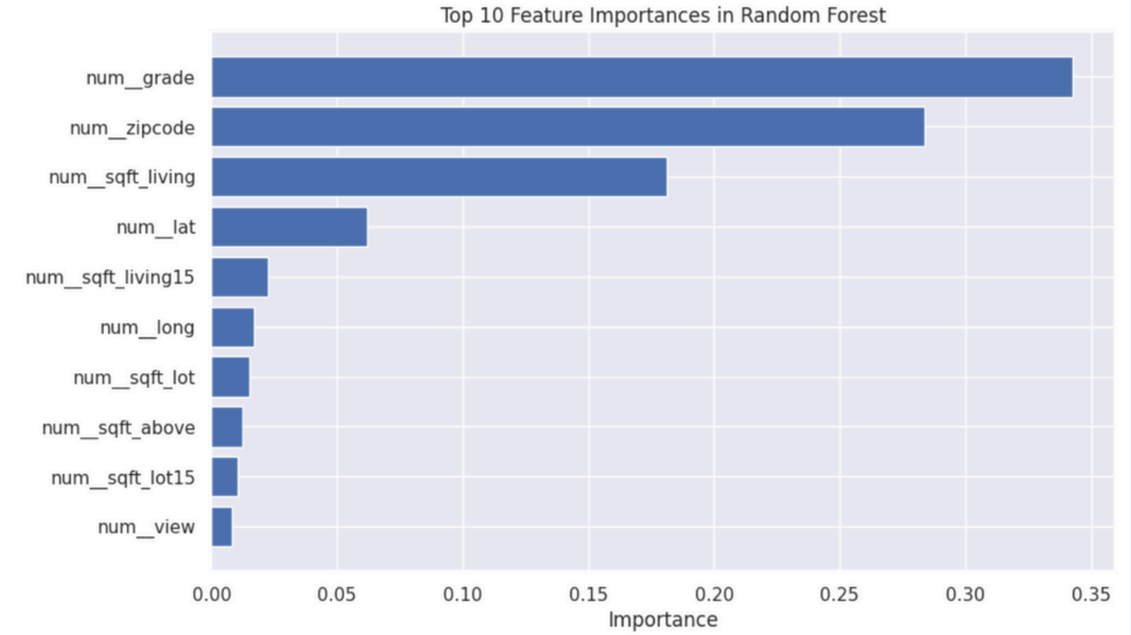
R^2 = 0.8738

- zipcode
- grade
- sqft_above
- sqft_living
- sqft_living15

Feature and Model Selection

Model 6: Random Forest

```
Model: Random Forest
  RMSE: 0.18645061718641293
  R²: 0.8737835485720991
Top Random Forest features: ['num__view' 'num__sqft_lot15' 'num__sqft_above' 'num__sqft_lot'
 'num__long' 'num__sqft_living15' 'num__lat' 'num__sqft_living'
 'num__zipcode' 'num__grade']
Top 10 Features by Importance:
               Feature  Importance
8           num__grade       0.343
13        num__zipcode       0.284
2     num__sqft_living       0.181
14            num__lat       0.062
16  num__sqft_living15       0.022
15           num__long       0.017
3        num__sqft_lot       0.015
9      num__sqft_above       0.012
17     num__sqft_lot15       0.010
6            num__view       0.008
```



- capture complex nonlinear relationships
- aggregate multiple decision trees, better resist noise and outliers in the data
- evaluate the importance of features and automatically select the features that contribute most to the prediction results

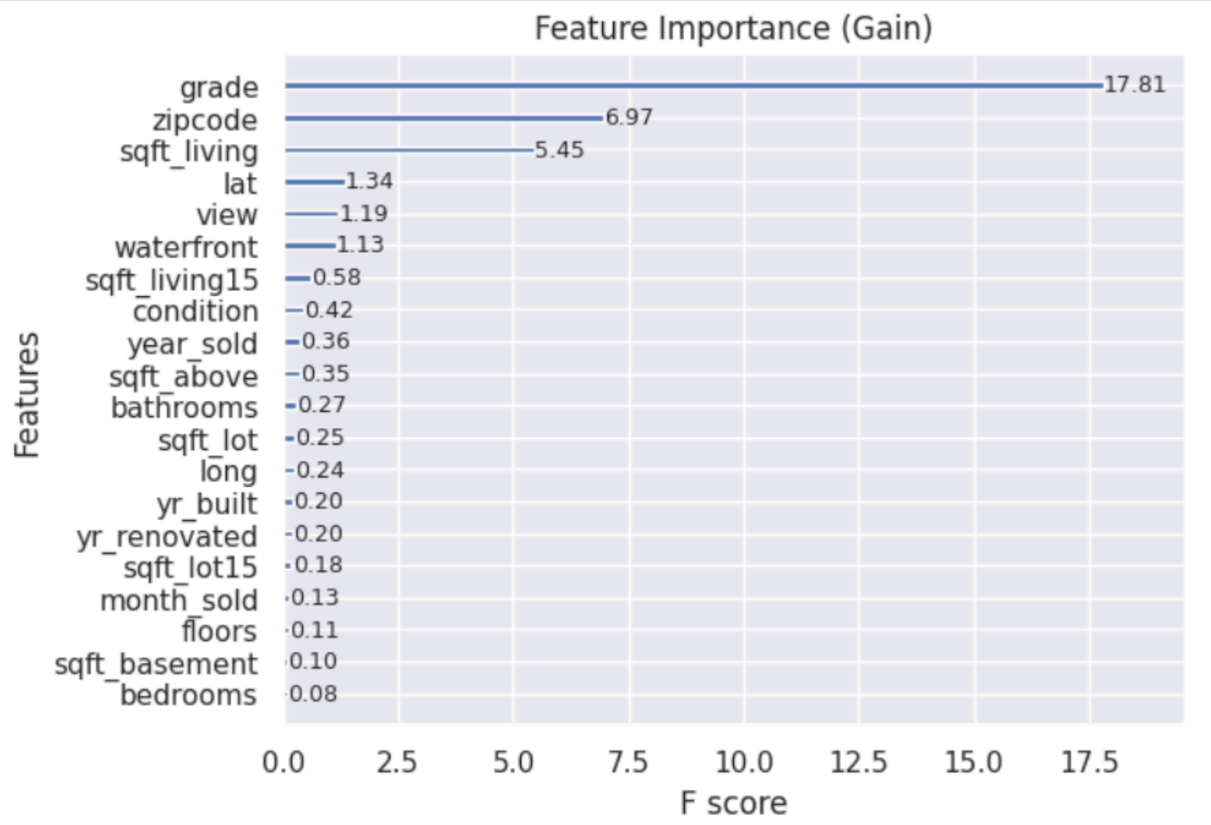
RMSE = 0.1865

R^2 = 0.8738

- grade
- zipcode
- sqft_living
- lat
- sqft_living15

Feature and Model Selection

Model 7: XGBoost



- ensemble multiple weak learners to improve the model's predictive performance
- fast training and prediction
- built-in regularization to prevent overfitting
- capture complex nonlinear relationships

RMSE = 0.15829

R^2 = 0.91027

- grade
- zipcode
- sqft_living
- lat
- view

Final Model Performance

Model	Description	Feature Select	RMSE	R^2
1	Linear Regression	all features without id	0.1863	0.8740
2	Linear Regression with Multicollinearity Handling	without sqft_above, sqft_basement, sqft_living15, bathrooms	0.1920	0.8670
3	Linear Regression with Multicollinearity Handling and Removal of Insignificant Variables	without sqft_above, sqft_basement, sqft_living15, bathrooms, floors, sqft_lot15	0.1919	0.8670
4	Lasso	top5 features: zipcode, sqft_living, grade, sqft_living15, lat	0.1865	0.8737
5	Ridge	top5 features: zipcode, grade, sqft_above, sqft_living, sqft_living15	0.1865	0.8738
6	Random Forest	top5 features: grade, zipcode, sqft_living, lat, sqft_living15	0.1865	0.8738
7	XGBoost	top5 features: grade, zipcode, sqft_living, lat, waterfront	0.1583	0.9103

- RMSE: The average error between the predicted values and the actual values when predicting the logarithm of house prices is approximately 0.1583.
- R^2: Approximately 91.03% of the variation in house prices can be explained by the features in the model.

Findings, and Learnings

Findings: Important Features for Determining the House Price

- Grade: [level of construction and design]

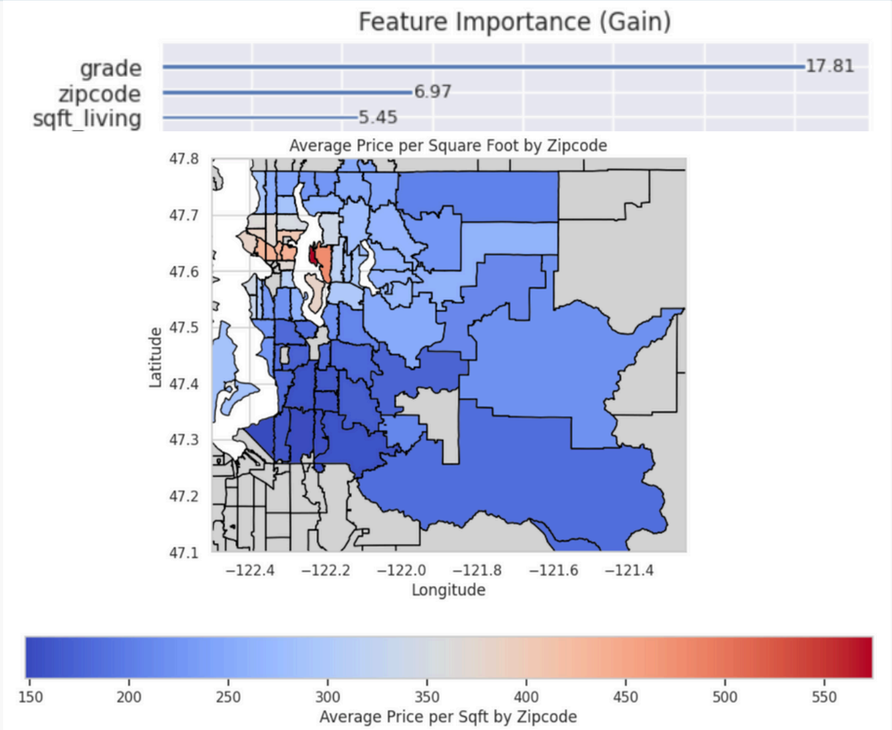
It is reasonable that a house in good construction and design quality will be more attractive to consumers, resulting in a higher price for sale.

- Zipcode: [the average price per square meter in the area]

Represent the geographic location, areas close to city centers, lakes, or transportation hubs typically have higher property values.

- sqft_living: [Square footage of the apartments interior living space]

The larger the house, the greater amount of money customers need to pay.



Learning:

- Apply the knowledge learned to real-world problems. Understand the basic workflow and tools of data analysis.
- Combine real-world considerations with analysis.

For example, for zero values in the bedrooms or bathrooms, instead of simply removing or replacing, we consider whether such cases might realistically exist. In this dataset, houses with zero bedrooms or bathrooms actually have a large area, so they are considered unreasonable.

Reference

Developer Ashish (2021). Predict house prices in King County USA using Machine Learning!! Data Science Regression Project. [online] YouTube. Available at: https://www.youtube.com/watch?v=7j0TZ45tJmo [Accessed 26 Oct. 2024].

MIFTAHUL ADIB (2024). Housing Price Regression: Top 4%. [online] Kaggle.com. Available at: https://www.kaggle.com/code/miftahuladib/housing-price-regression-top-4/notebook [Accessed 26 Oct. 2024].

NOUSSAIR MIGHRI (2024). House Price Prediction: Top 4 %. [online] Kaggle.com. Available at: https://www.kaggle.com/code/noussairmighri/house-price-prediction-top-4 [Accessed 26 Oct. 2024].Wang, Y. and Zhao, Q. (2022).

House Price Prediction Based on Machine Learning: a Case of King County. [online] ResearchGate. Available at: https://www.researchgate.net/publication/363370412_House_Price_Prediction_Based_on_Machine_Learning_A_Case_of_King_County [Accessed 25 Oct. 2024].

Thank you