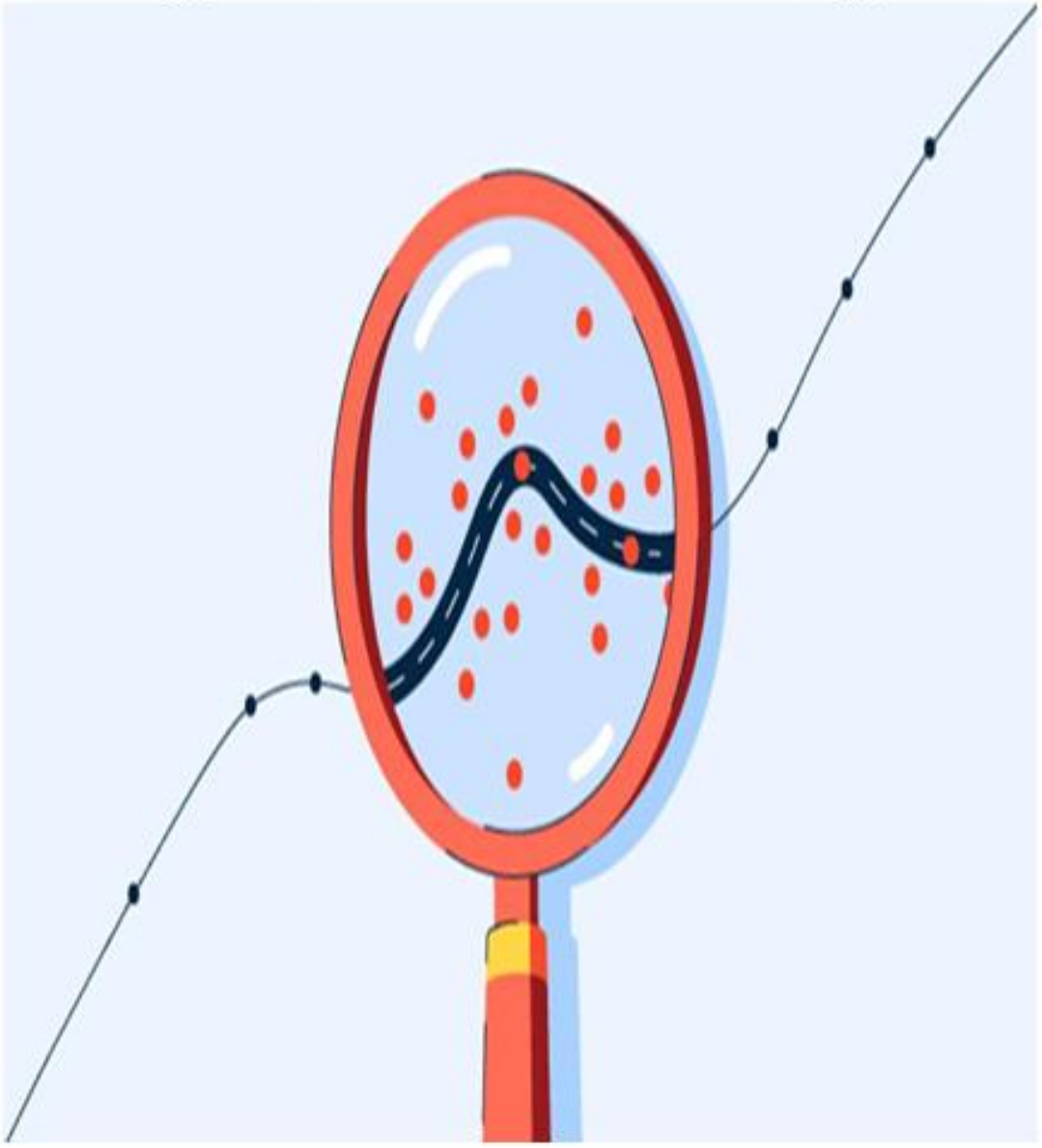
# Regression Analysis



**Presented by:**
Vindeshwari
Hariom
Satyam
Naveen
Sujan
Shivani

**Guided by:**
Prof. Monika Bhattacharjee

# REGRESSION ANALYSIS


## Acknowledgement

We would like to express our special thanks of gratitude to our respected teacher Professor Monika Bhattacharjee who has given us this golden opportunity to do this wonderful project on the topic “Regression Analysis on the real world dataset” which has helped us to do a lot of research and we have come to know a deep understanding of the topic.

Secondly, we would like to thank our senior Sonam and friends which has helped a lot in finishing this project within a limited time.

Lastly, we would like to thank everyone for the wonderful completion of this project. Special thanks to Kaustav.
Thank you ,

Vindeshwari Prasad Maurya
HariOm Gupta
Sujan Mandal
Naveen Kumar
Satyam Singh Kurmi
Shivani Singh

# CHAPTER-1

## Objective

The primary objective of this regression analysis is to construct a robust predictive model capable of accurately estimating house prices based on a variety of structural, locational, and temporal features. Drawing from a dataset that includes variables such as the number of bedrooms and bathrooms, living area and lot size, property condition, year built or renovated, and geographic identifiers like city and state, the goal is to explore and model the underlying relationships between these predictors and the target variable “price”. Through a methodical exploration of the data, rigorous preprocessing, and careful feature engineering, the project aims to ensure that the resulting model demonstrates both reliability and strong predictive performance. By adhering to established regression methodologies and validating the model through diagnostics and performance evaluation, the analysis seeks to uncover meaningful insights into the housing market and support informed decision-making for buyers, sellers, and real estate professionals .

.

# CHAPTER-2

## Dataset Description

**About the Dataset**: The dataset contains 4600 rows and 17 Columns. Out of which 7 columns are numerical and rest all are categorical. The first five rows of the dataset are given below:

| index | date | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | sqft_above | sqft_basement | yr_built | yr_renovated | street | city | statezip | country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2014-05-02 00:00:00 | 313000.0 | 3.0 | 1.5 | 1340 | 7912 | 1.5 | 0 | 0 | 3 | 1340 | 0 | 1955 | 2005 | 18810 Densmore Ave N | Shoreline | WA 98133 | USA |
| 1 | 2014-05-02 00:00:00 | 2384000.0 | 5.0 | 2.5 | 3650 | 9050 | 2.0 | 0 | 4 | 5 | 3370 | 280 | 1921 | 0 | 709 W Blaine St | Seattle | WA 98119 | USA |
| 2 | 2014-05-02 00:00:00 | 342000.0 | 3.0 | 2.0 | 1930 | 11947 | 1.0 | 0 | 0 | 4 | 1930 | 0 | 1966 | 0 | 26206-26214 143rd Ave SE | Kent | WA 98042 | USA |
| 3 | 2014-05-02 00:00:00 | 420000.0 | 3.0 | 2.25 | 2000 | 8030 | 1.0 | 0 | 0 | 4 | 1000 | 1000 | 1963 | 0 | 857 170th Pl NE | Bellevue | WA 98008 | USA |
| 4 | 2014-05-02 00:00:00 | 550000.0 | 4.0 | 2.5 | 1940 | 10500 | 1.0 | 0 | 0 | 4 | 1140 | 800 | 1976 | 1992 | 9105 170th Ave NE | Redmond | WA 98052 | USA |

About the columns: We can see that most of the columns are Categorical and 7 columns seem to be numerical. The datatypes of the columns are given below.

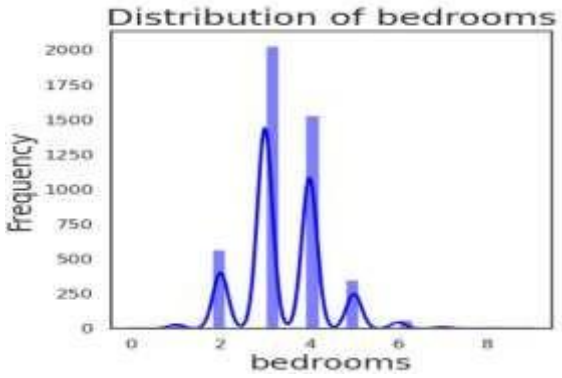
| Column | Type | Category |
|---|---|---|
| index | Integer | Numerical |
| date | Datetime | Temporal |
| price | Float | Numerical |
| bedrooms | Float | Categorical |
| bathrooms | Float | Categorial |
| sqft_living | Integer | Numerical |
| sqft_lot | Integer | Numerical |
| floors | Float | Categorical |
| waterfront | Integer | Categorical |
| view | Integer | Categorical |
| condition | Integer | Categorical |
| sqft_above | Integer | Numerical |
| sqft_basement | Integer | Numerical |
| Yr_built | integer | Numerical |
| Yr_renovated | integer | Numerical |
| Street | string | Categorical |
| city | string | Categorical |
| statezip | string | Categorical |
| country | string | Categorical |

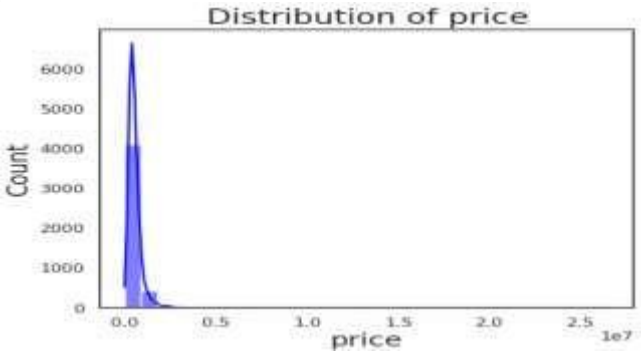
# Basic Information about data

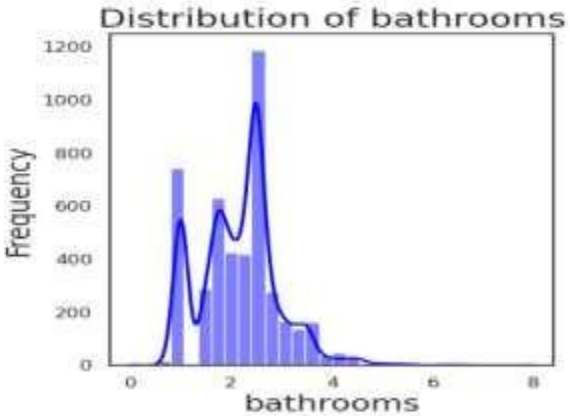
| | |
|---|---|
| date | 70 |
| price | 1741 |
| bedrooms | 10 |
| bathrooms | 26 |
| sqft_living | 566 |
| sqft_lot | 3113 |
| floors | 6 |
| waterfront | 2 |
| view | 5 |
| condition | 5 |
| sqft_above | 511 |
| sqft_basement | 207 |
| yr_built | 115 |
| yr_renovated | 60 |
| street | 4525 |
| city | 44 |
| statezip | 77 |
| country | 1 |

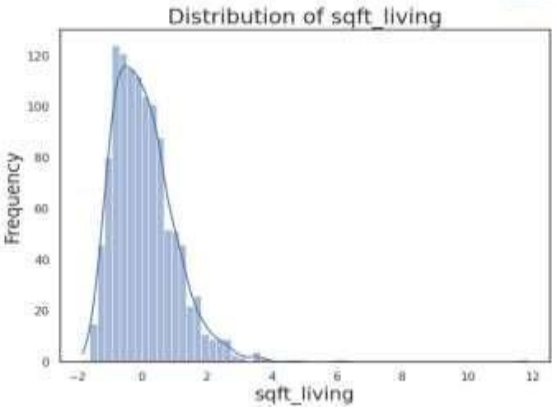
| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | sqft_above | sqft_basement | yr_built | yr_renovated |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 4.600000e+03 | 4600.000000 | 4600.000000 | 4600.000000 | 4.600000e+03 | 4600.000000 | 4600.000000 | 4600.000000 | 4600.000000 | 4600.000000 | 4600.000000 | 4600.000000 | 4600.000000 |
| mean | 5.519630e+05 | 3.400870 | 2.160815 | 2139.346957 | 1.485252e+04 | 1.512065 | 0.007174 | 0.240652 | 3.451739 | 1827.265435 | 312.081522 | 1970.786304 | 808.608261 |
| std | 5.638347e+05 | 0.908848 | 0.783781 | 963.206916 | 3.588444e+04 | 0.538288 | 0.084404 | 0.778405 | 0.677230 | 862.168977 | 464.137228 | 29.731848 | 979.414536 |
| min | 0.000000e+00 | 0.000000 | 0.000000 | 370.000000 | 6.380000e+02 | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 370.000000 | 0.000000 | 1900.000000 | 0.000000 |
| 25% | 3.228750e+05 | 3.000000 | 1.750000 | 1460.000000 | 5.000750e+03 | 1.000000 | 0.000000 | 0.000000 | 3.000000 | 1190.000000 | 0.000000 | 1951.000000 | 0.000000 |
| 50% | 4.609435e+05 | 3.000000 | 2.250000 | 1980.000000 | 7.683000e+03 | 1.500000 | 0.000000 | 0.000000 | 3.000000 | 1590.000000 | 0.000000 | 1976.000000 | 0.000000 |
| 75% | 6.549625e+05 | 4.000000 | 2.500000 | 2620.000000 | 1.100125e+04 | 2.000000 | 0.000000 | 0.000000 | 4.000000 | 2300.000000 | 610.000000 | 1997.000000 | 1999.000000 |
| max | 2.659000e+07 | 9.000000 | 8.000000 | 13540.000000 | 1.074218e+06 | 3.500000 | 1.000000 | 4.000000 | 5.000000 | 9410.000000 | 4820.000000 | 2014.000000 | 2014.000000 |

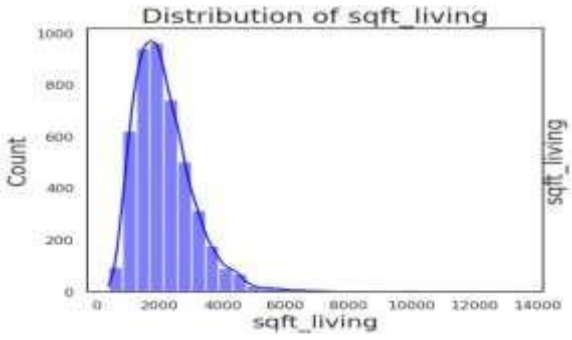
# Univariate Data Analysis:

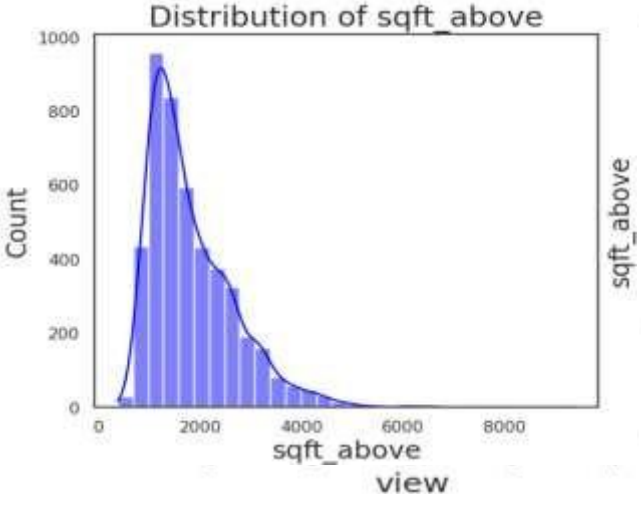


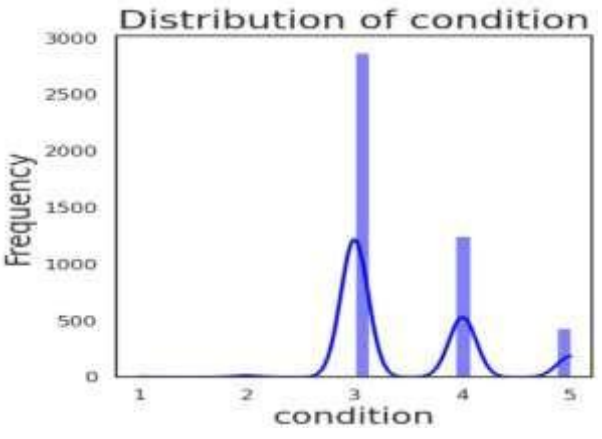



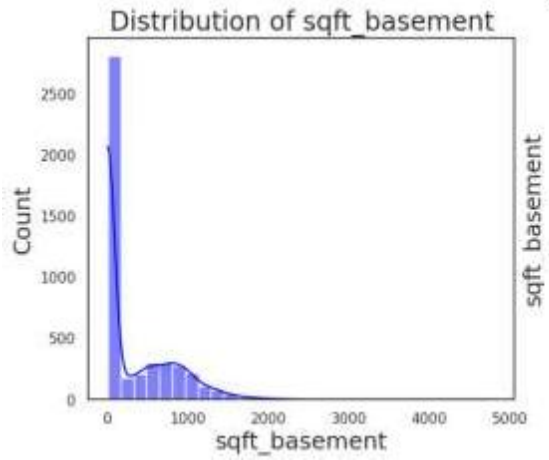



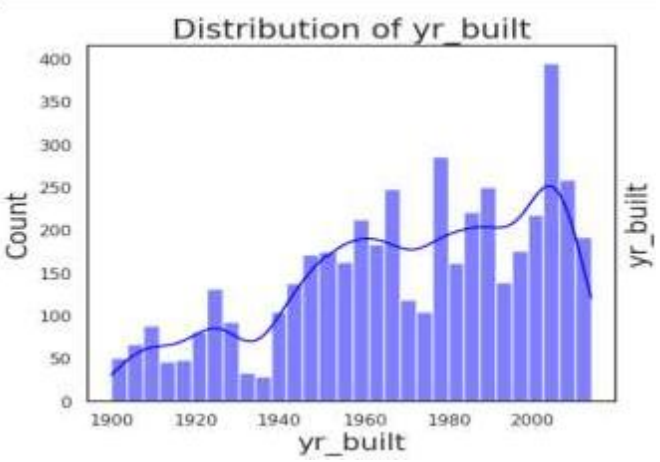



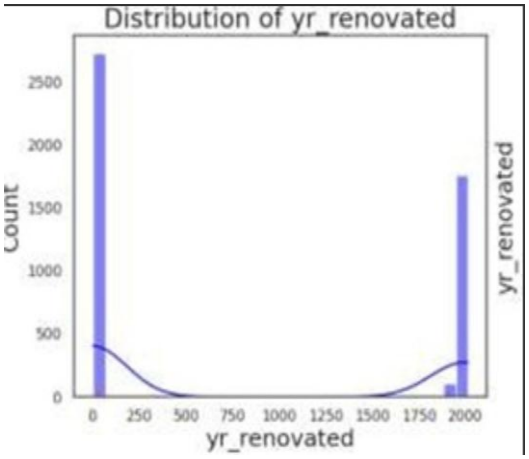



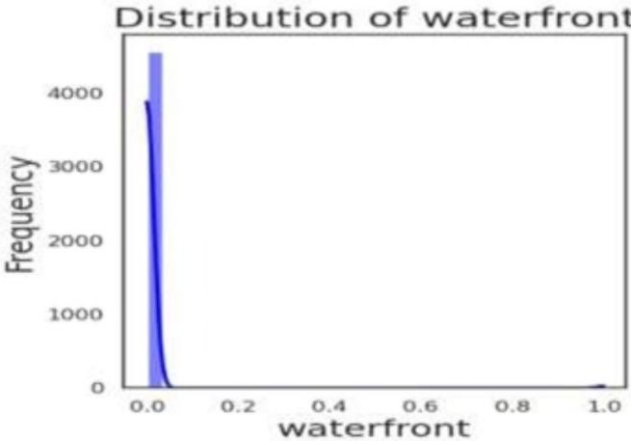




Distribution of sqft_basement
Count
sqft_basement
0
500
1000
1500
2000
2500
0
1000
2000
3000
4000
5000


Distribution of yr_built
Count
yr_built
0
50
100
150
200
250
300
350
400
1900
1920
1940
1960
1980
2000


Distribution of yr_renovated
Count
yr_renovated
0
500
1000
1500
2000
2500
0
250
500
750
1000
1250
1500
1750
2000


Distribution of waterfront
Frequency
waterfront
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0

# Classification of Data

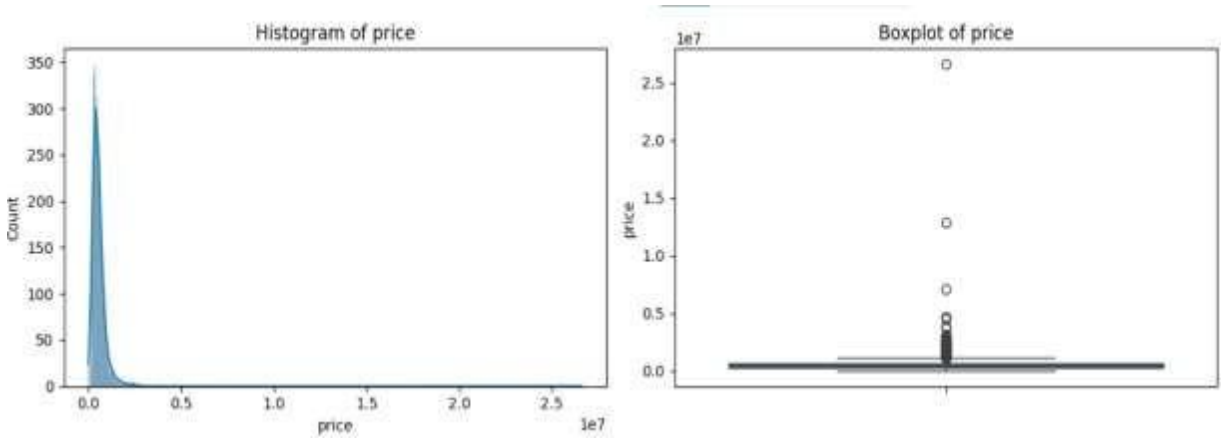
**Numerical Data:**
Price,sqft_living,sqft_lot,floors,sqft_above, sqft_basement, yr_built, yr_renovated

Numerical data typically shows a continuous distribution in the histogram, without clear separations between bars. The histogram have a smooth curve or shape, indicating the distribution of values across a range. There may be some variation in the height of bars, but overall the distribution appears continuous

**Categorical Data:**
waterfront, view, condition, street, city, statezip, country, bathrooms ,bedrooms .
Categorical data typically shows distinct bars or spikes in the histogram, indicating different categories or levels. The histogram have a discrete distribution with clear separation between the bars. Each bar represents the frequency or count of observation in each category .

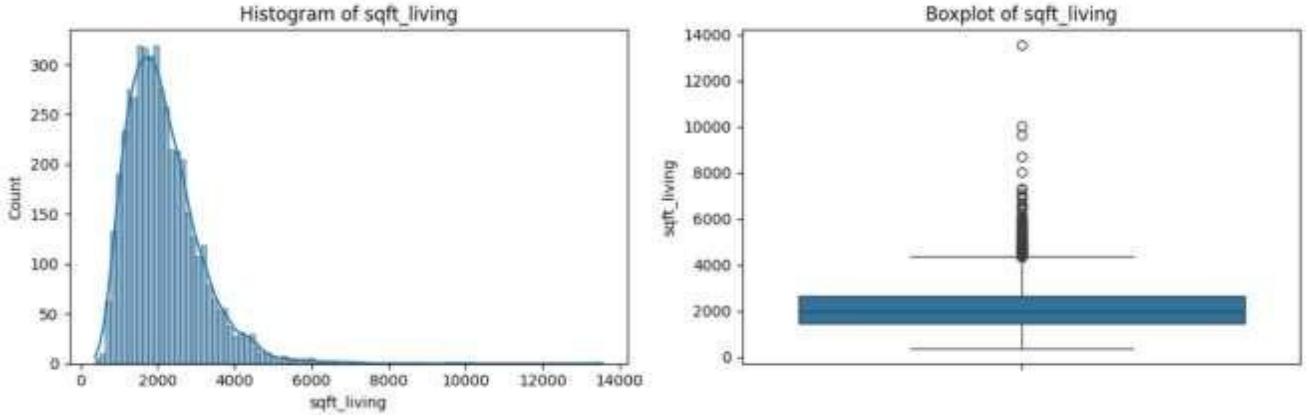
# Analysis of Numerical Columns

## Price




**Variable price** exhibits a right-skewed distribution, as shown in the histogram. The box plot reveals several high-end outliers, indicating the presence of extremely expensive properties or potential anomalies in the dataset.

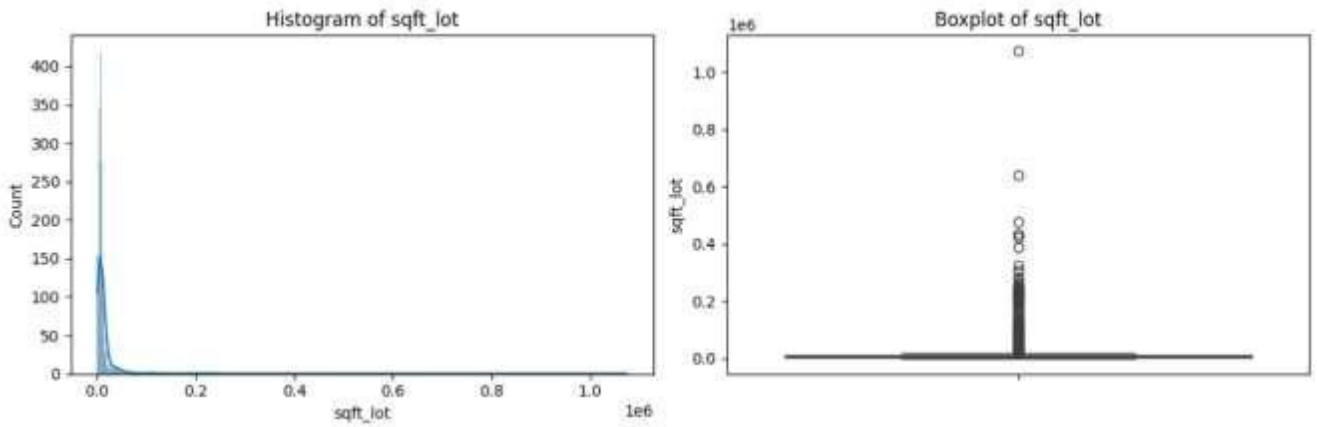
## Sqft_living




**Variable sqft_living** displays a right-skewed distribution, with a long tail toward higher values. The box plot indicates numerous high-end outliers, suggesting the presence of very large homes in the dataset.

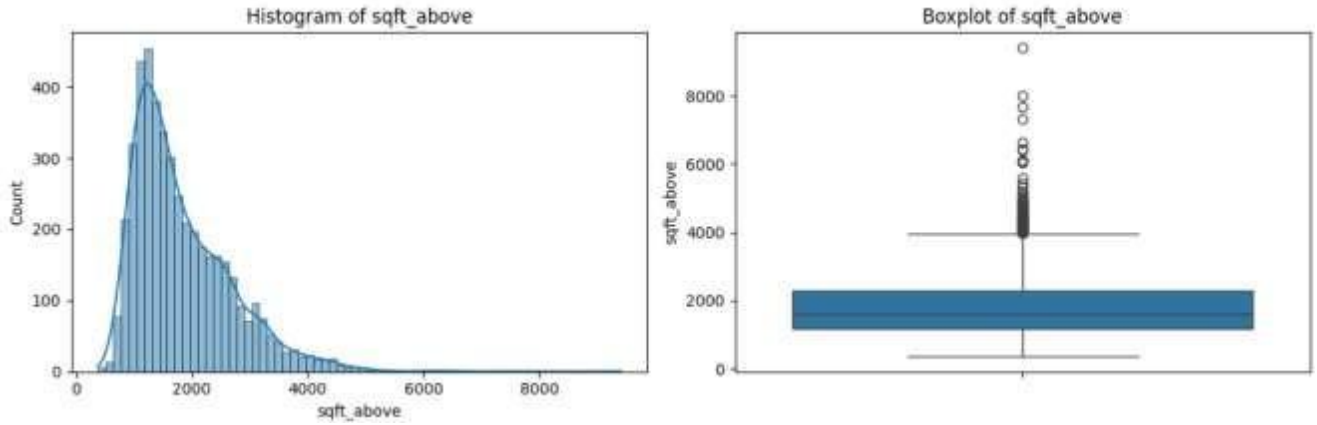
**Sqft_lot**

# Sqft_lot



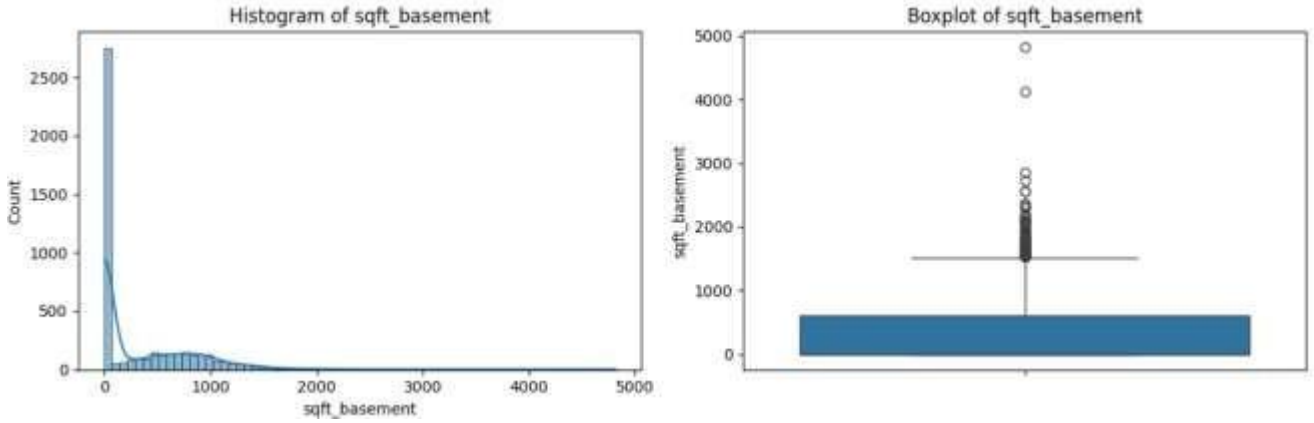

The variable **`sqft_lot`** exhibits a highly right-skewed distribution, as shown below in the histogram, with most properties having smaller lot sizes. The boxplot reveals a significant number of high-end outliers, suggesting the presence of unusually large lot sizes in the dataset.

# Sqft_above




The variable **sqft_above** shows a right-skewed distribution, with most values concentrated at lower ranges. The box plot highlights several high-end outliers, indicating the presence of unusually large homes that may influence statistical analyses or model performance .

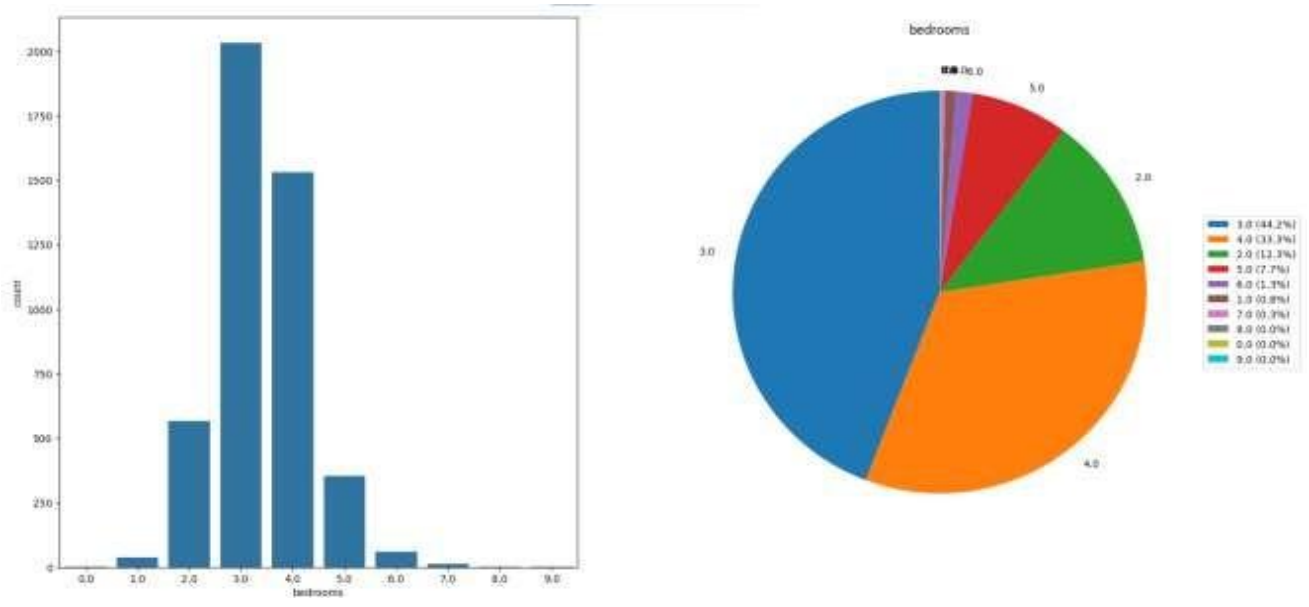
# Sqft_basement




Variable **sqft_basement** exhibits a right-skewed distribution, with a high concentration of zero values. The box plot reveals numerous high-end outliers, indicating the presence of properties with unusually large basement areas.

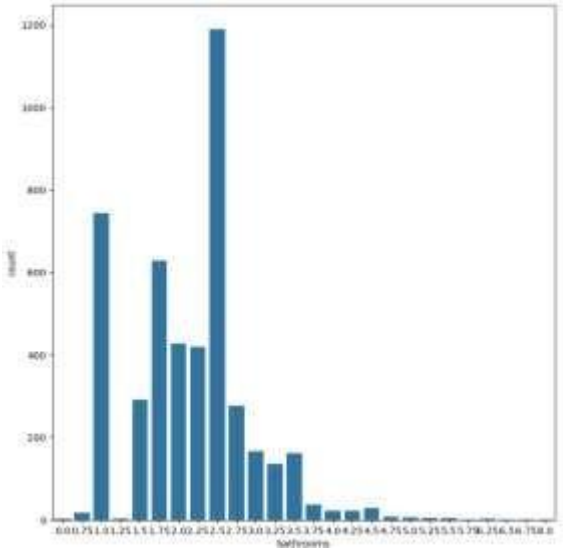
# Analysis of Categorical Variable

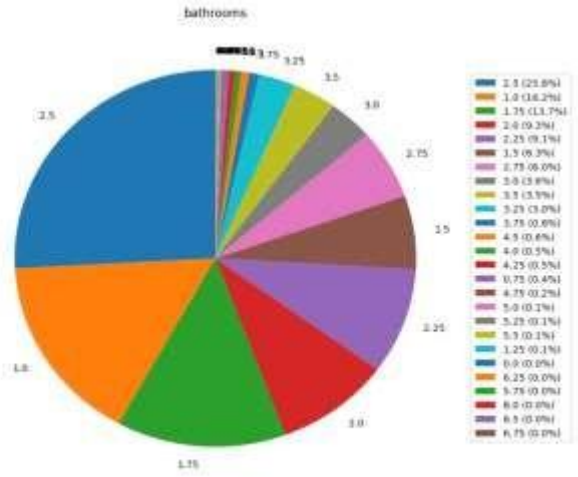
# Bedrooms



Based on the countplot and pie chart, the variable **bedrooms** exhibits an uneven distribution among its categories. Categories 3 and 4 are most frequent, accounting for 44.2% and 33.3% respectively.
Categories 2 and 5 follow with 12.3% and 7.7%. Categories 0, 1, 6, 7, 8, and 9 are minimally represented, each below 2%. This indicates that most homes have 3 or 4 bedrooms, while extreme values are rare.

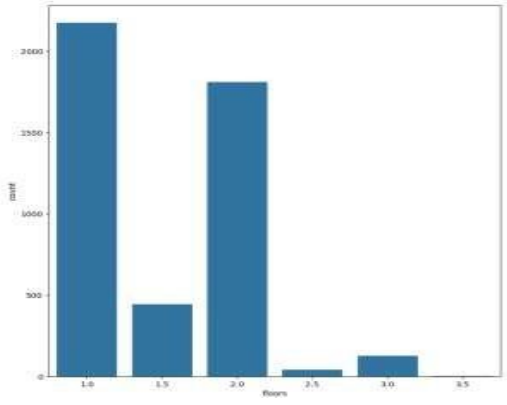
# Bathrooms

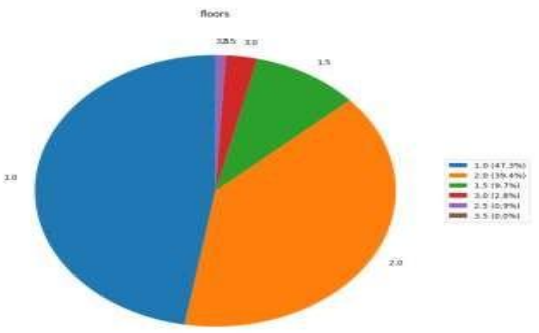





Based on the countplot and pie chart, the variable **`bathrooms`** displays a varied distribution with a peak at 2.5 bathrooms (25.8%), followed by 1.0 (16.2%) and 1.75 (13.7%). Other frequent values include 2.25 (9.1%) and 2.0 (9.0%). Higher values above 3.5 are rare, each contributing less than 1%. This suggests that most properties have between 1 and 2.5 bathrooms, while extreme counts are uncommon.

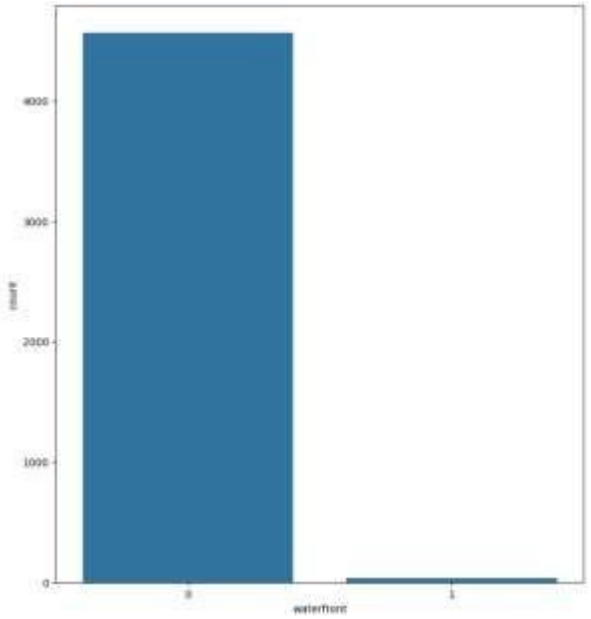
# Floors

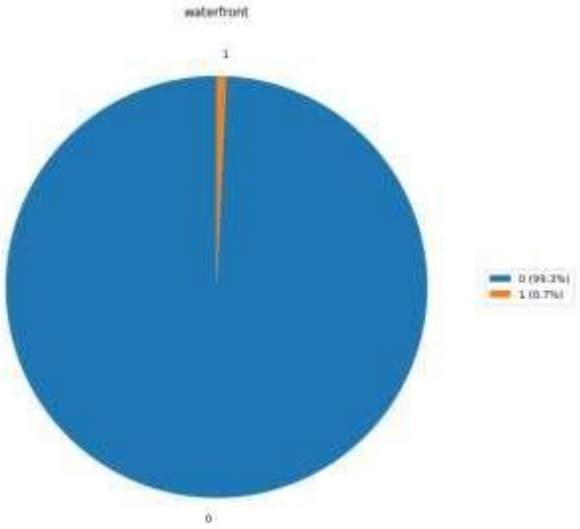




The histogram and pie chart show that the variable **"floors"** is heavily concentrated at 1.0 and 2.0, with 47.3% and 39.4% of the data falling in these categories, respectively. Other values like 1.5, 3.0, and 2.5 make up a small proportion of the data, indicating a skewed distribution with a few outliers.

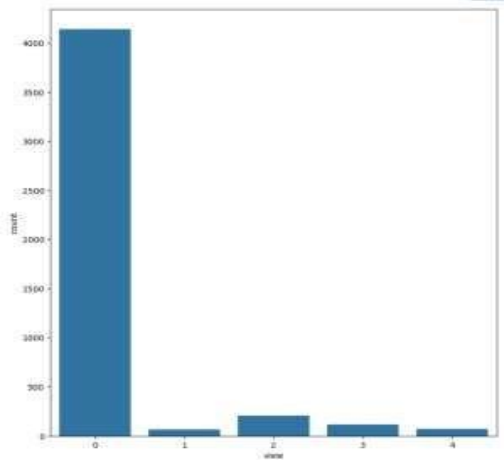
# Waterfront

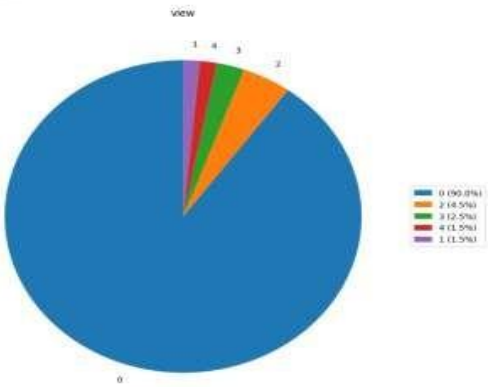



The histogram and pie chart show that the **"waterfront"** variable is heavily imbalanced, with 99.3% of the data classified as 0 (not on the waterfront) and only 0.7% classified as 1 (on the waterfront), indicating a highly skewed distribution.

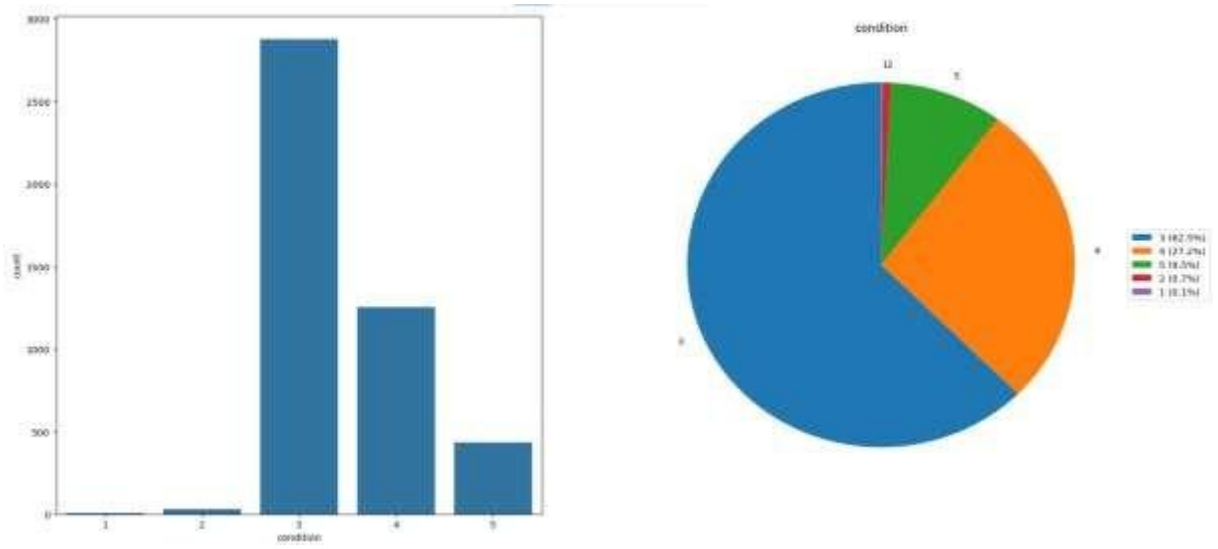
# View



Based on the countplot and pie chart, it's evident that the **"view"** is predominantly composed of category 0, which accounts for 90% of the occurrences. Categories 2, 3, and 4 make up a small portion, while categories 1 and 4 are relatively rare. This suggests that "view" is heavily skewed towards category 0, with other categories being uncommon occurrences in the dataset.

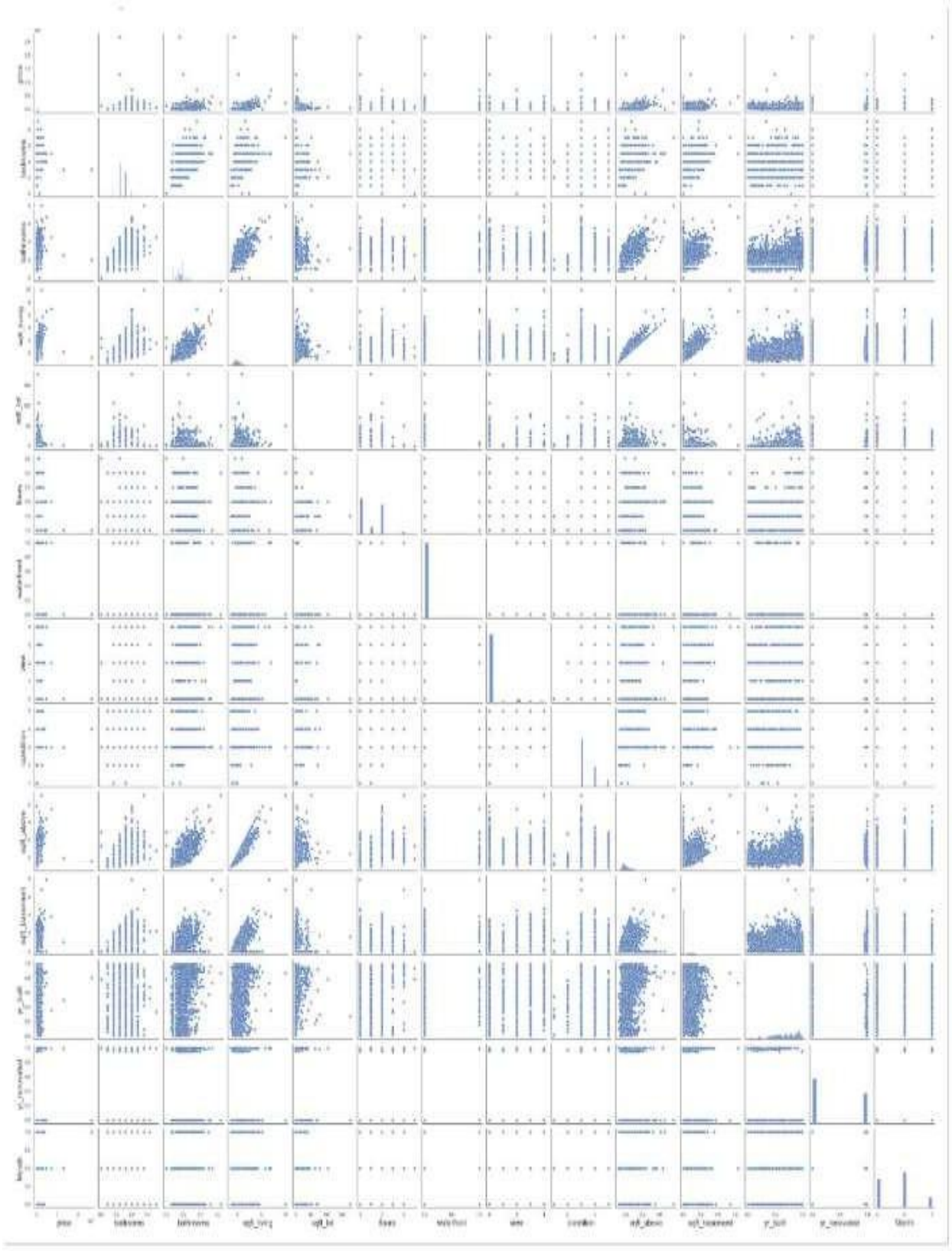
# Condition

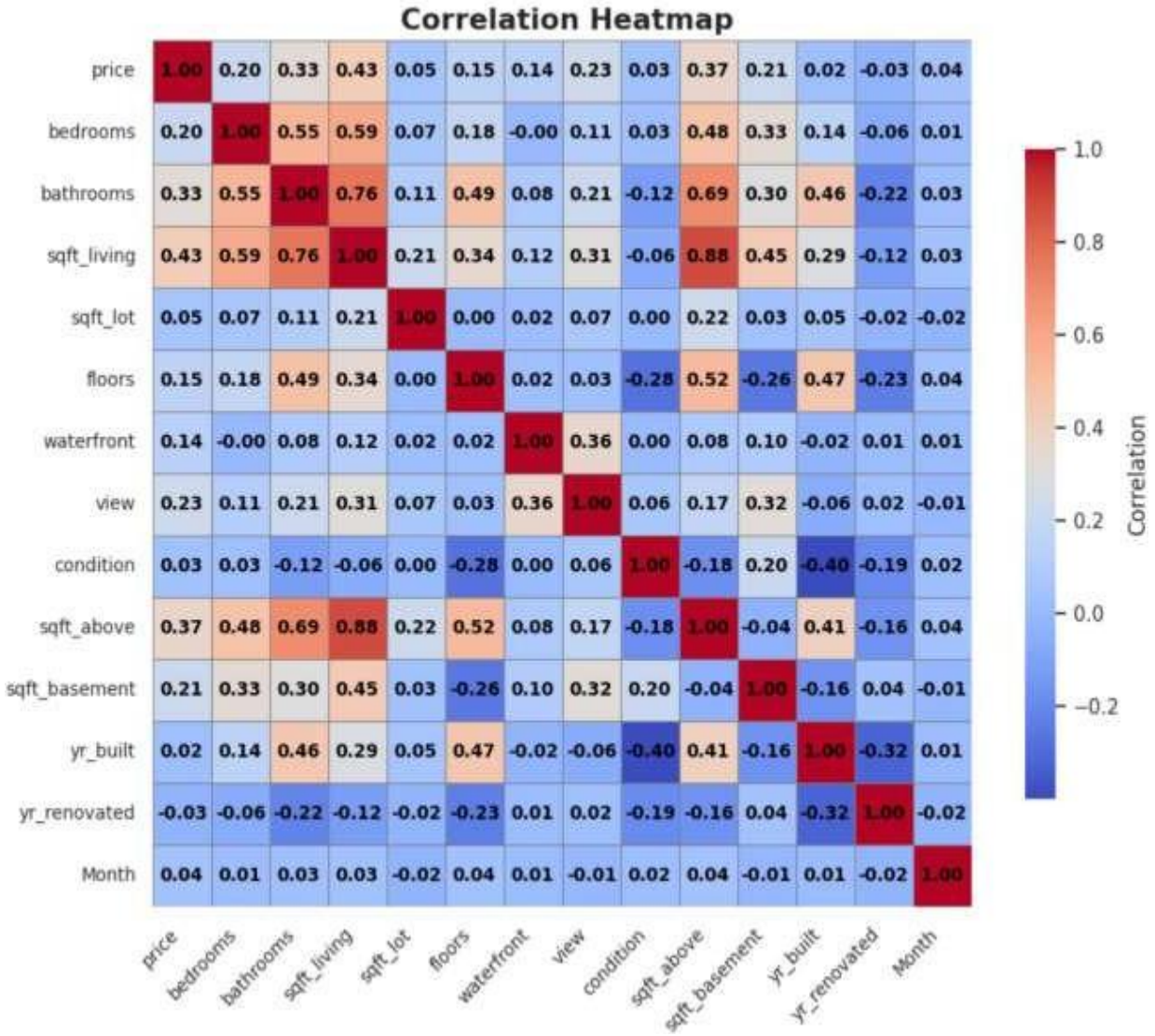


Based on the countplot and pie chart, it's clear that the variable **"condition"** is predominantly composed of category 3, which accounts for 62.5% of the data. Categories 4 and 5 make up 27.2% and 9.5%, respectively, while categories 1 and 2 are quite rare, making up only 0.1% and 0.7% of the dataset. This indicates that the "condition" variable is towards category 3, with categories 1 and 2 being outliers .

# Bivariate Data Analysis

- **Pairplot:**






Based on the correlations matrix, we can draw several inferences:

Strong Positive Correlations:
sqft_living & bathrooms (0.76): Larger houses tend to have more bathrooms.

sqft_living & sqft_above (0.88): As expected, total living area and above-ground area are highly correlated.

sqft_above & bathrooms (0.69): More above-ground space often includes more bathrooms.

price & sqft_living (0.43): Price increases with living area, a moderate positive correlation.

price & bathrooms (0.33): More bathrooms generally indicate higher price.

price & sqft_above (0.37): Above-ground square footage contributes to price.

Moderate Positive Correlations:
bedrooms & sqft_living (0.59): More bedrooms typically mean more square footage.

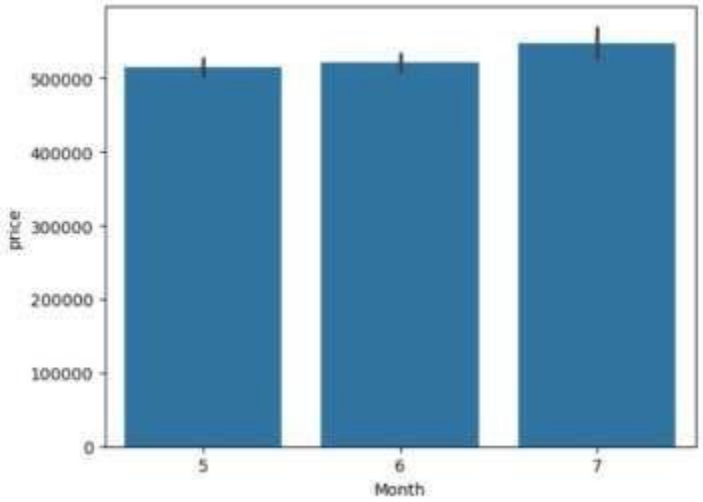
bedrooms & bathrooms (0.55): Homes with more bedrooms tend to have more bathrooms.

price & view (0.23): A better view tends to increase price.

price & waterfront (0.26): Waterfront properties are typically more expensive

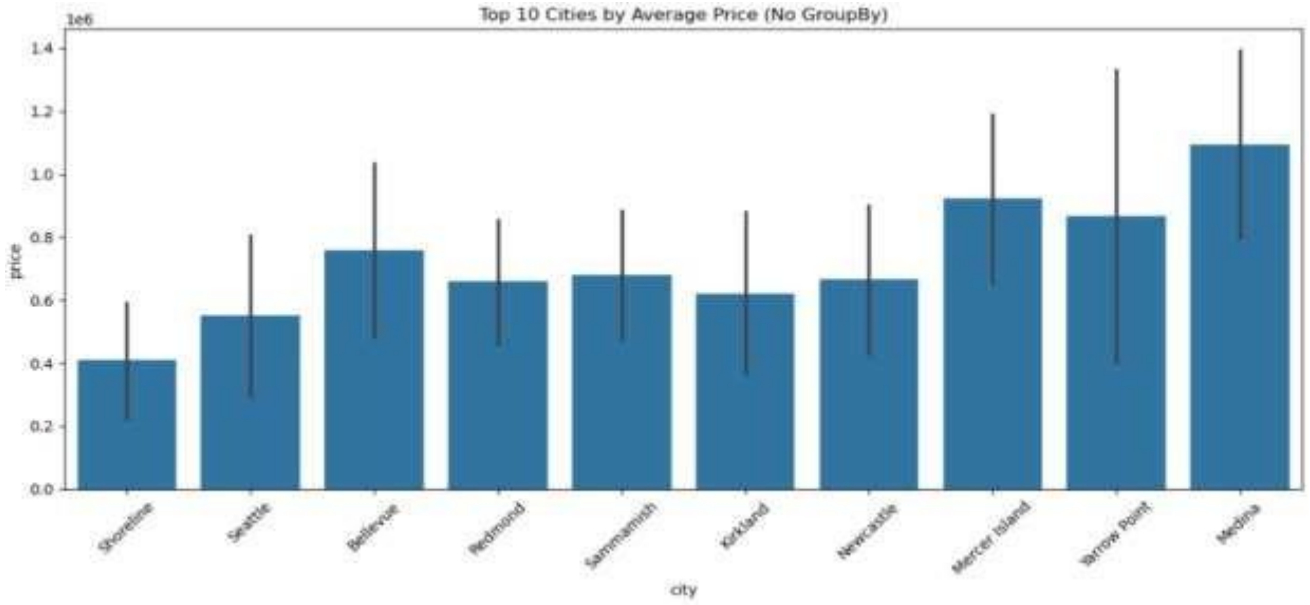
# Bivariate Analysis of Price vs Categorical Variables

## Price Vs Month



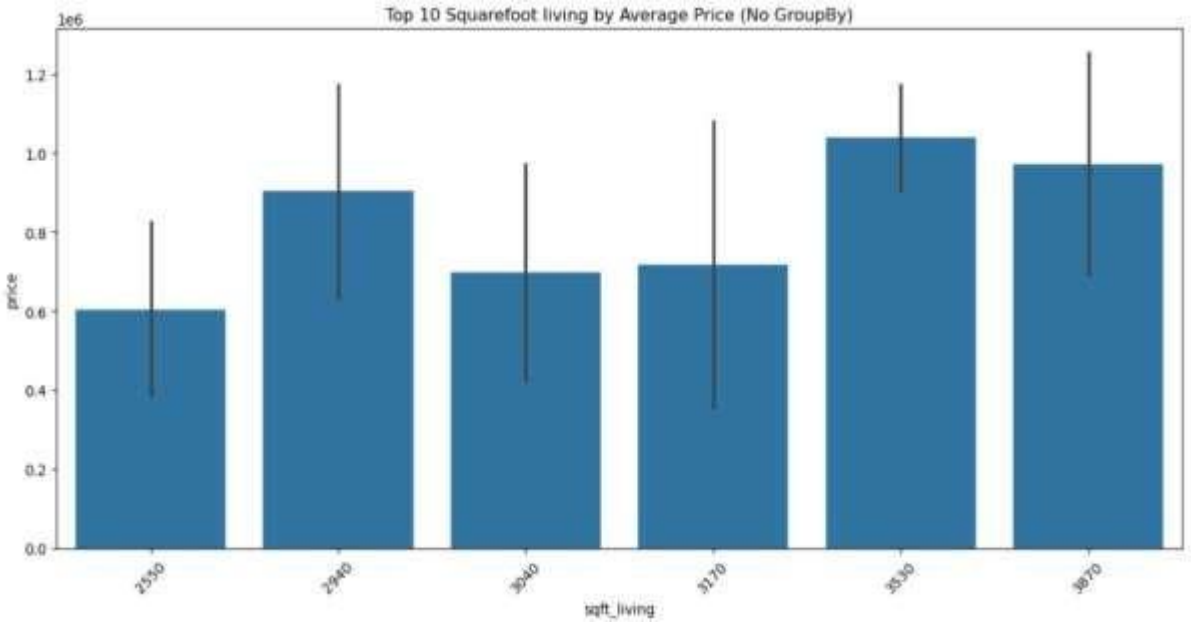

The bar plot shows that the average house prices for May, June, and July are similar, with prices around 500,000. The small error bars indicate low variability in prices across these months, suggesting stability in house prices during this period.

## Price Vs City



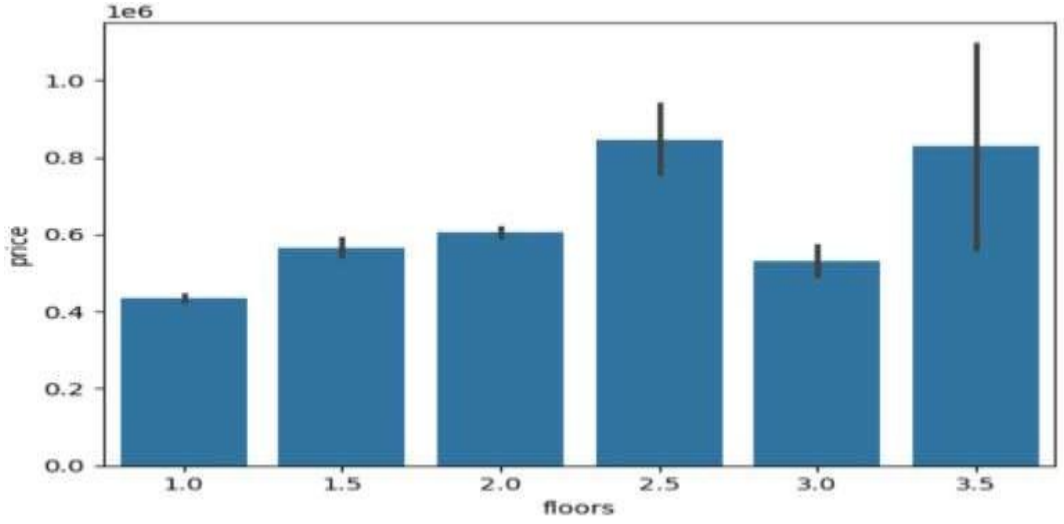
The bar plot shows that Medina has the highest average house price, followed by Yarrow Point and Mercer Island. The error bars indicate moderate variability in prices, suggesting some fluctuation within cities, but overall, the trends remain consistent

## Price Vs sqft_living



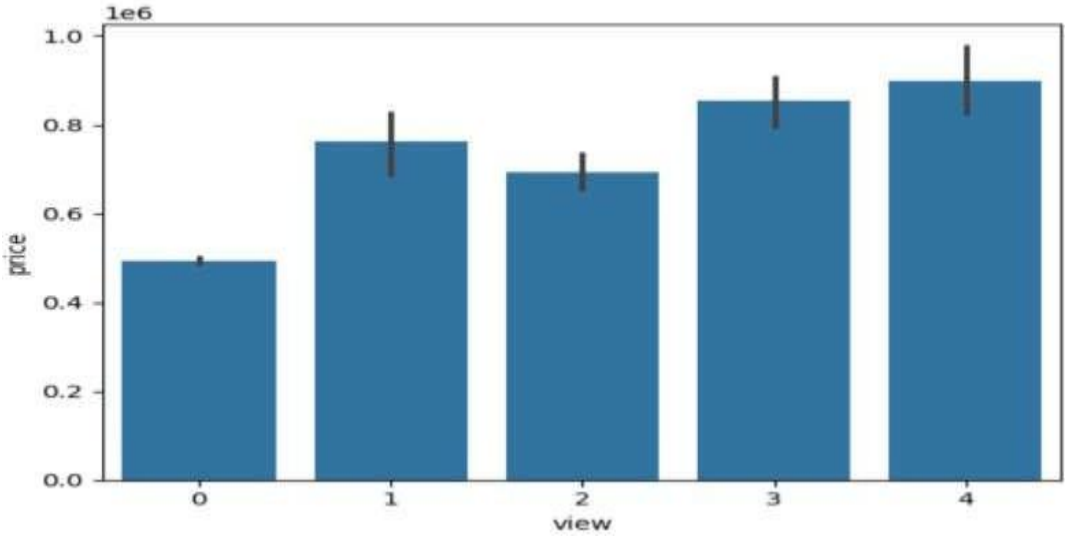

The bar plot shows that the average house prices increase with the square footage of the living space, with larger homes (e.g., 3530 and 3870 sqft) having higher average prices. The error bars suggest moderate variability in prices for each square footage category, indicating that while larger homes tend to have higher prices, there is still some fluctuation within each group.

## Price Vs floors



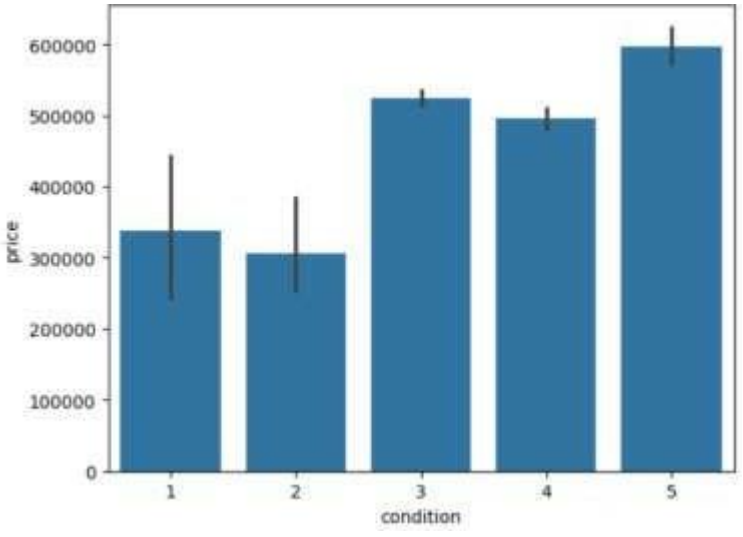
The bar plot shows the average house prices for different floor counts, with prices increasing as the number of floors rises. The error bars indicate that there is some variability in prices for each floor category, but the overall trend shows higher prices for homes with more floors. Homes with 3.5 floors have the highest average price, while those with fewer floors, especially 1.0 floor, have lower prices.

## Price Vs View




The bar plot shows the average house prices for different view ratings, with prices generally increasing as the view rating improves. The error bars indicate some variability in prices within each view category, but the overall trend suggests that better views are associated with higher house prices. Homes with a view rating of 4 have the highest average price, while those with a rating of 0 have the lowest.

## Price Vs Condition



The bar plot displays the average prices for different conditions, with prices generally increasing as the condition improves. The error bars indicate some variability in prices within each condition category, but the overall trend suggests better conditions are linked to higher prices. Homes with a condition of 5 have the highest average price, while those with a condition of 1 have the lowest.

# Univariate Regression

# Price vs Square foot living

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.422 |
| Model: | OLS | Adj. R-squared: | 0.421 |
| Method: | Least Squares | F-statistic: | 842.2 |
| Date: | Tue, 15 Apr 2025 | Prob (F-statistic): | 1.68e-139 |
| Time: | 10:47:32 | Log-Likelihood: | -15749. |
| No. Observations: | 1156 | AIC: | 3.150e+04 |
| Df Residuals: | 1154 | BIC: | 3.151e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

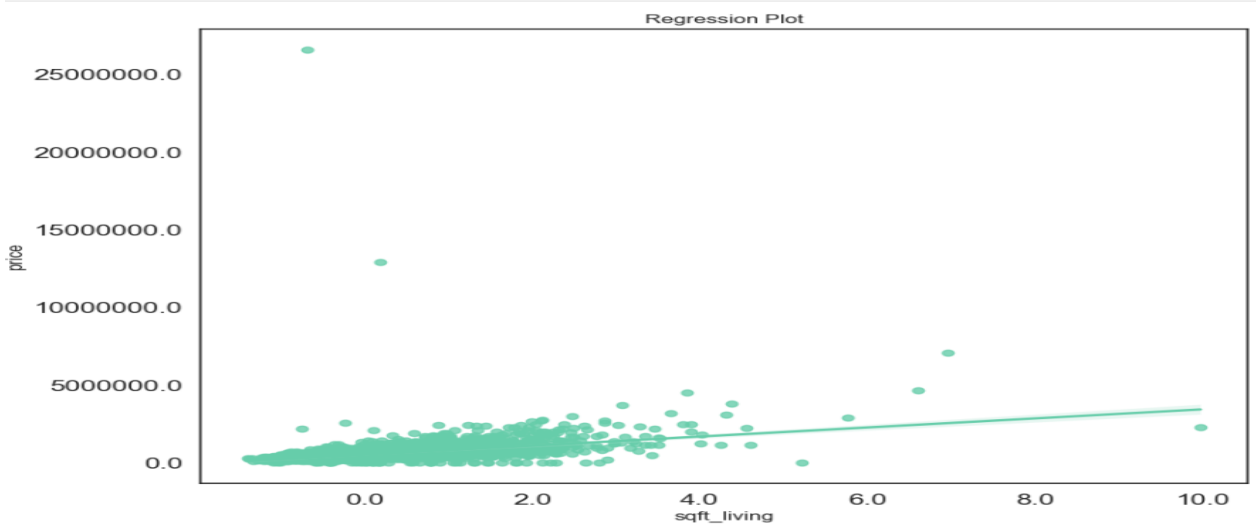
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.451e+05 | 6030.555 | 90.389 | 0.000 | 5.33e+05 | 5.57e+05 |
| sqft_living | 2.065e+05 | 7116.077 | 29.022 | 0.000 | 1.93e+05 | 2.2e+05 |

| | | | |
|---|---|---|---|
| Omnibus: | 492.363 | Durbin-Watson: | 1.933 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 20905.952 |
| Skew: | 1.238 | Prob(JB): | 0.00 |
| Kurtosis: | 23.686 | Cond. No. | 1.32 |

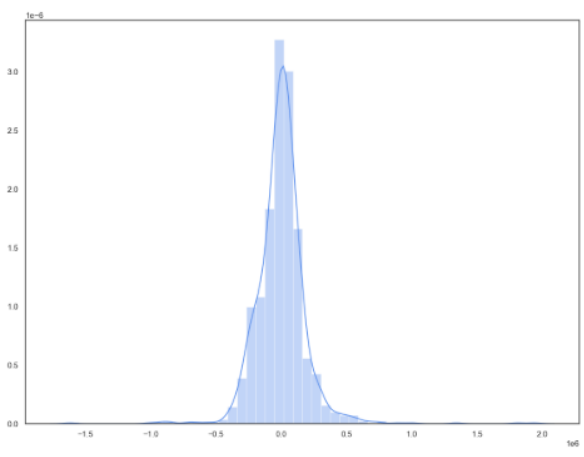
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

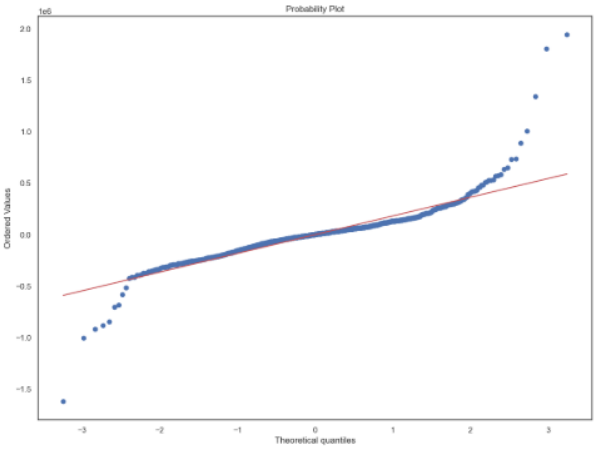
163

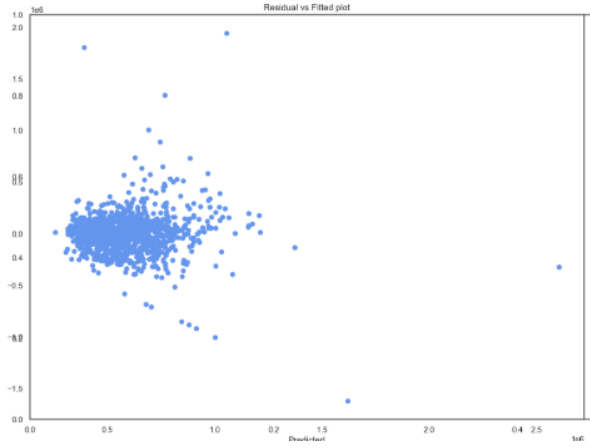
A simple linear regression was conducted to examine the impact of living area (sqft_living) on house prices. The model yielded an R-squared value of **0.422**, indicating that about **42.2%** of the variability in house prices is explained by the size of the living space.The coefficient for sqft_living is **206,500**, suggesting a significant positive relationshipwhich means larger homes tend to have higher prices. The **p-values** for both the intercept and sqft_living are **highly significant (p < 0.001)**, confirming the reliability of the results.

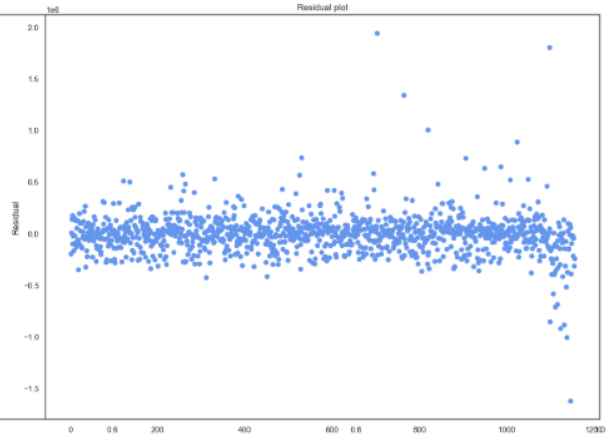



The plot shows a positive linear relationship between sqft_living and house price. As the living area increases, the price also increases. Most points are close to the fitted line, indicating a good model fit, though some outliers are present at higher values.





Model diagnostic plots were used to evaluate the assumptions of the linear regression model.

1. Residual distribution (top left):
   The histogram shows that the residuals are approximately normally distributed and centered around zero. This supports the normality assumption, although a slight skew is observed.
2. Q-Q plot (top right):
   The residuals generally follow the 45-degree line, indicating they are roughly normally distributed. Some deviations at the tails suggest the presence of a few outliers.
3. Residuals vs. fitted values (bottom left):
   The residuals are randomly scattered around zero, supporting the assumption of linearity. However, the spread increases at higher fitted values, suggesting potential heteroscedasticity (non-constant variance).
4. Residuals vs. observation order (bottom right):
   The residuals do not show any specific pattern over time, indicating there is no autocorrelation or time-based trend in the model errors.

# Price VS Square foot Above

OLS Regression Results

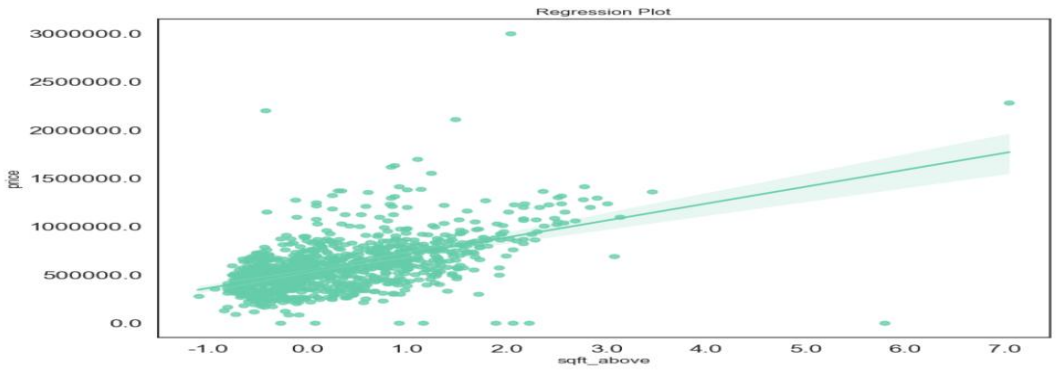
| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.295 |
| Model: | OLS | Adj. R-squared: | 0.294 |
| Method: | Least Squares | F-statistic: | 482.3 |
| Date: | Tue, 15 Apr 2025 | Prob (F-statistic): | 1.37e-89 |
| Time: | 10:47:39 | Log-Likelihood: | -15864. |
| No. Observations: | 1156 | AIC: | 3.173e+04 |
| Df Residuals: | 1154 | BIC: | 3.174e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.36e+05 | 6850.888 | 78.237 | 0.000 | 5.23e+05 | 5.49e+05 |
| sqft_above | 1.752e+05 | 7976.605 | 21.960 | 0.000 | 1.6e+05 | 1.91e+05 |

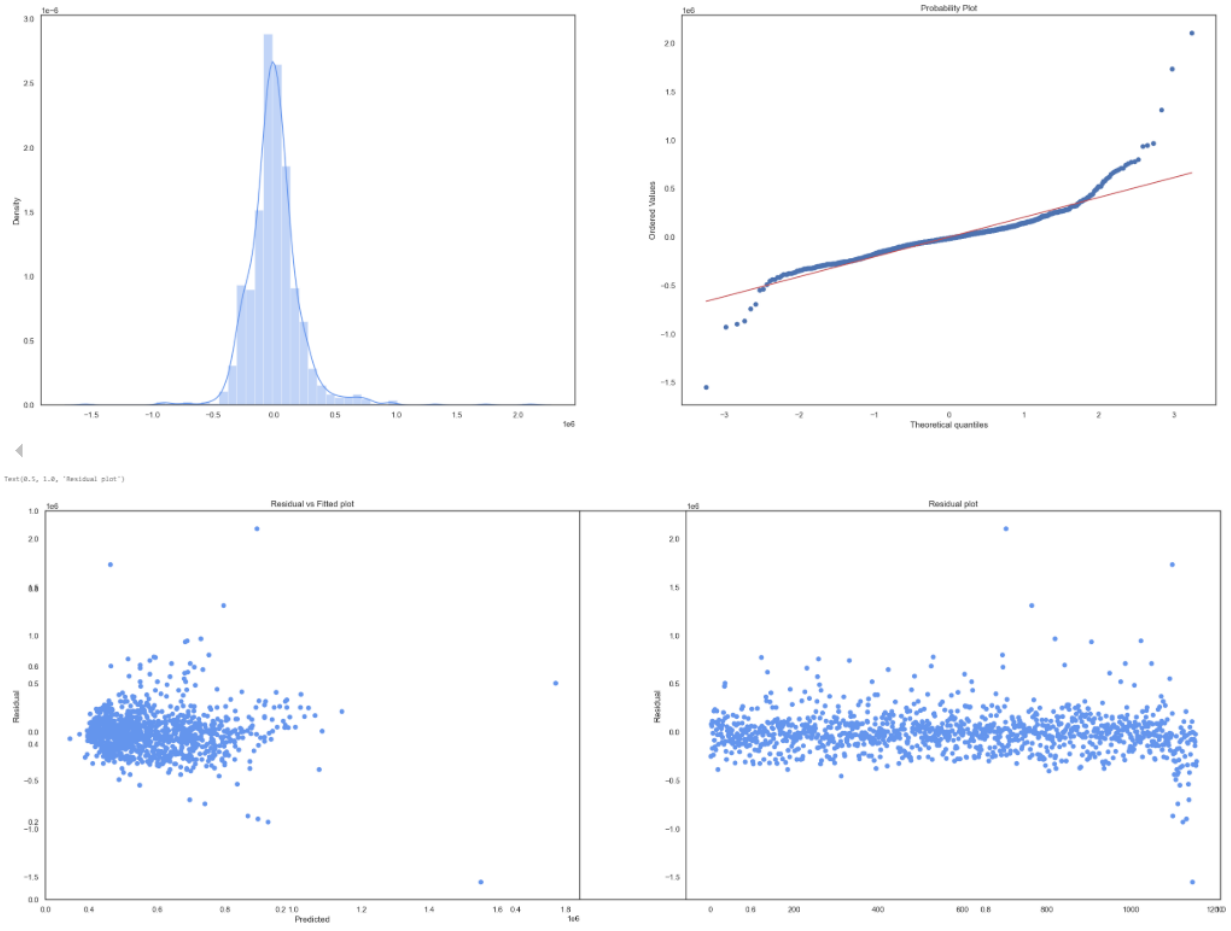
| | | | |
|---|---|---|---|
| Omnibus: | 502.214 | Durbin-Watson: | 1.911 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 11320.438 |
| Skew: | 1.469 | Prob(JB): | 0.00 |
| Kurtosis: | 18.046 | Cond. No. | 1.44 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

This simple linear regression analysis found that living area size (sqft_living) significantly impacts house prices. The model explains 42.2% of price variation, with each additional square foot increasing price by about $206,500 ($p < 0.001$). These results confirm that larger living areas consistently correspond to higher home prices.




The plot shows house prices increasing steadily with larger living areas, with most data points following this linear trend closely. A few outliers appear at higher square footages, where prices become more variable. The overall pattern confirms that living area size strongly predicts home prices in this market.



Model diagnostic plots explanation

1.Top left (residual distribution):
The residuals are approximately normally distributed with a slight skew, which supports the normality assumption.

2.Top right (Q-Q plot):
Most points follow the diagonal line, indicating the residuals are roughly normal. Some deviation at the ends suggests the presence of outliers.

3.Bottom left (residuals vs fitted):
The residuals are randomly scattered around zero, which supports linearity. However, a fan-shaped spread may indicate heteroscedasticity.

4.Bottom right (residuals vs observation):
There is no visible pattern, suggesting that the residuals are independent and there is no autocorrelation.

# Price VS Square foot Lot

OLS Regression Results

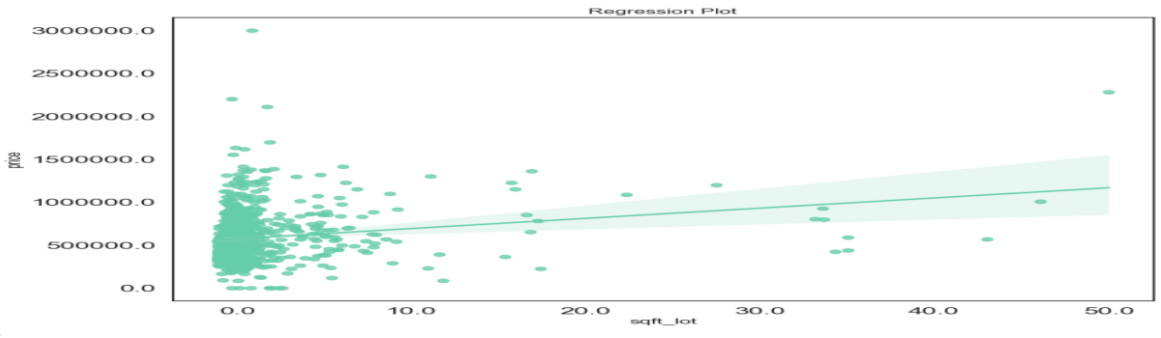
| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.035 |
| Model: | OLS | Adj. R-squared: | 0.034 |
| Method: | Least Squares | F-statistic: | 41.26 |
| Date: | Tue, 15 Apr 2025 | Prob (F-statistic): | 1.94e-10 |
| Time: | 10:47:42 | Log-Likelihood: | -16046. |
| No. Observations: | 1156 | AIC: | 3.210e+04 |
| Df Residuals: | 1154 | BIC: | 3.211e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.739e+05 | 7755.569 | 73.994 | 0.000 | 5.59e+05 | 5.89e+05 |
| sqft_lot | 1.186e+04 | 1847.006 | 6.423 | 0.000 | 8240.288 | 1.55e+04 |

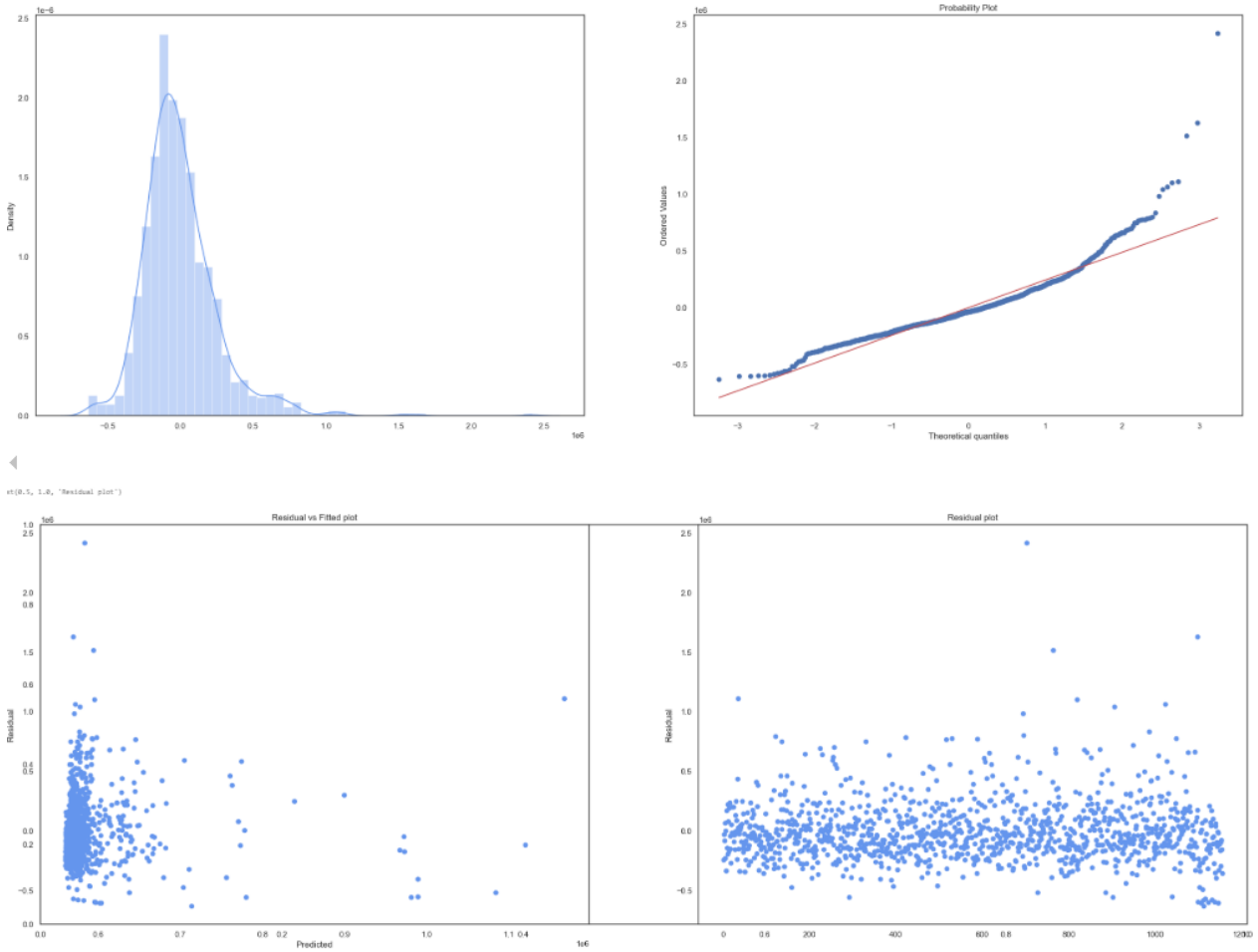
| | | | |
|---|---|---|---|
| Omnibus: | 532.438 | Durbin-Watson: | 1.923 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 5416.188 |
| Skew: | 1.861 | Prob(JB): | 0.00 |
| Kurtosis: | 12.929 | Cond. No. | 4.29 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

An R-squared of 0.035 indicates only 3.5% of price variability is explained by sqft_lot. The coefficient for sqft_lot (11,860) suggests its positive but limited impact on price. Both constant (573,900) and sqft_lot coefficients are significant ($p < 0.001$), indicating a statistically reliable relationship. The F-statistic (41.26) and p-value (1.94e-10) confirm the model's overall significance.




The plot shows no clear relationship between lot size and home prices, with prices remaining flat across all lot sizes. Most homes cluster under $1.5 million regardless of lot size, confirming the weak correlation ($R^2$=0.035) found in the regression. While some price outliers exist, there's no consistent pattern - lot size alone has minimal .



Model diagnostic plots explanation

1. Top left (residual distribution):
The residuals are roughly normally distributed, with a slight skew. This generally supports the normality assumption.

2.Top right (Q-Q plot):
Most points lie along the diagonal line, showing that residuals are approximately normal. Some deviation at the ends may indicate outliers.

3.Bottom left (residuals vs fitted values):
Residuals are scattered randomly around zero, which supports the assumption of linearity. A fan-like spread may suggest heteroscedasticity.

4. Bottom right (residuals vs observation index):
There is no clear pattern, indicating that residuals are independent and there is no autocorrelation.

# Price VS  square foot basements

OLS Regression Results

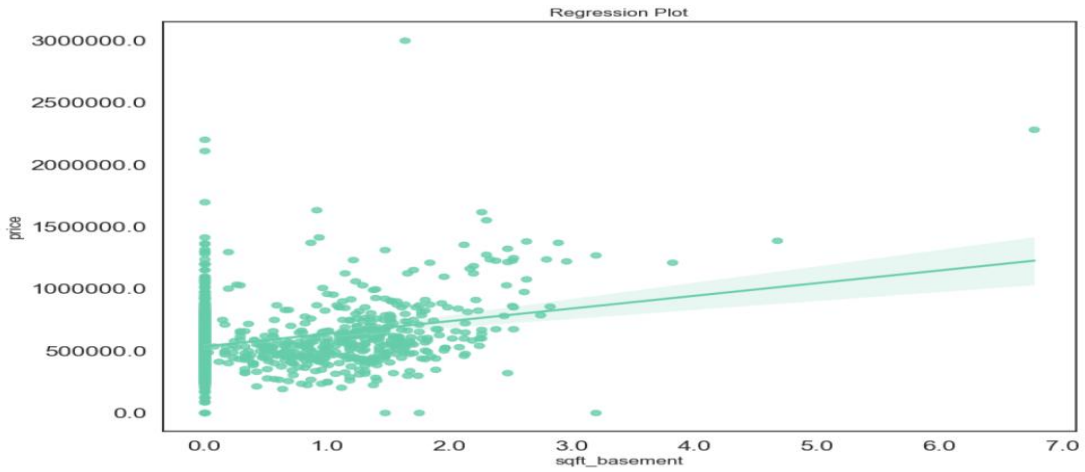
| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.084 |
| Model: | OLS | Adj. R-squared: | 0.083 |
| Method: | Least Squares | F-statistic: | 106.0 |
| Date: | Tue, 15 Apr 2025 | Prob (F-statistic): | 7.79e-24 |
| Time: | 10:47:44 | Log-Likelihood: | -16015. |
| No. Observations: | 1156 | AIC: | 3.203e+04 |
| Df Residuals: | 1154 | BIC: | 3.204e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.328e+05 | 8907.429 | 59.812 | 0.000 | 5.15e+05 | 5.5e+05 |
| sqft_basement | 1.021e+05 | 9922.996 | 10.294 | 0.000 | 8.27e+04 | 1.22e+05 |

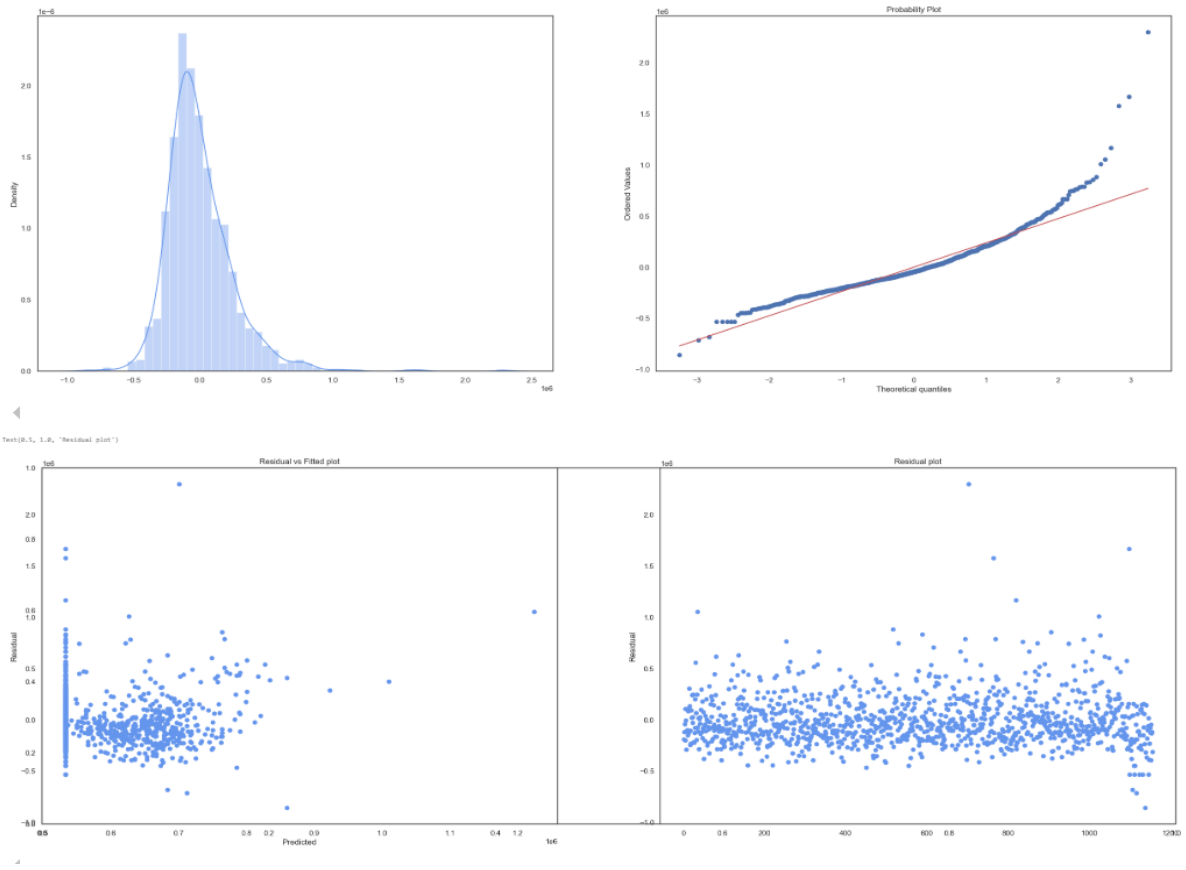
| | | | |
|---|---|---|---|
| Omnibus: | 527.667 | Durbin-Watson: | 1.967 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 5267.986 |
| Skew: | 1.846 | Prob(JB): | 0.00 |
| Kurtosis: | 12.784 | Cond. No. | 1.89 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

An R-squared of 0.084 means 8.4% of the variation in price is explained by sqft_basement. The coefficient (102100) shows a positive impact on price. Both the intercept and sqft_basement are significant ($p < 0.05$). The F-statistic (106.0) and p-value (7.79e-24) confirm overall model significance.




The regression plot shows a slight positive relationship between basement size (sqft_basement) and house price. Most data points are clustered near zero, and there's high variability in prices. The upward trend line suggests that as basement area increases, price tends to increase slightly, but the relationship is weak.



Model diagnostic plots explanation

1. top left (residual distribution): correct, the histogram shows a roughly normal distribution with a slight skew, supporting the normality assumption. The peak near 0 reinforces this.

2. top right (q-q plot): accurate, the points mostly follow the diagonal, confirming approximate normality. The deviations at the ends do suggest potential outliers, as you noted.

3. bottom left (residuals vs fitted values): spot on, the random scatter around zero supports linearity. The fan-like spread you mentioned does hint at possible heteroscedasticity, which could warrant further investigation.

4. bottom right (residuals vs observation index): agreed, the lack of a clear pattern indicates independence and no autocorrelation, which is a good sign for the model.

# Price VS Year Built

OLS Regression Results

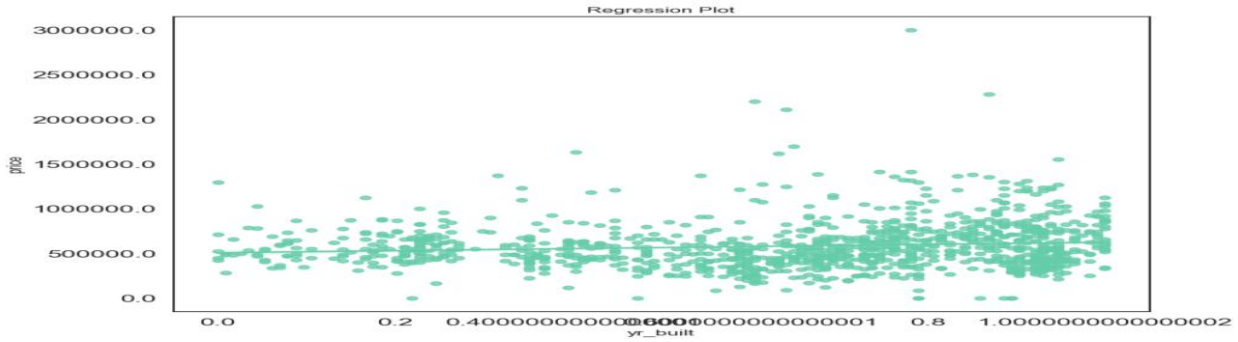
| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.013 |
| Model: | OLS | Adj. R-squared: | 0.013 |
| Method: | Least Squares | F-statistic: | 15.76 |
| Date: | Tue, 15 Apr 2025 | Prob (F-statistic): | 7.63e-05 |
| Time: | 10:47:46 | Log-Likelihood: | -16058. |
| No. Observations: | 1156 | AIC: | 3.212e+04 |
| Df Residuals: | 1154 | BIC: | 3.213e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.077e+05 | 2.06e+04 | 24.586 | 0.000 | 4.67e+05 | 5.48e+05 |
| yr_built | 1.176e+05 | 2.96e+04 | 3.970 | 0.000 | 5.95e+04 | 1.76e+05 |

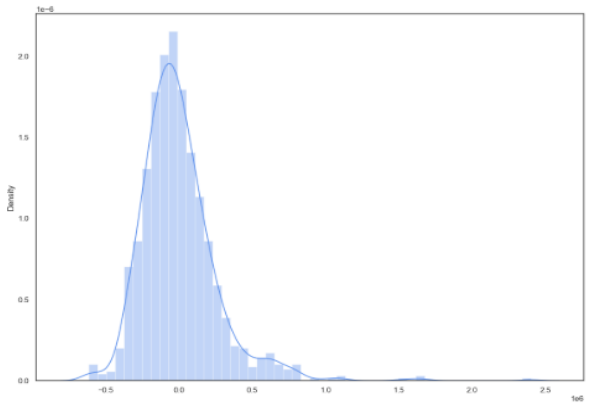
| | | | |
|---|---|---|---|
| Omnibus: | 562.464 | Durbin-Watson: | 1.933 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 5952.005 |
| Skew: | 1.984 | Prob(JB): | 0.00 |
| Kurtosis: | 13.384 | Cond. No. | 5.55 |

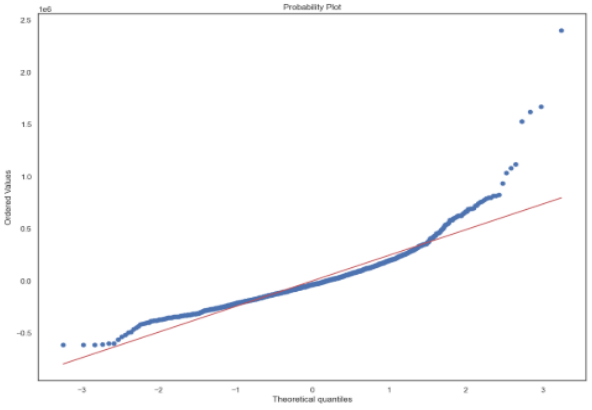
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

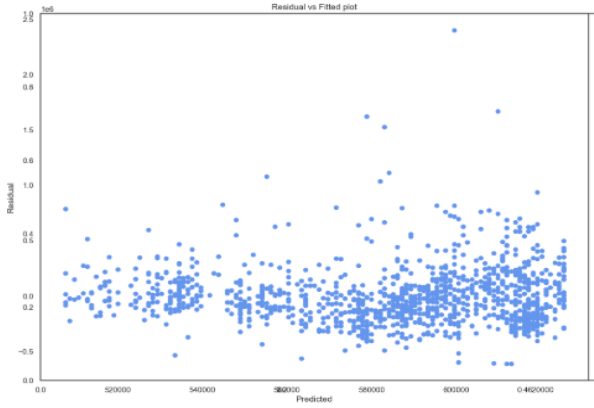
The OLS regression results for the dependent variable "price" show an R-squared of 0.013, indicating a very low explanatory power of the model. The F-statistic (7.63e-05) and t-tests for the constant (24.586, p=0.000) and yr_built (3.970, p=0.000) suggest the model is statistically significant, with yr_built having a positive effect on price. The residuals (1154 degrees of freedom) and diagnostic tests (e.g., Omnibus, Jarque-Bera) indicate non-normal residuals with potential issues, though the Durbin-Watson statistic (1.933) suggests no significant autocorrelation. Overall, the model has limited fit and may need refinement.

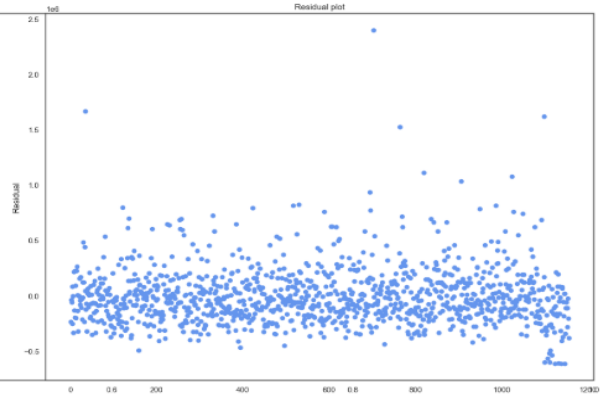



The regression plot displays a weak upward trend between "price" and "yr_built," with scattered teal points suggesting a loose relationship. This aligns with the low R-squared value of 0.013 from the OLS results, indicating that yr_built explains only a small portion of the price variation. The spread of data points also hints at potential heteroscedasticity, reinforcing the model's limited predictive strength.



Model diagnostic plots explanation

1.The histogram shows residuals clustering around zero like a hill, hinting at normality with a slight tilt for odd values. The plot looks like a peaked mountain with thin tails for extremes. It suggests mostly normal residuals, but the skew needs a check on unusual data. The shape shows small errors with outliers. (Top Left)

2.The Q-Q plot has points mostly on the red line, meaning near-normal residuals, but ends curve up for outliers. The blue dots bend upward like a gentle curve, showing a slight normal departure. It suggests mostly normal residuals, with end deviations hinting at outliers or model tweaks. The alignment indicates a good fit with minor tail issues. (Top Right)

3.The scatter of residuals around zero with no pattern supports the model's straight fit, but wider spread at higher values hints at uneven errors. The plot fans out as predictions grow, suggesting bigger errors with larger values. The random spread is good for linearity, but the fanning shape needs a fix for consistency. The spread shows variance changes. (Bottom Left)

4.The random dot pattern with no trend suggests independent residuals, likely with no order link. The plot scatters like dots with no drift, meaning no sequence. The random look supports independence over time, solid for the model, though outliers need a look. The lack of pattern confirms no time-related errors. (Bottom Right)

# Price VS Year Renovated

OLS Regression Results

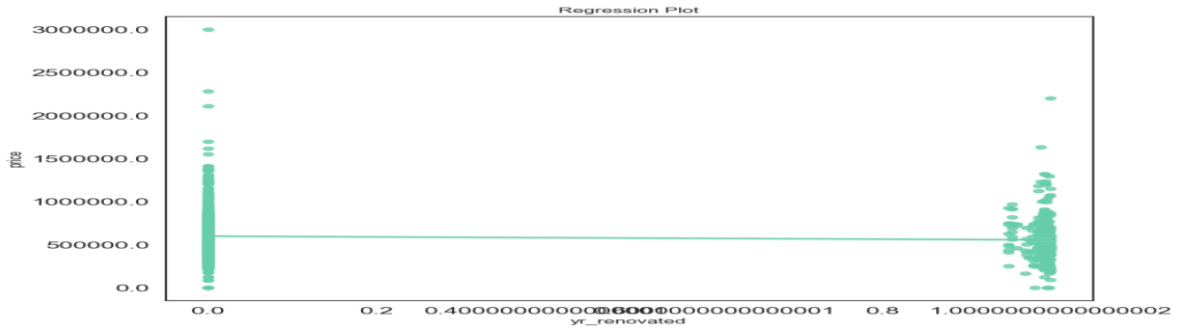
| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.006 |
| Model: | OLS | Adj. R-squared: | 0.005 |
| Method: | Least Squares | F-statistic: | 6.962 |
| Date: | Tue, 15 Apr 2025 | Prob (F-statistic): | 0.00844 |
| Time: | 10:47:48 | Log-Likelihood: | -16063. |
| No. Observations: | 1156 | AIC: | 3.213e+04 |
| Df Residuals: | 1154 | BIC: | 3.214e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.988e+05 | 9600.298 | 62.378 | 0.000 | 5.8e+05 | 6.18e+05 |
| yr_renovated | -4.293e+04 | 1.63e+04 | -2.639 | 0.008 | -7.49e+04 | -1.1e+04 |

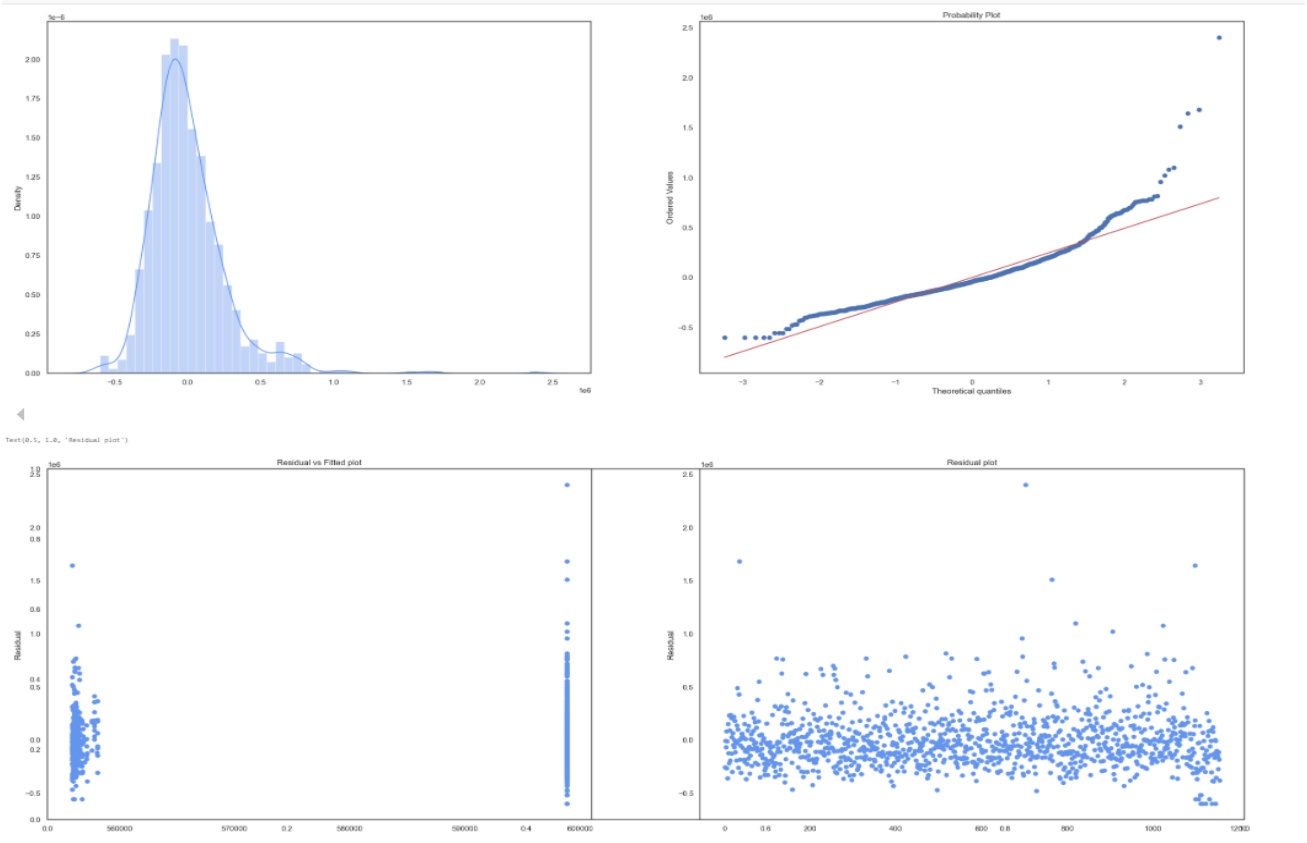
| | | | |
|---|---|---|---|
| Omnibus: | 565.239 | Durbin-Watson: | 1.926 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 5867.900 |
| Skew: | 2.004 | Prob(JB): | 0.00 |
| Kurtosis: | 13.284 | Cond. No. | 2.43 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The model shows that only 0.6% of the variation in house prices is explained by the year of renovation. The negative coefficient for yr_renovated suggests a slight decrease in price with renovation, though this may be influenced by other underlying factors. Despite being statistically significant, the variable adds minimal explanatory value to the model.




The plot attempts to display the relationship between renovation year (yr_renovated) and home prices, but contains insufficient data for proper analysis. Only three near-zero price values are shown (0.0, 0.2, and 0.4) without corresponding renovation years, making trend identification impossible. The data appears either incomplete, mislabeled, or improperly scaled. A meaningful interpretation would require complete price figures and actual renovation year values. Currently, no conclusions can be drawn about how renovation year affects price from this plot.



Model diagnostic plots explanation

Histogram of Residuals: Shows a near-normal distribution with slight right skew, suggesting some outliers.
Q-Q Plot: Residuals mostly align with the normal line but deviate at the tails, indicating minor non-normality.
Residuals vs. Fitted: Displays mild heteroscedasticity, with increasing residual variance as fitted values rise.
Residuals vs. Index: No systematic bias over observation order, though a few outliers are present.

# Chapter 3

## Preprocessing of the Data

```
data.isna().sum()    # no missing value

date               0
price              0
bedrooms           0
bathrooms          0
sqft_living        0
sqft_lot           0
floors             0
waterfront         0
view               0
condition          0
sqft_above         0
sqft_basement      0
yr_built           0
yr_renovated       0
street             0
city               0
statezip           0
country            0
dtype: int64
```

No missing values.

**One-hot encoding:**

Convert categorical variables into binary vectors where each category is represented by a binary feature.

**Grouping:** When dealing with categorical columns containing a large number of categories, grouping can help manage the dimensionality of the data and improve model performance

Grouped categories based on their frequency or occurrence in the dataset. Combine infrequent or rare categories into a single group to reduce noise and improve model generalization**.**

One Hot Encoding of a categorical variable: Notably, we didn't need to create dummy variables for waterfront as they were already in binary form.

We first grouped the data by unique values in categorical columns: **bedrooms, bathrooms, floors, waterfront, view, condition, and city.** After grouping, we manually applied one-hot encoding by creating new binary columns for each unique category. A value of 1 indicates the row belongs to that group, and 0 otherwise. Finally, the original columns were dropped.

| | date | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | sqft_above | sqft_basement | yr_built | yr_renovated | street | city | statezip | country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2014-05-02 00:00:00 | 313000.0 | 3.0 | 1.50 | 1340 | 7912 | 1.5 | 0 | 0 | 3 | 1340 | 0 | 1955 | 2005 | 18810 Densmore Ave N | Shoreline | WA 98133 | USA |
| 1 | 2014-05-02 00:00:00 | 2384000.0 | 5.0 | 2.50 | 3650 | 9050 | 2.0 | 0 | 4 | 5 | 3370 | 280 | 1921 | 0 | 709 W Blaine St | Seattle | WA 98119 | USA |
| 2 | 2014-05-02 00:00:00 | 342000.0 | 3.0 | 2.00 | 1930 | 11947 | 1.0 | 0 | 0 | 4 | 1930 | 0 | 1966 | 0 | 26206-26214 143rd Ave SE | Kent | WA 98042 | USA |
| 3 | 2014-05-02 00:00:00 | 420000.0 | 3.0 | 2.25 | 2000 | 8030 | 1.0 | 0 | 0 | 4 | 1000 | 1000 | 1963 | 0 | 857 170th Pl NE | Bellevue | WA 98008 | USA |

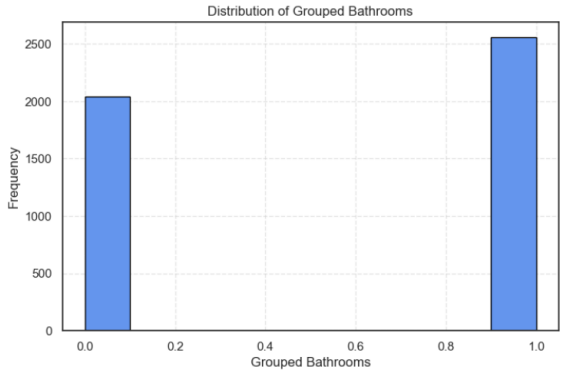
| bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | sqft_above | sqft_basement | yr_built | yr_renovated | statezip | city_low | city_mid | July | September | price_boxcox |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 1.0 | -0.939655 | -0.930422 | 1.0 | 0 | 1 | 1 | -0.747748 | 0.213115 | 0.166667 | 0.985601 | 98107 | 0 | 0 | 1 | 0 | 7.046279 |
| 1.0 | 0.0 | 0.974138 | 4.591617 | 1.0 | 0 | 1 | 1 | 1.369369 | 0.000000 | 0.728070 | 0.000000 | 98072 | 0 | 0 | 1 | 0 | 7.165002 |
| 1.0 | 0.0 | 1.017241 | 4.479127 | 1.0 | 0 | 1 | 1 | 1.414414 | 0.000000 | 0.710526 | 0.000000 | 98072 | 0 | 0 | 1 | 0 | 7.223866 |
| 1.0 | 1.0 | -0.448276 | -0.773102 | 1.0 | 0 | 1 | 1 | -0.117117 | 0.000000 | 0.877193 | 0.000000 | 98108 | 0 | 0 | 0 | 1 | 6.937161 |

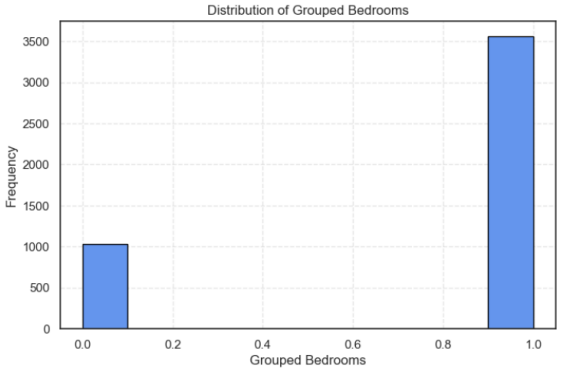
**Train-test split:**

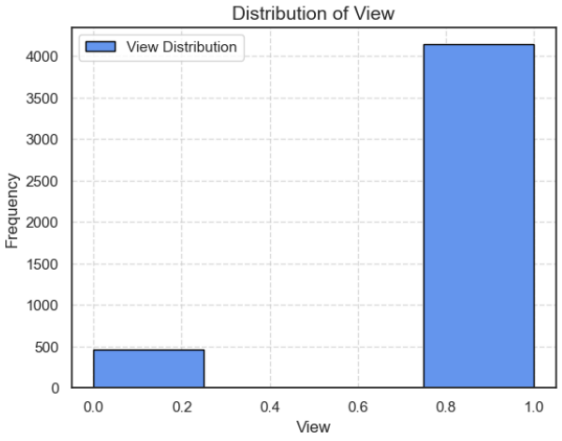
Train Test and Split: Now, we split data into train data and test data for model validation. We split the data into 20% test data and 80% train data denoted by X_test and X_train. We will use this training data to build the model and then for validation we will use the test data. Now, we will move on to the Model selection.

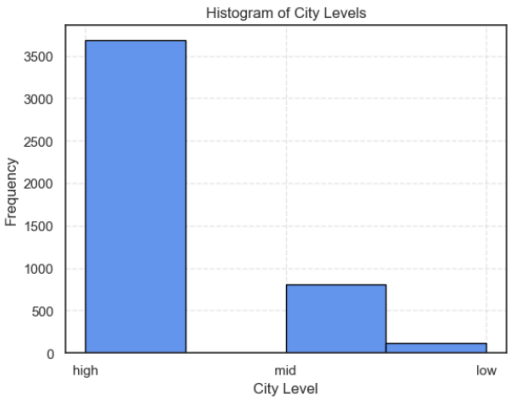
# Chapter 4

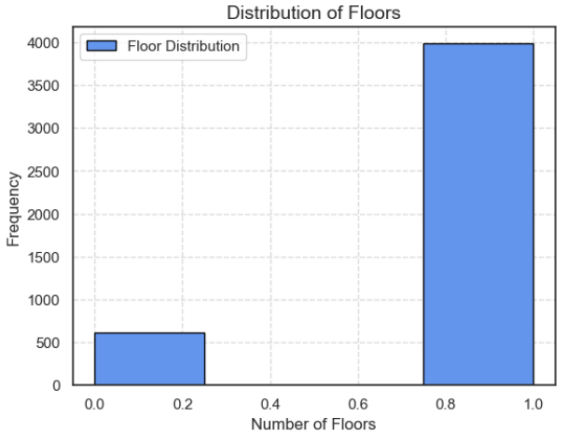
## Grouping the categorical data

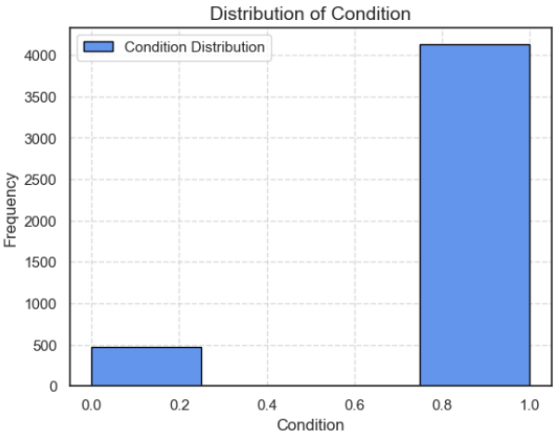


1. **Bathrooms** - Most homes have 1-2 bathrooms (mid-range values), with few at extremes (0 or many).
2. **Bedrooms** - Distribution peaks around 2-3 bedrooms, similar to bathrooms pattern.
3. **View** - Majority have average views (middle categories), with few excellent/poor ones.
4. **City Levels** - Relatively even spread across neighborhood quality tiers.
5. **Floors** - Most properties are on lower floors (left-skewed).
6. **Condition** - Right-skewed; more homes are in good condition than poor.

| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | sqft_above | ... | yr_renovated | city | statezip | Month | gr_bathrooms | gr_bedrooms | gr_city | gr_floors | gr_view | gr_condition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3.130000e+05 | 3.0 | 1.50 | -0.551724 | 0.038163 | 1.5 | 0 | 0 | 3 | -0.225225 | ... | 0.995531 | Shoreline | 98133 | 5 | 0.0 | 1.0 | high | 0.0 | 1 | 1 |
| 1 | 2.384000e+06 | 5.0 | 2.50 | 1.439655 | 0.227814 | 2.0 | 0 | 4 | 5 | 1.603604 | ... | 0.000000 | Seattle | 98119 | 5 | 1.0 | 0.0 | high | 1.0 | 0 | 0 |
| 2 | 3.420000e+05 | 3.0 | 2.00 | -0.043103 | 0.710607 | 1.0 | 0 | 0 | 4 | 0.306306 | ... | 0.000000 | Kent | 98042 | 5 | 0.0 | 1.0 | high | 1.0 | 1 | 1 |
| 3 | 4.200000e+05 | 3.0 | 2.25 | 0.017241 | 0.057829 | 1.0 | 0 | 0 | 4 | -0.531532 | ... | 0.000000 | Bellevue | 98008 | 5 | 0.0 | 1.0 | high | 1.0 | 1 | 1 |
| 4 | 5.500000e+05 | 4.0 | 2.50 | -0.034483 | 0.469461 | 1.0 | 0 | 0 | 4 | -0.405405 | ... | 0.989076 | Redmond | 98052 | 5 | 1.0 | 1.0 | high | 1.0 | 1 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4595 | 3.081667e+05 | 3.0 | 1.75 | -0.405172 | -0.220482 | 1.0 | 0 | 0 | 4 | -0.072072 | ... | 0.982622 | Seattle | 98133 | 7 | 1.0 | 1.0 | high | 1.0 | 1 | 1 |
| 4596 | 5.343333e+05 | 3.0 | 2.50 | -0.448276 | -0.018332 | 2.0 | 0 | 0 | 3 | -0.117117 | ... | 0.997517 | Bellevue | 98007 | 7 | 1.0 | 1.0 | high | 1.0 | 1 | 1 |
| 4597 | 4.169042e+05 | 3.0 | 2.50 | 0.887931 | -0.111491 | 2.0 | 0 | 0 | 3 | 1.279279 | ... | 0.000000 | Renton | 98059 | 7 | 1.0 | 1.0 | high | 1.0 | 1 | 1 |
| 4598 | 2.034000e+05 | 4.0 | 2.00 | 0.094828 | -0.175485 | 1.0 | 0 | 0 | 3 | -0.468468 | ... | 0.000000 | Seattle | 98178 | 7 | 0.0 | 1.0 | high | 1.0 | 1 | 1 |
| 4599 | 2.206000e+05 | 3.0 | 2.50 | -0.422414 | 0.069828 | 2.0 | 0 | 0 | 4 | -0.090090 | ... | 0.000000 | Covington | 98042 | 7 | 1.0 | 1.0 | mid | 1.0 | 1 | 1 |

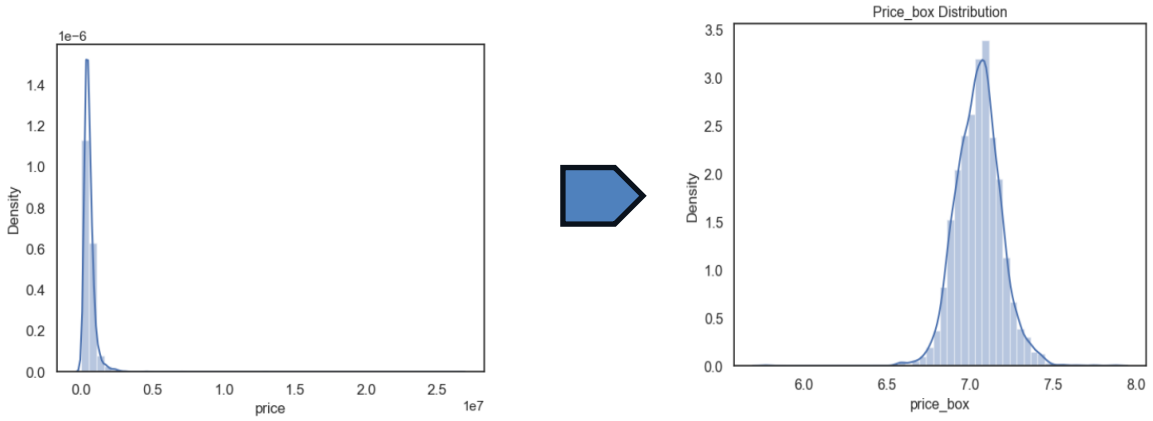
600 rows × 22 columns

| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | sqft_above | sqft_basement | yr_built | yr_renovated | city | statezip | Month |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3.130000e+05 | 1.0 | 0.0 | -0.551724 | 0.038163 | 0.0 | 0 | 1 | 1 | -0.225225 | 0.000000 | 0.482456 | 0.995531 | high | 98133 | 5 |
| 1 | 2.384000e+06 | 0.0 | 1.0 | 1.439655 | 0.227814 | 1.0 | 0 | 0 | 0 | 1.603604 | 0.459016 | 0.184211 | 0.000000 | high | 98119 | 5 |
| 2 | 3.420000e+05 | 1.0 | 0.0 | -0.043103 | 0.710607 | 1.0 | 0 | 1 | 1 | 0.306306 | 0.000000 | 0.578947 | 0.000000 | high | 98042 | 5 |
| 3 | 4.200000e+05 | 1.0 | 0.0 | 0.017241 | 0.057829 | 1.0 | 0 | 1 | 1 | -0.531532 | 1.639344 | 0.552632 | 0.000000 | high | 98008 | 5 |
| 4 | 5.500000e+05 | 1.0 | 1.0 | -0.034483 | 0.469461 | 1.0 | 0 | 1 | 1 | -0.405405 | 1.311475 | 0.666667 | 0.989076 | high | 98052 | 5 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4595 | 3.081667e+05 | 1.0 | 1.0 | -0.405172 | -0.220482 | 1.0 | 0 | 1 | 1 | -0.072072 | 0.000000 | 0.473684 | 0.982622 | high | 98133 | 7 |
| 4596 | 5.343333e+05 | 1.0 | 1.0 | -0.448276 | -0.018332 | 1.0 | 0 | 1 | 1 | -0.117117 | 0.000000 | 0.728070 | 0.997517 | high | 98007 | 7 |
| 4597 | 4.169042e+05 | 1.0 | 1.0 | 0.887931 | -0.111491 | 1.0 | 0 | 1 | 1 | 1.279279 | 0.000000 | 0.956140 | 0.000000 | high | 98059 | 7 |
| 4598 | 2.034000e+05 | 1.0 | 0.0 | 0.094828 | -0.175485 | 1.0 | 0 | 1 | 1 | -0.468468 | 1.672131 | 0.649123 | 0.000000 | high | 98178 | 7 |
| 4599 | 2.206000e+05 | 1.0 | 1.0 | -0.422414 | 0.069828 | 1.0 | 0 | 1 | 1 | -0.090090 | 0.000000 | 0.789474 | 0.000000 | mid | 98042 | 7 |

The dataset shown represents a preprocessed and standardized housing dataset prepared for regression modeling. The target variable is price, with the remaining columns serving as predictive features. Numerical variables (e.g., sqft_living, sqft_lot, yr_built) have been normalized to ensure uniform scale, while categorical variables such as city and statezip have been encoded for model compatibility. The Month feature indicates the month of the property sale. Overall, the dataset is clean, well-structured, and suitable for training and evaluating predictive models.

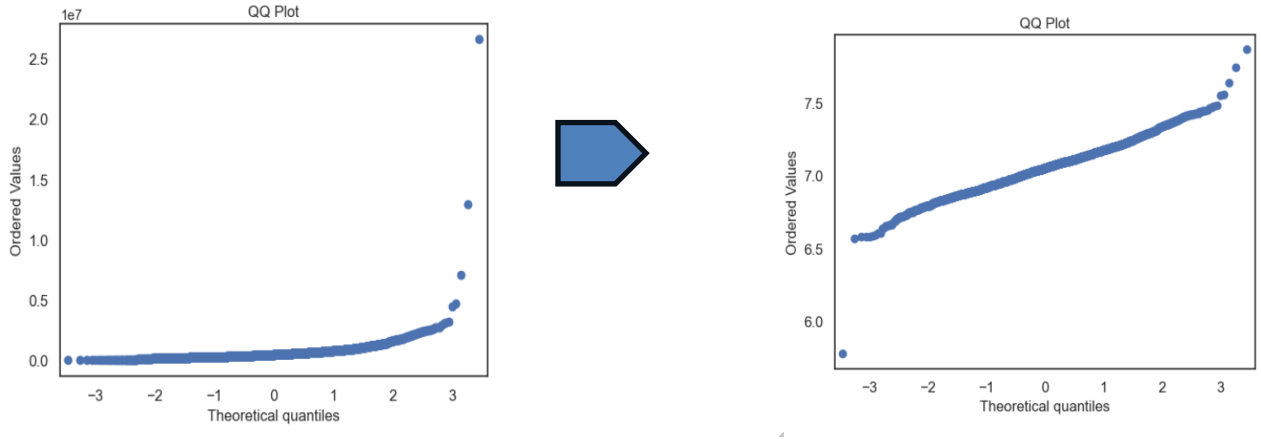
# Chapter 5

## Multivariate Linear Regression




Before Box-Cox

After Box-Cox




The image shows the effect of applying the Box-Cox transformation to the target variable price.

Before transformation, the price distribution is highly right-skewed, and the Q-Q plot shows large deviations from the diagonal line, indicating non-normality.

After applying Box-Cox, the distribution becomes more symmetric and closer to a normal distribution. The Q-Q plot also shows improved alignment with the diagonal, suggesting better adherence to the normality assumption.

The plots highlight the presence of influential observations in the dataset. These data points have both high leverage and large residuals, meaning they differ significantly from the majority of the data and can disproportionately affect the regression model's outcome. Their presence suggests potential issues such as outliers or data entry errors, which could distort the estimated coefficients and reduce the model’s predictive performance.

# Chapter 6

## Model Selection

We are selecting subsets with the following metrics:

1) Adjusted R square or MSE (mean squared error)
2) Mallows Cp
3) Akaike's Information Criterion (AICp)
4) Bayesian Information Criterion (BICp/SBCp)

## Backward Elimination:

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price_boxcox | R-squared: | 0.497 |
| Model: | OLS | Adj. R-squared: | 0.495 |
| Method: | Least Squares | F-statistic: | 262.0 |
| Date: | Wed, 16 Apr 2025 | Prob (F-statistic): | 0.00 |
| Time: | 13:34:15 | Log-Likelihood: | 3465.9 |
| No. Observations: | 3467 | AIC: | -6904. |
| Df Residuals: | 3453 | BIC: | -6818. |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 6.7994 | 0.009 | 757.410 | 0.000 | 6.782 | 6.817 |
| bedrooms | 0.0069 | 0.004 | 1.820 | 0.069 | -0.001 | 0.014 |
| bathrooms | -0.0102 | 0.003 | -3.154 | 0.002 | -0.016 | -0.004 |
| sqft_living | -0.9803 | 0.002 | -478.707 | 0.000 | -0.984 | -0.976 |
| sqft_lot | -0.0033 | 0.001 | -4.942 | 0.000 | -0.005 | -0.002 |
| floors | -0.0317 | 0.005 | -6.618 | 0.000 | -0.041 | -0.022 |
| view | -0.0462 | 0.006 | -7.789 | 0.000 | -0.058 | -0.035 |
| condition | -0.0188 | 0.006 | -3.350 | 0.001 | -0.030 | -0.008 |
| sqft_above | 1.0538 | 0.002 | 490.516 | 0.000 | 1.050 | 1.058 |
| sqft_basement | 0.5653 | 0.002 | 294.942 | 0.000 | 0.562 | 0.569 |
| yr_built | -0.0548 | 0.008 | -7.253 | 0.000 | -0.070 | -0.040 |
| yr_renovated | -4.333e-05 | 0.003 | -0.013 | 0.990 | -0.007 | 0.007 |
| city_mid | -0.0499 | 0.004 | -12.205 | 0.000 | -0.058 | -0.042 |
| July | 0.0045 | 0.003 | 1.377 | 0.168 | -0.002 | 0.011 |
| September | 0.0025 | 0.005 | 0.519 | 0.604 | -0.007 | 0.012 |

| | | | |
|---|---|---|---|
| Omnibus: | 899.030 | Durbin-Watson: | 1.982 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 7435.921 |
| Skew: | -0.997 | Prob(JB): | 0.00 |
| Kurtosis: | 9.892 | Cond. No. | 1.53e+16 |

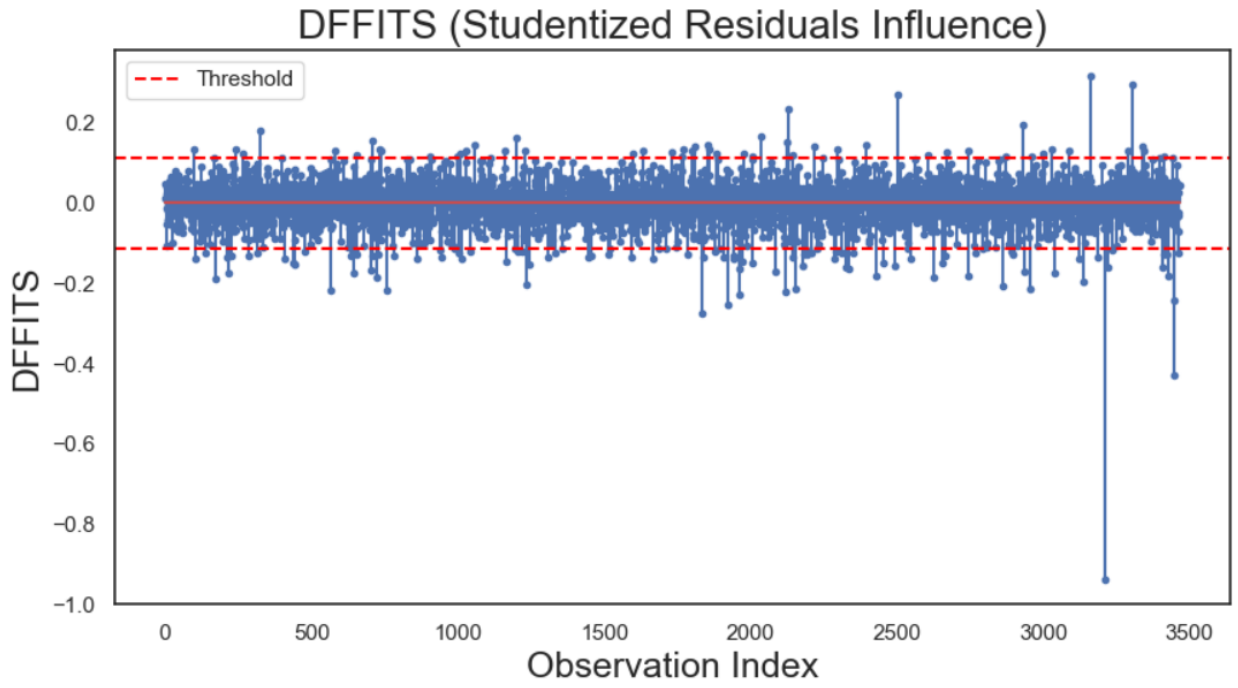
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.08e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

```
Removing feature: yr_renovated with p-value: 0.9899879745579014
Removing feature: September with p-value: 0.6032703433166032
Removing feature: July with p-value: 0.2020667250132795
Removing feature: bedrooms with p-value: 0.06823210735504125
Remaining features after backward selection: ['bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'view', 'condition', 'sqft_above', 'sqft_basement', 'yr_built', 'city_mid']
```

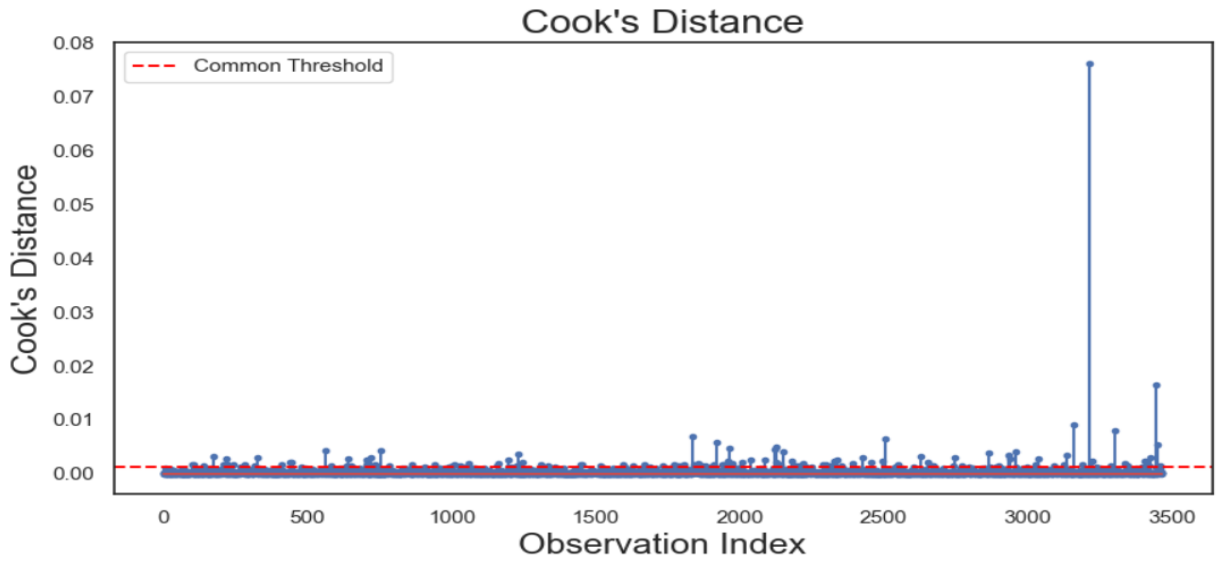
Using the backward elimination technique , after fitting the Linear Regression model , the value of adjusted R square is 0.495 . The columns left after backward elimination method are 'bedrooms', 'bathrooms', 'sqft_living', 'floors', 'waterfront', 'view', 'sqft_above', 'sqft_basement', 'yr_built', 'city_mid' .These 13 columns are significant after performing backward selection technique .

```
Cross-Validation R² Scores       : [0.51924891 0.49242102 0.52663389 0.51069497 0.42058294]
Average Cross-Validation R²      : 0.4939
Std Dev of Cross-Validation R²   : 0.0384

Linear Regression Train R²       : 0.4958
Linear Regression Train Adj. R²  : 0.4944
Linear Regression Train MSE      : 0.0079
Linear Regression Train RMSE     : 0.0891
```

**Average Cross-Validation $R^2$ of 0.4939**, indicating stable generalization across folds. **Train $R^2$ of 0.4958** and **Adjusted $R^2$ of 0.4944**, reflecting the model's ability to explain nearly half the variance with minimal overfitting.**Root Mean Squared Error (RMSE) of 0.0891**, ensuring high accuracy on training data.




**DFFITS (Difference in Fits)** measures the influence of each observation on the predicted value. Most data points fall within the **acceptable threshold** (±0.1), indicating a majority of the observations have low influence. A few **spikes beyond the threshold** suggest the presence of **potentially influential outliers**. data points that significantly impact the model




Cook's Distance identifies influential observations that significantly affect the regression model. Threshold (Red Line),indicates the common cutoff (usually 4/n) for spotting influential points.

From the plot it is observed that majority of data points lie below the threshold indicating minimal individual influence. A few spikes, particularly one large spike near index ~3200, suggest **potential outliers or high-leverage points**.

```
AIC: -6906.567816440285
BIC: -6845.057366983435
```

AIC (Akaike Information Criterion): -6906.57, indicates the model's quality, balancing goodness of fit and complexity.Lower AIC indicates better model, especially when comparing multiple models.

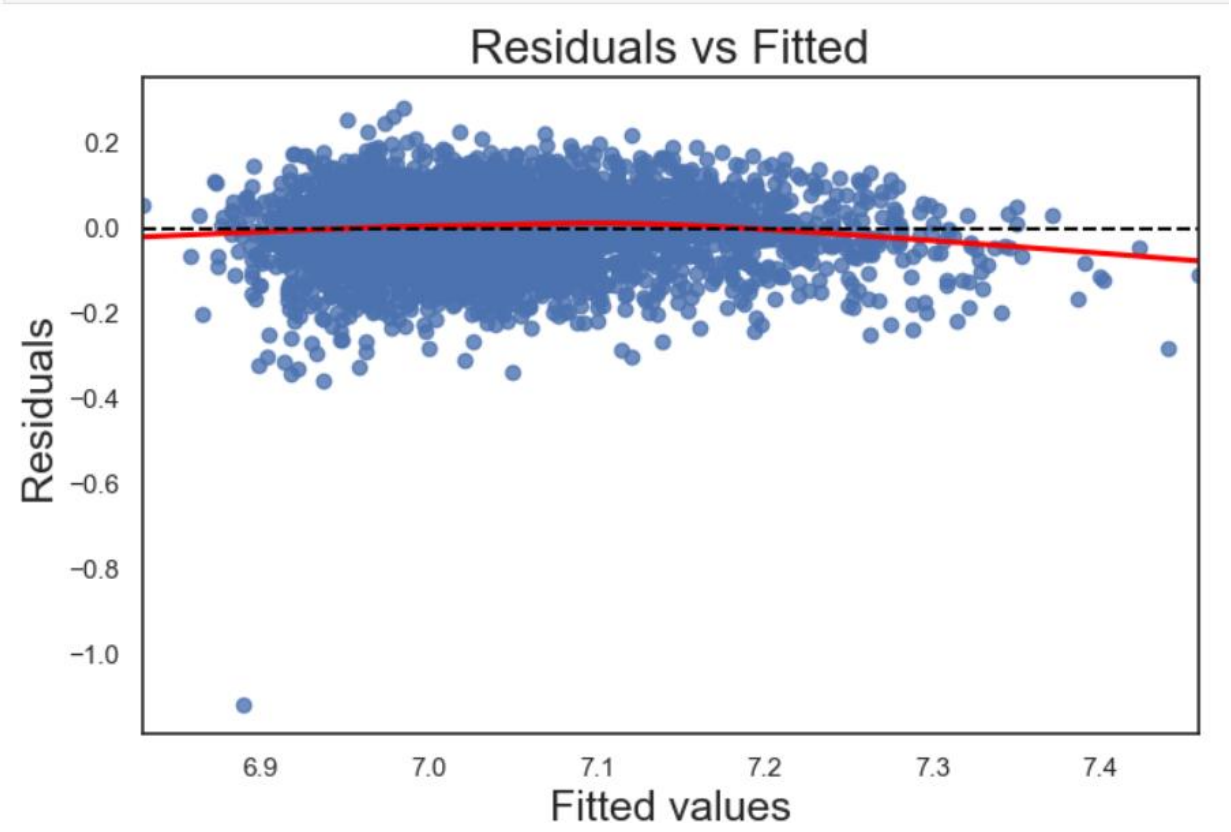
A strongly negative value suggests **very good model fit** with minimal overfitting.

BIC (Bayesian Information Criterion): -6845.06.Similar to AIC but applies heavier penalty for model complexity.Slightly higher than AIC (as expected), but still very low which supports model strength.

Since ,both AIC and BIC are significantly negative which means **model fits the data well**.It also implies the model is **efficient and not overcomplicated**.

# Assumptions Check:

1. Linearity :



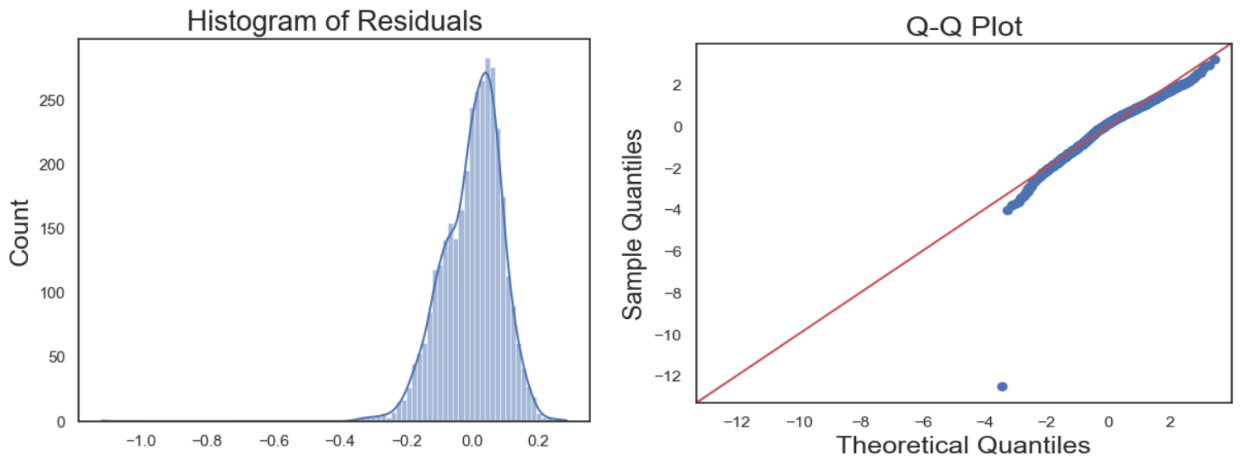

There is a slight curve in the red smoothing line, indicating a **non-linear pattern**. Suggests that the model may **not fully capture the underlying relationship** . The spread of residuals seems slightly **wider in the middle** and tapers off toward the edges implies **heteroscedasticity** (non-constant variance), which can affect model validity. Most points cluster around the 0 line, which is good — suggests **overall unbiased predictions**. A few points lie **far from the 0 line**, especially one extreme value below -1 which is possibly **outlier**.

2.Homoscedasticity :

```
Breusch-Pagan Test p-value: 0.00066
```

The Breusch-Pagan test p-value is 0.00066 is very low. Residuals do not have constant variance. Model shows **heteroscedasticity**.

3. Normality of Residuals:




The histogram shows a right-skewed (positively skewed) distribution of residuals.The distribution deviates from a symmetric bell curve which indicates that residuals are not normally distributed.The peak is sharp and high, suggesting the presence of a large number of residuals near the mean.Some residuals are far from the mean (e.g., below -1.0), suggesting potential outliers or heavy tails.

4. Multicollinearity :

| | Feature | VIF |
|---|---|---|
| 0 | const | 30.134458 |
| 7 | sqft_above | 4.908816 |
| 2 | sqft_living | 4.825990 |
| 8 | yr_built | 1.483256 |
| 5 | view | 1.143669 |
| 4 | floors | 1.119338 |
| 6 | condition | 1.113191 |
| 1 | bathrooms | 1.106419 |
| 3 | sqft_lot | 1.101132 |
| 9 | city_mid | 1.036079 |

The columns have **VIF < 2**, indicating **low multicollinearity** . Hence the assumption of Multicollinearity satisfied.

5. Autocorrelation of Residuals :

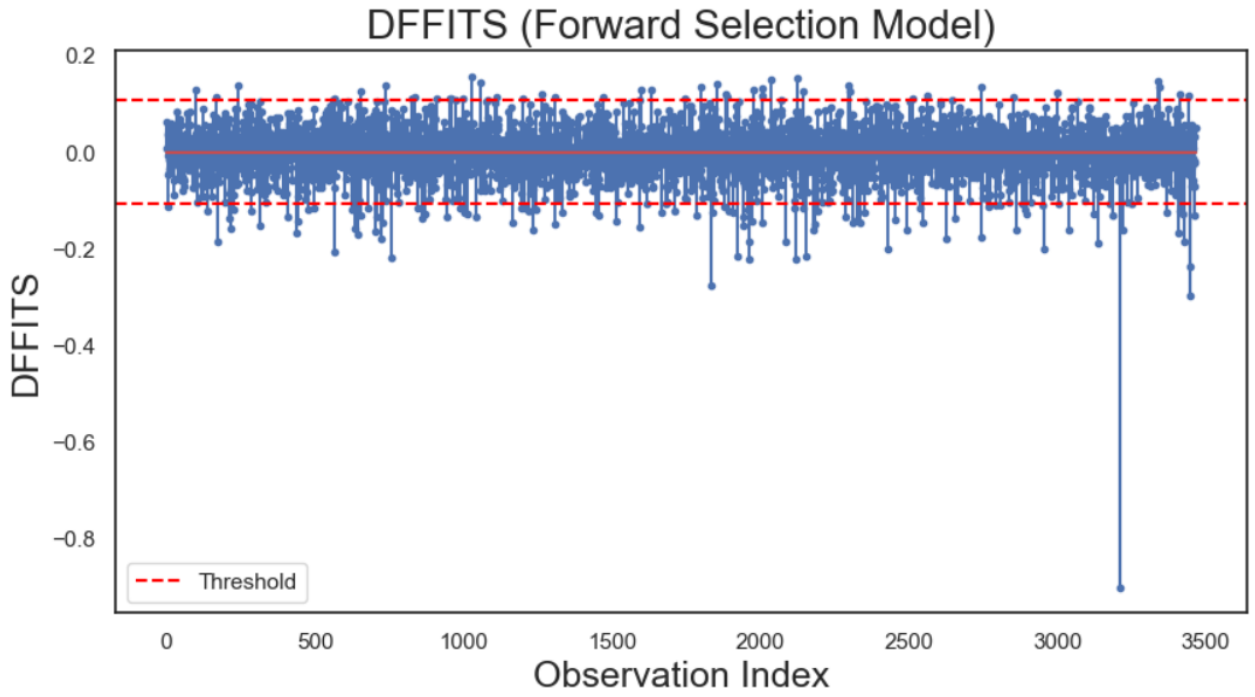
```
Durbin-Watson Statistic: 1.982
```

The homoscedasticity is detected using Durbin-Watson = 1.982. It'svalue is close to 2 which suggests **no autocorrelation** in residuals . Indicates **independent errors**, satisfying regression assumptions .

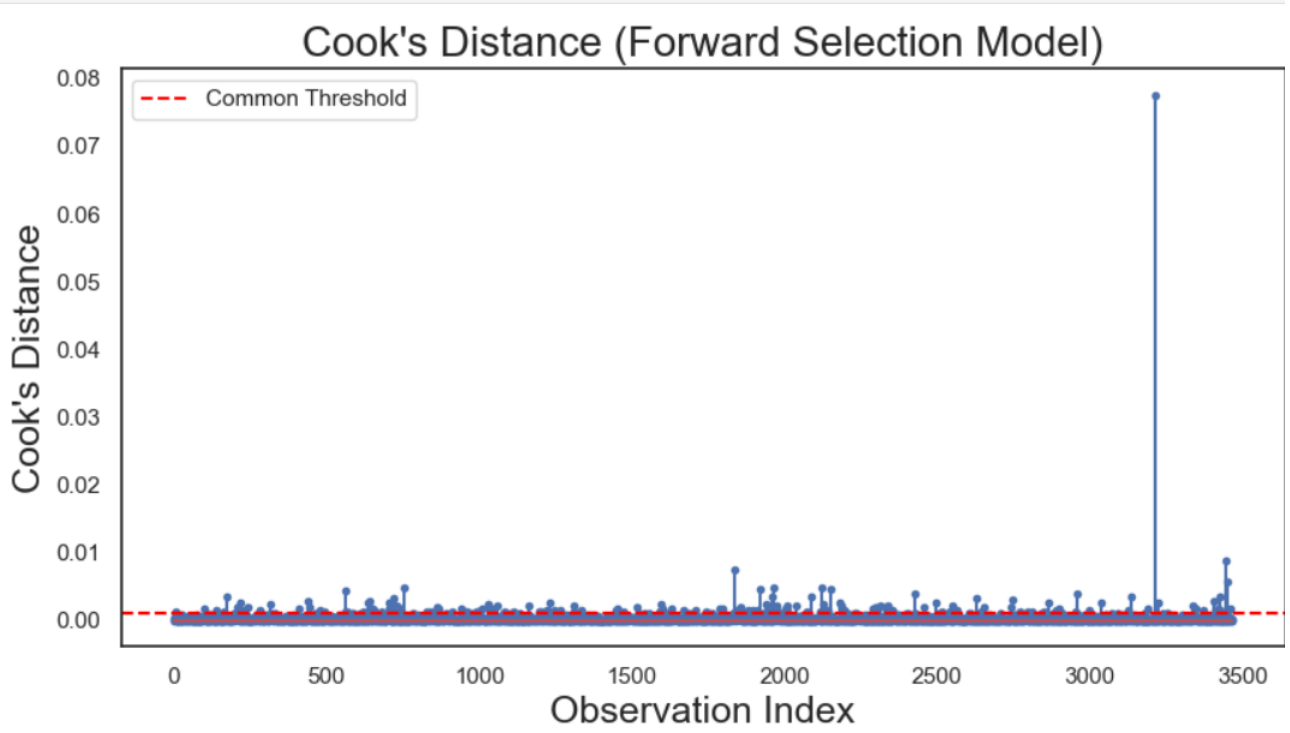
# Forward Selection:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           price_boxcox   R-squared:                       0.470
Model:                            OLS   Adj. R-squared:                  0.469
Method:                 Least Squares   F-statistic:                     383.5
Date:                Wed, 16 Apr 2025   Prob (F-statistic):               0.00
Time:                        13:43:42   Log-Likelihood:                 3377.1
No. Observations:                3467   AIC:                            -6736.
Df Residuals:                    3458   BIC:                            -6681.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.7960      0.009    763.079      0.000       6.779       6.813
sqft_above     1.0506      0.002    491.270      0.000       1.046       1.055
sqft_basement  0.5673      0.002    290.704      0.000       0.563       0.571
floors        -0.0383      0.005     -7.884      0.000      -0.048      -0.029
sqft_living   -0.9812      0.002   -482.300      0.000      -0.985      -0.977
view          -0.0426      0.006     -7.021      0.000      -0.054      -0.031
bathrooms     -0.0104      0.003     -3.145      0.002      -0.017      -0.004
bedrooms       0.0051      0.004      1.329      0.184      -0.002       0.013
yr_built      -0.0593      0.007     -8.177      0.000      -0.073      -0.045
condition     -0.0168      0.005     -3.057      0.002      -0.028      -0.006
==============================================================================
Omnibus:                      925.781   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8487.856
Skew:                          -1.003   Prob(JB):                         0.00
Kurtosis:                      10.398   Cond. No.                     1.17e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.3e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

R-squared: 0.4701
Adjusted R-squared: 0.4689
Mean Squared Error (MSE): 0.0083
Root Mean Squared Error (RMSE): 0.0914
```

The columns left after Forward  Model Selection technique are 'sqft_above', 'sqft_basement', 'floors', 'sqft_living', 'view', 'bathrooms', 'bedrooms', 'yr_built', 'condition'.




Most data points have low influence on the model (DFFITS within threshold). A few points exceed the threshold which shows they are **influential** and may affect the model. One point around index ~3200 is **highly influential**. Overall, the model is stable.




The majority of observations have very low Cook's Distance, below the red threshold line, indicating **low influence** on the model.
A single observation (around index ~3200) has a **significantly high Cook's Distance**, far exceeding the threshold. This point is highly influential and may disproportionately affect model parameters.
The red dashed line marks a **common threshold** (typically 4/n), used to flag influential observations. Overall, the model seems **stable**.

```
AIC (Forward Model): -6736.112430911573
BIC (Forward Model): -6680.753026400408
```

Indicates the model's overall fit and penalizes for complexity.
**Lower AIC** suggests a **better balance between accuracy and simplicity**. A highly negative value implies a **strong model fit**.BIC (Bayesian Information Criterion): -6680.75
The negative BIC confirms **good model performance with reasonable complexity**.

# Stepwise Selection:

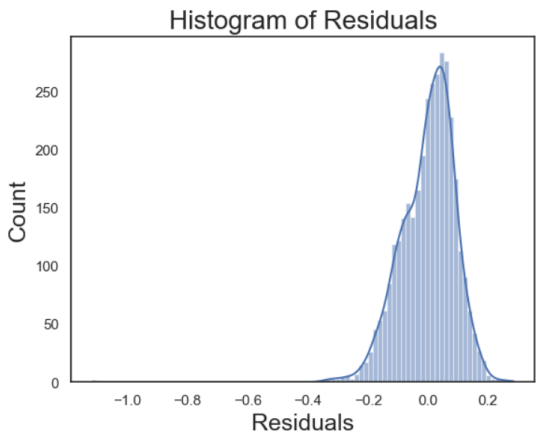
OLS Regression Results

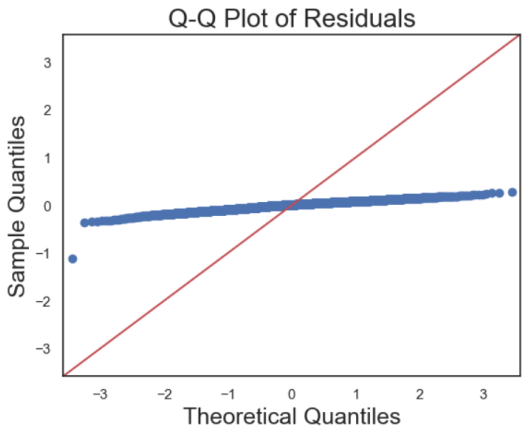
| | | | |
|---|---|---|---|
| Dep. Variable: | price_boxcox | R-squared: | 0.493 |
| Model: | OLS | Adj. R-squared: | 0.491 |
| Method: | Least Squares | F-statistic: | 335.7 |
| Date: | Wed, 16 Apr 2025 | Prob (F-statistic): | 0.00 |
| Time: | 13:45:18 | Log-Likelihood: | 3452.7 |
| No. Observations: | 3467 | AIC: | -6883. |
| Df Residuals: | 3456 | BIC: | -6816. |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| yr_renovated | -0.0005 | 0.003 | -0.146 | 0.884 | -0.007 | 0.006 |
| const | 6.8027 | 0.009 | 772.084 | 0.000 | 6.785 | 6.820 |
| sqft_above | 1.0525 | 0.002 | 500.758 | 0.000 | 1.048 | 1.057 |
| sqft_basement | 0.5657 | 0.002 | 294.714 | 0.000 | 0.562 | 0.569 |
| floors | -0.0339 | 0.005 | -7.098 | 0.000 | -0.043 | -0.025 |
| sqft_living | -0.9825 | 0.002 | -490.147 | 0.000 | -0.986 | -0.979 |
| view | -0.0471 | 0.006 | -7.927 | 0.000 | -0.059 | -0.035 |
| city_mid | -0.0508 | 0.004 | -12.413 | 0.000 | -0.059 | -0.043 |
| bathrooms | -0.0106 | 0.003 | -3.288 | 0.001 | -0.017 | -0.004 |
| bedrooms | 0.0059 | 0.004 | 1.565 | 0.118 | -0.001 | 0.013 |
| yr_built | -0.0525 | 0.008 | -6.931 | 0.000 | -0.067 | -0.038 |
| condition | -0.0186 | 0.006 | -3.291 | 0.001 | -0.030 | -0.008 |

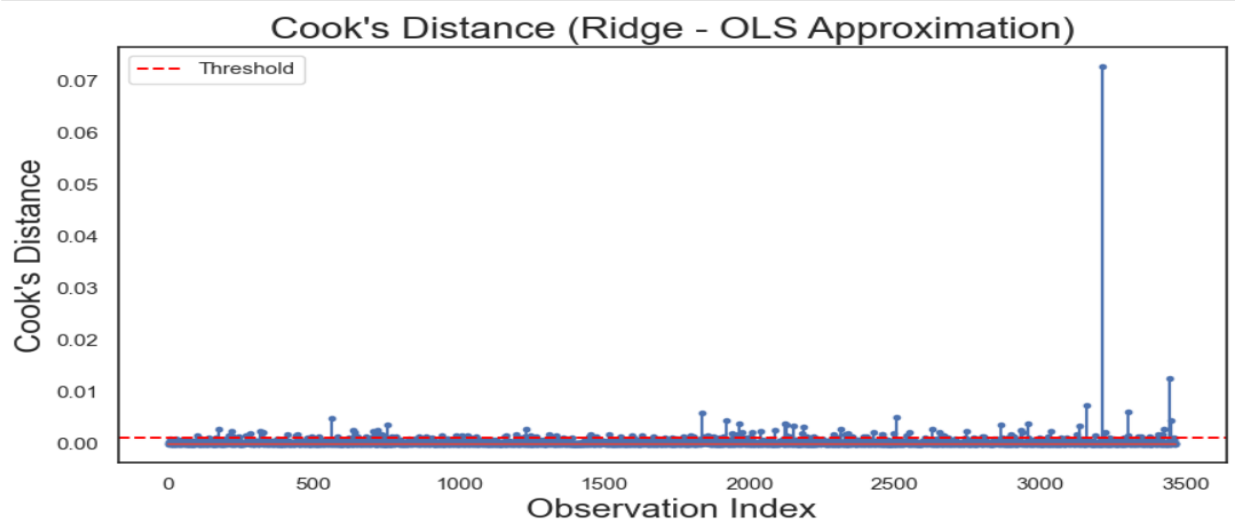
| | | | |
|---|---|---|---|
| Omnibus: | 911.486 | Durbin-Watson: | 1.984 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 7677.381 |
| Skew: | -1.009 | Prob(JB): | 0.00 |
| Kurtosis: | 10.005 | Cond. No. | 2.02e+16 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.52e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.



Residuals are not normally distributed. Q-Q plot shows strong deviation from the line. Histogram is right-skewed.




**Cook's Distance** shows how much each point influences the model.Most points have low influence (below the red threshold line). A few points, especially around index 3300, are **highly influential**. Ridge regression reduces influence overall, but some outliers still exist.

```
Approx AIC (Ridge via OLS): -6903.8084956945495
Approx BIC (Ridge via OLS): -6817.693866454959
```

AIC is -6903.81 and BIC is -6817.69 are both very negative which indicates a **good model fit**.

# Random Forest

```
R² Random Forest : 0.4932697587511653
Adjusted R² Random Forest : 0.48282171253984907
Mean Squared Error (MSE) Random Forest : 0.009259632283938191
```

$R^2$ is 0.4932 that means model explains 49% of the variance that is somewhat moderate fit. Adjusted $R^2$ is 0.4828. This slight drop shows some predictors may not add much value. MSE is 0.00925 that means prediction errors are low.

# Bibliography

**References :**

Notes provided by Monika Ma'am .
Introduction to Linear Regression Analysis by Montgomery , Peck & Vining
Applied Linear Statistical Model by Kutner , Neter .

**Website:**

w.w.w. google .com