|  |
| --- |
| House price analysis in king county |
| Chintan Jikkar |



# Introduction

The housing market plays a pivotal role in both individual financial planning and broader economic trends. King County, home to vibrant urban centers and serene suburban landscapes, offers a diverse range of housing options. Understanding the factors that influence house prices in this region is essential for buyers seeking value, sellers aiming for competitive pricing, and policymakers working to maintain housing affordability. This report delves into the critical determinants of house prices in King County, utilizing data analysis techniques to uncover trends and relationships.

The dataset used in this analysis provides detailed information on house prices and related attributes such as size, location, and structural features. These variables allow us to explore the variability in prices and housing characteristics. For instance, how do living area, lot size, and other size-related metrics correlate with price variability? Identifying these relationships will provide a foundation for deeper analysis and insights.

One key objective of this study is to develop a robust regression model to predict house prices based on these attributes. This model will help pinpoint the factors most strongly associated with pricing and identify their relative importance. Additionally, the analysis will address potential multicollinearity, ensuring that the model's predictions are both accurate and reliable.

The study also examines specific hypotheses that reflect common assumptions in the housing market. First, we test whether houses with waterfront access command higher average prices than those without. This insight could highlight the premium associated with such unique features. Second, we explore the relationship between house age and price, assessing whether older homes tend to have lower valuations. To analyze this, we create an “age” variable based on the year built, aligning with the timeframe of 2014–2015.

Through this analysis, we hope to contribute valuable insights into the dynamics of King County’s housing market, offering data-driven guidance for future decision-making in the real estate sector.

# Descriptive Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Price | Sqft Living | Sqft Lot | Sqft Above | Sqft Basement |
| Mean | $ 540,088.14 | 2079.90 | 15106.97 | 1788.39 | 291.51 |
| Median | $ 450,000.00 | 1910.00 | 7618.00 | 1560.00 | 0.00 |
| Mode | $ 450,000.00 | 1300.00 | 5000.00 | 1300.00 | 0.00 |
| Standard Deviation | 367118.70 | 918.42 | 41419.55 | 828.07 | 442.56 |
| Variance | 134776142225.57 | 843494.65 | 1715579393.30 | 685702.94 | 195863.61 |
| Minimum | $ 75,000.00 | 290.00 | 520.00 | 290.00 | 0.00 |
| Maximum | $ 7,700,000.00 | 13540.00 | 1651359.00 | 9410.00 | 4820.00 |
| Kurtosis | 34.59 | 5.24 | 285.08 | 3.40 | 2.72 |
| Skewness | 4.02 | 1.47 | 13.06 | 1.45 | 1.58 |

Key Insights:

* The Price has high variability depicted by the huge standard deviation of 367118.70 with a vast range from $75,000 to $7,700,000. Kurtosis suggests fatter tails, meaning there are many extreme outliers. Also, the Skewness to the right shows a few high-priced houses pulling the average upward.
* The high standard deviation in the SqFt Lot also shows that lot size varies drastically with a range of 520sqft to 1651359sqft.
* There are quite a few houses that lack a basement with the mean being closer to the median and mode.

# Regression Analysis

The regression model used in this analysis aims to predict house prices in King County based on various factors. Initially, the full model included the following variables: bedroom, sqft\_living, sqft\_lot, floors, waterfront, view, grade, sqft\_above, sqft\_basement, renovation, latitude, longitude and sqft\_living of 15 neighboring houses. These variables were selected because they are expected to influence the price of a property, whether through direct attributes of the house or external factors like neighboring properties.

However, due to the constraints in Excel’s regression tool, which limits the number of variables that can be included in the model to 16, we narrowed the model down to the following variables: bedroom, sqft\_living, floors, waterfront, view, sqft\_above, sqft\_basement, latitude, longitude, sqft\_living of 15 neighboring houses, age and renovation. Despite reducing the number of variables, this subset still captures many of the key attributes that are likely to affect house prices in King County.

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.8138 |
| R Square | 0.6623 |
| Adjusted R Square | 0.6621 |
| Standard Error | 213421.3317 |
| Observations | 21613.0000 |

The **Multiple R** value of 0.8138 suggests a strong positive correlation between the independent variables and the dependent variable (house price). The **R Square** value of 0.6623 indicates that approximately 66.23% of the variance in house prices can be explained by the selected variables. This means the model has a good explanatory power, though there is still a considerable amount of unexplained variance. The **Adjusted R Square** value of 0.6621 adjusts the R square value for the number of predictors in the model, confirming that the model’s performance remains consistent despite the reduced variable set.

The ANOVA table provides a statistical analysis of the variance in house prices explained by the regression model. It shows that the overall regression model is highly significant:

* **F-Statistic**: 2647.2343
* **Significance F**: 0.0000

The p-values for all variables are 0.0000, suggesting that all the predictors are highly significant in explaining the variability in house prices. The coefficients for each variable represent the expected change in the house price for a one-unit change in the respective predictor, assuming all other variables are held constant.

The formula for the regression model can be represented as:

***Price*** *= -54,889,947.6144*

*- 35,384.3187 \* Bedrooms*

*+ 215.9468 \* Sqft\_Living*

*+ 30,872.1882 \* Floors*

*+ 492,112.7982 \* Waterfront*

*- 326,859.3017 \* V1*

*- 214,240.1304 \* V2*

*- 243,303.4499 \* V3*

*- 172,450.7011 \* V4*

*+ 44.9325 \* Sqft\_Above*

*- 242,183.6914 \* SqD1*

*- 258,079.0890 \* SqD2*

*+ 632,828.0089 \* Latitude*

*- 206,302.4893 \* Longitude*

*+ 71.6265 \* Sqft\_Living15*

*+ 1,634.2203 \* Age*

*+ 51,497.3713 \* Reno*

**Hypothesis Testing**

**Hypothesis 1:**

**Null Hypothesis (H₀):** The average price of houses with waterfront is less than or equal to the average price of houses without waterfront.   
**Alternative Hypothesis (Hₐ):** The average price of houses with waterfront is higher than the average price of houses without waterfront.

* **Approach Used:** A two-sample t-test was conducted to compare the mean prices of houses with and without waterfront. The t-test was conducted assuming unequal variances between the two groups.
* **Result:** The p-value for the test was **0** (less than the significance level of 0.05), indicating that the difference between the two groups is statistically significant.
* **Conclusion:** We reject the null hypothesis (H₀) and accept the alternative hypothesis (Hₐ), concluding that houses with waterfront have a significantly higher average price compared to those without waterfront.

**Hypothesis 2:**

**Null Hypothesis (H₀):** The average price of older houses is greater than or equal to the average price of newer houses.   
**Alternative Hypothesis (Hₐ):** Older houses have a lower average price than newer houses.

* **Creation of Age Variable:** The age of each house was calculated by subtracting the year the house was built (yr\_built) from the year of sale (2014 or 2015). Houses were categorized as **older** if their age exceeded 30 years, and as **newer** otherwise.
* **Approach Used:** A two-sample t-test was performed to compare the mean prices of older and newer houses, assuming unequal variances between the two groups.
* **Result:** The p-value for the test was **0** (less than the significance level of 0.05), indicating a statistically significant difference in the prices of older and newer houses.
* **Conclusion:** We reject the null hypothesis (H₀) and accept the alternative hypothesis (Hₐ), concluding that older houses have a significantly lower average price compared to newer houses.

**Summary of Findings:**

Both hypotheses were tested, and the results showed strong evidence to reject the null hypotheses. We conclude that:

1. Houses with waterfront are, on average, priced higher than houses without waterfront.
2. Older houses are, on average, priced lower than newer houses.

These results align with expectations based on housing market trends and preferences.