Seattle Housing Prices EDA

Author: Dean Hawes Date: July 2025

$90,000
Min of Price

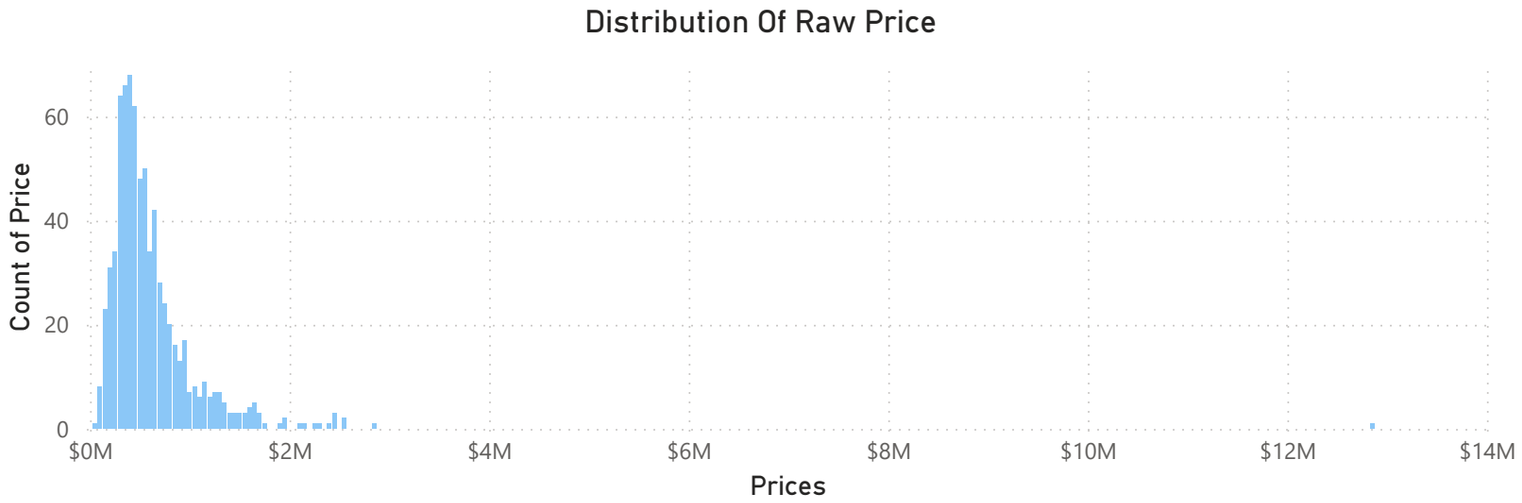
$12,899,000
Max of Price

$582,257
Average of Price

$490,000
Median of Price

1402
Count of Price

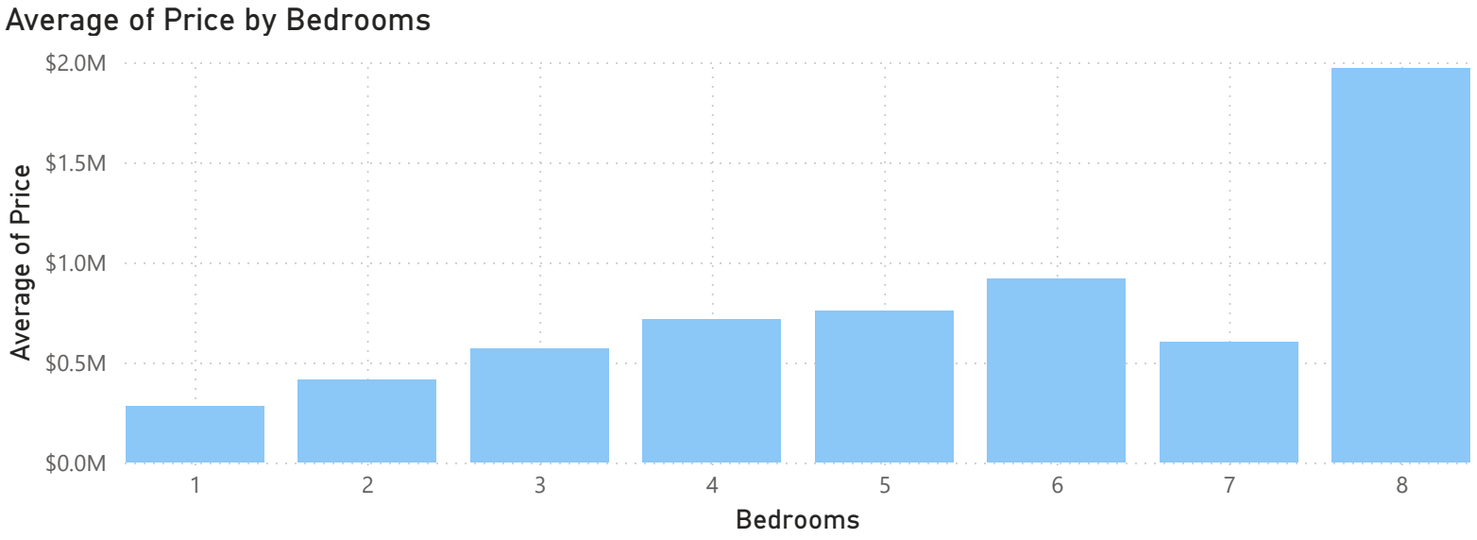
Skewness: 13.18
Kurtosis: 223.18



Bedrooms
All

Bathrooms
All

Statezip
All



📈 **Price is heavily right-skewed** → most houses are clustered in lower price ranges with a few extremely high-priced outliers.

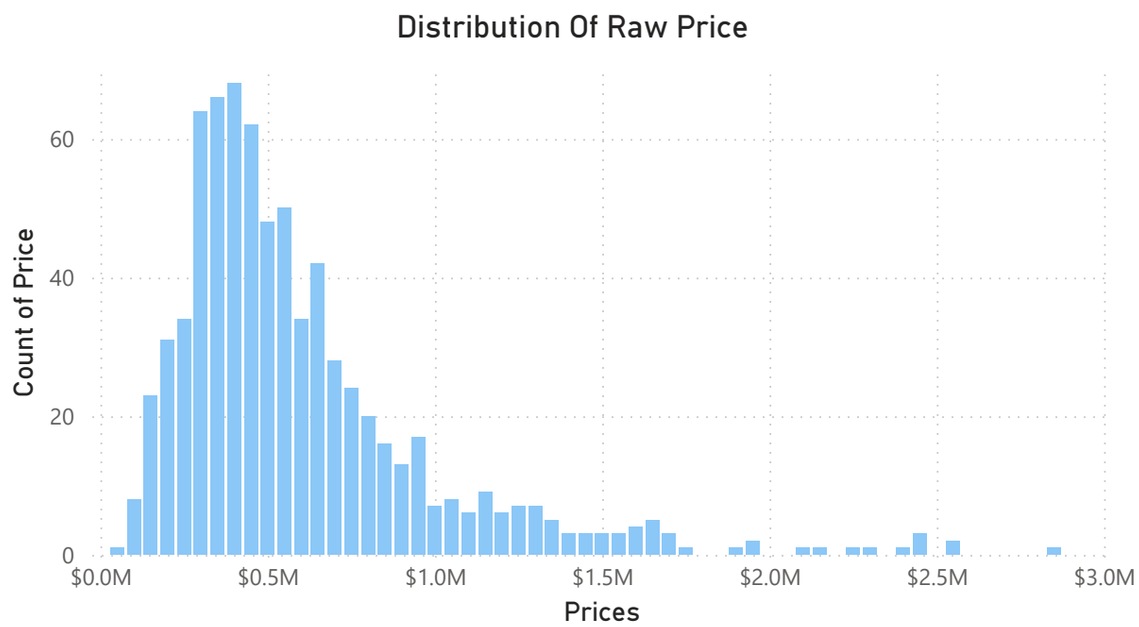
📊 **High skewness & kurtosis** confirm the distribution is not normal — outliers and extreme values stretch the tail far to the right.

⚠️ **Why it matters:** Many statistical models (like linear regression) assume that the target variable is approximately normal to produce reliable predictions.

🔍 **Outliers distort the mean** → can bias model results and weaken predictive power.

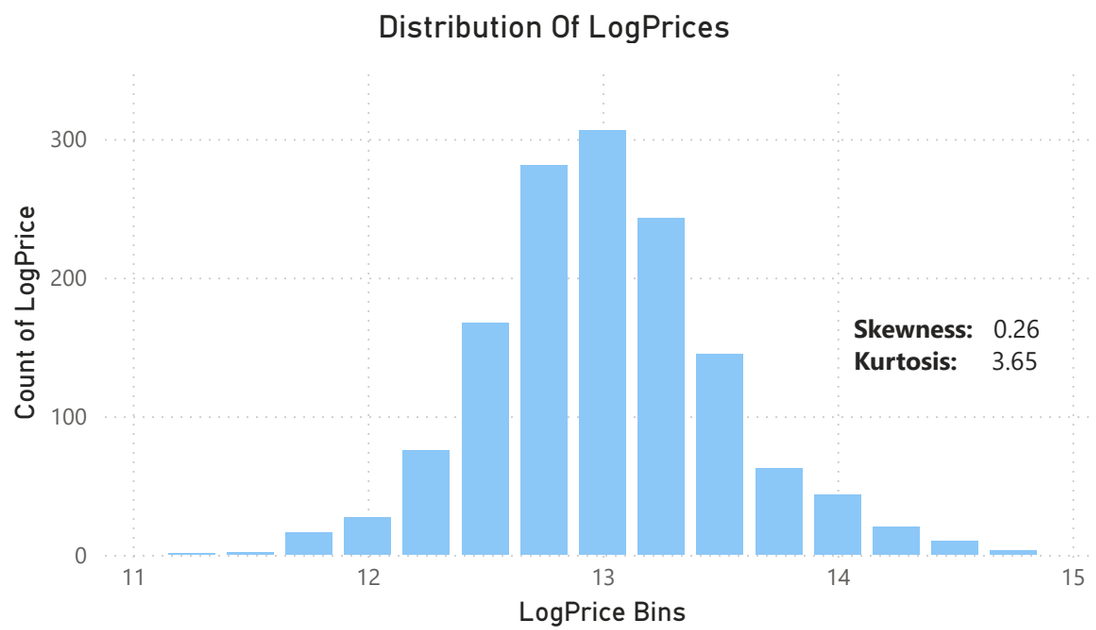
🔑 **Solution:** Remove or limit outliers and apply a **log transform** to stabilize variance, reduce skewness, and approximate a normal distribution.

✅ **Benefit:** Transformed prices better satisfy model assumptions, improving accuracy and interpretability.



Prices

$50,000 | $12,688,003.72015993



0.25	13.13	13.10	0.50
Variance	Mean	Median	Standard deviation

✅ **Log-Transformed Price Insights**

📊 The **log-transformed prices** now show an **approximately normal** distribution.

🔄 **Skewness is low**, meaning the data is **more symmetrical** with no long tail.

📉 **Kurtosis is low**, showing **fewer extreme outliers** than the raw prices.

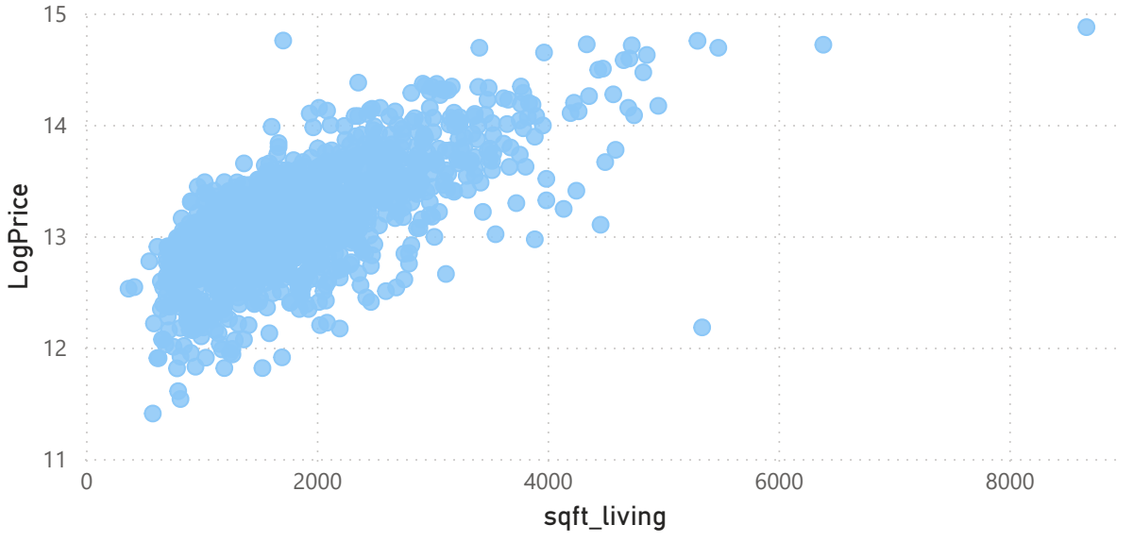
🧮 A **normal-like shape** means **linear models** fit the data more reliably.

📏 This improves key assumptions like **constant variance** and **normality of residuals**.

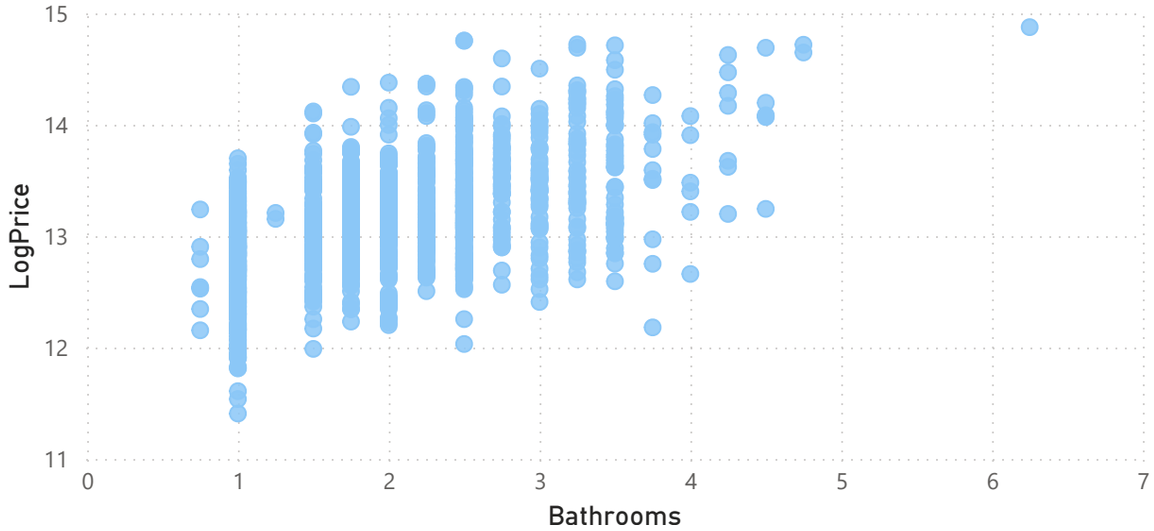
📈 The **linear regression** can now make **better predictions** and give **trustworthy insights**.

MODEL TRAINING

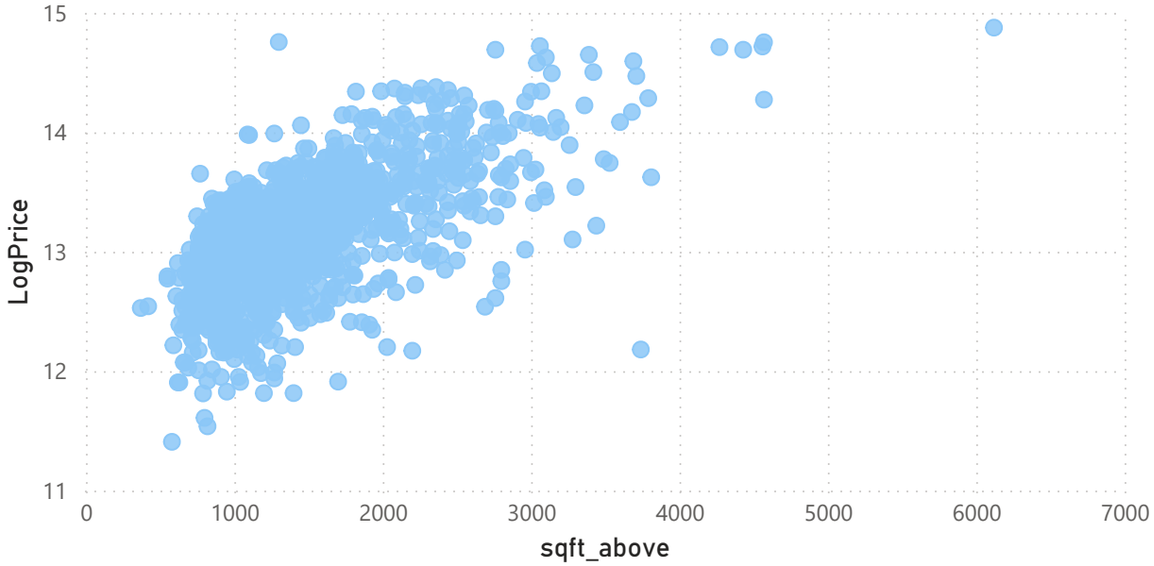
House SQFT vs LogPrice



Bathrooms vs LogPrice



Lot SQFT vs LogPrice



✅ **Modeling Results**

🧮 **Selected key features** (bedrooms, bathrooms, house size, lot size, statezip) that correlate with price.

🗂️ **Split the data** into 80% training and 20% testing sets for fair evaluation.

🤖 **Trained a linear regression model** using the log-transformed price.

📊 **Model metrics: *please see my Jupyter notebook for more insight**

- ✅ **MAE:** 0.22
- ✅ **MSE:** 0.08
- ✅ **R^2:** ~0.59

🔍 Shows the model explains a **good portion of price variation**, but there's **room for improvement**.