Table of Contents

KING COUNTY HOUSE PRICES PREDICTION

1.0 OVERVIEW

This study focuses on using King County House Sales dataset to analyze house sales in a county in the northwest using multiple linear regression modeling. Through an in-depth analysis of King County's house sales data, our objective is to equip local real estate agencies with the knowledge required to address their clients' inquiries, specifically homeowners. This involves identifying the key factors that significantly affect the selling prices of houses and quantifying the extent of their impact on how house improvements may raise the projected value of their properties and the amount of this increase. Local real estate agents frequently encounter questions about the impact of home renovations on house values, the significance of waterfront views, and more. By analyzing King County's house sales data, we aim to assist local real estate agencies, investors, and the general population in providing meaningful answers to these homeowner inquiries. Our objective is to identify the key factors that shape house prices and quantify their influence.

In this industry-driven endeavor, we embark on an analytical journey through King County's real estate domain where our mission is to employ advanced data science techniques to uncover the intricate factors that drive property sales and influence prices.

1.1 Background Information

King County is one of three Washington counties that are included in the Seattle–Tacoma–Bellevue metropolitan statistical area in the United States. It covers an area of 5980 square kilometers with a total of 39 towns and cities. According to Wikipedia, the population as of 2020 was 2,269,675. The average sale price of a home in King County was 815K US dollars in August 2022, up 5.2% since the previous year. The average sale price per square foot in King County is 481k US dollars, up 3.2% since the previous year.

In August 2022, King County home prices were up 5.0% compared to the previous year, selling for a median price of 815K US dollars. On average, homes in King County sell after 15 days on the market compared to 6 days in the previous year year. There were 2,744 homes sold in August 2022, down from 3,877 in 2021. (https://www.redfin.com/county/118/WA/King-County/housing-market).

King County has witnessed a vibrant real estate market, further accelerated by economic factors and lifestyle changes. The housing market in this county is competitive, and homeowners looking to buy or sell properties frequently seek guidance from local real estate agencies. One common concern among homeowners is the potential value added by home renovations. This project aims to provide data-driven insights into this issue, helping homeowners make informed decisions on their properties. The King County House Sales dataset serves as a valuable resource for this analysis.

1.2 Domain

The domain of this project is the real estate industry, particularly focusing on local real estate agencies that work with homeowners in a northwestern county. These agencies serve as intermediaries for homeowners looking to sell their properties or individuals seeking to purchase homes in the region. The project addresses the homeowners' need for accurate information about how home renovations can impact property values, enabling them to make informed decisions in a highly competitive and evolving real estate market.

1.3 Challenges and solutions

Some of the challenges to be expected in this project include:

i) Competitive Pricing: Pricing a property competitively is crucial. Overpricing can deter potential buyers, while under-pricing can lead to a loss of profit.

ii) Location and Neighbourhood Analysis: Property value is highly dependent on the location and the neighborhood’s quality. Homeowners may not fully understand how their property's location affects its value.

iii) Data-Driven Marketing: To attract buyers, it's essential to have data-driven marketing strategies. This includes knowing which property features to highlight in listings and marketing materials.

To curb the above challenges, we can use regression models to analyze the dataset and determine the optimal listing price for a property based on its characteristics and market conditions. This ensures that homeowners receive offers that align with their property's true value.

Also, we can utilize geospatial data and the dataset to provide insights into how location-specific factors impact property values. This can help homeowners position their properties effectively in the market.

Finally, we can use regression analysis to identify the key features that contribute significantly to property value, allowing them to tailor their marketing strategies and highlight these features to potential buyers.

By addressing these challenges and providing data-driven solutions, the real estate agency can better serve their stakeholders—homeowners looking to sell their properties—and enhance their reputation as a trusted advisor in the real estate industry. This approach will lead to more successful property transactions and increased client satisfaction.

2.0 BUSINESS UNDERSTANDING

2.1 Problem statement

The real estate agency, serving as the intermediary for homeowners seeking to either purchase or sell homes within King County, faces a complex challenge that revolves around the precision of assessing the multitude of factors influencing house prices in the area. Specifically, the agencies need to identify the key predictors of house prices and provide data-driven insights to homeowners, enabling them to make informed decisions about home renovations. The primary objective is to maximize the estimated value of homes based on these key predictors. To do so, we must address the following questions:

i) What are the primary factors influencing home prices?
ii) Is a waterfront view a price influencer?
iii) The relationship between the number of bedrooms and home prices?
iv) Does the condition of the house impact it's price?
v) Is the value of the house affected by its square foot?

2.2 Audiences

Our Audiences include:

- Real Estate Professionals: Industry experts seeking insights into market dynamics.
- Investors: Those aiming to make data-informed investment decisions.
- General Public

- Home Owners and Buyers

2.3 Objectives

Our objectives include:

Main:

1. To predict house prices in King County.

Develop a robust predictive model to estimate house prices accurately based on a range of factors, including property attributes, location, and other key features.

Others:

2. To identify key house prices determinants.

To identify and quantify the most significant factors that influence house prices in the King County housing market. This analysis will help homeowners and real estate professionals understand what drives property values.

3. To analyze the impact of waterfront view on house prices

Investigate whether properties with waterfront views tend to have higher or lower sale prices and assess whether areas exhibit unique pricing trends.

4. To Determine if the condition of a house affects its price.

We want to understand if the state and overall well-being of a property plays a significant role in determining its market value.

5. To Explore the influence of house grade on its price.

Grade is an indicator of construction and design quality on house. Exploring this on the houses within in King County housing market will help us determine whether higher-grade houses tend to have higher selling prices.

By achieving these objectives, the real estate agency will empower homeowners with the knowledge and guidance needed to maximize the estimated value of their homes. This will not only benefit homeowners but also enhance the agency's reputation as a trusted partner in the real estate industry, leading to more successful property transactions and satisfied clients.

3. 0 DATA UNDERSTANDING AND PREPARATION

3.1 Data Understanding

To successfully address the business problem of providing advice to homeowners about how home renovations might increase the estimated value of their homes, it's crucial to gain a comprehensive understanding of the King County House Sales dataset. This understanding involves:

1. Data Source: The dataset, "kc_house_data.csv," is the primary source of information. It contains records of house sales in King County and provides insights into various aspects of each property.
2. Column Descriptions: The descriptions of column names, as provided in "column_names.md," serve as a guide to understanding the dataset's attributes. These descriptions offer insights into the type of information each column holds.
3. Data Fields: An exploration of the dataset's columns is essential to understand what information is available. Key fields may include:
    - ID: A unique identifier for each property.
    - Date: The date of the sale.
    - Price: The sale price of the property.
    - Bedrooms and Bathrooms: The number of bedrooms and bathrooms in the property.
    - Sqft Living and Sqft Lot: The square footage of living space and the total lot size.
    - Floors: The number of floors in the property.
    - Waterfront: A binary indicator (0 or 1) for whether the property is waterfront.
    - View: A rating of the view from the property.
    - Condition: The overall condition of the property.
    - Grade: A rating of the overall grade of the property.
    - Sqft Above and Sqft Basement: The square footage above and below ground.

- Year Built: The year the property was built.
- Year Renovated: The year of any renovations.
- Zipcode: The property's zipcode.
- Latitude and Longitude: Geographical coordinates of the property.
- Square Footage of Living Space in 2015: The square footage of living space in 2015.
- Lot Square Footage in 2015: The total lot size in 2015.
- Basement Finished Percentage: The percentage of the basement that is finished.
- Age: The age of the property at the time of sale.

4. Data Distribution: Understanding the distribution of data within each column is crucial. This involves examining summary statistics, such as mean, median, standard deviation, and the range of values.
5. Data Quality: Assess the quality of the data, including identifying missing values, outliers, and any data cleaning or preprocessing requirements. Data quality is critical for building a robust regression model.
6. Data Relationships: Explore relationships between different attributes. For example, how do the number of bedrooms or bathrooms relate to the price of a property? Are there any apparent correlations between variables like condition, grade, and price?
7. Temporal Aspects: Consider how time-related factors, such as the date of sale and the year of construction or renovation, may impact property prices. Are there any noticeable trends over time?
8. Geospatial Analysis: Analyze the geographic components, such as latitude, longitude, and zipcode, to understand how location influences property values within the county.
9. Potential Additional Data Sources: Consider if external data sources, such as economic indicators or local real estate market trends, can provide supplementary insights to enhance the analysis.

By thoroughly understanding the data, the real estate agency can make informed decisions about how to approach the task of advising homeowners on home renovations and their potential impact on property values in King County. This foundational knowledge will guide the subsequent steps of data analysis and modeling.

3.2 Data Preparation / Cleaning

Data preparation and cleaning are crucial steps to ensure the accuracy and reliability of the analysis. In this project, we will perform the following data preparation and cleaning tasks:

1. Data Loading: Load the King County House Sales dataset from the provided CSV file (kc_house_data.csv) into a suitable data analysis environment or tool (e.g., Python with Pandas).
2. Handling Missing Values: Identify and handle missing values in the dataset. Depending on the extent of missing data, you may choose to remove rows with missing values, impute missing values using appropriate methods, or use domain knowledge to fill in missing information.
3. Data Type Conversion: Examine the data types of each column and convert them as needed. For instance, ensure that date information is in a datetime format, and numerical columns are appropriately encoded.
4. Outlier Detection and Handling: Identify and address outliers in the dataset. Outliers can significantly impact the results of the analysis, so it's important to decide whether to remove or transform them based on domain knowledge and statistical techniques.
5. Data Exploration: Conduct exploratory data analysis to gain a deeper understanding of the dataset and identify any patterns or relationships between variables.

Preparing and cleaning the data will lay a strong foundation for the subsequent multiple linear regression analysis, ensuring that the results are accurate, meaningful, and actionable for providing advice to homeowners about the impact of home renovations on the estimated value of their homes.

4.1 DATA ANALYSIS AND MODELING

4.1 DATA ANALYSIS

Data analysis is the process of inspecting, cleaning, transforming, and modeling data to discover meaningful patterns, draw inferences, and support decision-making. It is a crucial step in any project because it helps in drawing inferences that can answer research questions or make predictions.

The aspects of Data analysis carried out in this project include:

1. EDA: Present key insights and visualizations that enabled a better understanding of the data. Provide the relevant summary statistics, distributions and correlations of the data points present.
2. Feature Selection: Identify the independent variables to use for building a regression model. and handle missing values in the dataset.

Findings

Under EDA, data distributions were identified for each of the variables that were listed as columns in the data set. Most columns were skewed either to the left or right indicating a presence of outliers in the data.

Feature Selection was done by the use of a correlation Heatmap that outlined the correlation of the variables with each other. A positive correlation hints at a good relationship whereas a negative correlation hints on an opposite relationship.

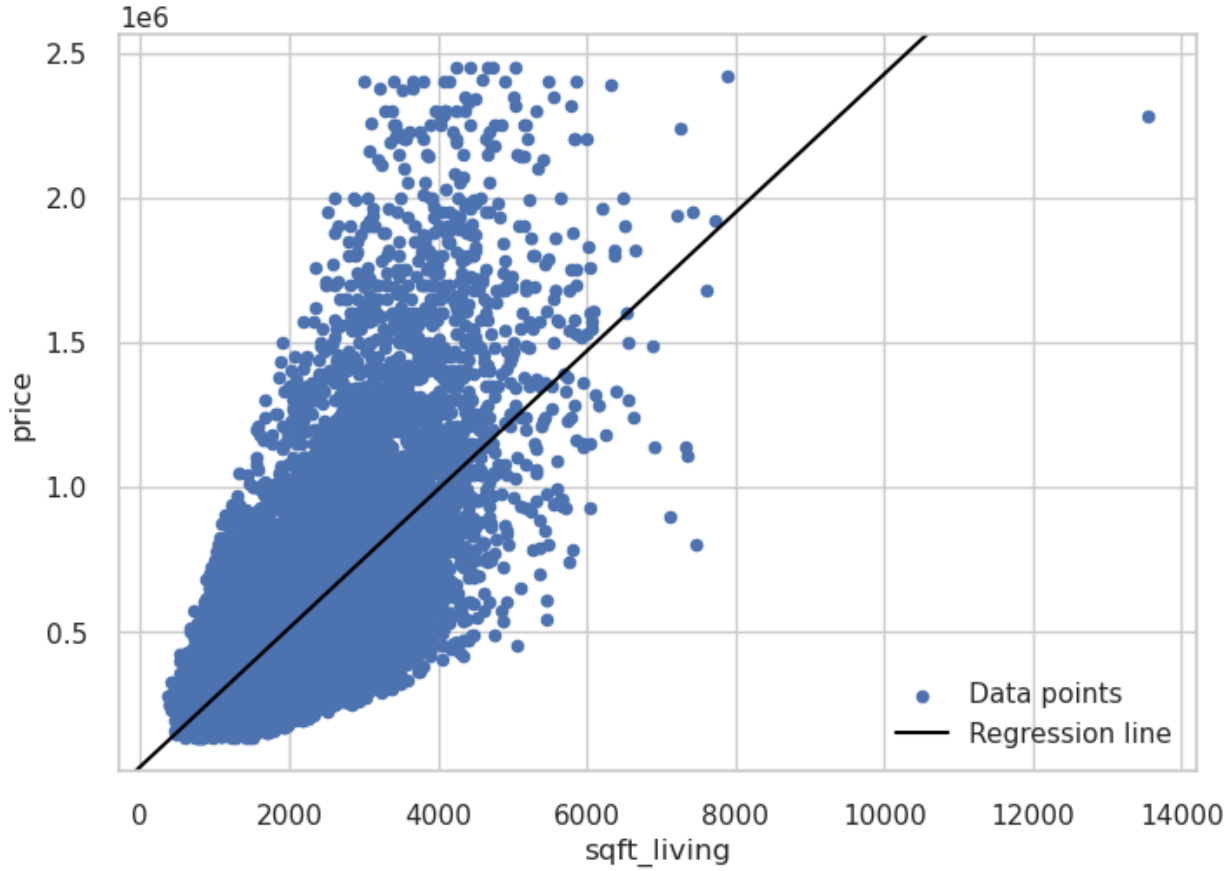
The selected columns were columns with the highest correlation; sqft_living- 0.704652, grade- 0.667224, bathrooms-0.527350

4.2 MODELING

Modeling is a cornerstone of regression and data science, serving as a powerful tool for extracting meaning from data, making predictions, uncovering hidden patterns, and facilitating data-driven decision-making. These models enable us to gain valuable insights, optimize processes, and enhance our understanding of complex systems.

The square foot living column had the highest linear correlation with price and was used as the first linear model.

Creation of a linear regression model was based on evaluating an independent variable that has



the highest number of correlation with price.

The model was statistically insignificant based on the r-squared value of 47. It also had a very small F-statistic to indicate that the grade alone is not enough to account for the varying house prices in KIng County.

The square foot living coefficient is about 239, suggesting that as the living space increases price tends to increase by $239 per square foot.

This infers that although the grading of the house matters, it is not enough to conclude that it is a key determinant of price.

On model evaluation we achieved a higher MSE on the train data compared to the test data.

MULTIPLE LINEAR REGRESSION

This introduces multiple independent variables and allows us to use them to identify whether they collectively influence the outcome of a target variable. In our case we are trying to analyze a multiple of factors that may be key determinants of price.

We are taking into account that the results of the previous model don't quite explain the data well. The multiple linear regression model uses various columns to try and achieve a significant model: sqft_living, grade, number of bathrooms, bedrooms, floors, condition, waterfront and square foot lot.

On plotting a multiple regression model it achieved an r squared value of 57, which means that the model explains more of the variance in the target variable.

The model also proved significance through an F-statistic of 2672 and a corresponding p-value of 0.00.

The model intercept was however uninterpretable and to change this zero centering was done to all the variables that were used in the multiple regression model.

Zero centering allows all the variable values to be accounted for as means of the whole column.

Applying zero entering to all the predictor columns used in the model shifted the intercept from -7.049e+05 to $528969 which meant that for a house with an average number of bedrooms , bathrooms, square footage living area and square footage lot area , floors and condition; we approximate the price to be around $528969, which is more or less the mean the price column.

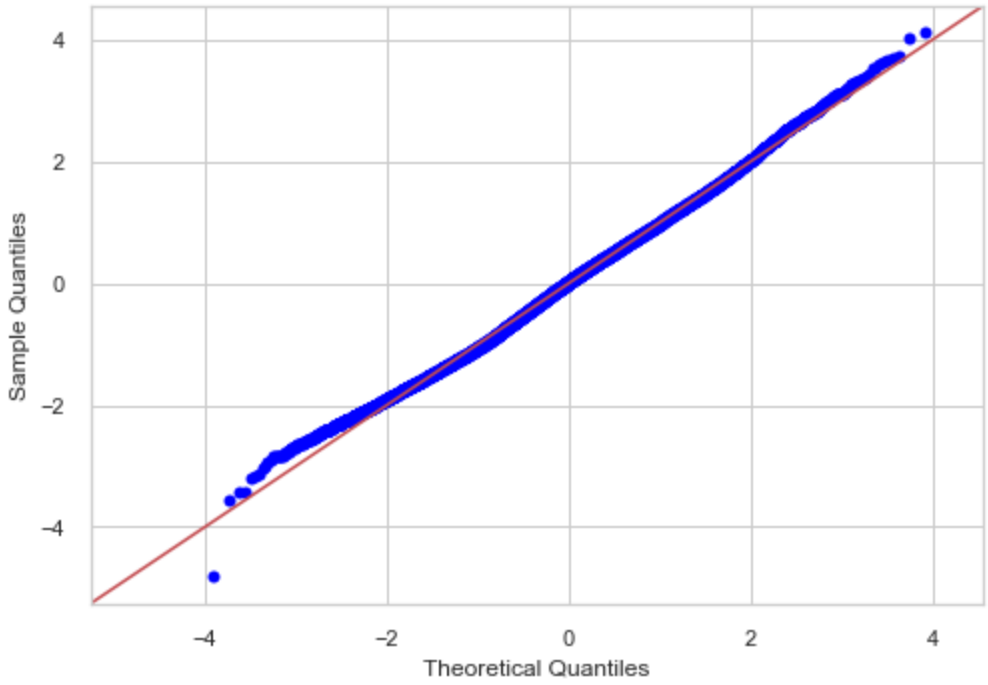
Check for Normality and heteroscedasticity

A check for normality is a statistical test or visual inspection used to assess whether a dataset follows a normal distribution, also known as a Gaussian distribution or a bell-shaped curve. The normal distribution is characterized by several key properties, including:

- Symmetry: The distribution is symmetric, with the mean, median, and mode being equal.
- Bell-shaped: The data forms a symmetric, unimodal (single peak) shape resembling a bell.
- Mean and Standard Deviation: The distribution is fully characterized by its mean (average) and standard deviation.

For the multiple regression model Log transformation was applied to normalize the data and treat heteroscedasticity. Log transforming a model can be a useful technique when dealing with data that is not normally distributed or exhibits heteroscedasticity (unequal variance) and can help address these issues. Log transformations are a specific type of mathematical transformation that can make the data more approximately normal or stabilize the variance.

By logging the model results normality and heteroscedasticity was achieved.



MODEL EVALUATION

Model evaluation involves assessing the performance of a trained model to determine how well it generalizes unseen data and whether it meets desired criteria for a specific problem.

Here's how training and test data are used in model evaluation:

1. Data Splitting:

The dataset is typically divided into two subsets: the training dataset and the test dataset.The training dataset is used to train the machine learning model. It is the data that the model uses to learn patterns, relationships, and make adjustments to its parameters.The test dataset is kept separate and is used to evaluate the model's performance. It represents data that the model has not seen during training.

2. Model Training:

During the training phase, the machine learning model learns to make predictions or classifications based on the features (input variables) in the training data.The model iteratively adjusts its internal parameters to minimize the difference between its predictions and the actual target values in the training data.

3. Model Testing:

After the model is trained, it is evaluated using the test dataset. The model makes predictions on the test data, and these predictions are compared to the actual target values.The evaluation metrics (e.g., Mean Squared Error, Accuracy, Precision, Recall) are computed to assess how well the model performs on the test data.

4. Generalization Assessment:

The primary goal of using separate training and test datasets is to assess the model's ability to generalize to new, unseen data. A model that performs well on the training data but poorly on the test data may be overfitting (fitting the noise in the training data). A well-generalizing model should provide reasonable predictions on the test data, indicating that it has learned meaningful patterns and relationships rather than memorizing the training data.

5. Performance Metrics:

The choice of performance metrics depends on the problem type. For regression tasks, metrics like Mean Squared Error (MSE) and R-squared are common. For classification tasks, metrics like Accuracy, Precision, Recall, and F1-score are used. The model's performance on the test data is typically summarized using one or more of these metrics.

6. Comparing Models:

If you are considering multiple models, you can use the same test data to compare their performance. This helps you select the best-performing model for your specific problem.

For our models:

1. The baseline model:

Train MSE: 50975099105 while Test MSE was: 4969866928. These values suggest that the baseline model, which likely uses "sqft_living" as the sole predictor, has relatively high errors when predicting house prices.

The model's predictions are quite off from the actual prices, both in the training and test datasets. This indicates that the baseline model may not capture the underlying patterns in the data and may not be a good model for the given problem. Further model improvement or exploration may be necessary to achieve better predictive accuracy.

2. The log transformed model:

Train MSE:0.11317968205719839

A lower Train MSE is generally better. A Train MSE of 0.113 indicates that, on average, the model's predictions on the training data are off by about 0.113 units when squared. Lower values mean the model's predictions are closer to the actual values.

Test MSE of : 0.11317968205719839

A Test MSE of 0.111 indicates that, on average, the model's predictions on the test data (which may also be in log-transformed form) are off by about 0.111 units when squared.

Lower MSE values indicate that the model's predictions are more accurate and closely match the original log-transformed target values.

5.0 CONCLUSION AND RECOMMENDATION

5.1 Conclusion

- From our data we can conclude that price is mainly determined by waterfront, sqft_living, grade.
- Most houses have an average of 3 bedrooms.
- For a house with an average number of bedrooms, bathrooms, sqft_living, area, sqft_lot area, waterfront, grade, floors and condition we would expect the price to be approximately $528969. Most houses in our data are of average condition.

5.2 Recommendation

1. Implement a pricing strategy that considers factors such as the specific location along the waterfront and any unique features of the property. Different waterfront properties may command different price points. A higher price for houses with waterfront views provides an opportunity for real estate agencies to specialize and cater to a niche market
2. Encourage sellers to invest in property improvements and maintenance to increase property value. Provide guidance on home staging and presentation to attract buyers since most of the houses are of average condition.
3. Larger living spaces (sqft_living) can command higher prices. Sellers of properties with spacious living areas should emphasize this feature in listings. Buyers should prioritize properties that offer the desired living space.
4. Buyers interested in high-quality properties should prioritize those with a higher grade. Sellers should invest in improving the grade of their properties to increase market appeal