PREDICTING HOUSE PRICES IN KING COUNTY: A MULTIPLE LINEAR REGRESSION APPROACH.

PROJECT BY:

- Kelvin Rotich
- Grace Mutuku
- Joy Ogutu
- Peter Otieno
- Shuaib Mahamud

INTRODUCTION

Welcome to this presentation on predicting house prices in King County, Washington. In this project, we leverage a Multiple Linear Regression model to gain insights into the dynamic real estate market of this thriving region. Our analysis is based on a comprehensive dataset from King County, which encompasses a multitude of factors influencing property prices.

The primary purpose of our project is to develop a predictive model that accurately estimates house prices in King County. We aim to provide valuable insights to various stakeholders, including homeowners, real estate agents, investors, and developers, to help them make informed decisions regarding property pricing

In the following sections, we will delve into our project's methodology, key findings, and the model's performance in predicting house prices. Thank you for joining us in this exploration of King County's real estate landscape.

PROJECT OVERVIEW

King County’s Real Estate Background

The real estate market in King County, Washington, is known for its dynamism and diversity. This thriving area boasts a robust economy driven by tech giants like Amazon, Microsoft, and Boeing, which continually attract a large workforce. However, King County's real estate market, while vibrant, is not without its complexities and challenges. One defining characteristic of King County's real estate market is the ever-present demand for homes. Thus, there is a growing interest in sustainable housing options. This robust demand has contributed to an ongoing conundrum: affordability. As more individuals seek housing, the supply-demand balance skews, leaving many residents struggling to find reasonably priced homes. As King County experiences urban expansion, the competition extends beyond traditional housing. The development of sustainable, modern, and environmentally friendly housing solutions is another front where stakeholders compete.

The King County real estate market is highly competitive, with various stakeholders such as developers, online platforms, and established real estate companies vying for their share of the market. To navigate the complexities of this environment effectively, it's crucial for local stakeholders to understand the dynamics of the real estate market.

To comprehend this intricate landscape, stakeholders must understand the factors influencing property prices, most of which align with the columns provided in our dataset:

1. Property-specific attributes, including location, size, condition, and amenities.
2. Market dynamics, which encompass supply and demand, interest rates, and broader economic conditions.
3. External factors, such as neighborhood characteristics and government policies.

In response to these challenges, our project seeks to develop a predictive multilinear regression model, utilizing the dataset at hand. Real estate agents can provide more accurate pricing guidance and develop effective marketing strategies. Homeowners can make informed decisions when pricing their properties, and investors and developers can identify promising opportunities to maximize their returns.

As we delve into the details of our predictive model, we will explore the methodology, key findings, and the model's performance. We will uncover how specific features and factors, such as the number of bedrooms, quality of views, and location, influence house prices in King County.

Problem Statement

In the dynamic real estate market of King County, Washington, where economic conditions, housing demand, and external influences drive property prices, the importance of accurate pricing cannot be overstated. The ever-present demand for homes, fueled by the presence of major tech companies and a stream of workers, creates a competitive environment. However, the challenge of accurately pricing properties in this competitive landscape can sometimes lead to overpricing. Sellers, eager to maximize their returns in a high-demand market, may set initial prices that are higher than what the market can sustain. This overpricing can, in turn, deter potential buyers and extend the time properties spend on the market. Therefore, the need for accurate pricing models that consider all relevant factors, including property conditions becomes paramount in ensuring that homes in King County are competitively priced, facilitating smoother transactions for both buyers and sellers.

Objectives

1. *Objective:* To explore and analyze the impact of numeric attributes on house prices in King County.
   Identify which numeric features have the most significant influence on pricing.Provide insights into how each unit increase or decrease in these attributes affects the final sale price. Generate recommendations for homeowners, buyers, and investors to optimize property attributes and investments based on numeric data.

2. *Objective:* To investigate the influence of categorical attributes on house prices.
   Determine which categorical features, such as being on a waterfront or having a high-grade rating, command premium prices. Provide recommendations on how to leverage these categorical attributes to maximize property values. Assist stakeholders in making informed decisions based on categorical data.

3. *Objective:* To create a precise property valuation model that calculates the cost of homes depending on a range of characteristics.
   This model will utilize a property's characteristics, including but not limited to bedrooms, square footage, and more. By carefully selecting and incorporating these features, we intend to build a model that accurately reflects the diverse attributes influencing house prices.

DATA UNDERSTANDING

This dataset is housed in the *kc_house_data.csv* file within the project's data folder and the columns outlined in the accompanying *column_names.md* file, include:

- *id* - Unique identifier for a house
- *date* - Date house was sold
- *price* - Sale price (prediction target)
- *bedrooms* - Number of bedrooms
- *bathrooms*- Number of bathrooms
- *sqft_living* - Square footage of living space in the home
- *sqft_lot* - Square footage of the lot
- *floors* - Number of floors (levels) in house
- *waterfront* - Whether the house is on a waterfront
- *view* - Quality of view from house
- *condition* - How good the overall condition of the house is. Related to maintenance of house.
- *grade* - Overall grade of the house. Related to the construction and design of the house.
- *sqft_above* - Square footage of house apart from basement
- *sqft_basement* - Square footage of the basement
- *yr_built* - Year when house was built
- *yr_renovated* - Year when house was renovated
- *zipcode* - ZIP Code used by the United States Postal Service
- *lat* - Latitude coordinate
- *long* - Longitude coordinate

- *sqft_living15* - The square footage of interior housing living space for the nearest 15 neighbors
- *sqft_lot15* - The square footage of the land lots of the nearest 15 neighbors.

By gaining a deep understanding of these columns and their relationships, we aim to pave the way for accurate property pricing and well-informed decisions for homeowners, real estate agents, and investors alike.

The dataset contains 21 columns and 21, 597 entries with sale prices and details of the houses sold from *2nd May 2014* to *27th May 2015* .

DATA CLEANING AND PREPARATION

Data preparation is a crucial stage in this project for a number of reasons:

- *Feature Engineering:* New features might be developed or current ones modified in order to improve analysis.
- *Handling Missing Data:* Analysis results can be greatly impacted by missing data. In order to achieve a robust analysis, handling missing values must be decided, whether through imputation, deletion, or other suitable approaches.
- *Outlier detection:* For statistical validity, outliers must be found and dealt with. We can use methods like visual inspection or statistical testing to find outliers and handle them correctly with the help of data preparation.

In the data preparation phase, several important actions were taken to ensure the dataset was ready for exploratory data analysis (EDA) and subsequent modeling:

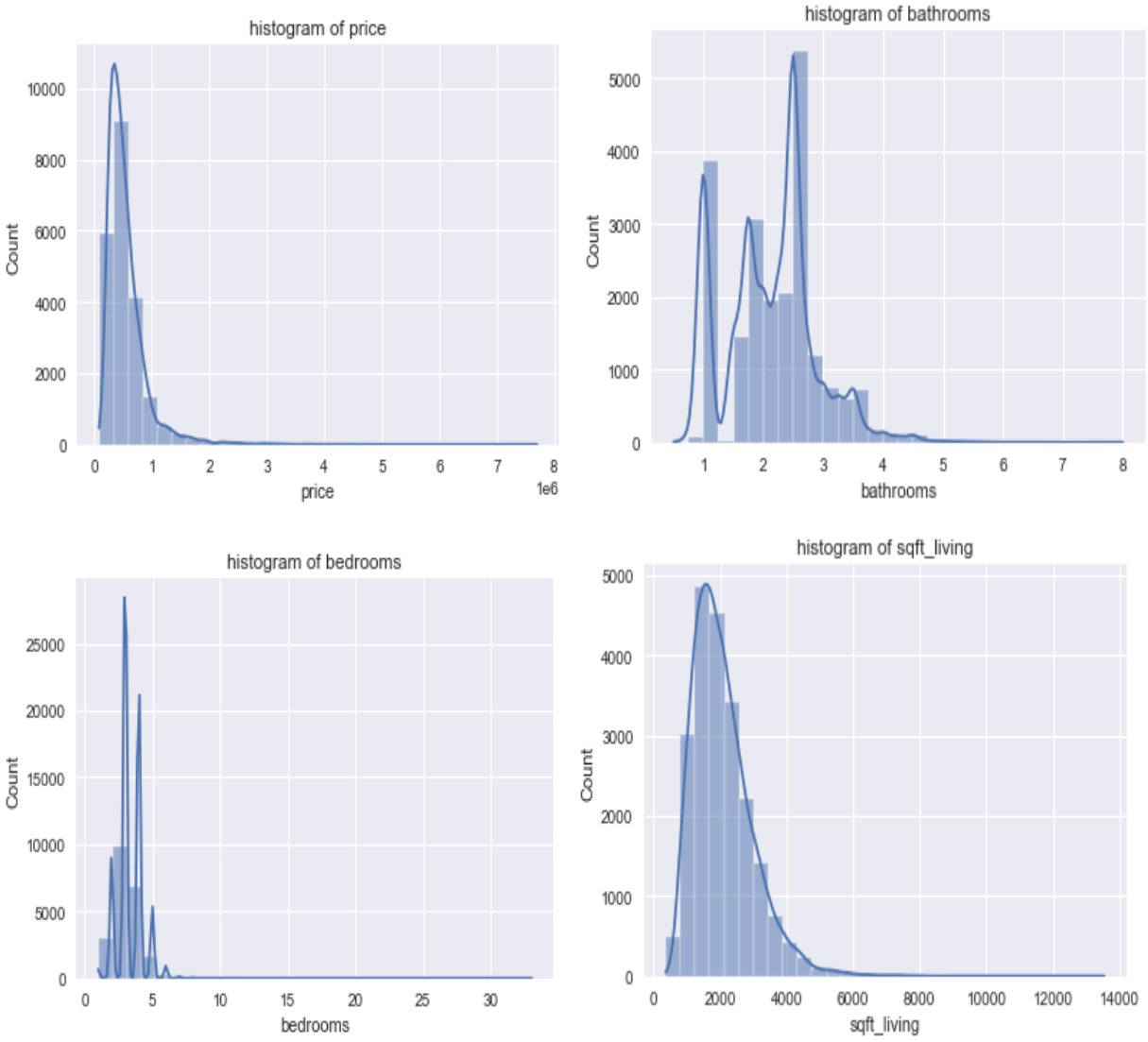
1. Null value counts and percentages were initially checked, revealing that most columns had no null values except for *'waterfront'*, *'view'* and *'yr_renovated'*, each with less than 30% missing data. Rather than dropping these columns, null values were replaced: *'waterfront'* and *'view'* were set to "unknown," while *'yr_renovated'* was imputed with the most common value, "0."
2. Duplicate entries were identified based on the *'id'* column, resulting in 177 flagged duplicates. Further investigation revealed that some houses were sold multiple times at different prices and times. These duplicate entries were retained for feature engineering. The *'date'* column was converted to a datetime format, and a new *'seasons'* column was introduced. We also created a new column called 'house_age_lv' which shows whether the house was newly built, houses built 25 years ago or more and houses built 50 years or more. This was made as a result of feature engineering on the 'yr_built' column.
3. Outliers were retained, as they represented genuine property attributes with valuable pricing information. Placeholder values were also addressed, with the *'?'* placeholder in the *'sqft_basement'* column replaced with *'0.0'*, the most common value.

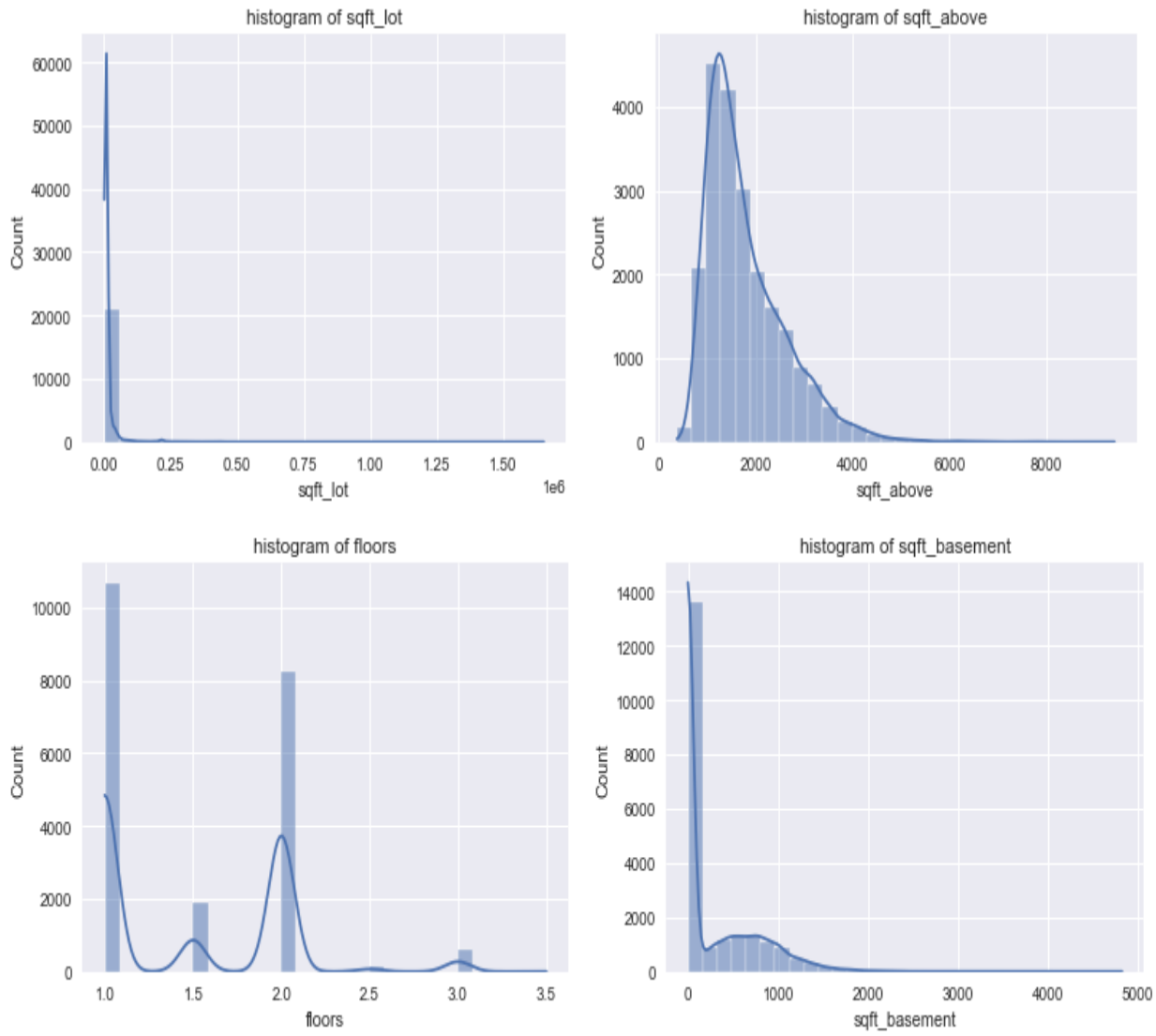
After these preparations, the dataset was confirmed to be ready for EDA, ensuring that the data was cleaned, imputed, and structured for meaningful analysis and model development.

EXPLORATORY DATA ANALYSIS

A crucial turning point in our investigation occurs during the data analysis stage, which enables us to extract significant patterns and insights from the compiled dataset. We used univariate, bivariate, and multivariate exploratory data analysis (EDA) methodologies in a comprehensive manner to properly analyze the data. Let's dive into our analysis.

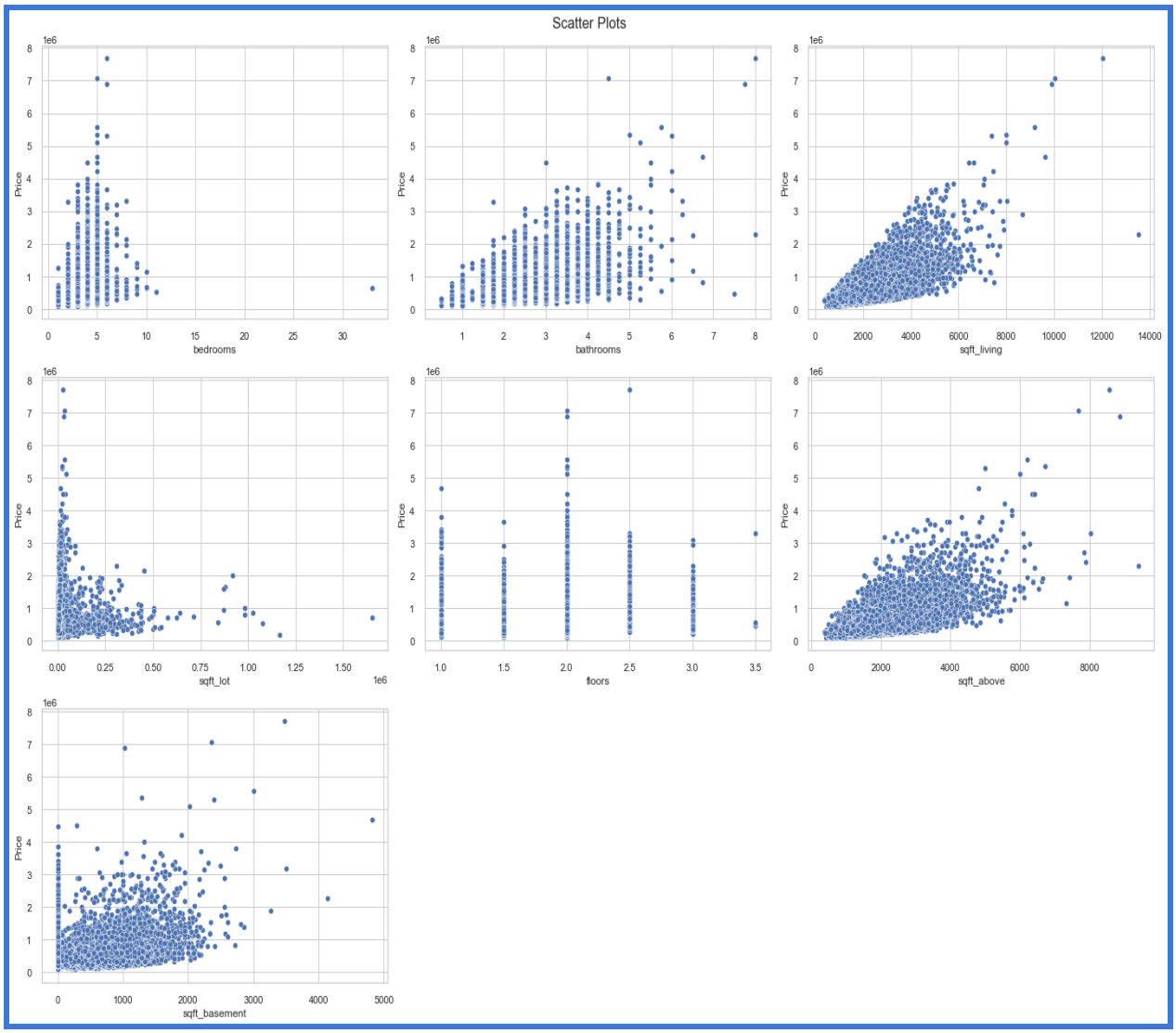
Objective: **To explore and analyze the impact of numeric attributes on house prices in King County.**





The above histograms provide concise representation of how data is spread across different values and helps reveal underlying patterns.
For the mentioned columns – *'price'*, *'bedroom'*, *'bathrooms'*, *'sqft_living'*, *'sqft_above'* and *'sqft_basement'*. The key observation is that these histograms exhibit a right-skewed pattern, as most of the values are concentrated on the left side with a tail extending to the right. This skewness indicates that, for these attributes, the majority of properties fall within a specific range or have certain characteristics, but there are relatively few properties with exceptionally high values.

This information is essential for understanding the price distribution and the distribution of other property characteristics. It guides the project's analysis and decision-making, particularly when dealing with properties that deviate from the norm in terms of pricing and attribute values.



A scatter plot's primary purpose is to provide a clear and concise representation of how two numeric variables interact with each other, which can help uncover patterns, correlations, or anomalies in the data.

In the context of this project, scatter plots are utilized to examine the relationships between the '*price*' (target variable) and various attributes such as '*bathrooms*', '*sqft_living*', '*sqft_above*', '*sqft_basement*', '*bedroom*', '*sqft_lot*', and '*floors*'.

Here's how are the observations related to these columns using scatter plots in relation to price:

- *'bathrooms'*, *'sqft_living'*, *'sqft_above'*, *'sqft_basement'* with *'price'*:
  Scatter plots for these attributes in relation to price reveal a linear correlation. As the number of bathrooms, the square footage of living space, the square footage above ground, or the square footage of the basement increases, there is a corresponding increase in the property price. This indicates a positive linear relationship, suggesting that these attributes have a direct impact on the pricing of houses.

- *'bedroom'* and *'sqft_lot'* with *'price'*:
  In contrast, scatter plots for the bedroom count and square footage of the lot in relation to price show a different pattern. These attributes are more clustered around a specific range and do not portray a clear linear relationship with price. This suggests that, for these attributes, there may be other factors influencing pricing, and they are not as directly correlated with the price.

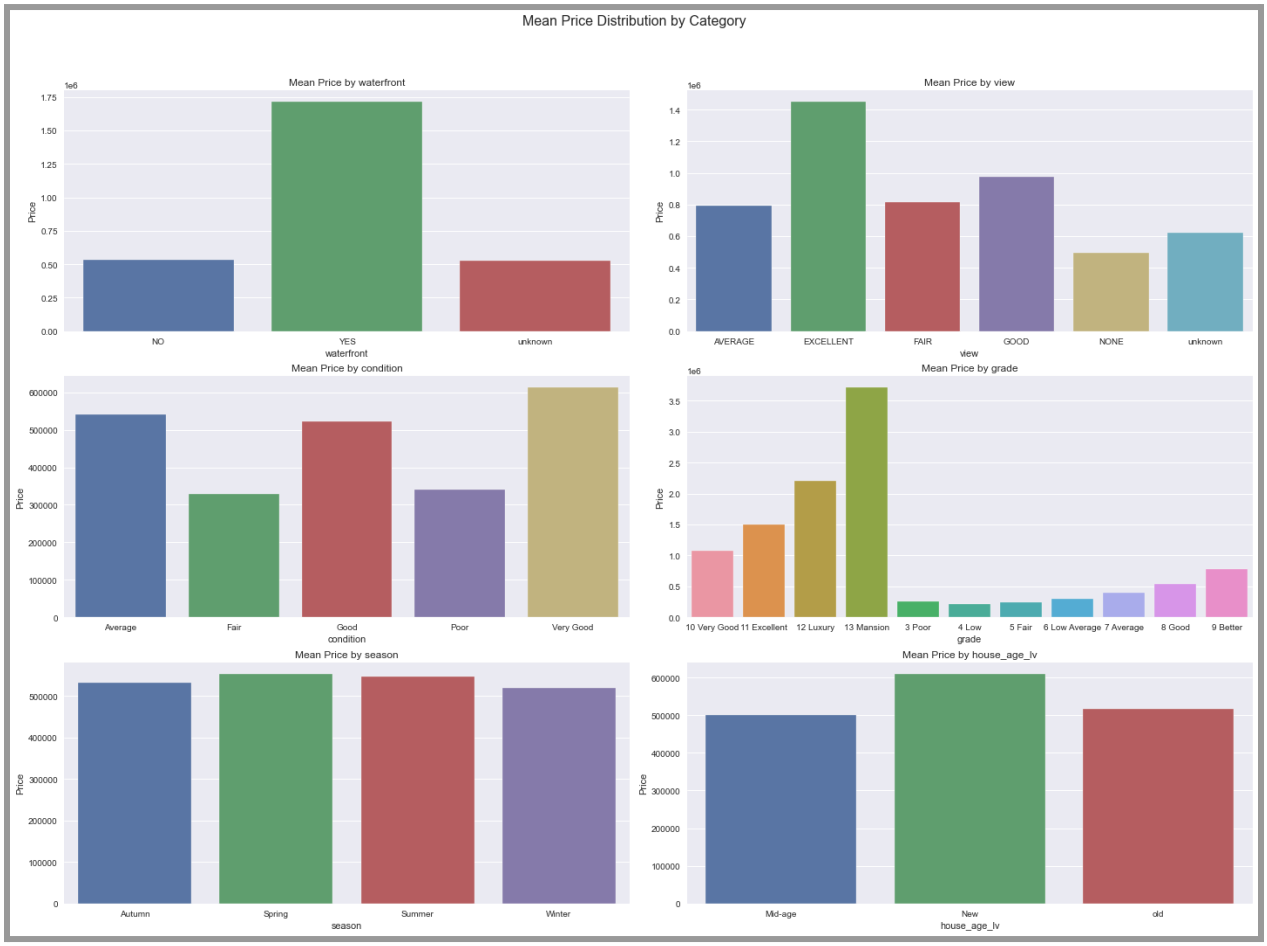
- *'floors'* with *'price'*:
  When examining the number of floors in relation to price, the scatter plot may reveal distinct clusters that resemble straight vertical lines. This indicates that the number of floors is a discrete variable, and there are specific price points associated with each discrete value of the floors.

Objective: **To investigate the influence of categorical attributes on house prices.**

Count plots enable us to quickly discern the prevalence of different values in the dataset. This information aids in understanding the distribution of categorical data in the following columns– '*waterfront*', '*view*', '*condition*', 'grade', '*season*' and '*house_age_lv*' to identify class imbalances, selecting relevant features for modeling, making comparisons across categories, and extracting valuable insights from the data.

- Many houses in the dataset have waterfront locations, indicating a strong appeal for waterside properties.
- A significant portion of houses in the dataset lacks impressive views, suggesting that it may not be a prevalent feature.
- Houses in average condition are numerous, indicating a balance between well-maintained and less well-maintained properties.
- Houses with an average grade of 7 are prevalent, possibly indicating a standard level of construction and design.
- A substantial number of houses were sold in the spring, reflecting a season of increased real estate activity.
- The dataset predominantly contains older houses, highlighting the longevity of certain properties within King County.

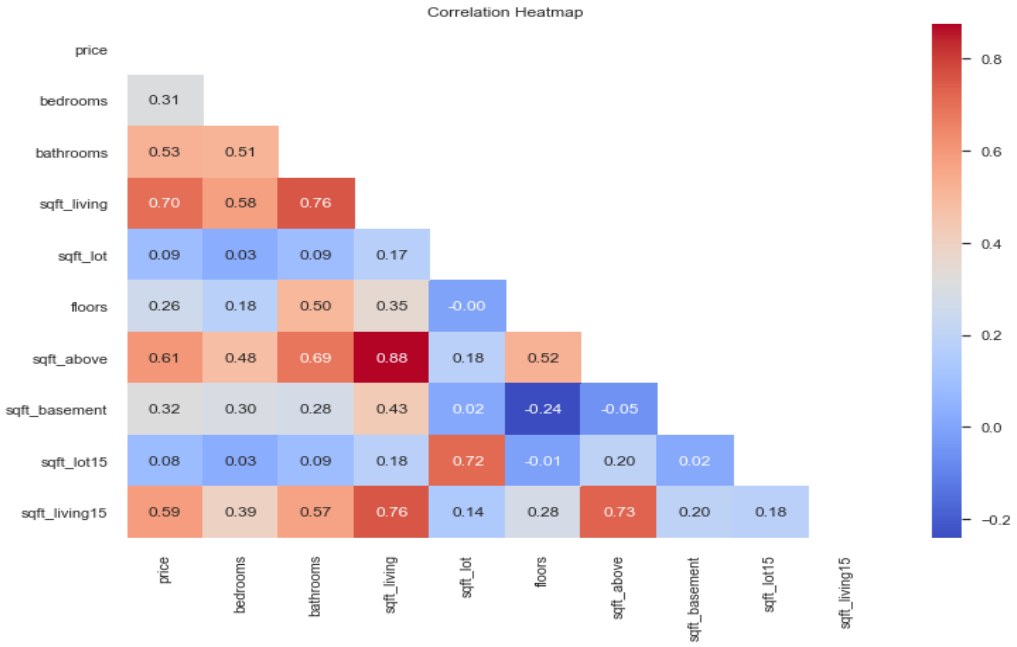
We plotted a bar graph to compare the different categorical columns to the mean prices and here's the observation:



- Houses located on waterfront properties fetched higher prices compared to those without waterfront views, underscoring the premium associated with waterside locations.
- Properties with an excellent view commanded higher sale prices compared to those without exceptional views, highlighting the value of scenic vistas in the real estate market.
- Houses in very good condition were sold at higher prices, suggesting that well-maintained homes are more desirable to buyers.

- Homes with a grade rating of 13, indicative of mansion-level construction and design, were associated with higher sale prices.
- Houses sold in the spring season generally commanded higher prices, possibly reflecting increased real estate activity during this time.
- Newer houses were sold at higher prices, indicating that buyers are willing to pay a premium for recently constructed properties.

A heatmap was utilized to determine which feature exhibits the highest correlation with the target variable, *'price*.' The analysis revealed that the *'sqft_living'* column possesses the most substantial correlation of *0.70* with price, indicating that it is a strong candidate to be the independent variable in the baseline model.



MODELLING

In our modeling phase, we are embarking on the exciting journey of creating a predictive model for house prices in King County, Washington. Our journey begins with a key observation from the correlation matrix: the *'sqft_living'* column exhibits the highest correlation with the target variable, *'price*.' As a result, we have chosen *'sqft_living'* as our predictor variable for the baseline model. This selection is grounded in the strong positive correlation that *'sqft_living'* shares with the price, signifying its potential as a powerful predictor of house prices. Our modeling endeavors will build upon this insight to construct a robust and effective predictive model that can provide accurate and valuable pricing recommendations for both homeowners and real estate stakeholders.

Baseline Model Summary

Based on the results of our baseline model, it's evident that the predictive power of the *'sqft_living'* column alone is limited. The baseline model suggests that only about 49% of the variance in house prices is accounted for by the square footage of living space (*'sqft_living'*). While this initial model is a good starting point, it falls short of providing accurate price predictions. Therefore, our strategy is to enhance the model's predictive capability by introducing additional features. This approach aims to capture a more comprehensive set of attributes that influence house prices in King County, thus improving the model's accuracy and usefulness.

Second Model Summary

In the second model, we have made significant improvements. This model is statistically significant, as indicated by the F-statistic p-value being less than 0.05, and it can explain approximately 50.96% of the variance in sales prices.

One limitation of the second model is that some of the p-values associated with the coefficients were found to be insignificant. Therefore, these coefficients might not be contributing significantly to the model's predictive power. To address this limitation and refine the model, further steps are required. These steps may include standardization and incorporate categorical columns, which can provide a more comprehensive understanding of the housing market and lead to even more accurate predictions.

Third Model Summary

The third model demonstrates high statistical significance with a very low F-statistic p-value of 0.0, indicating its overall robustness.This model has the ability to explain approximately 66.7% of the variance in house prices. This represents a substantial improvement in predictive power compared to earlier models.To ensure the robustness of our upcoming

model, we recognized the importance of addressing multicollinearity by dropping highly correlated features. The forthcoming model, with its refined set of predictors, is poised to offer improved insights into the determinants of house prices.

Fourth Model Summary

The fourth model represents a significant improvement in addressing multicollinearity, as we decided to drop the highly correlated columns, *'sqft_living15'* and *'sqft_above'*. This strategic move results in a more robust model, offering an enhanced understanding of the independent effects of our predictor variables. The model's R-squared value has maintained the same percentage of approximately 66.5% of the variance in house prices.

Objective: **To create a precise property valuation model that calculates the cost of homes depending on a range of characteristics.**

Final Model Summary

Here's a summary of the model results:

- Dependent Variable: Price
- R-squared: 0.665
- Adj. R-squared: 0.665
- F-statistic: 2254.0
- Prob (F-statistic): 0.00

Model Coefficients:

- In the regression summary provided, the constant term (const) has a coefficient of about $837,157, which represents the expected sale price of an average house when all other predictor variables are also average.
- '*bedrooms*': Each additional bedroom is associated with a decrease of approximately $25,780 in the sale price.
- '*bathrooms*': Each additional bathroom contributes around $37,160 to the sale price.
- '*sqft_living*: An extra square foot of living space increases the sale price by roughly $122,000.
- 'floors': Each additional floor adds approximately $25,490 to the sale price.
- '*sqft_basement*': More square footage of basement space increases the sale price by about $21,050.
- '*sqft_lot15*': An increase of 1 square foot in the lot size is associated with a decrease of roughly $12,970 in the sale price.

- *'waterfront_YES'*: Houses with a waterfront view are priced significantly higher, with an increase of about $722,000.
- *'house_age_lv_New'*: Newer houses are priced around $65,200 lower than middle-aged houses on average.
- *'house_age_lv_Old'*: Older houses are priced approximately $159,300 higher than middle-aged houses on average.
- *Grade Categories* (11 Excellent, 12 Luxury, 13 Mansion, 3 Poor, 4 Low, 5 Fair, 6 Low Average, 7 Average, 8 Good, 9 Better): Higher-grade categories lead to higher sale prices.

Model Evaluation:

- R-squared (R^2): An R-squared of 0.665 suggests that approximately 66.5% of the variance in house prices is explained by the model. In other words, the model captures 66.5% of the variability in house prices based on the included features.
- Adjusted R-squared (Adj. R^2): An adjusted R-squared of 0.665 in this model suggests that the explanatory variables still account for about 66.5% of the variance while adjusting for the model's complexity.
- F-statistic: In this case, the F-statistic is 2254.0, which is a high value. A high F-statistic implies that the model is statistically significant, indicating that at least one of the independent variables has a significant effect on the dependent variable.
- Prob (F-statistic): A very low p-value (in this case, 0.00) suggests that the overall model is statistically significant.

In summary, this regression model explains a substantial portion of the variance in house prices (R-squared) and is **statistically significant** (F-statistic). It indicates that the included features are valuable in predicting house prices.
Key predictors, including the number of bedrooms, bathrooms, living space, floors, basement space, and waterfront view, emerge as influential drivers of house prices. Grade and house age also exhibit substantial impacts on pricing. In particular, having a waterfront view significantly increases house prices, with an approximate increment of $722,000.

The model demonstrates its capability to make reasonably accurate predictions, with a Mean Absolute Error (MAE) of around $140,537 and a Root Mean Square Error (RMSE) of approximately $212,637. These metrics validate the model's effectiveness in estimating house prices and its potential utility for real estate professionals and stakeholders operating in the dynamic King County market.

CONCLUSION

In conclusion, several key factors significantly influence house prices in King County, Washington:

- Waterfront: Homes with a waterfront view have the most substantial positive impact on their prices. This feature is highly desirable and commands a premium.
-
- House Grade: The quality and grade of the house play a crucial role in determining its price. Properties with higher grades, such as "Mansion" and "Luxury," have significantly higher values. Quality construction and design are valued by buyers.
- Square Footage: More living space, including basements, has a positive effect on house prices. Larger homes tend to command higher values in the real estate market.
- Bathrooms and Floors: Additional bathrooms and floors in a house contribute positively to its price. These features offer convenience and comfort, which are reflected in the property's value.
- Lot Size: Surprisingly, larger lot sizes, especially Lot 15, have a negative impact on prices. This suggests that smaller, more manageable lots are preferred and can even result in higher property values.
- House Age: Older houses tend to be more expensive than newer ones. This might be due to historical or architectural significance associated with older properties. Buyers are willing to pay a premium for such houses.
- Bedrooms: An increase in the number of bedrooms is associated with lower house prices. This finding may reflect buyer preferences, as larger homes with more bedrooms might cater to a different market segment.

Understanding these determinants is essential for both buyers and sellers in King County's real estate market. Buyers can make informed decisions about property features, while sellers can set appropriate prices based on their property's characteristics. It's also valuable information for real estate professionals and investors seeking to maximize returns and navigate this competitive market effectively.

RECOMMENDATIONS

Based on the analysis of house price determinants in King County, we offer the following recommendations for buyers, sellers, and real estate professionals:

- *Prioritize Waterfront Properties:* If you're a buyer looking for an investment or a dream home, consider waterfront properties. These offer not only a beautiful living environment but also a significant potential for property value appreciation.
- *Enhance House Quality:* As a seller, focus on improving the quality of your property. Consider renovations or upgrades to increase the house grade, which will positively impact your selling price. Emphasize any unique design features or high-quality construction.
- *Highlight Square Footage:* When selling a property, make sure to highlight the square footage of the living space, including any basements. Potential buyers often place great importance on having enough space for their needs.
- *Consider Additional Bathrooms and Floors:* If you're planning to invest in a property, consider adding more bathrooms or additional floors to enhance its value. These features are attractive to many buyers and can result in a higher selling price.
- *Optimize Lot Sizes:* If you have a property with a larger lot size, consider the possibility of subdividing it into smaller, more manageable lots. Smaller lots, especially those similar in size to Lot 15, appear to be preferred in the market and can potentially lead to better prices.
- *Value Older Homes:* Older homes often come with historical and architectural charm. Sellers can emphasize these unique features to attract buyers who appreciate the character and history of older properties.
- *Optimize Bedroom Layouts:* Be mindful of the number of bedrooms in a property. Consider how they are laid out and whether the layout is appealing. Effective bedroom design can help maintain property appeal and value.

These recommendations take into account the key factors that influence house prices in King County. By considering these suggestions, buyers and sellers can make more informed decisions, and real estate professionals can better assist their clients in navigating this competitive real estate market

LIMITATIONS

The analysis does come with certain limitations and constraints:

- *Data Constraints:* The analysis relies on available data, potentially missing critical variables that could influence real estate pricing.

- *External Variables:* Economic shifts and government policies were excluded, which can have a substantial impact on the real estate market.
- *Simplified Model:* The model assumes linear relationships, neglecting potential nonlinear interactions that could exist in the data.

These limitations should be kept in mind when interpreting the results and making real estate decisions.

NEXT STEPS

- *Incorporate Economic Indicators:* To enhance the accuracy of market trend predictions, consider integrating economic indicators into the model. Variables such as local employment rates, income levels, and interest rates can offer valuable insights into housing market dynamics.
- *Advanced Predictive Models:* While the linear regression model has provided valuable insights, consider exploring advanced machine learning techniques such as gradient boosting and neural networks. These models can handle complex relationships and interactions within the data, potentially leading to more precise price forecasts.

By implementing these recommendations, stakeholders in the King County real estate market can further refine their models and decision-making processes, ultimately improving their ability to navigate this dynamic and competitive environment.