King County House Price Predictions Using Regression Models

Group Members

1. Doreen Wanjiru - Group Leader
2. Esther Francis
3. Gregory Mikuro
4. Celine Chege
5. Ian Korir

Scan to access the GitHub Repository

Introduction

- In the competitive real estate market, accurately pricing homes is essential.
- However, traditional methods, relying on personal opinions and limited housing options, often lead to errors and prolonged property searches.
- Additionally, catering to diverse client needs, such as first-time buyers and downsizing retirees, further complicates the process for agents.
- **Problem Statement**: The King County, real estate market, lacks accurate home pricing tools and data-driven insights into factors influencing property values, hindering informed decision-making for homeowners, sellers, and real estate agents.
- This project aims to address these challenges by leveraging data-driven approaches to refine pricing strategies and enhance client communication

Research Objectives

Main Objective

Empower real estate agents with data-backed pricing tools to optimize listing strategies, improve client communication, and maximize seller outcomes.

Other Objectives

1. Examine the features that have the most significant impact on home prices for effective marketing and negotiation strategies.
2. Develop models using the King County Housing dataset to predict home prices based on various features accurately.
   - A. Create a model for house price prediction with the metric accuracy being an R-squared of above 0.800 that can provide price predictions for potential listings based on key property characteristics.
   - B. Create a model for price range prediction with the metric accuracy being an R-squared of above 0.800 that can establish realistic price ranges for properties based on their features, enhancing agents' negotiation strategies.
3. Provide actionable insights to real estate agents to assist them in pricing homes accurately, understanding factors influencing property values, and advising homeowners on targeted renovations.

Data Overview & Methodology

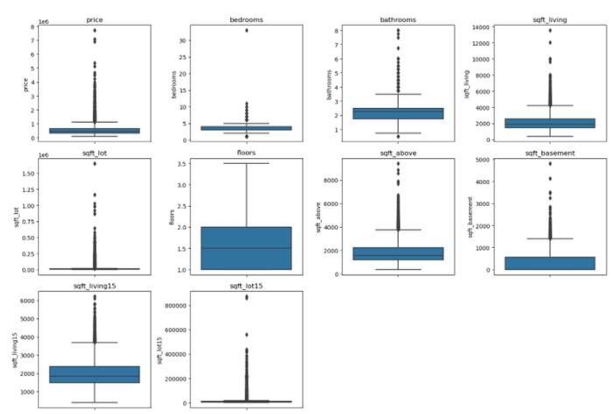
- Dataset - King County House Sales dataset (kc_house_data.csv).
- Methodology – Quantitative Statistical Analysis and Predictive Modeling

Features (Columns) Used and Their Relevance

- id: Unique identifier for each house sale record. May not be directly used for modeling, but essential for data cleaning and reference.
- date: Date of the house sale. Useful for time-based analysis, filtering by timeframe, or creating features related to seasonality.
- price: The target variable – the outcome we aim to predict.
- bedrooms: Number of bedrooms, essential for accommodating buyer needs.
- bathrooms: Number of bathrooms, impacting convenience and value.
- sqft_living: Square footage of interior living space, a major price driver.
- sqft_lot: Square footage of the land parcel, affecting lot size and potential use.
- floors: Number of floors in the house, a possible indicator of layout and space.
- waterfront: Binary variable indicating whether the property has waterfront access, a highly desirable feature in the region.
- view: Rated view quality of the property, a potential value-adding aspect.
- condition: Overall condition of the house, likely affecting price and renovation needs.
- grade: Overall grade assigned to the housing unit based on King County grading system. Understanding the details of this grading system is crucial.
- sqft_above: Square footage of the house excluding the basement.
- sqft_basement: Square footage of the basement, if present.
- yr_built: Year the house was originally built, indicating age.
- yr_renovated: Year of the last renovation, if applicable. Influences condition and potential for further updates.
- zipcode: Geographic location, potentially related to market dynamics and neighborhood desirability.
- lat: Latitude coordinate, useful for mapping or finer-grained location analysis.
- long: Longitude coordinate, used in conjunction with latitude.
- sqft_living15: Living space of homes in the neighborhood (15 nearest neighbors). Can provide insight into local market comparisons.
- sqft_lot15: Lot size of homes in the neighborhood (15 nearest neighbors).

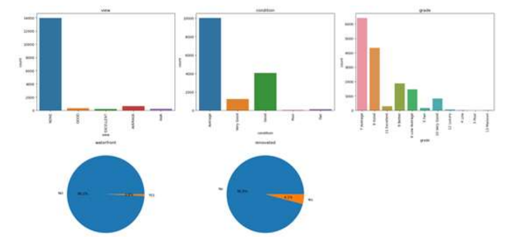
Data Preparation

Issue	Resolution
Correct Data Types	Converted 'date' column to datetime format and 'sqft_basement' to float using the pd.to_datetime() and pd.to_numeric() functions with appropriate format and error handling.
Missing Values	Removed rows with missing values in specified columns and replaced missing or zero values in 'yr_renovated' with 'yr_built' values using boolean masking and the fillna() method.
Add New Columns	Added new columns including 'year_sold', 'month_sold', 'house_age', 'renovation_age', and 'season' by extracting year and month from 'date' column, calculating house age, years since renovation, and season based on month sold.
Drop Unnecessary Columns	Dropped 'id', 'zipcode', and 'date' columns from the dataset using the drop() method with the appropriate axis parameter.
Check for Outliers	Created box plots for specified numerical columns to visualize the distribution of data and identify outliers. Identified outliers that represent legitimate data points and decided not to remove them to avoid introducing bias and maintain the model's generalizability.

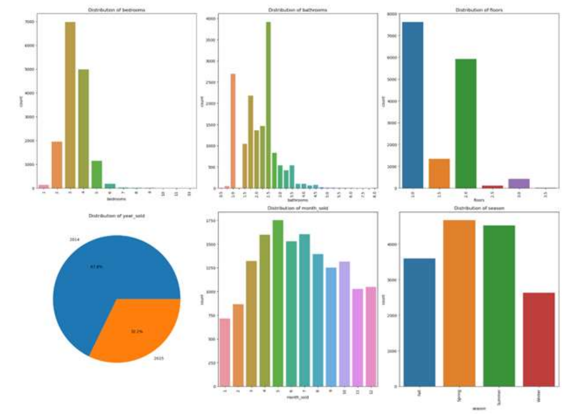


Exploratory Data Analysis – (Univariate Analysis)

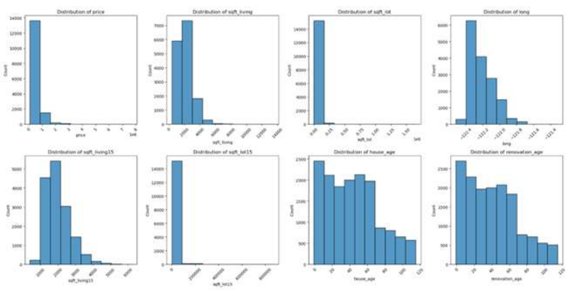
Categorical Features



Discrete Numerical Features

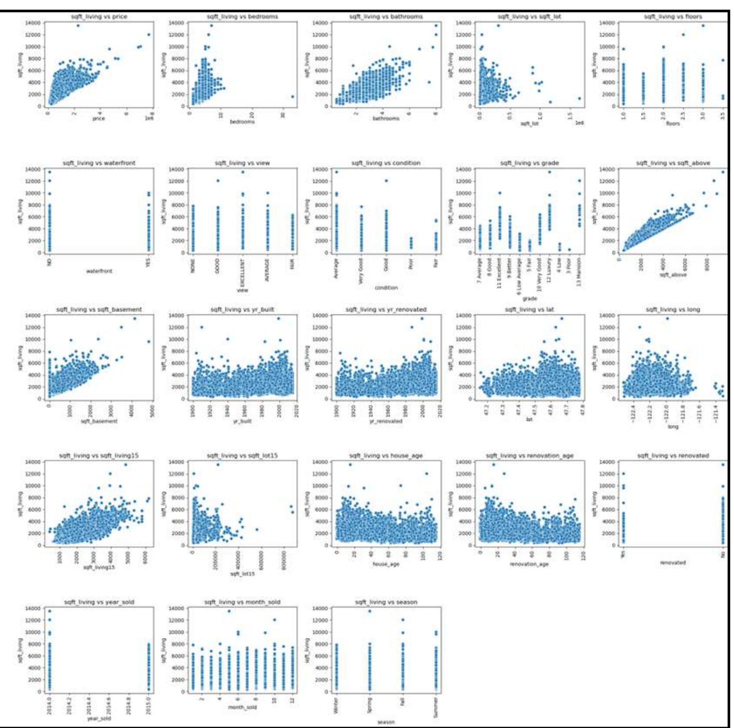


Continuous Numerical Features



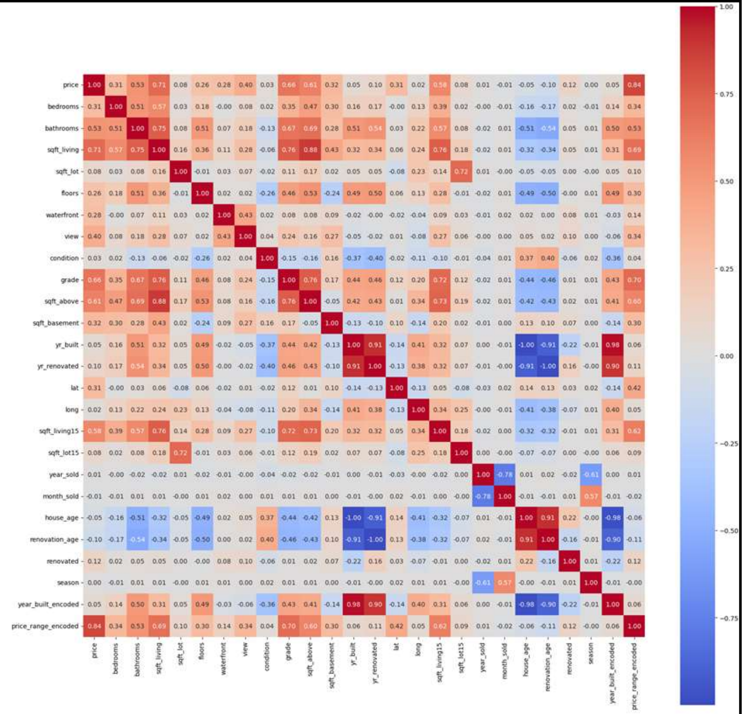
Exploratory Data Analysis – Price (IV) and Independent Variables (Bivariate Analysis)

- Positive correlations are observed between price and square footage, bedrooms, bathrooms, lot size, floors, and basement square footage, indicating that more extensive features tend to increase house prices.
- While some scatter plots reveal transparent linear relationships between variables, others show no discernible pattern, indicating variations in correlation strength across different pairs of variables.
- Outliers are present in the data, representing data points significantly deviating from the overall trend, and clusters of points suggest subgroups within the data, highlighting the need for further analysis or modeling to understand underlying patterns.



Exploratory Data Analysis – Correlations between all Variables (Bivariate Analysis)

- The correlation coefficients range from -1 to 1, where values near 1 indicate strong positive correlation, close to -1 indicate strong negative correlation, and around 0 suggest little to no correlation.
- Sqft_living and price have the highest positive correlation (0.71), implying that the living space correlates with the price of the property.
- The grade variable strongly correlates with both sqft_living and sqft_above, suggesting that higher-graded houses tend to have larger living spaces and more above-ground square footage.



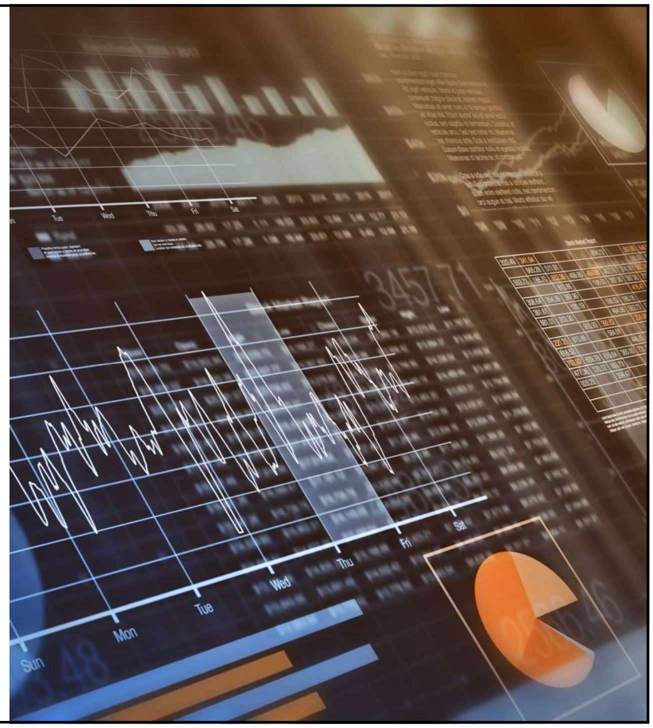
Exploratory Data Analysis – Price Vs Longitude and Latitude (Multivariate Analysis)

1. Spatial Clusters: Areas around 47.5, -122, 2 and 47.7, -122.1 exhibit clusters of higher-priced real estate (indicated by red dots). These locations likely correspond to desirable neighborhoods or central districts with elevated property values.

2. Geographical Variation: As latitude and longitude change, real estate prices fluctuate significantly. Understanding these spatial patterns can inform decisions related to property investment, urban planning, and market analysis.

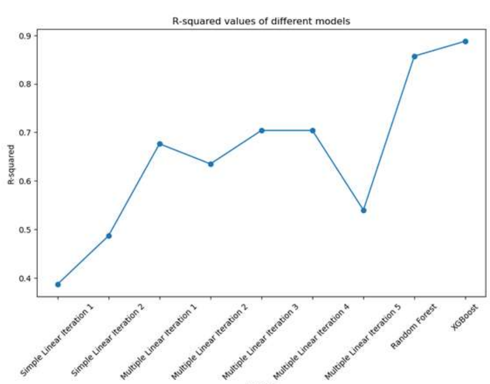
Modeling Overview

- The data preprocessing phase involved feature engineering and encoding categorical variables.
- Outliers were identified and removed based on scatterplot analysis during Exploratory Data Analysis.
- Correlation matrix and Variance Inflation Factors were utilized to detect multicollinearity among features, followed by data transformation and scaling to normalize distributions and ensure equal contribution of all features during model training.
- Simple linear regression, multiple linear regression, random forest, and XGBoost were iteratively employed, with adjustments made to enhance model accuracy through various iterations.

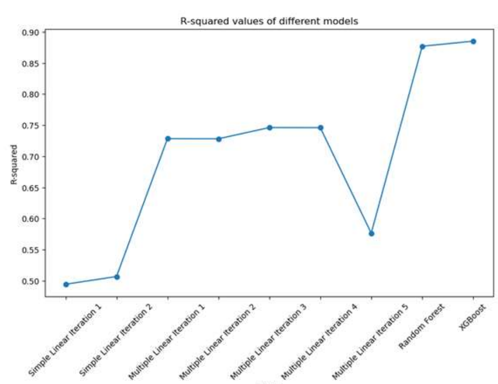


Findings – Performance of the Regression Models

Prices Prediction Models



Price Ranges Prediction Models



Deployment

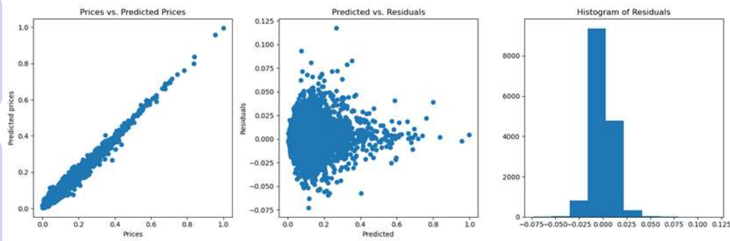
XGBoost Price Prediction Model: Achieved exceptional performance in explaining the variability in the target variable, indicating superior predictive accuracy.

Efficient Model Evaluation: Demonstrated robustness through low error metrics, indicating minimal overfitting and strong generalization capability.

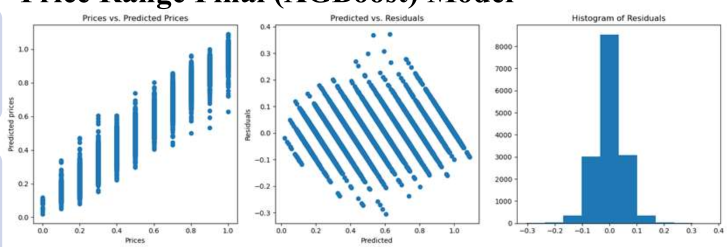
XGBoost Price Range Prediction Model: Similarly excelled in predicting categorical price ranges, signifying robust classification performance.

Visualizations: Observing scatterplots and histograms revealed a linear distribution of prices vs. predicted prices and random residuals, validating the models' accuracy and reliability.

Prices Final (XGBoost) Model



Price Range Final (XGBoost) Model



Conclusion

Accurate Pricing Guidance: Achieved high model accuracy (R-squared > 0.800), empowering agents with precise pricing predictions for effective listing strategies and client communication.

Key Feature Identification: Identified 'sqft_living', 'grade', and 'view' as pivotal factors influencing home prices, enabling tailored marketing and negotiation strategies to highlight property attributes.

Enhanced Marketing Strategies: Equipped agents to craft compelling listing descriptions and showcase amenities that resonate with buyers, leveraging insights to attract premium prices.

Strategic Renovation Recommendations: Provided data-driven guidance on cost-effective renovations that maximize property value, empowering agents to advise clients on strategic improvements.

Empowering Data-Driven Decisions: Overall, equipped real estate professionals with actionable insights into housing market dynamics, enabling data-driven decisions that optimize pricing strategies and enhance client satisfaction.

Recommendations & Next Steps

Recommendations

Action	Responsibility	Evaluation
Foster Collaboration	Data science & real estate	Schedule bi-weekly meetings from May 1st, 2024. Assess impact on model refinement.
Continuous Model Refinement	Data science & analytics	Implement monthly refinement sprints starting May 1st, 2024. Evaluate accuracy improvements.
Invest in Advanced Analytics Tools	IT & executive leadership	Allocate resources by July 1st, 2024. Measure adoption and efficiency gains.

Next Steps

Dynamic Data Pipeline

Develop automated pipeline for real-time data retrieval & preprocessing.

Ensure models stay current for accurate insights.

Interactive Consumer Dashboard

Design user-friendly dashboard for easy access to pricing predictions & market insights.

Enhance user engagement & decision-making with intuitive visualizations.

Time-Sensitive Data Research

Conduct research for up-to-date housing data, including market trends & economic indicators.

Anticipate future changes in the real estate landscape & adjust models accordingly for continued relevance & reliability.