Phase 2 Group 5 Project

Participants

1. Laura Kanda
2. Cynthia Jerono
3. George Chira
4. Ruth Kioko
5. Angela Maina

COUNTY REAL ESTATE CONSULTING COMPANY

A Comprehensive Analysis Using Multiple Linear Regression Models

TABLE OF CONTENTS

´Introduction

´Objectives

´Data Analysis and Cleaning

´Exploratory Data Analysis (EDA)

´Identified Key Features with Highest Correlations to House Prices

´Model Development

´Model Evaluation

´Recommendations

INTRODUCTION

- Accurately predicting house prices is crucial for home-buyers and home-sellers to make informed decisions in the real estate market.
- This project aims to equip homeowners with insights of the housing market in King County, Washington, by analyzing various features.
- The features used to analyze house prices and develop a regression model were; Living space, quality grade, and the number of bathrooms.

KEY OBJECTIVES

01

Develop Accurate Predictive Models for House Prices: Create and evaluate multiple linear regression models.

02

Identify Key Factors Influencing House Prices: Analyze various features to determine their impact.

03

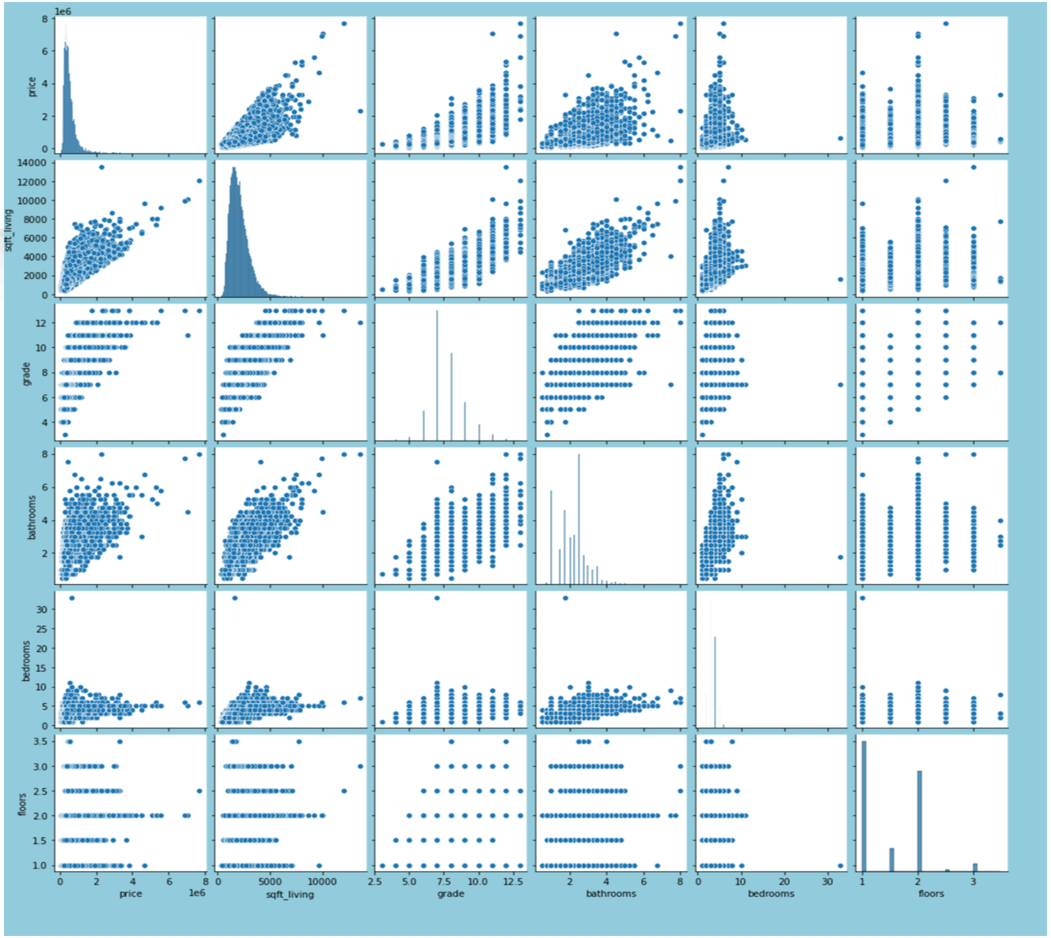
Provide Actionable Recommendations for Property Value Enhancement: Based on model results and feature analysis

04

Guide Homeowners and Real Estate Professionals: Optimize property quality and features to increase market value

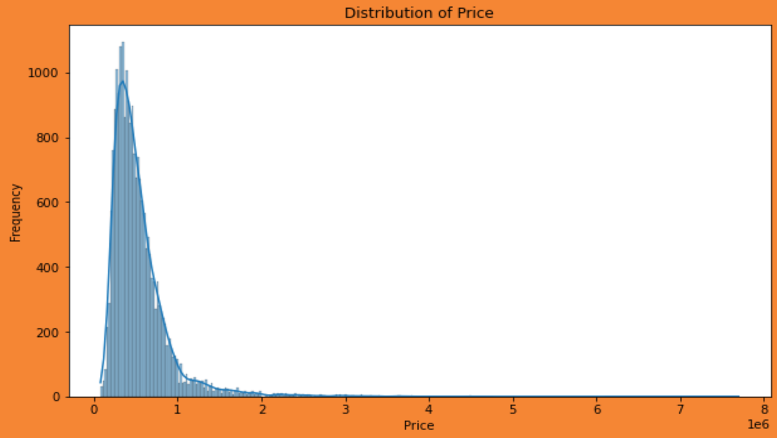
DATA UNDERSTANDING

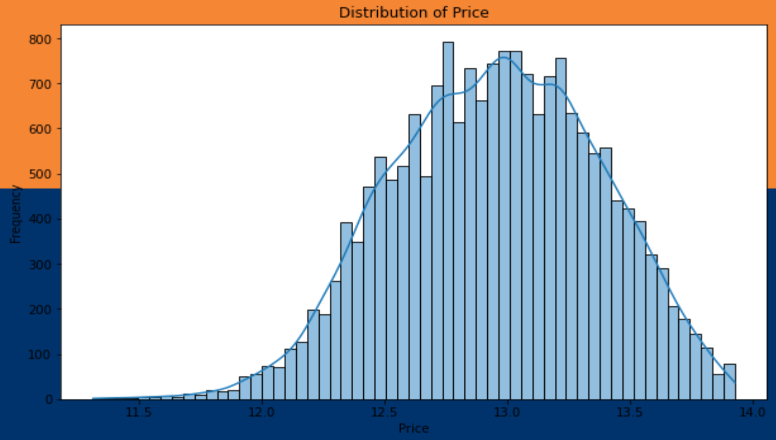
- **The dataset used consisted of properties sold between 2014 and 2015**
- **The columns used to make scatter plot models were :**
    - a. **Price -** is prediction target
    - b. **Bedrooms Number -** number of Bedrooms per House
    - c. **Bathrooms Number -** number of bathrooms per house
    - d. **sqft_livingsquare -** footage of the home
    - e. **Grade -** overall grade given to the housing unit, based on King County grading system
- From the scatter plots, linear relationships were determined, patterns identified and outliers detected



DATA CLEANING

- Loading and Cleaning: Imputing missing values, removing outliers and removing duplicate values
- Example: Dealing with missing values in the Waterfront column- imputation method was used by finding the mode of houses with waterfronts
- Below is a price distribution before and after cleaning and normalization



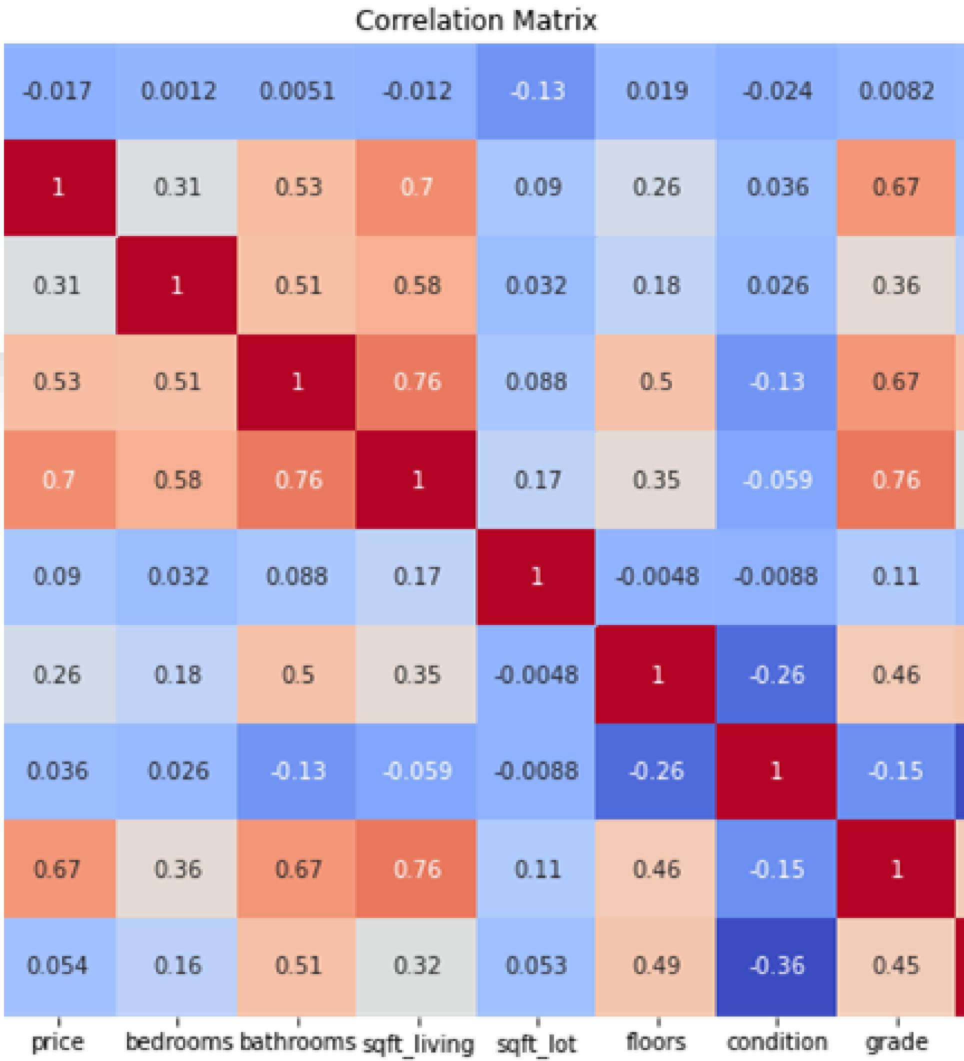


EDA- CHECKING FOR CORRELATION

- ´The heatmap on the right shows correlation of the selected features with price

- sqft_living , grade and bathrooms have the highest effect on price for they have a correlation > 0.5

- Condition, among other features have the low correlation thus not being used in our final model



Correlation Matrix

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	grade
-0.017	0.0012	0.0051	-0.012	-0.13	0.019	-0.024	0.0082
1	0.31	0.53	0.7	0.09	0.26	0.036	0.67
0.31	1	0.51	0.58	0.032	0.18	0.026	0.36
0.53	0.51	1	0.76	0.088	0.5	-0.13	0.67
0.7	0.58	0.76	1	0.17	0.35	-0.059	0.76
0.09	0.032	0.088	0.17	1	-0.0048	-0.0088	0.11
0.26	0.18	0.5	0.35	-0.0048	1	-0.26	0.46
0.036	0.026	-0.13	-0.059	-0.0088	-0.26	1	-0.15
0.67	0.36	0.67	0.76	0.11	0.46	-0.15	1
0.054	0.16	0.51	0.32	0.053	0.49	-0.36	0.45

FEATURE SELECTION

Criteria: Correlation coefficient above 0.5

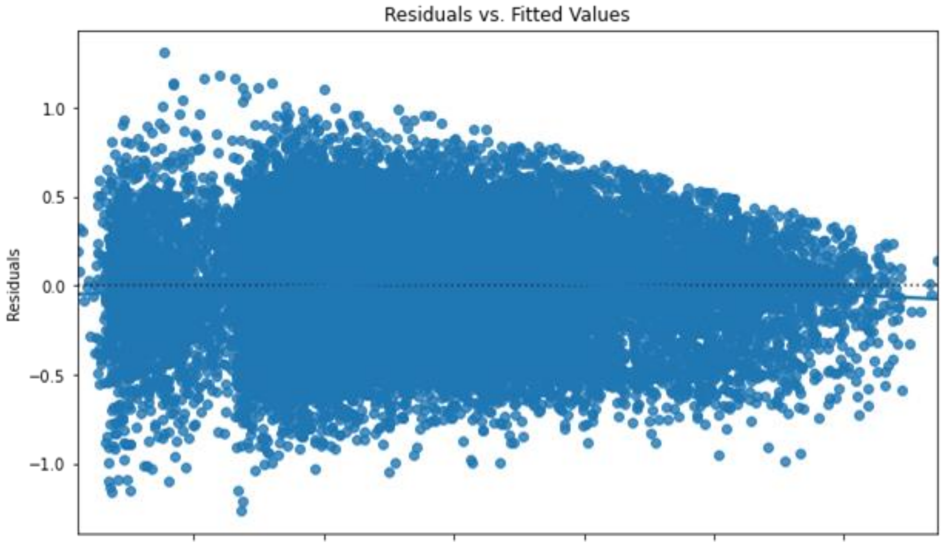
Selected Features: Living space, grade, number of bathrooms

Process: Based on correlation coefficients

Feature	Correlation Coefficient against Price
price	1
grade	0.667967
bathrooms	0.525912
bathrooms	0.525912
bedrooms	0.308795
floors	0.256811
sqft_lot	0.089879
sqft_lot	0.089879
yr_built	0.053952

MODEL DEVELOPMENT

- before model development was done, heteroscedasticity was tested and confirmed .
- We had **4 linear regression models** with an approach of **increasing complexity**
- This was done by Incrementally adding features
- Our goal was to Identify the most significant predictors.



MODEL EVALUATION

The Mean Absolute Error (MAE) and R^2 were the metrics used to assess the models performance.

4 models were created :

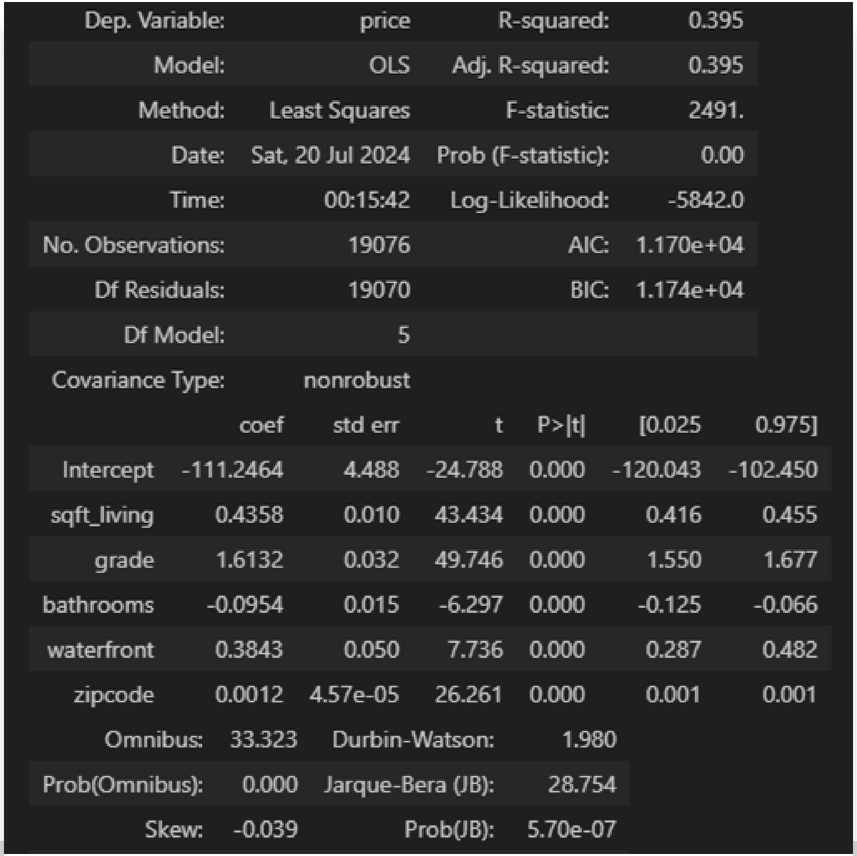
-Model 1(sqft_living & price): MAE = 0.291, R^2 = 0.305

-Model 2(sqft_living, grade, price): MAE = 0.373, R^2 = 0.372

-Model 3 (Sqft_living, grade, bathrooms, price): MAE= 0.273, R^2 = 0.374

-Model 4 (all featured variables): MAE= 0.26 , R^2 = 0.395

The best performing model with the lowest MAE and highest R^2 was Model 4. It was further used to make predictions



```
     Dep. Variable:            price          R-squared:      0.395
            Model:              OLS      Adj. R-squared:      0.395
           Method:    Least Squares          F-statistic:      2491.
             Date: Sat, 20 Jul 2024  Prob (F-statistic):       0.00
             Time:         00:15:42       Log-Likelihood:    -5842.0
 No. Observations:            19076                  AIC:  1.170e+04
     Df Residuals:            19070                  BIC:  1.174e+04
         Df Model:                5
  Covariance Type:        nonrobust
                 coef    std err        t   P>|t|    [0.025    0.975]
  Intercept -111.2464      4.488  -24.788   0.000  -120.043  -102.450
sqft_living    0.4358      0.010   43.434   0.000     0.416     0.455
      grade    1.6132      0.032   49.746   0.000     1.550     1.677
  bathrooms   -0.0954      0.015   -6.297   0.000    -0.125    -0.066
 waterfront    0.3843      0.050    7.736   0.000     0.287     0.482
    zipcode    0.0012   4.57e-05   26.261   0.000     0.001     0.001
       Omnibus: 33.323    Durbin-Watson:     1.980
Prob(Omnibus):   0.000  Jarque-Bera (JB):    28.754
          Skew: -0.039          Prob(JB):  5.70e-07
```

**RECOMMENDATIONS

1.Enhance Property Quality: Invest in improving the overall quality (grade) of properties. High-quality materials and design standards lead to substantial returns

2. Optimize Living Space: Increase living space (sqft_living) thoughtfully, ensuring additional space enhances functionality and appeal without unnecessary expansions.

3. Balanced Feature Development: Aim for a balanced approach in adding features like floors, bathrooms, and bedrooms. Focus on usability, aesthetics, and overall appeal to avoid potential negative impacts on house prices.

Implementing these recommendations helps stakeholders understand the factors influencing house prices and make informed decisions to enhance property value effectively.

CONCLUSION

Best Predictive Model: Model 4, using all features, is the most accurate and robust for predicting house prices. It balances prediction accuracy and explanatory power effectively.

Key Influencing Features: Grade is the most influential features positively affecting house prices. Enhancing property quality and optimizing living space can significantly increase property values.