PREDICTING HOUSE PRICES IN KING COUNTY

A MULTILINEAR REGRESSION APPROACH

AGENDA

Topic one : INTRODUCTION

Topic two :PROBLEM STATEMENT

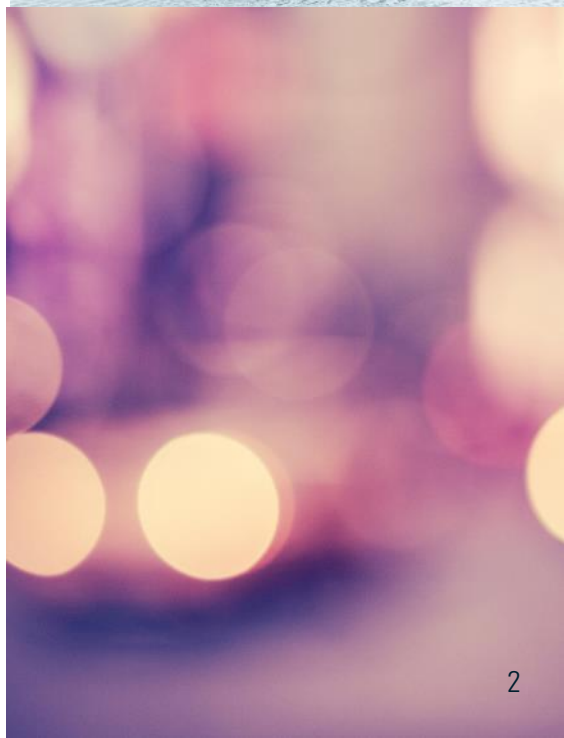
Topic three : OBJECTIVES

Topic four : ANALYSIS

Topic five : MODELLING

Topic six : CONCLUSION

Topic seven : RECOMMENDATION



INTRODUCTION

- Precisely forecasting house prices is a significant challenge globally, particularly in densely populated cities such as Seattle, where market fluctuations are constant. Given the dynamic nature of real estate prices, our interest lies in identifying the most influential features for predicting house prices accurately.

- We are looking into a dataset encompassing home selling prices in King County, Washington. Our approach involves constructing a multiple regression model, with the house price variable as the target, and carefully chosen features serving as explanatory variables.

PROBLEM STATEMENT

The Real Estate Agency needs to offer effective guidance for prospective home sellers seeking to enhance their property's value before listing it on the market. Specifically, there is a demand for personalized recommendations concerning improvements to specific features within homes

OBJECTIVES

To identify house features that are highly related to price.

To find the relationship between the house area/sizes and the price

Obtain how the price relates to the condition and grade.

To find the relationship between price and house features i.e., bedrooms, bathrooms .

To identify the relationship between price and location.

DATA CLEANING & PREPARATION

Handling missing values

Handledmissing values by filling in missing data points or removing rows/columns with missing values

Handling outliers

Handled outliers by analyzing and decide whether to remove or transform outliers based on the context of the analysis.

177 entries were flagged as duplicates. Further investigation revealed that some houses were sold multiple times at different prices and times.

Duplicate Entries

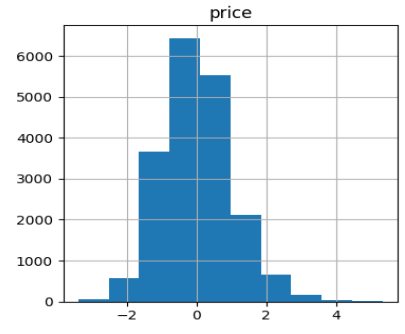
Dealt with categorical variables by transforming them to numerical variables which the model can make sense of.

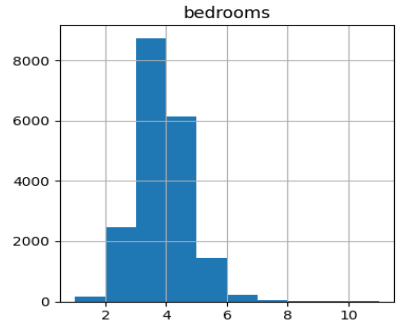
One Hot Encoding

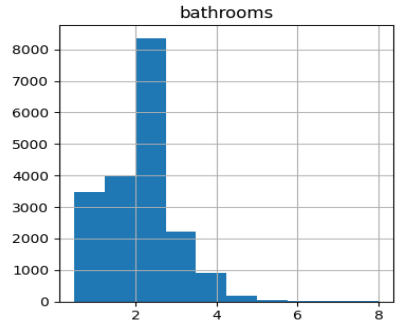
DATA ANALYSIS

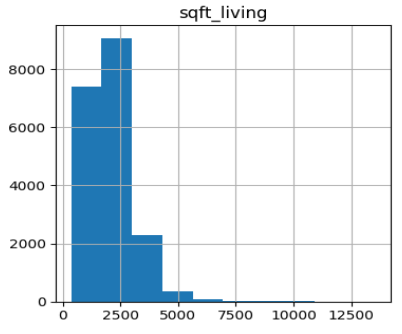
Visualizing the distributions of each data column.

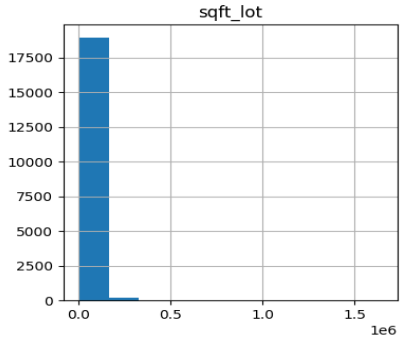
It can be seen that majority of the features are not normally distributed.

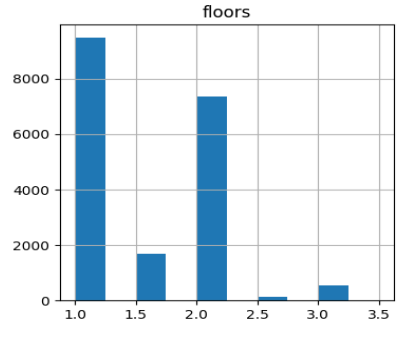


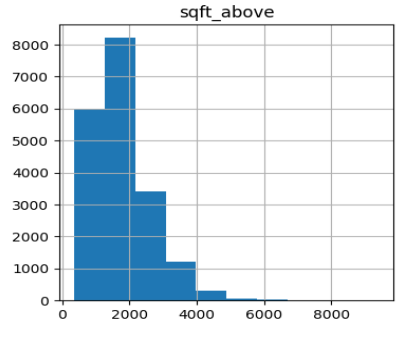


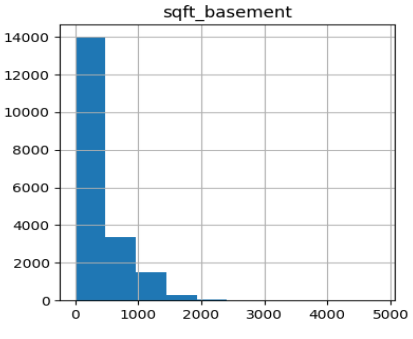


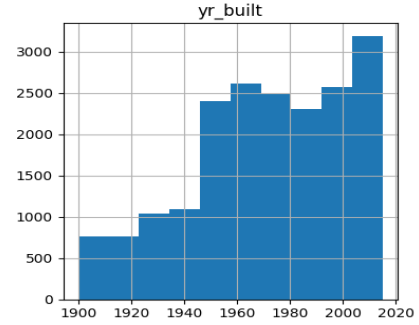


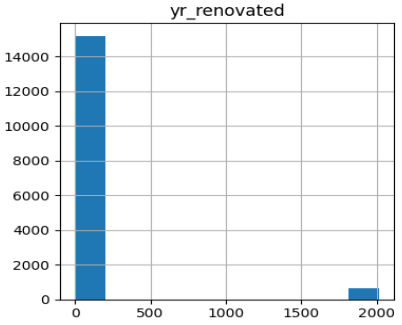


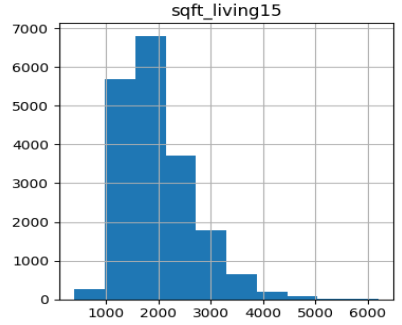


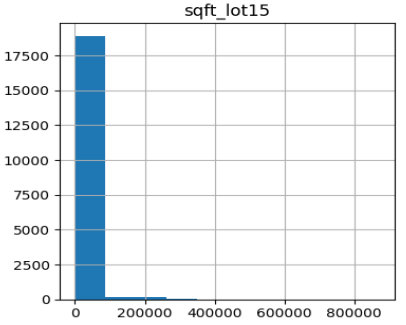


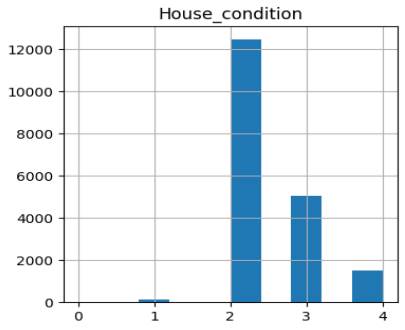








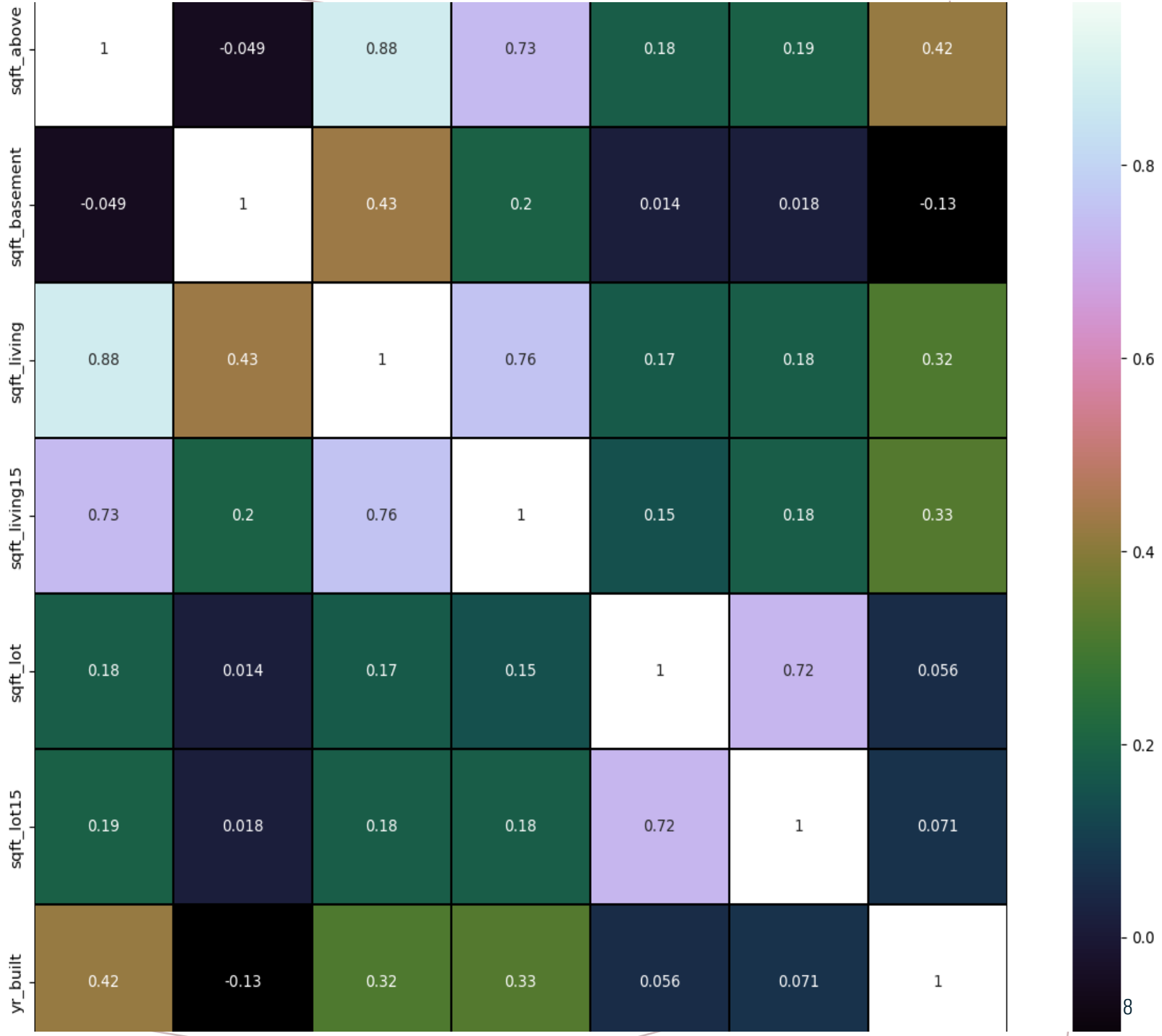




Data Analysis

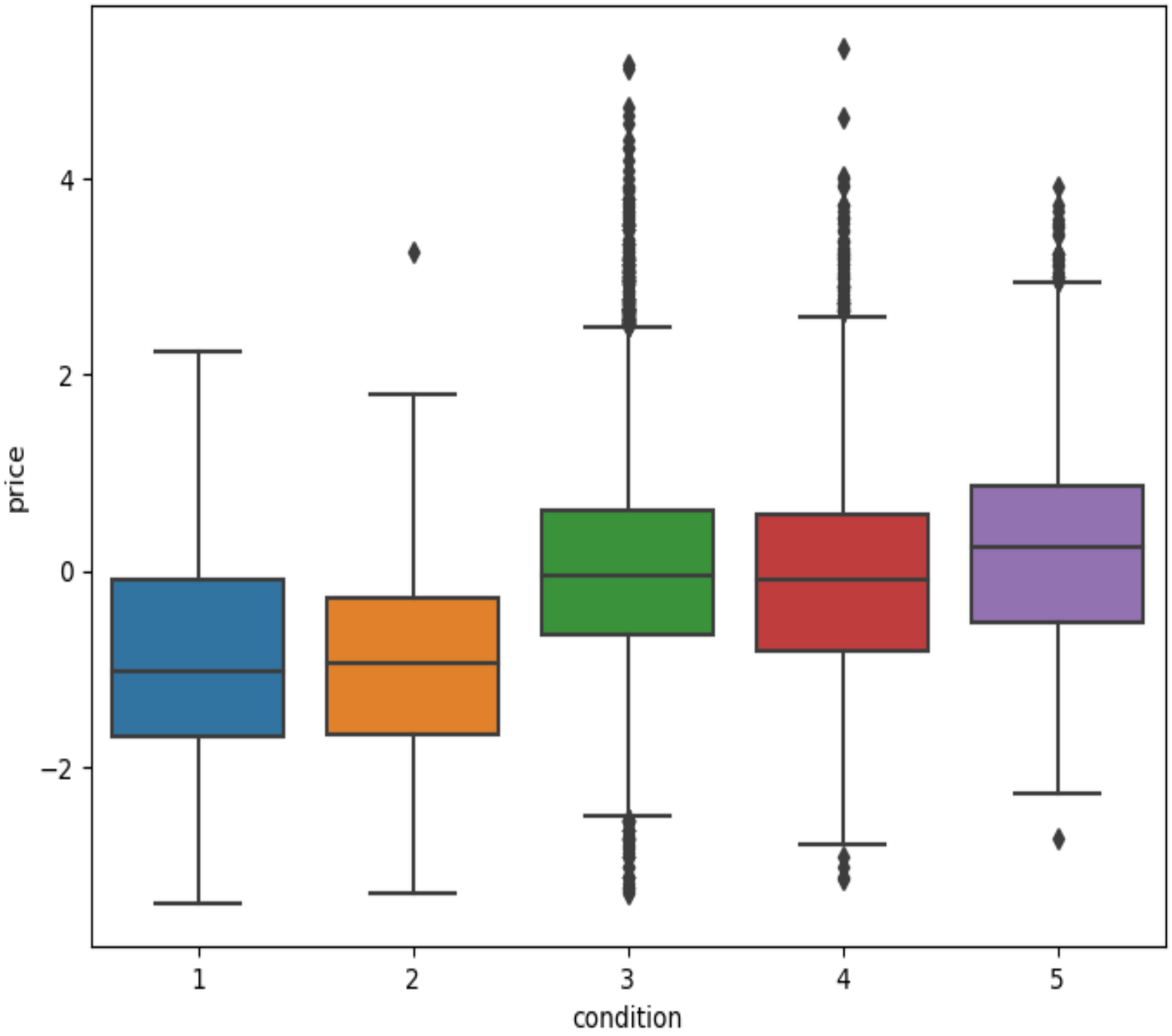
The visualisation of correlations between each variable to understand the relationship between the features and select features with a higher correlation to use in our model.

- Correlation is a measure of linear relationship between two variables.
- When the value in the matrix is high this shows a strong relationship.
- The correlation matrix was developed before modelling.
- The square footage above and the square footage of living were the most strongly correlated features



DATA ANALYSIS

RELATIONSHIP BETWEEN HOUSE CONDITION AND PRICE - THERE IS A SLIGHT INCREASE IN PRICE FOR A CONDITION RATING OF 3 OR HIGHER. - THERE ARE MORE OUTLIER PRICINGS FOR THE HIGHER RATING



MODELLING AND REGRESSION: MODEL 1

OLS Regression Results

Dep. Variable:	price	R-squared:	0.486
Model:	OLS	Adj. R-squared:	0.486
Method:	Least Squares	F-statistic:	1.814e+04
Date:	Thu, 26 Oct 2023	Prob (F-statistic):	0.00
Time:	22:22:11	Log-Likelihood:	-20914.
No. Observations:	19220	AIC:	4.183e+04
Df Residuals:	19218	BIC:	4.185e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.5730	0.013	-122.957	0.000	-1.598	-1.548

MODELLING AND REGRESSION: MODEL 1

Here are the observations based on the results:

R-squared Value: The R-squared value is 0.486. This indicates that approximately 48.6% of the variation in house prices is explained by the square meter of living space. In other words, the model with this single independent variable is moderately effective in explaining the variation in house prices.

F-statistic: The F-statistic is 1.814e+04. This statistic is used to test the overall significance of the regression model. A high F-statistic value suggests that the model is statistically significant, which is the case here. It means that at least one of the independent variables (in this case, 'sqmt_living') is related to the dependent variable ('price').

Coefficients: The coefficient for the constant (intercept) is approximately -1.5730. This is the estimated house price when the square meter of living space is zero. In this context, it doesn't have a practical interpretation. The coefficient for 'sqmt_living' is approximately 0.0081. It indicates that, on average, for every additional square meter of living space, the house price increases by 0.0081 units.

P-values: The p-values associated with both the intercept and 'sqmt_living' are very close to zero (0.000). Low p-values indicate that both variables are statistically significant in predicting house prices. The model indicates that square meter of living space ('sqmt_living') is a significant predictor of house prices, and for each additional square meter of living space, the house price increases by 0.0081 units. However, it's important to consider the specific context of the data and the assumptions of the regression model when interpreting these results. Additionally, this model uses only one independent variable to predict house prices, and there may be other important factors that contribute to price variation.

MODELLING AND REGRESSION: MODEL 2

OLS Regression Results

Dep. Variable:	price	R-squared:	0.648
Model:	OLS	Adj. R-squared:	0.648
Method:	Least Squares	F-statistic:	2948.
Date:	Thu, 26 Oct 2023	Prob (F-statistic):	0.00
Time:	22:22:11	Log-Likelihood:	-17266.
No. Observations:	19220	AIC:	3.456e+04
Df Residuals:	19207	BIC:	3.466e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-69.1575	9.062	-7.632	0.000	-86.919	-51.396
sqft_living	0.0003	5.56e-05	5.815	0.000	0.000	0.000
sqft_above	-0.0001	5.55e-05	-2.177	0.029	-0.000	-1.21e-05
zipcode	0.0009	9.1e-05	9.511	0.000	0.001	0.001
bathrooms	0.1562	0.010	15.371	0.000	0.136	0.176
condition	0.0878	0.007	12.101	0.000	0.074	0.102

MODELLING AND REGRESSION: MODEL 2

Here are some observations based on the results:

R-squared (R^2): The R-squared value is 0.648, indicating that the model explains approximately 64.8% of the variance in the target variable (price). This suggests that the selected features have a reasonably good explanatory power. Adjusted R-squared: The adjusted R-squared is 0.648, which adjusts the R-squared value based on the number of features. It is slightly lower than R-squared, as expected when more features are included.

F-statistic: The F-statistic tests the overall significance of the model. With a very high F-statistic of 2948 and a p-value close to zero, it suggests that the model is statistically significant.

Coefficients (coef): The coefficients represent the estimated impact of each feature on the target variable. For example, x1, x2, x3, etc., are the selected features, and their corresponding coefficients indicate the estimated change in price associated with a one-unit change in the feature. Negative coefficients suggest a negative relationship, while positive coefficients indicate a positive relationship.

P-values (P>|t|): The p-values associated with each coefficient test the null hypothesis that the coefficient is equal to zero. A low p-value (typically below 0.05) suggests that the feature is statistically significant in predicting the target variable

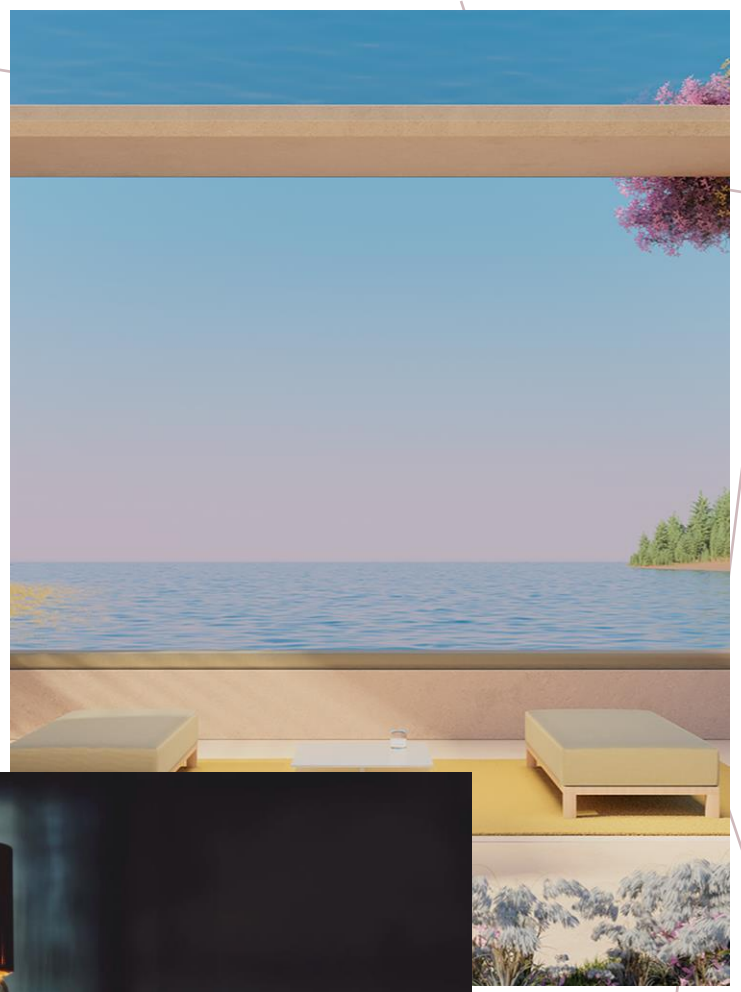
CONCLUSION

- Collectively, square footage, grade, and the number of bathrooms emerge as strong indicators of a house's price within King County. Individuals aspiring to maximize their property's value, dedicating efforts to enlarging the square footage and enhancing the overall construction quality is very important. When extending the square footage, it's advisable to contemplate the addition of extra bathrooms, as this analysis suggests a positive correlation between the number of bathrooms and the price.
- Nonetheless, the model does possess certain constraints. It necessitated log-transforming some variables to meet regression assumptions, implying that any new data fed into the model would require similar preprocessing. Furthermore, given the regional disparities in housing prices, the model's generalizability to data from other counties may be constrained. Eliminating outliers from the dataset might result in the model struggling to accurately predict properties with exceptionally high or low values.
- In prospective analyses, it would be beneficial to investigate the most influential predictors of home prices outside of King County, as well as for properties with extraordinary price levels

RECOMMENDATIONS

Focus on number of bathrooms- Our analysis strongly suggests that investing in kitchen and bathroom renovations tends to have a significant positive impact on the estimated value of homes. Upgrading fixtures, modernizing appliances, and enhancing the overall aesthetics of these spaces can notably relevant. -

Consider Functional Space Utilization: Maximizing the use of available space, especially in the living areas, can significantly enhance the appeal of a home. Our analysis indicates that optimizing the utilization of existing square footage, without necessarily increasing it, can positively influence the estimated value of the property. Creating multipurpose spaces and ensuring a seamless flow throughout the house are key consideration



THE TEAM

ALLEN MAINA

CELIAJOY OMIAH

CYNTHIA KAREKO

MARION JELIMO

WILLIAM OMBALLA

THANK YOU