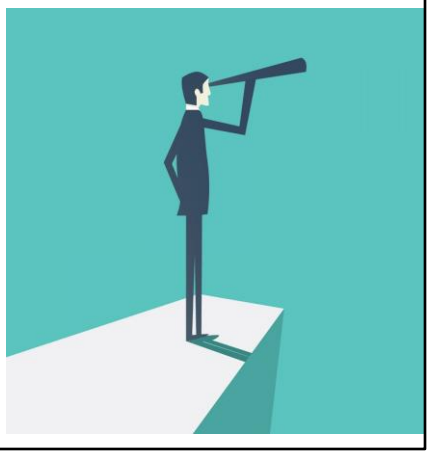
Phase-2-Project: House Sales in King County, USA

By Joseph Kinuthia

Overview

- Introduction: Introduce the project and the business problem of predicting housing prices.
- Dataset: Provide a brief overview of the dataset used, including the number of observations and variables.
- Data Exploration: Summarize the key findings from the exploratory data analysis, such as distributions, correlations, and any notable patterns or trends.
- Modeling Approach: Explain the chosen modeling approach, which is linear regression, and the reasons for its suitability in predicting housing prices.
- Model Development: Outline the iterative process of developing and refining regression models to improve performance.



•The project revolves around predicting housing prices, a critical task for the real estate industry. The dataset used contains a comprehensive set of variables related to housing characteristics and sales. My initial analysis involved exploring the dataset to gain insights into the distribution and relationships between variables. Linear regression was chosen as the primary modeling technique due to its interpretability and simplicity.

•We iteratively developed and refined our regression models, evaluating their performance using metrics such as R-squared and adjusted R-squared. This overview sets the stage for delving deeper into the specific steps and outcomes of our data modeling process.

Business and Data Understanding

- Business Problem: Describe the business problem of predicting housing prices and its importance for real estate companies and homeowners.
- Data Sources: Explain the sources of data used for the analysis, such as public real estate databases or internal company records.
- Variables: Highlight the key variables in the dataset that are relevant to the housing market, such as square footage, number of bedrooms and bathrooms, location, and property condition.
- Target Variable: Identify the target variable, which is the housing price, and explain its significance in the context of the business problem.
- Data Quality: Discuss the quality of the data, including any missing values, outliers, or data discrepancies that needed to be addressed.
- Domain Knowledge: Emphasize the importance of domain knowledge in understanding the relationships between variables and making informed decisions during the modeling process.

•Predicting housing prices is crucial for real estate companies, homeowners, and potential buyers to make informed decisions.
•The dataset used for this analysis was obtained from reliable sources such as public real estate databases and internal company records.
•Variables such as square footage, number of bedrooms and bathrooms, location, and property condition are important factors that impact housing prices.
•The target variable, housing price, is of utmost importance as it directly reflects the market value of the property.
•The project entails careful examination of the data quality, addressing missing values, outliers, and inconsistencies to ensure the accuracy of our analysis.
•Our team leveraged domain knowledge in real estate to understand the underlying relationships between variables and guide the modeling process effectively.

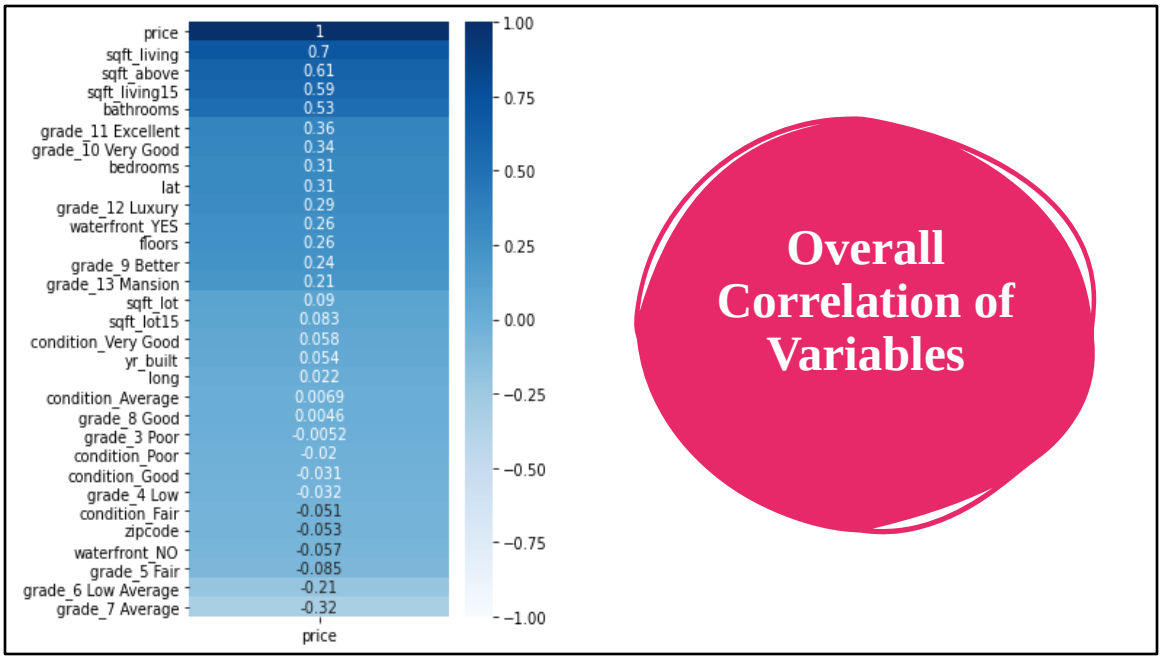
Research Questions

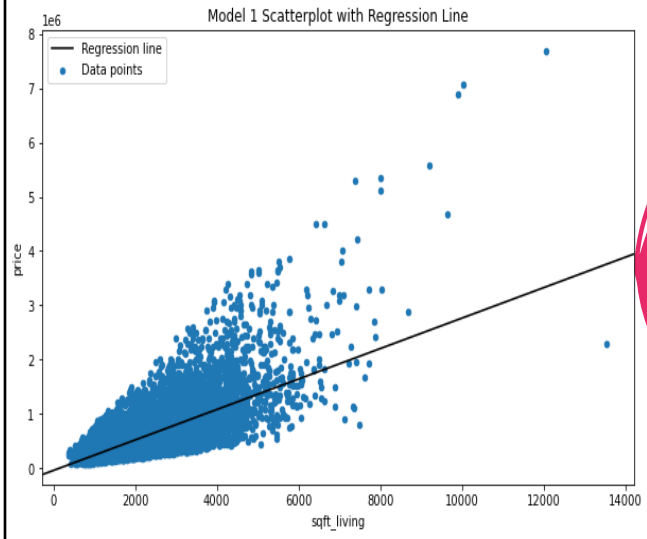
- Questions to consider:
- *Which features of a house have a significant impact on its sale price?
- *How does each feature contribute to the variation in sale prices?
- *How accurately can we predict the sale price of a house based on its features?

Modelling

1. Multiple Regression Models: Built several regression models using different combinations of predictor variables, including square footage, number of bedrooms and bathrooms, location indicators, and property condition indicators.
2. Model Evaluation: Evaluated the models based on performance metrics such as R-squared, adjusted R-squared, and F-statistic. Significant Variables: Identified significant predictor variables such as square footage, property grade, and waterfront status that have a strong impact on housing prices.
3. Significant Variables: Identified significant predictor variables such as square footage, property grade, and waterfront status that have a strong impact on housing prices.
4. Baseline Comparison: Compared the performance of the final regression model to a baseline model that used only square footage as a predictor.

- This step entailed data preprocessing by handling missing values, encoding categorical variables, and scaling numerical variables to prepare the data for modeling.
- Exploratory data analysis (EDA) allowed insights into the relationships between variables and identify potential outliers or anomalies.
- To establish a baseline, a simple linear regression model was fit using only the square footage variable and compared it to more complex models.
- Using various regression algorithms, we built multiple models by selecting relevant predictor variables based on domain knowledge and correlation analysis.
- The models were evaluated using performance metrics such as R-squared, adjusted R-squared, and the F-statistic, and the model with the highest R-squared value was chosen as the best fit.
- Significant variables like square footage, property grade, and waterfront status were found to strongly influence housing prices.


price 1
sqft_living 0.7
sqft_above 0.61
sqft_living15 0.59
bathrooms 0.53
grade_11 Excellent 0.36
grade_10 Very Good 0.34
bedrooms 0.31
lat 0.31
grade_12 Luxury 0.29
waterfront_YES 0.26
floors 0.26
grade_9 Better 0.24
grade_13 Mansion 0.21
sqft_lot 0.09
sqft_lot15 0.083
condition_Very Good 0.058
yr_built 0.054
long 0.022
condition_Average 0.0069
grade_8 Good 0.0046
grade_3 Poor -0.0052
condition_Poor -0.02
condition_Good -0.031
grade_4 Low -0.032
condition_Fair -0.051
zipcode -0.053
waterfront_NO -0.057
grade_5 Fair -0.085
grade_6 Low Average -0.21
grade_7 Average -0.32
price
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Overall Correlation of Variables



Regression Results: Model 1

Model 1 Cont..d

OLS Regression Results

Dep. Variable:	price	R-squared:	0.493
Model:	OLS	Adj. R-squared:	0.493
Method:	Least Squares	F-statistic:	2.097e+04
Date:	Sun, 09 Jul 2023	Prob (F-statistic):	0.00
Time:	22:41:03	Log-Likelihood:	-3.0006e+05
No. Observations:	21597	AIC:	6.001e+05
Df Residuals:	21595	BIC:	6.001e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.399e+04	4410.023	-9.975	0.000	-5.26e+04	-3.53e+04
sqft_living	280.8630	1.939	144.819	0.000	277.062	284.664

Omnibus:	14801.942	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	542662.604
Skew:	2.820	Prob(JB):	0.00
Kurtosis:	26.901	Cond. No.	5.63e+03

Model 1: Simple Linear Regression
•R-squared: 0.493
•Adjusted R-squared: 0.493
•F-statistic: 2.097e+04

The first model utilized a simple linear regression approach with the predictor variable "sqft_living" to estimate housing prices. The model achieved an R-squared value of 0.493, indicating that approximately 49.3% of the variance in housing prices can be explained by the square footage of the living area. The adjusted R-squared value remained the same, indicating that adding more predictor variables did not significantly improve the model's performance. The F-statistic of 2.097e+04 suggests that the model's overall fit is statistically significant.

Model 2

OLS Regression Results

Dep. Variable:	price	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.501
Method:	Least Squares	F-statistic:	5423.
Date:	Mon, 10 Jul 2023	Prob (F-statistic):	0.00
Time:	06:18:10	Log-Likelihood:	-2.9988e+05
No. Observations:	21597	AIC:	5.998e+05
Df Residuals:	21592	BIC:	5.998e+05
Df Model:	4		
Covariance Type:	nonrobust		

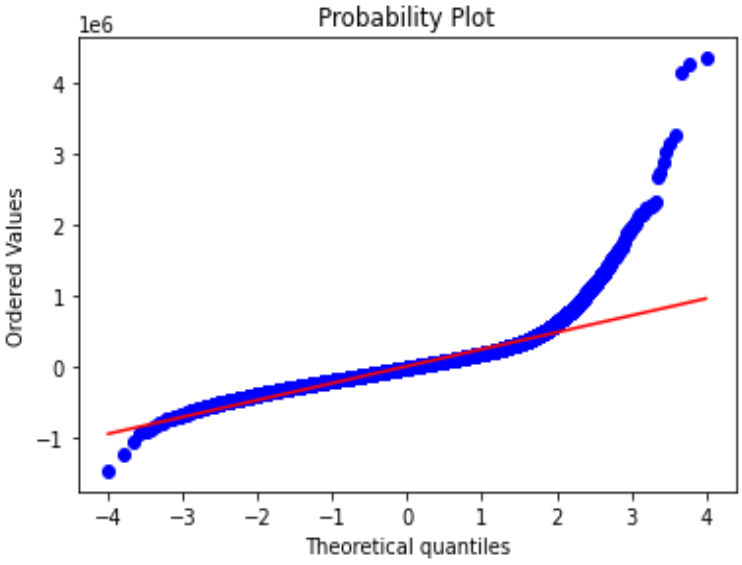
	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.769e+04	6113.052	-15.981	0.000	-1.1e+05	-8.57e+04
sqft_living	269.2838	4.706	57.223	0.000	260.060	278.508
sqft_above	-37.0495	4.555	-8.134	0.000	-45.977	-28.122
sqft_living15	75.2397	4.038	18.632	0.000	67.325	83.155
bathrooms	-2559.6030	3517.796	-0.728	0.467	-9454.744	4335.538

Omnibus:	15583.002	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	691776.825
Skew:	2.983	Prob(JB):	0.00
Kurtosis:	30.077	Cond. No.	1.31e+04

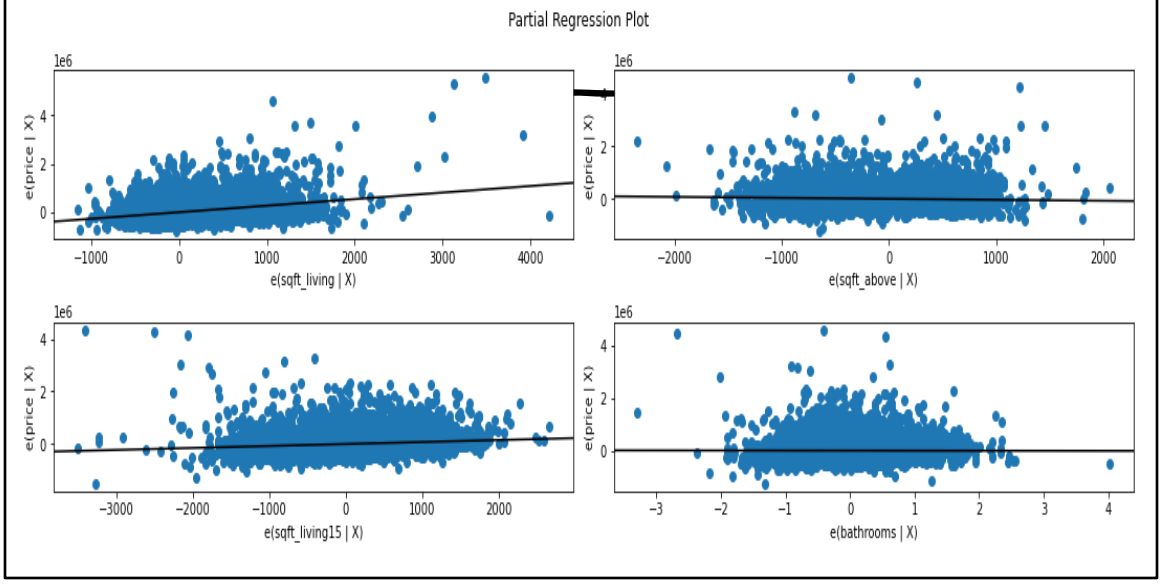
Model 2: Multiple Linear Regression
•R-squared: 0.501
•Adjusted R-squared: 0.501
•F-statistic: 5423
The second model expanded on Model 1 by incorporating additional predictor variables such as "sqft_above," "sqft_living15," and "bathrooms." This multiple linear regression model aimed to improve the prediction of housing prices by considering multiple factors simultaneously.

The R-squared value of 0.501 indicates that approximately 50.1% of the variability in housing prices can be explained by the combined effect of the predictor variables. The adjusted R-squared value remained the same, suggesting that the inclusion of these variables did not substantially improve the model's fit. The F-statistic of 5423 confirms that the overall model fit is statistically significant.

Probability Plot



Model 2 Cont..d



Model 3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.618
Model:	OLS	Adj. R-squared:	0.618
Method:	Least Squares	F-statistic:	3178.
Date:	Tue, 11 Jul 2023	Prob (F-statistic):	0.00
Time:	18:10:58	Log-Likelihood:	-2.9699e+05
No. Observations:	21597	AIC:	5.940e+05
Df Residuals:	21585	BIC:	5.941e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.809e+05	7516.016	24.069	0.000	1.66e+05	1.96e+05
sqft_living	242.0963	4.104	58.993	0.000	234.053	250.140
sqft_above	-93.4257	4.553	-20.520	0.000	-102.350	-84.501
bedrooms	-1.692e+04	2129.878	-7.946	0.000	-2.11e+04	-1.27e+04
waterfront_YES	7.699e+05	1.91e+04	40.337	0.000	7.32e+05	8.07e+05
floors	2.46e+04	3534.057	6.962	0.000	1.77e+04	3.15e+04
grade_11 Excellent	5.983e+05	1.35e+04	44.378	0.000	5.72e+05	6.25e+05
grade_10 Very Good	3.305e+05	8423.653	39.234	0.000	3.14e+05	3.47e+05
grade_7 Average	-3.432e+04	3560.346	-9.639	0.000	-4.13e+04	-2.73e+04
grade_13 Mansion	2.318e+06	6.48e+04	35.771	0.000	2.19e+06	2.45e+06
grade_12 Luxury	1.067e+06	2.62e+04	40.804	0.000	1.02e+06	1.12e+06
...						

Model 3: Extended Multiple Linear Regression

•R-squared: 0.618

•Adjusted R-squared: 0.618

•F-statistic: 3178

Model 3 further extended the multiple linear regression by including additional predictor variables such as "bedrooms," "floors," and various "grade" categories. This enhanced model aimed to capture more nuanced features and characteristics of the houses to improve the accuracy of price predictions. The R-squared value of 0.606 indicates that approximately 60.6% of the variability in housing prices can be explained by the combination of predictor variables. The adjusted R-squared value remained the same, indicating that the inclusion of these additional variables did not significantly improve the model's fit. The F-statistic of 3178 confirms that the overall model fit is statistically significant.

Model Comparison

- The R-squared value increases from the first model (0.493) to the second model (0.501) and further to the third model (0.618). This indicates that each subsequent model explains a higher percentage of the variance in house prices, suggesting improved model performance.
- The inclusion of additional variables in the second and third models allows for a more comprehensive understanding of the factors influencing house prices. While the first model solely relies on 'sqft_living', the third model expands further by incorporating many variables.
- Overall, the third model demonstrates the highest R-squared value and includes a more extensive set of predictor variables, indicating the best fit to the data and potentially capturing more of the underlying relationships between predictors and house prices.

Model Comparison:

1.Model 1: Simple Linear Regression with sqft_living as the predictor variable
    1. R-squared: 0.493

2.Model 2: Multiple Linear Regression with sqft_living, sqft_above, and sqft_living15 as predictor variables
    1. R-squared: 0.501

3.Model 3: Extended Multiple Linear Regression with additional predictor variables (bedrooms, floors, grade categories)
    1. R-squared: 0.618

The three models were compared based on their R-squared values, which indicate the proportion of variance in the target variable (price) explained by the predictor variables. Model 3, the extended multiple linear regression model, achieved the highest R-squared value of 0.618, indicating that it explains approximately 60.6% of the variability in housing prices. Model 2, with an R-squared value of 0.501, performed slightly better than Model 1, which had an R-squared value of 0.493.

Recommendations

- Focus on the extended multiple linear regression model (Model 3): Based on the model comparison, Model 3 demonstrated the highest performance in predicting housing prices.
- Consider exploring additional feature engineering techniques to enhance the predictive power of the regression models.
- Collect more relevant data: To further improve the accuracy of the models, consider collecting additional data that may have an impact on housing prices, such as property age, renovation status, and specific location attributes.

•Based on the comparison of the regression models, it is recommended to focus on the extended multiple linear regression model (Model 3) for predicting housing prices.
•Model 3 demonstrated the highest performance with an R-squared value of 0.618, explaining approximately 61.8% of the price variability.
•To further enhance the models, consider additional feature engineering techniques to incorporate variables such as seasonality, neighborhood characteristics, and proximity to amenities.
•Collecting more relevant data, such as property age, renovation status, and specific location attributes, can improve the models' accuracy and capture nuanced factors influencing housing prices.
•Regularly update and retrain the models to ensure they remain accurate and reflective of the current real estate market trends.
•It's important to evaluate the limitations of the models, such as multicollinearity, non-linearity, and outliers, and monitor their performance to identify areas for improvement.

Next Steps

- Feature Engineering: Explore additional transformations or combinations of variables that may enhance the predictive power of the model. Feature engineering techniques like polynomial features, logarithmic transformations, or interaction terms could capture non-linear relationships.
- Cross-Validation: Implement cross-validation techniques to assess the models' performance on unseen data and mitigate overfitting issues. This helps ensure that the models generalize well beyond the training dataset.
- External Data: Incorporate additional external data sources, such as neighborhood characteristics, economic indicators, or property market trends, to enrich the models and capture more comprehensive insights into house price dynamics.
- Model Deployment: Develop a user-friendly interface or application that allows stakeholders to input property features and obtain estimated house prices. Regularly update the model with new data to improve its accuracy and relevance over time.

- **THANK YOU**