Project Description.

ANALYSIS OF HOUSE SALES IN A KING'S COUNTY

Business Understanding.

Primetime Realtors situated in the heart of North Western County acts as the conduit for transforming home ownership aspirations into tangible realities. Committed to unwavering excellence and employing data-driven methodologies, the agency aspires to lead the way in achieving optimal pricing and facilitating successful real estate endeavors. Its overarching objective is to surpass traditional limitations by leveraging technology and analytical insights to revolutionize the real estate landscape as we perceive it.

Business Problem

In our role at a real estate agency, we are examining data from the Kings County House Sales dataset to advise our agency on strategies to boost home values in Kings County through renovations. Our goal is to identify the most impactful renovation factors that can enhance a home's value. By pinpointing these factors, our agency can effectively guide homeowners in maximizing their profits when selling their homes

Objectives

1. To identify key factors that significantly influence house prices in the Northwestern county

2. To develop an optimal pricing strategy using a robust multiple linear regression model.

3. To help improve the agency's annual revenue by leveraging the analytical insights and pricing strategy developed through this project.

The Data Understanding

This project uses the King County House Sales dataset, which can be found in kc_house_data.csv in the data folder in this assignment's GitHub repository. The description of the column names can be found in column_names.md in the same folder. As with most real-world data sets, the column names are not perfectly described, so you'll have to do some research or use your best judgment if you have questions about what the data means.

Data Cleaning Next, we chose to clean our data by dropping unnecessary columns that would not help us achieve the result we were aiming for. These columns included:

Id, date, waterfront, view, grade, sqft_above, sqft_basement, yr_renovated, zipcode lat, long sqft_living15, sqft_lot15. After dropping these columns, the remaining columns that we were experimenting with were:

Price, bedrooms, bathrooms, sqft_living, sqft_lot floors, condition, Waterfronts.

After examining the data types of each column, we identified the only non-numeric column as "condition." To handle this, we transformed the "condition" column using one-hot encoding, which split it into subcategories: cond_avg, cond_fair, cond_good, cond_poor, and cond_verygood, all of which were converted to float data types.

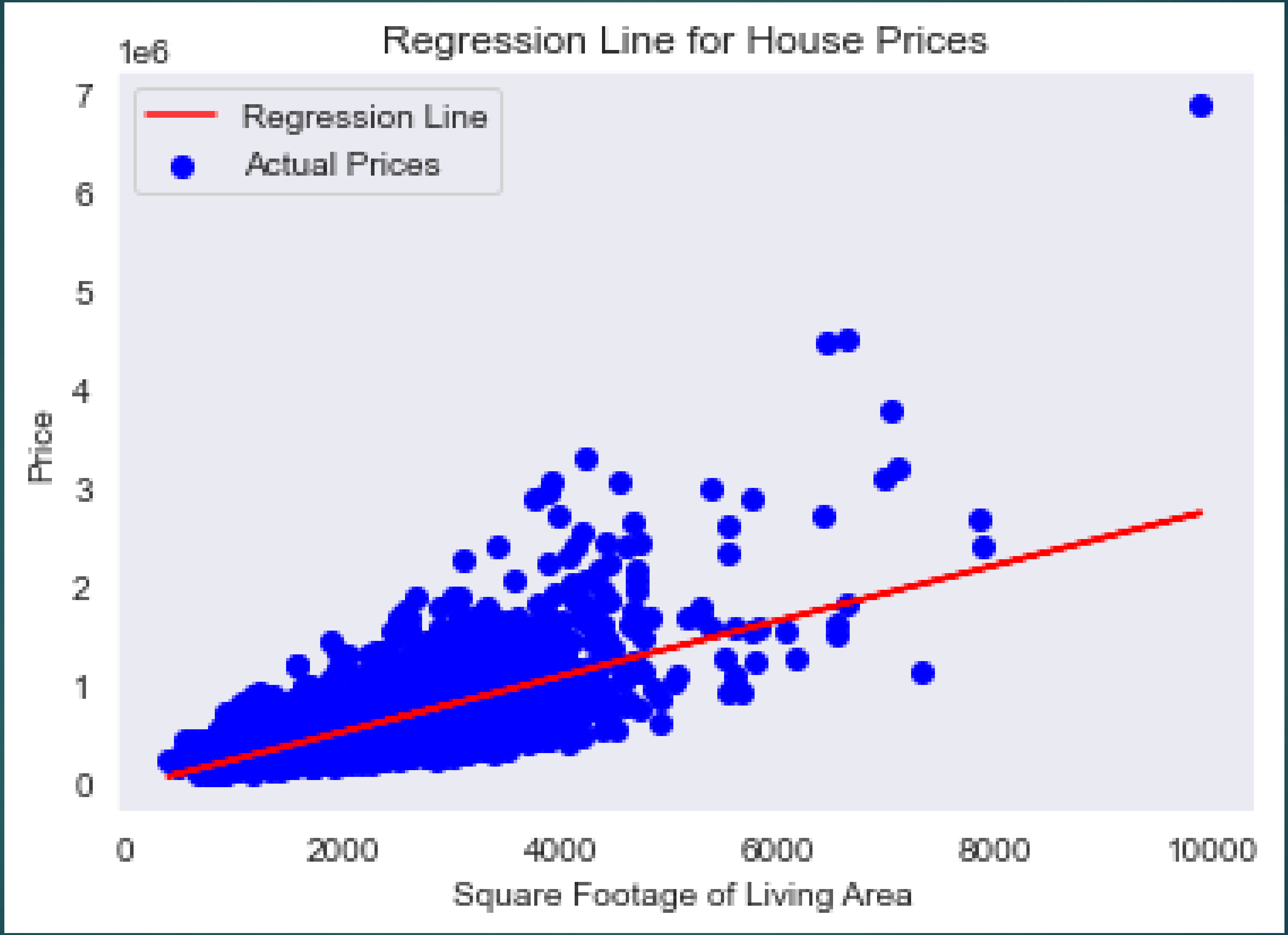
Next, we categorized the remaining columns into separate arrays based on whether they were continuous or categorical variables. we then proceeded to analyze each array individually: for categorical variables, we generated histograms, while for continuous variables, we created a scatter matrix to explore their relationships further.

Data Preparation

To ensure the integrity of our target variable, "price," we employed the train-test split method to normalize it, separating it from the original dataframe and assigning it to "y," while designating the remaining variables as "X".

Following the normalization of the target variable, we constructed a heat map to visualize the correlations between all columns and the target variable, "price." This heat map revealed that "sqft_living" exhibited the highest correlation. To validate this correlation further, we utilized cross-validation, which demonstrated a minimal discrepancy of about 0.01 between the training and validation scores, indicating a robust model.

Lastly, in the preparation phase, we crafted two models: one showcasing the variables ranked by their correlation strength with the target variable and another displaying a scatter matrix of all variables. This approach aimed to highlight any non-normal distributions, thereby guiding the normalization process for the modeling stage.

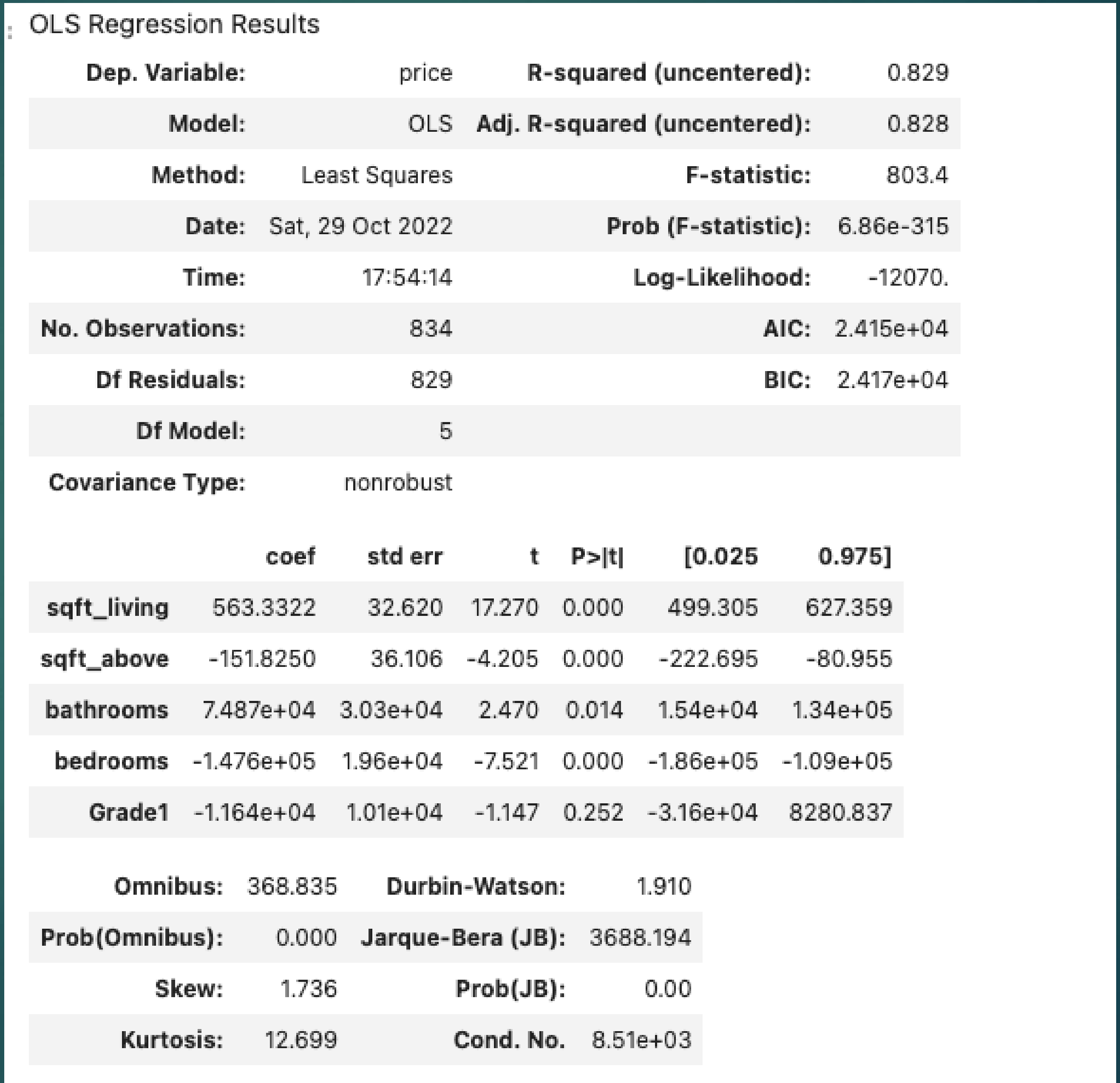

Regression Line for House Prices
1e6
7
6
5
4
3
2
1
0
Price
Regression Line
Actual Prices
0
2000
4000
6000
8000
10000
Square Footage of Living Area

Modeling

The modeling process began with a simple linear regression, examining the relationship between square footage of living space and house prices. However, the initial model showed a very low R-squared value, indicating that other factors beyond living space significantly influence house prices. To address this limitation, OneHot Encoding was introduced in the second iteration to handle categorical variables, resulting in an improved model performance. Subsequently, Log transformation was applied in the third iteration to reduce skewness and further enhance the model's R-squared value.

In the final model, a multiple linear regression was employed, incorporating additional features to Insignificant variables were identified and dropped, resulting in the highest R-squared value and the best fit of residuals to a normal distribution. Overall, the iterative process led to a significantly improved understanding of housing price determinants and a more accurate predictive model.

Regression Results.



OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.829
Model:	OLS	**Adj. R-squared (uncentered):**	0.828
Method:	Least Squares	**F-statistic:**	803.4
Date:	Sat, 29 Oct 2022	**Prob (F-statistic):**	6.86e-315
Time:	17:54:14	**Log-Likelihood:**	-12070.
No. Observations:	834	**AIC:**	2.415e+04
Df Residuals:	829	**BIC:**	2.417e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sqft_living	563.3322	32.620	17.270	0.000	499.305	627.359
sqft_above	-151.8250	36.106	-4.205	0.000	-222.695	-80.955
bathrooms	7.487e+04	3.03e+04	2.470	0.014	1.54e+04	1.34e+05
bedrooms	-1.476e+05	1.96e+04	-7.521	0.000	-1.86e+05	-1.09e+05
Grade1	-1.164e+04	1.01e+04	-1.147	0.252	-3.16e+04	8280.837

Omnibus:	368.835	**Durbin-Watson:**	1.910
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3688.194
Skew:	1.736	**Prob(JB):**	0.00
Kurtosis:	12.699	**Cond. No.**	8.51e+03

Conclusion.

The multiple linear regression model is better than the simple linear regression model because The multiple linear regression model has a slightly higher R-squared value (0.498) compared to the simple linear regression model (0.498 vs. 0.473). A higher R-squared value indicates that the multiple linear regression model explains a larger proportion of the variance in the target variable (house prices) compared to the simple linear regression model.

Recommendation

1. Expanding Dataset Variety: Divide the dataset into categories based on property size (small, medium, large) to develop specialized models tailored to different types of properties. Increase dataset diversity by incorporating additional data on various property sizes or from neighboring counties.

2. Optimizing Property Pricing: Utilize the multiple linear regression model to fine-tune property pricing strategies. Utilize features like square footage, location (zipcode, latitude, longitude), and overall condition (grade, waterfront, view) to accurately evaluate property values and establish competitive yet profitable listing prices.

3. Focusing on Key Property Attributes: Spotlight and prioritize features that significantly affect property value, such as living space (sqft_living), bedroom and bathroom count, construction quality (grade), and proximity to amenities (waterfront, view). Emphasize these features in marketing materials to attract targeted buyer segments effectively.

4. Investing in Renovation and Upgrades: Identify properties with potential for value enhancement based on regression coefficients (e.g., sqft_above, sqft_basement). Consider strategic renovations and upgrades to maximize ROI and appeal to discerning buyers.

Model Improvement.

1. Feature Engineering:

1. Explore additional relevant features that might influence house prices, such as the number of bedrooms, bathrooms, location factors, amenities, and neighborhood characteristics.
2. Conduct thorough data analysis and research to identify potential predictors that have a strong correlation with house prices.

2. Address Multicollinearity:

1. Investigate the presence of multicollinearity among the predictor variables, especially in the multiple linear regression model.
2. Use techniques such as variance inflation factor (VIF) analysis to identify and mitigate multicollinearity by removing highly correlated predictors or employing dimensionality reduction techniques.

3. Model Assumptions:

1. Validate the assumptions of the regression models, including linearity, homoscedasticity, normality of residuals, and independence of errors.
2. Apply appropriate transformations or adjustments to the data to meet these assumptions if necessary.

4. Regularization Techniques:

1. Implement regularization techniques like Ridge regression or Lasso regression to prevent overfitting and improve model generalization, especially in cases of high-dimensional data or multicollinearity.

By Paul, Pascalia, Harry, Ronney

ANY QUESTION???????