"Predicting House Prices: Data-Driven Insights for Star Real Estate Agency"

Joby Varghese
July 4, 2023

Summary

Analysis of the KC House dataset reveals the following observations that Star Real Estate can take into consideration:

1. Larger living areas correlate with higher prices. Consider this when marketing and valuing properties.
2. Newer properties tend to be more expensive. Highlight the age of the house when listing and pricing it.
3. Houses with multiple floors, especially 2-4 floors, have higher prices. Factor this into property valuations.
4. More bedrooms are associated with higher prices. Highlight the number of bedrooms in property listings.
5. Waterfront views increase house prices. Emphasize this feature when valuing waterfront properties.

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

1. Star Real Estate aims to conduct a comprehensive analysis of the property market in King County, gaining valuable insights into market trends, buyer preferences, and investment opportunities.
2. The agency seeks to implement predictive modeling to accurately forecast the indicative sale price of properties listed in their portfolio. This modeling approach will leverage historical data, market factors, and property attributes to provide reliable price estimates.
3. Building on the predictive model, Star Real Estate intends to enhance its services by offering a range of benefits to its customers. These include but are not limited to, accurate Pricing, market insights, targeted marketing strategies etc.

Data

For this analysis, a dataset encompassing historical property sales in King County was utilized. The dataset covers the period from May 2014 to May 2015 and includes various features associated with individual properties. Among the notable variables considered in this analysis are:

1. Price: The sale price of the property.
2. Total Bedrooms: The total number of bedrooms in the property.
3. Total Bathrooms: The total number of bathrooms in the property.
4. Sqft_Living Square: The square footage of the living area.
5. Sqft_Lot Square: The square footage of the lot or land area.
6. Waterfront view: Indicates whether the property has a view of the waterfront.
7. Grade: The grade or overall quality of the property.
8. Year Built: The year when the property was constructed.

Methods

1. The dataset used for analysis had minimal null values, which were appropriately addressed to ensure the reliability of the results.
2. A correlation matrix was generated to examine the relationships between the price variable and the independent variables. This analysis helped to identify the strength and direction of the correlations, providing valuable insights into the factors influencing property prices.
3. To enhance the simplicity and minimize potential noise or redundancy in the model, the following variables were dropped from the dataset: ['id','date','view','yr_renovated', 'zipcode','lat', 'long', 'sqft_lot15','sqft_living15'].
4. A baseline model for Multi Linear Regression was developed using the predictor variables without any transformation. This model served as a starting point to understand the initial relationship between the predictors and the target variable, allowing for further analysis and improvement of the model.
5. Further, the dataset underwent several transformations at different intervals to improve the model performance. These iterations involved applying various techniques such as log transformations, dummy variable encoding, and normalization to the dataset. Each transformation aimed to enhance the relationship between the predictor variables and the target variable, leading to more accurate predictions.

Results

ITERATION 1 - BASELINE MODEL

OLS Regression Results

Dep. Variable:	price	R-squared:	0.646
Model:	OLS	Adj. R-squared:	0.646
Method:	Least Squares	F-statistic:	3937.
Date:	Tue, 04 Jul 2023	Prob (F-statistic):	0.00
Time:	23:23:17	Log-Likelihood:	-2.9618e+05
No. Observations:	21597	AIC:	5.924e+05
Df Residuals:	21586	BIC:	5.925e+05
Df Model:	10		
Covariance Type:	nonrobust		

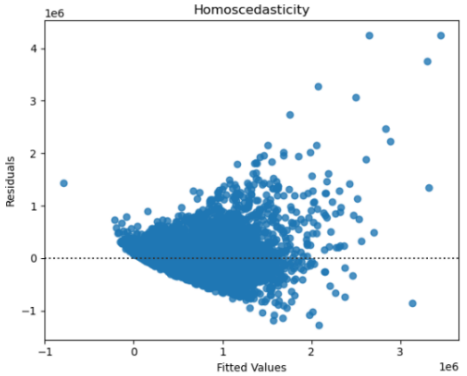
	coef	std err	t	P>|t|	[0.025	0.975]
const	6.625e+06	1.31e+05	50.477	0.000	6.37e+06	6.88e+06
bedrooms	-4.25e+04	2050.316	-20.729	0.000	-4.65e+04	-3.85e+04
bathrooms	4.849e+04	3513.755	13.801	0.000	4.16e+04	5.54e+04
sqft_living	120.3876	2.290	52.563	0.000	115.898	124.877
sqft_lot	-0.2288	0.037	-6.187	0.000	-0.301	-0.156
floors	2.64e+04	3778.167	6.988	0.000	1.9e+04	3.38e+04
waterfront	7.529e+05	1.84e+04	40.993	0.000	7.17e+05	7.89e+05
condition	1.839e+04	2492.859	7.378	0.000	1.35e+04	2.33e+04
grade	1.311e+05	2176.858	60.240	0.000	1.27e+05	1.35e+05
sqft_above	52.2943	2.202	23.748	0.000	47.978	56.611
sqft_basement	68.0933	2.766	24.614	0.000	62.671	73.516
yr_built	-3812.9436	67.363	-56.603	0.000	-3944.980	-3680.907

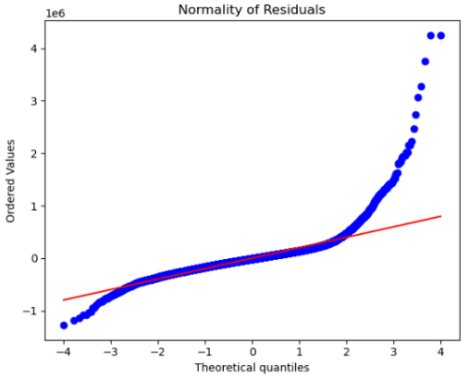
Omnibus:	15929.891	Durbin-Watson:	1.976
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1028236.936
Skew:	2.956	Prob(JB):	0.00
Kurtosis:	36.282	Cond. No.	1.23e+17

The adjusted R-squared value of 64.6% indicates that approximately two-thirds of the variation in house prices can be explained by the independent variables in the model. This suggests that the model has a moderate level of predictive power.

Overall, the model suggests that the number of bedrooms, bathrooms, square footage, condition, grade, waterfront view, and other variables have a significant influence on the property price in the King County real estate market.

However, there is an illogical representation for the predictor variable 'bedrooms'. according to the model, an increase of one bedroom is associated with a decrease in house price by that amount. Typically, one would expect an increase in the number of bedrooms to positively impact the price of a house. This contradictory result warrants further investigation and consideration.





Results

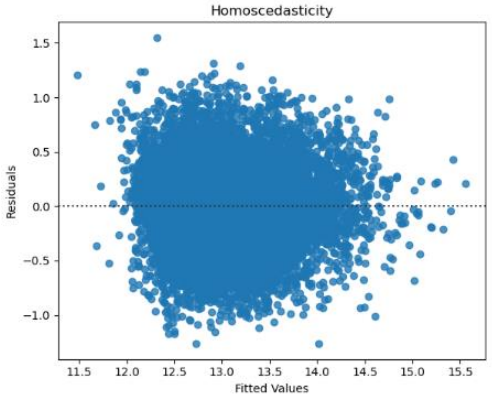
ITERATION 2

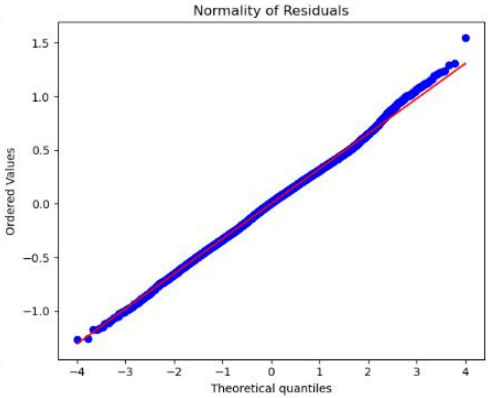
OLS Regression Results

Dep. Variable:	price	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.614
Method:	Least Squares	F-statistic:	1639.
Date:	Tue, 04 Jul 2023	Prob (F-statistic):	0.00
Time:	23:23:37	Log-Likelihood:	-6494.9
No. Observations:	21597	AIC:	1.303e+04
Df Residuals:	21575	BIC:	1.321e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.4881	0.080	118.777	0.000	9.331	9.645
sqft_living	0.6418	0.010	62.248	0.000	0.622	0.662
sqft_lot	-0.0362	0.003	-11.688	0.000	-0.042	-0.030
bedrooms_2_3	-0.1145	0.008	-14.639	0.000	-0.130	-0.099
bedrooms_3_4	-0.1509	0.009	-16.316	0.000	-0.169	-0.133
bedrooms_5plus	-0.1648	0.012	-13.711	0.000	-0.188	-0.141
bathrooms_1_2	-0.0152	0.006	-2.502	0.012	-0.027	-0.003
bathrooms_3_4	0.1007	0.008	12.030	0.000	0.084	0.117
bathrooms_4_5	0.1656	0.020	8.377	0.000	0.127	0.204
bathrooms_5plus	0.2734	0.043	6.325	0.000	0.189	0.358
floors_2_3	0.1680	0.013	13.113	0.000	0.143	0.193
floors_3_4	0.2289	0.124	1.849	0.064	-0.014	0.472
grade_3_5	-1.0524	0.066	-15.991	0.000	-1.181	-0.923
grade_5_7	-0.9596	0.020	-47.322	0.000	-0.999	-0.920
grade_7_9	-0.6579	0.018	-37.324	0.000	-0.692	-0.623
grade_9_11	-0.3049	0.017	-18.042	0.000	-0.338	-0.272
waterfront_1.0	0.5914	0.027	21.517	0.000	0.537	0.645
condition_2	-0.2060	0.026	-8.035	0.000	-0.256	-0.156
condition_3	-0.0548	0.006	-9.934	0.000	-0.066	-0.044
condition_5	0.0720	0.009	7.876	0.000	0.054	0.090
yr_built_1950_2000	-0.2778	0.006	-44.233	0.000	-0.290	-0.265
yr_built_2000_2021	-0.3120	0.008	-38.037	0.000	-0.328	-0.296

Omnibus:	25.870	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30.404
Skew:	0.006	Prob(JB):	2.50e-07
Kurtosis:	3.183	Cond. No.	661.





Results

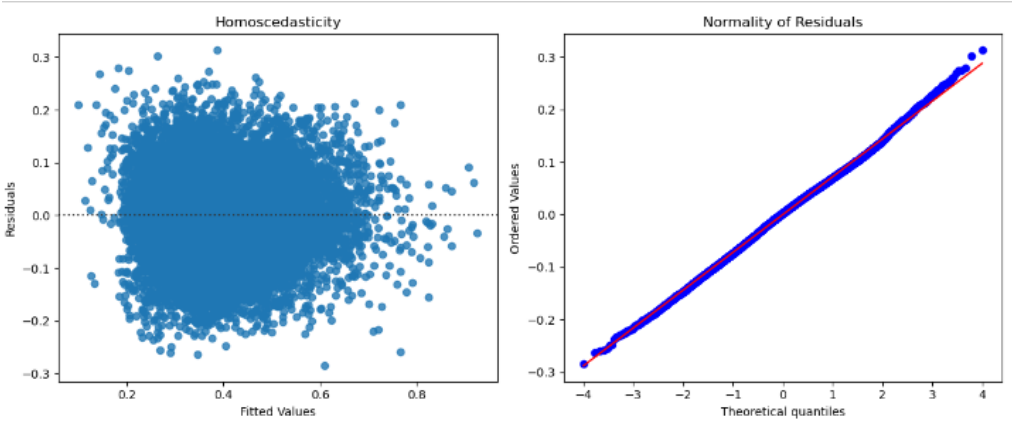
ITERATION 3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.604
Model:	OLS	Adj. R-squared:	0.604
Method:	Least Squares	F-statistic:	1648.
Date:	Tue, 04 Jul 2023	Prob (F-statistic):	0.00
Time:	23:23:58	Log-Likelihood:	26142.
No. Observations:	21597	AIC:	-5.224e+04
Df Residuals:	21576	BIC:	-5.207e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3587	0.006	59.900	0.000	0.347	0.370
sqft_living	0.5288	0.008	64.991	0.000	0.513	0.545
sqft_lot	-0.1209	0.005	-23.598	0.000	-0.131	-0.111
property_age	0.0262	0.001	39.880	0.000	0.025	0.028
bedrooms_2_3	-0.0360	0.002	-21.274	0.000	-0.039	-0.033
bedrooms_3_4	-0.0439	0.002	-21.765	0.000	-0.048	-0.040
bedrooms_5plus	-0.0513	0.003	-19.476	0.000	-0.056	-0.046
bathrooms_1_2	-0.0087	0.001	-6.296	0.000	-0.011	-0.006
bathrooms_3_4	0.0260	0.002	14.109	0.000	0.022	0.030
bathrooms_4_5	0.0424	0.004	9.720	0.000	0.034	0.051
bathrooms_5plus	0.0679	0.010	7.126	0.000	0.049	0.087
floors_2_3	0.0422	0.003	14.923	0.000	0.037	0.048
floors_3_4	0.0488	0.027	1.786	0.074	-0.005	0.102
grade_3_5	-0.2132	0.015	-14.700	0.000	-0.242	-0.185
grade_5_7	-0.2009	0.004	-45.047	0.000	-0.210	-0.192
grade_7_9	-0.1476	0.004	-37.901	0.000	-0.155	-0.140
grade_9_11	-0.0653	0.004	-17.526	0.000	-0.073	-0.058
waterfront_1	0.1260	0.006	20.767	0.000	0.114	0.138
condition_2	-0.0369	0.006	-6.534	0.000	-0.048	-0.026
condition_3	-0.0044	0.001	-3.595	0.000	-0.007	-0.002
condition_5	0.0178	0.002	8.844	0.000	0.014	0.022

Omnibus:	15.257	Durbin-Watson:	1.964
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15.706
Skew:	-0.049	Prob(JB):	0.000389
Kurtosis:	3.088	Cond. No.	93.6



Results

ITERATION 4

OLS Regression Results

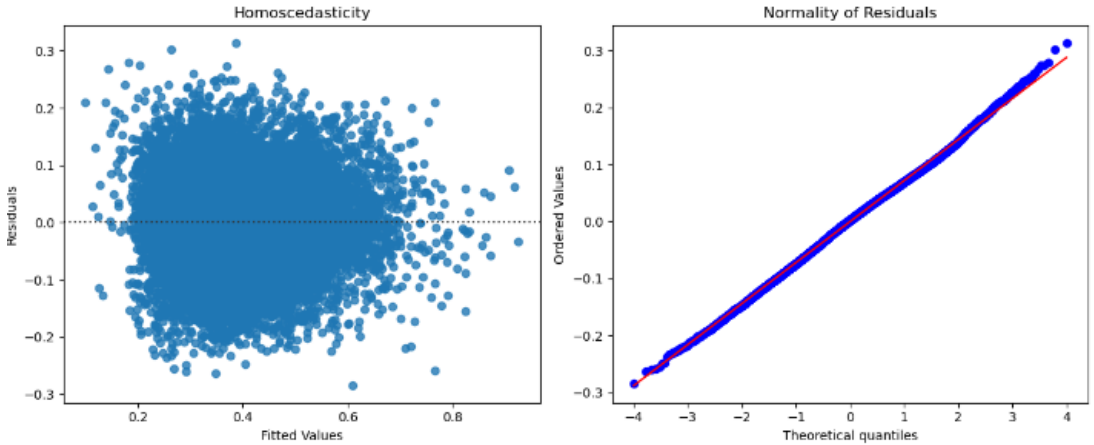
Dep. Variable:	price	R-squared:	0.604
Model:	OLS	Adj. R-squared:	0.604
Method:	Least Squares	F-statistic:	1569.
Date:	Tue, 04 Jul 2023	Prob (F-statistic):	0.00
Time:	23:24:14	Log-Likelihood:	26142.
No. Observations:	21597	AIC:	-5.224e+04
Df Residuals:	21575	BIC:	-5.206e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3070	0.011	28.164	0.000	0.286	0.328
sqft_living	0.5288	0.008	64.627	0.000	0.513	0.545
sqft_lot	-0.1209	0.005	-23.571	0.000	-0.131	-0.111
property_age	0.0262	0.001	39.851	0.000	0.025	0.028
floors_2_3	0.0422	0.003	14.922	0.000	0.037	0.048
floors_3_4	0.0488	0.027	1.785	0.074	-0.005	0.102
grade_3_5	-0.2130	0.015	-14.140	0.000	-0.243	-0.184
grade_5_7	-0.2009	0.004	-45.045	0.000	-0.210	-0.192
grade_7_9	-0.1476	0.004	-37.895	0.000	-0.155	-0.140
grade_9_11	-0.0653	0.004	-17.525	0.000	-0.073	-0.058
waterfront_1	0.1260	0.006	20.746	0.000	0.114	0.138
condition_2	-0.0369	0.006	-6.533	0.000	-0.048	-0.026
condition_3	-0.0044	0.001	-3.595	0.000	-0.007	-0.002
condition_5	0.0178	0.002	8.844	0.000	0.014	0.022
bedrooms_1_2	0.0513	0.003	19.473	0.000	0.046	0.056
bedrooms_2_3	0.0153	0.002	7.455	0.000	0.011	0.019
bedrooms_3_4	0.0073	0.002	3.780	0.000	0.004	0.011
bathrooms_1_2	-0.0083	0.009	-0.934	0.350	-0.026	0.009
bathrooms_2_3	0.0004	0.009	0.049	0.961	-0.017	0.018
bathrooms_3_4	0.0264	0.009	2.889	0.004	0.009	0.044
bathrooms_4_5	0.0428	0.010	4.280	0.000	0.023	0.062
bathrooms_5plus	0.0684	0.013	5.207	0.000	0.043	0.094

Omnibus:	15.260	Durbin-Watson:	1.964
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15.711
Skew:	-0.049	Prob(JB):	0.000388
Kurtosis:	3.089	Cond. No.	97.7

1. Based on the results of Iteration 4, the model exhibits an adjusted R-squared value of 0.604, indicating that it explains approximately 60.4% of the variance in house prices.
2. The F-statistic of 1569.0 with a very low p-value suggests the overall significance of the model.
3. Several variables show statistically significant coefficients with p-values less than 0.05, including sqft_living, sqft_lot, property_age, floors_2_3, grade categories, waterfront, condition categories, and bedroom categories. These variables play a crucial role in determining house prices, as indicated by their coefficients.
4. The variables floors_3_4, bathrooms_1_2, and bathrooms_2_3 have p-values greater than 0.05, suggesting that they are not significant in explaining the variation in house prices.
5. Overall, the reliability of Iteration 4 is strengthened by its robust statistical measures, significant variable coefficients, and satisfactory diagnostic tests.

Results



K-Fold Validation Score

```
Mean Train Mean Squared Err
or: 0.00496
Mean Test Mean Squared Erro
r: 0.00497
```

The scatter plot shows the relationship between the predicted values and the residuals of the model. In this plot, the data points are evenly spread around the line of best fit, indicating a roughly constant spread of residuals. This pattern suggests that the assumption of homoscedasticity, which assumes constant variance, holds true in this case.

The Q-Q plot of the residuals shows that the residuals in the current model closely aligns with the diagonal line, it suggests that the residuals follow a normal distribution. This alignment indicates that the assumption of normality for the residuals is reasonable.

Recommendations

1. A key factor in determining house prices is the living area. The coefficient for this variable indicates that larger living areas tend to have higher prices. Consider this feature when marketing properties and estimating a home's value.
2. When determining house prices, it is important to consider the property's age. According to the coefficient for this variable, newer properties are generally more expensive. The age of the house should be highlighted in property listings and considered when pricing the property.
3. During the evaluation of house prices, the number of floors (especially houses with floors 2-4) should be considered. The variables are statistically significant, indicating that houses with more than one floor sell for more. When valuing and marketing properties, consider the number of floors as a desirable feature.
4. The number of bedrooms is an important factor to consider in determining house prices. Three variables are statistically significant, indicating that properties with more bedrooms tend to be more expensive. When estimating house values, emphasize the number of bedrooms in property listings.
5. House prices are influenced by the presence of a waterfront view. In this variable, the coefficient is positive and highly significant, indicating that houses with waterfront views tend to command higher prices. Consider this feature when estimating the value of waterfront properties and highlight it in property descriptions.

Thank You!

Email: grace.joby@gmail.com
GitHub: @VargheseJoby
LinkedIn: linkedin.com/in/joby-varghese-516