Housing in King County

Author: Christine Li

Summary

The objective of this project is to create a regression model to predict home sale pricing in King County.

Using this multiple regression model, homeowners can then make informed decisions about targeted renovations for certain property features to increase the estimated value of their homes.

Data

This project will use the King County House Sales dataset. This data set contains information on historic home sales and property features within the target real estate of King County.

Image on the right contains the title and descriptions of the data set.

- id - unique identifier for a house
- dateDate - house was sold
- pricePrice - is prediction target
- bedroomsNumber - of Bedrooms/House
- bathroomsNumber - of bathrooms/bedrooms
- sqft_livingsquare - footage of the home
- sqft_lotsquare - footage of the lot
- floorsTotal - floors (levels) in house
- waterfront - House which has a view to a waterfront
- view - Has been viewed
- condition - How good the condition is (Overall)
- grade - overall grade given to the housing unit, based on King County grading system
- sqft_above - square footage of house apart from basement
- sqft_basement - square footage of the basement
- yr_built - Built Year
- yr_renovated - Year when house was renovated
- zipcode - zip
- lat - Latitude coordinate
- long - Longitude coordinate
- sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors
- sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

Methods (OEMIN)

1. Obtain the data
    a. Clean the data
    b. Deal with missing values
2. Explore the data
    a. Initial data analysis and visualisation
3. Model the data
    a. Select the features and target variable
    b. Baseline model
        i. Deal with categorical variables and implement dummies
        ii. Check for multicollinearity
        iii. Check it meets the multiple regression assumptions
    c. Second model iteration
        i. Deal with multicollinearity
        ii. Log transform and standardise continuous variables
    d. Third model iteration
        i. Train/Test Split Model Validation
4. Interpret the data

Results (Final Model)

R^2 value: 0.55

55% of the variance in the target variable price can be explained by the predictor features.

p value: 0

The p-value is lower than 0.05 so we can reject the null hypothesis.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.553
Model:	OLS	**Adj. R-squared:**	0.553
Method:	Least Squares	**F-statistic:**	4013.
Date:	Sat, 10 Feb 2024	**Prob (F-statistic):**	0.00
Time:	23:35:26	**Log-Likelihood:**	-6090.9
No. Observations:	16197	**AIC:**	1.219e+04
Df Residuals:	16191	**BIC:**	1.224e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	11.6881	0.034	345.508	0.000	11.622	11.754
sqft_living_log	0.2281	0.006	38.437	0.000	0.216	0.240
sqft_lot_log	-0.0376	0.003	-12.436	0.000	-0.044	-0.032
bedrooms	-0.0315	0.004	-7.695	0.000	-0.040	-0.023
bathrooms	-0.0227	0.006	-3.798	0.000	-0.034	-0.011
grade	0.1978	0.004	53.026	0.000	0.191	0.205

Omnibus:	84.909	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	86.118
Skew:	0.177	**Prob(JB):**	1.99e-19
Kurtosis:	2.952	**Cond. No.**	109.

Conclusions

Square footage and grade are the best predictors of a house's price in King County. We can see that these variables are positively related to price and therefore to get the best price for the house, renovations should be done in these areas.

These features share:

- Strong linear relationship with price
- Have relatively low multicollinearity,
- Low p-value rejecting the null hypothesis
- Positive coefficient.

Homeowners should target their renovations into increasing the square footage and consider increasing number of bathrooms/bedrooms as well as and improving the overall quality of the house.

Limitations

Some of the challenges:

- Some of the variables had to be log-transformed to satisfy the multiple regression assumptions and therefore any new data would have to go through the same process.
- Looking at the original distribution of the variables, there was some outliers that were not removed from the model. In the future when applying new data, further consideration would need to be made on whether it is worth removing the outliers or not as outliers inherently can skew and influence the results.

Thank You