King County House Price Prediction

Phase 2 Project

Introduction

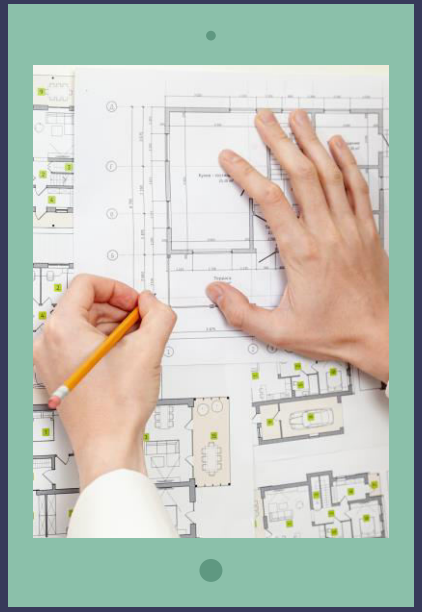
STAKEHOLDER

Real estate agency that builds and sells homes

BUSINESS PROBLEM

The need to provide advice to the Real Estate company about how home features might influence the price of a house.

Background



- The global real estate market size was valued at USD 3.69 trillion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 5.2% from 2022 to 2030 – Grand View Research.
- Real Estate players need to be in a position to price their units correctly based on features.
- Real estate players need to know which features to include in the units they build that will generate the best returns in sales.

Problem

How do we price houses correctly?

How do we know which features to include in a house to attract the most demand?

Solution

Come up with a regression model that can be used to predict house prices & which shows which house feautures stimulate demand the most.

Main Objective

To predict house prices for a real estate agency using a regression model

Specific Objectives

House Prices

To predict house prices for a real estate agency using a regression model

House Features

To determine which features to build into houses to stimulate demand for houses

The Experimental Design

01
Business Understanding

02
Data Understanding

03
Data Preparation

04
Modelling

05
Evaluation

06
Deployment

Exploratory Data Analysis

Partial Linear Regression of All Variables Against Price

	ind_var	r_squared	intercept	slope	p-value	normality (JB)
0	price	1.000000	7.003109e-11	1.000000e+00	0.000000e+00	1.592032e+06
1	bedrooms	0.095073	1.296487e+05	1.217896e+05	0.000000e+00	1.194538e+06
2	bathrooms	0.275766	1.046688e+04	2.504851e+05	0.000000e+00	8.829729e+05
3	sqft_living	0.492865	-4.386760e+04	2.808067e+02	0.000000e+00	5.435339e+05
4	sqft_lot	0.008038	5.281697e+05	7.951601e-01	8.061525e-40	1.147191e+06
5	floors	0.065939	2.791338e+05	1.746950e+05	0.000000e+00	1.255497e+06
6	waterfront	0.070932	5.316534e+05	1.130871e+06	0.000000e+00	9.213256e+05
7	view	0.157884	4.955515e+05	1.904825e+05	0.000000e+00	1.030223e+06
8	condition	0.001324	4.701380e+05	2.054423e+04	8.719407e-08	1.136508e+06
9	grade	0.445507	-1.057041e+06	2.085999e+05	0.000000e+00	2.043215e+06
10	sqft_above	0.366709	5.973768e+04	2.686437e+02	0.000000e+00	7.291583e+05
11	sqft_basement	0.104871	4.618236e+05	2.688033e+02	0.000000e+00	8.892866e+05
12	yr_built	0.002914	-7.905034e+05	6.751304e+02	1.999215e-15	1.142142e+06
13	yr_renovated	0.015988	5.304219e+05	1.156398e+02	9.976356e-78	1.085812e+06
14	sqft_living15	0.342663	-8.310674e+04	3.137541e+02	0.000000e+00	1.911706e+06
15	sqft_lot15	0.006799	5.260169e+05	1.109393e+00	6.321007e-34	1.141778e+06

Square Foot of the Living Area Has The Highest Linear Relationship With Price at 0.49

Multicollinearity Test

	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	sqft_living15	sql
bedrooms	1.000000	0.515884	0.576671	0.031703	0.175429	-0.006582	0.079532	0.028472	0.356967	0.477610	0.303093	0.391638	0.
bathrooms	0.515884	1.000000	0.754665	0.087740	0.500653	0.063744	0.187737	-0.124982	0.664983	0.685326	0.283770	0.568634	0.
sqft_living	0.576671	0.754665	1.000000	0.172826	0.353949	0.103818	0.284611	-0.058753	0.762704	0.876586	0.435043	0.756420	0.
sqft_lot	0.031703	0.087740	0.172826	1.000000	-0.005201	0.021604	0.074710	-0.008958	0.113621	0.183510	0.015286	0.144608	0.
floors	0.175429	0.500653	0.353949	-0.005201	1.000000	0.023698	0.029444	-0.263768	0.458183	0.523889	-0.245705	0.279885	-0.
waterfront	-0.006582	0.063744	0.103818	0.021604	0.023698	1.000000	0.401857	0.016653	0.082775	0.072074	0.080588	0.086463	0.
view	0.079532	0.187737	0.284611	0.074710	0.029444	0.401857	1.000000	0.045990	0.251321	0.167648	0.276947	0.280439	0.
condition	0.028472	-0.124982	-0.058753	-0.008958	-0.263768	0.016653	0.045990	1.000000	-0.144674	-0.158202	0.174105	-0.092824	-0.
grade	0.356967	0.664983	0.762704	0.113621	0.458183	0.082775	0.251321	-0.144674	1.000000	0.755917	0.168392	0.713202	0.
sqft_above	0.477610	0.685326	0.876586	0.183510	0.523889	0.072074	0.167648	-0.158202	0.755917	1.000000	-0.051963	0.731864	0.
sqft_basement	0.303093	0.283770	0.435043	0.015286	-0.245705	0.080588	0.276947	0.174105	0.168392	-0.051963	1.000000	0.200355	0.
sqft_living15	0.391638	0.568634	0.756420	0.144608	0.279885	0.086463	0.280439	-0.092824	0.713202	0.731864	0.200355	1.000000	0.
sqft_lot15	0.029244	0.087175	0.183286	0.718557	-0.011269	0.030703	0.072575	-0.003406	0.119248	0.194047	0.017276	0.183192	1.

Sqft_above, sqft_living, grade, sqft_living15 and bathrooms show high correlation(over 0.75) with each other. Some variables will be dropped.

Model 1 OLS Results

OLS Regression Results

Dep. Variable:	price_log	**R-squared:**	0.625
Model:	OLS	**Adj. R-squared:**	0.624
Method:	Least Squares	**F-statistic:**	543.4
Date:	Sun, 26 Mar 2023	**Prob (F-statistic):**	0.00
Time:	14:05:38	**Log-Likelihood:**	-20076.
No. Observations:	21613	**AIC:**	4.029e+04
Df Residuals:	21546	**BIC:**	4.082e+04
Df Model:	66		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4047	0.615	-0.658	0.511	-1.610	0.801

R Squared of 0.625

Model 2 OLS Results

OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.631
Model:	OLS	Adj. R-squared:	0.630
Method:	Least Squares	F-statistic:	550.9
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	0.00
Time:	11:58:28	Log-Likelihood:	-19881.
No. Observations:	21613	AIC:	3.990e+04
Df Residuals:	21545	BIC:	4.044e+04
Df Model:	67		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4487	0.610	-0.736	0.462	-1.644	0.746

R Squared of 0.631

Model 3 OLS Results

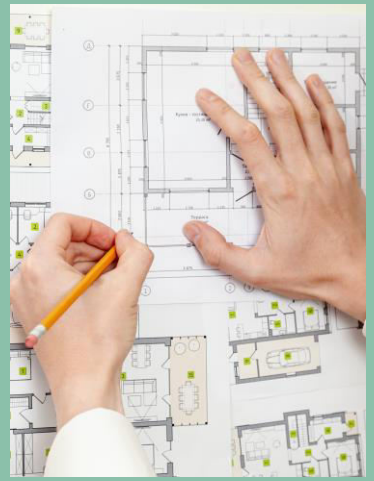
OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.634
Model:	OLS	Adj. R-squared:	0.633
Method:	Least Squares	F-statistic:	548.8
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	0.00
Time:	18:24:08	Log-Likelihood:	-19805.
No. Observations:	21613	AIC:	3.975e+04
Df Residuals:	21544	BIC:	4.030e+04
Df Model:	68		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6590	0.608	-1.084	0.278	-1.850	0.532
sqft_living_log	0.4273	0.012	36.703	0.000	0.404	0.450
sqft_lot_log	0.0027	0.011	0.250	0.803	-0.019	0.024
sqft_living15_log	0.1980	0.007	28.326	0.000	0.184	0.212

R Squared of 0.634

Conclusion



- The final model has an adjusted R-squared value of 0.634.
- With adjustments to the categorical variables and the continuous variables we can more accurately predict our dependent variable 'price'.
- For the continuous variables, we applied log transformations to the continuous data for them to appear more normal.
- Normalization was also applied to reduce the effect of outliers and reduce the values to be between 0-1.
- To handle categorical variables we used one hot encoding.
- The three best predictors for the sale price of a house are square footage of the living area, number of bathrooms/bedrooms and condition of the house

Business Recommendations

- To increase the sale price of a house, we can;
- 1. Increase the amount of bathrooms.
- 2. Increase the number of bedrooms of the property.
- 3. Renovate the property to help improve the condition of the house.
- 4. Increase the square footage of the living area.