King County House Prices Predictions

Benjamin Jaron

TABLE OF CONTENTS

01 OVERVIEW

- Our aim is to deliver our client a data driven approach to predict home prices in King County
- Linear regression modeling to address client's needs

BUSINESS PROBLEM

- Real estate company purchases homes and tries to flip at premium
- Wants to use data on house sales to see if the purchase and sale price are profitable

DATA OVERVIEW

30,311

Non-null entries

25

Columns; 23 aside from price and id to serve as predictors

0.61

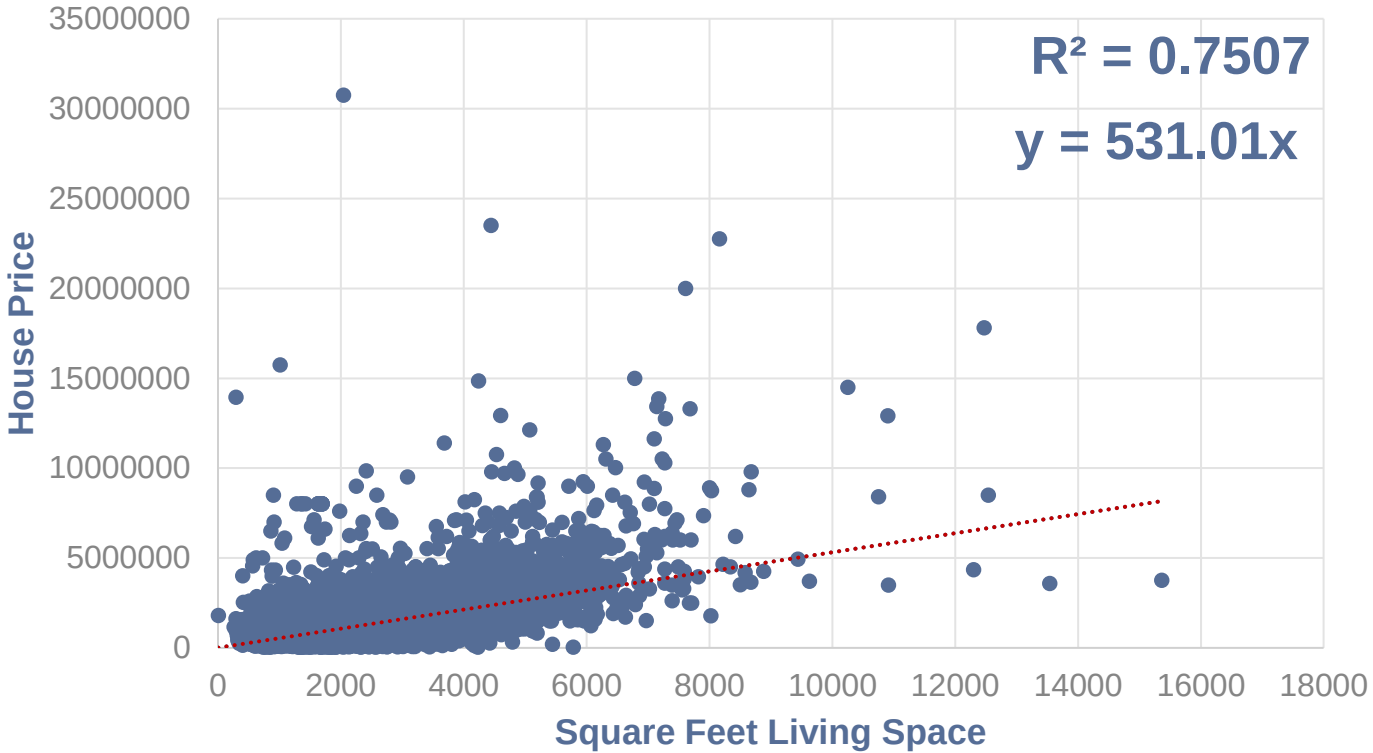
Positive correlation between sqft_living and price

02 REGRESSION

- Start with simple linear regression as baseline model
- Build a multiple regression and test if it beats baseline

Baseline Model

Price vs. Square Feet of Living Space



$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

Variable	Column Name	Coefficient Value
x_1	Bedrooms	-60,300
x_2	Sqft Living Space	102
x_2^2	Sqft Living Space Squared	0.06
x_3	Floors	-89,580
x_4	Sqft Garage	-242
x_5	Sqft Patio	169
x_6	Year Built	177
x_7	Grade	207,100
x_8	View	313,400

Regression Results

04

$$y = -60,300x_1 + 102x_2 + 0.06x_2^2 - 89,580x_3 - 242x_4 + 169x_5 + 177x_6 + 207,100x_7 + 313,400x_8$$

$$R^2 = 0.78$$

Our model explains 78% of total variance in house price and all coefficients are statistically significant

ASSUMPTIONS OF LINEARITY

LINEAR

Unfortunately, statistical testing showed that our model was not linear

INDEPENDENCE OF VARIABLES

In our modeling process, we worked hard to elimante all variables that were highly correlated

NORMALLY DISTRIBUTED RESIDUALS

Statistical testing showed us that our residuals were in fact normally distributed

HOMOSKEDASTICITY

The scatter of the errors (difference between actual and predicted values) is even

RECOMMENDATION

05

USE OUR MODEL, BUT CAUTIOUSLY

- Our model picks up on a decent chunk of variance in home prices
- It is statistically significant
- However it isn't linear
- Good practice to use intuitive real estate methods and use model to check


THE END