KINGS COUNTY PROJECT

Project Overview

For this project, we will use regression modeling to analyze house sales in a northwestern county.

Objective

I have to build a simple linear regression that can help to predict the value of the house King county, WA houses

Benefits

Being able to predict the fluctuation of home prices. Future homeowners; renters, investors, and businesses will all benefit from the outcome of this project. They will be able to appropriately plan and make informed decisions about purchases in regards to real estate.

Methodology

- We first needed to clean the provided data sets for accurate analysis.
- We leveraged the strength of the pandas library in order to ease ourselves the burden of cleaning the data sets using complex methods.

The kings county dataset contains a wealth of information about the price, size, location, condition and various other features of houses in Washington’s King County. In this presentation, I’ll present how I built a multiple linear regression model in Python to predict house prices

Column Names and descriptions for Kings County Data Set

- id - unique identified 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

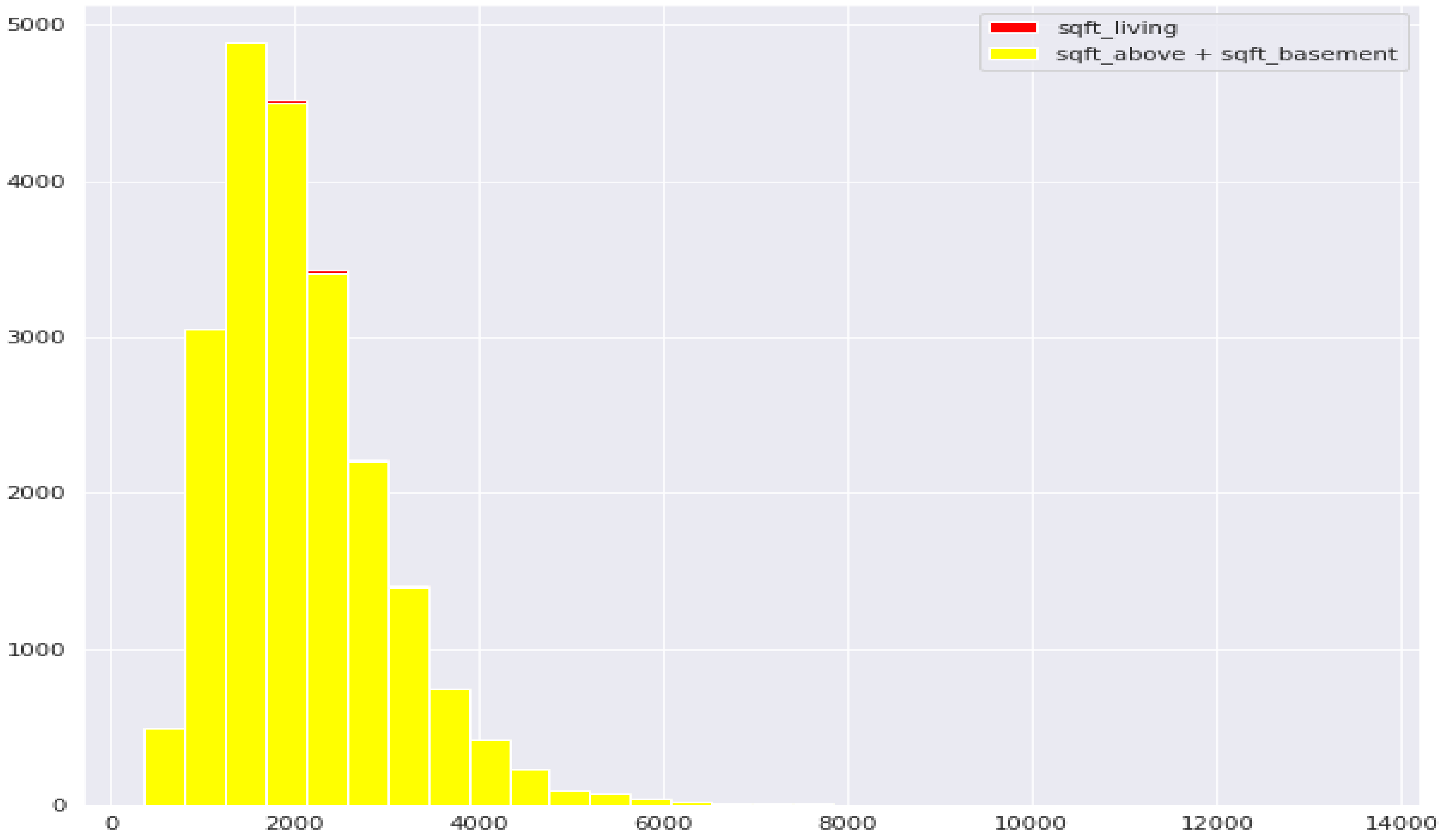
EDA

Majority of the columns were deemed acceptable for analysis and some were transformed for analysis. Especially categorical columns.

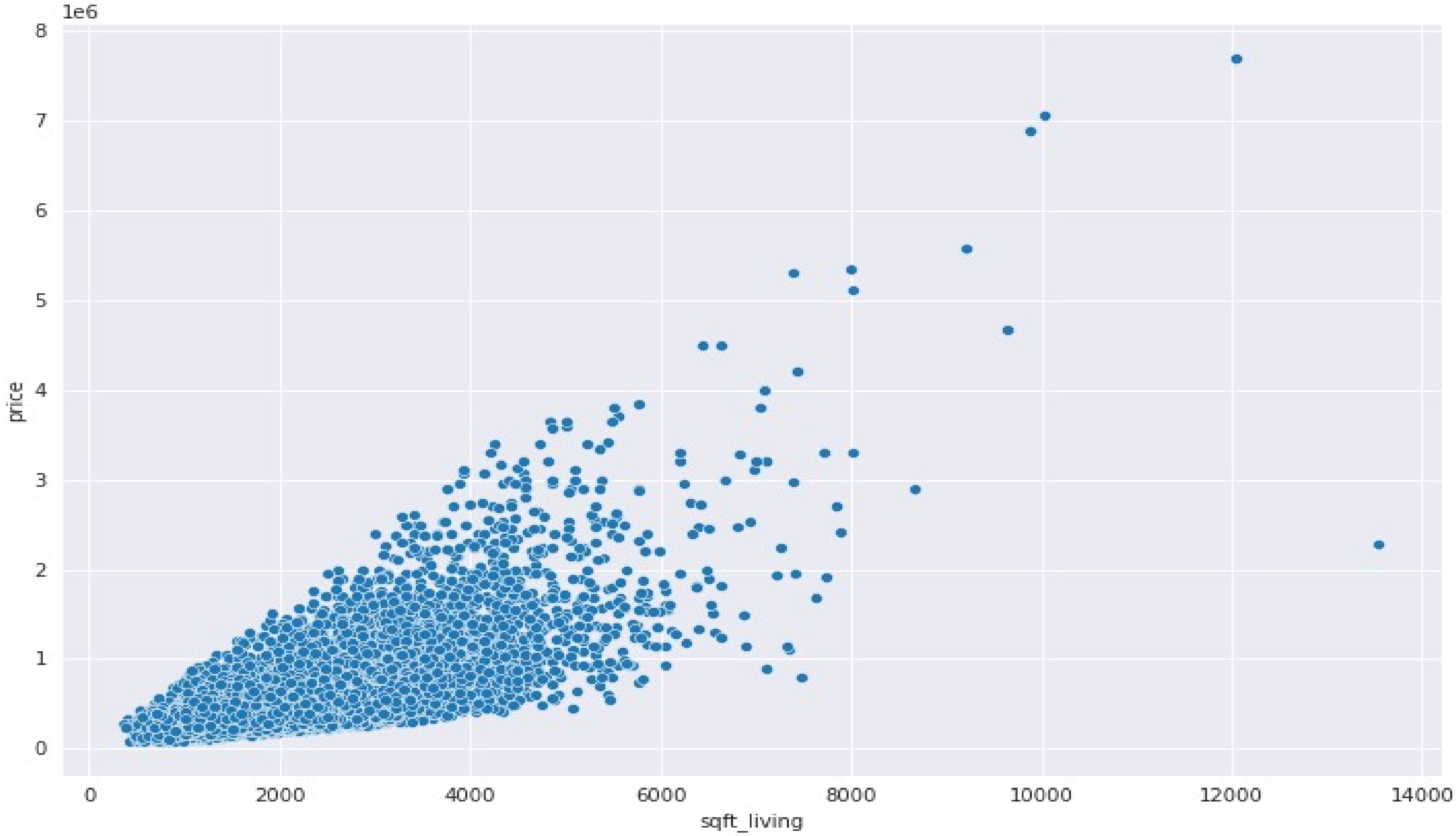
Patterns

- Sqft_living

  It was discovered that this variable had an additive relationship with two other plots ie. sqft above and sqft_basement.

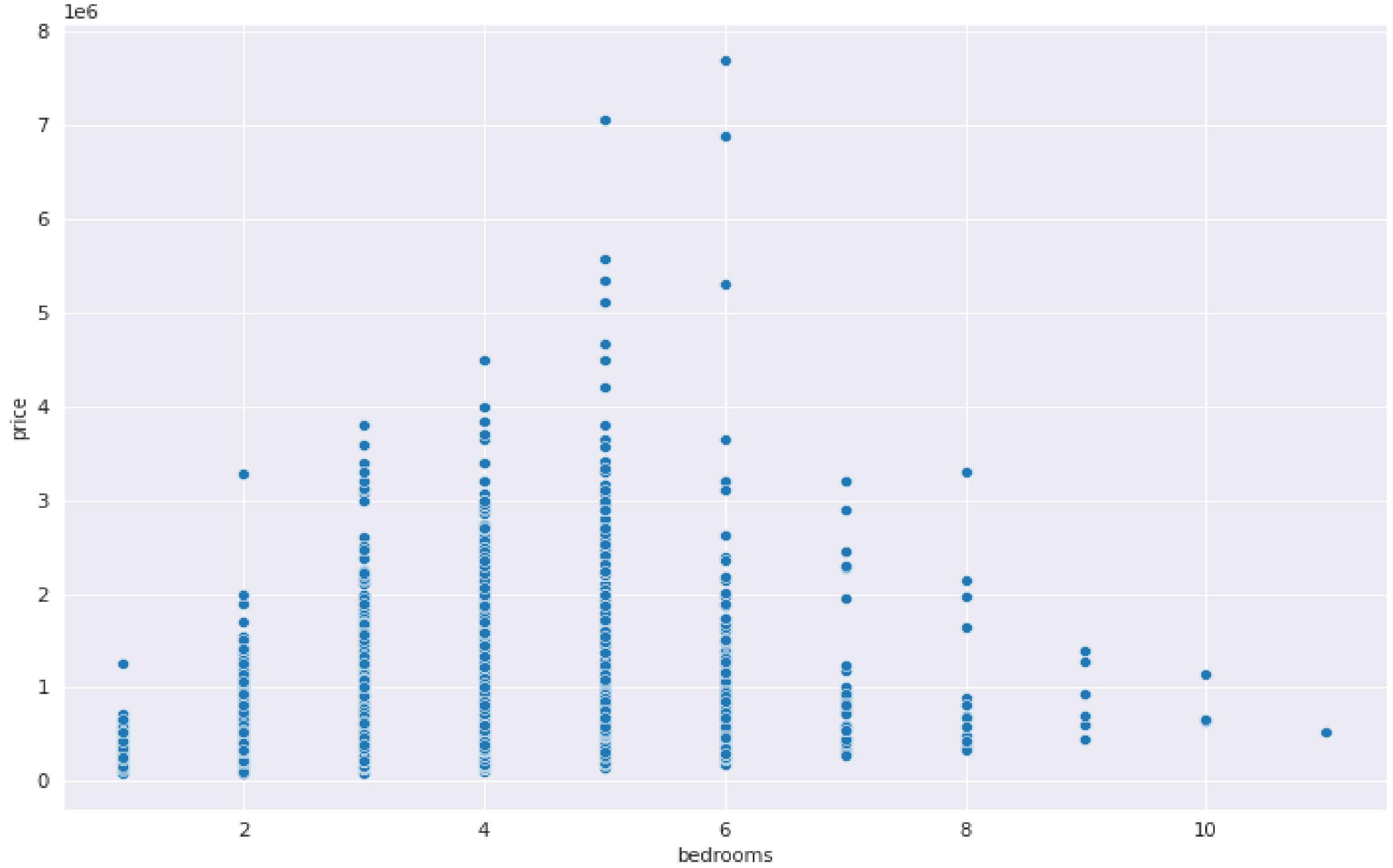

sqft_living
sqft_above + sqft_basement
0
1000
2000
3000
4000
5000
0
2000
4000
6000
8000
10000
12000
14000

The variable also appeared to have a linear relationship with the price.


1e6
8
7
6
5
4
3
2
1
0
price
0
2000
4000
6000
8000
10000
12000
14000
sqft_living

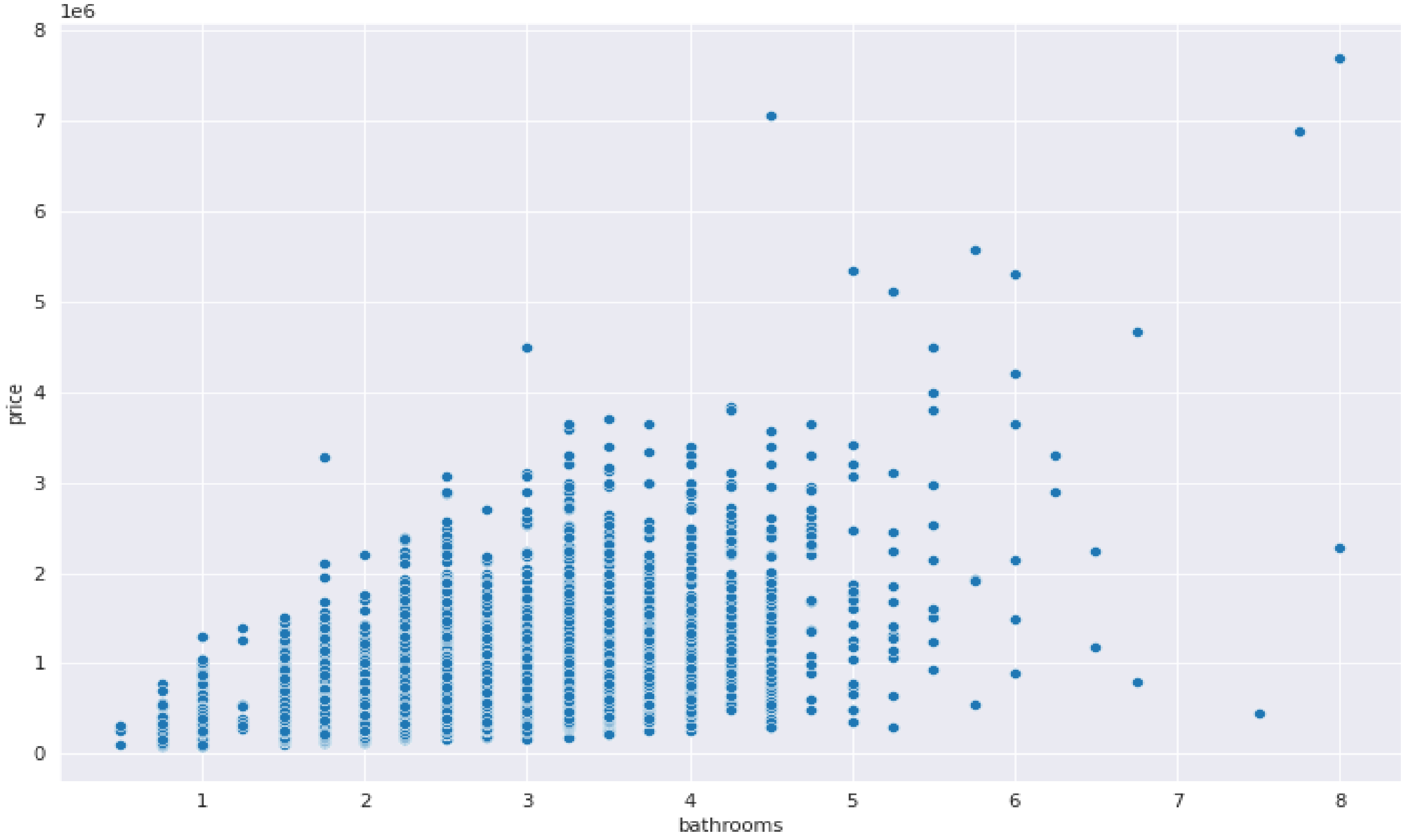
- Bedrooms

  It was noted that 5 bedroomed homes sold at higher prices.


1e6
8
7
6
5
4
3
2
1
0
price
2
4
6
8
10
bedrooms

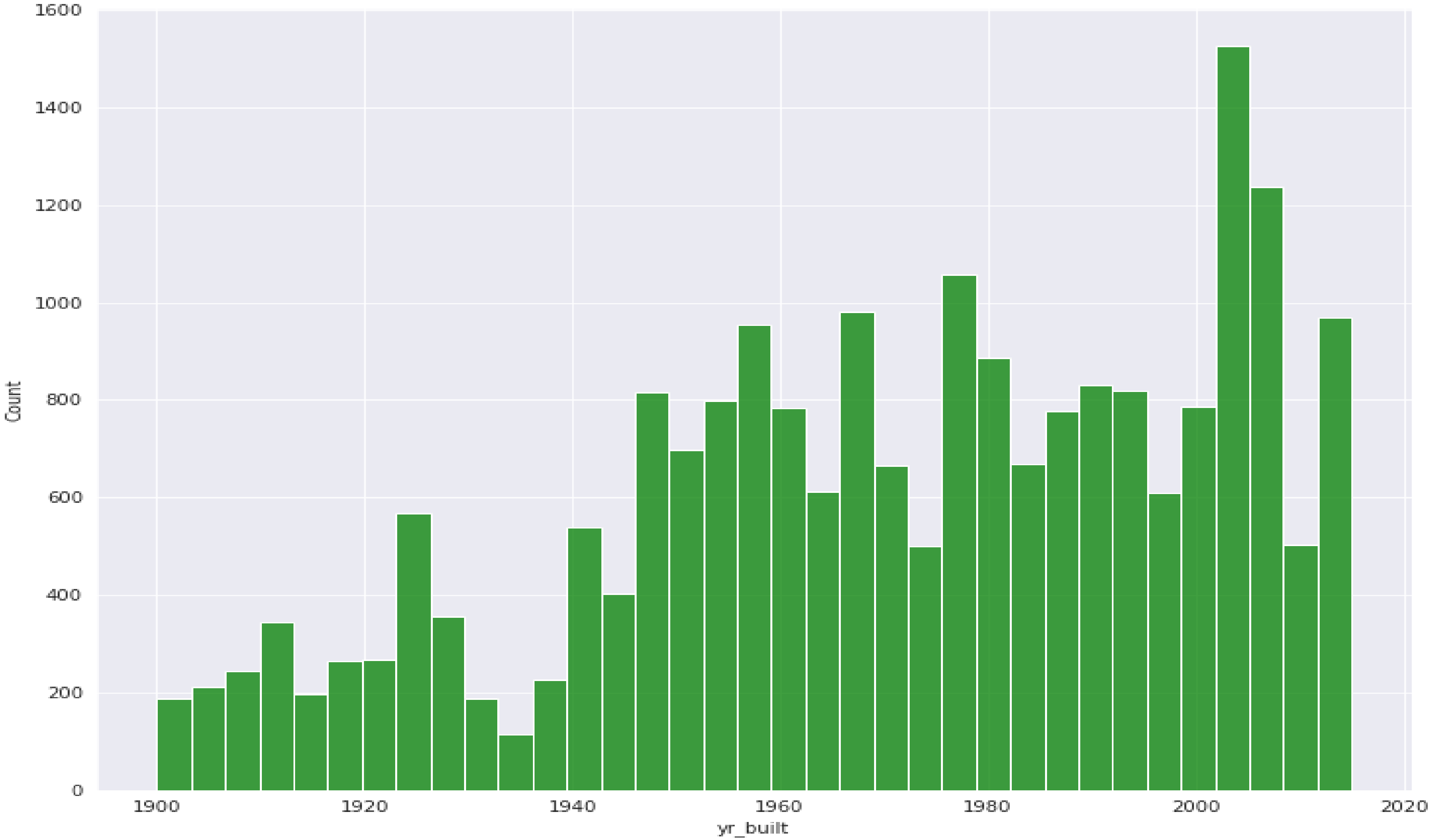
- Bathrooms

  It was observed that prices of homes kept increasing as the number of bathrooms increased.


1e6
8
7
6
5
4
3
2
1
0
price
1
2
3
4
5
6
7
8
bathrooms

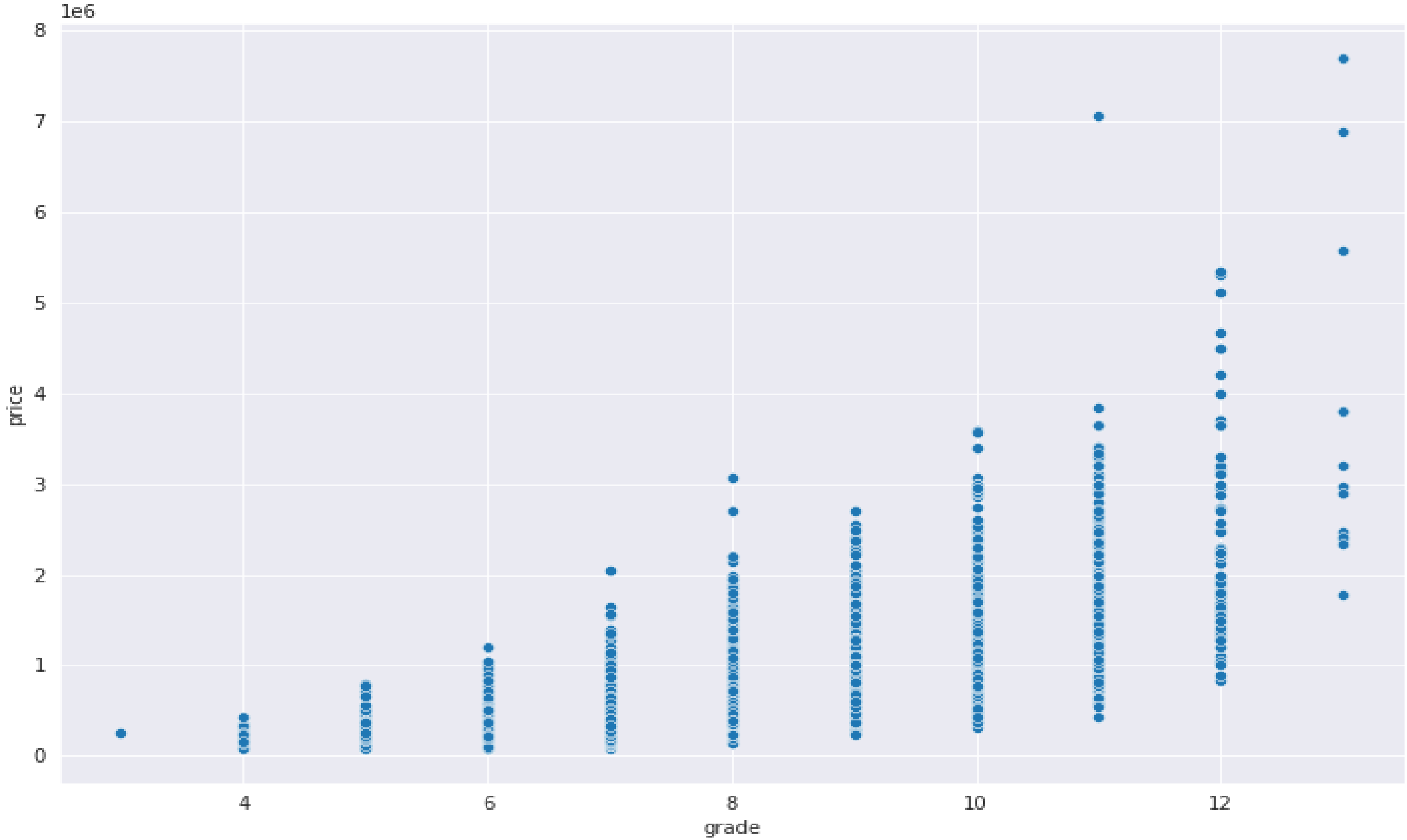
- Houses built

  It a appears there is an increasing in house constructions since 1900. This could be an indicator of increasing demand.


1600
1400
1200
1000
800
600
400
200
0
Count
1900
1920
1940
1960
1980
2000
2020
yr_built

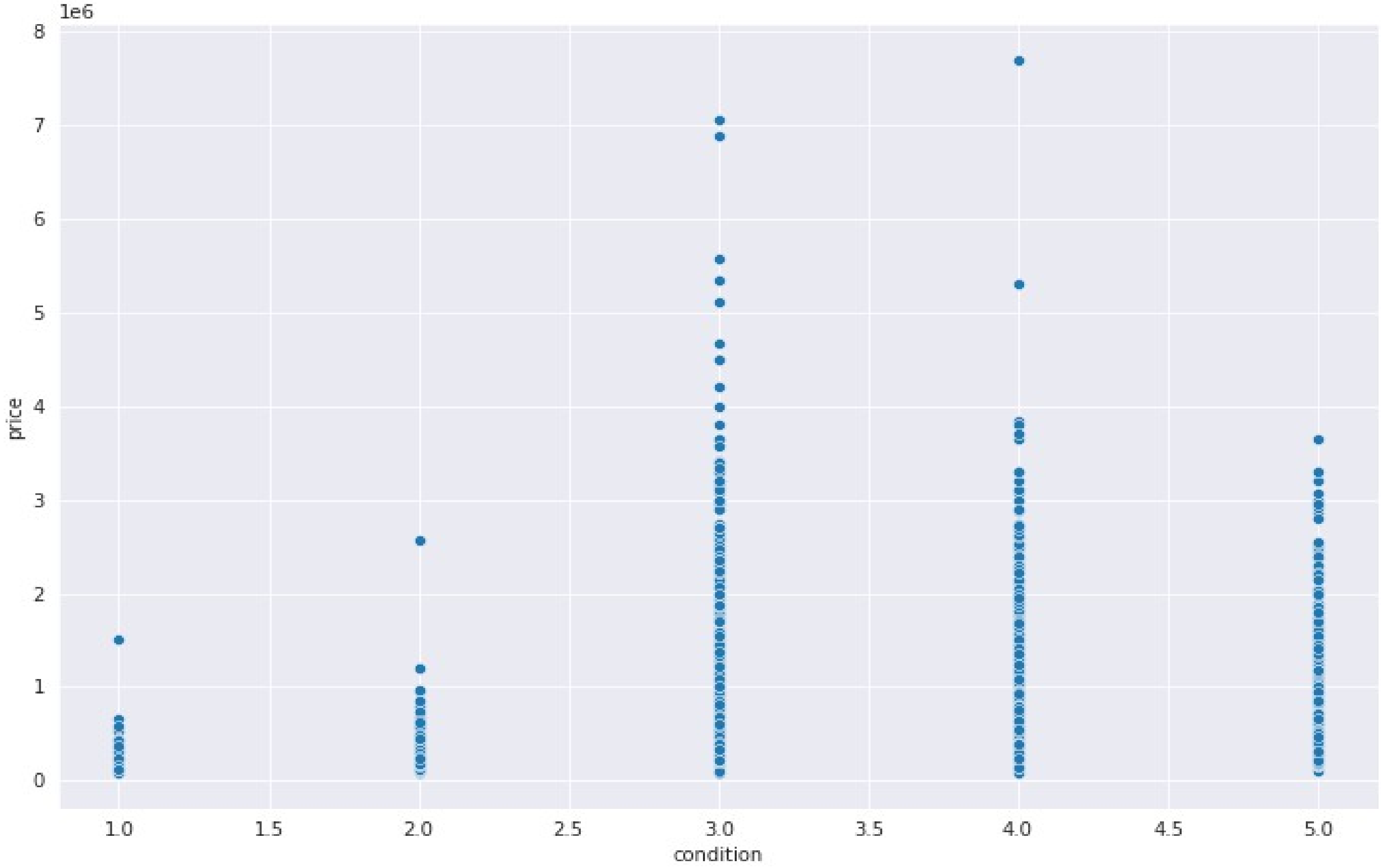
- Grades

  There also appeared to be some evidence that the prices of homes increased as the grade ratings were higher.


1e6
8
7
6
5
4
3
2
1
0
price
4
6
8
10
12
grade

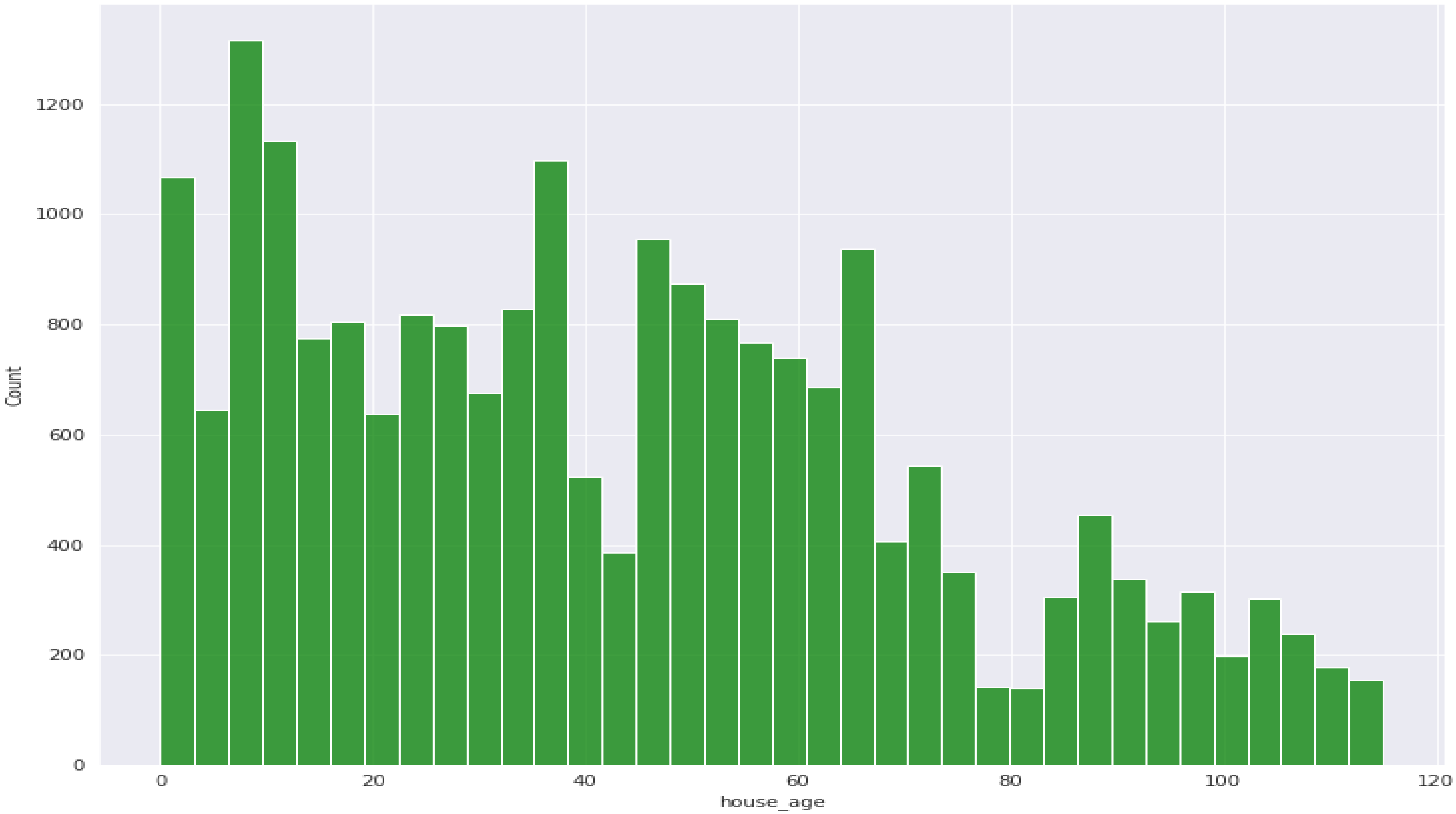
- Condition

  There was evidence that average condition homes sold at higher prices compared to good condition homes, which would go against the believed norms.


1e6
8
7
6
5
4
3
2
1
0
price
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
condition

- House age

  Newer homes have higher sale count compared to old homes with low sales count.



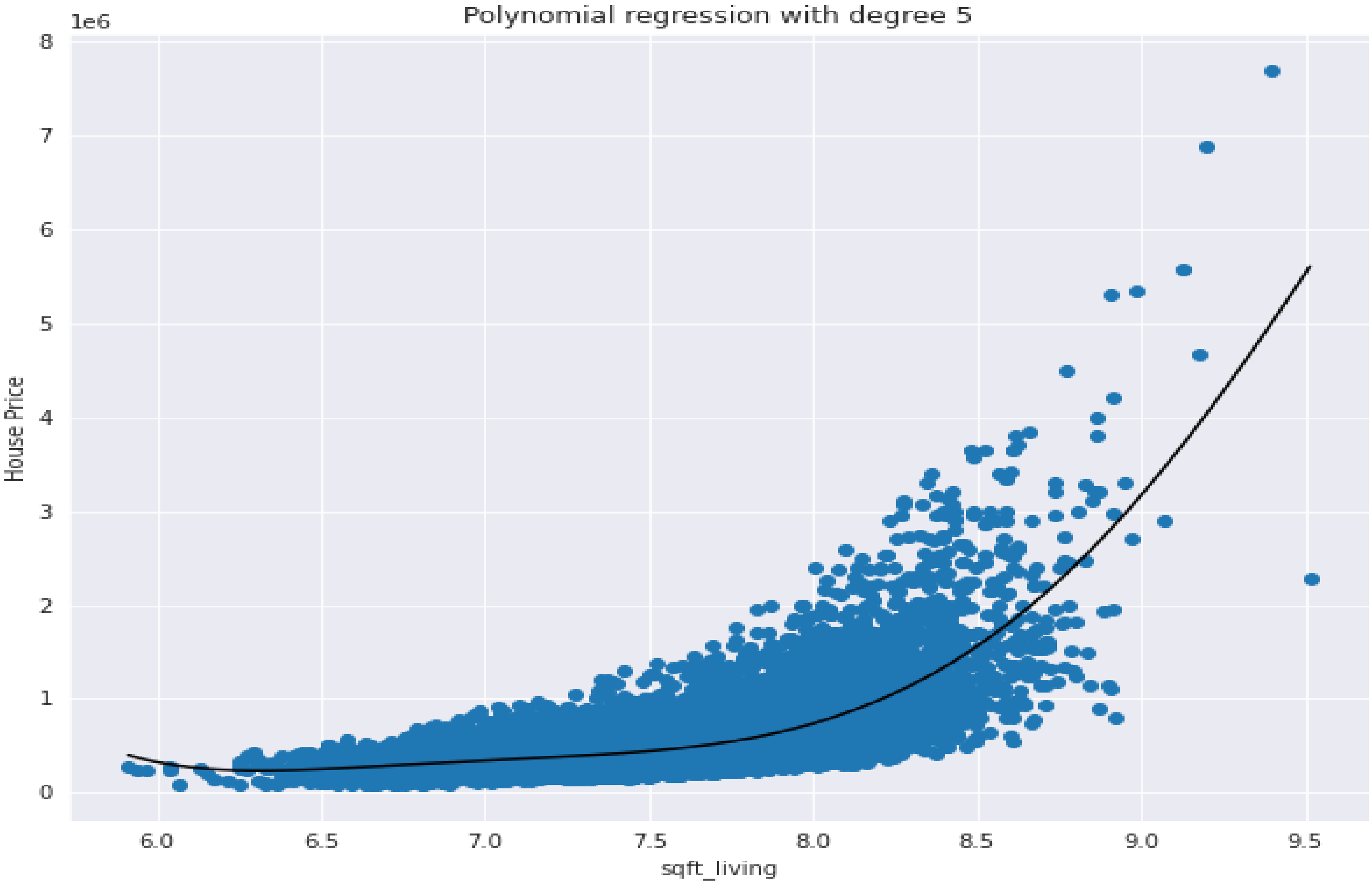
Count
0
200
400
600
800
1000
1200
0
20
40
60
80
100
120
house_age

Modeling

- Model 1

  I fitted a model of price ~ sqft_living^5

  It appeared that the relationship between price and sqft_living could be described by a polynomial relationship of order 5.


Polynomial regression with degree 5
1e6
8
7
6
5
4
3
2
1
0
House Price
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
sqft_living

- Final Model

  The best performing model however was one which utilised almost the full range of the variables made available to it, with an accuracy of about 61 percent, I know something to not loose sleep over… The model coefficients were as follows:

- (115945.80931477417, 'sqft_living'),
- (12545.092542462195, 'sqft_lot'),
- (51687.56909943093, 'waterfront'),
- (55629.24827197146, 'sqft_lot15')
- -15799.593453615063, 'one_two_bedrooms'),
- (141678.80167770002, 'ov_6_bedrooms'),
- (56700.013842632754, 'three_four_bedrooms'),
- (1029036.7756971495, 'four_five_bathrooms'),
- (-261915.8921985035, 'one_bathroom_below'),
- (-117480.47357278458, 'over_5_bathrooms'),
- (535617.1348683423, 'two_three_bathrooms'),
- (61185.57093468894, 'house_age'),

- (-164758.10856546738, 'average_quality_grade'),
- (60871.22502908802, 'poor_quality_grade'),
- (140237.29272080344, 'condition_1'),
- (1895.7740967030672, 'condition_2'),
- (94119.67043393666, 'condition_3'),
- (-152127.02855298377, 'condition_4'),
- (-54011.262696223486, 'one_floor'),
- (10789.5569907682, 'three_above_floors'),
- (363273.01157300675, 'two_floors')

- From the model, it was discovered that houses in Kings county were being underpriced by about -1208$

Recommendations

Despite the unreliability of the model, if these analysis is anything to go by, I would recommend that the stake holders focus on the following areas:

- Square foot size of the home
- Bigger size lot homes
- Waterfront homes
- Over 2 bed roomed homes
- 3 storied homes and above

These factors have a positive impact on the eventual pricing of the homes. Hence an investment on these factors will be beneficial to firms and home owners.

THE END

questions?