Multi Linear Regression for Predicting Housing Prices

Max Zimmerman

Who Is This Model For?

How can we use linear regression to help create the simplest model?

- How many variables should be used?

- How accurate does the model need to be?

- Should new variables be created?

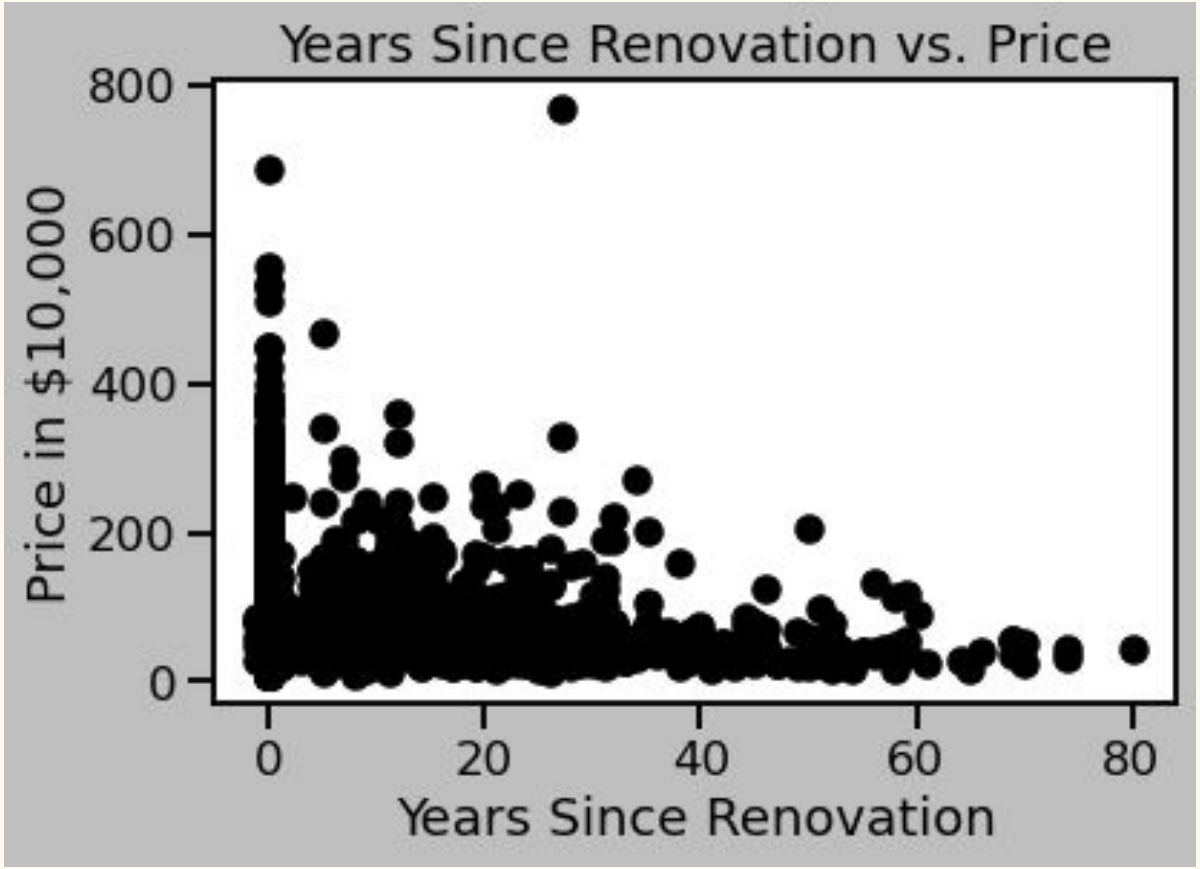
Engineered Variables:

- Years Since Renovation

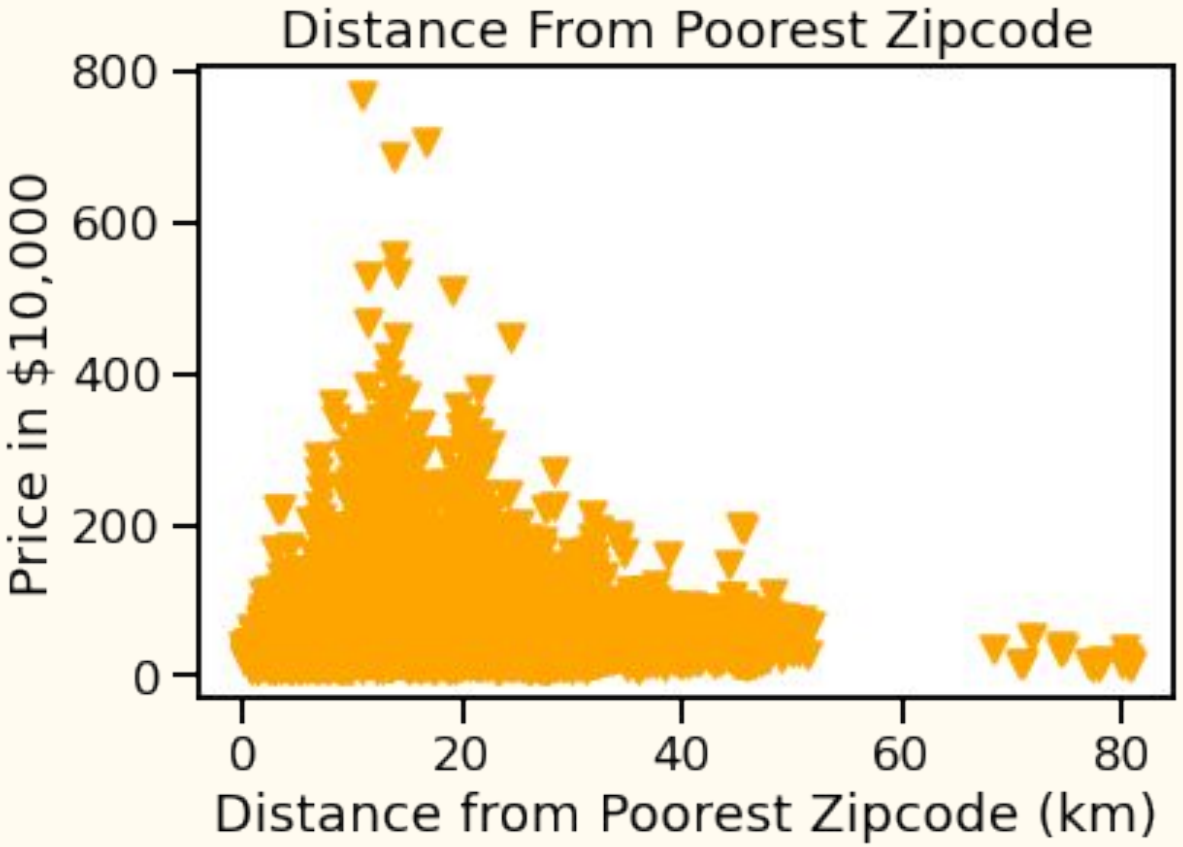
- Distance from Richest Zip Code

- Distance from Poorest Zip Code

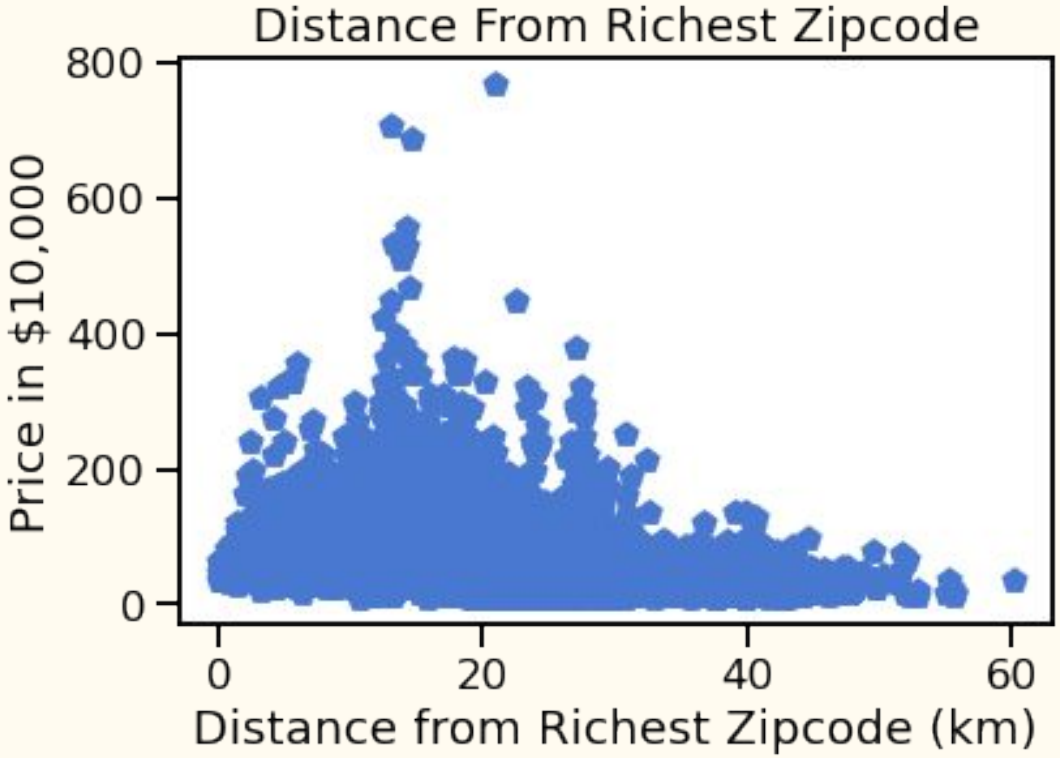
Year Since Renovation:



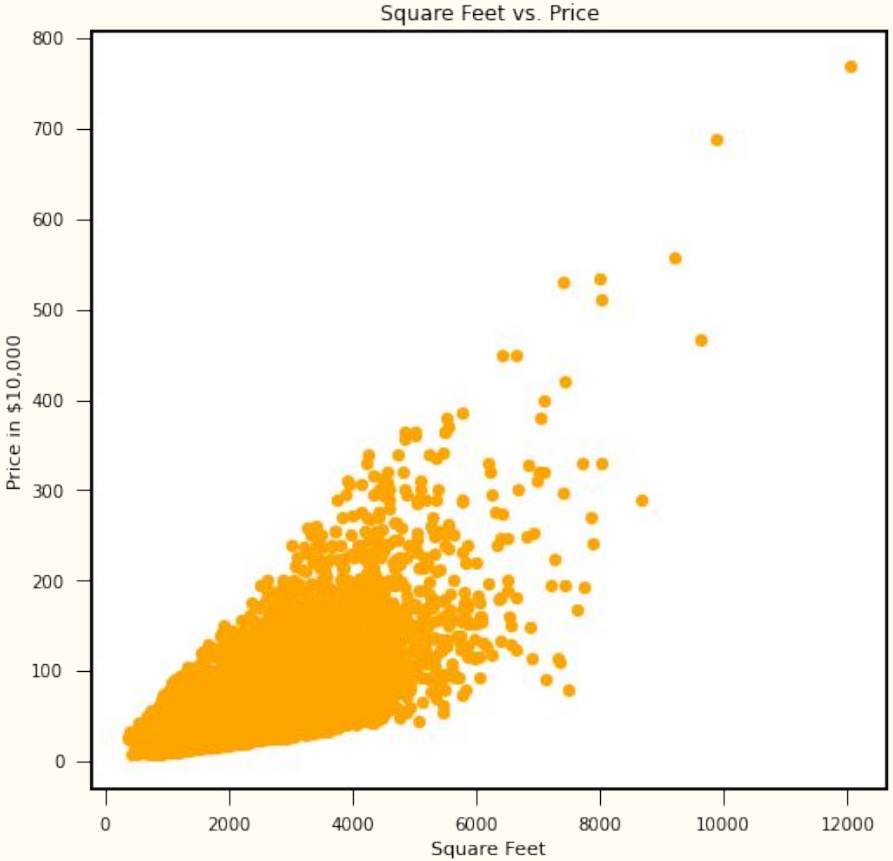
- Distance from Poorest Zip Code

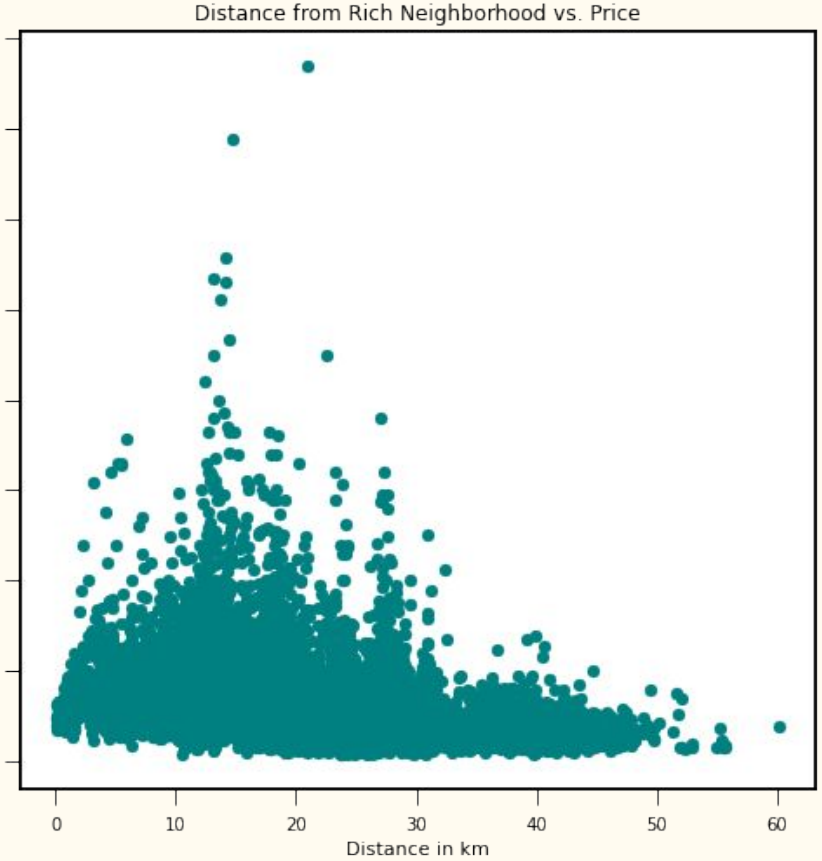


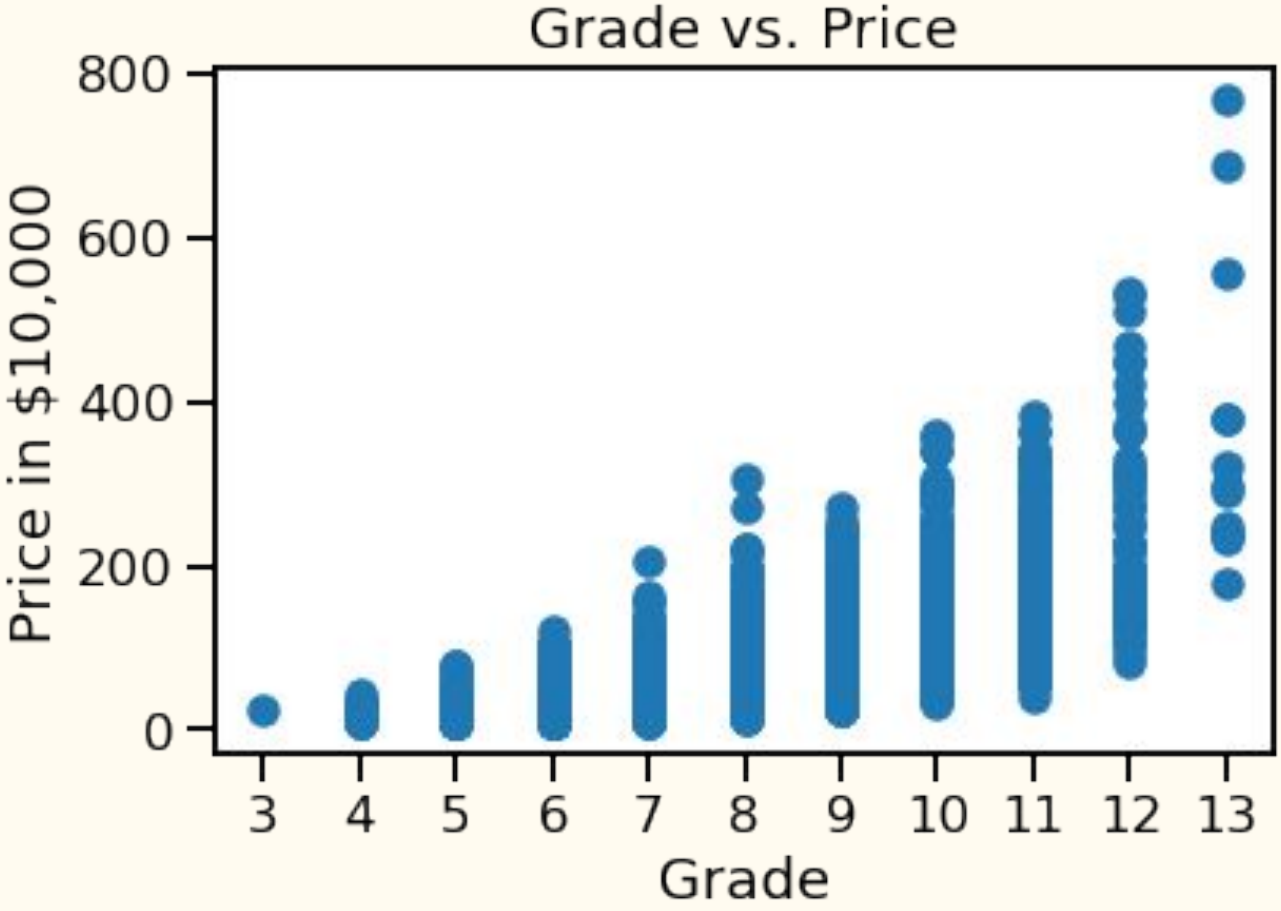
Distance From Richest Zip code



Chosen Variables:


Square Feet vs. Price
Price in $10,000
800
700
600
500
400
300
200
100
0
0
2000
4000
6000
8000
10000
12000
Square Feet


Distance from Rich Neighborhood vs. Price
0
10
20
30
40
50
60
Distance in km


Grade vs. Price
Price in $10,000
800
600
400
200
0
3
4
5
6
7
8
9
10
11
12
13
Grade

Explaining the Model:

Const = -369,700
Sqft_living = 175.21
Zip_distance_rich = -6,317
Grade = 89050

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.7941	0.022	530.597	0.000	11.750	11.838
sqft_living	0.0002	4.28e-06	48.060	0.000	0.000	0.000
zip_distance_rich	-0.0160	0.000	-58.788	0.000	-0.017	-0.015
grade	0.1532	0.003	45.618	0.000	0.147	0.160

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.697e+05	1.67e+04	-22.088	0.000	-4.02e+05	-3.37e+05
sqft_living	175.2152	3.225	54.325	0.000	168.893	181.537
zip_distance_rich	-6316.9923	205.589	-30.726	0.000	-6719.970	-5914.015
grade	8.905e+04	2532.009	35.168	0.000	8.41e+04	9.4e+04

Example:

2000 Square Foot House

Grade: 7

2 Kilometers Away From Richest Zip code

$$-369{,}700 + 7(89050) + 175.21(2000) + 2(-6{,}317) =$$

$591436

Recommendations:

Improvements for the Future:

- Many waterfront data missing
- Minimal amounts of data for bigger houses with 6 or more bedrooms
- More information on location/school districts

Thank You!