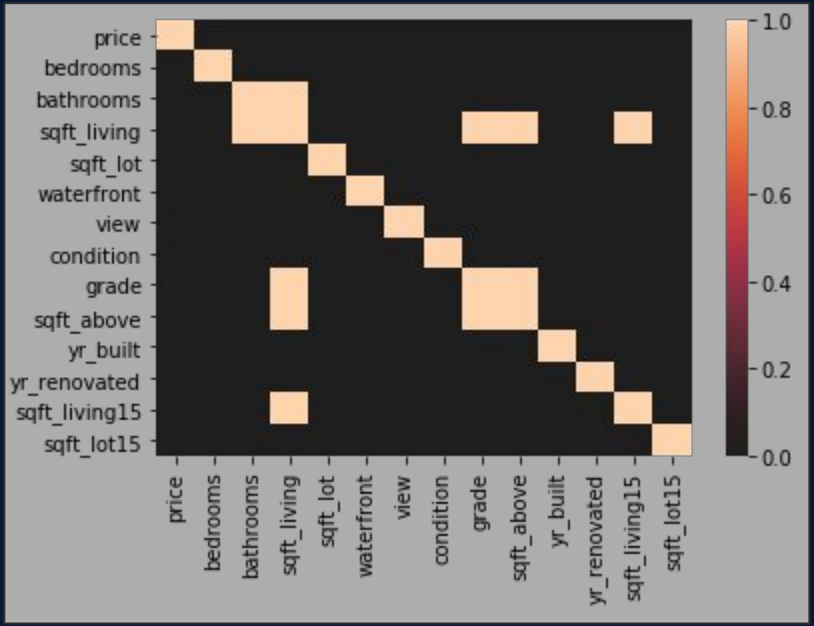
Hello, my name is Rob and this was the thought process of putting my model together, which I have named NostROBdamus

Rob Blatt - Flatiron School Data Science 042219 - Mod 1 Project

Decisions Were Made

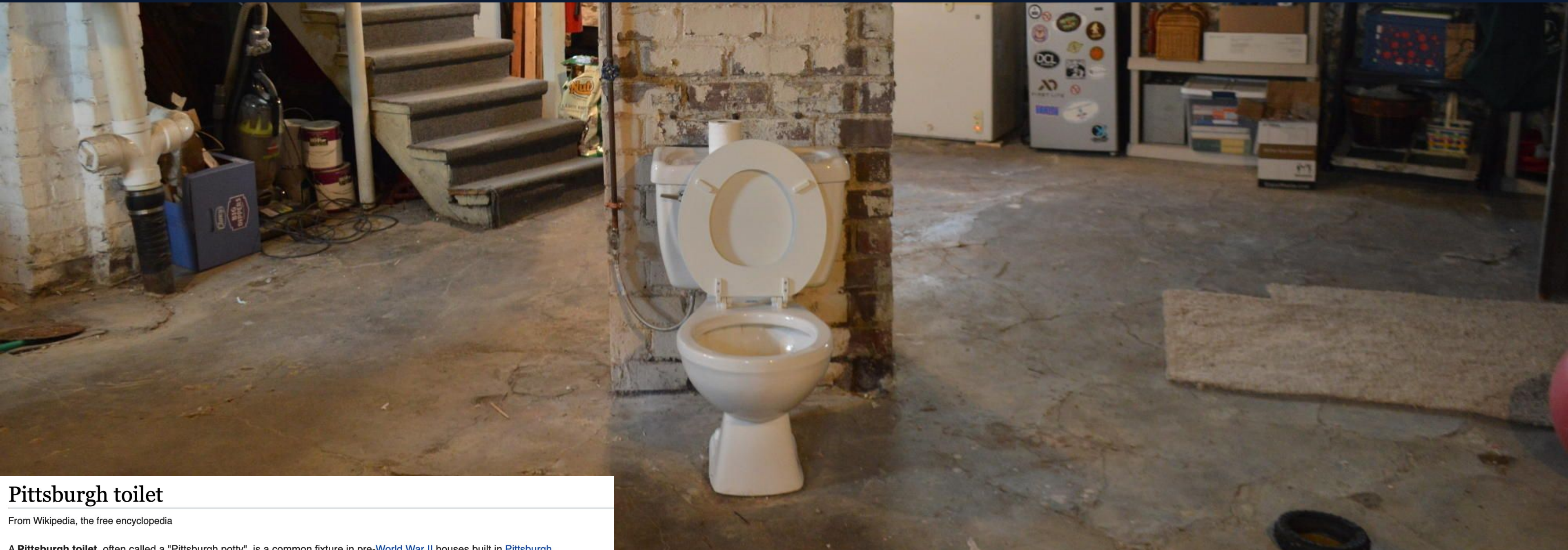
Right away I dropped the following:

- zipcode
- lat
- long
- floors
- id



sqft_living also looks untrustworthy

DOES THIS LOOK LIKE A REAL BATHROOM TO YOU?

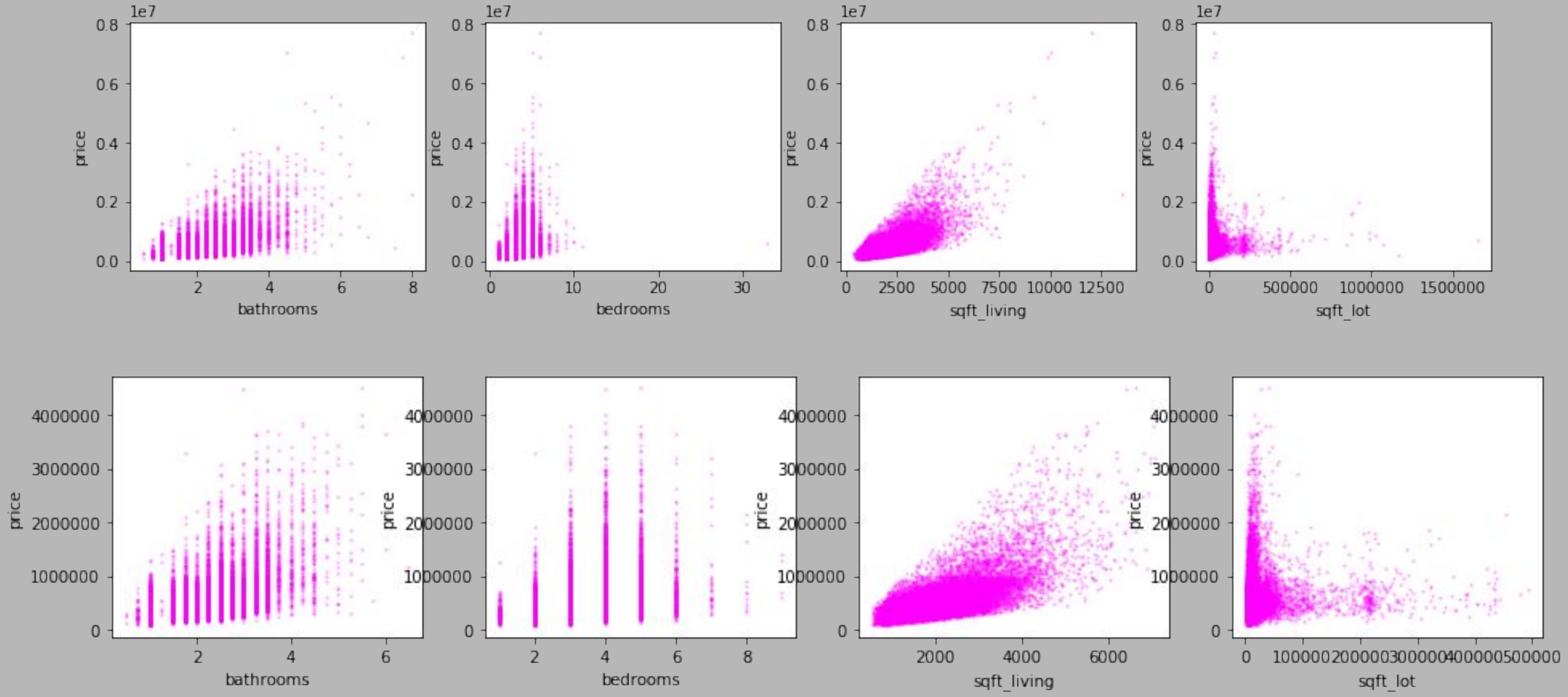


Pittsburgh toilet

From Wikipedia, the free encyclopedia

A **Pittsburgh toilet**, often called a "Pittsburgh potty", is a common fixture in pre-World War II houses built in Pittsburgh

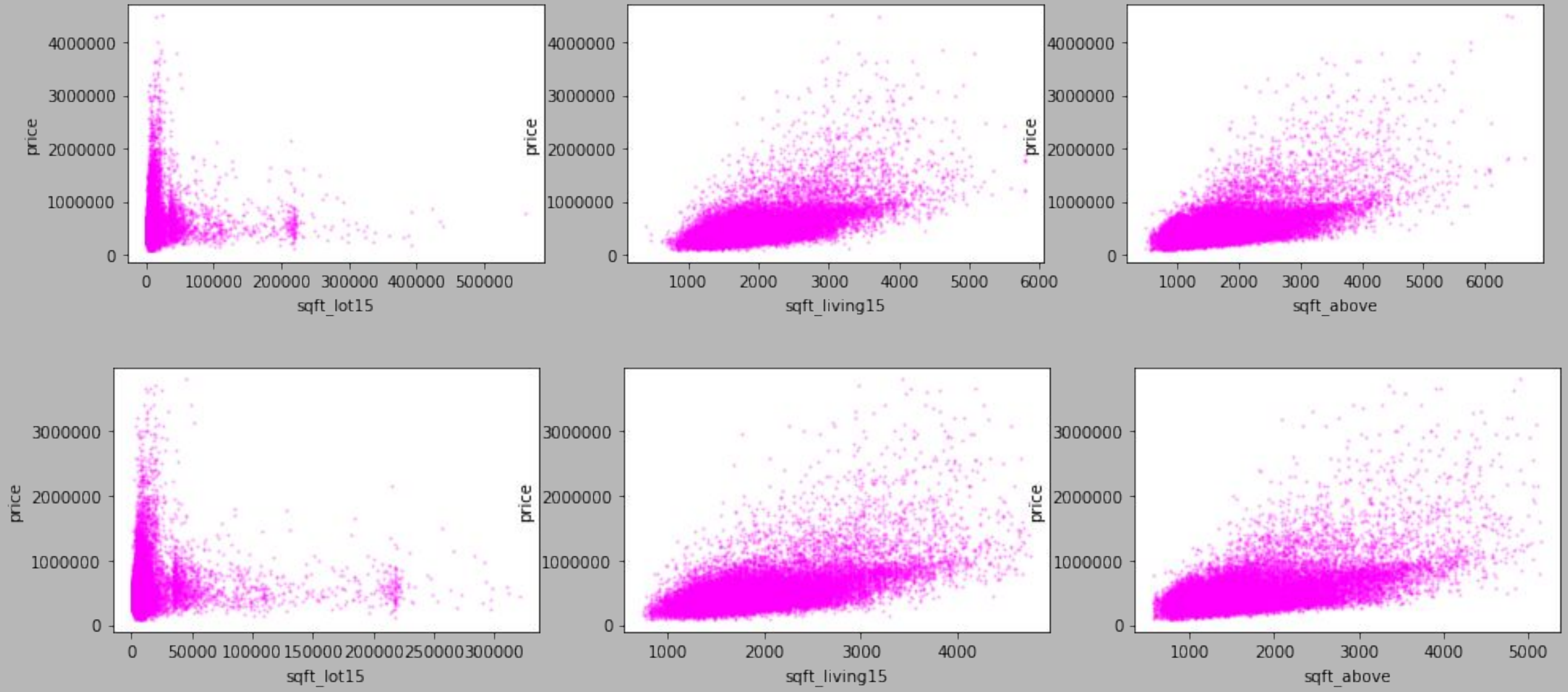
I removed the top and bottom 0.001%



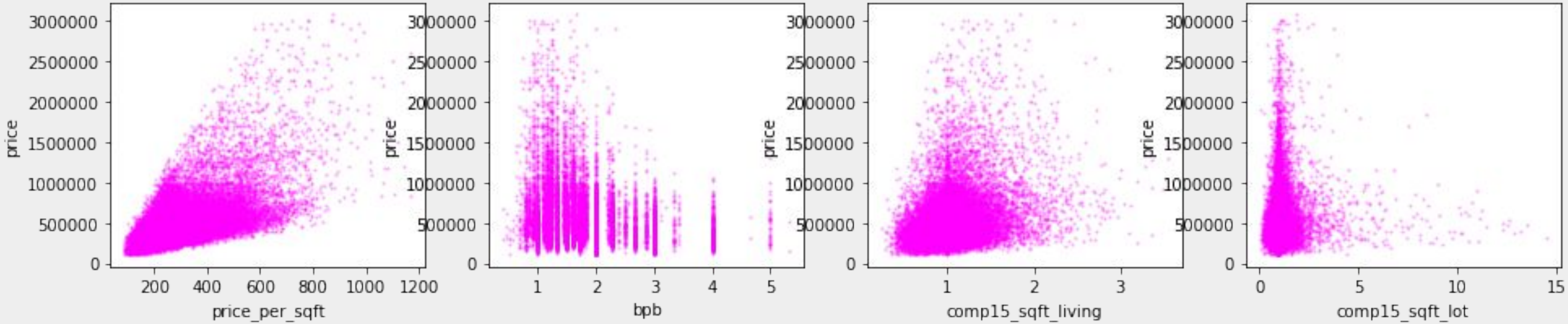
THIS IS REAL WORLD DATA

Homes don’t have 33 bedrooms or
4.5 bedrooms and 0.5 bathrooms

I removed the top and bottom 0.001%



Hi, my name is Rob and meet my new features:



price_per_sqft = Price Per Square Foot (above ground)
bnb = Bedrooms per Bathroom
comp15_sqft_living = sqft_living / sqft_living15
Comp15_sqft_lot = sqft_lot / sqft_lot

All the null values were in variables I would consider to be Boolean, so I set them all to zero.

I did a bunch of work with the year renovated data, creating “years since renovation” but ultimately I couldn’t make it helpful.

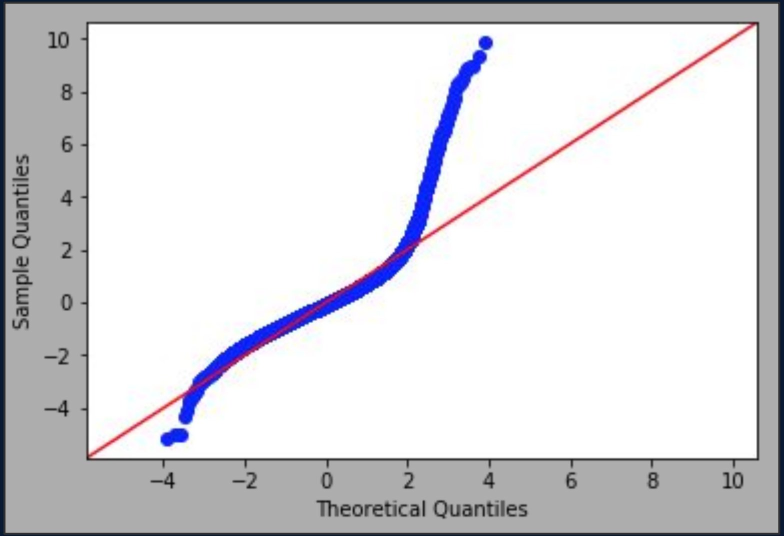
I thought I could make something cool and useful if I combined **view**, **condition**, **grade**, and **waterfront**.

I even gave it a name. “Shiny.”

It was not helpful.

R-squared of waterfront, view, condition, grade, price_per_sqft, & bpb was 0.733

Mean price	540,296
Waterfront == 1	1,717,214
View != 0	928,651
Condition > 4	612,577
Grade > 7	714,926



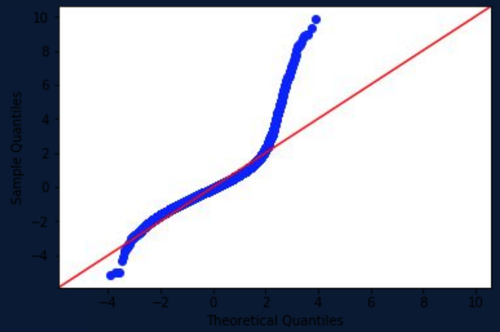
not thrilled with the QQ plot

A few of us were talking about the crazy idea of **log transforming the price** if the QQ plot was looking like an S.

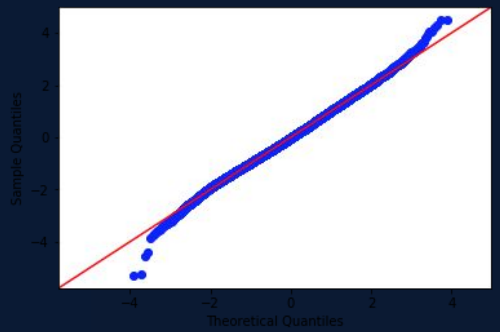
So I did!

I also **log transformed the price_per_sqft** because it was looking uneven.

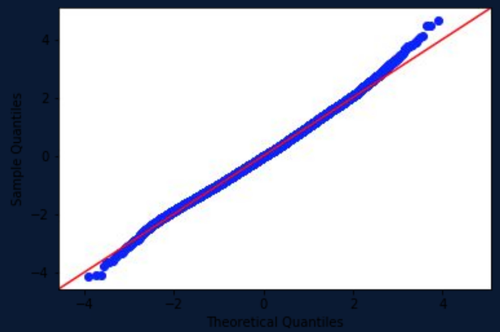
After the price log transforming,
removing waterfront did not change the R-squared



R-square: 0.733

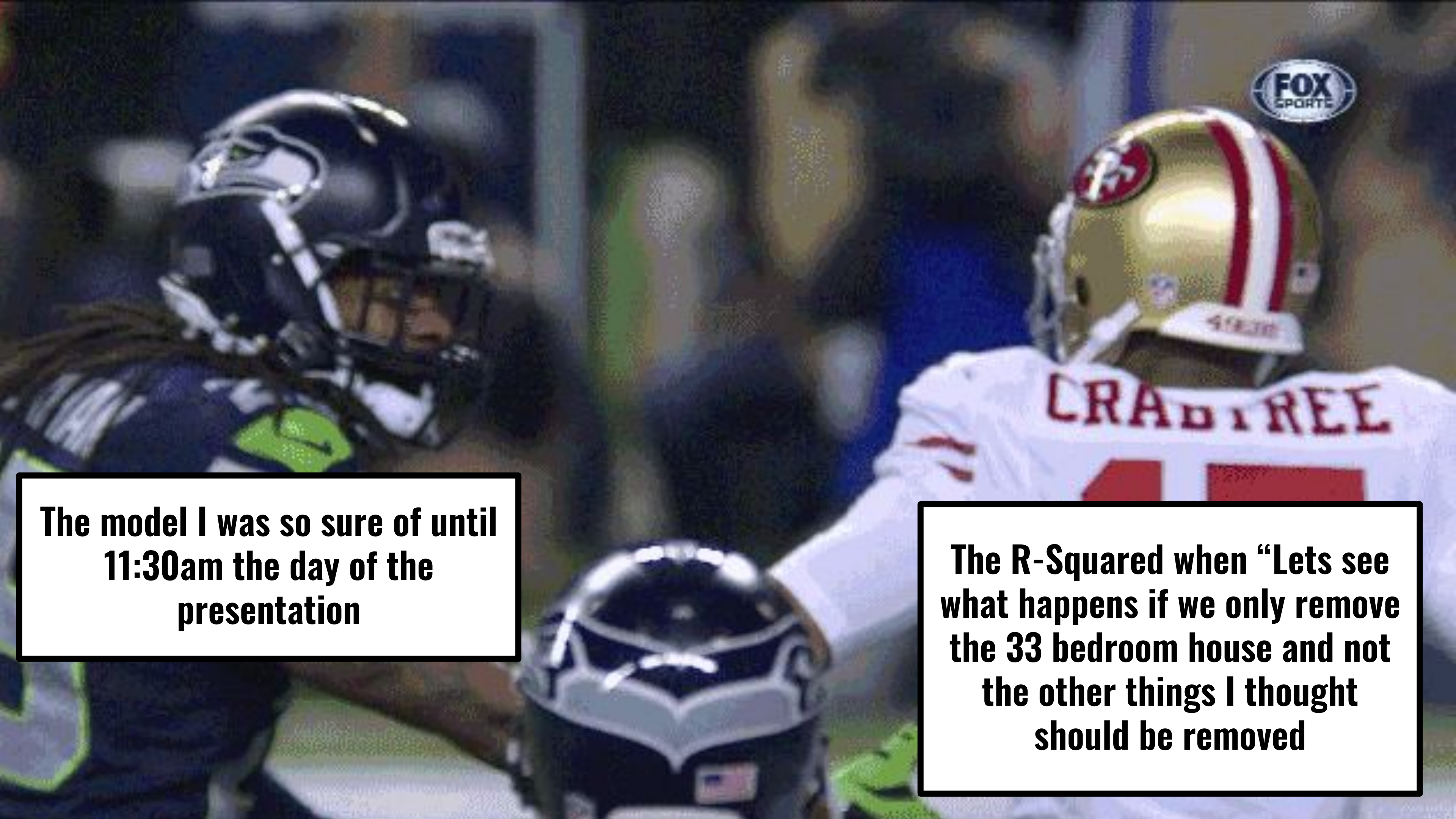


R-square: 0.785
Log transform: price



R-square: 0.785
Log transform:
price_per_sqft


OXI
CLEAN


The model I was so sure of until 11:30am the day of the presentation
The R-Squared when “Lets see what happens if we only remove the 33 bedroom house and not the other things I thought should be removed

Dep. Variable:	price	**R-squared:**	0.783
Model:	OLS	**Adj. R-squared:**	0.783
Method:	Least Squares	**F-statistic:**	1.535e+04
Date:	Wed, 08 May 2019	**Prob (F-statistic):**	0.00
Time:	09:42:08	**Log-Likelihood:**	454.51
No. Observations:	21269	**AIC:**	-897.0
Df Residuals:	21263	**BIC:**	-849.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3277	0.027	272.930	0.000	7.275	7.380
view	0.0322	0.002	13.986	0.000	0.028	0.037
condition	0.0757	0.009	8.133	0.000	0.057	0.094
grade	0.2745	0.002	168.405	0.000	0.271	0.278
price_per_sqft	0.6294	0.004	161.469	0.000	0.622	0.637
bpb	-0.0221	0.003	-7.993	0.000	-0.028	-0.017

Omnibus:	135.720	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	144.722
Skew:	0.169	**Prob(JB):**	3.75e-32
Kurtosis:	3.220	**Cond. No.**	165.

Removing "outliers"

Not removing "outliers"

Dep. Variable:	price	**R-squared:**	0.793
Model:	OLS	**Adj. R-squared:**	0.793
Method:	Least Squares	**F-statistic:**	1.659e+04
Date:	Wed, 08 May 2019	**Prob (F-statistic):**	0.00
Time:	11:16:04	**Log-Likelihood:**	245.38
No. Observations:	21595	**AIC:**	-478.8
Df Residuals:	21589	**BIC:**	-430.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.2723	0.027	274.101	0.000	7.220	7.324
view	0.0336	0.002	14.838	0.000	0.029	0.038
condition	0.0820	0.009	8.824	0.000	0.064	0.100
grade	0.2791	0.002	177.209	0.000	0.276	0.282
price_per_sqft	0.6304	0.004	162.249	0.000	0.623	0.638
bpb	-0.0187	0.003	-6.836	0.000	-0.024	-0.013

Omnibus:	142.332	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	162.723
Skew:	0.149	**Prob(JB):**	4.62e-36
Kurtosis:	3.304	**Cond. No.**	163.


HANG IN THERE
WE GOAT THIS!