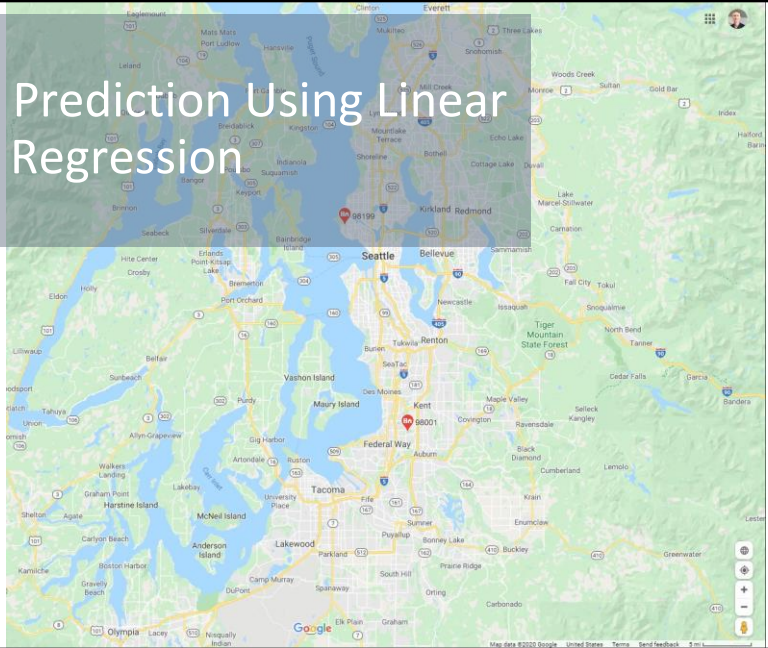
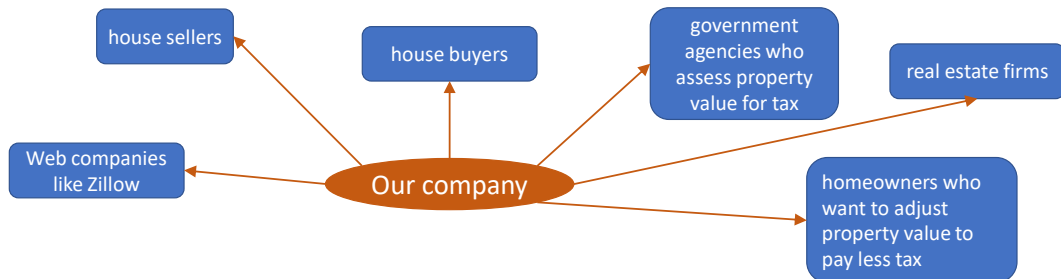


House Price Prediction Using Linear Regression

Pierre-Olivier Ariston
April 2020



Problem Statement and business value

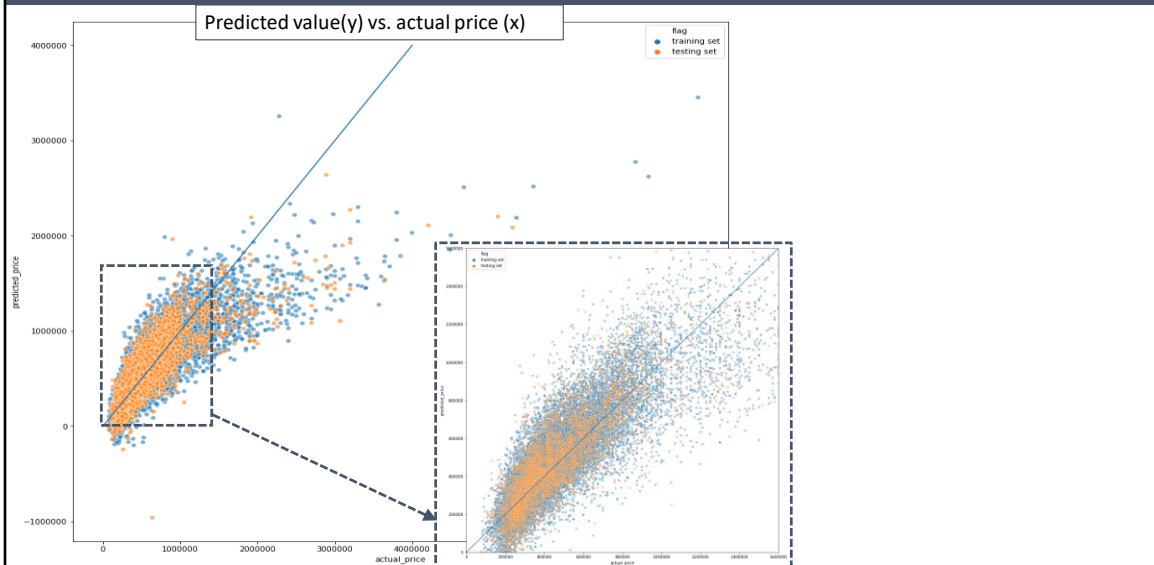


Value we bring: Expertise in prediction of house selling prices.
How we are different: **Superior accuracy** of our models

We are a small company trying to find our place in the housing market ecosystem. The **value** we bring is our expertise in house price prediction. Our **business model** is to sell our predictions to house sellers, house buyers, real estate firms, companies like Zillow, government agencies who assess property value for tax, homeowners who want to adjust property value to pay less tax

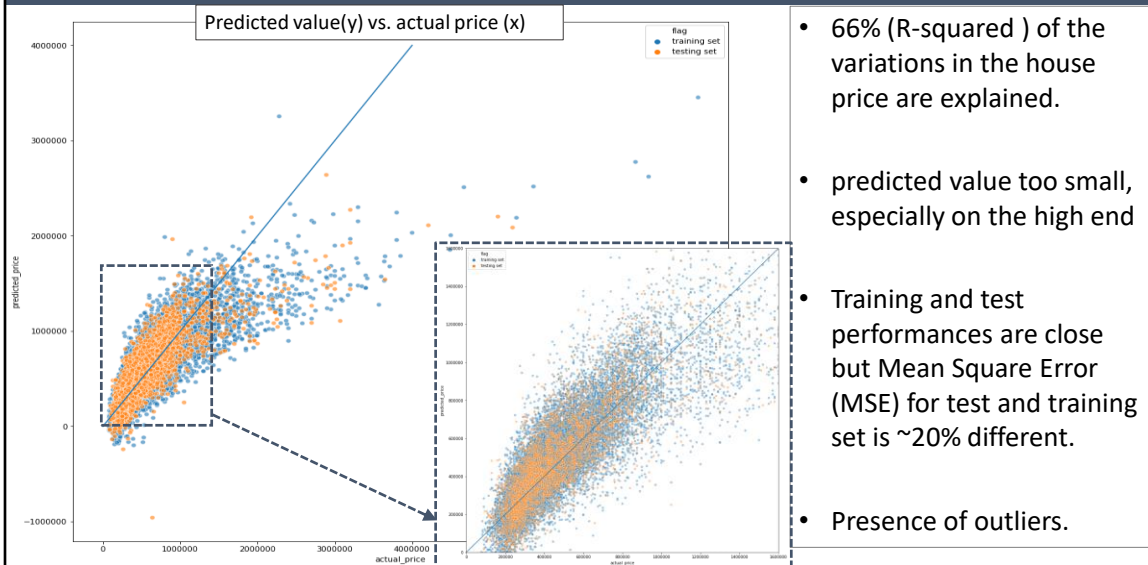
The **value** we bring is our expertise in house market value prediction. As a new company, we need a **very accurate model** that sets us apart from competition, establishes our reputation and makes us gain new clients.

Lessons learnt from our MVP (Minimum Viable Product)



scatter plot of training and test sets' predictions vs. actuals
Ideally, all points should be on the line $y=x$ displayed for reference

Lessons learnt from our MVP (Minimum Viable Product)



Problem for prediction of expensive houses: inaccurate predictions+ test performances very different from training differences.

=> Need to treat expensive houses separately.

Presence of outliers is evidenced by the prediction errors visualization. They are likely to be responsible for a good part of the inaccuracy of the model. Needs to be addressed in forthcoming work.

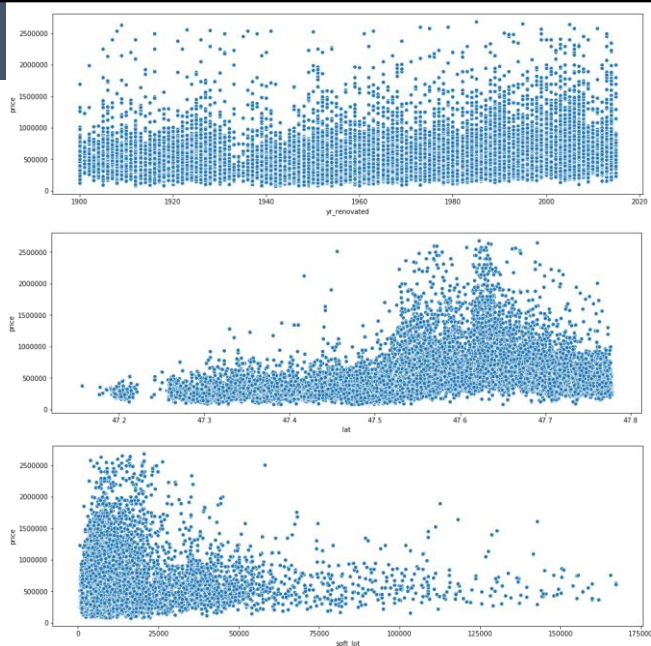
Phase 2 Workflow

Graphs of price (y) vs features **yr_renovated** (top), **lat** and **sqft_lot** (bottom).

Main issue:

Variations in price are much more complex than just linear or polynomial.

However, their correlation with price is significant: **yr_renovated**: 9%, **lat**: 35%, **sqft_lot**: 12%



At this point we are ready to move into fitting a multivariate linear model to the data, but we have issues with 3 features that contradict the assumptions of this process: **yr_renovated**, **lat**, **sqft_lot**.

These 3 predictors have correlation levels with the target price that are low to medium.

The main issue with all three of them is that the target price doesn't vary linearly with any of them.

What we propose is to create a model with and without them and to compare the performances of the 2 models

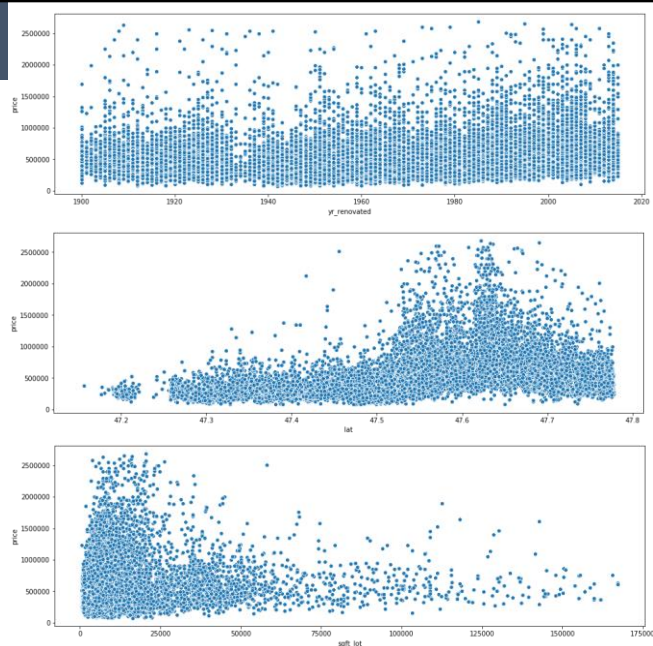
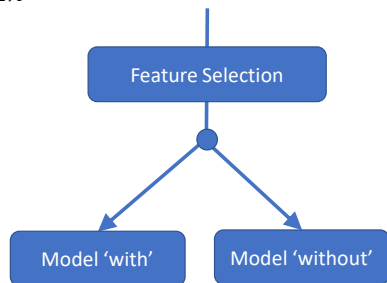
Phase 2 Workflow

Graphs of price (y) vs features **yr_renovated** (top), **lat** and **sqft_lot** (bottom).

Main issue:

Variations in price are much more complex than just linear or polynomial.

However, their correlation with price is significant: **yr_renovated**: 9%, **lat**: 35%, **sqft_lot**: 12%



At this point we are ready to move into fitting a multivariate linear model to the data, but we have issues with 3 features that contradict the assumptions of this process: **yr_renovated**, **lat**, **sqft_lot**.

These 3 predictors have correlation levels with the target price that are low to medium.

The main issue with all three of them is that the target price doesn't vary linearly with any of them.

What we propose is to create a model with and without them and to compare the performances of the 2 models

Dep. Variable: priceR-squared: 0.730

Model: OLSAdj. R-squared: 0.729

Method: Least SquaresF-statistic: 1263

Date: Sun, 19 Apr 2020Prob (F-statistic): 0.00

Time: 10:52:57Log-Likelihood: -1502.3

No. Observations: 16884AIC: 3259

DF Residuals: 16850BIC: 3545

DF Model: 34

Covariance Type: nonrobust

coefstd errtP>|t|[0.0250.975]

const-62.91780.015-4450.0000.000-64.916-61.319

sqft_tot-2.139e-071.52e-07-1.4000.159-5.19e-078.36e-08

sqft_above-0.00330.012-0.2670.790-0.1670.160

yr_built-0.00320.000-31.2790.000-0.003-0.003

lat1.36960.01686.9690.0001.3301.401

sqft_living_ratio15-0.23890.008-28.5880.000-0.255-0.223

sqft_living_ratio150.02440.01120.2840.0000.0040.048

bed_2-0.00580.003-2.4180.016-0.1010.091

bed_3-0.15300.023-6.6060.000-0.198-0.108

bed_4-0.19060.024-8.0120.000-0.237-0.144

bed_5-0.22400.025-8.9680.000-0.273-0.175

bed_6-0.27310.031-8.8430.000-0.333-0.213

bed_7-0.31840.060-5.3200.000-0.430-0.201

bath_20.33470.1582.1230.0340.0260.044

bath_100.20080.1541.3240.188-0.0210.583

bath_1250.349160.1802.0530.0440.0100.771

bath_150.26270.1541.7060.088-0.0360.560

bath_1750.28800.1541.8800.064-0.0170.587

bath_200.30960.1542.0090.0450.0060.612

bath_2450.26200.1541.6910.087-0.0090.536

bath_240.28460.1541.8520.068-0.0070.587

bath_2750.30110.1542.2760.0230.0400.563

bath_300.36840.1542.3730.0180.0640.680

bath_3200.44050.1552.8810.0040.1420.748

bath_340.40480.1582.5420.0110.1020.706

bath_3750.46840.1583.0170.0020.2430.695

bath_400.51870.1573.2940.0010.2090.823

bath_4250.62750.1593.9500.0000.3160.939

bath_450.46700.1583.0020.0020.1790.756

bath_4750.61270.1543.9710.0000.2720.954

bath_500.68330.1723.9470.0000.3251.001

bath_5250.59840.1953.0710.0020.2170.982

floor_100.01540.0080.1730.8690.0040.026

floor_200.00820.0070.1410.8890.0010.016

floor_250.10950.0254.3540.0000.0590.160

floor_300.16490.01411.9130.0000.1360.193

waterf_100.58840.02820.9210.0000.5330.644

Crosscheck: 202.654Durbin-Watson: 2.007

Prob(Omnibus): 0.000Jarque-Bera (JB): 506.467

Skew: 0.119Prob(JB): 1.50e-104

Kurtosis: 3.002Cond. No.: 7.42e+06

Phase 2 Results. Models
'with' | 'without'

• Adding the 3 questionable features enabled to model much more of the variations in house prices.

Dep. Variable: priceR-squared: 0.973

Model: OLSAdj. R-squared: 0.971

Method: Least SquaresF-statistic: 655.3

Date: Sun, 19 Apr 2020Prob (F-statistic): 0.00

Time: 10:52:57Log-Likelihood: -564.5

No. Observations: 16884AIC: 1.104e+04

DF Residuals: 16850BIC: 1.130e+04

DF Model: 33

Covariance Type: nonrobust

coefstd errtP>|t|[0.0250.975]

const-6.46070.15820.4720.6380.5680.513

sqft_tot-0.28840.015-19.2480.000-0.314-0.256

sqft_living_ratio15-1.25840.010-24.7440.000-0.277-0.236

sqft_living_ratio150.02840.0130.89320.3680.0041.095

bed_3-0.00890.012-0.71270.4780.001-0.0150.013

bed_3-0.26440.029-9.2700.000-0.320-0.203

bed_4-0.20750.020-10.1490.000-0.355-0.240

bed_5-0.32510.031-10.5510.000-0.386-0.265

bed_6-0.32600.038-8.6000.000-0.404-0.254

bed_7-0.29520.069-4.2910.000-0.430-0.160

bath_20.46330.1732.0780.0070.1240.802

bath_100.30950.1682.3260.0200.0610.702

bath_1250.36250.2051.9110.066-0.0100.739

bath_150.21520.1581.3750.081-0.0140.646

bath_1750.23820.1681.9540.061-0.0010.687

bath_200.33190.1681.9780.0480.0030.661

bath_2250.29570.1681.7590.079-0.0340.625

bath_250.22450.1681.4350.151-0.0880.571

bath_2750.32530.1681.9320.053-0.0060.646

bath_300.39630.1682.3130.0250.0260.867

bath_3250.42300.1682.5100.0120.0030.756

bath_350.33980.1682.3500.0190.0060.727

bath_3750.40540.1712.3420.0190.0110.700

bath_400.44500.1722.5630.0080.1180.779

bath_4250.59150.1743.3920.0010.2500.933

bath_450.46770.1732.6970.0070.1260.808

bath_4750.46870.1972.4720.0130.1010.873

bath_500.46250.1682.3420.0190.0010.871

bath_5250.59640.2172.6060.0090.1400.892

floor_100.23210.01023.8230.0000.2130.251

floor_20.04450.0080.5140.0000.0260.061

floor_250.20960.0219.5330.0000.2360.938

floor_300.24480.01714.7330.0000.2140.280

waterf_100.54210.03515.6010.0000.4740.610

Crosscheck: 1.514Durbin-Watson: 1.985

Prob(Omnibus): 0.445Jarque-Bera (JB): 1.537

Skew: 0.018Prob(JB): 0.841

Kurtosis: 2.968Cond. No.: 3.56e+03

Phase 2 Results. Models 'with' | 'without'

- Adding the 3 questionable features enabled to model much more of the variations in house prices.

Interpretation of statsmodel diagnostics with and without the 3 questionable features:

1. With: R-squared adjusted = 73%. It means that 73% of the variations of the target variable price are predicted by the model.

Without: R-squared adjusted is much smaller: 57%

This is consistent with F statistic being much higher with (1263) than without (685). Adding the 3 features enable to model more of the variations of the target variable.

2. From the coefficients report (with), we can tell that sqft_lot is actually not a statistically significant predictor. Its p-value is 16%.

3. With: JB test value is high (506) and the p-value is extremely close to zero. The Null hypothesis for the JB test is that the residuals are normally distributed. This hypothesis should be rejected. One other assumption for linear regression is not true !

Without : JB test value is much lower and the pvalue is 44%. which means that the Null hypothesis can not be rejected. We can be confident that the residuals are normally distributed!

This is consistent with the Skewness and Kurtosis values found for the residuals: they

are much closer to the ideal values of a normal distribution(0 and 3 respectively) for the model without than the model with.

This is also confirmed by the Omnibus test's probability

4. Some issues with the categorical bathroom variable for both models. P-values >5%. Need to change bathroom variable..25 bathroom doesn't make sense.

Dep. Variable: price

R-squared: 0.730

Model: OLS

Adj. R-squared: 0.728

Method: Least Squares

F-statistic: 1263

Date: Sun, 19 Apr 2020

Prob (F-statistic): 0.00

Time: 10:52:57

Log-Likelihood: -1502.3

No. Observations: 16884

AIC: 3278

DF Residuals: 16867

BIC: 3545

DF Model: 30

Covariance Type: nonconstant

coef

std err

t

Prob

[0.025

0.975]

const

-62.9178

0.015

-4458.0

0.000

-64.916

-61.319

sqft_lot

-2.139e-07

1.52e-07

-1.400

0.159

-5.19e-07

8.36e-08

sqft_above

0.0033

0.012

0.287

0.820

-0.017

0.040

yr_built

-0.0032

0.000

-31.279

0.000

-0.003

-0.003

lat

1.3696

0.016

86.969

0.000

1.330

1.401

sqft_living_ratio15

0.2389

0.008

28.588

0.000

0.225

0.253

sqft_living_ratio15

0.0244

0.011

2.124

0.034

0.004

0.045

bed_2

-0.0058

0.003

-2.418

0.016

-0.010

-0.001

bed_3

-0.1530

0.023

-6.606

0.000

-0.198

-0.108

bed_4

-0.1906

0.024

-8.012

0.000

-0.237

-0.144

bed_5

-0.2240

0.025

-8.968

0.000

-0.273

-0.175

bed_6

-0.2731

0.031

-8.840

0.000

-0.333

-0.213

bed_7

-0.3184

0.060

-5.320

0.000

-0.430

-0.201

bed_8

0.3347

0.158

2.123

0.034

0.020

0.644

bath_1

0.2088

0.154

1.324

0.188

-0.021

0.563

bath_1_25

0.3818

0.186

2.053

0.044

0.010

0.753

bath_1_5

0.2527

0.154

1.638

0.098

-0.039

0.549

bath_1_75

0.2880

0.154

1.880

0.064

-0.017

0.587

bath_2_0

0.3086

0.154

2.009

0.045

0.008

0.612

bath_2_25

0.2820

0.154

1.801

0.077

-0.009

0.580

bath_2_5

0.2846

0.154

1.852

0.068

-0.007

0.587

bath_2_75

0.3011

0.154

2.276

0.023

0.040

0.563

bath_3_0

0.3684

0.154

2.373

0.018

0.064

0.680

bath_3_25

0.4405

0.155

2.881

0.004

0.142

0.748

bath_3_5

0.4048

0.156

2.642

0.009

0.102

0.708

bath_3_75

0.5484

0.156

3.517

0.000

0.243

0.856

bath_4_0

0.5187

0.157

3.294

0.001

0.209

0.823

bath_4_25

0.6275

0.159

3.950

0.000

0.316

0.939

bath_4_5

0.4670

0.158

2.962

0.002

0.170

0.764

bath_4_75

0.6127

0.154

3.981

0.000

0.272

0.954

bath_5_0

0.6833

0.172

3.947

0.000

0.325

1.001

bath_5_25

0.5984

0.195

3.071

0.002

0.217

0.982

floor_1_0

0.0154

0.008

0.173

0.869

0.009

0.022

floor_2_0

0.0082

0.007

0.141

0.889

0.000

0.016

floor_2_5

0.1095

0.025

4.358

0.000

0.059

0.200

floor_3_0

0.1649

0.014

11.913

0.000

0.136

0.192

waterf_1_0

0.5884

0.028

20.921

0.000

0.533

0.644

Prob(Omnibus): 0.000

Durbin-Watson: 2.007

Skew: 0.119

Jarque-Bera LRB: 506.467

Kurtosis: 3.000

Prob(LRB): 0.49e-105

Cond. No.: 7.42e+05

Dep. Variable: price

R-squared: 0.570

Model: OLS

Adj. R-squared: 0.570

Method: Least Squares

F-statistic: 685.1

Date: Sun, 19 Apr 2020

Prob (F-statistic): 0.00

Time: 10:52:57

Log-Likelihood: -5454.5

No. Observations: 16884

AIC: 1.104e+04

DF Residuals: 16867

BIC: 1.130e+04

DF Model: 33

Covariance Type: nonconstant

coef

std err

t

Prob

[0.025

0.975]

const

5.4687

0.188

29.472

0.000

5.088

5.843

sqft_above

-0.2884

0.015

-19.248

0.000

-0.314

-0.256

sqft_living_ratio15

0.2584

0.010

24.744

0.000

0.237

0.280

sqft_living_ratio15

0.0263

0.013

2.052

0.044

0.004

0.048

bed_2

-0.0089

0.003

-3.127

0.001

-0.015

-0.003

bed_3

-0.2644

0.029

-9.270

0.000

-0.320

-0.208

bed_4

-0.2975

0.020

-10.149

0.000

-0.365

-0.240

bed_5

-0.3260

0.031

-10.551

0.000

-0.386

-0.265

bed_6

-0.3292

0.038

-8.600

0.000

-0.404

-0.254

bed_7

-0.2952

0.069

-4.291

0.000

-0.430

-0.160

bath_1

0.4633

0.173

2.678

0.007

0.124

0.802

bath_1_25

0.3985

0.188

2.128

0.032

0.061

0.729

bath_1_5

0.2162

0.181

1.198

0.231

-0.144

0.646

bath_1_75

0.3282

0.188

1.744

0.081

-0.041

0.697

bath_2_0

0.3319

0.188

1.759

0.084

-0.063

0.681

bath_2_25

0.2957

0.188

1.579

0.079

-0.034

0.625

bath_2_5

0.2842

0.188

1.493

0.151

-0.088

0.571

bath_2_75

0.3253

0.188

1.732

0.083

-0.036

0.616

bath_3_0

0.3963

0.188

2.113

0.035

0.026

0.867

bath_3_25

0.4730

0.188

2.510

0.012

0.083

0.758

bath_3_5

0.3988

0.188

2.100

0.039

0.006

0.727

bath_3_75

0.5054

0.192

2.642

0.009

0.119

0.791

bath_4_0

0.4905

0.172

2.853

0.003

0.150

0.830

bath_4_25

0.5815

0.174

3.362

0.001

0.250

0.913

bath_4_5

0.4677

0.173

2.697

0.007

0.126

0.808

bath_4_75

0.4687

0.187

2.472

0.013

0.101

0.873

bath_5_0

0.4828

0.188

2.542

0.016

0.091

0.874

bath_5_25

0.5864

0.217

2.689

0.009

0.140

0.982

floor_1_0

0.2321

0.101

2.303

0.023

0.033

0.231

floor_2_0

0.0405

0.031

1.314

0.000

0.026

0.061

floor_2_5

0.2046

0.021

9.533

0.000

0.236

0.358

floor_3_0

0.2448

0.017

14.713

0.000

0.214

0.276

waterf_1_0

0.5421

0.035

15.601

0.000

0.474

0.610

Prob(Omnibus): 0.000

Durbin-Watson: 1.985

Skew: 0.018

Jarque-Bera LRB: 1.837

Kurtosis: 2.988

Prob(LRB): 0.641

Cond. No.: 3.56e+03

Phase 2 Results. Models 'with' | 'without'

• Adding the 3 questionable features enabled to model much more of the variations in house prices.

• Normality of residuals: That is a second assumption for linear regression which is not true for the model 'with'! The other one is the linear variation of price.

- Adding the 3 questionable features enabled to model much more of the variations in house prices.
- Normality of residuals: That is a second assumption for linear regression which is not true for the model 'with'! The other one is the linear variation of price.

Interpretation of statsmodel diagnostics with and without the 3 questionable features:

1. With: R-squared adjusted = 73%. It means that 73% of the variations of the target variable price are predicted by the model.

Without: R-squared adjusted is much smaller: 57%

This is consistent with F statistic being much higher with (1263) than without (685). Adding the 3 features enable to model more of the variations of the target variable.

2. From the coefficients report (with), we can tell that sqft_lot is actually not a statistically significant predictor. Its p-value is 16%.

3. With: JB test value is high (506) and the p-value is extremely close to zero. The Null hypothesis for the JB test is that the residuals are normally distributed. This hypothesis should be rejected. One other assumption for linear regression is not true !

Without : JB test value is much lower and the pvalue is 44%. which means that the Null hypothesis can not be rejected. We can be confident that the residuals are normally distributed!

This is consistent with the Skewness and Kurtosis values found for the residuals: they

are much closer to the ideal values of a normal distribution(0 and 3 respectively) for the model without than the model with.

This is also confirmed by the Omnibus test's probability

4. Some issues with the categorical bathroom variable for both models. P-values >5%. Need to change bathroom variable..25 bathroom doesn't make sense.

Dep. Variable: price **R-squared:** 0.730

Model: OLS **Adj. R-squared:** 0.728

Method: Least Squares **F-statistic:** 1263.1

Date: Sun, 19 Apr 2020 **Prob (F-statistic):** 0.00

Time: 10:52:57 **Log-Likelihood:** -1582.3

No. Observations: 16884 **AIC:** 3258

DF Residuals: 16850 **BIC:** 3245

DF Model: 33

Covariance Type: nonconstant

	coef	std err	t	Prob	[0.025	0.975]
const	62.9178	0.015	4458.0	0.000	-64.516	61.319
sqft_lot	-2.138e-07	1.52e-07	-1.408	0.159	6.19e-07	8.36e-06
sqft_above	0.0023	0.012	0.200	0.847	-0.020	0.025
yr_built	-0.0032	0.000	-31.279	0.000	-0.003	-0.003
lat	1.3696	0.016	86.969	0.000	1.339	1.401
sqft_basement	-0.2389	0.008	-28.588	0.000	-0.255	-0.223
sqft_hvz	0.0244	0.011	20.284	0.000	0.004	0.048
sqft_hvz2	-0.0058	0.003	-2.418	0.016	-0.101	0.091
bed_1	-0.1530	0.023	-6.606	0.000	-0.198	-0.108
bed_2	-0.1896	0.024	-8.012	0.000	-0.237	-0.144
bed_3	-0.2240	0.025	-8.968	0.000	-0.273	-0.175
bed_4	-0.2731	0.031	-8.840	0.000	-0.333	-0.213
bed_5	-0.3184	0.060	-5.326	0.000	-0.436	-0.201
bed_6	0.3347	0.158	2.123	0.034	0.026	0.644
bed_7	0.2088	0.154	1.324	0.188	-0.021	0.563
bed_8	0.3493	0.186	1.879	0.064	0.000	0.698
bed_9	0.2627	0.154	1.706	0.088	-0.039	0.569
bed_10	0.2080	0.154	1.368	0.174	-0.097	0.507
bed_11	0.3086	0.154	2.000	0.046	0.000	0.612
bed_12	0.2620	0.154	1.697	0.097	-0.050	0.569
bed_13	0.2646	0.154	1.712	0.088	-0.057	0.587
bed_14	0.3011	0.154	1.952	0.053	0.000	0.603
bed_15	0.3684	0.154	2.373	0.018	0.064	0.680
bed_16	0.4405	0.155	2.861	0.004	0.142	0.748
bed_17	0.4048	0.156	2.642	0.009	0.102	0.708
bed_18	0.5484	0.156	3.517	0.000	0.243	0.855
bed_19	0.5187	0.157	3.294	0.001	0.209	0.823
bed_20	0.6275	0.159	3.950	0.000	0.316	0.939
bed_21	0.4670	0.156	2.982	0.002	0.176	0.758
bed_22	0.6127	0.154	3.971	0.000	0.302	0.924
bed_23	0.6833	0.172	3.947	0.000	0.325	1.001
bed_24	0.5984	0.195	3.071	0.002	0.217	0.982
bed_25	0.6754	0.198	3.413	0.000	0.284	1.066
bed_26	0.6882	0.197	3.493	0.000	0.296	1.080
bed_27	0.6885	0.195	3.530	0.000	0.296	1.080
bed_28	0.6885	0.195	3.530	0.000	0.296	1.080
bed_29	0.6885	0.195	3.530	0.000	0.296	1.080
bed_30	0.6885	0.195	3.530	0.000	0.296	1.080
bed_31	0.6885	0.195	3.530	0.000	0.296	1.080
bed_32	0.6885	0.195	3.530	0.000	0.296	1.080
bed_33	0.6885	0.195	3.530	0.000	0.296	1.080
bed_34	0.6885	0.195	3.530	0.000	0.296	1.080
bed_35	0.6885	0.195	3.530	0.000	0.296	1.080
bed_36	0.6885	0.195	3.530	0.000	0.296	1.080
bed_37	0.6885	0.195	3.530	0.000	0.296	1.080
bed_38	0.6885	0.195	3.530	0.000	0.296	1.080
bed_39	0.6885	0.195	3.530	0.000	0.296	1.080
bed_40	0.6885	0.195	3.530	0.000	0.296	1.080
bed_41	0.6885	0.195	3.530	0.000	0.296	1.080
bed_42	0.6885	0.195	3.530	0.000	0.296	1.080
bed_43	0.6885	0.195	3.530	0.000	0.296	1.080
bed_44	0.6885	0.195	3.530	0.000	0.296	1.080
bed_45	0.6885	0.195	3.530	0.000	0.296	1.080
bed_46	0.6885	0.195	3.530	0.000	0.296	1.080
bed_47	0.6885	0.195	3.530	0.000	0.296	1.080
bed_48	0.6885	0.195	3.530	0.000	0.296	1.080
bed_49	0.6885	0.195	3.530	0.000	0.296	1.080
bed_50	0.6885	0.195	3.530	0.000	0.296	1.080
bed_51	0.6885	0.195	3.530	0.000	0.296	1.080
bed_52	0.6885	0.195	3.530	0.000	0.296	1.080
bed_53	0.6885	0.195	3.530	0.000	0.296	1.080
bed_54	0.6885	0.195	3.530	0.000	0.296	1.080
bed_55	0.6885	0.195	3.530	0.000	0.296	1.080
bed_56	0.6885	0.195	3.530	0.000	0.296	1.080
bed_57	0.6885	0.195	3.530	0.000	0.296	1.080
bed_58	0.6885	0.195	3.530	0.000	0.296	1.080
bed_59	0.6885	0.195	3.530	0.000	0.296	1.080
bed_60	0.6885	0.195	3.530	0.000	0.296	1.080
bed_61	0.6885	0.195	3.530	0.000	0.296	1.080
bed_62	0.6885	0.195	3.530	0.000	0.296	1.080
bed_63	0.6885	0.195	3.530	0.000	0.296	1.080
bed_64	0.6885	0.195	3.530	0.000	0.296	1.080
bed_65	0.6885	0.195	3.530	0.000	0.296	1.080
bed_66	0.6885	0.195	3.530	0.000	0.296	1.080
bed_67	0.6885	0.195	3.530	0.000	0.296	1.080
bed_68	0.6885	0.195	3.530	0.000	0.296	1.080
bed_69	0.6885	0.195	3.530	0.000	0.296	1.080
bed_70	0.6885	0.195	3.530	0.000	0.296	1.080
bed_71	0.6885	0.195	3.530	0.000	0.296	1.080
bed_72	0.6885	0.195	3.530	0.000	0.296	1.080
bed_73	0.6885	0.195	3.530	0.000	0.296	1.080
bed_74	0.6885	0.195	3.530	0.000	0.296	1.080
bed_75	0.6885	0.195	3.530	0.000	0.296	1.080
bed_76	0.6885	0.195	3.530	0.000	0.296	1.080
bed_77	0.6885	0.195	3.530	0.000	0.296	1.080
bed_78	0.6885	0.195	3.530	0.000	0.296	1.080
bed_79	0.6885	0.195	3.530	0.000	0.296	1.080
bed_80	0.6885	0.195	3.530	0.000	0.296	1.080
bed_81	0.6885	0.195	3.530	0.000	0.296	1.080
bed_82	0.6885	0.195	3.530	0.000	0.296	1.080
bed_83	0.6885	0.195	3.530	0.000	0.296	1.080
bed_84	0.6885	0.195	3.530	0.000	0.296	1.080
bed_85	0.6885	0.195	3.530	0.000	0.296	1.080
bed_86	0.6885	0.195	3.530	0.000	0.296	1.080
bed_87	0.6885	0.195	3.530	0.000	0.296	1.080
bed_88	0.6885	0.195	3.530	0.000	0.296	1.080
bed_89	0.6885	0.195	3.530	0.000	0.296	1.080
bed_90	0.6885	0.195	3.530	0.000	0.296	1.080
bed_91	0.6885	0.195	3.530	0.000	0.296	1.080
bed_92	0.6885	0.195	3.530	0.000	0.296	1.080
bed_93	0.6885	0.195	3.530	0.000	0.296	1.080
bed_94	0.6885	0.195	3.530	0.000	0.296	1.080
bed_95	0.6885	0.195	3.530	0.000	0.296	1.080
bed_96	0.6885	0.195	3.530	0.000	0.296	1.080
bed_97	0.6885	0.195	3.530	0.000	0.296	1.080
bed_98	0.6885	0.195	3.530	0.000	0.296	1.080
bed_99	0.6885	0.195	3.530	0.000	0.296	1.080
bed_100	0.6885	0.195	3.530	0.000	0.296	1.080
bed_101	0.6885	0.195	3.530	0.000	0.296	1.080
bed_102	0.6885	0.195	3.530	0.000	0.296	1.080
bed_103	0.6885	0.195	3.530	0.000	0.296	1.080
bed_104	0.6885	0.195	3.530	0.000	0.296	1.080
bed_105	0.6885	0.195	3.530	0.000	0.296	1.080
bed_106	0.6885	0.195	3.530	0.000	0.296	1.080
bed_107	0.6885	0.195	3.530	0.000	0.296	1.080
bed_108	0.6885	0.195	3.530	0.000	0.296	1.080
bed_109	0.6885	0.195	3.530	0.000	0.296	1.080
bed_110	0.6885	0.195	3.530	0.000	0.296	1.080
bed_111	0.6885	0.195	3.530	0.000	0.296	1.080
bed_112	0.6885	0.195	3.530	0.000	0.296	1.080
bed_113	0.6885	0.195	3.530	0.000	0.296	1.080
bed_114	0.6885	0.195	3.530	0.000	0.296	1.080
bed_115	0.6885	0.195	3.530	0.000	0.296	1.080
bed_116	0.6885	0.195	3.530	0.000	0.296	1.080
bed_117	0.6885	0.195	3.530	0.000	0.296	1.080
bed_118	0.6885	0.195	3.530	0.000	0.296	1.080
bed_119	0.6885	0.195	3.530	0.000	0.296	1.080
bed_120	0.6885	0.195	3.530	0.000	0.296	1.080
bed_121	0.6885	0.195	3.530	0.000	0.296	1.080
bed_122	0.6885	0.195	3.530	0.000	0.296	1.080
bed_123	0.6885	0.195	3.530	0.000	0.296	1.080
bed_124	0.6885	0.195	3.530	0.000	0.296	1.080
bed_125	0.6885	0.195	3.530	0.000	0.296	1.080
bed_126	0.6885	0.195	3.530	0.000	0.296	1.080
bed_127	0.6885	0.195	3.530	0.000	0.296	1.080
bed_128	0.6885	0.195	3.530	0.000	0.296	1.080
bed_129	0.6885	0.195	3.530	0.000	0.296	1.080
bed_130	0.6885	0.195	3.530	0.000	0.296	1.080
bed_131	0.6885	0.195	3.530	0.000	0.296	1.080
bed_132	0.6885	0.195	3.530	0.000	0.296	1.080
bed_133	0.6885	0.195	3.530	0.000	0.296	1.080
bed_134	0.6885	0.195	3.530	0.000	0.296	1.080
bed_135	0.6885</					

are much closer to the ideal values of a normal distribution(0 and 3 respectively) for the model without than the model with.

This is also confirmed by the Omnibus test's probability

4. Some issues with the categorical bathroom variable for both models. P-values >5%. Need to change bathroom variable..25 bathroom doesn't make sense.

Phase 2 Results. Models 'with' | 'without'

Dep. Variable:	price	R-squared:	0.730
Model:	OLS	Adjusted R-squared:	0.720
Method:	Least Squares	F-statistic:	1263.0
Date:	Sun, 19 Apr 2020	Prob (F-statistic):	0.00
Time:	10:52:57	Log-Likelihood:	-1502.3
No. Observations:	16884	AIC:	3259
DF Residuals:	16850	BIC:	3545
DF Model:	34		
Covariance Type:	nonconstant		

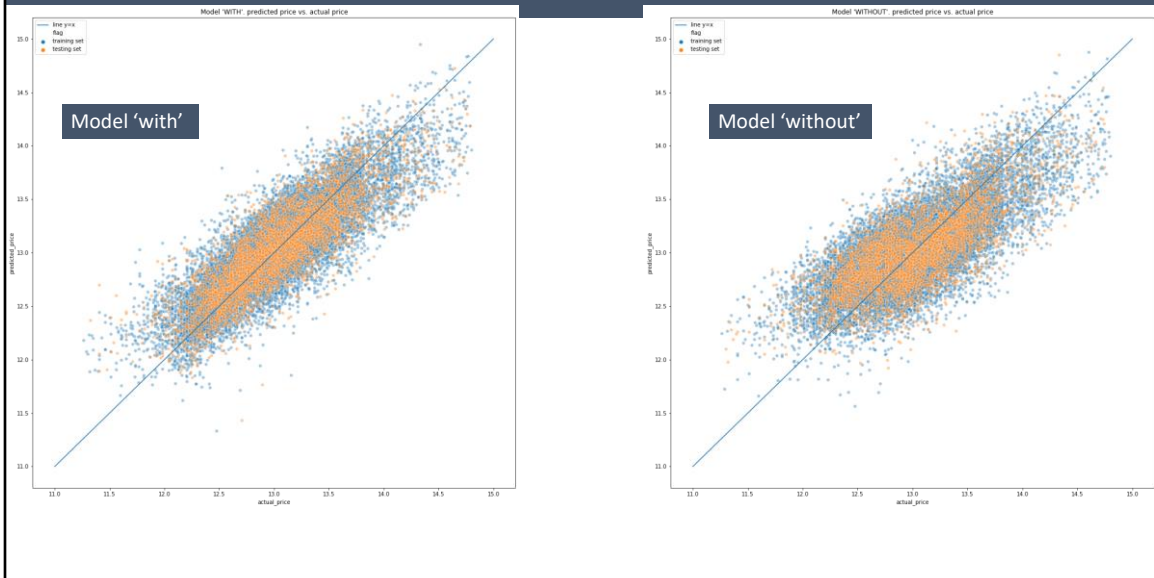
	coef	std err	t	Pr> t	[0.025	0.975]
const	52.9178	0.015	44.000	0.000	-44.516	41.319
sqft_lot	-2.138e-07	1.52e-07	-1.400	0.160	-6.19e-07	8.36e-08
sqft_above	0.0033	0.012	0.267	0.794	-0.017	0.023
yr_built	-0.0032	0.000	-31.279	0.000	-0.003	-0.003
lat	1.3696	0.016	86.969	0.000	1.330	1.401
sqft_living_ratio15	-0.2389	0.008	-28.588	0.000	-0.255	-0.223
sqft_living_ratio15	0.0244	0.011	20.284	0.000	0.004	0.048
sqft_living_ratio15	-0.0058	0.003	-2.418	0.016	-0.010	-0.001
bed_1	-0.1530	0.023	-6.606	0.000	-0.198	-0.108
bed_2	-0.1896	0.024	-8.012	0.000	-0.237	-0.144
bed_3	-0.2240	0.025	-8.968	0.000	-0.273	-0.175
bed_4	-0.2731	0.031	-8.840	0.000	-0.333	-0.213
bed_5	-0.3184	0.030	-10.520	0.000	-0.430	-0.201
bed_6	0.3347	0.158	2.123	0.034	0.026	0.644
bed_7	0.2088	0.154	1.324	0.189	-0.021	0.563
bed_8	0.3816	0.186	2.050	0.040	0.010	0.753
bed_9	0.2627	0.154	1.706	0.089	-0.039	0.569
bed_10	0.2080	0.154	1.364	0.173	-0.017	0.587
bed_11	0.3086	0.154	2.006	0.046	0.006	0.612
bed_12	0.2020	0.154	1.300	0.191	-0.009	0.586
bed_13	0.2946	0.154	1.910	0.060	0.007	0.587
bed_14	0.3011	0.154	2.278	0.023	0.040	0.563
bed_15	0.3684	0.154	2.373	0.019	0.064	0.680
bed_16	0.4405	0.155	2.881	0.004	0.142	0.748
bed_17	0.4048	0.156	2.642	0.009	0.102	0.706
bed_18	0.5484	0.156	3.517	0.000	0.243	0.856
bed_19	0.5187	0.157	3.294	0.001	0.209	0.823
bed_20	0.6275	0.159	3.950	0.000	0.316	0.939
bed_21	0.4670	0.156	2.982	0.002	0.176	0.758
bed_22	0.6127	0.154	3.921	0.000	0.272	0.954
bed_23	0.6833	0.172	3.947	0.000	0.325	1.001
bed_24	0.5934	0.165	3.571	0.000	0.217	0.962
bed_25	0.6754	0.168	4.019	0.000	0.340	1.010
bed_26	0.6802	0.167	4.073	0.000	0.345	1.015
bed_27	0.6805	0.167	4.073	0.000	0.345	1.015
bed_28	0.6805	0.167	4.073	0.000	0.345	1.015
bed_29	0.6805	0.167	4.073	0.000	0.345	1.015
bed_30	0.6805	0.167	4.073	0.000	0.345	1.015
bed_31	0.6805	0.167	4.073	0.000	0.345	1.015
bed_32	0.6805	0.167	4.073	0.000	0.345	1.015
bed_33	0.6805	0.167	4.073	0.000	0.345	1.015
bed_34	0.6805	0.167	4.073	0.000	0.345	1.015
bed_35	0.6805	0.167	4.073	0.000	0.345	1.015
bed_36	0.6805	0.167	4.073	0.000	0.345	1.015
bed_37	0.6805	0.167	4.073	0.000	0.345	1.015
bed_38	0.6805	0.167	4.073	0.000	0.345	1.015
bed_39	0.6805	0.167	4.073	0.000	0.345	1.015
bed_40	0.6805	0.167	4.073	0.000	0.345	1.015
bed_41	0.6805	0.167	4.073	0.000	0.345	1.015
bed_42	0.6805	0.167	4.073	0.000	0.345	1.015
bed_43	0.6805	0.167	4.073	0.000	0.345	1.015
bed_44	0.6805	0.167	4.073	0.000	0.345	1.015
bed_45	0.6805	0.167	4.073	0.000	0.345	1.015
bed_46	0.6805	0.167	4.073	0.000	0.345	1.015
bed_47	0.6805	0.167	4.073	0.000	0.345	1.015
bed_48	0.6805	0.167	4.073	0.000	0.345	1.015
bed_49	0.6805	0.167	4.073	0.000	0.345	1.015
bed_50	0.6805	0.167	4.073	0.000	0.345	1.015
bed_51	0.6805	0.167	4.073	0.000	0.345	1.015
bed_52	0.6805	0.167	4.073	0.000	0.345	1.015
bed_53	0.6805	0.167	4.073	0.000	0.345	1.015
bed_54	0.6805	0.167	4.073	0.000	0.345	1.015
bed_55	0.6805	0.167	4.073	0.000	0.345	1.015
bed_56	0.6805	0.167	4.073	0.000	0.345	1.015
bed_57	0.6805	0.167	4.073	0.000	0.345	1.015
bed_58	0.6805	0.167	4.073	0.000	0.345	1.015
bed_59	0.6805	0.167	4.073	0.000	0.345	1.015
bed_60	0.6805	0.167	4.073	0.000	0.345	1.015
bed_61	0.6805	0.167	4.073	0.000	0.345	1.015
bed_62	0.6805	0.167	4.073	0.000	0.345	1.015
bed_63	0.6805	0.167	4.073	0.000	0.345	1.015
bed_64	0.6805	0.167	4.073	0.000	0.345	1.015
bed_65	0.6805	0.167	4.073	0.000	0.345	1.015
bed_66	0.6805	0.167	4.073	0.000	0.345	1.015
bed_67	0.6805	0.167	4.073	0.000	0.345	1.015
bed_68	0.6805	0.167	4.073	0.000	0.345	1.015
bed_69	0.6805	0.167	4.073	0.000	0.345	1.015
bed_70	0.6805	0.167	4.073	0.000	0.345	1.015
bed_71	0.6805	0.167	4.073	0.000	0.345	1.015
bed_72	0.6805	0.167	4.073	0.000	0.345	1.015
bed_73	0.6805	0.167	4.073	0.000	0.345	1.015
bed_74	0.6805	0.167	4.073	0.000	0.345	1.015
bed_75	0.6805	0.167	4.073	0.000	0.345	1.015
bed_76	0.6805	0.167	4.073	0.000	0.345	1.015
bed_77	0.6805	0.167	4.073	0.000	0.345	1.015
bed_78	0.6805	0.167	4.073	0.000	0.345	1.015
bed_79	0.6805	0.167	4.073	0.000	0.345	1.015
bed_80	0.6805	0.167	4.073	0.000	0.345	1.015
bed_81	0.6805	0.167	4.073	0.000	0.345	1.015
bed_82	0.6805	0.167	4.073	0.000	0.345	1.015
bed_83	0.6805	0.167	4.073	0.000	0.345	1.015
bed_84	0.6805	0.167	4.073	0.000	0.345	1.015
bed_85	0.6805	0.167	4.073	0.000	0.345	1.015
bed_86	0.6805	0.167	4.073	0.000	0.345	1.015
bed_87	0.6805	0.167	4.073	0.000	0.345	1.015
bed_88	0.6805	0.167	4.073	0.000	0.345	1.015
bed_89	0.6805	0.167	4.073	0.000	0.345	1.015
bed_90	0.6805	0.167	4.073	0.000	0.345	1.015
bed_91	0.6805	0.167	4.073	0.000	0.345	1.015
bed_92	0.6805	0.167	4.073	0.000	0.345	1.015
bed_93	0.6805	0.167	4.073	0.000	0.345	1.015
bed_94	0.6805	0.167	4.073	0.000	0.345	1.015
bed_95	0.6805	0.167	4.073	0.000	0.345	1.015
bed_96	0.6805	0.167	4.073	0.000	0.345	1.015
bed_97	0.6805	0.167	4.073	0.000	0.345	1.015
bed_98	0.6805	0.167	4.073	0.000	0.345	1.015
bed_99	0.6805	0.167	4.073	0.000	0.345	1.015
bed_100	0.6805	0.167	4.073	0.000	0.345	1.015
bed_101	0.6805	0.167	4.073	0.000	0.345	1.015
bed_102	0.6805	0.167	4.073	0.000	0.345	1.015
bed_103	0.6805	0.167	4.073	0.000	0.345	1.015
bed_104	0.6805	0.167	4.073	0.000	0.345	1.015
bed_105	0.6805	0.167	4.073	0.000	0.345	1.015
bed_106	0.6805	0.167	4.073	0.000	0.345	1.015
bed_107	0.6805	0.167	4.073	0.000	0.345	1.015
bed_108	0.6805	0.167	4.073	0.000	0.345	1.015
bed_109	0.6805	0.167	4.073	0.000	0.345	1.015
bed_110	0.6805	0.167	4.073	0.000	0.345	1.015
bed_111	0.6805	0.167	4.073	0.000	0.345	1.015
bed_112	0.6805	0.167	4.073	0.000	0.345	1.015
bed_113	0.6805	0.167	4.073	0.000	0.345	1.015
bed_114	0.6805	0.167	4.073	0.000	0.345	1.015
bed_115	0.6805	0.167	4.073	0.000	0.345	1.015
bed_116	0.6805	0.167	4.073	0.000	0.345	1.015
bed_117	0.6805	0.167	4.073	0.000	0.345	1.015
bed_118	0.6805	0.167	4.073	0.000	0.345	1.015
bed_119	0.6805	0.167	4.073	0.000	0.345	1.015
bed_120	0.6805	0.167	4.073	0.000	0.345	1.015
bed_121	0.6805	0.167	4.073	0.000	0.345	1.015
bed_122	0.6805	0.167	4.073	0.000	0.345	1.015
bed_123	0.6805	0.167	4.073	0.000	0.345	1.015
bed_124	0.6805	0.167	4.073	0.000	0.345	1.015
bed_125	0.6805	0.167	4.073	0.000	0.345	1.015
bed_126	0.6805	0.167	4.073	0.000	0.345	1.015
bed_127	0.6805	0.167	4.073	0.000	0.345	1.015
bed_128	0.6805	0.167	4.073	0.000	0.345	1.015
bed_129	0.6805	0.167	4.073	0.000	0.345	1.015
bed_130	0.6805	0.167	4.073	0.000	0.345	1.015
bed_131	0.6805	0.167	4.073	0.000	0.345	1.015
bed_132	0.6805	0.167	4.073	0.000	0.345	1.015
bed_133	0.6805	0.167	4.073	0.000	0.345	1.015
bed_134	0.6805	0.167	4.073	0.000	0.345	1.015
bed_135	0.6805	0.167	4.073	0.000	0.345	1.015
bed_136	0.6805	0.167	4.073	0.000	0.345	1.015
bed_137	0.6805	0.167	4.073	0.000	0.345	1.015
bed_138	0.6805	0.167	4.073	0.000	0.345	1.015
bed_139	0.6805	0.167	4.073	0.000	0.345	1.015
bed_140	0.6805	0.167	4.073	0.000	0.345	1.015
bed_141	0.6805	0.167	4.073	0.000	0.345	1.015
bed_142	0.6805	0.167	4.073	0.000	0.345	1.015
bed_143	0.6805	0.167	4.073	0.000	0.345	1.015
bed_144	0.6805	0.167	4.073	0.000	0.345	1.015
bed_145	0.6805	0.167	4.073	0.000	0.345	1.015
bed_146	0.6805	0.167	4.073	0.000	0.345	1.015
bed_147	0.6805	0.167	4.073	0.000	0.345	1.015
bed_148	0.6805	0.16				

are much closer to the ideal values of a normal distribution(0 and 3 respectively) for the model without than the model with.

This is also confirmed by the Omnibus test's probability

4._Some issues with the categorical bathroom variable for both models. P-values >5%.
Need to change bathroom variable..25 bathroom doesn't make sense.

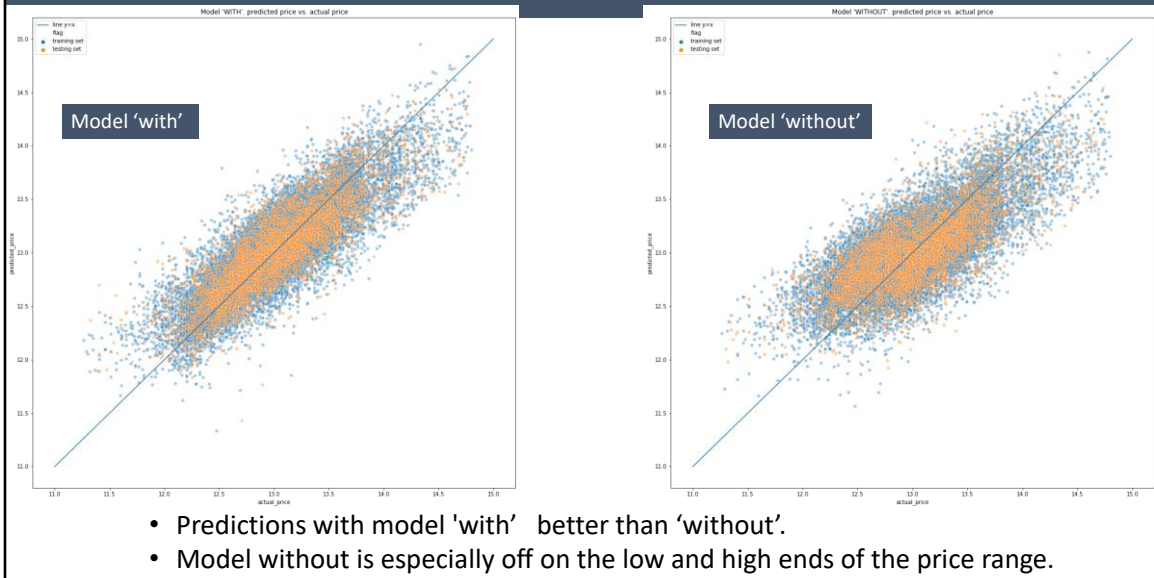
Phase 2 Results. Residuals.



A perfect model would predict the right value so all points would be on the line $y = x$.

- Predictions with model 'with' better than 'without'.
- Model without is especially off on the low and high sides of the price range.
- Observation is in agreement with the Rsquared values of the 2 models

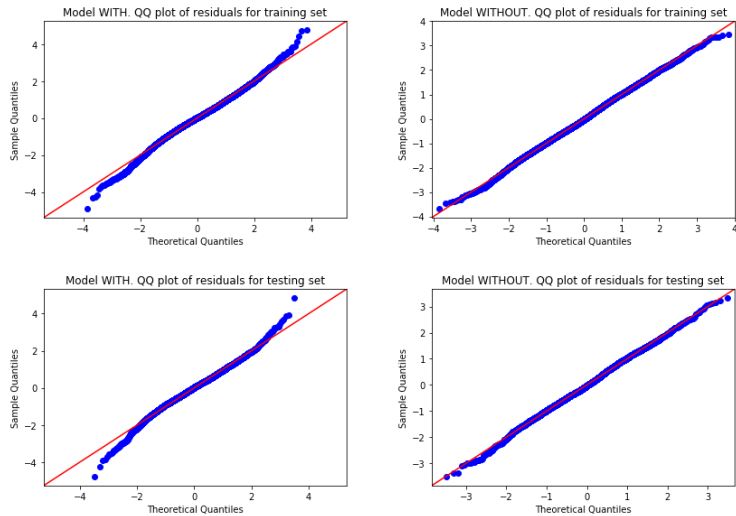
Phase 2 Results. Residuals.



A perfect model would predict the right value so all points would be on the line $y = x$.

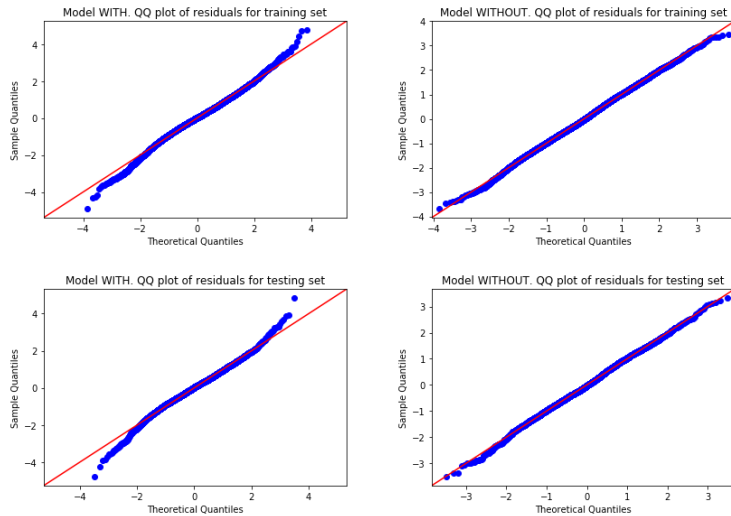
- Predictions with model 'with' better than 'without'.
- Model without is especially off on the low and high sides of the price range.
- Observation is in agreement with the Rsquared values of the 2 models

Phase 2 Results. Normality of Residuals.



The QQ-plots show clearly that the residuals for the model 'without' are closer to a normal distribution.

Phase 2 Results. Normality of Residuals.

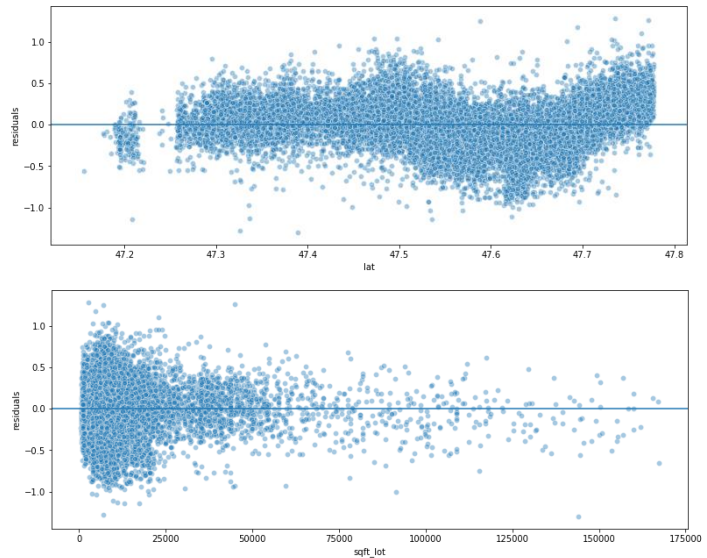


Residuals for the model 'without' are closer to a normal distribution.

The QQ-plots show clearly that the residuals for the model 'without' are closer to a normal distribution.

Phase 2 Results. Heteroscedasticity

Scatterplots of residuals with respect to latitude (lat) and sqft_lot for the model 'with'



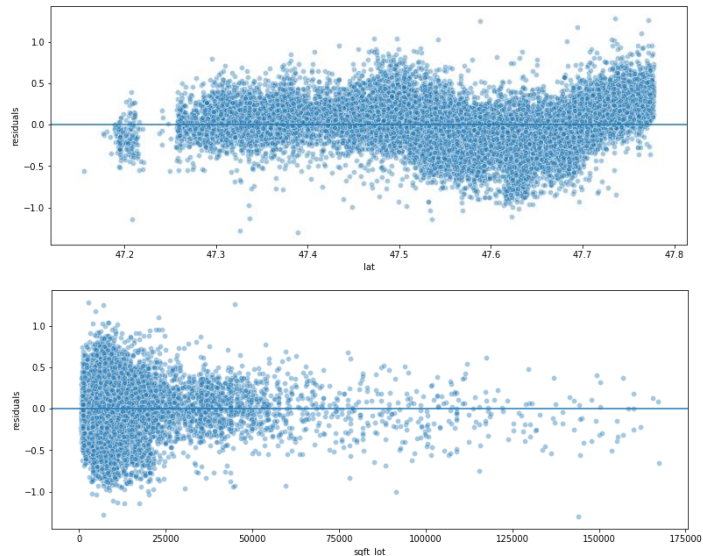
The scatterplots of residuals with respect to the main predictors show satisfactory homoscedasticity for the model without.

For the model with, the worst features for homoscedasticity are latitude (lat) and sqft_lot

Phase 2 Results. Heteroscedasticity

Scatterplots of residuals with respect to latitude (lat) and sqft_lot for the model 'without'

- Heteroscedasticity of the model 'with'.
- This is the third violation of the assumptions for a linear regression model



The scatterplots of residuals with respect to the main predictors show satisfactory homoscedasticity for the model without.

For the model with, the worst features for homoscedasticity are latitude (lat) and sqft_lot

Phase 2 Results. Summary. Comparison models WITH and WITHOUT

Model 'WITH' yr_renovated, lat and sqft_lot	Model 'WITHOUT'
Variations in price are not linear or polynomial for 3 var. yr_renovated, lat and sqft_lot	
As is, sqft_lot is a bad predictor of price.	
Issues with the bathroom variable	
More accurate predictions	Less accurate predictions
Normality of residuals: wrong.	Residuals are distributed normally
Heteroscedasticity	Homoscedasticity

Future Work

- Start over from model 'WITHOUT'.
- Understand and use 'sqft_lot' for prediction.
Consider interactions with other variables.

(Steps 1 and 5 Business Understanding & Feature Engineering)

- Fix 'bathroom' feature. Eliminate if it doesn't help predictions.

(Step 5. Feature Engineering)

- Use powers of variable 'lat' and polynomial regression to fix heteroscedasticity.

(Step 6. Predictive modeling)

- Work on 'mansion' price prediction as a separate project.



Start over from model 'WITHOUT'.

Understand and use sqft_lot for prediction

(Steps 01 and 05 Business Understanding & Feature Engineering)

Current hypothesis: the sqft_lot and yr_building have a strong interaction. New houses built on old small lots have very high price even if lot is small.

Maybe also interaction with 'grade'

Fix 'bathroom' feature. Eliminate if it doesn't help predictions

(step 5. Feature Engineering)

Remove .25 intervals. Only .5. maybe redundant with bedrooms. Predict with feature bathroom and without. Compare

Use powers of variable 'lat' and polynomial regression to fix heteroscedasticity.

(Step 6. Predictive modeling)

Thank you!

Questions?