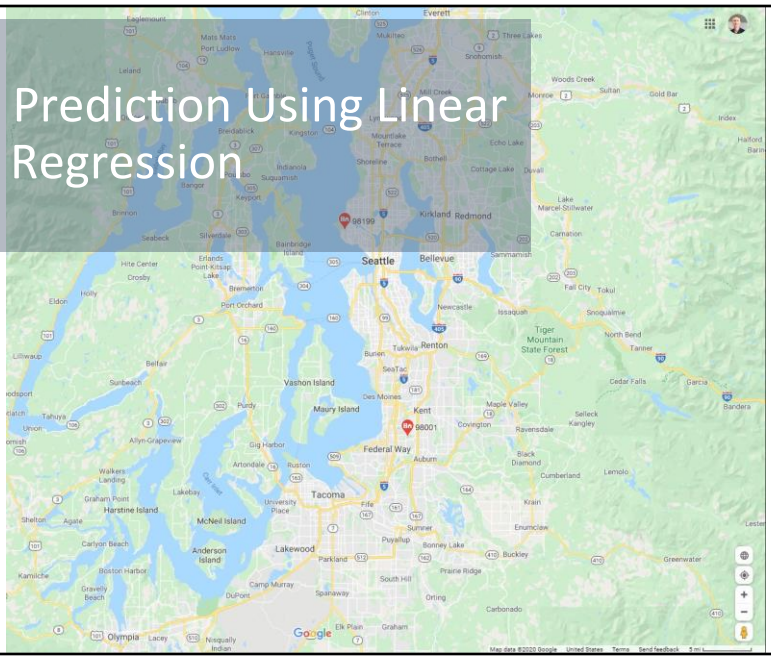
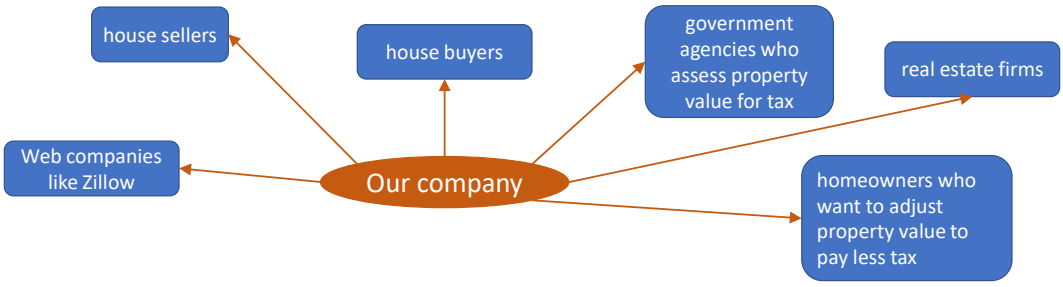
House Price Prediction Using Linear Regression



Pierre-Olivier Ariston

April 2020

Problem Statement and business value

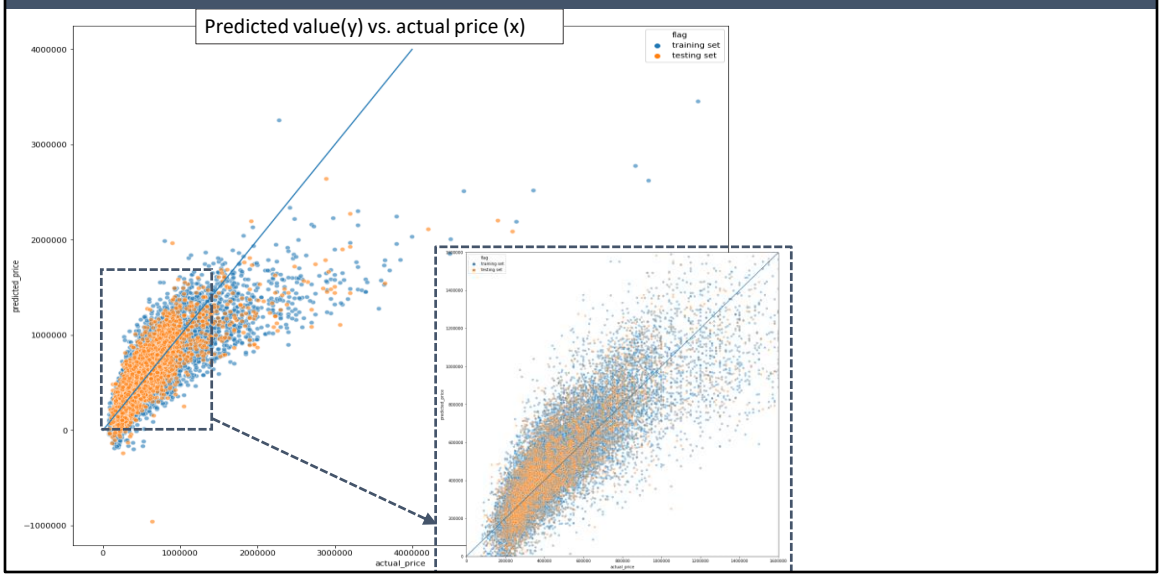


Value we bring: Expertise in prediction of house selling prices.
How we are different: **Superior accuracy** of our models

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

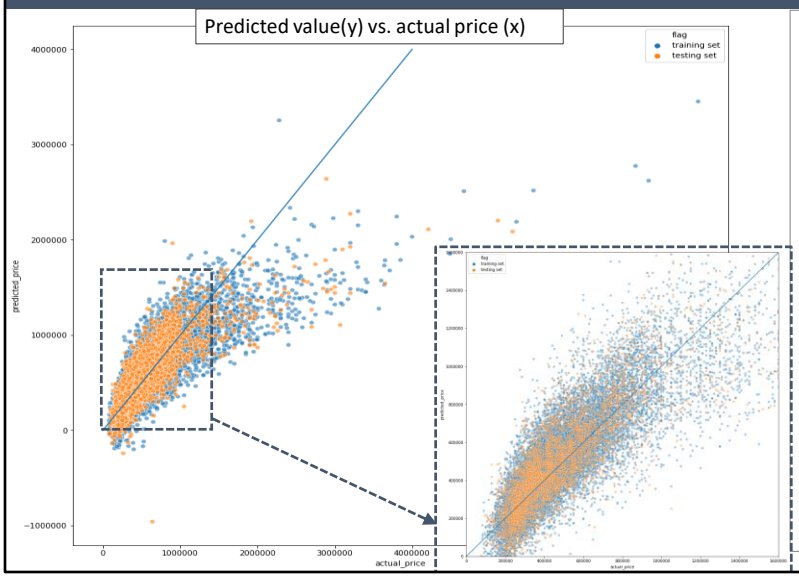
The **value** we bring is our expertise in house market value prediction.
As a new company, we need a **very accurate model** that sets us apart from competition, establishes our reputation and makes us gain new clients.

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)



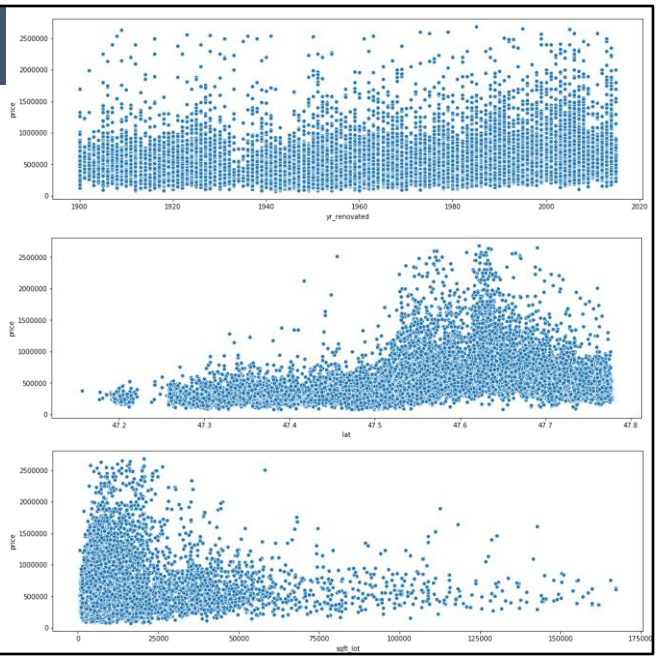
- 66% (R-squared) of the variations in the house price are explained.
- predicted value too small, especially on the high end
- Training and test performances are close but Mean Square Error (MSE) for test and training set is ~20% different.
- Presence of outliers.

Problem for prediction of expensive houses: inaccurate predictions+ test performances very different form 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%

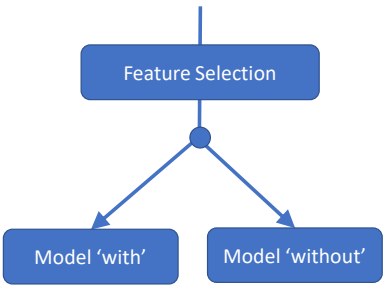


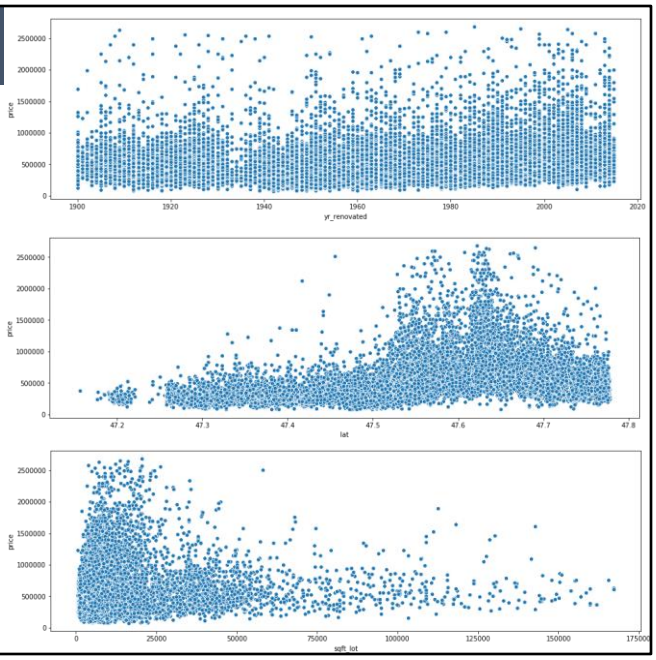
At this point we are ready to move into fitting a multivariate linear model to the data, but we have issues with 3 features that contradict the assumptions of this process: yr_renovated, lat, sqft_lot.
These 3 predictors have correlation levels with the target price that are low to medium.
The main issue 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

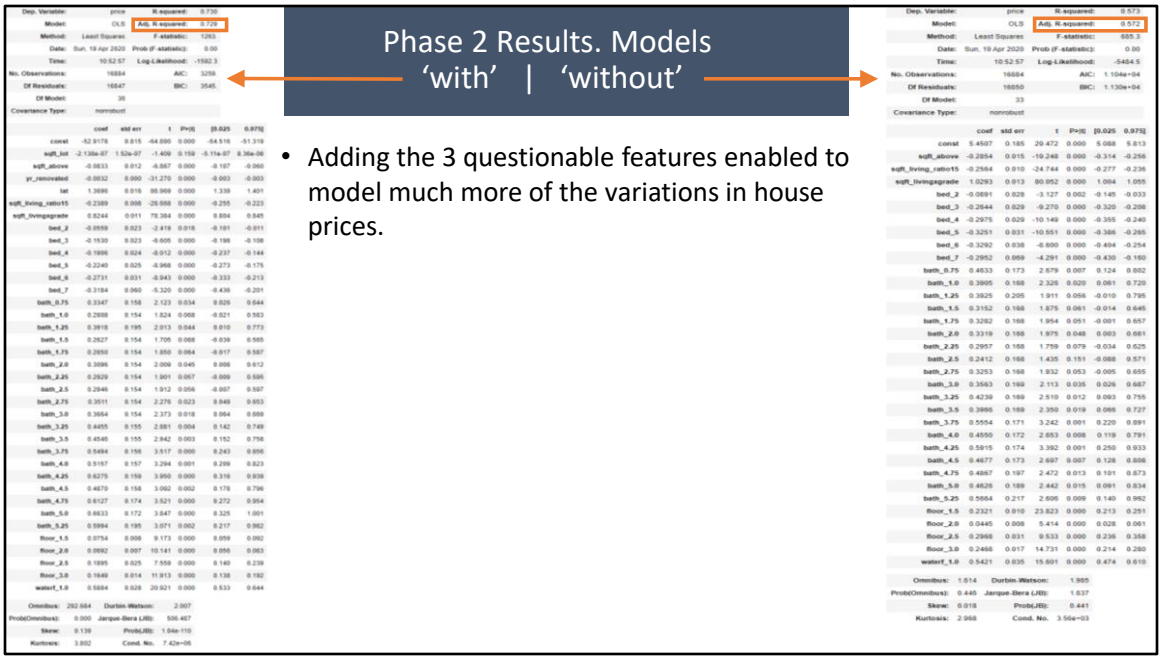
Graphs of price (y) vs features **yr_renovated** (top), **lat** and **sqft_lot** (bottom).

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



- 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:

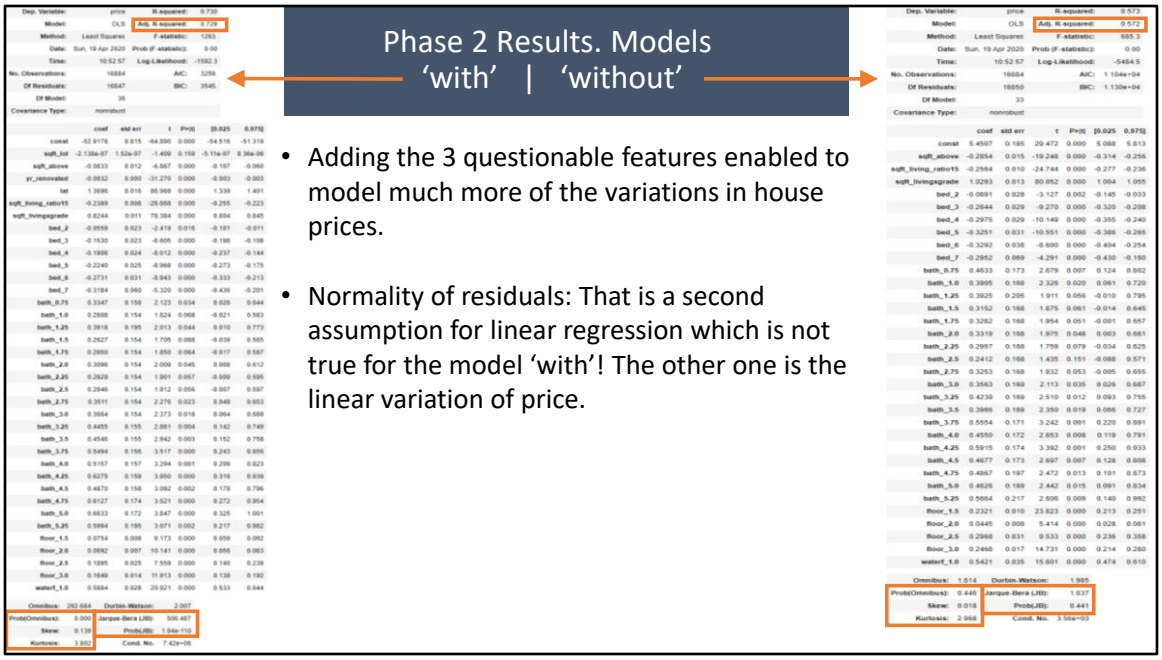
1. With: R-squared adjusted = 73%. It means that 73% of the variations of the target variable price are predicted by the model.
Without: R-squared adjusted is much smaller: 57%
This is consistent with F statistic being much higher with (1263) than without (685).
Adding the 3 features enable to model 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.



- 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:

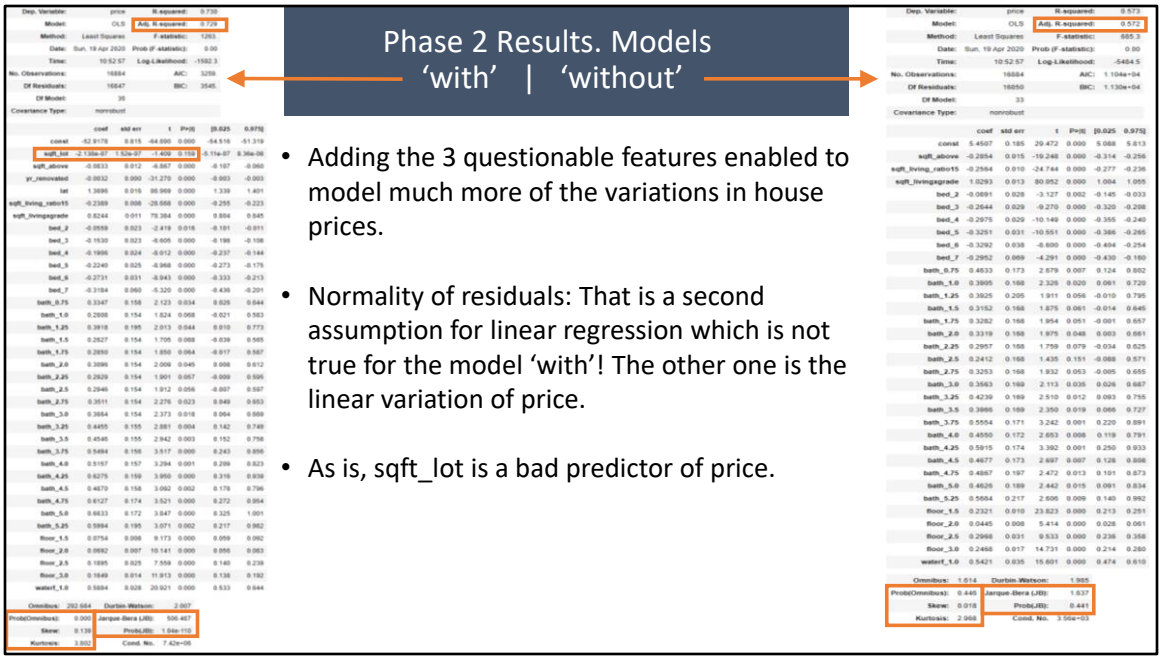
1. With: R-squared adjusted = 73%. It means that 73% of the variations of the target variable price are predicted by the model.
Without: R-squared adjusted is much smaller: 57%
This is consistent with F statistic being much higher with (1263) than without (685).
Adding the 3 features enable to model 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.



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.
- As is, sqft_lot is a bad predictor of price.

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

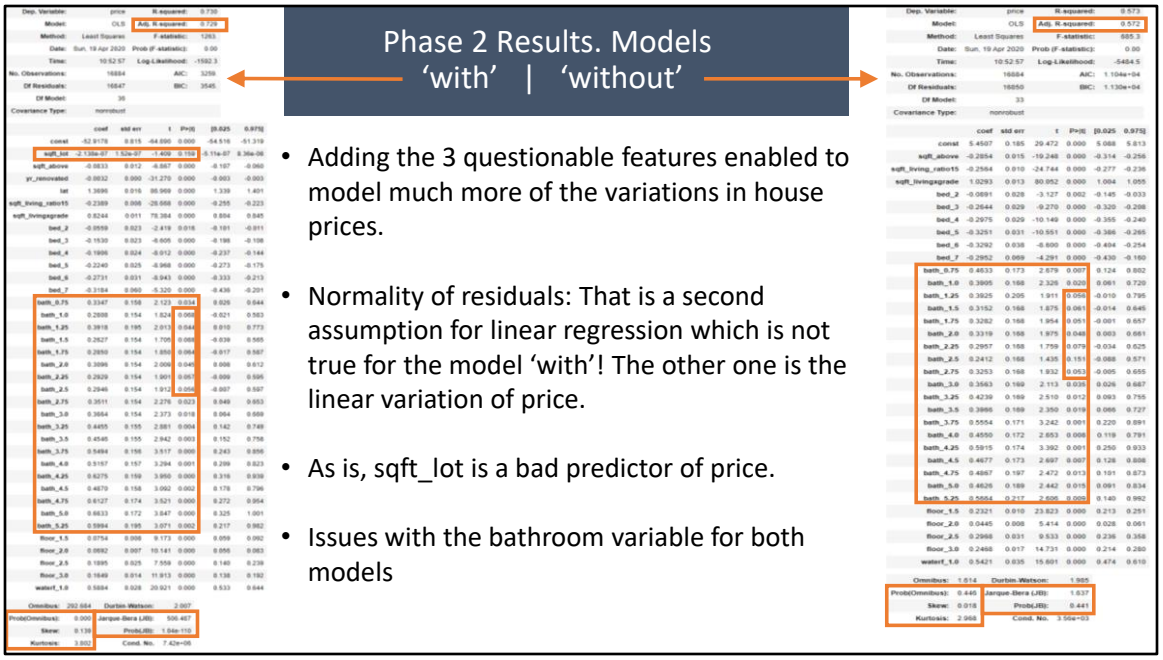
1. With: R-squared adjusted = 73%. It means that 73% of the variations of the target variable price are predicted by the model.
Without: R-squared adjusted is much smaller: 57%
This is consistent with F statistic being much higher with (1263) than without (685).
Adding the 3 features enable to model 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.



- 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.
- As is, sqft_lot is a bad predictor of price.
- Issues with the bathroom variable for both models

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.

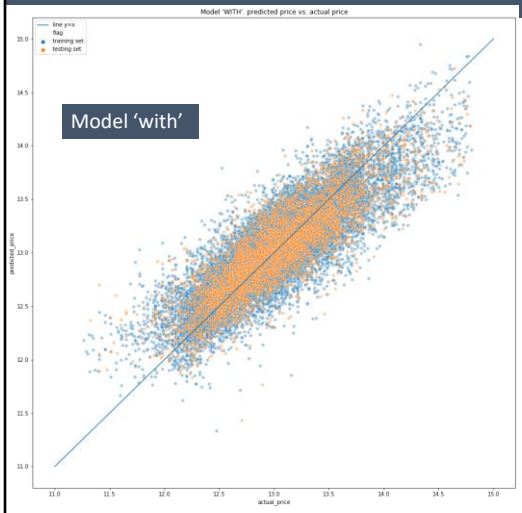
2. From the coefficients report (with), we can tell that sqft_lot is actually not a statistically significant predictor. Its p- value is 16%.

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.

Phase 2 Results. Residuals.



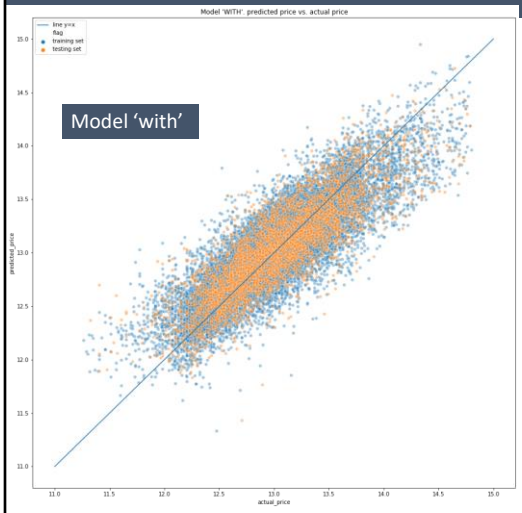
Model ‘with’

Model ‘without’

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.



Model ‘with’

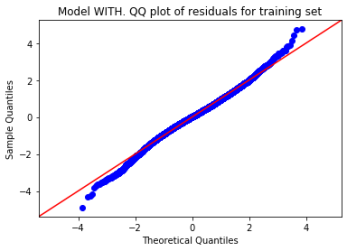
Model ‘without’

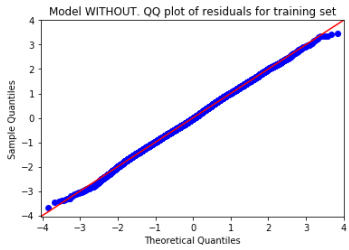
- Predictions with model 'with’ better than ‘without’.
- Model without is especially off on the low and high ends of the price range.

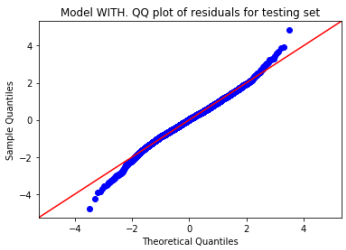
A perfect model would predict the right value so all points would be on the line y= x.

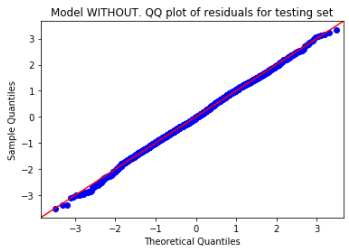
- Predictions with model 'with’ better than ‘without’.
- Model without is especially off on the low and high sides of the price range.
- Observation is in agreement with the Rsquared values of the 2 models

Phase 2 Results. Normality of Residuals.



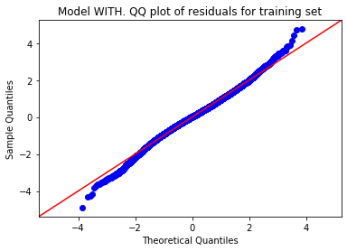


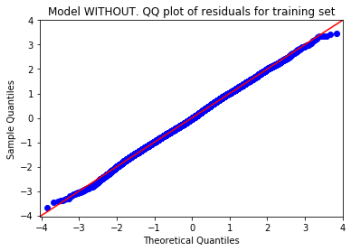


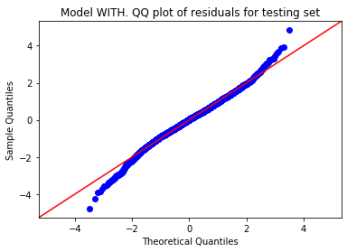


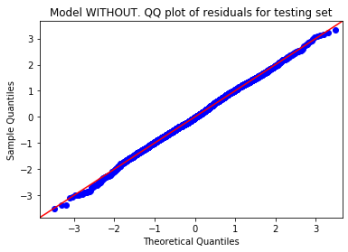
The QQ-plots show clearly that the residuals for the model 'without' are closer to a normal distribution.

Phase 2 Results. Normality of Residuals.







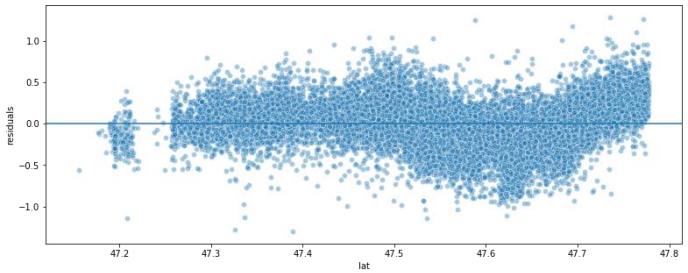


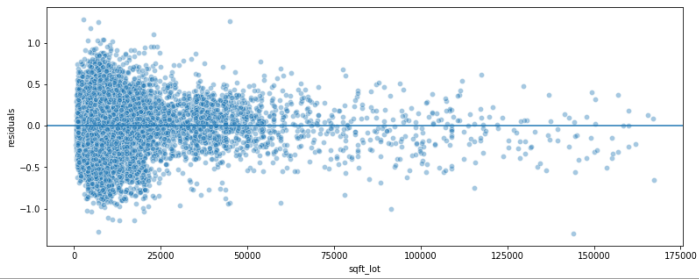
Residuals for the model 'without' are closer to a normal distribution.

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’



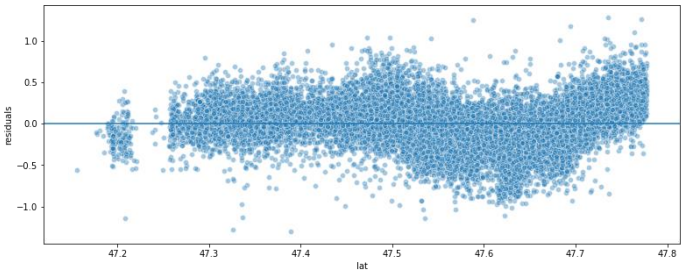


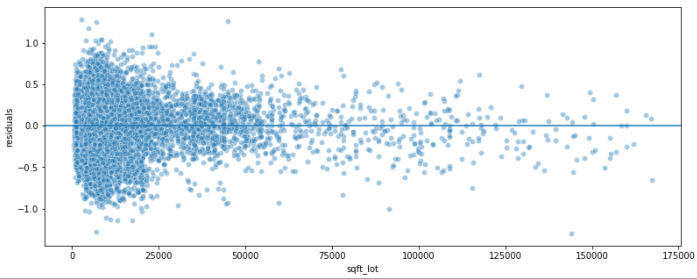
The scatterplots of residuals with respect to the main predictors show satisfactory homoscedasticity for the model without.
For the model with, the worst features for homoscedasticity are latitude (lat) and sqft_lot

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?