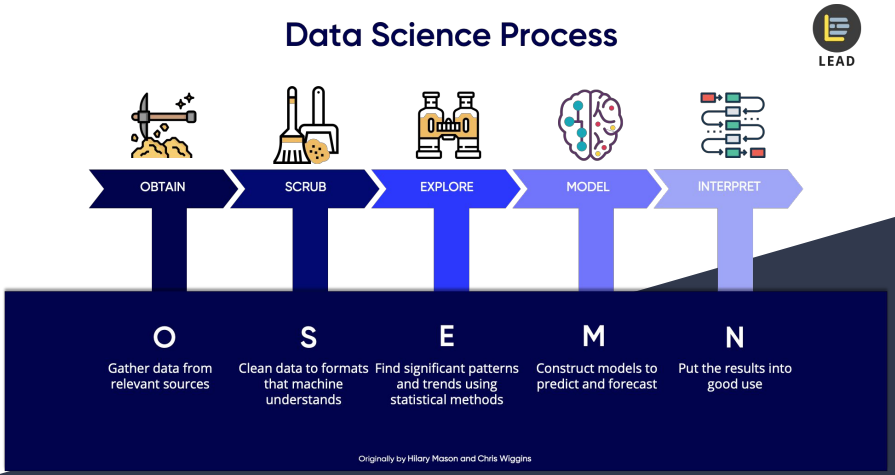
Multilinear Regression with King County Housing Data

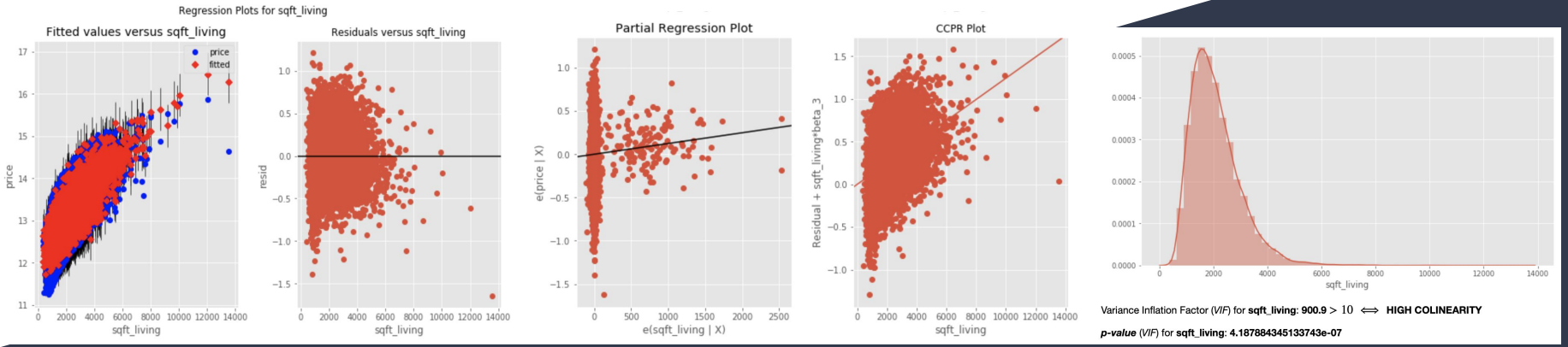
A Strategic Guide to Increase Market Value

Approach: The OSEMN Model



EDA

Leveraging Regression Diagnostics to Identify Required Transformations of Predictors

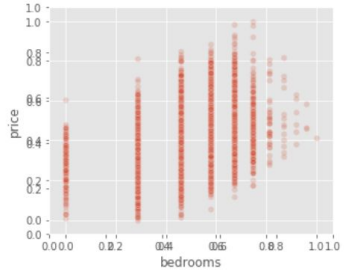


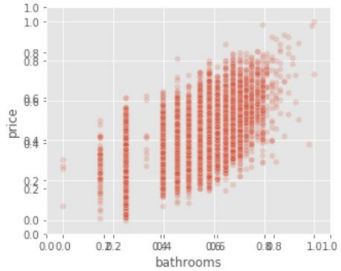
EDA

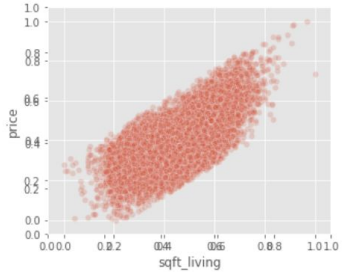
Linearity Study

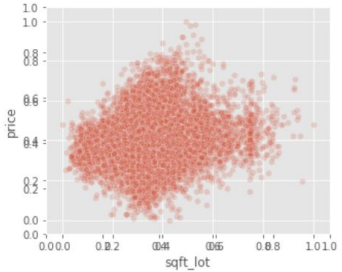
Scatter Plots:

- **target**: price
- **feature set**: Index(['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'lat', 'long', 'sqft_living15', 'sqft_lot15'], dtype='object')



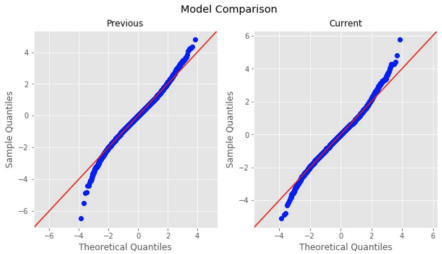






EDA

Iterative Analysis and Model Comparison



Summary

1. **Indep.**:
   A. Previous: ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long', 'sqft_living15', 'sqft_lot15']
   B. Current: ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'lat', 'long', 'sqft_living15', 'sqft_lot15']
2. DROPPED Indep.: ['zipcode']
3. R^2 :
   A. Previous: 0.7711554350458103
   B. Current: 0.7704646634831092
4. ΔR^2 : -0.0006907715627011424 (0.09% worse/reduction)
5. **Adjusted** R^2 :
   A. Previous: 0.7709167652405391
   B. Current: 0.7702385858051957
6. Δ (R^2 – **Adj.** R^2): -1.2592127357713423e-05 (105.57% better/reduction)
7. **rmse_train__and__rmse_test**:
   A. Previous: (0.252066606293786 47, 0.2536284341828196) $\implies \Delta$: 0.0015618278890331405
   B. Current: (0.05493267955260844, 0.05525435215613465) $\implies \Delta$: 0.0003216726035262102
8. $\Delta\Delta$ **rmse_train__and__rmse_test**: -0.0012401552855069303 (485.53% better/reduction)
9. **Condition No.**:
   A. Previous: 214442432.5595656
   B. Current: 106.68524661124478
10. Δ **Condition No**: -214442325.874319 (201004768.11% better/reduction)

We see that our R^2 has only slightly dropped after feature-tuning with the guidance of regression diagnostics.

$RMSE$ and $\Delta RMSE$ have dropped quite a bit.

$\Delta RMSE$ has improved by a factor of 5!

By virtue of the **drop in *Condition Number***, we see that **multicolinearity has been almost completely mitigated**!

This model manifests a kurtosis that is clearly platykurtic. It's not ideal but at least it's not leptokurtic. We have significantly improved kurtosis in residuals at the low-end tail but it is slightly worse on the high-end. Overall, however, distribution of residuals (kurtosis) has improved. But, we may be able to do more to produce a more mesokurtic distribtion in residuals.

This is a MAJOR improvement.

Which hypothesis?

1. clearly, there is a linear relationship between *log-transformed* **price** vs. *transformed* and/or *scaled continuous* features.
2. clearly, $\Delta RMSE$ has decreased!
3. clear, multicolinearity has improved (decreased Condition No.)

Optimized Feature Selection

Don’t Guess: Combinatorics, Dynamic Programming, and Cross Validation using K-Folds

Cross-validating $\binom{17}{15} = 136$ combinations of 15 features (out of 17) over 5 folds using score **condition_no__and__pvals__and__rsquared__and__rsquared_adj__and__rmse__and__delta_rmse** and target cond. no = 100...

```
new best condition_no__and__pvals__and__rsquared__and__rsquared_adj__and__rmse__and__delta_rmse score: (63.21149436
159284, 0.7693242562035698, 0.7691238033995625, [7.904829390219449e-23, 1.4138479423339623e-26, 9.399083713104544e-
126, 0.00550991905996695366, 3.218895787943788e-14, 2.134984133012824e-77, 0.0, 0.018369447725632718, 5.799293785069
347e-271, 2.3352570742093205e-11, 0.0, 1.2411533720270234e-96, 1.3937937921398377e-11, 1.9757945635472998e-76, 2.65
85584627253977e-101], 0.05506275444854354, 0.000589235987488955), from feature-set combo: ['bedrooms', 'bathrooms',
'sqft_living', 'sqft_lot', 'floors', 'condition', 'grade', 'sqft_above', 'yr_built', 'yr_renovated', 'lat', 'sqft_l
iving15', 'sqft_lot15', 'waterfront', 'view']
```

DISCARDED 133 15-feature combinations that were not based on prior optimal feature-combo ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'condition', 'grade', 'yr_built', 'yr_renovated', 'lat', 'sqft_living15', 'sqft_lot15', 'waterfront', 'view']

cv_selection chose the best of 1 15-feature combinations that met the constraints (out of 3 considered)

2 15-feature combinations (out of 3 considered) failed to meet the constraints

Cross-validating $\binom{17}{16} = 17$ combinations of 16 features (out of 17) over 5 folds using score **condition_no__and__pvals__and__rsquared__and__rsquared_adj__and__rmse__and__delta_rmse** and target cond. no = 100...

DISCARDED 15 16-feature combinations that were not based on prior optimal feature-combo ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'condition', 'grade', 'sqft_above', 'yr_built', 'yr_renovated', 'lat', 'sqft_living15', 'sqft_lot15', 'waterfront', 'view']

2 16-feature combinations (out of 2 considered) failed to meet the constraints

Cross-validating $\binom{17}{17} = 1$ combinations of 17 features (out of 17) over 5 folds using score **condition_no__and__pvals__and__rsquared__and__rsquared_adj__and__rmse__and__delta_rmse** and target cond. no = 100...

1 17-feature combinations (out of 1 considered) failed to meet the constraints

cv_selected best condition_no__and__pvals__and__rsquared__and__rsquared_adj__and__rmse__and__delta_rmse = (63.21149436159284, 0.7693242562035698, 0.7691238033995625, [7.904829390219449e-23, 1.4138479423339623e-26, 9.399083713104544e-126, 0.00550991905996695366, 3.218895787943788e-14, 2.134984133012824e-77, 0.0, 0.018369447725632718, 5.799293785069347e-271, 2.3352570742093205e-11, 0.0, 1.2411533720270234e-96, 1.3937937921398377e-11, 1.9757945635472998e-76, 2.65855846272539 77e-101], 0.05506275444854354, 0.000589235987488955)

cv_selected best feature-set combo (15 of 17 features) ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'condition', 'grade', 'sqft_above', 'yr_built', 'yr_renovated', 'lat', 'sqft_living15', 'sqft_lot15', 'waterfront', 'view'] based on condition_no__and__pvals__and__rsquared__and__rsquared_adj__and__rmse__and__delta_rmse scoring method with target cond. no. 100

starting feature-set:['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'lat', 'long', 'sqft_living15', 'sqft_lot15', 'waterfront', 'view']

cv_selection suggests dropping ['long', 'sqft_basement'].

The Final Model

Mathematically Justified Predictive Power:

1. Collinearity Mitigated (low Condition Number)
2. 10 features
3. High R^2 = 0.766
4. Not overfit: virtually non-existent difference between R^2 and Adjusted R^2: 0.000136

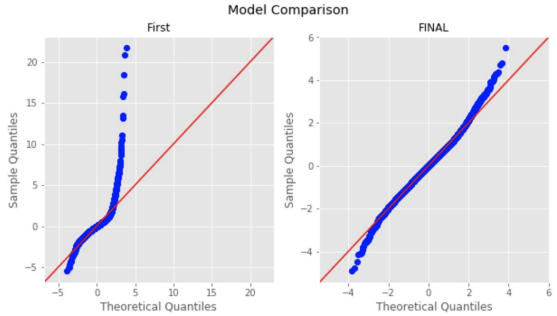
The Final Model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.766
Model:	OLS	**Adj. R-squared:**	0.766
Method:	Least Squares	**F-statistic:**	5661.
Date:	Tue, 14 Jan 2020	**Prob (F-statistic):**	0.00
Time:	19:33:44	**Log-Likelihood:**	25380.
No. Observations:	17278	**AIC:**	-5.074e+04
Df Residuals:	17267	**BIC:**	-5.065e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0979	0.003	-30.620	0.000	-0.104	-0.092
bathrooms	0.0628	0.005	11.932	0.000	0.052	0.073
sqft_living	0.2220	0.007	29.878	0.000	0.207	0.237
floors	0.0355	0.003	13.972	0.000	0.031	0.041
condition	0.0530	0.003	19.342	0.000	0.048	0.058
grade	0.3750	0.006	59.906	0.000	0.363	0.387
yr_built	-0.1039	0.002	-44.584	0.000	-0.109	-0.099
lat	0.1808	0.002	91.818	0.000	0.177	0.185
sqft_living15	0.1429	0.006	24.636	0.000	0.131	0.154
waterfront	0.1136	0.005	21.462	0.000	0.103	0.124
view	0.0361	0.002	23.499	0.000	0.033	0.039

Omnibus:	331.104	**Durbin-Watson:**	2.006
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	582.119
Skew:	0.154	**Prob(JB):**	3.93e-127
Kurtosis:	3.845	**Cond. No.**	36.7



Summary

1. **Indep.**:
   A. First: ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long', 'sqft_living15', 'sqft_lot15']
   B. FINAL: ['bathrooms', 'sqft_living', 'floors', 'condition', 'grade', 'yr_built', 'lat', 'sqft_living15', 'waterfront', 'view']
2. DROPPED Indep.: ['bedrooms', 'sqft_lot', 'sqft_above', 'sqft_basement', 'yr_renovated', 'zipcode', 'long', 'sqft_lot15']
3. R^2 :
   A. First: 0.6998109446922831
   B. FINAL: 0.7662667906628482
4. ΔR^2 : 0.06645584597056509 (109.5% better/increase)
5. **Adjusted** R^2 :
   A. First: 0.6994978672836535
   B. FINAL: 0.7661314265525007
6. **Δ (**R^2 **— Adj.** R^2**)**: -0.0001777329828208948 (231.29% better/reduction)
7. **rmse_train__and__rmse_test**:
   A. First: (201452.07674705895, 202982.19517924954) $\implies$ Δ : 1530.1184321905894
   B. FINAL: (0.0554623929476181 9, 0.0558063968074748) $\implies$ Δ : 0.0003440038601292897
8. ΔΔ **rmse_train__and__rmse_test**: -1530.1180881867292 (444796878.62% better/reduction)
9. **Condition No.**:
   A. First: 214442432.5595656
   B. FINAL: 36.71229854965147
10. Δ **Condition No**: -214442395.84726706 (584116062.0% better/reduction)

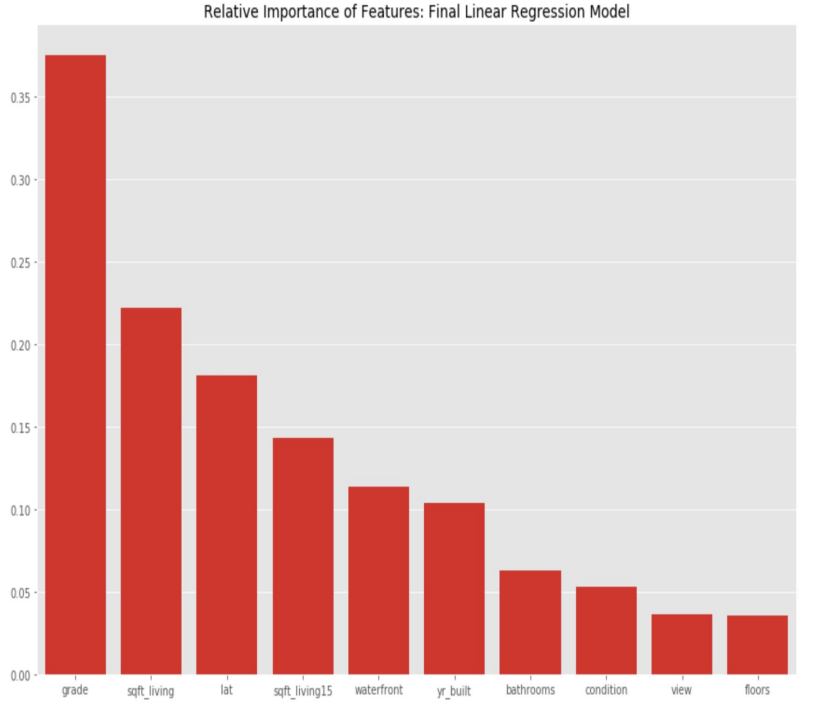
Solving a “Real” Problem

Using the Model

Question 1:

Which features that have the most impact could FEASIBLY be addressed by the seller to increase the sale price of his/her home?

Answering Question 1



But which of those features can a seller *feasibly* address?

Let's first exclude those features about which the seller can do nothing:

1. **lat**: the seller can not uproot his home and relocate it - we assume these are not mobile homes.
2. According to the definitions provided, **sqft_living15** is the square footage of surrounding homes; although its value certainly does influence the sale price of a given home, since it refers to a measure of homes different than the seller's, he can do nothing to change it.
3. **waterfront**: the seller cannot install a river or lake next to his home.
4. **yr_built**: the seller cannot go back in time and change the date of when his house was first built.

The remaining list of features are those that the seller can feasibly address to increase the sale price of his home, in order of importance in determining the sale price of a home in King County:

1. **grade**: since **grade** far outweighs other features in importance, the best strategy should be centered around getting the County Assessor to reassess the seller's home with a higher grade (than the grade originally assessed). But what basis would justify reassessment at a higher grade? The next most important feature seems to be inline with reason.
2. **sqft_living**: livable square footage comes next in terms of importance in sale price of a home; therefore, the strategy should involve renovating the home to increase LIVEABLE square footage.
3. **bathrooms**: bathrooms apparently form a special sub-category of liveable square footage that suggest that bathroom square footage is of the premium variety of liveable square footage.
4. **condition**: it seems reasonable that renovating a home by virtue of adding liveable square footage (other than bathroom), as well as bathroom square footage should improve the condition of the home, and thereby improve the condition. But this is of course dependent on the Assessor and is probably correlated somewhat to **grade**.
5. **view**: having the home viewed AFTER renovation and officially having it listed as viewed in the MLS apparently makes a difference in perceived value. But this should be done after renovation is complete.
6. **floors**: of course adding another floor to the seller's home should increase its value. But is this feasible? It is a large undertaking and could be treated almost equivalently to simply building another home separately. Therefore, given the cost and effort involved to add an entire floor, the suggestion is to avoid this. Also, adding another floor has less impact than the above.

Final list of features that can be directly or indirectly addressed by the seller to increase sale price

1. **sqft_living**
2. **bathrooms**
3. **condition**
4. **grade**
5. **view**

Experiment

Given a hypothetical set of values for the 10 features upon which the model is based, compute the predicted increase in sale price by "enhancement" of the strategic features.

Suppose a hypothetical seller wants to sell his house with the following "initial conditions":

1. **grade** := 4
2. **sqft_living** := 2357 $ft.^2$
3. lat: *<since this is a hypothetical scenario, we just assign the average>*
4. sqft_living15 := 2400
5. waterfront := 0
6. yr_built := 1968
7. **bathrooms** := 2.5
8. **condition** := 3
9. **view** := 0
10. floors := 1

Solving the “Experiment”

Predictors after Scaling/Transformation:

	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	0.1	0.514359	0.5	0.653658	0.0	0.591304	0.580482	0.5	0.0	0.0

1 rows x 10 columns

$\widehat{price}$ **(scaled/transformed):**

	price
0	0.239047

1 rows x 1 columns

Unscaled (but still transformed) $\widehat{price}$**:**

	price
0	12.362232

1 rows x 1 columns

Final: unscaled, not transformed $\widehat{price}$**:**

	price
0	233802.496511

1 rows x 1 columns

Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357	47.46675	2400	0	1968	2.5	3	0	1

Answer: $233,802.49

Building on the Experiment

Question 2:

Suppose the seller wants to get a higher price? What is his/her best course of action?

Answering Question 2

Let's observe how/if **price** increases based on incrementally "enhancing" the identified strategic features.

Let's start with renovating the home via new liveable square-footage construction - e.g. adding new bedroom. The effect will be to increase **sqft_living**.

Building on the Experiment

Question 2a:

By how much can the seller increase the market value of the home if he has another livable room added to his home, say, for example, a modest 10 ft. x 10 ft. (100 $ft.^2$) bedroom?

Answering Question 2a

Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357	47.46675	2400	0	1968	2.5	3	0	1
1	236569.479986	4	2457	47.46675	2400	0	1968	2.5	3	0	1

Thus, by adding an additional $100 ft.^2$ bedroom, the model predicts the sale price of his home will increase by: $2766.98 dollars.

That's actually not much, is it? Considering how much it costs to pay a contractor to build this add-on, the seller will likely be in the red. So far, simply adding an extra $100 ft.^2$ of liveable square footage doesn't seem worth the trouble.

BUT, suppose that the contractor offers the seller the deal of his lifetime to only charge for *his* cost of building materials provided the seller commits to at least 200 sq. ft. of construction, at the contractor's usual cost of 100 dollars per sq. ft. (with a 100 sq. ft. minimum).

That is, if the seller commits to a minimum of 200 sq. ft., for a minimum of 20,000 dollars, the contractor will only charge for the *contractor-cost* of the building materials after that.

Let's suppose for all intents and purposes, building materials *cost the contractor* roughly $20 per square foot. This will include the basics only, such as structural materials - i.e. wood, nails, etc. - as well as drywall and of course eletrical materials to wire the new room. Like I said, it's the deal of a lifetime for the seller.

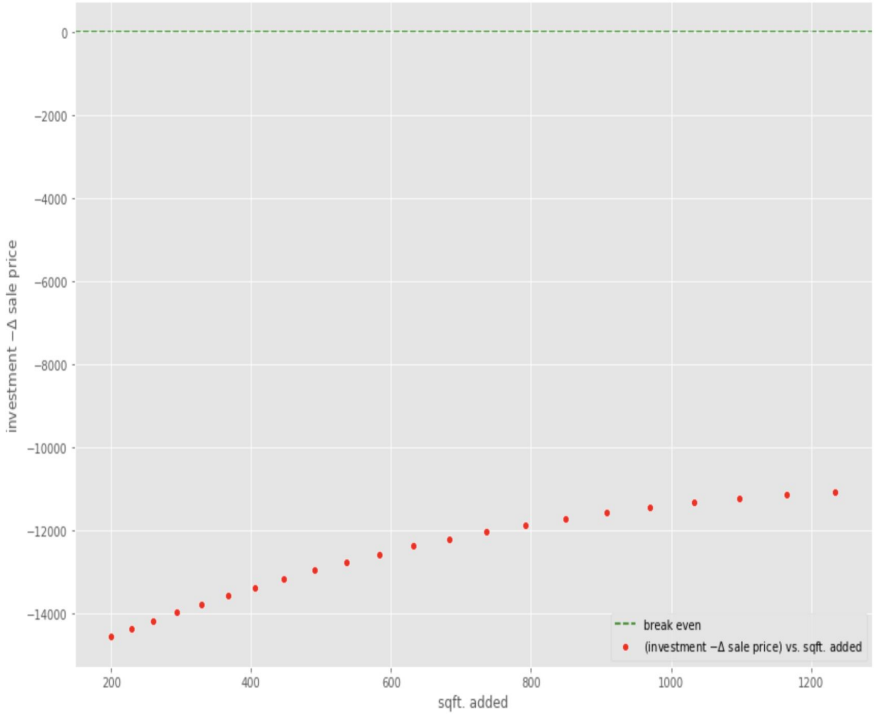
Finally, let's impose the constraint that the seller has a cut-off budget of $40,000 to invest in this deal.

Building on the Experiment

Question 2b:

Can this be accomplished within the seller's budget and, if so, how much will the seller have to invest in order to break even (based on the predicted market value of the renovated home) if he takes advantage of this deal?

Answering Question 2b



Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.000000	47.46675	2400	0	1968	2.5	3	0	1
1	263425.392259	4	3591.969696	47.46675	2400	0	1968	2.5	3	0	1

2 rows x 11 columns

1234.97 $ft.^2$ added

yields Δ sale price of home: $29622.9

requires $40699.39 investment

profit = $-11076.5

Sorry! The investment required to break even based on the predicted price is either infinite or has exceeded the seller's budget ($40000) limit threshold!

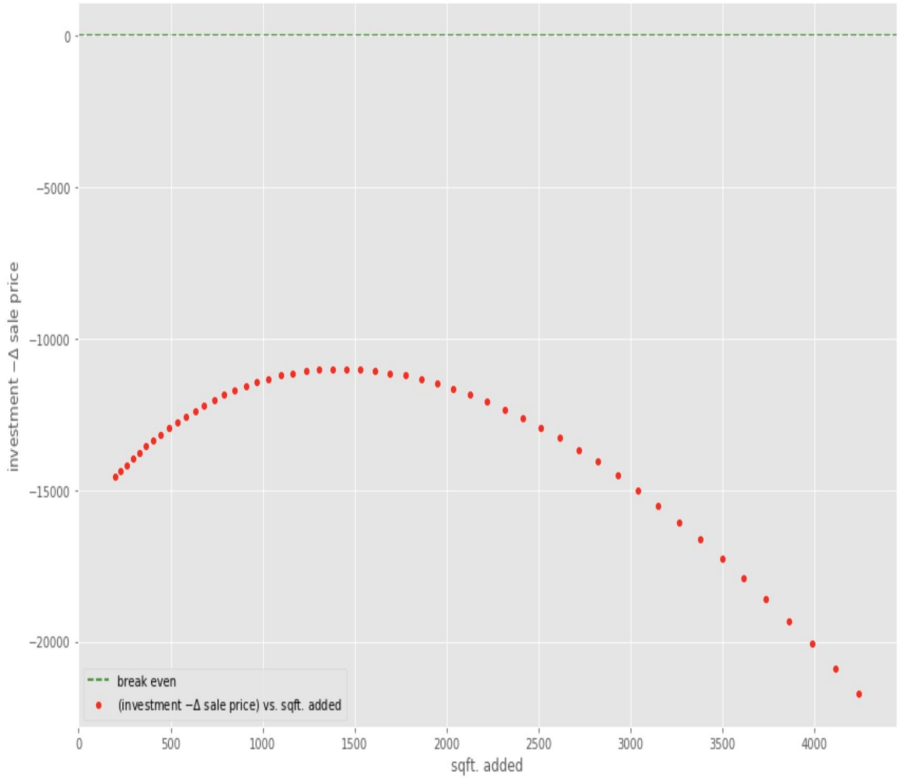
We see that the seller's budget of $40,000 was exceeded after adding about 1200 $ft.^2$ and only netted a profit in sale price of $29,622.90 for a TOTAL profit after investing $40,699.39 in the red at $-11076.5.

It's not looking like a wise investment so far at the outset.

But what if the seller simply dumped more more money into adding more liveable square footage? Will it ever be possible to break even or net an overall profit? (HINT: we will see that it will matter whether or not some of that square footage is a bathroom or not.)

Let's try the model again with a budget of $100,000 just to see...

Answering Question 2b (continued)



Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.000000	47.46675	2400	0	1968	2.5	3	0	1
1	312952.082432	4	6600.497834	47.46675	2400	0	1968	2.5	3	0	1

2 rows x 11 columns

4243.5 $ft.^2$ added

yields Δ sale price of home: $79149.59

requires $100869.96 investment

profit = $-21720.37

Sorry! The investment required to break even based on the predicted price is either infinite or has exceeded the seller's budget ($100000) limit threshold!

YIKES! NOPE! It is definitely not possible.

In fact, it begins to cost the seller more and more to yield less and less difference in sale price.

By the way, the point of "no return", after which the seller's loss will forever increase, is when sqft. added is:

```
g[1][g[0].index(max(g[0]))]
```

```
1454.8225099390856
```

Any square footage beyond 1454 (without enhancing any other feature) results in losses to the seller that just continue to increase!

Thus, **adding liveable square footage (of the non-bathroom variety) and doing nothing more does not look like a wise strategy.**

Answering Question 2b (continued)

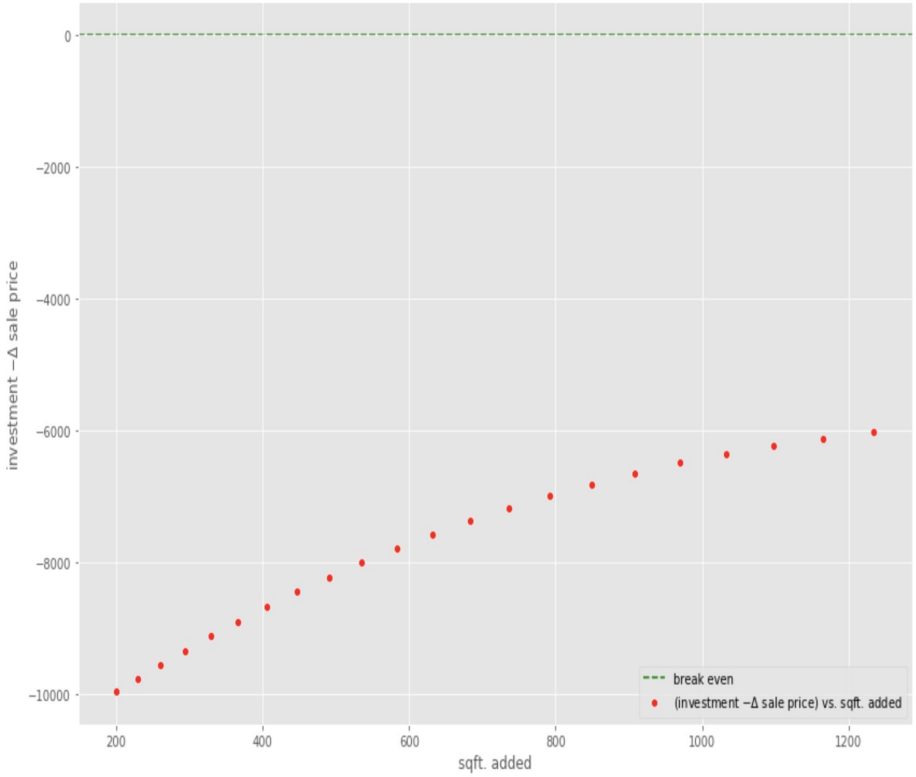
But hold on a minute! We noted initially that our model suggests that **bathrooms** is a strategic feature.

What if we allocated some of the construction effort for liveable square footage toward bathrooms?

Let's leave the numbers in our problem statement as is but do just that - i.e. require the contractor to build bathrooms out of some of that new square footage to be added. Why not? It will require more building materials but does not add any additional square footage per se.

Let's start with converting some of that new square footage to a single bathroom without a shower - i.e. addding **bathrooms** = 0.5.

Answering Question 2b (continued)



Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.000000	47.46675	2400	0	1968	2.5	3	0	1
1	268467.802997	4	3591.969696	47.46675	2400	0	1968	3.0	3	0	1

2 rows x 11 columns

1234.97 $ft.^2$ added

yields Δ sale price of home: $34665.31

requires $40699.39 investment

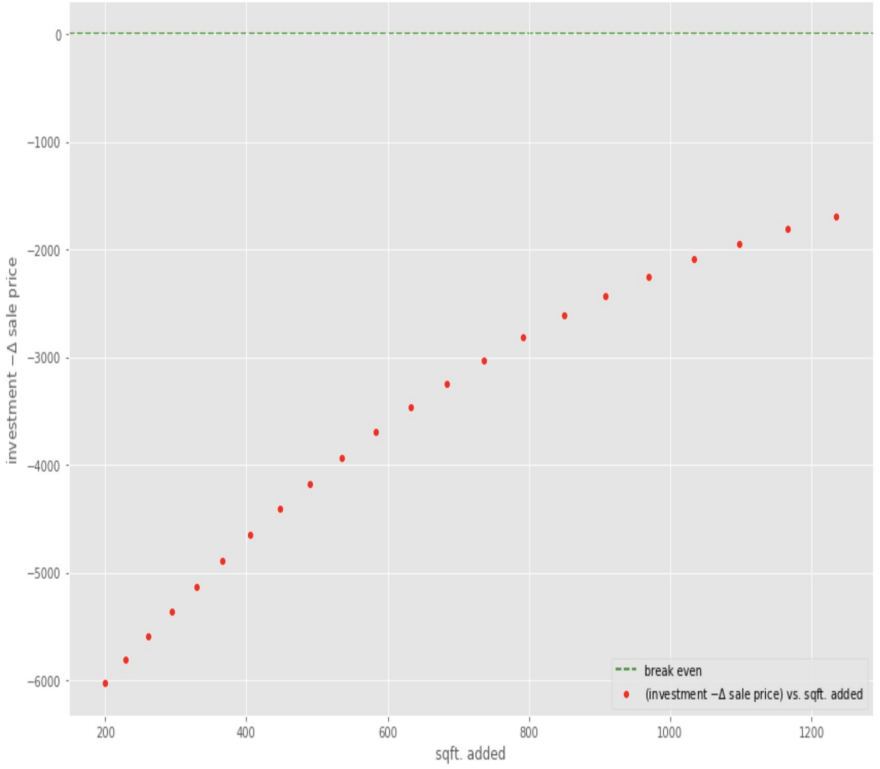
profit = $-6034.09

Sorry! The investment required to break even based on the predicted price is either infinite or has exceeded the seller's budget ($40000) limit threshold!

Classifying some of that square footage as a half-bathroom improved the situation by about $5000 (the first overall profit was $-11076.5)! This looks promising.

Let's add a shower to that bathroom and see how it turns out.

Answering Question 2b (continued)



Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.000000	47.46675	2400	0	1968	2.5	3	0	1
1	272806.328861	4	3591.969696	47.46675	2400	0	1968	3.5	3	0	1

2 rows x 11 columns

1234.97 $ft.^2$ added

yields Δ sale price of home: $39003.83

requires $40699.39 investment

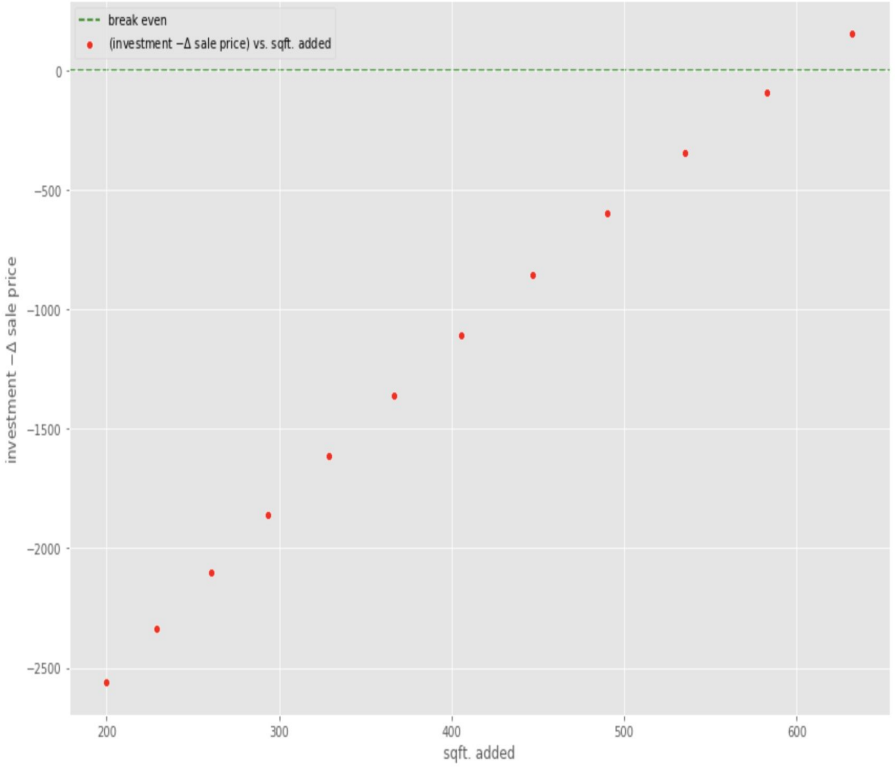
profit = $-1695.56

Sorry! The investment required to break even based on the predicted price is either infinite or has exceeded the seller's budget ($40000) limit threshold!

That still does not break even for the seller. BUT look at the overall proft! The seller is only in the red by $-1695.56 and the trend of this curve appears to be approaching the break-even point.

So let's allocate more of that square footage to a new half-bathroom. So, **bathrooms** = 1.5.

Answering Question 2b (continued)



Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.000000	47.46675	2400	0	1968	2.5	3	0	1
1	262599.256132	4	2989.126984	47.46675	2400	0	1968	4.0	3	0	1

2 rows x 11 columns

632.13 $ft.^2$ added

yields Δ sale price of home: $28796.76

requires $28642.54 investment

profit = $154.22

EUREKA!

A $28,642.54 investment is well below the seller's budget of $40,000 and results in breaking even, with a tiny profit of $154.22! So far, taking the contractor up on his offer to build an add-on constituting an additional 632 $ft.^2$ *with some of that LIVEABLE square-footage allocated to a full bathroom plus a half-bathoom* looks like a GREAT investment.

Suppose that now the seller asks his realtor to update his listing, which results in his home being **viewed** almost immediately.

Building on the Experiment

Question 3:

How much will the fact that his home has now been viewed affect the predicted sale price?

Answering Question 3

Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.00	47.46675	2400	0	1968	2.5	3	0	1
1	262599.331162	4	2989.13	47.46675	2400	0	1968	4.0	3	0	1
2	309976.961027	4	2989.13	47.46675	2400	0	1968	4.0	3	1	1

3 rows x 11 columns

Thus, subsequently having his home officially viewed (and listed as such) after adding some *livable* sq. ft. (with some of it reserved for 1.5 new bathrooms), the model predicts the sale price of his home will increase again by 47377.63 dollars!

That's a total profit of 47531.92 dollars (AFTER his initial investment of $28642.54)! AMAZING!

Finally, after having this work done and having his home listed in the MLS as viewed, the seller requests the King County Assessor to officially reassess his home, with the hope that either the **condition** or the **grade** is upgraded. Suppose first that this results in an increase in the **condition** of his home from 3 to 4.

Building on the Experiment

Question 4:

How much will the fact that his home has been officially upgraded from 3 to 4 in condition affect the predicted sale price?

Answering Question 4

Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.00	47.46675	2400	0	1968	2.5	3	0	1
1	262599.331162	4	2989.13	47.46675	2400	0	1968	4.0	3	0	1
2	309976.961027	4	2989.13	47.46675	2400	0	1968	4.0	3	1	1
3	329415.942799	4	2989.13	47.46675	2400	0	1968	4.0	4	1	1

4 rows x 11 columns

Thus, the official upgrade in *condition* from 3 to 4 results in an increase of 19438.98 dollars from the last predicted sale price

That's a total profit of 66970.91 dollars (after his initial investment)!!!

Finally, suppose Lady Luck has smiled upon the seller, resulting in the County Assessor also upgrading the value of the **grade** of his home from 4 to 5.

Building on the Experiment

Question 5:

How much will the fact that his home has been officially upgraded from 4 to 5 in grade affect the predicted sale price?

Answering Question 5

Predicted *price*, with associated predictors:

	price	grade	sqft_living	lat	sqft_living15	waterfront	yr_built	bathrooms	condition	view	floors
0	233802.496511	4	2357.00	47.46675	2400	0	1968	2.5	3	0	1
1	262599.331162	4	2989.13	47.46675	2400	0	1968	4.0	3	0	1
2	309976.961027	4	2989.13	47.46675	2400	0	1968	4.0	3	1	1
3	329415.942799	4	2989.13	47.46675	2400	0	1968	4.0	4	1	1
4	391330.597372	5	2989.13	47.46675	2400	0	1968	4.0	4	1	1

5 rows x 11 columns

Thus, the official upgrade in *grade* from 4 to 5 results in a virtually UNBELIEVABLE increase of $61914.65 from the last predicted sale price

That's a TOTAL profit of $128885.56 (after his initial investment)!!!

Concluding the Experiment

Sit back and absorb that for a second.

It seems unbelievable but that's what the model predicts and we can rely on this prediction, at least stastically speaking.

But, in order to garner that amazing profit, the seller had to do the following, in order:

1. spend an initial investment of $28,642.54 to to have his home renovated by adding an additional 632 $ft.^2$ of livable square footage (increase in **sqft_living**)
   A. note that it was absolutely fundamental to his success to have additional **bathrooms** (1.5 to be exact) built
   B. simply adding non-liveable square-footage or liveable square footage alone without bathrooms would not do the trick
2. make the effort to get the home viewed by potential buyers AFTER the renovation - this may mean additional hidden fees from a realtor since it has the added caveat that the home's **view** flag must be set in the MLS
3. make the effort to get the home reassessed by the King County Assessor - note that there may be additional hidden fees in this effort as well - which will hopefully result in an increased value in **condition** and, MOST IMPORTANTLY, **grade**.

Project Conclusion

Future Considerations: *All the way Back to EDA and Model Building…*

Given the amount of detail and time I spent on mitigating collinearity and manual principal component analysis, I did not spend time on dealing with outliers. It is a trade-off I willingly acquiesced to, especially considering the reliability in the final model produced. But, in the future, I would definitely like to invest some time in a proper study of outliers.

References

Lau, Dr. C. H. (2019). 5 Steps of a Data Science Project Lifecycle. Retrieved from https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492