Zillow Regression Analysis to Inform Purchase Decisions

March 10, 2023

By Andrew Levinton

Overview

- Business Problem Discussion
- Data used to conduct study
- Data Preparation Methodology
- Model Validity Parameters
- Preliminary (Baseline Model) - Flaws, and plans for improvement
- Methodology for improvement of data
- Model 2&3 - Showing the addition of categorical variables and data cleaning
- Presentation, Interpretation, and Recommendations from Final model
- Business questions model can answer
- Plan for Future work

The Data - KC Housing Dataset - Link Below
https://info.kingcounty.gov/assessor/DataDownload/default.aspx.

Columns from dataset

The full list of columns with descriptions from the data can be located in the readme file of the repository.

Length

- 30,155 Data Points
- After Nulls, outliers, and data cleaning approximately 28,004 data points remain.
- ~7% of the data is removed from data cleaning as a result.

Data Timeline

- All house ages are within the years of 1900-2022.
- All house sales in the dataset are in the years of 2021-2022.

Business Problem:

- Zillow is looking to find ways to manage its inventory to curb future costs and understand how to improve pricing.
- Zillow has decided to hire a consulting data scientist to give recommendations on how to enter and behave within the target market.

Business Understanding

- Zillow seeks to focus on the real estate market of the pacific northwest.

- Before looking for inventory, Zillow needs to understand how to determine the opportunity cost.

- Some parameters to look at are: Housing age, location, condition, and attributes like square footage.

Data Preparation

- Data must be numerical (float or int) in order to be utilized in a linear model.
- Categorical variables
  - Data columns with measurable quantities are converted to integers or floats.
  - Categorical variables that are not measurable undergo OneHotEncoding
- Ensure there are no missing values.
- The target variable 'price' is eventually square-root-transomed as part of the model fitting process.

Assumptions for Model and Checks

Assumption check:

- Is it linear?
- Is it normal?
- Is it homoscedastic?
- Is it Multicollinear?

To check for assumptions, look at:

- Scatter plots
- Histograms
- QQ Plots
- Correlation Coefficients
- Statsmodel p-values to test if the feature is statistically significant
- Variance Inflation Factor
- Durbin-Watson Score

Model #1 - Numerical Data only

Concerns/Observations

- The skew score is 10.060, indicating that this model is heavily skewed. (outside of the acceptable -2 to 2)
- Model contains some independent variables yielding a pvalue greater than 0.05, indicating statistical insignificance.
- R-Squared value is .514, indicating 51.4% can be explained by this model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.514
Model:                            OLS   Adj. R-squared:                  0.514
Method:                 Least Squares   F-statistic:                     1814.
Date:                Tue, 07 Mar 2023   Prob (F-statistic):               0.00
Time:                        13:10:41   Log-Likelihood:            -4.3109e+05
No. Observations:               29200   AIC:                         8.622e+05
Df Residuals:                   29182   BIC:                         8.624e+05
Df Model:                          17
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -6.16e+07      4e+06    -15.396      0.000   -6.94e+07   -5.38e+07
bedrooms      -1.136e+05   5091.194    -22.322      0.000   -1.24e+05   -1.04e+05
bathrooms      9.389e+04   7527.079     12.474      0.000    7.91e+04    1.09e+05
sqft_living     207.5950     17.071     12.161      0.000     174.135     241.055
sqft_lot          0.2667      0.063      4.265      0.000       0.144       0.389
floors        -1.476e+05   9568.312    -15.421      0.000   -1.66e+05   -1.29e+05
condition      5.315e+04   5778.105      9.198      0.000    4.18e+04    6.45e+04
grade          2.149e+05   5521.008     38.916      0.000    2.04e+05    2.26e+05
sqft_above      270.4146     17.425     15.519      0.000     236.262     304.568
sqft_basement    80.8679     12.893      6.272      0.000      55.596     106.140
sqft_garage    -164.9199     18.061     -9.131      0.000    -200.320    -129.520
sqft_patio      193.5427     16.684     11.600      0.000     160.841     226.244
yr_built      -2899.2445    190.203    -15.243      0.000   -3272.051   -2526.438
yr_renovated     68.9239      9.331      7.386      0.000      50.634      87.214
lat            1.344e+06   2.68e+04     50.165      0.000    1.29e+06     1.4e+06
long          -1.822e+04   3.04e+04     -0.599      0.549   -7.78e+04    4.13e+04
month          1.957e+04   1.28e+04      1.529      0.126   -5515.883    4.47e+04
day_of_year   -1215.8907    420.005     -2.895      0.004   -2039.120    -392.662
==============================================================================
Omnibus:                    46855.092   Durbin-Watson:                   1.915
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         91559041.827
Skew:                          10.060   Prob(JB):                         0.00
Kurtosis:                     276.586   Cond. No.                     6.92e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.92e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Improvements to be made to model #1:

1 - Pvalues	• Dropping of variables with p_values greater than 0.05
2 - Outlier Removal	• Removal of Outliers to address skewness
3 - Categorical Data	• Addition of Categorical Variables, to be one hot encoded

Model #2 - Numerical and Categorical Data

Changes made to model:

- Addition of categorical variables. Some changed to booleans/numerical values, some onehotencoded.
- Removal of outliers from data

Observed Changes/Concerns:

- Improved R-Squared: 62.2%
- Skew Score dramatically improved: 0.577
- Still variables with pvalue > 0.05

OLS Regression Results

Dep. Variable:	price	R-squared:	0.622
Model:	OLS	Adj. R-squared:	0.622
Method:	Least Squares	F-statistic:	1534.
Date:	Tue, 07 Mar 2023	Prob (F-statistic):	0.00
Time:	13:17:55	Log-Likelihood:	-3.9347e+05
No. Observations:	28004	AIC:	7.870e+05
Df Residuals:	27973	BIC:	7.873e+05
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.891e+07	2.1e+06	-13.763	0.000	-3.3e+07	-2.48e+07
bedrooms	-1.261e+04	2647.400	-4.764	0.000	-1.78e+04	-7423.552
bathrooms	3.438e+04	3907.140	8.800	0.000	2.67e+04	4.2e+04
sqft_living	137.7492	8.974	15.350	0.000	120.160	155.338
sqft_lot	0.3540	0.036	9.708	0.000	0.283	0.426
floors	-2.695e+04	4927.866	-5.468	0.000	-3.66e+04	-1.73e+04
condition	5.94e+04	2922.086	20.326	0.000	5.37e+04	6.51e+04
grade	1.469e+05	2888.537	50.859	0.000	1.41e+05	1.53e+05
sqft_above	98.6873	9.222	10.701	0.000	80.611	116.763
sqft_basement	9.0654	6.729	1.347	0.178	-4.123	22.254
sqft_garage	-14.9218	9.345	-1.597	0.110	-33.238	3.394
sqft_patio	52.5853	8.886	5.918	0.000	35.168	70.002
yr_built	-2128.0746	98.282	-21.653	0.000	-2320.712	-1935.437
yr_renovated	29.5335	4.828	6.117	0.000	20.070	38.997
lat	1.305e+06	1.35e+04	96.806	0.000	1.28e+06	1.33e+06
long	2.434e+05	1.59e+04	15.356	0.000	2.12e+05	2.74e+05
month	1.988e+04	6406.701	3.104	0.002	7326.324	3.24e+04
day_of_year	-1128.1162	210.292	-5.365	0.000	-1540.298	-715.934
sewer_PRIVATE RESTRICTED	1.742e+05	1.37e+05	1.268	0.205	-9.5e+04	4.43e+05
sewer_PUBLIC	5.498e+04	6113.519	8.993	0.000	4.3e+04	6.7e+04
sewer_PUBLIC RESTRICTED	-2.283e+04	2.17e+05	-0.105	0.916	-4.48e+05	4.02e+05
heat_source_Electricity/Solar	-3.437e+04	4.12e+04	-0.834	0.404	-1.15e+05	4.64e+04
heat_source_Gas	3.284e+04	4947.829	6.638	0.000	2.31e+04	4.25e+04
heat_source_Gas/Solar	1.564e+05	3.38e+04	4.634	0.000	9.03e+04	2.23e+05
heat_source_Oil	-1.553e+04	7536.189	-2.060	0.039	-3.03e+04	-756.860
heat_source_Oil/Solar	-4.439e+04	1.53e+05	-0.290	0.772	-3.45e+05	2.56e+05
heat_source_Other	9.011e+04	7.06e+04	1.276	0.202	-4.83e+04	2.29e+05
waterfront	1.227e+05	1.82e+04	6.751	0.000	8.71e+04	1.58e+05
nuisance	-2.687e+04	5007.274	-5.366	0.000	-3.67e+04	-1.71e+04
view	6.205e+04	2654.342	23.375	0.000	5.68e+04	6.72e+04
greenbelt	9.809e+04	1.19e+04	8.255	0.000	7.48e+04	1.21e+05

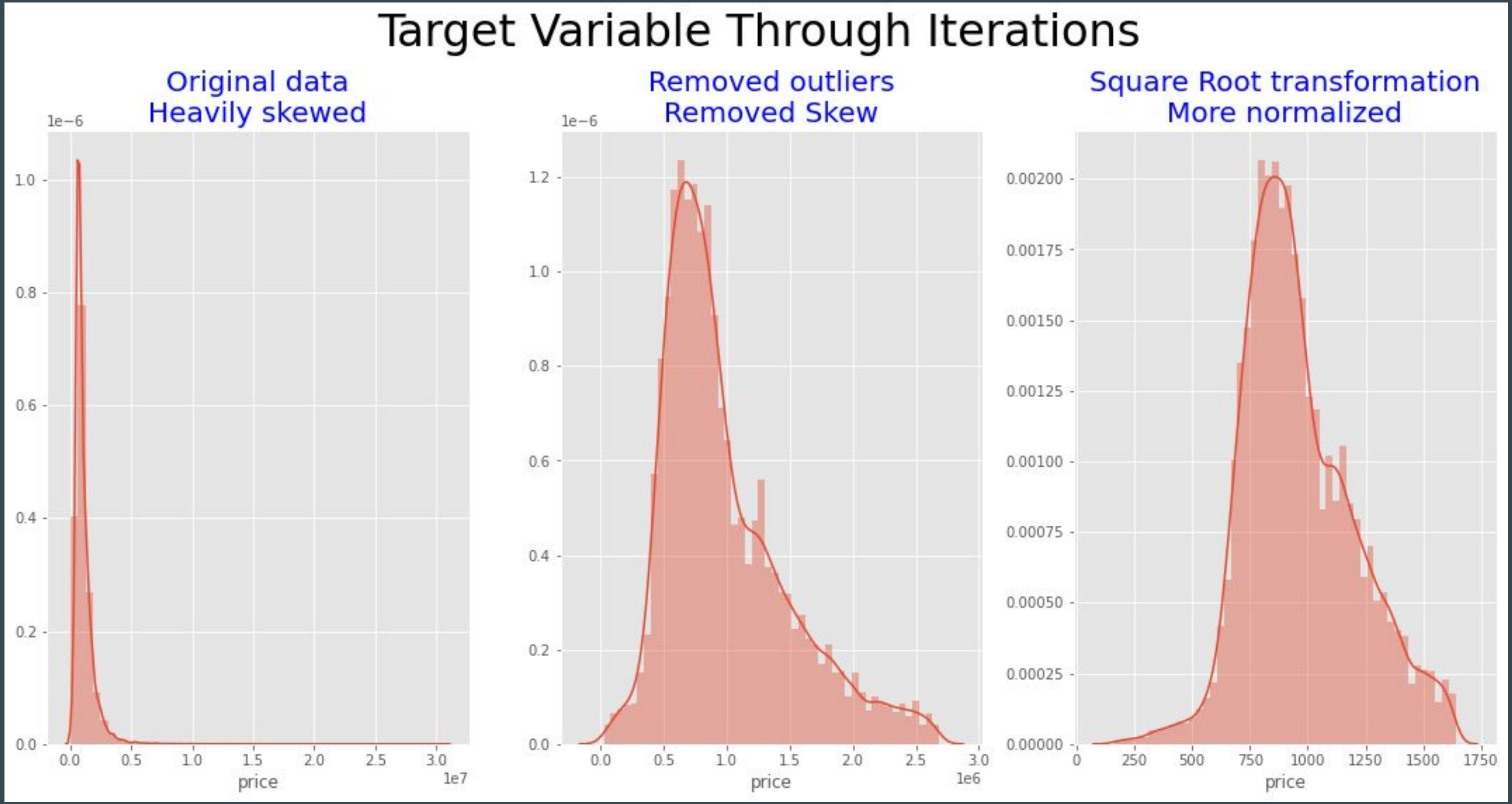
Omnibus:	3918.983	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20352.924
Skew:	0.577	Prob(JB):	0.00
Kurtosis:	7.014	Cond. No.	6.60e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.6e+07. This might indicate that there are strong multicollinearity or other numerical problems.

Improvements to be made to model #2:

1 - Data Transformation	• Dropping of variables with p_values greater than 0.05
2 - Pvalues	• Dropping Pvalues > 0.05
3 - Categorical Data	• Addition of Categorical Variables, to be one hot encoded, specifically the waterfront
4 - Dropping of collinear data	• Checking and dropping variables with high variance inflation factors to address collinearity

Target Variable Transformation - Square root of price



Final Model - Numerical and Categorical Data with waterfronts

Changes made to model:

- Addition of waterfront data - one hot encoded.
- Square root transformation to normalize data
- Dropping all variables of p-value > 0.05, high VIF (> 10)

Observed Changes/Concerns:

- All variables are statistically significant (pvalue <0.05)

OLS Regression Results

Dep. Variable:	price	R-squared:	0.626
Model:	OLS	Adj. R-squared:	0.625
Method:	Least Squares	F-statistic:	1949.
Date:	Tue, 07 Mar 2023	Prob (F-statistic):	0.00
Time:	17:27:21	Log-Likelihood:	-1.7948e+05
No. Observations:	28004	AIC:	3.590e+05
Df Residuals:	27979	BIC:	3.592e+05
Df Model:	24		
Covariance Type:	nonrobust		

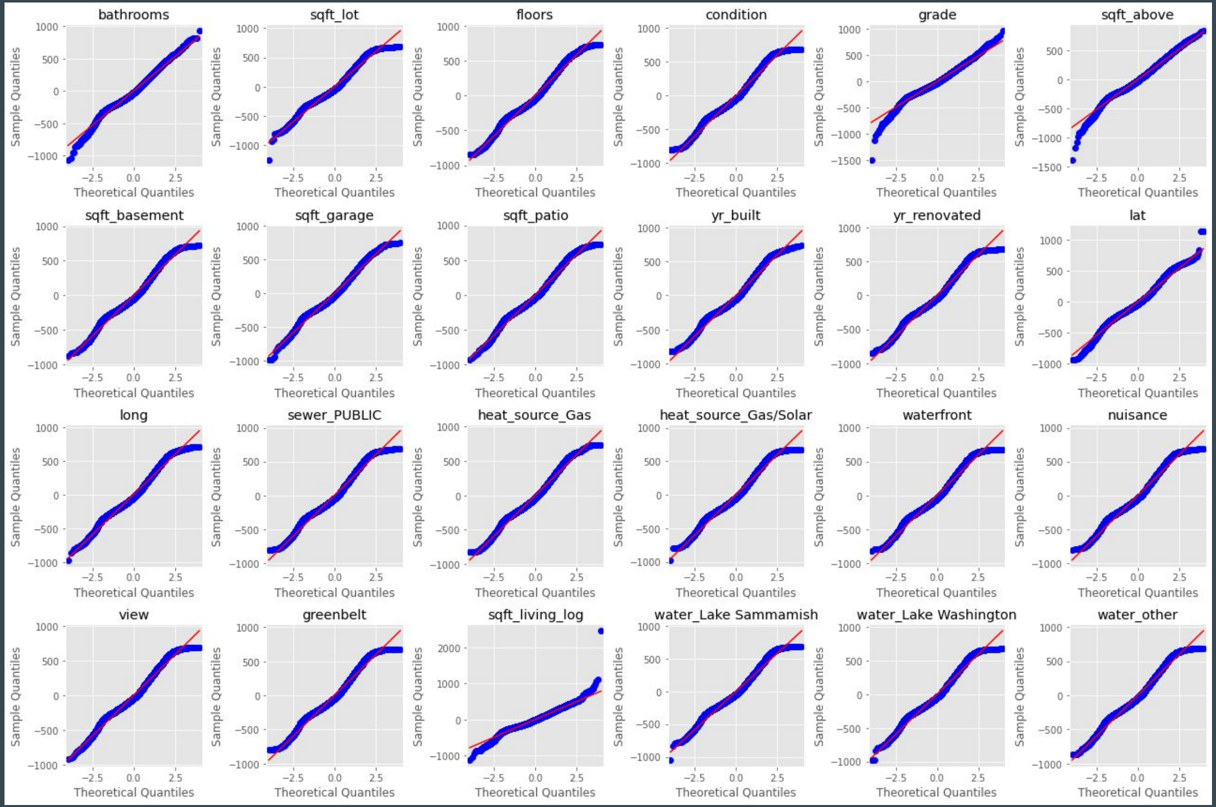
	coef	std err	t	P>|t|	[0.025	0.975]
const	929.1767	3.411	272.414	0.000	922.491	935.862
bathrooms	17.8321	1.475	12.086	0.000	14.940	20.724
sqft_lot	10.7765	0.965	11.166	0.000	8.885	12.668
floors	-5.7693	1.272	-4.537	0.000	-8.262	-3.277
condition	23.4380	0.988	23.717	0.000	21.501	25.375
grade	69.7639	1.447	48.208	0.000	66.927	72.600
sqft_above	63.9056	2.476	25.815	0.000	59.053	68.758
sqft_basement	17.4519	1.510	11.554	0.000	14.491	20.412
sqft_garage	-3.1052	1.227	-2.531	0.011	-5.509	-0.701
sqft_patio	7.7906	0.986	7.903	0.000	5.858	9.723
yr_built	-26.3626	1.462	-18.037	0.000	-29.227	-23.498
yr_renovated	6.5683	0.945	6.948	0.000	4.715	8.421
lat	100.3684	1.006	99.778	0.000	98.397	102.340
long	11.7060	1.139	10.281	0.000	9.474	13.938
sewer_PUBLIC	5.4272	1.075	5.050	0.000	3.321	7.533
heat_source_Gas	9.3428	0.976	9.570	0.000	7.429	11.256
heat_source_Gas/Solar	3.5477	0.884	4.014	0.000	1.816	5.280
waterfront	7.0093	0.958	7.315	0.000	5.131	8.887
nuisance	-5.6582	0.903	-6.268	0.000	-7.428	-3.889
view	23.0579	0.997	23.117	0.000	21.103	25.013
greenbelt	7.5699	0.898	8.427	0.000	5.809	9.331
sqft_living_log	15.1773	2.535	5.987	0.000	10.209	20.146
water_Lake Sammamish	137.7419	6.033	22.833	0.000	125.918	149.566
water_Lake Washington	-22.6046	7.618	-2.967	0.003	-37.537	-7.672
water_other	34.1416	3.534	9.660	0.000	27.214	41.069

Omnibus:	3707.653	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35770.221
Skew:	-0.301	Prob(JB):	0.00
Kurtosis:	8.504	Cond. No.	21.4

VIFS and QQplots

Variable	VIF
bathrooms	2.820882
sqft_lot	1.206917
floors	2.095106
condition	1.265521
grade	2.708681
sqft_above	7.941105
sqft_basement	2.955260
sqft_garage	1.949919
sqft_patio	1.259183
yr_built	2.767781
yr_renovated	1.158123
lat	1.311003
long	1.645893
sewer_PUBLIC	1.494486
heat_source_Gas	1.235073
heat_source_Gas/Solar	1.012068
waterfront	1.189304
nuisance	1.056119
view	1.288704
greenbelt	1.045747
sqft_living_log	8.325701
water_Lake Sammamish	1.133158
water_Lake Washington	1.159873
water_other	1.007988

All VIFS < 10, most < 3



Residuals against model appear to meet linearity assumption

Conclusion and Interpretation

- Latitude of the house coefficient suggests that houses located further north tend to have higher prices. Water proximity, with the Lake Sammamish coefficient suggests that houses located near this lake tend to have much higher prices than other houses. Lake Washington variable has a negative coefficient, indicating that houses located near this lake tend to have lower prices than other houses.
- Grade, square footage of the house apart from basement, and the condition all have coefficients greater than 20. The number of bathrooms, square footage of the basement, and the size of the view from the house are also important, with coefficients greater than 15.
- The presence of a nuisance nearby have negative coefficients, indicating that houses located near nuisances tend to have lower prices. The year the house was built has a negative coefficient, suggesting that older houses tend to have lower prices. With that, the longitude indicates that houses further West are cheaper as well.

- Overall, these results suggest that there are many factors that contribute to the price of a house, and that location, house size and quality, and the presence of nearby amenities all play important roles in determining the square root of house prices.

Recommendations

- **Look at properties that are near Lake Sammish or that are further north that also is accompanied with a waterfront.**
- **Since the grade, condition, and number of bathrooms appear positively correlated to the price it would make sense to try and buy older homes in the aforementioned areas as older homes tend to be cheaper in terms of price.**
- **Taking these homes and ensuring the grade and condition are of high quality through either pre-assessed purchases or renovations, along with possibly adding bathrooms can raise the price for resell value.**
- **Houses towards the west as well as ones that present nuisances clearly result in lower prices, so my recommendation would be to avoid buying houses that fit these parameters as it may result in "holding the bag" scenarios leading to longer times held with inventory.**

Questions model can answer:

- Should the house be on a waterfront?
- How far north should the houses be?
- What age should the house have?
- What level of renovations need to be performed on the houses, and when?
- Will the house price be affected by common nuisances? (eg. noise, construction, bugs)
- How far west should the house be before one should lose interest of the purchase?

Future Work

- In the future work, it is worth revisiting the value of the homes on the remaining waterfronts and seeing if there is any statistical significance. More exploration is needed but was not ready to be presented at this time.
- The views that are highlighted in the column_names.md documentation can be explored and onehotencoded and could be a potential candidate feature.
- Jarque Beras score and outliers of the dataset should be further explored. The use of 3 standard deviations from the mean being the metric for outliers could be expanded slightly as it appears this was still affected in a major way.
- Any independent variables that presented with a Variance Inflation factor above 5 should be looked at again to see if multicollinearity is an issue with these particular variables.