KING COUNTY HOUSING

REGRESSION ANALYSIS

Kyunghwan William Kim

May 27, 2022

Table of Contents

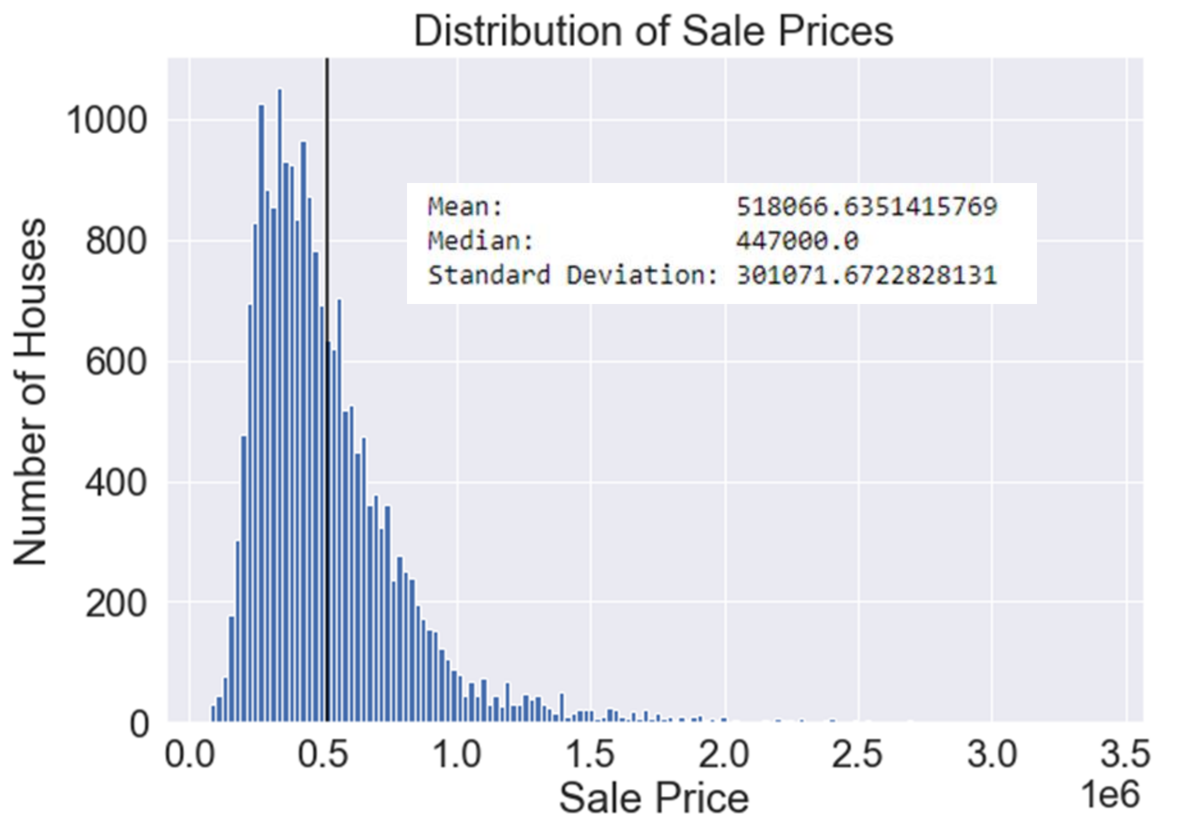
- **01: Overview**
- **02: Business Problem**
- **03: Methodology**
- **04. Recommendations**
- **05: Modeling**
- **06. Summary**
- **07. Future Work**

Overview

- A young couple is planning on selling their home, they want to increase the home value as much as possible but have limited capital for renovations. The couple decided to use Multiple Linear Regression Modeling to analyze and predict house sales in King County based on certain features or variables, so that they can be used to make profitable decisions.
- After careful evaluation and various iterations of our linear regression models, we have determined that square feet of living space, building grade, and number of bathrooms are the most correlated with a higher selling house price.

Business Problem

What features have the greatest impact on the price of a house?



```
Mean:                518066.6351415769
Median:              447000.0
Standard Deviation:  301071.6722828131
```

METHODOLOGY

Step 1: Acquire/Import Data

Step 2: Understand Business needs

Step 3: Exploratory Data Analysis – Explore and clean data

Step 4: Prepare Data for Modeling

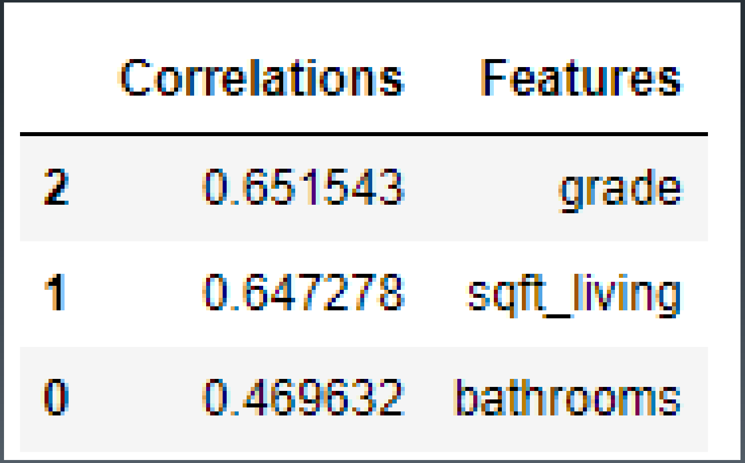
Step 5: Modeling

Step 6: Evaluate and Verify Modeling results

Recommendations

What features have the highest impact to the home price?

The grade, square-footage of living space, and the number of bathrooms are the features with the highest impact.

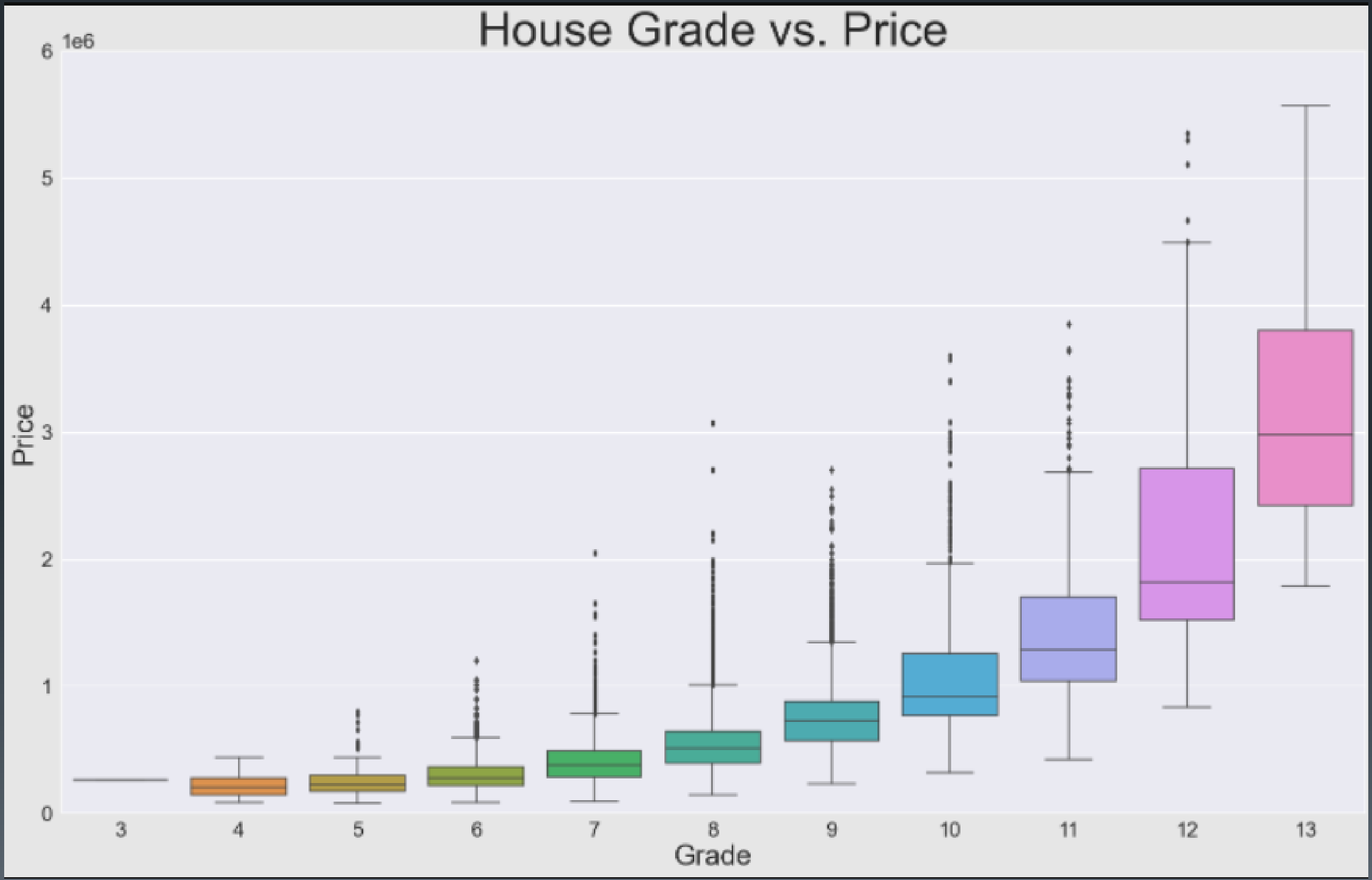


	Correlations	Features
2	0.651543	grade
1	0.647278	sqft_living
0	0.469632	bathrooms

Recommendation #1: Grade

Houses with high quality construction grades corresponds with house with higher values

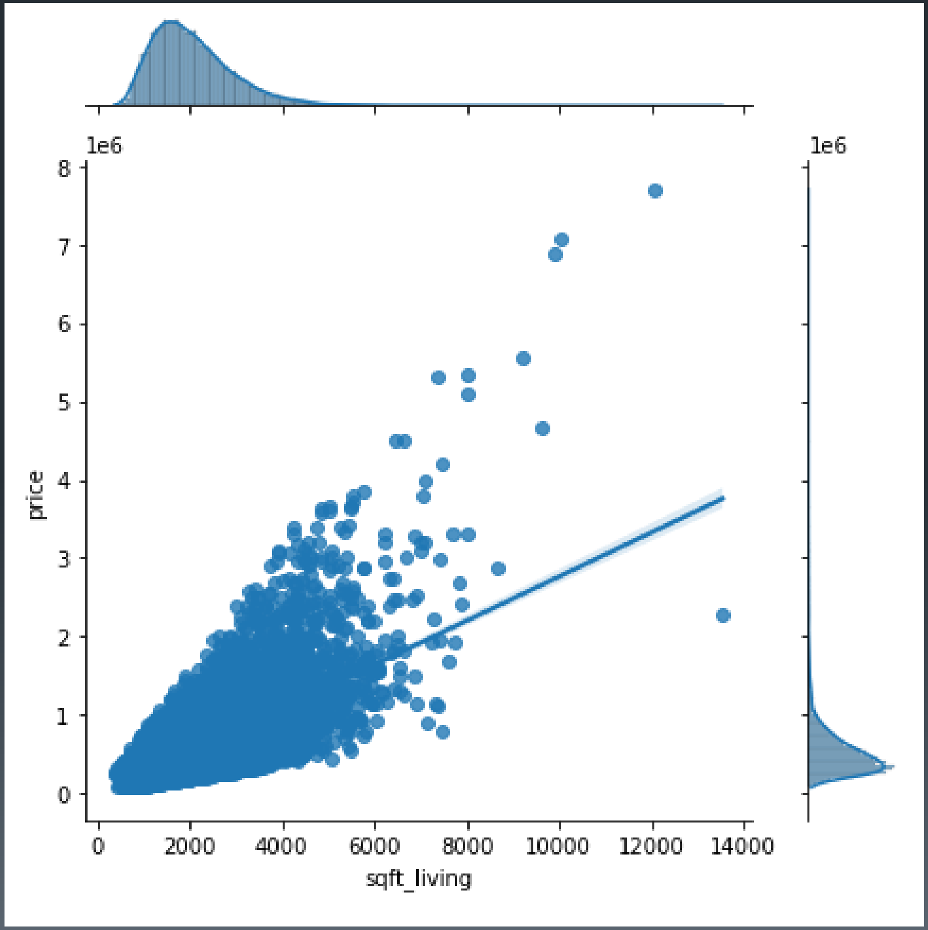
Grade +1 = Increase of about $134,307.61



Recommendation #2: sqft living

Houses with more sqft of living corresponds with houses with higher values

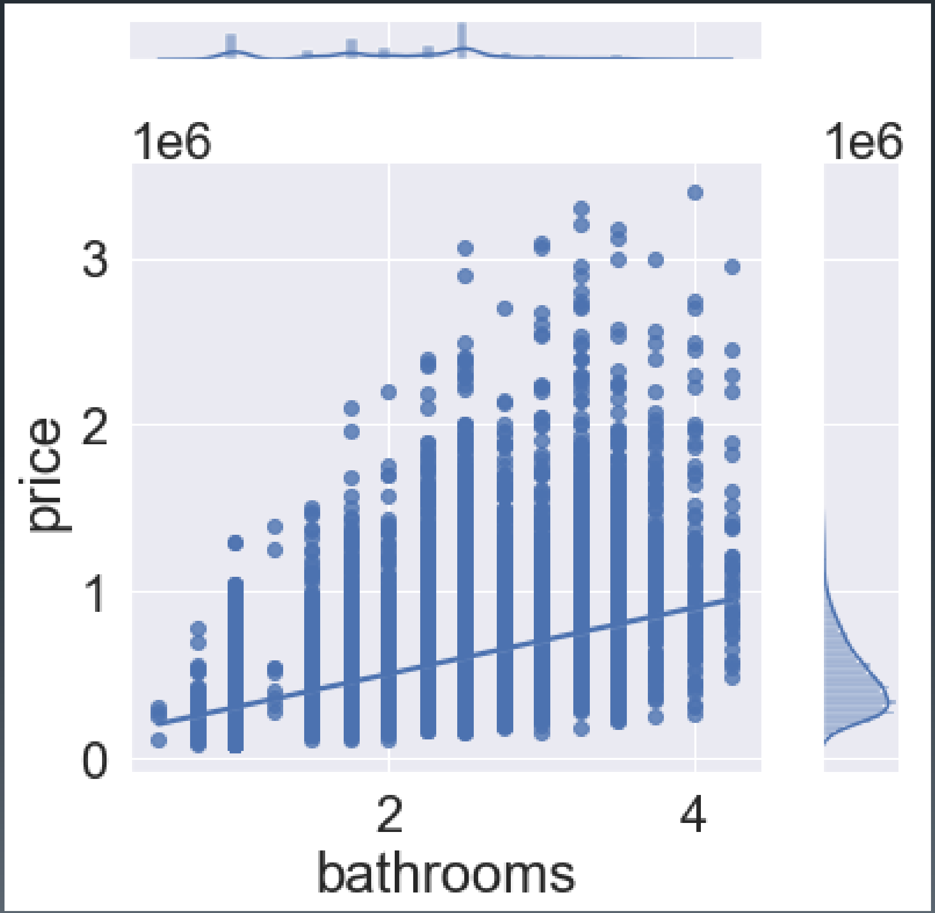
Sqft_livng + 1 = Increase of about $134.68



Recommendation #3: Bathrooms

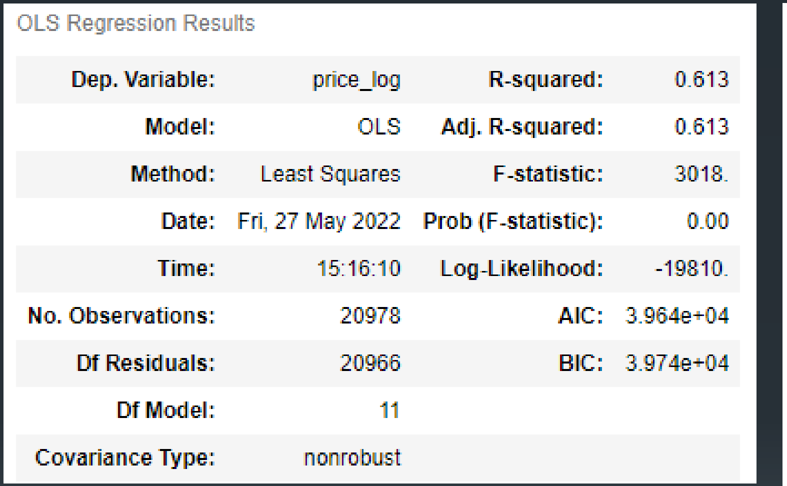
Increasing the number of bathrooms will help increase the house price

Bathroom + 1 = Increase of about $45,405.00



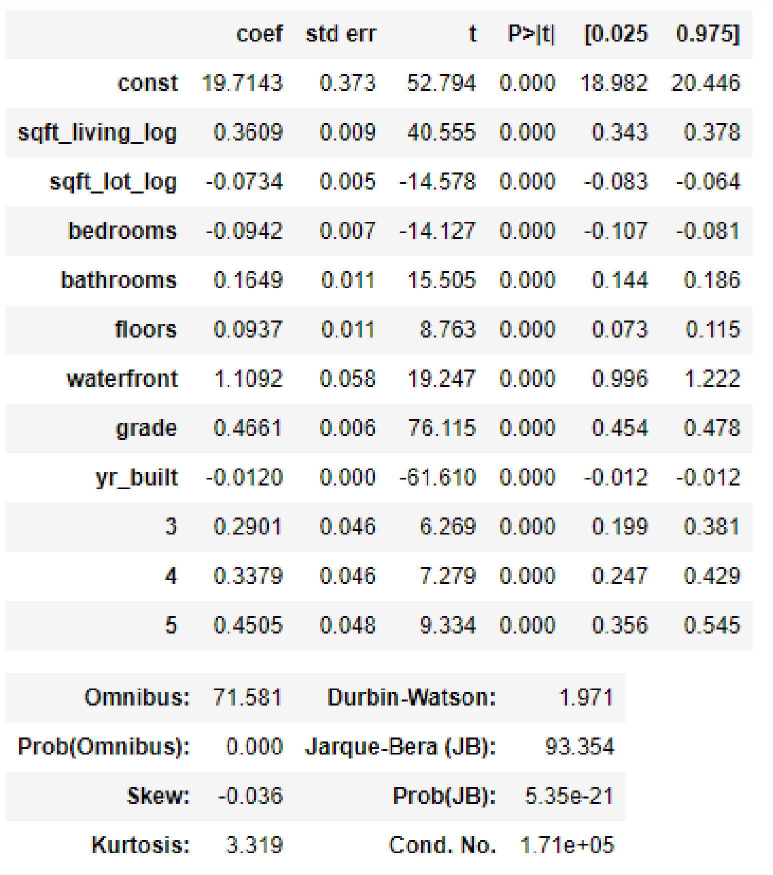
Modeling

After many iterations our final model’s r-squared value was 0.613, indicating that the model can account for about 61% of the variability of price around its mean.



OLS Regression Results

Dep. Variable:	price_log	**R-squared:**	0.613
Model:	OLS	**Adj. R-squared:**	0.613
Method:	Least Squares	**F-statistic:**	3018.
Date:	Fri, 27 May 2022	**Prob (F-statistic):**	0.00
Time:	15:16:10	**Log-Likelihood:**	-19810.
No. Observations:	20978	**AIC:**	3.964e+04
Df Residuals:	20966	**BIC:**	3.974e+04
Df Model:	11		
Covariance Type:	nonrobust		



	coef	std err	t	P>|t|	[0.025	0.975]
const	19.7143	0.373	52.794	0.000	18.982	20.446
sqft_living_log	0.3609	0.009	40.555	0.000	0.343	0.378
sqft_lot_log	-0.0734	0.005	-14.578	0.000	-0.083	-0.064
bedrooms	-0.0942	0.007	-14.127	0.000	-0.107	-0.081
bathrooms	0.1649	0.011	15.505	0.000	0.144	0.186
floors	0.0937	0.011	8.763	0.000	0.073	0.115
waterfront	1.1092	0.058	19.247	0.000	0.996	1.222
grade	0.4661	0.006	76.115	0.000	0.454	0.478
yr_built	-0.0120	0.000	-61.610	0.000	-0.012	-0.012
3	0.2901	0.046	6.269	0.000	0.199	0.381
4	0.3379	0.046	7.279	0.000	0.247	0.429
5	0.4505	0.048	9.334	0.000	0.356	0.545

Omnibus:	71.581	**Durbin-Watson:**	1.971
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	93.354
Skew:	-0.036	**Prob(JB):**	5.35e-21
Kurtosis:	3.319	**Cond. No.**	1.71e+05

Summary

Our recommendations: What key features increase the value of a house?

1. **Build or renovate a higher quality home that will produce a higher house price.**

2. **Build a larger house – increase the square footage of a house**

3. **Increase the number of bathrooms**

Together, square footage, grade and bathrooms are the best predictors of a house's price in King County. Homeowners who are interested in selling their homes at a higher price should focus on expanding square footage and improving the quality of construction. When expanding square footage, homeowners should consider building additional bathrooms, as this analysis suggests that number of bathrooms is positively related to price.

Future Work

A good next step here would be to start trying to figure out why our outliers behave the way they do. Maybe there is some information we could extract from the text features that are currently not part of the model

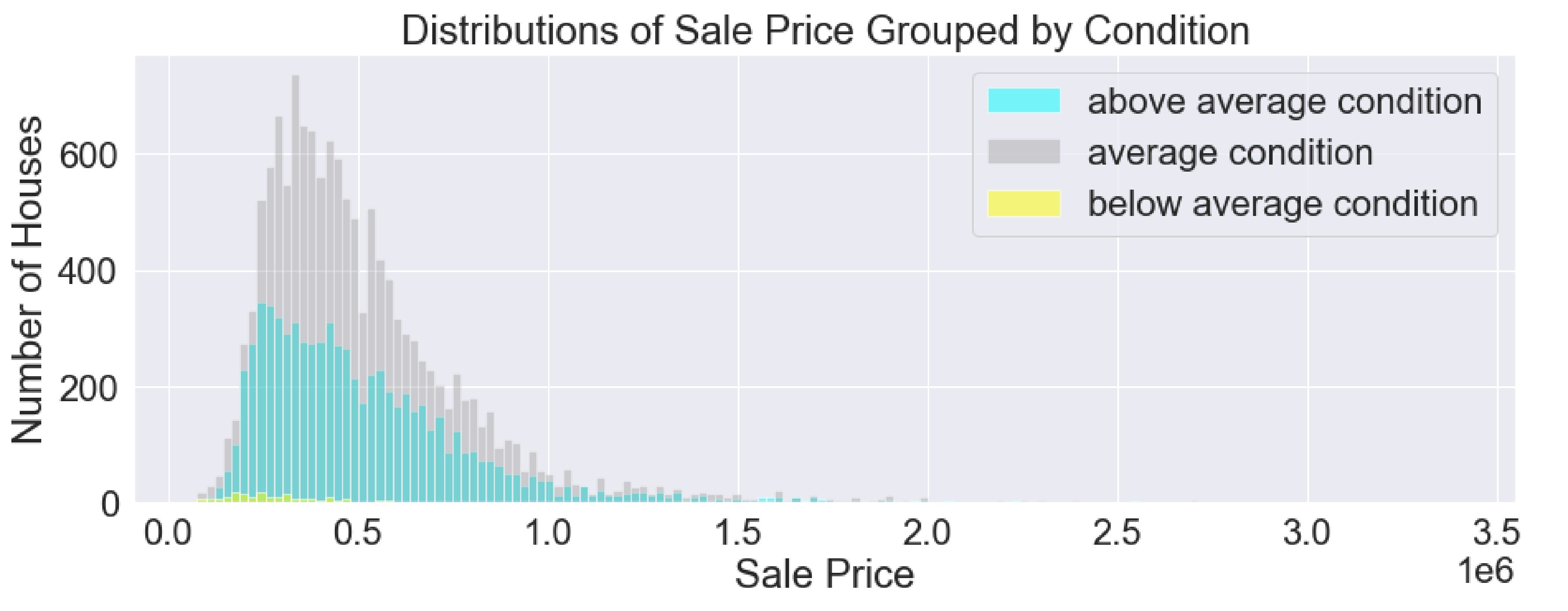
We can also try to improve the accuracy by reducing the noise in the data. Additionally we can simplify the model by grouping low-impact categorical features.

Thank you!

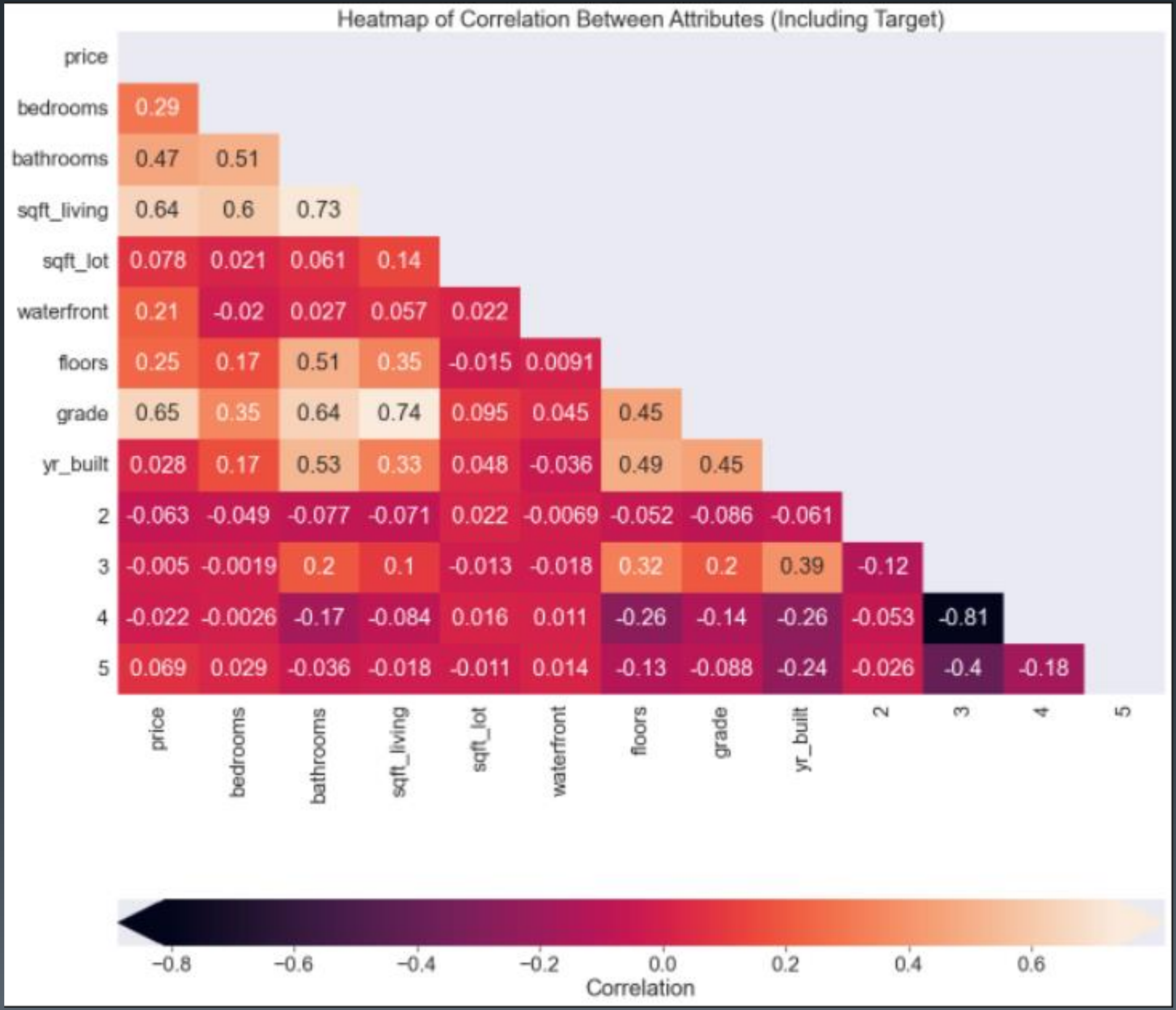
Kyunghwan William Kim
khwilliamkim@outlook.com

*Please see our appendix for further investigation of the topics discussed.

APPENDIX – Condition feature

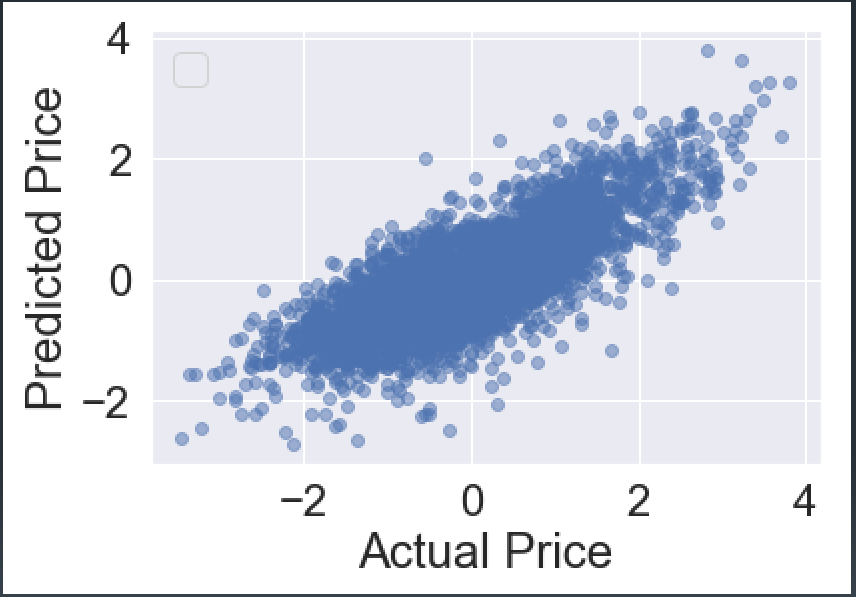


APPENDIX – Correlation Heatmap

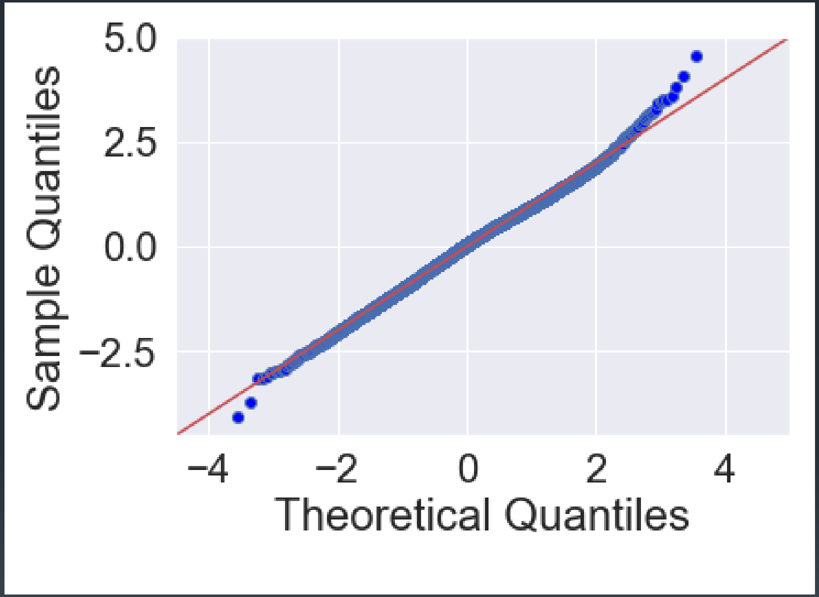


	price	bedrooms	bathrooms	sqft_living	sqft_lot	waterfront	floors	grade	yr_built	2	3	4	5
price													
bedrooms	0.29												
bathrooms	0.47	0.51											
sqft_living	0.64	0.6	0.73										
sqft_lot	0.078	0.021	0.061	0.14									
waterfront	0.21	-0.02	0.027	0.057	0.022								
floors	0.25	0.17	0.51	0.35	-0.015	0.0091							
grade	0.65	0.35	0.64	0.74	0.095	0.045	0.45						
yr_built	0.028	0.17	0.53	0.33	0.048	-0.036	0.49	0.45					
2	-0.063	-0.049	-0.077	-0.071	0.022	-0.0069	-0.052	-0.086	-0.061				
3	-0.005	-0.0019	0.2	0.1	-0.013	-0.018	0.32	0.2	0.39	-0.12			
4	-0.022	-0.0026	-0.17	-0.084	0.016	0.011	-0.26	-0.14	-0.26	-0.053	-0.81		
5	0.069	0.029	-0.036	-0.018	-0.011	0.014	-0.13	-0.088	-0.24	-0.026	-0.4	-0.18	

APPENDIX – Model Validation

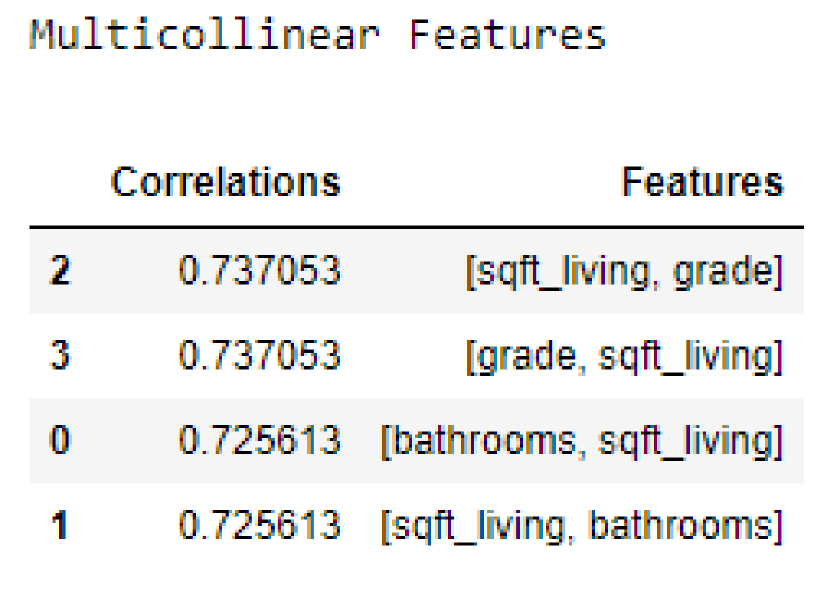


*Our model has a linear relationship



*Normality assumption is satisfied

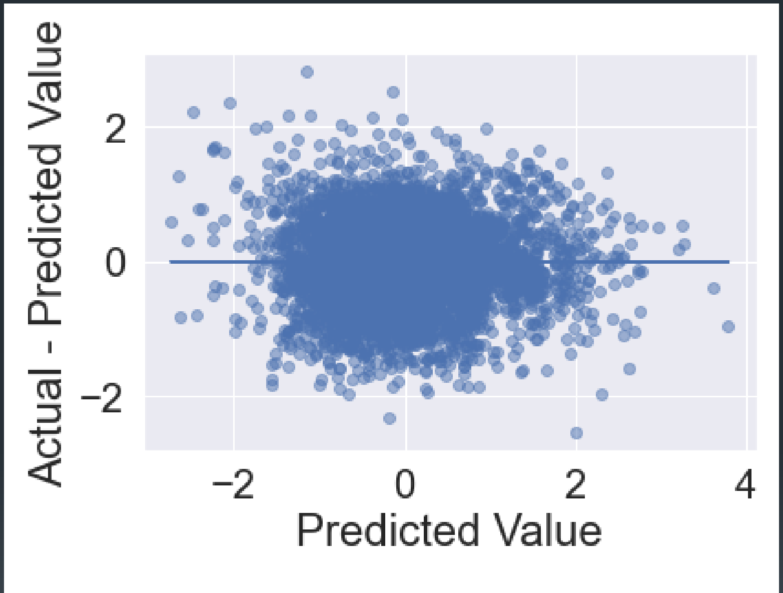
APPENDIX – Model Validation



```
Multicollinear Features
```

	Correlations	Features
2	0.737053	[sqft_living, grade]
3	0.737053	[grade, sqft_living]
0	0.725613	[bathrooms, sqft_living]
1	0.725613	[sqft_living, bathrooms]

***Correlations are under 0.75, meaning that we don’t have high multicollinearity**



***Our model passed the homoscedasticity test.**