Module 2 final project

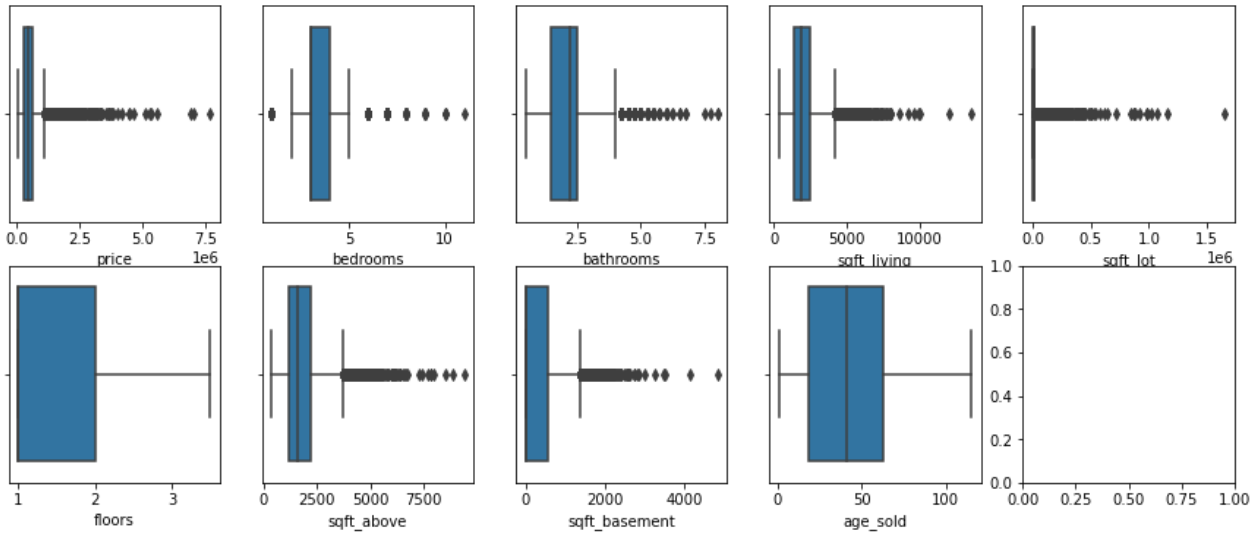
Zhiqiang Sun

- **Business understanding**
- We have had the house selling records for the last few years. With these data, I want to build a model in which I can use the features in the data about the house to predict the price. In this case, we can guide both the seller and buyer to their business. The seller can use the model to predict the selling price of their house and if they need to do any renovation before selling their home. The buyer can have some suggestions about which kind of house they can afford based on their budget. To the details goal：

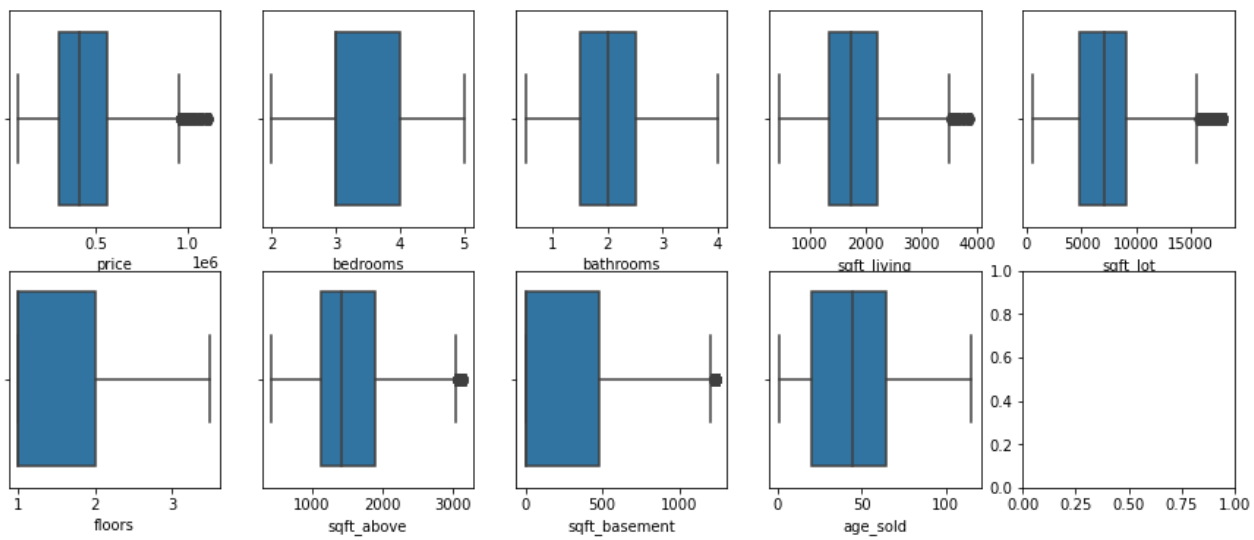
1. polish the data which have no meaning or is null to the price.
2. remove the features which do not contribute to the house price.
3. check if there are some high correlated features in which some of them can be removed.
4. build the linear regression model.
5. check how the features can contribute to the house change.

Remove the outlier data

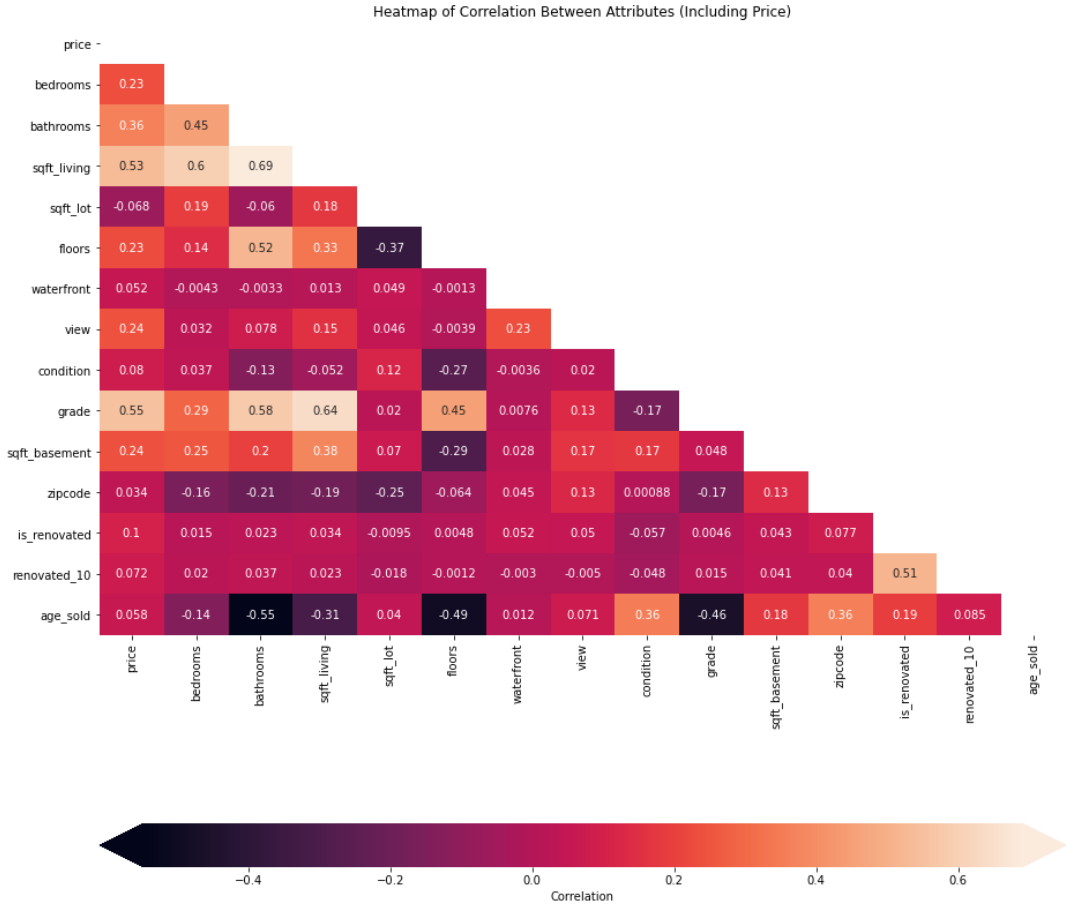
before



after



Find the feature best correlated to the price after remove the high correlated features.



Pre fit the data to with linear regression and remove the high p value features (0.05)

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.533
Model:	OLS	**Adj. R-squared:**	0.533
Method:	Least Squares	**F-statistic:**	1003.
Date:	Sat, 06 Nov 2021	**Prob (F-statistic):**	0.00
Time:	05:04:40	**Log-Likelihood:**	-1.6225e+05
No. Observations:	12297	**AIC:**	3.245e+05
Df Residuals:	12282	**BIC:**	3.246e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.322e+06	2.49e+06	0.933	0.351	-2.55e+06	7.2e+06
bedrooms	-1.786e+04	1960.903	-9.106	0.000	-2.17e+04	-1.4e+04
bathrooms	2.235e+04	3078.018	7.261	0.000	1.63e+04	2.84e+04
sqft_living	103.8368	3.834	27.086	0.000	96.322	111.351
sqft_lot	-7.3467	0.429	-17.118	0.000	-8.188	-6.505
floors	1.846e+04	3442.038	5.362	0.000	1.17e+04	2.52e+04
waterfront	1.43e+05	3.59e+04	3.983	0.000	7.26e+04	2.13e+05
view	3.082e+04	2246.465	13.717	0.000	2.64e+04	3.52e+04
condition	2.043e+04	1982.824	10.304	0.000	1.65e+04	2.43e+04
grade	1.045e+05	1933.535	54.029	0.000	1.01e+05	1.08e+05
sqft_basement	11.0299	4.511	2.445	0.014	2.188	19.872
zipcode	-305.0042	253.600	-1.203	0.229	-802.099	192.091
is_renovated	-6014.9547	8429.046	-0.714	0.475	-2.25e+04	1.05e+04
renovated_10	5.219e+04	1.56e+04	3.351	0.001	2.17e+04	8.27e+04
age_sold	2685.7902	60.005	44.759	0.000	2568.171	2803.410

Omnibus:	619.304	**Durbin-Watson:**	2.024
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	895.400
Skew:	0.462	**Prob(JB):**	3.68e-195
Kurtosis:	3.945	**Cond. No.**	2.64e+07

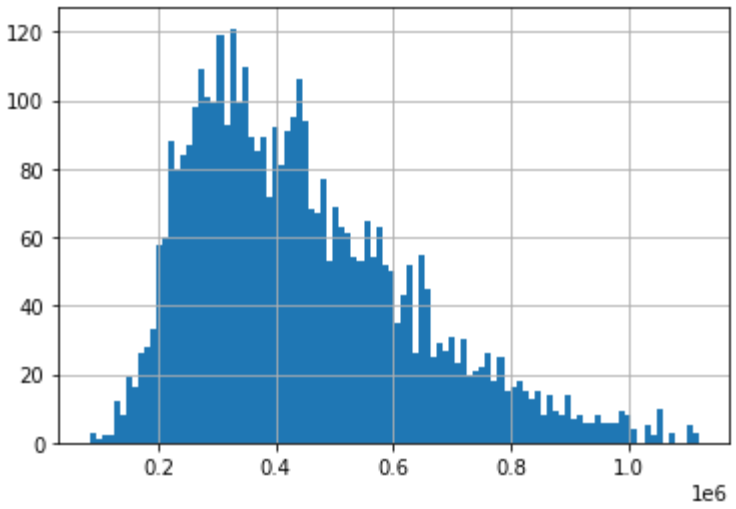
Final fitting

```
bedrooms          -17734.819592
bathrooms          22333.279123
sqft_living          104.197345
sqft_lot              -7.266691
floors             17895.940478
waterfront        140605.432066
view               30502.795377
condition          20867.253746
grade             104396.240619
sqft_basement         10.140670
renovated_10       46690.148414
age_sold            2655.946868
Name: Coefficients, dtype: float64

Intercept: -670009.6123770761
```

MSE value 130331

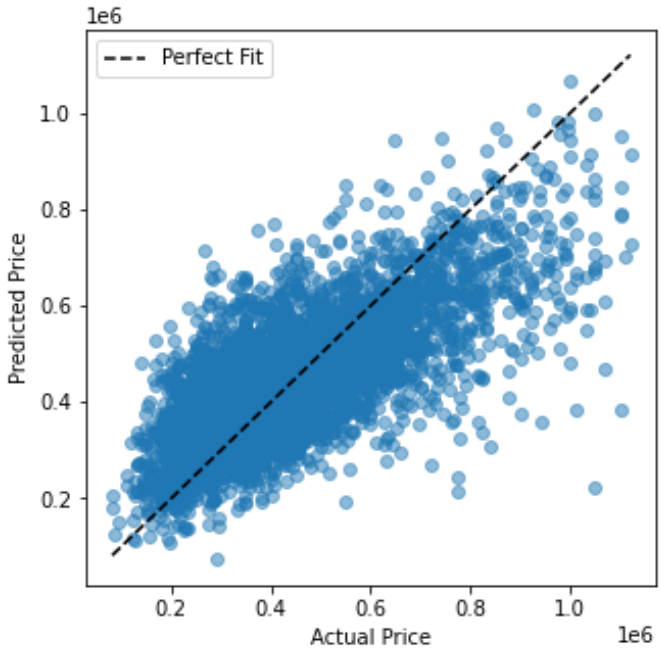
Mean of test price 445683

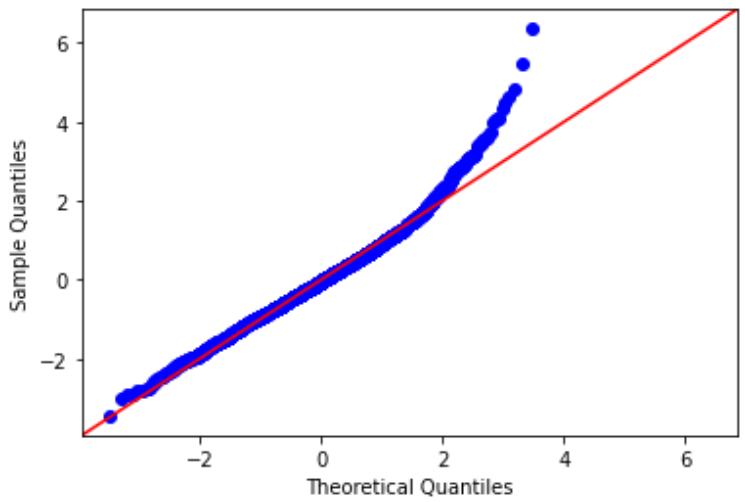


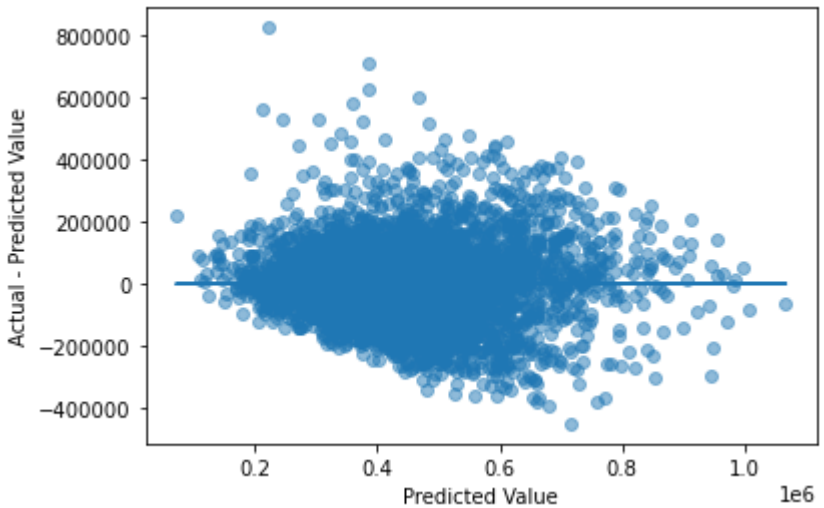
Train score: 0.53218577738026

Validation score: 0.536160764472546

Validation


1e6
Perfect Fit
Predicted Price
Actual Price
1e6


Sample Quantiles
Theoretical Quantiles


Actual - Predicted Value
Predicted Value
1e6

Summary

- Our model predicted well the house price on many of the features. The Coefficients for each features are bedrooms -17734, bathrooms 22333, sqft_living 104, sqft_lot -7, floors 17895, waterfront 140605 , view 30502, condition 20867 , grade 104396, sqft_basement 10 , renovated_10 46690, age_sold 2655,
- To the buyer, they can estimate the price of the house base on the features of the house. To the seller, if they want to sell the house in a better value, they can try to renovate the house and make waterfront if possible. They can also do something to improve the grade level of the house which can also increase the house value dramatically.