Linear Regression Project Report

Jingxian Li Wendeng Hu

1. Description of dataset:

a. Resource:

i. Link: https://www.kaggle.com/shree1992/housedata

ii. Brief description: The dataset we use is a house price dataset from Kaggle. Several variables including number of bedrooms, square foot of living, floors, condition are listed to predict house price. 34% of the records are house locates at Seattle, and we only include these records in our analysis.

b. Dimension

i. 1573 records of Seattle house price

ii. 18 variables including house price

c. Variable description

Name	Description	Category
date	Datetime information of when this property is sold	Datetime
price	Selling price	Ratio
bedrooms	Bedroom number of this property	Ordinal
bathrooms	Bathroom number of this property	Ordinal
sqft_living	Total area of this property measured in square feet	Ratio
sqft_lot	Parking lot area measured in square feet	Ratio
floors	Number of floors of this property	Ordinal
waterfront	Waterfront rating of this property	Ordinal
view	View rating of this property	Ordinal
condition	Condition of the property	Ordinal
sqft_above	Above ground area measured in feet	Ratio
sqft_basement	Basement are measured in square feet	Ratio
yr_built	Year this property is built	Numeric
yr_renovated	Year this property is renovated	Numeric
street	Street of this property	Nominal
city	City this property locates at	Nominal
statezip	State abbreviation of this property and zip number	Nominal
country	Country this property belongs to	Nominal

Table 1 Variable description

2. Data pre-processing

a. Drop variables:
Date was dropped because the datetime of selling has nothing to do with the house price. Only records in Seattle were chosen so city and country were also dropped. Although statezip and street can be treated as categorical data and neighborhood do have significant influence on house price, they have too many unique values, which would cause a lot of problems in the process of dummy and model. As a result, we decided to drop them at first.

b. Variable manipulation
Yr_built and yr_renovated are two variables that are manipulated into other form. When we considered the built time of a property, we actually considered its age, that is the present time minus built age. It is the same with yr_renovated. So, we transform them into another two variables, yr_age and renovated_age. Yr_age means the time between now and the year property built, while yr_renovated means the time between now and the year property renovated. When yr_age equals to yr_renovated, it means that this house has never been renovated and when yr_renovated equals to zero, we can see that this property has been renovated this year.

c. Ordinal data
We didn't do any of the transformation on ordinal data. We will put them directly into the model.

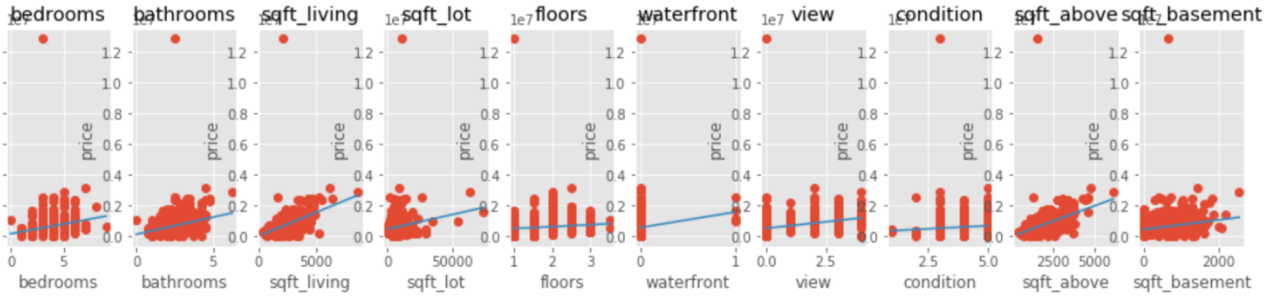
3. Problem Statement and Methods Summary

a. Problem statement
    i. How do house condition related variables influence house price in Seattle?

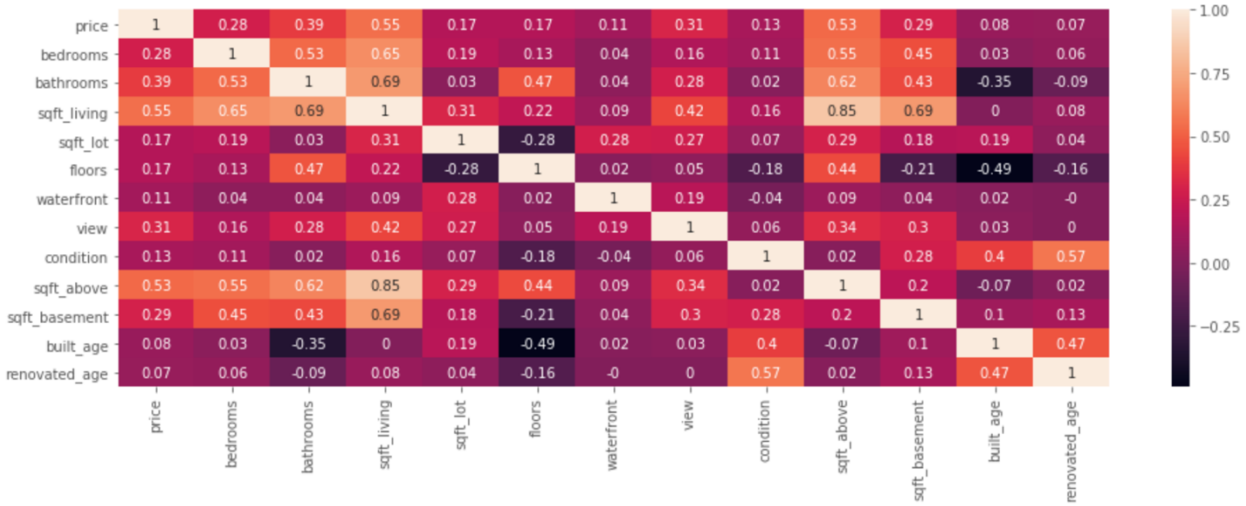
b. Methods Summary
    i. Linear regression model
        1. Exploratory analysis
        2. Model setup and preliminary evaluation (Normality and fitted value vs. residual)
        3. Drop influential points by combining hat matrix, internal and external residual, DFFITS, cook's distance and model reevaluation
        4. Model selection:
            a. Stepwise
            b. Best subset regression
            c. Using Mallow's Cp, AIC, BIC, adjusted R square
        5. Compare best model in stepwise and best subset regression. Test on non-equal variance, multicollinearity and solve related problems
        6. Choose the final model based on the above criteria and do the model evaluation
    ii. Logistic model
        1. Test on the best model of former linear regression and drop/add variables if necessary
        2. Begin with the full model to remove or add variables for logistic

4. Exploratory analysis

a. Correlation



Graph 1 Scatterplot for predictor and response



Graph 2 Correlation matrix in heat map

We can see from the above correlation heatmap and plot of every single predictor with response that number of bedrooms, number of bathrooms, sqft_living, sqft_above has a strong linear pattern with price. Among them, sqft_living ($\rho = 0.55$), sqft_above ($\rho = 0.53$) have moderate correlation with price, bathrooms ($\rho = 0.28$) and bedrooms ($\rho = 0.39$) are also very important. Apart from that, sqft_living and sqft_above shows strong correlation ($\rho = 0.85$), which may lead to multicollinearity in later regression. This makes sense because square feet of living is made up of square feet of above, square feet above and square feet of the parking lot. Sqft_above and price, bedroom number, bathroom number are moderately correlated. Renovated_age and condition are also moderately correlated ($\rho = 0.57$). built_age and renovated_age are weakly correlated ($\rho = 0.47$).

5. Model setup and preliminary evaluation

a. Preliminary model setup

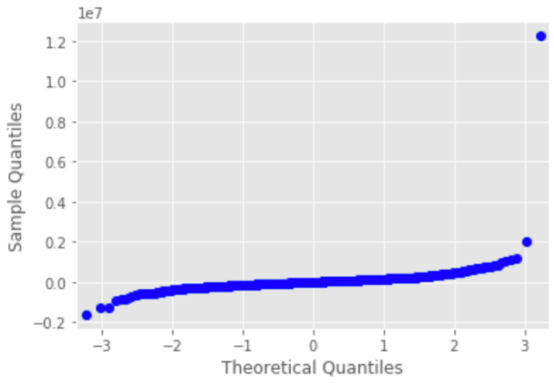
Dep. Variable:	price	**R-squared:**	0.348
Model:	OLS	**Adj. R-squared:**	0.344
Method:	Least Squares	**F-statistic:**	75.88
Date:	Sat, 12 Oct 2019	**Prob (F-statistic):**	1.13e-136
Time:	20:47:04	**Log-Likelihood:**	-22414.
No. Observations:	1573	**AIC:**	4.485e+04
Df Residuals:	1561	**BIC:**	4.492e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.473e+05	6.89e+04	-2.137	0.033	-2.82e+05	-1.21e+04
bedrooms	-6.378e+04	1.28e+04	-4.984	0.000	-8.89e+04	-3.87e+04
bathrooms	5.028e+04	2.01e+04	2.499	0.013	1.08e+04	8.97e+04
sqft_living	186.6921	13.966	13.368	0.000	159.299	214.086
sqft_lot	-5.0517	2.893	-1.746	0.081	-10.726	0.622
floors	7703.4543	2.46e+04	0.313	0.754	-4.05e+04	5.59e+04
waterfront	5.171e+05	1.99e+05	2.599	0.009	1.27e+05	9.07e+05
view	4.186e+04	1.28e+04	3.259	0.001	1.67e+04	6.71e+04
condition	3.669e+04	1.62e+04	2.261	0.024	4863.516	6.85e+04
sqft_above	189.7193	18.739	10.124	0.000	152.962	226.476
sqft_basement	-3.0272	20.836	-0.145	0.885	-43.897	37.843
built_age	1732.1039	380.087	4.557	0.000	986.569	2477.639
renovated_age	-473.8032	356.572	-1.329	0.184	-1173.215	225.608

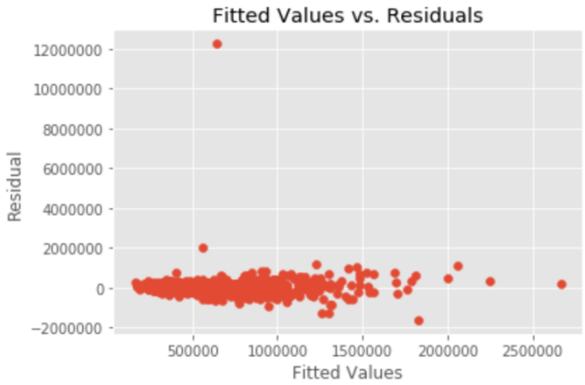
Omnibus:	3718.734	**Durbin-Watson:**	1.979
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	35674914.911
Skew:	22.536	**Prob(JB):**	0.00
Kurtosis:	739.395	**Cond. No.**	1.00e+16

Table 2 preliminary model summary

b. Model evaluation: Normality QQ plot and residual plot



Graph 3 QQ plot for normality check

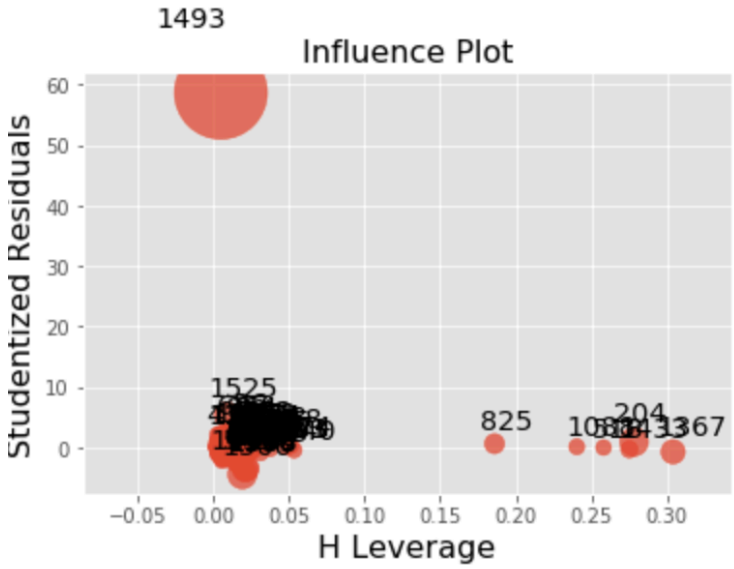


Graph 4 Residual plot

We first included all the predictors in the model. We can see from this graph that the adjusted R square is only 34.4%. Floors, sqft_basement and renovated_age are the three predictors that are not significant.

There's a severe normality problem as in graph 3 the normality plot is not approximate to 45° diagonal line. On the fitted values versus residual plot, points are not randomly distributed between the zero, rather, they appear aggregated near zero. And we can see there are some potential influential points. No matter it's the statistic or the graph, the preliminary model is not a very good one. So, the next step for us is to delete the influential points and see if normality and residual plot will show better results.

c. Drop influential points



Graph5 influential plot

We first draw the influential point graph and then used five methods taught in class to identify influential points: hat matrix, cook's distance, DFFITS, internal and external residuals. We combine the results together and all together dropped 63 records (out of 1753). After that, we refit the model and found R square from around 34.8% improve to around 65.13% and normality also improved.

6. Model selection

We performed model s4.8election through best subset selection and forward selection. From each method, we chose one model as the best and then compared these two models for a better one.

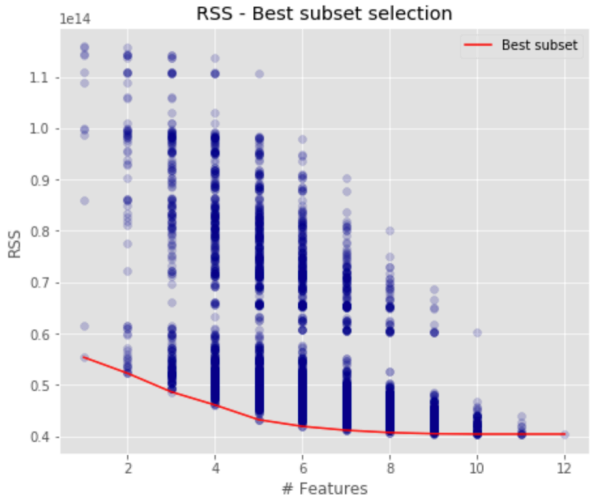
a. Best subset selection

We used best subset selection to generate models containing different combination of predictors. For every number of features, we chose the model with highest adjusted R square. And from these models, we calculated their Mallow's Cp, AIC, BIC and selected the top five models under each of the standard. We found out that

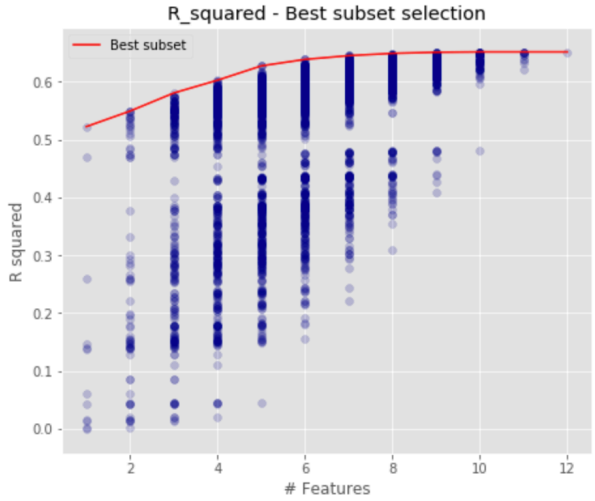
Mallow's Cp, AIC and adjusted R square have similar results for the top five, but we finally chose to use BIC as our standard.

	numb_features	RSS	R_squared	features	min_RSS	max_R_squared
2	1	5.537360e+13	0.522256	(sqft_living,)	5.537360e+13	0.522256
2	1	5.537360e+13	0.522256	(sqft_living,)	5.537360e+13	0.522256
38	2	5.226082e+13	0.549112	(sqft_living, sqft_above)	5.226082e+13	0.549112
39	2	5.226082e+13	0.549112	(sqft_living, sqft_basement)	5.226082e+13	0.549112
72	2	5.226082e+13	0.549112	(sqft_above, sqft_basement)	5.226082e+13	0.549112
191	3	4.863913e+13	0.580358	(sqft_living, floors, built_age)	4.863913e+13	0.580358
620	4	4.608045e+13	0.602434	(sqft_living, floors, view, built_age)	4.608045e+13	0.602434
1374	5	4.321288e+13	0.627174	(sqft_living, sqft_lot, view, sqft_above, buil...	4.321288e+13	0.627174
1376	5	4.321288e+13	0.627174	(sqft_living, sqft_lot, view, sqft_basement, b...	4.321288e+13	0.627174
1520	5	4.321288e+13	0.627174	(sqft_lot, view, sqft_above, sqft_basement, bu...	4.321288e+13	0.627174
1836	6	4.191392e+13	0.638381	(bedrooms, sqft_living, sqft_lot, view, sqft_a...	4.191392e+13	0.638381
1838	6	4.191392e+13	0.638381	(bedrooms, sqft_living, sqft_lot, view, sqft_b...	4.191392e+13	0.638381
1982	6	4.191392e+13	0.638381	(bedrooms, sqft_lot, view, sqft_above, sqft_ba...	4.191392e+13	0.638381
2550	7	4.116081e+13	0.644879	(bedrooms, bathrooms, sqft_living, sqft_lot, v...	4.116081e+13	0.644879
2552	7	4.116081e+13	0.644879	(bedrooms, bathrooms, sqft_living, sqft_lot, v...	4.116081e+13	0.644879
2696	7	4.116081e+13	0.644879	(bedrooms, bathrooms, sqft_lot, view, sqft_abo...	4.116081e+13	0.644879
3357	8	4.070457e+13	0.648815	(bedrooms, bathrooms, sqft_living, sqft_lot, v...	4.070457e+13	0.648815
3359	8	4.070457e+13	0.648815	(bedrooms, bathrooms, sqft_living, sqft_lot, v...	4.070457e+13	0.648815
3477	8	4.070457e+13	0.648815	(bedrooms, bathrooms, sqft_lot, view, conditio...	4.070457e+13	0.648815
3907	9	4.048536e+13	0.650706	(bedrooms, bathrooms, sqft_lot, view, conditio...	4.048536e+13	0.650706
4033	10	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356
4034	10	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356
4058	10	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_lot, floors, view, ...	4.041011e+13	0.651356
4084	11	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356
4085	11	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356
4088	11	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356
4091	11	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_lot, floors, waterf...	4.041011e+13	0.651356
4094	12	4.041011e+13	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356

Table 3 best subset selection initial result



Graph 6 RSS change in best subset selection



Graph 7 R square change in best subset selection

	numb_features	RSS	R_squared	features	min_RSS	max_R_squared	C_p	AIC	BIC	R_squared_adj	index
4058	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_lot, floors, view, ...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	22
4034	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	21
4033	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	20
3907	9	40485359090426	0.650706	(bedrooms, bathrooms, sqft_lot, view, conditio...	4.048536e+13	0.650706	26845421917	1.005123	1.036560	0.648633	19
4088	11	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26866118741	1.005898	1.044321	0.648823	25

Table 4 Top five for Mallow's Cp

	numb_features	RSS	R_squared	features	min_RSS	max_R_squared	C_p	AIC	BIC	R_squared_adj	index
4034	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	21
4033	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	20
4058	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_lot, floors, view, ...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	22
3907	9	40485359090426	0.650706	(bedrooms, bathrooms, sqft_lot, view, conditio...	4.048536e+13	0.650706	26845421917	1.005123	1.036560	0.648633	19
4088	11	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26866118741	1.005898	1.044321	0.648823	25

Table 5 Top five for AIC

	numb_features	RSS	R_squared	features	min_RSS	max_R_squared	C_p	AIC	BIC	R_squared_adj	index
3907	9	40485359090426	0.650706	(bedrooms, bathrooms, sqft_lot, view, conditio...	4.048536e+13	0.650706	26845421917	1.005123	1.036560	0.648633	19
3477	8	40704567934301	0.648815	(bedrooms, bathrooms, sqft_lot, view, conditio...	4.070457e+13	0.648815	26954066510	1.009191	1.037135	0.646963	18
3359	8	40704567934301	0.648815	(bedrooms, bathrooms, sqft_living, sqft_lot, v...	4.070457e+13	0.648815	26954066510	1.009191	1.037135	0.646963	17
3357	8	40704567934301	0.648815	(bedrooms, bathrooms, sqft_living, sqft_lot, v...	4.070457e+13	0.648815	26954066510	1.009191	1.037135	0.646963	16
4034	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	21

Table 6 Top five for BIC

	numb_features	RSS	R_squared	features	min_RSS	max_R_squared	C_p	AIC	BIC	R_squared_adj	index
4058	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_lot, floors, view, ...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	22
4034	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	21
4033	10	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26831114026	1.004587	1.039518	0.649054	20
4091	11	40410108053829	0.651356	(bedrooms, bathrooms, sqft_lot, floors, waterf...	4.041011e+13	0.651356	26866118741	1.005898	1.044321	0.648823	26
4088	11	40410108053829	0.651356	(bedrooms, bathrooms, sqft_living, sqft_lot, f...	4.041011e+13	0.651356	26866118741	1.005898	1.044321	0.648823	25

Table 7 Top five for adjusted R square

Based on BIC, model price ~ sqft_living + sqft_above + built_age + sqft_lot + view + bedrooms + bathrooms + condition + renovated_age was chosen. There were nine predictors in this model with Mallow's Cp = 26845421917, AIC = 1.005123, BIC = 1.03656 and adjusted R square = 0.648633.

b. Forward selection

	features	RSS	R_squared	numb_features	C_p	AIC	BIC	R_squared_adj
1	[sqft_living]	55373604054273	0.522256	1	36321770150	1.359928	1.363421	0.521942
2	[sqft_living, sqft_above]	52260820828020	0.549112	2	34316943131	1.284865	1.291851	0.548520
3	[sqft_living, sqft_above, built_age]	49407837919706	0.573726	3	32482365337	1.216176	1.226656	0.572886
4	[sqft_living, sqft_above, built_age, sqft_lot]	46225100861872	0.601186	4	30431697013	1.139397	1.153369	0.600137
5	[sqft_living, sqft_above, built_age, sqft_lot,...	43212883382895	0.627174	5	28492771532	1.066801	1.084267	0.625948
6	[sqft_living, sqft_above, built_age, sqft_lot,...	41913916567245	0.638381	6	27676552907	1.036241	1.057200	0.636953
7	[sqft_living, sqft_above, built_age, sqft_lot,...	41160810764306	0.644879	7	27218041369	1.019074	1.043526	0.643241
8	[sqft_living, sqft_above, built_age, sqft_lot,...	40704567934301	0.648815	8	26954066510	1.009191	1.037135	0.646963
9	[sqft_living, sqft_above, built_age, sqft_lot,...	40485359090426	0.650706	9	26845421917	1.005123	1.036560	0.648633
10	[sqft_living, sqft_above, built_age, sqft_lot,...	40410108053829	0.651356	10	26831114026	1.004587	1.039518	0.649054
11	[sqft_living, sqft_above, built_age, sqft_lot,...	40410108053829	0.651356	11	26866118741	1.005898	1.044321	0.648823
12	[sqft_living, sqft_above, built_age, sqft_lot,...	40410108053829	0.651356	12	26901123456	1.007208	1.049125	0.648590

Table 8 Forward selection result

Forward selection gave us view on how each predictor's entering will influence the model. We used BIC to choose model and our final choice was: price ~ bedrooms + bathrooms + sqft_lot + view + condition + sqft_above + sqft_basement + built_age + renovated_age. This model contained basically the same predictors as the one that was chosen through best subset selection. The only difference was that in the model chosen through best subset selection, sqft_living and sqft_basement were included whereas in the model chosen through forward selection, sqft_above and sqft_basement were included. Actually, these two predictors contained the same information on the area of house, however, the one that in best subset selection might cause multicollinearity because sqft_living = sqft_above + sqft_basement. Our assumption was proved in later test of multicollinearity.

7. Model diagnosis and comparison: for the two models chosen from best subset selection and forward selection.

a. Model chosen through best subset regression

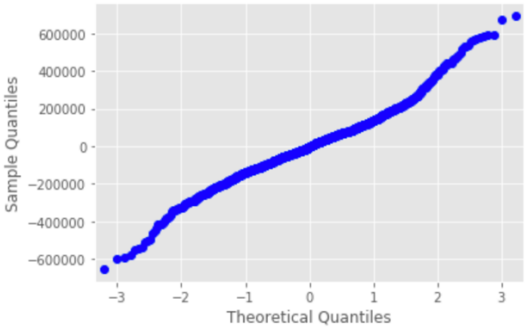
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.651
Model:	OLS	**Adj. R-squared:**	0.649
Method:	Least Squares	**F-statistic:**	313.8
Date:	Sat, 12 Oct 2019	**Prob (F-statistic):**	0.00
Time:	20:47:26	**Log-Likelihood:**	-20478.
No. Observations:	1526	**AIC:**	4.098e+04
Df Residuals:	1516	**BIC:**	4.103e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.025e+05	2.81e+04	-3.641	0.000	-1.58e+05	-4.73e+04
sqft_living	159.3758	12.935	12.322	0.000	134.004	184.748
sqft_above	187.1158	14.001	13.365	0.000	159.653	214.578
built_age	1893.6752	161.877	11.698	0.000	1576.149	2211.201
sqft_lot	-15.1546	1.622	-9.341	0.000	-18.337	-11.972
view	5.662e+04	6043.432	9.369	0.000	4.48e+04	6.85e+04
bedrooms	-4.605e+04	5961.664	-7.724	0.000	-5.77e+04	-3.44e+04
bathrooms	4.388e+04	8915.182	4.922	0.000	2.64e+04	6.14e+04
condition	3.663e+04	7326.051	4.999	0.000	2.23e+04	5.1e+04
renovated_age	-456.4847	159.330	-2.865	0.004	-769.015	-143.955

Omnibus:	77.780	**Durbin-Watson:**	1.914
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	210.959
Skew:	0.235	**Prob(JB):**	1.55e-46
Kurtosis:	4.760	**Cond. No.**	4.26e+04

Table 9 Model summary of model chosen from best subset regression



Graph 8 Model normality

Results for heteroskedasticity:
{'LM Statistic': 245.69779101234948, 'LM-Test p-value': 8.072455576263355e-48, 'F-Statistic': 32.325514724392406, 'F-Test p-value': 2.888609629789586e-52}

The null hypothesis is $\gamma 1 = \gamma 2 = \gamma 3 = \ldots\ldots = \gamma n$. The null hypothesis was rejected with p value less than 0.05. We choose to believe the alternative hypothesis, there was heteroskedasticity problem. We use weighted least square method to fit the model again.

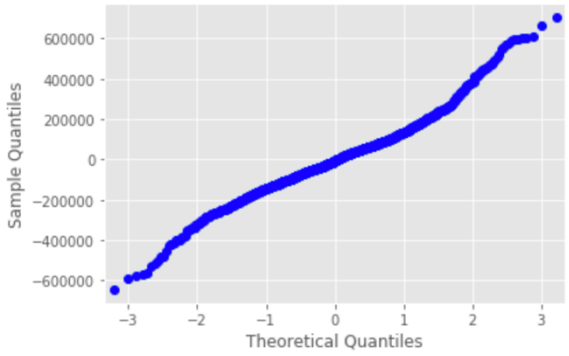
WLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.929
Model:	WLS	**Adj. R-squared (uncentered):**	0.928
Method:	Least Squares	**F-statistic:**	2195.
Date:	Sat, 12 Oct 2019	**Prob (F-statistic):**	0.00
Time:	20:47:26	**Log-Likelihood:**	-20485.
No. Observations:	1526	**AIC:**	4.099e+04
Df Residuals:	1517	**BIC:**	4.104e+04
Df Model:	9		
Covariance Type:	nonrobust		

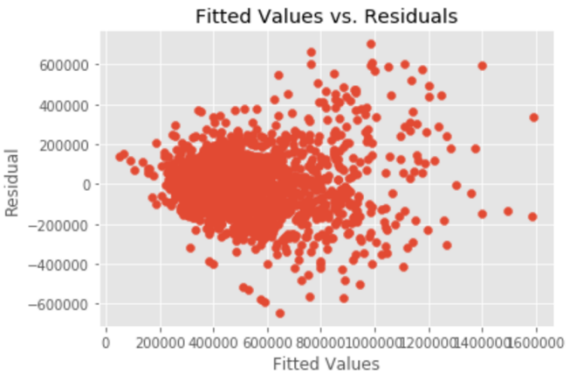
	coef	std err	t	P>|t|	[0.025	0.975]
sqft_living	175.6848	12.183	14.420	0.000	151.787	199.583
sqft_above	171.4887	13.380	12.816	0.000	145.243	197.735
built_age	1781.4132	159.554	11.165	0.000	1468.443	2094.383
sqft_lot	-16.2972	1.598	-10.198	0.000	-19.432	-13.162
view	5.763e+04	6061.395	9.508	0.000	4.57e+04	6.95e+04
bedrooms	-5.144e+04	5798.302	-8.871	0.000	-6.28e+04	-4.01e+04
bathrooms	3.445e+04	8565.057	4.022	0.000	1.76e+04	5.13e+04
condition	1.77e+04	5184.699	3.415	0.001	7533.665	2.79e+04
renovated_age	-265.2984	151.035	-1.757	0.079	-561.559	30.962

Omnibus:	94.838	**Durbin-Watson:**	1.915
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	241.355
Skew:	0.339	**Prob(JB):**	3.89e-53
Kurtosis:	4.826	**Cond. No.**	1.34e+04

Table 10 Using WLS to fit model chosen from best subset regression



Graph 9 Model normality



Graph 10 Residual plot

Result for multicollinearity:

	VIF Factor	features
0	5.050210	sqft_living
1	3.232208	sqft_above
2	1.811425	built_age
3	1.233463	sqft_lot
4	1.150662	view
5	1.857179	bedrooms
6	2.702225	bathrooms
7	1.681896	condition
8	1.707410	renovated_age

There existed multicollinearity problem for sqft_living as we expected. Let's look at the model chosen through forward selection. As in this model, sqft_living was replaced by sqft_above, there shouldn't be multicollinearity problem.

b. Model chosen through forward selection

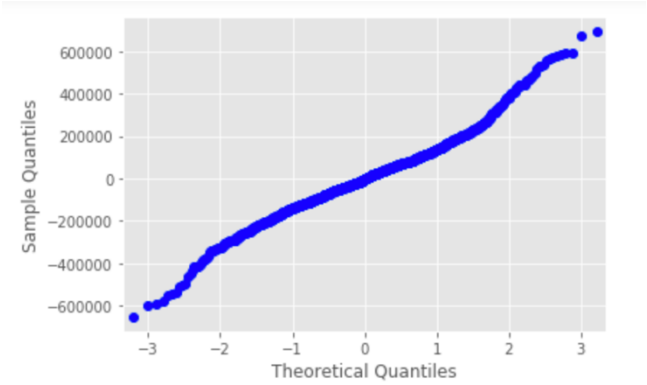
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.651
Model:	OLS	**Adj. R-squared:**	0.649
Method:	Least Squares	**F-statistic:**	313.8
Date:	Sat, 12 Oct 2019	**Prob (F-statistic):**	0.00
Time:	20:47:26	**Log-Likelihood:**	-20478.
No. Observations:	1526	**AIC:**	4.098e+04
Df Residuals:	1516	**BIC:**	4.103e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.025e+05	2.81e+04	-3.641	0.000	-1.58e+05	-4.73e+04
bedrooms	-4.605e+04	5961.664	-7.724	0.000	-5.77e+04	-3.44e+04
bathrooms	4.388e+04	8915.182	4.922	0.000	2.64e+04	6.14e+04
sqft_lot	-15.1546	1.622	-9.341	0.000	-18.337	-11.972
view	5.662e+04	6043.432	9.369	0.000	4.48e+04	6.85e+04
condition	3.663e+04	7326.051	4.999	0.000	2.23e+04	5.1e+04
sqft_above	346.4916	11.448	30.268	0.000	324.037	368.946
sqft_basement	159.3758	12.935	12.322	0.000	134.004	184.748
built_age	1893.6752	161.877	11.698	0.000	1576.149	2211.201
renovated_age	-456.4847	159.330	-2.865	0.004	-769.015	-143.955

Omnibus:	77.780	**Durbin-Watson:**	1.914
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	210.959
Skew:	0.235	**Prob(JB):**	1.55e-46
Kurtosis:	4.760	**Cond. No.**	4.11e+04

Table 10 Model summary of model chosen from forward selection



Graph 11 Model normality

Results for Heteroskedasticity:
{'LM Statistic': 245.69779101233138, 'LM-Test p-value': 8.072455576334401e-48, 'F-Statistic': 32.32551472438958, 'F-Test p-value': 2.8886096298197573e-52}

The null hypothesis is $\gamma 1 = \gamma 2 = \gamma 3 = \ldots\ldots = \gamma n$. The null hypothesis was rejected with p value less than 0.05. We chose to believe the alternative hypothesis, there was heteroskedasticity problem. Weighted least square method was chosen.

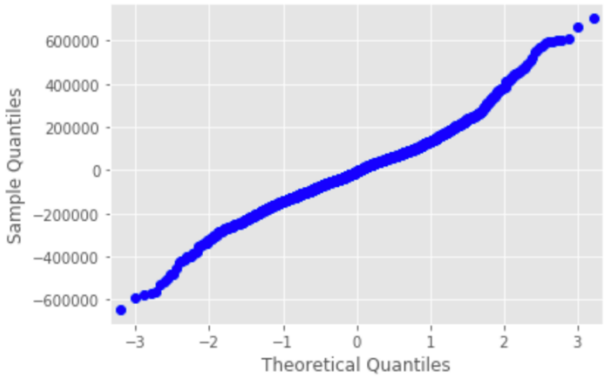
WLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.929
Model:	WLS	**Adj. R-squared (uncentered):**	0.928
Method:	Least Squares	**F-statistic:**	2195.
Date:	Sat, 12 Oct 2019	**Prob (F-statistic):**	0.00
Time:	20:47:26	**Log-Likelihood:**	-20485.
No. Observations:	1526	**AIC:**	4.099e+04
Df Residuals:	1517	**BIC:**	4.104e+04
Df Model:	9		
Covariance Type:	nonrobust		

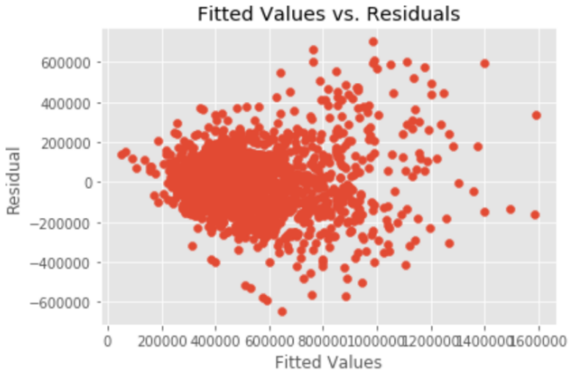
	coef	std err	t	P>|t|	[0.025	0.975]
bedrooms	-5.144e+04	5798.302	-8.871	0.000	-6.28e+04	-4.01e+04
bathrooms	3.445e+04	8565.057	4.022	0.000	1.76e+04	5.13e+04
sqft_lot	-16.2972	1.598	-10.198	0.000	-19.432	-13.162
view	5.763e+04	6061.395	9.508	0.000	4.57e+04	6.95e+04
condition	1.77e+04	5184.699	3.415	0.001	7533.665	2.79e+04
sqft_above	347.1736	11.492	30.209	0.000	324.631	369.716
sqft_basement	175.6848	12.183	14.420	0.000	151.787	199.583
built_age	1781.4132	159.554	11.165	0.000	1468.443	2094.383
renovated_age	-265.2984	151.035	-1.757	0.079	-561.559	30.962

Omnibus:	94.838	**Durbin-Watson:**	1.915
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	241.355
Skew:	0.339	**Prob(JB):**	3.89e-53
Kurtosis:	4.826	**Cond. No.**	1.29e+04

Table 11 WLS to fit model chosen from forward selection



Graph 12 Model normality



Graph 13 Residual plot

	VIF Factor	features
0	1.857179	bedrooms
1	2.702225	bathrooms
2	1.233463	sqft_lot
3	1.150662	view
4	1.681896	condition
5	2.160901	sqft_above
6	1.705948	sqft_basement
7	1.811425	built_age
8	1.707410	renovated_age

There's a significant improve in R square after we use weighted least square. Heteroskedasticity problem still exists but multicollinearity problem doesn't exist. However, we found that renovated_age's p value is bigger than 0.05, which means we do not have enough evidence to reject the null hypothsis. So we fit another wls model removing the renovated_age variable.

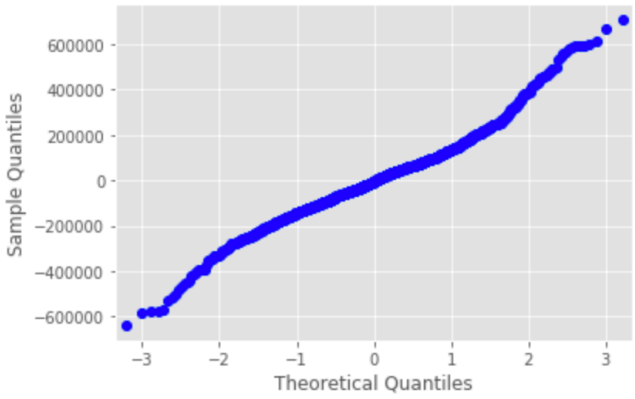
WLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.929
Model:	WLS	**Adj. R-squared (uncentered):**	0.928
Method:	Least Squares	**F-statistic:**	2466.
Date:	Sun, 13 Oct 2019	**Prob (F-statistic):**	0.00
Time:	20:58:56	**Log-Likelihood:**	-20487.
No. Observations:	1526	**AIC:**	4.099e+04
Df Residuals:	1518	**BIC:**	4.103e+04
Df Model:	8		
Covariance Type:	nonrobust		

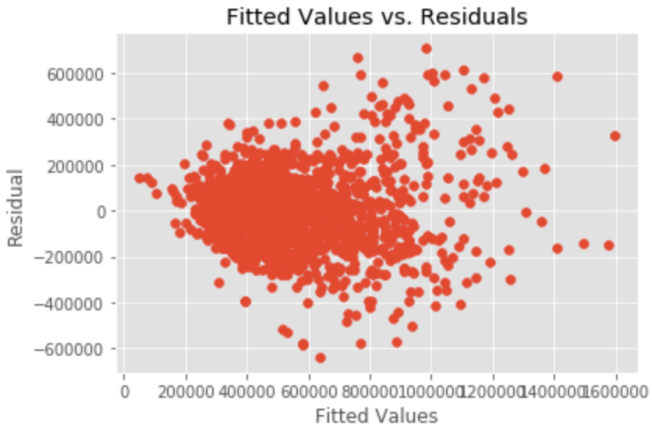
	coef	std err	t	P>|t|	[0.025	0.975]
bedrooms	-5.029e+04	5765.101	-8.723	0.000	-6.16e+04	-3.9e+04
bathrooms	3.607e+04	8520.873	4.234	0.000	1.94e+04	5.28e+04
sqft_lot	-15.9204	1.585	-10.046	0.000	-19.029	-12.812
view	5.782e+04	6064.602	9.534	0.000	4.59e+04	6.97e+04
condition	1.438e+04	4829.810	2.977	0.003	4903.480	2.39e+04
sqft_above	345.9855	11.480	30.138	0.000	323.467	368.504
sqft_basement	173.3794	12.121	14.304	0.000	149.604	197.155
built_age	1710.5551	154.476	11.073	0.000	1407.545	2013.565

Omnibus:	91.210	**Durbin-Watson:**	1.918
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	232.682
Skew:	0.323	**Prob(JB):**	2.98e-51
Kurtosis:	4.801	**Cond. No.**	1.28e+04

Table 12 WLS to fit model chosen from forward selection (remove waterfront)



Graph 14 Model normality



Graph 15 Residual plot

	VIF Factor	features
0	1.856316	bedrooms
1	2.702217	bathrooms
2	1.228938	sqft_lot
3	1.149493	view
4	1.299483	condition
5	2.155463	sqft_above
6	1.705351	sqft_basement
7	1.651529	built_age

For this time, all of the predictors are statistically significant.

8. Final choice of model and interpretation

	sum_sq	df	F	PR(>F)
bedrooms	1.568697e+12	1.0	58.463042	3.655453e-14
bathrooms	6.482917e+11	1.0	24.160889	9.822449e-07
sqft_lot	2.252834e+12	1.0	83.959834	1.590419e-19
view	2.392358e+12	1.0	89.159710	1.322097e-20
condition	4.562428e+11	1.0	17.003506	3.933206e-05
sqft_above	2.429479e+13	1.0	905.431348	2.225425e-156
sqft_basement	4.091188e+12	1.0	152.472628	1.902968e-33
built_age	3.446323e+12	1.0	128.439431	1.229123e-28
Residual	4.070457e+13	1517.0	NaN	NaN

Table 13 ANOVA table for final model

So, our final choice of model is the model chosen from forward selection. price ~ bedrooms + bathrooms + sqft_lot + view + condition + sqft_above + sqft_basement + built_age + renovated_age. Using weighted least square method, adjusted R square is 92.8%. All the variables in the model are significant in t test. The null hypothesis for t test is that $\beta = 0$, and small p-value indicates that we reject null hypothesis, which means $\beta \neq 0$. So, all the predictors have significant influence on the model. And in summary, we can see p-value for F test = 0.00 < 0.05, which indicates at least one $\beta i \neq 0$.

ANOVA test was further performed. The null hypothesis for the anova test was $\beta = 0$ $, and\ the\ alternative\ hypothesis\ for$ anova test was $\beta \neq 0$. All of the F statistics and its associated p value show significant.

9. Logistic regression

House price was categorized into two categories: above median and below median, respectively as high price house and low price house. Two methods for selecting best model of logistic regression were performed.

Method I

The predictors in our best linear model were included at first.

Model 1: price_class ~ bedrooms + bathrooms + sqft_lot + view + condition + sqft_above + sqft_basement + built_age + renovated_age

Generalized Linear Model Regression Results

Dep. Variable:	price_class	No. Observations:	1526
Model:	GLM	Df Residuals:	1516
Model Family:	Binomial	Df Model:	9
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-681.27
Date:	Sun, 13 Oct 2019	Deviance:	1362.5
Time:	17:11:03	Pearson chi2:	1.59e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.7001	0.528	-12.700	0.000	-7.734	-5.666
bedrooms	-0.3381	0.104	-3.258	0.001	-0.541	-0.135
bathrooms	0.4923	0.151	3.262	0.001	0.196	0.788
sqft_lot	-0.0002	3.17e-05	-7.426	0.000	-0.000	-0.000
view	0.4049	0.124	3.258	0.001	0.161	0.648
condition	0.3915	0.120	3.251	0.001	0.155	0.628
sqft_above	0.0035	0.000	13.252	0.000	0.003	0.004
sqft_basement	0.0018	0.000	7.556	0.000	0.001	0.002
built_age	0.0237	0.003	8.217	0.000	0.018	0.029
renovated_age	-0.0040	0.003	-1.522	0.128	-0.009	0.001

Table 14 logistic regression summary of model 1 in Method 1

The null hypothesis for z test is $\beta = 0$ and the alternative hypothesis for z test is $\beta \neq 0$. P value for renovated_age was large than 0.05, which means renovated_age was statistically insignificant. Renovated_age was removed from our model and we get new model:

Model2: price_class ~ bedrooms + bathrooms + sqft_lot + view + condition + sqft_above + sqft_basement + built_age

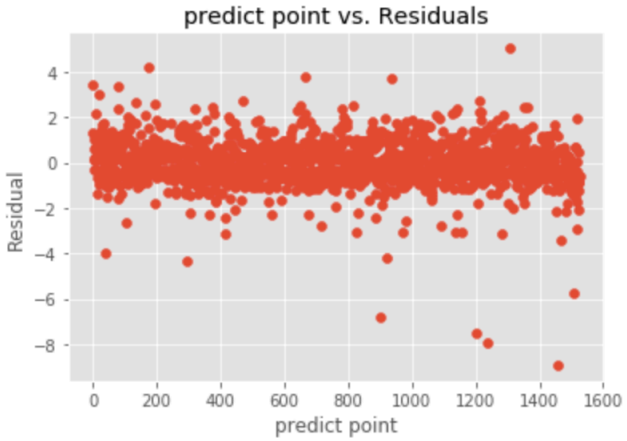
Generalized Linear Model Regression Results

Dep. Variable:	price_class	No. Observations:	1526
Model:	GLM	Df Residuals:	1517
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-682.44
Date:	Sun, 13 Oct 2019	Deviance:	1364.9
Time:	17:11:03	Pearson chi2:	1.57e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.4405	0.495	-13.009	0.000	-7.411	-5.470
bedrooms	-0.3303	0.103	-3.192	0.001	-0.533	-0.127
bathrooms	0.4865	0.151	3.232	0.001	0.191	0.782
sqft_lot	-0.0002	3.14e-05	-7.327	0.000	-0.000	-0.000
view	0.4104	0.124	3.299	0.001	0.167	0.654
condition	0.3041	0.105	2.891	0.004	0.098	0.510
sqft_above	0.0034	0.000	13.209	0.000	0.003	0.004
sqft_basement	0.0018	0.000	7.551	0.000	0.001	0.002
built_age	0.0224	0.003	8.195	0.000	0.017	0.028

Table 15 logistic regression summary of model 2 in Method 1

Based on the summary table above, all of the predictors are statistically significant.



Graph 16 Residual plot

Residuals plot for Model2 was plotted. Most of the points fall into the range of -3 to 3, which was one of the indicators that showed the model that we build was good.

Deviance test was further performed. The null hypothesis was Model 2 should be selected. The alternative hypothesis was Model 1 should be selected. The difference between this two model's deviance was 2.4. The critical value (chi-square with 1 degree of freedom and 0.05 is 3.84). 2.4 is less than 3.84. We failed to reject the null hypothesis.

The prediction accuracy was calculated based on the whole data set. We got accuracy at 78.96%.

Method II

Full model was built at first.

Model 1: price_class ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfront + view + condition + sqft_above + sqft_basement + built_age + renovated_age

Generalized Linear Model Regression Results

Dep. Variable:	price_class	No. Observations:	1526
Model:	GLM	Df Residuals:	1515
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-679.74
Date:	Sat, 12 Oct 2019	Deviance:	1359.5
Time:	21:06:32	Pearson chi2:	1.58e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.1958	0.604	-11.909	0.000	-8.380	-6.012
bedrooms	-0.3376	0.104	-3.250	0.001	-0.541	-0.134
bathrooms	0.4403	0.154	2.861	0.004	0.139	0.742
sqft_living	0.0017	0.000	12.223	0.000	0.001	0.002
sqft_lot	-0.0002	3.51e-05	-5.896	0.000	-0.000	-0.000
floors	0.3191	0.182	1.757	0.079	-0.037	0.675
waterfront	1.482e-15	1.22e-16	12.130	0.000	1.24e-15	1.72e-15
view	0.4083	0.125	3.277	0.001	0.164	0.652
condition	0.3918	0.121	3.244	0.001	0.155	0.628
sqft_above	0.0016	0.000	9.395	0.000	0.001	0.002
sqft_basement	0.0002	0.000	0.989	0.323	-0.000	0.000
built_age	0.0255	0.003	8.292	0.000	0.019	0.032
renovated_age	-0.0041	0.003	-1.576	0.115	-0.009	0.001

Table 16 logistic regression summary of model 1 in Method 2

The null hypothesis for Z test is $\beta = 0$, and the alternative hypothesis for Z test is $\beta \neq 0$. Based on the summary table, sqft_basement and renovated_age were not significant in Z test. There's no strong enough evidence showing that $\beta \neq 0$. sqft_basement and renovated_age were removed and a new model was fitted.

Model2: price_class ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfront + view + condition + sqft_above + built_age

Generalized Linear Model Regression Results

Dep. Variable:	price_class	No. Observations:	1526
Model:	GLM	Df Residuals:	1517
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-682.44
Date:	Sat, 12 Oct 2019	Deviance:	1364.9
Time:	21:23:48	Pearson chi2:	1.57e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.4405	0.495	-13.009	0.000	-7.411	-5.470
bedrooms	-0.3303	0.103	-3.192	0.001	-0.533	-0.127
bathrooms	0.4865	0.151	3.232	0.001	0.191	0.782
sqft_living	0.0018	0.000	7.551	0.000	0.001	0.002
sqft_lot	-0.0002	3.14e-05	-7.327	0.000	-0.000	-0.000
waterfront	4.861e-16	7.3e-16	0.666	0.506	-9.45e-16	1.92e-15
view	0.4104	0.124	3.299	0.001	0.167	0.654
condition	0.3041	0.105	2.891	0.004	0.098	0.510
sqft_above	0.0017	0.000	6.721	0.000	0.001	0.002
built_age	0.0224	0.003	8.195	0.000	0.017	0.028

Table 17 logistic regression summary of model 2 in Method 2

Based on the summary table above, waterfront should be removed since it was not statistically significant with a p value at 0.506. A new model was fitted further.

Model3:
price_class~bedrooms+bathrooms+sqft_living+sqft_lot+floors+view+condition+sqft_above
+built_age

Generalized Linear Model Regression Results

Dep. Variable:	price_class	No. Observations:	1526
Model:	GLM	Df Residuals:	1517
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-682.44
Date:	Sun, 13 Oct 2019	Deviance:	1364.9
Time:	17:11:03	Pearson chi2:	1.57e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.4405	0.495	-13.009	0.000	-7.411	-5.470
bedrooms	-0.3303	0.103	-3.192	0.001	-0.533	-0.127
bathrooms	0.4865	0.151	3.232	0.001	0.191	0.782
sqft_living	0.0018	0.000	7.551	0.000	0.001	0.002
sqft_lot	-0.0002	3.14e-05	-7.327	0.000	-0.000	-0.000
view	0.4104	0.124	3.299	0.001	0.167	0.654
condition	0.3041	0.105	2.891	0.004	0.098	0.510
sqft_above	0.0017	0.000	6.721	0.000	0.001	0.002
built_age	0.0224	0.003	8.195	0.000	0.017	0.028

Table 18 logistic regression summary of model 3 in Method 2

Based on the summary table above, all of the predictors were statistically significant.

Deviance test for full model and model 3 was performed. The null hypothesis was model 3 should be selected. The alternative hypothesis was full model should be selected. The difference in deviance between full model and reduced model was 5.4. The critical value(chi-square with 4 degree of freedom and a =0.05) was 9.49. 5.4 is less than 9.49. We failed to reject the null hypothesis. Model 3 was selected.

And finally, we use this mode to do prediction, the accuracy is 78.96%.

10. Model summary and interpretation
    a. Linear model

The best linear model that we select is:
price ~ bedrooms + bathrooms + sqft_lot + view + condition + sqft_above + sqft_basement + built_age + renovated_age.

Interpretation:
Based on the summary table, an interested thing was noticed, when the number of bedrooms increase by 1, the house price will decrease 5.144e+04.The reason for that may be, when people want to buy a house, indexes like view, condition, sqft_basement maybe more important than number of bedrooms.When the number of bathrooms increase by 1 unit, the house price will increase by 3.445e+04.When sqft_loft increase by 1 unit, house price will decrease by -15.2972.When the level of view increase by 1 level, the house price will increase 5.763e+04. When the condition increase by 1 level, the house price will increase 1.77e+04. When the sqft_above increase by 1 unit, house price will increase 347.1736. When sqft_basement increase by 1 unit, house price will increase 175.6848. When the built_age increase by 1 year, house price will increase 1781.4132. When the renovated_age increase by 1 year, house price will decrease 265.2984.
View, number of bedrooms and number of bathrooms have the most influence on house price. Sqft_lot has the lowest influence on house price.

    b. Logistic model

Logit(π) = bedrooms + bathrooms + sqft_lot + view + condition + sqft_above + sqft_basement + built_age.

For logistic regression, we used the similar model as the linear regression model. When predictors have a significant influence, they also have a significant influence in logistic regression model. The view, bedrooms and bathrooms have significant influence on the odds.