Assignment03

Assignment

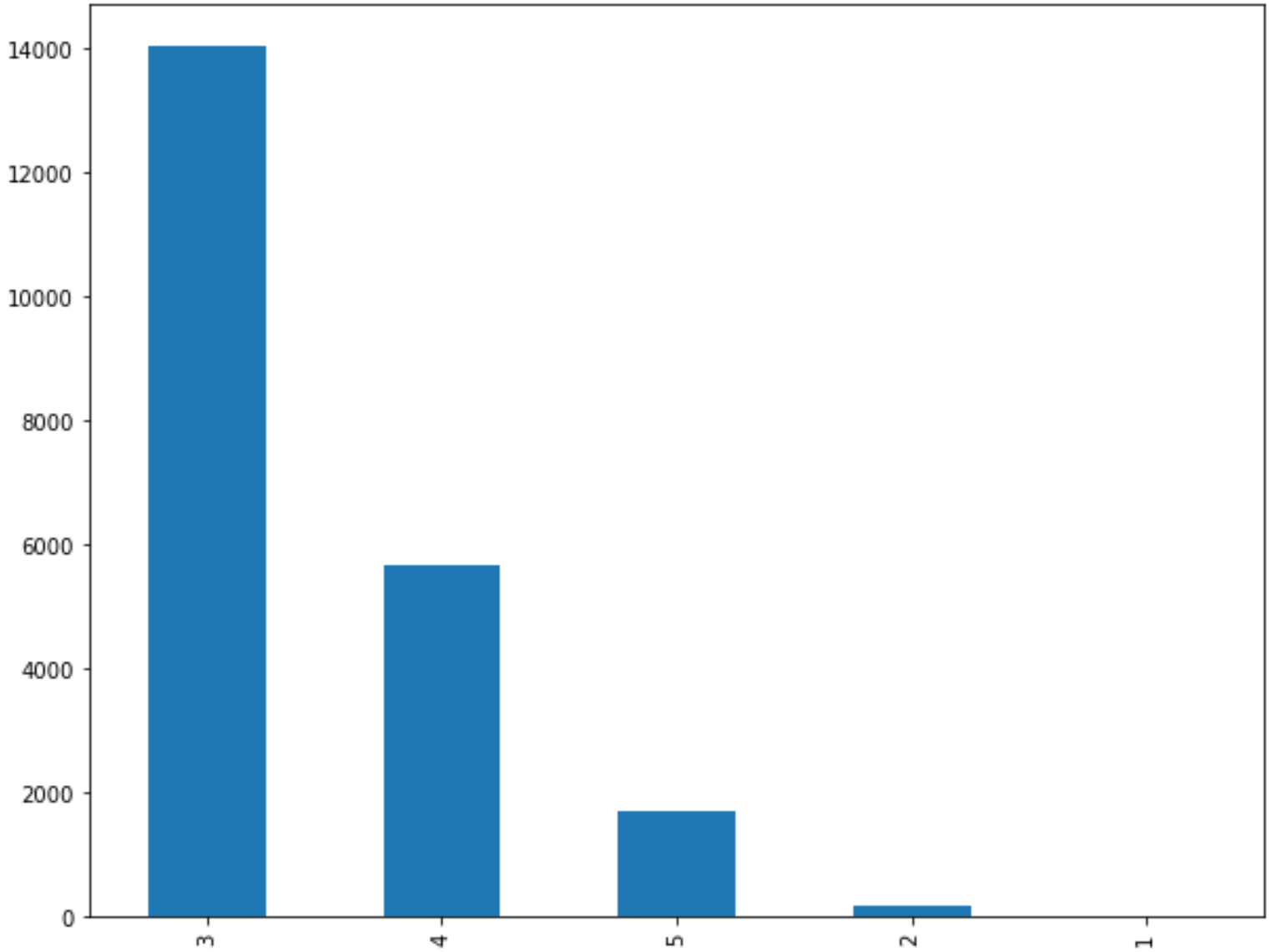
Assignment 03

- Add categorical variables to variable set
  - ‘view’, ‘condition’, ‘grade’
  - Interpret the results

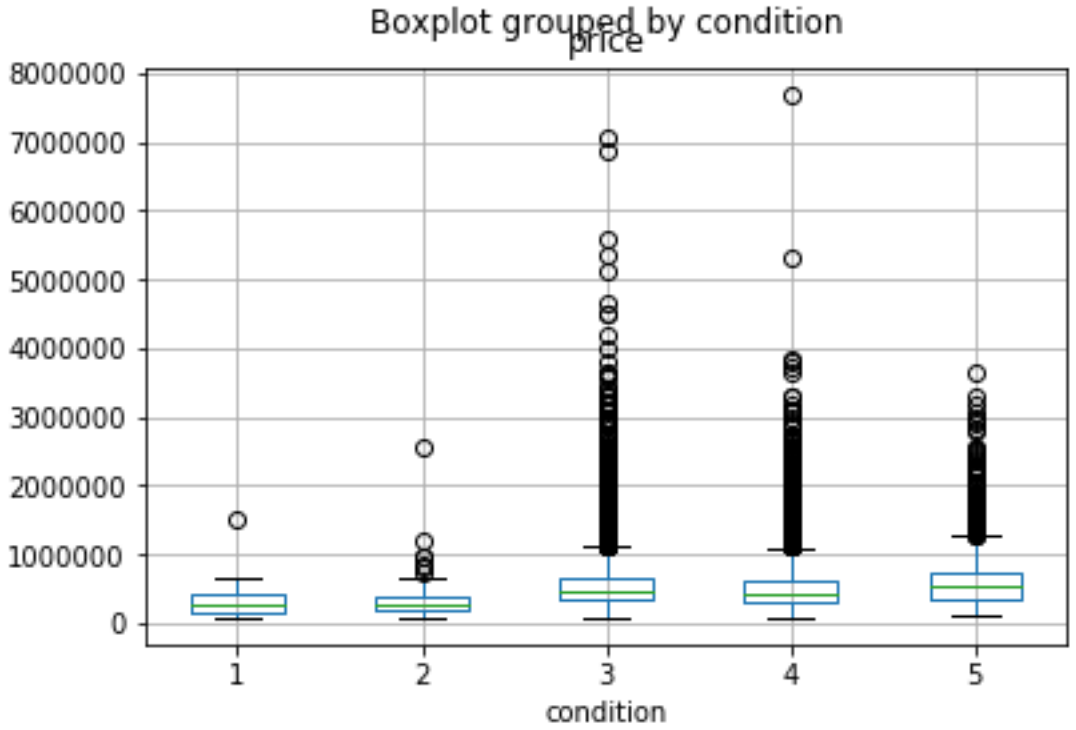
- Ideas to utilize zipcode, lat, and long
  - Without other resources
    - Data manipulation approach based on these three variables
  - With other resources
    - Additional useful information for these three variables

- Illustrate your ideas using Power Point
  - Some students have to create a video clip to explain their results and ideas

Condition

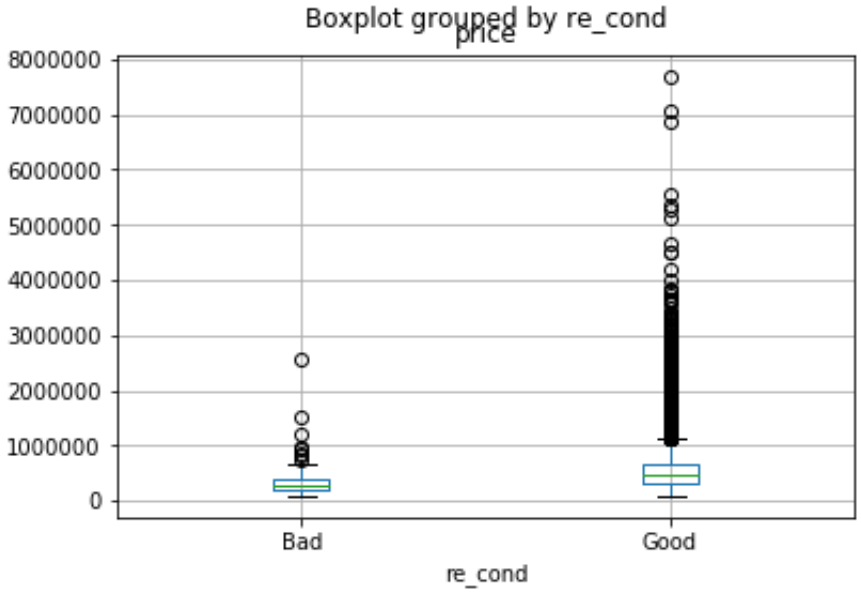


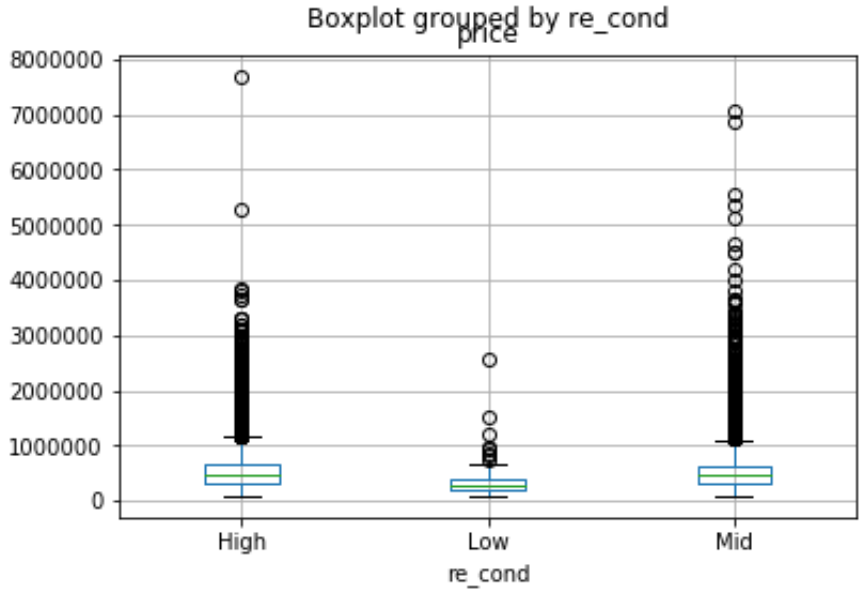
Condition


Boxplot grouped by condition
price
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
1
2
3
4
5
condition

Re-Binning

- Good and Bad
  - Good if condition> 2
- High, Mid, Low
  - High if condition > 3
  - Mid if condition= 3





OLS Model 1

OLS Regression Results

Dep. Variable:	price	R-squared:	0.598
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	1716.
Date:	Tue, 22 Sep 2020	Prob (F-statistic):	0.00
Time:	10:26:26	Log-Likelihood:	-2.3834e+05
No. Observations:	17290	AIC:	4.767e+05
Df Residuals:	17274	BIC:	4.768e+05
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.124e+06	1.7e+05	36.029	0.000	5.79e+06	6.46e+06
bedrooms	-5.814e+04	2395.751	-24.269	0.000	-6.28e+04	-5.34e+04
bathrooms	6.384e+04	4234.975	15.075	0.000	5.55e+04	7.21e+04
sqft_lot	-0.0092	0.061	-0.150	0.881	-0.129	0.110
floors	5.747e+04	4563.913	12.592	0.000	4.85e+04	6.64e+04
waterfront	6.74e+05	2.12e+04	31.745	0.000	6.32e+05	7.16e+05
sqft_above	240.1351	4.352	55.183	0.000	231.605	248.665
sqft_basement	243.6856	5.424	44.931	0.000	233.055	254.316
yr_built	-3256.4363	85.181	-38.230	0.000	-3423.400	-3089.472
yr_renovated	17.5940	4.695	3.747	0.000	8.391	26.797
sqft_living15	94.8250	4.129	22.964	0.000	86.731	102.919
sqft_lot15	-0.7240	0.091	-7.967	0.000	-0.902	-0.546
cond_2	7.765e+04	5.4e+04	1.437	0.151	-2.83e+04	1.84e+05
cond_3	1.182e+05	5.02e+04	2.353	0.019	1.97e+04	2.17e+05
cond_4	1.277e+05	5.03e+04	2.541	0.011	2.92e+04	2.26e+05
cond_5	1.682e+05	5.05e+04	3.328	0.001	6.91e+04	2.67e+05

Omnibus:	11985.490	Durbin-Watson:	2.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	607402.206
Skew:	2.753	Prob(JB):	0.00
Kurtosis:	31.510	Cond. No.	4.94e+06

OLS Model 2

OLS Regression Results

Dep. Variable:	price	R-squared:	0.598
Model:	OLS	Adj. R-squared:	0.597
Method:	Least Squares	F-statistic:	1975.
Date:	Tue, 22 Sep 2020	Prob (F-statistic):	0.00
Time:	10:27:21	Log-Likelihood:	-2.3835e+05
No. Observations:	17290	AIC:	4.767e+05
Df Residuals:	17276	BIC:	4.768e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.386e+06	1.62e+05	39.402	0.000	6.07e+06	6.7e+06
bedrooms	-5.814e+04	2397.825	-24.247	0.000	-6.28e+04	-5.34e+04
bathrooms	6.571e+04	4225.364	15.552	0.000	5.74e+04	7.4e+04
sqft_lot	-0.0078	0.061	-0.127	0.899	-0.127	0.112
floors	5.836e+04	4564.724	12.786	0.000	4.94e+04	6.73e+04
waterfront	6.751e+05	2.12e+04	31.777	0.000	6.33e+05	7.17e+05
sqft_above	239.7818	4.355	55.062	0.000	231.246	248.318
sqft_basement	244.5173	5.426	45.061	0.000	233.881	255.153
yr_built	-3321.7808	84.351	-39.381	0.000	-3487.117	-3156.445
yr_renovated	16.1085	4.691	3.434	0.001	6.913	25.304
sqft_living15	94.1772	4.131	22.798	0.000	86.080	102.274
sqft_lot15	-0.7268	0.091	-7.991	0.000	-0.905	-0.549
re_cond_Low	-6.997e+04	1.91e+04	-3.656	0.000	-1.07e+05	-3.25e+04
re_cond_Mid	-1.767e+04	4233.380	-4.174	0.000	-2.6e+04	-9373.403

Omnibus:	11925.661	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	596060.711
Skew:	2.736	Prob(JB):	0.00
Kurtosis:	31.239	Cond. No.	4.62e+06

Summary

```
train,test =train_test_split(house, test_size=0.2, random_state=30)
```

- Model 1
  - MSE for test set = 49845975570.53953
- Model 2
  - MSE for test set = 49791693775.603615

Dummy Variables

- If there are too many dummy variables compared to the number of samples, training a model may not work well
  - If there are other variables that can describe the characteristics of each category of the specific categorical variable
  - If there are too many categories, reduce the number of categories by checking the relationship between the variable and the target