Phase 2 Project Presentation

NAME: PETER KIGOTHO WAITI

GROUP: DSFP COHORT 15

~Moringa School

HOUSE SALES PREDICTION IN KING COUNTY USING REGRESSION MODELLING

Business Problem

- A real estate Agency need to provide prospective home sellers with guidance on how to improve the value of their home prior to listing, including the predicted increase in value expected based on improvements to particular features(renovation).

Business Question?

- What features of their home can prospective home sellers change or improve to increase the value of their home, and by amount could this increase be specific to certain features?

OBJECTIVES

- Import the required libraries
- Load the given data
- Inspect the data
- Perform data cleaning
- Begin regression modelling
- Ask relevant questions that need to be answered in the form of visualizations.
- Derive conclusions and recommendations.

DATA USED

- This project uses the King County House Sales dataset.
- The dataset contains information about the sale of each house as well as the number of predictor variables.

- price - Sale price (prediction target)
- date - Date house was sold
- bedrooms - Number of bedrooms
- bathrooms - Number of bathrooms
- Sqft-living - Square footage of living space in the home
- Sqft-lot - Square footage of the lot
- floors - Number of floors (levels) in house
- waterfront - Whether the house is on a waterfront
- view - Quality of view from house
- condition - How good the overall condition of the house is. Related to maintenance of house
- grade - Overall grade of the house. Related to the construction and design of the house
- Sqft-above - Square footage of house apart from basement
- Sqft-basement - Square footage of the basement
- Yr.-built - Year when house was built
- Yr.-renovated - Year when house was renovated
- zip code - ZIP Code used by the United States Postal Service

Data Cleaning

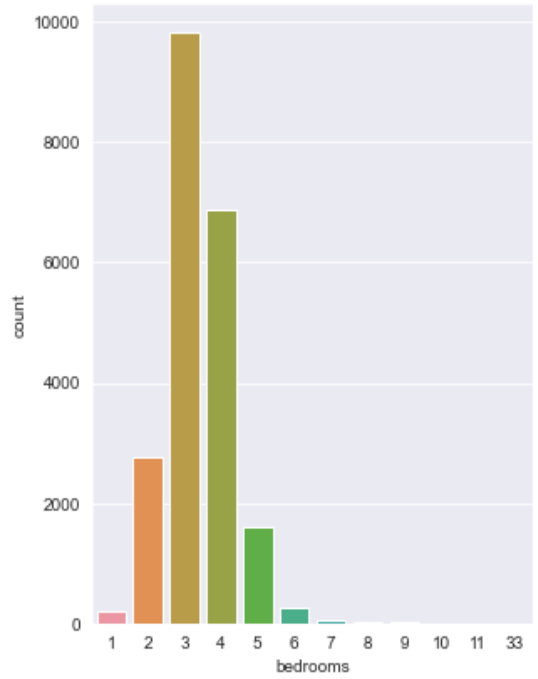
- Ensuring that each column has the correct data type.
- Check the percentage of missing values
- Drop or replace null values in the affected columns or rows.
- The replace technique used was filling the null values using the median.
- Why median?

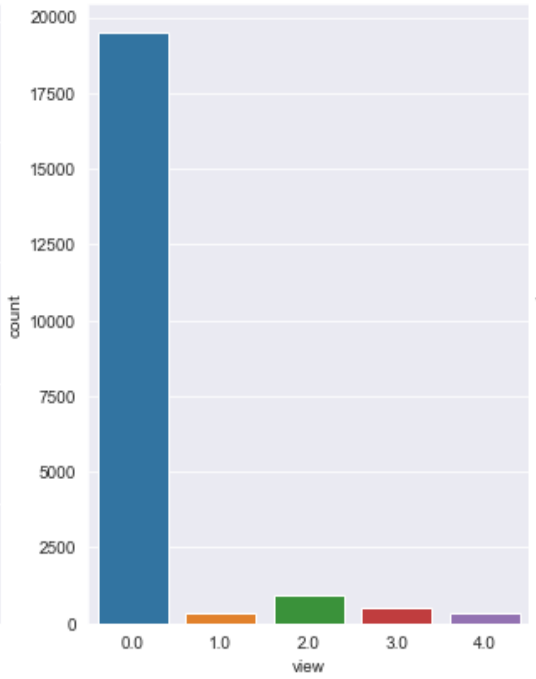
  The choice was informed by the fact that the mean could be affected by outliers and thus could offset our values or rather may end up giving a false impression.
- Check for duplicates.

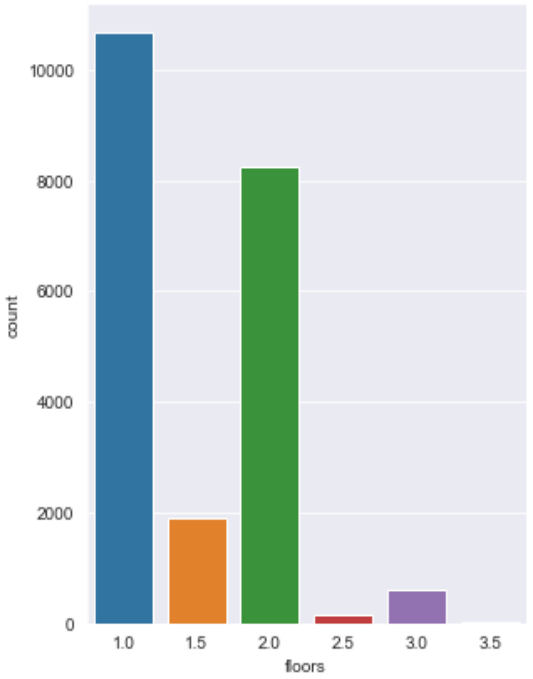
EXPLORATORY DATA ANALYSIS

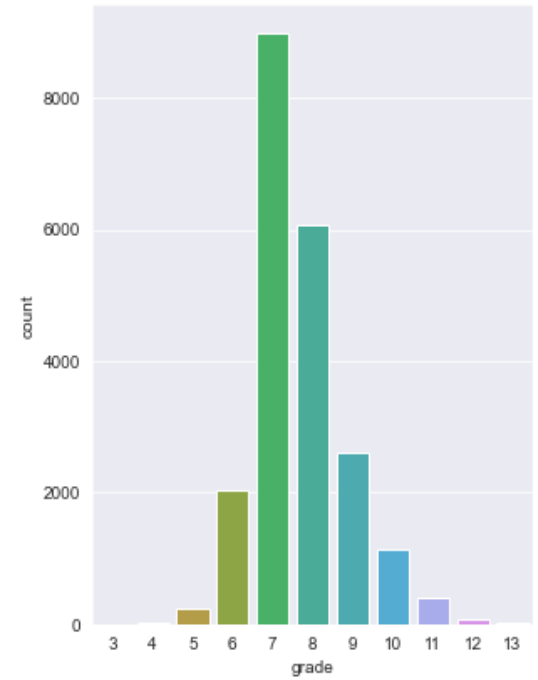
Univariate Analysis

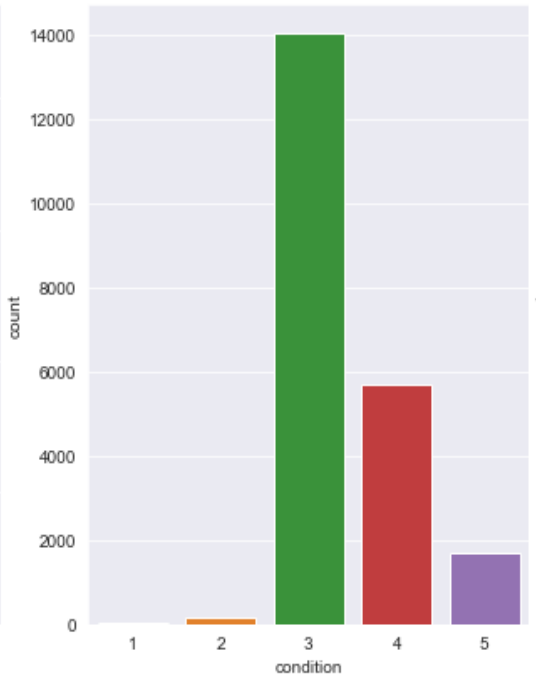
- From the distribution above we can see that most houses range from $0 to around $1.7. We can also see that there are some outliers in the distribution.

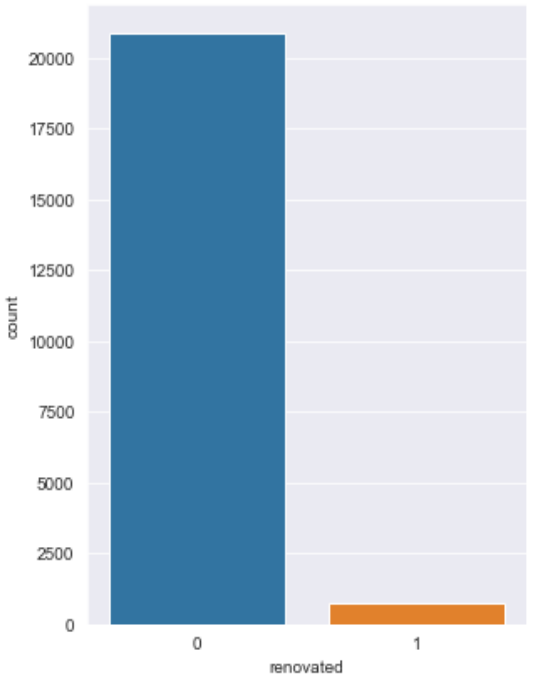








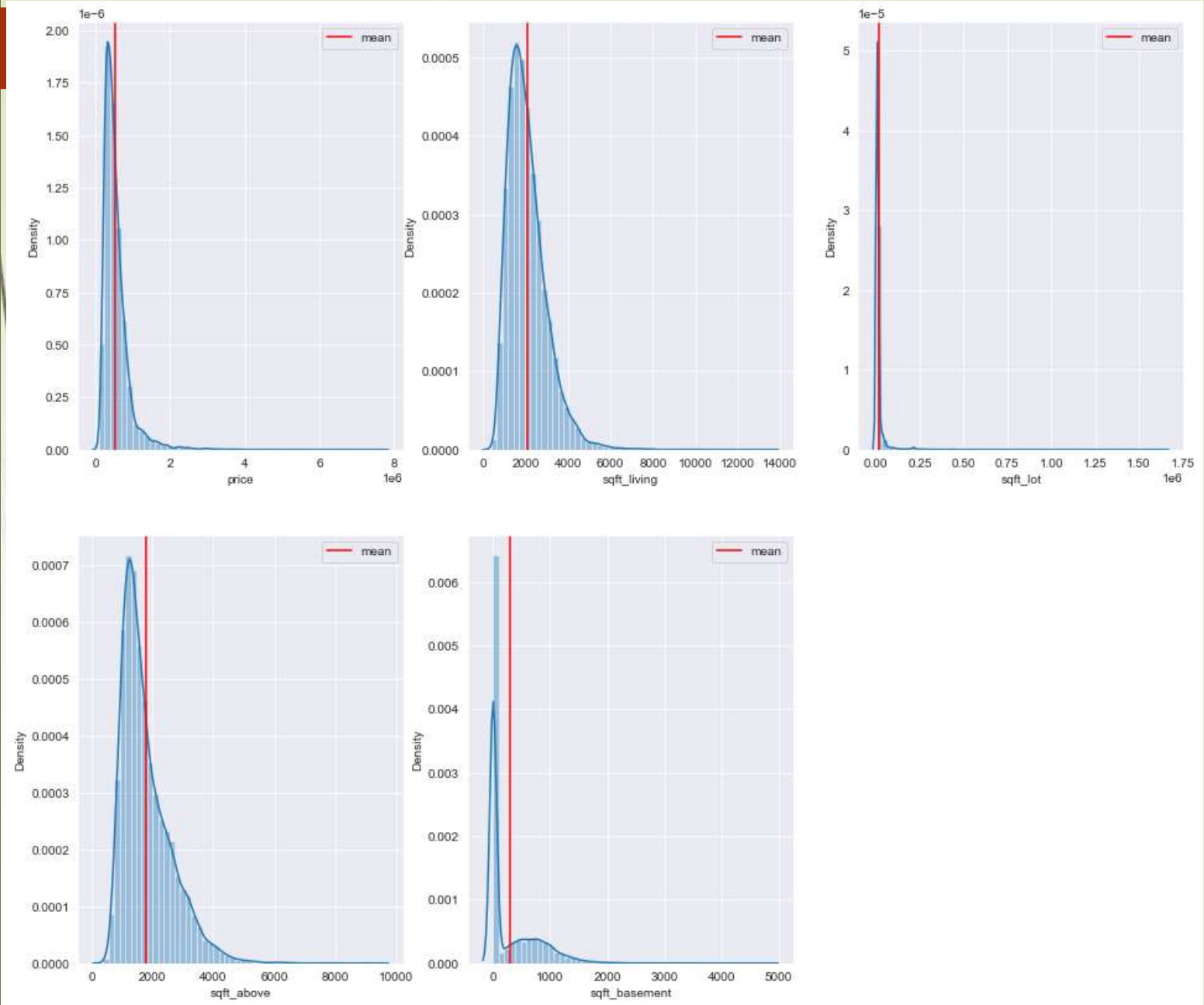




EXPLORATORY DATA ANALYSIS

Univariate Analysis (categorical variables)

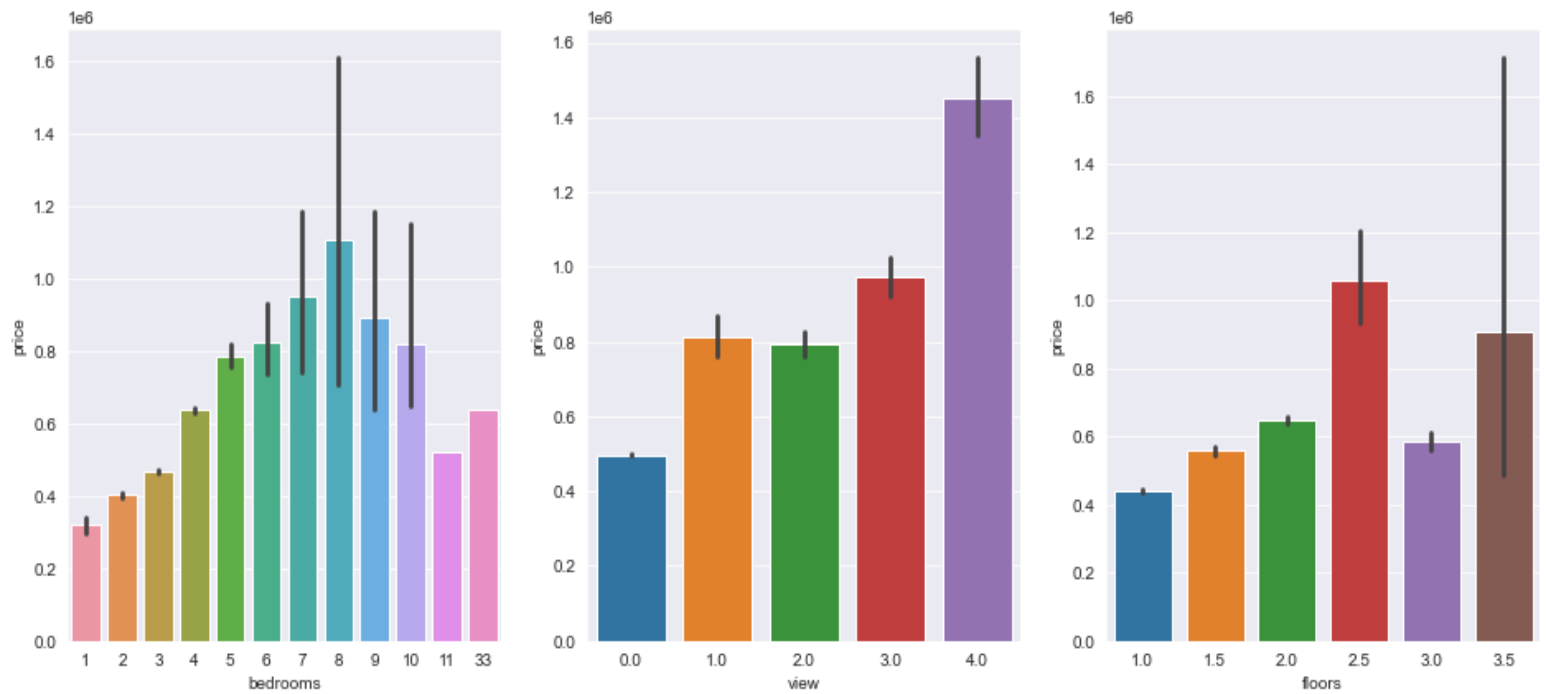
- From the distribution we can see that 3 and 4 bedroom houses are with the most count.
- Most houses have low quality of view. Most houses have an average grade and an average condition.
- 1 floor and 2 floor houses have the highest count.
- Most houses range from $0 to around $1.7. We can also see that there are some outliers in the distribution.

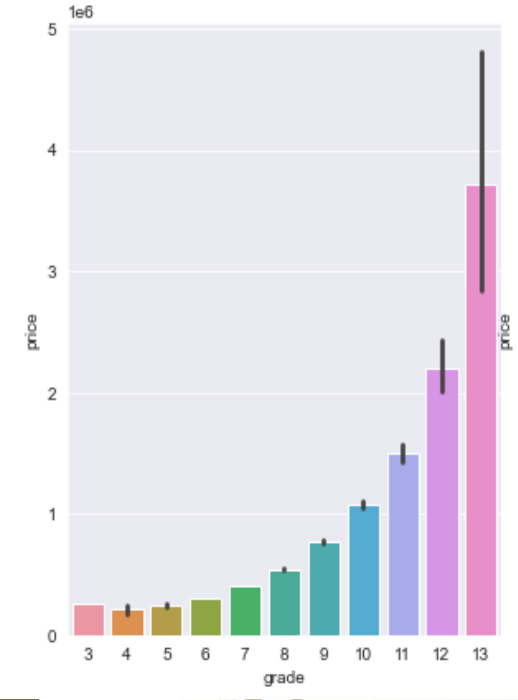


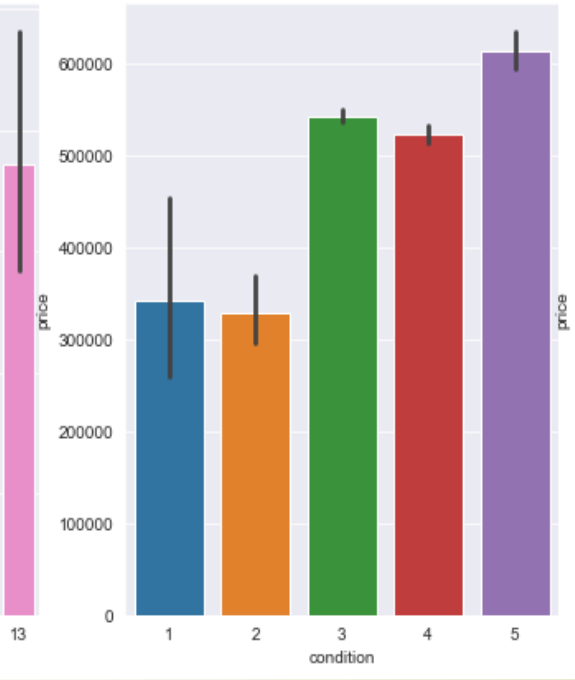
Q1: What is the distribution of all numerical features?

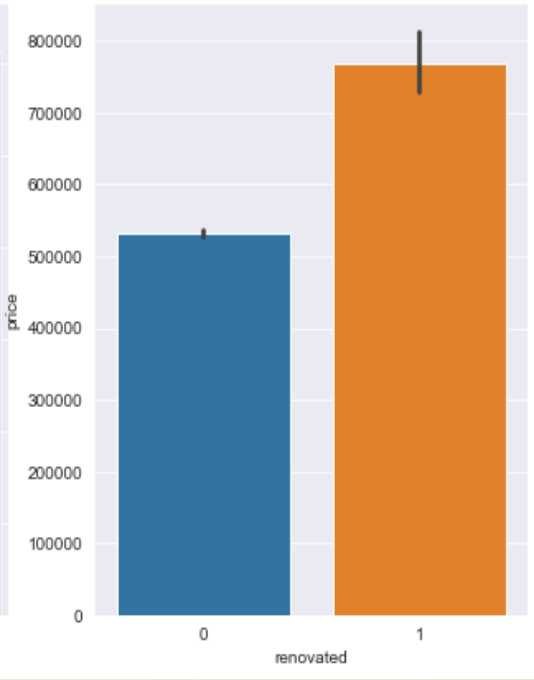
This is to check the mean and distribution of the numerical columns.

EXPLORATORY DATA ANALYSIS









Bivariate Analysis

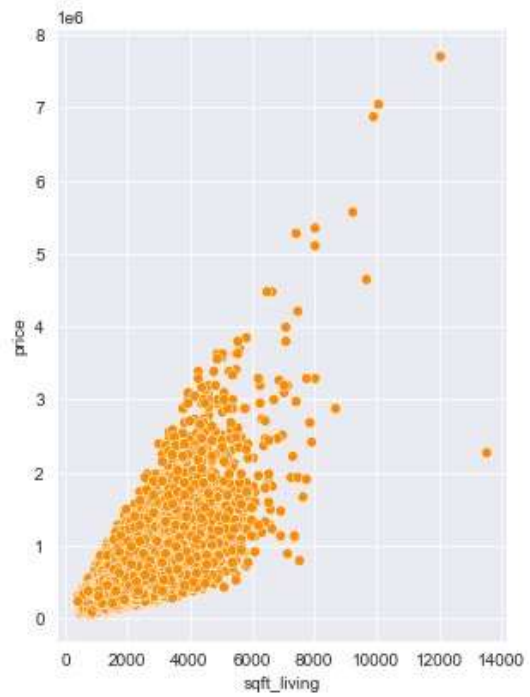
Q2: What is the relationship between price and categorical variables?

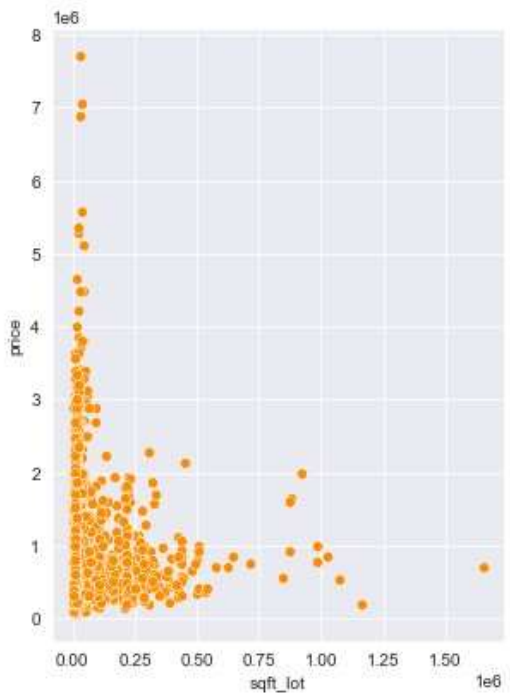
The plots show the relationship between price and categorical variables.

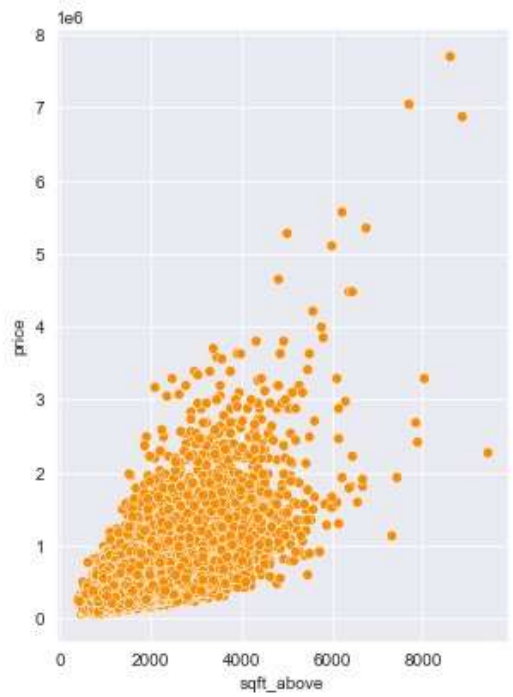
From the plots, we can see that as the number of bedrooms, floors, views, conditions, and grades increase, the price also increases.

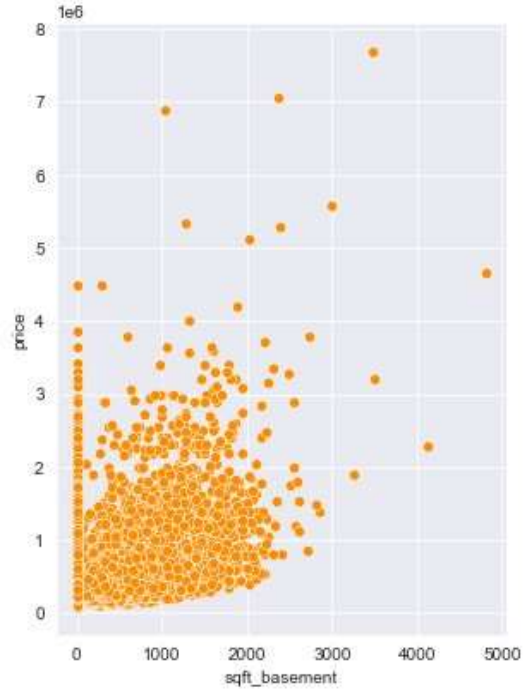
Renovation also increases house prices.

EXPLORATORY DATA ANALYSIS









Bivariate Analysis

Q3: What is the relationship between price and continuous variables?

Sqft_living, above, and basement all have a linear relationship with price.

Sqft_loft has a minimal relationship with price.

EXPLORATORY DATA ANALYSIS

Bivariate Analysis

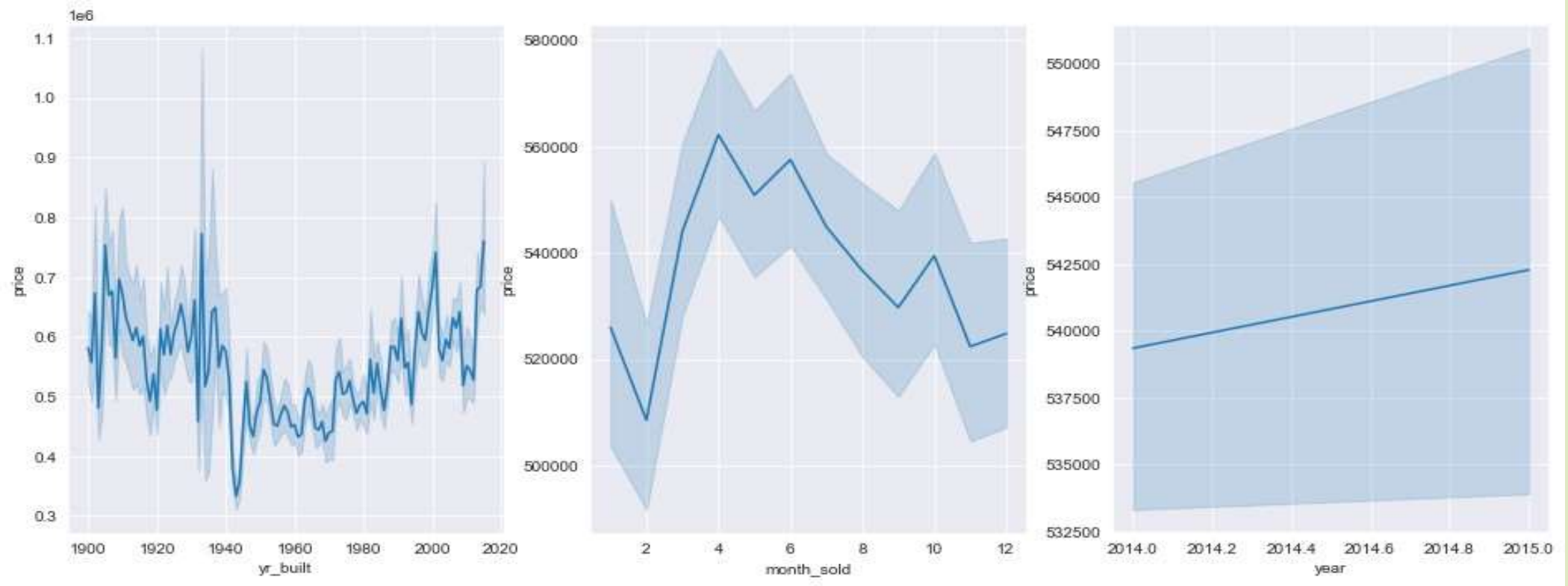
Q4: What is the relationship between price and time-series data?

From the graphs below;

I. There is no relationship between price and year built.

II. Months 4 and 6 had houses with the highest price.

III. There is no relationship between the price and year sold.

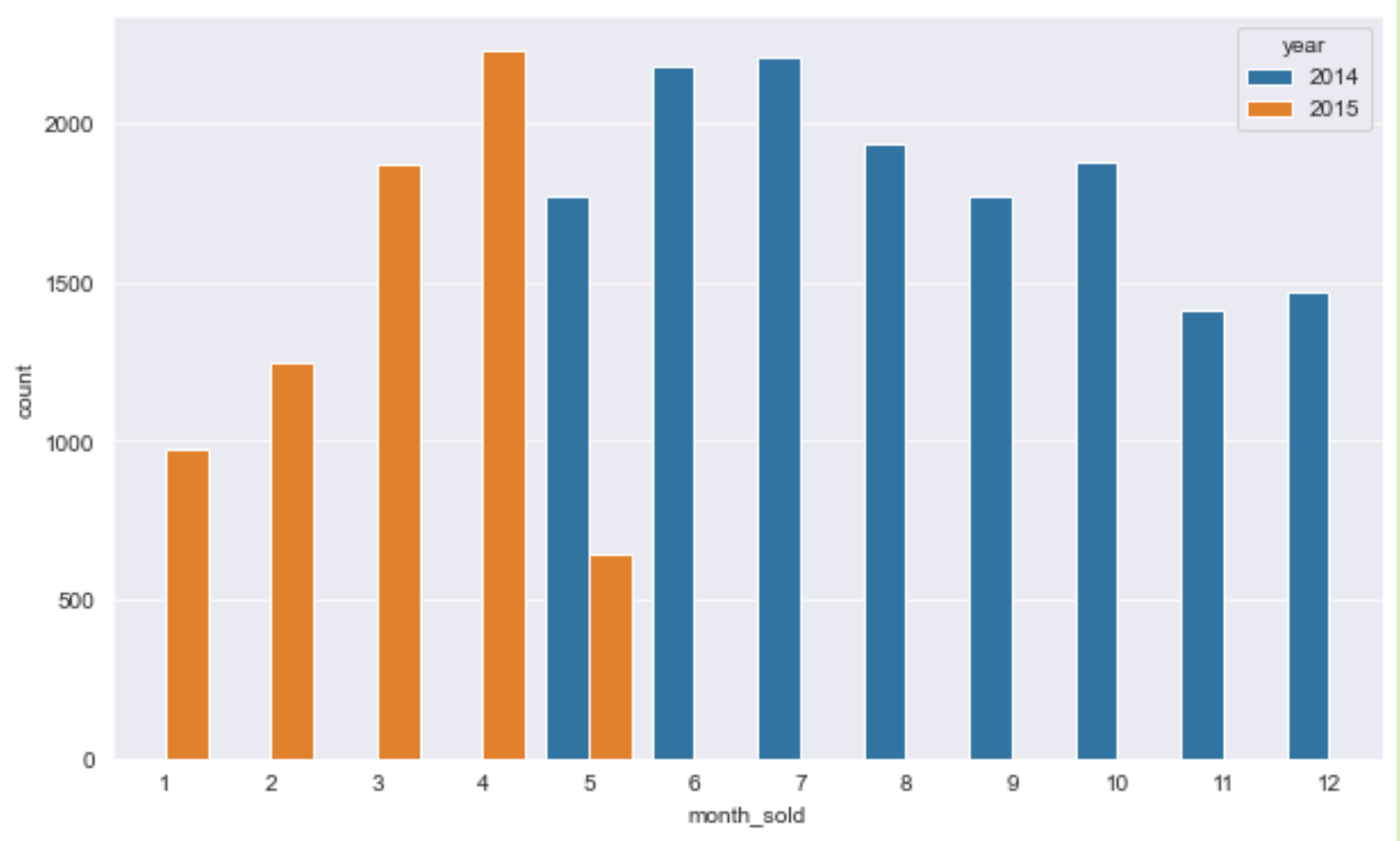


EXPLORATORY DATA ANALYSIS

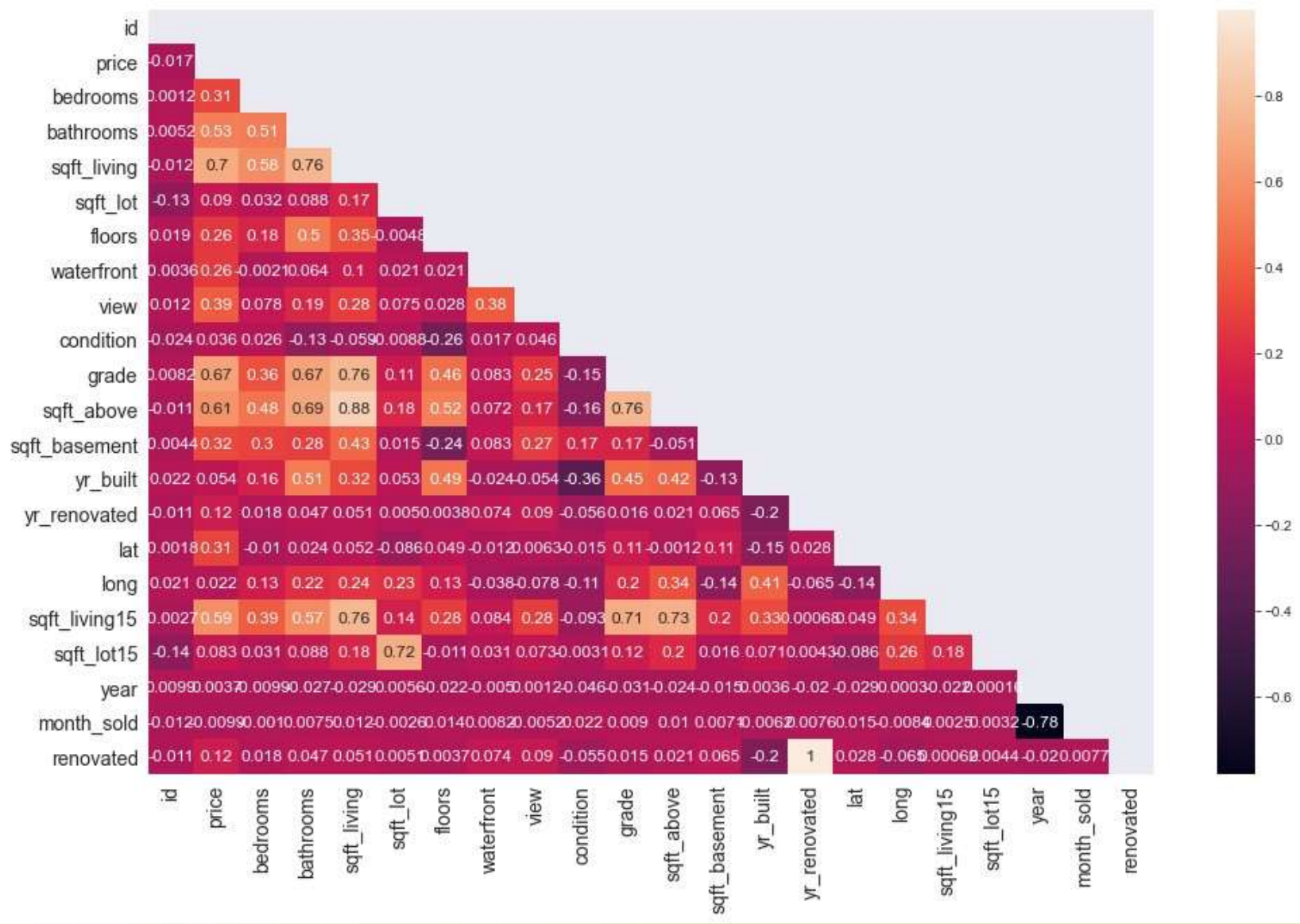
Bivariate Analysis

Q5: Which month and year did the houses sell the most?

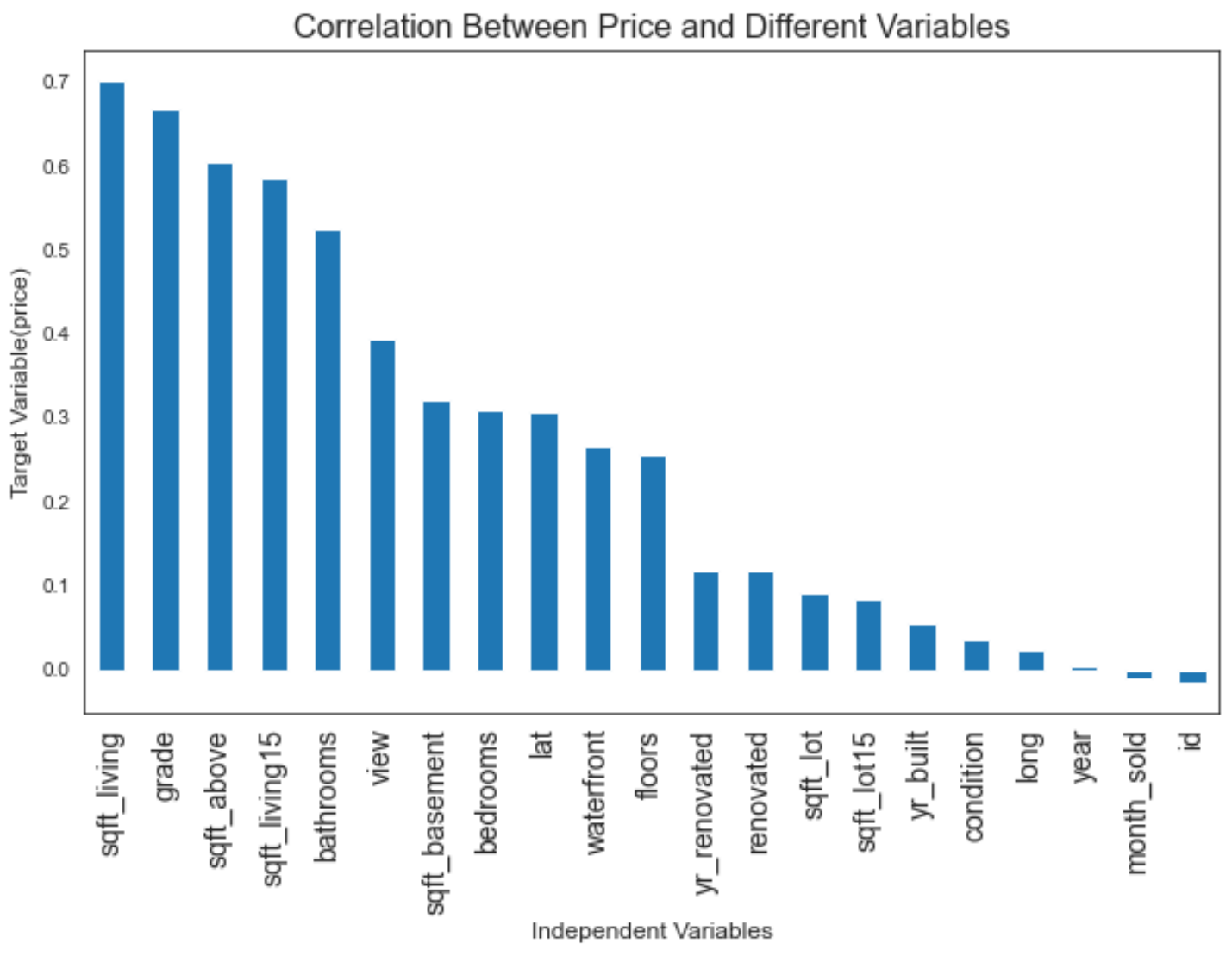
Upon further analysis, the house sales began in May 2014 till May 2015.



Q6: What is the correlation between the variables of the given dataset?



Q7: What is the correlation between price(target variable) and independent variables?



	price
price	1.000000
sqft_living	0.701917
grade	0.667951
sqft_above	0.605368
sqft_living15	0.585241
bathrooms	0.525906
view	0.393497
sqft_basement	0.321108
bedrooms	0.308787
lat	0.306692
waterfront	0.264306
floors	0.256804
yr_renovated	0.117855
renovated	0.117543
sqft_lot	0.089876
sqft_lot15	0.082845
yr_built	0.053953
condition	0.036056
long	0.022036
id	0.016772
month_sold	0.009928
year	0.003727

Checking for Multicollinearity

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	...	sqft_basement	yr_built	yr_renovated	lat	lor
id	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	Fal
price	False	True	False	False	True	False	False	False	False	False	...	False	False	False	False	Fal
bedrooms	False	False	True	False	False	False	False	False	False	False	...	False	False	False	False	Fal
bathrooms	False	False	False	True	True	False	False	False	False	False	...	False	False	False	False	Fal
sqft_living	False	True	False	True	True	False	False	False	False	False	...	False	False	False	False	Fal
sqft_lot	False	False	False	False	False	True	False	False	False	False	...	False	False	False	False	Fal
floors	False	False	False	False	False	False	True	False	False	False	...	False	False	False	False	Fal
waterfront	False	False	False	False	False	False	False	True	False	False	...	False	False	False	False	Fal
view	False	False	False	False	False	False	False	False	True	False	...	False	False	False	False	Fal
condition	False	False	False	False	False	False	False	False	False	True	...	False	False	False	False	Fal
grade	False	False	False	False	True	False	False	False	False	False	...	False	False	False	False	Fal
sqft_above	False	False	False	False	True	False	False	False	False	False	...	False	False	False	False	Fal
sqft_basement	False	False	False	False	False	False	False	False	False	False	...	True	False	False	False	Fal
yr_built	False	False	False	False	False	False	False	False	False	False	...	False	True	False	False	Fal
yr_renovated	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	Fal
lat	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	Fal
long	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	Tru
sqft_living15	False	False	False	False	True	False	False	False	False	False	...	False	False	False	False	Fal
sqft_lot15	False	False	False	False	False	True	False	False	False	False	...	False	False	False	False	Fal
year	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	Fal
month_sold	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	Fal
renovated	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	Fal

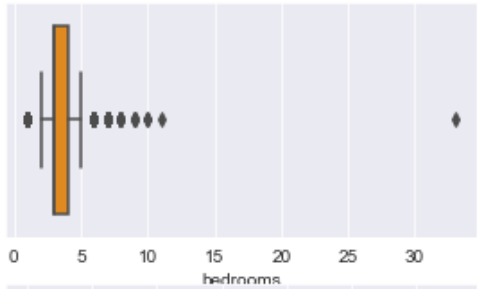
	cc
pairs	
(yr_renovated, renovated)	0.999968
(sqft_above, sqft_living)	0.876448
(year, month_sold)	0.782325
(grade, sqft_living)	0.762779
(sqft_living15, sqft_living)	0.756402
(grade, sqft_above)	0.756073
(sqft_living, bathrooms)	0.755758

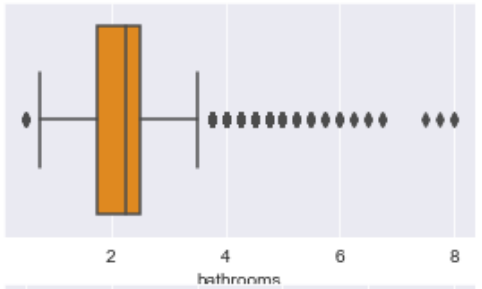
Q8: Which columns to drop?

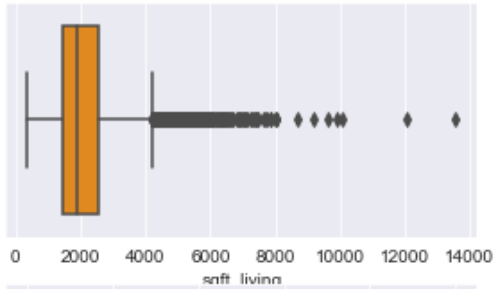
- From the table above I realized that **sqft-living** has **high correlation** with **bathrooms, grade, sqft- above,** and **sqft-living15.** Also, I noticed **grade** has high correlation with **sqft-living, sqft-above.** From this observation, I think I will drop both **sqft-living and grade** since they might cause problems of Multicollinearity to the model.

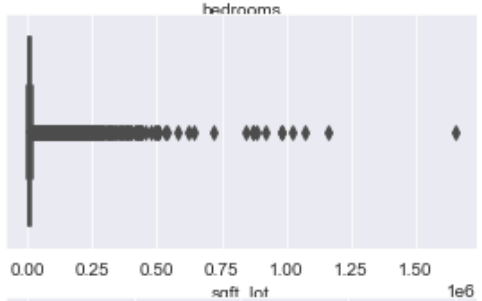
- since id doesn't have any correlation with house prices we drop the column.
- I dropped the "zip code" column, because if I feed it to our model it would be considered as a continuous value although it isn't.
- I dropped the date column since we already split the date into month and year sold.

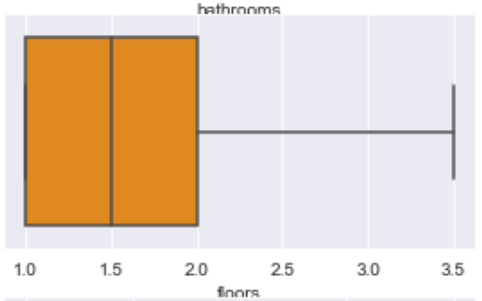
EXPLORATORY DATA ANALYSIS

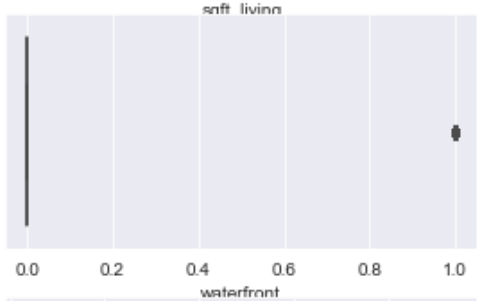


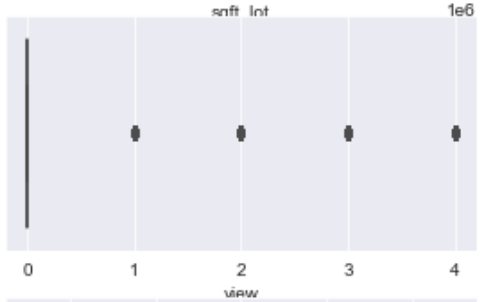


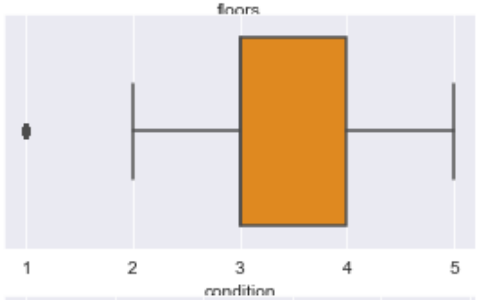


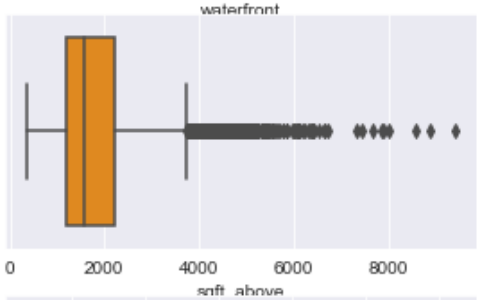


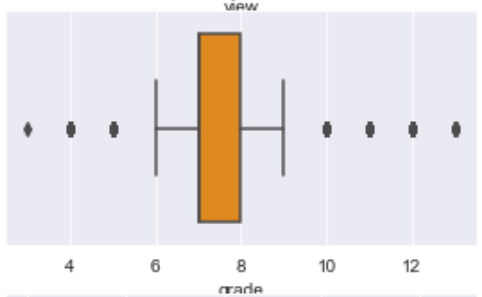


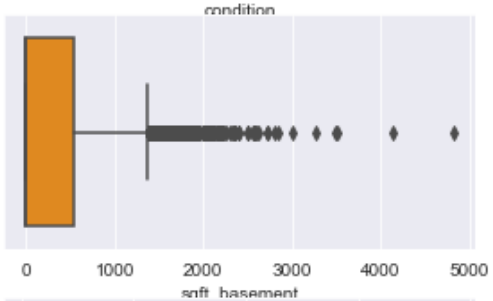


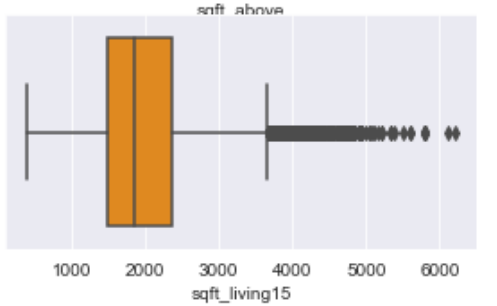


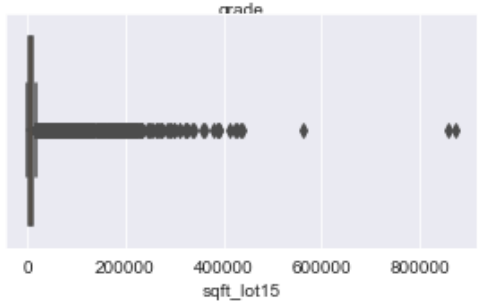


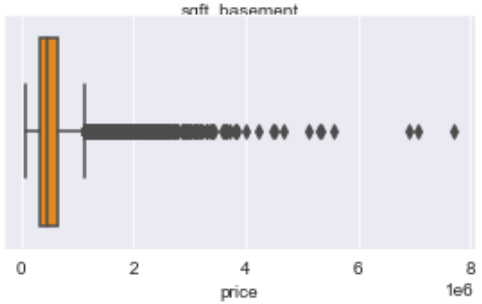










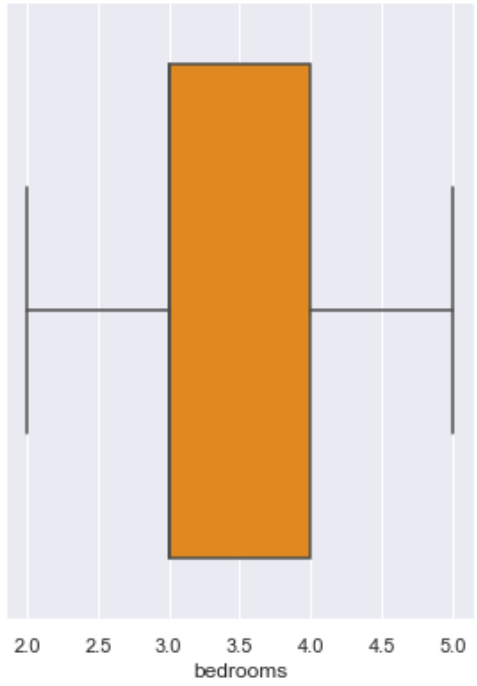


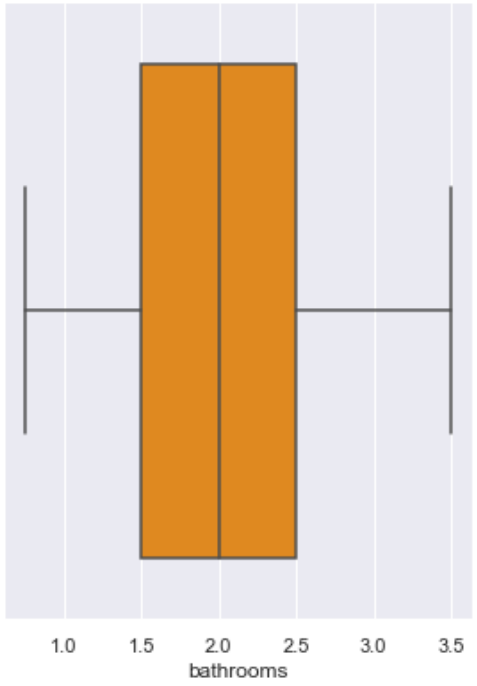
Check for Outliers

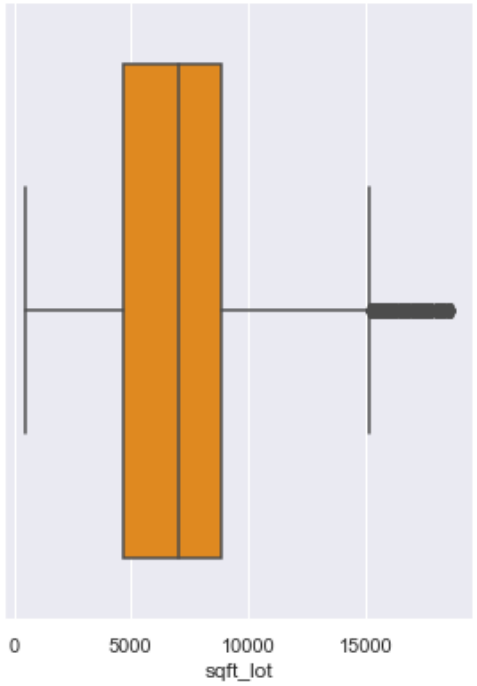
From these charts, most of the data contain outliers.

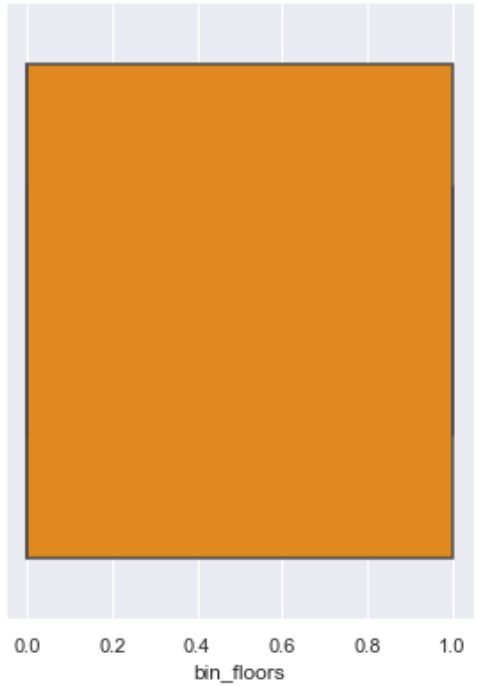
Having outliers can alter the result of the model therefore they need to be removed.

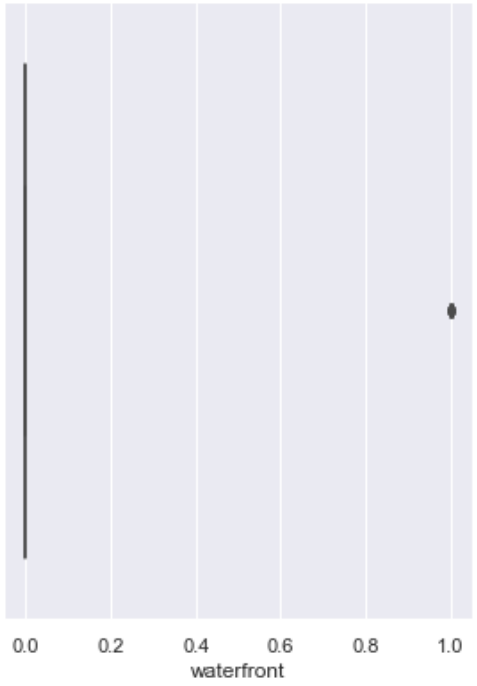
EXPLORATORY DATA ANALYSIS

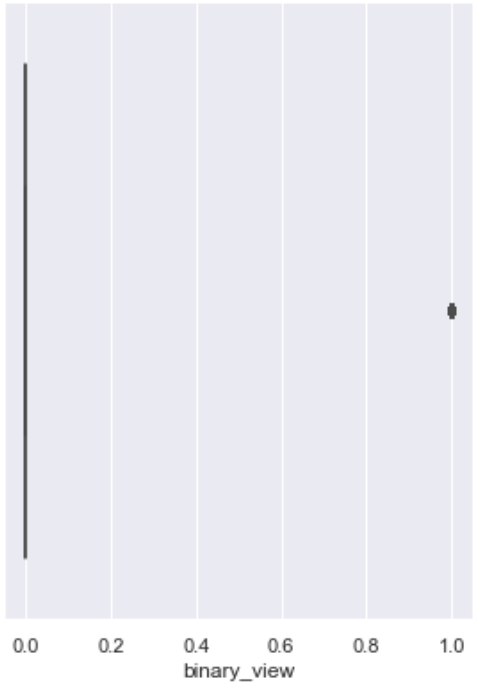












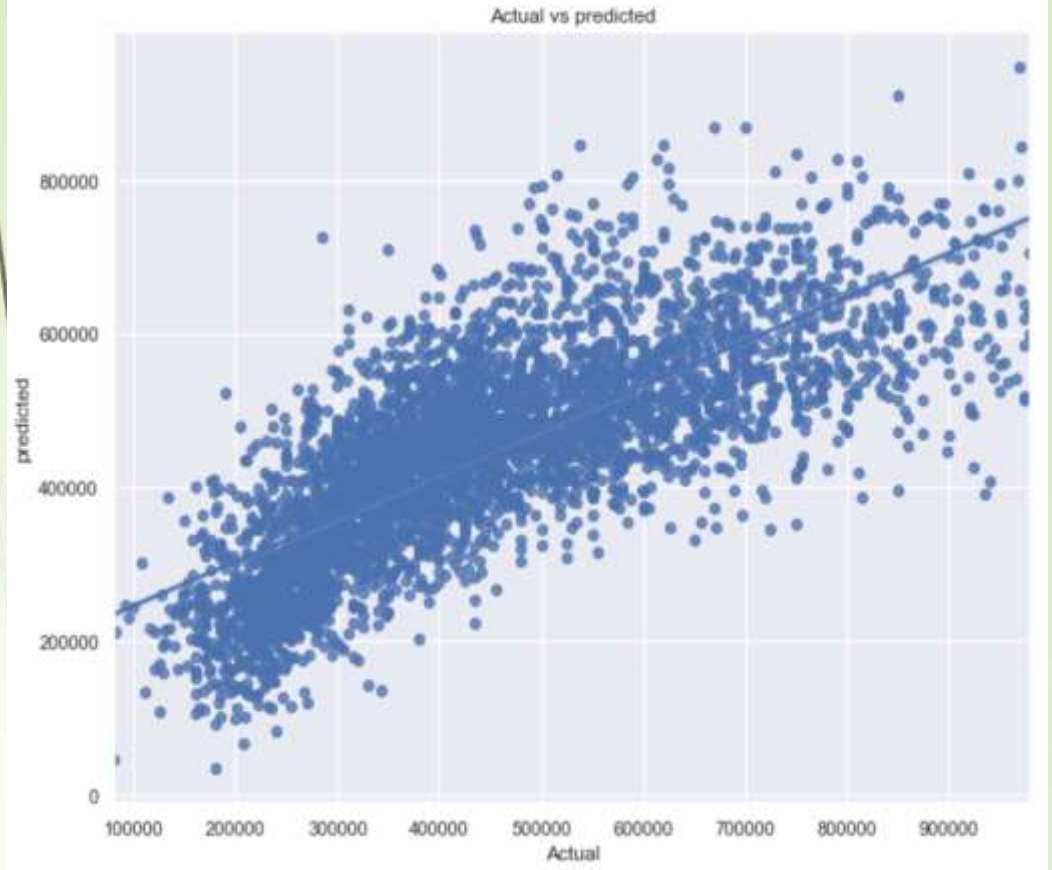
Check for outliers after removal

The charts show that there are no outliers as they have all been removed.

Regression Modelling

- To tackle the business question, the variables are prepared to be fed into the model.
- Train the model by feeding it with independent variables as ‘y’ and the dependent variable as ‘x’.
- In one of the models, renovations is included and in the other, it is excluded.
- Prediction is then performed to obtain results.

LINEAR REGRESSION MODELLING



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.579
Model:	OLS	**Adj. R-squared:**	0.578
Method:	Least Squares	**F-statistic:**	1524.
Date:	Tue, 05 Jul 2022	**Prob (F-statistic):**	0.00
Time:	19:32:47	**Log-Likelihood:**	-1.5996e+05
No. Observations:	12222	**AIC:**	3.199e+05
Df Residuals:	12210	**BIC:**	3.200e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.429e+05	1057.690	418.780	0.000	4.41e+05	4.45e+05
x1	7007.3102	1066.625	6.570	0.000	4916.557	9098.064
x2	6.875e+04	1871.327	36.740	0.000	6.51e+04	7.24e+04
x3	3.901e+04	1207.233	32.312	0.000	3.66e+04	4.14e+04
x4	8.466e+04	1094.219	77.373	0.000	8.25e+04	8.68e+04
x5	-1.056e+04	1220.632	-8.649	0.000	-1.29e+04	-8164.262
x6	3.818e+04	1625.570	23.489	0.000	3.5e+04	4.14e+04
x7	-2.176e+04	1289.117	-16.877	0.000	-2.43e+04	-1.92e+04
x8	1.29e+04	1065.481	12.105	0.000	1.08e+04	1.5e+04
x9	1.861e+04	1102.407	16.883	0.000	1.65e+04	2.08e+04
x10	3964.2001	1058.545	3.745	0.000	1889.284	6039.117
x11	4168.2373	1533.222	2.719	0.007	1162.879	7173.596

Omnibus:	758.130	**Durbin-Watson:**	1.972
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1010.575
Skew:	0.570	**Prob(JB):**	3.60e-220
Kurtosis:	3.827	**Cond. No.**	3.46

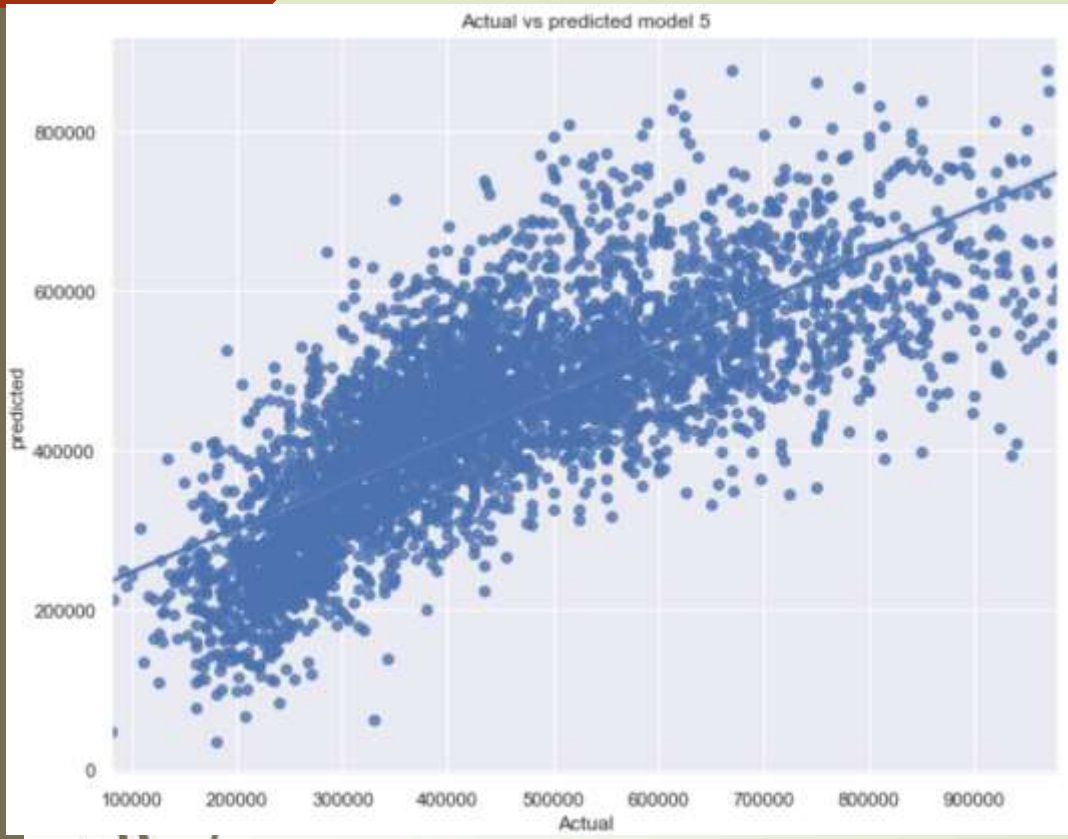
Out[90]:

	index	PREDICTIONS	ACTUAL VALUES	error
0	0	402299.393582	394000.0	-8299.393582
1	1	504469.523846	700000.0	195530.476154
2	2	661818.993613	870000.0	208181.006387
3	3	420110.497760	445000.0	24889.502240
4	4	465656.695421	450000.0	-15656.695421

In [91]:
```
performance['error'].abs().mean()
```

Out[91]: 90571.12200332498

LINEAR REGRESSION MODELLING



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.574
Model:	OLS	**Adj. R-squared:**	0.573
Method:	Least Squares	**F-statistic:**	1643.
Date:	Tue, 05 Jul 2022	**Prob (F-statistic):**	0.00
Time:	19:33:02	**Log-Likelihood:**	-1.6003e+05
No. Observations:	12222	**AIC:**	3.201e+05
Df Residuals:	12211	**BIC:**	3.202e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.429e+05	1063.974	416.307	0.000	4.41e+05	4.45e+05
x1	7607.5917	1071.802	7.098	0.000	5506.690	9708.493
x2	6.967e+04	1880.915	37.038	0.000	6.6e+04	7.34e+04
x3	3.979e+04	1212.661	32.813	0.000	3.74e+04	4.22e+04
x4	8.495e+04	1100.460	77.196	0.000	8.28e+04	8.71e+04
x5	-1.116e+04	1226.869	-9.095	0.000	-1.36e+04	-8752.923
x6	3.691e+04	1631.826	22.622	0.000	3.37e+04	4.01e+04
x7	-2.155e+04	1296.668	-16.623	0.000	-2.41e+04	-1.9e+04
x8	1.916e+04	1108.031	17.289	0.000	1.7e+04	2.13e+04
x9	4014.3685	1064.826	3.770	0.000	1927.140	6101.597
x10	4541.7673	1542.019	2.945	0.003	1519.165	7564.369

Omnibus:	812.584	**Durbin-Watson:**	1.971
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1092.690
Skew:	0.596	**Prob(JB):**	5.31e-238
Kurtosis:	3.851	**Cond. No.**	3.45

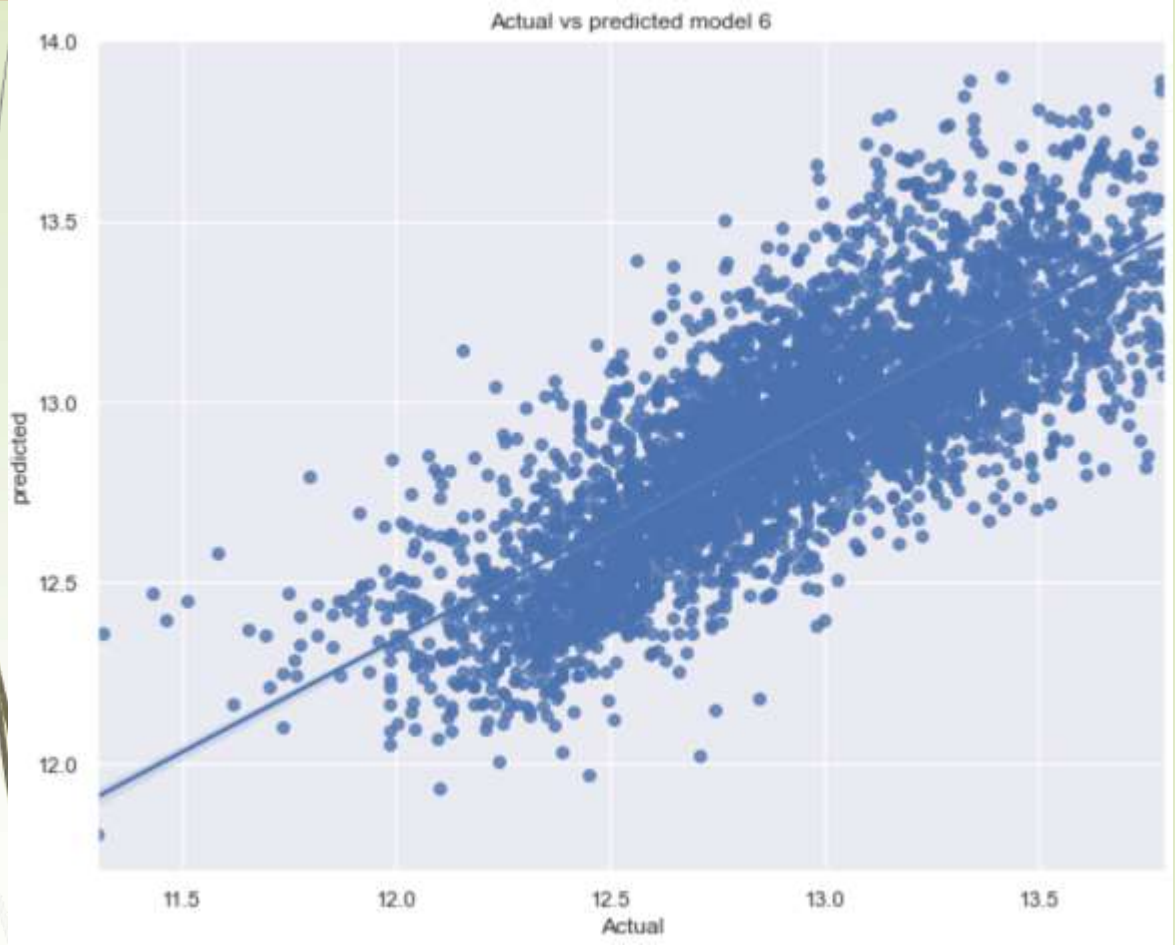
	PREDICTIONS	ACTUAL VALUES	error
8138	402830.697976	394000.0	-8830.697976
15491	507284.436673	700000.0	192715.563327
20451	666792.849878	870000.0	203207.150122
21179	421900.789950	445000.0	23099.210050
376	468365.459426	450000.0	-18365.459426
...	...	...	...
11671	506545.153303	500000.0	-6545.153303
13899	688298.614154	480000.0	-208298.614154
18020	650857.213944	585000.0	-65857.213944
6681	323585.935826	350000.0	26414.064174
5918	449638.341424	575000.0	125361.658576

4074 rows × 3 columns

```
performance5['error'].abs().mean()
```

```
90860.18515098584
```

LINEAR REGRESSION MODELLING



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.624
Model:	OLS	**Adj. R-squared:**	0.624
Method:	Least Squares	**F-statistic:**	2028.
Date:	Tue, 05 Jul 2022	**Prob (F-statistic):**	0.00
Time:	19:33:04	**Log-Likelihood:**	-684.56
No. Observations:	12222	**AIC:**	1391.
Df Residuals:	12211	**BIC:**	1473.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.9171	0.002	5577.671	0.000	12.913	12.922
x1	0.0161	0.002	6.900	0.000	0.012	0.021
x2	0.1523	0.004	37.194	0.000	0.144	0.160
x3	0.0937	0.003	35.511	0.000	0.089	0.099
x4	0.2181	0.002	91.067	0.000	0.213	0.223
x5	-0.0151	0.003	-5.652	0.000	-0.020	-0.010
x6	0.0899	0.004	25.300	0.000	0.083	0.097
x7	-0.0566	0.003	-20.057	0.000	-0.062	-0.051
x8	0.0383	0.002	15.866	0.000	0.034	0.043
x9	0.0145	0.002	6.238	0.000	0.010	0.019
x10	0.0159	0.003	4.734	0.000	0.009	0.022

Omnibus:	142.052	**Durbin-Watson:**	1.969
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	200.299
Skew:	-0.147	**Prob(JB):**	3.20e-44
Kurtosis:	3.554	**Cond. No.**	3.45

	PREDICTIONS	ACTUAL VALUES	error
8138	12.858362	12.884106	0.025744
15491	13.061402	13.458836	0.397433
20451	13.452331	13.676248	0.223917
21179	12.906882	13.005830	0.098947
376	12.985235	13.017003	0.031768
...	...	...	...
11671	13.076565	13.122363	0.045799
13899	13.532767	13.081541	-0.451225
18020	13.402968	13.279367	-0.123601
6681	12.629390	12.765688	0.136298
5918	12.927189	13.262125	0.334937

4074 rows × 3 columns

```
performance6['error'].abs().mean()
```

```
0.19975417315287974
```

Conclusion

- Considering fewer houses have been renovated, the model showed an increase in price when renovation was included in the model.
- The increase in price was by 294$.
- Various independent variables have a linear relationship with price. These include grade, condition, view, number of bedrooms, and the size of the house.
- In conclusion, it is advisable to renovate a majority of the houses as the house prices will also rise.