John Kundycki

Predicting Real Estate Prices

King County Housing Dataset

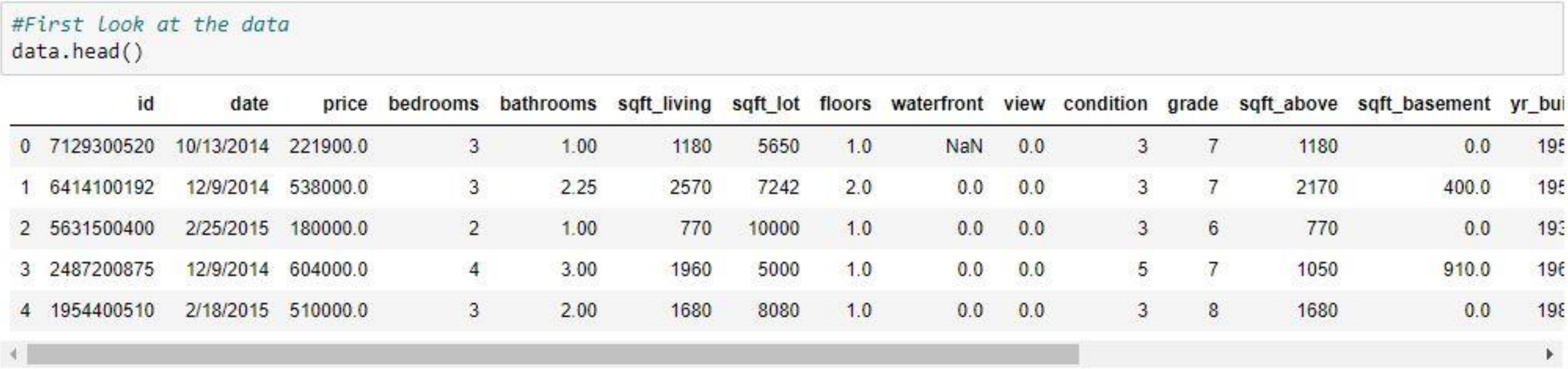
Data Science Life Cycle: OSEMN

OSEMN:

Obtain - Gather Data from relevant resources
Scrub - Clean data to formats that machine understands
Explore - Find significant patterns and trends using statistical methods
Model - Construct models to predict and forecast
Interpret - Put the results into good use

OBTAIN

Provided data on housing prices and different features for houses in King County, Seattle, WA

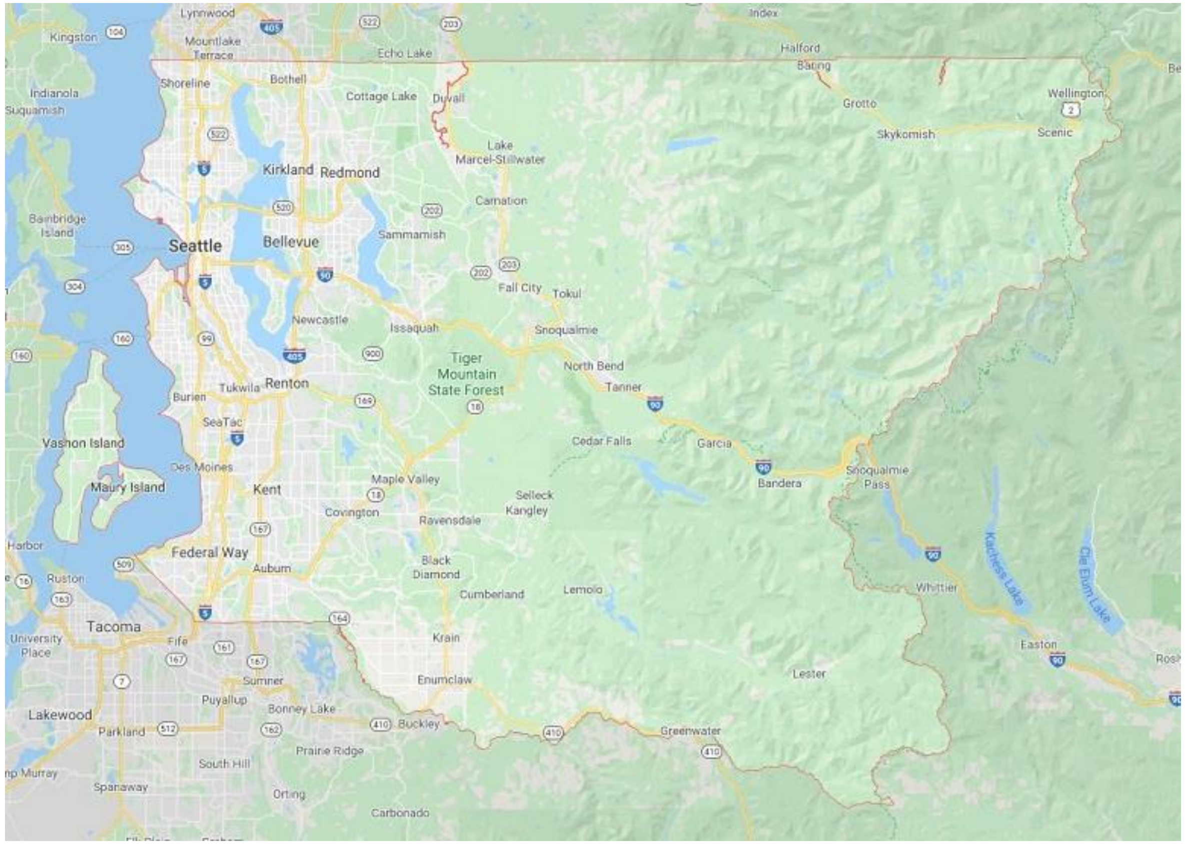


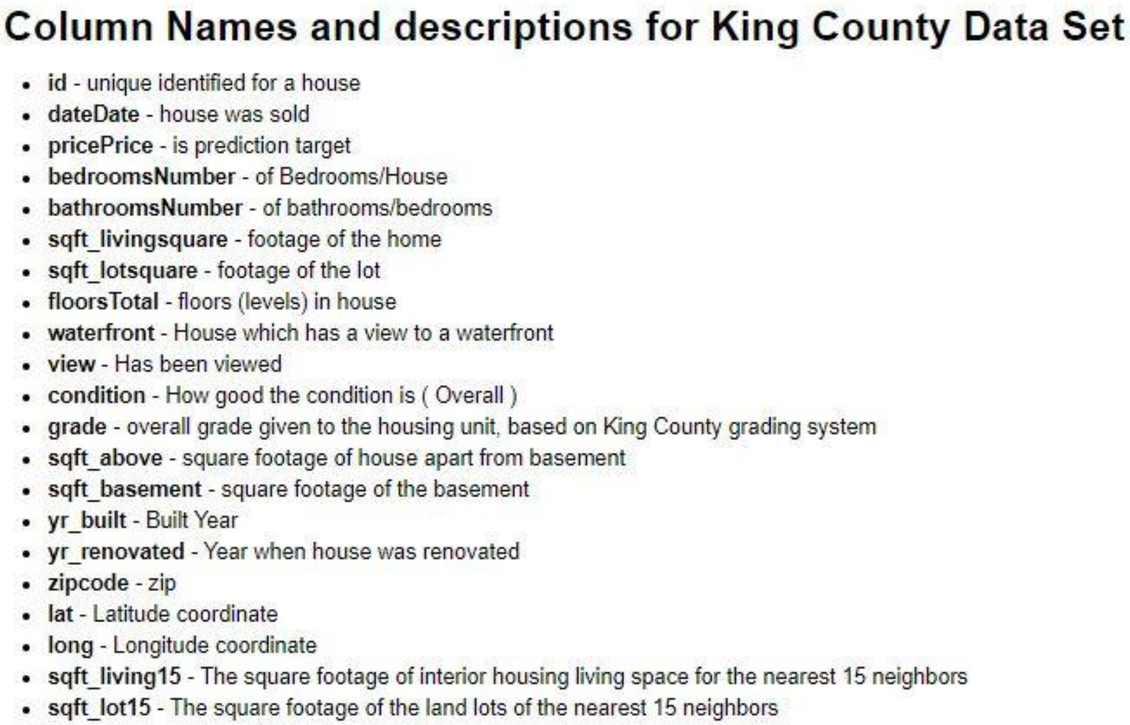
```
#First look at the data
data.head()
```

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_bui
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN	0.0	3	7	1180	0.0	195
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	3	7	2170	400.0	195
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	3	6	770	0.0	193
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	5	7	1050	910.0	196
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	3	8	1680	0.0	198

OBTAIN

More information about the dataset



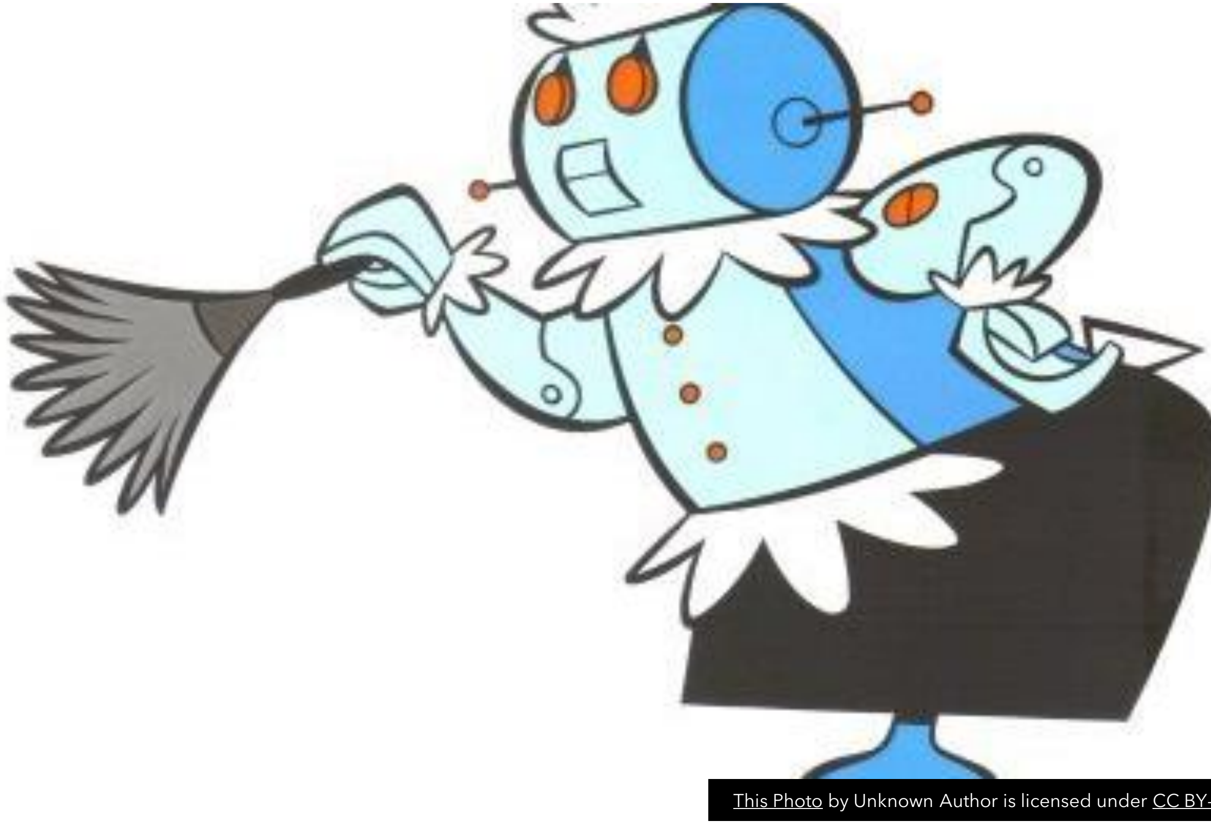


Column Names and descriptions for King County Data Set

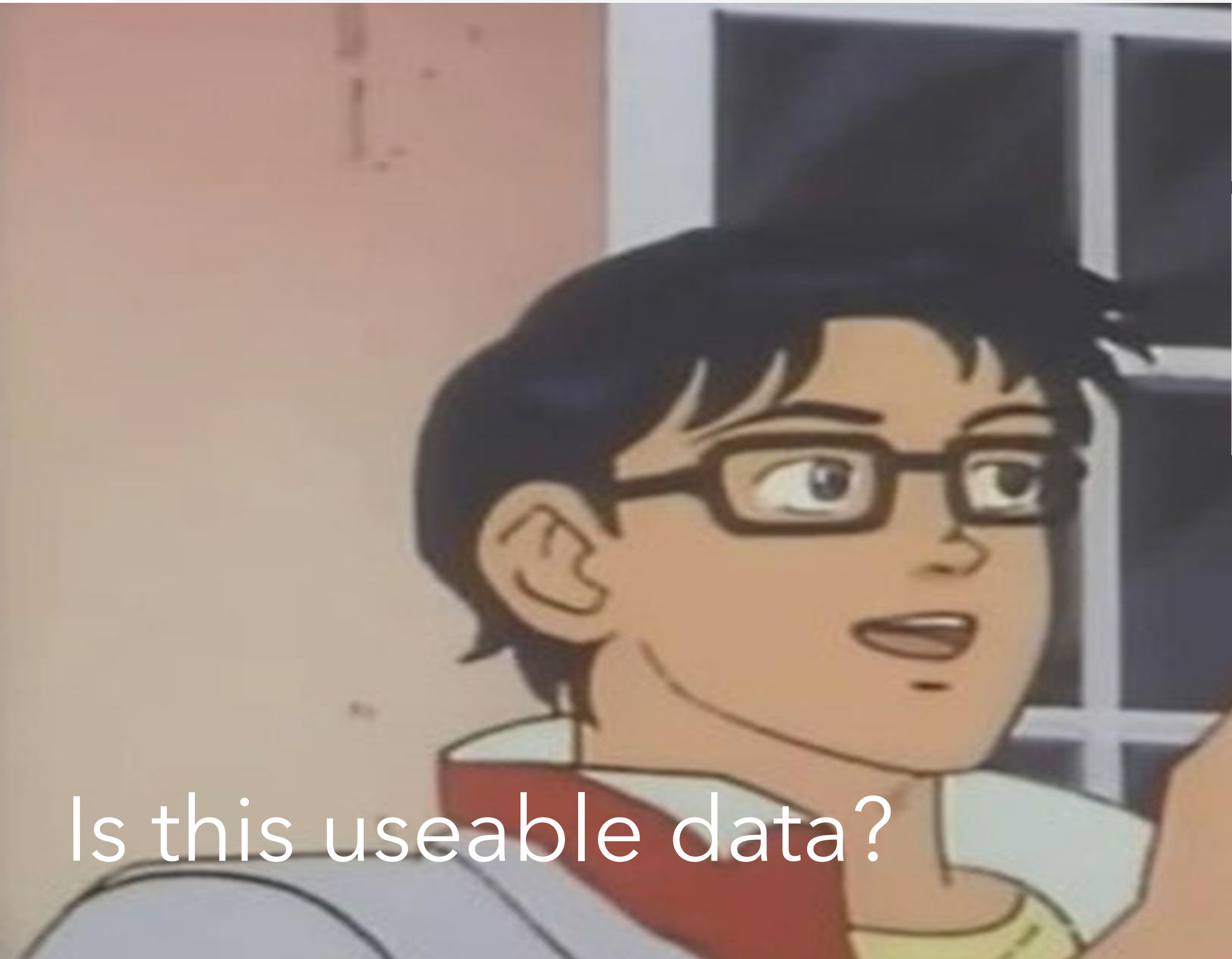
- **id** - unique identified for a house
- **dateDate** - house was sold
- **pricePrice** - is prediction target
- **bedroomsNumber** - of Bedrooms/House
- **bathroomsNumber** - of bathrooms/bedrooms
- **sqft_livingsquare** - footage of the home
- **sqft_lotsquare** - footage of the lot
- **floorsTotal** - floors (levels) in house
- **waterfront** - House which has a view to a waterfront
- **view** - Has been viewed
- **condition** - How good the condition is (Overall)
- **grade** - overall grade given to the housing unit, based on King County grading system
- **sqft_above** - square footage of house apart from basement
- **sqft_basement** - square footage of the basement
- **yr_built** - Built Year
- **yr_renovated** - Year when house was renovated
- **zipcode** - zip
- **lat** - Latitude coordinate
- **long** - Longitude coordinate
- **sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors

SCRUB

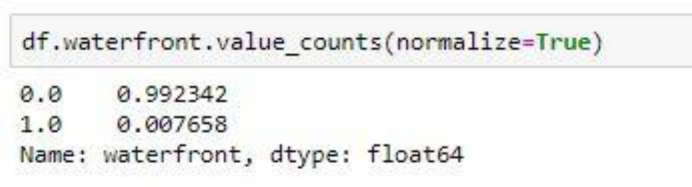
- Cleaning Data means fixing problems with the data that are preventing the information from being interpreted by the user and the machine.



SCRUB

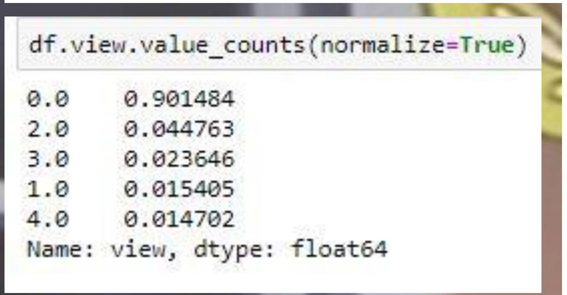


Is this useable data?



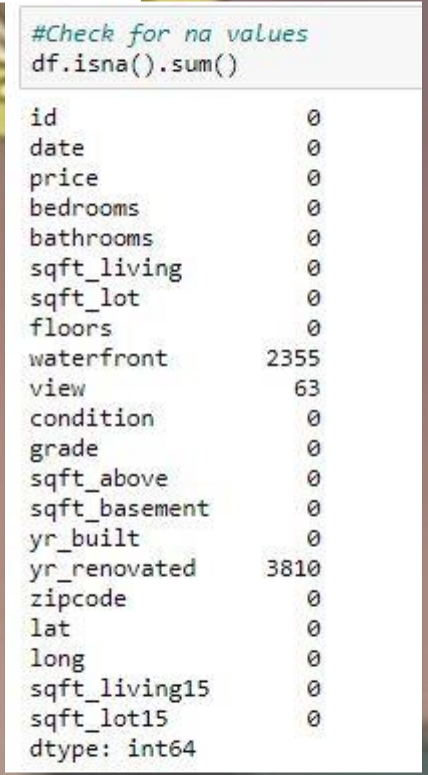
```
df.waterfront.value_counts(normalize=True)

0.0    0.992342
1.0    0.007658
Name: waterfront, dtype: float64
```



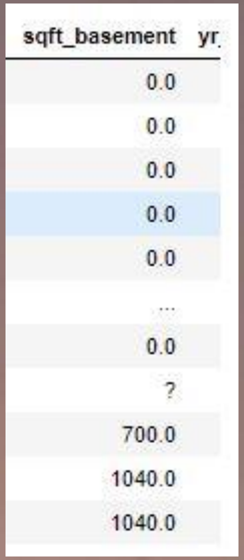
```
df.view.value_counts(normalize=True)

0.0    0.901484
2.0    0.044763
3.0    0.023646
1.0    0.015405
4.0    0.014702
Name: view, dtype: float64
```



```
#Check for na values
df.isna().sum()

id                  0
date                0
price               0
bedrooms            0
bathrooms           0
sqft_living         0
sqft_lot            0
floors              0
waterfront       2355
view               63
condition           0
grade               0
sqft_above          0
sqft_basement       0
yr_built            0
yr_renovated     3810
zipcode             0
lat                 0
long                0
sqft_living15       0
sqft_lot15          0
dtype: int64
```



sqft_basement	yr
0.0	
0.0	
0.0	
0.0	
0.0	
...	
0.0	
?	
700.0	
1040.0	
1040.0	

EXPLORE

ASK MEANINGFUL QUESTIONS ABOUT THE DATA

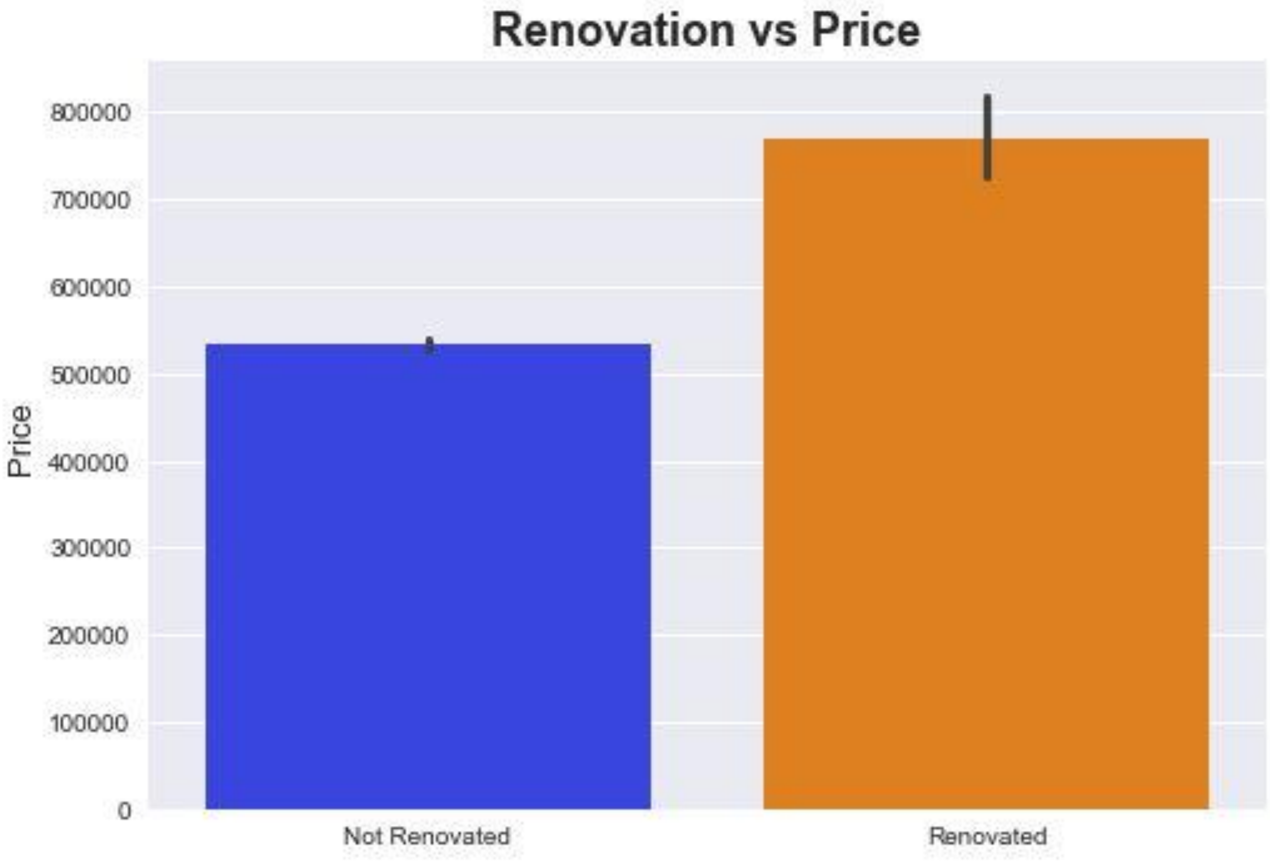
MULTICOLINEARITY

SEE HOW FEATURES RELATE TO PRICE

EXPLORE

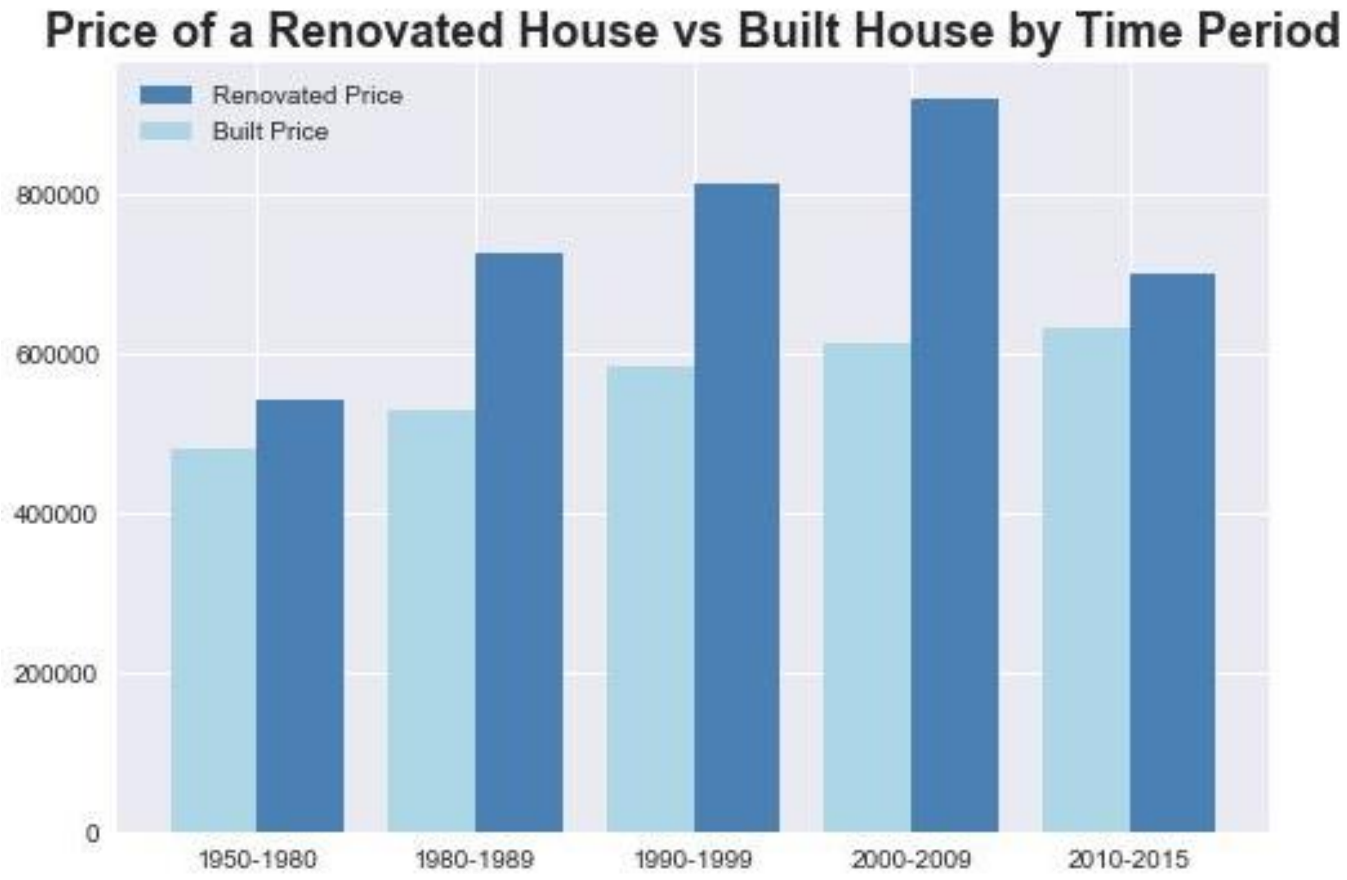
- Question 1: Does renovation have a noticeable effect on price?
- Question 2: Is there a difference in price between a house built in a given time period versus a house renovated in that same time period?
- Question 3: Is there a difference in price based on geographical location in King's County? And if so, why?

Question 1: Does renovation have a noticeable effect on price?



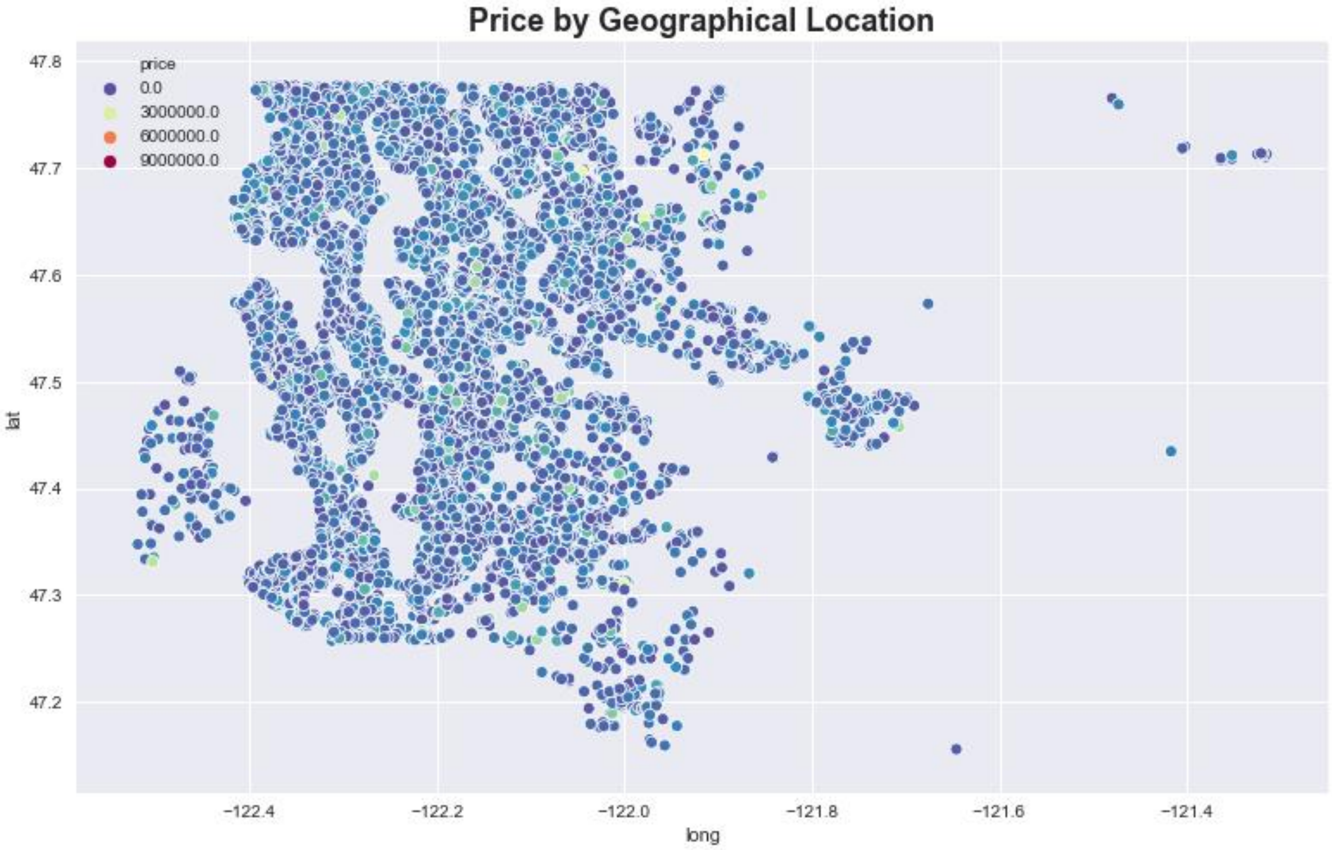
- Renovating a house in King County has a mean price increase of $237,423 or 144%

Question 2: Is there a difference in price between a house built in a given time period versus a house renovated in that same time period?

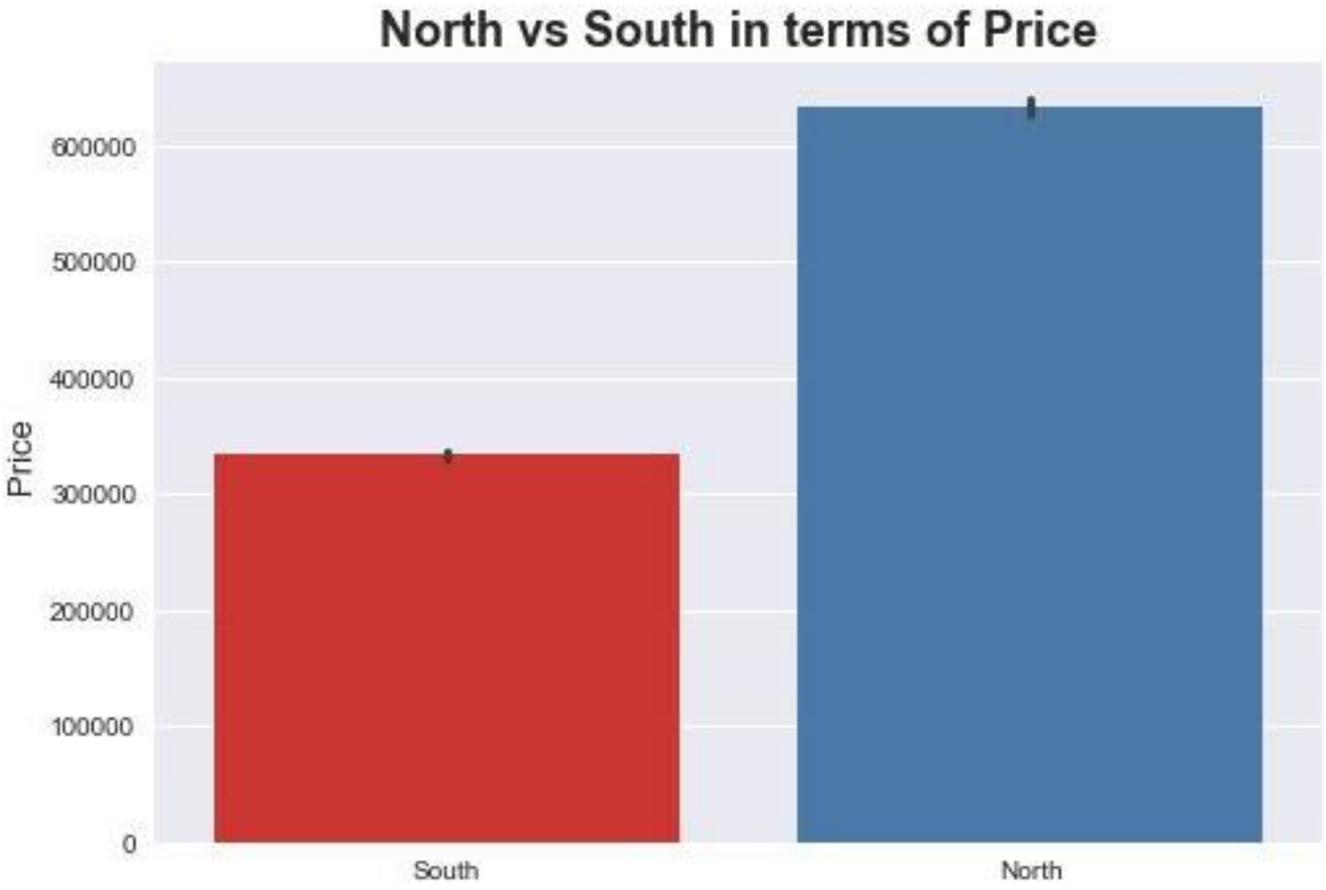


- Significantly larger price for renovated houses
- Gap between renovated and built houses seems to be growing
- 2010-2015 might need more data points to observe this trend

Question 3: Is there a difference in price based on geographical location in King's County?

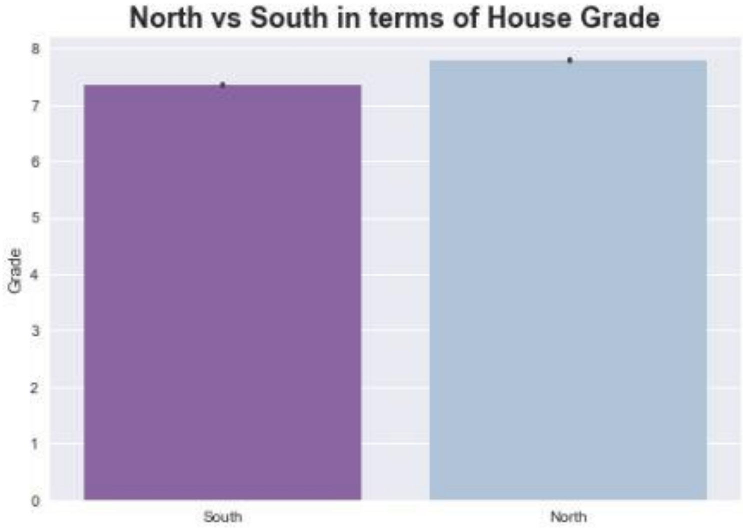


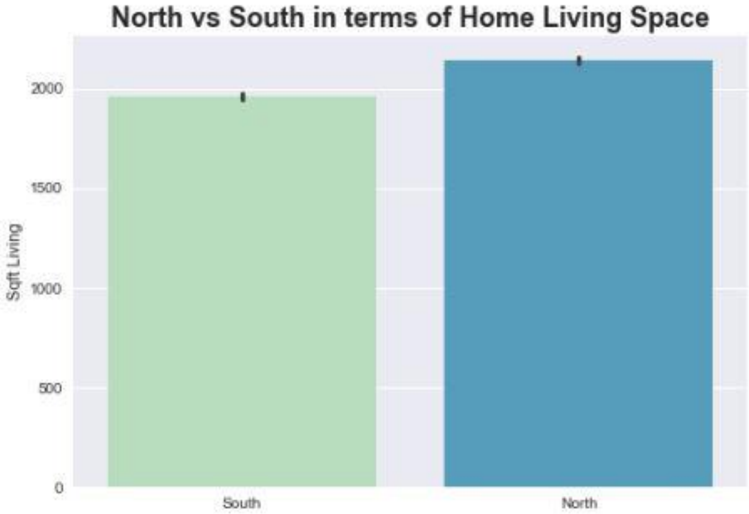
Question 3: Is there a difference in price based on geographical location in King's County?

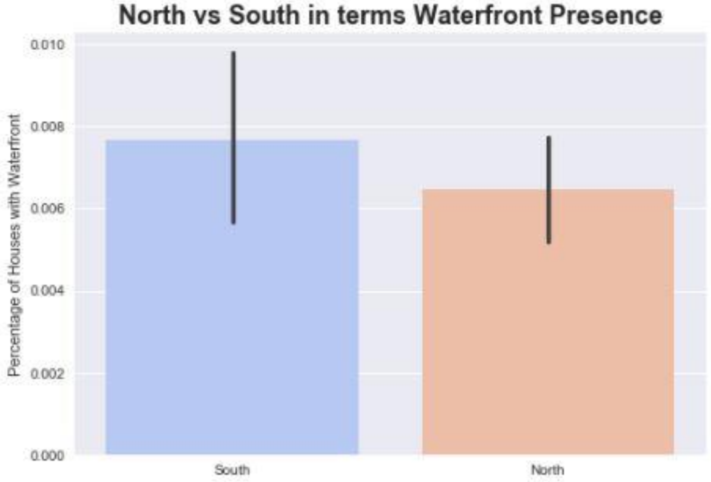


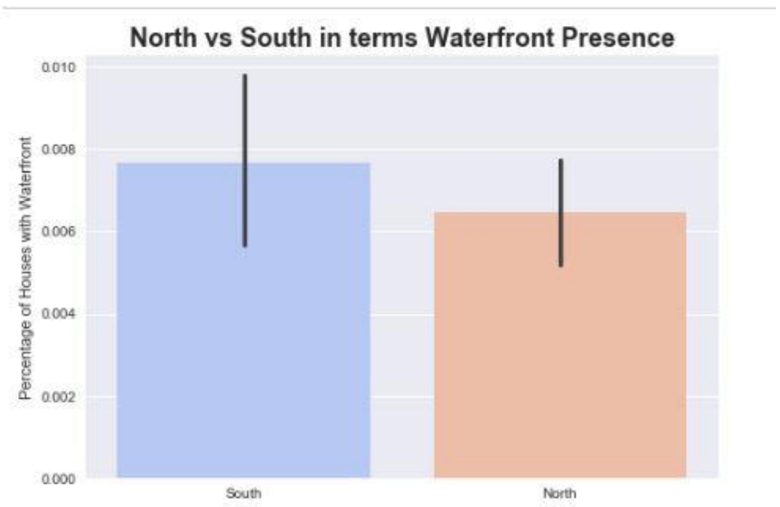
- Houses in the Northern half of Seattle are worth 190% those of in the Southern half – Almost double the price

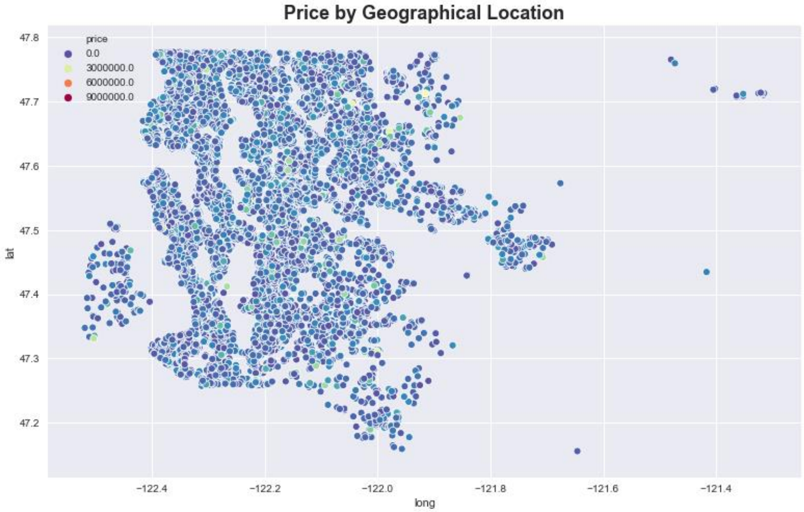
Question 3 (cont.): Why is there a difference in price between Northern and Southern homes in Seattle?

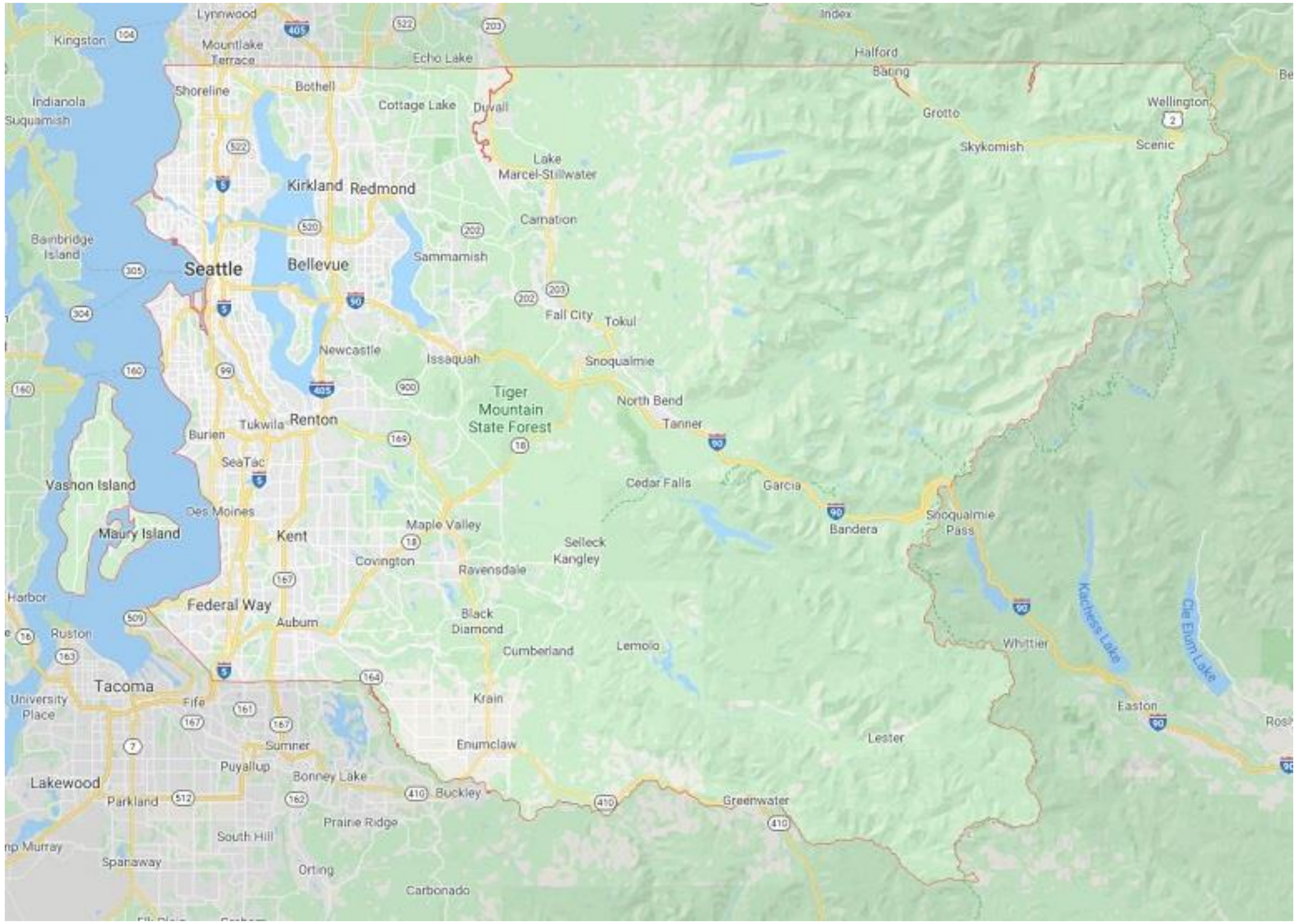




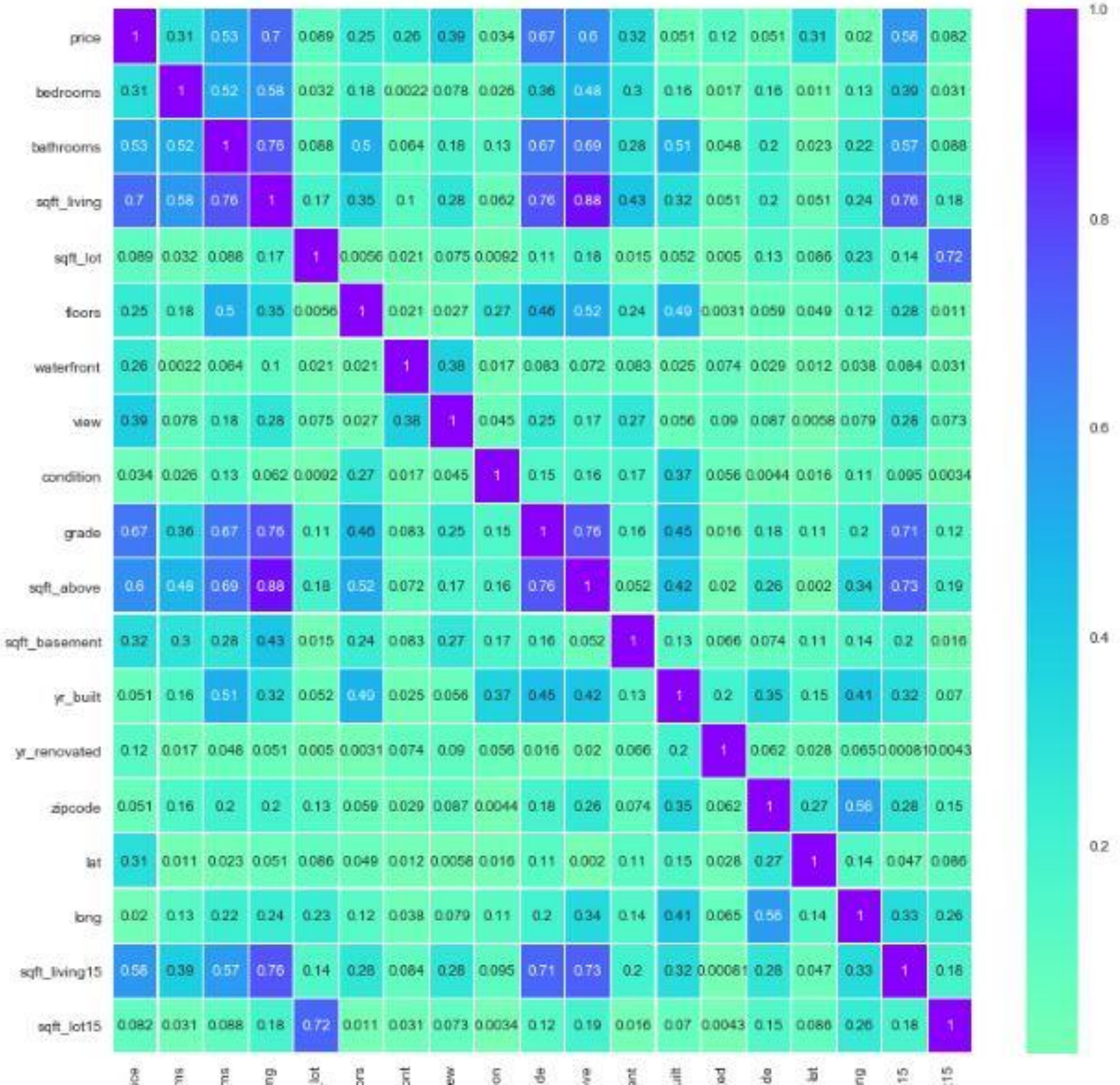



North vs South in terms Waterfront Presence
Percentage of Houses with Waterfront
0.010
0.008
0.006
0.004
0.002
0.000
South
North

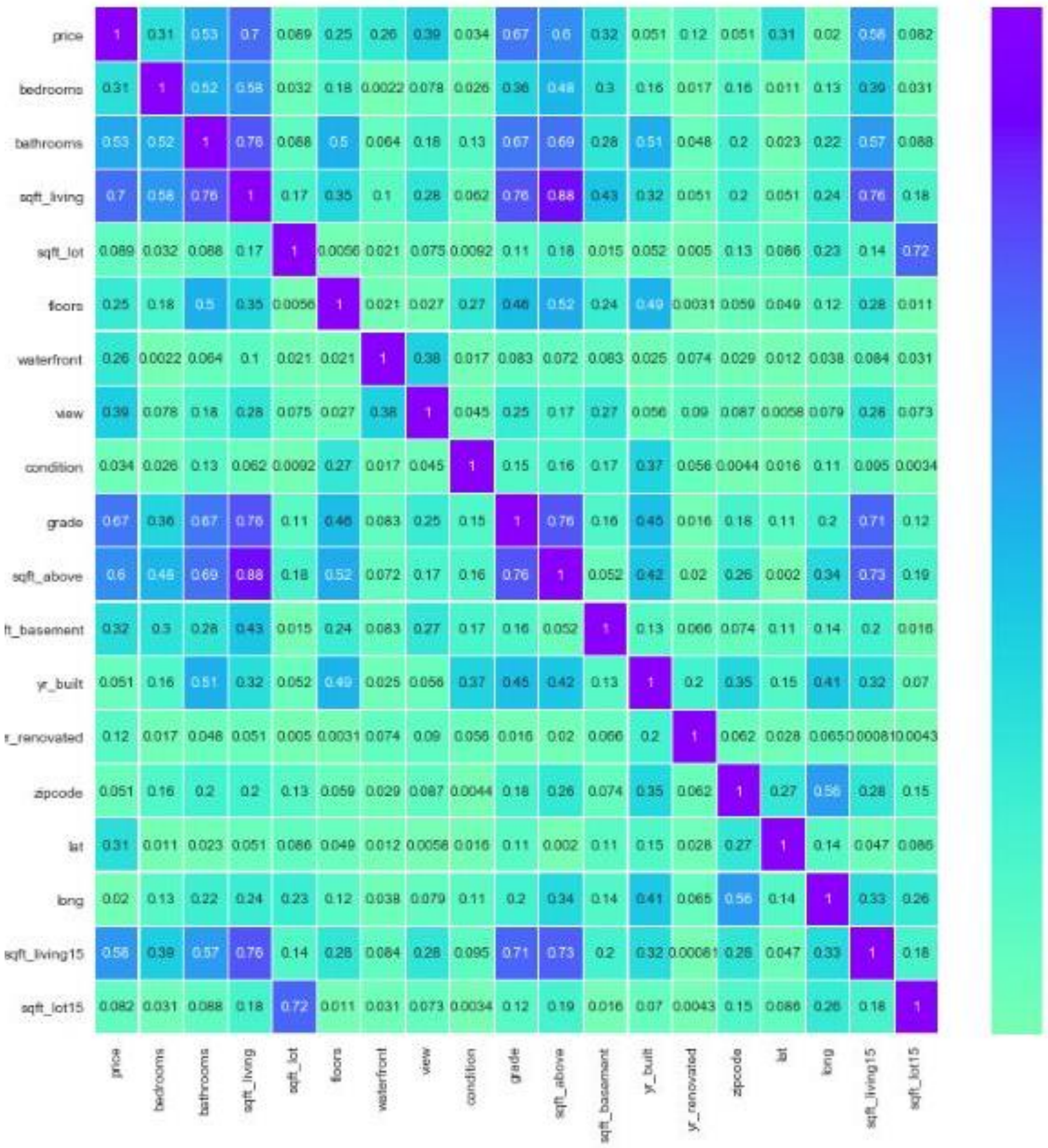

Price by Geographical Location
price
0.0
3000000.0
6000000.0
9000000.0
lat
47.8
47.7
47.6
47.5
47.4
47.3
47.2
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
long


Lynnwood
Kingston
Mountlake Terrace
Echo Lake
Index
Halford
Baring
Shoreline
Bothell
Cottage Lake
Duvall
Indianola
Suquamish
Grotto
Skykomish
Wellington
Scenic
Lake Marcel-Stillwater
Kirkland
Redmond
Carnation
Bainbridge Island
Seattle
Bellevue
Sammamish
Fall City
Tokul
Snoqualmie
Newcastle
Issaquah
Tiger Mountain State Forest
North Bend
Tanner
Burien
Tukwila
Renton
SeaTac
Vashon Island
Cedar Falls
Garcia
Bandera
Snoqualmie Pass
Des Moines
Maury Island
Kent
Maple Valley
Covington
Ravensdale
Selleck
Kangley
Harbor
Federal Way
Auburn
Black Diamond
Ruston
Cumberland
Lemolo
Whittier
Kachess Lake
Cle Elum Lake
Tacoma
Fife
University Place
Krain
Easton
Sumner
Enumclaw
Lester
Lakewood
Puyallup
Bonney Lake
Parkland
Buckley
Greenwater
Prairie Ridge
South Hill
Spanaway
Orting
Carbonado

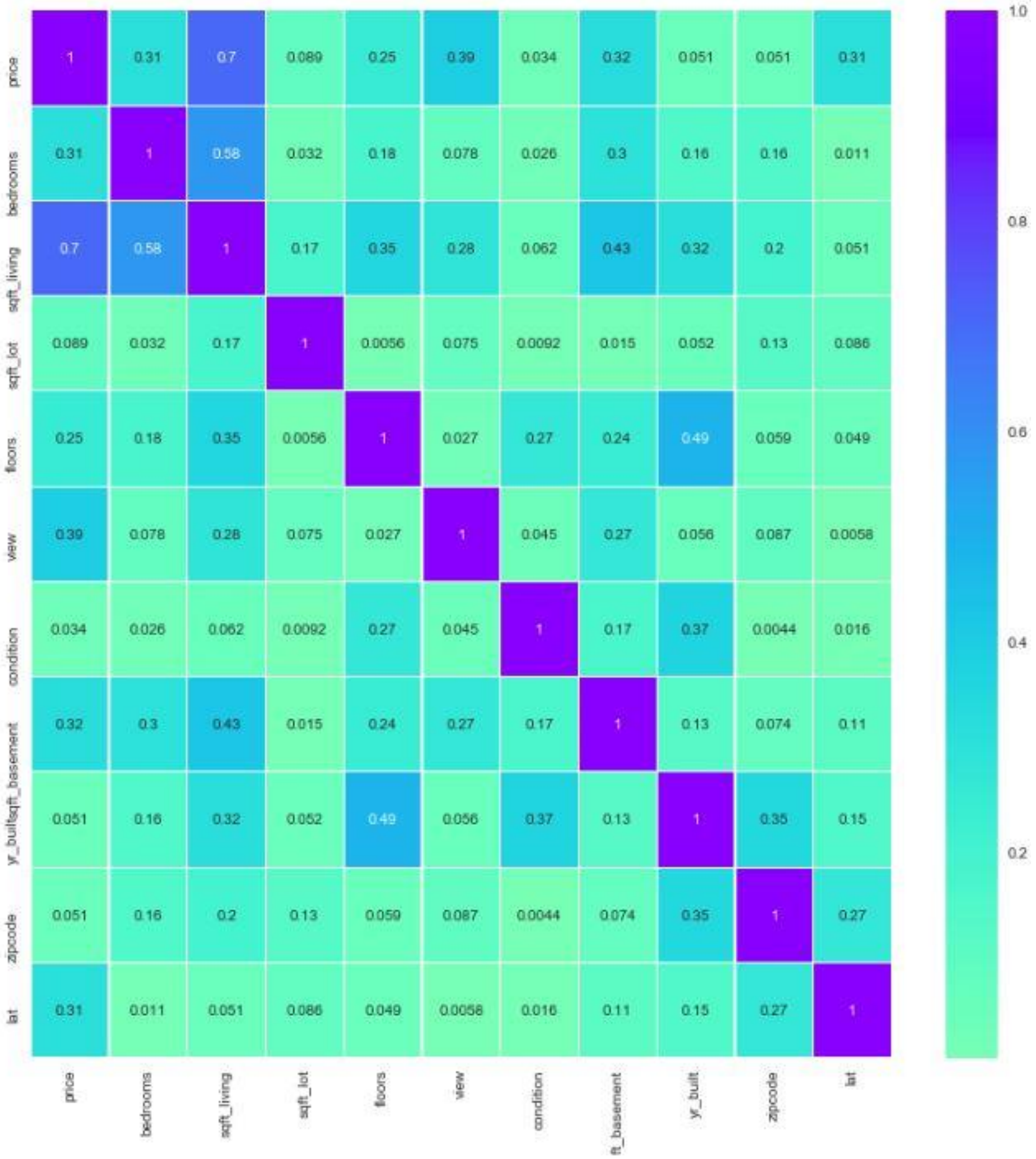
Multicolinearity



	ice	ms	ms	ng	lot	ors	ont	ew	on	de	ve	ent	uilt	ted	de	lat	ng	15	15
price	1	0.31	0.53	0.7	0.089	0.25	0.26	0.39	0.034	0.67	0.6	0.32	0.051	0.12	0.051	0.31	0.02	0.56	0.082
bedrooms	0.31	1	0.52	0.58	0.032	0.18	0.0022	0.078	0.026	0.36	0.48	0.3	0.16	0.017	0.16	0.011	0.13	0.39	0.031
bathrooms	0.53	0.52	1	0.76	0.088	0.5	0.064	0.18	0.13	0.67	0.69	0.28	0.51	0.048	0.2	0.023	0.22	0.57	0.088
sqft_living	0.7	0.58	0.76	1	0.17	0.35	0.1	0.28	0.062	0.76	0.88	0.43	0.32	0.051	0.2	0.051	0.24	0.76	0.18
sqft_lot	0.089	0.032	0.088	0.17	1	0.0056	0.021	0.075	0.0092	0.11	0.18	0.015	0.052	0.005	0.13	0.086	0.23	0.14	0.72
floors	0.25	0.18	0.5	0.35	0.0056	1	0.021	0.027	0.27	0.46	0.52	0.24	0.49	0.0031	0.059	0.049	0.12	0.28	0.011
waterfront	0.26	0.0022	0.064	0.1	0.021	0.021	1	0.38	0.017	0.083	0.072	0.083	0.025	0.074	0.029	0.012	0.038	0.084	0.031
view	0.39	0.078	0.18	0.28	0.075	0.027	0.38	1	0.045	0.25	0.17	0.27	0.056	0.09	0.087	0.0058	0.079	0.28	0.073
condition	0.034	0.026	0.13	0.062	0.0092	0.27	0.017	0.045	1	0.15	0.16	0.17	0.37	0.056	0.0044	0.016	0.11	0.095	0.0034
grade	0.67	0.36	0.67	0.76	0.11	0.46	0.083	0.25	0.15	1	0.76	0.16	0.45	0.016	0.18	0.11	0.2	0.71	0.12
sqft_above	0.6	0.48	0.69	0.88	0.18	0.52	0.072	0.17	0.16	0.76	1	0.052	0.42	0.02	0.26	0.002	0.34	0.73	0.19
sqft_basement	0.32	0.3	0.28	0.43	0.015	0.24	0.083	0.27	0.17	0.16	0.052	1	0.13	0.066	0.074	0.11	0.14	0.2	0.016
yr_built	0.051	0.16	0.51	0.32	0.052	0.49	0.025	0.056	0.37	0.45	0.42	0.13	1	0.2	0.35	0.15	0.41	0.32	0.07
yr_renovated	0.12	0.017	0.048	0.051	0.005	0.0031	0.074	0.09	0.056	0.016	0.02	0.066	0.2	1	0.062	0.028	0.065	0.00081	0.0043
zipcode	0.051	0.16	0.2	0.2	0.13	0.059	0.029	0.087	0.0044	0.18	0.26	0.074	0.35	0.062	1	0.27	0.56	0.28	0.15
lat	0.31	0.011	0.023	0.051	0.086	0.049	0.012	0.0058	0.016	0.11	0.002	0.11	0.15	0.028	0.27	1	0.14	0.047	0.086
long	0.02	0.13	0.22	0.24	0.23	0.12	0.038	0.079	0.11	0.2	0.34	0.14	0.41	0.065	0.56	0.14	1	0.33	0.26
sqft_living15	0.56	0.39	0.57	0.76	0.14	0.28	0.084	0.28	0.095	0.71	0.73	0.2	0.32	0.00081	0.28	0.047	0.33	1	0.18
sqft_lot15	0.082	0.031	0.088	0.18	0.72	0.011	0.031	0.073	0.0034	0.12	0.19	0.016	0.07	0.0043	0.15	0.086	0.26	0.18	1

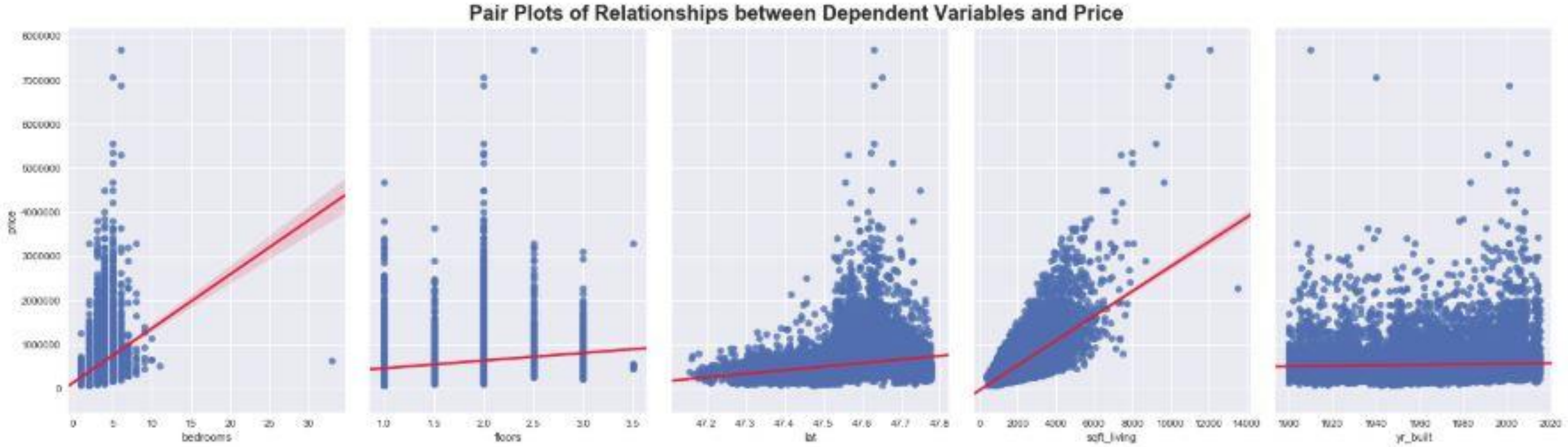


	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
price	1	0.31	0.53	0.7	0.089	0.25	0.26	0.39	0.034	0.67	0.6	0.32	0.051	0.12	0.051	0.31	0.02	0.56	0.082
bedrooms	0.31	1	0.52	0.58	0.032	0.18	0.0022	0.078	0.026	0.36	0.48	0.3	0.16	0.017	0.16	0.011	0.13	0.39	0.031
bathrooms	0.53	0.52	1	0.76	0.088	0.5	0.064	0.18	0.13	0.67	0.69	0.28	0.51	0.048	0.2	0.023	0.22	0.57	0.088
sqft_living	0.7	0.58	0.76	1	0.17	0.35	0.1	0.28	0.062	0.76	0.88	0.43	0.32	0.051	0.2	0.051	0.24	0.76	0.18
sqft_lot	0.089	0.032	0.088	0.17	1	0.0056	0.021	0.075	0.0092	0.11	0.18	0.015	0.052	0.005	0.13	0.086	0.23	0.14	0.72
floors	0.25	0.18	0.5	0.35	0.0056	1	0.021	0.027	0.27	0.46	0.52	0.24	0.49	0.0031	0.059	0.049	0.12	0.28	0.011
waterfront	0.26	0.0022	0.064	0.1	0.021	0.021	1	0.38	0.017	0.083	0.072	0.083	0.025	0.074	0.029	0.012	0.038	0.084	0.031
view	0.39	0.078	0.18	0.28	0.075	0.027	0.38	1	0.045	0.25	0.17	0.27	0.056	0.09	0.087	0.0058	0.079	0.28	0.073
condition	0.034	0.026	0.13	0.062	0.0092	0.27	0.017	0.045	1	0.15	0.16	0.17	0.37	0.056	0.0044	0.016	0.11	0.095	0.0034
grade	0.67	0.36	0.67	0.76	0.11	0.46	0.083	0.25	0.15	1	0.76	0.16	0.45	0.016	0.18	0.11	0.2	0.71	0.12
sqft_above	0.6	0.48	0.69	0.88	0.18	0.52	0.072	0.17	0.16	0.76	1	0.052	0.42	0.02	0.26	0.002	0.34	0.73	0.19
ft_basement	0.32	0.3	0.28	0.43	0.015	0.24	0.083	0.27	0.17	0.16	0.052	1	0.13	0.066	0.074	0.11	0.14	0.2	0.016
yr_built	0.051	0.16	0.51	0.32	0.052	0.49	0.025	0.056	0.37	0.45	0.42	0.13	1	0.2	0.35	0.15	0.41	0.32	0.07
r_renovated	0.12	0.017	0.048	0.051	0.005	0.0031	0.074	0.09	0.056	0.016	0.02	0.066	0.2	1	0.062	0.028	0.065	0.00081	0.0043
zipcode	0.051	0.16	0.2	0.2	0.13	0.059	0.029	0.087	0.0044	0.18	0.26	0.074	0.35	0.062	1	0.27	0.56	0.28	0.15
lat	0.31	0.011	0.023	0.051	0.086	0.049	0.012	0.0058	0.016	0.11	0.002	0.11	0.15	0.028	0.27	1	0.14	0.047	0.086
long	0.02	0.13	0.22	0.24	0.23	0.12	0.038	0.079	0.11	0.2	0.34	0.14	0.41	0.065	0.56	0.14	1	0.33	0.26
sqft_living15	0.56	0.39	0.57	0.76	0.14	0.28	0.084	0.28	0.095	0.71	0.73	0.2	0.32	0.00081	0.28	0.047	0.33	1	0.18
sqft_lot15	0.082	0.031	0.088	0.18	0.72	0.011	0.031	0.073	0.0034	0.12	0.19	0.016	0.07	0.0043	0.15	0.086	0.26	0.18	1



	price	bedrooms	sqft_living	sqft_lot	floors	view	condition	ft_basement	yr_built	zipcode	lat
price	1	0.31	0.7	0.089	0.25	0.39	0.034	0.32	0.051	0.051	0.31
bedrooms	0.31	1	0.58	0.032	0.18	0.078	0.026	0.3	0.16	0.16	0.011
sqft_living	0.7	0.58	1	0.17	0.35	0.28	0.062	0.43	0.32	0.2	0.051
sqft_lot	0.089	0.032	0.17	1	0.0056	0.075	0.0092	0.015	0.052	0.13	0.086
floors	0.25	0.18	0.35	0.0056	1	0.027	0.27	0.24	0.49	0.059	0.049
view	0.39	0.078	0.28	0.075	0.027	1	0.045	0.27	0.056	0.087	0.0058
condition	0.034	0.026	0.062	0.0092	0.27	0.045	1	0.17	0.37	0.0044	0.016
sqft_basement	0.32	0.3	0.43	0.015	0.24	0.27	0.17	1	0.13	0.074	0.11
yr_built	0.051	0.16	0.32	0.052	0.49	0.056	0.37	0.13	1	0.35	0.15
zipcode	0.051	0.16	0.2	0.13	0.059	0.087	0.0044	0.074	0.35	1	0.27
lat	0.31	0.011	0.051	0.086	0.049	0.0058	0.016	0.11	0.15	0.27	1

Relationship between house features and Price



Pair Plots of Relationships between Dependent Variables and Price

MODEL

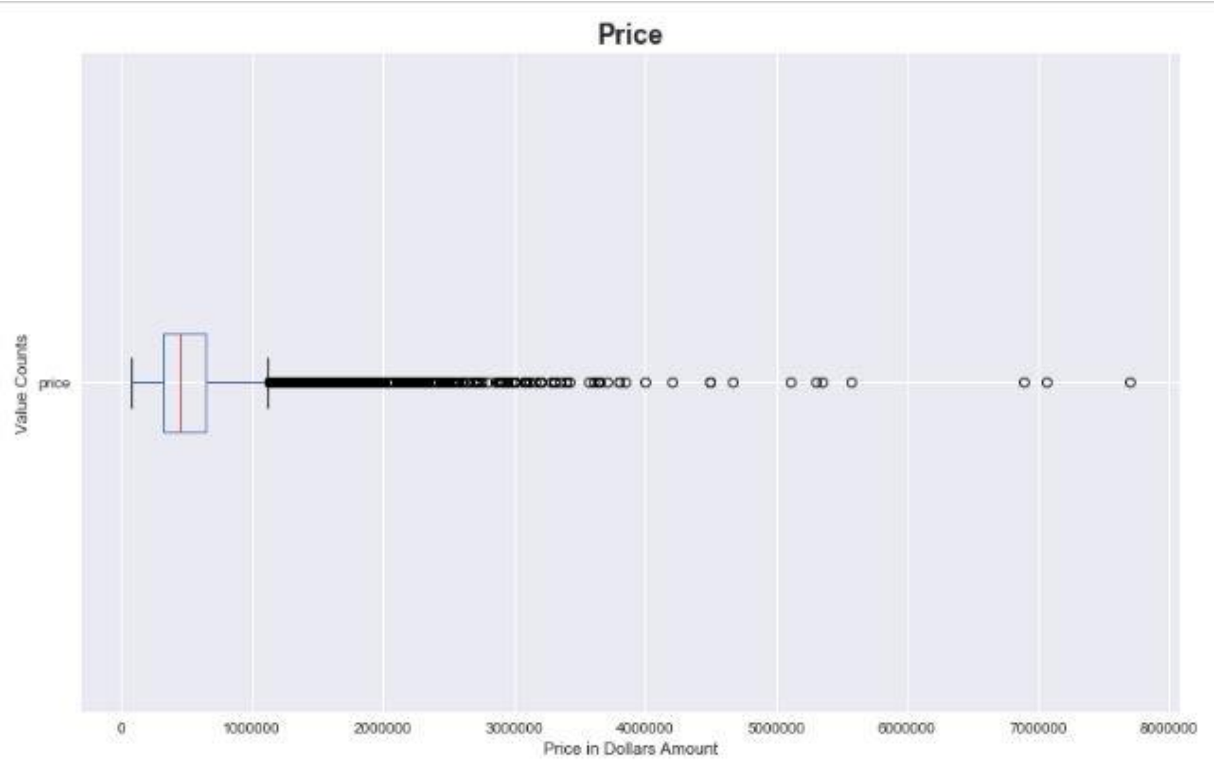
LINEARITY: The relationship between the independent and dependent variables needs to be linear

NORMALITY: The linear regression analysis requires all variables to be normal.

MULTICOLINEARITY: There is little to no multicolinearity in the data.

HOMOSCEDASTICITY: Residuals must be equal across the regression line.

MODEL

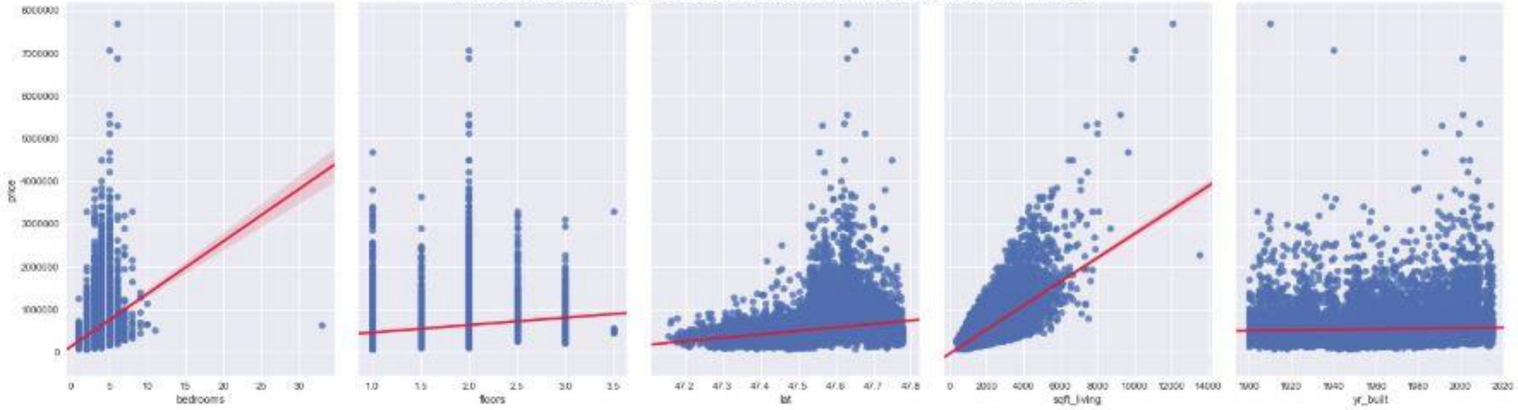


In the linear relationships, there was a skewness due to outliers high in price.

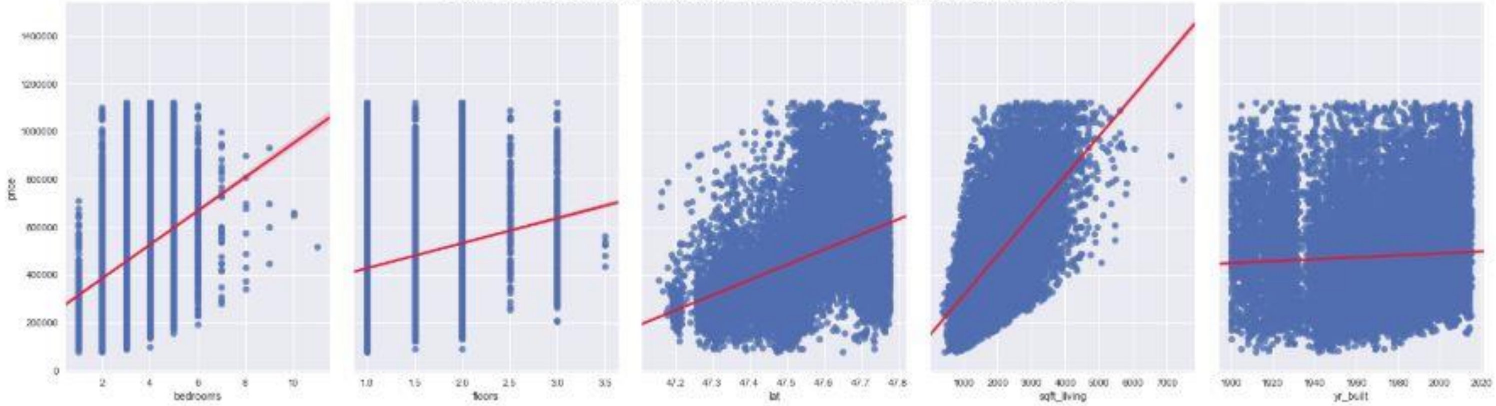
Cut outliers at the upper whisker – $1,120,000

Linear Relationships before and after removal of Price Outliers

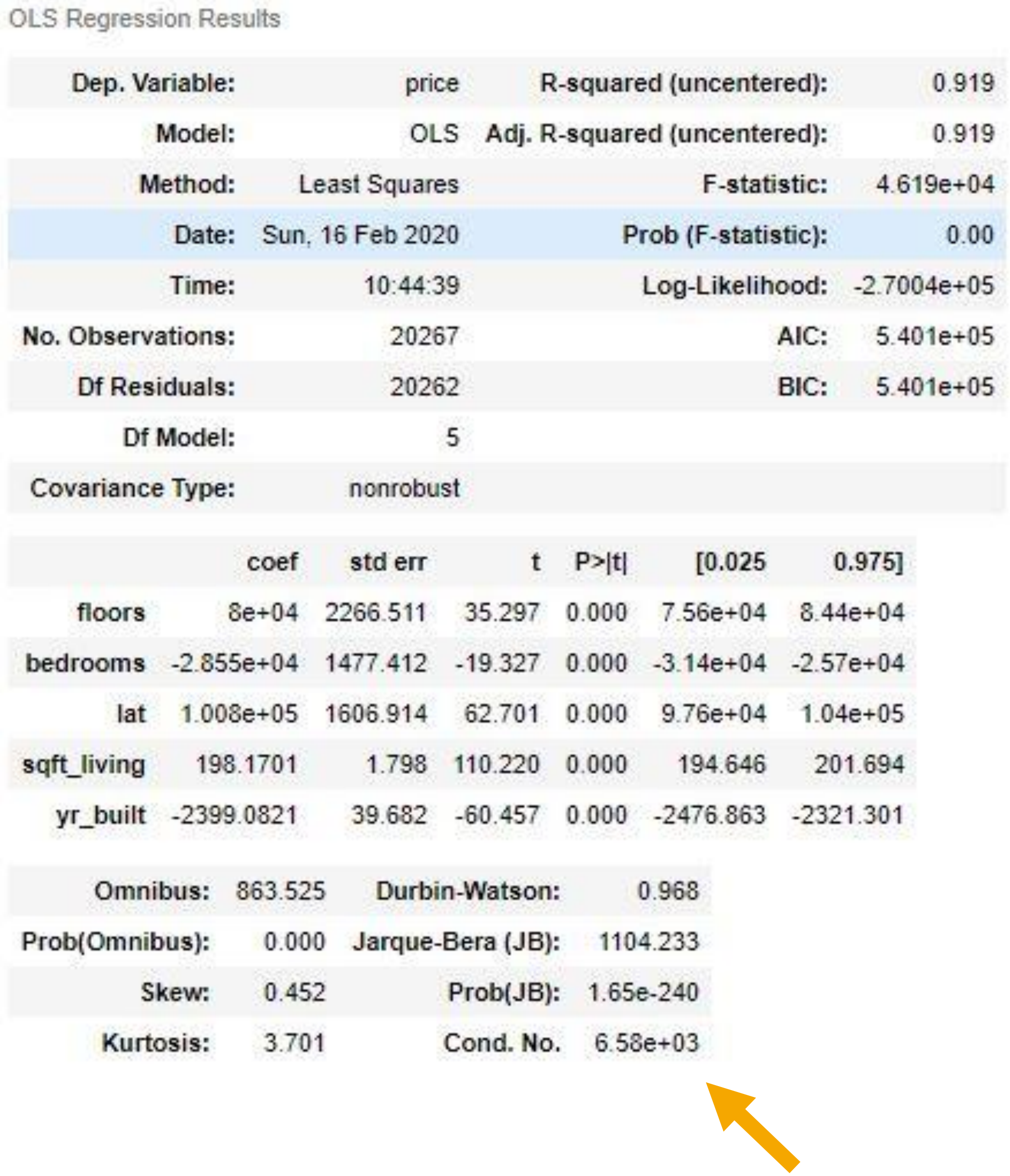
Pair Plots of Relationships between Dependent Variables and Price



Pair Plots of Relationships between Dependent Variables and Price



Multicolinearity: Round 2



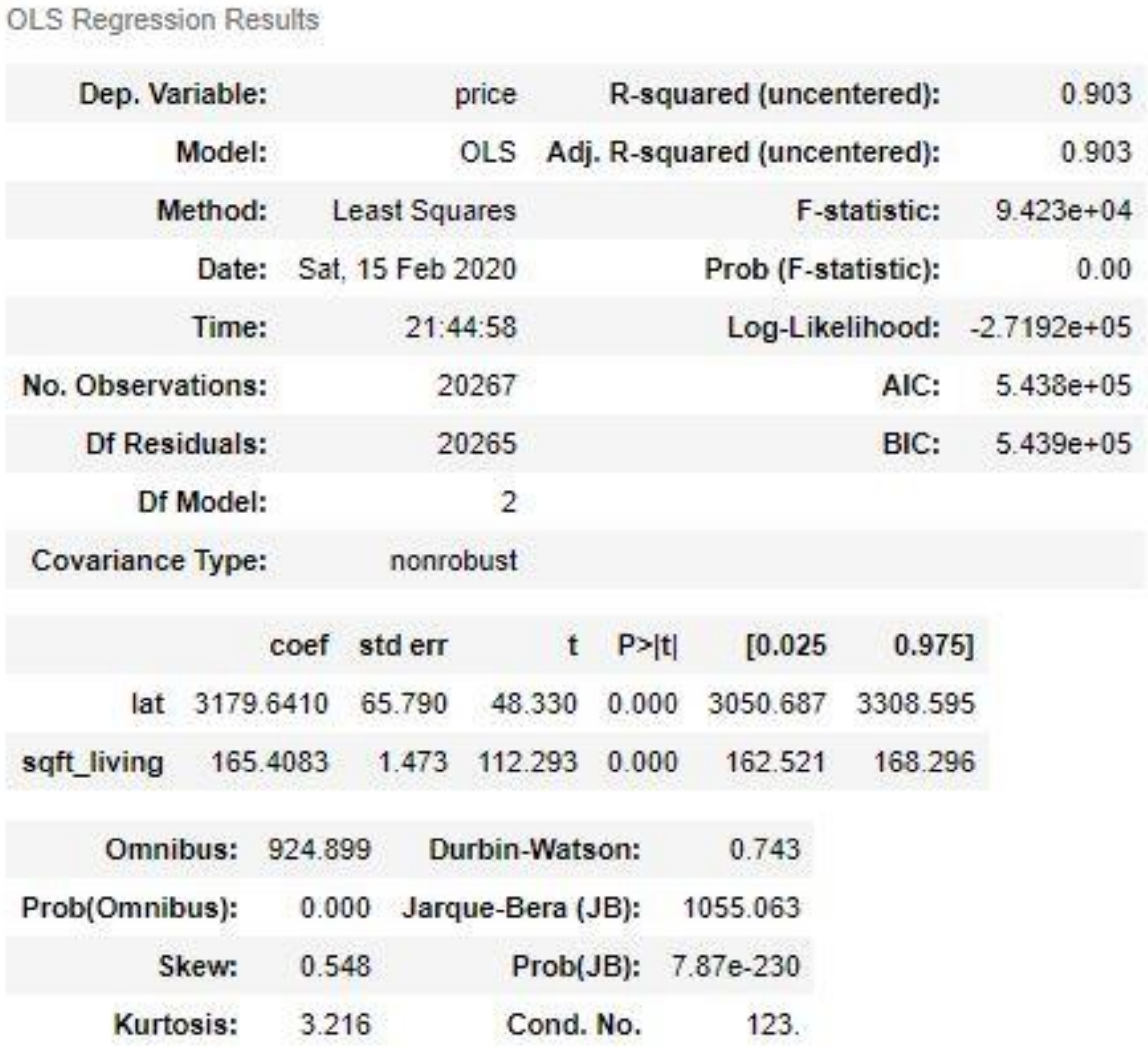
OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.919
Model:	OLS	Adj. R-squared (uncentered):	0.919
Method:	Least Squares	F-statistic:	4.619e+04
Date:	Sun, 16 Feb 2020	Prob (F-statistic):	0.00
Time:	10:44:39	Log-Likelihood:	-2.7004e+05
No. Observations:	20267	AIC:	5.401e+05
Df Residuals:	20262	BIC:	5.401e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
floors	8e+04	2266.511	35.297	0.000	7.56e+04	8.44e+04
bedrooms	-2.855e+04	1477.412	-19.327	0.000	-3.14e+04	-2.57e+04
lat	1.008e+05	1606.914	62.701	0.000	9.76e+04	1.04e+05
sqft_living	198.1701	1.798	110.220	0.000	194.646	201.694
yr_built	-2399.0821	39.682	-60.457	0.000	-2476.863	-2321.301

Omnibus:	863.525	Durbin-Watson:	0.968
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1104.233
Skew:	0.452	Prob(JB):	1.65e-240
Kurtosis:	3.701	Cond. No.	6.58e+03

Multicolinearity: Round 2



OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.903
Model:	OLS	**Adj. R-squared (uncentered):**	0.903
Method:	Least Squares	**F-statistic:**	9.423e+04
Date:	Sat, 15 Feb 2020	**Prob (F-statistic):**	0.00
Time:	21:44:58	**Log-Likelihood:**	-2.7192e+05
No. Observations:	20267	**AIC:**	5.438e+05
Df Residuals:	20265	**BIC:**	5.439e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
lat	3179.6410	65.790	48.330	0.000	3050.687	3308.595
sqft_living	165.4083	1.473	112.293	0.000	162.521	168.296

Omnibus:	924.899	**Durbin-Watson:**	0.743
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1055.063
Skew:	0.548	**Prob(JB):**	7.87e-230
Kurtosis:	3.216	**Cond. No.**	123.

MODEL

LINEARITY: THE RELATIONSHIP BETWEEN THE INDEPENDENT AND DEPENDENT VARIABLES NEEDS TO BE LINEAR

NORMALITY: THE LINEAR REGRESSION ANALYSIS REQUIRES ALL VARIABLES TO BE NORMAL.

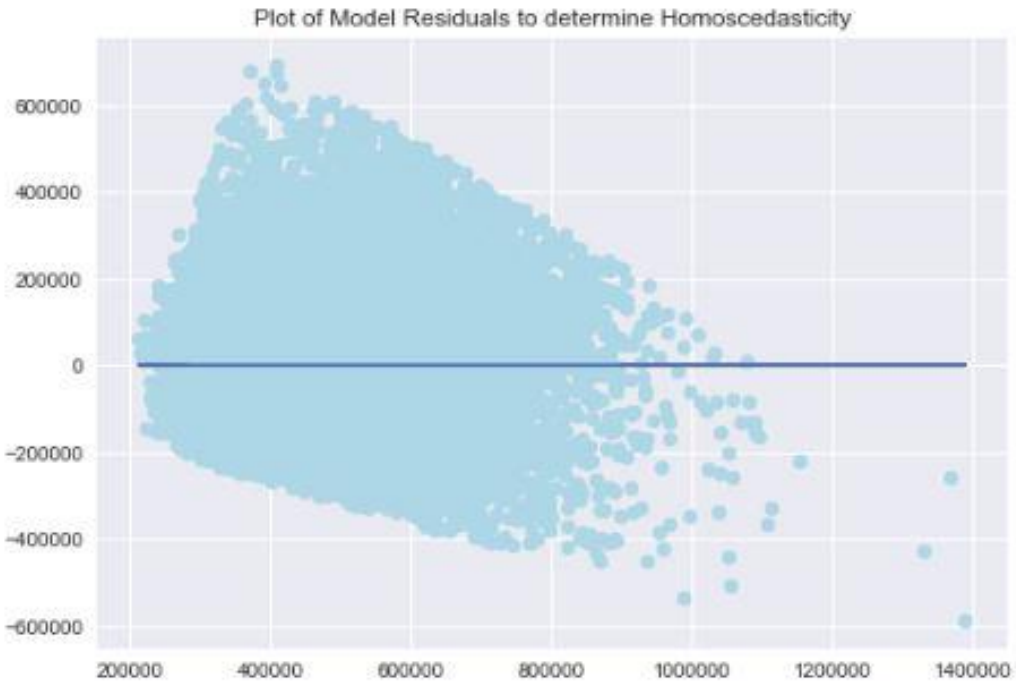
MULTICOLINEARITY: THERE IS LITTLE TO NO MULTICOLINEARITY IN THE DATA.

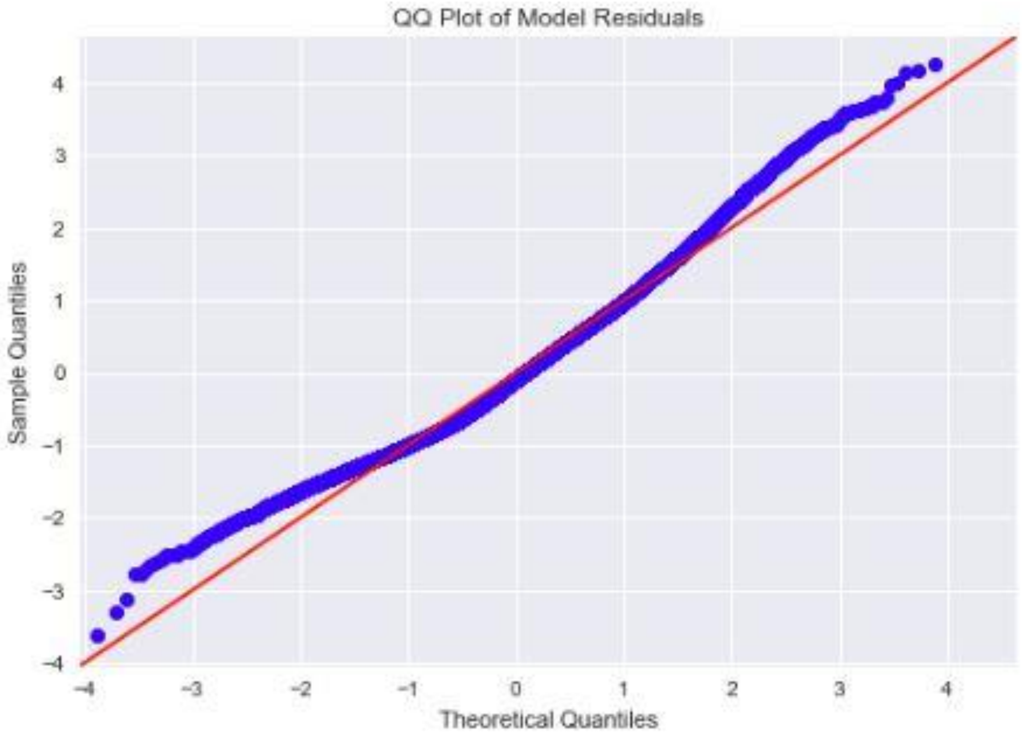
HOMOSCEDASTICITY: RESIDUALS MUST BE EQUAL ACROSS THE REGRESSION LINE.

Assumptions that can only be tested after the model has been created

Homoscedasticity

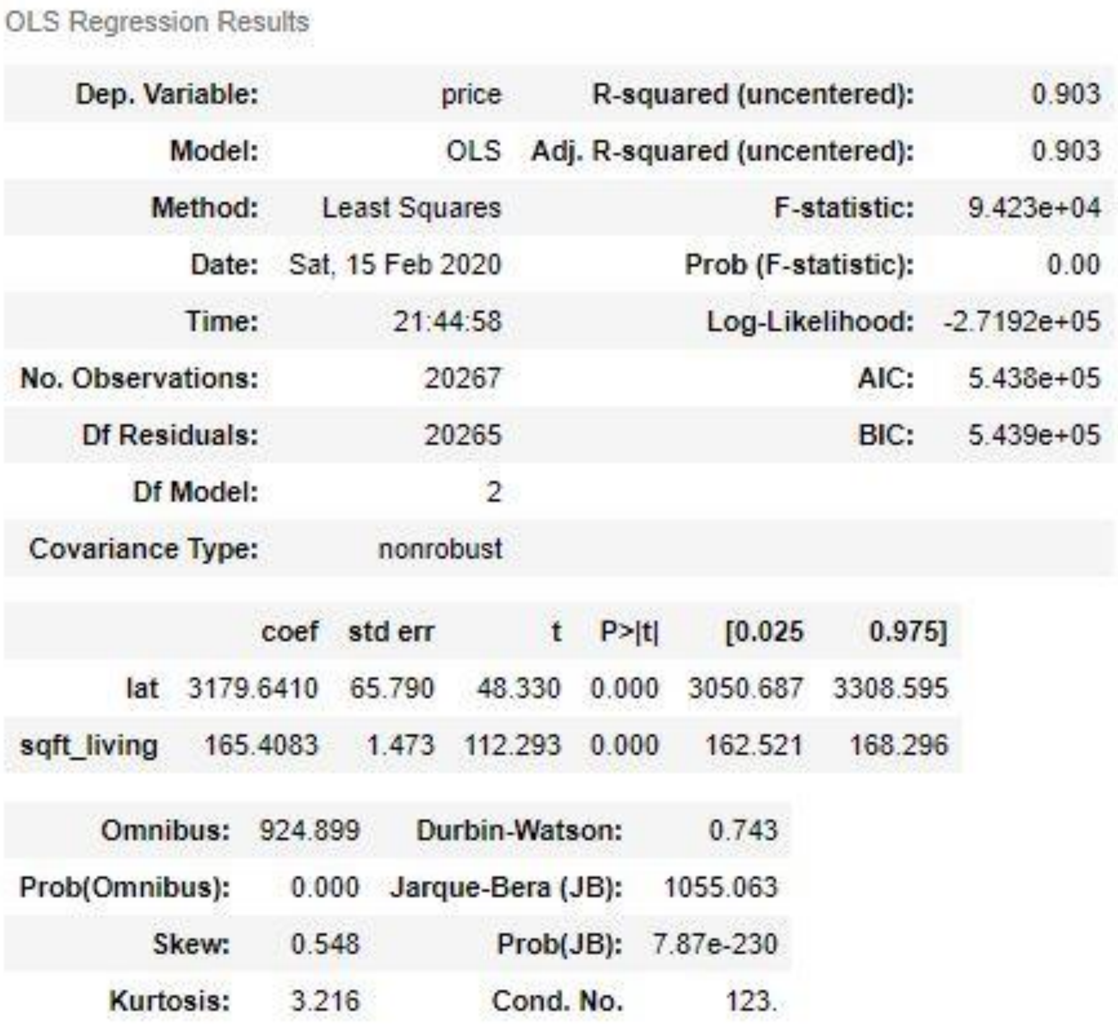
Normality





INTERPRET

- **Scores:**
- R-squared: The proportion of the variance in the dependent variable that is predictable from the indenpendent variables. Approximately 90% of the variance in the price of a house can be predicted from the model.
- F-Statistic: The F-statistic is the ratio of the mean regression sum of squares divided by the mean error of sum squares. With a probability of 0, the null hypothesis that the fit of the intercept only model and my model are equal is rejected.
- Coefficients: The coefficients show the relationship between the independent variable and the dependent variable.
- p-Values: The p-values are all 0, meaning the indepedent variables are statistically signficant to the model.
- Skew: The skewness is positive which implies a skewness towards the right. With a value between -0.5 and 0.5, the data is pretty symmetrical.
- Kurtosis: The Kurtosis score is about 3, implying that the data fits a normal distribution.
- Jarque-Bera: The Jarque-Bera test is a goodness-of-fit test. Because the Jarque-Bera score is large, this indicates that errors are not normally distributed.
- Condition Number: Measures multicollinearity, the number is large which implies multicollinearity.



OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.903
Model:	OLS	Adj. R-squared (uncentered):	0.903
Method:	Least Squares	F-statistic:	9.423e+04
Date:	Sat, 15 Feb 2020	Prob (F-statistic):	0.00
Time:	21:44:58	Log-Likelihood:	-2.7192e+05
No. Observations:	20267	AIC:	5.438e+05
Df Residuals:	20265	BIC:	5.439e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
lat	3179.6410	65.790	48.330	0.000	3050.687	3308.595
sqft_living	165.4083	1.473	112.293	0.000	162.521	168.296

Omnibus:	924.899	Durbin-Watson:	0.743
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1055.063
Skew:	0.548	Prob(JB):	7.87e-230
Kurtosis:	3.216	Cond. No.	123.

Questions

John Kundycki

Predicting Real Estate Prices

King County Housing Dataset