KINGS COUNTY HOUSING DATA

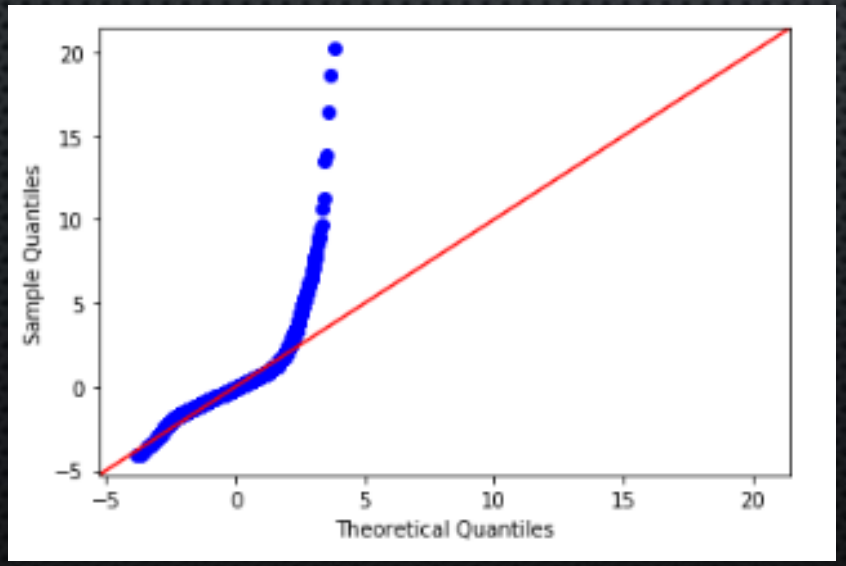
ISABEL JOSEPH

Data Cleaning

- Using the kings county housing data set, I completed a data analysis in order to make predictions about housing price.
- Initially I started off by looking through the data set, and completed a thorough clean, removing errors in the data and dealing with nan values. This initial clean brought down our housing entries from 21597 to 21419.

BASELINE MODEL

- I INITIALLY CREATED A BASELINE MODEL WITH ALL THE POSSIBLE FEATURES.
- TO CHECK THE NORMALITY ASSUMPTION I RAN A Q-Q PLOT



Dep. Variable:	price	R-squared:	0.645
Model:	OLS	Adj. R-squared:	0.645
Method:	Least Squares	F-statistic:	2395.
Date:	Fri, 27 Mar 2020	Prob (F-statistic):	0.00
Time:	10:59:33	Log-Likelihood:	-2.3520e+05
No. Observations:	17135	AIC:	4.704e+05
Df Residuals:	17121	BIC:	4.705e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.573e+06	1.52e+05	43.385	0.000	6.28e+06	6.87e+06
id	-2.031e-06	5.95e-07	-3.411	0.001	-3.2e-06	-8.64e-07
bedrooms	-4.73e+04	2436.832	-19.409	0.000	-5.21e+04	-4.25e+04
bathrooms	5.084e+04	3948.410	12.876	0.000	4.31e+04	5.86e+04
sqft_living	181.3219	3.770	48.097	0.000	173.933	188.711
sqft_lot	-0.2875	0.044	-6.483	0.000	-0.374	-0.201
floors	1.809e+04	3936.252	4.596	0.000	1.04e+04	2.58e+04
waterfront	7.23e+05	2.07e+04	34.964	0.000	6.82e+05	7.63e+05
condition	2.235e+05	1.39e+04	16.049	0.000	1.96e+05	2.51e+05
grade	1.321e+05	2447.544	53.962	0.000	1.27e+05	1.37e+05
yr_built	-3870.8568	76.587	-50.542	0.000	-4020.975	-3720.738
cond_2	-2.062e+05	3.84e+04	-5.366	0.000	-2.81e+05	-1.31e+05
cond_3	-4.344e+05	2.18e+04	-19.929	0.000	-4.77e+05	-3.92e+05
cond_4	-6.456e+05	1.56e+04	-41.501	0.000	-6.76e+05	-6.15e+05
cond_5	-8.344e+05	2.03e+04	-41.044	0.000	-8.74e+05	-7.95e+05

Omnibus:	13385.528	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1006968.380
Skew:	3.187	Prob(JB):	0.00
Kurtosis:	40.010	Cond. No.	2.76e+25

REFINED MODEL

- I THEN REFINED THE MODEL, TO INCLUDE ONLY THE INFLUENTIAL PREDICTOR VARIABLES.
- I REMOVED ALL OF THE NEGATIVE COEFFICIENTS IN THE INITIAL MODEL, AND I ALSO RAN A VARIANCE INFLATION FACTOR TEST, TO TEST FOR MULTICOLLINEARITY BETWEEN THE VARIABLES. I RECEIVED VIFS OF OVER 10 WITH CONDITION AND ALL THE PREDICTORS THAT WERE DUMMY VARIABLES OF CONDITION.

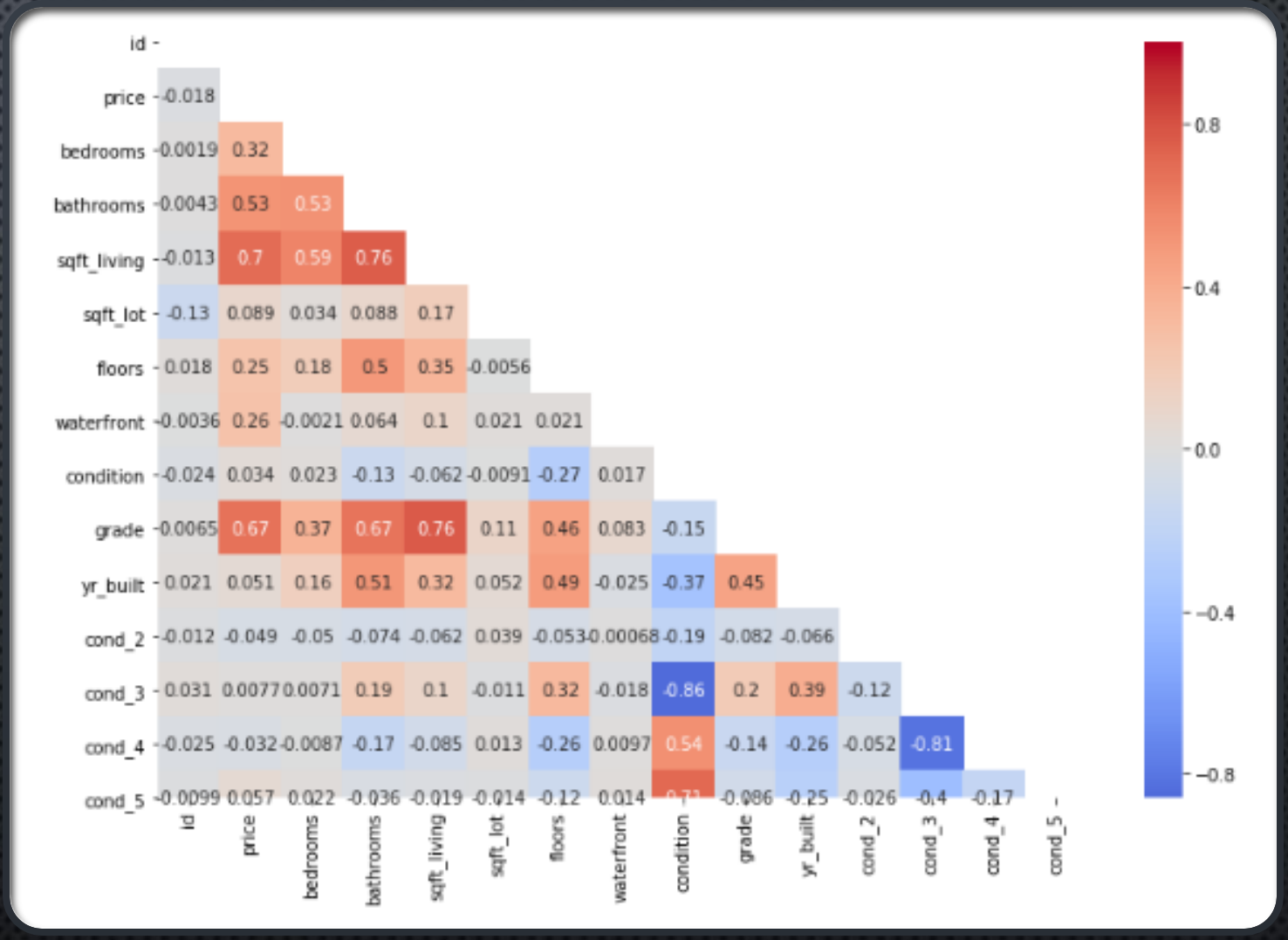
Dep. Variable:	price	R-squared:	0.584
Model:	OLS	Adj. R-squared:	0.584
Method:	Least Squares	F-statistic:	2676.
Date:	Fri, 27 Mar 2020	Prob (F-statistic):	0.00
Time:	10:59:48	Log-Likelihood:	-2.3656e+05
No. Observations:	17135	AIC:	4.731e+05
Df Residuals:	17125	BIC:	4.732e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.728e+05	6.49e+04	-8.820	0.000	-7e+05	-4.46e+05
bathrooms	-2.435e+04	3970.317	-6.133	0.000	-3.21e+04	-1.66e+04
sqft_living	182.9532	3.667	49.893	0.000	175.768	190.141
floors	-1.975e+04	4155.549	-4.754	0.000	-2.79e+04	-1.16e+04
waterfront	8.133e+05	2.23e+04	36.532	0.000	7.7e+05	8.57e+05
condition	-2.557e+04	1.41e+04	-1.807	0.071	-5.33e+04	2161.350
grade	1.183e+05	2559.436	46.204	0.000	1.13e+05	1.23e+05
cond_2	-2.617e+04	4.14e+04	-0.632	0.527	-1.07e+05	5.49e+04
cond_3	-4.431e+04	2.21e+04	-2.004	0.045	-8.76e+04	-979.129
cond_4	3.231e+04	8864.624	3.645	0.000	1.49e+04	4.97e+04
cond_5	1.413e+05	7983.332	17.696	0.000	1.26e+05	1.57e+05

Omnibus:	13377.675	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	932723.914
Skew:	3.211	Prob(JB):	0.00
Kurtosis:	38.569	Cond. No.	3.00e+19

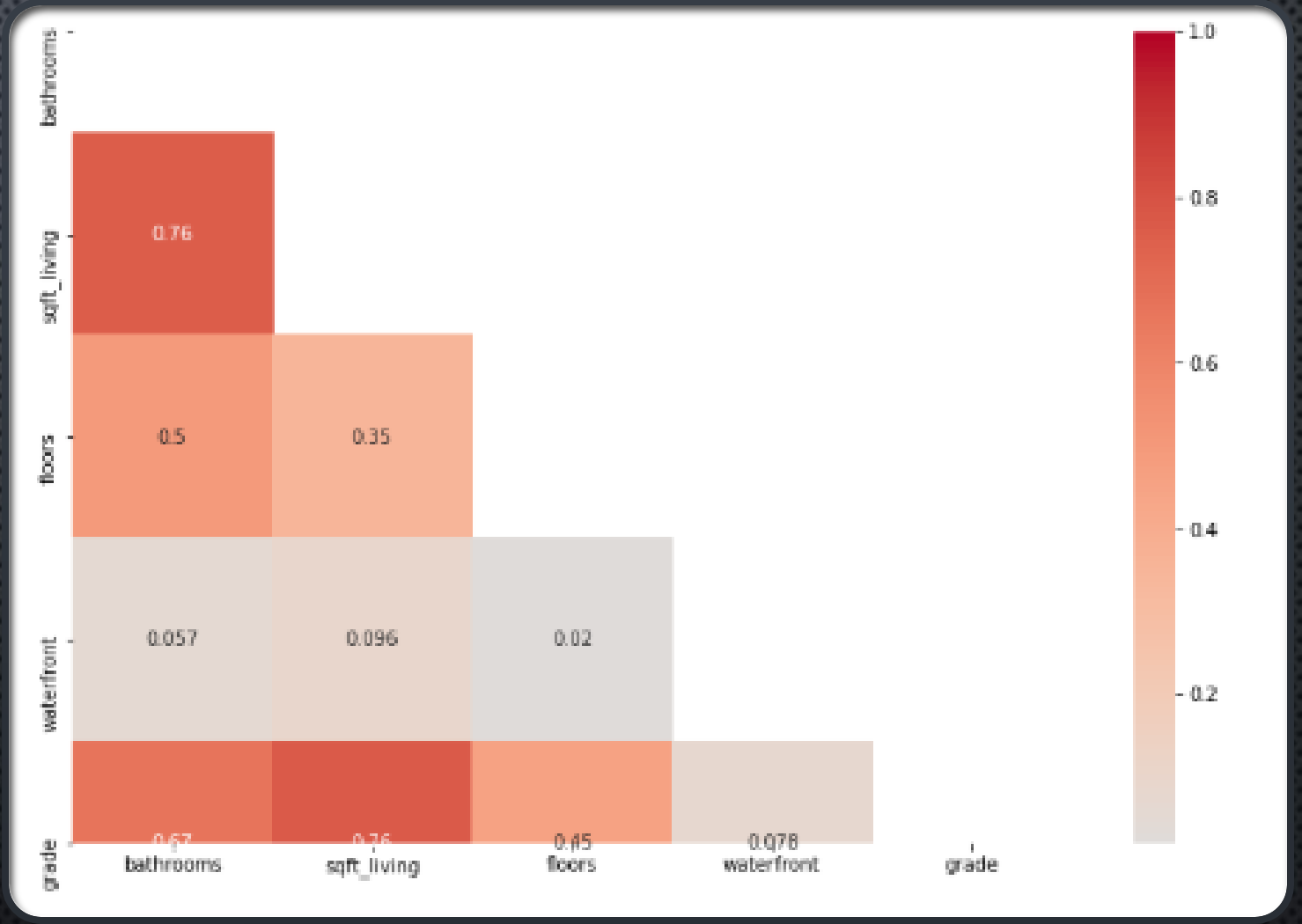
WHAT FEATURES HAVE AN EFFECT ON PRICE?

- Our target variable is price, so according to this correlation heatmap, I found that the highest correlating variable to price is sqft_living (0.7), followed by grade (0.67) and bathrooms (0.53).

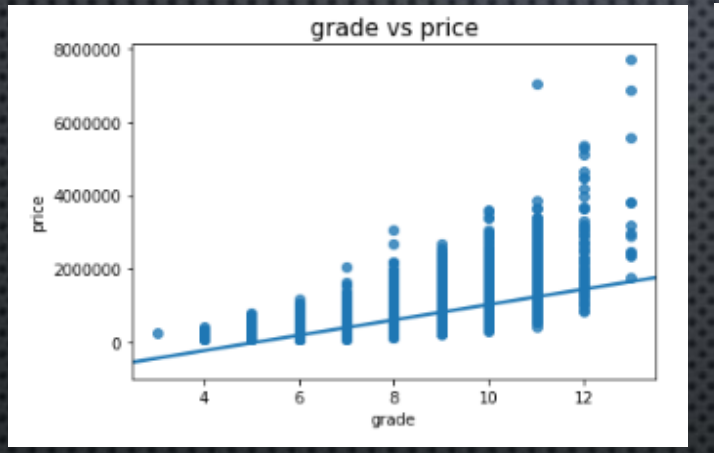


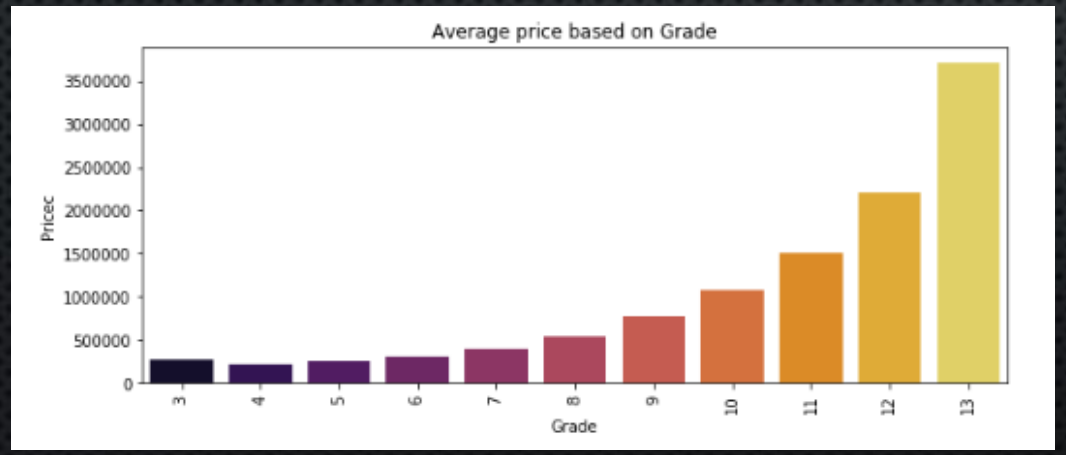
IS THERE MULTICOLLINEARITY BETWEEN FEATURES?

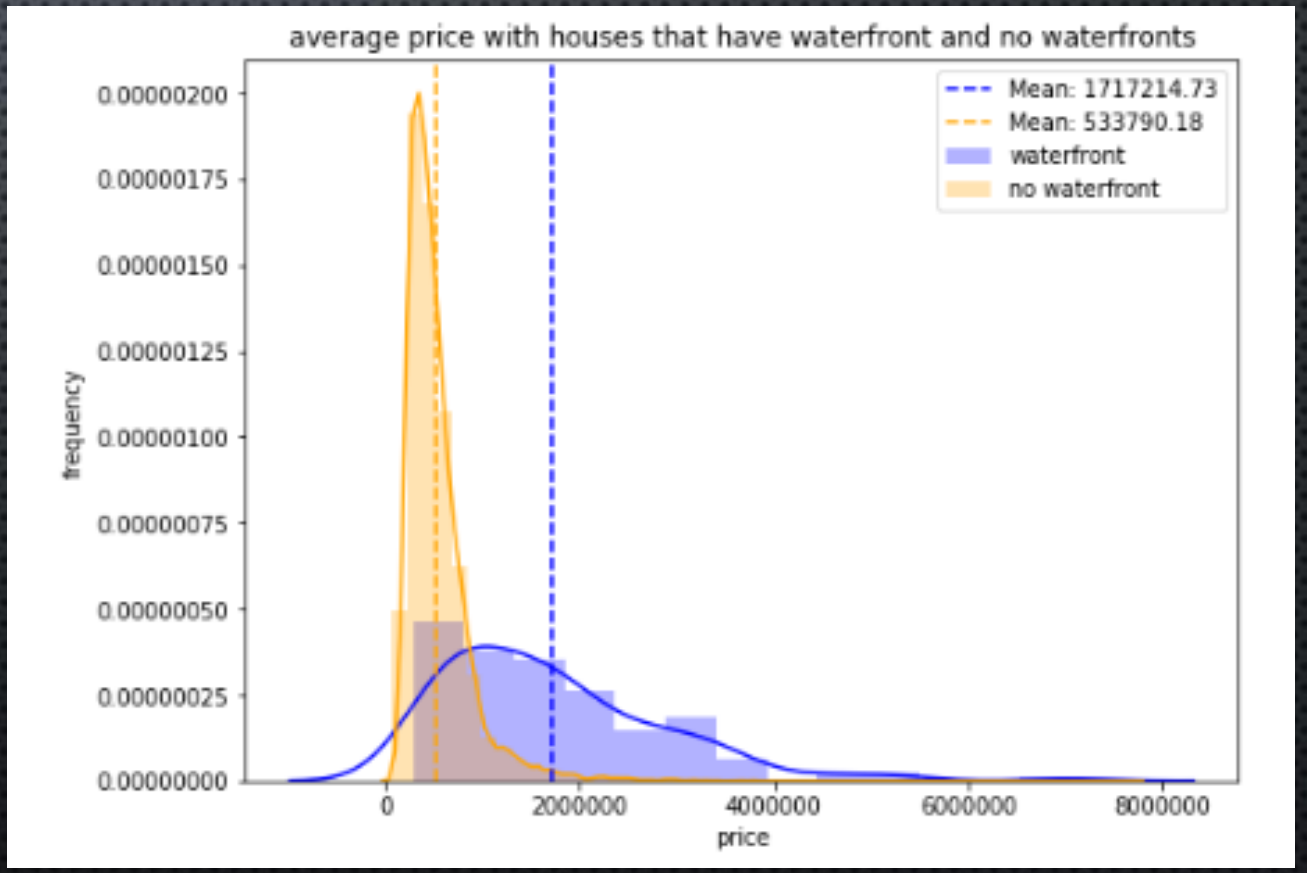
- WE CAN SEE FROM THIS HEATMAP THAT THERE IS COLLINEARITY BETWEEN SQFT_LIVING AND BATHROOMS, AND DUE TO THEIR COLLINEAR RELATIONSHIP THEY WILL BE REMOVED AS A VARIABLE IN THE REGRESSION MODEL. THERE IS ALSO MULTICOLLINEARITY BETWEEN SQFT_LIVING AND BEDROOMS.



DOES GRADE AND WATERFRONTS HAVE AN EFFECT ON PRICE?

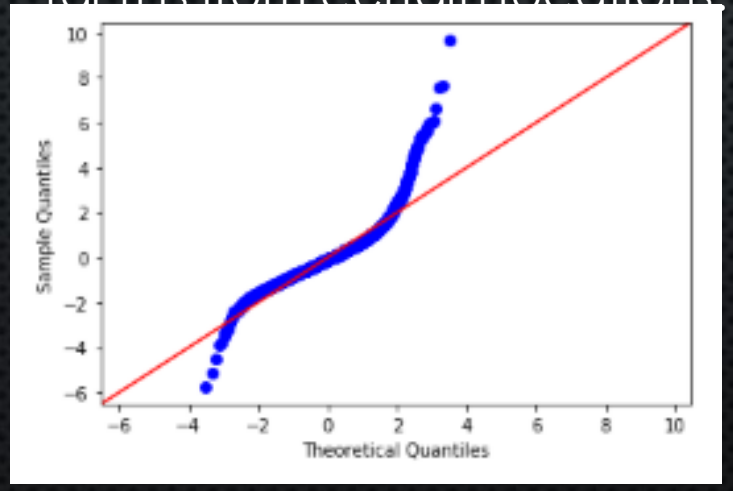






FINAL MODEL- HOW DOES SQFT_LIVING, WATERFRONT AND GRADE AFFECT PRICE?

- After removing all of the unnecessary features, my predictor variables were narrowed down to sqft_living, grade and waterfront.
- Although the r2 squared values received was reasonable, this possible could be improved with access to other data such as when the house was sold, location and how far it is from certain locations.



Dep. Variable:	price	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.575
Method:	Least Squares	F-statistic:	1933.
Date:	Fri, 27 Mar 2020	Prob (F-statistic):	0.00
Time:	11:00:01	Log-Likelihood:	-58941.
No. Observations:	4284	AIC:	1.179e+05
Df Residuals:	4280	BIC:	1.179e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.918e+05	2.78e+04	-17.675	0.000	-5.46e+05	-4.37e+05
sqft_living	178.4286	5.991	29.783	0.000	166.683	190.174
waterfront	9.781e+05	4.38e+04	22.338	0.000	8.92e+05	1.06e+06
grade	8.517e+04	4686.333	18.175	0.000	7.6e+04	9.44e+04

Omnibus:	1741.410	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17353.794
Skew:	1.659	Prob(JB):	0.00
Kurtosis:	12.285	Cond. No.	2.84e+04

S= sqft_living, W= waterfront, G= grade

price = 178S + 978068W + 85175G - 491775

Future work/recommendations:

- Using locations of the houses to determine if those improve the models
- Obtaining more data regarding other benefits a house may have such as distance from schools, and the city, and to see if they improve the model

- For intverstors I recommend that you invest in houses that have waterfronts, as there is a significant increase of price of houses with waterfronts.
- In addition houses that are typically higher in grade also tend to sell for a much higher price.

THANK YOU!