HOUSE SALES

Analysis about house prices in northwestern country

SUMMARY

In this analysis , we used a set of data about house prices in Northwestern country to analyze **what factors drives the house prices** in that area which will be used by real estate agency to advice potential sellers.

OUTLINE

- Business Problem
- Data
- Methods
- Results
- Conclusions

BUSINESS PROBLEM

To provide analysis on **what factors drives the house prices in Northwestern country** to Local Real estate Agency for them to advice potential sellers.

DATA

The analysis is based on a large data set of approximately **21500** housing sales. The data includes many different types of information about each houses. They were cataegorized in to two different types of data:

1. Continuous Data- Sqft Living, Sqft Lot, Price etc
2. Categorical Data –Bedrooms,Bathrooms, Grade, Condition, View etc.

METHODS

1.Data preparation and cleaning

- Understanding the available Data (Scatter plot)
- Dropping few Features as they are not highly related to house prices- (date, lat, long, zipcode, view)

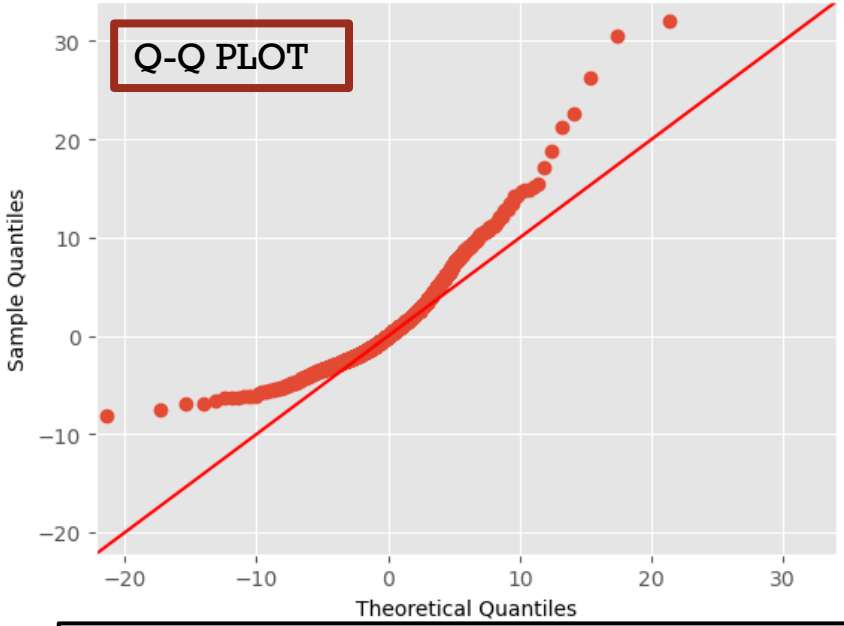
2. Regression modelling

- A baseline model and two Iteration models after to find the best line of fit to predict the house prices in future
- Model Assumptions to Verify the model

2. Model Validation

- Validate the model to see how well the model is generalizing to future cases

BASELINE MODEL



Dep. Variable:	price	R-squared:	0.661
Model:	OLS	Adj. R-squared:	0.660
Method:	Least Squares	F-statistic:	1439.
Date:	Sun, 09 Jul 2023	Prob (F-statistic):	0.00

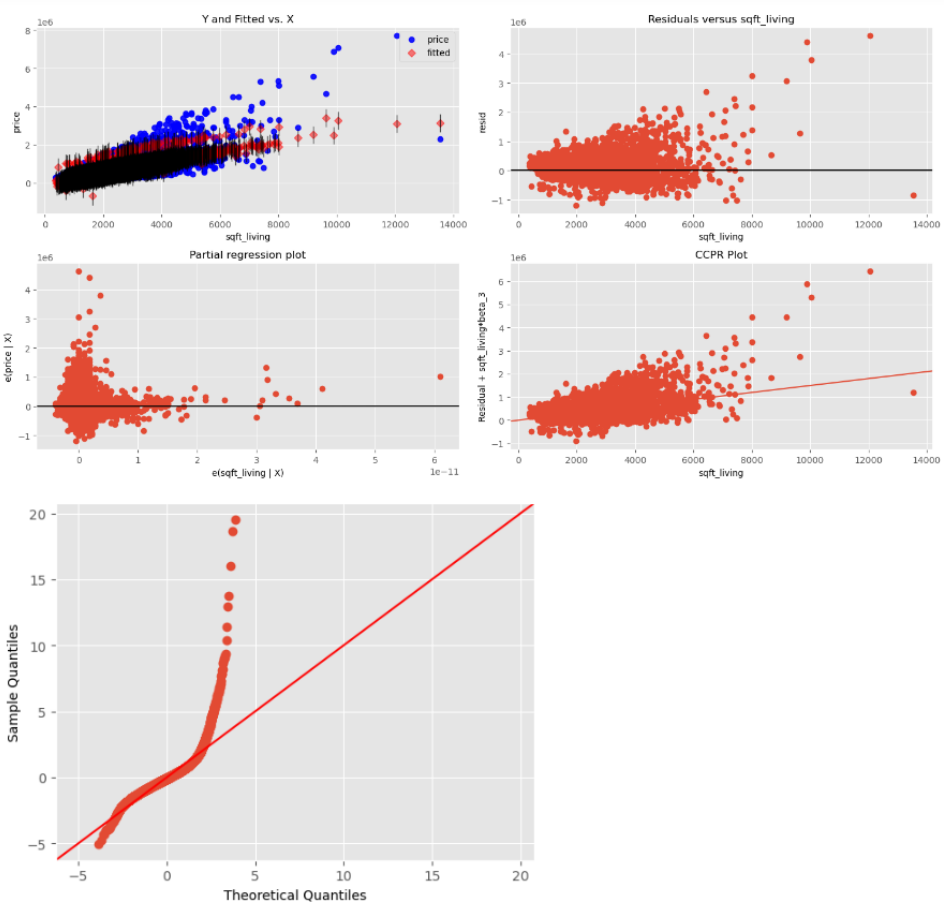
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.69e+06	3.67e+05	15.486	0.000	4.97e+06	6.41e+06
bedrooms	-5281.1550	3722.058	-1.419	0.156	-1.26e+04	2015.140
bathrooms	-1.342e+04	6208.230	-2.162	0.031	-2.56e+04	-1250.215
sqft_living	-3.604e+06	2.35e+05	-15.370	0.000	-4.06e+06	-3.14e+06
sqft_lot	-7.568e+05	4.34e+04	-17.448	0.000	-8.42e+05	-6.72e+05
floors	-9.743e+04	1.01e+04	-9.636	0.000	-1.17e+05	-7.76e+04
condition	8.424e+04	4557.942	18.482	0.000	7.53e+04	9.32e+04
grade	1.381e+05	4524.620	30.531	0.000	1.29e+05	1.47e+05
sqft_basement	1.041e+06	5.95e+04	17.478	0.000	9.24e+05	1.16e+06
waterfront	8.686e+05	2.84e+04	30.538	0.000	8.13e+05	9.24e+05
sqft_above	508.2250	12.398	40.992	0.000	483.921	532.529

R-square is 0.66 which is good but Q-Q plot of the Residual is a curved line suggesting residuals have a non-normal distribution

MODEL ASSUMPTIONS

- Homoscedasticity Plot shows the dependent variable is unequal across the range of values of the independent variable. (Cone-like shape)

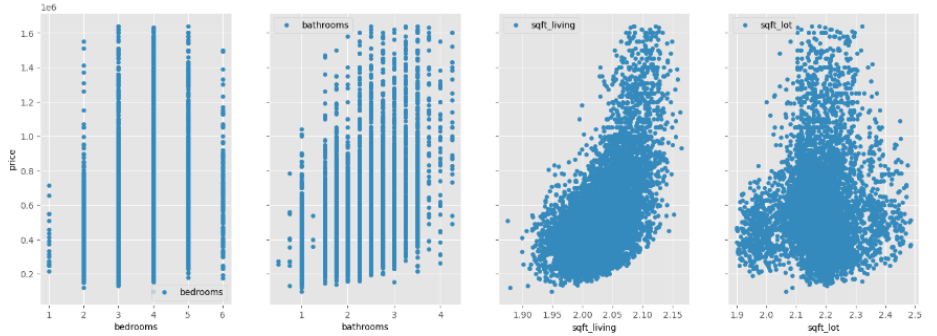
- Normality plots of the Model Residual shows that residual have non normal Distribution
- This suggests we will need to do log Transformation and remove outliers in the next Iteration

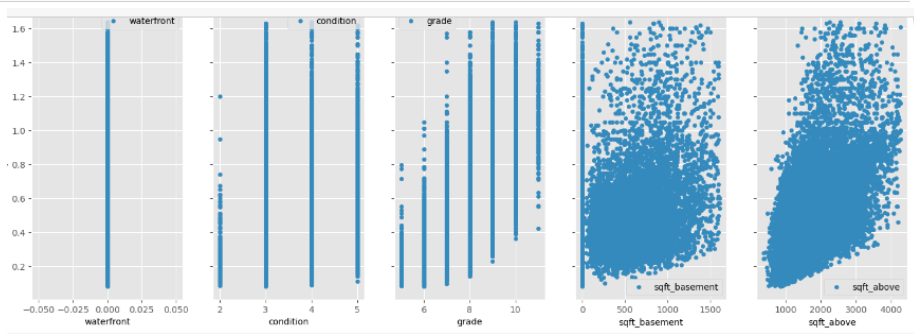


Sq.Ft living Residual Plots

ITERATION 2

- Outliers removed
- Identifying Categorical Data using Scatter plots.(**bedrooms, bathroom, waterfront, condition, grade**)
- Created Dummies for Categorical data
- Derive new variables from dummies by adding them **(Bedrooms2-3, Bedroom>5, Bath<2, Bath2-5, Grade>8)**
- Checking Multicollinearity and dropping one of them



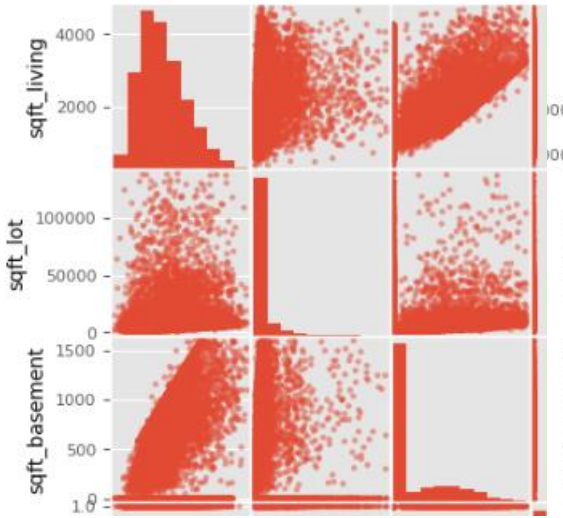


cc

pairs	
(bath1_2, bath2_5)	0.999660
(Bed2_3, Bed4_6)	0.982528
(sqft_living, sqft_above)	0.855987
(condition_3, condition_4)	0.817182

ITERATION 2

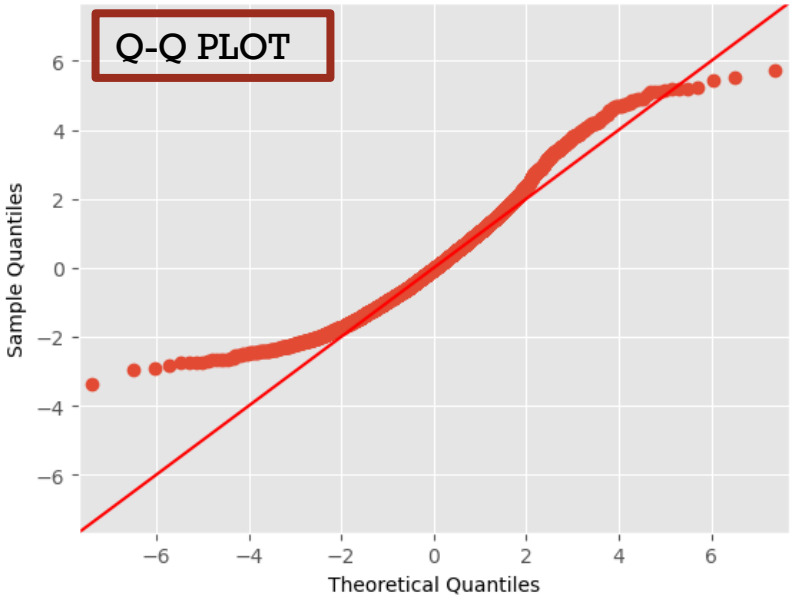
- Log Transformation on Continuous Variables
- The Variables were skewed as shown here
- After log Transformation, the skewness of distribution improved



Before Log Transformation

After Log Transformation

ITERATION 2



Dep. Variable:	price	R-squared:	0.451
Model:	OLS	Adj. R-squared:	0.451
Method:	Least Squares	F-statistic:	702.3
Date:	Sun, 09 Jul 2023	Prob (F-statistic):	0.00

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.374e+06	1.21e+05	-36.299	0.000	-4.61e+06	-4.14e+06
sqft_living	9.017e+06	2.16e+05	41.692	0.000	8.59e+06	9.44e+06
sqft_lot	-1.168e+06	6.73e+04	-17.350	0.000	-1.3e+06	-1.04e+06
sqft_basement	-9.884e+05	6.97e+04	-14.176	0.000	-1.13e+06	-8.52e+05
condition_4	3.747e+04	5568.255	6.729	0.000	2.66e+04	4.84e+04
condition_5	1.15e+05	8116.578	14.174	0.000	9.91e+04	1.31e+05
Bed4_6	-1.537e+04	5640.171	-2.724	0.006	-2.64e+04	-4309.061
bath2_5	-3.003e+04	5900.270	-5.089	0.000	-4.16e+04	-1.85e+04
grade8_11	1.196e+05	6140.500	19.475	0.000	1.08e+05	1.32e+05

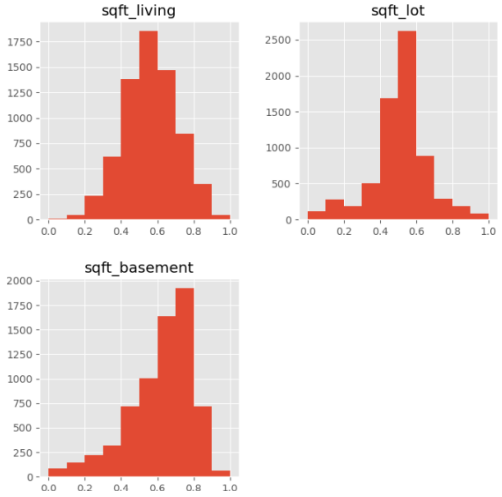
Adjusted R-square is 0.451 which is less compared to baseline model but Q-Q plot of the Residual is better leaning towards the fit suggesting residuals have a a better normal distribution suggesting the fit of the model is better.

ITERATION 3

- Apply Feature Scaling and Min-Max Scaling on the continuous Variable
- Very minor difference in Skewness but still better Distribution after Scaling

- Drop Variable(Sq.ft Basement)as it contains Nan Values and not very relevant

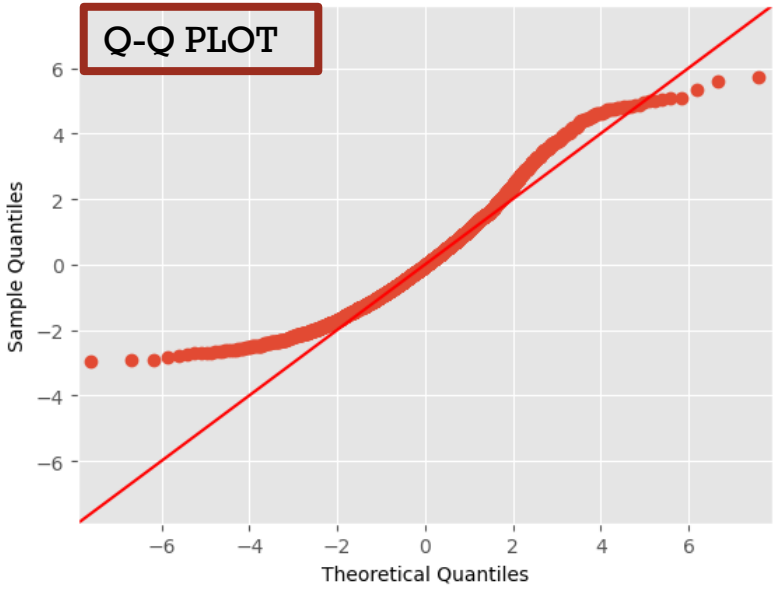
Before Scaling



After Scaling

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sqft_living   6842 non-null   float64
 1   sqft_lot      6842 non-null   float64
 2   price         6842 non-null   float64
 3   Bed4_6        6842 non-null   float64
 4   bath2_5       6842 non-null   float64
 5   grade8_11     6842 non-null   float64
 6   condition4_5  6842 non-null   float64
dtypes: float64(7)
memory usage: 374.3 KB
```

ITERATION 3



Dep. Variable:	price	R-squared:	0.430
Model:	OLS	Adj. R-squared:	0.429
Method:	Least Squares	F-statistic:	858.2
Date:	Sun, 09 Jul 2023	Prob (F-statistic):	0.00

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.249e-16	0.009	1.37e-14	1.000	-0.018	0.018
sqft_living	0.5651	0.014	41.407	0.000	0.538	0.592
sqft_lot	-0.2014	0.010	-19.275	0.000	-0.222	-0.181
condition4_5	0.0881	0.009	9.404	0.000	0.070	0.106
Bed4_6	-0.0354	0.011	-3.345	0.001	-0.056	-0.015
bath2_5	-0.0399	0.011	-3.615	0.000	-0.062	-0.018
grade8_11	0.2539	0.011	22.595	0.000	0.232	0.276

Adjusted R-square is 0.429 which is less compared to previous models but Q-Q plot of the Residual is the almost the same as Iteration 2
The coefficients are used from this model for predicting the house prices

RESULTS

Based on Iteration 3 Model Summary, following predictions are made:

- There is positive relation ship between Price and Sqft.living, Condition and Grade
- For each unit increase in Sq.ft living, there will 0.56 increase in price
- For each unit increase in condition, there will 0.08 increase in price
- For each unit increase in Grade, there will 0.25 increase in price
- There is negative relationship between Price and Sq.ft lot, bedrooms >4, and Bath>2

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.249e-16	0.009	1.37e-14	1.000	-0.018	0.018
sqft_living	0.5651	0.014	41.407	0.000	0.538	0.592
sqft_lot	-0.2014	0.010	-19.275	0.000	-0.222	-0.181
condition4_5	0.0881	0.009	9.404	0.000	0.070	0.106
Bed4_6	-0.0354	0.011	-3.345	0.001	-0.056	-0.015
bath2_5	-0.0399	0.011	-3.615	0.000	-0.062	-0.018
grade8_11	0.2539	0.011	22.595	0.000	0.232	0.276

MODEL VALIDATION

The model validation was done using Train/Split method where 70% of sample used for training and 30% for testing.
The Mean square error of the residuals of both the samples are below:

```
Train Mean Squared Error: 0.5732431710926651
Test Mean Squared Error:  0.5617337437752656
```

There does not seem to be a big difference between the train and test MSE! Thus we can say the model is generalizing well to future cases and is the best line of Fit.

THANK YOU!

EMAIL: e.gajanani9@gmail.com
GITHUB: @Gajas9
LINKEDIN: LINKEDIN.COM/IN/GAJA-SANCHAYAN/