House Price Regression Modelling Project

Nat Berryman

Summary

The purpose of this regression model is to predict the house prices in King County by analysing the King County House Sales dataset.

Outline

- Business Problem
- Cleaning the Data
- Exploratory Data Analysis
- Models
- Conclusions

Business Problem

How can we predict the house price sales in King County?

In order to solve this problem, I intended to answer the below questions:

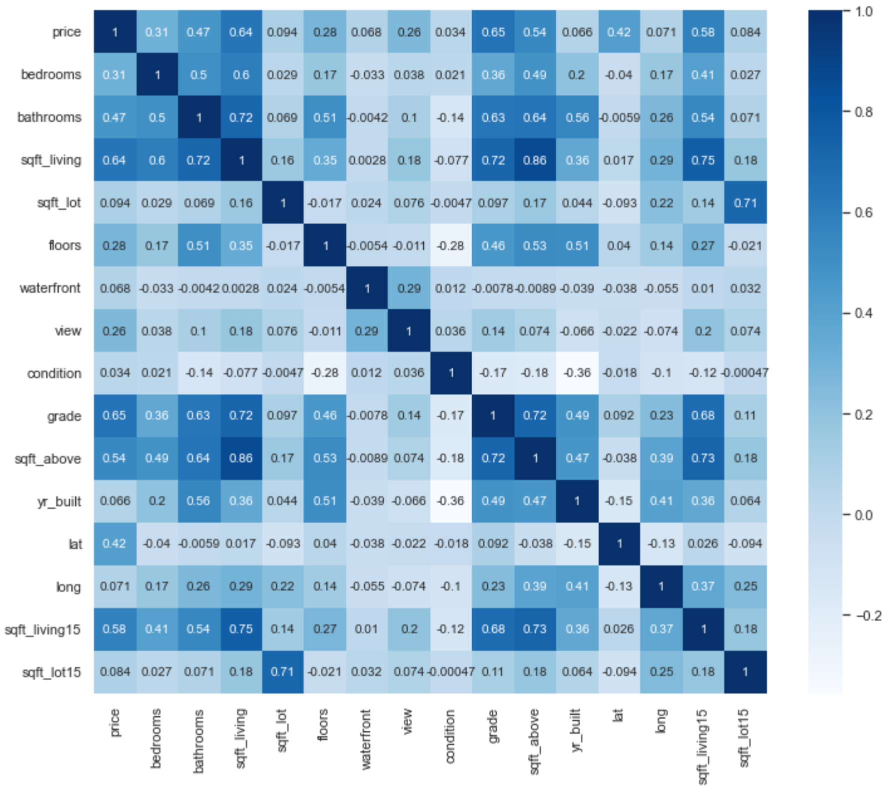
1. Does location impact sale price?
2. Does the size of the house impact sale price?
3. Does quality of the house impact sale price?

Cleaning the Data

- Dropped unnecessary data
- Replaced or removed null values
- Narrowed data to only included houses with <6 bedrooms
- Using the empirical formula I removed outliers
- Addressed multicollinearity
- Split data set between continuous and categorical data
- Binned Grade into Low, Average and High

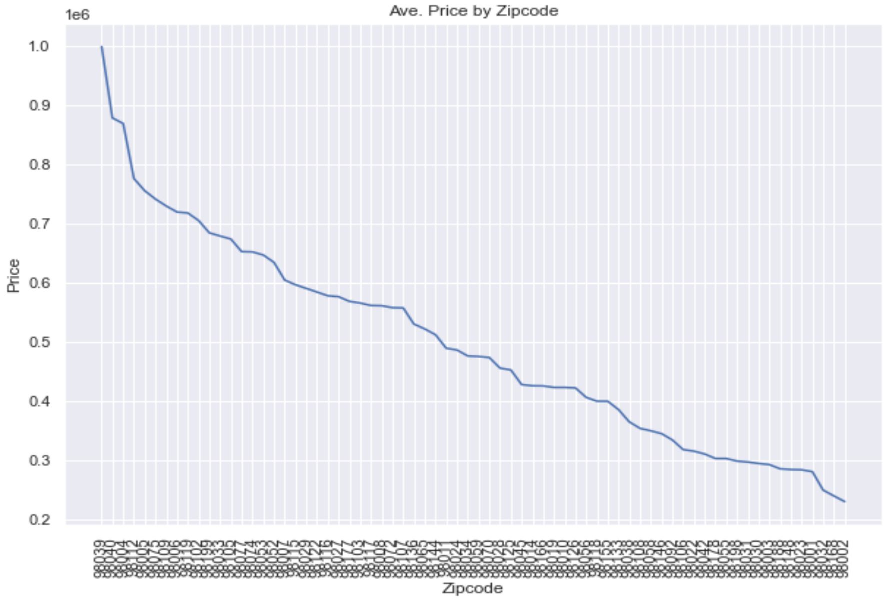
Exploratory Data Analysis

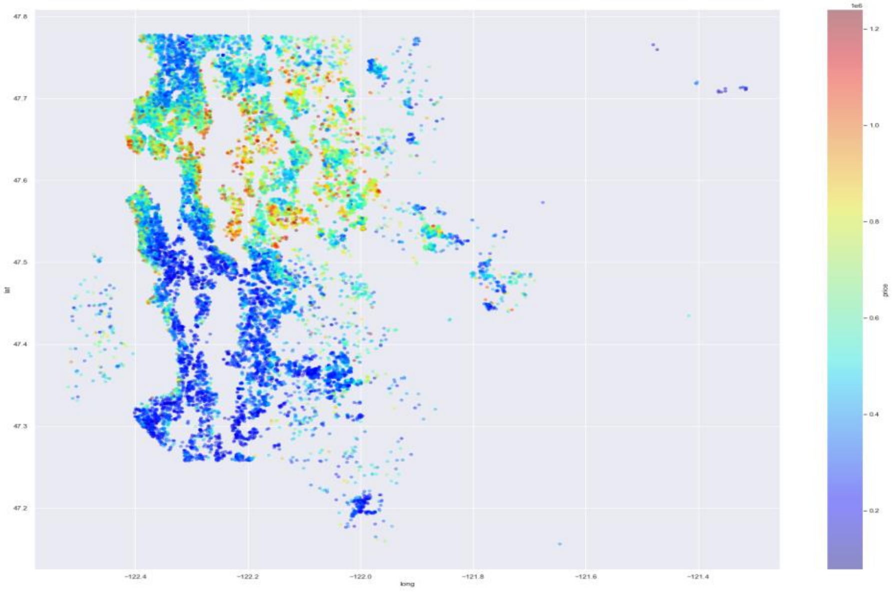
- Key Features include:
  - Bathrooms
  - Square Foot living space
  - Grade
  - Latitude
  - Square Foot Above
  - Square Foot Living 15 (neighbors)
- Notes:
  - Zip code excluded as data-type is a string
  - Latitude correlates with price more than Longitude



Exploratory Data Analysis

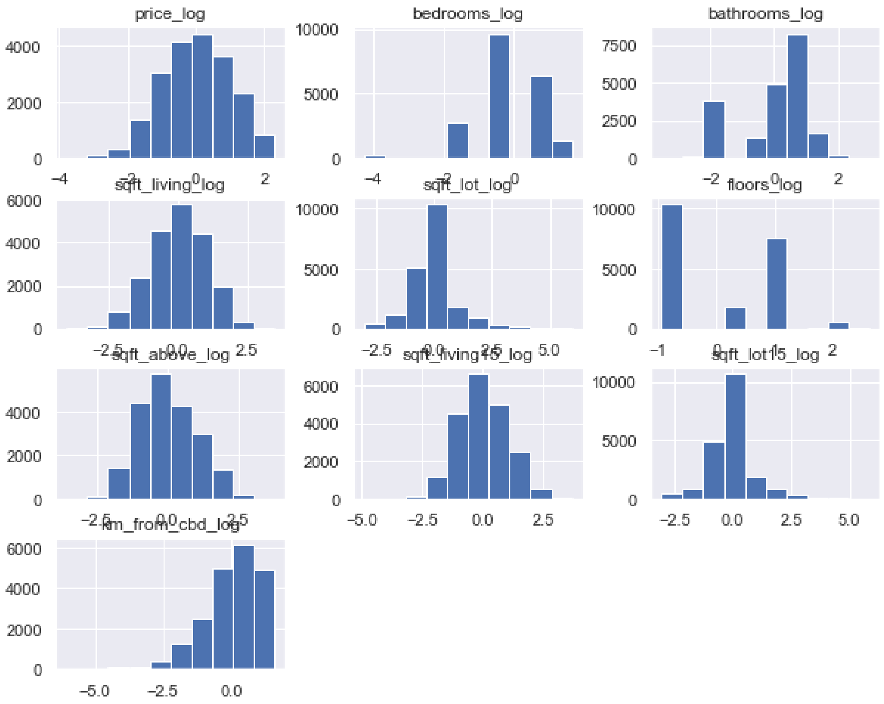
- Created price vs zip code graph to explore price distribution across zip codes and then plotted to a heatmap.
- Using these visualization I created a new variable – Distance from CBD





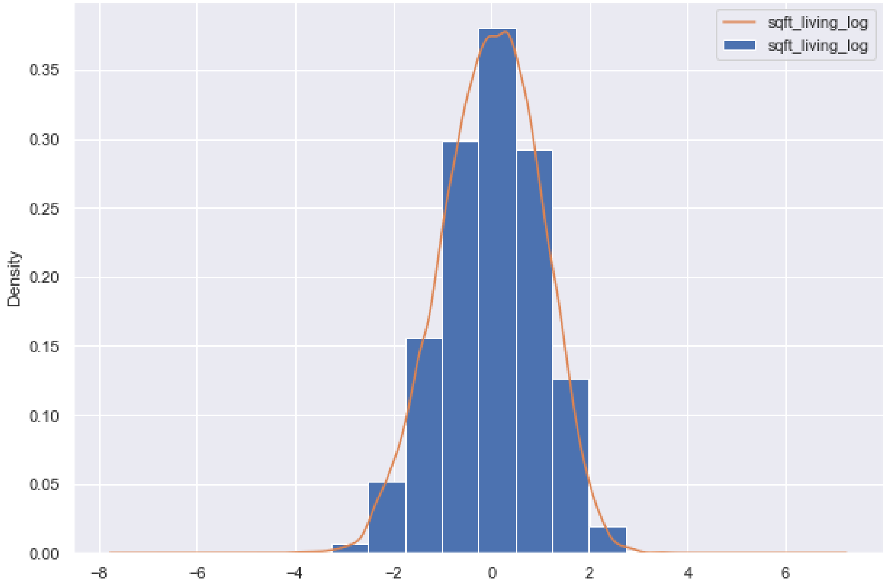
Exploratory Data Analysis

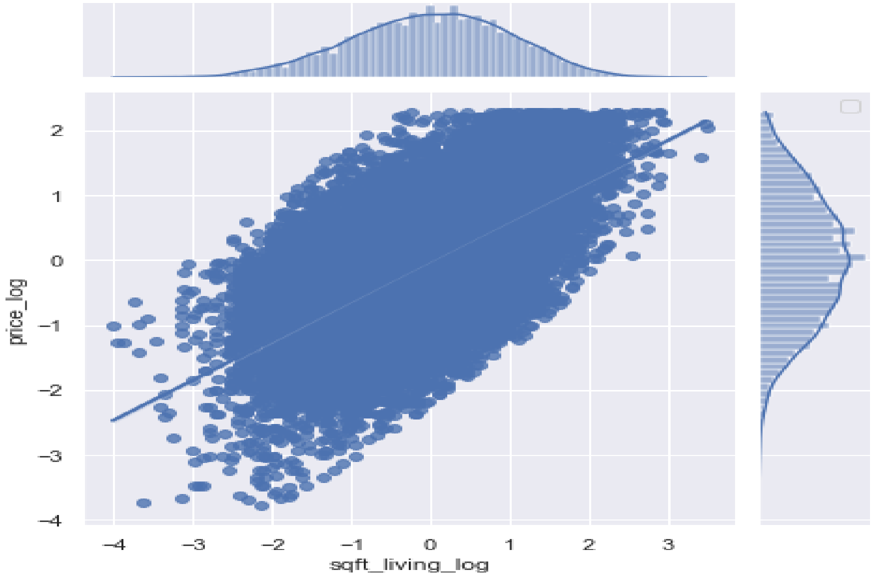
- Used mean normalization to standardise the data
- Sqft living, sqft lot, sqft above, sqft living 15, sqft lot 15 appear good
- Km from CBD is negatively skewed



Exploratory Data Analysis

- Used KDE plot and joint plot to explore data





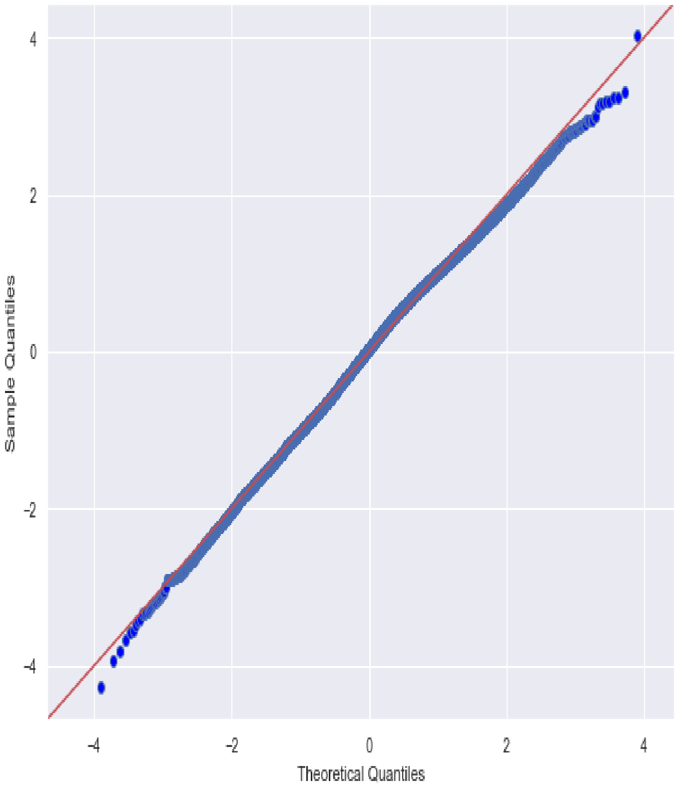
Model 1

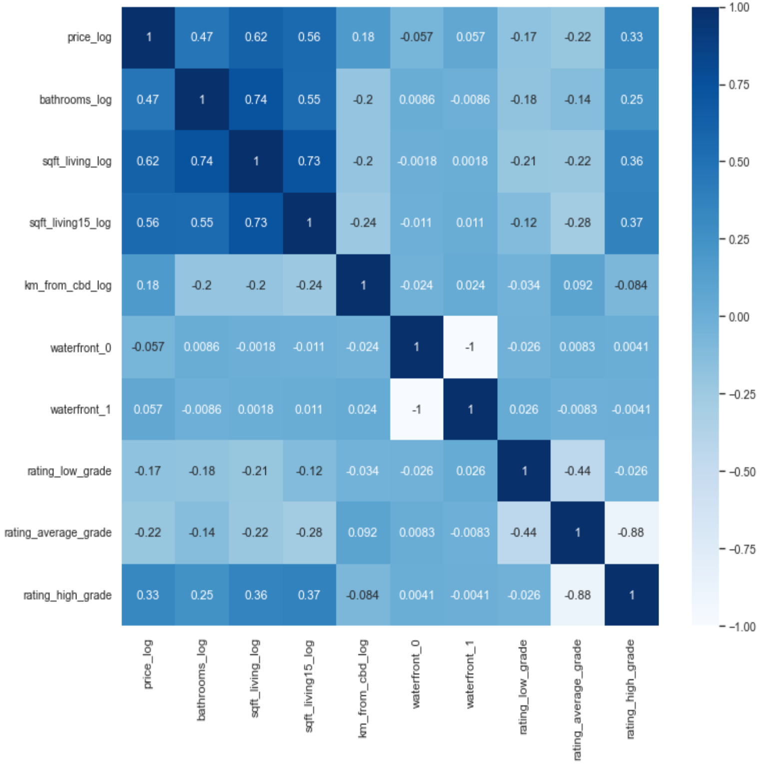
OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.527
Model:	OLS	Adj. R-squared:	0.527
Method:	Least Squares	F-statistic:	3244.
Date:	Fri, 03 Jun 2022	Prob (F-statistic):	0.00
Time:	10:23:16	Log-Likelihood:	-21322.
No. Observations:	20407	AIC:	4.266e+04
Df Residuals:	20399	BIC:	4.272e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0273	0.012	2.268	0.023	0.004	0.051
bathrooms_log	0.0632	0.007	8.810	0.000	0.049	0.077
sqft_living_log	0.3877	0.009	43.476	0.000	0.370	0.405
sqft_living15_log	0.2796	0.007	38.529	0.000	0.265	0.294
km_from_cbd_log	0.3396	0.005	67.982	0.000	0.330	0.349
waterfront_1	0.8624	0.089	9.685	0.000	0.688	1.037
rating_low_grade	-0.3423	0.034	-10.215	0.000	-0.408	-0.277
rating_average_grade	-0.0493	0.013	-3.929	0.000	-0.074	-0.025
rating_high_grade	0.4189	0.021	19.581	0.000	0.377	0.461

Omnibus:	92.698	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	90.193
Skew:	-0.144	Prob(JB):	2.60e-20
Kurtosis:	2.848	Cond. No.	7.34e+15





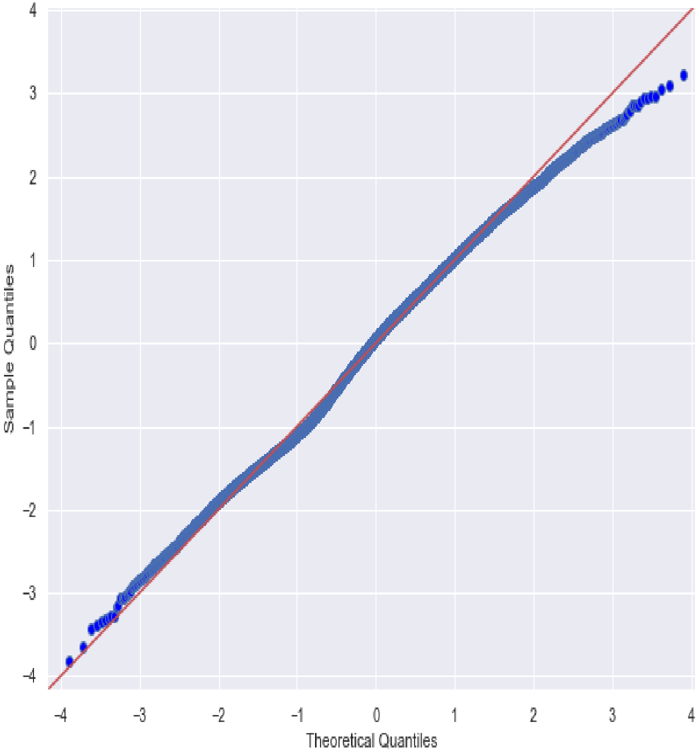
Model 2

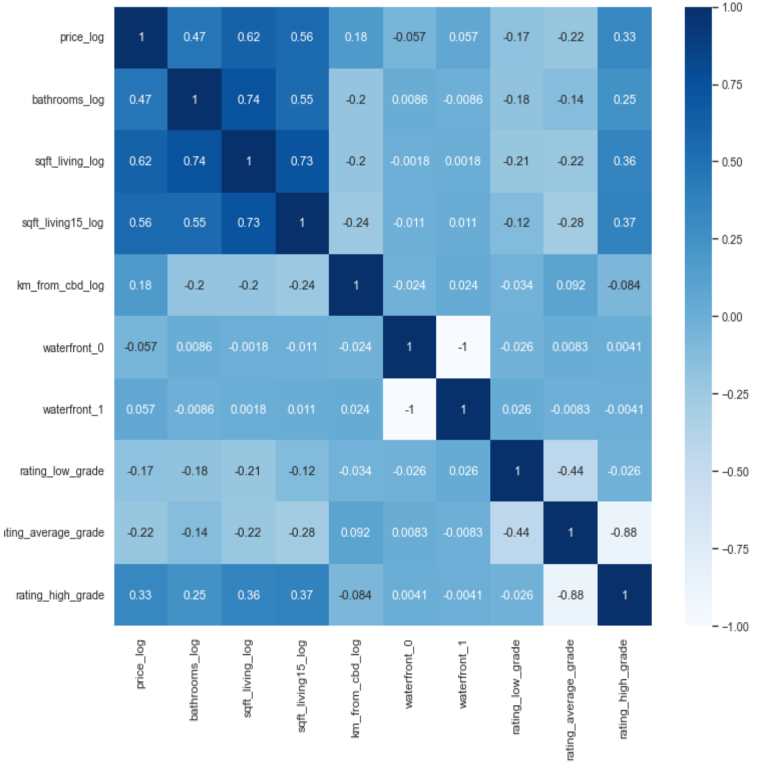
OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.413
Model:	OLS	Adj. R-squared:	0.413
Method:	Least Squares	F-statistic:	3592.
Date:	Fri, 03 Jun 2022	Prob (F-statistic):	0.00
Time:	10:23:18	Log-Likelihood:	-23516.
No. Observations:	20407	AIC:	4.704e+04
Df Residuals:	20402	BIC:	4.708e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0235	0.006	-4.259	0.000	-0.034	-0.013
bathrooms_log	0.0280	0.008	3.521	0.000	0.012	0.044
sqft_living_log	0.4062	0.010	41.272	0.000	0.387	0.425
sqft_living15_log	0.2075	0.008	25.980	0.000	0.192	0.223
rating_high_grade	0.4723	0.027	17.611	0.000	0.420	0.525

Omnibus:	305.924	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	187.564
Skew:	-0.072	Prob(JB):	1.87e-41
Kurtosis:	2.553	Cond. No.	7.71





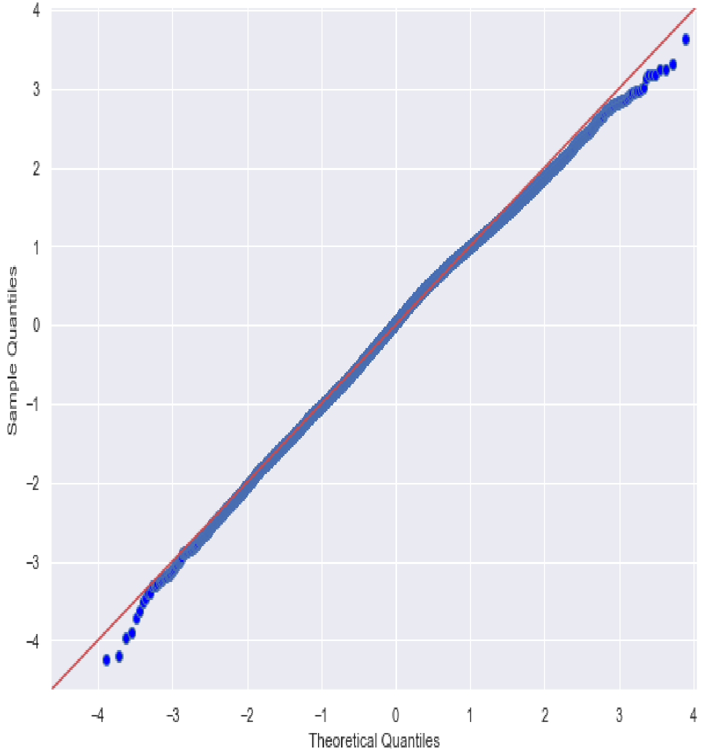
Model 3

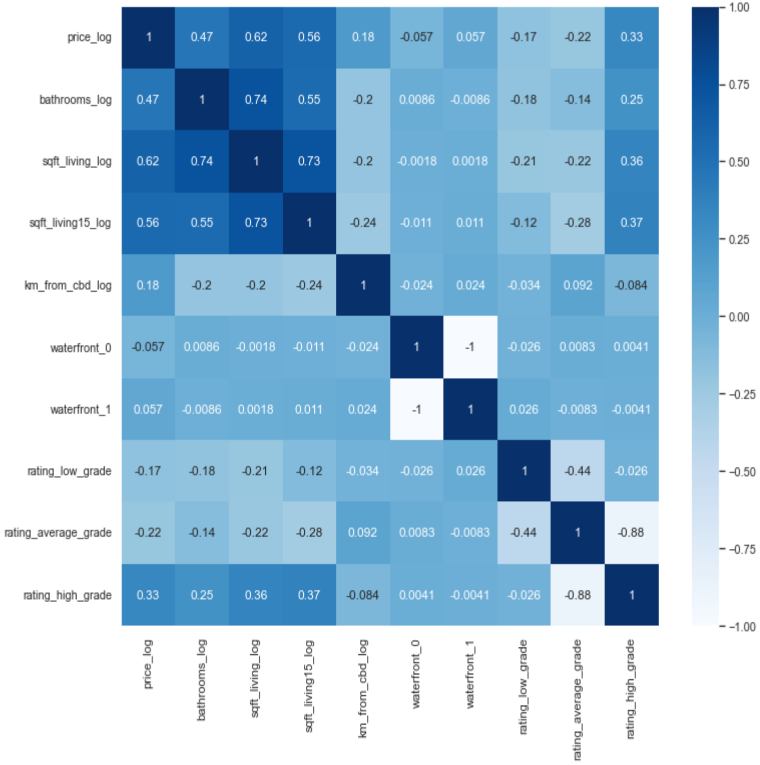
OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.524
Model:	OLS	Adj. R-squared:	0.524
Method:	Least Squares	F-statistic:	4485.
Date:	Fri, 03 Jun 2022	Prob (F-statistic):	0.00
Time:	10:23:18	Log-Likelihood:	-21390.
No. Observations:	20407	AIC:	4.279e+04
Df Residuals:	20401	BIC:	4.284e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0229	0.005	-4.593	0.000	-0.033	-0.013
sqft_living15_log	0.2791	0.007	38.373	0.000	0.265	0.293
sqft_living_log	0.3956	0.009	44.604	0.000	0.378	0.413
km_from_cbd_log	0.3434	0.005	68.749	0.000	0.334	0.353
rating_high_grade	0.4590	0.024	18.991	0.000	0.412	0.506
bathrooms_log	0.0645	0.007	8.973	0.000	0.050	0.079

Omnibus:	92.684	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	91.160
Skew:	-0.149	Prob(JB):	1.60e-20
Kurtosis:	2.863	Cond. No.	7.86

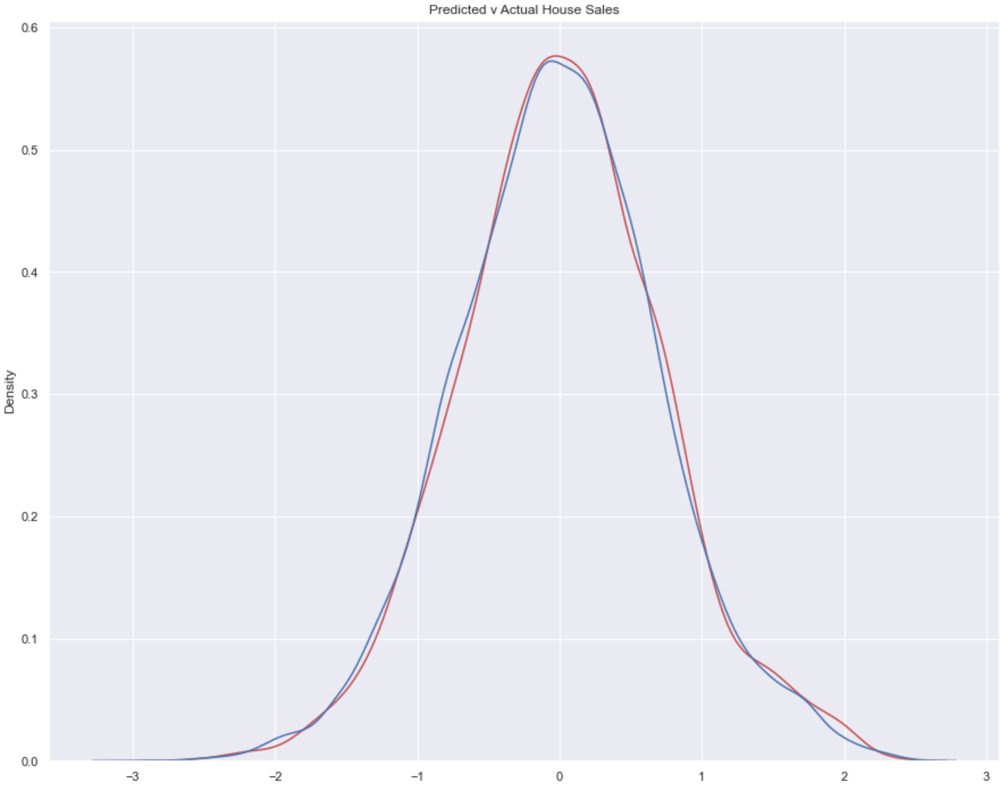




	price_log	bathrooms_log	sqft_living_log	sqft_living15_log	km_from_cbd_log	waterfront_0	waterfront_1	rating_low_grade	rating_average_grade	rating_high_grade
price_log	1	0.47	0.62	0.56	0.18	-0.057	0.057	-0.17	-0.22	0.33
bathrooms_log	0.47	1	0.74	0.55	-0.2	0.0086	-0.0086	-0.18	-0.14	0.25
sqft_living_log	0.62	0.74	1	0.73	-0.2	-0.0018	0.0018	-0.21	-0.22	0.36
sqft_living15_log	0.56	0.55	0.73	1	-0.24	-0.011	0.011	-0.12	-0.28	0.37
km_from_cbd_log	0.18	-0.2	-0.2	-0.24	1	-0.024	0.024	-0.034	0.092	-0.084
waterfront_0	-0.057	0.0086	-0.0018	-0.011	-0.024	1	-1	-0.026	0.0083	0.0041
waterfront_1	0.057	-0.0086	0.0018	0.011	0.024	-1	1	0.026	-0.0083	-0.0041
rating_low_grade	-0.17	-0.18	-0.21	-0.12	-0.034	-0.026	0.026	1	-0.44	-0.026
rating_average_grade	-0.22	-0.14	-0.22	-0.28	0.092	0.0083	-0.0083	-0.44	1	-0.88
rating_high_grade	0.33	0.25	0.36	0.37	-0.084	0.0041	-0.0041	-0.026	-0.88	1

Conclusions

- Model 3 provided most reliable result with R^ of 0.524
- Selected features all statistically significant with p-value <0.05
- sqft living15 coef – 0.2791
- sqft living coef – 0.3956
- distance from CBD coef - 0.3434
- Bathrooms coef – 0.0645
- high grade rating coef– 0.4590
- These Coef figures mean for unit increase in any one of these variables there was in increase in price by ~0.3 units.



Thank You!

Email: nathaniel.berryman@gmail.com
GitHub: @natberr
LinkedIn: linkedin.com/in/nathaniel-berryman/