King County sales price Analysis for for real estate investor

ANALYSIS BY SHRADHA WADDEPALLI

Business Problem

I am creating a model to predict the sales prices for real estate investor in King County

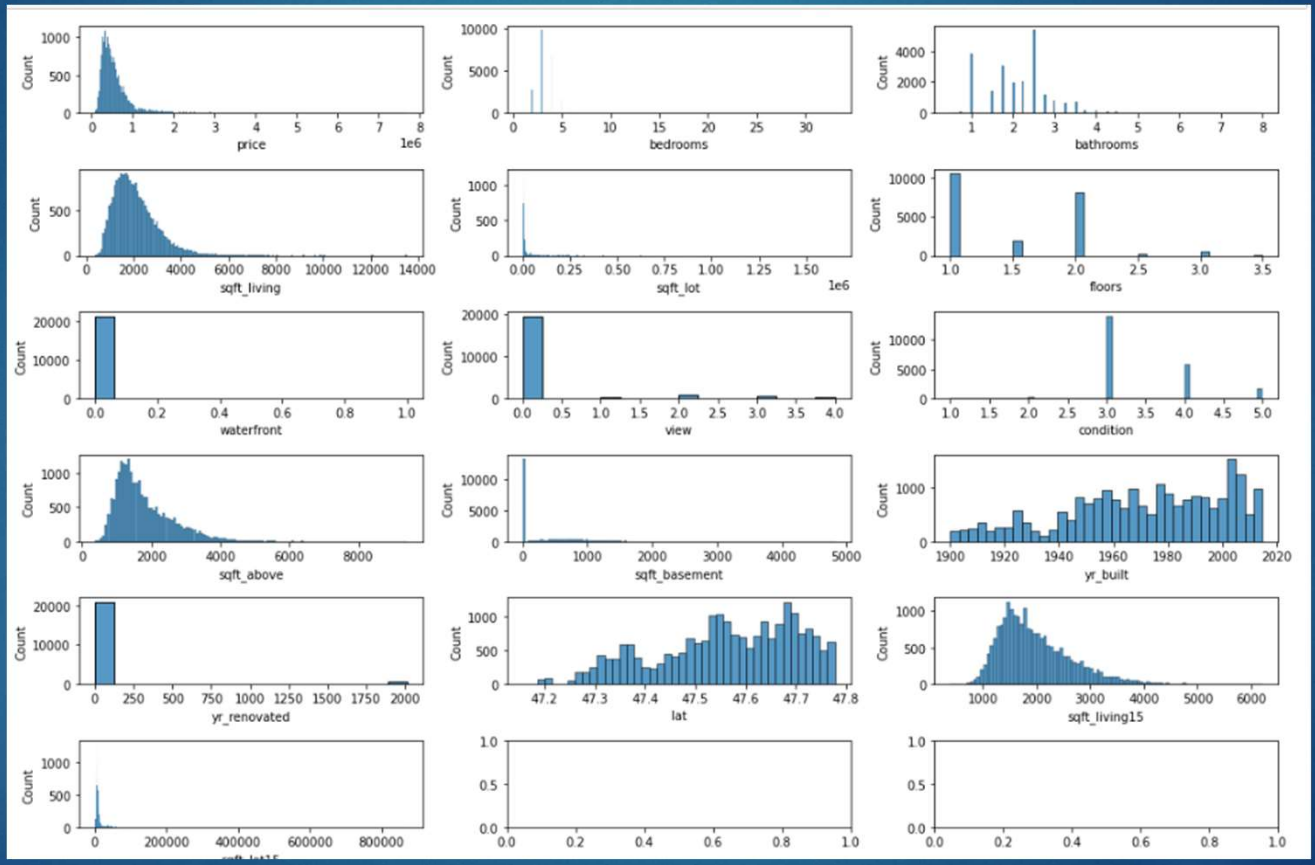
Data

King County dataset has 21,597 rows and 17 columns consisting of important information number of factors such as bedrooms, bathrooms, square foot area etc.

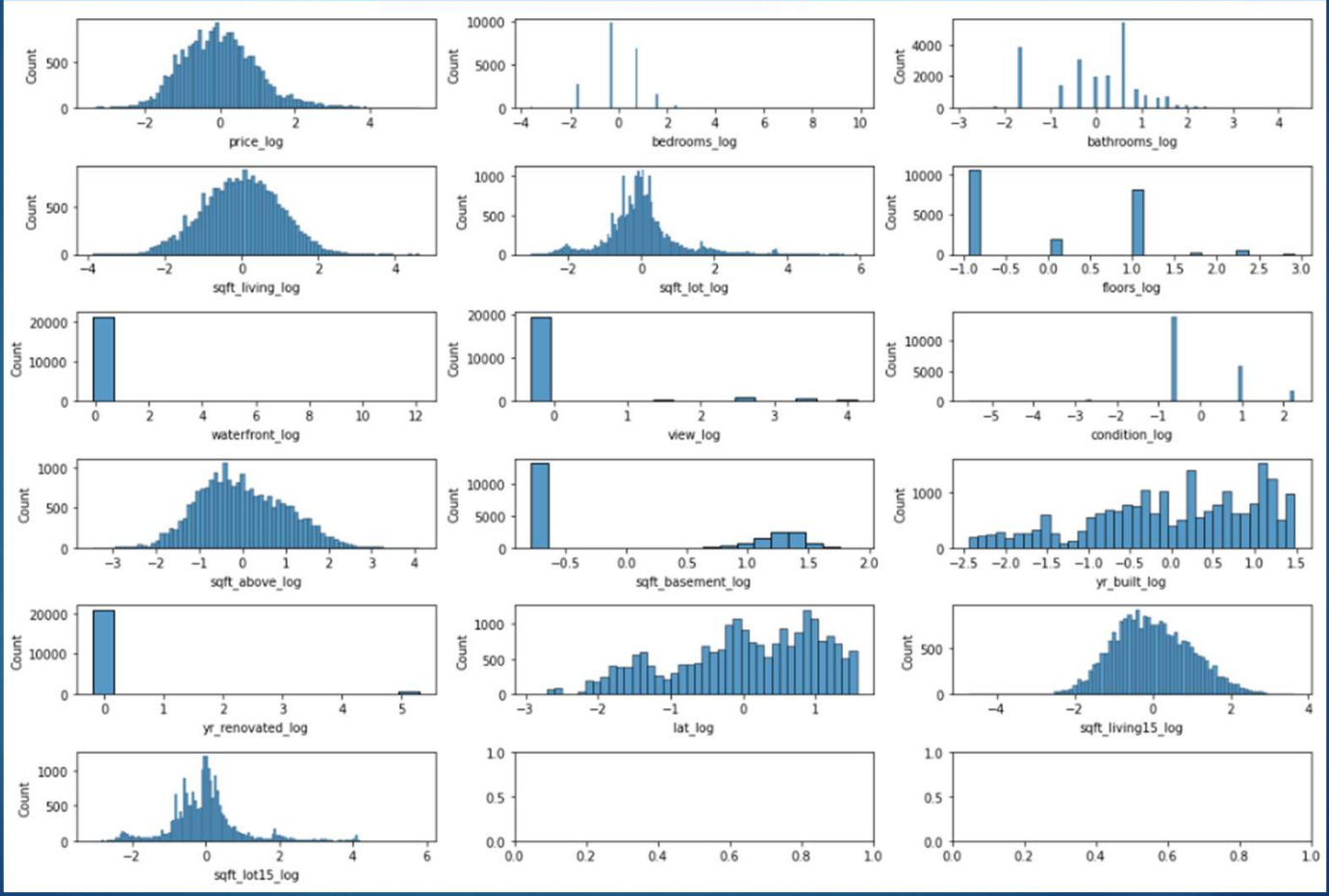
Data cleaning: Square foot basement, year renovated, waterfront and View columns needed to be cleaned to be able to proceed with analysis.

After recognizing continuous variable and discreet variables. I performed One Hot Encoding, which produced additional columns to be a total of 88 columns

EDA- continuous variables



EDA- Log transformed and Normalised



Pearson's correlation >0.6 variables

- Splitting Train and test with a ratio of 75% - 25%
- Created simple regression model with these three variables.
- Also predicted values within train and test data

	index	price_log
2	sqft_living_log	0.674829
13	sqft_living15_log	0.607167
86	grade	0.703720

Statistical parameters for analysis

R squared
adjusted R squared
Mean squared error

Cross Validation to refine model using 10 splits.

Simple regression model with three variables

- Microsoft should produce movies with either of following Directors.

	fit_time	score_time	test_r2	train_r2	test_neg_mean_squared_error	train_neg_mean_squared_error	n_features	dataset
0	0.001935	0.001218	0.547638	0.551303	-0.449934	-0.448647	3	simple

Final model

- Microsoft should get the stories written by either of following Writers or combination of writers

	fit_time	score_time	test_r2	train_r2	test_neg_mean_squared_error	train_neg_mean_squared_error	n_features	dataset
0	0.001935	0.001218	0.547638	0.551303	-0.449934	-0.448647	3	simple
1	0.004135	0.001094	0.375939	0.381973	-0.619628	-0.617950	7	with 7 more variables
2	0.043893	0.002366	0.860045	0.863197	-0.138990	-0.136787	88	with 88 more variables

Summary

Coefficient of determination : R Squared is between 0-1, higher is better.

MSE : lower the better.

	fit_time	score_time	test_r2	train_r2	test_neg_mean_squared_error	train_neg_mean_squared_error	n_features	dataset
0	0.001935	0.001218	0.547638	0.551303	-0.449934	-0.448647	3	simple
1	0.004135	0.001094	0.375939	0.381973	-0.619628	-0.617950	7	with 7 more variables
2	0.043893	0.002366	0.860045	0.863197	-0.138990	-0.136787	88	with 88 more variables

Email: shradha.waddepalli@gmail.com
GitHub: @shrwad