Zillow House Prediction

Springboard Data Science Career Track Program

Ly Nguyen

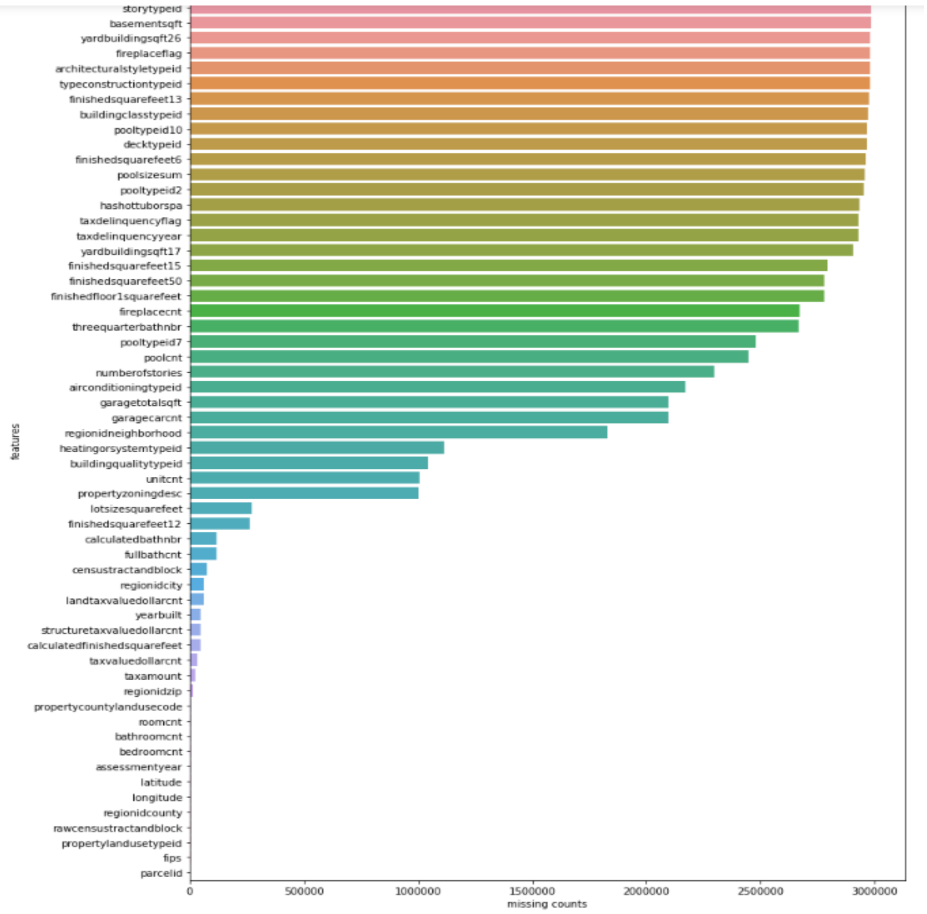
Introduction

- Selling and buying house demands are increased year by year
- Investors are interested in seeing accurate house price prediction before they actually buy or sell their properties

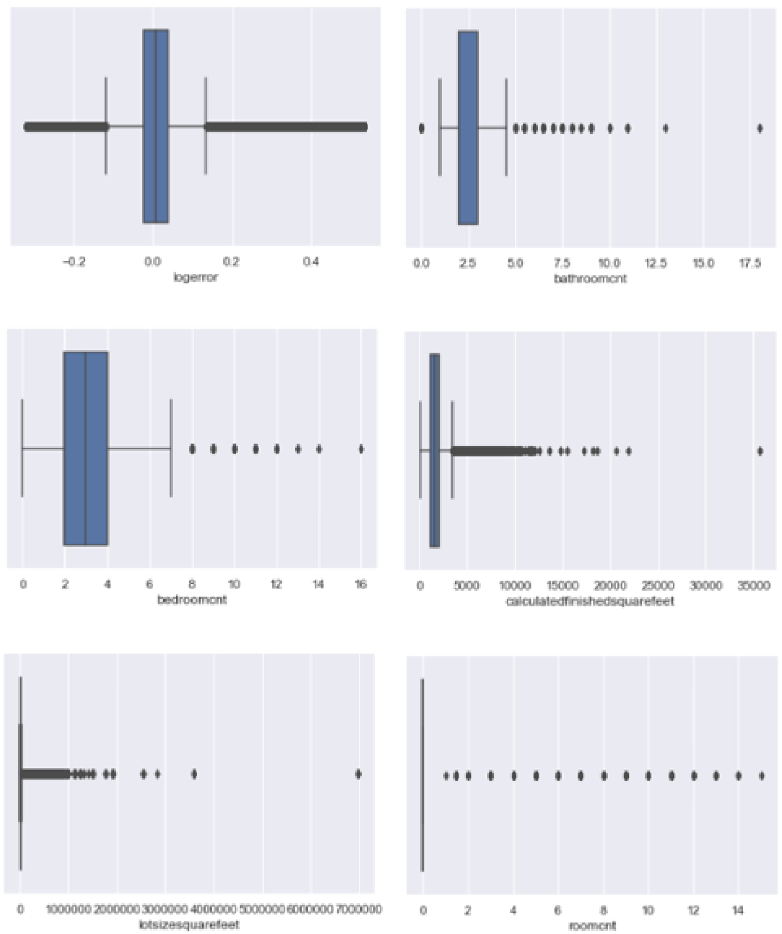
Data

- Data source: Zillow
- Data characteristics:
    - Train data: contains 77613 log errors of prediction from Jan to Sep 2017
    - Properties data: includes 2985217 rows and 58 columns. It include data from 3 counties in California: Los Angeles, Orange, Ventura
    - 50% of the columns have more than 65% of missing data

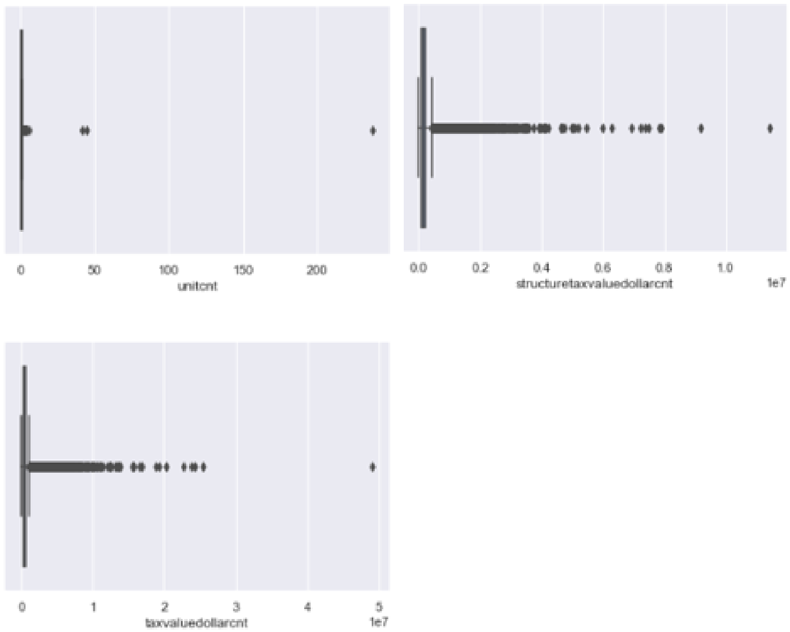
Missing data



Outliers

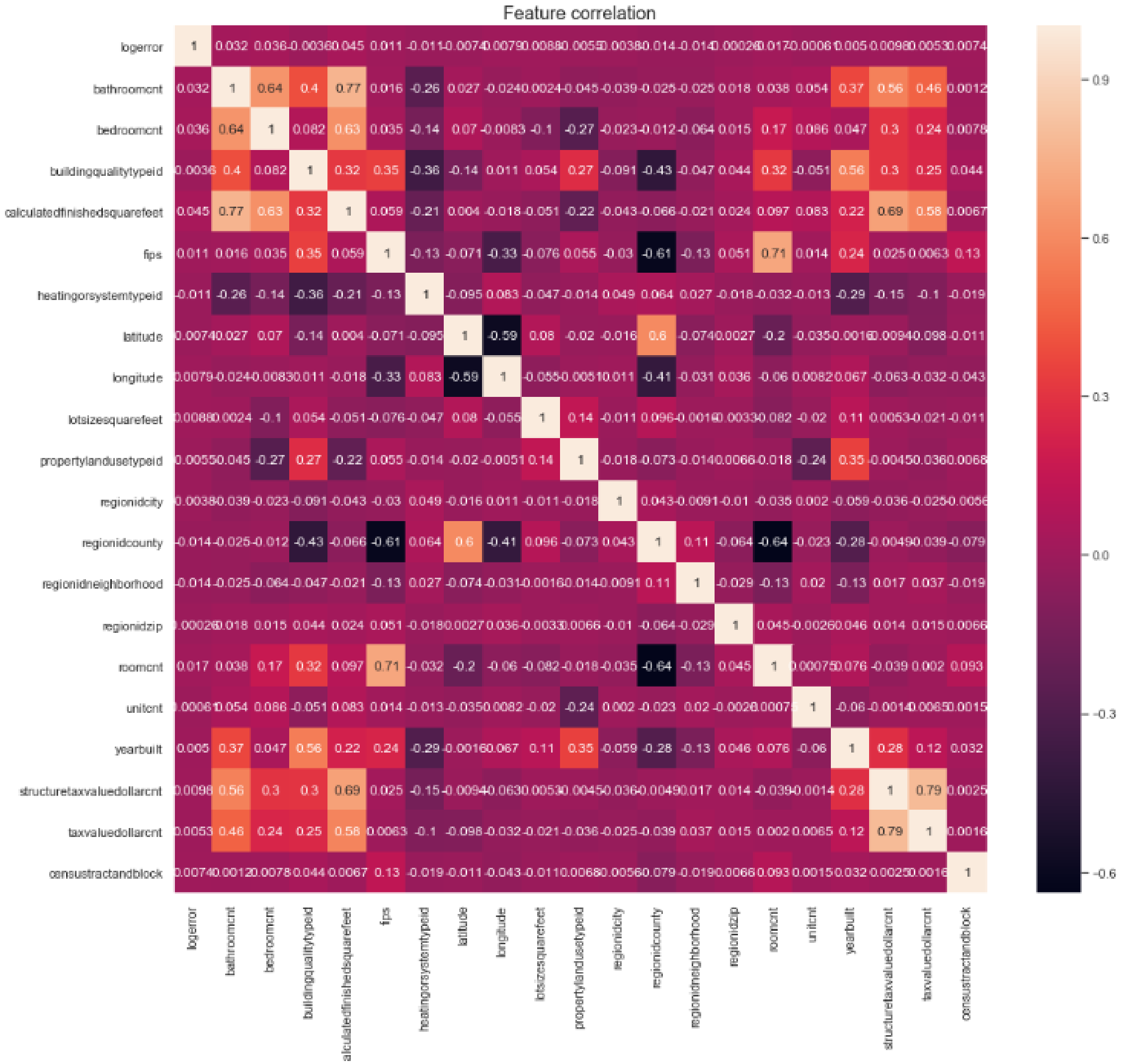


logerror
bathroomcnt
bedroomcnt
calculatedfinishedsquarefeet
lotsizesquarefeet
roomcnt

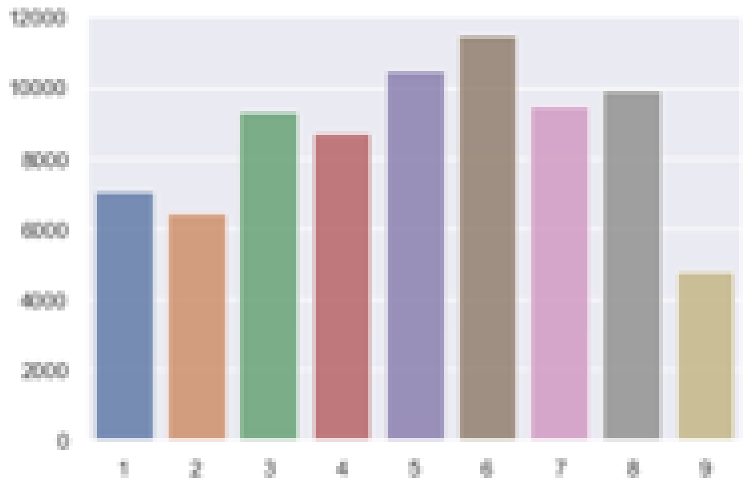


unitcnt
structuretaxvaluedollarcnt
1e7
taxvaluedollarcnt
1e7

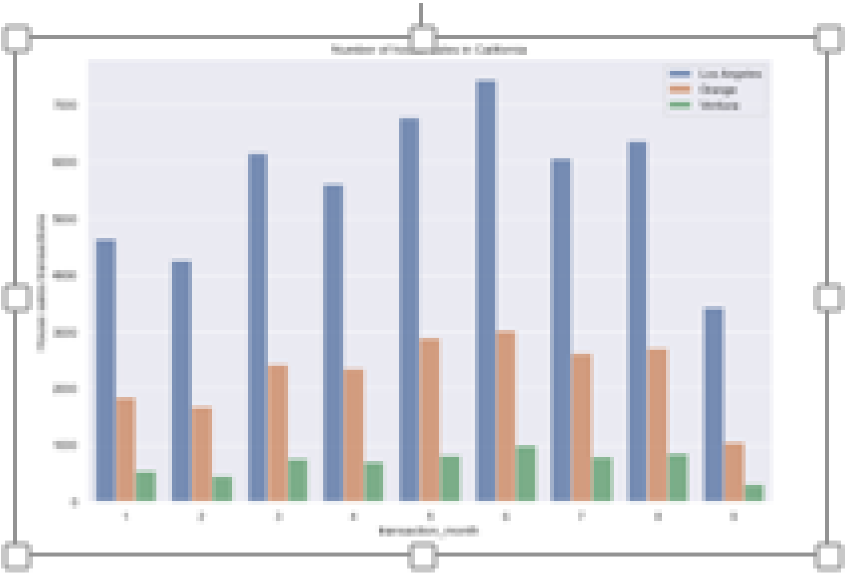
Correlation Analysis



Exploratory Analysis

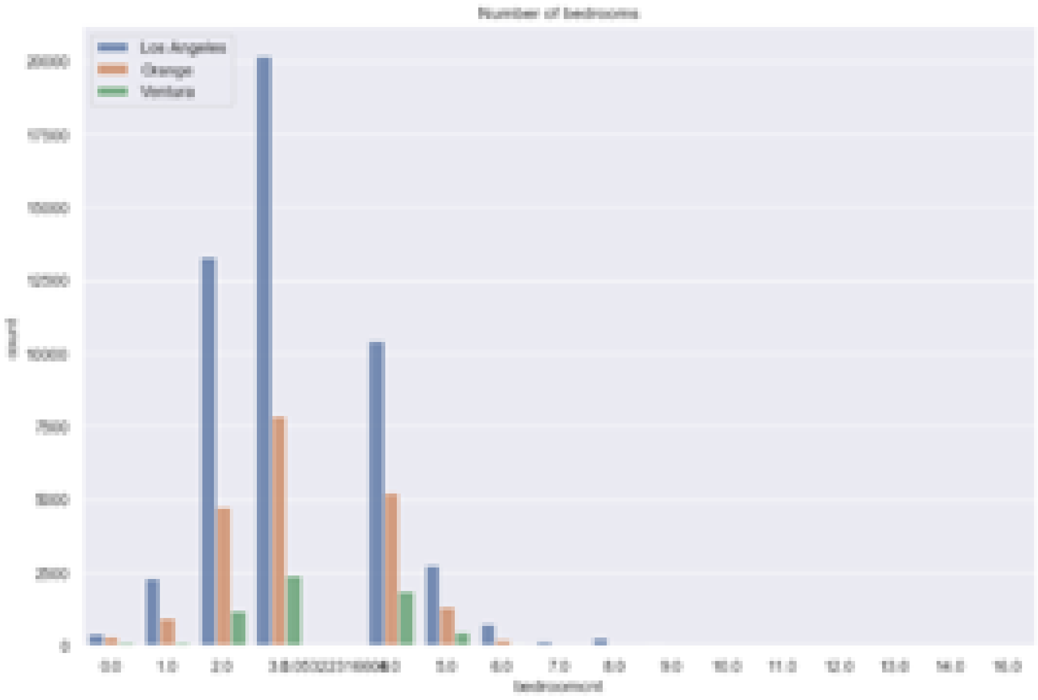


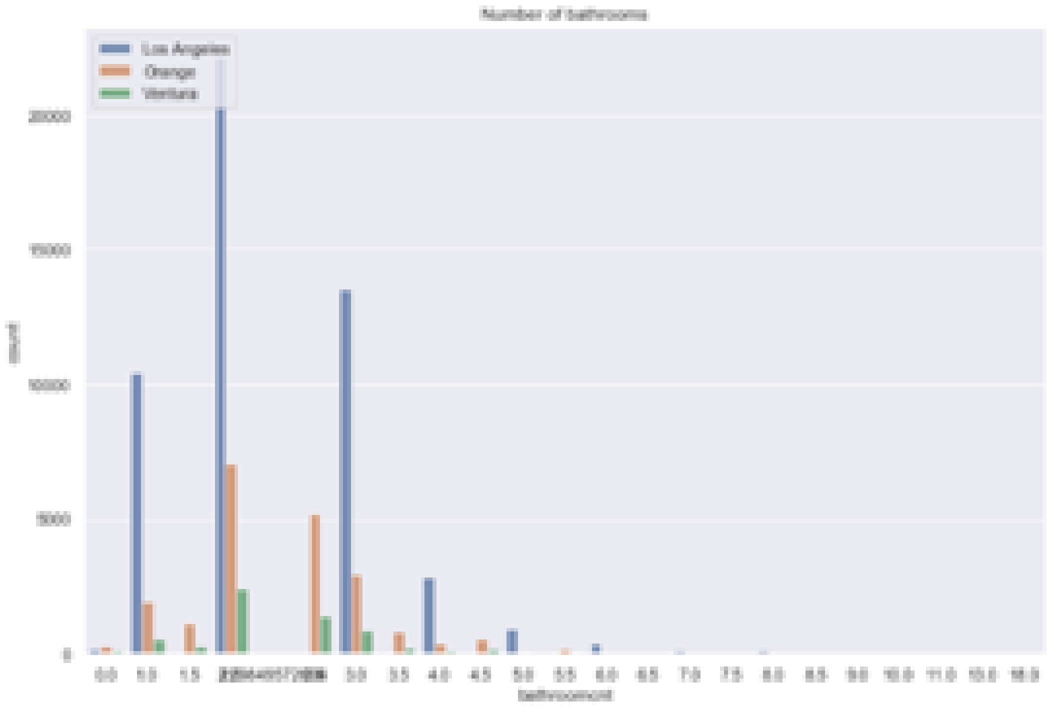
Number of house sale in 2017



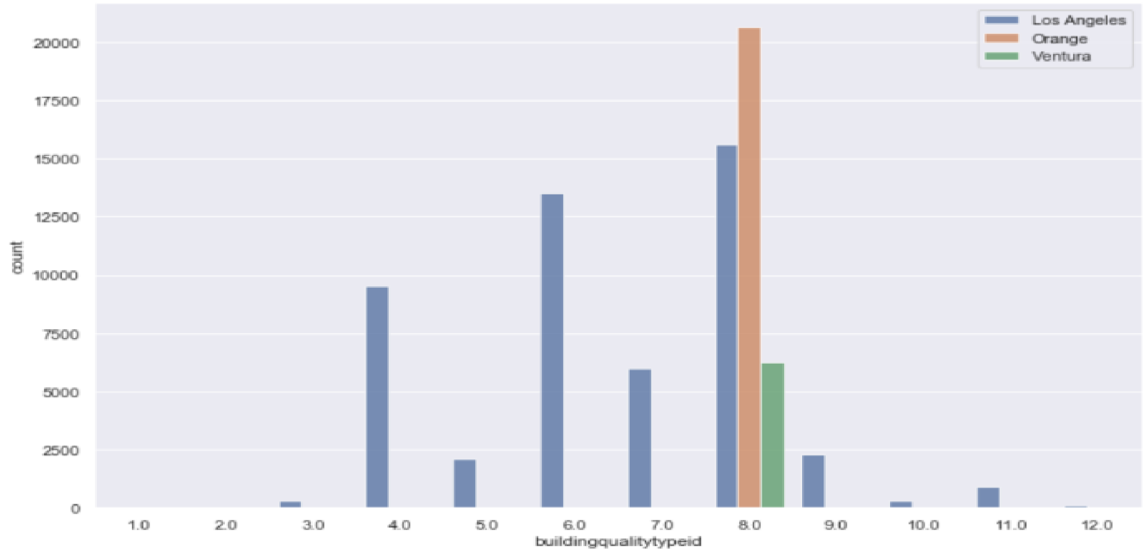
House Sales in California Counties

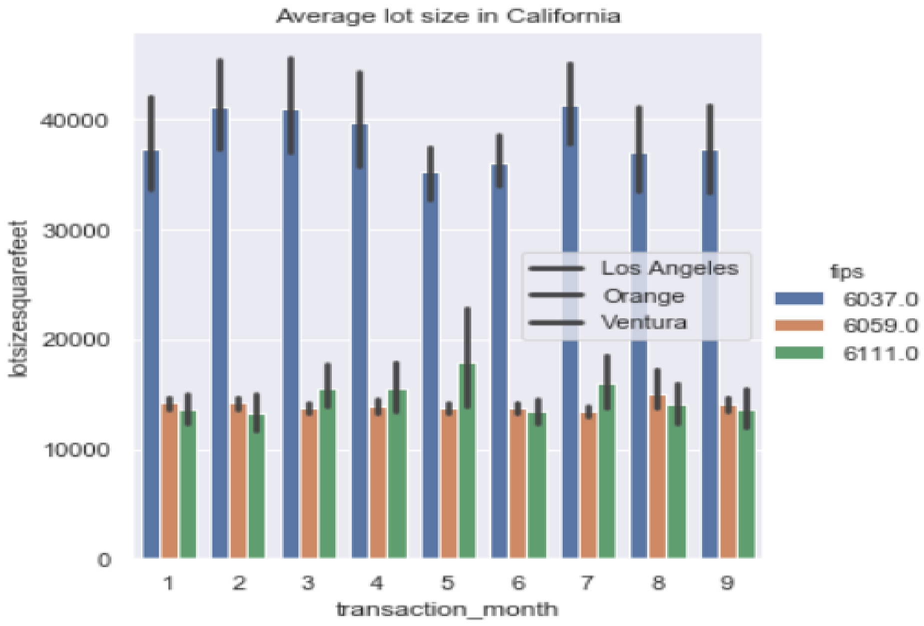
Exploratory Analysis

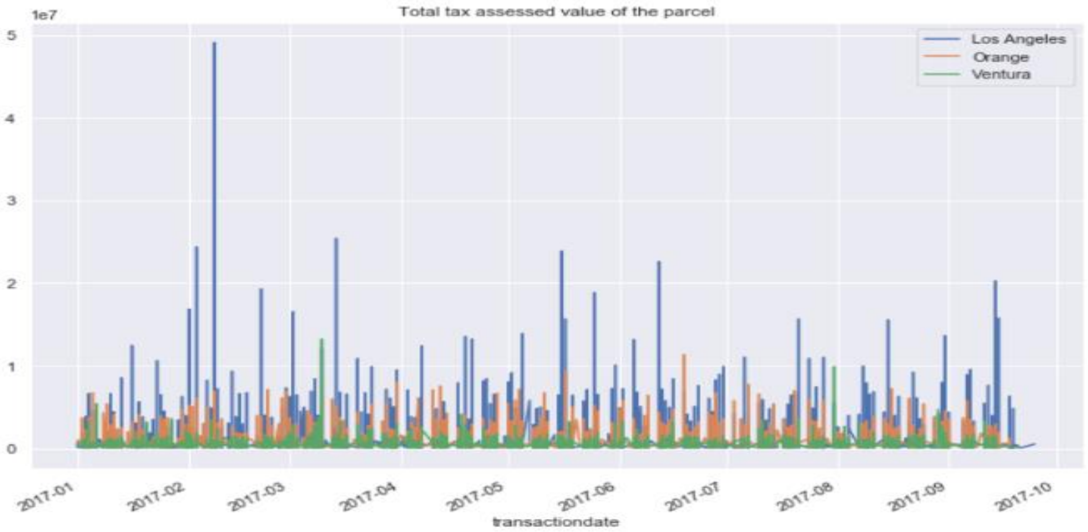


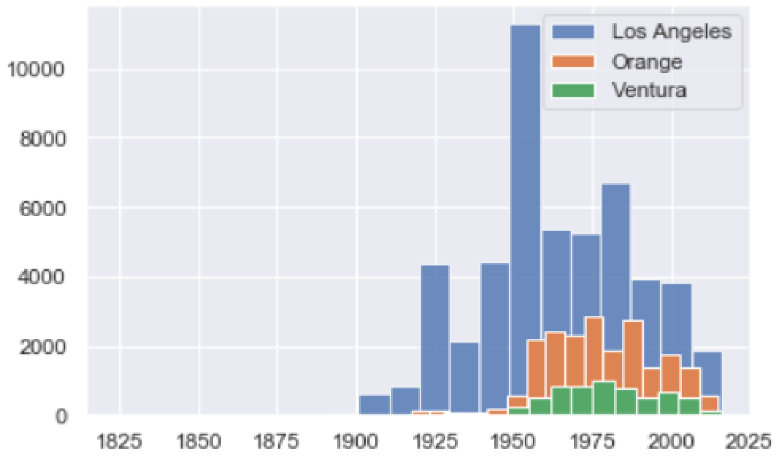


Building quality of house sale



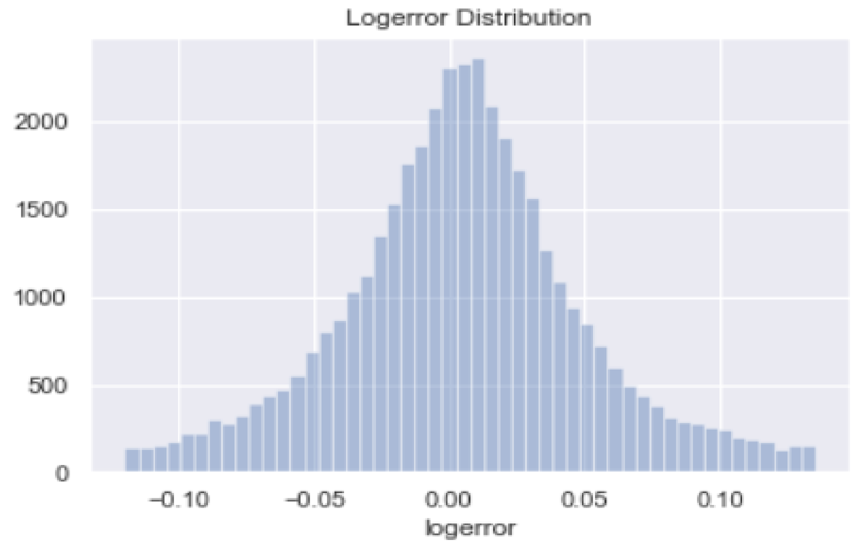






Log error

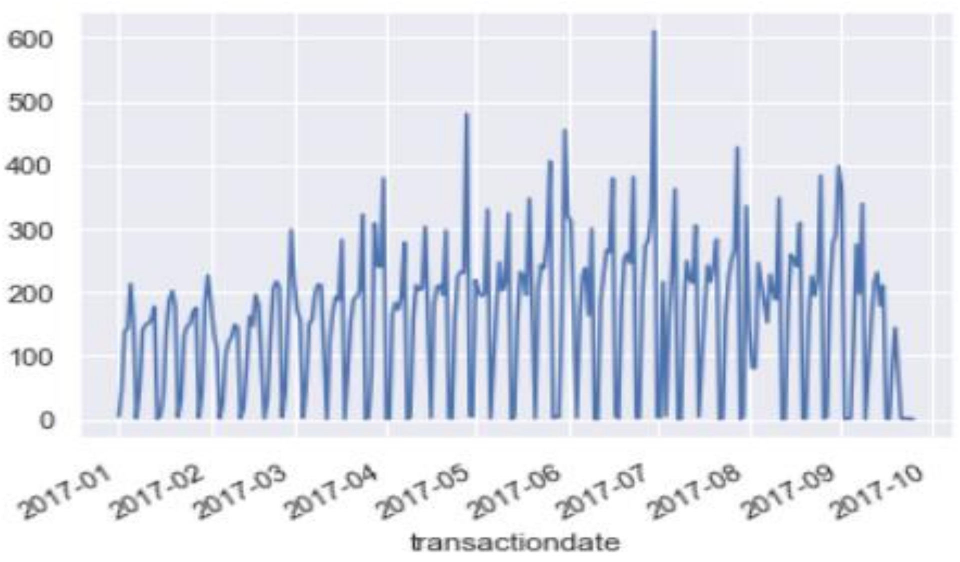

0.05
0.00
−0.05
−0.10
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
transactiondate


Logerror Distribution
2000
1500
1000
500
0
−0.10
−0.05
0.00
0.05
0.10
logerror

Statistical Analysis

How does the total tax assessed value of the parcel change over time?

Is there any trend with the year when the houses were built?



Data

	Los Angeles	**Orange**	**Ventura**
Training set	20657	7932	2894
Test set	5690	2066	798

bathroomcnt	bedroomcnt	calculatedfinishedsquarefeet	fips	lotsizesquarefeet	roomcnt	unitcnt	yearbuilt	structuretaxvaluedollarcnt	taxvaluedollarcnt
1.0	2	1465	6111	12647	5	1	1967.0	88000	464000
2.0	3	1243	6059	8432	6	1	1962.0	85289	564778
3.0	4	2376	6037	13038	0	1	1970.0	108918	145143
2.0	3	1492	6111	903	6	1	1982.0	198640	331064
1.0	2	738	6037	4214	0	1	1922.0	18890	218552

Evaluation Metrics

- Mean Squared Error(MSE)
- Root-Mean-Squared-Error (RMSE)
- Mean-Absolute-Error (MAE)
- R^2 or Coefficient of Determination

Evaluation Metrics

- **Mean Squared Error** is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model. As it squares the differences, it penalizes even a small error which leads to over-estimation of how bad the model is. It is preferred more than other metrics because it is differentiable and hence can be optimized better. It is always a non-negative number. Values closer to zero represent a smaller error.
- **Root Mean Square Error (RMSE)** is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors. This implies that RMSE is useful when large errors are undesired.
- **Mean Absolute Error:** MAE is the absolute difference between the target value and the value predicted by the model. The MAE is more robust to outliers and does not penalize the errors as extremely as MSE. MAE is a linear score which means all the individual differences are weighted equally. It is not suitable for applications where you want to pay more attention to the outliers.
- **R^2 Error:** Coefficient of Determination or R^2 is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R squared value is always between 0 and 1, and that the best value is 1.0.

Evaluation Metrics

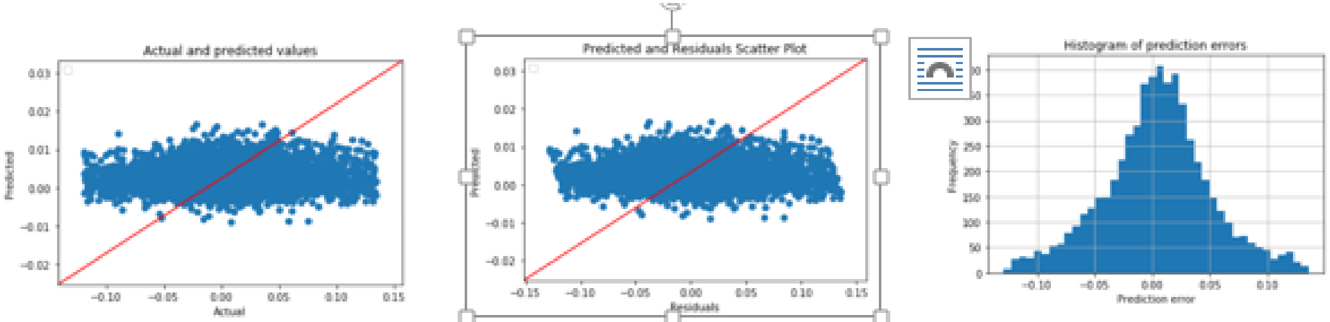
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

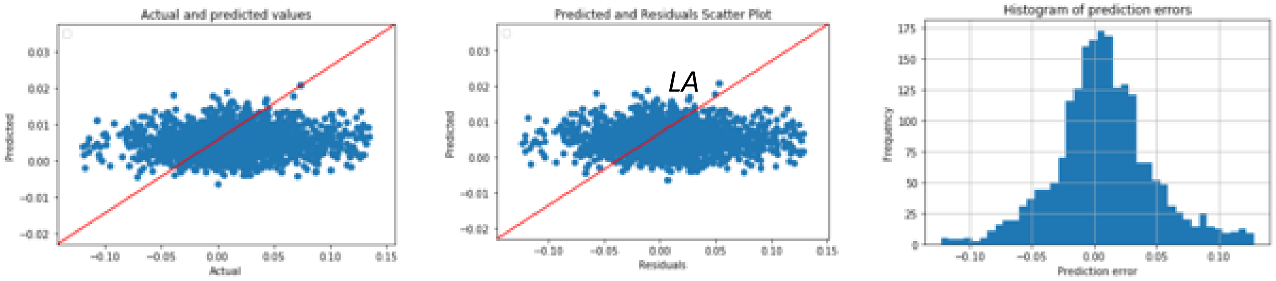
$$MAE = \overset{\text{Divide by the total number of data points}}{\boxed{\frac{1}{n}}} \underset{\text{Sum of}}{\sum} \underbrace{\left| \overset{\text{Actual output value}}{\boxed{y}} - \overset{\text{Predicted output value}}{\boxed{\widehat{y}}} \right|}_{\text{The absolute value of the residual}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

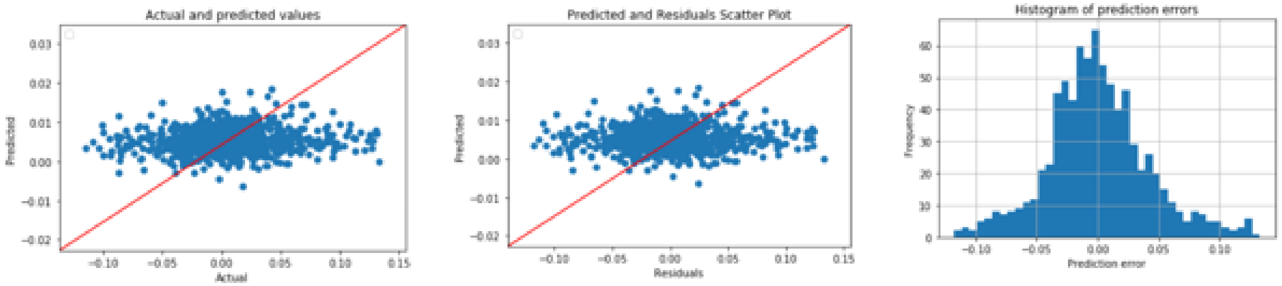
$$\mathbf{r} = \frac{n(\Sigma\, xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Linear Regression





Orange County

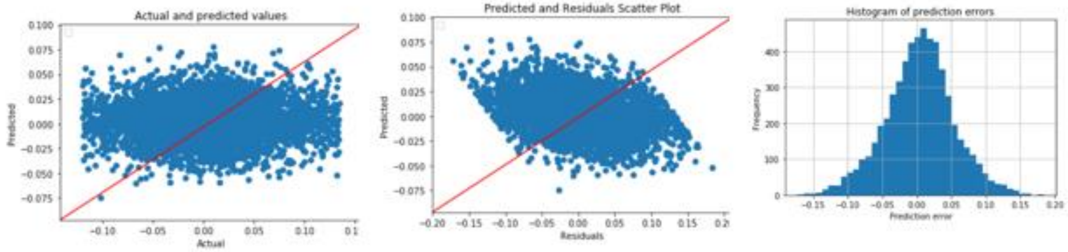


Ventura County

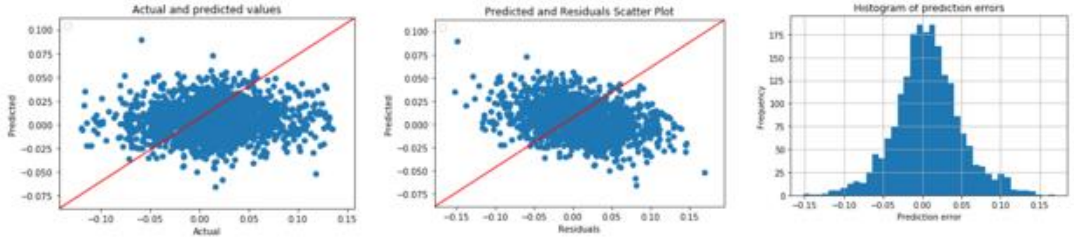
Linear Regression

	R2	MSE	RMSE	MAE	MAPE
LA	-0.0028	0.002	0.04	0.0356	48
Orange	-0.0278	0.002	0.04	0.0298	7.2
Ventura	0.0079	0.002	0.04	0.0312	-6.5

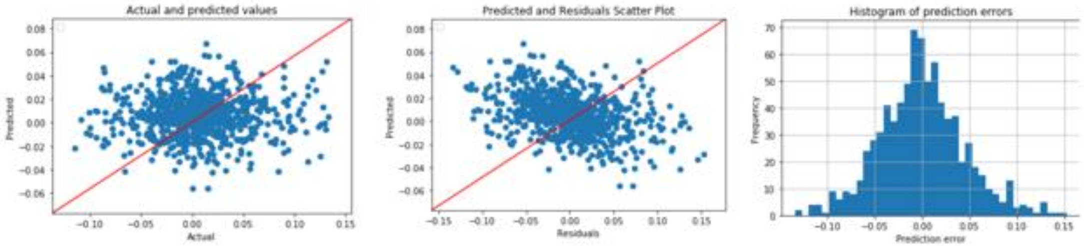
Random Forest



Los Angeles



Orange



Ventura

Random Forest

	R2	MSE	RMSE	MAE	MAPE
LA	- 0.15	0.002	0.04	0.0387	178
Orange	-0.15	0.002	0.04	0.0321	92
Ventura	-0.16	0.002	0.04	0.0344	96

Random Forest Results

Hyperparameter Tuning

- n_estimators = number of trees in the foreset
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node
- bootstrap = method for sampling data points (with or without replacement)

Best hyperparameters:

- {'n_estimators': 400, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}

Results

- mape = 100 * np.mean(errors / test_labels)
- Accuracy = 100 - MAPE

		Accuracy
Linear Regression	LA	51%
Random Forest		-78%
Hyper Parameter		53%
Linear Regression	Orange	92%
Random Forest		7%
Hyper Parameter		76%
Linear Regression	Ventura	106%
Random Forest		3%
Hyper Parameter		95%

Conclusion

- The model we have developed in this report have automated some level of human manual analysis. With sufficient data, the model will be easily to automate higher level analysis and reduce a lot of time and efforts for real estate experts.
- There are still limitations in the model in regards to the terms of timeline, percentage of growth, time decay, and seasonal factors therefore the model results are not very good at prediction of house price log error.
- There are several approach that consider time series forecasting like ARIMA model