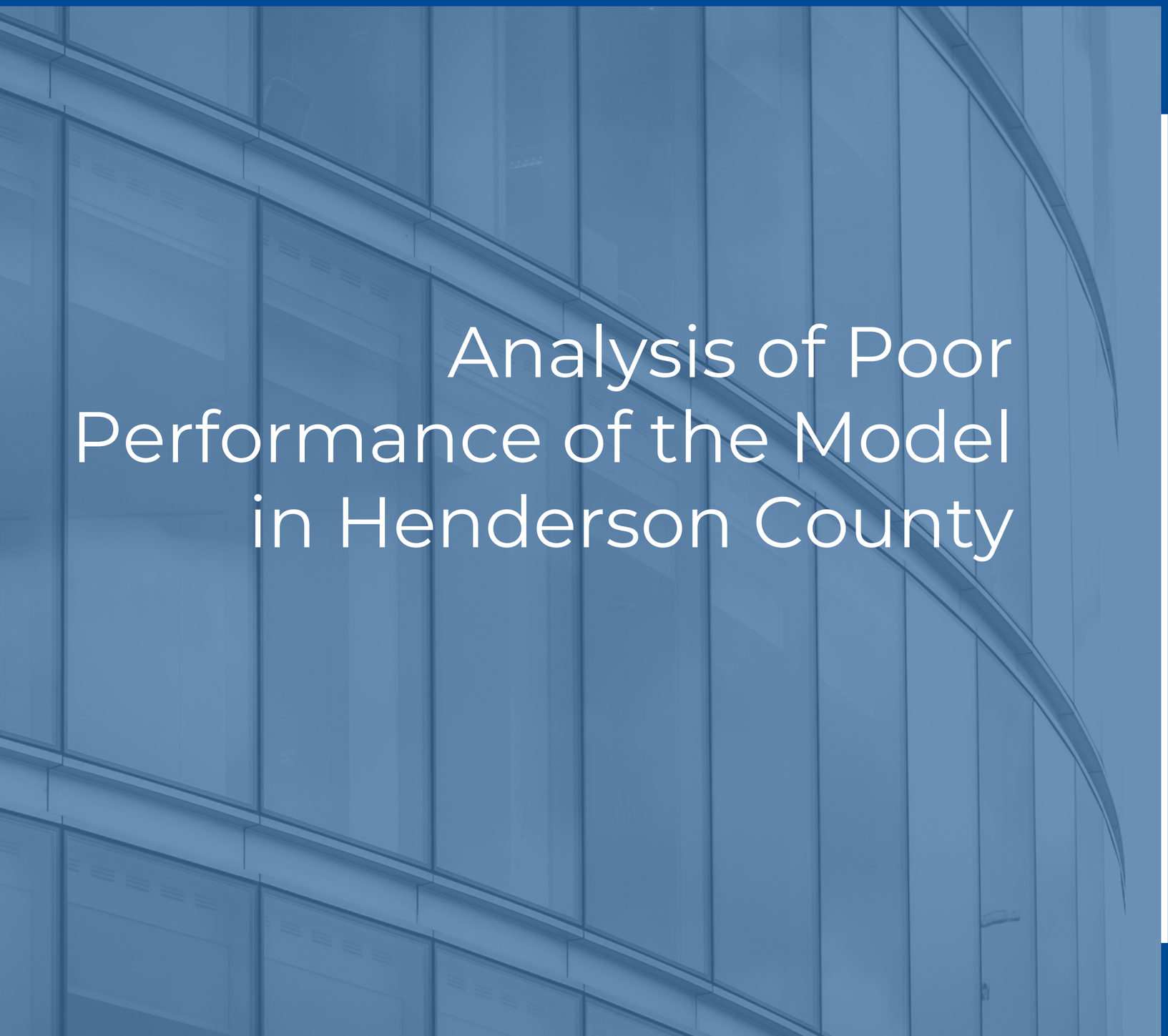


Analysis of Poor Performance of the Model in Henderson County

MAY 8TH 2019

Windfall Data Analyst Take-Home Submission

PRESENTED BY:
ANASTASIA SIMPSON

DATA

home_values_texas.csv

Contains information on properties in selected counties

county_model_performance.csv

Contains Model Performance Measurements

Problem

Poor
Property Valuation
Model Performance
in Henderson
County

0.49
MAE

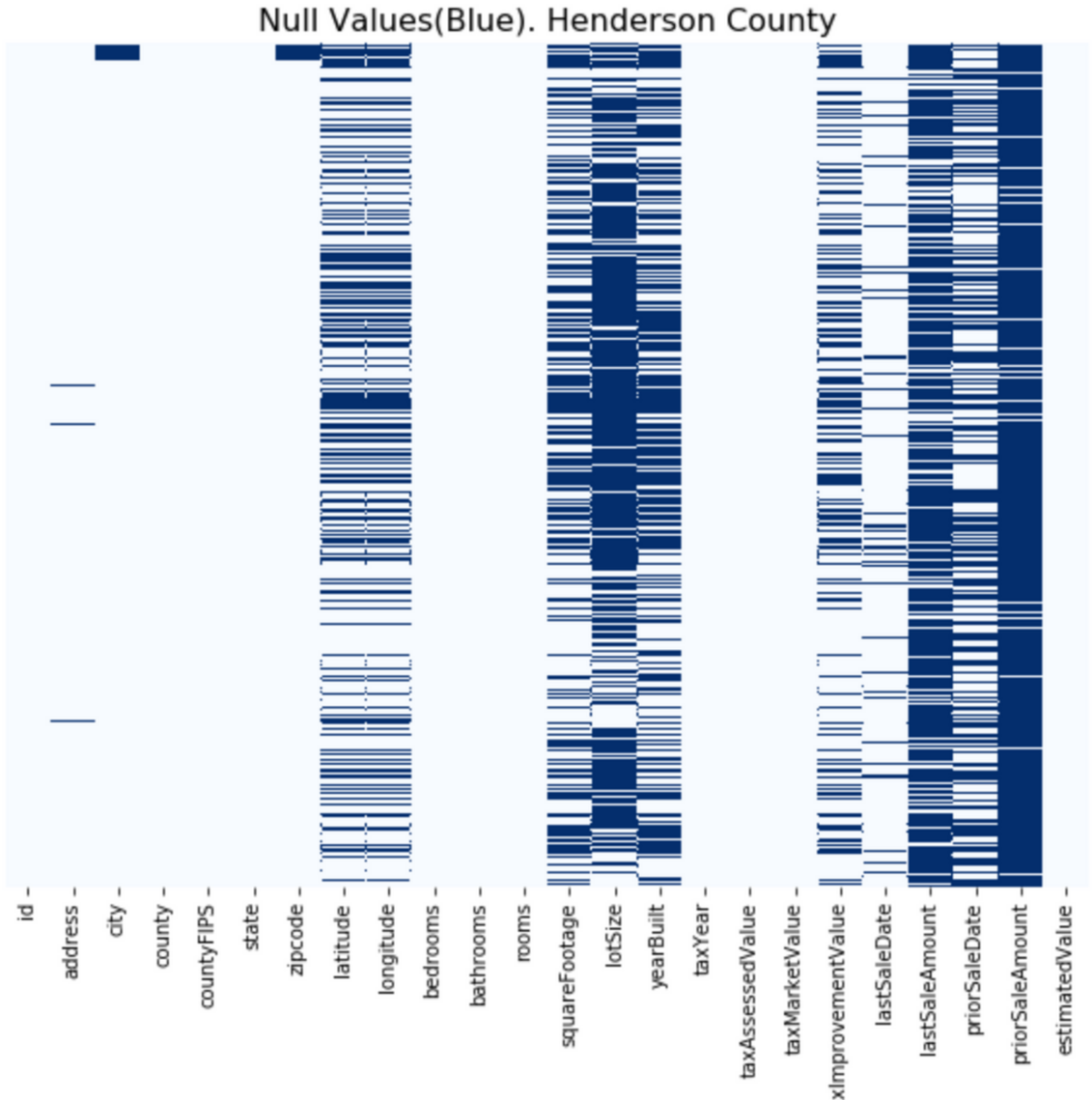
Observed Data Issues

- Data Accuracy
- Data Completeness
- Outliers
- Multicollinearity of Predictors

Data Accuracy

	rooms	bedrooms	bathrooms	lastSaleAmount	priorSaleAmount
count	53920.0	53920.0	53920.000000	12058.000000	2727.000000
mean	0.0	0.0	1.188807	3804.653840	12384.719839
std	0.0	0.0	2.179751	25175.840323	39009.188896
min	0.0	0.0	0.000000	0.000000	0.000000
25%	0.0	0.0	0.000000	0.000000	0.000000
50%	0.0	0.0	0.000000	0.000000	0.000000
75%	0.0	0.0	1.500000	0.000000	0.000000
max	0.0	0.0	13.000000	925000.000000	524885.000000

Data Completeness



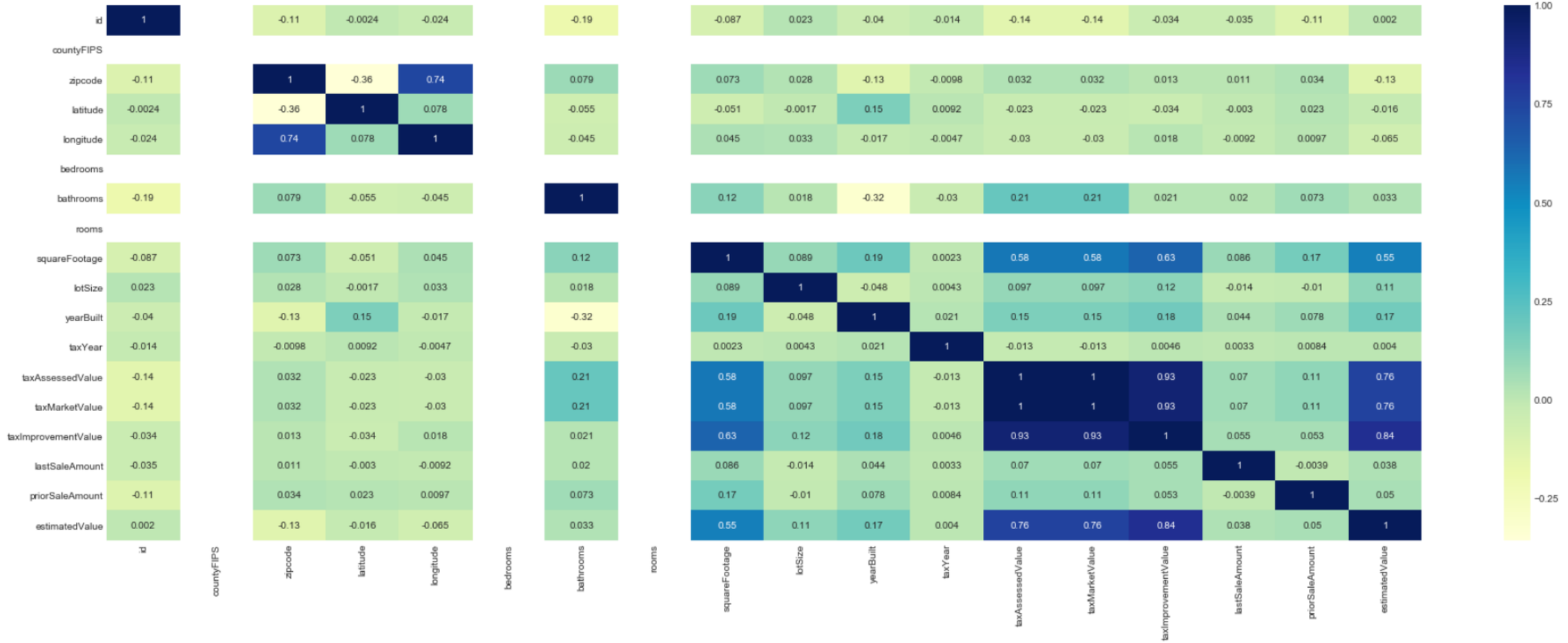
RECOMMENDATIONS

DATA ACCURACY & COMPLETENESS

- Replace zero values with NaN
- Deploy the model w/ adjustment
- Remove features with incorrect values from the model
- Collect more data


Detecting Outliers
Target Variable
Henderson County
Outlier
Median Estimated Price
Q75 + (IQR*1.5)
Mean + 3*std
Henderson County
1
145000
374000
576777
1000000
2000000
4000000
5000000
5466093
Estimated Price

Multicollinearity of Predictors



Recommendations

Outliers:

- Consider different outlier treatments
- Try models that are robust to outliers (i.e. regression tree model))
- Remove Extreme Outliers
- Deploy an additional model for outlier houses

Multicollinearity:

- Feature engineering (i.e. AVG of tax features)

POSSIBLE ADDITIONAL FEATURES

- Location Specific
- Luxury Housing
- Number of Yelp Reviews
- Building Energy Consumption

Thank You