FLORIDA HOUSING PROJECT

Kyle Morris

DSC 630

May
2020

INTRODUCTION

On Analysis of the Housing Market in Florida

- Where our Data came from
- How we validated and cleaned our Data.
- How the most important features were selected.
- How the model was created.
- How the parameters of our model looked.
- What challenges arose?
- What did our final model look like?
- How did the model compare to the actual data?
- What can we conclude?

PROBLEM

Given the attributes of a home, can we predict the fair market value of the residence?

- The information came from the Census bureau directly.
- While data is available for all 50 states, the District of Columbia, and Puerto Rico, this model is focused on the Florida portion.
- The 2018 data is utilized. While we can adjust for inflation and incorporate multiple year's worth of data in the analysis, only the most recent data will be used.

DATA SOURCE

2018 American Community Survey by the US Census Bureau

DATA CLEANING

3 categories of variables were identified for the purposes of missing numbers:

0 As Valid Value

- ['BDSP', 'HUGCL', 'NPP', 'NR', 'NRC', 'PARTNER', 'PSF', 'R18', 'R60', 'R65', 'SRNT', 'SSMC', 'SVAL']
- -1 was used for missing values since 0 was a valid response.
- These columns were categorical variables coded as numbers and so the mean would not make sense.

Mean Appropriate

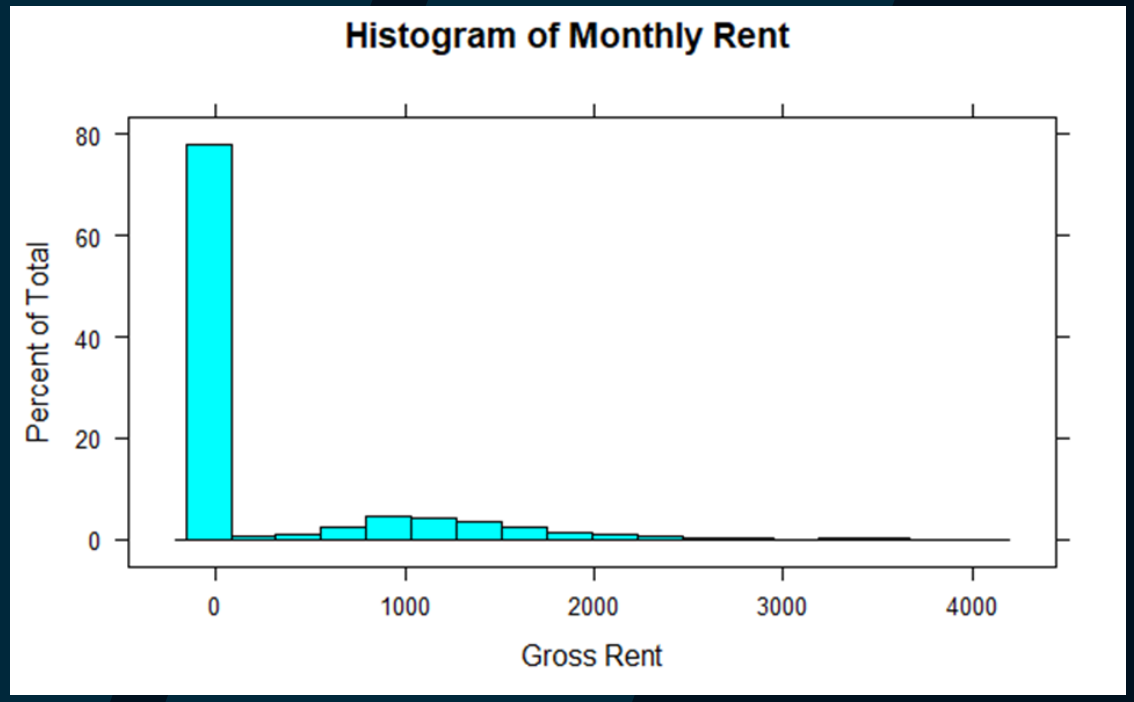
- 'FINCP', 'OCPIP', 'SMOCP']
- Since these were actual values (such as monthly rent, average electricity, etc.) and 0 would not be appropriate, the mean was used for these.
- By using the mean, we would avoid creating outliers that would impact the analysis.

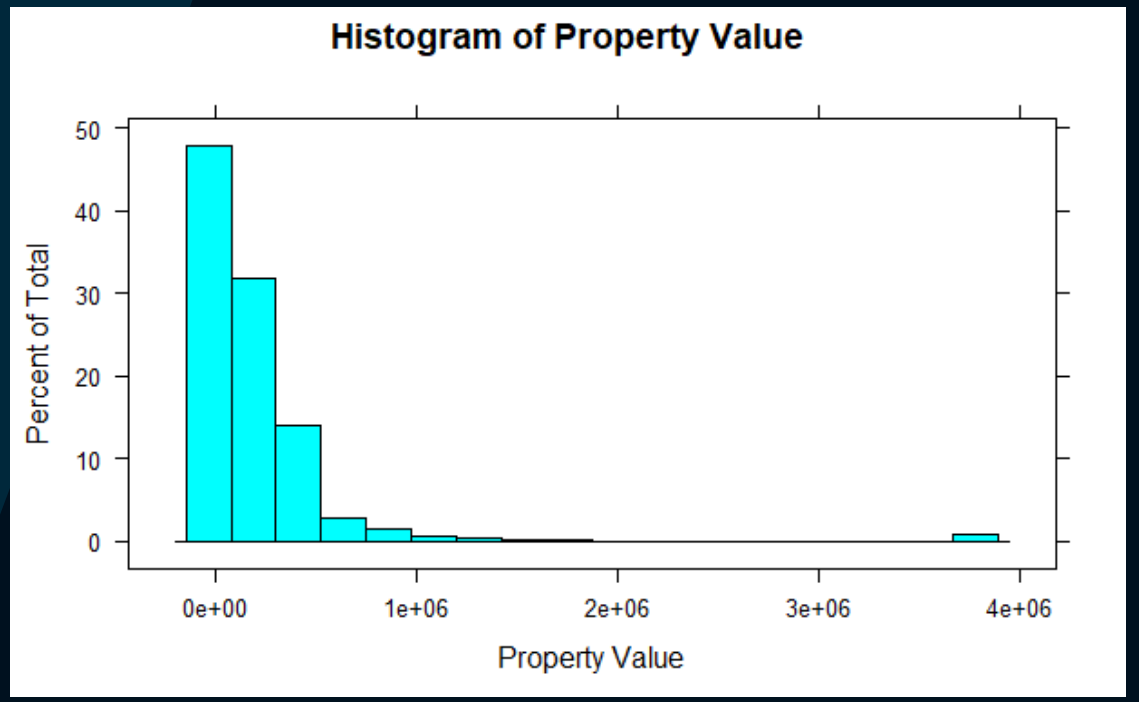
All Others

- 0 was an appropriate value for missing values and so it was used.

INITIAL GRAPHS

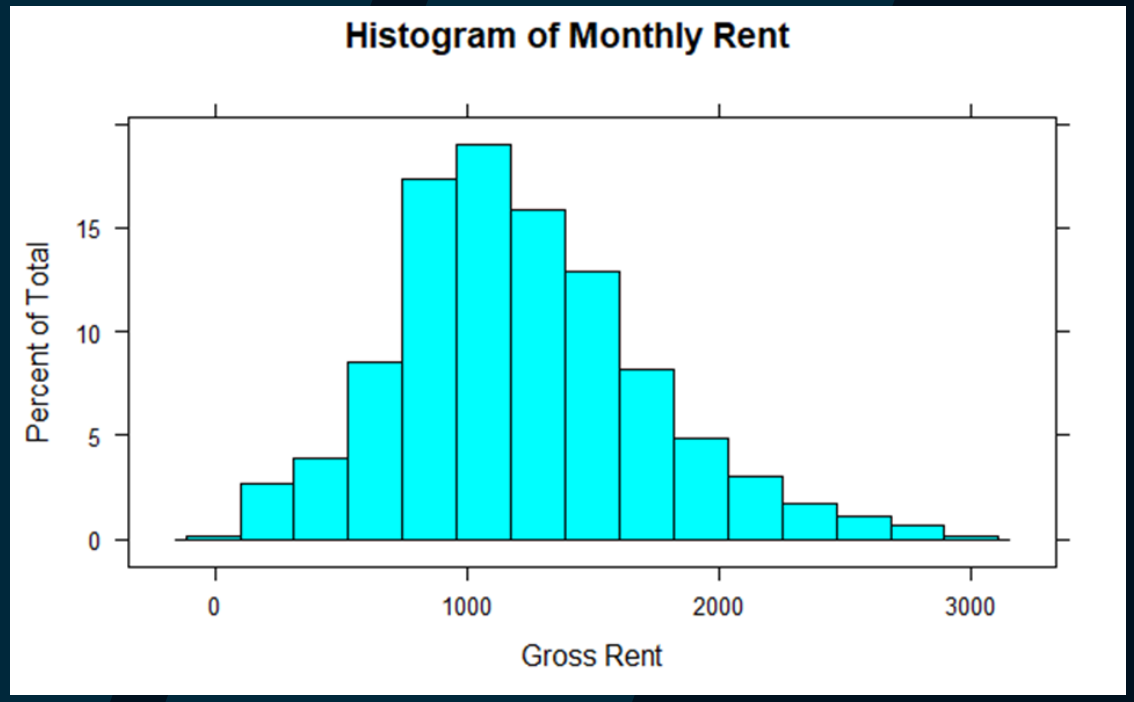
Gross Rent had a large spike at 0, while Property Value was positively skewed.

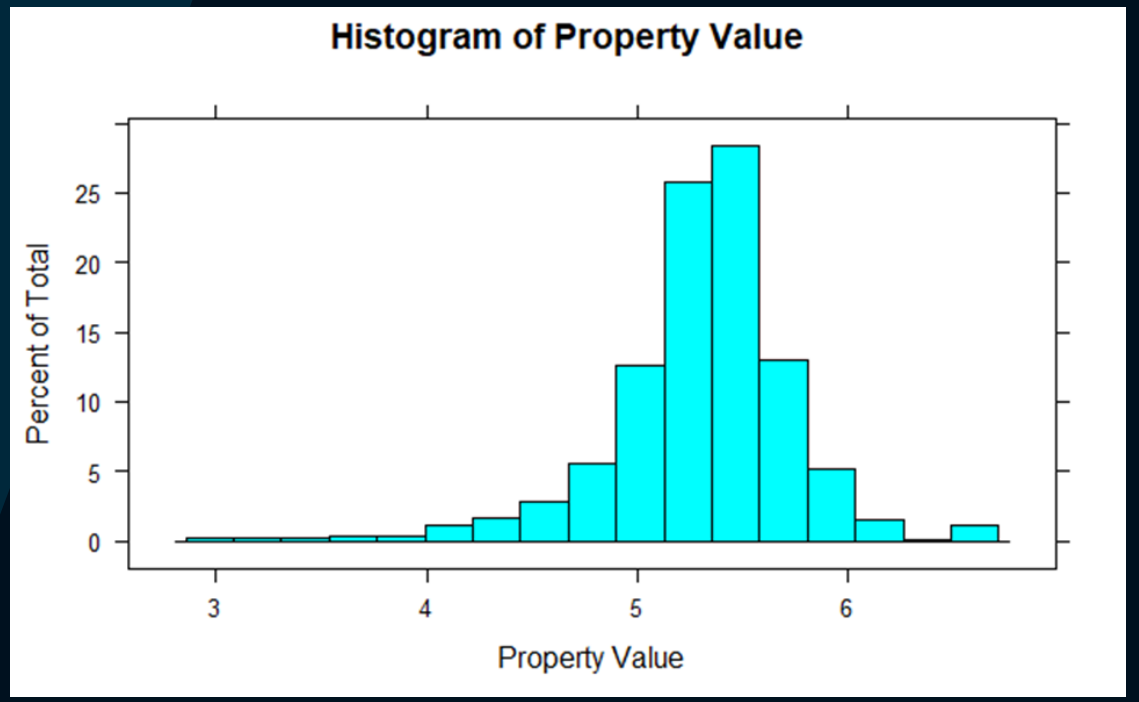




POST CLEANUP

After the changes, the data was much more normally distributed.

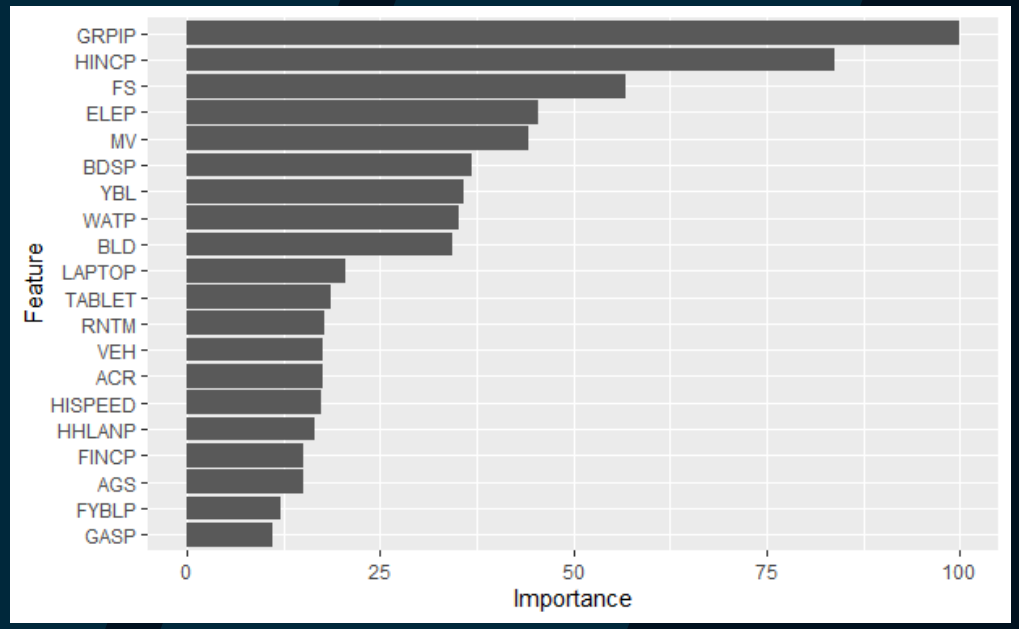




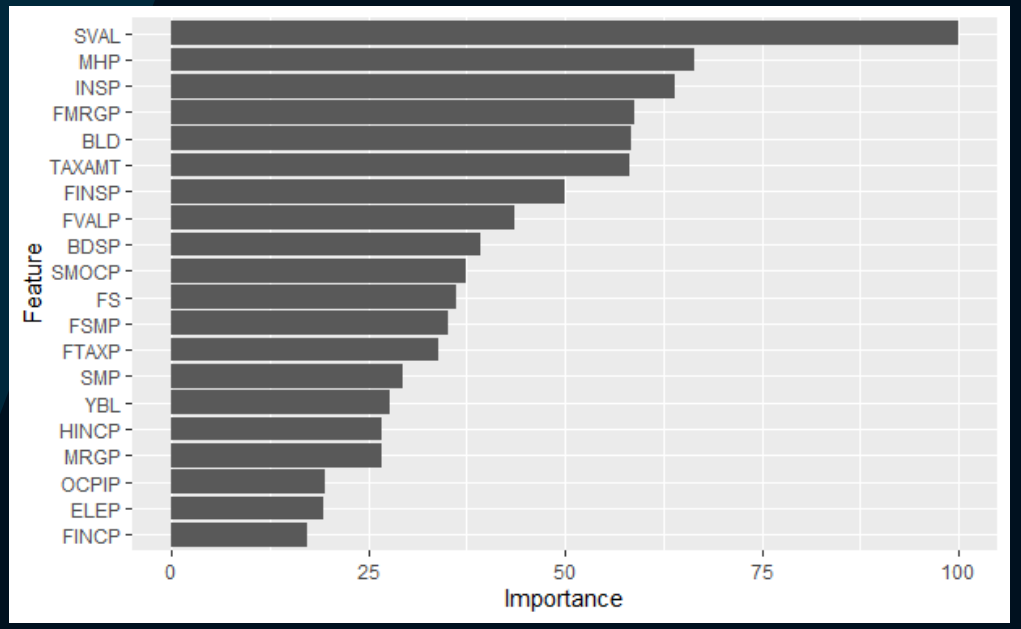
FEATURE SELECTION

- Varimp was used to classify the important features.
- Only the ones most important to determining the value will be used in the model.

Monthly Rent



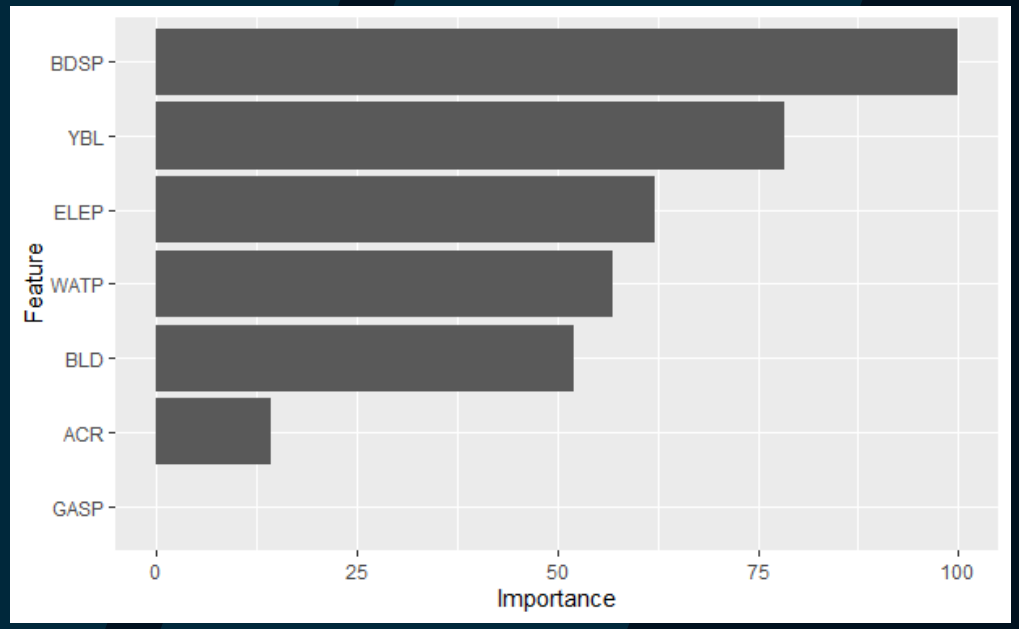
Property Value



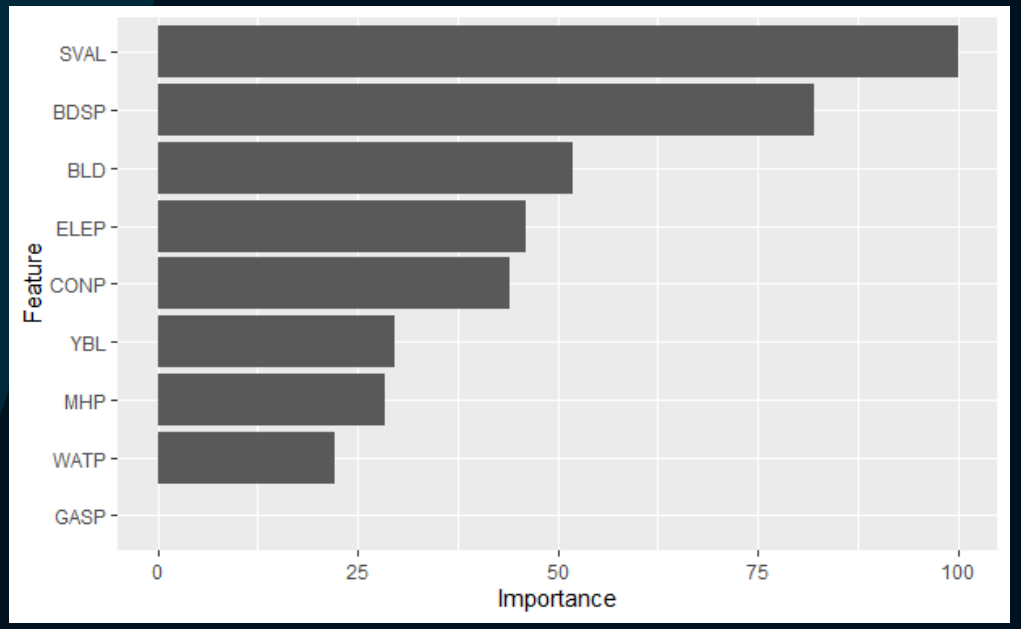
FEATURE SELECTION FIXED

- Some variables had no effect or were the dependent variable under a different format.
- These are the previous charts, but only showing the columns deemed most important.

Monthly Rent



Property Value



MODEL CREATION

Two models were created, one for predicting Monthly Rent, the other for predicting Property Value.

- Linear Regression was used for both models.
- Linear Regression is useful due to its linearity and the results are easy to interpret – each coefficient represents the weight that that column has on the final results.

MODEL COEFFICIENTS

Final coefficient outputs from the model. Each coefficient shows the weight of that column in the final value.

Property Value Model									
Intercept	SVAL	BLD	MHP	BDSP	YBL	CONP	ELEP	WATP	GASP
5.30326487	0.149084	0.102204	-0.04792	0.10835	0.042578	0.080054	0.063853	0.036906	0.015506

Rent Model							
Intercept	BDSP	ELEP	YBL	WATP	BLD	ACR	GASP
1216.251757	150.4366	92.09602	98.6556	77.07377	125.7296	45.76685	9.230444

CHALLENGES

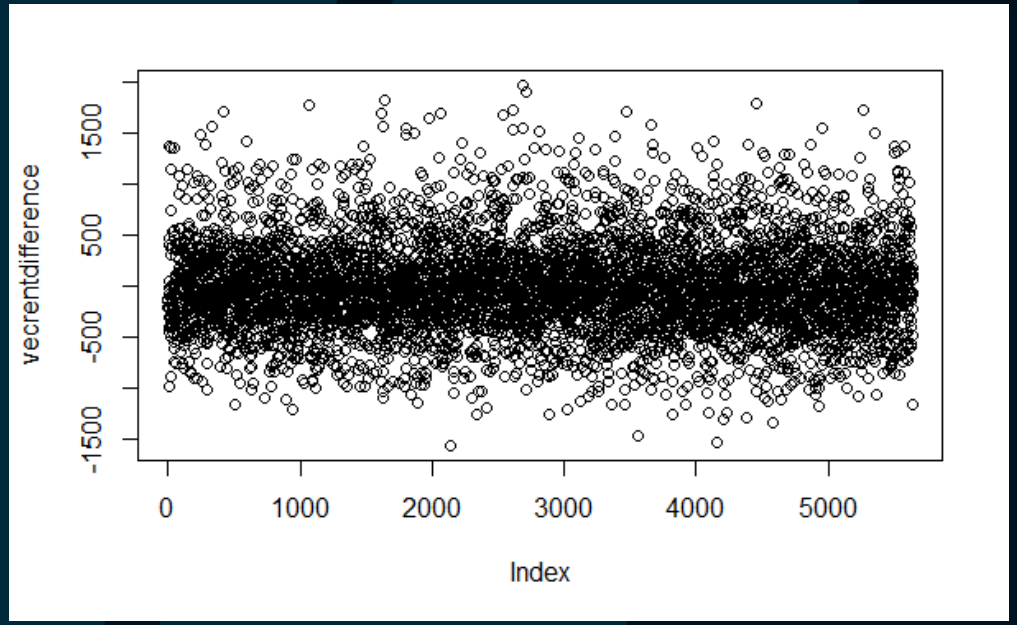
What was done to optimize the model?

- Missing data and how to handle it was a huge factor, given that the data came from self-submitted surveys.
- Some of the data wasn't normalized, which was a requirement for analysis and so transformation techniques were applied.
- Data came in one set, and in order to assess the power of a model we need to be able to compare the predicted output to the actual values, so some records were withheld as a testing set.

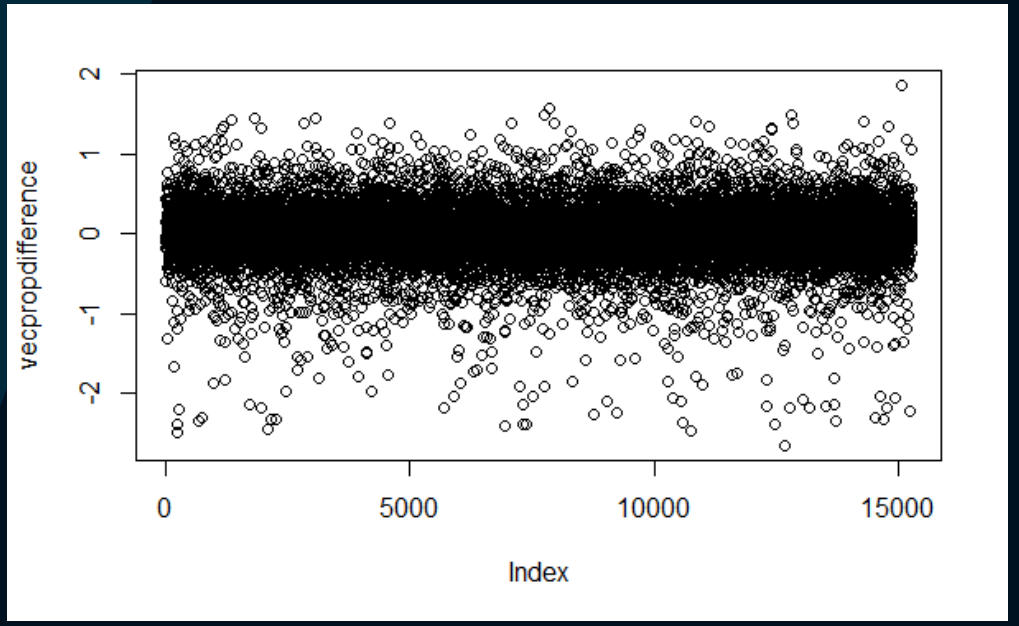
RESULTS

- These graphs show the spread of the actual costs versus the predicted cost.
- The majority of predictions are within ±500 or ±1.

Monthly Rent



Property Value



CONCLUSIONS

Useful tool to classify a property as overvalued, undervalued, or just right.

- The vast majority of predictions fell in line with what we expected.
- Data can easily be adapted for other states.
- Model can be refined over time – when 2019 data comes out, can easily be run again to adapt to changes in the housing market over time.
- Care must be taken not to use the model as gospel.

THANK YOU

Kyle Morris

Email

kymorris@my365.Bellevue.edu