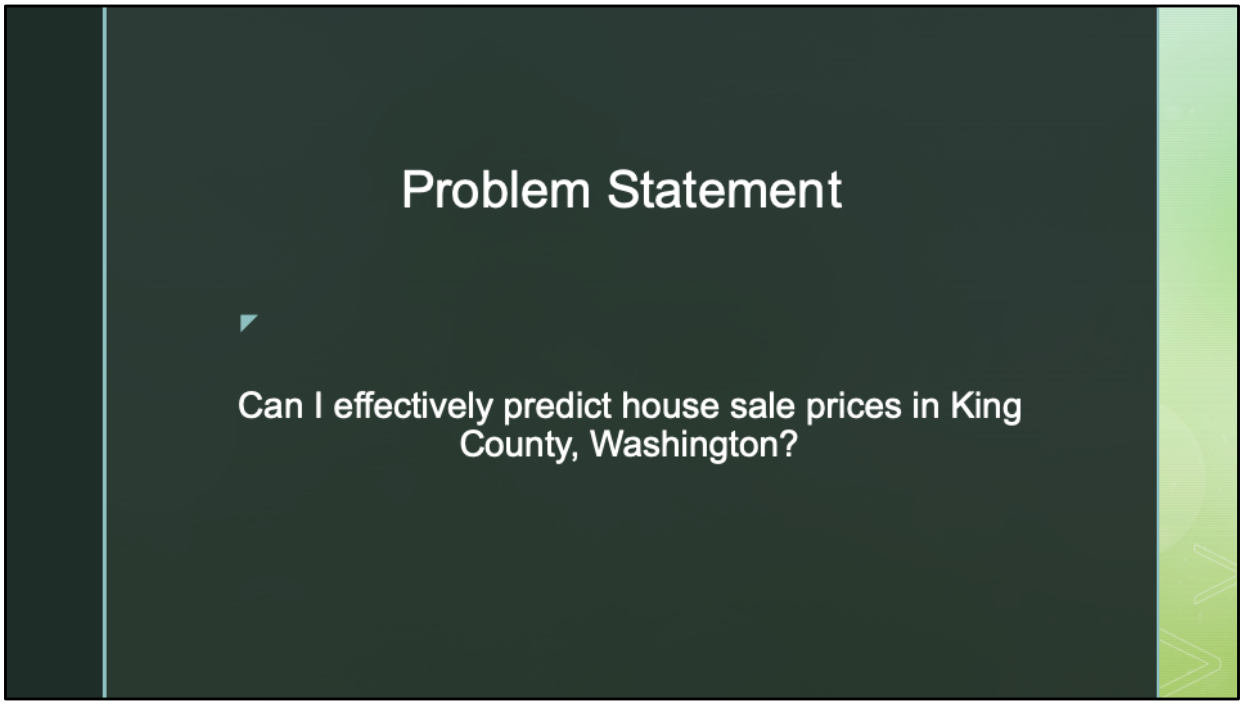
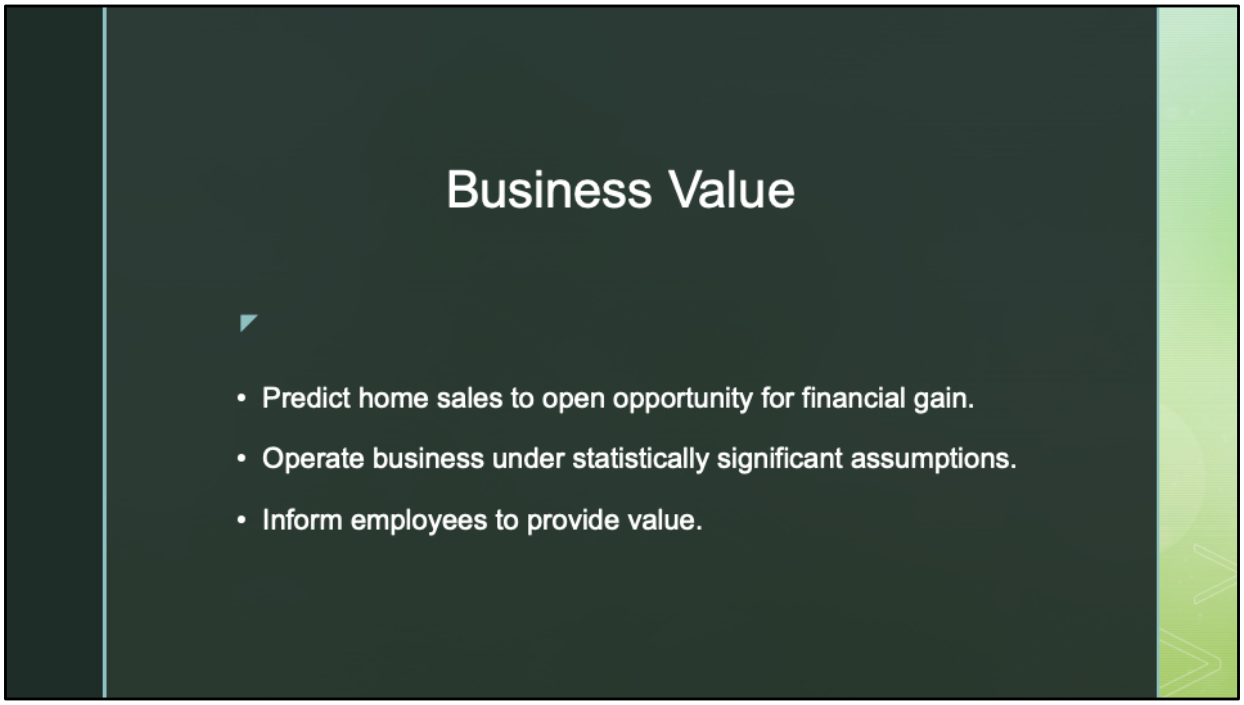
Module 2
Final Project

Michael Mathews Jr

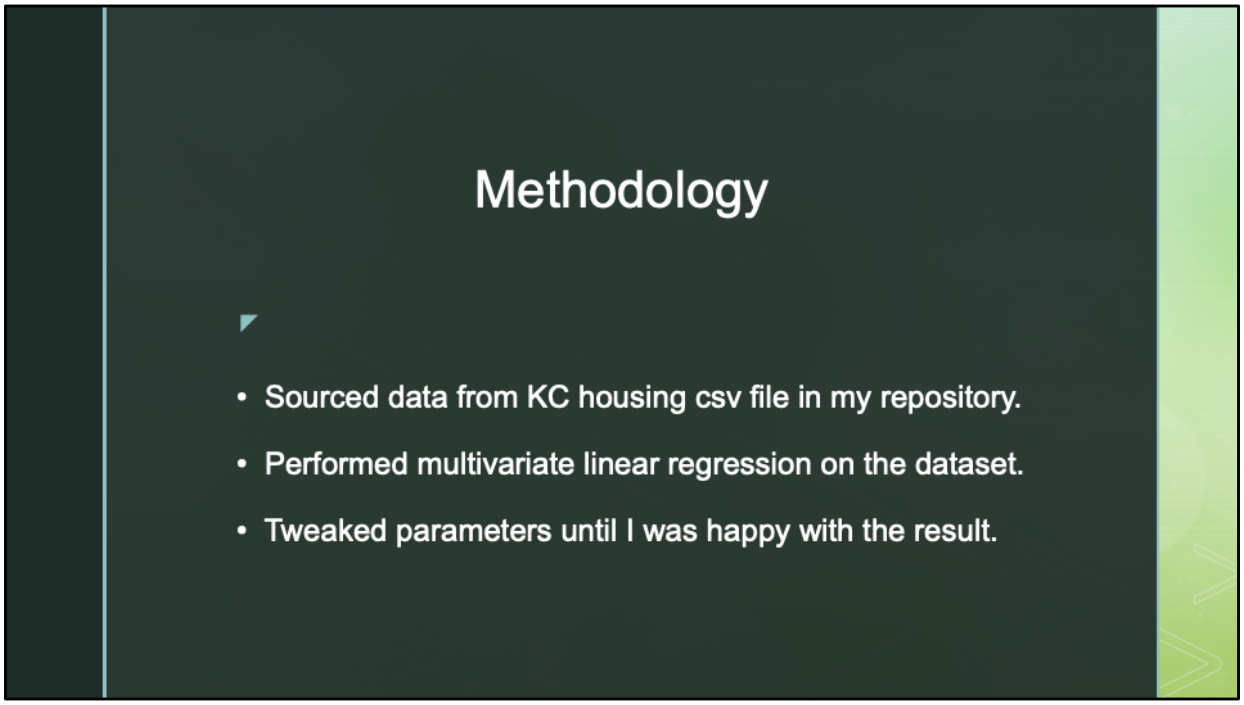
Online-DS-PT-120919



Can I use multivariate linear regression to predict house sale prices in the given dataset? How can I lower my residuals to an acceptable level in order to push my model to production?



- The accuracy of the model should allow the business to leverage the information for financial gain.
- Making business decisions using advanced statistics leaves little room for gut feelings and opinions. Get a statistical advantage over the competition by utilizing a machine learning model.
- Make the whole business aware of the information and why it is significant. Allow employees to incorporate the information into their day to day.



- Please check my repo for the dataset used in this analysis.
- I used multivariate linear regression to build the model. This model has its strengths and weaknesses. It does not handle outliers well so I removed them. It is sensitive to feature transformation so I spent a lot of time testing different combinations of transformation and scaling.
- I tried to tweak every variable in my model to achieve the best adjusted R2 score.

Top 12 feature contributions in my model.

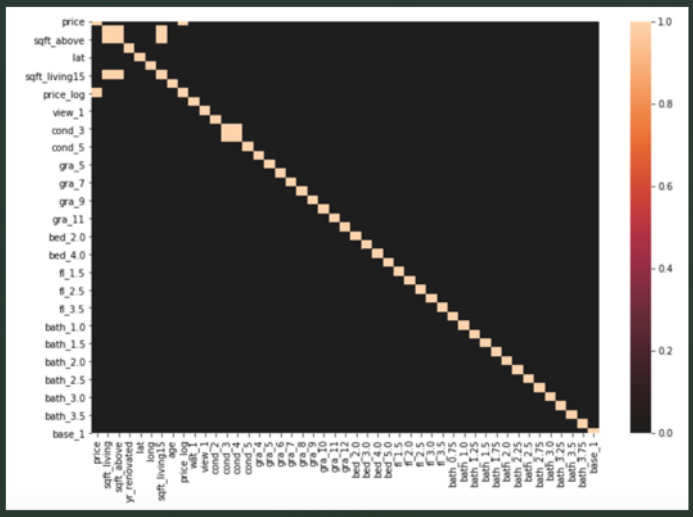
Bathroom dummy variables dominate feature contribution in my model.

Predictor	Coefficient(log)	Coefficient	Contribution(%)
lat	0.453064	2.838336	5.349553
sqft_living15	0.188779	1.544468	2.910936
bath_3.75	0.158404	1.440137	2.714297
sqft_living	0.145804	1.398955	2.636680
bath_3.5	0.144195	1.393784	2.626933
wat_1	0.135987	1.367688	2.577749
bath_3.25	0.134878	1.364199	2.571173
bath_1.25	0.129517	1.347464	2.539633
bath_3.0	0.126681	1.338694	2.523102
bath_2.75	0.124645	1.332432	2.511301
bath_2.5	0.117988	1.312162	2.473097
bath_2.0	0.117711	1.311326	2.471521

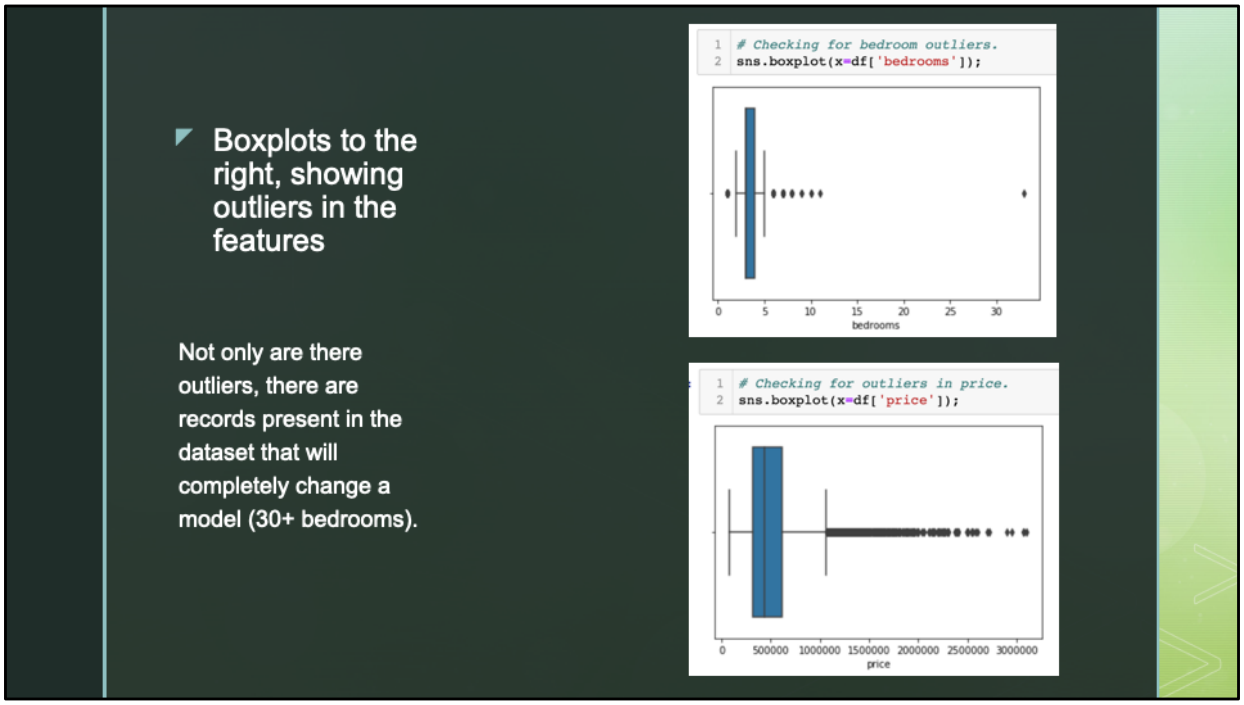
There are a lot of dummy variables. These include bathrooms, bedrooms, grade, floor, condition and more. There is a lot of opportunity for the dummy variables to contribute to the model, yet only the bathroom dummies seem to rank high.

Heatmap describing feature correlations

The yellowish squares show a correlation of absolute value > 0.70.



There are not a lot of highly correlated features. Although a few features are correlated with each other, it didn't seem to impact the model. In fact, the more features included, correlated or not, raised the adjusted R2 score of the model.

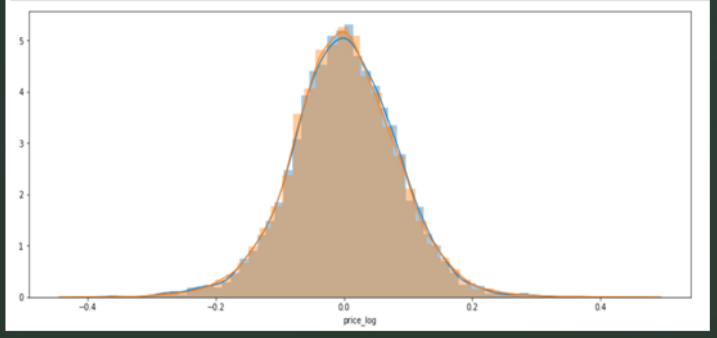


Handling outliers can be a tricky situation because I believe more data to train a model the better. Another reason to watch how many records you drop is the outliers aren't necessarily bad data, and assuming this, these events really did happen. In my model I played it safe and removed more than enough outliers.

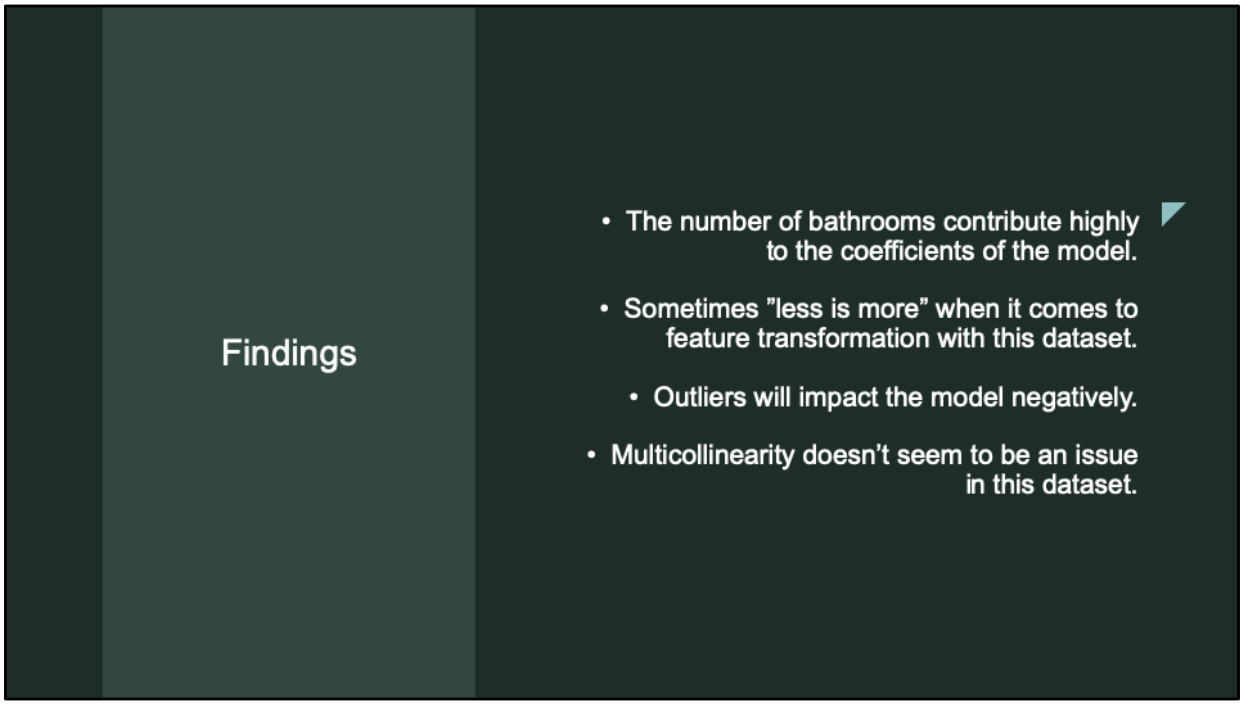
- Histogram plot showing distribution of residuals

The residuals between the train and test targets seem to be normally distributed, which is what you want.

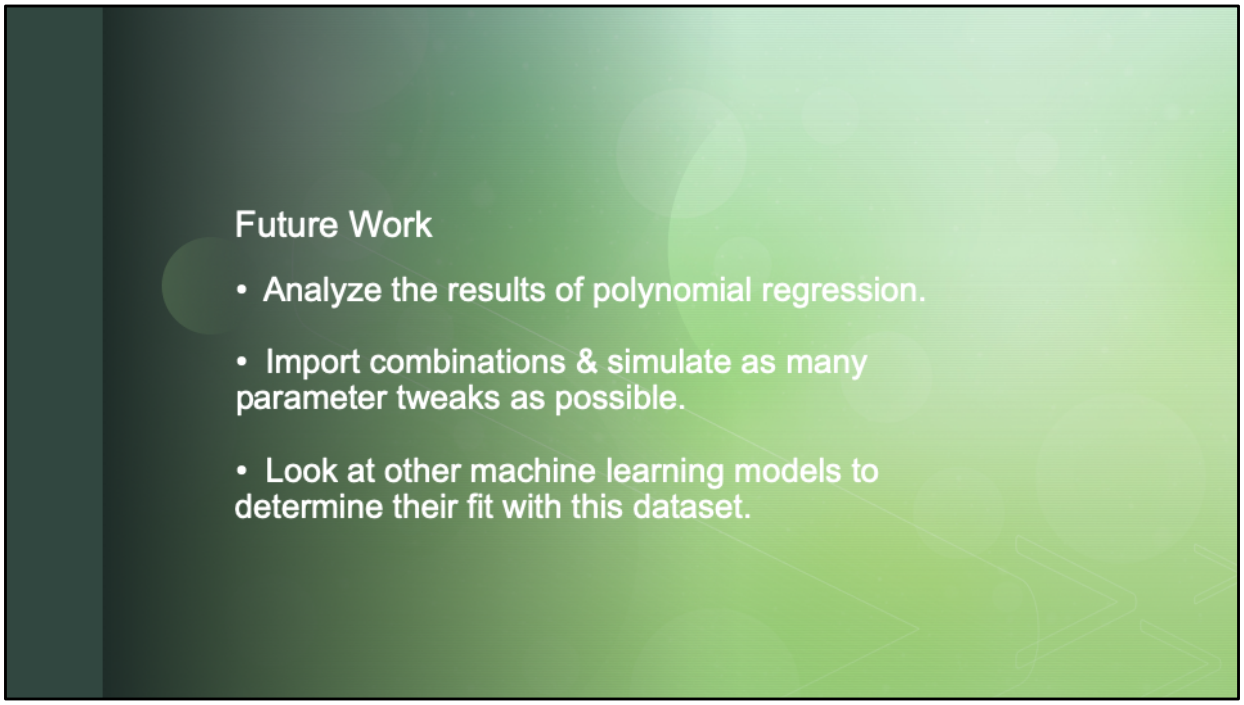
```
# Distribution of residuals

plt.figure(figsize=(16,6))
sns.distplot(test_residuals)
sns.distplot(train_residuals)
plt.show();
```



I like this visualization which shows the distribution of R2. Just asking for the R2 score works but a plot like this helps me understand what a residual is and what it looks like when I use a seaborn visual.



- Bathrooms make up 8 of the top 12 contributors (by percentage) for coefficients in the model.
- This dataset seems to react poorly to "over processing". Meaning log transformations don't improve the model, and in some cases, can increase the residuals. The same can be said for the categorical features, I'm not confident that the one hot encoding technique I deployed helped much at all.
- Outliers will impact this multivariate linear regression model negatively. Its important to spend time testing the model after removing different levels of outliers.
- I couldn't find any evidence to suggest multicollinearity is negatively impacting the model. In fact, the more features I deployed, the more my adjusted R2 score increased, regardless of feature correlation.



- Polynomial regression on certain features may or may not have an affect on the adjusted R2, but I'd like to explore those options in the future.
- Use python's "combinations" library to automate some of the tedious testing when tweaking the makeup of my features and target variable.
- I'm curious as to how other machine learning models perform with this dataset. I look forward to circling back and applying new models I learn in the future.


Thank You
Questions?
Comments?
Feedback?