

Zillow Regression Report

by Matthew Dalton

MARCH 24, 2021





Agenda

- 1 Executive Summary
- 2 Data Wrangling
- 3 Exploration and Stats
- 4 Modeling and Results
- 5 Tax Rate Distribution
- 6 Conclusion

Executive Summary



Goals

- Analyze drivers that assist in the prediction of property tax value using the 2017 Zillow data set



Tax Rate Distribution

Properties are within 1 of 3 counties in California:

- LA County
 - Tax rate: 1.294 %
 - Tax amount: \$5013.56
- Orange County
 - Tax rate: 1.188 %
 - Tax amount: \$5471.98
- Ventura County
 - Tax rate: 1.147 %
 - Tax amount: \$4979.44



Top Drivers of value

Square feet



Bathrooms

Bedrooms



Age

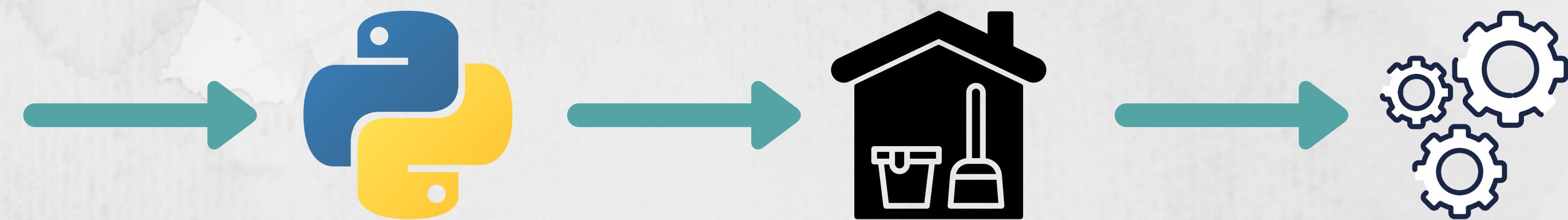


Conclusions

- Polynomial Regression model:
 - RMSE 213,616
 - R squared: 38.1%
- Stats:
 - Square feet
 - $r = 0.5296$
 - bathrooms
 - $r = 0.4431$
- LA County Tax Rate is significantly different from total population

Data Wrangling

Pull data 2017 Zillow data
from the Codeup SQL
Database



Bring data to python to
familiarize with the data:
shape, data types, basic stats,
number of null values, id
categorical and continuous
variables, etc.

Clean data: removing duplicates,
nulls, changing data types, etc.
Engineered data into a
useable form and size for
exploration by removing
outliers, creating new
features from existing
ones, etc.

Exploration and Stats

Feature Selection:

Used **Select K Best** and **RFE** to assist in selecting best features for my model.

Top 4 :

- bathrooms
- bedrooms
- square_feet
- age_of_house

Correlation Test

Hypothesis:

- The H0 is that there is no correlation between the feature and tax_value
- The H1 is that there is a correlation between the feature and tax_value

Results:

- **Move forward with the H1 with all features**

Feature	r	p-value
bathrooms	0.4431	0.0
bedrooms	0.2590	0.0
square_feet	0.5296	0.0
beds_and_baths	0.3871	0.0
beds_per_sqft	-0.3647	0.0
baths_per_sqft	-0.152	0.0

One Sample T-test

Hypothesis:

- The H0 is that there is no difference in the means of the LA County Tax rates.
- The H1 is that the LA County Tax Rate is a different mean than the entire population.

Results:

- **Move forward with H1**
- t-statistic: 46.5759
- p-value = 0.0

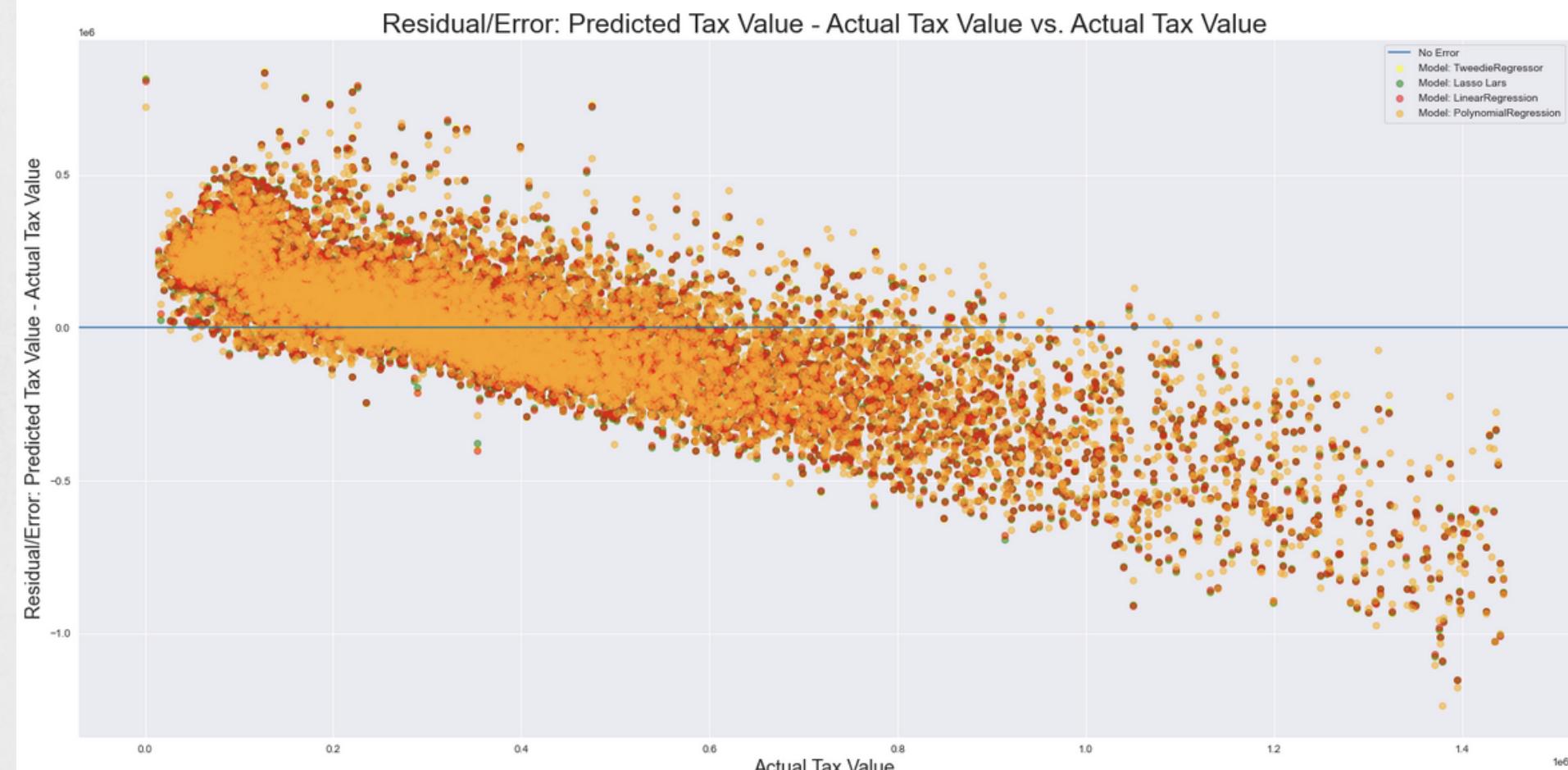
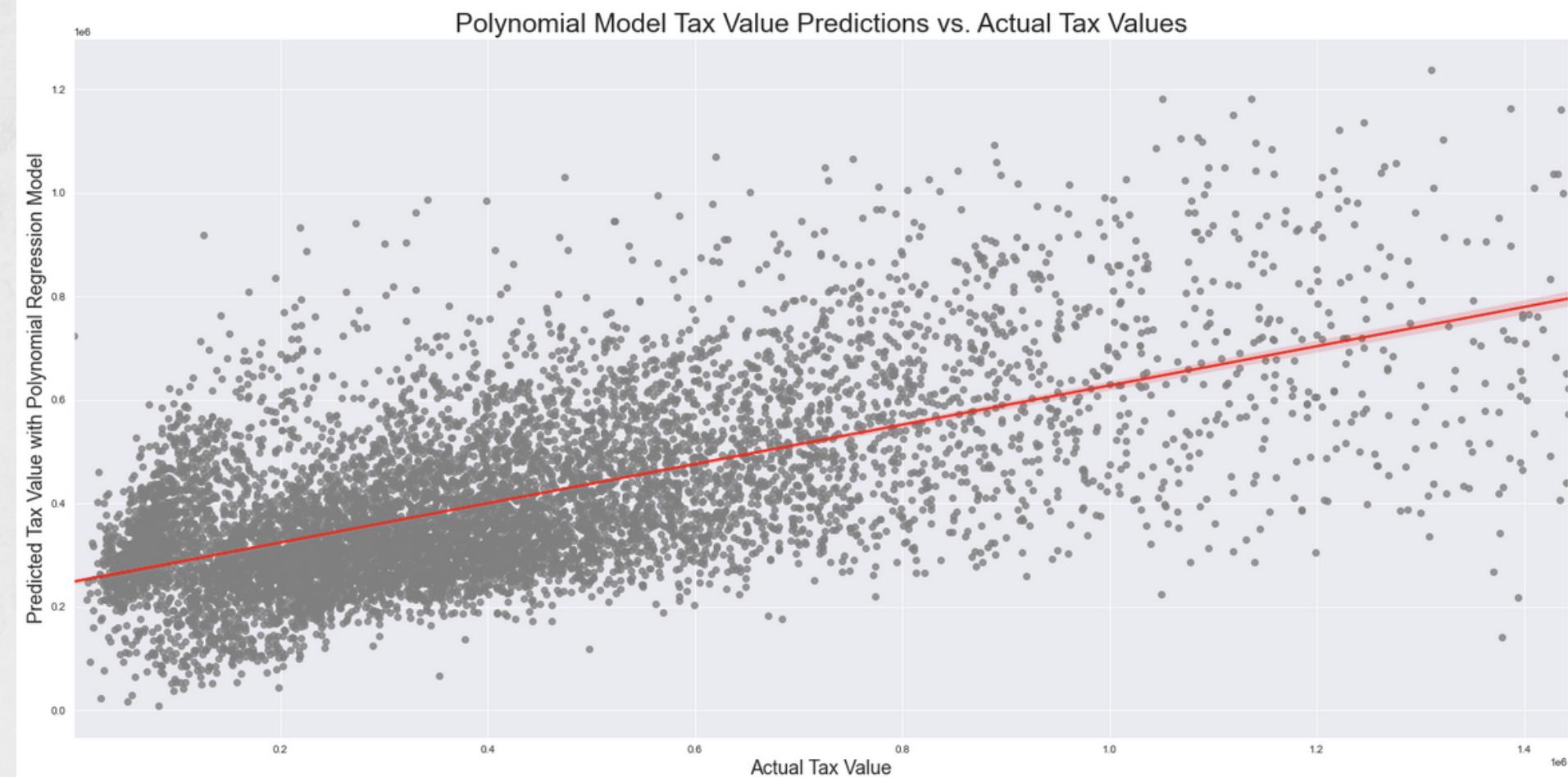
Modeling and Results

My best model was the **Polynomial Regression model**

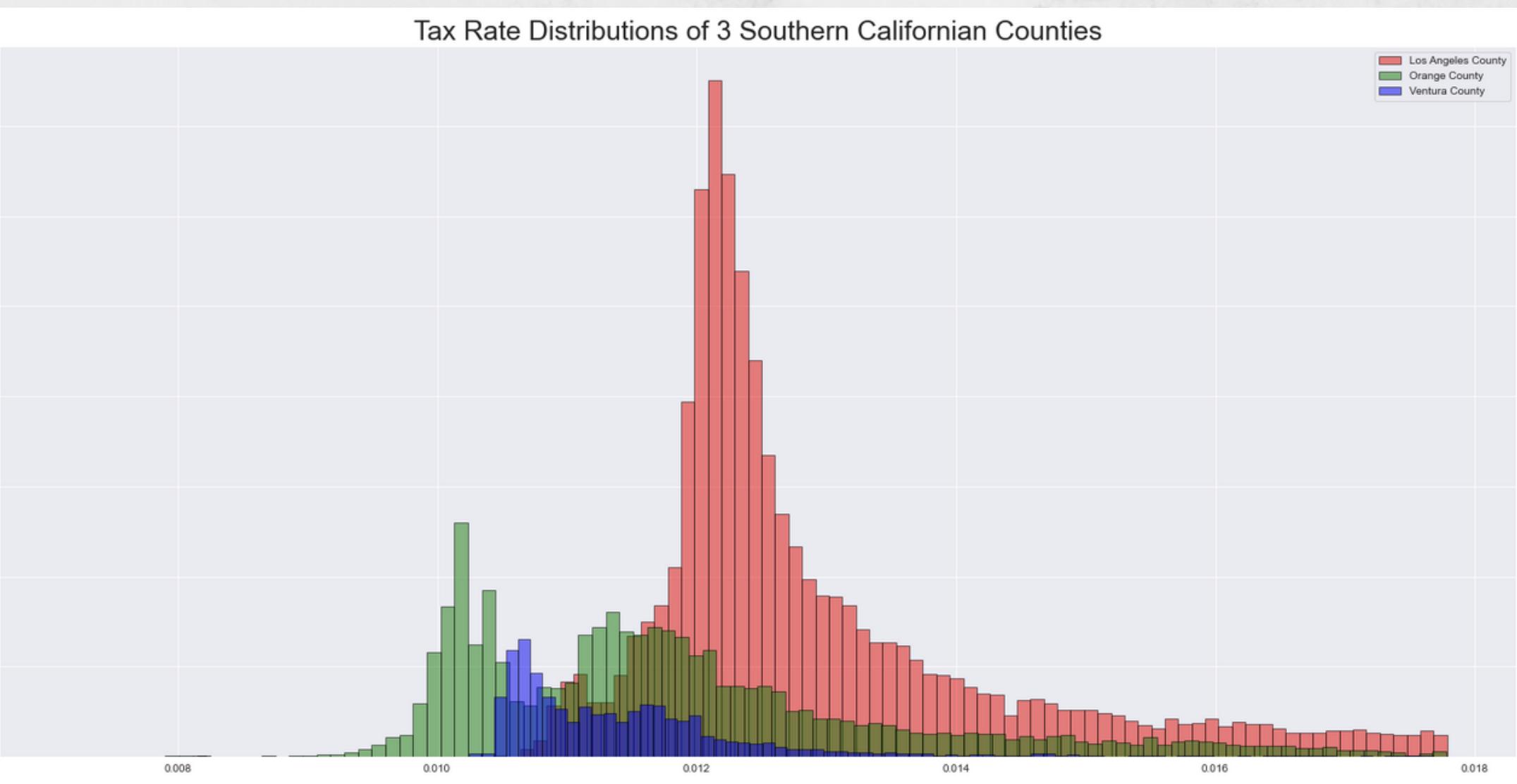
- RMSE dropped from the baseline of 272,150 to 213,616
 - **RMSE improvement of 58,534**
- **R2 Value of the model, 38.1%**, indicates there is an improvement over the baseline but not a very strong fit overall.

Interpretations:

- My model begins by **overvaluing** properties less than **~\$150K**
- Then appears to be most accurate from ~\$150K to ~\$500K
- Then begins to **undervalue** properties over **~\$500K**



Tax Rate Distribution



- LA County
 - Tax rate: **1.294 %**
 - Tax amount: **\$5,013.56**
- Orange County
 - Tax rate: **1.188 %**
 - Tax amount: **\$5,471.98**
- Ventura County
 - Tax rate: **1.147 %**
 - Tax amount: **\$4,979.44**

Conclusion

Stats

- All H0's are rejected
- **Square feet** and number of **bathrooms** are good drivers of the property value
- A large amount of features are lacking data
- LA County's tax rate is significantly different from all 3 counties

Model

- I created a **Polynomial Regression** model that out performs the baseline RMSE by 58,534 and fits the data with an **R squared value of 38.1%**.
- Not a great model, with more time I will explore more features and improve upon this model

Appendix



For complete details of the
Zillow project visit my Github
and take a look at the
corresponding README.md file.



Connect with me on Linkedin
and maybe we can collaborate
on future projects