Predicting Home Values

By: Caitlyn Carney
03/23/3031

1 Executive summary slide

2 Overview

3 Preparing Data

4 Exploring Data

5 Modeling

6 Conclusion

7 Appendix

Zillow

Big Idea

The number of bedrooms, bathrooms and amount of square footage affects home value prices in southern California.

Goals

Create a regression model to predict home value in southern California as accurately as possible.

The Process

- Gather and clean Zillow data
- Explore data and finding what it means.
- Evaluating and modeling the data.

Key Findings

Appraisal value in southern California is driven by number of bedrooms, bathrooms, as well as square footage.

Takeaways

Square feet, bedrooms, and bathrooms do have an affect the appraisal value of homes in southern California.

Zillow

Prepare Data

After gathering the data from the Zillow database in the Codeup Sequel server I needed to clean it up like crazy!

How unseemly!

parcelid	id	airconditioningtypeid	architecturalstyletypeid	basementsqft
14634203	2026522	1.0	NaN	NaN
11721753	616260	NaN	NaN	NaN
11289917	2061546	1.0	NaN	NaN
11637029	2554497	1.0	NaN	NaN
11705026	1834372	NaN	NaN	NaN

So many blank values!

```
airconditioningtypeid        26358
architecturalstyletypeid     38481
basementsqft                 38555
                               ...
taxdelinquencyyear           37314
censustractandblock            144
```

I needed something easier to handle with columns not missing half of their values.

Zillow

Prepare Data

I ended up going from 38,582 rows to 32,542 rows by cleaning.

For more detail please see prepare.py on GitHub

In order to clean it I:

- Dropped columns with less than 35 thousand non null values.
- Dropped unhelpful columns.
- Renamed columns.
- Handled outliers in some columns.
- Drop left over null values

bathrooms	bedrooms	square_feet	fips	full_baths
2.0	3.0	1125.0	6059.0	2.0
2.0	3.0	1316.0	6037.0	2.0
2.0	3.0	1458.0	6037.0	2.0
2.0	3.0	1766.0	6037.0	2.0
1.0	2.0	1421.0	6037.0	1.0

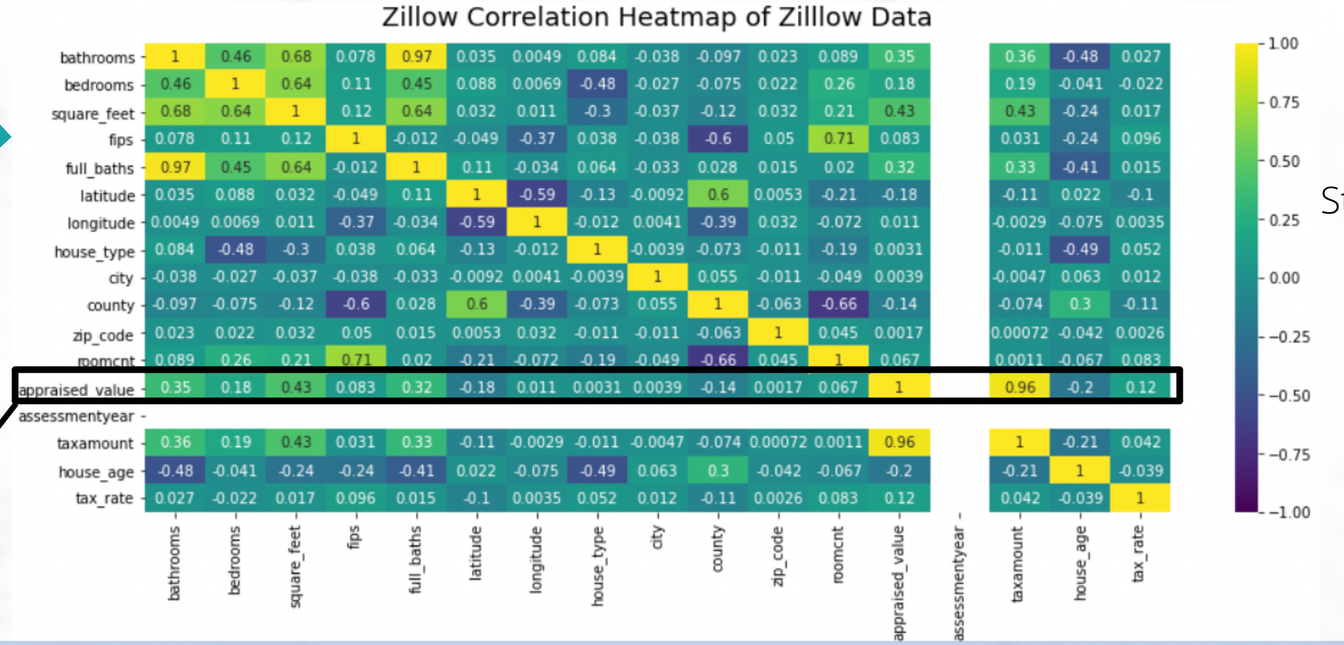
Now thats what I call clean!

Zillow

Prepare Data

After this, I focused my data onto theses features, split my data, and scaled the split data.
For further detail please see prepare.py on GItHub

Decision time! What columns to focus on!



Strongest correlations to the target
- square_feet 0.43
- bathrooms 0.35
- bedrooms 0.18
- latitude -0.18
- county -0.14
- house_age 0.13
- tax_rate 0.12

I will be focusing on square feet, bedrooms, and bathrooms today.

Zillow

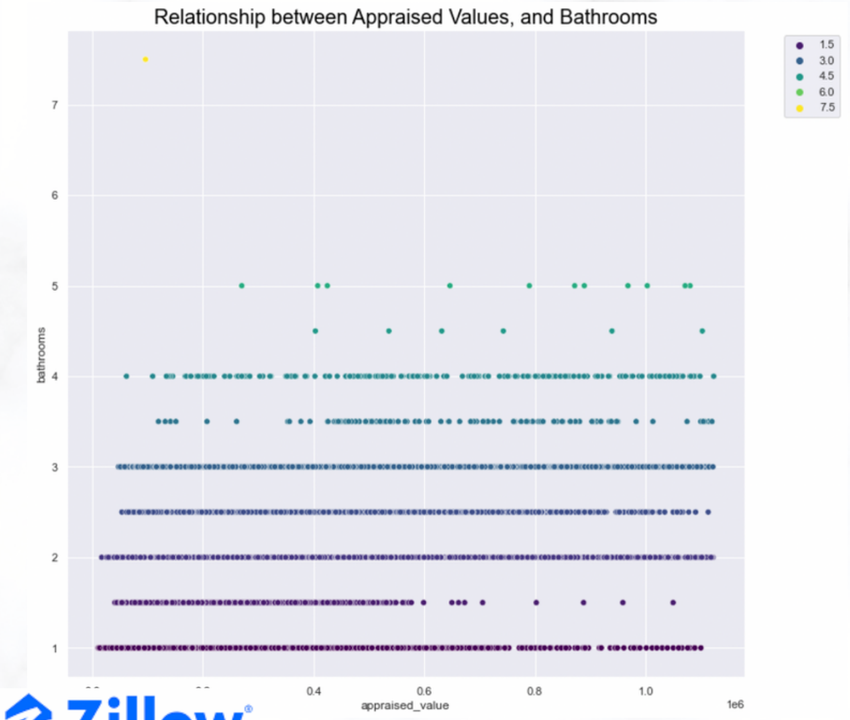
Explore Data

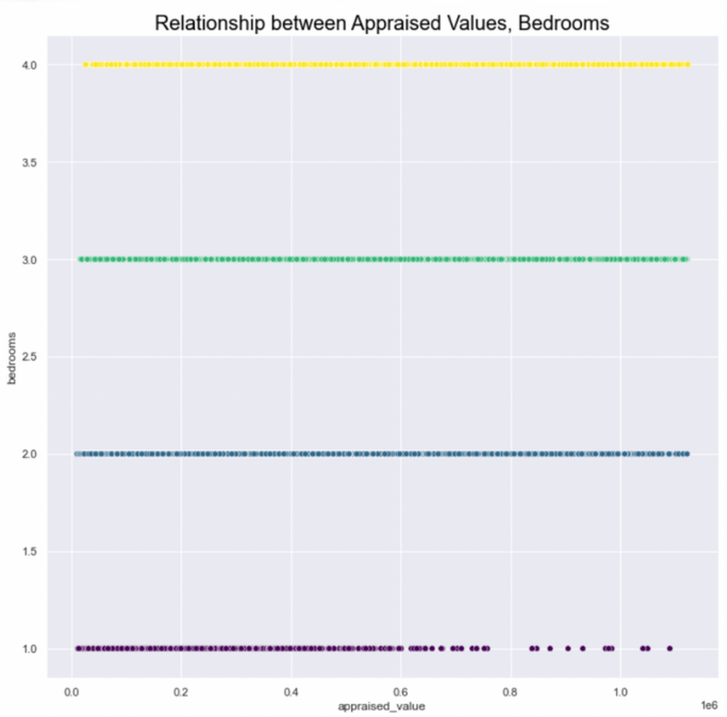
Now that the data is clean and easy to read, I needed to take a deep dive into the data and find relationships, abnormalities, etc.

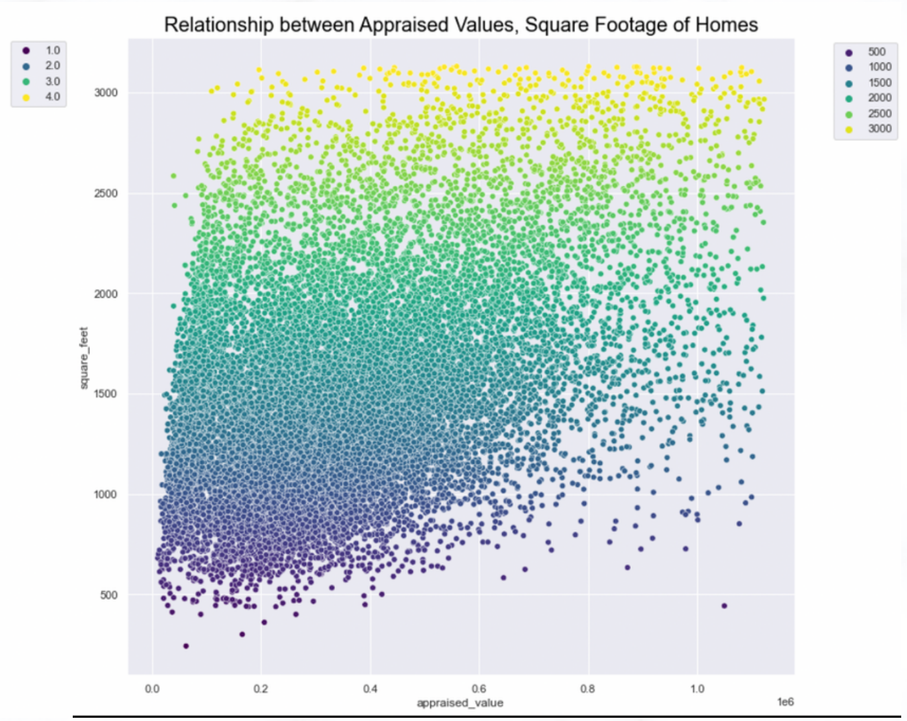
Using Scatterplots to see the Relationships to Appraised Value

- There seems to be a noticeable difference in the amount of homes with a low appraised value and low square footage vs. homes with a higher appraised value's square footage

- Homes with 1 bedroom seem to drop off when appraised values get higher while the 2+ bedroom houses seem to continue rising with the appraised value



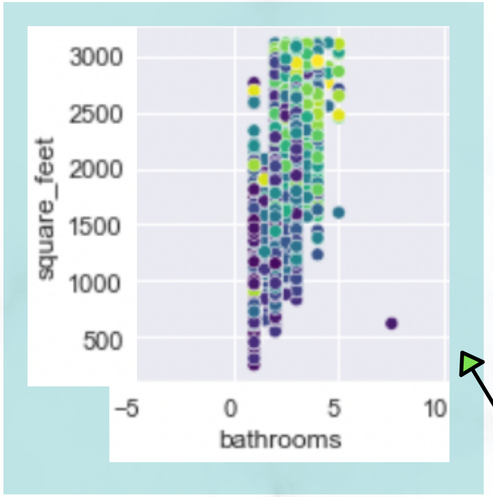




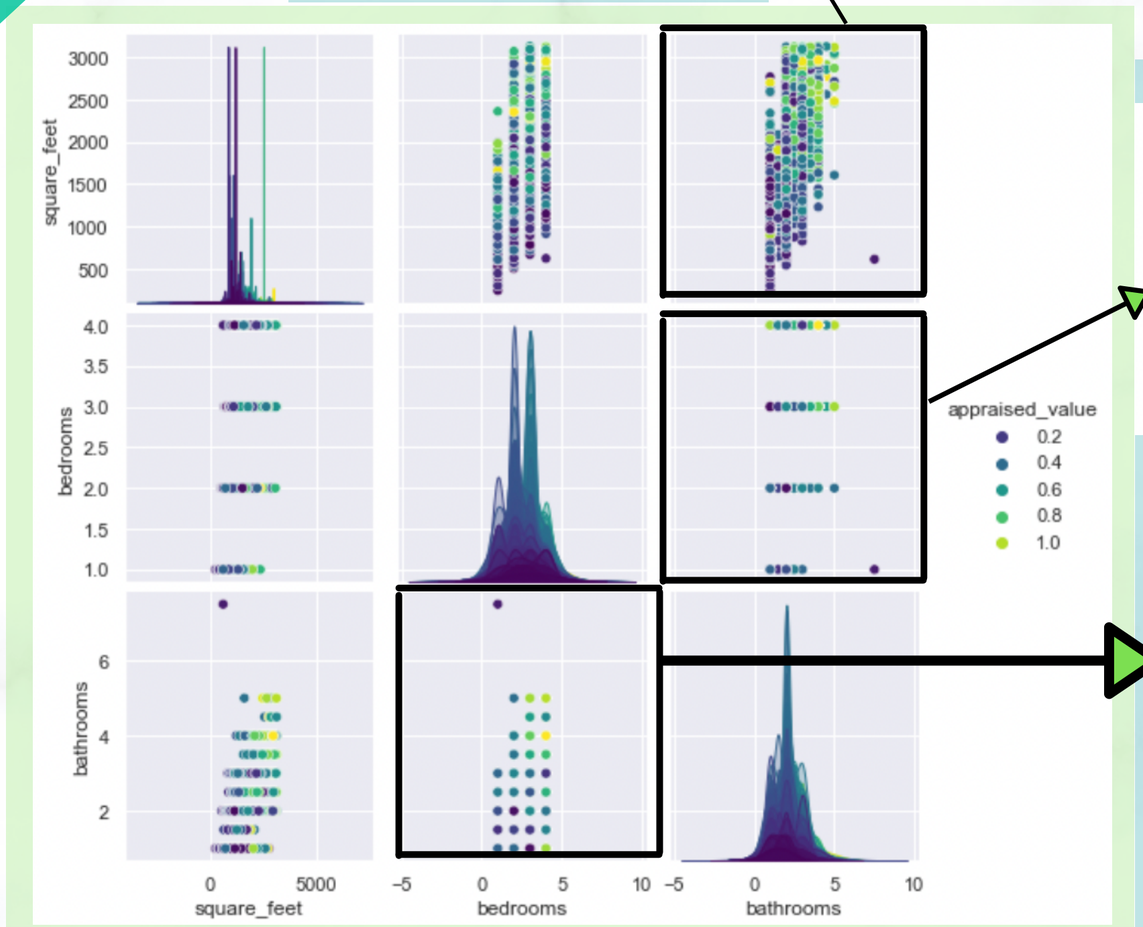
Zillow

Explore Data

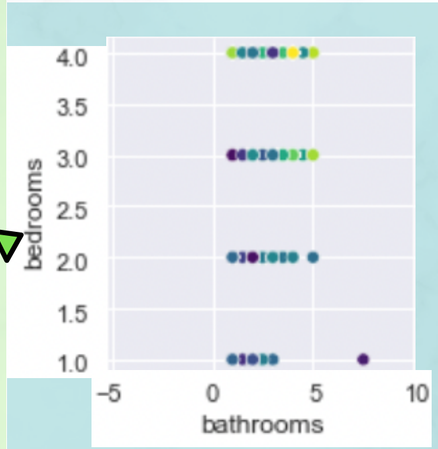
Here I wanted to see how appraised value was affected by the relationship between other feature relationships.



- It seems that homes with more bathrooms and larger amount of square feet tend to be appraised at a higher amount.
- The yellow which indicated the higher appraisal rates, are more heavily indicated the higher the bathroom number and amount of square feet gets.

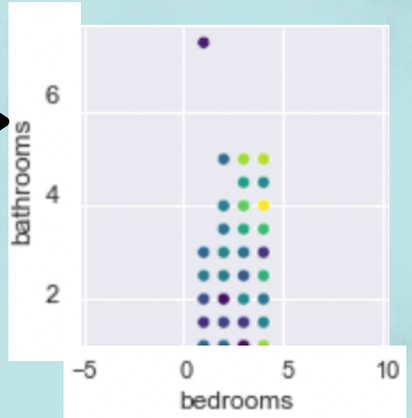


We can see a gradual increase in appraisal price in the bedrooms and bathrooms relationships.



- When both features are high we move into higher appraisal values

This is indicated by the green in the top right.



- The lower amount of bedrooms and bathrooms, the lower the appraisal value gets

This is indicated by the dark blue on the bottom left.

Zillow

Explore Data

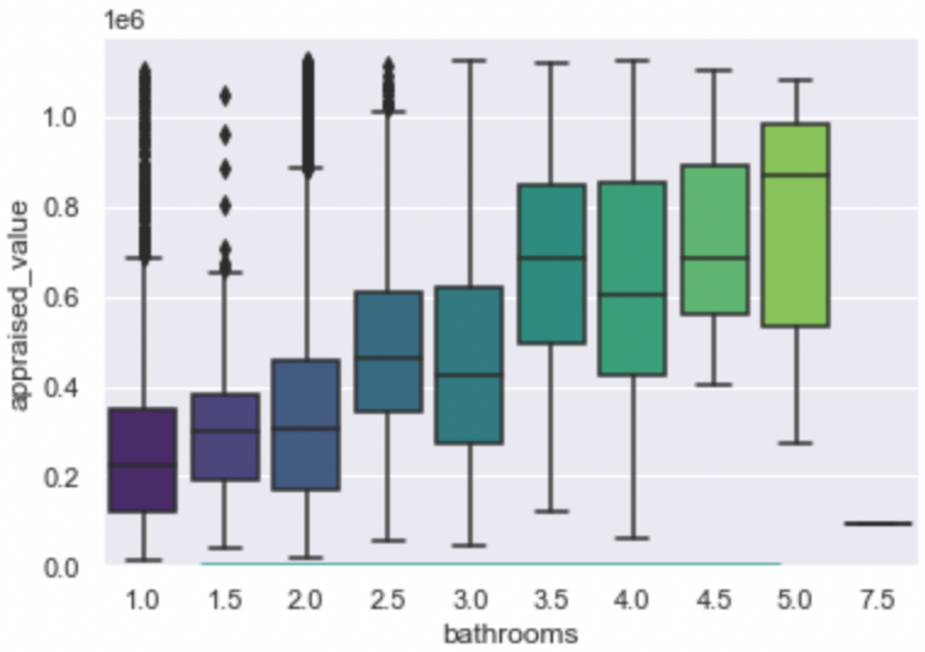
Take a deep dive into the bathroom feature. I ran a correlation test for this features.
For further detail please see explore.py on GitHub

Null Hypothesis:

"There is no correlation between number of bathrooms and appraised value."

Alternative Hypothesis

"There is a correlation between number of bathrooms and appraised value."



The correlation between Bathrooms and the Appraised value is: 0.356

The P value between Bathrooms and Appraised Value is: 0.0

I reject the null hypothesis and now move forward with our alternative hypothesis

Zillow

Explore Data

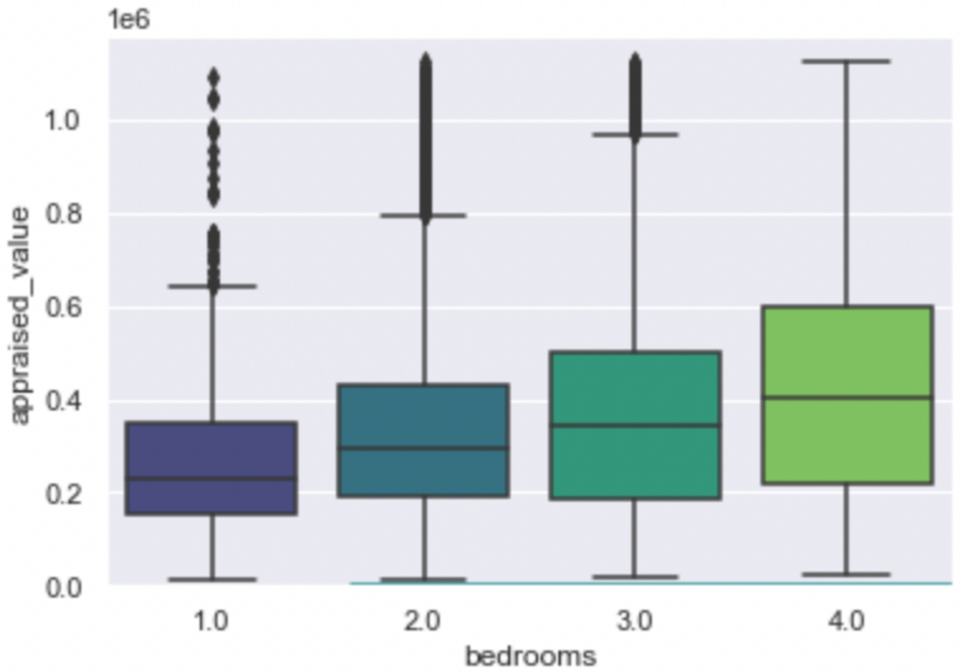
Take a deep dive into the bedroom feature. By running a correlation test.
For further detail please see explore.py on GitHub

Null Hypothesis:

"There is no correlation between number of bedrooms and appraised value"

Alternative Hypothesis

"There is a correlation between number of bedrooms and appraised value."



The correlation between Bathrooms and the Appraised value is: 0.177

The P value between Bathrooms and Appraised Value is: 7.71

I reject the null hypothesis and now move forward with our alternative hypothesis

Zillow

Explore Data

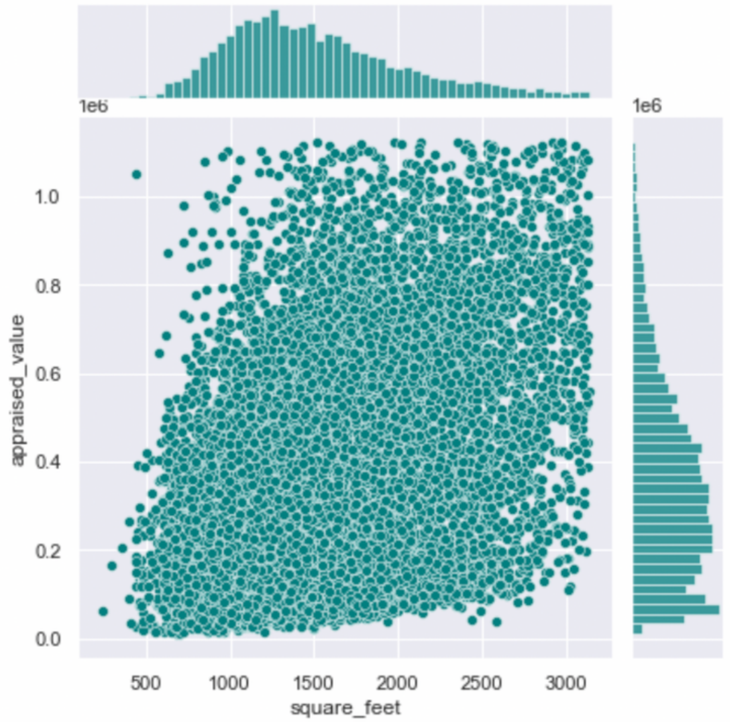
Take a deep dive into the square footage. By running a correlation test.
For further detail please see explore.py on GitHub

Null Hypothesis:

"There is no correlation between a homes square footage and appraised value."

Alternative Hypothesis

"There is a correlation between square feet and appraised value."



The correlation between Bathrooms and the Appraised value is: 0.432

The P value between Bathrooms and Appraised Value is: 0.0

I reject the null hypothesis and now move forward with our alternative hypothesis

Zillow

Model

Now that I have gotten to know the data, it is now time to evaluate and create models! For further detail please see the evaluate.py and model.py on GitHub

My predicted mean was : 364,215.53

My predicted median was : 329,107

I will move forward using my predicted median!

SSE = 911614726222345.5

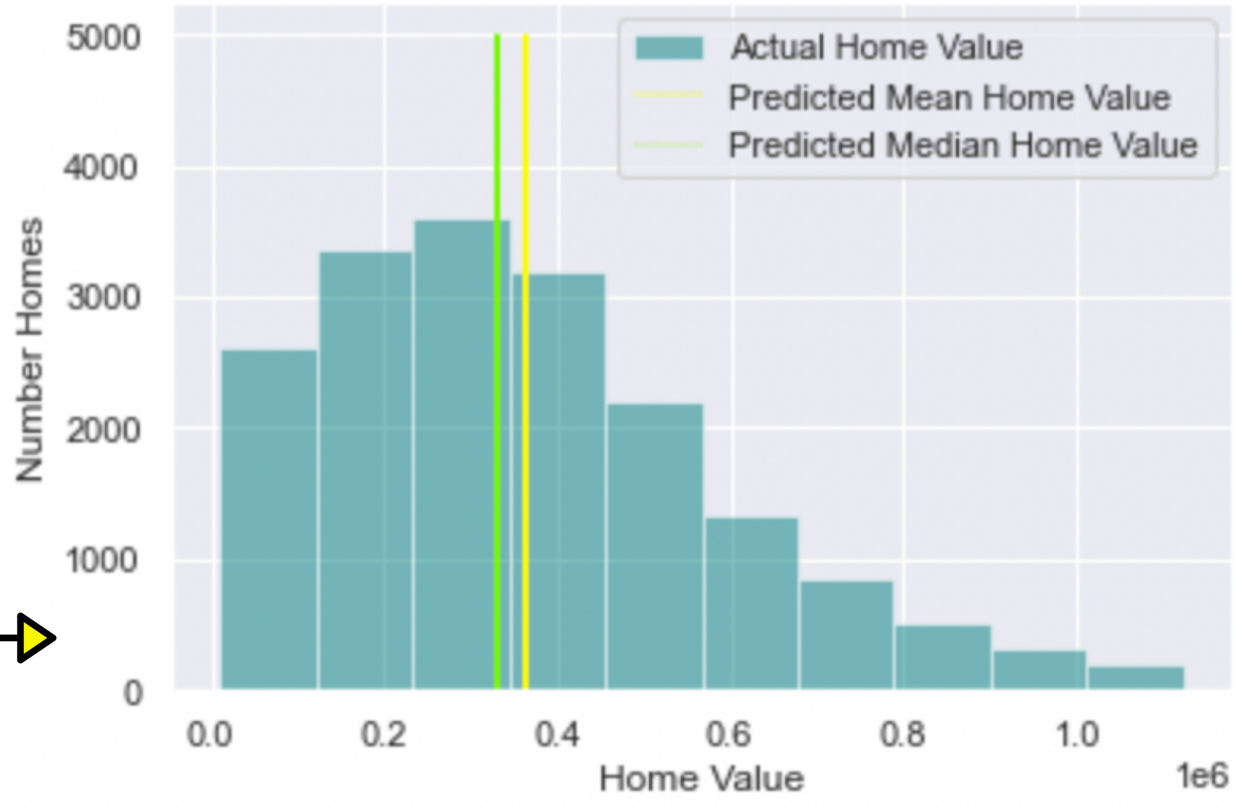
SSE Baseline = 941263447006771.6

MSE = 28013481845.687

MSE baseline = 28924572767.709

RMSE = 167372.285

RMSE baseline = 170072.257



Zillow

Model

Now I take a look at each model and determine the best one!

For further detail please see the evaluate.py and model.py on GitHub

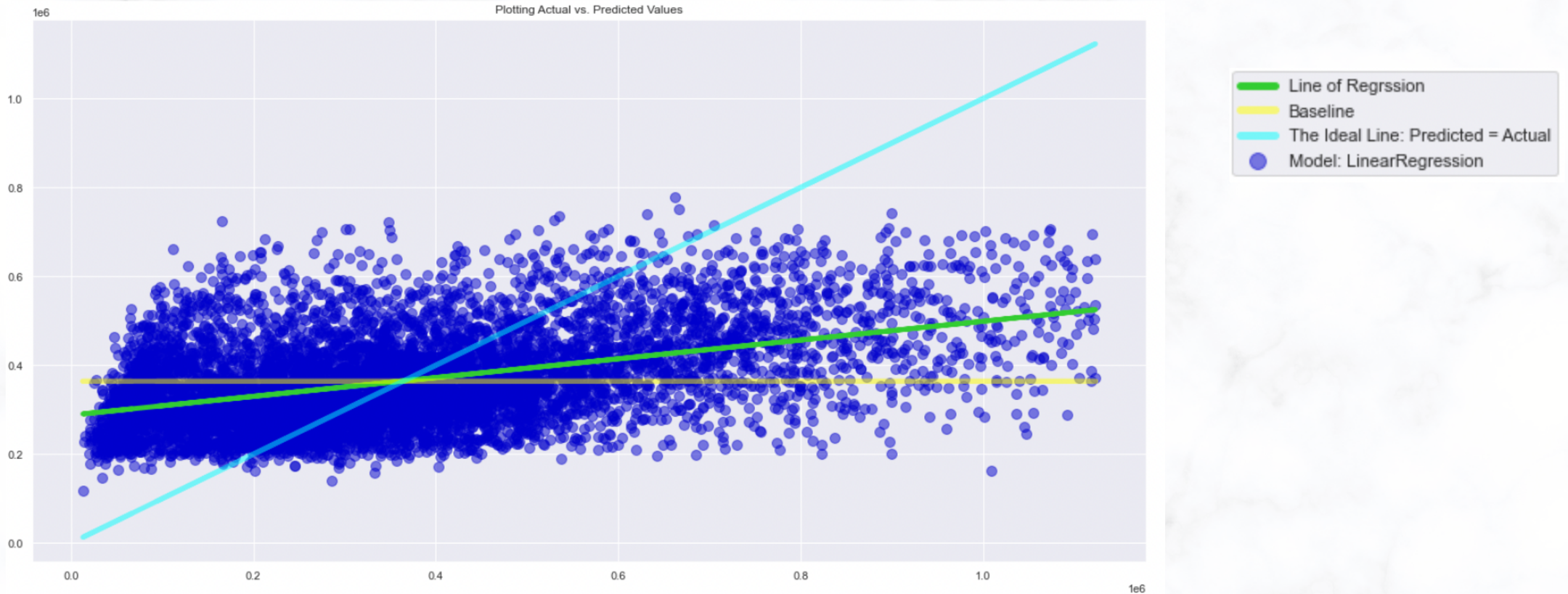
LinearRegression Model
In-Sample: 201813.737
Out-of-Sample: 202696.951

Lasso Model
In-Sample: 20364445665.7664
Out-of-Sample: 20542786732.111

Tweedie Model
In-Sample: 25826248340.196
Out-of-Sample: 26018336332.939

Polynomial Model
In-Sample: 20298379492.33372
Out-of-Sample: 20476129727.74332

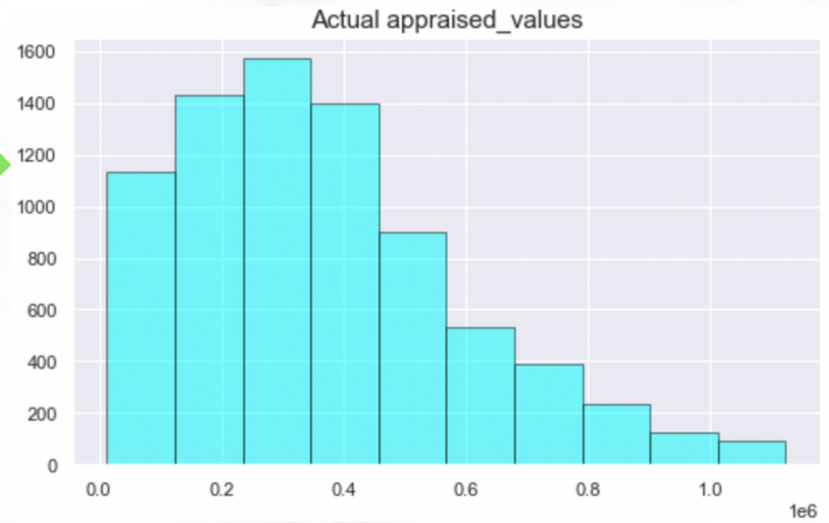
RMSE = 167372.285
RMSE baseline = 170072.257

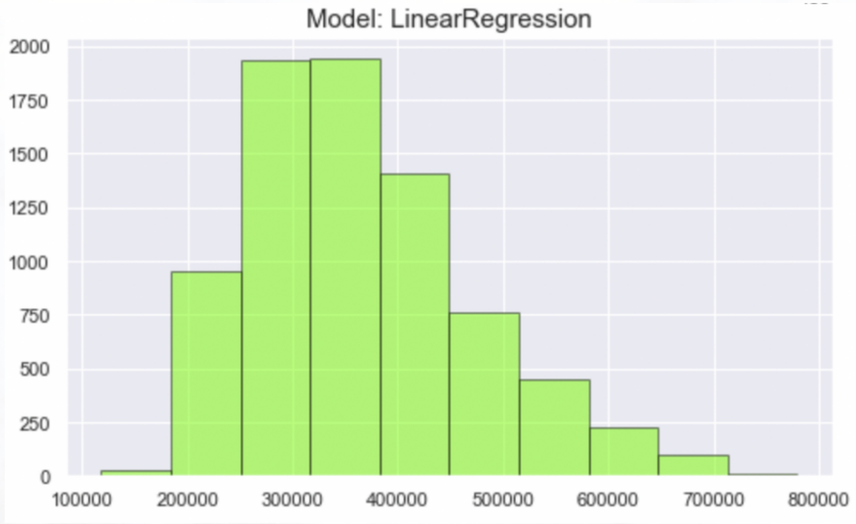


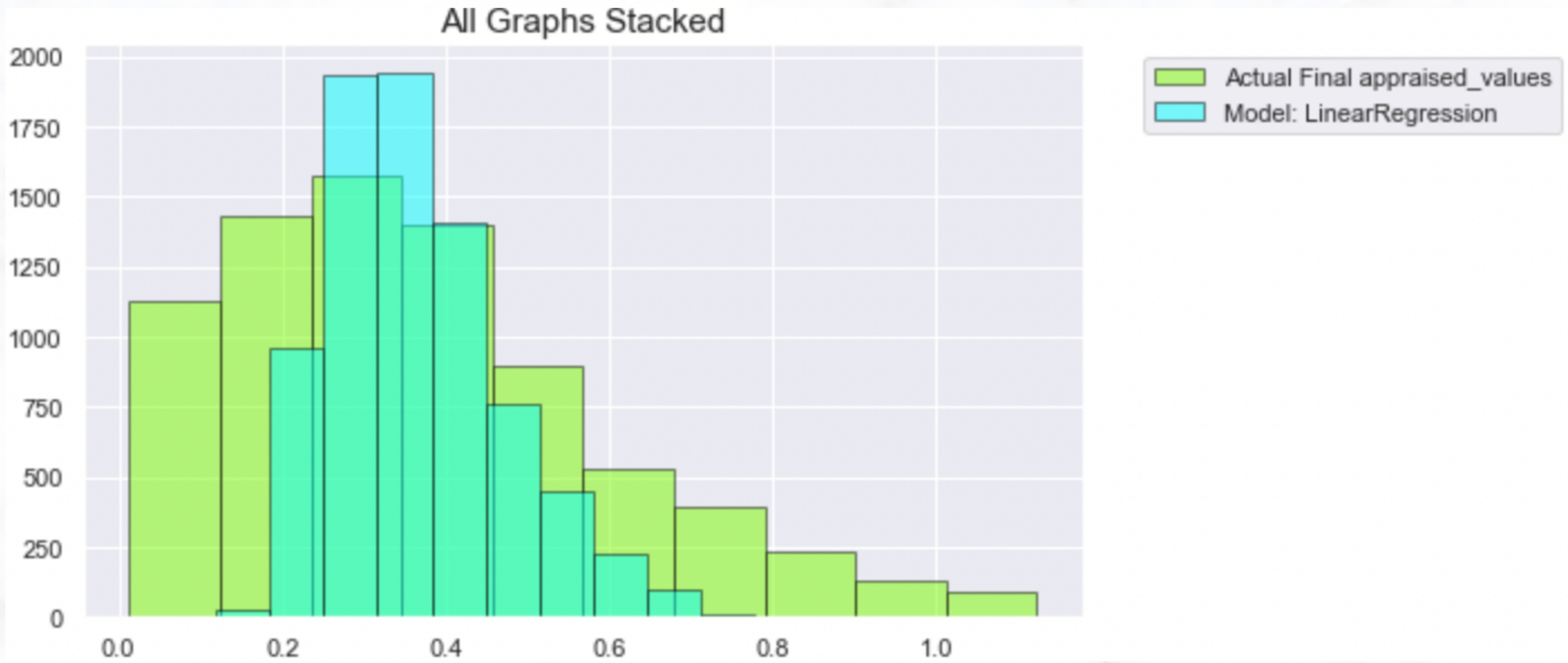
Zillow

Model

Here I visualize the actual appraised value vs. predicted appraised value.
For further detail please see the evaluate.py and model.py on GitHub



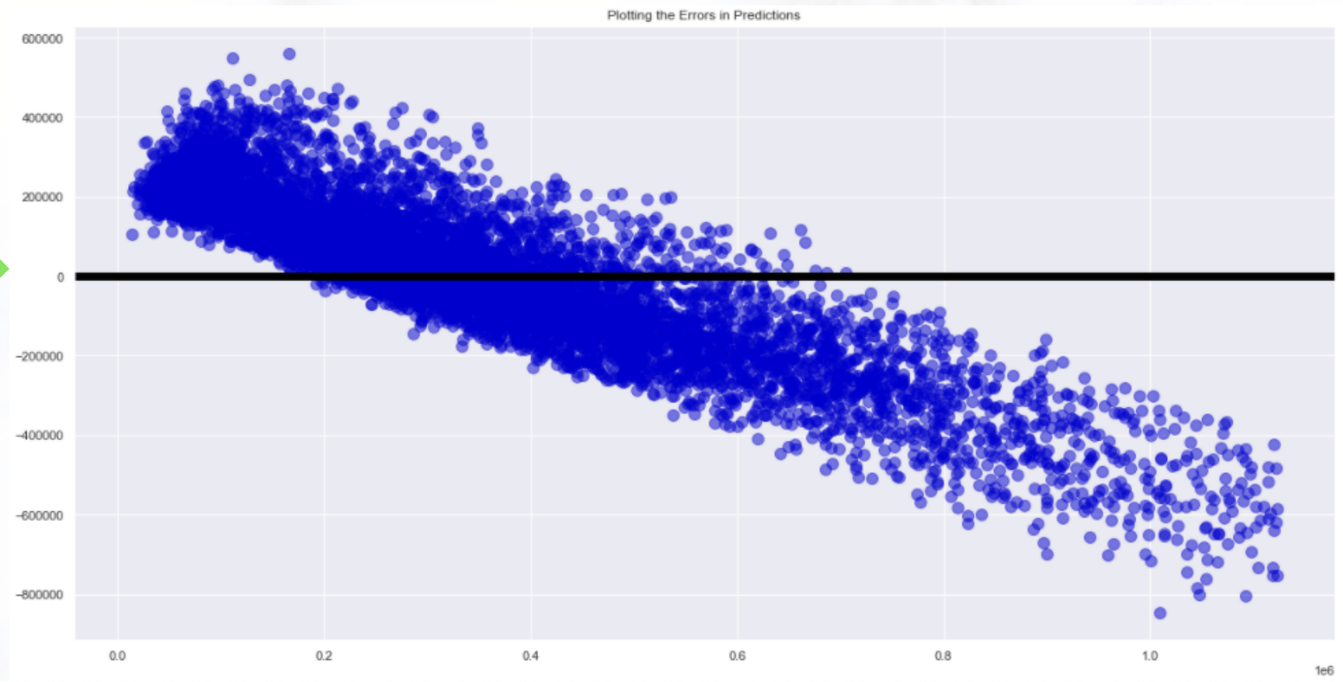


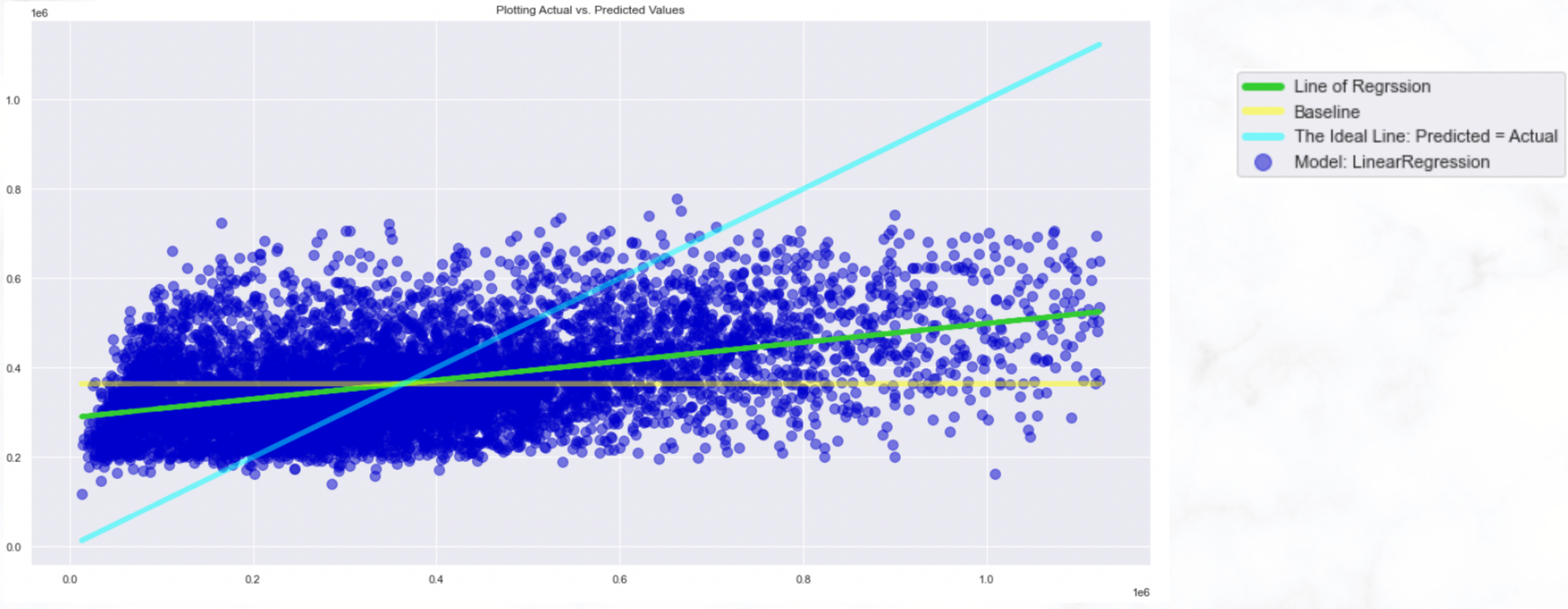


Zillow

Model

Here I visualize the errors in my prediction model and actual vs. predicted.
For further detail please see the evaluate.py and model.py on GitHub





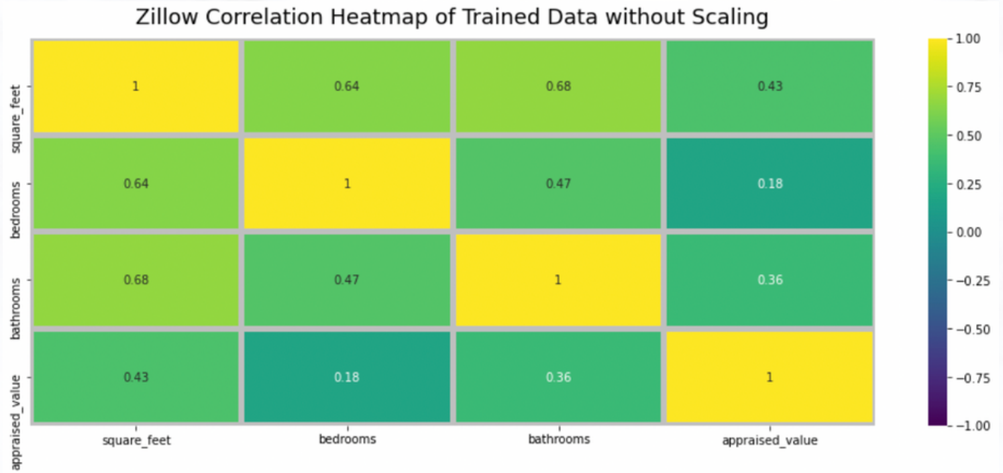
Zillow

Conclusion

By gethering, cleaning, exploring, and modeling the Zillow data; I learned a lot.

In southern California, a homes appraised price is affected in many ways.

Square feet, number of bedrooms, and the number of bathrooms are just to name the top 3 drivers.



By looking at these top 3 drivers we can see a correlation between them and appraisal values.

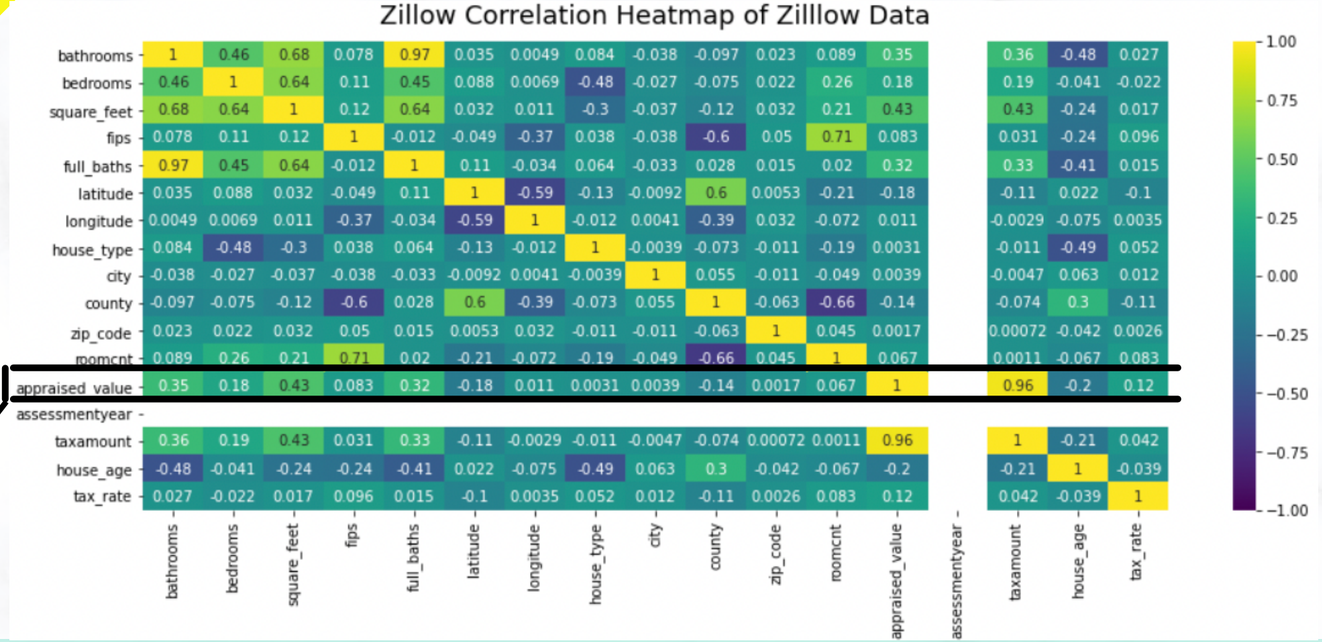
Zillow

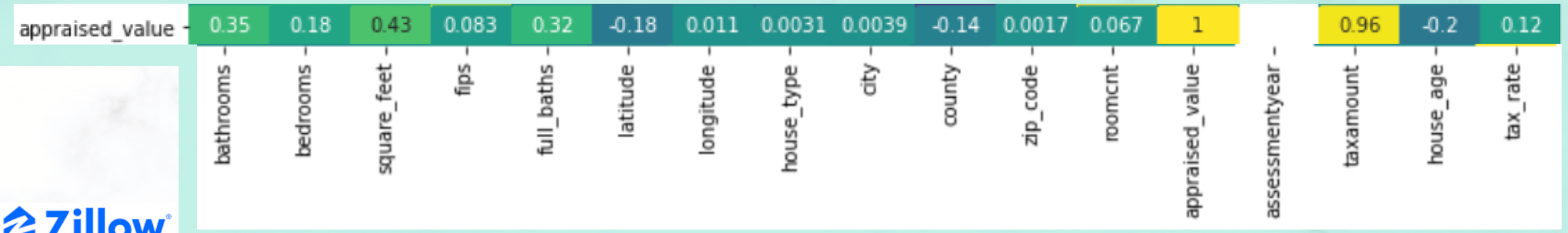
Conclusion

By gethering, cleaning, exploring, and modeling the Zillow data; I learned a lot.

Even though these three drivers give a good prediction of appraisal values, we can't discount the possibility that features such as zip code, city, etc. have a role in the pricing as well.

Although these other features may have a smaller correlation that doesn t mean that together dont affect appraisal value in a big way.





Zillow

Appendix

Follow these links for more information on the creation of the project, steps taken, and myself.

Link to the GitHub Repository

Zillow Home Value Prediction Modeling

- README.md
- acquire.py
-prepare.py
- explore.py
- evaluate.py
- model.py
- project-final.ipynb

Link to the Trello Board

Zillow Predicting Trello Board

Link to the my Linkedin

Caitlyn Carney