CAPSTONE PROJECT 1 PRESENTATION

STUDENT: **NATHAN ZHANG**

MENTOR: **MISAEL MANJARRES**

PROJECT OVERVIEW

1. Context Familiarization (10%)
2. Data Wrangling (35%)
3. Exploratory Analysis (30%)
4. In-Depth Analysis w/ ML (20%)
5. Result & Conclusion (5%)

** #'s in parenthesis represent time allocation*

1. CONTEXT FAMILIARIZATION

- Zillow provides real estate information for consumers
- Zillow wants to push the accuracy of home value estimates (Zestimate) further
- Out goal is to predict residual errors, defined as log(Zestimate) - log(Sale Price)
- Zillow does not provide actual residual errors from later transactions, so we will look at R^2 instead

1. CONTEXT FAMILIARIZATION

- Properties_2016.csv – all the homes with property features for 2016
- Properties_2017.csv – all the homes with property features for 2017 (as of 10/2/2017)
- Train_2016.csv – transaction data from 1/1/2016 to 12/31/2016
- Train_2017.csv – transaction data from 1/1/2017 to 9/15/2017 (as of 10/2/2017)

1. CONTEXT FAMILIARIZATION

- Properties data includes homes in greater LA area (2,985,217)
  - Homes in Properties_2016 & _2017 are the same ones
- Train data contains logerrors in each year (90,275 & 77,613)
- We only use Properties data w/ corresponding logerrors (90,275 + 77,613)

2. DATA WRANGLING

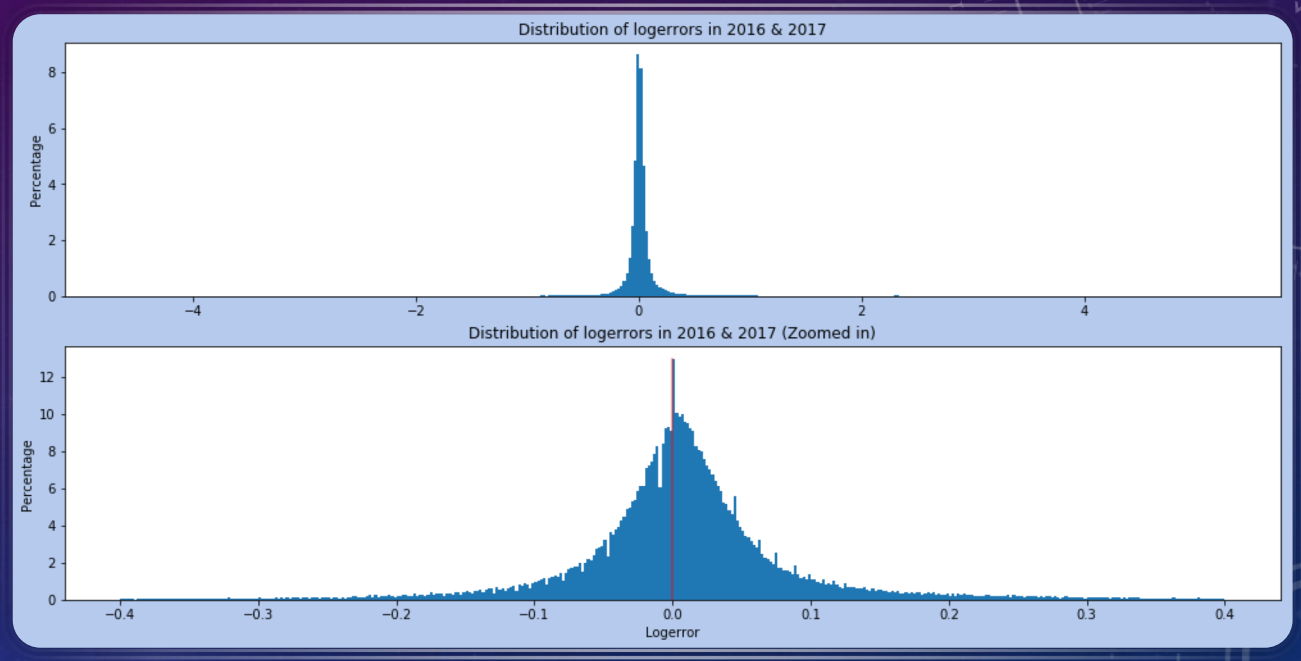
- Identify different types of features
  - Numerical: #bedroom, #bathroom, basement size, etc.
  - Categorical: air conditioning system, construction type, architecture style, etc.
  - Other: transaction date, longitude & latitude, etc.
- Find & handle missing values
  - Fill w/ 0's: #fireplace, #pool, etc.
  - Fill w/ means: most numerical variables
  - Fill w/ "missing": most categorical variables

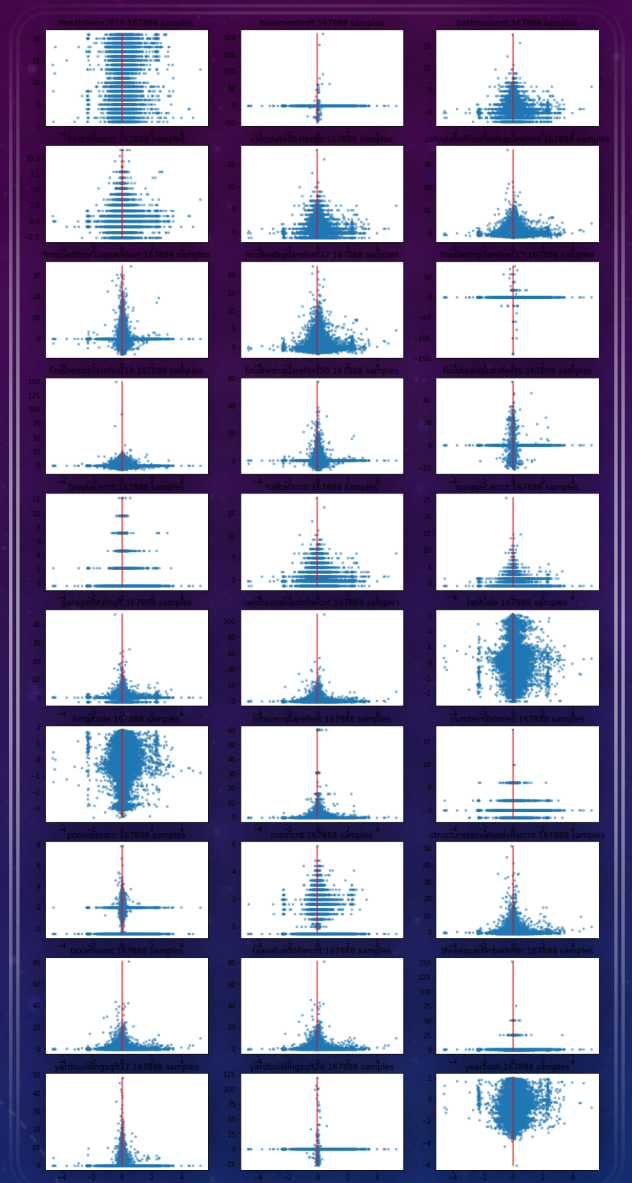
3. EXPLORATORY ANALYSIS

- Target variable
- Numerical variable
- Categorical variable
- Other variable

3. EXPLORATORY ANALYSIS
TARGET VARIABLE

- Mostly near zero
  - 95% logerrors within ±.24
- Thin but long tails
  - Only 0.4% logerrors > 1 or < -1
  - Max = 5.3, Min = -4.7
- Slightly positive-skew





3. EXPLORATORY ANALYSIS
NUMERICAL VARIABLES

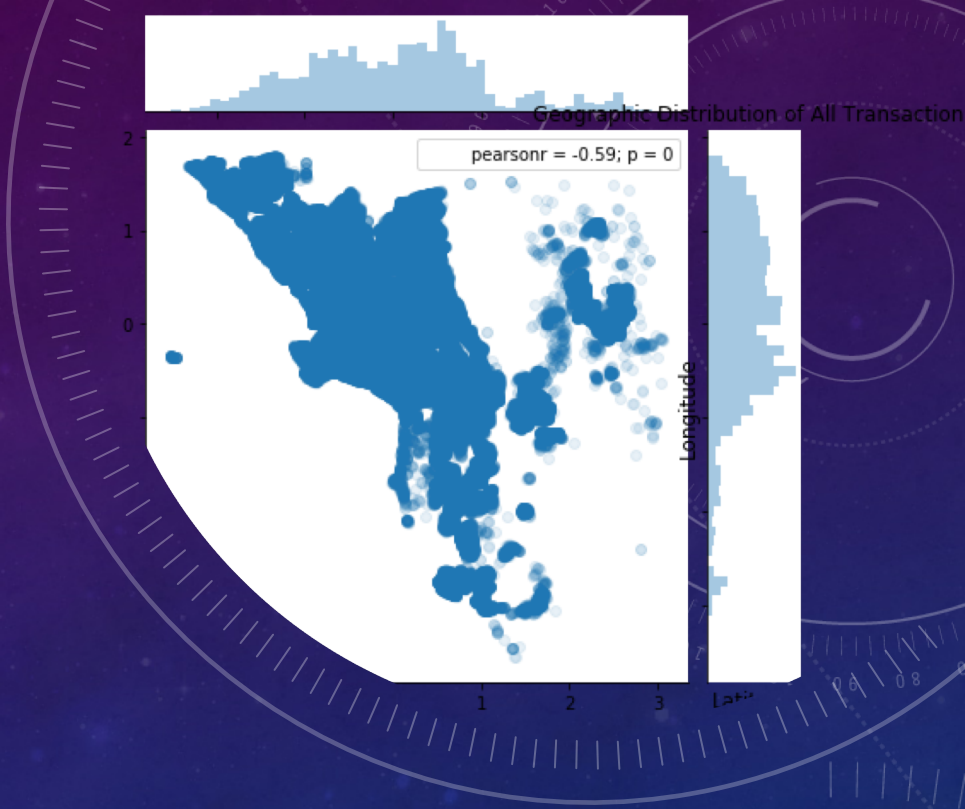
- No obvious trend from visual
- Very weak correlation w/ target variable
  - Due to large # of zero-ish values
- Weak correlation between variables
- 20 variables w/ low p-value from OLS are selected

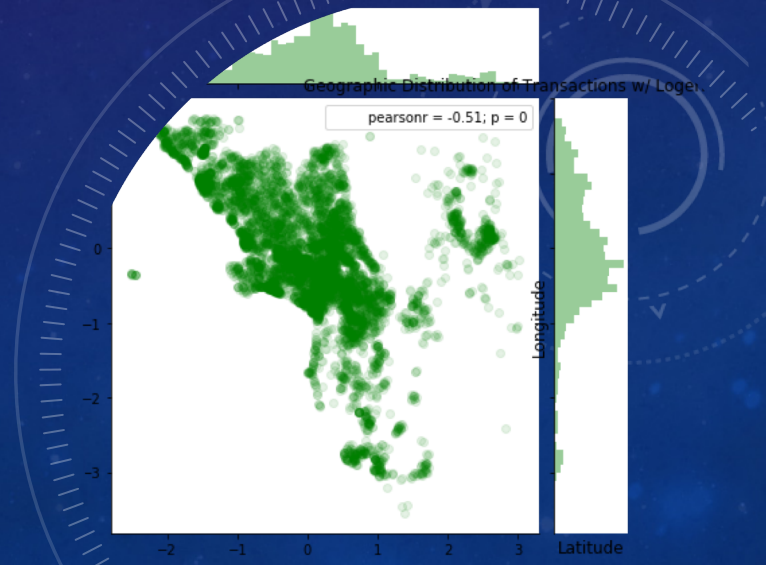
3. EXPLORATORY ANALYSIS
CATEGORICAL

- Created 3000+ dummies from 27 variables
  - Fitting OLS w/ a 167000 x 3000 data frame is a pain ☹
- 10 variables are statistically helpful
  - E.g. Pool, City, Building Quality Type, Class Type, etc.
  - They generate ~600 dummies

3. EXPLORATORY ANALYSIS OTHER

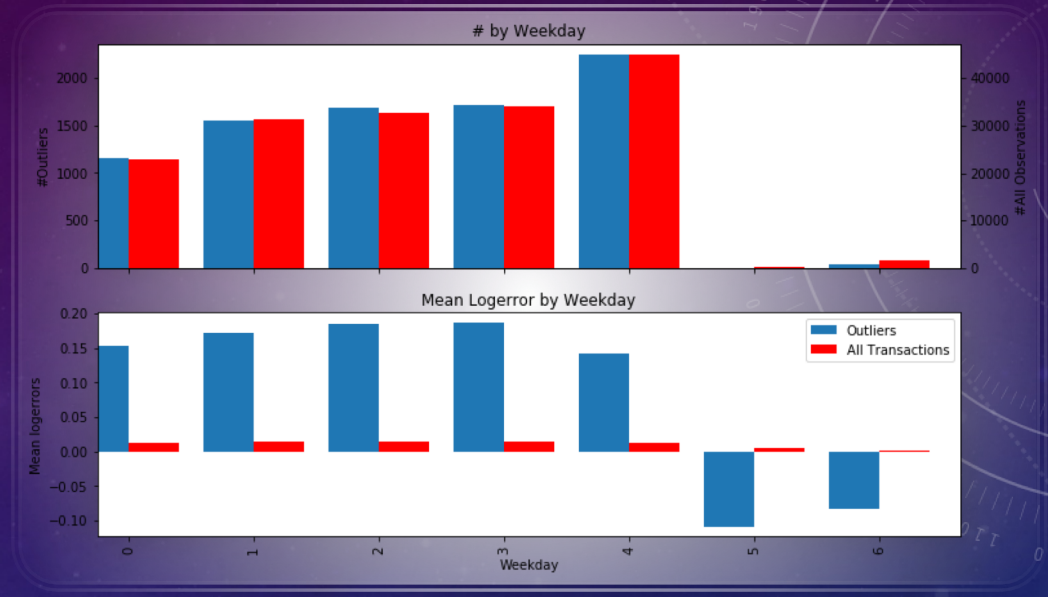
- Longitude & Latitude
  - Not statistically helpful
  - Excluded from later analysis
  - Zip code is a better alternative
    - more specific and useful

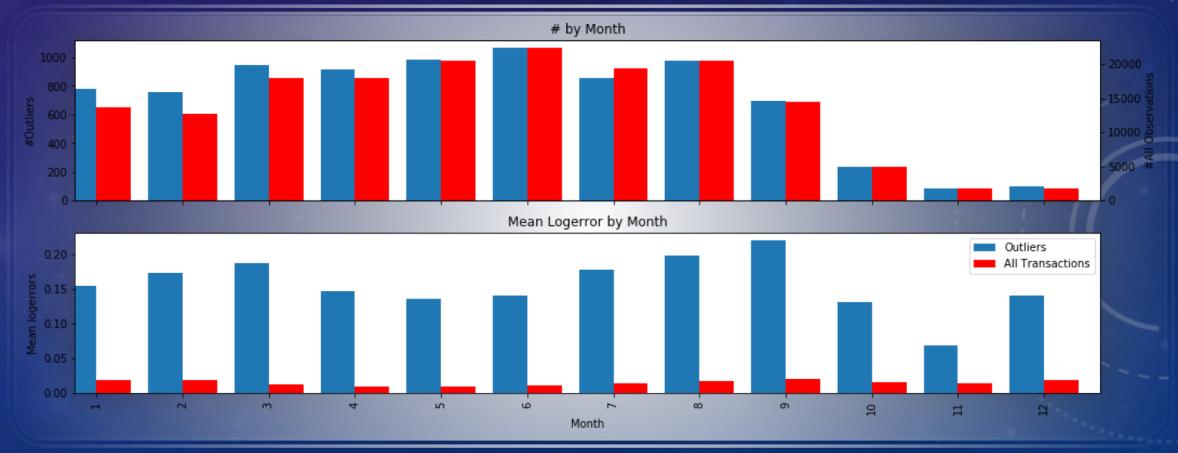




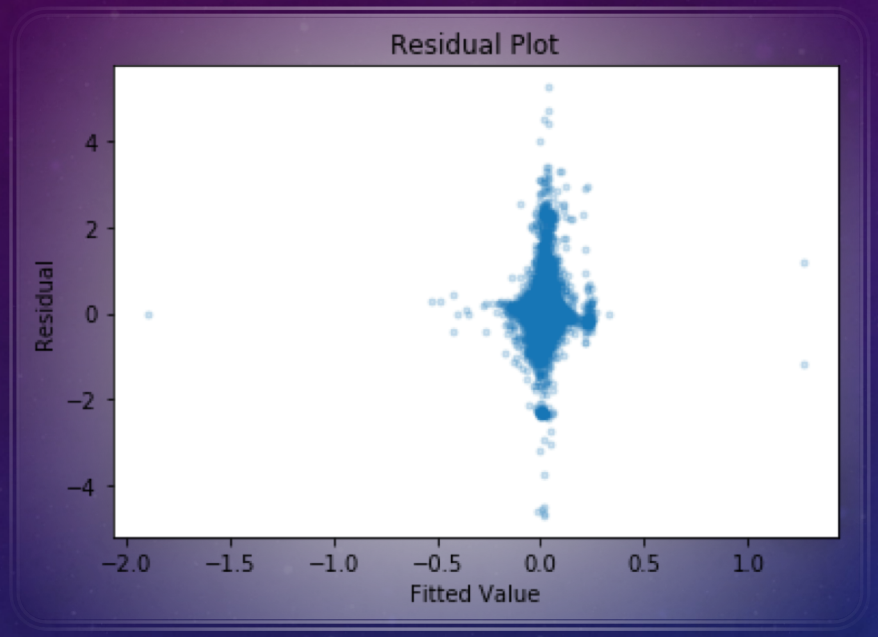
3. EXPLORATORY ANALYSIS OTHER

- Transaction Date (extract 3 features)
  - Months_Since_2015
    - A time series factor reflect error inflation
  - Weekday_7, Q_2
    - 2 categorical variables
    - Homes transacted on Sunday or during 2nd quarter are underestimated





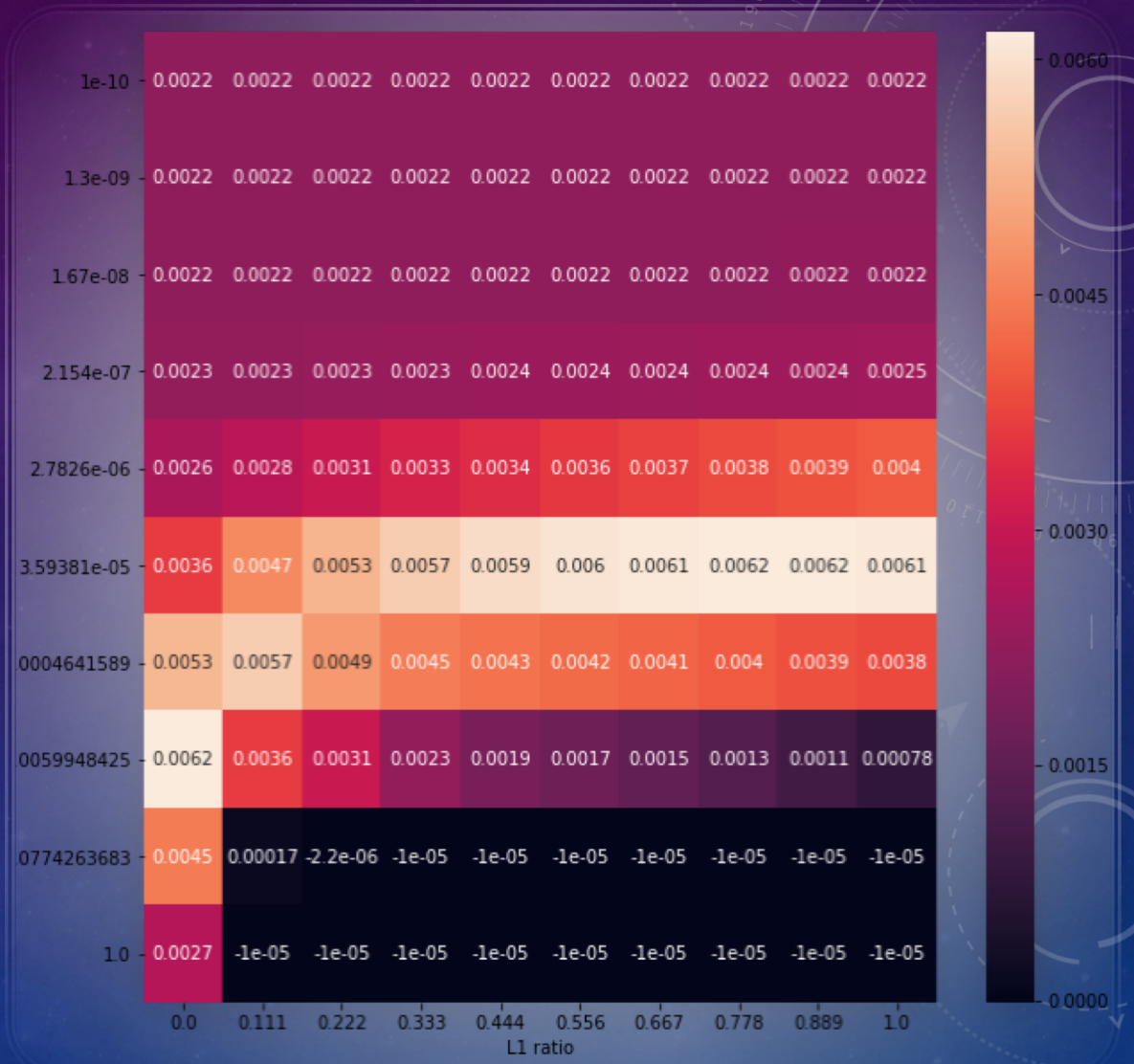
3. EXPLORATORY ANALYSIS PUT ALL TOGETHER

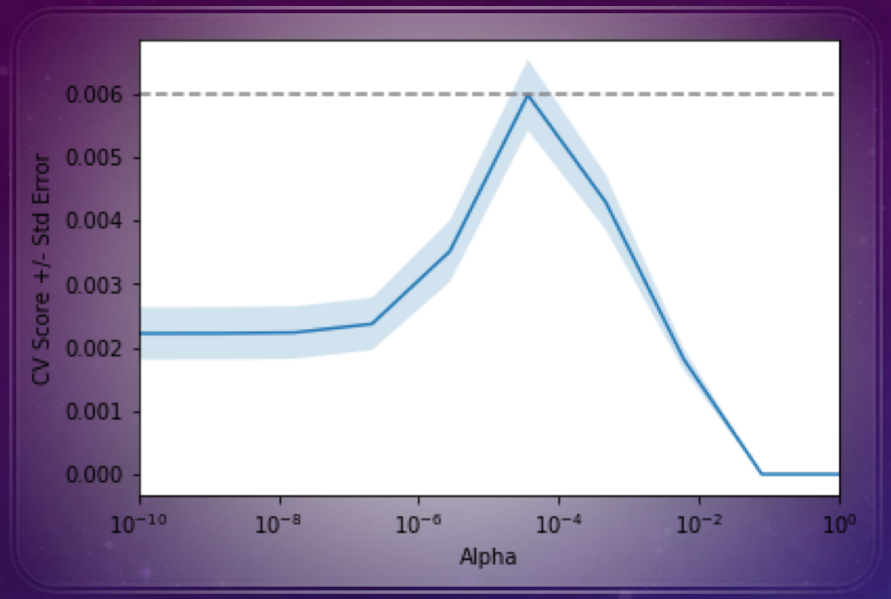


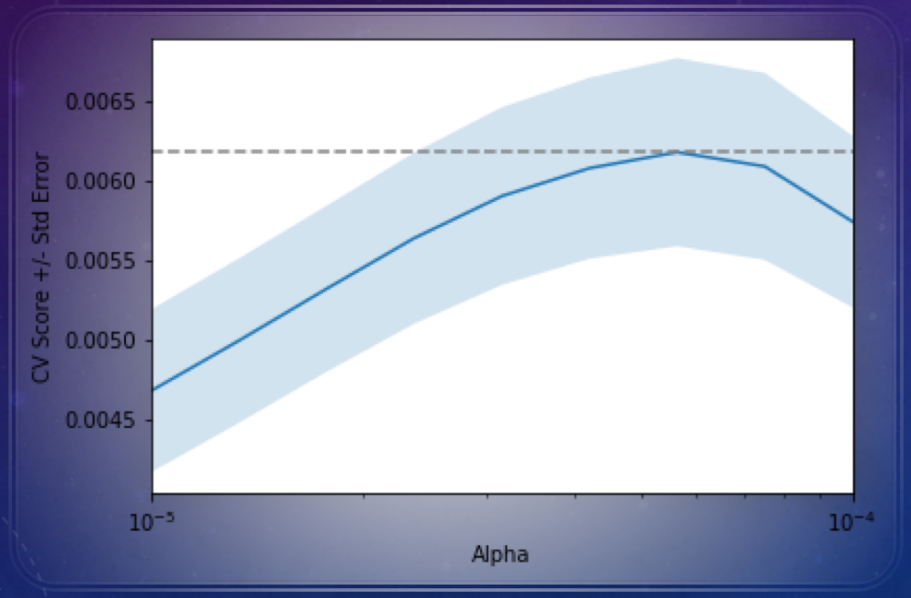
- OLS w/ all numerical & categorical parameters (621 in total)
- No significant trend can be identified
- Hard to improve significantly w/o external data

4. MACHINE LEARNING ANALYSIS

- Split into train & test sets (70% vs 30%)
- Stick to OLS for the nature of problem
- Add L1 and L2 regularization (Elastic Net)
  - Hyperparameter tuning: Alpha & L1 Ratio
  - 2-D Heat Map (took 17 hrs!)
  - Higher sensitivity to Alpha







4. MACHINE LEARNING ANALYSIS

- Alpha Tuning
  - We are more sensitive to Alpha
  - Only 10 values from 0 to 10^{-10} in a 10x10 grid
  - Further Tuning give us optimized Alpha = $4.22 * 10^{-5}$
  - Best R^2 score is around .0061

5. RESULT & CONCLUSION

- Conclusions
  - Found 25 parameters that can statistically help avoid mispricing
  - In most cases, Zestimate is a good estimate of home values
  - Outliers are almost inevitable by using current data
- Suggestions
  - Look into the largest outliers and find the exact reasons of over- and under-valuation
  - Consider adding new features based on investigations
  - Encourage users to provide more information and avoid missing values