California Housing Prices

Using Data Science to Determine Key Housing Values Indicators

What Is this Data Set?

- Kaggle housing data
- Median house prices for California districts derived from the 1990 census.
- “although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.”

But weren’t you going to do a music data set?

Challenges with the data

What have we gotten ourselves into

- The data was really dirty
- Missing values in some fields
- No data dictionary
- Random Categorical field

```
longitude             20640
latitude              20640
housing_median_age    20640
total_rooms           20640
total_bedrooms        20433
population            20640
households            20640
median_income         20640
median_house_value    20640
ocean_proximity       20640
dtype: int64
```

Data Scrubbing

It’s actually pretty cathartic

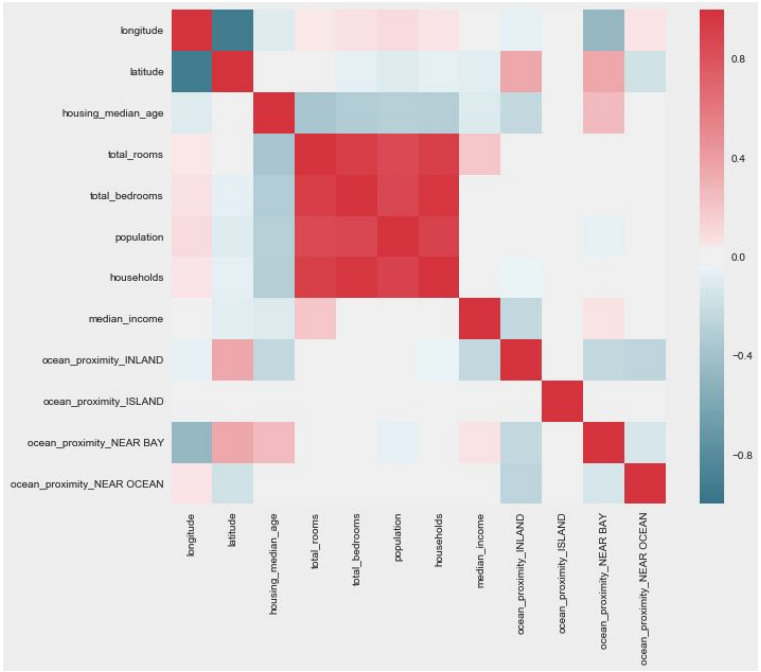
Mr. Clean, Mr. Clean

- First I got rid of any null or 0 values
- Then I imputed the mean
- Then we (Nico helped) figure out what to do with that Ocean_Proximity column
    - Some of it isn't even ocean related

Data Analysis

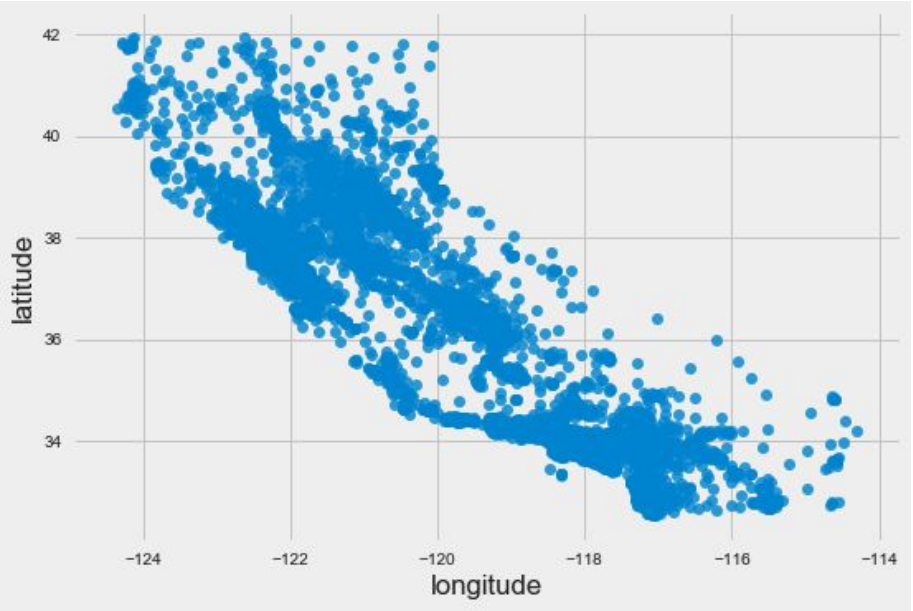
Exploratory Analysis

- data.describe() to view sample values
- Train_test_split to prepare for analysis
- Correlation map to test for multicollinearity



Machine Learning

- Trial and error - trying different values for features
- Regression not Classification (not binary, looking to predict a number)
  - I tried both - not going to lie
- MSE = .55



Learnings

- LOCATION, LOCATION, LOCATION
- Most important feature is Longitude
- The closer it was to water the higher the value of housing
    - This isn't that surprising, but now it's data backed
- Outside of location and proximity to water, the only other feature that mattered was house median age.

Next Steps

Ensemble Regression Trees

Determine key areas in CA for the best purchasing decision

Questions?

BRITNEE FOREMAN DAT-NYC 11.14.17