Doing More with Less: Achieving Near State-of-the-Art Results on Real Estate Valuations

Taylor Kilian
taylor.kilian@gmail.com

Presentation Outline

1. **Introduction and Motivation**
2. Target Datasets
    a. Airbnb prices (rentals)
    b. Assessor values (property sales)
3. Feature Addition to Datasets
    a. Geospatial data
    b. Vectorization method
4. Machine Learning Models
5. Model Evaluation
6. Conclusions

Project Motivation

$30,000,000,000,000 (trillion)

- Total U.S. housing stock
- Real estate is the largest tangible asset in U.S.

$470,000,000,000 (billion)

- Total amount renters paid for housing in 2016

Establish a **fair market price**:

- The renter/buyer doesn’t overpay
- Seller/lessor can find new renter/buyer quickly
- Affects almost every citizen in the nation, and can better aid investment

Presentation Outline

1. Introduction and Motivation
2. **Target Datasets**
    a. **Airbnb prices (rentals)**
    b. **Assessor values (property sales)**
3. Feature Addition to Datasets
    a. Geospatial data
    b. Vectorization method
4. Machine Learning Models
5. Model Evaluation
6. Conclusions

Target Datasets

Standard Features (Baseline Model)

Airbnb Dataset (~5900 listings):

- 25 features
  - Beds, baths, capacity, availability, review scores, reviews per month, etc.

- 58 more features
  - Categorically encoded neighborhoods and housing type

- Ignored text features
  - Save review text for later

Assessor Dataset (~300,000 properties):

- 8 features
  - Beds, bath, rooms, stories, units, area (sqft), percent ownership, and property age

Target Datasets (add Geospatial data)

Baseline	Geospatial Data

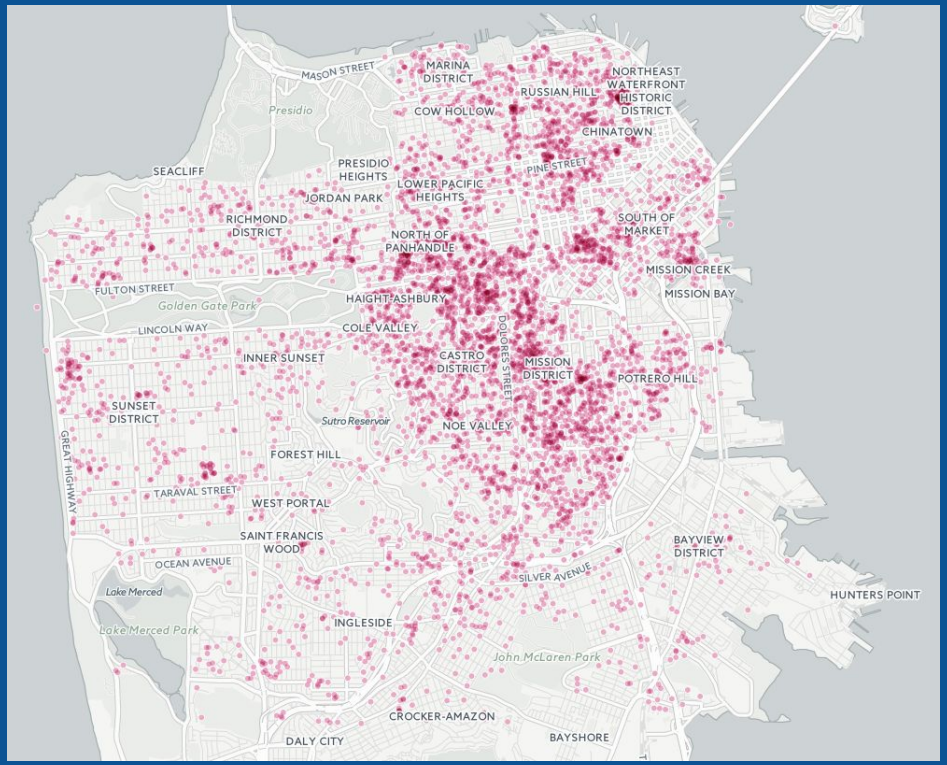
Airbnb
 	
83 features
	- Income and poverty levels - 2 features
	  - By SF census tracts
	- Crime Incidents - 9 features
	  - Larceny/theft, vehicle theft, drug/narcotic, vandalism, burglary, robbery, loitering, non-criminal, weapon laws
	- 311 requests - 7 features
	  - Non-emergency complaints or requests
	  - Street and sidewalk cleaning, damaged property, graffiti, street defects, sidewalk or curb, color curb, sfha requests
	- Parks - 1 feature
	  - Proximity to parks
	- Noise - 5 features
	  - Proximity to loud areas

SF Assessor 8 features	

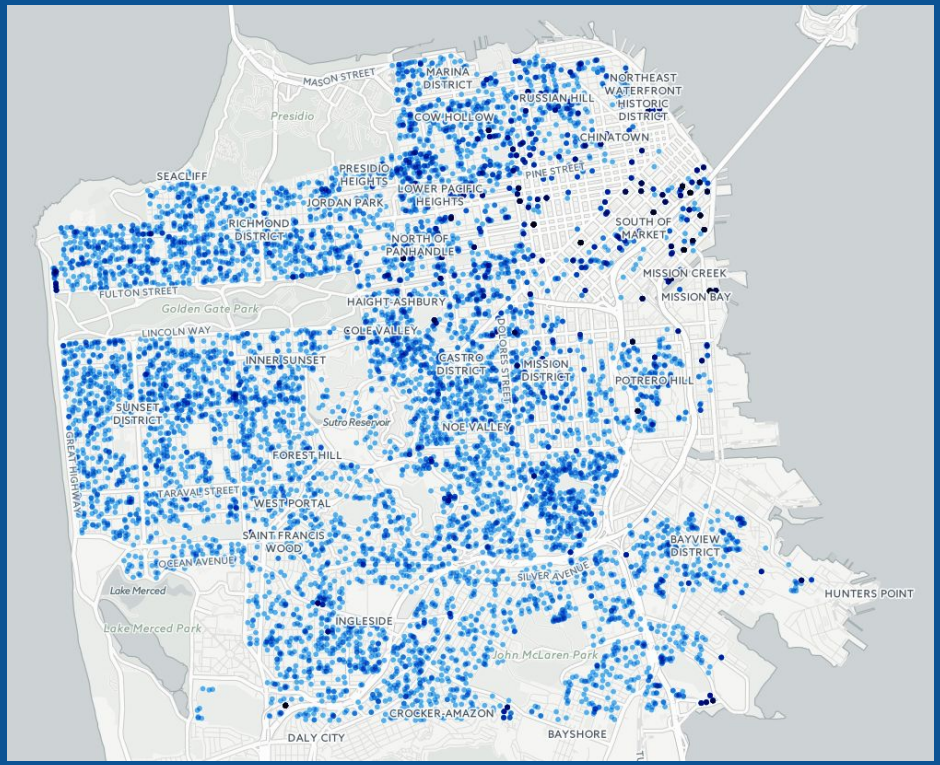
Target Datasets (add Text and Image Data)

Baseline	Geospatial Data	Text	Images
Airbnb 83 features	- Income and poverty levels - 2 features - By SF census tracts - Crime Incidents - 9 features - Larceny/theft, vehicle theft, drug/narcotic, vandalism, burglary, robbery, loitering, non-criminal, weapon laws - 311 requests - 7 features - Non-emergency complaints or requests - Street and sidewalk cleaning, damaged property, graffiti, street defects, sidewalk or curb, color curb, sfha requests - Parks - 1 feature - Proximity to parks - Noise - 5 features - Proximity to loud areas	- Tf-idf vectorization of Airbnb reviews over a year - Sentiment analysis of Airbnb reviews	- Use listing thumbnail images to predict residuals
SF Assessor 8 features		- No text available	- Incorporate Google Street View images into model

Target Maps

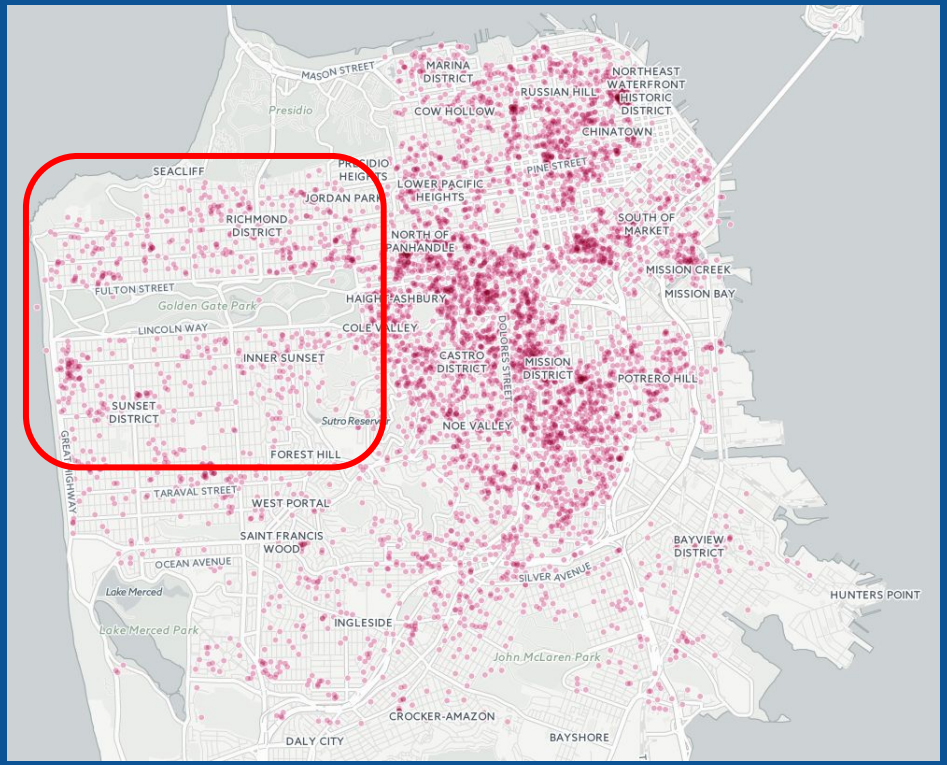


Airbnb Locations
~5,900 listings

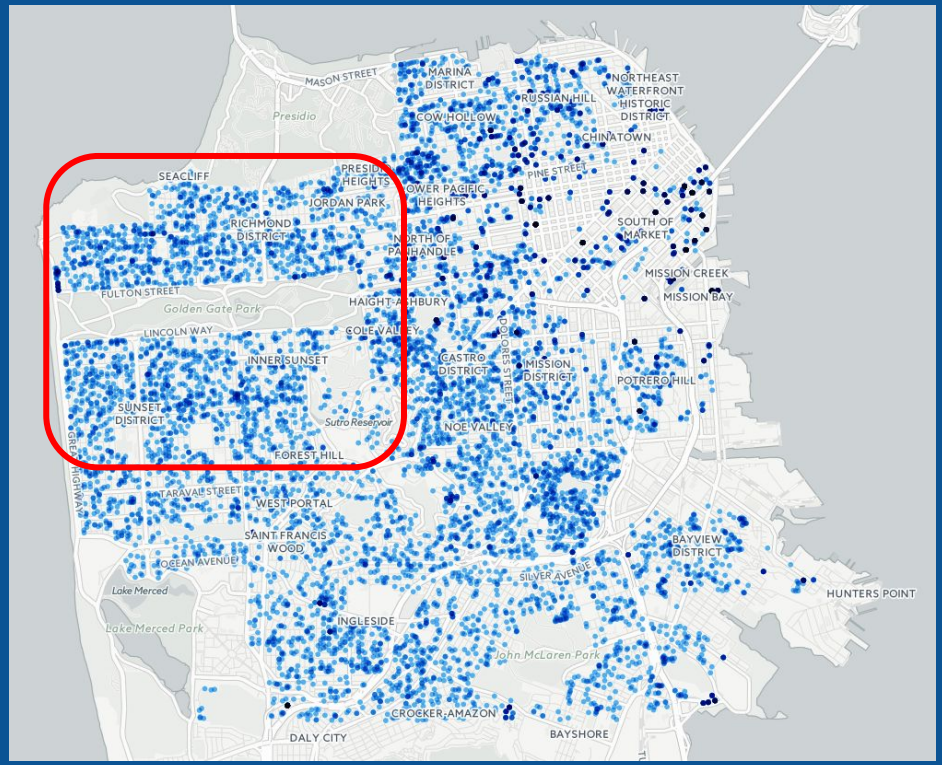


Property (Assessor) Locations
10,000 properties (small sample)

Target Maps

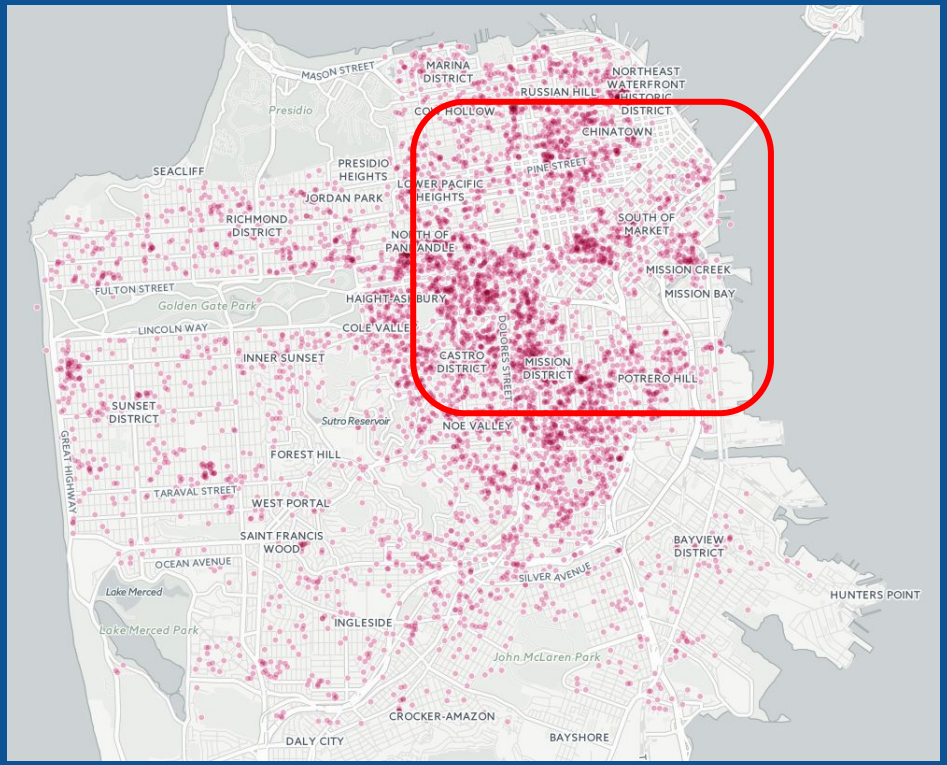


Airbnb Locations
~5,900 listings

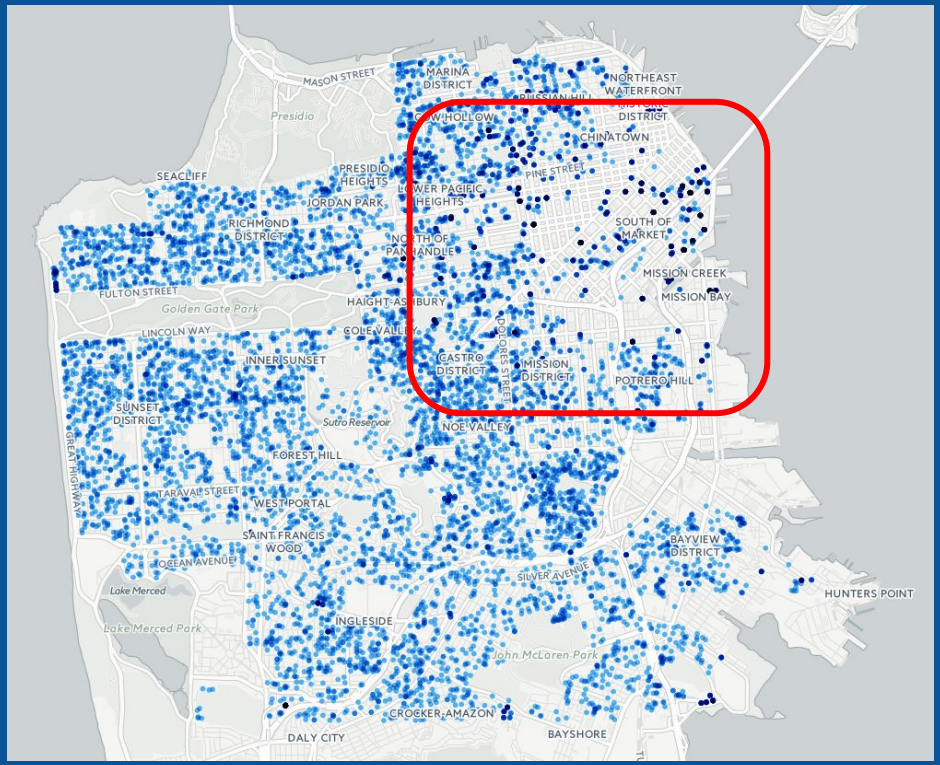


Property (Assessor) Locations
10,000 properties (small sample)

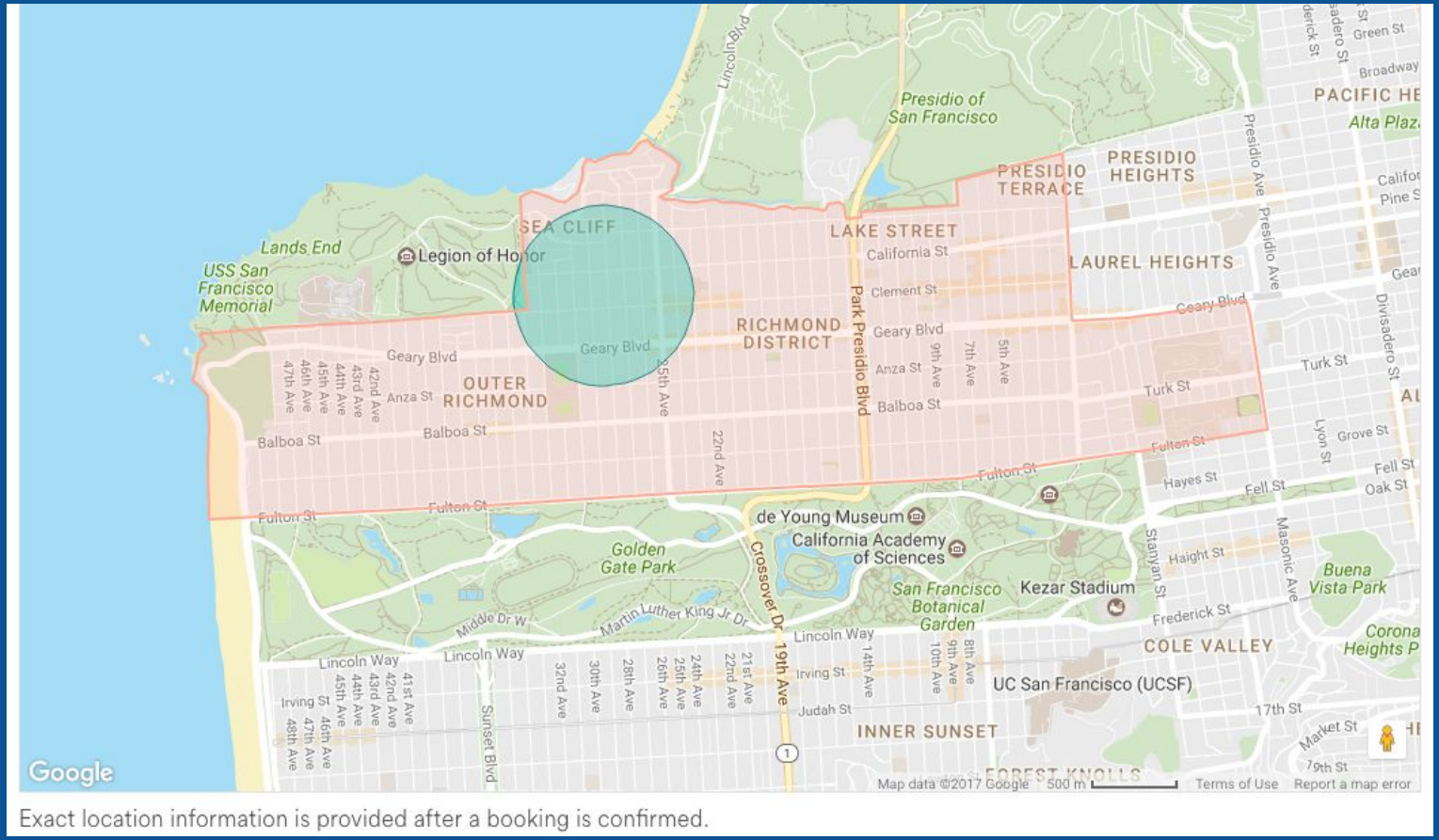
Target Maps



Airbnb Locations
~5,900 listings



Property (Assessor) Locations
10,000 properties (small sample)

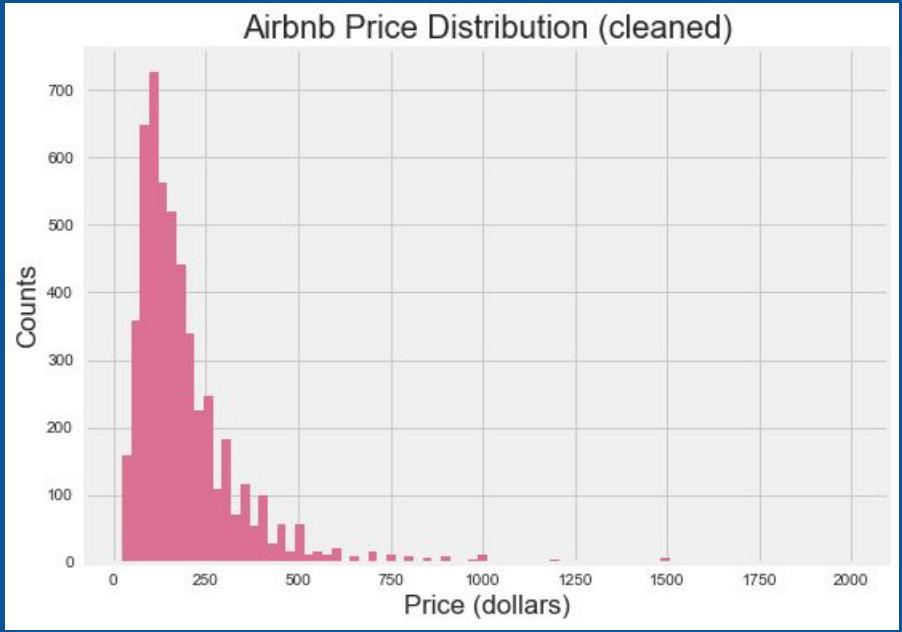


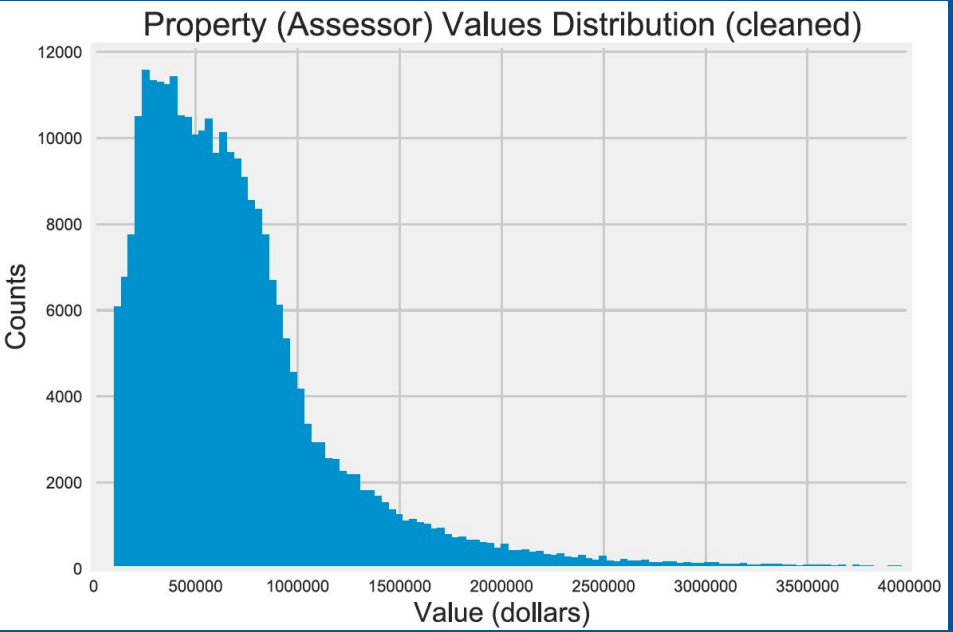
Exact location information is provided after a booking is confirmed.

Airbnb anonymizes rental locations, making geospatial relationships more ambiguous

- This might create meaningless geospatial features

Target Distributions





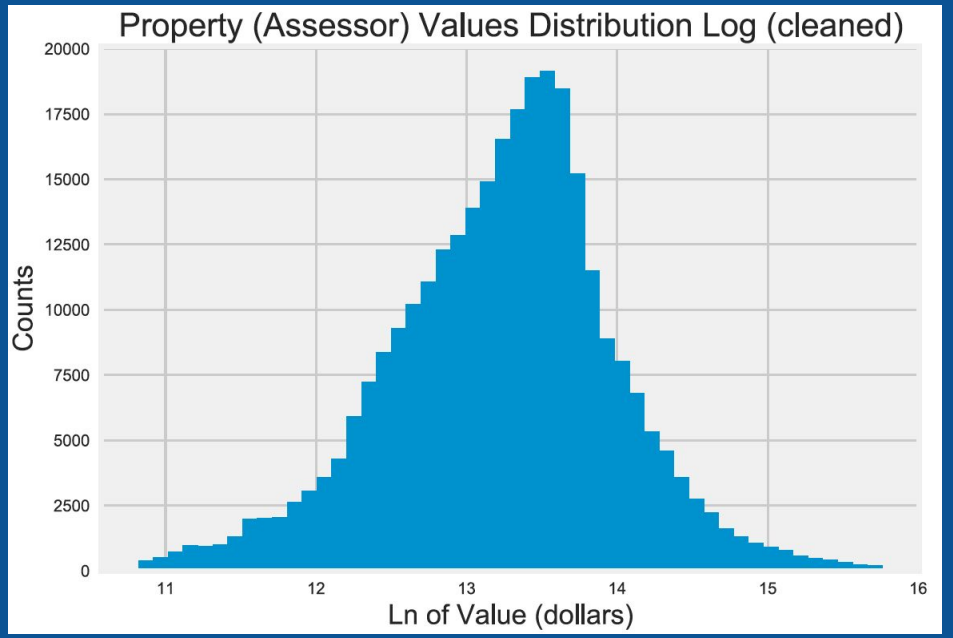
Airbnb Prices

Units with > 0.2 reviews/month

Property Assessor Valuations

Total value between $100,000 and $7 million

Log Transformed Distributions

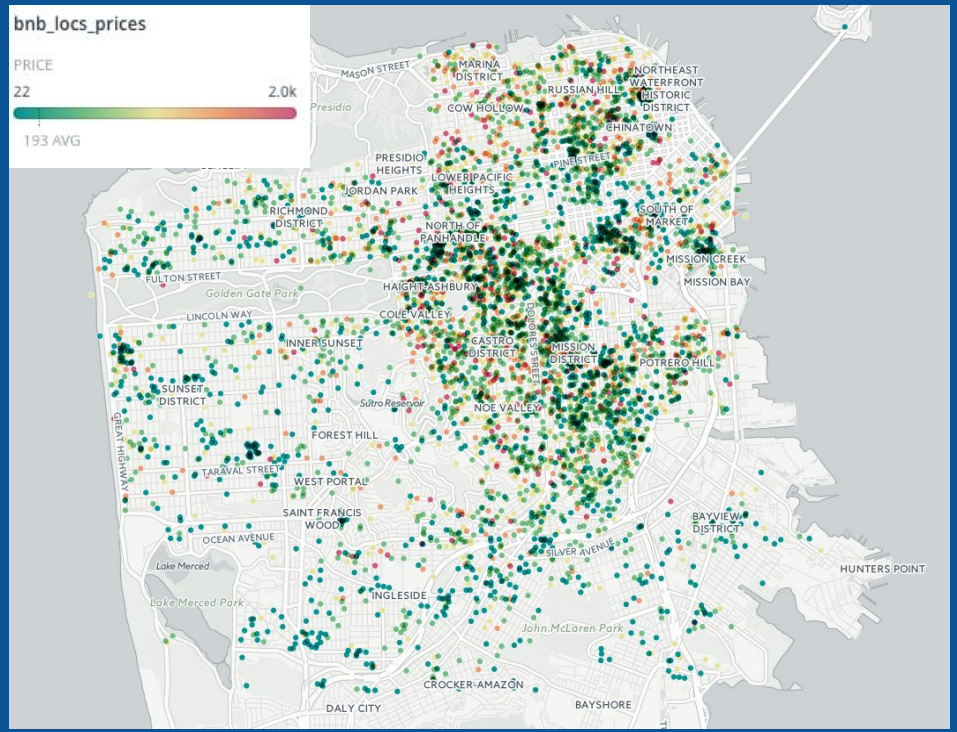


Airbnb Prices

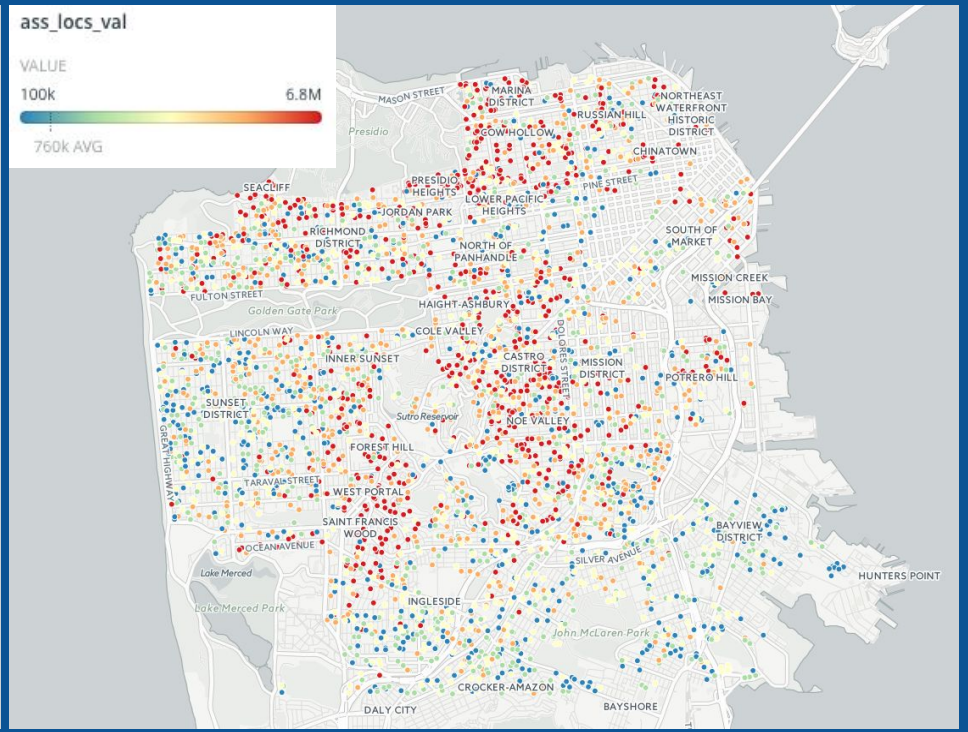
Property Assessor Valuations

Transformed to improve their model predictability

Target Spatial Distribution - Mapped by Price/Value



Airbnb listings shaded by price

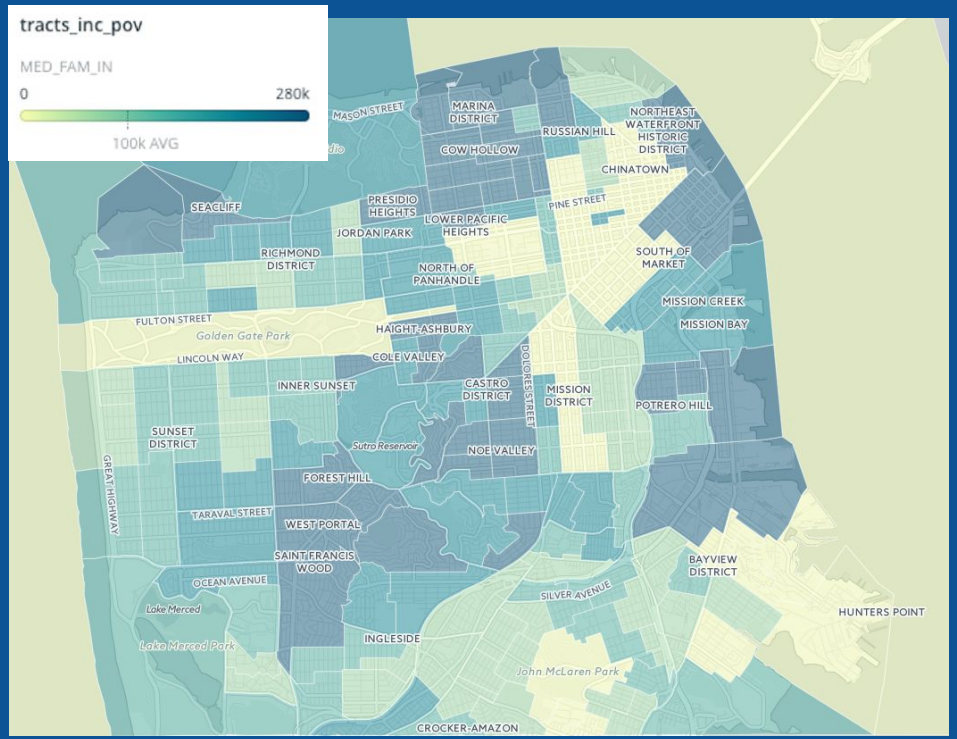


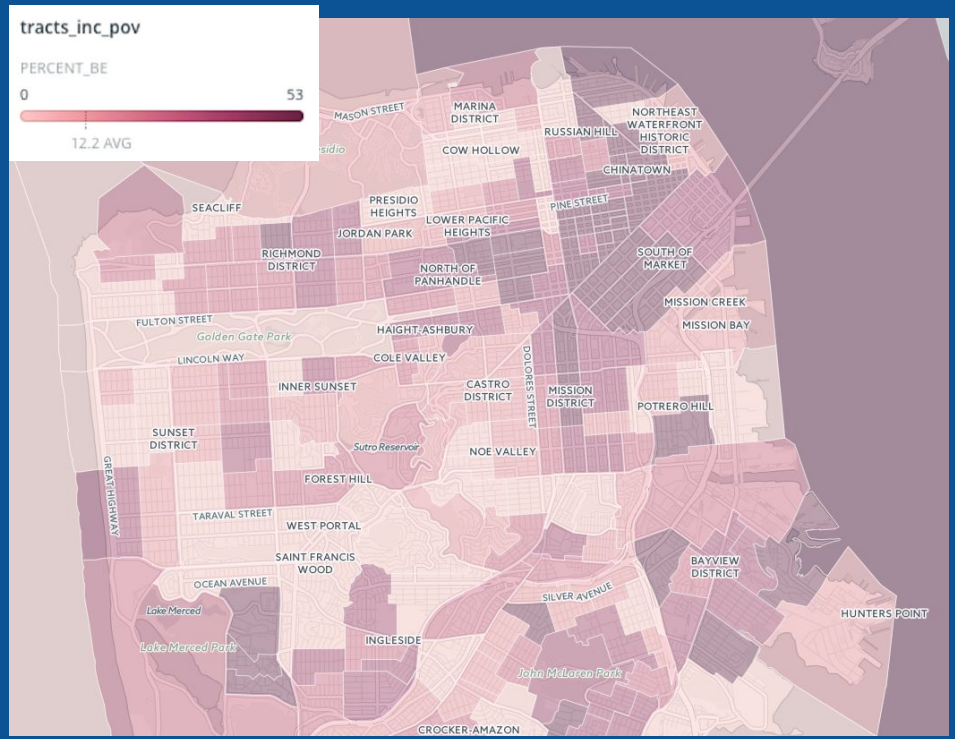
Assessor properties shaded by value

Presentation Outline

1. Introduction and Motivation
2. Target Datasets
    a. Airbnb prices (rentals)
    b. Assessor values (property sales)
3. **Feature Addition to Datasets**
    a. **Geospatial data**
    b. Vectorization method
4. Machine Learning Models
5. Model Evaluation
6. Conclusions

Geospatial Data - Income/Poverty%



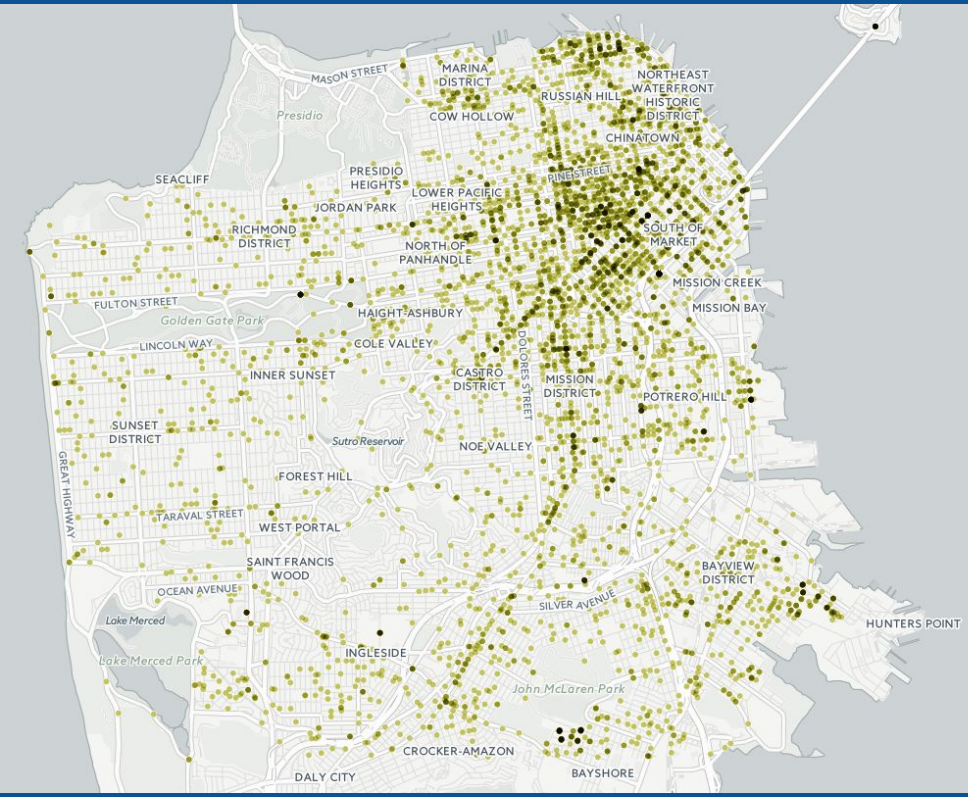


Census Tracts - median family income

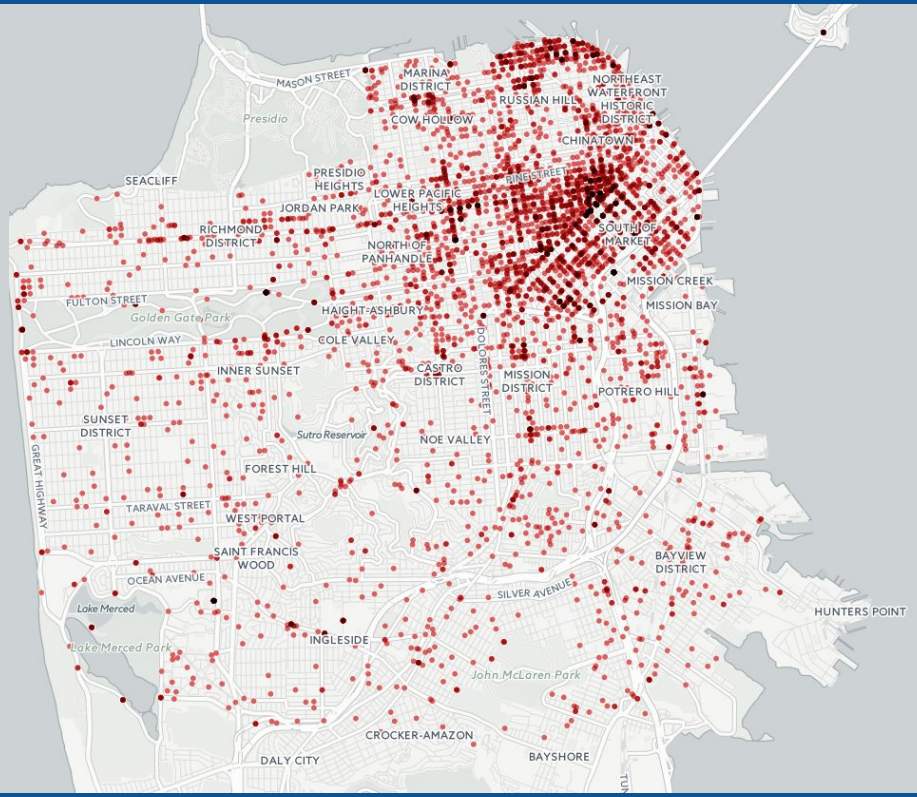
Census Tracts - % below poverty level

Geospatial Data - Crime Incidents

Crime - Vandalism



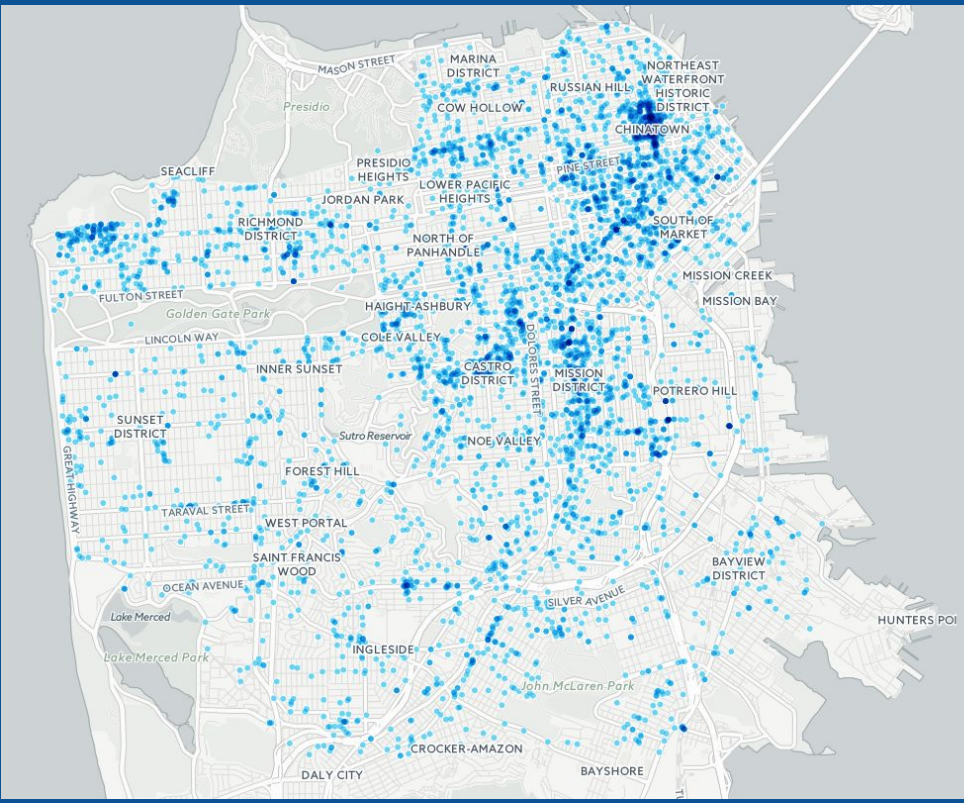
Crime - Larceny/ Theft



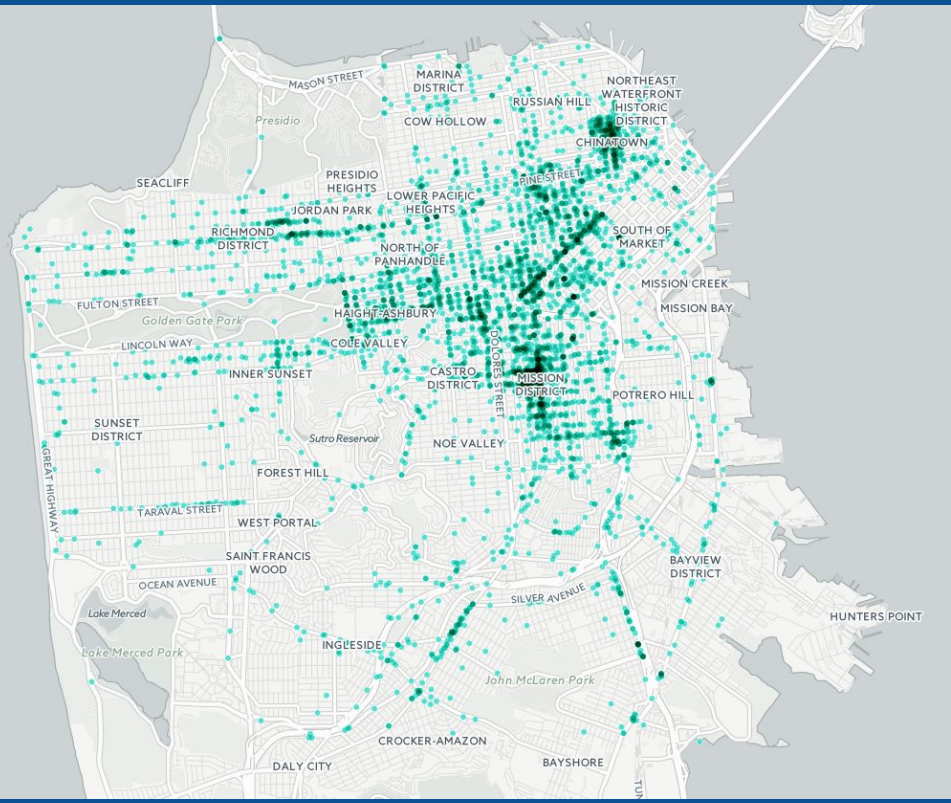
- Many categories available - subjectively chose 9 groups to use in model

Geospatial Data - 311 Requests

311 Requests - Sidewalk and Curb Repair



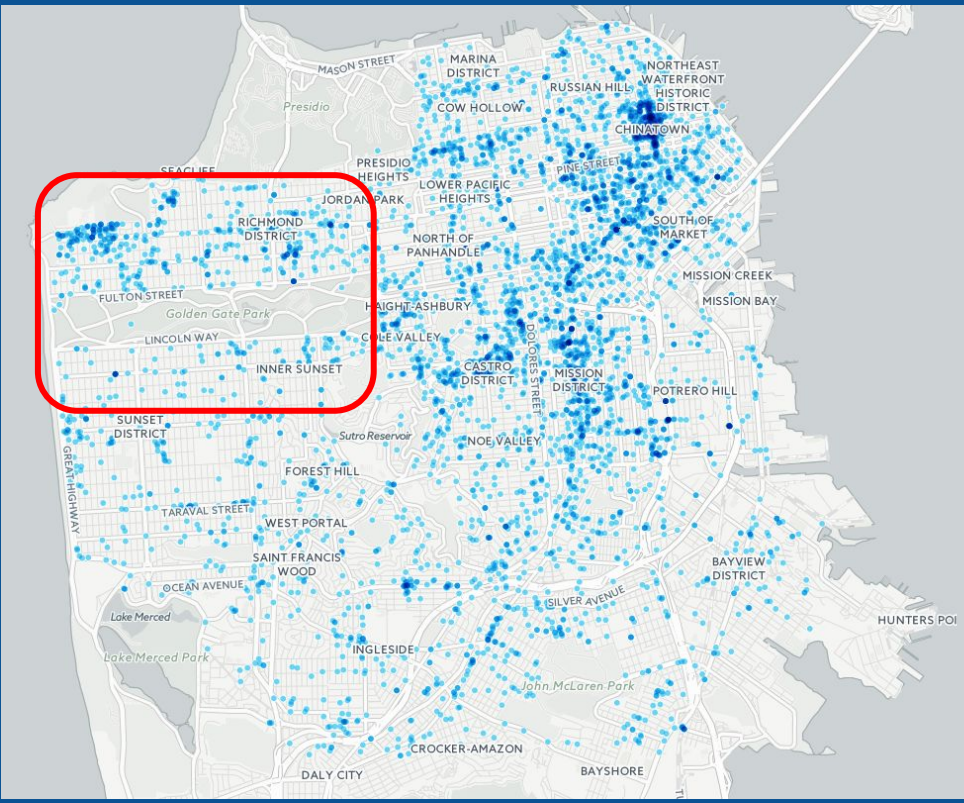
311 Requests - Graffiti



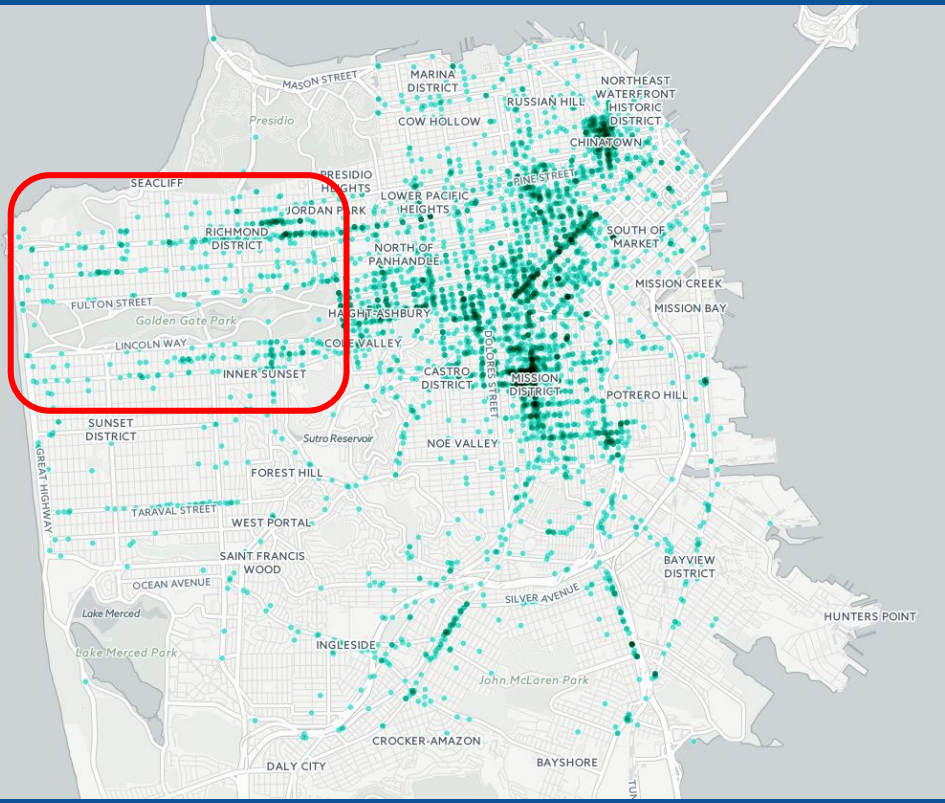
- Many categories available - subjectively chose 7 groups to use in model

Geospatial Data - 311 Requests

311 Requests - Sidewalk and Curb Repair



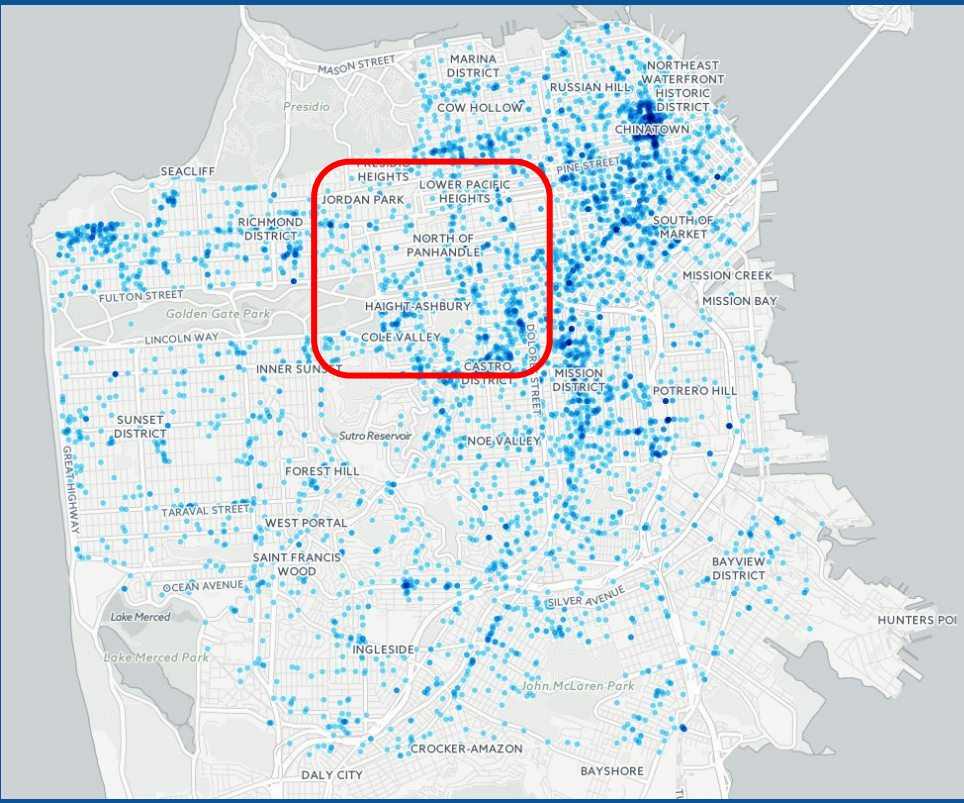
311 Requests - Graffiti



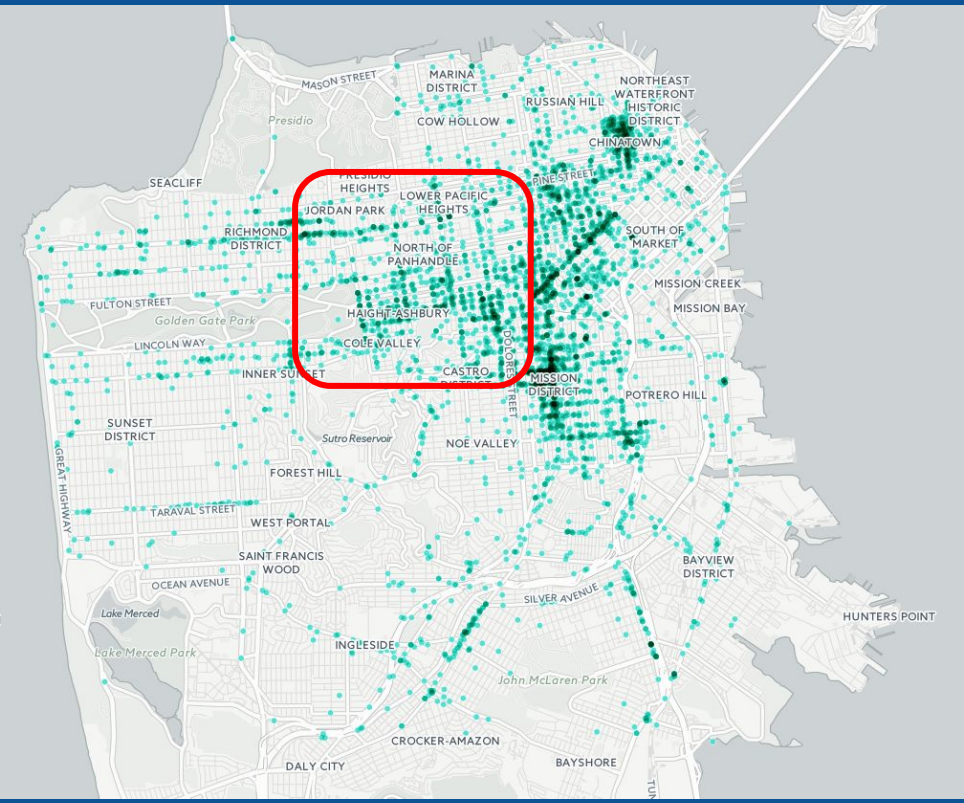
- Many categories available - subjectively chose 7 groups to use in model

Geospatial Data - 311 Requests

311 Requests - Sidewalk and Curb Repair



311 Requests - Graffiti

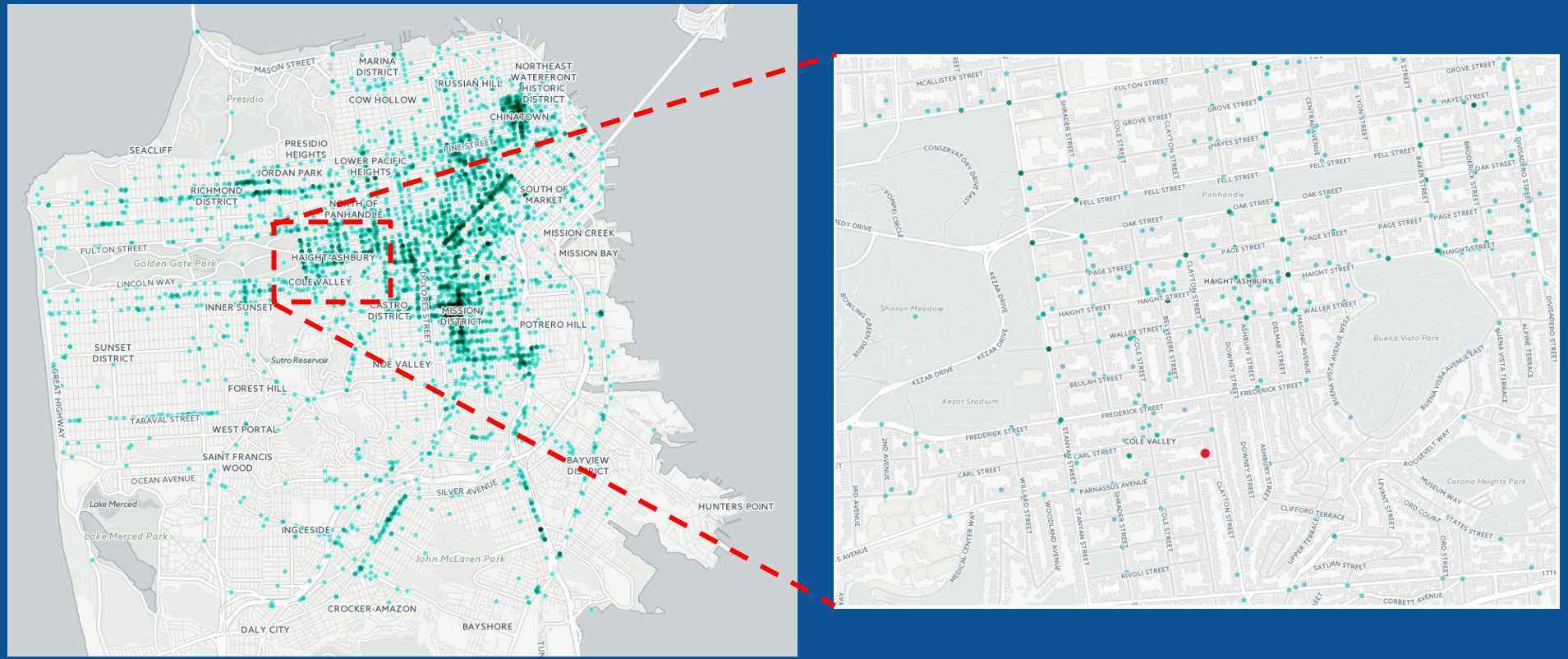


- Many categories available - subjectively chose 7 groups to use in model

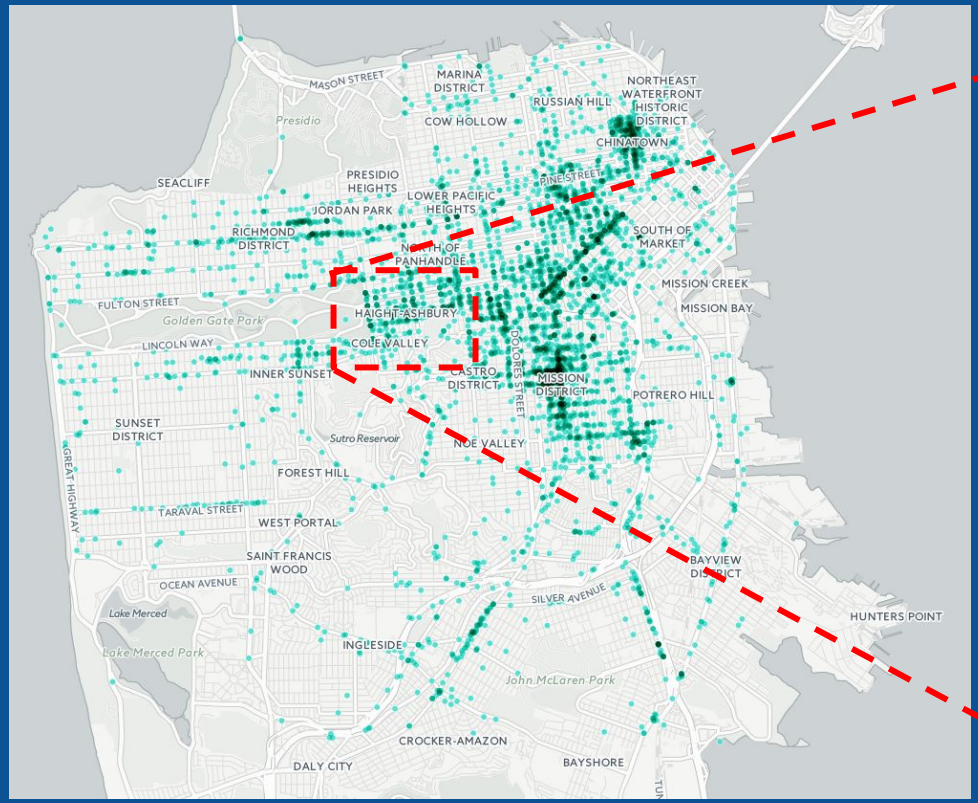
Presentation Outline

1. Introduction and Motivation
2. Target Datasets
    a. Airbnb prices (rentals)
    b. Assessor values (property sales)
3. **Feature Addition to Datasets**
    a. Geospatial data
    b. **Vectorization method**
4. Machine Learning Models
5. Model Evaluation
6. Conclusions

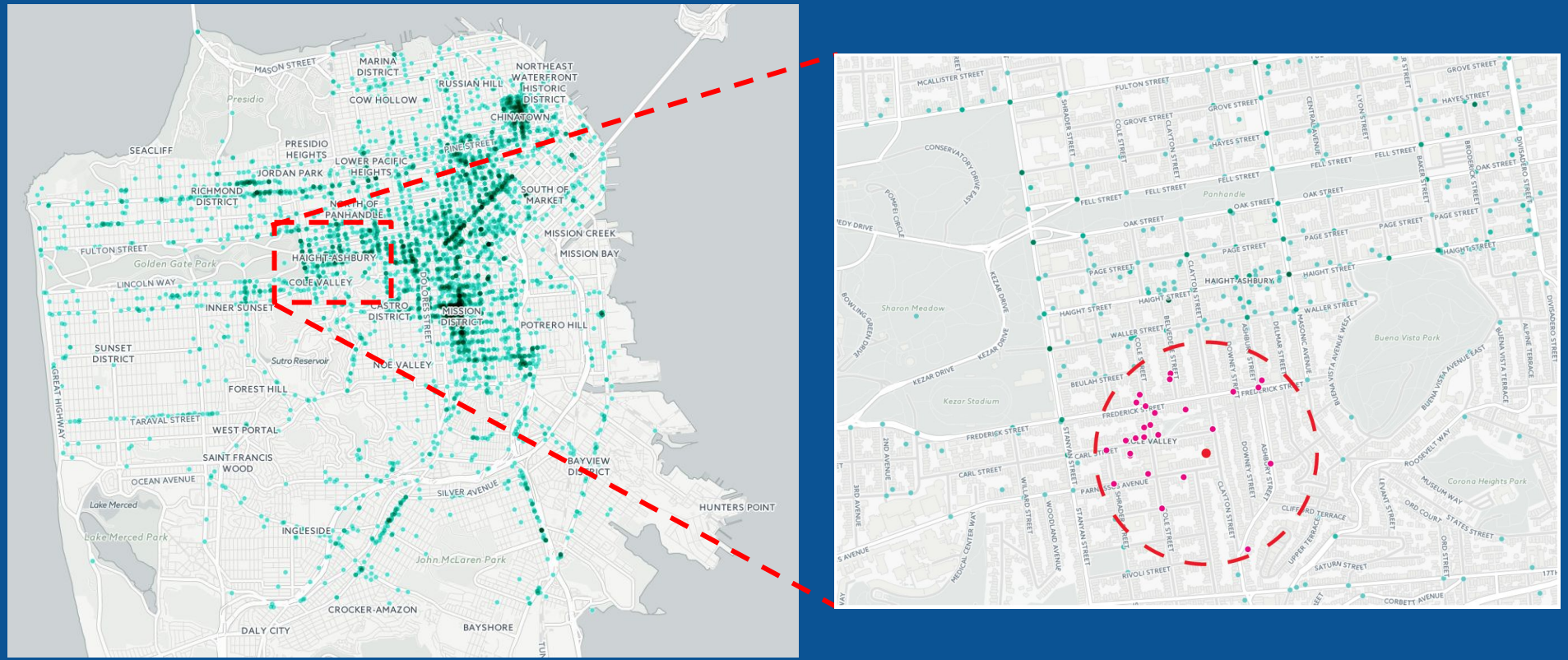
Geospatial Data - Vectorization Method



Geospatial Data - Vectorization Method

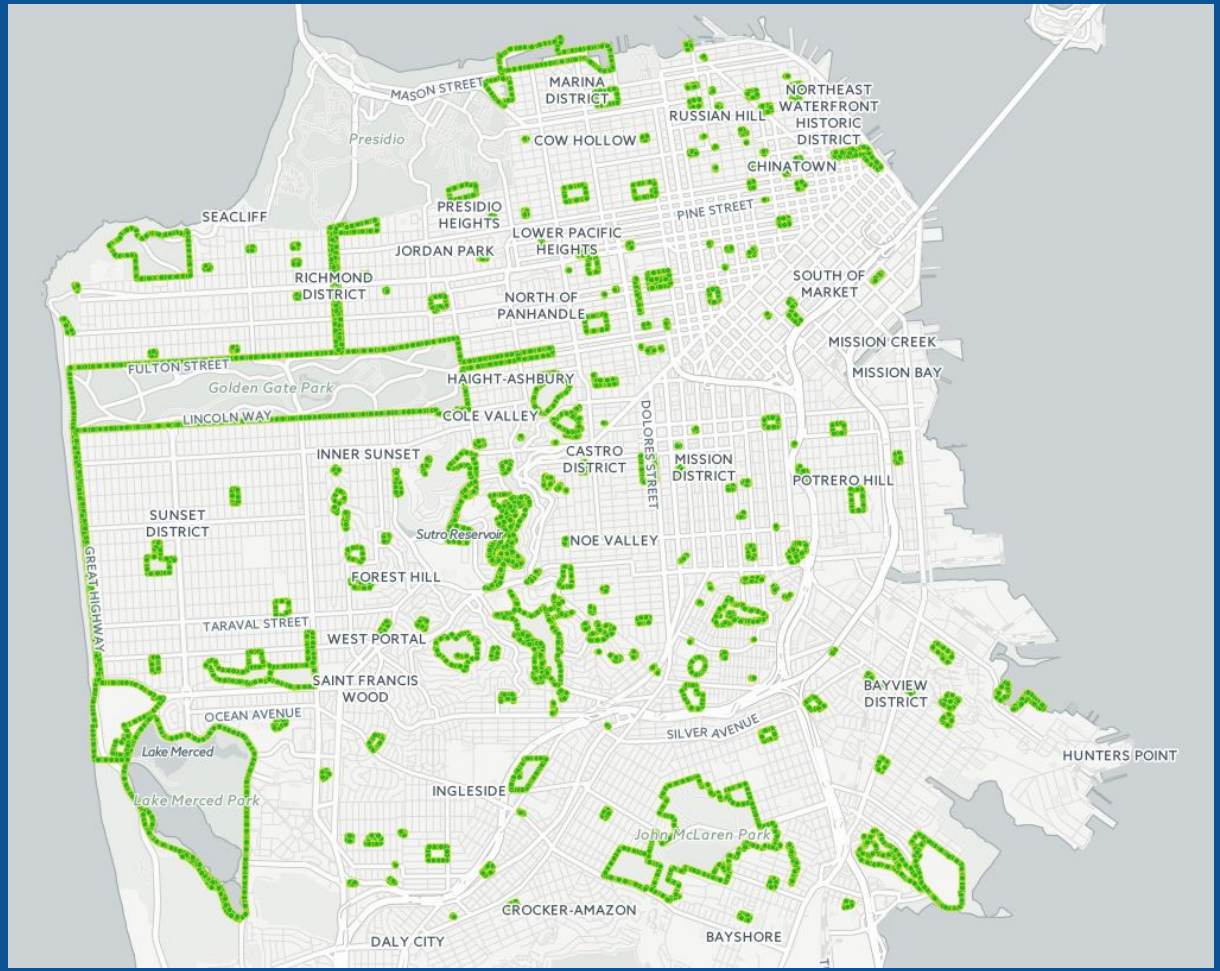


Geospatial Data - Vectorization Method

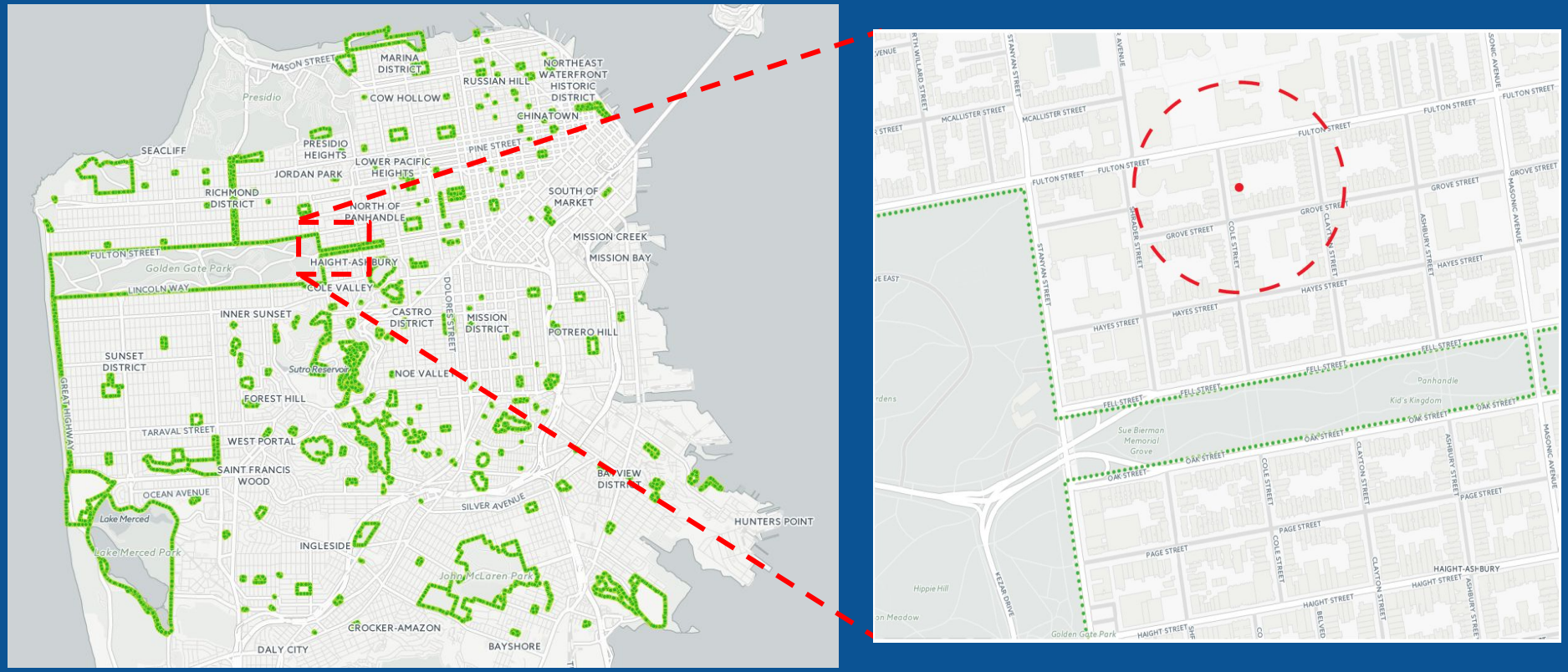


Geospatial Data - Parks

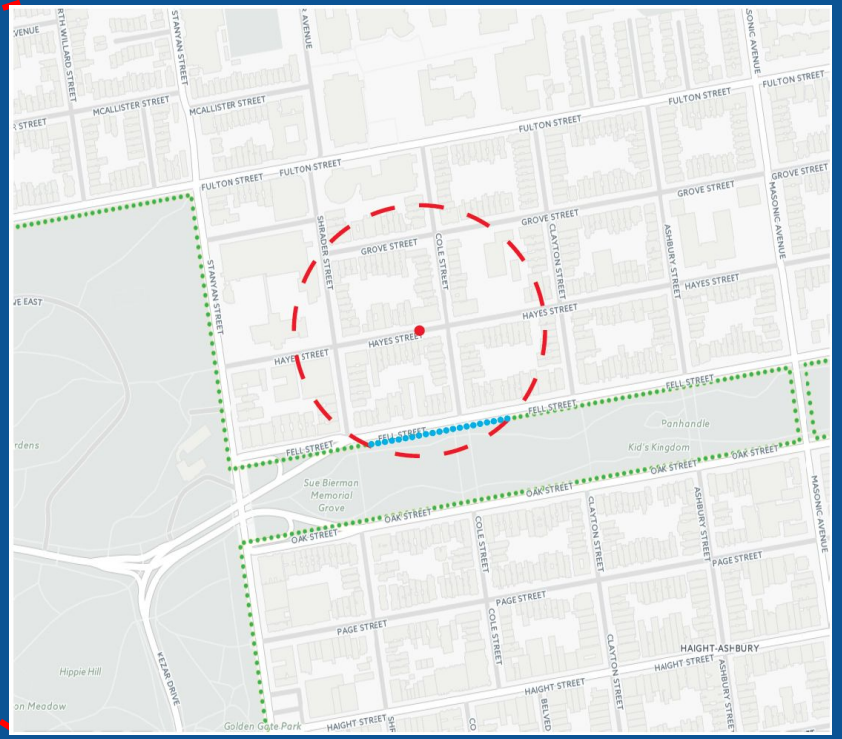
- Park shapefiles
  - Points were added with regular frequency along boundaries using QGIS
  - Use similar method to incorporate as feature into datasets



Geospatial Data - Vectorization Method



Geospatial Data - Vectorization Method



Geospatial Data - Noise Level

SF Dept. of Public Health dataset - parsed with QGIS, same method used for parks

76 Ldn (Loudest Noise)

72 Ldn (Lower Noise)

- 5 different noise levels, each added as feature using boundary counts (like parks)

Presentation Outline

1. Introduction and Motivation
2. Target Datasets
    a. Airbnb prices (rentals)
    b. Assessor values (property sales)
3. Feature Addition to Datasets
    a. Geospatial data
    b. Vectorization method
4. **Machine Learning Models**
5. Model Evaluation
6. Conclusions

Machine Learning Models

Both Airbnb and Assessor datasets tested with multiple models

- Linear regression (with lasso and ridge regularization)
- Ensemble methods:
  - Random Forests
  - Extremely Randomized Trees
  - Gradient Tree Boosting
  - XGBoost

- Chose best performing models for grid searches to find best model parameters (applying cross validation)

Model Scoring Metrics

Zillow Rent Estimate Model:

- We will used same scoring metric to compare results

Metro	Homes with Rent Zestimates	Within 5% of Rent Price	Within 10% of Rent Price	Within 20% of Rent Price	Within Zestimate Range	Median Error
United States	98,746,479	28.5%	49.7%	74.3%	74.6%	10.0%
Atlanta, GA	2,169,091	28.7%	51.8%	77.1%	73.4%	9.5%
Baltimore, MD	956,488	33.0%	53.2%	81.4%	73.6%	9.1%
Boston, MA	1,416,056	31.1%	50.5%	75.2%	75.1%	9.8%
Chicago, IL	3,207,001	27.4%	48.3%	74.8%	76.0%	10.3%
Cincinnati, OH	690,244	31.4%	50.7%	73.6%	74.7%	9.6%
Cleveland, OH	754,060	30.6%	50.1%	76.0%	75.5%	10.0%
San Francisco, CA	1,254,830	30.4%	53.1%	78.4%	74.5%	8.9%

Rent zestimates use percentage within range of actual price

- Median Error Percentage (ME%) - half of predictions are within ME% of true value

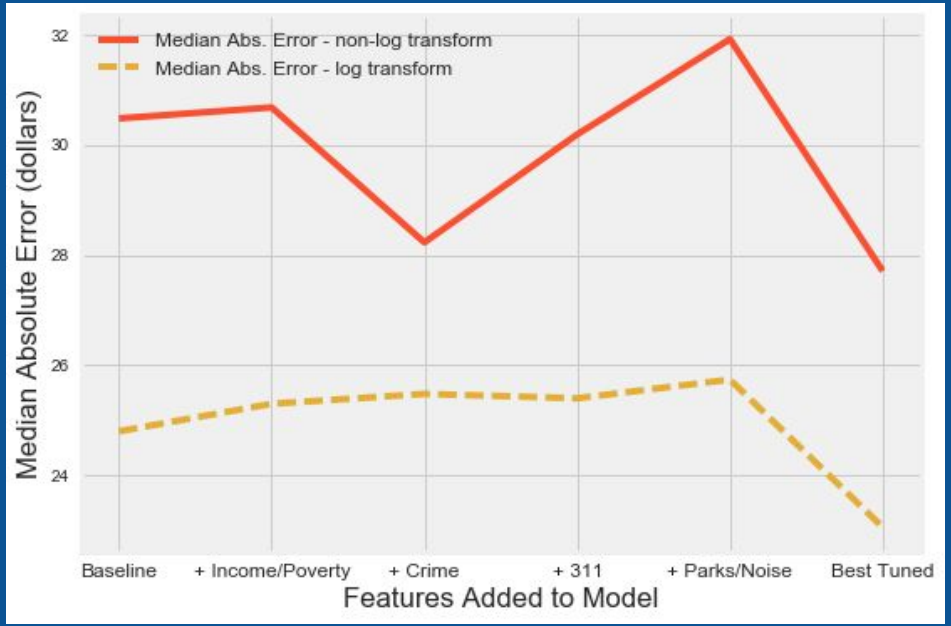
Also used Median Absolute Error - error in dollars

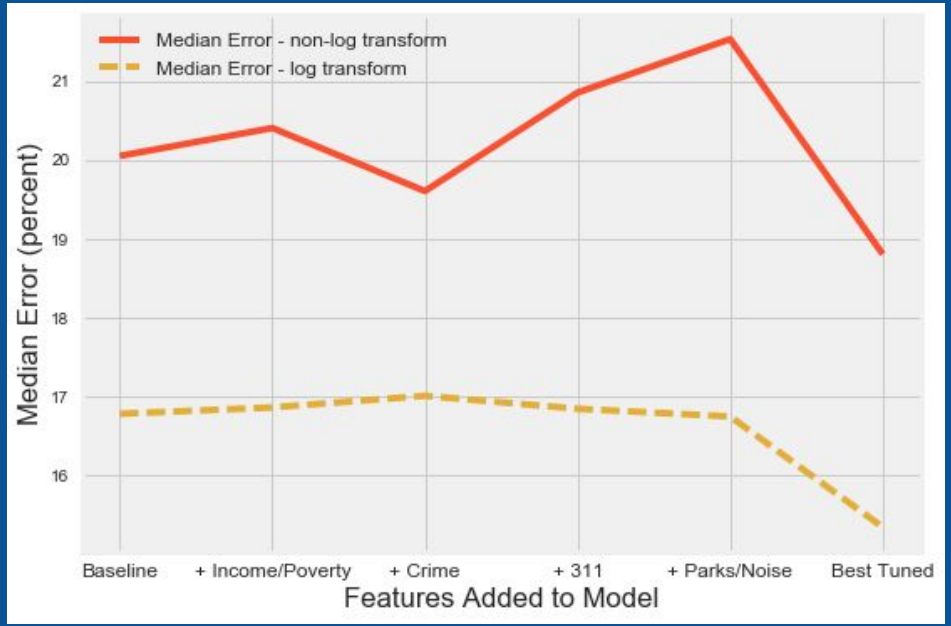
Presentation Outline

1. Introduction and Motivation
2. Target Datasets
    a. Airbnb prices (rentals)
    b. Assessor values (property sales)
3. Feature Addition to Datasets
    a. Geospatial data
    b. Vectorization method
4. Machine Learning Models
5. **Model Evaluation**
6. Conclusions

Airbnb Rent Prices - Model results

Progressively add geospatial data - what is effect on model error?





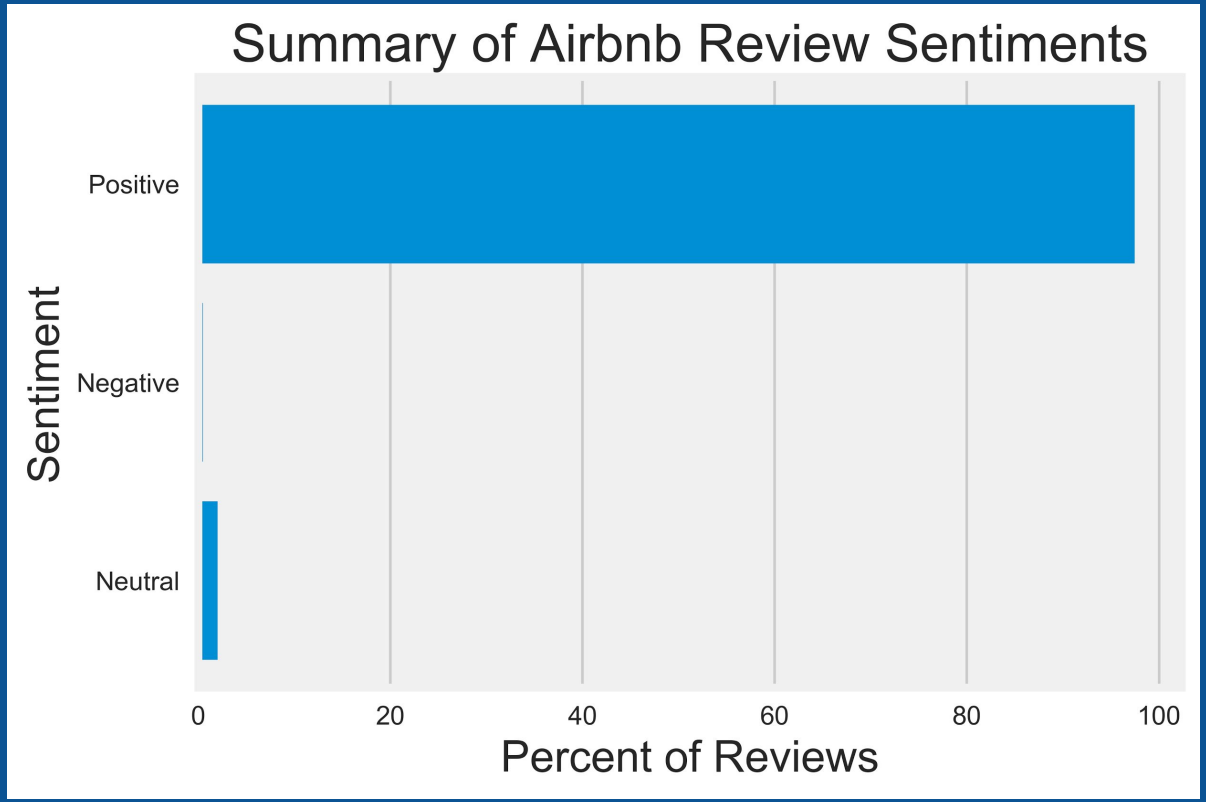
Anonymized location data make geospatial correlations meaningless

Airbnb Rent Prices - Text and Sentiment Analysis

NLP analysis of text from reviews does not enhance predictions

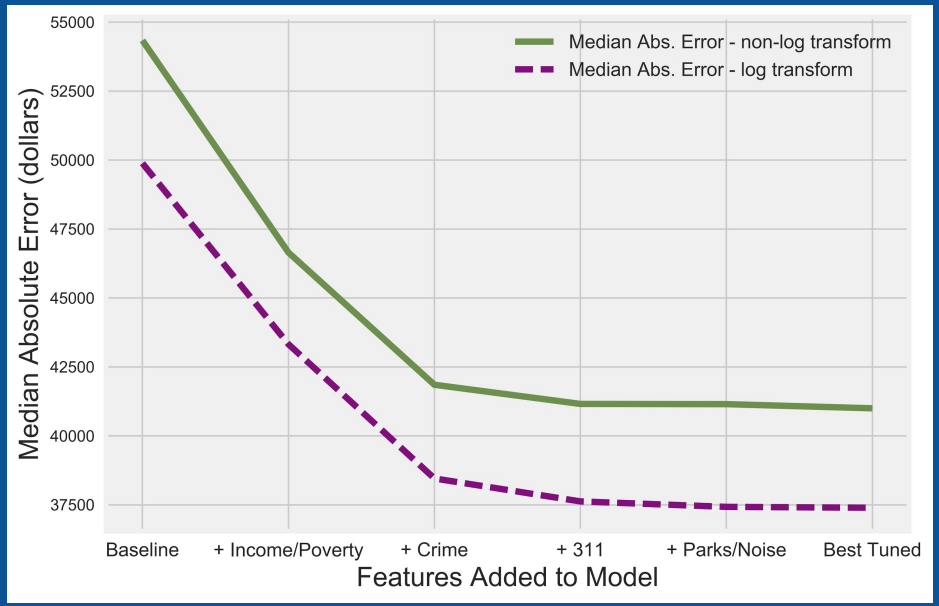
Reviews are inherently biased and overwhelmingly positive

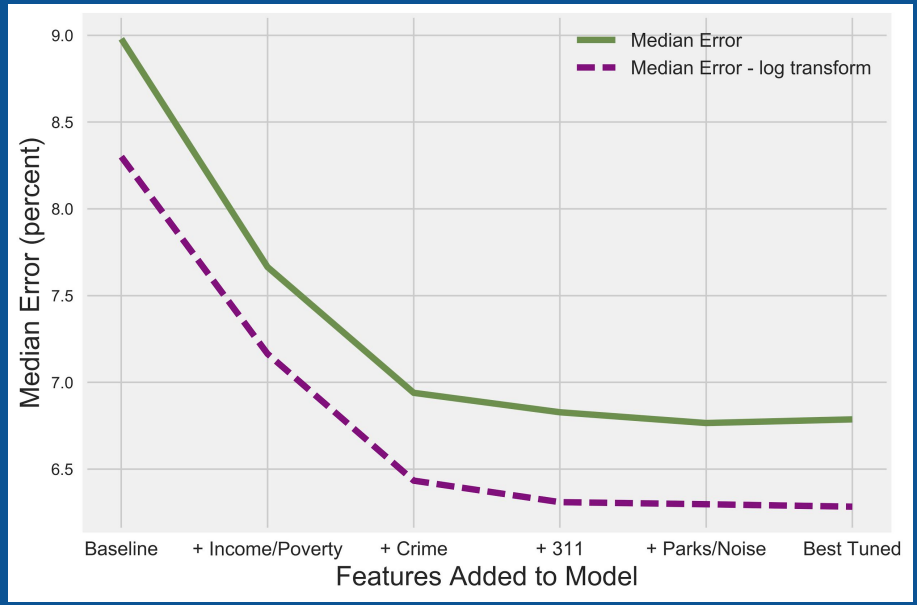
- 97.5% are positive
- 0.5% are negative
- 2.0% are neutral



Assessor Property Values

Progressively add geospatial data - what is effect on model error?





More precise location data

- Geospatial data contribute more when spatial resolution of target is high

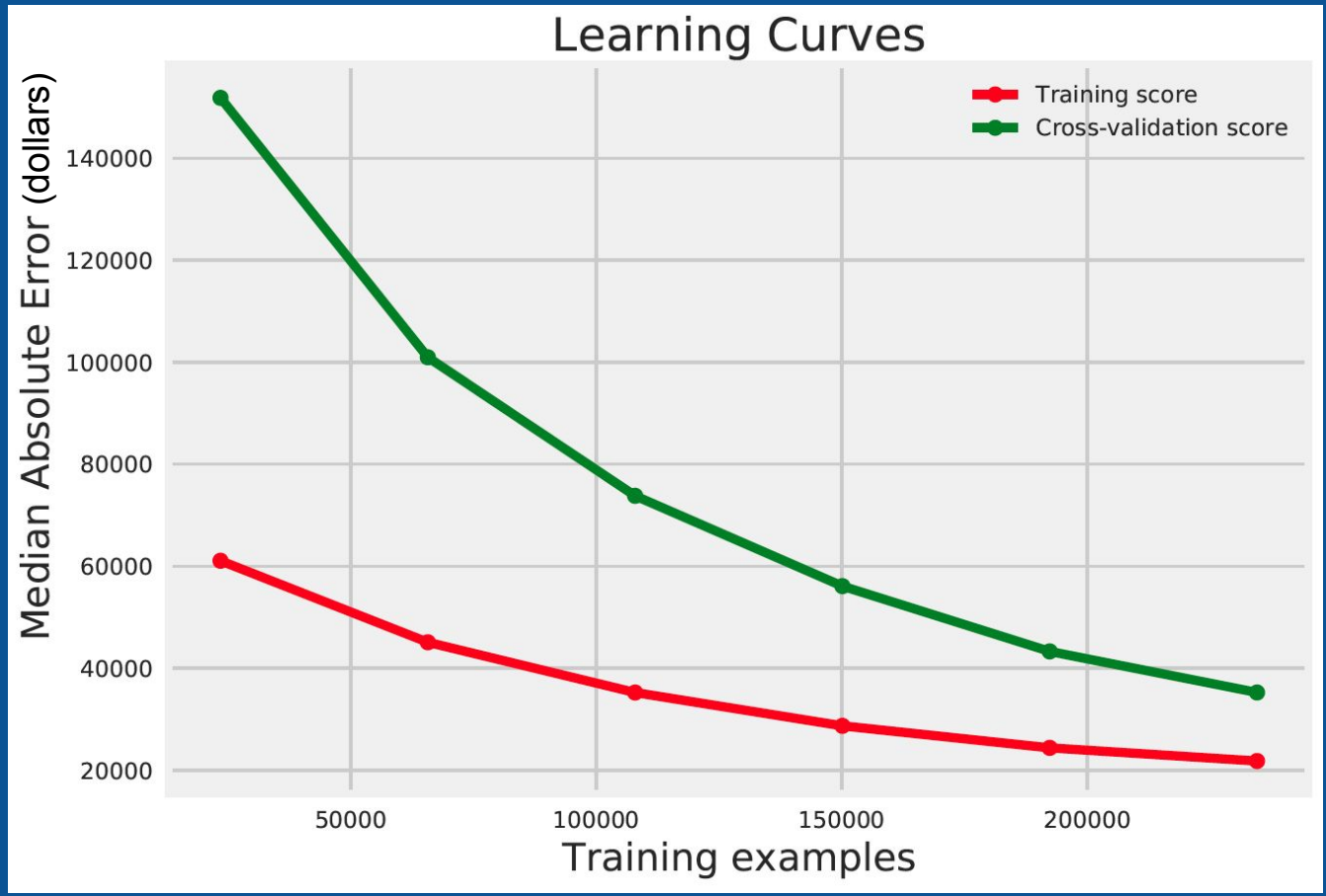
Assessor Property Values - Learning curves

Random Forest

Low variance/bias

More data could improve accuracy of predictions

Caveat - like Airbnb rentals, these property are not actual sale prices



Comparison to Zillow Values

Airbnb:

~5200 units, ~$21 (MAE)

	Median Error Percentage
Zillow Model (Rents)	8.9%
This Model (Airbnb)	15.0%
Difference	+ 6.1%

Properties (Assessor):

~300k units, ~$37,000 (MAE)

	Median Error Percentage
Zillow Model (Sales)	5.7%
This Model (Assessor)	6.3%
Difference	+ 0.6%

Not quite there, need more data and approaches:

- Add time-series data
- Feature engineering
- Learning curve shows more data will improve accuracy!

Presentation Outline

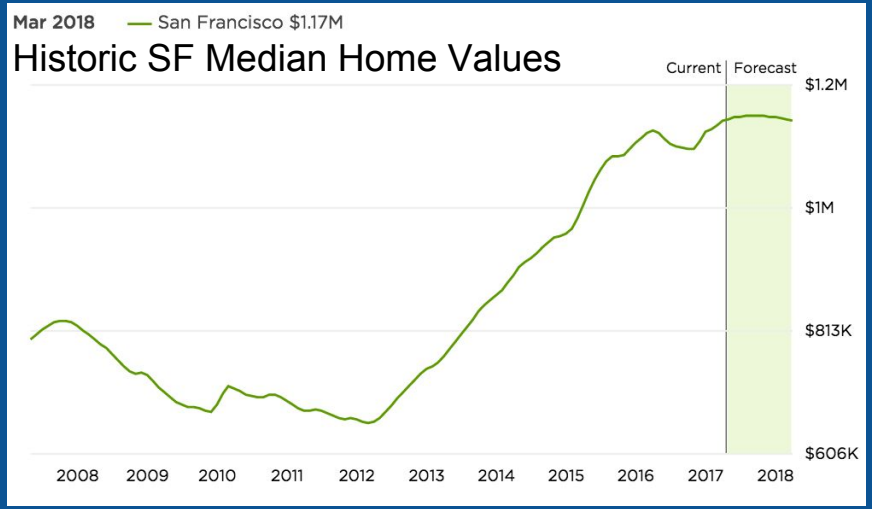
1. Introduction and Motivation
2. Target Datasets
    a. Airbnb prices (rentals)
    b. Assessor values (property sales)
3. Feature Addition to Datasets
    a. Geospatial data
    b. Vectorization method
4. Machine Learning Models
5. Model Evaluation
6. **Conclusions**

Conclusions:

- Easier to model property values than rental prices (especially Airbnb)
- Large free geospatial datasets can have significant impact on pricing models
- Simple to apply data to model, easy to iterate

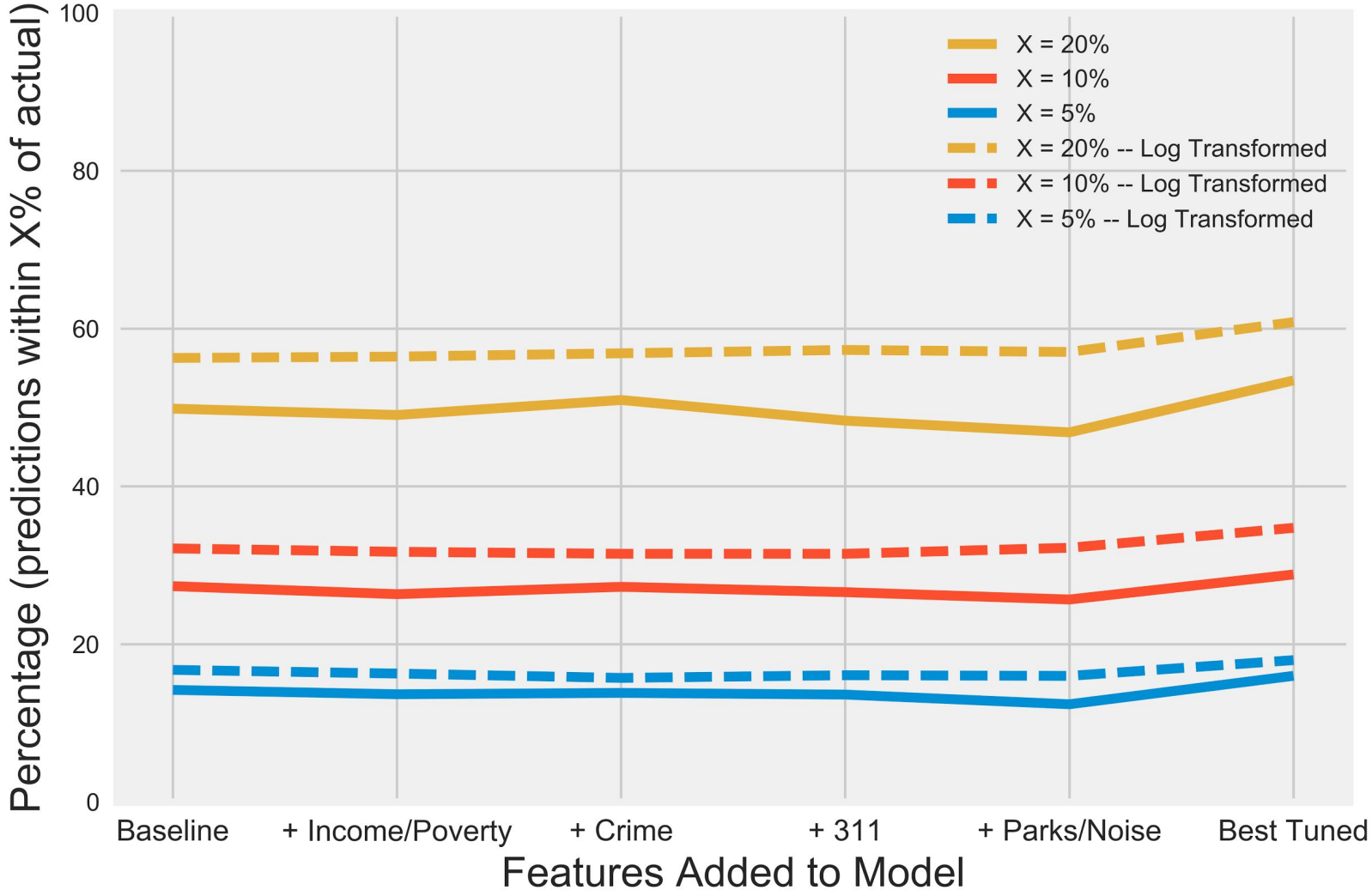
Next steps:

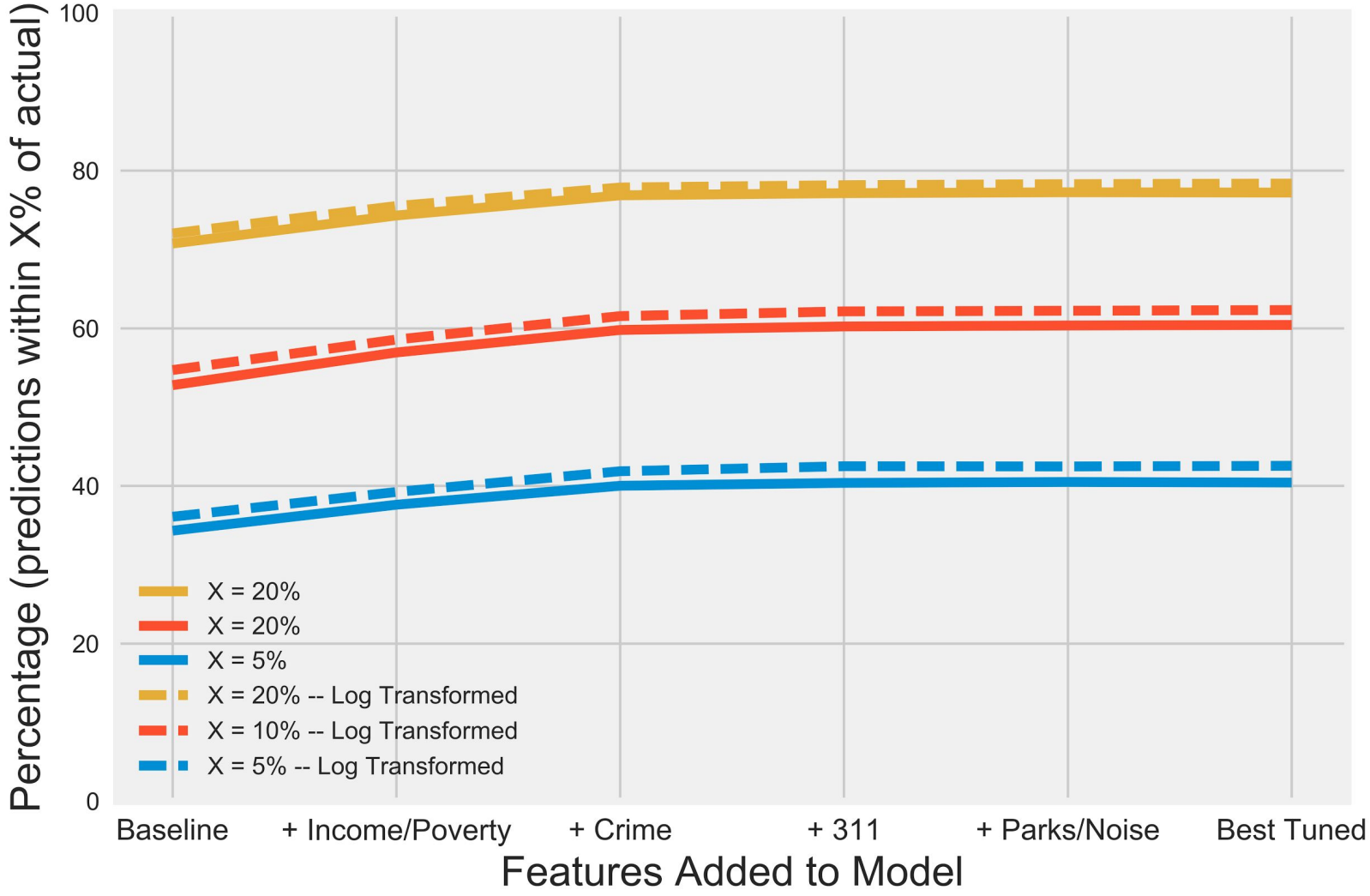
- Adjust parameters of geospatial data
- Incorporate time into model
- Acquire additional data
- Incorporate street images into model
- Obtain actual transaction data



Thanks to: GA - DSI-5 classmates and instructors,
SF Dept. of Public Health, Inside Airbnb, SF Open Data

Email: taylor.kilian@gmail.com


Percentage (predictions within X% of actual)
100
80
60
40
20
0
X = 20%
X = 10%
X = 5%
X = 20% -- Log Transformed
X = 10% -- Log Transformed
X = 5% -- Log Transformed
Baseline
+ Income/Poverty
+ Crime
+ 311
+ Parks/Noise
Best Tuned
Features Added to Model


Percentage (predictions within X% of actual)
100
80
60
40
20
0
X = 20%
X = 20%
X = 5%
X = 20% -- Log Transformed
X = 10% -- Log Transformed
X = 5% -- Log Transformed
Baseline
+ Income/Poverty
+ Crime
+ 311
+ Parks/Noise
Best Tuned
Features Added to Model