Impact of Future Housing Developments in SF

Jasmine Teo, Sally Fang, Anton Bosneaga, Priya Reddy, Arsh Hothi

Table of contents

Background & Motivation

ADUs

AHDBP

Housing in San Francisco

... is expensive.

Several factors have contributed to this problem:

- a limited supply of housing
- a thriving tech industry that has led to increased demand
- restrictive zoning laws that make it difficult to build new housing units.

Home SF

SOMA

https://www.sfchronicle.com/projects/2022/fixing-san-francisco-problems/Housing
https://sfplanning.org/central-soma-plan
https://sfplanning.org/accessory-dwelling-units
https://sfplanning.org/ahbp
https://sfplanning.org/home-sf

Research Question

Can we predict the value of new developments in San Francisco from historical property valuations?

San Francisco has implemented several policies and programs aimed at increasing the production of new housing units, particularly affordable units. These efforts have helped increase the supply of new housing in the city, but the demand for housing continues to outpace supply, leading to high prices and limited affordability.

Our research question for this project is whether we can predict the value of new developments in San Francisco based on historical property valuations.

The Data

Assessor Historical Secured Property Tax Rolls

DataSF - City of San Francisco's Open Data Platform

Includes all legally disclosable information (location, value of property, property characteristics) from July 1, 2007 to June 30, 2018 for San Francisco

Police Department Incident Reports

DataSF - City of San Francisco's Open Data Platform

This dataset includes incident reports that have been filed as of present starting 2013. These reports are filed by officers or self-reported by members of the public using SFPD's online reporting system.

San Francisco City Survey Data

DataSF - City of San Francisco's Open Data Platform

The City Survey asks residents to indicate their usage and satisfaction with city services and infrastructure like libraries, Muni, public safety, and street cleanliness. The City Survey was conducted every year from 1996 to 2004, and biennially from 2005 onward.

Data Cleaning: Housing Valuation

Encoded Zip Codes

Extracted and formatted coordinate data to join with San Francisco ZIP Codes.geojson using GeoPandas

Selected Data

Kept only the relevant features, and data from 2013-2019 (by the Closed Roll Year)

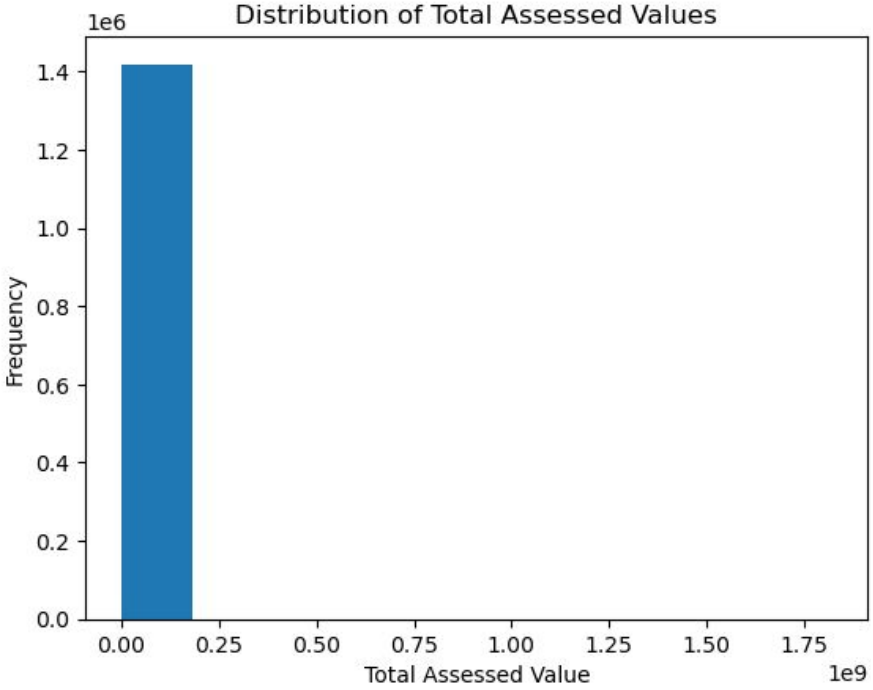
Calculated the Outcome Variable

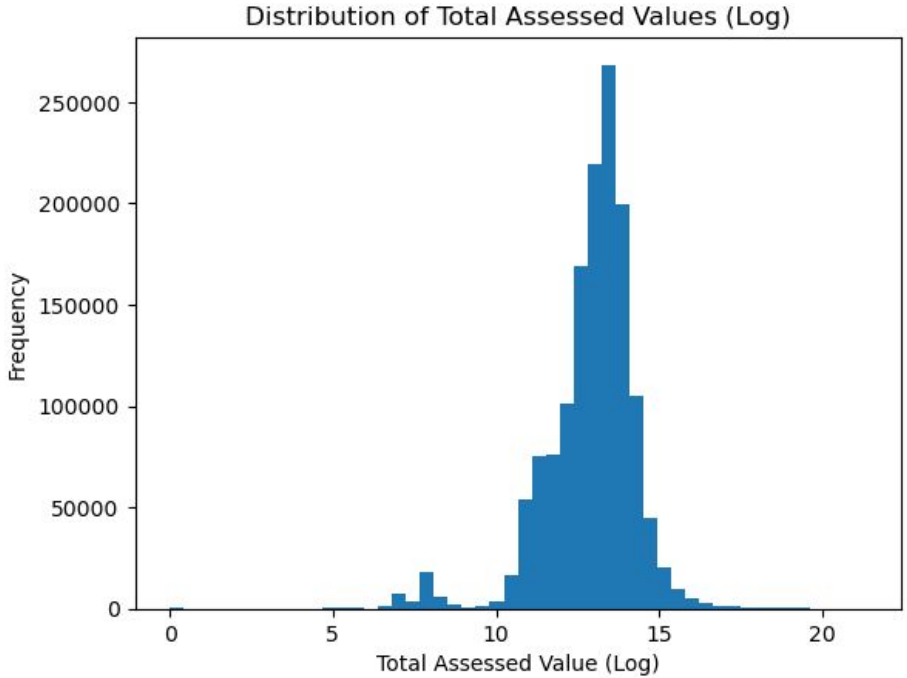
Total Assessed Value = Assessed Personal Property Value + Assessed Improvement Value

Logged Total Assessed Value

Valuation exhibits a right-skewed distribution because of a few expensive lots, we will instead look at the log

Total Assessed Values





Data Cleaning: Police Incidents

Combined 2 Datasets

Combined 2013-2018 and 2018-Present together, dropped duplicated data for 2018 using Incident ID (unique identifier)

Selected Data

Kept only the relevant features and columns we are interested for EDA and later modeling

Encoded Zip Codes

Extracted and formatted coordinate data to join with San Francisco ZIP Codes.geojson using GeoPandas

Data Cleaning: City Survey

Selected Columns of Interest

Selected questions relating to cleanliness of the neighborhood, safety (both at day and night), household density, and zip code

Aggregated Features

Question about cleanliness changed over time, aggregated ordinal columns per response

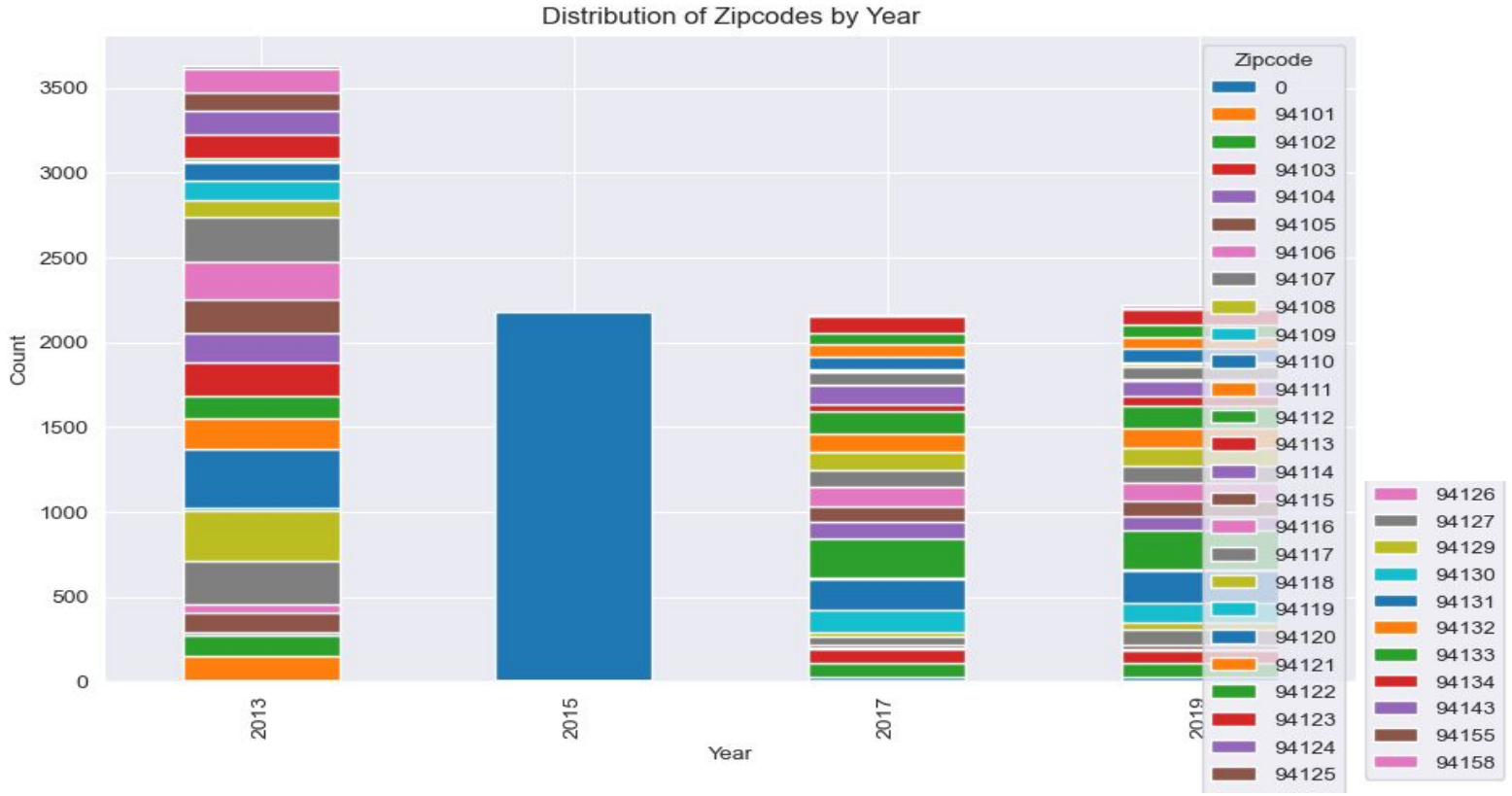
Filtered by Valid Zip Code

Filtered for valid zip codes and years 2013-2019. Found that there are no zip codes available for 2015, and we only data of every other year

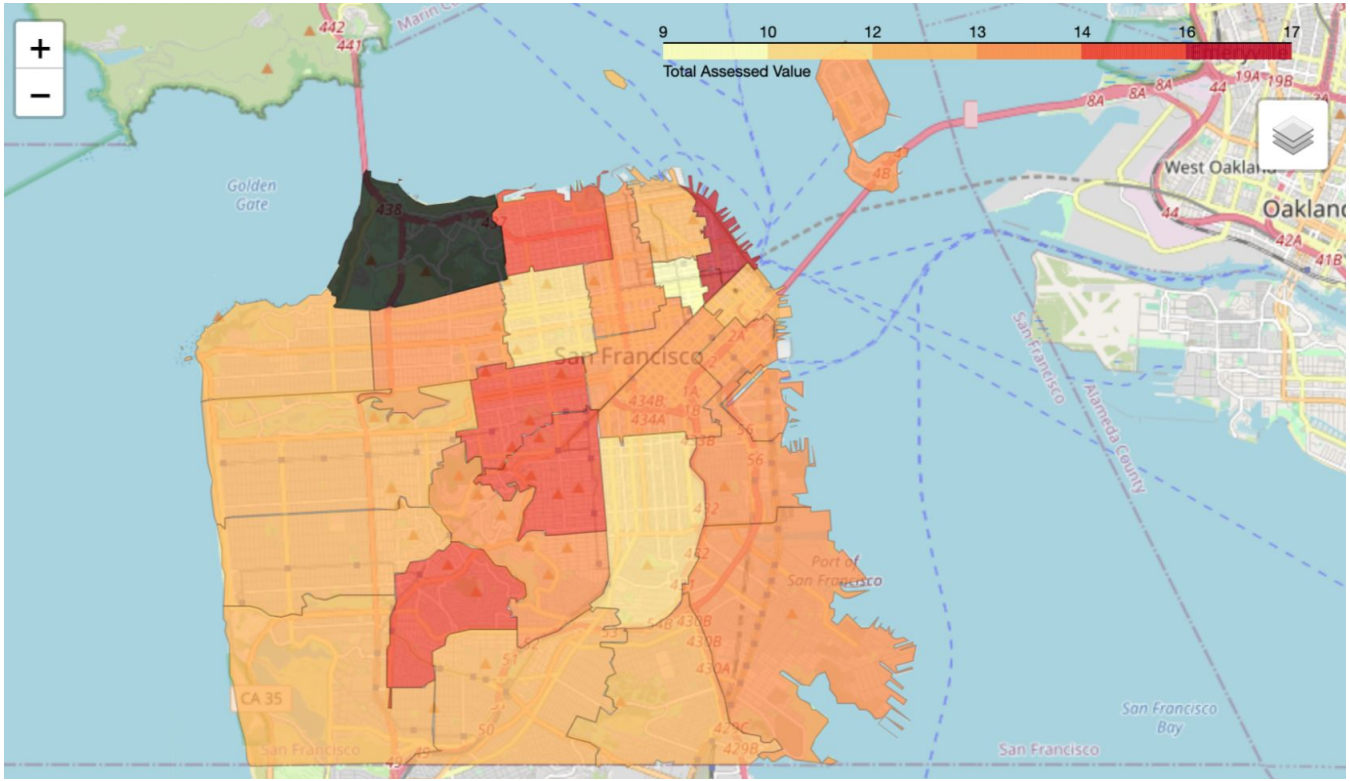
Filled in Gap Year Data

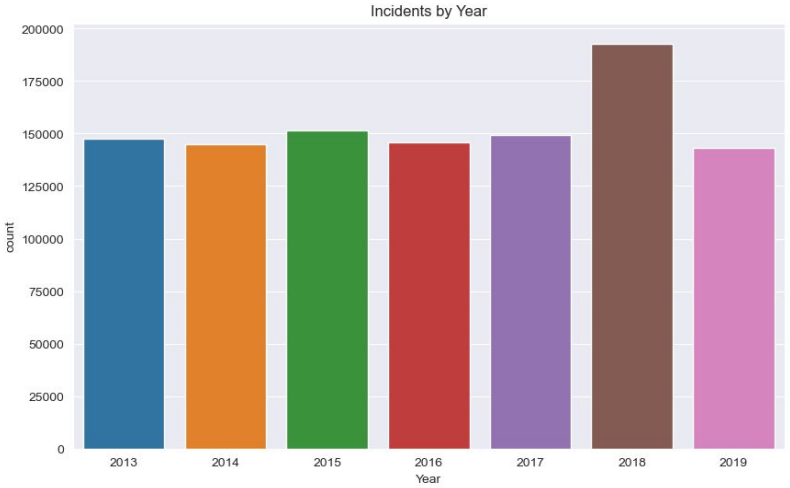
Filled in gap years with previous year data (i.e. fill in 2014 & 2015 & 2016 data with 2013 data)

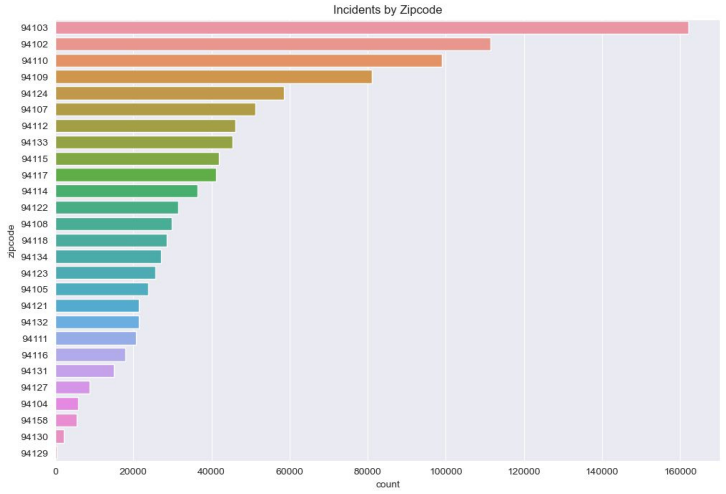
Distribution of Zipcodes

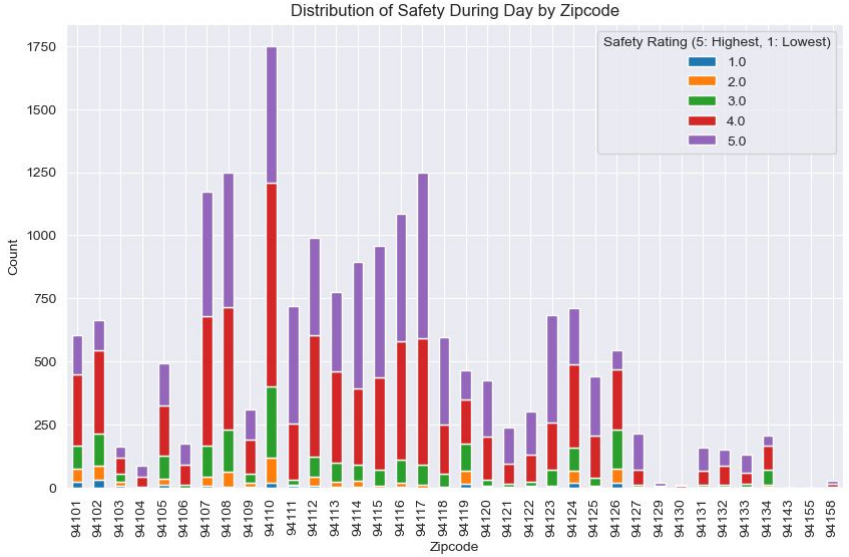


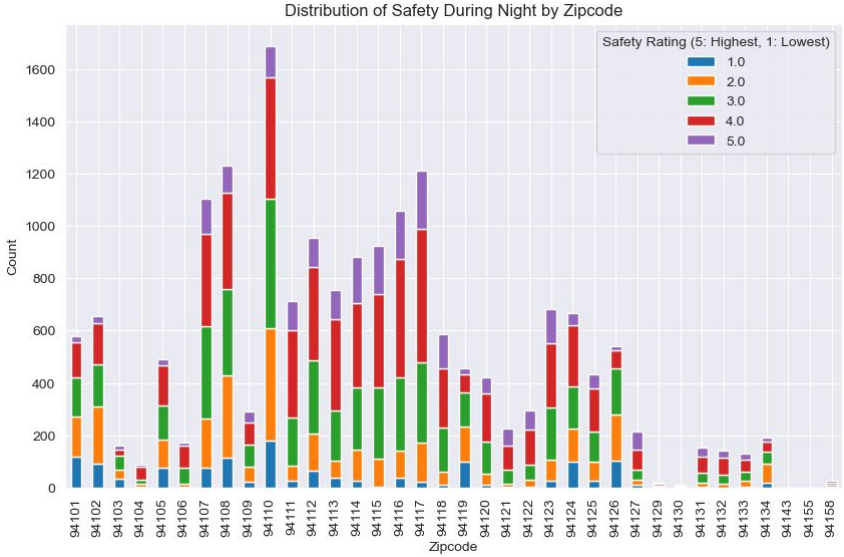
Distribution of Properties (and their values) by Zipcode




Incidents by Year
200000
175000
150000
125000
100000
75000
50000
25000
0
count
2013
2014
2015
2016
2017
2018
2019
Year


Incidents by Zipcode
94103
94102
94110
94109
94124
94107
94112
94133
94115
94117
94114
94122
94108
94118
94134
94123
94105
94121
94132
94111
94116
94131
94127
94104
94158
94130
94129
zipcode
0
20000
40000
60000
80000
100000
120000
140000
160000
count


Distribution of Safety During Day by Zipcode
Safety Rating (5: Highest, 1: Lowest)
1.0
2.0
3.0
4.0
5.0
Count
Zipcode
0
250
500
750
1000
1250
1500
1750
94101 94102 94103 94104 94105 94106 94107 94108 94109 94110 94111 94112 94113 94114 94115 94116 94117 94118 94119 94120 94121 94122 94123 94124 94125 94126 94127 94129 94130 94131 94132 94133 94134 94143 94155 94158


Distribution of Safety During Night by Zipcode
Safety Rating (5: Highest, 1: Lowest)
1.0
2.0
3.0
4.0
5.0
Count
Zipcode
0
200
400
600
800
1000
1200
1400
1600
94101 94102 94103 94104 94105 94106 94107 94108 94109 94110 94111 94112 94113 94114 94115 94116 94117 94118 94119 94120 94121 94122 94123 94124 94125 94126 94127 94129 94130 94131 94132 94133 94134 94143 94155 94158

Putting it all together:
Merging the Data

Police Data

Year	Year of Interest
Zip code	Zip code crimes took place in
Incident Count	Number of Incidents for that year and zip

House Value Data

Year

Zip Code

 # Bathrooms
 # Bedrooms
 # Stories
 # Units
+
+
+

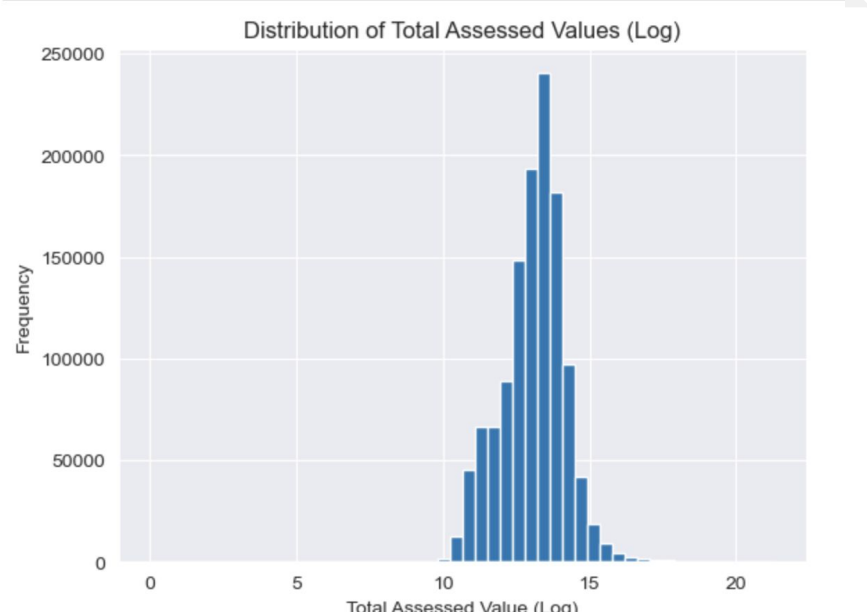
Survey Data

Year	Year of interest
Zip Code	Zipcode of interest
Safe Day	Avg score for day safety
Safe night	Avg Score for night safety
Household Size	Avg Household Size
Cleanliness	Avg Score for cleanliness of Zip Code

Multivariable Linear Regression

Why Linear Regression?

- Supervised learning model that has **high interpretability** used to predict continuous variables (such as price)
- Based on research we did on common methodologies for predicting housing prices it seems that linear regression models **are typically standard**
- Based on initial exploratory plots we can also see **linear relationship** between features and price



Implementation Steps

- Feature Engineering
- Split data into train, validation, and test sets
- Hyperparameter tuning
- Evaluate against baseline

Multivariable Linear Regression

Model Tuning

- Feature selection:
  - All Features
  - Only features with a correlation >0.5 or <-0.5
-
- Switch to Random Forest

Performance Metrics

Model	Train RMSE	Validation RMSE	Test RMSE
1. LM 1 (all vars)	339,933,381,452.719	308,289,496,854.8	NA
2. LM 2 (only high corr vars)	246,863,620,528.9301	254,280,097,924.99	252,689,331,870.50
3. Median Predictor (baseline)	**7,850,186.20**	**9,624,777.43**	**7,221,440.69**

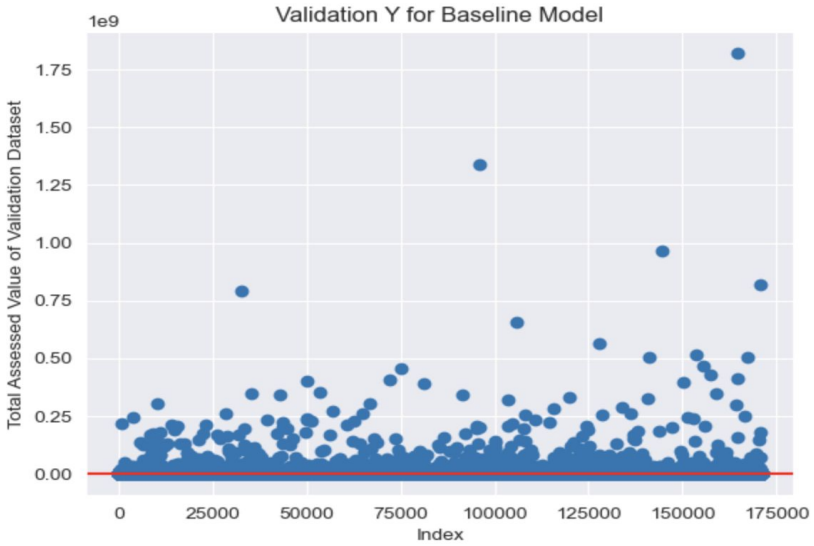
Baseline: Median Predictor

Poor Performance of Linear Model

- Use median of *Log of Total Assessed Value* as baseline predictor
- No hyperparameter tuning needed

Performance Metrics

RMSE on Train Data	7,850,186.20
RMSE on Validation Data	9,624,777.43
RMSE on Test Data	7,221,440.69



Random Forest Regressor

Why Random Forest?

- ***M*ultiple decision tree models** built and each prediction is combined to make final predictions
- **Handles non-linear relationships** between features & target variable, *Log of Total Assessed Value*
- **Deals well with missing values and outliers**, providing robust predictions with relatively low risk of overfitting

Feature Selection

- Including irrelevant/redundant features can introduce **noise & unnecessary complexity**
- **Selected Features:** *Number of Bathrooms, Number of Rooms, Zipcode_94123, Number of Units, Zipcode_94105, Zipcode_94124, household_size, Zipcode_94112 Zipcode_94134','Zipcode_94122, Zipcode_94116, cleanliness, Year, Year Property Built, Zipcode_94104, Zipcode_94115, Number of Bedrooms. Zipcode_94109, Zipcode_94111, Number of Stories, Zipcode_94114, Zipcode_94103, Zipcode_94108, safe_day_binary*

Random Forest Regressor

Hyperparameter Tuning

- n_estimators: number of decision trees to include in the random forest
  - Ran random forest on values ranging from 10 to 40
- max_depth: controls the maximum depth of each decision tree in the random forest
  - No large divergence between train and validation RMSE
  - Set max_depth to default value

Performance Metrics

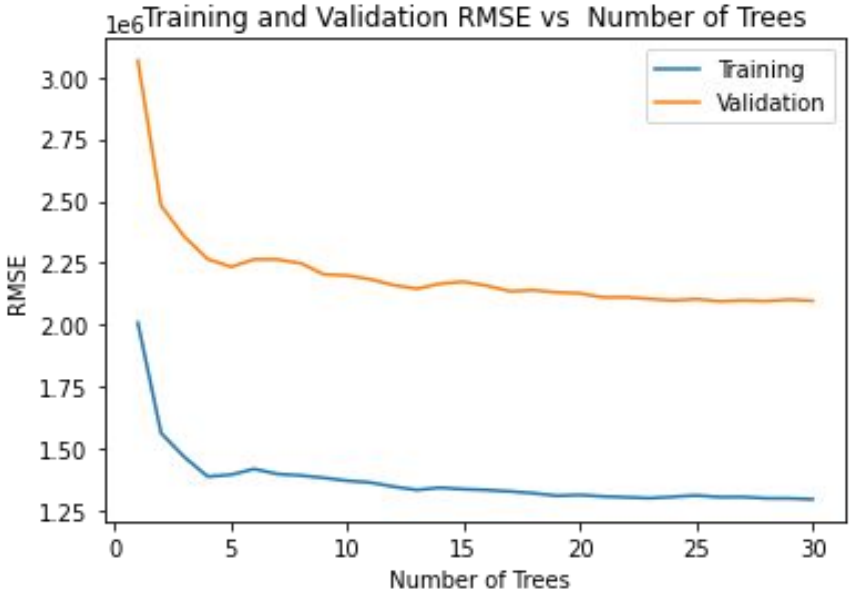
Model	Train RMSE	Validation RMSE	Test RMSE
1. RF (n_estimators = 10)	1,370,849.90	2,199,539.04	NA
2. RF (n_estimators = 20)	**1,313,411.46**	**2,127,457.25**	**2,107,751.89**
3. Median Predictor	**7,850,186.20**	**9,624,777.43**	**7,221,440.69**

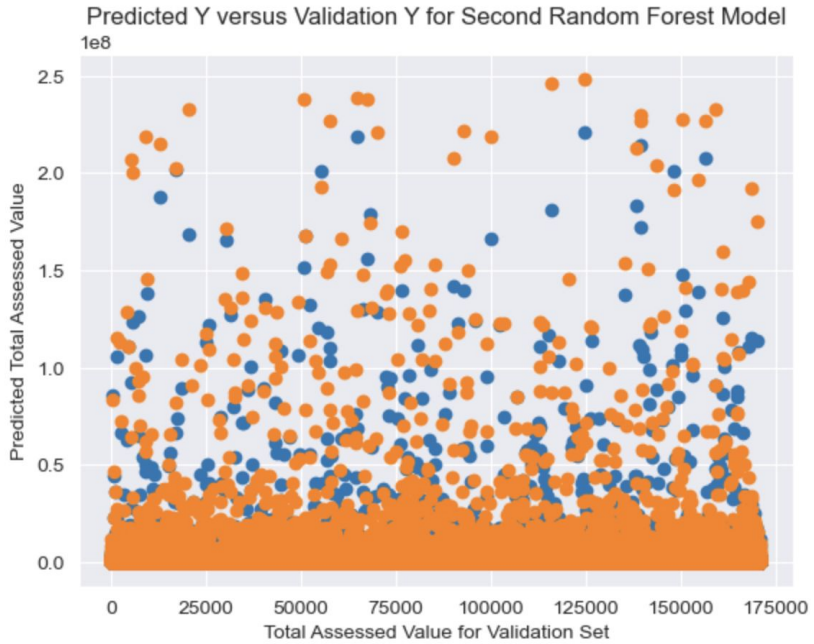
Final Model

Baseline

Random Forest Regressor

n_estimators = 20





Results

Model	Train RMSE	Validation RMSE	Test RMSE for Tuned Model
First Linear Model	339933381452.719	308289496854.8	N/A
Second Linear Model	246863620528.9301	254280097924.99942	168664180981.5098
Baseline Model (Median)	4703213.630206855	4713523.811974167	4406240.195313015
First Random Forest	1370849.9003143772	2199539.0445950455	N/A
Second Random Forest (Final)	1313411.4599670158	2127457.2478843774	2107751.893084619

Conclusions

Random Forest Models are Unreasonably Effective

- We explored multiple flavors of both linear models and more complex NN models
- Our final Random Forest model does a reasonably good job of modeling our data
- It is vastly superior to the median baseline model and more basic linear models
- Our Random Forest Model with 20 n_estimators and default max_depth yields the best results among tested models

Loss of Explainability

- Due to the complexity of the relationships in our dataset we have had to sacrifice explainability
- Final model potentially useful as a predictor
- But not useful as an explainable model

Future Work

Better Data

- More granular data could yield better results
- Data not reliant on city appraisals
- More detailed data (currently neighborhood data), perhaps from custom surveys
- Incorporate additional data to better separate data points (zoning, unit level data, ownership status)

Better Models

- With more granular data we could attempt to find a fitting linear model
- Alternatively, search for a more flexible FNN or CNN model to take advantage of more rich data

Known Biases

Aggregation

- Zip code level aggregation was necessary to provide a common scale for all of the available datasets
- We are losing some granularity and implementing implicit averaging by focusing on the zip code level

Missing Data

- In order to populate our zip code level records with missing data we have extrapolated neighborhood level data down to the zip code level
- This further exacerbates the averaging the impact of the affected features
    - E.g. safety, neighborhood satisfaction, etc.

Thanks!

Any Questions ?

Team Contributions

Notebooks:

Met up to collaborate on notebooks 4-5 times for work sessions.

EDA : Arsh, Anton

Data Cleaning: Priya, Sally, Jasmine

Data Merging: Priya, Sally, Jasmine

Model Making + Results: Priya, Sally, Jasmine

Cleaning up notebook : All

Slides:

Motivation: Arsh

Data (police, city survey, valuation):
- Introduce Raw data - Arsh
- Explain EDA -Arsh
- Cleaning - Sally
- Explain merging of data - Priya

Baseline Model (Linear Model 1,2 - Priya & Median - Jasmine)

Final Model (Random Forest 1,2 - Jasmine)

Conclusions: Anton

Contribution: ALL