Allegheny County Home Value Index Case Study

For a Fidelity Management and Research Interview

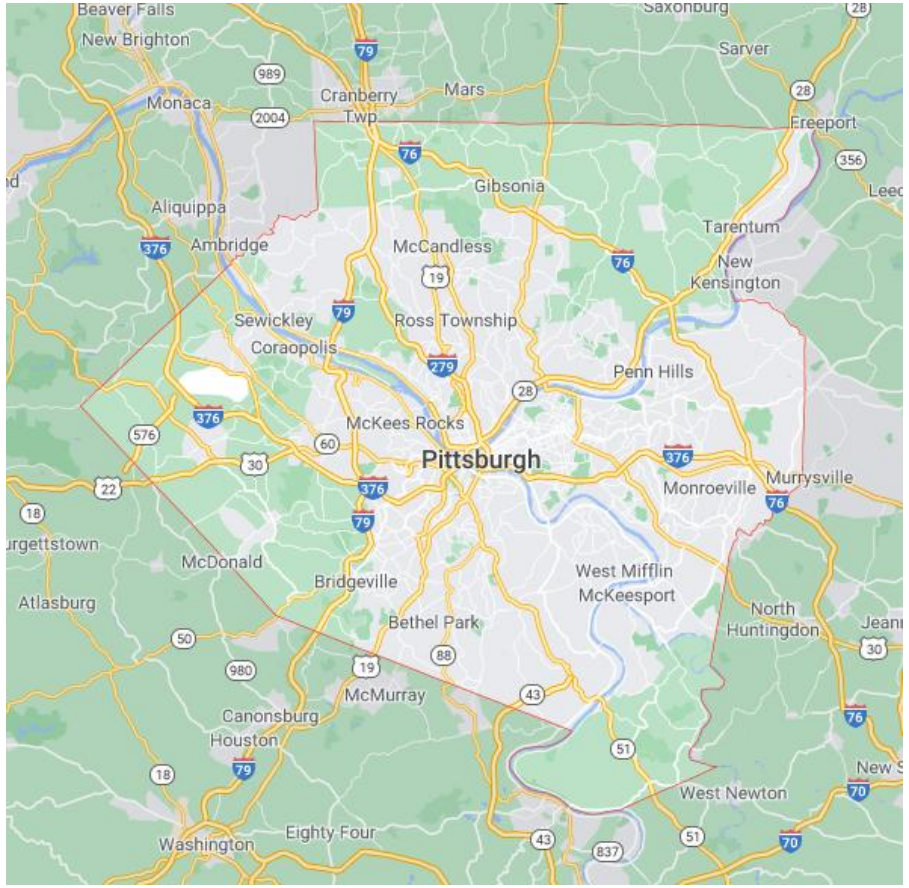
Matheus C. Fernandes (PhD Candidate - Harvard University)

2/11/2021

Please find more information and full code at https://git.fer.me/fidelity-interview

About Allegheny County

- Located in southwest Pennsylvania
- State's second-most populated county
- City of Pittsburgh is in center
- With 446 bridges, Pittsburgh has more bridges than any other city in the world*



*https://uncoveringpa.com/facts-about-pittsburgh

About the Allegheny County Dataset

- 86 columns (potential features)
- 580,997 property assessments
- Columns contain information on:
  - Property features: bedrooms, bathrooms, fireplace etc.
  - Location: physical address and location codes
  - Different levels of assessment values – county, local for building and land

About the Allegheny County Dataset

- Has numerical, categorical, and binary variables
- Dataset contains repetitive information i.e. descriptions of codes
- Contains administrative information i.e. deed recording information and legal descriptions

Goal of Analysis

Develop a monthly “Allegheny County Home Value Index” (HVI) to understand key features of the market.

Create model to gain insights for investment opportunities.

Structure of Data Science Workflow

Data Cleaning

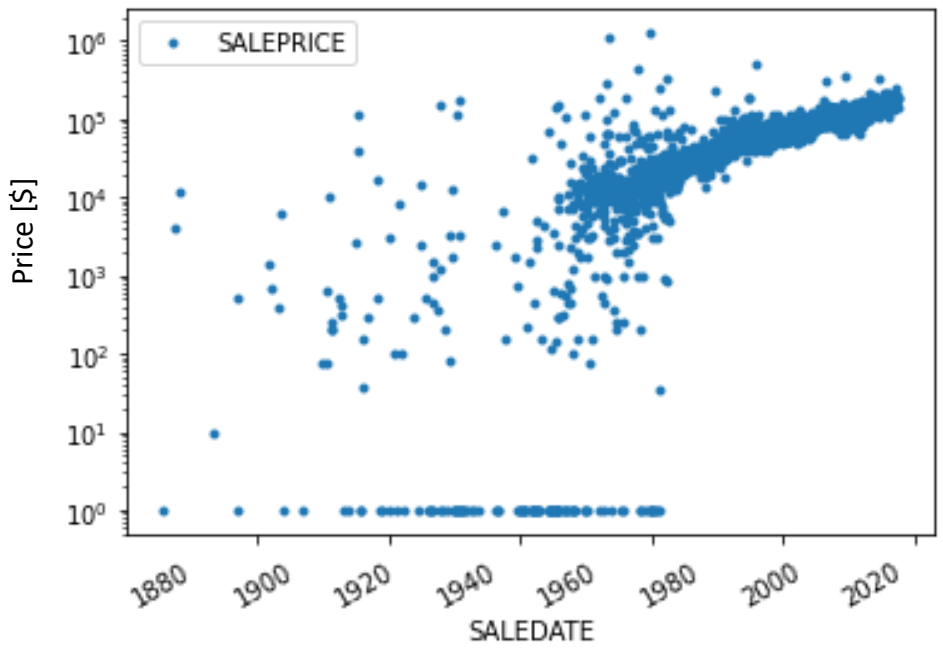
Exploratory Data Analysis

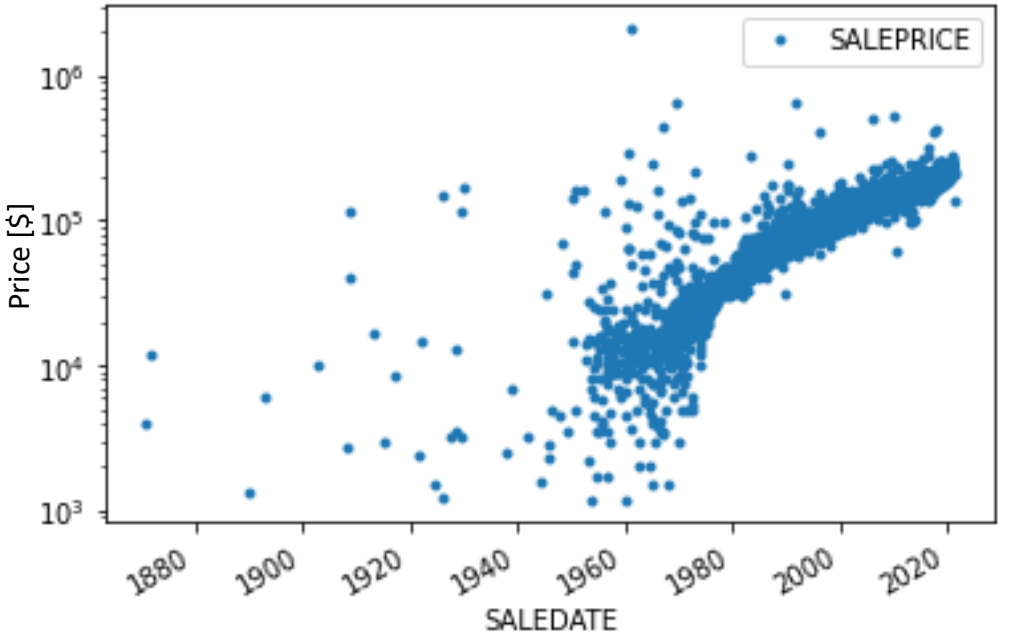
Model Exploration and Selection

Computing Home Value Index

Missing Data and Data Types

Removing false data





Missing Data and Data Types

- Remove data missing important information such as:
  - Sale price
  - Sale date
  - Sale price is 0 or unreasonably low (<$1000)

Missing Data and Data Types

- Imputed missing data depending on datatype
  - Creating a new category of unknown, zero, or boolean
  - Replacing mean of data for continuous variables
  - Imputing information from different column

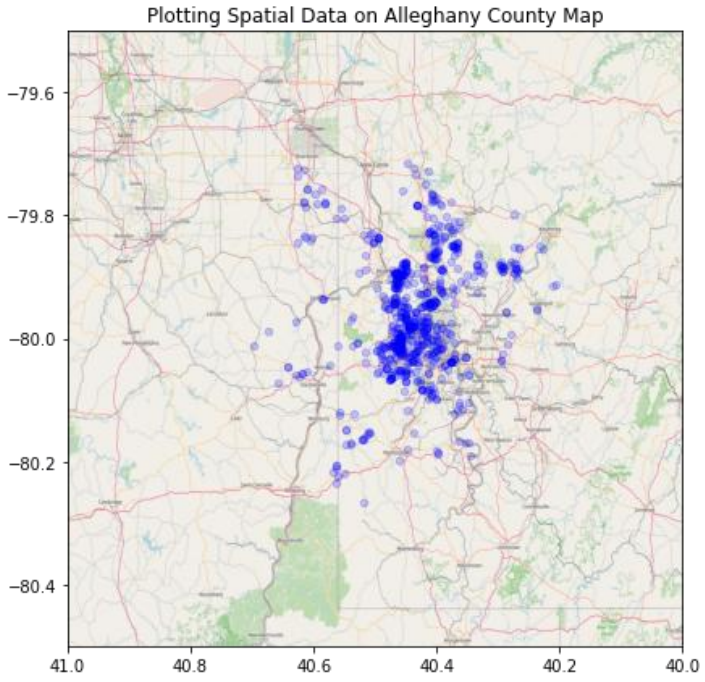
Missing Data and Data Types

- Converted variable types to increase bit efficiency
  - From 64 bit to 32bit and 8 bit
  - Reduced memory ~380MB to ~120MB without loosing information
- Converted date types to numerical
- Standardize input data

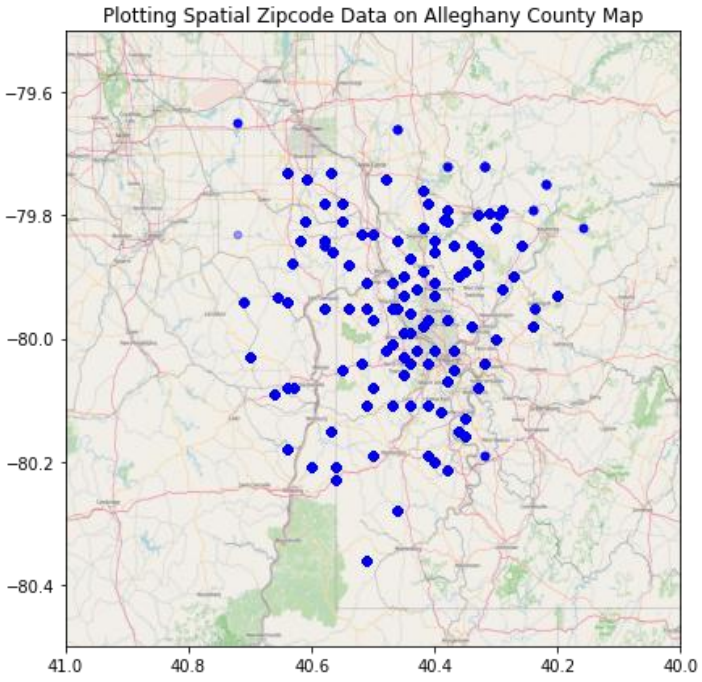
Feature Engineering

Geolocation Exploration

Address Granularity



Zip code Granularity



Structure of Data Science Workflow

Data Cleaning

Exploratory Data Analysis

Model Exploration and Selection

Computing Home Value Index

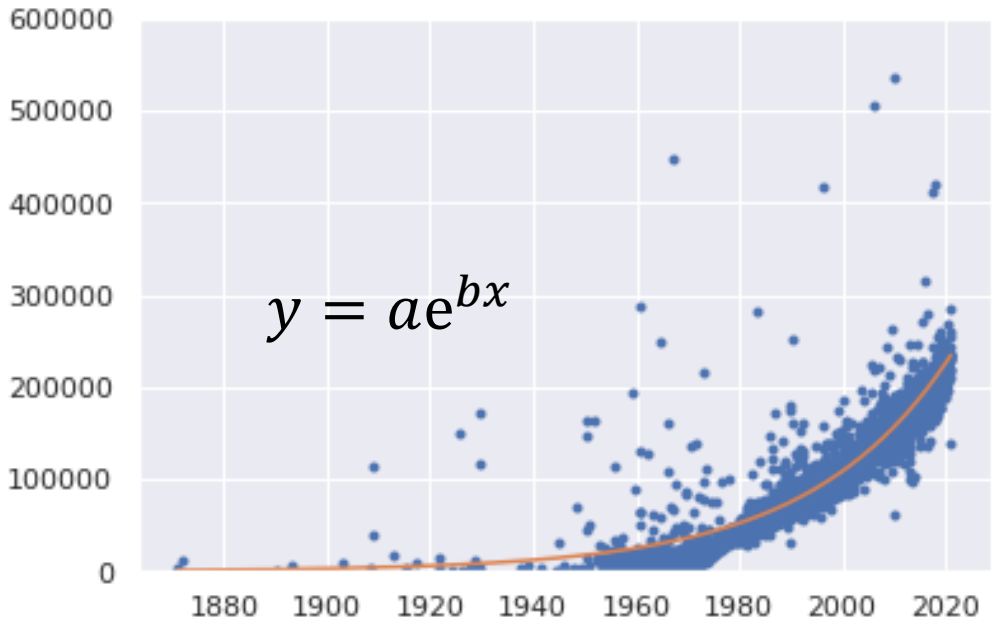
Martingale?

Is housing prices in Allegheny county a martingale?

$$\mathbf{E}(|X_n|) < \infty$$

$$\mathbf{E}(X_{n+1} \mid X_1, \ldots, X_n) = X_n$$

By fitting an exponential line, we see that the expectation follows an exponential growth.



Exploratory Data Analysis

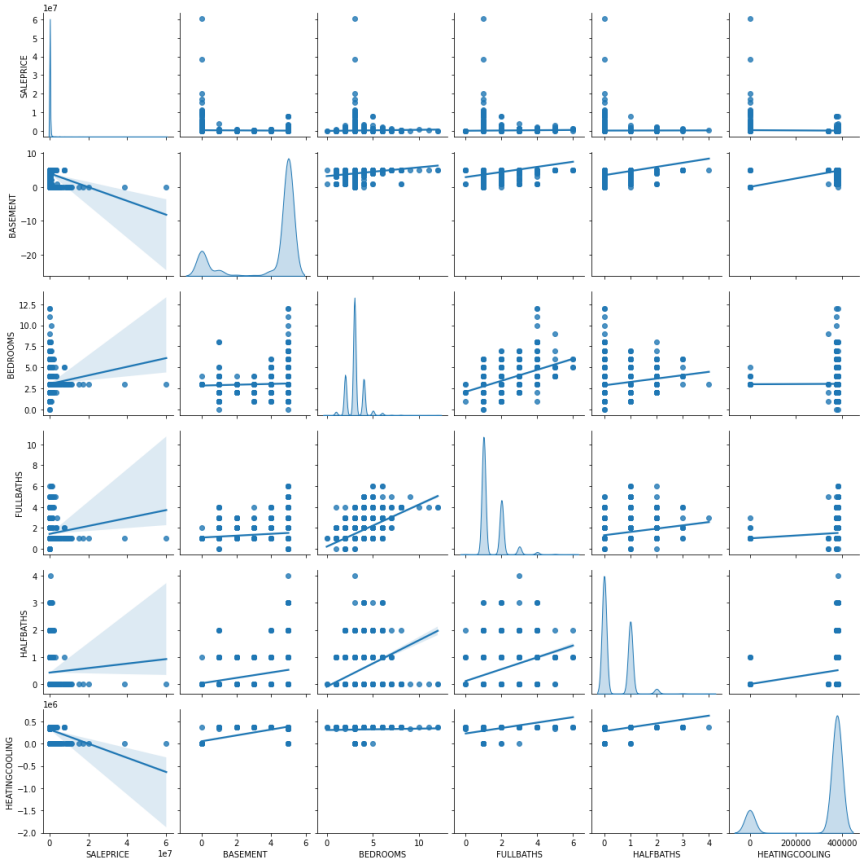
Correlation Matrix

- Provides information on correlation of variables
- No correlation does not mean no useful information
- High correlation between variables means potential redundancy between those variables.

Exploratory Data Analysis

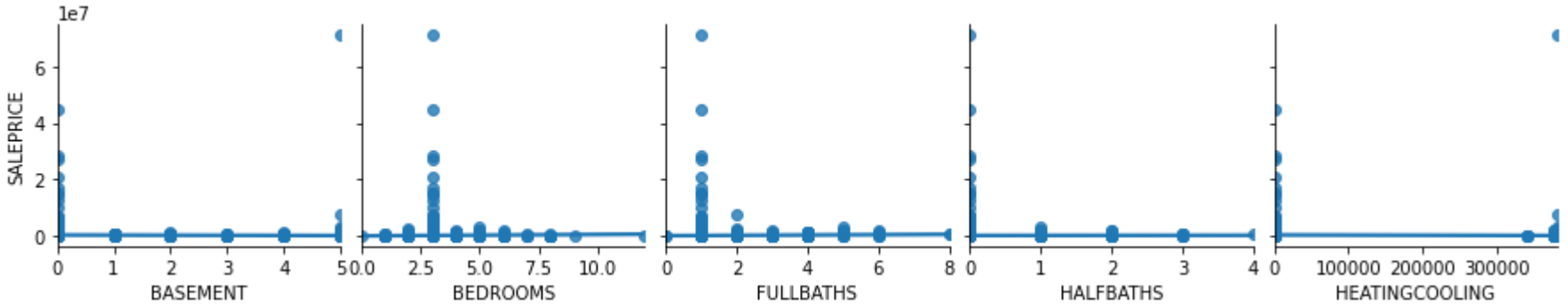
A deeper dive into the data

- How does each variable depends on the other
- Look for trends in the data
- How does property sale price depends on each feature



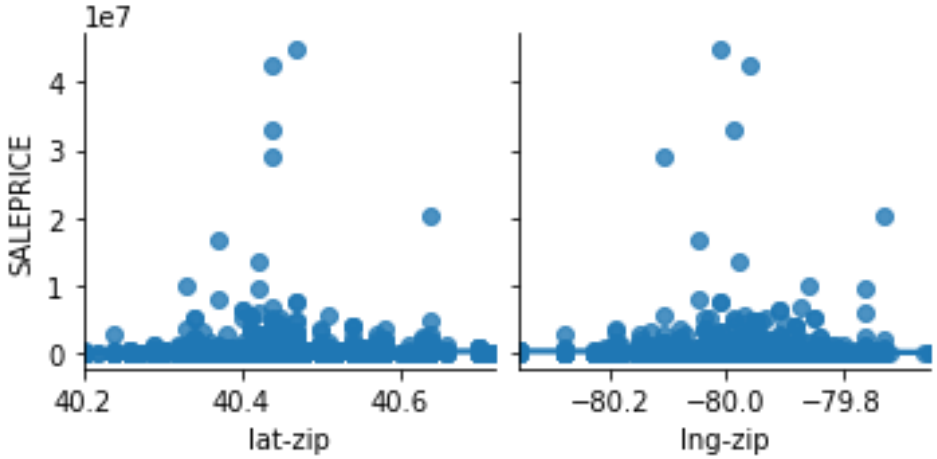
Exploratory Data Analysis

How do house properties impact pricing?



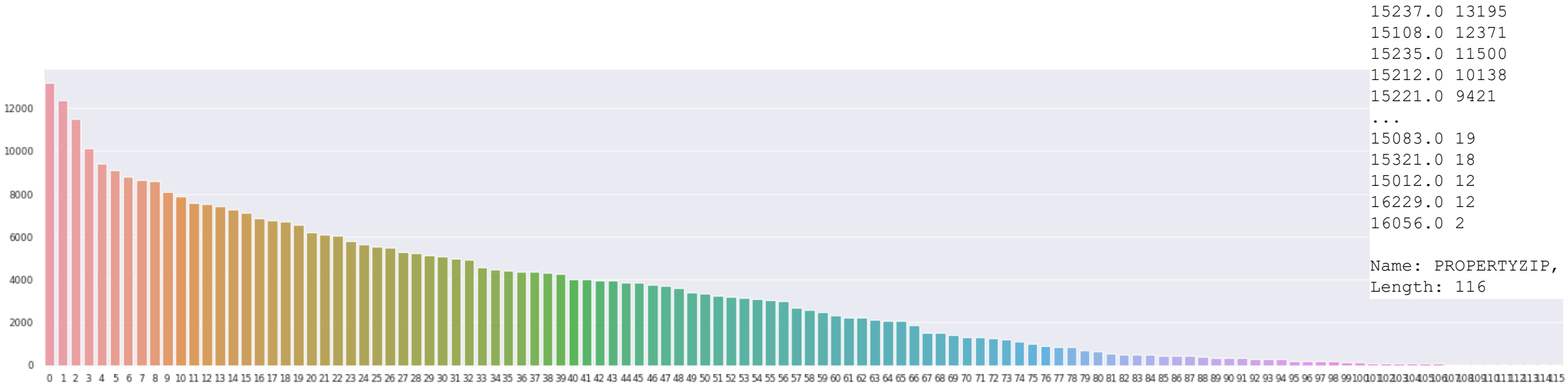
Exploratory Data Analysis

How does location of properties impact pricing?



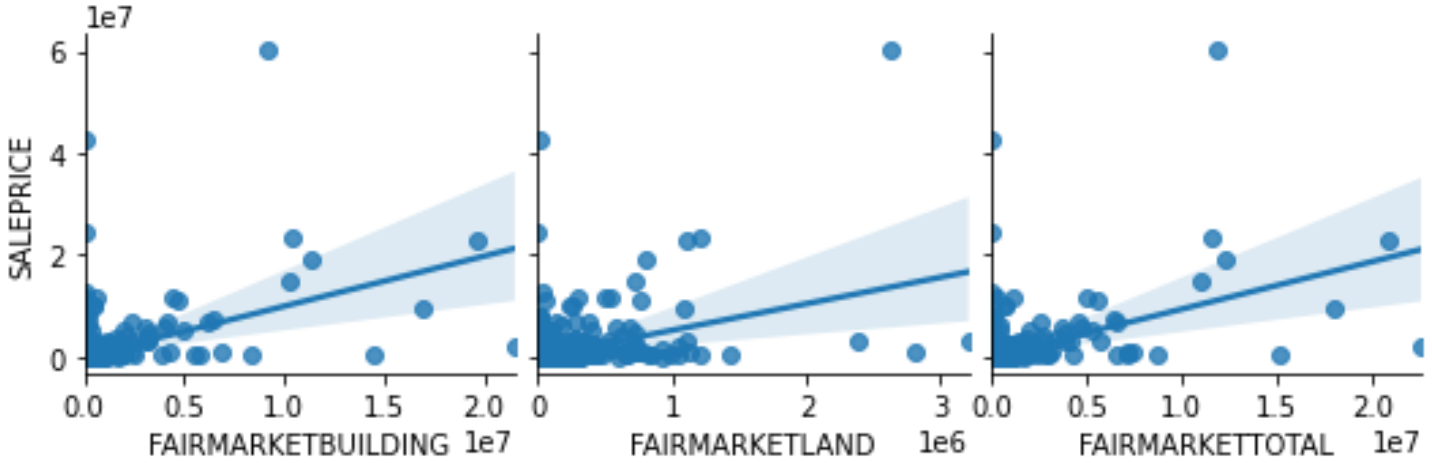
Exploratory Data Analysis

What is the distribution of the data across different locations?



Exploratory Data Analysis

How do the assessments impact pricing?



Structure of Data Science Workflow

Data Cleaning

Exploratory Data Analysis

Model Exploration and Selection

Computing Home Value Index

Model Information

Model Goal: Predict valuation of existing homes for a variable sale date.

Target variable: Sale Price

Input variables: Sale date and important features that provide information on the valuation of a property at a certain date

Model Information

Model Assumptions:

- No information on the buyers side (demand)
- No listing prices or spread of ask/bid
- No information on interest rates
- No demographic information
- No information on the economy
- No refined information on location
- Based on assessments only from 2021 (dataset)
- Discrete daily sampling

Model Choices

Seek these regression model characteristics:

- Deal with sparse data
- Good for dealing with categorical and numerical data
- Efficient at training (limited computational resources on my end)
- Scalable to potentially adding more data in the future

Model Choices

Model	Train Score	Test Score
Linear Model: LassoCV	0.573	0.495
Support Vector Machine (SD)	0.201	0.007
Ensemble: Random Forest	0.900	0.523
Ensemble: Bagging	0.930	0.588
Ensemble: Adaptive Boosting (SD)	0.829	0.380
Ensemble: Extreme Gradient Boosting	0.814	0.767

*Scores are measured using R2 Score: 1-(sum of square residuals/total sum of squares)
SD = sampled dataset

What is XGBoost?

- Decision-tree-based ensemble model
- Uses gradient boosting framework
  - Converting weak to strong learner through sequential learning
  - Gradient descent algorithm
- Great for small-to-medium structured/tabular data
- Boosting optimized for software and hardware parallelization

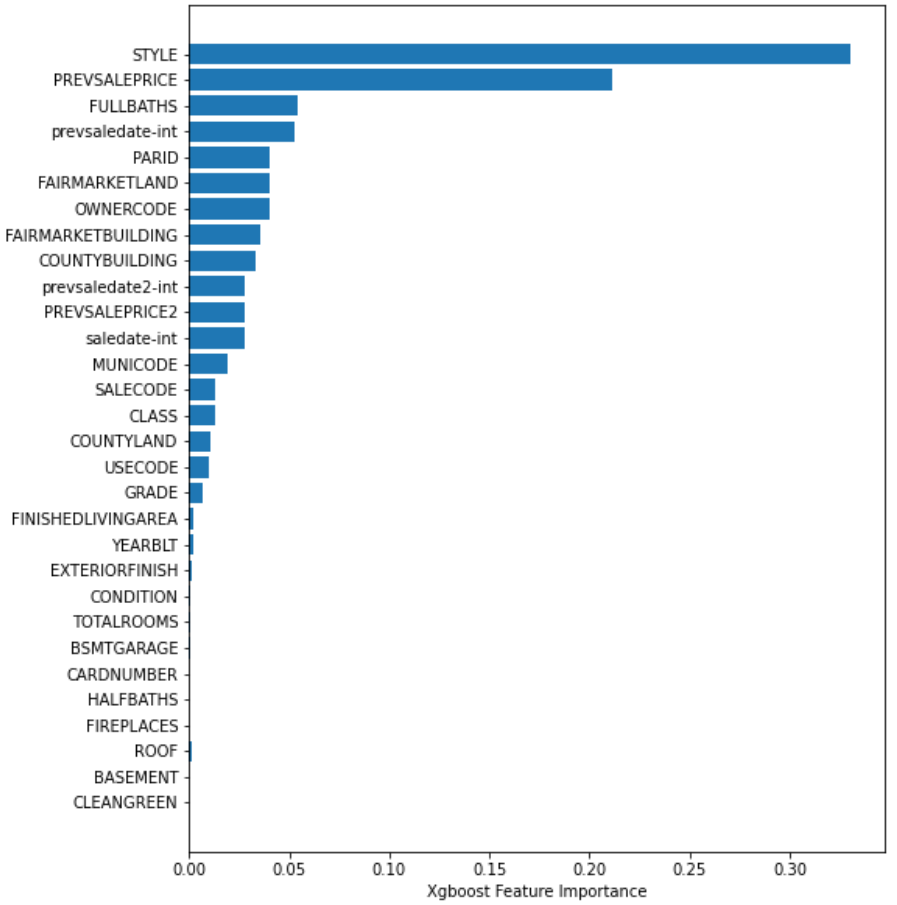
Another Level of Feature Selection

Feature Importance

- Based on Gini importance

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

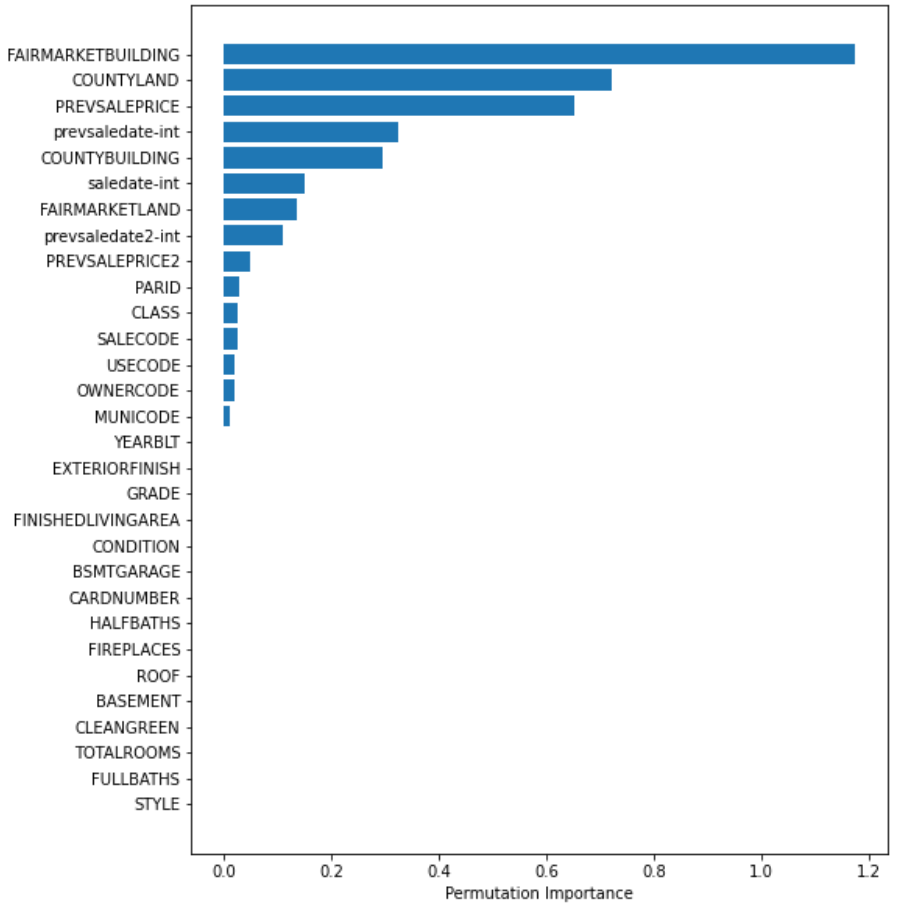
- The higher the importance the more crucial it is for prediction



Another Level of Feature Selection

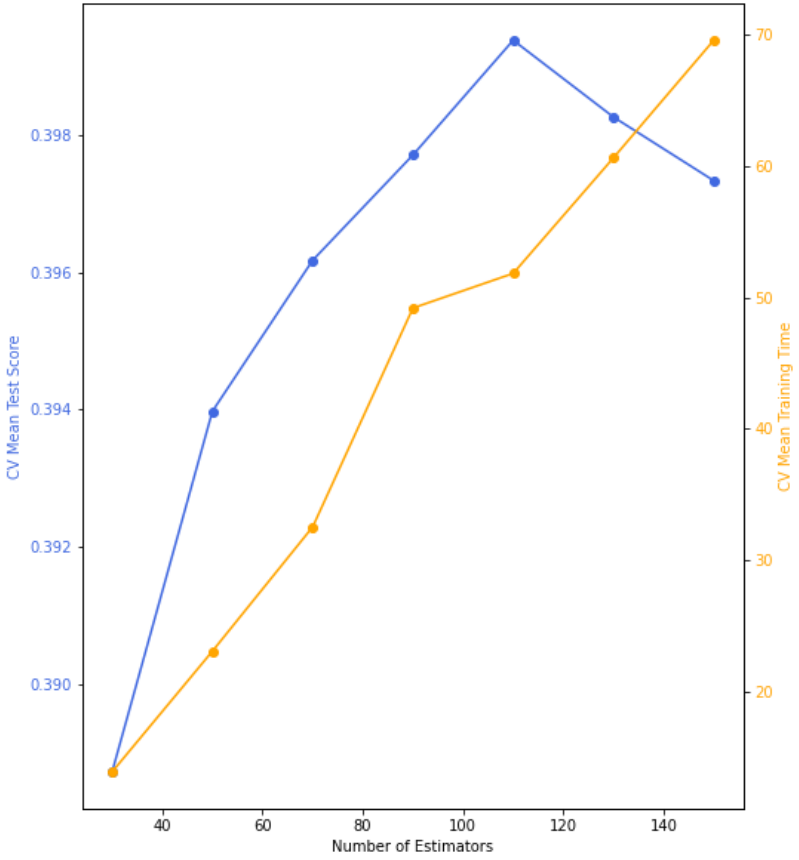
Permutation Importance

- Compute score of model
- For each feature shuffle column and compute score for corrupted dataset
- The higher the importance the more crucial a particular feature



XGBoost Hyperparameter Tuning

- For computational efficiency, only tuned 1 param – number of estimators
- Use cross-validation with 4-way split
- A bias-variance balance is obtained at n=110, with underfitting before and overfitting after
- As n increases computational time increases



Structure of Data Science Workflow

Data Cleaning

Exploratory Data Analysis

Model Exploration and Selection

Computing Home Value Index

Home Value Index

What does it mean? (based on Zillow's definition)

- Insight on typical expected home values
- Insight on housing market at a given time
- Appreciation over time

Home Value Index

How is it defined:

$$I_{t-1} = \frac{I_t}{1 + A_t}$$

Home Value Index

How is it defined:

$$a_{i,t} = \frac{z_{i,t} - z_{i,t-1}}{z_{i,t-1}}$$

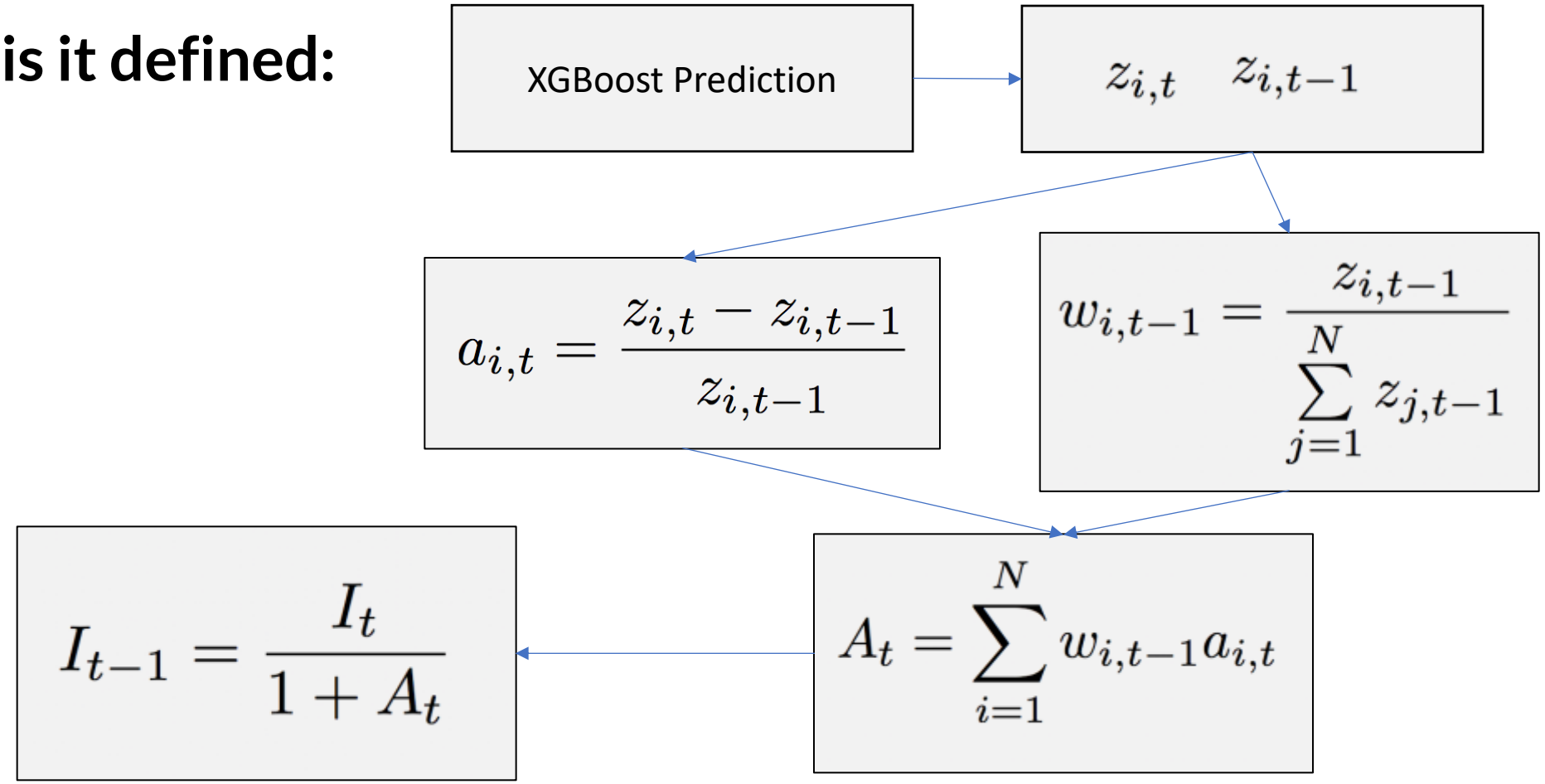
$$w_{i,t-1} = \frac{z_{i,t-1}}{\sum_{j=1}^{N} z_{j,t-1}}$$

$$A_t = \sum_{i=1}^{N} w_{i,t-1} a_{i,t}$$

$$I_{t-1} = \frac{I_t}{1 + A_t}$$

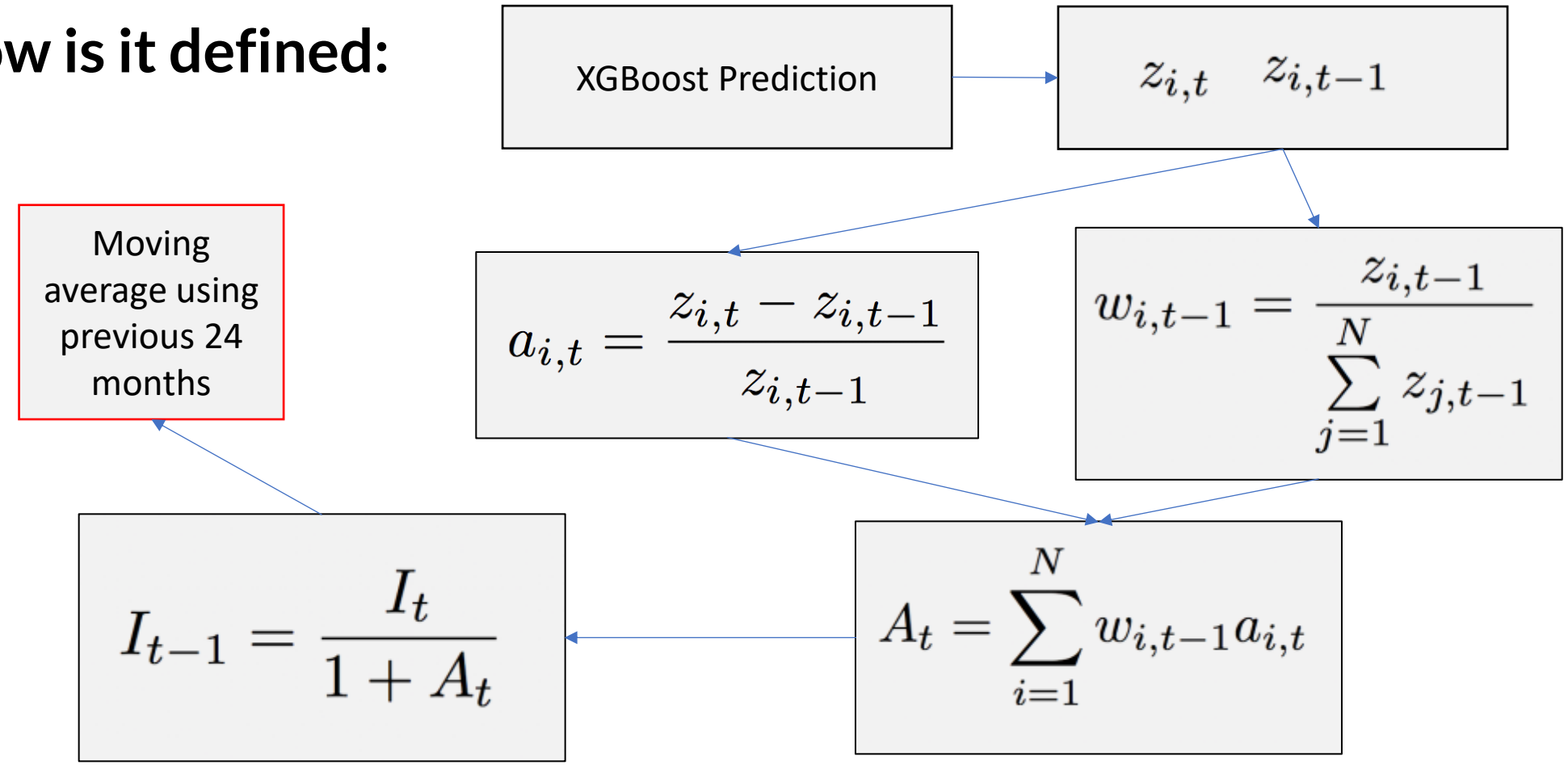
Home Value Index

How is it defined:



Home Value Index

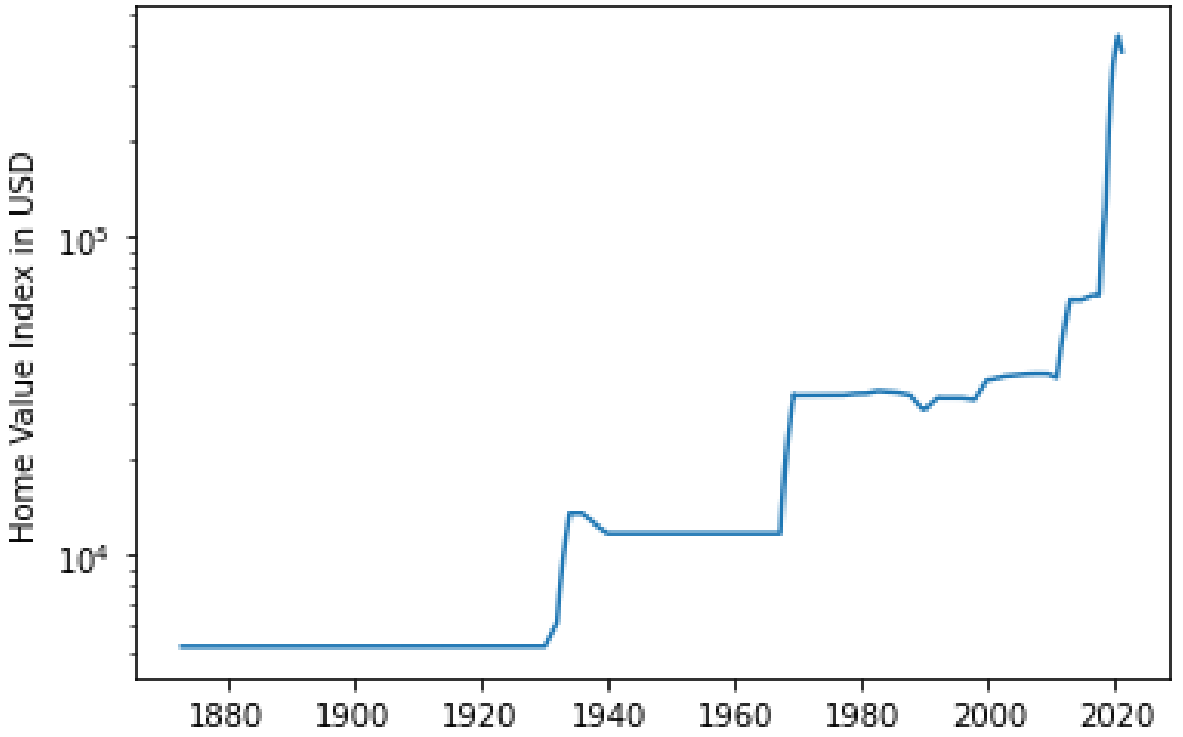
How is it defined:



Results

- Monthly time series of index

- Extract key economic features
  - Late 1930s housing boom
  - 1990 housing crisis
  - Dot-com bubble
  - 2007 housing bubble
  - 2007 housing recovery
  - 2020 pandemic



Conclusions and Recommendations

- Model needs further development and validation
- Must relax certain model assumptions
- Allegheny county real estate has historically been a good long-term investment
- Based on recent trends I would advise not investing in the housing stock of Allegheny county

Model Extension

- Create monthly HVI for each separate Zip code and create an index based on that
- Implement a better geolocation scheme to refine location
- Account for economic data such as interest rates into the model
- Include additional sale data and historical assessment data from multiple listing service (MLS)

Thank you for the opportunity!

For a Fidelity Management and Research Interview

Matheus C. Fernandes (PhD Candidate - Harvard University)

Please find more information and full code at https://git.fer.me/fidelity-interview