Data Science Case Study

Allegheny County Property Assessment Dataset

Fidelity Sr. Data Scientist Candidate Interview

Yinsen Miao, Ph.D.
Feb 15, 2021

About me

Yinsen Miao, Ph.D.

Professional Experience

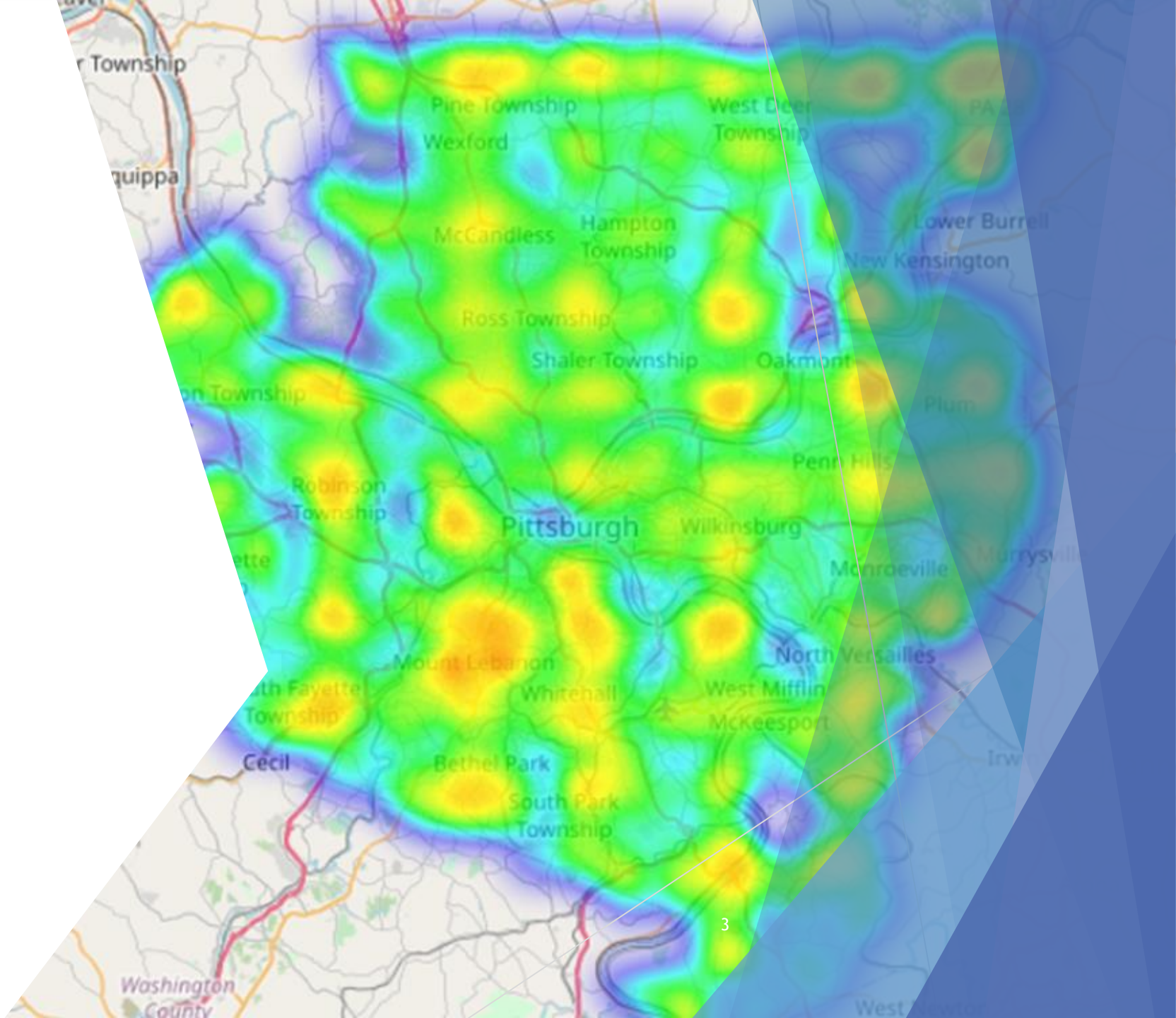
Shell	AI, Quantitative Researcher	Jan 2020 – current, Houston TX
Shell	Data Scientist	June 2017 – May 2019, Houston TX
Quantlab Financial	Quantitative Research Intern	June 2019 – Aug 2019, Houston TX

Education

Ph.D. in Statistics	Rice University	Dec 2019, Houston TX
M.A. in Statistics	Rice University	Dec 2014, Houston TX
B.S. in Applied Math	Minzu University of China	June 2013, Beijing, China

Content

- Project Introduction
- Workflow:
  - Data Preparation
  - Feature Engineering
  - ML model
  - Home Value Index
  - Portfolio Optimization
- Summary and Future Direction



Project Introduction

Scenario: U.S. government launched a website data.gov which provide high value, machine readable dataset. With the Allegheny County Property Assessments Dataset from Data.gov, we would like to generate valuable insights for investing by applying modern data science techniques.

Questions to answer:

1. Is housing price in Allegheny county a martingale?
2. Design and test a simple monthly “Allegheny County Home Value Index” using the data set.
3. Design and test an investment strategy with $5 million budget: make your investment to homes appear on the market starting from Jan 1, 2016 and check the resulting value of your investments by Nov 30, 2020.

Why housing price is not a martingale?

- What is martingale?

Consider a stochastic process, $X_1, X_2, \cdots, X_n$. It is a martingale if :

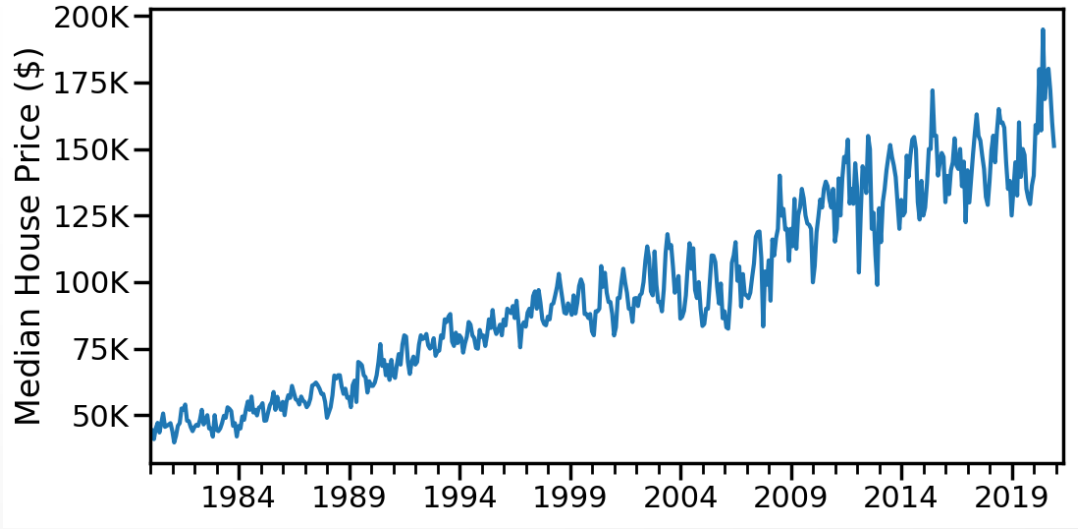
1. $\mathbb{E}[X_n] < \infty$ for all n i.e. X_ns are integrable.
2. $X_n \in \mathcal{F}_n$ for all n, i.e. X_ns are adapted, we can "observe" X_n at time n.
3. $\mathbb{E}[X_{n+1} \mid X_1, X_2, \cdots, X_n] = X_n$ for all n.

- How about housing price?

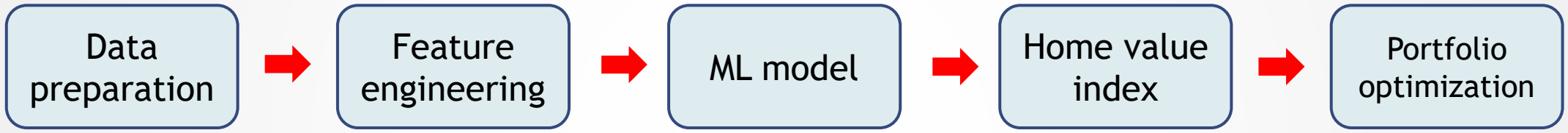
  Does the expectation of future price is the same as the present price?

  The time series of median price shows

  Finite value, trend, seasonality



End-to-end Machine Learning (ML) Pipeline

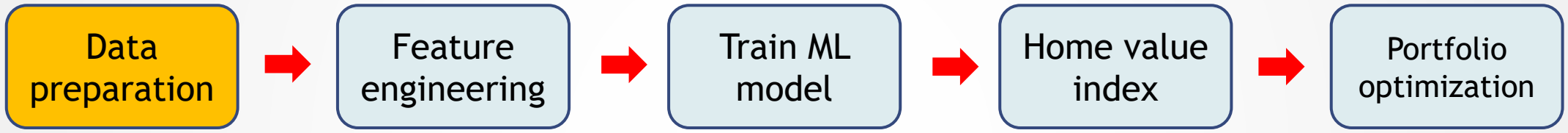


- ✓ Data collection
- ✓ Valid transaction filtering
- ✓ Market segmentation

- ✓ Training/testing split
- ✓ LightGBM
- ✓ Shapley ML Interpretation
- ✓ Optuna hyperparam optimize

- ✓ Henderson MA Filter
- ✓ Seasonal adjustment
- ✓ Mean variance optimization

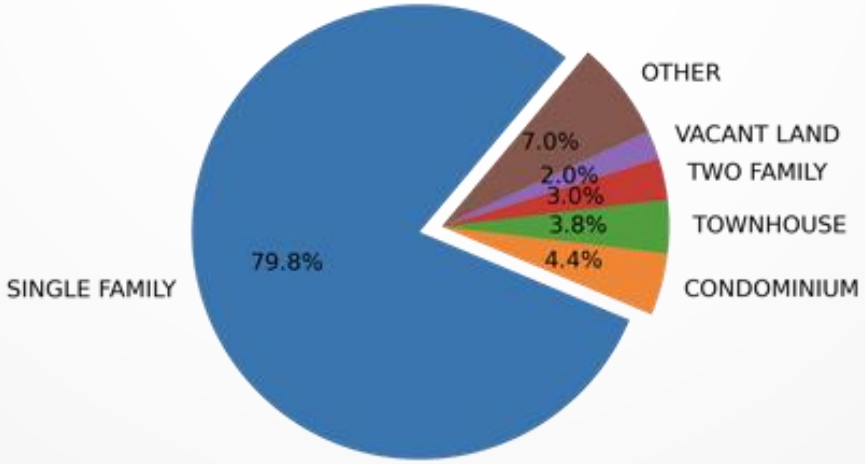
ML Pipeline – Data Preparation



- ✓ Data collection
- ✓ Valid transaction filtering
- ✓ Market segmentation

Data Preparation – Assessment Data Overview

- ~ 581K properties, 1.15 million sales records and 86 variables.
- Up to 3 historical sale transactions for each property.
- Only ~30% property's most recent transactions are considered as valid sales.
- Among the properties with valid sale records, the top 5 market segments are: Single Family (79.80%), Condom (4.42%), Townhouse (3.79%), Two Family (2.97%) and Vacant Land (2.02%).
- Due to the limitation of sales data, we will focus on building pricing model for Single family.



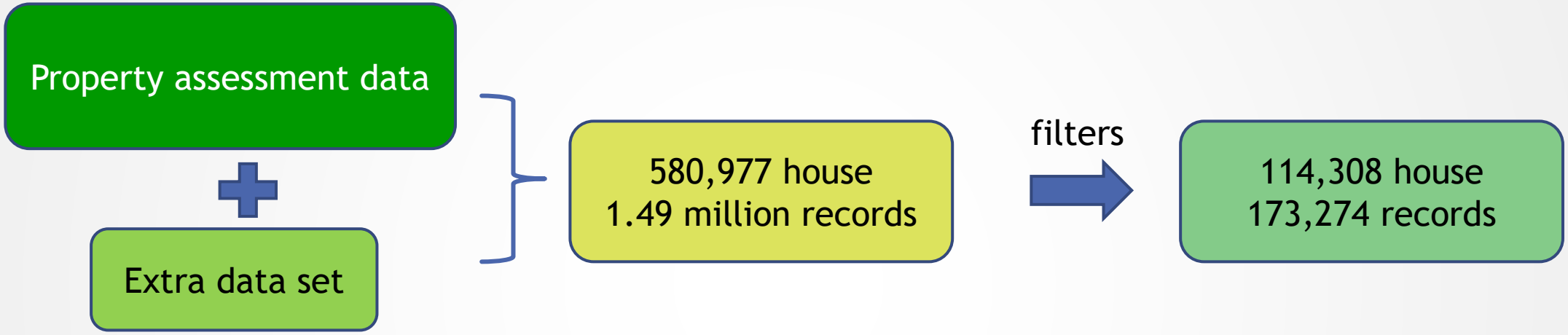
Data Preparation – Assessment Data Limitation

Data limitation:

- Missing key information.
- Only up to three transaction per property. (properties are infrequently traded)
- Invalid sale transaction cannot reflect fair market.

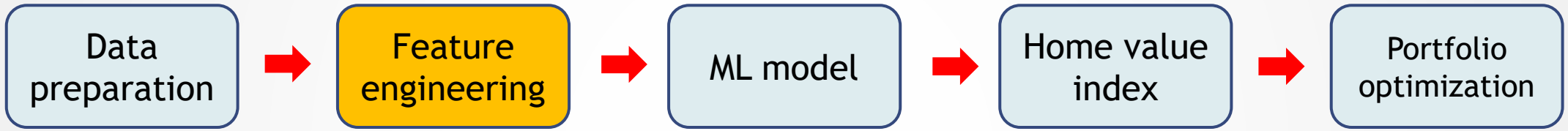
Extra data set	Provided information
Property Sale Transactions (post 2013)	All transactions records after 2013
School District Ranking (2020)	School district ranking (grade from A+ to D)
Poor Properties Survey (2016)	Estimated distressed housing percentage in community
Older Properties Survey (2016)	Estimated older housing percentage in community
Vacant Properties Survey (2016)	Estimated vacant housing percentage in community
Anxiety Medication Survey (2016)	Anxiety medications usage data in community

Data Preparation – Data Processing



1. Merge all useful information from extra data set to property assessment data.
2. Filters applied to data set.
   - **Valid sale transaction** only.
   - **Single family** only.
   - **Residential properties** only.

ML Pipeline – Feature Engineering



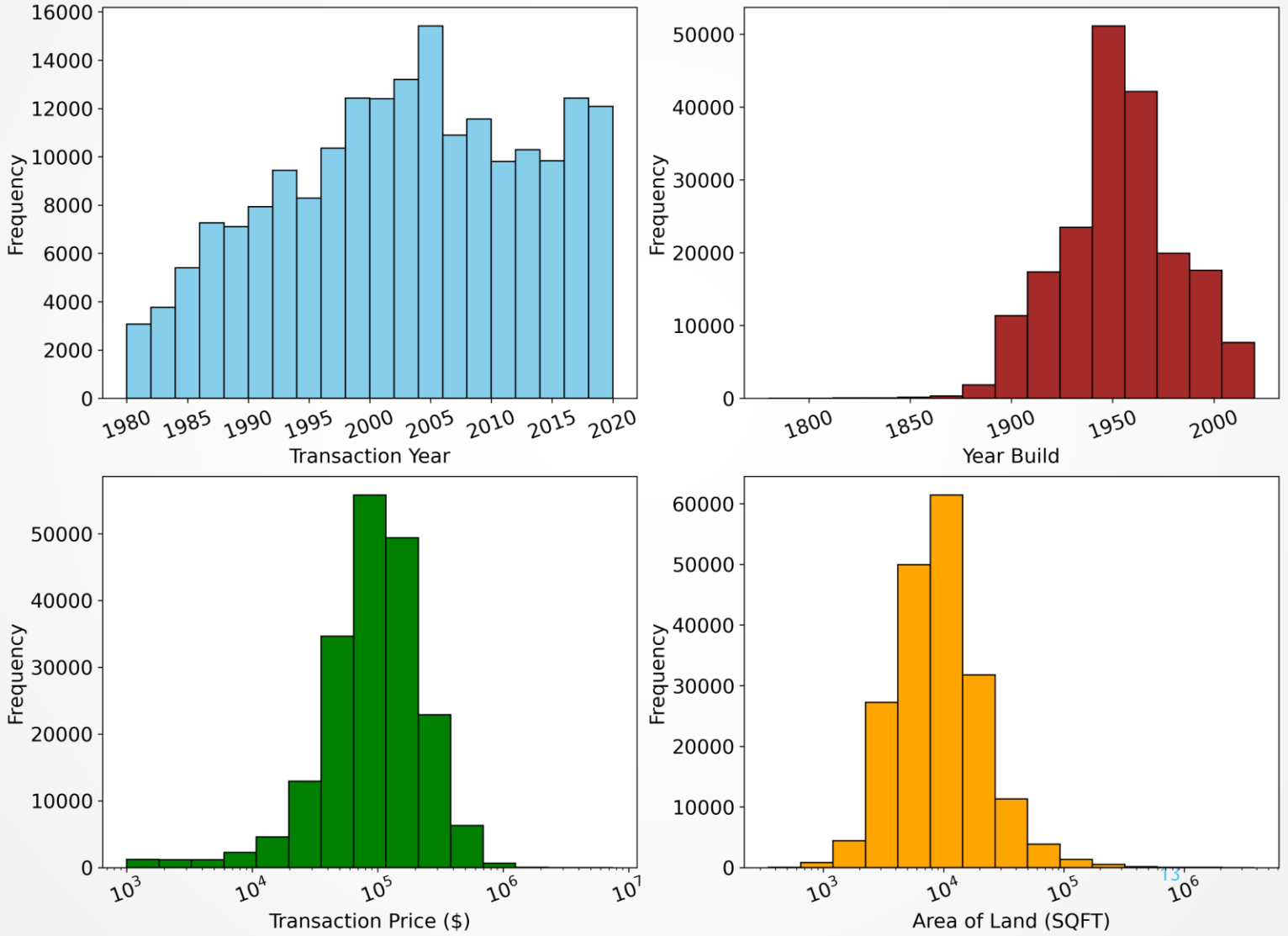
Feature Engineering – features overview

Variable Type	Variable (p=25)
Continuous variable	▪ TIME ▪ YEARBLT, STRORIES, BEDROOMS, ADJUSTBATHS, BSMTGARAGE, FIREPLACES, BASEMENT, LOGLOTAREA, LOGFINISHEDLIVINGAREA, PRICEPERSQFT ▪ LATITUDE, LONGITUDE ▪ ANXIETY, OLD, POOR, VACANT
Ordinal variable	▪ GRADERANK, CDURANK, SCHOOLRANK
Nominal variable	▪ NEIGHCODE, EXTFINISH_DESC, STYLEDESC, MONTH, TIERS
Target	▪ LOGPRICE

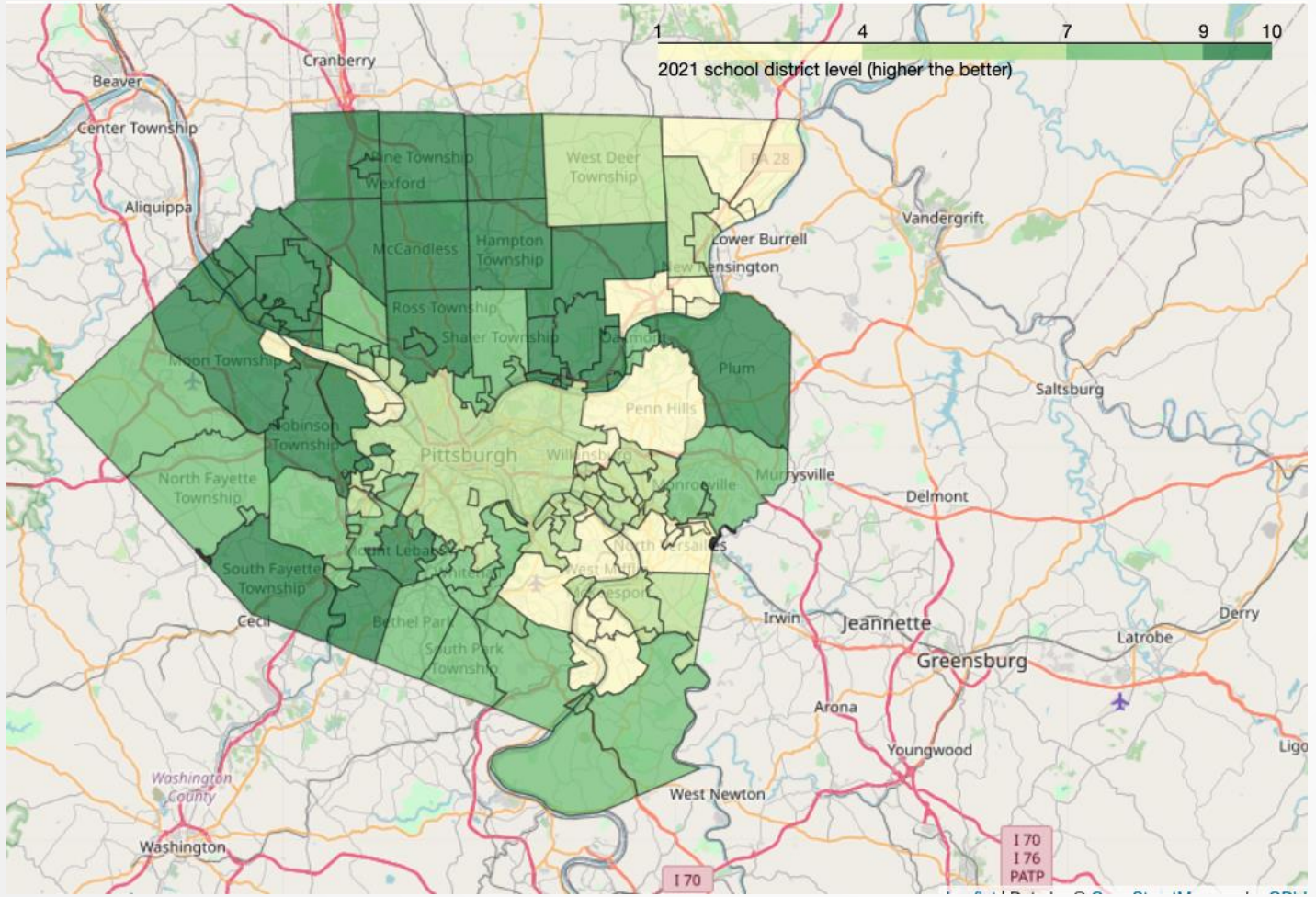
Logarithm transformed variables
Variables created by calculation
Variables created by categorization

- ❖ TIME = SALEDATE – 19900131, in month
- ❖ PRICEPERSQFT = FAIRMARKETTOTAL / TOTAREA

Feature Engineering – Feature Visualization



Feature Engineering – Feature Visualization

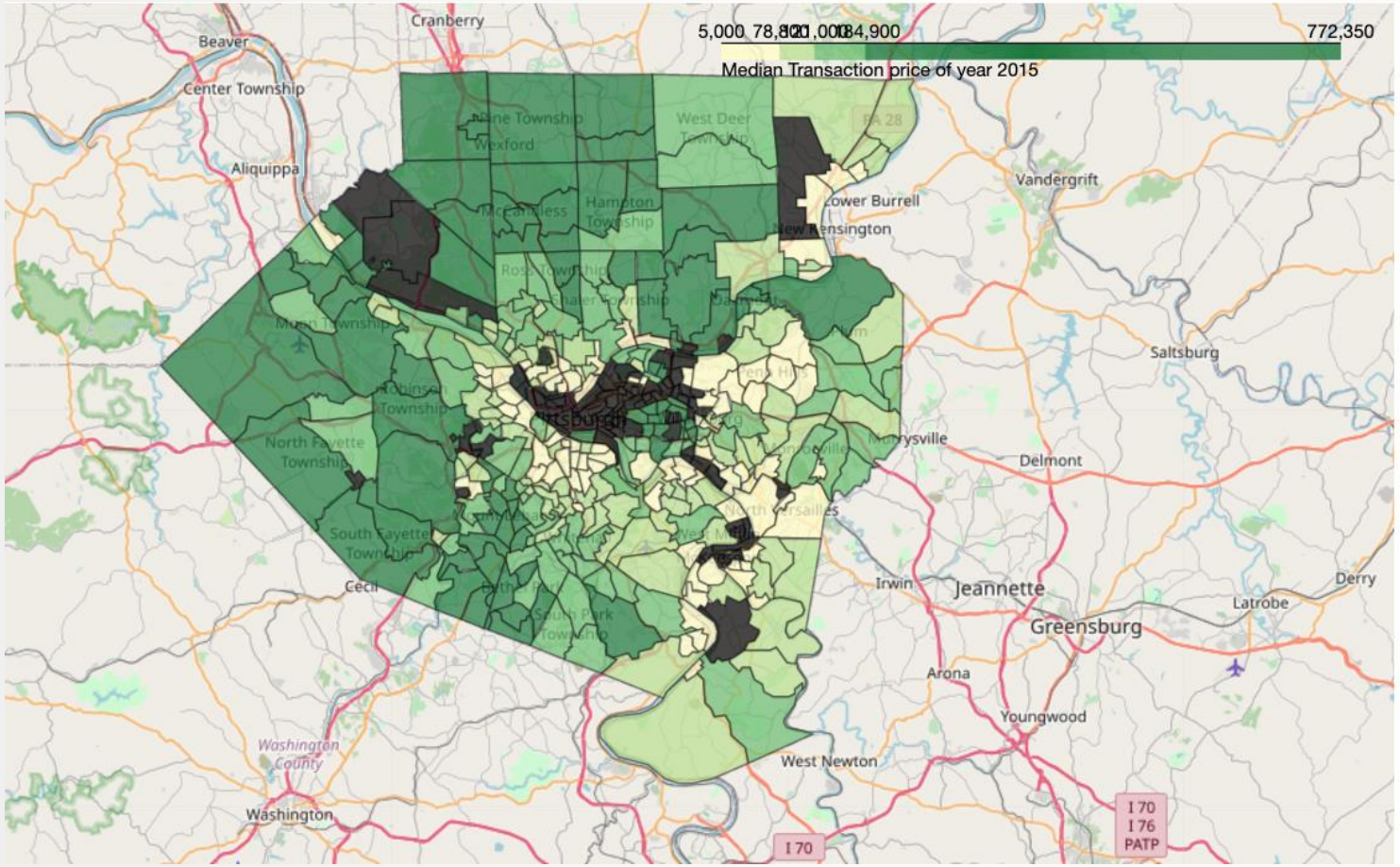


School District by Municipal

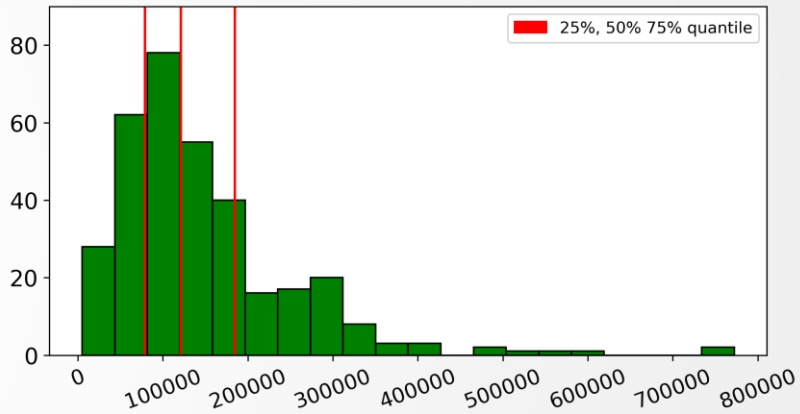
Grade	Score*
A+, A, A-	10, 9, 8
B+, B, B-	7, 6, 5
C+, C, C-	4, 3, 2
D	1

*score, the higher the better

Feature Engineering – Feature Visualization



Median transaction price in 2015 by community



ML Pipeline – ML model

- ✓ Training/testing split
- ✓ LightGBM
- ✓ Shapley ML Interpretation
- ✓ Optuna hyperparam optimize

Spatial Temporal Modeling via LightGBM

- Let z_{it} be the value for the ith house at time t.
- Use LightGBM as a functional approximator to model each house's appreciation trajectory.
- Use Bayesian optimization search for the hyperparameter ϑ^* that minimizes the RMSE on the validation set. 1000 trials were performed.

Appreciation of value in time.

$$z_{i,t} = F_\theta\left(\text{  }, t; \vartheta\right)$$

Hyperparameters

$i \in \{1, \cdots, 114308\}$
$t = \{19900131 \cdots 20201231\}$

House features such as location, number of rooms etc.

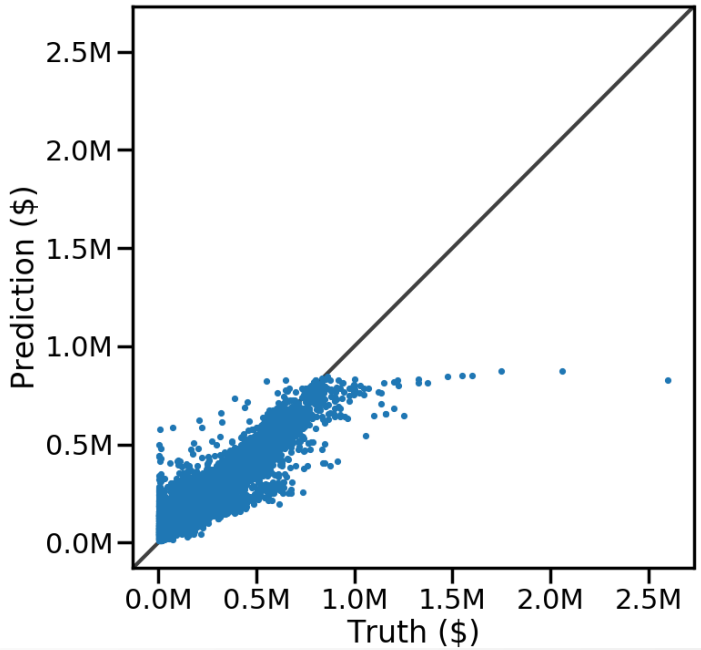
Training LightGBM

- Randomly split the cleaned housing sales data into training set (80%) and validation set (20%)

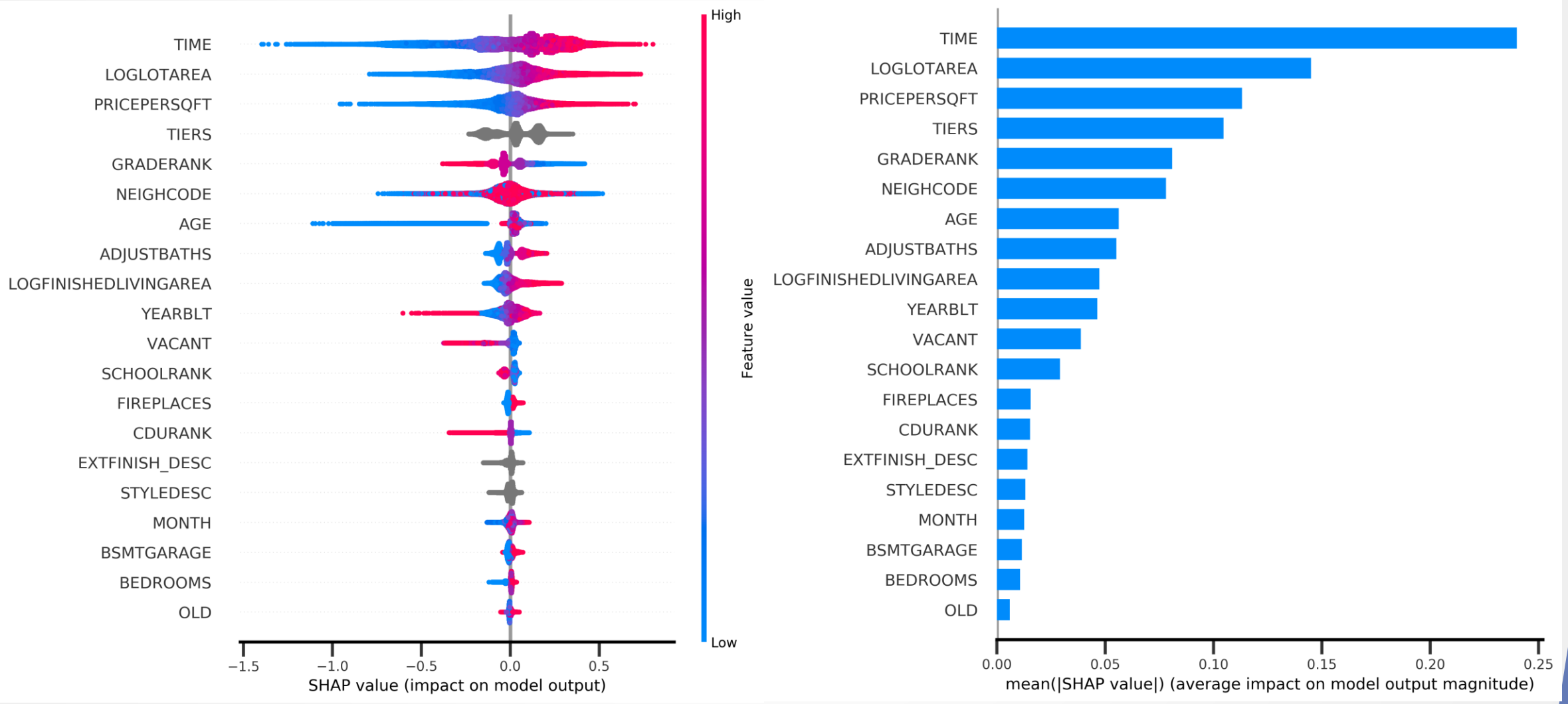
Split	Count	RMSE	Pearson
Training	138,619	3.567	0.902
Validation	34,655	3.870	0.882

RMSE is scaled with a unit of $10K.

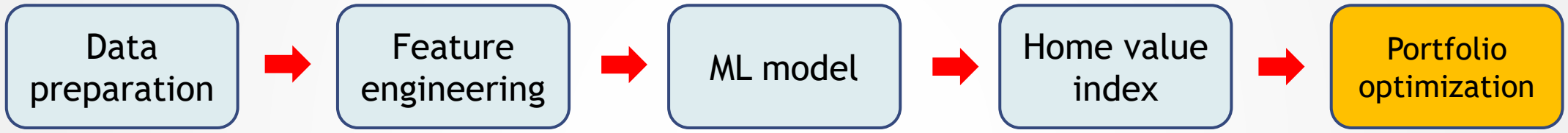
Prediction versus the truth on the validation dataset.



Interpret LightGBM models via the Shapley Values



ML Pipeline – ML model



- ✓ Zillow Index methodology
- ✓ Henderson MA Filter
- ✓ Seasonal adjustment
- ✓ Mean variance optimization

Zillow Home Value Index (ZHVI) Methodology

Given our estimation of z_{it}, we follow Zillow's HVI mythology and consider the total market appreciation as a weighted average of each home's appreciation in the property's universe.

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

ML Estimated value appreciation of home i from time $t-1$ to t.

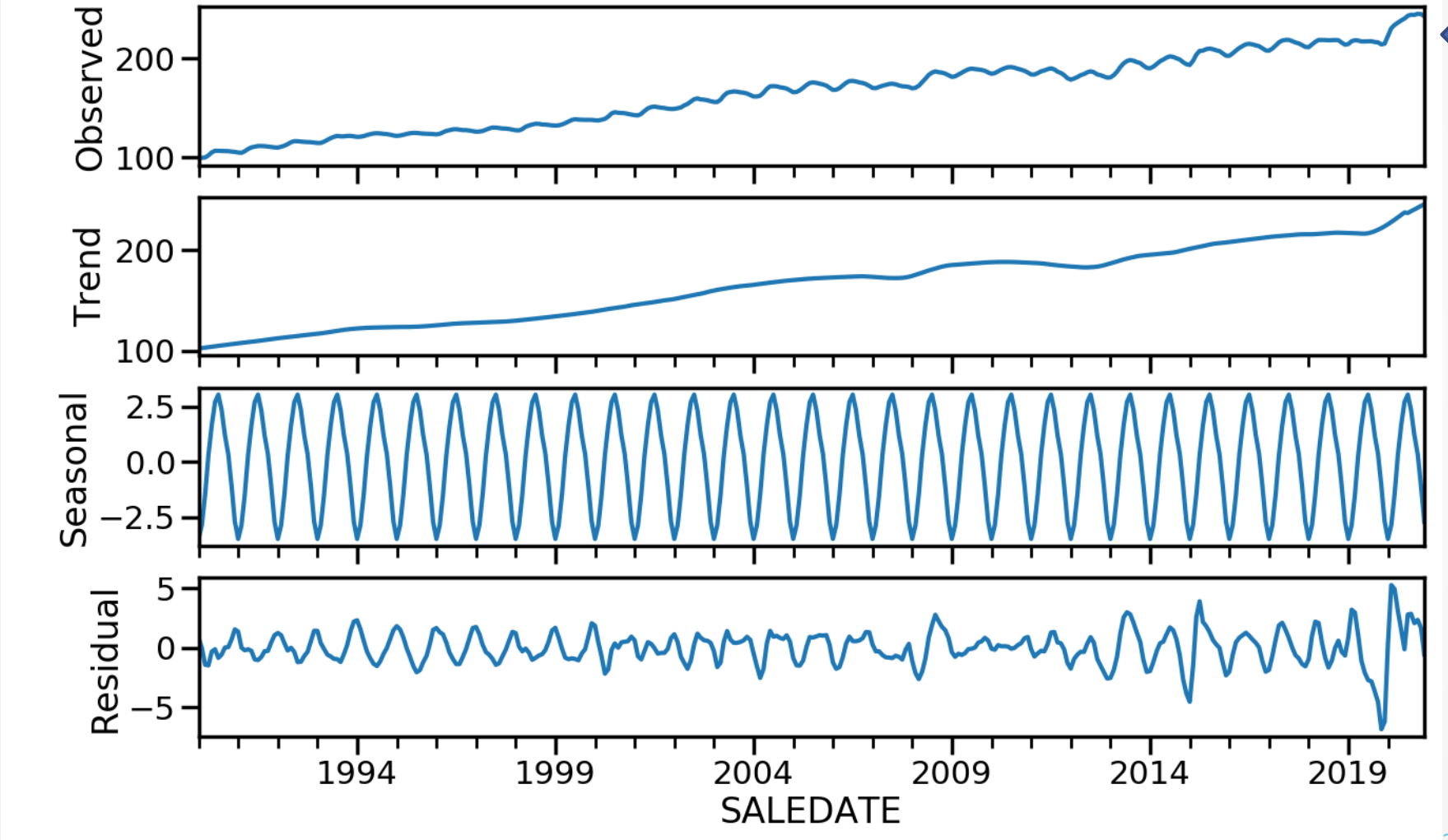
$$w_{i,t-1} = \frac{z_{i,t-1}}{\sum_{j=1}^{N} z_{i,t-1}}$$

Home i 's share of the total market value

$$R_t = \sum_{i=1}^{N} w_{i,t-1}\, r_{i,t}$$

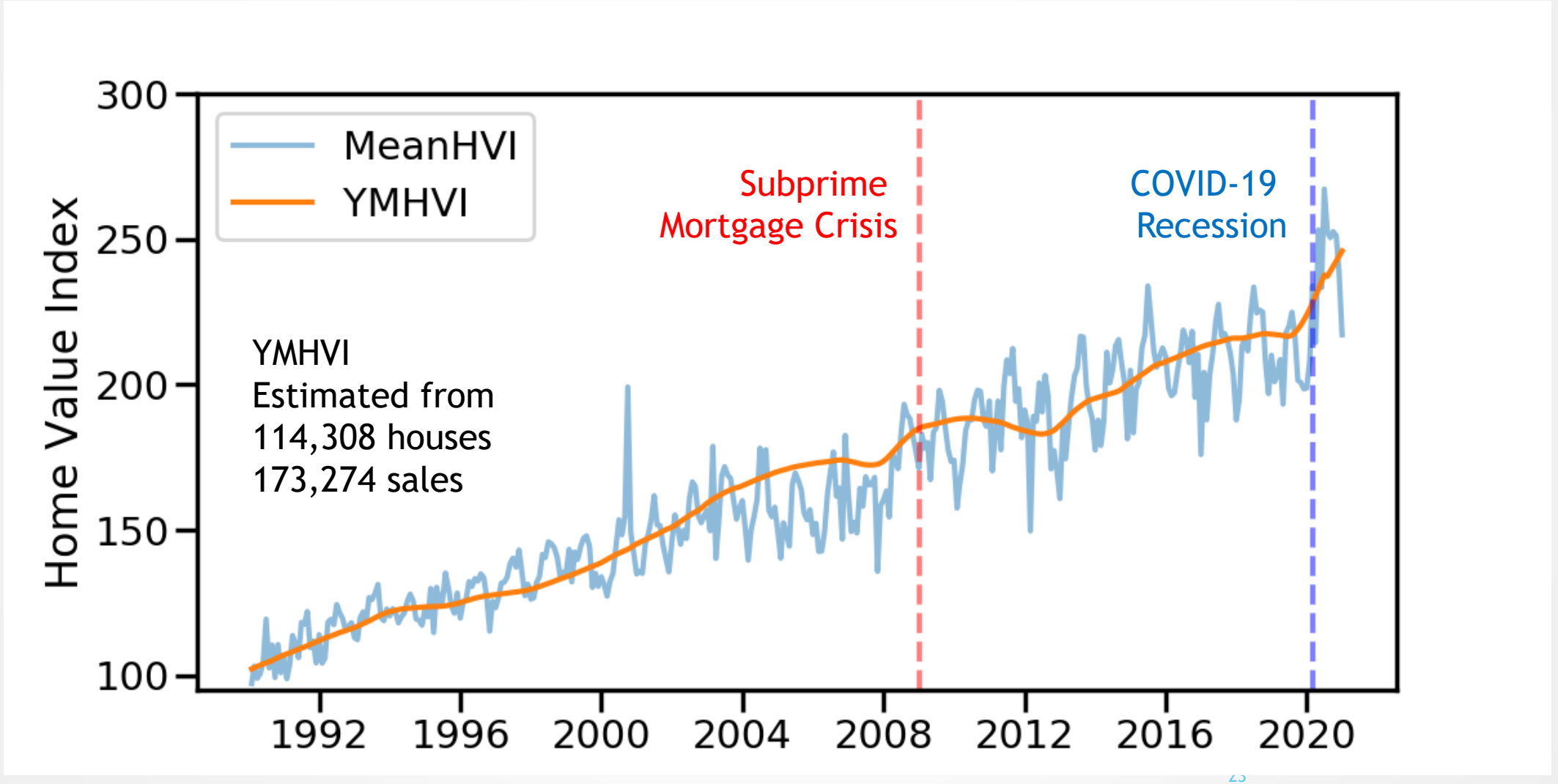
ML Estimated value appreciation of the total market from time $t-1$ to t.

Allegheny County Home Value Index (HVI)



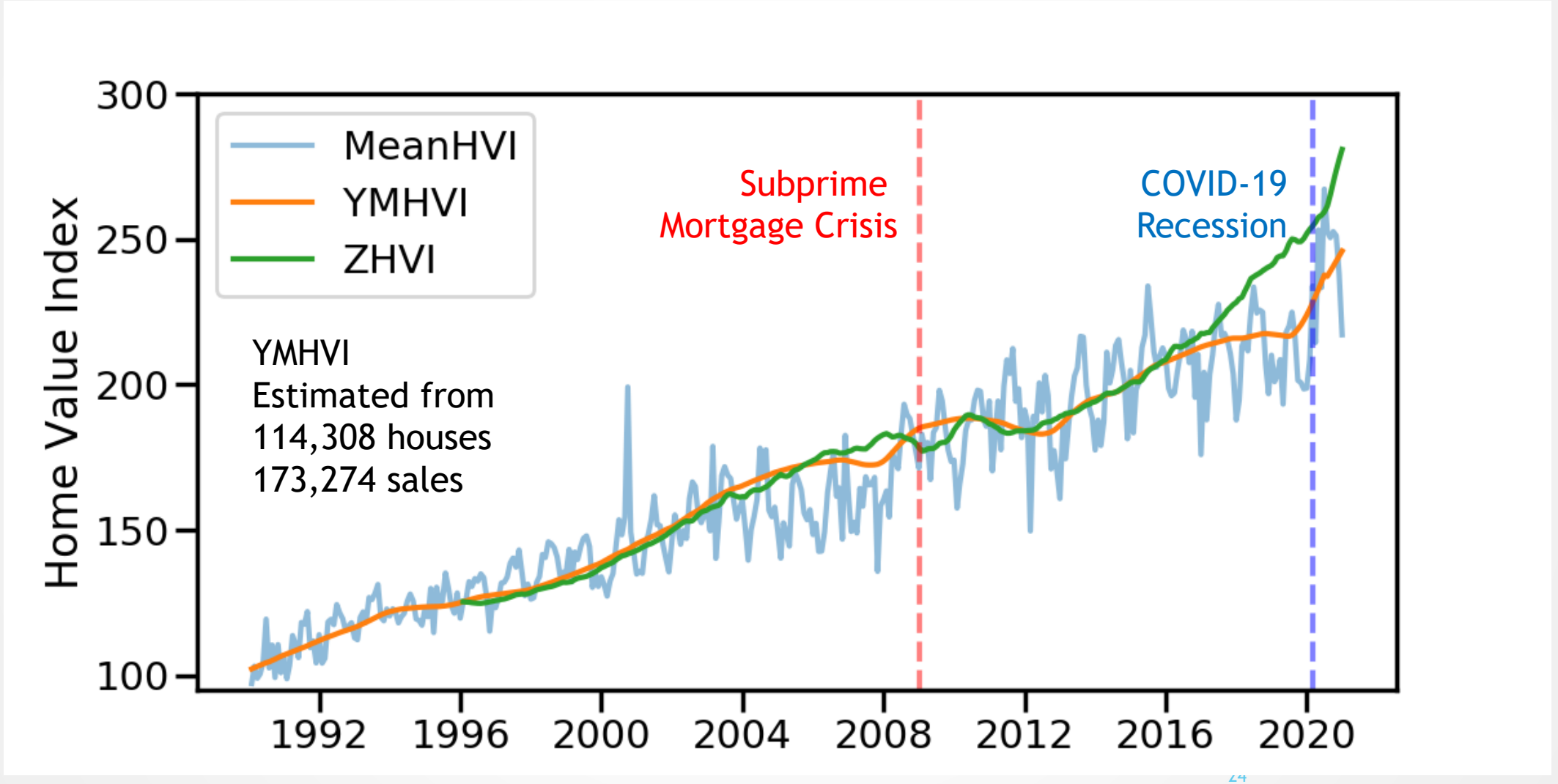
Used Henderson MA Filter to smooth the raw signals.

Allegheny County Home Value Index (HVI)



Assume the HVI level to be 100 on 1990-01-31.

Allegheny County HVI with Zillow (ZHVI)



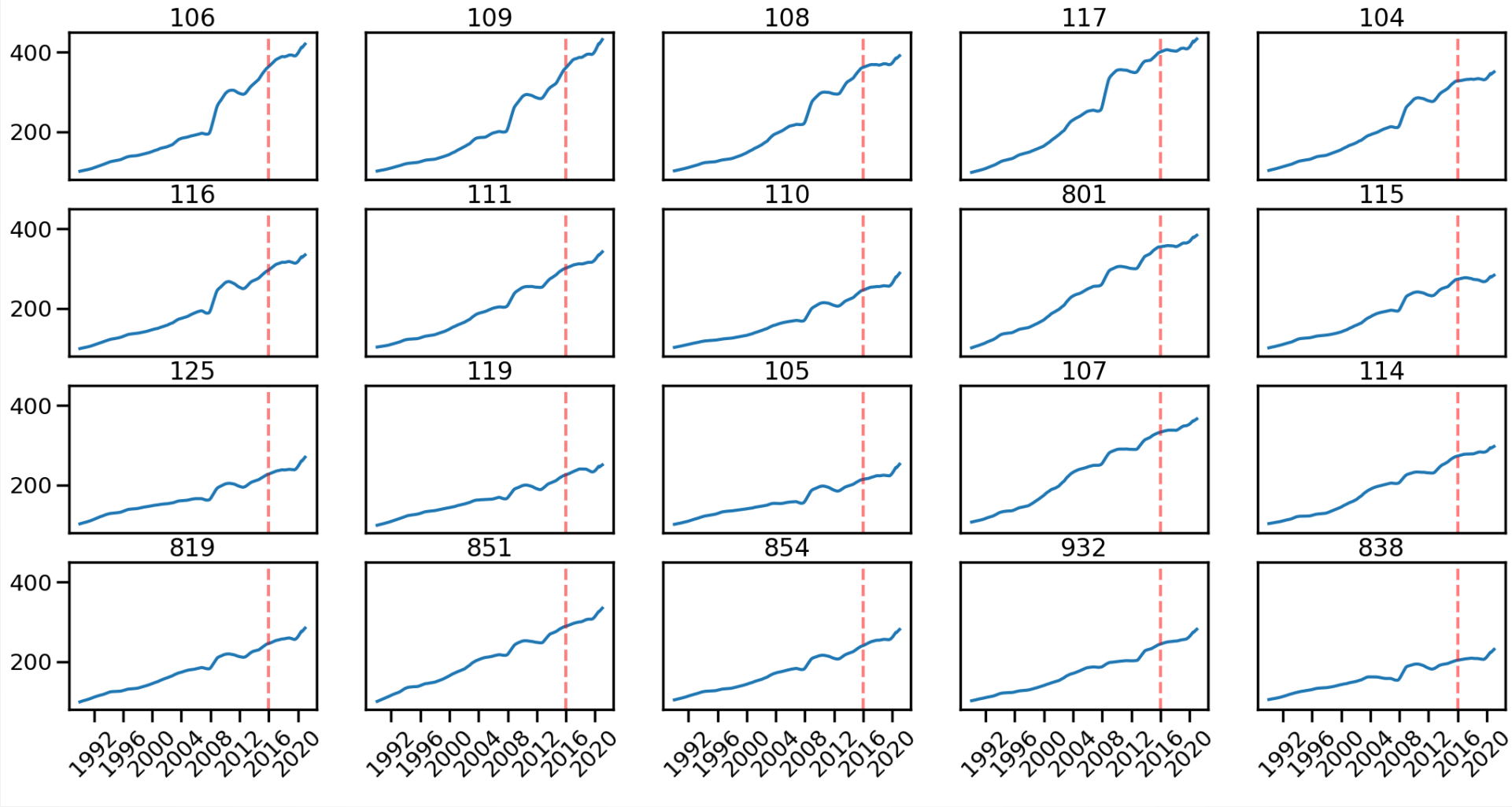
Assume the HVI level to be 100 on 1990-01-31.

Return & Risk Summary for the **Top 20 MUNICODEs**

MUNICODE	Count	Return (%)	Risk (%)	MUNICODE	Count	Return (%)	Risk (%)
106	160	6.59	9.41	125	124	3.21	5.41
109	210	6.24	8.16	119	2882	3.18	4.61
108	276	5.36	6.79	105	166	3.13	5.87
117	384	4.77	8.16	107	237	3.10	3.22
104	176	4.67	6.55	114	2815	3.09	3.17
116	454	4.56	8.32	819	988	3.05	4.65
111	664	4.15	4.55	851	229	3.04	3.72
110	1289	3.91	5.34	854	793	2.90	4.44
801	137	3.61	4.08	932	599	2.81	3.24
115	1119	3.59	5.44	838	207	2.57	6.34

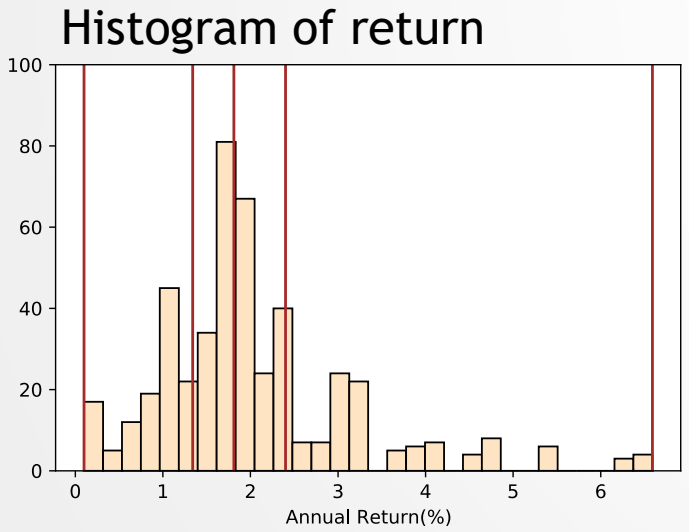
Geometric average and std of annual returns from 2006-01-31 to 2015-12-31

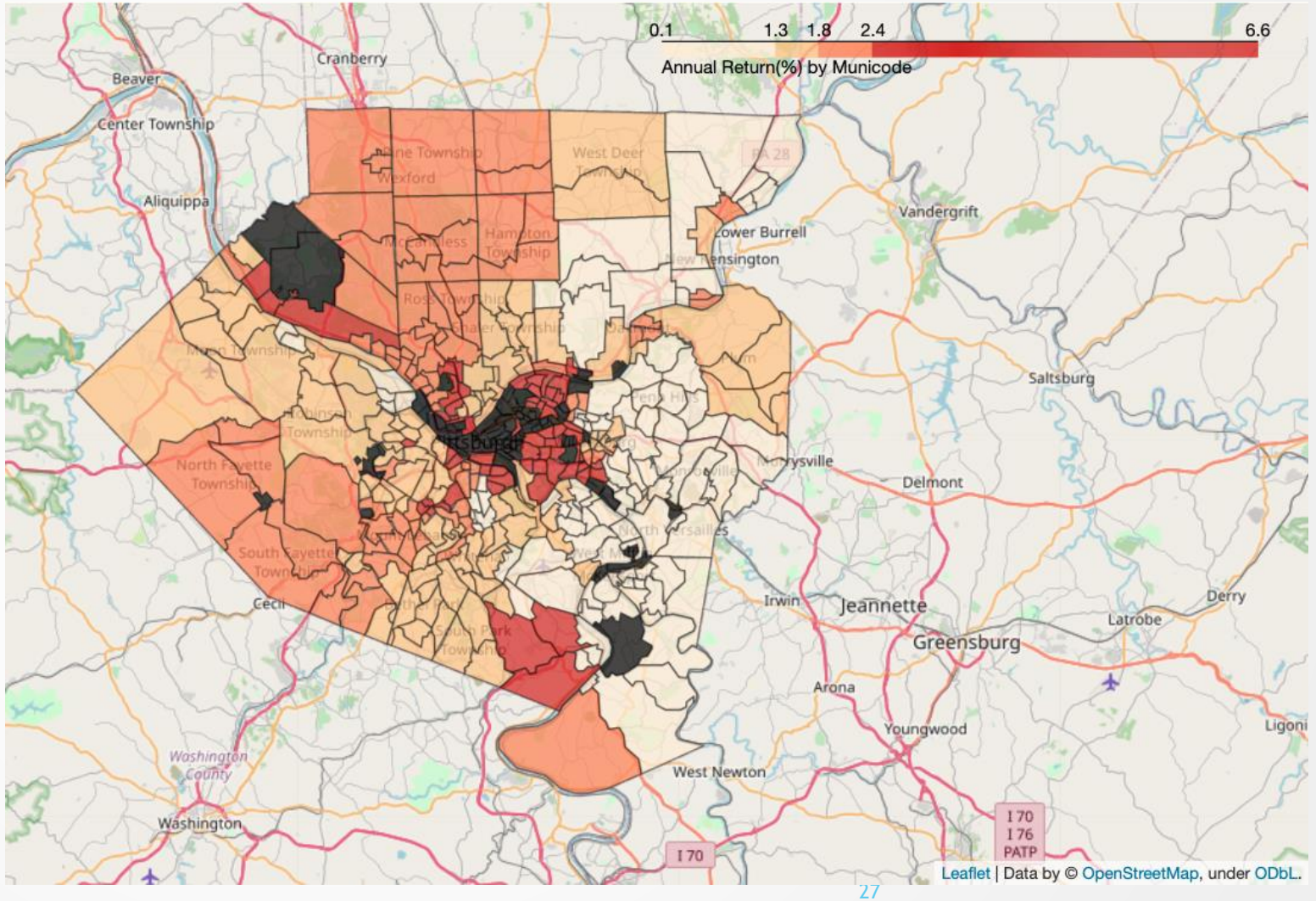
HVI for the **Top 20** MUNICODEs (by avg. return)



Ranked by the geometric average of annual returns from 2006-01-31 to 2015-12-31

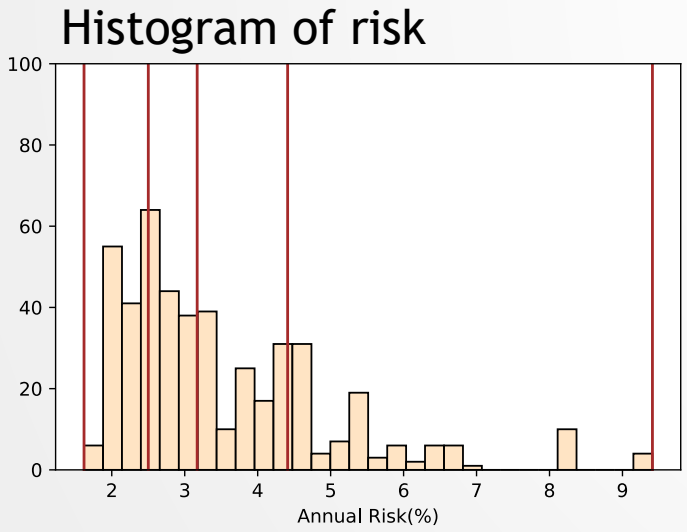
Geospatial distribution of Annual Return (%)

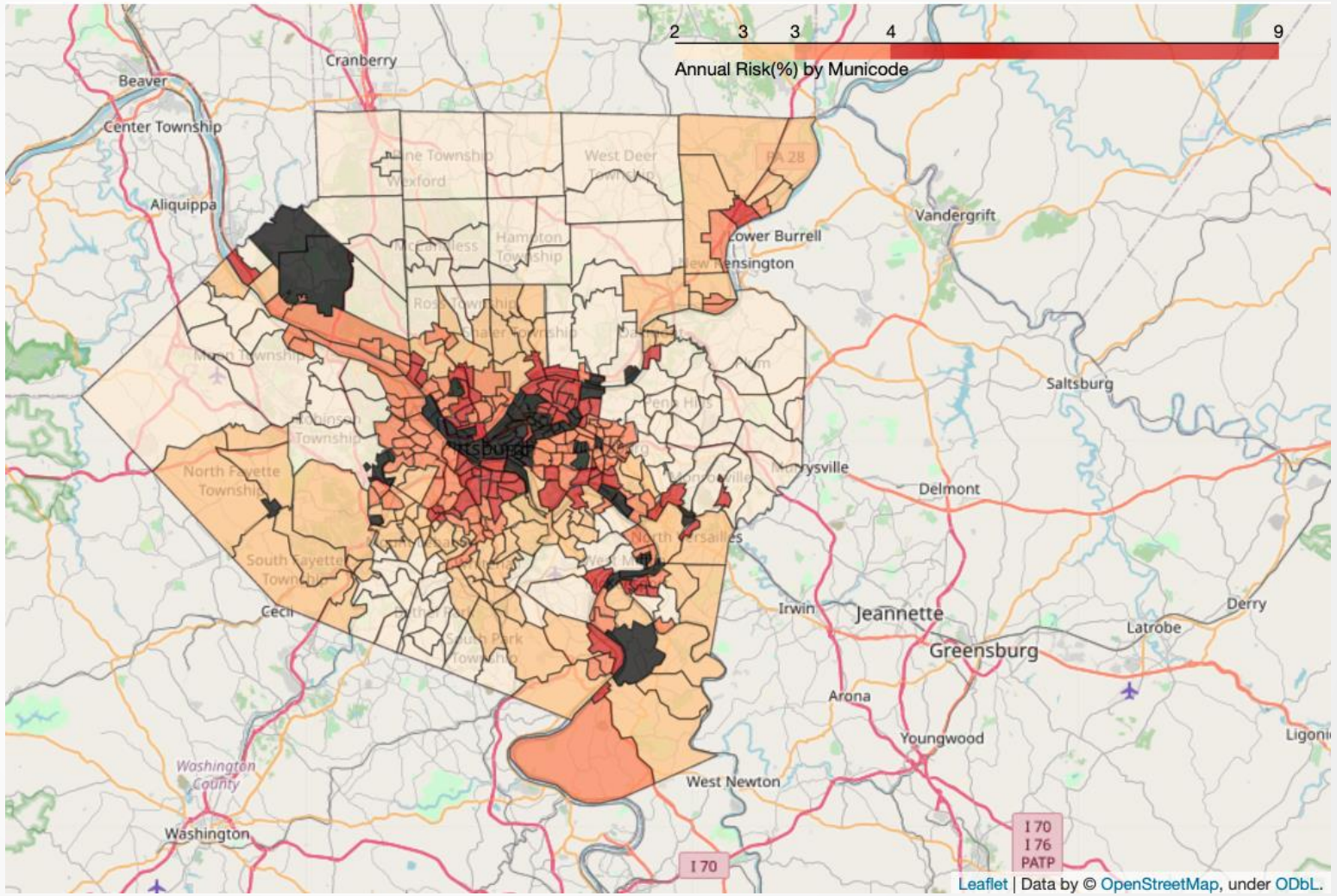




Geometric average and std of annual returns from 2006-01-31 to 2015-12-31

Geospatial distribution of Annual Risk (%)



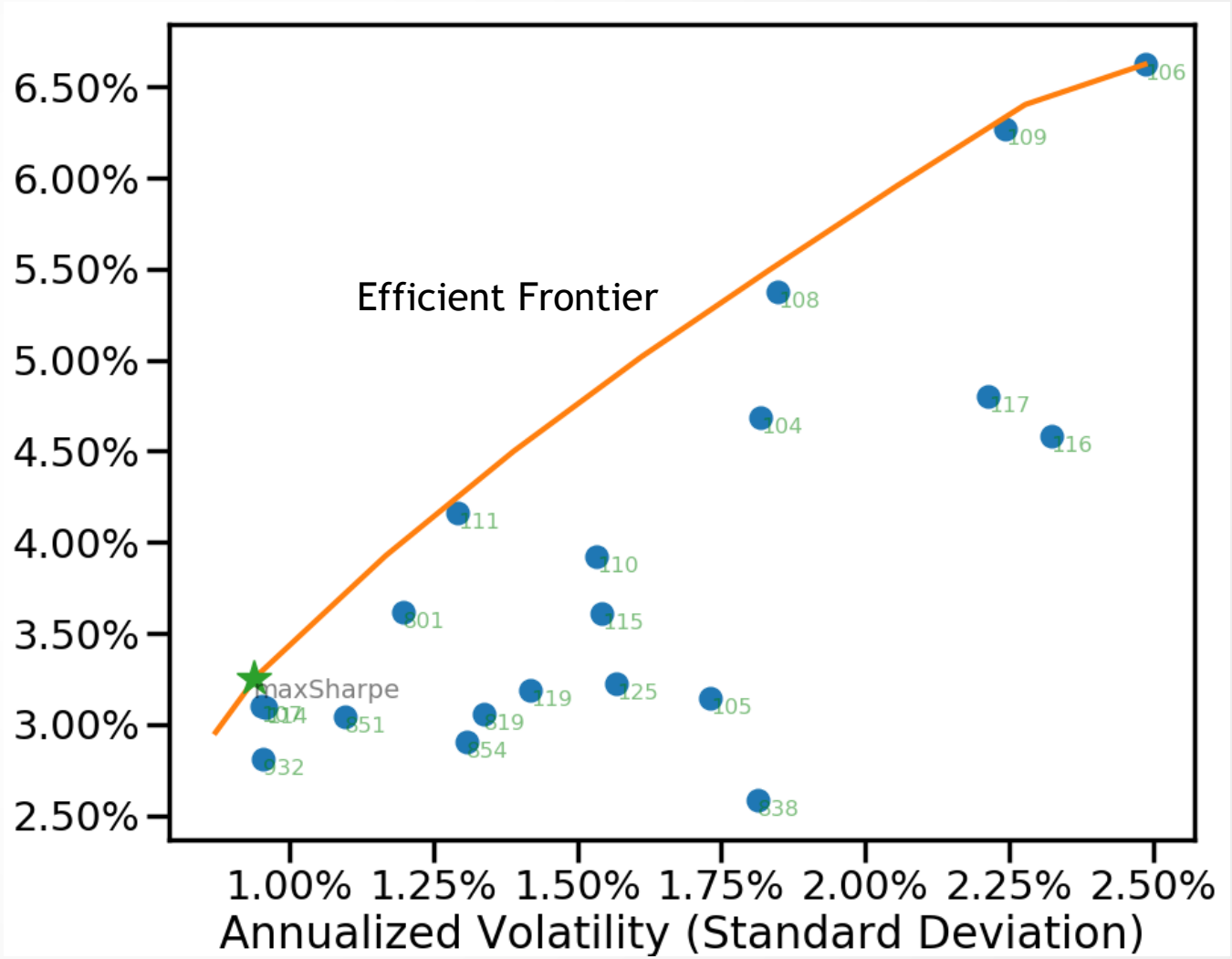


Geometric average and std of annual returns from 2006-01-31 to 2015-12-31

Mean-Variance Efficient Frontier

We compute optimal assets allocation weight from MVO.

$$\begin{cases} \underset{w}{\text{maximize}} & w^\top \mu - \dfrac{\delta}{2} w^\top \Sigma w \\ \text{subject to} & \displaystyle\sum_{i=1}^{n} w_i = 1 \\ & 0 \le w_i \le 1, \ \forall i = \{1, 2, \ldots, n\}. \end{cases}$$



Portfolio Optimization Results

Strategy	Cumulative Value ($)	Cumulative Return	Annual Return	Sharpe Ratio	Maximum Drawdown
Target Risk 1pct	$5.79 million	15.83%	2.92%	2.37	-0.23%
Equal Weighting	$5.64 million	12.83%	2.42%	1.57	-0.44%
Min Variance	$5.61 million	12.22%	2.29%	1.92	-0.17%
Max Return	$5.78 million	15.63%	2.86%	2.23	-0.57%
Max Sharpe	$5.64 million	12.49%	2.35%	2.14	-0.09%
Risk Parity	$5.64 million	12.73%	2.39%	1.60	-0.34%
Overall Market	$5.92 million	18.32%	3.42%	1.98	-0.42%

Asset Allocation

106	109	108	111	801
$504,034	$2,316,692	$528,993	$1,483,404	$166,875

Summary

- LightGBM was used to build a Spatial Temporal model for the Single Family property in Allegheny.
- Zillow Home Value Index (ZHVI) methodology was applied to build HVI both county and each municipal levels.
- Given a portfolio of HVI indexes, we chose the top 20 municipals with the highest averaged annual returns (for the past ten year).
- The mean-variance optimization was used to find the optimal asset allocation.
- We further verify the investment performance by evaluating cumulative return, Sharpe ratio etc.

Thank you for your attention!

Question?