Backup Slides

Outlier Detection

Clustering

Feature Engineering

Nick Park

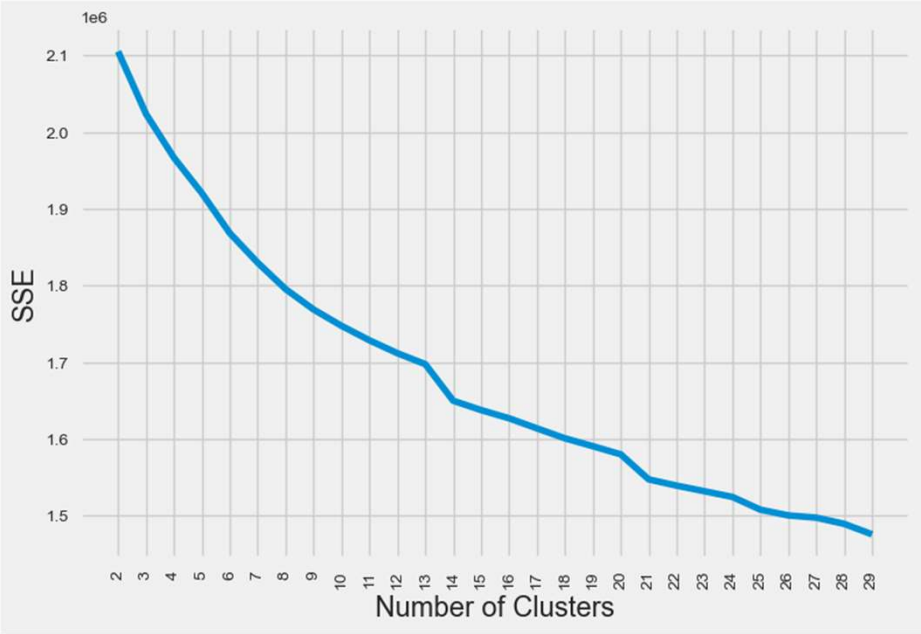
Significant Outliers

- Top 5 zip codes based on Cook's Distance
  - 98571 (2021) - Grays Harbor County, Washington (Pacific Beach)
  - 94129 (2021) - San Francisco, California (Presidio of San Francisco)
  - 02199 (2013) - Boston, Massachusetts (Prudential Center)
  - 02199 (2018)
  - 02199 (2014)

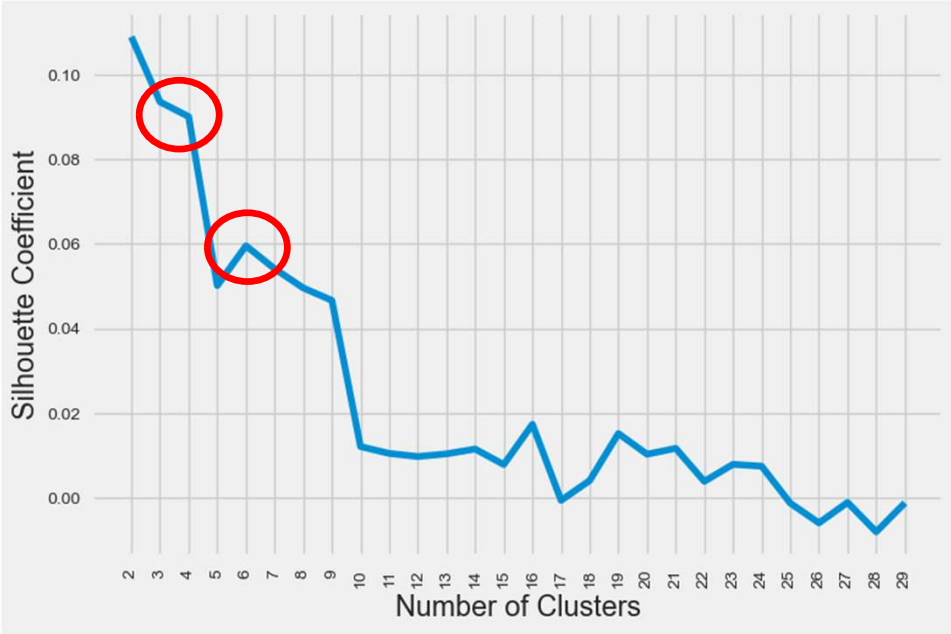
		Cluster 1	Cluster 2	Cluster 3
Characteristic		Average population + higher income + higher education	Large population + average income + lower education	Low population + low income
Avg. Population		13,852 *(Compared to Cluster 1)*	19,086 (+38%)	5,277 (-72%)
Avg. Median Income		77,138	68,965 (-11%)	55,110 (-29%)
Geographical Mobility	Less than High School	844	1,705 (+102%)	558
	High School	2,428	3,335 (+37%)	1,302
	Some Degree	2,641	3,853 (+46%)	1,065
	Bachelor	2,147	2,589 (+21%)	434
	Graduate	1,456	1,522 (+5%)	245

Clustering

Sum of Squared Error (SSE)



Silhouette Coefficient



Clustering

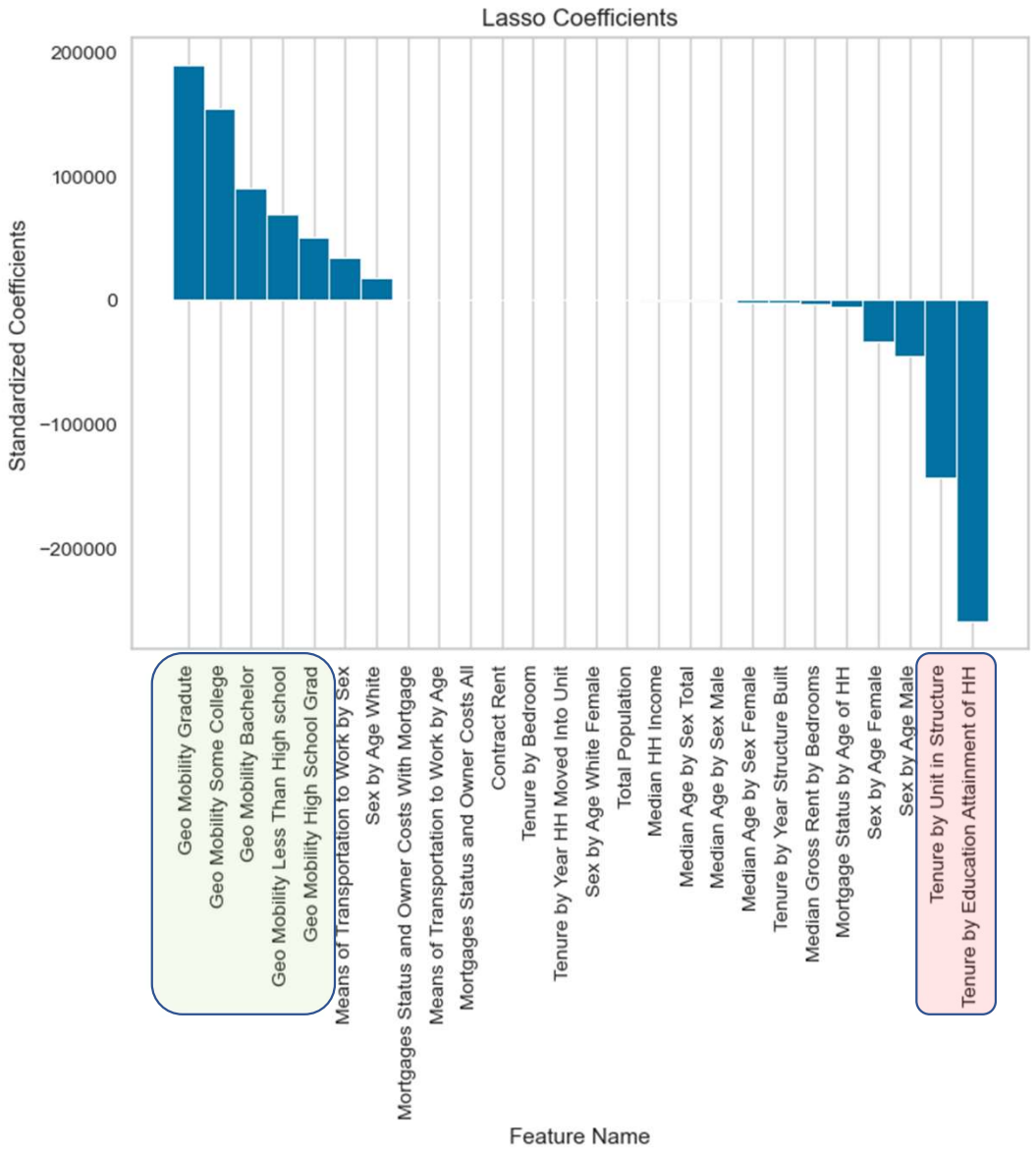
- Used data from 02/2015 ~01/2023
  - Too many missing values prior to 02/2015 (more than 17%)
  - Monthly percent change in price was used
  - Log(x+1) and standardization were applied
- Used K-means
  - Computationally inexpensive compared to hierarchical clustering

Feature Engineering

- Imputed missing/negative values with each year's median value for:
  - Median Age by Sex (All/Male/Female)
  - Median Household Income
  - Median Gross Rent by Bedrooms
    - All null values from 2011- 2014
    - Linear interpolation from later years
- Log(X+1) transformation prior to standardization
- Improved MAE by $30,769 (95,487 → 64,718)

Before Feature Engineering

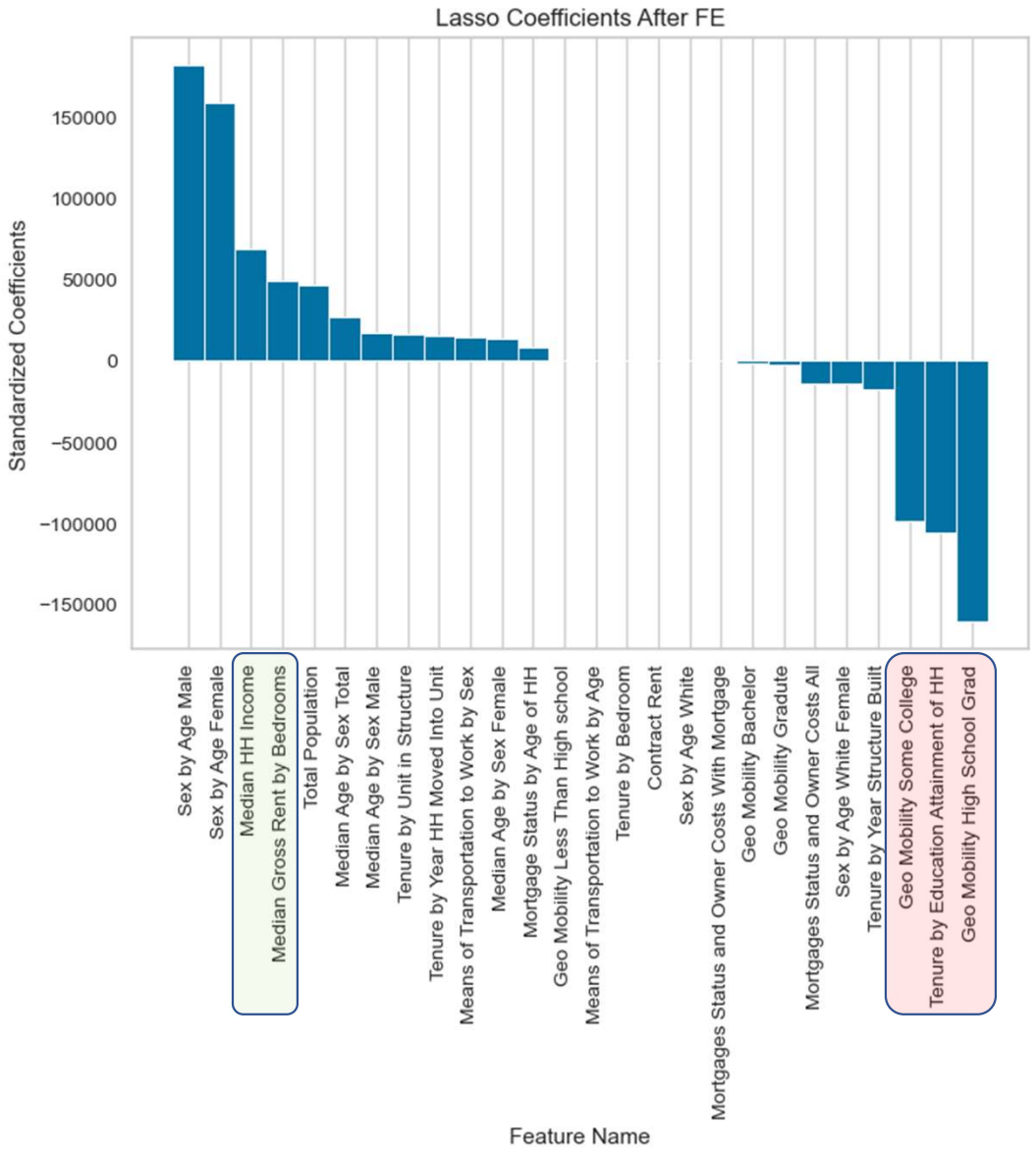
MAE = 95,487



After Feature Engineering

MAE = 64,718

- 30,769 less than the first model



Others

- Used January of each year from 2011 to 2021 for modeling
  - Census data was available only from 2011 to 2021
  - For missing values, performed linear interpolation
- Used LASSO regression model
  - Feature selection to identify the driving/ factors