Single-Family Home Flips by Census Tract

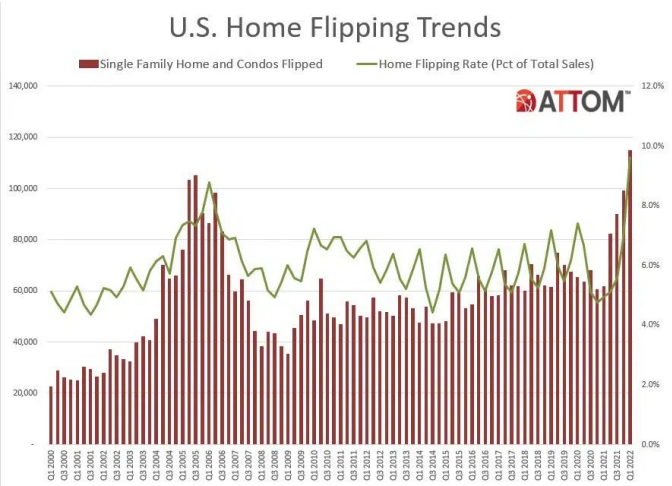
a research report composed through the aid of clustering and dimensionality reduction in methods of unsupervised learning

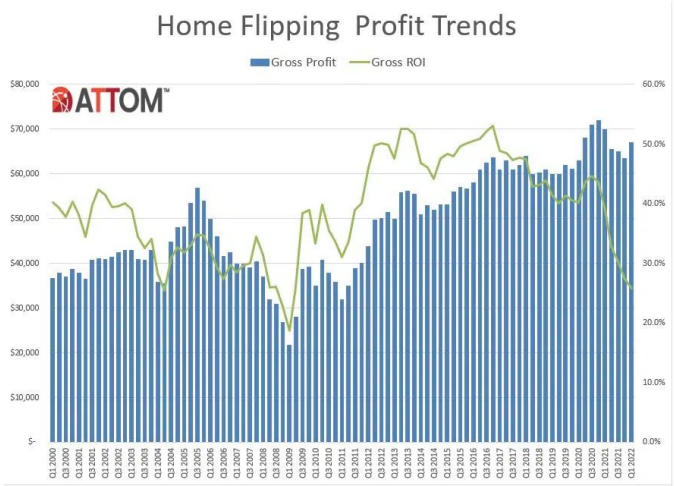
Overview and Problem Statement

- Home flipping is a highly profitable area of real estate to be involved in
  - More control over the marketability of the home
  - A niche sector in an extremely competitive field of work
- In some cases, this process may also give rise to gentrification
  - Displacement of less wealthy inhabitants of gentrified areas
  - Perhaps it could be valuable to organize the data in a way that reflects differences such as this
- **Our Question Today:** How many different groups (or, in a sense, “tiers” of home flipping) can recent data be effectively divided into?
  - Let’s attempt to visualize this sorting to better understand how the data is compiled

Cultural Fit: The Dramatically Changing Housing Market

- Due to the COVID-19 pandemic, the housing market skyrocketed
  - Why? → work from home
    - Larger interest in moving due to no commute
    - Sellers gained an advantage, which promoted the option of home flipping
- Sales and profits for home flipping have been up since 2014:





(ATTOM, June 2022)

Data Analyzed: FLIPS

- *Before clustering:*

  dropped [**'OBJECTID'**] and [**'YEAR'**] from the data

  - ➢ separated the target variable, [**'FLIPS'**], as well

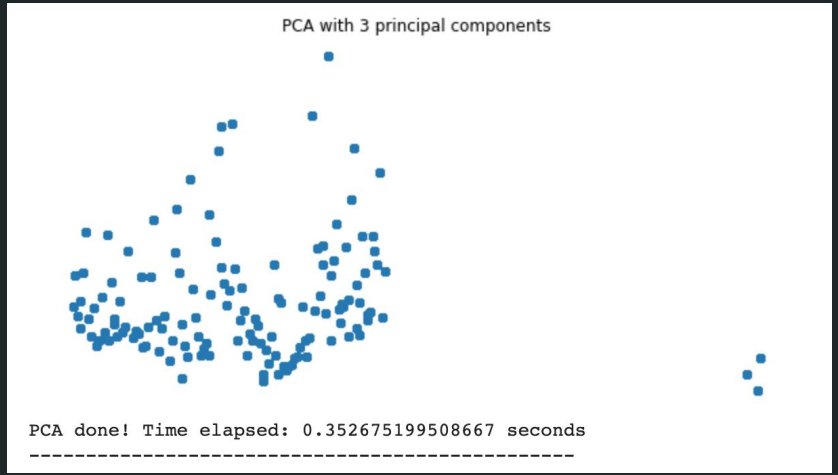
- Raw Data Available Here

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2025 entries, 0 to 2024
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   OBJECTID        2025 non-null   int64
 1   TRACT           2025 non-null   int64
 2   YEAR            2011 non-null   float64
 3   FLIPS           2010 non-null   float64
 4   Shape__Area     2025 non-null   float64
 5   Shape__Length   2025 non-null   float64
dtypes: float64(4), int64(2)
memory usage: 95.0 KB
```

	OBJECTID	TRACT	YEAR	FLIPS	Shape__Area	Shape__Length
0	1	2500	2006.0	0.0	0.000118	0.052050
1	2	2600	2006.0	3.0	0.000149	0.050714
2	3	5600	2006.0	3.0	0.000357	0.102853
3	4	6800	2006.0	0.0	0.000086	0.038414
4	5	6000	2006.0	1.0	0.000157	0.066840
5	6	5900	2006.0	0.0	0.000217	0.078608
6	7	5700	2006.0	1.0	0.000596	0.153638
7	8	4900	2006.0	0.0	0.000131	0.049355
8	9	5801	2006.0	1.0	0.000218	0.088168
9	10	4800	2006.0	1.0	0.000160	0.059893
10	11	4700	2006.0	0.0	0.000238	0.075637

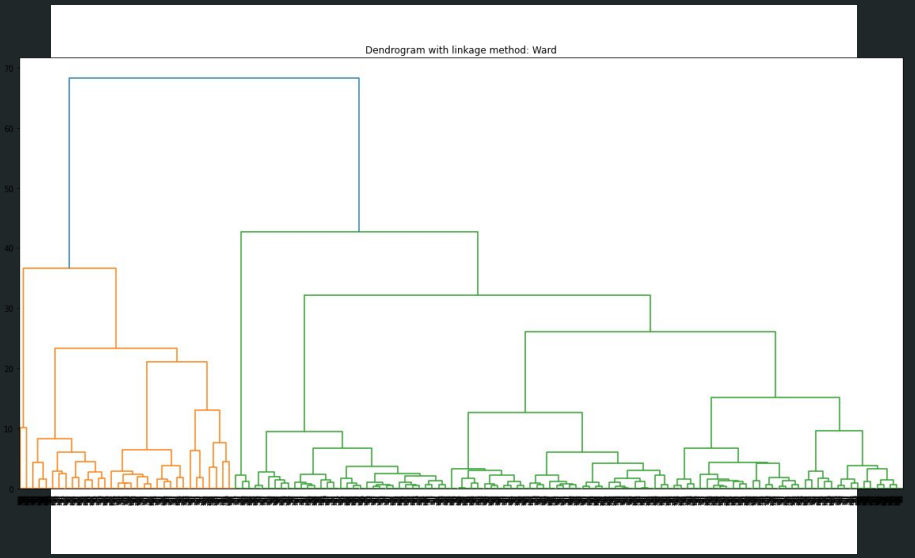
Visualization of Data

- Dimensionality reduced best by PCA ➔



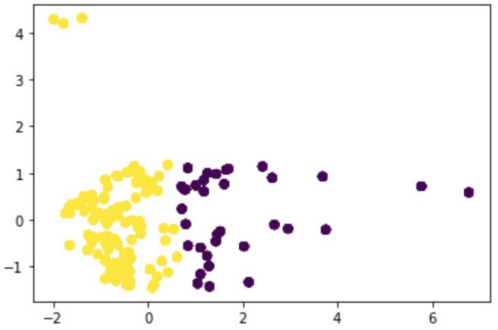
Results: Hierarchical with Ward linkage

- **Adjusted Rand Index**: 0.021
  - **Silhouette**: 0.487



Was the [‘TRACT’] column an outlier of sorts?

- Much greater value for standard deviation than other tested variables
- Inconsistent manner of scattering among data visualizations and clustering
  - K-means at its seeming best still separated more than half of the data inaccurately (see below)
  - t-SNE improved in performance as `perplexity` increased
- There appears to be 3 principal components in the data
  - Most of the clustering techniques seem to exhibit 3 main clusters
  - Clusters beyond the 3rd seem to often contain only 1 data point



Comparing k-means clusters against the data:

col_0	0	1
row_0		
0	164	346
1	814	701

Thank you!

Questions, comments, or recommendations for discussion?