Fund That Flip- Business Intelligence Challenge

Exercise #3: Creative Analysis

Sean Yang

Introduction

- Zillow's ordinal measure (0 to 10), the Market Health Index, conveys the current health of a region's housing market in relation to other markets within the United States.
    - To elaborate on such , if a metro area has a value of 8 on the Market Health Index, the metro is healthier than 80% of all metro areas covered by Zillow in the US.
- The Market Health Index is formed from ten different metrics. To capture the recent and continued rebounds from housing value troughs the following are included in the dataset: the month-over-month and year-over-year change in ZHVI, the percentage of homes selling for a gain, as well as the percentage change in the Zillow Home Value Index (ZHVI) forecasted for the coming year.

Decision-Making

- The given instructions for the following exercise were quite vague, understanding that it was designed to avoid limiting us and our approach. After considering several approaches to the creative task through careful examination over the multiple datasets provided, it was of my best interest to pursue the dataset in which contains all information on **cities** in the U.S.
- The reason behind this is simply due to the fact that cities are more versatile in terms of inclusiveness to the other region types (metros, zip codes, counties, and states). To elaborate, cities define a metro, multiple cities can share the same zip code, many cities can be grouped within a county and state. This will prove beneficial to the machine learning algorithm I have prepared for this exercise.

Cluster Analysis- What & Why?

- In the most simple explanation possible, a cluster analysis (or cluster) is the execution of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- The reason why I chose to perform such task onto the dataset provided by Zillow was to analyze the similarities and differences found within various sub-groups of which are distributed by ML. In doing so, the goal is to see if there are any patterns that may help us discern any influential factors that contribute to a city's respective market health index that may/may not have been previously acknowledged.

Purpose of Analysis

- “A data analyst reviews data to identify key insights into a business's customers and ways the data can be used to solve problems. They also communicate this information to company leadership and other stakeholders.” - Danielle Gagnon
- The following questions to be answered by potential discovery:
  - Why/How are each city within each given cluster related to one another?
  - Which cluster of cities feature the highest average for market health index (MHI)?
  - Which factors determine a higher MHI, if any?

Data Wrangling

- A common issue when evaluating all five datasets provided on data.world was the fact that there was a lot of missing values/data.
- 28,650 missing values (~19% of entire dataset)
- 10,958 rows * 19 columns worth of missing values in entire dataset (208,202 total)
- This was still the lowest percentage of missing values amongst all five datasets which was a positive note (state dataset had ~31%, highest percentage)

```
[3] #Find the total number of missing values from the entire dataset
    city_df.isnull().sum().sum()

    39608
```

Data Wrangling (Cont'd)

- This screenshot reveals which the distribution of missing values amongst all columns within the dataset

```
[ ] #Find the total number of missing values per column
    city_df.isnull().sum()
```

RegionType	0
RegionName	0
City	0
State	0
Metro	545
CBSATitle	545
SizeRank	79
MarketHealthIndex	0
SellForGain	3812
PrevForeclosed	9541
ForeclosureRatio	8671
ZHVI	79
MoM	79
YoY	79
ForecastYoYPctChange	2408
NegativeEquity	1353
Delinquency	1353
DaysOnMarket	106

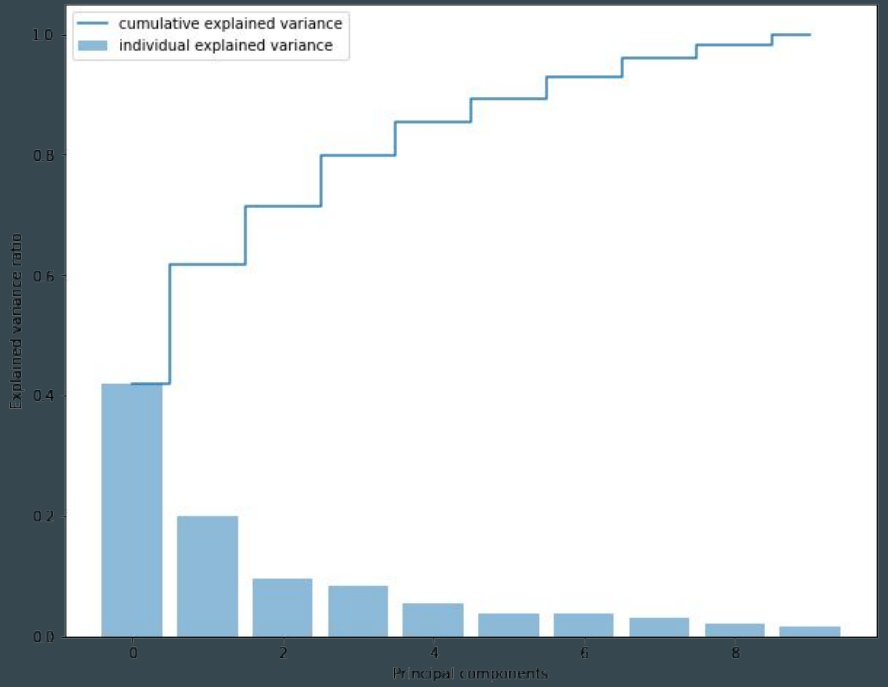
Data Wrangling (Cont'd)

- Decided it was best to get rid of nine columns (string columns so clustering computations can be executed + market health index to prevent clustering with MHI influence)
- Dropped all rows that contained a missing value, rather than impute rows with "educated guess".
- This process left me with 951 rows and 10 columns to work with

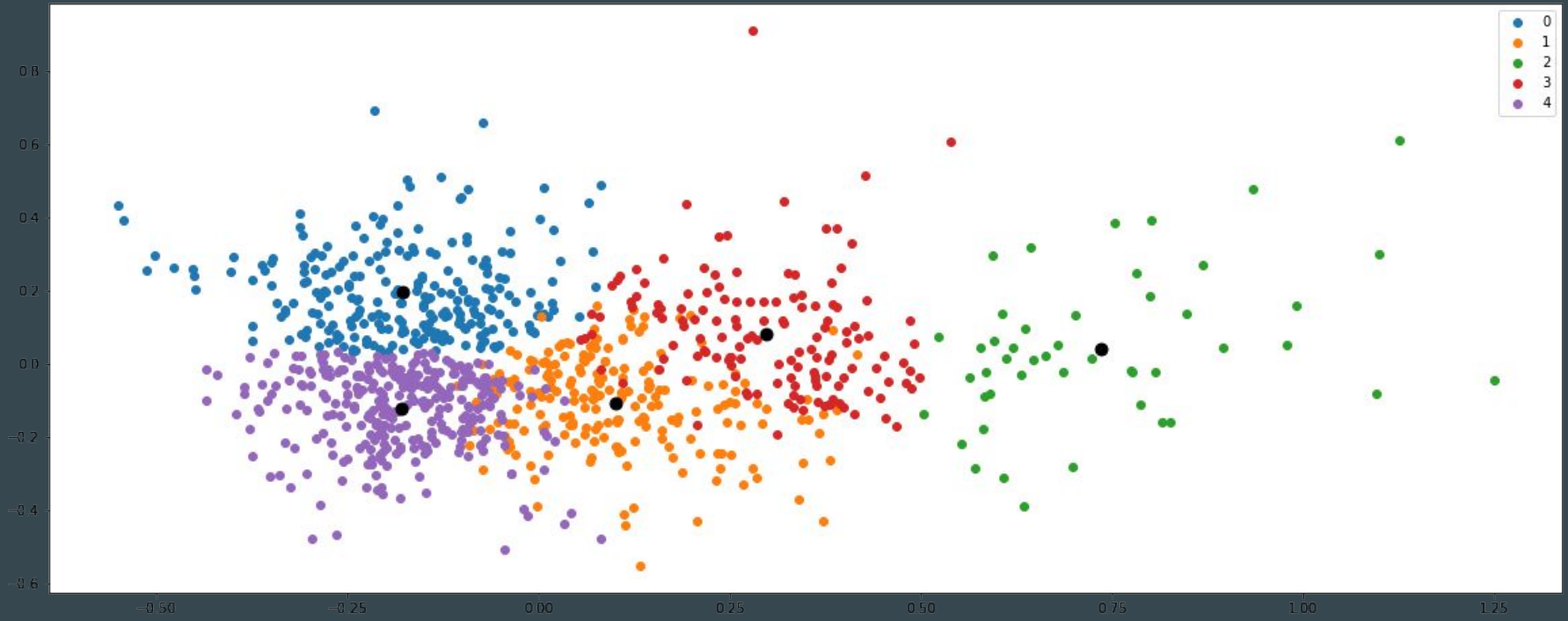
```
city_df = city_df[['RegionType' ,'RegionName', 'State', 'MarketHealthIndex', 'SellForGain', 'PrevForeclosed', 'ForeclosureRatio',
                   'ZHVI', 'MoM', 'YoY', 'ForecastYoYPctChange', 'NegativeEquity', 'Delinquency', 'DaysOnMarket']]
city_df = city_df.dropna()
city_df

drop_columns = ['RegionType', 'RegionName', 'State', 'MarketHealthIndex']
city_train = city_df.drop(columns=drop_columns)
city_train
```

Principal Components

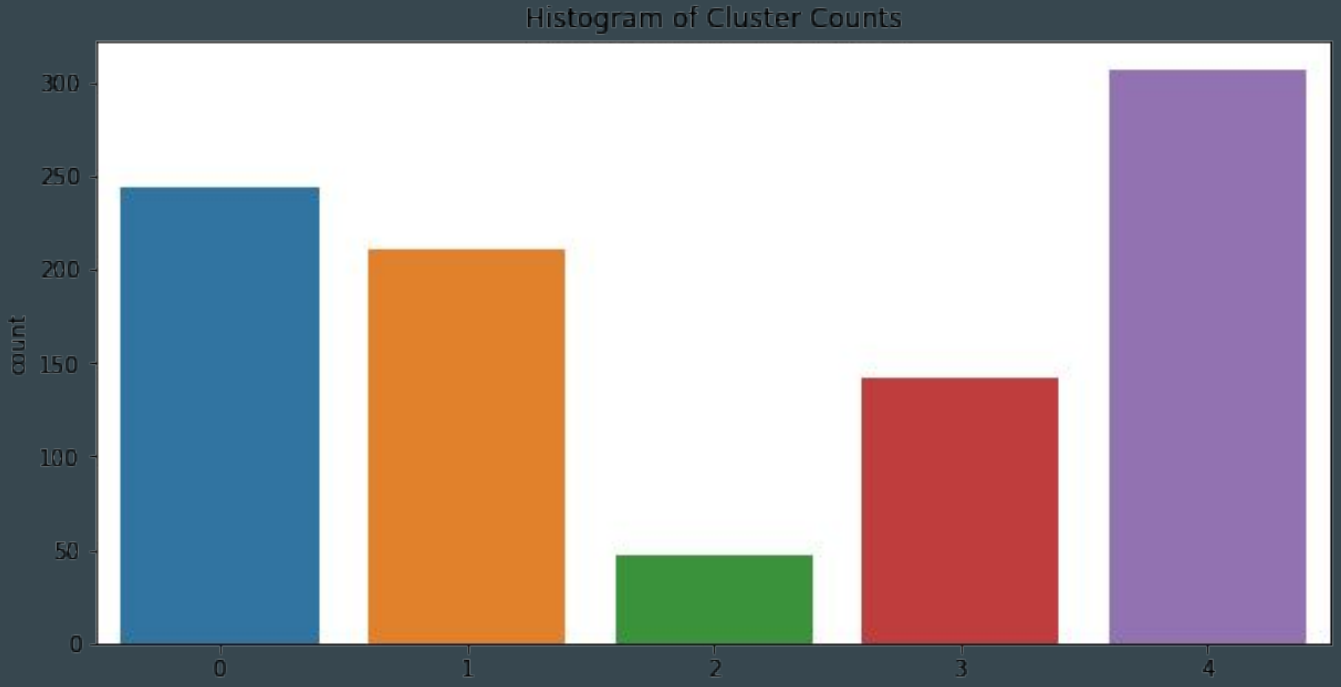
- Principal Component Analysis (PCA) is a technique used for reducing dimensionality of datasets, increasing interpretability but minimizing the loss of information at the same time
- Rule of Thumb: select number of PCs that add up to ~80% variance (3 PCs)



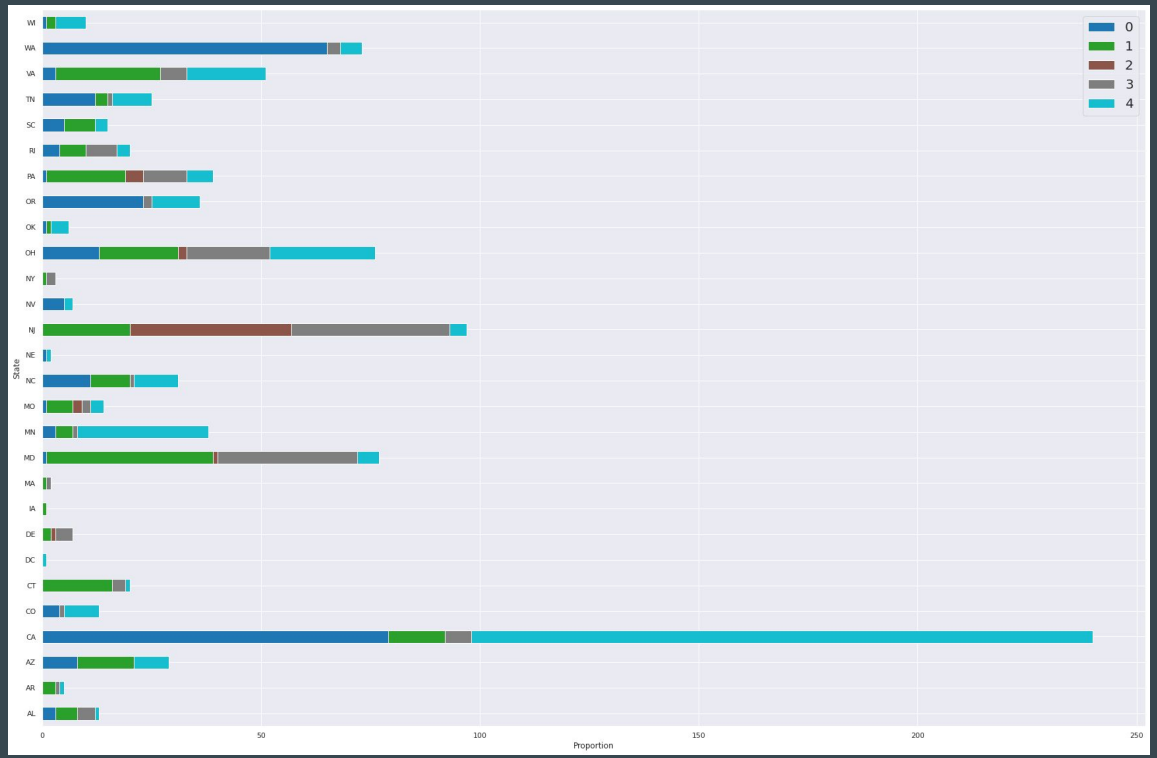
Clusters (Five) + Cluster Centroids



Prevalence of Clusters



Proportion of Clusters v. State



Average Measures Across All Five Clusters

	cluster	SellForGain	PrevForeclosed	ForeclosureRatio	ZHVI	MoM	YoY	ForecastYoYPctChange	NegativeEquity	Delinquency	DaysOnMarket
0	0	93.00	5.85	2.77	286744.26	0.71	9.94	0.05	0.09	0.03	60.61
1	1	76.92	4.63	2.14	268467.30	0.33	3.29	0.03	0.11	0.04	86.02
2	2	64.84	13.12	8.62	170631.91	0.29	2.81	0.02	0.21	0.14	125.01
3	3	80.28	10.69	5.12	208357.75	0.53	5.27	0.03	0.15	0.09	93.35
4	4	93.29	3.45	1.32	455320.52	0.34	5.31	0.02	0.06	0.02	64.11

Visualization of MHI Distribution v. Cluster

Average MHI Amongst All Clusters

```
[26] print('cluster 0 avg mhi:', cluster_0['MarketHealthIndex'].mean())
     print('cluster 1 avg mhi:', cluster_1['MarketHealthIndex'].mean())
     print('cluster 2 avg mhi:', cluster_2['MarketHealthIndex'].mean())
     print('cluster 3 avg mhi:', cluster_3['MarketHealthIndex'].mean())
     print('cluster 4 avg mhi:', cluster_4['MarketHealthIndex'].mean())

     cluster 0 avg mhi: 7.1446865677893845
     cluster 1 avg mhi: 4.013506114254426
     cluster 2 avg mhi: 0.7653594187477912
     cluster 3 avg mhi: 2.106410134469897
     cluster 4 avg mhi: 7.03735554111553
```

Cluster 0- Healthiest & Most Desirable

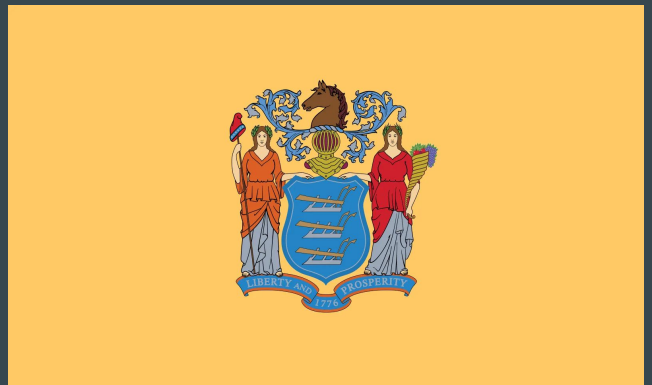
- **Highest** average MHI (7.1446865677893845), highest MoM (0.71) and YoY (9.94), highest Forecast YoY Change (0.05), least amount of average # of days on the market (60.61)
- 2nd highest Sell For Gain average, 2nd highest ZHVI
- 244 total cities, 20 unique states
- Predominantly features most cities from Pacific region (WA, OR, and CA)
    - CA (79) and WA (65) with the most cities, including my hometown! (Woodinville, WA)

Cluster 1- The Median Group

- 3rd highest average MHI (4.013506114254426), 3rd lowest average # of days on the market (86.02), 3rd highest ZHVI
- 2nd lowest SellForGain average, 2nd lowest Foreclosure Ratio
- 211 total cities, 22 unique states
- Predominantly features most cities from East Coast (MD, VA, NJ, PA)
  - MD (38) and VA (24) with the most cities

Cluster 2- Missed Payments/Least Desirable

- **Lowest** average MHI (0.76535941874779 12), **Lowest** ZHVI (17,0631.91)
- **Highest** average # of days on the market (125.01), **Highest** Delinquency average **Highest** Foreclosure Ratio (8.62)
- 47 total cities, 6 unique states
- Predominantly features most cities from NJ (~79% of all cities)
    - NJ (37) and PA (4) with the most cities
    - 76% (73/96) of cities ranked in the 90th percentile for Delinquency are located in NJ



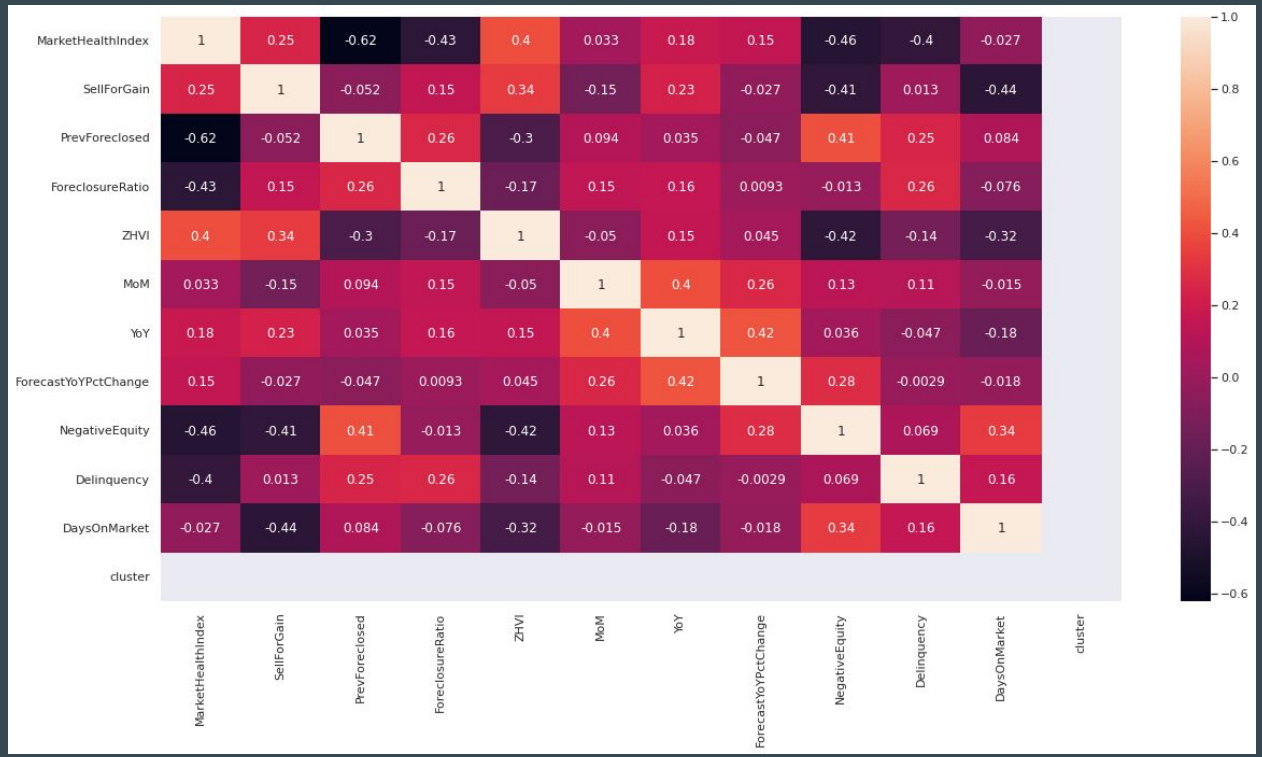
Cluster 3- Essentially, Cluster 2 With Less Bias

- 2nd lowest MHI (2.106410134469897), 2nd lowest ZHVI
- 2nd highest average # of days on the market (93.35), 2nd highest Foreclosure Ratio (5.12)
- 142 total cities, 20 unique states
- Predominantly features most cities from East Coast (NJ, MD, RI, PA)
    - NJ (36) and MD (32) with the most cities
    - 32 of the 36 cities in NJ are found in the 90th percentile amongst all cities in delinquency
        - 8/32 cities in this cluster from MD come from the same percentile group
        - While only one other state (MD) coming from the same percentile group was placed in cluster two, nine other states (MD, DE, CT, NY, PA, CO, MA, RI, and WA) are included in cluster three

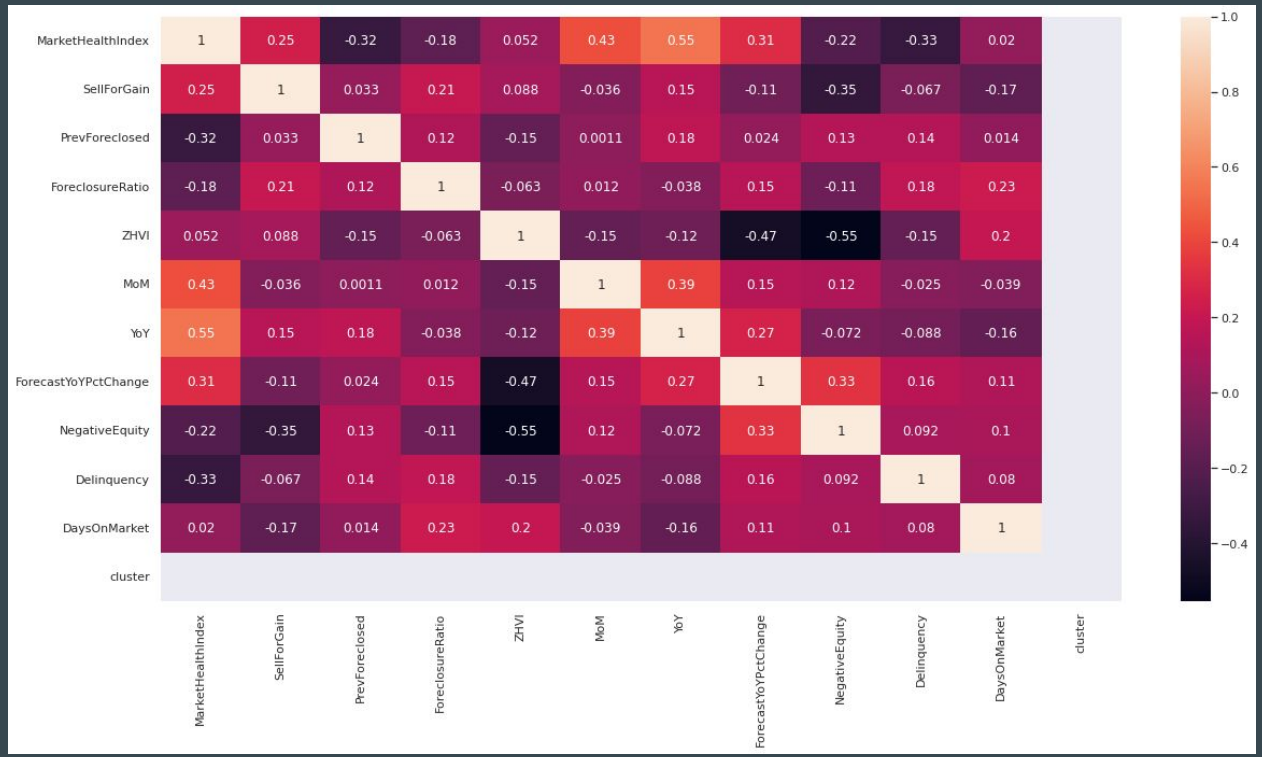
Cluster 4- Cluster 1 With a Little More Sun (California Cluster)

- **Highest** ZHVI, 2nd highest average MHI (7.037355554111553)
- **Lowest** Foreclosure Ratio, **Lowest** Negative Equity, **Lowest** Delinquency, 2nd lowest average # of days on the market (64.11)
- 307 total cities, 24 unique states
- Predominantly features most cities from California
    - CA (142) and MN (30) with the most cities

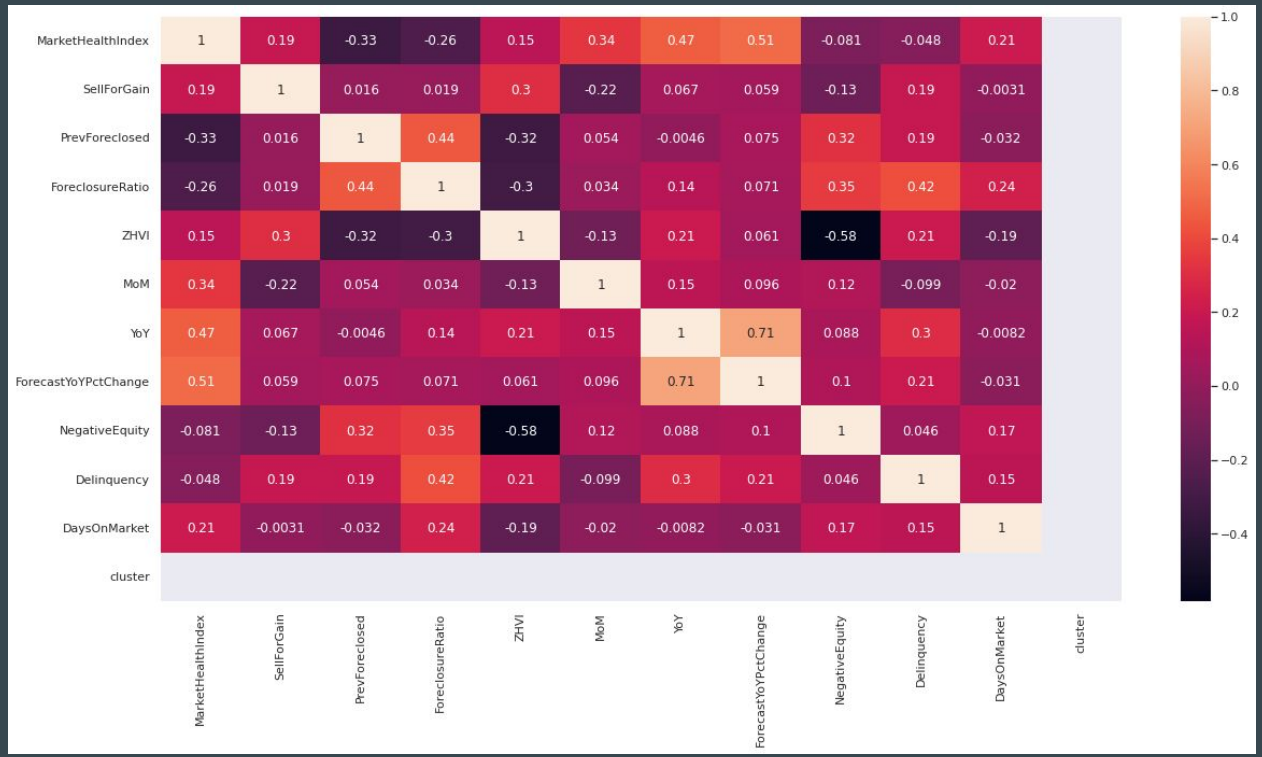
Correlations(?)- Heat Map of Cluster 0



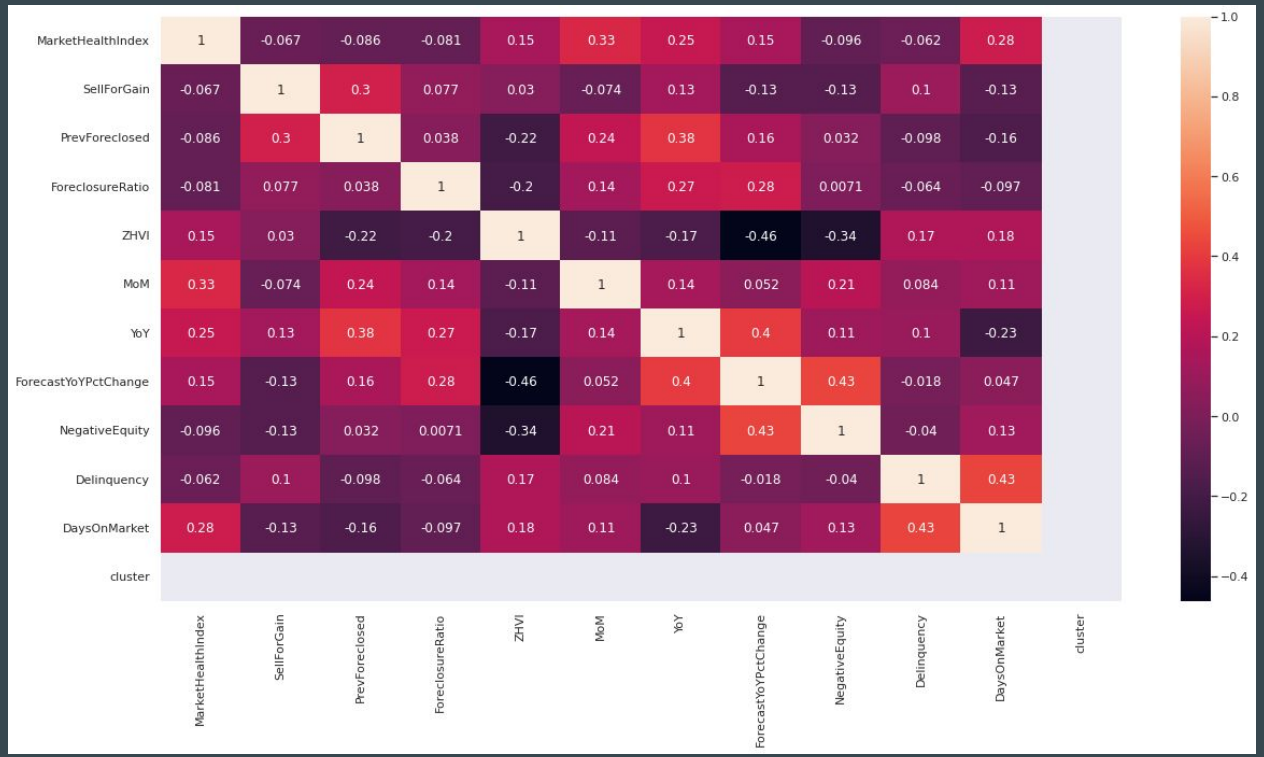
Correlations(?)- Cluster 1



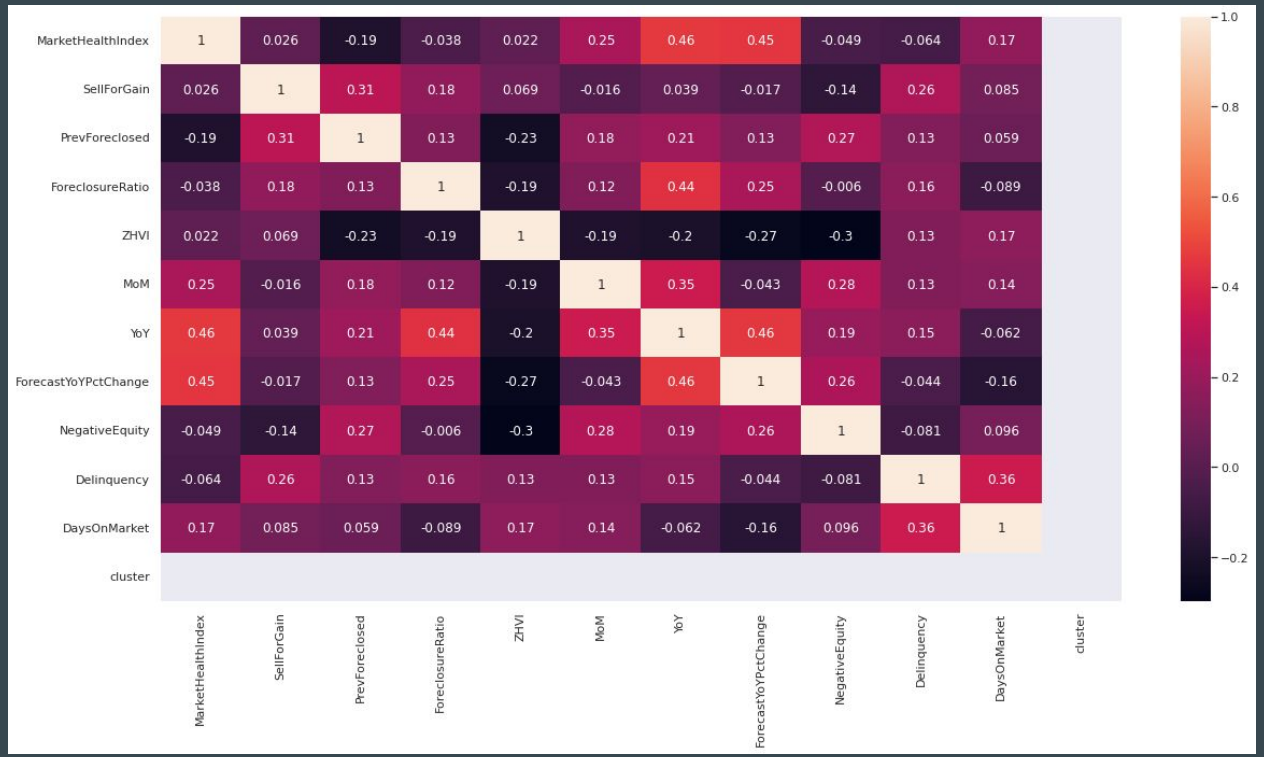
Correlations(?)- Heat Map of Cluster 2



Correlations(?)- Heat Map of Cluster 3



Correlations(?)- Heat Map of Cluster 4



Final Insights

- *Why/How are each city within each given cluster related to one another?*
  - Based on MHVI, Delinquency/Foreclosure Ratio, and State
- *Which cluster of cities feature the highest average for market health index (MHI)?*
  - Cluster 1 (CA, WA, and OR)
  - May feature the highest MHI based on the west coast featuring: strong economic growth, job creation/abundance in career opportunities, weather, climate/culture, various levels of living cost
- *Which factors determine a higher MHI, if any?*
  - While there were several correlation coefficients (a statistical measure of the strength of the relationship between the relative movements of two variables) whose magnitude ranged from 0.3-0.5 (low correlation), the four heat maps did not reveal a single correlation that was worth mentioning and diving deeper into (any correlation worth acknowledging should at least be >= 80%).
  - For this exact reason, we cannot make any concluding statements on what factors may have contributed to a higher MHI.

Reflection

- **Caveat**: Because there was a lot of missing data featured in this dataset, I find my reports to be somewhat biased in the story-telling, as the data wrangling procedure left me with roughly 8.6% of the original dataset.
- If I had more time/redo this project again, I think I would have either performed another clustering on top of the clusters that were discovered or just increased the number of clusters from my original code instead. The reason for this is because clusters 0 and 4, as well as clusters 2 and 3 were too similar in characteristics.
- Furthermore, I enjoyed working on this project and the learning lessons that came with the whole process. I hope that this was enough to showcase my skills/capabilities of a potential candidate for the FTF team and hope to hear good news post-evaluation!
- Click here to access the full code to this analysis project on my GitHub!

Citation/Resources

- The dataset used for this project was obtained from data.world
- The software used to perform this code is through the Google Colab platform (Python language)