Relocating to Massachusetts - Analysis of the Neighborhoods

...

Niru Udayakumar

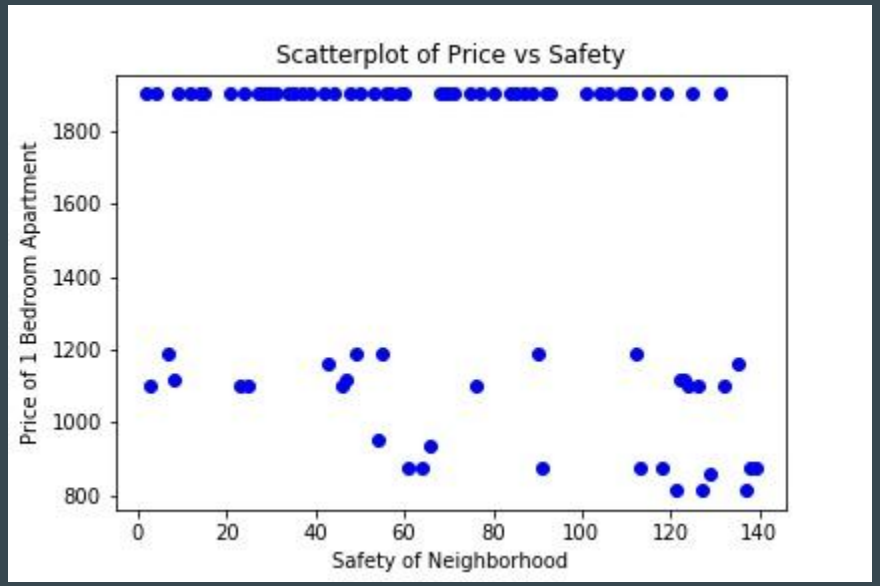
Problem Statement

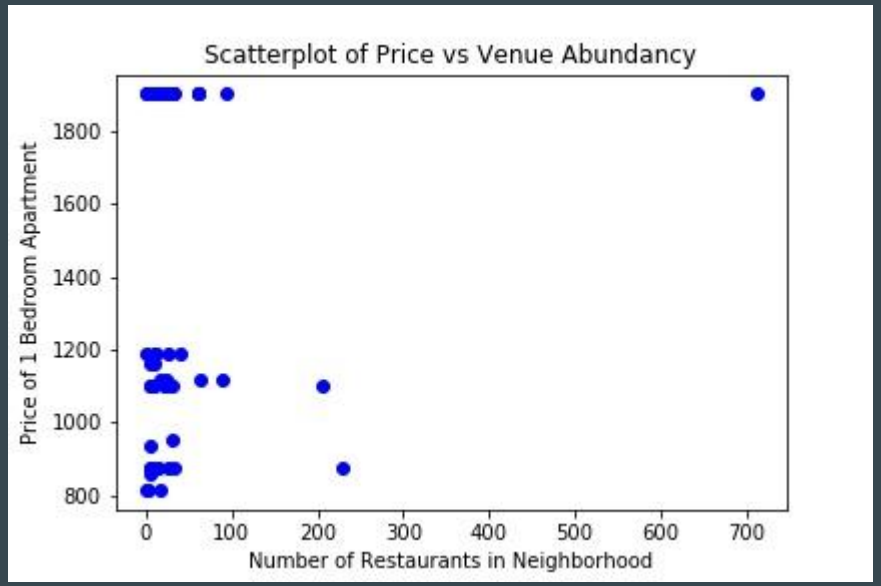
- Relocation is a very stressful issue but happens very frequently for work
  - 2.8-4.5M people relocate for work every year in the US since 1999
  - 62% of workers say they would relocate for work
- Factors to consider:
  - Housing costs
  - Safety of neighborhood
  - Nearby venues
- This will be collect, clean, analyze, and cluster neighborhoods in Massachusetts based on similarities between them from the factors listed above

Data Sources

- Four main data sources used:
- Neighborhood location data
  - URL: https://www.zipcodestogo.com/Massachusetts/
  - All the zip codes in the state and their associated neighborhoods
- House pricing data
  - URL: https://www.rentdata.org/states/massachusetts/2020
  - Pricing data for a 1 bedroom apartment in each neighborhood
- Safety data
  - URL: https://www.safewise.com/blog/safest-cities-massachusetts/
  - Safety ranking and rates of violent and property crimes in each neighborhood
- Foursquare venue data
  - Focuses on restaurant and gym venues near the neighborhood, based off personal preference

Exploratory Correlation Analysis





- Looking for correlation between housing costs and safety (left) or number of venues (right)
- No clear correlation seen - further emphasizes need for machine learning model

Input Data for Model

- K means clustering used to cluster the data
- Following table (snippet) shows the inputs used to cluster the neigborhoods

	Zip Code	Neighborhood	1 BR Apartment ($)	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
0	1001	Agawam	875	113	3.6	14.8	12	1
1	1002	Amherst	875	64	3	4.3	9	6
2	1003	Amherst	875	64	3	4.3	9	4
3	1004	Amherst	875	64	3	4.3	9	7
4	1040	Holyoke	875	139	9.7	40.5	14	2
5	1060	Northampton	875	118	4	18.9	2	2
6	1061	Northampton	875	118	4	18.9	15	3
7	1063	Northampton	875	118	4	18.9	15	4

Results of k means Clustering Model

- Table snippet shows cluster number for each neighborhood/zip code

	Zip Code	Neighborhood	Cluster Number	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
0	1001	Agawam	1	875	113	3.6	14.8	12	1
1	1002	Amherst	1	875	64	3	4.3	9	6
2	1003	Amherst	1	875	64	3	4.3	9	4
3	1004	Amherst	1	875	64	3	4.3	9	7
4	1040	Holyoke	1	875	139	9.7	40.5	14	2
5	1060	Northampton	1	875	118	4	18.9	2	2
6	1061	Northampton	1	875	118	4	18.9	15	3
7	1063	Northampton	1	875	118	4	18.9	15	4

Analysis of Clusters

- Mean of each input for the individual clusters, as well as the mean for the full data set

	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
mean, cluster 0	1900.000	94.985	3.788	14.206	20.646	15.415
mean, cluster 1	868.667	119.250	6.887	21.054	14.333	3.667
mean, cluster 2	1140.100	85.100	3.655	11.540	14.750	3.650
mean, cluster 3	1117.000	8.000	0.200	4.800	30.000	12.000
mean, cluster 4	1900.000	73.643	2.482	10.139	11.786	5.893
mean, cluster 5	1065.381	104.238	4.914	16.195	15.190	6.667
mean, full table	1533.585	94.321	4.135	14.392	16.730	9.308

Analysis of Clusters - Comparison to Full Data Set Averages

- Gradient indicating the percent variation from the mean of each input compared to the mean values of the full data set

	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
mean, cluster 0	23.89%	0.70%	-8.40%	-1.29%	23.41%	65.61%
mean, cluster 1	-43.36%	26.43%	66.56%	46.29%	-14.32%	-60.61%
mean, cluster 2	-25.66%	-9.78%	-11.61%	-19.82%	-11.83%	-60.79%
mean, cluster 3	-27.16%	-91.52%	-95.16%	-66.65%	79.32%	28.92%
mean, cluster 4	23.89%	-21.92%	-39.98%	-29.55%	-29.55%	-36.69%
mean, cluster 5	-30.53%	10.51%	18.84%	12.53%	-9.20%	-28.38%

Summary of Clusters - Pros and Cons

Cluster Number	Pros	Cons
0	- Average safety - Plenty of venues	- More expensive than average
1	- Cheapest neighborhood	- Poor safety rating - Slightly lower than average # of restaurants - Very few gyms
2	- Cheaper than average - Slightly safer than average	- Slightly lower than average # of restaurants - Very few gyms
3	- Cheaper than average - Extremely safe - Plenty of venues	- Only 1 zip code falls within this cluster
4	- No radical characteristics (none are far higher or lower than the average)	- More expensive than average - Slightly lower than average safety ranking - Fewer than average # of venues
5	- Cheaper than average - Slightly safer than average	- High property crime rate - Slightly lower than average # of restaurants - Fewer than average # of gyms

Conclusions

- K means clustering was an effective means of analyzing the neighborhoods as it is unsupervised and allows for clustering
- Future work should change the number of clusters from 6 to 5 to eliminate cluster 3
    - Cluster 3 only has one neighborhood encompassed
- Future work should also include performing this same analysis for other US states to which people often relocate for work, such as New York and California