Battle Of Neighborhoods IBM Capstone Project

Jack Perng
07/2020

Introduction

- **Problem**: Where would be a good location to setup an acupuncture clinic in San Jose?
- **Target Audience**: Licensed California acupuncturists living in San Jose, trying to determine an ideal neighborhood to start their own business

- **Goal**: Explore San Jose neighborhoods (ZIP codes) based on:
  - Population
  - Per capita income
  - Existing acupuncture clinics
  - Crime rate
  - Unemployment rate (age 25+)
  - Bachelor degree percentage (age 25+)
  - Median home price

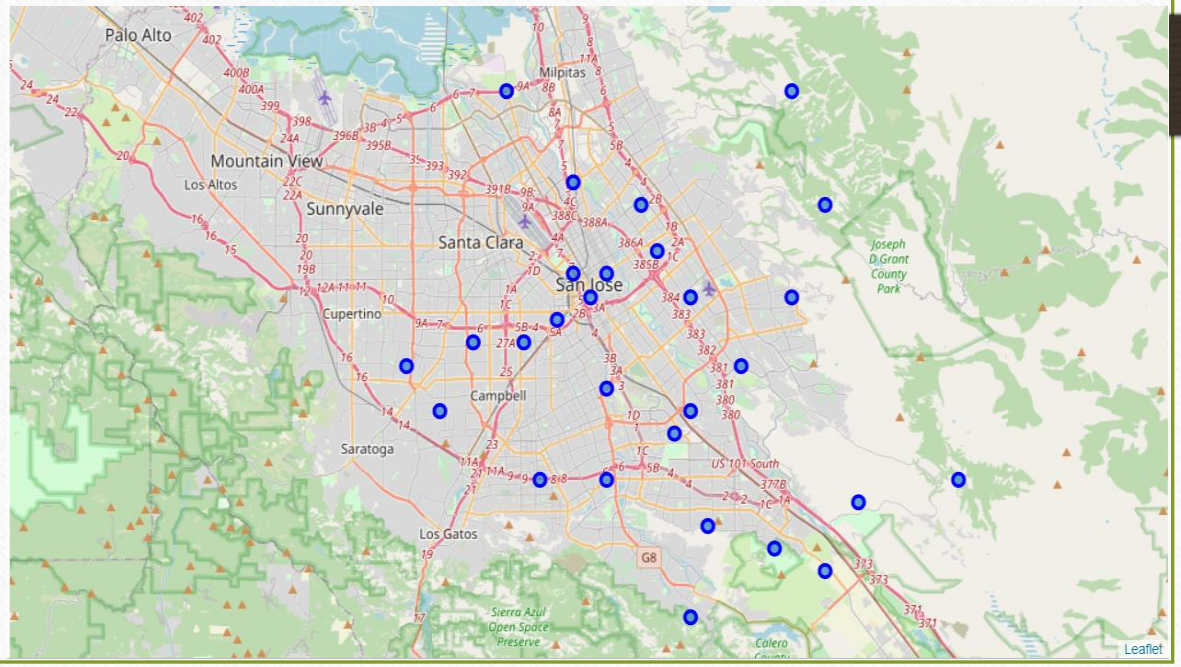
Data

- **Kaggle dataset**
  - U.S. ZIP codes
  - Geo-coordinates (latitude & longitude)
  - Estimated Population
  - Total Wages
- **Foursquare location data**
  - Acupuncture clinics in San Jose
  - Note: only returns maximum 50 results
- **ADT Security Services**
  - Total Crime Rate
- **City-Data**
  - Unemployment Rate (age 25+)
  - Bachelor Degree or Higher Percentage (age 25+)
- **Zillow**
  - Median Home Price

Methodology

- Data Preparation
  - Clean up Kaggle dataset
  - Remove non-San Jose, CA ZIP codes
  - Remove P.O. Box ZIP codes
  - Remove ZIP codes with missing data
  - **28** ZIP codes after final processing

- Folium Map of Neighborhood



Methodology

- Data Aggregation
  - Combine data from all sources

	Zipcode	Lat	Long	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
0	95110	37.34	-121.90	12621.0	29036.412963	2	91	4.9	32.4	820
1	95111	37.28	-121.83	43578.0	19872.887374	0	32	6.1	19.6	774
2	95112	37.34	-121.88	34111.0	26143.931606	1	85	6.2	35.0	849
3	95113	37.33	-121.89	1049.0	36152.631077	0	86	4.5	72.8	763
4	95116	37.35	-121.85	35357.0	17645.394519	0	37	6.9	16.6	712

Methodology

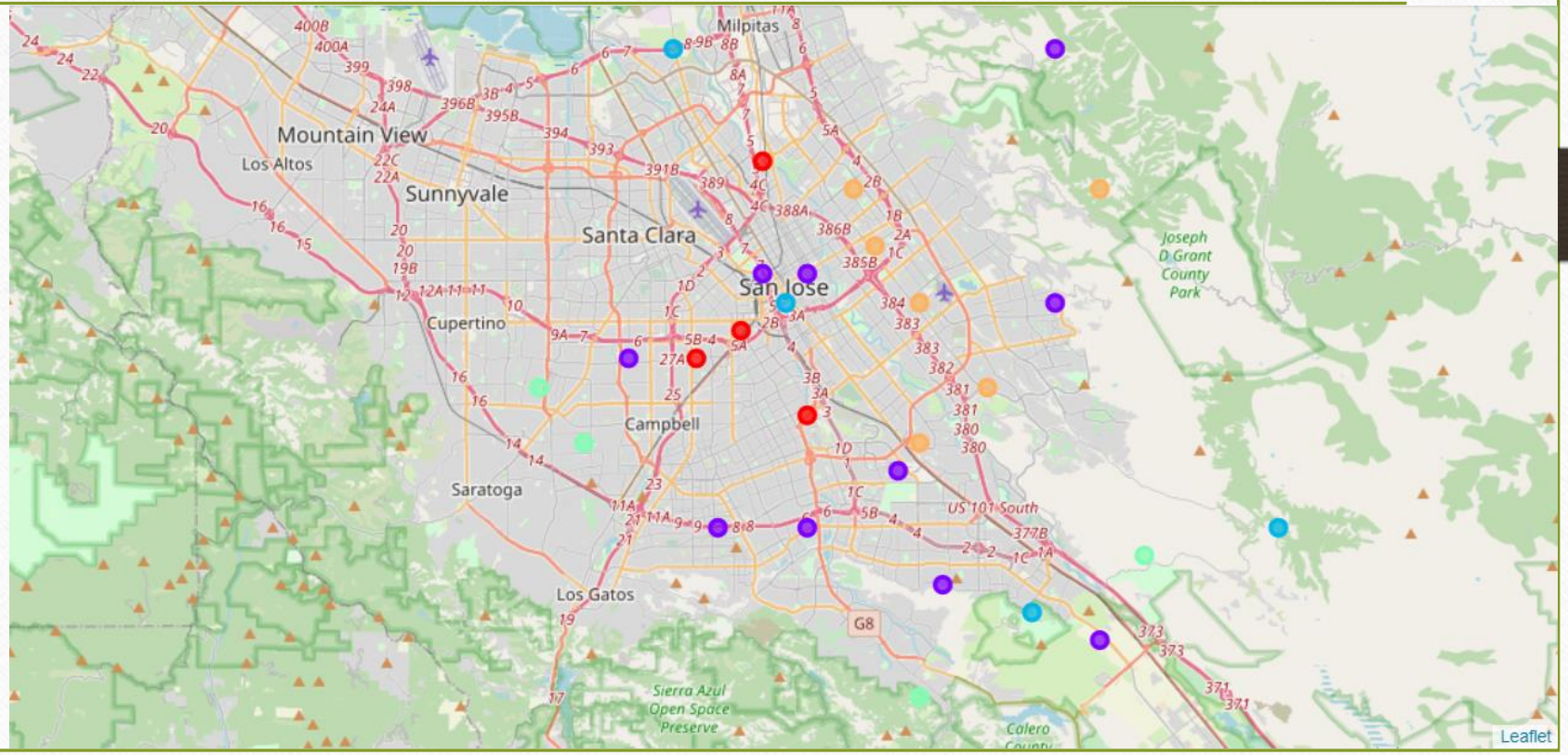
- Data Transformation
  - Data columns have different scales
  - Normalized using **Standard Scalar**
- Modeling
  - **K-means Clustering** algorithm for segmentation and clustering
  - Number of clusters k = 5

	Zipcode	Lat	Long	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
0	95110	37.34	-121.90	-1.123828	-0.517780	0.418548	0.980054	-0.084071	-0.829971	-0.915354
1	95111	37.28	-121.83	1.265132	-1.470805	-0.558064	-1.231762	1.002381	-1.610097	-1.098995
2	95112	37.34	-121.88	0.534561	-0.818604	-0.069758	0.755124	1.092919	-0.671508	-0.799580
3	95113	37.33	-121.89	-2.016842	0.222321	-0.558064	0.792612	-0.446221	1.632299	-1.142910
4	95116	37.35	-121.85	0.630715	-1.702469	-0.558064	-1.044320	1.726682	-1.792939	-1.346512

normalized columns

Results

Clustered Neighborhood Map



Cluster 0

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
13	95125	41048.0	43425.502022	5	92	4.8	54.0	1331
14	95126	23076.0	36464.311362	4	93	4.8	52.0	1025
16	95128	25327.0	33020.557231	9	80	3.6	43.8	1159
19	95131	24403.0	37677.994919	6	55	4.4	55.9	1077

- Affluent cluster: high per capita wages and home prices, low unemployment rate
- Crime rate rather high
- **Stiff competition**: large number of clinics

Cluster 1

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
0	95110	12621.0	29036.412963	2	91	4.9	32.4	820
2	95112	34111.0	26143.931606	1	85	6.2	35.0	849
5	95117	22030.0	30240.189696	1	21	5.5	45.5	1275
6	95118	26249.0	33438.090632	0	65	4.3	46.6	1128
11	95123	50481.0	32382.124086	1	67	4.0	41.1	946
12	95124	39234.0	38621.629938	0	69	4.4	53.0	1275
20	95132	34344.0	30705.994118	0	65	5.9	45.7	1151
24	95136	35078.0	33568.082074	0	68	4.7	46.7	961
26	95139	5634.0	36503.131523	0	40	6.2	47.7	904
27	95148	37541.0	30403.984470	0	61	5.7	39.0	1026

- Contains most neighborhoods
- All column attributes are average
- Sparse number of clinics

Cluster 2

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
3	95113	1049.0	36152.631077	0	86	4.5	72.8	763
7	95119	8171.0	35339.883368	0	128	4.2	41.5	953
22	95134	12670.0	51631.955722	0	92	3.0	75.8	951
23	95135	17221.0	42079.977063	0	121	4.7	61.3	1169

- Affluent neighborhoods: high per capita wages, low unemployment rate, reasonably well-educated
- Lower home prices
- Crime rate high
- Nonexistent competition

Cluster 3

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
8	95120	33486.0	50890.606821	1	52	3.1	71.3	1508
17	95129	32839.0	41750.644569	0	53	3.7	72.9	1751
18	95130	10841.0	36396.430311	0	64	3.7	55.7	1419
25	95138	15421.0	57789.241554	0	42	5.3	59.2	1213

- Wealthiest neighborhood: high per capita wages and home prices
- Well-educated and low unemployment rates
- Low crime rates
- Few competition

Cluster 4

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
1	95111	43578.0	19872.887374	0	32	6.1	19.6	774
4	95116	35357.0	17645.394519	0	37	6.9	16.6	712
9	95121	30427.0	25115.081638	1	59	6.0	29.6	852
10	95122	41936.0	16719.609953	2	29	6.4	15.1	726
15	95127	46641.0	22820.174160	0	45	5.5	23.4	792
21	95133	20337.0	26582.617544	0	24	7.3	35.3	870

- Poorest neighborhood: Low per capita wages, low home prices, high unemployment rate, low bachelor degree percentages
- Low crime rate (somewhat counterintuitive result)
- Weak competition

Discussions

- Clusters to avoid
  - Cluster 0: very stiff competition
  - Cluster 4: not wealthy neighborhood
- Cluster 1 considered “safe bet”, with overall average attributes
- Cluster 3 top choice
  - Wealthy, highly-educated, weak competition, low crime rates
- Cluster 2 second choice
  - Wealthy, highly-educated, weak competition, **higher crime rates**

Conclusions and Future Directions

- Based on collected demographic data, San Jose neighborhoods were clustered to determine ones most favorable for opening an acupuncture clinic.
- Ideas for improvement
  - Include commercial rent prices
  - Obtain more complete and accurate acupuncture clinic data
  - Focus on important crime rate data (violent as opposed to non-violent crimes)