New Gym in New York

Luca Benini

Agenda

- Business Problem
- Data Science Solution
- Data Gathering
- Data Analysis
- Results
- Conclusion
- Future Development

Business Problem

- The fitness industry has become hyper-competitive
- 8 out of 10 Gym will fail in their first year.
- Choosing the right location can be fundamental
- The ideal location has good business opportunity, but also
- …the ideal location has little or no competitor

Data Science Solution

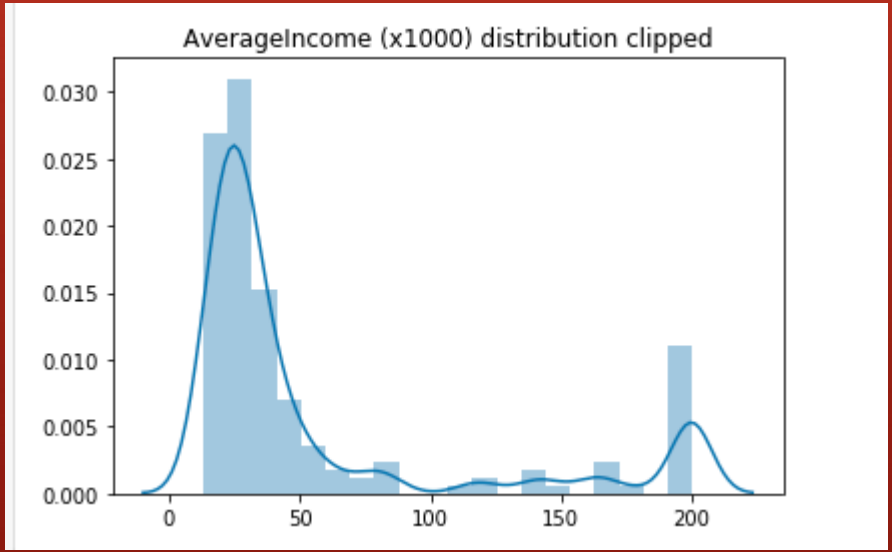
- Some areas are better location for a gym
- Similar area have similar business opportunities
- Some of these areas will be already exploited by competitor
- Identify areas similar to successful gym hotspot but with no gym
- Cluster and Select

Data Gathering

- Census Data (Population and Zip Codes)
- IRS Data (Income)
- Foursquare Data (Gym business distribution)
- Foursquare Data (Venues Frequencies)
- Geographical Data (Shape, Locations, and Areas)

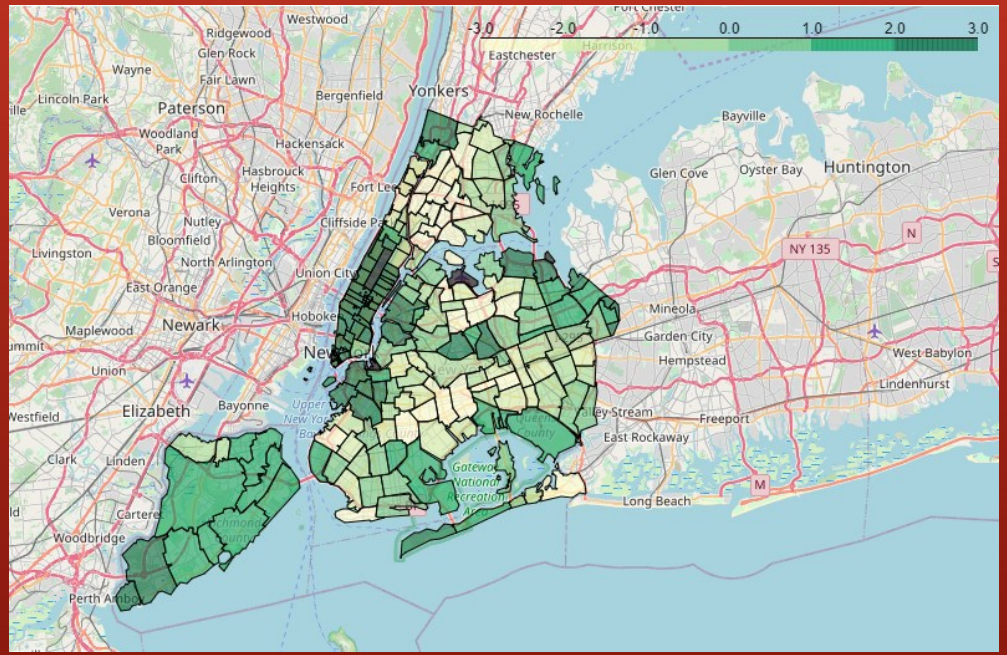
Analysis - Income

- Income is not uniformly distributes
- We clip the data >200.000$/year to 200.000$/year



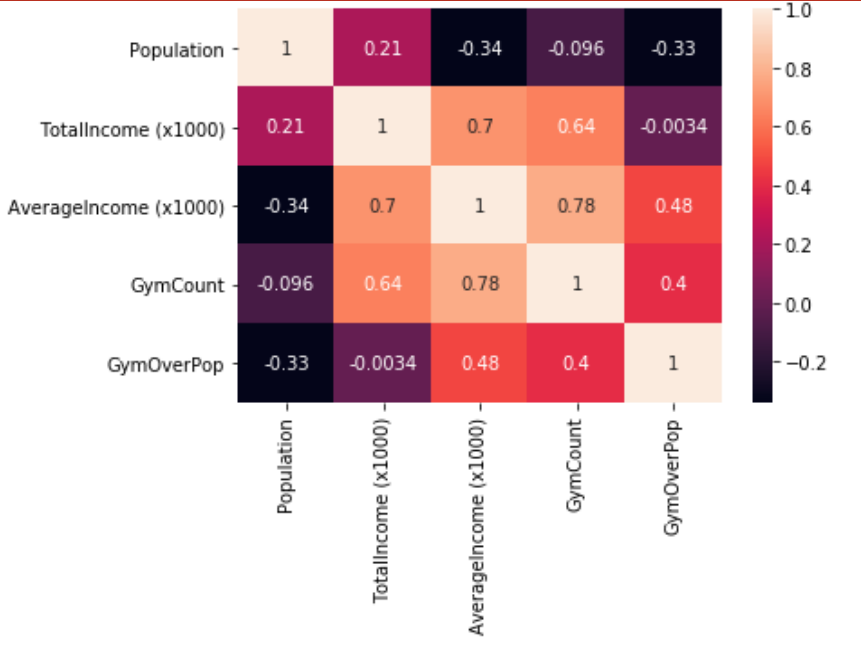
Analysis - IncomeIndex

- We cut by quartile to introduce a more generic index



Analysis - Correlation

- Zip Code with higher income have higher Gym Count



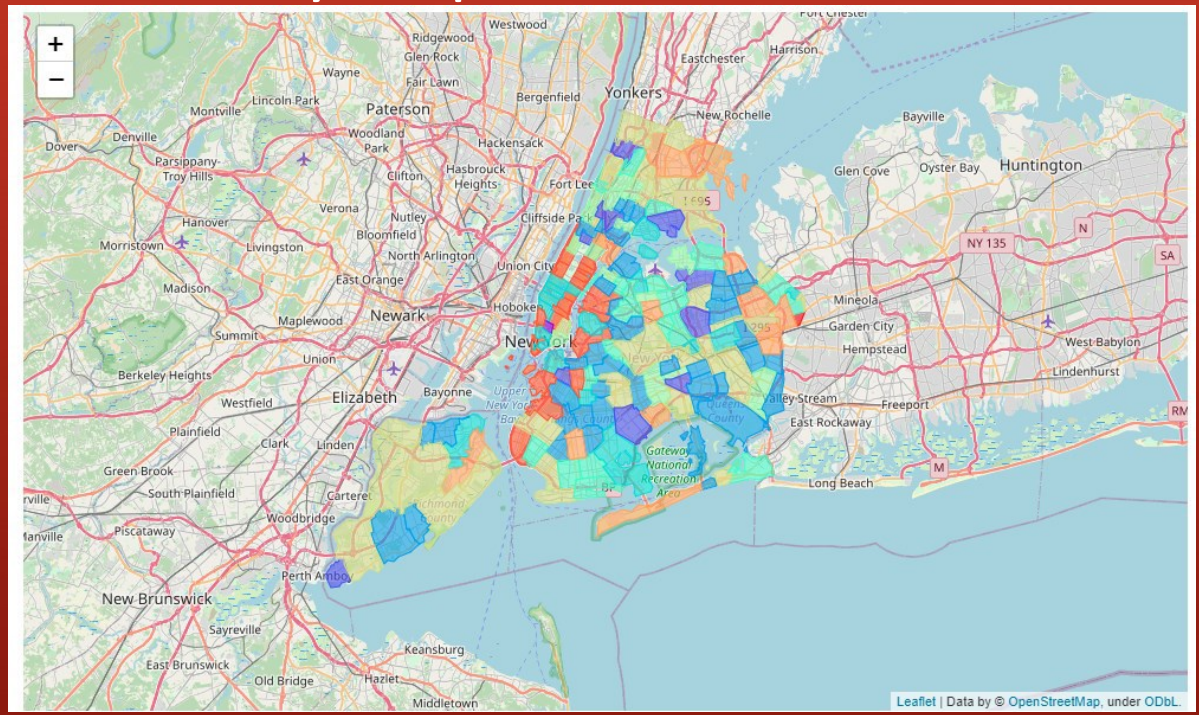
Analysis - Clustering

- We cluster the zip codes using venues frequency and income (all normalized)

- We have selected k=10 clusters, some will be degenerates (outlier zip code)

Analysis - Clustering

- Zip Code similarity map



Results

- A spefic cluster(8) has a very high gym median gym count…
- … but the min is 1!
- Our objective!

Cluster	count	mean	std	min	25%	50%	75%	max
0	1.0	30.000000	NaN	30.0	30.00	30.0	30.00	30.0
1	11.0	8.363636	9.330303	0.0	2.00	5.0	9.50	30.0
2	38.0	5.763158	6.478381	0.0	1.25	3.0	8.75	30.0
3	21.0	18.428571	13.643942	0.0	2.00	29.0	30.00	30.0
4	20.0	8.550000	5.942488	1.0	4.00	8.0	11.00	27.0
5	15.0	10.800000	8.945869	0.0	4.00	9.0	14.00	29.0
6	36.0	3.694444	5.338911	0.0	0.75	2.0	5.00	29.0
7	16.0	2.875000	2.305790	0.0	1.00	2.0	5.25	6.0
8	24.0	26.125000	8.131060	1.0	27.00	29.5	30.00	30.0
9	1.0	4.000000	NaN	4.0	4.00	4.0	4.00	4.0

Conclusion

- 11231 and 10021 are similar to high gym density zip codes…
- …but have a very low gym rate
- Higher the income better the business
- 10021 is our result!

ZipCode	Population	IncomeIndex	GymCount	GymOverPop
11232	27723	-2	4	0.004252
10021	102078	2	27	0.007794
11209	69840	1	24	0.010126
11211	85089	0	30	0.010389
11231	32974	2	13	0.011617
10024	61414	3	26	0.012475
10023	62206	3	29	0.013737
10128	59856	3	30	0.014769
11238	48965	2	27	0.016248
10003	53673	3	30	0.016470

Future Development

- Include budget consideration (not all Zip Code have the same rents)
- Explore the negative correlation between population and gym count