Effect of venues on average salaries of residents in the area

Denis Rakhmanov

October 18, 2019

Table of contents

1. Introduction 1
1.1. Business problem 1
2. Data 2
2.1. Loading the geographical data 2
2.2. Collecting Foursquare Data 2
3. Methodology 2
3.1. Correlation between amount of venues and income 2
3.2. Clusterization....................................6
3.3. Venues effect on average salary 8
4. Results 8
5. Discussion....................................8
6. Conclusion 9

1. Introduction

No matter where we live we are surrounded by different venues. They take their part in shaping us one way or another. People who live in one area might have different sets of values, maybe the availability of different services and place around them has its toll.

However, I'll try to look into to this connection the other way around: is there a correlation between per capita income of families who live in that area and surrounding them venues. Is poor regions have enough services available to them? Are venues in rich neighborhoods greatly different from those in less rich.

1.1. Business problem

The main purpose of this research is to look into sociological correlation between types of venues and people living near them.

The results of this research might be helpful for those who looks into sociologic values of neighborhoods and tries to look into their economy.

2. Data

Because I will need the data about average income per neighborhood I decided to go with Chicago. Earlier in this source we had a link to Chicago website with all the necessary information about the city, including:

- average income and hardship index per community, available in .csv file
- geojson file for creating the map and getting the coordinates of communities using the shapely Python library
- other information you can find interesting.

Another source of data will be Foursquare API.

Using the search endpoint we will look for most common places instead of recommended ones that are returned by explore endpoint.

Also we will be searching by several categories to split our request into multiple ones to not overflow our results with only one most common type in this area and to not hit the limit that is returned by the API.

2.1. Loading the geographical data

Since geojson file has the information of shapes of communities but not the actual center of community we will load the polygonal information into shapely python module to get the required coordinates of each community.

2.2. Collecting Foursquare Data

After having required coordinates we used FoursquareAPI to gather information about venues in each area.

We made API calls with “category” parameter. This was done by two reasons: to not include certain types of venues that are seems irrelevant and not to hit the limit returned by the API, leaving us with only partial information about location.

After that we had to do some cleaning, combining several similar categories of venues into one to decrease the list of features in our dataset. For example, we combined Sport fields and Stadiums into one category. After that our list of features shortened from over 300 to 107

3. Methodology

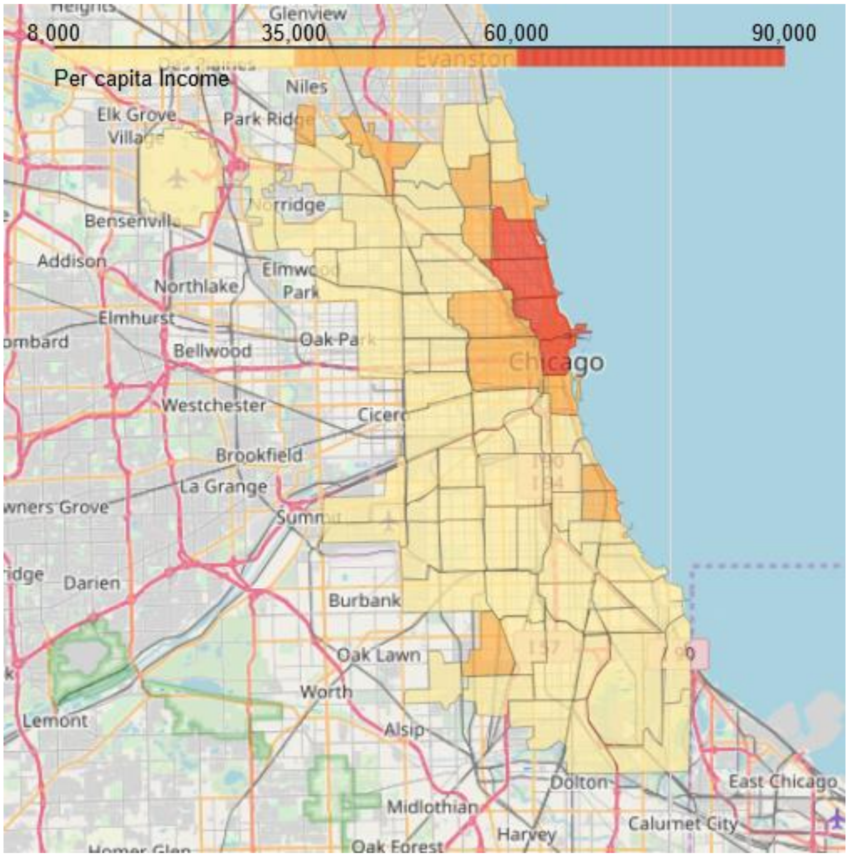
First we will look into possibility or correlation between the average income in communities and the amount(density) of venues near its centers and calculate basic coefficients using regression model.

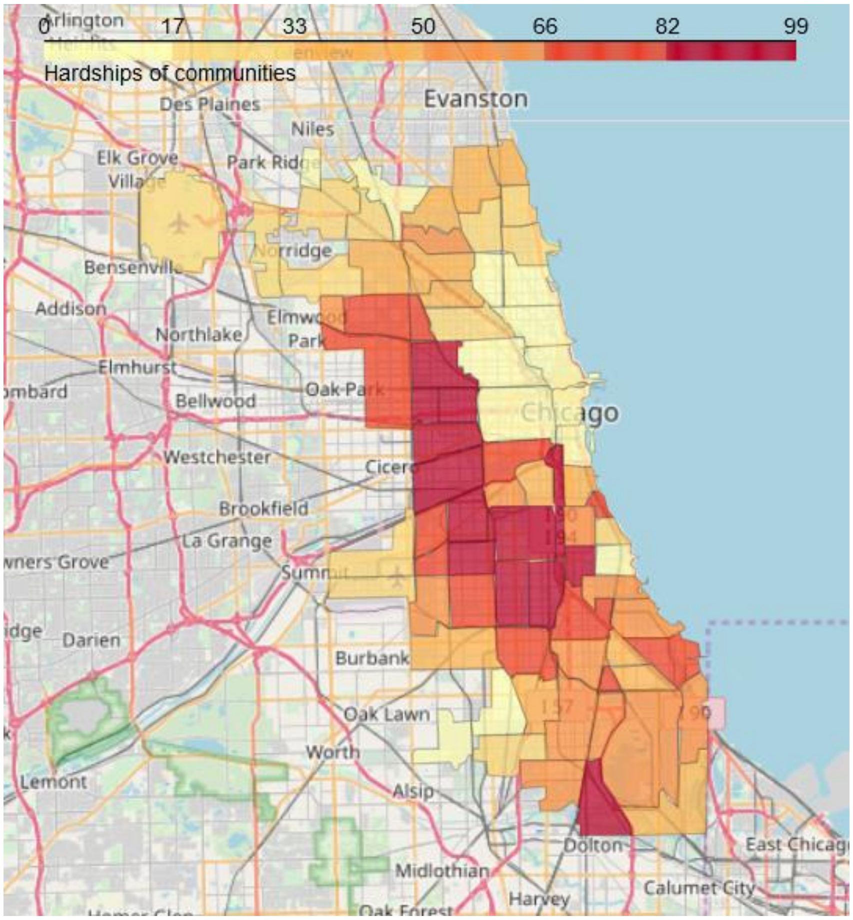
Then we will try to use clusterization and see how clusters correspond with income and hardship of community.

Then we will try to see which venues might affect income of families in neighborhoods more.

3.1. Correlation between amount of venues and income

Looking at geographical data we can see that areas with highest average income located near the city Harbor, while areas with hard living conditions located right beside them with "middle areas" on the outskirts of the city.

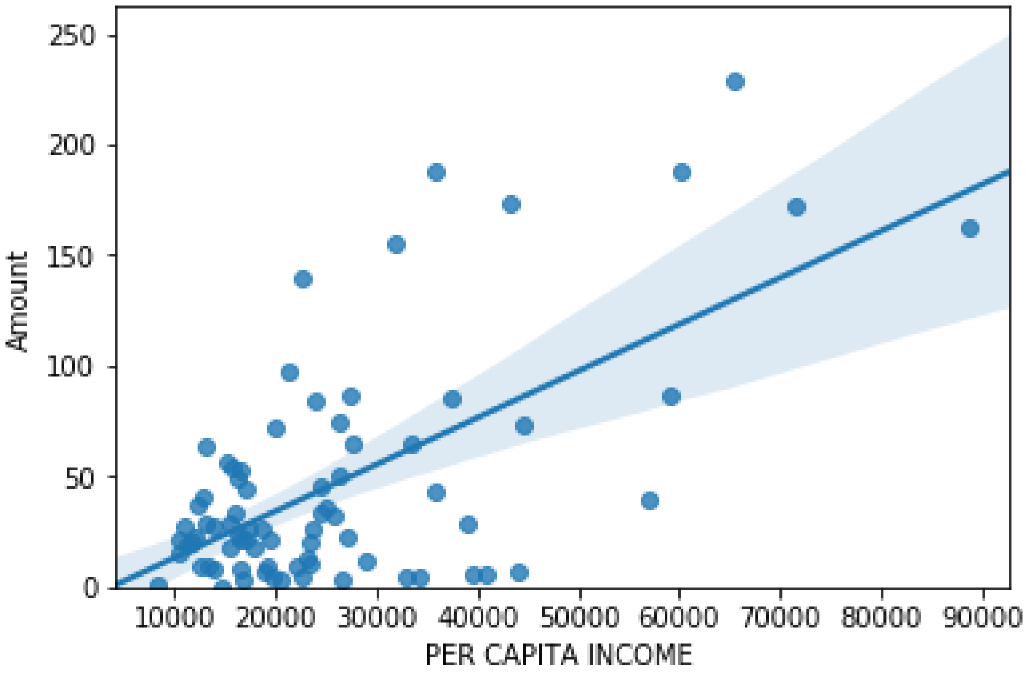

8,000
35,000
60,000
90,000
Per capita Income
Glenview
Evanston
Niles
Elk Grove Village
Park Ridge
Norridge
Bensenville
Addison
Northlake
Elmwood Park
Elmhurst
Lombard
Bellwood
Oak Park
Chicago
Westchester
Cicero
Brookfield
La Grange
Downers Grove
Summit
Darien
Burbank
Oak Lawn
Worth
Lemont
Alsip
Dolton
East Chicago
Midlothian
Calumet City
Harvey
Oak Forest
Homer Glen


0
17
33
50
66
82
99
Hardships of communities
Arlington Heights
Glenview
Des Plaines
Evanston
Niles
Elk Grove Village
Park Ridge
Norridge
Bensenville
Addison
Northlake
Elmwood Park
Elmhurst
Lombard
Bellwood
Oak Park
Chicago
Westchester
Cicero
Brookfield
La Grange
Downers Grove
Summit
Darien
Burbank
Oak Lawn
Worth
Lemont
Alsip
Dolton
East Chicago
Midlothian
Calumet City
Harvey
Oak Forest
Homer Glen

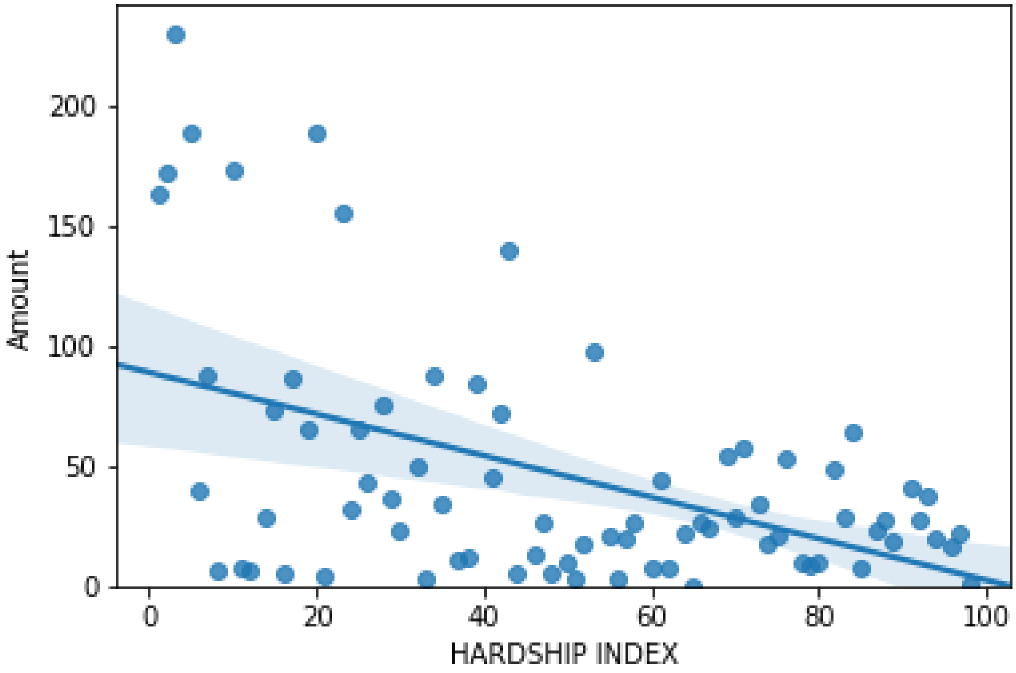
Lets look into correlation between Amount of venues and Average income. We also include hardship index, available in dataset from Chicago website – an index based not only on income but other features like Percent of households below poverty, percent of adults of high school diploma and others.

	PER CAPITA INCOME	HARDSHIP INDEX	Amount
PER CAPITA INCOME	1.000000	-0.849167	0.629972
HARDSHIP INDEX	-0.849167	1.000000	-0.483058
Amount	0.629972	-0.483058	1.000000

We see, that the correlation between income and amount is positive and big enough to suspect some relations. Hardship index isn't good enough for such conclusion.



Using linear regression model, we discovered that R2-Score of this distribution is about 0.4. Which is not great.



Then we counted the most common type of venues in each community and looked into what types were most common in low-income communities and high-income ones.

COMMUNITY AREA NAME	PER CAPITA INCOME	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
SOUTH LAWNDALE	10402	Daily Conveniences	Mexican Restaurant	IT	Sports	Nightlife
FULLER PARK	10432	Religious Place	Daily Conveniences	Fast Food	Food	Parks and Rec
WEST GARFIELD PARK	10934	Food	Religious Place	Daily Conveniences	Beauty	Tea and Desserts
WEST ENGLEWOOD	11317	Religious Place	Daily Conveniences	Professional Places	Post Office	Education
ENGLEWOOD	11888	Healthcare	Beauty	Fast Food	Business Service	Gym

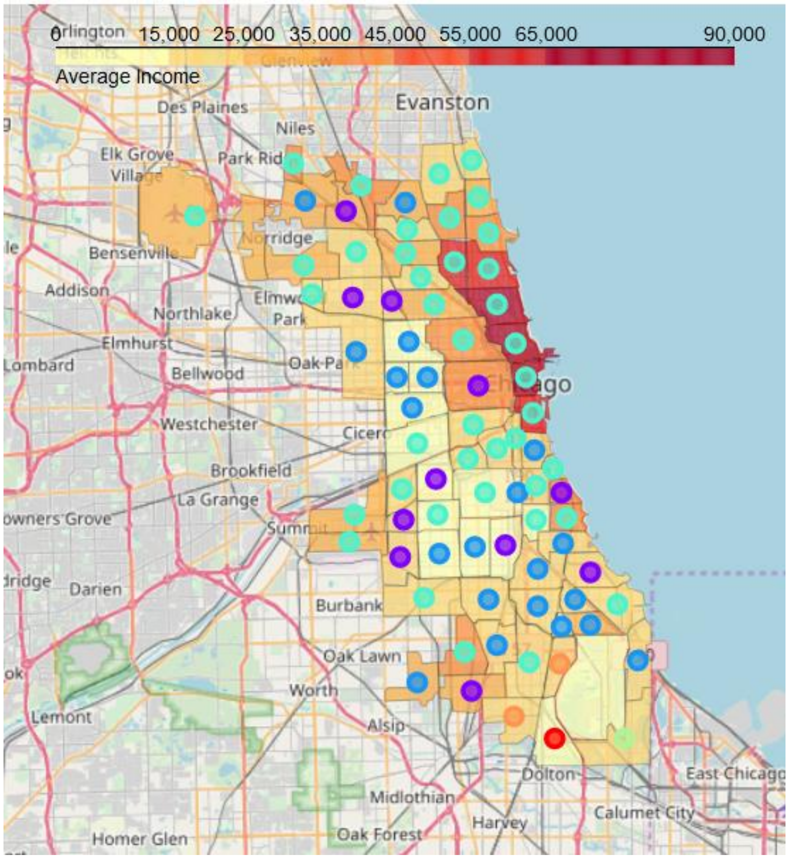
COMMUNITY AREA NAME	PER CAPITA INCOME	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
NEAR NORTH SIDE	88669	Professional Places	Healthcare	Landmark	selfimprovement	Arts Venue
LINCOLN PARK	71551	Bar	Daily Conveniences	Landmark	Professional Places	Fast Food
LOOP	65526	College and Univercity	Bar	Professional Places	Landmark	Fast Food
LAKE VIEW	60058	Bar	Professional Places	Landmark	Daily Conveniences	Entertainment
NEAR SOUTH SIDE	59077	Landmark	Gym	Professional Places	Healthcare	Parks and Rec

We can see from this result:

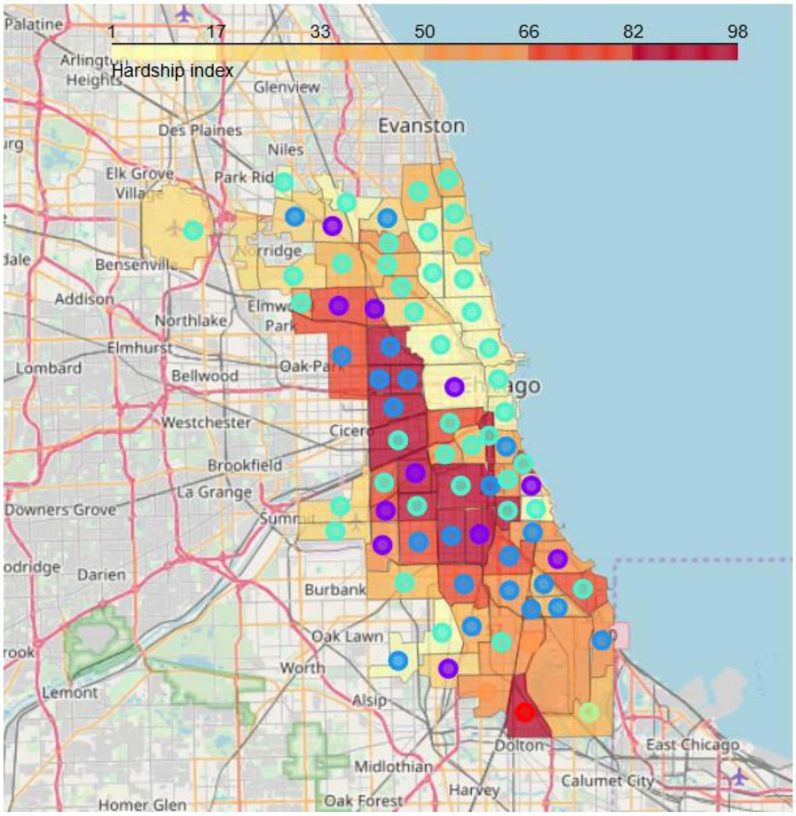
- in poor areas there are less venues than in more prosperous. and while their numbers are small more often than not we can see Daily necessities like Laundries or dry cleaners, food (and fast food) venues and places of religion
- in Top 5 wealthy areas Landmarks and Professional places (places where people work like offices, business centers and business practices) are in top 5
- In wealthy areas Bars and Gyms are more popular than in poorer areas

3.2. Clusterization

So, we see, that there might be a correlation between types of venues and the average income of residents. Lets look, how clusters built on venues data will look like if we combine them with geographical data



And again, but with hardship index



3.3. Venues effect on average salary

What if we built a regression model to see if venues around might affect salaries of residents (or, probably, attract people with different incomes)? What kind of venues would make the most impact?

Based on our dataset the venues with the most impact are: Summer Camp, Bus Line, Hotel, Men's Store, Harbor / Marina, Waste Facility, Fire Station, Cemetery, Stables, Mediterranean Restaurant. Venues with least impact are: Pawn Shop, Nature Preserve, Fair, Courthouse, Entertainment, Entertainment and education, Bookstore, Business Service, Library, Money Services. And venues with most negative impact are: Eastern European Restaurant, Entertainment Service, Toys and Games, Print and press, Casino, Latin American Restaurant, Post Office, Police Station, IT, Video Store

However, R2 Score is only 0.33, which makes it hard to trust these results. Especially when we got Waste Facility and Cemetery among the venues that attract high-income people to live near them.

4. Results

By looking at a "Amount of venues/Average Income" plot we see, that there's definitely a positive regression. However, low R2 Score can mean that the correlation may be present to only certain categories of venues while other types just add noise.

By looking at most common types of venues we saw that low-income neighborhood have more of venues aimed towards daily needs while high income communities have more recreational venues, like landmarks or Bars. However, clusterisation model didn't prove that types of venues correspond with income of residents in that area.

From regression model "Average Income based on venues" we have strange results. Among venues with the most effect were 'Cemetery' and 'Waste Facility'. However, the R2 score is too low to get any certainty from this result.

5. Discussion

Although we get some results suggesting that there might be a possibility of a correlation between average salary in community and types of venues in that area, the error scores aren't great.

Errors might come from the fact that we gather only venues within 300m radius of the center of community area, but we have no idea how incomes are distributed within a community itself. The areas that make the most impact on average salary of community might not be around the center.

Also, 300m is not enough, but we had reached a limit that can be returned by Foursquare API using it. To overcome these shortcomings, we need to increase radius of our search while splitting categories even further to have more results returned by our API calls.

Another area that made an impact of our results is the size of our dataset. The amount of features is greater than the amount of areas we considering. It would be better to have smaller areas within communities to take into consideration, as this will also take care of distribution of income within communities

6. Conclusion

We have shown that there's a positive correlation between average Income in the neighborhood and the amount of venues. Even though the R2 Score is pretty low we also saw that the most common venues in wealthy and poor neighborhoods are different.

We can look into how different categories correlate with income, maybe we got low R2 Score because we also included some categories that just add noise to our picture.

Our dataset consists of 77 communities of Chicago, and while we decreased the amount of categories of venue in our set it is still greater that the set itself which creates bigger errors in our calculations.

To be more certain we need bigger set: either to split communities of Chicago into smaller areas, but this would require access to income data not available in public access, or to include other cities into the set, but this can lead to bigger errors due to possible cultural differences between cities.