Obesity and Venues in Chicago

Table of contents

1. Introduction
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

1. Introduction

There is a public concern regarding with overweight and obesity all around the world. Nowadays, there is a huge percentage of americans suffering with obesity and children are also in danger. Factors such as bad eating habits, sedentarism, and the fast living pace are increasing the junk food consume and decreasing physical activity.

In this project we will try to analyze similarities between community areas in the city of Chicago, according with their percentage of population suffering from obesity, in terms of the most popular venues in each area. The main objective is to verify whether the venues are affecting people's weight. This report will be targeted to the Government (Specially Health Area), as well as to the owners of this venues so that they can also take corrective actions. This information is also very useful for Centers for Disease Control and Prevention (CDC), which is an entity in charge of protect community health. They are developing a project called *Communities Putting Prevention to Work (CPPW),* which is an initiative designed to “make healthy living easier by promoting environmental changes at the local level”[1].

For the CDC, it would be interesting to have an analysis that allows them to take decisions regarding venues that are affecting positively or negatively the population's weight. If it is possible to verify whether there exists a relationship between the venues and the amount of obese people, some action plans can be taken to improve food quality in determined zones and create incentives for people so that they can make exercise actively. It would be cheaper to prevent this behavior from the source, rather than spending money in gastric bypassing, surgeries and heart diseases treatments needed because of obesity.

---

[1] https://www.cdc.gov/nccdphp/dch/programs/communitiesputtingpreventiontowork/communities/profiles/both-il_chicago.htm

The main idea with this project would be to compare venues in the community areas and by using clustering verify if there is a pattern in data that can be related with high/low obesity levels in each area, so that the Government can take any action e.g. giving rewards for restaurants to control portions and food quality, subsidize gym subscriptions for people, and so on.

2. Data

Based on our problem, we have to consider the following parameters:

- Percentage of Obesity in each Community Area
- Which are the most popular venues per area regarding food and exercise
- How similar are the community areas according to their most popular venues
- How similar are the results from community areas' venues to the percentage of obesity per community area.

We will need to use data that describes the percentage of obesity per Community Area, that is going to be retrieved from the Chicago Health Atlas. Data should come with the name of the community area, and the obesity percentage, e.g. Community Area Name: Englewood, Obesity Percentage: 30%. Then, we need to retrieve the most popular venues per Community Area, as well as it's category to see if the type of food, or place for exercising can affect the obesity rates. For obtaining this information we will use Foursquare API. In this case, we will only consider two types of venues:

- Food
- Outdoors and Recreation

The idea is to retrieve the 100 most popular venues with their name, category, latitude and longitude e.g. Name: Hai Yen, Category: Vietnamese Restaurant Latitude: 41.97328297968702, longitude: -87.65705585461119

Having this information we can compare the different categories and use the K-means clustering algorithm to verify how similar are the community areas and if it exists a relation between the category of food and Outdoors and Recreation venues, and the obesity percentage per community area. We will also use geographic data in GeoJson format to display the geographic limits of community areas, and use a palette to display colors according to their obesity percentages. This data can be found in the Chicago Data Portal. For obtaining longitude and latitude per Community Area we will use Nominatim API.

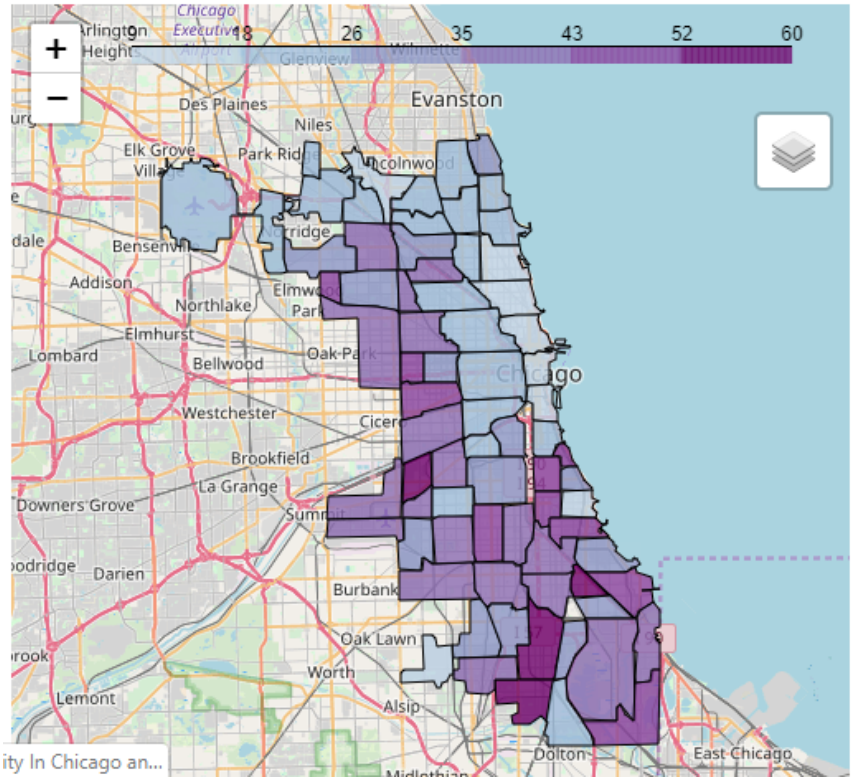
2.1 Obesity Data

First thing we have done is to obtain data related with obesity. This data can be found in https://www.chicagohealthatlas.org/indicators/adult-obesity. We are adding this data, which is in

excel spreadsheet format, to this repository. This file is called 'Adult_Obesity.xlsx'. There is a lot of information that will not be necessary to use, for example old information that comes from 2001. Our interest is in the last report, so the period of 2015-2017 per community area was taken. Also we retrieved from this data only the weight percentage, as well as the information of the Community Area, as we are not evaluating any specific ethnicity, or age.

In order to visualize this data in a Choropleth map, We have obtained boundaries per community area in geojson format from the Chicago Data portal in https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6. We will also add this data to the repository, with some modifications to allow us to work with it.

We can see the resulting map in the following graph:



Then, we have added another column to the dataframe with information of latitude and longitude for each community area. This data was obtained by using Nominatim.

This was the final structure of the dataset:

	Geo_Group	Geo_ID	Weight_Percent	lat	lon
0	ROGERS PARK	1	33.7	42.010531	-87.670748
1	WEST RIDGE	2	24.0	42.003548	-87.696243
2	UPTOWN	3	20.9	41.966630	-87.655546
3	LINCOLN SQUARE	4	20.3	41.890910	-87.625348
4	NORTH CENTER	5	18.1	41.956107	-87.679160
5	LAKEVIEW	6	16.8	41.943919	-87.654076
6	LINCOLN PARK	7	15.8	41.921699	-87.647832
7	NEAR NORTH SIDE	8	13.8	41.900033	-87.634497
8	EDISON PARK	9	19.7	42.005733	-87.814004
9	NORWOOD PARK	10	20.9	41.985590	-87.800577

2.2 Foursquare

We have examined two different types of venues considered to be the ones with more influence related with obesity. These venues are Restaurants and Exercising places. First we have obtained the 100 most popular restaurants per Community Area, by using category_Id='4d4b7105d754a06374d81259'. This category id shows all types of restaurants.

	CA	Name	Category	Lat	Lon
0	ROGERS PARK	El Famous Burrito	Mexican Restaurant	42.0104	-87.6742
1	ROGERS PARK	Taqueria & Restaurant Cd. Hidalgo	Mexican Restaurant	42.0116	-87.6745
2	ROGERS PARK	Smack Dab Bakery	Bakery	42.0045	-87.6731
3	ROGERS PARK	J.B. Alberto's Pizza	Pizza Place	42.0079	-87.6651
4	ROGERS PARK	Smack Dab	Bakery	42.0093	-87.6662

After having obtained the 100 most popular restaurants per community area, we did the same with physical activity, by retrieving the 100 most popular venues in the category Outdoors and Recreation, with id= '4d4b7105d754a06377d81259'.

	CA	Name	Category	Lat	Lon
0	ROGERS PARK	Loyola Park	Park	42.0114	-87.6624
1	ROGERS PARK	Loyola Beach	Beach	42.009	-87.6595
2	ROGERS PARK	Leone Beach	Park	42.0131	-87.6636
3	ROGERS PARK	Pottawattomie Park	Park	42.0151	-87.6769
4	ROGERS PARK	Howard Beach Park	Playground	42.0189	-87.6641

3. Methodology

In the first step we have obtained the 100 most popular venues in **restaurants** and **outdoors and recreation** categories by using the foursquare API, as well as **Obesity Percentage** from Chicago Health Atlas portal, to see if this type of venues are affecting obesity index in the community areas of Chicago.

With all the data collected, we have analyze the different sub-categories of venues and obtain the most popular ones per Community Area. Initially, we made a separate analysis with restaurants and another with sports by creating clusters to see if there is an independent relation between each of them and the Obesity values. We will display this information in the choropleth map, so that we can identify visually how this values can be related.

Finally, we have merged Restaurants and Outdoors and Recreation data, to see the influence of both categories in the obesity index by creating clusters as we did in the previous step.

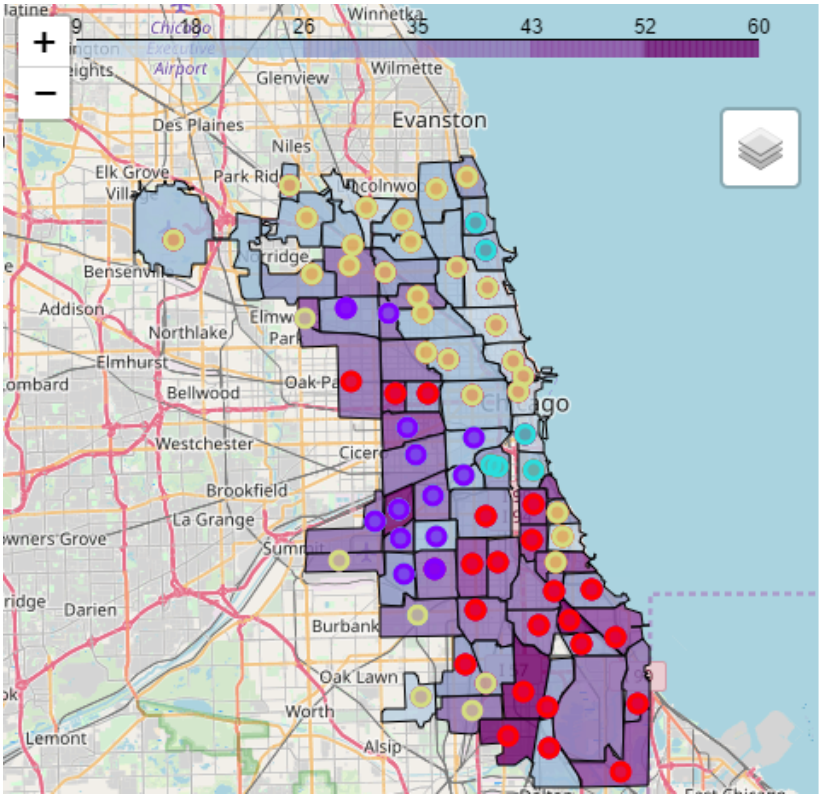
3.1 Restaurants

We have started by checking which are the most popular categories. For this, we needed to use one hot encoding to count the values per category, per Community Area. the obtained data was very sparse and we needed a general overview per Community Area, so the mean per Community area per restaurant subcategory was calculated.

With this information we have found the 10th most popular food categories per Community Area:

CA	1 most popular	2 most popular	3 most popular	4 most popular	5 most popular	6 most popular	7 most popular	8 most popular	9 most popular	10 most popular
ALBANY PARK	Sandwich Place	Pizza Place	Mexican Restaurant	Korean Restaurant	Asian Restaurant	Middle Eastern Restaurant	Chinese Restaurant	Diner	Thai Restaurant	Café
ARCHER HEIGHTS	Mexican Restaurant	Fast Food Restaurant	Pizza Place	Sandwich Place	Taco Place	Café	Donut Shop	Seafood Restaurant	American Restaurant	Fried Chicken Joint
ASHBURN	Sandwich Place	Mexican Restaurant	Pizza Place	Fast Food Restaurant	Seafood Restaurant	Fried Chicken Joint	American Restaurant	Donut Shop	Breakfast Spot	Chinese Restaurant
AUBURN GRESHAM	Fast Food Restaurant	Sandwich Place	Fried Chicken Joint	Seafood Restaurant	American Restaurant	Chinese Restaurant	Mexican Restaurant	Donut Shop	Pizza Place	Wings Joint
AUSTIN	Fast Food Restaurant	Fried Chicken Joint	Pizza Place	Sandwich Place	American Restaurant	Seafood Restaurant	BBQ Joint	Café	Chinese Restaurant	Italian Restaurant

Having obtained these values, we used K-means clustering algorithm to see how similar are our community areas regarding with restaurants. We have chosen 4 clusters, because we want to classify community areas in 4 regions according to their weight percentage. Then we created a new dataframe with Community Areas, Weight Percentage, Latitude, Longitude, Cluster number and their most popular venues. With this information, the clusters are shown in the following map:



We can see that there exists a relationship between the less obese region and the type of venues. We have also analyzed the type of restaurant venues per cluster as follows:

Cluster	**Types of Predominating Restaurants**
0 - Red	Fast Food Restaurants, Sandwich and Pizza places.
1 - Purple	Mexican Restaurants, Sandwiches and Pizza places.
2 - Blue	Chinese and vietnamese food, Mexican Restaurants and Pizza.
3 - Yellow	Sandwich and pizza, different countries cuisine restaurants e.g. Chinese Restaurant, Italian Restaurant, Thai Restaurant.

In cluster 3, we can see that also Sandwich and pizza places are predominating. Nevertheless, in the top ten of this Community Areas, there are a lot of different countries cuisine restaurants e.g. Chinese Restaurant, Italian Restaurant, Thai Restaurant, and so on, which implies more elaborate plates and less fast food ingredients, in other words, better quality ingredients.

For evaluating the relation between this clusters and obesity, we have used quantiles to divide the data in 4 equal parts according to obesity percentage and then verify in each group, which cluster is predominating:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Group 1 (<Q1)	0	1	2	16
Group 2 (Q1 to Q2)	5	4	2	7
Group 3 (Q2 to Q3)	7	3	1	7
Group 4 (Q3<)	10	5	1	3

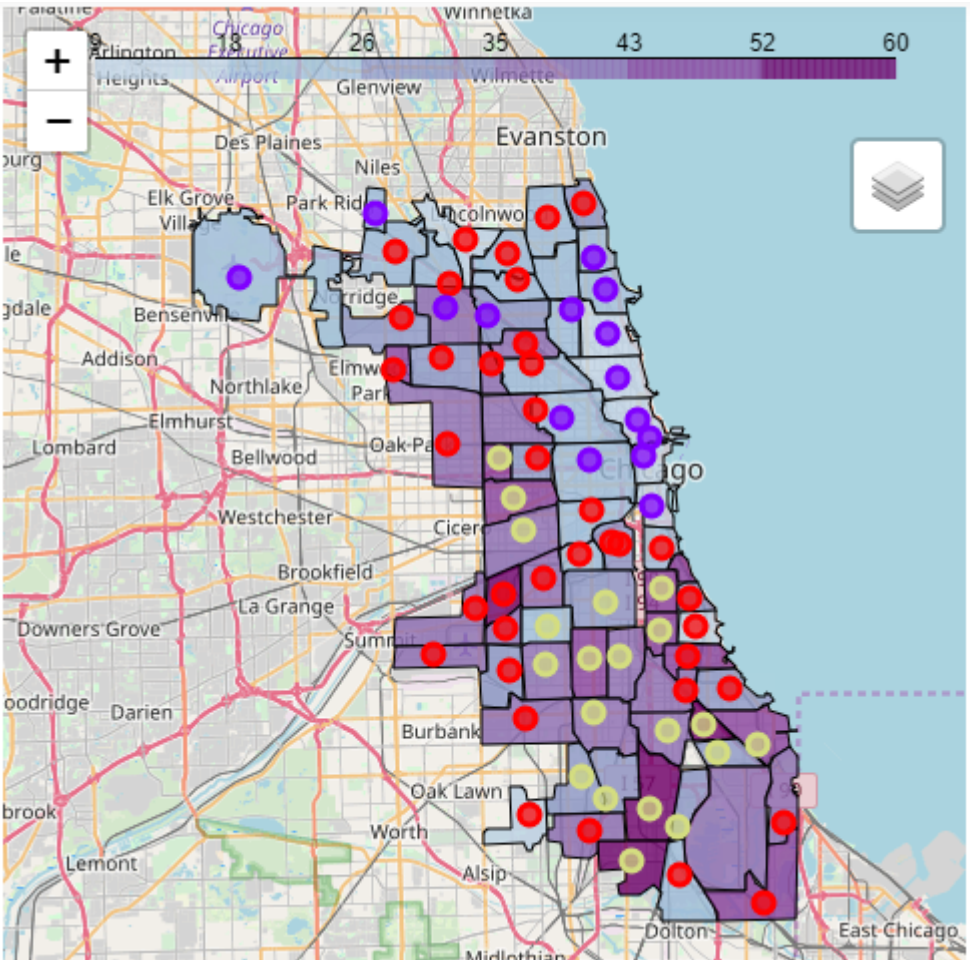
We have seen that in the Areas with lower weight percentage, the predominating cluster is 3. In the second group there is not a very clear pattern, we can see that values are similarly distributed. In group number 3 , it still have not a clear pattern, indeed there is a balance between 0 and 3 clusters, which are opposite in terms of subjective food quality, and finally in group 4, the predominating cluster is 0.

3.2 Outdoors and Recreation

Also for this type of venues, we needed to use one hot encoding to count the values per category, per Community Area. The obtained data was very sparse and we needed a general overview per Community Area, so the mean per Community area per restaurant subcategory was calculated. With this information we have found the 10th most popular outdoors and recreation categories per Community Area:

CA	1 most popular	2 most popular	3 most popular	4 most popular	5 most popular	6 most popular	7 most popular	8 most popular	9 most popular	10 most popular
ALBANY PARK	Park	Gym	Gym / Fitness Center	Martial Arts Dojo	Yoga Studio	Athletics & Sports	Baseball Field	Tennis Court	Garden	Gymnastics Gym
ARCHER HEIGHTS	Park	Gym / Fitness Center	Gym	Tree	Athletics & Sports	Dog Run	Basketball Court	Botanical Garden	Golf Course	Gym Pool
ASHBURN	Park	Gym	Gym / Fitness Center	Baseball Field	Martial Arts Dojo	Soccer Field	Yoga Studio	Athletics & Sports	Golf Course	Weight Loss Center
AUBURN GRESHAM	Park	Gym / Fitness Center	Gym	Golf Course	Martial Arts Dojo	Weight Loss Center	Athletics & Sports	Soccer Field	Baseball Field	Basketball Court
AUSTIN	Park	Gym / Fitness Center	Yoga Studio	Gym	Playground	Athletics & Sports	Weight Loss Center	Baseball Field	Garden	Golf Course

After obtaining the previous information, we used again K-means clustering algorithm. We created the new dataframe with Community Areas, Weight Percentage, Latitude, Longitude, Cluster number and their most popular Outdoor and Recreation venues. With this information, the clusters are shown in the following map:



We can see that there exists a relation between the less obese region and the type of recreation venues. We also analyzed the type of venues per cluster as follows:

Cluster	**Types of Predominating Outdoor and recreation venues**
0 - Red	Park, Gym, Baseball Field
1 - Purple	Gym, Park, Gym/Fitness Center, Yoga Studio
2 - Blue	Trail, farm and forest (only one CA)
3 - Yellow	Park, Gym, Golf, Baseball field, Basketball

For evaluating the relation between the obtained clusters and obesity, we used quantiles again to make our comparisons:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Group 1 (<Q1)	7	11	0	1
Group 2 (Q1 to Q2)	11	3	0	4
Group 3 (Q2 to Q3)	13	1	0	4
Group 4 (Q3<)	7	0	1	11

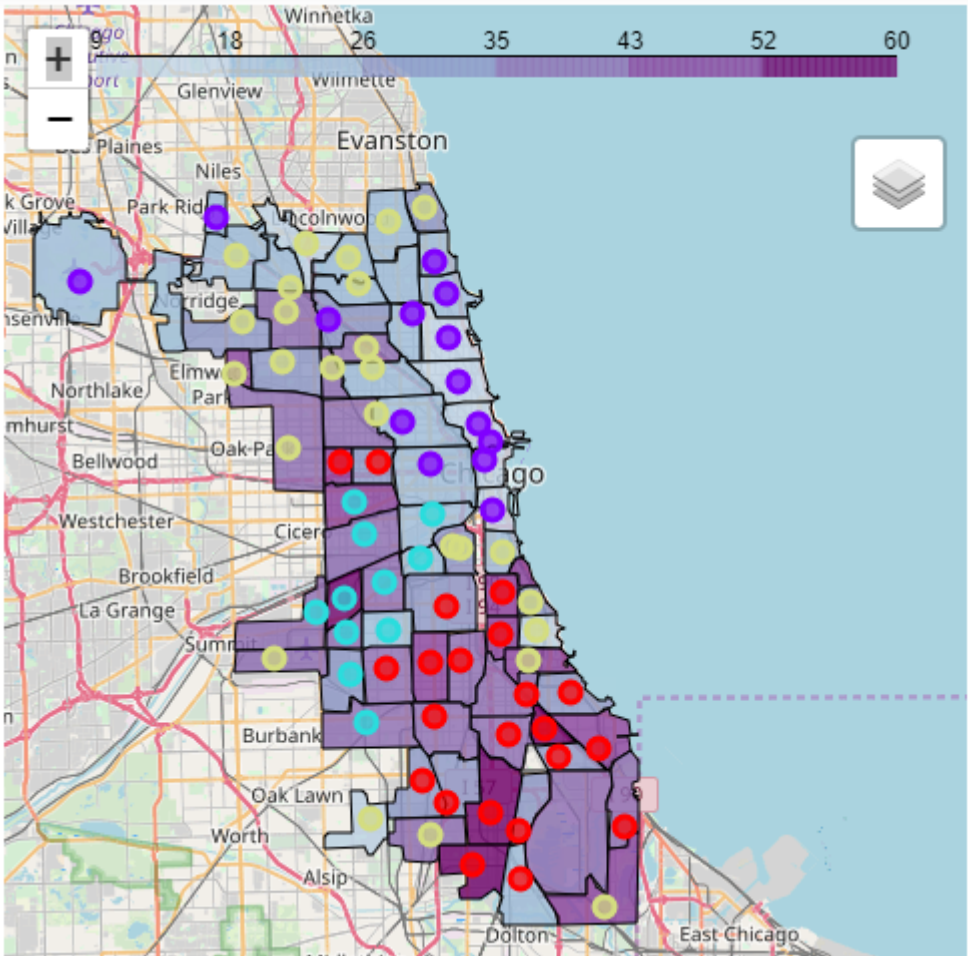
We have seen that in the Areas with less obesity rates, the predominating cluster is 1, where gym and fitness centers are the most popular venue, over parks. In contrast, where obesity predomines, most popular venues ar park, basketball and baseball fields. Nevertheless the obtained values for Outdoor and Recreation venues are not as clear as what we have obtained with restaurants, so now we will analyze both of them at the same time.

3.3 Merging “Restaurants” and “Outdoor and Recreation” Categories

In this case we have merged the already processed datasets with means per category per community area from restaurants and recreation venues, so we did not need to repeat this process. This data helps us to obtain again the 10th most popular venues between both categories, as follows:

Weight_Percent	lat	lon	cluster	1 most popular	2 most popular	3 most popular	4 most popular	5 most popular	6 most popular	7 most popular	8 most popular	9 most popular	10 most popular
42.3	41.750474	-87.664304	0	Park	Fast Food Restaurant	Sandwich Place	Fried Chicken Joint	Seafood Restaurant	Gym / Fitness Center	American Restaurant	Chinese Restaurant	Mexican Restaurant	Gym
55.7	41.745035	-87.588658	0	Park	Sandwich Place	Fast Food Restaurant	Fried Chicken Joint	Gym / Fitness Center	Gym	Pizza Place	Chinese Restaurant	Donut Shop	Athletics & Sports
32.6	41.718153	-87.671767	0	Park	Sandwich Place	Pizza Place	Fried Chicken Joint	Fast Food Restaurant	Gym / Fitness Center	Golf Course	Donut Shop	Mexican Restaurant	Seafood Restaurant
31.1	41.730035	-87.579213	0	Park	Fast Food Restaurant	Sandwich Place	Mexican Restaurant	Fried Chicken Joint	Gym / Fitness Center	Gym	Harbor / Marina	Pizza Place	Southern / Soul Food Restaurant
39.5	41.741145	-87.612548	0	Park	Sandwich Place	Fast Food Restaurant	Fried Chicken Joint	Gym / Fitness Center	Seafood Restaurant	Gym	Chinese Restaurant	Pizza Place	Mexican Restaurant

Then we have created clusters based on the mean of each category. Clusters are shown in the following map:



We can also see that there exists a relation between both types of venues and obesity. We analyzed the type of venues per cluster as follows:

Cluster	**Types of Predominating Venues**
0 - Red	Park, Fast Food Restaurant, Sandwich place
1 - Purple	Gym, Park, Gym/Fitness Center, Yoga Studio, Sandwich place
2 - Blue	Park, Mexican Restaurant, Gym/Fitness Center
3 - Yellow	Park, Gym, Various restaurants

With quantiles we evaluated again the relation between the obtained clusters and obesity:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Group 1 (<Q1)	0	11	1	7
Group 2 (Q1 to Q2)	6	3	3	6
Group 3 (Q2 to Q3)	7	0	3	8
Group 4 (Q3<)	9	1	4	5

We have seen that in the first group of areas with less obesity rates, the predominating cluster is 0, where gym, parks and fitness centers are more popular than food venues. On the other hand, in areas where obesity predomines, most popular venues are parks and then all types of fast food, before Gym. We can say that indeed venues are affecting (positively or negatively) the obesity rates in all community areas.

4. Results

Our analysis has given us some interesting results. We have found that food venues are in fact influencing the way in which people live and eat. In Community Areas where most popular food venues are fast food, fried chicken and sandwiches, there are the highest obesity percentages. On the contrary, in places where most popular food is traditional cuisine from different countries, there are lower obesity percentages.

If we examine Outdoor and Recreation venues, we can see that people tends to be in a better shape where the most popular venue is the Gym or Gym/Fitness center. This places must be specialized in creating routines to improve people's wellness and helping in weight loss. In places where the most popular venue is the park, maybe people is not going there only for exercising, but also for relaxing or rest, so go to the park does not guarantee that any person is necessarily making exercise.

If we combine both of the categories, we can see that even when most popular places are Outdoor and Recreation, it is common to find Park in the first place and then different types of fast food restaurants in the subsequent places, where there is a higher obesity index. In contrast, in places where the most popular venues are park and gym places, and then food, they tend to be in a better shape.

5. Discussion

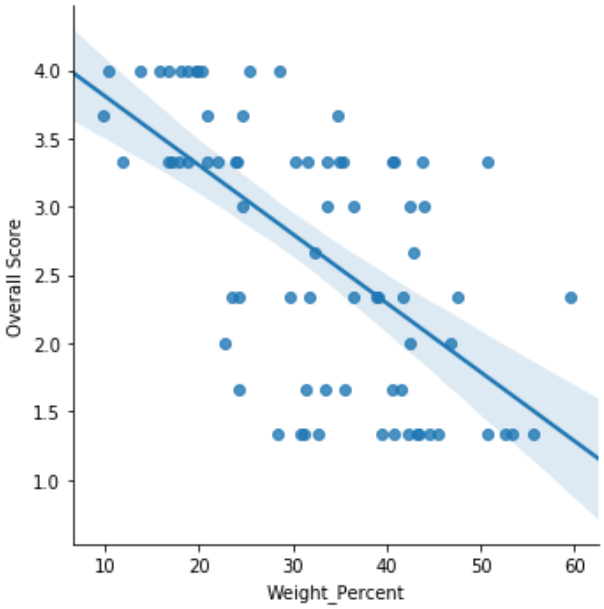
To summarize the results, we have made a qualitative subjective classification from 4 as the best, to 1 as the worst, according to the clusters we are obtaining:

points	food	recreation	food and recreation
4	Cluster 3	Cluster 1	Cluster 1
3	Cluster 2	Cluster 0	Cluster 3
2	Cluster 1	Cluster 3	Cluster 2
1	Cluster 0	Cluster 2	Cluster 0

Then a table was created with this values per Community Area to obtain the mean of this subjective punctuation:

Geo_Group	Weight_Percent	food	recreation	food and recreation	Overall Score
NEAR SOUTH SIDE	9.7	3	4	4	3.666667
LOOP	10.4	4	4	4	4.000000
HYDE PARK	11.9	4	3	3	3.333333
NEAR NORTH SIDE	13.8	4	4	4	4.000000
LINCOLN PARK	15.8	4	4	4	4.000000
FOREST GLEN	16.7	4	3	3	3.333333
LAKEVIEW	16.8	4	4	4	4.000000
MOUNT GREENWOOD	17.1	4	3	3	3.333333

For better visualization, let's see the results in a graph. The following graph shows that while ***Weight_Percent*** increases, the ***Overall Score*** decreases. Even when the linearity is not strong, we can see that the Overall Score is High when the weight percent is less. This is showing the relation between venues and obesity. Nevertheless it is necessary to remind that this subjective classification is merely done to show a relation, but we cannot assure which food or venues are good or bad. Anyway, we can find a contrast between the results at the beginning of the table and the ones at the end in our sorted table. the tendency is to find higher scores where there are low weight percentage values.



There is also something else that we could do in a next opportunity, as a recommendation, which is to compare this results also with the salary income per region in Chicago, because sometimes certain venues e.g. sushi restaurants are harder to find in places where there are lower incomes. This can also affect the type of venues, food quality, access to Gym and so on.

6. Conclusion

This project was aimed to find how venues could affect obesity rates per Community Area in Chicago, so that organizations like CDC can offer solutions knowing that there is a problem in certain community areas with the food that they are consuming, and the lack of physical activity.

In this project we have found that there exists a relation between venues and obesity, in places where people go mostly to gym, they tend to be thinner than where people goes mostly to parks. We have also seen that when it exist different varieties of food, which is also more elaborated (perhaps less fried), obesity rates are lower than in places where the most popular venues are fast food and sandwiches.

This insights are aimed to give tools to the government and CDC to create programs that incentivize people to exercise, perhaps creating more affordable gyms for community areas with less concurrency to these places, as well as make agreements with the owners of fast food venues to help their clients to eat healthier and better.