Capstone Project - The Battle of the Neighborhoods

Obesity vs Restaurants and Income

Yu-Ling Hsiao June 2, 2019

1. Introduction

Obesity has become a threat to our civilization. The problems result from obesity are both physically and mentally harmful, like cardiovascular diseases, diabetes, or depression. This can put pressure on not only individuals but also the health care or medical system of the government.

There are different factors that might lead to this problem, for example, environment, lifestyle or habit. Here in this project, I want to identify whether the types of restaurants of the given area have a relationship with the obesity rate, and further examine if the obesity rate and personal income are related.

I will use data science powers to find the correlation between types of restaurants, obesity rate, and personal income. Hopefully, we can gain insight that helps tackle the obesity problem at its roots for the stakeholders, including the government, society and every person.

2. Data

Based on the problem, we need to examine the following factors:

- obesity rate
- types of popular restaurants
- personal income

Because of the availability of the dataset on the Internet, we can only get the city-level obesity rate and county-level personal income data.
Following sources will be needed to generate the required information:

- obesity among adults of 500 cities using csv file from **https://catalog.data.gov/dataset/500-cities-city-level-data-gis-friendly-format-845f9**
- the list of categories of the most checked-in restaurants in the city using **Foursquare API**
- the personal income of the counties using csv file from **https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas**
- the county name of the given cities using csv file from **https://simplemaps.com/data/us-cities**

For data cleansing, here is what I do for four dataset:

- For the obesity rate one, I read the csv file into a dataframe and remained only the obesity rate and its relevant location data. Here is a sample shot of the data.

	Location	State	City	Obesity	Population	Latitude	Longitude
187	Fremont, CA	CA	Fremont	15.3	214089	37.527869	-121.984122
291	Milpitas, CA	CA	Milpitas	15.7	66790	37.433870	-121.892083
259	Boulder, CO	CO	Boulder	16.6	97385	40.027551	-105.251518
380	Irvine, CA	CA	Irvine	16.7	212375	33.678011	-117.773633
304	Union City, CA	CA	Union City	17.0	69516	37.602839	-122.018966

- For the most checked-in food venues from Foursquare, I used the latitude and longitude to find the 50 food venues from each city. Then transformed the venues into dataframe and remain only the ones that appear in our obesity data because the Foursquare API might include the data from nearby cities. Finally, I deleted the venue categories which have less than 20 venues and the cities that have less than 30 venues.

	state	city	categories	name	lat	lng	Location
1	CA	Sacramento	Sushi Restaurant	Kura Revolving Sushi Bar	38.593780	-121.419040	Sacramento, CA
2	CA	Sacramento	Food Court	The Bank	38.581611	-121.497487	Sacramento, CA
5	CA	Sacramento	American Restaurant	Yard House	38.581382	-121.501619	Sacramento, CA
6	CA	Sacramento	Steakhouse	Echo & Rig	38.581521	-121.499316	Sacramento, CA
15	CA	Sacramento	American Restaurant	Esquire Grill	38.695456	-121.592825	Sacramento, CA

- For personal income data, I first downloaded the csv file and replaced the state name with its abbreviation, then merged the country and state column as one column.

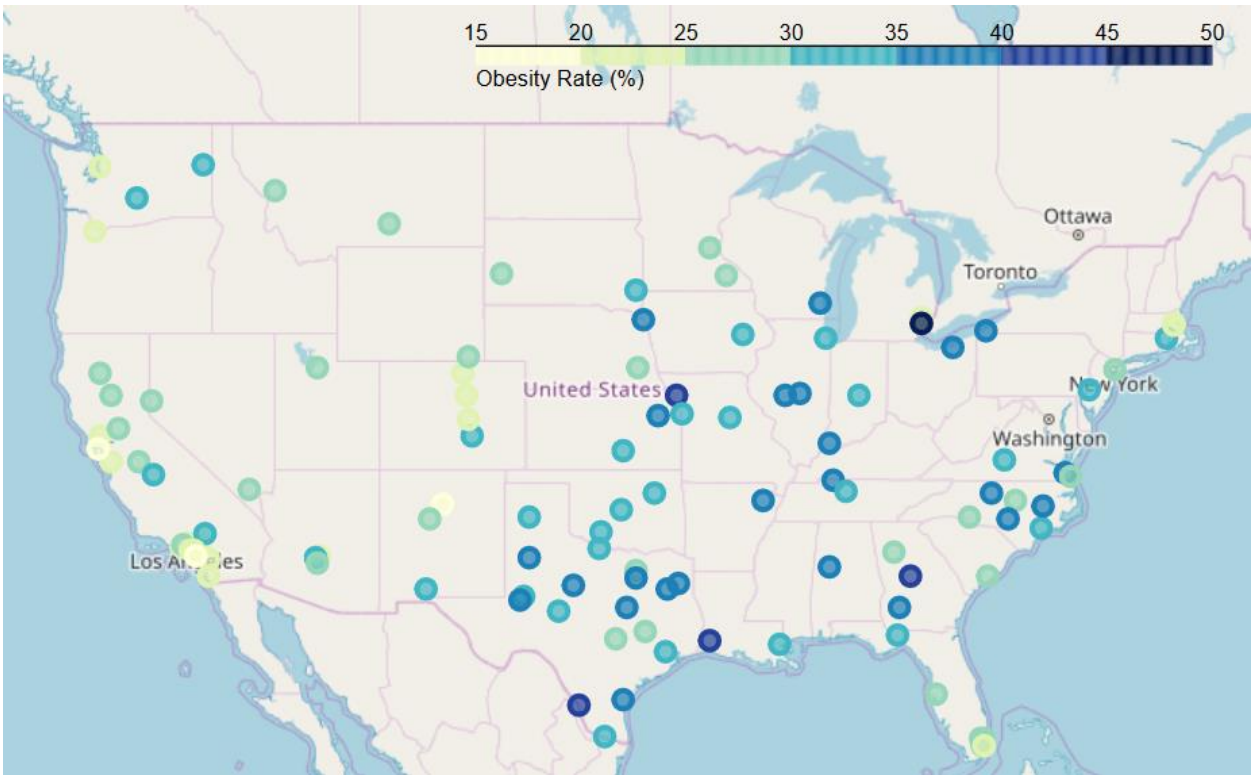
	County	Income
0	Autauga, AL	40484
1	Baldwin, AL	44079
2	Barbour, AL	33453
3	Bibb, AL	30022
4	Blount, AL	33707

- For city and county data, I also downloaded the file and drop unnecessary column except state, city and county ones. Then I merged the state and city columns as one location column for later use.

	county_name	Location
0	Pierce	Prairie Ridge, WA
1	Skagit	Edison, WA
2	Lewis	Packwood, WA
3	Kitsap	Wautauga Beach, WA
4	Kitsap	Harper, WA

3. Methodology

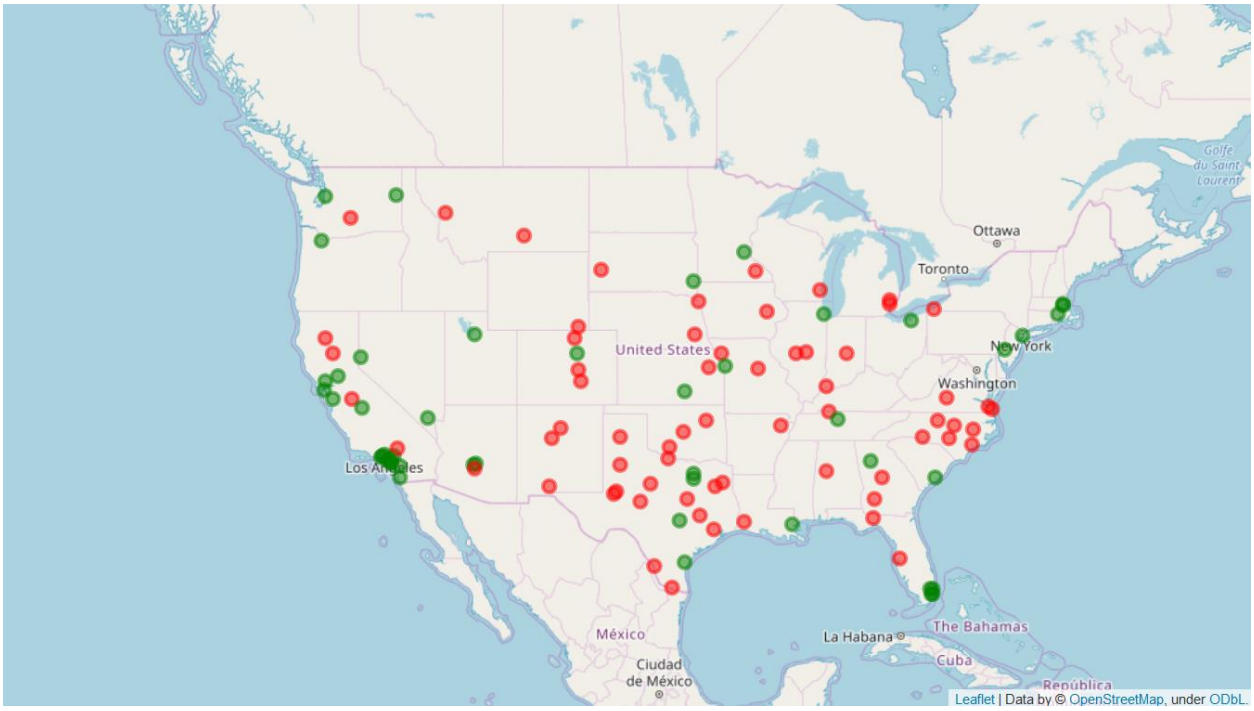
Let's have a rough idea of how our obesity dataset looks like.



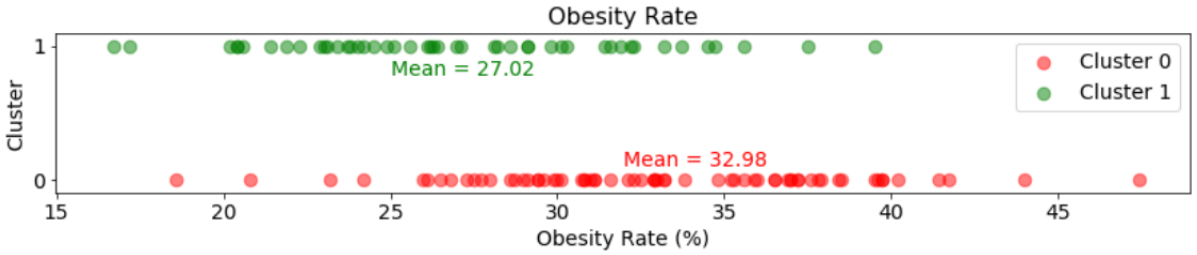
I used one hot encoding to transform the food venues data to create a dataframe that the columns are presented as categories and the rows presented as each venue, and marked the value with one for the category of that venue. Then grouped the dataframe by city to get the numbers of each venue category for one city and sorted the top 10 most common venues by the desending order.

	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abilene, TX	Fast Food Restaurant	Mexican Restaurant	Burger Joint	Tex-Mex Restaurant	Chinese Restaurant	Restaurant	Italian Restaurant	Wings Joint	Bakery	Donut Shop
1	Albany, GA	Fast Food Restaurant	Mexican Restaurant	Sandwich Place	Seafood Restaurant	Steakhouse	Wings Joint	Italian Restaurant	Burger Joint	Restaurant	Buffet
2	Albuquerque, NM	Fast Food Restaurant	Mexican Restaurant	Food Court	Wings Joint	Seafood Restaurant	Donut Shop	Hot Dog Joint	Pizza Place	Restaurant	BBQ Joint
3	Amarillo, TX	Fast Food Restaurant	Burger Joint	Mexican Restaurant	BBQ Joint	Restaurant	Thai Restaurant	Steakhouse	Sandwich Place	Breakfast Spot	Cafeteria
4	Anaheim, CA	Bakery	Italian Restaurant	Burger Joint	Middle Eastern Restaurant	Ice Cream Shop	Diner	Cajun / Creole Restaurant	Fast Food Restaurant	Mexican Restaurant	Mediterranean Restaurant

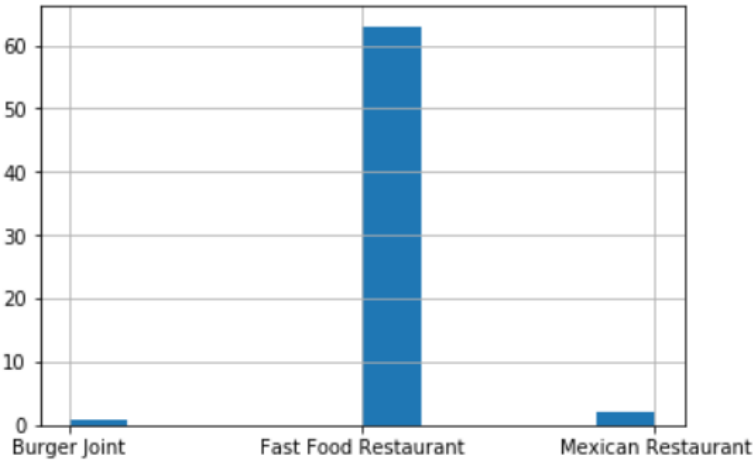
For analysis, I chose K-means to classify the cities to two cluster. Each is expected to have similar composition of the food venue categories. And the following picture shows the cluster result. Red is cluster 0, green is cluster 1.



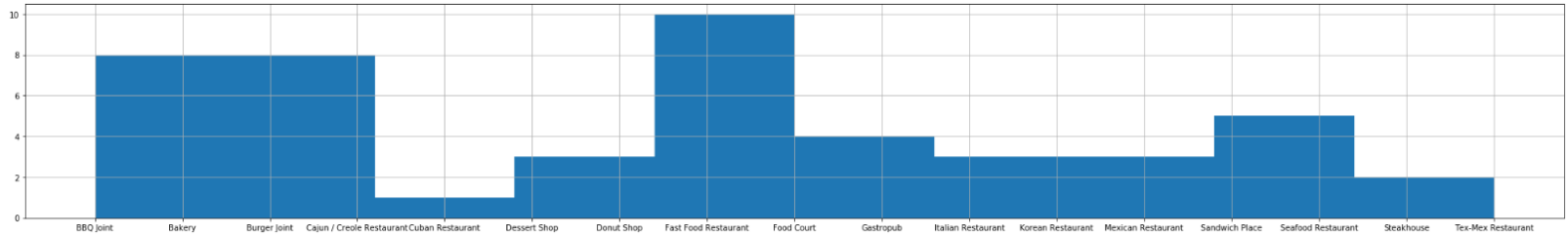
Here is the scatter chart of two clusters. It is obvious that cluster 1 is generally having lower obesity rate than cluster 2, with the mean rate of 27.02, while the other one is 32.98.



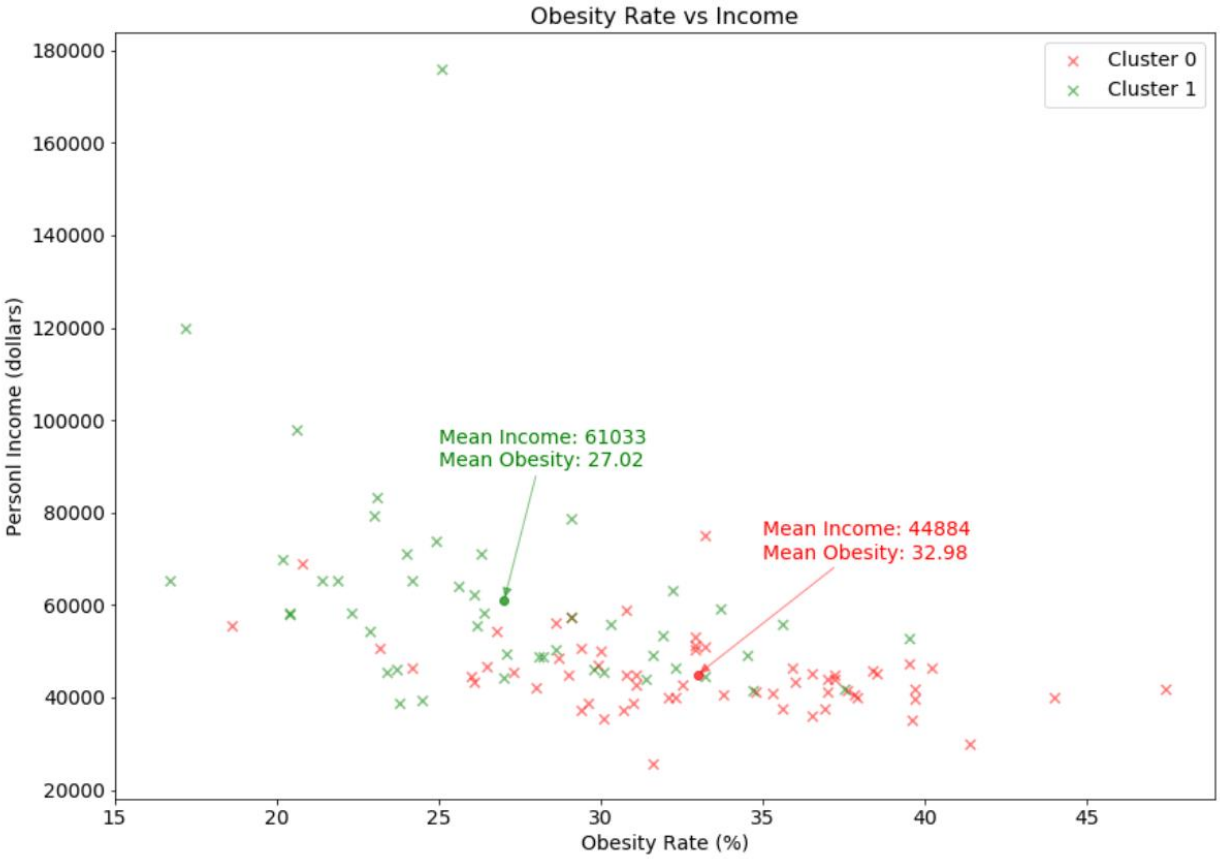
I summarized the data to see what is the mode for the 1st common venue column in both column. For cluster 0, we can see that the Fast Food Restaurant has overwhelming number.



For Cluster 1, the distribution of the categories is very diverse.



Finally, I merged the income data with the cluster based on county and plot the annual personal income and obesity rate on scatter chart. The points of cluster 0 are scattering at the lower right part, with the mean income of 44884, while cluster 1's are at the relatively upper left part, with the mean income of 61033.



As we can see, the personal income might be an influencing factor of obesity rate.
The lower the income is, the higher chance of obesity one city has. On the contrary, the richer cities have lower obesity rate.

4. Results

Our analysis shows that income and eating habits might have something to do with obesity. By gathering the most checked-in food venues in cities from Foursquare, we clustered those who share similar types of food venues into two groups. Then from the calculation of the mean obesity from two clusters, we knew the cities which the most popular food venue is fast food restaurant are usually more overweight. Moreover, as we look into the personal income data, we know that the cities which are wealthier usually have normal weight. Especially when we look into the lower obesity rate cluster, we can see the cities within the ones that are bigger, more developed, like Los Angeles, New York, San Francisco, Seattle, Portland, Philadelphia, Phoenix, Denver, and Dallas.

5. Discussion

Obviously, obesity is not a personal problem. Sometimes people would think those who are overweight are lazy, lack the ability of self-control, or have other navigate behaviors. But as our analysis shows, obesity is might a social environment problem. People who live in poverty might go to a fast food restaurant more often because it is more affordable for them, and result in a higher chance of obesity.

If the government wants to solve this problem, it is not only an issue of proposing a healthy lifestyle or provide diet program to people, it might be a poverty problem. And we should tackle this problem at its root like considering how to reduce the gap between rich and poor.

6. Conclusion

Purpose of this project was to identify if there is a relationship between the obesity rate and types of restaurants or personal income so our stakeholders can take action to solve this civilization disease. By clustering the cities to two groups which share similar characteristics and compare to obesity rate and personal income, we find a tendency that the more popular the fast food restaurant is in a city, the higher obesity rate and lower personal income it has. Contrarily, the higher income cities are usually not that obese and have a diverse choices of food venues.

While rather take obesity as a personal problem, every stakeholder might consider paying attention on how to make everyone to not only affordable for fast food restaurant when they eat out.