Opening a Fitness Center Based on Regional Health Patterns

1. Introduction/Business Problem

The purpose of this report is to utilize freely available data to determine the optimal location for opening a new fitness center. In this case, the optimal location would be a highly unhealthy region of the United States where there are little to no fitness centers within a reasonable traveling distance. In 2015-2016, about 39.8% of adults in America were obese[1], while about 18.5% of American children ages 2-19 years old were obese[2]. About 9.4% of the US population has been diagnosed with diabetes[3]. Since this data was sampled and presented, these numbers have continued to rise every year. Opening a fitness center in a highly unhealthy location where the most people would benefit from it could help slowly reverse this trend. If successful, the process could be repeated for more unhealthy areas, as to benefit as many people as possible.

The stakeholders in this project consist of unhealthy people, which for the purposes of this report, will be Americans who are obese and/or have diabetes. This can also be extended to their families and friends, as they do not want to have any unhealthy loved ones. This could also possibly extend to medical offices and hospitals, as a decrease in unhealthy Americans would likely lead to a little less stress for them.

Based on this information, the question being asked here is: *Can we use available data to determine optimal locations for opening new fitness centers?*

2. Data

The data being used for this report comes from FourSquare.com, and the USDA ERS (United States Department of Agriculture Economic Research Service). The USDA ERS hosts a dataset[4] on their website called the Food Environment Atlas, which details data on demographics (such as age and race), access to nearby grocery stores and fast food restaurants, SNAP benefits, health/physical activity rates, income level, and much more. The main set of data from this dataset that will be used here is the data in the

Health section; more specifically, information on physical activity rates, and adult obesity/diabetes rates from 2008 and 2013. Other data from the dataset will be used as well for comparison and analysis at times.

FourSquare.com is an online platform where users can search for different types of venues in their current location, such as different restaurants, shopping locations, and entertainment options. The FourSquare API will be leveraged in order to pull and visualize this data; specifically, data regarding nearby fitness centers will be utilized to help determine optimal locations for opening new fitness centers.

3. Methodology

As mentioned above, the main dataset used in this project was the USDA's Food Environment Atlas (hereafter referred to as the "FEA"). This dataset was downloaded as a Microsoft Excel file and converted to a Pandas dataframe; the FEA contains multiple spreadsheets detailing numerous facts about each recorded county, including demographics, population, local farms/farmers' markets, grocery stores, and SNAP-authorized stores, among other data. The main spreadsheet from this dataset that was used for this report was the health spreadsheet; it included data on adult obesity/diabetes rates for the years 2008 and 2013, the number of recreational/fitness facilities in the county, and the percentage change of recreational/fitness facilities in the county from the years 2009-2014. Note that this was only the main data used for the project; there were numerous other categories of data included in the health spreadsheet, such as percentage of physically active high schoolers. This data was excluded from the analysis, as it was not relevant to this analysis.

After removing empty entries from the dataframe, some inferential statistics were calculated for the dataframe; these statistics included the mean, standard deviation, minimum and maximum values for all 3,014 counties in the data. This can be seen below in Figure 1:

	FIPS	Adult Diabetes Rate (2008)	Adult Diabetes Rate (2013)	Adult Obesity Rate (2008)	Adult Obesity Rate (2013)	Recreation/Fitness Facilities (2009)	Recreation/Fitness Facilities (2014)	Recreation/Fitness Facilities % Change (2009-2014)
count	3014.000000	3014.000000	3014.000000	3014.000000	3014.000000	3014.000000	3014.000000	3014.000000
mean	30523.340080	9.886297	11.207465	28.878202	30.958228	10.062376	10.470471	-5.754836
std	15126.881443	2.056662	2.472564	3.697387	4.513673	30.484952	34.245433	50.332383
min	1001.000000	3.000000	3.300000	11.700000	11.800000	0.000000	0.000000	-100.000000
25%	19013.500000	8.500000	9.500000	27.100000	28.300000	0.000000	0.000000	-25.000000
50%	29208.000000	9.700000	11.100000	29.100000	31.200000	2.000000	2.000000	0.000000
75%	45088.500000	11.300000	12.800000	30.900000	33.800000	7.000000	6.000000	0.000000
max	56045.000000	18.200000	23.500000	43.700000	46.300000	738.000000	845.000000	400.000000

Figure 1: Inferential statistics for the data.

The FIPS column here can be safely ignored, as these are just the federal codes for each county. As seen above, the average adult obesity rate was approximately 28.88% in 2008, and 30.96% in 2013, showing a general increase in obesity rates; the same can be said for adult diabetes rates as well. What should also be noted here is that the average amount of recreation/fitness facilities among all counties was almost the exact same from 2009 to 2014, whereas there was an average 5.75% decrease overall in recreation/fitness facilities. This fact could be correlated with the rise in obesity over time, but there is not enough hard data to support this.

The dataframe was then sorted in descending order of adult obesity rate in 2008, and the top 5 counties were analyzed for trends and similarities. The top 5 counties are shown in Figure 2.

	FIPS	State	County	Adult Diabetes Rate (2008)	Adult Diabetes Rate (2013)	Adult Obesity Rate (2008)	Adult Obesity Rate (2013)	Recreation/Fitness Facilities (2009)	Recreation/Fitness Facilities (2014)	Recreation/Fitness Facilities % Change (2009-2014)
31	1063	AL	Greene	18.2	21.0	43.7	46.3	0	0	0.0
1414	28027	MS	Coahoma	15.9	16.8	42.7	42.9	1	1	0.0
1426	28051	MS	Holmes	15.6	17.5	42.2	46.1	0	0	0.0
1427	28053	MS	Humphreys	15.7	17.3	42.1	41.6	0	0	0.0
1432	28063	MS	Jefferson	15.3	16.9	41.8	44.5	0	0	0.0

Figure 2: The top 5 counties for adult obesity rate in 2008.

The adult obesity rates for 2008 and 2013 for each of the top 5 counties can be visualized with the Seaborn library as line graphs; these graphs are shown in Figure 3.

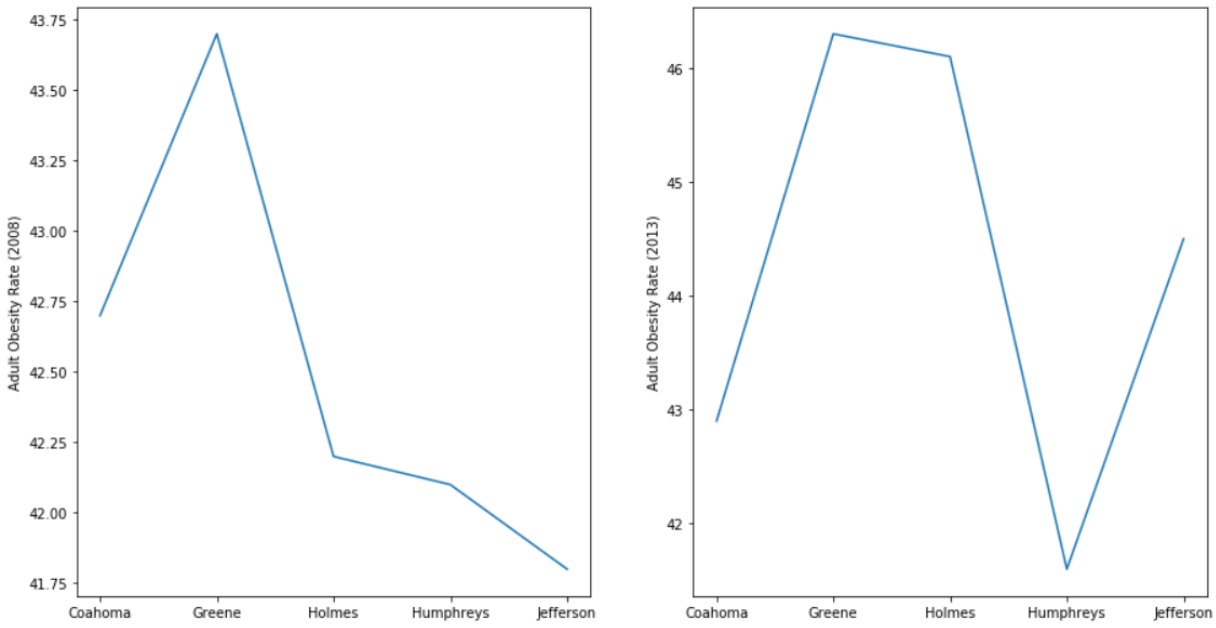

Figure 3: Visualization of the top 5 counties for adult obesity rate in 2008 and 2013.

The percentage change in obesity and diabetes rates between 2008 and 2013 can be calculated and added to the dataframe, as shown in Figure 4.

	FIPS	State	Adult Diabetes Rate (2008)	Adult Diabetes Rate (2013)	Adult Obesity Rate (2008)	Adult Obesity Rate (2013)	Recreation/Fitness Facilities (2009)	Recreation/Fitness Facilities (2014)	Recreation/Fitness Facilities % Change (2009-2014)	% Change Diabetes Rate	% Change Obesity Rate
County											
Greene	1063	AL	18.2	21.0	43.7	46.3	0	0	0.0	2.8	2.6
Coahoma	28027	MS	15.9	16.8	42.7	42.9	1	1	0.0	0.9	0.2
Holmes	28051	MS	15.6	17.5	42.2	46.1	0	0	0.0	1.9	3.9
Humphreys	28053	MS	15.7	17.3	42.1	41.6	0	0	0.0	1.6	-0.5
Jefferson	28063	MS	15.3	16.9	41.8	44.5	0	0	0.0	1.6	2.7

Figure 4: Percent change in diabetes and obesity rates between 2008 and 2013.

This updated dataframe shows that the highest increases in both diabetes and obesity came from Holmes, Jefferson and Greene counties.

The top 50 obese counties from the list were then clustered into 10 separate clusters using k-means clustering, a machine learning algorithm that groups data together into clusters based on similar features. The data was normalized and then clustered, as shown in Figure 5:

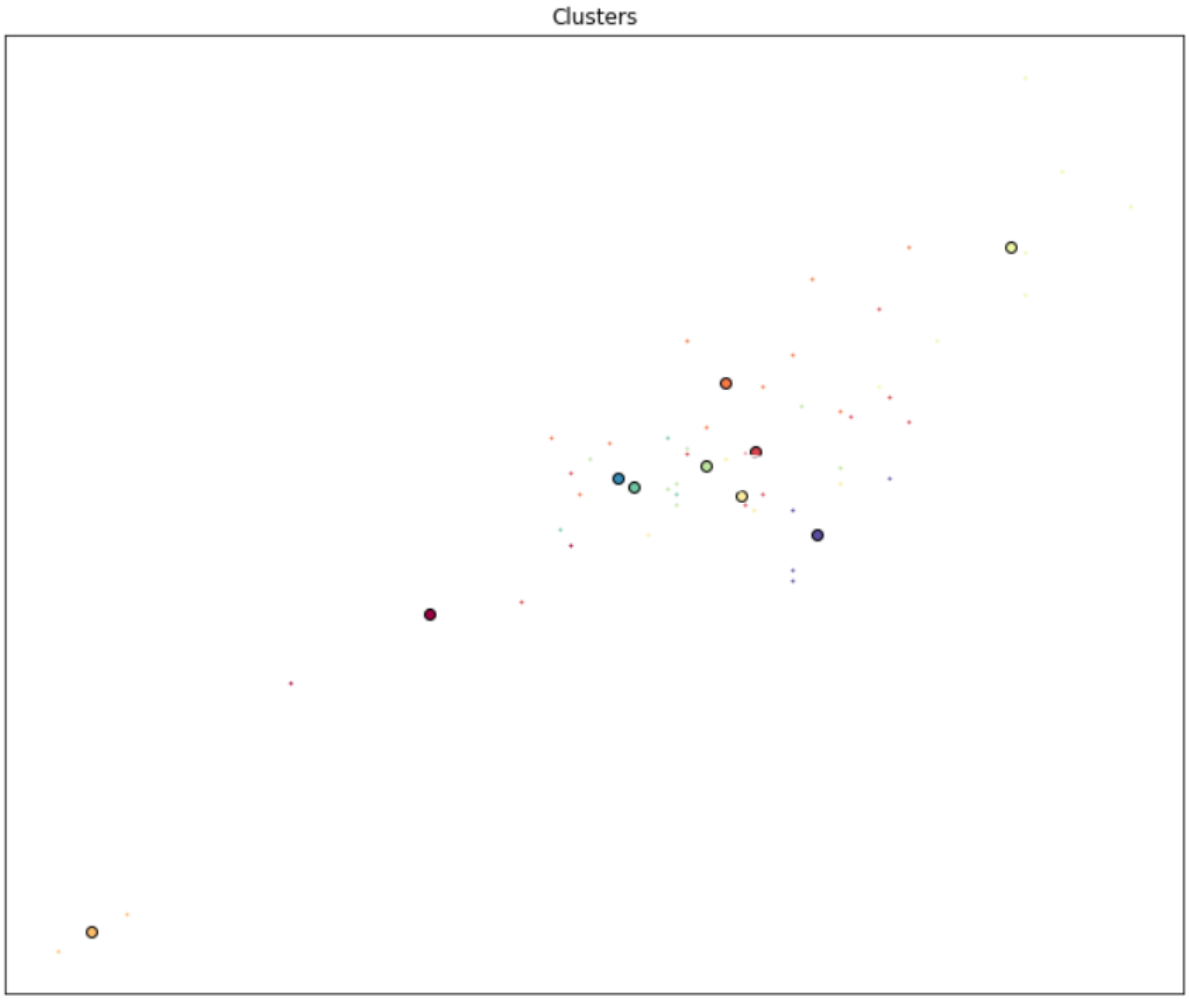

Figure 5: The data, normalized and clustered, plotted on a graph.

These clusters were then plotted on a map of the United States using the Folium library. The latitude and longitude for each of the top 50 counties was taken from each county's respective Wikipedia page, added to the dataframe, then plotted using Folium as shown in Figure 6:

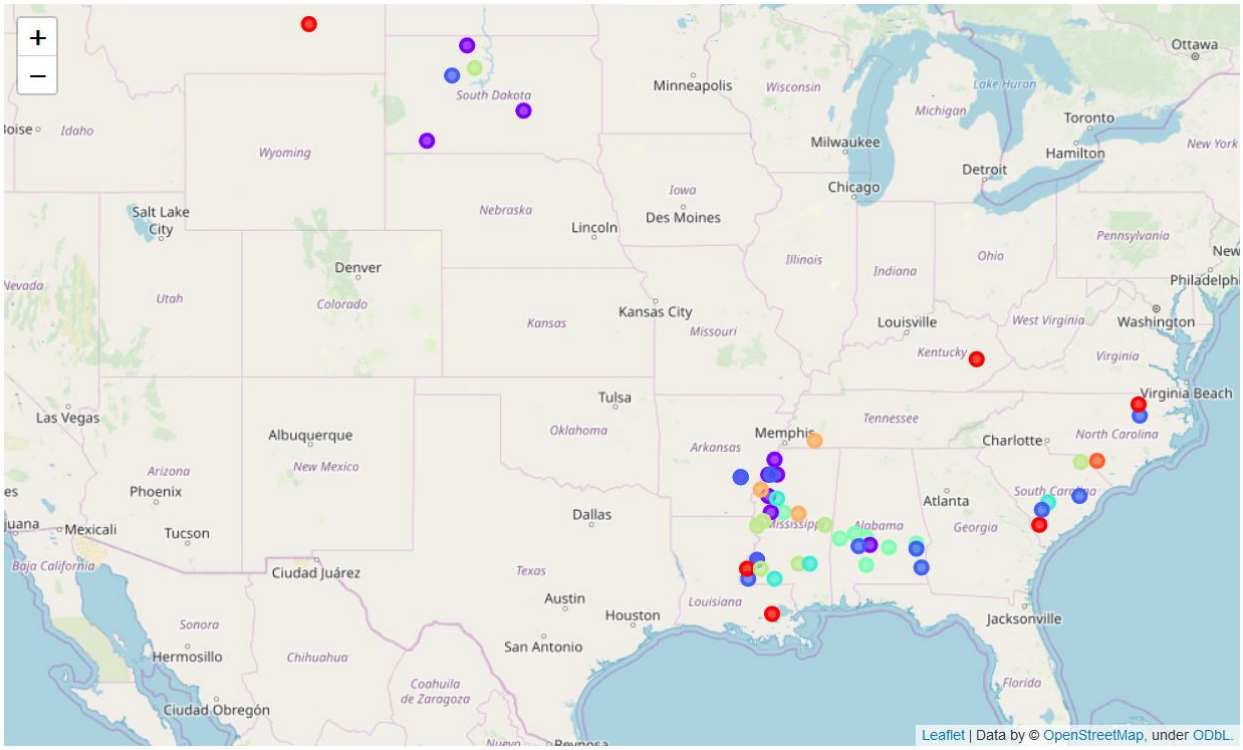

Figure 6: All the clusters plotted on an interactive map of the United States using Folium.

The FourSquare API was utilized to explore Greene County in Alabama, which was the county with the highest adult obesity rate in 2008. FourSquare is a location service that allows people to review and recommend venues to other users that are in their current location. Using FourSquare, Greene County was explored for the presence of fitness centers within a 100 kilometer radius (100 kilometers was the maximum range FourSquare would allow to be searched from an API call); the resulting analysis was stored in a dataframe, shown in Figure 7.

	name	categories	lat	lng
0	PieLab	Café	32.704189	-87.595226
1	Valley Grande Mexican Grill	Mexican Restaurant	32.482748	-87.000929
2	Yogurt Mountain	Ice Cream Shop	33.196479	-87.526792
3	Bryant-Denny Stadium	College Football Field	33.207683	-87.550563
4	The Fresh Market	Grocery Store	33.200300	-87.525970

Figure 7: Dataframe containing 100 venues in Greene County, Alabama, and their respective categories.

Scanning through the dataframe, there were no gyms or fitness centers found anywhere in Greene County.

4. Results

From the analysis, a number of observations can be made regarding the 5 counties with the highest adult obesity rates:

- Most of the time, adult obesity rates increased from 2008 to 2013.
- The diabetes rate is less than half the obesity rate for each county, so we can conclude that the two statistics are not correlated and likely have a correlation coefficient close to 0. In other words, just because an adult is obese does not mean they are going to get diabetes as well. Interestingly though, diabetes rates still increased regardless for each county, sometimes more than obesity rates increased.
- Out of all 5 counties, only one had a fitness facility; the rest had none for that entire period. Similarly, all 5 counties had a 0% change in fitness facilities from 2009-2014. There is a good chance this is correlated with high obesity rates.
- The 2 states represented in the top 5 counties are both adjacent to each other, and are both located in the southern United States.

Looking at the United States map plotted with Folium, the vast majority of the 50 counties with highest obesity rates are located in the southern/southeastern United States. While this is an interesting trend, there is not enough data to assume any sort of correlation here, especially since this subset of the data is so small compared to the entire dataset.

The counties were divided into 10 separate clusters based on their similarities to the other counties. For example, Greene, Holmes, Jefferson, Wilcox, and Perry Counties were all clustered into group 8, likely due to the fact that they all had some of the largest obesity rates for 2013, and also had no fitness centers anywhere in the county.

5. Discussion

Based on the above analysis and observations noted, a list of potential candidates for a fitness center has been generated; however, there are still many other factors not discussed here that would need to be taken into consideration before deciding to build a fitness center here, such as average citizen income, crime levels in the area, county funding, and whether citizens actually *want* a fitness center nearby (among other factors). For these reasons, a list of the top counties analyzed here should be submitted to stakeholders, rather than only one single county. My recommendation would be to begin by submitting a list of the top 10 counties for adult obesity rate, and letting stakeholders select the optimal location from this list based on the factors listed above. Should this list not be suitable enough for stakeholders, the next 10 counties in the dataframe can be submitted for consideration.

6. Conclusion

This project aimed to find the optimal location in the United States for opening a fitness center based on available health and local data; while no location was singled out, a list of suitable locations was put together to be submitted to stakeholders for deliberation and approval. Adult obesity and diabetes rates, amount of local recreation/fitness centers, and percent changes in obesity and diabetes rates for each county were analyzed and determined to be the most important features. Machine learning was used to cluster similar counties together, and were then visualized on a map of the United States.

7. References

[1] Adult Obesity Facts, Centers for Disease Control and Prevention. August 13, 2018. Retrieved from https://www.cdc.gov/obesity/data/adult.html

[2] Childhood Obesity Facts, Centers for Disease Control and Prevention. August 13, 2018. Retrieved from https://www.cdc.gov/obesity/data/childhood.html

[3] National Diabetes Statistics Report, Centers for Disease Control and Prevention. February 24, 2018. Retrieved from https://www.cdc.gov/diabetes/data/statistics/statistics-report.html

[4] Data Access and Documentation Downloads, United States Department of Agriculture Economic Research Service. March 27, 2018. Retrieved from https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/