Obesity and City Composition

Exploring the Impact of Environmental Stimuli

Amy Peniston

July 21, 2019

1 Introduction

In America, obesity has become an epidemic. According to the Centers for Disease Control and Prevention, more than 30% of adults over the age of 18 are obese [1]. Despite growing recognition of the severity of the problem, including significant research and investment in public health initiatives, the obesity rate continues to rise. Now, as we begin to feel the societal impacts of this phenomenon, we have to ask ourselves: what can be done to combat obesity?

Obesity is a medical condition in which an individual carries excess weight or body fat resulting in a body mass index (BMI) of 30 or above. It is linked to an increased risk for a variety of serious health conditions including hypertension, type 2 diabetes, stroke and heart disease [2]. Unfortunately, the obesity epidemic has wide-ranging ramifications, affecting not only individuals' well-being but also burdening society with added medical costs and decreased productivity. Research indicates that the total annual economic impact of obesity is in excess of $215 billion [3]. It is clear that society must actively address this growing health crisis or suffer deadly and expensive consequences.

Obesity is fundamentally caused by a prolonged caloric excess; this physiological imbalance is in turn impacted by an individual's external surroundings. As environments have changed such that high levels of physical activity are no longer required and calorie dense foods are available in abundance, there has been a marked trend towards sedentary lifestyles and overeating. Indeed, obesity is the body's natural response to the comfortable environment that humanity has engineered.

Obesity researchers, health organizations and public policy makers are three groups who are particularly interested in understanding how environmental factors such as easy access to fast food are facilitating society's weight gain. The million dollar question is whether we are able to modify the environment to help limit the extent of this growing epidemic. While it is unlikely that we will be able to eliminate obesity completely by adding or removing certain external stimuli, there is the potential for encouraging small behavioral changes, which over time can add up to a significant improvement in public health.

In this study, I examine the link between obesity and the environment. To do this, I apply various data science and machine learning techniques to a comprehensive location dataset, comparing and contrasting findings for regions with the highest and lowest obesity rates. Capital cities are then clustered by their composition and I examine the differences in the average obesity rates between clusters. I have tailored the direction and methodology of this research in the context of understanding what makes an environment "healthy".

The findings of this study would be of particular interest to government agencies and public policy makers looking to understand and prevent obesity. If a measurable connection does indeed exist between obesity and city composition, it might be possible to identify "at risk" environments that encourage lifestyle choices leading to decreased levels of physical activity and poor diet. Ultimately, this research could be used to seed the growth of healthier cities by helping to engineer surroundings that are optimized for our well-being.

2 Data

2.1 Sources

First, I leveraged data collected by the Behavioral Risk Factor Surveillance System (BRFSS) via stateofobesity.org to visualize the prevalence of obesity in the United States. Specifically I was interested in obesity rates, that is, the percent of adults aged 18 years and older who have obesity. Using historical BRFSS statistics dating back to 2000, I confirmed the increasing rates of obesity and identified the most and least obese states.

Cognizant of Foursquare and Mapquest API request limits, I decided to restrict my analysis to one area in each state, namely the capital city. I scraped city and zip code data for all states from Wikipedia and City-Data.com. Each zipcode was converted into coordinate form (latitude/longitude) using the Mapquest geocoding API.

Finally, I utilized the Foursquare API to gather location data within a five mile radius of each geocoded zipcode, thus capturing venues within a small neighborhood. I then calculated the frequency of venue categories and created a list of the top 10 venue categories for each sub area. Results were grouped to produce venue category frequencies by city, which were used for statistical testing and machine learning clustering algorithms.

2.2 Cleaning

State, capital city, zipcode and coordinate data were scraped from several websites, cleaned and combined into one dataframe. I decided to drop District of Columbia as this region was inconsistently referenced across sources.

As the absolute value of obesity rate can be difficult to interpret, I created another feature in the dataset to represent the rank of states in terms of obesity rate, where the value of 1 represented the most obese state.

With both the Mapquest geocoding and Foursquare location APIs, I batched requests into smaller chunks to bypass daily rate limiting. This was necessary in order to get the required data while utilizing a free-tier developer account.

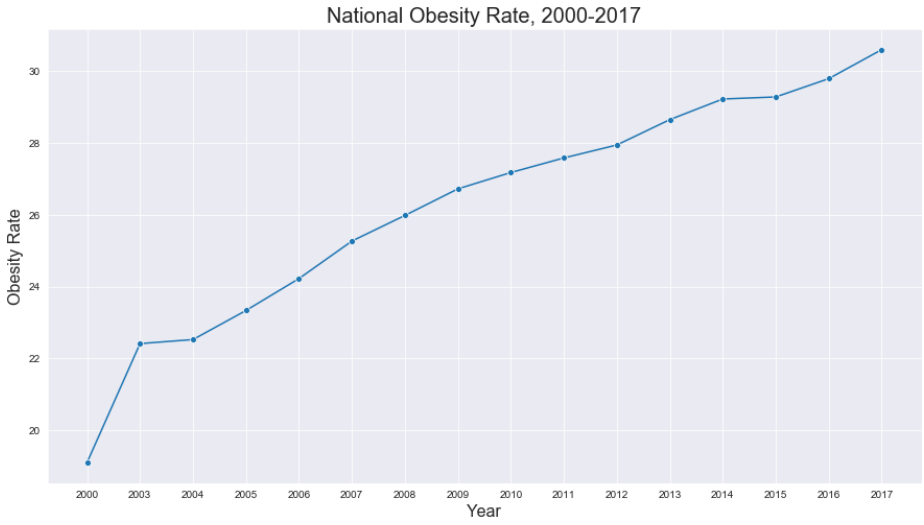
One hot encoding was used to process venue categories and ultimately create a venue category frequency dataframe for use with the K-Means clustering algorithm.

Following an initial round of clustering, I identified several states (Vermont, Maine and South Dakota) as potential outliers. I decided to exclude these states and re-run K-Means clustering.

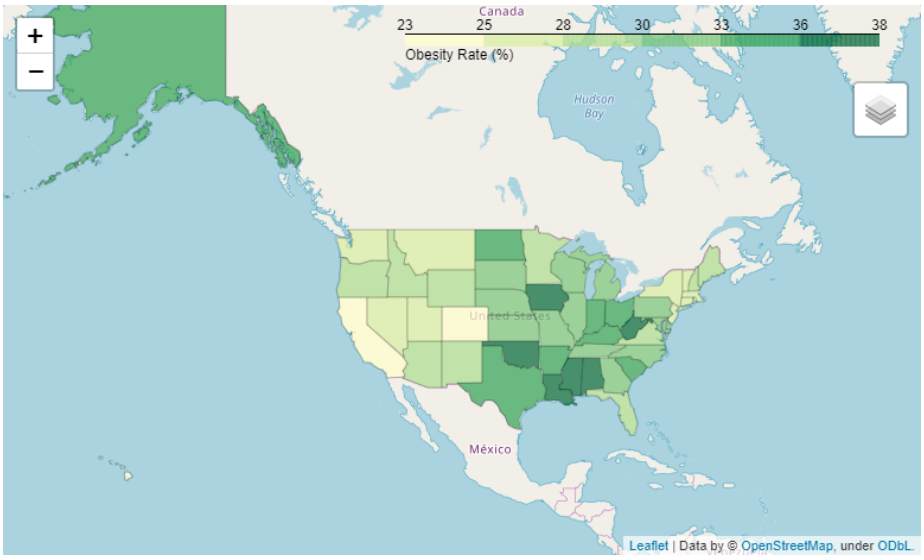
Finally, I selected a handful venue categories, including fast food restaurants, gyms, gym/fitness centers, trails, recreation centers, grocery stores, supermarkets and health food stores, and ran k-Means clustering using this subsection of data.

3 Methodology

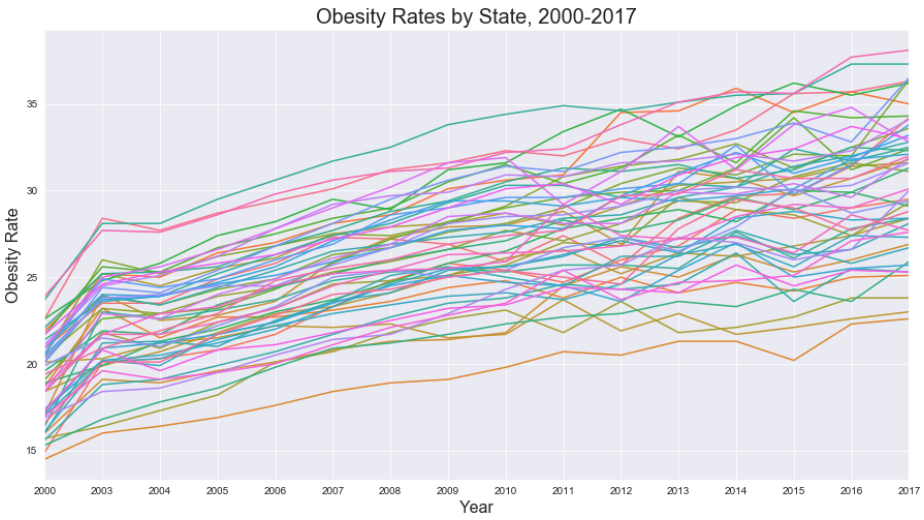
The goal of this study was to better understand obesity and investigate the potential linkage between obesity and environmental stimuli such as fast food restaurants. The first step was to visualize obesity rate data on a national level and to confirm that we are indeed seeing increasing rates of obesity across the US.



Obesity data was then used to create a choropleth map to visualize the variation in 2017 obesity rates between states.



Likewise, state level data indicates an increasing trend in obesity rates, confirming what was seen on a national level.



I then identified the most/least obese states by sorting the dataframe by obesity rate in descending order and viewing the top/bottom records.

Most Obese States

	State	Capital City	Abbreviation	2017 Obesity Rate	State Rank
0	West Virginia	Charleston	WV	38.1	1
1	Mississippi	Jackson	MS	37.3	2
2	Oklahoma	Oklahoma City	OK	36.5	3
3	Iowa	Des Moines	IA	36.4	4
4	Alabama	Montgomery	AL	36.3	5

Least Obese States

	State	Capital City	Abbreviation	2017 Obesity Rate	State Rank
45	Utah	Salt Lake City	UT	25.3	47
46	Montana	Helena	MT	25.3	48
47	California	Sacramento	CA	25.1	49
48	Hawaii	Honolulu	HI	23.8	50
49	Colorado	Denver	CO	22.6	51

I decided to focus my analysis on capital cities, breaking each city into its constituent zipcodes. Foursquare location data was used to determine the prevalence of different types of venues within a five mile radius of each zipcode, which I used as a proxy for the size of a small neighborhood. From these frequencies, I summarized the top 10 venue categories for each of the capital cities. This enabled me to make qualitative observations such as whether the category of fast food restaurant appeared in a city's top 10 list.

A central question in obesity research is whether there exists a correlation between certain environmental stimuli and obesity. In this study, I examined two specific environmental stimuli: fast food restaurants and gyms. The Pearson correlation coefficient was calculated to determine the extent to which these variables are linearly related to obesity rate. The Seaborn library was then used to generate regression plots to visualize the relationship between variables.

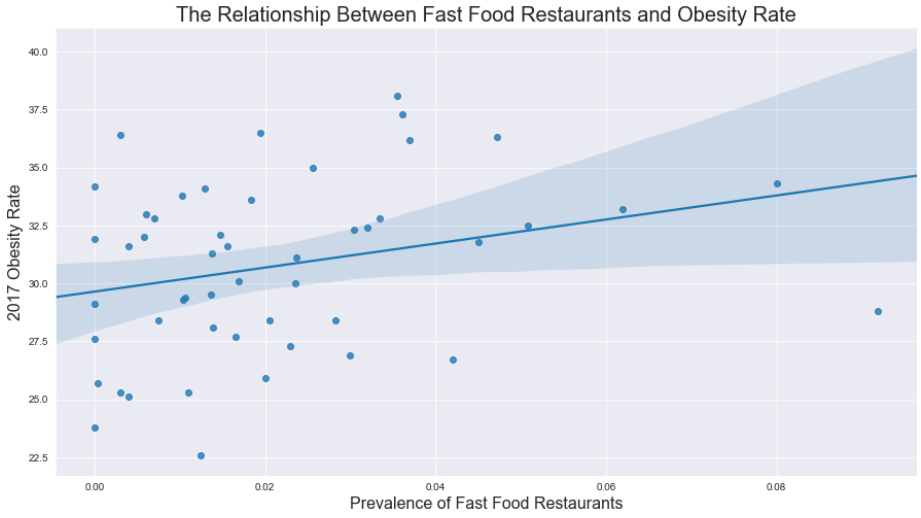
Since the Pearson correlation coefficients indicated only a weak positive relationship between variables, I also used a residuals plot to assess whether a non-linear model might be appropriate.

Finally, I used the Scikit-learn library to perform multiple rounds of clustering using the K-Means algorithm. The optimal number of clusters was determined using the elbow method, when appropriate. Clusters were viewed on a map and examined in terms of the average obesity rate and state rank within each cluster. In the final iteration of clustering, a subset of city features were used and resulting clusters were viewed in relation to a state obesity rate choropleth map.

4 Results

The purpose of this analysis was to investigate whether obesity rates are linked to environmental stimuli. After examining the contents of Foursquare's location data, I chose two environmental stimuli which are commonly associated with health, namely fast food restaurants and gyms/fitness centers.

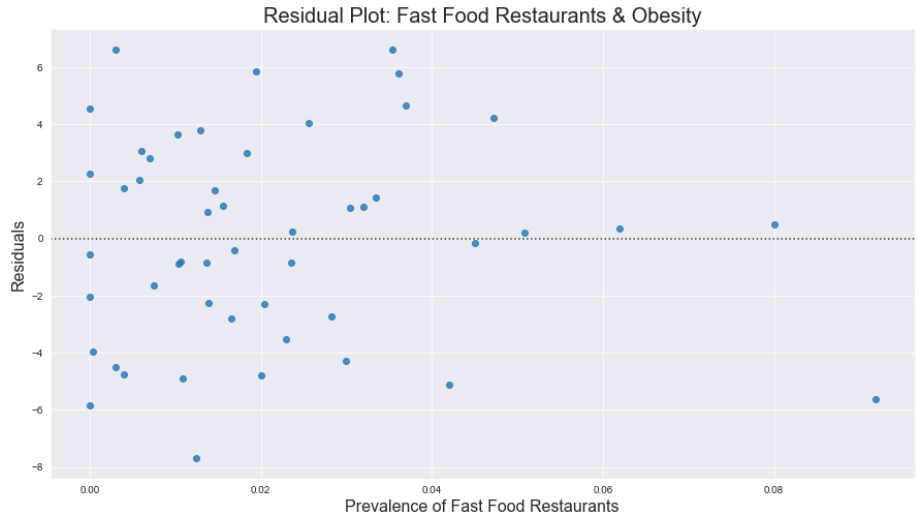
The prevalence (frequency) of the two selected environmental stimuli were plotted against the 2017 obesity rate for each state. These graphs indicate that there is a very weak positive relationship between the frequency of fast food restaurants and obesity. On the other hand, there appears to be no correlation between the frequency of gyms and obesity. These findings were confirmed by the Pearson correlation coefficients.

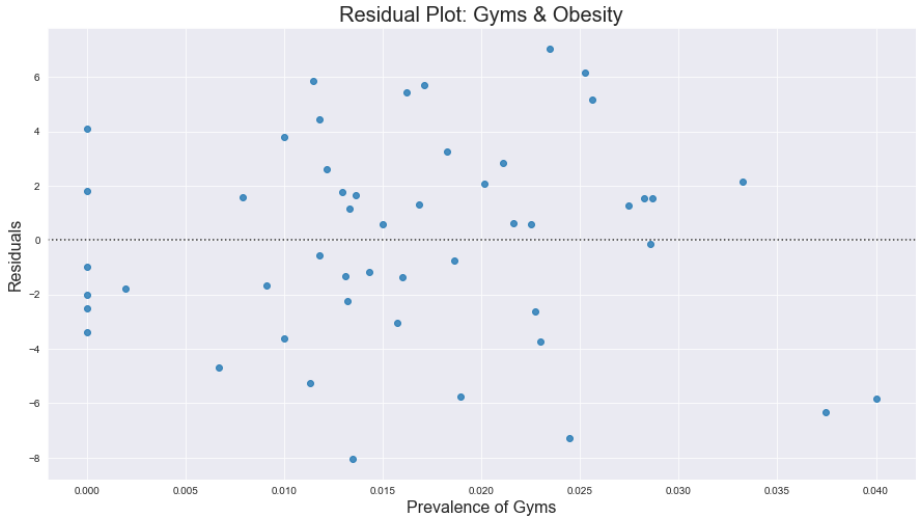


Pearson Correlation Coefficient Matrix

	2017 Obesity Rate	State Rank	Fast Food Restaurant	Gym + Fitness
2017 Obesity Rate	1.000000	-0.989262	0.277398	0.105826
State Rank	-0.989262	1.000000	-0.278843	-0.121166
Fast Food Restaurant	0.277398	-0.278843	1.000000	0.025494
Gym + Fitness	0.105826	-0.121166	0.025494	1.000000

I then created residual plots to determine whether there is a non-linear relationship between the selected environmental stimuli and obesity. Since residuals are evenly distributed around the horizontal axis, I concluded that a non-linear relationship does not exist between variables.





Next, I looked at the top 10 most and least obese states, counting the number of times "fast food restaurant" appeared in each city's top venue category list:

- "Fast food restaurant" appeared in the top 10 venue categories for six out of the 10 most obese states.
- "Fast food restaurant" appeared in the top 10 venue categories for two out of the 10 least obese states.

Lastly, I applied a K-Means clustering algorithm to various segments of the

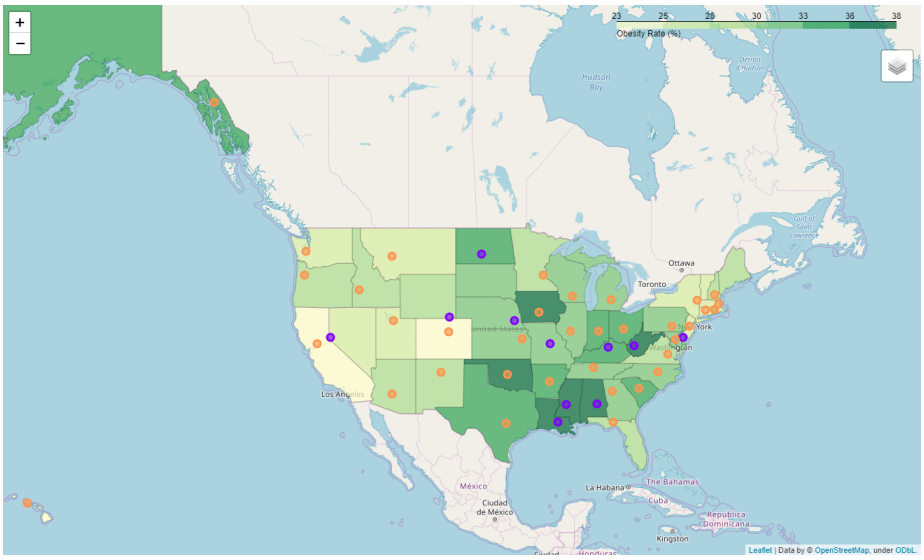
venue frequency data. After identifying and excluding three outlier states (Vermont, Maine and South Dakota), the algorithm produced two distinct clusters, however the difference in average obesity rate between clusters was negligible (6%).

	Cluster	Avg Obesity Rate	Avg State Rank	# States	States (Rank)
0	0	31.94	22	20	[Alabama (5), Connecticut (43), Delaware (23), Florida (36), Illinois (27), Kansas (18), Kentucky (8), Louisiana (6), Massachusetts (45), Michigan (19), Mississippi (2), Missouri (17), Nebraska (15), Nevada (44), New Jersey (42), North Dakota (13), Oklahoma (3), Rhode Island (30), West Virginia (1), Wyoming (35)]
1	1	30.01	29	27	[Alaska (9), Arizona (31), Arkansas (7), California (49), Colorado (52), Georgia (24), Hawaii (50), Idaho (33), Indiana (12), Iowa (4), Maryland (26), Minnesota (38), Montana (48), New Hampshire (39), New Mexico (37), New York (46), North Carolina (20), Ohio (11), Oregon (32), Pennsylvania (25), South Carolina (10), Tennessee (16), Texas (14), Utah (47), Virginia (29), Washington (40), Wisconsin (21)]

For the last round of clustering, I selected eight out of the total 488 unique venue categories on which to apply K-Means. Features included the frequency of fast food restaurants, gyms, gyms/fitness centers, trails, recreation centers, grocery stores, supermarkets and health food stores. This resulted in two clusters with average obesity rates that differed by 11.4%, an improvement over clustering using all of the features.

	Cluster	Avg Obesity Rate	Avg State Rank	# States	States (Rank)
0	0	33.45	15	11	[Alabama (5), Delaware (23), Kentucky (8), Louisiana (6), Mississippi (2), Missouri (17), Nebraska (15), Nevada (44), North Dakota (13), West Virginia (1), Wyoming (35)]
1	1	30.03	29	36	[Alaska (9), Arizona (31), Arkansas (7), California (49), Colorado (52), Connecticut (43), Florida (36), Georgia (24), Hawaii (50), Idaho (33), Illinois (27), Indiana (12), Iowa (4), Kansas (18), Maryland (26), Massachusetts (45), Michigan (19), Minnesota (38), Montana (48), New Hampshire (39), New Jersey (42), New Mexico (37), New York (46), North Carolina (20), Ohio (11), Oklahoma (3), Oregon (32), Pennsylvania (25), Rhode Island (30), South Carolina (10), Tennessee (16), Texas (14), Utah (47), Virginia (29), Washington (40), Wisconsin (21)]

Clusters were overlaid on a choropleth map to visualize whether obesity rate data is captured by the different clusters.



5 Discussion

My goal for this exploratory research project was to utilize data science and machine learning techniques and free third-party data sources to investigate obesity rates across the US. I began with a naive hypothesis that certain environmental stimuli, namely fast food restaurants and gyms, would be positively and negatively correlated, respectively, with obesity.

While my results indicate a weak positive relationship between the frequency of fast food restaurants and obesity, it is undoubtedly true that obesity is a complex societal phenomenon. Many different factors contribute to the decreased levels of physical activity and poor diet that play a role in the rising levels of obesity. Honing in on the important environmental features will enable more accurate classification of cities based on their composition. One goal might be to utilize machine learning to cluster cities into "High Risk", "Medium Risk" and "Low Risk" categories, thus enabling public policy makers and health organizations to address problematic regions by implementing targeted outreach or community programs.

On a qualitative level, this study illustrates one method for classifying the composition of cities using Foursquare location data. I have attempted to capture the concept of a neighborhood by limiting searches to a five mile radius around each zipcode. While this undoubtedly reduced the chance that locations were included more than once, I would suggest that further refinement is necessary to address overlap between defined subareas. API request limits aside, it would be interesting to sweep location data for entire cities, rather than

individual zipcodes, to perform statistical analysis on a larger scale.

6 Conclusion

In summary, this exploratory analysis, while simplistic, lays the groundwork for a data science-driven approach to assessing environmental risk factors. Although I have focused on obesity specifically, other population statistics such as lifespan and happiness could be incorporated to create an overall "health" metric for individual cities. Further research would be required to identify the collection of city features which are the best predictors for such a metric. Ultimately, improved knowledge of how we are shaped by our environment could be used to engineer surroundings that are optimized for our well-being.

References

[1] C. for Disease Control and Prevention, "National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity." `https://www.cdc.gov/nccdphp/dnpao/data-trends-maps/index.html`. [Online; accessed 06-July-2019].

[2] W. contributors, "Obesity in the United States. In Wikipedia, The Free Encyclopedia." `https://en.wikipedia.org/w/index.php?title=Obesity_in_the_United_States&oldid=904132800`. [Online; accessed 06-July-2019].

[3] R. Hammond and Levine, "The economic impact of obesity in the united states," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. Volume 3, p. 285–295, 2010.