Exploring the relation between obesity and types of venues by in selected counties in US,

Coursera Capstone,

C. Antonescu

Introduction:

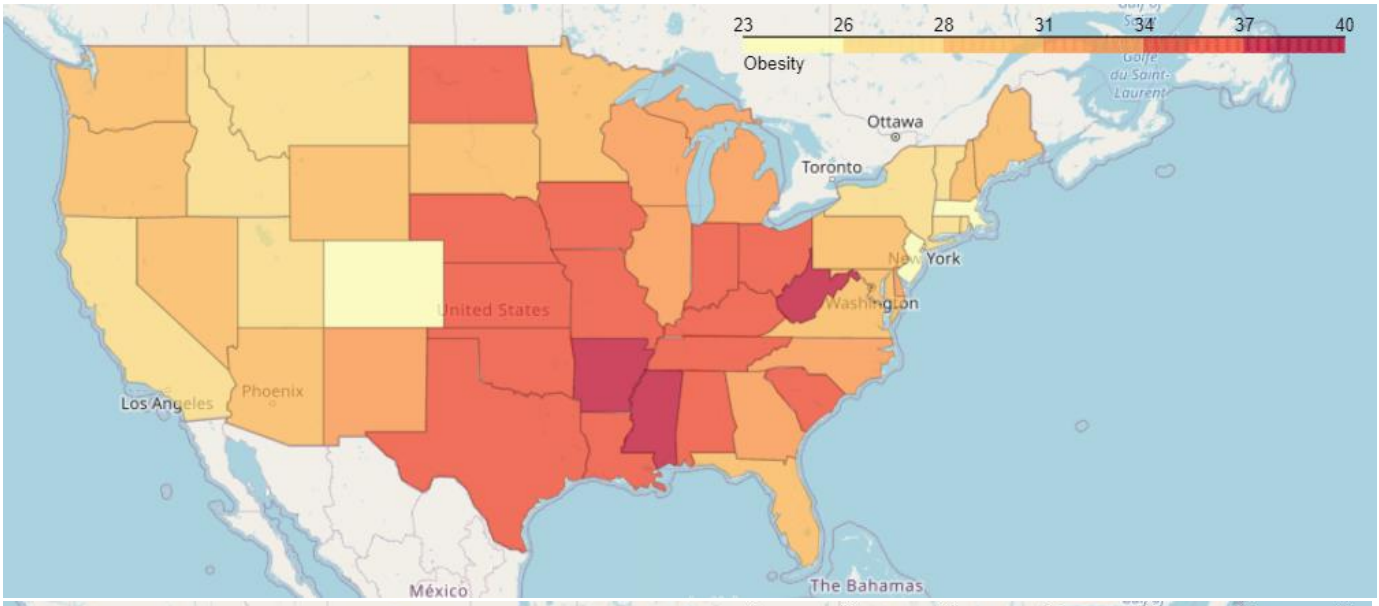
It is now common knowledge that obesity is a result of a few factors, like physical activity and nutrition. Government health agencies, like the Center for Disease Control (CDC) maintain a repository of data and maps showing the prevalence of obesity in different regions of US. They also keep data on risk factors for obesity, such as nutrition and physical activity and related to that, the proximity to places of physical activity. Other large private healthcare organizations also have published data on obesity and availability of physical activity, by county, and this data is available to the public for download. However, one exploration that seems to be missing, is the type of venues that are present, in areas of high incidence of obesity (and more limited availability of physical activity) versus areas where obesity is low and places where physical activity is easily available. In my project, I plan to explore the differences at regional level between the mix of venues, in areas of high versus low obesity. The purpose for this investigation of whether there are significant differences in the venue mix, is to understand whether the type of data that Foursquare API provides can be used in further research.

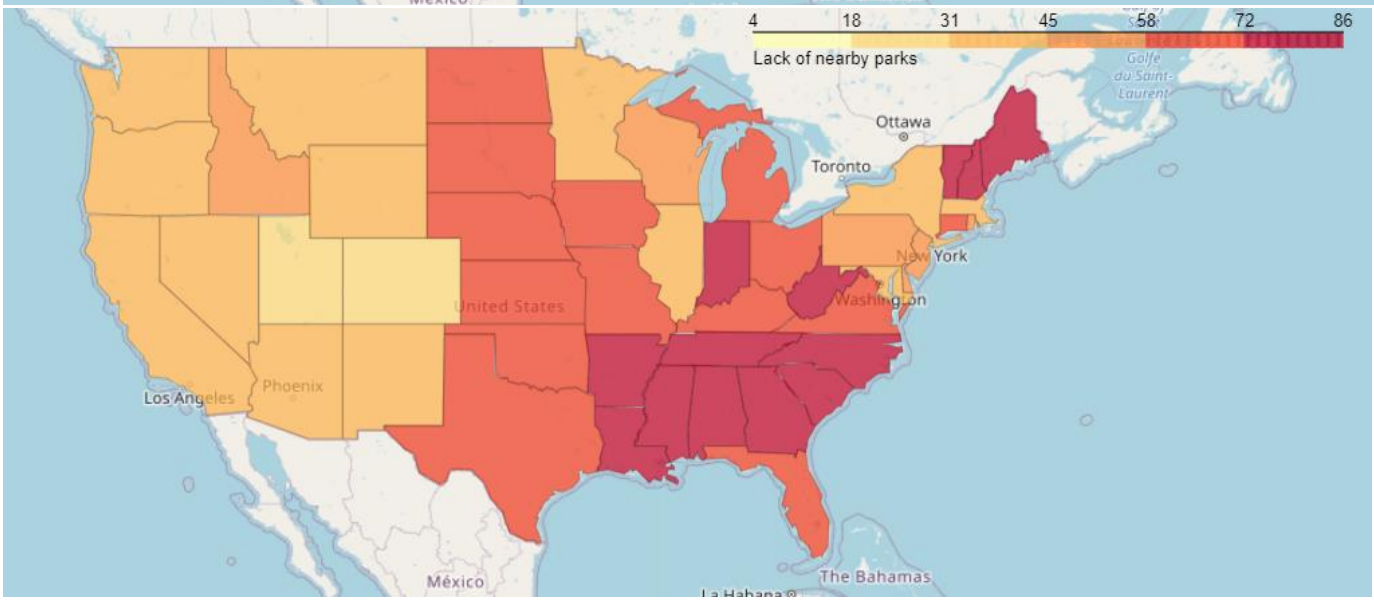
Data:

The obesity and obesity risk factors data that I will use comes from CDC and the Robert Wood Johnson foundation which are available to the public for download. The BR FSS data sets that are available on the CDC website are based on telephone surveys of more than 400,000 adult interviews each year. For the types of venues I will be using the Foursquare data which is available through the API. The Robert Wood Johnson foundation data comes from many different sources. I will also be using data for creating maps, data for matching counties (which is how healthcare organizations publish their data) with zip codes (which is how I will obtain data from the Foursquare API). Additionally, in order to allow avoiding looking at outliers and comparing very large with very small states, I will use data that shows the population in each state. Because of the sheer number of venues in each state, and the API limits for FourSquare, I am not going to be able to analyze full state data, but will focus instead on the comparing the top and bottom few counties from few states that have high versus low prevalence of obesity.

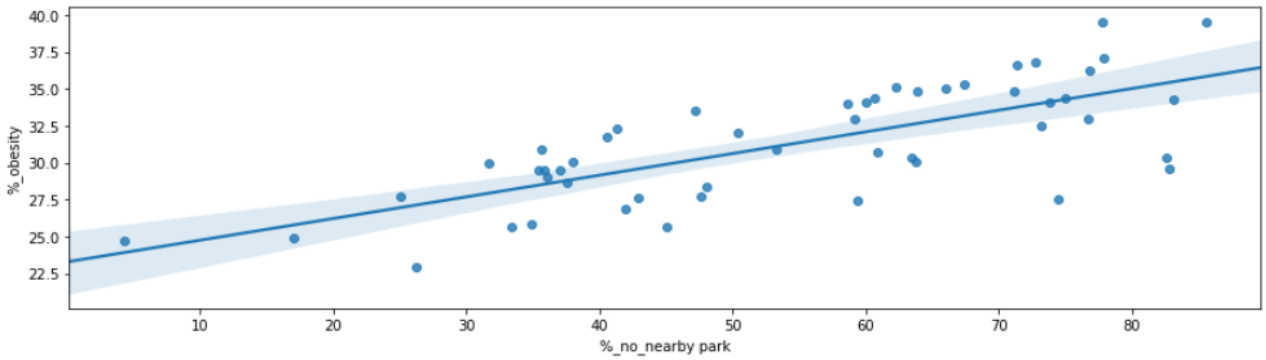
Results:

The two maps shown here outline the incidence of Obesity (in the first map) and the incidence of surveyed individuals who answered not having a park within ½ mile. Having a nearby park was viewed by the surveyors as being equivalent to having access to a place where physical activity can be performed.





While the correlation is not perfect, we can see that many of the darker areas are somewhat matched. Indeed, looking at the scatter plot below we can estimate a correlation, as shown by the regression line.



The limits of the data access in the FourSquare application implies that we won't be able to asses all venues in all states. So, we will be focus at the top and bottom states. We can however focus on the states where the identified lack of access to parks and obesity are found in ranked lists.

This is list of top and bottom states ranked in terms of obesity and access.

Park within 1/2 mile		
Abbr	State	% access
SC	South Carolina	16.9
AR	Arkansas	22.1
MS	Mississippi	22.2
AL	Alabama	23.2
TN	Tennessee	25
NV	Nevada	64.6
MA	Massachusetts	66.6
OR	Oregon	68.3
CO	Colorado	73.8
UT	Utah	74.9

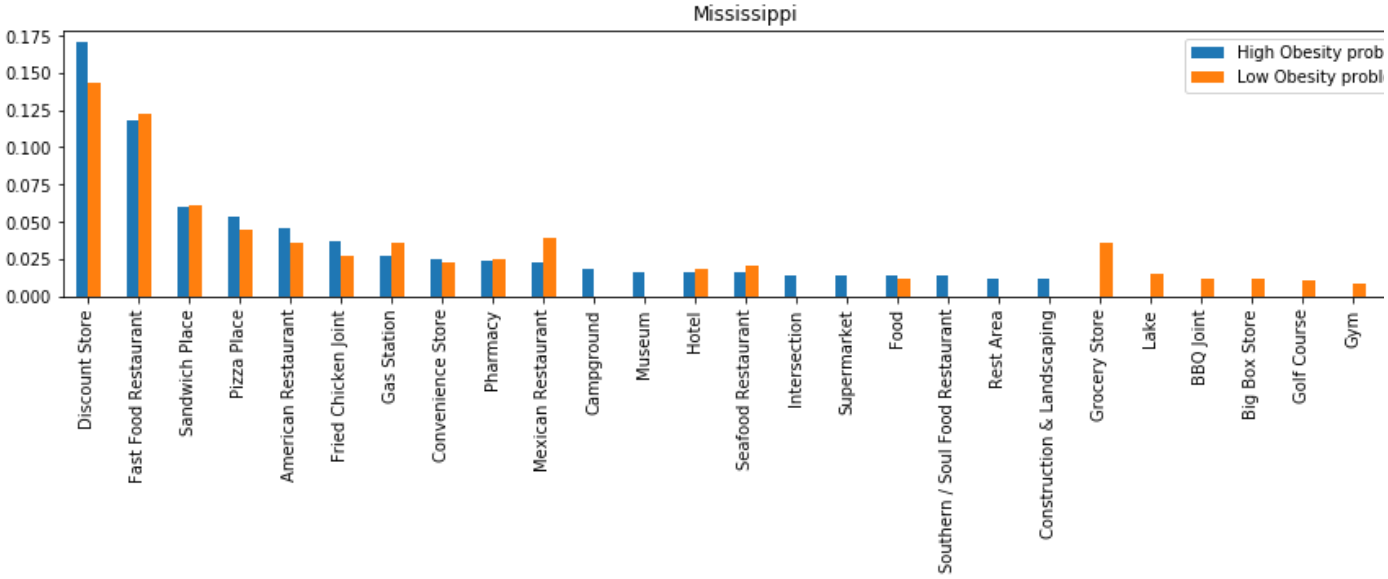
Obese (BMI 30.0 - 99.8)		
Abbr	State	% obese
MS	Mississippi	39.47
AR	Arkansas	37.12
LA	Louisiana	36.76
KY	Kentucky	36.62
AL	Alabama	36.24
CO	Colorado	22.95
MA	Massachusetts	25.69
CT	Connecticut	27.42
UT	Utah	27.75
ID	Idaho	28.36

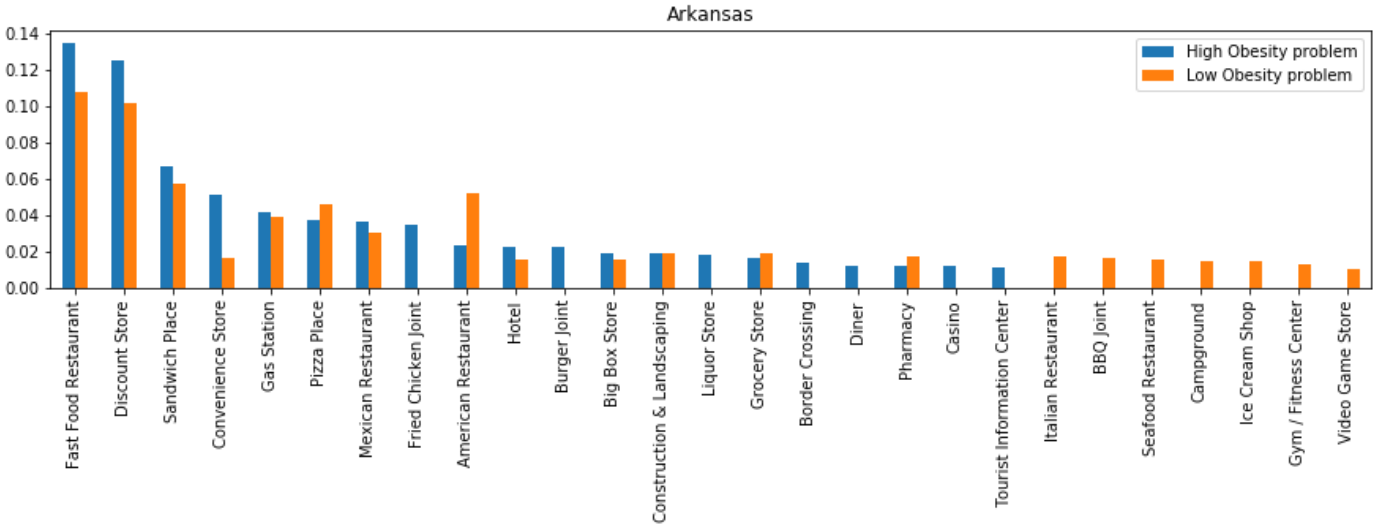
We will get compare a list of top frequent venues in the zip codes mapped to the top and bottom counties from the highlighted states : AR, MS, AL, CO, MA, UT.

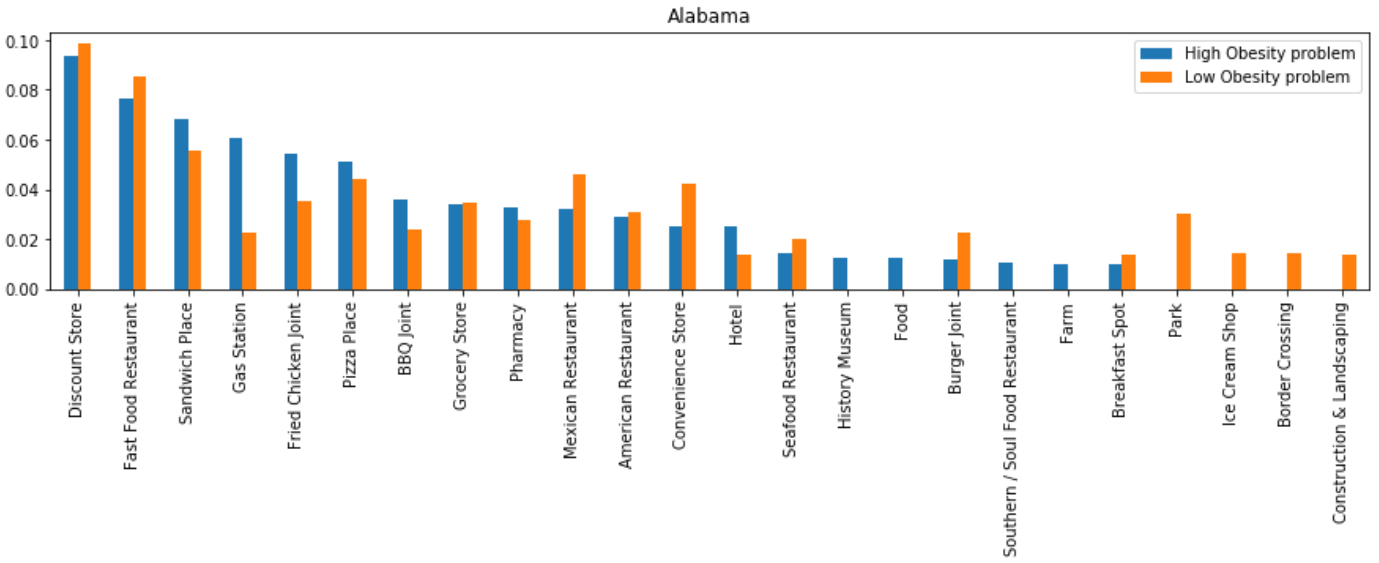
In order to get the most significant data we will be again limiting our analysis to only counties with population in the 25the to 75th percentile range. We will select the 5 counties with higher vs lowest obesity prevalence.

We will then compare the 20 top venue categories from the counties with the highest percentage of obesity vs the states with lowest prevalence of obesity

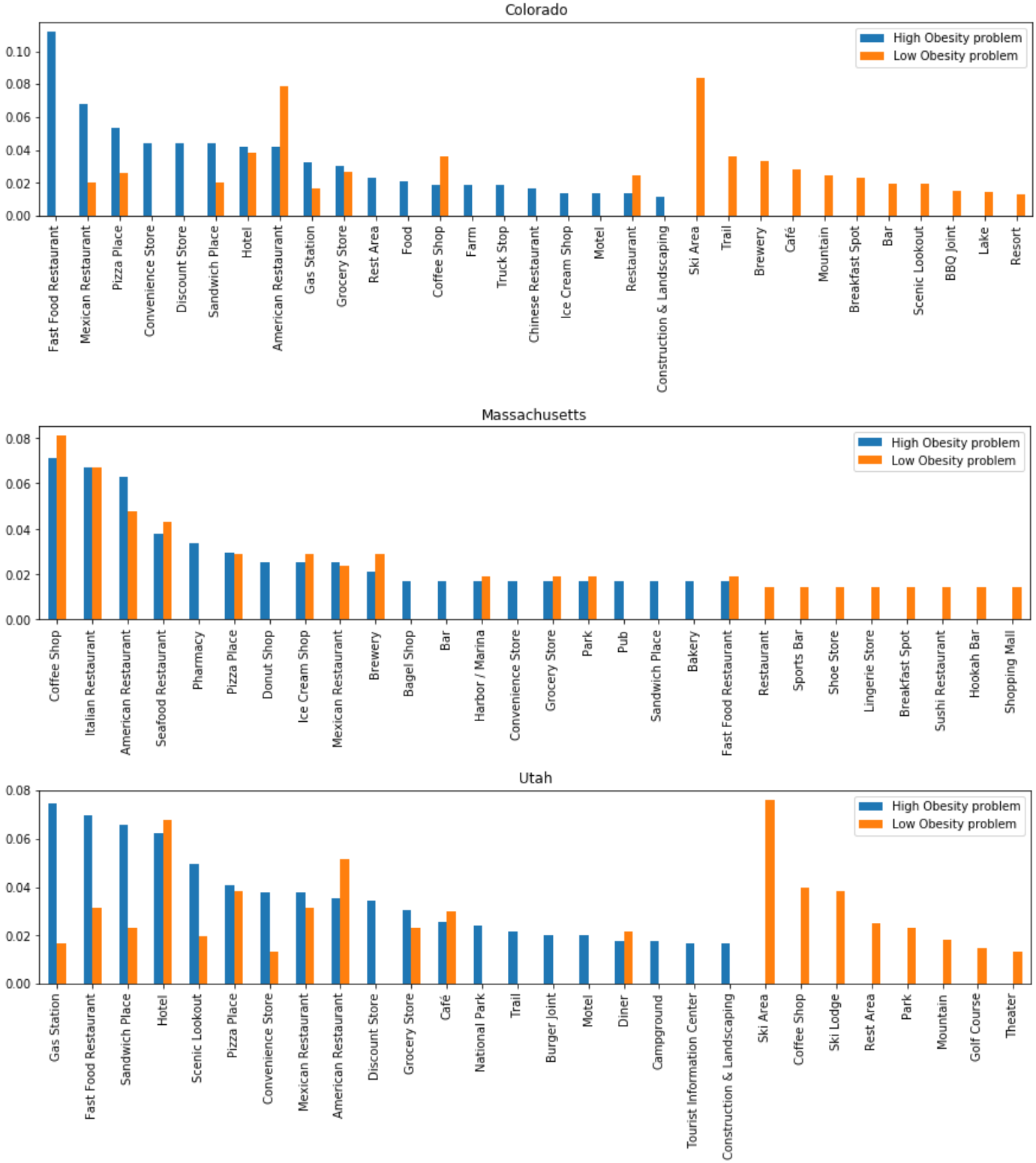
States with high prevalence of obesity







States with low prevalence of obesity



The results above show that in states were prevalence of obesity is high, there are a lot of discount stores, fast food restaurants, sandwich places, pizza places, and the counties with high prevalence of obesity vs low prevalence of obesity don't have a much different ranking between ranking between these.

States of UT and CO are somewhat different because of the opportunities for outdoor activities provided by the mountain areas. In these states fast food restaurants, sandwich places, discount stores, pizza places are associated with high obesity.

In Massachusetts however the number of fast food restaurants, pizza places and discount stores is quite low in general (ranking in the second half as frequency both in the high and low obesity areas) .

Limitations:

Some limitations of this study include the limited amount of data that was analyzed because of the API limits, however I am grateful to FourSquare, considering this data was provided for free. Other limitations include the fact that some of the data published by CDC was collected in different years and so, we have to assume that the mix of venues has not change much over time. Finally, another limitation is that I used the FourSquare API , however there are multiple providers of venue data currently available, and for a more reliable analysis one could consider using multiple sources of data.

Conclusion: Analysis of venue data provided by API's such as FourSquare can provide interesting insights into factors associated with Obesity.