Tackling the problem of Obesity in the United States

Abhishek Devarakonda, Abhinav Pathak, Luisa Chu, Nisanth Dheram, Sahil Gupta, Vaishnavi Jala

Introduction:

Obesity is common, serious and costly! A lot has been said and written about this condition which is one of the leading causes of heart disease, the biggest killer in the United States. According to the Centre for Disease Control and Prevention (CDC), more than one-third (34.9% or 78.6 million) of U.S. adults are obese. Approximately 17% (or 12.7 million) of children and adolescents aged 2—18 years are obese. Obesity-related conditions include heart disease, stroke, type 2 diabetes and certain types of cancer, some of the leading causes of preventable death. The estimated annual medical cost of obesity in the U.S. was $147 billion in 2008 U.S. dollars; the medical costs for people who are obese were $1,429 higher than those of normal weight. CDC's State Public Health Actions to Prevent and Control Diabetes, Heart Disease, Obesity and Associated Risk Factors and Promote School Health provided multiple initiatives to combat obesity for every state. Though these programs have gone a long way in taking on this problem head-on, we believe that a more targeted and efficient approach could be to initiate such programs at a county level by identifying clusters of similar counties.

The ultimate goal of our analysis is to identify clusters of counties which share a similar obesity and food environment profile which can then be targeted with programs to reduce obesity. To do this, we tried to understand the impact of food environment factors such as prevalence of fast food restaurants, nutrition stores, recreation facilities, poverty rate, etc. on obesity and identify which among these factors most significantly impact the obesity rate in each county. The dataset that we used was obtained from **data.gov** and provides information for the different food environment factors at a county level. We have analyzed the data for 2010 as this was year which had data for majority of the factors.

We analyzed five questions which provide a layout of the data and trends at a county as well as state level en route to identifying clusters of counties:

1. **What are the top and least 5 obese and diabetic states?**
2. **How does the rate of obesity vary for Metro and Non-Metro counties?**
3. **How do the different food environment factors collectively affect obesity and which among them have the most statistically significant impact?**
4. **How can we cluster and profile counties based on their obesity rates and the significant food environment factors identified above?**

Related Work:

Indiana Wesleyan University analyzed the factors explaining Obesity in the Midwest using multiple regression in 2015. The research analyzed 737 counties in Midwestern states of Indiana, Iowa, Illinois, Minnesota, Missouri, Ohio, and Wisconsin. We used county obesity information for the United States for our research. The university analysis was restricted to adult population with factors focusing on food,

recreation facilities and demographic factors. Our analysis explained how various stores like grocery stores, convenience stores, nutrition stores etc. in a county affect its obesity percentage. While their model accounted for only 33% of the variation in obesity, we were able to explain 48% variation in obesity. The analysis in Wesleyan University was confined to identify the factors affecting obesity, which was not very insightful. We have segmented counties into different obesity groups and analyzed their profiles and thereby providing a deeper understanding of the counties.

North Dakota State University of Agriculture and Applied Science has conducted a research in 2005 which primarily focused towards finding factors that contributed for adult obesity. The research focused on overnutrition with foods high in fat and sugar coupled with low physical activity attributes. The analysis also included socioeconomic factors like education and income and includes food and milk prices in grocery stores. The goal was to estimate differential price effects of addictive foods on normal, overweight and obese people. The work analyzed the prevalence of obesity by constructing a logit model on various states in US. The study was limited due to the unavailability of pricing data at a customer transaction level. In our analysis, we combined adult and child obesity and modelled overall obesity. The study analyzed adult obesity at state level whereas our analysis was at a county level. Since we analyzed relatively more information, the results obtained are more meaningful and useful at tackling the obesity issue.

Data cleaning and treatment:

The raw dataset (MS Excel file) we obtained from data.gov contained data for the different variables spread across different sheets of the file. There was one meta sheet which contained descriptions for all the variables along with their specific sheet locations. To consolidate these variables into one master dataset, we did the following:

- Created a datasheet descriptor file containing the variables which are of interest to us with the respective sheet names where they are located
- Iterated through each unique sheet as given in the descriptor file and obtain a list of all variables of interest available in that sheet
- Collected data for the variables obtained in each iteration and appended it to the master dataset

The main advantage of creating this process was that if we wanted to add a new variable to the master dataset, all we had to do was to add its description and location to the descriptor file and the code would take care of adding it to the master dataset thereby eliminating any manual intervention and reducing the possibility of errors.

We did not have data for some of the variables available for the year 2010, but data was available for other years before and after 2010 (2007,08,12 or 2013 in most cases). Such variables included the no. of recreation facilities, no. of fast food restaurants, no. of full service restaurants, no. of WIC stores, no. of SNAP stores, no. of grocery stores, no. of supercenters & club stores, no. of convenience stores and no. of specialized food centers. As a workaround, we used **linear interpolation** by taking the years of data available either side of 2010 and estimating the values for 2010.

Next, to account for differences in population among the different counties, we normalized the variables which varied based on county population size by converting them to values per thousand people.

Finally, obesity rate was available as a separate variable for adults and children. Since most of the variables were at an overall population level per county irrespective of age, we decided that finding obesity rate for the entire population would make for a much more accurate analysis. Hence, we obtained obesity rate for the entire population by consolidating the obesity rates for adults and children making use of their respective population sizes.

To ensure that missing data and null values would not affect our analysis, we discounted those counties which had data missing either for obesity rate or any of the variables that we considered for this analysis.

Data Description:

Our final master dataset has **3,143 observations** of the following variables that we considered at a county level (all variables for 2010):

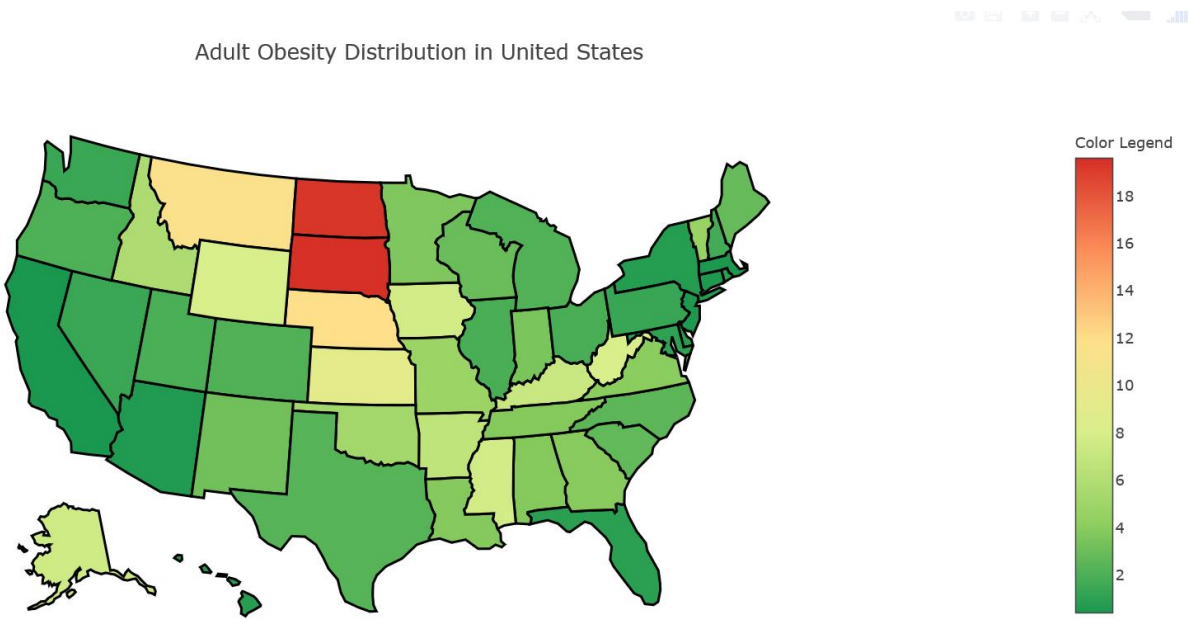
Variable	Description
PCT_OBESE10	% Obesity
PCT_DIABETES10	% Diabetes
FFRPTH10	Fast-Food restaurants per thousand people
FSRPTH10	Full service restaurants per thousand people
RECFACPTH10	No. of recreation facilities per thousand people
PCT_NHWHITE10	Percentage of people who are White
PCT_NHBLACK10	Percentage of people who are Black
PCT_HISP10	Percentage of people who are Hispanic
PCT_NHASIAN10	Percentage of people who are Asian
PCT_65OLDER10	Percentage of people aged 65 years or older
PCT_18YOUNGER10	Percentage of people aged 18 years or younger
POVRATE10	Poverty rate
MEDHHINC10	Median household income
SNAPSPTH10	SNAP-authorized stores per thousand people
WICSPTH10	WIC-authorized stores per thousand people
METRO13	Metro/Non metro county classification

PCT_LACCESS_HHNV10	% of population with low access to stores
PCT_LACCESS_POP10	% of households with no car & low access to store
GROCPTH10	No. of grocery stores per thousand people
SUPERPTHC10	No. of supercenters & club stores per thousand people
CONVSPTH10	No. of convenience stores per thousand people
SPECSPTH10	No. of specialized food stores per thousand people

**SNAP: Supplemental Nutrition Assistance Program*
**WIC: Women, Infants and Children*

Analysis and Results:

To start off we looked at a heat map of the distribution of *adult* obesity rate across all the states of the United states:



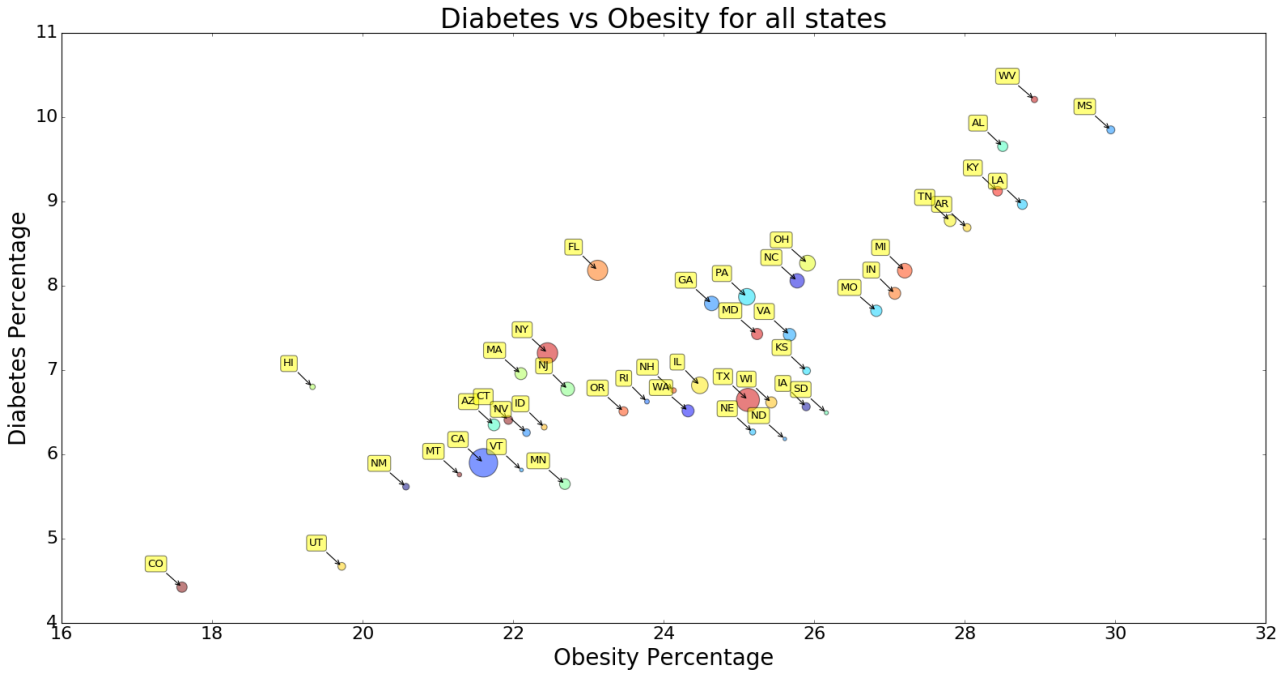
1. What are the top and least 5 obese and diabetic states?

We analyzed the trend of obesity and diabetes rates in all the 44 states which we had data available for to identify the most and least obese and diabetic states Following are the top 5 and least 5 obese and diabetic states as per our analysis:

Rank	Obesity	Diabetes
1	Mississippi	West Virginia
2	West Virginia	Mississippi
3	Louisiana	Alabama
4	Alabama	Kentucky
5	Kentucky	Louisiana

Rank	Obesity	Diabetes
44	Colorado	Colorado
43	Hawaii	Utah
42	Utah	New Mexico
41	New Mexico	Minnesota
40	Montana	Montana

The expectation was that the trend must be very similar as obesity and diabetes rates are likely to be correlated with each other given how obesity is a major cause of diabetes. We observe that this does indeed seem to be the case as many of the states appear commonly in both the lists of highest and least obese and diabetic states. We tried to further visualize the relationship using the bubble scatter plot shown below (bubble size proportional to population of the state):

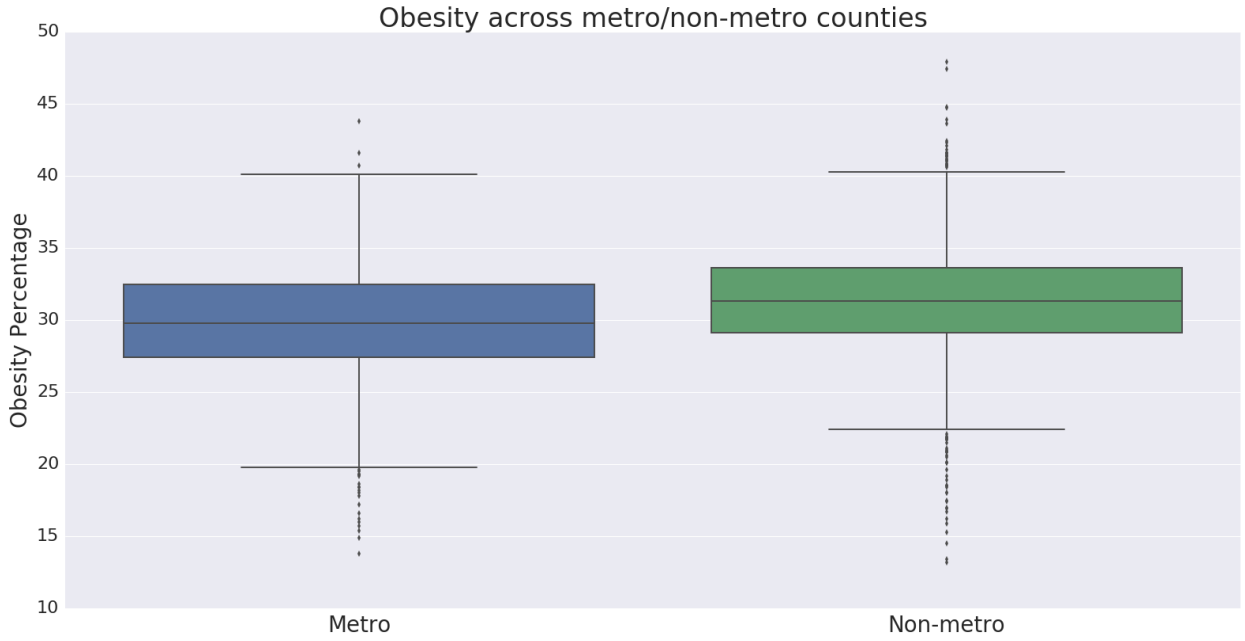


The chart clearly illustrates a high correlation between obesity and diabetes rates as expected. Thus an added bonus of trying to tackle the problem of obesity is a possible reduction in the diabetes rate as well.

As a side note, it is interesting to observe that Minnesota ranks among the least diabetic states and also among the lower obese states. Probably the large number of people we see running or cycling on the roads everyday goes some way in explaining why that seems to be the case!

2. How does the rate of obesity vary for Metro and Non-Metro counties?

We hypothesize that Metropolitan counties are likely to have a higher obesity rate given the expectation of higher number of fast food and full service restaurants among other factors. Interestingly, we observe that obesity rates are higher in non-metropolitan counties, albeit by a small margin. Following is a box and whisker plot showing the distribution of obesity rates across the two types:



We will try and explain the reasons behind this finding further down the line when we run a regression model to identify the most significant variables which impact obesity rate.

3. How do the different food environment factors collectively affect obesity and which among them have the most statistically significant impact?

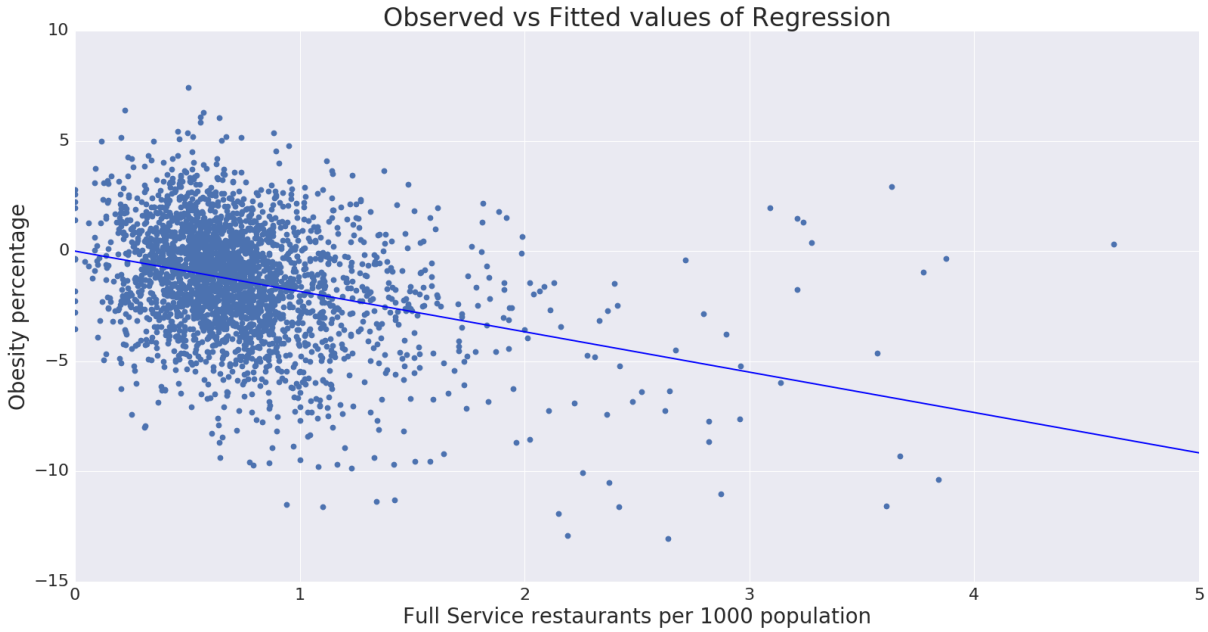
Given the set of variables we have at our disposal which might have an impact on the obesity rate, it is likely that not all of them will have a considerable impact and we need to understand which among them are statistically significant for explaining the obesity rate of each county. Gaining an understanding of the most significant factors can potentially help us better design programs targeted at reducing obesity and modify existing programs.

We used a **linear regression model** to analyze which factors are most significant in collectively explaining the obesity rate and their relative importance. The following table shows the list of variables which have a significant impact and those that are found to have a very low impact and are to be removed from the model:

Variable	Significant (Yes/No)
Poverty rate	No
Median household income	Yes
Percentage of people who are White	Yes
Percentage of people who are Black	Yes
Percentage of people who are Hispanic	Yes
Percentage of people who are Asian	Yes
Percentage of people aged 65 years or older	No
Percentage of people aged 18 years or younger	No
Metro/Non metro county classification	Yes
Fast-Food restaurants per thousand people	No
Full service restaurants per thousand people	Yes
No. of recreation facilities per thousand people	Yes
SNAP-authorized stores per thousand people	Yes
No. of grocery stores per thousand people	Yes
No. of supercenters & club stores per thousand people	Yes
No. of convenience stores per thousand people	No
No. of specialized food stores per thousand people	No
WIC-authorized stores per thousand people	No

The model takes into account the combined impact of all the variables together rather than just their individual correlations with obesity and some of the variables will not have a significant impact on obesity rate in the presence of other variables. Since we are interested in the combined impact off all the variables taken together, we will only consider the variables which we obtain from the regression model for our analysis.

A sample regression line of our model for one of the variables which shows the observed values of obesity rate vs the values obtained from the model is shown below:



Following is how the obesity rate varies with the most significant variables as obtained from our model:

- With an increase of one supercenter per 1000 people, obesity rate **increases by 7.3%**
- With an increase of one recreational facility per 1000 people, obesity rate **decreases by 4.6%**
- With an increase in one SNAP store per 1000 people, obesity rate **increases by 2%**
- With an increase of one full service restaurant per 1000 people, obesity rate **reduces by 1.8%**
- With an increase of one grocery store per 1000 people, obesity **decreases by 1.8%**
- Obesity rate **reduces by 0.2%** with a 1% increase in the number of Asians
- Metro counties are **0.2% less obese** than non-metro counties which is consistent with the result we obtained previously
- Obesity rate **reduces by 0.1%** with a 1% increase in the number of Hispanics
- With an increase of $1000 in Median household income, obesity rate decreases by **0.06%**
- Obesity rate **reduces by 0.02%** with a 1% increase in the number of Blacks
- Obesity rate **reduces by 0.02%** with a 1% increase in the number of Whites

Interestingly, we find that obesity rate seems to be increasing as the number of full service restaurants increases. With an increase of one restaurant per 1000 people, obesity rate **reduces by 1.8%.**

Another observation of note that we find from our model results is that out of the nutrition stores, which include SNAP & WIC stores, only the number of SNAP stores per 1000 people is found to have a

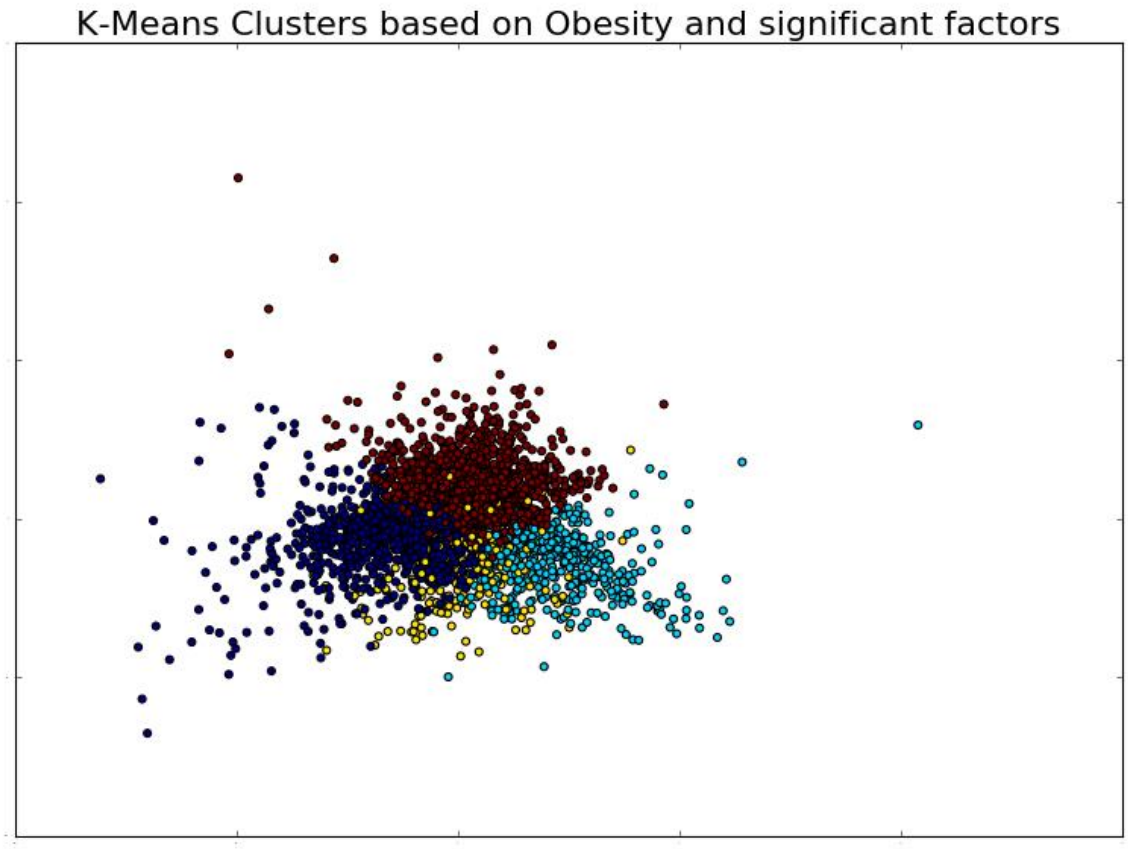
significant impact on obesity rate and it does seem to have an unwanted side effect of increasing the obesity rate. With an increase in one SNAP store per 1000 people, obesity rate is found to **increase by 2%.**

We then went on to leverage this model to predict the missing values for counties which did not have data available for all variables, to make our analysis more robust.

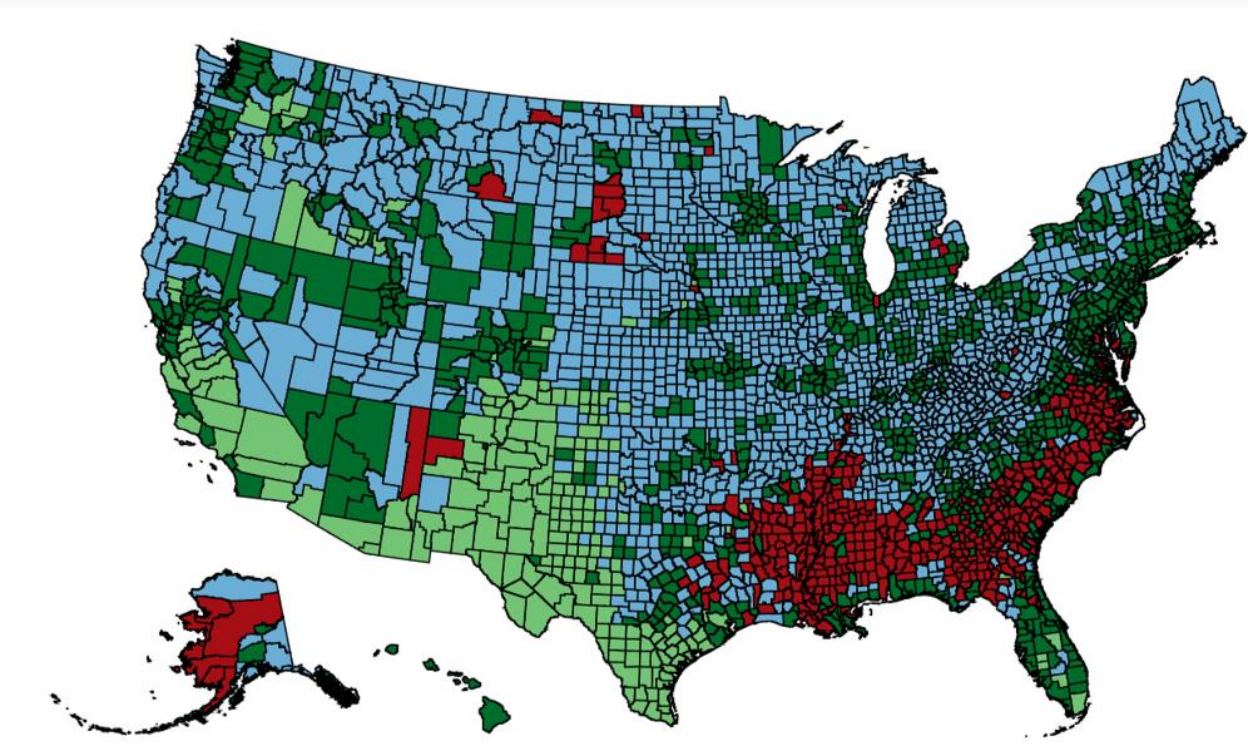
4. **How can we cluster and profile counties based on their obesity rates and the significant food environment factors identified above?**

Having identified the most significant factors which have an impact on obesity rate, we then try to form segments of counties which have a similar profile based on these variables. We used **K-means clustering** to cluster similar counties together. K-means uses Euclidian distance method to calculate the centroids of the cluster. So we normalized all the variables to bring them to a similar scale to ensure that no one variable biases the Euclidian distance calculation resulting in inaccurate clusters. Normalization was done using z-score transformation to transform the variables to a normal distribution. We tried different combinations of clusters to finally fixate on 4 clusters which was giving the best result.

We used Principle Component Analysis (PCA) algorithm to visualize the 12 dimensional clusters in a 2D graph. Following is the distribution of clusters that were obtained:



We also plotted the counties on the United States map segregating them according to the clusters they belonged to. Vincent Following is the map along with the counties with the counties colored differently based on the clusters 1-4 as indicated in the legend:



**Color varies from dark green (Cluster 1) to dark green (Cluster 4)*

The characteristics of each cluster as described by the means for each variable are as follows:

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Obesity rate	24.74%	24.79%	27.10%	30.10%
Full service restaurants per thousand people	0.75	0.77	0.91	0.48
No. of recreation facilities per thousand people	0.04	0.11	0.07	0.04
Percentage of people who are White	43.77%	79.69%	89.87%	52.33%
Percentage of people who are Black	3.39%	7.27%	2.16%	36.90%
Percentage of people who are Hispanic	49.07%	7.71%	4.27%	4.16%
Percentage of people who are Asian	1.42%	2.44%	0.47%	0.61%
Median household income	$39,405.69	$53,248.54	$40,122.33	$34,548.65

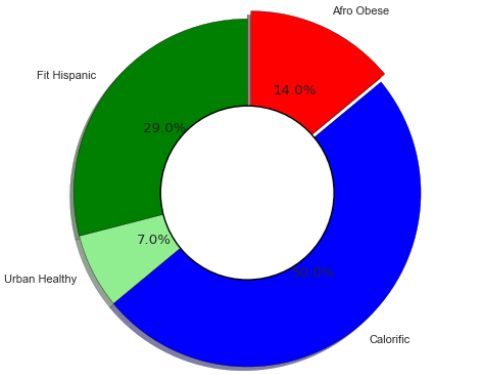
SNAP-authorized stores per thousand people	0.82	0.55	0.82	1.06
No. of grocery stores per thousand people	0.24	0.18	0.32	0.24
No. of supercenters & club stores per thousand people	0.01	0.02	0.02	0.02

We defined each cluster by observing their most dominant characteristics and labeled each of them accordingly:

Cluster 1: 'Fit Hispanic' counties

These counties are mainly characterized by a lower than average obesity rate - the lowest obesity rate among all the 4 clusters. Surprisingly, given that they have the least obesity rate, the number of recreational facilities per 1000 people is also one of the least among the clusters. Another characteristic is the extremely high percentage of Hispanics in comparison to the other cluster which is close to 50%. These counties also have the least number of supercenters & club stores per 1000 people.

Distribution of Counties across clusters



Cluster 2: 'Urban Healthy' counties

These counties are dominantly characterized by a high median household income which is the highest among all 4 clusters. This could be due to the fact this cluster has the highest percentage of metro counties. Also, the number of recreational facilities per 1000 is the highest among all the clusters. These counties have a lower than average obesity rate.

Cluster 3: 'Calorific' counties

These counties have high availability of full service restaurants and grocery stores per 1000 people among all the four counties. They have a higher than average obesity rate.

Cluster 4: 'Afro Obese' counties

These counties have a predominant Afro American population and also the highest obesity rate among the 4 clusters. The median household income is the least and as a possible result of that the number of full service restaurants per 1000 people is also the least among all the 4 clusters. These counties though, have the highest number of SNAP-authorized stores per 1000 people.

Conclusion:

This paper presents the health, demographic and socioeconomic factors which significantly explain obesity variation in different counties in United States for 2010. Multiple linear regression has been used to identify the factors which collectively affect obesity. While full service restaurants and recreational facilities are increasing obesity, nutrition stores like SNAP stores are negatively affecting obesity. Obesity is decreasing with increase in grocery stores where as it is reducing with supercenters. Since there are

more than 3000 counties, K-Means clustering has been used to segment clusters based on obesity and factors obtained from regression. The four clusters obtained from K-Means are Fit Hispanic, Urban Healthy, Calorific and Afro Obese. Fit Hispanic counties have lower obesity percentage with higher percentage of Hispanic population. Urban Healthy counties have more proportion of metro cities with lower average obesity percentage. Calorific counties have relatively higher obesity percentages with more full service restaurants and grocery stores. Afro Obese have the highest obesity percentages and also a significant proportion of Afro Americans. Based on the clusters obtained, we were able to identify the major factors responsible for obesity within these clusters.

Since obesity is a serious issue in the United States, there is a scope for additional research. This analysis can be further developed by introducing food and price variables like diet schedule, price and tax of food/drinks in stores and vending machines.

References

1. Matti, Josh and Kim, Hansol (2013) "Factors Explaining Obesity in the Midwest: Evidence from Data," Undergraduate. *Economic Review: Vol. 10: Iss. 1, Article 1.*

2. Economic Factors Affecting the Increase in Obesity in the United States: Differential Response to Price. *Major Professors: Dr. Dragan Miljkovic and Dr. William Nganje.*