ANLY699: Applied Project in Analytics

Impact of public parks in the neighborhood on health outcomes

Project Report

Subhash Pemmaraju
7-26-2020

Contents

INTRODUCTION

Background and Significance

Public recreational infrastructure investment is some of the hardest investment to justify. It involves significant expenditure of taxpayer dollars with no direct tangible benefits unlike investments in infrastructure like roads, hospitals, schools etc. These investments have a clear economic impact as a healthy, well-educated society with access to markets far and wide leads to faster economic growth. With public recreational infrastructure, the link is not as clear, although it is intuitive that using them can improve your health and well-being. Recreational infrastructure in this context refers to public parks. While the benefits of having public parks in the neighbourhood are clear to most people – opportunities for relaxation and exercise, there have been meagre attempts to relate access to public parks to positive health outcomes. While there is extensive literature on how public infrastructure like roads, schools etc. improve economic and educational outcomes and how access to public recreational facilities improves physical activity in the neighbourhood, there is very little research on how public parks can lead to positive health outcomes. This paper is an attempt at filling that void. The question that this paper attempts to address is the influence of public parks access on health. The significance of this question has a lot to do with the policy implications. Evidence based research pointing to a positive correlation between these variables can influence policy decisions in the years to come.

REVIEW OF LITERATURE

Existing literature about public parks and impact on health outcomes is very limited. Bulk of the research is around the short-term relationship between public recreational facilities and physical activity. Most studies start off with the assumption that physical activity leads to better health outcomes and then try to arrive at a positive correlation between different measures of public recreational facility usage and measures of physical activity. Literature is scarce on a direct link between ease of public park usage and health indicators. This study is intended to test the hypothesis that ease of access to public parks should, in the long term, result in positive health outcomes for the community as individuals improve their physical activity and improve their health. As such, there should be a positive correlation between measures of access to public parks and health indicators. Below is a summary of existing literature on this subject.

Stewart et al. (2018)[1] explore how presence of parks in the home neighbourhood contribute to physical activity. They examine data of 635 Seattle area residents and data from physical activity trackers and other instruments using a mixed effects negative binomial regression model. They find that every additional park in the neighbourhood leads to a 9% increase in physical activity.

Cohen et al. (2013)[2] conduct a randomized control trial (RCT) of 51 parks/recreational facilities in Los Angeles to examine the impact of community engagement in park development on physical activity. They find a statistically significant increase in the level of physical activity in the groups where the community was actively engaged in development decisions.

Floyd et al. (2011)[3] study the relationship between park use and park-based physical activity among children. They use SOPARC (System for observing play and recreation in communities) direct observation technique to observe physical activity in Durham, NC. They find strong evidence of increased physical activity among children with more access to parks. They also find that formal structured park activities encourage sustained physical activity among pre-school children.

Cohen et al. (2009)[4] conducted a descriptive study on why some parks are more used than others. They studied a sample of parks from South California serving around 4 million residents. They use a combination of surveys and direct observations and conclude that perception of safety and a lack of park specific attractions involving physical activity are major deterrents of active usage.

Scott et al. (2007)[5] study the impact of access to recreational facilities, including public parks as predictors of physical activity among adolescent girls. Using self-reported data of the 1,367 participants in the study along with measures of physical activity measured using activity trackers, they find that the number of facilities within half mile of girls' homes are strong indicators of perception of ease of access. However, they fail to find any strong correlation between ease of access and greater physical activity except for basketball facilities. There they find that every additional court results in 3% more physical activity.

Rung, Mowen and Cohen (2005)[6] develop a conceptual model of how to think about the impact of parks on public health. They explore the following categories of benefits from park usage – physical health, mental health, social benefits through cohesion and integration, economic

benefits and environmental benefits. They also discuss theoretical understanding of the barriers to effective park usage including ease of access, aesthetics, safety, maintenance etc.

As can be seen from the papers summarized above, bulk of the literature on this subject is based on surveys/trials in specific neighbourhoods within the US and are targeted at specific sections of the population like children or carried out in a different context. This paper aims to look at nationwide data and attempts to tease out the relationship between access to public parks and health outcomes.

RESEARCH OBJECTIVE

Research Question/Hypothesis

The specific research question that this study hopes to address is the following: “Does having more public parks in the neighbourhood improve health indicators of the population living in those neighbourhoods?”

This is a very relevant and important question. In the modern world, avenues for physical activity become very important as most white-collar jobs are very sedentary in nature. Public infrastructure that facilitates physical activity involves a significant investment on the part of the local government and has consequences in terms of higher taxes on the residents to finance these investments. Given that public parks are free for everyone to access, quantifying the benefit to the community becomes difficult and, an inability to quantify it makes the case for spending that money that much harder to build. This study can add to the literature by providing a quantifiable, tangible link between access to public parks and improved physical health. This will help policy makers build the case for an investment into public infrastructure like parks.

Correlational research can be used to answer this question. If a direct positive correlation can be found between prevalence of public parks and positive health outcomes, then it supports the hypothesis. It can also involve descriptive elements when comparing health outcomes between two different neighborhoods with varying demographics.

Overall Objective

The overall objective of this study is to find evidence of positive health outcomes driven by easier access to recreational facilities – particularly public parks.

Specific Aims

1. To find positive correlation between ease of access to public parks and indicators of health
2. To ensure that the finding is robust and holds across various segments of the data – demographic, geographic, varying health indicators
3. To ensure that the results are directionally consistent and statistically significant
4. To quantify the impact of how much improvement in specific health indicators would there be for every increment in the measure for ease of access

METHOD AND DESIGN

Overview

The research problem is going to be evaluated by first gathering the data from different sources and identifying the key variables for the analysis. The next step is to conduct cluster analysis and regression analysis to arrive at the relationship between these variables and how they interact and finally, report the findings. The detailed process is explained below.

Sources of Data

There are two main sources of data for this research paper. The first set of data has to do with "Percentage of population that resides within half mile of a park"[7]. This is a US county level dataset and is the preferred measure for ease of access to parks. The source of this data is the Center for Disease Control and Prevention, National Environmental Public Health Tracking Network. The second set of data has to do with indicators of health and associated demographic variables[8]. This data comes from a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute. They have compiled this dataset from a wide variety of sources including the American Community Survey, Bureau of Labor Statistics, American Medical Association, Centers for Disease Control and Prevention and other survey data.

Data Management

For the purpose of this analysis, the variable of preference to measure ease of access to parks is "% of population that resides within half mile of a park", which is one single dataset with information for all US counties.

For health indicators, we have a dataset corresponding to each state within the US. The dataset has the following set of information at each individual county level:

1. Health Outcomes – 4 different variables covering length of life and quality of life
    a. Yrs_plr: "Years of potential life lost before age 75 per 100,000 population (age-adjusted)" [9]
    b. perc_fair_poor_health[9]: "Percentage of adults reporting fair or poor health (age-adjusted)."
    c. avg_phy_unh_days[9]: "Average number of physically unhealthy days reported in past 30 days (age-adjusted)."
    d. avg_mental_unh_days[9]: "Average number of mentally unhealthy days reported in past 30 days (age-adjusted)."
2. Health Behaviours – 5 different variables covering topics like use of addictive substances, exercise, and diet related variables
    a. perc_smokers[9]: "Percentage of adults who are current smokers."
    b. perc_obese[9]: "Percentage of the adult population (age 20 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m2."
    c. food_env_ind[9]: "Index of factors that contribute to a healthy food environment, from 0 (worst) to 10 (best)."
    d. perc_phy_inact[9]: "Percentage of adults age 20 and over reporting no leisure-time physical activity."
    e. perc_excess_drink[9]: "Percentage of adults reporting binge or heavy drinking."
3. Clinical Care – 3 different variables covering access to healthcare and quality of care provided
    a. perc_uninsured[9]: "Percentage of population under age 65 without health insurance."
    b. physician_ratio[9]: "Ratio of population to primary care physicians."
    c. preventable_hosp_rate[9]: "Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees."

4. Social and environmental variables – 7 different variables covering topics like education, employment, community support and safety
    a. hs_grad_rate[9]: “Percentage of ninth-grade cohort that graduates in four years.”
    b. perc_college[9]: “Percentage of adults ages 25-44 with some post-secondary education.”
    c. perc_unemp[9]: “Percentage of population ages 16 and older unemployed but seeking work.”
    d. perc_child_poverty[9]: “Percentage of people under age 18 in poverty.”
    e. 80_20_inc_ratio[9]: “Ratio of household income at the 80th percentile to income at the 20th percentile.”
    f. perc_single_parent[9]: “Percentage of children that live in a household headed by single parent.”
    g. median_inc[9]: “The income where half of households in a county earn more and half of households earn less.”
5. Demographic variables – 5 variables covering Age, gender, race etc.
    a. perc_lt_18[9]: “Percentage of population below 18 years of age.”
    b. perc_gt_65[9]: “Percentage of population ages 65 and older.”
    c. perc_black[9]: “Percentage of population that is non-Hispanic Black or African American.”
    d. perc_female[9]: “Percentage of population that is female.”
    e. perc_rural[9]: “Percentage of population living in a rural area.”
6. Additional variables – We look at additional variables that do not belong to above categories such as % with food insecurity[9] defined as “Percentage of population who lack adequate access to food”.

This dataset is vast and comprehensive and the first step in managing this data needs to be to reduce it to a manageable dataset with only the key variables of importance for this study. The way that will be done as a part of this research is to look at each individual variable within the broad segments identified above and address the following key questions:

- Does the variable represent best a measure of health that would be impacted by more physical activity?
- Does a correlation plot between this variable and the independent variable (% of population within half mile of a park) support that there is a relationship?

- Does the variable have enough data with minimal missing information and does the data looks biased or skewed in any way? This will be verified by tracking % of missing data and plotting histograms for distribution of the data.

Similar analysis is carried out for other independent variables that will be included in addition to the core ease of access variables. The intent behind using these other independent variables is to control for the effects of population differences. The socio-economic variables, environmental factors, demographics are used to control for the differences.

The final step in the data management process is to combine all the key variables from above for all 50 states and 3,000 odd counties into one comprehensive dataset.

Data Analysis Strategies

There are two ways to analyse this dataset to address the problem. The first way is to conduct a nationwide study and examine all the counties at once and include all the independent variables at once in one big linear regression model. The challenge with this approach is that there may be too many independent variables and two wide a difference between each record in the dataset to provide any meaningful insight into the data. It may even result in no statistically significant correlation between the independent and dependent variables.

The second approach and the approach of this paper would be to divide the US county level data into 3 or more clusters. The clusters would be decided based on demographic and environmental variables that are more likely to influence health factors. This method of cluster analysis of demographic factors is supported by literature. Wallace et al (2019)[10] apply k-means clustering analysis to US county level demographic data and divide it into 8 clusters. They find that prevalence of obesity was highest in semi-urban, low to middle income clusters and lowest in young, urban, middle to high income clusters. They also find that smoking prevalence was lowest in the young, urban, middle to high income clusters. Similarly, Rayward et al (2017)[11] examine physical activity and sleep quality among a group of participants from Australia. They conduct cluster analysis and divide the participants into 4 clusters, influenced by socioeconomic variables as well as gender, age and education. The reason behind conducting such cluster analysis is that certain health indicators/outcomes are more influenced by demographic factors. For instance, Scheinker, Valencia, Rodriguez (2019)[12] look at US county level data and find that demographic factors explain 45% of the variation in obesity prevalence and socioeconomic factors explain 33%. So, the idea is to be able to segment the 3,000 odd US counties into several clusters with similar

demographic attributes. This can be verified by plotting distributions of the attributes overall and at each cluster level to quantify within cluster and across cluster differences.

This will be done as follows:

1. Gather all the socio-economic and demographic variables - % with college education, % unemployed, median income, % below 18 years of age, % above 65 years of age, % females, % African American
2. We first apply principal component analysis to reduce these dimensions into a smaller manageable set of dimensions
3. We identify which of the principal components contribute to explaining more than 80% of the variance; these variables become the final list of PCs for subsequent analysis
4. Secondly, the principal component analysis is followed is K-means clustering; scale/normalize each of the PCs in the list above
5. Calculate the distance matrix and plot it to get a visual sense of the distance matrix
6. Apply the K-means clustering algorithm on the normalized dataset starting with 2 clusters
7. Repeat the above for 3/4/5 clusters and for each of these, view the cluster plot to determine how these clusters behave
8. Use the elbow method to determine the optimal number of clusters for this paper
9. The result of this process is the segmentation of the county level data into a few clusters with similar socio-economic and demographic characteristics

The next step in the analysis to analyse the variables within each cluster. This would include histograms to capture the distribution of the variables, box plots to understand population differences. This step is crucial in really being able to understand the data and answer questions like

- what is the median age of this cluster?
- What is the median income/distribution of income?
- Is there over-representation of a gender or racial group within this cluster?

The next step is to examine variables that have not made it to the cluster analysis but could still influence health indicators beyond ease of access to parks. For example, ease of access to health care as measured by the ratio of population to primary care physicians or % of uninsured. A recent study by Zhang et al. (2018)[13] on association of medical access and health outcomes in China found that inadequate access to health care was associated with higher risk of disability, cognitive

problems, and higher mortality. They also find that the relationship is stronger in women than in men. This has nothing to do with ease of access to parks. Similarly, measures of safety such as # of violent crimes/100,000 population could also influence people's health behaviours. For example, Christian et al. (2017)[14] conducted a survey-based study on residents in Perth and found that perception of safety from crime improved total minutes/week of walking by 7 to 22 minutes. Price et al. (2015)[15] examined the correlation between neighbourhood crime and measures of children's physical activity and found that children living in the lowest quartile levels of crime neighbourhoods report 40 minutes more moderate to vigorous physical activity than the highest quartile of crime neighbourhoods. Similarly, Janke, Popper and Shields (2015)[16] examine the impact of violent crime on walking behaviour in England and find that violent crime in the neighbourhood is a strong deterrent on walking and leads to a sharp drop in overall physical activity. These variables are then going to make it into the within-cluster regression model of ease of access to parks against measures of health as control variables.

The regression model is then developed for each individual cluster. Care will also be taken to ensure that potential interaction effects between the independent variables are taken into account. The model will look like something below:

$Y_i = A + BX_{1i} + CX_{2i} + DX_{3i} + EX_{2i}*X_{3i} + \ldots.$

Statistical tests and regression diagnostics are then conducted to ensure robustness of results. The analysis and tests that need to be conducted include:

1. Residual vs fitted plots to check linearity and homoscedasticity and also run tests for it
2. Q-Q plots to test normality assumption of residuals
3. Cook's distance to evaluate influence of outlier observations on regressions
4. Iterating on the regression to drop statistically insignificant variables across clusters; this can be done by looking at p-values of coefficients in the regression model

The final step is to compare the results across clusters and examine the relationship between the variables. This step will help address the specific aims of this research paper – what is the relationship between ease of access and health indicators, is it consistent across clusters, is it statistically significant and what is its quantification.

ANALYSIS AND RESULTS

Data Selection

The first step is to consider all the variables in each of the broad categories identified in this paper and reduce them to one or two critical variables for each category.

Primary independent variable: park_access (% of people living within half mile of a park)

Figure 1a: Distribution of % of people living within half mile of a park

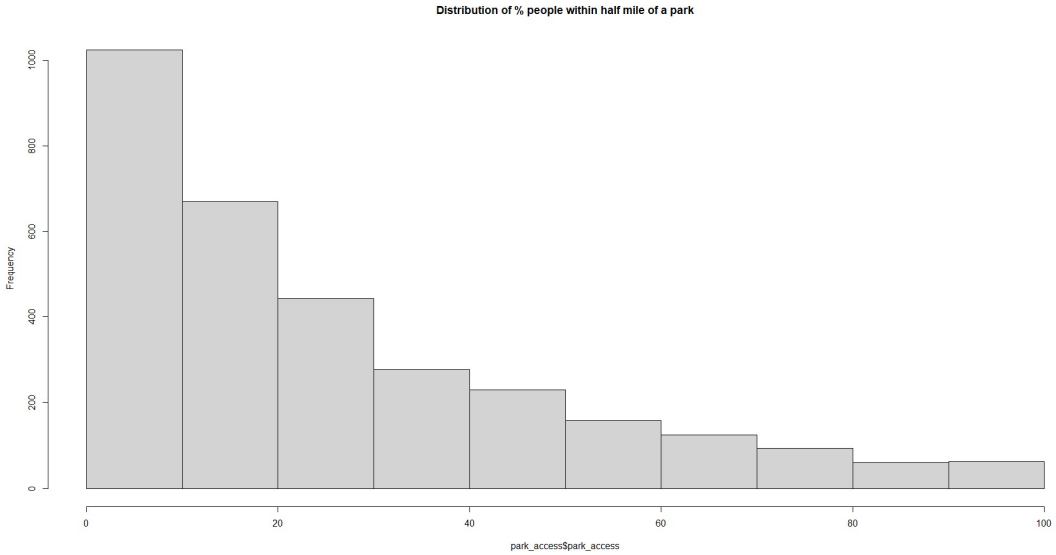


Table 1a: Statistical summary of park access variable

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	7.00	18.00	25.94	38.00	100.00

% of people living within half a mile of a park is skewed to the right with a long right tail. On average 18% of people live within half mile of a park. However, the distribution is spread across the spectrum with some counties having no parks within half a mile of the population and some having 100% of the population living within half mile of a park.

Health Outcomes (Dependent Variable):

Figure 1b: Distribution and correlation plot of health outcome variables

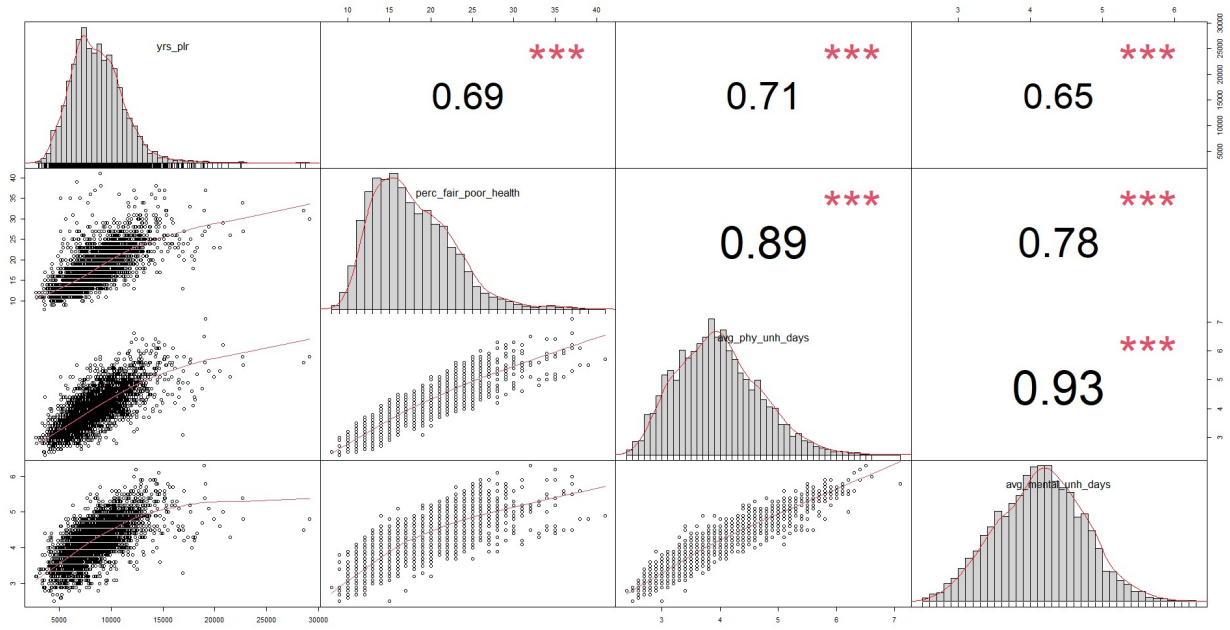


Table 1b: Statistical summary of health outcome variables

	yrs_plr	perc_fair_poor_health	avg_phy_unh_days	avg_mental_unh_days
	Min. : 2731	Min. : 8.00	Min. :2.400	Min. :2.500
	1st Qu.: 6793	1st Qu.:14.00	1st Qu.:3.500	1st Qu.:3.700
	Median : 8310	Median :17.00	Median :3.900	Median :4.200
	Mean : 8583	Mean :17.93	Mean :3.991	Mean :4.168
	3rd Qu.:10077	3rd Qu.:21.00	3rd Qu.:4.400	3rd Qu.:4.600
	Max. :29138	Max. :41.00	Max. :7.100	Max. :6.300
	NA's :293	NA	NA	NA

From Figure 1a and Table 1a, we can see that all of the variables are highly correlated with each other, so it is enough to include just one of the variables. Among the variables, “Years of potential life lost” has 293 missing values. Given the high correlation between % with fair/poor health and physically unhealthy/mentally unhealthy days, we will be using % with fair/poor health is the variable we will be including as a health indicator for subsequent analysis.

Health Behaviours:

Figure 1c: Distribution and correlation of health behaviour variables

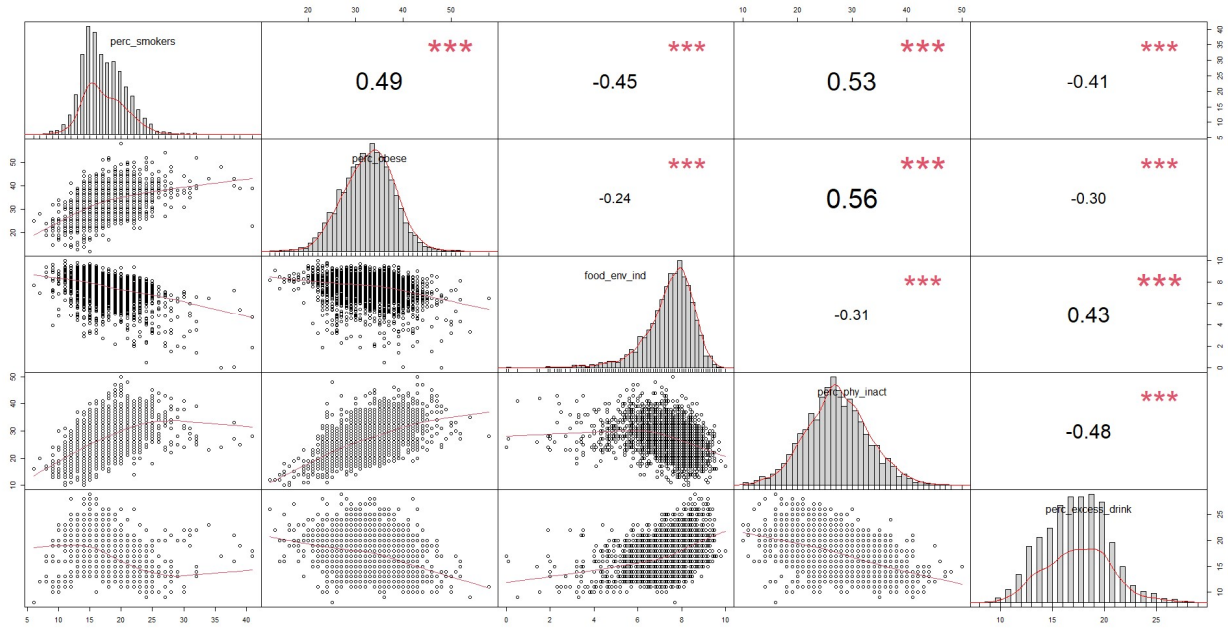


Table 1c: Statistical summary of health behaviour variables

	perc_smokers	perc_obese	food_env_ind	perc_phy_inact	perc_excess_drink
	Min. : 6.00	Min. :12.0	Min. : 0.000	Min. :10.00	Min. : 8.00
	1st Qu.:15.00	1st Qu.:29.0	1st Qu.: 6.900	1st Qu.:24.00	1st Qu.:15.00
	Median :17.00	Median :33.0	Median : 7.600	Median :27.00	Median :18.00
	Mean :17.47	Mean :32.9	Mean : 7.451	Mean :27.42	Mean :17.51
	3rd Qu.:20.00	3rd Qu.:37.0	3rd Qu.: 8.200	3rd Qu.:31.00	3rd Qu.:20.00
	Max. :41.00	Max. :58.0	Max. :10.000	Max. :50.00	Max. :29.00
	NA	NA	NA's :19	NA	NA

Among the highly correlated variables for health behaviours, we chose % of obese to be the variable of choice to represent this category due to no missing data and approximately normally distributed variable.

Clinical Care:

Figure 1d: Distribution and correlation of clinical care variables

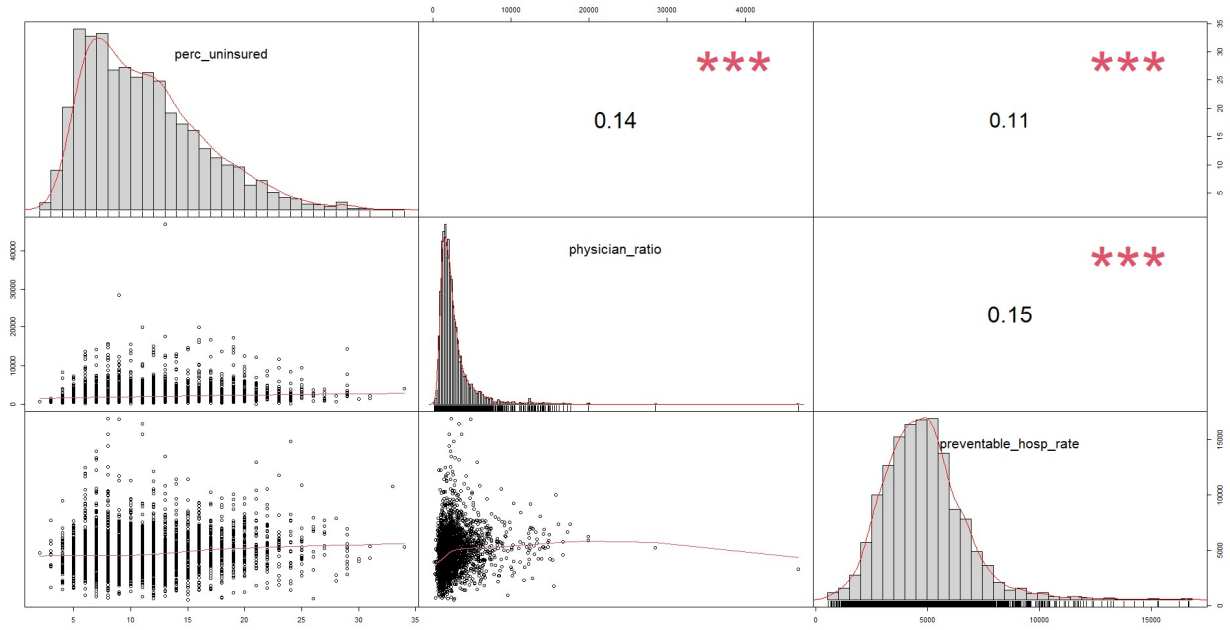


Table 1d: Statistical summary of clinical care variables

	perc_uninsured	physician_ratio	preventable_hosp_rate
	Min. : 2.00	Min. : 88	Min. : 536
	1st Qu.: 7.00	1st Qu.: 1372	1st Qu.: 3613
	Median :11.00	Median : 2003	Median : 4710
	Mean :11.48	Mean : 2640	Mean : 4859
	3rd Qu.:14.00	3rd Qu.: 3001	3rd Qu.: 5802
	Max. :34.00	Max. :46784	Max. :16851
	NA's :1	NA's :147	NA's :43

Physician ratio is a highly skewed variable. Preventable hospitalization rate has several missing values. Therefore, we use % uninsured as the variable that point to availability of clinical care.

Socio-economic variables:

Figure 1e: Distribution and correlation of socio-economic variables

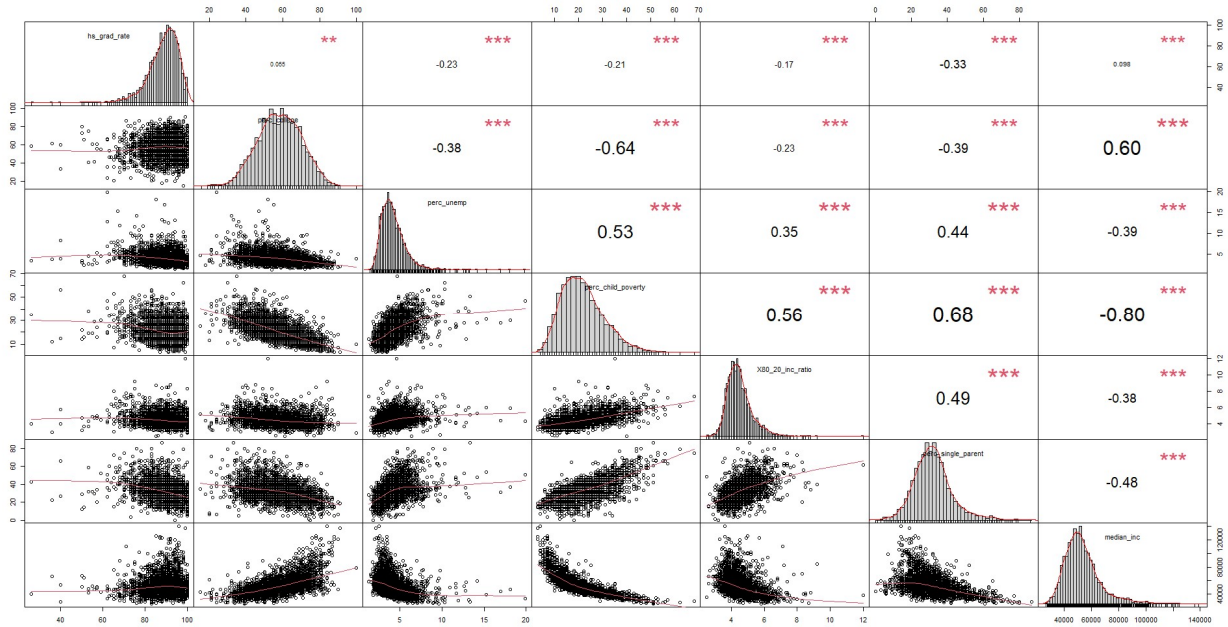


Table 1e: Statistical summary of socio-economic variables

	hs_grad_rate	perc_college	perc_unemp	perc_child_poverty
	Min. : 26.00	Min. : 15.00	Min. : 1.300	Min. : 3.00
	1st Qu.: 85.00	1st Qu.: 50.00	1st Qu.: 3.100	1st Qu.:15.00
	Median : 90.00	Median : 58.00	Median : 3.900	Median :20.00
	Mean : 88.76	Mean : 57.89	Mean : 4.133	Mean :21.16
	3rd Qu.: 94.00	3rd Qu.: 66.00	3rd Qu.: 4.800	3rd Qu.:26.00
	Max. :100.00	Max. :100.00	Max. :19.900	Max. :68.00
	NA's :96	NA	NA's :1	NA's :1

	X80_20_inc_ratio	perc_single_parent	median_inc
	Min. : 2.500	Min. : 0.00	Min. : 25385
	1st Qu.: 4.000	1st Qu.:26.00	1st Qu.: 43681
	Median : 4.400	Median :32.00	Median : 50568
	Mean : 4.513	Mean :32.36	Mean : 52794
	3rd Qu.: 4.900	3rd Qu.:38.00	3rd Qu.: 58848
	Max. :12.000	Max. :87.00	Max. :140382
	NA's :2	NA's :2	NA's :1

High school graduation rate is heavily skewed and has several missing values. So, we do not consider it for further analysis. % with college education appears more symmetrical bell shaped while the rest of the variables are skewed to the right.

Demographic variables:

Figure 1f: Distribution and correlation of demographic variables

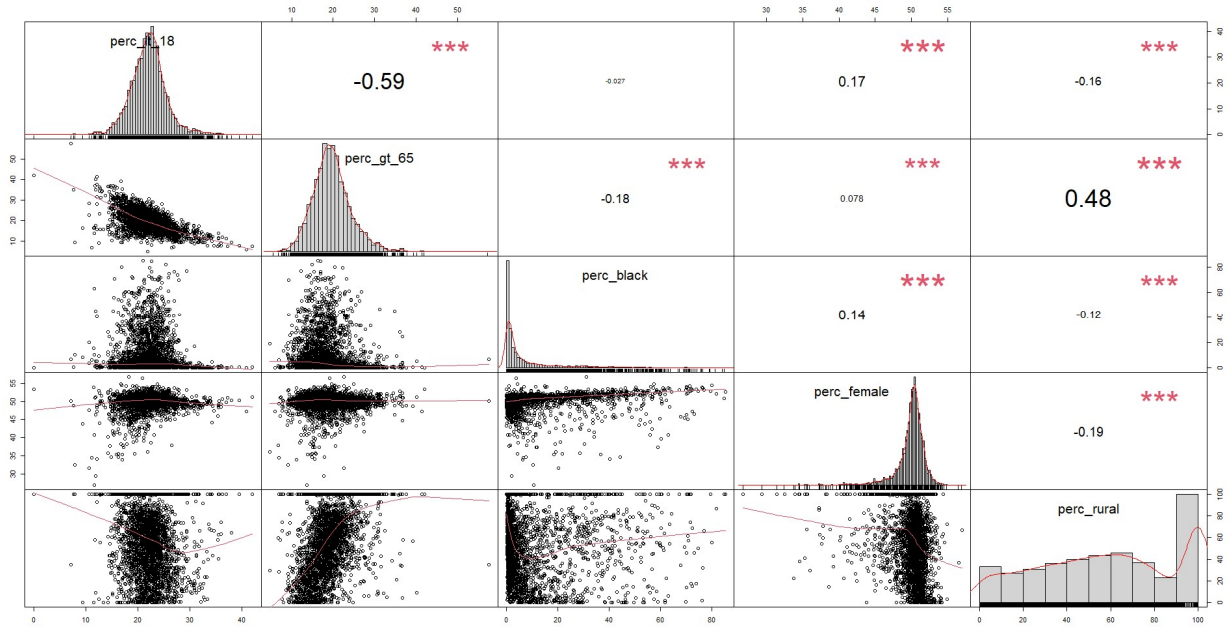


Table 1f: Statistical summary of demographic variables

	perc_lt_18	perc_gt_65	perc_black	perc_female	perc_rural
	Min. : 0.00	Min. : 4.80	Min. : 0.000	Min. :26.80	Min. : 0.00
	1st Qu.:20.00	1st Qu.:16.20	1st Qu.: 0.700	1st Qu.:49.40	1st Qu.: 33.25
	Median :22.10	Median :18.90	Median : 2.200	Median :50.30	Median : 59.50
	Mean :22.07	Mean :19.27	Mean : 8.994	Mean :49.89	Mean : 58.58
	3rd Qu.:23.90	3rd Qu.:21.80	3rd Qu.:10.200	3rd Qu.:51.00	3rd Qu.: 87.80
	Max. :42.00	Max. :57.60	Max. :85.400	Max. :56.90	Max. :100.00
	NA	NA	NA	NA	NA's :7

% less than 18 appears symmetrical bell shaped. % greater than 65 years appears slightly skewed to the right. % black appears heavily skewed to the right with a sharp peak and a long right tail. % female is skewed to the left with majority of counties having more than 50% females but with a long left tail. % rural similarly has a left skew.

Data Analysis

Now that all the variables for subsequent analysis have been identified, the next step is to examine the combined dataset as a whole and the impact of ease of access to parks on health outcomes.

First step in data analysis is to divide the data into manageable clusters with similar socio-economic and demographic indicators. There are a total of 6 socio-economic variables and 5

demographic variables. Several of these variables, as can be seen from the correlation plots in Figure 1e and 1f are highly correlated to each other. So before carrying out the cluster analysis, the dimensionality of the data needs to be reduced. For that principal component analysis is performed.

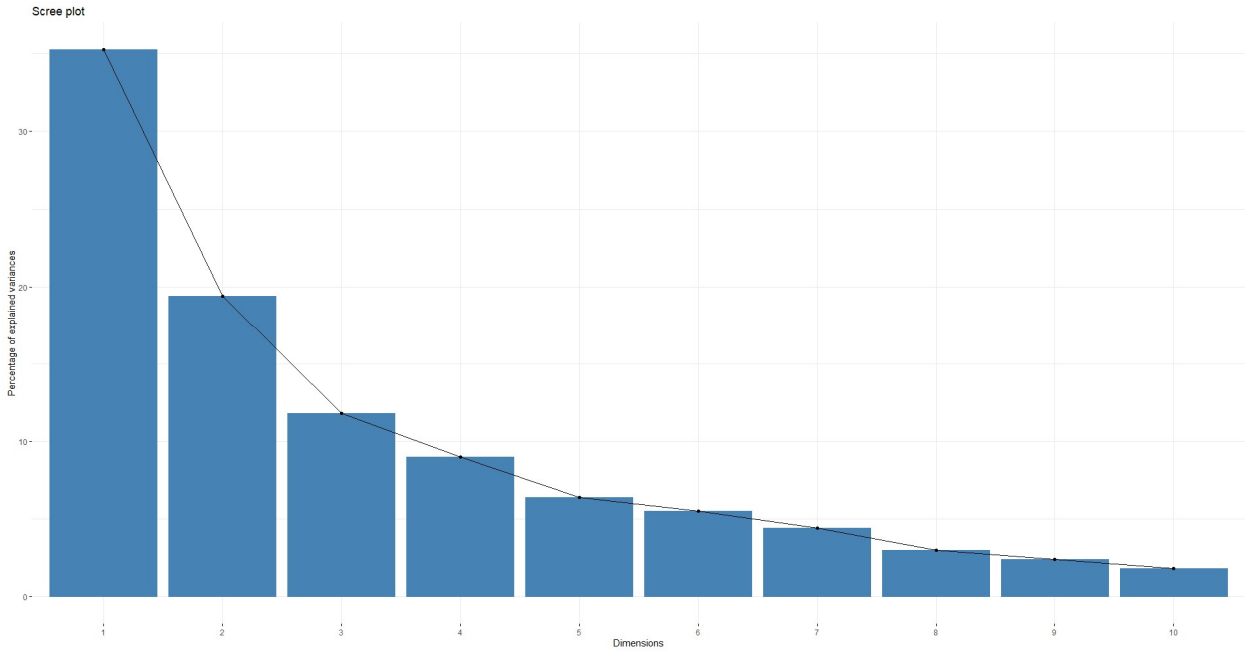
Principal Component Analysis:

Principal components of the 11 variables are described below.

Table 2a: Importance of components

Metric	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	1.97	1.46	1.14	0.99	0.84	0.78	0.70	0.57	0.51	0.45	0.34
Proportion of Variance	0.35	0.19	0.12	0.09	0.06	0.05	0.04	0.03	0.02	0.02	0.01
Cumulative Proportion	0.35	0.55	0.66	0.75	0.82	0.87	0.92	0.95	0.97	0.99	1.00

Figure 2a: Scree Plot of PCs



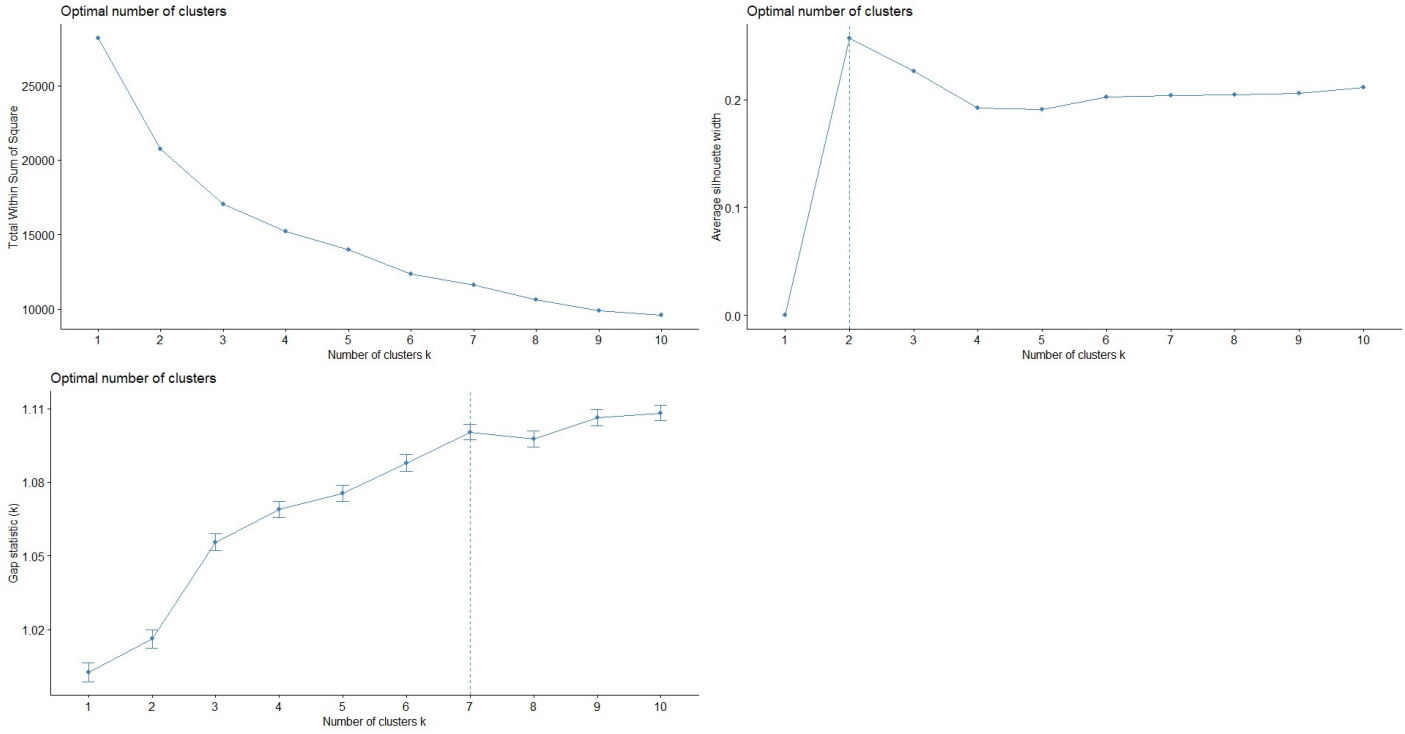
From the principal components, we can see that PC1 accounts for 35% of the variance. PC2 – PC5 account for 19%, 12%, 9%, 6% of the variance respectively. Together they account for over 80% of the variance. So, we will ignore all the other components beyond 5 for the rest of the analysis. Therefore, we can see that through principal component analysis, 11 socio-economic/demographic variables have been reduced to 5 manageable dimensions accounting

for over 80% of the variation in the data. The next step is to use these 5 PCs to divide the data into clusters.

Cluster Analysis:

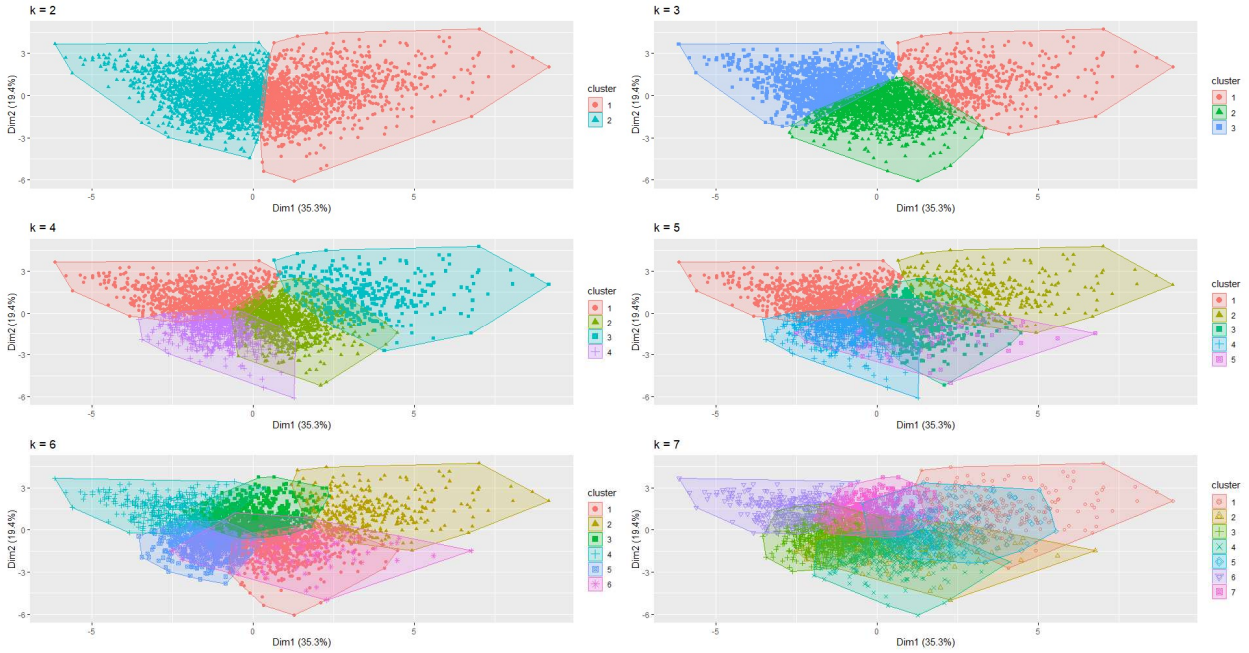
We use “within cluster sum of squares (WSS)”, “silhouette method” and “gap statistic” to carry out cluster analysis of the 5 principal components. The optimal number of clusters by different methods are shown below.

Figure 3a: Optimal number of clusters



From the above plots, we can see that the optimal number of clusters is between 2 and 7 depending on the method we use. Therefore, we try a few combinations of clusters and try to visualize whether clear delineation exists between the clusters.

Figure 3b: Cluster plot for number of clusters ranging from 2 to 5



From the above plot in Figure3b, we can see that 3 clusters seem to be the optimal as any clusters beyond that, there is no clear delineation of the clusters. Therefore, **number of optimal clusters = 3**. The number of variables per cluster is :

Cluster 1: 1,759

Cluster 2: 1,176

Cluster 3: 196

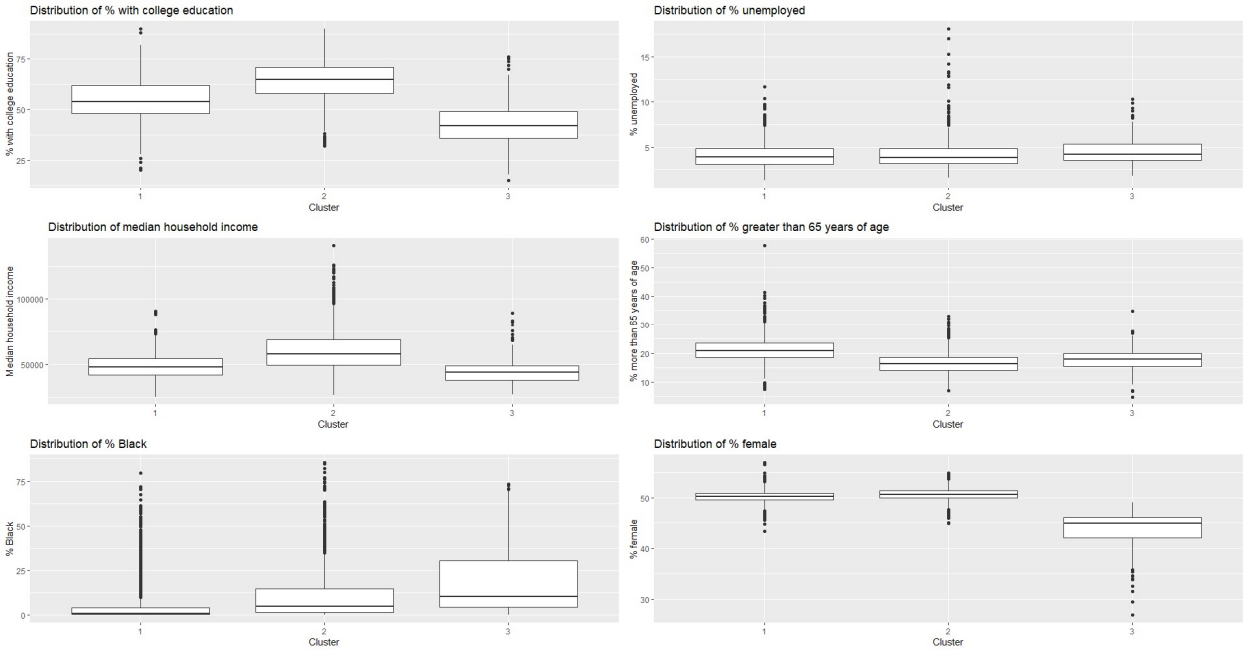
Table 3a: Median values of the socio-economic/demographic variables by cluster

Cluster	% College	% Unemp.	% child pov.	80/20%ile inc ratio	% single parent	Med. Income	% LT 18	% GT 65	% Black	% Female	% Rural
1	54	3.9	22	4.3	31	47985	22	20.7	1	50.2	78.4
2	65	3.8	17	4.5	33	57803	22.6	16.3	5	50.6	26.9
3	42	4.2	26	4.6	35	43704	19.2	17.8	10.3	44.9	67.4

From Table 3a, we can see that the data is divided into 3 clusters with differences in some key variables. Cluster 2 has the highest % of college educated (65%) and cluster 3 has the lowest (42%). Similarly, cluster 3 has the highest unemployment (4.2%), highest child poverty (26%) and lowest median income ($43.7K). It also has the highest percentage of African-Americans and lowest percentage of women. The cluster which has the highest median income (cluster 2) also has the least % of rural.

The differences among some of the “key variables” can be understood better through a series of box plots for the variables above as shown in Figure 3c below.

Figure 3c: Boxplots of key variables in the socio-economic/demographic clusters

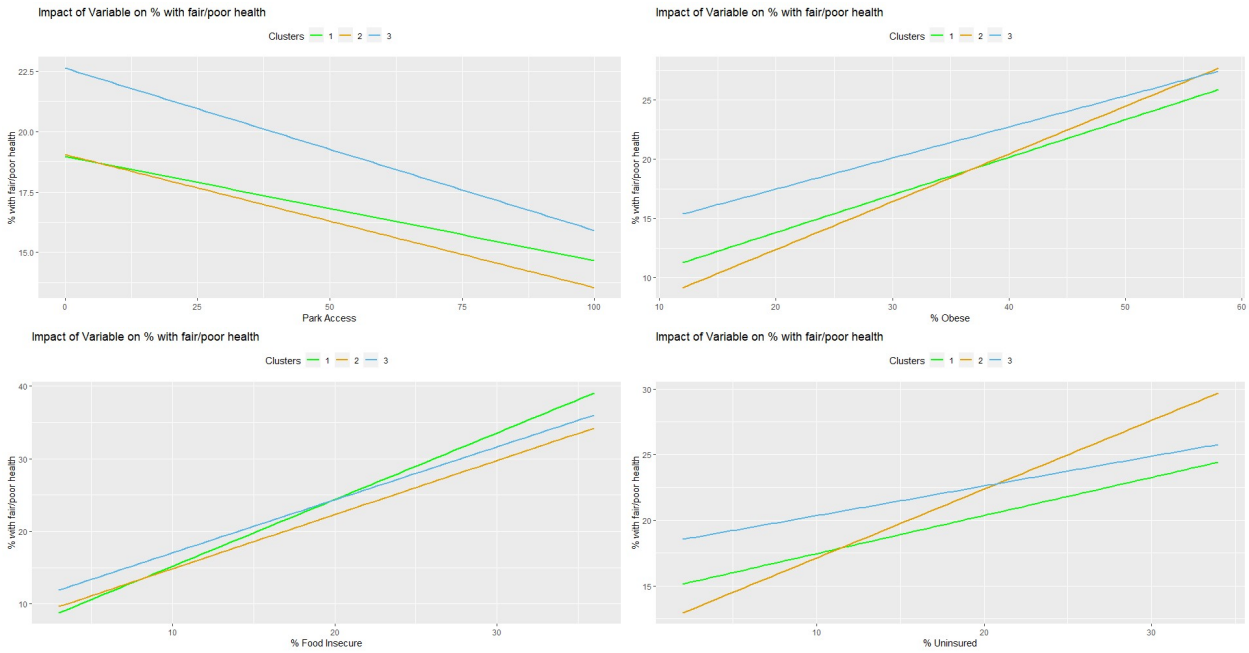


Next step is to carry out regression analysis by cluster for the entire dataset.

Regression Analysis:

To begin with, the variables outside of socio-economic/demographic variables that weren’t used in the principal component and clustering analysis are examined for their relationship to health outcomes within each cluster. The plot can be seen below in Figure 4a.

Figure 4a: Regression plot of % with fair/poor health (health outcome) against dependent variables – Park Access, % Obese, % Food Insecure and % Uninsured



From the plots, we already observe certain interesting trends:

- The magnitude of the slope of the lines varies by clusters indicating that the strength of the relationship varies by cluster
- The directionality of each line is the same for each variable across clusters, i.e. direction of relationship is consistent across clusters for each variable
- Park Access: We can see from the first plot that improvement in access to parks reduces % of people with fair/poor health which is a positive indication of the primary relationship being examined in this paper
- % Obese: As the % of obese in the population increases, the % of people with fair/poor health increases which makes intuitive sense as obesity causes negative health outcomes
- % Food insecure: More is the food insecurity, more is the % of population reporting fair/poor health which is also intuitive
- % Uninsured: As a larger % of population is uninsured, there are more % of people with fair/poor health. This also makes sense because lack of insurance makes healthcare unaffordable or very expensive and results in poor health outcomes

Elaborate regression models are now developed for these variables accounting for interaction effects between variables as well.

It can be seen from Figure 1b that % of population with fair/poor health is not normally distributed but is positively skewed with a long right tail. Therefore, the log of this variable is taken for subsequent analysis.

Cluster 1:

Regression Model 1: Log(%_fair_poor_health) =

C0 + C1*park_access + C2*(%_obese)*(%_food_insecure)*(%_uninsured)

Table 5a: Regression model 1 output

term	estimate	std.error	statistic	p.value
(Intercept)	1.0497122	0.2587713	4.056525	0.0000520
park_access	-0.0013429	0.0001838	-7.307172	0.0000000
perc_obese	0.0183221	0.0075640	2.422260	0.0155253
perc_food_insecure	0.1126990	0.0192742	5.847147	0.0000000
perc_uninsured	0.0742076	0.0187580	3.956060	0.0000792
perc_obese:perc_food_insecure	-0.0007839	0.0005483	-1.429792	0.1529551
perc_obese:perc_uninsured	-0.0006172	0.0005548	-1.112505	0.2660737
perc_food_insecure:perc_uninsured	-0.0048400	0.0013288	-3.642298	0.0002781
perc_obese:perc_food_insecure:perc_uninsured	0.0000428	0.0000382	1.120706	0.2625668

From the model 1 output, we can see that the interaction terms are statistically significant. However, we can see that % obese is not as statistically significant by itself and in any interaction term. Therefore, we attempt another model by dropping obesity variable.

Regression Model 2: Log(%_fair_poor_health) =

C0 + C1*park_access + C2*(%_food_insecure)*(%_uninsured)

Table 5b: Regression model 2 output

term	estimate	std.error	statistic	p.value
(Intercept)	1.6489884	0.0396359	41.60338	0
park_access	-0.0018079	0.0001784	-10.13699	0
perc_food_insecure	0.0892102	0.0030363	29.38158	0
perc_uninsured	0.0520810	0.0030269	17.20578	0
perc_food_insecure:perc_uninsured	-0.0033806	0.0002220	-15.23112	0

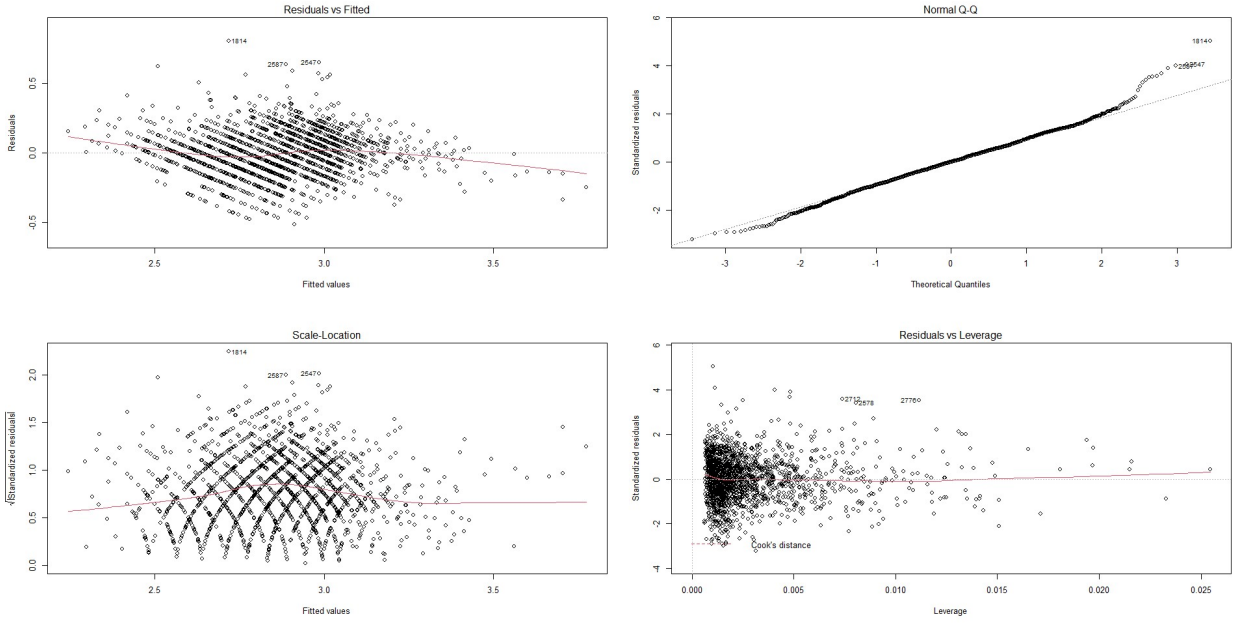
Model 2 has all the coefficients statistically significant and the coefficient of park_access is negative indicating that improvement in park access leads to fewer people with fair/poor health. Furthermore, we can see that as expected, more food insecurity and more the % of uninsured population, more are the % of people with fair/poor health. Interestingly, there is an interaction

effect between the two and the negative coefficient indicates that these two variables are positively correlated, and the interaction term removes the potential double counting arising from including two positively correlated variables. It makes sense that % uninsured and % food insecure are positively correlated because people who are not getting enough to eat are unlikely to be able to afford insurance. We can see model 2 is the superior one in terms of predictive power as can be seen from the ANOVA test below in Table 5c.

Table 5c: Model comparison

```
Analysis of Variance Table

Model 1: log_perc_fph ~ park_access + perc_obese * perc_food_insecure * perc_uninsured
Model 2: log_perc_fph ~ park_access + perc_food_insecure * perc_uninsured
  Res.Df    RSS Df Sum of Sq      F               Pr(>F)
1   1750 43.133
2   1754 45.233 -4   -2.1008 21.309 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5a: Regression diagnostics for Model 2



From residual vs fitted and scale-location plots, we can see that the residuals do not exhibit any evidence of heteroskedasticity. Furthermore, from the Q-Q plot, we can see that non-normality is not a huge problem for this model. Finally, the residuals vs leverage plot shows us that outliers do not play an outsized role in this model.

Cluster 2:

Regression Model 1: Log(%_fair_poor_health) =

C0 + C1*park_access + C2*(%_obese)*(%_food_insecure)*(%_uninsured)

Table 6a: Regression model 1 output

term	estimate	std.error	statistic	p.value
(Intercept)	1.8184986	0.1861065	9.7712787	0.0000000
park_access	-0.0005596	0.0002089	-2.6790457	0.0074870
perc_obese	0.0043792	0.0057371	0.7633190	0.4454274
perc_food_insecure	0.0348289	0.0155909	2.2339331	0.0256763
perc_uninsured	0.0450151	0.0159464	2.8229039	0.0048399
perc_obese:perc_food_insecure	0.0006332	0.0004613	1.3725842	0.1701454
perc_obese:perc_uninsured	0.0001022	0.0004817	0.2122159	0.8319756
perc_food_insecure:perc_uninsured	-0.0016461	0.0012580	-1.3085445	0.1909463
perc_obese:perc_food_insecure:perc_uninsured	-0.0000241	0.0000365	-0.6602088	0.5092501

From the model 1 output, we can see that the interaction terms are statistically significant. However, we can see that % obese is not as statistically significant by itself and in any interaction term. Therefore, we attempt another model by dropping obesity variable.

Regression Model 2: Log(%_fair_poor_health) =

C0 + C1*park_access + C2*(%_food_insecure)*(%_uninsured)

Table 6b: Regression model 2 output

term	estimate	std.error	statistic	p.value
(Intercept)	1.9240262	0.0362890	53.019497	0
park_access	-0.0016330	0.0001937	-8.430923	0
perc_food_insecure	0.0611029	0.0028573	21.384605	0
perc_uninsured	0.0481926	0.0030273	15.919412	0
perc_food_insecure:perc_uninsured	-0.0024525	0.0002273	-10.790623	0

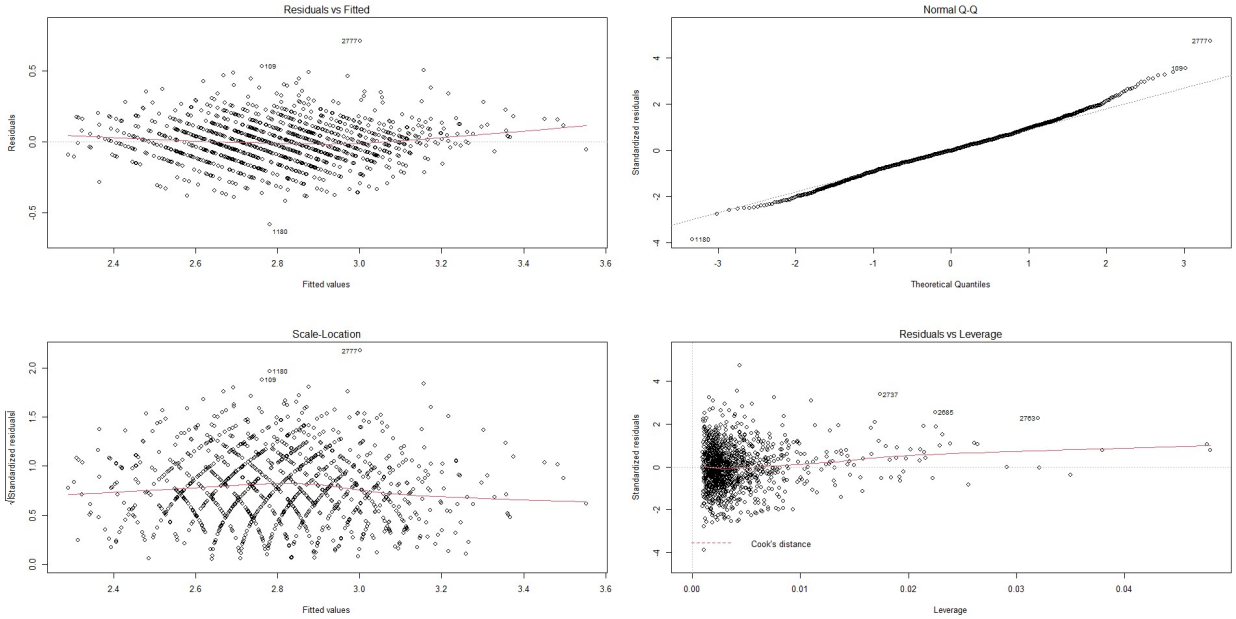
Model 2 has all the coefficients statistically significant and the coefficient of park_access is negative indicating that improvement in park access leads to fewer people with fair/poor health. Similar to cluster 1, the other coefficients exhibit expected behavior. We can see model 2 is the superior one in terms of predictive power as can be seen from the ANOVA test below in Table 6c.

Table 6c: Model comparison

```
Analysis of Variance Table

Model 1: log_perc_fph ~ park_access + perc_obese * perc_food_insecure * perc_uninsured
Model 2: log_perc_fph ~ park_access + perc_food_insecure * perc_uninsured
```

```
  Res.Df    RSS Df Sum of Sq    F               Pr(>F)
1   1167 24.104
2   1171 26.764 -4   -2.6604 32.2 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6a: Regression diagnostics for Model 2



From residual vs fitted and scale-location plots, we can see that the residuals do not exhibit any evidence of heteroskedasticity. The Q-Q plot shows non-normality is not a huge problem for this model. Finally, the residuals vs leverage plot shows us that outliers do not play an outsized role in this model.

Cluster 3:

Regression Model 1: Log(%_fair_poor_health) =

C0 + C1*park_access + C2*(%_obese)*(%_food_insecure)*(%_uninsured)

Table 7a: Regression model 1 output

term	estimate	std.error	statistic	p.value
(Intercept)	1.1668640	0.8114764	1.4379518	0.1521191
park_access	-0.0016018	0.0004822	-3.3220594	0.0010747
perc_obese	0.0192695	0.0217114	0.8875278	0.3759349
perc_food_insecure	0.1102472	0.0603411	1.8270669	0.0692838
perc_uninsured	0.0922645	0.0478983	1.9262554	0.0555889
perc_obese:perc_food_insecure	-0.0010319	0.0015818	-0.6523413	0.5149821

perc_obese:perc_uninsured	-0.0011210	0.0013181	-0.8504436	0.3961661
perc_food_insecure:perc_uninsured	-0.0054483	0.0035420	-1.5382154	0.1256863
perc_obese:perc_food_insecure:perc_uninsured	0.0000671	0.0000953	0.7039513	0.4823387

From the model 1 output, we can see that the intercept term is not statistically significant. interaction terms are also not statistically significant. %uninsured and % food insecure are significant at 10%. Therefore, we attempt another model by dropping obesity variable and then evaluating the outcome.

Regression Model 2: Log(%_fair_poor_health) =
C0 + C1*park_access + C2*(%_food_insecure)*(%_uninsured)

Table 7b: Regression model 2 output

term	estimate	std.error	statistic	p.value
(Intercept)	1.8015076	0.1165593	15.455721	0.000000
park_access	-0.0017560	0.0004612	-3.807632	0.000189
perc_food_insecure	0.0763568	0.0076064	10.038471	0.000000
perc_uninsured	0.0554357	0.0076184	7.276549	0.000000
perc_food_insecure:perc_uninsured	-0.0032320	0.0005100	-6.337288	0.000000

Model 2 has all the coefficients statistically significant and the coefficient of park_access is negative indicating that improvement in park access leads to fewer people with fair/poor health. Similar to cluster 1 and 2, the other coefficients exhibit expected behaviour. We can see model 2 is no better than model 1 in terms of predictive power as can be seen from the ANOVA test below in Table 7c.

Table 7c: Model comparison

```
Analysis of Variance Table

Model 1: log_perc_fph ~ park_access + perc_obese * perc_food_insecure * perc_uninsured
Model 2: log_perc_fph ~ park_access + perc_food_insecure * perc_uninsured
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    187 4.3306
2    191 4.3836 -4 -0.053042 0.5726 0.6828
```

Comparing model 1 and model 2, interaction terms which were insignificant before become significant in model 2 after dropping % obesity, while predictive power does not really improve. Therefore, we try another model dropping interaction terms from model 2.

Regression Model 3: Log(%_fair_poor_health) =
C0 + C1*park_access + C2*(%_food_insecure) + C3*(%_uninsured)

Table 7d: Regression model 2 output

term	estimate	std.error	statistic	p.value
(Intercept)	2.4589789	0.0582964	42.180638	0.0000000
park_access	-0.0018545	0.0005057	-3.666834	0.0003179
perc_food_insecure	0.0307522	0.0027038	11.373838	0.0000000
perc_uninsured	0.0088958	0.0022241	3.999813	0.0000904

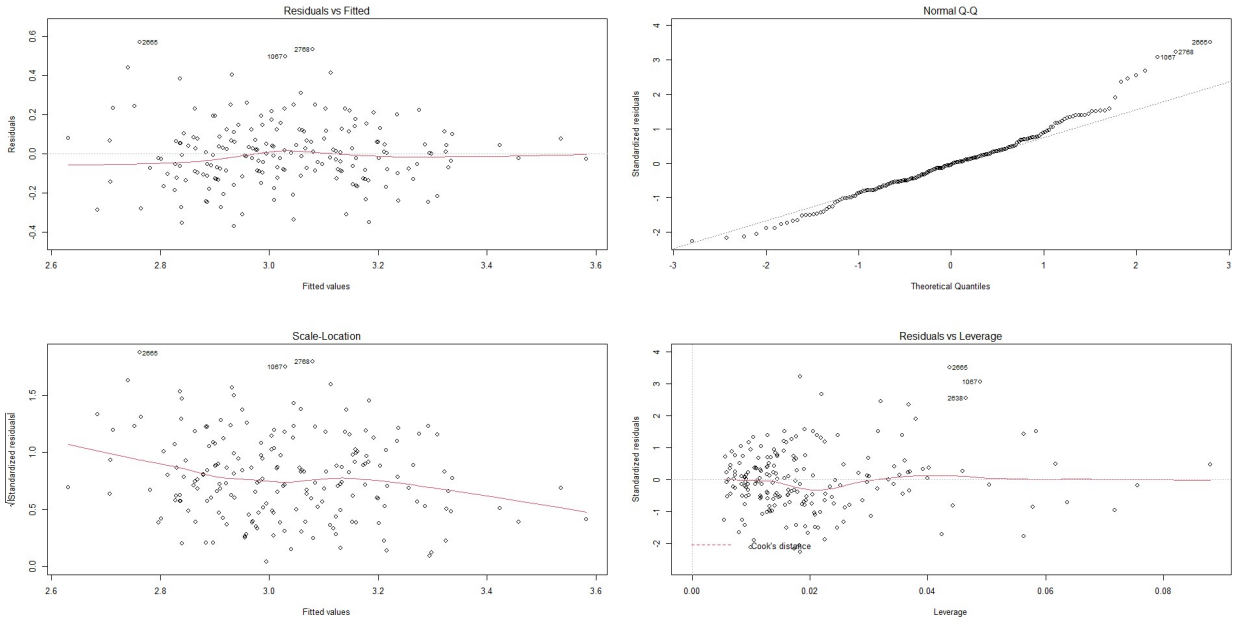
Model 3 has all the coefficients statistically significant and the coefficient of park_access is negative indicating that improvement in park access leads to fewer people with fair/poor health. Like cluster 1 and 2, the other coefficients exhibit expected behaviour. There is no interaction term unlike earlier models. We can see model 3 is better than model 1 and model 2 in terms of predictive power as can be seen from the ANOVA test below in Table 7e.

Table 7e: Model comparison

```
Analysis of Variance Table

Model 1: log_perc_fph ~ park_access + perc_obese * perc_food_insecure * perc_uninsured
Model 2: log_perc_fph ~ park_access + perc_food_insecure + perc_uninsured
  Res.Df    RSS Df Sum of Sq      F       Pr(>F)
1    187 4.3306
2    192 5.3054 -5  -0.97478 8.4184 0.0000003342 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Analysis of Variance Table

Model 1: log_perc_fph ~ park_access + perc_food_insecure * perc_uninsured
Model 2: log_perc_fph ~ park_access + perc_food_insecure + perc_uninsured
  Res.Df    RSS Df Sum of Sq      F         Pr(>F)
1    191 4.3836
2    192 5.3054 -1  -0.92174 40.161 0.000000001643 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparing model 1, model 2 and model 3, model 3 is the best and the diagnostics are shown below.

Figure 7a: Regression diagnostics for Model 3



From residual vs fitted and scale-location plots, we can see that the residuals do not exhibit any evidence of heteroskedasticity. The Q-Q plot shows non-normality is not a huge problem for this model. Finally, the residuals vs leverage plot shows us that outliers do not play an outsized role in this model.

Results

The results from all the analysis is summarized below:

1. The national county level data can be divided into 3 clusters based on socio-economic and demographic data
2. Each cluster differs in terms of % black population, median income, % older than 65, % female, % living in rural population, % completed college and % unemployed
3. Coefficients of the key variables for each of the regression is shown below:

Cluster	%_park_access	%_food_insecure	%_uninsured	%_food_insecure*%_uninsured
1	-0.0018079	0.0892102	0.0520810	-0.0033806
2	-0.0016330	0.0611029	0.0481926	-0.0024525
3	-0.0018545	0.0307522	0.0088958	N/A

4. From the above regression coefficients, we can see that for every 1% increase in the percentage of people living within half a mile of a park, log (percentage of people with fair/poor health) drops by roughly 0.2%.

a. The relationship holds across clusters and varies only slightly in magnitude
b. The dependent variable depends not just on ease of access to parks but also other variables such as % of population without insurance and % of population with food insecurity with % of population with fair or poor health increasing with increase in these variables

SIGNIFICANCE AND CONCLUSION

Significance

This study can add to the body of knowledge in a significant way in that very few studies exist which study the impact of access to public parks on long term health outcomes. They are mostly limited to impact of park usage on physical activity and assume a positive link between physical activity and better health outcomes. If this study does indicate that easier access to public parks leads to positive health outcomes and if this finding is consistent across clusters with a wide variety of demographic differences, then that is a strong finding and can support additional research into this link. This paper can add to the body of literature that provides the data-based backing needed to push through changes in the policy that would improve investment in public parks.

Conclusion

From this paper, we can overall conclude the following:

- There is evidence for ease of access to parks improving health outcomes across the US
- The evidence is consistent at varying levels of income, education, age, access to medical care, and race and gender difference
- There are other factors that influence health outcomes such as unhealthy behaviours, food security, pre-existing health conditions etc.
    - But the presence of these factors does not negate the finding. Rather, the relationship between ease of access to parks and health outcomes is still maintained and just as impactful as in the absence of these factors
- Roughly 0.2% reduction in log % of people with fair/poor health happens for every 1% increase in % of people living within half a mile of a park

REFERENCES

[1] Stewart O, Moudon A, Littman A, Seto E, Saelens B. Why neighborhood park proximity is not associated with total physical activity. Health & Place. 2018;52:163-169.

[2] Cohen DA, Han B, Derose KP, Williamson S, Marsh T, McKenzie TL. Physical activity in parks: a randomized control trial using community engagement. American Journal of Preventive Medicine. 2013 Nov; 45(5): 590-7

[3] Floyd MF, Bocarro JN, Smith WR, Baran PK, Moore RC, Cosco NG, Edwards MB, Suau LJ, Fang K. Park-based physical activity among children and adolescents. American Journal of Preventive Medicine. 2011 Sep; 41(3): 258-265

[4] Cohen DA, Marsh T, Williamson S, Derose KP, Martinez H, Setodji C, McKenzie TL. Parks and physical activity: Why are some parks used more than others?. American Journal of Preventive Medicine. 2010 Jan; 50(Suppl 1): S9-12

[5] Scott MM, Evenson KR, Cohen DA, Cox CE. Comparing perceived and objectively measured access to recreational facilities as predictors of physical activity in adolescent girls. Journal of Urban Health. 2007 May; 84(3): 346-359

[6] Bedimo-Rung AL, Mowen AJ, Cohen DA. The significance of parks to physical activity and public health: a conceptual model. American Journal of Preventive Medicine. 2005 Feb; 28(2): 159-168

[7] National Environmental Public Health Tracking Network Query Tool [Internet]. Ephtracking.cdc.gov. 2019 [cited 14 November 2019]. Available from: https://ephtracking.cdc.gov/DataExplorer/#/

[8] [Internet]. Countyhealthrankings.org. 2020 [cited 26 July 2020]. Available from: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

[9] Variable descriptions sourced as-is from:
[Internet]. Countyhealthrankings.org. 2020 [cited 26 July 2020]. Available from: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

[10] Wallace, M., Sharfstein, J., Kaminsky, J. and Lessler, J. (2019). Comparison of US County-Level Public Health Performance Rankings With County Cluster and National Rankings. JAMA Network Open, 2(1), p.e186816.

[11] Rayward, A., Duncan, M., Brown, W., Plotnikoff, R. and Burton, N. (2017). A cross-sectional cluster analysis of the combined association of physical activity and sleep with sociodemographic and health characteristics in mid-aged and older adults. Maturitas, 102, pp.56-61.

[12] Scheinker, D., Valencia, A. and Rodriguez, F. (2019). Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Machine Learning Models. JAMA Network Open, 2(4), p.e192884.

[13] Zhang, X., Dupre, M., Qiu, L., Zhou, W., Zhao, Y. and Gu, D. (2018). Age and sex differences in the association between access to medical care and health outcomes among older Chinese. BMC Health Services Research, 18(1).

[14] Christian, H., Knuiman, M., Divitini, M., Foster, S., Hooper, P., Boruff, B., Bull, F. and Giles-Corti, B. (2017). A Longitudinal Analysis of the Influence of the Neighborhood Environment on Recreational Walking within the Neighborhood: Results from RESIDE. Environmental Health Perspectives, 125(7), p.077009.

[15] Kneeshaw-Price, S., Saelens, B., Sallis, J., Frank, L., Grembowski, D., Hannon, P., Smith, N. and Chan, K. (2015). Neighborhood Crime-Related Safety and Its Relation to Children's Physical Activity. Journal of Urban Health, 92(3), pp.472-489.

[16] Janke, K., Propper, C. and Shields, M. (2016). Assaults, murders and walkers: The impact of violent crime on physical activity. Journal of Health Economics, 47, pp.34-49.

APPENDICES

Appendix 1: Code

```
# Load all the required packages
library(readr)
library(psych)
library(tidyverse)
library(Hmisc)
library(stats)
library(car)
library(dae)
library(corrplot)
library(dplyr)
library(moments)
library(ggplot2)
library(bios2mds)
library(factoextra)
library(PerformanceAnalytics)
library(gridExtra)

# Set Working Directory
setwd("C:/Users/subha/Desktop/GRAD695")


# Health Outcomes data analysis
health_out<-read.csv("health_outcomes.csv")
# Subset the data to only look at county data
health_out_cty <- subset(health_out, County!="")
```

```
#Summary statistics of health outcomes
knitr::kable(summary(health_out_cty[,c(4:7)]))

my_data <- health_out_cty[,c(4:7)]

# Correlation plot of health outcomes
chart.Correlation(my_data, histogram=TRUE, pch=19)



# Health Behaviors data analysis
health_beh<-read.csv("health_behaviours.csv")
health_beh_cty <- subset(health_beh, County!="")
summary(health_beh_cty[,c(4:8)])

#Summary statistics of health behaviours
knitr::kable(summary(health_beh_cty[,c(4:8)]))

my_data1 <- health_beh_cty[, c(4:8)]

# Correlation plot of health behaviours
chart.Correlation(my_data1, histogram=TRUE, pch=19)



# Clinical Care access data analysis
health_acc<-read.csv("clinical_care.csv")
health_acc_cty <- subset(health_acc, County!="")
summary(health_acc_cty[,c(4:6)])

# Summary statistics of clinical care access
knitr::kable(summary(health_acc_cty[,c(4:6)]))

my_data2 <- health_acc_cty[, c(4:6)]

# Correlation plot of clinical care access
chart.Correlation(my_data2, histogram=TRUE, pch=19)
```

```
# Socio-economic data analysis
health_soc<-read.csv("socio_economic_data.csv")
health_soc_cty <- subset(health_soc, County!="")

# Summary of socio-economic data
knitr::kable(summary(health_soc_cty[,c(4:10)]))

my_data3 <- health_soc_cty[, c(4:10)]

# Correlation plot of socio-economic data
chart.Correlation(my_data3, histogram=TRUE, pch=19)



#demographic data analysis
health_demo<-read.csv("demographic_data.csv")
health_demo_cty <- subset(health_demo, County!="")

# Summary of demographic data
knitr::kable(summary(health_demo_cty[,c(4:8)]))

my_data4 <- health_demo_cty[, c(4:8)]

# Correlaton plot of demographic data
chart.Correlation(my_data4, histogram=TRUE, pch=19)



health_ind<-read.csv("final_datasetV1.csv")
health_ind_cty <- subset(health_ind, County!="")

park_access<-read.csv("data_191957.csv")
#head(park_access)

hist(park_access$park_access, main = "Distribution of % people within half mile of a park")
knitr::kable(summary(park_access$park_access))
summary(park_access$park_access)

# Merge the park access and other data
```

```
merged_data<-merge(park_access, health_ind_cty, by.x = "countyFIPS", by.y = "FIPS", all.x = TRUE)

summary(merged_data)

clust_data <- merged_data[, c(6,9:12, 14:24)]
clust_data1 <- clust_data[complete.cases(clust_data),]



#Principal Component Analysis
x<-prcomp(clust_data1[,c(6:16)], retx=TRUE, center=TRUE, scale=TRUE)
summary(x)

# Scree Plot
fviz_screeplot(x)
biplot(x,scale=0, cex=1.3)

pcs<-data.frame(x$x[,1:5])

data1<-data.frame(clust_data1,pcs)



set.seed(123)

# function to compute total within-cluster sum of square

pa<-fviz_nbclust(data1[,c(17:21)], kmeans, method="wss")
pb<-fviz_nbclust(data1[,c(17:21)], kmeans, method="silhouette")
pc<-fviz_nbclust(data1[,c(17:21)], kmeans, nstart = 25,  method = "gap_stat", nboot = 50)

grid.arrange(pa,pb, pc, nrow=2)

# Plot the different clusters
k2 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 2, nstart = 25)
k3 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 3, nstart = 25)
k4 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 4, nstart = 25)
k5 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 5, nstart = 25)
k6 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 6, nstart = 25)
```

```
k7 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 7, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 5")
p5 <- fviz_cluster(k6, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 6")
p6 <- fviz_cluster(k7, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 7")

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)



final <- kmeans(scale(data1[, c(17:21)]), 3, nstart = 25)

#Print the centers, size
print(final$centers)
print(final$size)



final_dataset0<-data.frame(data1,final$cluster)
final_dataset<- final_dataset0[complete.cases(final_dataset0),]



#Summarize the median data for each cluster
final_dataset %>%
  group_by(final.cluster) %>%
  summarise(across(c(6:10), median))



final_dataset %>%
  group_by(final.cluster) %>%
  summarise(across(c(11:16), median))



# Box Plots
```

```
x1<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_college))+geom_boxplot()+
ggtitle("Distribution of % with college education")+xlab("Cluster")+ylab("% with college education")

x2<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_unemp))+geom_boxplot()+
ggtitle("Distribution of % unemployed")+xlab("Cluster")+ylab("% unemployed")

x3<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=median_inc))+geom_boxplot()+
ggtitle("Distribution of median household income")+xlab("Cluster")+ylab("Median household income")

x4<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_gt_65))+geom_boxplot()+
ggtitle("Distribution of % greater than 65 years of age")+xlab("Cluster")+ylab("% more than 65 years of
age")

x5<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_black))+geom_boxplot()+
ggtitle("Distribution of % Black")+xlab("Cluster")+ylab("% Black")

x6<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_female))+geom_boxplot()+
ggtitle("Distribution of % female")+xlab("Cluster")+ylab("% female")

grid.arrange(x1, x2, x3, x4, x5, x6, nrow = 3)


# Regression Plots for non socio-economic/demographic variables
y1<-
  ggplot(final_dataset, aes(x=park_access, y=perc_fair_poor_health, color=factor(final.cluster),
shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "Park Access",
color="Clusters", shape = "Clusters")

y2<-
  ggplot(final_dataset, aes(x=perc_obese, y=perc_fair_poor_health, color=factor(final.cluster),
shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
```

```
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "% Obese", color="Clusters", shape = "Clusters")

y3<-
  ggplot(final_dataset, aes(x=perc_food_insecure, y=perc_fair_poor_health, color=factor(final.cluster), shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "% Food Insecure", color="Clusters", shape = "Clusters")

y4<-
  ggplot(final_dataset, aes(x=perc_uninsured, y=perc_fair_poor_health, color=factor(final.cluster), shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "% Uninsured", color="Clusters", shape = "Clusters")

grid.arrange(y1, y2, y3, y4, nrow = 2)



# Regression analysis by clusters
final_dataset$log_perc_fph<-log(final_dataset$perc_fair_poor_health)
cluster1<-subset(final_dataset, final.cluster == 1)
cluster2<-subset(final_dataset, final.cluster == 2)
cluster3<-subset(final_dataset, final.cluster == 3)
```

```
chart.Correlation(final_dataset[,c(1:5,14)], histogram=TRUE, pch=19)


# Cluster 1 analysis
c1_model1<-lm(log_perc_fph~park_access+perc_obese*perc_food_insecure*perc_uninsured,
data=cluster1)
kable(tidy(c1_model1))

c1_model2<-lm(log_perc_fph~park_access+perc_food_insecure*perc_uninsured, data=cluster1)
kable(tidy(c1_model2))
anova(c1_model1, c1_model2)

par(mfrow=c(2,2))
plot(c1_model2)


# Cluster 2 analysis
c2_model1<-lm(log_perc_fph~park_access+perc_obese*perc_food_insecure*perc_uninsured,
data=cluster2)
kable(tidy(c2_model1))

c2_model2<-lm(log_perc_fph~park_access+perc_food_insecure*perc_uninsured, data=cluster2)
kable(tidy(c2_model2))
anova(c2_model1, c2_model2)

par(mfrow=c(2,2))
plot(c2_model2)


# Cluster 3 analysis
c3_model1<-lm(log_perc_fph~park_access+perc_obese*perc_food_insecure*perc_uninsured,
data=cluster3)
kable(tidy(c3_model1))

c3_model2<-lm(log_perc_fph~park_access+perc_food_insecure*perc_uninsured, data=cluster3)
kable(tidy(c3_model2))
anova(c3_model1, c3_model2)
```

```
par(mfrow=c(2,2))
plot(c3_model2)

c3_model3<-lm(log_perc_fph~park_access+perc_food_insecure+perc_uninsured, data=cluster3)
kable(tidy(c3_model3))
anova(c3_model1, c3_model3)
anova(c3_model2, c3_model3)

par(mfrow=c(2,2))
plot(c3_model3)
```