Connecting School Performance to Public Parks

Applied Data Science Capstone Project

Dan Ritter, 2020/08/21

Introduction/Business Problem

Purpose

This project aims to identify and prioritize improvement opportunities for neighborhood parks in Green Bay, Wisconsin, USA.

Audience

The target audience for this analysis is a community action group working toward improved elementary education outcomes for the city population. The group includes representation from the school district, local government, business community leaders, and citizen stakeholders. The group wants to know if they should pursue a park improvement initiative.

Background

Successful school outcomes are the result of a complex system of factors. Many factors are within a school's control, such as student-to-teacher ratio or highly qualified teachers. Many more factors, however, are outside a school's control, such as parent engagement, neighborhood safety, and food security. To effectively address this complex issue, a complex solution is required, with interventions applied to internal and external factors concurrently. It is unrealistic to expect a school district to create meaningful and lasting improvements to student outcomes without outside assistance addressing external factors. The external factor examined by this analysis is equitable access to outdoor recreational activities provided by public parks. The greenspaces and playgrounds found in public parks promote the holistic development of healthy children through physical activity, environment exploration, and unstructured, self-directed play.

Objective 1: Proof of Principle

Based on observations made from the data, is this idea worth pursuing? The primary purpose of this analysis is to determine whether or not to pursue a park improvement initiative. In other words, is there any measurable relationship between school outcomes and access to parks.

- Hypothesis 1: Quantity of nearby parks is related to higher school performance.
- Hypothesis 2: Size of nearby parks is related to higher school performance.
- Hypothesis 3: Specific park amenities are related to higher school performance.

Objective 2: Identify Recommendations for Action

If the determination is made to pursue the initiative, which park improvements should be prioritized?

- Which neighborhoods are in need of park improvement/development?
- What specific improvements might provide the most benefit to park users?

Data

To understand the current state of elementary **schools** in Green Bay, the following data will be pulled from the Wisconsin Department of Public Instruction ("DPI") website:

- Names and locations of elementary schools in the Green Bay Area Public School District ("GBAPS").
- School accountability reports from the last seven years. These reports include a variety of performance metrics. The data will be used to select neighborhoods in greatest need of intervention. This data can also be used as a target for machine learning models seeking to explain factors related to school performance.

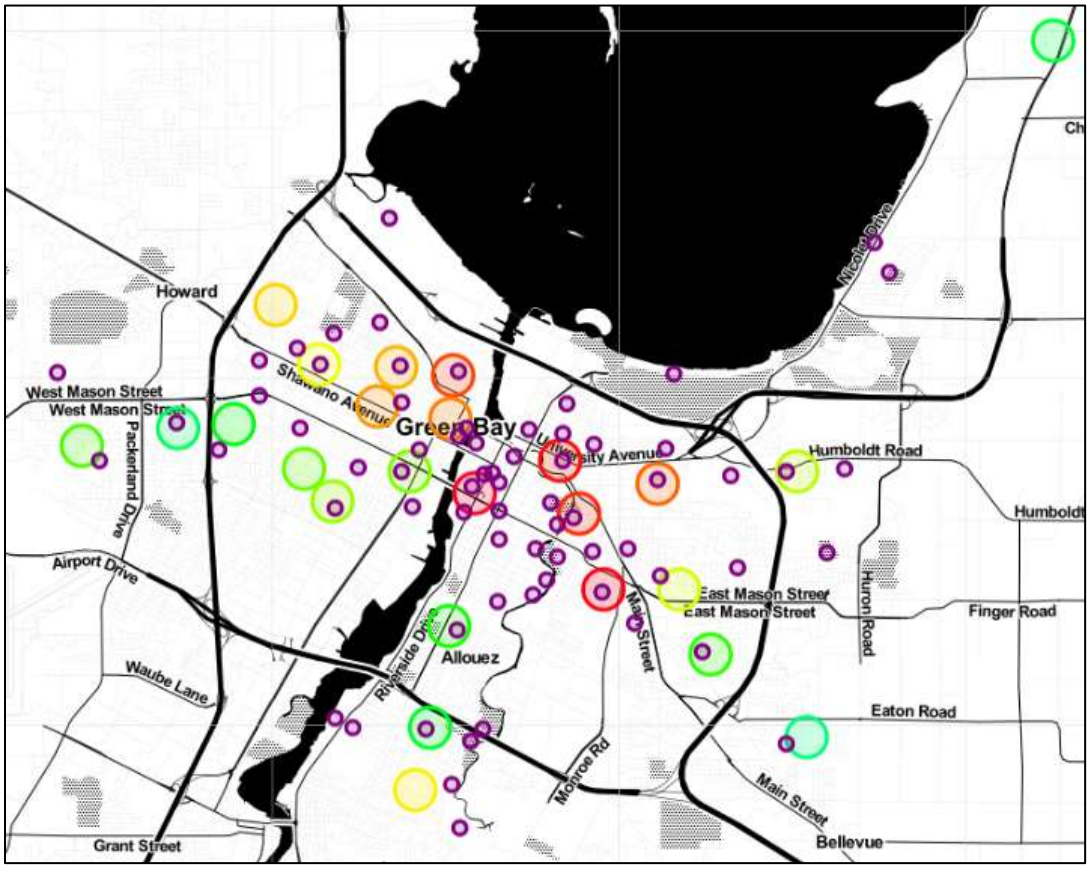
To understand the current state of **parks** in Green Bay, the following data will be retrieved from the Foursquare API and the Green Bay Department of Parks, Recreation and Forestry website.

- Names and locations of parks in the proximity of elementary schools. Park distance from school will be a feature considered for modeling. (Foursquare)
- User ratings of parks. This will serve as a metric of park quality, another feature to be considered for modeling. (Foursquare)
- Amenities found at each park in Green Bay. (Dept of Parks)

Figure 1: Map of Green Bay Schools and Parks

Schools are indicated with large circles, and shaded according to performance metrics; red = low-performing, yellow = average-performing, green = high-performing.

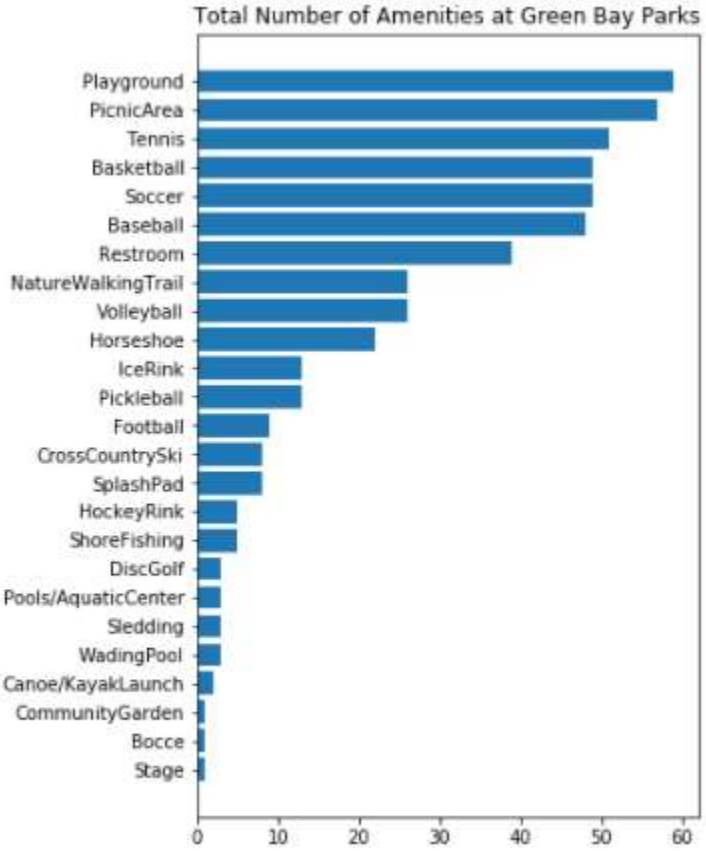
Parks are indicated with small purple circles.



Upon visual inspection, there appears to be a trend of decreasing scores for schools closer to the center of the city. In addition, poverty levels and crime rates tend to rise in neighborhoods closer to the center of the city. These two factors likely have an impact on students' school performance.

We can also see there tends to be a high concentration of parks near the lowest performing schools. It appears Hypothesis 1 (Quantity of nearby parks is related to higher school performance.) will likely be rejected. Parks near downtown tend to be smaller, so this may support Hypothesis 2 (Size of nearby parks is related to higher school performance).

Figure 2: Total Count of Each Type of Amenity at Green Bay Parks



Looking at Figure 2, we can see that some park amenities are more common, such as playgrounds, picnic areas, tennis courts, basketball courts, soccer fields and baseball diamonds. Since they are so prevalent, I think it is likely that they will not be related to school performance.

Other amenities are rarer, such as disc golf courses, pools, sledding hills, and wading pools. Might one of the amenities in the bottom half of the list be related to higher school performance?

Methodology

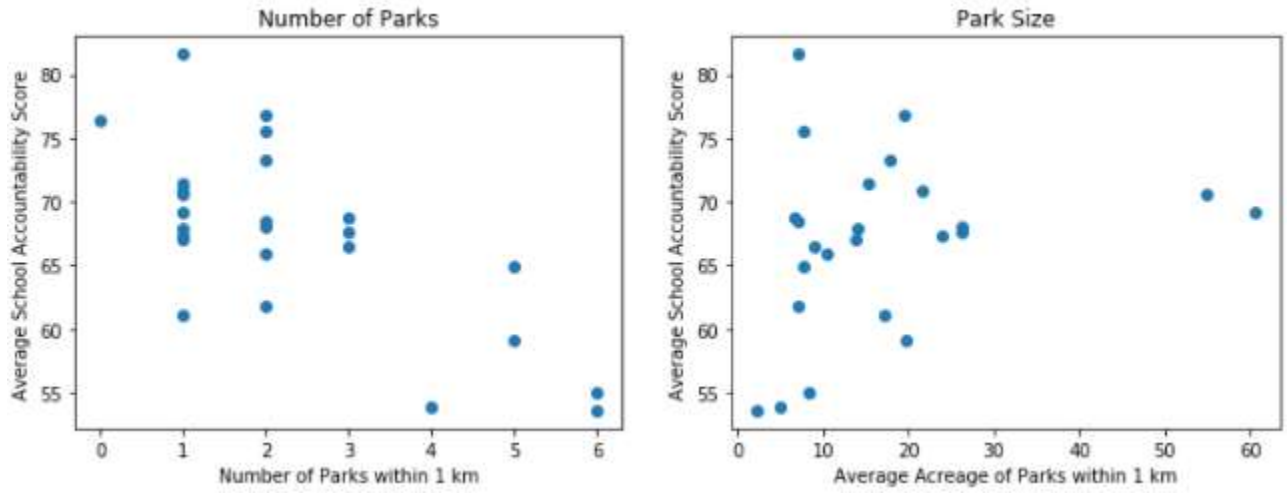
I will use Multiple Linear Regression to test the impact features have on school performance scores.

First, I need to select features for modeling. I need to find features that are individually correlated to the target variable, but not too strongly correlated with each other. (If there are multiple features that are strongly correlated to the target and each other, it might be appropriate to create a composite variable.) The potential features to be considered are:

- Total number of parks within 1 kilometer of school
- Total park acreage within 1 kilometer of school
- Average acreage per park with 1 kilometer of school
- Average distance from school to parks within 1 kilometer of school
- Total number of each amenity (25 amenities in total) within 1 kilometer of school

Let's check out a couple features with scatter plots. One hypothesis I have, based on the maps above, is a "quality over quantity" sort of dynamic exists. Based on the map in Figure 1, I'm guessing there will be a negative correlation between school performance and number of parks, and a positive correlation between school performance and average nearby acreage.

Figure 3: Plots of Number of Parks, Park Size against School Performance Metric



There is definitely a negative correlation with the number of parks, but park size does not seem to be significantly correlated. Following is a full list of features with a correlation of >0.3 or <-0.3.

Potentially significant positive correlations:

- Sledding: 0.350813

Potentially significant negative correlations:

- Nearby Parks: -0.707351 (Figure 3)
- Basketball: -0.486927
- Nature Walking Trail: -0.314951
- Picnic Area: -0.730287
- Playground: -0.655987
- Horseshoe: -0.611832
- Restroom: -0.457380
- Volleyball: -0.441440

- Pools/Aquatic Center: -0.301798
- Stage: -0.404332

The next step is to inspect a correlation table of the features listed above, and inspect for multicollinearity.

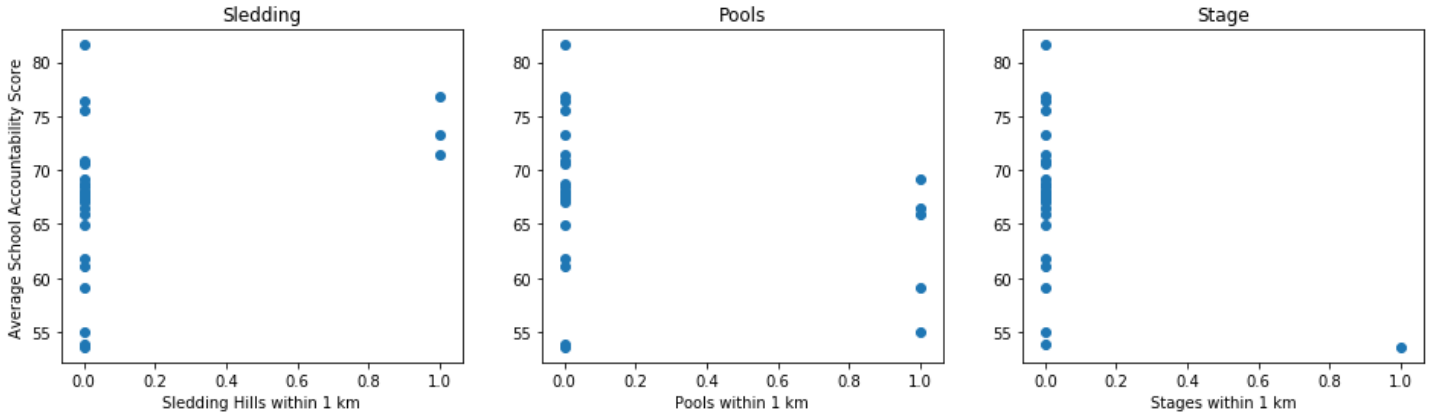
Figure 4: Correlation Table of Potential Features for Modeling

	NearbyParks	Basketball	NatureWalkingTrail	PicnicArea	Playground	Horseshoe	Restroom	Volleyball	Pools/AquaticCenter	Sledding	Stage
NearbyParks	1.000000	0.743742	0.538833	0.968510	0.839101	0.753249	0.728086	0.512125	0.328780	-0.146891	0.457355
Basketball	0.743742	1.000000	0.320056	0.748476	0.918808	0.809693	0.721006	0.484799	0.373182	-0.049764	0.025392
NatureWalkingTrail	0.538833	0.320056	1.000000	0.555361	0.431552	0.487818	0.248350	-0.165928	0.250000	-0.020515	0.059536
PicnicArea	0.968510	0.748476	0.555361	1.000000	0.851816	0.776353	0.707416	0.532133	0.381597	-0.187886	0.406158
Playground	0.839101	0.918808	0.431552	0.851816	1.000000	0.853917	0.761446	0.549511	0.329276	-0.125278	0.134427
Horseshoe	0.753249	0.809693	0.487818	0.776353	0.853917	1.000000	0.740140	0.581302	0.349482	-0.118300	0.008917
Restroom	0.728086	0.721006	0.248350	0.707416	0.761446	0.740140	1.000000	0.708842	0.211195	-0.046216	0.134120
Volleyball	0.512125	0.484799	-0.165928	0.532133	0.549511	0.581302	0.708842	1.000000	0.145693	0.017934	0.257742
Pools/AquaticCenter	0.328780	0.373182	0.250000	0.381597	0.329276	0.349482	0.211195	0.145693	1.000000	-0.184637	-0.102062
Sledding	-0.146891	-0.049764	-0.020515	-0.187886	-0.125278	-0.118300	-0.046216	0.017934	-0.184637	1.000000	-0.075378
Stage	0.457355	0.025392	0.059536	0.406158	0.134427	0.008917	0.134120	0.257742	-0.102062	-0.075378	1.000000

Multicollinearity exists in several of these features. Many of these variables are strongly correlated with NearbyParks (Basketball, NatureWalkingTrail, PicnicArea, Playground, Horseshoe, Restroom, Volleyball). This makes sense, they are the most common park amenities and are found at most parks. Since they are found at most parks, I will keep the NearbyParks variable but remove the others from consideration.

Three variables other than NearbyParks are still under consideration: Pools, Sledding, and Stage. Let's inspect each of these with scatter plots.

Figure 5: Plots of Sledding Hills, Pools, Stages against School Performance Metric



Since there is only one instance of Stage, it doesn't make sense to keep that as a feature, but I'll keep the other two.

The final list of features for modeling is:

- Total number of nearby parks
- Nearby sledding hills
- Nearby pools or aquatic centers

Now I will perform linear regression to see how the features interact with the target variable. My objective, at this point, is simply to measure what relationship may exist. I'm not trying to predict school performance scores, so I will not split the data into training and testing sets. However, if I had a larger training set, I probably would split it.

OLS Regression Results

Dep. Variable:	AvgOverallScore	R-squared:	0.564
Model:	OLS	Adj. R-squared:	0.502
Method:	Least Squares	F-statistic:	9.057
Date:	Sat, 25 Jul 2020	Prob (F-statistic):	0.000481
Time:	19:32:35	Log-Likelihood:	-73.444
No. Observations:	25	AIC:	154.9
Df Residuals:	21	BIC:	159.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	73.2510	1.824	40.155	0.000	69.457	77.045
NearbyParks	-2.7709	0.645	-4.294	0.000	-4.113	-1.429
Sledding	5.2529	3.134	1.676	0.109	-1.264	11.770
Pools/AquaticCenter	-0.6899	2.667	-0.259	0.798	-6.236	4.856

Omnibus:	0.876	Durbin-Watson:	1.323
Prob(Omnibus):	0.645	Jarque-Bera (JB):	0.178
Skew:	0.167	Prob(JB):	0.915
Kurtosis:	3.243	Cond. No.	9.78

I have decided to only consider features with a p-value of less than 0.05. The only significant feature is NearbyParks.

Results

Hypothesis 1: Access to parks is related to higher school performance.

Result: Not Supported

In fact, a significant negative correlation exists between the number of nearby parks and school performance. In other words, low-performing schools tend to have more parks nearby. This appears to be due in part to the high density of very small parks located in the city center, where lower-performing schools also tend to be located.

Hypothesis 2: The size of parks, not the quantity, is related to higher school performance.

Result: Not Supported

There was no significant correlation between the two variables. Small and mid-sized parks were equally likely to be near low- or high-performing schools. The largest parks tended to be around average-performing schools.

Hypothesis 3: Specific park amenities are related to higher school performance.

Result: Not Supported

Many common park amenities, such as playgrounds or basketball courts, are so prevalent that there is not a meaningful difference between the parks themselves and individual park amenities. Are few exceptions are Sledding Hills and Pools/Aquatic Centers, which were both positively correlated with higher school performance, but not significantly.

Discussion

Recommendation 1: Data Development Agenda

It is recommended that the City of Green Bay collect, store, and make publicly-available detailed information on its Parks, Recreation and Forestry holdings and assets. This would include an assessment of park amenities, date of last improvement, planned future improvements, cost of amenities.

It is unfortunate that Foursquare had insufficient rating data for Green Bay parks, because a measure of park quality is major gap in this analysis. Schools in dense urban areas are clearly nearby several parks, but the current state of those parks is not accounted for. Playgrounds which are old, outdated, or in disrepair could potentially have an adverse impact on neighborhood kids. Furthermore, more detailed and publicly-available on parks would be helpful in determining the impact of parks on the communities they are in. Such data would allow us to examine what impacts are observed immediately following park improvements. Do school performance scores rise? Does crime decrease?

Recommendation 2: Focus on Winter Activities

Continue to investigate the impact of sledding hills.

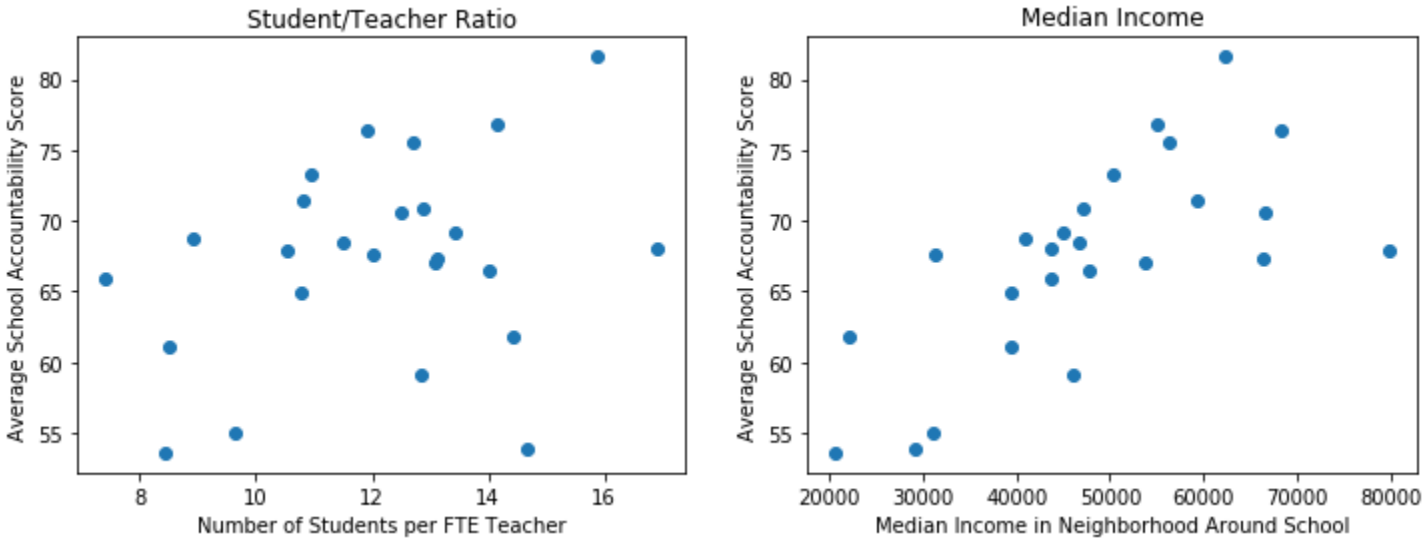
Although sledding hills ended up not passing the test for statistical significance, it was close. In Green Bay, snow may be on the ground for 4-6 months each year. This has a major impact on access to park amenities. Most publicly-available, outdoor, winter activites require expensive equipment, such as cross-country skiiing, hockey, or ice skating. Sledding is unique in its accessibility to a wider range of socioeconomic levels. Furthermore, if parks are to have an impact on school performance, it seems important that meaningful park amenities would be available when school is actually in session (the winter).

Other questions...

How does this park data compare to other school or community factors?

I found two more variables that might have a significant impact on school performance: Student/Teacher Ratio, and Median Income of families in neighborhood around each school. Below are the scatter plots and regression results for these two variables.

Figure 6: Plots of Student/Teacher Ratio, Neighborhood Median Income against School Performance Metric



OLS Regression Results

Dep. Variable:	AvgOverallScore	R-squared:	0.574
Model:	OLS	Adj. R-squared:	0.514
Method:	Least Squares	F-statistic:	9.448
Date:	Wed, 29 Jul 2020	Prob (F-statistic):	0.000376
Time:	08:16:39	Log-Likelihood:	-73.143
No. Observations:	25	AIC:	154.3
Df Residuals:	21	BIC:	159.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.6371	5.787	7.714	0.000	32.603	56.671
MedianIncome	0.0003	7.04e-05	4.148	0.000	0.000	0.000
Sledding	5.1614	3.087	1.672	0.109	-1.258	11.581
StudentTeacherRatio	0.6734	0.429	1.569	0.132	-0.219	1.566

Omnibus:	0.876	Durbin-Watson:	2.202
Prob(Omnibus):	0.645	Jarque-Bera (JB):	0.850
Skew:	-0.256	Prob(JB):	0.654
Kurtosis:	2.255	Cond. No.	2.93e+05

Median Income is definitely a significant factor, but Student/Teacher Ratio is actually (slightly) less significant than nearby Sledding Hills. Perhaps more reason to give more consideration to accessible winter activities.

Conclusion

Objective 1: Proof of Principle = Rejected

At this time, it is the recommendation of this analysis that the community action group not pursue a park improvement initiative as a targeted means of improving school performance. Instead, it is recommended the group pursue impacts related to Median Income level. Questions to consider in future analysis include:

- Is this the right data to answer this question? Incorporating time-based data may be more appropriate.
- Is the Overall School Accountability Score an accurate measure of school performance? Does it represent child well-being in a meaningful way?
- What would be a more meaningful metric to represent the well-being of neighborhood children?