Overpopulation is Atlanta Public Schools

James Hauth

7/13/2020

1. Introduction

1.1. Problem

The city of Atlanta is home to half a million residents and is serviced by 10 public high schools. Unfortunately, these schools and populations are not evenly represented across city limits. Many students may need to travel larger distances to get to their public high school than students who may be districted in another high school. Furthermore, some high schools may have to cater to a larger population than other high schools, which can lead to overcrowding or under crowding depending on the school.

1.2. Proposals

In this project we will try to find an optimal location for a new high school. Specifically, this report will be targeted to public school lawmakers in **Atlanta, GA**. There are already several public high schools in the city of Atlanta, but they do not represent equal portions of the population or equal areas of the city. We will be attempting to find a location for a new school that can reduce overcrowding in Atlanta Public Schools.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by the city

2. Data

2.1. Population data

The information needed for this project is related to the population of the city of Atlanta, the neighborhoods within the city, and the demographics of the schools themselves. Data was retrieved from the 2010 census data found online on Arcgis.

2.2. Data formatting and cleaning

Information used in this project was saved into two pandas DataFrames. Schools contained information on each school, including its name, physical coordinates, and address. NeighDF includes information on neighborhoods, including its name, centroid coordinates, population, and assigned school district. A third DataFrame, withPop was created, but is functionally identical to Schools and a new “proposed” school.

Data on BEST Academy High School was removed from the set, as it shares a campus with Coretta Scott King Young Women's Leadership Academy. They are both single gendered high schools. For these reasons, the two schools were considered as one.

2.3. Data Examples

2.3.1.

	School	Address	Latitude	Longitude
0	Benjamin Elijah Hays High School	3450 Benjamin E Mays Dr SW Atlanta, Georgia 30...	33.737973	-84.500985
1	BEST Academy High School	1190 Northwest Drive NW, Atlanta, Georgia 30318	33.788677	-84.479339
2	Booker T. Washington High School	45 Whitehouse Drive SW, Atlanta, Georgia 30314...	33.754066	-84.420035
3	Coretta Scott King Young Women's Leadership Ac...	1190 Northwest Drive NW, Atlanta, GA 30318	33.788677	-84.479339
4	Daniel McLaughlin Therrell High School	3099 Panther Trail Southwest Atlanta, Georgia	33.699190	-84.490096

2.3.2.

	Neighborhood	Lat	Long	Pop	Assigned School
0	Arden/Habersham, Argonne Forest, Peachtree Bat...	33.830723	-84.398380	2672.0	
1	Peachtree Heights East, Peachtree Hills	33.820586	-84.381416	3736.0	
2	Peachtree Heights West	33.832133	-84.388578	4874.0	
3	Buckhead Forest, South Tuxedo Park	33.846813	-84.383542	3372.0	
4	Chastain Park, Tuxedo Park	33.865428	-84.398157	3423.0	

3. Methodology

3.1. In this project, we are focused on determining the optimal location for opening a new public high school. In order to do this, we make a few assumptions. First, we assume that populations are equally distributed within neighborhoods. Second, that there are equal percentage distributions of high school aged minors among all the neighborhoods. Finally, that an entire neighborhood is assigned to one school.

3.2. The first step in our process is to plot each school and create neighborhood clusters for each school based upon the distance each neighborhood is away from each school. We will use a version of K means clustering for this step.

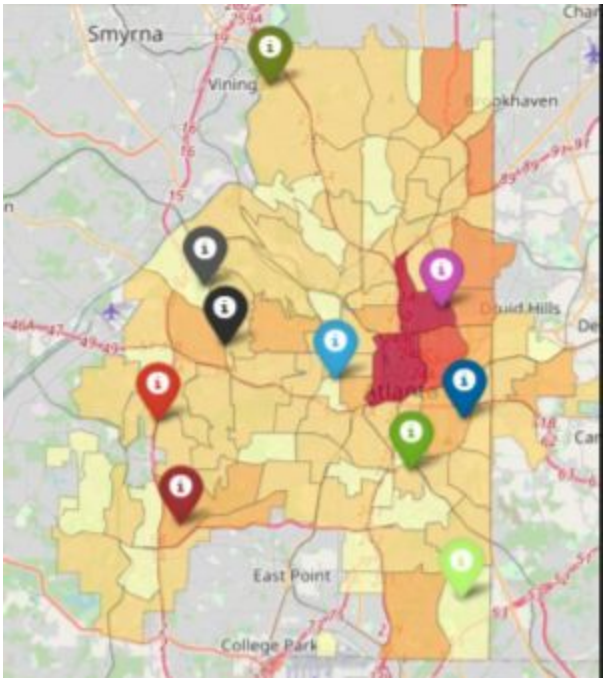
3.3. From there, we will analyze the proportion of the population that each school represents. This will help us determine which schools can be

considered overpopulated and would therefore need another school to handle the population.

3.4. Finally, we will be adding one new school and using many iterations of K means clustering to determine which location would be ideal for a new school.

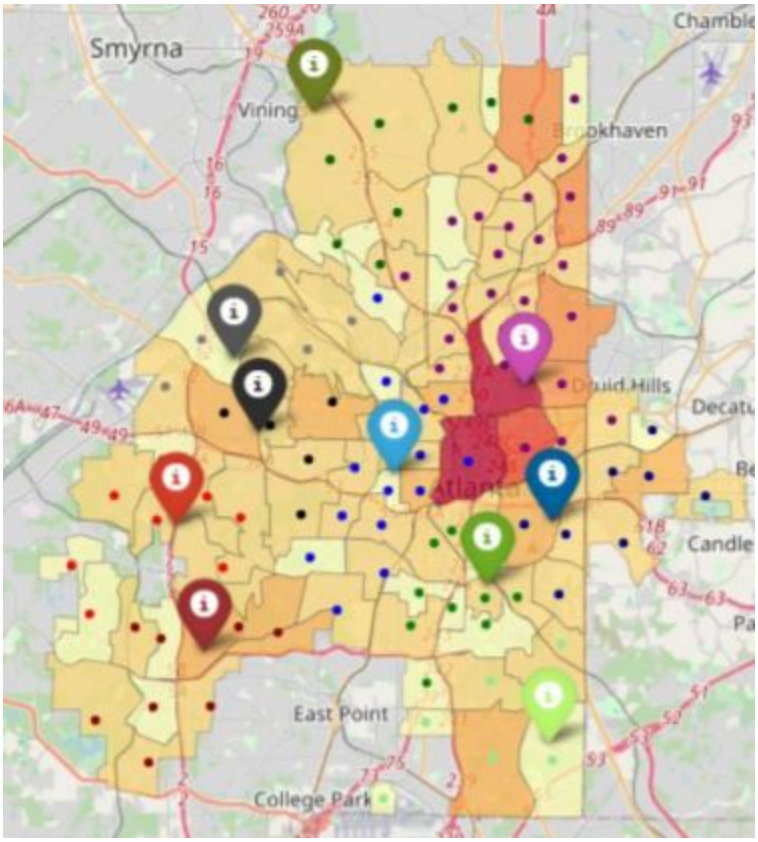
4. Analysis

4.1. We define a function plotPops() that plots the relevant neighborhood population data and a function plotSchools(schools) that plots the locations of the schools saved in the DataFrame argument "schools". These functions will be useful later when we need to see how our recommendations change school clustering. We begin to format our data into a useful DataFrame. We remove all irrelevant information and derive neighborhood centroids from the GeoJson data. The DataFrame we create has information on each neighborhood pertaining to its centroid latitude and longitude, population, and what school it is to be assigned to. We initialize the Assigned Schools to blank values. They will be updated later.



We create a DataFrame withProp which is the same as the previously created schools DataFrame, except that it now includes one new school called "Proposed Location. The location is initialized in Brookhaven, as this area intuitively seemed desolate of schools.

Functions are defined for updating school assignments for each neighborhood using the assign_members function and for plotting the neighborhoods on the map, using the assigned school to differentiate neighborhoods by color.

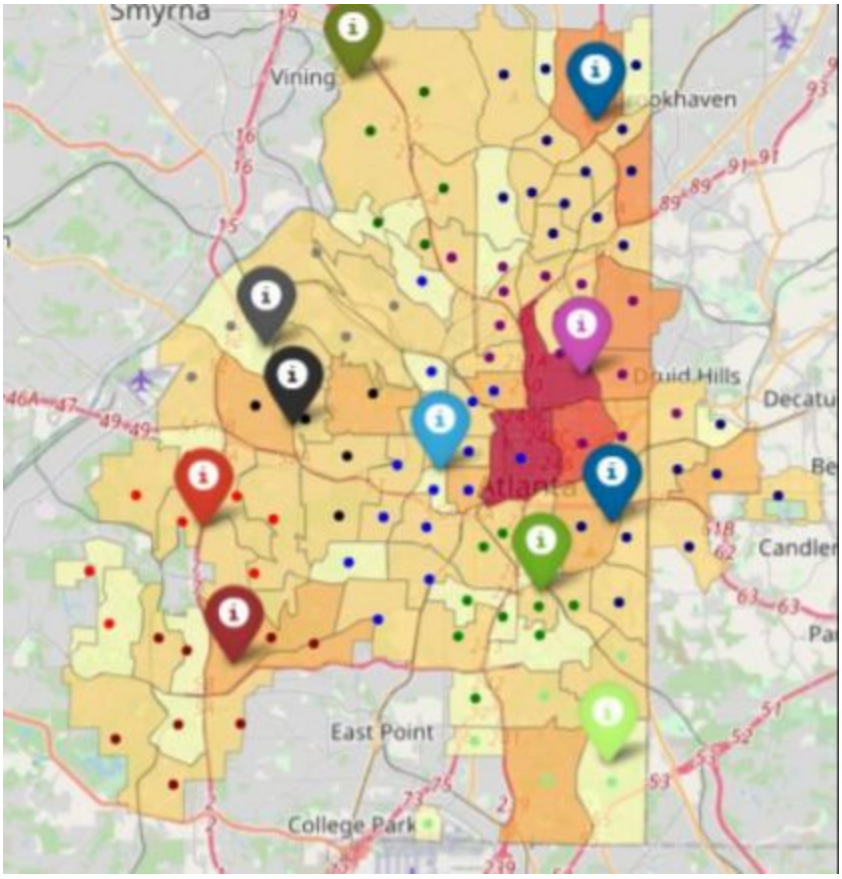


Now that neighborhoods have been clustered for school assignment, we would like to know the population each school must represent based upon these assignments. As we can see, there is an unequal distribution of the population of Atlanta represented by each school. In particular, Henry W. Grady High School must represent almost twice the population of the next largest school, Booker T. Washington High School. There is therefore cause for opening a new school

	Pop
Assigned School	
Benjamin Elijah Hays High School	27831.0
Booker T. Washington High School	65969.0
Carver High School	35005.0
Coretta Scott King Young Women's Leadership Academy	20479.0
Daniel McLaughlin Therrell High School	31189.0
Frederick Douglass High School	20451.0
Henry W. Grady High School	127585.0
Maynard H. Jackson High School	39539.0
North Atlanta High School	31203.0
South Atlanta High school	20752.0

We now use out previously defined functions to add the initial Proposed Location to the map and cluster based upon its current location. Of course, this location is initialized outside city limits and will therefore need to be updated.

	Pop
Assigned School	
Benjamin Elijah Hays High School	27831.0
Booker T. Washington High School	65969.0
Carver High School	35005.0
Coretta Scott King Young Women's Leadership Academy	20479.0
Daniel McLaughlin Therrell High School	31189.0
Frederick Douglass High School	20451.0
Henry W. Grady High School	107515.0
Maynard H. Jackson High School	39539.0
North Atlanta High School	20779.0
Proposed Location	30494.0
South Atlanta High school	20752.0



As we can see, the addition of this new school at its initial location already reduces the burden on Henry W. Grady by 20,000 students. We now create a function to move the location of the Proposed Location to the average coordinates of its cluster. We now iterate the update_newSchool function 500 times so that it may converge on an ideal cluster location.

	Pop
Assigned School	
Benjamin Elijah Hays High School	27831.0
Booker T. Washington High School	65969.0
Carver High School	35005.0
Coretta Scott King Young Women's Leadership Academy	20479.0
Daniel McLaughlin Therrell High School	31189.0
Frederick Douglass High School	20451.0
Henry W. Grady High School	84131.0
Maynard H. Jackson High School	39539.0
North Atlanta High School	17356.0
Proposed Location	57301.0
South Atlanta High school	20752.0

Now that a new location has been found for a school, we look at our population data to determine if this addition is successful in relieving the burden on Henry W. Grady. We can see that we are successful, having reduced Henry W. Grady's population representation by over 40,000 people. The new proposed location will represent approximately 57,000 people. Below, we determine the address for the new school.

```
3596, North Stratford Road Northeast, Buckhead, Atlanta,
```

5. Results and Discussion

5.1. Our purpose in this project was to determine the necessity and location of a new public high school in the Atlanta area. Our analysis shows that

while several high schools in Atlanta represent a manageable portion of the population, Henry W. Grady High School is in a highly populated area and must represent larger portions of the city than many of the other schools. Specifically, the north-east region of the city seems to be barren of public high schools.

5.2. In order to address this issue, we added a new public high school location and used K means clustering to find a location that would help alleviate this population burden. A location at or around 3596 North Stratford Road NW, Buckhead Atlanta was determined to be the suggested location. Of course, there may be unknown reasons why this area does not have a high school located nearby. There could be private high schools that represent the population, there may be schools outside city limits, or there may be low barriers to quick transportation to and from other high schools to this area. Therefore, recommendations made should be used as a starting point for further analysis.

6. Conclusion

6.1. The purpose of this project was to analyze Atlanta school and population information in order to make a recommendation to policymakers regarding the need for a new school within city limits. We identified highly population density neighborhoods and clustered them to their nearest high school. From their we determined which schools would have been overpopulated based upon these clusters. All this information in addition to general visual

distribution of the public schools led us to determine which area was most in need of a high school. The north east region of the city was chosen and a new proposed school location was determined using K means clustering. This process led to