Joshua Walker
Professor Teplovs
SI 330

Final Project Report

Motivation:
My project involves visualizing the 2017 4-year graduation rates of all public high schools in the state of Michigan by location. In a sociology class I took last year, we discussed the reasons and implications behind high/low graduation rates. I wanted to analyze this at a state level (with a particular focus on the metro Detroit area) to determine any concentration of successful/unsuccessful high schools. There are a couple questions I wanted to answer:

- Where are the most notable areas that exhibit a large disparity in graduation rates?
- Are schools that have very high or very low graduation rates isolated to certain geographical areas? Or spread somewhat evenly across the state?

Data sources:

- 2017 Michigan Graduation/Dropout Rate by District and School
    - Due to the redesign/restructuring of the Michigan School Data website, I am unable to locate the specific CSV file I downloaded. I have included the original CSV with just the first 1000 records in the zip file.
    - Size: 19.5MB
    - Format CSV -> pandas
    - Access method: via MI School Data website
    - Important variables/columns in dataset: BuildingName (name of high school), CountyName, GraduationRate
    - Contained no information about each school's location beyond what county it was in
- Retrieving the location (lat/long) of Michigan public high schools
    - Access method: Google Places API Text Search Service, using request URLs containing queries of school name, county, and state
    - Return format: JSON
    - Important variables contained: location latitude and longitude
    - Number of records queried/retrieved: 1089 queried, 1058 retrieved

Data Manipulation Methods:
Michigan School Data (2017 Michigan Graduation/Dropout Rate by District and School):

All of this data originated as a CSV so I first loaded it into a pandas dataframe.

The original dataset contained much more information than what I needed so a lot of the manipulation was related to only getting the data that I needed. It broke down the graduation rates per school by race/gender, so I filtered this to include data for all students. Data was given based on 4, 5, and 6 year graduation rates, but I was only interested in 4-year rates. The first set of rows was dedicated to showing rates at a statewide basis but I was not interested in this data so they were removed.

I discovered there was missing data in the set-- some rows had “0000” as the building name, and some rows showed graduation rates of 0%. I made a safe assumption that no school has a 0% graduation rate and removed these rows.

I did these manipulations through dataframe selection/recreation as we have done in class. For example: `df = df[df.RateYear == '4-Year']` to create a dataframe from the existing one that only showed 4-year graduation rates.

To prepare for my API query, I had to adjust the `BuildingName` for each school. They initially each had trailing characters that contained spaces, paraenthesis, and a 5-digit number. To clean this up, I used:
`df = df.assign(BuildingName_clean=df.BuildingName.str[:-8])`
to remove these trailing characters and get just the name of the building as a string.

At the end of this, I cleaned up the dataframe to only show the necessary columns:
`CountyName, BuildingName_clean, GraduationRate.`

<u>Google Places API Text Search and dataset combination:</u>
Now that I had the data from the first set ready, I had to get the correct queries together that would be sent to the Google Places API. The Places API works just like a search in Google Maps: you type in a text query about a location/building name and and list of results are returned. For the API, the returned data is formatted as JSON and includes information such as address, icon, rating, type, and location. I was most interested in the location data, which is returned in an object of latitude and longitude. This is the information I wanted so I could create a geographic heat map.

I made a for loop that went through each row in the graduation rate dataframe to query the Places API. I formatted the text query so that it was “{name}, {county}, MI.” If at least

one result was found, the latitude and longitude from the first result were placed in two new columns corresponding to lat/long. Here's a snippet of the code for this loop:

```
for i in df_cleaned.index:
    name = df_cleaned.loc[i, 'BuildingName_clean']
    county = df_cleaned.loc[i, 'CountyName']
    query = f'{name}, {county} County, MI'
    r = requests.get(f'https://maps.googleapis.com/maps/api/place/textsearch/json?query={query
    res = r.json()
    if res['results']:
        df_cleaned.loc[i, 'latitude'] = str(res['results'][0]['geometry']['location']['lat'])
        df_cleaned.loc[i, 'longitude'] = str(res['results'][0]['geometry']['location']['lng'])
    else:
        df_cleaned.loc[i, 'latitude'] = '0'
        df_cleaned.loc[i, 'longitude'] = '0'
```

I had to add error catching because sometimes there was nothing returned. I assigned a lat/long of 0/0 for these which were later filtered out. Of the 1089 queries made, 1058 returned valid latitude/longitude values. I was worried that Google could have returned a location that ended up being way off, like out of state. The heat map that I created confirmed that every result was at least in Michigan. There is a possibility that the location returned for some schools still may not have been accurate, but I am confident that this only happened a couple times, if at all. The only instance this probably could have occurred is if there were multiple high schools with a very similar name in the same county.

The resulting dataframe is the complete, combined dataset. Here is the head of that dataframe:

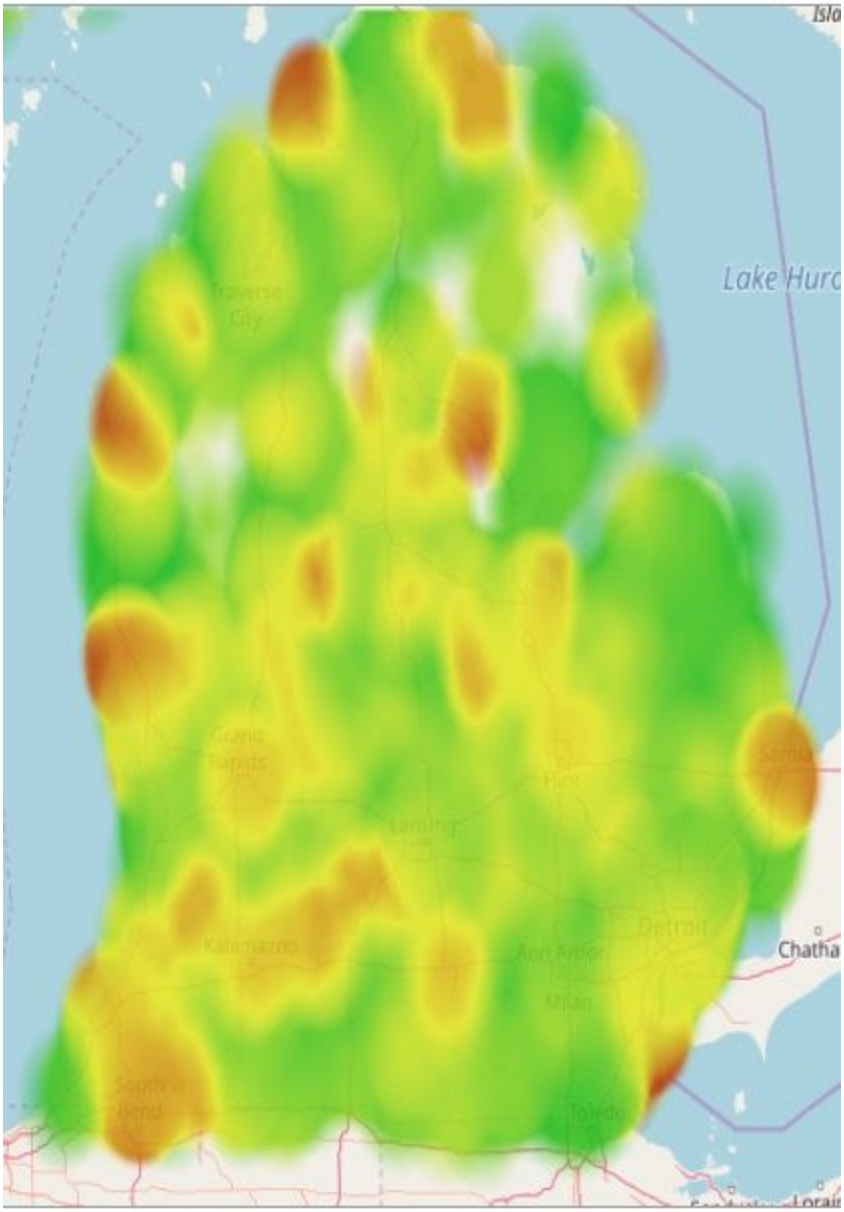
	CountyName	BuildingName_clean	GraduationRate	latitude	longitude
0	Alcona	Alcona Community High School	91.38	44.6572637	-83.405481
1	Alger	Burt Township School	100.00	46.6683084	-85.9787546
2	Alger	Munising High and Middle School	93.02	46.4173573	-86.6626345
3	Alger	Superior Central School	85.71	46.3532337	-86.9717661
4	Allegan	Plainwell High School	96.89	42.436614	-85.65339449999999

The final manipulation was renaming the building name and graduation rate columns to name and value so that it was in a format that OpenHeatMap understood.

Analysis and Visualization:

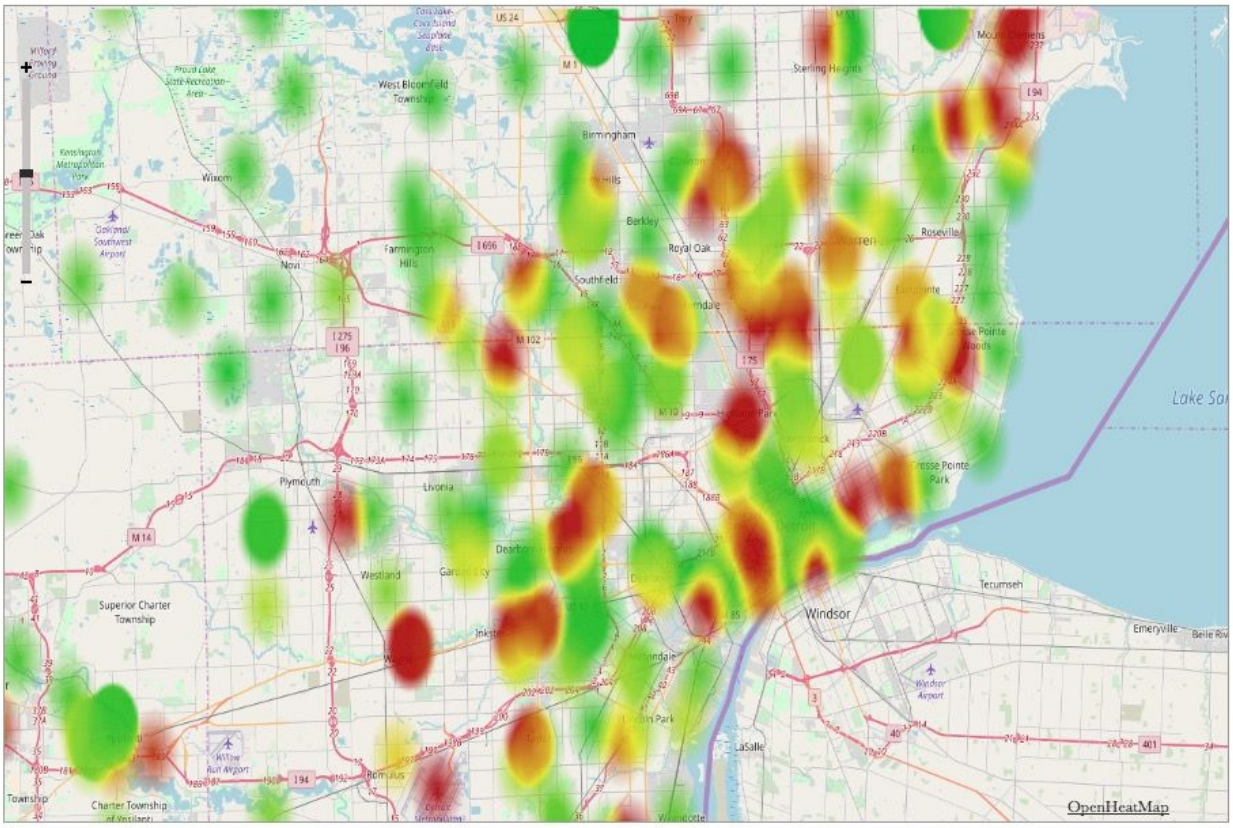
For my analysis and visualization I used OpenHeatMap. I converted my final dataframe to a CSV file that was imported through the OpenHeatMap webapp. I used a scale of 35-100% and used traditional “heat” colors on a scale of red, yellow, and green; red represented the lowest rates and green the highest.

Here’s the visualization of the Lower Peninsula. The Upper Peninsula is sparsely populated so I focused on analysis of the LP.



Overall, the graduation rates seem spread fairly evenly across the state. I adjusted the "blob size" until I was able to make note of distinct areas of low performing schools. Of these, I determined the area known as downriver Detroit to be the worst, followed by Ludington/Manistee and Charlevoix/Petosky.

I am from southeastern Michigan so I was interested in analyzing the map of the metro Detroit area specifically, which happens to also be the highest concentration of high schools in the state.



There are many more red areas than I would have expected. My hypothesis was that the city limits of Detroit would be red/yellow and the surrounding areas green. There are many "pockets" of dark red surrounded by bright green, corresponding to very low and very high graduation rates. There is a large disparity in graduation rates in many areas throughout Detroit and its near suburbs.

My hypothesis was based on my own knowledge and perception of the wealth of communities in SE Michigan, and while some of the results were surprising, other confirmed my guesses. In the areas northwest of Detroit (like Farmington Hills, West Bloomfield, Novi, Northville) graduation rates are consistently very high.

There is a pitfall in the generation of the heat map that I noticed upon close inspection. I discovered that some small, alternative schools have very low graduation rates and that this can misrepresent the rest of the area immediately surrounding it. I don't believe this significantly altered the results but it is certainly occasionally present.

Conclusion:

Overall, I found this project to be very interesting and I was surprised by some of the results. I was able to identify and analyze a number of areas in Michigan with clusters of low performing high schools. I believe this visualization could be useful for the Michigan Department of Education to determine key areas where action could be taken at a level other than school or district-specific.