Hyper-Localized Analysis of New York City's Lead Poisoning Epidemic

Data Science and Analytics

Virtual SLC
2020-2021
PID: 205101

Table of Contents

Introduction

Brooklyn's lead poisoning is (7) times higher than Manhattan's and 12 times higher than Staten Island's, showing the ineffectiveness of New York City's decades-long attempts to identify and solve factors relating to disproportionate lead-poisoning (Culliton, 2019). New York City's lead poisoning management team needs comprehensive, localized data analysis that provides a geographical and personal profile for individuals who are at particular risk of lead poisoning. Our analysis seeks to magnify each New York City borough by analyzing which neighborhoods are disproportionately impacted, and identifying the root causes of such disproportionality. By specifically narrowing down high-risk neighborhoods and individuals, New York City can save time and resources while canvassing homes for lead poisoning more effectively. All in all, the holistic goal of this project is to facilitate New York City's campaign to eradicate lead poisoning among minors by identifying the most high-risk populations, based on several corroborating factors.

Purpose

Exposure to these high levels of lead poisoning can lead to life-long implications such as permanent memory loss, tremors, and hallucinations (Schneyer & Pell, 2017). Over the past two (2) decades, New York City has dramatically fought its lead poisoning epidemic; however contemporary studies argue that the rate at which lead poisoning is decreasing has stagnated, severely affecting upwards of eleven-thousand children (Ferré-sadurni, 2019). As New York City's comptroller, Scott Stringer, has argued that "any lead poisoning of our children must be treated as a five-alarm fire, but the city isn't utilizing basic tools," New York City's lead poisoning management team has been unable to effectively extinguish lead poisoning, as in 2019 alone, 9,099 residences went unchecked for lead poisoning (Ferré-sadurni, 2019). A breadth of factors contributes to New York City's lead poisoning—old housing, industrial waste, and lead-ridden soil. To fight this multifactorial epidemic, the city has no room for overlooking critical residences in their campaign to eliminate lead poisoning. New York City's canvassing protocols can be vastly improved by clearly defining the most pertinent sections of the city. By analyzing New York City's children < 6 lead poisoning data with corroborating factors (income, poverty rate) and on a scale as small as neighborhoods, we seek to elucidate the high for correlations within specific boroughs. Additionally, external factors (income, poverty rate, water consumption, etc.) can be appended to analyze the damage of certain factors on elevated lead blood levels within children under the age of six.

Methods

1. NYCOpenData: We utilized New York City's open-source database to sift through datasets concerning lead poisoning among children in New York City. By conditioning datasets for relevance and the amount of redacted data, we utilized datasets that were recent (within the past decade) and also accounted for many different variables that may create outliers within the data.

2. Point2Homes: Point2Homes provides small-scale data on several neighborhoods throughout New York City, making it convenient for finding statistics on poverty rate, average income, and median income for the respective neighborhood.

3. NYC Open Data Maps: NYCOpenData provides numerous shapefiles and GeoJSON mappings of New York City's layout. To construct geographical maps, we imported NYC shapefiles that contain information about the city's neighborhoods and their polygonal bounds respective to each other.

Principal Component Analysis: Principal Component Analysis is used to reduce a data set with many variables into one with only 2 components. First, it uses elimination to standardize the data to two different axes. Then, it scales and shapes that data according to a matrix to create covariances in the data.

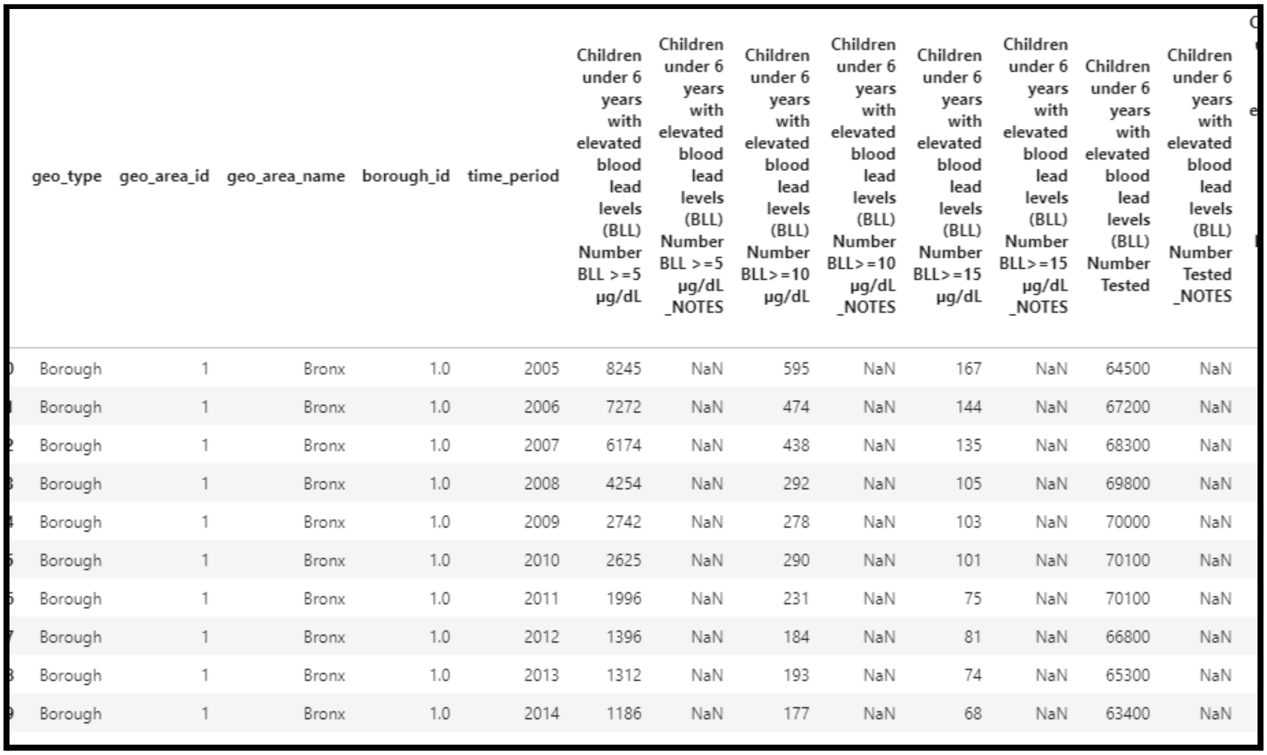
Heatmap: Heatmaps are one of the best ways to visualize data on a map in a way that allows viewers to easily understand how the data connects to a geographical context.

Scatter Plots: Scatter plots do a good job of demonstrating data of two numeric values and with a lot of variance in data points. It makes it easy to read and understand what the data is showing and the correlation between the two factors.

Line Plots: Line plots were used in this analysis to apply regression algorithms and observe future trends. The most notable example is the graph displaying New York City's lead poisoning cases over time, with an accompanying exponential regression.

Distribution Plots: A distribution plot is used to show how the data is spread across a certain value. This two-part graph demonstrates how close a certain value actually is to the hypothesized line. For our project, this analyzed the distribution of lead poisoning in the boroughs/neighborhoods in New York City and how it related to the poverty rate using a bell curved line and outliers in an interval.

Results



	geo_type	geo_area_id	geo_area_name	borough_id	time_period	Children under 6 years with elevated blood lead levels (BLL) Number BLL >=5 µg/dL	Children under 6 years with elevated blood lead levels (BLL) Number BLL >=5 µg/dL _NOTES	Children under 6 years with elevated blood lead levels (BLL) Number BLL>=10 µg/dL	Children under 6 years with elevated blood lead levels (BLL) Number BLL>=10 µg/dL _NOTES	Children under 6 years with elevated blood lead levels (BLL) Number BLL>=15 µg/dL	Children under 6 years with elevated blood lead levels (BLL) Number BLL>=15 µg/dL _NOTES	Children under 6 years with elevated blood lead levels (BLL) Number Tested	Children under 6 years with elevated blood lead levels (BLL) Number Tested _NOTES
0	Borough	1	Bronx	1.0	2005	8245	NaN	595	NaN	167	NaN	64500	NaN
1	Borough	1	Bronx	1.0	2006	7272	NaN	474	NaN	144	NaN	67200	NaN
2	Borough	1	Bronx	1.0	2007	6174	NaN	438	NaN	135	NaN	68300	NaN
3	Borough	1	Bronx	1.0	2008	4254	NaN	292	NaN	105	NaN	69800	NaN
4	Borough	1	Bronx	1.0	2009	2742	NaN	278	NaN	103	NaN	70000	NaN
5	Borough	1	Bronx	1.0	2010	2625	NaN	290	NaN	101	NaN	70100	NaN
6	Borough	1	Bronx	1.0	2011	1996	NaN	231	NaN	75	NaN	70100	NaN
7	Borough	1	Bronx	1.0	2012	1396	NaN	184	NaN	81	NaN	66800	NaN
8	Borough	1	Bronx	1.0	2013	1312	NaN	193	NaN	74	NaN	65300	NaN
9	Borough	1	Bronx	1.0	2014	1186	NaN	177	NaN	68	NaN	63400	NaN

We started our data analysis with a dataset that documents the number of positive lead poisoning tests in a sample size of New York City children under six each year in all boroughs and numerous neighborhoods. Each neighborhood documents the number of children that had levels of blood levels of lead poisoning (BLL) above 5 - 15 µg per each year between 2005-2016. In this data analysis, we focused on BLL greater or equal to 5 because the CDC declared 5 µg of lead in the blood as the minimum amount needed for considerable developmental threats.

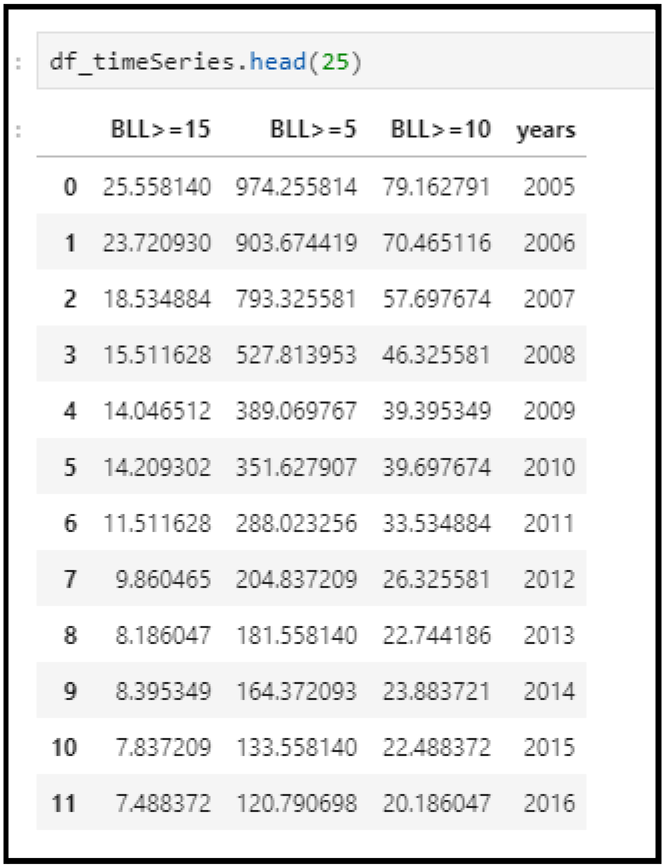
Data Engineering Algorithm #1 (requires Repl.it)

Real-world data is not pre-processed for data analysis. These three operations depicted in the code help us extract some context of NYC's lead problem.

1. First, we removed New York City from the "geo_area_name" column.
    a. New York City is not a borough. This set of 12 data points represents holistic data, which is not helpful for a localized analysis.
2. Second, we averaged the data points of BLL >= 5 Per Year

  a. For example: Collect each data point with "time_period" 2005. Average these data points.

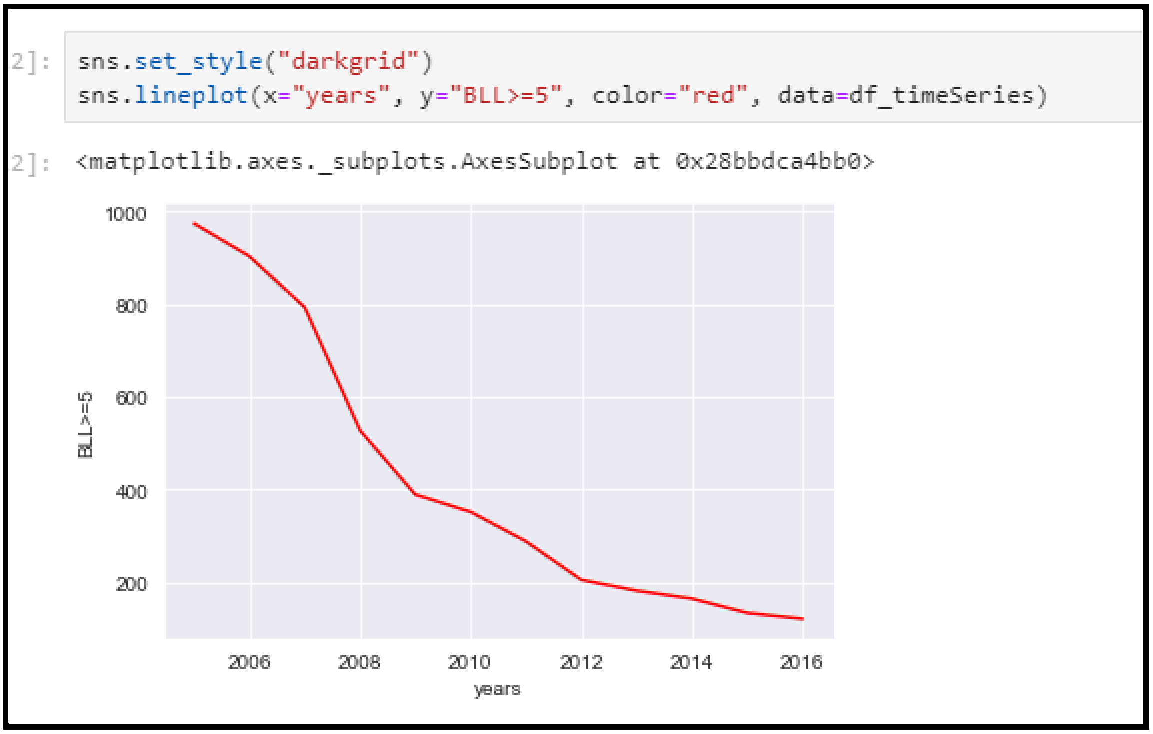
3. Last, we created a new dataframe (table) representing a “time series.”
  a. This allows us to observe lead poisoning rise/decline/stagnancy over time



```
df_timeSeries.head(25)
```

	BLL>=15	BLL>=5	BLL>=10	years
0	25.558140	974.255814	79.162791	2005
1	23.720930	903.674419	70.465116	2006
2	18.534884	793.325581	57.697674	2007
3	15.511628	527.813953	46.325581	2008
4	14.046512	389.069767	39.395349	2009
5	14.209302	351.627907	39.697674	2010
6	11.511628	288.023256	33.534884	2011
7	9.860465	204.837209	26.325581	2012
8	8.186047	181.558140	22.744186	2013
9	8.395349	164.372093	23.883721	2014
10	7.837209	133.558140	22.488372	2015
11	7.488372	120.790698	20.186047	2016

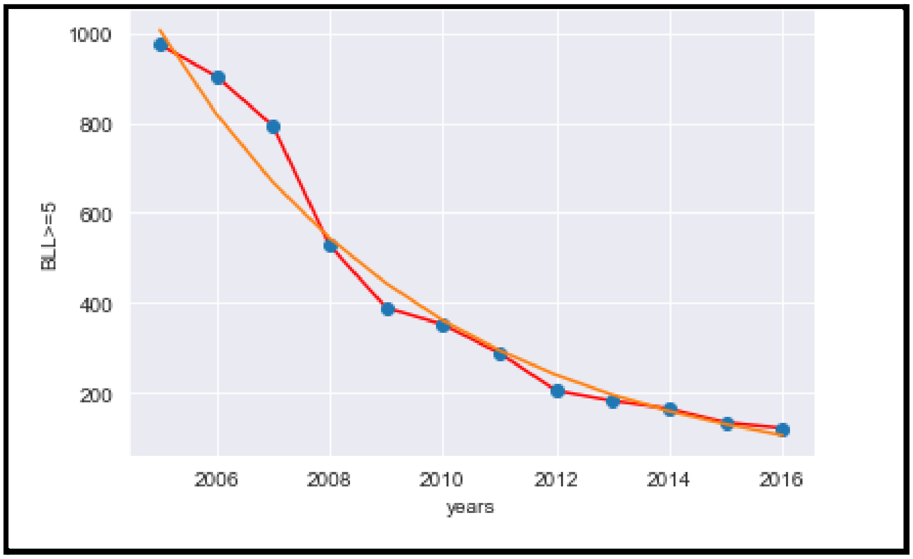
This time series data frame has the average number of children with a specified BLL for every year (2005-2016).



```
2]: sns.set_style("darkgrid")
    sns.lineplot(x="years", y="BLL>=5", color="red", data=df_timeSeries)

2]: <matplotlib.axes._subplots.AxesSubplot at 0x28bbdca4bb0>
```

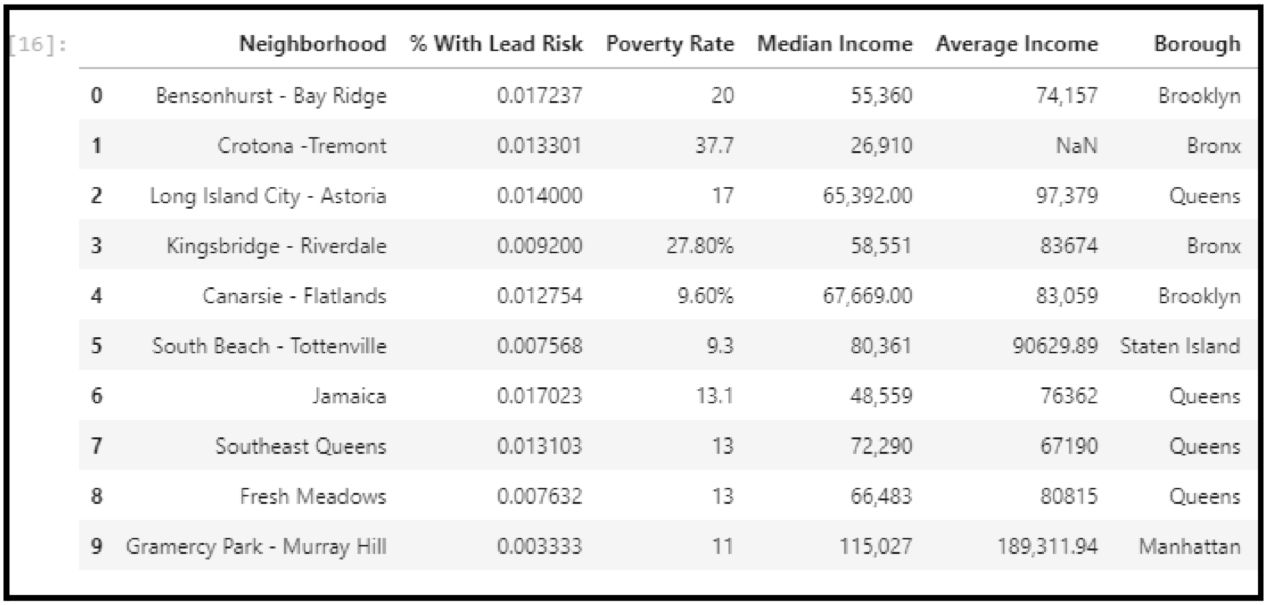
We are only interested in children with a BLL >= 5. As such, we graphed the children with a BLL >= 5 over the course of 2005 - 2016. The data indicates that NYC has done a tremendous job handling the lead epidemic from 2005-2012. However, their progress has significantly slowed in recent years.



After fitting the data onto an exponential curve, the flattening of the decrease in lead poisoning cases becomes more evident. The flattened curve between 2014-2016 marks that NYC's lead poisoning epidemic is now flattening, but not dying.

Data Engineering #2 Code (requires Repl.it)

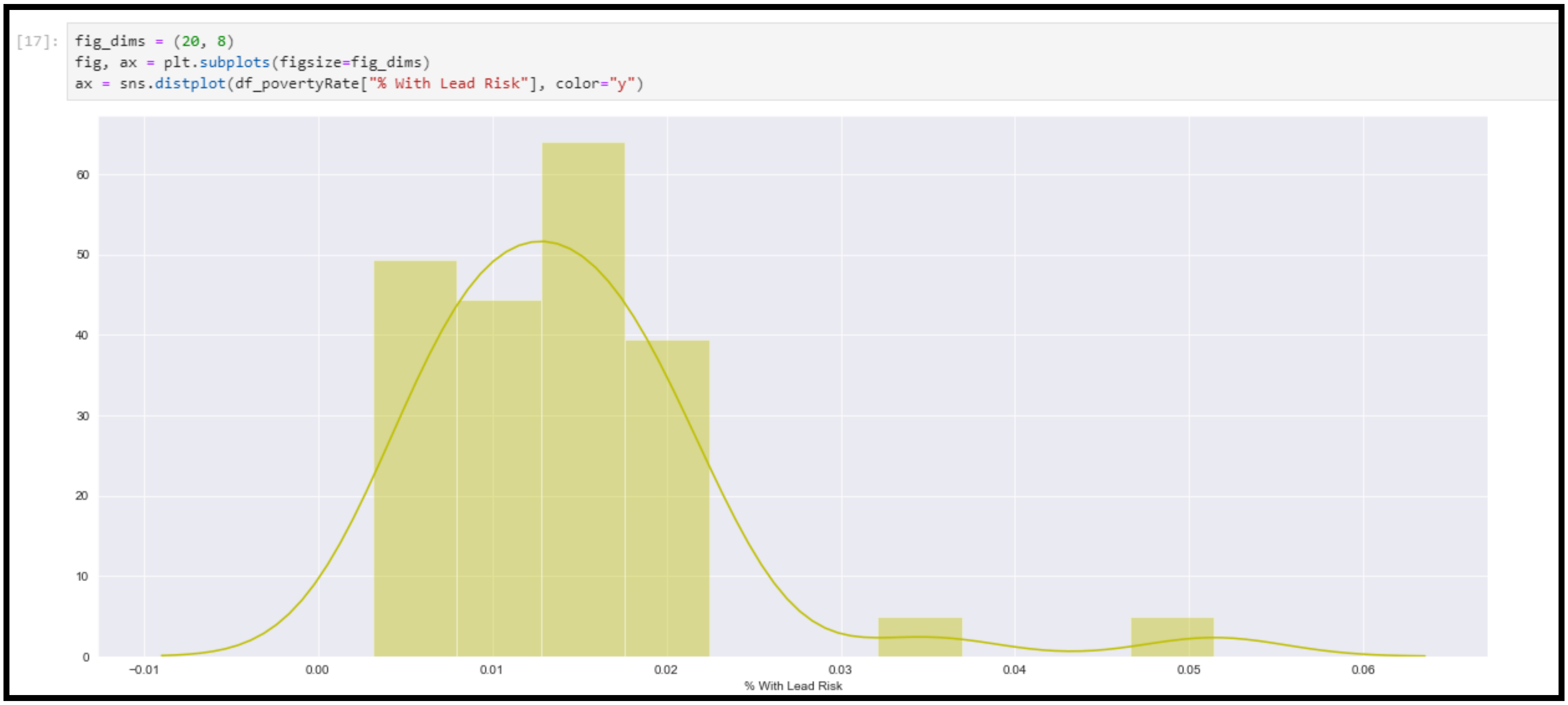
This section of data engineering converts each number of poisoning cases within a certain neighborhood to a percentage of its population. This conversion allows us to meaningfully and ethically compare the level of lead poisoning per neighborhood.



[16]:

	Neighborhood	% With Lead Risk	Poverty Rate	Median Income	Average Income	Borough
0	Bensonhurst - Bay Ridge	0.017237	20	55,360	74,157	Brooklyn
1	Crotona -Tremont	0.013301	37.7	26,910	NaN	Bronx
2	Long Island City - Astoria	0.014000	17	65,392.00	97,379	Queens
3	Kingsbridge - Riverdale	0.009200	27.80%	58,551	83674	Bronx
4	Canarsie - Flatlands	0.012754	9.60%	67,669.00	83,059	Brooklyn
5	South Beach - Tottenville	0.007568	9.3	80,361	90629.89	Staten Island
6	Jamaica	0.017023	13.1	48,559	76362	Queens
7	Southeast Queens	0.013103	13	72,290	67190	Queens
8	Fresh Meadows	0.007632	13	66,483	80815	Queens
9	Gramercy Park - Murray Hill	0.003333	11	115,027	189,311.94	Manhattan

Along with corroboration of data on poverty rate, median income, and average income, this dataset represents the output of the previous data engineering: each neighborhood matched up with its percentage of population with lead poisoning complications [in children]. This dataset contains 43 neighborhoods across New York City's five boroughs. The borough is indicated by the 'borough' column.

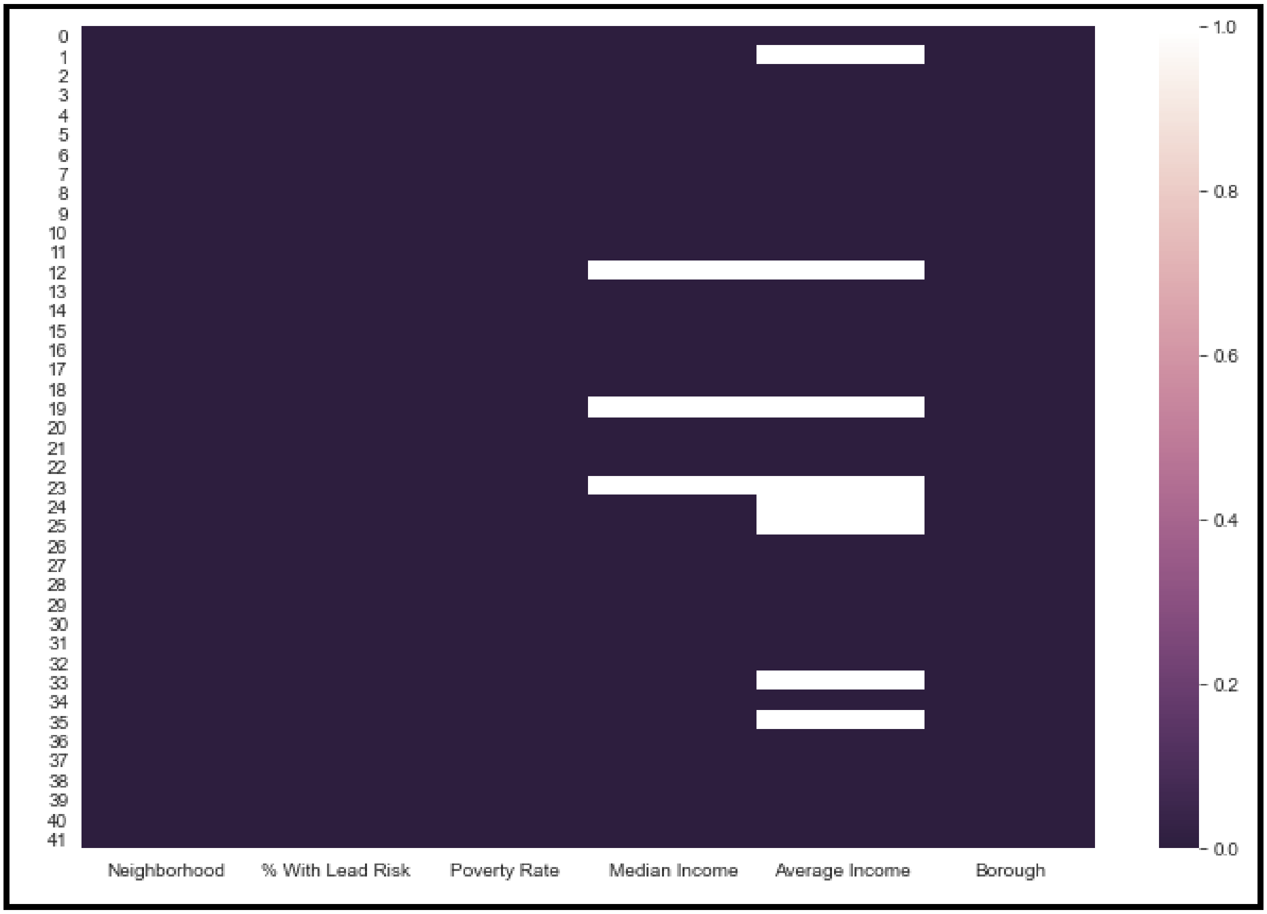


```
fig_dims = (20, 8)
fig, ax = plt.subplots(figsize=fig_dims)
ax = sns.distplot(df_povertyRate["% With Lead Risk"], color="y")
```

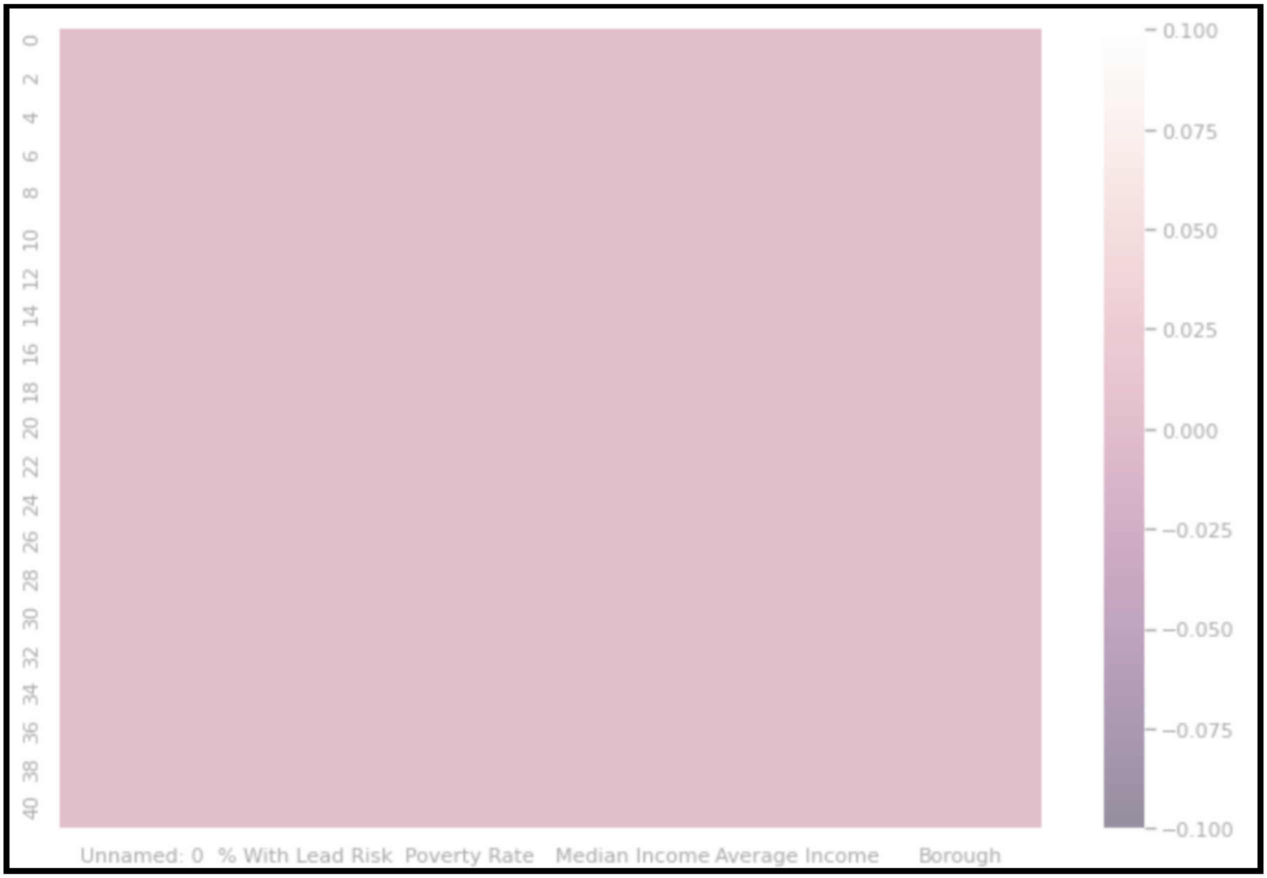
This plot graphs the distribution of the neighborhood % Lead Risk. The bell curve lies between the intervals [0%, ~2.5%]. This graph confirms the presence of outliers, indicated by the humps in this dataset lie beyond 3% on the x-axis.

Data Engineering #3 (requires Repl.it)

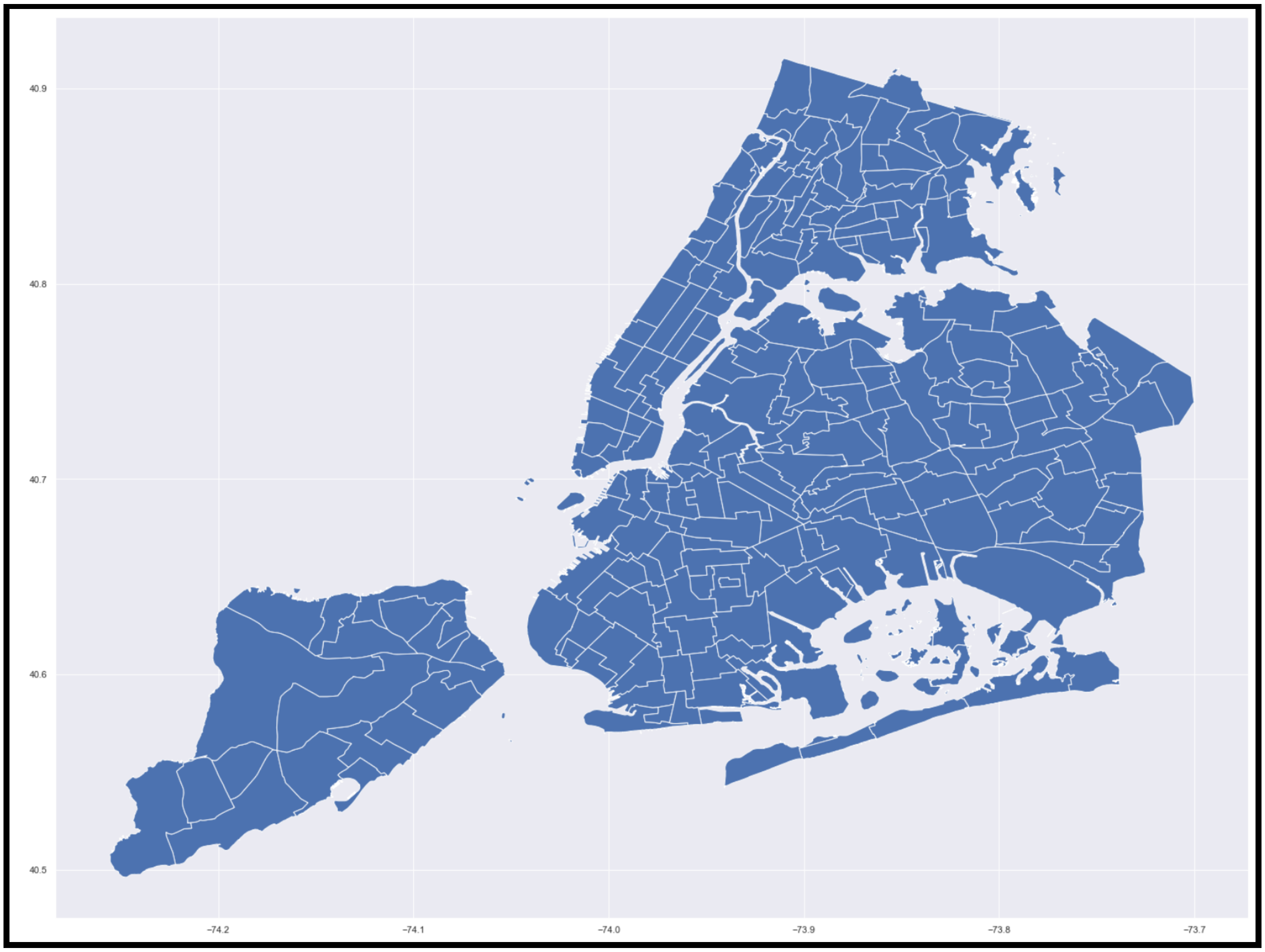
This section of data engineering was more simple. The dataset depicted above had some missing values. First, we graphed a heatmap to observe which columns lacked the most data.



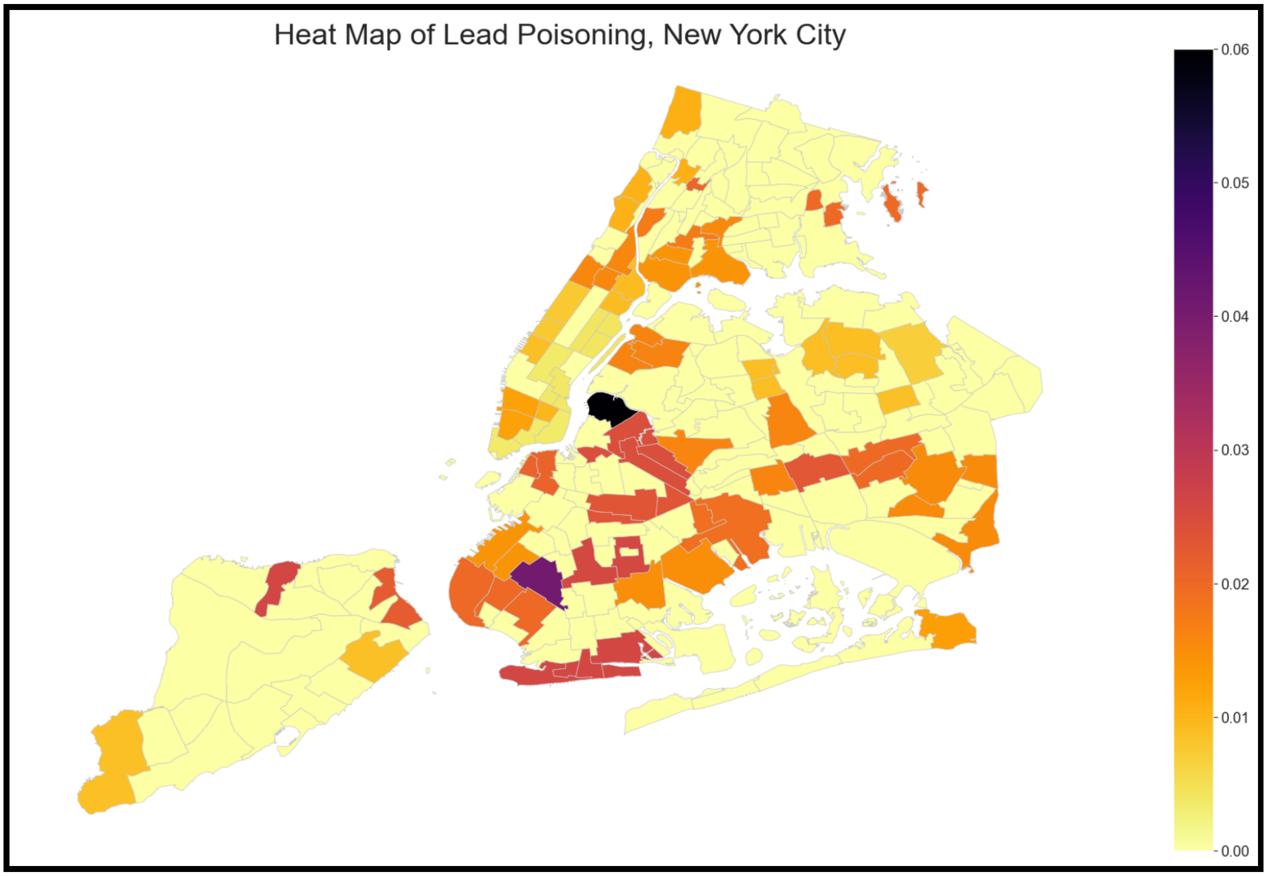
There is a sizable portion of missing data in the columns "Average Income" and "Median Income." Missing data is denoted by white patches within each column. Purple confirms that adequate data exists within the column.



After completing more research, we appended these data points into our dataset. The resultant heat map appeared as so, confirming we sufficiently filled in the sources of missing data. Now that we have complete, accurate, and comparable data, we can visualize NYC's lead risk geographically.



Using data embedded with shapefiles, we plotted New York City partitioned according to its 259 neighborhoods.



We plotted the % Lead Risk [range : 0% to ~0.06%] on the heatmap. The bright-yellow neighborhoods represent areas where there was no data on lead poisoning risks.

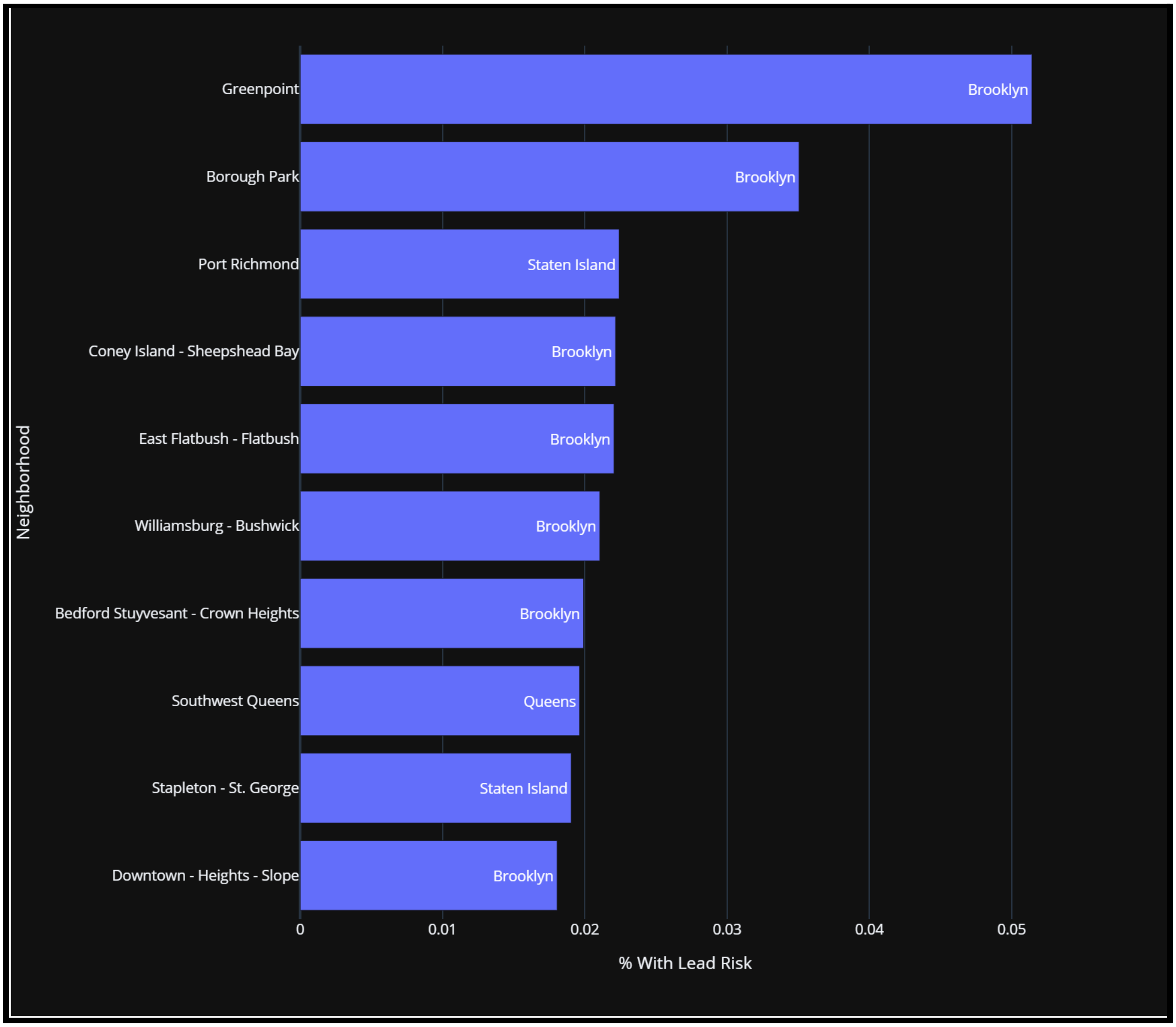
This graph depicts some notable disparities in lead poisoning throughout New York City. Brooklyn is disproportionately hit hardest by lead poisoning cases, with dense pockets of neighborhoods that lie on the more dangerous-side of the spectrum scattered throughout central and western Brooklyn.

Other than Brooklyn, South-Central and East Queens both elucidate a belt of high-risk areas within the borough. In Manhattan, the borough with the fewest number of high-risk neighborhoods, also has three neighborhoods in its north (The Harlems) that have medium to high risks of lead poisoning.

Since Staten Island was underrepresented in the dataset, only a few neighborhoods were included in the heatmap. Our team observed Staten Island's North Shore and noticed all three neighborhoods had concerning levels of lead poisoning prevalence. After researching lead poisoning in Staten Island, we discovered that children in Staten Island's North Shore are among the most affected populations by lead poisoning in New York City.

In essence, the geographic lead poisoning areas of interest we identified from this heatmap include: Manhattan's Harlems, the majority of Brooklyn, Southeast Queens, Staten Island's North Shore, and South Bronx. These hotspots serve as a guide that New

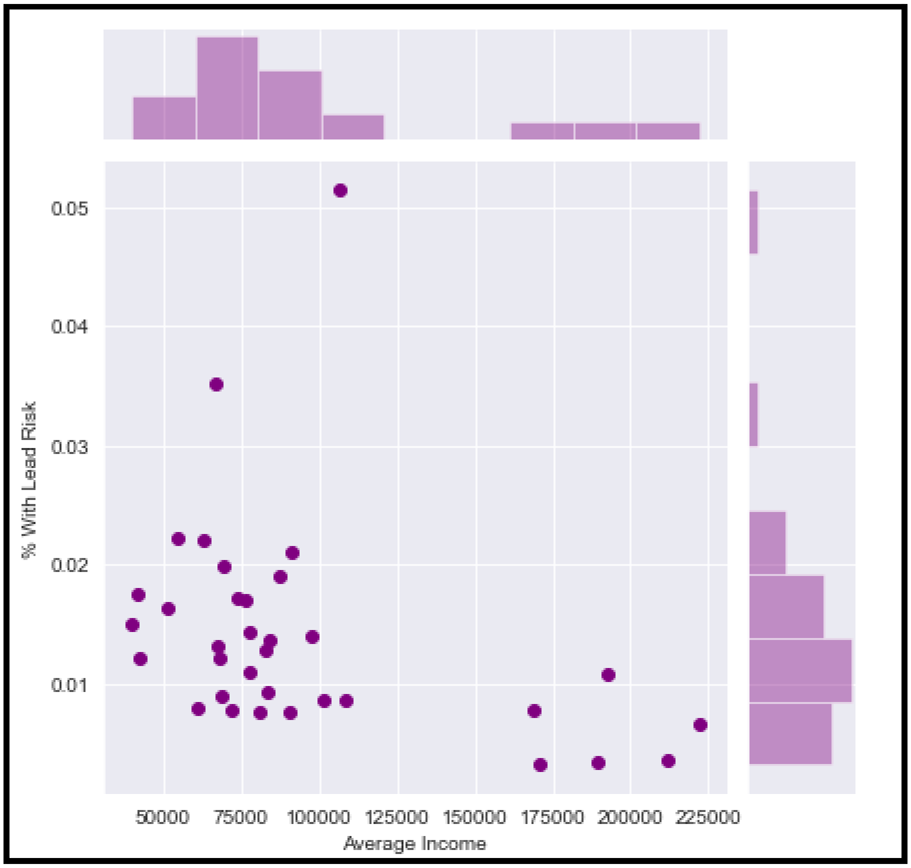
York City's government must follow to target the areas in which lead poisoning affects the most children.



To increase the scope of the analysis and observe which neighborhoods in particular fall in the top ten most affected areas in New York City, we created a labeled bar graph. Brooklyn has seven out of the ten most affected neighborhoods, with Staten Island following in with two neighborhoods and Queens with one.

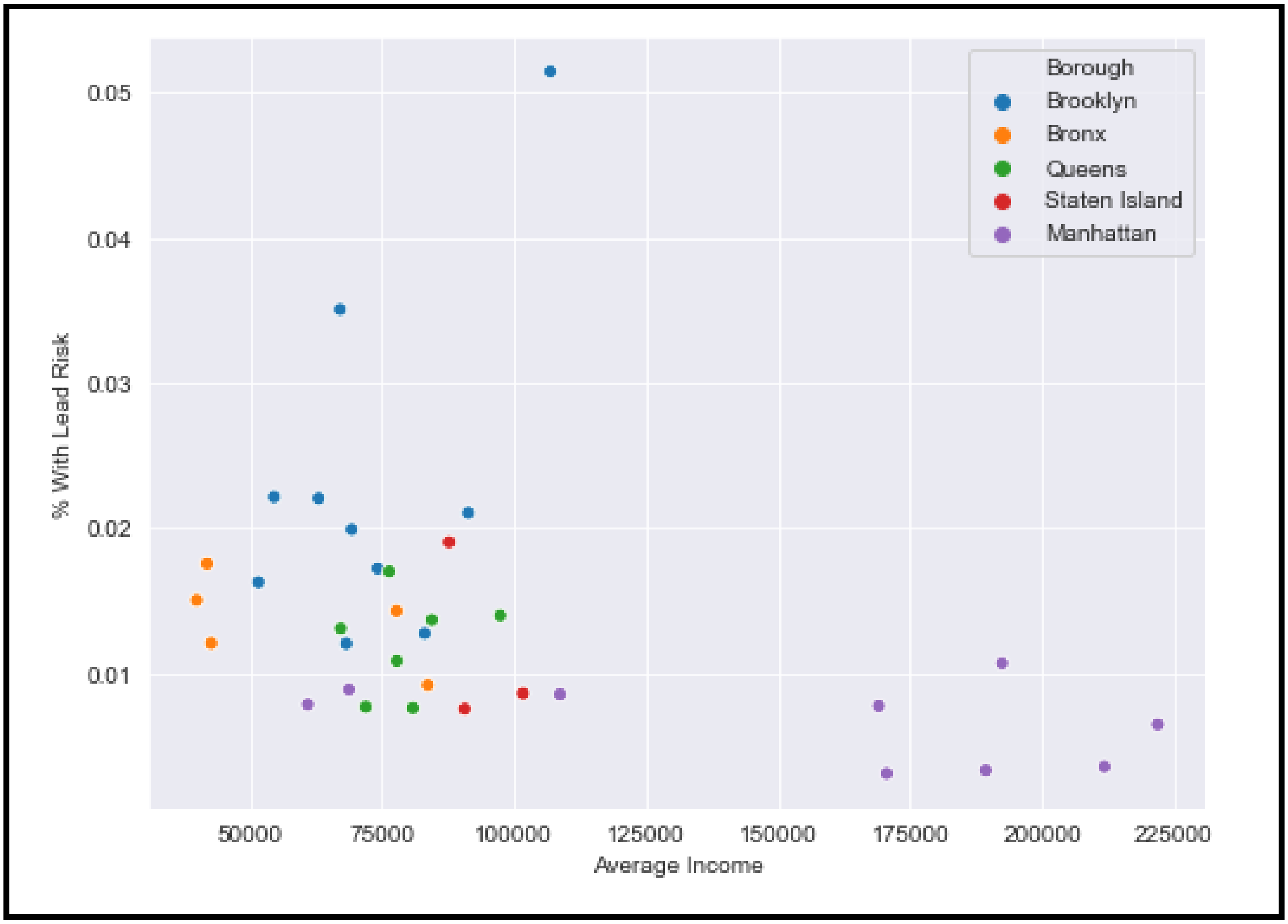
This data adds further depth to the geographical analysis because it solidifies that Brooklyn's is overrepresented in the realm of child lead poisoning cases. The next step is to observe if poverty rate and average/median income influence these disproportions associated with Brooklyn and smaller sections of NYC neighborhoods (Southeast Queens, Harlem, etc).

Data Synthesis: Poverty, Income, and Lead Poisoning:

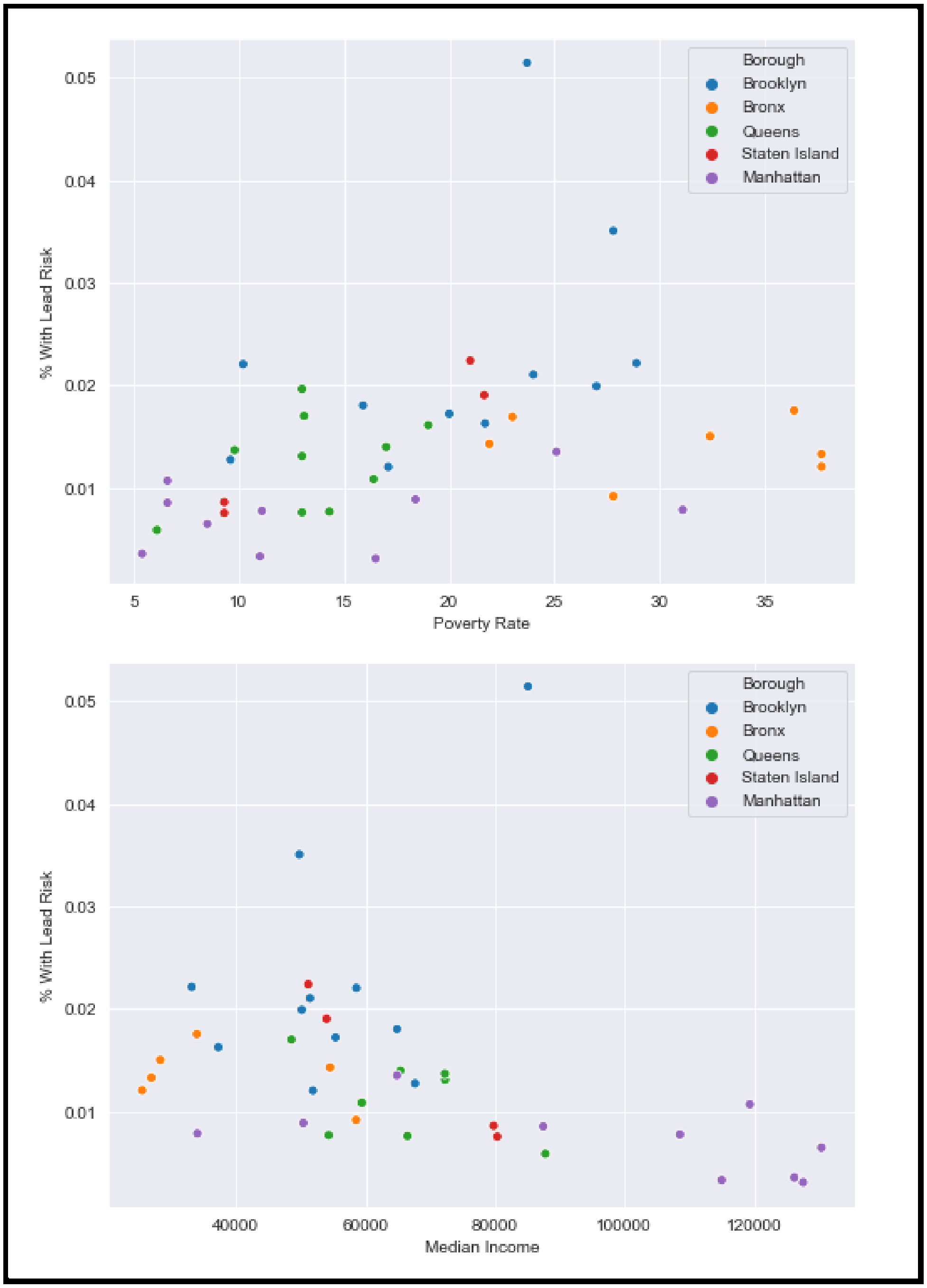


X

In reference to Figure X, we plotted the '% Lead Risk' and 'Average Income' against each other. Although the data points express no significant relationship, there was a distinctive feature in this graph—clusters. The lefternmost cluster has an extremely high average income and a low overall lead poisoning risk. We speculated that this cluster could be the affluent, lead-safe neighborhoods of Manhattan. As such, we graphed the subsequent with a color code to observe clusters through the scope of groups of neighborhoods and boroughs.

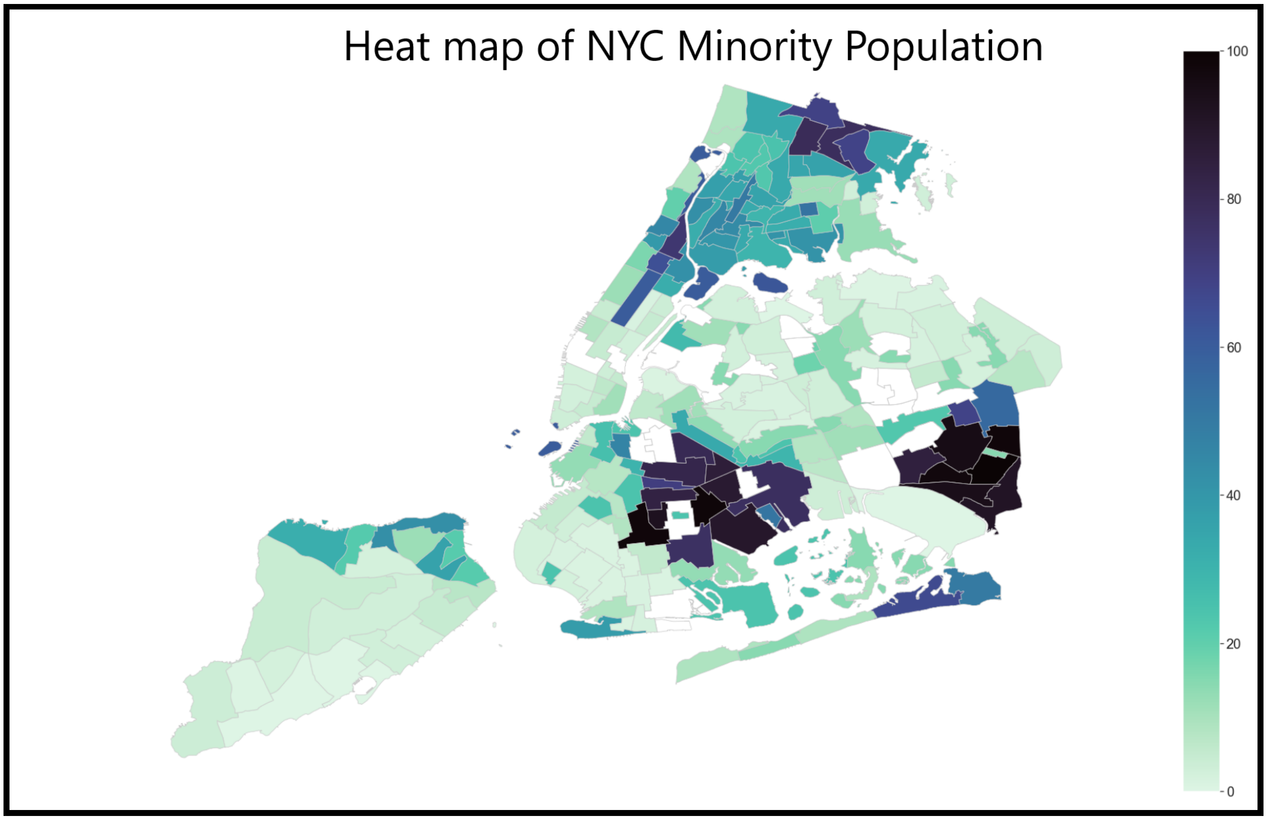


This graph is the same as the prior one, with the exception that it is now colored coded according to borough. There are a few distinct clusters within the graph, most notably the lefternmost cluster that is confirmed as belonging to Manhattan. There is also a cluster of blue points representing Brooklyn that demonstrate low average income and relatively higher lead risk. From this graph, we can clearly identify two clusters belonging to Brooklyn and Manhattan that observe opposite characteristics. Between these two clusters, there is also a patch of green points above the 0.01 threshold, representing Southeast Queens and its middle-class yet lead-struck neighborhoods.

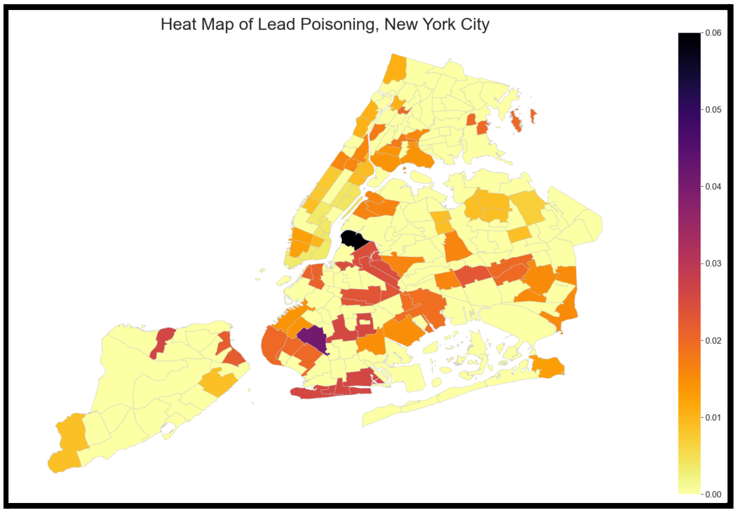


These two graphs demonstrate ‘Poverty Rate’ and ‘Median Income’ VS. ‘% Lead Risk.’ Brooklyn and Manhattan still inhabit distinctive regions in the graph, telling us that there are certain disproportions between NYC’s boroughs.

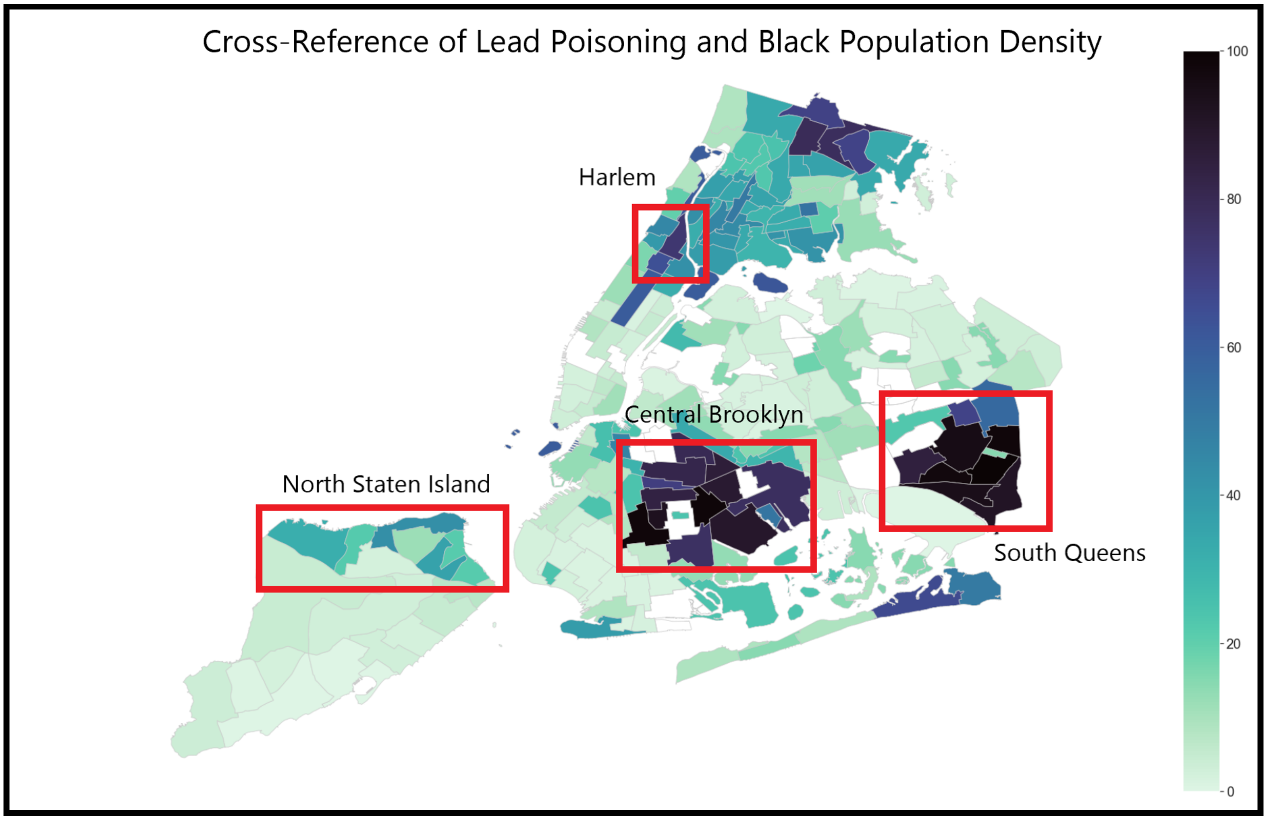
Race Demographics Analysis



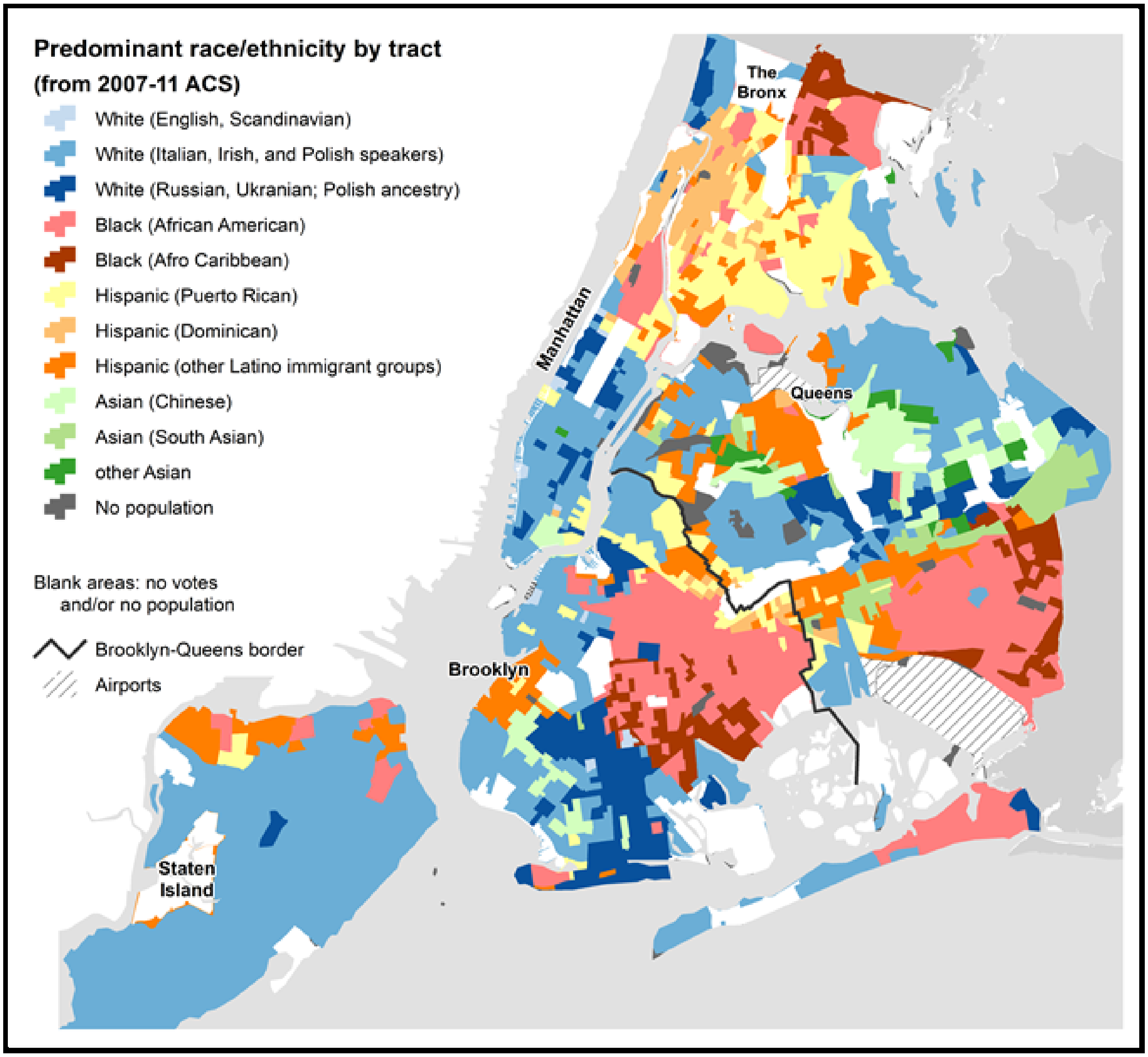
This graph marks a third, tiny section of our analysis where we wanted to explore the possibility of NYC's black community being disproportionately affected by lead poisoning. This heatmap represents neighborhoods with their color associated with the percentage of black residents.



Compare this heatmap with the heatmap concerning lead poisoning in New York City. Specific areas such as Brooklyn, Southeast Queens, Harlem, and Northern Staten Island on the demographics heatmap are consistent with hotspots on the lead poisoning heatmap.

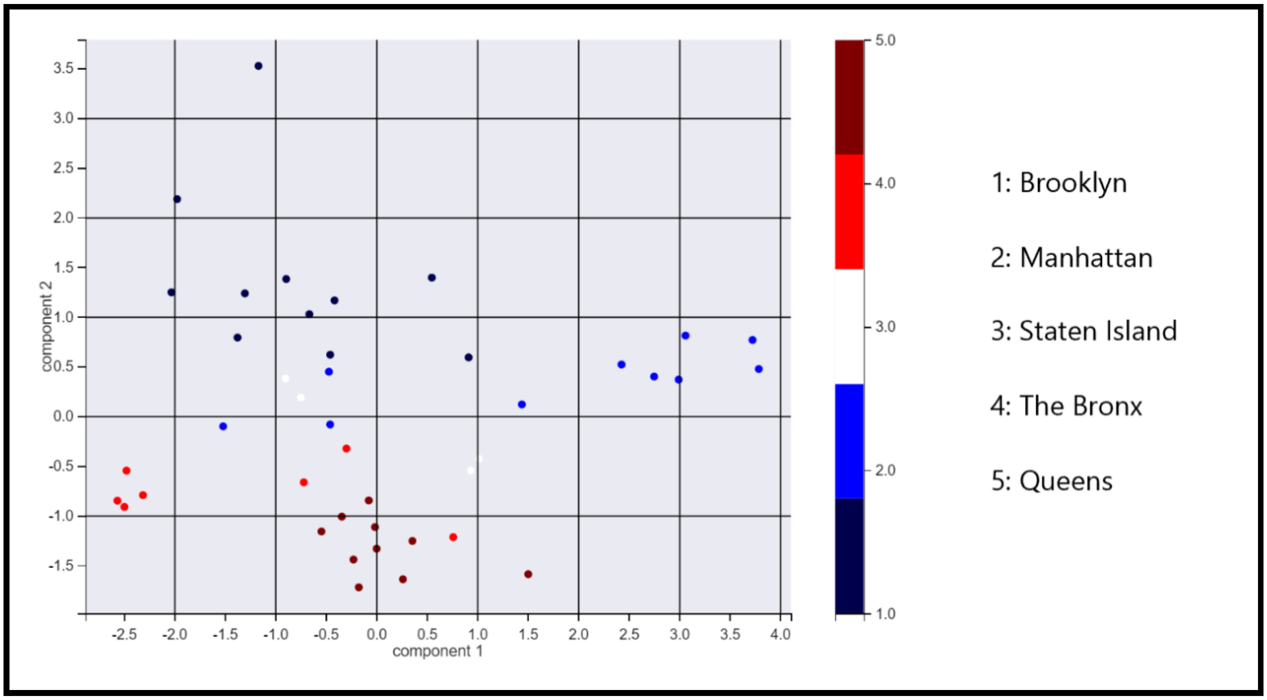


Identified in this heatmap are the areas in which the black population and high-risk lead poisoning neighborhoods overlap. Thus, NYC's black community is disproportionately in the presence of lead poisoning, an issue that must be addressed to NYC's lead poisoning management team. However, this analysis was not complete until we compared NYC's lead poisoning hotspots with a heatmap of NYC's racial/ethnic enclaves.



As depicted in this heatmap, areas not identified in the cross-reference between NYC's lead poisoning and black population proportions are also home to significant Asian and Latino populations, such as Northern Staten Island and South Brooklyn. When cross-referenced with white-majority, neighborhood groupings throughout NYC such as Central and Southern Staten Island and South Manhattan all lack significant lead poisoning proportions. Thus, minorities appear to be disproportionately affected by child lead poisoning.

Principal Component Analysis Code



To sum up our analysis, we conducted a **Principal Component Analysis (PCA).** As referenced in our methods, the PCA algorithm takes a dataset with a principal component (in our case, lead poisoning risk) and compresses every subsequent factor into a graph emphasizing these factors' relations with the principal component. In our dataset, there were four factors (poverty rate, average income, median income, black population) to fit onto the principal component. Using the PCA algorithm, we condensed this 5-column dataset into a 2-column dataset, with column one representing the principal component and column two representing the compressed factors. PCA is useful for determining whether clusters observed in exploratory data analysis are significant to all factors in the graph, essentially a holistic cluster analysis.

After graphing the PCA, we applied a numeric-color code to the data points with each number signifying a borough. The legend on the left indicates which color represents which borough. As depicted by the graph, the clusters differentiating each borough stayed when all factors were applied to the principal component, lead risk. We acknowledge an area in the middle of the graph where some of the data points merge together. However, the majority of data points within Queens, Brooklyn, and Manhattan fit into distinct clusters. Hence, this analysis confirms poverty rate, average income, median income, black population contribute to the disproportion of lead poisoning cases in New York City's boroughs. For NYC to develop a system in which they can profile individuals and neighborhoods for lead-risk, they must consider a broader range of factors beyond housing age to combat lead poisoning.

Additional standout information is within the section of the graph where several data points from different boroughs collide. Three of the data points representing Manhattan in this section are the three Harlems. This key point bolsters our analysis because Harlem is historically low-income, with a sizable black population. Hence, we found its data points on the graph to be disconnected from the more affluent areas of Manhattan. NYC must consider these factors on a neighborhood level basis as well, because the average lead poisoning/income of one borough does not properly represent all of its neighborhoods.

Conclusion

As confirmed by the Principal Component Analysis, disproportion exists between neighborhoods and boroughs in New York City. Lead poisoning in NYC disproportionately affects areas with high minority populations and lower income. Accordingly, areas such as Central Brooklyn, Northern Staten Island, Southeast Queens, and Harlem all display the highest lead poisoning risks. Our analysis can be used in several different ways and scopes. Foremost, NYC's lead poisoning management team can use this analysis to observe the worst affected individual neighborhood and focus their efforts there. Additionally, since median income, average income, poverty rate, and minority population are confirmed to influence child lead poisoning risk, New York City can use this new knowledge to design algorithms that can geographically and personally develop profiles to target high-risk families and/or neighborhoods. In the scope of boroughs, NYC officials also have further evidence confirming that neighborhoods such as Brooklyn suffer a significantly higher proportion of child lead poisoning than Manhattan. In essence, our analysis can be viewed in several dimensions, and in every aspect provide a unique perspective on NYC's child lead poisoning crisis.

Next Steps

New York City lacks an effective framework to record lead poisoning, leading to huge gaps in data. Even worse, the city severely lacks the ability to take affirmative action. In fact, Scott Stringer, the city's comptroller, has identified many agencies that have missed crucial opportunities to protect children from the harm of lead poisoning. As New York City develops definitive tests to collect lead poisoning data on a holistic scale, we will be able to integrate deeper analysis through the usage of a machine learning algorithm. More data means that our algorithm will better utilize factors such as age, geographic location, income, and race to quantify an individual's risk of lead poisoning. Moreover, to help New York City's lead poisoning team better evaluate hotspot areas, we seek to implement minority-specific data to find and distinguish lead poisoning trends within the African, Mexican, and Asian community.

References

Census demographics at the NEIGHBORHOOD TABULATION area (nta) level: NYC open data. (2015, February 17). Retrieved March 21, 2021, from https://data.cityofnewyork.us/City-Government/Census-Demographics-at-the-Neighborhood-Tabulation/rnsn-acs2

Culliton, K. (2019, April 24). Brooklyn leads city in Child lead poisoning cases, data shows. Retrieved March 21, 2021, from https://patch.com/new-york/brooklyn/brooklyn-leads-city-child-lead-poisoning-cases-data-shows

Department of Health and Mental Hygiene. (2018, July 12). Children under 6 yrs with elevated blood lead levels (bll): NYC open data. Retrieved March 21, 2021, from https://data.cityofnewyork.us/Health/Children-Under-6-yrs-with-Elevated-Blood-Lead-Leve/tnry-kwh5

Ferré-sadurní, L. (2019, September 26). 11,168 children tested positive for Lead. the city didn't inspect the homes. Retrieved March 21, 2021, from https://www.nytimes.com/2019/09/26/nyregion/nyc-lead-exposure.html

Liotta, P. (2019, September 26). Kids on North shore among those most often found with elevated blood lead levels in NYC. Retrieved March 21, 2021, from https://www.silive.com/news/2019/09/kids-on-north-shore-among-those-most-often-found-with-elevated-blood-lead-levels-in-nyc.html

Schneyer, J., & Pell, M. (2017, November 14). Lead poisoning lurks in scores of New York neighborhoods. Retrieved March 21, 2021, from https://www.reuters.com/investigates/special-report/usa-lead-newyork/