Lead Levels in NYC Children

By: Sokol Sheri, Josue Avendano, Daryl Tobierre, Fjodi Hyzoti

Summary

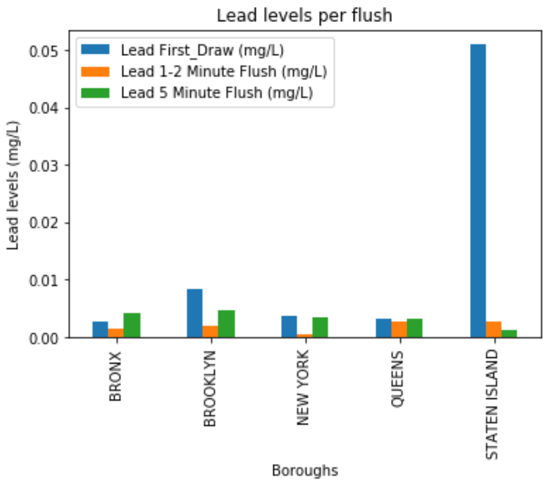
For this project, our group decided to explore lead-levels in New York City. The question that prompted this research was do lead levels in certain boroughs in New York influence elevated lead-levels in children. In particular, we were curious to see if kids obtain these elevated blood-lead levels by contracting them from park-fountains in their area. The three datasets that we used for this research were:

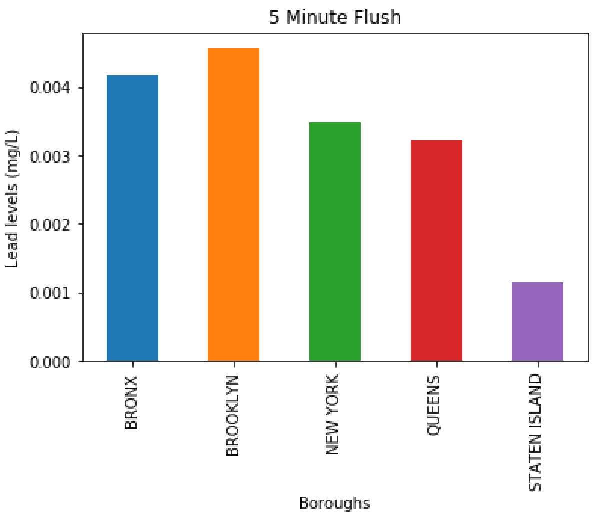
- https://data.cityofnewyork.us/Health/Children-Under-6-yrs-with-Elevated-Blood-Lead-Leve/tnry-kwh5
- https://data.cityofnewyork.us/Environment/Free-Residential-at-the-tap-Lead-and-Copper-Data/k5us-nav4
- https://data.cityofnewyork.us/Environment/Drinking-Fountains/ph76-k6qa

The first link is a dataset for kids under 6 years old with elevated blood-lead levels. This dataset includes geographic locations of each kid. The second link shows lead and copper levels by geographic locations. The third link shows drinking fountains in the city. Overall, our research is looking to find a relationship between all datasets. After completing our research, we have found reason to believe that there is a direct relationship between water fountains, high-blood lead levels in children under 6, and lead-levels in residential households that all point to particular areas having a direct impact on children's blood-lead levels.

The dataset, ‘Free-Residential-at-the-tap’, analyzed lead levels in residential areas from January 1st 2014- December 31st 2017. In order to keep data consistent with other datasets we are using, we filtered the data to only include January 1st 2014- December 31st 2016.This data was collected via test-kits that households used in order to obtain lead-level statistics. The data is broken up in three different tests. The tests were ‘first draw’, ‘1-2 minute flush’, and ‘5 minute flush’. Essentially, testers would draw samples of lead before running their taps, then they would test again after running water through their taps for 1-2 minutes, and finally by testing after running the taps for 5 minutes. So after analyzing all of the data, we decided that all we needed from this dataset were the columns pertaining to boroughs, date collect, first draw, 1-2 minute flush, and 5 minute flush data. This data was essential because we first wanted to realize which areas in New York had high levels of lead and where they stand amongst each other in this regard.

For the most part the data did not need much cleaning. However, we needed to fix the ‘borough’ column. The problem that we found was that data for Queens was not defined as ‘Queens’, instead it was defined by the town in Queens. For example, Breezy Point would be found as a column value instead of Queens. This needed to be changed for grouping purposes. Another issue we found was that some values would be entered as lowercase and uppercase. We needed to change this for consistency purposes. In order to clean the data, we first used a lambda function to change all of the columns in borough to uppercase. After switching all of the values to uppercase, we extracted the values of the columns and turned them from a data series to a list. From there, we checked each value in a loop. The loop checked to see if the values were not in the five boroughs. If the value was not equal to one of those boroughs, we would convert them to Queens. After successfully converting all of the values, we then added the values to a new column called BOROUGHS. After this, we extracted the BOROUGHS column along with all the lead data columns and created a new dataframe. From there we grouped the BOROUGHS with each test data column and found the mean for each. The following results were obtained:





The first bar graph shows each borough's mean lead levels per test. As seen, something strange is made apparent when looking at Staten Island's first draw, mean score. After doing research, we found that Staten Island had an outlier. The outlier was residential first draw score of 9.655, by far the largest. We believe this is why the score is so much higher than the rest.

15223601	STATEN ISLAND	10304	07/23/2015	07/31/2015	9.655

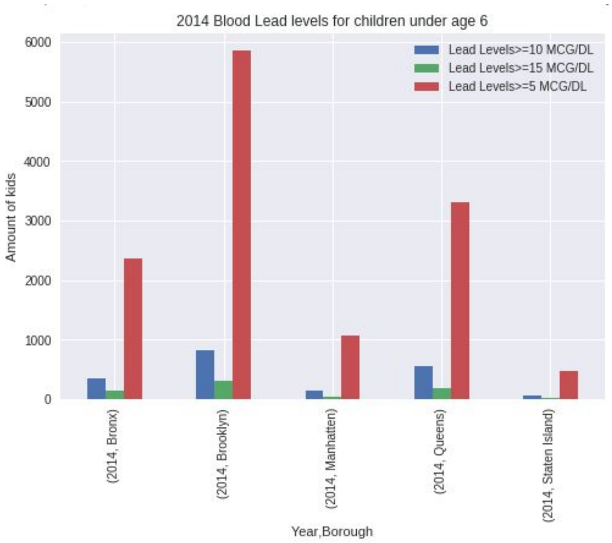
After calculating the mean without the outlier value for Staten Island, we get 0.0059154929577464755.

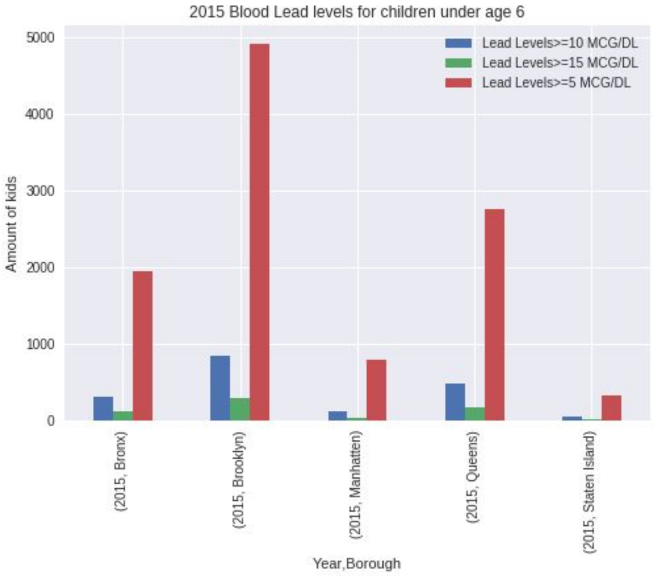
The second bar graph shows the means for the five minute flush. To us, this test is the most important because it shows that after 5 minutes of testing, the lead still exists. Based on these results you can see that Brooklyn leads in 5 minute flush, then Bronx, Manhattan, Queens, and Staten Island in that order.

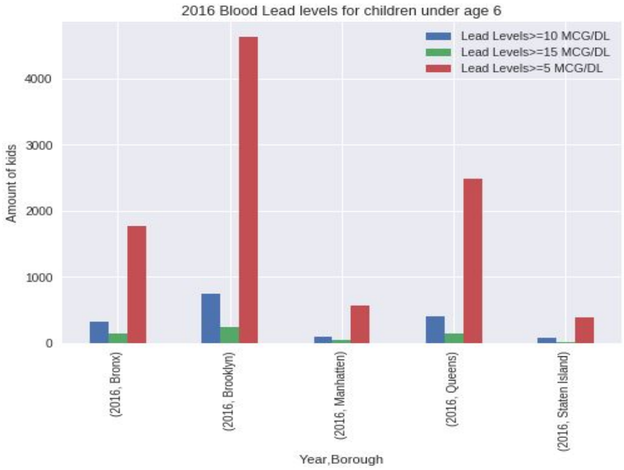
The dataset, "Children_Under_6_yrs_with_Elevated_Blood_Lead_Levels__BLL_", analyzed blood lead levels in children under the age of 6. In order to keep as up to date as we can, we only used years from 2014-2016. In accordance to the New York State Department of Health, the average amount of lead a child under 6 years of age should have is 2 micrograms per deciliter. Going forward the data then divides into different ranges: 5 - child has a little more lead than most children, 10 - the child's lead level is high and requires action, 15- child lead level is quite high and needs immediate action.

With that in mind, the Children _Under_6_yrs_with_Elevated_Blood_Lead_ Levels__BLL_ dataset categorized children that had blood lead levels greater than or equal to 5, 10, or 15. This data was very clean and easy to work with because the boroughs had a borough id number, and the years were written as numbers (i.e. 2014, 2015, 2016). To make our data presentable, we changed the borough ID number to the names of the boroughs themselves. In order to make the columns into names that were easier to work with, we placed the dataset into a dataframe and chose the columns that we wanted to work with. From there, we changed the names in our dataframe.

We then compared the amount of children under the age of 6 year by year to show which borough had the most children who were affected by elevated blood lead levels and our results are as follows:



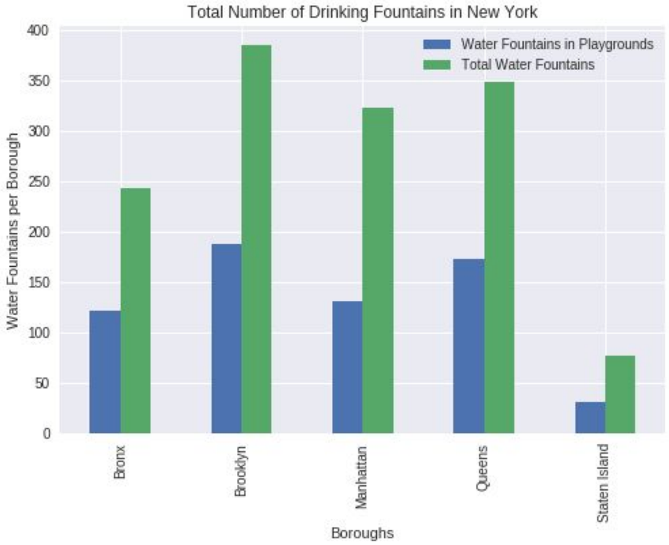




These results all point to Brooklyn having the highest number of children under 6 years old having elevated blood lead levels, with Queens coming in second.

The ‘Drinking Fountains” data set, provided by the Department of Parks and Recreation, contains records of drinking fountain counts that were observed through individual site inspections. In total there are 1,375 sites recorded containing at least 1 water fountain. The data had a unique ‘Site ID’ but it also had a more descriptive ‘Site Name.’ The site name included keywords such as beach, field, court, park, playground, and plaza, to name a few. The borough column of our data set used keywords to identify them such as B, Q, M, X, and R. Our first step in cleaning and preparing the data was to replace the keywords with the exact borough names. Then afterwards, in order to further organize our data we decided to replace the values of site name with a descriptor instead. We first observed the data and saw that various words and phrases were used to describe the same type of place. For example, we repeatedly used loc to find the site name that contains words like ‘playground’, ‘plgd’, and ‘p/g’ and replaced them with ‘Playground’. We repeated the same steps for categories like Plaza, and Beach.

Since our data deals with lead levels in children we thought that analyzing and focusing on water fountains in areas where large amounts of children frequent would be ideal. That’s why after cleaning our data we were able to easily filter by water fountains per borough in playgrounds. We also were able to count how many drinking fountains there are in each borough. Out of our results, we found that the boroughs with the most public water fountains are located in playgrounds in Brooklyn and Queens.



This research does match our hypothesis that the boroughs with the most playground drinking fountains would have the most children with elevated lead blood levels. Brooklyn is #1 in public and playground drinking fountains, residential water lead levels in 5-minute flush, and elevated blood lead levels for each threshold. The second highest in drinking fountains is Queens; it has the second highest number of children with elevated blood lead levels and the second highest residential lead draw from the 1-2 minute flush. The causes behind these findings can be attributed to several typical reasons:

- Old buildings with pipes high in lead content that were acceptable when constructed
- Poor water filtration and corrosion control that exposes lead
- A lack of standards and enforcement capability for private residential buildings

Old buildings and water systems are a reasonable, if not convenient explanation for the findings here. However, if we take into account the apparent correlation between the number of drinking fountains and the number of children with elevated blood lead levels, there is a gap in the coverage that reasoning provides. Some onus of responsibility would fall into the city’s lap. And while nothing conclusive can be stated, it is perhaps worth further study.