Air Quality in Marginalized Communities

By Will Munson

The purpose

Air pollution tends to be a serious issue in large urban environments. What makes this issue worse is poorer areas tend to have higher levels of air pollution than wealthier areas. As a result, those who live in those communities tend to be more vulnerable to various forms of cancer, respiratory problems, and even slower brain development in some children. Most studies indicate PM2.5 at or below 12 μg/m3 is considered healthy with little to no risk from exposure. If the level goes to or above 35 μg/m3 during a 24-hour period, the air is considered unhealthy and can cause issues for people with existing breathing issues such as asthma. For this, I will be taking a closer look at the average Fine Particulate Matter (PM2.5) levels from 2008 to 2020, and see how they correlate with New York City's demographics for each year.

Tools and Packages

- Started with two Excel spreadsheets
- R Studio was used to load and manipulate the data
  - Packages used:
    - Tidyverse - filtering and removing unnecessary fields
    - Lubridate - changing values from char to date
    - Ggplot2 - proper visualizations
    - Broom - used for obtaining numeric values for Cook’s Distance and residuals

Datasets used

Air Quality data obtained from data.gov for New York City

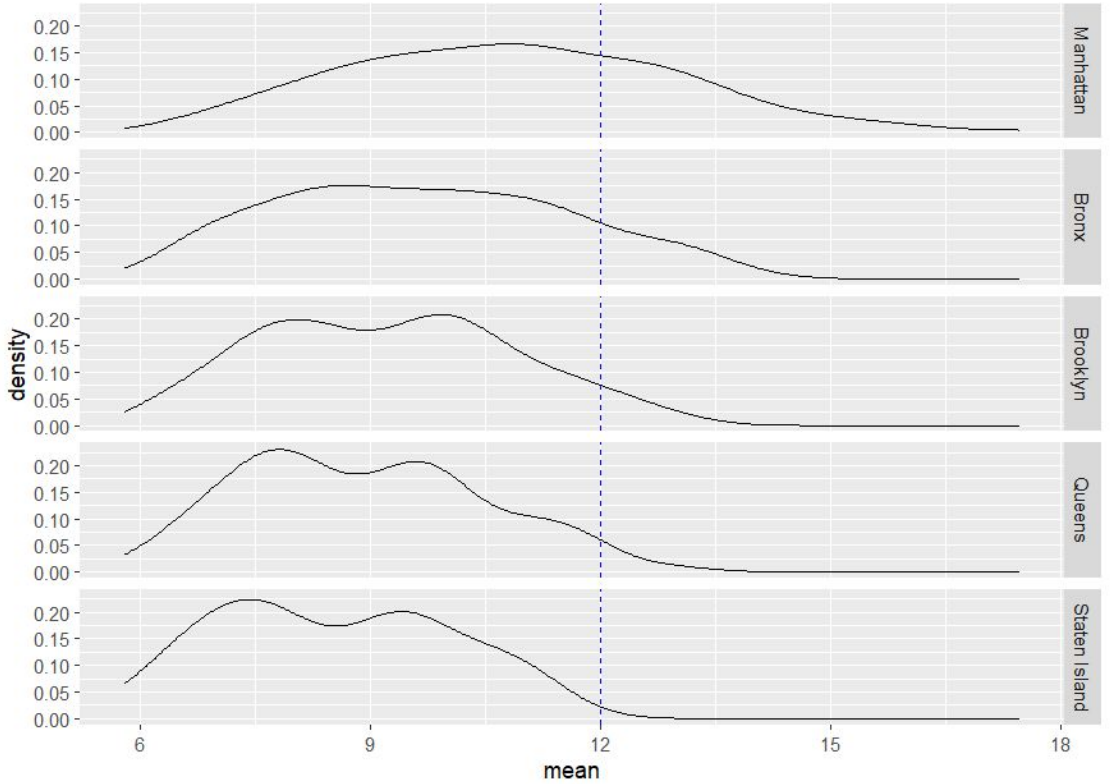
NYU Fuhrman Center Demographic Data for New York City, broken down by boroughs, Public Use Measurement Area, and community district

Roadblocks and Issues

- One of the biggest issues with this dataset is the fact that none of the most basic demographic data was filled in for areas broken down by community district.
- Had to copy and paste data for population density from the PUMA data into their corresponding districts, meaning some districts had to have matching numbers for density (example, Manhattan districts 1 & 2)

Distributions

The averages for every district each year, from 2008 to 2020.

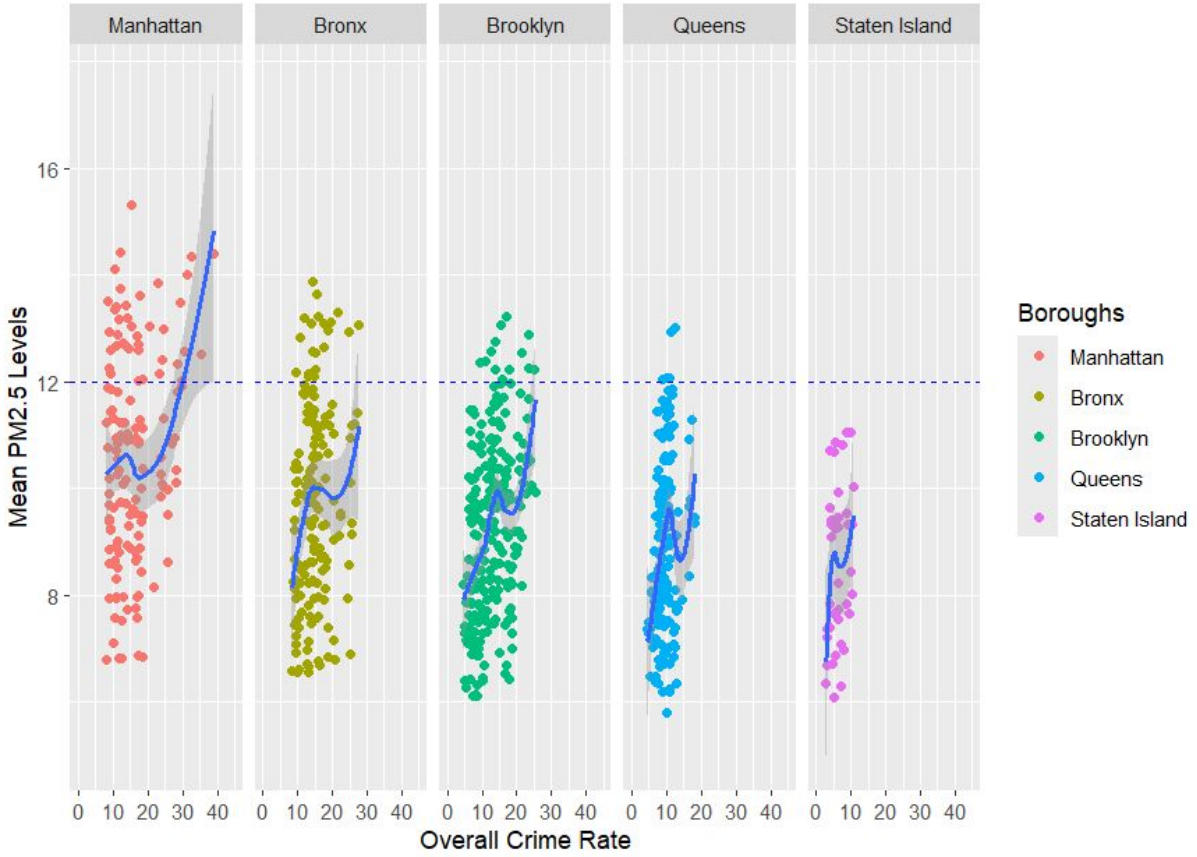


Mean pollution VS. Crime

The Five Boroughs:

1. Manhattan
2. The Bronx
3. Brooklyn
4. Queens
5. Staten Island

As shown on the graph, most areas of Manhattan and the Bronx appear to be above 12 ppm on average. In addition, neighborhoods with higher crime also appear to have higher levels of PM2.5.



The variables

While the initial dataset had a combined total of 124 variables, only one third of them contained data, which was then whittled down to a total of 17 variables. Many of these were excluded due to lack of useful data (proximity to a park or subway), or did not have enough significance for the linear model.

Factors used in our linear model

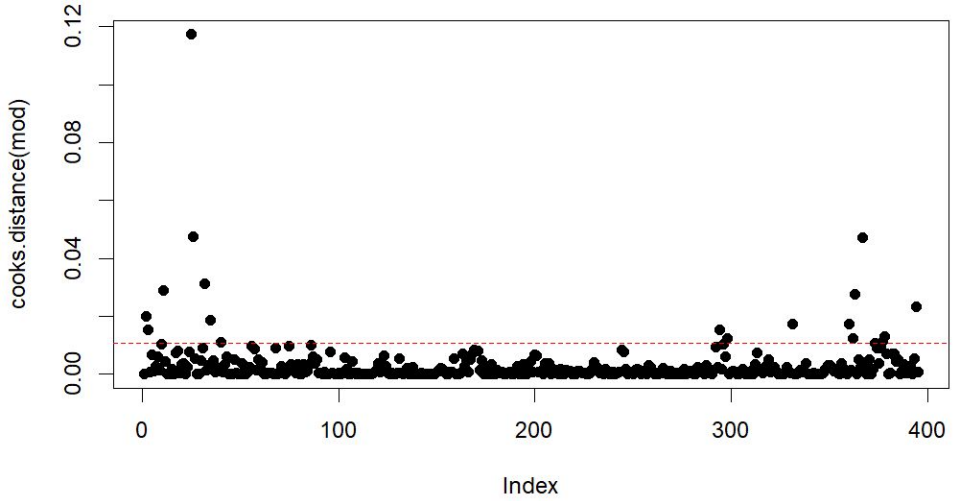
Mean	Response variable - Yearly mean PM2.5 levels
Crime_all_rt	Total crime rate
Hpi_al	Average price changes for all homes
Lp_all	total number of properties that had mortgage foreclosure actions initiated against them.
med_r_4f	Median rent for 2-4 family homes
med_r_ot	Median rent for multifamily buildings
nb_permit_res_units	Total units authorized by new residential building permits
pct_prof_ela	Percent of students (4th grade or below) performing at or above grade level in the arts/english
Pfn_fam14condo	Total pre-foreclosure notices issued to homeowners of 1-4 family homes
pfn_fam14condo_rate	Rate of pre-foreclosure notices issued to homeowners of 1-4 family homes
priv_evic_filing_rt	Eviction filings per rental unit
priv_evic_filings	Total eviction filings
total_viol_rate	Total housing violation rates
volume_1f	Amount of transactions made towards 1 family homes
volume_4f	Amount of transactions made toward 2-4 family homes
Volume_al	Amount of transactions made toward all properties
Population_density	Population density per square mile (copied and pasted from the PUMA data)

Starting Results

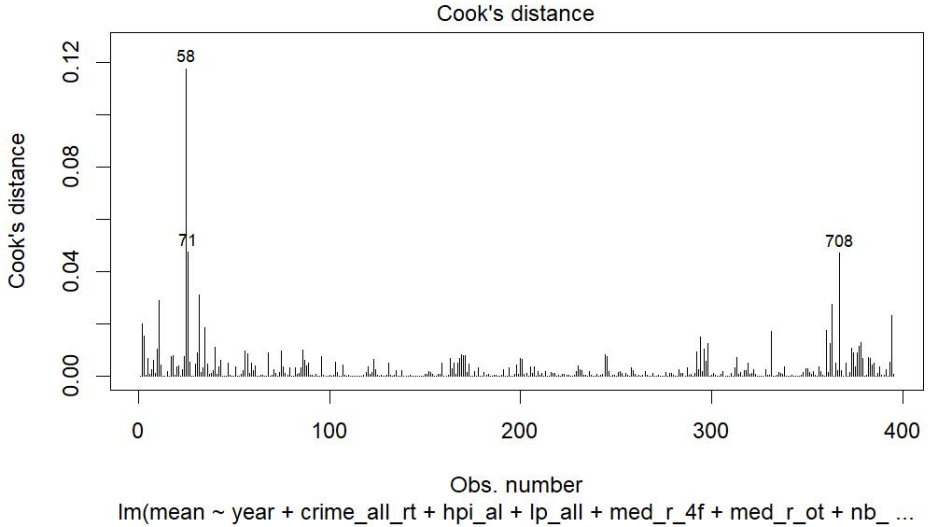
Before beginning the data cleaning process, I had to find a linear model that would be most fitting for the dataset, then used Cook's Distance to outline where there were outliers

```
Call:
lm(formula = mean ~ year + crime_all_rt + hpi_al + lp_all + med_r_4f +
    med_r_ot + nb_permit_res_units + pct_prof_ela + pfn_fam14condo +
    pfn_fam14condo_rate + priv_evic_filing_rt + priv_evic_filings +
    total_viol_rate + volume_1f + volume_4f + volume_al + population_density,
    data = Demo2.5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55079 -0.36823 -0.02885  0.31585  2.26314

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.803e+01  1.006e+00  27.853  < 2e-16 ***
year                -1.197e-03  7.112e-05 -16.835  < 2e-16 ***
crime_all_rt         5.155e-02  5.672e-03   9.088  < 2e-16 ***
hpi_al              -6.496e-04  5.166e-04  -1.258 0.209330
lp_all              -1.969e-03  3.574e-04  -5.509 6.71e-08 ***
med_r_4f            -1.383e-07  5.920e-08  -2.336 0.020017 *
med_r_ot             4.290e-07  3.965e-07   1.082 0.280021
nb_permit_res_units  5.870e-05  4.567e-05   1.285 0.199516
pct_prof_ela         6.160e-03  4.692e-03   1.313 0.190037
pfn_fam14condo       4.032e-04  1.092e-04   3.692 0.000255 ***
pfn_fam14condo_rate  9.074e-03  1.730e-03   5.245 2.62e-07 ***
priv_evic_filing_rt -4.692e-03  1.350e-03  -3.476 0.000568 ***
priv_evic_filings    4.363e-05  3.119e-05   1.399 0.162666
total_viol_rate     -4.489e-04  3.550e-04  -1.264 0.206868
volume_1f           -2.313e-03  3.147e-04  -7.350 1.24e-12 ***
volume_4f           -3.252e-03  4.530e-04  -7.180 3.74e-12 ***
volume_al            1.402e-03  2.069e-04   6.776 4.77e-11 ***
population_density   3.874e-03  2.151e-03   1.801 0.072455 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5515 on 377 degrees of freedom
  (371 observations deleted due to missingness)
Multiple R-squared:  0.8308,    Adjusted R-squared:  0.8232
F-statistic: 108.9 on 17 and 377 DF,  p-value: < 2.2e-16
```

Outlier tests - Cook’s Distance

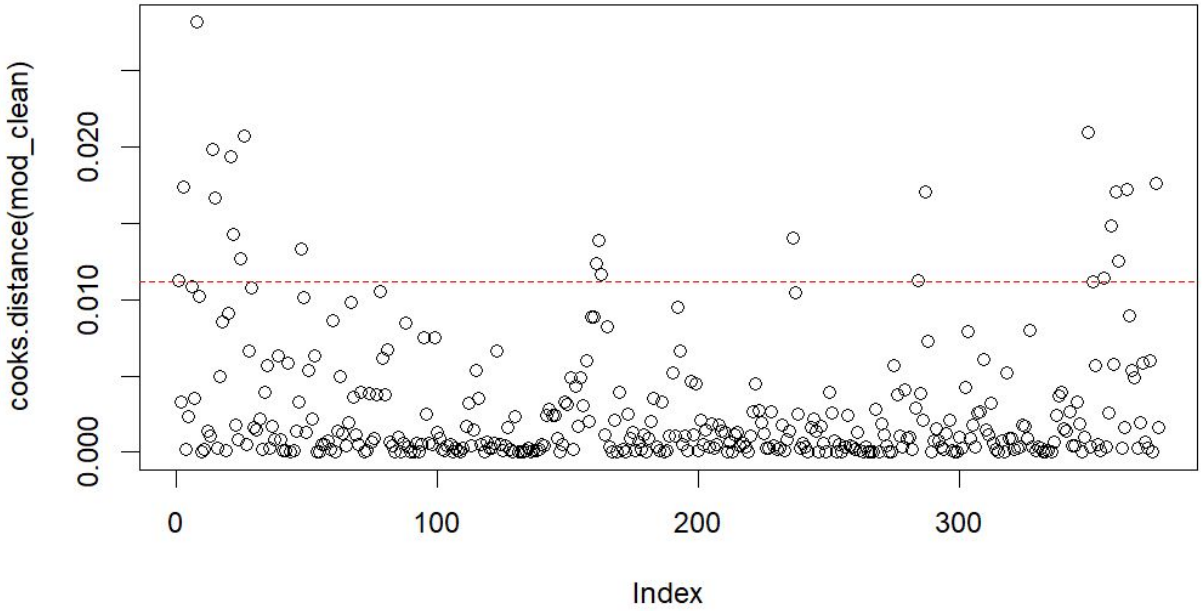


First Cook’s Distance test - quite a few outliers



Point 58 - Midtown in 2013
Point 71 - Stuyvesant Town/Turtle Bay in 2013

Cleaning process



Test 2 - Reduced the number of points from 395 to 376. This will probably be a good stopping point

Cleaned Model

After cleaning the model, we find that the number of transactions towards homes have much higher significance, and foreclosure actions have higher significance. Meanwhile, number of permits, academic performance and even population density appear to be obsolete.

```
Call:
lm(formula = mean ~ ., data = mod_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59782 -0.35218 -0.01163  0.31042  1.23799

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.823e+01  9.452e-01  29.861  < 2e-16 ***
year               -1.184e-03  6.677e-05 -17.731  < 2e-16 ***
crime_all_rt        4.818e-02  8.343e-03   5.775 1.67e-08 ***
hpi_al             -5.794e-04  4.984e-04  -1.163   0.2458
lp_all             -2.750e-03  3.861e-04  -7.122 5.86e-12 ***
med_r_4f           -8.900e-08  6.529e-08  -1.363   0.1737
med_r_ot            4.713e-07  3.944e-07   1.195   0.2328
nb_permit_res_units 1.434e-05  4.609e-05   0.311   0.7560
pct_prof_ela       -2.651e-04  4.347e-03  -0.061   0.9514
pfn_fam14condo      5.524e-04  1.135e-04   4.869 1.69e-06 ***
pfn_fam14condo_rate 9.079e-03  1.634e-03   5.556 5.38e-08 ***
priv_evic_filing_rt -5.386e-03 1.303e-03  -4.134 4.45e-05 ***
priv_evic_filings   6.098e-05  2.904e-05   2.100   0.0364 *
total_viol_rate    -2.703e-04  3.247e-04  -0.832   0.4057
volume_1f          -2.739e-03  3.066e-04  -8.934  < 2e-16 ***
volume_4f          -3.365e-03  4.239e-04  -7.939 2.65e-14 ***
volume_al           1.649e-03  2.028e-04   8.132 7.02e-15 ***
population_density -1.028e-03  2.056e-03  -0.500   0.6174
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4955 on 358 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8368
F-statistic: 114.1 on 17 and 358 DF,  p-value: < 2.2e-16
```

Improvements from data cleaning

Multiple R-Squared increased from 0.8308 to 0.8442

Adjusted R-Squared increased from 0.8232 to 0.8368

Standard Error reduced from 0.5515 to 0.4955

Conclusions

Judging by the results, we can indeed conclude that air pollution appears to be more of a problem in areas where crime is high, housing is more troubling, and people are less inclined to even buy/rent a home.

Works Cited

Air Quality. .xlsx, https://data.cityofnewyork.us/api/views/c3uy-2p5r, data.cityofnewyork.us, 11 Apr. 2022, https://catalog.data.gov/dataset/air-quality.

McNally, Charles. *CoreData.Nyc Data Profile*. .xlsx, 20 Dec. 2022, https://furmancenter.org/coredata/userguide/methodology.

Fine Particles (PM 2.5) Questions and Answers. June 2023, https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm.