Tackling Environmental Justice with One Backpack in New York Neighborhoods

Katie Colasono
Group 159

Abstract

This report examines environmental justice disparities in New York City boroughs by focusing on PM2.5 concentration levels. Dr. Audrey Gaudel and her colleagues in New York City measured street-level air quality by transporting air-quality equipment in backpacks. One of the main goals of this report was to establish if the CDC's environmental justice ranking accurately captured PM2.5 concentration levels and could be a reliable source to inform communities. A second goal was to determine if non-white and white census tracts could be grouped based on CDC's environmental justice index (EJI) and social vulnerability ranking (SVR). The hypothesis was that non-white tracts would have substantially higher EJIs and SVRs than white tracts.

Utilizing k-means clustering and exploratory analysis, this study reveals notable disparities between white and non-white census tracts, consistently exhibiting higher EJI and SVRs. However, comparisons between EJI's predicted PM2.5 levels and actual measurements from the backpack campaign highlight discrepancies, suggesting potential inaccuracies in EJI's rankings.

The findings underscore the importance of refining the clustering algorithms to access the environmental justice rankings accurately. Further research is needed to enhance the comparability of EJI's PM2.5 concentration levels with on-the-ground measurements. Addressing these disparities is crucial for advocating for equitable environmental policies and interventions to safeguard public health and promote environmental justice in urban areas like New York.

Introduction

Last summer, a University of Colorado Boulder's scientist Dr. Audrey Gaudel partnered with researchers in New York City to get detailed readings of air pollutants at street level. The goal was to capture fine particulate matter and ozone readings for multiple neighborhoods in New York City. Ozone is a pollutant at ground level that forms when other pollutants react with the sunlight[1]. Fine particulate matter, also known as PM2.5, consists of particles smaller than 2.5 microns in diameter[1]. PM2.5 has been linked to differing forms of cancer because of its ability to penetrate human lungs. Gaudel and her colleagues measured air quality in about 30 walks over 22 days using backpacks filled with air quality equipment. They focused on areas with known poor air quality and covered areas in Manhattan, the Bronx, Brooklyn, and Queens.

One of the goals of this project is to give New York City residents more information to feel empowered to fight for environmental or air quality justice. Environmental justice is when populations with lesser capacity to influence environmental decisions have an inequitable share of the negative impacts of the environment[2]. The CDC has started to provide air quality and environmental justice rankings for many US census tracts through their environmental justice index

[1] https://cires.colorado.edu/news/video-cires-researchers-tackle-air-quality-streets-new-york
[2] https://www.atsdr.cdc.gov/placeandhealth/eji/docs/EJI-2022-Documentation-508.pdf

(EJI). It has provided these rankings in an interactive map[3], as well. The overall EJI is calculated based on three modules: environmental burden, social vulnerability, and health vulnerability.

The EJI suggests that neighborhoods in New York City can change drastically in their environmental index score (or other modules) by the separation of one street. For example, Figure 1 below shows two different census tracts parallel to each other – one slightly more north than the other. The left image shows a neighborhood with an EJI of 0.40 and a social vulnerability rank of 0.12. The right image, the slightly more northern neighborhood, has a substantially higher EJI rank of 0.89 (2.2x higher) and a social vulnerability rank of 0.96 (8x higher!) than the slightly more south neighborhood. However, all the air pollution metrics are relatively the same. Both neighborhoods have an Ozone of 0.75 and a PM2.5 of 0.62.

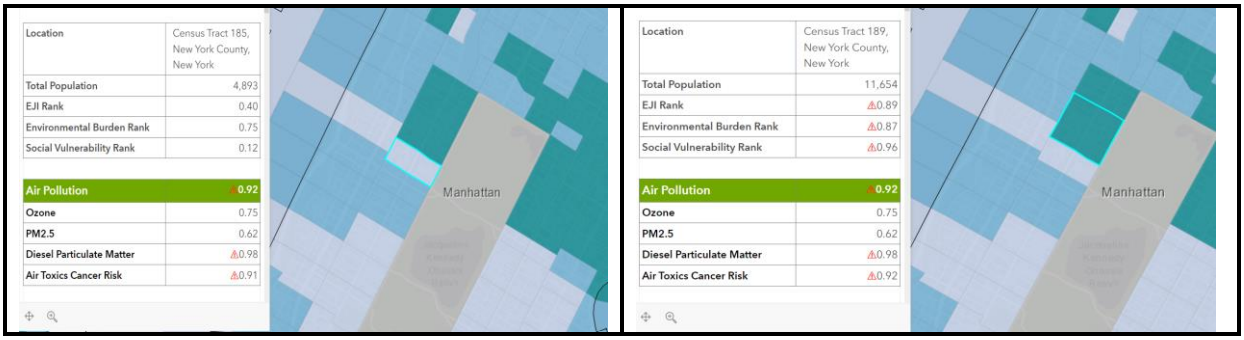


Figure 1: Images were taken from the CDC environmental justice index interactive map.

The following questions are posed: is the EJI capturing the right ozone and PM2.5 levels for each census tract? Are the PM2.5 measurements from the New York backpack campaign and EJI's PM2.5 levels comparable? How do the EJI ranks compare amongst New York boroughs (i.e., Queens, Kings, the Bronx, and Manhattan)? Are census tracts with a higher social vulnerability rank paired with higher EJI ranks? These questions will be answered in this report through exploratory analysis and k-means clustering.

Methodology

Part I: Data

Two data sets are the focus of this analysis: the PM2.5 concentration measurements from the backpack campaign in New York City completed by Dr. Gaudel and her colleagues and the socioeconomic rank data from the EJI. The measurements for PM2.5 concentrations were collected every 15 seconds – and the corresponding latitude and longitude for where the PM2.5 levels occurred. For example, Figure 1 is an image created by Dr. Baron, a researcher on the project, of a walk in Manhattan on August 4th, 2023. The scale on the right shows the concentration levels of PM2.5, where blue and green are low levels of PM2.5 while orange and red are high levels. For this specific walk, the concentration levels for PM2.5 were around 30 $\mu g\ m^{-3}$. For this analysis, each walk is categorized into its respective borough. If a walk crossed into several boroughs it was

[3] https://onemap.cdc.gov/portal/apps/sites/#/eji-explorer

categorized into a borough based on the latitude and longitude data. For future analysis, it would be ideal to be able to categorize each walk into its respective census tract to get a more granular analysis.

The EJI data contains many variables around socio-economic demographics and pollution metrics. The EJI has a metric for the annual mean days above PM2.5 regulatory standards for each census tract. This data was sourced directly from the CDC which provided predicted not actual measurements of the PM2.5 concentrations for each borough. These *predicted* values that the EJI uses in its ranking metrics are from 2014-2016 and thus are not relevant observations of the air quality in New York. To be able to compare EJI's PM2.5 concentrations to the backpack campaign the CDC *predicted* values from July to Aug 2016 were used rather than the annual mean days above regulatory standards.

Figure 1: Map of a walk in Manhattan on August 4th, 2023. Map created by Dr. Baron, a researcher on the project.

The other variables of interest are the percentage of minority persons in each census tract. A minority person defined by EJI includes anyone who is not white/non-Hispanic. For the remainder of this report, this minority group will be referred to anyone as non-White. Each census tract is defined as non-White vs White census tract if the percent of non-White individuals in that tract is higher than average for that borough. The remaining variables of interest are the EJI and social vulnerability ranks.

Part II: K-means Clustering

Two models for the clustering analyses are examined: a basic and a more complex model. The basic model includes three variables: the percent of non-White individuals, the Social Vulnerability Rank, and the Environmental Justice Index for each census tract. The second model includes the following:

i. EJI rank
ii. % of non-White individuals
iii. Estimate of individuals below the 200% poverty line
iv. No high school degree
v. Estimate of the number of renters
vi. Households that make 75,000 or less
vii. Estimate of individuals with no internet
viii. Estimate of the percent of individuals with asthma
ix. A flag indicating that the census tract has more individuals with cancer than average
x. Estimate of the percent of individuals with diabetes

xi. Percentage of individuals reporting not good mental health
xii. An indicator for a census tract with higher-than-average health issues

All these variables were deemed of interest as they have high correlations with the EJI rank. To improve upon this model, it might be useful to conduct PCA to find the most relevant variables to include in the k-means clustering analysis.

It is usually not necessary to split the data into training and testing samples because a k-means algorithm is usually not used for prediction purposes. However, these models may be used as a basis for clustering algorithms for future census tracts or boroughs. Thus, the data is split into a 70/30 split for testing and training samples. A k-means clustering analysis is conducted for each borough to determine if the boroughs have different groupings or clusters. The data is scaled, and the optimal number of clusters is found through the silhouette coefficient. This coefficient is a measure of how similar a data point is within a cluster compared to other clusters. For most models run in this report the optimal cluster is two. To further identify if there is a difference between the two clusters the group means for each variable is calculated and compared for significance.

Exploratory Data Analysis

Of the questions previously posed in this report, two of them can be explored through graphs. The first question is: how do the EJI ranks compare amongst New York boroughs? For this report, we are interested in three New York boroughs – Manhattan (i.e., New York Central), Kings, and the Bronx. These are the areas where Dr. Gaudel and her colleagues collected data in New York. For comparison, the Queens borough will also be added because it is near the other boroughs of interest. Figure 2 displays the EJI rankings for each of the four boroughs split between white (brown) and non-white (green) census tracts. Every borough has higher EJI ranks for non-white census tracts. Manhattan has the largest disparity. Manhattan also has the lowest median for white census tracts compared to the other three boroughs. It is important to note that Manhattan also has the largest population of white individuals among the four boroughs. The Bronx has the highest median for non-white census tracts while the other three boroughs have relatively close medians.

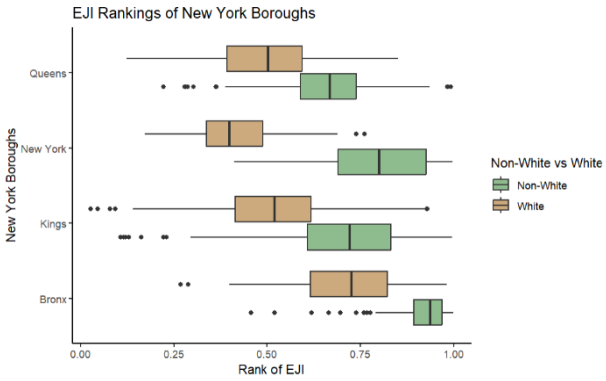


Figure 2: EJI rankings of the four New York boroughs by white and non-white census tracts.

It seems that there are distinct differences in EJI scores between boroughs and white vs non-white census tracts – but is this difference noticeable in the social vulnerability rank, as well? It is clear from Figure 3, that non-white census tracts have a higher social vulnerability rank than white tracts. However, there is a substantial spread for the Queens and Kings boroughs – with some possible outliers. Once again, Manhattan has the largest difference between non-white and white census tracts. For example, the non-white census tract has a median of approximately 0.75 while the white census tract has a median of approximately 0.20. The Bronx also has the highest median for non-white census tracts close to 1.

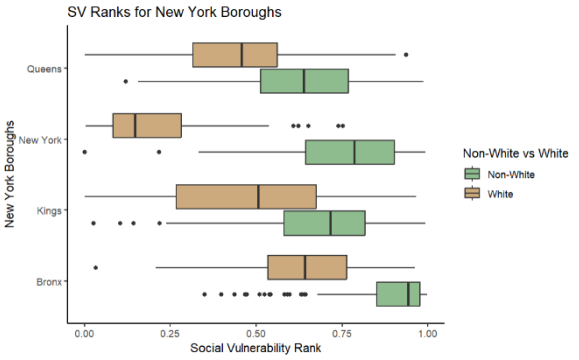


Figure 3: SV rankings of the four New York boroughs by white and non-white census tracts.

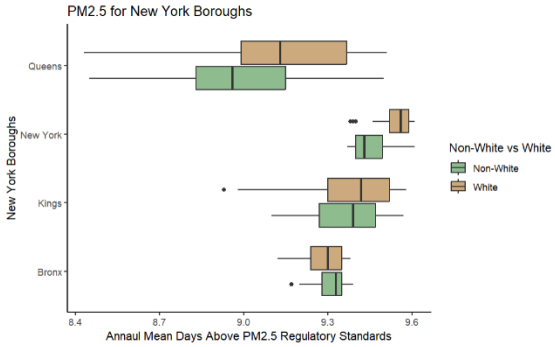


Figure 4: Annual mean days above PM2.5 regulatory standards for 2014 – 2016.

If a census tract a has high EJI then it is likely that the SV ranking will also be high – and Manhattan is a prime example of this relationship. But are PM2.5 concentrations also high for non-white census tracts, as well? As mentioned previously the EJI uses the annual mean days above the regulatory standard as its metric for PM2.5. Figure 4 displays the annual mean days above PM2.5 regulatory standards for each borough. Each borough spends roughly 9-9.5 days above the regulatory standard per year. For reference, the regulatory standard for PM2.5 in 2016 was roughly ~12 $\mu g\ m^{-3}$. From Figure 4, there is no clear difference in the number of days spent above the regulatory stand per year between non-white and white census tracts. Queens does have the largest spread and spends about ~0.5 days less than the other boroughs above the standard. But does the distribution for the predicted concentration levels match the concentration levels from the backpack campaign? The short answer is no – and is shown in Figures 5 and 6.

The data used to get the annual mean days above the regulatory standard is from the CDC's predicted amount for PM2.5 from 2014-2016. Figure 5 is a histogram of the CDC's predicted values for PM2.5 from July to August 2016. Unfortunately, the CDC's historical data for their predicted values did not go past 2016 – so not all the data used by the EJI could be included in Figure 5. To be able to compare the PM2.5 concentration levels between the CDC and the backpack campaign, only July and August are included because the campaign was conducted in July and August.

Looking at the raw data for the predicted PM2.5 concentration levels from July to August in Figure 5, the concentration levels are between 0 and 20. The distribution of the PM2.5 concentration levels does not differ substantially by borough. All the boroughs have 6-10 days at 7.5 $\mu g\ m^{-3}$ – with the Bronx having the most days at this concentration level. However, the concentration levels are far greater in the backpack campaign and range between 0 and 400 (Figure 6). Measurements found of

400 $\mu g\, m^{-3}$ are not frequent but do happen. In Figure 6, Manhattan experiences higher PM2.5 concentration levels than the Bronx and Kings boroughs. There are also higher recorded frequencies of PM2.5 in Manhattan because more walks happened in Manhattan than in the Bronx and Kings boroughs. While this is not a complete picture of the PM2.5 concentration level, it does indicate that the EJI PM2.5 concentration levels are not relevant.

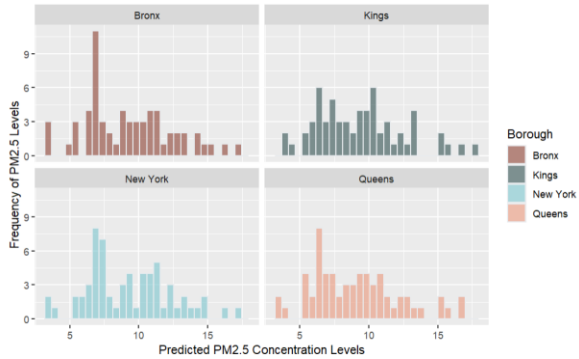


Figure 5: Predicted PM2.5 concentration levels for July and August 2016

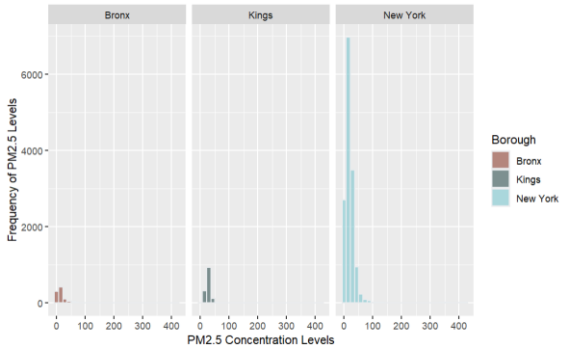


Figure 6: PM2.5 concentration levels from the backpack campaign.

RESULTS

A total of eight cluster analyses were conducted – one basic and one complex for each borough. Each borough required an optimal number of two clusters for the basic model. The same is true for the complex model, except the optimal number for the Queens borough was 5. Table 1 summarizes the performance of each model when using the testing data. Manhattan had the best mode performance out of all the other boroughs. The variance explained within each cluster was 77.2% for the basic model and 60.7% for the complex model. The average silhouette width was decent, as well. The Queens' clusters had the worst performance for the average silhouette width for both the basic and complex models. However, the complex model explained a decent amount of variance (58.8%) within each cluster but still needs improvement because of the small silhouette width (0.26).

The differing performance between the boroughs indicates that more refinement of the k-means clustering algorithm is needed. It is also possible that each borough has factors that require different algorithms. The basic model has the best performance for all the boroughs and could be the recommended model to build off for future research.

	Kings		**Queens**		**Manhattan/New York Central**		**Bronx**	
	Basic	Complex	Basic	Complex	Basic	Complex	Basic	Complex
Optimal Clusters	2	2	2	2	2	2	2	2
Within Cluster SS	51.8%	**37.9%**	**49.0%**	58.8%	**77.2%**	**60.7%**	66.4%	46.4%
Average Silhouette Width	0.44	0.3	**0.43**	**0.26**	0.64	**0.5**	**0.68**	0.45

Table 1: Summarized output from all the k-means clustering models.

The k-means clustering algorithm did group the census tracts into mostly non-white vs white clusters and this is shown by the clustering plot in Figure 7 and the summary statistics in Table 2. From the clustering plot, the blue cluster represents the non-white census tracts, and the orange/pink cluster represents the white census tracts. The white cluster has some variation while the non-white has very little and is clustered close together. While the sample size is relatively small, there is a statistically significant difference between the EJI and SV ranks for non-white and white census tracts in these two clusters. For example, the non-white census tract has a 1.4x higher EJI than the white census tract. The PM2.5, a measurement from EJI, not the backpack campaign, is relatively similar – but the white census tracts have a higher PM2.5. Again, it would be interesting in future research to be able to pair the EJI and other socio-demographics from the EJI database with the backpacking campaign to see if the clusters group differently or if the summary statistics change. See the appendix for the remaining cluster plots and tables.

	Cluster 1 (white)	*Cluster 2 (non-white)*
N	19	80
PM2.5	9.38	9.32
EJI	0.65	0.92
SVR	0.58	0.91

Table 2: Summary statistics from cluster groups for the Bronx.

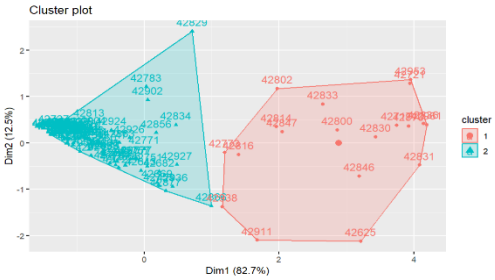


Figure 7: Bronx basic k-means clustering model.

Conclusions

In general, the k-means clustering algorithms need more refinement. The models explain variation within clusters relatively well but need improvement. Each borough may need a modified version of the clustering algorithm. Another improvement upon the model is to conduct PCA to find the most relevant variables to include in the k-means clustering analysis.

A key takeaway is that the non-white and white census tracts are grouped into separate clusters – and within those clusters, the non-white tracts had statistically significantly higher EJI and SV ranks than the white tracts. The second takeaway is that the backpack campaign clearly shows that PM2.5 concentration levels are much higher than represented by the EJI predicted levels. A more granular comparison of the PM2.5 levels from EJI and the backpack campaign by matching longitude and latitude coordinates will provide a more accurate comparison of how well the EJI is capturing PM2.5 levels.

Lessons Learned

I really enjoyed this class! I definitely have learned a lot of different techniques for analyzing data and being able to answer important questions. I feel more confident in my abilities to implement different machine learning methods. However, I do feel like the lectures could have more applied examples. It would also be helpful if the presentation and the paper were due on the same day rather than the presentation being due first. It allows for more time to complete the project.

Credits

Dr. Gaudel – Thank you for the data and the project idea!

Bibliography

CIRES researchers tackle air quality from the streets of New York. (2023, October 13). Retrieved from CIRES: https://cires.colorado.edu/news/video-cires-researchers-tackle-air-quality-streets-new-york

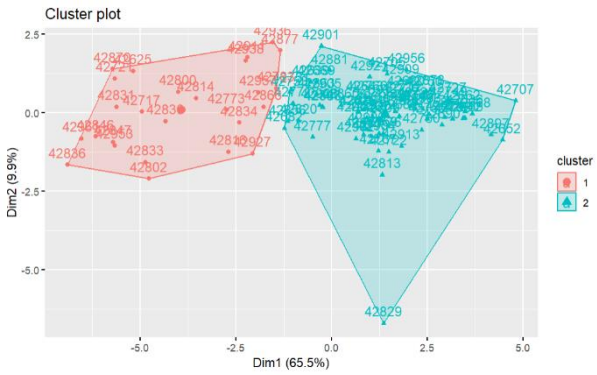
Enivronmental Justice Index Explorere. (2024). Retrieved from CDC: https://onemap.cdc.gov/portal/apps/sites/#/eji-explorer

Technical Documentation for the Environmental Justice Index 2022. (2022). Retrieved from CDC: https://www.atsdr.cdc.gov/placeandhealth/eji/docs/EJI-2022-Documentation-508.pdf

Appendix

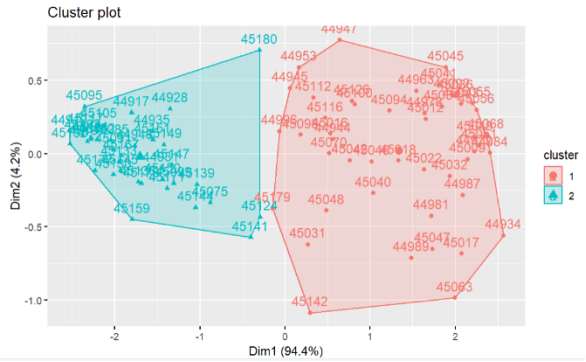
Bronx – Complex Model

	Cluster 1 (white)	*Cluster 2 (non-white)*
N	26	73
EJI	0.70	0.93
Below 200% Poverty	24.8%	58.5%
No Highschool Degree	16.1%	31.1%
% Renter	52.1%	88.7%
% making $75K or less	35.3%	56.3%
% No Internet	53.6%	74.4%
% with Asthma	10.6%	12.6%
Flag for Cancer	0.34	0.01
% with Diabetes	11.4%	16.0%
Flag for Health Issues	1.31	3.21
% with Bad Mental Health	11.9%	16.8%



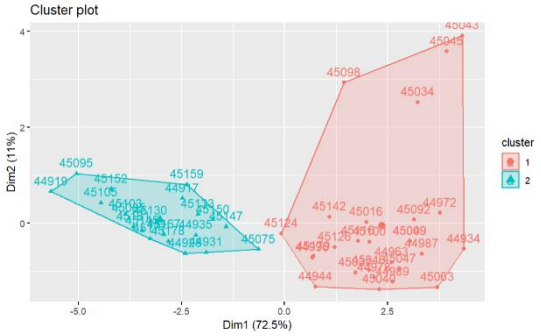
New York – Basic Model

	Cluster 1 (white)	*Cluster 2 (non-white)*
N	47	37
PM2.5	9.54	9.44
EJI	0.45	0.86
SVR	0.26	0.82



New York – Complex Model

	Cluster 1 (white)	*Cluster 2 (non-white)*
N	27	21
EJI	0.46	0.88
Below 200% Poverty	17.7%	51.9%
No Highschool Degree	4.6%	26.7%
% Renter	67.0%	91.9%
% making $75K or less	25.2%	45.7%
% No Internet	25.8%	73.2%
% with Asthma	9.09%	11.3%
Flag for Cancer	0.15	0
% with Diabetes	6.0%	14.0%
Flag for Health Issues	0.26	2.29
% with Bad Mental Health	10.2%	15.1%



Kings – Basic Model

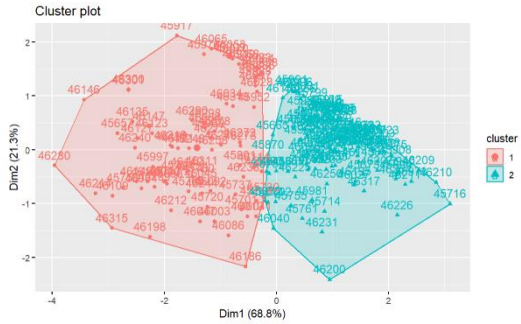
	Cluster 1 (white)	*Cluster 2 (non-white)*
N	79	145
PM2.5	9.40	9.37
EJI	0.44	0.73
SVR	0.36	0.73

Kings – Complex Model

	Cluster 1 (white)	*Cluster 2 (non-white)*
N	114	110
EJI	0.48	0.78
Below 200% Poverty	25.0%	49.8%
No Highschool Degree	12.0%	23.6%
% Renter	54.1%	76.1%
% making $75K or less	32.4%	49.4%
% No Internet	37.6%	69.8%
% with Asthma	9.30%	11.3%
Flag for Cancer	0.14	0.04
% with Diabetes	8.61	13.1
Flag for Health Issues	0.38	1.95
% with Bad Mental Health	11.2	15.1

Queens – Basic Model

	Cluster 1 (white)	*Cluster 2 (non-white)*
N	81	111
PM2.5	9.06	9.01
EJI	0.46	0.69
SVR	0.37	0.68



Queens – Complex Model

	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
N	11	46	64	17	54
EJI	0.52	0.64	0.45	0.76	0.66
% non-white	30.5%	97.8%	54.6%	93.8%	82.8%
Below 200% Poverty	16.2%	22.8%	22.6%	55.7%	38.8%
No Highschool Degree	10.7%	14.9%	11.3%	36.5%	22.4%
% Renter	21.5%	29.2%	49.8%	81.6%	60.9%
% making $75K or less	24.1%	31.6%	32.8%	49.3%	44.7%
% No Internet	46.4%	39.2%	34.0%	74.0%	48.6%
% with Asthma	8.7%	11.1%	8.2%	10.2%	8.7%
Flag for Cancer	1	0	0	0	0
% with Diabetes	11.1%	13.7%	9.8%	15.0%	12.5%
Flag for Health Issues	1.6	2.3	0.05	1.7	0.6
% with Bad Mental Health	10.0%	12.7%	10.5%	15.3%	12.4%