Nicole Mis
CS 504
5/3/19

Examining the Relationship Between the Environment and Our Health

As the world's population continues to grow and the world becomes increasingly urbanized, cities will continue to swell with people and expand into surrounding areas. About 55% of the world's population already lives in urban areas but this proportion is expected to increase to 68% by 2050 (United Nations). Cities, often those in developing countries, have gotten a bad reputation for being unhealthy, filled with grime, crime and disease. For this project, I wanted to explore the health of Boston and how the city landscape affects our health. More specifically, I wanted to see if I could pinpoint certain areas of the city that have more waste sites and less green spaces, and analyze how this has affected its residents. Past literature has focused on how toxic substances found in our air, soil, and water has affected our health. With this in mind, I decided to try to explore this relationship using data for Boston.

As a first step in my project, I started researching what data was available to me for Boston. However, I found that getting health data is extremely difficult, most likely because it is hard to collect this type of data. I eventually found data from 500 Cities: Local Data for Better Health, which is a project funded by the Robert Wood Johnson Foundation in conjunction with the CDC foundation. It is a complete dataset that includes 2016 and 2015 model-based small estimates for 27 measures of chronic disease related to unhealthy behaviors, health outcomes, and use of preventive services. Data from the Behavioral Risk Factor Surveillance System (BRFSS), Census Bureau, and American Community Survey were used to generate these measures. It is important to keep in mind that I am using this dataset as my primary source for health data. Next, I wanted to find a dataset that would capture how cities often have more pollution and waste than more rural areas. I searched for data about air quality, soil quality, water quality but often this data is not provided for every area within a city but rather is provided for a larger region like a whole city or whole region of the country. Instead, I collected information on waste sites within the city of Boston.

To collect information on waste sites, I found several data sources provided by the Massachusetts government about sites that produce waste or oil. One of these datasets lists sites that are designated as "hazardous waste generators" and is provided by the Massachusetts Department of Environmental Protection. These "hazardous waste generators" appear to be businesses like

pharmaceutical companies or auto body shops. Another dataset that I retrieved, called Oil and/or Hazardous Material Sites with Activity and Use Limitations (AUL), had a list of the approximate location of oil or hazardous material release or disposal sites. The locations of these sites are approximations and had not been verified by the Department of Environmental Protection. The third source I used to gather data on waste was called Tier Classified Oil and/or Hazardous Material Sites which contained information about oil and/or hazardous material disposal sites that had been reported and Tier Classified under M.G.L. Chapter 21E and the Massachusetts Contingency Plan (MCP). Tier I or II sites are those where permanent cleanup of the site had not been accomplished within a year of it being reported to MassDEP. There are also a variety of other factors, including the site's complexity, the type of contamination, and the potential for human or environmental exposure to the contamination, that are considered when designating waste sites. Some sites are automatically given a Tier I classification if they pose an imminent hazard or affect public water supplies. Those three sources were combined to create a complete hazardous waste data set.

To create a more robust analysis, I also collected data about crime, green spaces, income, population and schools in the city of Boston. I retrieved crime data from Boston's data portal. Additionally, I found data on pre-k through 12th grade schools from Boston Maps, which is a data portal that contains geospatial datasets. Additionally, income and population data were retrieved from the Census Bureau for every census tract within Boston. These were the main datasets that I used for my analysis.

After retrieving these datasets, I conducted a series of data manipulations to transform this data into a more useable form. First, I had to create a new dataset from the health data I retrieved from the CDC because every row was a different health outcome, unhealthy behavior, or preventative behavior. However, for my purposes, I needed to aggregate all of these health outcomes for a certain census tract. Therefore, I first filtered out unhealthy behavior and preventative behavior because my analysis was not focused on those measures. Thus, I was only focusing on illnesses that weren't necessarily linked to a certain behavior. Then, I created a new dataset where every row was a different census tract which had twelve columns with each column documenting a different health outcome such as cancer. Next, I also had to make some adjustments to the waste site dataset. For instance, I had to find the coordinates for some of the waste sites by using their addresses because this information wasn't provided in the original source. I also had to get the census tract of where these waste sites were located.

In project 1, I tried to match up Boston zip codes with the number of hazardous sites, median income, crime rate, and health issues to see whether there was a correlation between these factors. I wanted to determine whether zip codes with more health problems had more waste sites. I also retrieved income, crime, and population data for every zip code because I wanted to get a sense of the demographics of the area because these factors could also have an effect on the health of that area. Therefore, I created a data set that counted all of the waste sites for every zip code in Boston. I also added the average income for every zip code to this data set so the final dataset had the total number of waste sites and average income for every zip code within Boston. However, in project 2, I decided to look at data at the census tract level which are smaller geographic units than zip codes because the health data from the CDC was already recorded at this geographic level.

In project 2, I added an open space dataset which contained information about all the green spaces within Boston and a data set about the public and non-public pre-k through 12^{th} grade schools in Boston. The open space data set gave coordinates of the boundaries of the green spaces so I found the centroids of all that green spaces so that it would be easier to compare distances between green spaces and waste sites. The last transformation I created was the most drastic because it combined the number of crimes committed in each of the census tracts with the amount of health problems, number of open spaces and schools, population and average income for that census tract. This transformation ensured that I had a consistent and comprehensive data set of all the key factors for every census tract, thus allowing me to solve a constraint satisfaction problem and conduct a statistical analysis.

Now that my data sets were in a useable form, I began to experiment with constraint satisfaction and optimization techniques to analyze this data. I decided that I wanted to conduct some type of linear regression on my data to see whether there was a relationship between these factors. Thus, I first decided to use gradient descent as an optimization technique to find parameters that minimized the sum-of-squares cost function for linear regression. Gradient descent is computationally faster than regular linear regression so that is why I first decided to try it. In order for my model to converge, I had to set my learning rate to a very small number. However, after conducting gradient descent, I found that my parameters were very small and also that gradient descent doesn't calculate any of the standard deviations, p-values or t-statistics for the coefficients which makes it harder to measure the significance of the coefficients. Therefore, in order to get more meaningful estimates, I decided to use the statsmodels API package in python which allows you to run linear regression. I was then able to run the following model:

$$y_{disease} = \beta_0 + \beta_1 x_{crime} + \beta_2 x_{waste} + \beta_3 x_{openspace} + \beta_4 x_{income} + \beta_5 x_{population} + e$$

The outcome variable is the number of diseases that occurred in every census tract. The data I retrieved from the CDC had prevalence rates for each disease so I took this proportion and multiplied it by the number of people living in that census tract to get the total number of disease occurrences. I transformed the data in this way because I thought that having proportions as my outcome variable could make my regression biased. So instead, I ran the regression with the number of diseases as my outcome variable. Additionally, using linear regression with probabilities is problematic because the model may find fitted values that are impossible, and this could make my predicted values extend outside of the 0 and 1 bound. This makes interpreting the coefficients from the model harder. Upon further research, I found that using a logistic regression may be a better fit because my outcome variable is bounded between 0 and 1. So I ran the same regression as the linear one but using logistic regression. Finally, I also found that there is regression called a beta regression which is a type of regression that can be used when your outcome variable is a proportion. The beta regression is a better model for this data because data from the unit interval is typically heteroskedastic, displaying more variation around the mean than toward the extremes, and this model accounts for this. Additionally, proportions are typically asymmetric and so using the normality assumption is not always accurate. Therefore, this led me to run the beta regression model using cancer prevalence rate as my outcome variable and average disease prevalence rate. I decided to use cancer and average probability of disease as my outcome variables because I didn't transform cancer in any way but I did take the average of all the prevalence rates to create one overall prevalence rate for every census tract. Therefore, to ensure that I wasn't getting results that were biased due to my transformations, I also used cancer as my outcome variable. By running all of these regressions, I hoped to create an accurate model that captured any possible relationships between my variables.

In the event that waste sites did have a large impact on health, I wanted to find the locations of pre-existing waste sites that had the lowest impact on the Boston area. In order to do this, I ran K-means to find the top waste clusters in Boston using the haversine metric because it determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Next, I found the average distance from these waste sites to all schools and green spaces. I also found the population of the zip code where all of these waste sites were located. I then ranked each of these waste sites on

each of these features and summed up the ranks to find an overall score for each waste site. I then was able to rank these waste sites from lowest to highest impact.

The third and final analysis I conducted on these datasets included calculating correlation coefficients and a custom scoring metric. I computed correlation coefficients between cancer occurrences and waste, income, crime, and green spaces. Additionally, I computed correlation coefficients between total disease occurrences and waste, income, crime and green spaces. Finally, I calculated the correlation coefficient between open space and income. In addition to correlation coefficients, I created a custom health scoring measure so that I could get a sense of whether there were areas within Boston that were healthier than others. In order to get this metric, I first normalized all of the factors between 0 and 1 and multiplied them by their corresponding correlation coefficient which was first normalized between 0 and 1. However, by deriving the weight of each factor using the correlation coefficient, I am not factoring in the significance of these factors with the health outcome so using the correlation coefficients may not be an accurate way to weight all of these factors. Therefore, you should keep these shortcomings in mind when interpreting the results of this metric.

The results of my analysis showed that it is difficult to draw a causal relationship between waste sites and health outcomes. My linear regressions (figures 1-2) show that the estimates are either insignificant or the coefficients are very close to zero and the corresponding standard errors are small too. This could indicate that there is no relationship between these variables or that the model is not correct. Consequently, I ran a logistic regression (figure 3) that showed that all of the coefficients were statistically insignificant, again demonstrating that there is no relationship. Finally, I ran two beta regressions (figures 4-5) - one with probability of disease as the outcome variable and one with probability of cancer as the outcome. The beta regressions show that population and income have an effect on health because the coefficients are somewhat greater than 0 and are statistically significant at the 5% confidence level. These results are surprising because the regression shows that there is a positive coefficient for income, meaning that an increase in income also increases the prevalence of disease. However, there could be omitted variables that are influencing this result. For population, the coefficient was also positive which makes more sense because having a larger population within an area could detrimentally affect someone else's health. For instance, second hand smoke could contribute to someone else getting lung cancer who wouldn't have otherwise gotten it if they had been in a less crowded area.

The scatterplots (figure 6) reinforce the regression results by showing that there is only a slight correlation between these factors. Based on the correlation coefficients, disease prevalence and crime have the highest correlation. This could be due to unobserved variables that affects a person's health such as unhealthy behaviors like drinking. Open spaces and disease prevalence also have a positive correlation which is surprising because it is not immediately apparent why more open spaces can cause a higher prevalence of diseases. Again, there could be confounding variables affecting this correlation. I also generated a map that displays the custom health metric (figure 7) that I calculated for every census tract. This map shows that there are pockets of unhealthy areas that are mostly dispersed along the southwest border of Boston. The other map I included shows the average disease prevalence rate (figure 8) for every census tract and I also plotted the clusters of waste sites that I found with their corresponding rank. If you go to my website, you can click on the map markers and look at the rank of each waste cluster with 1 being the worst rank or the least optimal waste site. This map shows that the locations of these waste sites do somewhat correspond with the custom health metric I created in the first map, possibly providing evidence that not all locations within Boston are equally healthy. Overall, since there are a lot of factors that can impact health and which are not considered in these models, more analysis needs to be done in order to determine with certainty the relationship between these variables.

In terms of future work, I think it would be interesting to also look at preventive and unhealthy behaviors within an area to see whether populations within certain demographics are more likely to behave in a certain way. These behaviors can also have a huge impact on someone's health, and so they are important to consider. Similarly, it could be interesting to look at air and water quality of certain regions within the United States and whether this correlates with the health of the population within that area. However, this would have to be conducted on a macro level so that there is more variation within air and water quality. Someone could potentially look at air and water quality of all the major cities and compare and contrast the healthiness of the people in those cities. You would also want to factor in the demographic features of these populations to create a more comprehensive analysis. Overall, there are a million different ways someone could go about analyzing the complicated and complex relationship between our environment and our health. With this analysis, I hoped to provide just one approach to tackling this problem.

Appendix

Figure 1 – Linear Regression

$$y_{health\ occurrences} = \beta_0 + \beta_1 x_{crime} + \beta_2 x_{waste} + \beta_3 x_{openSpace} + \beta_4 x_{income} + \beta_5 x_{population} + e$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.359
Model:	OLS	Adj. R-squared:	0.343
Method:	Least Squares	F-statistic:	22.97
Date:	Thu, 02 May 2019	Prob (F-statistic):	2.79e-18
Time:	19:47:59	Log-Likelihood:	651.35
No. Observations:	211	AIC:	-1291.
Df Residuals:	205	BIC:	-1271.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0084	0.002	4.558	0.000	0.005	0.012
x1	7.466e-05	0.000	0.382	0.703	-0.000	0.000
x2	-0.0002	0.000	-1.398	0.164	-0.000	7.29e-05
x3	6.24e-08	3.85e-08	1.620	0.107	-1.35e-08	1.38e-07
x4	0.0009	0.000	2.257	0.025	0.000	0.002
x5	3.747e-06	4.66e-07	8.049	0.000	2.83e-06	4.67e-06

Omnibus:	13.450	Durbin-Watson:	1.729
Prob(Omnibus):	0.001	Jarque-Bera (JB):	14.467
Skew:	0.638	Prob(JB):	0.000722
Kurtosis:	3.124	Cond. No.	9.36e+04

Figure 2 – Linear Regression

$$y_{cancer\ occurrences} = \beta_0 + \beta_1 x_{crime} + \beta_2 x_{waste} + \beta_3 x_{open\ space} + \beta_4 x_{income} + \beta_5 x_{population} + e$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.066
Model:	OLS	Adj. R-squared:	0.043
Method:	Least Squares	F-statistic:	2.903
Date:	Thu, 02 May 2019	Prob (F-statistic):	0.0148
Time:	19:49:13	Log-Likelihood:	-1025.5
No. Observations:	211	AIC:	2063.
Df Residuals:	205	BIC:	2083.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0480	5.228	-0.009	0.993	-10.356	10.260
x1	1.4584	0.553	2.639	0.009	0.369	2.548
x2	0.6385	0.359	1.778	0.077	-0.069	1.346
x3	-8.358e-05	0.000	-0.767	0.444	-0.000	0.000
x4	-1.5839	1.109	-1.428	0.155	-3.771	0.603
x5	0.0012	0.001	0.890	0.374	-0.001	0.004

Omnibus:	275.176	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13726.024
Skew:	5.767	Prob(JB):	0.00
Kurtosis:	40.792	Cond. No.	9.36e+04

Figure 3 – Logistic Regression

$$y_{health\ occurrences} = \beta_0 + \beta_1 x_{crime} + \beta_2 x_{waste} + \beta_3 x_{openSpace} + \beta_4 x_{income} + \beta_5 x_{population} + e$$

Logit Regression Results

Dep. Variable:	y	No. Observations:	211
Model:	Logit	Df Residuals:	205
Method:	MLE	Df Model:	5
Date:	Thu, 02 May 2019	Pseudo R-squ.:	inf
Time:	20:09:40	Log-Likelihood:	-5.9385
converged:	True	LL-Null:	0.0000
		LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.2727	1.635	-3.224	0.001	-8.478	-2.067
x1	0.0020	0.106	0.019	0.985	-0.206	0.210
x2	0.0065	0.074	0.087	0.931	-0.140	0.152
x3	5.055e-06	2.45e-05	0.206	0.837	-4.3e-05	5.31e-05
x4	0.0019	0.209	0.009	0.993	-0.408	0.412
x5	48.4328	38.743	1.250	0.211	-27.503	124.369

Figure 4 – Beta Regression

$$y_{disease\ prevalence} = \beta_0 + \beta_1 x_{crime} + \beta_2 x_{waste} + \beta_3 x_{open\ space} + \beta_4 x_{income} + \beta_5 x_{population} + e$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.180
Model:	OLS	Adj. R-squared:	0.160
Method:	Least Squares	F-statistic:	9.027
Date:	Thu, 02 May 2019	Prob (F-statistic):	9.01e-08
Time:	20:16:47	Log-Likelihood:	-278.40
No. Observations:	211	AIC:	568.8
Df Residuals:	205	BIC:	588.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.955e-16	0.063	-6.26e-15	1.000	-0.125	0.125
x1	0.1175	0.070	1.688	0.093	-0.020	0.255
x2	-0.0733	0.066	-1.110	0.269	-0.204	0.057
x3	0.1114	0.065	1.724	0.086	-0.016	0.239
x4	0.3267	0.068	4.809	0.000	0.193	0.461
x5	0.1599	0.065	2.449	0.015	0.031	0.289

Omnibus:	2.080	Durbin-Watson:	1.569
Prob(Omnibus):	0.353	Jarque-Bera (JB):	1.940
Skew:	0.147	Prob(JB):	0.379
Kurtosis:	2.634	Cond. No.	1.66

Figure 5 – Beta Regression

$$y_{cancer\ prevalence} = \beta_0 + \beta_1 x_{crime} + \beta_2 x_{waste} + \beta_3 x_{open\ space} + \beta_4 x_{income} + \beta_5 x_{population} + e$$

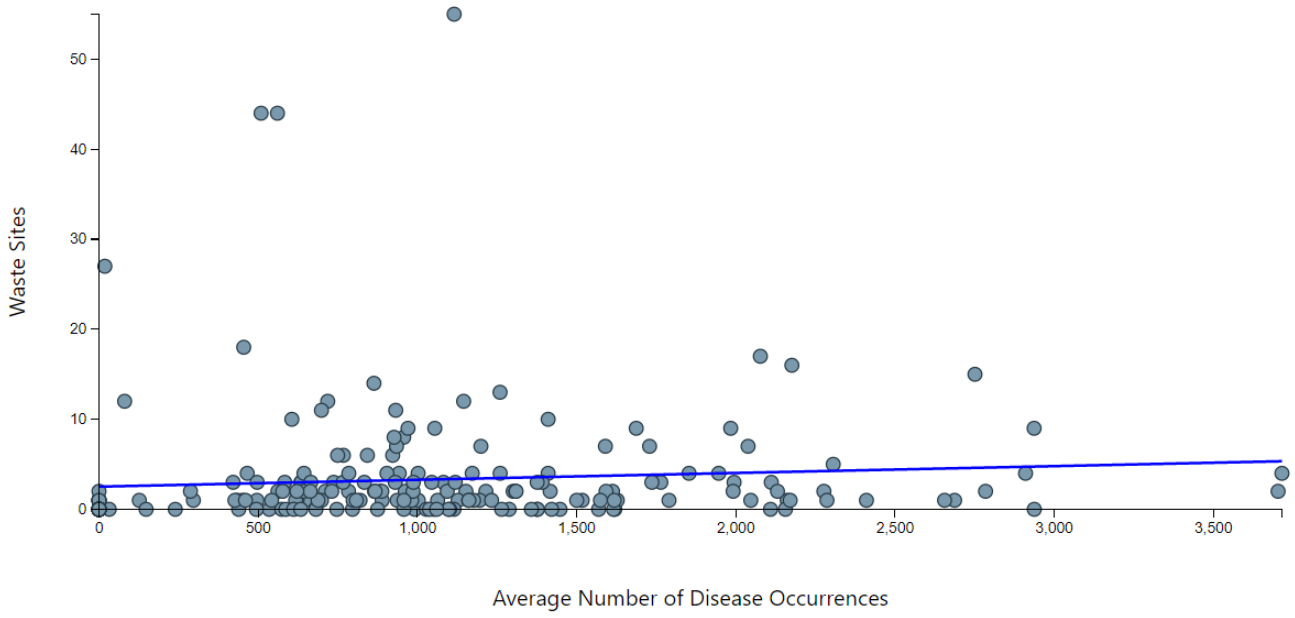
OLS Regression Results

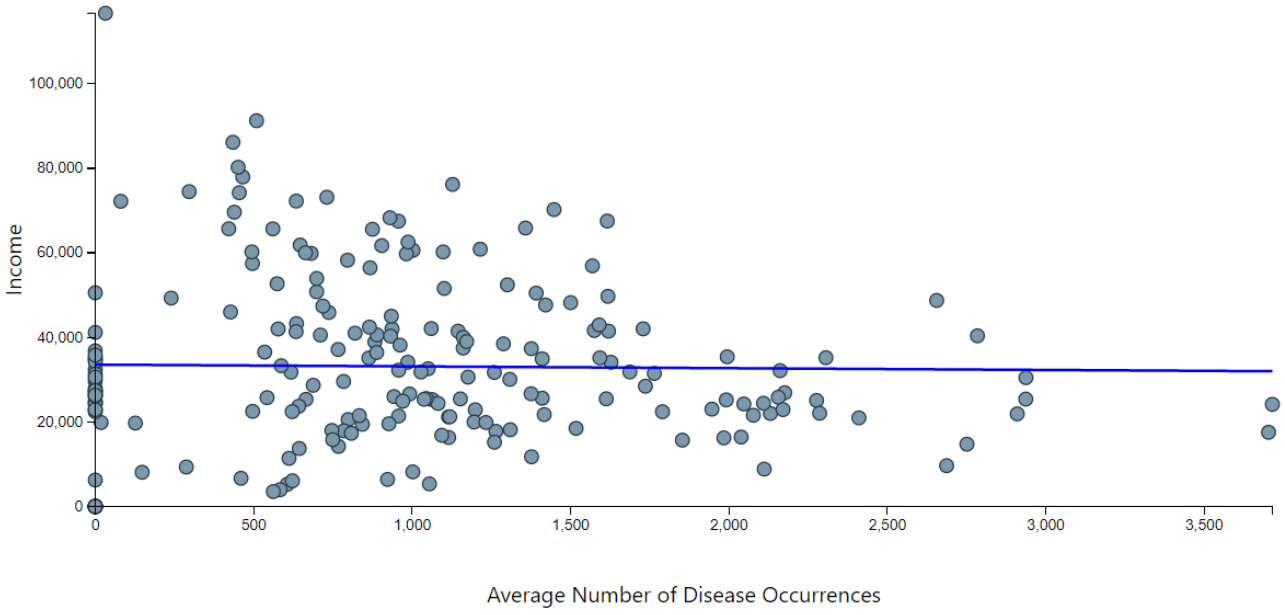
Dep. Variable:	y	R-squared:	0.021
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.8677
Date:	Thu, 02 May 2019	Prob (F-statistic):	0.504
Time:	20:22:13	Log-Likelihood:	-297.19
No. Observations:	211	AIC:	606.4
Df Residuals:	205	BIC:	626.5
Df Model:	5		
Covariance Type:	nonrobust		

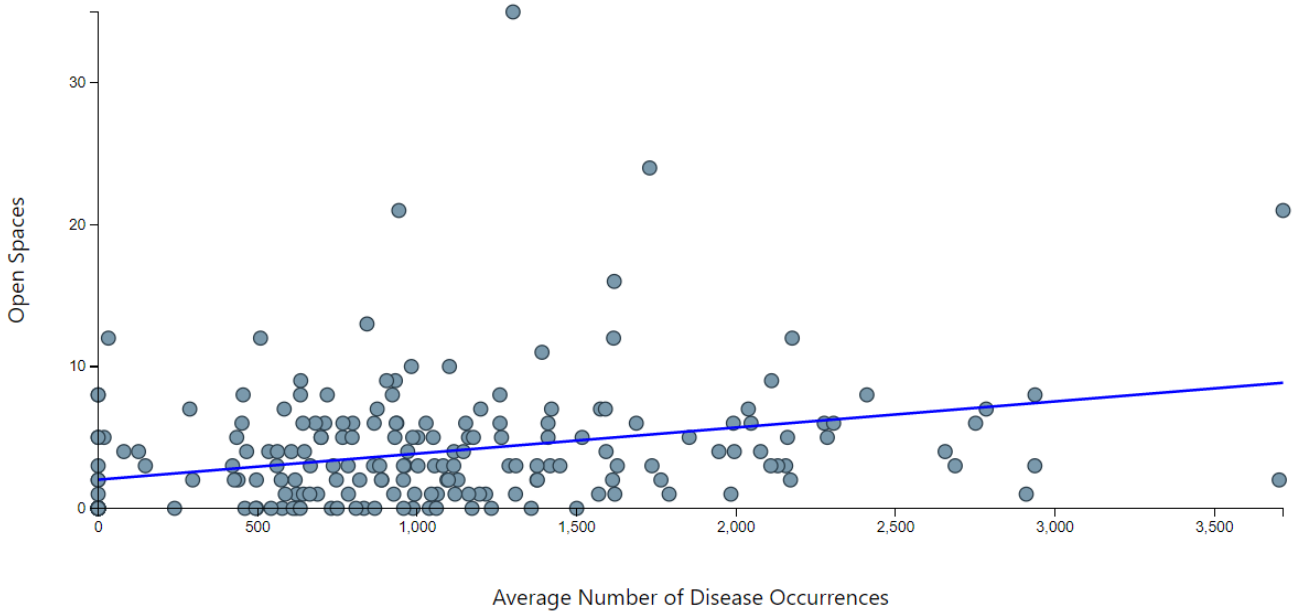
	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.334e-17	0.069	-7.72e-16	1.000	-0.136	0.136
x1	-0.0402	0.076	-0.528	0.598	-0.190	0.110
x2	0.0220	0.072	0.304	0.761	-0.120	0.164
x3	0.0338	0.071	0.479	0.632	-0.105	0.173
x4	0.0215	0.074	0.290	0.772	-0.125	0.168
x5	0.1393	0.071	1.952	0.052	-0.001	0.280

Omnibus:	275.937	Durbin-Watson:	1.694
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12294.731
Skew:	5.868	Prob(JB):	0.00
Kurtosis:	38.507	Cond. No.	1.66

Figure 6 – Scatter Plots




Income
100,000
80,000
60,000
40,000
20,000
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
Average Number of Disease Occurrences


Open Spaces
30
20
10
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
Average Number of Disease Occurrences

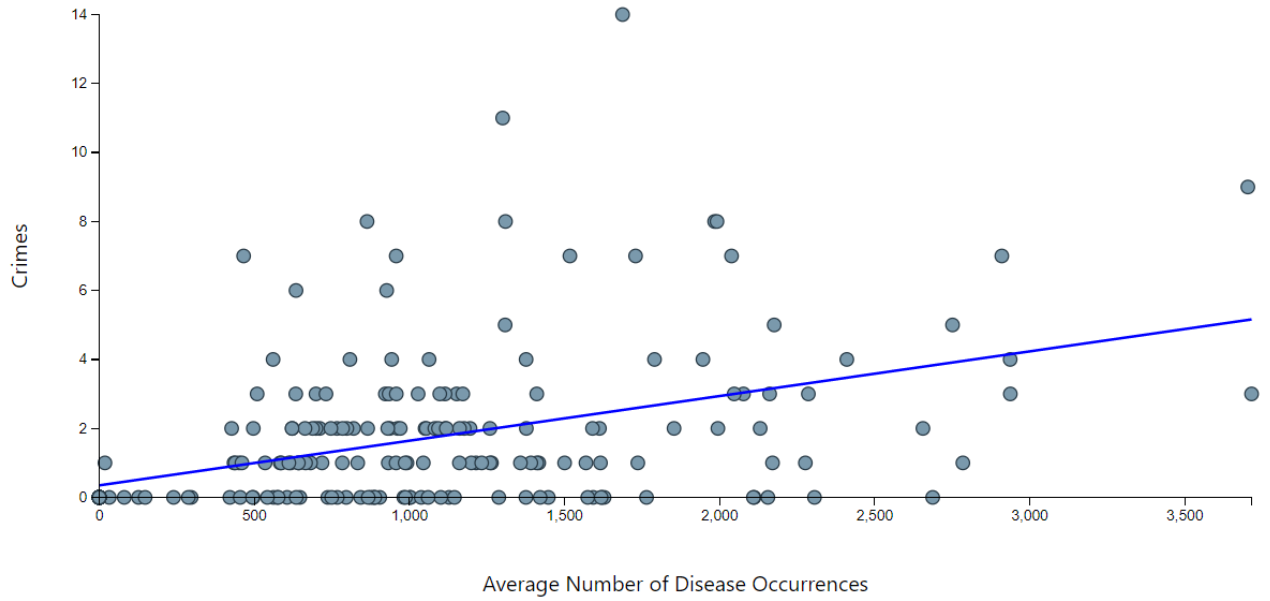


Figure 7 – Map of Custom Health Scoring Metric

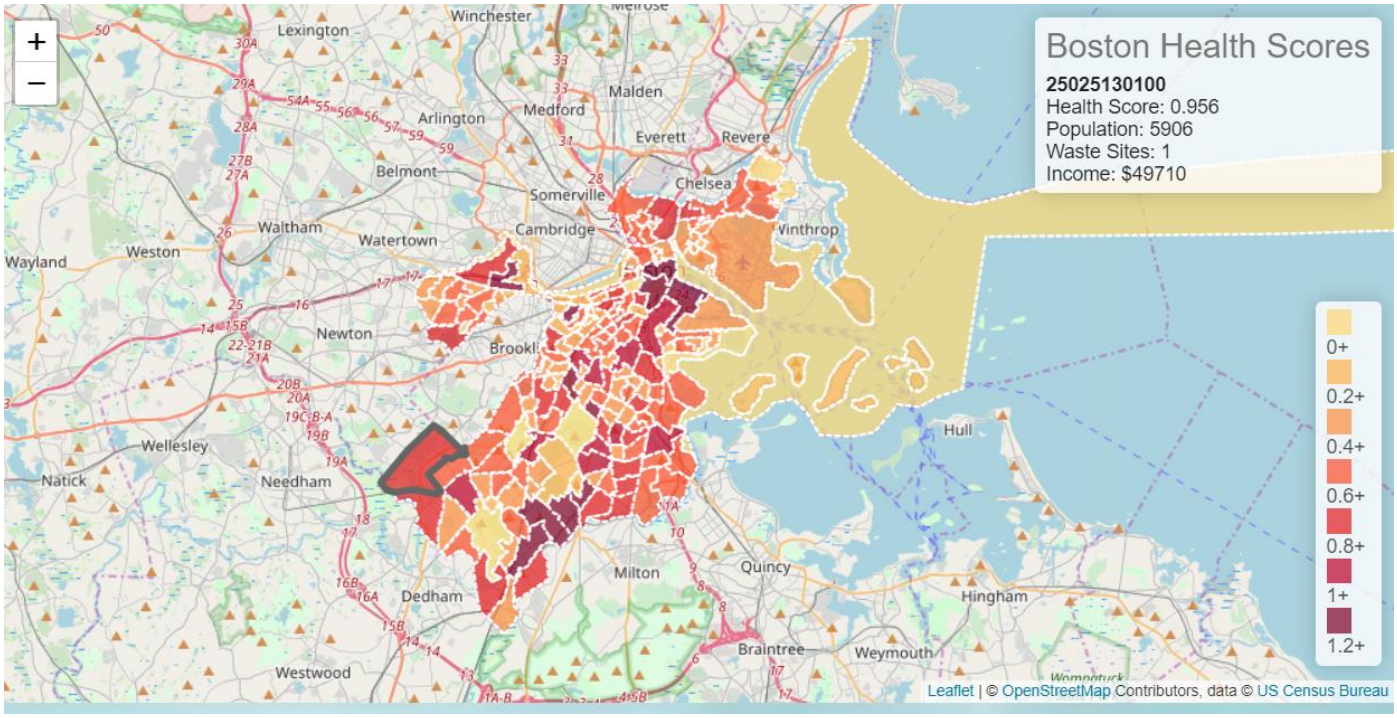


Figure 8 – Disease Prevalence with Optimal Waste Sites