Data Open

Nial Daly,Evan Hurley O'Dwyer,Bodgan Blaga, Luuk

November 2018

1 Topic Question

The goal of public health is to understand on a broad level, the contributing factors towards good health and ill health at a population level. The main end-goal of public health analysis is the identification of "easy targets" for increasing the health of citizens - the low hanging fruit. Much of the global increase in health over the past decades has been driven by data-driven initiatives. It is essential to understand what factors influence health if we wish to enact public policy to increase health. Understanding health at this level is notoriously difficult for many reasons. Given the variety of factors that modify disease-risk including family genetics, race, diet, environmental pollution, smoking, alcohol use and much more, it can be extremely difficult to distinguish genuine causation from correlation (something that has been often mentioned in the literature surrounding epidemiological nutrition studies).

In this challenge, we attempt to unravel some of the factors going into health for members of New York state. Our topic question is as follows: What are the main factors, including socioeconomic and environmental, that impact on the health of New York state citizens?

2 Executive Summary

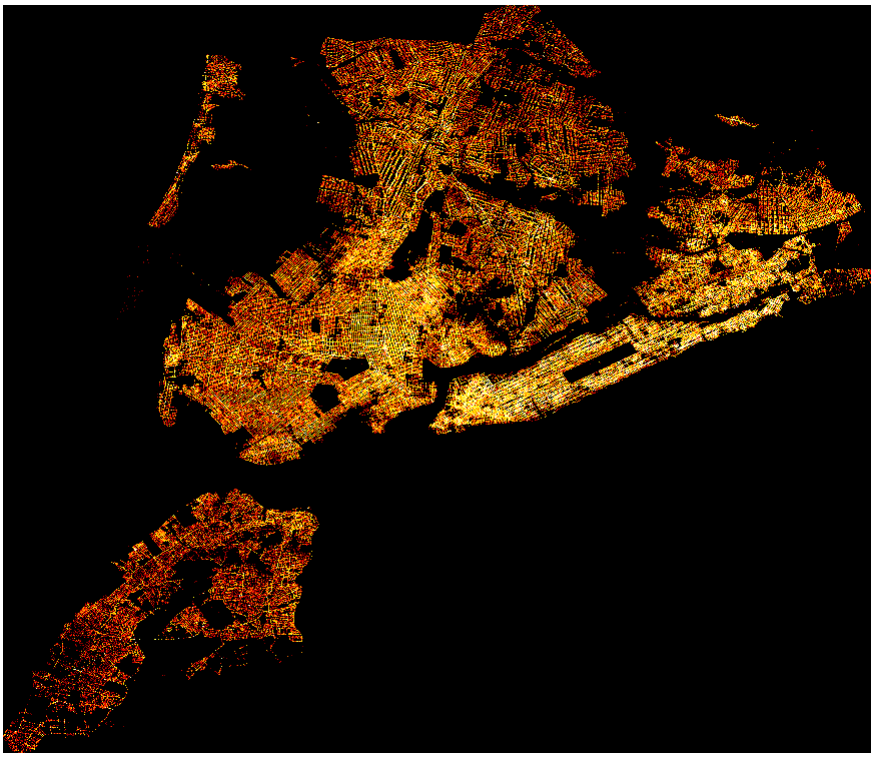
In summary, our findings were as follows:

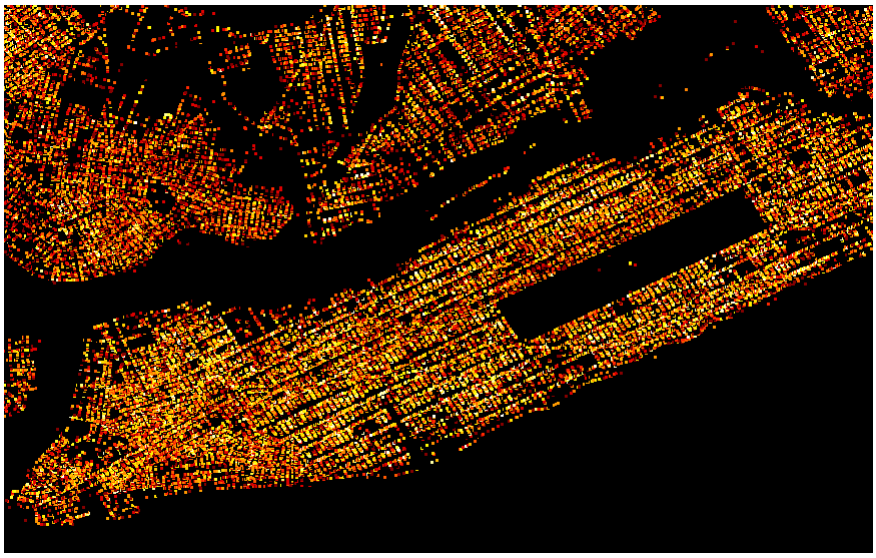
- High county median income is associated with a lower risk of a variety of obesity-associated conditions
- More specifically, low income counties had statistically significantly higher rates of overweight/obesity, stroke/heart attack, as well as diabetes.
- Radon exposure was positively correlated with lung cancer incidence and overall cancer incidence, however this was not strictly significant.

3 Initial Exploration

3.1 311 Calls

The 311 service database was a large database with roughly 1 million records. Each row represented a call to the 311 service, with an indication of the reason for the call, as well as the coordinates for the location of the call and the timestamp associated with it. The variety and volume of locations from which calls were made is vast enough that one can map the entire state through the calls (for example, one can clearly make out Central Park):





Predictably, the brightest area is represented by the main city area by Manhattan, where the population density is highest, although in general the entire area of the city is reasonably well marked by calls.

3.2 Radiation exposure and cancer

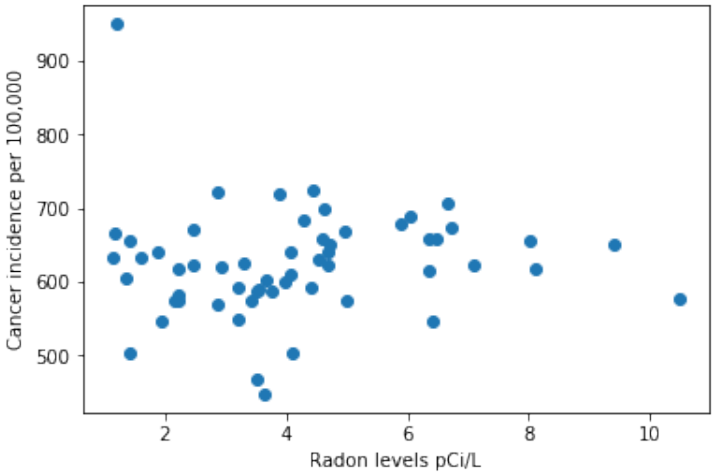
It has long been known that high exposure to background radiation is linked to the development of cancer. According to the Environmental Protection Agency, radon-exposure induced lung cancer causes 21,000 deaths a year in the US. Residential exposure to radon has been shown to increase the risk of lung cancer to that similar of passive smokers. With this background information in mind, we decided to look at the data to see if we could identify increased cancer rates in counties of New York with higher levels of radon.

3.2.1 Exploratory Data Analysis

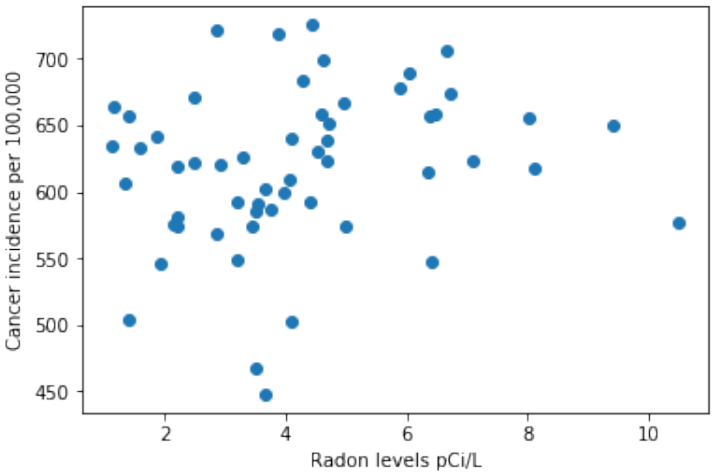
The radiation dataset contained much information on radiation levels in various areas of New York. In order to keep the scope of the search short, we focused specifically on radon exposure, as radon is the most commonly implicated radioactive pollutant with respect to public health. Having looked at the data, there were a little more than 30 measurements for radium (the element responsible for radon), which was not very helpful when comparing with the health outcomes for more than 50 counties. We used a publicly available dataset from https://www.health.ny.gov/environmental/radiological/radon/county.htm to get the information on radon for each county.

The community health data was filtered for cancer-related information. This

allowed us to visualize the relationship between radon levels and cancer incidence (all causes):



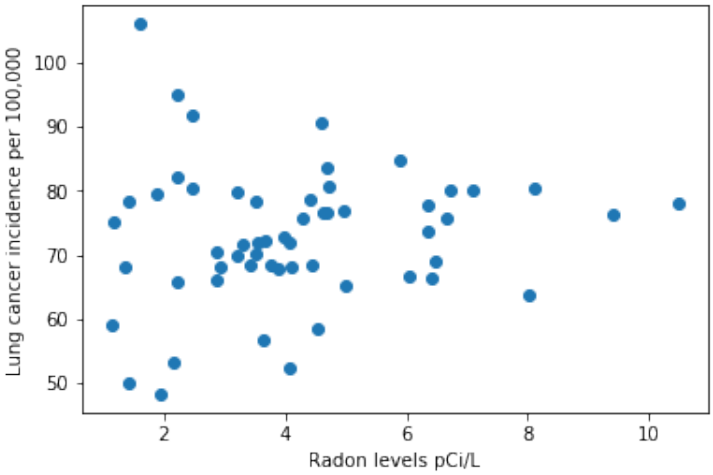
There was one noticeable outlier, which upon further investigation represented Hamilton, a county of New York which had a population of fewer than 5000 people, so it was removed from the data (all other counties had populations much higher). The correlation coefficient for the two variables before removing this outlier was .04, however after removing the outlier, the correlation coefficient was .20, relatively much higher, with the following scatter:



The positive correlation coefficient between the two variables is in line with epidemiological studies which indicate an increased risk of cancer associated with radon exposure, however some limitations are present. The correlation

coefficient was associated with a p-value of .14, which is not strictly significant. Of course, radon exposure is only one of many causes of cancer, so one could not expect a very high correlation coefficient when so many other variables may be at play in influencing cancer risk in various counties (diet, lifestyle factors, etc). Also, the risk of cancer increases with time, so the 2012 levels of cancer may not be modelled well by the 2018 levels of radon gas.

When investigating purely with respect to lung cancer, we get an even lower correlation coefficient, .11, which is again inconclusive with respect to confirming the link between county radon levels and cancer rates:



3.3 Fast-food restaurants per neighbourhood in New York City

First of all we needed to select the specific data on which we wanted to perform our analysis. Most importantly we needed to know the density of fast food restaurants in the particular communities within the state of New York for which we were provided the data set. The table "food venues" in the data set contained the column "category", so the we first filtered all the ones labeled "fast food" however this gave not enough data since we noticed that restaurants like McDonald's weren't labeled "fast food", therefore we decided to filter the data by name and selected the data of the top 10 fast food restaurants in the US, as given in reference [[**?**]] as reported by the specific corporations. This gave us 2228 fast food restaurants in the city New York.

We needed to group those by borough (NTA) in New York, the way we did this was by creating the polygons of the table "geographic" of the NTA and then the ideas was check if the given longitude and latitude of the specific restaurant was inside this polygon. Unfortunately we were not able to plot the polygons

or check if the restaurants were in a given NTA.

To calculate the density of fast food restaurants we wanted to divide the number of restaurants in the same NTA by the people in that NTA.

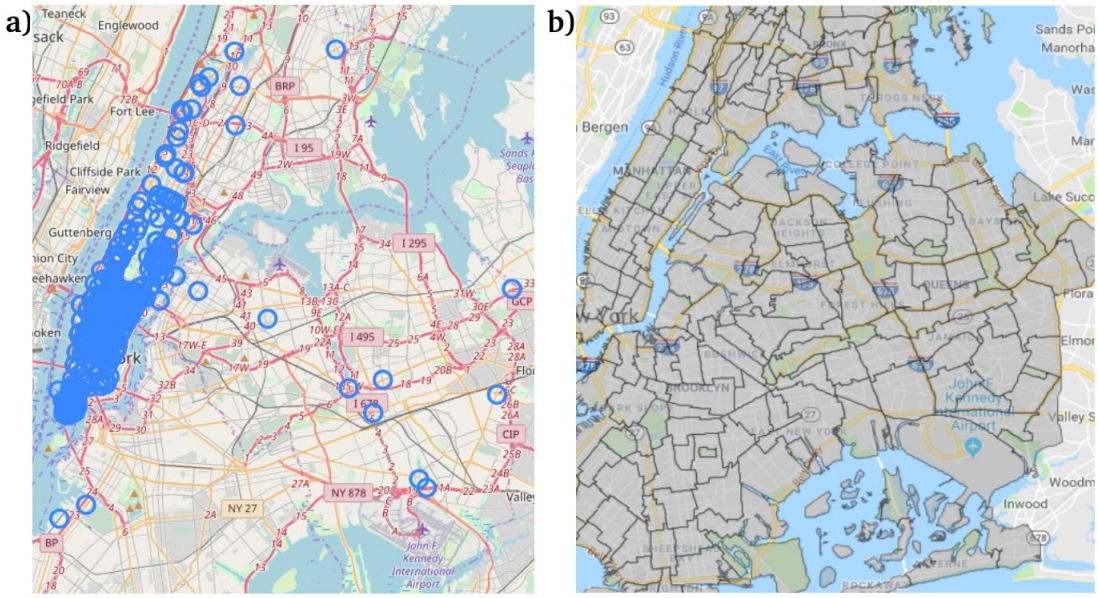
$$\rho_{fastfood} = \frac{\#fastfoodrestaurantsinNTA}{\#peopleinNTA} \tag{1}$$

where $\rho_{fastfood}$ is the density of fast-food restaurants. We normalised this data to get not too small values

$$X_norm = \frac{X}{\sqrt{(\sum_{n=1}^{N} X_n^2)}} \tag{2}$$

The result we get for this is a column with $\rho_{fastfood}$ per NTA. Then we use the "demographics city" table to get the column "mean income" in those states. Unfortunately we didn't have time to test the correlation between those two tables.

The results are shown in the figure below.



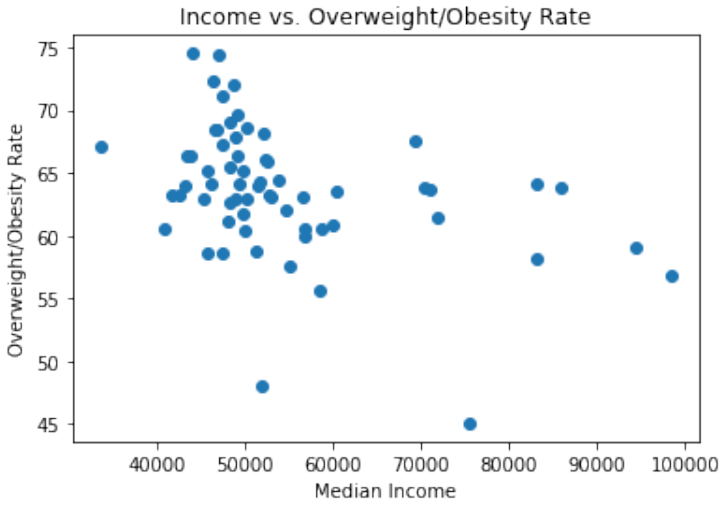
From these two maps we can see that there is no correlation between high income neighbourhoods like Manhattan or low income neighbourhoods like the Bronx and the number of fast food restaurants.

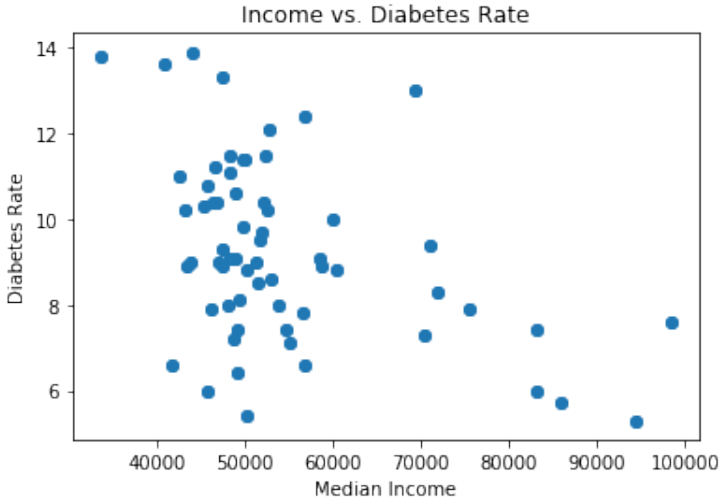
3.4 Income and Health

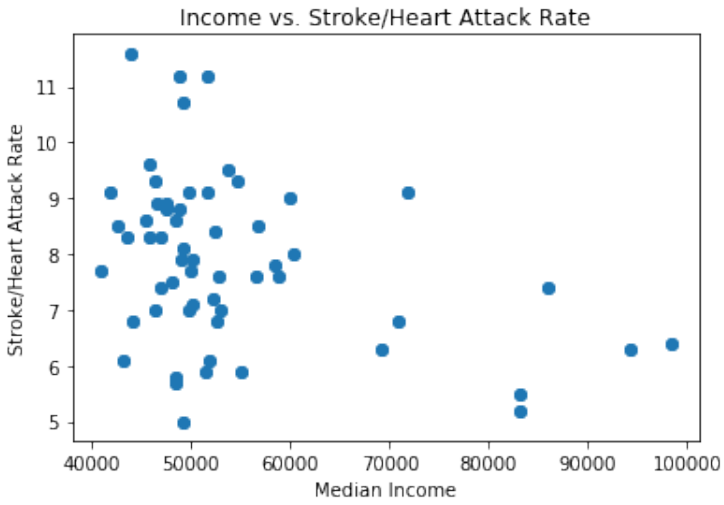
As part of our investigation into the factors determining the health outcomes at a public level, we investigated the relationship between income at a county level and various health conditions.

Among the health conditions analysed were overweight/obesity (defined by a

BMI > 25), diabetes and stroke. The following scatter plots give an indication of the relationship between the median county income and disease risk:







Respectively, the correlation coefficients were -.39, -.44 and -.36, i.e. among all conditions, low income is associated with increased risk. All three correlations were statistically significant (respective p-values being 0.001, 3.0e-07 and 8.4e-05).

One reason for this correlation is the fact that people in higher-income families have access to better education which in turn results in better choices regarding nutrition and diet. Having more awareness about what is good and what is not good to eat helps prevent certain diseases, especially the ones we pointed out. Another reason could be the access to higher-quality food. People from lower-income families, especially from poor neighbourhoods don't have access to quality food and so are more prone to diseases.

4 Conclusion and limitations

There were several assumptions that were limitations during our investigation. Given more time, we would like to investigate whether hidden variables were also involved in certain associations. For example, different races have different risks of certain diseases such as diabetes, etc. A full analysis would need to take other variables such as this into account. As another example, education may play a role in diet, however we did not have data relating to education differences between counties.

In conclusion, our recommendations are:

- Government initiatives to reduce income inequality may help health outcomes
- Further research should be carried out to investigate radon emissions and cancer

- Initiatives to reduce overweight and obesity should be positively supported
- No positive correlation between fast food restaurants and income (based on limited study).