The impact of the food quality and rent prices on the residents' health

Jason Chang, Yusuke Kikiuchi, Jiying Zou

Topic Question: What factors are correlated with worsening health in NYC, and can we predict NY residents' health condition from housing qualities?

Ever since the 19th century, the Statue of Liberty in New York's coast has been a shining distant symbol of immigration and freedom for thousands of immigrants abroad. From then until now, although the demographics and purposes of immigrants have largely shifted, one thing has remained the same: the allure of the Empire State as an epitome of a cosmopolitan, rapidly advancing place to live. Young dreamers in finance and economics rush over to the infamous Wall Street, tourists crowd the city streets daily sporting "I Heart NY" shirts, and long-established families flourish in the bright economic playground.

As the population of New York rises, so does a plethora of public health concerns. In a land where opportunities may be infinite, resources are not. Some cities have it worse than others -- as can be seen in the map below, much more of the population is crowded into Manhattan than other boroughs, with the densest regions reaching over 85 people per acre:

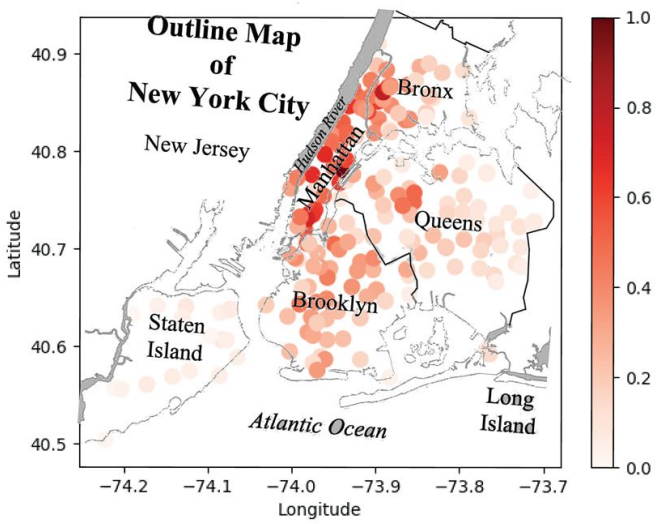

Fig.1 Population density in New York City. Darker red indicates higher density, in terms of residents per acre.

Not only are some places being crowded like a pigpen, it is not cheap to live in these areas either! Rent has become less affordable since 2014, with family-oriented living arrangements witnessing an overall increase. This puts a lot of long-term pressure on the proportion of the

residential population likely to stay long-term (in comparison to students or other short-term residents), which could translate into health implications:

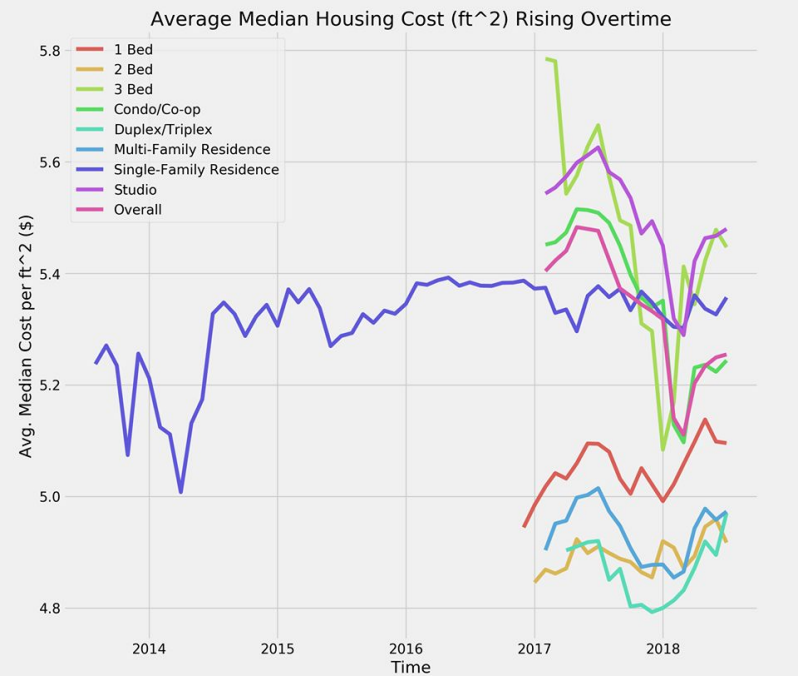

Fig 2. Median rent prices (per square feet) are on the rise for family-friendly places (single-family residences, duplexes, triplexes, multi-family residences, and 2-bedrooms).

Faced with raising rent prices, some residents opt to skip meals in order to save money to pay rent. There is evidence that people living in major cities such as NYC experience better health when their wages are increased (hungerfreenyc.org), but with the economic pressure nowadays, this does not always happen. It is widely accepted that malnutrition resulting from lack of nourishment will lead to dire health consequences. The implication is that people's health suffer by choosing less nutritious lifestyles in order to save money to pay ever increasing rent.

On the other side of the story, restaurants also suffer from land becoming more and more expensive. Restaurants thus may be incentivized to spend less time and resources on sanitation and upholding good practices so that they can maximize their profits and pay their own rents. In fact, there is an undeniable rise in the proportion of restaurant inspections citing critical violations from 2009 to 2017:

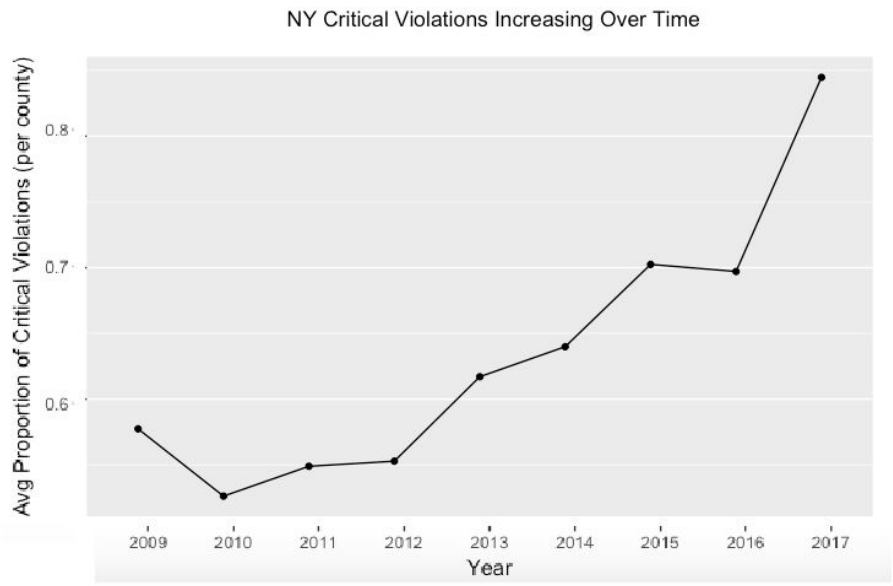

Fig 3. The average proportion of restaurant inspections returning with critical violations per county (for ~40 counties) has increased rapidly from less than 60% to almost 90%.

This is similarly problematic in that for those who choose to eat at restaurants, which many young full-timers do, they are not only ingesting food of lower nutritional value than something home cooked, but they are also exposed to some type of sanitation or proper food preparation issue that also puts their health at risk.

From consideration of housing variables and we were able to predict the health of residents with 40% accuracy.

Technical Summary

We first sought to explore some characteristics of the New York area, to see if there are correlations between certain factors driving public health concerns.

1) Population Concentration

From the “demographics_city” table, we derived the average population density (people_per_acre) for certain geographical locations. We grouped by nta_code and took the average latitude and longitude for each NTA region to plot a map overlay of the population density per acre. This gave us an idea that Manhattan was amongst the most crowded boroughs of the New York City area.

2) Rent Price

We realized that most people devote a significant proportion of their salary to rent. We pulled rent data for several different types of housing from the Zillow website and was able get the average of median price per square foot for rental listings over the past decade or so. This external data set showed us the significant trends in rent prices over time, and confirmed our hypothesis that rent has been on the rise in the past decade (although we were not able to find out why rent dropped recently, the quick bounceback does not annul our hypothesis).

3) Restaurant Inspections

To build out the other (restaurant) side of the story, we looked at the "food_service_establishment_inspections" table to aggregate the total proportion of restaurants with critical violations over time. We looked at some of the non-critical violations and deemed that the critical violations were more likely to affect people's long term health, so we worked with the latter. Statistical rigor was taken to make sure that the number of boroughs covered were relatively the same, and earlier years when inspection information was scarce and 2018 when inspection information was incomplete were thrown out to maintain data integrity. The proportion was calculated rather than a raw number in order to control for different numbers of inspections performed, and all sample sizes for inspection numbers were large enough (in the hundreds) to make argumentative sense and be robust against outliers.

4) Health Outcome

We struggled to come up with an outcome variable indicative of short-term residential health, but ultimately decided to create a weighted sum of the 15 health indicators in the "community_health" table. The weights were manually created through evaluation based on the top causes of premature death in 2015 in NY, as can be found here:

	Number of deaths and age-adjusted death rate					
	Total Deaths	#1 Cause of Death	#2 Cause of Death	#3 Cause of Death	#4 Cause of Death	#5 Cause of Death
2015	Total Deaths 53,786 587.0 per 100,000	Heart Disease 17,212 186.2 per 100,000	Cancer 12,618 138.5 per 100,000	Pneumonia and Influenza 2,101 22.9 per 100,000	Unintentional Injury 1,952 21.6 per 100,000	Stroke 1,932 21.2 per 100,000

Fig.4 Source: https://apps.health.ny.gov/public/tabvis/PHIG_Public/lcd/reports/#state

Health indicators such as cardiovascular disease are ranked as high as 5, and and indicators that contribute less to premature death are ranked lower. The rankings are as shown below, and these are used in a weighted sum multiplied with the prevalence of each indicator, in each county.

To measure the health situation, we defined a metric called "health score"
The definition of the health score is given by

(health score) = $\sum_i w_i p_i$,

where the values of w_i's can be found in the table below and p_i is the percentage of the ith indicators in the "community_health" dataset (if the original data is given in rate, it is converted into the corresponding percentage).

w_i's are determined based on the impact from the food quality, the bigger, the higher weight.

https://apps.health.ny.gov/public/tabvis/PHIG_Public/lcd/reports/#state

indicators	w_i
Family Planning/Natality Indicators	2
Cancer Indicators	6
Oral Health Indicators	1
Maternal and Infant Health Indicators	2
Injury Indicators	3
Socio-Economic Status and General Health Indicators	5
Cardiovascular Disease Indicators	6
Child and Adolescent Health Indicators	2
Obesity and Related Indicators	6
Cirrhosis/Diabetes Indicators	5
HIV/AIDS and Other Sexually Transmitted Infection Indicators	2
Respiratory Disease Indicators	4
Tobacco, Alcohol and Other Substance Abuse Indicators	5
Communicable Disease Indicators	4
Occupational Health Indicators	1

5) Machine Learning

With the 8-year price-per-square-feet data acquired from Zillow, we first isolated and extracted only the data associated with the counties from New York state. Subsequently, we attempted to construct a regression model by applying method of Support Vector Classification(SVC) to solve regression problem, also known as Support Vector Regression(SVR). We decided to implement this regression algorithm in comparison with other methodologies since SVR because of its dependency on only the subset of the input data due to the fact that cost function for SVR model construction ignores any training data close to the model prediction. SVR also allows us to classify high-dimensional data provided by the monthly time-series data provided by Zillow within the 8-year time interval. The objective of the experiment is to use housing market price to predict the health condition of the local residents. Again we hypothesize that in areas with rising housing price, local residents tend to sacrifice part of food expense to pay for the expensive rent. To measure the health quality of the local resident, we used the unique weighted health measurement as described *a priori*. We tested our regression model by splitting the dataset into 20% training set and 80% testing set. In the end, our methodology yields an accuracy of approximately 0.42867.

As for the future optimization of the accuracy of our model, we realized our original Zillow dataset was missing values in the middle time period. Hence, a more complete housing market dataset should improve the accuracy.

Conclusively, our analysis and experiment have illustrated that housing quality is a good predictor for the health and socioeconomic condition of local residents.

Limitations

References

1. https://www.hungerfreenyc.org/sites/default/files/atoms/files/2016%20Annual%20Hunger%20Survey%20Report%20Final.pdf