The Significance of the U.S. Government Sanitation Regulations on Public Health

- TheDataOpen 2018 Project

Prepared by: *Yufeng Wang, Frank Han, Hanming Lu*

University of Waterloo
May 12th, 2018

Executive Summary

Mahatma Gandhi once said: It is health that is real wealth and not pieces of gold and silver. Unfortunately, while the US is a leader in military, economy, and cutting-edge technologies, recent reports show a downward health trend in New York State. Some people argue that the cause of current health issues is subject to food quality control, and others claiming that the root cause is primarily due to high health care costs.

Many policy makers have attempted to improve public health through decreasing health care costs, which only tries to mitigate the problems when they occur, while neglecting the reasons why they occur in the first place. To find out the substantial reasons behind public health problems, we came up with three specific questions: What are the impacts of enforcing sanitation regulations? Does efficiency in regulation enforcement affect the outcome? And what factors have the highest predictive power on the major health level indicator?

Our analysis consists of two parts: exploratory data analysis and visualization and predictive analysis based on regression model. From these two sub-analyses, we generated three key findings. First, a higher restaurant inspection frequency can effectively control the amount of critical sanitation violations in the local restaurant. On top of that, high efficiency in health-related government departments can significantly reduce the amount of critical violations. In addition, both XGB and GLM models illustrate that income level and critical violations have the greatest predictive power on cancer with a great statistically significance in both.

1 Introduction

In order to understand and improve the governmental initiatives and socioeconomic factors, we want to investigate the factors for local health and focus on the effects of government regulations on major health level indicators. Since each state maintains its own health indicator, statewide actions naturally have a pronounced effect on the performance of that state. Specifically, the question we seek to answer is: What are the impacts of existing sanitation regulations and how can government improve? Moreover, what factors have the highest predictive power on the health level indicator? Does efficiency in execution affect the result? We use our analysis to make statewide policy recommendations that will improve the New York State's performance as a whole.

2 Exploratory Data Analysis and Visualization

Sanitation level in public areas is critical to public health since diseases can spread out easily in public areas compared to private places. Although restaurants are not public areas, a critical sanitation violation in a restaurant can instantly affect all clients through food or air pollution. Therefore, restaurant sanitation regulation is one of the most critical public health regulations, on which we will conduct a county-based analysis and visualization.

2.1 Restaurant Inspection Frequency vs. Critical Violation

To answer the questions: "What are the impacts of enforcing sanitation regulations?", we dived into the 311 service requests and food establishment inspections datasets, along with demographics information as a supplementary dataset. First of all, we performed data wrangling by filtering out invalid entries (e.g. invalid city), aggregating all entries by city, mapping them to the specific county they locate in, and then generated the restaurant inspection per capital on the left-hand side along with the average critical violation per 100 inspections on the right-hand side.

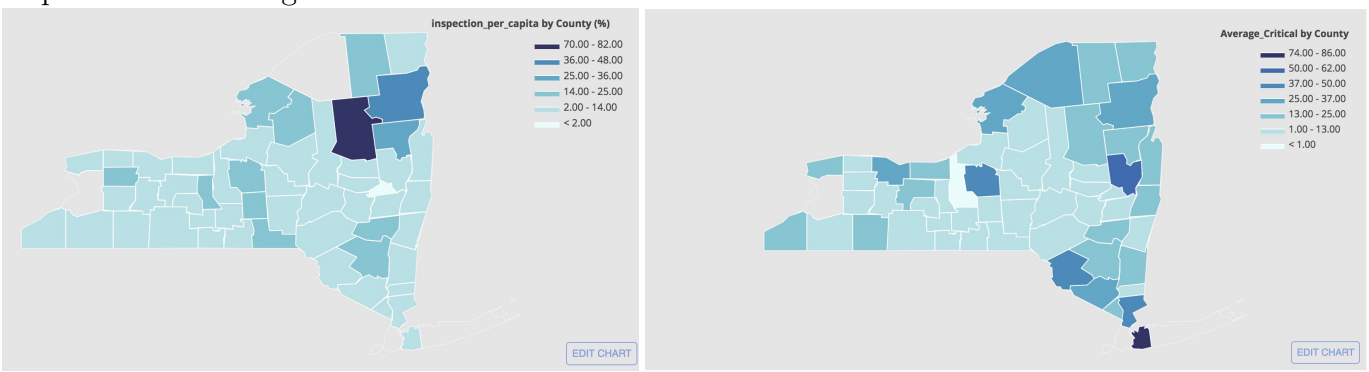

Figure 1: Inspection and Violation Distribution

By comparing these two plots, it illustrates a negative relationship between restaurant inspection per capita and average critical violation per 100 inspections. For counties that have a lower inspection per capita, such as the counties at the right bottom corner, we observe that they generally have a higher critical violation per inspection. This relationship

demonstrates that sanitation regulation enforcement from government is not only necessary but critical to restaurant sanitation control. Counties with less inspection frequency are more likely to have restaurants which perform illegal conducts. Therefore, enforcing sanitation regulations does have a critical impact on public health.

2.2 Request Handling Efficiency vs. Critical Violation

We want to analyze how the 311 request handling affects the number of critical food violations. Rather than the frequency of the request, we use the average resolve time to evaluate with the efficiency of resolution. We filter out the invalid entries(e.g. invalid city name, pending request) and categorize them to their corresponding city. Our hypothesis is that restaurants are less likely to violate sanitation regulations if the request resolution is more efficient. We averaged the response time from each city and ranked them in order, as shown in figure 3. The mean response time is around 700 minutes with little variance. However, in some cities, the waiting time is significantly higher. Does the long waiting time contribute partially to the sanitation violation?

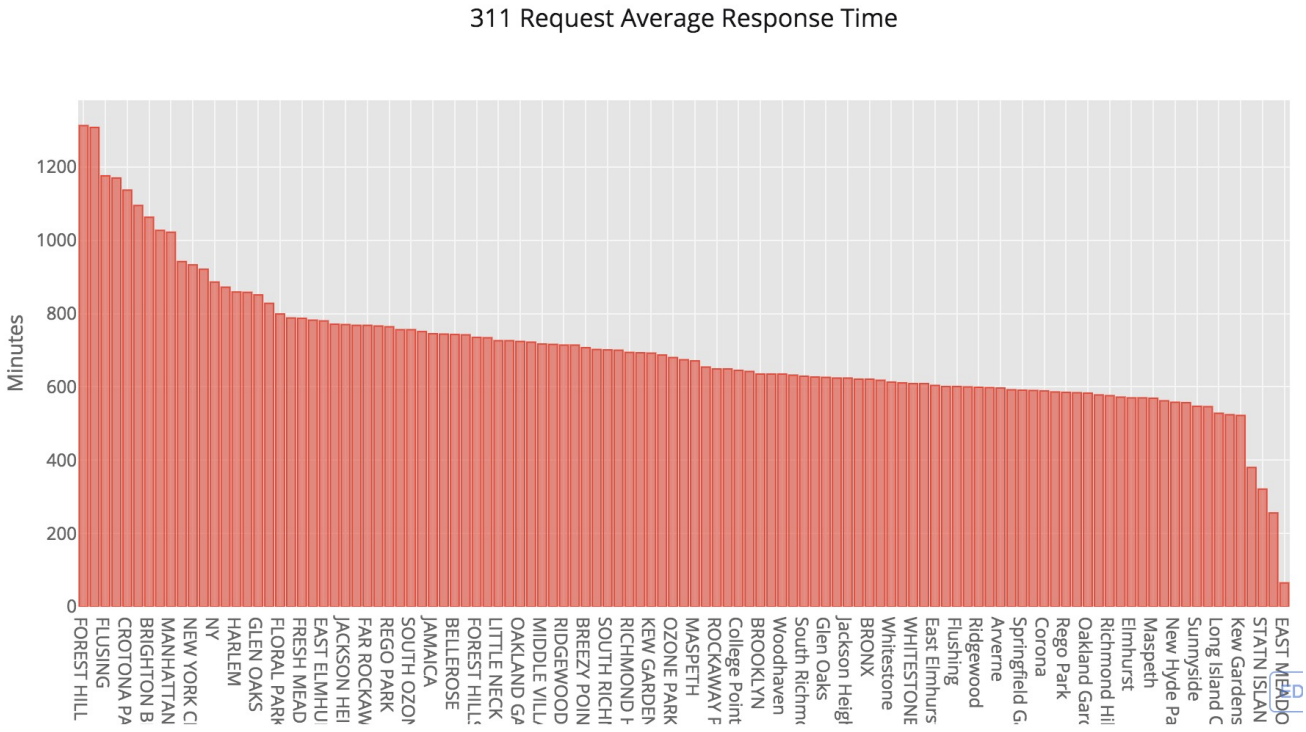

Figure 2: Average Request Waiting Time

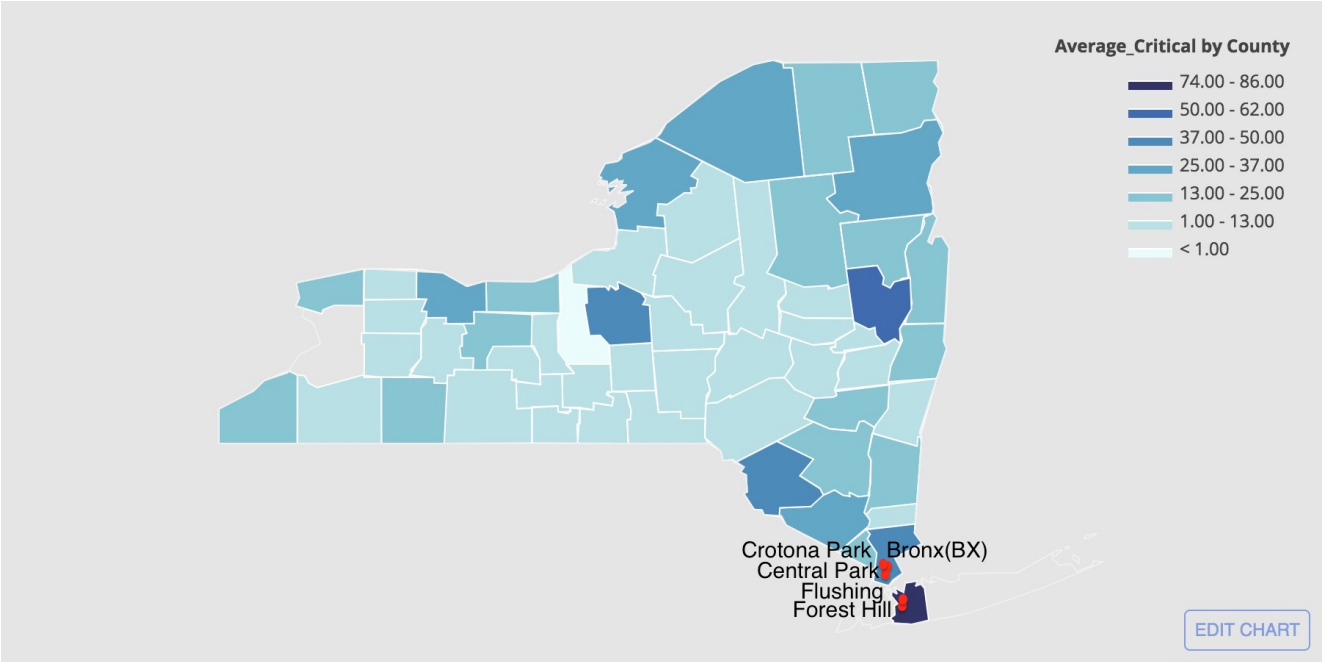

Figure 3: Average Critical Violations with Cities

We select a few cities with highest responses such as Forest Hill and then plot them on figure 1 to get figure 3. We can see that the areas with the longest average response time locates at the county with highest average critical food violations. Cities with fast response time usually ends up with less critical issues. This does prove our hypothesis. By observation, it shows the efficiency of health related department plays a very important role in preventing food violation and can be something that government can improve at.

3 Regression Analysis - Factors of Community Health

Our next goal is to look at possible socioeconomic factors and government-related activities and build regression models to quantify the effect of each factor on the health-quality level. Each sample in our model corresponds to each county in the New York State.

3.1 Target and Feature Selection

To evaluate the health-quality level, we look at all the health indicators available for use. Our goal is to use a universal health indicator that does not only apply on a specific condition or a group of residents. Therefore, from all the candidates, we select **"cancer incidence per 100,000 households"** as our target variable for the model. We do admit that this limits the use of the model since it only looks at cancer as a target, but within the time limit we have, we believe it is the best reprensentitive as a health indicator.

For feature selection, since the topic of interest is government related, we look at a set of features that government can potentially design specific policies for. Specifically, we look at **mean household income, critical violation ratios, non-critical violation ratio, private coverage, public coverage**. The definitions of these features are:

- **Mean household income**: Average income of all the households in the county
- **Critical violation ratio**: Ratio between the number of inspections with critical violations and the total number of inspections in the county
- **Non-critical violation ratio**: Ratio between the number of inspections with non-critical violations and the total number of inspections in the county
- **Public coverage**: Percentage of households in the county who are covered by public insurance policies

Before applying these features in the model, we perform a correlation analysis to see if features for the model are not unexpectedly strongly correlated with each other. See Table 1.

Correlation	Income	Critical	Non-critical	Public Coverage
Income	1	-0.090	-0.015	0.164
Critical	-0.090	1	0.558	0.153
Non-critical	-0.015	0.558	1	0.086
Public coverage	0.164	0.153	0.086	1

Table 1: Correlation Matrix

As we observe, none of features we select is strongly correlated to the other one. Critical vs. non-critical has the highest correlation of 0.558, but that is rather expected in the sense that restaurants that have critical violations are very likely to have other non-critical violations. Moreover, we would like to see if we are in the correct direction, so we perform the correlation analysis betwee target and features so we may able to filter out completely uncorrelated features. See Table 2.

	Correlation vs. Cancer percentage
Income	-0.130
Critical	0.169
Non-critical	0.037
Public coverage	0.087

Table 2: Correlation Table versus Target Variable

From the table, we have some evidence that all the features are correlated with the target variable. Therefore, we proceed with the target and features to the modelling process.

3.2 Model Architectures

We attempt to use two models, boosting tree and generalized linear model, for this specific problem. The reasoning behind using these two models is to investigate the importance of each feature in the model, and to quantify the effect of an increase/decrease in a feature to the response variable.

For the boosting tree model, we implement the model using the package **xgboost** for its overall robustness and flexibility among all boosting tree packages. For the generalized linear model, we implement an ordinary least square regression so that we are able to predictively measure and interpret the influence of a feature.

3.3 Results

After training the aforementioned two models on the dataset, we would like to see which feature is the most important among the four. We measure the importance of each feature by calculating its F score within the model. See Figure 4.

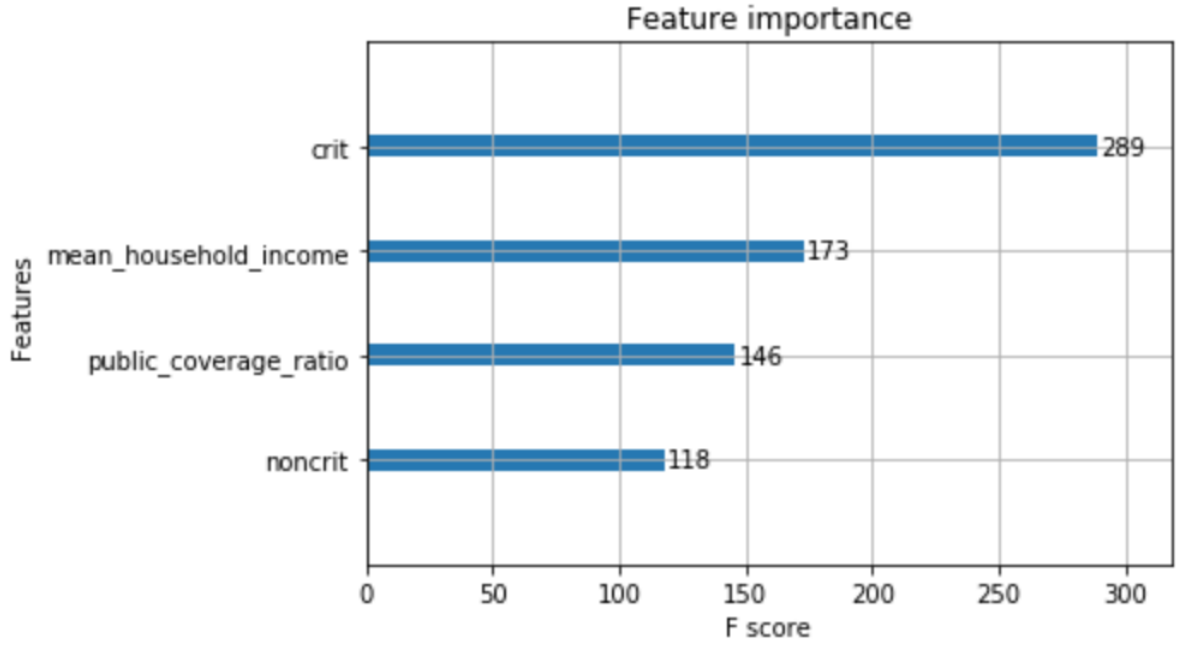

Figure 4: Feature importance

From the figure, we observe that **critical violation ratio** is the most important feature among the four, and its score is significantly higher than the other three.

We proceed to train the generalized linear model on the normalized dataset and the weights of each feature are computed. See Table 3.

Feature	Weight (in unit of e-04)
Critical	3.5
Non critical	-0.76
Income (in 10k)	-2.15
Public Insurance Coverage	-2.48

Table 3: Weights of features

Due to the time constraint, we are unable to estimate the outcome for all the improvements. However we made a predictive model for the effect that improved critical violation ratio can have on reducing the number of local health issues. Quantitatively, by reducing 10% of the critical incident, we can effectively reduce the 0.7%. This analysis is rather shorthanded on numerical value, but this still agrees with the result from the previous xgboost model.

4 Conclusion

From the scope of our data sets, we see that government regulations and their enforcement have a statistically significant impact on public health. Specifically, a higher restaurant inspection frequency can effectively reduce the amount of critical sanitation violations in general. In addition, the efficiency of health-related departments also has substantial impacts on the chance of restaurant critical violation. On top of that, both XGB and GLM models demonstrate that income level and critical violations have the greatest predictive power on cancer on a county-basis.

In conclusion, since income level is not a factor that the government can easily modify, it further suggests government to focus on restaurant sanitation violations. we recommend that the New York State government to focus on enforcing sanitation regulations on a higher frequency and improving efficiency in health-related departments to improve the general public health.

With limited government resources and overwhelming amount of restaurants, it is indeed difficult to do multiple inspections and immediate responses for all of them. However, we suggest some future research on how the type of restaurants and frequency of violations are correlated, as government can use the correlation to regulate the opening of local restaurants as a source of the sanitation problem, and essentially improve the local health quality level.