Analysis of restaurant grading and 311 complaints in NYC

Adam Kelbl

Introduction

This paper documents the process of my analysis of the NYC restaurant inspection and 311 complaints datasets as a part of the Datathon entry task.

I approached this assignment as a process of learning working with data analytical tools and further developing my analytical thinking. Because I'm still not too fluent in Python and its libraries, I used AI tools like ChatGPT and GitHub Copilot to help me write some parts of the code. I am however doing my best to learn this language to be able to use it for working with data.

Goal

The goals of this analysis are to explore and describe the distribution of restaurant inspection grades in NYC and to investigate whether there is any statistical relationship between bad restaurant grades and volume of public health and hygiene complaints on the New York 311 line.

Basic exploratory analysis

First, I analyzed the NYC Restaurant Inspection dataset, focusing on grade distribution across the city, its boroughs and ZIP codes. Each restaurant receives a grade of A, B, or C, ranked from best to worst.

i) Grades across the city

I first calculated the overall distribution across NYC. The pie chart below shows that 77 % of restaurants received grade A, while 23 % were graded B or C.

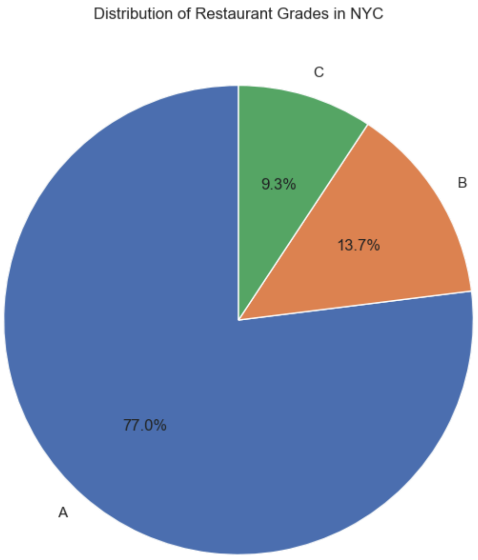

Figure 1: Distribution of restaurant grades in NYC

ii) Grades across the NYC boroughs

New York City is divided into 5 boroughs, I computed the distribution of grades in each of them.

Borough	A	B	C
Bronx	74.6	16.6	8.7
Brooklyn	77.0	13.7	9.2
Manhattan	79.1	12.2	8.7
Queens	73.9	15.0	11.1
Staten Island	80.4	14.1	5.6

Table 1: Distribution of restaurant grades across NYC boroughs

Across all boroughs, the distribution is relatively consistent, with about 75 % of restaurants graded A and 25 % receiving a B or C. Staten Island has the best inspection results among NYC boroughs, with the highest share of A grades (80.4 %) and the lowest share of C grades (5.6 %). Queens has the worst inspection results among NYC boroughs, with the lowest share of A grades (73.9 %) and the highest share of C grades (11.1 %).

iii) Distribution across ZIP codes

The dataset includes over 200 ZIP codes and I calculated the grade shares for each. The detailed data is available in the Jupyter notebook attached, but it's not possible to visualize it here. This area could be explored further in future work, potentially with a deeper dive into spatial patterns.

Correlation analysis

Using Pearson's and Spearman's correlation coefficients, I checked whether there is any relationship between bad restaurant grading (B/C) and number of hygiene and health related complaints across NYC ZIP codes. Pearson correlation coefficient is 0.136, and Spearman correlation coefficient is 0.089. These results show only a very weak connection between the share of bad restaurant grades and the number of complaints. This suggests that bad restaurant grades and complaints don't strongly go together and other factors might affect both. This is confirmed by the following graph of the relationship between these two variables:

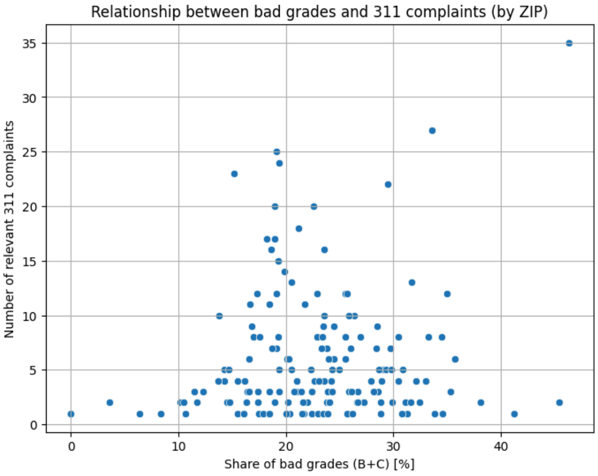

Figure 2: Relationship between bad grades and 311 complaints

Modeling the relationship with Linear Regression

Next, I tried to model the relationship between the number of complaints and share of restaurants with bad grades using Linear Regression model via the OLS method. The model is simple, using only one explanatory variable:

$$\text{Number\_of\_complaints} = \beta_0 + \beta_1 \times \text{Percentage\_of\_bad\_grades} + \varepsilon$$

The results are following:

Variable	**Coefficient**	**Std Error**	**P-value**
Constant	3.455	1.498	0.022
bad_share	0.108	0.063	0.086

R-squared	0.018
Adj. R-squared	0.012

Table 2: Results of the Linear Regression model

As we can see, the results are not so exciting, the R^2 coefficient is only 0.018, meaning the model accounts for only 1.8% of the dependent variable's variance, and the p-value of the only explanatory variable is 0.086, indicating the relationship is not statistically significant.

Summary

This analysis looked at whether bad restaurant grades in NYC are linked to more health complaints through the 311 system. Most restaurants (77%) get an A grade, with small differences between boroughs – Staten Island does best and Queens worst.

The data shows almost no connection between bad restaurant grades and complaint numbers. The correlation is very weak, and a regression model could only explain 1.8% of what drives complaints. The relationship isn't statistically significant.

Possible additional work

I have done my best to get the most out of the data. However I know that there are many more ways to explore it and gain valuable insights, I am simply not capable of doing that. What follows are my ideas on possible future analysis of these datasets:

- Accessing the data in a different way: Using number of complaints per capita/per restaurant etc
- Using more features and explanatory variables in the LR model and possibly using more advanced models
- Taking full advantage of the ZIP code data and using it for spatial analysis/clustering analysis, possibly discovering problematic areas