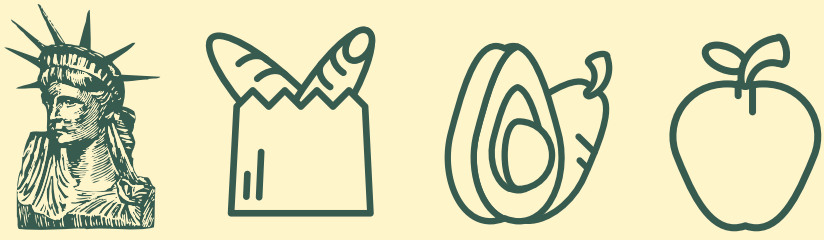
JACQUELINE
LESSOFF

Spring 2022

NEW YORK CITY RESTAURANT INSPECTION

Digital Economics



DATA SETS

zip_size.csv

zipcode_pop.csv

nyc-zip-codes.csv

Restaurant API

Rodent API

Note: to run this code, it is necessary to obtain a key on NYC open data in order to access the API for the rodent and restaurants dataset, and to obtain the most recent data. Click here for instructions on creating a key

The main data used in this project comes from the DOHMH New York City Restaurant Inspection Result dataset from NYC Open Data. The data has approximately 350,000 rows, and is updated on a daily level. The python notebook is connected directly to the API, allowing for an automated feed. The main data set provides every violation citation from every full or partial inspection 3 years prior to the most recent inspection. For the purpose of this analysis, I have limited the data to 2021, 2020, and 2019. Each row corresponds to a single violation in an inspection, thus, each inspection can yield multiple rows. Restaurants that have gone out of business are not included in the data set, as it only includes restaurants with an 'active' status.

This project leverages additional data to provide more information regarding the neighborhood of each restaurant. I connected to the Rodent Inspection live data set (updated daily) from NYC open data to capture the number of rodent reports in each zip code in 2021, 2020, and 2019. Additionally, I obtained zip-code level population and land area from the US Census Bureau. Finally, I connected to a NYC city administrative file to map zip codes to neighborhoods.

OBJECTIVES

Descriptive Analysis

Causal Analysis

This project has two main objectives: descriptive and causal analysis.

- Because of the complex nature of the datasets, extensive EDA is necessary to understand trends and relationships, as well as how the data is distributed.
- After thoroughly exploring the data, I construct a Directed acyclic graph, and conduct chi squared independence tests to explore the existence of dependencies between variables

Ultimately, the purpose of this analysis is to provide restaurant inspectors and owners detailed information regarding the patterns and relationships in food code violations to create a strategy to prevent and identify them in the future.

DATA LIMITATIONS AND CONCERNS

The data contained in this data set comes from several large administrative data systems. For this reason, it contains many complexities and required a good amount of feature engineering. There are several missing values that are simply the result of data entry or transfer errors.

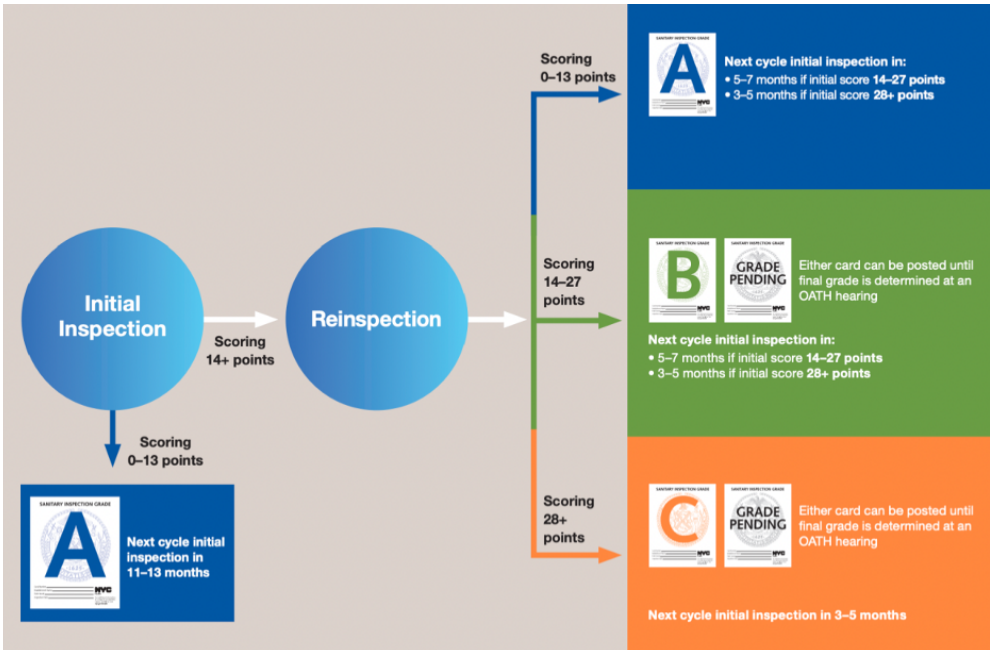
Additionally, there are records for restaurants that have applied for a permit, but have not been inspected. For the purpose of this analysis, these records are not included, and are indicated in the data by a start date of '1/1/1900'

Finally, it is possible for restaurants to appeal their score, a process that can take several months. For this reason, scores that are current today may be changed in the following weeks or months. Thankfully, as we have a live connection to the data source, the code can be rerun to reflect these changes as needed.

DATA PREP AND FEATURE ENGINEERING

Extensive data preparation and feature engineering was conducted to address concerns in the data, as well as to prepare the data for analysis. Firstly, it was necessary to address the missing values in the restaurant dataset. In terms of missing values, the grade field revealed the highest amount of missing values. This is due to the fact that not every inspection is 'gradable', depending on the inspection type and action.

Due to the missing values in this field, I created a proxy value for 'grade' that is based on the total score of the examination. The total score is based on the sum of penalties from the aggregation of food violations in a particular exam. For this analysis, it is helpful to create a category for these scores, as the traditional 'A,B,C' grade is what is typically used to judge the success or failure of an exam. I followed the guidance from the NYS Department of Health to create an 'A,B,C' grade from each examination's score.



Aside from the 'score' variable, it is documented that most missing values are missing completely at random and caused by the improper joining of several administrative data sets. As the percentage of remaining missing data in relevant fields were less than 1%, I decided to drop these values.

In addition to the handling of missing values, I decided to reduce the number of categories for the **cuisine_description** and **violation_code** variables in order to more effectively perform EDA and construct bayesian networks.

To create categories for cuisine description, there were several methods that I considered while creating categories, including price range, food preparation methods, and ethnicity. After reviewing the 86 values for cuisine_description, I constructed categories based on my judgement. As this field will impact results, it would be advised to have an experienced food inspector construct these categories, or use NLP techniques such as word2vec to create these categories. However, given the constraints, my judgement will suffice.

For the violation_code description, I examined the violation hierarchy from the NYC government website, and extracted the first two numbers to represent the violation category. This reduced the number of violation categories from 41 to 11.

Additional Datasets

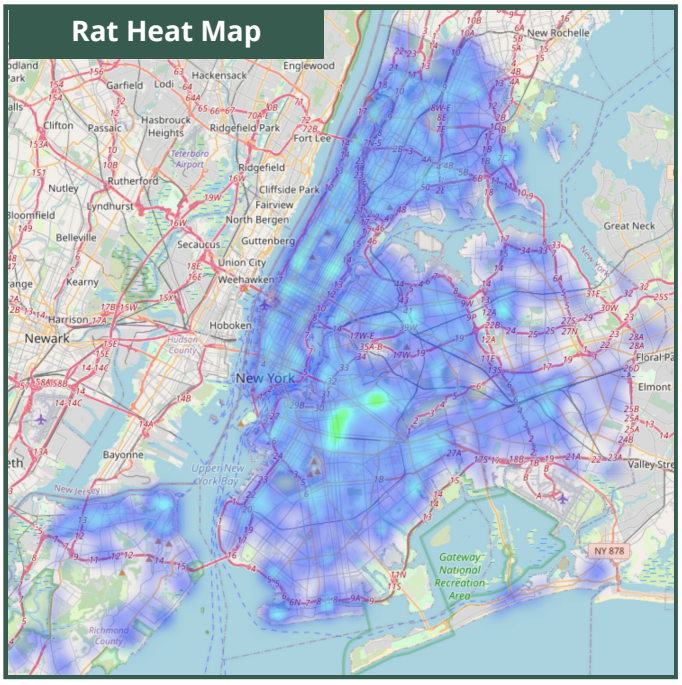
In addition to the variables provided in the main dataset, I also wanted to include fields to represent the population density of each neighborhood, as well as the prevalence of rodents in each neighborhood.

With the rodent data set, I obtained the number of rodent complaints from 2019,2020, and 2021 and grouped the count of complaints by zipcode. I then grouped the zipcodes by neighborhood divided this value by the total area to get the number of complains per km squared. I did the same with 2018 NYC population data (the most recent obtainable source), first grouping by neighborhood, and then dividing by area squared in each neighborhood.

EXPLORATORY AND DESCRIPTIVE ANALYSIS

In order to provide a representation of the rat and restaurant density of NYC, I created two map visualizations. This representation will allow the user to obtain a better understanding of the relative geography of the city, which will be important when understanding and interpreting results. The first map shows NYC restaurant dataframe, with each dot representing a business. Each dot color represent the proxy grade variable that I created; grey corresponds to A, while yellow and red correspond to B and C respectively. While the images attached in this report are small, an interactive html file is attached that allows zooming to specific neighborhoods and streets.

Rat Heat Map

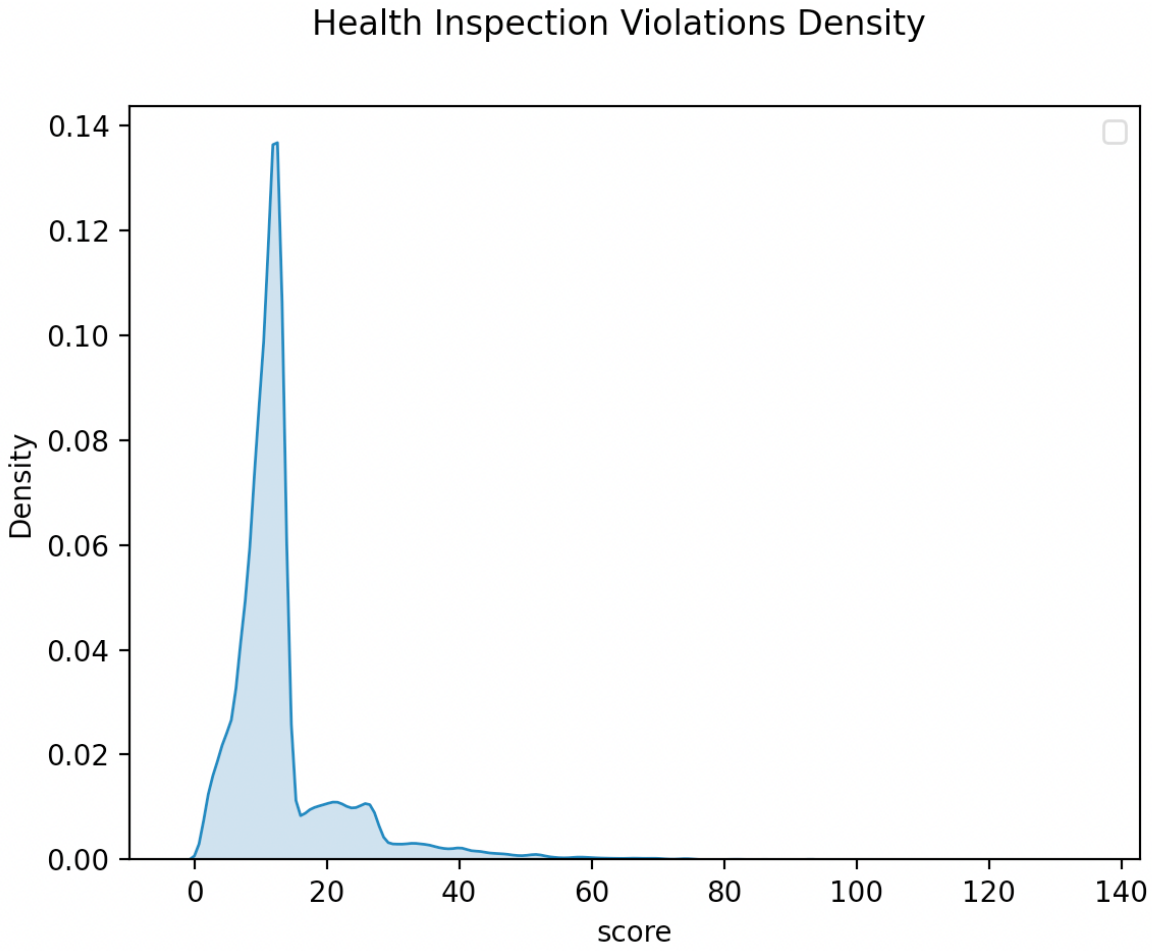


Restaurant Map

While the maps provided do not provide detailed statistical information regarding the data, they enable the user to better interpret the data, and pose relevant questions, potentially resulting in an improved approach to identify and prevent violations. For example, by viewing the 'rat density' map, one may ask why there is a hot spot of activity near the neighborhoods of Bushwick/Bedstuy, which could be related to the rapid gentrification and resulting construction in the area. Another peculiarity can be viewed in the lack of rat activity in midtown, a popular tourist hotspot. After researching this phenomenon, I found that midtown lacks the residential space that rats prefer. This, combined with low rat mobility, makes midtown surprisingly rat free.

While these observations are anecdotal, they lead to important questions and hypotheses relating to the data. For instance, by viewing this map, I may hypothesize that rat density may be correlated with violations relating not only to rodents, but also construction, as they may both be linked to gentrification. Moreover, I can hypothesize that rodent violations and density may also be correlated to population and garbage disposal management. While this may be a logical conclusion to some, it provides a way to identify possible explanations to certain occurrences.

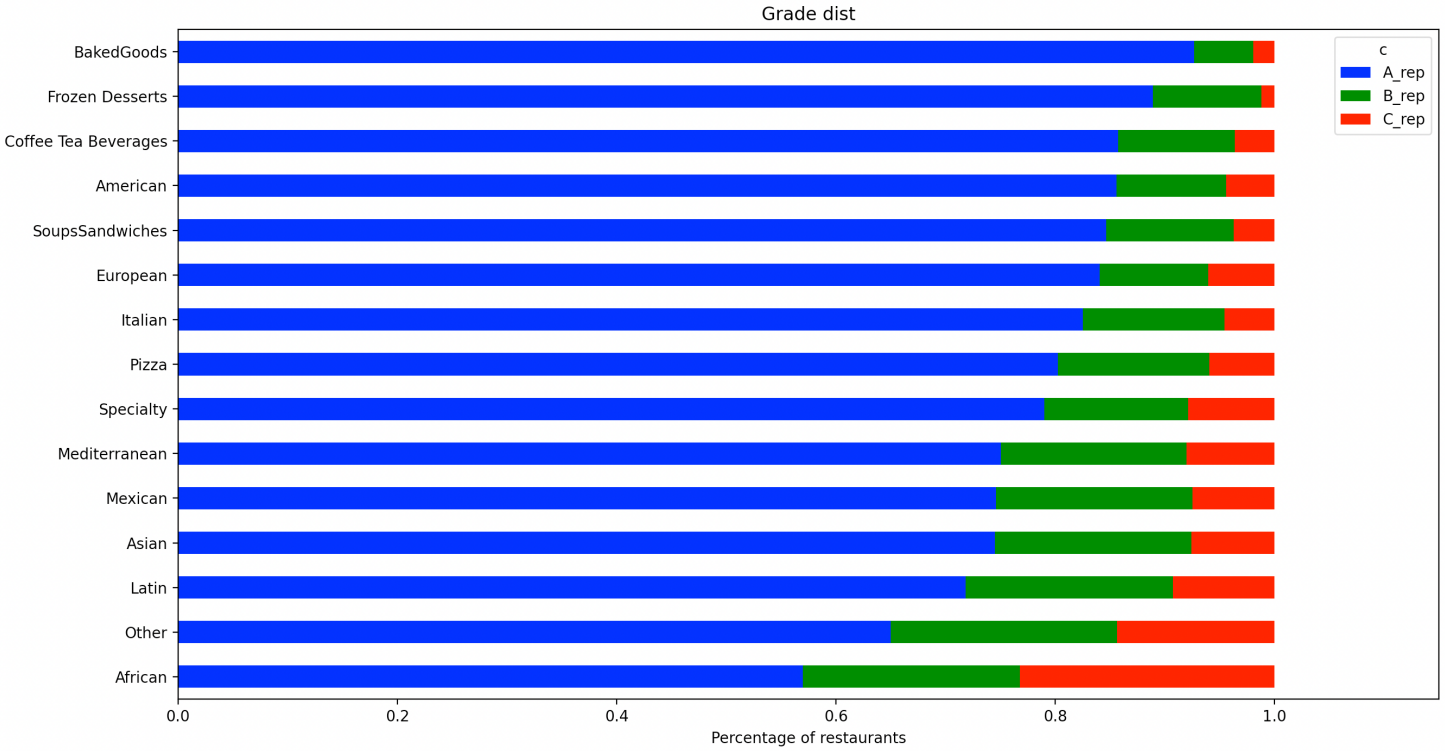
Score Distribution



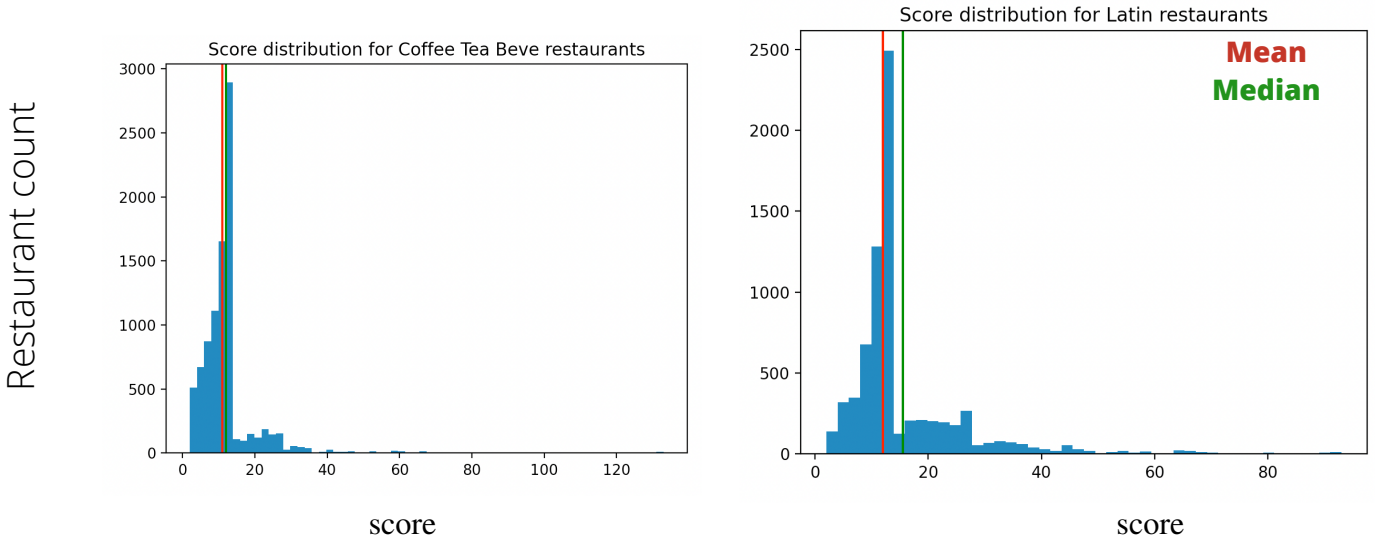
This map shows the score distribution for all NYC restaurants in the dataframe. The mean score of inspections is a 20, or a B. However, it is clear that there is an abnormal distribution of the scores. There is an apex at the score of 12, which represents the mode of the data.

The shape of this distribution is interesting when we consider the NYC grading criteria. As referenced earlier, the boundary between an A and a B is 13, and the boundary between B and C is 27. There are peaks located just before these points in the data, which indicates that inspectors may be reluctant to give a harsher grade, and may account for that while grading. This graph poses questions on what violations would result in the jump from a 13 to 14, or from a 27 to a 28. Additionally, this distribution also has a long tail, indicating that once a restaurant reaches a C, they will accumulate points quickly.

Cuisine Analysis



After obtaining the overall distribution for the scores among NYC restaurants, I took a deeper look at how grades and scores are attributed based on cuisine description. In this chart, I sorted cuisine types by the count of restaurants with A ratings. I then looped through the top 15 cuisine types to obtain separate distributions for score, attaching two examples below, with the rest appended to this document. The distribution of grades gives an indication on what cuisine types are more likely to receive a lower grade. Many of these are intuitive, as the worse grades may correspond with cooking techniques that are more complicated or prone to violations



The two cuisine types I highlighted in this report are coffee/tea restaurants and Latin American restaurants. As expected, both cuisine types peak at the border between an A and a B, although the Latin American restaurant distribution has a longer tail, and a few outliers between 60 and 70 points. Intuitively, it makes sense that several restaurant serving meals would score higher than a restaurant serving coffee and tea. However, as seen on the overall grade distribution chart, it appears that foods that are considered 'ethnic' (Asian, Latin, African food) are consistently scored lower than other food types such as European or American food. With the proper resources, it would be advised to investigate what factors play into these differences, and whether the inspectors are potentially biased, and are more critical of certain cuisines. The difference between restaurants may also be caused by other factors, such as if the owner is able to understand English, which could impact compliance with the regulations. While this analysis is not concerned with predicting inspection outcomes, this may be a point that should be investigated before fitting a model for prediction.

METHODOLOGY

After exploring the data and obtaining a better understanding of the different variables, I wanted to find how the variables were associated with one another, and explore potential causal dependency between variables. To do this, I implemented a Bayesian graphical model to build a Directed Acyclic Graph (DAG).

Before implementing this model, I needed to do further preprocessing to prepare my data for the graph. I decided to group the data by restaurant, and pivoted the data to create a way to analyze the dependencies between the violations themselves. The new dataframe I obtained contained the following variables:

2	**Indicator Variable for presence of at least one Time and Temperature Control for Safety (TCS) Food Temperature violations in most recent exam**
3	**Indicator Variable for presence of at least one Food Source violations in most recent exam**
4	**Indicator Variable for presence of at least one Food Protection violations in most recent exam (including evidence of pests)**
5	**Indicator Variable for presence of at least one Facility Design and Construction violations in most recent exam**
6	**Indicator Variable for presence of at least one Food Worker Hygiene and Other Food Protection violations in most recent exam**
7	**Indicator Variable for Obstruction of Department personnel violations in most recent exam**
8	**Indicator Variable for presence of at least one Garbage, Waste Disposal and Pest Management violations in most recent exam**
9	**Indicator Variable for presence of at least one Food Protection violations in most recent exam**
10	**Indicator Variable for at least one Facility Maintenance violations in most recent exam**
19	**Indicator Variable for at least one Organic Container violations in most recent exam**
22	**Indicator Variable for at least one Poster/Signage violations in most recent exam**
exam_count	**Count of exams in 2019,2020,2021**
vcount	**Average number of violations per exam from 2019,2020,2021**
zipcode	**ZCTA code of neighborhood**
rat_dens	**Rats per square km in Restaurant's neighborhood divided into 2 categories by quantile**
pop_dens	**Residents per square km in Restaurant's neighborhood divided into 4 categories by quantile**
neighborhood	**Neighborhood of restaurant**

METHODOLOGY

Conceptually, the purpose of building a DAG is to understand the causality's direction and understand which variable influences which variable. This is achieved by holding one variable constant and observing the effect.

It is possible to construct a DAG by creating all possible graphs and scoring the graph on its fit. This method is computationally difficult, thus, I decided to use a constraint based structure learning technique.

I used a chi-square test to identify independencies in the data set. This process uses a chi squared test statistic, relying on conditional hypotheses to understand dependencies between variables in the model. I set the p-value at 0.05, a standard level to test significance. This is the probability of observing a certain chi squared statistic under **the null hypothesis that variables are independent given other variables**. To calculate edge strength between nodes, I implemented an independence test with a chi squared test.

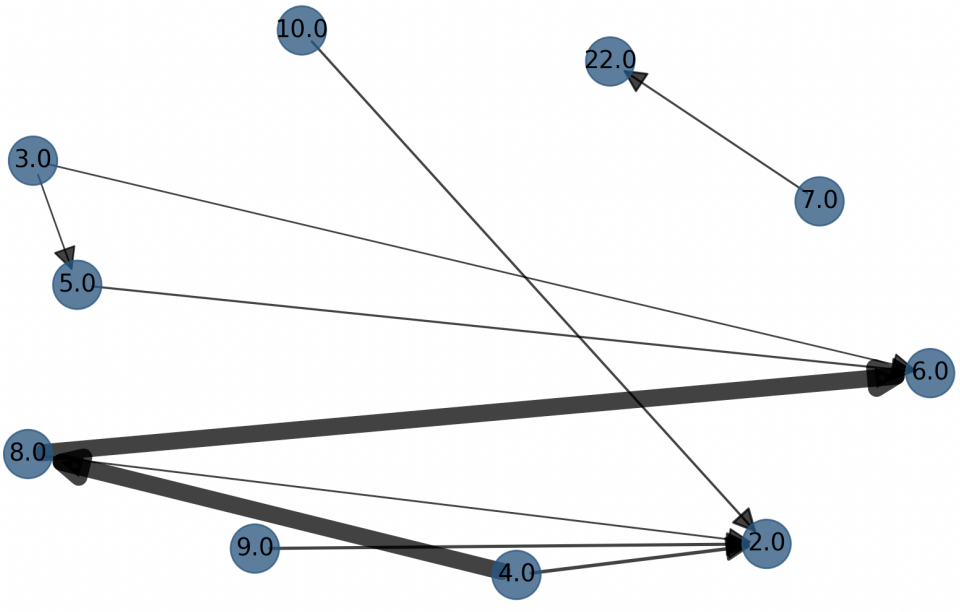
chi squared formula

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The formula for chi-square is relatively simple, with c being degrees of freedom, 'O' being the observed value, and 'E' being the expected value. As a result, larger differences between expected and actual data produce a larger value, meaning that a larger chi-squared value means there is a larger probability of a difference between the sample data and the null hypothesis of independence.

I first created a DAG using only the violation variables, to get a better understanding of how the presence of certain violations affect others without external factors. Afterwards, I added in the variables of neighborhood, rat_dens, exam_count, vcount, zipcode, and pop_density.

Violation Analysis



The detected DAG consists of 10 nodes out of the 11 variables, with the variable 19 (violations related to organic container misuse) not included because of lack of relavent dependencies.

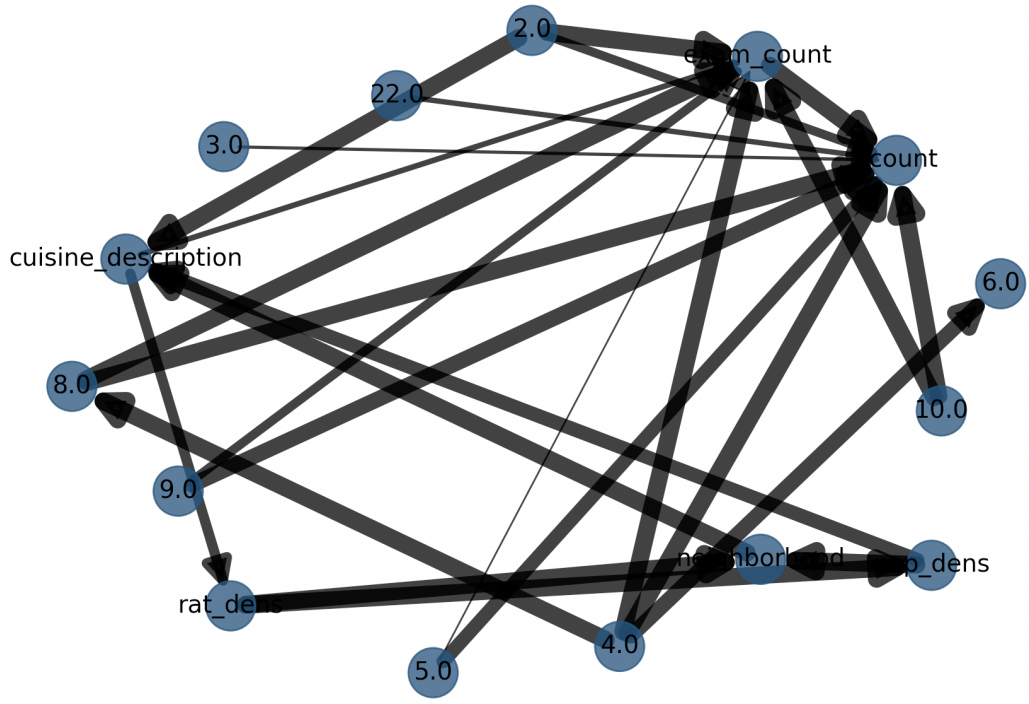
Notably, the graph shows that violation 4 (**food protection violation including presence of pests**) has a strong causal dependency on 8 (**garbage, waste disposal and pest management violations**), which has a strong causal dependency on 6 (**food worker hygiene and other food protection violations**). Intuitively, this relationship makes sense, as unprotected food and evidence of pests (included in violation 4) would be expected to be related to violations regarding the use of pesticides, exterminators, and garbage disposal (violation 8). Violation 6 is concerned with the decency of food contact surfaces, equipment and the hygiene of the worker.

Initially, the direction of the dependencies may not be intuitive, as I expected violations in pest management and garbage disposal (8) to cause unprotected food and evidence of pests (4). However, this can be explained when considering the nature of an inspection. If an inspector finds evidence of pests, there is almost certainly an issue with the pest management protocol in a restaurant. For this reason, the presence of violation 4 may serve as an indicator to an inspector that there is the presence of violation 8

source	target	p_value	chi_square
8	6	8.51574e-135	610.545
4	8	0	11311.4
9	2	3.25103e-16	66.6449
5	6	1.14862e-07	28.1058
3	6	0.000663944	11.5876
3	5	0.000680988	11.5404
10	2	1.00921e-09	37.307
4	2	5.25296e-19	79.331
7	22	3.3254e-07	26.0509

To provide further information, I present the results of the independence test on the violation data. Notably, the significance level between violations 4 and 8 is zero. While this result seems extreme, it likely captures the phenomenon that it is extremely unlikely, and potentially forbidden to have a violation relating to pests without having a violation relating to pest control, resulting in the two variables having dependency in this sample.

Additional Variables



The variables exam_count and vcount capture the restaurants history, but also act as a proxy for overall performance. In the case of exam_count lower scored restaurants are required to have more frequent examinations, and in the case of vcount, the average number of violations per examination indicates overall performance/score. For this reason, these variables have many relationships and arrows pointing towards them.

After adding in the additional variables, several observations can be made on the relationships between variables. Notably (and as expected), rat_density, neighborhood, and population density are highly related, and including all three variables is likely redundant. It is also revealing that rat density does not show a direct causal relationship to violations regarding pests. Cuisine description is shown to be highly related to these variables, likely because certain ethnicities are highly concentrated in certain neighborhoods in New York City. Additionally, Cuisine Description has a causal relationship to violation 2.0, **Time and Temperature Control for Safety (TCS)**. And, in return, violation 2 has a strong causal violation to both exam_count and vcount, indicating that it is an important violation. As this violation is one of the most likely to cause illness, it is considered a critical violation with a high penalty, and may explain why certain cuisines have lower scores.

It is also notable that once adding in the variables vcount and exam_count, the direction of the dependency between 4, 6, and 8 has changed. Before, 4 had a causal relationship to 6, who had a causal relationship to 8. Now, violation 4 shows evidence of causality towards both 6 and 8, and 6 no longer has a causal relationship to 8. This could be due to the fact that without including exam_count and vcount, the presence of 6 may have been representing some of the properties of restaurant performance and history. With the presence of these variables, 6 no longer has a causal effect on 8, and 4 has a causal effect towards 6. For reasons such as this, it is important to run these analyses with many different subsets of variables.

source	target	p_value	chi_square
cuisine_description	exam_count	1.06E-62	595.29
cuisine_description	rat_dens	8.61E-123	678.101
exam_count	vcount	0	2075.16
5	vcount	6.89E-169	833.677
5	exam_count	2.92E-11	55.2197
pop_dens	cuisine_description	9.25E-172	1099.94
pop_dens	neighborhood	0	68589
neighborhood	cuisine_description	0	9436.66
9	exam_count	3.91E-107	501.082
9	vcount	1.02E-193	949.57
2	exam_count	0	1686.56
2	vcount	2.70E-104	530.814
2	cuisine_description	2.70E-217	1128.03
10	exam_count	0	1553.57
10	vcount	0	2258.14
4	exam_count	0	2058.75
4	vcount	0	1700.65
4	6	3.36E-189	860.743
4	8	0	11311.4
8	exam_count	0	2006.42
8	vcount	0	2748.93
rat_dens	neighborhood	0	22863
rat_dens	pop_dens	0	12082.2
22	vcount	5.29E-59	316.043
3	vcount	9.13E-31	179.264

Compared to the previous results, we can see that the p-value for the relationship between 4 and 6 is slightly smaller than the prior relationship between 8 and 6, and the chi_square is higher. Thus the causal relationship between the new pair is stronger.

Additionally, there are several p-values of zero in this analysis that confirm the dependencies between neighborhood, pop_dens, and rat_dens as expected.

It is also unsurprising that exam_count and average violation counts are highly dependent on each other, as an increased frequency of examinations can result in more violations per exam, as the restaurant has less time and resources to prepare.

When considering which violations are the most significant in terms of the effects on inspection results, one can begin by identifying the variables that show a direct strong causal relationship to exam_count and vcount can help identify the violations that have a large impact on overall performance. For the sake of this analysis, I will focus on vcount, as it more directly captures violation history. According to this analysis, the variables that have a p value of 0 with vcount are: 8, 10, 4, 2, sorted in a descending order by chi squared magnitude. However, a large part of the relationship between 8 and vcount is influenced by the extremely strong causal relationship from violation 4. In future analyses, it would be interesting to exclude violation 8 to see how this affects other relationships.

Another interesting observation concerns the relationship between violation 2 and exam_count and vcount. Violation 2 has a strong causal effect on exam_count, but a much weaker effect on vcount. This possibly indicates that the presence of a violation 2 (Time and Temperature Control for Safety (TCS)) can result in a lower grade, causing more frequent examinations. However, the fact that there is a weaker causal effect between violation 2 and average violations per exam may indicate that the presence of this violation causes a lower score regardless of the number of violations.

Key Takeaways and Recommendations

The analysis presented clarified and confirmed a few theories about restaurant inspection results, while also presenting a few unexpected results

First of all, it is unsurprising that neighborhood is closely related to population density and rat density. However, I was surprised to find that there were no causal relationships identified between variables related to location and violations related to rodents. Along with the extremely strong relationship between pest prevention methods and pest violations, one can conclude that location cannot be blamed for pest infestations, and resources focused on adequate pest extermination and prevention will yield positive results.

Secondly, it is also not surprising that violations related to time and temperature control have a strong causal relationship to the performance of a restaurant. Violations in this category are even identified as 'CRITICAL', or a 'Public Health Hazard'. The strong causal relationship between violation 2 and cuisine type is also relatively unsurprising (an example being the score distribution of Coffee shops and Latin restaurants). Additionally, as mentioned earlier in the report, restaurants that are more 'ethnic' are shown to receive lower grades. Therefore, I would advice 'ethnic' restaurants to focus their attention on this violation, as it is connected to cuisine_description, and as the presence of this violation is more likely to cause a grade drop regardless of the presence of other violations. Additionally, as cuisine types are strongly related to both neighborhood and violation 2, it is possible that there are geographic clusters of ethnic restaurants that all receive lower scores due to this violation. While this is entirely unconfirmed, it might be interesting to explore this area, as a community-targeted outreach program relating to temperature violations could reduce the number of violations, improving scores and reducing the need for additional examinations.

Future Work

While this analysis involved substantial data processing, exploration, and analysis, there are several areas that could be improved upon with more time and resources. In terms of the causal analysis, I would have liked to spend more time on feature selection and categorization of continuous variables. I also would have liked to create more DAGs with different subsets of variables to better understand their relationships.

I would have liked to analyze each violation separately rather than relying on the simplified code from the hierarchal structure to get a better understanding of the effects of each violation.

Additionally, while most of the cuisine_descriptions were based on ethnicity/cultural factors, and thus could be used as a proxy for ethnicity, there were some categories (such as coffee/tea restaurants) that were unrelated. For this reason, in the future I would recommend using an indicator variable for whether a category is related to an ethnicity.

Ultimately, as I was starting from scratch and exploring an extremely large data set, I wasn't able to focus on a specific area. The results from both the causal and descriptive analysis lead to

thousands of other questions and areas to explore. If I were to extend this project, I would want to explore the effects of gentrification in violations of long running restaurants, as gentrification is one of the most prevalent problems in New York City. By analyzing the effects on restaurants, it would be possible to inform policy to help these businesses rather than spending resources on frequent inspections and causing closures.