Investigating NYC Health Inspections and Income

due April 20, 2022 at 11:59 PM

Alex Pieroni, Ali Gaviser, Marina Chen, Nick Reddy

20 April 2022

Introduction and Data

Topic and Motivation

We are interested in this data because food safety is important for a dense area such as NYC. The data set is huge, allowing freedom of exploration. We chose health scores and grades because they are widely applicable to each restaurant and borough (since every restaurant received a grade/score). Although New York City is considered a homogeneous metropolitan area, there is depth and variety among, and even within, boroughs. We wanted to highlight these nuances. Furthermore, it's also important to reveal discrepancies caused by income to expose inequalities; we wanted to show that socioeconomics can have a correlation with many facets of life, especially food and health.

The study cited below, by Meltzer, et al. (2015), supports the importance of exploring NYC restaurant data stated above. They found that while health inspections in restaurants do increase overall grades for all restaurants over time, benefiting the health of consumers, the result of a health grade can negatively impact the establishment's immediate sales. For less affluent restaurant owners or cultural cuisines that may require more preparation, the restaurant market may suffer in terms of diversity. Thus, the correlation between income, health grade/score, and borough is important to further investigate how wealth is intertwined with the location of restaurants and their scores.

Dataset: NYC Restauraunt Inspections

The data is from the New York Open Data Portal, sourced from TidyTuesday. Only restaurants in an active role are included. The dataset contains every sustained violation citation from every full or special program inspection as collected by the City of NYC Department of Health (www1.nyc.gov/). There are 345,036 obs. of 26 variables; the vast majority were collected from 2015-2022, though two observations are from the year 1900. The data encompasses a wide variety of boroughs, scores, cuisine types, and zipcodes. There are both a letter and number score of the restaurant's hygiene. We will add an income variable from the 2019 US Census Bureau.

Restaurants are rated on a scale starting at 0. A score between and including 0-23 receives an A health grade, 24-27 scores correlate to a B health grade, and scores of 28 and above indicate a C health grade. A lower score indicates a better health grade. Therefore, A is best, then B, then lastly, C.

Sources:

https://github.com/rfordatascience/tidytuesday/tree/master/data/2018/2018-12-11

https://www1.nyc.gov/site/doh/business/food-operators/letter-grading-for-restaurants.page#:~:text=Since%202010%2C%20New%20York%20City,seen%20by%20people%20passing%20by

https://www.census.gov/programs-surveys/acs/news/updates/2019.html

http://www.appam.org/assets/1/7/Impact_paper_9-30-15.pdf

Research Question:

What boroughs have the highest and lowest health ratings, and how does that correlate to median income?

Hypotheses:

We hypothesize that NYC restaurants in Manhattan, a richer borough, will have better health grades. We believe that boroughs with higher median income will have higher health grades. Conversely, we expect the Bronx, a borough with a lower median income, will have the worst health grades. Overall we expect better health grades to be correlated with higher median income and vice versa.

Methodology

We made three data sets from our main data. The first one drops the non-applicable values in the dataset, and selected the variables of borough, grade, and score.

The second data set selects the borough and grade, and organizes each grade with A = 1, B = 2, and C = 3. From there, we found the mean grade number on the same 1-2-3 scale by each borough and added the median income from the 2019 US Census Bureau. We dropped non-applicable values.

We only looked at grades A, B, and C, because we found that other investigations only used these grades, and there are far more data points correlated with these grades. These are also the grades that the NYC government uses on their website.

The third data set selects the borough and score, dropping non-applicable values. The scores are given based on sanitary inspection. A score from 0-13 is an A, from 14-27 is a B, and 28 or more points is a C. From there, we found the mean score by each borough and added the median income from the 2019 US Census Bureau.

FIGURE 1:
Health Grade Letter Ratings in Each NYC Borough

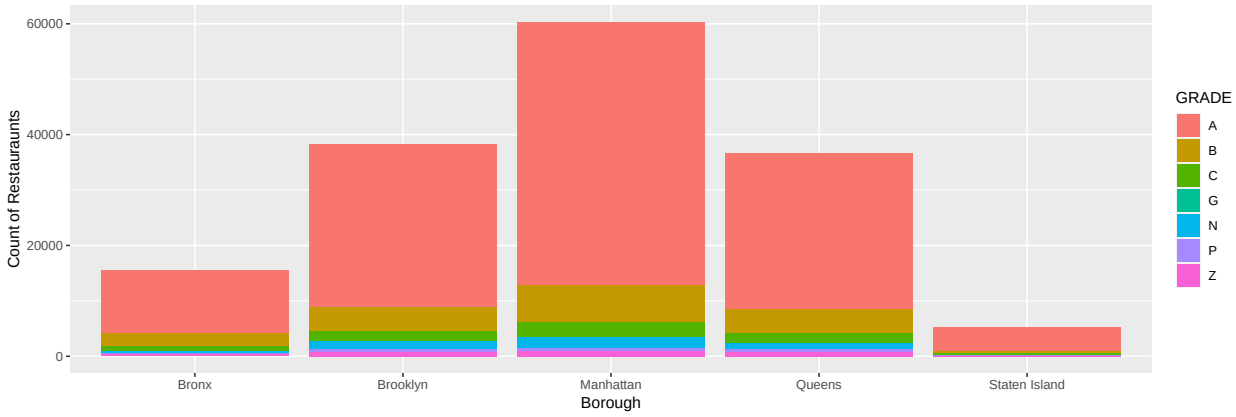


Figure 1 is a histogram of the letter Health Grades for each borough. We used a histogram to show the count of each grade while using different colors. This allows us to show the number of restaurants in each grade category in each borough. This visualization shows that Manhattan has the most amount grade A restaurants and also has the most amount of total restaurants in the data set with a little more than 60,000 observations. On the other hand, Staten Island has around 5,000 restaurants in this data set. The difference in the amount of observations for each borough will be a factor worth noting when analyzing our results.

FIGURE 2:
Percent Health Grade Letter Ratings in Each NYC Borough

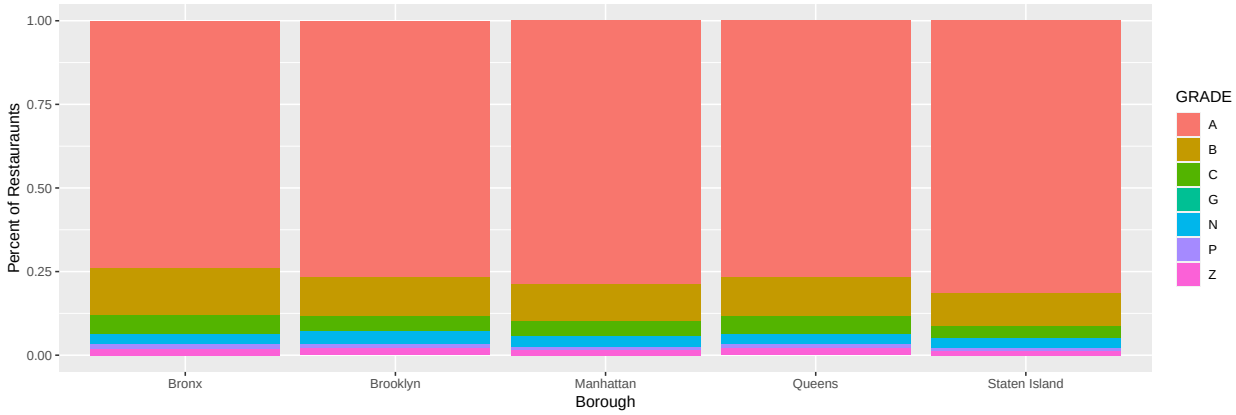


Figure 2 is similar to Figure 1, but now shows the frequency of letter grades by borough instead by percentages. This figure shows that Staten Island has the highest percentage of grade A restaurants, which was not seen in Figure 1. This allows us to compare each borough's health ratings relatively, which is also important in our an analysis because of the differences in number of observations for each borough.

FIGURE 3:
Health Grade faceted by Borough

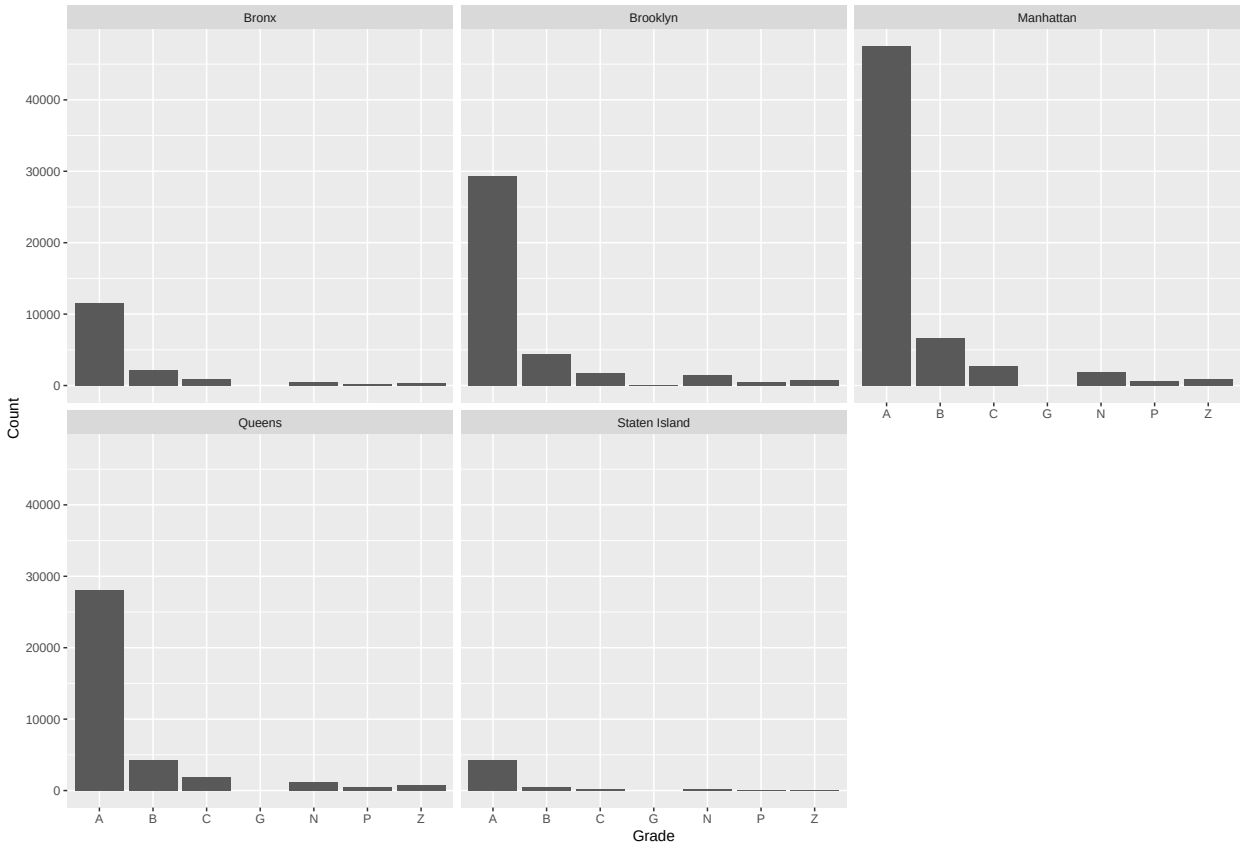
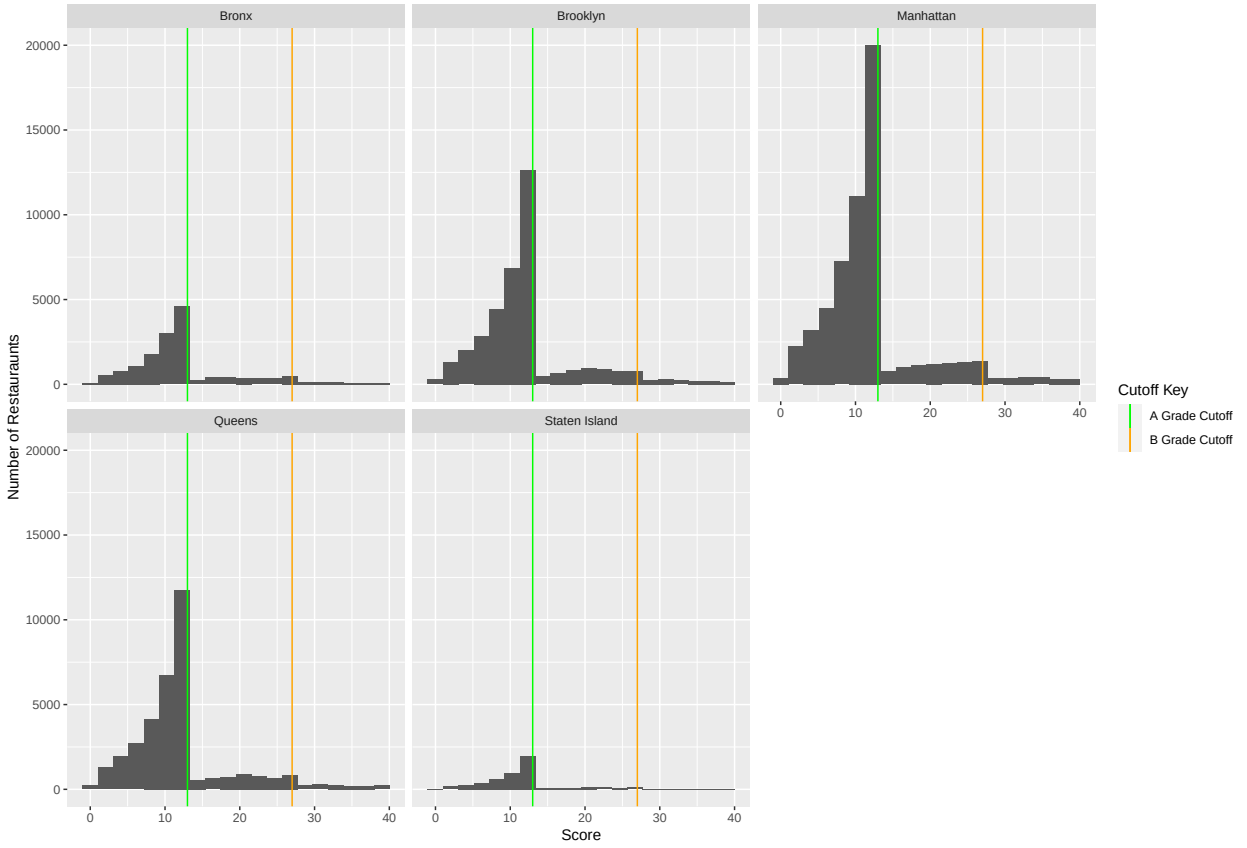


Figure 3 illustrates the frequency of each health letter grade, faceted by borough. This would allow us to see the spread of health ratings across each borough. It is one method to determine a correlation between restaurant health ratings and the restaurant's borough.

FIGURE 4:
Frequency of Scores in All Boroughs



Like Figure 3, Figure 4 is a histogram of health ratings faceted by borough. However, instead of using letter grades, we graphed scores. The faceting allows the easy comparison of each borough. We can see that although each borough has a different frequency of each score, the distribution is quite similar. We chose to do scores because it adds more information about restaurant health ratings as compared to grades. This is because each grade is determined within a range of scores. The green vertical lines are at a score of 13, the cutoff for an A letter grade. The orange lines indicate the cutoff for a B rating at 27. A restaurant with a score of 28 or above is given a C grade.Each borough has a peak (of varying size) extremely close to the A/B cutoff point. There are very few high quality A ratings. The graph shows that Manhattan has the most A ratings, however, it also shows that there are more scores recorded in Manhattan in general. However, it's worth noting how much higher the A scores are (particularly by the A/B barrier) considering the B scores are not present at that much higher of a frequency in comparison to the other boroughs. We decided to cutoff the range of scores at 40. We did this because although there were scores above 40, proportionally there were not enough scores to actually see them in the histogram. Cutting off the C scores at 40 gives each letter grade a range of 13, making the graph proportional.

FIGURE 5:
Correlation Between Income and Mean Health Grade
Colored by Each Borough

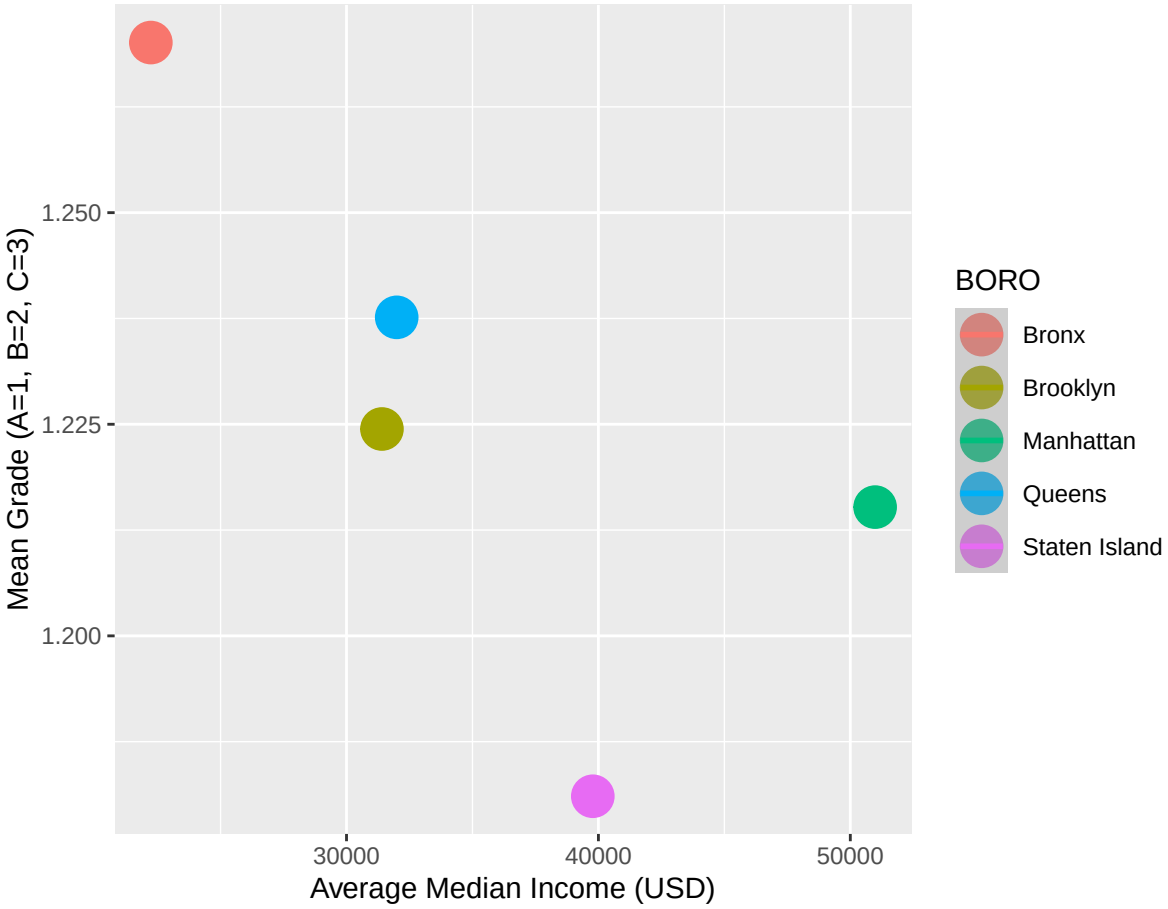
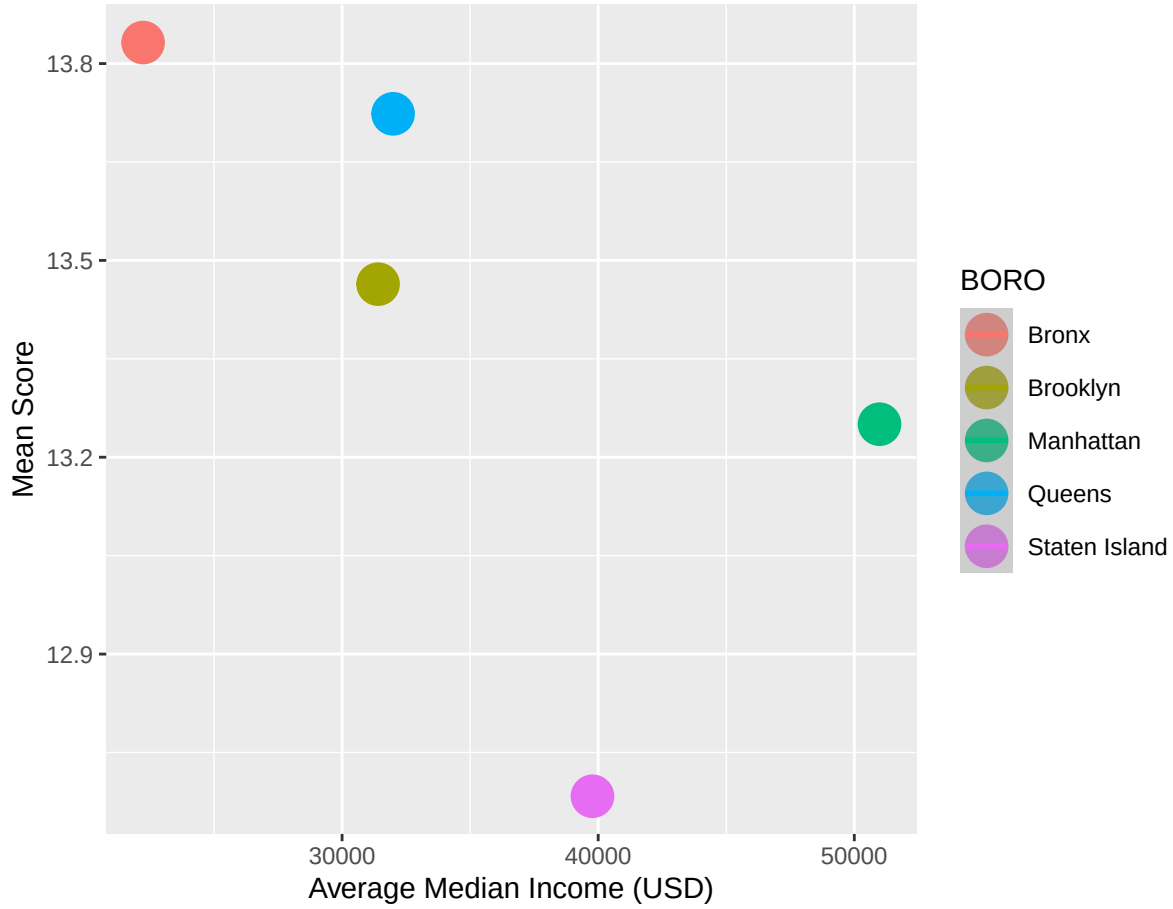


Figure 5 above was chosen to clearly show the relationship between mean health grades and median income for each borough. A scatter plot is the best choice because it clearly shows the relationship between two numerical variables, and we can use color to indicate the borough that each point represents. The Bronx has the highest mean health grade (closest to a B) and the lowest median income, while the Manhattan borough has the highest median income and the second lowest mean grade (second closest to an A). These correlations are what we expected, as indicated in our hypothesis. It's worth noting that Staten Island has the lowest mean grade by a notable difference, and the second highest median income. This could also correlate to what we expected in our hypothesis, but it may also be a product of the fact that Staten Island had the fewest recorded data/number of restaurants in the data set.

FIGURE 6:
Correlation Between Income and Mean Score
Colored by Each Borough



Similarly to Figure 5, Figure 6 correlates mean score with the average median income (USD) colored by the borough. This allows us to investigate the question using the mean score. The cutoff for an A score is 13, and the range of the mean scores is from 12.69 to 13.83.

```
## # A tibble: 1 x 1
##   mean_grade
##        <dbl>
## 1       1.23
```

The mean grade of restaurants across the five boroughs is between an A and a B. The mean grade is 1.22, with A being 1 and B being 2; therefore, the mean grade is actually closer to an A than a B.

```
## # A tibble: 5 x 3
##   BORO          mean_grade Income
##   <chr>              <dbl>  <dbl>
## 1 Bronx               1.27  22232
## 2 Queens              1.24  31992
## 3 Brooklyn            1.22  31406
## 4 Manhattan           1.22  50985
## 5 Staten Island       1.18  39777
```

While the mean for all restaurants across all boroughs is 1.22, the mean grade in Manhattan is 1.21. This is slightly lower than the total mean, leading us to believe that restaurants in Manhattan have a higher grade

on average. In other words, in comparison to all of New York City, the Manhattan restaurants have a mean grade (in this sample) that is slightly closer to an A grade.

As mentioned above, the mean for all restaurants across all boroughs is 1.22. The mean grade in the Bronx, however, is 1.27. This is higher than both the mean for New York City restaurants, and it is actually the highest mean grade out of each individual borough. Therefore, the Bronx has a mean grade that is closer to B than any other borough. The health ratings in this dataset are lowest for the Bronx out of all five boroughs.

```
## # A tibble: 5 x 3
##   BORO          mean_score Income
##   <chr>              <dbl>  <dbl>
## 1 Bronx               13.8  22232
## 2 Queens              13.7  31992
## 3 Brooklyn            13.5  31406
## 4 Manhattan           13.3  50985
## 5 Staten Island       12.7  39777
```

Similarly to the results from our mean score, our summary results tell us that the Bronx has the highest (and therefore least healthy) mean score, while Staten Island has the lowest mean score. This borough also has the fewer data points when compared to Manhattan.

Results

To answer what boroughs have the highest and lowest health ratings, we will find summary statistics for each boro.

Null Hypothesis 1

Null Hypothesis: The mean health score of Manhattan restaurants is equal to the mean of all other boroughs' restaurant scores.

Alternative Hypothesis: The mean health score of Manhattan restaurants is less than the mean of all other boroughs' restaurant scores.

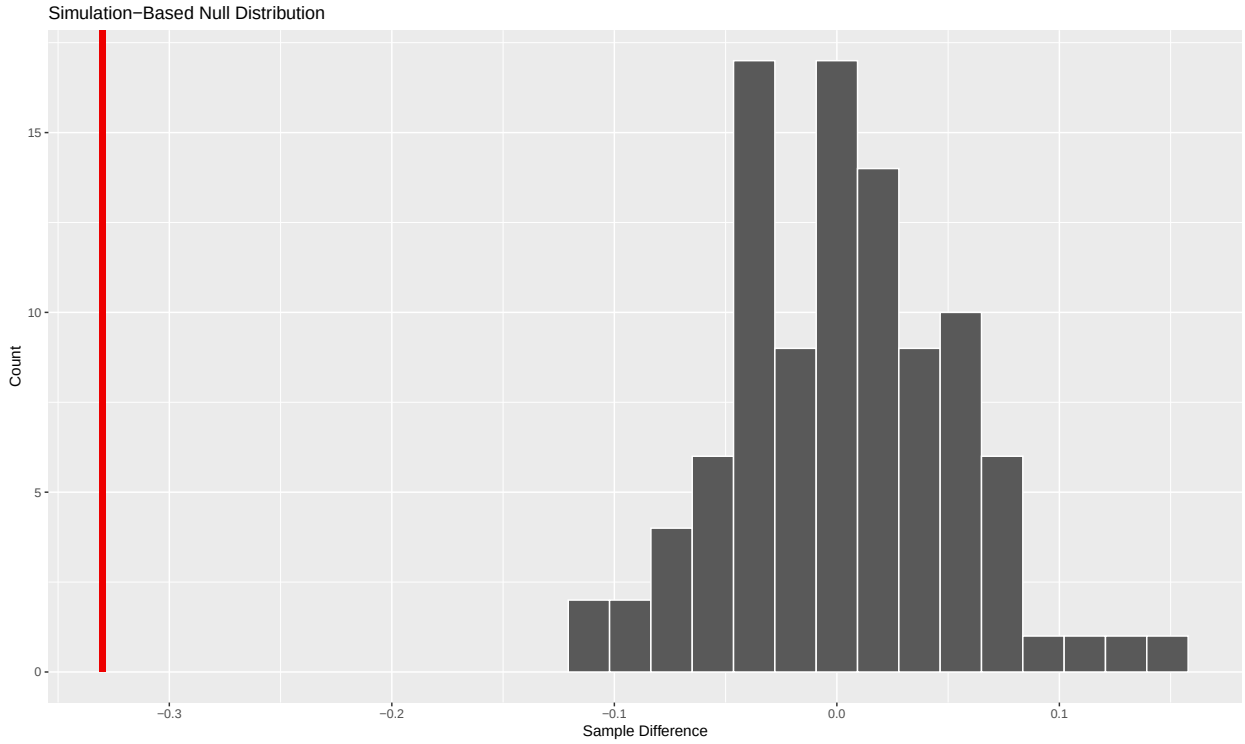
Note: the closer the score is to 0, the less violations and more up to code a restaurant is

M: Manhattan

O: Other Boroughs

$H_0 : \mu_M = \mu_O$

$H_1 : \mu_M < \mu_O$




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The p-value for this simulation is 0, so $p < \alpha$. Since the p-value is less than the significance level, we reject the null hypothesis that the mean health score of Manhattan restaurants is equal to the mean of all other boroughs' restaurant score. Instead, we favor of the alternative hypothesis that the mean health score of Manhattan restaurants is less than the mean of all other boroughs' restaurant scores. We used 100 reps because it was sufficient for our data set and our computers ran out of memory with anything over 100.

Now we will perform hypothesis testing for the Bronx.

Null Hypothesis 2

Null Hypothesis: The mean health score of Bronx restaurants is equal to the mean of all other boroughs' restaurant scores.

Alternative Hypothesis: The mean health score of Bronx restaurants is greater than the mean of all other boroughs' restaurant scores.

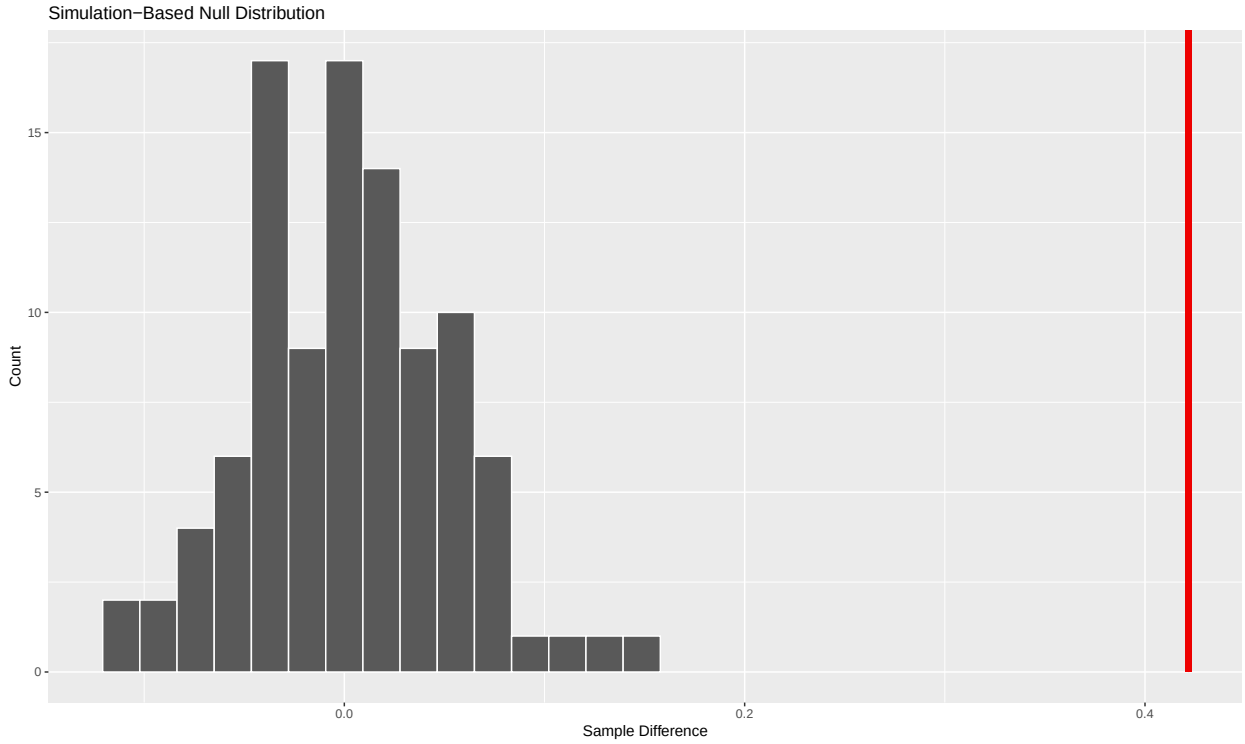
Note: the closer the score is to 0, the less violations and more up to code a restaurant is

B: Bronx

O: Other Boroughs

$H_0 : \mu_B = \mu_O$

$H_1 : \mu_B > \mu_O$




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The p-value for this simulation is also 0, so $p < \alpha$. Since the p-value is less than the significance level, we reject the null hypothesis that the mean health score of Bronx restaurants is equal to the mean of all other boroughs' restaurant score. Instead, we favor of the alternative hypothesis that the mean health score of Bronx restaurants is greater than the mean of all other boroughs' restaurant scores.

Linear Regression

```
## # A tibble: 2 x 5
##   term           estimate  std.error statistic   p.value
##   <chr>             <dbl>      <dbl>     <dbl>     <dbl>
## 1 (Intercept)  1.30       0.0455         28.6  0.0000939
## 2 Income      -0.00000212 0.00000124     -1.70 0.188
```

For our linear regression model, we are trying to predict the mean grade of a borough from the median income of the borough. Interpreting the intercept, for a borough with a median income of 0 dollars, on average the mean grade is 1.3, which in this context does not make sense because a borough will never have a median income of 0 dollars. Interpreting the slope, for an increase in median income by 1 dollar, on average the mean grade is expected to decrease by .00000212 dollars. However, for an alpha of .05, the p-value for this predictor, .183, is not significant, therefore there is not enough evidence to conclude a relationship exists between median income and mean grade and are not necessarily correlated.

We compared the mean health score for restaurants in Manhattan to the mean health score for restaurants in the other four boroughs. Our data in the methodology section shows that Manhattan has the highest income of the five boroughs.

Discussion

Conclusions:

The boroughs that have the highest and lowest health ratings were Manhattan and the Bronx, respectively. The mean grade for all of New York City's boroughs collectively was 1.22. Manhattan's was 1.21, which is the second closest to an A in comparison to all of New York City. Staten Island had a mean grade of 1.18, the lowest (and best) health grade of the five boroughs. Staten Island also has the second highest median income. On the other hand, the Bronx had a mean grade of 1.27, which is closer to a B than any other borough. The same trend holds for mean health scores. Our overall hypothesis was also supported: income was indeed correlated with which borough had the highest mean health grade, and vice versa. However, these correlations are just summary statistics of the data set. Our hypothesis test is more informative. The hypothesis test for Manhattan having a lower mean score than the other boroughs yielded a significant p-value of 0 at the $\alpha = 0.05$ significance level. The hypothesis test for the Bronx having a higher mean score than the other boroughs yielded a significant p-value of 0 with $\alpha = 0.05$. Therefore we were able to reject the null hypothesis in both cases.

Limitations:

We rejected the null hypothesis both of our hypothesis tests because the p-values were significant, but when we ran our linear regression, the p-value was not significant. This means we should be hesitant about the validity of our null hypothesis rejection, warranting further investigation into our research topic. Our hypothesis test also used 1,000 repetitions; 10,000 reps were too high for a data set of this size, the test wouldn't run. This may have impacted our results as well, but it was an unavoidable choice.

In terms of the data, it was collected between 2015 and 2022, with a couple entries actually collected in 1900. The health ratings could have drastically changed over this time period. This could have potentially increased or decreased the correlations we observed. Moreover, while data sets with a lot of information can be helpful, however our data set had so much information that it was almost too much to work with. A lot of the restaurants had NA values, some of the scores didn't make sense, etc. This forced us to filter out some data points in order to produce figures and data that we could actually analyze. Cutting the data made the most sense, however, it could have impacted our results. Lastly, although out of our control, we don't know how health ratings are given or biases that might go into it. Some health grades may not be deserved for one reason or another.

Further Research:

In the future we would like to compare health grades by smaller divisions than borough, perhaps going into New York City neighborhoods as well as comparing scores to the neighborhood's mean incomes. Additionally, we would like to add in another variable: cuisine. It would be interesting to see how health ratings change by cuisine in general, as well as across boroughs. Perhaps the racial and ethnic makeup of a borough may impact the ratings of a particular group's cuisine.

In terms of further developing our findings, in the future we would like to try to assess the true mean health scores with confidence interval within each borough and New York as a whole, and then compare that to the median income within each borough. This would be a further development since we just assessed the health grades of Manhattan and the Bronx relative to the other boroughs.