Project Draft

due March 27, 2022 by 11:59 PM

Alex Pieroni, Ali Gaviser, Marina Chen, Nick Reddy

11 April 2022

Load Packages

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
install.packages("tidytuesdayR")
```

Load Data

```
## Warning: One or more parsing issues, see `problems()` for details
```

The data comes from tidytuesday.

Introduction and Data

Dataset: NYC Restauraunt Inspections

The source of this data is from the New York Open Data Portal. Only restaurants in an active role are included. The dataset contains every sustained violation citation from every full or special program inspection. This data contains 345,036 obs. of 26 variables. The data set focuses on NYC Restaurant inspections in a wide variety of neighborhoods, scores, cuisine type, and zipcode. There is both a letter and number score of the restaurant's hygiene.

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

Research Questions:

What boroughs have the highest and lowest health ratings?

Hypotheses:

We hypothesize that NYC restaurants in richer neighborhoods will have higher health grades.

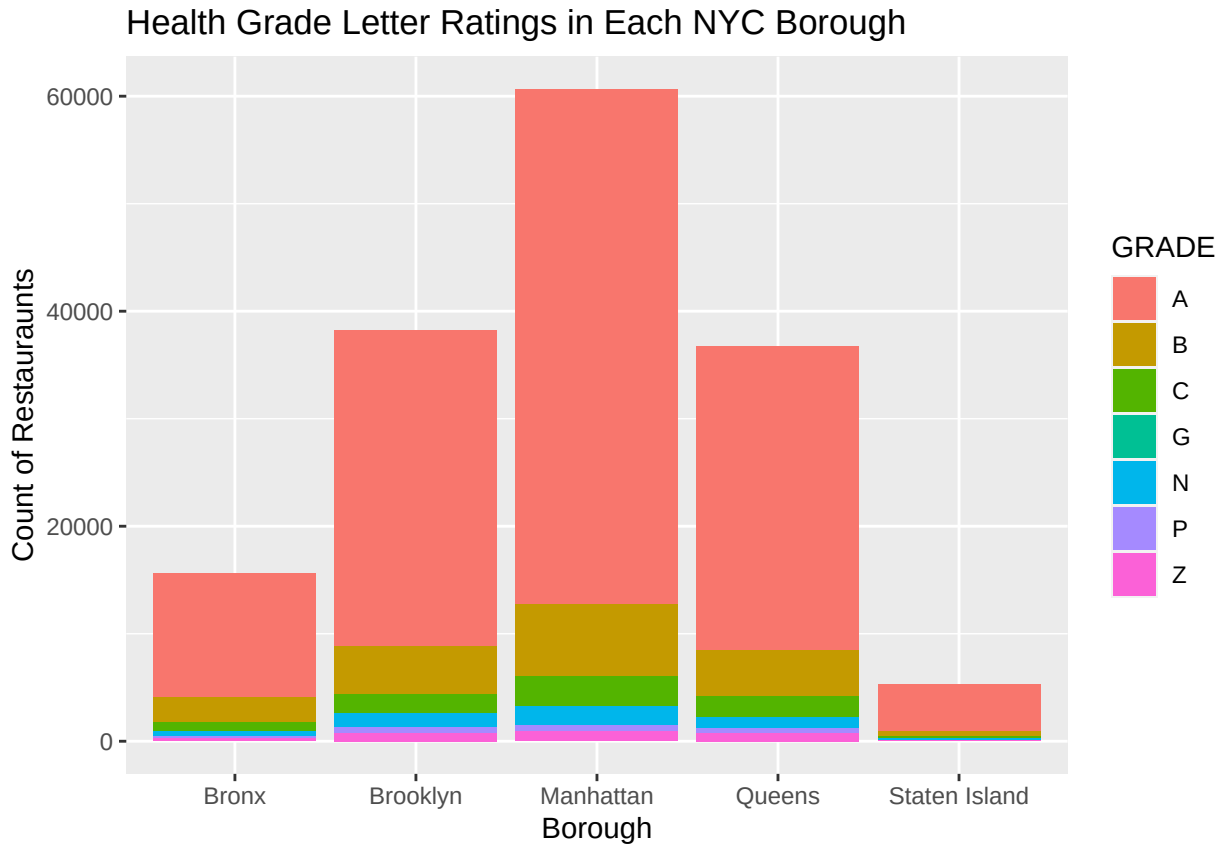
Glimpse

```
## Rows: 313,540
## Columns: 26
## $ CAMIS                  <dbl> 50017126, 41466918, 50007379, 50068576, 500156~
## $ DBA                    <chr> "BRAVO PIZZA", "NUEVO MEXICO MEXICAN RESTAURAN~
## $ BORO                   <chr> "Manhattan", "Brooklyn", "Brooklyn", "Queens",~
```

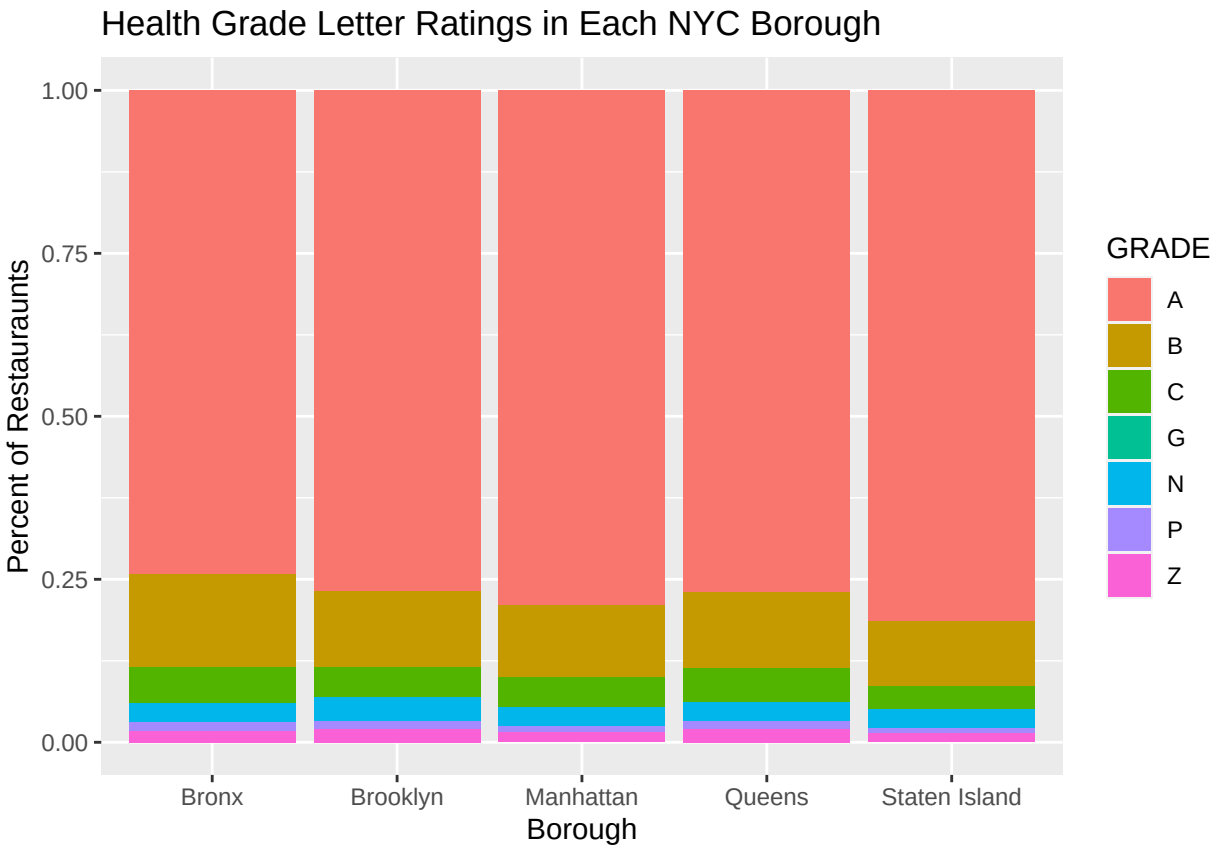
```
## $ BUILDING                <chr> "6", "489", "2781", "2707", "4617", "4617", "5~
## $ STREET                  <chr> "EAST   42 STREET", "5 AVENUE", "SHELL ROAD", ~
## $ ZIPCODE                 <dbl> 10017, 11215, 11223, 11101, 11203, 11203, 1123~
## $ PHONE                   <dbl> 2128674960, 7188320050, 7187690001, 9292083100~
## $ `CUISINE DESCRIPTION`   <chr> "Pizza", "Mexican", "Soups/Salads/Sandwiches",~
## $ `INSPECTION DATE`       <chr> "06/01/2017", "08/09/2019", "03/20/2019", "06/~
## $ ACTION                  <chr> "Violations were cited in the following area(s~
## $ `VIOLATION CODE`        <chr> "04L", "10F", "02G", "09C", "10F", "10F", "20D~
## $ `VIOLATION DESCRIPTION` <chr> "Evidence of mice or live mice present in faci~
## $ `CRITICAL FLAG`         <chr> "Critical", "Not Critical", "Critical", "Not C~
## $ SCORE                   <dbl> 19, 9, 9, 19, 13, 13, NA, 13, 8, 8, NA, 14, NA~
## $ GRADE                   <chr> NA, "A", "A", "B", "A", "A", NA, "A", "A", "A"~
## $ `GRADE DATE`            <chr> NA, "08/09/2019", "03/20/2019", "06/27/2019", ~
## $ `RECORD DATE`           <chr> "04/13/2022", "04/13/2022", "04/13/2022", "04/~
## $ `INSPECTION TYPE`       <chr> "Cycle Inspection / Initial Inspection", "Cycl~
## $ Latitude                <dbl> 40.75332, 40.66759, 40.58450, 40.74824, 40.641~
## $ Longitude               <dbl> -73.98055, -73.98760, -73.97451, -73.94135, -7~
## $ `Community Board`       <dbl> 105, 306, 313, 402, 317, 317, 303, 103, 301, 3~
## $ `Council District`      <chr> "04", "39", "47", "26", "45", "45", "36", "01"~
## $ `Census Tract`          <chr> "008200", "013900", "037402", "001900", "08400~
## $ BIN                     <dbl> 1035342, 3022922, 3423929, 4005129, 3112291, 3~
## $ BBL                     <dbl> 1012760069, 3010230004, 3072330210, 4004320025~
## $ NTA                     <chr> "MN20", "BK37", "BK26", "QN31", "BK91", "BK91"~
```

Project Draft

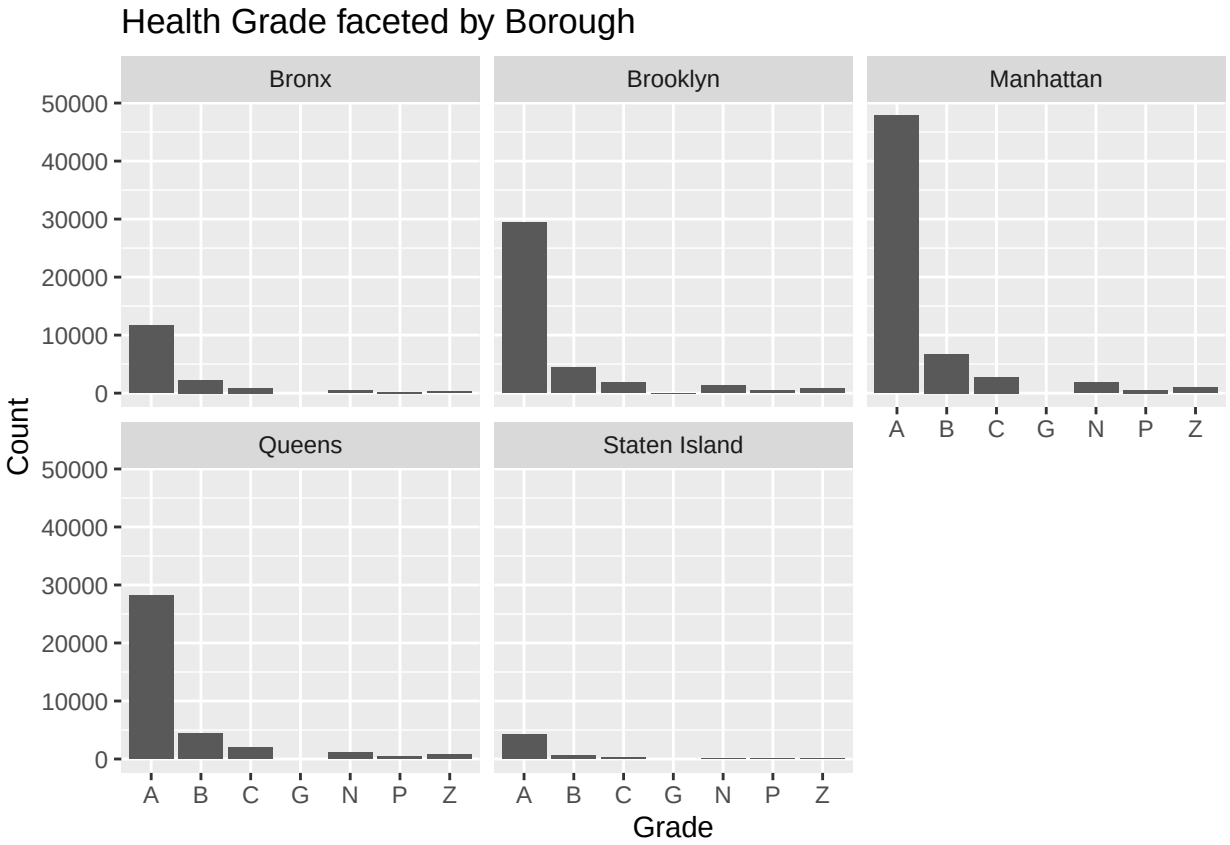
Methodology



This graph is a histogram of the letter Health Grades for each borough. Choosing a graph for this research question was somewhat difficult considering the large number of data points. We originally grouped the data by zipcode, but considering the huge number of zipcodes we decided to change the division to boroughs. We used a histogram to show the count of each grade while using different colors, this allowed ys to show the amount of restaurants in each grade category in each borough to account for the larger number of restaurants in certain areas. The color coding and graph choice allows us to compare each borough and grades with counts and combines a huge amount of data into an extremely clear and organized chart. This graph allows us to clearly assess any correlation there may be between frequency of letter health grades depending on the borough of the restaurant.

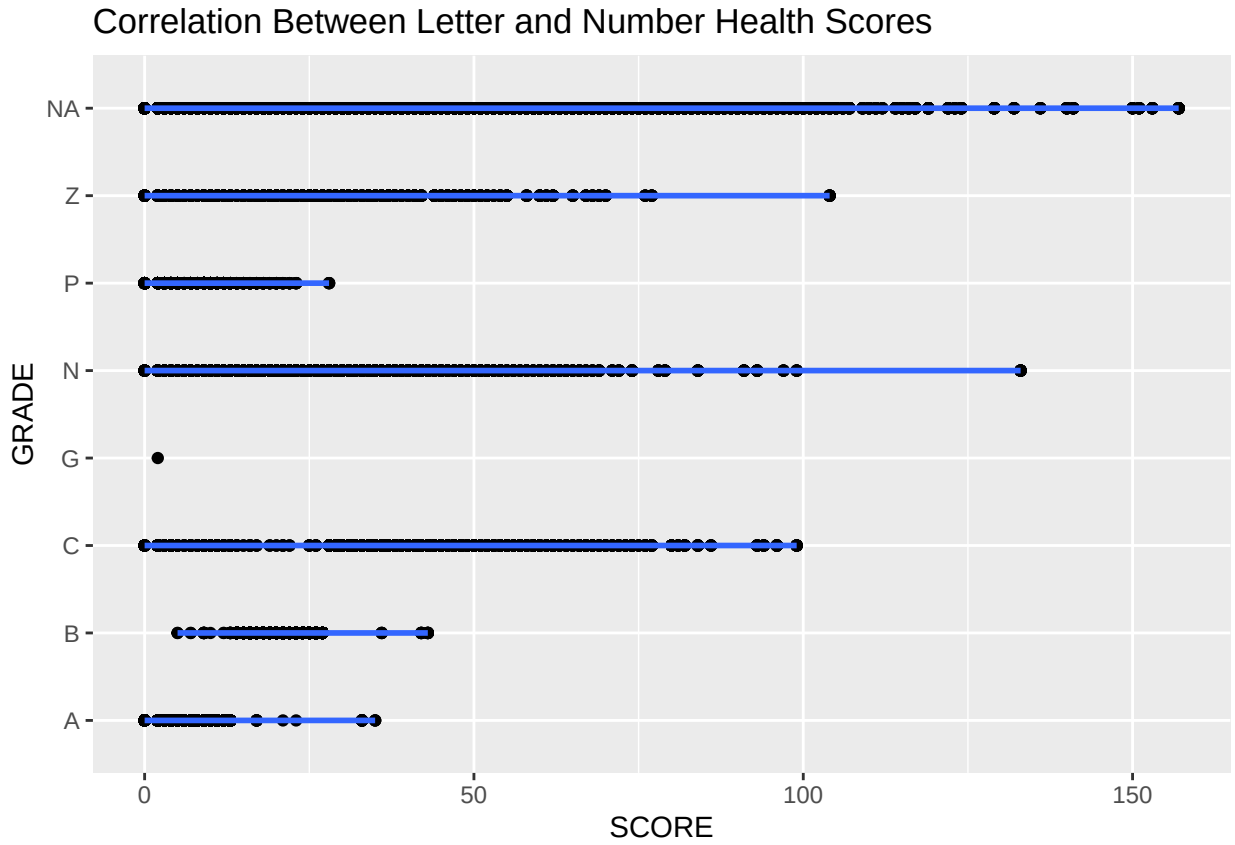


This bar graph, in comparison to the one above, shows the frequency of letter grades by borough instead by percentages. This allows us to compare each borough's health ratings relatively, which is also important in our an analysis.



We chose to do bar graphs illustrating the frequency of each health letter grade and faceted by borough This would allow us to see the spread of health ratings across each borough very clearly and see the numbers more clearly as well. This visualization is helpful in answering our question. It is one method to determine a correlation of any kind between restaurant health ratings and the restaurant's borough.

```
## # A tibble: 156,505 x 3
## # Groups:   BORO [5]
##    BORO      GRADE SCORE
##    <chr>     <chr> <dbl>
##  1 Brooklyn  A         9
##  2 Brooklyn  A         9
##  3 Queens    B        19
##  4 Brooklyn  A        13
##  5 Brooklyn  A        13
##  6 Manhattan A        13
##  7 Brooklyn  A         8
##  8 Brooklyn  A         8
##  9 Brooklyn  A         9
## 10 Manhattan A        13
## # ... with 156,495 more rows
```



Since the graph illustrates that the "Score" and "Grade" a restaurant receives is not perfectly linearly connected, we cannot just get the mean Score to see how restaurants in New York are doing overall.

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

The mean grade is between an A and a B, with A being 1 and B being 2. It is 1.22.

Results

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

```
## # A tibble: 313,540 x 2
## # Groups:   BORO [6]
##    BORO  GRADE
##    <chr> <chr>
##  1 0     <NA>
##  2 0     A
##  3 0     C
##  4 0     A
##  5 0     A
##  6 0     A
##  7 0     <NA>
##  8 0     <NA>
##  9 0     <NA>
```

```
## 10 0     <NA>
## # ... with 313,530 more rows
```

```
## # A tibble: 1 x 1
##   mean_grade
##        <dbl>
## 1       1.21
```

While the mean for all restaraunts in all boros is 1.22, the mean in Manhattan is 1.21. This is slightly lower than the total mean, leading us to believe that restaraunts in Manhattan have a higher grade on average.

```
## # A tibble: 1 x 1
##   mean_grade
##        <dbl>
## 1       1.27
```

#Discussion

The borough we found to have the highest health rating was Manhattan (1.21, less than the overall mean for all of NYC). We tentatively suggest that the Bronx has the poorest health ratings due to its 1.26 mean grade. The graphs support the order of best to worst grades being Manhattan, Queens, Brooklyn, Bronx, and Staten Island.

Some issues with our methods were the use of dplyr calculations rather than potential p-tests or any other methods that would yield statistically significant results. We will look into this for the next step in the project. The data is reliable/valid as NYC restaurant checks are done by the City of NYC Department of Health; they are unnanounced and so restaurants don't have time to prepare and fabricate a facade of health. This information was found at www1.nyc.gov/. As well, it is in the DOH's best interest to be honest about food inspection. However, it's not entirely impossible that corruption does exist. The analysis was not totally appropriate as we did not fully explore the research questions.

If we were to start over with the project, we would have looked further at the breadth of the data. Because there are too many zipcodes to analyze them all, it was much more intelligent to study overall boroughs. We also would have checked the type of each variable (string vs numerical) so as to more efficiently observe the correlations without having issues with choosing variables themselves.