Shrina Parikh

Assignment 2

URL:

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Source used to validate changes to ‘Grade’ column:

https://www1.nyc.gov/site/doh/business/food-operators/letter-grading-for-restaurants.page

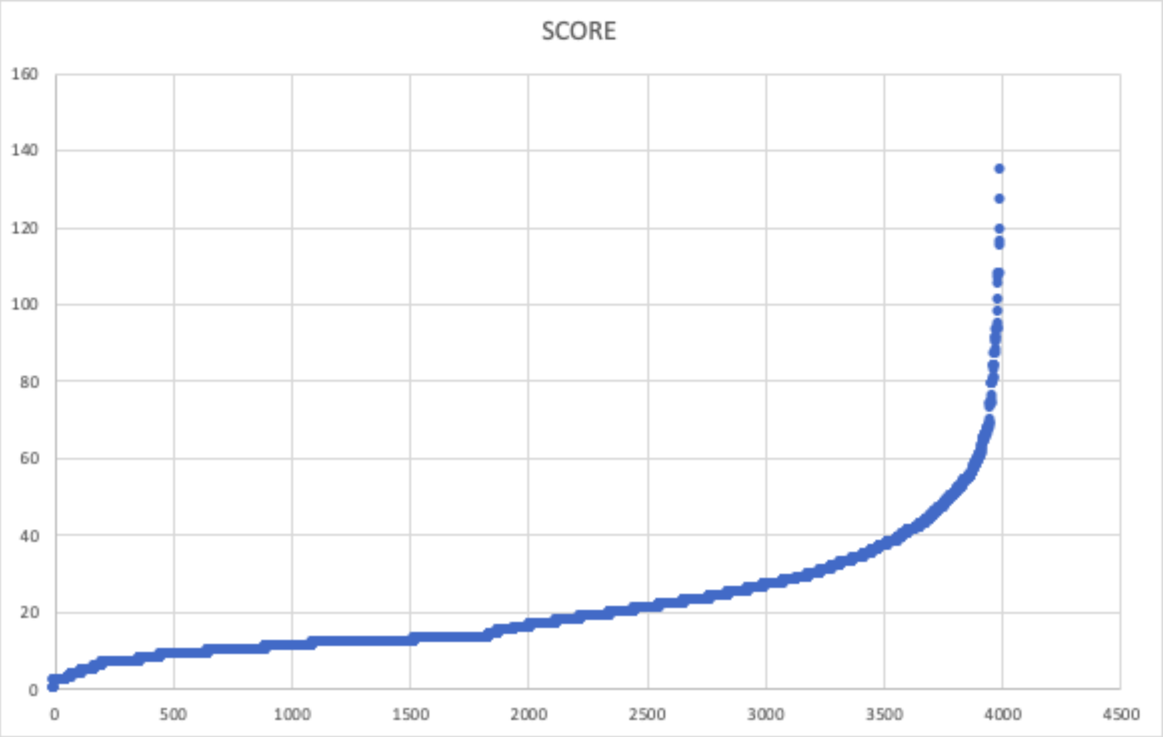
Although I wanted to do work with data focused on politics or world poverty/state of women/etc., I had trouble finding viable datasets, but I will keep looking. I found this dataset detailing health violations across NYC restaurants, and I found it especially interesting considering the frequency of health violations. I trust this data because NYC Open Data was a recommended site to use for this project and because the NYC Open Data consists of data released by NYC city government, a reliable source. This dataset could be used to see if restaurants with citations receive less citations over time. The dataset can tell us about how common various health violations are in NYC over all as well as within each borough. Moreover, the data can also be used to analyze if certain types of violations are more common to certain neighborhoods, boroughs, and/or cuisines.

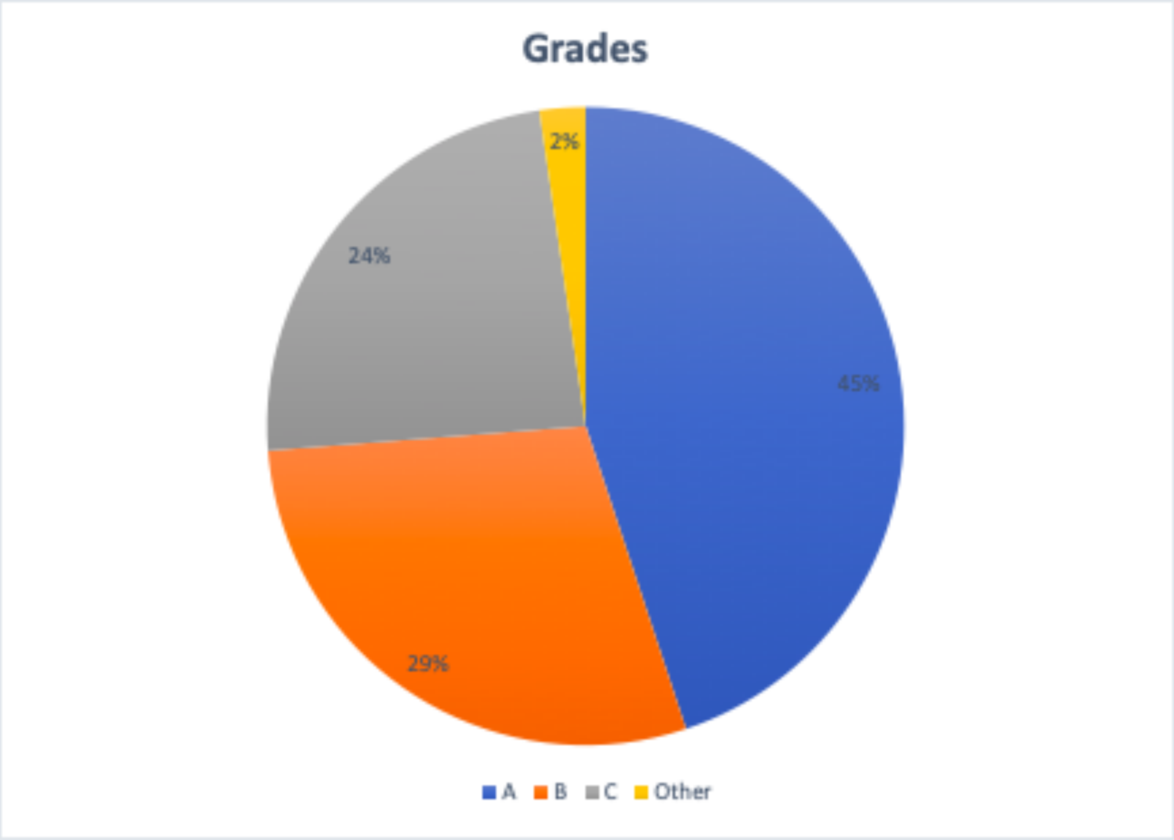
I eliminated rows without valid any recorded violations, populated some empty fields, eliminated irrelevant columns, put together 3 columns to create an entire address field, removed data entries with a missing score, and populated the entire ‘Grade’ column based on the score. The resulting dataset was about 200,000 rows (originally about 400,000), but I only wrote the first 4000 data entries to the output file for the purposes of this assignment. The changes are detailed and labeled in my python file as well.

In order to examine the data, I ran a few basic statistics. The average score was 20.72725, which is a little higher than the median, 16. Thus, the scores in this data sample skew higher. The most frequent score within this sample is 12. The max and the min are 135 and 0, respectively. A 0 score means no violations, so I (in theory) should go back into my python script and remove rows with 0 scores. Additionally, 135 seems like an outlier. I also calculated the averages within this sample based on the grade scores. The average scores for grade A, B, and C restaurants are ~9.93, ~20.677, and ~42.45, respectively.

To create the following pie chart (below), I used the following formula to count the frequency of A,B, and C grades: COUNTIF(K1:K4001,"A"), replacing "A" with the appropriate character. To calculate the frequency of "Other" grades (there are 'Z' grade ratings and I don't know where they are coming from; 'P' grade ratings mean pending), I simply subtracted the sum of the previously found frequencies from 4000. I also plotted a scatter plot on the SCORE column (with an empty X-axis field, so the X-axis in the plot below can be ignored). This graph tells us that there are a number of outliers toward the upper range of the scores.

I think this data could be interesting to learn more about restaurant health and safety in New York City, although it would need to be cleaned more. For example, the 'Z' values in the 'GRADE' column should be handled somehow as well as the outlier values. However, I have reached out to data services and am continuing to look for a dataset I find more interesting for future projects.


SCORE
160
140
120
100
80
60
40
20
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500


Grades
2%
24%
45%
29%
■ A ■ B ■ C ■ Other