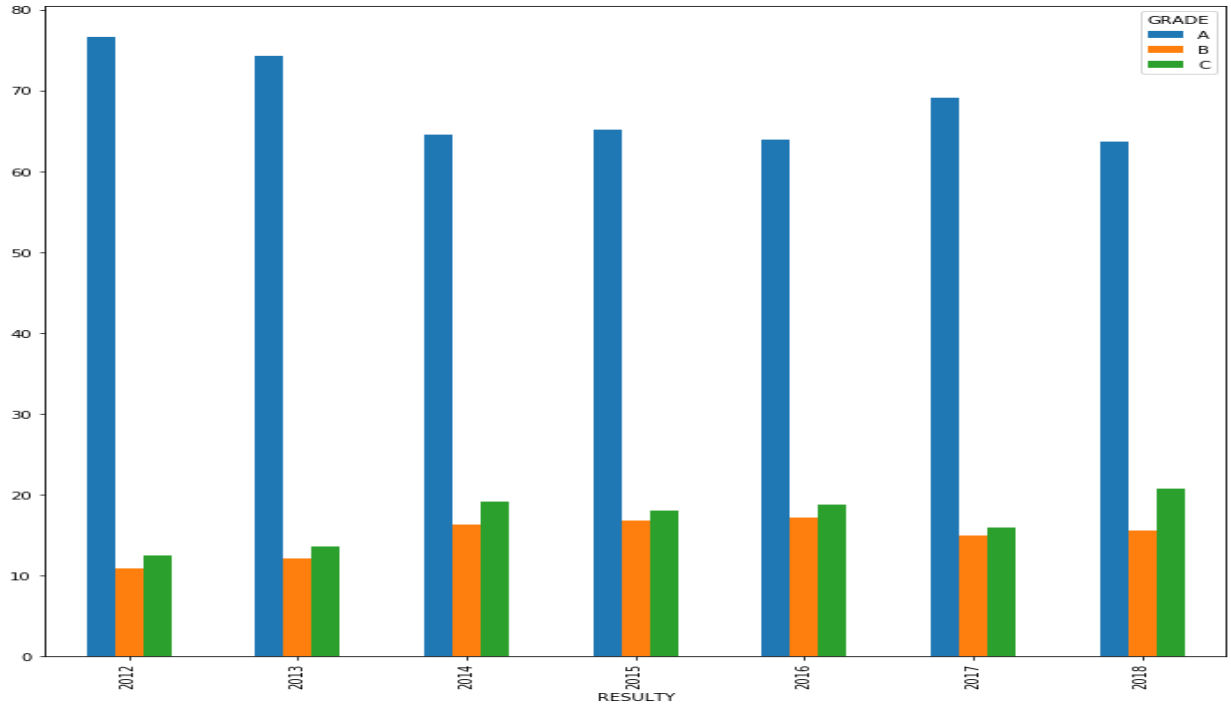
This file explains general methodology of the solutions and the visualizations generated by the solutions. For code details, please refer to the .ipynb file

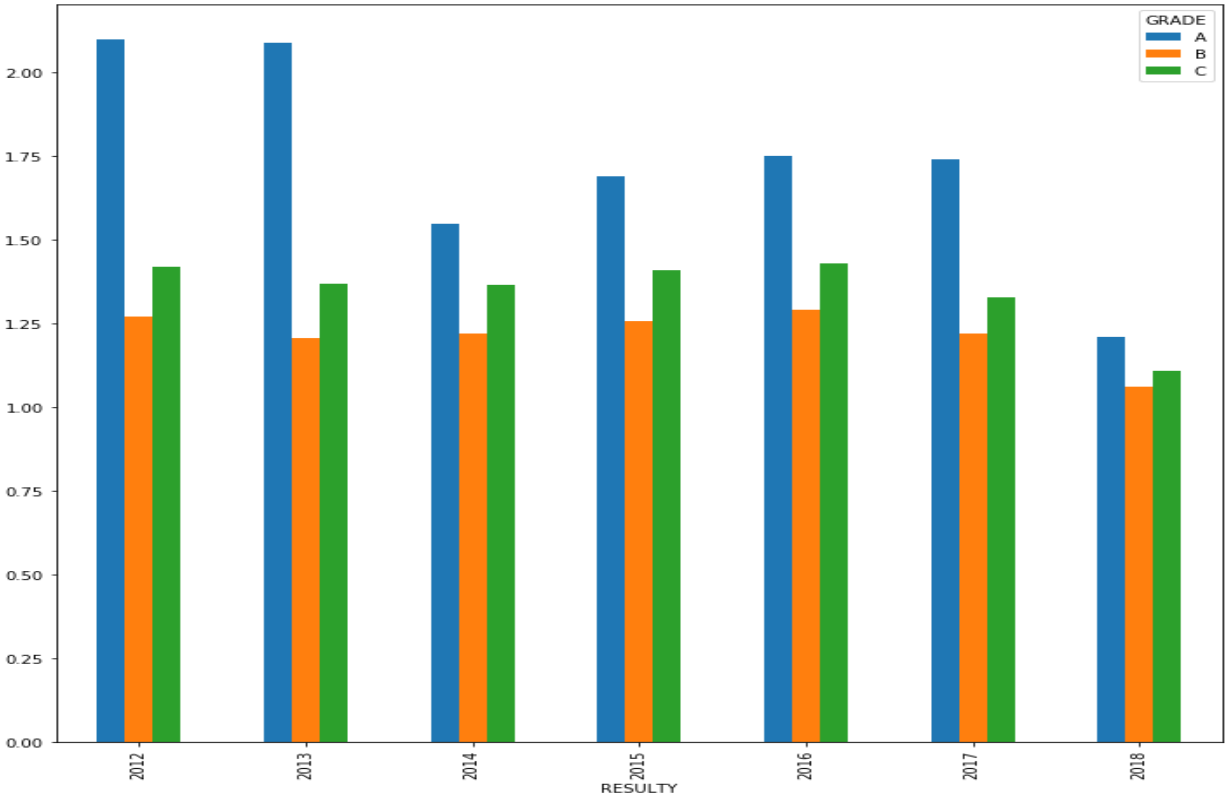
- Is the City conducting inspections effectively? Are poorer performing establishments inspected more often than better performing establishments? What are the trends over time?

Methodology:
I used the grades dataset for this problem because I could take advantage of the "GRADE" column to differentiate poor and good food establishment. I firstly analyze the percentage of inspections for establishments with different grades in distinctive years. For example, if there were 1000 rows labels as "A", "B" and "C" respectively in year 2012, then the percentage of grade "A" establishments inspected in 2012 was 1/3. We can see from the picture below, the percentage of good establishments being inspected is high and decreases slowly in years while the percentage of inferior establishments is relatively low and increases slowly. According to this general picture, the city isn't conducting inspections very effectively since grade A establishments have been inspected the most frequently. The trend has been getting better slowly over time.



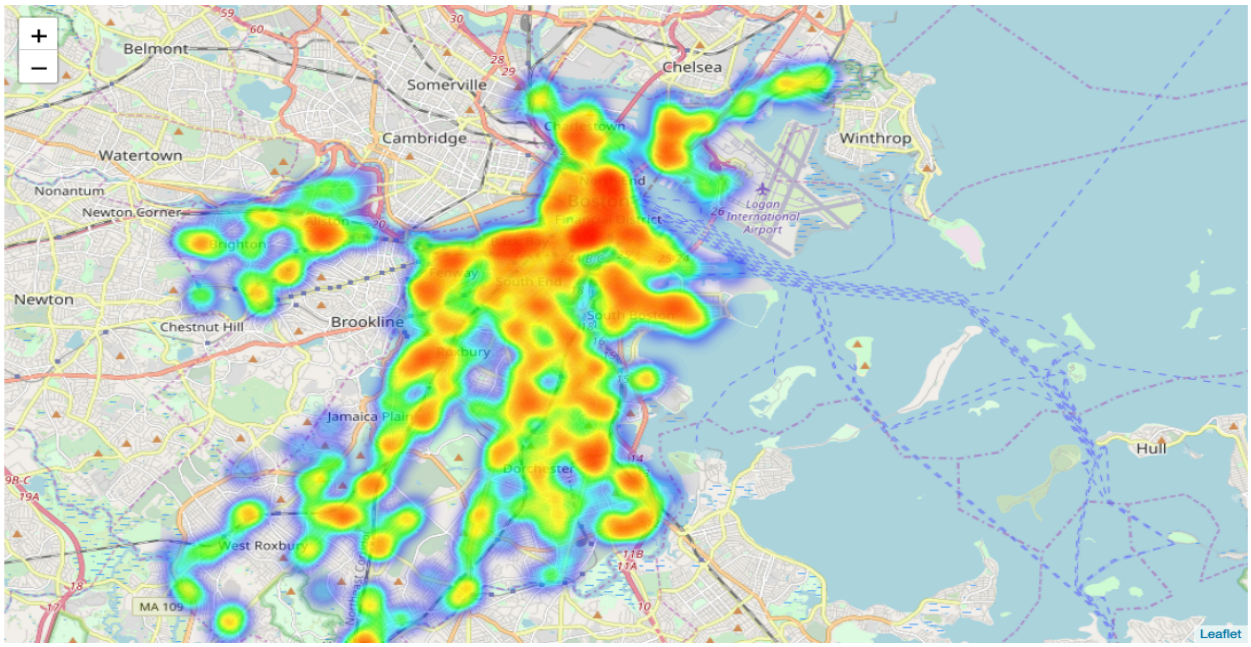
One concern of the analysis above is that grade A establishments are actually the majority of the establishments and thus would be inspected most of the time. So, I did further analysis to check average inspection times per grade over distinctive years. The formulas is (total establishments of grade x in year y / unique establishments of grade x in year y). For instance, if there were 1000 rows labeled as grade A 2012, and there were 500 distinctive establishments labeled as A in 2012, then the average inspection time for grade A establishments in 2012 is 1000/500=2. We can see from the result below that Grade A establishments still have the highest averages, but the gap between good and bad establishments are getting significantly closer over the years.

Therefore, the city hasn't been conducted the inspections very effectively in the past but tend to improve the situation over time.



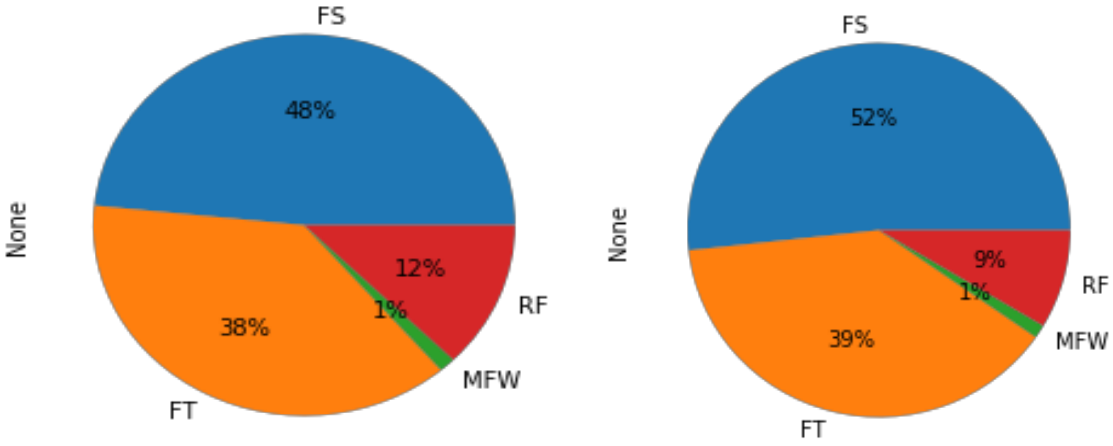
- Are there commonalities or trends among establishments that perform poorly?

Methodology: For this problem, I'm interested in the distribution of the poorly performed establishments and the categories of these establishments. I treat the grade C establishment as poorly performance indicator and merge the grades table to the violations table so that I can use column features from both datasets at the same time. After merging two tables, I selected C grade establishments and delete duplicated location so that a single establishment wouldn't be counted multiple times. Then I manipulated the latitude and longitude of all poorly performing establishments to draw a heat map as below:



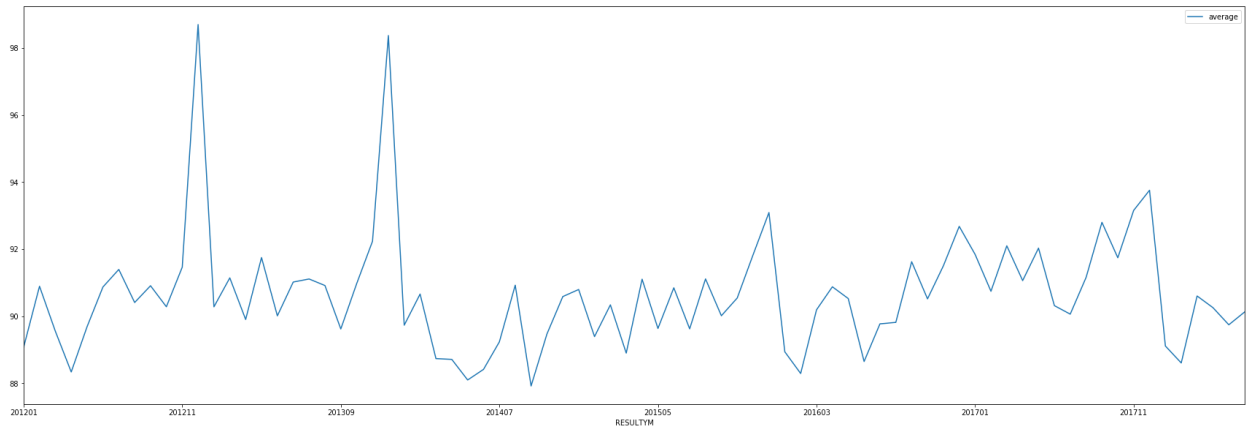
The red section means densely populated poor performance establishments. We can tell the location similarities of bad establishments. In the .ipynb file, we can scroll the scope up and down to have clearer look of bad establishments and this feature makes it more intuitive for government officials to focus on certain areas instead of using the dull latitude and longitude in the dataset.

In addition, I studied the categories of the poorly performing establishments. The left pie chart displays the percentage of all establishments categories while the picture on the right shows the percentage of bad establishments (grade C) categories. We can tell that FS and FT (restaurants-typed) have more bad establishments than normal and RF (grocery-typed) has less bad establishments than normal.



Do establishments improve their performance on inspections over time?

Methodology: For this problem, I consider investigating the average scores of all establishments in monthly manner. In order to so, I manipulate the grades dataframe to calculate the mean scores for all establishments in each month and finally draw the time series plot:



As we can see from the plot, the performance of establishments doesn't increase over time, rather it generally moves up and down between score 88 and 94.