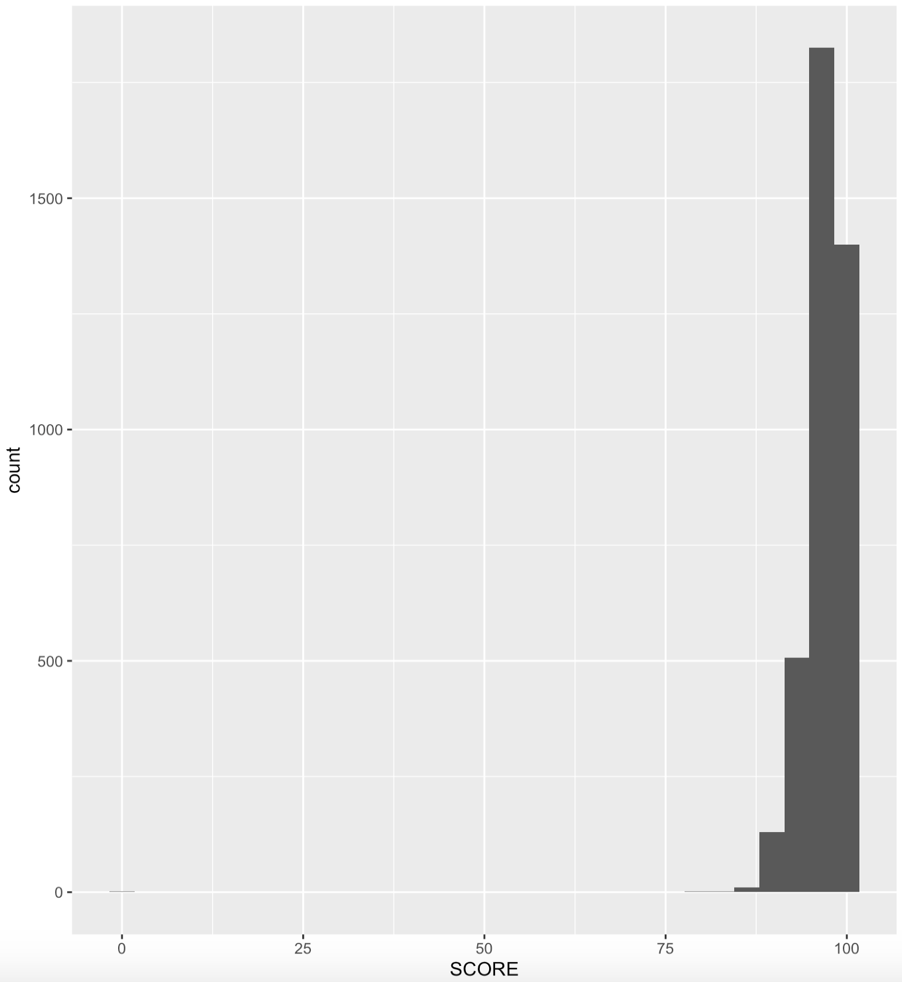
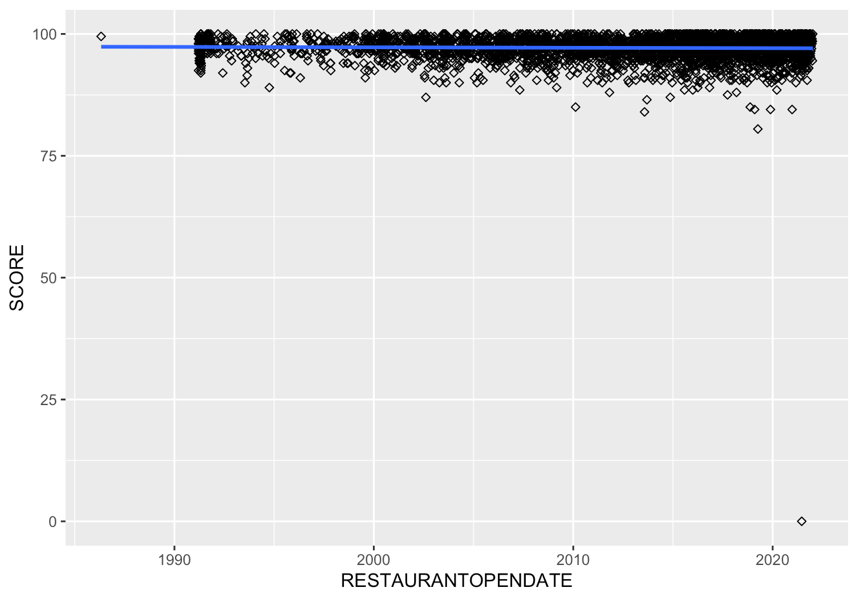
Repository: <https://github.com/cassandrariggs/plan372-hw2>

1.Visualize the overall distribution of inspection scores using a histogram. [1 point]



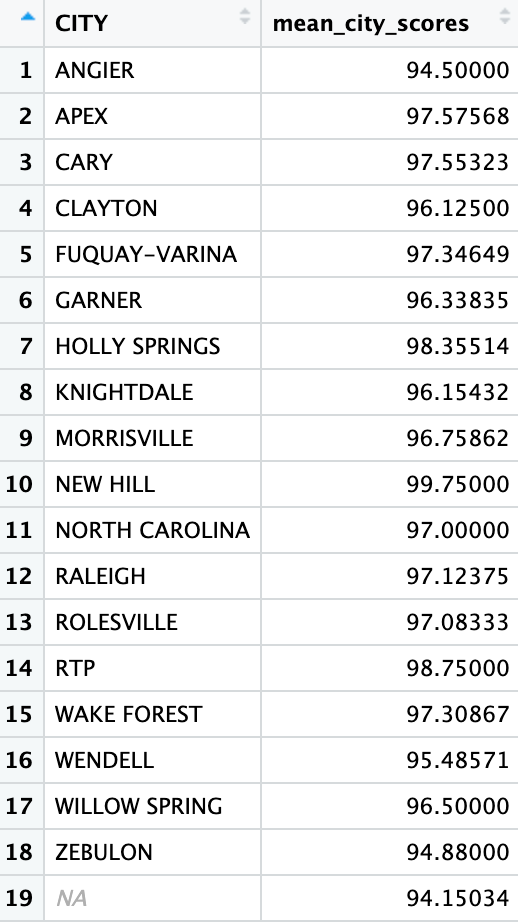
2. Some restaurants have been in business much longer than others. Is there any trend in   
terms of how highly older vs. newer restaurants score on their inspections? [0.5 points]

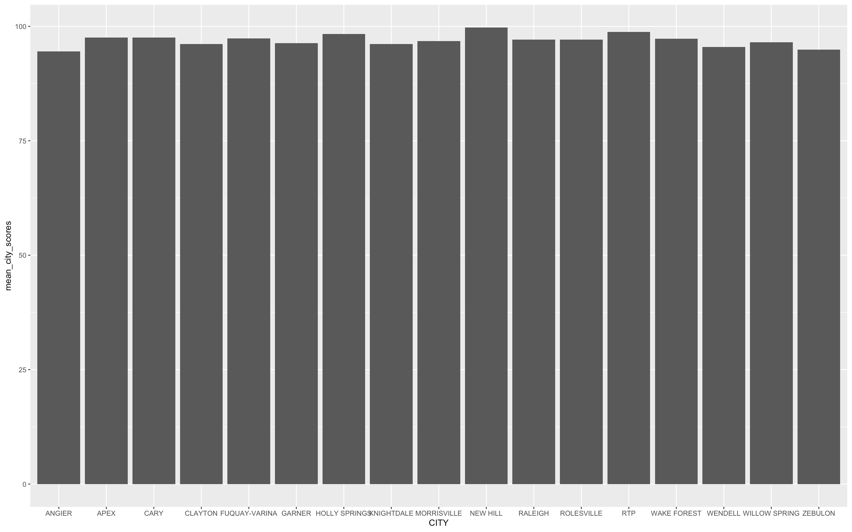




No, there seems to be no correlation between the restaurant open date and their score as determined by the chart and the small r^2 of 0.0006440471. The restaurant open date column was missing 296 data points which were omitted from the graph, which could possibly be affecting the correlation positively or negatively between the restaurant open date and score depending on the scores’ missing years. Unless the scores are very low for either younger or older restaurants though, it should not majorly affect the overall correlation because the r^2 is very low at .064%

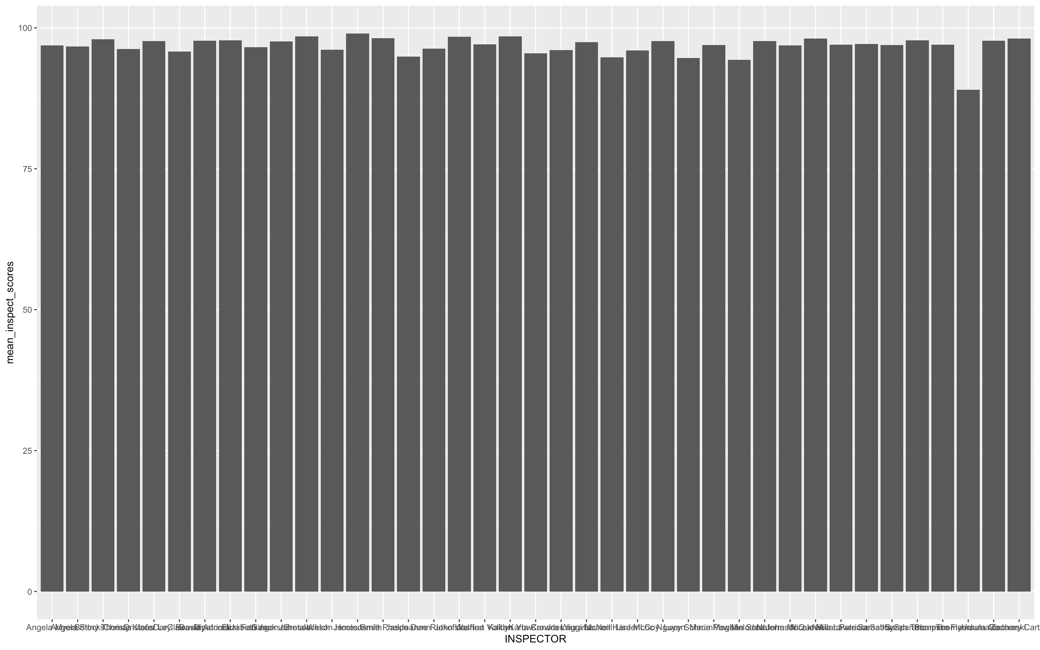
3. Wake County is the most populous county in North Carolina, and there are many cities   
in it. Do the inspection scores vary by city? Note that the city column contains some   
differently spelled city names; make sure to clean those up so that there is only one   
estimated value per city. The recode function that we used for creating a   
weekend/weekday variable in the SFpark exercise will be useful here, and you may also   
be interested in the str\_to\_upper function. [1 point]

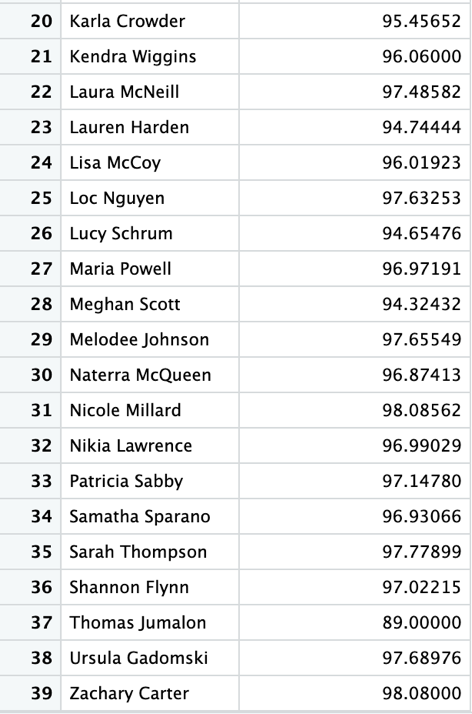




After recording the names of the cities, I developed a bar chart, which displayed some varying scores among cities. While not large, there is some difference between New Hill’s 99.75, the highest, which is almost a perfect 100 for an entire city, vs Angier’s 94.5, the lowest, which is only slightly above an A-. The observation “NORTH CAROLINA” only has one data point, which I omitted because they did not specify the city. So, because there was only 1 data point for this, it most likely will not greatly affect the other data if removed. There are also 296 missing city observations that were omitted from the plot, which have the potential to change the mean scores for the cities as they could for the restaurant’s age. The NA category has the lowest score, which means depending on how dispersed among the cities the missing scores are from, they could impact a city’s score majorly or not at all, which should be noted in all analyses looking at mean city inspection scores in North Carolina as well as further investigated to determine the true effect.

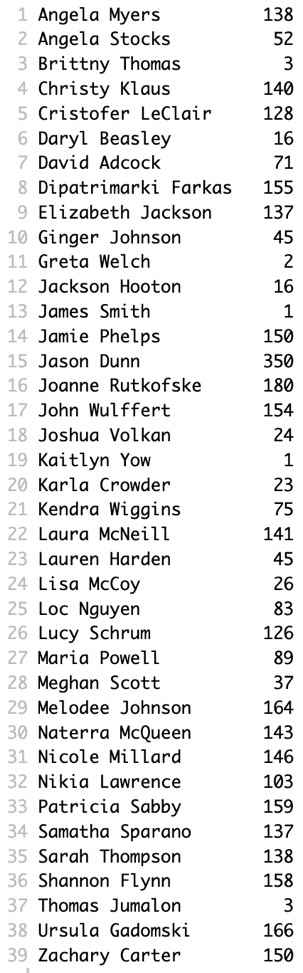
4. Wake County employs a whole team of inspectors. It is possible that some inspectors   
may be more thorough than others. Do inspection scores vary by inspector? [0.5 points]

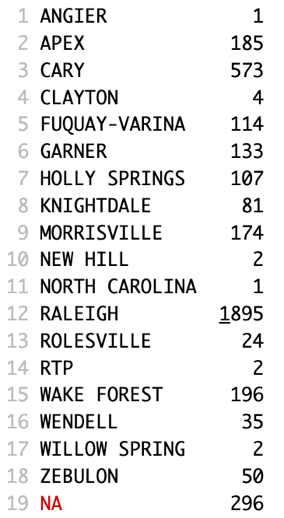




There is some variability between the inspector’s mean scores. A 99 and an 89 is a large difference in the food industry and can cost them business. The 89 does seem to be an outlier though as most of the other scores are 95 and above and should be further inspected to determine the cause of the low score and the effect on the data. I did not remove this outlier because there is a number of reasons why it could be much lower, so it should not be counted out yet. There is a possibility the inspectors with scores around 95 and 96 are more thorough than the inspectors with higher scores around 99 since there is some variability.

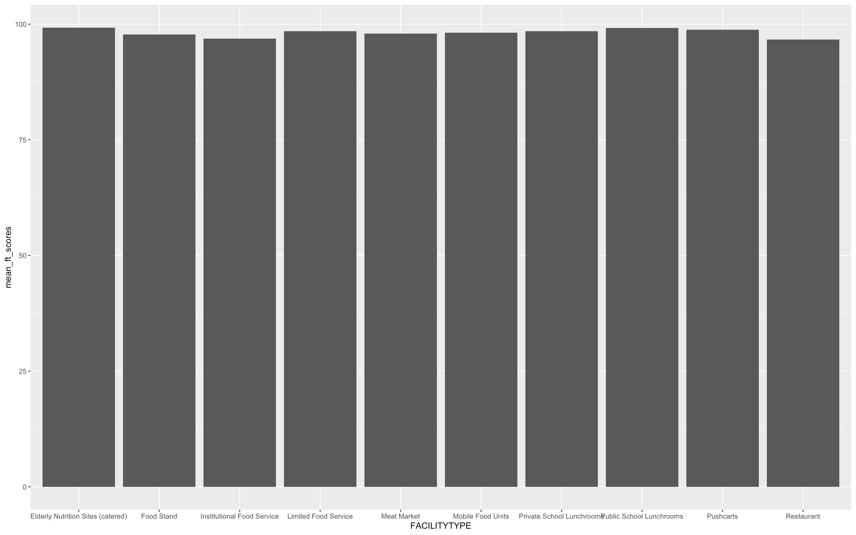
5. It is possible that some extreme results from the previous questions are due to small   
sample sizes in a particular city, for a particular inspector, or in a particular time period.   
Look at the sample sizes in each of your groups. Do you think this is an explanation for   
the results you came to above? [0.5 point]

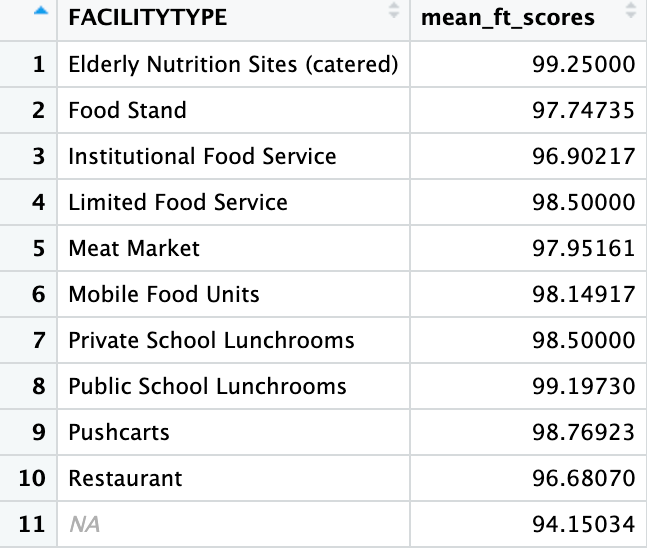




I believe the sample sizes are definitely playing a role in some of the mean scores for the cities and inspectors. The samples range from 1 – 350 for inspectors and 1 – 1895 for cities, which is a large difference. The main outlier for the inspectors’ scores has one of the smallest sample sizes, meaning his small sample size might not represent what his mean score will be after gathering a larger sample size, therefore, there’s a possibility it is not a good representation and is skewing the data as a whole. This can be explained by the large sample size theory in a distribution that as the sample size increases, the sample mean becomes closer to the population mean.

6. The data file contains records for many types of food-service facility (e.g. restaurants,   
food trucks, etc.). Are the scores for restaurants higher than other types of facility? [0.5   
point]

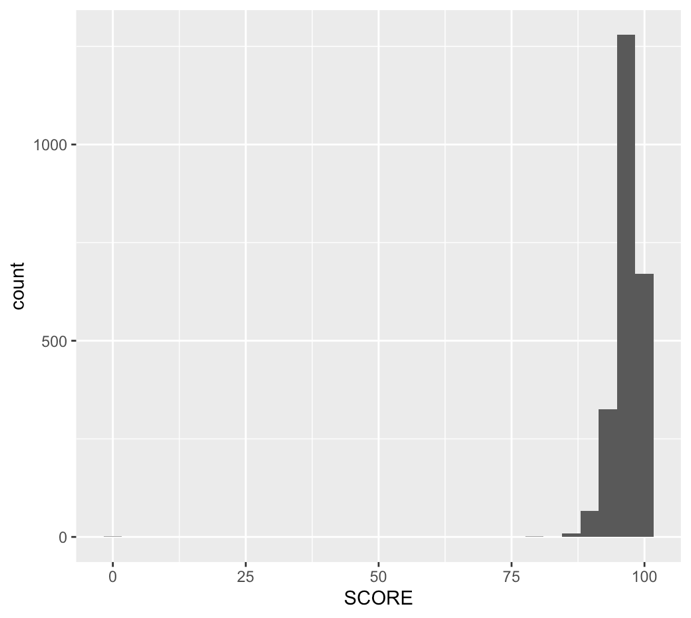




Restaurants are actually on the lower side of the scores when compared to the rest of the facilities. There are 296 NA observations though with a much lower score than the rest, which means removal could possibly have a big effect on the overall scoring averages for all of the facilities. If the missing values were all restaurants, the restaurant category would be lower, but if none of them were restaurants, there is a potential that restaurants would be moved up in the rankings if lower scores were dispersed among the other facilities. Further investigation into the missing values would be needed to determine the effect of NA removal on mean facility scores.

7. Since restaurants are where the general public is most likely to interact with the food-  
service system, Wake County Public Health is particularly interested in sanitation in   
restaurants. Repeat the analyses above (1-5) for restaurants specifically. [2 points]

1.



Restaurant only histogram of score distribution.

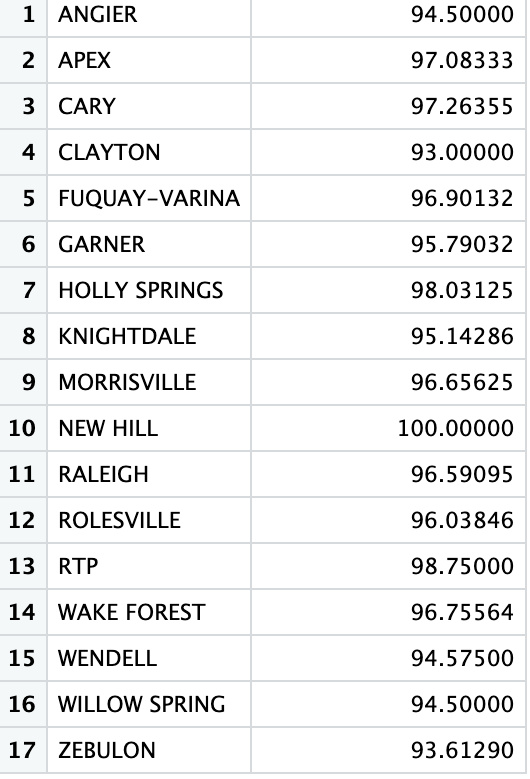
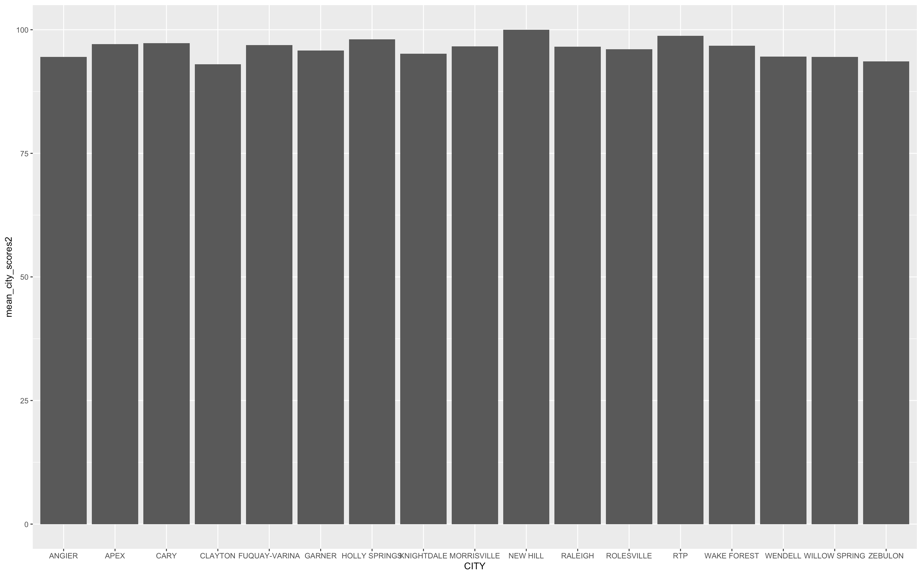
2.





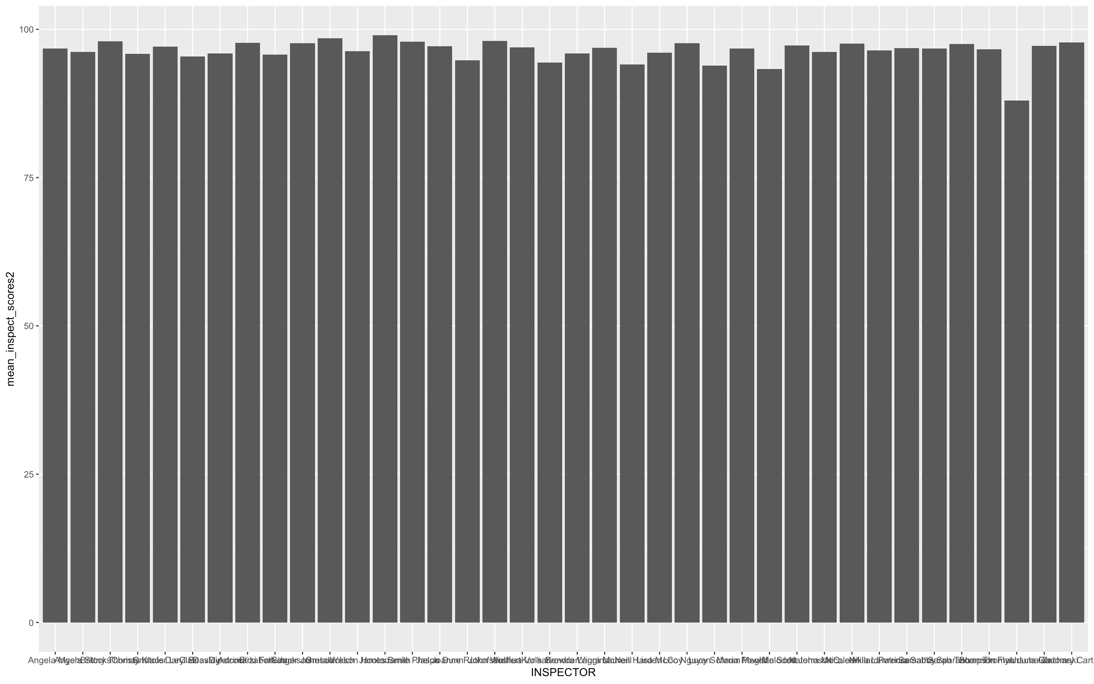
There is even less of a correlation between restaurant open date and score than between all facilities and open date. One big difference is that there is no missing data for restaurants only, meaning the missing open dates were for other facilities.

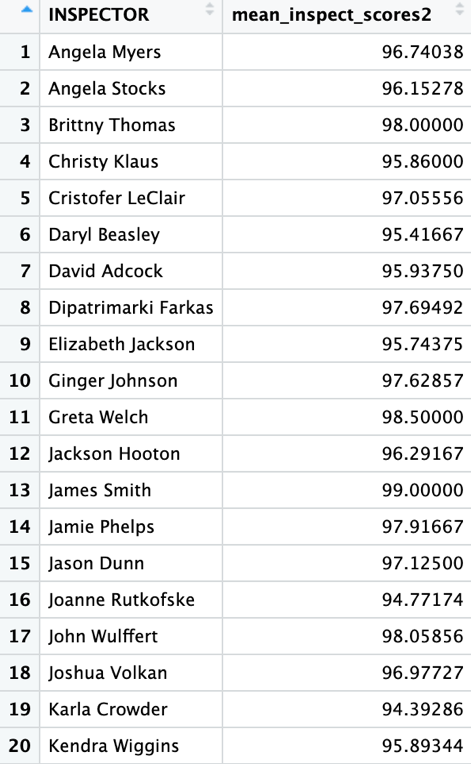
3.



There is some variation among only restaurant scores in different cities. The range is larger than when all the facilities are included. There was no missing data.

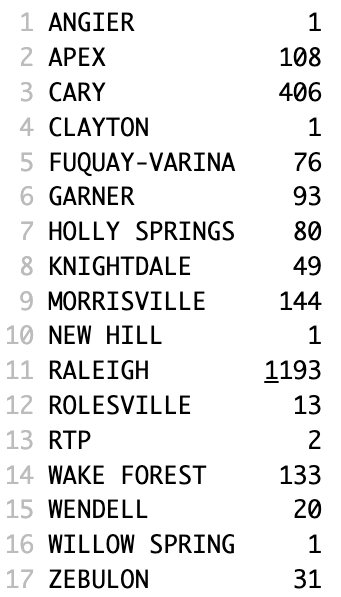
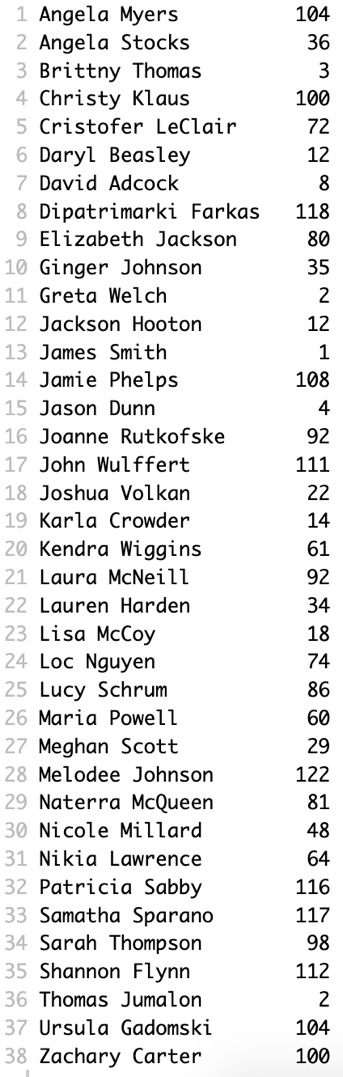
4.





There is some variation among inspectors’ scores for only restaurants. The range is also slightly larger than when all the facilities are included. There was no missing data.

5.



As for all the facilities, the sample size ranges are large, which could have caused some variation among the scores. The lower sample sizes may not be a good representation of the population (Inspectors or cities) as a whole.