Data Storytelling Report

After the food_business dataset is loaded, I started with .info() and .describe() to observe the general information on the data. There are 3 numeric columns and 6 string columns. For my project, stars rating, review count, and population are important variables. I decided to visualize the distribution on these three variables.

From the distribution, percentiles, and mean of the stars rating, we can see most ratings falls between 3.0 to 4.0 stars. One of the interesting discoveries is the relation graph between review count and stars rating is very similar with the distribution of stars rating.

The numeric and visualized general review count distribution both showing there are many outliers and most of the counts fall under 100. To further exploration, I “zoom in” to review counts under 100, then do the numeric and visualized result again. The result under 100 counts showing most review counts fall under 40 counts.

The distribution of population overthrows my initial assumption that more population means more food businesses. The graph shows no correlation between the number of food businesses and population. This might due to lack of spaces to open businesses. High population doesn’t necessarily mean bigger land. Since I plan on using population to determine the review counts range, I plot population and review count to see if there’s relation between the two variables. I plotted 4 graphs, all with population as x axis and review count as y axis. At first, the graphs don’t seem to show any correlation between the two, until I plot the sum of review counts with population. The graph is roughly showing more population does end up getting more review counts in total.

After researching the relation between population and review counts, I suspect that whether the city is a tourist place could also be a key factor. Tourists don’t count in the population but they do leave reviews. Thus, I decided to observe the relation between review count and states. I first plotted the distribution of state, and found there is only one food business in KY, MD and TX in the dataset. Then I made a graph with median of review count and a graph with sum of review count that both based on states. The graph with sum of review count based on state shows more food businesses does end up with more review count in total. I also list out the cities in top 5 and bottom 5 review counts and population respectively. Comparing the results, I found it interesting that there are 4 cities in both top 5 list, but only one in both bottom 5 list.

In conclusion, the cities with large population could indicate it’s a big city thus review counts in total is higher. The cities with less population could indicate it’s a town or smaller city thus review count in total tends to be less.