YELP HEALTH INSPECTION SCORES ANALYSIS

By Chaitra Subramaniam

Objective:

The dataset I will be using is the Yelp data about Local Inspector Value Entry Specification (LIVES) from 2012. https: //www.yelp.com/healthscores. The dataset provides valuable information about businesses, health inspections scores they received, and reasons for health inspections. Even though the Yelp data is available for numerous counties, I picked 5 (Anchorage, Fortworth, Evanston, SF, and Boulder) as they span different coasts of the United States, have various sizes, and have differing mean income groups.

The overarching question I want to ask is how do scores differ by county and can we find any reasons as to why they do? This is interesting because my presumption is that the more expensive the county, the better the food standards will be and therefore, a higher score. However, I want to cut out this bias and try to understand the relationship between county and score using data science.

Key Takeaways of EDA:

Figure 1:

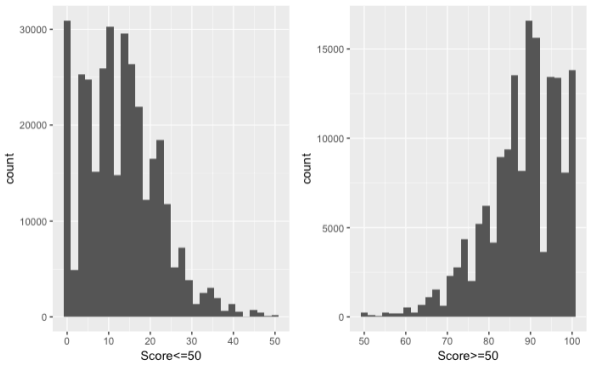

Figure 2:

1. There are significant distinctions in scores by county (Evanston has the highest mean score, Fortworth has the lowest)

2. Distribution of the scores are higher towards the lower and upper ends (0,100). Very few scores actually score 50 (or average) – *Fig 1*

3. The most common words used in the descriptions differ by county but are not super representative of why the counties were scored a certain way (mostly talk about food, surfaces etc.) but it hard to reach a conclusive judgment on quality (no words like bad, good etc.)

4. Mean county scores don't change over years. If the county receives a certain score one year, they are likely to receive the same the next. This is interesting because you'd expect the counties that score less to be able to improve and bump their scores or counties that score high to have "bad years" where they don't score that well – *Fig 2*