Geovanna Karina Meier

# EDA and Statistical Inference

## Predicting Critical Health and Safety Violations at Food Establishments

The general hypothesis for this project is that inspection results and passing rates are different across months, food categories, and the establishment's star ratings, business ratings, and review count.

The null hypothesis is that holding everything else constant, the probability of receiving a passing grade, is the same for all food establishments, independent of the month the inspection occurred, the type of food, Yelp's stars ratings, business ratings, or review counts. The alternative hypothesis is that there are statistically significant differences in compliance rates across months and food categories.

Two variables in particular were considered to be significant in terms of explaining Downgrades. The first one is the month the inspection occurs. The hypothesis is that seasonal changes in weather may affect food handling, preparation standards and increase the likelihood of flies and mosquitos around food. For instance, higher temperatures during the summer months may result in higher food violations as higher temperatures may increase the growth of bacteria if food is not prepared or consumed promptly or food surfaces.

The second variable is the type of food. The assumption is that establishments with buffet options and/or raw food menus may have higher food violations due to stricter standards of temperature and food handling.

The first step was to test if there was a difference in compliance rate by season. Thus, the hypotheses are:

$H_0$: There is no significant difference in compliance rates by season </p>

$H_A$: compliance rate of establishments in warmer months $\neq$ compliance rate of establishments during cooler months

$\alpha = 0.05$

Based solely on the inspections results data, the EDA shows that more inspections are scheduled during Spring months with a slightly higher of non_compliance results. Using the frequentist approach, compliance rates were compared across seasons.

Based on all the inspection results, the compliance rate in winter is slightly higher compared to other months with 88.7%. Compliance in Spring is 87.93%, 87.89% during Fall and Summer has the lowest rate at 87.4%. Again, this is only slightly lower compared to winter, which has the highest compliance rate.

In order to see the likelihood of getting these inspection results again, the data for winter and summer were randomly sampled, ignoring the season ‘labels’ to see if

there was a statistically significant difference between the compliance results obtained during each season. Using a t-test, the results showed that after 10,000 simulations, the difference is statistically significant at the 0.05 level and non compliance rates are higher during the Summer months.

The comparisons were also run on the merged dataset, which includes the inspections results and Yelp data, and the results were the same, obtaining a p-value of 0.001.

****

Considering the type of food, the hypotheses are:

$H_0$: There is no significant difference in compliance rates by food type.

$H_A$: compliance rate among establishments varies between food type.

$\boldsymbol{\alpha} = 0.05$

Looking at the distribution of compliance rates across different establishment types, Yelp data showed that establishments categorized as 'Japanese' had a higher number of 'non compliant' results followed by Chinese, Mexican, Seafood and Thai food type categories. Considering their compliance proportions, establishments categorized by Yelp as Japanese had a 65.76% compliance rate, which was the lowest, followed by Chinese with 68.1%.

****

A comparison of Yelp's stars rating and the inspection results, showed that the vast majority of restaurants with daily four and five stars rating averages have positive compliance results.

**Correlations**

A heatmap of the correlation coefficients showed that inspection demerits (total violations values) and violation counts are closely related to the inspection results, with inspection demerits having a higher negative correlation with the results. While not all food safety violations have the same impact on the inspection results, as different violation type carry different weights, in general, a higher violation count results in higher demerits and a non compliance inspection result.

**EDA Conclusions**

Thus far, the analysis indicates that the inspection month and the type of restaurant can help prioritize inspection frequencies and optimize schedulings.