Team 3 Project 1 Proposal

Project Title:

What did you eat?

Relationship between LA County restaurant health violations and yelp ratings/reviews.

Using yelp API we are interested in determining if there is a correlation between star rating and health violations in the County of LA restaurants. Some of the questions we are interested in answering are:

* Is there a correlation between star rating and health violations?
* Do high starred restaurants show higher propensity for health violations?
* Do lower starred restaurants show a higher rate of health violations in comparison to higher starred restaurants?
* What areas in the County have higher propensity of health violations in restaurants?
* What type of restaurants, Mexican, Chinese, American, etc. show higher/lower health violations?

This information can be very useful to us foodies who not only care about savoring a good meal, but also care about the quality of the food we consume in this great metropolis.

Data:

We will also use Yelp API data to get the restaurant star rating.

<https://www.yelp.com/dataset>

Team 3 Members:

* Allison
* Binqian
* Maritza
* Orlando

Tasks:

1. Get proposal approved
2. Get API data from yelp
3. Clean up API data
4. Standardize data
5. Data Exploration/Analysis
6. Data Visualization
7. Conclusion/Report
8. Presentation

Team 3 Project Summary

Project Idea:

Initially we wanted to find a correlation between the yelp star ratings and health violation data for restaurants in LA County. However, as a result of the limited data we could get from yelp, we decided to change course in our project and relied solely on the health violation data. Based on the new direction of the project we developed the following new hypothesis:

* How many violations are committed per year?
* Which month exhibits the highest number of violations?
* Is there a correlation between the size of the restaurant and propensity to commit violations?
* Which restaurants commit the highest and lowest violations?

Findings:

How many violations are committed per year?

Based on our analysis restaurants in LA County committed 260,685 violations in the 4-year span between 2015-2018. Based on the data, 2016 was the year where more violations were committed (See Figure 1).

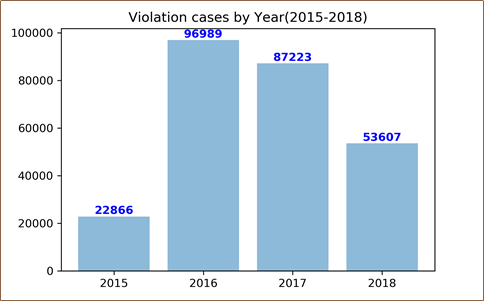


Figure 1 - LA County restaurant health violation cases by year (2015-2018).

Which month exhibits the highest number of violations?

We also wanted to explore if there were variations in the total number of violations and the month. In order to do this analysis we had to take the average violations for each month for the 4 year span. Since the data was incomplete for years 2015 and 2018, we ensured that average was calculated accordingly to the number of months that had data in the 4-year span. Our analysis showed that on average more violations are committed in the month of June than any other month of the year (See Figure 2). Conversely, November had the lowest average of violations of all the months.

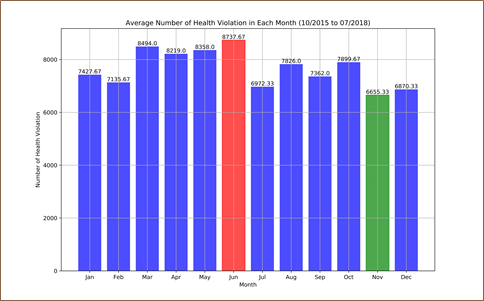


Figure 2 - Average number of health violations in each month for LA County restaurants (10/2015 to 7/2018).

Is there a correlation between the size of the restaurant and propensity to commit violations?

The LA County Data categorized all the restaurants that committed violations by seating capacity. The restaurants were categorized as follows:

* RESTAURANT (0-30)
* RESTAURANT (31-60)
* RESTAURANT (61-150)
* RESTAURANT (151 + )

Based on our analysis smaller restaurants with a seating capacity of 1-30 seats had the largest total number of health violations (See Figure 3). Conversely, larger sized restaurants with 151+ seats exhibited the lowest number of violations for the 4 year span.

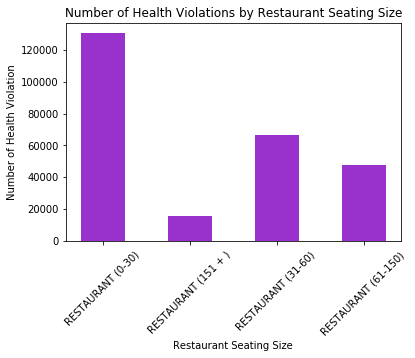


Figure 3 - Number of Health Violations by Restaurant Seating Size for LA County.

Which restaurants commit the highest number of violations?

Next we wanted to get from the dataset which are the restaurants or restaurant chains with the most violations. The results showed that larger venues for sporting events committed the most violations, Dodger Stadium being the number one offender (See Figure 4).

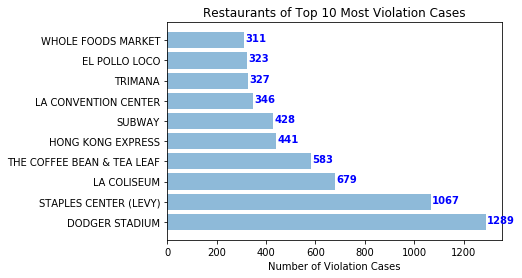


Figure 4 - Top 10 Restaurants in LA County that committed the most violations.

Other Findings:

In addition to categorizing restaurants by seating capacity, restaurants were also labeled as High Risk, Moderate Risk, and Low Risk. Although, we were not able to determine what was the criteria for each of those categories we made the assumption that based on the number of violations the County labeled the restaurants accordingly. We wanted to explore this further and plotted the frequency of the label for the different restaurants thus showing us which are the restaurants are the riskiest of committing violations and being unsafe. As a result, the restaurant that was the riskiest was Dodger Stadium due to the volume of violations committed by that establishment (Figure 5). Coffee Bean was the safest of all the restaurants analyzed.

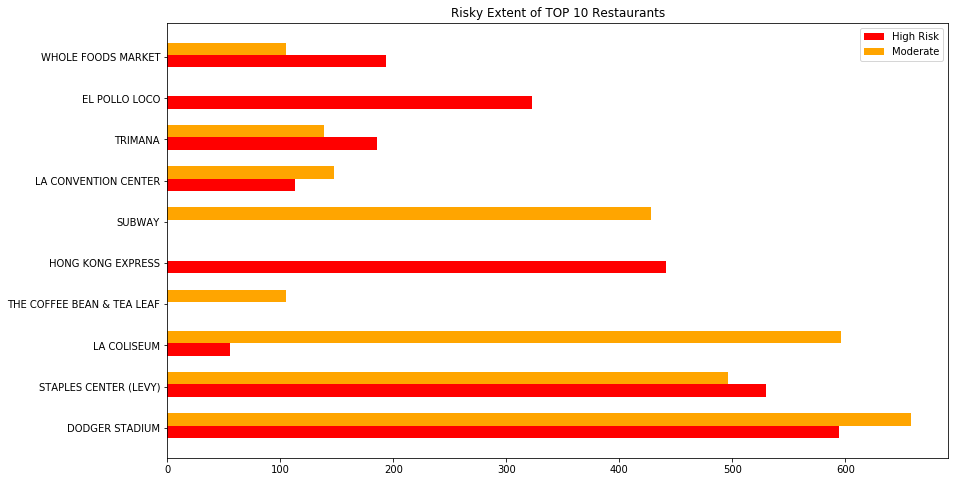


Figure 5 - Top 10 restaurants exhibiting risk factor for committing health violation. Red bar shows high risk, while orange bar shows moderate risk.

Finally, we also wanted to see what the average health score for restaurants with different seating capacity was. The results show that the average score for restaurants of all seating capacities was 91.3 and the difference of the average for each of the categories was statistically insignificant (Figure 6).

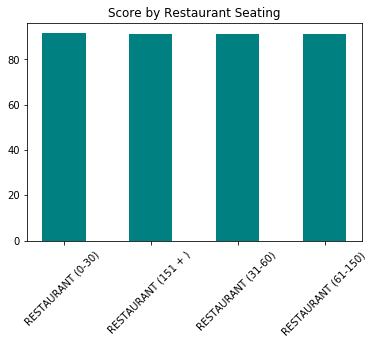


Figure 6 - Average health score for restaurants in LA County by seating size.

Conclusion:

Our analysis provided us with interesting insights about the characteristics of restaurants that commit violations in LA County.

However, we need to perform more advanced statistical analysis to determine if there are correlations that can provide us with more predictive insights.

Moreover, we also think that further analysis by linking the local demographic data and geographic data can help in finding unobserved variables among the various restaurants.

Also, conducting a survey to get consumer perception, will help us in doing hierarchical regression to find more correlations with health violations and consumer sentiment.