Health Inspection Dataset Analysis

Team Health Inspection

Ryan Permenter, Rachel Podemski, Clarence Robinson, Brett Thompson

26 February 2021

OVERVIEW

This dataset contains details of health inspections and scores in the greater Dallas area dating back to October 2016. There are a total of 51,679 individual inspections, with each report containing the name of the establishment inspected, physical location, date the inspection was conducted, overall score for the inspection, and point deduction for the individual violations. The data is provided by the Dallas Code Compliance Services Department and is updated monthly on the Dallas Open Data website.

Questions to be asked from this dataset include: (1) Is there a trend of a specific restaurant's inspection scores over time? (2) Do inspection scores vary by region? (3) Is there a correlation between the number of violations and total inspection score? (4) Where are the "best/worst" places to eat in the greater Dallas area?

DATA & MODELING APPROACH

The original dataset was available in multiple forms (CSV, JSON, API). Multiple processes were utilized to extract and visualize the data to answer specific questions. After confirming cleanliness of the data, JSON data was extracted from the Socrata Open Data API (SODA) for the ability to filter, query, and aggregate. Additionally, the CSV containing the most recent 51,000 health inspections was downloaded from the Dallas Open Data website. A column was added to the CSV to calculate the total number of violations per restaurant, then all additional columns were deleted so that only "Establishment Name", "Inspection Score", and "Total Number of Violations" remained. No additional pre-analysis work was required to use this dataset.

Several different methods were utilized to answer specific questions from the dataset: (1) A line graph was designed from JSON data extracted with the SODA API and developed through Plotly to visualize trends of restaurant health inspection scores over time. (2) A similar approach was implemented, and a bar graph was developed through Plotly to visualize inspection scores filtered by zip code. (3) A scatter plot was designed with Plotly using data from the CSV file to visualize the impact on inspection score versus total violations. (4) A heatmap was generated with Leaflet.js to visualize areas in Dallas that had a concentrated amount of either "high" or "low" inspection scores.

INSPIRATIONS

The inspiration for this project was to learn more about how inspections scores are calculated and how well restaurants in the Dallas area are performing. The colors red and yellow were utilized when visualizing our data because studies have shown that the two colors encourage an increase in appetite.

RESULTS OF DATA ANALYSIS

(i) *Is there a trend of a specific restaurant's inspection scores over time?*

Health Inspections are conducted at-least every six-months. There are a few instances where a health inspection is conducted multiple times that will be discussed later. The line graph below shows a base trend of two restaurant's (OLIVE GARDEN and CHIPOTLE) over the past four years.
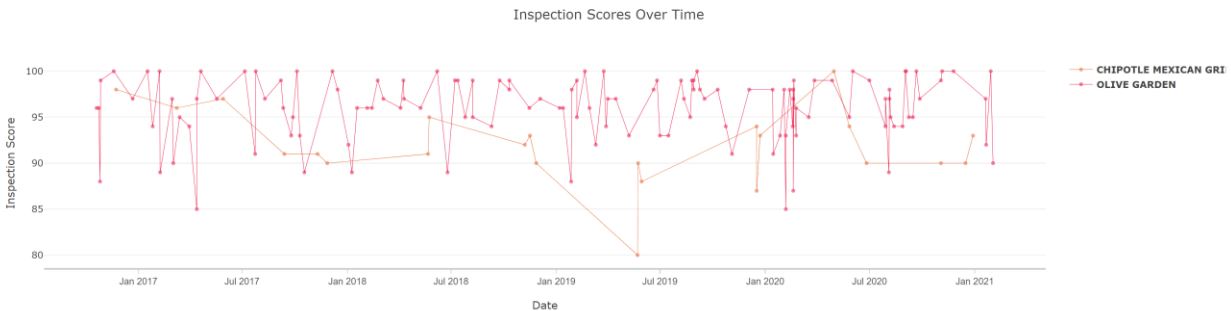

(ii) *Do inspection scores vary by region?*

The data is aggregated by zip code and the results shown are the average inspection score for each respective zip code in the Greater Dallas area. The results are sorted from lowest average inspection scores to highest, and a filter is applied to visualize any average results below a 90 overall score.
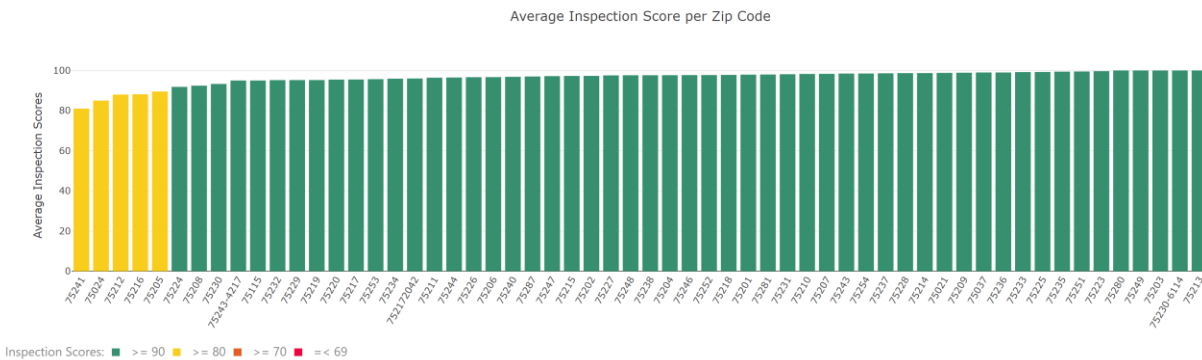

There are a total of three types of inspections seen the dataset. They are (1) Routine, (2) Follow-Up, and (3) Complaint. A routine inspection occurs at least once every 6-months for each restaurant. A Follow-Up inspection occurs within one-week and is the result of poor sanitation issues or low overall inspection scores. A Complaint inspection occurs as the result of a complaint from a patron regarding general sanitation/hygienic practices/illness investigation, smoking, or other incidents. The results of each type of inspection were filtered into buckets visualized in a bar graph.

Average Inspection Score per Type of Inspection
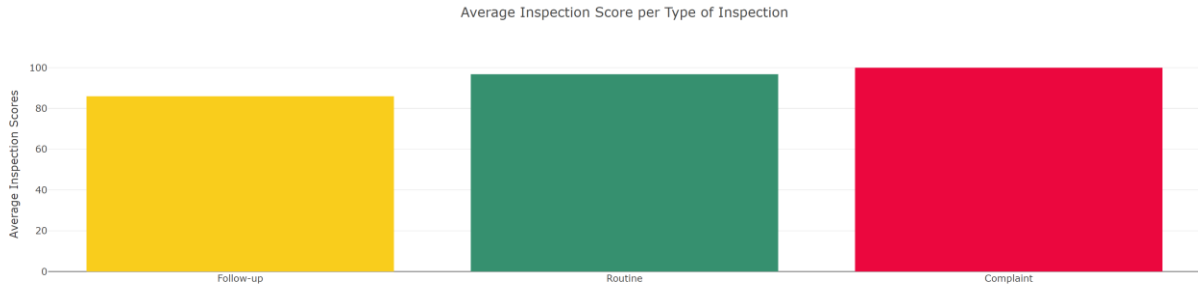
(iii) *Is there a correlation between total number of violations and overall inspection score?*

Most violations have a standard 1-point deduction in overall score. There are a few instances when more than a 1-point deduction can occur if a violation is extreme enough.

Inspection Score vs Violations Scatter Plot
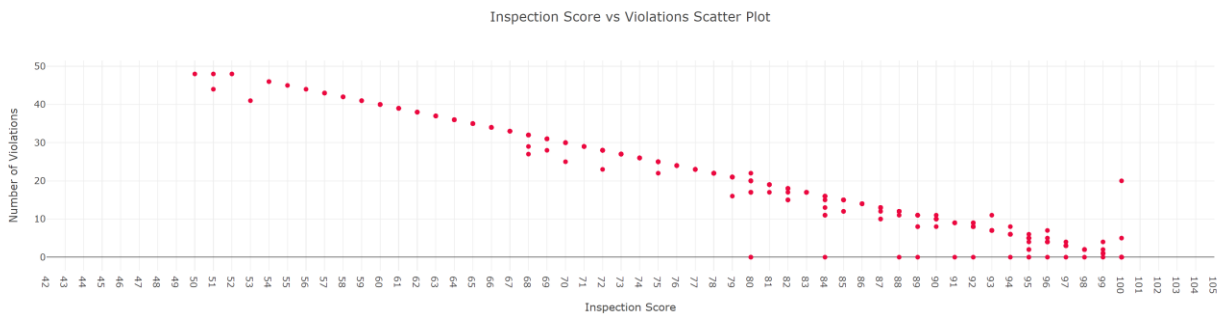
(iv) *Where are the best/worst places to eat in Dallas?*

The heatmap visualizes areas of Dallas where lower/higher inspection scores are concentrated. Red areas on the map denote locations that have a larger concentration of low inspection scores.

Last 5,000 Health Inspections in the year
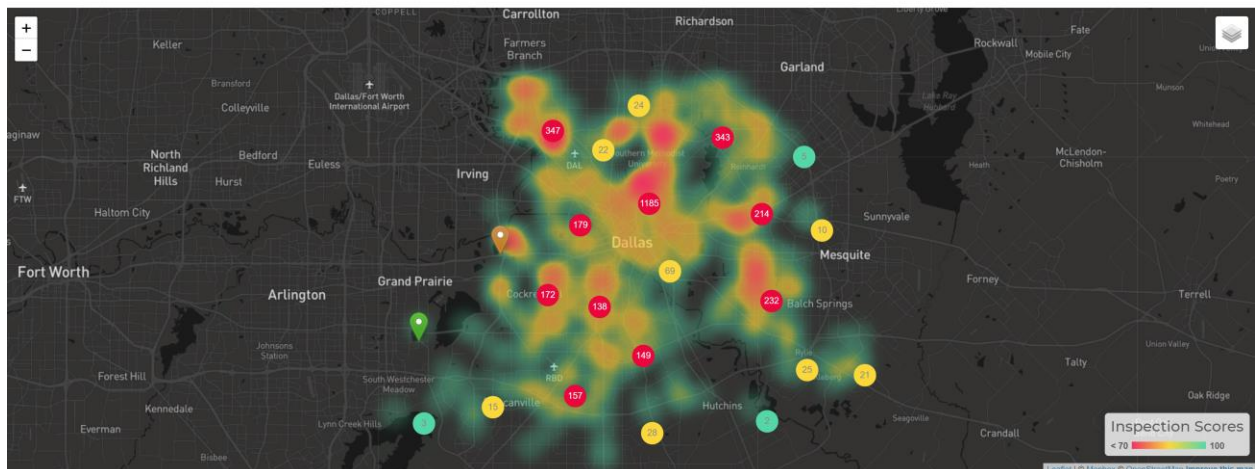
CONCLUSIONS

The food establishments at the intersection of Knox Street & IH-75 have had consistently lower health inspection scores compared to the greater Dallas area for the past 4-years.

The food establishments located to the Northwest of Dallas Love Field also tend to have lower inspection scores in the past 4-years.

Taking a sample size of 5,000 data points, on average most zip codes have an inspection score equal to or greater than 90.

When looking at the reason for the inspection, with a sample size of 5,000 data points, if the inspection occurred because of a "Complaint", they tend to have a higher average inspection score than a "Routine" or "Follow-Up" inspection.

A lower inspection score usually resulted in a higher score during the "Follow-Up" inspection.

Violation and inspection scores are inversely correlated. Meaning the more total violations a restaurant has, the lower the overall inspection score will be.

RECOMMENDATIONS

The recommendation based on the information found was to avoid restaurants specifically near Knox Street, since inspection scores were low around that area for the past four years. If a restaurant has an inspection score in the 70's, while that is a passing grade, the number of violations and reasons for those violations would deter customers from eating at those restaurants.

LIMITATIONS/FUTURE WORK

Some limitations that were encountered, was that the data was not as clean as initially perceived. It appeared that some inspectors when providing the data for the restaurant's inspection did not provide the correct the information (some zip codes were missing or had an extra digit), and that the information was not always uniform (chain restaurants would sometimes list the establishment number, and others would not).

For future work, adding a filter for each region of Dallas, to get a better visualization of where each zip code is located would provide more insight. Also, adding a count function for the bar graphs, to better see how many restaurants are located within that zip code.

WORKS CITED

1. "**Leaflet** — An Open-Source Javascript Library For Interactive Maps". *Leafletjs.Com*, 2021, https://leafletjs.com/. Accessed 08 Feb 2021.

2. "**Pointhi/Leaflet-Color-Markers**". *Github*, 2021, https://github.com/pointhi/leaflet-color-markers/blob/master/js/leaflet-color-markers.js. Accessed 15 Feb 2021.

3. "**Restaurant and Food Establishment Inspections (October 2016 to Present)**" https://www.dallasopendata.com/City-Services/Restaurant-and-Food-Establishment-Inspections-City/vcg4-5wum