Ellie Stevens and Ryan Kraft

Matthew Zook

GEO 310

13 December 2021

Lexington Restaurant Health Score Data Analysis

**Introduction**

Everyone has to eat. No matter what, no matter how, it is a necessary part of all life in the world. Having good, clean, and safe food is an extremely important part of everyday life all over the world. When food is unclean and prepared in an unsanitary manner, food borne illnesses can spread, cause people to become sick and can even lead to further complications. For this project, we decided to examine the inspection scores for food establishments in Lexington, KY to see if there were any areas of concern when it came to providing food that was safe to eat.

A health/inspection score is, in short, a score given to restaurants based on the cleanliness of the store, how food is stored, handled, cooked, and prepared for the consumer to eat. There is more that goes into this based on handling of specialized equipment, the presence of a pest or a pest infestation, and other various violations that can show up in a restaurant, kitchen, supermarket, or other food establishment. By examining the inspection scores, we get a better understanding of how clean the establishment is and if it is safe to eat or buy food from there or if we should avoid certain establishments and go somewhere else entirely.

For this research project we came up with four questions to help us think about the cleanliness of food establishments in Lexington. Research question #1: How does the opening date of a restaurant affect the health score of the restaurant? We chose this question because there are many restaurants and other establishments in Lexington that have been around for many years. Codes have changed over those years and as a result could cause discrepancies in what restaurants are set up to deal with and are prepared for. Research Question #2: Is there a difference in health score between fast food restaurants and non-fast food restaurants in Lexington? When you think of fast food versus non-fast food restaurants, there is a difference in the service, quality of food, and it is overall just a very different experience. Because there are so many differences between the two kinds of restaurants, it could be possible that they approach health and food safety precautions in a different manner. Research Question #3: How do Yelp ratings and number of reviews relate to the health score of the restaurant? For this question we wanted to bring in outside data because public perception of a restaurant can make or break the success of a restaurant. We also wanted to see if the public could tell if a restaurant was not up to cleanliness standards, because it would be interesting to know if that was something an everyday person would observe. Research Question #4: Does the location of restaurants in Lexington correlate to their health score? How does the income status of the area around the restaurant affect the restaurant's health score? For this question we wanted to know if there were geographical elements that came into play when it came to determining a food establishment's health score.

**Data and Methodology**

The data primarily being used for this study was taken from the Lexington-Fayette County Health Department website. On the website there is a downloadable excel file that contains the inspection scores of every food establishment in Lexington. For each food establishment, the excel file contains the establishment number, premise name, premise address, reporting area, inspection type, score, inspection date, violation list, whether or not there was a follow up required, the opening date, the R F Inspection, status, and EHS name. Not all of these columns of data were used, and many of the columns had to be aggregated in order to work with the data and answer the research questions.

One of the biggest issues with this data set was that if a restaurant had to have a follow up inspection of some sort, then there would be a separate row for each follow up inspection. This meant that one restaurant could have multiple lines of data, so the data needed to be sorted and aggregated in order to create one row of data for each restaurant (the most for one restaurant was four follow up inspections, making a total of five rows of data for one restaurant). Five columns were added to the sheet, and then the spreadsheet was sorted by establishment number so restaurants that had multiple inspections would have all of their inspections grouped together. The five columns were labelled Score1, Score2, Score3, Score4, and Score5, and then we went through for each restaurant and condensed them down to one row per restaurant by adding the inspection scores into columns Score1-Score5 in order of inspection date. For example, Joella’s score a 90 on their first inspection and a 97 on their follow up inspection, so Score1 was listed as 90, Score2 was listed as 97, and Score3-Score5 were left blank. The scores were then averaged to create an average score for each food establishment, and this was the inspection score used for a majority of the tests.

 The next aggregation done for the entire dataset was food establishment type (labeled as RestaurantType in the spreadsheet). The five classifications for food establishment type were 01\_FastFood, 02\_NonFastFood, 03\_Cafeteria, 04\_Private, 05\_Grocery/Convenience. In order to be classified as 01\_FastFood, the food needed to be mass produced, likely precooked and only needing assembly, the restaurant had to be a chain and it needed to have a drive thru. There were some exceptions made for the drive thru requirement because not all locations for a chain will have a drive through, but if the chain itself often has a drive thru then it could be considered fast food. For example, the Starbucks on Plaudit Place in Lexington does not have a drive thru, but it is still considered fast food because many of the other Starbucks locations in Lexington do have a drive thru. 02\_NonFastFood was classified as a restaurant available to the general public, with no membership or extra form of payment to attend the restaurant, it also would not meet the requirements to be classified as an 01\_FastFood, and the primary business has to be as a restaurant. 03\_Cafeteria is classified as a restaurant or food service that resides in a school, prison, place of work, or other where food is not made to order. Cafeterias may or may not be available to the general public. 04\_Private food establishments include catering businesses or locations that only provide catering, places that require a membership to attend, and places that are open to the public but require a fee to enter because the primary business is not food. 05\_Grocery/Convenience are gas stations or grocery stores where food can be bought, and the primary business is not being a restaurant.

Outside data was gathered from Yelp and was added to each restaurant that was classified as 01\_FastFood and 02\_NonFastFood. To gather the Yelp data, location was set to Lexington and we searched for each restaurant and recorded the Yelp rating and how many people left reviews. To assure accuracy with restaurants, and making sure all of the data was matched up correctly, the address in Yelp was compared to the address listed in the inspection score dataset to make sure they matched. Any restaurants that were not found in Yelp were labelled as MISSING in the dataset.

Research Question #1 Methods

 To answer this question, the data first had to be aggregated by how old the restaurant was. The Opening Date column in the original dataset contained Day-Month-Year, so we split the column into three separate columns for day, month, and year. For the purposes of this statistical analysis, the day and month column were removed, and we focussed on the years for aggregation. The sheet was then sorted by years so that the years could be aggregated into more broad groupings of years. For the aggregation, we used five year increments starting at restaurants that had been open for one year or more and added to a column labeled TotalOpenYears. The aggregated groups were <1 Year, 1-5 Years, 6-10 Years, 11-15 Years, 16-20 Years, 21-25 Years, 26-30 Years, and >30 Years.

 After the aggregation was completed, we calculated average score, median score, sample size (n), standard deviation, and the coefficient of variation. All of this data was compiled into a summary table to keep all of the data together. The hypothesis for this question was that the older restaurants would have lower average inspection scores than the newer restaurants, so we wanted all of the inspection scores for each restaurant to be gathered together to make comparisons. We ultimately decided to use the average score for each age of restaurant grouping because there were medians that were repeated between each grouping and it was not as clear that there was any difference at all in inspection scores between the groups. The average was then used to calculate the coefficient of variation because we wanted to be able to compare skewness of the groups. The only issue that we really encountered with this method is that there were three restaurants that were missing their opening date.

Research Question #2 Methods

 For this question we primarily worked with the data that was classified 01\_FastFood and 02\_NonFastFood. In the summary table we collected the average, median, sample size (n), standard deviation, and coefficient of variation for all food establishment types so the research question could be expanded if need be. In addition to the general summary table, we used the two sample difference of proportions test to compare the proportion of 01\_FastFood and 02\_NonFastFood restaurants that had an inspection score less than 96 to the proportion that had an inspection score greater than or equal to 96. 96 was chosen as the midpoint being used for this comparison of data because the average score for all 01\_FastFood and 02\_NonFastFood was 95.85, which rounded would be 96. While this may not be the most accurate midpoint to use, we felt that it would best represent the comparison of 01\_FastFood and 02\_NonFastFood. This test could be done again with 90 being the midpoint for the test because anything below a 90 would be considered a B rating for inspection scores. In order to most easily determine what scores were above and below a 96, the spreadsheet was sorted by the AverageScore column and everything less than 96 was labelled as <96Score and everything greater than or equal to 96 was labelled as >96Score in a separate column titled AggregateScore.

 One problem with using the two sample difference of proportion tests was that we did have to decide what the midpoint of the data would be. Another issue that we ran into was only being able to compare two types of food establishments at a time. There are five different kinds of food establishments that we determined in this data set and only two of them can be compared to one another at a time using this method, so that is why we also collected the averages and standard deviations as well.

Research Question #3 Methods

 As discussed earlier, Yelp reviews were pulled directly from Yelp for 01\_FastFood and 02\_NonFastFood because they were not in the original dataset. The data was only gathered for 01\_FastFood and 02\_NonFastFood because they were the primary focus of research question #2 and were the most readily available to find on yelp. For 03\_Cafeteria and 04\_Private, most of those were not on Yelp, and for 05\_Grocery/Convenience the rating applied more to the store than to the food at the store from observations of the review comments. Also, the data for 03\_Cafeteria, 04\_Private, and 05\_Grocery/Convenience was removed from the dataset for this question in order to prevent any issues with calculations.

 To answer the research question, we created a summary table with three sections: one section for 01\_FastFood and 02\_NonFastFood combined, one section for just 01\_FastFood and one for 02\_NonFastFood. The summary table was organized by the rows being Yelp score 1-5, and then the columns being Yelp Count (how many restaurants had that rating), PercentofCount (what percentage the Yelp Count made up of the sample size), and Average Score (the average score for all restaurants with that Yelp rating score). This allowed for the comparison of how many restaurants fell into each Yelp rating score and what the average inspection score was for each yelp rating score. This biggest issue with this data collection was making a summary table that was easily readable. When the calculations were first input they were not easily readable.

 The other calculations done for this research question was calculating R and R2 to determine correlation between Yelp rating and inspection score as determined by the Health Department. To find R and R2 for the data, three different scatter plots were created in google sheets, and from those scatter plot charts the R and R2 were included in the formatting of the chart. The three charts created were for 01\_FastFood and 02\_NonFastFood combined, 01\_FastFood, and 02\_NonFastFood, Yelp rating vs. inspection score. The biggest issue with these calculations is that most of the data fall between 90-100 for the inspection score, but that will be gone into greater detail in the findings section.

Research Question #4 Methods

Does the location of restaurants in Lexington correlate to their health score? How does the income status of the area around the restaurant affect the restaurant's health score? This question was answered by creating a map of the Lexington area. By doing so, it helped create a picture of the city of lexington to see where the lowest scores in Lexington would be located. To do so, we needed to collect some basic information and shapefiles. The basic information we used was the median income status of each household in the Lexington area by census tract as a base layer for the data. This was done by collecting some already made shapefiles in 2018 by the US census department and then joining the data and shapefiles together. After all data was collected it was then put into QGIS version 3.16. Since the addresses of the restaurants were not previously geocoded and were only given the addresses, we were able to do some geocoding in QGIS by the plugin MMQGIS. This plugin is limited however to only 2,500 locations at a time to geocode and it is a very slow process to geocode. Luckily, we only had about 1,000 or so points in the dataset but it did take a few minutes to geocode. For the median household income and the census tract we had to form a join between the two, doing so created a better visual of the data and were able to create a choropleth map from the median income values and the average health scores for each restaurant. With the highest income being a darker green and the lowest income being barely green or close to white in color. This also goes for the health scores with them being the darker the color, the higher in score they are and the lighter they are, they in turn have a lower score. However, this did not work on first try since the data was in a text form and to create it into a choropleth it has to be a string/integers. Therefore, we had to refactor the attributes table in QGIS from the processing toolbox to make the categories we needed (Health score and Median Income) into integers rather than text.

**Findings**

Findings for Question #1

*Hypothesis: Older restaurants will have a lower inspection score because of the age and greater likelihood of the building being out of date and not up to code.*

****

**Chart 1**

The average score for each grouping of years showed that there was not a consistent or significant decrease in inspection score as a restaurant increased in age. Only 5.1% of the food establishments in the entire dataset had an inspection score that was less than 90, and only 22% of the establishments had an inspection score less than 95. There was no evidence to prove the hypothesis correct. As seen in Chart 1, there was a slight decrease in average inspection score from establishments open <1 Year to establishments open 11-15 years, but in the next grouping, which is 16-20 years, the average increased by just over a full point, which interrupted the trend.

 For this question we also calculated the coefficient of variation for each grouping to see if any of the groupings of years had a greater dispersion of the data. Almost all of the coefficients of variation were between 0.032-0.037, with the outliers with greater dispersion being 6-10 years (0.044) and 21-25 years (0.04). All of these numbers show that the dispersion is relatively consistent throughout all of the groups and the averages show similar information.

Findings for Question #2

*Hypothesis: Non-fast food restaurants will have a lower health score compared to fast food restaurants.*

We were able to determine from conducting a Two Sample Difference of Proportions Test that non-Fast food restaurants did in fact have a lower health score compared to fast food. With fast food having a proportion of 0.302 (51 out of 169) of restaurants being lower than a health score of 96, and a proportion of 0.698 (118 out of 169) being above a health score of 96. While non- Fast food restaurants had a proportion of 0.406 (182 out of 448) of restaurants being lower than a health score 96, and a proportion of 0.594 (266 out of 448) of restaurants being higher than a health score of 96. This in turn gave us a 0.8% chance of making a type one error in our dataset, meaning we can reject the null hypothesis of it having no relation whatsoever and us accepting our original hypothesis of Non- fast food restaurants having a lower health score compared to fast food restaurants.

Findings for Question #3

*Hypothesis: Lower Yelp score will correlate with lower inspection scores for fast food and non fast food restaurants.*

 

**Chart 2**

 For this question, we looked at R and R2 to determine the strength of the correlation that exists between inspection scores and Yelp ratings. Chart 2 shows a scatter plot of Inspection scores (labeled as Health Score in the chart) versus Yelp ratings for 01\_FastFood restaurants and 02\_NonFastFood restaurants combined. R for this correlation is 0.55 and R2 is 0.003, which leads us to the conclusion that there is not a strong correlation between Yelp rating and inspection scores for 01\_FastFood restaurants and 02\_NonFastFood restaurants. We repeated this process for 01\_FastFood restaurants and 02\_NonFastFood restaurants to see if individually they would have stronger correlations to their Yelp scores.

****

**Chart 3**

Chart 3 shows the correlation between inspection scores and Yelp ratings for 01\_FastFood restaurants only. The R for this relationship is 0.045 and R2 is 0.002. This shows an even lower correlation than shown in Chart 2.

****

**Chart 4**

 Chart 4 shows the correlation between inspection scores and Yelp ratings for 02\_NonFastFood restaurants only. The R for this relationship is 0.13 and R2 is 0.017. This is the strongest correlation of the three different correlation comparisons, but it still would not be considered to be a strong correlation because R is so low.

 The main reason the correlations are so low for all of these relationships between inspection scores of restaurants and their respective Yelp ratings is because of the clustering of data between 90 to 100. As mentioned in the findings section for question #2, there is a large majority of the inspection scores that are greater than 90. For just 01\_FastFood restaurants and 02\_NonFastFood restaurants, 6.8% have an inspection score below 90 and 28.9% have and inspection score below 95. This leaves the data clustered making little room for there to be any correlations between the Yelp rating and inspection scores.

Findings for Question #4

*Hypothesis: Areas of lower income will have lower average health scores overall compared to areas with higher income.*

By creating the map and producing it in a choropleth style, we were able to determine it is mostly vice versa from our original hypothesis. In the downtown area of Lexington we can see that those census tracts have the highest scores of income while the outer areas of Lexington tended to have lower scores. This could be because the downtown area is mostly business oriented and finer dining, with the people living in these areas mostly renting and therefore don’t have a high income. With the outer census tracts being more suburban neighborhoods like Tates Creek, Hamburg, and then the horse farms being the richest areas the farther out you go. However in the southwest region of the city, it tends to show an even spread of low scoring restaurants to high scoring restaurants. These areas of Lexington are also heavy retail areas of the city with Nicholasville road in particular being heavily busy with consumers since the Fayette mall is right there and also other outlet malls and supermarkets close by to it. With all of these factors we were able to determine that our hypothesis was not necessarily correct with the lower household income of the city being where the lowest health scores would be.



**Map 1**

**Conclusions**

From our research we were able to determine quite a few things. Overall, Lexington is an extremely clean city when it comes to food. Most of the restaurants did in fact have a high score between 90 and 100, with an exception of a few outliers in the data. We do not believe that if the restaurant had an incredibly low score they would not be open and therefore not available to the public to consume. The city also does a considerably good job on keeping their information up-to-date and available to the public, with the most recent data being from September of 2021. We were not able to conclude that our hypotheses were fully correct but we were able to give a better look at the city and how health scores are determined and affected within Lexington.

With question #1 showing that there was no real evidence to prove our hypothesis to be true. Question #2 showing that proving that we were correct in our hunch that non-fast food restaurants had a lower health score compared to regular fast food, this in turn showing that fast food restaurants have a lot more of a rigorous cleaning, handling, storaging, and cooking guidelines so their food is acceptable by not only the city but by also the company’s standards. Question #3 showed that there was no real correlation between yelp scores and health scores since most of the restaurants had high scores in the 90’s giving very little variance in the results since they were all clustered in one spot. Lastly, with question #4 showing us that there is not a full relation between location of a restaurant and it’s health score since most of the higher ranking restaurants were located in the downtown area, a much lower income area compared to the rest of the city.

If we were able to do this research again, there are a few things we would have done differently. One thing we would have done differently is we would’ve looked more at the map’s statistical table and tried to see if higher income areas had higher health scores but that would have involved a lot more data cleaning and filing to make sure everything is organized in a more specific way. This could be done by creating a pivot table and then graphing the data out to see if there were any correlations between income and inspection scores. Another part of the research we may have done differently would be looking into the specific violations that restaurants received and if the violations changed or remained the same for restaurants that had to have a follow up inspection. There were some restaurants that appeared to have multiple inspections within the same day because of a violation that could be quickly resolved and bring their inspection score, so it would be interesting to look at those specific violations and how common and fixable they are. The last change we would have made would have been to compare multiple years worth of data to see if there was any sort of shift in inspections score because of the Covid-19 pandemic or if there were any standards that changed and caused inspection scores to change.

**Works Cited**

Bureau, US Census. “Cartographic Boundary Files - Shapefile.” *Census.gov*, 8 Oct. 2021, https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.

*Explore Census Data*, https://data.census.gov/cedsci/table?q=United+States&t=Income+%28Households%2C+Families%2C+Individuals%29&g=0500000US21067%241400000&y=2019&d=ACS+5-Year+Estimates+Detailed+Tables&tid=ACSDT5Y2019.B19013&tp=true.

“Food Protection and Restaurant Inspections.” *Lexington-Fayette County Health Department*, published September 15, 2021, lexingtonhealthdepartment.org/food-protection/. Downloaded October 31, 2021.

*Yelp.* Yelp, 2004, yelp.com. Accessed November-December 2021.

From the Class Readings:

“Explanatory Statistics” and “Descriptive Statistics” by Visser and Jones.

“Correlation”

“Two-Sample Tests”