OSM CMPT 353 Project - Examining Vancouver's Local Areas and Restaurant Ratings

Vincent Fang (vgfang), Yamato Nakahara (ynakahar)

August 14, 2021

Problem

Vancouver, BC has many restaurants throughout the city and there is a significant amount of data entries in the given in the provided `amenities-vancouver.json.gz` file. We were initially inspired to do analysis based on the example problem examining if there were parts of the city with more chain restaurants than non-chain restaurants.

We refined this example problem by limiting the results to the City of Vancouver's 22 local areas and analyzing the results within those areas, rather than areas in Greater Vancouver. We chose this because information on the area boundaries were provided on the City of Vancouver's website. Other refinements included expanding the scope of the problem and doing analysis on the user ratings of restaurants provided on Google Maps.

Using data analysis tools, we wanted to determine if certain local areas in the city had a higher proportion of well-rated restaurants. We also wanted to examine if there was a significant difference in average ratings between franchises and independently owned restaurants. Lastly, we wanted to investigate if there was a meaningful correlation between the proportion of franchise restaurants in an area and the average ratings in that area.

Data Refinement and Gathering

The data set we initially started with was the `amenities-vancouver.json.gz` [1] file given to us on the project page. As there may be possible differences in the answers to our questions depending on the city, we have chosen to only use the data for restaurants within Vancouver instead of analyzing the entirety of Greater Vancouver.

To filter the data to only restaurants within Vancouver, we have used a data set that the City of Vancouver which provides the geographic boundaries of Vancouver's 22 local areas in the form of a JSON file. These local areas are neighborhoods such as "Point Grey" or "Killarney." The JSON file, `local-area-boundary.json` [2], uses lists of latitude and longitude values to form polygons of geographic points to represent the boundaries. Locations that did not fall into any of the 22 local areas were excluded from the data set to limit the locations to the City of Vancouver.

Using the `shapely` library in Python, the lists were converted into polygon objects in which they can be easily checked if a point was within its boundary. The latitude and longitude values in the `vandata` data set were used to determine the name of local area each location belonged to.

To limit our data to restaurants only, we used the *amenity* column in the data set and filtered it so that our data set would only contain amenities in Vancouver that were attributed as: cafe, pub, restaurant, fast_food, bar.

To get the ratings data for each location, we used the Google Places API [3]. The data needed for this analysis is the "Atmosphere Data" which contains both the Google Maps user rating and the number of user ratings for a location. The ratings are in the format $n/5$, where $1 \leq n \leq 5$. The API can take inputs for location name, latitude and longitude and returns a JSON object with the rating data.

In our analysis, we will be comparing two groups of restaurants: franchises and non-franchises. Restaurants were considered to be franchises if they had the 'wikidata' tag in the original dataset. This makes sense as most franchises will be large enough to have their own Wikipedia page. To avoid errors in classification, restaurants with 5 or more locations were also considered franchises, just in case they lacked the tag.

Methods

For all of our T-tests, an alpha of 0.05 was chosen, meaning the null hypothesis H_0 will be rejected if $p < 0.05$.

ANOVA Analysis

The first question we decided to look at in our analysis was whether or not the average rating for the stores in each area were different in each area. We have decided to rely on the Central Limit Theorem in our assumptions that the ratings are somewhat normal. Then we used the function `f_oneway` from the `scipy.stats` package to perform the initial ANOVA analysis across all of the areas. We repeated this analysis twice to see if there was any impact difference between the ratings of all stores combined, and the ratings of all of the non-franchise stores. We also decided to perform a separate analysis where areas with less than 25 total ratings were excluded from the analysis to account for the lower limit of the Central Limit Theorem.

Heat maps

As the number of stores in each area is an important part of our analysis, we decided to use heat maps to visualize it. In this analysis we decided to use the two groups of only franchises and only non-franchises so that it is easy to see the distribution of the two groups across the city. To accomplish this we used a jupyter notebook along with the package `gmaps`to create the heat maps. Along with refined data set we have been using, we also used the `amenities-vancouver.json.gz` [1] data set to plot out the boundaries for each area on the data set. The jupyter notebook has following heat maps: the heat map for only franchises, the heat map including both franchises and non-franchises without weights on the franchises and the heat map including both groups with weights on the franchises. As our cut-off for considering a restaurant to be a franchise is somewhat arbitrary, the weights on the franchises will show the location of major franchises like Starbucks more clearly on the heat map.

Mean Ratings Calculation and Mann-Whitney U Test

To determine the mean ratings for franchises and non-franchises, we used pandas to group the data and aggregate the average ratings for each group.

To determine if the difference between the two group's ratings was significant, we used two-tailed Mann-Whiteney U test from `stats.mannwhitneyu`. It is a non-parametric test that can be used to decide if samples from one group tends to be larger than another. It is appropriate for determining franchises ratings tend to be lower or higher.

Correlation/Linear Regression

To calculate the correlation between franchise proportion and average rating in an area, we used a linear regression test using `stats.linregress`. We assumed a correlation would most likely be linear based on the scatterplot we produced. The franchise restaurants proportion is calculated by finding the quotient of the number of franchise restaurants in the area and the total number of restaurants in the area. For the regression, we decided to remove local areas that had fewer than 25 restaurants because they can be inaccurate when it comes to the franchise restaurants proportion value.

Results

The results of the ANOVA analysis that examined if the average rating of the restaurants were different in each area gave the following results along with the question the analysis is answering below. This results show that the different areas do have significant different average ratings.
"Do different areas in Vancouver have different average ratings?" p-value: $6.19e-16$
"Do different areas in Vancouver have different average ratings when franchises are not included?" p-value: $1.77e-18$
"Do different areas in Vancouver have different average ratings when areas with less than 25 total ratings are filtered out?" p-value: $9.51e-15$
"Do different areas in Vancouver have different average ratings when franchises are not included and areas with less than 25 total ratings are filtered out?" p-value: $1.69e-13$

The heat map that separates the franchises and non-franchises without weights on the franchises is the following in Fig.1:

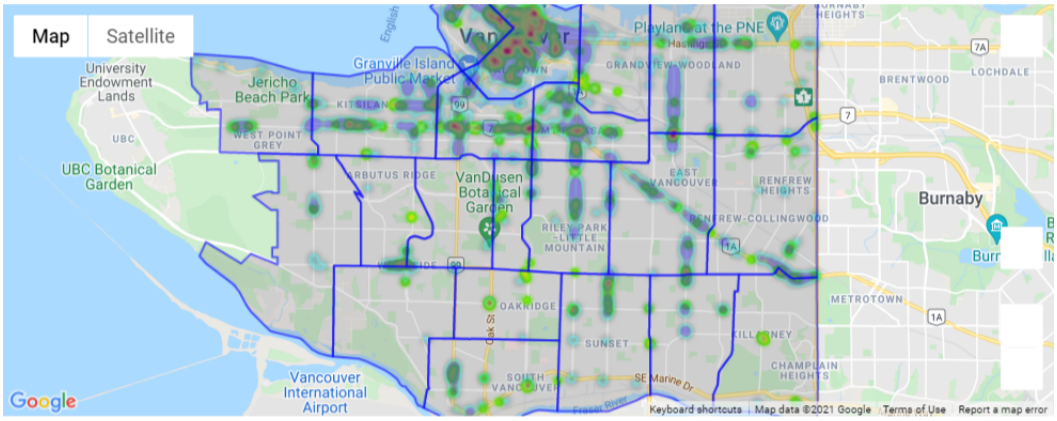

Figure 1: Heat map of the concentration of franchises and non-franchises throughout Vancouver. Franchises are measured from low to high concentration with the range of colors from light green to red. Non-franchises are measured from low to high concentration with the range of colors from aqua to blue. The franchises in this heat map are not weighted.

For determining the difference between franchise and non-franchise ratings, the mean rating for the 339 Franchise restaurants was 3.841593 and the mean rating of the 1683 non-Franchise restaurants was 4.210873. From our Mann-Whitney U analysis we found that the p-value for our test was $3.1241840082221806e-68$ suggesting that the means are meaningfully different.

The linear regression produced a p-value of 0.002379887506451116 which suggests that the slope is non-zero. The slope value was found to be -1.35223980692476 which is a negative correlation. The linear regression formula was $rating = 4.34 - 1.35x$, where x is the franchise restaurants proportion in that area.

For the plot, a colormap was included on the right to show the number of ratings for each point.

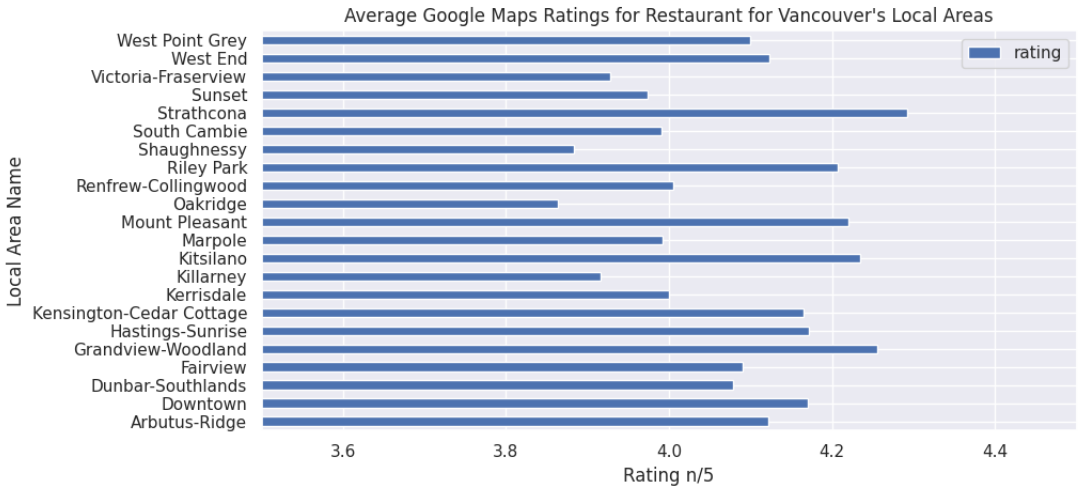

Figure 2: Bar graphs of the average google maps ratings for restaurants in each area in Vancouver.

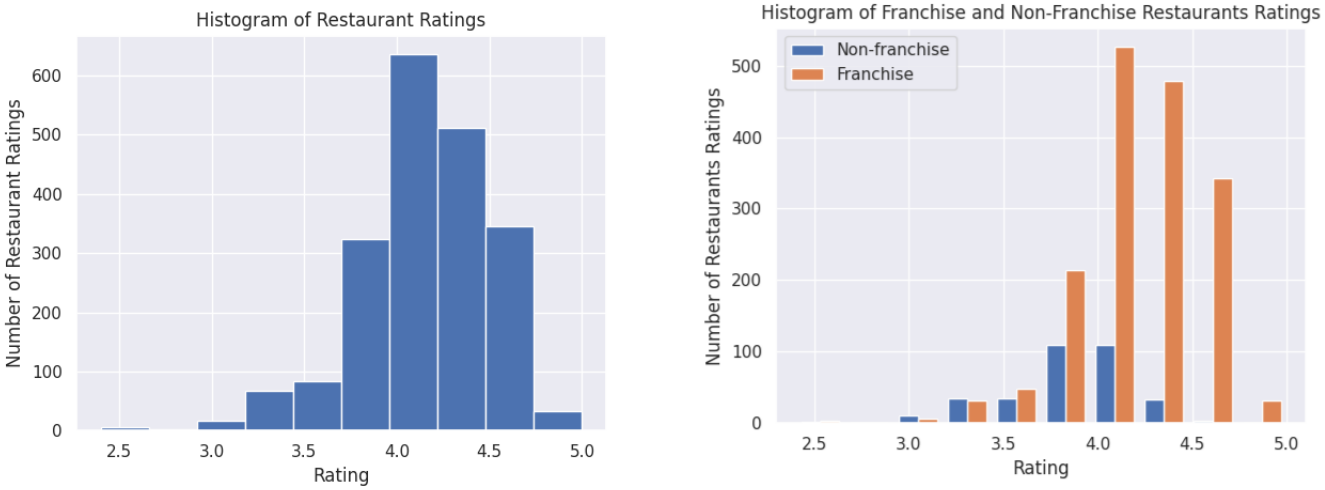

Figure 3: Left: Histogram containing all of the ratings for the restaurants in the area. Right: Histogram for the ratings for franchises and non-franchises.

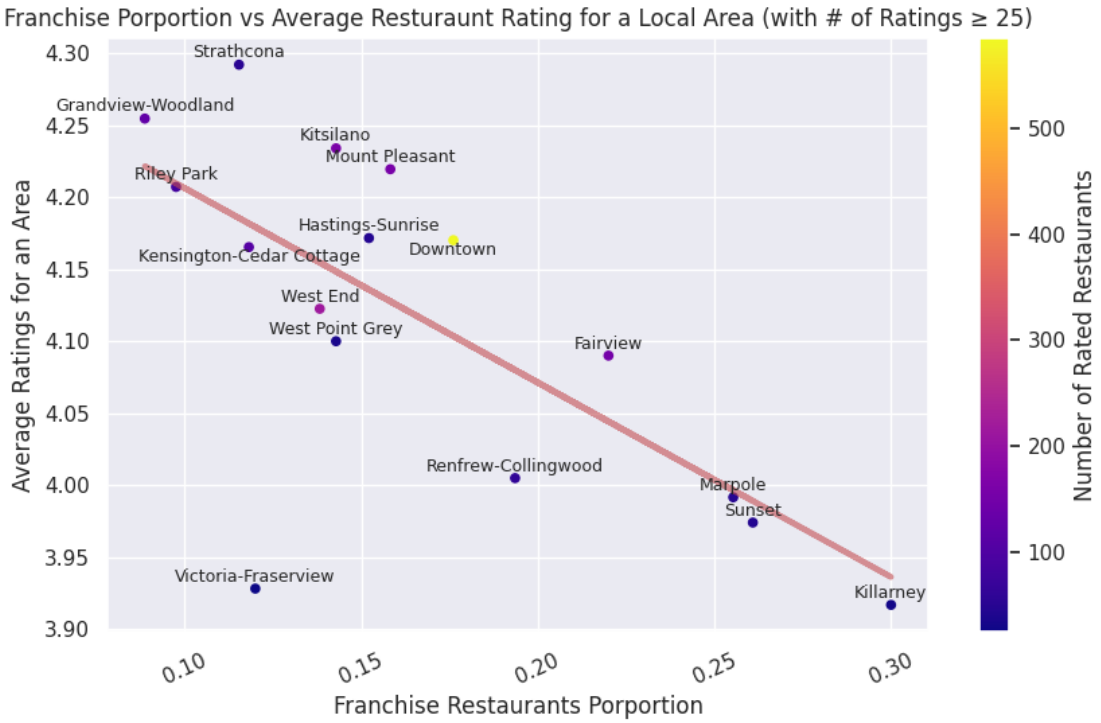

Figure 4: Plot containing the average rating per area against the proportion of franchises in the area. The dots are the actual data points while the line is the fitted linear regression line. The number of rated restaurants in each area can be seen by the color which is coded by the gradient on the right.

Discussion

Looking at at the results of our analysis, it is clear to see that many of our questions do have a defined answer. Regarding the ANOVA analyses, we have clear p-values that reject the null hypothesis in each of the questions asked and effectively answer yes to the four questions posed from those results.

While the heat maps may be unlabelled, it is easy to see when inspecting the heat map at how some areas can easily be considered outliers in the number of restaurants in the area or the total number of ratings for the restaurants in the area. It can clearly be seen that every area that goes further from Downtown, which is the area with the high concentration of restaurants in the north, gradually decreases in the number of restaurants present in the area.

According to our Mann-Whitney U analysis, non-franchise restaurants generally have a higher rating than franchise restaurants. This is not surprising to us as many franchise restaurants specialize in fast food. The difference appears small with the means being 4.21 and 3.84, but the the looking at the histograms of restaurant ratings in Fig.3, the difference is is significant. Almost all ratings are in between 3 and 5, so a 0.37 difference is significant.

For the correlation between franchise proportion in an area and the average rating, we found a negative slope. This suggests that with more franchise restaurants relative to total restaurants, the average restaurant rating in that area decreases. This is consistent with our previous finding that suggested that franchise restaurants have a significantly lower average rating.

Limitations and Problems

The foremost problem in this analysis is the dataset. Certain local areas like "Oakridge" have less than 25 restaurant entries whereas the area "Downtown" has almost 500. The number of rated restaurants are very different between areas, but in the analysis they are treated the same. This could be because `amenities-vancouver.json.gz` lacks locationsor that certain areas are smaller and have fewer restaurants.

For determining which restaurants are franchises, the method we used may allow for some outliers, as there are likely restaurants that could be considered a franchise that do not have Wikipedia data and have fewer than 5 restaurants in Vancouver.

For the Google Places API [3], there is a limitation when it comes to the cost of using the API. For retrieving the ratings data, the price for a monthly volume range below $100,000$ is $22.00 per 1000 API calls. If this process was called every time the analysis program was tested during development with approximately 2000 data points, the costs would be significant. This limitation meant that retrieving the ratings data had to be triggered manually in the data pipeline to mitigate costs. A separate file python file was used to read the data set from a `.csv` file and and write a modified `.csv` file. The total cost ended up to be $250.00, though it was paid for through credits from Google's free trial.

Additionally, the Google Places API calls take a considerable amount of time, with 2000 calls taking 18 minutes. This is not a problem for the analysis stage, as the `.csv` file will be already produced.

Conclusion

From our data analysis, we came to the following conclusions:

1. There are meaningful differences in average restaurant ratings in Vancouver's local areas.
2. Non-franchise restaurants are generally rated higher than franchise restaurants in Vancouver.
3. There is a slight negative correlation with the relative number of Franchise restaurants and average restaurant ratings in Vancouver's local areas.

If we were to do this project again, we would be more observant of the limitations of the data set early on, such as the limited number of restaurants in certain areas. Also we spent a lot of time getting the API working and making sure we didn't overspend. In the future, we would probably be more vary of the costs of using various APIs and their runtime overhead.

Project Experience Summary

Vincent Fang, (vgfang)

1. Constructed an appropriate list of related topic questions and data science methods to get the project focused on a concrete goal.
2. Researched and applied the cumbersome Google Places API to find restaurant ratings to facilitate an interesting data analysis project.
3. Described the problem, procedure, and results of the data science analysis to produce a cohesive and meaningful report.

Yamato Nakahara (ynakahar)

1. Applied the gmaps package in a jupyter notebook to use the Google Maps API to plot heat maps and the polygons of each area onto the same figure.
2. Conducted ANOVA Analysis using Python on all the regions in Vancouver to obtain p-values that rejected the null hypothesis that the mean of the measured values were the same.
3. Used LaTeX to write a report using our results and figures to answer the questions in the report.

References

[1] Baker, G. (n.d.). *Project Topic: OSM, Photos, and Tours.* https://coursys.sfu.ca/2021su-cmpt-353-d1/pages/ProjectTour

[2] City of Vancouver. (n.d.). *Local area boundary.* https://opendata.vancouver.ca/explore/dataset/local-area-boundary/information/

[3] Google LLC. (n.d.). *Places API Usage and Billing* https://developers.google.com/maps/documentation/places/web-service/usage-and-billing