Vancouver Neighbourhood Recommender - Final Project Report

CMPT 353 - D100: Computational Data Science

Professor Gregory Baker

Summer 2021

Vancouver Neighbourhood Group

Members:

James Braun - jrb16 - 301300957

Jonghyeok Kim - jka236 - 301418058

Myckland Matthew - mma221 - 301416413

Problems Addressed:

According to Canada's 2016 Census, over 40% of people in Vancouver are foreign-born. If you include people born in Canada but not Vancouver, then a majority of Vancouver residents are from somewhere else. When people relocate to Vancouver, they all face the same question: which neighborhood is the best to live in? This question is difficult to answer because it depends on one's situation and preferences. For example, young people might wish for entertainment venues nearby, whereas new families may want to live in a safe and quiet neighbourhood.

For this project, we developed a model in Python using various libraries that suggests the best neighborhoods to live in for people planning to relocate to Vancouver. The model scores each of Vancouver's 22 neighbourhoods according to 10 criteria we believe are important to potential home buyers. They include: proximity to Skytrain, home price, safety, and local restaurant quality. The final 6 criteria are the amount of amenities nearby for 6 different categories: artistic or cultural landmarks, automotive services, educational or childcare facilities, places to eat or drink, healthcare clinics, and financial institutions. The model then collects inputs from the user regarding their preferences and combines it with the calculated scores to produce the top 5 recommended neighbourhoods to live in. If the user would like to look at the data themselves, we also provided a visualization map written in html of all the amenity locations and scores for the user to peruse and modify to their own liking.

In addition, for this project we used functions from the SciPy and statsmodels libraries to statistically analyze the neighbourhoods based on all 10 criteria to determine if any neighbourhoods are unique, even after normalizing certain criteria for population differences. For example, we answered the question: Do the mean monthly number of violent crimes per 10,000 residents differ by neighbourhood? Another question we answered was: Does the proportion of healthcare clinics in Kitsilano differ from its proportion of Vancouver's population?

Data Used:

First, we needed the geographic location of each neighbourhood. The neighbourhood names were recorded manually according to Wikipedia's article on Vancouver's 22 neighbourhoods.[1] We used Google Maps' Geocoding API[2] to collect a single latitude/longitude point to represent each neighbourhood and used these points in some analyses. The API search result was not just a single point; it required extra work to extract the latitude and longitude.

We also used the Geocoding API to get the precise location of every Skytrain station in Metro Vancouver. The names of the Skytrain stations were scraped from Wikipedia[3] with a script we made, and we also had to use regular expressions to clean up extraneous punctuation marks.

[1] en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver
[2] developers.google.com/maps/documentation/geocoding/overview
[3] en.wikipedia.org/wiki/List_of_Vancouver_SkyTrain_stations

To gather the home price data, we manually recorded average sale price per square foot figures from redfin.ca for each neighbourhood since the site prohibited automatic scraping.

The population data came in CSV format from Abundant Housing Vancouver,[4] who recorded past census data, but only until 2016. So, we used the average yearly population change rates from 2011 to 2016 to extrapolate until 2021. See **Appendix** for an example calculation.

The amenities data came in JSON format courtesy of Professor Baker who selected data from the OpenStreetMap (OSM) project. However, it required quite a bit of work in Pandas to transform it into a more appropriate form for our data analysis. First, we selected a wide variety of amenity types and split them into six distinct categories that we thought were important to potential home buyers. The six categories were:

1. Education or childcare facilities like colleges or kindergartens.
2. Financial institutions such as banks and currency exchanges.
3. Artistic or culturally-significant landmarks like places of worship, theatres, and community centres.
4. Healthcare clinics such as hospitals, dentist offices, and pharmacies.
5. Places to buy food or drinks like bars, restaurants, and vending machines.
6. Automotive service locations such as car washes and gas stations.

Following that, we used the latitude and longitude of each amenity combined with geometric polygon data of Vancouver's neighbourhood boundaries to determine which neighbourhood each amenity belonged to. This was accomplished using the Shapely package. We then tabulated how many amenities of each category were located in each neighbourhood. The neighbourhood boundary data came from the Open Data Portal on the City of Vancouver's website.[5]

The crime data originated from the Vancouver Police Department's website[6] in a CSV file that included the date, neighbourhood, and type of all reported crimes in Vancouver from 2003 to 2021. To analyze how safe each neighbourhood is, we used data on violent crimes (homicides, offences against a person, or residential break and enters) from 2016-2020. Since every neighbourhood differed in population, we decided to calculate and compare crime rates per 10,000 residents using the extrapolated population values.

To get data on restaurant quality, we used Yelp's Fusion API[7] to collect the Yelp rating for each of the top 50 most reviewed restaurants in each neighbourhood, so each neighbourhood had 50 ratings, one for each restaurant, and each rating ranged from 1 to 5 in 0.5 increments (e.g. 2.0 or 3.5, but not 2.37). The API wouldn't allow us to use the neighbourhood boundaries we used for the amenity counts, so instead for each neighbourhood we searched for restaurants within 2km of its single point representation that was returned from Google Maps' Geocoding API.

[4] www.abundanthousingvancouver.com/research

[5] opendata.vancouver.ca/explore/dataset/local-area-boundary/information/

[6] geodash.vpd.ca/opendata/

[7] www.yelp.com/fusion

Analysis Techniques:

Note: all distance calculations were performed using the haversine formula[8] *to get the distance between two latitude/longitude points in metres.*

To start our neighbourhood analysis, we decided to score each neighbourhood based on distance to Skytrain. To accomplish this, for each neighbourhood we used a single point to represent it and found the distance to the closest Skytrain station. This gave us a sample of 22 distances, and this sample had a mean and a standard deviation. If a neighbourhood's Skytrain distance was within a half standard deviation of the mean, then it got a score of 3 out of 5. If it was between a half and one standard deviation, it got either a 2 or a 4 out of 5, depending on if it was below (4) or above (2) the mean. If it was further than one standard deviation, it got a 1 or 5.

We used the exact same mean/standard deviation method to score neighbourhoods based on average sale price per square foot, where lower prices resulted in higher scores.

Before we scored based on amenities, we decided to analyze if the neighbourhood affected the number of amenities for our six recorded amenity categories, even after accounting for population differences. Our null hypotheses were that neighbourhood had no impact; all amenity counts matched expectations given the differing populations. The alternative hypotheses were that the ratio of amenity counts for each neighbourhood differed from what was expected (i.e. neighbourhood had an impact). We decided to use one-way chi-square tests to see if our results were statistically significant. This required calculating an expectation on the number of amenities for each neighbourhood and amenity category combination. See **Appendix** for an example calculation, but in short it depends on each neighbourhood's relative population. Before we started any analyses, we decided to use the standard p-value of 0.05 to test for significance. Also, since we performed 6 tests, we decided to use a Bonferroni correction and so in actuality our threshold was $p < 0.05/6 = 0.0083$ to reduce the Type I error rate.

Following the chi-squared tests, for each amenity category we looked at every neighbourhood and tried to determine which neighbourhoods had a statistically significant different amenity count from what was expected given their relative populations. We accomplished this using two-sided binomial tests to determine the probability under the null hypothesis that the given neighbourhood's amenity count was as extreme or was more extreme as it was. For each test, the null hypothesis was that the proportion of amenities within a neighbourhood was equal to that neighbourhood's proportion of Vancouver's total population. The alternative hypothesis was that these two proportions were not equal. Also for each binomial test, the number of successes was the neighbourhood's amenity count for that category, the number of trials was the total amenity count for that category across all of Vancouver, and the probability of success was the relative population of the given neighbourhood. See **Appendix** for an example binomial test.

[8] en.wikipedia.org/wiki/Haversine_formula

These binomial tests gave 132 p-values which we used to calculate the amenity scores. If $p < 0.05/22 = 0.00227$, then this showed significance even after a Bonferroni correction (22 = # of neighbourhoods = # of binomial tests for each amenity category) and the neighbourhood got a 1 or 5 (depending on if the actual amenity count is less than (1) or greater than (5) the expected count) for that particular amenity category because there was overwhelming evidence that the neighbourhood's amenity count differed from its expectation. If $0.00227 \leq p < 0.05$, then the neighbourhood got a 2 or 4 since that showed a moderate amount of evidence that the amenity count differed from what was expected. If $p \geq 0.05$, then the neighbourhood got a 3 because then there's not enough evidence to say the count was different. See **Appendix** for a scoring example.

To analyze the crime data, we used monthly violent crime rates per 10,000 residents from 2016 to 2020. Before performing an ANOVA test to see if neighbourhoods differed in safety, we created a histogram for each to see if monthly crime rates were normal-enough to proceed. Figure 2a) from **Visualizations** shows that some of the histograms were right-skewed. Therefore, we took the square roots of the monthly crime rates and made another batch of histograms (see Figure 2b). The square-rooted data was much more normal looking, so we proceeded with ANOVA. We decided to forgo any normality or equal variance tests since we had 60 pretty normal-looking data points per neighbourhood, so by the Central Limit Theorem this was safe to do. For the ANOVA test, we again used $\alpha = 0.05$ to determine if the means of the square roots of the monthly crime rates between any two neighbourhoods differed. Then we performed Tukey's Honest Significant Difference (HSD) test again on the transformed data to determine which neighbourhoods had different means. Following that, we generated the safety scores. To calculate them, we created a min-max scaler and gave it each neighbourhood's mean monthly violent crime rate (not square-rooted). The scaler calculated a float from 0.00 to 5.00 for each neighbourhood where the lower the mean monthly crime rate, the higher the safety score.

Before we analyzed the Yelp restaurant ratings data, we wanted to make sure we had balanced data. That is how we came to the 2km radius parameter for the Yelp API (plus the fact that neighbourhood radii typically are a bit smaller than 2km according to Google Maps); any smaller and some neighbourhoods would have had less than 50 restaurants. We also could have looked at the top 50 best reviewed or 50 closest restaurants, but we chose the top 50 most reviewed since we believed that gave the most accurate snapshot of the restaurant quality in each neighbourhood. Like with the crime data, before performing any ANOVA test, we created histograms of the Yelp ratings to see if they were normal looking. Looking at Figures 3a) and 3b), we can see that not every neighbourhood's histogram looked exactly normal, but we thought it looked normal enough, and since we had 50 data points for each neighbourhood, we felt good about proceeding without any normality or equal variance tests. We then conducted an ANOVA test to see if the mean restaurant Yelp rating between any two neighbourhoods differed. After that, we did a post-hoc analysis with Tukey's HSD to see which neighbourhoods had different means. To generate a score for restaurant quality, we simply used the mean Yelp rating of the top 50 most reviewed restaurants in each neighbourhood since that was already on a 1-5 point scale.

Results:

The Skytrain distance scores varied nicely and as expected. More urban places like Downtown got high scores, while more suburban ones like West Point Grey received low scores.

The price scores also varied as one would expect. Posh neighbourhoods like Shaughnessy were the most expensive, while more blue-collar ones like Hastings-Sunrise were the cheapest.

The one-way chi-square tests all showed with a high degree of certainty ($p < 0.0083$ for all six tests, see Figure 6) that neighbourhood impacted the amenity counts for all six categories, even after factoring in population differences. However, one of the one-way chi-square test assumptions is that all the expected counts are at least 5.[9] This was an issue because the expected counts for some of the amenity categories and some of the smaller neighbourhoods were below 5. We didn't believe this was an issue for any category except for education and childcare, since in every other category, the expected counts were at least 5 for at least 68% of the neighbourhoods. However, for the education category, none of the expected counts were above 5, so we decided that, despite the low *p*-value, we failed to reject that particular null hypothesis.

Looking closer at the binomial test *p*-values for the amenities in Figure 7 and the resulting scores in Figure 1, we see some interesting results. First, from the arts/culture category we see that Strathcona is the place to be for those kinds of amenities with the sole 5/5 score. Second, in terms of the car service category, we see a healthy mix of scores. Downtown and West End are on top, each with a 5/5, while Renfrew-Collingwood is the only neighbourhood that got a 1/5. Third, looking at the education/childcare category we see that only Downtown Vancouver had a statistically significant amenity count with its 5/5 score. Fourth, for food/drink we see scores ranging from 1-5. Downtown and Fairview are areas with many places to eat, while more residential neighbourhoods like Shaughnessy and Arbutus Ridge don't have a lot of options. Fifth, for healthcare, Downtown had the most amenities and only a few other neighbourhoods had a score not equal to 3. Sixth, for the money category, we see a similar story with Downtown having a lot of amenities and only a few other neighbourhoods with statistically significant counts. From all these results, we see two things in particular: one that Downtown Vancouver has a lot of amenities relative to its population, and two for many amenity categories, our α's were set too low to differentiate between many different neighbourhoods (but since we're all good data scientists, we aren't going to re-run our analyses with higher α's).

The *p*-value for the crime ANOVA test was virtually 0, so we can conclude that at least two neighbourhoods have different means of the square roots of monthly violent crime rates. From Tukey's HSD post-hoc results in Figure 4, we can't conclude that many neighbourhoods have different means, but we can definitely conclude that Strathcona is the most dangerous neighbourhood. For safety scores, Strathcona got a 0, while the next lowest was 3.31. This makes sense as it is home to the notorious Downtown Eastside, the roughest part of Vancouver by far.

[9] docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html

The *p*-value for the Yelp ANOVA test was incredibly small ($p < 10^{-25}$), so we can conclude that at least two neighbourhoods have different mean restaurant Yelp ratings. Looking at Tukey's HSD plot in Figure 5, it seems like Marpole has the worst restaurants while Downtown Vancouver and the West End have the best. Despite this, there was a lot of overlap between different neighbourhoods, so we were unable to conclude that many pairs of neighbourhoods have different means. The restaurant quality scores ranged from 3.27 to 4.02.

From the example neighbourhood recommender run-through in Figure 8 and many others we performed during testing, we usually saw Downtown Vancouver at or near the top, probably due to how well it did in every amenity category.

Taking a closer look at the three map visualizations in Figure 9, we noticed how amenities seem to closely follow major arterial roads like Broadway, Kingsway and Granville St.

Limitations:

One major limitation was on the amount of data points we could get for each neighbourhood. Vancouver has 22 neighbourhoods, and that made it hard to get enough data points for each measure to get statistically significant results. This caused our issues with the chi-square assumptions and probably why every education/childcare score except for Downtown Vancouver's is 3/5. The OSM data provided a lot of Vancouver amenities, but we noticed that it was missing a number of them as well, like Windermere Secondary School. We had a similar problem with the Yelp data. The API restricted us to only 50 ratings per search, and we feel if that limit was higher, we could have distinguished between more pairs of neighbourhoods in terms of restaurant quality; there was a lot of overlap in our Yelp post-hoc plot, Figure 5.

Another limitation was how sometimes we had to represent each neighbourhood as a singular point rather than as a 2D polygon. This occurred with the Skytrain analysis and the Yelp rating collection, where instead we used the latitude/longitude point given by Google Maps' Geocoding API. This wasn't a terrible representation of each neighbourhood, but depending on where the point was located, it would change the results. For example, in one neighbourhood the point might be next to a Skytrain station, while in another the point might be in the geographic centre. In that case, it might not be fair to conclude that one neighbourhood is "closer" to the Skytrain, or that the restaurants within a 2km radius of the point are indicative of the whole neighbourhood's restaurant quality. Given more time, we would have investigated ways to either always represent neighbourhoods as 2D polygons, or at least find a better representation.

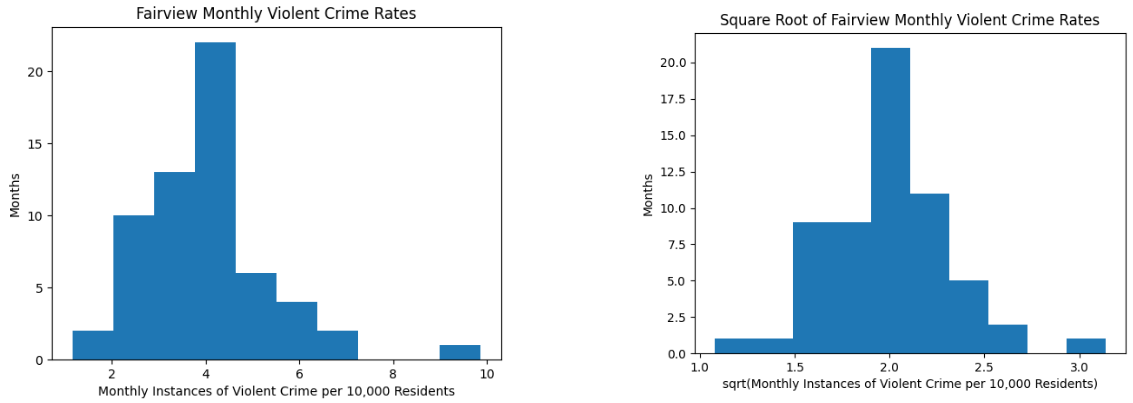
Finally, some of the data that we originally wanted to use for this project ended up being unattainable. This limited our analysis in several ways, namely in predicting and comparing home prices in each neighborhood. At first we planned to use an API to collect real estate price data, but every API we found could only be used by people who had an official realtor license, or by those who paid ~$400. In the end, we resorted to manually scraping very basic average sale price data from redfin.ca, since even an automatic scraper was banned on the site.

Visualizations:

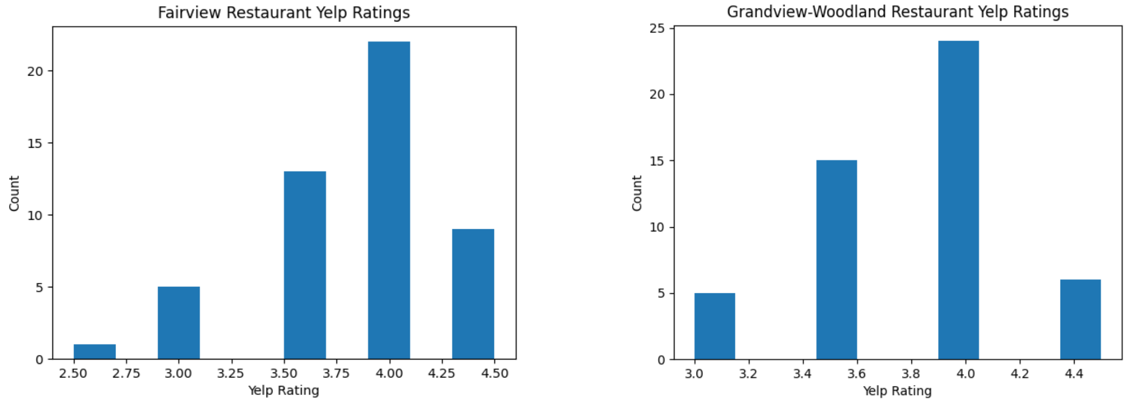
Scores out of 5 for each Neighbourhood and Measurement: (Figure 1):

neighbourhood	skytrain	price	arts/culture	cars	education/childcare	food/drink	healthcare	money	safety	yelp
Arbutus Ridge	2	3	3	3	3	1	3	3	4.76	3.75
Downtown	5	2	4	5	5	5	5	5	3.31	4.02
Dunbar-Southlands	1	3	3	2	3	1	3	3	4.87	3.69
Fairview	4	3	4	4	3	5	3	4	4.91	3.83
Grandview-Woodland	3	4	3	3	3	5	3	3	4.11	3.81
Hastings-Sunrise	3	5	3	3	3	1	3	3	4.69	3.79
Kensington-Cedar Cottage	3	3	3	3	3	1	3	3	4.81	3.90
Kerrisdale	2	2	3	3	3	1	3	3	4.50	3.39
Killarney	3	5	2	2	3	1	3	3	5.00	3.40
Kitsilano	2	2	3	3	3	5	4	3	4.83	3.77
Marpole	4	3	3	3	3	1	3	3	4.62	3.27
Mount Pleasant	4	2	3	5	3	4	3	3	4.61	3.92
Oakridge	5	3	3	3	3	1	3	3	4.24	3.60
Renfrew-Collingwood	5	5	2	1	3	1	2	2	4.83	3.69
Riley Park	3	3	3	2	3	3	3	3	4.54	3.83
Shaughnessy	3	1	3	3	3	1	3	3	3.53	3.67
South Cambie	4	1	3	3	3	3	3	3	4.43	3.77
Strathcona	4	4	5	3	3	4	3	3	0.00	3.86
Sunset	3	5	3	2	3	1	3	3	4.92	3.51
Victoria-Fraserview	2	5	3	3	3	1	3	2	5.00	3.60
West End	3	3	3	5	3	5	3	3	4.39	4.02
West Point Grey	1	1	3	3	3	3	3	3	4.64	3.60

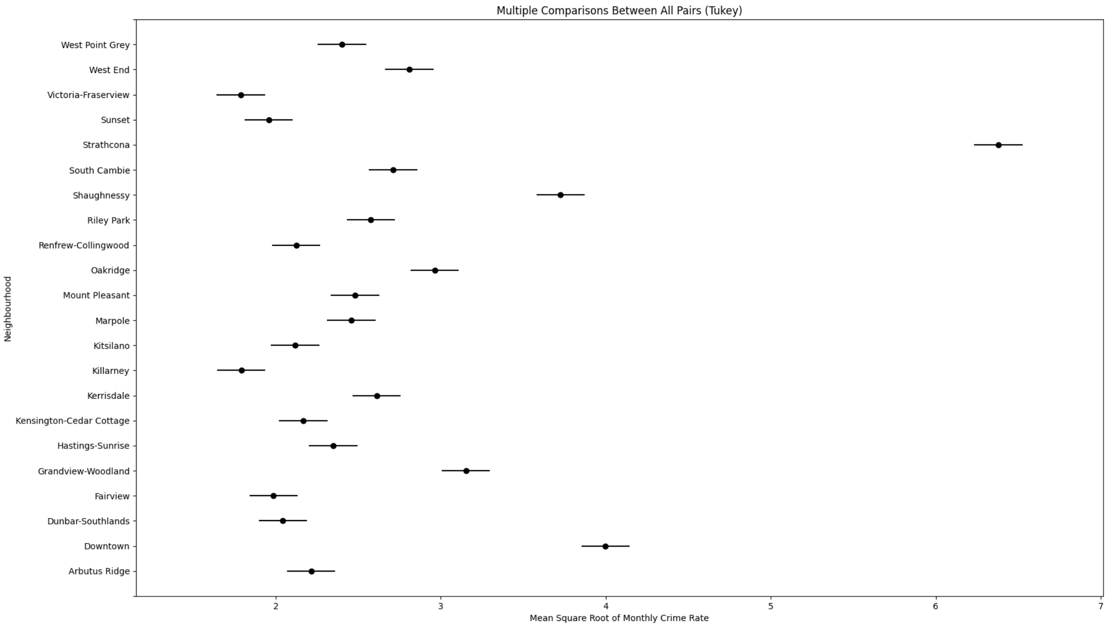
Example Histograms of Crime Rates Before & After Transformation (Figures 2a) & 2b)):



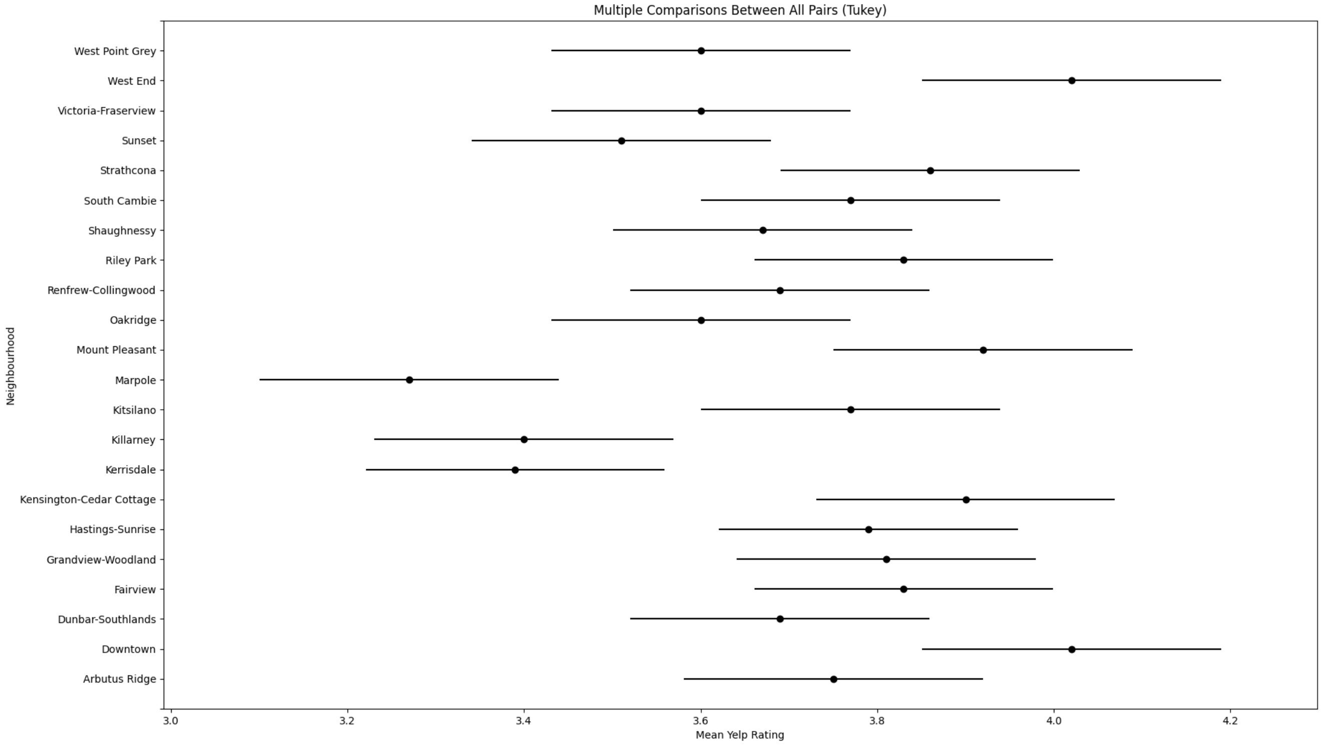
Example Histograms of Yelp Ratings (Figures 3a) & 3b)):



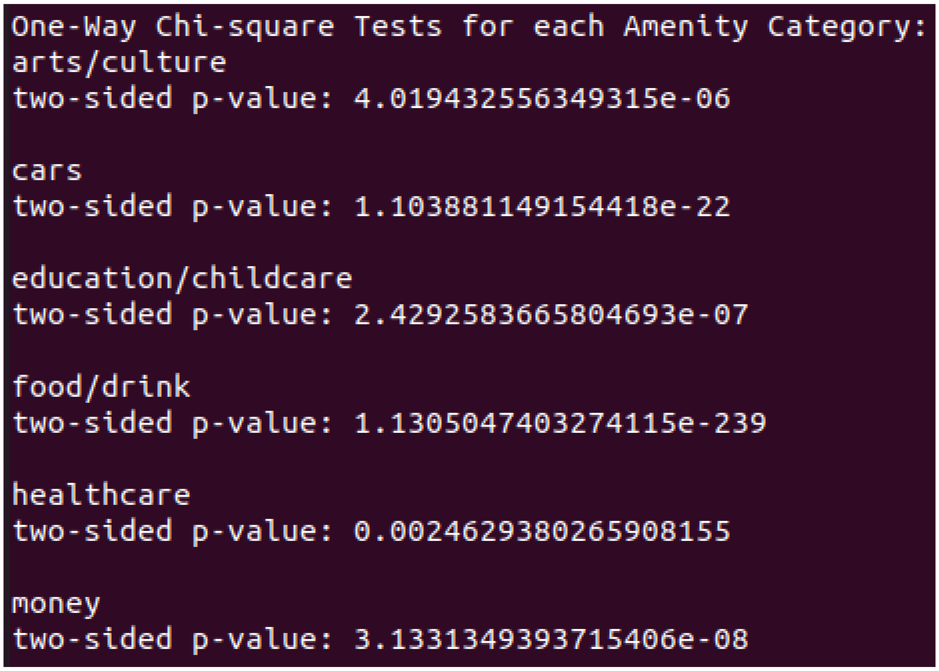
Result of Post-Hoc Analysis of Crime Rates (Figure 4):



Result of Post-Hoc Analysis of Yelp Restaurant Ratings (Figure 5):

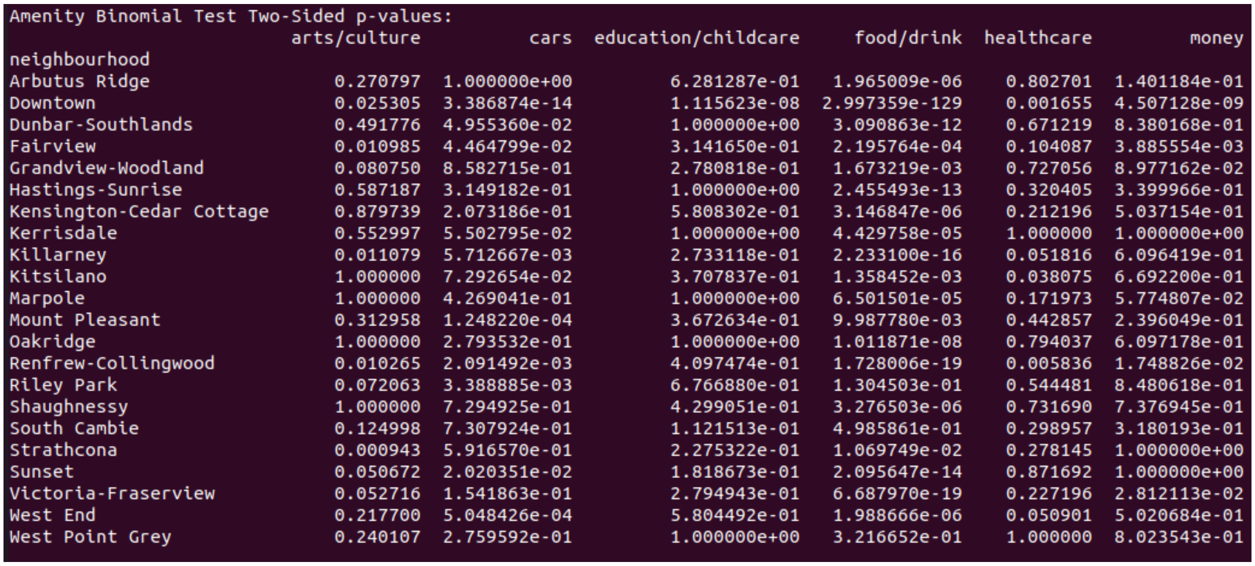


Results of One-Way Chi-Square Tests (Figure 6):



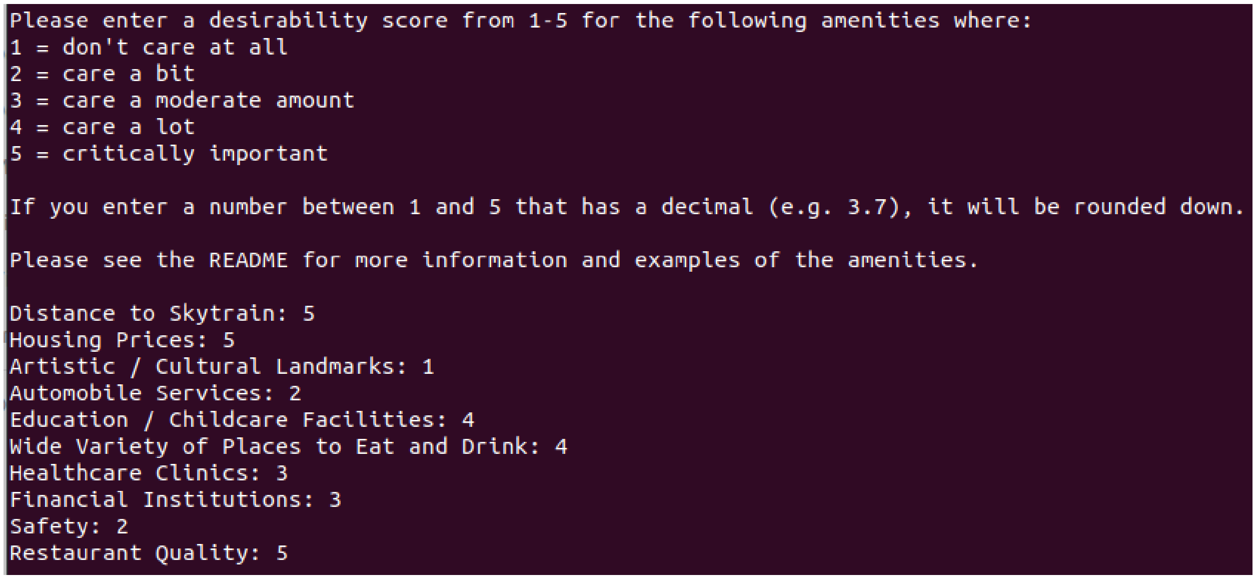
```
One-Way Chi-square Tests for each Amenity Category:
arts/culture
two-sided p-value: 4.019432556349315e-06

cars
two-sided p-value: 1.103881149154418e-22

education/childcare
two-sided p-value: 2.4292583665804693e-07

food/drink
two-sided p-value: 1.1305047403274115e-239

healthcare
two-sided p-value: 0.0024629380265908155

money
two-sided p-value: 3.1331349393715406e-08
```

Results of Two-Sided Binomial Tests (Figure 7):

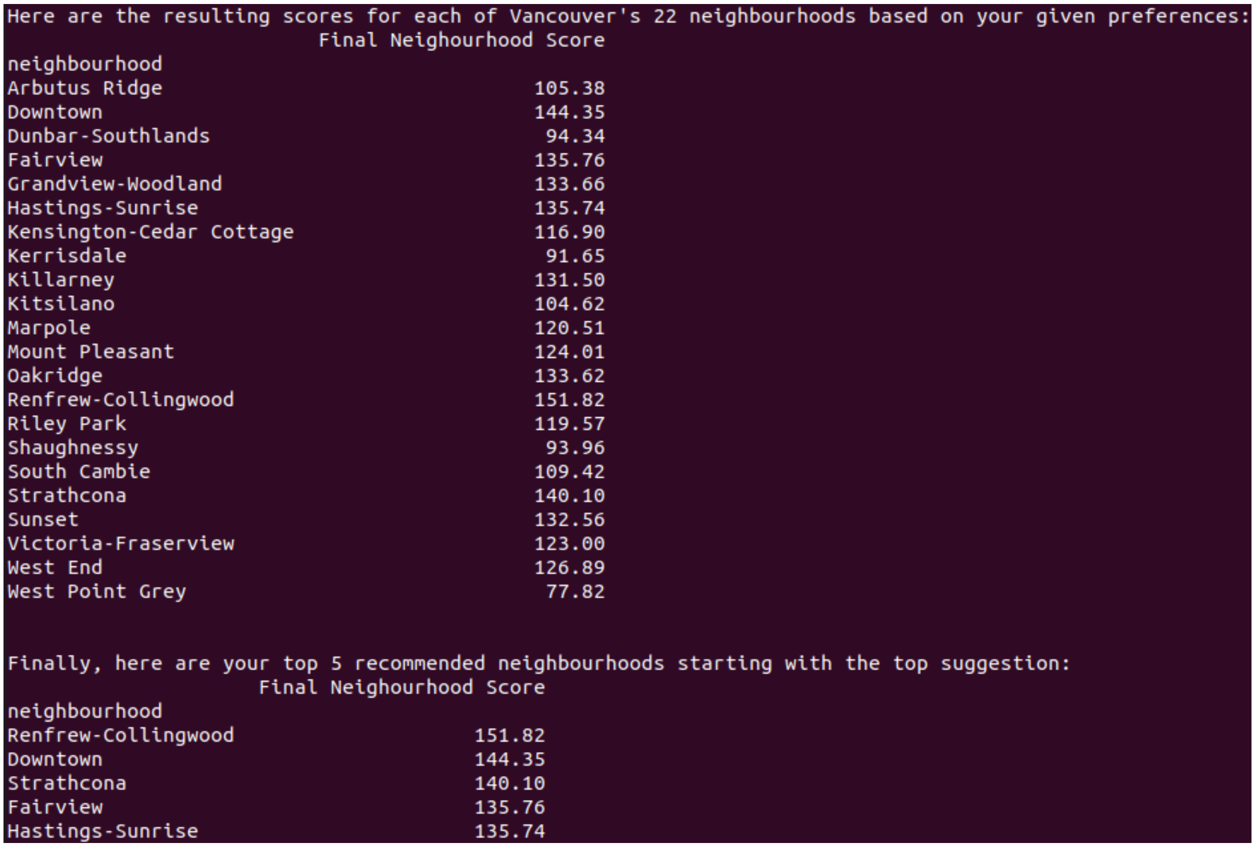


```
Amenity Binomial Test Two-Sided p-values:
                          arts/culture          cars  education/childcare     food/drink  healthcare         money
neighbourhood
Arbutus Ridge                 0.270797  1.000000e+00         6.281287e-01   1.965009e-06    0.802701  1.401184e-01
Downtown                      0.025305  3.386874e-14         1.115623e-08  2.997359e-129    0.001655  4.507128e-09
Dunbar-Southlands             0.491776  4.955360e-02         1.000000e+00   3.090863e-12    0.671219  8.380168e-01
Fairview                      0.010985  4.464799e-02         3.141650e-01   2.195764e-04    0.104087  3.885554e-03
Grandview-Woodland            0.080750  8.582715e-01         2.780818e-01   1.673219e-03    0.727056  8.977162e-02
Hastings-Sunrise              0.587187  3.149182e-01         1.000000e+00   2.455493e-13    0.320405  3.399966e-01
Kensington-Cedar Cottage      0.879739  2.073186e-01         5.808302e-01   3.146847e-06    0.212196  5.037154e-01
Kerrisdale                    0.552997  5.502795e-02         1.000000e+00   4.429758e-05    1.000000  1.000000e+00
Killarney                     0.011079  5.712667e-03         2.733118e-01   2.233100e-16    0.051816  6.096419e-01
Kitsilano                     1.000000  7.292654e-02         3.707837e-01   1.358452e-03    0.038075  6.692200e-01
Marpole                       1.000000  4.269041e-01         1.000000e+00   6.501501e-05    0.171973  5.774807e-02
Mount Pleasant                0.312958  1.248220e-04         3.672634e-01   9.987780e-03    0.442857  2.396049e-01
Oakridge                      1.000000  2.793532e-01         1.000000e+00   1.011871e-08    0.794037  6.097178e-01
Renfrew-Collingwood           0.010265  2.091492e-03         4.097474e-01   1.728006e-19    0.005836  1.748826e-02
Riley Park                    0.072063  3.388885e-03         6.766880e-01   1.304503e-01    0.544481  8.480618e-01
Shaughnessy                   1.000000  7.294925e-01         4.299051e-01   3.276503e-06    0.731690  7.376945e-01
South Cambie                  0.124998  7.307924e-01         1.121513e-01   4.985861e-01    0.298957  3.180193e-01
Strathcona                    0.000943  5.916570e-01         2.275322e-01   1.069749e-02    0.278145  1.000000e+00
Sunset                        0.050672  2.020351e-02         1.818673e-01   2.095647e-14    0.871692  1.000000e+00
Victoria-Fraserview           0.052716  1.541863e-01         2.794943e-01   6.687970e-19    0.227196  2.812113e-02
West End                      0.217700  5.048426e-04         5.804492e-01   1.988666e-06    0.050901  5.020684e-01
West Point Grey               0.240107  2.759592e-01         1.000000e+00   3.216652e-01    1.000000  8.023543e-01
```

Example Run-through of Neighbourhood Recommender (Figures 8 a) & 8b)):

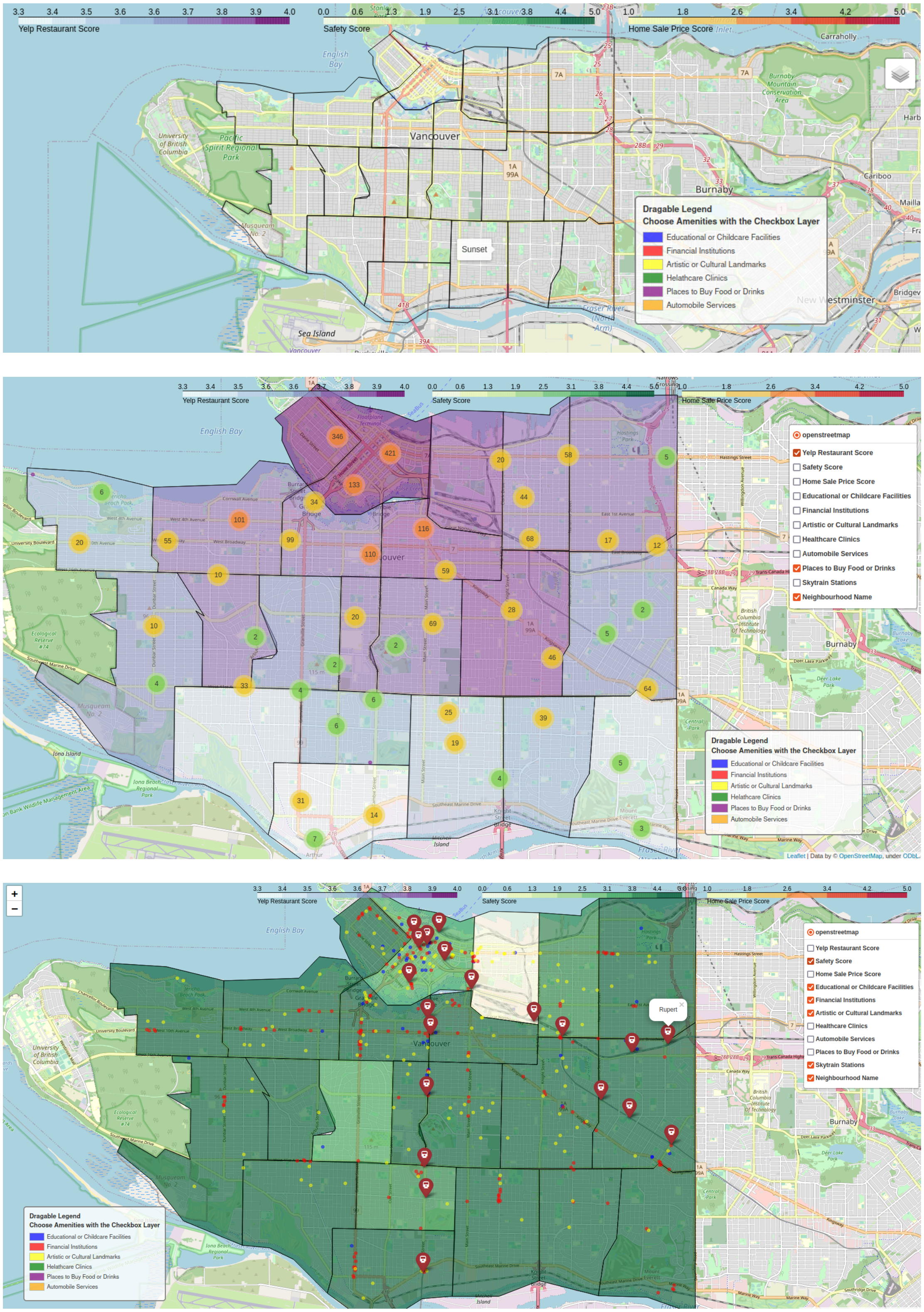


```
Please enter a desirability score from 1-5 for the following amenities where:
1 = don't care at all
2 = care a bit
3 = care a moderate amount
4 = care a lot
5 = critically important

If you enter a number between 1 and 5 that has a decimal (e.g. 3.7), it will be rounded down.

Please see the README for more information and examples of the amenities.

Distance to Skytrain: 5
Housing Prices: 5
Artistic / Cultural Landmarks: 1
Automobile Services: 2
Education / Childcare Facilities: 4
Wide Variety of Places to Eat and Drink: 4
Healthcare Clinics: 3
Financial Institutions: 3
Safety: 2
Restaurant Quality: 5
```



```
Here are the resulting scores for each of Vancouver's 22 neighbourhoods based on your given preferences:
                          Final Neighourhood Score
neighbourhood
Arbutus Ridge                               105.38
Downtown                                    144.35
Dunbar-Southlands                            94.34
Fairview                                    135.76
Grandview-Woodland                          133.66
Hastings-Sunrise                            135.74
Kensington-Cedar Cottage                    116.90
Kerrisdale                                   91.65
Killarney                                   131.50
Kitsilano                                   104.62
Marpole                                     120.51
Mount Pleasant                              124.01
Oakridge                                    133.62
Renfrew-Collingwood                         151.82
Riley Park                                  119.57
Shaughnessy                                  93.96
South Cambie                                109.42
Strathcona                                  140.10
Sunset                                      132.56
Victoria-Fraserview                         123.00
West End                                    126.89
West Point Grey                              77.82


Finally, here are your top 5 recommended neighbourhoods starting with the top suggestion:
                     Final Neighourhood Score
neighbourhood
Renfrew-Collingwood                    151.82
Downtown                               144.35
Strathcona                             140.10
Fairview                               135.76
Hastings-Sunrise                       135.74
```

Examples of Map Visualizations of Results (Figures 9a), 9b), & 9c)):



Appendix:

Population Extrapolation Example:

If the population in a neighbourhood in 2016 was 12,000 and in 2011 was 10,000, then the average yearly population change rate was (12,000 / 10,000)^(1/5) = 1.037 or an increase of 3.7% per year. In that case, we estimated 2017's population to be 12,000 * 1.037 = 12,445, 2018's population to be 12,445 * 1.037 = 12,907, etc.

Expected Amenity Count Example:

If there are 2,000 eating establishments in Vancouver, and Kerrisdale has 2% of Vancouver's total population, then we expect that Kerrisdale would have 2,000 * 0.02 = 40 eating establishments.

Amenity Count Binomial Test and Scoring Example:

Using Kerrisdale as an example again, let's say the OSM data says it contains 25 eating establishments. Then, for a binomial test we have the following hypotheses:

Null: true proportion of eating establishments here compared to all of Vancouver = 0.02

Alternative: true proportion of eating establishments ≠ 0.02 (so two-sided test)

We also have the following parameters:

Number of Successes: 25

Number of Trials: 2,000

Probability of Success: 0.02

This gives a *p*-value of 0.0132. Since 25 < 40 (the expected number of eating establishments) and $p < 0.05$, but $p > 0.05/22 = 0.00227$, that would give Kerrisdale a food/drink amenity score of 2 out of 5.

Project Experience Summaries:

James

- Collected geographic data of Vancouver's 22 neighbourhoods and 53 Skytrain stations using Google Maps' Geocoding API, which allowed my team to analyze distances between neighbourhoods and points of interest.
- Transformed over 17,000 records of amenity data using the Pandas library to properly format it for statistical analyses. Co-wrote Python code to categorize each amenity into one of six major categories and identify which neighbourhood it belonged to.
- Performed over 100 statistical analyses such as chi-squared and binomial tests using the SciPy Python package to determine the effect neighbourhood had on the amount of amenities nearby relative to the local population. Assigned each neighbourhood six scores, one for each amenity category.
- Programmed a Python script which utilized user inputs and the computed neighbourhood scores to recommend the user the best neighbourhood to live based on their preferences.
- Co-authored a 12-page report on this project that will help readers understand my group's methods and findings.

Jonghyeok

- Proposed the project topic and subdivided the project into 5 areas of interest - Skytrain score, home sale price score, amenity score, safety score, and restaurant quality - such that the problem could be answered with data analysis.
- Acquired demographic and geographic data of Vancouver's 22 neighbourhoods from official government data sources. Contributed to writing code that used the Shapely package to determine which neighbourhood each amenity belonged to, which allowed my team to calculate the amenity score of each neighbourhood.
- Cleaned and transformed over 750,000 crime records along with population data from 2016 to 2020 to conduct data analysis. Performed statistical analyses such as ANOVA and Tukey's HSD tests to compare the safety of each neighbourhood.
- Wrote Python code which plotted the choropleth map of safety score and plotted roughly 3,000 amenity locations by using the Folium package so that users can visually understand the outcome of the analyses.

Myckland

- Proposed an additional data source, Yelp Fusion's REST API, which enriched the quality of the data analyses with over 1,000 restaurant ratings.
- Performed ETL on the Yelp data using Pandas to format it for further analysis.
- Conducted ANOVA and Tukey's HSD tests on the Yelp data using the SciPy package to determine if restaurant quality differs by neighbourhood.
- Wrote Python code using Folium functions to plot a Choropleth map of Yelp restaurant ratings, Skytrain station locations, and other points of interest to visualize results.
- Provided the draggable legend for the amenity points and created restaurant clusters to reduce cluttering, which improved user experience when using the map visualization.