CMPT 353 Final Project: Vancouver Transit Exploration Task

Jason Cai, Rebekah Wong (Group: “Mute Jammers”)

1. Introduction

Public transit in Vancouver has been consistently ranked as one of the best worldwide; as of 2023, TransLink was announced as the 22nd best public transit system internationally, and also placed fourth and first within North America and Canada respectively [1]. However, as Vancouverites, the public transit rankings within Vancouver neighbourhoods may be more relevant to us. Additionally, public transit is a means of convenience for the general populace, so changes in the population of neighbourhoods in Vancouver may also lead to fluctuations in perceived transit quality around different regions of the city. Hence, this Vancouver transit exploration task aims to answer the following questions:

1. To what extent has Vancouver’s transit quality changed over the years within different neighbourhoods?
2. Is there a correlation between the perceived transit quality and the population of different Vancouver neighbourhoods?
3. Is there a correlation between the perceived transit quality and transit ridership within Vancouver?

2. Data

i. Data Characteristics

The City of Vancouver Intangible Transit Costs dataset (data/transit.zip) [2]—which coincidentally originated from Simon Fraser University—was used as a starting point and remains as a central component to the transit exploration task. The data contained transit quality scores over five-year intervals corresponding to 992 smaller subsections of Vancouver, represented as different geometry shapes that could be mapped out with GeoPandas [3]—a library based on Pandas, but better suited for geographic information system data—as shown in Figure 1. This dataset was vital to have acquired, as most of the analyses were dependent on the transit quality scores to answer the first research question outlined in the Introduction section.

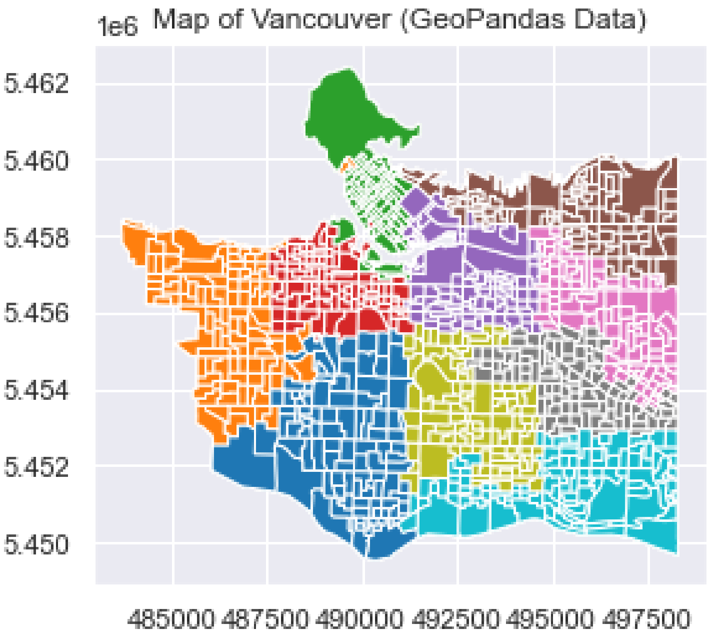

Figure 1. Map of Vancouver from GeoPandas in the transit quality dataset.

Other relevant data included Vancouver's neighbourhood boundaries represented in a similar geometric shape format as the transit quality dataset, as defined by the City of Vancouver (data/local-area-boundary.zip) [4], as well as the most recent 2016 census information from Statistics Canada and compiled by the City of Vancouver pertaining to not only population counts, but also more specific demographic information of different neighbourhoods in Vancouver (data/CensusLocalAreaProfiles2016.csv) [5] to help answer the second research question.

Finally, TransLink ridership data (data/translink-historic-ridership-trend.csv) [6] with annual data collected from 2003 to 2022 of the total number of public transit boardings (measured in millions) was also taken to answer the third research question. All of the aforementioned data were directly downloaded from various websites to use in the project.

ii. Data Cleaning

For the transit quality dataset, our data cleaning steps after reading the zip file included renaming column headers with more descriptive names, and extracting the relevant years to our analysis (2010, 2015, and 2020), as future years up to 2040 were also measured in the dataset—these were filtered out and excluded, as the future readings were presumably extrapolated data that may have been useful for what the data was originally collected for, but not for this transit exploration project. Most notably, however, the uncleaned transit quality data from 2010 to 2020 had a minimum value of 0.5, a maximum value of 4765.2, a mean of 446.5, and a median of 204.6. As shown in Figure 2, the resulting histogram was evidently heavily right-skewed.

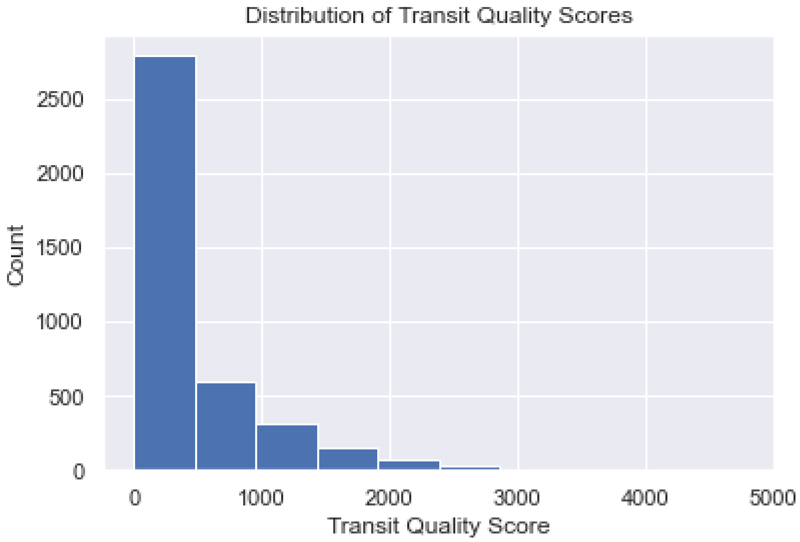

Figure 2. Histogram depicting the distribution of original transit quality scores.

Thus, the square root and logarithm were subsequently applied to the transit quality scores separately, to determine which modification resulted in the least skewed data. Figure 3 and Figure 4 depict the updated histograms of transit quality counts after applying the square root and logarithm respectively.

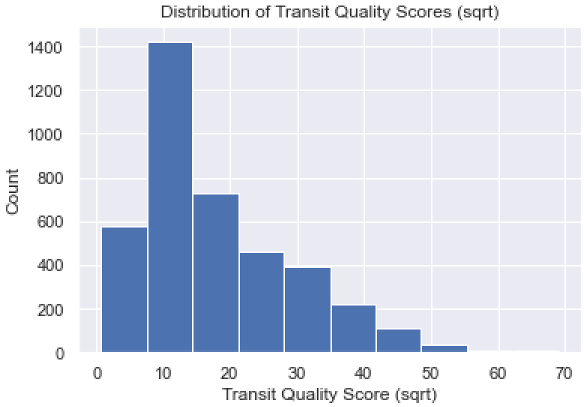

Figure 3. Histogram depicting the distribution of square root transit quality scores.

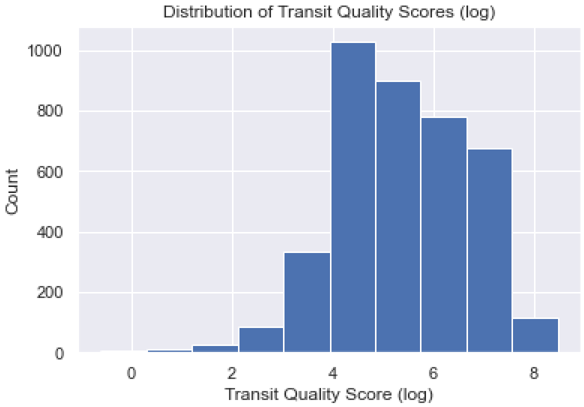

Figure 4. Histogram depicting the distribution of logarithm transit quality scores.

The minimum value for the square root data was 0.7, the maximum value was 69.0, the mean was 17.8, and the median was 14.3. On the other hand, the minimum value for the logarithm data was -0.6, the maximum value was 8.5, the mean was 5.4, and the median was 5.3. Conceptually, the transit quality after applying the logarithm was the easiest to digest, as the numbers could have reasonably fit a range from 0 to 10. Thus, 1 point was added to all of the logarithm values for the final scoring, as a means of properly fitting the scores on a scale out of 10.

The 2016 census dataset was messier to clean compared to the others, but involved extracting the neighbourhoods and their respective population count from columns rather than rows and placing said data into a new dataframe. All other census data was unneeded for the project and thus omitted in the cleaning process. In addition, the population counts were strings rather than ints, and in a format such that thousands were separated by a comma (e.g. 15,295), so the commas had to be stripped before the population counts were appropriately casted as ints.

The census data was later combined with the neighbourhood boundaries shown in Figure 5 to connect the mapped neighbourhoods to their corresponding population. Entity resolution was required in this part, as one of the neighbourhoods were labelled as "Arbutus-Ridge" in the census data, but "Arbutus Ridge" without the hyphen in the area boundaries data, although it was evident enough that both of these names belonged to the same entity.

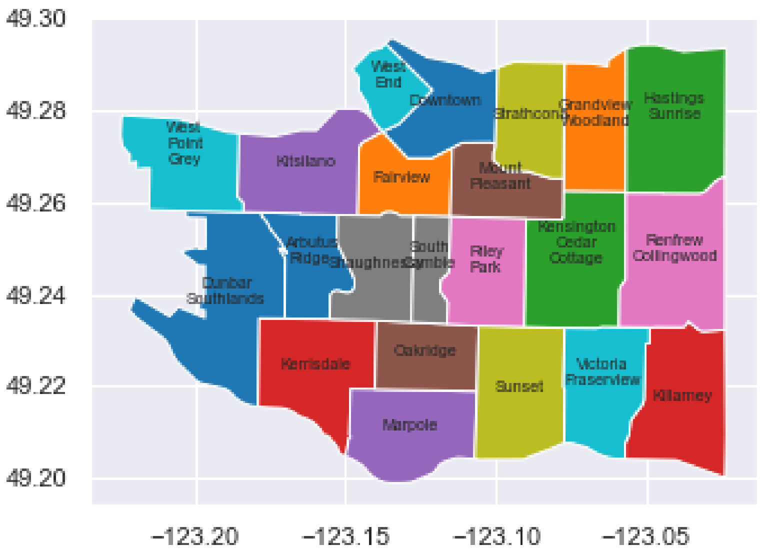

Figure 5. Labelled Vancouver neighbourhoods map from the area boundaries dataset.

Most importantly, the 992 geometry shapes representing different subsections of Vancouver from the transit quality dataset had to be linked to their corresponding neighbourhood. Initially, this was done by sorting the shape values, as shapes closer in value would theoretically be closer on the map as well. Groups of sorted shapes were printed out and cross-referenced with the preexisting boundaries from the aforementioned Figure 5, and manually assigned neighbourhoods as an estimate. Samples of the assignment process are illustrated in Figure 6 and Figure 7, while Figure 8 contains the full map of boundary estimates across Vancouver neighbourhoods. The neighbourhood and population data from the census dataset was later joined with the transit quality data based on the newly established neighbourhood assignments resulting from the cleaning process.

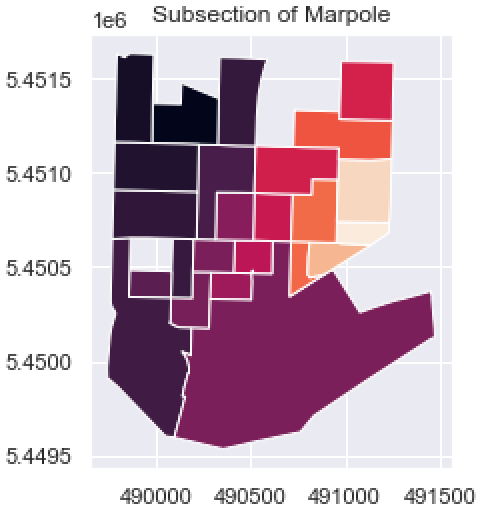

Figure 6. Map of a subsection identified as Marpole, cross-referenced with the neighbourhood boundaries.

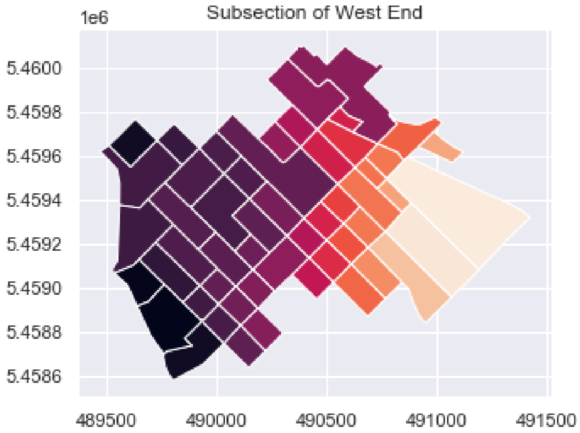

Figure 7. Map of a subsection identified as West End, cross-referenced with the neighbourhood boundaries.

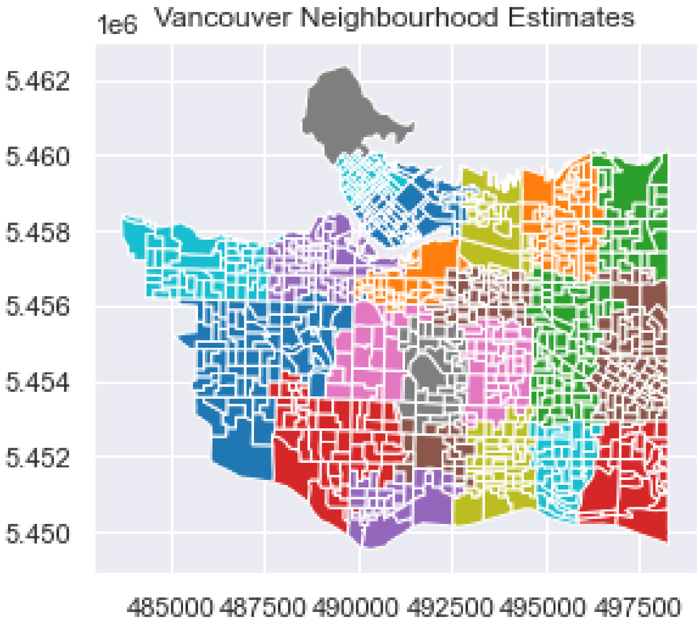

Figure 8. Vancouver neighbourhood boundary estimates based on subsection cleaning.

However, a cleaner algorithm was later used by using the Shapely library to translate, scale, and normalize every geometric shape into such that they overlaid one another. This process involved calculating the centroid of all geometric shapes, as well as the average distance of the centroids of every shape in the common centroid to keep both maps within the same scale.

3. Techniques

Analyzing the transit data involved aggregating the unskewed transit quality scores into averages per geometry shape region over the years, which was then used to plot a choropleth map to represent the range of transit quality scores found in different subsections of Vancouver, as seen in Figure 9. The same aggregation process was also applied to data filtered by 2010, 2015, and 2020 to visualize any changes in the choropleth map.

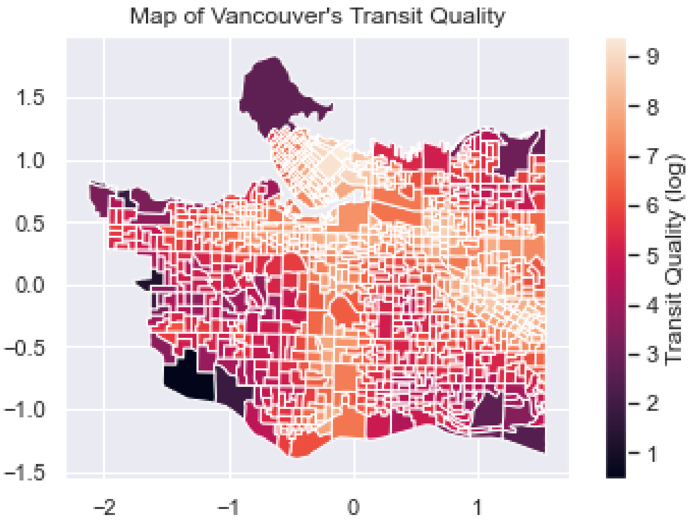

Figure 9. Choropleth map of logarithm transit quality scores averaged from 2010-2020.

The Pearson correlation (*r*) test was also conducted to determine the correlation between transit quality and other variables such as neighbourhood population and public transit ridership. Additionally, linear regression was used to visually determine whether there were any linear trends in the relationship between transit quality and neighbourhood population.

4. Results

The transit quality scores across 2010, 2015, and 2020 have remained relatively consistent over the years, with minimal changes in quality scores and neighbourhood rankings. Table 1 lists the minimum, maximum, mean, and median scores of all subsections of Vancouver across years. On average, Vancouver's public transit system scores a 6/10 in terms of quality, although because the data was modified to be unskewed during cleaning, these results are relative. The subsection with the best transit quality was consistently within Downtown, while the subsection with the worst transit quality was consistently within Dunbar-Southlands.

	Minimum	**Maximum**	**Mean**	**Median**
2010	0.39	9.39	6.20	6.10

	(Dunbar-Southlands)	(Downtown)		
2015	0.40 (Dunbar-Southlands)	9.40 (Downtown)	6.37	6.36
2020	0.43 (Dunbar-Southlands)	9.47 (Downtown)	6.43	6.44

Table 1. Minimum, maximum, mean, and median transit quality score (/10) over years.

Meanwhile, Table 2 showcases the neighbourhood rankings from year to year. As previously mentioned in Table 1, the Downtown neighbourhood consistently ranked first in terms of having the best transit quality in Vancouver, whereas there was more variation and fluctuation in the ranking of other neighbourhoods. Similarly and unfortunately, Dunbar-Southlands was also consistently perceived as the worst neighbourhood to transit in.

2010 Neighbourhood Ranking	**2010 Transit Quality**	**2015 Neighbourhood Ranking**	**2015 Transit Quality**	**2020 Neighbourhood Ranking**	**2020 Transit Quality**
1. Downtown	8.14	1. Downtown	8.15	1. Downtown	8.20
2. Renfrew-Collingwood	7.46	2. Fairview	7.77	2. Fairview	7.81
3. West End	7.26	3. Renfrew-Collingwood	7.46	3. Renfrew-Collingwood	7.50
4. Strathcona	7.17	4. West End	7.26	4. West End	7.32
5. Kensington-Cedar Cottage	7.05	5. Mount Pleasant	7.24	5. Mount Pleasant	7.29
6. Fairview	6.80	6. Strathcona	7.18	6. Strathcona	7.24
7. Mount Pleasant	6.79	7. Kitsilano	7.09	7. Kitsilano	7.15
8. South Cambie	6.79	8. Kensington-Cedar Cottage	7.07	8. Kensington-Cedar Cottage	7.11
9. Marpole	6.75	9. South Cambie	6.80	9. South Cambie	6.84
10. Oakridge	6.45	10. Marpole	6.46	10. Marpole	6.53
11. Grandview-Woodland	6.29	11. Oakridge	6.30	11. Oakridge	6.37

12. Kitsilano	6.09	12. Grandview-Woodland	6.08	12. Grandview-Woodland	6.16
13. Shaughnessy	5.96	13. Shaughnessy	5.99	13. Shaughnessy	6.03
14. Hastings-Sunrise	5.76	14. West Point Grey	5.51	14. West Point Grey	5.59
15. West Point Grey	5.49	15. Riley Park	5.49	15. Hastings-Sunrise	5.56
16. Riley Park	5.43	16. Hastings-Sunrise	5.48	16. Riley Park	5.54
17. Kerrisdale	5.28	17. Kerrisdale	5.32	17. Kerrisdale	5.37
18. Sunset	5.09	18. Arbutus-Ridge	5.29	18. Arbutus-Ridge	5.36
19. Victoria-Fraserview	4.87	19. Sunset	5.10	19. Sunset	5.20
20. Killarney	4.79	20. Victoria-Fraserview	4.87	20. Victoria-Fraserview	4.99
21. Arbutus-Ridge	4.72	21. Killarney	4.79	21. Killarney	4.91
22. Dunbar-Southlands	4.46	22. Dunbar-Southlands	4.51	22. Dunbar-Southlands	4.57

Table 2. Ranking of average neighbourhood transit quality scores over time.

Expanding the choropleth map from Figure 9, similar maps have also been plotted for the 2010, 2015, and 2020 transit quality data in Figure 10, Figure 11, and Figure 12 respectively to illustrate the not-so noticeable changes in transit quality. Therefore, Vancouver's transit quality has slightly changed over the years within different neighbourhoods, but not enough to warrant a significant difference.

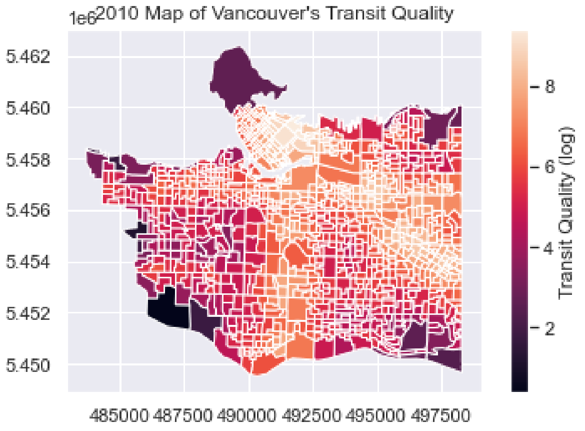

Figure 10. Choropleth map depicting logarithm transit quality values from 2010.

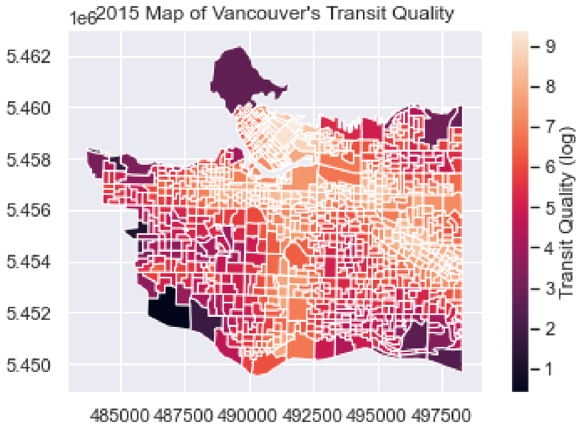

Figure 11. Choropleth map depicting logarithm transit quality values from 2015.

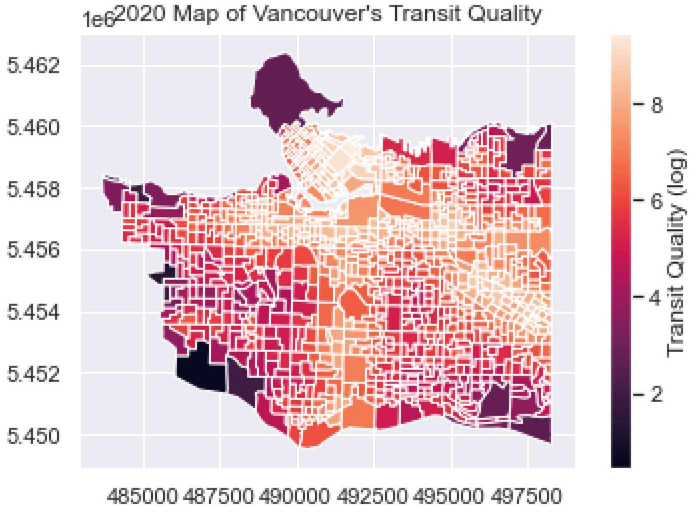

Figure 12. Choropleth map depicting logarithm transit quality values from 2020.

As for the correlation between transit quality and neighbourhood population, the Pearson coefficient resulted in $r = 0.466206$. Given that r-values closer to ±1 are linearly related, and r-values closer to 0 have little to no linear relationship, the ~0.5 r-value in between could be interpreted as having a weak linear relationship between quality score and population. Figure 13 conveys this more clearly with a best-fit prediction line from linear regression, and while the results are not as linear as they could be, there is still a modest relationship between these two variables. After all, there are popular and bustling areas of Vancouver that are often travelled to, such as Downtown, regardless of the neighbourhood one resides in.

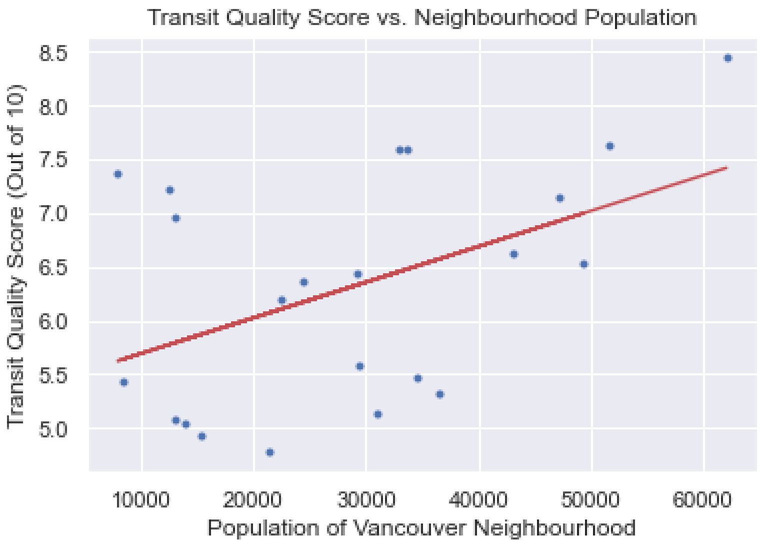

Figure 13. Scatter plot and best fit line of transit quality vs. neighbourhood population.

Finally, there was also very little correlation perceived between the transit quality and transit ridership within Vancouver. The Pearson coefficient resulted in r = -0.62739 in this case, so the relationship is once again quite modest between the two variables. One may expect for transit quality to affect ridership in a way such that transit quality would have decreased if ridership had too, but external variables that disrupt public transit usage also exist. For instance, transit quality has remained quite consistent throughout the years, but public transit ridership noticeably plummeted in 2020—this also marked the start of the coronavirus pandemic. Regardless of how good or poor Vancouver's transit quality was, transit quality seemed to be completely independent of ridership as it was not TransLink's fault that fewer people used public transit. Health and sanitary reasons due to the pandemic were justifiably attributed as a cause for the ridership decrease [7] rather than the transit quality scores found in the exploration task.

Interestingly, a recent article earlier in the week showcased a choropleth map of Vancouver transit times, and visually, the Downtown region had by far the shortest transit times [8], which corresponds to Downtown also consistently having the best transit quality in our study as previously shown in Figure 9 and Table 2. Preliminarily, the efficiency (or lack thereof) in transit time could be one such factor that contributes to a neighbourhood's transit quality based on these initial observations.

5. Limitations

A major problem of this project occurred in the data cleaning stage, as the original transit quality dataset only contained the geometry shapes that mapped out the choropleth map of Vancouver, but with no indication of which neighbourhood each piece belonged to. While there was another existing dataset that contained the neighbourhood boundaries of Vancouver [4], the map itself was unfortunately in a completely different scale than the map found in the transit quality data. Our inexperience with GeoPandas and working with geographic information initially led us to a brute force solution, but was eventually modified as outlined in the Data section with a more elegant solution in calculating centroid distances with subsequent scaling and normalizing. If we had more time, we would have ideally prioritized improving the more elegant algorithm to categorize the geometric shapes to their corresponding neighbourhoods to further improve the accuracy of our results, as the newer solution also contained unrecognizable neighbourhood pieces near the edge of the boundaries that were dropped in the interest of time.

Another limitation was found in the availability of Vancouver neighbourhood census data in recent years, as the City of Vancouver only had data published for 2001, 2006, 2011, and 2016. While the 2011 and 2016 census data could have been mapped to the 2010 and 2015 transit quality data as estimates, the lack of 2021 data could have, in retrospect, been remedied by using machine learning tools to predict future census data based on past records. Currently, the 2016 census data was used as a midpoint between 2010 and 2020 for the averaged neighbourhood transit quality and population data found in the output file, neighbourhood_transit_quality.csv. However, the general trend of transit quality score vs. neighbourhood population may have remained relatively similar regardless.

Similarly, the TransLink data for annual journeys (data/translink-annual-journeys.csv) [6] remained unused as it only contained data from 2018 to 2022 without any of the preceding years. While TransLink defines ridership as the number of boardings, many trips likely involve transferring buses and/or SkyTrains to reach the final destination. In this case, the journeys dataset condenses the multiple boardings as a single journey, which may be a better representative of public transit ridership. Machine learning tools would have been helpful to determine the relationship between boardings and journeys, and work backwards to predict what past records may have looked like as a more accurate representation of public transit usage.

6. Project Experience

Jason Cai:

- Spearheaded transit and traffic data analysis brainstorming, quickly narrowing down topics of interest for the project.
- Conducted statistical tests to check for correlations between transit quality and total annual ridership for all available data timeframes
- Normalized choropleth and zone maps for neighborhood matching and labeling to prepare for neighborhood transit quality analysis and identification, which were also used for neighborhood population and transit quality correlation analysis study
- Added UX improvements for the zone maps for better readability

Rebekah Wong:

- Modified heavily skewed GeoPandas transit quality data (ranging from 0 to thousands) to more understandable scores (out of 10) to graph onto choropleth maps for visualizing regions with better or worse transit quality across Vancouver
- Conducted statistical tests to ascertain the correlation (or lack thereof) between transit quality and neighbourhood population of different regions of Vancouver
- Analyzed the implications of transit quality scores over time in relation to population and external factors to determine whether other variables affect transit quality
- Spearheaded written report to convey relevant information to target audience of technically literate coworkers/manager

References

[1] K. Chan, “Vancouver’s TransLink is North America’s 4th-best public transit system,” Urbanized | Daily Hive, https://dailyhive.com/vancouver/vancouver-translink-north-america- best-public-transit (accessed Aug. 4, 2023).

[2] B. Zuehlke, “City of Vancouver Intangible Transit Costs,” Federated Research Data Repository, https://doi.org/10.25314/5e94d820-678e-4d3a-9a97-51fb730d5cf5 (accessed Aug. 4, 2023).

[3] “Geopandas 0.13.2,” GeoPandas, https://geopandas.org/en/stable/ (accessed Aug. 4, 2023).

[4] “Local area boundary,” City of Vancouver Open Data Portal, https://opendata.vancouver.ca/explore/dataset/local-area-boundary/information/ (accessed Aug. 4, 2023).

[5] “Census local area profiles 2016,” City of Vancouver Open Data Portal, https://opendata.vancouver.ca/explore/dataset/census-local-area-profiles-2016/information/ (accessed Aug. 4, 2023).

[6] “Accountability Centre - Ridership,” TransLink, https://www.translink.ca/plans-and-projects/data-and-information/accountability-centre/ridership (accessed Aug. 4, 2023).

[7] N. Griffiths, “Covid-19: How a pandemic changed transit in Metro Vancouver,” Vancouver Sun, https://vancouversun.com/news/local-news/covid-19-how-a-pandemic-changed-transit-in-metro-vancouver (accessed Aug. 4, 2023).

[8] A. Ali, “Colourful map makes it easier to understand Vancouver Transit Times,” Urbanized | Daily Hive, https://dailyhive.com/vancouver/time2reach-vancouver-transit-map (accessed Aug. 4, 2023).