CISC 451 – Assignment 1

Data Understanding – Planning Bus Routes

October 5th, 2020

Gavin McClelland – 10211444
Marshall Cunningham - 20249991

1.0 Introduction

To effectively plan bus routes for the City of Kingston, the current state of public transportation and urban landscape must be analyzed first. This assignment uses datasets that are made publicly available by the City of Kingston to visually analyze the dynamic of the transit service. Bearing in mind that these datasets are provided with limited description, they are open to interpretation and assumptions must be made to perform appropriate analyses. These assumptions will be detailed further in this paper when they are introduced. A variety of tools were used to perform analyses on these datasets. An exhaustive list of all datasets and software tools is provided, along with any installation instructions where applicable.

2.0 Prerequisites

2.1 Datasets

All datasets used in this assignment were taken from the Open Data Kingston portal [1]. Each explored dataset is listed in the table below, accompanied by a brief description, and a direct link to where they can be accessed or downloaded. All datasets are among those recommended in the assignment instructions.

Table 1 - Datasets used for analysis

Dataset	Description
Transit Data – October.xlsx	Provided data of people riding Kingston busses in October 2017
cycling-facilities.csv	City-wide cycling infrastructure
driveways.csv	All municipal driveways
neighbourhoods.csv	Different neighbourhood boundaries within the city
parking-areas.csv	Driven portion of public or private parking lots
points-of-interest.csv	Specific point locations of common interest within the city
transit-gtfs-routes.csv	General Transit Feed Specification (GTFS) for static bus routes
transit-gtfs-stops.csv	General Transit Feed Specification (GTFS) for individual stops

2.2 Software Packages and Tools

Any code written to create visuals and drive the analytics process was written in Python. The main packages used include numpy, pandas, plotly, matplotlib, turfpy, geojson, and Jupyter. For brevity, an exhaustive list of required Python modules will be included in the file requirements.txt and can be installed using the command “pip install -r requirements.txt”. Code written for questions 2 and 4 is located in the Jupyter Notebook “EDA-Q2-Q4.ipynb”, and code for question 3 is located in “Q3.ipynb”. Additionally, the script “Q1-Heatmap.py” was used to create the heatmap in support of question 1. A Power BI dashboard file “a1.pbix” was mainly used for exploration, and includes some simple visuals referenced in this document.

For visualizing geographical data, the primary software tool of choice was the web application found on the Open Data Kingston portal. This tool supports the ability to visualize the datasets in question without any third-party software. Furthermore, efforts were made to create simple geographical visuals with Microsoft Power BI, which can be downloaded from the Microsoft Store, or online [2]. Lastly, some visuals were created in Python by leveraging the plotly open source graphing library, and seaborn which was included in the previously discussed list of Python modules.

3.0 Overview of Analytics Process

The analytics process began by exploring the initially provided dataset (Transit Data – October.xlsx) and cleaning it accordingly for preliminary analysis (understand what was being given to work with). First, the records including erroneous data from malfunctioning GPS antennae were removed. This was first done by removing records with a value of zero (0) for latitude and longitude, but it was found that some outliers remained. These were removed by thresholding the coordinates of interest even further by determining boundaries. So, any record with a latitude less than 44.0 or greater than 45.0 was removed, while any record with a longitude value of less than -77.0 or greater than -76.0 was removed. Once cleaned, simple visuals were first created to explore basic relationships in the given dataset. First, the general usage of public transit over the course of October was analyzed. This is shown in the below day-by-day plot (Figure 1). Notice that the days with a lower volume of passengers are on weekends.

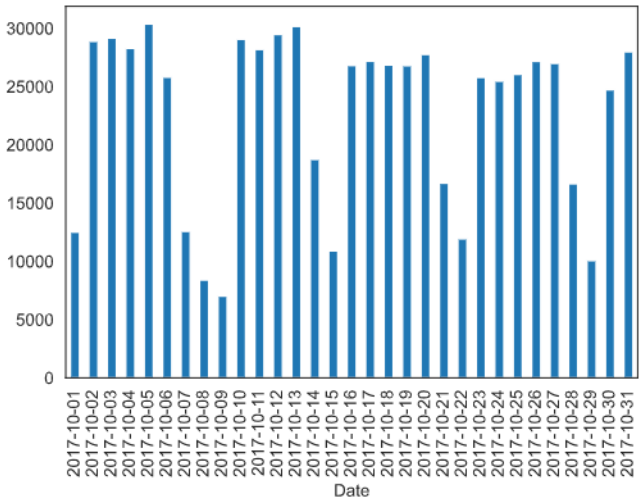

Figure 1 - Transit usage (passengers) by day through the month of October 2017

The next relationship modeled was the most frequently travelled route. Figure 2 depicts the number of passengers that travelled each route in the provided dataset. This shows that the most frequently travelled route is 701, and it can also be noted that routes servicing peak demand in certain areas (routes 8 and 13) are infrequently travelled.

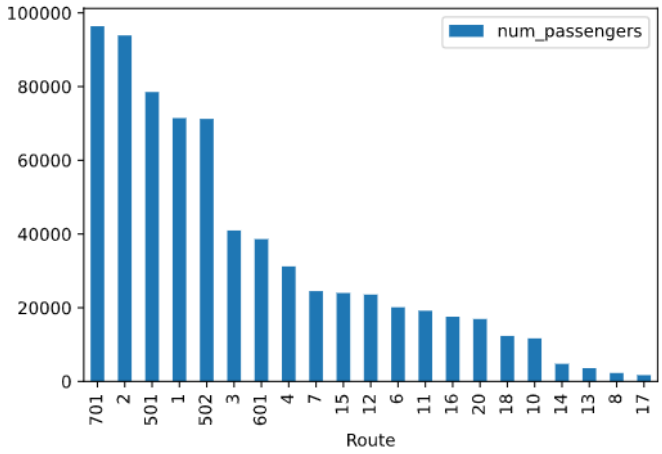

Figure 2 - Most frequently travelled routes on Kingston public transit

After synthesizing some simple metrics and establishing a level of comfort, the four questions provided in the assignment description served as motivation for any remaining analyses. These questions will be discussed in further detail and visual responses will be provided.

3.1 Accessibility of Bus Service

To address the accessibility of the bus service to all Kingston residents (via bicycle, parks, or on foot), an understanding of the city geography must be established. To visualize the position of certain services throughout the city, the datasets of interest are cycling facilities, neighbourhoods, and parking areas. A low-fidelity visual was created using these three datasets in Power BI, shown below in Figure 3.

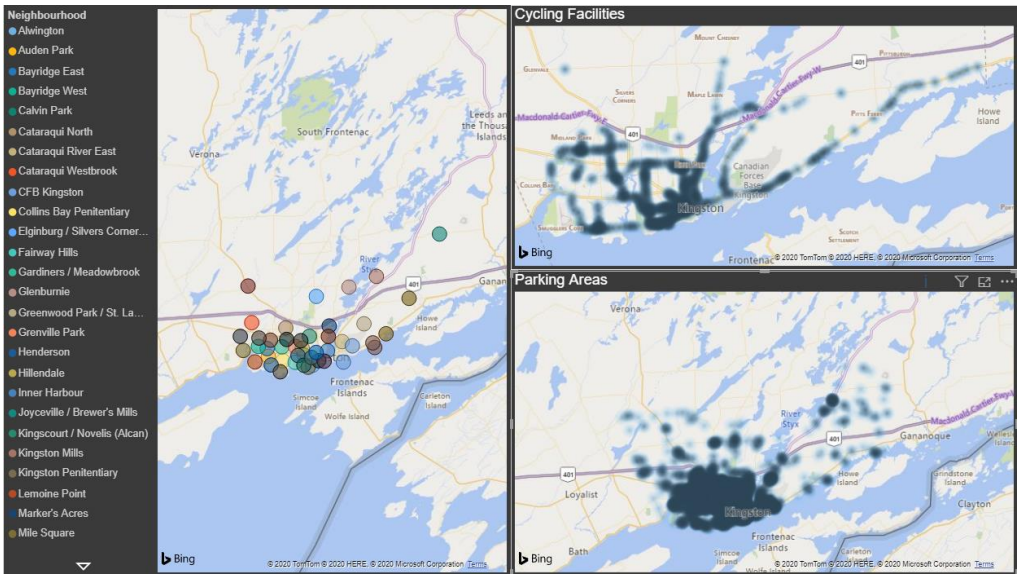

Figure 3 - Density of cycling facilities and parking areas in Kingston juxtaposed with the centroid of each neighbourhood.

Here we can see that the general density of the cycling facilities and parking areas closely follows the location of neighbourhoods. A more sophisticated visual was created using the same datasets and plotly is shown below in Figure 4, which includes the unique pick-up locations in the assignment outline.

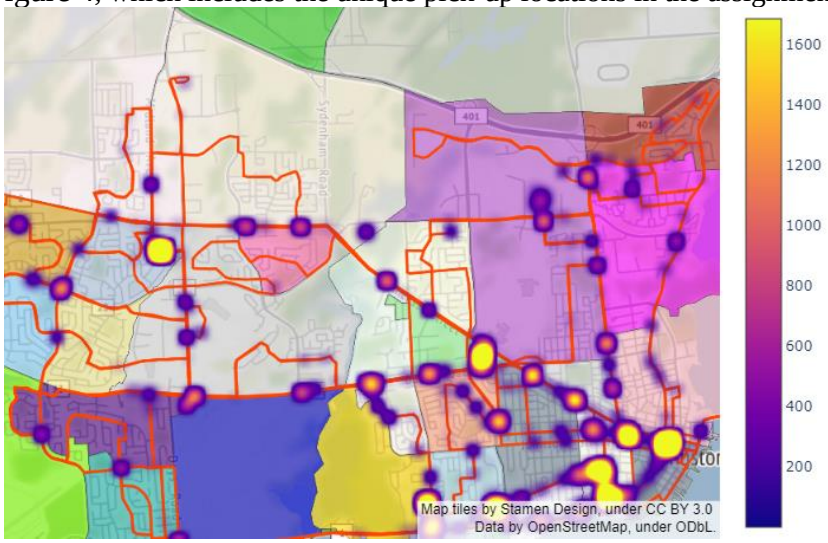

Figure 4 – Heatmap showing the density of where passengers were picked up super imposed on separate neighbourhoods and separate routes.

This image is zoomed in to provide appropriate context and shows that the high-density pick-up locations are at shopping malls (middle left), and downtown (bottom-right). From these figures, it can be suggested that the bus service is generally accessible across all Kingston neighbourhoods by bicycle, or by driving to a parking area. However, the accessibility of the bus service becomes less accessible by foot in the northern and eastern neighbourhoods, such as Sharpton, Kingston Mills, and Joyceville, where there exists a lower density of passengers, cycling and parking areas. This is further supported by Figure 5 constructed on the Open Data Kingston web application, showing the number of bus stops in each general vicinity (bubbles), super-imposed on each neighbourhood (gray), and cycling facilities (red paths).

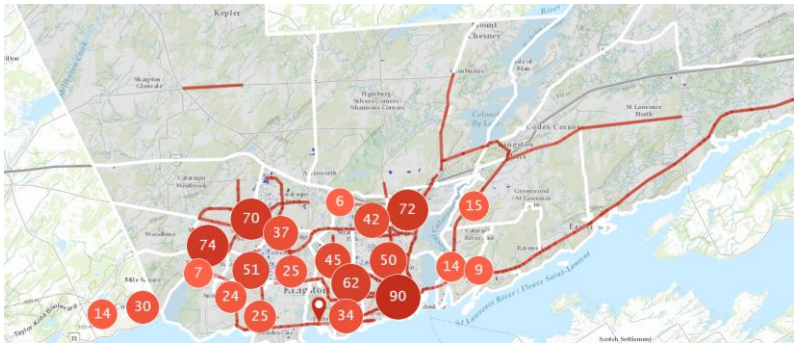

Figure 5 - Number of bus stops in Kingston super imposed on neighbourhoods.

3.2 Redundancy of Existing Bus Routes

To assess whether any bus routes are independent, the provided features must be considered. All that is provided is the time a passenger is picked up, their coordinates, and which route that bus was taking. From the given dataset, it remains unknown when passengers board the bus, but not when they exit. As such, it would be difficult to look at when busses are full, and if overlapping routes accommodate for the number of passengers needing to take a certain route to get from stop to stop. So, to make use of the provided features, the approach was to look at how much overlap existed between bus routes. To do so, the dataset was used to create a data-frame containing all the latitude and longitude coordinates each route made a stop at. Then, a correlation matrix was constructed to explore whether the pair-wise coordinates of one route correlated with the pair-wise coordinates of another. The resulting figure is shown below in Figure 6.

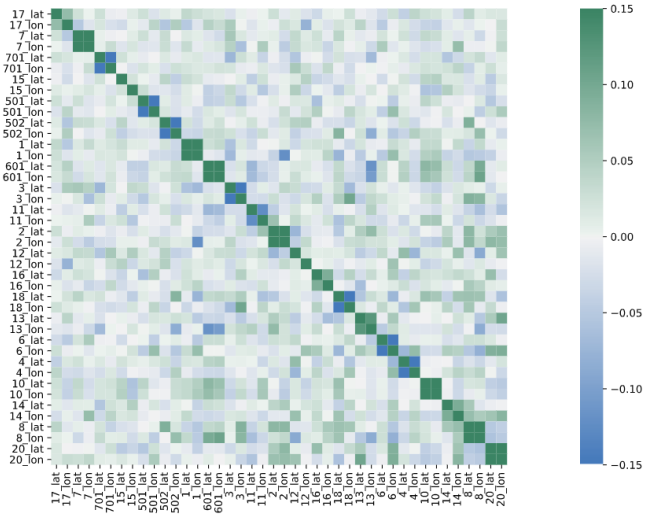

Figure 6 - Correlation of latitude and longitude values across all bus routes seen in the provided dataset.

This does not appear to provide any valuable insights. Note that a threshold had to be created on the interval (-0.15,0.15) to make the differences noticeable, as all correlation values were low—generally within (-0.3,0.3). These correlation values were also normalized. A few exceptions that are somewhat noticeable are the slightly positive correlation between routes 601 and 10. There is also a slight negative correlation between routes 601 and 13. This is clearly due to the fact that all coordinates are closely grouped in similar city regions and perhaps there is a more appropriate way to pre-process the coordinate data more effectively. A cruder approach to seeing if any bus routes are redundant and should be cancelled, was visualizing the paths each route takes, and seeing where they overlap. A visual was created using the Open Data Kingston web application to super-impose each of the 23 bus routes on top of each other. This is shown below in Figure 7.

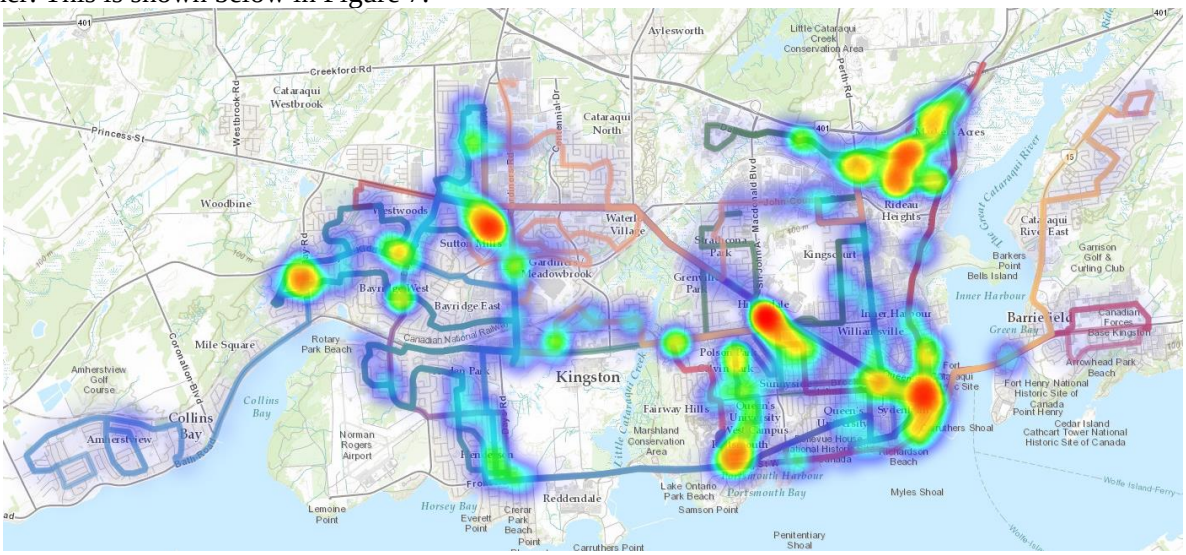

Figure 7 – Heatmap of bus stop density imposed on overlapping bus routes.

The motivation for this figure was to see where the areas with a high number of bus stops are, and if there are an appropriate number of busses to service them. For example, the east portion of the figure shows a low density of bus stops, and two independent routes to service each neighbourhood. Visually, redundancy (or overlapping routes) only appears to occur in high traffic areas. A case example of this is the Cataraqui Centre Mall, which has approximately 10 different routes passing through. From magnifying Figure 7, the result in Figure 8 shows this relatively clearly.

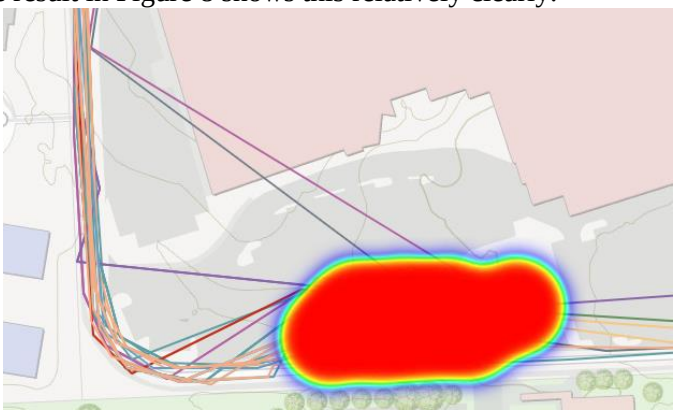

Figure 8 - A high number of bus stops servicing an area in Kingston with several passengers (Cataraqui Centre)

To assess redundancy more accurately at the mathematical level, knowing when passengers get off the bus would allow for the demand at each stop to be better understood in relation to bus capacity. In other words, since it is unknown when busses are full, it is difficult to hypothesize if routes are redundant in

high-demand areas. The ideal circumstance would be to maximize the interaction between areas of high demand, without a negative impact on accessibility. Studies have been conducted in this space using a distance-based approach in which the collection of routes was considered as a network of nodes [3]. If more time were available, and a denser feature space was provided, further work would investigate this problem using this approach.

3.3 Increasing Bus Routes to Reduce Gas Emissions

Assessing the potential reduction in greenhouse gas emissions by replacing all vehicle commuting with city busses requires following:

- The number of busses required to service the new transit commuters.
- The bounds of the routes in which the new busses will be traveling.
- The amount of CO2 emitted from the new busses.
- The current amount of CO2 emitted from vehicles commuting in Kingston.

Given the complexity of the problem, several assumptions have been made to reasonably simplify the modeling. The assumptions involve the number of commuters, approximations for cartesian distances, and physical attributes of motor vehicles. The assumptions are as follows:

1. All people work in downtown Kingston, for simplicity the downtown transfer station is taken as the destination coordinates for every person's commute.
2. Each driveway has two commuters and two cars; driveway coordinates are taken from the "driveways" dataset from the City of Kingston.
3. The proposed bus routes run from the centroid of each neighborhood to the downtown transfer point.
4. A conversion factor of 1.3 is used to compensate for the use of Euclidean distances.
5. The capacity of a transit bus is 76 [4].
6. The average fuel efficiency of the commuting vehicles is 8.9L/100km (11.286 km/l) [5].
7. Fuel efficiency of a transit bus is 3.26 mpg (1.4878 km/l) [6].
8. All cars use gasoline as fuel, producing 2.29 kg of CO2 per liter of consumption [7].
9. All transit busses use diesel as fuel, producing 2.66 kg CO2 per liter of consumption [7].

3.3.1 Number of Required Busses

The number of busses required to service all additional commuters was found by tallying the number of driveways in each neighborhood. Assuming there are two commuters per driveway, and the maximum capacity of a transit bus is 76 people, the total number of busses required per neighborhood is shown in Figure 9.

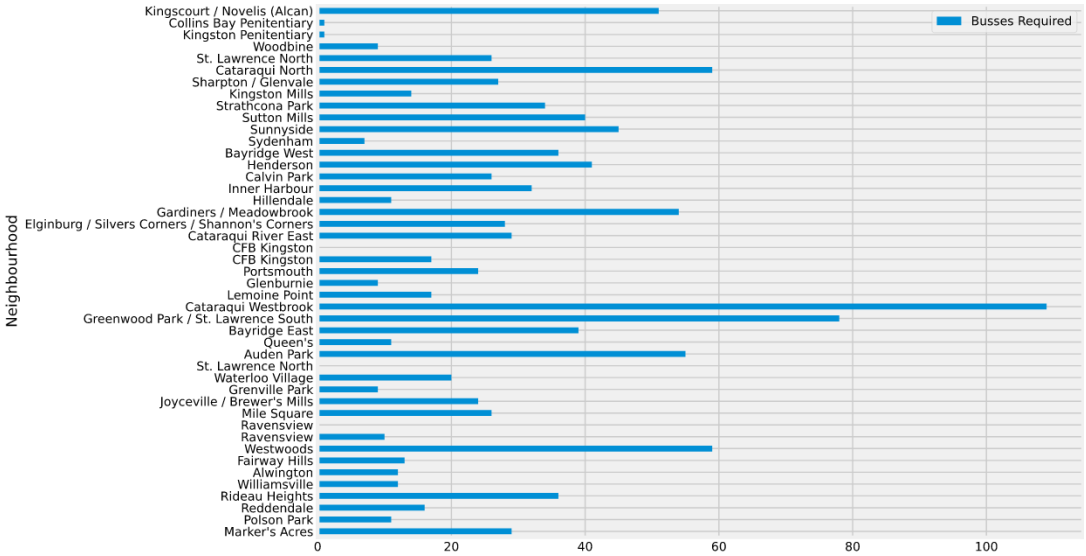

Figure 9 - Number of busses per neighbourhood required to transport every dailiy commuter in Kingston.

Interestingly, three of the neighbourhoods in Kingston contain zero driveways, and therefore require no bus routes to accommodate additional commuting.

3.3.2 Distance Traveled by Proposed Routes

To simplify the problem, the proposed bus routes will run between the centroid of each neighborhood and the downtown transfer point. This should allow all working individuals in Kingston the opportunity to commute via the transit system, instead of their regular commute. A map of the proposed routes is displayed below in Figure 10 where the line thickness is directly proportional to the number of busses required by each route.

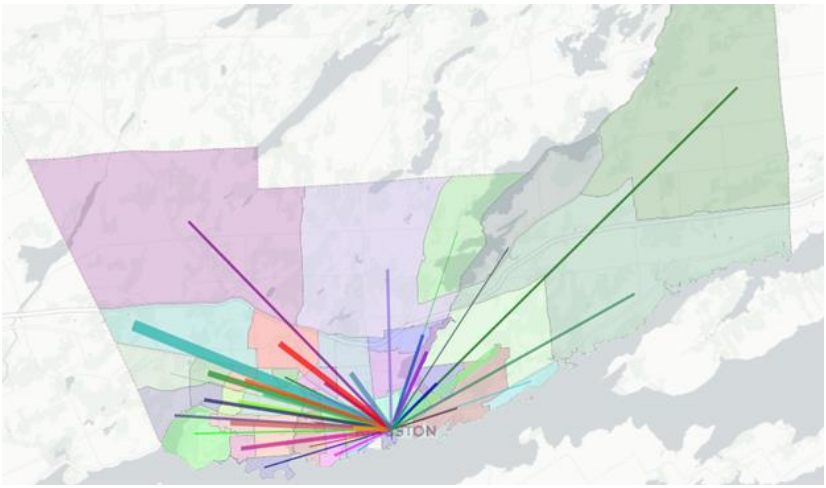

Figure 10 - Proposed transit routes to accommodate new commuters. Line thickness is proportional to number of busses.

3.3.3 Fuel Consumption and CO2 Emissions

Given the Euclidean distance of each proposed bus route (d), the cartesian conversion factor (C), the number of busses required to service peak demand (n), the fuel efficiency of a transit bus (E), and the CO2 emissions for each liter of diesel fuel consumed (D), the total greenhouse gas emissions per day (2 trips) of the proposed bus routes was calculated using the formula below.

$$2 \cdot \frac{\sum d \cdot C \cdot n}{E} \cdot D$$

The greenhouse gas emissions for the current state was calculated similarly, however the distance was found by summing the Euclidean distances of every driveway to the downtown transfer point. Additionally, the fuel was assumed to be gasoline, the fuel efficiency of the Canadian vehicle fleet was used, and n=2 (two cars per driveway). A comparison of the estimated fuel consumption and greenhouse gas emissions of current state and the proposed solution is shown in Figure 11.

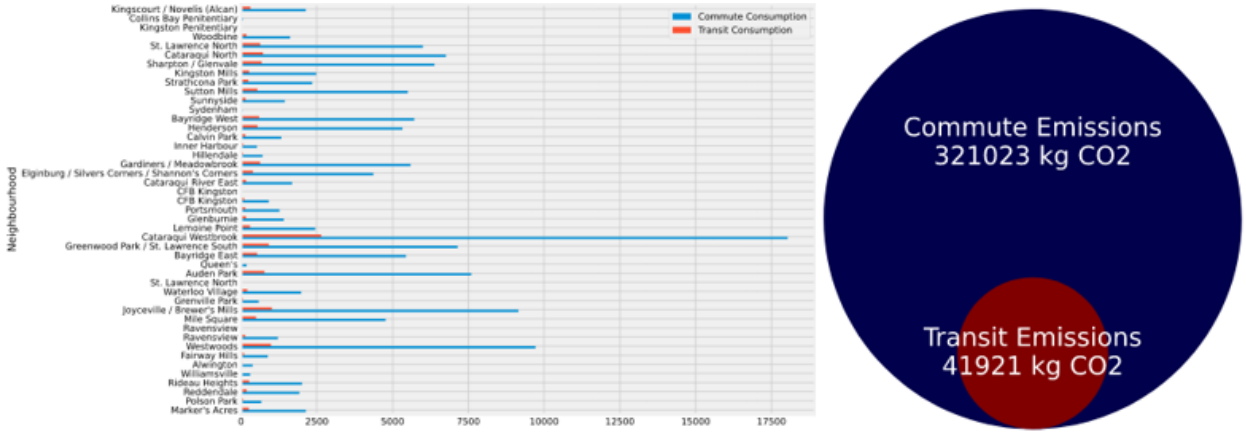

Figure 11 - Breakdown of daily fuel consumption by neighbourhood (left) and total daily CO2 emissions (right) for vehicle commuting and proposed transit commuting.

3.4 Additional Interesting Information

While it was indicated that 8% of the records in the provided dataset were from malfunctioning GPS antennae, it was found that among these records, some were either latitudinally or longitudinally displaced from the recorded stop location. A simple visual using Power BI was created to show this behaviour—it creates a triangle among erroneous data.

Figure 12 - Visually displaying errors indicated in the provided dataset

While this would not be helpful in better-planning bus routes, this could indicate the ability to collect cleaner data with this behaviour in mind. Other than this interesting error, the only other interesting thing of note was the fact that routes used in peak demand areas (8 and 13) did not service a high amount of passengers (Figure 2), and this was likely due to the limited use and hours of operation (see Figure 13 below).

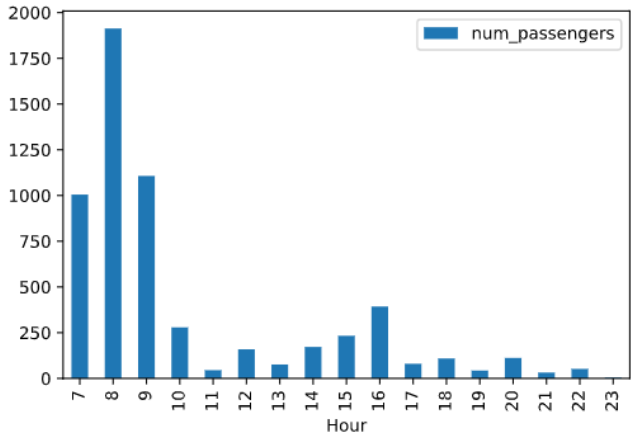

Figure 13 - Usage of routes 8 and 13 by hour, where they are used to service peak demand

This is interesting that the use of these routes appears to be during commuting hours (morning and early evening). This appears to be a concept already implemented by the City of Kingston to plan the public transportation network more effectively.

Conclusion and Assessment of Findings

This assignment features an analysis of different aspects of the Kingston Public Transit using the provided dataset of passengers in October 2017 in conjunction with datasets from the Open Data Kingston portal. The accessibility of the bus service across all neighbourhoods was first analyzed from a geographic perspective. This allowed for the density of stops made to be juxtaposed with cycling facilities, parking areas, and neighbourhoods to visualize the neighbourhoods in which the bus service is less accessible. Next, the redundancy of bus routes was analyzed by correlating their coordinates with each other to identify overlap. With few conclusive redundancies discovered, future approaches would ideally have access to when passengers leave the bus, in addition to performing a distance-based analysis instead of pure correlation and geographic analysis. Another analysis was performed on the hypothesis that all working Kingston residents worked downtown, and that public transit would replace their personal daily commutes. With many assumptions made, the daily CO2 emissions from commutes would be 321,023 kg per day, versus 41,921 kg per day by adding more bus routes as an alternative. Lastly, a few interesting details in the dataset were that the erroneous coordinates from the malfunctioning GPS antennae that were not (0,0) were displaced either longitudinally or latitudinally from the actual coordinates. Also, routes 8 and 13—used in peak demand areas—are mainly used during commute hours, suggesting an initiative from the City to better plan public transit has already been put in place.

References

[1] City of Kingston, "Welcome -- Open Data Kingston," [Online]. Available: https://opendatakingston.cityofkingston.ca/pages/welcome/. [Accessed 17 September 2020].

[2] Microsoft, "Downbloads | Microsoft Power BI," 2020. [Online]. Available: https://powerbi.microsoft.com/en-us/downloads/. [Accessed 19 September 2020].

[3] D. E. M, S. Li and A. T. Murray, "Identifying bus stop redundancy: A gis-based spatial optimization approach," *Computers, Environment and Urban Systems,* vol. 36, no. 5, p. 445–455, 2012.

[4] Fantastic Offence, "City | Transit Buses Dimentions and Drawings | Dimentions.com," 2020. [Online]. Available: https://www.dimensions.com/element/city-transit-buses. [Accessed 28 September 2020].

[5] Canada Energy Regulator, "CER Market Snapshot: How does Canada rank in terms of vehicle fuel economy?," [Online]. Available: https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/market-snapshots/2019/market-snapshot-how-does-canada-rank-in-terms-vehicle-fuel-economy.html#:~:text=In%202017%2C%20Canada%27s%20average,kilometres%20(L%2F100km).&text=In%20comparison%2C%20fuel%20. [Accessed 1 October 2020].

[6] US Department of Energy, "Alternative Fuels Data Center: Maps and Data - Average Fuel Economy by Major Vehicle Category," [Online]. Available: https://afdc.energy.gov/data/10310. [Accessed 1 October 2020].

[7] Natural Resources Canada, "AutoSmart Factsheet," 2014. [Online]. Available: https://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/oee/pdf/transportation/fuel-efficient-technologies/autosmart_factsheet_6_e.pdf. [Accessed 1 October 2020].