Big Data Science and Engineering Solution for Transit Performance Analytics

Carson K. Leung(), Ngoc Pham, Yixi Wu, Mohammadafaz Munshi, Vrushil Patel
University of Manitoba, Winnipeg, MB, Canada
Carson.Leung@UManitoba.ca

***Abstract*—Bus transit is an important component of the day-to-day activities of many people. It provides a cost-effective and convenient way for individuals to commute to work, school, and other destinations. Bus transit is a vital mode of transportation for students, as it enables them to commute to and from their educational institutions. Delays in bus schedules can have severe consequences—such as missing exams, meetings, and other important engagements—in daily activities of city residents. Hence, in this paper, we present a data science solution for mining and transportation analytics on public transit on-time performance data. Knowledge discovered from these data helps improve public transit performance, and thus enhance rider experience in a city. This helps build a smart city. To elaborate, our solution adapts frequent pattern mining, which identify uncover variations in transit performance across various neighborhoods. Through identifying significant findings, we establish correlations to determine the factors contributing to bus delays in specific areas. Improving the bus arrival or departure time can have a positive impact on the overall usability and attractiveness of bus transit for commuters since people are more likely to use buses when they can rely on them to arrive on time and get to their destinations promptly. Our solution also provides users features to visualize the discovered knowledge about the bus departure time in all the neighborhoods at different times of the day. Evaluation results on real-life public transit data from a Canadian city demonstrated the practicality of our data science solution towards the building of smart city.**

***Index Terms*—Bus transit, Data mining, Relational database, Neighborhoods, Frequent pattern mining, Association rules, Transit performance, Analysis, Transportation analytics, Smart city.**

I. Introduction

Maintaining good performance for the public transit system is essential for many people. For instance, in many Canadian cities (e.g., Winnipeg, Toronto, Montreal, Ottawa), the weather is cold in almost one-third of the year and the cities are covered in snow. On top of that, a large percentage of city residents do not have convenient access to public transit already [1]. This means that bad transit performance will result in the city residents suffering longer in the cold winter. The motivation behind conducting the data mining of Transit On-time performance data to find neighborhoods that report late or early bus departure times is to address the issue of unreliable public transportation. Late or early bus departure times can cause inconvenience and frustration for commuters, which can have a negative impact on their daily lives. Our goal is to use data mining—in particular—frequent pattern mining to identify areas that experience frequent bus deviating from scheduled departure time and try to uncover factors contributing to these events. We can use the interesting association rules from the mining to improve the operation of bus transit. This can contribute to the enhancement of public transit system in these cities resulting in a more efficient and reliable public transportation system that meets the needs of its users.

In this paper, our *key contributions* include our data science solution for mining and transportation analytics on public transit on-time performance data. Our solution discovers neighborhoods with poor transit performance. It also identifies the routes that are responsible for the bad performance in the neighborhoods based on frequent pattern mining. It provides users with visualizations of the data and analyze patterns—such as the relationship between the density of bus stops and transit performance. Evaluation on real-life public transit data from a Canadian city demonstrated the practicality of our data science solution towards the building of smart city.

The rest of the paper is structured in four sections. Section II explains the background and related works. Section III describes our data science solution. Evaluation results are reported in Section IV. Finally, Section V concludes our work.

II. Background and Related Work

In today's time, a lot of data is accessible to us. Transportation data is one of them. With the advances in technologies, this data is used for mining and researching purposes [2]. Many related works analyze public transportation data to predict bus arrival or departure times. A lot of research has been conducted to predict bus schedules but very few of them produced accurate data. For instance, Audu et al. [3] predicted delays in bus arrival time in Toronto city in Canada. Researchers used a number of machine learning techniques, including decision trees, random forests, and artificial neural networks, to analyze and predict traffic patterns in urban areas. Different cities were studied for comparison and to find ways to reduce traffic congestion.

Another existing work is mining data of bus transit during the pre-COVID-19 era and COVID-19 era [4]. The researchers analyzed the open public transit data for Winnipeg city to see how often the buses were full and could no longer take more passengers, which were referred as a 'pass-up'. Frequent and sequential pattern mining were used to mine information about the occurrence of pass-ups during different times of the year, month, and week, and hence to find out ways to reduce pass-ups.

In contrast, our goal is to study the bus transit data and compare neighborhoods based on the on-time departure of buses. Buses are frequently late because of many factors such as weather conditions, time of day, time of the year, pandemics, and type of neighborhood such as residential areas or industrial areas. Using the frequent pattern mining, we come up with interesting association rules that point out neighborhoods where buses frequently deviate from its scheduled times. In our evaluation, we also compare delays in bus departure times in 2021 (during the COVID-19 pandemic) and 2022 (after the COVID-19 pandemic). In 2021, people stay indoors due to safety protocols, and very few people use bus transit to travel. Buses tend to delay when in-person activities pick up and more people use public transit which is evidenced in the year 2022 after the COVID-19 pandemic.

III. Our Data Science Solution

We design and build a data science solution by capturing important data and information about public bus performance. It usually includes:

- Bus Stop Number;
- Route Number;
- Route Name;
- Route Destination, represents the direction of the route;
- Day type, represents either weekdays or weekends;
- Scheduled Time, represents the expected scheduled departure time;
- Deviation, represents the deviation between scheduled departure time and actual departure time in seconds (so that negative deviations indicate late buses and positive deviation indicates early buses);
- Location, contains geometric datatype POINT (latitude, longitude) representing the GPS of the bus stop.

A. Preprocessing Data

In the first phase of preprocessing data, we create a (relational) database with 7 tables: 5 tables which each has a primary key, and 2 joined tables which are datasets of Transit On-time Performance with reduced number of columns. We find it beneficial to have a well-structured database, as it aids us in effectively visualizing the data at a later stage:

- `route` table with route_id as a primary key, attributes include: route_name;
- `route_destination` table with route_id as a primary key, attributes include: route_destination;
- `stop` table with stop_id as a primary key, attributes include: latitude, longitude, geometry, nbh_id, ward_id;
- `neighborhood` table with nbh_id as a primary key, attributes include: nbh_name, geometry;
- `ward` table with ward_id as a primary key, attributes include: ward_name, geometry, councillor;
- `2021` joined table comprises of 12 archive monthly datasets of Transit On-time Performance in 2021;
- `2022` joined table comprises of 12 archive monthly datasets of Transit On-time Performance in 2022.

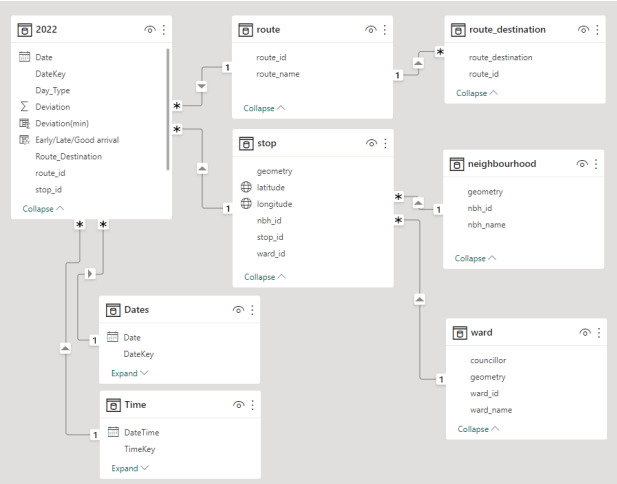

Fig. 1. Schema using on-time performance datasets.

TABLE I
Time-of-Day Category

Category	Condition	Hour Range
Morning	05:00 - 09:00	≥ 5 && < 9
Work hours	09:00 - 16:00	≥ 9 && < 16
Peak hours	16:00 - 19:00	≥ 16 && < 19
Evening	19:00 - 23:00	≥ 19 && < 23
Night	23:00 - 05:00	≥ 23 || < 5

The second phase is to perform an additional transformation on the Transit On-time Performance Data. To improve the results of our frequent pattern mining, we decide to segment our dataset based on five different time-of-day categories. Our main objective in using frequent pattern mining is to identify frequent correlations within the data, specifically patterns that indicate when buses are usually late. By correlating the time of day with instances of bus lateness, we hope to identify which times of day and routes may be performing poorly in certain areas. Our dataset includes a column labeled Time, which uses the ISO 8601 format. We create a Python script that leveraged the `datetime` and `pandas` libraries. Through this script, we are able to extract the time-of-day from the Time column and assign each entry to one of the five predefined categories as outlined in Table I. By doing so, we are able to generate a more granular understanding of Transit On-time Performance that accounts for the time-of-day, as well as other factors such as routes and geographical locations.

We classify the bus departure times into three categories based on the deviation attribute. As shown in table II, we label any bus departure as ”Late” if its deviation from the expected time is late more than or equal to -180 seconds (3 minutes). If the deviation is more than +60 seconds (1 minute), we categorize it as an ”Early” departure. Otherwise, we consider the deviation is a ”Good Arrival”. These categories are consistent with the metrics used by some cities (e.g., Canadian city of Winnipeg) to determine whether a bus is late or early. We utilize the departure classification to identify

TABLE II
DEVIATION CATEGORY

Category	Condition	Deviation (sec)
Early	More than 1 min	> 60
Good Arrival	3 min late to 1 min early	> -180 && $<= 60$
Late	More than 3 min	≤ -180

Algorithm 1 Define a neighborhood of a stop

1: Read in 1st geodataframe: neighborhood shapefile
2: Keep nbh_id, nbh_name, geometry (MULTIPOLYGON datatype) columns
3: Read in 2nd geodataframe: stop csv file
4: Keep stop_id, longitute, latitute columns
5: Use `geopandas` to create geometry (POINT datatype) column based on longitude, latitude
6: Use `geopandas's sjoin()` function to spatial join two geodataframes, "within" as predicate
7: Find a stop and its associated neighborhood by stop_id

instances of bus deviation and pass the resulting data as input to our algorithm for further analysis.

To identify neighborhoods with inadequate public transportation, we associate the bus stops with their respective neighborhoods. As previously discussed, the Transit On-time Performance Data supplies an attribute indicating bus stop location in a POINT geometry format (latitude, longitude). Furthermore, the Winnipeg Open Data Portal offers a Neighborhood dataset using a different geometry datatype, MULTIPOLYGON, which outlines neighborhood boundaries. We employ Python's `geopandas` package to spatial merge the two geodataframes. The pseudo codes are provided in Algorithm 1.

Figure 2 displays a map of Winnipeg which is segmented into different wards and neighborhoods, with all stops marked in red dots. Algorithm 1 is employed to determine the location of a specific stop, Figure 2 shows an example of stop #50140 denoted with a black dot. The algorithm identifies that this stop is situated in the Southland Park neighborhood, which is assigned an identifier 1196.

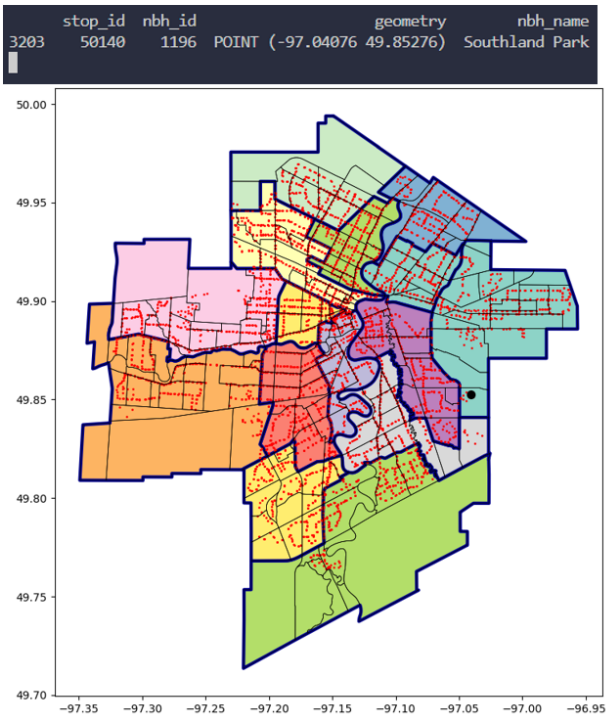

Fig. 2. Winnipeg map with 237 neighborhoods, 15 wards and 5218 bus stops. Stop #50140 is located in Southland Park neighborhood shown as a black dot.

Algorithm 2 Get neighborhood late ratio

1: **for** each record in dataset **do**
2: **if** record is late **then**
3: late[neighborhood] ← late[neighborhood] + 1
4: **end if**
5: total[neighborhood] ← total[neighborhood] + 1
6: **end for**
7: **for** each neighborhood **do**
8: ratio[neighborhood] ← late[neighborhood] / total[neighborhood]
9: **end for**
10: Output ratio[neighborhood]

B. Processing Data

After cleaning our data, we perform frequent pattern mining to discover which {neighborhood, routes, time-of-day, route destination, "Late"} frequently occur. We first mine the frequent pattern month by month because each month has records ranging from 4 million to 8 million rows. This method helps us identify the frequent pattern in each month so that later on, we can visualize specific routes and areas across the whole year. We set the minimum support to 0.15% and the minimum confidence to 30%. We generated lots of association rules and some interesting rules are evaluated in Section IV.

It captures neighborhoods with a high frequency of late buses, ignoring the individual ratio for each neighborhood itself. To further elaborate, neighborhoods like South Portage has more than 40 bus routes that go through it, while some residential neighborhoods, such as Bridgwater, have less than 5 routes that go through it. Frequent pattern mining will ignore those residential neighborhoods because the frequency of those neighborhoods being late is low, but it does not signify the probability of buses being late in those neighborhoods is low. Therefore, we also create scripts and visualization to show the ratio of late buses in each neighborhood.

Additionally, we also reuse this algorithm for bus routes by changing our variable to route_id, which identifies routes that have a low number of schedules but a high ratio of being late. Notice that Winnipeg Transit does provide route_id ratio graph for the past year in their website. However, creating our own graph also enable us to filter the ratio by different category such as by {time-of-day, route destination, Month}, which greatly help us analyze the data in depth.

C. Visualizing Data

Our solution also visualizes data for easy analysis of transit data. Figure 3 and Figure 4 are some of the graphs, which help users visualize and analyze the data.

IV. Evaluation

To evaluate our solution, we applied it to real-life data from the Canadian city of Winnipeg. All buses operated by Winnipeg Transit are equipped with a Global Positioning System (GPS) that records the departure time of the bus from each stop. The system logs the exact time the bus is scheduled to leave the stop and compares it to the actual time it departs. This Transit On-time Performance Data is available for download and updated monthly.

In April 2020, the Winnipeg Transit system has changed significantly due to an introduction of a new rapid transit line namely BLUE[1]. To accommodate the BLUE line, many other routes have been decommissioned and some have been replaced with new ones. Hence, we used 24 monthly archive datasets comprised of two full years 2021 and 2022 to strengthen the reliability and accuracy of our analysis.

A. Visual Knowledge Discovery with Column Charts

When applying our solution to real-life data. We observed the following: In 2022, East Elmwood, Victoria Crescent and Bridgwater Centre neighborhoods are the top three neighborhoods having the highest late bus percentage, which are more than 40%, as compared to its early buses and on-time buses, as depicted in Figure 3. These are mainly residential neighborhoods and we will do a more in-depth analysis in the below section of our paper. Moreover, in the same year, Figure 4 shows that there are four routes namely 89, 86, 45, 56 having a high late bus ratio with more than 40% each. Interestingly, these routes mainly pass through the downtown and Transcona residential areas.

B. Association Rules

Figure 5 shows some of the examples of interesting association rules that we find by running frequent pattern mining for each month. We focus on filtering interesting association rules that have Late/Early as consequences. In this way, we can infer what frequent patterns usually cause the Late or Early buses. We interpret our finding by the following: for example, from the third association rule in the figure, we can infer that if you are taking Route 47 during work hours (9 am to 4 pm) to the University of Manitoba via Downtown during 2022 September, there is a 49.78% chance that your bus is going to be late. In September 2022, if you are trying to go downtown via the Crescent Park neighborhood by taking Route 60, there is a 68.85% chance that your bus is going to be late, which is an impressive bad performance. Although we do not show all the results generated from frequent pattern mining, there are some single frequent items such as ”neighborhood: South Portage”, ”neighborhood: University” and ”neighborhood: St. James Industrial”. These are single frequent patterns that occur over and over for each month in the result. These single frequent items make sense because they are popular areas that require public transit services, thus becoming high frequency of late buses neighborhoods to go through. On top of that, there are some association rules that have regular occurrences for multiple months, all of which are related to the University, Downtown and Polo Park (which is near St. James Industrial).

Interestingly, South Portage is a special case. It appears in all the ”Late”, ”Good Arrival” and ”Early” frequent patterns. We believe that South Portage serves as the central bus hub for downtown, with buses travelling in both directions, making it the first and last stop for these buses. Therefore, South Portage increases its chance of having early buses (first stop) and late buses (last stop). With the interesting association rules that we

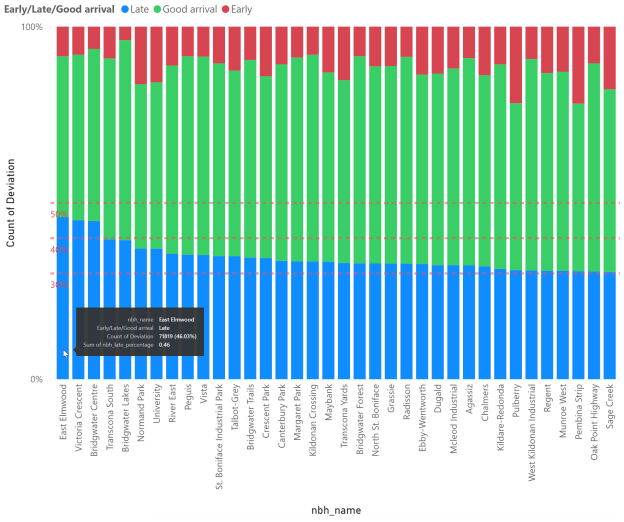

Fig. 3. 2022 Early/Good Arrival/Late ratio for neighborhoods with the top percentage of having late buses.

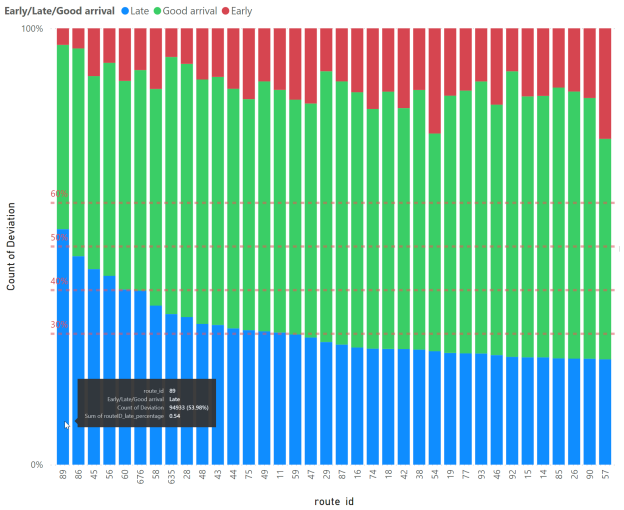

Fig. 4. 2022 Early/Good Arrival/Late ratio for routes with the top percentage of having late buses.

[1]https://info.winnipegtransit.com/en/service/blue-rapid-transit/

```
LHS: ['Work Hours', 'Centre Street via Bridgwater', '676'] -- RHS: ['Late']
Support: 0.0015147442632755385
Confidence: 0.4316026788674491
2022 September

LHS: ['Neighbourhood: Crescent Park', 'Downtown', '60'] -- RHS: ['Late']
Support: 0.0016706635529351807
Confidence: 0.688501007001055
2022 September

LHS: ['Work Hours', 'University of Manitoba via Downtown', '47'] -- RHS: ['Late']
Support: 0.0039007748257834 65
Confidence: 0.4977875449172927
2022 September

LHS: ['Downtown', 'Work Hours', '60'] -- RHS: ['Late']
Support: 0.0021947153987610084
Confidence: 0.36162042814629936
2021 July

LHS: ['Afteroon Peak Hours', 'Neighbourhood: South Portage'] -- RHS: ['Late']
Support: 0.0028706244821872 41
Confidence: 0.3376383262187661
2021 September

LHS: ['77', 'Polo Park', 'Afteroon Peak Hours'] -- RHS: ['Late']
Support: 0.0020471718124830 85
Confidence: 0.4667987073908058
2021 November
```

Fig. 5. Interesting Association Rules Example

find, we narrow down the area and bus routes that we need to dive deeper into and use visualization to analyze and visualize our data.

C. Analysis Results: Early and Late Deviations

To carry out evaluations in the context of public transit, several obstacles need to be addressed. Our collective effort involves deciding if we should treat both Late and Early deviations as subpar performance or solely focus on Late deviations. Furthermore, we strive to identify the appropriate aggregation method for our analysis, choosing between the Sum of Deviations or the Count of Deviations.

We classify deviation into three categories—i.e., Early, Good Arrival, and Late—as outlined in Section III-A, aligning with the Winnipeg Transit grouping metrics. As passengers, missing a bus due to an early departure also signifies inadequate performance. If the bus is delayed, passengers may also arrive late at their intended destination, possibly at another stop for transferring to a different bus line. Hence, passengers potentially will miss the subsequent buses, resulting in a domino effect. Therefore, optimal performance occurs when the bus leaves within the Good Arrival window. Consequently, we will take into account both Early and Late deviations in the upcoming outcomes for our analysis.

We aim to validate our analysis by comparing some results with the paper "Data Mining on open public transit data for transportation analytics during Pre-COVID-19 Era and COVID-19 Era." To do this, we first apply the Sum of Deviation. The results reveal in Figure 6 that the highest poor performance across all neighborhoods and routes occurred in December 2022 (19.5 million minutes) and November 2021 (10.9 million minutes). This aligns with the paper's prediction that pass-ups increase during winter, likely due to severe weather conditions [4], thus resulting in a decrease in transit performance. Additionally, we observe that in 2022, as in-person activities resumed and COVID-19 restrictions eased,

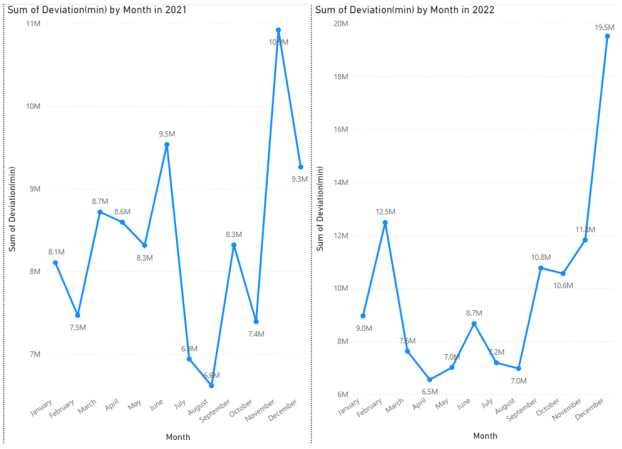

Fig. 6. 2021 Sum of Deviation by Month (left) and 2022 Sum of Deviation by Month (right).

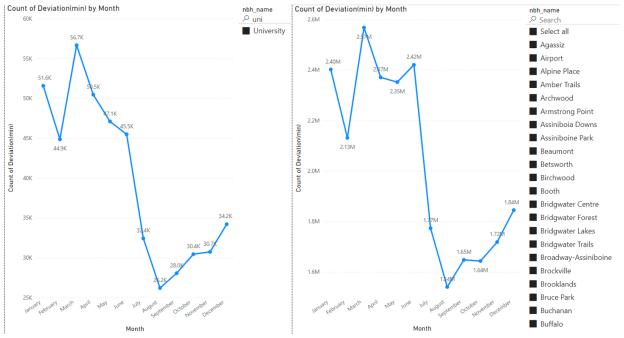

Fig. 7. 2021 Count of Deviation by Month in all neighborhoods and in University.

the highest poor performance figure nearly doubled compared to 2021.

Regarding the Count of Deviation, besides the South Portage neighborhood, we believe that the University neighborhood may be an area with poor transit performance, which led to the introduction of the BLUE line in April 2020. Figure 7 demonstrates that the Count of Deviation pattern in the University neighborhood closely resembles that of all neighborhoods, suggesting it plays a significant role in contributing to the overall Count of Deviation. However, this similarity is not evident when using the Sum of Deviation as the measurement unit, as depicted in Figure 8. Hence, we place greater confidence in the Count of Deviation compared to the Sum of Deviation and will employ the Count of Deviation for further analysis.

Having determined that the Count of Deviation is the suitable metric for our frequent pattern analysis, we successfully identify the top five neighborhoods of subpar transit performance, also known as the highest Count of Deviation, as depicted in the bar charts in Figure 9 and Figure 10. Intriguingly, the top five remain consistent for both 2021 and 2022:

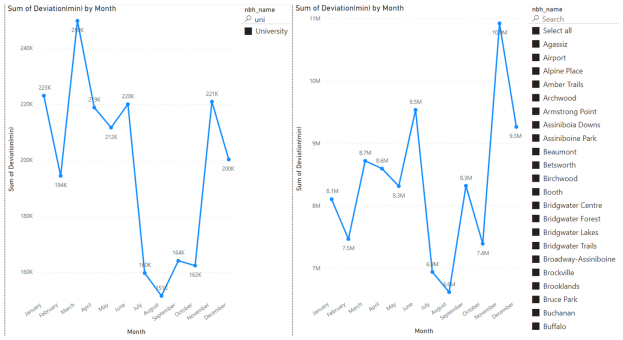

Fig. 8. 2021 Sum of Deviation by Month in all neighborhoods and in University.

- South Portage
- University
- St. James Industrial
- Regent
- Windsor Park

Our objective is to identify the particular routes responsible for poor transit performance in neighborhoods. We focus on the top five neighborhoods, investigating and examining the routes with the highest Count of Deviation as shown in table format in Figure 9 and Figure 10.

Our observations for 2021 can be found in Table III and Fig. 9. Those for 2022 can be found in Table IV and Fig. 10.

Interestingly, while South Portage seems to have the poorest transit performance, Route 19 - Windsor Park via Drake consistently exhibits the highest Count of Deviation in both 2021 and 2022. This indicates subpar transit performance in the Windsor Park neighborhood, suggesting it requires transit improvements more urgently than South Portage. Another notable observation is that five other routes pass through the Windsor Park neighborhood, namely Routes 16, 50, 57, 75, and 96. However, enhancing mainly Route 19 - Windsor Park via Drake could substantially improve the transit situation in the Windsor Park neighborhood.

During both 2021 and 2022, Route 14 - Ferry Road and Route 47 - University of Manitoba via Downtown persistently stand out as the primary contributors to the highest Count of Deviation in their respective neighborhoods, St. James Industrial and Regent. This suggests that enhancing these two routes could lead to an improvement in neighborhood transit performance.

According to Statistics Canada, in 2016, 34.6% of Winnipeg commuters used sustainable transportation, while 13.6% utilized public transit [5]. This indicates that the primary commuting methods in Winnipeg metropolitan area involve more public transportation as compared to cars and heavy trucks [5]. The BLUE rapid transit line was introduced in April 2020 to increase capacity and maintain consistent service to and from downtown. Although some routes were decommissioned due to the BLUE rapid transit line, coordinating a single route is easier than managing multiple routes entering and leaving downtown[2]. Nonetheless, Winnipeg Transit took this into account and extended Route 47 to Pembina to compensate for those decommissioned routes, such as Route 160, 170, 180. According to Singh et al. (2022), the implementation of the new Bus Rapid Transit (BRT) service in Winnipeg resulted in increased accessibility to essential services across the entire corridor, which is an advantage for residents during the COVID-19 pandemic [6]. The positive impact of the BLUE rapid transit line is more evident in 2021 when it was the main contributor to the Count of Deviation in the South Portage area, however, as commuters adjusted to the BLUE line, it was no longer the main factor in 2022, with Route 60 taking over as the primary contributor to bad performance. Route 60 appeared three times in two subpar transit performance areas (South Portage and University) in 2022 and seems to perform poorly in both directions, whereas it was not problematic in 2021. Since Route 47 - University of Manitoba via Downtown was a newly extended route in 2021, commuters were still getting accustomed to it, resulting in poor performance in the University area. However, by 2022, Route 47 performance improved, and again Route 60 - University became the problematic route. In April 2020, Winnipeg Transit attempted to increase the frequency of Route 60 during weekdays, but no significant improvement has been observed. Thus, it is reasonable to conclude that enhancing Route 60 in both directions would substantially improve transit performance in both Downtown and University areas.

D. Analysis Results: Early and Late Deviation Ratio

Another perspective that we analyze the data is to look at the late ratio of neighborhoods. Through the visualization, we identify neighborhoods that have a high ”Late” ratio of buses. For the South Portage neighborhood, its ratio is quite good with just 24.56% of having late buses. As we mentioned before, South Portage frequency is high because lots of bus routes pass through South Portage. Yet having a relatively low ratio also makes sense because it is at the heart of the downtown area and the city knows that and handles it well. Interestingly, we find that the late bus performance in South Portage is quite similar to the late bus performance in the whole city for each month, as we can see in Figure 11.

However, some neighborhoods perform badly in terms of the ratio of late buses. For example, one neighborhood that stands out to us is Bridgwater. From Figure 3 in Section III-C, we can see that Bridgwater Center neighborhood has a roughly 45% chance of having late buses for the year 2022, which is a really bad performance. From a CBC News article, a student who lives in Bridgwater ”eventually gave up on riding the bus and started driving to campus from Bridgwater, shelling out more than $600 a year to park.” [7]. We discover that, for the top 10 neighborhoods having late buses in terms of ratio, 8 out of 10 are residential neighborhoods (except University and Transcona South). There are 7 out of 8 of those residential neighborhoods have less than 5 different bus

[2]https://info.winnipegtransit.com/en/service/blue-rapid-transit/

TABLE III
2021 BAD TRANSIT PERFORMANCE NEIGHBORHOODS & BAD ROUTES

Neighborhood	Box color	Route ID and Route Destination
South Portage	Purple	BLUE - Downtown, 60 - Downtown
University	Green	60 - Downtown, 47 - UofM via Downtown
St. James Industrial	Red	14 - Ferry Road
Regent	Blue	47 - UofM via Downtown
Windsor Park	Black	19 - Windsor Park via Drake

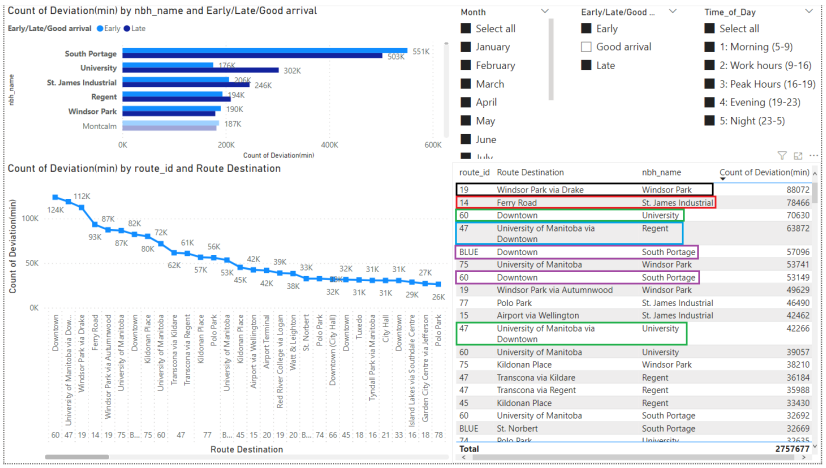

Fig. 9. 2021 Neighborhood ranking with routes.

TABLE IV
2022 BAD TRANSIT PERFORMANCE NEIGHBORHOODS & BAD ROUTES

Neighborhood	Box color	Route ID and Route Destination
South Portage	Purple	60 - Downtown, BLUE - Downtown
University	Green	60 - UofM via Downtown, 60 - Downtown
St. James Industrial	Red	14 - Ferry Road
Regent	Blue	47 - UofM via Downtown
Windsor Park	Black	19 - Windsor Park via Drake

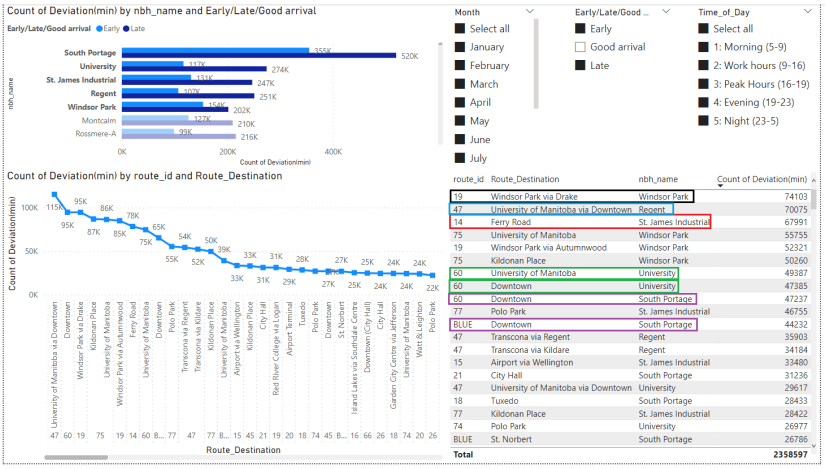

Fig. 10. 2022 Neighborhood ranking with routes.

routes, while only Peguis has 6 different bus routes. Although we are not sure what the actual cause is, we can say that residential neighborhood tends to perform poorly. We presume that the city pays less attention on improving the residential neighborhoods, as they may assume that people living in those areas own cars and are less reliant on public transit.

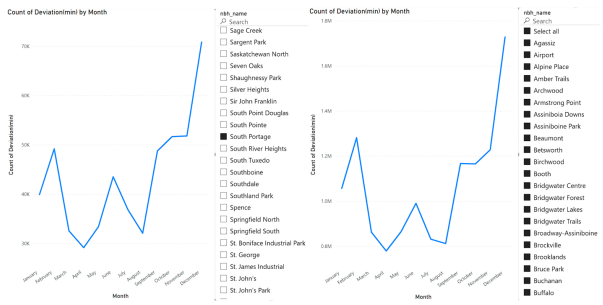

Fig. 11. 2022 South Portage Late Buses Count (left) vs Winnipeg Late Buses Count (right) by month

Nevertheless, the combination of few available buses and a high likelihood of encountering late buses is quite unpleasant to residents who live in a residential neighborhood and rely on public transportation. Moreover, it is worth noting that the lack of reliable and accessible public transportation options in residential neighborhoods can have a significant impact on the well-being and quality of life of those who reside there, particularly for individuals who do not own a car, such as the elderly, low-income families, and people with disabilities. This reinforces the need for greater investment in public transportation infrastructure and services in residential areas, to ensure that all members of the community have access to safe, reliable, and affordable transportation options.

We find another interesting pattern when analyzing bus deviation across neighborhoods in both 2021 and 2022: there is a higher likelihood of late buses during Afternoon Peak Hours compared to Morning Peak Hours, as illustrated in Figure 12. While the reason for this behavior is uncertain, it is possible that the city prioritizes morning commutes over afternoon ones since it is important for people to get to work on time. However, not all individuals work a standard 9 to 5 schedule, and this inconsistency can be frustrating for those who rely on public transportation during afternoon peak hours. Another guess is that it might be because people usually get off from work at the same time and popular areas like downtown and St. James st are crowded with people. Buses tend to start late at these places, resulting in delaying their arrival time along the way. However, whatever the reasons are, it highlights the need for improvements in public transit performance during these times.

E. Analysis Results: Bus stop density

Based on our analysis using visualization, while residential areas have fewer routes, they contain a higher number of bus stops when compared to areas such as highways like Pembina Highway. This aligns with the city's objective of trying to provide convenient transit experiences for its residents. We notice that there is at least one bus stop for every street in residential areas, which can be convenient for commuters. However, having a higher density of bus stops may negatively affect bus performance. The distance between each bus stop

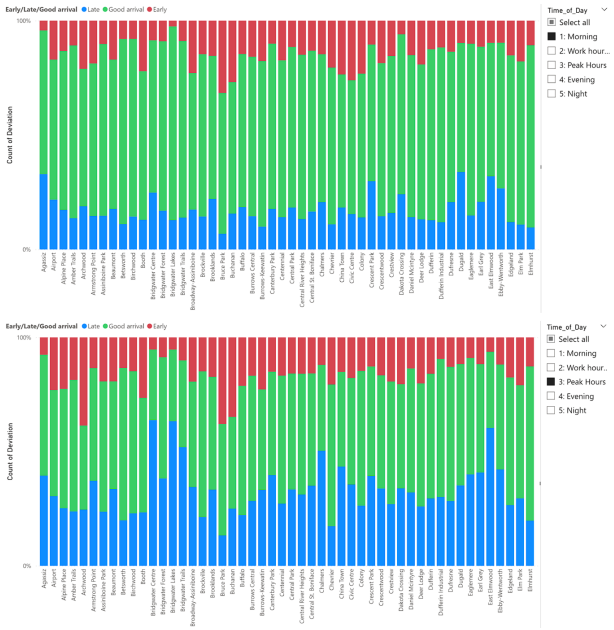

Fig. 12. 2022 Late Ratio For neighborhoods Morning (Upper) vs Afternoon Peak Hours (Bottom)

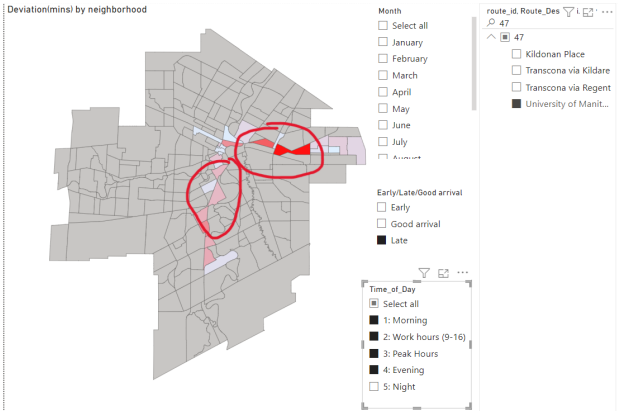

Fig. 13. Route 47 going to UofM via Downtown has more frequent late buses in the residential areas

has a significant impact on the performance of transit systems [8]. Our analysis shows that buses are frequently running late in areas with a higher number of bus stops. In other words, as the density of bus stops increases, the likelihood of buses running late also increases.

As shown in Figure 13, Route 47, which goes to the University of Manitoba, has a higher density of late buses in residential areas. Specifically, in the Regent area, which is predominantly residential, there are a large number of bus stops. As the bus moves away from Regent towards Pembina Highway, it begins to catch up to its schedule, and the frequency of late buses decreases, which aligns with our analysis.

The reason for the poor performance of Route 47 in areas with a high density of bus stops is straightforward. When

a bus has to make more stops, it spends more time on the route. Passengers getting on or off the bus at each stop may take longer to board the bus, especially when the bus is nearly full or there are many passengers waiting at the stop. Frequent stops in the same area can negatively impact the bus's performance. The visualization results indicate that Route 47 is more prone to being late in areas with a higher concentration of bus stops, but experiences fewer delays on Pembina Highway, where there are fewer stops.

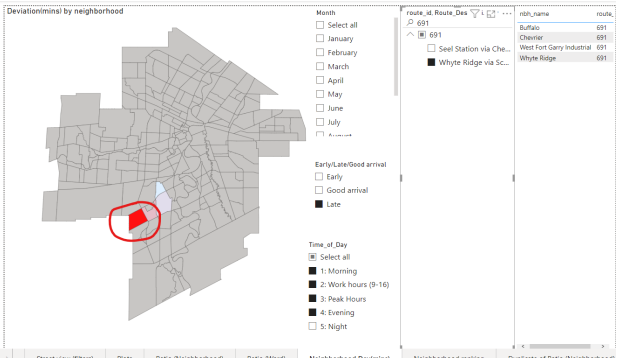

Fig. 14. route 691 going to Whyte Ridge via Scurfield has more frequent red buses (darker area) with more bus stops

Fig. 15. Stop locations for Route 691, Source: [9]

This reasoning can be applied to all buses, even to those that are performing well. For example, Route 691 going to Whyte Ridge via Scurfield is generally on time, but when it does run late, we observe that the areas where it gets delayed have a higher density of bus stops than other areas. As seen in Figure 15 [9], more bus stops for this route are towards the lower left side of this route. This observation supports the inference that buses tend to experience delays more frequently in areas with a higher number of bus stops. The visualization results in Figure 14 provide evidence that supports this conclusion.

Based on our observation, we can infer that a higher density of bus stops in an area negatively impacts the bus's on-time performance. Increasing the spacing between bus stops by 8 percent will result in passengers having to walk long distances to reach the bus stops, however, this change is expected to improve the running time of the buses[10]. Therefore, we

TABLE V
2021 TOP 5 WORST PERFORMING BUSES

Route Number	Route Destination
60	Downtown
18	Tuxedo
47	UofM via Downtown
77	Polo Park
18	Garden City Center via Jefferson

TABLE VI
2022 TOP 5 WORST PERFORMING BUSES

Route Number	Route Destination
47	UofM via Downtown
60	Downtown
77	Polo Park
18	Tuxedo
77	Kildonan Place

recommend reducing the number of bus stops in residential areas to improve overall transit performance. Specifically, we suggest increasing the distance between each bus stop in residential areas.

It is important to note that night hours are an exception to the previous observation. During this time, there are very few people using public transit, especially near midnight. However, some buses are still observed to be late during this time. We notice a pattern where buses frequently run late near the starting point of their route during the night. For example, Route 60 going downtown is often late near the University of Manitoba area, but it catches up with its schedule further up the route. We hypothesize that this is due to bus bunching, where the driver intentionally starts late from the starting point. Since there are fewer stops at night, the bus will naturally reach the first few stops on time or even early, but be late for subsequent stops. We believe this is because it is better to be late than early since waiting for an early bus can be more frustrating for commuters as the only option now is to catch the next bus. One solution to this could be to increase bus frequency during the night time, but this will not be financially viable for the city.

F. Analysis Results: Late Deviation

Our aim in using frequent pattern mining is to discover interesting correlations between entities. To achieve this, we focus on analyzing non-singleton frequent itemsets. These itemsets represent combinations of items that occur frequently together. One such itemset we study consisted of the {route ID, route destination} showing instances of a particular bus getting delayed on a specific route. This itemset is particularly interesting to us because a route typically involves a bus travelling in both directions, whereas this itemset provided data for a bus travelling in a single direction. For example, if a bus goes from destination A to B and from B back to A, the {route ID, route destination} itemset would give us results for either A to B or B to A if any of those trips are frequent.

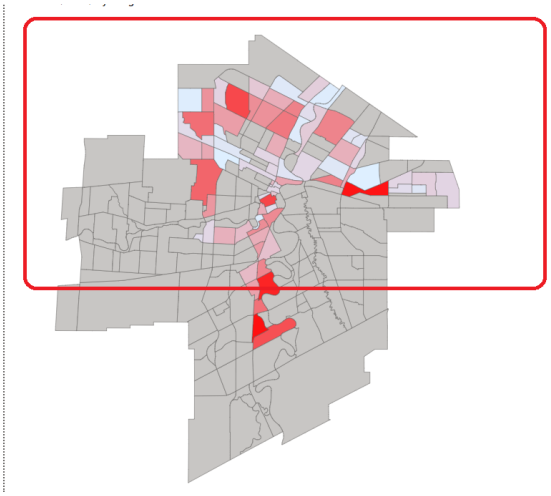

Fig. 16. 2022 - Worst performing buses mostly pass through the northern part of the city

The findings for both years are almost similar, and although we only list the top five routes, two trends emerged from our results. Firstly, most of the routes are in the Northeast or Northwest side of the city, except for Route 60. Secondly, nearly all of the identified buses travel through residential neighborhoods. As previously mentioned, buses passing through residential areas have poor performance, so it is logical that the frequently late buses identified by frequent pattern mining are primarily from residential neighborhoods.

Figure 16 reveals that the buses that are performing poorly are mostly delayed in the Northeast or Northwest areas of the city. These areas are indicated by the red blocks, which represent the neighborhoods that the buses pass through. However, there are a few red blocks in the Southern part of the city due to Routes 60 and 47. Route 47, however, is mostly delayed on the Northeast side of the city. The frequent pattern mining identifies buses that are frequently delayed, and it is possible that there are other routes that perform poorly but are not caught by the algorithm because they have fewer scheduled buses. This point is further discussed in the paper. The buses identified by frequent pattern mining are likely to have more passengers relying on them than other routes, which is why the city has planned to schedule these routes more frequently. Therefore, it is necessary to improve the transit routes associated with the Northeast and Northwest areas of the city.

V. Conclusions

In this paper, we presented our data science solution, which mines and analyzes on-time performance data of public transit data. By integrating the transit performance data with the neighborhood data, we were able to identify the correlation between poorly performing buses and the areas they operate in. We also extracted valuable information—such as the neighborhoods where the buses pass through and the specific times of day when delays are more likely to occur. To further analyze this data, we utilized frequent pattern mining to discover frequent patterns and association rule mining to discover interesting association rules. Our results revealed important insights, such as the areas of the city with the highest probability of late buses and the worst-performing buses and neighborhoods. We extended our analysis by visualizing the data, which helped us identify hidden patterns. Evaluation results on real-life data from the Canadian city of Winnipeg demonstrated practicality of our solution. As *ongoing and future work*, we explore ways to further enhance our data mining and knowledge visualization.

References

[1] K. Wiebe, "Measuring Winnipeggers' convenient access to public transit," Feb. 2018. DOI: https://www.iisd.org/publications/brief/measuring-winnipeggers-convenient-access-public-transit.

[2] C. Leung, P. Braun, C. Hoi, J. Souza, and A. Cuzzocrea, "Urban analytics of big transportation data for supporting smart cities," Aug. 2019, pp. 24–33, ISBN: 978-3-030-27519-8. DOI: 10.1007/978-3-030-27520-4_3.

[3] A.-R. Audu, A. Cuzzocrea, C. Leung, *et al.*, "An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city," Jan. 2020, pp. 224–236, ISBN: 978-3-030-22353-3. DOI: 10.1007/978-3-030-22354-0_21.

[4] C. K. Leung, Y. Chen, S. Shang, *et al.*, "Data mining on open public transit data for transportation analytics during pre-covid-19 era and covid-19 era," in *Advances in Intelligent Networking and Collaborative Systems*, Aug. 2020, pp. 133–144. DOI: https://doi-org.uml.idm.oclc.org/10.1007/978-3-030-57796-4_13.

[5] Statistics Canada, "Commuters using sustainable transportation in census metropolitan areas," in *2016 census in brief (2017)*, Nov. 2017. DOI: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016029/98-200-x2016029-eng.cfm.

[6] S. S. Singh, R. Javanmard, J. Lee, J. Kim, and E. Diab, "Evaluating the accessibility benefits of the new brt system during the covid-19 pandemic in Winnipeg, Canada," Feb. 2022. DOI: https://doi.org/10.1016/j.urbmob.2022.100016.

[7] E. Brass, "Winnipeggers call for more reliable bus service as gas prices continue to surge," Jan. 2022. DOI: https://www.cbc.ca/news/canada/manitoba/winnipeg-transit-efficient-1.6491239.

[8] S. I. Chien and Z. Qin, "Optimization of bus stop locations for improving transit accessibility," 2004. DOI: https://doi.org/10.1080/0308106042000226899.

[9] "Line route 691 - Winnipeg Transit - bus schedules." DOI: https://moovitapp.com/winnipeg_mb-1142/lines/691/19683323/4311338/en.

[10] E. Mylona, M. Savelonas, and D. Maroulis, "Effects of bus stop consolidation on passenger activity and transit operations," Jan. 2006. DOI: https://doi.org/10.1177/0361198106197100104.