A Regression Analysis of Winnipeg Transit On-time Performance in Relation to Traffic Volume

Group #19:

Jason Tran
Juhee Kim
Tommy Wu
Nurida Karimbaeva
Tanisha Turner

Introduction

The Team

Juhee Kim

Jason Tran

Tommy Wu

Nurida Karimbaeva

Tanisha Turner

This project is part of the course COMP 4710: Introduction to Data Mining for Fall 2021

About Our Project

Goal

- *Analyze and model On Time Performance of Buses in Winnipeg using information about the traffic and other factors*

Serves many useful purposes:

- *Understand which factors can influence bus delays and by how much*
- *Predict future bus delays*

Source Datasets

- Winnipeg Transit On time Performance Dataset
  - *Information about **time deviation** of bus arrivals at bus stops across the city of Winnipeg*
- Permanent Count Station Traffic Counts
  - *Information about **traffic volumes** recorded (for every 15 minutes) at 8 traffic count stations across the city of Winnipeg*
- Other Datasets:
  - *Locations of Bus stops*
  - *Street Mappings*
  - *Construction sites and road closure*

Idea

- Match the information of schedule deviation and nearby traffic volumes for every time period (e.g) at all bus stops of interest
- Problem becomes: Fitting a regression model with
  - *Explanatory variables: Bus stop information, Traffic volume and Timestamp*
  - *Response variable: Mean on-time performance*
- Use Machine Learning models to fit the data
  - *LinearSVC, KNeighborsRegressor, DecisionTreeRegressor, RandomForestRegressor*
- Analyze the results
  - *Whether there exists a relationship*
  - *Which features are important*

Data Preprocessing

Main Process:

- Choose bus stops of interest.
  - *Nearby one of the 8 traffic count stations (e.g. within 2 km)*
- Each bus stop have its intrinsic values
  - *Location (distance to the traffic count stations), Street (street name, direction facing, number of lanes, etc.), Stop Number*
- For each of these bus stops we then collect the information for
  - *Traffic volume (in each direction and in total)*
  - *The average schedule deviation*

  ...over a period of time (e.g. 1 month) and at a specific rate (e.g. every 30 minutes)

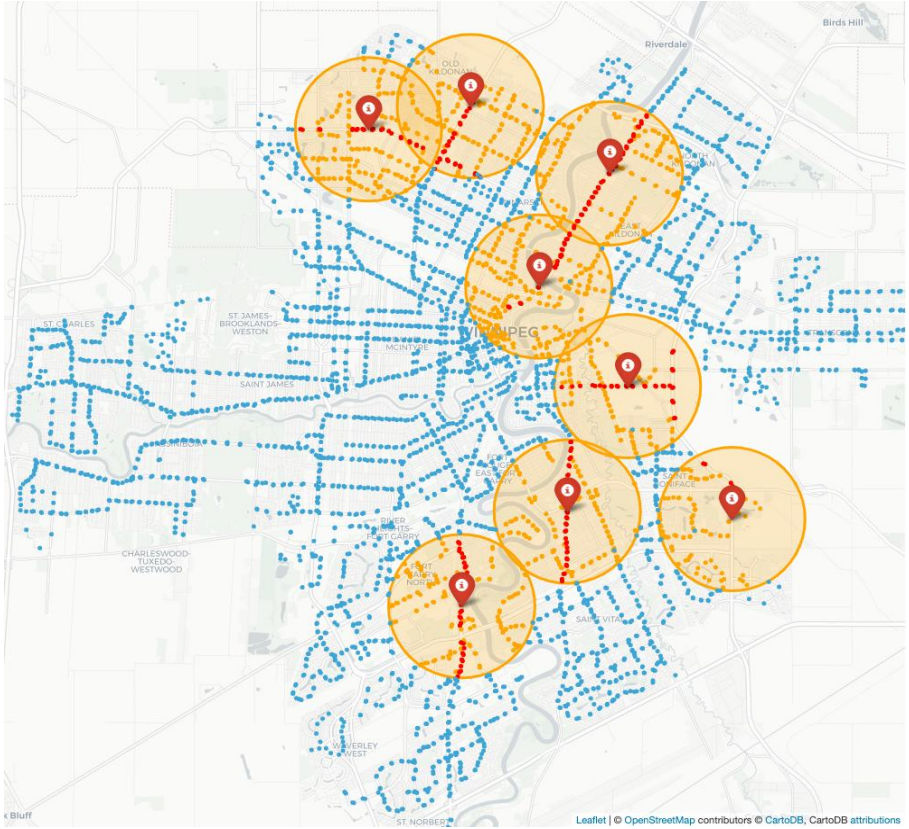
Bus Stops Coverage

Traffic Count Stations

- 8 count stations (red marker) are located on the city's main streets: *McPhillips, Henderson, Disraeli, Pembina, St. Mary's, Inkster, Marion, & Lagimodiere.*

Bus Stops

- ***5155 Bus Stops in total*** *across Winnipeg (blue, orange and red)*
- ***1559 Bus Stops are within 2 km*** *to one of the 8 traffic count stations (orange and red)*
- ***226 Bus Stops are on the same street*** *as the traffic count stations (red)*



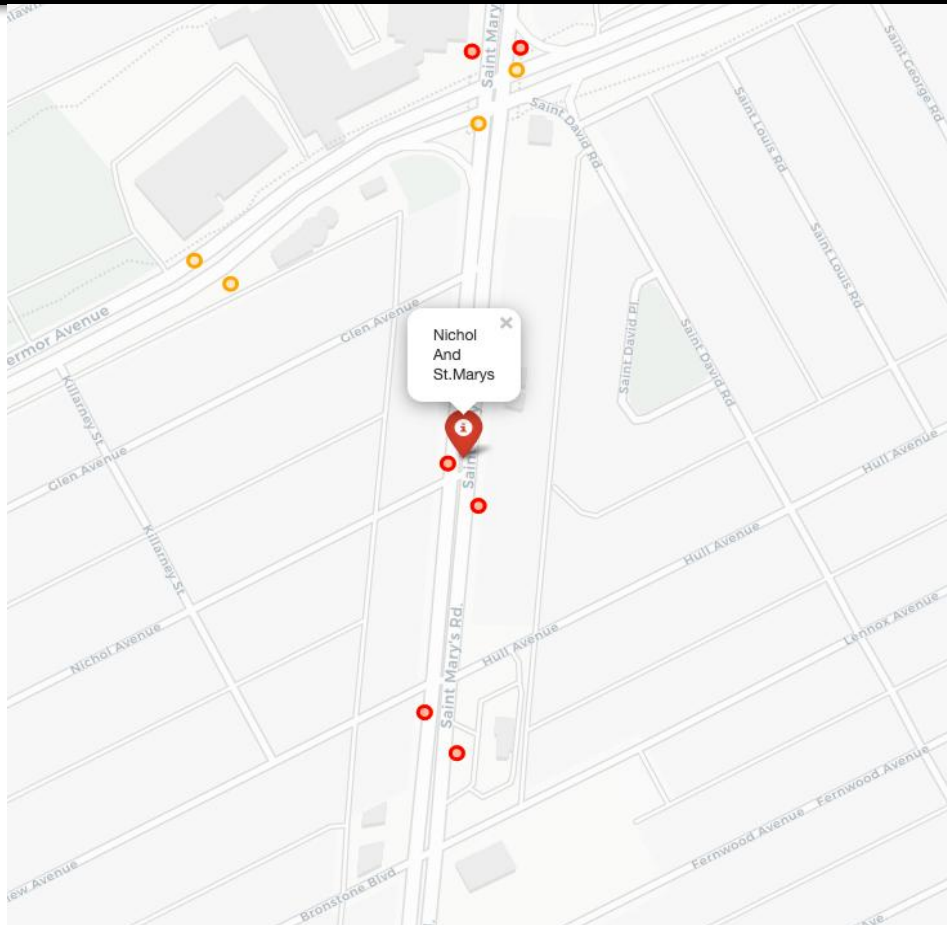
Bus Stops Features

Unique features
- *Stop Number (#35643)*
- *Location (49.863325,-97.422145)*

Street
- *Street name (St Mary's)*
- *Direction (Northbound)*
- *Number of Lanes (3)*

Relative to Traffic Count station
- *Distance to the station (41.456)*
- *Same Street? (Yes/1)*

These features are static!

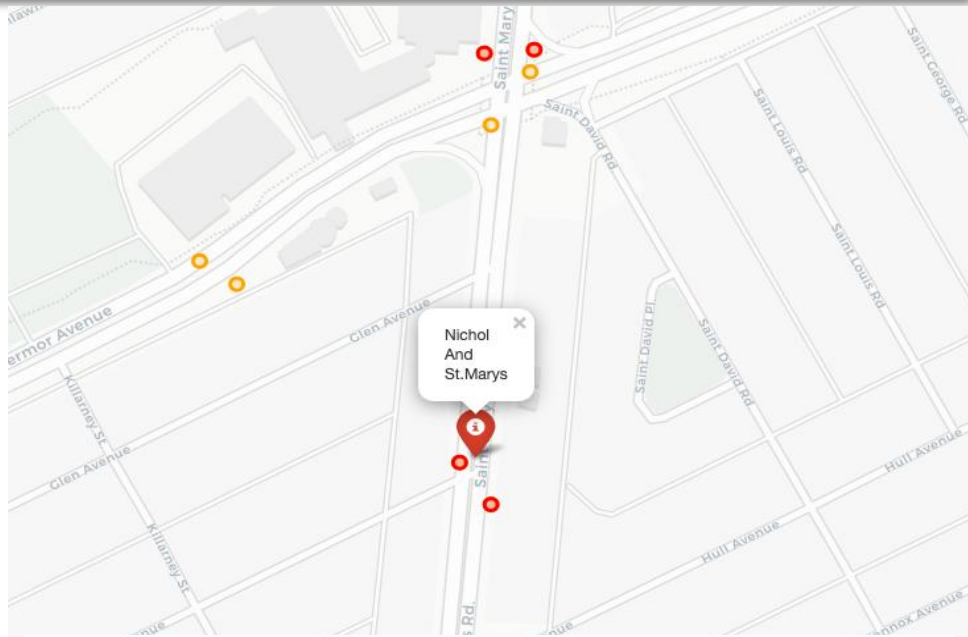


Traffic Volume Features

Location
- *Station name & location*
- *Street name*

Traffic Volume
- *Traffic each direction (N-S or E-W)*
- *Total traffic*

Timestamp
- *Time Interval: 15 minute*
- *Absolute time value*
- *Cyclical feature: Hour of day, Day of week, Day of year*



Timestamp	Northbound	Southbound	Eastbound	Westbound	Total	Street
2021-08-01 00:15:00	14.0	28.0	0.0	0.0	42	St Mary's
2021-08-01 00:30:00	12.0	21.0	0.0	0.0	33	St Mary's
2021-08-01 00:45:00	16.0	18.0	0.0	0.0	34	St Mary's
2021-08-01 01:00:00	10.0	17.0	0.0	0.0	27	St Mary's
2021-08-01 01:15:00	16.0	14.0	0.0	0.0	30	St Mary's

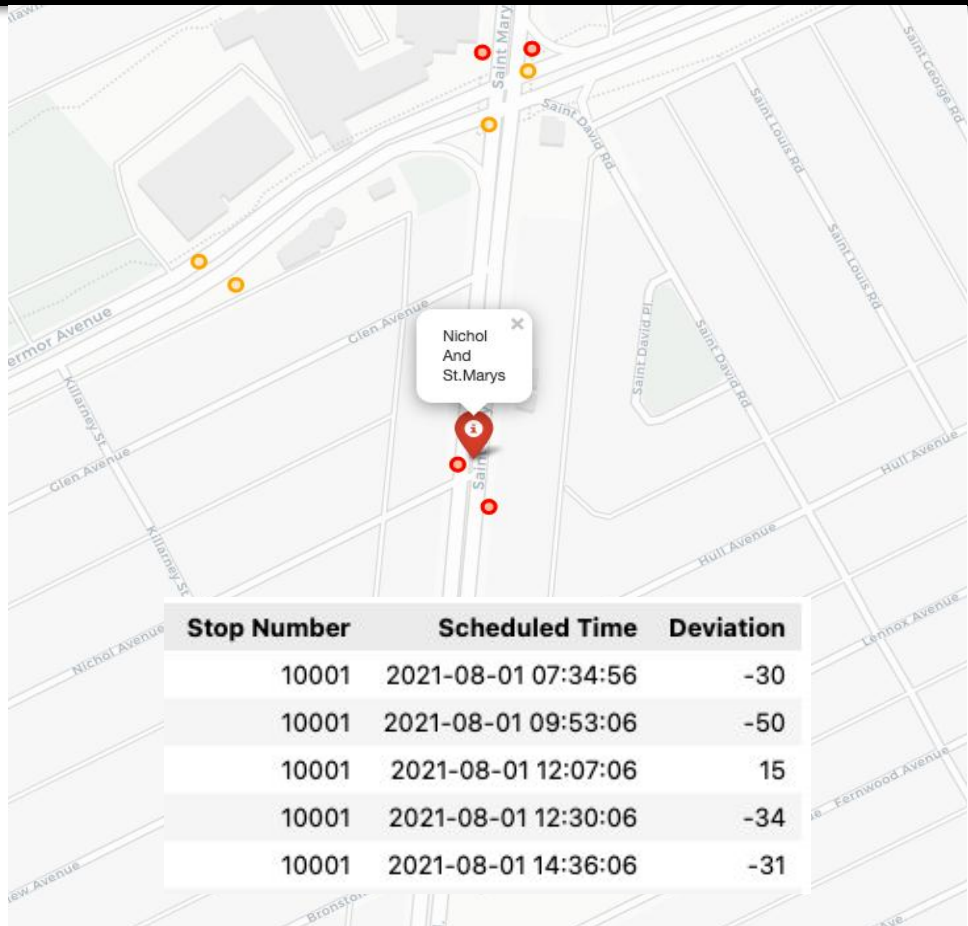
Bus On Time Performance

Bus Stop
- *Stop Number*

Schedule Deviation
- *Deviation = Arrival time - Scheduled time*

Timestamp
- *Exact scheduled time*

We will aggregate these data over a period of time to match with the traffic volume



Stop Number	Scheduled Time	Deviation
10001	2021-08-01 07:34:56	-30
10001	2021-08-01 09:53:06	-50
10001	2021-08-01 12:07:06	15
10001	2021-08-01 12:30:06	-34
10001	2021-08-01 14:36:06	-31

Resulting Data

Finally, our data will look like this

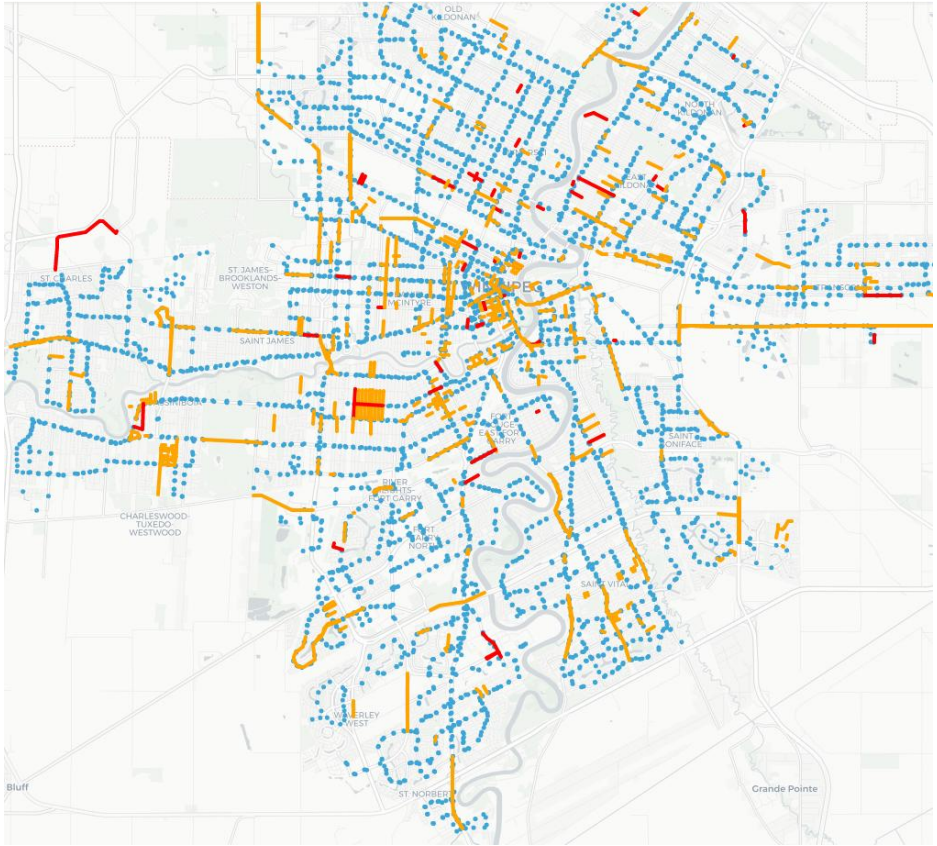
	Stop Number	Street	Site	Distance	Same Street	Directional	Total	Arrivals	Average OTP	Timestamp	Time of Day	Day of Week	Day of Year	Time value	Number of Lanes
0	30199	Leila	McPhillips	421.070384	0	0.0	418.0	2.0	-327.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
1	30205	Leila	McPhillips	1689.003285	0	0.0	418.0	2.0	-381.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
2	30206	McGregor	McPhillips	1741.524735	0	0.0	418.0	2.0	-378.5	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
3	30207	Partridge	McPhillips	1691.791193	0	0.0	418.0	1.0	-28.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
4	30209	McGregor	McPhillips	1718.386291	0	0.0	418.0	1.0	-188.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
5	30211	McGregor	McPhillips	1727.524447	0	0.0	418.0	1.0	-177.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
6	30212	McGregor	McPhillips	1743.271888	0	0.0	418.0	2.0	-328.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
7	30213	McGregor	McPhillips	1744.183619	0	0.0	418.0	1.0	-170.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
8	30214	McGregor	McPhillips	1766.327244	0	0.0	418.0	2.0	-330.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
9	30215	McGregor	McPhillips	1781.773222	0	0.0	418.0	1.0	-159.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
10	30216	McGregor	McPhillips	1815.182063	0	0.0	418.0	2.0	-332.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
11	30217	McGregor	McPhillips	1836.007740	0	0.0	418.0	1.0	-149.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
12	30218	McGregor	McPhillips	1861.761376	0	0.0	418.0	2.0	-333.5	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
13	30220	McGregor	McPhillips	1912.490434	0	0.0	418.0	1.0	-161.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
14	30221	McGregor	McPhillips	1942.335511	0	0.0	418.0	2.0	-336.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
15	30227	McGregor	McPhillips	1996.575295	0	0.0	418.0	1.0	-151.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	4.0
16	30354	Leila	McPhillips	1569.336930	0	0.0	418.0	2.0	-375.0	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
17	30355	Leila	McPhillips	1307.279722	0	0.0	418.0	2.0	-357.5	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
18	30356	Leila	McPhillips	1164.119545	0	0.0	418.0	2.0	-347.5	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0
19	30357	Leila	McPhillips	997.106894	0	0.0	418.0	2.0	-337.5	2021-08-01 02:00:00	2.0	6	213	1627783200000000000	2.0

Construction Sites (Supplementary)

Location
- *Street*
- *Boundaries, Average location*

Duration of closure
- *Start and End date*
- *Operation hours during the day*
- *Complete*

Caveat:
- *Only have information for active construction sites*
- *Since we are looking back at past data, some construction sites that were active then but had since finished may not appear in the data set*



Resulting Data with Construction Sites

With construction features, our data will look like this

	Stop Number	Street	Site	Distance	Same Street	Directional	Total	Arrivals	Average OTP	Timestamp	Time of Day	Day of Week	Day of Year	Time value	Number of Lanes	Nearby Constr	Same Street Constr
0	30001	McPhillips	McPhillips	1991.190701	1	204.0	619.0	1.0	70.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	3.0	0.0	0.0
1	30199	Leila	McPhillips	421.070384	0	0.0	619.0	2.0	-0.5	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	0.0	0.0
2	30207	Partridge	McPhillips	1691.791193	0	0.0	619.0	1.0	35.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
3	30209	McGregor	McPhillips	1718.386291	0	0.0	619.0	1.0	-9.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
4	30211	McGregor	McPhillips	1727.524447	0	0.0	619.0	1.0	-16.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	4.0	2.0	0.0
5	30213	McGregor	McPhillips	1744.183619	0	0.0	619.0	1.0	-6.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	4.0	2.0	0.0
6	30215	McGregor	McPhillips	1781.773222	0	0.0	619.0	1.0	7.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	4.0	2.0	0.0
7	30217	McGregor	McPhillips	1836.007740	0	0.0	619.0	1.0	3.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	4.0	2.0	0.0
8	30219	Jefferson	McPhillips	1935.619242	0	0.0	619.0	1.0	-60.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
9	30220	McGregor	McPhillips	1912.490434	0	0.0	619.0	1.0	-13.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	4.0	2.0	0.0
10	30222	Jefferson	McPhillips	1919.077528	0	0.0	619.0	1.0	0.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
11	30223	Jefferson	McPhillips	1756.643725	0	0.0	619.0	1.0	-49.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
12	30224	Jefferson	McPhillips	1762.260324	0	0.0	619.0	1.0	3.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
13	30225	Jefferson	McPhillips	1599.574955	0	0.0	619.0	1.0	-39.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
14	30226	Jefferson	McPhillips	1572.320342	0	0.0	619.0	1.0	-11.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	2.0	0.0
15	30227	McGregor	McPhillips	1996.575295	0	0.0	619.0	1.0	-1.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	4.0	2.0	0.0
16	30340	Templeton	McPhillips	1961.367923	0	0.0	619.0	1.0	-842.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	0.0	0.0
17	30341	Templeton	McPhillips	724.264234	0	0.0	619.0	2.0	-229.0	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	0.0	0.0
18	30343	Templeton	McPhillips	843.353201	0	0.0	619.0	2.0	-230.5	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	0.0	0.0
19	30344	Garden Park	McPhillips	981.023401	0	0.0	619.0	2.0	-230.5	2021-11-01 06:00:00	6.0	0	305	1635746400000000000	2.0	0.0	0.0

Methodology

- Supervised Machine Learning:
  - Training - Testing split: 75% - 25%
  - Randomized test data
- Regression Models:
  - LinearSVR
  - DecisionTreeRegressor
  - KNeighborsRegressor
  - RandomForestRegressor
- Evaluation:
  - R2 score
  - Results are average of 5 random states

Key Findings

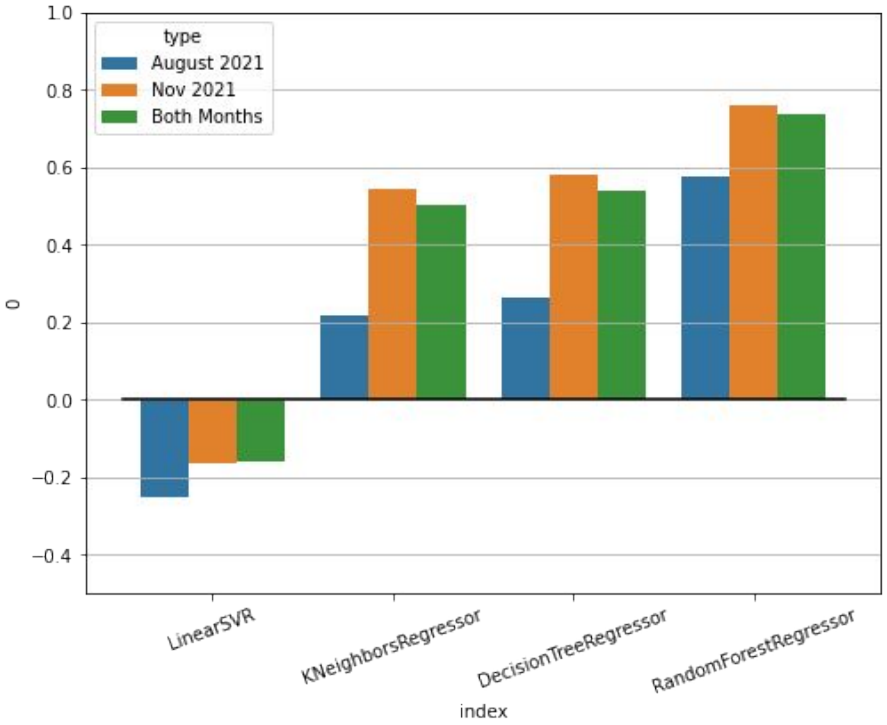
- Regression works
  - RandomForestRegressor is the best model for our data, and can predict with up to an r2 score of 0.79 on an average case
  - DecisionTreeRegressor and KNeighborsRegressor generally performed decently
  - Linear SVR produces extremely low score, an indication that the relationship is not a linear one
- Key features for performance:
  - Time factors (Time value, Time of Day, Day of Week and Day of Year) as well as Stop Number were shown to be key contributors to the regression models, as evidenced by the lower score in their absence.
  - The regression score is also significantly improved when trained and tested on only bus stops that are on the same street versus bus stops that are on different street to the count station, across all non-linear models.

Models' Performance

Observations

- Best Performer: RandomForestRegressor (0.759)
- November 2021 Data is better suited to our model. All models performed worse when August 2021 data were involved (Aug & Combined dataset).
- Thus moving forward, we will be using Nov 2021 data as a basis for comparison

R2 score	Aug 2021	Nov 2021	Combined data
LinearSVR	-0.253967	-0.165791	**-0.157980**
KNeighborsRegressor	0.219058	**0.544445**	0.504456
DecisionTreeRegressor	0.264441	**0.582471**	0.540938
RandomForestRegressor	0.574442	**0.759289**	0.737900

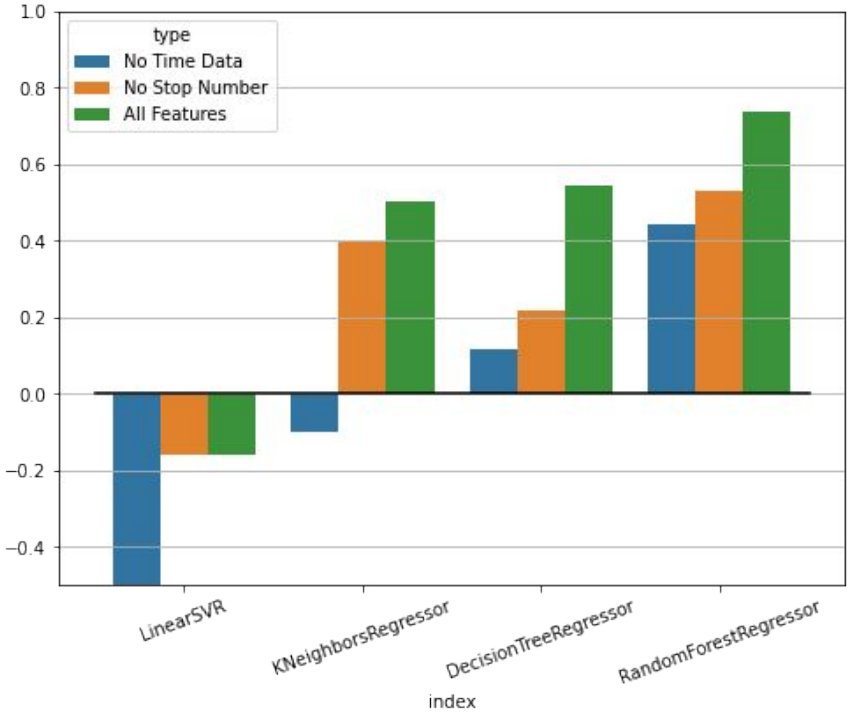


Time and Stop Data

Observations

- Removing temporal features from the regression model drastically reduces the performance of all models.
- Similarly, removing bus stop identifier may not really seem unreasonable, but they are shown here to be a contributory factor.

	No Time Data	No Stop Number	All Features
LinearSVR	-1.536164	**-0.157980**	**-0.157980**
KNeighborsRegressor	-0.100097	0.395829	**0.504456**
DecisionTreeRegressor	0.114709	0.217776	**0.542167**
RandomForestRegressor	0.442889	0.527975	**0.737917**

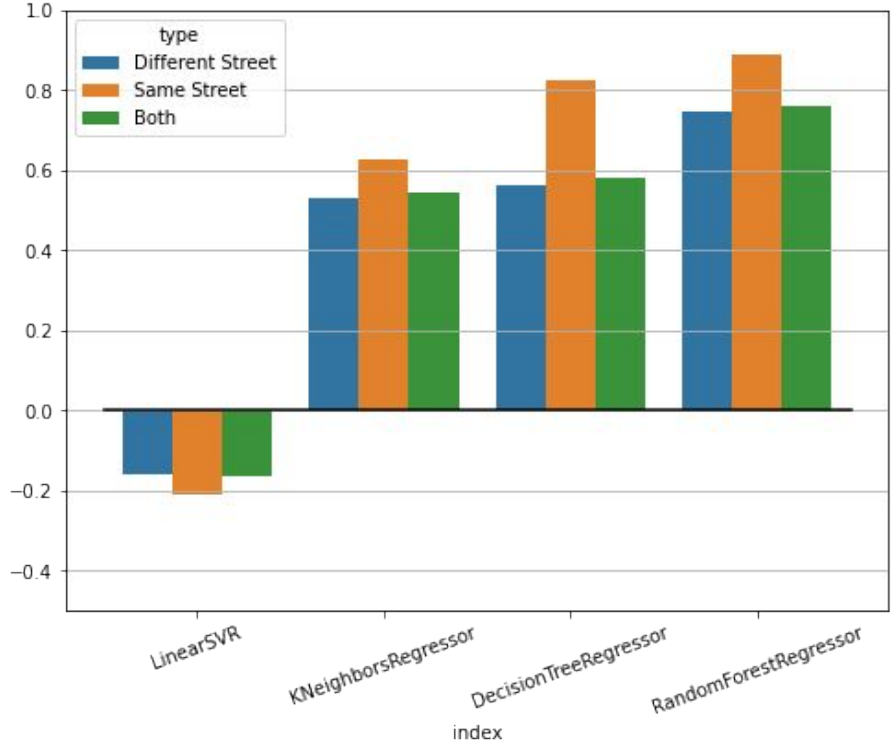


Same Street vs Different Street

Observations

- Bus stops on the same street show a stronger relationship between the explanatory variables and the response variable, as indicated by a sharp improvement and a very good overall regression score
- Bus stops on different streets show a slightly weaker relationship between the variables than the overall data

	Different Street	Same Street	Both
LinearSVR	**-0.159971**	-0.210019	-0.165791
KNeighborsRegressor	0.530451	**0.625394**	0.544445
DecisionTreeRegressor	0.564255	**0.822840**	0.582471
RandomForestRegressor	0.746875	**0.887484**	0.759289



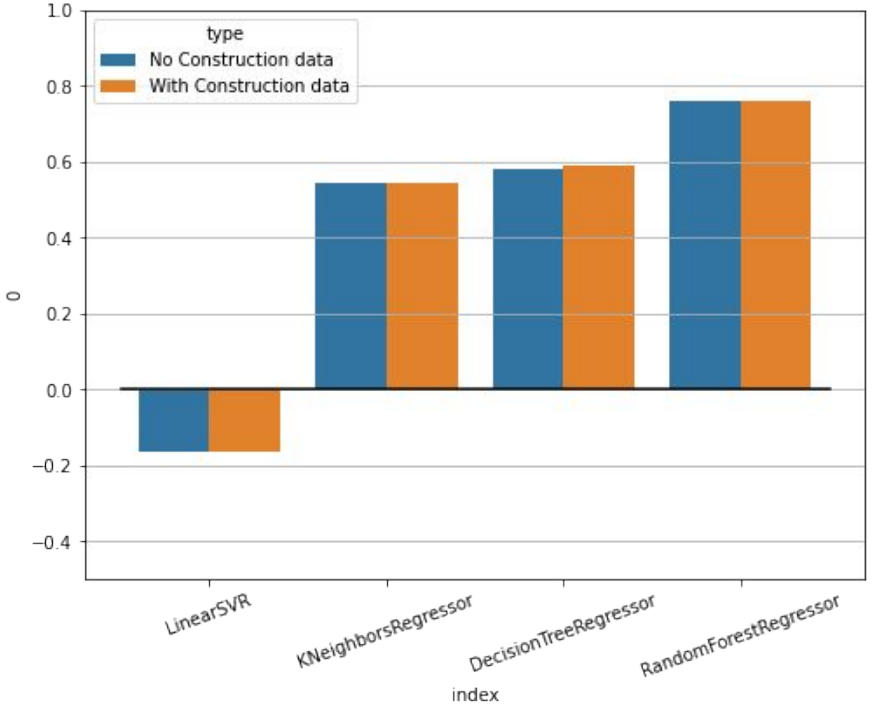
Other Findings

Construction data were not significant to the model

Weak relationship due to some likely reasons:

- Bus stops that are directly affected by the lane closure will not produce any data for us
- Bus stops are more frequently connected by their routes than the streets they are on.

	With Construction data	Without Construction data
LinearSVR	-0.165791	-0.165791
KNeighborsRegressor	0.544414	**0.544445**
DecisionTreeRegressor	**0.591031**	0.582471
RandomForestRegressor	**0.761905**	0.759289



Other Findings

Future Predictions did not Work

- We also tried to see if we can predict bus delays using knowledge of prior patterns
  - *For example, we attempted to predict the bus delays for several hour periods during the 28th of November, having trained the model on data of the preceding events (i.e. the prior 4 weeks)*
- The results were poor and inconsistent, however.

2021-11-28	**12:00**	**13:00**	**14:00**	**15:00**	**16:00**	**17:00**	**…**
R2 score	-0.2022	0.1334	-4.1245	-0.6151	-0.3556	0.0258	…

- Likely because the datasets were not large enough
  - *Data size constraint: Not able to process data that are too large*
  - *Not able to generalize for a complex process like future prediction*

Obstacles

- The findings suggest that the more data the model has access to, the better it becomes.
  - *This could help improve our regression model even further and more importantly, the ability for the model to predict future delays based on traffic volume*
- However, the amount of data that would need to be mined for this process would be immense
  - *Data for Transit On Time Performance for the entire year of 2020 was 91 million entries (11.67 GB)*
- Thus we were constrained to working with only 1 to 2 months worth of data.
  - *This might have led to uninteresting findings when it comes to a process as complex as predicting future delays*

Future Work

Nevertheless, the results still look promising.

We expect that this work can be further improved upon in several ways, such as

- *Training the model on more data (e.g. 1 year of data) to make better future predictions*
- *Analyzing the prediction data to see which time of day account for the most error.*
- *Incorporating bus route and it's associated stops with the on-time-performance*
- *Taking emergencies such as car accidents, floods, and malfunctioning stop lights into account*

Thank you

We would like to express our very great appreciation to our professor Dr. Carson K. Leung for providing us with his incredibly valuable guidance and feedback during the development of this project. It was instrumental in the completion of our project.

We would also like to thank our TAs: Evan Madill, Adam Pazdor, and Qi Wen for their great assistance to our research.