Uncovering causality on TTC Line 2 Delays

Elric Lazaro

12/23/2020

Code and data supporting this analysis is available at:
https://github.com/ElricL/TTC-Line-2-Delay-Causality.git

Abstract

This study seeks to find root causalities on TTC Line 2 Delays in hopes of improving the subway system. Through the use of propensity score matching, simulated subway entry data was analyzed based on travel time. Evidence of cold seasonal times and peak days of the week were found from observing the logistic regression model used to calculate the propensity scores. However no significant findings were found when evaluating the matching which suggests that a focus shift is required, either by looking at the system more as a whole or diving in deeper.

Keywords

TTC, Toronto, Subway System, Propensity Score Matching, Delay, Causal inference

1. Introduction

One of the many daily challenges students who did not live by UTSG faced was commuting with major delays. This struggle can also be seen from those that work in downtown Toronto as well. A Regular daily commuting route often consists of many services such as MiWay buses, TTC Subway, Go Buses, and many more. Yet all of which are very susceptible to delays. TTC Subways often receive backlash, with the service often requiring train exchanges, emergency stops, maintenance, and slowdowns. Despite all the recent improvements such as Presto gates and more accessibility elevators, to this day many delay issues are still apparent with TTC subways. These problems have desensitized many TTC Subway users who often simply leave home, school, or work early to beat any possible major delays. While one paper may not be enough to fully understand TTC's complex subway system, it is still important to start uncovering some of the underlying problem's that we have come to accept today.

This study seeks to delve deeper into what variables can likely cause Subway delays and how the delays overall affect the subway user's travel time, which remained largely unclear from previous study TTC Subway Delay Cause Analysis (Lazaro, 2020). Propensity score matching (PSM) will be used to identify any major causalities on Subway Delays on TTC Line 2 stations. The advantage of PSM is that it will allow us to compare observation outcomes and consider large number of variables without it having a large effect on our sample size.

The observed data will be simulated based on observed patterns from Toronto's open data on TTC Subway Delays along with the study, TTC Subway Delay Cause Analysis (Lazaro, 2020). More on how the data is simulated will be further expanded upon in the Methodology section.

2. Methodology

2.1 Data

The entries in the simulated data represents an individual entering a station and their travel time onward. Each row or entry consist of the month, day, station, bound (East or West), whether there is a delay or not, and the travel time. Likelier events are prioritized and have higher chance to appear in the simulated data. The target population for this dataset is all TTC Subway users with the frame population being the simulated users observed/generated in the dataset

Business days, school terms, and the reported busiest stations from Urbanized (Chan, 2019) were taken into consideration into which values are more likely to appear for month, day, and station. The sample size totals to 500,000 observations given the simulated data is meant to represent a year. Whether the individual is delayed or not depends on the generated month, day, and station.

Utilizing the 2019 TTC Subway delays dataset observed from the study, TTC Subway Delay Cause Analysis (Lazaro, 2020), I've ranked the distinct values for the three variables based on the number of occurrences. The rankings for each variables can be found in the appendix with the lowest number having lowest number of occurrences to highest number having highest number of occurrences. Note that the values with number of occurrences that are tied or quite close to each other will be assigned the same ranking. The rankings are applied for each entries depending on what values are generated and then the total ranking is used to calculate a softmax probability for the chance of delay. Note bound was not used in this probability formula as both East and West had very close number of occurrences. Since we used the rank for month, day, and station to determine the probability of delay, we are enforcing the more common occurrences observed from Toronto's dataset to have higher chance on having delays in our simulation. For instance we can see in Figure 1 that the simulated data relatively has the higher numbered rank stations from the Toronto data to have highest amount of delay occurrences.

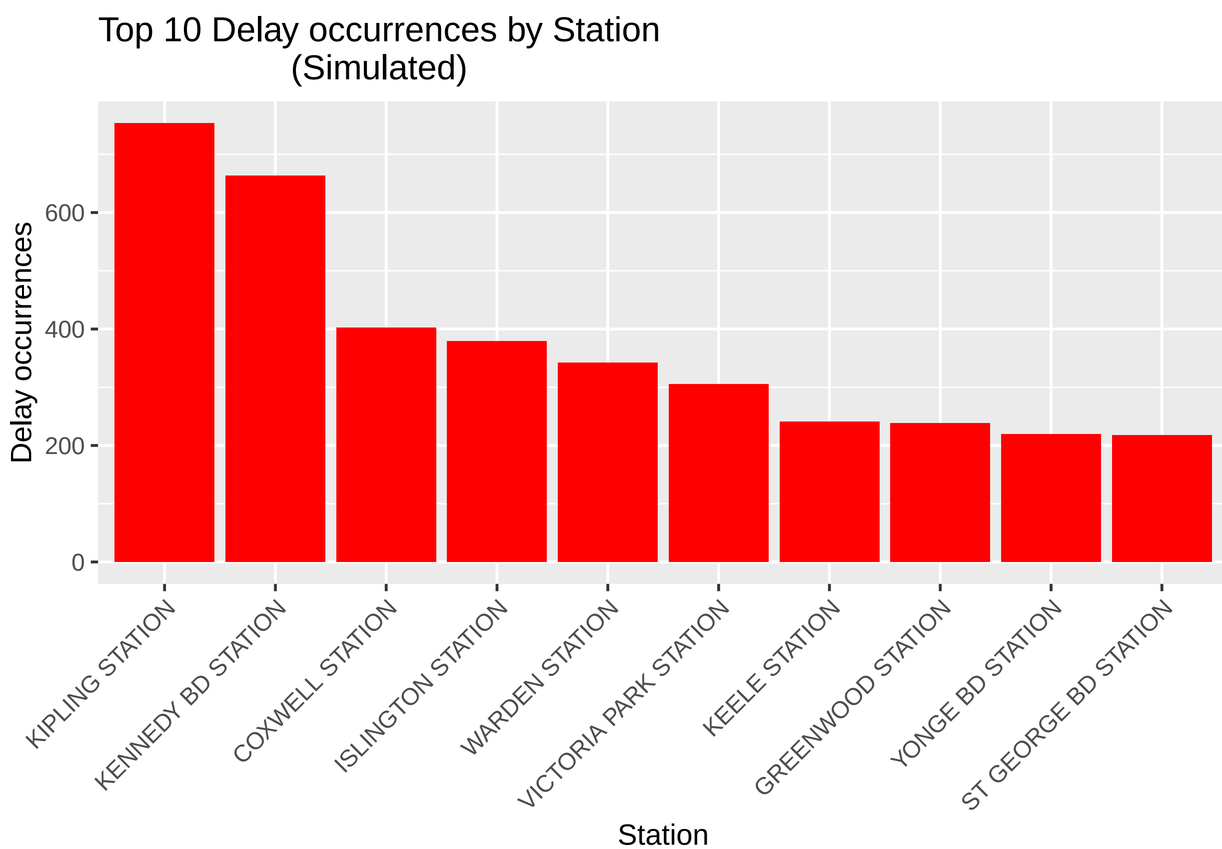

Figure 1: Top 10 Delay occurrences by Station

The softmax probability is calculated simply by $p = \frac{Sum\ of\ rankings}{exp(50)}$. Note that the probability of not getting delayed would then be $1 - p$.

Lastly to simulate the travel time, the normal distribution was chosen. The normal distribution is used to take advantage on the dataset's large size, specifically the sampling distribution of the mean will approach closer to a normal distribution. The distribution applied has a mean (μ) 20 minutes for non delayed and 20 + 6.82 minutes (travel time mean plus average delay time) when delayed. With the average delay times per station from TTC Subway Delay Cause Analysis (Lazaro, 2020), the mean delay of 6.82 was calculated. Check table 5 for the average delay time by Month calculated from the previous study. As for the standard mean, 20 was roughly estimated from the TTC Subway Travel time Chart (Flack, 2019). To see examples of entries in the data, please refer to Table 8 in the appendix.

With the simulated data generated, we now have our dataset which we'll be studying through PSM methodology. The final two variables generated are significant as delay will be our Treatment with travel time being the outcome of interest for our PSM model which is further explained in the 2.2 Model section.

2.2 Model

Propensity Score matching allows us to compare and evaluate the outcome of our observations to determine any significant variables. Firstly, we want to see the propensity of someone getting delayed and then match based on that. A logistic regression model will be used to determine the propensity score for each observations. Aside from determining propensity scores, the model will also be used to observe p-values of the independent variables based on delay to determine any significant relationships. Specifically, check if the variables are less than 0.05 to signify a relationship. The variable Delayed will be modeled based on the independent variables Month, Day, Station, and Bound:

$$
\begin{aligned}
log(\frac{(p)}{1-(p)}) = \beta_0 + \beta_1 X_{Month:\ January} + \\
\beta_2 X_{Month:\ February} + ... + \\
\beta_1 2 X_{Month:\ December} + \beta_1 3 X_{Day:\ Monday} + \\
\beta_1 4 X_{Day:\ Tuesday} + ... + \\
\beta_1 9 X_{Day:\ Sunday} + \beta_2 0 X_{Station:\ KIPLINGSTATION} + \\
\beta_2 1 X_{Station:\ ISLINGTONSTATION} + ... + \\
\beta_{22} X_{Station:\ KENNEDYBDSTATION} + \beta_{23} X_{Bound:\ E} + \\
\beta_{24} X_{Bound:\ W}
\end{aligned}
$$

Once the logistic model is calculated, it will be forecasted onto the dataset, essentially calculating the probability of delay for each individual's entry to the subway. A new dataset will then be created by gathering those that have gotten delayed and match them with those that have not gotten delayed but have similar or same propensity scores.

To determine how the delays and other observed variables affect the travel time, a linear regression will be used on the new reduced-matching dataset:

$$
\begin{aligned}
y = \beta_0 + \beta_1 X_{Month:\ January} + \\
\beta_2 X_{Month:\ February} + \ldots + \\
\beta_1 2 X_{Month:\ December} + \beta_1 3 X_{Day:\ Monday} + \\
\beta_1 4 X_{Day:\ Tuesday} + \ldots + \\
\beta_1 9 X_{Day:\ Sunday} + \beta_2 0 X_{Station:\ KIPLINGSTATION} + \\
\beta_2 1 X_{Station:\ ISLINGTONSTATION} + \ldots + \\
\beta_{22} X_{Station:\ KENNEDYBDSTATION} + \beta_{23} X_{Bound:\ E} + \\
\beta_{24} X_{Bound:\ W} + \beta_{25} X_{Delayed}
\end{aligned}
$$

This model will allow us how the treatment variable, delay, will affect the travel time as well as see any causalities with the other independent variables. To discover any significant variables, we will observe the p-value testing and seeing if it is less than 0.05 as a result of the model.

3. Results

As a result of the the applying probabilities based on rankings for the delay, in table 1 we can see there were 3,461 entries out of 500,000 observations that have experienced a delay. Calculating the number of delay occurrences out of 500,000 observations, we can see in theory there is approximately 0.7% chance to experience delay every time one enters a Line 2 subway station throughout the year.

Delayed	number of occurences
Not Delayed	496539
Delayed	3461

Table 1: Number of delay occurrences and those not delayed.

When applying the logistic regression model to determine propensity scores, p-values for the independent variables based on delay was observed. Looking at the values seen in Table 2, we can see that some of the months such as December and each days of the week play a large role in the likeliness of a delay occurrence. However with the ranking probability logic applied, stations overall do not play much significance on delays.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-28.1378703	821.2133514	-0.0342638	0.9726668
MonthAugust	-2.2739916	0.3923522	-5.7957913	0.0000000
MonthDecember	-17.8810392	450.3680676	-0.0397032	0.9683298
MonthFebruary	2.3863526	0.1107212	21.5528119	0.0000000
MonthJanuary	4.3113125	0.1106776	38.9538028	0.0000000
MonthJuly	-3.5636673	0.7148853	-4.9849497	0.0000006
MonthJune	-3.1398109	0.5868599	-5.3501882	0.0000001
MonthMarch	1.1777352	0.1181936	9.9644600	0.0000000
MonthMay	-0.1342046	0.1674350	-0.8015322	0.4228236
MonthNovember	-17.9430474	451.6330083	-0.0397293	0.9683090
MonthOctober	-17.9245475	449.6940339	-0.0398594	0.9682052
MonthSeptember	-3.9625573	0.7146958	-5.5443967	0.0000000
DayMonday	-2.1548780	0.2092213	-10.2995146	0.0000000
DaySaturday	-3.0374521	0.4557846	-6.6642267	0.0000000
DaySunday	-3.8906363	0.7126303	-5.4595440	0.0000000
DayThursday	2.9916479	0.0891323	33.5641441	0.0000000
DayTuesday	2.9577194	0.0892892	33.1251580	0.0000000
DayWednesday	-0.0567639	0.1088986	-0.5212550	0.6021892
StationBAY STATION	0.0076941	1168.9576768	0.0000066	0.9999947
StationBROADVIEW STATION	-0.0219590	1162.3688969	-0.0000189	0.9999849
StationCASTLE FRANK STATION	0.0564155	2405.4917353	0.0000235	0.9999813
StationCHESTER STATION	-0.0506982	2436.2323555	-0.0000208	0.9999834
StationCHRISTIE STATION	-0.0347211	1168.6258582	-0.0000297	0.9999763
StationCOXWELL STATION	19.4044180	821.2133446	0.0236290	0.9811486
StationDONLANDS STATION	-0.0974270	2426.0715275	-0.0000402	0.9999680
StationDUFFERIN STATION	-0.0313501	1159.5275143	-0.0000270	0.9999784
StationDUNDAS WEST STATION	0.0169916	1161.8818833	0.0000146	0.9999883
StationGREENWOOD STATION	16.5028123	821.2139533	0.0200956	0.9839671
StationHIGH PARK STATION	0.0654694	1168.4017118	0.0000560	0.9999553
StationISLINGTON STATION	18.3585226	821.2133524	0.0223554	0.9821645
StationJANE STATION	0.0310856	1164.5501904	0.0000267	0.9999787
StationKEELE STATION	14.4907221	821.2139481	0.0176455	0.9859217
StationKENNEDY BD STATION	20.5897700	821.2133395	0.0250724	0.9799972
StationKIPLING STATION	22.5034064	821.2133396	0.0274026	0.9781386
StationLANSDOWNE STATION	-0.0025342	1165.7601416	-0.0000022	0.9999983
StationMAIN STREET STATION	-0.0079423	1162.6582567	-0.0000068	0.9999945
StationOLD MILL STATION	-0.1130151	2407.6330831	-0.0000469	0.9999625
StationOSSINGTON STATION	0.0645537	1168.5271581	0.0000552	0.9999559
StationPAPE STATION	0.0185679	1166.0934231	0.0000159	0.9999873
StationROYAL YORK STATION	0.0227634	1164.3965860	0.0000195	0.9999844
StationRUNNYMEDE STATION	-0.0323758	1159.9779772	-0.0000279	0.9999777
StationSHERBOURNE STATION	0.0083263	1163.0800374	0.0000072	0.9999943
StationSPADINA BD STATION	-0.0161599	1164.4617996	-0.0000139	0.9999889
StationST GEORGE BD STATION	13.4610277	821.2139475	0.0163916	0.9869220
StationVICTORIA PARK STATION	17.1391024	821.2133827	0.0208705	0.9833490
StationWARDEN STATION	17.4343252	821.2133712	0.0212300	0.9830622
StationWOODBINE STATION	0.0093888	1162.1276513	0.0000081	0.9999936
StationYONGE BD STATION	14.5558094	821.2135416	0.0177248	0.9858584
BoundW	-0.0398310	0.0461716	-0.8626740	0.3883168

Table 2: Summary of Logistic Regression Model, used for Propensity Score logic.

Given there are 3,461 delay occurrences, the reduced-matched dataset consists of 6922 observations to be evaluated. A linear model was used to determine relationships between the dependent variable travel time and the independent variables, Month, Day, Station, Bound, and Delayed. When looking at table 3, we can see most variables hold no to little significance, due to having p-values greater than 0.05. However, there appears to be a relationship to travel time and the month of July with a negative coefficient of -2.37. Given a normal distribution with different mean of 26.82 was applied for delay occurrences during simulation (20 for non delays), there is a relationship with travel time and delays with positive coefficient of 6.8.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.9596623	0.2084609	95.7477621	0.0000000
MonthAugust	-0.1887109	0.5560649	-0.3393684	0.7343426
MonthFebruary	0.0167083	0.1479489	0.1129326	0.9100872
MonthJanuary	-0.0546206	0.1447816	-0.3772621	0.7059904
MonthJuly	-2.3652837	1.0173328	-2.3249851	0.0201016
MonthJune	-0.1620776	0.8339760	-0.1943433	0.8459128
MonthMarch	0.0708362	0.1616071	0.4383236	0.6611655
MonthMay	0.0449783	0.2338650	0.1923258	0.8474927
MonthSeptember	-1.1748897	1.0168148	-1.1554608	0.2479418
DayMonday	0.2464894	0.2840986	0.8676193	0.3856330
DaySaturday	0.5587904	0.6410926	0.8716220	0.3834450
DaySunday	-0.6821841	1.0094167	-0.6758201	0.4991775
DayThursday	0.0413585	0.0860442	0.4806657	0.6307694
DayTuesday	-0.0099921	0.0865511	-0.1154476	0.9080937
DayWednesday	0.1031579	0.1121133	0.9201219	0.3575412
StationGREENWOOD STATION	0.4820181	1.4306980	0.3369112	0.7361941
StationISLINGTON STATION	-0.3036891	0.2486407	-1.2213977	0.2219772
StationKEELE STATION	-0.0256381	1.4302212	-0.0179260	0.9856984
StationKENNEDY BD STATION	0.0480130	0.1409094	0.3407370	0.7333120
StationKIPLING STATION	0.0259049	0.1374027	0.1885325	0.8504648
StationST GEORGE BD STATION	-2.6321906	1.4419764	-1.8254048	0.0679832
StationVICTORIA PARK STATION	-0.1340740	0.4033664	-0.3323877	0.7396066
StationWARDEN STATION	-0.1752085	0.3532713	-0.4959603	0.6199382
StationYONGE BD STATION	-0.8542482	0.8331909	-1.0252731	0.3052703
BoundW	-0.0183284	0.0488904	-0.3748863	0.7077565
Delayed1	6.8002932	0.0518498	131.1536585	0.0000000

Table 3: Summary of Linear Regression Model, used for evaluating propensity score matching.

4 Discussion

4.1 Summary

In this study, Line 2 subway station entries along with travel time were simulated based on delay occurrence rankings for each variable and travel time from TTC Travel time Chart (Flack, 2019). Propensity score values are used to compare and match each similar observations. The values are calculated by calculating a logistic regression model from the simulated data and forecast it onto each observations. Entries with delays are then matched to those without delays based on similar propensity scores. The logistic model was also used to observe causality on delays by observing the evaluated p-values. With the reduced dataset with matched entries, we then examine the effect of getting delayed on travel time through evaluating the linear regression model and examine the p-values for any significant relationships.

4.2 Conclusion

If the simulated data is reproducible, the 0.7% chance calculated may generalize the percentage of delay occurrences. This means that whenever one enters a Line 2 subway station intending to take the train, one approximately has a chance of 0.7% to experience a delay. Note this is generalized for the entire year and may vary from month to month and other aspects that have not been studied. While 0.7% may seem small, given that user such as a student or worker that commutes on a daily basis, the chance of delay may no longer seem unlikely.

After having observed the p-values from the initial logistic regression model, we can conclude that looking at each stations independently holds no significance to the chance of delay. It appears that the problem of delays cannot directly be connected to a specific station. This means that an improvement cannot simply be made on a station that may be lacking in performance rather one would need to look at the system of line 2 subways as a whole. In contrast, we have observed that weekdays and some of the months hold high significance. For weekdays we can see that the weekends, Saturday and Sunday, have the smallest chance of having delays given their small coefficients. Given that there's less Subway users on the weekends, this could tell us that the current system right now may not have the optimal support for large amount of Subway users for Line 2. For the months that shows to have significance with delays, one can observe that the months in winter season have positive coefficients while those in other seasons have negative. This supports the finding from previous study, TTC Subway Delay Cause Analysis (Lazaro, 2020), that the Line 2 system may need some improvement when it comes to colder weather.

With the propensity matching evaluated, there were no relationships found for the outcome travel time aside from the variables Delayed and the month of July. Delayed however holds no particular interest given it is part of the process of simulating travel time as observations that are delayed are enforced to have a longer travel time. July has a negative coefficient which results in having faster travel times. This may mean that the Subway in summer generally fares well in comparison to other seasons. However given that other months had p-values greater than 0.05, the relationship between July and travel time remains inconclusive and may have appeared by chance due to the dataset being simulated. While we may not have learned much causality from conducting PSM, this could provide a lesson that a shift of focus may be needed. For instance, it may be problematic that looking at the stations independently given they are connected. It may be important to look at the TTC system more as a whole or look at more specific aspects.

4.3 Weaknesses and Next Steps

The data for this study is largely simulated, hence it may not generalize well to an equivalent data gathered from a survey. The probability for each value during the simulation process was evaluated under assumptions and as well as the observations from another dataset. Conducting studies on datasets related to the probabilities that were assumed may result into a more accurate simulated table. While it is very expensive it is also possible to conduct an experiment or a survey for each station.

This study still uses Toronto's open data on TTC Subway delays for reference on calculating the probability of delays. Unfortunately the dataset contains uninterpretable values which had to be removed which could result in skewing significant characteristics and creating bias observations. Instead of removing, it is possible to contact staff responsible of the dataset for consultance on how to clean up the dataset.

Having only studied line 2, only a portion of the TTC subway system was studied. Therefore this paper does not generalize well to the overall system. This decision was made since each bounds have different characteristics such as different trains and regions. Fortunately this study can be reproduced for each line and we can then find any common patterns between each study to generalize characteristics of the TTC subway system as a whole.

In the end not many causalities were identified from PSM methodology. This may be due to the methodology's problematic nature on heavy reliance with complete randomization. This randomization may lead to increase in imbalance. To see that this study's lack of causal findings is a result of imbalance, we can check other matching methods such as coarsened exact matching (King et al., 2019).

5 Appendix

Station	n	ranking
LANSDOWNE STATION	74	1
RUNNYMEDE STATION	74	2
DUFFERIN STATION	86	3
PAPE STATION	86	4
SHERBOURNE STATION	89	5
CASTLE FRANK STATION	94	6
CHESTER STATION	101	7
BAY STATION	106	8
SPADINA BD STATION	115	9
BATHURST STATION	117	10
MAIN STREET STATION	117	11
HIGH PARK STATION	120	12
DONLANDS STATION	130	13
WOODBINE STATION	133	14
CHRISTIE STATION	141	15
BROADVIEW STATION	147	16
OSSINGTON STATION	147	17
DUNDAS WEST STATION	162	18
OLD MILL STATION	174	19
ROYAL YORK STATION	188	20
JANE STATION	208	21
ST GEORGE BD STATION	218	22
YONGE BD STATION	219	23
GREENWOOD STATION	238	24
KEELE STATION	241	25
VICTORIA PARK STATION	305	26
WARDEN STATION	342	27
ISLINGTON STATION	379	28
COXWELL STATION	402	29
KENNEDY BD STATION	663	30
KIPLING STATION	753	31

Table 4: Station rankings by number of delay occurrences. (From Toronto's open TTC Delay Dataset)

Month	n	average_delay_time	ranking
November	430	6.787645	1
December	454	7.099222	2
October	457	6.886827	3
September	483	6.795494	4
June	494	6.108153	5
July	541	7.457539	6
August	553	8.855856	7
April	568	6.809240	8
May	570	6.373494	9
March	579	6.544423	10
February	604	5.797320	11
January	720	6.281098	12

Table 5: Month rankings by number of delay occurrences. Also contains average delay time by month. (From Toronto's open TTC Delay Dataset)

Day	n	ranking
Sunday	640	1
Saturday	763	2
Monday	986	3
Friday	1005	4
Wednesday	1005	5
Thursday	1026	6
Tuesday	1028	7

Table 6: Day rankings by number of delay occurrences. (From Toronto's open TTC Delay Dataset)

Bound	n	ranking
E	3174	1
W	3279	2

Table 7: Bound rankings by number of delay occurrences. (From Toronto's open TTC Delay Dataset)

unique_id	Month	Day	Station	Bound	Delayed	travel_time
1	August	Saturday	CHRISTIE STATION	W	0	21.73145
3	December	Sunday	SHERBOURNE STATION	E	0	19.55909
10	November	Sunday	DUNDAS WEST STATION	E	0	20.79242
12	June	Friday	PAPE STATION	W	0	22.44426
15	August	Wednesday	OSSINGTON STATION	W	0	16.18878
16	December	Thursday	OSSINGTON STATION	E	0	21.08448

Table 8: Sample of the simulated station entry dataset.

References

Chan, K. (2019, August 08). Urbanized. Retrieved December 23, 2021, from https://dailyhive.com/toronto/ttc-toronto-subway-station-ridership-2018

Flack, D. (2019, May 11). These are the ideal travel times between TTC subway stations. Retrieved December 23, 2021, from https://www.blogto.com/city/2017/02/ideal-travel-times-between-ttc-subway-stations/

TTC. (2014, December). TTC Operating Statistics. Retrieved December 23, 2021, from https://www.ttc.ca/Coupler/Short_Turns/Operating Statistics/index.jsp

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27(4), 435-454. doi:10.1017/pan.2019.11

Lazaro, E. (2020). TTC Subway Delay Cause Analysis. TTC Subway Delay Cause Analysis.