CSC343 Term Project Proposal

Jonathan Gabe, Nikolas Till

November 19, 2020

1) Domain and Datasets

Our domain of interest is transportation in Toronto via the TTC. Specifically, we are interested in the common commuter experience of delays on the TTC. We will use data from Toronto open data portal to investigate delay by type of transportation, the effects on ridership and revenue and the typical causes of delays.
We found the following data sets on the open Toronto data pool to investigate these topics.

a) **TTC Delays on buses**
Source: https://open.toronto.ca/dataset/ttc-bus-delay-data/

The relevant data in this table is the reported date, time, vehicle, delay, location, route and incident. Vehicle is the bus identification number. The delay is the duration of the delay in minutes. Location is the intersection or bus stop where the delay occurred. The route is the TTC route that the bus followed and the incident is a description of the cause of the delay. Further research might be required to understand the descriptions of the cause of the delay. We will consult TTC and other resources as needed. The reported date is the date of the delay. We used the readMe accompanying the data to understand the meaning of these column labels. In order to avoid redundancy in our relations, we will extract vehicle, delay, reported date and time into its own relation since these are common attributes between TTC delays on buses, trains and streetcars. We will assign a unique delayID to each of the tuples in this new relation. We will also add a vehicle type attribute to our schema. The unique attributes of TTC delays on buses will form their own relation. See the schema section for more details.

b) **TTC Delays on trains**
Source: https://open.toronto.ca/dataset/ttc-subway-delay-data/

The relevant data in this table is the reported date, time, route, location, incident, delay, and vehicle. Many of these attributes are the same as the attributes described in part a) above. We used the readMe accompanying the data to understand the meaning of these column labels. The unique attributes of TTC delays on trains will form their own relation. See the schema section for more details.

c) **TTC Delays on streetcars**
Source: https://open.toronto.ca/dataset/ttc-streetcar-delay-data/

The relevant data in this table is the date, time, station, min delay, line, and vehicle. Many of these attributes are the same as the attributes described in part a) above. The station and line attributes represent the subway station and subway lines in the TTC map. We used the readMe accompanying the data to understand the meaning of these column labels. The unique attributes of TTC delays on streetcars will form their own relation. See the schema section for more details.

d) **TTC Ridership**
Source: https://open.toronto.ca/dataset/ttc-average-weekday-ridership/

The relevant data in this table is the year, period (i.e. month) and the value, which is the number of actual average weekly TTC riders by month (in 000s). We obtained this information from the readMe file that accompanied the data set. This data spans 2007-2020. We will only use 2019-2020 data for the sake of this project. Therefore, we will remove 2007-2018 data and remove the column describing each tuple as TTC ridership. Since this table is very similar to the TTC revenue table below, our schema will not duplicate the month and year attributes. Ridership and revenue will be grouped into the same relation, along with the attributes month and year. See the schema section for more details.

e) **TTC Revenue**
Source: https://open.toronto.ca/dataset/ttc-ridership-revenues/

This table is formatted similarly to the TTC ridership table above. The relevant data is the year, period and the value, which is actual revenues from TTC riders by month. These descriptions were obtained from the readMe accompanying the data set. Similar to above, we will remove 2007 - 2018 data since we are only using 2019-2020 data. We will filter the measure column to only include TTC ridership revenues and then delete this description column in order to clean up the data. As discussed above, we will add the revenue to the ridership table.

2) Questions

1. Which months of the year have the most delays and how does this impact TTC ridership and revenue?
2. What types of delays cause the longest delay? What are the most frequent types of delays?
3. Which streetcar, buses, and trains (by vehicle number) have the most delays due to mechanical issues?

3) Schema

Relations

- Vehicle(<u>number</u>, type)
  A tuple in this relation represents a TTC vehicle. number is the vehicle number of the vehicle. type is the type of vehicle.
- Delay(<u>delayId</u>, vehicleNumber, date, time, delayTime)
  A tuple in this relation represents a TTC delay instance. delayId is the unique id of a delay. vehicleNumber is the vehicle number of the vehicle that was delayed. date is the date of when the delay occurred. time is the time (24h) of when the delay occurred. delayTime is the amount of minutes the delay lasted.
- TrainDelay(<u>delayId</u>, station, line, delayTypeCode, description)
  A tuple in this relation represents a TTC train delay. delayId is the unique id of the delay. station is the name of the station where the train delay occurred. line is the subway line where the train delay occurred. delayTypeCode is the train delay type code for the train delay. description is the description of the delay type code.
- StreetcarDelay(<u>delayId</u>, location, route, streetcarDelayType)
  A tuple in this relation represents a TTC streetcar delay. delayId is the unique id of the delay. location is the description of the location where the streetcar delay occurred. streetcarDelayType is the streetcar delay type for the streetcar delay.
- BusDelay(<u>delayId</u>, location, route, busDelayType)
  A tuple in this relation represents a TTC bus delay. delayId is the unique id of the delay. location is the description of the location where the bus delay occurred. streetcarDelayType is the streetcar delay type for the bus delay.

- RidershipRevenue(<u>month, year</u>, ridership, revenue)
  A tuple in this relation defines the numbers of riders and the amount of revenue earned by the TTC in a given month and year. Month and year are the unique id of TTC data. ridership is the quantity of riders per month and revenue is the dollars earned by the TTC in a month.

Integrity Constraints

- Delay[vehicleNumber] $\subseteq$ Vehicle[number]
- Vehicle $\in$ {"train", "streetcar", "bus"}
- TrainDelay[delayId] $\subseteq$ Delay[delayId]
- TrainDelay[station] $\in$ {x | x is a valid TTC subway station}
- TrainDelay[line] $\in$ {"YU", "BD", "SHP", "SRT"}
- TrainDelay[delayTypeCode] $\in$ {x | x is a valid TTC delay type code}
- StreetcarDelay[delayId] $\subseteq$ Delay[delayId]
- BusDelay[delayId] $\subseteq$ Delay[delayId]

Please note there are 198 TTC delay type codes so we excluded them from the formal definition of the integrity constraint. The valid TTC delay type codes can be found in the ttc-subway-delay-codes.csv in the TTC Delays on trains dataset.