CSC343 project phase 1

Sagnik Adusumilli, Shimon Nauenberg

November 19, 2021

Domain

For our project, we chose to focus on cycling in Toronto. Cycling is a fun and enjoyable form of exercise, and has become an increasingly efficient method for travel in Toronto. We want to investigate how the city infrastructure such as bike parking, location of TTC stations, placement of bike lanes etc., influence the way cyclists commute in Toronto. The data sets we are using are dated from 2019 and earlier.

Data sources

- Bike shop data. This dataset contains locations and information of Bike shops in Toronto. We are only going to store the location and whether they offer rentals All other attributes are irrelevant to our research.
- Red light camera data. This dataset contains information about the traffic lights in Toronto. We are only going note the location of the each traffic light, as the other information present about the traffic lights are irrelevant with respect to our domain of interest.
- Bike parking data (high capacity parking)
- Bike parking data (post and rings)
- Bike parking racks.
- These 3 datasets regard bike parking in Toronto. From the different bike parking datasets we will record the location, capacity, and type of bike parking spots.
- Traffic volume at intersections. From this dataset we will record the location at which the traffic volume was measure and also record the total number of cars present.
- Bike share Toronto data. From this dataset we will get the location of the bike share stations and also compute the average check-in and checkout number for each station.
- TTC subway usage data We retain the average number of people traveling from each TTC subway station across different years. This will be used to check if there is any relationship between the number of bike share stations located nearer to the station and number of people utilizing those stations.
- Bike repaired station data:
  - spreadsheet of data
  - scraped from google maps

Additional learning Needed

- The bike share data is fairly complicated and has datasets spanning multiple years. We will have to read the metadata to understand how to navigate the dataset better.
- We will also have to use some R or python programming to clean the get some summary statistics of datasets that will be put into our relations later.

Data cleaning/transformations needed

- The datsets appear to have no missing values.
- However, There are several datasets that contain information about bike parking spots. We will attempt to combine all this data into one relation. There might be duplicate information present. We will look at the address and remove repeated data.

Questions

1. Do streets with bike lanes report more traffic volume? Could there be other causes for this? For example, These streets could also be having more traffic lights and therefore increasing the traffic volume.
2. Is there a relationship between the areas with more biking parking spots and less number of traffic lights? Do these areas also have increased number of bike share stations
3. Which streets are best for biking under different criteria, such as less traffic lights, parking availability etc? We could list good streets to bike on based on these different preferences
4. Are people using bike share stations to get to TTC subway stations? We can look at bike share stations that are closer to a subway station and see if people return more number of bikes to those stations compared to stations that are farther away.

Schema design

Relations

Relation Name	Description
Streets	Each tuple contains info about a street
Intersections	Each tuple represents an intersection of two streets
BikeShops	Each tuple contains info about a bike shop
SubwayStations	Each tuple contains info about a TTC subway station
BikeStation	Each tuple contains info about a Bike share station
ParkingSpots	Each tuple contains info about a bike parking spot

- Streets(stName, trrafficCountAvg,trafficlightCount, bikeStationCount, subwayStationCount, bikeShopCount)
- Intersections(stName1, stName2, avgtraffic)
- Bikeshops(stName, stNumber, hasRental)
- SubwayStations(stationName, stName, stNumber, hasRepairStand, hasBikeStation)
- BikeStations(id, stName, stNumber, checkinAvg, checkoutAVg)
- ParkingSpots(id, stName, stNumber, capacity, type)

Refer to the data dictionary table for the description of the attributes of the tuples

Integrity Constraints

we made all the street names have the same name in all the relations, so that we have less attributes overall

- Intersection[stName1] $\subseteq$ Streets[stName]
- Intersection[stName2] $\subseteq$ Streets[stName]
- SubwayStations[stName] $\subseteq$ Streets[stName]
- Bikeshops[stName] $\subseteq$ Streets[stName]
- BikeStation[stName] $\subseteq$ Streets[stName]

Data dictionary

Attribute	Description	Type
stName	Name of the street	TEXT
stNumber	Number describing where a structure is located on a street	UNISIGNED INT
traffiCountAvg	Average number of vehicles on the street at intersections	UNSIGNED FLOAT
trafficlightCount	Number of traffic lights on the street	UNSIGNED INT
bikeStation	Number of Bike share stations on the street	UNSIGNED INT
subwayStationCount	Number of TTC subway stations on the street	UNSIGNED INT
bikeShopCount	Number of bike shops on the street	UNSIGNED INT
street1	Name of the first street in an intersection	TEXT
street2	Name of the second street in an intersection	TEXT
hasRental	Does the bike shop give out bikes for rent	BOOLEAN
stationName	Name of the TTC subway station	TEXT
hasRepairStand	Binary attribute to indicate if the subway station has a bike repair stand	BOOLEAN
checkinAvg	Average Number of biked checked into a bike share station	UNSIGNED FLOAT
checkoutAvg	Average Number of biked checked out a bike share station	UNSIGNED FLOAT
capacity	Number of bikes the parking spot can hold	UNSIGNED INT
type	Type of parking spot	spottype (User Defined)
avgTraffic	average traffic in that intersection	UNSIGNED FLOAT
hasBikeSation	indicates the presence of a bike share station	BOOLEAN

Justification

1. Several datasets in our sources had a postal code along with full address to mark locations. We realized including both address and postal code to mark locations would be redundant. We discarded the postal code and used street number and street name as location marker. Here we made an assumption that all streets are named uniquely in the city.
   However, the main reason we used the street and street number to mark location is that they can be also be used to measure the proximity of location. For example, a bike station located at "24 Yonge st" would be closer to subway station located at "26th Yonge st" than a bike station located at "60 Yonge st". We realize that this system can only measure proximity of objects located on the same street. Ideally we would have used latitude and longitude of objects to measure proximity. Unfortunately many of the datasets we use does not have this information.

2. Many of the relations have the common attribute name stName and stNumber to denote street name and street number respectively. This was done to prevent creating additional attributes that mean the same thing. There is an exception to this however, we created different attribute titles for the names of objects, such as "stationName" and "repariStandName". The reason we did is because the names of these objects does not make it clear, to which object we are referencing. For example a station containing a repair stand will have the same address and name. So while writing queries we might confuse a station for a repair stand and vice versa. Hence we created separate attributes to avoid this confusion.

3. The Streets relation was created to able to compare streets based on different metrics. For example, we might want to know which street has the most bike parking spots, or has the most traffic volume at intersections. To compare different streets by different metrics, we created attributes for different metrics. These metrics are summary statistics that can be computed using SQL operations. However those operations are more expensive and it is inefficient to carry out those operations for each street. Hence we decided to carry compute the summary statistics and put them in the relation beforehand.

4. The relation Intersections was created keep track of street intersections. It could be used check if streets with similar properties could be linked together to make a route. For example if "Yonge Street" has a bike lane and "Bloor street" has a bike lane and they intersect. Then it could be part of biking route with bike lanes. We could keep linking streets this way to find out the largest route with bike lanes.

5. The Bikeshops relation was created to store the information about bike shops. One of the metrics we thought could be used to compare streets is the number of bike shops present on the street. Perhaps streets with more bike shops is a better choice for travelling in case of emergency repairs.

6. The Parkingspot relation was created to investigate the different places and ways to park a bike. We are not sure if the type of bike parking spot has any influence for cyclists. Perhaps cyclists prefer certain types of parking types because they are more secure. We may drop this attribute from our schema design if we don't find any interesting relationships.

7. The subwaysStaions relation is very important. Since we are looking at cycling as a form of transportation, we are interested in how it interacts with other public transportation systems in the city. Further, for this relation we wanted to include information about a repair stand to see if the existence of a repair stand encourages people to bike to and from such a station.

8. The RepairStand relation allows us to investigate if the locations of the repair stands encourage people to bike near those locations because they can be reassured that if a problem arises they will more likely be able to fix their issues.

9. The Bikestations relation gives us information about the use of BikeShareTO. This is Toronto's bike-share program. This is particularly valuable as it gives us information about people's starting and stopping points when they use the bikes. It also gives us information about the length of time they use the bikes. A limitation of this information is that people with memberships are usually limited to only a half hour of use at one time, but can park the bike and continue with their trip with a new bike. Therefore, we might not be able to get accurate information about the total usage.

10. The parkingSpots relation gives us insight into locations in which people bike. The stNumber gives us information about the closest address which will further allow us to determine if there are certain places where people decide to bike. Further, as there are different types of bike parking spots, we can determine which one's are most popular and in which location they are the most popular.