PA2: Manipulating and exploring a database

CIV8760E - Transport data management

Frédérick Chabot & François Delisle

October 6th, 2023

The aim of this practical assignment is to have you manipulate and explore transportation data using database software. You'll get hands-on experience using database manipulation software (SQLite or similar), spatial extension software (SpatiaLite), a geographic information system (GIS) such as QGIS, and data analysis software (Python and/or Excel).

1 Database

The data used concerns the csv file of the road collisions in Montreal between 2012 and 2021. It is a subset of data from the Société de l'Assurance Automobile du Québec (SAAQ), geo-referenced by the Ville de Montréal. As the dataset is quite voluminous, we strongly recommend starting analyses (testing your SQL queries, comparing with Excel, etc.) on subsets of the data (for example, taking only the values for 2020 using SQL queries). Please note, however, that the full database must be used in order to meet the mandates!

2 Mandates

This practical work is divided into four steps to put into practice different skills associated with data management in transportation.

Please note that any SQL queries used to extract the necessary information must be specified in the report. Please refer to the appendix for an example of the expected format. You may not include simple screenshots of your queries. It is therefore mandatory to use SQL to process/filter the data before extracting it into figures or tables. Mapping should be done using QSIS software. Finally, you are free to use any tool you like to make the graphs/tables.

2.1 Descriptive data analysis

For this part, you need to import the data into a database program (from the ".csv" file available) to extract the necessary information. Please present and comment on the following elements in tabular or graphical form:

- the number of collisions per year, month and weekday (including weekend days). It's important to make at least one graph for each (year, month and day);
- collision severity indicator (numerical value) per year and per weekday (including weekend days). It's important to make at least one graph for each (year and day);
- the severity of collisions (categorical value) according to an explanatory factor such as type of collision, type of road, pavement condition, weather conditions, etc. (severity as a function of ...). You must propose at least three graphs (three explanatory factors).

Note that you need to use assumptions (ideally from the literature) to transform the severity values from categorical to numerical. Also, be sure to use the accident severity variable numerically and not categorically when requested.

2.2 Spatial data analysis

This part also requires manipulation using database software. This analysis is carried out at the level of "limites administratives municipales" (municipal entities), a file of which is available on the Ville de Montréal's open data portal. It will also be necessary to use the geobase of the road network of the City of Montreal. The following questions can be partially answered using GIS (QGIS) or database software with spatial extension (SpatiaLite). Explaining the method used, please present a table or graph **(not a map)** and comment on the following elements:

- the number of collisions per entity with and without taking into account the length of the road network (sum of polyline lengths, i.e. a number of collisions per kilometer) per municipal entity;
- the number of collisions per entity with another variable taken into account;
- the average severity of collisions (numerical values) per municipal entity;
- the implications (e.g., number of serious/light injuries) of vulnerable users by municipal entity.

So you'll need to intersect (in other words, a form of "spatial join") certain layers in QGIS to extract a modified database. Next, you'll need to perform your processing (e.g. "GROUP BY") using SQL. Note that it is possible to make SQL queries directly in QGIS, so there's no need to export a layer from QGIS! Note also that taking the road network into account means calculating a new value which is the ratio of the number of collisions to the length of the network.

2.3 Cartographic presentation of various analyses

The aim of this final section is to present some of the previous elements on a map. Please present and comment on at least one of the following three elements (you may also propose another map if you consider it relevant):

- the spatial distribution of accidents as a heat map (tutoriel) specifying the parameters used;
- the spatial distribution of accidents reflecting their severity as a heat map, specifying the parameters used;

- the number of collisions per municipal entity, taking network length into account.

The map must include the essential elements for a good understanding. In addition, you should present the entire island of Montreal on one map, but don't hesitate to add maps that "zoom in" on certain parts of the city.

3 Modalities

This job is done with the same teams as the previous job. Please contact the laboratory manager if you have any problems. A report in PDF or Word format, not exceeding **20 pages**, must report on the mandates of this practical work. The due date is 11:59 p.m. on October 26, 2023. The file must be submitted in electronic format on moodle.

The name of the file must include the following nomenclature: EQ{team number}_TP{number of TP}_{semester of study (A, H or E)}{year of study}. For example, "EQ01_TP1_A23".

Particular attention will be paid to writing (spelling mistakes and such will be penalized, as will poor general organization of the work), counting for one point (5%) on the final mark.

Please consult the Writing guide for civil engineers available on Moodle in the Resources section.

4 Appendix

Example of an accepted query format:

```
SELECT * FROM thisTable GROUP BY columnName
```

If necessary, for section 2.2, you could use other data files concerning boroughs (entities), such as the census. On the Montréal open data portal, you can find 2011.