CSCI-4380 Database Systems, Project Memo

Erik Du, duy4@rpi.edu

Jinghua Feng, fengj3@rpi.edu

Shuo Han, hans8@rpi.edu

Yanting Wang, wangy45@rpi.edu

April 8, 2020

1   Team Members

As listed below, we have four members in our team.

- Erik Du, duy4@rpi.edu

- Jinghua Feng, fengj3@rpi.edu

- Shuo Han, hans8@rpi.edu

- Yanting Wang, wangy45@rpi.edu

2   Datasets

2.1   The location of the data

In our project, we will explore two datasets about New York City: Motor Vehicle Collisions and Daily Climate Data. The Motor Vehicle Collisions can be found at NYC OpenData,`https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95`. While the climate data can be accessed from National Centers for Environmental Information, `https://www.ncdc.noaa.gov/cdo-web/`.

2.2   Dataset description

The dataset of Motor Vehicle Collisions is updated daily and there may exist multiple tuples on each day. The whole dataset contains 1.67 millions of rows and 29 attributes. The attributes we may incorporate in our project are (Crash Date, Crash Time, Location, Number of Persons Injured, Number of Persons Killed, Vehicle type).

The weather dataset is also updated daily and it has one tuple each day. The size of this dataset depends on the date range of crash dataset we explore. The main attributes include (Date, Max Temperature, Min Temperature, Precipitation(Snow, Rain), Cloudiness).

2.3 Any relevant license information

As shown on website `https://opendata.cityofnewyork.us/overview/`, the access and use of crash dataset should follow all of the Terms of Use of NYC.gov `https://www1.nyc.gov/home/terms-of-use.page` and Privacy Policy `https://www1.nyc.gov/home/privacy-policy.page`. Weather dataset is available at no charge and can be either viewed online and downloaded with format of "csv". The access requires a basic registration for an API token as shown on website `https://www.ncdc.noaa.gov/cdo-web/faq#webServicesSection`.

2.4 How you plan to join the two datasets

The two datasets share the same attribute "Date", thus can be conditionally joined by this attribute. The aspects we could explore the datasets may include

- Are traffic crashes related with weather conditions such as precipitation, cloudiness?

- Compare crash frequency that occur on weekend, weekdays, holiday, etc.

- Compare the amount of precipitation accumulate in different seasons

3 Work Plan

04/06/2020 - 04/12/2020: Datasets downloading and setup database, python virtual environment properly. Complete `datasets.txt`, `retreive_data.py` and `requirements.txt`

04/13/2020 - 04/19/2020: Define schemas properly and load data into Postgres database. Complete `schema.sql` and `load_data.py`

04/20/2020 - 04/26/2020: Explore the data and relations. Complete `application.py`

04/27/2020 - 05/07/2020: Create User Interface, record the videos and submit the project. Complete `README.md`