Concordia University

Course CEBD1160 : Introduction to Big Data Technology

Instructor: Dr. Nesrine Zemirli

Student Name: Ilia Kassianenko

Student ID: S29059318

Date submitted: April 13 2018

CEBD 1160 Introduction to Big Data Technology

==========================

1.Pick application domain

Decription of projet:

The purpose of this projet is to identify the quantity of snow moved for the City of Montreal to a designated disposal area depot, for the 2016 to 2018 snowfalls seasons. Each depot has a specified capacity, except for a sewage depot, where the runoff sewage systems capacity is not specified. Also, the quantity of daily transactions per depot is considered to be the quantity of visits to unload snow at a specific depot. The daily frequency for a depot is obtained by dividing the quantity of visits for a depot by the total quantity of visits to all depots on a specific day.

Additionally, the amount of precipitation is recorded at the Montreal YUL international airport as an average depth in centimeters. Also, it is assumed for this project that the quantity of precipitation recorded at the Montreal YUL international airport is treater as the uniform distribution of precipation across the area, which is being consider for this analysis. Therefore, all locations serviced by vehicules with a contract ID are assumed to have received the same amount of precipitation as the aforementioned airport.

Each daily snow volume unloaded at each depot is divided by the recorded daily average depth of the precipitation. The result from the division is an order of magnitude estimate for the daily surface area that has received precipitation and that has been served by the fleet of vehicules.

Also, by dividing the quantity of visits by the surface area where snow was removed, a service ratio is obtained for a specific depot and day. The difference in time for a transaction between the snow loading and unloading time is considered to be the cycle time. The project herein does not consider the effect on the cycle time following reassignement of vehicules to depots.

Finally, this analysis will produce for each depot the daily frequency of use and the daily service ratio. Consequently, the analysis will show if constrained capacity depots are used more frequently daily than a sewer depot, and the analysis will show at which depot there is an high quantity of vehicules unloading snow.

--------------------------

1.1 Identify the six V:

Value: Yes

A plot of depots against a service ratio axis and a frequency axis will provide an order of magnitude perspective on underperforming depots.

Real-time data analytics can provide guidance to inform vehicules when to unload snow at a depot that is underused.

Visibility: Yes

All transaction and contract records are publicly available on the City of Montreal portal, which include the vendor name of each contract number that corresponds to a transaction. Also, the Montreal YUL international aiport precipitation records are publicly available.

Velocity: No, the data is pre-recorded before being visible.

Data is not visible in real-time. The data recorded for snow loading and unloading are recorded in real-time; however, the report with this information is not produced in real-time for use on the City of Monttreal portal. Also, precipitation is collected at a specified time interval. Thereafter, the daily average is available on the Government of Canada portal. Finally, snow loading transaction are only recorded during seasons with expected snowfall, which are chosen to be between September 1st and April 1st for this project. Consequently, data is not recorded during the entire year.

Volume: Yes

Even for non-sequential dates, there are several transactions ID for each dates and times. Furthermore, there a several transactions per depot and a contract ID is assigned to each transaction.

Variety: Yes

There are several input variables, such as transaction ID, dates and times, depot location, snow volumes unloaded, and snow amounts on the ground.

Veracity: No, the data is not fully accurate.

There are recording errors for the loading date, and there are notable differences between the cumulative volume of snow recorded at the disposal area at the depot does and the cumulative volume of snow recorded for all transactions. The former is resolved by using the unloading date, which is frquently within several hours from the loading date. However, the latter cannot be corrected and the amount of snow recorded for each transaction is used as is.

Furthermore, the amount of snowfall is recorded as missing on the Canadian Government climate records data records. The recorded amount of precipiation is used as is; there is a disclaimer that the data has undergone basic quality checking.

==========================

2. Select data set (one to two sources, in order to store on personal computer)

(volume unloaded) <http://donnees.ville.montreal.qc.ca/dataset/contrats-transaction-deneigement>

(depot capacity) <http://donnees.ville.montreal.qc.ca/dataset/depot-neige>

(optional) <http://donnees.ville.montreal.qc.ca/dataset/secteur-deneigement>

domain: <http://climate.weather.gc.ca/historical_data/search_historic_data_e.html>

(sub-domain) [http://climate.weather.gc.ca/climate\_data/daily\_data\_e.html?hlyRange=2008-01-08%7C2018-04-08&dlyRange=2002-12-23%7C2018-04-08&mlyRange=%7C&StationID=30165&Prov=QC&urlExtension=\_e.html&searchType=stnName&optLimit=yearRange&StartYear=2010&EndYear=2018&selRowPerPage=25&Line=3&searchMethod=contains&txtStationName=montreal&timeframe=2&Day=8&Year=2016&Month=1#](http://climate.weather.gc.ca/climate_data/daily_data_e.html?hlyRange=2008-01-08%7C2018-04-08&dlyRange=2002-12-23%7C2018-04-08&mlyRange=%7C&StationID=30165&Prov=QC&urlExtension=_e.html&searchType=stnName&optLimit=yearRange&StartYear=2010&EndYear=2018&selRowPerPage=25&Line=3&searchMethod=contains&txtStationName=montreal&timeframe=2&Day=8&Year=2016&Month=1)

===========================

3. Describe your data set

The City of Montreal provides a record for the volume of cubic meters of snow delivered to a snow disposal area. A transaction ID is recorded for each lot of snow delivered to the snow disposal area on specific day. The contract ID, snow loading and unloading times, depot ID are recorded for each transaction. A snow unloading date is analogous to a delivery date for a depot.

Based on the delivery dates, the periods under review are January 2016 through March 2016, September 2016 through February 2017, and December 2017 through March 2018.

There are several recorded transactions with an erroneous or illogical loading date; for these case, the unloading date is selected as the reference. Based on an analysis of a small sample, the variance between the loading and unloading date is not more than several hours. However, there is a transaction, when the loading and unloading times are two days apart.

For a cumulative volume of snow recorded at the disposal area at the depot does not equal to the cumulative volume of snow recorded for all transactions. Specifically, for the 2017 to 2018 season,

There is a variance of approximately 300,000 cubic meters of snow, and the variance for the 2016 to 2017 season is (270,000) cubic meters of snow.

Each vehicule associated with a contract ID is not bound to visit only one depot per day, based on review records for the 2017-2018 season.

--------------------------

The data from station 702S006 is collected for the data recorded at the Montreal/Pierre Elliott Trudeau international airport. An individual report will be manually generated for years 2016, 2017, and 2018, by downloading the CSV daily data. The command line, which is detailed on Government of Canada website for "Past weather and climate" under the "Get More Data" option, will not be used. Only total precipitation records will be considered for each day.

The amount of precipitation is used as is, without any consideration for the recorded amount of snow on the ground. This precipitation is defined as being rainfall and a water equivalent to the amount of snowfall. Consequently, a portion of the precipitation recorded is not snow by definition.

The recorded amount of snow on the ground is the defined as the average daily depth of the snow. I am considering this snow to be the net amount remaining between the effect of snow and rain fall and the effect of evaporation, without any mechanical or manual removal.

--------------------------

Residential snow removal contracts frequently have an effective start date of November 15th and an end date of April 15th, based on personal experience. However, based on the availabilty of records from the City of Montreal for snow disposal, the peridod under analysis will be September 1st through April 1st. This period will be considered for years 2016, 2017, and 2018 when observing the precipitations and snow records, irrespectively of the mean temperatureS between the aforementioned dates that are recorded at station 702S006.

===========================