Workflow

Building a database

Toronto data

Made one big table and then wrote a python script to cycle through csv's and bulk insert
Adding in data for indexes: day of week and time of day and then indexed: by start-end location name, day of week, and time of day `add_columns_indexes.sql`

Road shapes

Downloaded Center Lines from Toronto Open Data Portal. Inserted the monitoring stations into a PostGIS table with a sql script (`monitoring_stations.sql`) and then mapped them in QGIS

Holidays

Downloaded python module `workalendar` that contains holidays for 2015 for Canada (amongst others). I uploaded this to PostgreSQL with a python script `holiday_insert.py`. Since Friday travel before a long weekend tends to be different from "normal" I also added Fridays preceding Monday holidays to be excluded.

Descriptive Analysis

Identify timeperiods

I wanted to get a sense of which days of the week and timeperiods were similar. First I examined variance by day to see which days were similar while excluding holidays (`dow_ttvariance.sql`) . Based on this I came up with the following groupings: Monday, Mid-Week, Friday, Saturday and Sunday. (`daytypes.sql`)

Variation in median travel time by Time of Day

From these "day type" groupings, I wanted to examine how travel time varied over the day for each segment and each day type to see if segments were similar and in order to visually identify peaks for the different day types. For every minute of the day, the sql script took the median day's observation of the median travel time on that segment. Segments were divided between East and North-Bound ("Outbound") and West and South-Bound ("Inbound") (`to_stations_dir.sql`). Plotting was done using iPython notebook (`to_data.ipynb`)

Identifying number of slow days:

`slow_days.sql`

If I had more time

1. Automate timeperiod and day type discovery, instead of visual inspection write a script that would cycle through the number of time periods and day types and calculate the variance for every grouping to minimize within group variance and maximize it between groups.
2. Identify jams/slowdowns by identifying start and end times of travel times being above X times free flow travel time as well as peak delay for each event then identify whether they are recurring or unique occurences and where they might start.
3. Dive deeper into linking weather and travel time from Environment Canada data I got