Josh Castillo

Jaime Lopez

Stephen Rondain

Kingsley Wong

**CS 129.1 Final Project Report Paper**

**Big Data Problem**

Given a dataset of MRT Line 3 daily passenger traffic, how do we determine rush hours, and areas of congestion at a given time period. The outputs will be shown through different charts.

Here are the expected outputs that are to be shown in the charts:

**Definition of Terms**

Rush hour: Time period with the maximum in and out, or total traffic.

Congestion: Congestion is the amount of people inside a station at any given hour.

**Context**

According to Open Data Philippines, “the Metro Rail Transit Line 3 (MRT-3) is a Project Management Office under the Department of Transportation in charge for the operations of the MRT3 System in accordance with the Build-Lease-Transfer Agreement signed between the government and the private sector.” There is a dataset available for each year from 2012-2014 that shows the daily number of passengers who enter and exit a specific station at a specific time period (per hour).

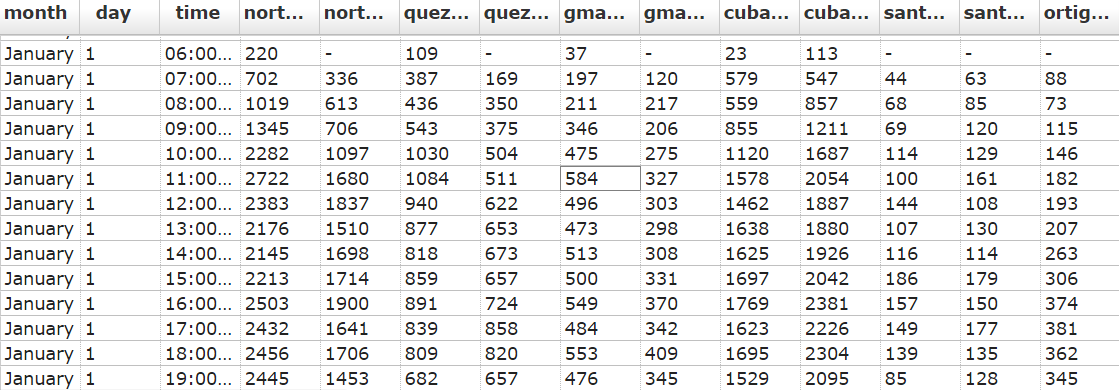
**Content**

The Metro Rail Transit Line 3 dataset includes 8,784 record entries which includes the following:

1. Passenger traffic in each station as an entry point
2. Passenger traffic in each station as an exit point
3. The time period (hourly basis)

Note: These records are organized in a way that each day of each month of 2012 has data on each station’s entry and exit point. Moreover, it is arranged per day starting on January 1, 2012 with time periods on an hourly basis.

Here is an example/preview of the dataset:

****

**Description of the Data Source: Open Data Philippines (data.gov.ph)**

<http://data.gov.ph/dataset/metro-rail-transit-line-3-passenger-traffic-daily>

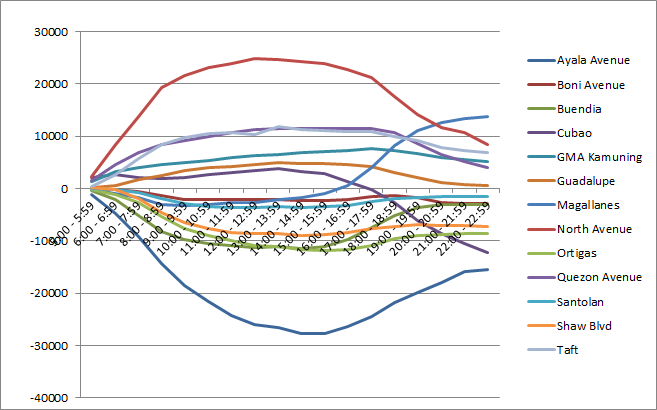
“Open Data Philippines aims to make all open government data accessible through one single portal, data.gov.ph. Open Data Philippines collects datasets from different government agencies, allowing users to find specific data from a continuously growing collection of public datasets. “

The Metro Rail Transit Line 3 has datasets which contain the daily passenger traffic from the different stations along line 3 in a day to day basis (month, day, time, [station]). The datasets span from the year 2012 to 2014. From this dataset we can extrapolate the possible rush hours and areas of congestion. The dataset is in default csv format and can be downloaded through the website. The team chose the dataset for 2012 as it is the most complete when compared to the other datasets available.

**Description, Explanation, and Discussion of the Output**

**Output 1 - Congestion**

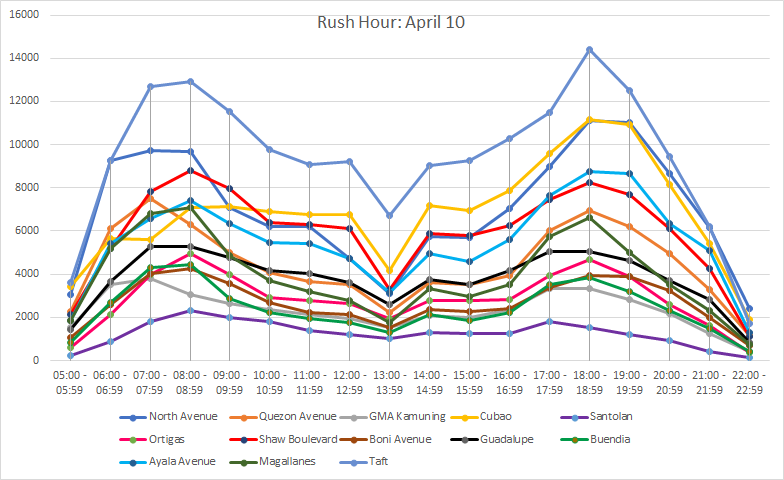
For congestion, outputs are grouped by date and station. For each day and station, it will output all the times from 5:00-22:59 and their corresponding congestion value. The congestion value is calculated from adding the amount of people entering a station minus the amount of people exiting the station. The congestion value is accumulated, starting with 0 at 5:00, and is not reset to 0 again until the next day. There are some cases where a negative value may be encountered, this simply means that the station is primarily a destination since the amount of people exiting the station is greater than the amount of people entering.



**Output 2 - Rush Hour**

For the measurement of rush hour, outputs are grouped by date and station. For each day and station, it will output all the times from 5:00-22:59 and their corresponding flow value. The rush hour is determined by the flow value which is measured as the aggregate of the entry and exit of persons per station, per time period. The time period with the highest flow will determine the rush hour for that particular station on that day.

The line graph below shows the corresponding amount of flow(Y Axis) per station(lines), per time period (X Axis). The “peaks” that fall onto each corresponding station indicate the amount of flow and the time period in which it occurred - basically signifying the rush hour for that particular station, on that specific date.

****

**Output 3 - Rush Hour Leaderboards**

By further map-reducing the output 2, we can get the amount of times an particular time period has been crowned ‘rush hour’.