KE5208: Sense Making and Insights Discover

CA Project

Documentation

|  |  |
| --- | --- |
| Team 9  Members: |  |
| Choong Yue Lin | A0150392L |
| Ho Kok Loon | A0150144W |
| Lee Tin Onn | A0150375J |
| Lee Wai Tong Peter | A0150406U |
| Lei Jun | A0150141B |
| Vincent Leung | A0150347L |

Contents

[1 Introduction 1](#_Toc466687568)

[1.1 Objectives 1](#_Toc466687569)

[1.2 Approach 1](#_Toc466687570)

[2 Strategies 2](#_Toc466687571)

[3 Insights Discovered 3](#_Toc466687572)

[3.1 Correlation between speed band and traffic incident 3](#_Toc466687573)

[3.2 Traffic Condition on Weekdays, Weekends and at arterial roads 5](#_Toc466687574)

[3.3 Traffic Condition versus Weather Condition 5](#_Toc466687575)

[4 EPN Design 7](#_Toc466687576)

[5 EPN Implementation 8](#_Toc466687577)

[6 Conclusion 11](#_Toc466687578)

[7 Reference 12](#_Toc466687579)

# Introduction

## Objectives

The objectives of this project is to make use of data science technology to reduce traffic jam conditions in Singapore roads, particularly at the arterial roads leading to expressways. The highlight of the project is to provide online real-time alerts to vehicle drivers when traffic starts to build up on these roads. In addition, the project will alert drivers when accidents occur on certain roads. The alert is generated by a set of Event Processing Networks (EPNs) which is designed based on insights discovered from past data. The insights will help the alert system to predict traffic conditions instead of just doing reporting. Various factors such as weather and peak hours will be taken into consideration.

## Approach

Prior to designing the alert system, the project analyses various sources of real-time data that attribute to traffic conditions. In addition to monitoring online traffic conditions, this project considers other factors that may cause traffic jams, such as heavy rain, electronic road pricing (ERP) gantries. The project collects a period of online data from various reliable sources, e.g. traffic conditions and ERP gantries from Land Transport Authority (LTA), and weather information from National Environment Agency (NEA). It uses analytical software tools to discover insights from the data collected in order to set criteria for the alerts.

After analysis is done, the project is designed with Event Processing Networks (EPNs) based on the set of criteria. The networks are then implemented as the alert system that can be used by vehicle drivers. This project demonstrates the implementation by developing a software prototype.

# Strategies

The project starts with identifying the factors that may cause the traffic conditions to fluctuate:

* Temporal factor
  + Traffic generally builds up at peak hour. It may be different on weekdays from weekends.
* Spatial factor
  + However, not all arterial roads are fully crowded, and different roads are crowded at different portion of peak hours. We keeps track of the locations of the arterial roads and the timing of the traffic. So this factor is related to temporal factor
* Weather
  + Whenever it rains, drivers slow down their vehicles for safety purpose. This will cause traffic to build up. But it may not happen to every road, depending on the road structure. By analysing the location of the rainfall, we can match it the traffic condition and identify if there is any correlation.
* ERP locations
  + Not all the arterial roads have traffic jam, but it is always perceives that those with ERP gantries will be crowded at the last few minutes of the operating hours. ERP locations thus become another possible factor.

The traffic condition data and weather data are all extracted externally from websites such as data.gov and weather.gov using web scraping scripts and form into their own respective data groupings. In addition, we retrieves traffic conditions from LTA but using [DataMall@MyTransport.sg](https://www.mytransport.sg/content/mytransport/home/dataMall.html) 1 . This is an LTA Open Data Initiative that provides a set of API for developers to download data on real-time basis. By collecting a few days (or even weeks) of data, we can analyse the data and identify some trends in it.

Based on the data collected, we analyse it by combining various factors. Speed is always the dependent variable in correlation with the various factors (independent variables). If there is high correlation between speed and the factors, then we can determine the speed band at a particular location at a particular point of time based on the factors. For example, whether the speed of most vehicles is low due to a number of road works and accidents or rainfall.

Combining all these various factors, we can design the EPN that specifies the various criteria. We can then specify what to react (e.g. what kind of alert message) when a particular event exceeds the pre-determined threshold at a particular decision, and show the output to the consumer.

# Insights Discovered

## Correlation between speed band and traffic incident

According to LTA2, a speed band can be categorized as below. For this project, when we receive a speed band of 1, we will regard it as a major jam.

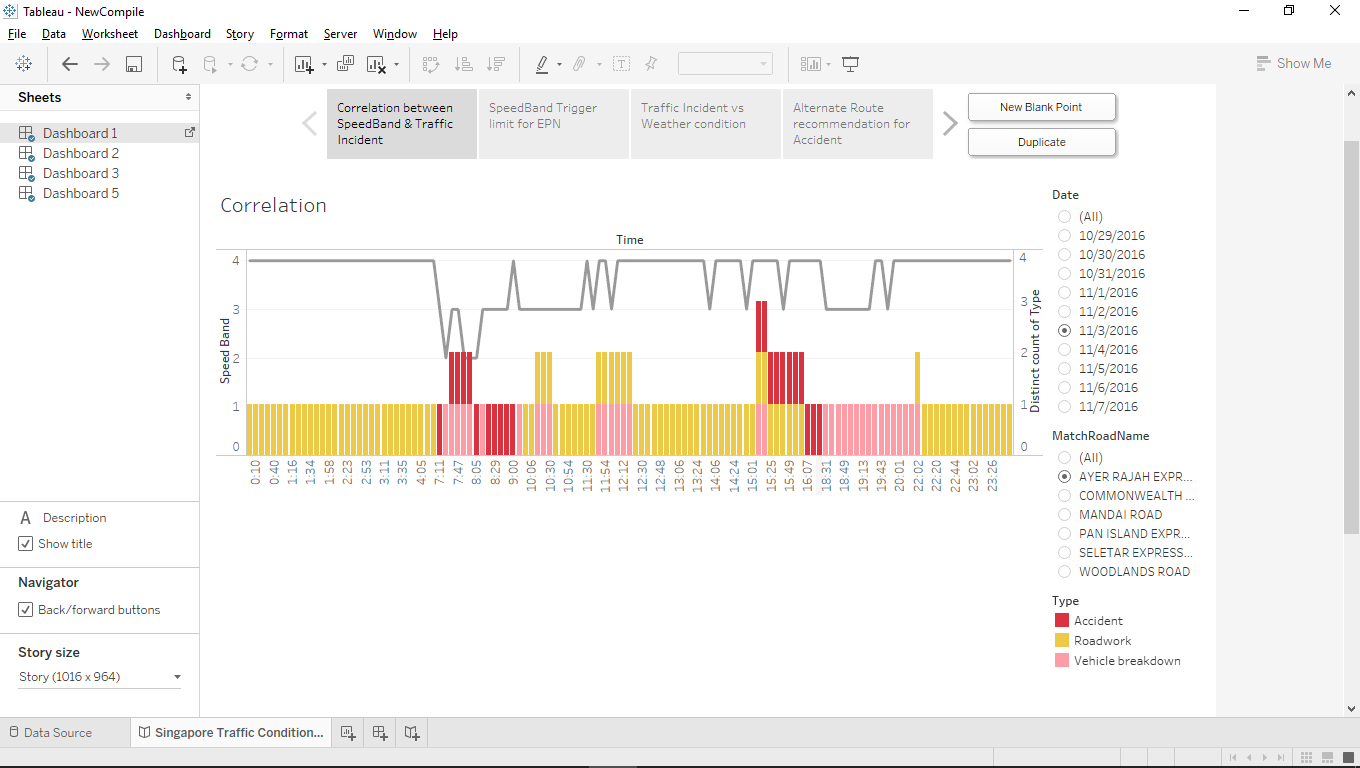
1 – indicates speed range from 0 < 20

2 – indicates speed range from 20 < 40

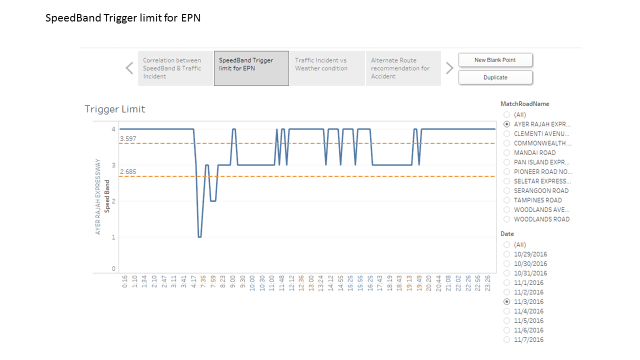
3 – indicates speed range from 40 < 60

4 – indicates speed range > 60

Using Tableau Desktop software, we constructed a graph that shows the correlation between speed band and traffic incident in a particular day, for a particular location. The example in the graph below shows that, on 3 Nov 2016, at Ayer Rajah Expressway, the traffic was smooth (the thick line at speed band 4) when there is no incident along the way. However, when there was road work, accident or vehicle breakdown during the peak hours, vehicles were tend to slow down at speed band 3. When two incidents (accident and vehicle breakdown) happened at morning peak hours (about 7am to 8am), the situation became worse and the speed band went as low as level 2.

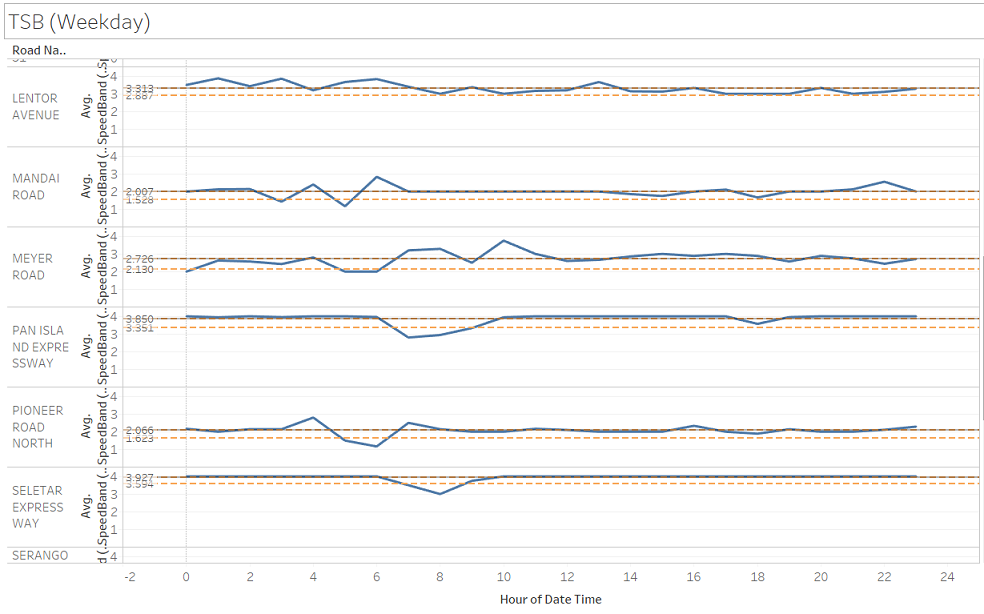


To identify when an event is considered beyond acceptable level (that is, an event falls beyond 99% probability), we compute the mean and 3 x the standard deviation (also known as six sigma) on the information collected over a period of time (as shown in the graph below). Any speed that goes beyond 3 x standard deviation from the mean will trigger an alert message. This rule is captured in the EPN when we design EPN in the next stage.



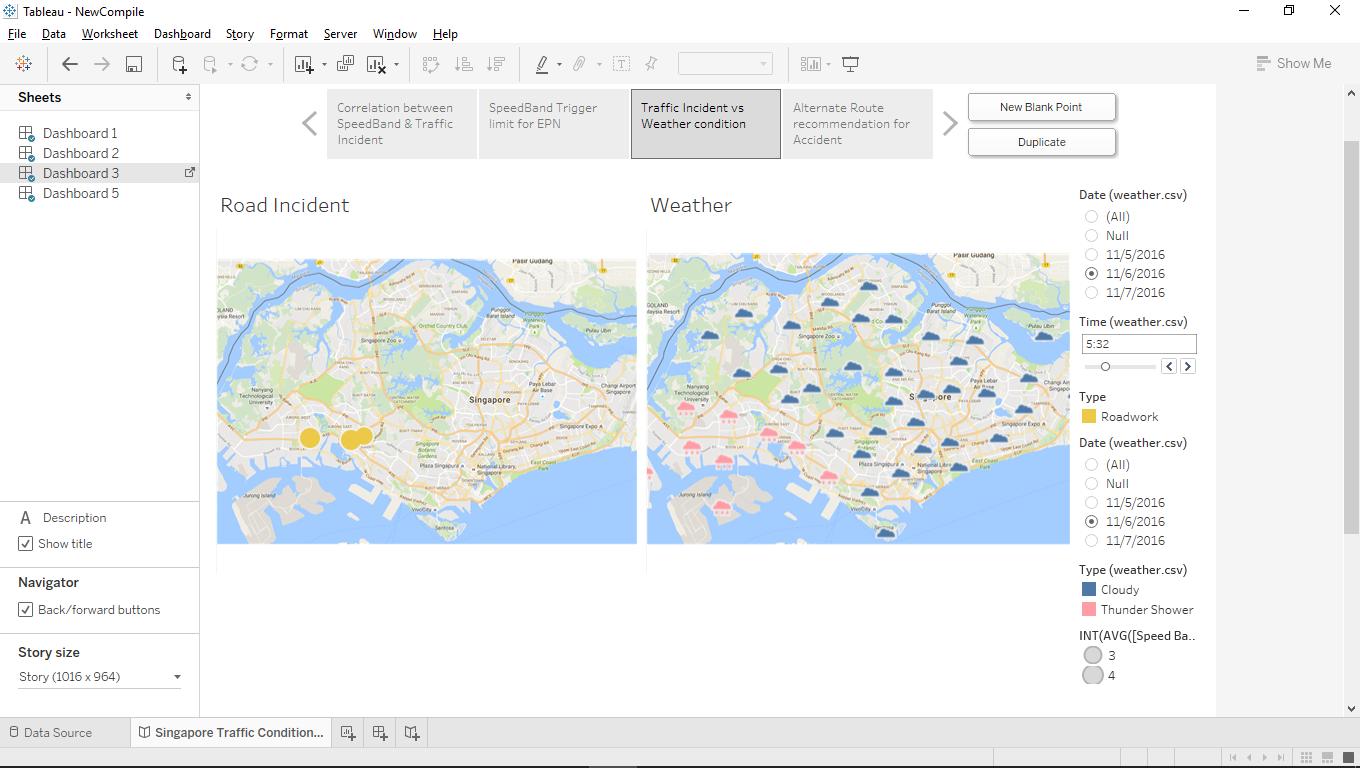
## Traffic Condition on Weekdays, Weekends and at arterial roads

Making use of the road conditions provided by LTA, we generate a graph that shows all the conditions at various arterial roads. In addition, we break it down into weekdays and weekends as the traffic conditions for these two periods are different.

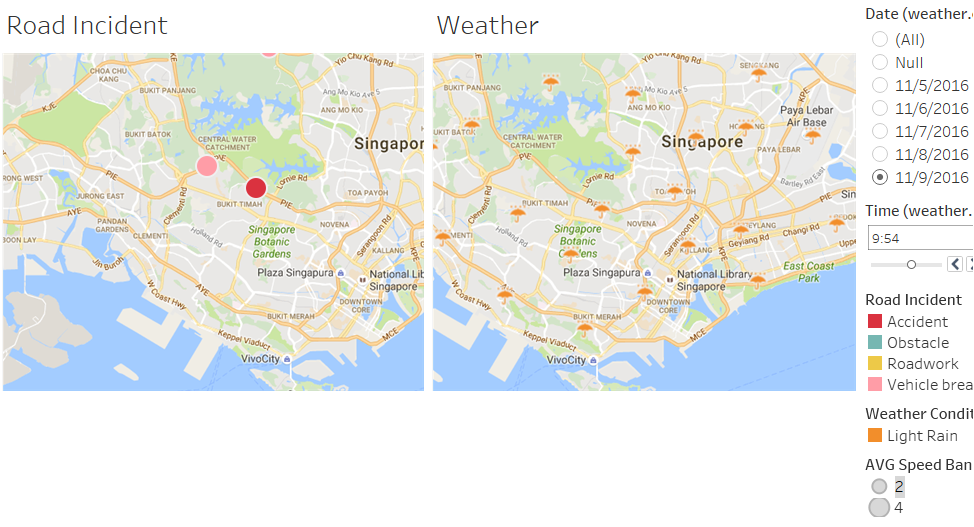


## Traffic Condition versus Weather Condition

Using Tableau Desktop, we analysed both the conditions of traffic and weather through visualization. In the example below, on 6 Nov 2016, there was thunder shower at the west south part of Singapore. In addition, there were some roadworks happening around the same area where there was shower. However, the speed band stayed at around 3rd and 4th level, showing that the traffic was satisfactory.

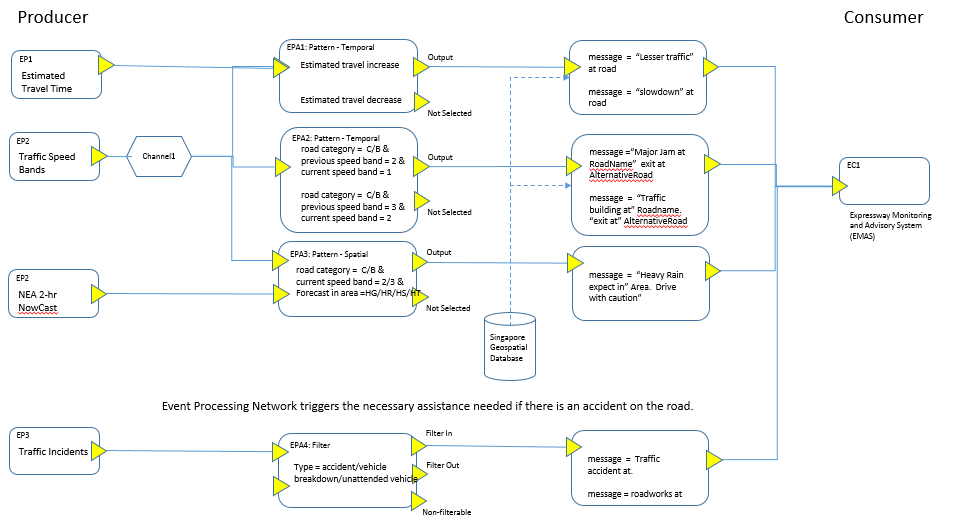


However, a few days later, on 9 Nov 2016, we noticed that vehicle breakdown and accidents caused more impact to road condition even though it was just light rain across the island. The speed band dropped to level 2. Weather seems to play less impact on road condition, unlike what we have perceived.



# EPN Design

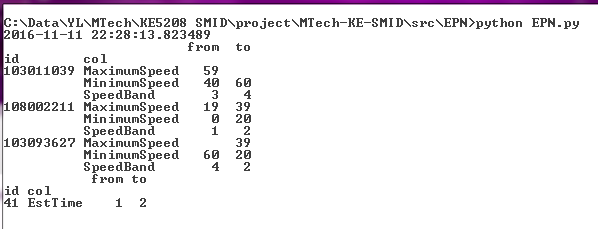
Based on the insights we have gathered, we design the Event Processing Network (EPN) that will trigger the messages at each of the individual arterial roads so as to improve road conditions at major expressways. Below is the complete EPN design.

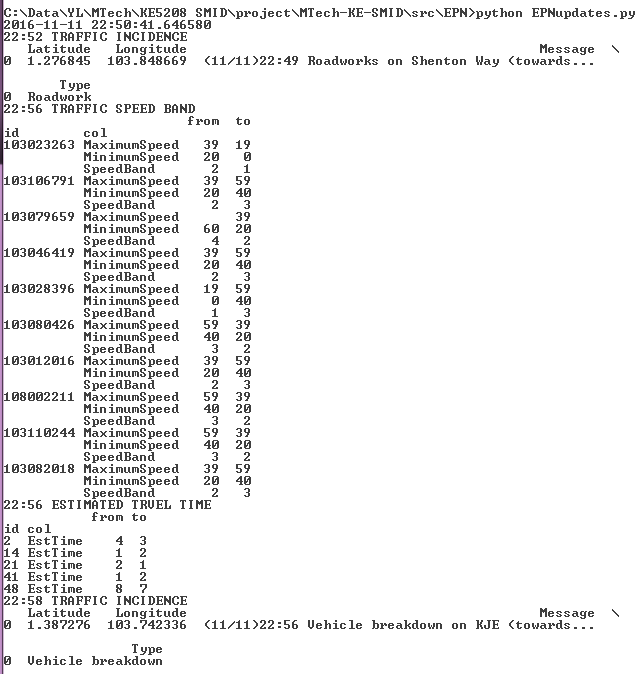


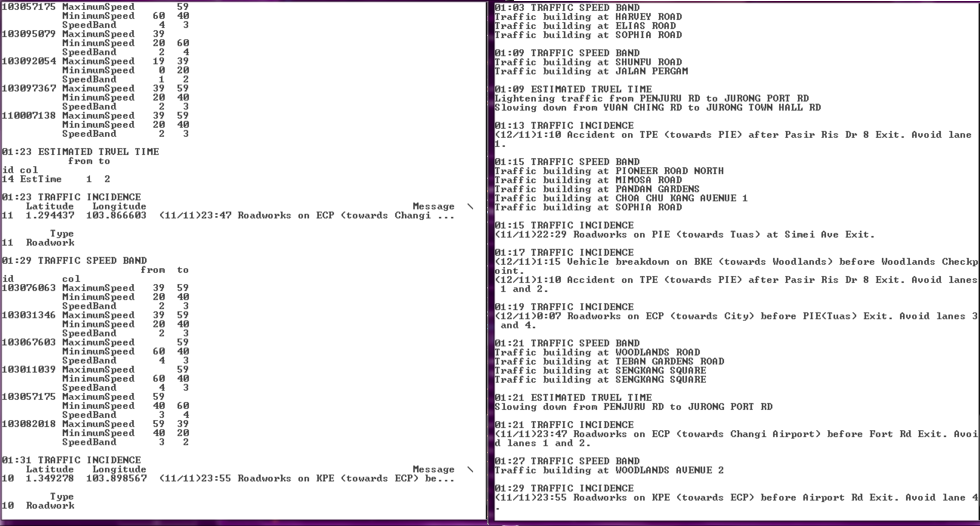
# EPN Implementation

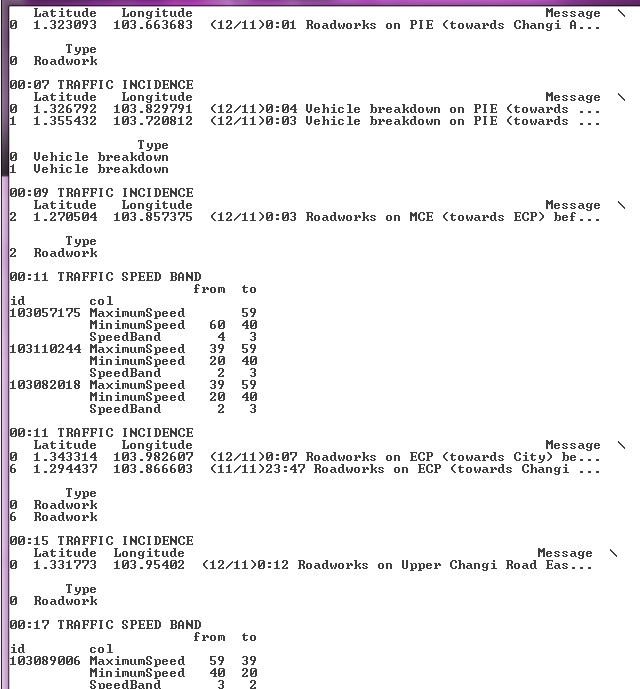
The EPN is implemented using Python coding. It takes in online traffic information in real time through the API available at datamall@LTA. As for the weather, we wrote Python program to do web scrapping from NEA website 3. The website forecasts weather every two hours.

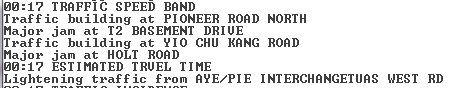
Our EPN Python program receives input from the above real-time sources, then processes each event and decides whether to trigger certain alert. Below are some screenshots that demonstrate the output generated by the EPN program after processing every incoming event.

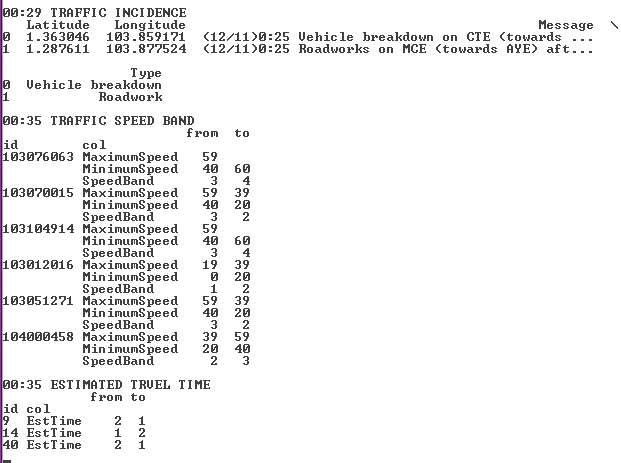


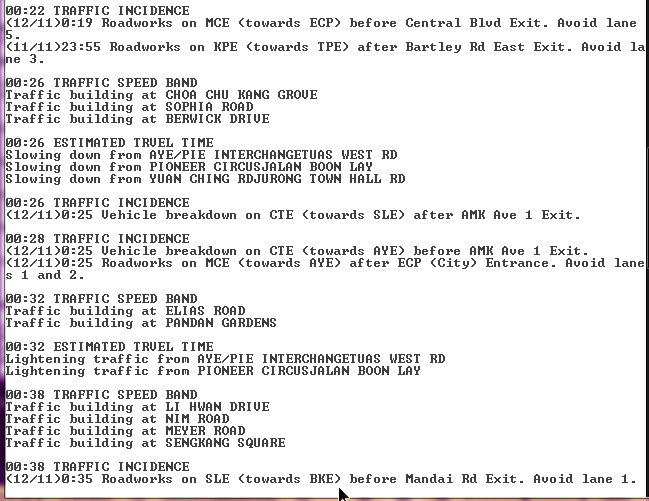










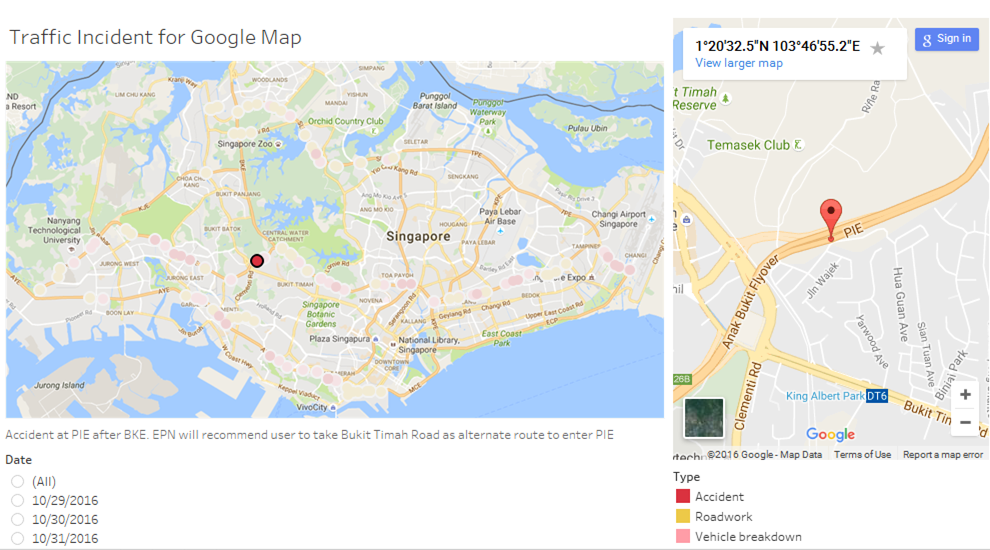


# Conclusion

From this exercise, the team acquired the opportunity to practise the various steps and approaches in designing and implementing an EPN right from capturing online data and discovering insights from this data.

Due to time constraint, the data we have collected is less than a month. As the duration is short, the insights we have identified is limited and may not be very accurate. For example, we can hardly identify the linear regression between traffic condition and rainfall amount as it did not always rain in the past few weeks and often at off-peak hours. This may have affected the accuracy of the EPN outcome. But we have tried to keep the EPN process as complete as possible.

Our project prototype focuses on implementing EPN according to our design closely. As such, a lot of attention is spent on following the rules stated in the design, deciding every event and triggering the appropriate alert. Hence, due to time constraint, we emphasize less on the user interface of the EPN output. If time permits, we can present the output on Google map and display alerts beside it, as illustrated with the following mock up using Tableau Desktop.



# Reference

1. datamall@LTA

<https://www.mytransport.sg/content/mytransport/home/dataMall.html>

1. LTA API user guide

<https://www.mytransport.sg/content/dam/mytransport/DataMall_StaticData/LTA_DataMall_API_User_Guide.pdf>

1. NEA 2-hourly forecast

<http://www.weather.gov.sg/weather-forecast-2hrnowcast-2/>