Data Wrangling Project Phase I

By : Hinam Mohit Mehra (Student number : 653167)

1. Title of Project : What effect does Weather have on Metropolitan Trams, and Trains Delivery and Punctuality ?

2. Domain : Transport and Weather.

3. What is the question you are seeking to answer? Who would be interested in an answer to this question and why? How might the information be used and who could it benefit?

I would like to know what effect weather conditions like rain, and heat waves can have on Melbourne's Public Transport system. This information will be useful to Public Transport Victoria (PTV), because if we want to reduce traffic congestion from the city and reduce Co2 emissions by encouraging people to choose public transport as their primary option to commute, then we will have to make public transport more reliable even in adverse weather conditions, especially in adverse weather conditions. We already know that Melbourne trains are overcrowded ("Melbourne Transport Crisis"), and this issue gets worse when trains or trams are delayed. My analysis will recognise the weather conditions that causes delays in public transport. Using that data, PTV can employ extra services on days, or even specific times of the day when those weather conditions are met.

4. In what respects will your answer to this question provide innovative information? (you do not want to have a question which is trivial, or for which the answer already publically exists and can be readily found).

My analysis aims to depict the extent to which weather causes disruptions to the normality of the public transport system. There are occasional PTV tweets which warn passengers of delay because of wet weather, or because of swelling up of tracks during a heat wave. But my answer aims to report that delays due to a certain type of weather happen constantly, and that PTV shouldn't dismiss these disruptions as a one type occurrence.

5. Datasets: What are the datasets (minimum of 2, maximum of 3) that you will use? Provide a brief description of the information in each and a link (URI) to where the dataset can be downloaded.

The PTV website releases Daily Observations about public transport's delivery and punctuality service. As part of Right to Information Act, i have emailed PTV asking them for daily observations for the year 2015. If that information will not be available, then i will access the Disruption data using PTV's timetable API (https://www.data.vic.gov.au/data/dataset/ptv-timetable-api). For weather conditions, i will use Melbourne's Rainfall and Temperature dataset (http://www.bom.gov.au/climate/data/index.shtml?bookmark=136).

6. What difficulties and challenges do you envisage in processing, integrating and visualising these datasets to answer your question?

The main challenge will be in obtaining the Daily Observations data for a year. If that is unsuccessful, then there will be a bigger challenge of accessing the PTV transport API in Python, since according to discussion forums on the web, the API documentation is missing some pieces.

7. In what ways will your processing, integration and visualization add value compared to having just the raw data?

   My visualizations will consist of graphs grouped by weather conditions. For example, a bar graph showing weather conditions that count as a heat wave on the y-axis, and daily delivery and punctuality on the left axis. This will show to what extent different weather phenomenons affect the normal public transport timetable. This will help us see a pattern, which wouldn't be possible with just raw data.

8. How much code (in Python) do you estimate will need to be written from scratch? What are the major Python libraries that you will make use of? What other other publically available code do you intend to use?

If i am successful in getting the data from PTV, then i would estimate writing under 100 lines of code in Python, using the CSV module to parse the data, and using Plotly to build graphs. Otherwise, if i have to access disruptions data from the timetable API, then i would estimate writing about 200 lines of code, using the JSON decoder module to obtain the data, and averaging out number of disruptions to form my own estimate about daily delivery and punctuality percentages.

References

"Melbourne Transport Crisis". *NewsComAu*. N.p., 2015. Web. 3 Apr. 2016.