Data Wrangling Project Phase 4

By : Hinam Mohit Mehra (Student number : 653167)

Title : What effect does Weather have on Melbourne's Metropolitan Trams' Performance?

Domain : Transport, Weather

Question : The analysis aims to find a relationship between poor performance of trams and weather. The analysis will be performed on each of the 23 routes currently in service to see how pronounced is effect of weather on their performance. And which weather conditions affect performance the most. Lastly, the consistency of this relationship between poor performance and weather will also be tested. A descriptive data mining model will be used to answer these questions.

Datasets :
3 datasets were used for this analysis.
The first dataset named 'Tram - Operational Performance 2015' was obtained in an XLSX via email from Yarra Trams. The dataset consists of daily punctuality and delivery percentages for the entire network and for each of the 23 tram routes currently in service. It also consists delivery percentages for the City Circle tram.

The Disruption API available on PTV's website only consisted of disruptions that was caused by accidents or services upgrades. And daily operational reports on their website was only available for the previous week, hence they were requested for the year 2015. Only 2015 was chosen to keep my analysis relevant in light of the changes and upgrades that have been made in the past. Network level data didn't consist of any missing values. Route level punctuality and delivery data consisted of 10 missing values each for route 12 from 2, Jan - 11, Jan and 2 missing values for route 11 from 14, Feb - 15, Feb. In addition, City Circle tram data had 33% of its values missing.

The second and third dataset consisted of Daily Rainfall and Daily Maximum Temperature respectively. Both of these datasets were obtained from http://www.bom.gov.au/ (rainfall - http://goo.gl/aIuX11 , max. temperature - http://goo.gl/uxxcn8) in CSV format.
The station Melbourne Olympic Park (Station number - 86338) was used because in comparison to other stations it had the least number of missing values i.e. 1 for rainfall data on 11, June and none for temperature data.

Rainfall and heat were considered as instances of bad weather. Because, wet weather makes gripping difficult thus slowing down tram speeds. And high temperature causes tracks to swell up and increases pressure on trams air conditioning systems resulting in power outages.

Pre-Processing :
Missing value methods :
No records were missing from maximum temperature data. Only 1 record was missing from rainfall data which was filled out as 0mm because rainfall on the previous and next day was also 0mm. The row was not deleted because on the corresponding day some of the tram routes showed poor performance hence, analysing this day was crucial.

As for Tram performance data, all the missing values for Route 12 and 11 were filled using the Network level performance data of that day, which was the average daily performance of all the routes. Missing values for City Circle trams wasn't filled out.

Column deletion :
City Circle tram's delivery column was deleted because 33% of its values were missing and a column for punctuality didn't exist either. In addition to that, the dependency on the city circle tram amongst Melbournians is not that high hence the data wasn't an important part of my analysis.
From all the datasets, the column specifying the Year was deleted because my analysis only focused on one year - 2015 and thus the column was occupying unnecessary space.
From the temperature and rainfall data, columns named 'Product Code', 'Bureau of Meteorology Station Number', 'Days of accumulation' and 'Quality' were additionally deleted.

Columns replaced :
In the Tram dataset, the 'Month' column had month names specified at the start of each month, otherwise the values were blank. On the other hand, the weather data had month numbers for each row. Hence, for easier integration and to provide more meaning to the data, the Month column in tram data was replaced by the month column from weather data.

Renaming Columns :
'% services on-time over length of route' was renamed to 'Punctuality' and '% timetable delivered' was renamed to 'Delivery' to make it easier to understand.

Converting Column Names (Example of issues encountered) :
Route names in Route level punctuality and delivery data were explicitly converted to string because whilst indexing the tram dataset with route name, python kept considering that as a column number rather than a column name.

Integration :
Each dataset was loaded as a dataframe variable using pandas.
For initial investigations, new dataframes were created by applying thresholds to network level tram data. Poor punctuality was defined as a percentage less than or equal to 77% and poor delivery was defined as a percentage less than or equal to 98%("Performance Monitoring - Public Transport Victoria"). These dataframes were then merged with weather dataframes (rainfall and maximum temperature) on columns 'Month' and 'Day'. The end result consisted of dataframes like 'Network Punctuality <= 77%, Rainfall >= 0mm', 'Network Delivery <= 98%, Rainfall >= 0mm', 'Network Punctuality <= 77%, Temperature >= 0°C' and 'Network Delivery <= 98%, Temperature >= 0°C'.
In addition to that, thresholds were applied to weather dataframes, rainfall greater than 0mm and temperature greater than 25°C which is the average temperature in summer("Weather In Melbourne - Climate & Temperature - Tourism Australia"). New weather dataframes were then merged with network level tram data on 'Month' and 'Day'.

The end result to this consisted of dataframes like ‘Network Punctuality, Rainfall > 0mm’, ‘Network Delivery, Rainfall > 0mm’, ‘Network Punctuality, Temperature > 25°C’ and ‘Network Delivery, Temperature > 25°C’.

For latter investigations, each route’s tram data (both punctuality and delivery) were filtered using the above mentioned thresholds. And it was merged with filtered weather data created using above mentioned thresholds. The two dataframes were again merged on columns ‘Month’ and ‘Day’. The end result consisted of dataframes like ‘Route Name, Punctuality <= 77%, Rainfall > 0mm’ , ‘Route Name,‘Delivery <= 98%,‘Rainfall > 0mm’, ‘Route Name, Punctuality <= 77%, Temperature > 25°C’ and ‘Route Name, Delivery <= 98%, Temperature > 25°C’.

Results :

Initial Visualization :

From the 8 merged dataframes obtained from integrations in initial investigations, 8 scatter plots were drawn. Scatter plots where thresholds were applied to tram’s performance variable are colored blue, and plots where thresholds were applied to weather are colored red.

Interesting Results from Initial Visualization :

Highest rainfall of the year, 30mm had a good punctuality and delivery rate of 88% and 99.5% respectively(figure 1,2). On the other hand, there were days where rainfall was 20mm, the punctuality rate was 70% and delivery rate was 98%(figure 3,4). But majority of high rainfall days did not experience poor performance. Most points on the scatter plot that have low performance lie between 1mm - 7mm(figure 1). Thus, Rainfall didn’t have consistent effect on tram performance.

On the other hand, the lowest punctuality rate 68% of the year was on a day with quite high temperature of 33°C(figure 5). But even then most poor punctuality points lie between 15°C - 25°C(figure 7).
The maximum temperature of the year 40°C had a high punctuality of 85°C but a low delivery of 98%.
The plot also shows that there were consistently low deliveries for high temperatures between 35°C - 40°C(figure 6, 8). Hence, delivery percentage is coupled with temperature but not punctuality.

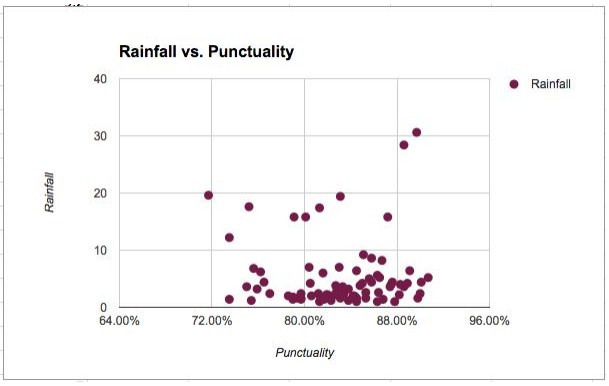

Figure 1: Rainfall > 0mm - Punctuality

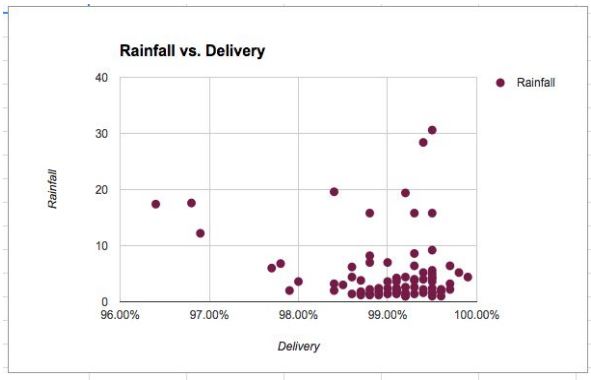

Figure 2: Rainfall > 0mm - Delivery

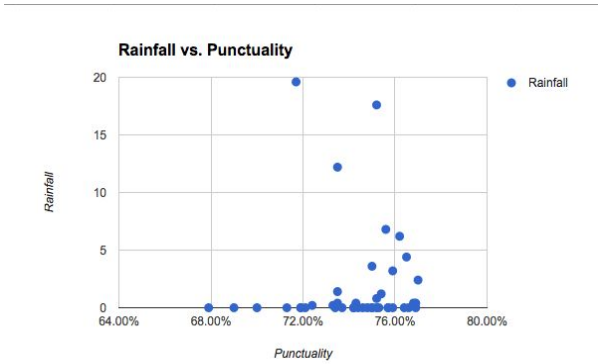

Figure 3: Rainfall - Punctuality <= 77%

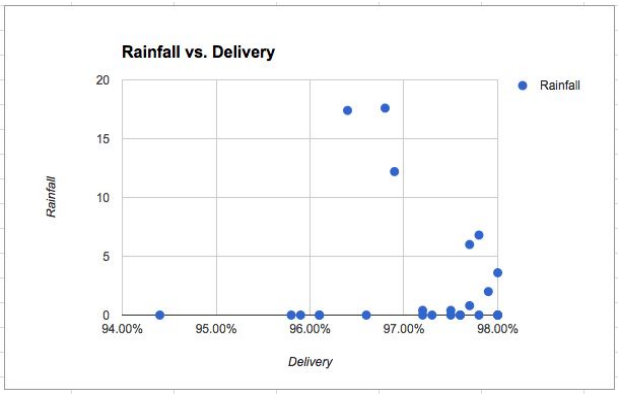

Figure 4: Rainfall - Delivery <= 98%

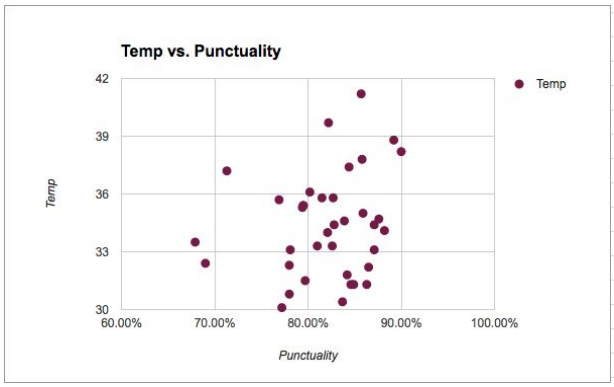

Figure 5: Temperature > 25°C - Punctuality

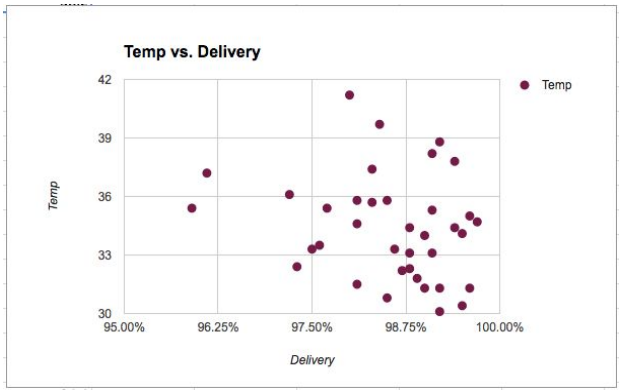

Figure 6: Temperature > 25°C - Delivery

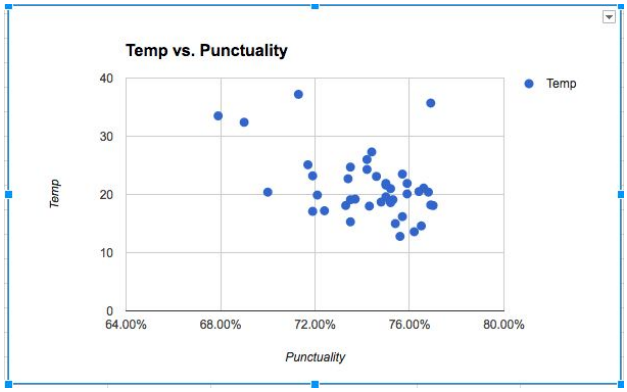

Figure 7: Temperature - Punctuality <= 77%

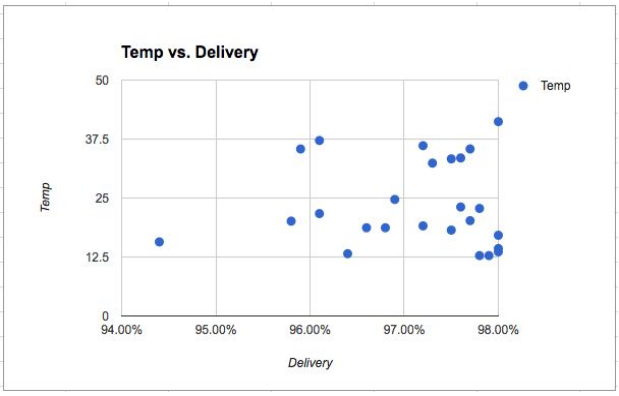

Figure 8: Temperature - Delivery <= 98%

Later Investigations :

Visualization :
Blue colored scatter plots using thresholds on performance attributes and weather were made for each route. These scatter plots can be found in the folder named ‘Plots’.

Clustering :
Using the integrations made earlier which resulted in rows where tram threshold and weather threshold was applied, number of rows, average weather value and max weather value was calculated for each route. After this was calculated for all the routes, k-means clustering was performed with k = 6. The aim of clustering was to divide the routes into groups and judge the impact of a weather condition on their performance attribute. The impact was divided into lowest, low to medium, medium, medium to high, high, highest. k = 6 was used instead of k = 3 because tram routes showed wildly different behaviour (intra cluster distances were high) and 3 groups wasn't enough to show the characteristic of each group. The following tables are the output of the clustering method. The impact is based on Centroid num rows and Centroid average, supported scatter plots for each route(Folder - Plots).

Punctuality - Rainfall

Route Num	Centroid num rows	Centroid Average	Centroid Max	Impact
12	6	0.8	2.2	Lowest
55, 57, 82	7.3	4.5	16.4	Low to Medium
1, 109, 11, 48, 5, 59, 6, 75, 86, 86	20.6	3.6	19.4	Medium
19, 64, 70	31	3.5	18.3	Medium to High
16, 3, 72, 67, 78	50.8	3.5	23.6	Highest
8	69	3.2	19.6	High

Punctuality - Maximum Temperature

64	31	30.5	41.2	High
1, 11, 12, 19, 55, 59	9.2	30.2	36.4	Low to Medium
16, 72, 8	42.7	29.9	38.7	Highest
109, 48, 5, 6, 75, 86	16.8	30.1	37.2	Medium
57, 82	3.5	29.6	34.7	Lowest
3, 67, 70, 78, 96	35.6	30.5	39.5	Medium to High

Delivery - Rainfall

70	24	4.5	30.6	High

1, 109, 11, 12, 55, 57, 59, 75, 82, 96	12.3	4.1	17.3	Low to Medium
3, 6, 64, 67, 78, 8	22.5	4.4	18.9	Medium
16	58	3.2	19.6	Low to Medium
19, 48, 5, 86	9.5	2.4	7.6	Lowest
72	44	3.7	19.6	Medium

Delivery - Maximum Temperature

109, 12, 48, 5, 67, 78, 8, 96	17.1	32	39	Medium
72	34	30.6	37.4	High
55, 57, 59, 6, 82, 86	7.2	31.8	37.8	Lowest
16	39	30.1	41.2	Highest
70	25	31.8	38.8	Low to Medium
1, 11, 19, 3, 64, 75	12.3	33	40.3	Medium to High

Correlation :
For every route, Pearson's coefficient was calculated. Since the coefficient is suppose to show downhill (as rainfall/temperature increases - performance decreases), minimum coefficient was calculated.

	Rainfall	Maximum Temperature
Punctuality	Min : -0.10717981895 Avg : 0.0162390584669	Min : -0.342654369193 (Route 96) Avg : -0.106706064846
Delivery	Min : -0.174437204418 Avg : -0.0233840406675	Min : -0.290976756897 (Route 96) Avg : -0.128538127595

Discussion :
From the Delivery - Rainfall clusters, it is evident that rainfall doesn't have significant impact on tram's delivery. However, the comparatively high correlation coefficient and the high mean, maximum temperature of the Delivery - Temperature centroid shows that tram's delivery and heat are highly coupled.
Punctuality on the other hand is moderately affected by high rainfall and is equally as affected as delivery by heat.

It is interesting to note that route number 12 has repeatedly been in the lower impact end of the clusters whereas route 16 has been in the highest impact end. Some other routes that have shown poor performance under bad

weather conditions are 72, 64, 67 and 6. It is also interesting to note that routes 16, 72, 64, 67 and 6 share majority of their tram route. Hence, maybe the tram route might need an upgrade.

Limitations :
PTV defines punctuality as reaching the stop not 59 seconds before and not 5 minutes after the designated time. Realistically, a tram that is 5 minutes late is considered punctual but is not reliable. Since, those 5 minutes might make you a miss a train that is now in 30 minutes. Hence, if PTV had adjusted its threshold I would have had a lot more poor performance data relevant to my reliability based analysis.

Value :
Looking at the raw data didn't tell us much. It just showed daily performance percentages and daily weather conditions. Whereas scatter plots helped determine coupling between trams performance metrics and weather. Clustering helped us group different tram routes and specify the impact of weather on each of these groups. Clustering also showed the bigger picture of why some trams were repeatedly showing poor performance. For more predictive based models in the future, it was important to derive a relationship between weather and tram performance. Correlation helped us achieve just that and showed there is no linear relationship between rainfall and performance, but there exists a weak linear relationship for temperature and performance for route 96. All of these visualizations helped us see a pattern which wouldn't have been possible with just raw data.

Challenges and Reflections :
Data Gathering :
Well, first it was challenging to understand the type of disruption data that I could get from PTV. It took a long time to get the required data since first an email was sent to get a developer key to access the API. But the API data wasn't enough so I had to look for other readily available datasets I could use. Hence, I finally emailed PTV to send me the data I wanted, which again took a long time to come back.

I wanted to get different, more accurate weather data for each tram route since each route is different and it starts and ends in different suburbs. But it was difficult to get weather data for each route and that would have resulted in a lot of a datasets and a high computational overhead.

Data Processing :
It was difficult to get tram data that purely or largely depended on weather, since there was no way to eliminate rows that were disrupted because of accidents , planned disruptions or ill customers.

Clustering :
The code for calculating clusters was quite challenging because dictionaries aren't ordered in python 2.7 (which is a necessary version for Pandas to work). Dictionaries with route name with keys was used so that cluster attributes could be meaningfully stored. There was no way that a route name could be matched to an index in the labels array with this method. Hence, instead of a dictionary, a two dimensional array was made where the first element was the route name and the second element was a tuple containing cluster attributes. This would mean that the index of the tram route would be preserved in the labels array.

Question Resolution :

From my analysis, weather has significant effect on the performance of trams, particularly high temperatures. Interestingly high rainfall doesn't have an impact on performance, but low to medium rainfall does. This effect is different for each route like non-existent for route 12 and highly pronounced on route 16, as evident from clustering. In addition, my question aimed to answer if this effect of weather is consistent or not. It is evident from correlation analysis, that only route 96's performance is weakly negatively correlated to weather. Hence, weather's effect is not consistent.
This information will be useful to PTV and Yarra Trams to help build a more reliable and faster network, especially in poor weather conditions. Yarra Trams has 700 tweets in a month about disruptions on the network, most of them for unspecified reasons. This analysis will help them judge the amount of disruptions that are caused by weather. A more faster network and reliable network will reduce traffic congestion in Melbourne CBD, and will encourage people to use public transport as their main option for commute which will in-turn reduce C02 emissions and improve sustainability. In addition to that, the issue of over crowdedness in trams can be reduced by identifying weather conditions that cause delays, and thus employing more trams or taking other steps to counteract for their reduced speeds.

Code :
The entire project was done in python from data processing to data visualization, clustering and correlation. 100% of the code written (files phase2.py and phase4.py) was written from scratch.

Pandas was used to process the data. First csv and xlsx files were read in as dataframes. Several methods were applied in the pre-processing like fillna() filled blank missing values, columns were removed using drop(), column names were renamed using rename, dataframes were merged using merge(), route names were converted to strings using rename() and str(). Four functions were created that would clean and fill tram and weather data.

Matplotlib was used to draw scatter plots. A function was created that would take a weather dataframe, tram performance dataframe, hex color and show a scatter plot with tram performance metric on x-axis and weather on y-axis.

Sklearn and Numpy were used for clustering. As mentioned in section 'Results', a tuple of three items - num rows, mean weather metric, max weather metric was calculated for each route. This tuple was converted into a numpy array using np.array() because cluster() method from Sklearn only accepts a numpy array as an argument. The clustering function returns an array of group numbers and centroids.

Scipy was used to find Pearson's coefficient to analyse correlation between performance metrics and weather.

Principles of abstraction were strongly followed while writing code. Functions were extensively used to reduce redundant code and avoid re-computation of unnecessary global variables.

Bibliography :
"Performance Monitoring - Public Transport Victoria". *Ptv.vic.gov.au*. N.p., 2016. Web. 23 May 2016.

"Weather In Melbourne - Climate & Temperature - Tourism Australia". *Australia.com*. N.p., 2016. Web. 23 May 2016.