Tales of London's Busiest and Quietest Stations

Author: Eu Meng Chong
Published on: 15th June 2020

Table of Contents

Introduction

With the world oldest metro system which is known as the London Underground, alongside with the Overground and privately owned commuter trains services, London has one of the most extensive rail networks in the world, consisting a total of 330 stations. Moreover, as mentioned in this webpage, London and Southeast England (counted together as their railway network are more integrated) has an annual passenger count of 1.216 billion between 2018 - 2019 and the trend is steadily increasing.

In this project, we would like to first look into the neighbourhoods surrounding these railway stations in London and classify them whether they are having high volume of trains stopping by or otherwise. Fortunately, the Department of Transport (DfT) has categorised all the stations in the UK (London included) based on that metric from A being the highest to F2 being the lowest. Moreover, as not all the neighbourhoods surrounding each station are made equal (i.e. some neighbourhoods are solely residential, others having more commercial or corporate area surrounding it), we will use the assumption which is the purpose of usage of each station is solely determined by venues closest to a it. For example, if there are no residential areas in a neighbourhood, then the passengers at the particular station could be using it to travel back home for work. This instance is one example of daily migrations of people in a city as populous as London.

What we would like to achieve from this project is to construct a classification model based on these data on each neighbourhood surrounding the stations to determine their respective volume of trains stopping by on their respective stations. In fact, this model has the potential to determine the factors which affect the ridership count on each train stations in London and could be used to pinpoint a profitable extensions of railway network in London.

Data

We will use the following sources to retrieve the abovementioned data:

1. https://en.wikipedia.org/wiki/List_of_London_railway_stations: This webpage provides the list of all the railway stations in London with their respective coordinates, operators, DfT Categories and the Boroughs in London. Here we shall use the Python package of BeautifulSoup to scrape the table from the webpage. Here are the first 5 rows of the table scrapped from the webpage.

	Station	Borough	Managed By	DfT Category	Coordinates
0	Abbey Wood	Greenwich	TfL Rail[1]	C	51.4915,0.1229
1	Acton Central	Ealing	London Overground	D	51.5088,-0.2634
2	Acton Main Line[2]	Ealing	TfL Rail	E	51.5169,-0.2669
3	Albany Park	Bexley	Southeastern	D	51.4358,0.1266
4	Alexandra Palace[3]	Haringey	Great Northern	D	51.5983,-0.1197

2. Next, we shall use the Foursquare API to explore venue types surrounding each station. As a matter in fact, Foursquare also label categories on each venue categories with a more refined sub-categories. We may find such list of categories with its corresponding Category ID here. Here are the example of categories we are interested to look at:-
    - Arts & Entertainment; 4d4b7104d754a06370d81259
    - College & University; 4d4b7105d754a06372d81259
    - Events; 4d4b7105d754a06373d81259
    - Food; 4d4b7105d754a06374d81259
    - Outdoors & Recreation; 4d4b7105d754a06377d81259
    - Professional & Other Places; 4d4b7105d754a06375d81259
    - Residence; 4e67e38e036454776db1fb3a
    - Shop & service; 4d4b7105d754a06378d81259
    - Travel & Transport; 4d4b7105d754a06379d81259

    Here, for each station, we shall only consider all the venues within 1000m radius surrounding it to be its neighbourhood as 1000m is a approximately a the distance travelled in a 15-minute walk, which makes such range a reasonable enough to identify the main usage of each stations.

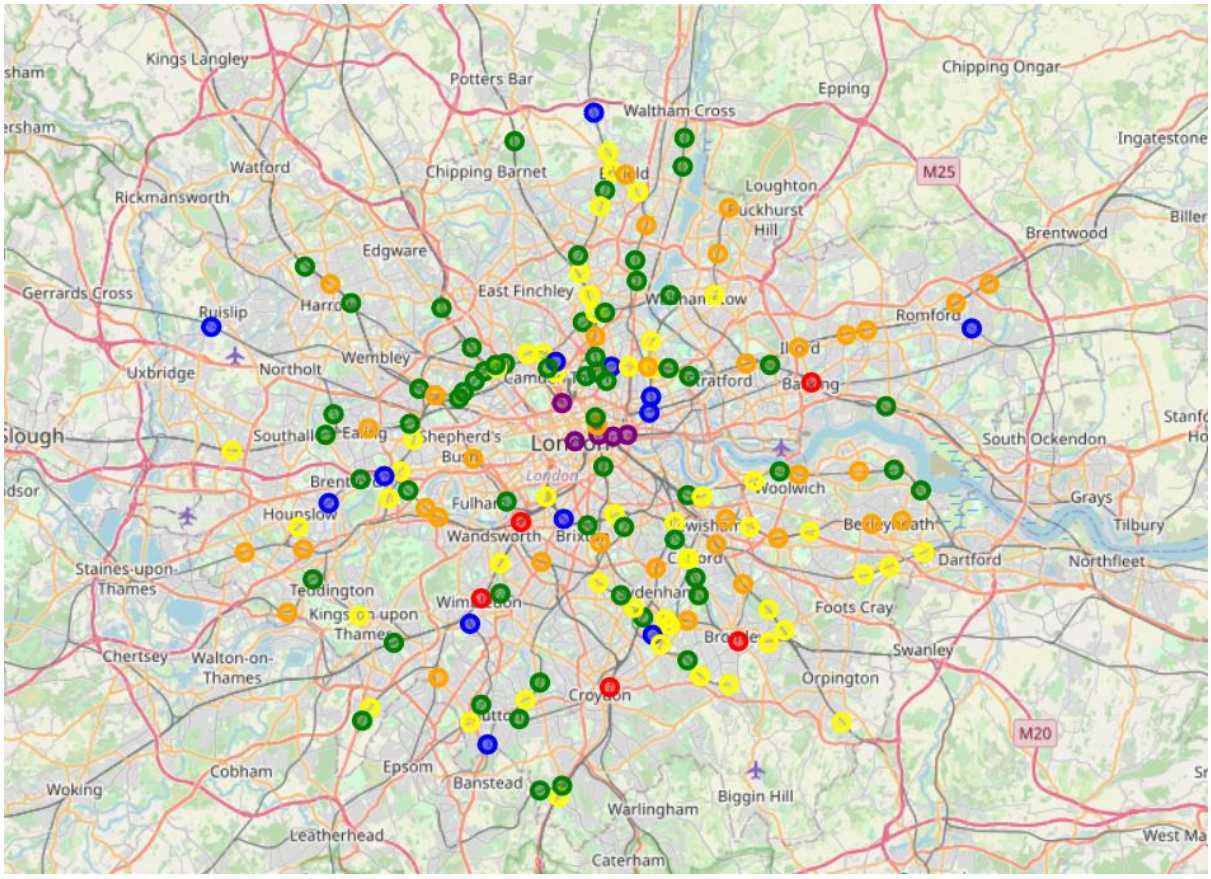
Methodology

Based on the category ID described in the last Section, we shall use the Foursquare API to query the number of venues with respect of each venue categories discussed above within the 1000m radius surrounding each station in London. By incorporating the totalResult value provided from the Foursquare with the table of all the stations in London, we get the following table (only the last 5 rows are shown):-

	Station	Borough	Managed By	DfT Category	Coordinates	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
344	Woodmansterne	Croydon	Southern	E	51.3192,-0.1539	3.0	0.0	0.0	6.0	3.0	2.0	9.0	1.0	1.0	3.0
345	Wood Street[48]	Waltham Forest	London Overground	D	51.5864,-0.0021	4.0	3.0	0.0	18.0	11.0	5.0	19.0	2.0	19.0	4.0
346	Woolwich Arsenal	Greenwich	Southeastern	C	51.4898,0.0694	4.0	11.0	0.0	22.0	7.0	10.0	28.0	4.0	42.0	15.0
347	Woolwich Dockyard[65]	Greenwich	Southeastern	E	51.4913,0.0536	3.0	4.0	0.0	16.0	4.0	5.0	17.0	3.0	30.0	9.0
348	Worcester Park[66]	Sutton	South Western Railway	C	51.3804,-0.2412	2.0	0.0	0.0	9.0	5.0	2.0	6.0	3.0	15.0	7.0

Exploratory Analysis and Data Cleaning

First and foremost, let's observe the table with the input by Foursquare. We have observed that almost half of the data has no input from Foursquare and hence, labelled as NaN (Not a Number). We need to remove these data as these data is practically unusable for the subsequent steps of this project (i.e. Exploratory Analysis, Modelling and Evaluation). Next, we have done some preliminary clean up on the resultant table, notably on the data with ambiguous DfT Categories and Borough outside of London (but practically considered to be in London). After the data is being cleaned, let's display the distribution of the stations with DfT Category labels on the map of London below.

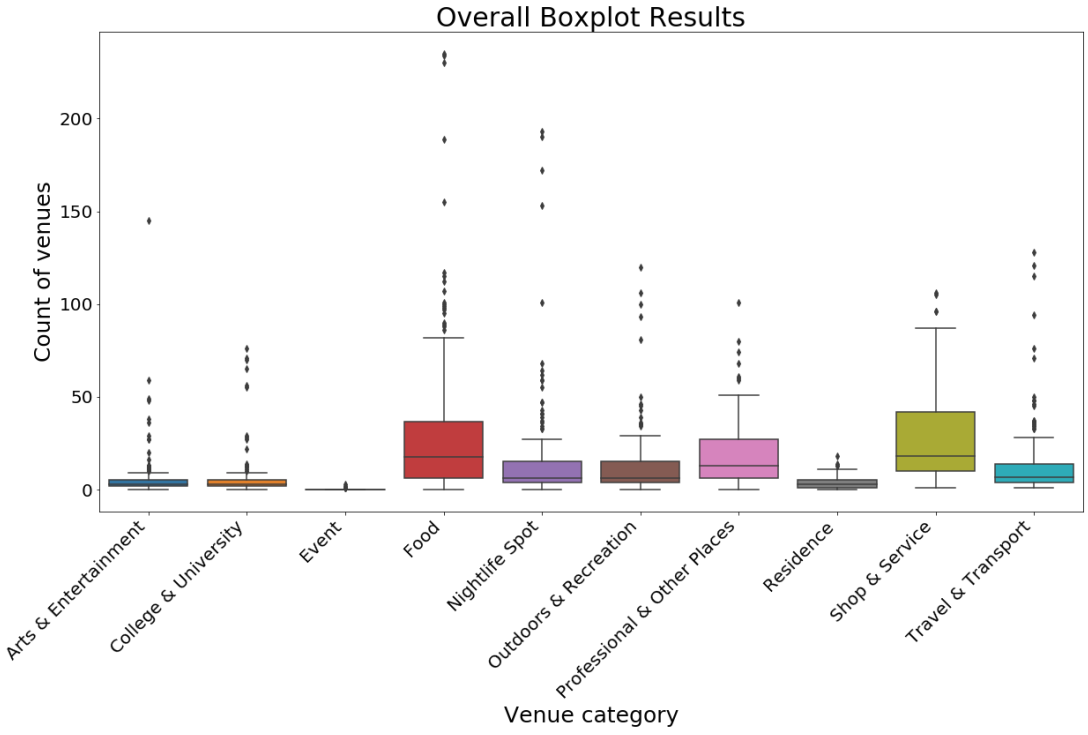


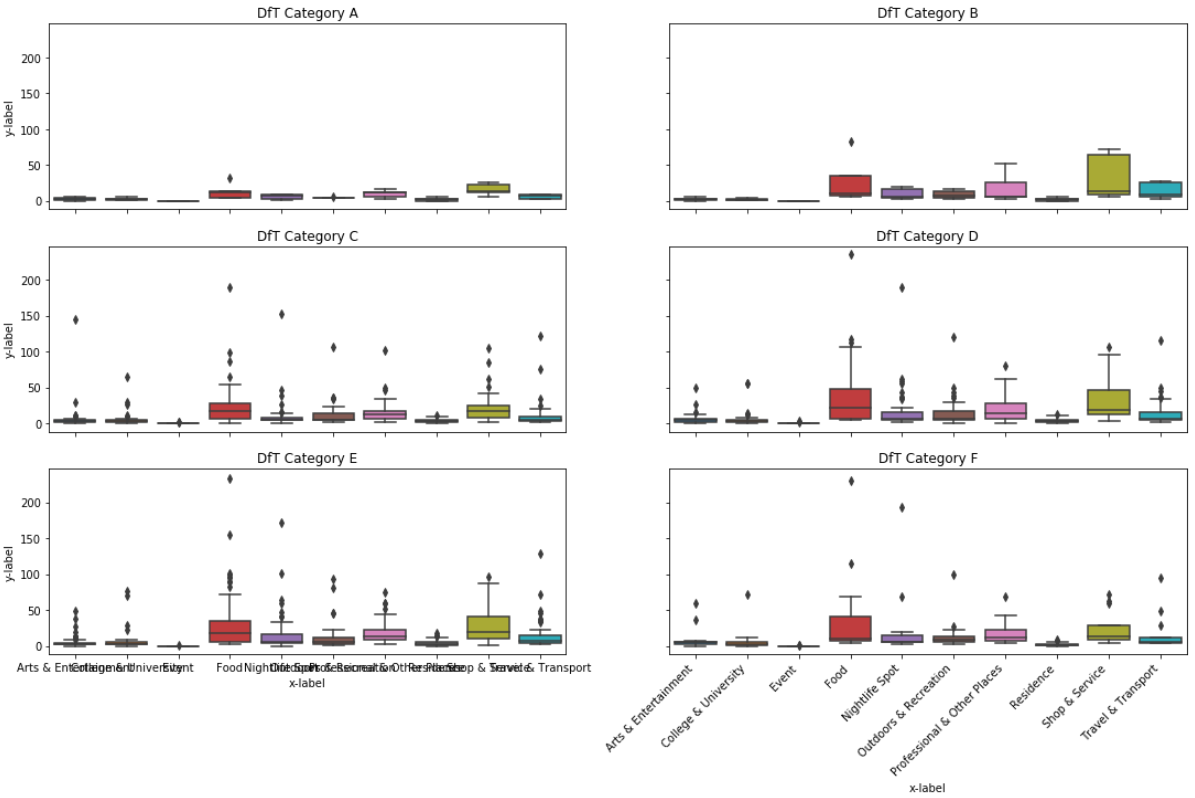
Note: The colour of the labels of each station are corresponding with its DfT Category, with

DfT Category	A	B	C	D	E	F
Colour	Purple	Red	Orange	Yellow	Green	Blue

Based on the labelling of the stations on the map, it appears that the DfT stations does not fit in the central / outer London paradigm, with DfT category B, E, F stations being scattered across central and Outer London. (Stations of DfT Category A is a notable exception.)

Now, we use boxplot to visualise at the overall distribution of the venue categories in the data. We then shall visualise the distribution based on each DfT categories.



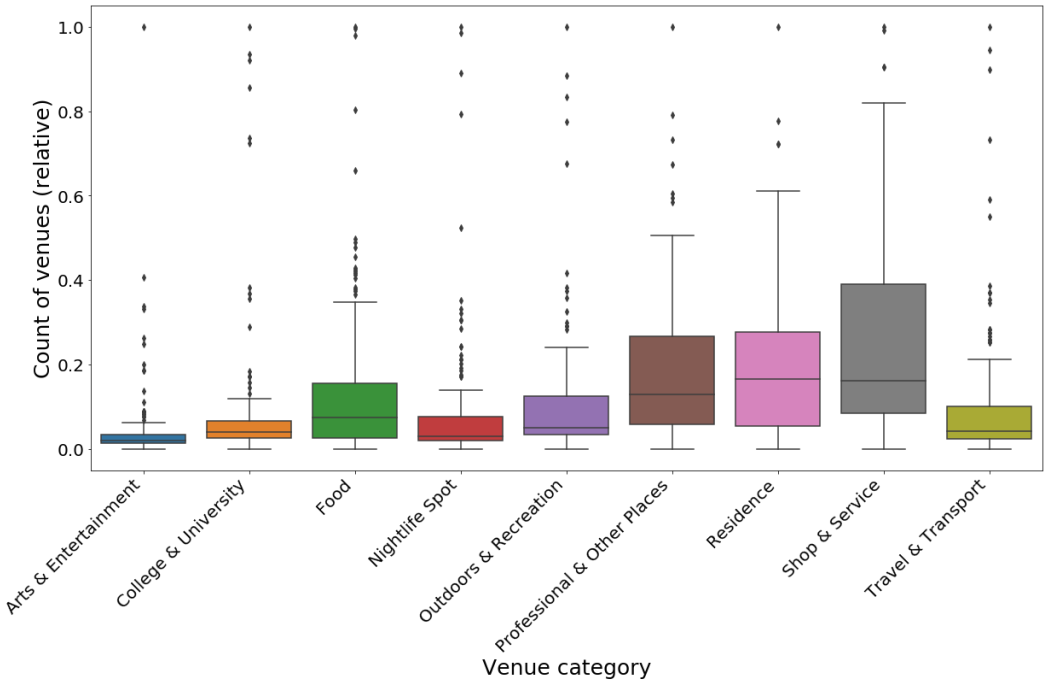


As the number of venues under ‘Events’ are almost zero, we shall discard the category from this point on. Observe that the number of venues under the categories of Food and Shop & Services are the highest under each DfT categories. Moreover, we decided to merge DfT categories A and B, Cand D, and lastly E and F together, reducing the number of unique DfT categories to 3 as their distributions seems similar.

Data Processing

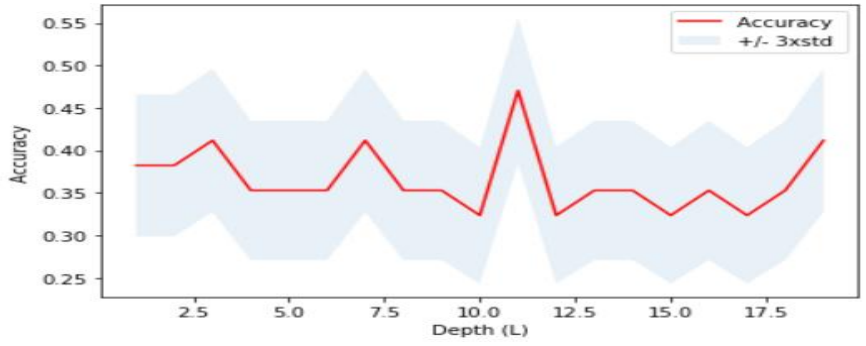
We use the Min-Max Scaling to normalise the number of venues on each venue categories, with 1 being the highest proportion in number of venues available under a specific venue category on each neighbourhood and 0 for otherwise. Here is the result on the overall data which is scaled accordingly.

	Station	Borough	Managed By	DfT Category	Coordinates	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Abbey Wood	Greenwich	TfL Rail	C&D	51.4915,0.1229	0.006897	0.013158	0.025532	0.025907	0.033333	0.029703	0.111111	0.057143	0.023622
1	Acton Central	Ealing	London Overground	C&D	51.5088,-0.2634	0.027586	0.052632	0.110638	0.062176	0.141667	0.297030	0.166667	0.295238	0.062992
2	Acton Main Line	Ealing	TfL Rail	E&F	51.5169,-0.2669	0.020690	0.026316	0.093617	0.020725	0.083333	0.217822	0.333333	0.171429	0.094488
3	Albany Park	Bexley	Southeastern	C&D	51.4358,0.1266	0.006897	0.026316	0.021277	0.025907	0.016667	0.059406	0.000000	0.161905	0.007874
4	Alexandra Palace	Haringey	Great Northern	C&D	51.5983,-0.1197	0.027586	0.052632	0.144681	0.072539	0.141667	0.267327	0.222222	0.428571	0.094488
5	Anerley	Bromley	London Overground	E&F	51.4125,-0.0651	0.082759	0.026316	0.157447	0.088083	0.100000	0.227723	0.166667	0.390476	0.070866

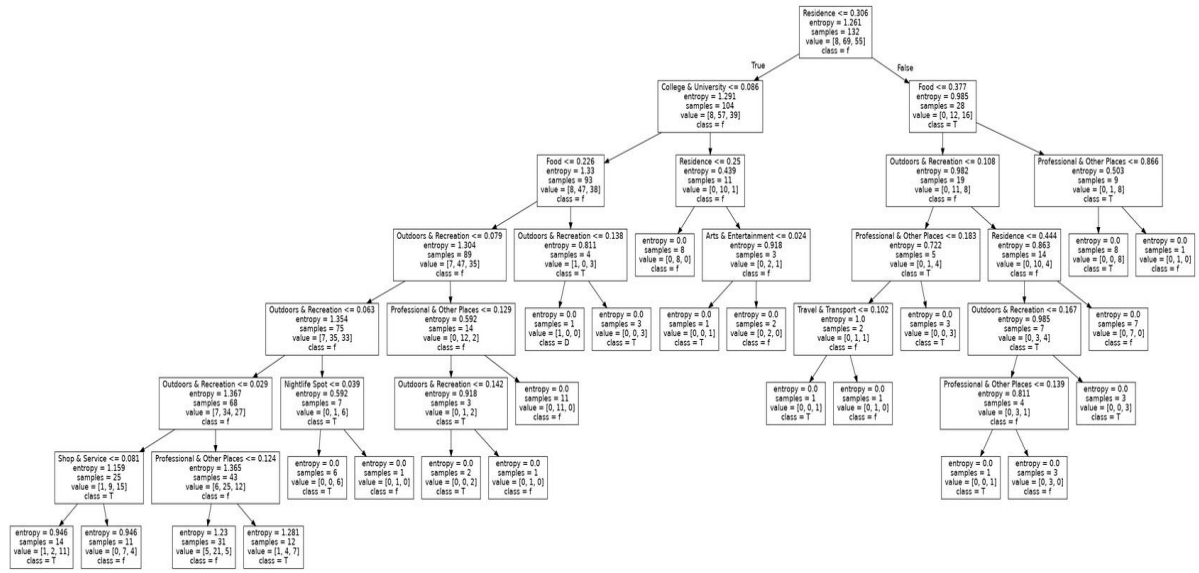


Modelling and Evaluation

We shall use Decision Tree to model the classification of DfT categories of each station based on the data on each venue category. In order to decide the ideal maximum depth of the decision tree, we consider the ones with the highest accuracy (which for this instance, we get the maximum depth of 11, see the graph below). But as the decision tree with depth of 11 is too complex and might overfit the data, we chose the one with 2nd highest accuracy (maximum depth of 7) instead.



Though the accuracy of this model is nowhere 50% which is shown above, by displaying the schematic of the our decision tree below, it is immediately obvious for us to pinpoint which criteria to distinguish the stations of DfT Categories of A&B and E&F only (ignoring the DfT Category C&D) by focusing on a few key venue categories. We shall elaborate it in the next section.



Results

Upon inspection of the decision tree we have generated above, key points have been taken on the following:-

1. For stations of DfT Categories A and B only:
    - They are not surrounded with higher density residential area (i.e. Residential Area density which is at most 31%).
    - They have less Outdoors and Recreational Areas surrounding it, with the density of at most 13.8%.
    - They are less likely to be surrounded by Professional and Other Places. This may indicate that there are higher possibility that the working class use these stations as interchange rather than exiting to work.
2. For stations of DfT Categories E and F only:
    - They are either surrounded by the higher density of residential areas, or
    - Almost surely not surrounded by Colleges or University as <9% of their neighbourhood are surrounded by Colleges and University (with >95% certainty).

Discussions

First and foremost, Foursquare data isn't all-encompassing. The highest number of venues are in the Food and Shop & Service categories and rather low on the residental Category, even in the high-density residental area of Camden, Edgware and so forth. Besides that, we have noticed that almost half (166 out of the total of 348) of the stations in Greater London do not have corresponding neighbourhood data in Foursquare, especially the residental areas and regions close to the London Green Belt.

Moreover, it is worth pointing out that the DfT Category itself is dependent to the Operator of the stations. If you could remember in the Exploratory Analysis and Data Cleaning, we can see that a station named Highbury & Islington carries two distinct DfT categories due to the fact that different operators are running on the same station as each operators have different statistics on frequency of trains and passenger count in their respective services. This is also hold true to stations which are very close to each other but operated by different Private companies. (Best example would be West Hampstead and West Hampstead Thameslink stations.) This may have adversely impacted the results obtained from the modelling and Evaluation process as they have different DfT Categories (D and E respectively) despite they both share the same neighbourhood. But as the projects aims to study the relationship between the composition of the neighbourhood and the DfT Category of each stations, Operators are not consider in this project.

Conclusion

Despite of the following shortcomings found in the Foursquare and the DfT Categories itself, this project has managed to provide a good insight on the neighbourhoods surrounding the stations which are either 'too busy' (DfT Categories A and B) and 'too quiet' (DfT Categories E and F). Moreover, with additional data combined (i.e. annual passenger count at each stations) or even treating the stations within a 1000m radius as a 'cluster' with combined passenger count, this project could yield a more accurate result.