TALES OF LONDON'S BUSIEST & QUIETEST STATIONS

Eu Meng Chong

What makes a Station Busy or Quiet? Was it due to its Neighourhood?

- Neighbourhood surrounding a railway station in London are not made equal:
  - Some are residential,
  - Some houses more corporate offices or financial hubs, or
  - Some could be a mixture of residential and working area (and the list goes on…)

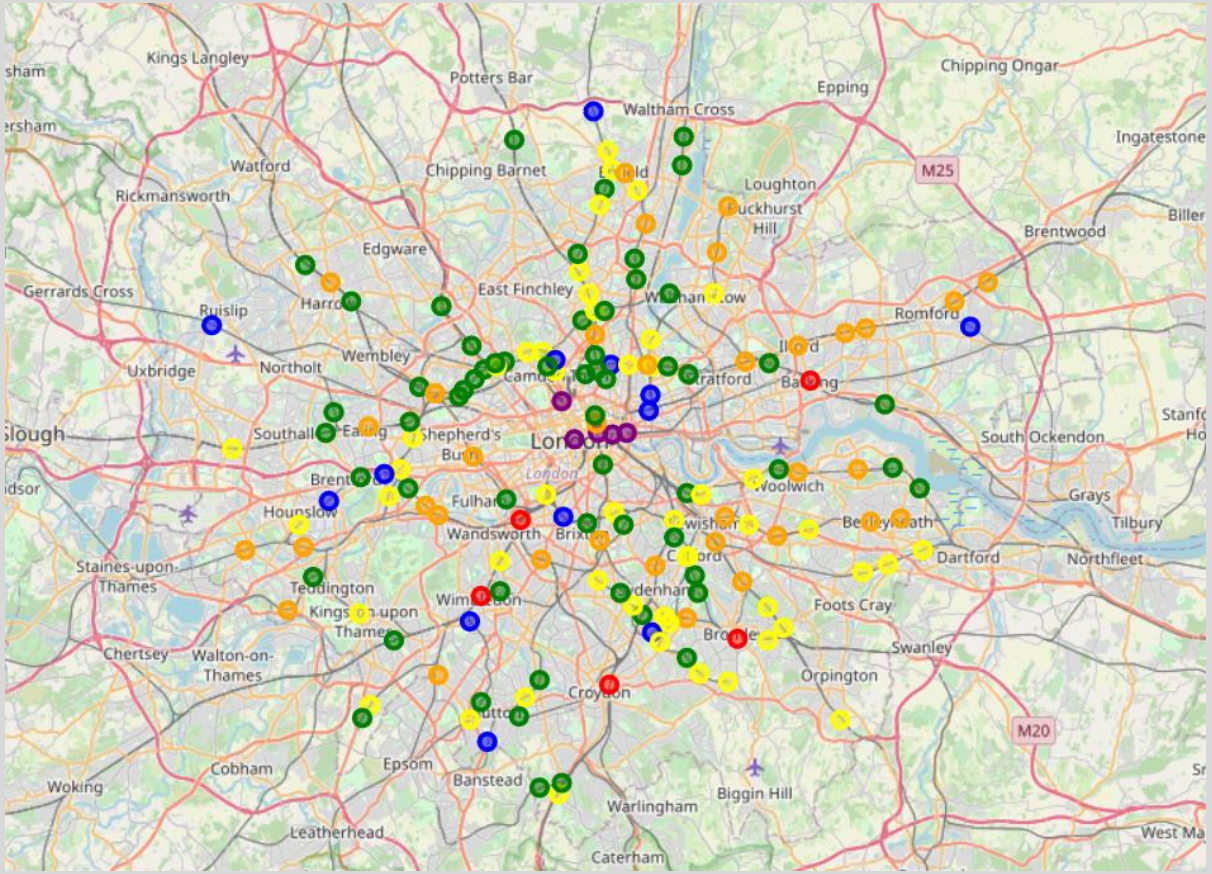
- Studying the Railway activities on each stations by understanding the Neighbourhood surrounding it provides potential benefits for the operators and the passengers.
  - For Operators:
    - To optimise passenger's satisfaction towards their service and the cost of operating the trains.
    - Providing opportunity to make railway extensions which are profitable.
  - For Passengers:
    - To optimise the frequency of trains at their local railway stations which provides a more efficient travel experience (notable example: skipping the stations which are more quiet to cut short the passenger's arrival time to their destinations.)

Data Acquisition and Cleaning

- We have taken 2 sources of data:-
  - https://en.wikipedia.org/wiki/List_of_London_railway_stations: This webpage contains a table of all the stations in London. We used 'BeautifulSoup' library to extract the table.
  - Foursquare API to query the number of venues within the 1000m radius from each station based on the 10 Venue Categories set by Foursquare.

- Data Cleaning
  - Not all neighbourhood surrounding each station have Foursquare data on its venues:
    - We drop the stations which have no corresponding neighbourhood data in Foursquare. This leaves us with 182 out 348 rows of data to be considered.
  - Ambiguous DfT Category on a station.
    - We fix the DfT Category based on the station's operator.

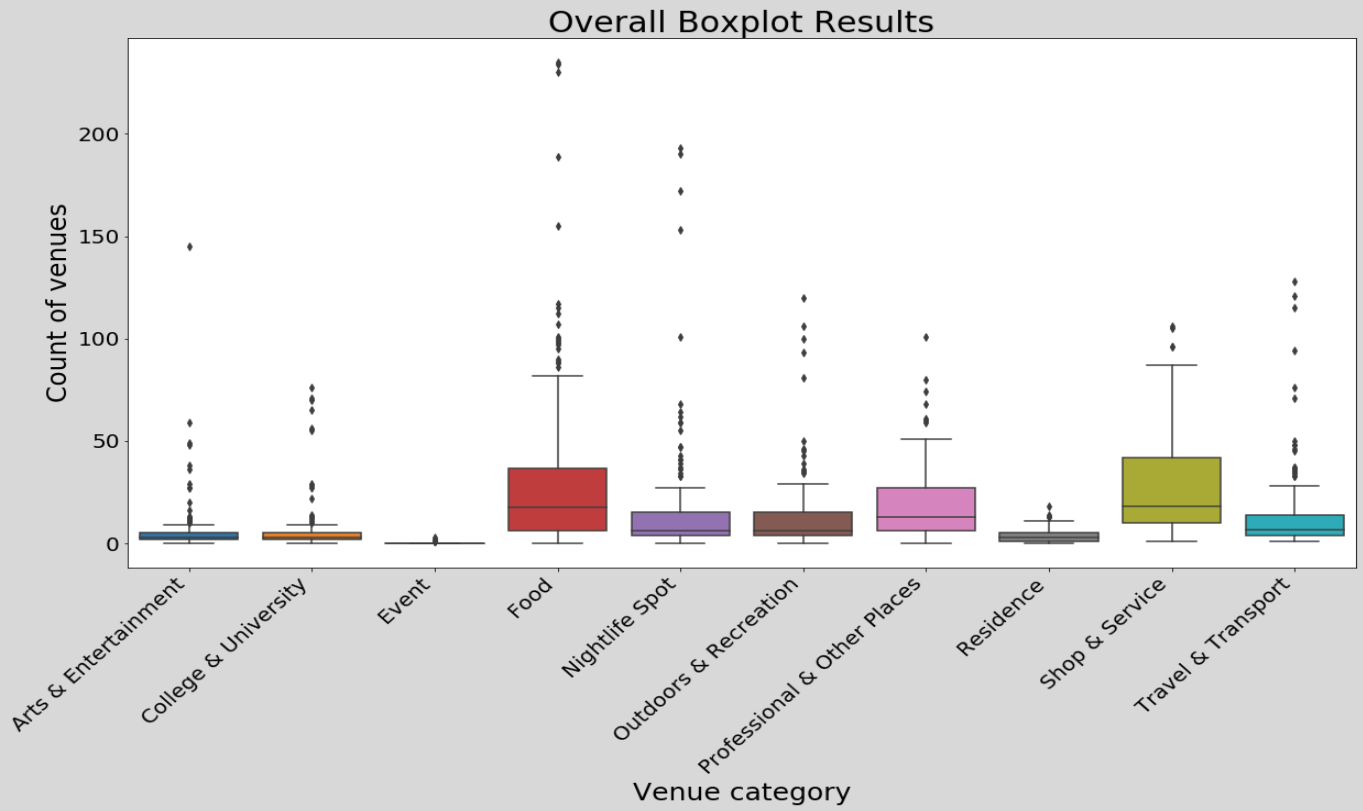
They are Scattered Everywhere!



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

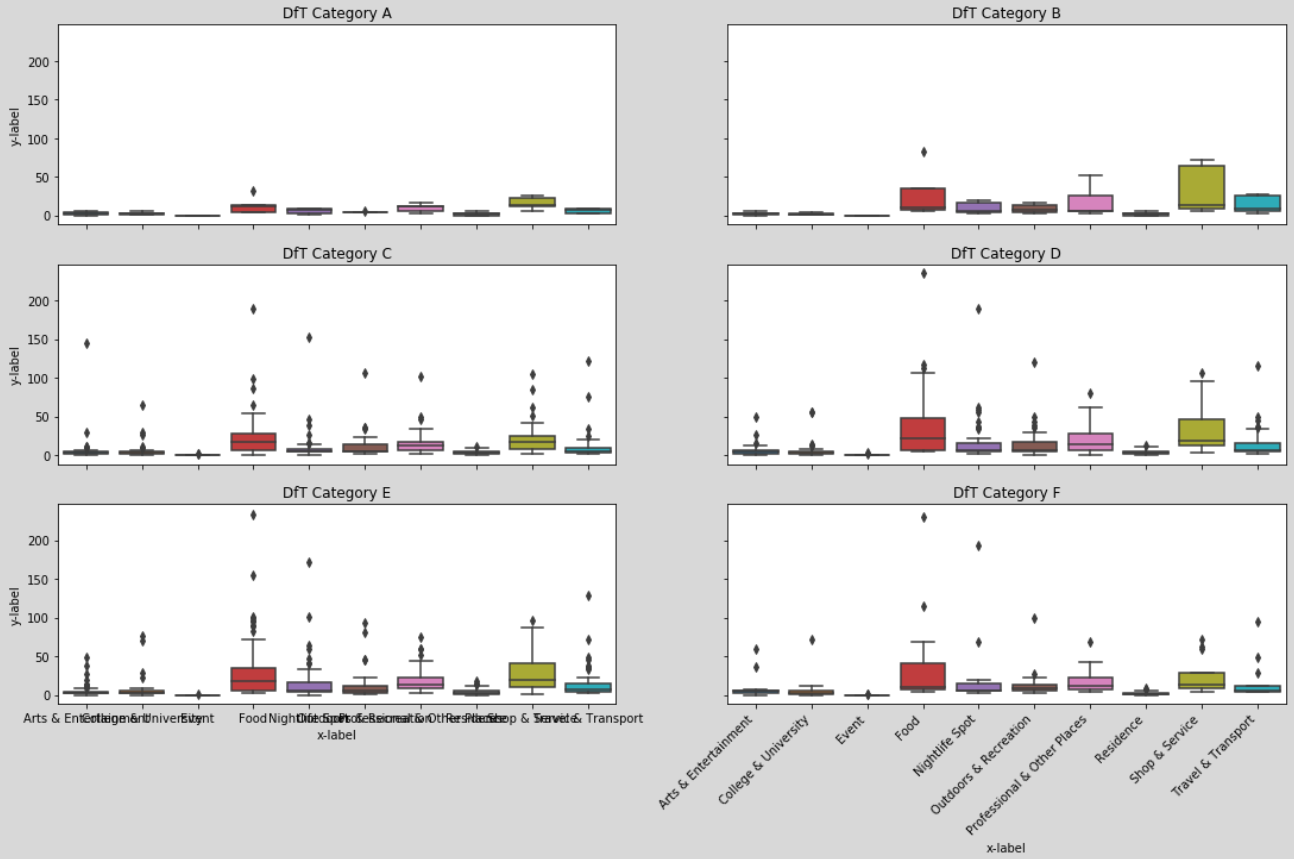
- Stations of DfT Categories B, D, E & F are scattered across central London and Outer London.
- The only exception: DfT Category A are all in central London.
- Here, we decided not to use the station's location as candidate attribute for modelling.

Distribution of Venues - Overall



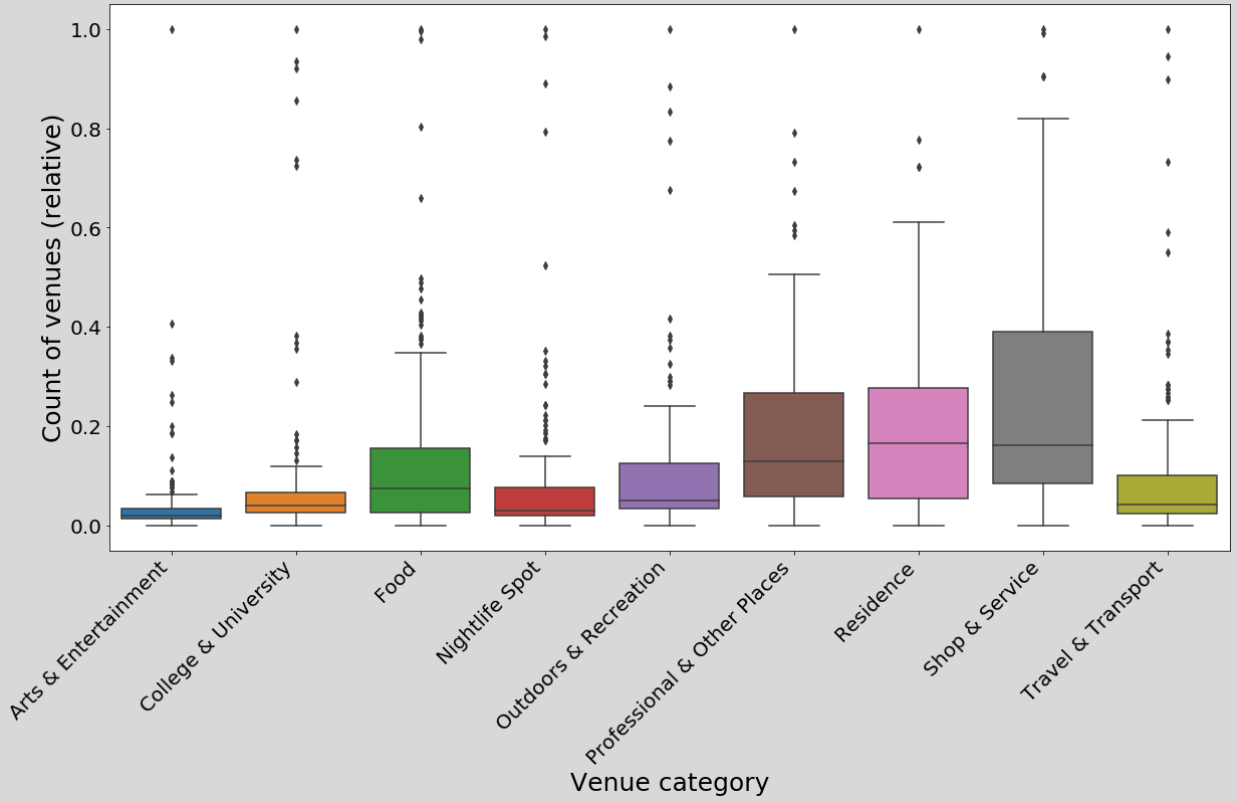
- The top two venue categories are:
  - Food
  - Shop & Service
- The venue category 'Event' has the least amount of venues.
- Apart from the outliers, the boxplot for 'Nightlife Spot' and 'Outdoors & Recreation' are similar.

Distribution of Venues – By DfT Category



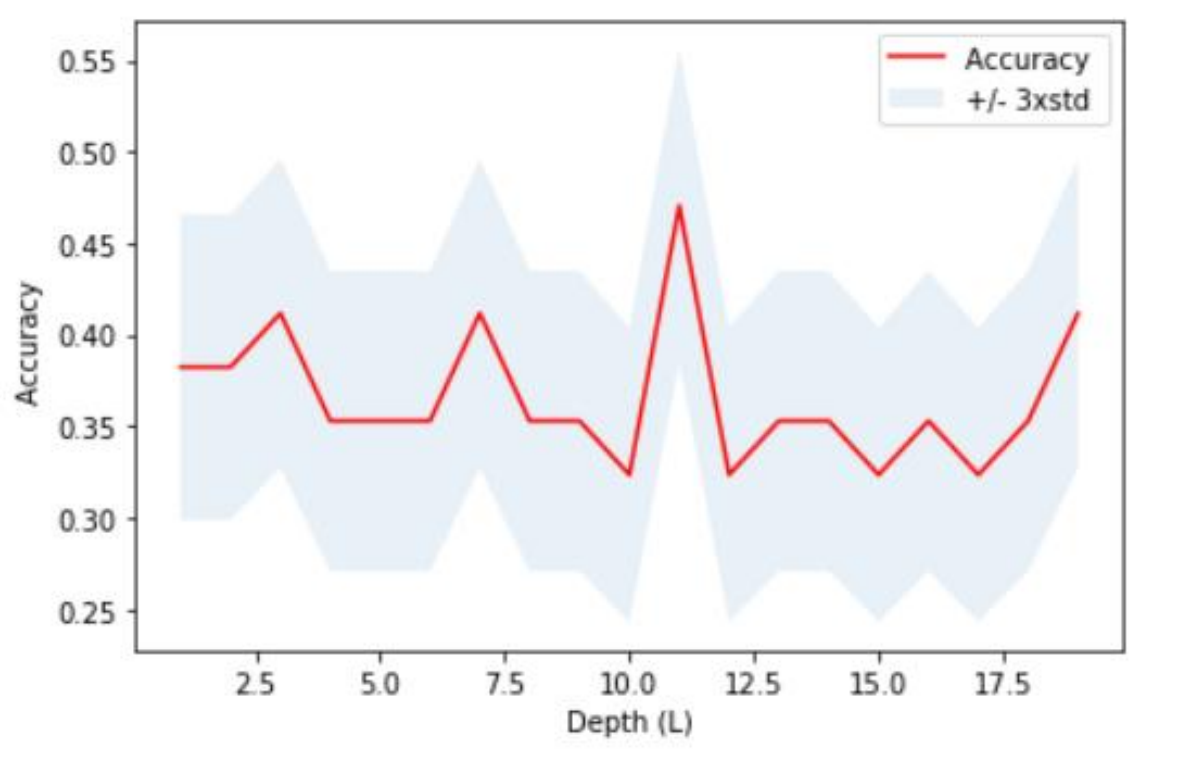
- As observed, the distribution on DfT Categories A & B are the least erratic.
- The 'Event' category is almost to none across all DfT categories. Hence it is safe for us to remove this category from the dataset.
- Due to the similarity of the distribution of DfT categories A and B, C and D, and lastly E and F, we shall merge these pairs together, forming 3 distinct DfT categories.

Before we Start…



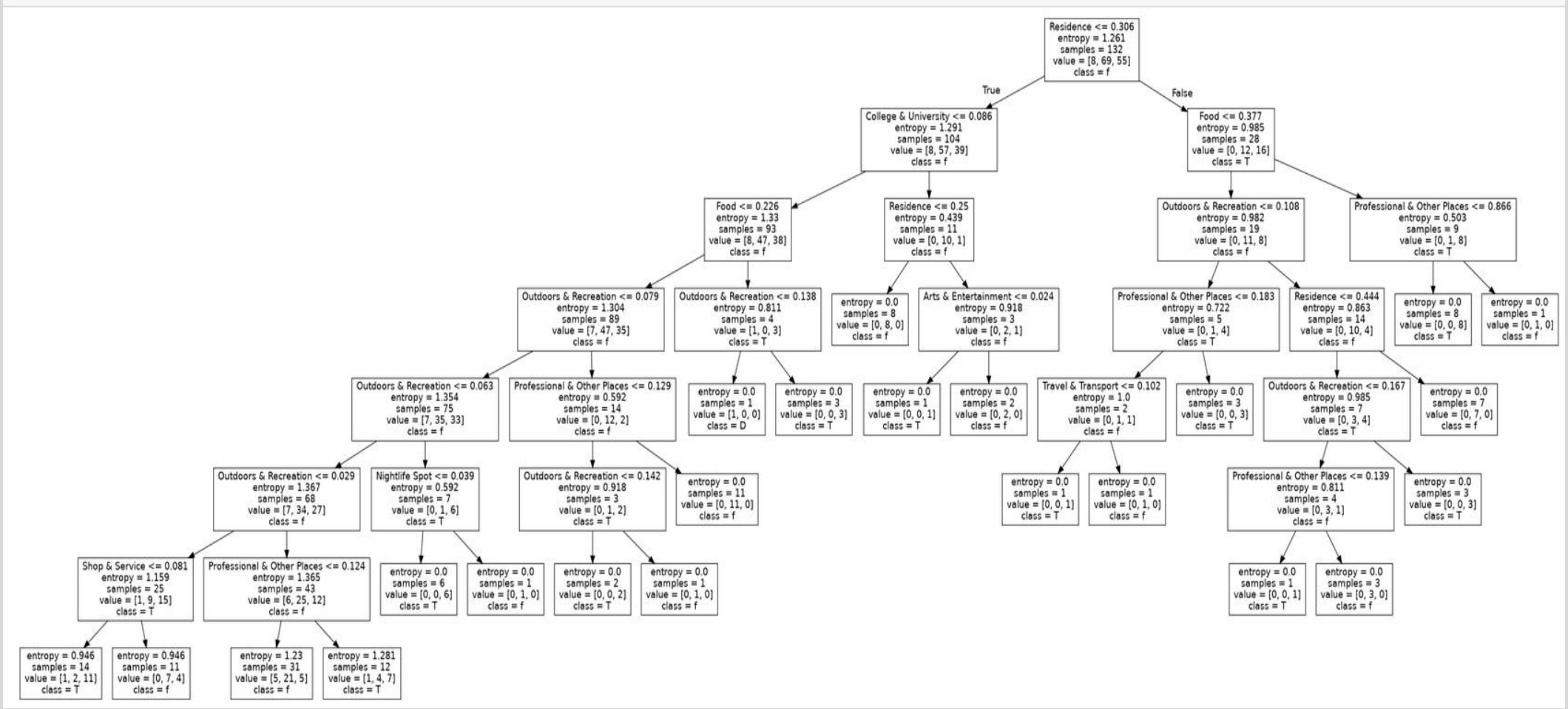
- We normalise the dataset by using the Min-Max Scaling
  - Where it is scaled between 0 and 1 instead of the total venues surrounding the railway stations.

Modelling – Decision Tree



- We shall use accuracy to determine the maximum depth which is optimal for our dataset.
- The model with maximum depth of 11 is the most accurate but too complex. We used the 2nd most accurate model (maximal depth of 7) instead.
  - To prevent overfitting
- Though this is not accurate, we still proceed to visualise the Decision Tree and gain any insights from the model.

Decision Tree Schematics



Insights: Extremities of DfT Categories can be Determined

DfT Categories A and B (Busiest ones) only:

- Density of Residential Area is less than 31%
- Density of Outdoors and Recreational Areas is less than 13.8%.

DfT Categories E and F (Quietest ones) only:

- Surrounded by the higher density of residential areas, or
- Very less likely to be surrounded by Colleges and Universities (< 13%)

Conclusion and Future Directions

- Despite the inaccuracy of the model, it has though successfully distinguish the 'busiest' (DfT Categories A&B) and the 'quietest' (DfT Categories E&F) stations in London with high accuracy and certainty.
- We may need to consider additional data to improve our model:
  - Annual passenger count on each station, reducing our dependence on DfT categories which may be misleading.
  - Using a better alternative to Foursquare API which are able collect venue data on the places in London where Foursquare API could not obtain.
  - Considering all the stations within a 1000m radius as a cluster and aggregate the data accordingly.
- We need to improve the accuracy of the model:
  - Reclassifying the modified DfT categories.
  - Using different Scaling methods to normalise the data.