1. Data acquisition and cleaning

1.1 Data sources

The data set with London's postcodes, Counties, Districts, Wards, Geospatial information, train stations and how close they are to the postcodes can be found at the following url https://www.doogal.co.uk/UKPostcodesCSV.ashx?area=London. To get venues in the wards, I used foursquare data.

1.2 Data cleaning

Data retrieved from the csv file was read into jupyter notebooks. For the most part the data set was clean except some postcodes were no longer in use, however since the purpose of this research is to use the codes as a unique identifier, this development did not affect the dataset so all postcodes were retained.

1.3 Feature selection

After cleaning the data, there were 320426 samples and 44 features. Eight of the 44 features required namely Postcode, Latitude, Longitude, County, District, Ward, Nearest station and Distance to nearest station were selected.

On selecting the rows in the 'E1 xxx' Postcode of 'City of London' district, I realized a redundancy between the County feature and District feature so the County feature was dropped since the one required county was already acquired. After all that, 82 samples and 7 features remained

From the foursquare data, venue name, venue categories, venue latitude and longitude were retrieved and used in analysis.