**Peer-graded Assignment:**

**Capstone Project - The Battle of Neighbourhoods**

**Final Report**

**Tony Holdroyd**

* **Introduction where we discuss the business problem and who would be interested in this project.**

A large digital services company, MultiCorp Ltd, are moving to the Dartford area of Kent in the UK

from Scotland, also in the UK. The company is relocating 100 members of its younger members of staff, a large percentage of whom have primary school age children. Each member of staff has been given a relocation allowance such that they will be able to sell their properties in Scotland and buy a house in the Dartford area.

Dartford is the principal town in the Borough of Dartford, Kent, England. It is located 18 miles (29 kilometres) south-east of Central London, and is situated adjacent to the London Borough of Bexley to its west. To its north, across the Thames estuary, is Thurrock in Essex, which can be reached via the Dartford Crossing.(ref: Wikipedia).

The members of staff who are relocating wish to be within the catchment area (allocated by post code) of a primary school, but also wish to have local facilities to enjoy when they are not at work.

So the problem for them is, how to choose a school with the best (most diverse and largest in number) local facilities, but also to choose areas with the least venues for those staff preferring a quiet neighbourhood. The problem is also that different staff will require different types of venues (for example: pubs/bars/restaurants or perhaps local shopping facilities or train stations)

The human resources department of the company have been tasked to address this problem and find a solution for the staff and they have hired you on a consultancy basis to provide the solution.

You decide to create loop-up tables for the staff so that they can visit and assess the schools but also decide if the neighbourhood is suitably well provided with the local facilities they need for their leisure time. You also decide to provide each member of staff with an interactive map of Dartford in the form of an HTML document which allows them to click on hotspots, one per school, which will bring up tooltips showing the school post code and address for navigational purposes.

* **Data (where we describe the data that will be used to solve the problem and the sources of that data.)**

Because the school names are in a stand-alone table provided by Dartford council on their website, the names can be exacted from there. The postcodes, addresses and phone numbers exist in separate html documents, one per school. Because of this it was decided to manually extract these data and add them to the names table with code. The required latitude and longitude, based on the post codes of the schools, were available as one CSV file for the whole of England. The latitude and longitude could therefore be added to the names table by merging on the imported post codes file and using an inner join. An intern was employed to do the manual extracting and recording.

This analysis produced, for example:

| **School Name** | **Post Code** | **Address** | **Phone** | **latitude** | **longitude** |
| --- | --- | --- | --- | --- | --- |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | Bean Primary School | DA2 8AL | School Ln, Bean, Dartford | 01474 833225 | 51.426347 | 0.289653 |

**Summary Data**

**The data required will be:**

The names of all the primary schools in the Dartford area (Neighbourhood)

e.g. The Anthony Roper Primary School

The addresses of the primary schools

e.g High St, Eynsford, Dartford

The post code of each of the primary schools (Neighbourhood latitude and longitude)

e.g. DA4 0AA

The latitude and longitude of each of the primary schools

e.g. 51.369382 , 0.213492

The facilities (venues) local to the primary schools:

The venue names.

e.g. The Malt Shovel

The venue categories

e.g. Pub

The venue location

e.g

The local facilities will be ascertained by use of the Foursquare API

The staff will be able to use the phone number to book visits to the school, then use the post code and address with their satnavs to visit and inspect the schools.

Staff will then examine the cluster analysis that will be performed on the facilities to determine their preferred neighbourhood.

The preferred neighbourhood might be one that contains different type of facilities. It will depend on the individual staff family, some may prefer a quiet neighbourhood, some a busy one. Some may want pubs and restaurants, some may want gym facilities and so on.

* **Methodology section where we discuss and describe any exploratory data analysis that we did, any inferential statistical testing that we performed, if any, and what machine learnings were used and why.**

First of all, an HTML page was found which listed all the schools in a table. This was manually saved to an HTML file which was then read into a Pandas dataframe with the pandas.read\_html() library utility. ([3])., called schools

The School Name had some strange characters appended to it, so the appropriate column was renamed. (Since at this stage all that was needed was the School Name, the Town and School Type columns were then dropped.) Also now the schools of type ‘Primary School’ were extracted ([4], [5])

As post codes, addresses and phone numbers for each school were needed as part of the solution to the Problem these empty columns were then added to the schools table. ([6]). It was discovered that these data were spread across a number of HTML pages, one per school, meaning that automated web scraping would take a long time to implement, and the decision was made to manually copy these data from the pages into the notebook, from where they were written into the schools dataframe . An intern was tasked with this rather tedious chore. ([7], [8])

Next a CSV file which contained all the postcodes in the UK together with their latitudes and longitudes was sourced, from the Internet of course, and this was read directly into a Pandas dataframe ([9]). The postcode column was renamed to Post Code, to match the schools dataframe. This was subsequently left inner-joined to the School dataframe using the Post Codes , this join uses the Post Codes from the School on the left and picks out the matching Post Codes from the dataframe on the right and thus the latitudes and longitudes for each school were added to the schools dataframe. ([10], [11], [12]).

The geopy API was next used to find the latitude and the longitude for use by the mapping software [(13)], and then the Foursquare API, together with the schools data used to generate an interactive HTML stand-alone map for use by the staff members when visiting the schools ([14]).

To prepare for use of the Foursquare API, the appropriate credentials were next hard-coded followed by the definition of a function that returns a dataframe of venues that are nearby to each of the names, passed in as names together with their latitudes and longitudes. A call to this function then with the appropriate arguments returned the venues nearby to each of the schools ([15], [16],[17], [18])).

So the nearby venues to each school were now available for further processing to present to the staff.

We were interested to know how many unique categories of events there were so this figure was determined next ([19]).

The total number of venues (venue count)for each school were then established and joined on to the schools dataframe([20]-[22])

Next the schools with the most and least nearby venues was established.([23], [24]).

It was then decided to analyse school neighbourhoods to see which of each of the venues were nearby, producing a wide and deep dataframe ([25]-[28]). Also the mean frequency of occurrence of each category of venue, per school, was determined.([29]-[30]).

Next the 10 most common venues, per school were determined as this directly addresses the business problem were were tasked with ([33],[34]).

The schools were then clustered into 4 clusters, by venue. These clusters were mapped and output in the form of dataframes per cluster ([35]-[47]).

This was the end of the analysis.

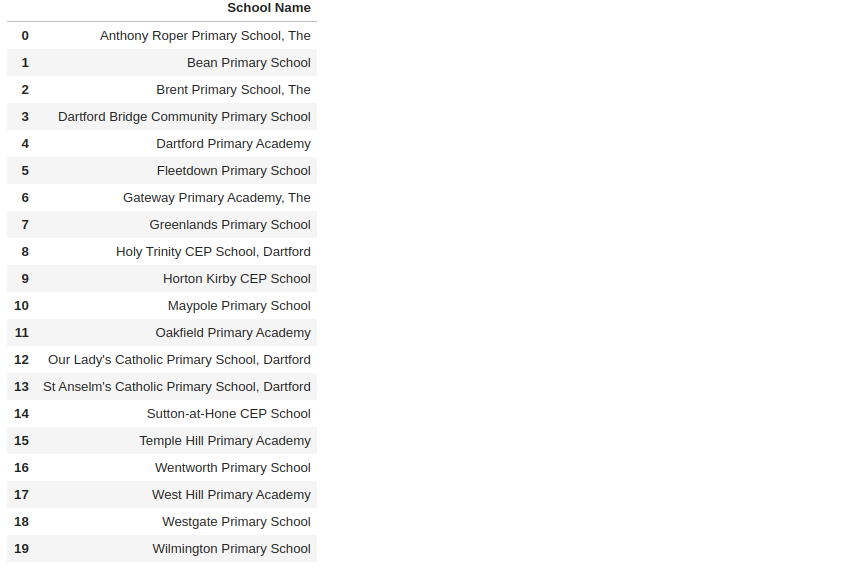
**Note**:

For checking purposes, the head of dataframes were printed at appropriate junctures, as can be seen in the Notebook (or occasionally the complete dataframe)

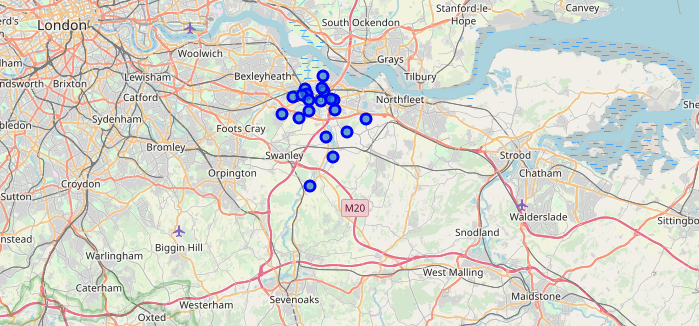
* **Results section where we discuss the results.**

There were see to be a total of 32 schools in the Dartford local council area of which 20 are Primary

These are:

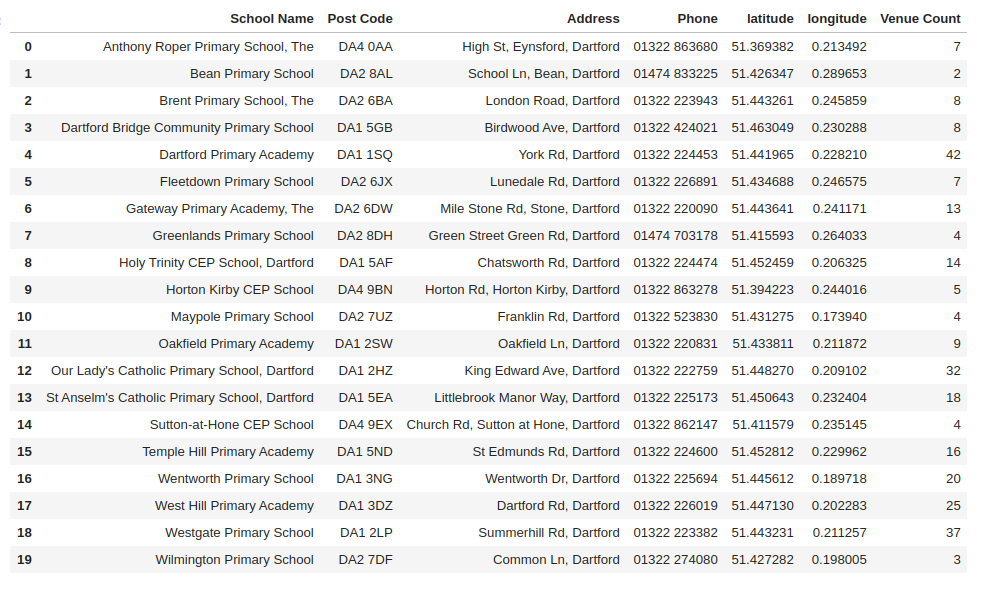
The latitude and longitude of Dartford were as follows:

And the (interactive HTML) map looks like:



The total number of unique facilities turned out to be 55; the area with the most facilities was Dartford Primary Academy, York Road, Dartford, DA1 1SQ (51.442, 0.264033) with 42 local facilities (venues); that with the least was Bean Primary School, School Lane, Bean Dartford, DA2 8AL (51.426347, 0.289653) with just 2 local facilities.

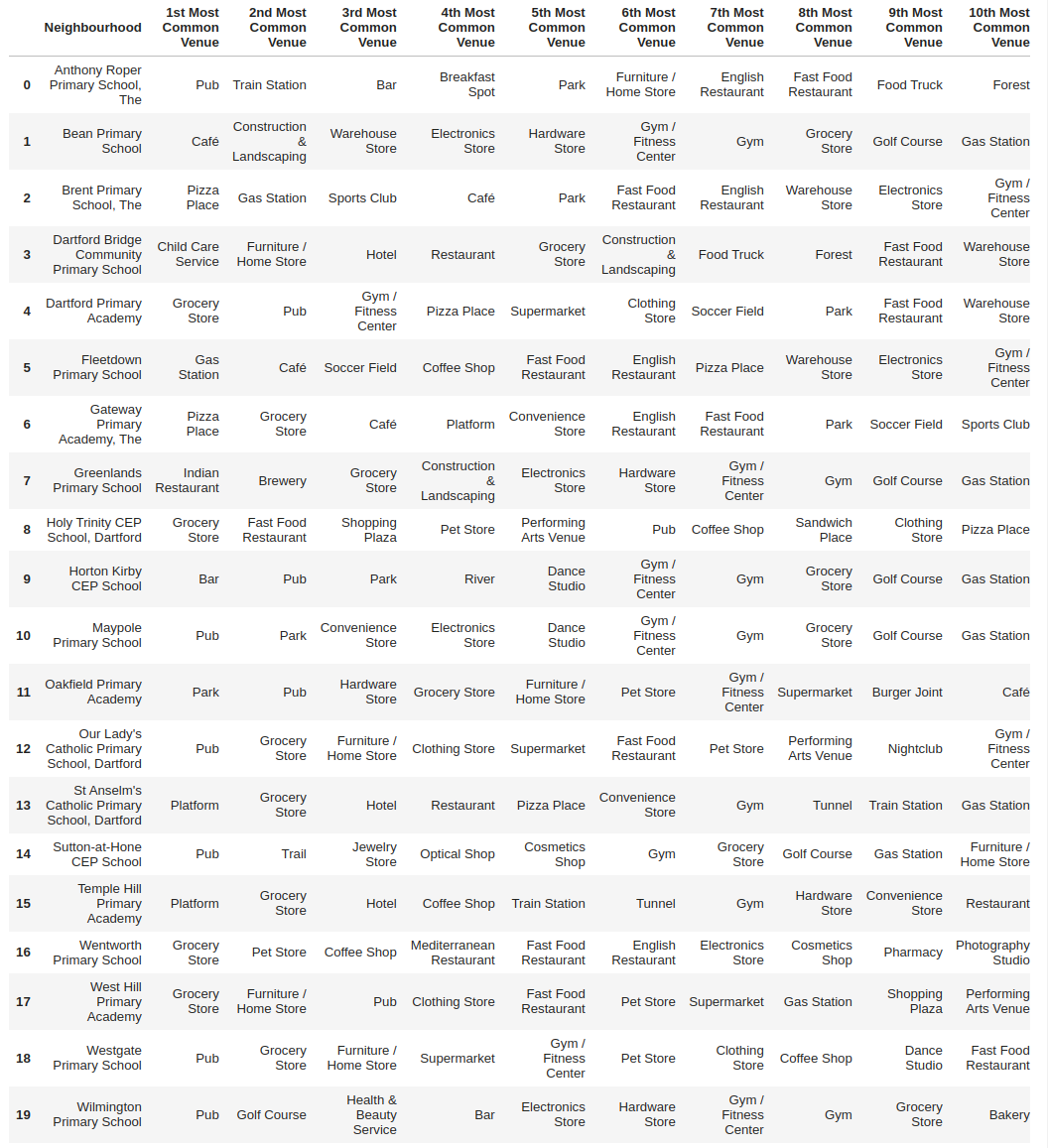
The venue counts are contained in the following (which dataframe also contains the full school details):



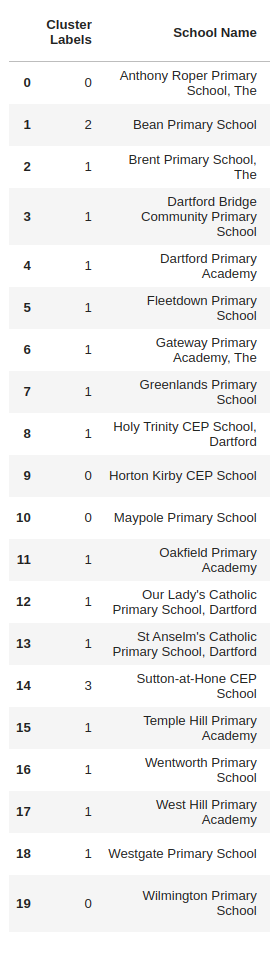
A sample of the schools dataframe with added venue details (Venue, latitude, longitude and Category is:

for ‘Neighbourhood, read ‘school’)

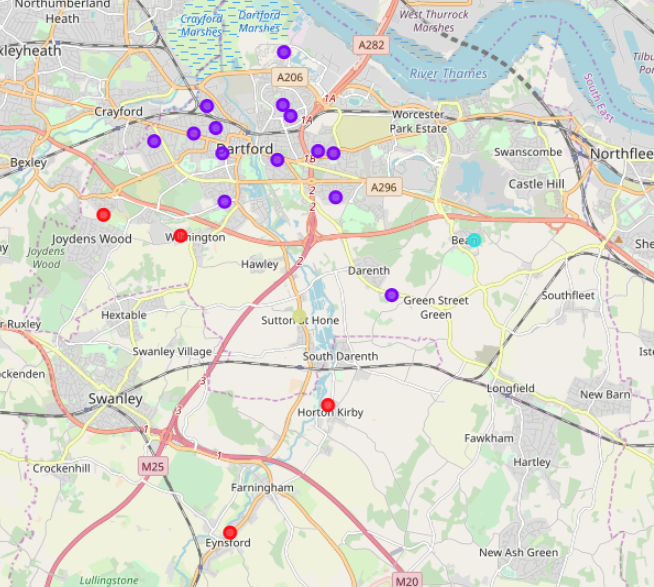
The dataframe containing the 10 most common venues, per school (neighbourhood):



The clusters are as follows:



And the cluster map is:



**( red:cluster 0, purple: cluster 1, blue: cluster 2 and yellow cluster 3)**

**The clusters were as follows:**

**Cluster 1: Good mix of venues:**



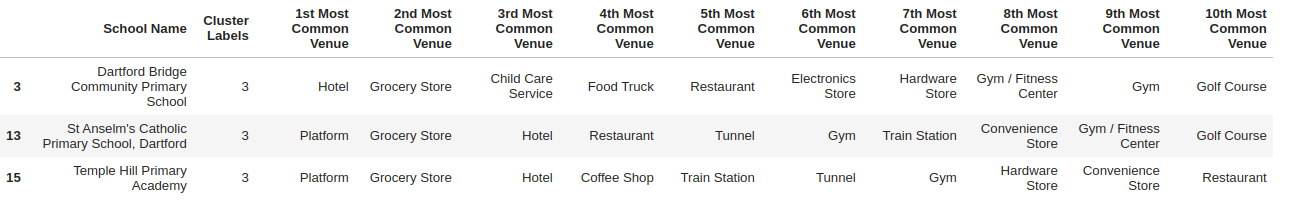
**Cluster 2: Mostly groceries and pubs/restaurants.**



**Cluster 3: Predominantly pubs and bars:**



**Cluster 4: Near stations , groceries and hotels:**



* **Discussion section where we discuss any observations we noted and any recommendations we can make based on the results**.

We are neutral about the best school for a particular member of staff since much will depend on that member of staff’s requirements of a school, its immediate neighbourhood and facilities. However, examination of the clusters gives a rough idea of the pool of schools from which schools may be chosen for further investigation by the member of staff in question. These are in the results section and also in my presentation to staff so I won’t copy them here as well.

Note that the longitudes are all quite close to 0, as was to be expected as Dartford is near to Greenwich in the UK through which 0 degrees of longitude (the Greenwich Meridian) passes.

* **Conclusion section where we conclude the report.**

This report has described and discussed the problems faced by staff relocating to an area they do not know and how they may choose a primary school based on the local facilities. The presentation was made to staff and they proceeded to investigate local primary schools for themselves by making use of the digital tools we have provided, as described in the presentation. Should the staff wish to further investigate individual schools, there are government (OFSTED) reports available at

* <https://www.compare-school-performance.service.gov.uk/schools-by-type?step=default&table=schools&parliamentary=Dartford&geographic=parliamentary&for=ofsted>

2. A link to our Notebook on our GitHub repository pushed showing our code.

***As GitHub was refusing to render my notebook ( a known problem with GitHub) I used nbviewer to render it and have also supplied this link as a fallback. It is still hosted on GitHub.***

**3. Presentation**

Please see link

Thank you for reading my report.