Project:

**Segmenting and Clustering Neighborhoods in UK**

**O**

ne of the issues when parents choose whether to settle in a district or area to live is whether a location has good schools or not and what are the surrounding neighbourhoods. Are the surroundings suitable for children? Take for an example, if a family settle in Bristol, most likely this family has made the decision to retire in this near-seashore city. In Bristol, when looking up its neighbourhood venues in Foursquare, many of them are bars, which is very much different from that of Oxford. Therefore, finding out schools and its surrounding could help parents making sensible choices when choosing a place to settle or for their children to study.

**Source of data**

The data came from the open data initiatives of UK government.

Schools data were downloaded from <https://ea-edubase-api-prod.azurewebsites.net/edubase/downloads/public/edubasealldata20210201.csv> , provided by Gov.uk. These data contained information about a school: its location, its establishment group, its religion, its Ofsted Rating, etc. There is no latitude and longitude data.

To get the latitude, longitude and nearby postcodes of a postcode, this API is used : <https://api.postcodes.io/> .

Venues data are from Foursquare.

**Methodology**

First, a complete list of postcodes with NHS info and a complete list of education establishment were downloaded from gov.uk. The shape of the data frame created is (1711275, 10), i.e. around 1.7 mil rows for all postcodes. The example looks like this. Because there is no latitude and longitude data, we need to figure out how to get those location data for each postcode.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Postcode** | **Positional\_quality\_indicator** | **Eastings** | **Northings** | **Country\_code** | **NHS\_regional\_HA\_code** | **NHS\_HA\_code** | **Admin\_county\_code** | **Admin\_district\_code** | **Admin\_ward\_code** |
| **0** | CR0 0AA | 10 | 539052 | 162028 | E92000001 | E19000003 | E18000007 | NaN | E09000008 | E05011471 |
| **1** | CR0 0AB | 10 | 539266 | 161728 | E92000001 | E19000003 | E18000007 | NaN | E09000008 | E05011471 |
| **2** | CR0 0AD | 10 | 539327 | 161477 | E92000001 | E19000003 | E18000007 | NaN | E09000008 | E05011471 |
| **3** | CR0 0AE | 10 | 539086 | 162071 | E92000001 | E19000003 | E18000007 | NaN | E09000008 | E05011471 |
| **4** | CR0 0AF | 10 | 539322 | 161810 | E92000001 | E19000003 | E18000007 | NaN | E09000008 | E05011471 |

The information of schools were downloaded from [gov.uk](http://gov.uk) and have the following format. The shape of the data frame is (48858, 139), which is also big. Partial data was shown here:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Postcode** | **URN** | **LA (name)** | **EstablishmentName** | **TypeOfEstablishment (name)** | **EstablishmentTypeGroup (name)** | **PhaseOfEducation (name)** | **Gender (name)** | **ReligiousCharacter (name)** | **OfstedRating (name)** | **Easting** | **Northing** |
| **0** | EC3A 5DE | 100000 | City of London | The Aldgate School | Voluntary aided school | Local authority maintained schools | Primary | Mixed | Church of England | Outstanding | 533498.0 | 181201.0 |
| **1** | EC2Y 8BB | 100001 | City of London | City of London School for Girls | Other independent school | Independent schools | Not applicable | Girls | None | NaN | 532301.0 | 181746.0 |
| **2** | EC4M 9AD | 100002 | City of London | St Paul's Cathedral School | Other independent school | Independent schools | Not applicable | Mixed | Church of England | NaN | 532160.0 | 181151.0 |
| **3** | EC4V 3AL | 100003 | City of London | City of London School | Other independent school | Independent schools | Not applicable | Boys | None | NaN | 531981.0 | 180844.0 |
| **4** | NW5 4LP | 100004 | Camden | Sherborne Nursery School | Local authority nursery school | Local authority maintained schools | Nursery | Mixed | Does not apply | NaN | 528515.0 | 184869.0 |

Initial exploration was done to understand what data were available for investigation. For example, we can study the amount of schools with different Ofsted Rating in a particular Town.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **County (name)** | **Postcode** | **EstablishmentName** |  |
| **Town** | **EstablishmentTypeGroup (name)** | **OfstedRating (name)** |  |  |  |
| **Oxford** | **Academies** | **Good** | 10 | 11 | 11 |
| **Outstanding** | 1 | 1 | 1 |
| **Requires improvement** | 3 | 3 | 3 |
| **Special Measures** | 4 | 4 | 4 |
| **Colleges** | **Good** | 1 | 1 | 1 |
| **Free Schools** | **Good** | 1 | 1 | 1 |
| **Independent schools** | **Good** | 3 | 3 | 3 |
| **Local authority maintained schools** | **Good** | 21 | 21 | 21 |
| **Outstanding** | 4 | 4 | 4 |
| **Requires improvement** | 1 | 1 | 1 |
| **Serious Weaknesses** | 2 | 2 | 2 |
| **Special Measures** | 2 | 2 | 2 |
| **Special schools** | **Good** | 1 | 2 | 2 |
| **Outstanding** | 2 | 2 | 2 |
| **Special Measures** | 1 | 1 | 1 |
| **Universities** | **Good** | 1 | 1 | 1 |

This table showed the top 20 towns with most schools.

|  |  |  |
| --- | --- | --- |
|  | **EstablishmentName** |  |
| **Town** | **Postcode** |  |
| **London** | **E14 9TS** | 16 |
| **Manchester** | **M29 8BS** | 14 |
| **Tyldesley** | **M29 8BS** | 11 |
| **Newcastle-upon-Tyne** | **NE4 8XJ** | 8 |
| **Peterborough** | **PE4 6HX** | 8 |
| **London** | **N16 5RP** | 8 |
| **Blackpool** | **FY3 9JL** | 8 |
| **Middlesbrough** | **TS3 8RD** | 7 |
| **Skelmersdale** | **WN8 8EH** | 7 |
| **Ipswich** | **IP3 0EW** | 7 |
| **South Croydon** | **CR2 8HD** | 7 |
| **Cambridge** | **CB4 2BD** | 7 |
| **Nottingham** | **NG8 5PN** | 7 |
| **Portsmouth** | **PO6 3JL** | 7 |
| **Milton Keynes** | **MK2 2HB** | 7 |
| **Chatham** | **ME5 0LB** | 7 |
| **Coventry** | **CV6 3BL** | 6 |
| **Newcastle-upon-Tyne** | **NE5 2LW** | 6 |
| **Thornton Heath** | **CR7 8BT** | 6 |
| **Coventry** | **CV2 1EQ** | 6 |

Since there is no latitude and longitude info in both data set, the data was obtained from an online API for each postcode. The returned data set is in JSON format that requires parsing to extract the necessary data. Query and results are stored in dataframe for processing.

query result

0 CR0 0AH {'postcode': 'CR0 0AH', 'quality': 1, 'easting...

1 CR0 0PA {'postcode': 'CR0 0PA', 'quality': 1, 'easting...

2 CR0 0PH {'postcode': 'CR0 0PH', 'quality': 1, 'easting...

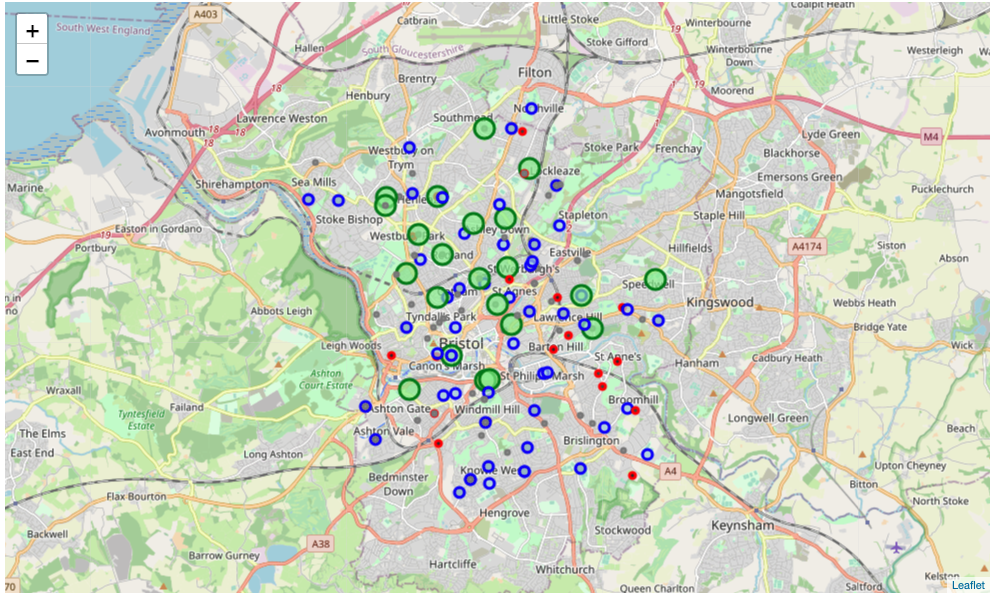
3 CR0 0EG {'postcode': 'CR0 0EG', 'quality': 1, 'easting...

4 CR0 1ND {'postcode': 'CR0 1ND', 'quality': 1, 'easting...

With all necessary info, a dataframe is created that contains school info, latitude, longitude and other relevant info.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **EstablishmentTypeGroup (name)** | **EstablishmentName** | **OfstedRating (name)** | **County (name)** | **Town** | **Postcode** | **country** | **postcode** | **admin\_district** | **latitude** | **longitude** |
| **0** | Local authority maintained schools | Fairchildes Primary School | Outstanding | Surrey | Croydon | CR0 0AH | England | CR0 0AH | Croydon | 51.334550 | -0.002455 |
| **1** | Local authority maintained schools | Addington High School | NaN | Surrey | Croydon | CR0 0AH | England | CR0 0AH | Croydon | 51.334550 | -0.002455 |
| **2** | Academies | Meridian High School | Requires improvement | Surrey | Croydon | CR0 0AH | England | CR0 0AH | Croydon | 51.334550 | -0.002455 |
| **3** | Academies | Fairchildes Primary School | NaN | Surrey | Croydon | CR0 0AH | England | CR0 0AH | Croydon | 51.334550 | -0.002455 |
| **4** | Academies | Meridian High School | Requires improvement | Surrey | Croydon | CR0 0AH | England | CR0 0AH | Croydon | 51.334550 | -0.002455 |

Next we drill into one of the cities to explore the rating and number of schools inside the city and create a visual map for further exploration. Schools are colored according to their Ofsted Rating

* Good: Blue
* Outstanding: Green
* Requires improvement: Red
* Others: Gray

The neighbourhoods were identified using an online API. As the results returned were in postcodes, they were converted to latitude and longitude before exploring further with Foursquare.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **name** | **categories** | **lat** | **lng** |
| **0** | The Old Market Assembly | Bar | 51.456784 | -2.579021 |
| **1** | The Volunteer Tavern | Pub | 51.457749 | -2.582045 |
| **2** | The Trinity Centre | Music Venue | 51.458080 | -2.576293 |
| **3** | 25A Old Market | Café | 51.455865 | -2.582173 |
| **4** | Cabot Circus | Shopping Mall | 51.458507 | -2.585154 |
| **5** | The Barley Mow | Pub | 51.453403 | -2.577350 |
| **6** | Harvey Nichols Restaurant & Bar | French Restaurant | 51.457233 | -2.586228 |
| **7** | BBB - Bristol Bear Bar | Gay Bar | 51.456473 | -2.579813 |
| **8** | Côte Brasserie | French Restaurant | 51.457400 | -2.587210 |
| **9** | Nando's | Portuguese Restaurant | 51.458234 | -2.584539 |
| **10** | Five Guys | Burger Joint | 51.458220 | -2.584666 |

A dataframe of neighbourhoods were created for making requests to Foursquare for further exploration. The following tables explore the count of venues per postcodes.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Neighbourhood Latitude** | **Neighbourhood Longitude** | **Venue** | **Venue Latitude** | **Venue Longitude** | **Venue Category** |
| **Neighbourhood** |  |  |  |  |  |  |
| **BS2 0DB** | 42 | 42 | 42 | 42 | 42 | 42 |
| **BS2 0DE** | 42 | 42 | 42 | 42 | 42 | 42 |
| **BS2 0DQ** | 40 | 40 | 40 | 40 | 40 | 40 |
| **BS2 0DR** | 45 | 45 | 45 | 45 | 45 | 45 |
| **BS2 0DS** | 46 | 46 | 46 | 46 | 46 | 46 |
| **BS2 0DT** | 42 | 42 | 42 | 42 | 42 | 42 |
| **BS2 0DU** | 40 | 40 | 40 | 40 | 40 | 40 |
| **BS2 0DX** | 40 | 40 | 40 | 40 | 40 | 40 |
| **BS2 9DT** | 43 | 43 | 43 | 43 | 43 | 43 |
| **BS2 9DY** | 45 | 45 | 45 | 45 | 45 | 45 |

Afterwards, neighbourhoods were explored. The following are the venues obtained from those postcodes. In this example, BS2 0DT was used as the centre to explore its neighbourhood. The following is the result obtained from Foursquare. Places with similar shops are clustered. There are 5 clusters in this example.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **country** | **Neighbourhood** | **Latitude** | **Longitude** | **Cluster Labels** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** | **4th Most Common Venue** | **5th Most Common Venue** | **6th Most Common Venue** | **7th Most Common Venue** | **8th Most Common Venue** | **9th Most Common Venue** | **10th Most Common Venue** |
| **0** | England | BS2 0DT | 51.457825 | -2.580188 | 2 | Pub | Café | Clothing Store | Gay Bar | Sushi Restaurant | Department Store | Gym / Fitness Center | Burger Joint | Music Venue | Pizza Place |
| **1** | England | BS2 9DT | 51.457653 | -2.580344 | 2 | Pub | Café | Clothing Store | Burger Joint | French Restaurant | Sushi Restaurant | Department Store | Gym / Fitness Center | Gay Bar | Music Venue |
| **2** | England | BS2 0DB | 51.457556 | -2.579882 | 1 | Pub | Café | Clothing Store | Gay Bar | Sushi Restaurant | Department Store | Gym / Fitness Center | Burger Joint | Music Venue | Pizza Place |
| **3** | England | BS2 0DS | 51.457786 | -2.580720 | 0 | Pub | Clothing Store | Café | Gay Bar | Burger Joint | French Restaurant | Sushi Restaurant | Coffee Shop | Italian Restaurant | Music Venue |
| **4** | England | BS2 0DU | 51.457621 | -2.579638 | 4 | Pub | Café | Gay Bar | Pizza Place | Music Venue | Burger Joint | Clothing Store | Gym / Fitness Center | Department Store | Fast Food Restaurant |
| **5** | England | BS2 0DQ | 51.457494 | -2.579766 | 4 | Pub | Café | Gay Bar | Pizza Place | Music Venue | Burger Joint | Clothing Store | Gym / Fitness Center | Department Store | Fast Food Restaurant |
| **6** | England | BS2 0DR | 51.457543 | -2.580702 | 0 | Pub | Clothing Store | Café | Gay Bar | Burger Joint | French Restaurant | Sushi Restaurant | Coffee Shop | Italian Restaurant | Music Venue |
| **7** | England | BS2 0DE | 51.457402 | -2.579968 | 1 | Pub | Café | Clothing Store | Gay Bar | Sushi Restaurant | Department Store | Gym / Fitness Center | Burger Joint | Music Venue | Pizza Place |
| **8** | England | BS2 0DX | 51.458232 | -2.579588 | 4 | Pub | Café | Gym / Fitness Center | Gay Bar | Pizza Place | Music Venue | Burger Joint | Clothing Store | Department Store | Fast Food Restaurant |
| **9** | England | BS2 9DY | 51.458480 | -2.580311 | 3 | Pub | Café | Gym / Fitness Center | Clothing Store | Gay Bar | Sushi Restaurant | Department Store | Burger Joint | Music Venue | Pizza Place |



As we can see from the result, most common venues are pub, cafe, clothing store and bar, which is expected as Bristol is a city for vacation and travellers. The result of data illustrate this point.

**Result**

Using one of the postcodes in City of Bristol as example, the result showed that the most common venue in the neighbourhoods are Pub, clothing stores and Cafe. This reflects the reality that Bristol is a travel city, good for retirement and vacation. However, the concentration of schools is quite different from that of Oxford. The result could serve as an initial supporting evidence for selecting an area to settle or study.

**Discussion**

This exercise is a framework for analysis. For further improvement, it would be interesting to see any spending figures or income figures added to the dataframe for analysis. For example, is there any relationship between income, study area and transport availability? What is the best place to study if a person do not want to own a car? This framework can be extended further to answer other interesting questions.