

Data

Data Sources

We are aiming to build a recommender system for neighbourhoods in a city. So the first step we need is to choose a city and then define the neighbourhoods within that city. I have chosen London as the city I will investigate. During the course we have already looked at the cities of New York and Toronto along with their neighbourhoods. The second step is to get venues in each neighbourhood.

Defining Neighbourhoods - Wikipedia

In London (and the whole of the United Kingdom) every place has a postcode. This is alphanumeric and varies in length between 6 and 8 characters (including a space separating the outward and inward code). The postcode is composed of several parts:

POSTCODE			
Outward code		Inward Code	
Area	District	Sector	Unit
SW	1W	0	NY

The Area in the Outward code indicates a large geographic region, a city or part of a large city. There are several Areas in London. The Outward code, i.e. the Area together with the District generally indicate a neighbourhood. We have chosen to use the Outward code as a single district. Note Outwards codes vary in size but geographically smaller Outward codes tend to be more densely populated so we can safely assume Outward codes identify neighbourhoods.

We obtained the London Area Postcodes from Wikipedia^[1]. There were 8 Area Codes for London (references ^[2] to ^[9]) each with several Districts making up a total of 170 Neighbourhoods. The neighbourhoods were stored in a dataframe London_df:

```
London_df.head()
```

	Postcode district	Post town	Coverage	Local authority area(s)
0	E1	LONDON	Algate	Tower Hamlets, Hackney, City of London
1	E1W	LONDON	Wapping	Tower Hamlets
2	E2	LONDON	Bethnal Green	Tower Hamlets, Hackney
3	E3	LONDON	Bow	Tower Hamlets, Newham
4	E4	LONDON	Chingford	Waltham Forest, Enfield, Epping Forest (Essex)

Finding Neighbourhood coordinates - FreeMapTools

Once we had the list of neighbourhoods, we had to find a datasource that would return the latitude and longitude of each postcode district. For this purpose we downloaded a csv file containing the outward code from [FreeMapTools.com](https://www.freemaptools.com/). Note data attribution statements and data licences in ref [10]. Once imported this file contained the outwards code and the latitude and longitude coordinates as shown below:

```
lon_coor.head()
```

	postcode	latitude	longitude
0	AB10	57.13514	-2.11731
1	AB11	57.13875	-2.09089
2	AB12	57.10100	-2.11060
3	AB13	57.10801	-2.23776
4	AB14	57.10076	-2.27073

These two tables were joined to show the London Neighbourhoods with their coordinates:

```
London_df.head()
```

	Postcode	City	Neighbourhood	Borough	Latitude	Longitude
0	E1	LONDON	Algate	Tower Hamlets, Hackney, City of London	51.51766	-0.05841
1	E1W	LONDON	Wapping	Tower Hamlets	51.50775	-0.05739
2	E2	LONDON	Bethnal Green	Tower Hamlets, Hackney	51.52939	-0.06080
3	E3	LONDON	Bow	Tower Hamlets, Newham	51.52789	-0.02482
4	E4	LONDON	Chingford	Waltham Forest, Enfield, Epping Forest (Essex)	51.62196	-0.00339

Finding Neighbourhood Venues - FOURSQUARE API

Once we had a list of neighbourhoods and their respective centre coordinates, the next step was to find information about the venues in each neighborhood. This was done using the FOURSQUARE API [11] as described in the course. We used the coordinates of each neighbourhood to find the first 100 venues within a 500 metre radius of the coordinates.

```
London_venues.head()
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Algate	51.51766	-0.05841	Mouse Tail Coffee Stories	51.519471	-0.058573	Coffee Shop
1	Algate	51.51766	-0.05841	Rinkoff's Bakery	51.519964	-0.053238	Bakery
2	Algate	51.51766	-0.05841	One Mile End	51.520151	-0.056136	Brewery
3	Algate	51.51766	-0.05841	Needoo Grill	51.517070	-0.062379	Indian Restaurant
4	Algate	51.51766	-0.05841	Lahore One	51.514725	-0.059399	Indian Restaurant

Data Cleaning

As there were 8 different postcode Areas in London each with a different Wikipedia page, we had to scrape all 8 different pages. The data cleaning process initial step involved removing non-geographic postcodes (special postcodes assigned to a particular building). The second step involved removing the text after “district:” in each neighbourhood as this just added unnecessary information about that particular neighbourhood. The third step involved giving names to all the subdistricts (e.g. SW1V) falling within a head district (SW1?). For example postcode district “SW1V” was assigned Borough or neighbourhood name “Victoria”. To complete this task you need context knowledge but this information is not necessary as we could simply refer to neighbourhoods by the postcode outward code (Area code together with District code). Once each of the 8 postcode Areas in London were cleaned they were all stacked together to form a large dataframe London_df as shown above.

The coordinates were added by joining on the postcode outward codes (Area code together with District code) from the csv file which returned the table shown earlier.

From the FOURSQUARE API all the venues in each neighbourhood were obtained using the coordinates. One hot encoding was performed on the venue categories. All venue categories were summed over the neighbourhood so as to keep a count of each venue category. For example there were 6 neighbourhoods with “Accessories Stores”:

```
London_grouped[London_grouped['Accessories Store']!=0.0]
```

	Neighbourhood	Accessories Store	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	A Cra Stc
13	Bloomsbury, British Museum, Southampton Row	1	0	0	0	0	0	0	0	0	0	2	1	
21	Charing Cross	1	0	0	0	0	0	0	0	0	0	1	2	
30	Covent Garden	1	0	0	0	0	0	0	0	0	0	0	0	
77	Leicester Square, St. Giles	1	0	0	0	0	0	0	0	0	0	0	0	
97	New Oxford Street	1	1	0	0	0	0	0	0	0	0	0	0	
106	Oxford Street (west)	1	0	0	0	0	0	0	0	0	1	6	0	

We can also see that Oxford Street (West) has 6 art galleries but this exploratory analysis will be developed later on.

References

- [1] Wikipedia contributors. (2021, February 26). London postal district. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:21, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=London_postal_district&oldid=1009140985
- [2] Wikipedia contributors. (2021, March 1). E postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=E_postcode_area&oldid=1009531161
- [3] Wikipedia contributors. (2020, December 18). EC postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=EC_postcode_area&oldid=994966749
- [4] Wikipedia contributors. (2021, January 21). N postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=N_postcode_area&oldid=1001787686
- [5] Wikipedia contributors. (2021, January 11). NW postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=NW_postcode_area&oldid=999732776
- [6] Wikipedia contributors. (2021, February 26). SE postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=SE_postcode_area&oldid=1009068149
- [7] Wikipedia contributors. (2021, February 28). SW postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=SW_postcode_area&oldid=1009466671
- [8] Wikipedia contributors. (2021, February 11). W postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=W_postcode_area&oldid=1006095827
- [9] Wikipedia contributors. (2021, January 7). WC postcode area. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:19, March 2, 2021, from https://en.wikipedia.org/w/index.php?title=WC_postcode_area&oldid=998980977
- [10] Freemaptools.com. (2021). Download UK Postcodes with Latitude and Longitude. Retrieved 17:19, March 2, 2021, from <https://www.freemaptools.com/download-uk-postcode-lat-lng.htm>
- Contains Ordnance Survey data © Crown copyright and database right 2020
 - Contains Royal Mail data © Royal Mail copyright and database right 2020
 - Source: Office for National Statistics licensed under the Open Government Licence v.3.0
- [11] FOURSQUARE API [Attribution | Usage Guidelines \(foursquare.com\)](#)