Locating a new venue in an unknown city

Using opening hours to determine foot traffic hot spots

Dave Corrigan-Kavanagh

August 8th 2019

1. Introduction

Investors frequently need to make judgements about an opportunity in an area that they are not familiar with. Local regulations, cultural norms, and geographic idiosyncrasies make it difficult to know the layout of a foreign city's busy spots during different times of day without spending significant time there.

The time-distributed foot traffic can be inferred from the opening times of the existing venues.

Investors, planners, and business partners would benefit from knowing the distribution of foot traffic throughout the day so as to avoid situations like:

- investing in a nightclub on a street filled mostly with daytime stores
- investing in a hardware store in the middle of the clubbing district
- opening any venue without being aware of the geography of the city

For example consider Manchester’s Gay Village. It has close proximity to shopping centres but predominantly hosts late-opening pubs and clubs (and some daytime opening cafes). The location would suggest suitability for a store, but the foot traffic tells us that this area is not as busy during the day. Similarly in cities on mountainous terrain (eg. Funchal), a venue placed 300 metres from a busy area may see low foot traffic due to being at the top of a steep road – something which is not readily obvious when looking at the map.

2. Data acquisition and cleaning

2.1 Source

Foursquare's API provides all of the required data for this task, including venue categories, latitude/longitude coordinates, and opening hours. The script only requires a set of latitude/longitude coordinates to identify the centre of the area of interest. In this script I have focussed on Manchester in the UK.

2.2 Acquisition

Careful consideration is made to work within the limits of the API's free service, which limits the number and type of queries that can be made, as well as the number of results each query returns. Two types of request are made to the API; first to create a dataframe of venues in the area, then to request the opening times of each of those venues.

To retrieve the venues a grid is generated using the lat/long coordinates of the *area of interest* as the centre, and the Foursquare API is queried for each point on the grid. The API returns a maximum of 30 venues for each query, so the grid is dense so as to ensure overlap and therefor coverage of all venues in the *area of interest*. The results of each query are combined in a dataframe, and the duplicates are removed.

Once the venues dataframe is complete the API is called again, this time requesting the opening hours for each venue. These are considered premium calls so only 500 are permitted each day. The script saves the results locally so that it can be run daily until opening hours are provided for every venue.

2.3 Cleaning and feature selection

Category data from Foursquare is verbose so it is simplified into categorical labels. Venues fitting certain categories are removed so as to reduce the number of API calls that will subsequently be made to request opening hours. This step is a compromise to accommodate the API limits, and would preferably be skipped if time or access to the data was more generous.

Any venues sitting outside the edges of the *area of interest* are also removed. To complete the venues dataframe certain attributes are removed as they are not needed. The dataframe consists of just under 2000 venues.

Opening hours are queried for each venue in turn, and the results are saved locally. This step is repeated over several days as it reaches the API limit each day. Each json densely presents the opening hours of a venue (for example one row can represent several days) so as each json is

returned, a short script flattens it into a dataframe, loops over the values to expand into a longer form, and appends that to another dataframe outside the loop, which contains the wide-form opening hours of every all venues.

The times are provided as strings. The script converts them to floats by converting the hour characters to integers, and scaling the minute characters up by 100/60. The resulting times are decimal analogs of the hours (1.5 == 1:30am, 17.75 == 5:45pm, etc.), and contain some time values greater than 24, which represent closing times during the following day.

Finally, the dataframe is reshaped to a narrow format, where each row represents a set of open/close times as opposed to a venue, and any rows without open/close times is removed. This step prepares the data to be converted to a numpy array later on.

3. Methodology

3.1 New city, new norms

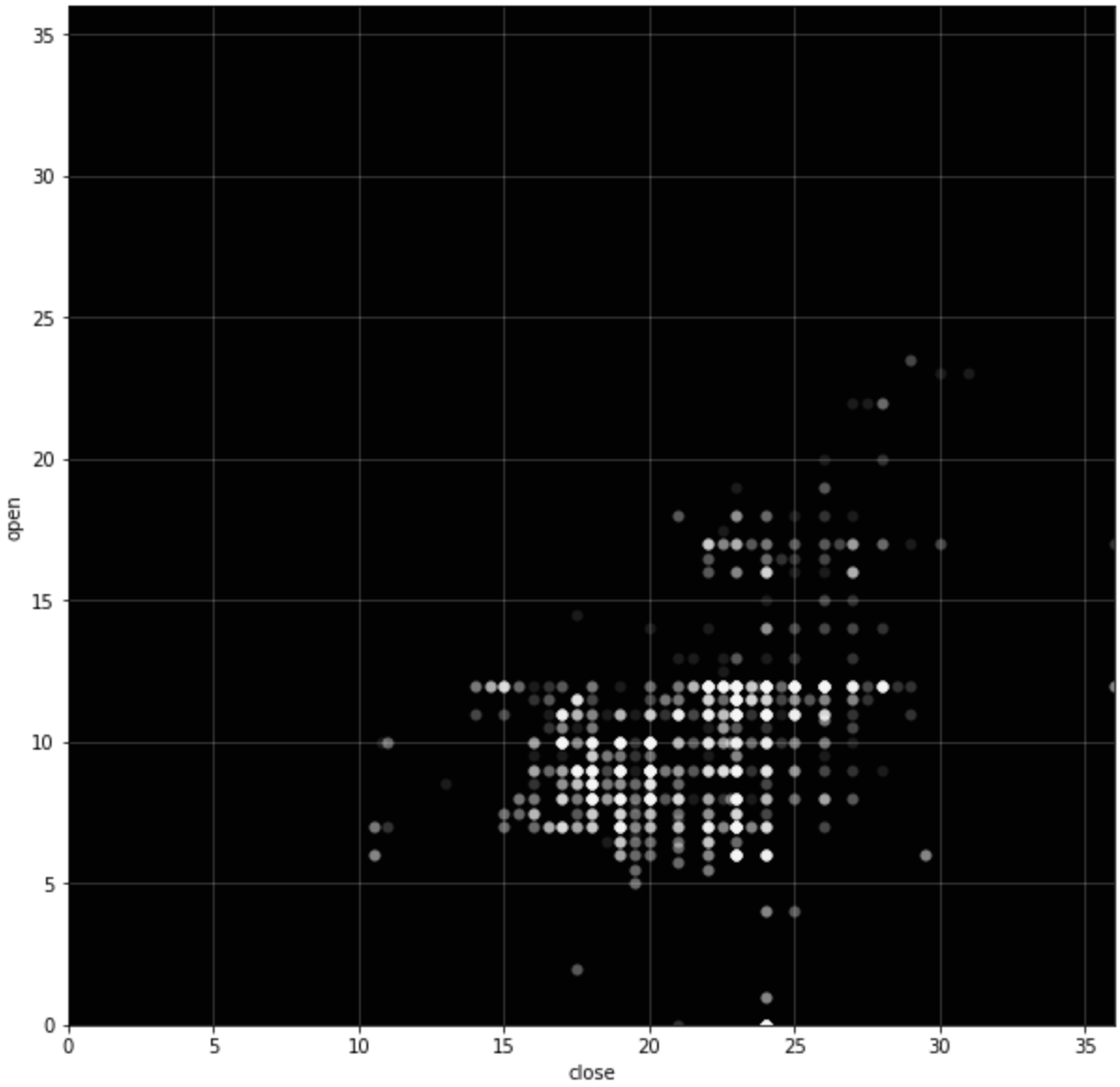
Local regulations and cultural norms result in different sets of typical opening hours in each city. The intention is that any city's venues could be analysed without pressing on it the assumptions formed by observing other cities. Unsupervised machine learning will be used to cluster the venues both by their opening hours and locations to predict foot traffic.

In the first section I have extracted the data. Next will be two phases of machine learning and exploratory analysis. Manchester was selected as the area of interest as I am intimately familiar with the city centre, and this allowed me to perform essential sanity checks during the exploratory analysis.

3.2 Clustering opening hours

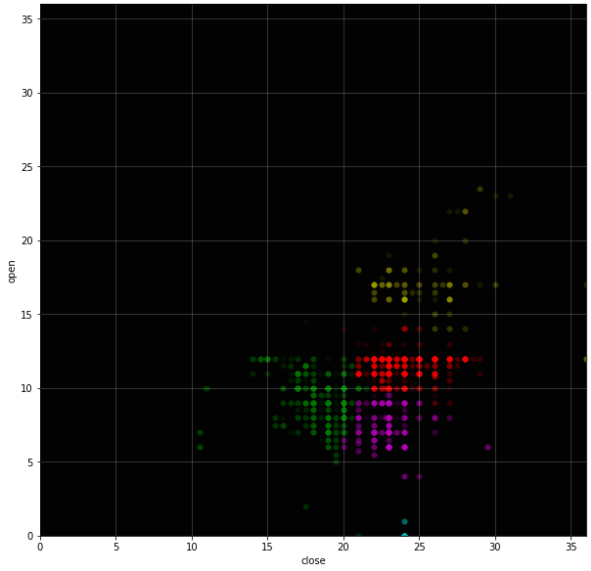
By selecting an unsupervised clustering technique, we can learn about the typical sets of opening hours without needing an awareness of local norms.

The data points are quite widespread:

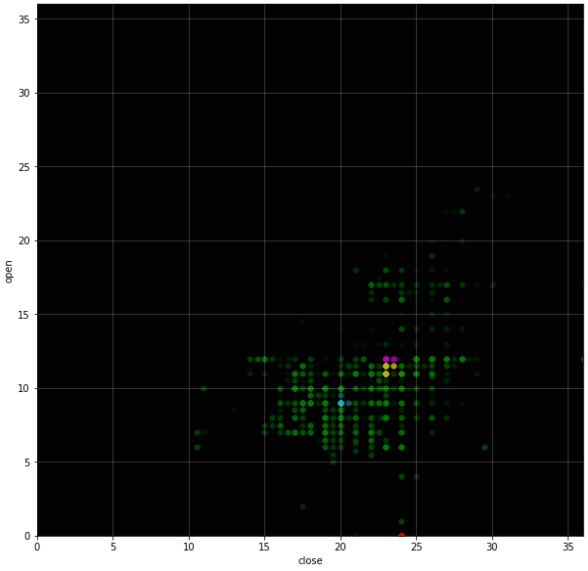


K-means and DBCSAN were attempting for this phase. K-means was favoured as it produces clusters suitable for separating out wide-spread data points. Since DBSCAN works by identifying high-density areas, it tended to produce very distinct clusters with a majority of outliers.

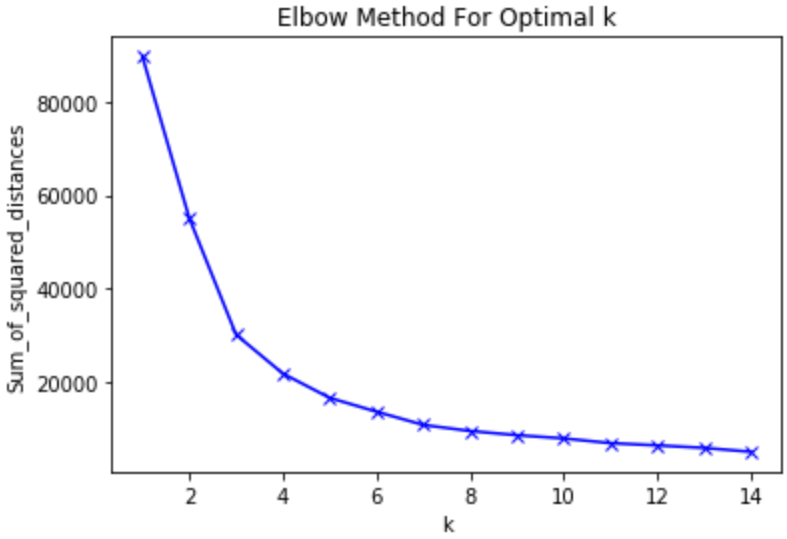
K-means



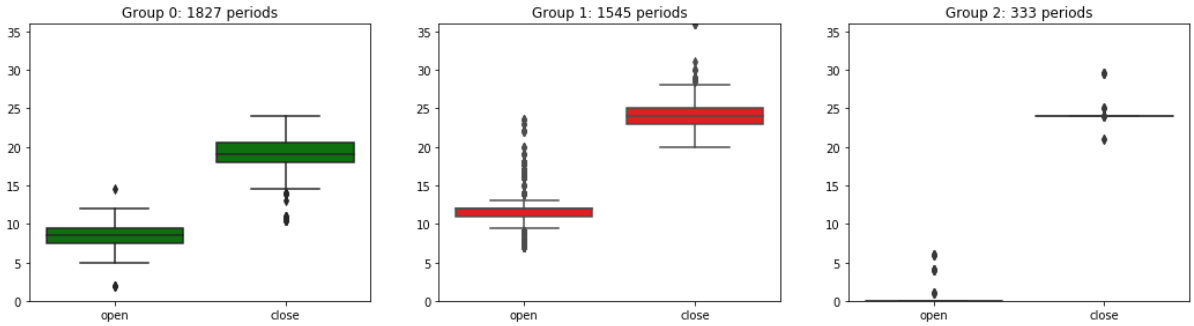
DBSCAN



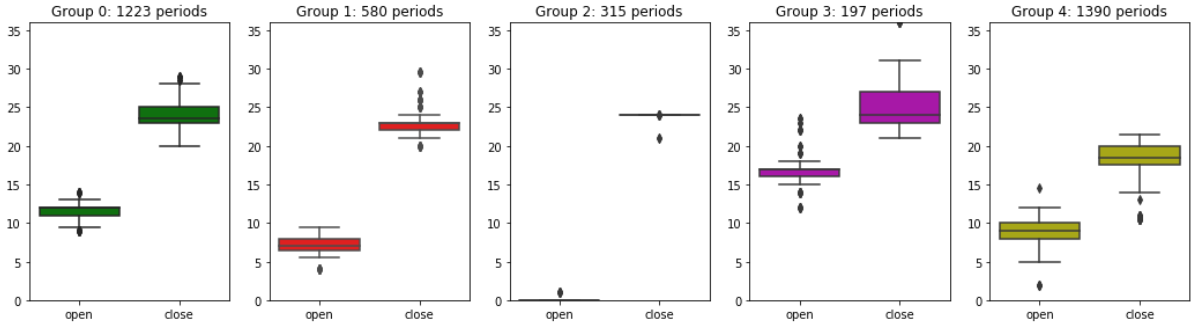
While assessing the K-means model, the elbow point suggested that k=3 would be the most suitable value, however k=5 produced more distinct clusters. This discrepancy is an area that would require more detailed investigation in the future in order to fully automate the process.



The elbow point suggests an optimal k value of 3.

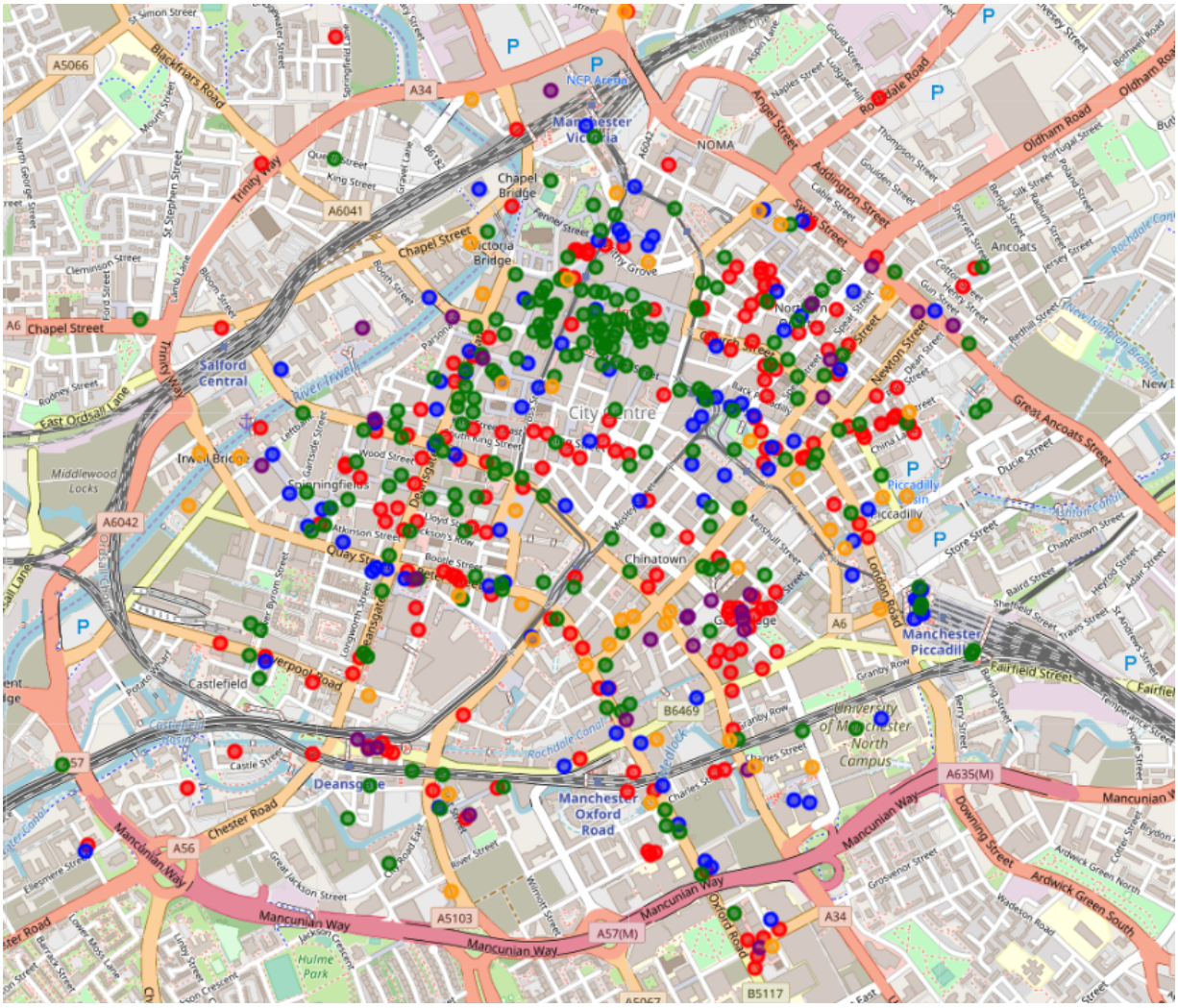


Each group at k=3 contains datapoints which are statistical outliers



At k=5 the groups become more distinct

Once the k-means model was trained the labels were applied to each set of opening hours. At this point we could check on the locations and types of venue that these groups represented. The map showed that some location clustering already existed for the opening hour groups.



Some distinct areas exist where most venues share their opening hours group

Extracting the times of each cluster along with the most common venue type acted as another sanity check against my knowledge of the area. The groups made important distinctions for example between a standard pub, and a gay bar.

1	2	3	4	5
12:00- 24:00 Pub	7:00 - 23:00 Grocery Store	9:00 - 18:00 Coffee Shop	17:00 -25:00 Gay Bar	0:00 - 24:00 Hotel

3.3 Clustering venues by locations

The next phase is to cluster the venues by their location. By filtering to a specific set of opening hours we can then check which parts of the city would experience foot traffic at those times. For example we can select group 1 (Pubs, open midday to midnight) to find out which areas are busy from the afternoon onwards.

For this phase DBSCAN was suitable as it finds areas of higher density. First a function needed to be created, which would accept two arguments: opening time and closing time. The function predicts an opening-hour group from the pre-trained k-means model, then selects only the venues matching that group. Next DBSCAN clustering is applied to the lat/long coordinates of the remaining venues to produce the location clusters of the venues.

Finally, once those clusters have been determined it is simple to produce a colour-coded map and some information on the venues inside them such as the type of venue, the lat/long central point, and an address at that location, which is retrieved via the geolocator API.

Below is a sample when the function was run with opening hour arguments 12 and 26:

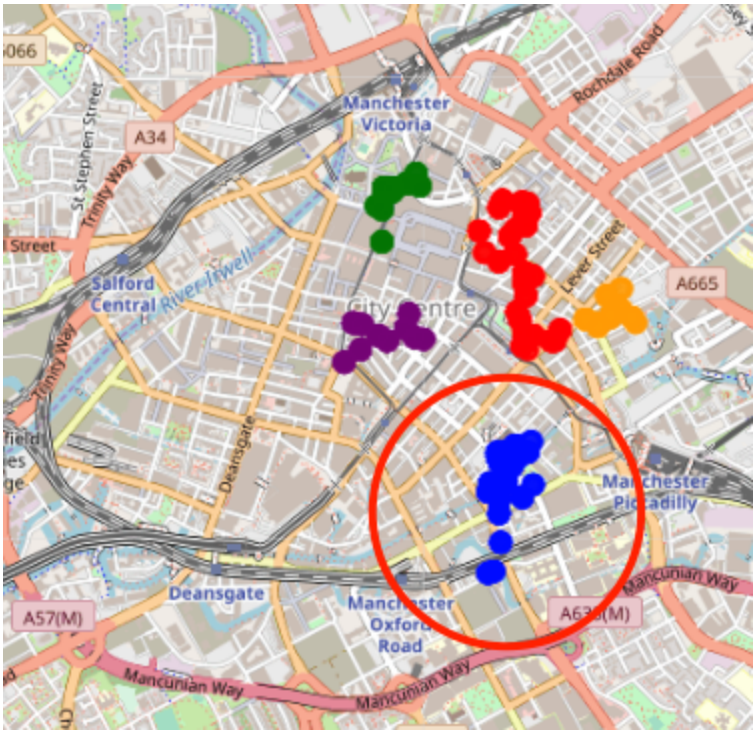
```
==========================================================================
Suitable areas around 53.480472, -2.246467 for a venue opening between 12:00 and 26:00
==========================================================================

Locally the most similar common opening times are from 11:00 to 24:00
Venues sharing these opening times are grouped around the areas below.
These areas should promise the greatest foot traffic during the selected opening hours.


Group A is centred around 53.48266374456969, -2.2365070173732184
--------------------------------------------------------------------------
Suggested address to investigate: Big Horn, Tib Street, Northern Quarter, City Centre, Manchester, Greater Manchester,
North West England, England, M4 1PQ, UK
Venues in this area with similar opening times are mostly of type 'Bar'


Group B is centred around 53.47651120079398, -2.2372693065625158
--------------------------------------------------------------------------
Suggested address to investigate: Lock87, 3, Brazil Street, Gay Village, City Centre, Manchester, Greater Manchester,
North West England, England, M1 3PJ, UK
Venues in this area with similar opening times are mostly of type 'Gay Bar'
```

Among other areas the function had correctly identified Manchester’s Gay Village simply by clustering the hours and location.



Group B in blue is a cluster of venues sharing late opening and very late closing times, located in Manchester's Gay Village.

4. Results

Given a centre for the area of interest we were able to obtain extensive data on existing venues in the area, and infer suitable locations for a venue purely on the basis of opening hours.

Our analysis showed that the area of interest contained 5 typical opening hour groups, and that particular categories of venue belong in each. It also showed that those venues tended to cluster together physically, allowing us to identify locations that would be suitable for a similar venue, and avoid researching areas where foot traffic during the opening hours would be low.

The script is arranged in such a way that a new area of interest could be selected for analysis, and any combination of opening hours tested for suitable locations. The final clusters are reached without supervision and the result gives both high level feedback (visual map) and more detailed information (the addresses at the centre of each cluster), which would form the basis for further investigation by the stakeholders.

5. Discussion

5.1 Pragmatically, the goal was achieved

The primary purpose of this project was achieved, though there are nuances outside the scope of the project that could be addressed to improve it.

5.2 Retrieving the data

Data retrieval formed the large bulk of this task, and redundancy was built in as an imperfect assurance of comprehensive coverage. Applying the notebook to a new area of interest would require several days of querying the Foursquare API and reaching the daily limit. A better approach might involve more permissive access to the data source so that a single query with a radius argument could return the results rather than using a large grid.

5.3 The optimal k value

It was troublesome that the apparent optimal k value (5) was not the elbow point (3). A better form of validation is clearly needed in order to automate that section of the analysis. Further investigation of opening hours in different areas might help to determine an acceptable proportion of outliers in the groups.

5.4 Checking against day of the week

In this analysis each venue was represented as a set of opening hours. If a venue opened 7 days a week it would be represented by 7 sets of opening times in the final dataframe, however if a venue opens only a day per week it will appear as only 1 set. This allowed for simpler analysis as there were no NaNs in the dataframe but did not allow for outliers or for clustering venues by their typical opening times across the week. A more sophisticated analysis would need to account for venues that are fully closed.

An example weakness of the current project would be certain parts of central London, which are busy on weekdays as they are almost exclusively composed of office buildings, but entirely empty on the weekends. As it stands, my current analysis would assess those areas as busy between 9:00 and 18:00 but fail to mention that the day of the week is important.

5.5 Better coordinate systems

A better coordinate system could have been used when creating the grid. The current solution was to find a grid that roughly covered a city centre, while in reality we would want to specify a radius around the area of interest. It was sadly outside the scope of this project to become versed in the different coordinate systems, but a conversion from lang/long to X/Y coordinates would have helped in querying the area.

6. Conclusion

The final function is imperfect but produces useful results and at no cost. The end result correctly identifies hot spots of foot traffic in the area purely on the basis of intended opening hours. It could certainly inform an investigation into a new area while successfully navigating cultural blind spots and a lack of local knowledge.