Searching for suitable location in Central London for a new Music Venue

Jakob Hegley

August 17th 2020

Introduction

1. Background

Our scenario is that we have been hired by an up and coming entertainment company to search through the central Greater London area for a suitable location for them to create a new music venue. This music venue will cater to a wide variety of musical tastes, so it won't need to be constricted by certain local genre preferences. At first glance it seems like music venues can be a gamble as to whether they can be profitable. Our employee came to the conclusion that the most important factor comes down to location. Getting acts to perform usually is never the issue, it's getting people to show up to your venue as opposed to others. London has always been a musical powerhouse, with several genres of music being invented within its bounds over the years. Finding a particular location where there is a lack of any sort of venue can be the key to any sort of success.

2. Problem

First and foremost, we need location data. In particular we need location data about music venues found in the Greater London area as well as a way to designate areas of London that may be of interest. We also might wish to understand the surrounding population values around each of these venues. With this we can then do some analysis to find patterns with these areas that will help us narrow down some final areas where a music venue could prosper.

3. Interest

What's important about this is that the methods we use here could be applied to any sort of exploratory analysis to find possible locations for a variety of venue types. Restaurants, sports venues, or even big corporate conglomerates could use these methods to find suitable locations for their next possible venue.

Data Acquisition and Cleaning

1. Data Sources

First off, we need to decide on what our locations are. We chose to use the London postcode areas as our general areas, and then their subdistricts as our actual "areas of interest." To grab a detailed list of these areas we shall scrape the Wikipedia page for the London postal districts (https://en.wikipedia.org/wiki/London_postal_district), in particular each Wikipedia page for each district, each of which this initial page links to. For example, for the "E" district we use the page https://en.wikipedia.org/wiki/E_postcode_area.

Next, we need the latitude and latitude values for each of these subdistricts. Thankfully these can be found from https://www.freemaptools.com/download-uk-postcode-lat-lng.htm. Now we have our areas of interest and their map references. We will also want to know the size of each district (in square meters) as well as their populations, which together will give us their population densities. I found this data at https://www.streetlist.co.uk/ and https://www.streetcheck.co.uk/postcode/alldistricts. I had issues grabbing the data from these two sites directly, so I collected the data collectively from each site by hand and created a single .csv file.

Finally, we shall use the FourSquare API to pull venue data within a certain radius (defined by the size of the district) for each postal district. We shall do our API calls within the Jupyter notebook itself. From this we will get the top 100 venues (as defined by FourSquare), as well as their general venue type/category and lat/long coordinates, within a radius about the latitude and longitude of each district.

2. Data Cleaning

We began by pulling all of the postcode district data from the Wikipedia pages into a single table. The included features were the "Postcode district" label, the post town label (which in our case was just LONDON), the "Coverage" which represents points of interest that are found in the postcode, and the "Local Authority area(s)" which are the London boroughs that each postcode district lies within. In the end we only desire the "Postcode district" and "Local Authority area(s)" features, so the others we dropped.

Next, we pulled in the latitude and longitude data in. This dataset however had the data for all of the postcodes of London, a large majority of which we did not need. We were only interested in those that are found in the Greater London area. We did a comparison between the sets and pulled only the lat/long data that matched with the postcodes from our other dataset. From there we combine these two data sets into a single DataFrame that has the "Postcode district," "Borough(s)" (a renaming of the "Local Authority area(s)"), the "Latitude," and "Longitude" features.

Now after some digging, we find that some of the data points refer to postcodes that actually do not represent geographical regions. We remove these data points by singling out the points that have “non-geographic” in their “Borough(s)” feature. We actually needed to delete a few extra points due to the structuring of the Wikipedia pages, in particular their “non-geographic” listing was under a different feature, and some points represented postcodes that were actually fictional (usually used on television).

Next, we pull in the file I created from hand that contains the features “Postcode district,” “Area (km2)” which has the surface area of each district in square kilometers, “Population,” and “Population Density.” Once again, we combine this set with our other set, using the “Postcode district” as a similar feature to facilitate the joining.

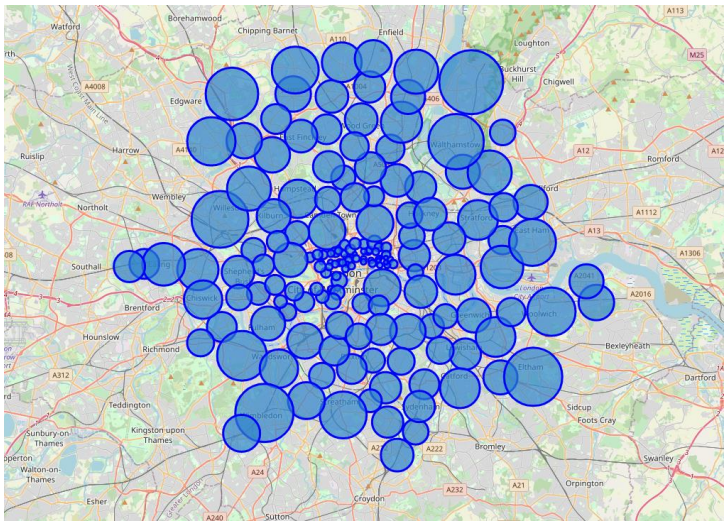
Pulling the FourSquare data using their API was rather simple, we just push our latitude and longitude coordinates in the APU call. The main point of interest I want to bring up is how I made each call using a different value for the RADIUS input. My thought process for this was that each postcode varies widely in their size, so if I just use a consistent value of say 500 meters, then some postcode districts will be widely undervalued in their venue options, while others will overlap so much that we will have many repeating venues, which will throw off our results. Thus, in the loop I used for each postcode district, I passed their “Area (km2)” value into the RADIUS input, and converted it to meters. Notice that I’m passing in an area, and turning it into a pure distance. My method here was to consider each postcode district as a perfect square with an area matching that found in their “Area (km2)” column. Then, I considered a circle inscribed within the square. We know that the ratio between the area of this square and the area of this inscribed circle is one-fourth pi. Thus, we use this ratio to turn the square area into the relative area of the circle, then we use the equation for the area of a circle to find the radius of this circle. Finally, we convert this area from kilometers to meters. This is the value that we pass, for each postcode district, into the RADIUS input of the API call.

From here we run our loop to grab the top 100 (or less depending on the district) venues from Foursquares' database for each postal district. We pass these results to a DataFrame with the following features: “Postcode,” “Postcode Latitude,” “Postcode Longitude,” “Venue,” “Venue Latitude,” “Venue Longitude,” and “Venue Category.” With this, we have all the necessary data we need to begin our data analysis.

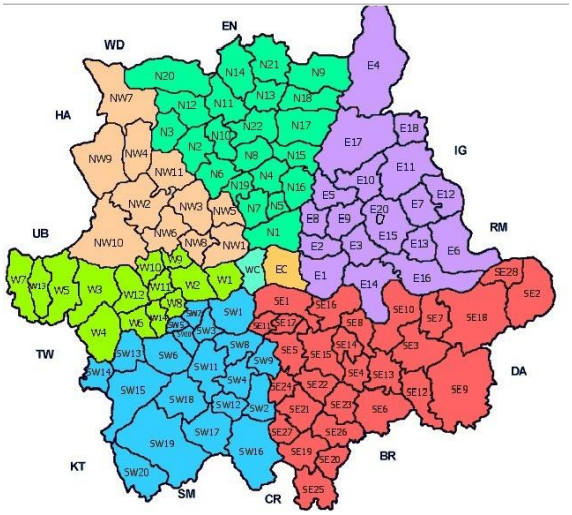
Exploratory Analysis and Methodology

1. Initial looks

First off, we did a general mapping, using the folium python package, of all 169 postal districts, showing the radii that we used for the API call. This shows us how much coverage we have in comparison. Recall that each of these circles are centered around our defined latitude and longitude coordinates for each district:



Let's compare this to an image of the actual boundaries:

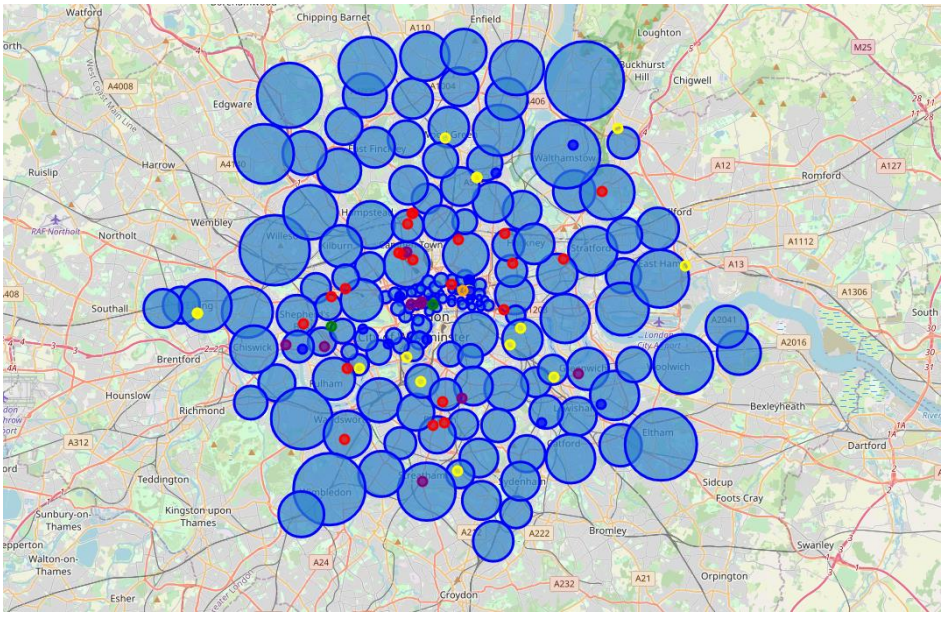


Overall, our radii appear to be a good measure of the majority of each district. We can definitely see from the central areas (which on the colored map they don't even separate because they are so small) that our choice of differing radii was a wise one.

2. Looking for Music Venues

Next, we look into the venues themselves. We did a look to see how many venues we got from each district, and were surprised to find that only about twenty percent of them actual had 100 (or above) venues to consider. While some of these are just the largest districts, which makes sense, we can also see that the one near the very center of London also have some of the highest numbers of venues, despite their small sizes. Next, we looked into the which venue categories we actually found. In total, we found 398 unique venue categories out of 9636 venues. From searching through these categories, we took note of which ones could be considered a "musical venue." We created a list of the following categories: Music Venue, Concert Hall, Opera House, Jazz Club, Piano Bar, and Performing Arts Hall. While some of these categories don't exactly match the style of venue our employer is trying to create, it doesn't hurt to get the full view of the music venue scene throughout our areas of interest.

We pulled the venues that matched these categories and came up with 61 data points. Several of these are duplicates due to the minor overlapping of the radii. We mapped these venues to get an idea of the spread of music venues within our areas of interest. We also included the district radii to get a general idea of where music representation might be lacking.



They are color schemed as follows: Red for Music Venue, Blue for Concert Hall, Green for Opera House, Purple for Jazz Club, Orange for Piano Bar, and Yellow for Performing Arts Hall. Right away we can see some potentially good areas of interest. It seems that the Northwestern area (NW) is almost completely missing any musical performance venues entirely. While the Southeastern (SE) section has some sorts of performance venues, they are only concert halls, a jazz club, and a performing arts hall. A more general music venue would

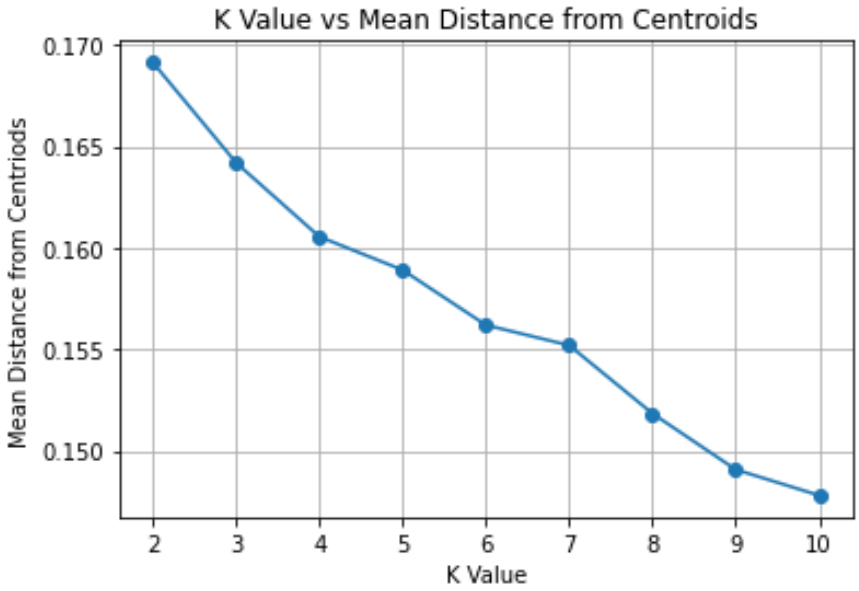
probably fit very well in there. Another area of interest is the most eastern edge of the West (W) postcode. We do see some Jazz clubs and a Concert Hall here but nothing more general. While there are some music venues nearby, they are still a decent enough distance away to leave an interesting gap worth exploiting. These three locations will most likely be our areas of choice. However, we still need to narrow down a specific postcode within each district area.

3. K-Means Clustering

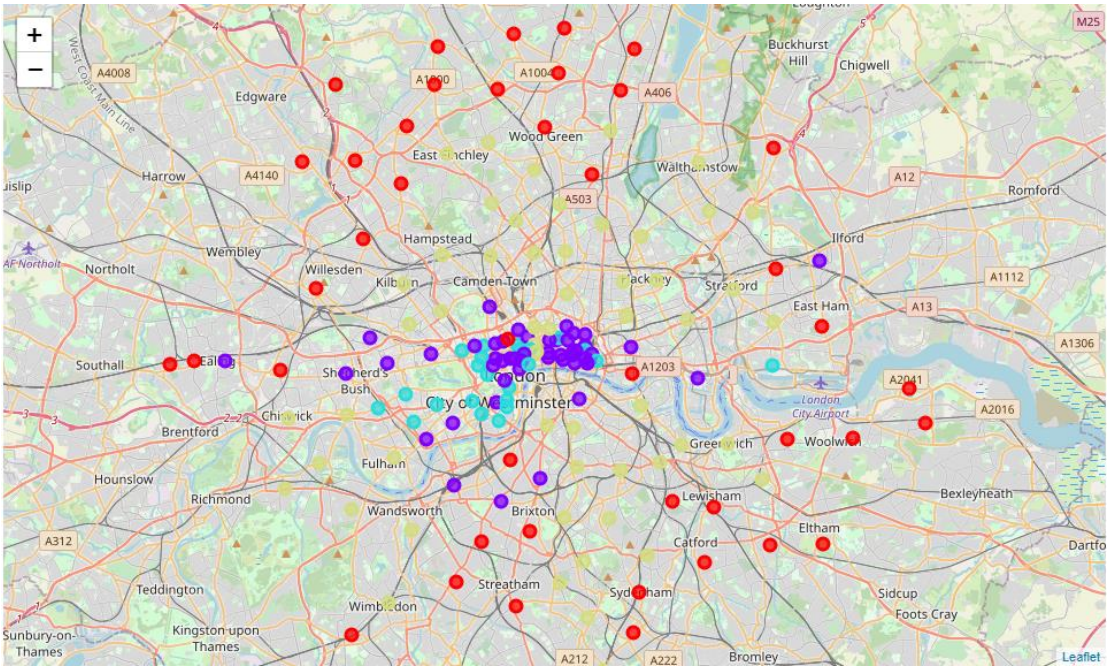
In order to get a better feel for the venue styling of each postcode district, we shall use the K-Means clustering algorithm. But before we even begin our analysis, we need to turn our data from categorical to numerical. K-means clustering involves calculating distances between values in some sort of numerical space, so we must find a way to give our venue data some sort of numerical value.

We achieved this via creating a frequency scale. We shall calculate the frequency of each venue category within each postal district. This was achieved by onehot encoding a new set of columns into our dataset, one for each venue category. Then for each venue we fill in the category column that venues' category falls into with a one, and every other category column is filled with a zero. Next, we sum up all of these columns for each postal district and then divide by the total amount of venues in that district (essentially taking the mean). This yields us a single row for each of the 169 postcode districts with each of their category columns filled with what percentage (in decimal form) that venue category takes up of the total amount of venues in that district. This gives us our necessary numerical measure of the venues that we can use to do our K-means clustering.

Before we can do our analysis though, we must first decide upon an appropriate k-value for our K-means algorithm. Recall that the k-value represents the number of clusters our algorithm will bin our data points into. There are a variety of methods to decide which k-value should be used. In our case we chose to use the "elbow method." This method involves running the K-means algorithm for multiple values of k, then calculating the average distance each data point is from its designated centroids center. We then graph these distances for each k-value. There should be a general negative slope to the result, which corresponds to the concept that more clusters mean each point will be much closer to their clusters centroid. What we are looking for in this graph is the first moment where the downward slope takes it first drastic upward turn (the negative slope becomes more positive by a significant amount). This is our "elbow point." I ran this for k-values between two and ten and graphed the results:



We can see that the first upward turn happens at a k-value of four. Thus, we chose this value to run our actual algorithm for analysis. Once the K-means algorithm was run with a k-value of four, we the cluster labels to a data frame that listed everything from our original set as well as the top five frequent venues in each postcode district. From here we mapped each district and color coated them based on with cluster they fell into:



Immediately some patterns can be seen. The red clusters primarily take up the outer edge of the districts, the yellow take up the mid perimeter, and the purple and blue take up the middle areas. To get a deeper look at what these clusters are tell us we looked at the points in each cluster.

4. Cluster Analysis and Discussion

I have included the first few points in each cluster to help visualize where my analysis came from.

1. Cluster 0 (Red):

	Postcode	Population Density (Person per km2)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	E18	6814.19	0	Italian Restaurant	Grocery Store	Coffee Shop	Bar	Supermarket
10	E1W	15889.09	0	Coffee Shop	Park	Italian Restaurant	Pub	Pizza Place
15	E6	8483.64	0	Hotel	Supermarket	Coffee Shop	Pub	Grocery Store
16	E7	15765.08	0	Grocery Store	Café	Hotel	Pub	Indian Restaurant
44	N11	7252.74	0	Grocery Store	Hardware Store	Electronics Store	Sporting Goods Shop	Café
45	N12	5971.68	0	Coffee Shop	Grocery Store	Supermarket	Café	Fast Food Restaurant
46	N13	8451.04	0	Grocery Store	Greek Restaurant	Park	Pub	Italian Restaurant
47	N14	5047.63	0	Pub	Grocery Store	Park	Gym / Fitness Center	Café
48	N15	14305.82	0	Café	Grocery Store	Coffee Shop	Bus Stop	Pizza Place
51	N18	6856.95	0	Pub	Turkish Restaurant	Grocery Store	Supermarket	Coffee Shop

With the abundance of Grocery stores and Coffee shops being the most common, I would suspect that these represent residential areas. Still plenty of Pubs and simple restaurants as well, but not a huge variety. Notice how these areas actually have some the most average spread of population densities, showing that it is more uniform. These would most likely not make good places for a music venue due to it not being supported by the businesses already in place. Notice how the Northwest (NW) area is almost entirely found in this cluster, which would explain its lack of musical performance venues.

2. Cluster 1 (Purple):

	Postcode	Population Density (Person per km2)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	E1	20226.51	1	Coffee Shop	Pub	Hotel	Café	Indian Restaurant
3	E12	8532.36	1	Indian Restaurant	Train Station	Restaurant	Gym / Fitness Center	Park
5	E14	12126.99	1	Italian Restaurant	Hotel	Coffee Shop	Park	Indian Restaurant
19	EC1A	11559.26	1	Pub	French Restaurant	Wine Bar	Coffee Shop	Italian Restaurant
21	EC1N	24377.78	1	Coffee Shop	Sandwich Place	Pub	Food Truck	Gym / Fitness Center
23	EC1V	22143.68	1	Coffee Shop	Food Truck	Pub	Italian Restaurant	Café
24	EC1Y	25431.03	1	Food Truck	Coffee Shop	Pub	Hotel	Gym / Fitness Center
25	EC2A	8827.27	1	Gym / Fitness Center	Italian Restaurant	Food Truck	Coffee Shop	English Restaurant
26	EC2M	7393.10	1	Sandwich Place	Coffee Shop	Gym / Fitness Center	Boxing Gym	Café
28	EC2R	9318.18	1	French Restaurant	Wine Bar	Coffee Shop	Modern European Restaurant	Sushi Restaurant
29	EC2V	7911.11	1	Coffee Shop	Steakhouse	Clothing Store	Italian Restaurant	Plaza
30	EC2Y	17125.00	1	Italian Restaurant	Art Gallery	Deli / Bodega	Indie Movie Theater	Café

This definitely seems to represent the highly economically prosperous downtown areas of central London. Tons of high-quality restaurants, simple food shops, and coffee shops mixed together, as well as a large variety of entertainment and shopping options. While these areas would be great to make a new music venue, as we already saw, these areas already have a plethora of musical performance venues, so it would be difficult to make something new prosperous.

3. Cluster 2 (Blue):

	Postcode	Population Density (Person per km2)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
7	E16	6054.73	2	Hotel	Grocery Store	Coffee Shop	Café	Pub
20	EC1M	22290.91	2	Beer Bar	Hotel	Italian Restaurant	Plaza	Spanish Restaurant
27	EC2N	4711.11	2	Café	Event Space	Hotel	Boxing Gym	German Restaurant
33	EC3N	5814.29	2	Hotel	Sandwich Place	Cocktail Bar	Coffee Shop	Salad Place
113	SW1A	2831.09	2	Outdoor Sculpture	Pub	Plaza	Hotel	Monument / Landmark
115	SW1H	22129.41	2	Coffee Shop	Hotel	Hotel Bar	Juice Bar	Café
116	SW1P	17479.78	2	Hotel	Coffee Shop	Café	Italian Restaurant	Restaurant
117	SW1V	25081.48	2	Hotel	Pub	Pizza Place	Turkish Restaurant	Park
118	SW1W	13237.04	2	Hotel	Italian Restaurant	Pub	Bakery	Coffee Shop
119	SW1X	13427.50	2	Hotel	Café	Pub	Plaza	Italian Restaurant
125	SW5	25092.21	2	Hotel	Pub	Café	Garden	Italian Restaurant
127	SW7	15056.71	2	Café	Hotel	Exhibit	Italian Restaurant	Science Museum

This cluster is dominated by hotels. Still plenty of restaurants to be seen, but we can see that these are huge tourist areas. The one outlier to the east can be explained by noticing its vicinity to an airport. Certainly, these could be areas of interest, but we need to consider music venues already in place.

4. Cluster 3 (Yellow):

	Postcode	Population Density (Person per km2)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	E10	9601.98	3	Pub	Grocery Store	Café	Park	Chinese Restaurant
2	E11	7031.32	3	Pub	Grocery Store	Park	Café	Restaurant
4	E13	14978.84	3	Pub	Café	Gym / Fitness Center	Convenience Store	Grocery Store
6	E15	8026.76	3	Pub	Grocery Store	Park	Hotel	Café
8	E17	7783.74	3	Pub	Grocery Store	Coffee Shop	Café	Pizza Place
11	E2	18578.21	3	Coffee Shop	Pub	Café	Wine Bar	Cocktail Bar
12	E3	11816.18	3	Pub	Coffee Shop	Café	Pizza Place	Grocery Store
13	E4	3436.87	3	Pub	Grocery Store	Coffee Shop	Park	Italian Restaurant
14	E5	11985.40	3	Pub	Café	Park	Coffee Shop	Grocery Store
17	E8	17146.57	3	Pub	Café	Coffee Shop	Cocktail Bar	Bakery
18	E9	8553.43	3	Pub	Coffee Shop	Café	Bakery	Italian Restaurant

This cluster is nothing but Pubs, Coffee shops, and simple restaurants as well as some parks and recreational spaces. This most likely just describes generally high populace areas where people go for simple food. These can be great areas of interest because of the high foot traffic they most likely get.

Results

With all of this analysis done we can narrow down our options to these several choices, in order of likelihood of success:

- **W1** (in particular W1K, W1J, W1S): This is our best choice. Relatively high population densities, with plenty of surrounding general venues to keep foot traffic flowing to spread interest. While there may be some nearby Jazz Clubs and a Concert Hall, this is a solid lack of a general music venue for quite a distance. Since most of this postcode area works, there is likely many real estate options as well.
- **SE3**: This one's biggest strength is its distance from any musical venues while still falling into cluster 3, which means that the nearby food options could be an advantage. The surrounding areas are also highly residential, so the music venue would cater to a large amount of people.
- **NW** (the northwest portion in particular): This is one that has one major flaw and one major strength. The strength is the utter lack of any music venues of any type whatsoever, making the desire for one probably rather high. Its biggest flaw is that it is almost entirely residential, so the music venue would have to create its buzz and support itself all on its own.

Conclusions:

Our final decision will be to look into available real estate options in the W1 area. This area has the perfect mix of economic success, potential for growth, and an acute lack of a general musical venue. The other areas shown in the results section show a good deal of possible success as well. It will all come down to what our employer is willing to deal with in regards to real estate options. It's quite possible that the W1 option will cost the most due to its vicinity to central London, while the SE3 and NW options might be more affordable due to them being in the outer edges.