IBM Data Science Capstone

Donut Venue Location Analysis

Report for final submission

James Cheung
2-27-2020

Contents

Introduction

Description and Background

While I was studying in university, my classmate, and friend, suggested we open a donut establishment in Glasgow city centre. Sadly, we both went our separate ways when we received offers for graduate programmes in our industry. The plan did not ever come to fruition; but I think this offers a brilliant opportunity to conduct some location analysis for a donut selling venue.

Glasgow is the most populous city in Scotland, and the third most populous city in the United Kingdom, as of the 2017 estimated city population of 621,020. The city has one of the highest densities of any locality in Scotland at 4023/km^2 with an average density of 3555/km^2. [1]

Location plays a big part to the success or failure of a donut shop in multitude of ways, from catching the attention of customers to daily foot-traffic.

Moreover, some factors which are linked to location may have a bigger impact on the final decision than others. For example, a great location must have affordable rent, no matter how much foot traffic it receives.

For this task, it is assumed that locations nearer the city centre will have a better visibility, higher foot-traffic and easy access. Of course, further market research would be required when considering a real-life location for donut shop.

Problem

Analysis will be conducted on the location of the current market of donut in Glasgow city centre, possibly to choose an optimal location where there are not already too many competing establishments. There will be a preference for locations nearest the Glasgow city centre area, after the first condition have been fulfilled.

This project will generate a map and suggest some promising locations based on market research.

Data acquisition and cleaning

Data sources

To display a visual map the Python library, Folium [2] was used to pull a tileset from providers such as OpenStreetMap, Mapbox, and Stamen. For boundary data for Glasgow, UK-GeoJSON [3] (maintained by martinjc) was used. This was divided into Output Areas. geojson.io [4] was used to find further information on each feature of the GeoJSON. Finally, the Foursquare API [5] was used to get the current venues in Glasgow.

Data cleaning

The boundary data provided by UK-GeoJSON is available in different divisions, these include Westminster Parliamentary Constituencies, Westminster Parliamentary Wards, Intermediate Data Zones, Data Zones, and Output Areas. The latter three are shown in figure 1, 2, and 3.

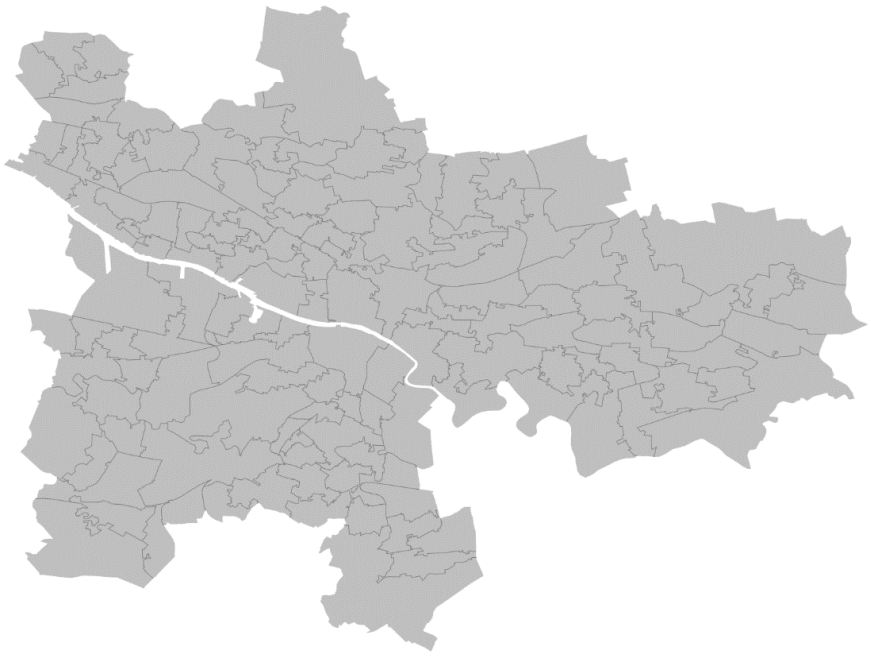

Figure 1 Intermediate Data Zones

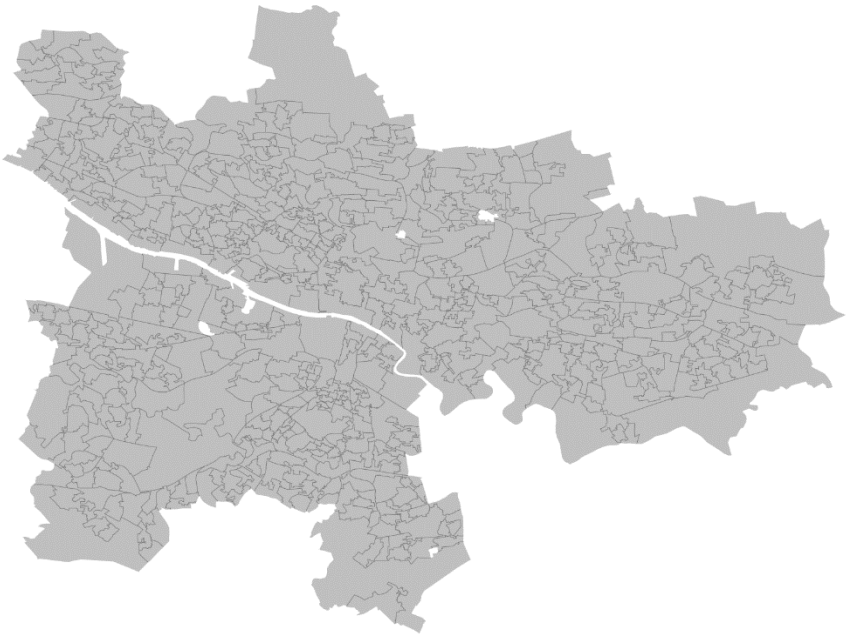

Figure 2 Data Zones

Figure 3 Output Areas

The resultant files were topoJSON files stored as JSON. Further processing was needed to acquire appropriate data.

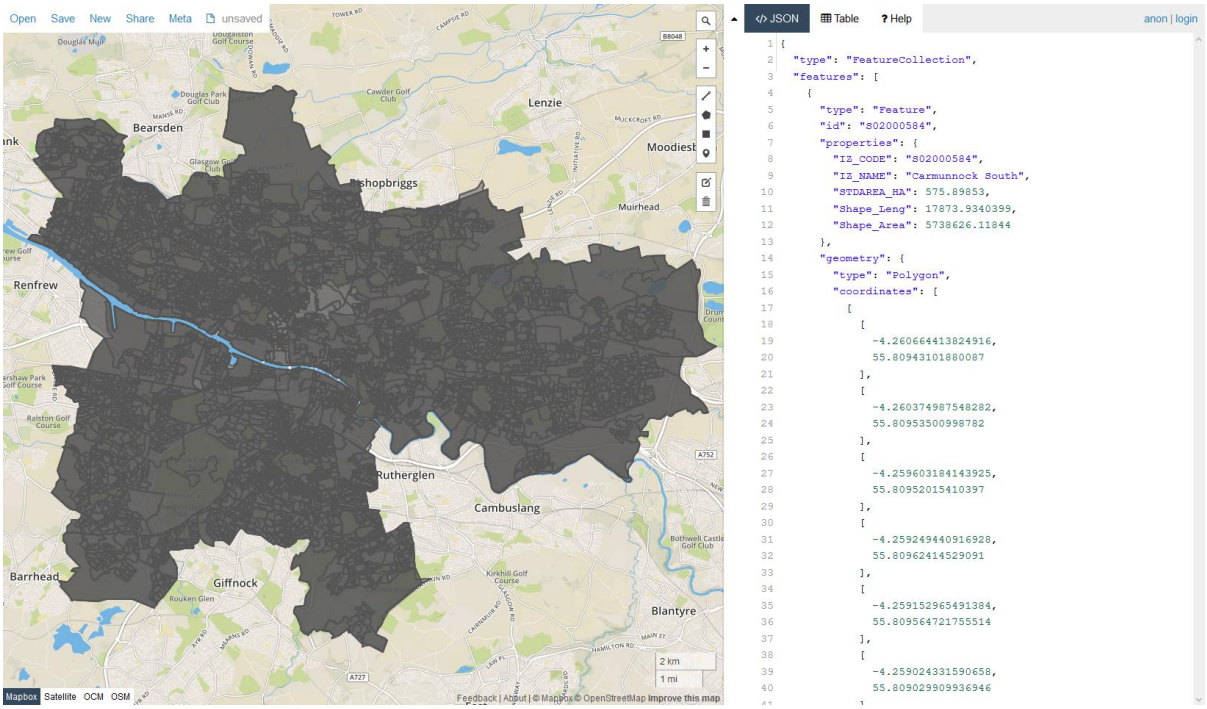

Figure 4 Screenshot of geojson.io interface with the Output Area data displayed.

The online tool geojson.io was used to process the JSON files. geojson.io processed the files and stored as geoJSON type; a format for encoding geographical data structures using the JavaScript Object Notation (JSON). All three sets of boundary data were preserved at this point.

All three datasets were divisions of the same area, Glasgow City. However, this project only requires the city centre. To reduce the data set and improve processing times, QGIS [6], a geographical data manipulation tool, was used to select postcodes beginning with G1 and G2 (feature masterpc in the GeoJSON). The final data was boundary data for just the Glasgow City Centre area and is illustrated in figure 5.

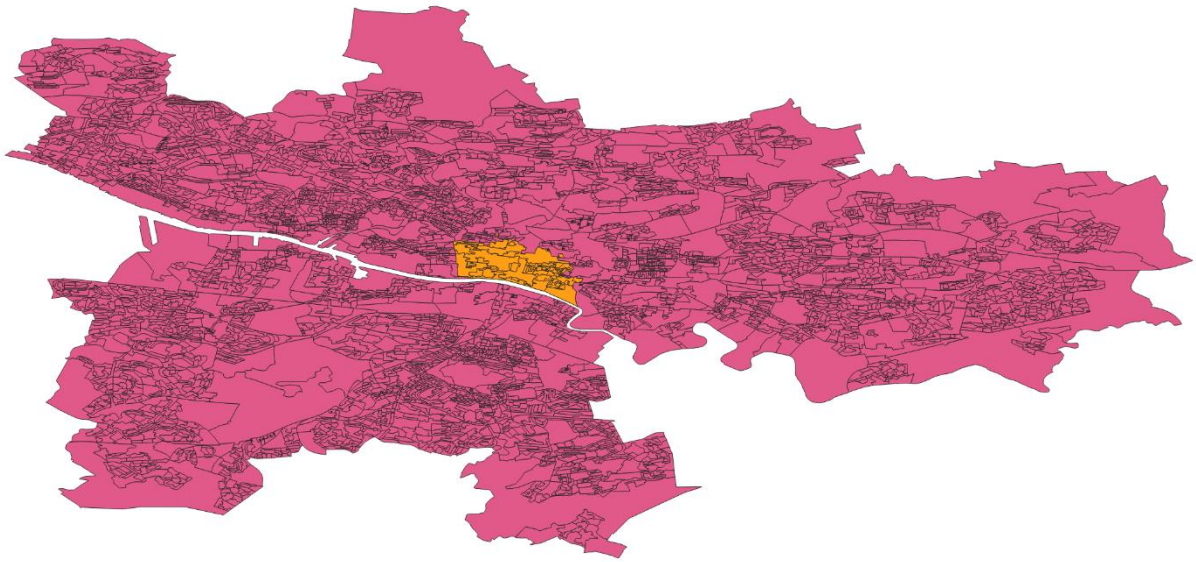

Figure 5 Geographical data manipulation tool QGIS with only G1 & G2 postcodes highlighted.

Methodology

Foursquare API Query

The main challenge when building the Foursquare API query is providing the right categories so the query returns specifically venues which would compete with this donut business. The categories can be found in the documentation [7]. Initially, all possible donut selling establishments were considered, this was found via the Foursquare app on Android using the search query “Donut Shop” in Glasgow. This returned donut shops, bakeries, cafes, sandwich shops as they all may contain donuts in their menu. After this query was made, the returned venues often did not sell donuts. When queried with only donut shops, the data set would be too small and did not include all venues which sold donuts. The final categories which were chosen were donut shops and bakeries as this provided the most accurate data.

Zoning

Boundary Data

To decide on the most appropriate division size, each of the GeoJSON datasets were displayed using the Python library Folium and visually judged based on the number on distribution across the city centre. The smallest division – Output Areas - was chosen to provide more specific results.

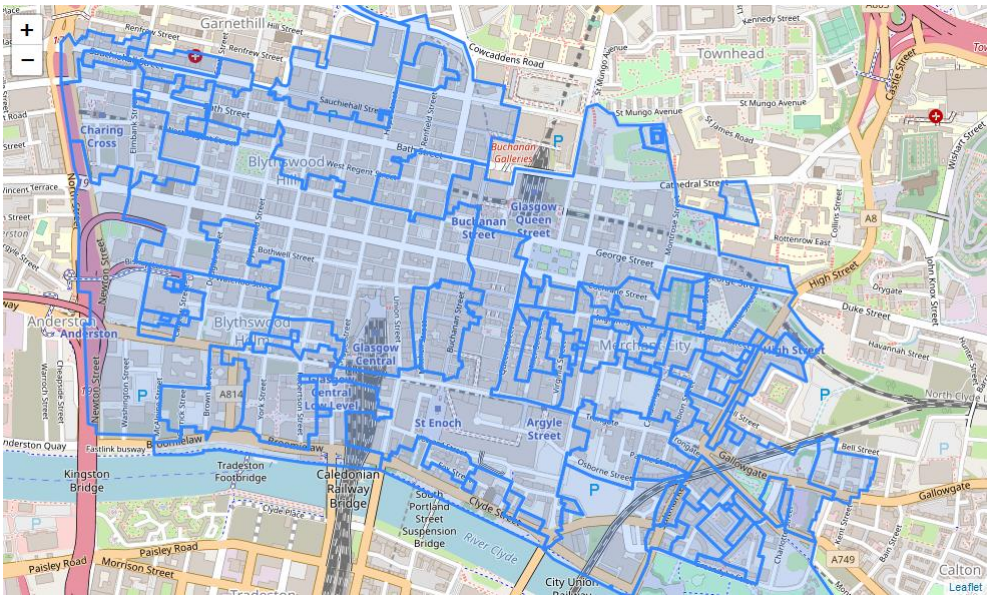

Figure 6 Output Areas rendered using Folium

The displayed GeoJSON is made up of features containing properties which provide coordinates of a polygon and additional properties including the postcode. The number of current competitors in each area is needed to proceed. However, the Foursquare API cannot take polygons as input, so the

simplest solution would be to convert each area from polygons to points by finding the centroid of each polygon. This was completed using QGIS and illustrated in figure 7.

Figure 7 QGIS map of centroids layer on top of the original city centre layer.

These centroids were then displayed on a map with Folium (figure 8).

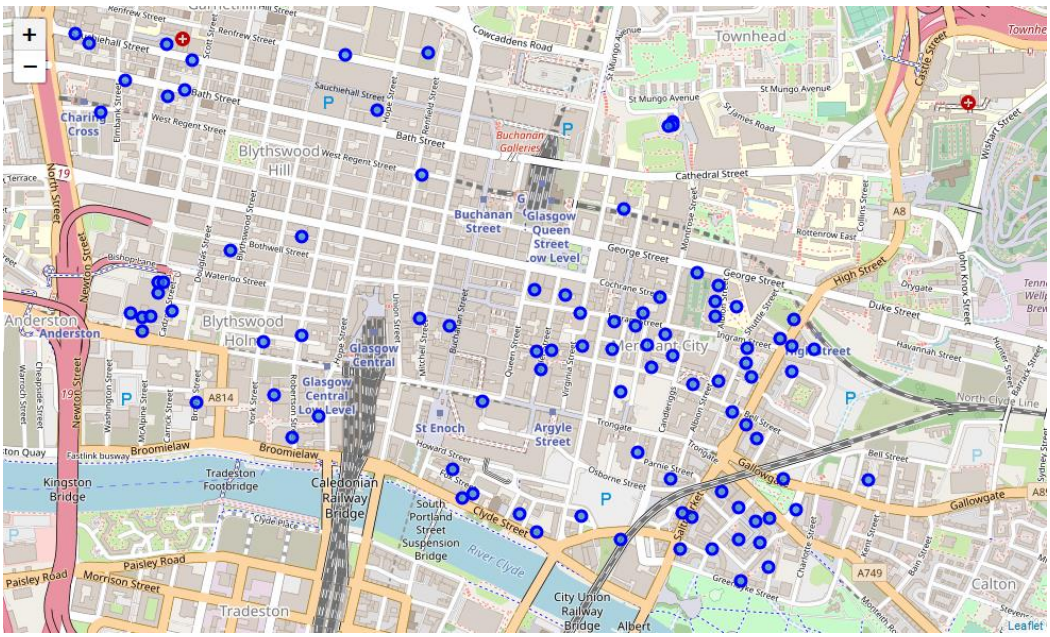

Figure 8 Centroids of Output Areas plotted on a map with Folium.

It is very apparent that these points are not evenly distributed across the city centre, with a higher density around the east and west sides. When the Foursquare API is used on these centroids a radius must be given which balances the difference of distribution. If the radius is too low, there is a risk of not retrieving venues between the search areas of points with sparse distribution; and if the radius is too high, this runs the risk of overlap of search areas which will return the same venue for two points. The radius value was experimented with to try and replicate the true map of venues.

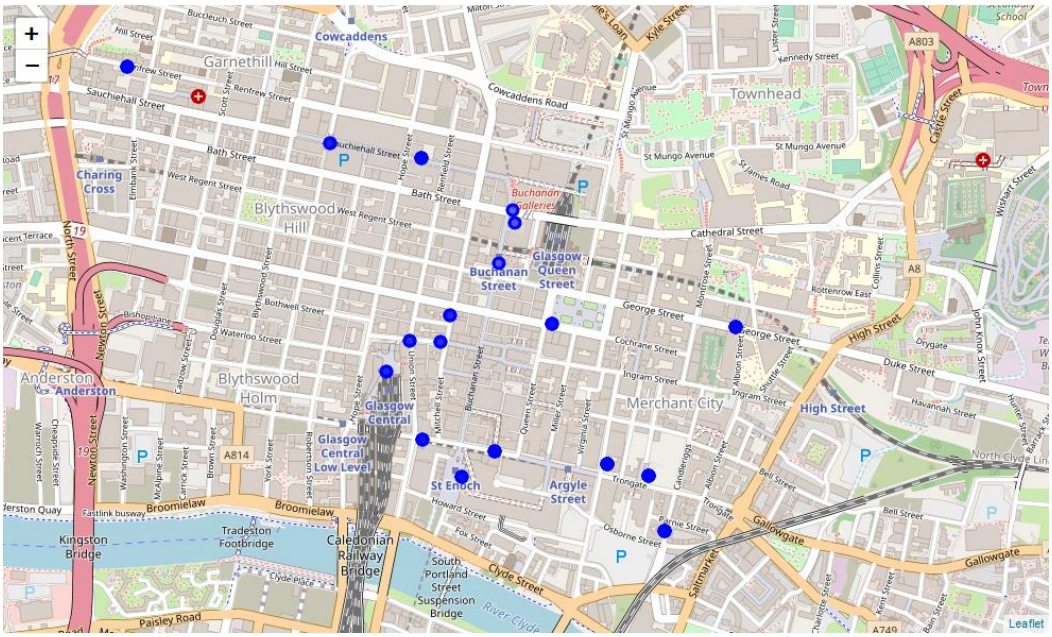

Figure 9 Venue map using centroids of Output Areas as input for Foursquare API with radius 200.

This was compared with the "true" map of venues across Glasgow City Centre, which was pulled using the Foursquare API using the centre coordinate of Glasgow City Centre as the point with radius 1500.

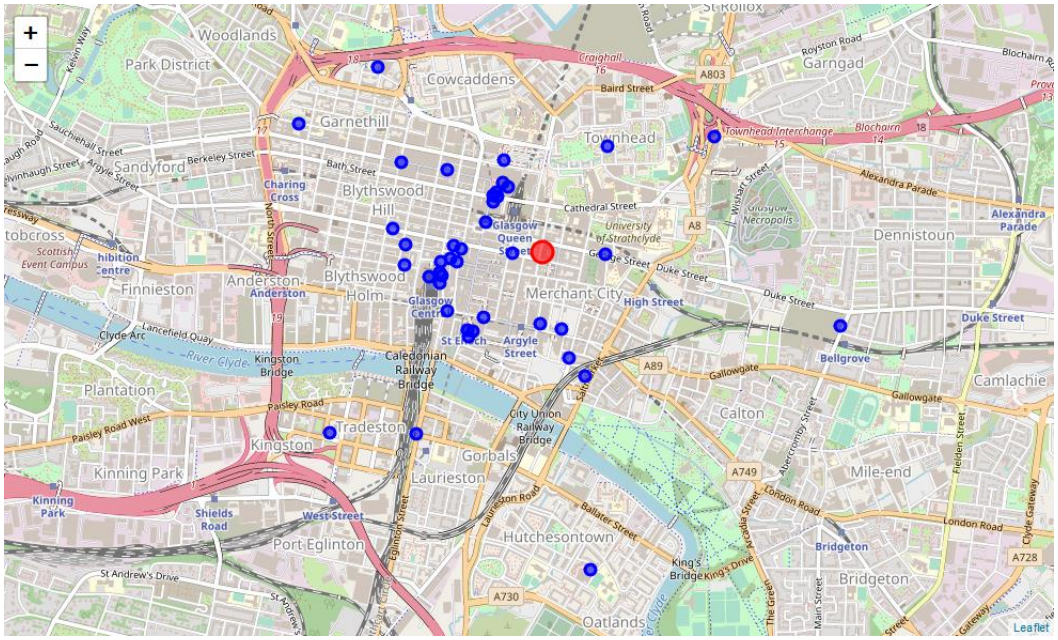

Figure 10 Existing competitor venues found using Foursquare API centred on Glasgow City Centre

Equidistant points

Another method to generate neighbourhood locations to search around was to use randomly allocated equidistant points.

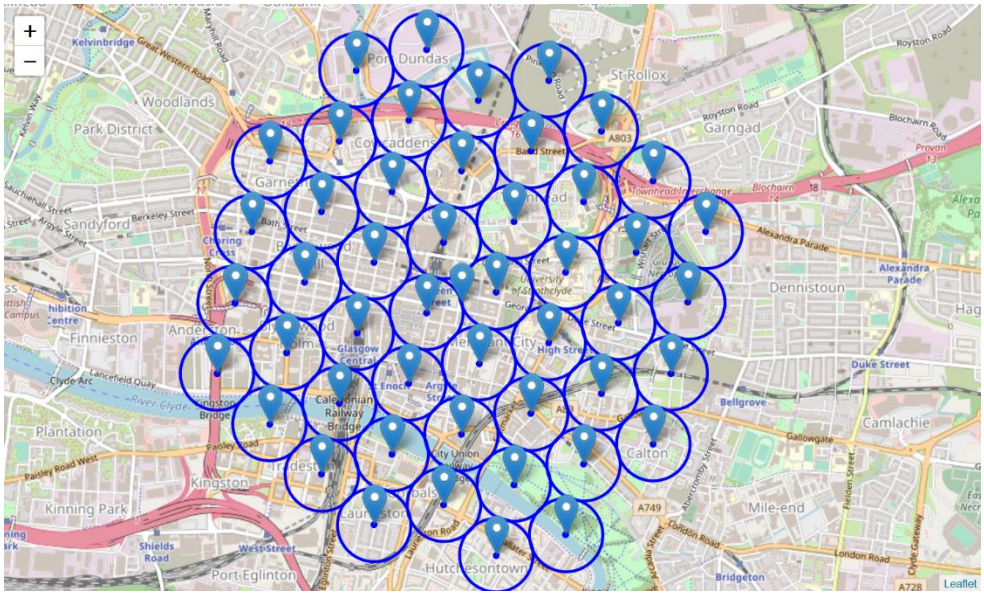

Figure 11 Randomly allocated equidistant points with radius 200.

This method roughly imitates a walking distance of 200m. If a customer is in a neighbourhood, they may only be attracted to go to venues within a short walking distance. Equidistant points also provided another advantage, as each point covers a circular area with radius 200m, a Foursquare API query with radius 200m would provide results very similar to the "true" map of venues.

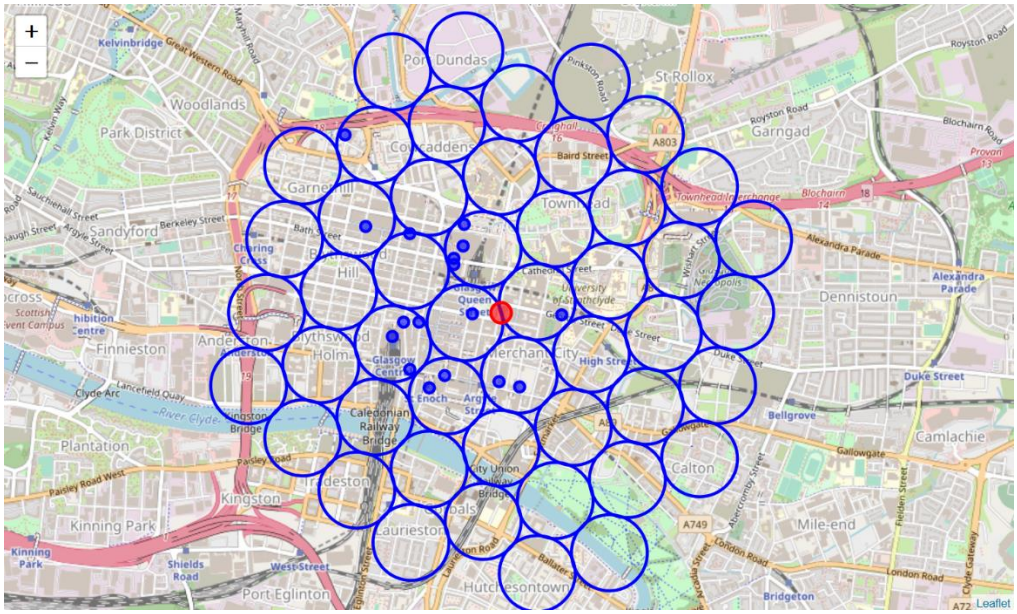

Figure 12 Donut/bakery venues found using equidistant points.

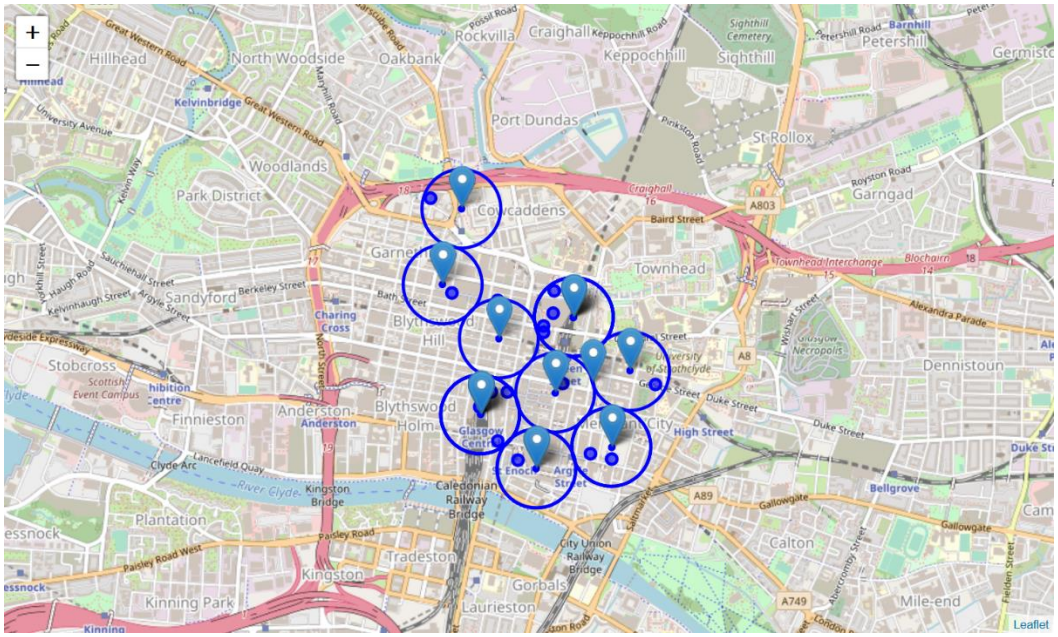

Figure 13 Neighbourhoods with venues within 200m radius.

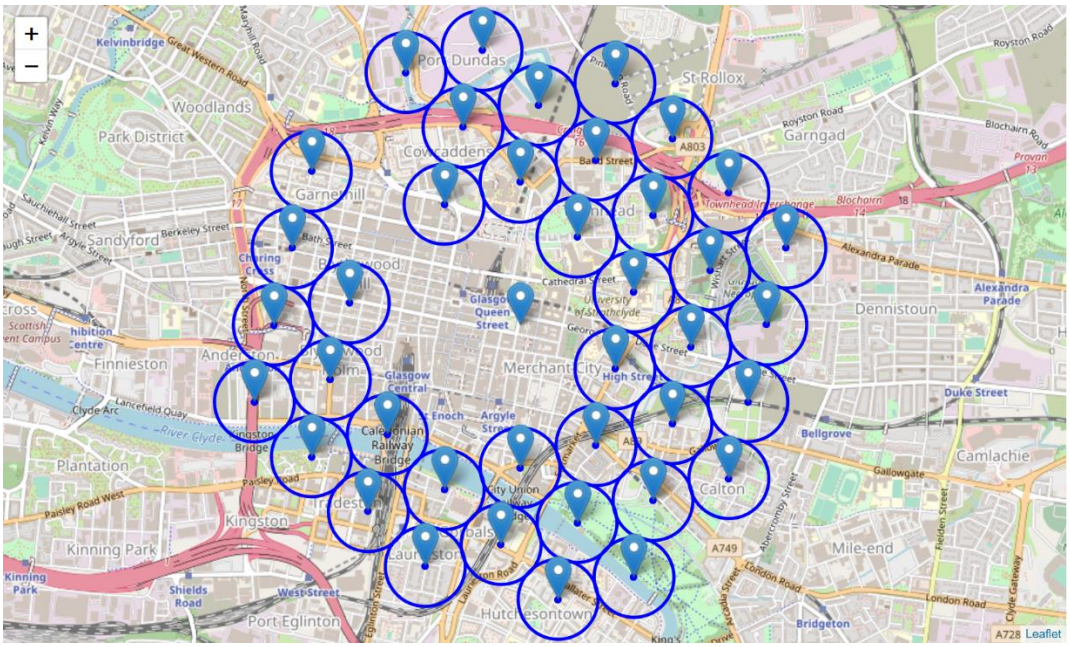

Figure 14 Empty neighbourhoods with no venues

By finding a map of all the empty neighbourhoods, the first condition was fulfilled. Figure 13 shows only the neighbourhoods with competitor venues and figure 14 displays only the empty neighbourhoods.

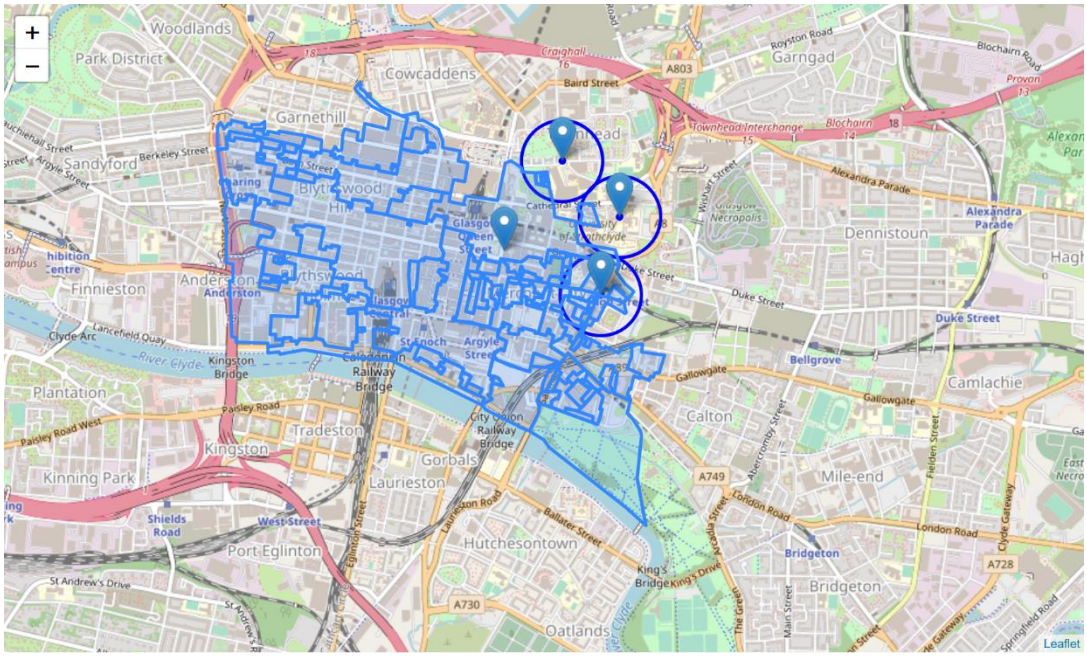

Figure 15 Three empty neighbourhoods closest to city centre.

The next condition was to find the neighbourhoods nearest the city centre. The distance of each neighbourhood from the city centre was found and sorted in ascending order. The top three neighbourhoods (three closest to city centre) are shown in figure 15.

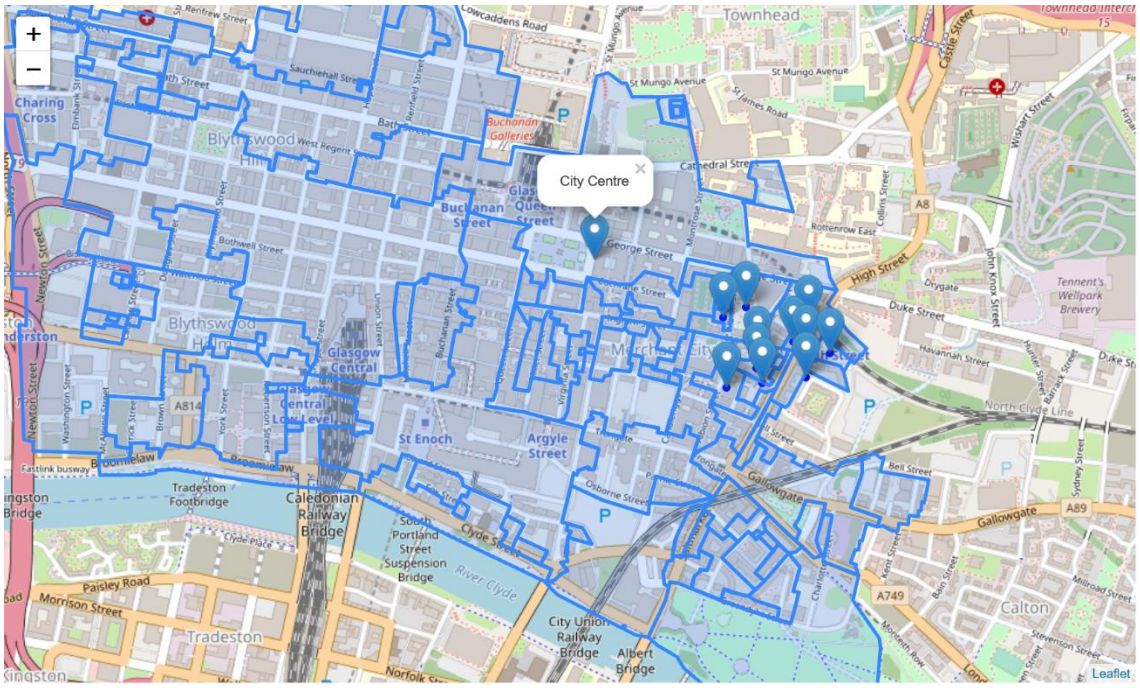

Figure 16 City centre neighbourhood with potential neighbourhoods marked.

From figure 15 it was clear only one potential neighbourhood was contained within the G1/G2 postcode category. All the postcode locations contained within this neighbourhood were found and plotted in figure 16.

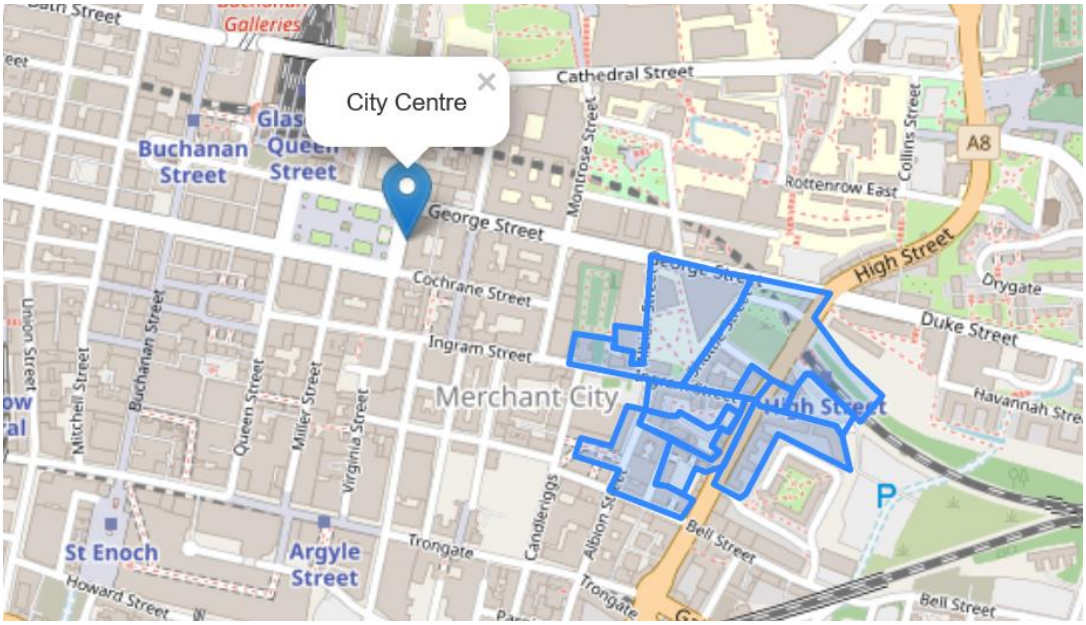

Figure 17 Potential postcode locations for new venue.

QGIS was used again to select the desired features using the boundary data; the results are in figure 17.

Results

The final map (figure 18) meets both conditions proposed in the problem section. This neighbourhood displays a neighbourhood with little competition and is also near the city centre.

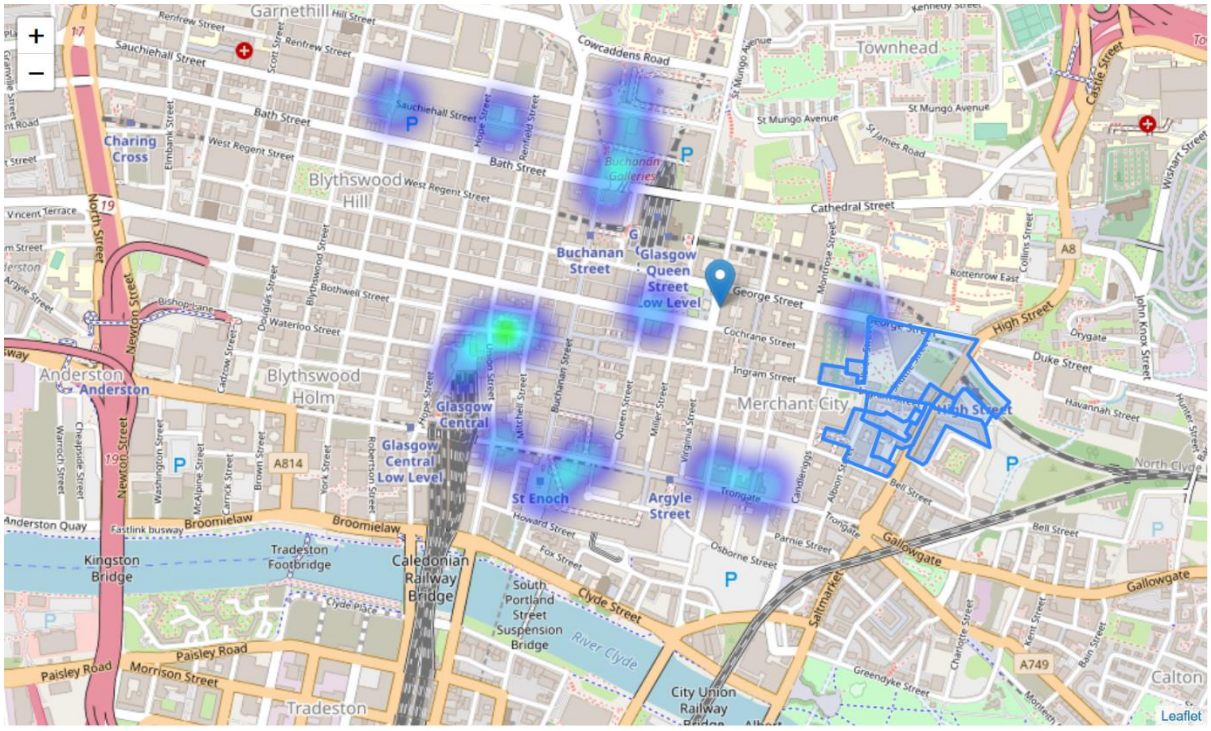

Figure 18 Potential postcode locations with a heatmap of the current venues.

Discussion

When it comes to choosing a venue for any business - even outside the donut business – there are many more factors to consider other than proximity to other similar businesses and distance from city centre. Arguably, factors such as foot-traffic and rent costs make a much larger impact on the success of a business than the two conditions set out in this project. Being further away from other similar businesses is not always an advantage as potential customers may be going to the neighbourhood of an already established venue and come to this new business if the original venue is closed or has a long queue. Rent costs vary widely even within the same neighbourhood, as donut venues do not require a kitchen, they can be contained within small units which may only be found in specific areas of the city. Foot traffic within Glasgow City Centre is heavily concentrated around Central Station and Queen Street Station, where commuters go to and from work every day. The proposed neighbourhood is outside this area. However, High Street Station is contained within the neighbourhood; which has quite significant foot traffic. One additional thing to consider is the geography of the city, the 200 radii neighbourhoods are a much better estimate of walking distance than centroid of boundary data, but this does not account for uphill ground, walking paths, and traffic. More complex data analysis of this problem would account for all these factors.

Conclusion

Analysis was conducted on the location of the current market of donut in Glasgow city centre, where an optimal location was found with little competition. The location is near the city centre. A map was generated from this data and further developments were discussed. The final map also shows a heatmap of the current venues to help visualise the density of venues. Further market research can be conducted with these results to find a suitable location for the new business.

References

[1] “Glasgow - Wikipedia,” 27 02 2020. [Online]. Available: https://en.wikipedia.org/wiki/Glasgow.

[2] “Folium 0.10.1 documentation,” Folium, 27 02 2020. [Online]. Available: https://python-visualization.github.io/folium/.

[3] martinjc, “UK-GeoJSON,” 27 02 2020. [Online]. Available: https://martinjc.github.io/UK-GeoJSON/.

[4] “geojson.io,” 27 02 2020. [Online]. Available: http://geojson.io/.

[5] Foursquare, “Foursquare Documentation,” Foursquare, 27 02 2020. [Online]. Available: https://developer.foursquare.com/docs.

[6] QGIS, “Welcome to the QGIS Project!,” QGIS, 27 02 2020. [Online]. Available: https://www.qgis.org/en/site/.

[7] Foursquare, “Foursquare Venue Categories,” Foursquare, 27 02 2020. [Online]. Available: https://developer.foursquare.com/docs/resources/categories.