Coursera Capstone Project – Applied Data Science

Gabriel Leow

Scope

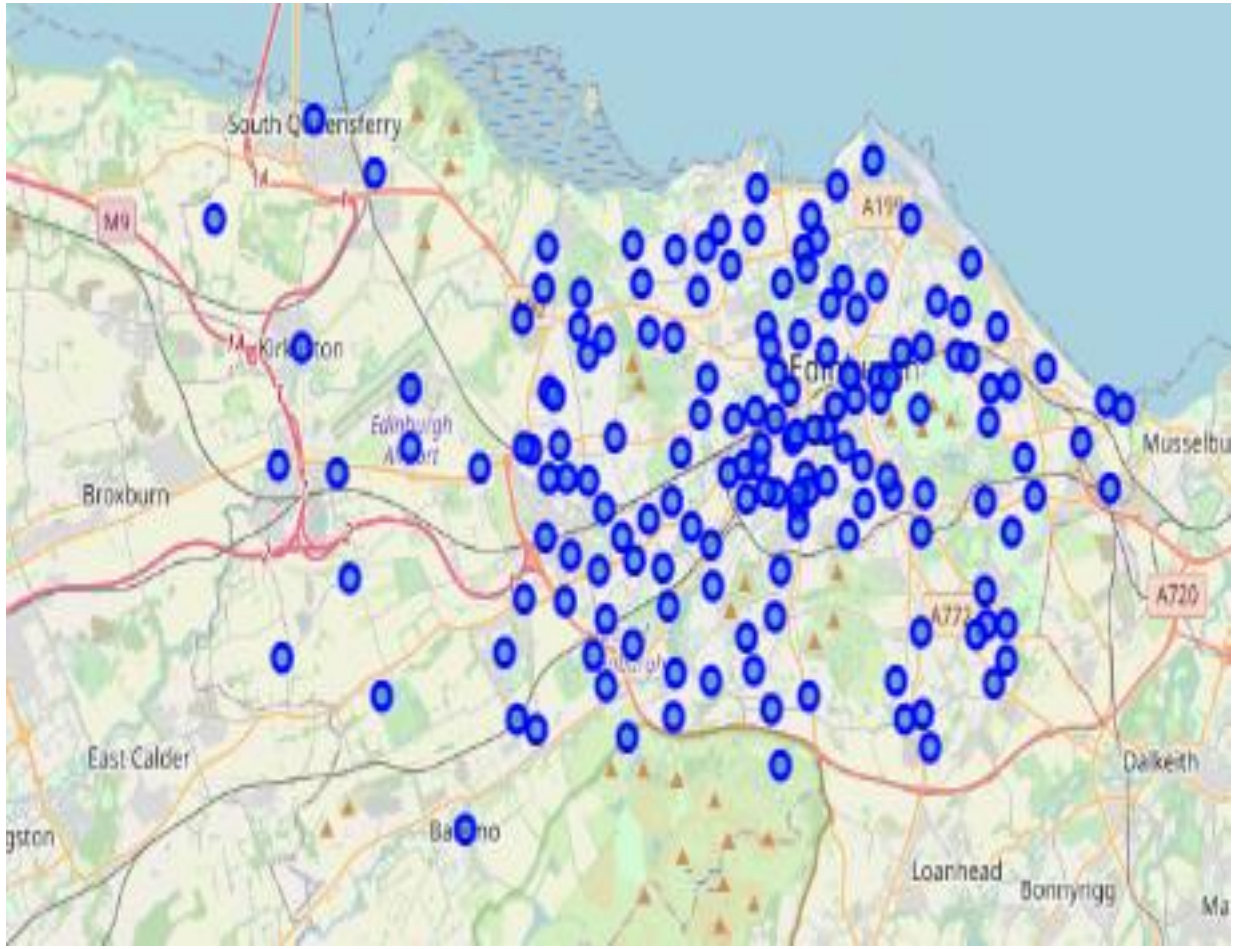
- Introduction and Business Problem
- Data – Neighbourhoods, Coordinates and Venue Data
- Methodology – Exploratory Data Analysis and K-Means Clustering
- Results and Discussion – Folium Visualisation and Cluster Analysis
- Conclusion

Introduction and Business Problem

- COVID-19 has imposed lockdowns around the world.
- While lockdowns may be lifted gradually, without a workable vaccine, human activity and human interaction is expected to remain low, with people staying at home.
- Grocery shopping is one of the few activities that are still considered essential, and people will want to leave their homes to do their grocery shopping, in order to maintain some semblance of normalcy.
- Supermarket Chains looking to expand their footprint in Edinburgh will want to find the prime location in order to maximise profitability.

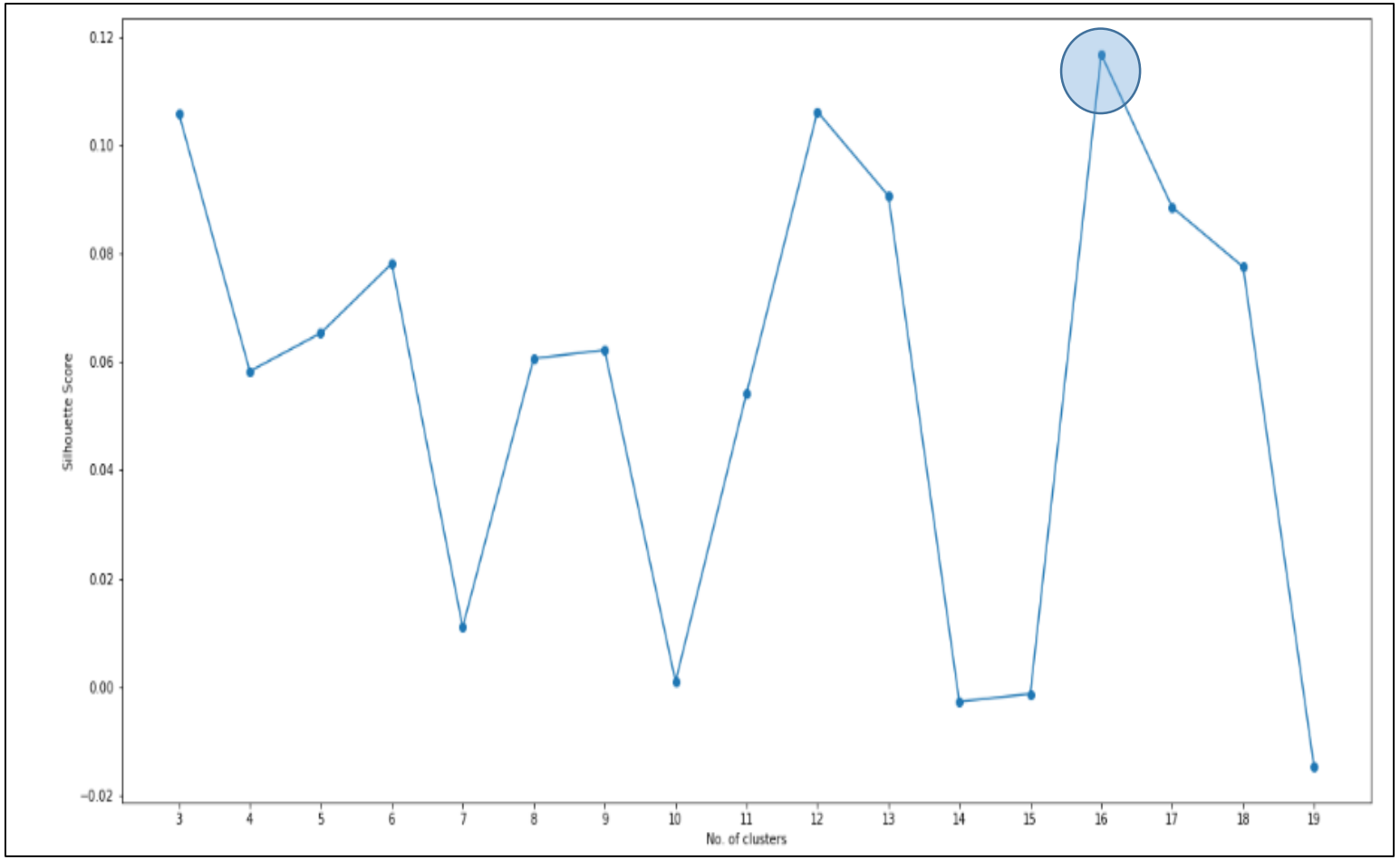
Data

- Our data sources are
  1. Wikipedia: List of Edinburgh Neighbourhoods
  2. Geocoder: GPS Coordinates of Edinburgh Neighbourhoods
  3. Foursquare: Venue data for each of the neighbourhoods.
- There are a total of 163 neighbourhoods in Edinburgh, 2 of which do not have venue data on Foursquare.
- Of the 161 neighbourhoods in Edinburgh, there are 5,245 unique venues and 248 unique venue categories.



Methodology – Use K-Means Clustering

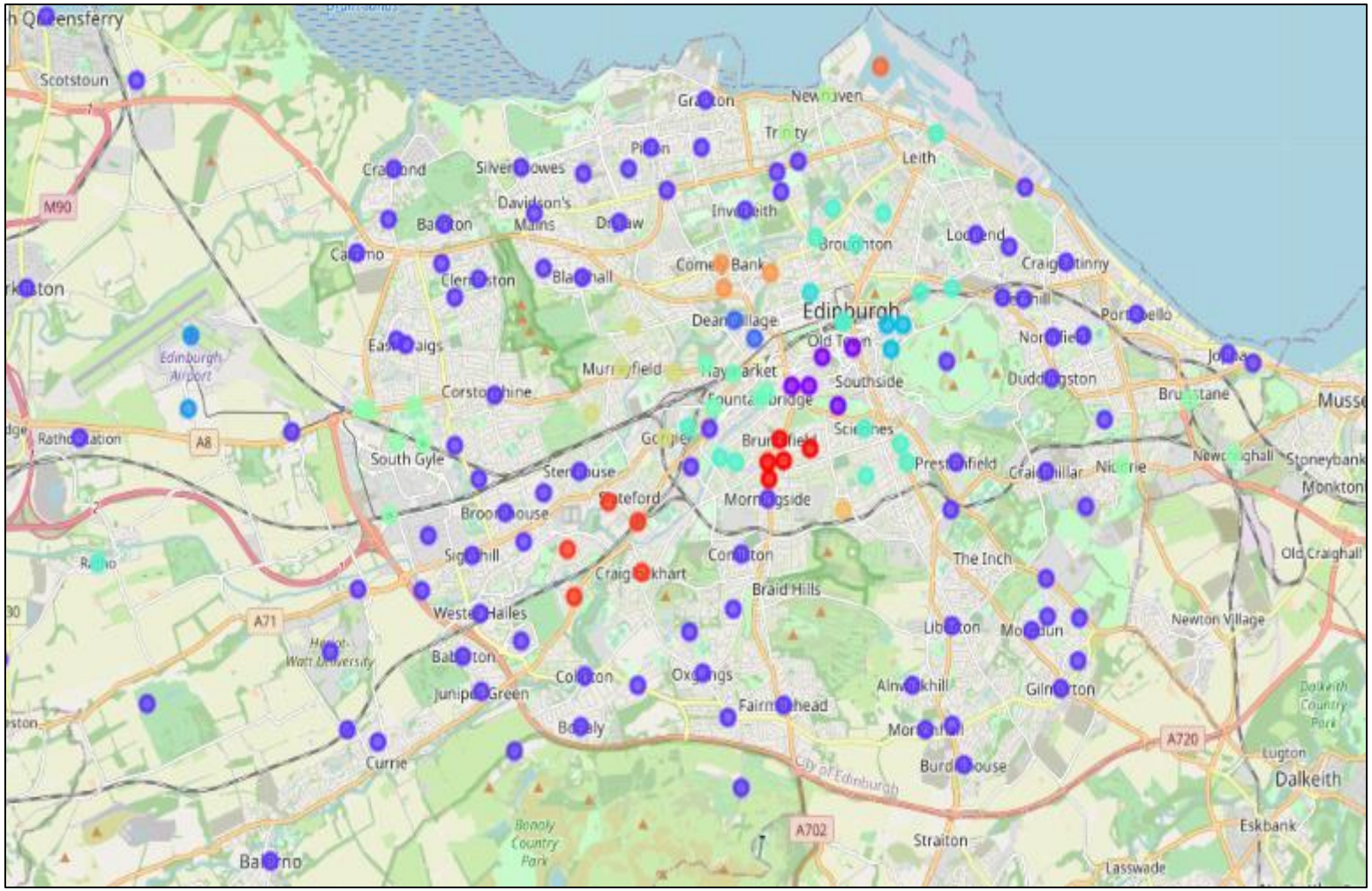
- Prepare Data for K-Means
  - One-Hot Encoding of Categorical Variables
  - Derive Top 10 Venue Categories for each Neighbourhood
  - For example: Abbeyhill's top 10 venue categories are – (1) Café; (2) Bar; (3) Grocery Store; (4) Park; (5) Bakery; (6) Hotel; (7) Pub; (8) Liquor Store; (9) Coffee Store; (10) Restaurant
  - Normalise Data
- Derive Optimal Number of Clusters using Silhouette Score
  - Optimal Number is **16** (see picture)



We then assign each neighbourhood a cluster number from 1 to 16.

Results and Discussion

- Each cluster is represented by a colour.
- Cluster 3 is the largest cluster with 94 neighbourhoods.



Results and Discussion

- By looking at the occurrence of Supermarkets and Grocery Stores in the top 10 venues categories of each clusters, we find that Cluster 11 and 16 have the highest occurrence of Supermarkets and Grocery Stores.

Cluster 11

```
[289]: C11 = edin_merged.loc[edin_merged['Cluster Labels'] == 10, edin_merged.columns[[0] + list(range(4, edin_merged.shape[1]))]]
C11
```

[289]:

	Neighbourhood	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
115	Newhaven, Edinburgh	Hotel	Café	Park	Supermarket	Grocery Store	Clothing Store	Food & Drink Shop	Sandwich Place	Bike Trail	Street Food Gathering
151	Trinity, Edinburgh	Pub	Park	Hotel	Café	Food & Drink Shop	Rugby Pitch	Bike Trail	Supermarket	Climbing Gym	Trail

Cluster 16

```
[266]: C16=edin_merged.loc[edin_merged['Cluster Labels'] == 15, edin_merged.columns[[0] + list(range(4, edin_merged.shape[1]))]]
C16
```

[266]:

	Neighbourhood	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
28	Chesser	Supermarket	Grocery Store	Gas Station	Auto Garage	Soccer Field	Bowling Alley	Fast Food Restaurant	Market	Tourist Information Center	Chinese Restaurant
41	Craiglockhart	Soccer Field	Stadium	Supermarket	Trail	Grocery Store	Gas Station	Tennis Court	Coffee Shop	Fast Food Restaurant	Market
91	Kingsknowe	Trail	Gas Station	Chinese Restaurant	Supermarket	Market	Train Station	Deli / Bodega	Department Store	Fountain	Forest
99	Longstone, Edinburgh	Supermarket	Museum	Fast Food Restaurant	Tourist Information Center	Market	Trail	Train Station	Chinese Restaurant	Grocery Store	Gym
143	Slateford	Grocery Store	Supermarket	Coffee Shop	Trail	Gym / Fitness Center	Bowling Alley	Soccer Field	Fast Food Restaurant	Nature Preserve	Gas Station

Results and Discussion

- Cluster 16 has a higher percentage of “grocery stores + supermarkets” versus Cluster 11.
- Supermarkets and Grocery Stores occupy the Top 3 places in Cluster 16, but 4^{th}-10^{th} places in Cluster 11.
- Cluster 16 is a more optimal cluster for the opening of a supermarket.

Cluster 16

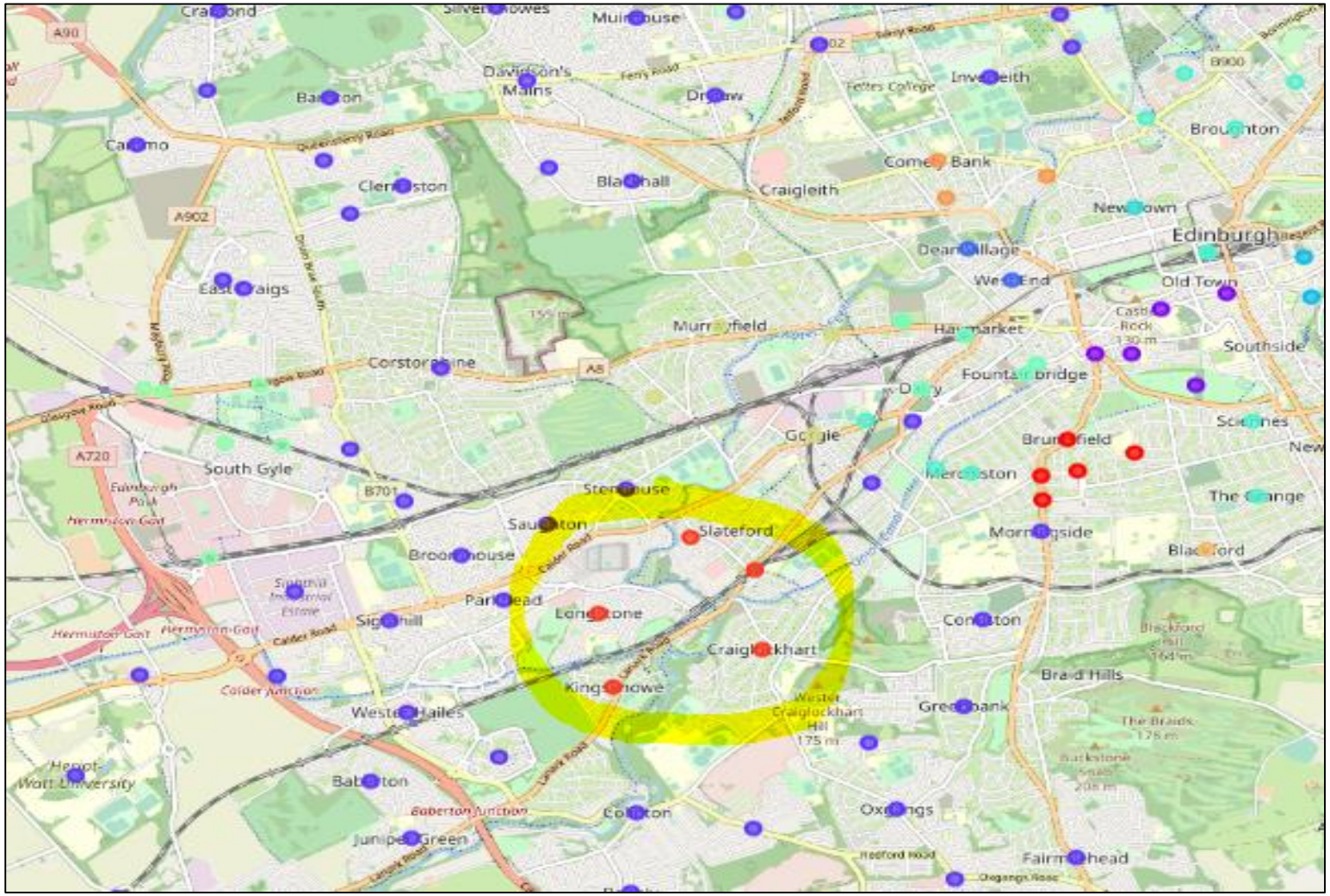
```
[266]: C16=edin_merged.loc[edin_merged['Cluster Labels'] == 15, edin_merged.columns[[0] + list(range(4, edin_merged.shape[1]))]]
C16
```

[266]:

	Neighbourhood	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
28	Chesser	Supermarket	Grocery Store	Gas Station	Auto Garage	Soccer Field	Bowling Alley	Fast Food Restaurant	Market	Tourist Information Center	Chinese Restaurant
41	Craiglockhart	Soccer Field	Stadium	Supermarket	Trail	Grocery Store	Gas Station	Tennis Court	Coffee Shop	Fast Food Restaurant	Market
91	Kingsknowe	Trail	Gas Station	Chinese Restaurant	Supermarket	Market	Train Station	Deli / Bodega	Department Store	Fountain	Forest
99	Longstone, Edinburgh	Supermarket	Museum	Fast Food Restaurant	Tourist Information Center	Market	Trail	Train Station	Chinese Restaurant	Grocery Store	Gym
143	Slateford	Grocery Store	Supermarket	Coffee Shop	Trail	Gym / Fitness Center	Bowling Alley	Soccer Field	Fast Food Restaurant	Nature Preserve	Gas Station

Results and Discussion

- One possible reason for Cluster 16 (yellow circle) is that it sits astride the main train track connecting Edinburgh westwards to Glasgow, which are the 2 biggest cities in Scotland.



Conclusion

- Given that COVID-19 is likely to persist, Supermarket chains looking to expand their footprint should look at opening their new outlet around Cluster 16 so as to maximise the potential for profitability.