APPLIED DATA SCIENCE CAPSTONE

CLUSTER ANALYSIS OF COMMUTER BELT LOCATIONS AROUND BRUSSELS, BELGIUM

Edward Christie

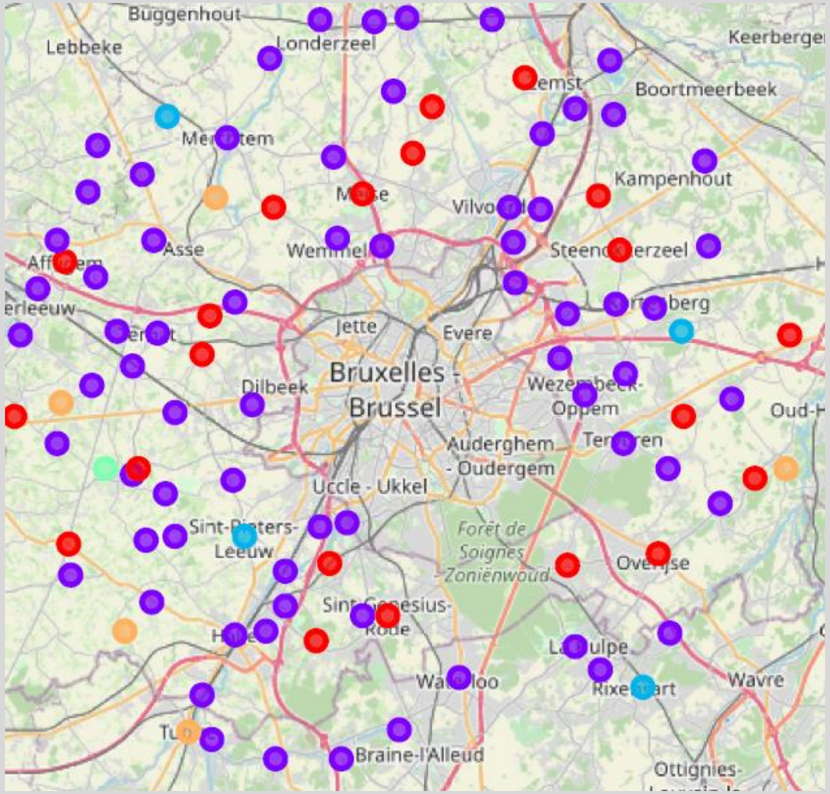
15 June 2020

Problem Statement and Approach

- **Illustrate how to leverage Foursquare data on commercial venues (such as shops and restaurants) using a K-Means clustering approach with real-world locations**
- Approach:
  - Formulate a business use case with a focus on specific cities or regions
  - Obtain geolocation data for the corresponding locations
  - Create a pandas data frame from the above
  - Make a corresponding venue data request from Foursquare
  - Integrate the venue data into the data frame
  - Carry out K-Means clustering
  - Display the results on a map
  - Discuss the results with respect to the intended business use case

Business Use Case: Brussels, Belgium

- **Geographical scope: commuter belt locations around the city** = at least X kilometres from the city centre, but not more than Y kilometres, from that centre
- Potential applications:
  - **Real Estate business**: help real estate agents and/or their clients to get a snapshot idea of the nature of various locations around a city centre
  - **Market analysis**: learn more about locations and how they differ; develop a basis for market gap analyses



Data Sources

- **Source 1: geolocation data for Belgium from open sources**
- From: github user jief
- Full csv format table covering all postal codes (zip codes) in the country – with location name and geographical coordinates
- Once the final choice of locations is made (based on selecting the commuter belt only), the geolocation data is passed in a request to Source 2

- **Source 2: data on commercial venues by location from Foursquare**
- As studied in the course, a request is passed to Foursquare using the free account credentials
- Several iterations were attempted in order to establish what subsets of venue categories would have enough data points for further analysis, while also enabling a certain thematic focus. Final choice: food and drink

Methodology

- **Step 1: apply Haversine formula to compute distances between locations**
- Many implementations in Python available
- Applied to original geolocation data frame to keep only locations that are between 7 km and 20 km from Brussels city centre
- This data frame is then passed into the Foursquare request → **get venue data**

- **Step 2: K-Means clustering and search for optimal number of clusters using ‘Elbow Graphs’**
- Focus only on food- and drink-related venues → 1084 venues, 68 venue categories, in 114 postal codes
- Locations with 5 venues or less are dropped → 1006 venues, 68 categories, in 80 postal codes
- Market segment variables also tested
- K-Means: see **Results I** slide

Market segment variables

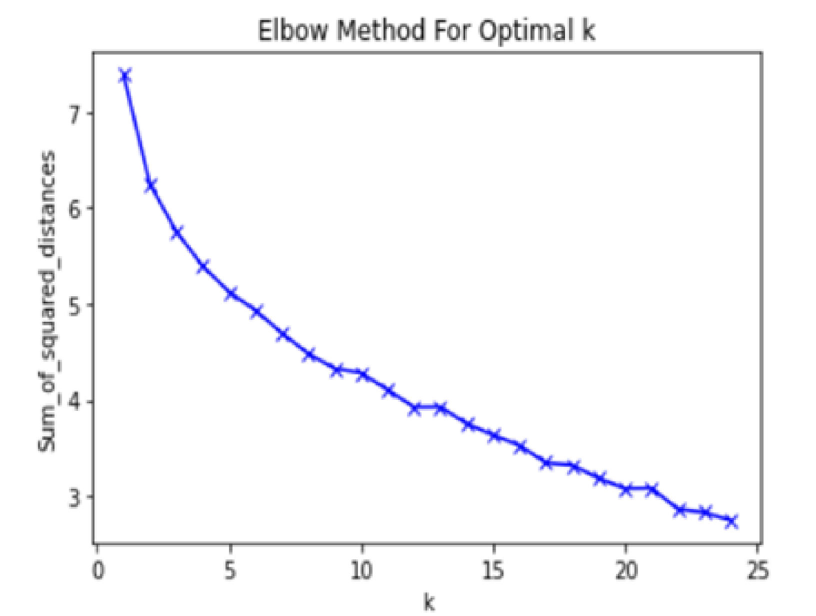
Market Segment	Venue Categories
	French Restaurant
	Cheese Shop
	Gourmet Shop
‘Status’	Wine Bar
	Gastropub
	Wine Shop
	Whisky Bar
	Snack Place
	Fast Food Restaurant
‘Budget’	Pizza Place
	Sports Bar
	Cafeteria
‘Neutral’	All other venue categories

	Municipality	African Restaurant	Argentinian Restaurant	Asian Restaurant	Bakery	Bar	B B
0	Affligem	0.0	0.0000	0.000000	0.000000	0.250000	0.
1	Asse	0.0	0.0000	0.000000	0.142857	0.142857	0.
2	Baardegem	0.0	0.0000	0.000000	0.000000	0.333333	0.
3	Beersel	0.0	0.0625	0.000000	0.125000	0.187500	0.
4	Beigem	0.0	0.0000	0.000000	0.000000	0.222222	0.
5	Bertem	0.0	0.0000	0.000000	0.000000	0.444444	0.
6	Borchtlombeek	0.0	0.0000	0.000000	0.142857	0.285714	0.
7	Braine-L'alleud	0.0	0.0000	0.166667	0.000000	0.166667	0.

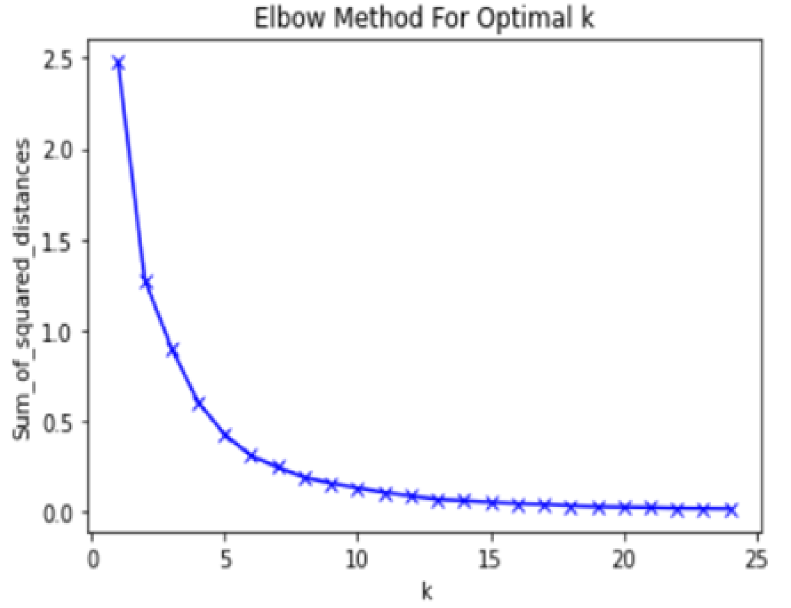
	Municipality	Status	Budget	Neutral
0	Affligem	0.000000	0.250000	0.750000
1	Asse	0.000000	0.000000	1.000000
2	Baardegem	0.000000	0.000000	1.000000
3	Beersel	0.062500	0.062500	0.875000
4	Beigem	0.111111	0.111111	0.777778
5	Bertem	0.111111	0.111111	0.777778
6	Borchtlombeek	0.000000	0.000000	1.000000
7	Braine-L'alleud	0.000000	0.000000	1.000000

Results I

- Approach 1: full data frame
- (1006 venues x 68 venue categories)



- Approach 2: reduced data frame
- (1006 venues x 3 market variables)

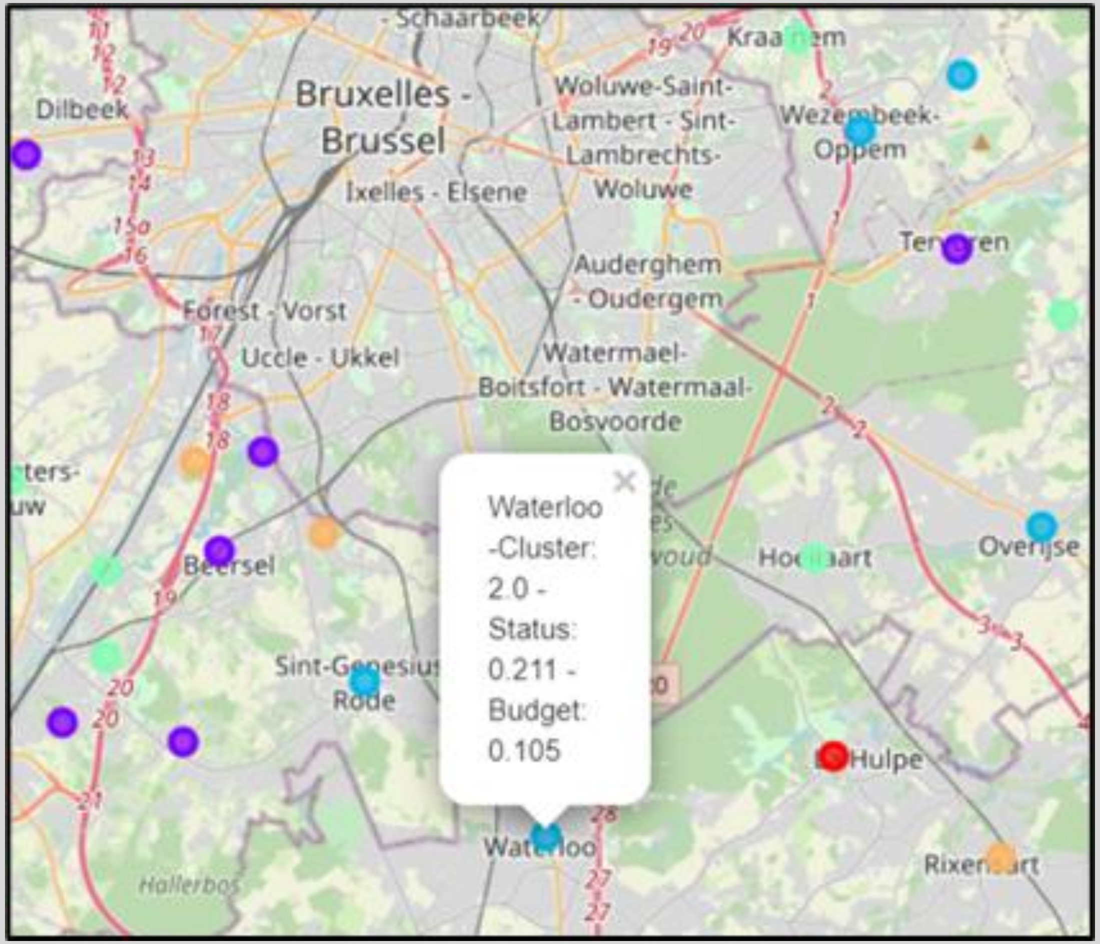


Results II

- Final choice: Approach 2 (market segment variables, not raw venue category data) – and then K-Means with k=5 clusters
- Visual inspection and of local knowledge → interpretation of the 5 resulting clusters

Cluster No	Tendency	Examples
0	Exceptionally high share of ‘Status’ venues	La Hulpe
1	High share of ‘Budget’ venues; no ‘Status’ venues	Tubize; Buizingen
2	High share of ‘Status’ venues, lower share of ‘Budget’ venues	Waterloo; Braine-le-Chateau
3	Low shares of both ‘Status’ and ‘Budget’ venues, including many with 0% under both	Braine-L’alleud; Halle
4	High share of ‘Budget’ venues, lower share of ‘Status’ venues	Rixensart

Presentation of results with Folium



As seen in the course

In addition, I edited the code in order to also display the scores (proportion of venues) that belong to the upmarket ('Status') and downmarket ('Budget') market segments

Discussion

- The main technical lesson from this assignment is that K-Means clustering can have certain limitations
- It is very easy to implement, but it may fail to reveal a clearly optimal number of clusters

- It may be better to combine K-Means with additional insights about the data (unless K-Means gives very compelling results from the start)
- In this case study, I used judgment and local knowledge to strongly reduce dimensionality and focus the analysis on a upmarket / down-market research question

Conclusions

- With some improvements, the approach initiated in this assignment could be further developed – e.g. a visualisation tool for real estate agencies
- A different direction to be explored could be to leverage the full extent of the Foursquare data to identify market gaps – e.g. use K-Means to detect outliers, as compared to their respective centroids – and see what venues might be “missing” from such locations