IBM Data Science Professional certification

CAPSTONE PROJECT:

Investing in Leuven’s neighborhoods

Eugene Ambe Ndamukong

April 12, 2020

CONTENTS

1 Introduction

As an international resident in Leuven, Belgium since 2013, I have witnessed the huge potential of the city as a business hub. The city, over the years, has been a huge attraction to international students who are registered in universities in Leuven and also to tourists who come to witness the ruins of world war 2 and other historic sites. All these large in-flock of people into the city also bring in money into local businesses like restaurants, cinemas, fast foods etc. However, these varying businesses are found in different neighborhoods. Some neighborhoods have more of a particular business than others. The distribution of different businesses in respective neighborhoods could serve a great interest to investors who are new to the city. The problem that this project aims to solve is to facilitate the investors' choice of where (neighborhood) to invest in and which type of business could do well in that neighborhood.

2 Data description

To solve the mentioned problem, geolocation data on all the businesses was obtained. Foursquare was the location data provider of choice from which the data for this analysis was retrieved. A total of 100 neighborhoods were randomly retrieved from the database such as Dijleterrassen, Dagelijkse Kost etc within a 2500m radius. Out of the 100 neighborhoods, there was a total of 130 business categories currently active in Leuven as seen in figure 1. The most occurring business categories as seen in the figure are Italian restaurant, Coffee shop, Burger joint etc. The frequency of the categories is proportional to the size of the words.

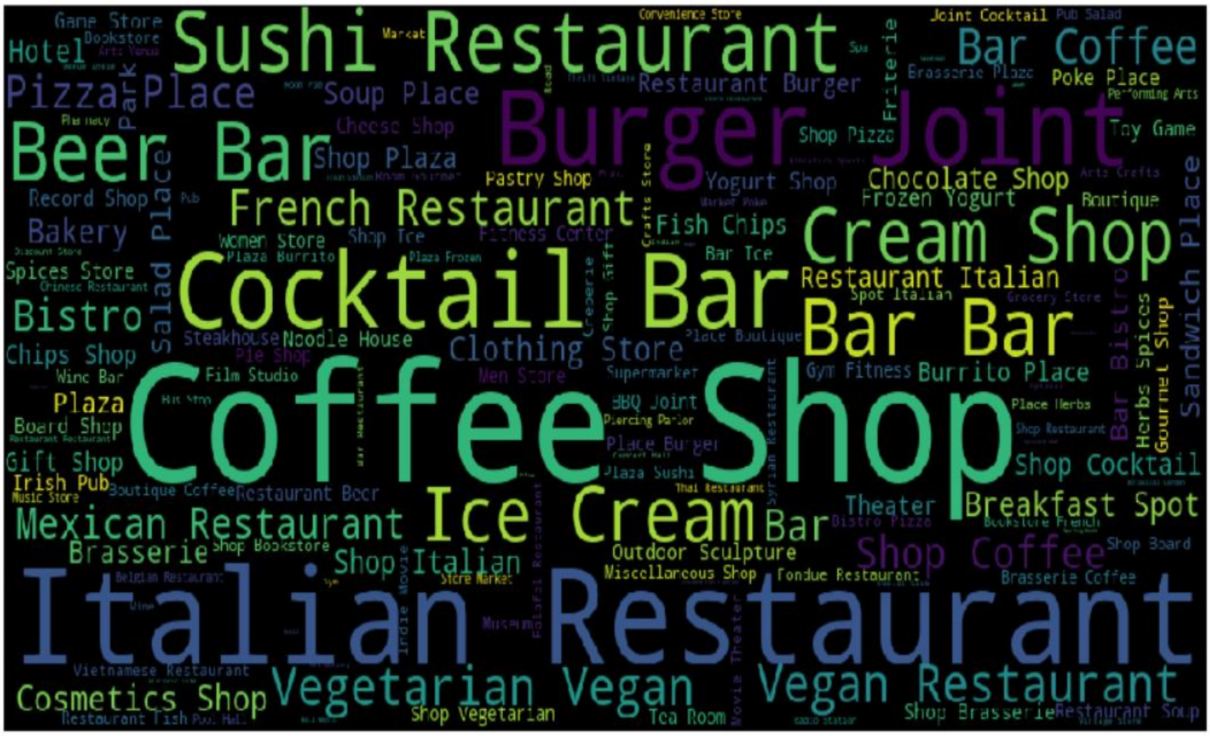

Figure 1: **Display of businesses categories which are proportional to their respective frequencies**

The geolocation of the different businesses was also available as latitude and longitude. These geolocations will be used to visualize the clustered neighborhoods on the geographical map of Leuven.

3 Methodology

3.1 Data preprocessing

The data which was initially obtained from Foursquare was received in a JSON format. Therefore, the information which was important to solve the business problem was extracted from the JSON file and saved into a dataframe which was easy for use. The information on neighborhoods extracted were *Neighborhood name, Neighborhood Latitude, Neighborhood Longitude, Business name, Business Latitude, Business Longitude, Business Category.* 130 dummy variables were created using binary coding out of the *business category*. The data was then grouped by neighborhoods and their means per neighborhood taken. The latter was then used for the clustering process.

3.2 K-Means clustering

K-means clustering was used to group the similar neighborhoods into clusters using the variables created in the pre-processing step. To decide how many centers for the clustering to use, several K-means models were created with 1 to 15 centers and their respective within sum of squares errors (SSE) were plotted against their number of centers (clusters). This can be seen in figure 2. The elbow point was not very distinct. However, this plot seem to show a great drop in the SSE from the 1 cluster to 5 clusters. After 5 clusters, the SSE seemingly didn't drop that much compared to the previous. Therefore, a K-means clustering model specified with 5 centers was implemented.

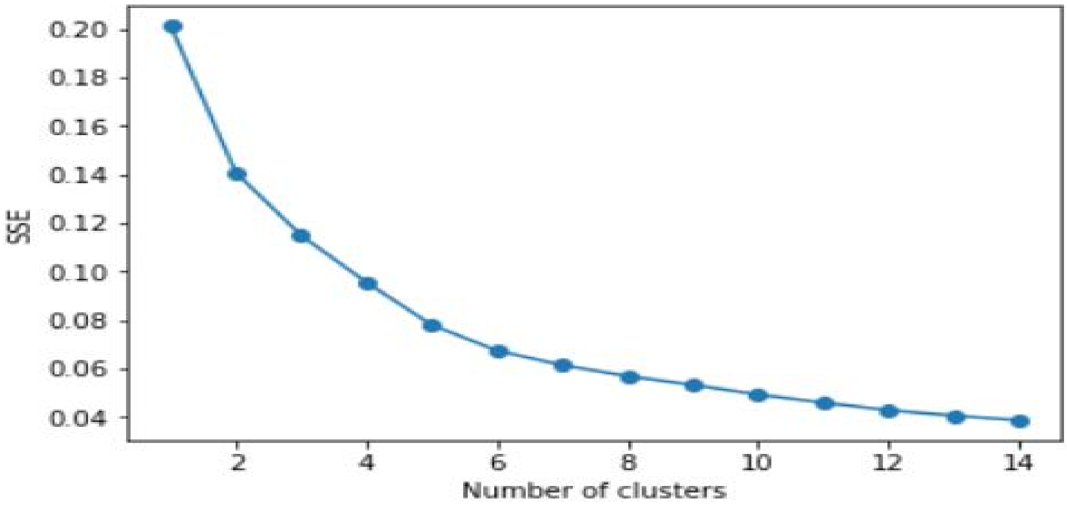

Figure 2: Within sum of square error(SSE) versus number of clusters

After the clustering process was done, the clusters were characterised by the 130 features. Out of the 130 features, some were quite influential in the creation of the different clusters and used in characterising the clusters. The features which were used in characterising the clusters were those with cluster centers (feature means) greater than the overall center or mean. Also, it should be noted that the overall center or mean is the mean of the whole data set used.

3.3 STATISTICAL TEST

After obtaining the clusters, it was necessary to test how different the cluster were. Multivariate analysis of variance (MANOVA) tests were used to test if the clusters were the same or if their respective means were the same. However, the dataset is multidimensional (more columns than rows) which could lead to matrix inversion problems while trying to perform a MANOVA test. For this reason, the dimensions were reduced using principal component analysis (PCA). The number of principal components to be used for the MANOVA test were obtained from the following figure 3 using the "elbow joint" method. Using the "elbow joint" method, 4 principal components were used as new dependent variables in the MANOVA test.

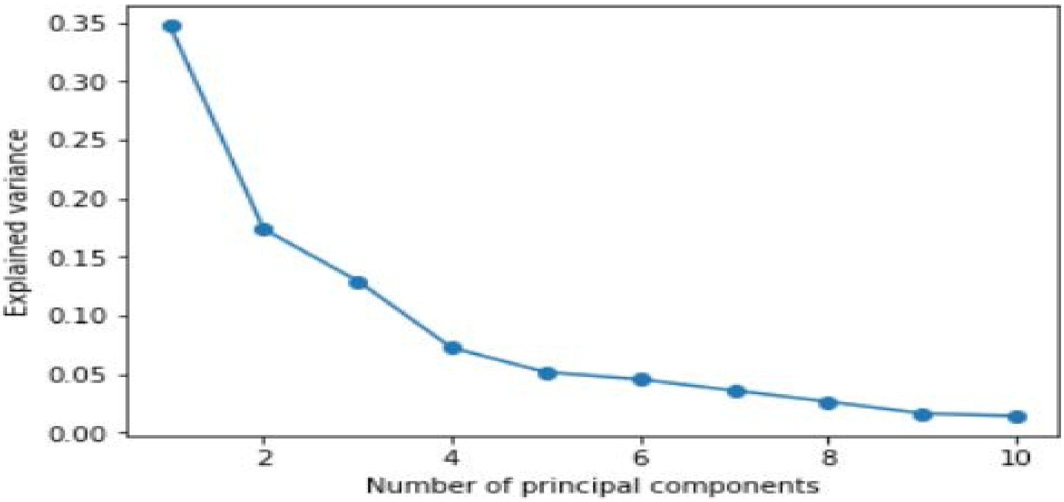

Figure 3: Explained variance versus number of principal components

4 RESULTS AND DISCUSSION

All 100 neighborhoods where grouped into the 5 clusters as seen in table 1. Cluster 4 had the most neighborhoods and cluster 1 had the least. As mentioned before, these number of neighborhoods are not the absolute number of neighborhoods in Leuven.

	Number of neighborhoods
Cluster 1	4
Cluster 2	25
Cluster 3	19
Cluster 4	46
Cluster 5	6

Table 1: **Distribution of neighborhoods across clusters**

The city of Leuven is a circular city defined by a "ring" road as seen in figure 4. This geometric circular structure of the city serves as a huge attraction to tourists. Furthermore, the city is also characterised by vegetation areas (green zones), streams (blue curvy lines) ,road systems (white cross lines) etc. This map was created using geolocation coordinates i.e latitude and longitudes of the neighborhoods. Clusters 1 (Red), 3 (Blue) and 5 (Orange) seem to more or less surround clusters 2 (Violet) and 4 (Light green). Clusters 2 and 4 are found at the centre of the city.

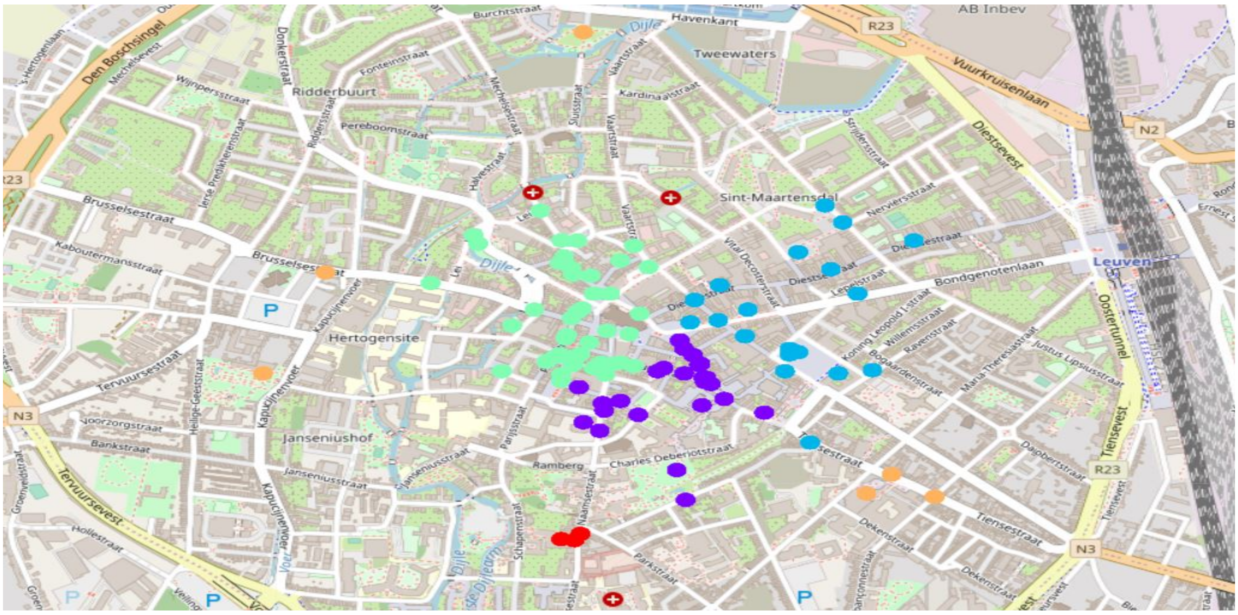

Figure 4: **The clustered neighborhoods have been colored: Cluster 1(Red), Cluster 2 (Violet), Cluster 3 (Blue), Cluster 4 (Light green), Cluster 5 (Orange)**

The clusters mentioned before were highly influenced by different features or variables (venue categories). Some features are highly influential than others. This influence was based on the cluster mean values per feature. The following table 2 shows the top 10 most influential features per cluster respectively. The top features shown here have been ordered in descending order i.e highest to lowest. All features can't be reported here since each cluster has over 30 features but they all can be found in the Jupyter notebook. Out of the top 10 most influential features, cluster 1 had a total of 34 features; cluster 2 had a total of 43 features; cluster 3 had a total of 55 features; cluster 4 had a total of 47 features and cluster 5 had a total of 68 features.

Clusters	**Top 10 features**
Cluster 1	Poke Place, Pharmacy, Irish Pub, Vietnamese Restaurant, Breakfast Spot, Chocolate Shop, Pool Hall, Theater, Noodle House
Cluster 2	Beer Bar, Poke Place, Breakfast Spot, Burrito Place, Board Shop, Restaurant, Frozen Yogurt Shop, Bookstore, Vietnamese Restaurant
Cluster 3	Clothing Store, Beer Bar, Cocktail Bar, Fondue Restaurant, Breakfast Spot, Poke Place, Belgian Restaurant, Restaurant, Frozen Yogurt Shop
Cluster 4	Beer Bar, Breakfast Spot, Gourmet Shop, Board Shop, Bookstore, Burrito Place, Pastry Shop, Miscellaneous Shop, Fish & Chips Shop
Cluster 5	Grocery Store, Pub, Indian Restaurant, Sushi Restaurant, Print Shop, Pool Hall, Lawyer, Laundromat, Electronics Store

Table 2: **Summary of top most influential features per cluster**

To make the characterisation of the clusters much easier, the business categories will be further grouped into groups such as Eating & drinking (e.g all restaurants and drinking spots), shopping (clothes, groceries etc), entertainment & fun (pool, cinemas etc) etc.

Cluster 1 consist of neighborhoods highly influenced eating spots, drinking spots, entertainment, shopping and health. The top eating categories are poke places, Vietnamese restaurant, breakfast spot , noodle house. Furthermore, in terms of health, pharmacy is seen to be also quite influential in this cluster of neighborhoods. Also, in terms of entertainment & fun, these neighborhoods are highly impacted by poll halls and theaters. Shopping on chocolate is also seen to be another highly influential category that contributes to the similarities among the neighborhoods.

Neighborhoods in cluster 2 are highly impacted by eating spots (poke place, breakfast spot, burrito place, restaurant), drinking spot (beer bar) and shopping (board, bookstore and frozen yoghurt). Unlike the first cluster, this cluster does not have entertainment and health among it's top 10 influential businesses.

Cluster 3's neighborhoods highly influential by shopping(clothing and frozen yogurt), drinking businesses (beer bar and cocktail bar) and eating businesses (restaurants). Unlike the first cluster, they are probably not highly influenced by health and entertainment activities in reference to it top 10 highly influential business categories.

The fourth cluster's neighborhoods are highly influenced by drinking & eating spots (beer bar, breakfast spot, burrito place) and shopping (gourmet shop, board shop, bookstore, pastry shop, fish & chips shops, miscellaneous shop) similarly to cluster 3. However, unlike cluster 3, cluster 4's business categories are different from that of the latter. Also, unlike cluster 1 and 5, the neighborhoods in this cluster are not highly influenced by the fun & entertainment activities,

The fifth cluster's neighborhoods were highly influenced by shopping (grocery store, print shop,electronic store), eating(restaurants),drinking(pubs), law service(lawyer), cleaning service (laundromat) and entertainment (pool hall). Among the 5 clusters, cluster 5 seem to be the most diversely influenced group of neighborhoods considering the top most influenced economic activities.

To further confirm that the clusters of neighborhoods obtained were quite distinct, MANOVA tests were used to test this. The following test results can be found in table 3. The degrees of freedom is the number of clusters (5) minus 1 and the F-test is the global test used to test the hypothesis. The F-test is obtained from the different multivariate tests. All tests test if the multivariate means are the same across the 5 clusters. Based on the obtained p-values (p-value <0.05), this hypothesis can be rejected and it can be concluded that the 5 clusters are indeed different.

Test name	**test value**	**degrees of freedom**	**F-value**	**P-value**
Wilks' lambda	0.8550	4.0000	4.0267	0.0046
Pillai's trace	0.1450	4.0000	4.0267	0.0046
Hotelling-Lawley trace	0.1695	4.0000	4.0267	0.0046
Roy's greatest root	0.1695	4.0000	4.0267	0.0046

Table 3: **MANOVA test**

5 CONCLUSION

The top venue or business categories which could interest investors can be found in table 2. These business categories seem to be the most economically active in respective neighborhoods. Based on the stated top economic activities, investors could profit if they invest in eating& drinking and shopping across all neighborhoods regardless of clusters. However, in case the investors are interested in investing in health (pharmacies), then neighborhoods in cluster 1 could provide a good niche. Also, in case investors are highly interested in fun & entertainment, neighborhoods in cluster 1 and 5. Also, there are other categories in which are present in one cluster's top economic activities such as health (pharmacies), cleaning service (laundromat) and absent in other. This absence could provide an opportunity for investors with monopoly tendencies to invest in neighborhoods where this category does not have a huge influence.