IBM Applied Data Science Capstone Project
Districts in Budapest by restaurant categories

Introduction

Budapest is a large Central-European city which accounts for more than 35% of the total GDP of Hungary, therefore, it has a central role in the Hungarian economic outlook (OECD, 2019). Hence, Budapest is not only the capital of Hungary, but the most important economic region of the country from the perspective of the services sector and tourism. In 2018, Budapest registered 12.5 million guests, which is a year-to-year increase of 5%. Moreover, it is expected that tourism will continue to flourish in the capital (Berende, 2019). On top of the growth in the number of visitors, Budapest also received the European Best Destination 2019 award, ahead of such cities as Athens and Florence ("Best places to travel in 2019", 2019). Tourism also has a great impact on the Hungarian economy as it contributes with 6.3% to GDP, which is composed by commercial accommodation & catering (amounts to 2% contribution to GDP) and investments in tourism ("Broader tourism sector generates more than 6% of GDP", 2019). The commercial catering sector is greatly sustained by the thriving restaurant market in Budapest, where one can find a variety of different cuisines. Nevertheless, there are defining restaurant categories one might find in the capital.

Business Problem

The purpose of this project is to segment districts in Budapest based on restaurant categories that can be found in the given district. A combination of descriptive statistics and machine learning techniques, such as clustering, will be used to support the aforementioned goal. The analysis will support the decision-making process of tourists and anyone that plan to dine in Budapest to look for a restaurant in the district that most suits their needs.

Data

To support the analysis, data will be collected from various sources. Firstly, a list of the districts of Budapest will be needed. Next, geographical data is needed for the respective districts. The geographical data will support the data visualization part of the project, especially the mapping of neighborhoods and venues. Lastly, restaurant categories data will be needed to perform the clustering on the neighborhoods.

To supply the data about the neighborhoods, Wikipedia will be used, which provides a list of districts with the corresponding neighborhoods in Budapest. Moreover, Wikipedia also provides a list of postal codes of the districts, which will be used in combination with geographical data from geodatos. Geodatos maintains the latitude and longitude coordinates of each district. Moreover, Foursquare API will be used to supply venues data from each neighborhood in Budapest.

The data from Wikipedia will be extracted using web scraping techniques with the pandas package. The geographical data from Geodatos will be extracted into a .csv file and joined to the main neighborhoods dataframe. Lastly, GET calls will be utilized from the Foursquare API based on the geographical data to have the venues data needed to perform the analysis.

Methodology

The research started with loading and wrangling the data about Budapest districts. Next, descriptive analysis was made to understand more deeply of the types of restaurants that exist in Budapest and their frequency in different districts. Lastly, k-means clustering was made in order to segment similar districts.

The data from Wikipedia had to be reorganized for the purpose of this analysis. First the 'Sights' column was removed, since it is out of the scope of this research to analyze data in this column. Next, the districts with no names had to be handled. In Budapest, some districts are called after it's number, therefore, the name for these cases had to be changed to the district number. Next, the district coordinate datapoints were added to the dataset. After the initial data wrangling, a geojson file was used from the work of Mór Kopronczay, which contained the district coordinates. This represents a crucial moment in further mapping of the clusters. The district coordinate datapoints were mapped together with the geojson to form a choropleth map with circle markers. However, it was noticed that the coordinate datapoints are often too close to the border of the districts, which would lead to issues when getting the restaurants data from Foursquare. Since restaurants would be gotten using the coordinate datapoints, if such datapoint is too close to a district border, then it is probable that venues that belong to the other district would have been called, since those lie in the defined radius for the GET call. Therefore, it was imperative to move the datapoints to a location where it’s within its own district and not close to any borders. Such alterations were made to the Districts: II, III, XX, XXII and XXIII. When getting the restaurants category venue data, a function was created that will call the GET request from Foursquare for each district based on a 500 meters radius, the name of the district and the coordinates of the coordinate datapoints. The newly found data was stored in a dataframe.

In order to understand the frequency distribution of different restaurant types in each district, first the data was one-hot encoded and then grouped by districts to find the mean value of each restaurant type in the separate districts. Later, the ten most common restaurant categories were stored in a dataframe for further analysis and clustering.

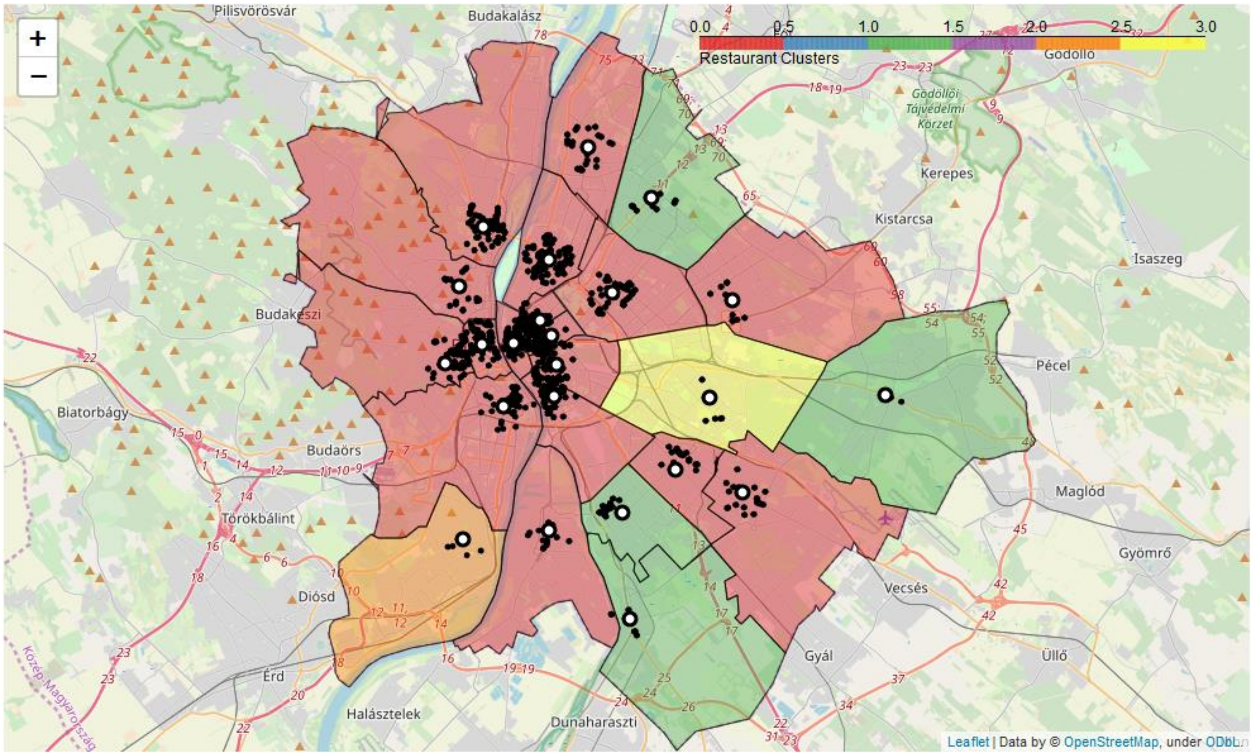
Before proceeding to fit a k-means model to the dataframe, the optimum value for k was determined. First the elbow method was used, which calculates the Within-Cluster-Sum of Squared Errors (WSS) for different values of k. One shall choose the k where the WSS starts to diminish. However, the elbow method did not effectively provide this research with the optimum k, since the slope of the WSS is too steep. Therefore, the silhouette method was used, which measures how similar a point is to its own cluster compared to other clusters. A high silhouette score value indicates that the point is placed in the correct cluster. Thus, one shall choose the k with the highest silhouette score. Based on this method, the optimal cluster number was determined to be four. Lastly, the model was fitted to our dataframe and the identified clusters were checked and mapped using Folium.

Results

It was found that the most frequent restaurant categories around the used coordinate datapoints are bakeries (9.9% of total), undefined restaurants (9.9%), cafés (7.5%), Hungarian (5.7%) and lastly Chinese restaurants (4.7%). One can find 71 unique restaurant categories in Budapest around the coordinate datapoints.

Based on the k-means analysis, four different clusters were identified. The first cluster contains 69% of the districts, which signifies the similarity of restaurant categories to be found in the city. Even though there is a large selection variety on the city, it can be stated that the categories distribution based on districts is rather homogenous, where the above mentioned 5 categories are commonly found in each district. The second cluster contains outer-city districts that are suburbanized. Here one will find mostly bakeries as this seems to be a common need for the population of these districts. The third and the fourth cluster contains one-one district which shows how these regions are somewhat different from the first two clusters. The most common restaurant categories are German restaurants and pizza places respectively. Nevertheless, only a small number of restaurants were gathered for these districts which can potentially lead to a skewed result in clustering.

The below plot visualizes the found clusters. The white circles represent the data points used the gather the restaurant categories data, while the black dots are the restaurants to be found near the coordinate data points. One can clearly see the homogenous nature of restaurant category distribution by districts of Budapest.



Plot 1: Districts of Budapest clustered by restaurant categories

Discussion

Budapest tends to have a homogenous nature in terms of restaurant categories that are to be found in each district respectively. Most of the districts appear in one cluster which is dominated by cafés, Hungarian restaurants and bakeries. Nevertheless, this research provides the reader with a comprehensive outlook on what restaurants to expect when visiting Budapest.

For future research in the topic, it is recommended to cluster by neighborhood centroids instead of district coordinates to be able to get more valid representation of each neighborhood in Budapest. Moreover, one would be able to find a higher number of restaurants which would lead to better clustering.

Discussion

This research attempted to cluster Budapest districts based on the restaurant categories to be found in each district. It is safe to conclude that there are many options for tourists and citizens of Budapest to dine. However, one will find a mostly heterogenous distribution of different types of cuisines, the most predominant ones being Hungarian and generic restaurants.

Bibliography

Berende, P. (2019). Hungary attracts record number of visitors. *welovebudapest.com*. Retrieved from https://welovebudapest.com/en/2019/02/13/hungary-attracts-record-numbers-of-visitors/

Best places to travel in 2019. (2019). Retrieved from https://www.europeanbestdestinations.com/european-best-destinations-2019/

Broader tourism sector generates more than 6% of GDP. (2019). Retrieved from https://bbj.hu/economy/broader-tourism-sector-generates-more-than-6-of-gdp_159634

OECD (2019). Regional Economy. *OECD*. Retrieved from https://stats.oecd.org/Index.aspx?DataSetCode=REGION_ECONOM