**Cuisine of Warsaw, Poland**

**Introduction**

The aim of this project is to analyse the market of restaurants in Warsaw, Poland and check which cuisines are most popular in the city. This information may be beneficial to people planning to open a new restaurant and help them decide what kind of food could be preferred by potential customers. To better understand the market it is important to get to know the competition. Future restaurant owners should know what type of food is popular in the neighbourhood or which cousins are trending in other similar districts but are not yet present in the one they are opening in.

**Data**

The list of Warsaw districts and their areas was checked in Wikipedia. The coordinates of Warsaw’s districts were found using geopy Python library which provides geocoding web services. This activity was based on district names and as a result latitudes and longitudes were gathered and joined with Wikipedia data. The information regarding restaurants was sourced from the Foursquare platform. The service provides location based experiences with diverse information about venues, users, photos, and check-ins. The combined data was used to visualize locations on a map, gather important statistics on top cuisines and apply K-Means clustering to discover certain groups of districts based on types of restaurants preferred.

**Methodology**

The first step of the analysis was to source a list of all districts in Warsaw. The information was found in Wikipedia and scraped from the webpage using pandas library. The data included name, population, area of each district. The area feature was used to calculate radius that required in the further steps. The geopy library provided coordinates of all districts and allowed building a map using folium library. The map shows areas that were analysed in terms of restaurants located within them. In the next step a function was built to source venues data that included: name, location and category. The venue category feature was used to determine a type of cuisine. Some of them had to be reclassified to match a more general group. In the next steps the data was manipulated to receive statistics of most common venues were selected for each district. Based on this information an attempt was made to group districts looking at the top five cuisines. This was possible thanks to the K-Means clustering algorithm. The method of vector quantization aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

**Result & Discussion**

As a result, four clusters were created. When analysed separately, it becomes quite clear which cuisines defined each group. Cluster 0 consists of districts that are commonly known as expensive locations and the most popular cuisines in these locations include: Italian and Japanese. American and Turkish cuisines are common in cluster 1 which includes old districts that are not well developed. Cluster 2 is built by developing districts inhabited by medium class. These people share interest in Italian, Japanese, American and Asian food. There is one district in cluster 3 that is different than all the other. People leaving there seem to prefer cuisine of the region which may be caused by the fact that the population of this district includes a significant number of immigrants from eastern European countries.

**Conclusion**

To sum up, the algorithm worked very well and not only grouped districts by popular cuisines but also brought a question about other possible similarities between them like welfare of people living there. This could be a great start point for further analysis.