**Introduction to the problem**

The city council of the city of Chicago is interested in whether there is a correlation between the types of crime that frequently occur in a community area and the popular venues that the area has to offer to its citizens. For example, in an area with many ATM, banks and jewelry stores, robbery and burglary may be popular types of crime to be committed. If there is such a correlation, we can predict the potential types of crime that may arise in a certain community area and, therefore, will have better solutions to prevent those from happening.

**Data**

The Foursquare location data is very useful for this problem since I can retrieve the popular venues in a specific community area. In addition, I will make use of following data set

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

This data set is publicly available and can be downloaded as a csv file. It contains all reported crimes in the city of Chicago from 2001 and is updated on a regular basis. According to this data set, there are over 7 million reported cases. For simplicity, I will only consider the cases that occur during and after 2015. This results in a subset with over 1 million cases. Each observation in the data set contains useful information about the case, including the primary type of crime, the community area in which it happened and the precise location in terms of longitude and latitude. Using this data set, wIwill know the longitude and latitude of a community area and, therefore, easily retrieve the list of venues for that area by making an API call to Foursquare.

**Methodology**

To solve this problem, I will mainly use K-means clustering. The analysis is divided into 2 parts:

1. In the first part, I make use of the crime data set supplied by the city of Chicago. Since the data set is big (more than 7 million observations), I read it chunk by chunk, each of which has 1 million observations. As mentioned earlier, I only consider a subset of this data set, which contains all cases during and after 2015. This results in a smaller data set with 1,338,577 observations. The most useful columns are the primary type of crime, the community area in which it happened and the precise location (longitude and latitude). To summarize, there are 77 community areas in the city of Chicago and 33 different types of crime that are reported.

For each community area, I count the number of occurences of each type of crime. I then divide each count by the total number of cases of the corresponding community area to get the percentage. Finally, I perform K-means clustering on this data frame with k = 5.

1. In the second part, I use the location data from Foursquare. Given the crime data set, I can approximate the location of each community area. In particular, for each area, I average the longitudes and the latitudes of all the cases that happened in that area. Then, I make an API call to Foursquare to retrieve a list of popular venues in each community area. It seems that some categories of venue have very low frequencies. Therefore, I decide only consider the categories that appear more than 10 times across all community areas. This reduces the number of categories from 225 to 40. This procedure gives me a data frame, in which each row corresponds to a community area and the number of occurences of each category of venue in that area. Based on this data frame, I perform a K-means clustering with k = 5.

In the first part, I divide the community areas into clusters based on the types of crime. In the second part, they are instead clustered according to what kinds of venue they have. By comparing the two clustering results, I can deduce whether there exists a correlation.

**Results**

Using the folium package, the community areas are displayed on a map and colored according to the clusters that they belong to. Based on visualization of the two clustering results, there seems to be no correlation between the types of crime and the popular venues that a community area has to offer.

Given that there are 77 community areas (numbered from 1 to 77) in the city of Chicago, we can form 77 x 76 / 2 = 2,926 unordered pairs (i,j), where i and j are distinct community areas (i is not equal to j). Out of these 2,926 pairs, the two clustering results agree on 1,644 pairs. In particular, given a pair (i,j), we say that the clustering results agree if they belong to the same cluster in both models or if they belong to different clusters in both models. This means that the clustering results agree on 56.2% of the total number of pairs. This percentage is close to 50%, which confirms our observation that there exists no correlation.

**Discussion**

Although the results show that there is no correlation between the types of crime and the popular venues in a community area, there are further works to be done to get a more accurate picture. For example, a formal statistical test can be conducted rather than an ad hoc approach as discussed in the previous section (56.2% might have been significantly different from 50% from a statistical point of view, provided that the total number of distinct, unordered pairs was large). We may also try different clustering algorithms (e.g. a non-parametric method like DBSCAN) to see if the results are consistent. There are a few other cities (e.g. San Francisco) that also make crime data publicly available. We can conduct the same study on these cities and compare the results.

**Conclusion**

The results of this study lead to the conclusion that there exists no correlation between thetypes of crime and the popular venues in a community area. However, additional analysis should be done to consolidate this conclusion.