2. Data acquisition and cleaning

2.1 Data sources

There is a clear relationship between a location, its venues available around, and the level of alms the population nearby are willing to provide. As an example, people would restrain giving alms near a shopping center, mainly for two reasons, because they usually pay by credit card and the amount of spare change is limited, and because security in those places restrict access to homeless people, even in the surroundings.

Mr. Florin has provided the following list of places that according to his experience, he gets the highest amount of alms (not necessarily in this order):

- Bakery
- Food & Drink Shop
- Pharmacy
- Betting Shop
- Scenic Lookout
- Ice Cream Shop
- Gourmet Shop
- Health Food Store
- Gift Shop
- Bookstore

In order to come up with a route that would optimize income, we would mainly need two types of data. On one hand, locations around Dublin, and on the other hand, all venues available from those locations, so we can gather the locations with the highest amount of venues listed before.

In Ireland counties are subdivided in a unique way, counties into baronies, baronies into parishes, and parishes into townlands. The townland (*baile bó* in Irish) is a unique feature of the Irish landscape and certainly existed long before the parishes and counties. An ancient division dating back to pre-Norman times, it is the common term or English translation for a variety of small local land units that varied in name and meaning throughout the island of Ireland.

For this project, we will get townland information from https://www.townlands.ie where there is a specific dataset here. The main objective is to get the coordinates (latitude and longitude) from each townland in Dublin's area.

We will also use Foursquare API (https://foursquare.com) as data provider for venues around a specific location, this information will be very useful for achieving our objective.

2.2 Data cleaning

The raw data set related to townlands contains 61.076 townlands and 27 parameters for each townland. Most of that data is irrelevant for the purpose of this project. Data cleaning

is needed for getting places only from Dublin's area and the parameters that can provide any meaningful information for the project's purpose. Once the data is filtered by county Dublin, we check for null values, and there is none.

Regarding the venues dataset from Foursquare, there is no issue, because we can obtain precisely the information we are looking for.

2.3 Feature selection

From the 27 features in the raw dataset, only 5 are necessary for the project's purpose. Those are:

- Name of the townland (NAME_TAG)
- County name (CO_NAME)
- Civil parish name (CP_NAME)
- Latitude (LATITUDE)
- Longitude (LONGITUDE)

From Foursquare, we will be interested in the following features for each Neighborhood:

- Venue name
- Venue address
- Venue latitude
- Venue longitude
- Venue Category