1. Introduction

1.1. Background

COVID-19 pandemic has marked a milestone in our lives: no matter your nationality, your job or your social status, this disease is being a major challenge for most of us. However, although this health and economic crisis has affected most of population world-wide, it is true that its impact has been much more serious in some countries or areas.

Factors such as ageing population, precarious jobs or economic models have been some of the reasons why some countries have suffered more infections and deaths and their economic recovery is likely to be more arduous.

Spain was one of the countries more affected during COVID-19 first wave, and factors mentioned above (ageing population, precarious jobs and its economic model) are thought to be the main reasons.

Despite the fact that Spain suffered a huge number of deaths, unemployment increase and serious hospital saturation due to COVID-19, the truth is this situation was unequal: though the situation was serious all around the country, it was critical in some areas.

Madrid, the capital city of Spain, was one of the areas where the crisis was uncontrollable: its large population (more than 3 million) and travellers (not only travellers from the rest of Spain, but also travellers from other countries) continuously coming and going are considered to be the principal causes.

Due to Madrid's great population, economic differences between its districts are notorious: whereas population of some of them is considered "working-class", the population of the rest is labelled as "upper class". It is confirmed that purchasing power is also a risk factor for COVID-19: people who live in areas with lower incomes are more likely to be infected. This is due to the fact that people who live in these areas often live with their whole family (grandparents, brothers and sisters, children, etc.), feel obliged to share a flat and expenses with flatmates and/or use public transport (which is usually overcrowded) instead of getting to their workplace in their own car.

Another risk factor for getting infected with COVID-19 is to frequent some venues with certain characteristics. Places which are indoors, usually overcrowded and/or where mask-faces are not obligatory, multiply the risk of infection.

Although everyone is exposed to COVID-19, some factors increase the risk of getting infected, and the ones mentioned in the two previous paragraphs are some of them. The study and the analysis this information can be useful to predict and avoid (or, at least, slow down) future outbreaks.

1.2. Stakeholders

As mentioned above, the place where you live and the venues you frequent can be risk factors for getting infected with COVID-19. Using the experience and the data acquired during the first wave of the pandemic, another similar situation can be avoided in the future.

Taking the districts of Madrid, their number of infections by COVID-19 and their most popular venues, these districts can be clustered in order to know which ones are more suitable to suffer another outbreak. Knowing it, governments will be able to take measures as soon as the number of infections starts to grow in any of these districts. Not only will this knowledge be useful for these specific districts, but it also can be applied in areas with similar characteristics.

Therefore, on the basis of a dataset which contains the number of COVID-19 infections in each district of Madrid, the first step of this project will consist of labelling them depending on the number of infections. The next step will be getting the most popular venues for each district: using the coordinates of each district and FourSquare, we will be able to add to each district its most popular venues. Finally, districts will be clustered by their most popular venues.

Having districts labelled by number of infections and clustered by their most popular venues, the type(s) of venue where an outbreak is more probable can be easily identified and restrictive measures can be sooner taken and customized for this/these kind(s) of venue.

2. Datasets

In order to get the aim of this project, three datasets will be used. Although each dataset will be shown in detail when loaded in the project, a brief description for each dataset is written below.

2.1. COVID-19 data by district

The first dataset, provided by the Community of Madrid[1], contains COVID-19 cumulative incidence on a specific date by town and, in the case of the city of Madrid, by district. Since this project will focus exclusively on the city of Madrid, only data relative to districts will be used.

This dataset can be downloaded from here:
https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos

2.2. Districts' coordinates

In order to visualize districts over a map and relate them with FourSquare information, it is needed to get the latitude and the longitude of each district. Due to the number of districts (only 21) and the impossibility to find their coordinates in any open data repository, coordinates of each district will be got "by hand" using Google and saving the result in a custom data set, which will contain the name of every district and its coordinates.

2.3. Most popular venues by district

Using the coordinates of each district, most popular venues in each one will be got through calls to FourSquare API.

[1] Spain is divided into 17 communities and the capital city of Spain, Madrid, is located in the Community of Madrid. Therefore, "the Community of Madrid" is not the same as "the city of Madrid". In this project, we will study the city of Madrid.