DATA DESCRIPTION AND ANALYTIC METHODOLOGIES

Data origins and descriptions

Geographical data: First, we are going to scrap the Wikipedia page that describes each Madrid district and neighborhood by using Beautiful Soup library in order to extract the corresponding table and arrange it as dataframe. Then we are going to extract the coordinates of each district geographical center with the aid of the Nominatim module of Geopy library. These coordinates will be subsequently added to the district dataframe, from where we will later plot in a map through the Folium library each district marker according to its coordinates. These markers will later be used to map district clustering results.

Real Estate data: In this work we will simplify the real estate factor by focusing on district property values represented as average price per square meter for each of them. These values are dynamic and many sources exist reporting reference values at different times. After reviewing several web sources, the chose to extract values from two sources. The first one is an article of El País-Cinco-Días reporting October 2019 square meter prices from, where a chart image was typed in an Excel spreadsheet. The second source came from the Real Estate site Idealista.com, from where we extract a table with 2021 values that are easily copied to an Excel spreadsheet. We chose to work with 2019 and 2021 data and not 2020's because the later was an odd year to the whole World. With these values we calculate price variations, average prices and normalized prices to improve our observations and include them as new features in our main dataset for further processes.

Safety data: From Madrid municipality data repository datos.madrid.es we will download 2021 monthly police intervention tables where they describe the number of interventions per district and intervention type. These tables are then loaded and mined in our Jupyter notebook. We totalize the number of cases per category and district with the help of Numpy library. Afterwards, we group the cases by similar categories. For example, the cases where people and patrimony were affected and guns seized, will be grouped as an unsafety category, because these crimes directly affect the local safety. Drug possession and consumption are to be grouped as a "drug" category. They don't necessarily affect the community safety but represent a factor that affect to a lesser or greater degree the quality of life of the community. From the "unsafe" and "drug" attributes we generate a proportion version of each of them, with respect to the corresponding total of cases. All these generated attibutes will be integrated to our main data frame.

Venues per district: Now it is time to review the main venues that the Madrid districts possess. We will extract them with the Foursquare API. Once extracted the venues per district, we then proceed to count them, count the different venue types per district and group these new attributes with the real estate and safety ones. Then we proceed to plot these attributes to observe their correlations and interdependency. Once completed this step we move to the analytic methods we have thought for this case study.

Analytic methodologies

Plots: Our mined datasets will be plotted to understand the attribute behaviors across the different districts and identify any existing trend or correlation among them. Through the plot observations we expect to get some initial insights about each district and understand the data inputs for the next analytic methods. To perform the plotting steps we mainly rely on Matplotlib Pyplot library.

Linear regression: One of the questions we want to answer is how the quantity and variety of venues, the unsafety and drug cases affect the property values per district. We will evaluate all and some of them to conclude what are the most important factors that might be related or affect the average district property values and how confident we are affirming this. For this method, we will rely on the Scikit-Learn's Linear Regression module.

K-Means clustering: Our next analytic step will be the clustering of the districts according to their different features such as property prices, safety and drug indicators, total and variety of venues and the quantity of each venue type. We select this method because it allows us to choose the quantity of classes that we want: A fair quantity that allows us to sufficiently differentiate the districts and keep the same class districts as similar as possible. We will use as analytic tools the "KMeans" module of the Cluster Scikit-Learn library to perform the clustering step and "cdistance" module from Geopy's Spacial Distance library to help us quantify the internal differences of the resulting clustered classes. Once again, we will then use the Folium library to plot the district markers know colored and labeled with their class output. Afterwards we are going to check the districts included in each class along with their attributes to discuss and understand what they have in common and how they can work for the different people profiles and interests.