Clustering of French Cities

# Introduction

## Background

France is divided into 18 regions : 12 in mainland France and 6 elsewhere (1 in Europe ( Corsica), 2 in the Caribbean, 1 in South America, and 2 near Africa. We usually talk about Metropolitan France (mainland France + Corsica) and Oversea France (France d'Outre-Mer).

Following the 2016 reform, France is divided as follows :

* 13 Metropolitan regions, including Corsica
* 5 Overseas regions
* 96 departments (Each region is subdivided in one or more department)
* and further subdivisions that we will not talk about here

This what the map of the regions looks like:

A close up of a map

Description automatically generated

## Problem Statement:

Before the 2016 reform, the regions were smaller, in fact Metropolitan France went from having 26 to just 13 regions. The debates were intense to know how the new cut would made and which regions would be merged.

The goal of this project would be to explore if by clustering the France cities **we are able to reproduce France regional division**. That is find if we can characterize the regions with common "features" found in the cities that make up the region.

This could be of particular interest for the **French Government** as, 4 years after the reform, it could justify or deny the choices made then.

# Data

In order to solve this problem we will use the list of french cities [here](https://sql.sh/736-base-donnees-villes-francaises)as well as climatic data from the [openweather api.](https://openweathermap.org/api/statistics-api" \t "_blank) Finally as required, we will use data from the Foursquare API to further categorize our cities.

## Cities

The cities data have been exported from the website into a csv file that can be found in the Data folder of the repository of this project. Note that the CSV has no headers. You can find a description of the columns at the original website but find below a summary of what the database contains.

* Various versions of the name of the cities
* Coordinates of teach city (so precious to call the Foursquare API)
* Population data (will certainly be very useful as well)
* Other geographical data (min and max altitudes, area)
* Department number (it would be interesting to check if the algorithm learn and find the relationship between regions and departments)

## Weather

We will use the openweather API to retrieve the weather predictions of the day for each city. From this API, we can get data such as minimum recorded, maximum recorded, average minimum, average maximum for the following climatic features :

* Temperature
* Pressure
* Humidity
* Wind
* Precipitation
* Clouds

## Venues

We will use the Foursquare API to retrieve the venues that can be found in each city. If you access this notebook from outside this course please check the [developer.foursquare.com](https://developer.foursquare.com/) website. In a nutshell, with this API you can get the following.

* What are the most popular venues are there in each city?
* How many of those are there?
* How many tips (recommendations, opinions from visitors) those venues received?

# Methodology

## Statistical Analysis, Scoping and Filtering

We used the pandas describe method to get a quick report on the data.

![A screenshot of a cell phone

Description automatically generated]()

We noted a few interesting things from this table:

* The cities names are not unique therefore we cannot use them as-is for indexing
* A department can contain up to 62 cities
* Except for the altitudes we do not have missing values
* The population, density and area data are very spread out, with the standard deviation being superior to the mean
* The area max seems to be an outlier

## Data Consolidation

## Preprocessing

## Modeling

# Results

# Discussion

# Conclusion