**1. Introduction**

**1.1 Business Problem**

It is important to determine a city’s, region’s or state’s population. This can be very helpful in tasks such as resources distribution, network allocation. Therefore, we would like in this project to make an analysis on US Population by State by clustering the different states based on their coordinates on one hand, and their positions on the other hand.

**2. Data Acquisition and cleaning**

**2.1 Data sources**

For this task, we will be using the US City Populations Dataset by Gabe Salzer (link to dataset : https://data.world/gmoney/us-city-populations). Data includes every city name (over 6k), its population and state. Numbers are from August 2015.

It is composed of 3 columns: State, City name and Population.

**2.2 Data Cleaning**

To achieve this task, we will only need the state and population columns, aswell as the latitude and longitude of each state. So we will delete the city column and group the dataframe by state, then sum the population for each city in that state.

There are no missing, or wrong values in the dataset. So we can move forward to the next step

**3. Exploratory Data Analysis**

**3.1 Most populated cities**

**Une image contenant capture d’écran

Description générée automatiquement**

Here we can see that New York, Los Angeles and Chicago are the most populated cities in 2015.

**3.2 Most populated States**

**Une image contenant capture d’écran

Description générée automatiquement**

The DC state has by far the most population. Followed by Nevada, Arizona and California

**3.3 Number of cities per state**

Une image contenant capture d’écran

Description générée automatiquement

We can see here that most states have between 0 and 50 cities. Less states have more than 50 cities. Only few states have more than 600 cities.

**4. States clustering**

We will now make a clustering of the different states, we want the clustering to be done taking into account two criteria’s : The coordinates of the state, and its population (the sum of the population of each city in this state).

**4.1 KMeans clustering**

We choose for this class a classic KMeans clustering with 5 clusters. The results are as follow:

Une image contenant texte, carte

Description générée automatiquement

We can see that a cluster has appeared in the East of America, this corresponds to states that are more populated than other states in the center of USA.

**5. Conclusion**

Through this project, we have applied all our previous knowledge to solve a problem which is USA states clustering based on population. We first prepared and transformed our data, then explored the data using exploratory data analysis. Next, we applied KMeans clustering to extract the different clusters and we finally projected them on a map of the USA.