|  |  |
| --- | --- |
|  | Applied Data Science Capstone |



Final Assignment

Global Cities Climate: Clustering Insights From a climatology perspective

(Using global climate data to cluster cities)

Christodoulou George

**Thessaloniki 2018**

Scenario – Business Problem

# 1 Introduction

Due to financial crisis over the last decade, there has been a huge population migration from weak economic countries to countries with stronger financial background and infrastructure. Choosing a country or a city to relocate had only one criterion: opportunity for self-employment and self-improvement aka setting the basis for starting a new life. Taking other parameters under consideration for relocation such as weather conditions or average quality of life was a luxury that were rarely considered, since individuals were on the brink of financial destruction and of course there weren’t enough job opportunities.

# 1.2 Business Problem

However, in the last decade, along with the economic crisis, some professions have been emerged, in which the physical presence of the employee in the company is not needed (remote jobs). Many companies also spread to other countries where the workforce was cheaper, creating more opportunities for people who wanted to emigrate and choose a job of their own specialization. What if an employee has two or more job opportunities in different countries / cities with similar financial earnings? Then additional criteria have to be taken into account to make the final decision. One of these criteria is the climatic conditions prevailing in a country / city. Knowing the climatic conditions of a city, how can an employee know which other cities have similar weather conditions? That would be quite helpful for the employee’s relocation problem.

Data

# 2.1 Data Description

The problem was described in the Scenario-Business problem section as city clustering under the scope of similar climatological weather conditions. The data used for solving this problem were searched and retrieved from various climatology sites such as weatherbase.com. Despite the fact that there are huge amounts of data regarding climatology reports in cities and countries, an excel file which contains a summary of climatological data per city or per country, was never been created. Thus, *I was forced to create my dataset from scratch*. The excel file consists of 108 rows and 14 columns. From 87 countries 108 cities were selected: 87 capitals and 21 other major/known city.

The columns include the following not climate information: Country, City, Latitude and Longitude. This will help us create folium dots on a map that represent all cities reported in the excel file.

The rest of the columns contain climatological variables such as Average Temperature, Average High Temperature, Average Low Temperature, Average Precipitation, Average Number of Days With Precipitation, Average Length of Day, Average Number of Days Above 30-32 Degrees C, Average Number of Days Below 0 C, Average Relative Humidity and finally Average Wind Speed. These will be our variables which will help us find similarities among the aforementioned 108 cities by feeding them to our clustering algorithm.

The goal is to create a colored dot world map, where each dot represents a city and each color similar climate characteristics.

Methodology

# 3.1 Data Preprocessing

For analysis and processing of the data, python programming language was used in combination with data processing, machine learning and visualization libraries such as pandas, matplotlib, numpy, folium, sklearn and seaborn.

## 3.1.1 Nan Values

Since the data was inserted manually into the excel file, I already had an idea regarding the variables’ datatype:

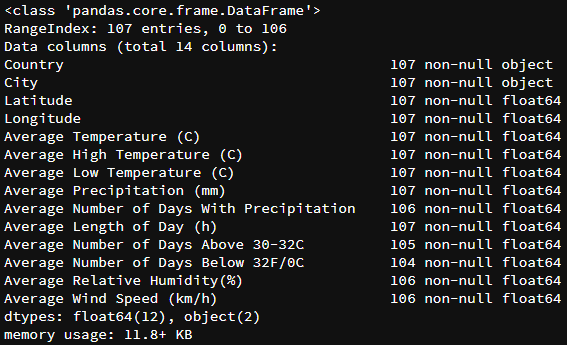


Figure 1

As we see from fig. 1 there are some values missing from certain columns (NaN Values). Total number of NaN values are 8. There are two ways handling the missing values. You can either just drop all rows containing NaN values or you can replace them with the average value of each column. The average replacement method was chosen.

## 3.1.2 Data Normalization

Since there is a great value variation for the variables in the columns dataset, normalization of ratings was a must. Normalized/rescaled values allow the comparison to some relative size variable. Thus, all columns regarding climatological data was normalized. Of course the first four columns including country, city and coordinates were left untouched.

## 3.1.3 Map Creation

Using the folium library, a world map was created. The zoom level was initiated so that the user could have an overall view of the entire globe.

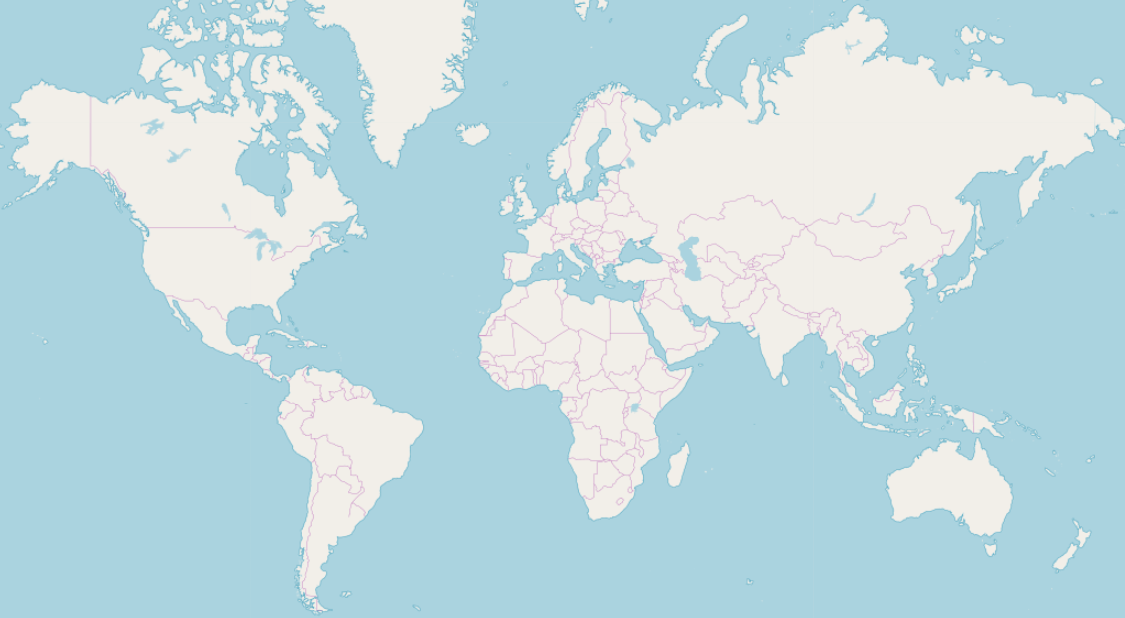


Figure 2

After that, I needed to make sure that the coordinates in my dataset were correct. So inserted 108 cities of my dataset onto the map and checked for any anomalies.

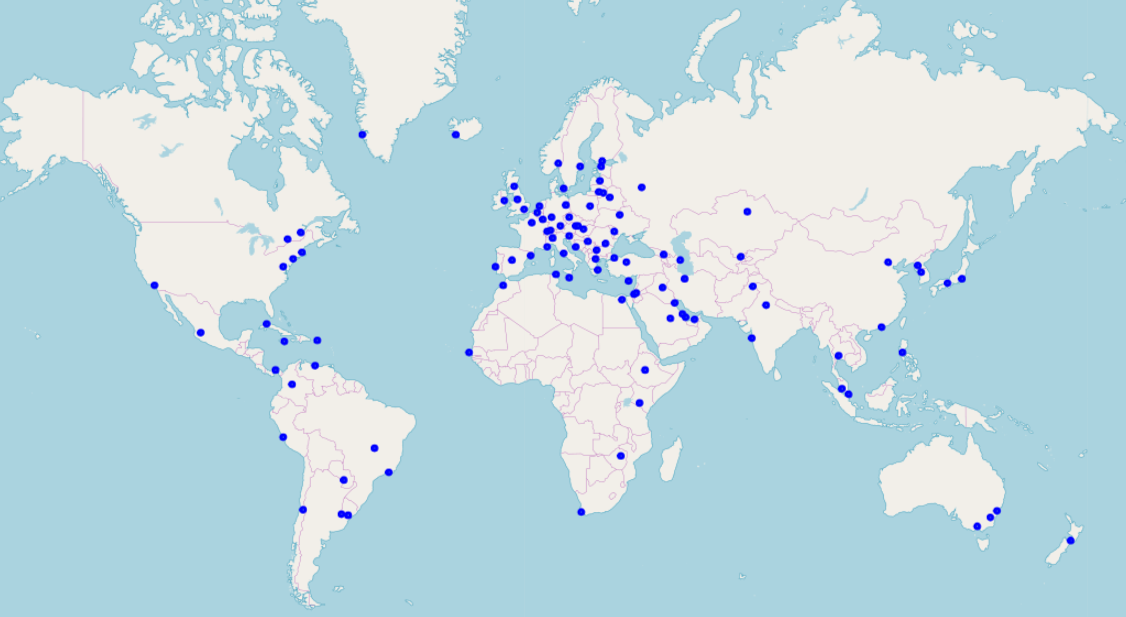


Figure 3

Cities were mapped without any errors, therefore the preprocessing part is over.

# 3.2 Clustering Process

## 3.2.1 The “elbow” method

The problem was defined from the start: Cluster cities according to their climatological behavior. However as in all clustering problems, we should define the exact number of clusters we should use. In order to address this issue, the “elbow method” was used. The “elbow” method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point.

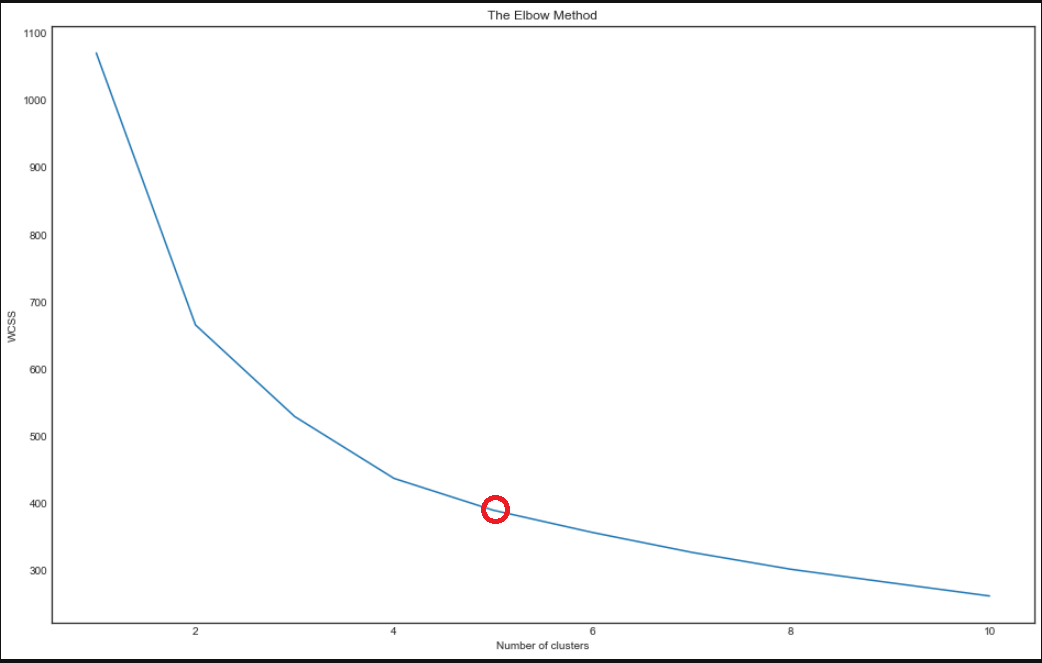


Figure 4: The elbow method

Thus it was decided that the optimal number of clusters should be 5.

### 3.2.2 Machine learning: K-Means algorithm

Using the K-means, provided by sklearn library, the climate data was fitted into the algorithm. K-means was chosen because it is one of the simplest and fastest unsupervised learning algorithms that solve clustering problems. For clustering purposes, a label from 0 to 4 was assigned to each city.

Results

# 4.1 Results

Since we cluster each city under a different label numbered from 0 to 4 according to their climatological similarity, it is time to visualize our results on to a world map:

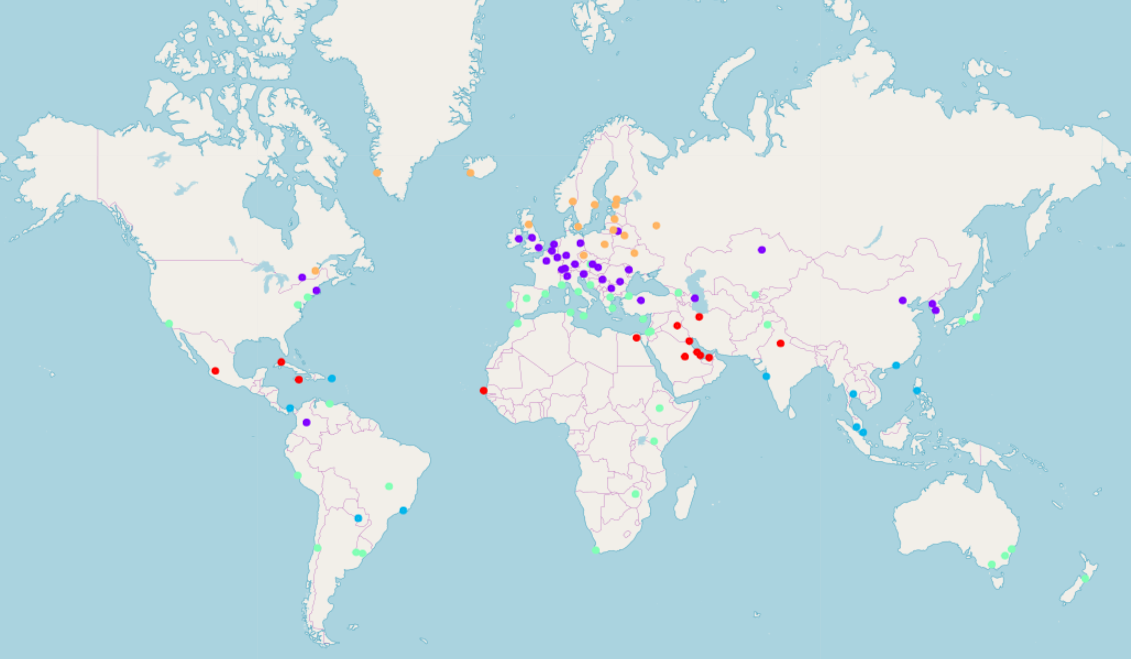


Figure 5: Clustering cities

|  |  |
| --- | --- |
| Color | Cluster Number |
| Red | 0 |
| Magenta | 1 |
| Blue | 2 |
| Tirquaz | 3 |
| Orange | 4 |

Since the map is interactive, the user can click on each dot, revealing the name of the city and its cluster.

From the map we can see the similarity between cities in southern Europe such as Athens, Rome, Barcelona etc, Africa such as Cape Town, Harare and Addis Ababa, Asia such as Tokyo and Islamabad, America such as New York, Brasilia and Buenos Aires and finally Australia’s Sidney, Melbourne and Canberra.

Clicking on other cities on the notebook, we can identify their names and the cluster to which they belong.

Discussion/Conclusion

# 5.1 Discussion – Future work

Being a citizen of Greece and having lived in both Athens and Thessaloniki (2 cities of my data) I notice that my algorithm has ranked these two cities in the same group. It is a fact that the climate in these two cities has several differences. Obviously if the grouping was done at country level, then these two cities would belong to different groups. In order to perform this, there should have been data from many cities which belong to the same country. But due to the fact that our data includes cities around the world, obviously the correlation of the Athens-Thessaloniki climate will be stronger than for example Athens-Oslo or Thessaloniki-Oslo. Therefore, our algorithm clustered Athens and Thessaloniki together.

Future work may include increasing the dataset, adding more cities and performing the clustering algorithm country wise. By doing so, interesting facts may emerge, like climate similarity but in this case from country perspective.

# 5.2 Conclusion

This project was about clustering cities according to their climate characteristics. A dataset was created from scratch, containing data acquired from several metrological-climatological sites. The dataset also included the city’s name, country of origin and its coordinates. After processing the data, it was fed to K-means machine learning algorithm, which grouped each city according to its climate characteristics. Finally, a different color was assigned to each cluster and cities were printed on a world map via dots.

It was very interesting observing the power of K-means clustering, contributing to the solution of relocation problem.