# Data collection

For this exercise I am working on a dataset on covid that was scraped via API calls and then cleaned. The dataset was loaded into a dataframe.

Table

Description automatically generated

Below is the info on the dataframe:

RangeIndex: 169 entries, 0 to 168

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 country 169 non-null object

1 ISO\_2 169 non-null object

2 ISO\_3 169 non-null object

3 Population (2020) 169 non-null int64

4 lat 169 non-null float64

5 lng 169 non-null float64

6 Density (P/Km²) 169 non-null int64

7 Confirmed 169 non-null int32

8 stringency\_index 169 non-null float64

9 Infection\_rate 169 non-null float64

dtypes: float64(4), int32(1), int64(2), object(3)

memory usage: 12.7+ KB

Below is the five number summary:

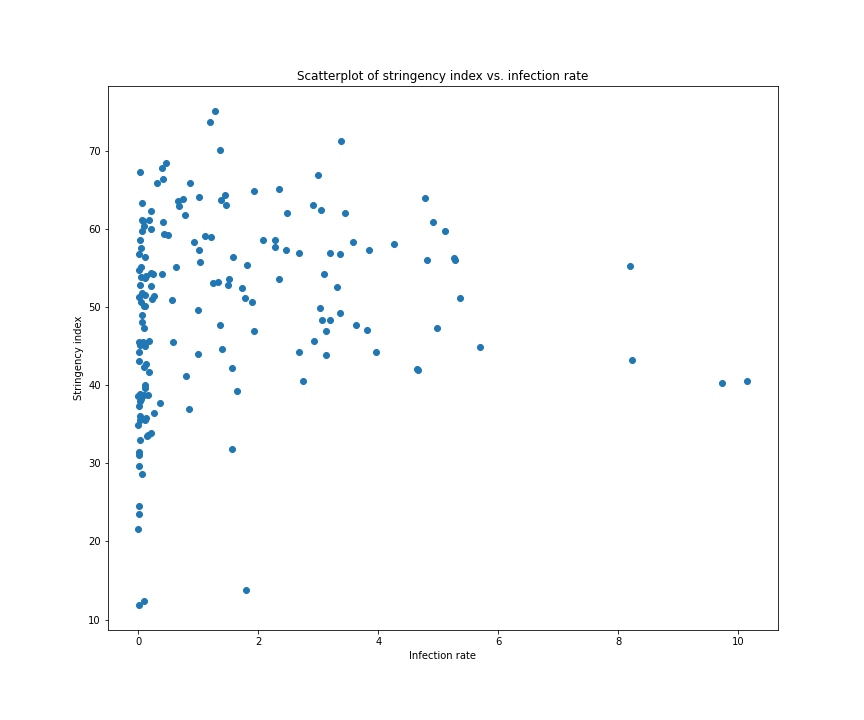
Table

Description automatically generated

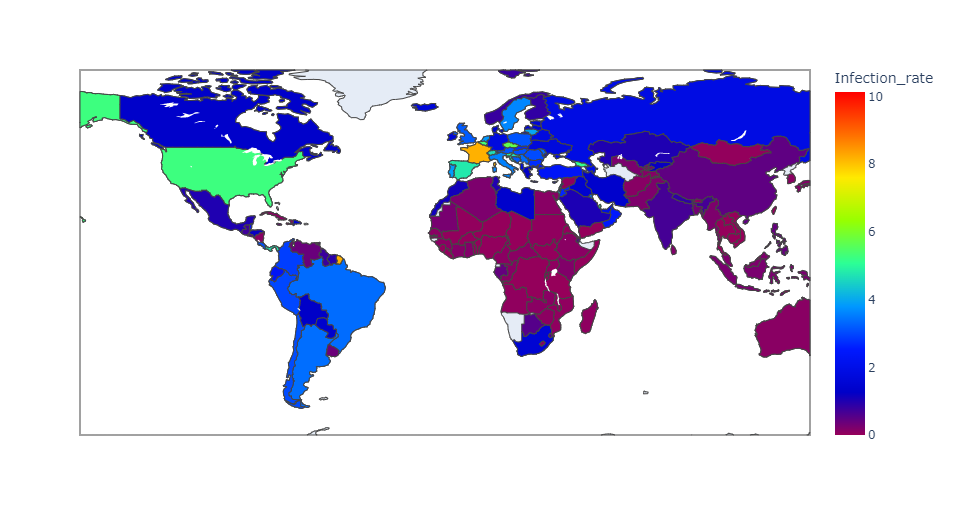
# Feature selection and Exploratory data analysis

We will try to create cluster based on stringency index and infection rate.

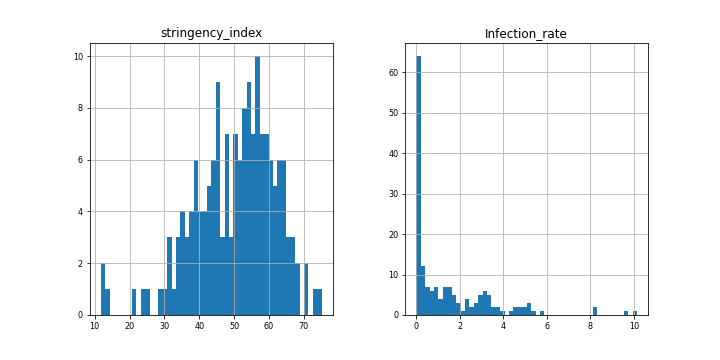
Below is the scatterplot for stringency index vs. infection rate:



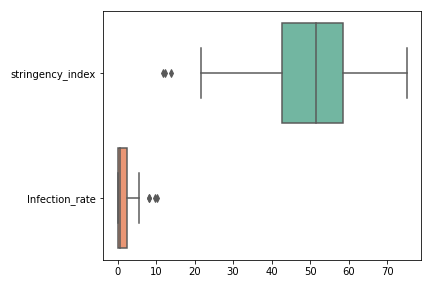
Below is a choropleth map with intensity on infection rates:



Below are the histograms of those 2 features:



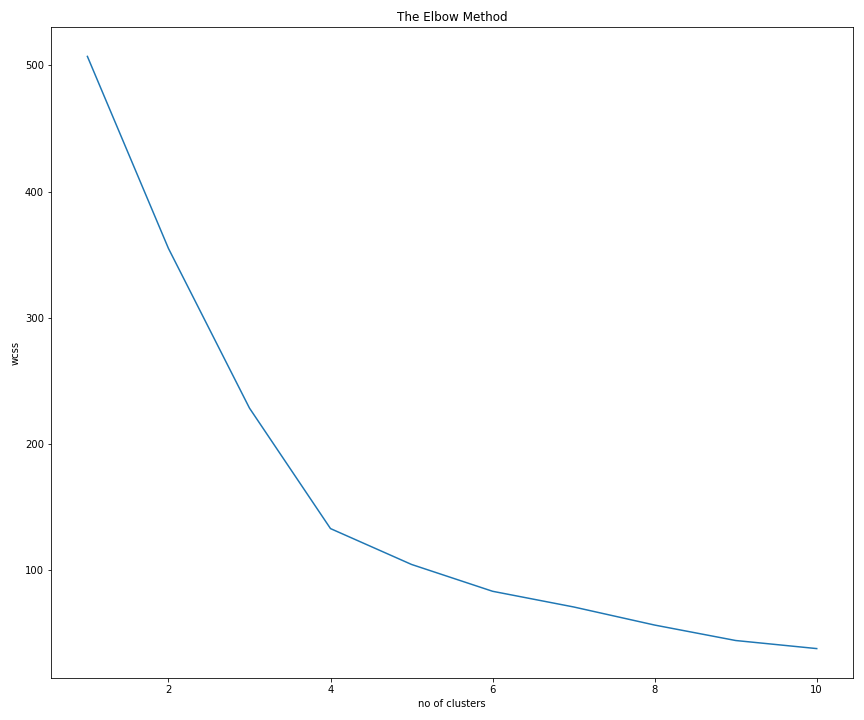
Below are the boxplots of those 2 features:



Even though I found a couple outliers in those 2 features, as they are not drastic or extreme, I used standard scaler to scale the data.

# K-means

Below is output for elbow method:



We are going to use k=6 for k means clustering.

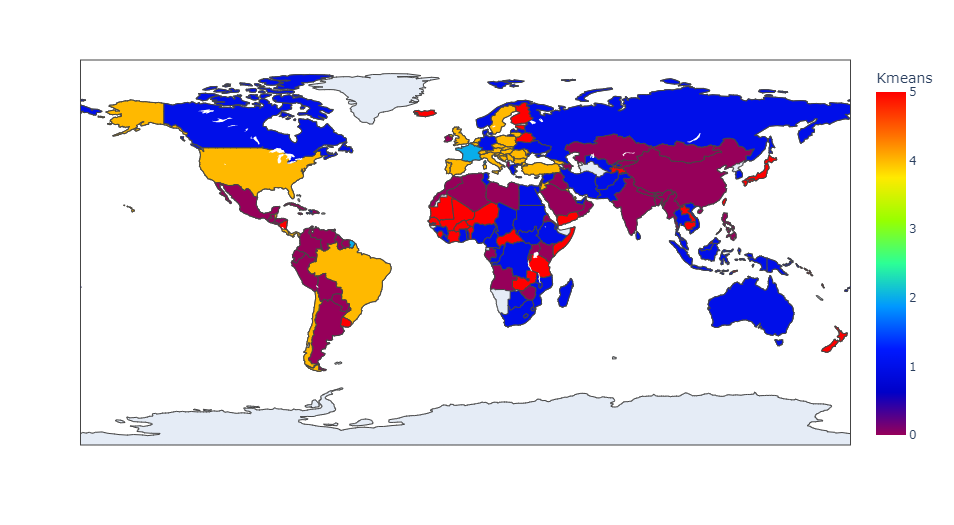
Below are the value counts for k means:

Table

Description automatically generated

Below is the scatterplot colored with k means:

# 

Below is a choropleth map with intensity on k means clustering labels:

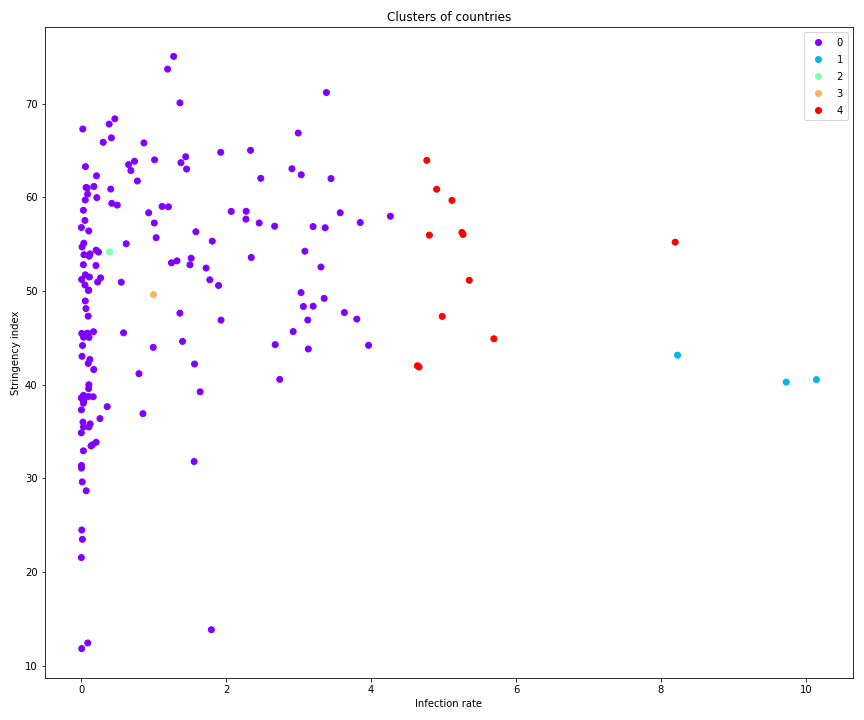
# Mean shift

I have used bandwidth = 1 as parameter which created 5 clusters.

Table

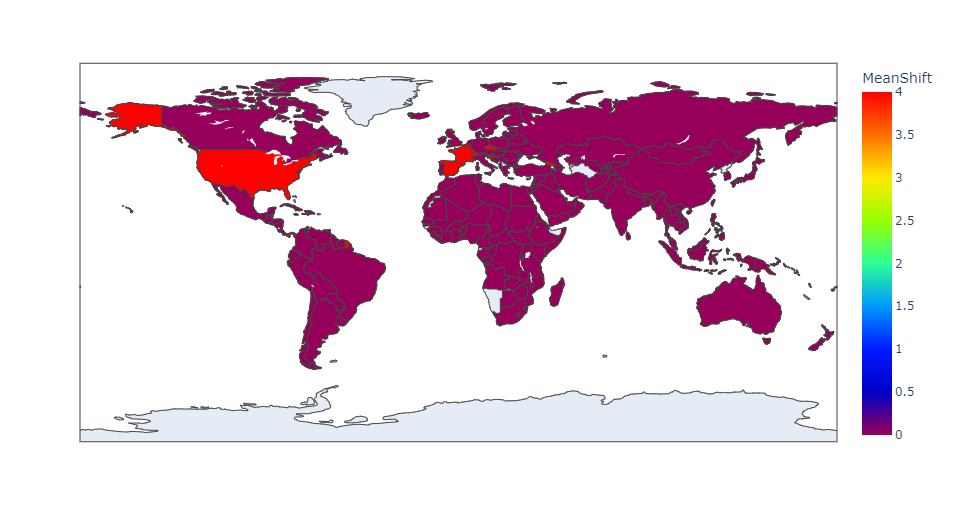
Description automatically generated

Below is the scatterplot colored with mean shift clusters:



With higher bandwidth, less clusters were being created and with lower bandwidth, more clusters were being created. Both cases, the clusters were not properly segregated because of overlaps.

Below is a choropleth map with intensity on mean shift clustering labels:



# Cross evaluation

Table

Description automatically generated

Looking at the counts and the graphs above, I noticed the mean shift is grouping majority of the records under one cluster (Mean shift label= 0), whereas K means is breaking that same group of records into 4 clusters (K means label 0,1,4,5). Apart from this, both techniques created 2 clusters for data points on the right side of the scatterplots. In this case, I would go with K means clustering as it showed more accuracy and convergence.