Prediction for the countries/areas which are most likely to be affected by the virus in the next decade

The Problem approach and the dataset description

Assumptions:

1. There has been lots of missing data in the 27 lines .Not sure what is the purpose of that data. I am not considering that data for the time being. Assume some of the missing data can be filled by the regions/countries likely effected
2. Not considering the Data with year format XXXX-XXXX which represents duration of year.
3. There is no description on the "STATUS" column
4. The output to be in an **csv** file having list of countries and effected region coordinated where occurrence time is less than 5 years

The Dataset

1. (42042, 15), 15 Columns of data
2. Categorical values
   ['VECTOR','SOURCE_TYPE','COUNTRY_ID','infection_source', 'infection_time','X','Y']

The Problem:

1. Since there is no clear distinction about a value to be predicted this is clearly an **Unsupervised problem.** Since we need to classify and segment the data as we do not know what value to predict and based on that we can provide the probability of region to be in which region.

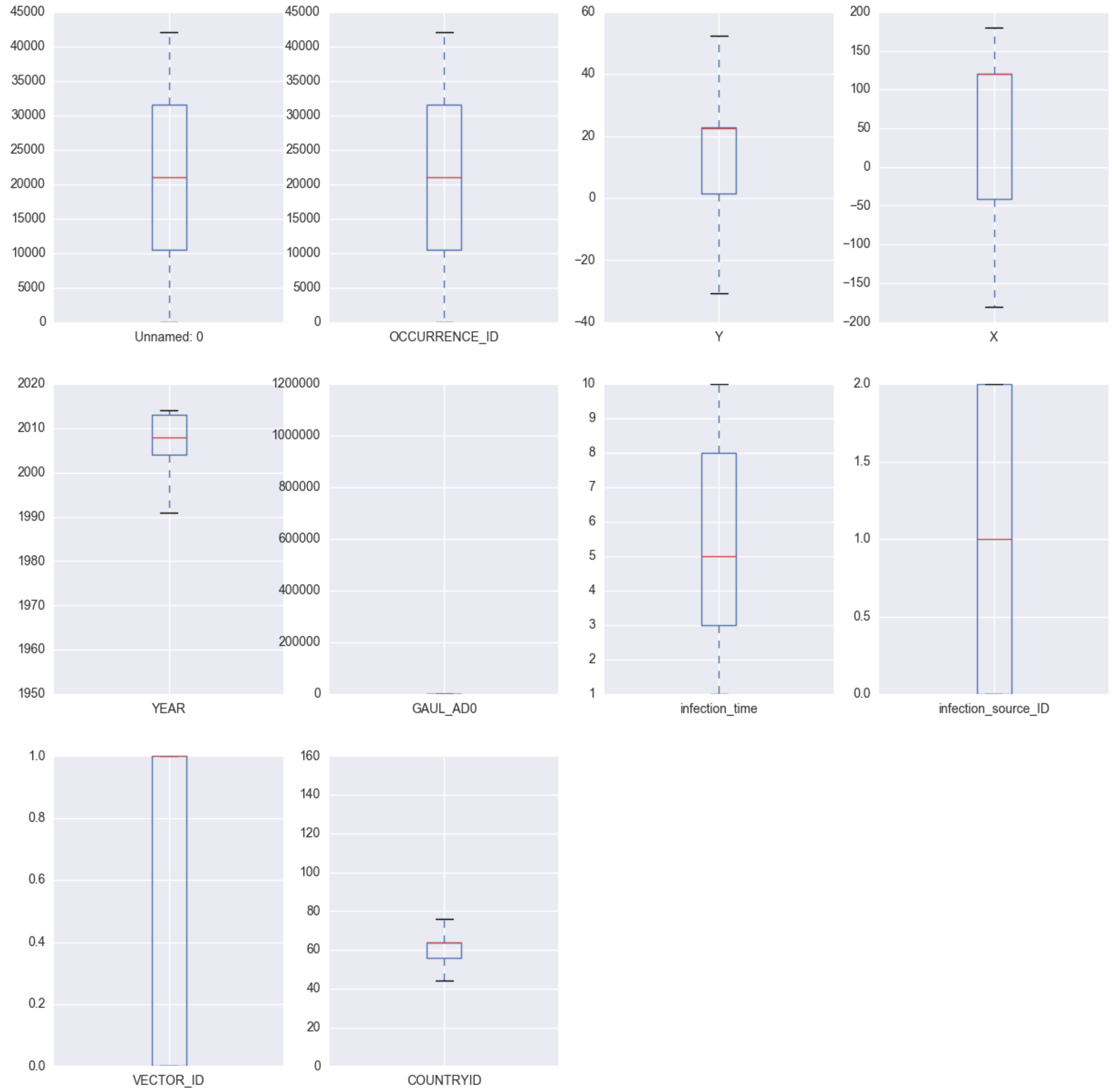
2.We can hence create cluster of the region (segment the regions) based on the High and Low probability of infection time.

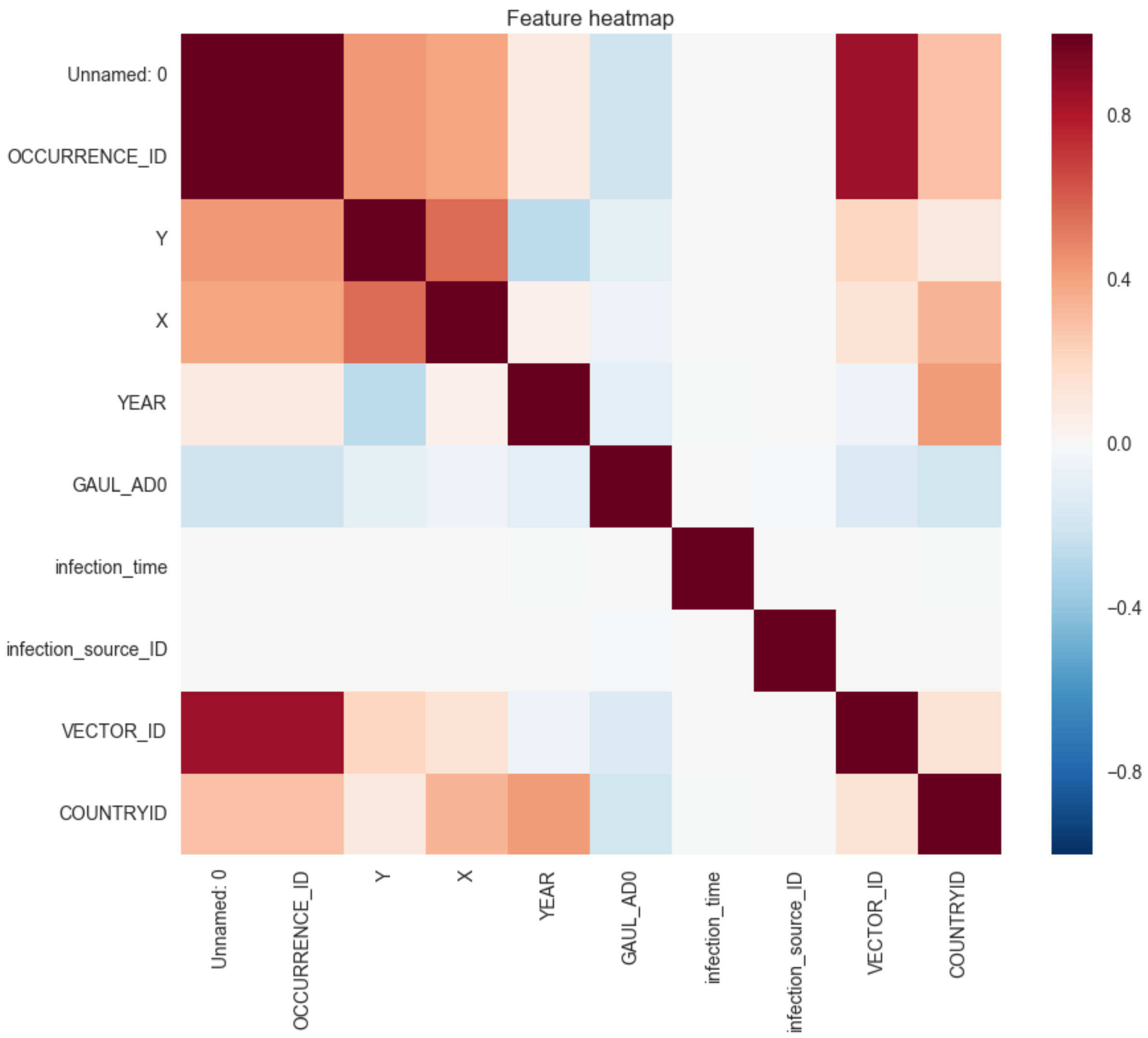
3. If the "infection time" is less than 5 years high likely to be in the High risk cluster and more than 5 will be in Low risk cluster.

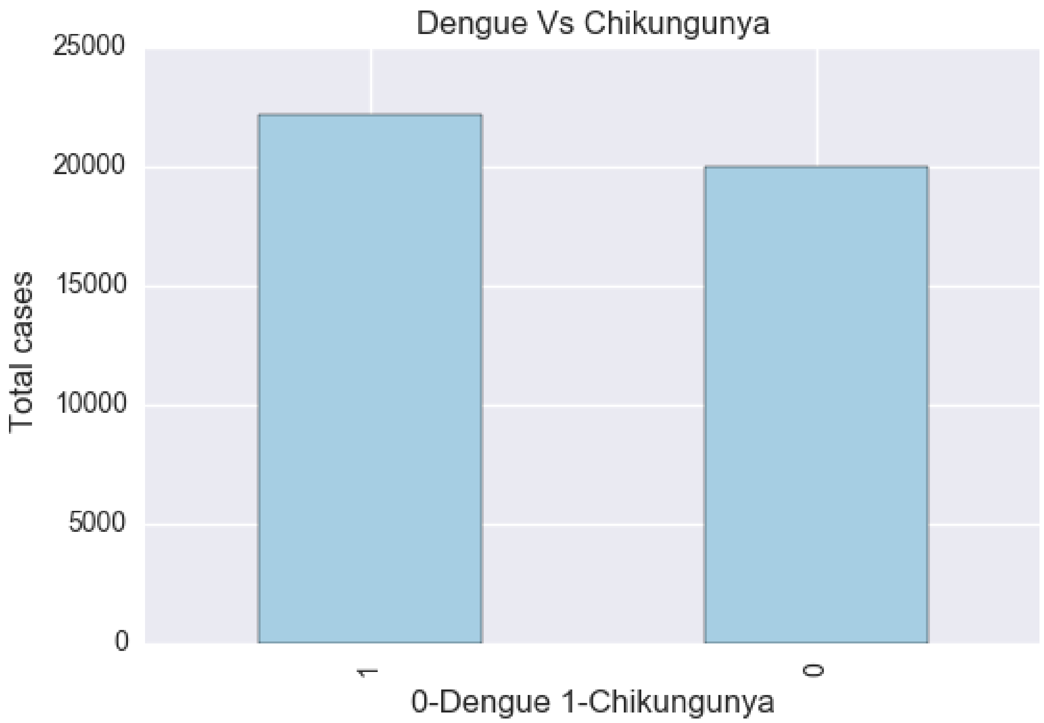
4.Using K-means (algorithm approach) we can find the no of best suited clusters and we can classify as "Risk region"

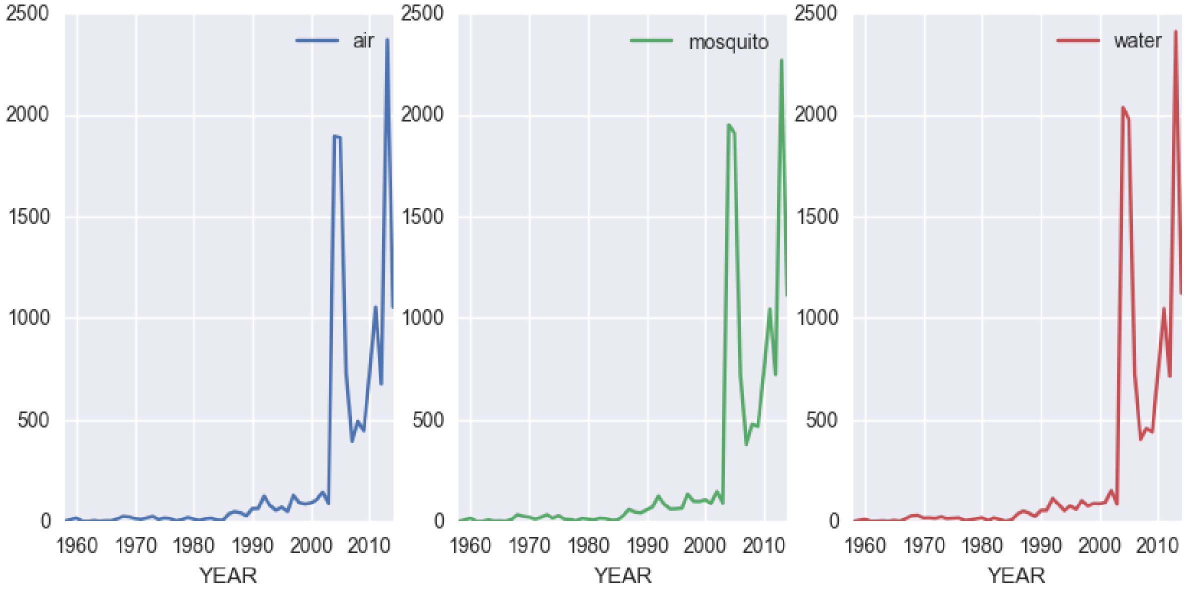
Solution approach:

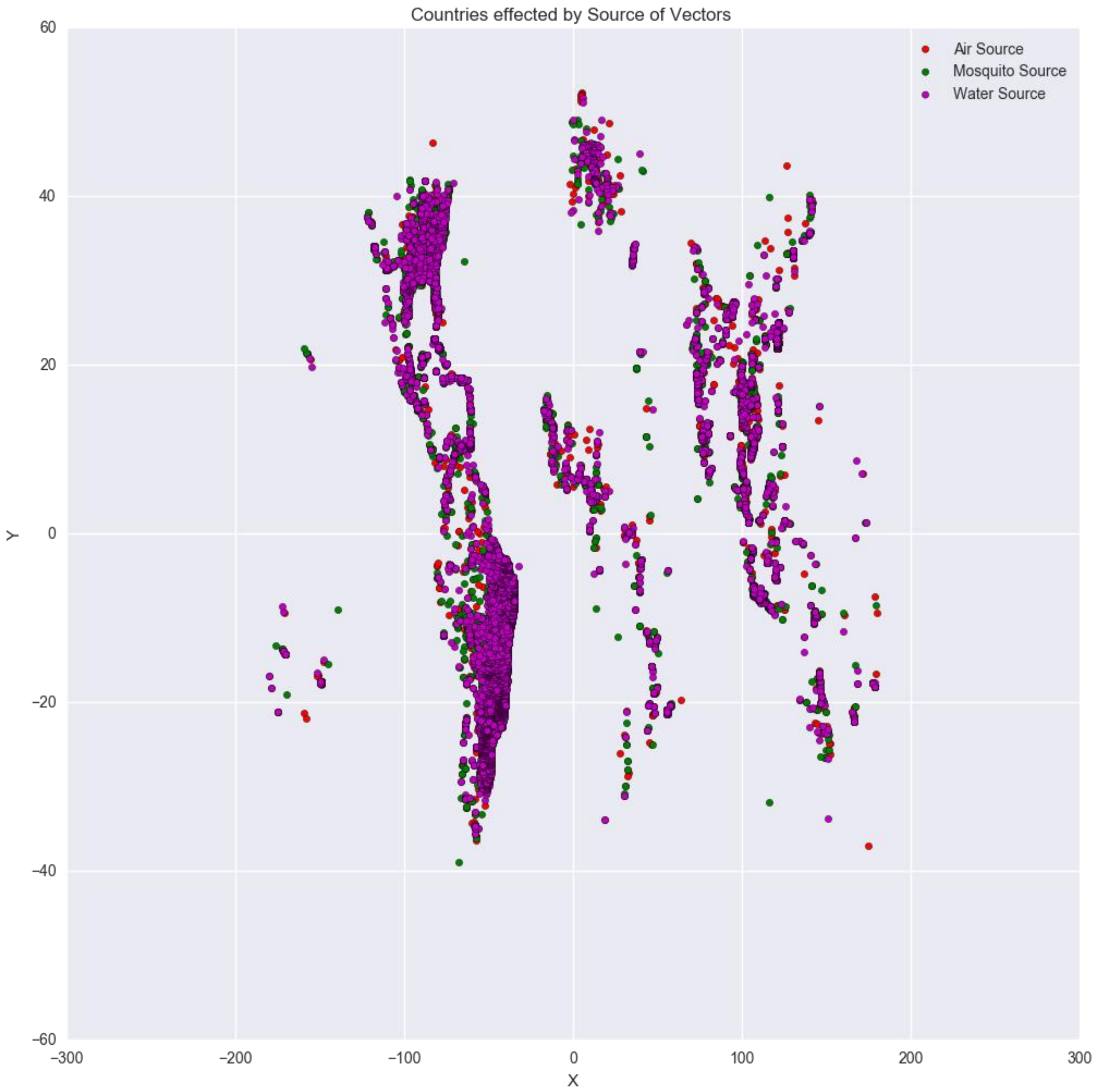
1. The Vector, infection-time ,infection-source has a direct relation with the disease occurrence as from the visualization graphs and how they have effected over the time.
   Aedes aegypti - Dengue
   Ae. Albopictus –Chickungunya

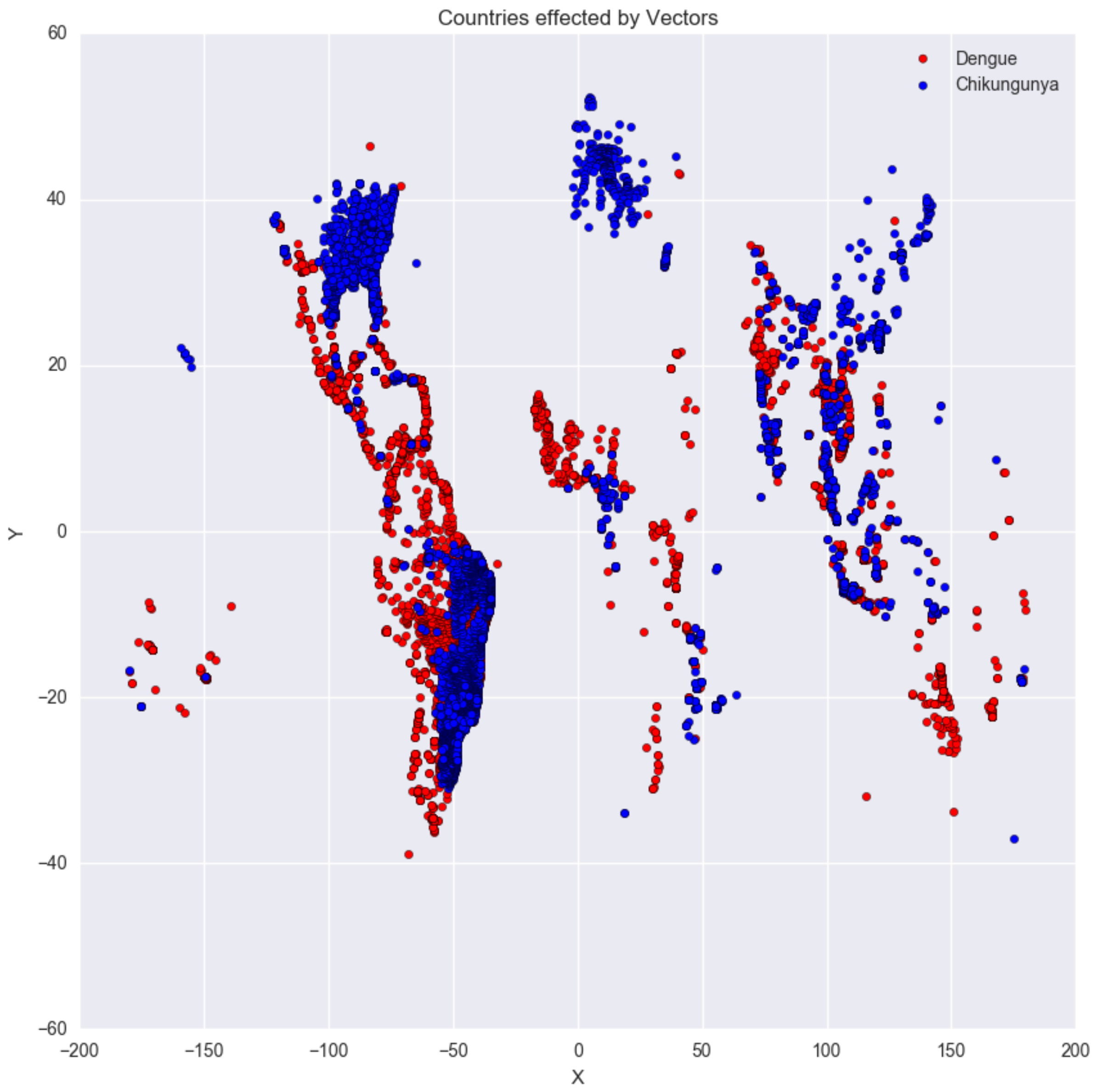

45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Unnamed: 0
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
OCCURRENCE_ID
60
40
20
0
−20
−40
Y
200
150
100
50
0
−50
−100
−150
−200
X
2020
2010
2000
1990
1980
1970
1960
1950
YEAR
1200000
1000000
800000
600000
400000
200000
0
GAUL_AD0
10
9
8
7
6
5
4
3
2
1
infection_time
2.0
1.5
1.0
0.5
0.0
infection_source_ID
1.0
0.8
0.6
0.4
0.2
0.0
VECTOR_ID
160
140
120
100
80
60
40
20
0
COUNTRYID

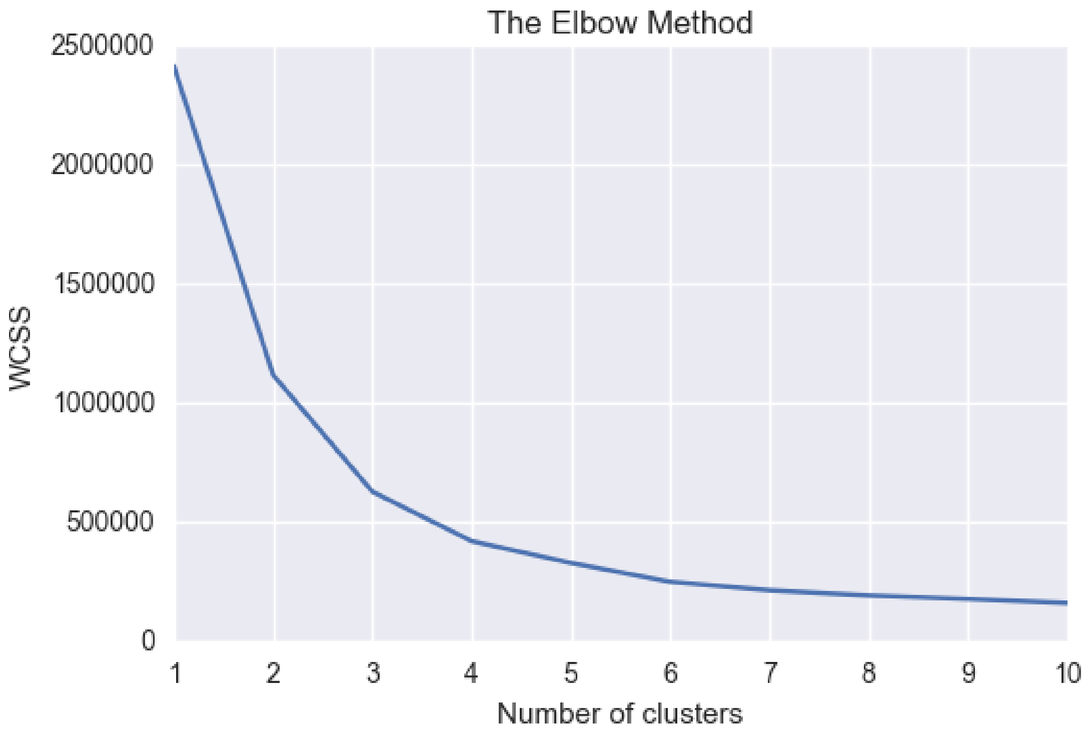

Feature heatmap
Unnamed: 0
OCCURRENCE_ID
Y
X
YEAR
GAUL_AD0
infection_time
infection_source_ID
VECTOR_ID
COUNTRYID
Unnamed: 0
OCCURRENCE_ID
Y
X
YEAR
GAUL_AD0
infection_time
infection_source_ID
VECTOR_ID
COUNTRYID
0.8
0.4
0.0
−0.4
−0.8


Dengue Vs Chikungunya
Total cases
25000
20000
15000
10000
5000
0
1
0
0-Dengue 1-Chikungunya

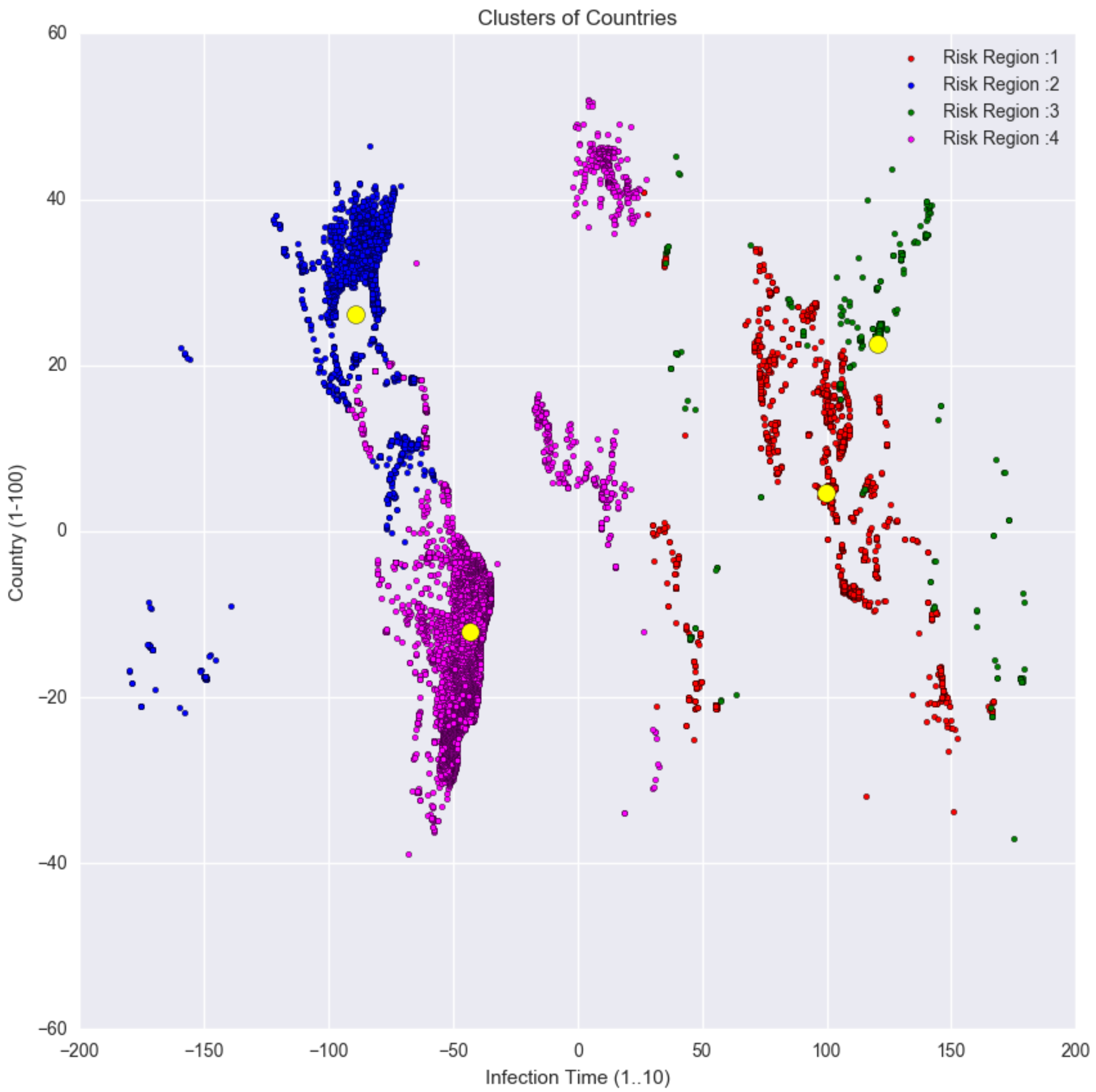

2500
2000
1500
1000
500
0
air
1960
1970
1980
1990
2000
2010
YEAR
mosquito
water


Countries effected by Source of Vectors
Air Source
Mosquito Source
Water Source
60
40
20
0
−20
−40
−60
Y
−300
−200
−100
0
100
200
300
X


Countries effected by Vectors
Dengue
Chikungunya
Y
X
60
40
20
0
−20
−40
−60
−200
−150
−100
−50
0
50
100
150
200



The Elbow Method
WCSS
2500000
2000000
1500000
1000000
500000
0
1
2
3
4
5
6
7
8
9
10
Number of clusters


Clusters of Countries
Risk Region :1
Risk Region :2
Risk Region :3
Risk Region :4
Country (1-100)
Infection Time (1..10)
60
40
20
0
−20
−40
−60
−200
−150
−100
−50
0
50
100
150
200