K-Means Clustering - Data Exploration

UN Data on Countries of the World

We are going to explore or dataset which we get in a csv format but may have missing values. We need to be able to drill down on useful dimensions to explore after cleaning up the data. Since we only have one observation per country, we may not have the option to use columns where there are many missing values as we are effectively going to drop many countries when we drop rows with missing values. But then how did we drop such rows before? Because in those cases there were many observations per individual entity and dropping some did not eliminate an entity altogether.

So first we import the data and explore the columns and types - this time rather than doing it manually we are going to use the facilities in our software to do that.

```
df = read.csv('./datasets/UN.csv')

# print the raw column information plus summary header
head(df)
```

```
##          country region  tfr contraception educationMale educationFemale
## 1    Afghanistan   Asia 6.90            NA            NA              NA
## 2        Albania Europe 2.60            NA            NA              NA
## 3        Algeria Africa 3.81            52          11.1             9.9
## 4 American.Samoa   Asia   NA            NA            NA              NA
## 5        Andorra Europe   NA            NA            NA              NA
## 6         Angola Africa 6.69            NA            NA              NA
##   lifeMale lifeFemale infantMortality GDPperCapita economicActivityMale
## 1     45.0       46.0             154         2848                 87.5
## 2     68.0       74.0              32          863                   NA
## 3     67.5       70.3              44         1531                 76.4
## 4     68.0       73.0              11           NA                 58.8
## 5       NA         NA              NA           NA                   NA
## 6     44.9       48.1             124          355                   NA
##   economicActivityFemale illiteracyMale illiteracyFemale
## 1                    7.2         52.800            85.00
## 2                     NA             NA               NA
## 3                    7.8         26.100            51.00
## 4                   42.4          0.264             0.36
## 5                     NA             NA               NA
## 6                     NA             NA               NA
```

```
# look at the types of each column explicitly
print('Individual columns - R data types')
```

```
## [1] "Individual columns - R data types"
```

```
str(df)
```

```
## 'data.frame':    207 obs. of  14 variables:
##  $ country              : Factor w/ 207 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ region               : Factor w/ 5 levels "Africa","America",..: 3 4 1 3 4 1 2 2 4 5 ...
##  $ tfr                  : num  6.9 2.6 3.81 NA NA 6.69 NA 2.62 1.7 1.89 ...
##  $ contraception        : int  NA NA 52 NA NA NA 53 NA 22 76 ...
##  $ educationMale        : num  NA NA 11.1 NA NA NA NA NA NA 16.3 ...
```

```
##  $ educationFemale       : num  NA NA 9.9 NA NA NA NA NA NA 16.1 ...
##  $ lifeMale              : num  45 68 67.5 68 NA 44.9 NA 69.6 67.2 75.4 ...
##  $ lifeFemale            : num  46 74 70.3 73 NA 48.1 NA 76.8 74 81.2 ...
##  $ infantMortality       : int  154 32 44 11 NA 124 24 22 25 6 ...
##  $ GDPperCapita          : int  2848 863 1531 NA NA 355 6966 8055 354 20046 ...
##  $ economicActivityMale  : num  87.5 NA 76.4 58.8 NA NA 74.4 76.2 65 74 ...
##  $ economicActivityFemale: num  7.2 NA 7.8 42.4 NA NA 56.2 41.3 52 53.8 ...
##  $ illiteracyMale        : num  52.8 NA 26.1 0.264 NA NA NA 3.8 0.3 NA ...
##  $ illiteracyFemale      : num  85 NA 51 0.36 NA NA NA 3.8 0.5 NA ...
```

Here we see that we have 14 columns with country and region being string types and the rest being floats. We also see that the country column has 207 values, ie this is data on 207 countries. The region columns also has 207 entries, but the rest of the columns have many missing entries, indicated by number of non-null values less than 207.

We see that tfr, lifeMale, lifeFemale and GDP, and infantMortality are the columns closest to 207. That is, if we use these columns we will only drop a few countries and not whole clusters as we might if we used educationMale and educationFemale. On the other hand were we to use educationMale and educatonFemale we would have to drop almost 2/3 of the data. So we focus on the columns with non-null values close to 207.

So our short list is now, country, region, tfr, lifeMale, lifeFemale and GDP, and infantMortality.

We suspect that there is clustering of lifeMale, lifeFemale and infantMortality according to GDP and we are going to pull out the heavy machinery of K-Means sofwtare to analyse this in detail and look at the clusters.

We don't know in advance how many clusters there will be which is different from the iris example where we had a 'species' label and there were three unique species.

So while using our KMeans software we will also look at some analytical measures to decide what the right number of clusters might be after looking at multiple such possibilities from 1 through 10 candidate clusters.

Finally, to be able to apply the KMeans modeling software we convert each field in our file to a scientific float format that the numerical algorithms expect.

Onward!