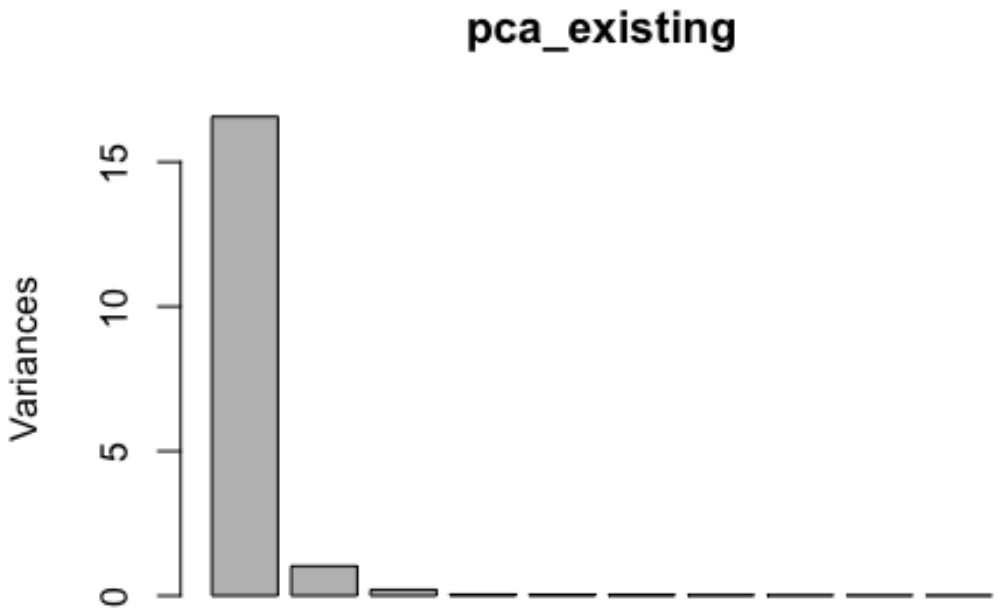
PURVACHAR_HW9_R.R

rashmimahesh

Mon Dec 3 16:09:48 2018

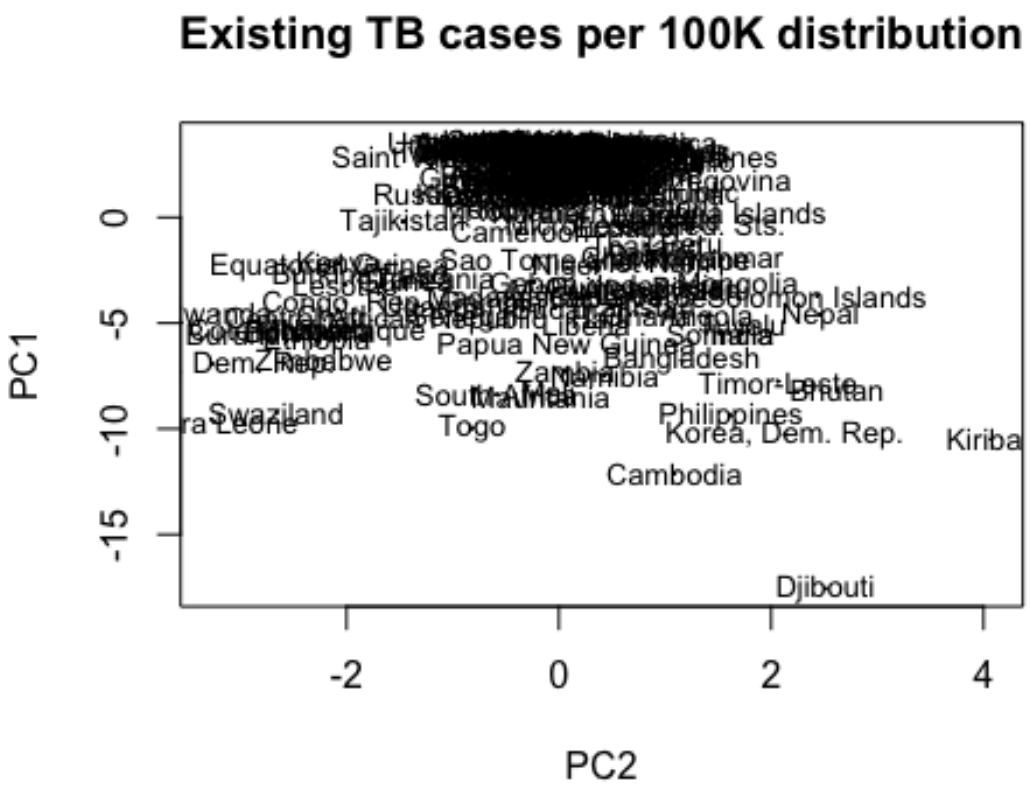
```
library(RCurl)

## Loading required package: bitops

existing_cases_file <-
getURL("https://docs.google.com/spreadsheets/d/1X5Jp7Q8pTs3KLJ5JBWKhncVACGsg5
v4xu6badNs4C7I/pub?gid=0&output=csv")
existing_df <- read.csv(text = existing_cases_file, row.names=1,
stringsAsFactor=F)
existing_df[c(1,2,3,4,5,6,15,16,17,18)] <-
  lapply( existing_df[c(1,2,3,4,5,6,15,16,17,18)],
          function(x) { as.integer(gsub(',', '', x) )})

pca_existing <- prcomp(existing_df, scale. = TRUE)

plot(pca_existing)
```



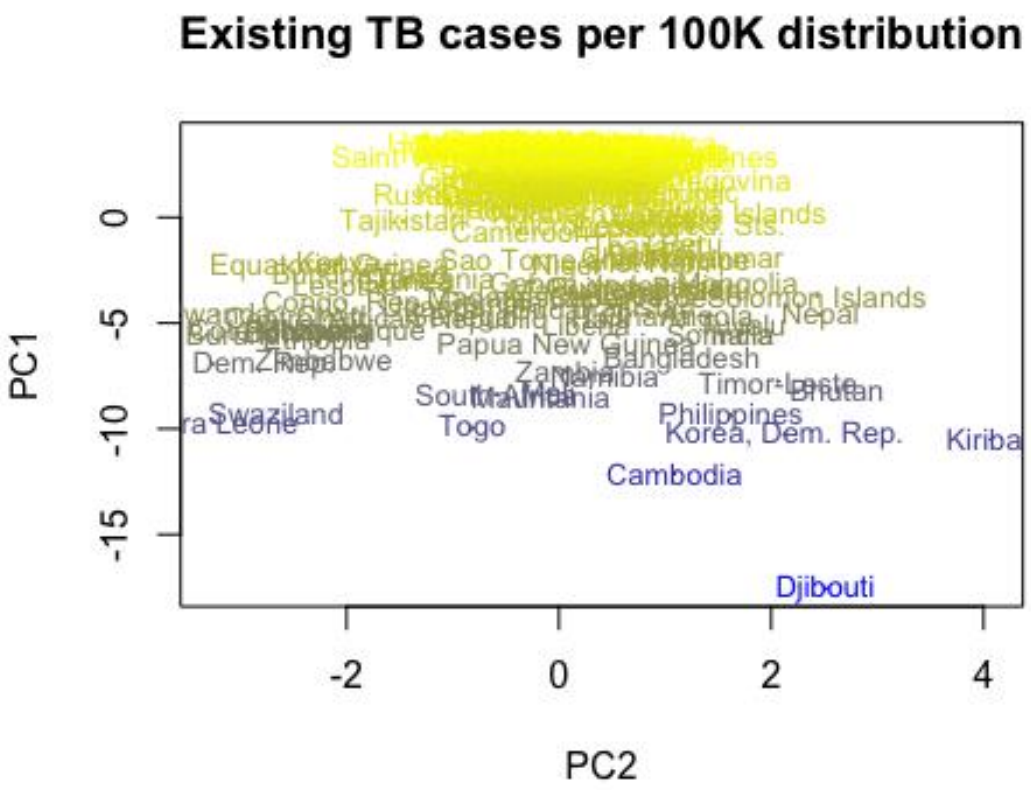

```
scores_existing_df <- as.data.frame(pca_existing$x)
# Show first two PCs for head countries
head(scores_existing_df[1:2])
```

```
##                      PC1          PC2
## Afghanistan    -3.490274  0.973495650
## Albania         2.929002  0.012141345
## Algeria         2.719073 -0.184591877
## American Samoa  3.437263  0.005609367
## Andorra         3.173621  0.033839606
## Angola         -4.695625  1.398306461
```

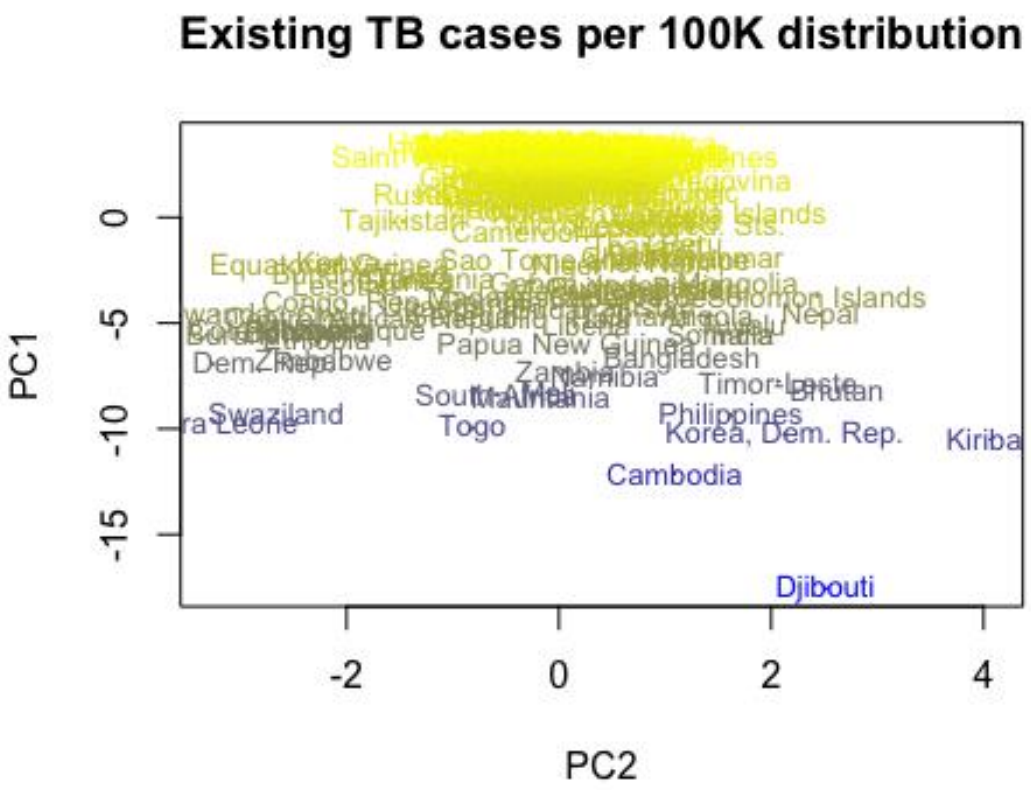
```
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid")
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8)
```



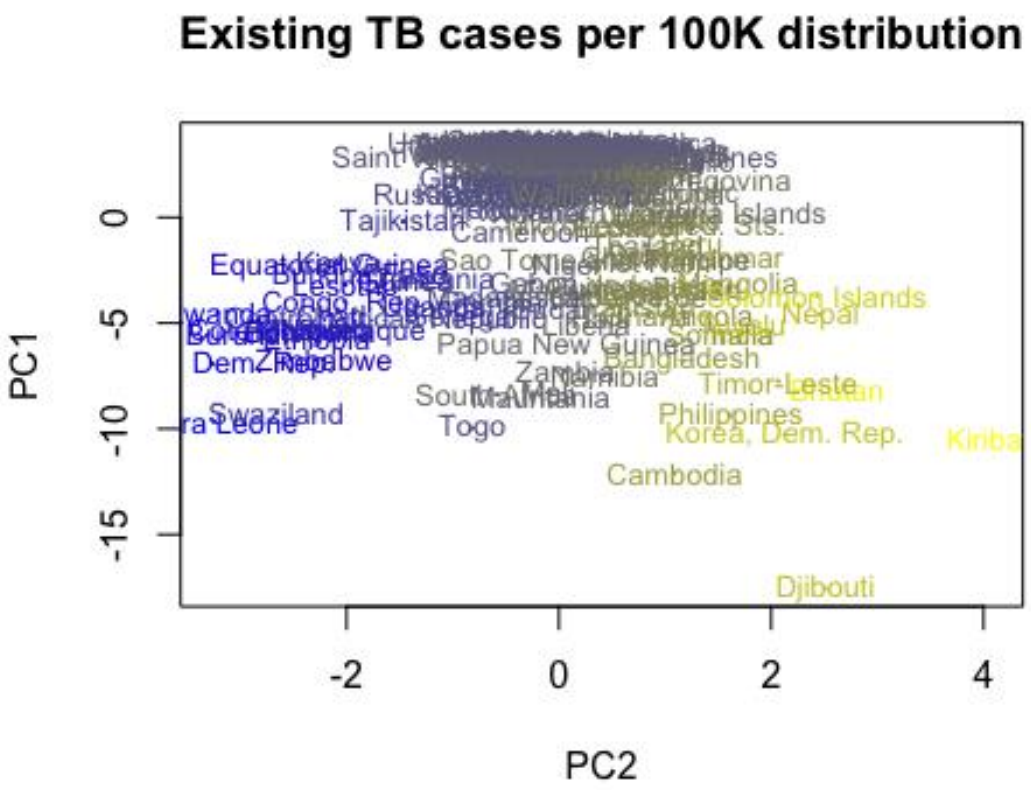

```
library(scales)
ramp <- colorRamp(c("yellow", "blue"))
colours_by_mean <- rgb(
  ramp( as.vector(rescale(rowMeans(existing_df),c(0,1)))),
  max = 255 )
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=colours_by_mean)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=colours_by_mean)
```



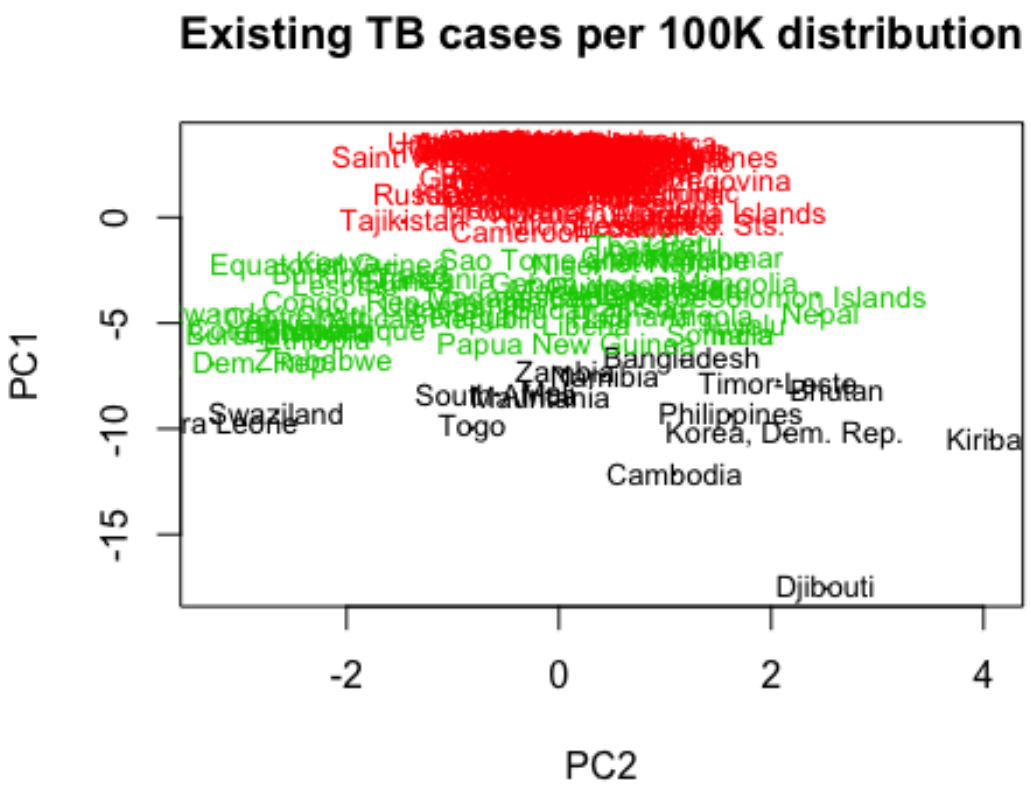

```
ramp <- colorRamp(c("yellow", "blue"))
colours_by_sum <- rgb(
  ramp( as.vector(rescale(rowSums(existing_df),c(0,1)))),
  max = 255 )
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=colours_by_sum)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=colours_by_sum)
```



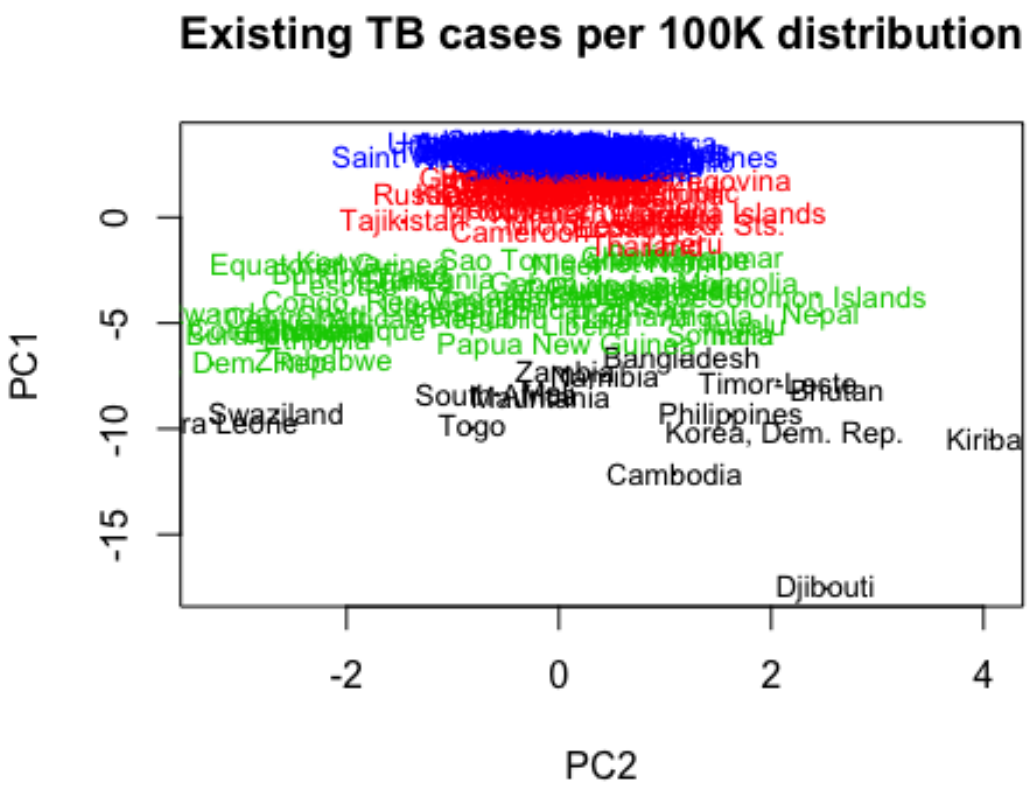

```
existing_df_change <- existing_df$X2007 - existing_df$X1990
ramp <- colorRamp(c("yellow", "blue"))
colours_by_change <- rgb(
  ramp( as.vector(rescale(existing_df_change,c(0,1)))),
  max = 255 )
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=colours_by_change)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=colours_by_change)
```



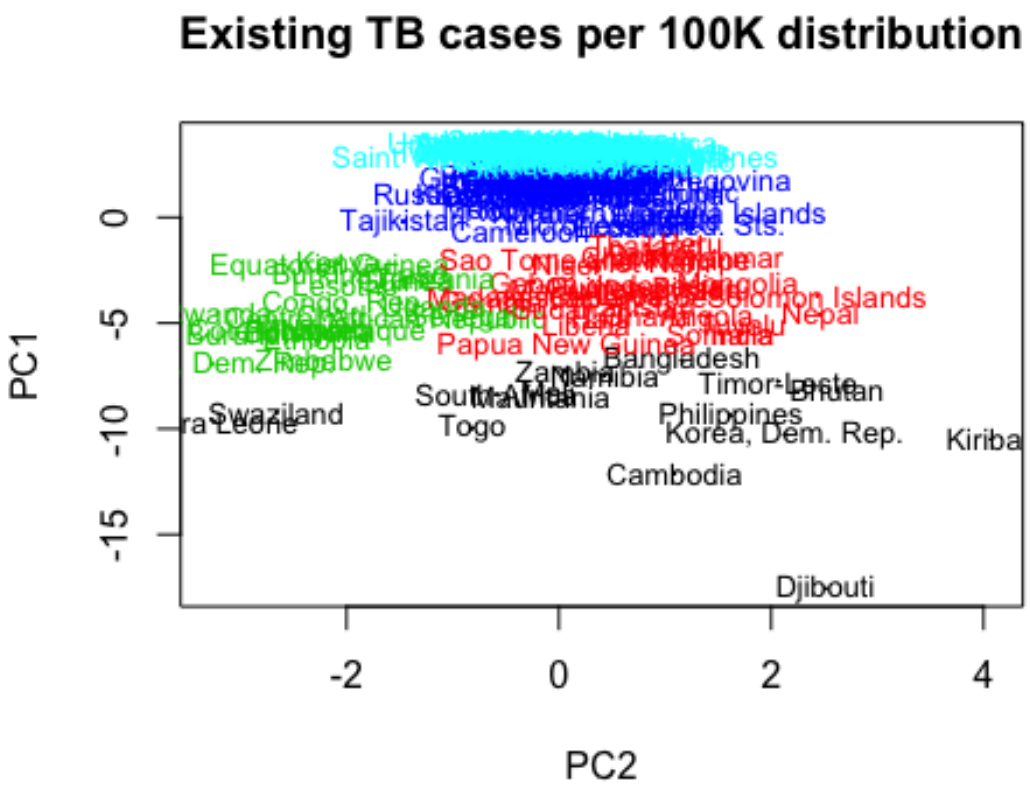

```r
# K clustering

set.seed(1234)
existing_clustering <- kmeans(existing_df, centers = 3)

existing_cluster_groups <- existing_clustering$cluster
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=existing_cluster_groups)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=existing_cluster_groups)
```



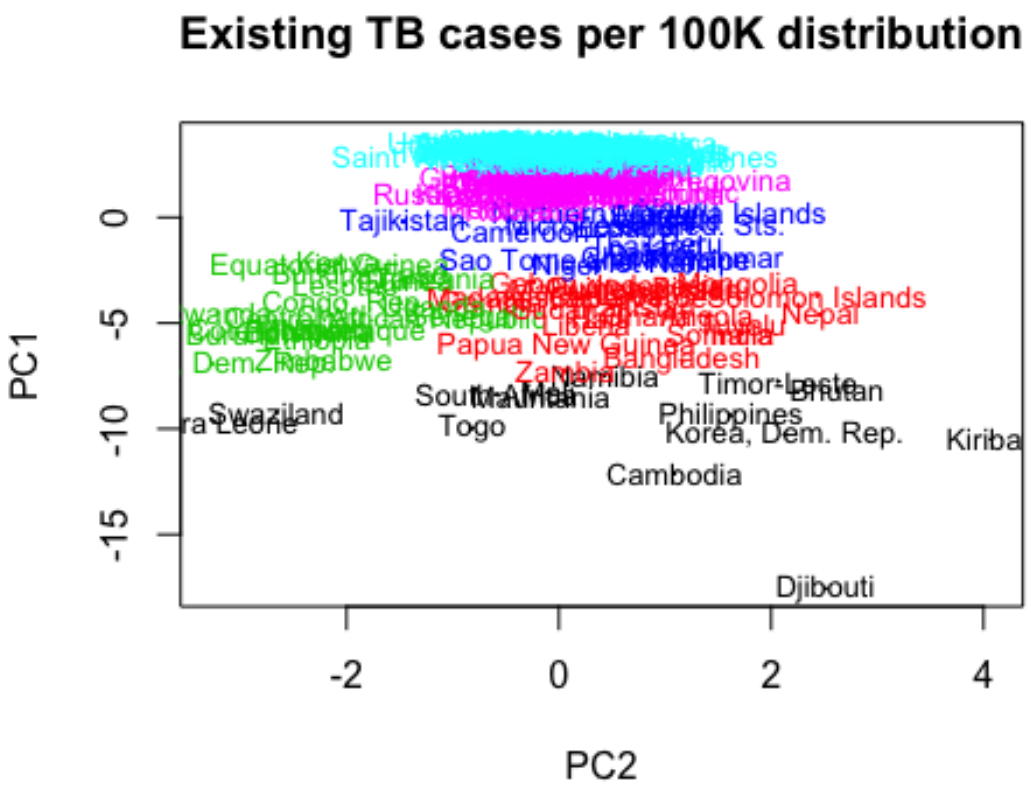

```r
set.seed(1234)
existing_clustering <- kmeans(existing_df, centers = 4)
existing_cluster_groups <- existing_clustering$cluster
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=existing_cluster_groups)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=existing_cluster_groups)
```



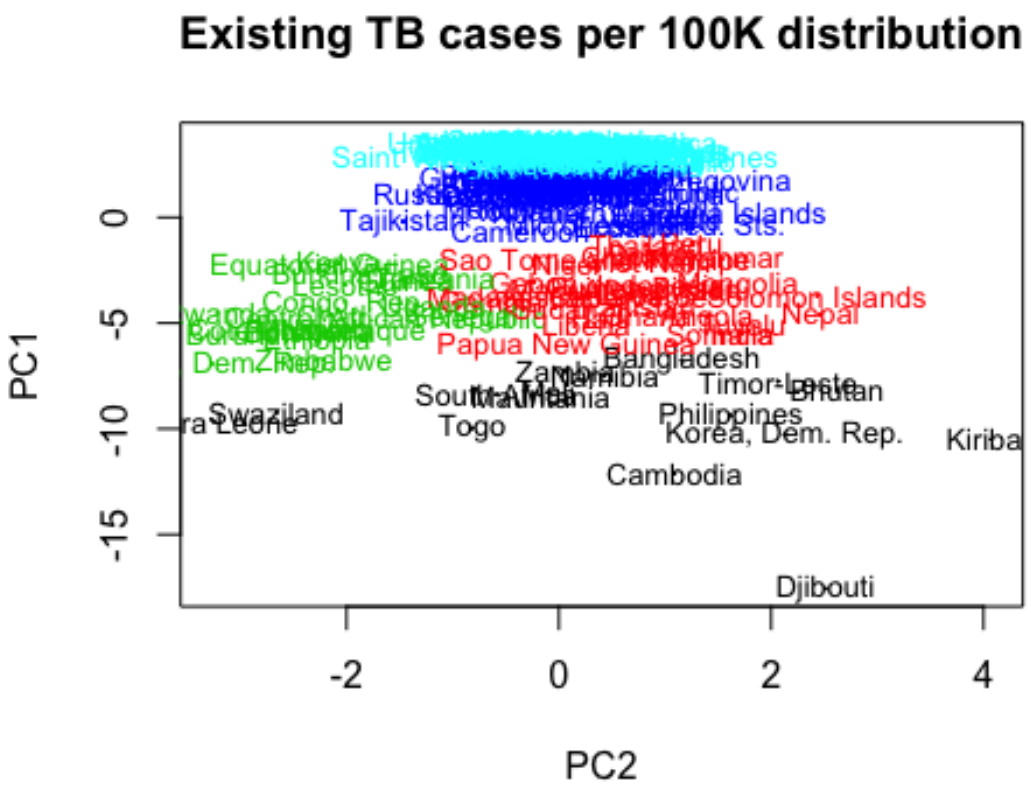

```r
set.seed(1234)
existing_clustering <- kmeans(existing_df, centers = 5)
existing_cluster_groups <- existing_clustering$cluster
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=existing_cluster_groups)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=existing_cluster_groups)
```




```r
set.seed(1234)
existing_clustering <- kmeans(existing_df, centers = 6)
existing_cluster_groups <- existing_clustering$cluster
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=existing_cluster_groups)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=existing_cluster_groups)
```

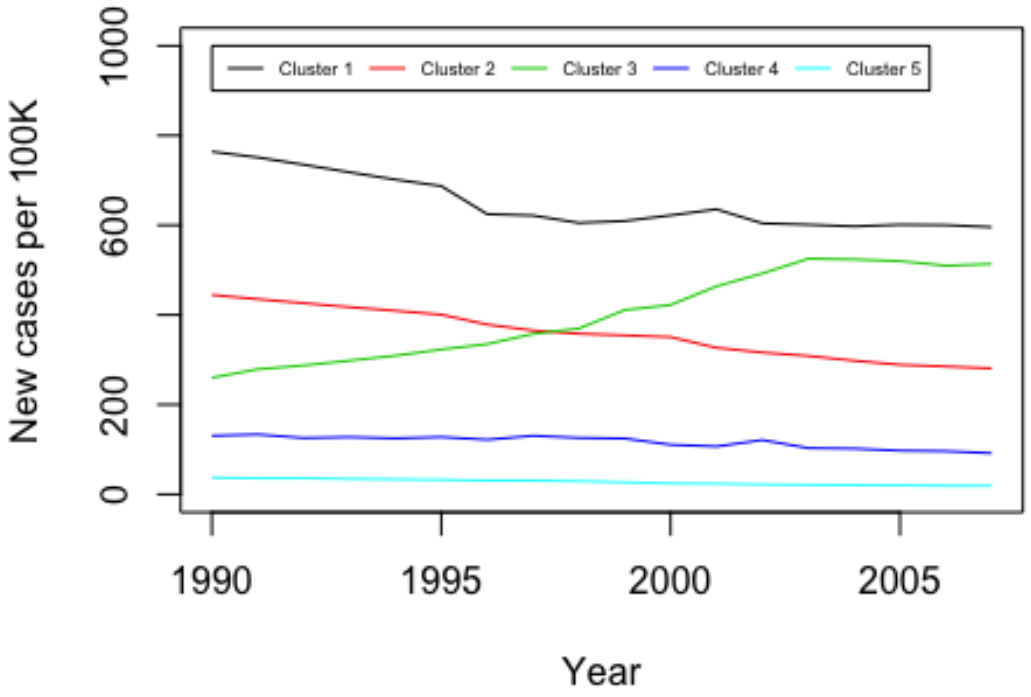



```r
set.seed(1234)
existing_clustering <- kmeans(existing_df, centers = 5)
existing_cluster_groups <- existing_clustering$cluster
plot(PC1~PC2, data=scores_existing_df,
     main= "Existing TB cases per 100K distribution",
     cex = .1, lty = "solid", col=existing_cluster_groups)
text(PC1~PC2, data=scores_existing_df,
     labels=rownames(existing_df),
     cex=.8, col=existing_cluster_groups)
```




```r
#Cluster Interpretation

existing_df$cluster <- existing_clustering$cluster
table(existing_df$cluster)

##
##  1  2  3  4  5
## 16 30 20 51 90

xrange <- 1990:2007
plot(xrange, existing_clustering$centers[1,],
     type='l', xlab="Year",
     ylab="New cases per 100K",
     col = 1,
     ylim=c(0,1000))
for (i in 2:nrow(existing_clustering$centers)) {
  lines(xrange, existing_clustering$centers[i,],
        col = i)
}
legend(x=1990, y=1000,
       lty=1, cex = 0.5,
       ncol = 5,
```

```
       col=1:(nrow(existing_clustering$centers)+1),
       legend=paste("Cluster",1:nrow(existing_clustering$centers)))
```




```
# Cluster 1
# Cluster 1 contains just 16 countries. These are:
rownames(subset(existing_df, cluster==1))
```

```
##  [1] "Bangladesh"       "Bhutan"           "Cambodia"
##  [4] "Korea, Dem. Rep." "Djibouti"         "Kiribati"
##  [7] "Mali"             "Mauritania"       "Namibia"
## [10] "Philippines"      "Sierra Leone"     "South Africa"
## [13] "Swaziland"        "Timor-Leste"      "Togo"
## [16] "Zambia"
```

```
# The centroid that represents them is:
existing_clustering$centers[1,]
```

```
##    X1990    X1991    X1992    X1993    X1994    X1995    X1996    X1997
## 764.0000 751.1875 734.9375 718.0625 701.6875 687.3125 624.7500 621.6250
##    X1998    X1999    X2000    X2001    X2002    X2003    X2004    X2005
## 605.1875 609.4375 622.0000 635.5000 604.2500 601.1250 597.3750 601.1250
##    X2006    X2007
## 600.2500 595.7500
```

```r
# Cluster 2
# Cluster 2 contains 30 countries. These are:
rownames(subset(existing_df, cluster==2))
```

```
##  [1] "Afghanistan"           "Angola"
##  [3] "Bolivia"               "Cape Verde"
##  [5] "China"                 "Gabon"
##  [7] "Gambia"                "Ghana"
##  [9] "Guinea-Bissau"         "Haiti"
## [11] "India"                 "Indonesia"
## [13] "Laos"                  "Liberia"
## [15] "Madagascar"            "Malawi"
## [17] "Mongolia"              "Myanmar"
## [19] "Nepal"                 "Niger"
## [21] "Pakistan"              "Papua New Guinea"
## [23] "Peru"                  "Sao Tome and Principe"
## [25] "Solomon Islands"       "Somalia"
## [27] "Sudan"                 "Thailand"
## [29] "Tuvalu"                "Viet Nam"
```

```r
# The centroid that represents them is:
existing_clustering$centers[2,]
```

```
##    X1990    X1991    X1992    X1993    X1994    X1995    X1996    X1997
## 444.5000 435.2000 426.1667 417.4000 409.2333 400.5667 378.6000 365.3667
##    X1998    X1999    X2000    X2001    X2002    X2003    X2004    X2005
## 358.0333 354.4333 350.6000 326.7333 316.1667 308.5000 297.8667 288.8000
##    X2006    X2007
## 284.9667 280.8000
```

```r
# Cluster 3
# This is an important one. Cluster 3 contains just 20 countries. These are:
rownames(subset(existing_df, cluster==3))
```

```
##  [1] "Botswana"                "Burkina Faso"
##  [3] "Burundi"                 "Central African Republic"
##  [5] "Chad"                    "Congo, Rep."
##  [7] "Cote d'Ivoire"           "Congo, Dem. Rep."
##  [9] "Equatorial Guinea"       "Ethiopia"
## [11] "Guinea"                  "Kenya"
## [13] "Lesotho"                 "Mozambique"
## [15] "Nigeria"                 "Rwanda"
## [17] "Senegal"                 "Uganda"
## [19] "Tanzania"                "Zimbabwe"
```

```r
# The centroid that represents them is:
existing_clustering$centers[3,]
```

```
##  X1990  X1991  X1992  X1993  X1994  X1995  X1996  X1997  X1998  X1999
## 259.85 278.90 287.30 298.05 309.00 322.95 335.00 357.65 369.65 410.85
```

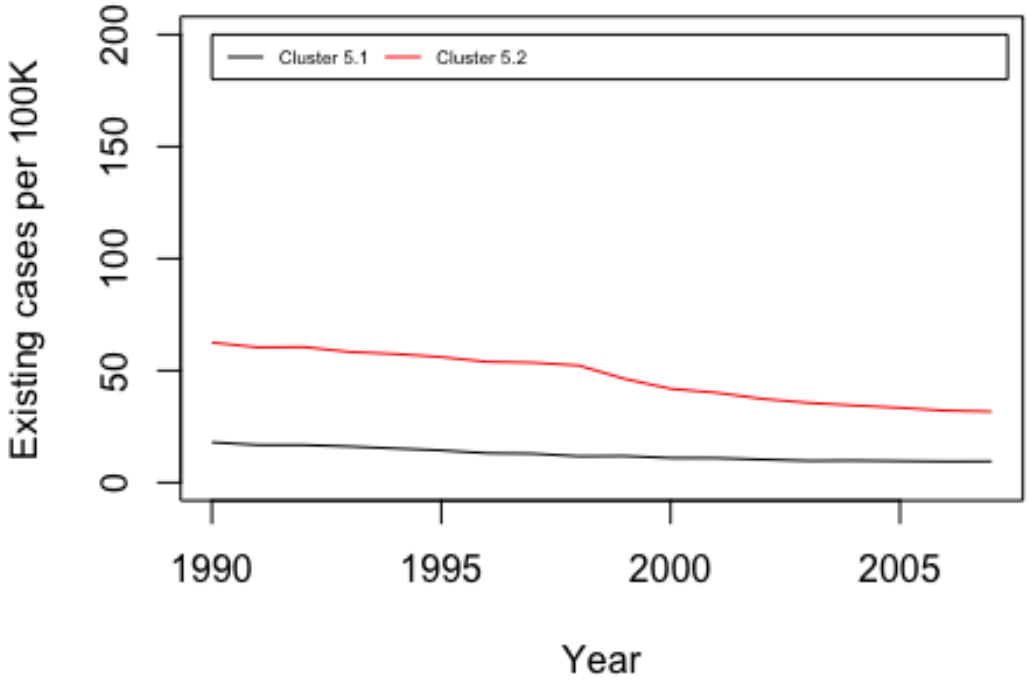
```
##  X2000  X2001  X2002  X2003  X2004  X2005  X2006  X2007
## 422.25 463.75 492.45 525.25 523.60 519.90 509.80 513.50

# Cluster 4
# The fourth cluster contains 51 countries.
rownames(subset(existing_df, cluster==4))

##  [1] "Armenia"                  "Azerbaijan"
##  [3] "Bahrain"                  "Belarus"
##  [5] "Benin"                    "Bosnia and Herzegovina"
##  [7] "Brazil"                   "Brunei Darussalam"
##  [9] "Cameroon"                 "Comoros"
## [11] "Croatia"                  "Dominican Republic"
## [13] "Ecuador"                  "El Salvador"
## [15] "Eritrea"                  "Georgia"
## [17] "Guam"                     "Guatemala"
## [19] "Guyana"                   "Honduras"
## [21] "Iraq"                     "Kazakhstan"
## [23] "Kyrgyzstan"               "Latvia"
## [25] "Lithuania"                "Malaysia"
## [27] "Maldives"                 "Micronesia, Fed. Sts."
## [29] "Morocco"                  "Nauru"
## [31] "Nicaragua"                "Niue"
## [33] "Northern Mariana Islands" "Palau"
## [35] "Paraguay"                 "Qatar"
## [37] "Korea, Rep."              "Moldova"
## [39] "Romania"                  "Russian Federation"
## [41] "Seychelles"               "Sri Lanka"
## [43] "Suriname"                 "Tajikistan"
## [45] "Tokelau"                  "Turkmenistan"
## [47] "Ukraine"                  "Uzbekistan"
## [49] "Vanuatu"                  "Wallis et Futuna"
## [51] "Yemen"

# The centroid that represents them is:
existing_clustering$centers[4,]

##     X1990     X1991     X1992     X1993     X1994     X1995     X1996
## 130.60784 133.41176 125.60784 127.54902 124.82353 127.70588 121.68627
##     X1997     X1998     X1999     X2000     X2001     X2002     X2003
## 130.50980 125.82353 124.45098 110.58824 106.60784 121.09804 103.01961
##     X2004     X2005     X2006     X2007
## 101.80392  97.29412  96.17647  91.68627

# Cluster 5
# The last and bigger cluster contains 90 countries.
rownames(subset(existing_df, cluster==5))

##  [1] "Albania"                  "Algeria"
##  [3] "American Samoa"           "Andorra"
##  [5] "Anguilla"                 "Antigua and Barbuda"
```

```
##  [7] "Argentina"                          "Australia"
##  [9] "Austria"                            "Bahamas"
## [11] "Barbados"                           "Belgium"
## [13] "Belize"                             "Bermuda"
## [15] "British Virgin Islands"             "Bulgaria"
## [17] "Canada"                             "Cayman Islands"
## [19] "Chile"                              "Colombia"
## [21] "Cook Islands"                       "Costa Rica"
## [23] "Cuba"                               "Cyprus"
## [25] "Czech Republic"                     "Denmark"
## [27] "Dominica"                           "Egypt"
## [29] "Estonia"                            "Fiji"
## [31] "Finland"                            "France"
## [33] "French Polynesia"                   "Germany"
## [35] "Greece"                             "Grenada"
## [37] "Hungary"                            "Iceland"
## [39] "Iran"                               "Ireland"
## [41] "Israel"                             "Italy"
## [43] "Jamaica"                            "Japan"
## [45] "Jordan"                             "Kuwait"
## [47] "Lebanon"                            "Libyan Arab Jamahiriya"
## [49] "Luxembourg"                         "Malta"
## [51] "Mauritius"                          "Mexico"
## [53] "Monaco"                             "Montserrat"
## [55] "Netherlands"                        "Netherlands Antilles"
## [57] "New Caledonia"                      "New Zealand"
## [59] "Norway"                             "Oman"
## [61] "Panama"                             "Poland"
## [63] "Portugal"                           "Puerto Rico"
## [65] "Saint Kitts and Nevis"              "Saint Lucia"
## [67] "Saint Vincent and the Grenadines"   "Samoa"
## [69] "San Marino"                         "Saudi Arabia"
## [71] "Singapore"                          "Slovakia"
## [73] "Slovenia"                           "Spain"
## [75] "Sweden"                             "Switzerland"
## [77] "Syrian Arab Republic"               "Macedonia, FYR"
## [79] "Tonga"                              "Trinidad and Tobago"
## [81] "Tunisia"                            "Turkey"
## [83] "Turks and Caicos Islands"           "United Arab Emirates"
## [85] "United Kingdom"                     "Virgin Islands (U.S.)"
## [87] "United States of America"           "Uruguay"
## [89] "Venezuela"                          "West Bank and Gaza"
```

```
# The centroid that represents them is:
existing_clustering$centers[5,]
```

```
##    X1990    X1991    X1992    X1993    X1994    X1995    X1996    X1997
## 37.27778 35.68889 35.73333 34.40000 33.51111 32.42222 30.80000 30.51111
##    X1998    X1999    X2000    X2001    X2002    X2003    X2004    X2005
## 29.30000 26.77778 24.35556 23.57778 22.02222 20.93333 20.48889 19.92222
```

```
##    X2006    X2007
## 19.25556 19.11111

# A Second Level of Clustering
# subset the original dataset
cluster5_df <- subset(existing_df, cluster==5)
# do the clustering
set.seed(1234)
cluster5_clustering <- kmeans(cluster5_df[,-19], centers = 2)
# assign sub-cluster number to the data set for Cluster 5
cluster5_df$cluster <- cluster5_clustering$cluster

xrange <- 1990:2007
plot(xrange, cluster5_clustering$centers[1,],
     type='l', xlab="Year",
     ylab="Existing cases per 100K",
     col = 1,
     ylim=c(0,200))
for (i in 2:nrow(cluster5_clustering$centers)) {
  lines(xrange, cluster5_clustering$centers[i,],
        col = i)
}
legend(x=1990, y=200,
       lty=1, cex = 0.5,
       ncol = 5,
       col=1:(nrow(cluster5_clustering$centers)+1),
       legend=paste0("Cluster 5.",1:nrow(cluster5_clustering$centers)))
```




```
rownames(subset(cluster5_df, cluster5_df$cluster==2))
```

```
##  [1] "Albania"                          "Algeria"
##  [3] "Anguilla"                         "Argentina"
##  [5] "Bahamas"                          "Belize"
##  [7] "Bulgaria"                         "Colombia"
##  [9] "Egypt"                            "Estonia"
## [11] "Fiji"                             "French Polynesia"
## [13] "Hungary"                          "Iran"
## [15] "Japan"                            "Kuwait"
## [17] "Lebanon"                          "Libyan Arab Jamahiriya"
## [19] "Mauritius"                        "Mexico"
## [21] "New Caledonia"                    "Panama"
## [23] "Poland"                           "Portugal"
## [25] "Saint Vincent and the Grenadines" "Samoa"
## [27] "Saudi Arabia"                     "Singapore"
## [29] "Slovakia"                         "Slovenia"
## [31] "Spain"                            "Syrian Arab Republic"
## [33] "Macedonia, FYR"                   "Tonga"
## [35] "Tunisia"                          "Turkey"
## [37] "United Arab Emirates"             "Venezuela"
## [39] "West Bank and Gaza"
```

```r
rownames(subset(cluster5_df, cluster5_df$cluster==1))
```

```
##  [1] "American Samoa"           "Andorra"
##  [3] "Antigua and Barbuda"      "Australia"
##  [5] "Austria"                  "Barbados"
##  [7] "Belgium"                  "Bermuda"
##  [9] "British Virgin Islands"   "Canada"
## [11] "Cayman Islands"           "Chile"
## [13] "Cook Islands"             "Costa Rica"
## [15] "Cuba"                     "Cyprus"
## [17] "Czech Republic"           "Denmark"
## [19] "Dominica"                 "Finland"
## [21] "France"                   "Germany"
## [23] "Greece"                   "Grenada"
## [25] "Iceland"                  "Ireland"
## [27] "Israel"                   "Italy"
## [29] "Jamaica"                  "Jordan"
## [31] "Luxembourg"               "Malta"
## [33] "Monaco"                   "Montserrat"
## [35] "Netherlands"              "Netherlands Antilles"
## [37] "New Zealand"              "Norway"
## [39] "Oman"                     "Puerto Rico"
## [41] "Saint Kitts and Nevis"    "Saint Lucia"
## [43] "San Marino"               "Sweden"
## [45] "Switzerland"              "Trinidad and Tobago"
## [47] "Turks and Caicos Islands" "United Kingdom"
## [49] "Virgin Islands (U.S.)"    "United States of America"
## [51] "Uruguay"
```