Comparison of World Country Clusters by HDI and Gender Equality and by Water Resources and Agriculture

Robert Hand and Rachel Jordan

May 2022

Data Cleanup

```
#data
data <- read.csv("water.csv", header = TRUE)

#rename the variables
colnames(data) <- c("country","year","country_area","arable_land","crop_area","pct_area_cultivated","to

#Arable land/country_area*100 = percent arable land
data <- data %>% mutate(pct_arable_land = (arable_land/country_area)*100)


#pct_crop_area
data <- data %>% mutate(pct_crop_area = (crop_area/country_area)*100)

#new data set of variables for water and agriculture clustering.
data_cluster <- data %>% select(pct_arable_land,pct_crop_area,pct_area_cultivated,ag_value_added_pct_gd
```

Imputation of Missing Data

```
#multiple imputation of missing values using mice package
data_cluster <- mice(data_cluster,m=10, seed = 50) #a big list of not useful output.
```

```
##
##  iter imp variable
##   1   1  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   2  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   3  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   4  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   5  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   6  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   7  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   8  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   9  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   1   10  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_
##   2   1  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   2  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   3  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   4  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
```

```
##   2   5  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   6  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   7  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   8  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   9  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   2   10  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_
##   3   1  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   2  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   3  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   4  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   5  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   6  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   7  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   8  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   9  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   3   10  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_
##   4   1  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   2  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   3  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   4  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   5  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   6  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   7  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   8  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   9  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   4   10  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_
##   5   1  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   2  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   3  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   4  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   5  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   6  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   7  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   8  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   9  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_i
##   5   10  pct_arable_land  pct_crop_area  pct_area_cultivated  ag_value_added_pct_gdp  pct_gva_from_
```

```
## Warning: Number of logged events: 300
```

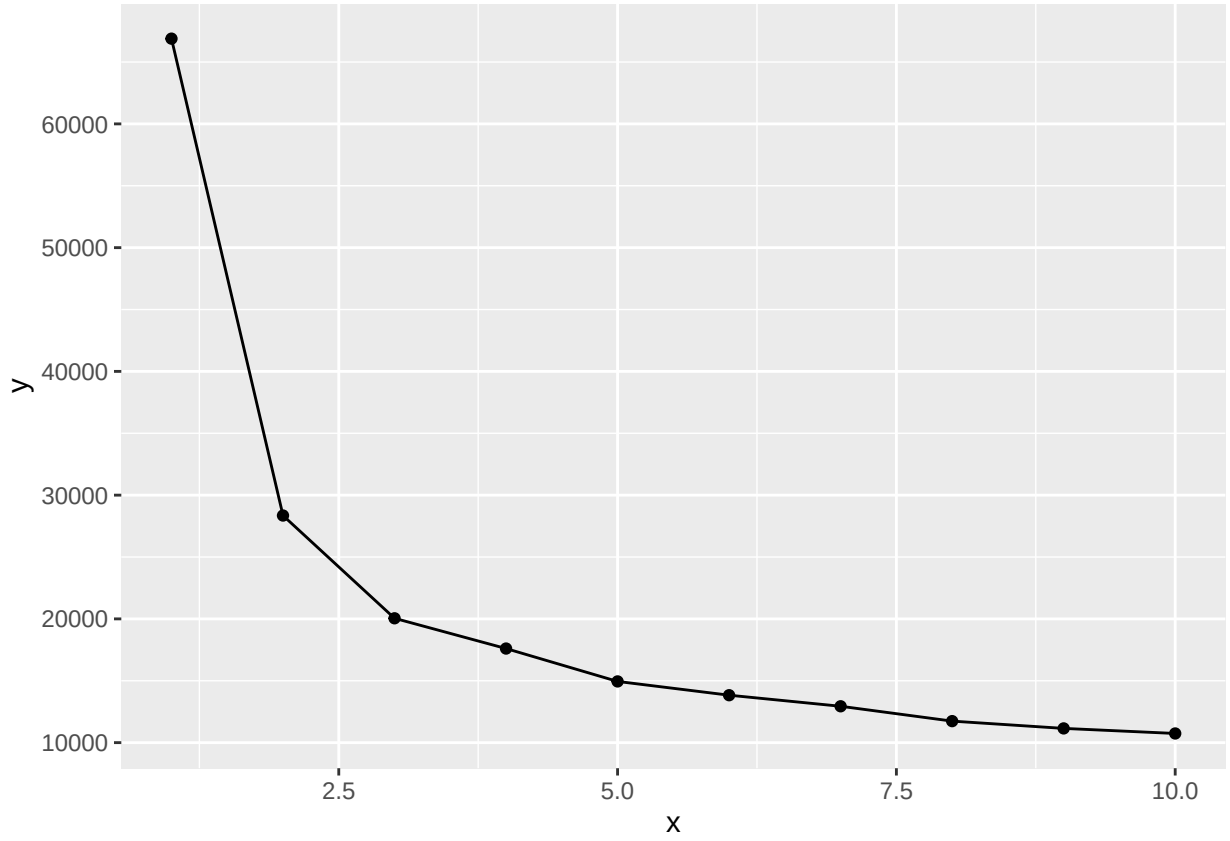
```
data_cluster <- complete(data_cluster,1)
```

Clustering

```
#scaling the data.
data_cluster<- scale(data_cluster,
center = FALSE, scale = TRUE)


#elbow plot to aid in deciding how many centers
wss <- sapply(1:10,
              function(k){kmeans(dist(data_cluster, method = "euclidean"), k)$tot.withinss})
data_frame(x = 1:length(wss), y = wss) %>%
  ggplot(aes(x, y)) +
```

```
  geom_point() +
  geom_line()
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

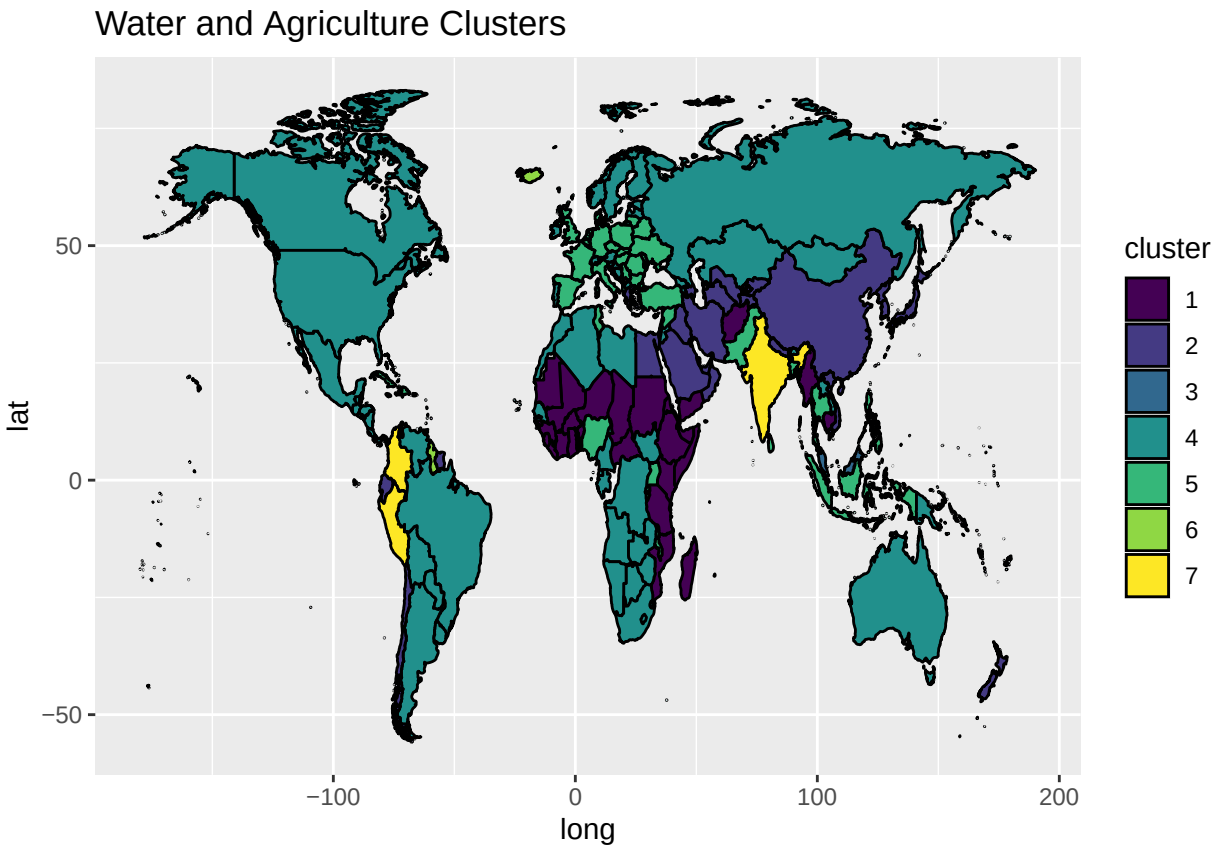



```
#using 7
clusters <- kmeans(data_cluster, centers = 7, nstart = 20)

country_clusters <- cbind(clusters$cluster,data$country)
country_clusters <- as.data.frame(country_clusters)
```

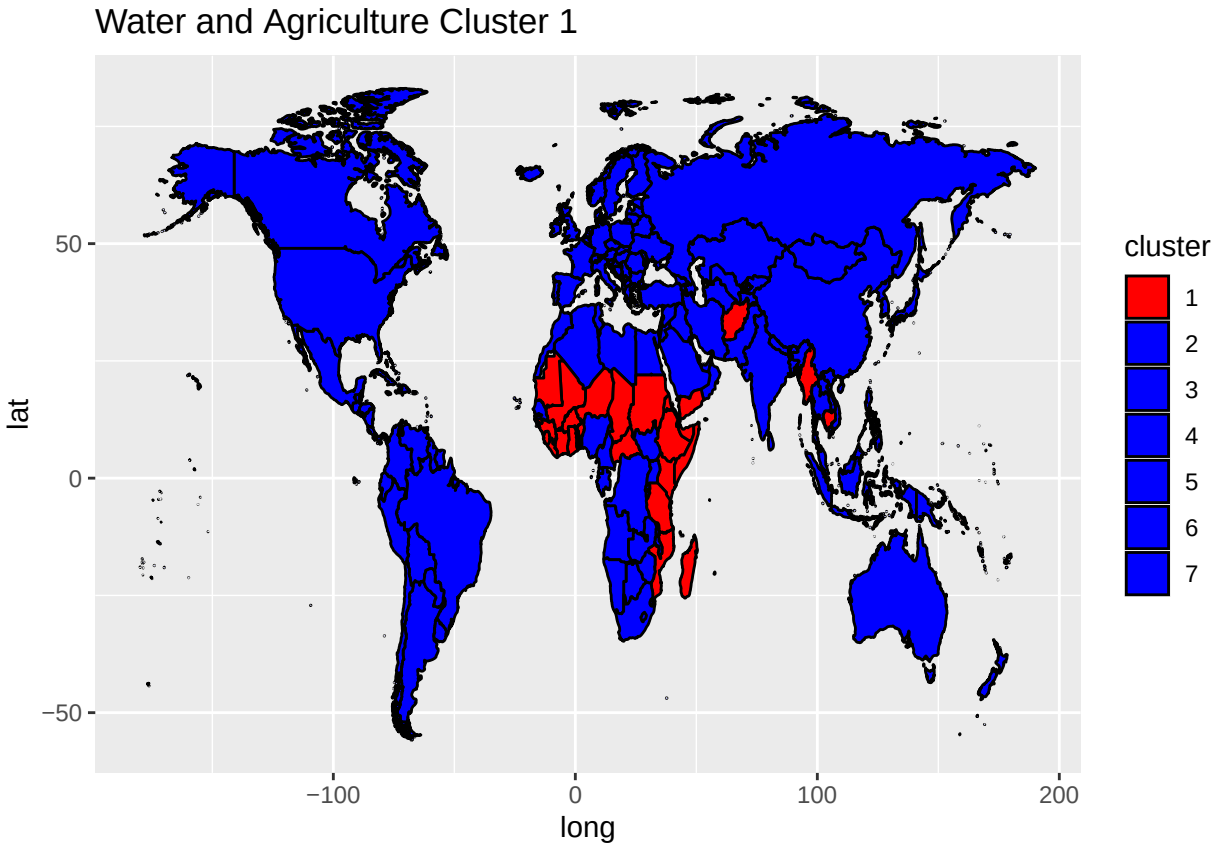
Mapping the clusters

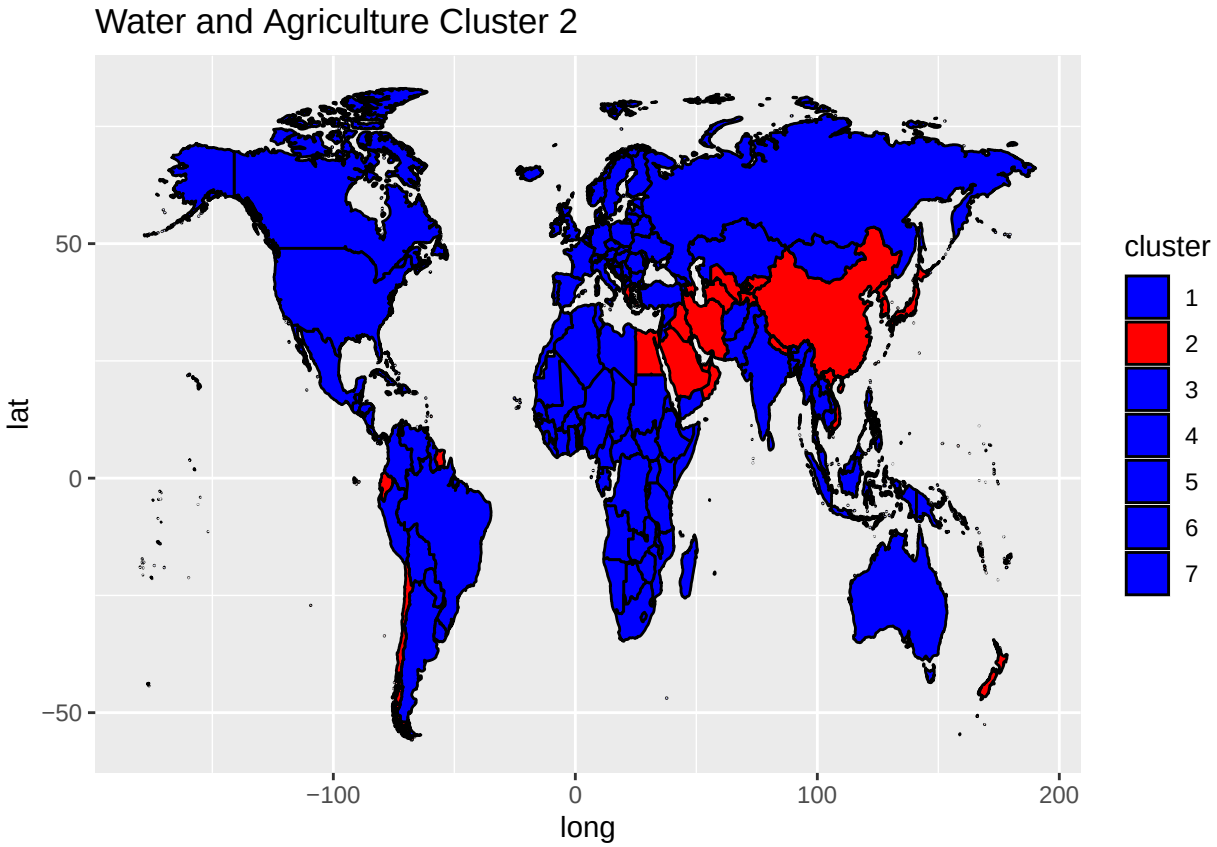
```
#plot clusters. Need to rename some of the countries for the join.

colnames(country_clusters) <- c("cluster","region")
country_clusters <- country_clusters %>% mutate(region = case_when(
  region == "United States of America"~"USA",
  region == "Russian Federation"~"Russia",
  region == "Venezuela (Bolivarian Republic of)"~"Venezuela",
  region == "Bolivia (Plurinational State of)"~"Bolivia",
```

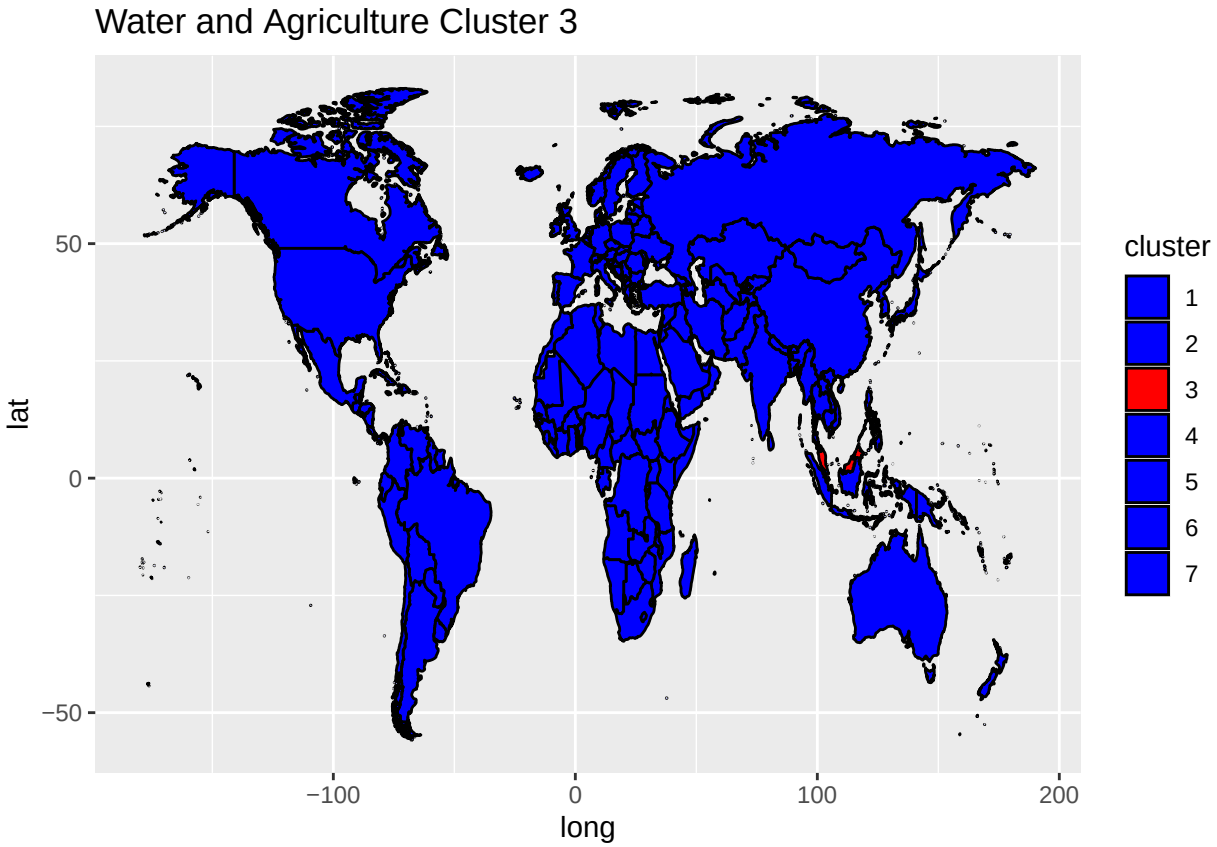
```
  region == "Czechia"~"Czech Republic",
  region == "Iran (Islamic Republic of)"~"Iran",
  region == "Antigua and Barbuda"~"Antigua",
  region == "Brunei Darussalam"~"Brunei",
  region == "Cabo Verde"~"Cape Verde",
  region == "Congo"~"Democratic Republic of the Congo",
  region == "CÃ´te d'Ivoire"~"Ivory Coast",
  region == "Democratic People's Republic of Korea"~"North Korea",
  region == "Eswatini"~"Swaziland",
  region == "Grenade"~"Grenada",
  region == "Holy See"~"Vatican",
  region == "Lao People's Democratic Republic"~"Laos",
  region == "Micronesia (Federated States of)"~"Micronesia",
  region == "Republic of Korea"~"South Korea",
  region == "Republic of Moldova"~"Moldova",
  region == "Saint Kitts and Nevis"~"st Kitts",
  region == "Saint Vincent and the Grenadines"~"Saint Vincent",
  region == "Syrian Arab Republic"~"Syria",
  region == "Trinidad and Tobago"~"Trinidad",
  region == "United Kingdom"~"UK",
  region == "United Republic of Tanzania"~"Tanzania",
  region == "Viet Nam"~"Vietnam",
  TRUE ~region
))


#create a map of the clusters.
map_data <- map_data("world")
map_data1 <- left_join(map_data,country_clusters, by="region")
map_data1 <- map_data1 %>% filter(!is.na(map_data1$cluster))
map_water <- ggplot(map_data1,aes(x=long, y=lat, group=group)) + geom_polygon(aes(fill=cluster), color =

map_water
```

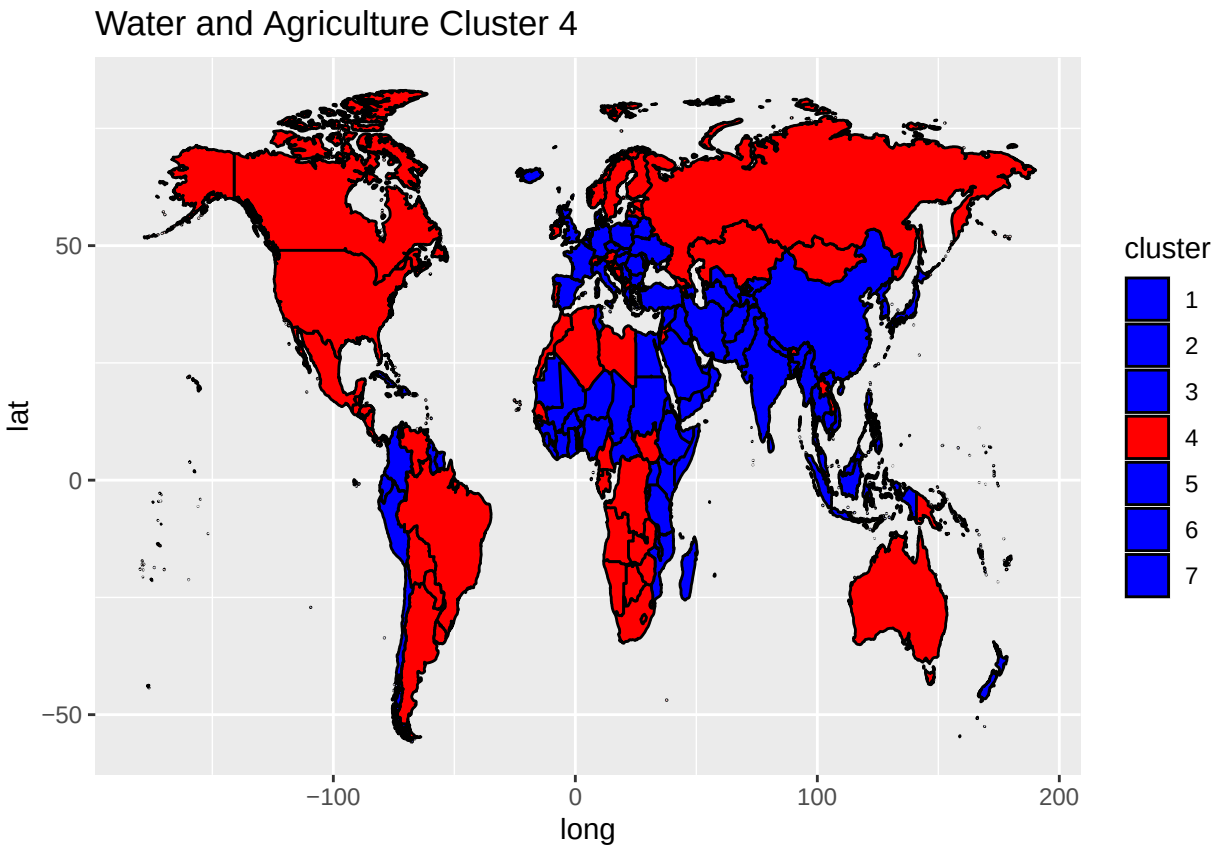


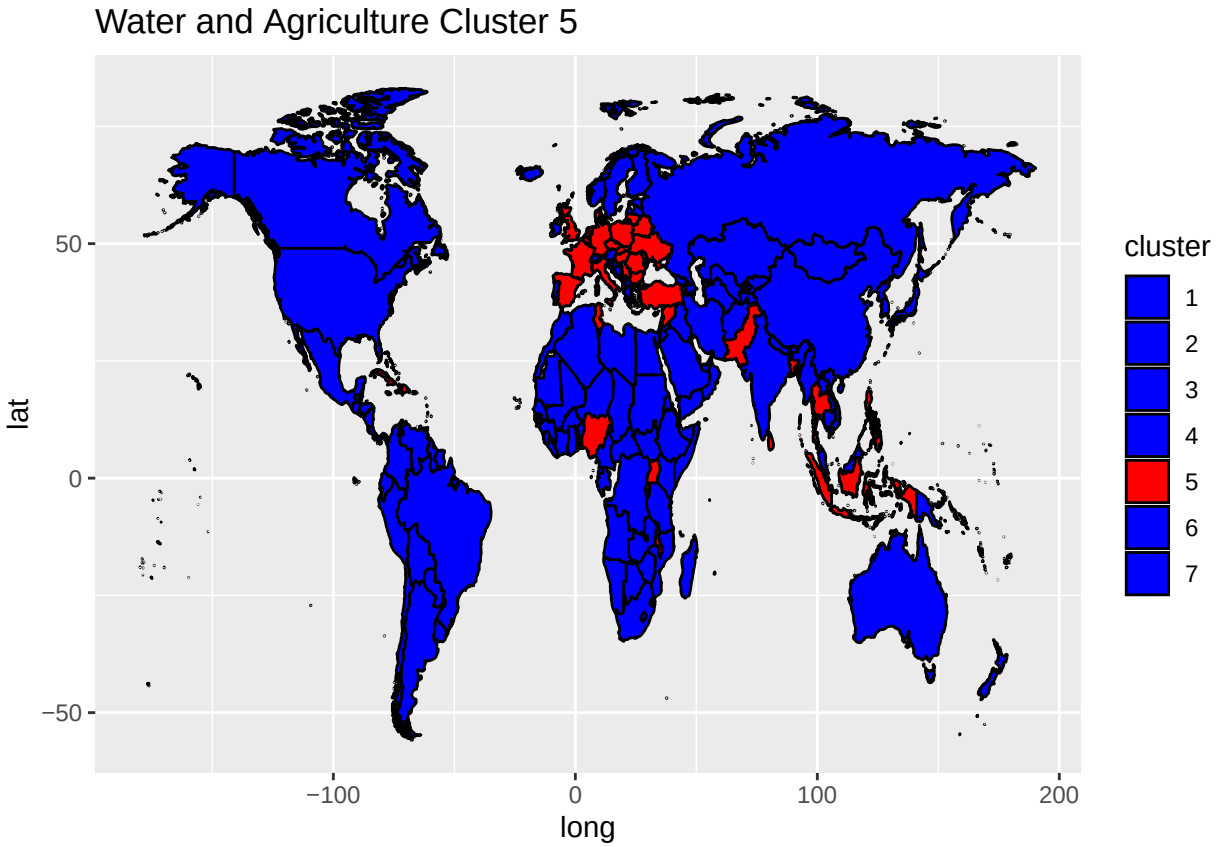

```
#it's also helpful visually to see each cluster in isolation.
x <- c(NA,NA,NA,NA,NA,NA,NA)

for (i in 1:7) {
x[i] <- "red"
x[-i] <- "blue"
print(ggplot(map_data1,aes(x=long, y=lat, group=group)) + geom_polygon(aes(fill=cluster), color = "blac
}
```

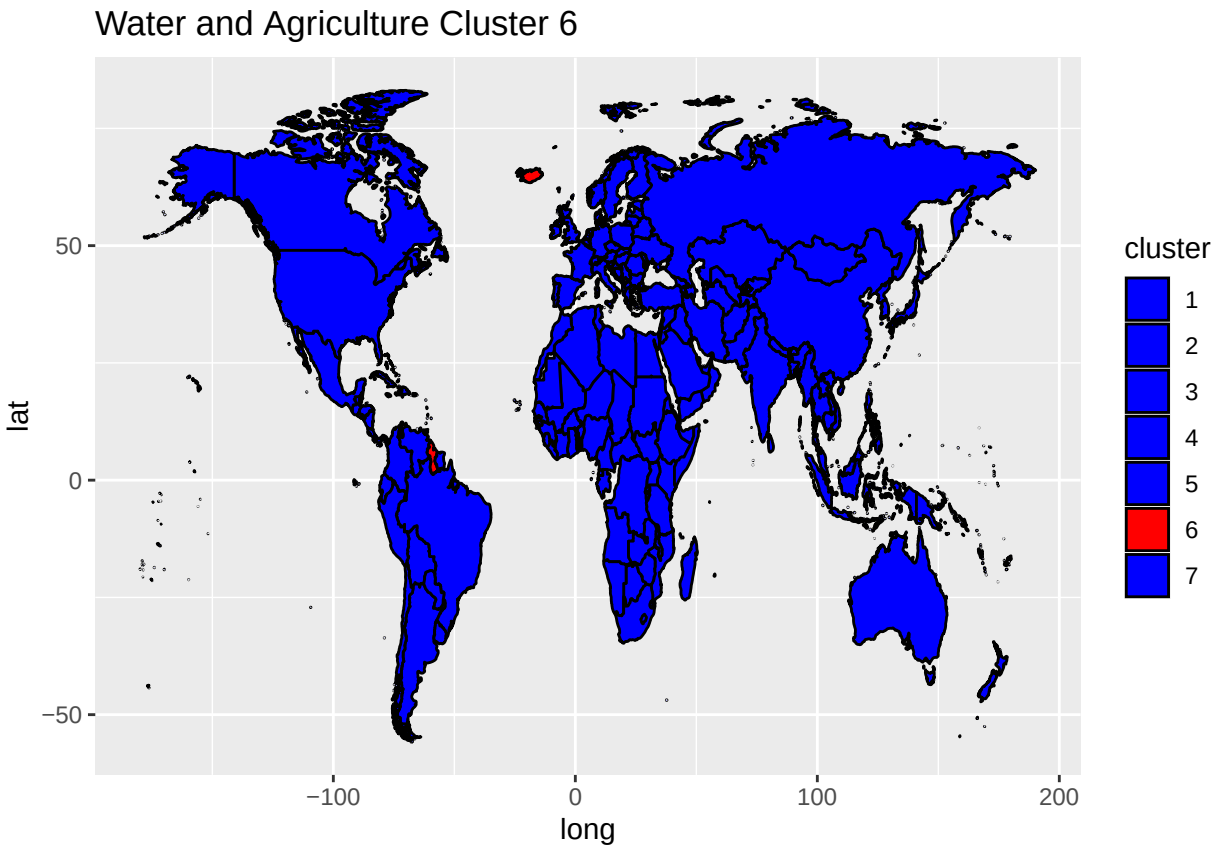

Water and Agriculture Cluster 1
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7

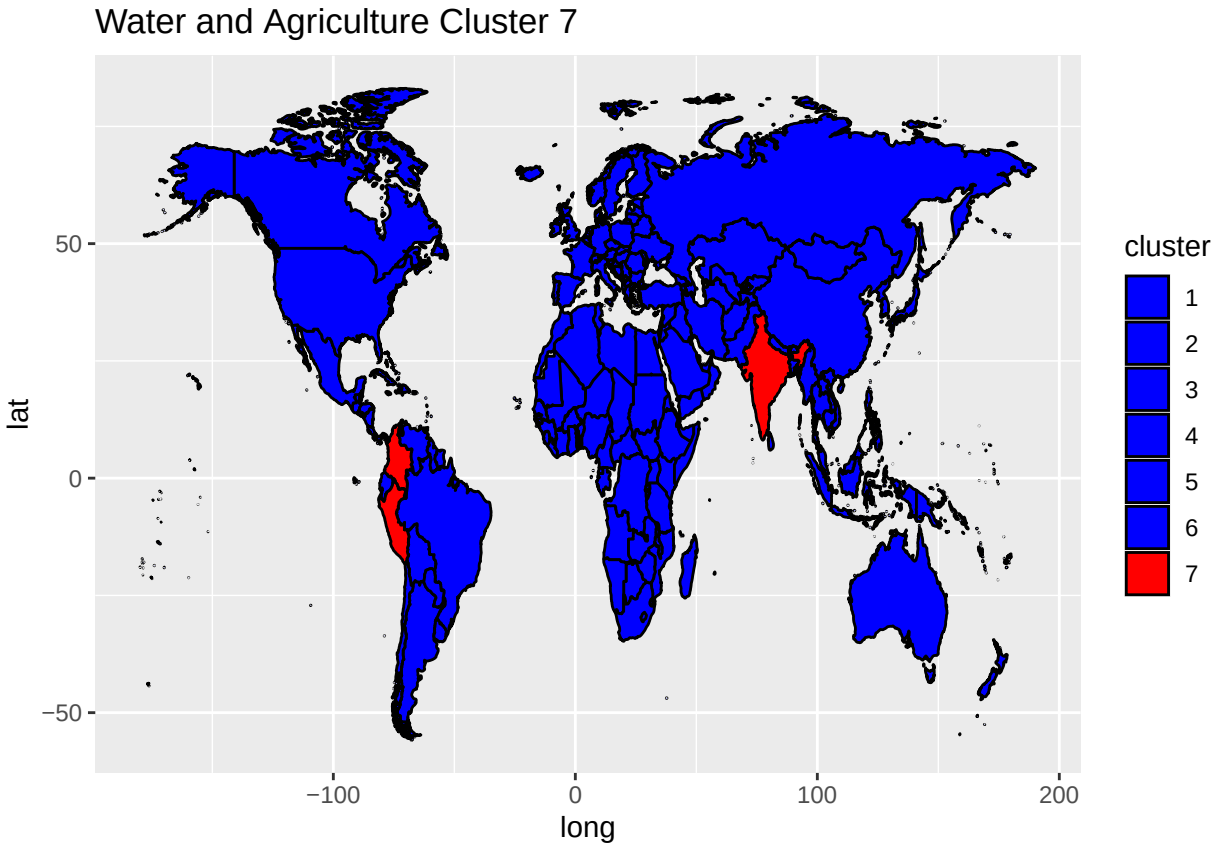

Water and Agriculture Cluster 2
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Water and Agriculture Cluster 3
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Water and Agriculture Cluster 4
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Water and Agriculture Cluster 5
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Water and Agriculture Cluster 6
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7




```
#View each cluster by country name.
countries <- lapply(split(country_clusters$region, country_clusters$cluster), sort)
maxl <- max(sapply(countries, length))
res <- sapply(countries, function(x) c(x, rep(NA, maxl - length(x))))
maxl <- max(sapply(countries, length))
country_names <- sapply(countries, function(x) c(x, rep(NA, maxl - length(x))))
country_names <- as.data.frame(res)
country_names %>% gt() %>% fmt_missing(columns = 1:7,missing_text = "")
```

1	2	3	4
Afghanistan	Azerbaijan	Comoros	Algeria
Benin	Bahrain	Dominica	Andorra
Burkina Faso	Barbados	Kiribati	Angola
Cambodia	Chile	Malaysia	Antigua
Central African Republic	China	Marshall Islands	Argentina
Chad	Cook Islands	Sao Tome and Principe	Armenia
Democratic Republic of the Congo	Djibouti	Tokelau	Australia
Eritrea	Ecuador	Tuvalu	Austria
Ethiopia	Egypt		Bahamas
Ghana	Greece		Belize
Guinea	Iran		Bhutan
Guinea-Bissau	Iraq		Bolivia
Ivory Coast	Japan		Bosnia and Herzegovina
Kenya	Kuwait		Botswana

Liberia
Madagascar
Mali
Mauritania
Mozambique
Myanmar
Niger
Sierra Leone
Solomon Islands
Somalia
Sudan
Tanzania
Yemen

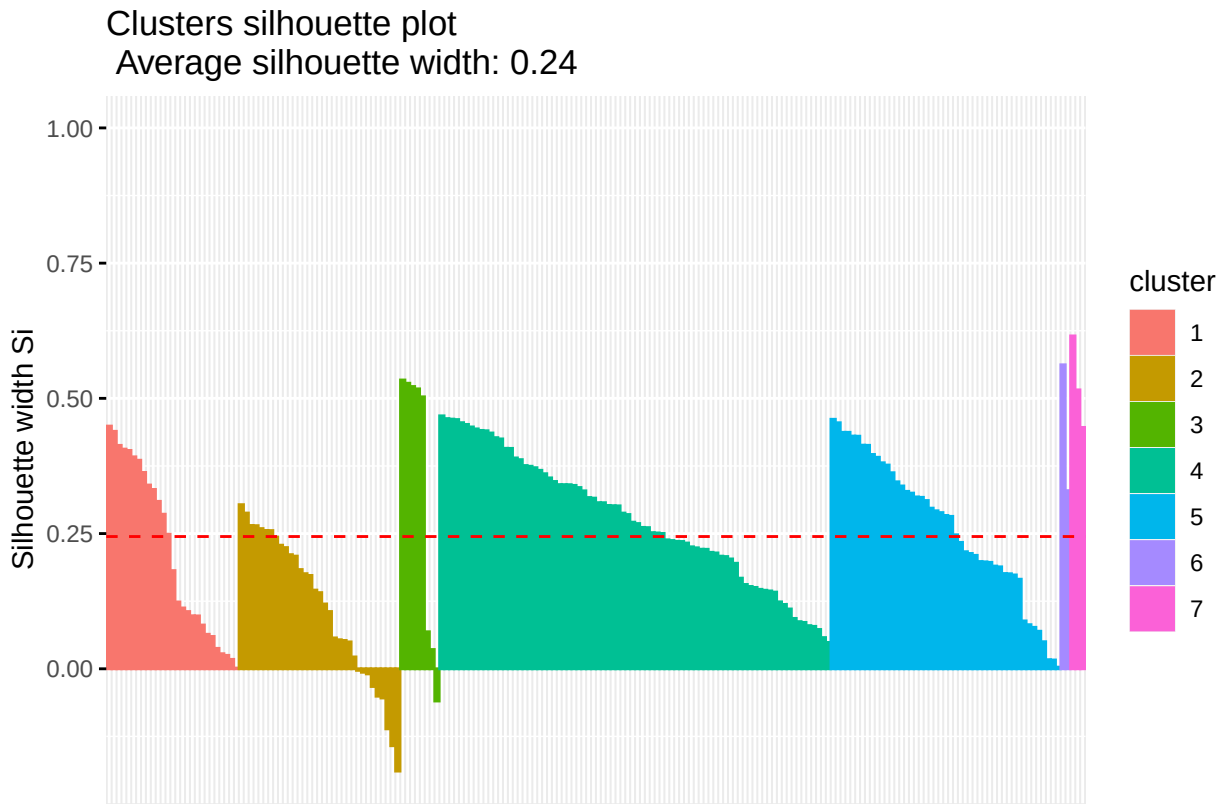
Kyrgyzstan
Lebanon
Micronesia
Nepal
New Zealand
North Korea
Oman
Palestine
Puerto Rico
Qatar
Saudi Arabia
South Korea
Suriname
Tajikistan
Turkmenistan
United Arab Emirates
Uzbekistan
Vanuatu
Vietnam

Brazil
Brunei
Cameroon
Canada
Cape Verde
Costa Rica
Croatia
Cyprus
Democratic Republic of the Congo
Equatorial Guinea
Estonia
Faroe Islands
Fiji
Finland
Gabon
Georgia
Grenada
Guatemala
Honduras
Ireland
Israel
Jamaica
Jordan
Kazakhstan
Laos
Latvia
Lesotho
Libya
Liechtenstein
Maldives
Mexico
Mongolia
Montenegro
Morocco
Namibia
Nauru
Nicaragua
Niue
North Macedonia
Norway
Palau
Panama
Papua New Guinea
Paraguay
Portugal
Russia
Saint Lucia
Saint Vincent
Samoa
Senegal
Seychelles
Singapore
Slovenia
South Africa

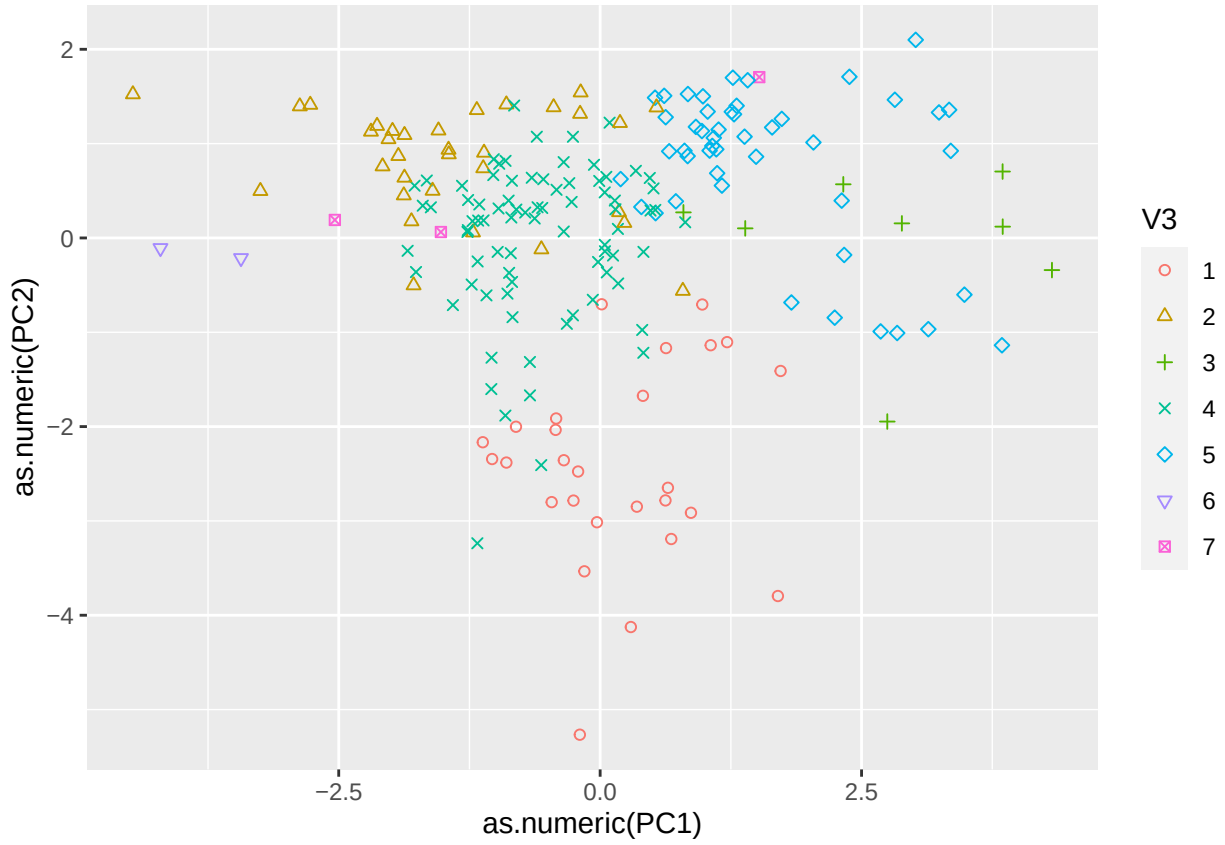
South Sudan
st Kitts
Swaziland
Sweden
Switzerland
Timor-Leste
Trinidad
Uruguay
USA
Venezuela
Zambia
Zimbabwe

```
#silhouette plot.
sil <- silhouette(clusters$cluster, dist(data_cluster))
fviz_silhouette(sil)
```

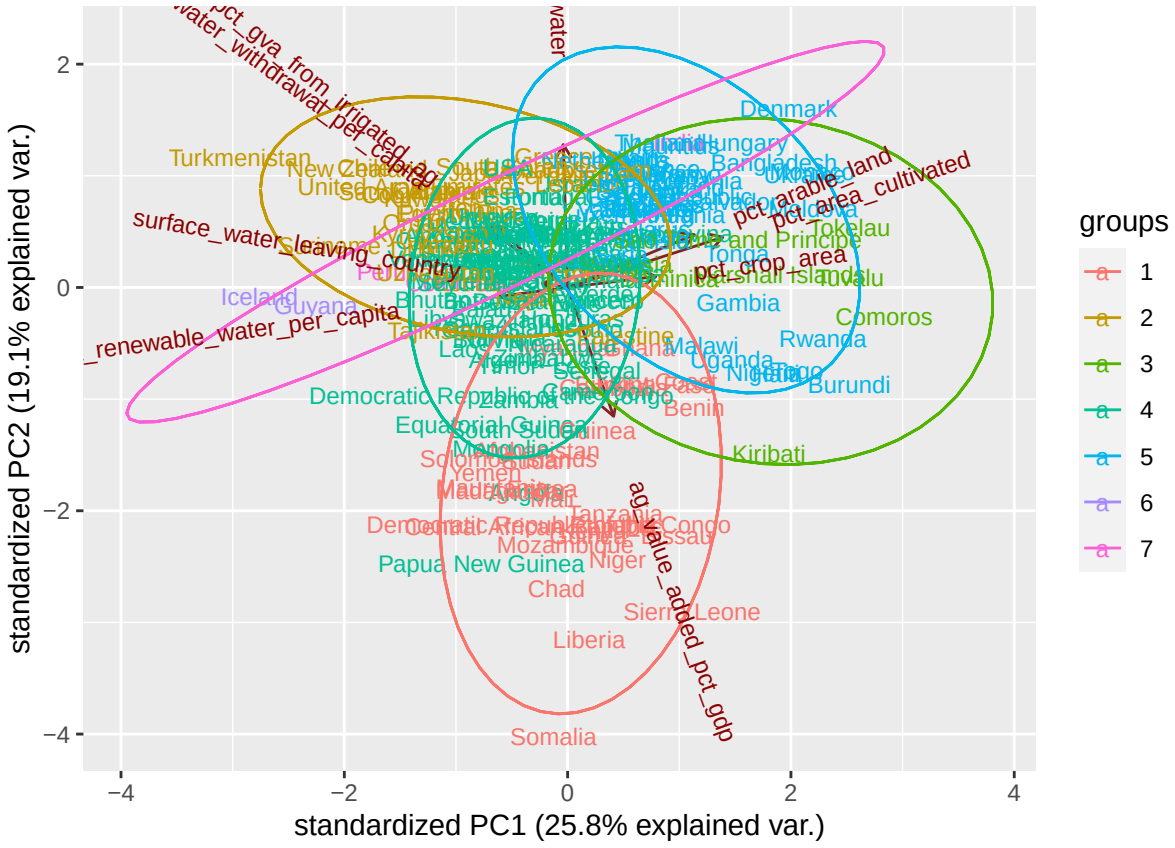
```
##   cluster size ave.sil.width
## 1       1   27          0.21
## 2       2   33          0.11
## 3       3    8          0.33
## 4       4   80          0.28
## 5       5   47          0.26
## 6       6    2          0.45
## 7       7    3          0.53
```



```
#do the prcomp mapping for first two principal components.
pca <- prcomp(data_cluster, scale. = TRUE)
pca_scores <- prcomp(data_cluster, scale. = TRUE)$x[,1:2]
pca_scores <- as.data.frame(cbind(pca_scores,country_clusters$cluster))
pca_scores <- pca_scores %>% mutate(PC1 = as.numeric(PC1))
pca_scores <- pca_scores %>% mutate(PC2 = as.numeric(PC2))
pca_scores %>% ggplot(aes(x=as.numeric(PC1), y=as.numeric(PC2), group=V3)) + geom_point(aes(shape=V3,co
```




```
#ggbiplot also helps visualize this better.
ggbiplot(pca, groups = country_clusters$cluster, labels=country_clusters$region, ellipse = TRUE, ellips
```

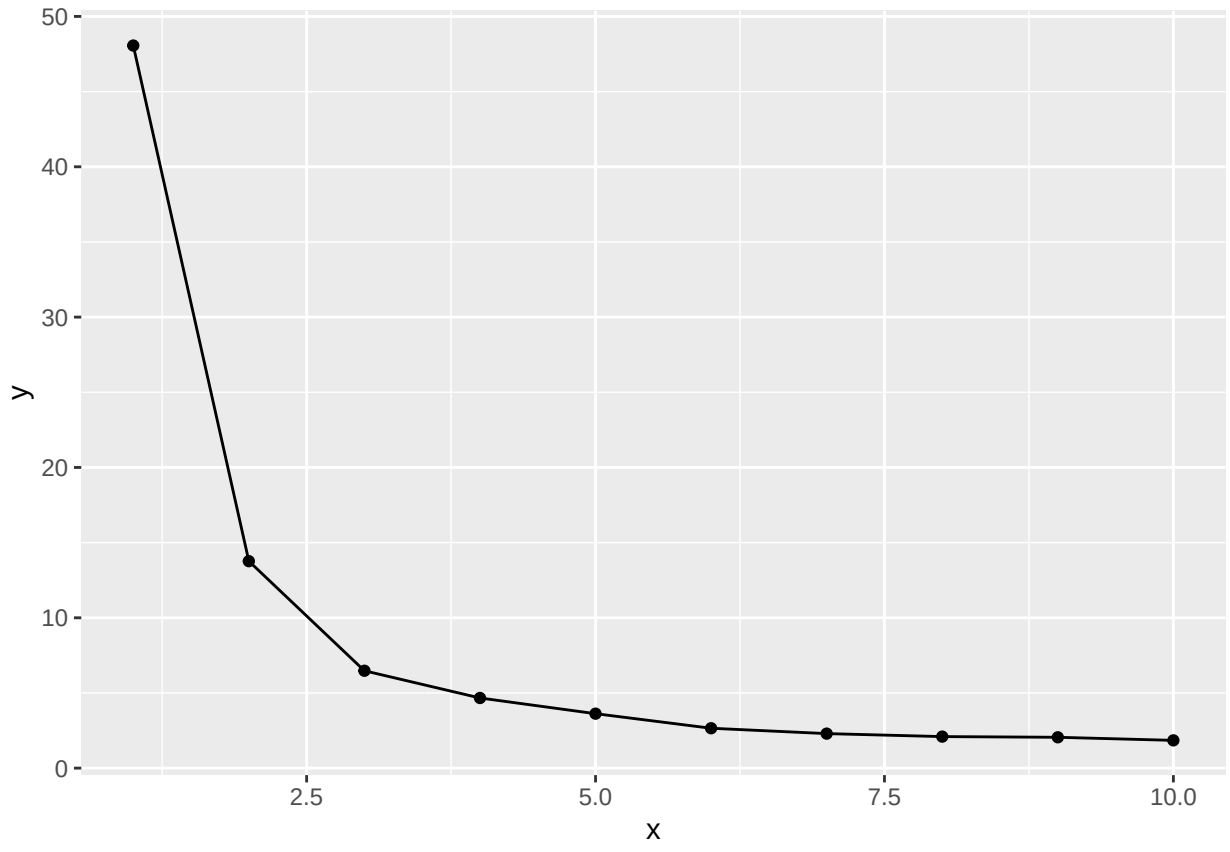



```
#look at how these align with clustering off some other variables of human development index and gender
other_variables <- data %>%
  select(hdi,gii)


#cluster on other variables. imputation, selection of centers, clustering.
other_imputation <- mice(other_variables)
```

```
##
##  iter imp variable
##    1   1  hdi  gii
##    1   2  hdi  gii
##    1   3  hdi  gii
##    1   4  hdi  gii
##    1   5  hdi  gii
##    2   1  hdi  gii
##    2   2  hdi  gii
##    2   3  hdi  gii
##    2   4  hdi  gii
##    2   5  hdi  gii
##    3   1  hdi  gii
##    3   2  hdi  gii
##    3   3  hdi  gii
##    3   4  hdi  gii
##    3   5  hdi  gii
##    4   1  hdi  gii
```

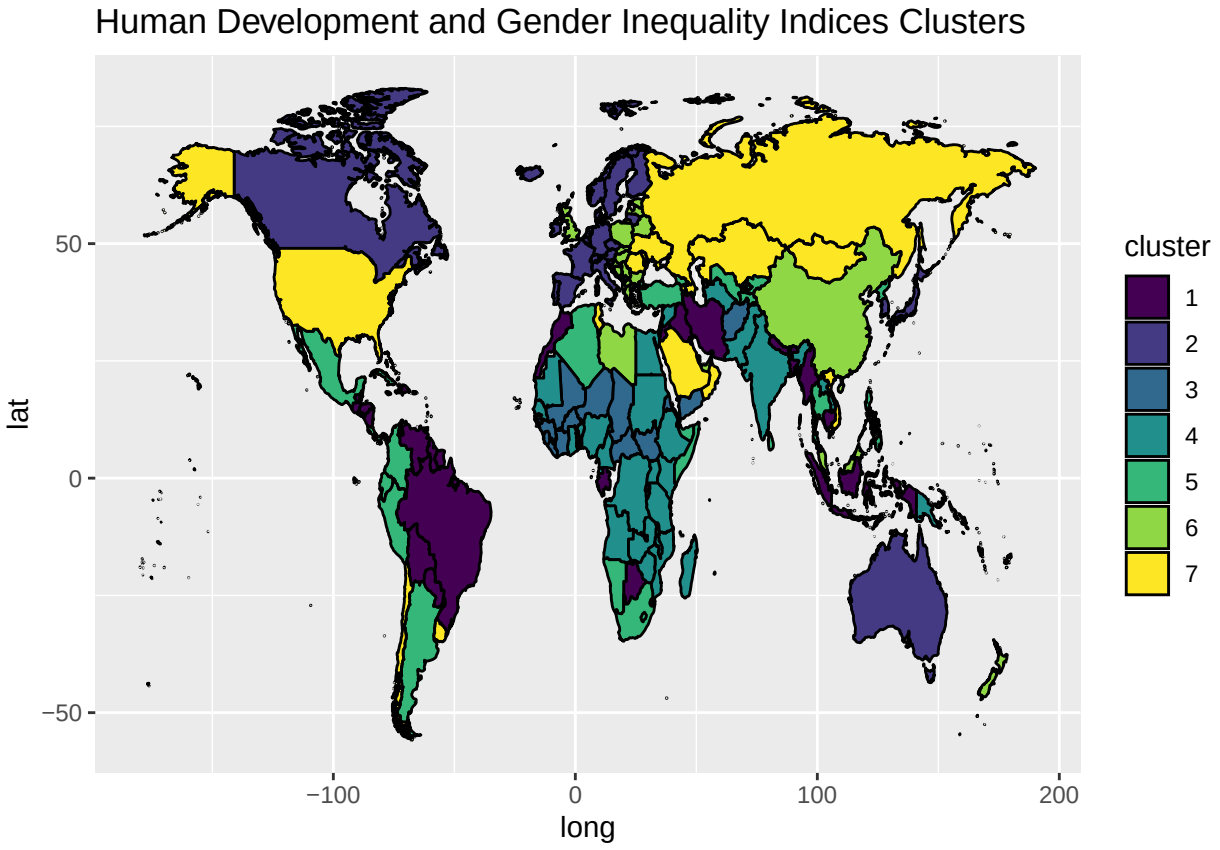
```
##   4   2  hdi  gii
##   4   3  hdi  gii
##   4   4  hdi  gii
##   4   5  hdi  gii
##   5   1  hdi  gii
##   5   2  hdi  gii
##   5   3  hdi  gii
##   5   4  hdi  gii
##   5   5  hdi  gii
```

```
other_imputed <- complete(other_imputation,1)
other_imputed_scaled <- scale(other_imputed,center=F)
wss2 <- sapply(1:10,
               function(k){kmeans(other_imputed_scaled, k)$tot.withinss})
data_frame(x = 1:length(wss2), y = wss2) %>%
  ggplot(aes(x, y)) +
  geom_point() +
  geom_line()
```

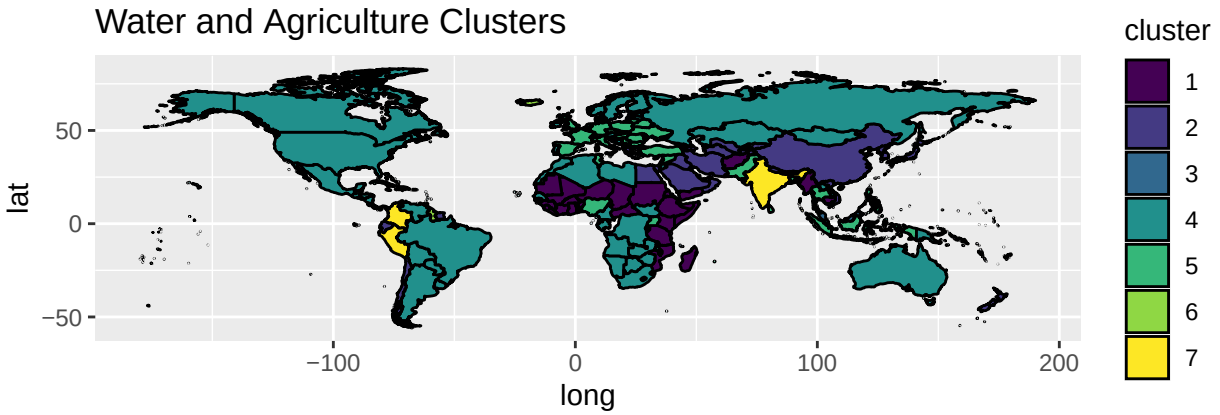


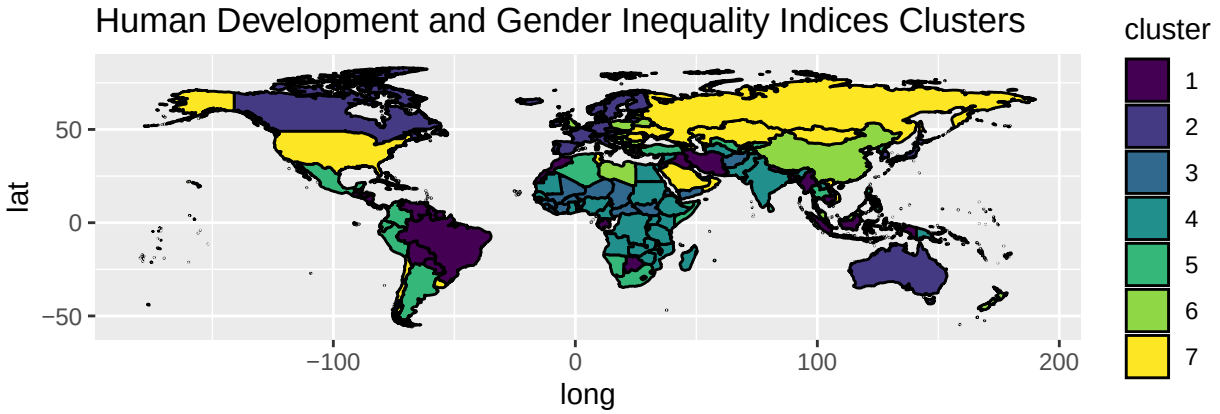

```
other_kmeans_k4 <- kmeans(other_imputed_scaled, centers =7 , nstart = 20)

other <- as.data.frame(cbind(other_kmeans_k4$cluster,data$country))


#to plot on map, need to rename some of the countries for the join.
```

```
colnames(other) <- c("cluster","region")
other <- other %>% mutate(region = case_when(
  region == "United States of America"~"USA",
  region == "Russian Federation"~"Russia",
  region == "Venezuela (Bolivarian Republic of)"~"Venezuela",
  region == "Bolivia (Plurinational State of)"~"Bolivia",
  region == "Czechia"~"Czech Republic",
  region == "Iran (Islamic Republic of)"~"Iran",
  region == "Antigua and Barbuda"~"Antigua",
  region == "Brunei Darussalam"~"Brunei",
  region == "Cabo Verde"~"Cape Verde",
  region == "Congo"~"Democratic Republic of the Congo",
  region == "CÃ´te d'Ivoire"~"Ivory Coast",
  region == "Democratic People's Republic of Korea"~"North Korea",
  region == "Eswatini"~"Swaziland",
  region == "Grenade"~"Grenada",
  region == "Holy See"~"Vatican",
  region == "Lao People's Democratic Republic"~"Laos",
  region == "Micronesia (Federated States of)"~"Micronesia",
  region == "Republic of Korea"~"South Korea",
  region == "Republic of Moldova"~"Moldova",
  region == "Saint Kitts and Nevis"~"st Kitts",
  region == "Saint Vincent and the Grenadines"~"Saint Vincent",
  region == "Syrian Arab Republic"~"Syria",
  region == "Trinidad and Tobago"~"Trinidad",
  region == "United Kingdom"~"UK",
  region == "United Republic of Tanzania"~"Tanzania",
  region == "Viet Nam"~"Vietnam",
  TRUE ~region
))


#create a map of the clusters.
map_data2 <- left_join(map_data,other, by="region")
map_data2 <- map_data2 %>% filter(!is.na(map_data2$cluster))
map_other <- ggplot(map_data2,aes(x=long, y=lat, group=group)) + geom_polygon(aes(fill=cluster), color =

map_other
```

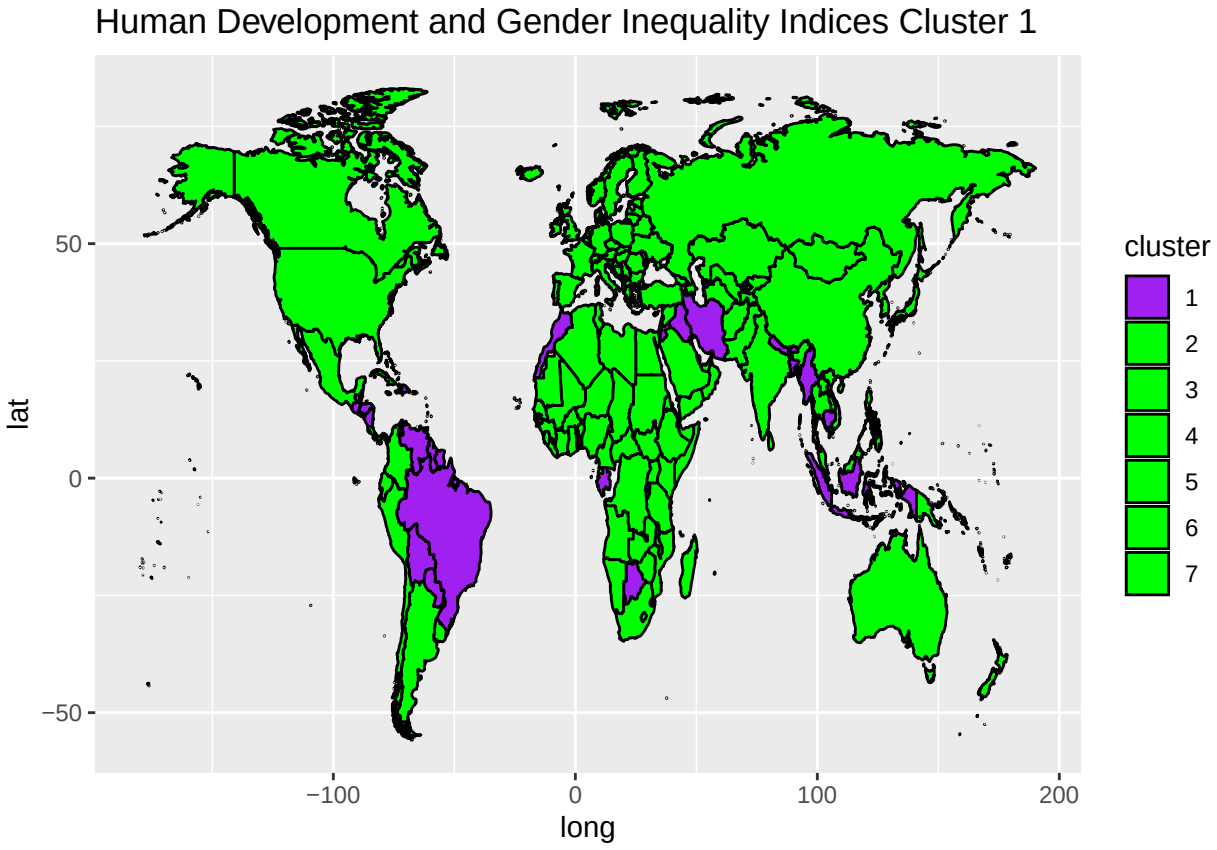


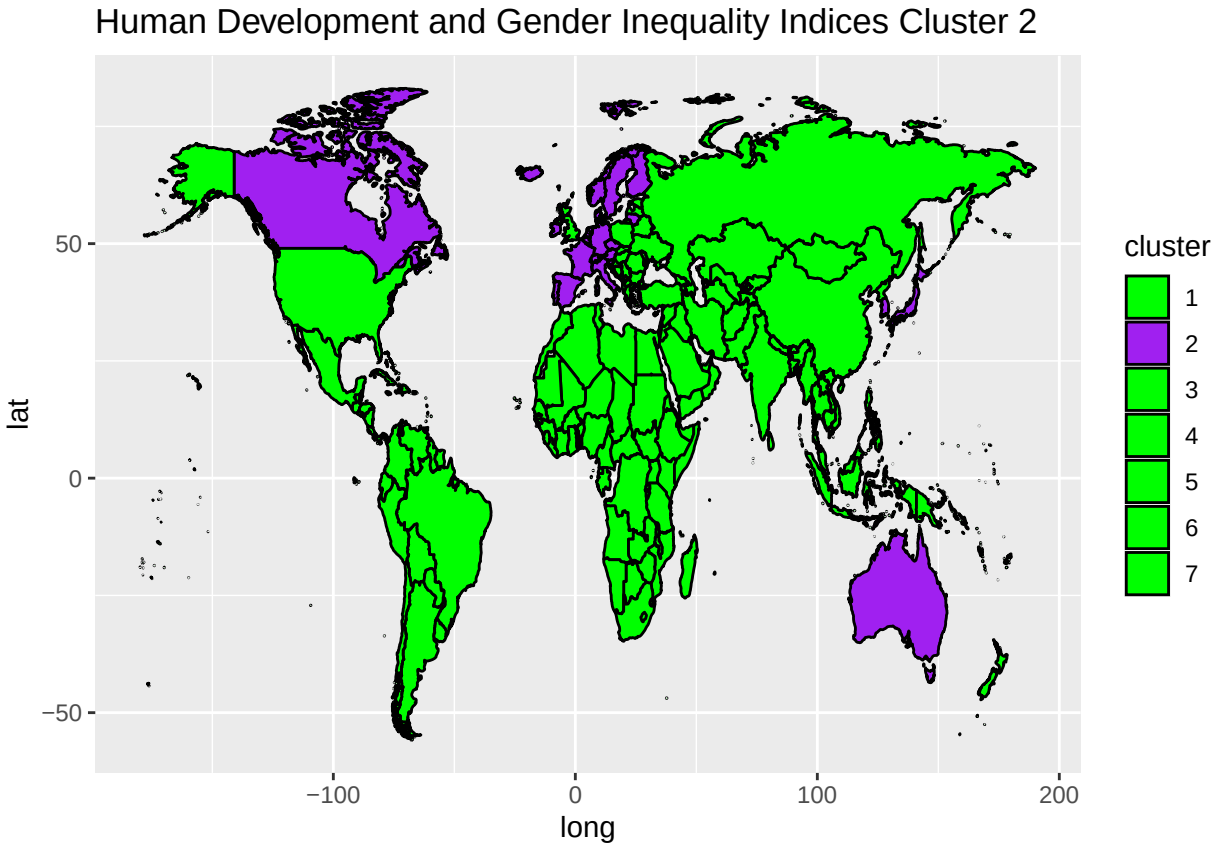

```
#look at the two maps side by side.
grid.arrange(map_water,map_other,nrow=2)
```

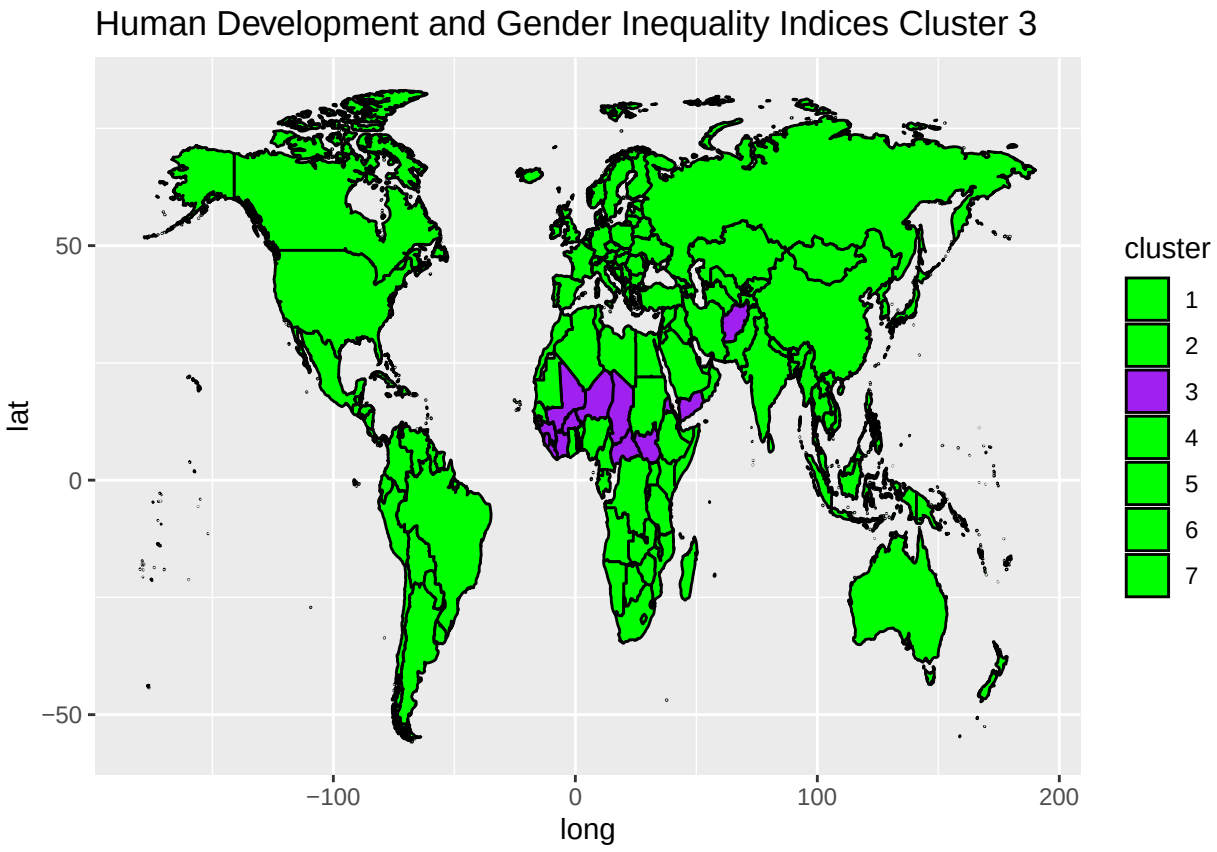


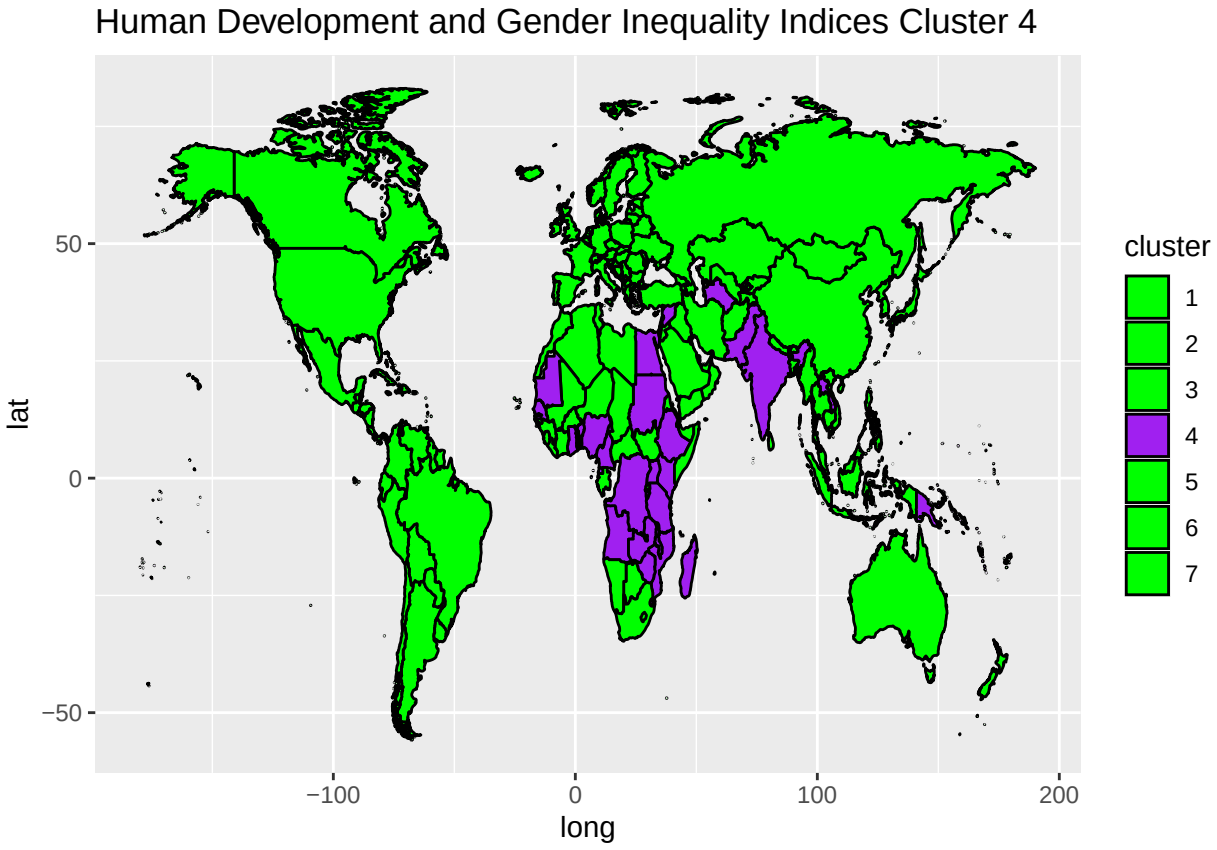


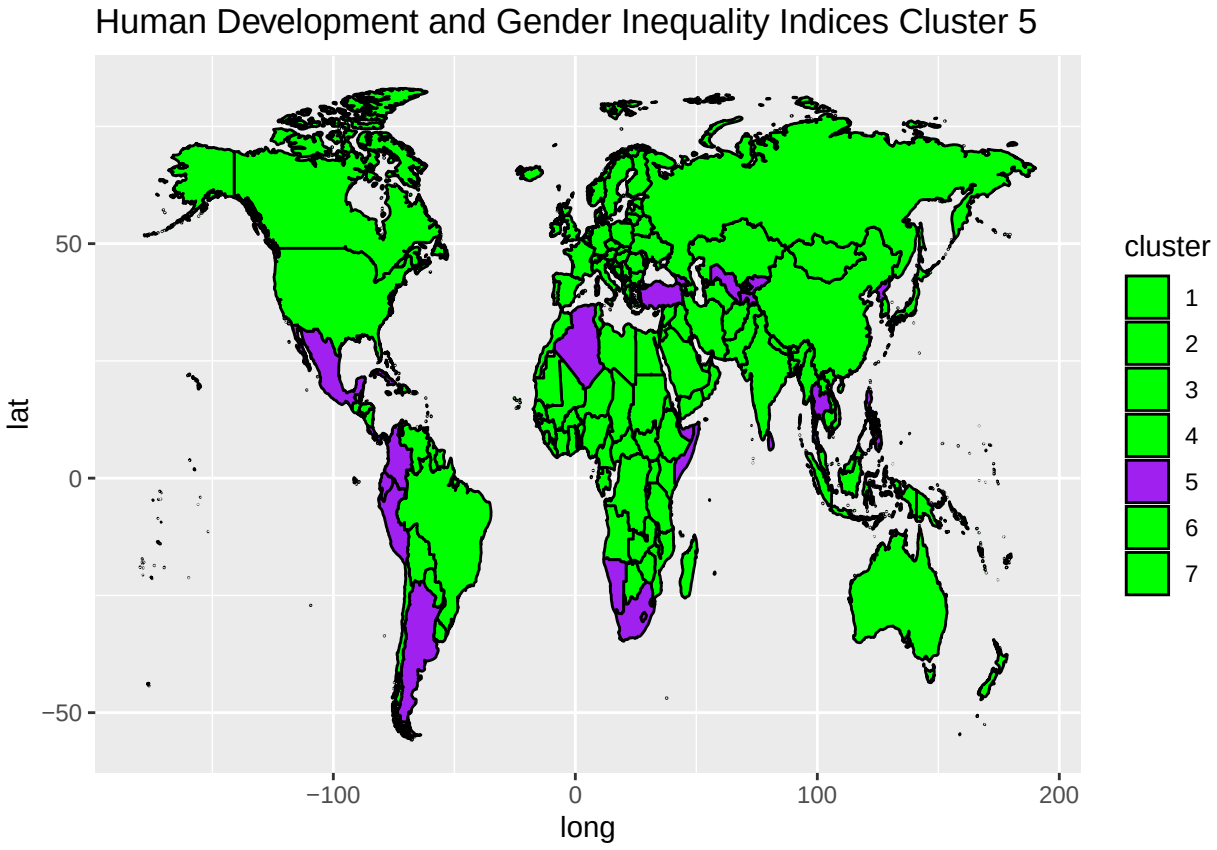

```
y <- c(NA,NA,NA,NA,NA,NA,NA)
#messing around with the mapping visualizations. Interested if we can show each of the clusters compare

for (i in 1:7) {
y[i] <- "purple"
y[-i] <- "green"
print(ggplot(map_data2,aes(x=long, y=lat, group=group)) + geom_polygon(aes(fill=cluster), color = "blac
}
```

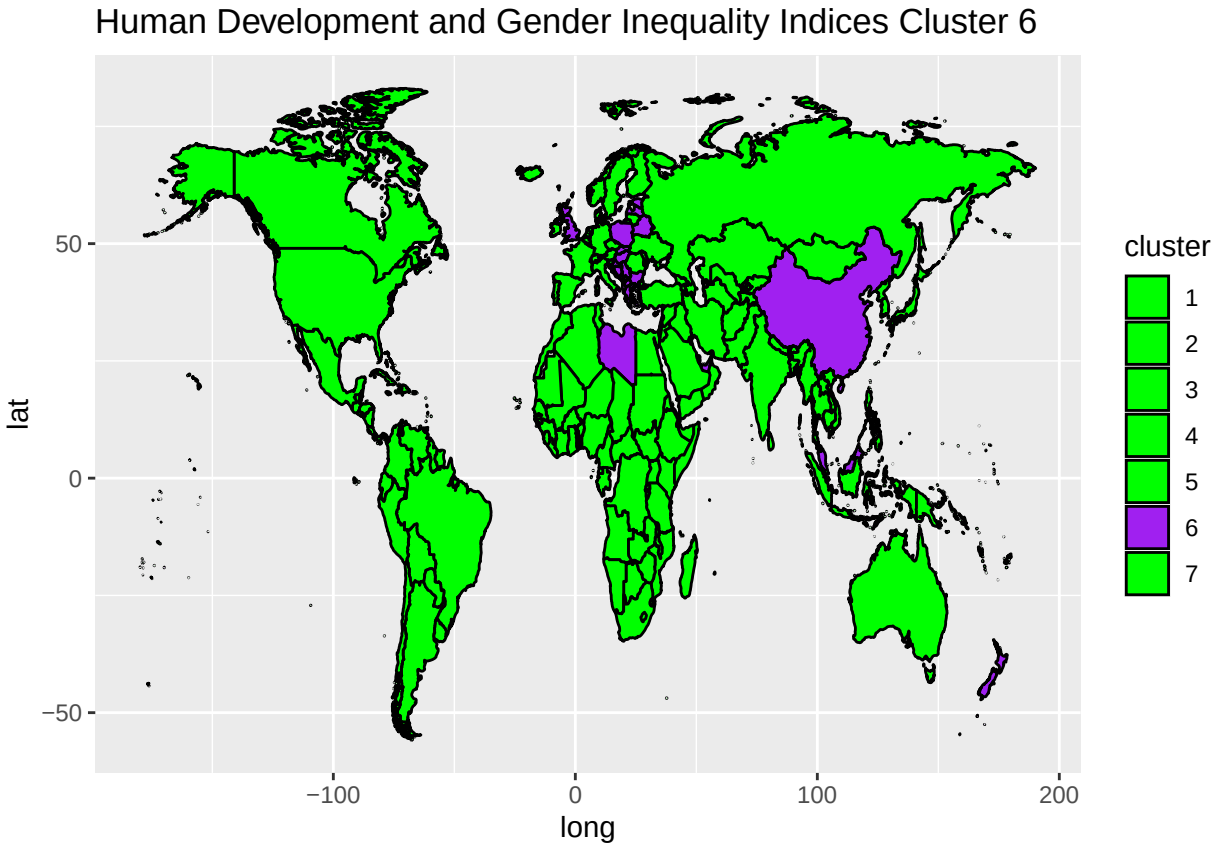

Human Development and Gender Inequality Indices Cluster 1
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7

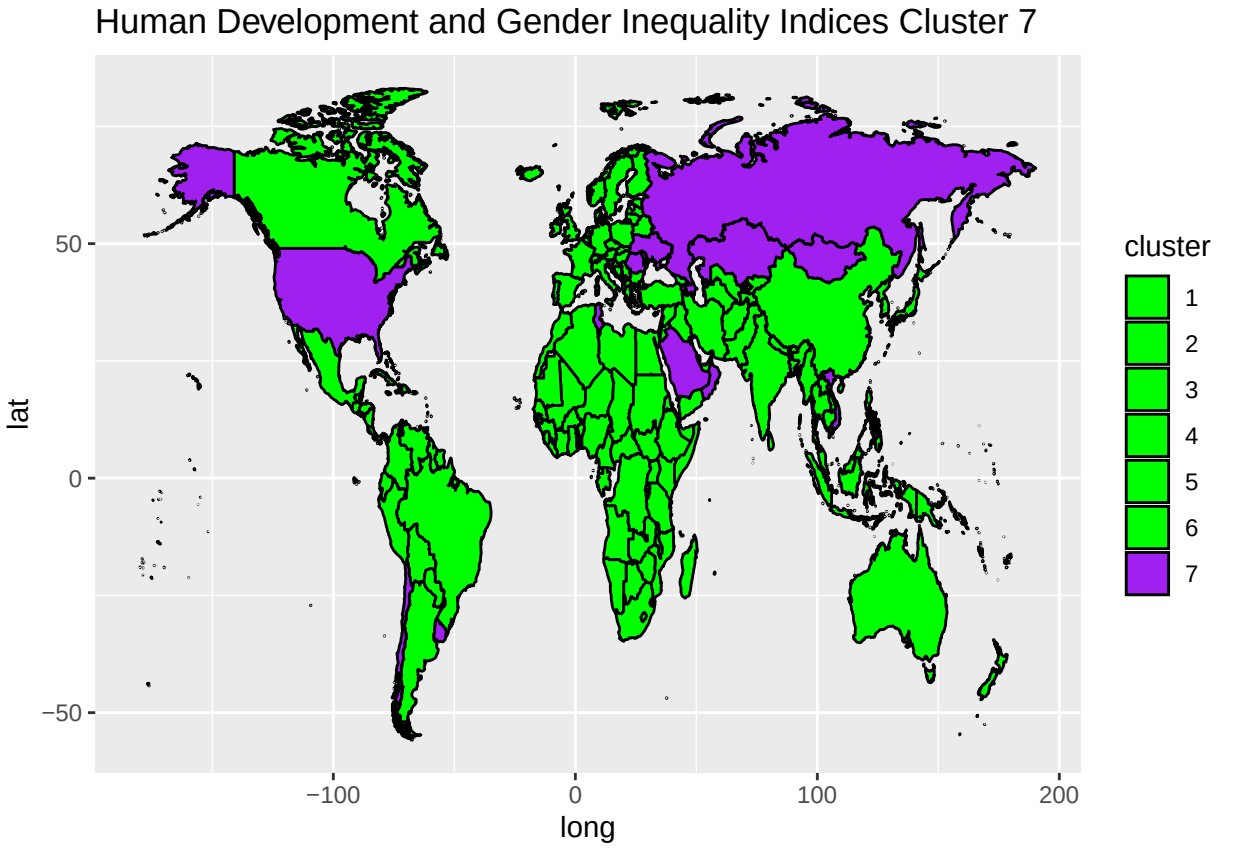

Human Development and Gender Inequality Indices Cluster 2
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Human Development and Gender Inequality Indices Cluster 3
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Human Development and Gender Inequality Indices Cluster 4
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Human Development and Gender Inequality Indices Cluster 5
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7


Human Development and Gender Inequality Indices Cluster 6
lat
50
0
−50
−100
0
100
200
long
cluster
1
2
3
4
5
6
7



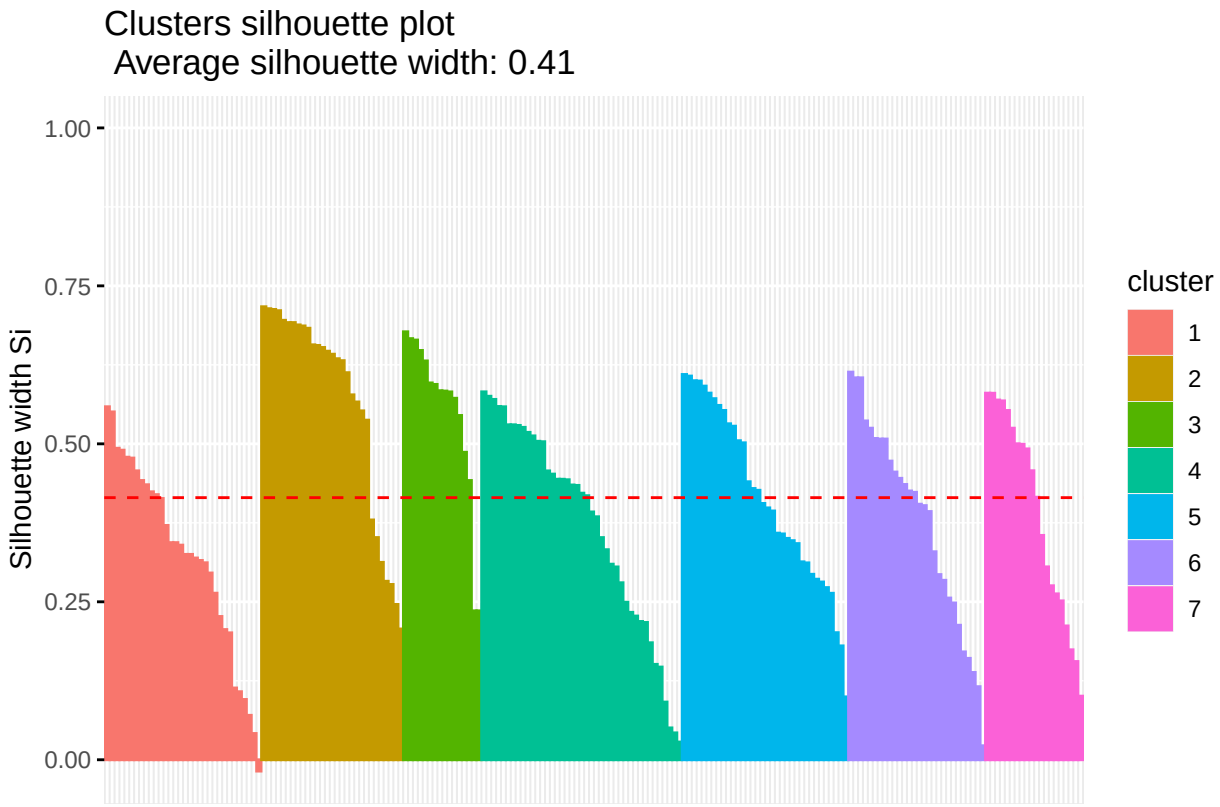

```
#look at these clusters by name.
names <- lapply(split(other$region, other$cluster), sort)
maxl <- max(sapply(names, length))
res <- sapply(names, function(x) c(x, rep(NA, maxl - length(x))))
maxl <- max(sapply(names, length))
names_index <- sapply(names, function(x) c(x, rep(NA, maxl - length(x))))
names_index <- as.data.frame(res)
names_index %>% gt() %>% fmt_missing(columns = 1:7,missing_text = "")
```

1	2	3	4	5
Bangladesh	Australia	Afghanistan	Angola	Algeria
Bhutan	Austria	Burkina Faso	Benin	Antigu
Bolivia	Belgium	Central African Republic	Burundi	Argent
Botswana	Brunei	Chad	Cameroon	Barbad
Brazil	Canada	Democratic Republic of the Congo	Comoros	Belize
Cambodia	Cyprus	Eritrea	Democratic Republic of the Congo	Colom
Cape Verde	Czech Republic	Gambia	Djibouti	Costa
Cook Islands	Denmark	Guinea	Egypt	Cuba
Dominica	Finland	Guinea-Bissau	Equatorial Guinea	Ecuad
Dominican Republic	France	Ivory Coast	Ethiopia	Fiji
El Salvador	Germany	Liberia	Faroe Islands	Georgi
Gabon	Iceland	Mali	Ghana	Jamai
Guatemala	Ireland	Niger	Haiti	Kuwai
Guyana	Israel	Sierra Leone	India	Kyrgyz

Honduras	Italy	South Sudan	Kenya	Leban
Indonesia	Japan	Yemen	Laos	Marsh
Iran	Lithuania		Lesotho	Maurit
Iraq	Luxembourg		Madagascar	Mexic
Jordan	Netherlands		Malawi	Namib
Kiribati	Norway		Mauritania	North
Morocco	Portugal		Micronesia	Panam
Myanmar	Singapore		Mozambique	Peru
Nepal	Slovenia		Nigeria	Philip
Nicaragua	South Korea		Pakistan	Puerto
Palestine	Spain		Papua New Guinea	Saint
Paraguay	Sweden		San Marino	Seyche
Qatar	Switzerland		Sao Tome and Principe	Somali
Rwanda	Tokelau		Senegal	South
Samoa	Tuvalu		Solomon Islands	Sri La
Suriname			Sudan	Tajikis
Timor-Leste			Swaziland	Thaila
Venezuela			Syria	Trinid
			Tanzania	Turkey
			Togo	Uzbek
			Tonga	
			Turkmenistan	
			Uganda	
			Vanuatu	
			Vatican	
			Zambia	
			Zimbabwe	

```
#silhouette plot.
sil <- silhouette(other_kmeans_k4$cluster, dist(other_imputed_scaled))
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1   32          0.32
## 2       2   29          0.57
## 3       3   16          0.55
## 4       4   41          0.37
## 5       5   34          0.41
## 6       6   28          0.37
## 7       7   20          0.39
```


Clusters silhouette plot
Average silhouette width: 0.41
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
4
5
6
7