Project 1

VADYM DUDARENKO

Imagine clustering

In this project, I decide to use unsupervised learning methods for image clustering. The goal is to cluster an image and find the main colors. I will compare two physical maps of countries such as Iceland and Cyprus. Colors can show surface relief and environment of each country

Data preprocessing

First step is to open a given image. To open .jpg file, JPEG library

```
library(jpeg)

image1 <- readJPEG("C:/Users/Vadym/Desktop/iceland.jpg")
image2 <- readJPEG("C:/Users/Vadym/Desktop/cyprus.jpg")
```

First, we represent the image in the form of a matrix of numbers to show it. The most suitable is RGB format. It helps to determine the amount of green, blue and red in this pixel of the picture. This is all easy to do, RGB represents the images in a 3-column matrix where the first column refers to the amount of red, the second refers to the amount of green, and the last refers to the amount of blue in the range 0-255.

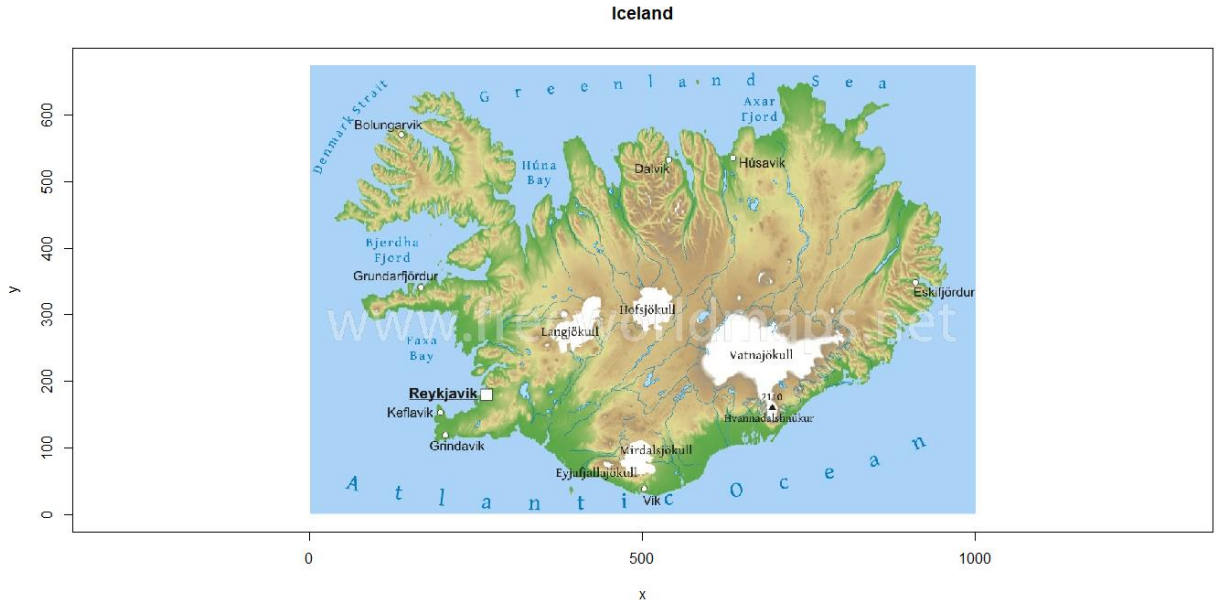
First, let’s check the dimension.

```
dm1 <- dim(image1);dm1[1:2]
[1]  673 1000
dm2 <-  dim(image2);dm2[1:2]
[1] 501 800
```

Size of the “iceland” image is equal to 673x1000 and the size of the “cyprus” image is equal to 501x800.

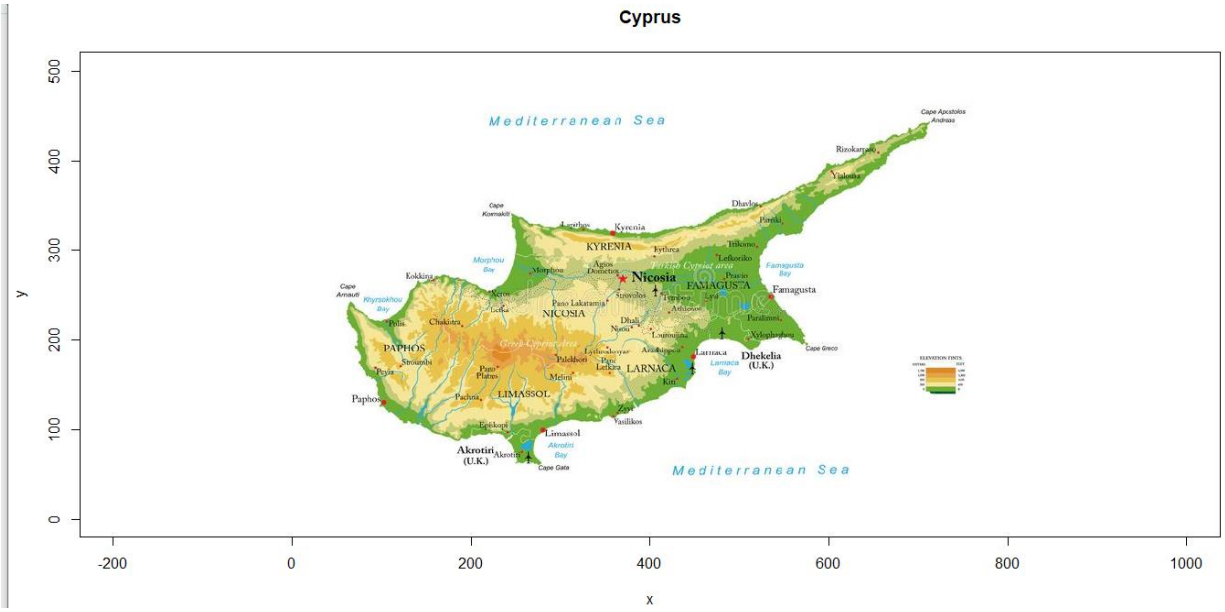
Now, we can change the format of the images from jpg to rgb and display them on the plot.

```
rgbImage1 <- data.frame(
  x=rep(1:dm1[2], each=dm1[1]),
  y=rep(dm1[1]:1, dm1[2]),
  r.value=as.vector(image1[,,1]),
  g.value=as.vector(image1[,,2]),
  b.value=as.vector(image1[,,3]))

plot(y ~ x, data=rgbImage1, main="Iceland",
     col = rgb(rgbImage1[c("r.value", "g.value", "b.value")]),
     asp = 1, pch = ".")
```




```
rgbImage2 <- data.frame(
  x=rep(1:dm2[2], each=dm2[1]),
  y=rep(dm2[1]:1, dm2[2]),
  r.value=as.vector(image2[,,1]),
```

```
  g.value=as.vector(image2[,,2]),
  b.value=as.vector(image2[,,3]))

plot(y ~ x, data=rgbImage2, main="Cyprus",
     col = rgb(rgbImage2[c("r.value", "g.value", "b.value")]),
     asp = 1, pch = ".")
```



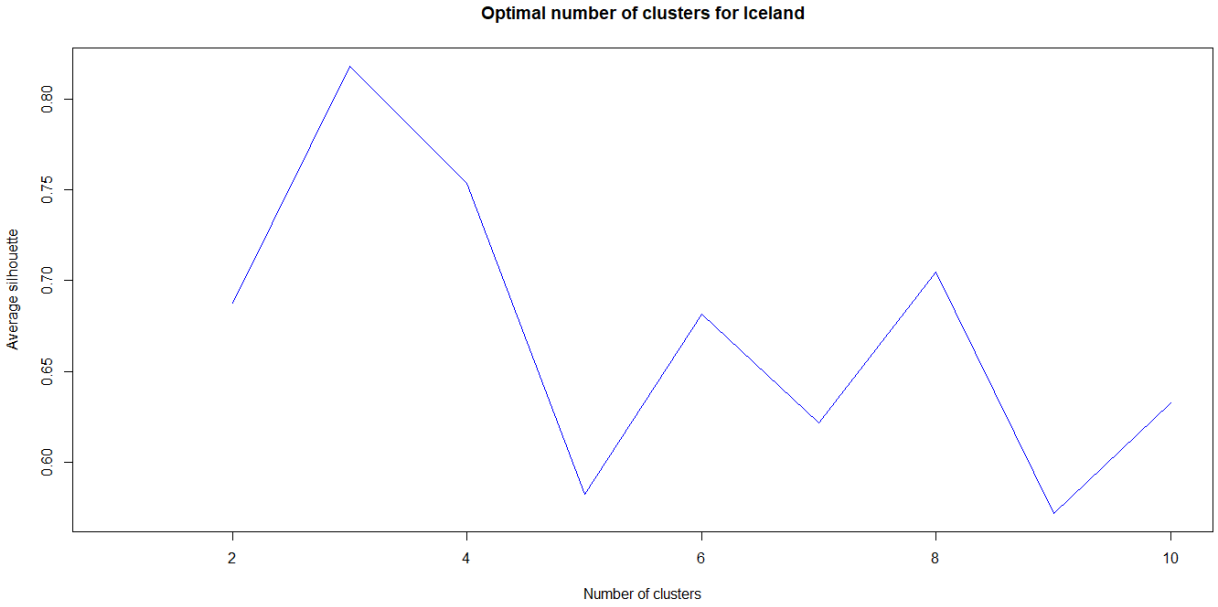
Optimal number of k-clusters

I will use Clara algorithm for image clustering. It is based on k-medoids PAM algorithm and it is optimal for large data sets. The size of the data sets of the images are 673 x 1000 x 3 and 501 x 800 x 3, which is 673000 and 400800, so they could be classified as large data sets.

First step is to find the optimal number of k - clusters for each image by comparing average silhouette width for every k. Silhouette ranges from -1 to 1 and it describes clustering consistency. A positive value means that the elements in cluster are correctly matched - objects in a given clusters are similar to each other and dissimilar to the objects of the surrounding clusters. The higher value, the better clustering.

I will use “cluster” library to run clara algorithm for 10 consecutive numbers to analyze the average silhouette width.

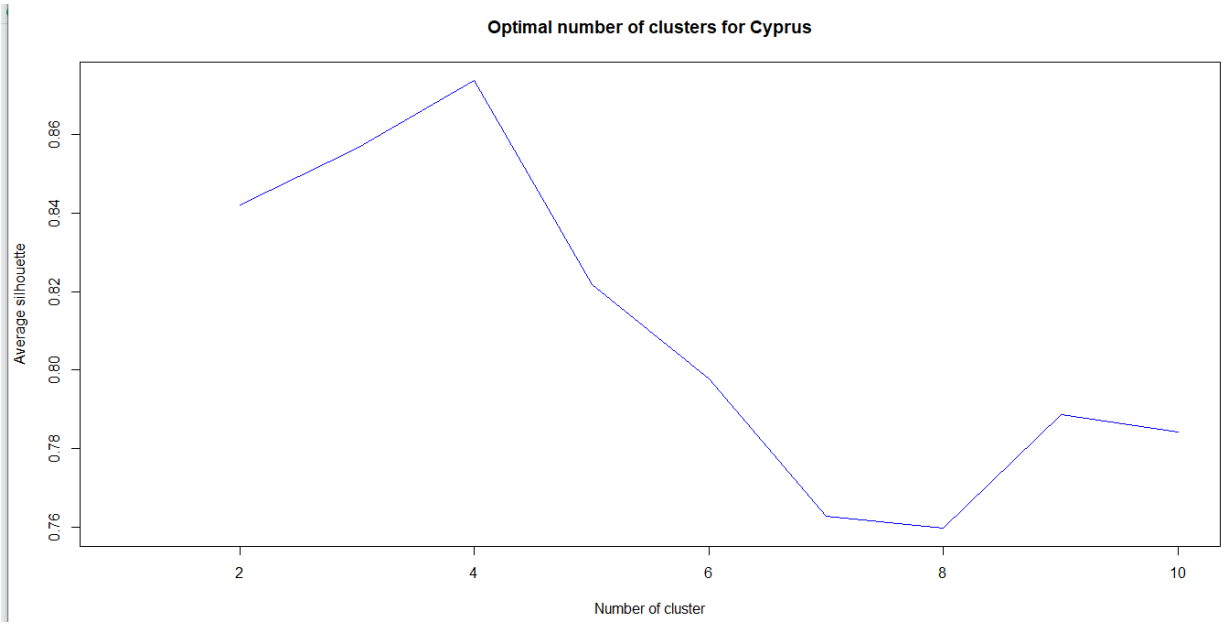
```
library(cluster)

n1 <- c()
for (i in 1:10) {
  cl <- clara(rgbImage1[, c("r.value", "g.value", "b.value")], i)
  n1[i] <- cl$silinfo$avg.width
}

plot(n1, type = 'l',
     main = "Optimal number of clusters for Iceland",
     xlab = "Number of clusters",
     ylab = "Average silhouette",
     col = "blue")
```



The results show that 3 clusters are optimal for the “Iceland” image.

Let’s repeat the process the “Cyprus” image.

```
n2 <- c()
for (i in 1:10) {
  cl <- clara(rgbImage2[, c("r.value", "g.value", "b.value")], i)
  n2[i] <- cl$silinfo$avg.width
}

plot(n2, type = 'l',
     main = "Optimal number of clusters Cyprus",
     xlab = "Number of cluster",
     ylab = "Average silhouette",
     col = "blue")
```



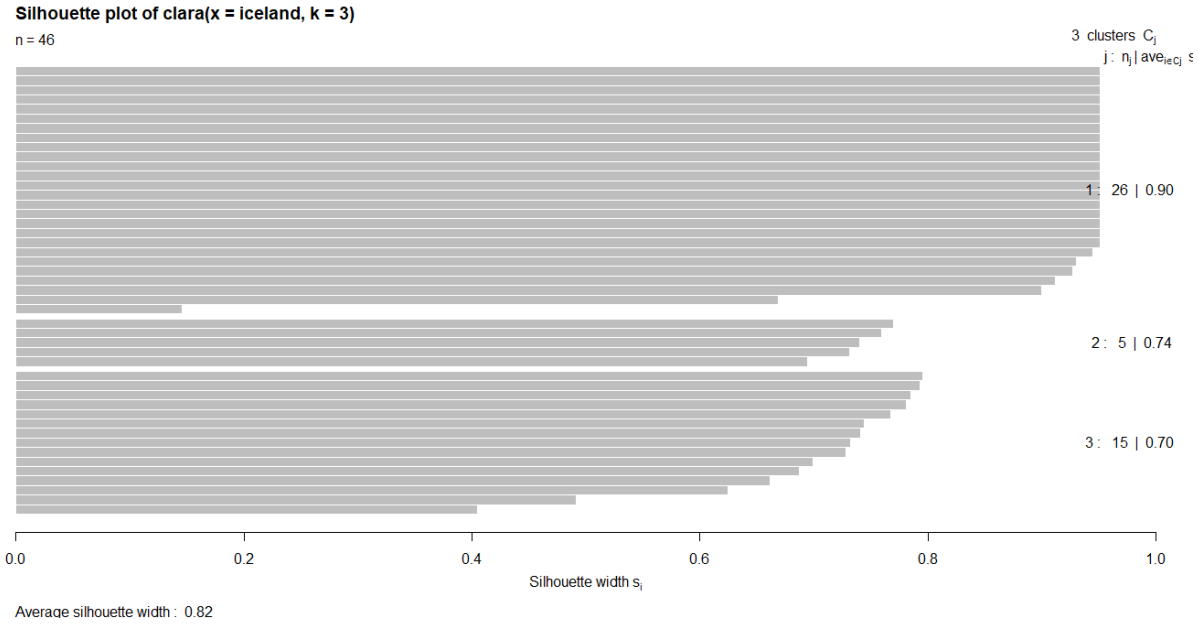
The results show that 4 clusters are optimal for the “Cyprus” image.

Running Clara algorithm

Now, let’s run clara with the given number of clusters.

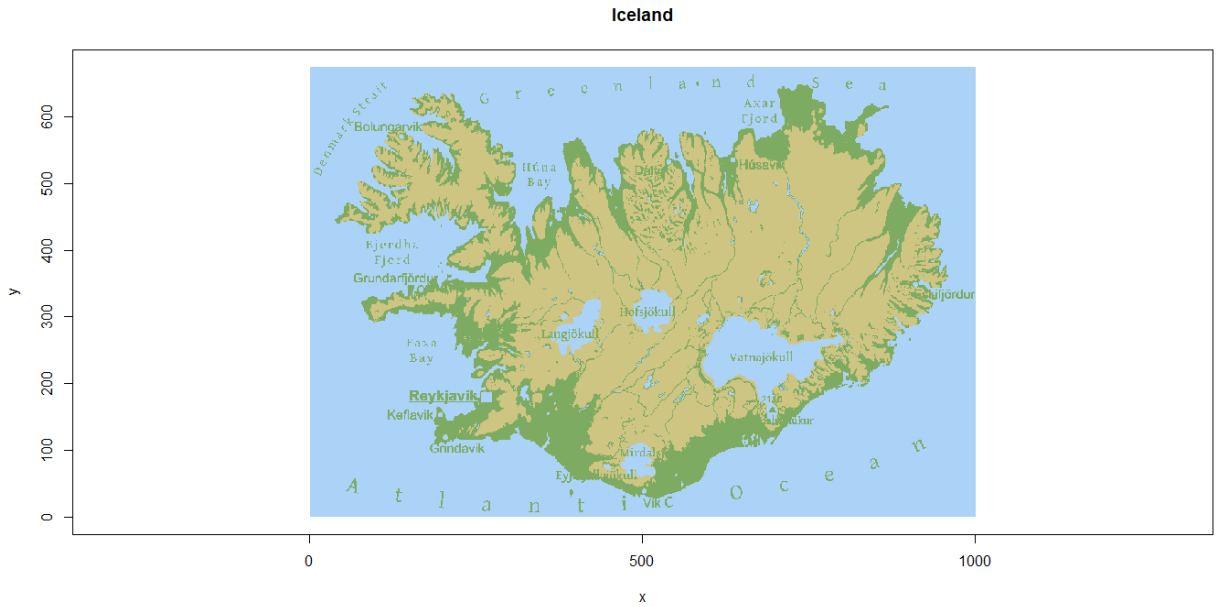
“Iceland” image:

```
iceland = rgbImage1[, c("r.value", "g.value", "b.value")]
clara <- clara(iceland, 3)
plot(silhouette(clara))
```



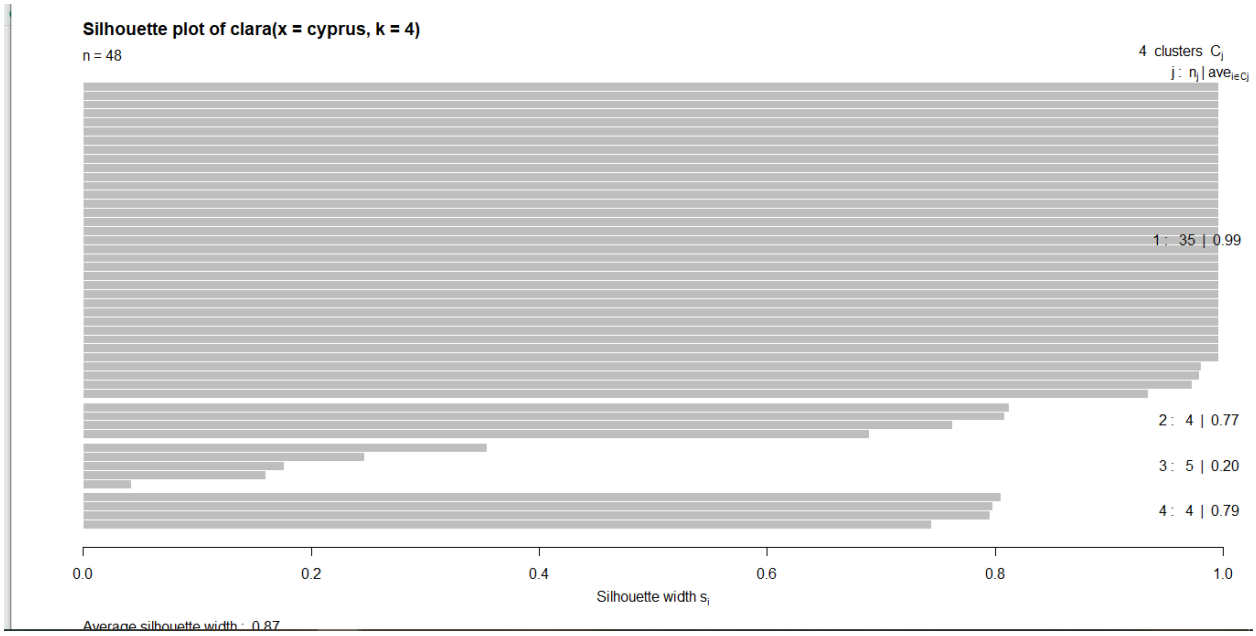
Let's see the clustered image on the plot with the usage of rgb() function, which creates colours from rgb values.

```
colours <- rgb(clara$medoids[clara$clustering, ])
plot(y ~ x, data=rgbImage1, main="Iceland",
    col = colours,
    asp = 1, pch = ".")
```

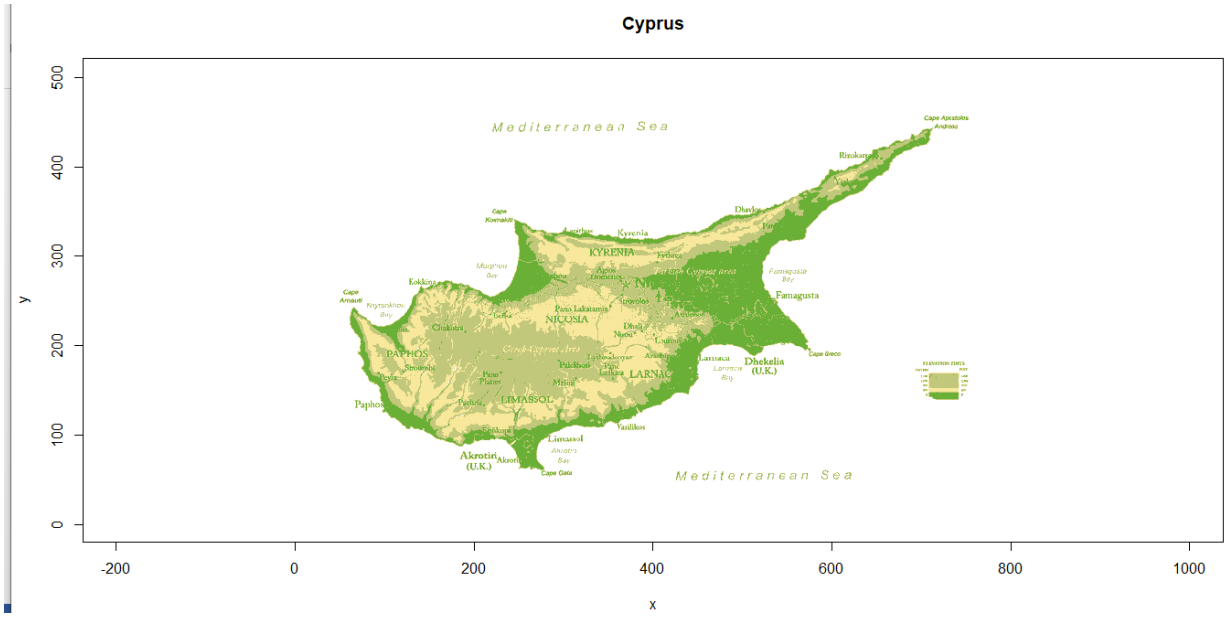


“Cyprus” image:

```
cyprus = rgbImage2[, c("r.value", "g.value", "b.value")]
clara2 <- clara(cyprus, 4)
plot(silhouette(clara2))
```




```
colours2 <- rgb(clara2$medoids[clara2$clustering, ])
plot(y ~ x, data=rgbImage2, main="Cyprus",
    col = colours2,
    asp = 1, pch = ".")
```



Finding main colors

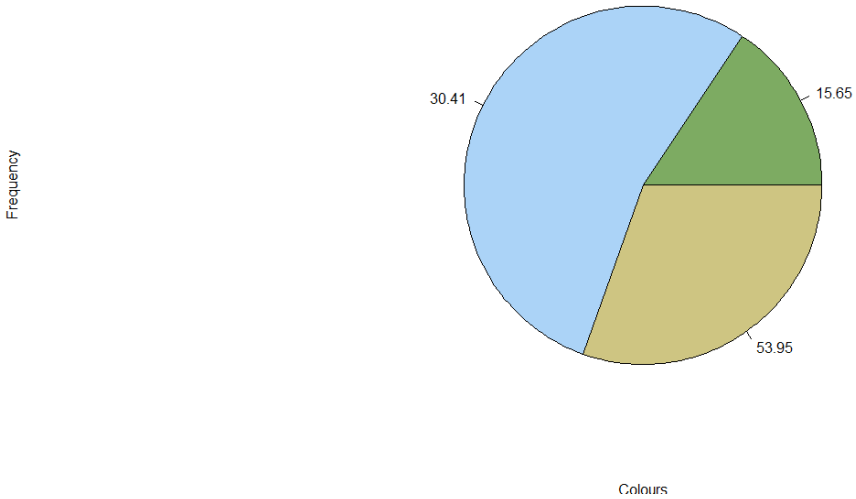
We can find the main colors of the image by counting clusters distribution. Let's use the output of rgb() function, which is given in a hexadecimal format and and represents colours.

Let's count the percentage colour distribution. Value of the colour frequency column is simply a size of the clusters.

```
dominantColours <- as.data.frame(table(colours))

max_col  <- max(dominantColours$Freq)/sum(dominantColours$Freq)
min_col  <- min(dominantColours$Freq)/sum(dominantColours$Freq)
medium_col <- 1-max_col - min_col

dominantColours$distribution <- round((c(min_col, medium_col, max_col) * 100),
2)
dominantColours
```

```
colours         Freq         distribution
1 #7DAB62 105309        15.65
2 #ABD3F7 363052        30.41
3 #CEC582 204639        53.95
```

```
dominantColours$colours <- as.character(dominantColours$colours)
pie(dominantColours$Freq, labels = dominantColours$distribution,
    col = dominantColours$colours,
    xlab = "Colours",
    ylab = "Frequency")
```



Looking at the graph, we can see that there are three primary colors, that is, the relief mainly consists of this. 30. 41 % - occupies a blue color, which means that this is the island and it is surrounded by some kind of ocean or sea. Also, looking at the picture, we can understand that there are water resources inside of territory of country, this is a small percentage that is included in 30.41.

15.65 % is the smallest percentage that shows plain and forested surfaces.

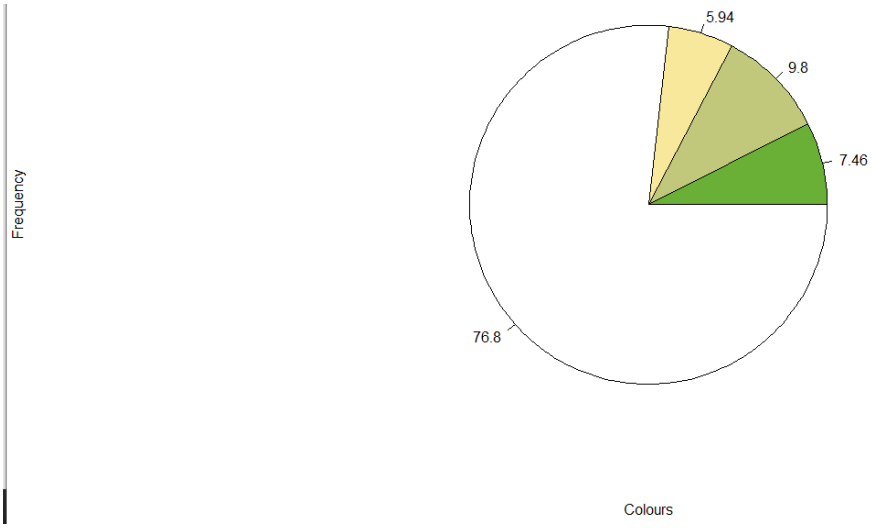
53.95 % - is a plateau, that is, a very low mountainous area, which is what a very light brown color tells us, it means that height is approximately no more 2,00 meters.

If we consider only the territory of the country (without an ocean), we can simply count it in the percentage, then 72.3 % is our plateau, 20.9 is a plain surface and 6.8 are rivers, lakes or other water resources. Of course, this is just assumption and not precise data, that we can get after clustering, but mostly this is correct information.

Let’s repeat the process for the “Cyprus” image.

```
dominantColours2 <- as.data.frame(table(colours2))

max_col2  <- (dominantColours2$Freq[1])/sum(dominantColours2$Freq)
min_col2  <- (dominantColours2$Freq[3])/sum(dominantColours2$Freq)
medium_col2 <- (dominantColours2$Freq[2])/sum(dominantColours2$Freq)
```

```
bg <- (dominantColours2$Freq[4])/sum(dominantColours2$Freq)
dominantColours2$distribution <- round((c(max_col2, medium_col2, min_col2, bg) * 100), 2)

dominantColours2

colours2       Freq    distribution
1  #6AB036  29899        7.46
2  #C1C87B  39281        9.80
3  #F8E89B  23821        5.94
4  #FFFFFF 307799       76.80


dominantColours2$colours2 <- as.character(dominantColours2$colours2)
pie(dominantColours2$Freq, labels = dominantColours2$distribution,
    col = dominantColours2$colours2,
    xlab = "Colours",
    ylab = "Frequency")
```



Looking at the graph, we can see that there are four primary colors, that is, the relief mainly consists of this. 76. 8 % - occupies a white color, which means that this is the island and it is surrounded by some kind of ocean or sea. Also, looking at the

picture, we can understand that there are water resources inside of territory of country, this is a small percentage that is included in 76.8%.
7.46 % is the smallest percentage that shows plain and forested surfaces.
9.8 % and 5.94% - a mountainous area. The colors tells us that it exist mountain territory with low height and higher which bright brown shows.

If we consider only the territory of the country (without an ocean or sea), we can simply count it in the percentage, then 65% is our mountains, 30.8 is a plain surface and 4.2 % are water resources.

Conclusion

We can compare two island. So, this is obvious that islands are surrounded by some kind of ocean or sea. When we consider only territory of country, we observe Iceland has bigger plateau territory, but Cyprus has more number of higher points what bright brown color indicates. Cyprus has more plain area on the about 10 % and also inside Iceland is more water resources.