Final project: Step 1

Danyu Zhang, Limingrui Wan, Daniel Alonso

December 13th, 2020

Contents

Importing libraries

```
library(dplyr)
library(ggplot2)
library(reshape2)
library(PerformanceAnalytics)
library(gridExtra)
library(stringr)
library(foreach)
library(MASS)
library(andrews)
library(mice)
library(factoextra)
library(corrplot)
library(plotrix)
```

Importing data

```
data <- read.csv('./data/data.csv')
head(data)
```

```
#>   X     continent             location total_cases new_cases new_cases_smoothed
#> 1 0          Asia          Afghanistan       41728        95              99.429
#> 2 1        Africa               Angola       11035       230             236.286
#> 3 2        Europe              Albania       21523       321             296.857
#> 4 3        Europe              Andorra        4888        63              80.429
#> 5 4          Asia United Arab Emirates      135141      1234            1272.429
#> 6 5 South America            Argentina     1183118      9598           11547.143
#>   total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 1         1544          3                3.143                1071.918
#> 2          286          2                2.571                 335.755
#> 3          527          9                6.714                7478.977
#> 4           75          0                0.429               63262.797
#> 5          497          1                2.429               13663.856
#> 6        31623        483              331.714               26177.623
#>   new_cases_per_million new_cases_smoothed_per_million total_deaths_per_million
#> 1                2.440                          2.554                   39.663
#> 2                6.998                          7.189                    8.702
#> 3              111.544                        103.154                  183.126
#> 4              815.376                       1040.944                  970.685
#> 5              124.767                        128.653                   50.251
#> 6              212.365                        255.492                  699.689
#>   new_deaths_per_million stringency_index population population_density
#> 1                  0.077             5.56   38928341              54.422
#> 2                  0.061               NA   32866268              23.890
#> 3                  3.127            50.93    2877800             104.871
#> 4                  0.000            59.26      77265             163.755
#> 5                  0.101            47.22    9890400             112.442
#> 6                 10.687            81.94   45195777              16.177
#>   median_age aged_65_older aged_70_older gdp_per_capita extreme_poverty
#> 1       18.6         2.581         1.337       1803.987              NA
#> 2       16.8         2.405         1.362       5819.495              NA
#> 3       38.0        13.188         8.643      11803.431             1.1
#> 4         NA            NA            NA             NA              NA
#> 5       34.0         1.144         0.526      67293.483              NA
#> 6       31.9        11.198         7.441      18933.907             0.6
```

```
#>   cardiovasc_death_rate diabetes_prevalence hospital_beds_per_thousand
#> 1               597.029                9.59                       0.50
#> 2               276.045                3.94                         NA
#> 3               304.195               10.08                       2.89
#> 4               109.135                7.97                         NA
#> 5               317.840               17.26                       1.20
#> 6               191.032                5.50                       5.00
#>   life_expectancy human_development_index development
#> 1           64.83                   0.498         low
#> 2           61.15                   0.581      medium
#> 3           78.57                   0.785        high
#> 4           83.73                   0.858   very high
#> 5           77.97                   0.863   very high
#> 6           76.67                   0.825   very high
```

Excluding smoothed columns as they are redundant transformations of other columns

```
removed_cols <- c('new_deaths_smoothed','new_cases_smoothed',
                  'new_cases_smoothed_per_million',
                  'total_cases_per_million','new_deaths_per_million',
                  'new_cases_per_million','total_deaths_per_million',
                  'aged_70_older','new_deaths','X')
data_n <- data
for (col in removed_cols) {data_n <- data_n[names(data_n) != col]}
```

Exploratory data analysis

Variable types

Categorical variables

- continent
- location
- development

Numerical variables

Discrete

- total_cases
- new_cases
- total_deaths
- new_deaths
- population

Continuous

- new_cases_smoothed
- new_deaths_smoothed
- total_cases_per_million
- new_cases_per_million
- new_cases_smoothed_per_million
- total_deaths_per_million
- new_deaths_per_million
- stringency_index
- population_density
- median_age
- aged_65_older
- aged_70_older
- gdp_per_capita
- extreme_poverty
- cardiovasc_death_rate
- diabetes_prevalence
- hospital_beds_per_thousand
- life_expectancy
- human_development_index

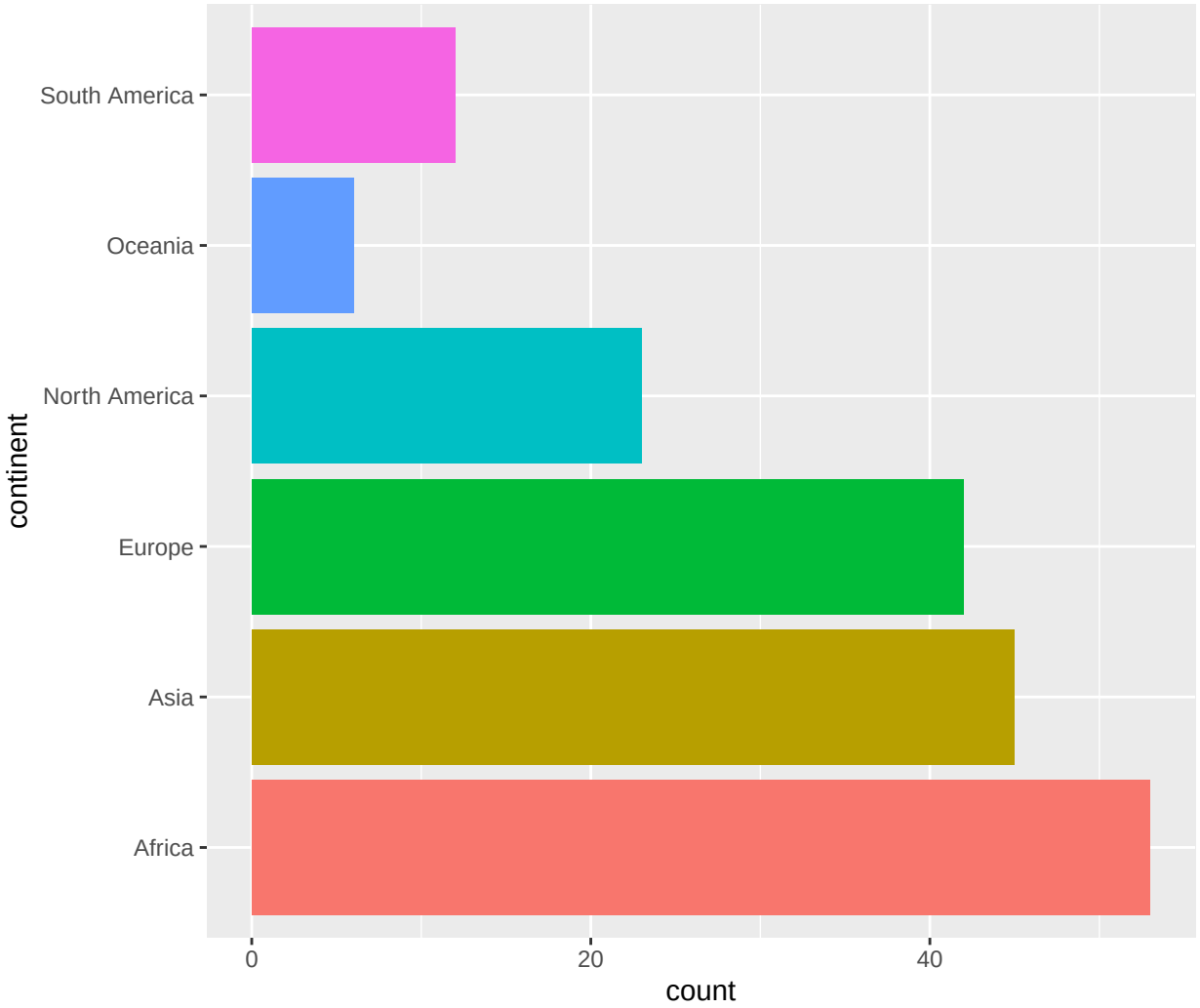
We select variables that we consider interesting to visualize, as the ones we haven't selected might be ralated to these or even ratios of them (in the case of total cases per million)

```
categorical <- c('location','continent','development')
interesting_vars <- c('total_cases','new_cases','total_deaths',
                      'stringency_index','population',
                      'population_density','median_age',
                      'aged_65_older','gdp_per_capita',
                      'extreme_poverty','cardiovasc_death_rate',
                      'diabetes_prevalence','hospital_beds_per_thousand',
                      'life_expectancy','human_development_index')
```

Plots with categorical variables

Countries per continent in the dataset

```
ggplot(data=data) +
    geom_bar(aes(fill=continent, y=continent), show.legend = FALSE)
```

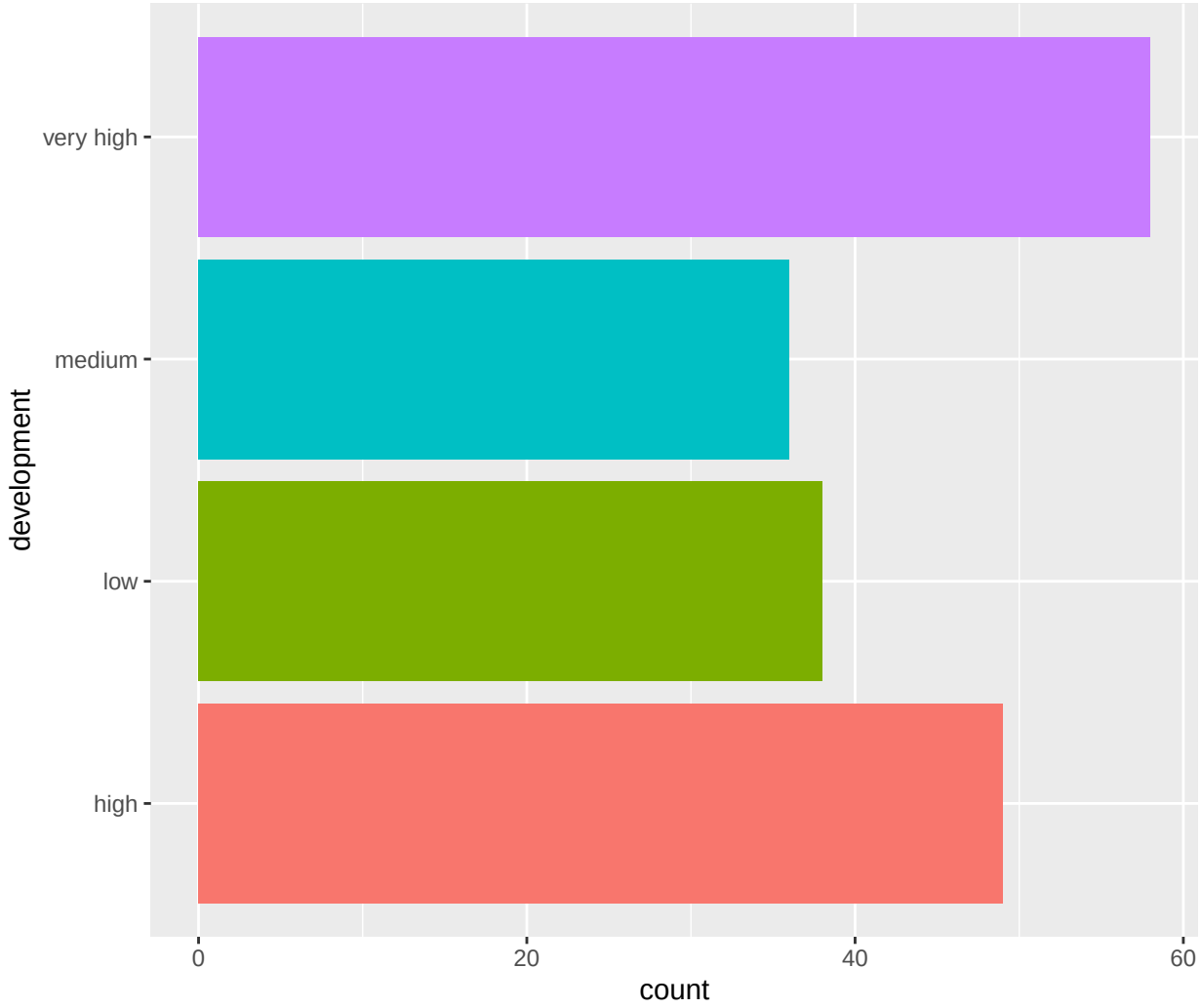


As a heads up, we will have more data points for continents with more countries, as there is an entry per country in the dataset. Therefore Africa will be our continent with the most entries, followed by Asia and Europe.

North america contains all of Central America and the Caribbean. Oceania contains some island countries and archipielagos in the pacific that belong to the continent but it is by far our continent with the least entries.

Amount of countries per HDI

```
ggplot(data=data) +
    geom_bar(aes(fill=development, y=development), show.legend = FALSE)
```



For development (which is a variable constructed from the *human_development_index* variable), we have an even amount of countries per HDI, with very high development countries being the largest group.

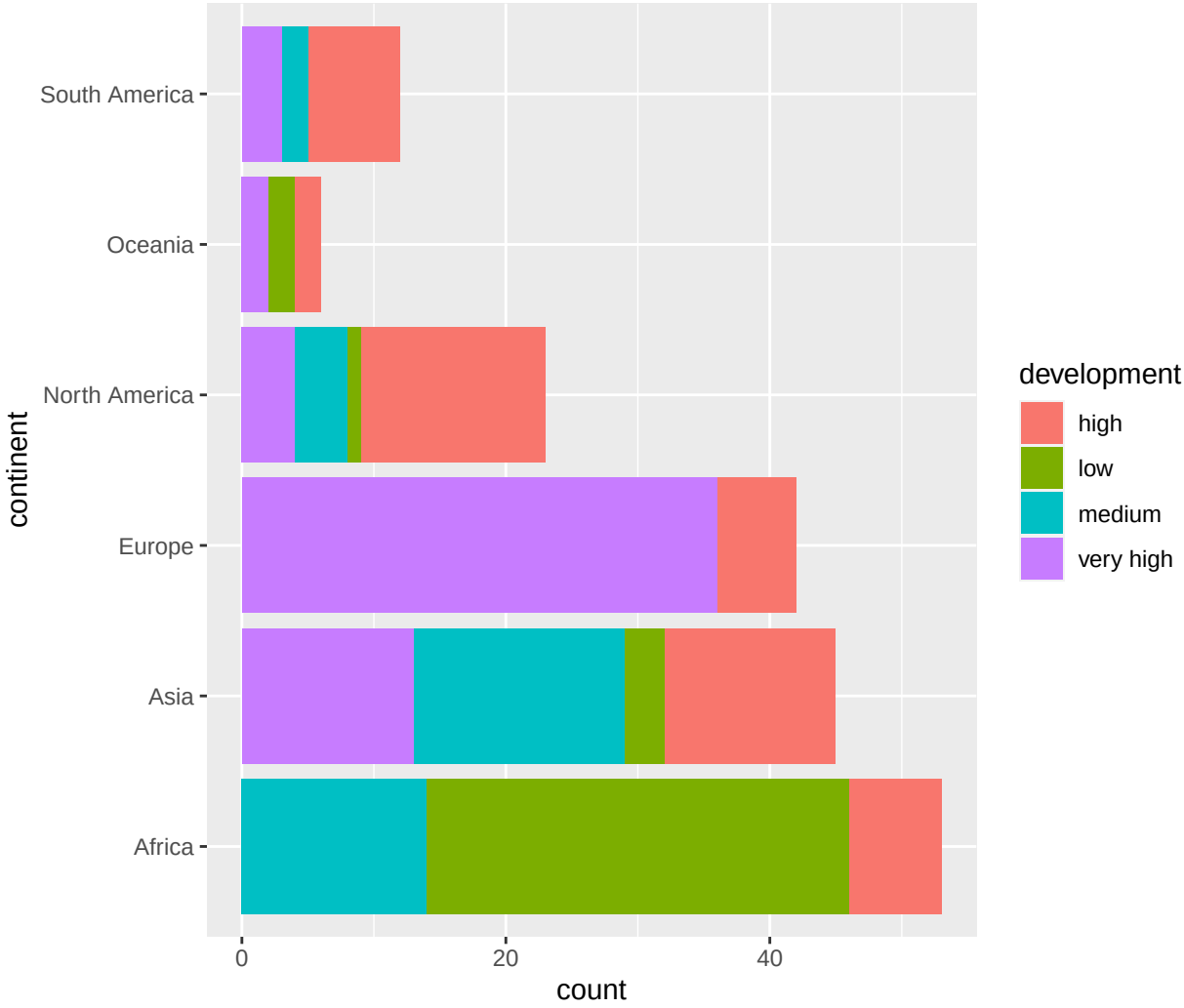
The *development* variable was constructed as follows from the *human_development_index* variable:

- **very high** for HDI of 0.800 and above
- **high** from 0.700 to 0.799
- **medium** from 0.550 to 0.699
- and **low** below 0.550.

We followed Wikipedia's criteria for the construction of this variable as it accurately represents and summarizes well HDI in 4 categories.

Countries per continent per HDI

```
ggplot(data=data) +
    geom_bar(aes(fill=development ,y=continent))
```



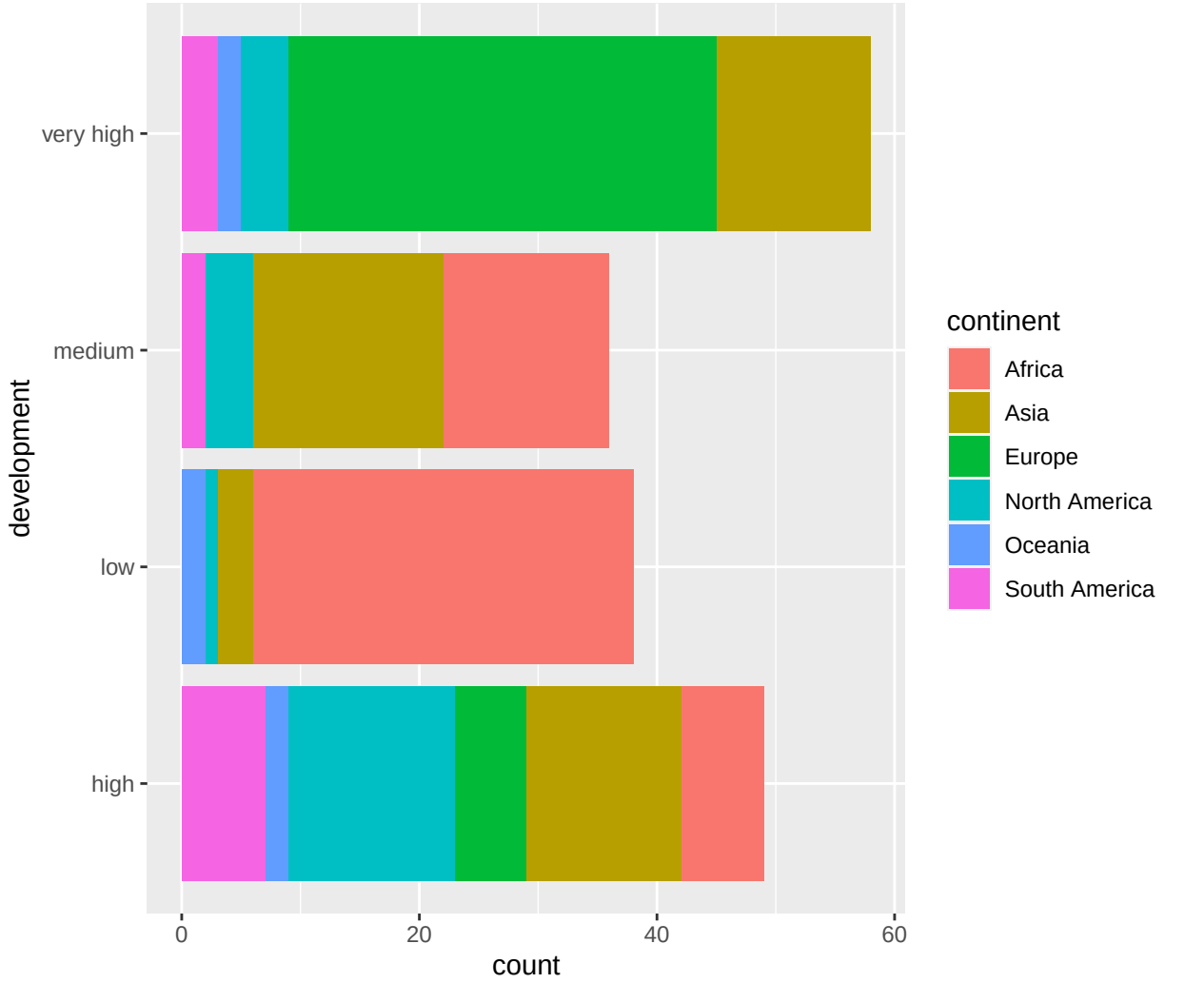
We can see here how many countries per continent correspond to which HDI using our constructed development variable. We can see Europe is fully composed of high and very high development countries. North america has most of its countries being high development countries, with a significant group of medium to low development countries, possibly located in Central America and the Caribbean. The very high development countries correspond to Bahamas, Barbados, Canada and the United States.

Africa is mostly composted of medium to low development countries and Asia is a bit of a tossup between high, very high and medium development countries.

To be fair, a country having high HDI does not imply that it isn't a developing country, as most countries with high HDI are, indeed, developing countries. This is simply a categorical representation of Human Development Index.

Proportions of HDI per continent

```
ggplot(data=data) +
    geom_bar(aes(fill=continent, y=development))
```



Looking at the opposite plot to the previous one we can see that most of the very high development countries are located in Europe and Asia. With not a single very high development country in Africa and with most of the low development countries located in Africa.

Medium development countries are mostly located in Asia, Africa and the Americas with not a single medium or low development country located in Europe.

Which again, does not suggest at all that perhaps poverty situations do not exist in the European continent. For instance, about 30% of the population of Albania live under 5.50 USD a day. This figure corresponds to nearly a third of the population and this particular trend shows up for other countries in the Balkan Peninsula along with Moldova and Romania (for more examples).

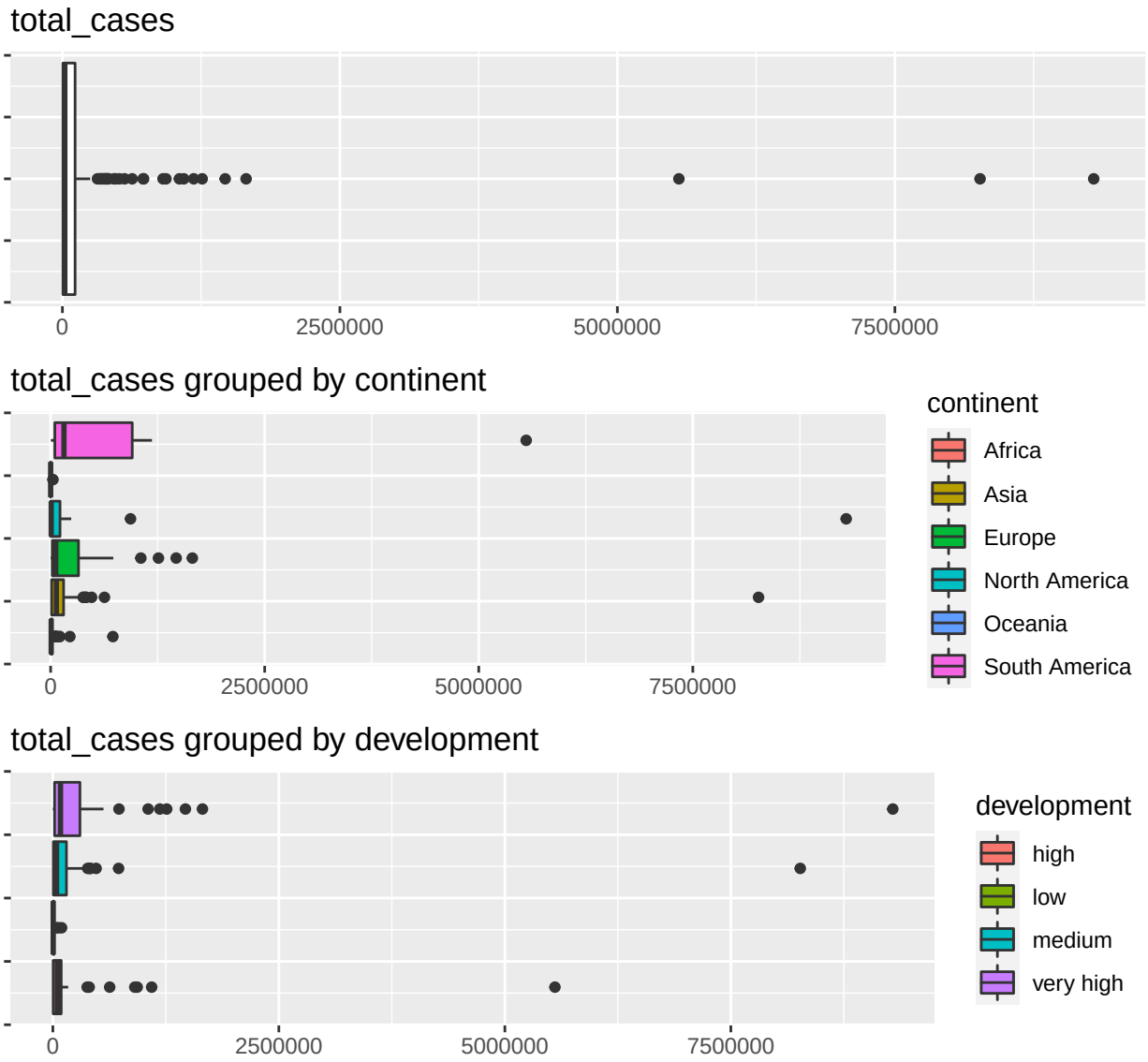
Plots with numerical variables

Function to plot quantitative variables

```
plots <- function(dataset ,col, type, density=TRUE, bins='default', xtick_angles='default') {
    var <- dataset %>% dplyr::select(col)
    if (bins == 'default') {bins = rep(10,3)}
    if (xtick_angles == 'default') {xtick_angles = rep(90,3)}
    if (type == 'boxplot') {
        p1 <- dataset %>% ggplot(aes(x=var[,1])) +
            geom_boxplot() +
            ggtitle(str_interp("${col}")) +
            theme(axis.title.x=element_blank(),axis.text.y=element_blank())
        p2 <- dataset %>% ggplot(aes(x=var[,1], fill=continent)) +
            geom_boxplot() +
            ggtitle(str_interp("${col} grouped by continent")) +
            theme(axis.title.x=element_blank(),axis.text.y=element_blank())
        p3 <- dataset %>% ggplot(aes(x=var[,1], fill=development)) +
            geom_boxplot() +
            ggtitle(str_interp("${col} grouped by development")) +
            theme(axis.title.x=element_blank(),axis.text.y=element_blank())
    } else if (type == 'hist') {
        p1 <- dataset %>% ggplot(aes(x=var[,1])) +
            geom_histogram(aes(y=..density..), bins=bins[1]) +
            geom_density() +
            ggtitle(str_interp("${col}")) +
            theme(axis.title.x=element_blank(),
                  axis.text.x = element_text(angle = xtick_angles[1]))
        if (density == FALSE) {
        p2 <- dataset %>% ggplot(aes(x=var[,1], fill=continent)) +
            geom_histogram(show.legend = FALSE,bins=bins[2]) +
            ggtitle(str_interp("${col} by continent")) +
            theme(axis.title.x=element_blank(),
                  axis.text.x = element_text(angle = xtick_angles[2])) +
            facet_wrap(~continent, nrow = 1)
        p3 <- dataset %>% ggplot(aes(x=var[,1], fill=development)) +
            geom_histogram(show.legend = FALSE,bins=bins[3]) +
            ggtitle(str_interp("${col} by development")) +
            theme(axis.title.x=element_blank(),
                  axis.text.x = element_text(angle = xtick_angles[3])) +
            facet_wrap(~development, nrow = 1)
        } else {
        p2 <- dataset %>% ggplot(aes(x=var[,1], fill=continent)) +
            geom_histogram(show.legend = FALSE,bins=bins[2],aes(y=..density..)) +
            geom_density(show.legend = FALSE) +
            ggtitle(str_interp("${col} by continent")) +
            theme(axis.title.x=element_blank(),
                  axis.text.x = element_text(angle = xtick_angles[2])) +
            facet_wrap(~continent, nrow = 1)
        p3 <- dataset %>% ggplot(aes(x=var[,1], fill=development)) +
            geom_histogram(show.legend = FALSE,bins=bins[3],aes(y=..density..)) +
            geom_density(show.legend = FALSE) +
            ggtitle(str_interp("${col} by development")) +
            theme(axis.title.x=element_blank(),
                  axis.text.x = element_text(angle = xtick_angles[3])) +
            facet_wrap(~development, nrow = 1)
        }}
    grid.arrange(p1,p2,p3, nrow=3)
}
```

Boxplots for total cases of COVID-19

```
plots(dataset=data, col='total_cases',type='boxplot')
```

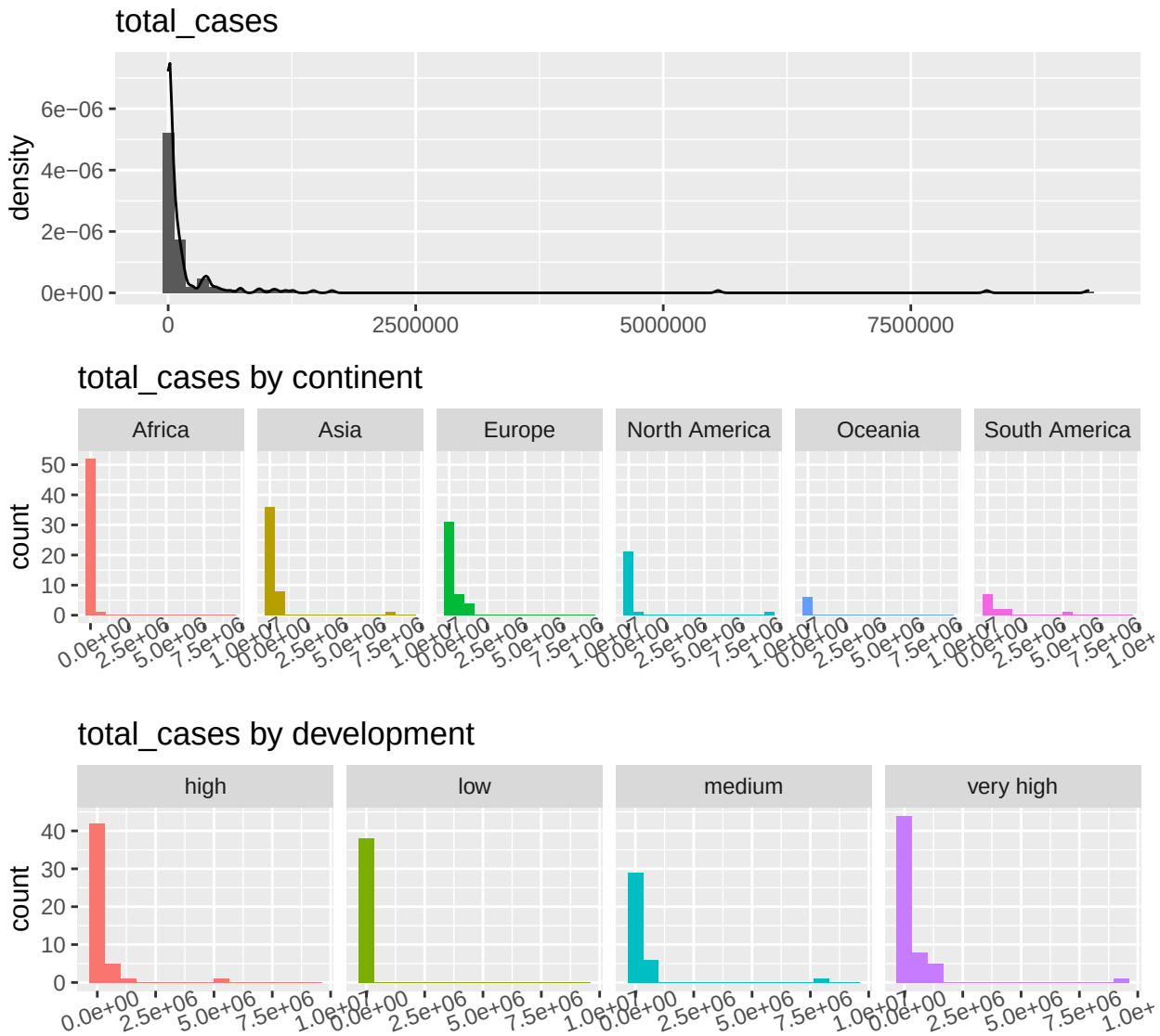


On the first box-plot of the variable, we can observe that the distribution of it is very right skewed, with some outliers. And the box-plots grouped by continents tell us that the country that has the most of the total cases is the US from continent North America which has a very high HDI. Meanwhile, the country that has the most of the total cases of Asia is India (second of the world), and has medium HDI. The third country that has the most of the cases is Brazil from South America with high HDI.

We can probably say that these three countries are the outliers for the variable total cases.

Histogram and kernel density for total cases of COVID-19

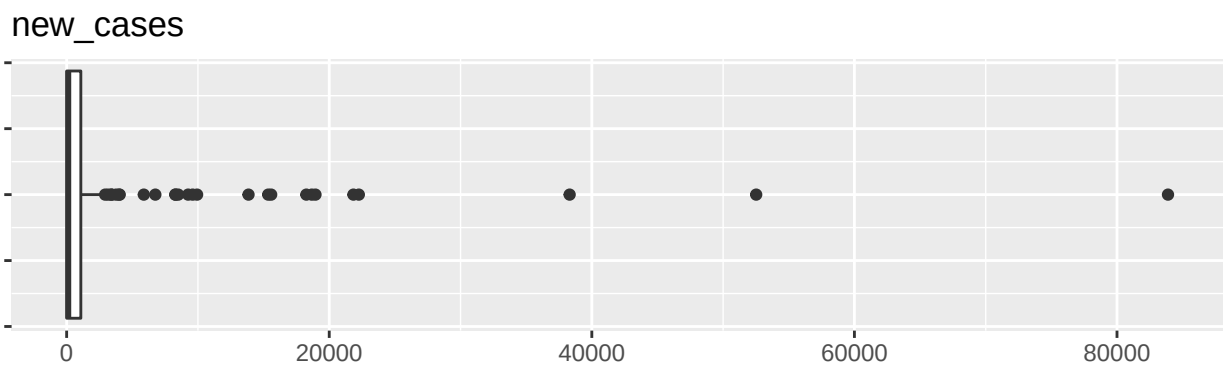
```
plots(dataset=data, col='total_cases',type='hist',
      density=FALSE, bins = c(80,15,15),xtick_angles=c(0,30,20))
```

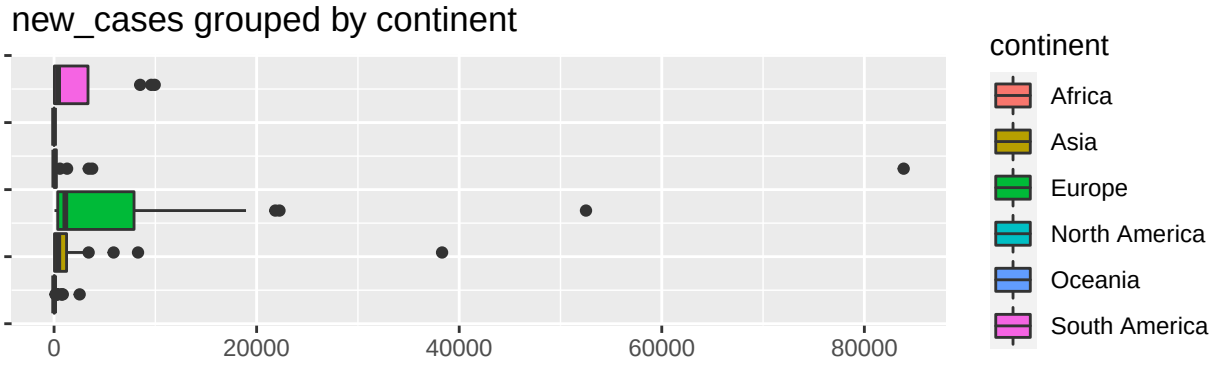


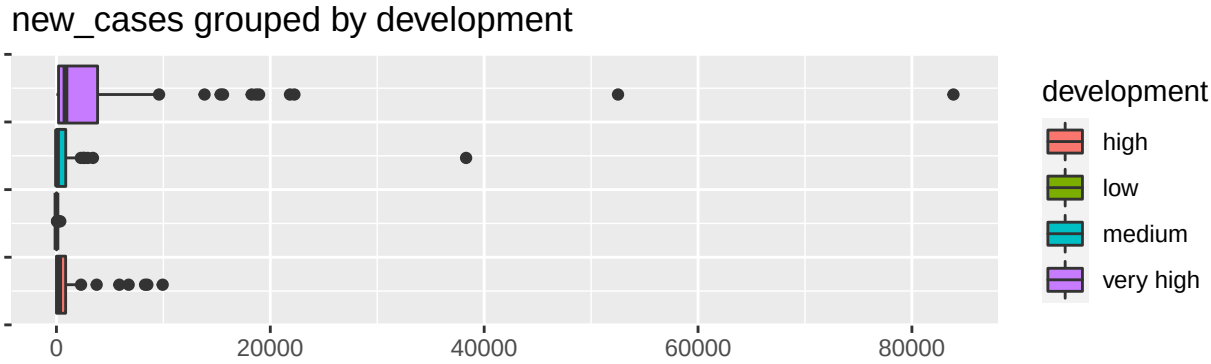
Observing this graph, we can confirm that the distribution is very right-skewed. Only Africa, Europe and Oceania don't have outliers. But it is probably because we don't have the dataset updated yet (we have the data-set updated on 3rd of November, 2020). About the development of different countries, we can't group the countries in terms of how they have developed by the total cases of COVID-19 they have.

Boxplots for new cases of COVID-19

```
plots(dataset=data, col='new_cases',type='boxplot')
```







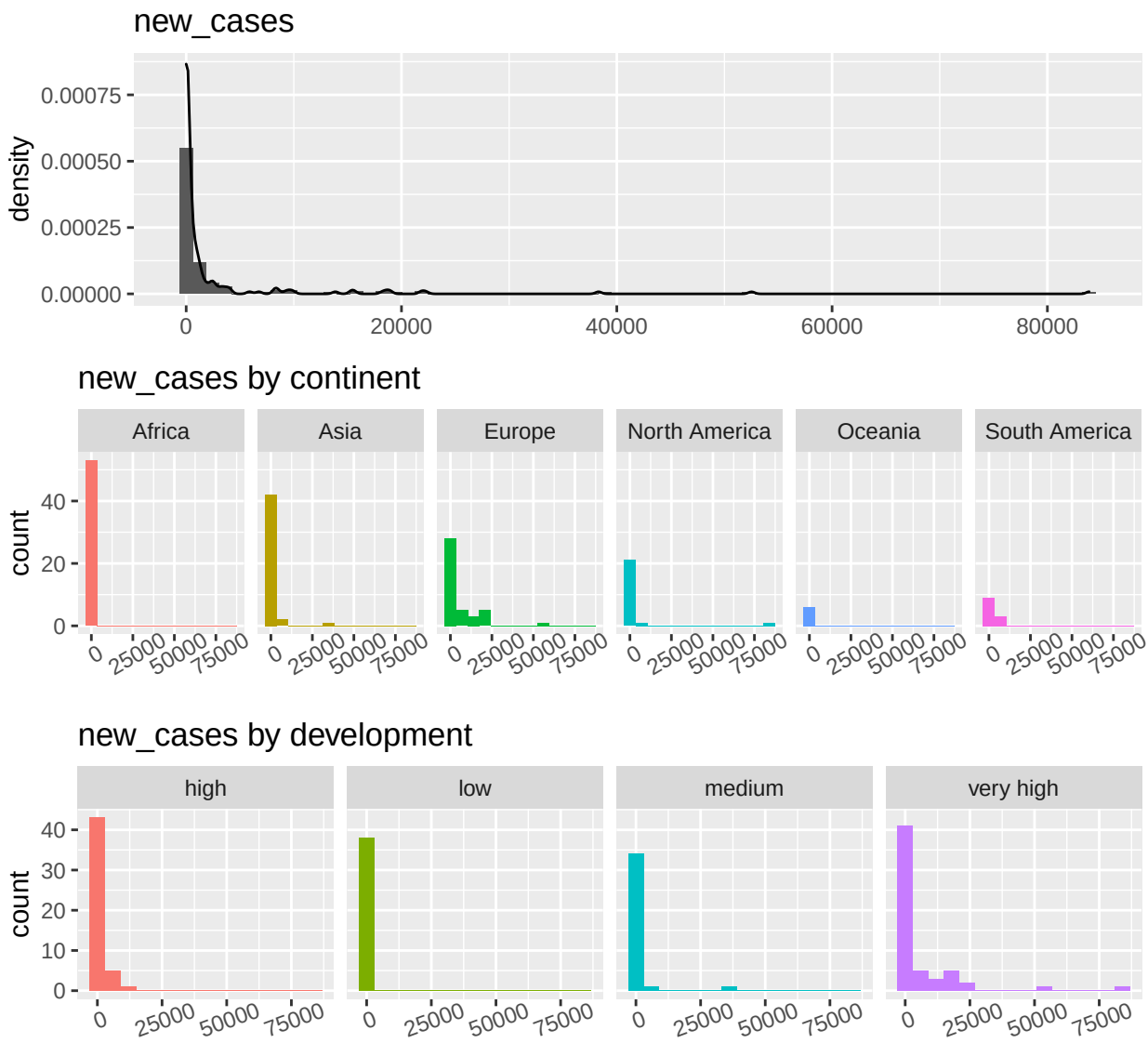
In the first box-plot above, we see that the distribution es very right skewed with some outliers. The country that has the most of the new cases is the US.

Observing the box-plot of new cases grouped by continent it is obvious that the country of North America that has the most of the new cases is the US. And the second country that has the most of the new cases is in Europe, France. Both of them have a very high Human Development Index. The third country that has the most of the new cases is India from Asia with medium HDI.

Another thing to mention is that the countries that have the most of the new cases is very related with the previous variable, which is the total cases, they have similar characteristics.

Histogram and kernel density for new cases of COVID-19

```
plots(dataset=data, col='new_cases',type='hist', density=FALSE,
      bins = c(70,13,15),xtick_angles=c(0,30,20))
```

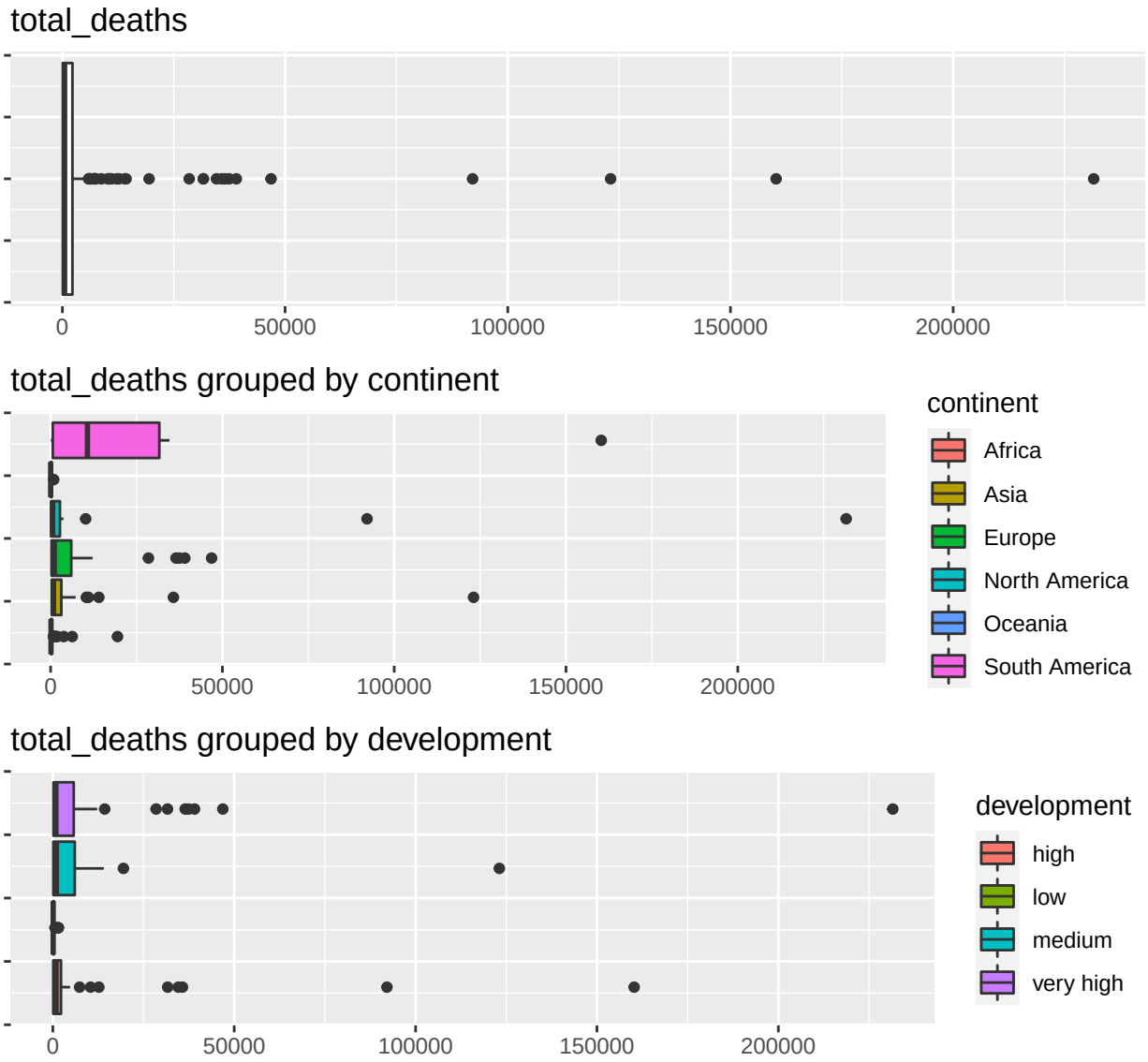


Observing this graph, we can confirm that the distribution is very right-skewed with some outliers (especially the US, France and India).

In the histograms of new cases by continent, we can see that Europe has the most dispersed distribution comparing to other continents, which means that the countries of Europe have very different values of new cases from each other. About the development of different countries, we can't group the coutries in terms of how they have developed by the new cases per day of COVID-19 they have, the distribution of countries that have a very high Human development Index have a some outliers.

Boxplots for total deaths due to COVID-19

```
plots(dataset=data, col='total_deaths',type='boxplot')
```

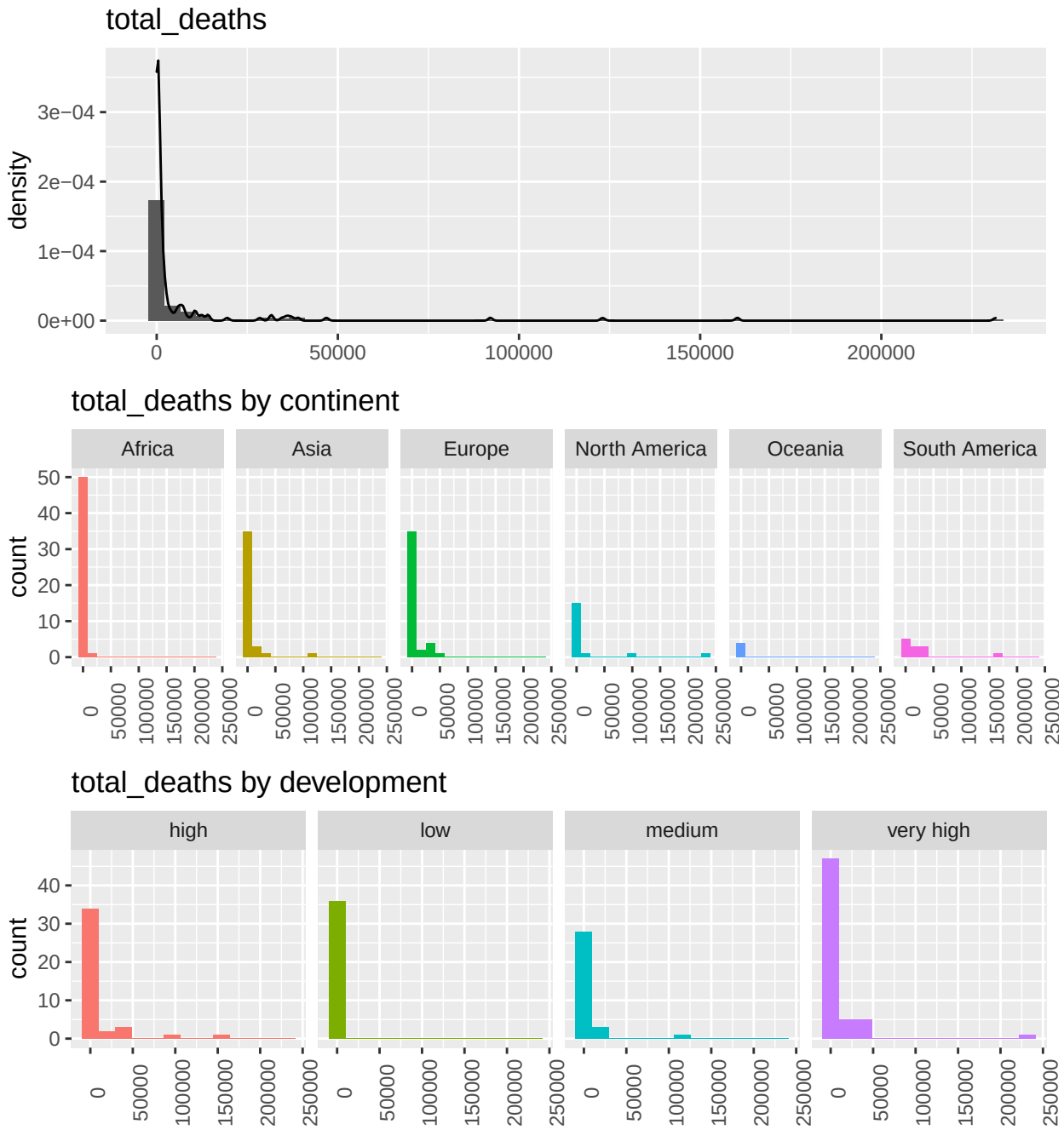


From the box-plots above we can say that this variable of new deaths is very likely distributed with the variables total cases and new cases. These three variables are all right-skewed and all have some outliers.

In this case, the country that has the most of the total deaths is still the US. It is obvious that the country of North America that has the most of the total deaths is the US. And the second country that has the most of the total deaths is in South America, Brazil. The third country that has the most of the total deaths is from Asia, India.

Histogram and kernel density for total deaths due to COVID-19

```
plots(dataset=data, col='total_deaths',type='hist', density=FALSE,
      bins = c(55,15,13),xtick_angles=c(0,90,90))
```

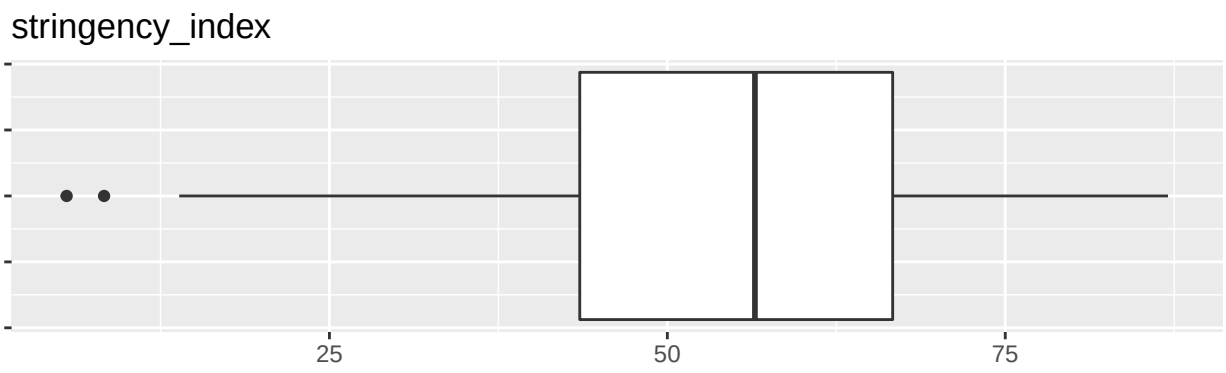


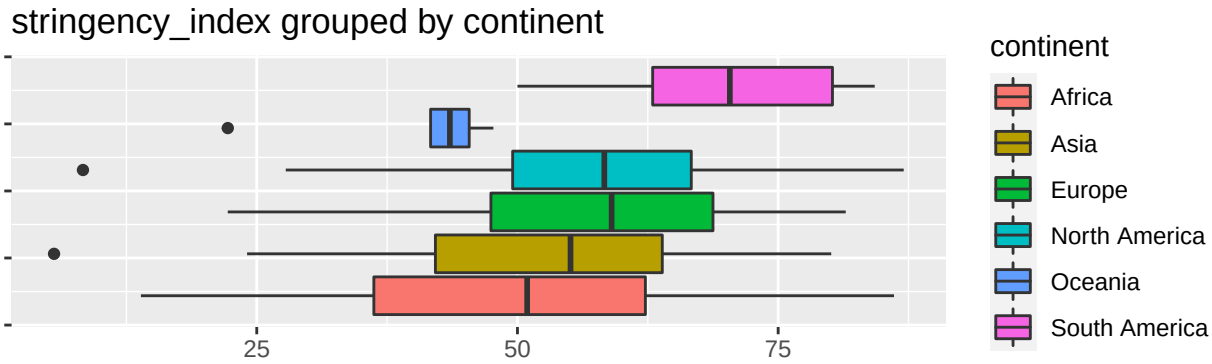
Again, the distribution is very right-skewed with some outliers (the US, Brazil and India).

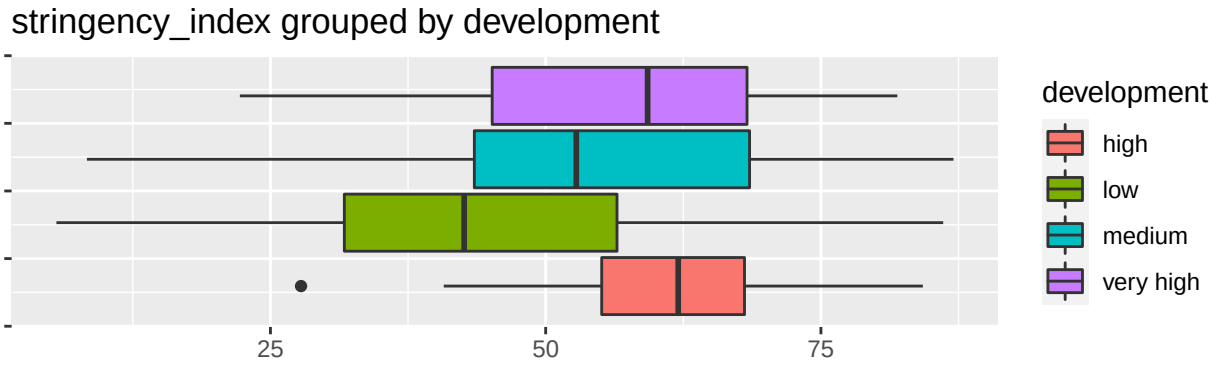
In the histograms of new cases by continent and histograms by development, the distributions are also very right-skewed, some of them have outliers. We still can't group the coutries in terms of how they have developed by the total deaths of COVID-19 they have, the distribution of countries that have a very high, high and medium Human development Index have a some outliers.

Boxplots for stringency index (how strict measures are)

```
plots(dataset=data, col='stringency_index',type='boxplot')
```







From the box-plot above we can say that the global distribution of stringency index is a little left skewed, but it is the most normally distributed until now. There are some countries that have really low stringency index, for example Afghanistan from Asia has stringency index equals to 5.56, or Nicaragua from North America with stringency index 8.33.

Observing the box-plot of stringency index grouped by continent we can see that South America has the most strict measurements and Oceania has the least strict measurements. The rest of them have similar distribution on stringency index. Look at the stringency index grouped by development we notice that the countries which have low HDI have less stringency index (using the criterion of quantiles).

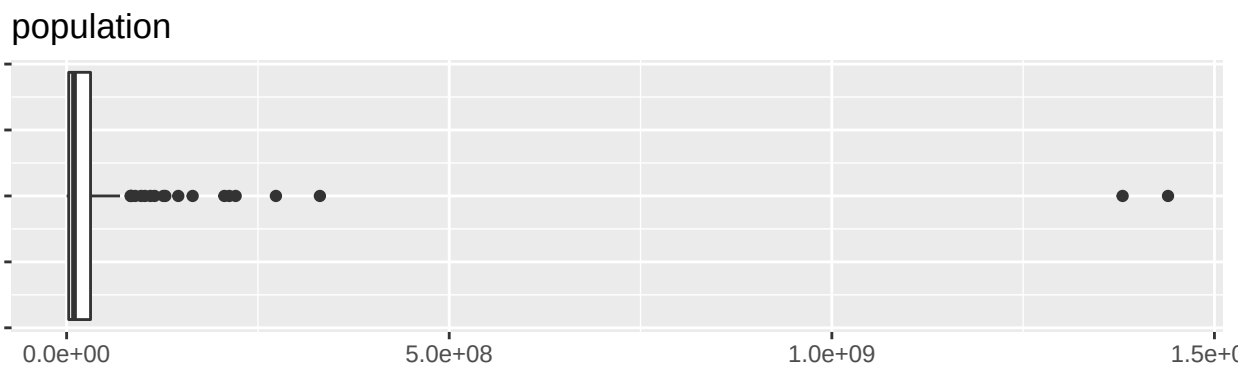
Histogram and kernel density for stringency index

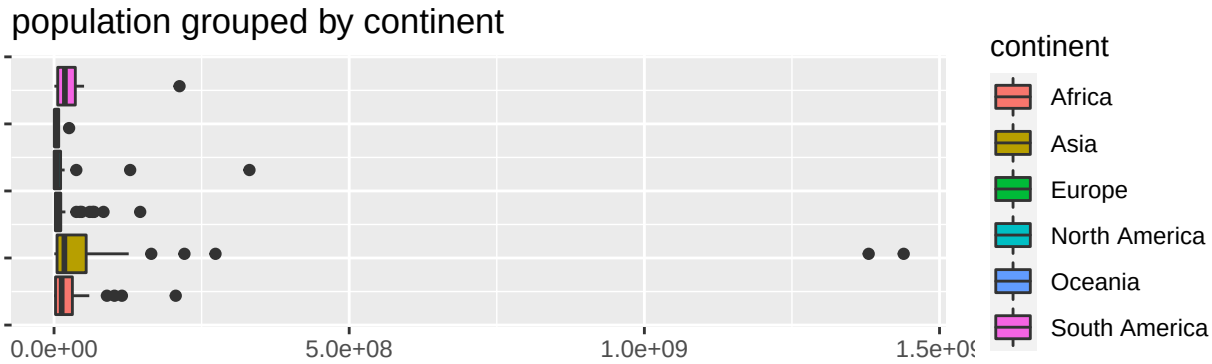
```
plots(dataset=data, col='stringency_index',type='hist',
      density=TRUE,xtick_angles=c(0,30,20))
```

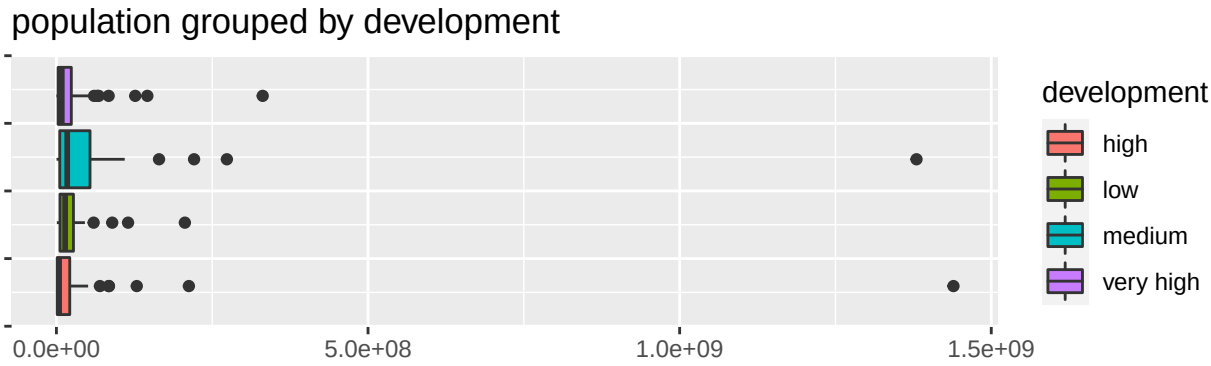
The distribution is quite symmetric distributed for different continents. Except for Oceania, there are some countries have really low stringency index; and South America, the distribution of this variable is quite left skewed. We can probably distinguish the countries with low HDI from others, these countries usually have less stringency index.

Boxplots for population

```
plots(dataset=data, col='population',type='boxplot')
```



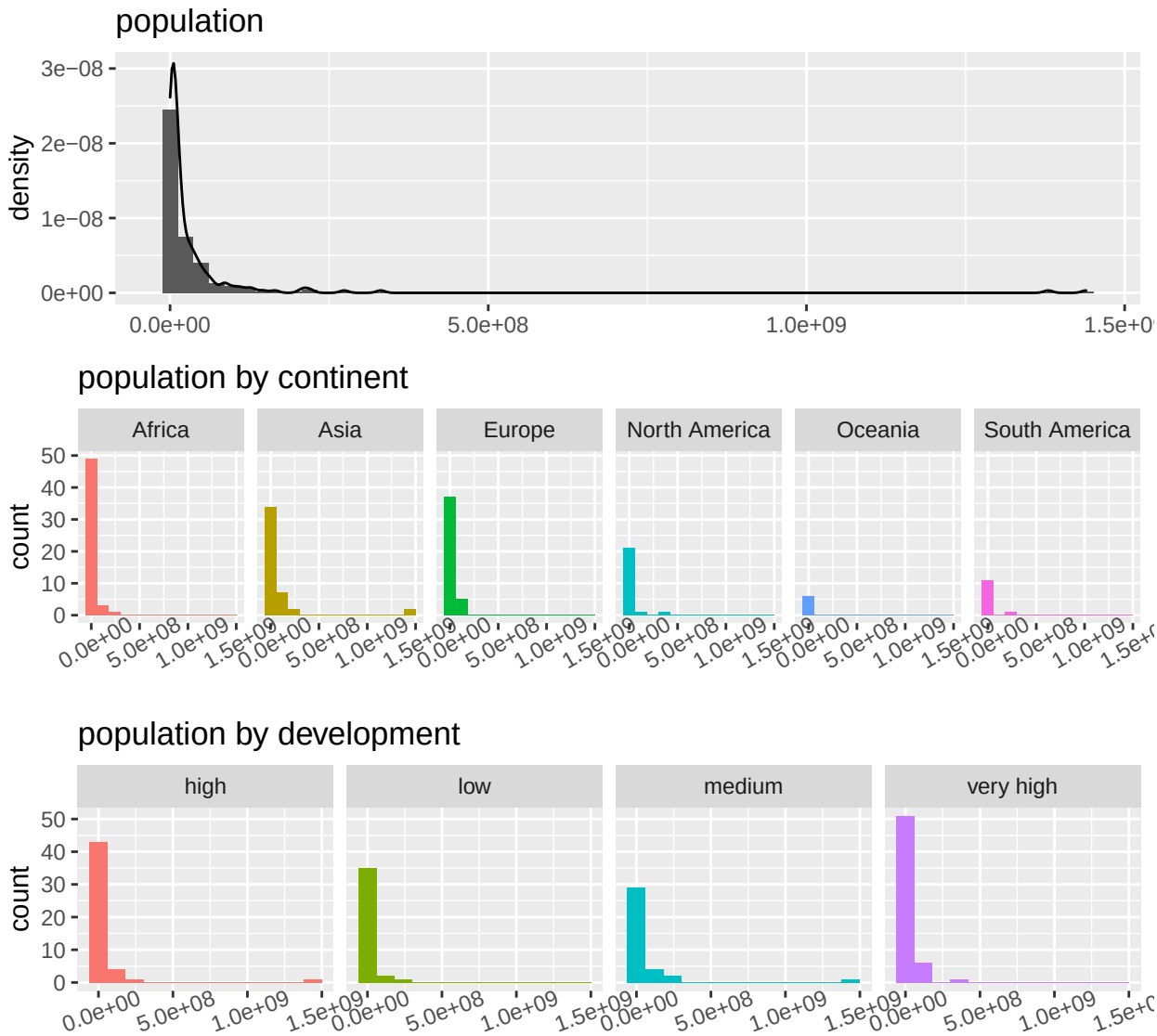




Observing the box-plot above we can see that the distribution of population is very right skewed, some countries have way more population than others. For example, China has the most population of all, India is the second, these two countries are most exaggerated outliersfrom the plot.

Histogram and kernel density for population

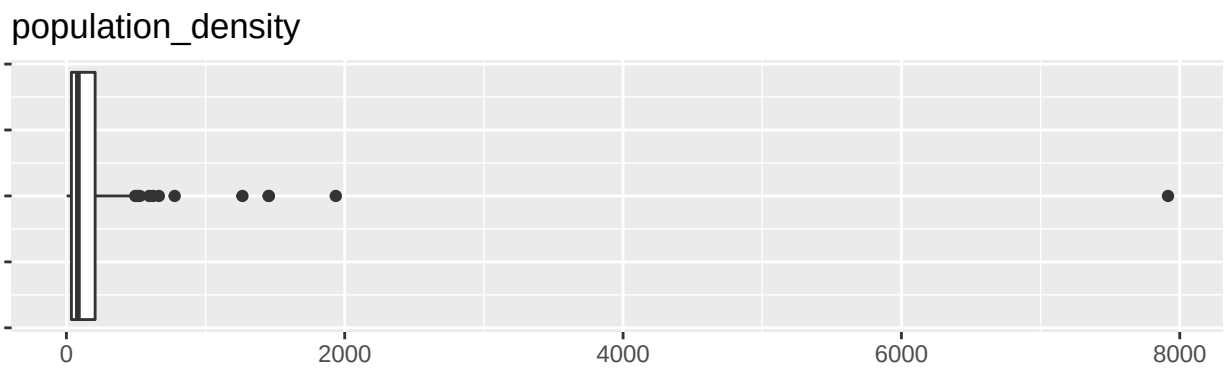
```
plots(dataset=data, col='population',type='hist', density=FALSE,
      bins = c(60,13,13),xtick_angles=c(0,30,20))
```

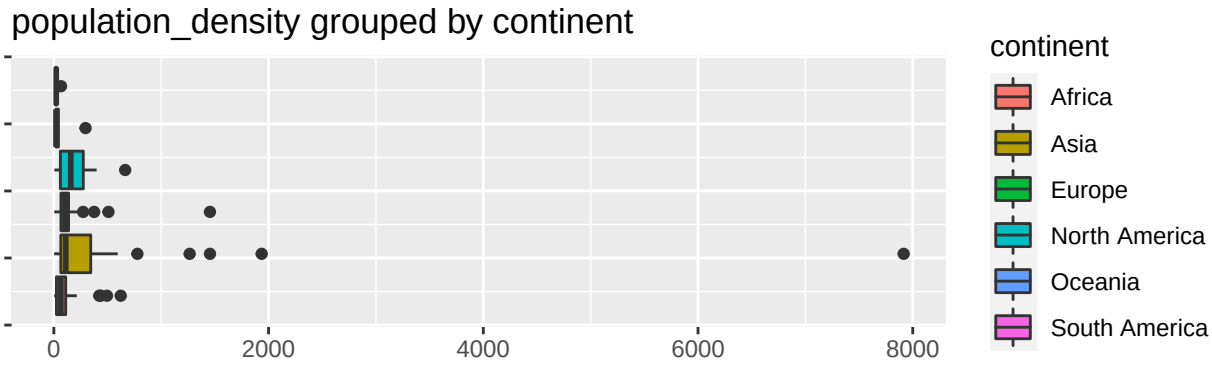


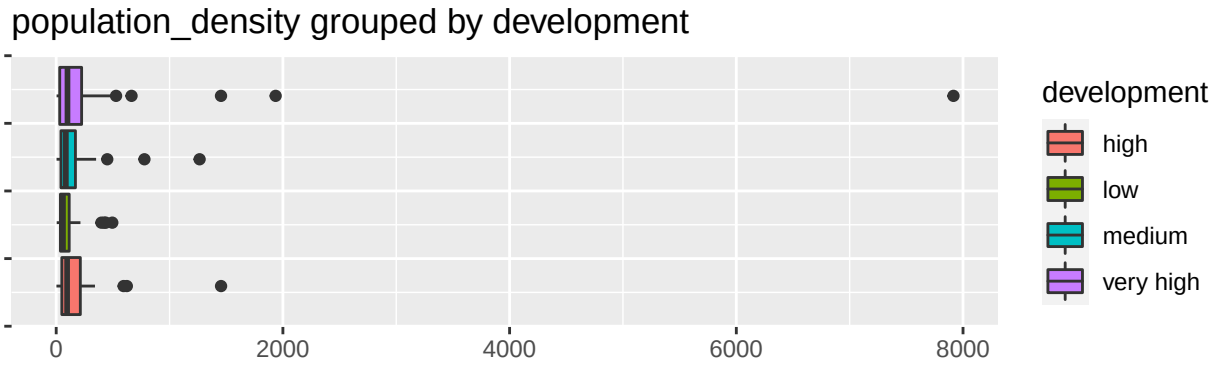
The distribution is very right-skewed, the population of each country is very different from others, but the variable population does not provide any information of whether the country has high HDI or not.

Boxplots for population density

```
plots(dataset=data, col='population_density',type='boxplot')
```



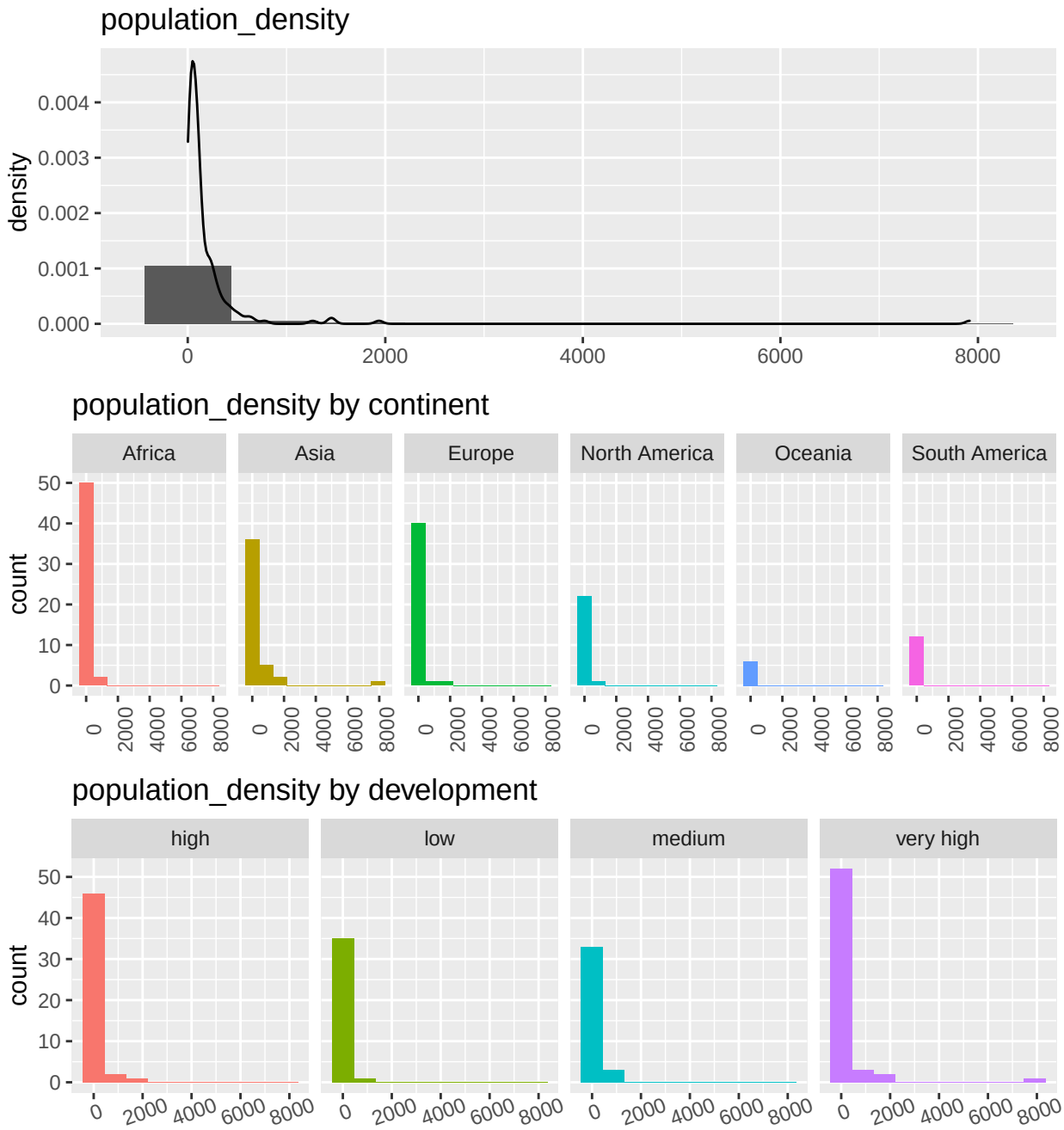




Population density is a measurement of population per unit area. Observing the previous box-plot above we can see that the distribution of population density is very likely distributed as population distribution, it is very right skewed, some countries have really high population density. For instance, Singapore has the most population density of all with a value of 7915.731, it is a small country of Asia with very high HDI.

Histogram and kernel density for population density

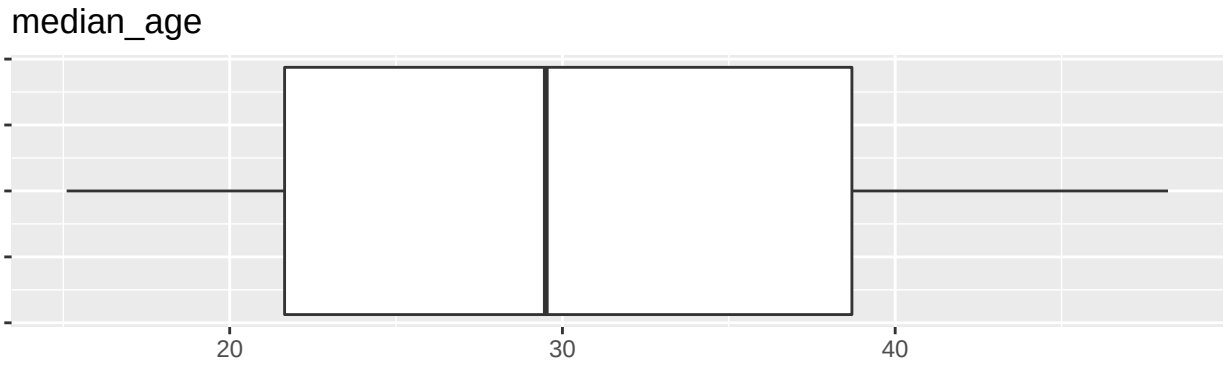
```
plots(dataset=data, col='population_density',type='hist',
      density=FALSE, xtick_angles=c(0,90,20))
```

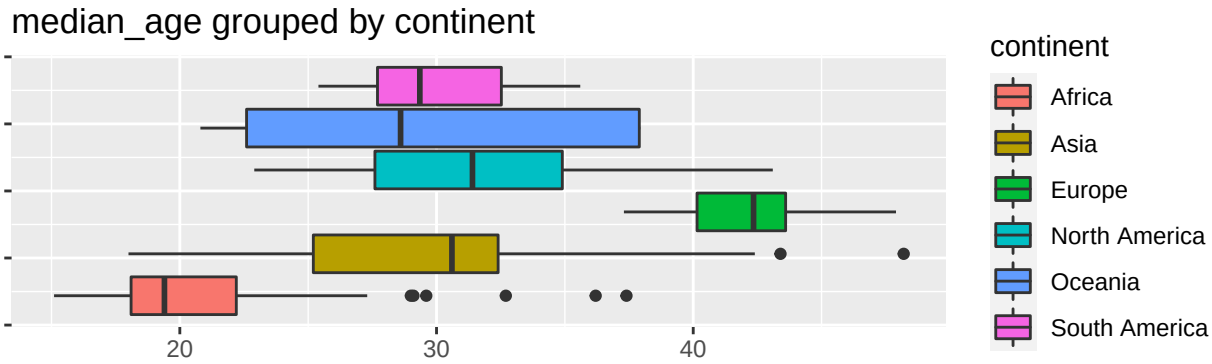


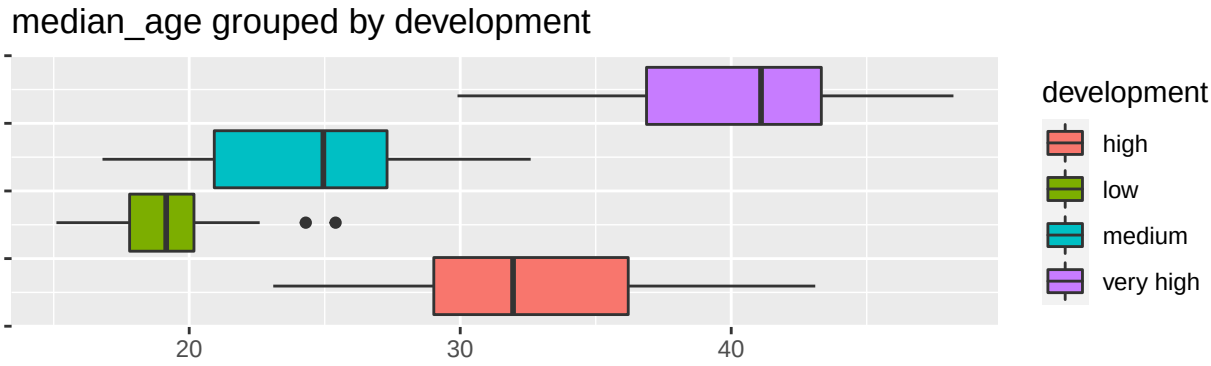
The distribution is very right-skewed, the population density of each country is very different from others. And the variable does not provide any information of whether the country has high HDI or not.

Boxplots for median age

```
plots(dataset=data, col='median_age',type='boxplot')
```







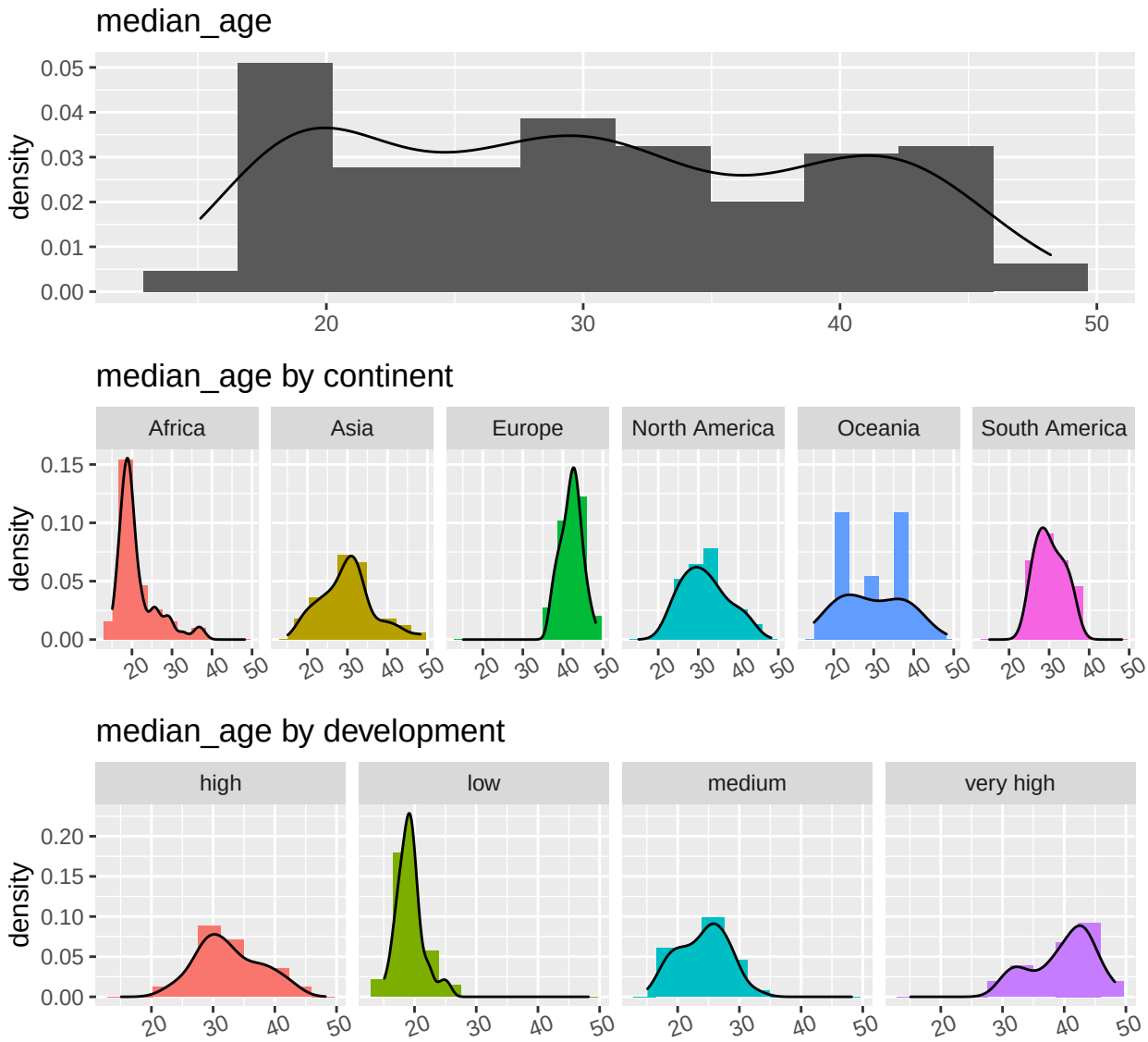
Observing the box-plot for the global median age we can notice that the distribution of it is quite symmetric. The majority of the median age of different countries is located between 20-40.

But from the grouped box-plots we can find something really interesting:

- For the box-plots grouped by continent we can see that the median age of Europe is larger (more than 40) than the rest of the continents while Africa has the least median age (less than 20) with some "outliers" that have similar median age as other continents.
- For the box-plots grouped by development we detect that usually higher developed a country, larger the median age, e.g. the countries that have very high HDI have median of median age more than 40, and the countries that have low HDI have median of median ge less than 20. From that, we can conclude that the majority of countries from Africa has low HDI while majority of countries from Europe has very high HDI.

Histogram and kernel density for median age

```
plots(dataset=data, col='median_age',type='hist',
      density=TRUE, xtick_angles=c(0,30,20))
```

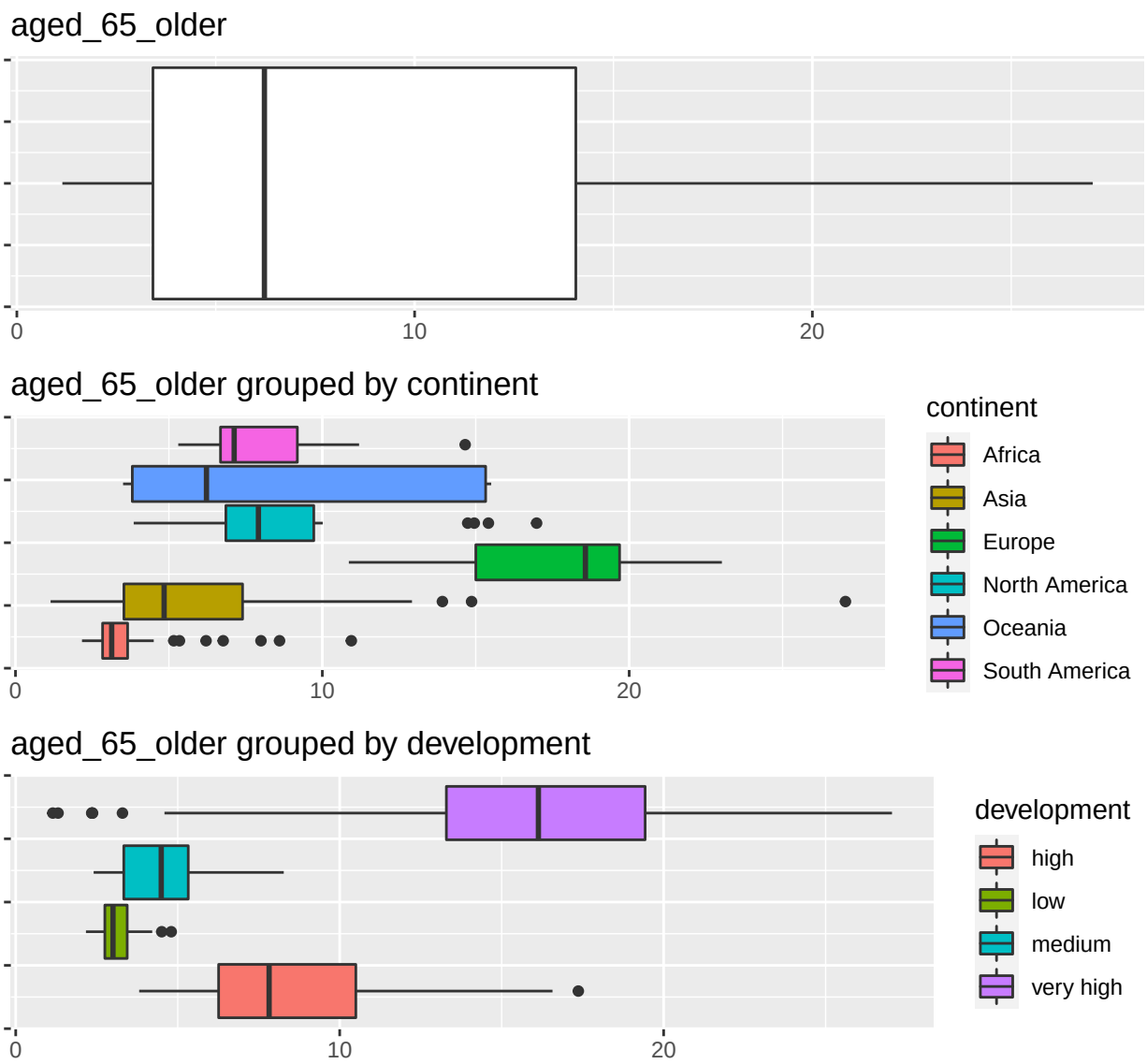


The distributions of median age for different continent are very different. The distribution of Africa is right skewed while others are symmetric.

The distributions of Asia, North America, Oceania and South America are more flat (platykurtic), and the distributions of Africa and Europe are more concentrated (leptokurtic).

Boxplots for the percentage of population aged 65 or older

```
plots(dataset=data, col='aged_65_older',type='boxplot')
```



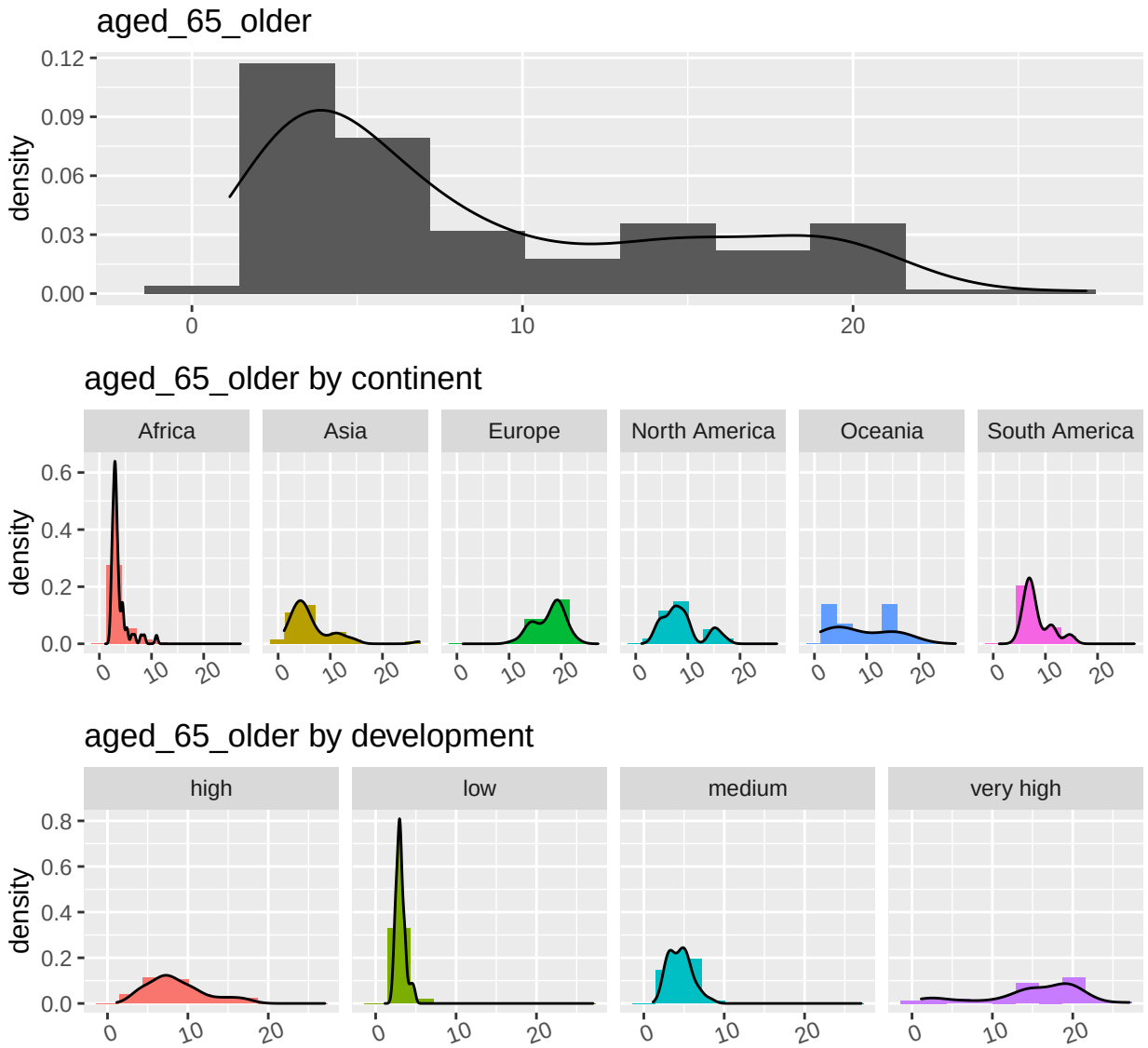
From the box-plot for the global percentage of population aged 65 or older we can notice that the distribution is right skewed. There are more than half of the countries have less than 10% of population of aged 65 or older.

However, the grouped box-plots give us more interesting results:

- In the box-plots grouped by continent we can see that the percentage of population aged 65 or older of Europe is larger (more than 10%) than the rest of the continents while Africa has the least median age (generally less than 10%) with some "outliers" that have similar values as other continents.
- In the box-plots grouped by development we detect that usually higher developed a country, larger the median age. These box-plots kind of give us the same information as the median age of each country, although this variable is not as clear as the previous one, median age.

Histogram and kernel density for the percentage of population aged 65 or older

```
plots(dataset=data, col='aged_65_older',type='hist',
      density=TRUE, xtick_angles=c(0,30,20))
```

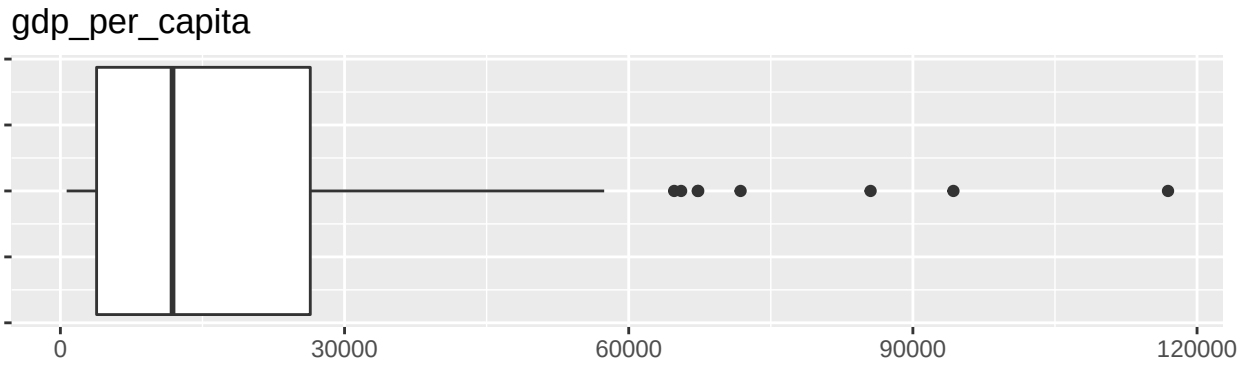


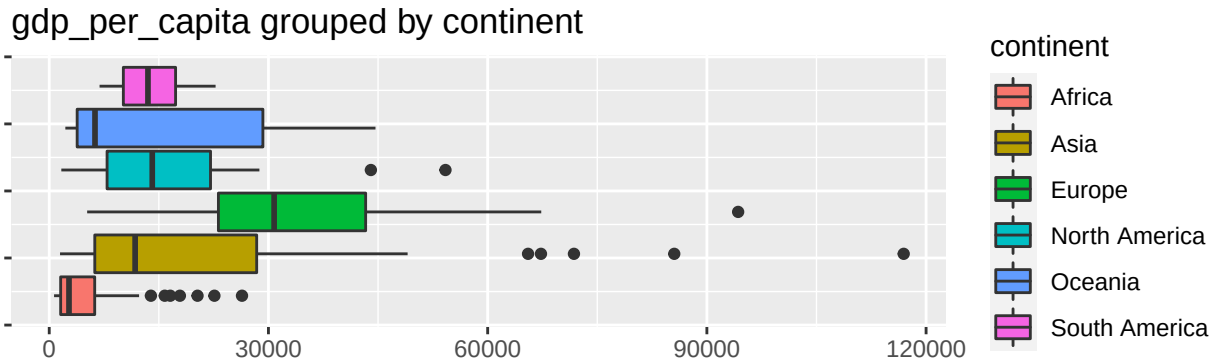
The distributions of this variable for different continent are very different. The distributions of Africa, Asia, North America and South America are right skewed while the distribution of Europe is left skewed.

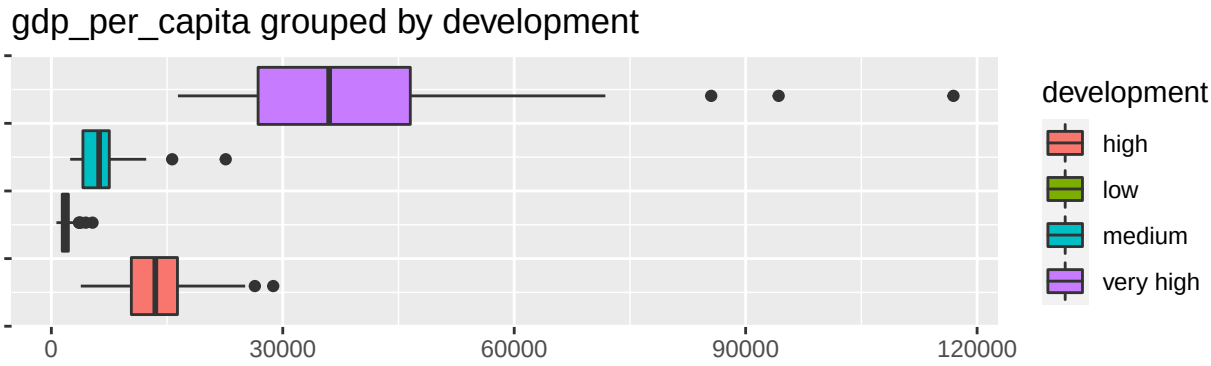
The distribution of Africa is more concentrated (leptokurtic) while others are more flat. The distributions of countries that have low HDI are more concentrated, and they usually have less percentage of population of aged 65 or older.

Boxplots for GDP per capita

```
plots(dataset=data, col='gdp_per_capita',type='boxplot')
```







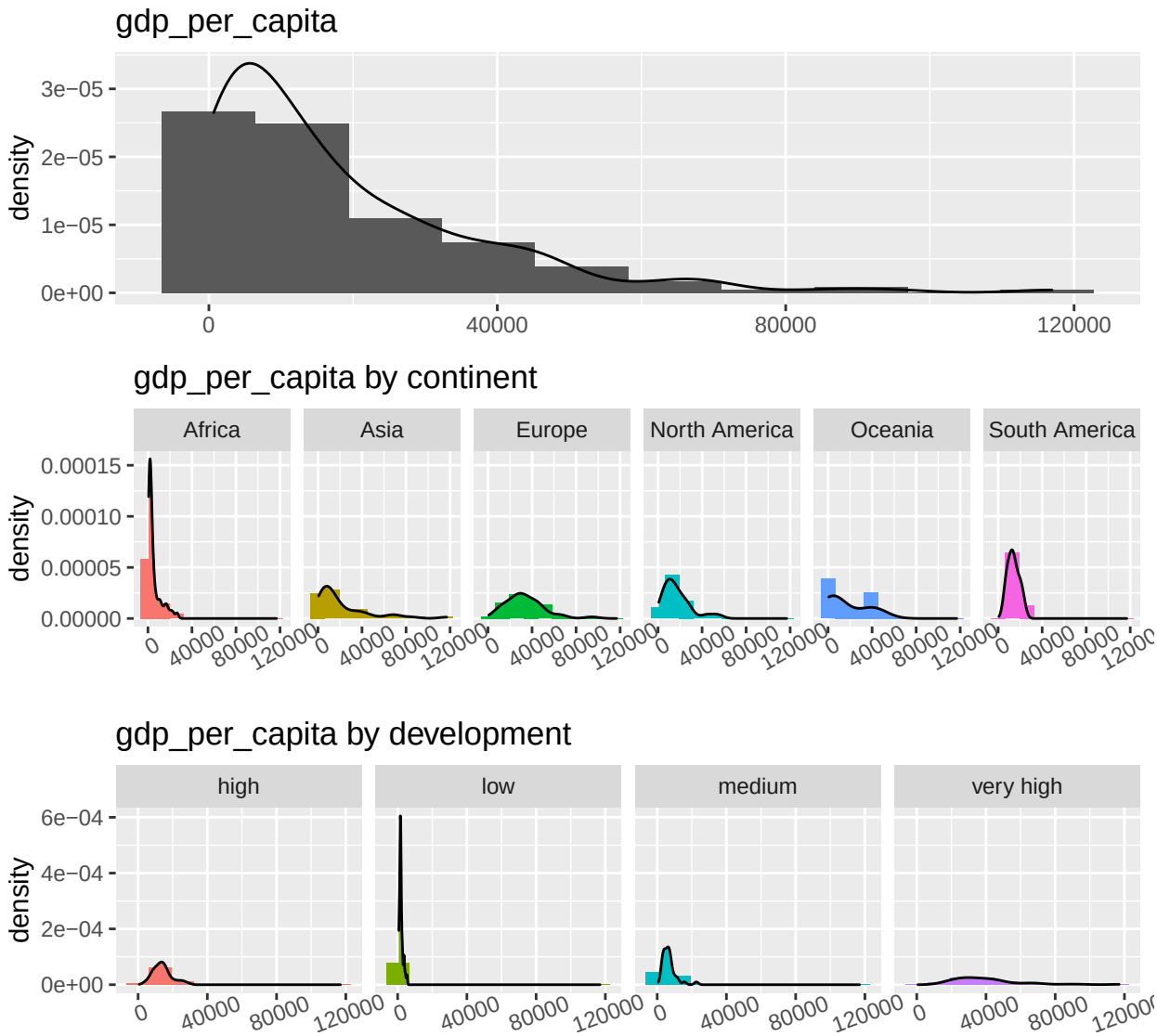
Observing the box-plot for the global GDP per capita we can see that the distribution is right skewed. The country that has the highest GDP per capita is Qatar from Asia, then comes Luxembourg and Singapore, all of them have a very high HDI.

The grouped box-plots provide more interesting conclusions:

- In the box-plots grouped by continent we observe that the GDP per capita of Europe is a little bit higher than the rest of the continents while Africa has the least median of GDP per capita with some "outliers" that have similar values as other continents.
- Nevertheless, the box-plots grouped by development give us more relevant information. We can more or less define whether a new country has very high, high, medium or low HDI by having its GDP per capita. Due to the clear difference of GPD per capita between the different levels of HDI. The countries that have very high HDI often have larger GDP per capita, and the countries with low HDI have less GDP per capita. There is a very clear correlation between these two variables.

Histogram and kernel density for GDP per capita

```
plots(dataset=data, col='gdp_per_capita',type='hist',
      density=TRUE, xtick_angles=c(0,30,20))
```

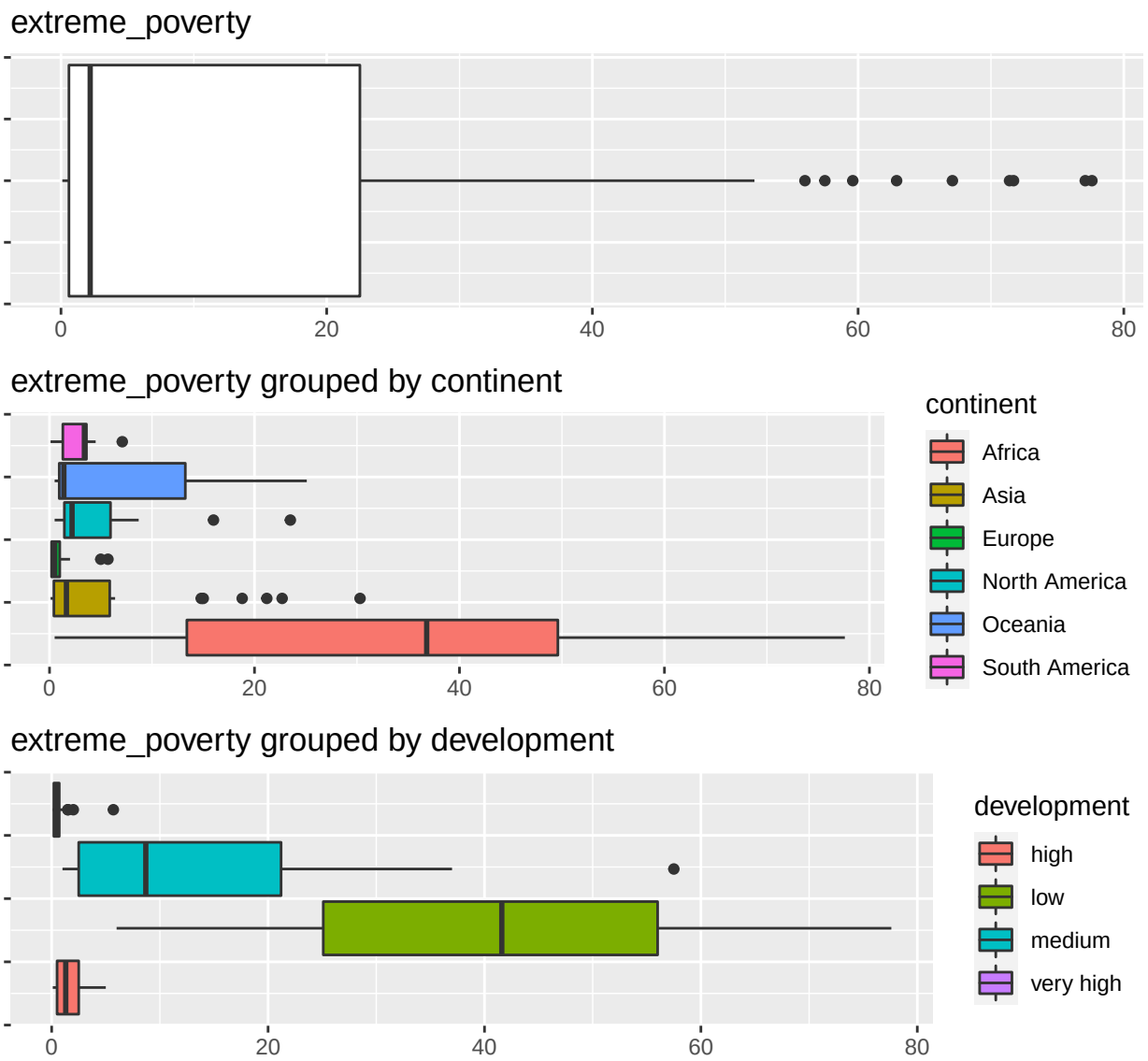


The global distribution of this variable is very right skewed. The distributions of Africa is more concentrated (leptokurtic) while others are more flat (platykurtic).

The distributions of countries that have low HDI are more concentrated, and they usually have less GDP per capita.

Boxplots for percentage of population in extreme proverty

```
plots(dataset=data, col='extreme_poverty',type='boxplot')
```



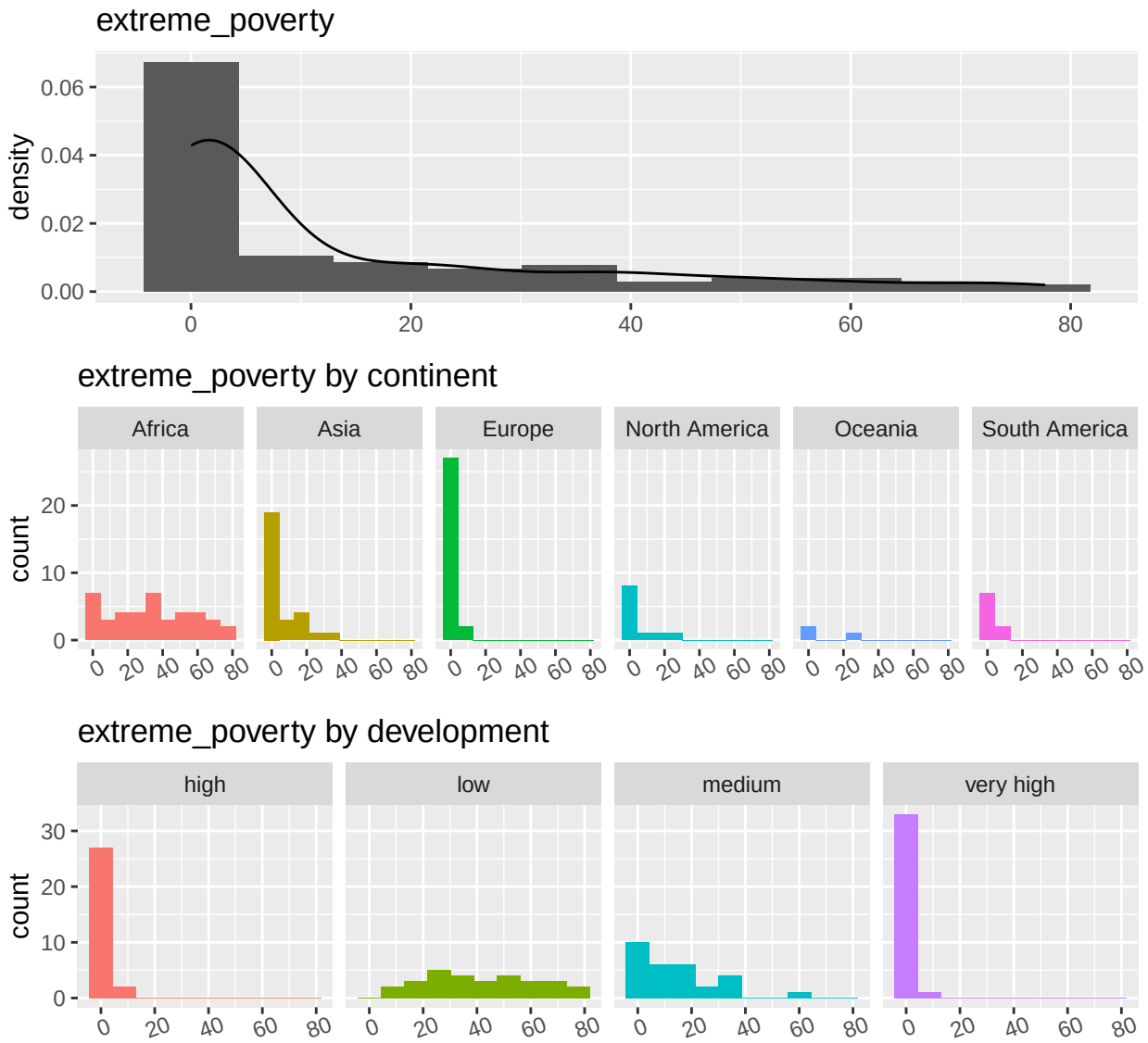
From the box-plot of the global extreme poverty we can observe that the distribution is right skewed. The country that has the highest extreme poverty is Madagascar, then comes Democratic Republic of Congo and Burundi, all of them are from Africa.

However, the grouped box-plots provide more interesting conclusions:

- The box-plots grouped by continent tell us that the median of the extreme poverty of Europe is the least of all the continents while Africa has the highest median extreme poverty.
- The box-plots grouped by development give us more important information. The countries that have higher index of extreme poverty often have low HDI while the countries with lower extreme poverty have higher HDI. There is a quite clear correlation between these two variables.

Histogram and kernel density for percentage of population in extreme proverty

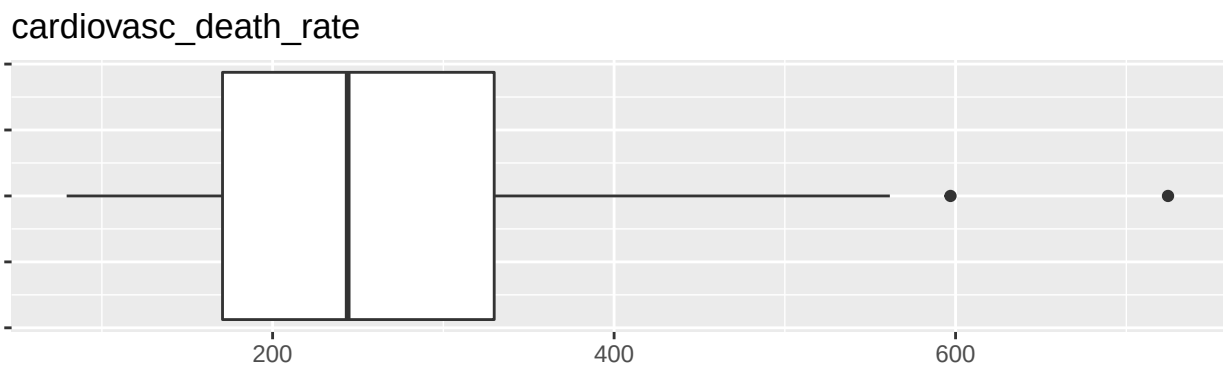
```
plots(dataset=data, col='extreme_poverty',type='hist',
      density=FALSE, xtick_angles=c(0,30,20))
```

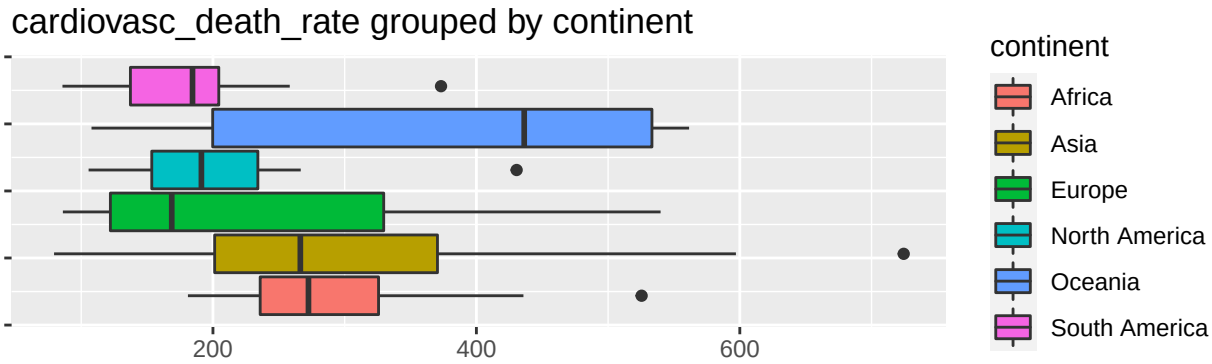


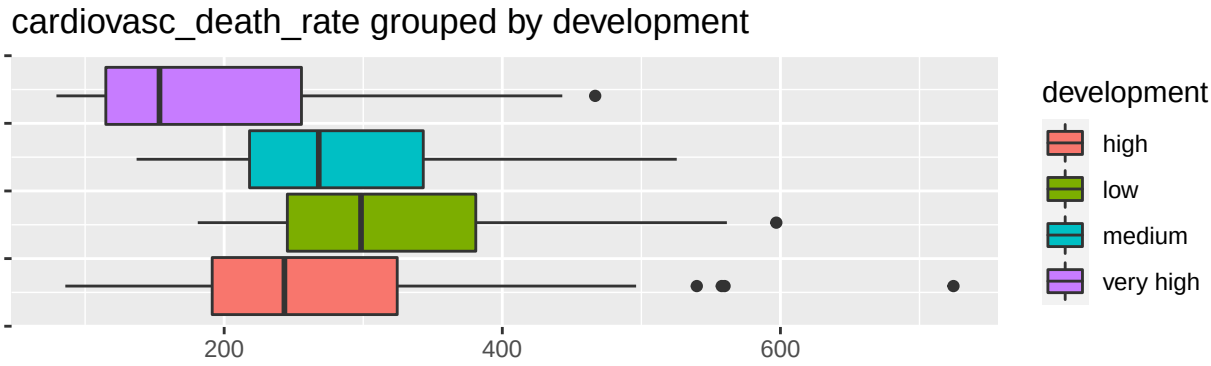
The global distribution of this variable is very right skewed. The distributions of Europe is more concentrated (leptokurtic) in low values while others are more flat (platykurtic). The distributions of countries that have very high HDI are more concentrated, and they usually have lower extreme poverty.

Boxplots for cardiovascular death rate

```
plots(dataset=data, col='cardiovasc_death_rate',type='boxplot')
```







It appears that most countries seem to have a cardiovascular death rate between 170.67 and 329.79 deaths per 100,000 inhabitants. With Uzbekistan being in the absolute extreme, with about 724 deaths by cardiovascular disease per 100,000 inhabitants.

Grouping by continent we see that Oceania seems to have the largest box (probably due to its lower amount of countries), with a few extreme cases per continent. On average the continent with the highest death rate due to cardiovascular disease is Oceania, followed by Asia. Seems like cardiovascular disease in the Americas could be a less common cause of death than in the rest of the world.

By development we can see a bit of a pattern, where the least developed a country is, the higher its cardiovascular death rate. However, even if we see this pattern, we can't confidently say that living in a less developed country makes an individual more likely to die from cardiovascular disease. There are definitely many other factors that affect such rate per HDI.

Histogram and kernel density for cardiovascular death rate

```
plots(dataset=data, col='cardiovasc_death_rate',type='hist',
      density=TRUE, bins=c(13,10,16), xtick_angles=c(0,0,0))
```

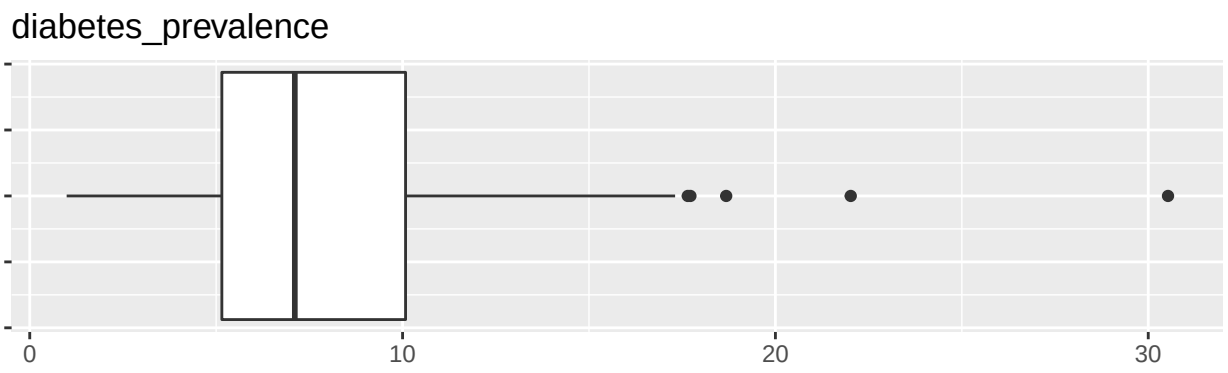
For cardiovascular death rate we see a similar story here than with the general boxplot. The larger concentration of countries clumps around the previously mentioned interval, and the distribution of the variable as is is somewhat normal-like with a relatively long left tail.

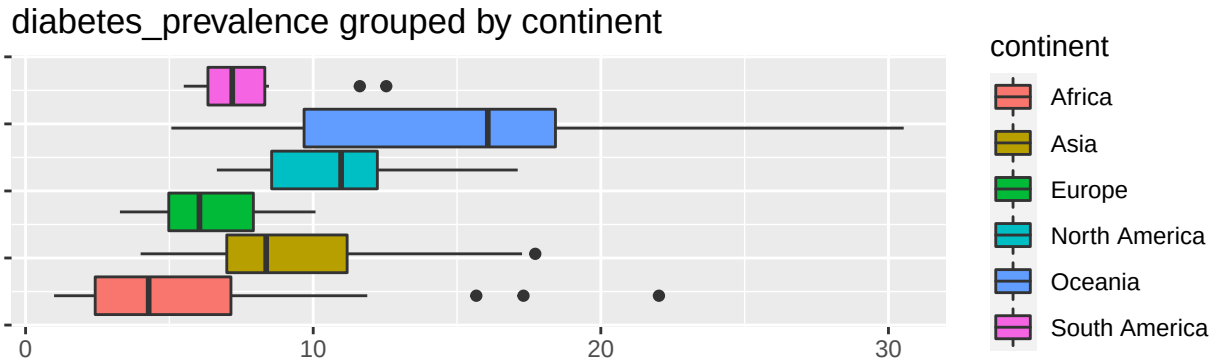
Per continent we see that Europe has a significant concentration of countries below 200, along with South America, which, on average, is the continent with the lowest death rate from cardiovascular disease according to our data. For Oceania we see a flat distribution with some high numbers and low numbers, of course, we know that there's less data points, therefore our main concentration below 200 corresponds to New Zealand and Australia, and the rest of the countries seem to have a higher death rate than the rest. Asia's left tail suggests a few other countries with a very high cardiovascular death rate like Uzbekistan.

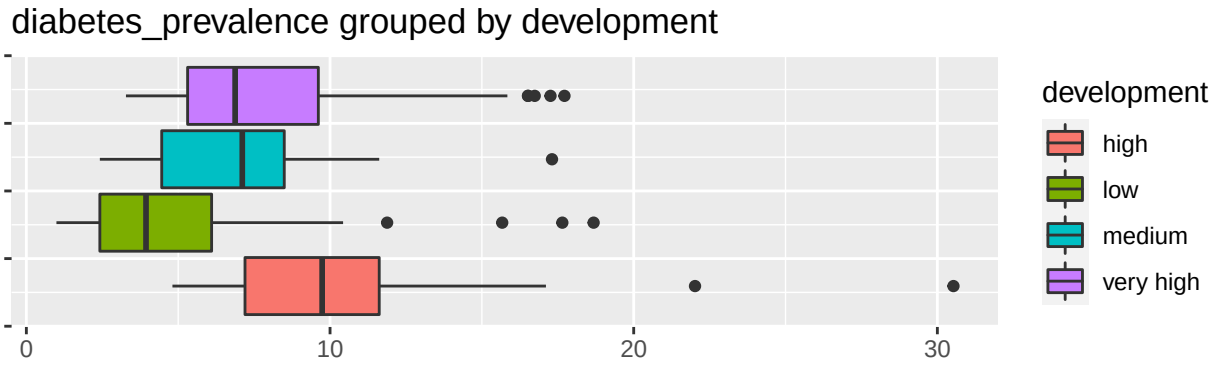
Looking at development we see the much higher concentration of low cardiovascular death rates for very high development countries. Which in general tend to have better healthcare. However while lower for low and medium development countries, we don't see too much of a difference between the two in terms of their distribution.

Boxplots for diabetes prevalence

```
plots(dataset=data, col='diabetes_prevalence',type='boxplot')
```







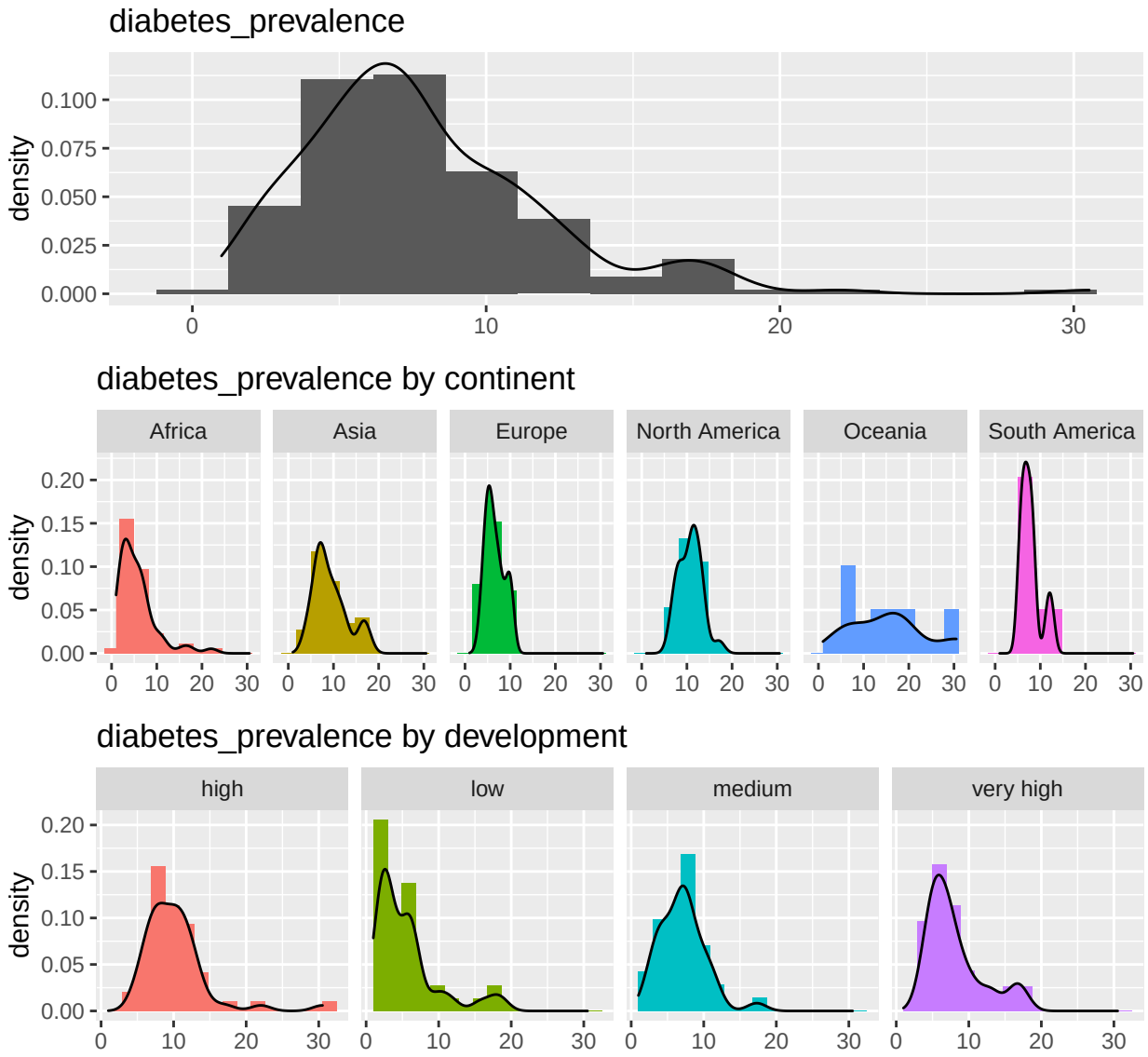
For diabetes prevalence we can see that most countries sit at a value of less than 10, but higher than 5. Some countries surpassing even 30%. These extreme values correspond to a few countries in Oceania and Africa.

The continent with the highest incidence as a proportion of its population seems to be Oceania. Which includes the top 3 countries with the highest amount of diabetics as a percentage of their population. With 30.53% for Marshall Islands. Although the values of the other top 2 countries are not included in our dataset, after some research, we found out that they're also 2 countries in Oceania. North America's diabetes incidence has nearly doubled in the past 20 years, therefore taking the spot 2 as the continent with the highest incidence with Asia, South America, Europe and Africa trailing behind.

We can, to an extent, see that higher development doesn't necessarily mean higher or lower diabetes prevalence and this might relate more to genetic composition and diet of the inhabitants.

Histogram and kernel density for diabetes prevalence

```
plots(dataset=data, col='diabetes_prevalence',type='hist',
      density=TRUE, bins=c(13,10,16), xtick_angles=c(0,0,0))
```



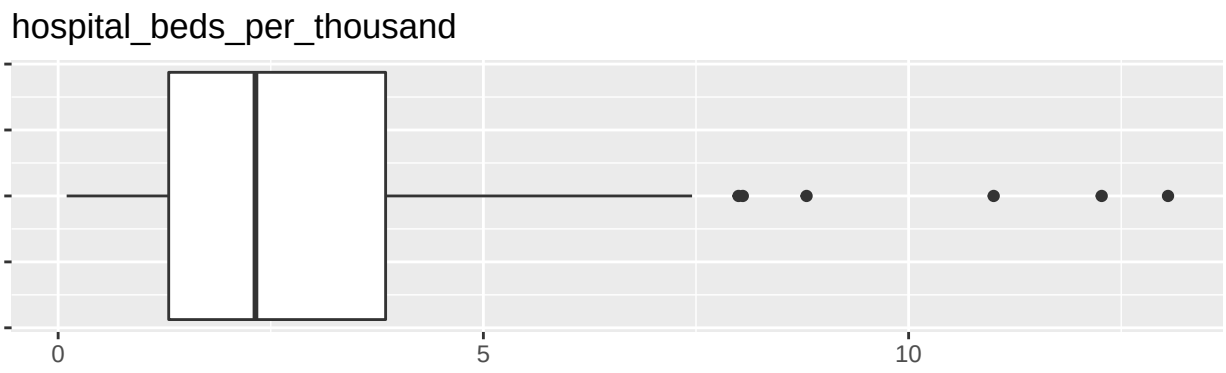
For the distribution of the data we see that it is resembles a normal distribution with a long left tail and the most countries clumped around the mean of ~7.9%.

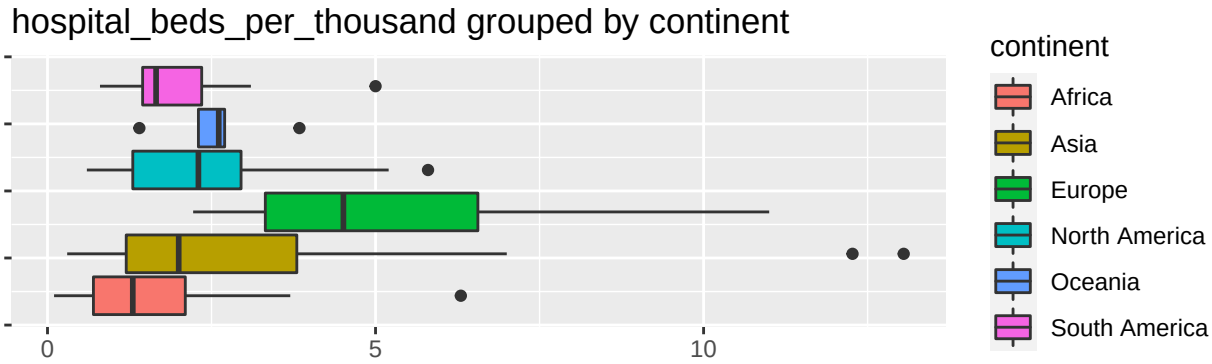
For each continent the incidence seems to be quite different, with some continents having a much higher incidence than others (for example Oceania vs Africa), however they all seem to clump around similar values.

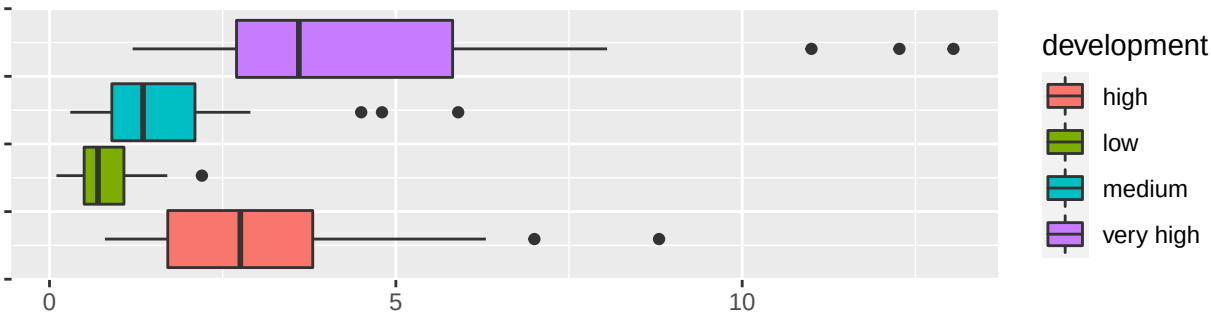
For the development we see the same we saw in the boxplots. Not much of a pattern or indication that there's any specific relationship between HDI and diabetes prevalence.

Boxplots for hospital beds per thousand inhabitants

```
plots(dataset=data, col='hospital_beds_per_thousand',type='boxplot')
```







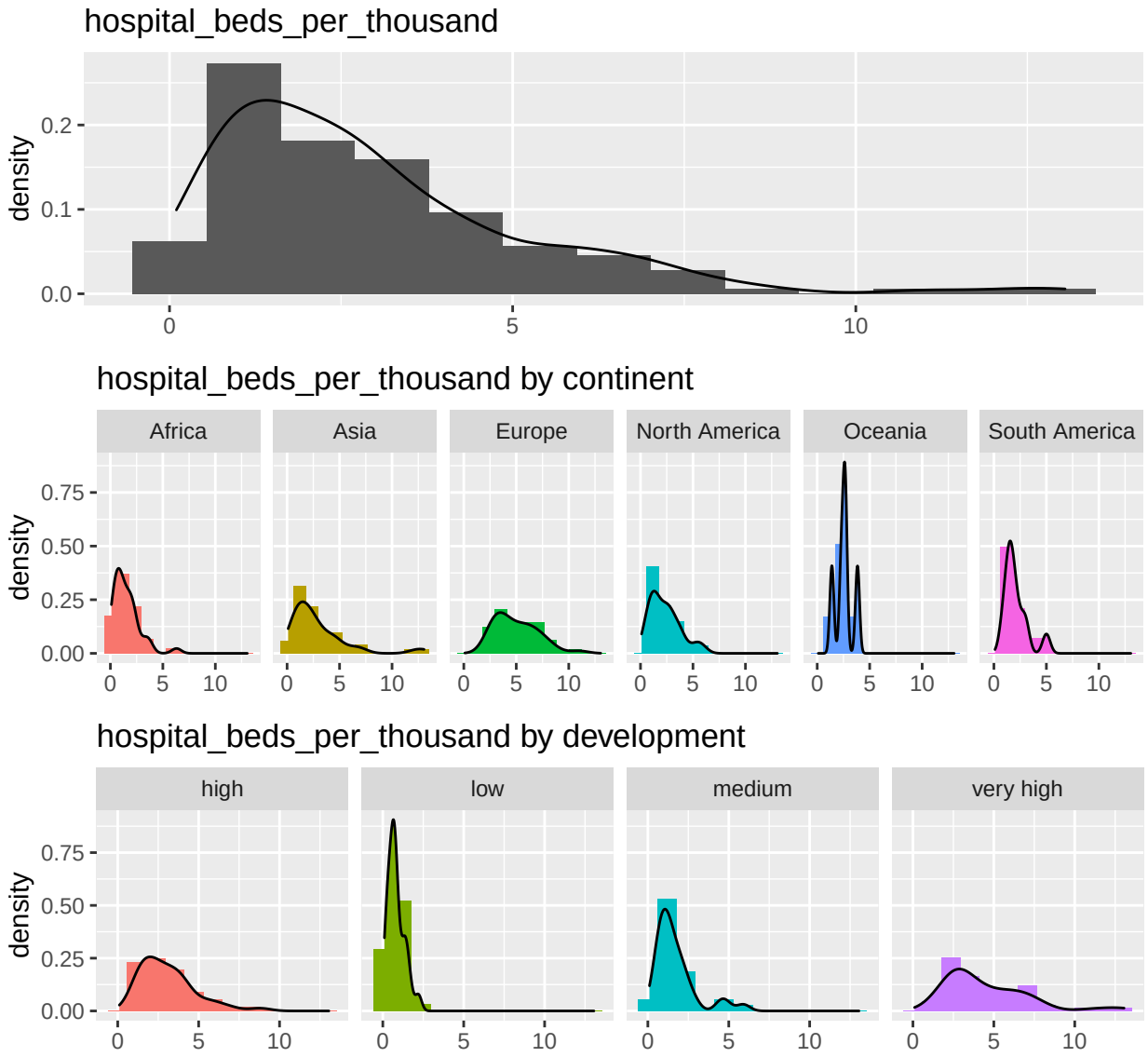
Looking at the hospital beds per thousand inhabitants variable boxplots we can see a few interesting things. We could use this variable as a measure of the quality of a healthcare system of a country. Where the higher the bed availability in hospitals is, the better the health system can cope with the demand for beds that a pandemic usually comes with. Especially with how widespread COVID-19 is.

We can see that some extremely underdeveloped countries have about 0.1-0.3 beds per thousand inhabitants, like it is the case with Mali and Niger. Some other countries like South Korea or Belarus have an extremely high capacity, with around 12 and 11 beds per thousand inhabitants respectively. However, even if the amount of beds per thousand inhabitants seems to be low, there's some countries with a suspicious seemingly low amount of beds, however, some of this are clearly just very highly populated countries.

For countries with high and very high HDI, there's a clear bias towards having greater bed capacity, however, this is not the case for all countries with that quality as there's clearly some countries with medium HDI that have a quite formidable bed capacity as well.

Histogram and kernel density for hospital beds per thousand inhabitants

```
plots(dataset=data, col='hospital_beds_per_thousand',type='hist',
      density=TRUE, bins=c(13,12,12), xtick_angles=c(0,0,0))
```



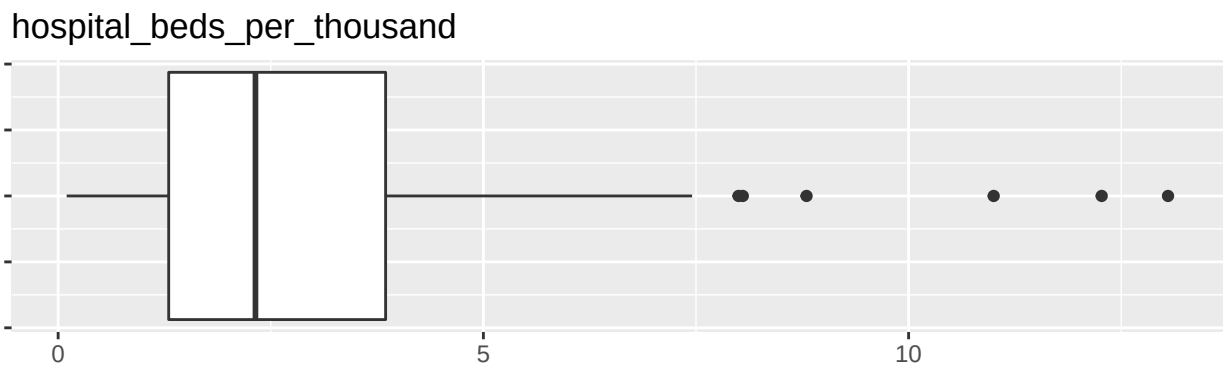
These plots tell a little bit of a different story to the boxplots. Where the largest concentration of countries is between 0 and 5 hospital beds per thousand inhabitants with an extremely scarce amount of countries with more than 10 beds per thousand inhabitants.

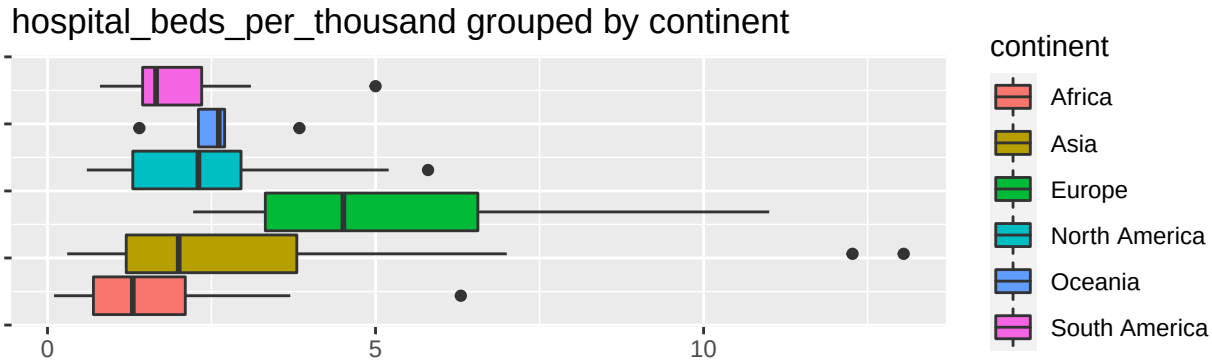
By segregating the data by continent we see that development does not necesarily mean greater healthcare capacity, with most continents boasting very similar numbers in this aspect whil esome like Asia, Africa, Europe and North america possessing some exceptions with extremely high numbers compred to the rest. However, yes, there's definitely a hint in continents with more developed countries (like Europe or some parts of Asia) which have a higher amount of beds, while Africa, which is predominantly composed of less developed countries tend to have a lower amount of beds.

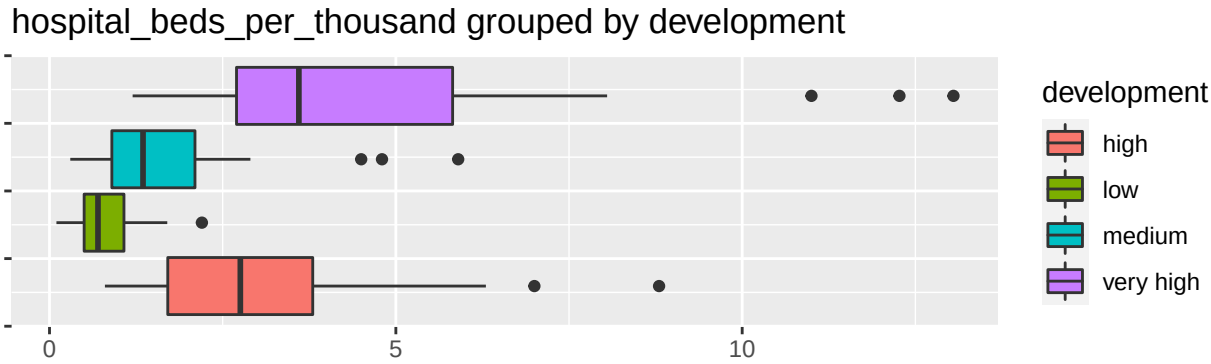
Finally, looking at development we see that it is rare for much less developed countries to have high bed capacity, while it is much easier for high to very high developed countries to have greater capacity. However, we can't confidently say that there's lots of exceptions to this 'rule'.

Boxplots for life expectancy

```
plots(dataset=data, col='hospital_beds_per_thousand',type='boxplot')
```







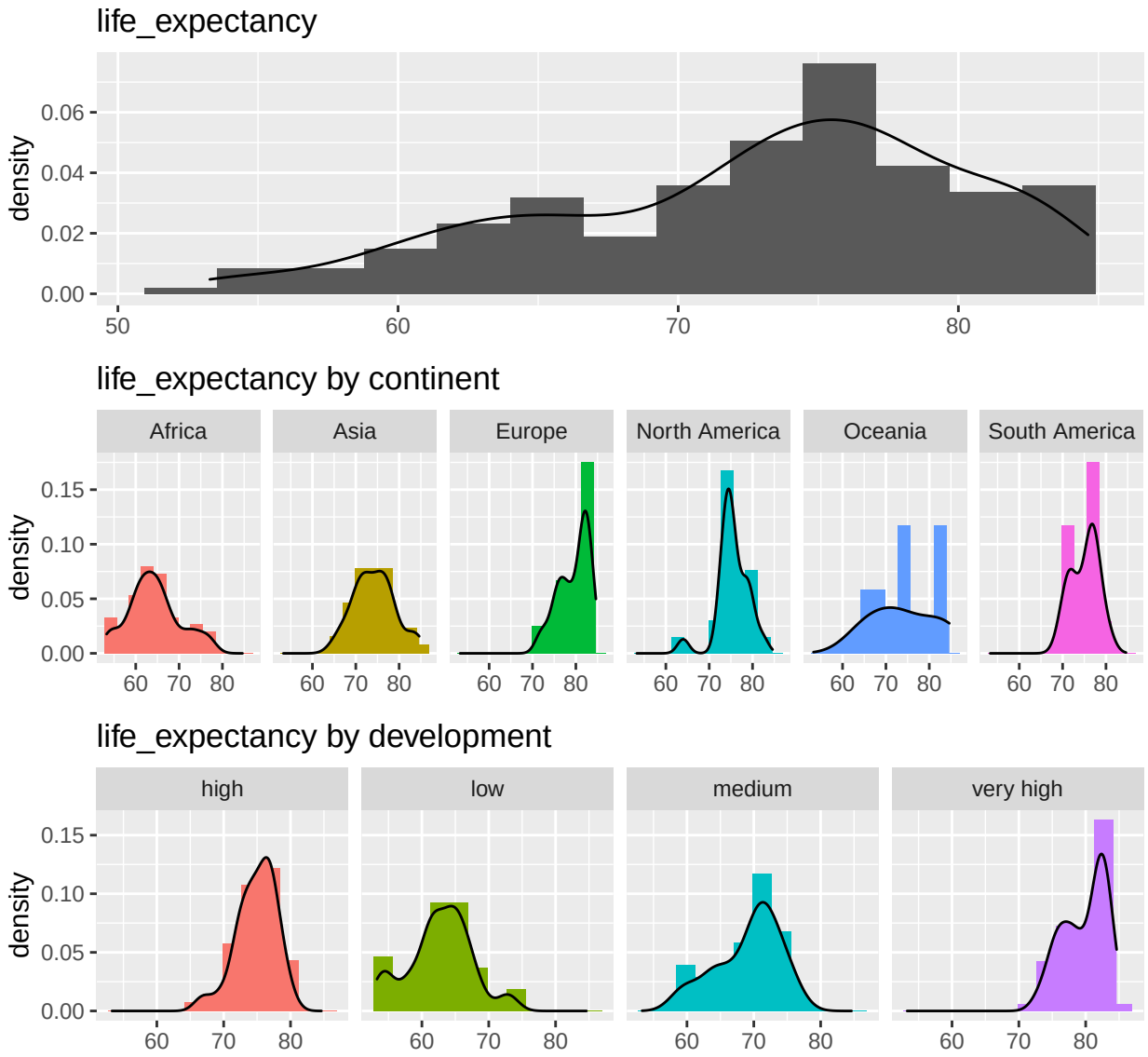
For life expectancy we can see moost countries sitting amove 66 years of age, with values going as low as 53.24 and as high as 84.63.

Africa has the lowest life expectancy while europe has the highest. The rest of the continents sit at roughly similar ranges.

Grouping by HDI, we can see that the most developed countries have a significantly higher life expectancy than those with low HDI. It clearly shows a strong positive correlation between them. Where the higher the life expectancy the higher the HDI. With very few exceptions.

Histogram and kernel density for life expectancy

```
plots(dataset=data, col='life_expectancy',type='hist',
      density=TRUE, bins=c(13,12,12), xtick_angles=c(0,0,0))
```



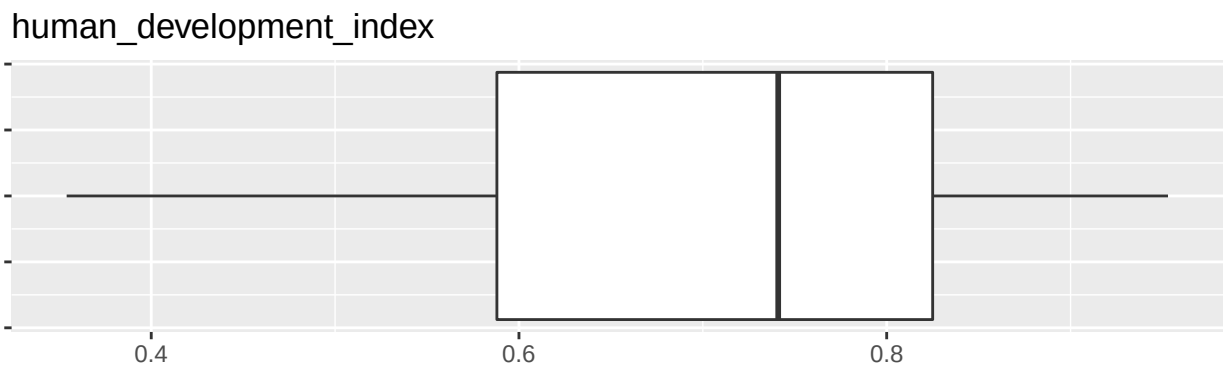
The general plot is somewhat left skewed, as most countries (about 80%) have a life expectancy higher than 65 years of age. Our density plot shows a strong concentration between 70 and 80 years of age, as this range covers the most nations.

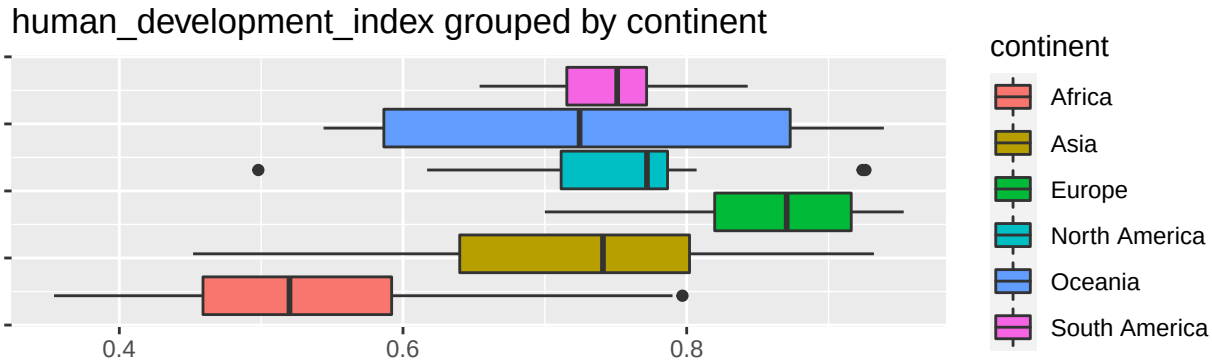
For each continent we see that europe shows a typically very high life expectancy while Africa shows a typically lower-than-average life expectancy for most countries with some exceptions. The rest of the continents sit at about the average life expectancy with some countries in Asia and North America at significantly higher-than-average numbers.

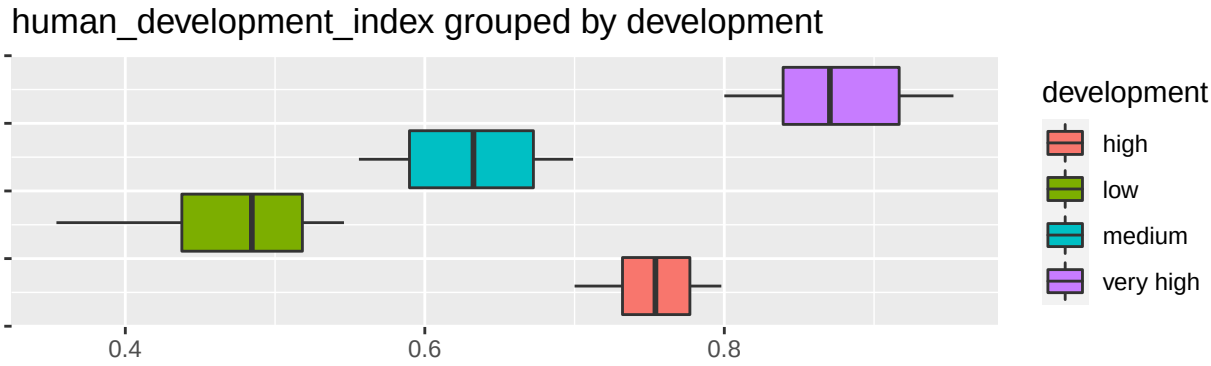
For HDI we can again see some of the strong correlation, where life expectancy for very highly developed nations seems to be also quite high and the same happens with less developed nations.

Boxplots for Human Development Index

```
plots(dataset=data, col='human_development_index',type='boxplot')
```







We can see most countries fall between 0.6 and 0.8, o ur median HDI is 0.741.

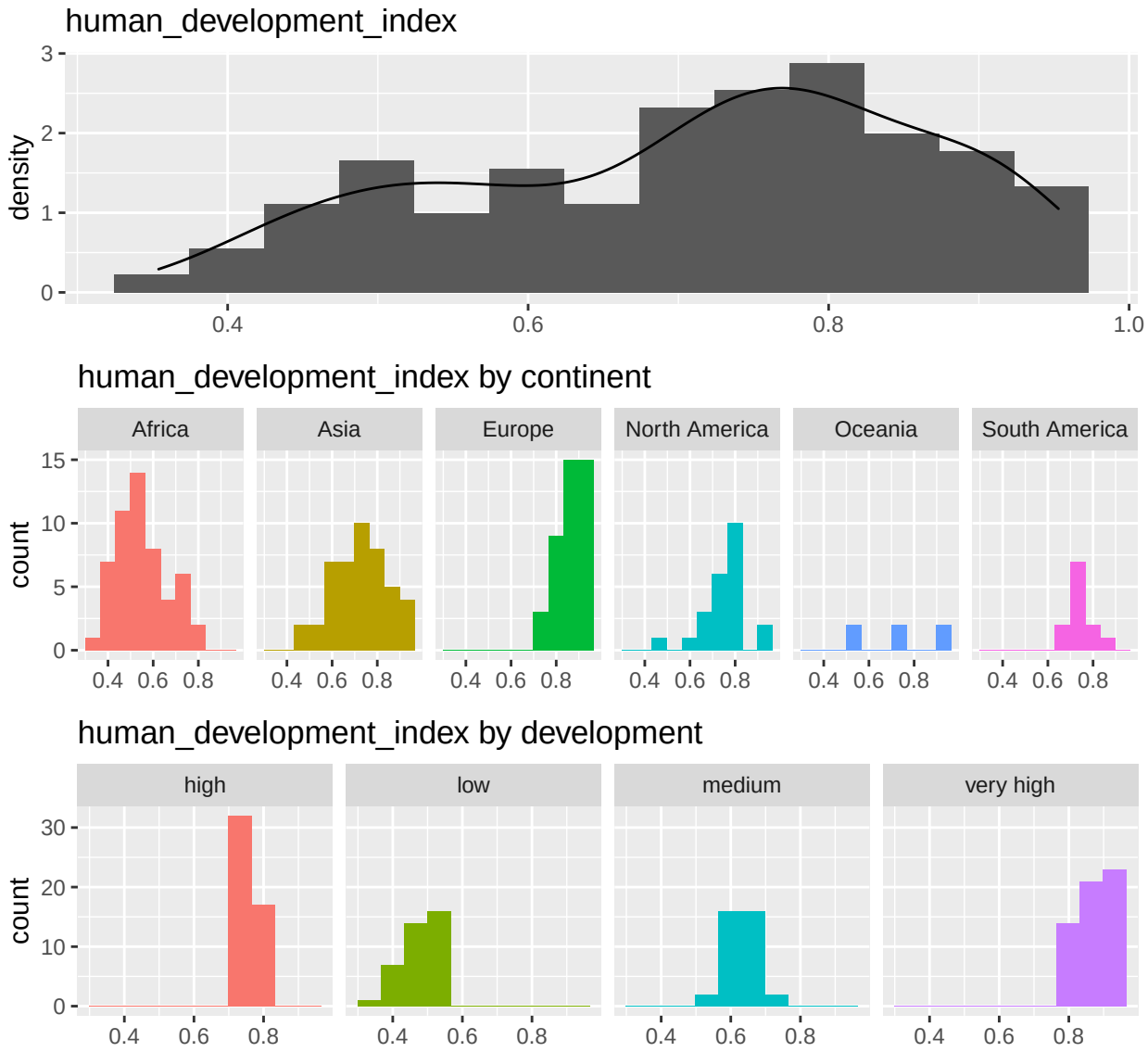
For continents we can see Africa lagging behind with most of its countries between 0.4 and 0.6 HDI, probably given the porverty situation in the continent.

The rest of the continents sit between 0.6 and 0.8 for most of its countries with North America having 2 very extreme outliers which are its minimum and maximum values (corresponding respectively to Haiti and USA). Europe is generally above 0.8.

As our development variable was constructed from the human_development_index variable, we can see that there's clearly marked bounds for each HDI range. The ranges are as follows: *very high* for HDI of 0.800 and above, *high* from 0.700 to 0.799, *medium* from 0.550 to 0.699 and *low* below 0.550.

Histogram and kernel density for Human Development Index

```
plots(dataset=data, col='human_development_index',type='hist',
      density=FALSE, bins=c(13,10,10), xtick_angles=c(0,0,0))
```



For the human development index we can see that the variable is somewhat left skewed, given that the average HDI is ~0.71, which most countries either match or are above of.

For the HDI per continent we can see that africa has a clear concentration below 0.6, given that most countries in Africa have a low HDI. South america and Asia tell a similar story, most countries are at or above 0.6. We can see that for North America there's a little concentration below 0.6 and most countries betweeen 0.6 and 0.8 as North America includes Central America and the Caribbean which tend to have a lower HDI than USA/Canada, which are towards the right of 0.8. Most european countries have a very high to high HDI, therefore the density plot is quite left skewed and most contries in Oceania have a lower-than-average HDI with the exception of New Zealand and Australia which are above 0.8.

Correlation and scatter plots

We define a function to set colors for categorical variables in a PCP plot:

```
colors <- function(cat_var, colors_vector) {
    kleuren <- as.numeric(as.factor(cat_var))
    foreach (i=1:length(kleuren), kleur=kleuren) %do% {
        kleuren[i] = colors_vector[kleur]
    }
    return(kleuren)
}
```

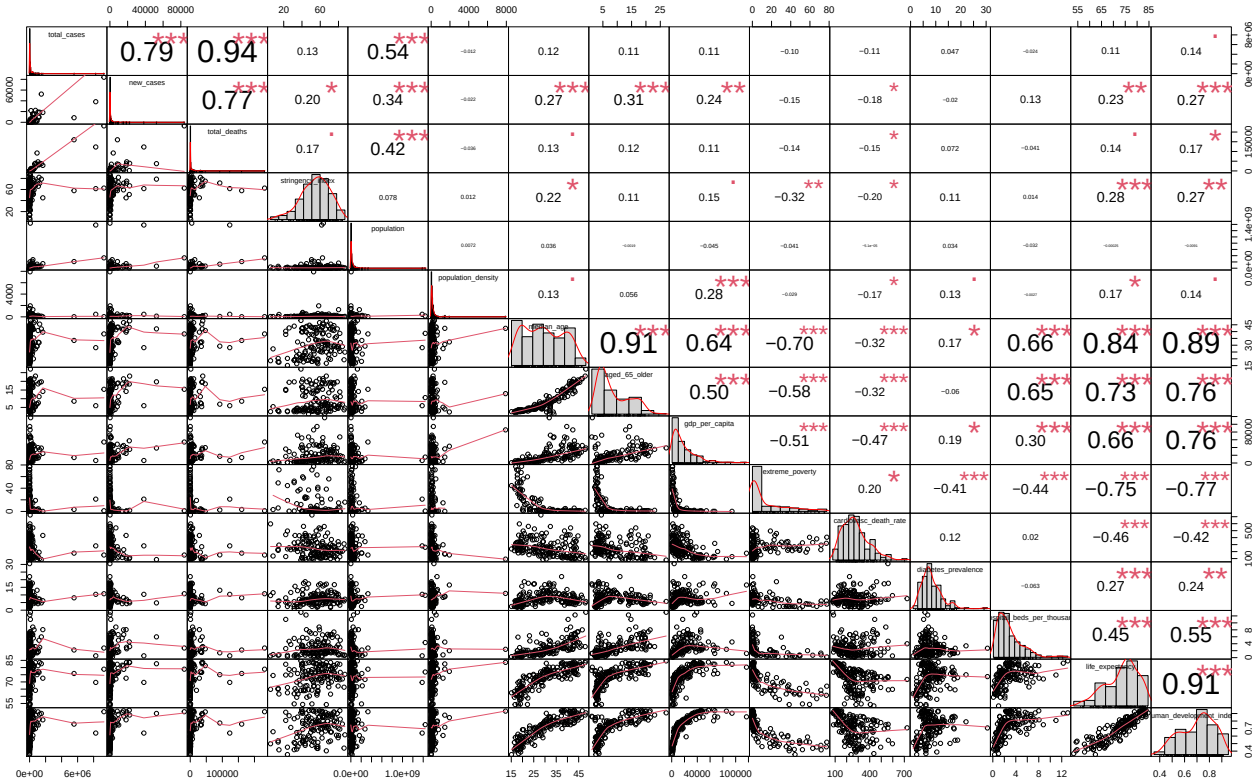
Colours we picked:

```
# setting colors development
color_1 <- "blueviolet"
color_2 <- "brown"
color_3 <- "seagreen"
color_4 <- "yellow3"
color_5 <- "black"
color_6 <- "deeppink1"
palette1 <- c(color_1,color_2,color_3,color_4)
palette2 <- c(color_1,color_2,color_3,color_4,color_5,color_6)
```

```
development_colors <- colors(data$development,palette1)
continent_colors <- colors(data$continent,palette2)
```

Correlation matrix + scatter plot

```
pa <- data_n %>% dplyr::select(interesting_vars)
chart.Correlation(pa, histogram=TRUE, pch=19, method="pearson")
```

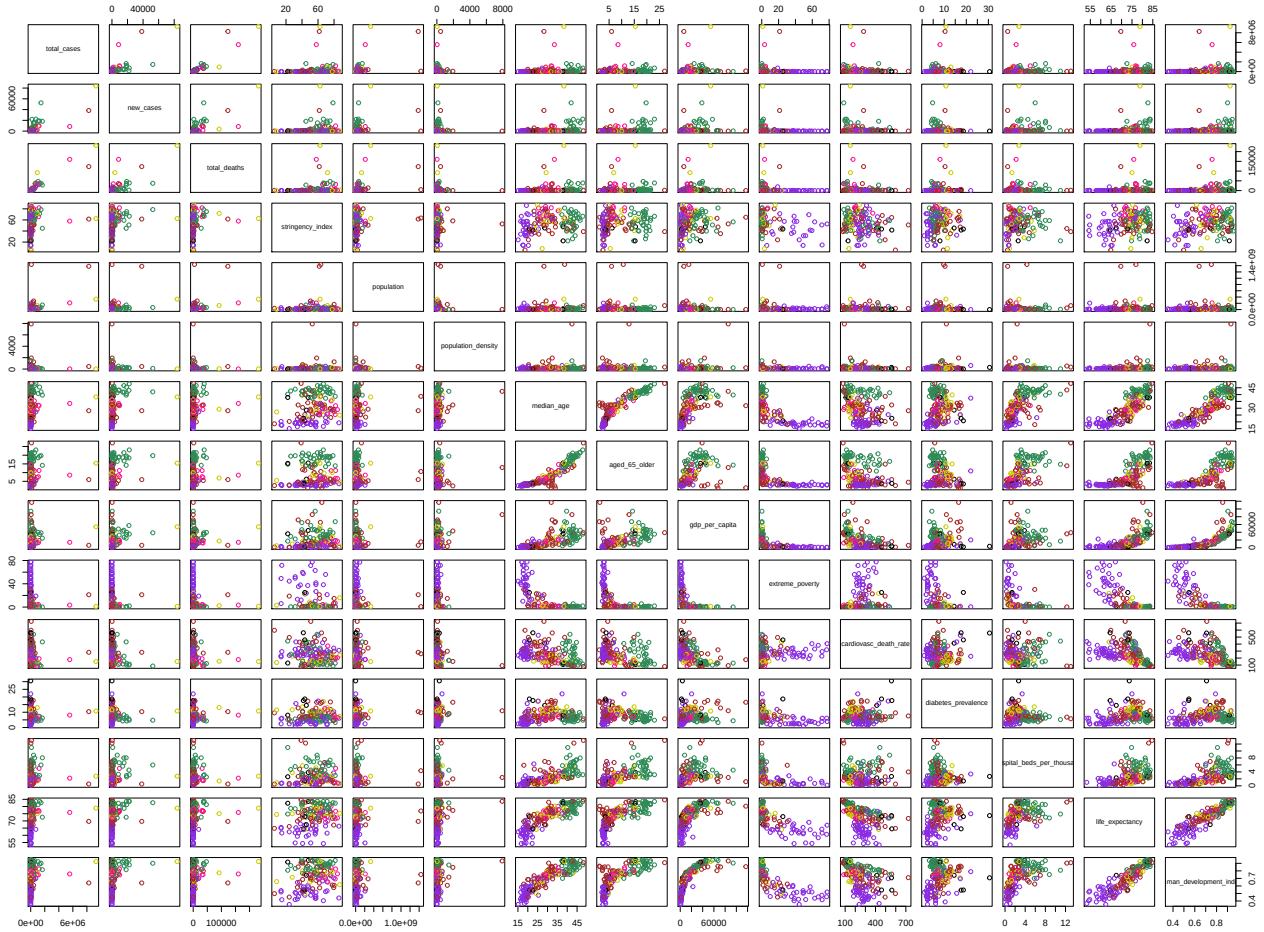


Looking at the variables which are interestingly correlated we have the following:

- *Very high correlation*:
  - **0.94**: Total deaths and total cases, not much to say here, it obviously makes sense that, during a pandemic of a deadly disease, we see correlation between total cases and total deaths.
  - **0.91**: median age and proportion of population aged 65 or older are both measures of age within a population, therefore makes sense to see them being highly correlated.
  - **0.91**: life expectancy and human development index, as life expectancy is a component of HDI it makes sense that these are correlated.
  - **0.89**: HDI and median age, we can see that clearly, highly developed countries tend to have an aging population, along with lower fertility rates.
  - **0.84**: life expectancy and median age. The higher the life expectancy of a population, the more developed the country is, then, the older its population is. This can definitely be framed from a different perspective though.
- *High to medium correlation*:
  - **0.79**: total cases and new cases, we can see that the correlation is high, however, not as much as we might expect, which tells us that outbreaks happening in certain countries differ from its total case count significantly, therefore we can have countries with massive total counts but very small daily counts for november 11th 2020.
  - **0.77**: new cases and total deaths show a similar trend to what previously described with total cases and new cases. We see that an active outbreak and death count of a country can differ significantly, however, they are usually somewhat correlated.
  - **-0.77**: extreme poverty and HDI, this is clear, as the more developed a country is, the least likely it is to have a significant portion of its population living under the poverty line.
  - **0.76**: HDI vs aged_65_older and HDI vs gdp_per_capita, this relates to median age vs HDI as well as to the logical idea that the higher the income of individuals in a population is, the higher the development of the nation they reside in tends to be.
  - **-0.75**: extreme poverty and life expectancy are clearly related, as the more developed a nation is, the least people under the poverty line it tends to have, and along with this, the healthier the population will be. People living under the poverty line can very commonly be malnourished or take less care of their health, therefore effectively reducing their life expectancy.
  - **0.73**: proportion of population aged 65 or older and life expectancy are correlated as well for obvious reasons, similar to median age vs life expectancy.
- *Other correlations worth mentioning*:
  - **0.66**: life expectancy and gdp per capita, which at first glance would suggest more correlation than simply 0.66, but this is enough to confidently state that there's definitely a strong relationship between life expectancy and gdp per capita, and definitely, the more a person earns, the longer they tend to live.
  - **0.66 and 0.65**: hospital beds per thousand inhabitands vs median age and vs aged_65_older. This suggests that the more healthcare availability there is, the generally older the population tends to be. Usually we would mention life expectancy which still boasts a somewhat significant correlation with hospital beds per thousand inhabitants (**0.45**), however, while a higher life expectancy tends to also mean a higher median age, or the other way around, we might want to mention that not necessarily because a country has a high life expectancy, it will have great healthcare availability.
  - **0.54**: population and total cases, one could assume that countries with a very high population would have more trouble managing a pandemic, as they tend to be larger and have more cities and larger cities, therefore making movement within the country a hugely difficult logistic problem. However, countries like China or Indonesia, with their massive populations have been able to manage the pandemic just fine. However, others like India, USA or European countries like Spain, France and Germany, have had a much harder time managing the pandemic.

Scatter plot matrix for interesting variables, grouped by continent

```
pairs(pa,pch=1,col=continent_colors)
```



Color coding:

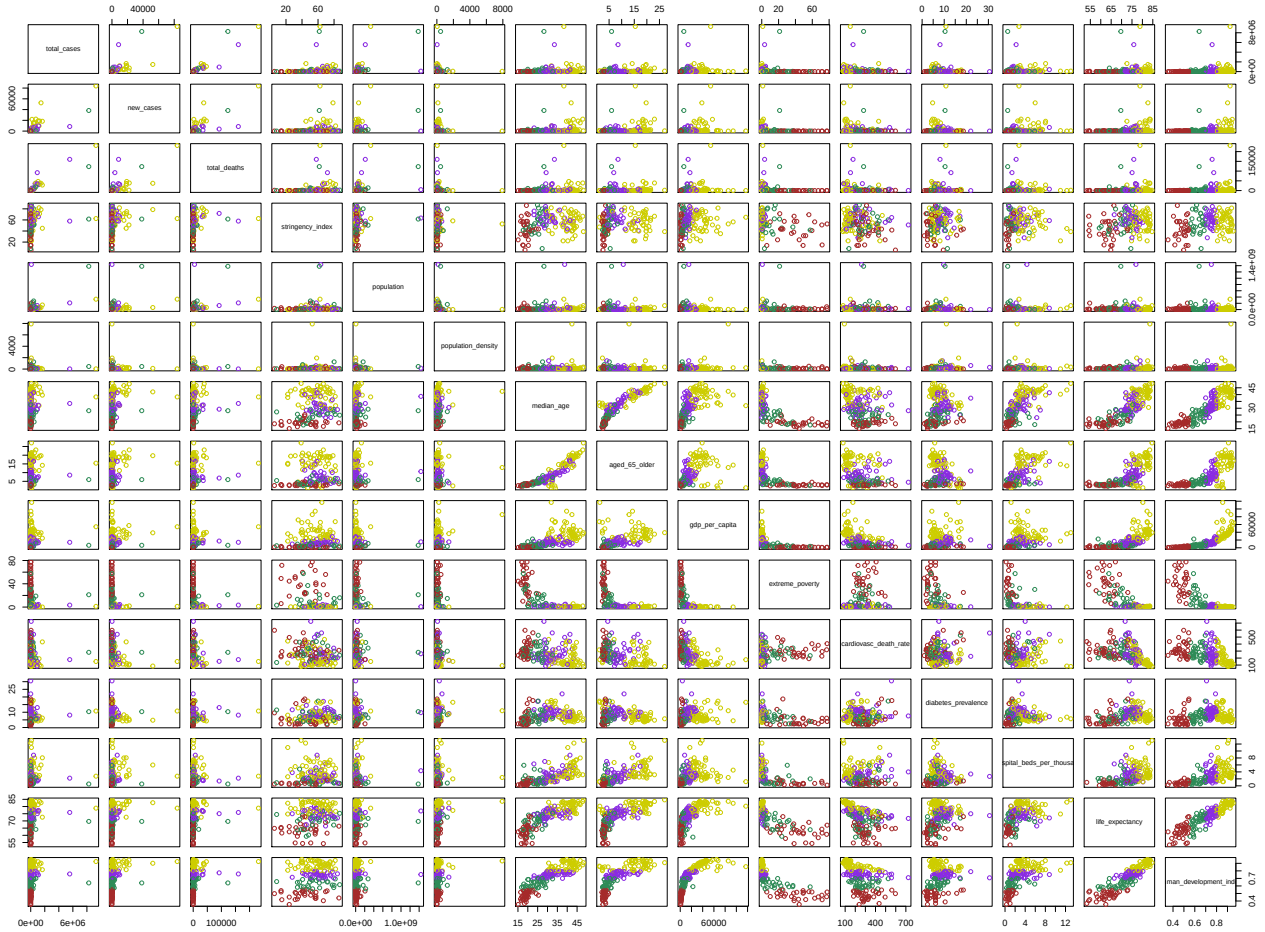
```
unique(cbind(data$continent, continent_colors))
#>                    continent_colors
#> [1,] "Asia"           "brown"
#> [2,] "Africa"         "blueviolet"
#> [3,] "Europe"         "seagreen"
#> [4,] "South America"  "deeppink1"
#> [5,] "North America"  "yellow3"
#> [6,] "Oceania"        "black"
```

Let's make a few observations from this:

- Usually Europe is at the top of high quality of life metrics like median age, HDI. gdp per capita and such. Europe, however, tends to be at the top (along with some other countries) of the cases and deaths charts.
- Usually Africa is at the bottom of high quality of life metrics. The rest of the continents are around the middle. However, it also tends to be at the bottom of case and death counts and stringency index.
- There's clearly defined groups in measures typically used to measure quality of life, things like HDI, life expectancy and similar plots against other variables of the same type, show a clear grouping distinction between continents.

Scatter plot matrix for interesting variables, grouped by development

```
pairs(pa,pch=1,col=development_colors)
```



Color coding:

```
unique(cbind(data$development, development_colors))
#>                  development_colors
#> [1,] "low"       "brown"
#> [2,] "medium"    "seagreen"
#> [3,] "high"      "blueviolet"
#> [4,] "very high" "yellow3"
```

Ignoring all correlations using human_development_index as it composes the development variable we can observe a few things:

- Even more so than with continents, quality of life measures are much more clearly defined. We see that countries with the highest development also top the cases correlations with basically everything else. They tend to be at the rightest, and and the top of such plots.

PCP Plot

We group variables by their skewness, while we have many right skewed variables, we group the rest of them in another PCP plot, to have a less crowded plot.

```
right_skewed <- c('extreme_poverty','gdp_per_capita',
                  'aged_65_older','population_density',
                  'total_cases','new_cases','population')
right_skewed <- data_n %>% dplyr::select(right_skewed)

others <- c('human_development_index','stringency_index',
            'life_expectancy','median_age','diabetes_prevalence',
            'cardiovasc_death_rate','hospital_beds_per_thousand')
others <- data_n %>% dplyr::select(others)
```

Right skewed variables PCP

Grouped by continent

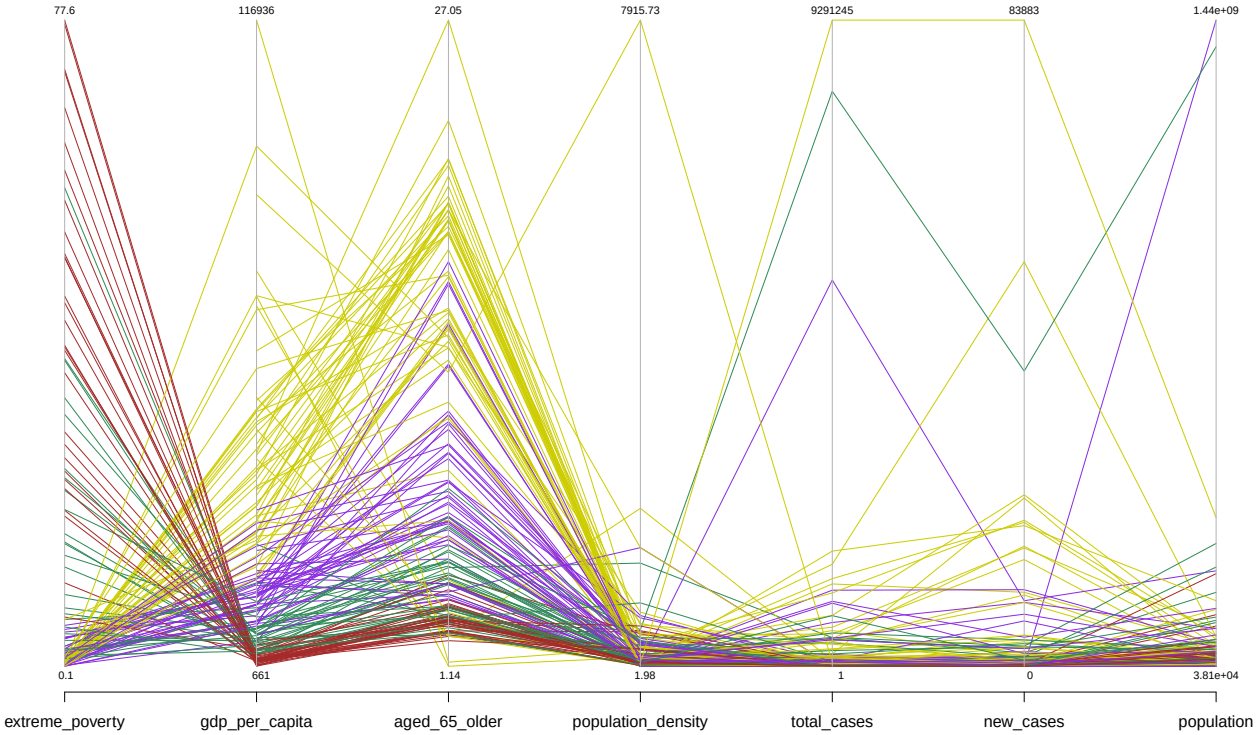
```
parcoord(right_skewed,var.label=TRUE, col=continent_colors)
```

For the PCP plot of our (mostly) right skewed variables we have the following observations:

- We can see the group differences clearly for the extreme poverty variable, where the top observations are mostly from Africa
- For gdp_per_capita it's clear that the trend is shifted entirely, whereas Africa was at the top of the extreme poverty observations, it is now at the bottom of the gdp_per_capita observations, along with Europe and a few countries from other countinents like Asia, North America and Oceania climbing to the top.
- aged_65_older shows a similar trend as gdp_per_capita which we confirmed in the correlation plot as somewhat correlated variables, along with life expectancy.
- Population density gets mashed down as we have a few extreme outlier countries with incredibly high population densities. The top observation being Singapore, with a staggering population density of 7915.73 people per square kilometer.
- Similar with population density, total cases is opaqued by the extreme total case numbers of USA, India and Brazil.
- Population is topped by China and India, which include represent over 1/4th of the world population.

Grouped by development

```
parcoord(right_skewed,var.label=TRUE, col=development_colors)
```



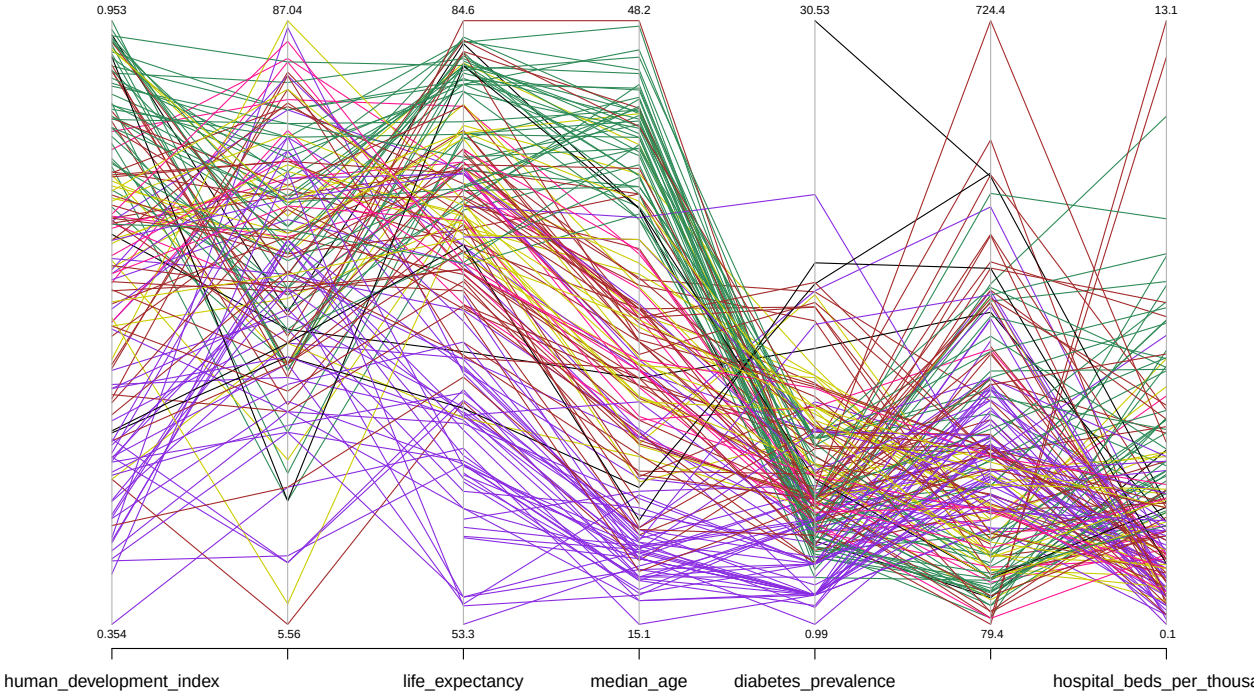
When segregating by development:

- Our extreme poverty countries tend to all be countries with low or medium HDI, however, low HDI countries represent the highest spots.
- gdp per capita and aged 65 or older are variables which were previously mentioned as being strongly correlated to HDI, therefore we can see how strong the correlation is here and how starkly it divides groups, where it is almost certain that the top positions of both these variables are taken by very high or high development countries.
- for total cases and new cases it is interesting to see that the seemingly top positions are all covered by the highest development countries. We could attribute this to more widespread testing, but also to much more widespread movements of people. People in low HDI countries tend to move much less and live much simpler lives usually as most inhabitants live in rural areas.

Other variables PCP

Grouped by continent

```
parcoord(others,var.label=TRUE, col=continent_colors)
```



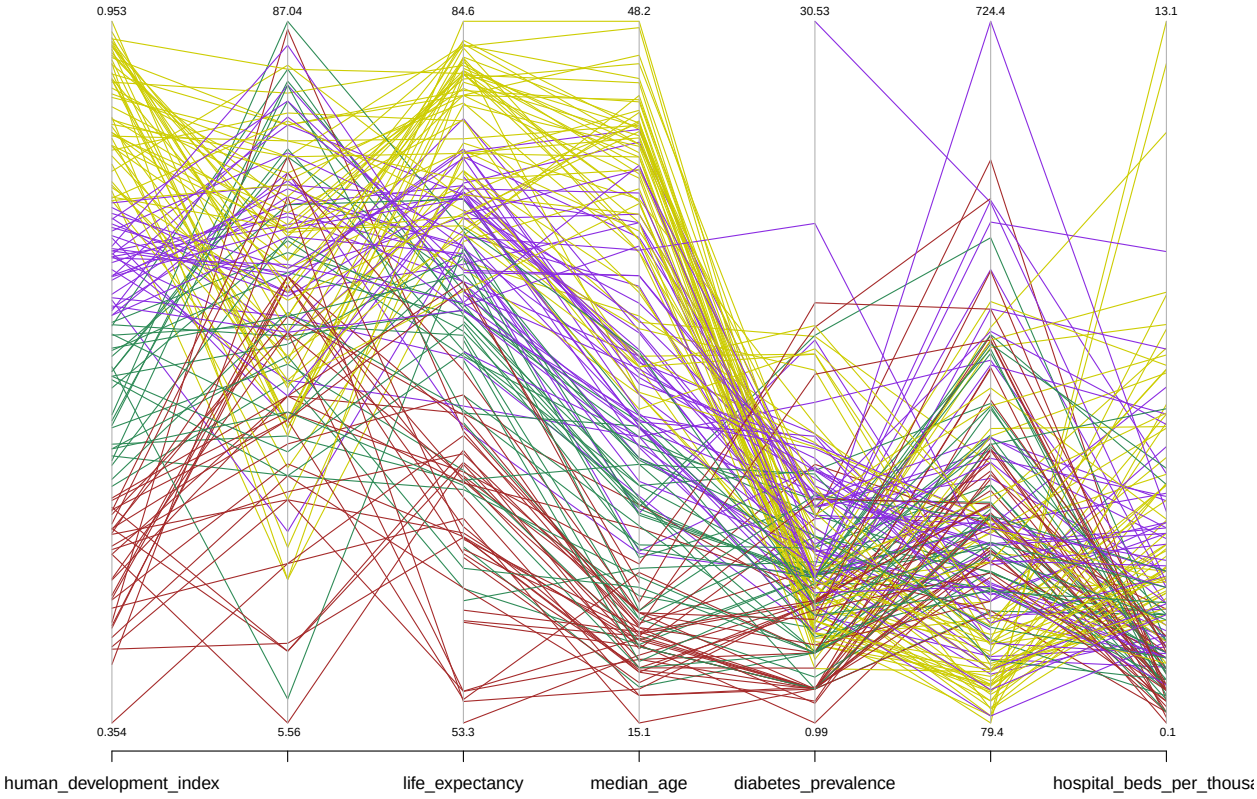
When looking at the rest of the variables, which were separated from the right skewed ones to allow for beter visibility of group segregation, we see some interesting things.

Grouping by continent we can point out a few things:

- Human development index is clearly significantly higher for Europe and some countries in North America, Asia and Oceania than the rest of the countries, while African countries usually take the bottom spots with some quite underdeveloped countries in Oceania, Asia and the Americas.
- Stringency index seems like a huge mashup, where the only discernible information we can take is that there's seemingly no way to tell which continents tend to have the strictest rules. Every country seems to do its own thing when it comes to restrictions.
- life expectancy tells much of the same story we saw on the correlation plot and the histograms, clearly, the lower the quality of life in a country, the lower its life expectancy will be. And as most quality of life metrics disfavour the least developed countries, life expectancy is not an exception here. We unfortunately also see Africa taking the lowest positions and very high development countries taking the highest, with Japan taking the top spot closely followed by Some European and Asian countries with very high HDI. Median age also follows this trend.
- For diabetes prevalence there's a bit of a mash, where genetic background and diet seem to probably matter more than the location of individuals. Prevalence seems to be the highest in Oceania and the lowest in Africa and Europe.
- No discernible group difference between continents for cardiovasc_death_rate
- For hospital_beds_per_thousand we can see that there's a tendency for countries in Africa to have significantly lower bed availability, and while more developed countries tend to have higher bed availability, there are definitely plenty of exceptions to this rule.

Grouped by development

```
parcoord(others,var.label=TRUE, col=development_colors)
```



When grouping by our HDI categories, we see the following:

- Groups are clearly divided in most categories with the exception of diabetes prevalence
- We see much of the same we had mentioned earlier related to stringency index, there's no clear grouping among developed or developing nations. All that's visible is that developed nations might be somewhat less flexible than average with restrictions, but there's plenty of exceptions to this hypothesis.
- For life expectancy and median age, as mentioned previously, there's a clear division among the most and the least developed nations, as development is quite related to these two variables.
- Hospital beds per thousand inhabitants shows much of the same, there's plenty of developed countries with lacking bed availability for new patients. However, there's clearly some of a tendency for highly developed nations to have better availability than nations with a lower HDI.

Principal Component Analysis (PCA)

Taking a subset of the dataset:

```
covid = data[c('continent','location','development','total_cases','new_cases','total_deaths','stringency
covid$continent=factor(covid$continent)
covid$development=factor(covid$development)
```

Imputation of missing values:

```
covid_imp=mice(covid,m=5,method = "cart")
covid_imp=complete(covid_imp)
```

```
covid = covid_imp
covid_quans=covid[,4:18] # quantitative variables
```

PCA segregating by development (measure of HDI)

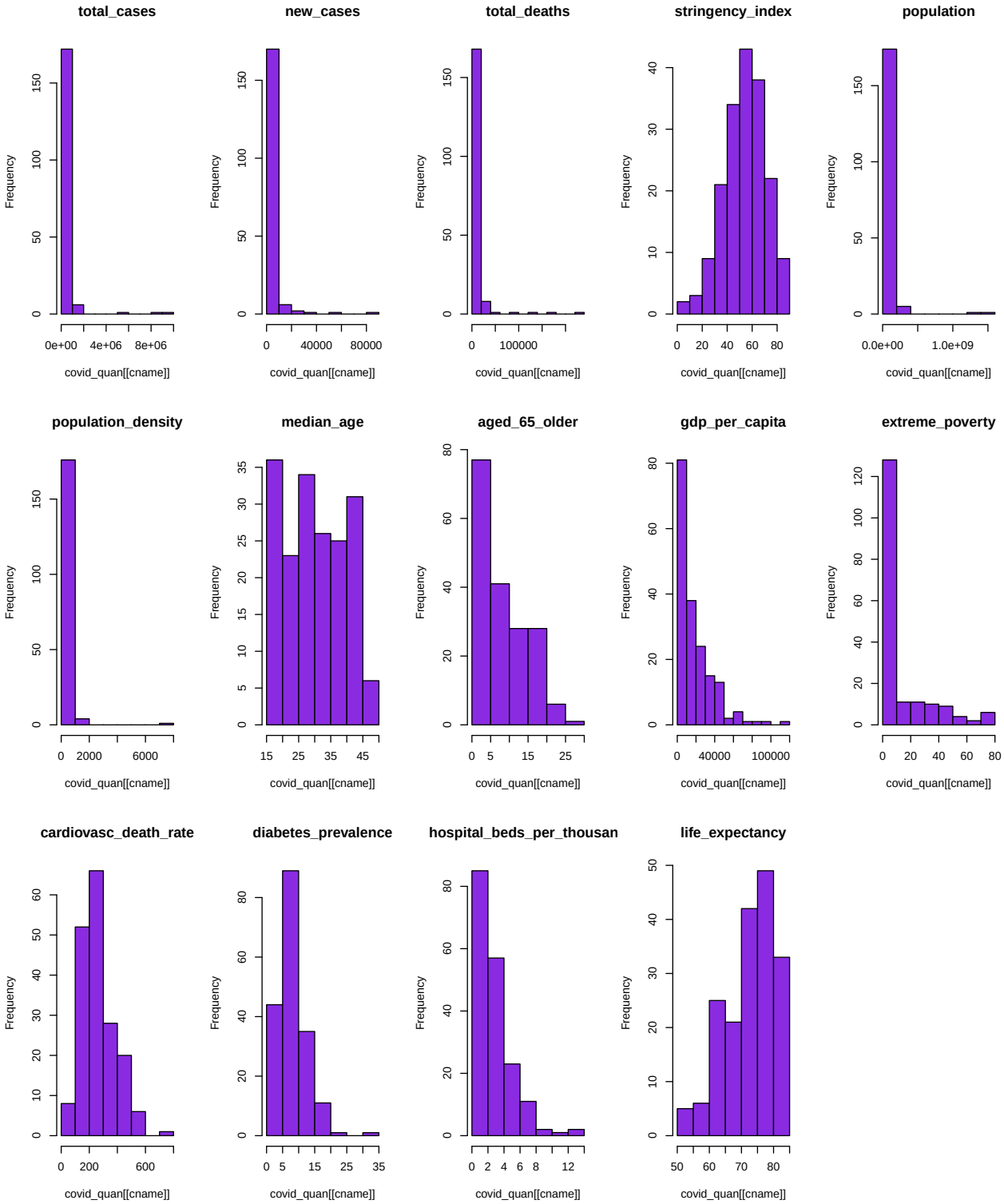
Avoiding adding *human_development_index* as it constructs our target variable *development*.

```
covid_quan <- covid_quans[,-15]
```

```
n <- nrow(covid_quan)
p <- ncol(covid_quan)
```

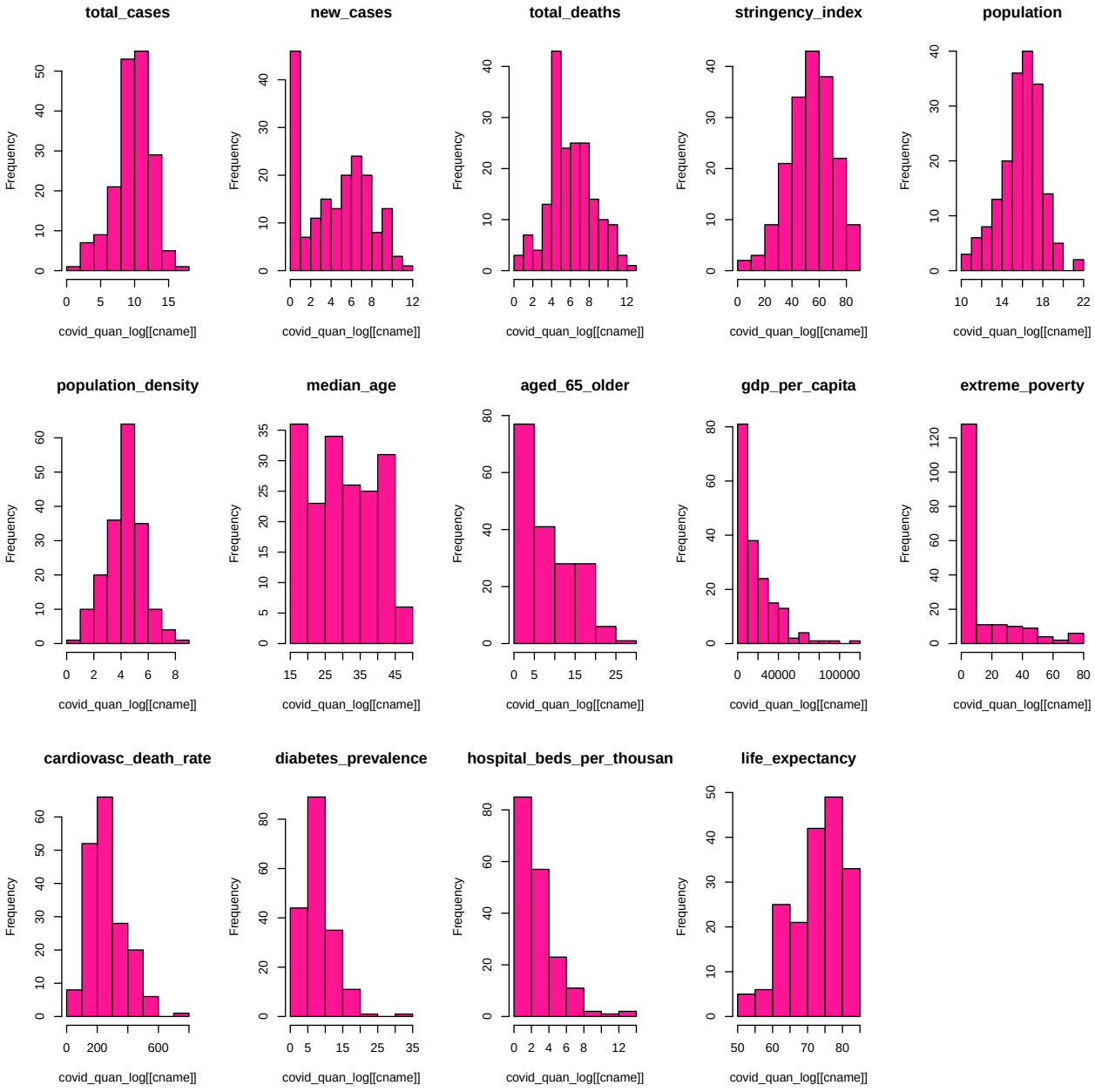
Checking the distribution of each variable using histogram

```
par(mfrow=c(3,5))
sapply(names(covid_quan),function(cname){hist(covid_quan[[cname]],main=cname,col = color_1)})
```



Transforming some of the variables using log because some of the variables are highly positively skewed

```
covid_quan_log=covid_quan
for (i in 1:6) {
    if (i != 4) {
        covid_quan_log[,i]=log(covid_quan[,i])
    }
}
covid_quan_log$new_cases[which(covid_quan_log$new_cases== -Inf )] = 0
par(mfrow=c(3,5))
sapply(names(covid_quan_log),function(cname){hist(covid_quan_log[[cname]],main=cname,col = color_6)})
```



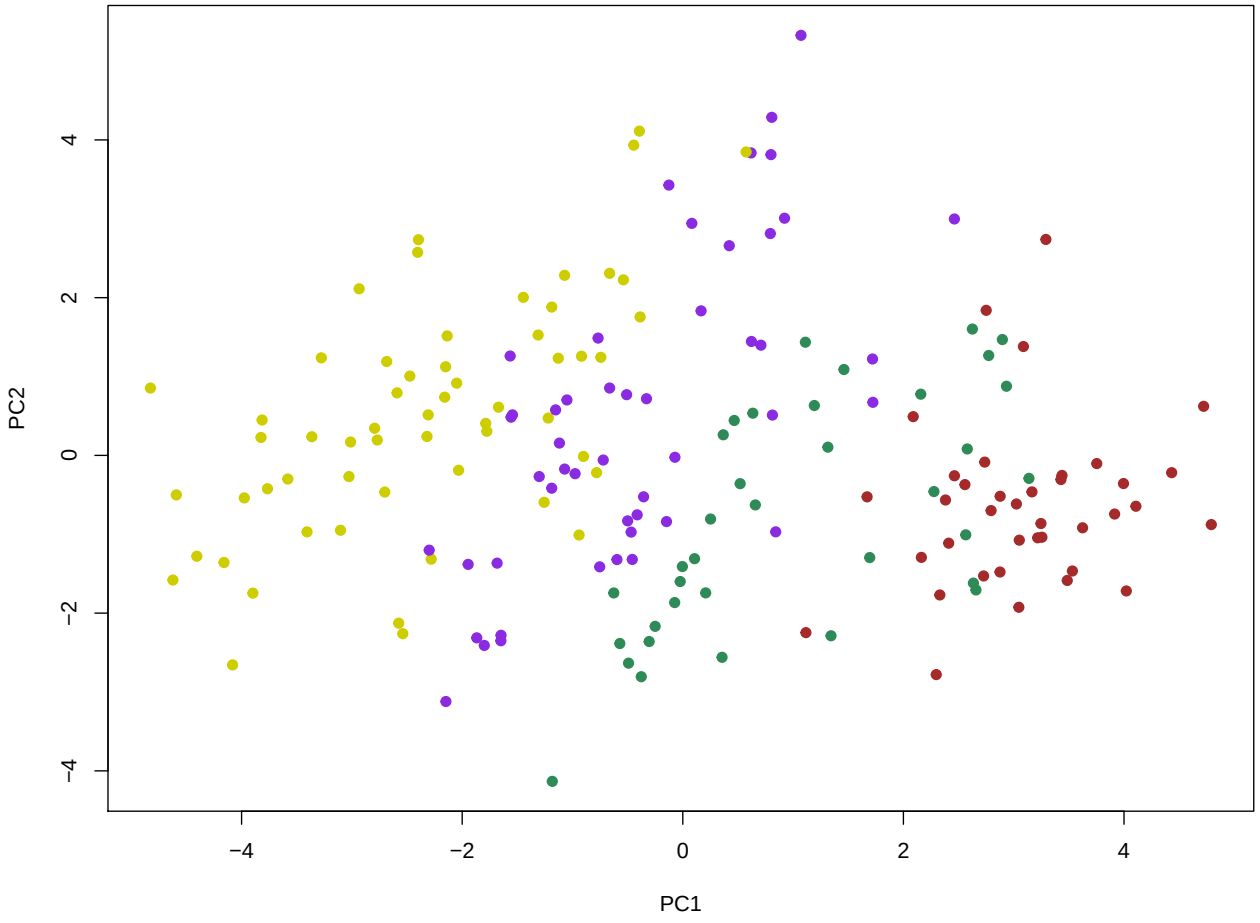
PC Scores:

```
X_pcs <- prcomp(covid_quan_log, scale=TRUE)
```

Plotting the first two PCs grouped by development

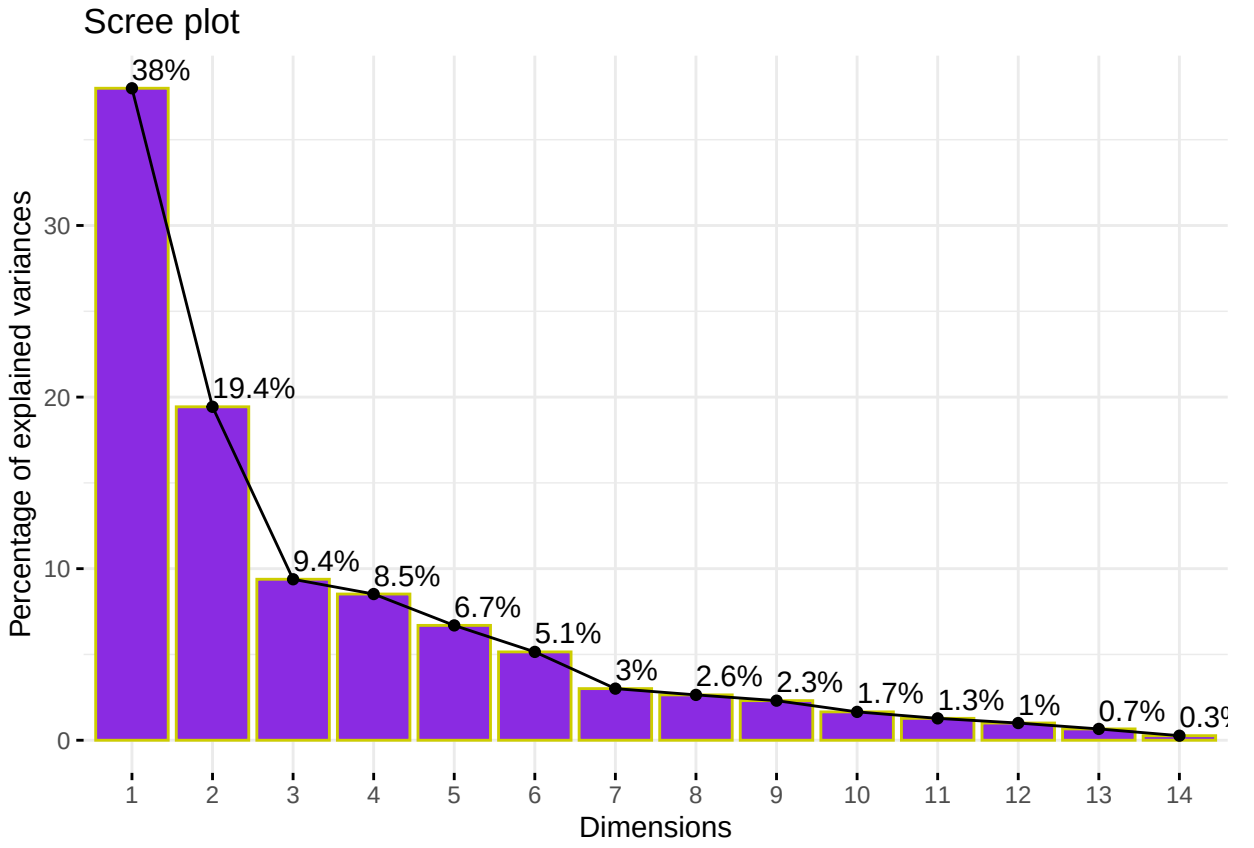
We can see here the principal components differentiate perfectly the different levels for 4 levels of HDI

```
par(mfrow=c(1,1))
plot(X_pcs$x[,1:2],pch=19,col=development_colors)
```



We can check the variance explained by each principal component, e.g. the first pincipal component explains 38% of the total variability and the second one explains 19.4%.

```
fviz_eig(X_pcs,ncp=17,addlabels=T,barfill=color_1,barcolor=color_4)
```



Checking eigenvalues and explained variance

Now we check the eigenvalues and the cumulative percentage of explained variance, and we have decided to take the first 5 principal components.

We will be using the 33% of the variables and will be keeping the 81.99% of the total information.

```
get_eigenvalue(X_pcs)
#>        eigenvalue variance.percent cumulative.variance.percent
#> Dim.1  5.31960453       37.9971752                   37.99718
#> Dim.2  2.72072808       19.4337720                   57.43095
#> Dim.3  1.31349884        9.3821346                   66.81308
#> Dim.4  1.19305544        8.5218246                   75.33491
#> Dim.5  0.93705417        6.6932441                   82.02815
#> Dim.6  0.72094710        5.1496221                   87.17777
#> Dim.7  0.42131302        3.0093787                   90.18715
#> Dim.8  0.37015731        2.6439808                   92.83113
#> Dim.9  0.32313440        2.3081029                   95.13924
#> Dim.10 0.23153384        1.6538132                   96.79305
#> Dim.11 0.17872329        1.2765949                   98.06964
#> Dim.12 0.14051815        1.0037011                   99.07334
#> Dim.13 0.09241939        0.6601385                   99.73348
#> Dim.14 0.03731242        0.2665173                  100.00000
```

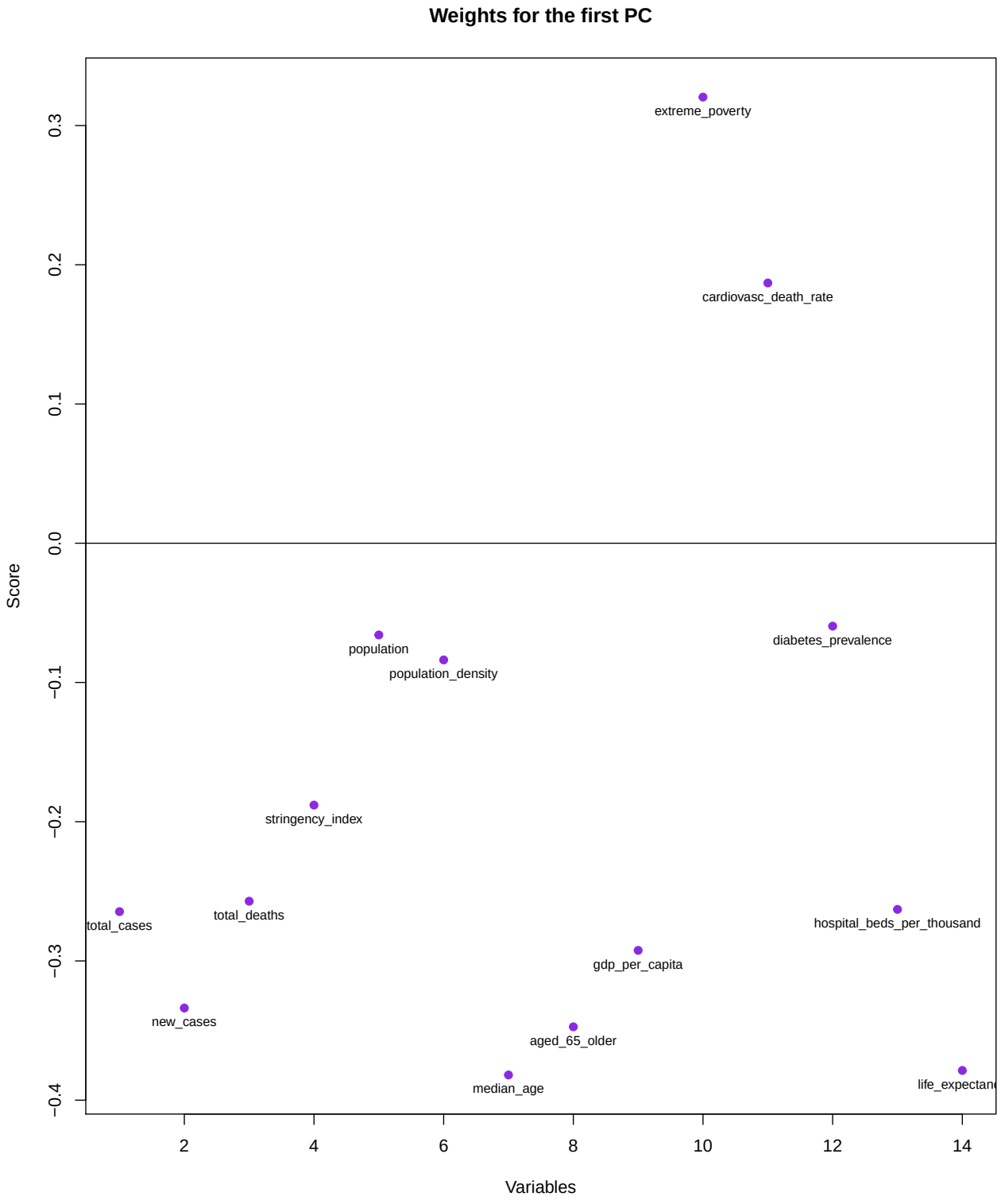
See the loadings of the first five principal components, loadings are the correlations between the original variables and the unit-scaled principal components, so we can examine the importance of each one of the original variables.

In this case, we see that the variables that are relevant for the first principal component are median_age (weight of -0.38833), life_expectancy (weight of -0.37843), etc; the variables that are important for the second principal component are population (weight of -0.52215), total_cases of COVID19 (weight of -0.42947), etc.

```
X_pcs$rotation[,1:5]
#>                                    PC1         PC2          PC3          PC4
#> total_cases                -0.26458778 -0.44002379  0.044449845 -0.0025742426
#> new_cases                  -0.33383071 -0.29386217  0.005974229  0.0742520415
#> total_deaths               -0.25711227 -0.42276130  0.072505132  0.1512397090
#> stringency_index           -0.18807378 -0.11091431  0.449292352  0.0945087682
#> population                 -0.06588023 -0.51687834 -0.069044789 -0.0007256816
#> population_density         -0.08380688  0.10941126  0.304420200 -0.2314136024
#> median_age                 -0.38190572  0.20149725 -0.172891359  0.0465889585
#> aged_65_older              -0.34732708  0.13664277 -0.348381375 -0.0801965062
#> gdp_per_capita             -0.29241247  0.17811156  0.098726649 -0.2472329599
#> extreme_poverty             0.32041383 -0.17820880 -0.126631057 -0.2834348210
#> cardiovasc_death_rate       0.18693083 -0.05148111 -0.256313480  0.6828378474
#> diabetes_prevalence        -0.05956384  0.25828293  0.468226963  0.5143413046
#> hospital_beds_per_thousand -0.26298934  0.15388118 -0.475024378  0.1629214705
#> life_expectancy            -0.37867607  0.18457158  0.076186855 -0.0412615430
#>                                    PC5
#> total_cases                 0.034834803
#> new_cases                   0.053361240
#> total_deaths               -0.003309426
#> stringency_index           -0.311914128
#> population                  0.204771791
#> population_density          0.860266721
#> median_age                  0.054616515
#> aged_65_older               0.061185721
#> gdp_per_capita             -0.125499589
#> extreme_poverty             0.131414144
#> cardiovasc_death_rate       0.216718307
#> diabetes_prevalence         0.129683627
#> hospital_beds_per_thousand  0.109690441
#> life_expectancy            -0.032928415
```

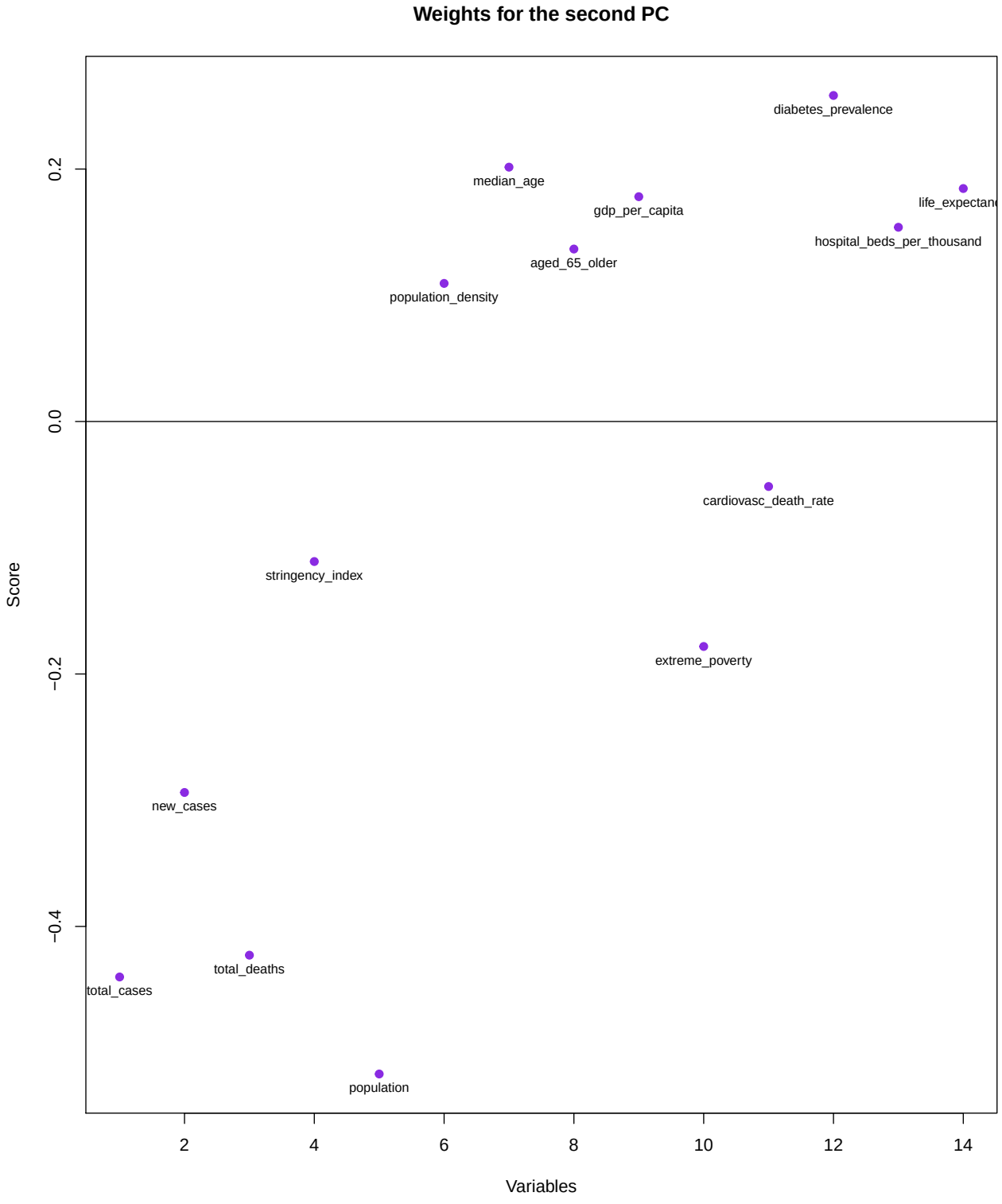
Plot of the first PC

```
plot(1:p,X_pcs$rotation[,1],pch=19,col=color_1,main="Weights for the first PC", xlab="Variables",ylab="S
abline(h=0)
text(1:p,X_pcs$rotation[,1],labels=colnames(covid_quan_log),pos=1,col=color_5,cex=0.75)
```



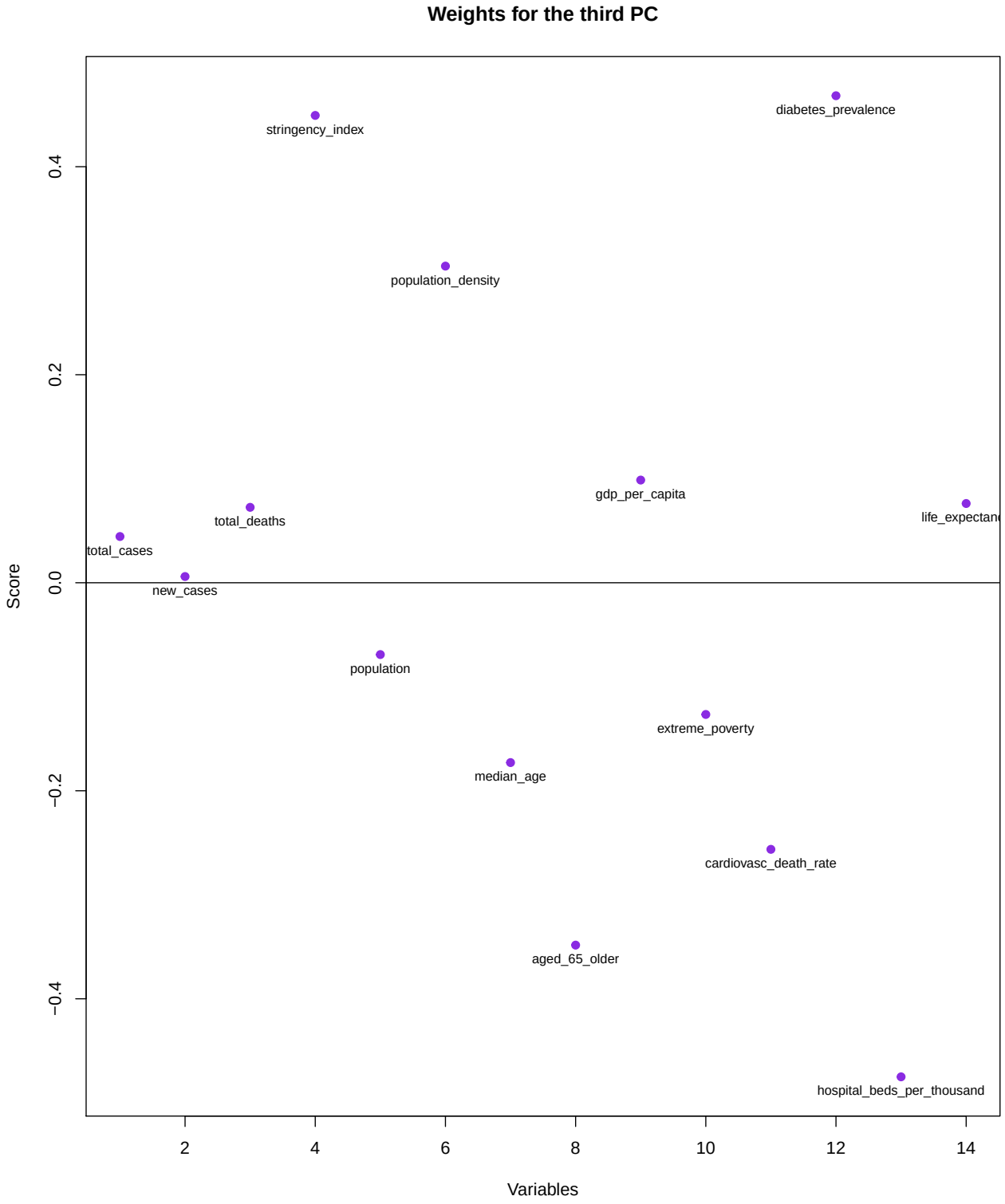
Plot of the second PC

```
plot(1:p,X_pcs$rotation[,2],pch=19,col=color_1,main="Weights for the second PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,2],labels=colnames(covid_quan_log),pos=1,col=color_5,cex=0.75)
```



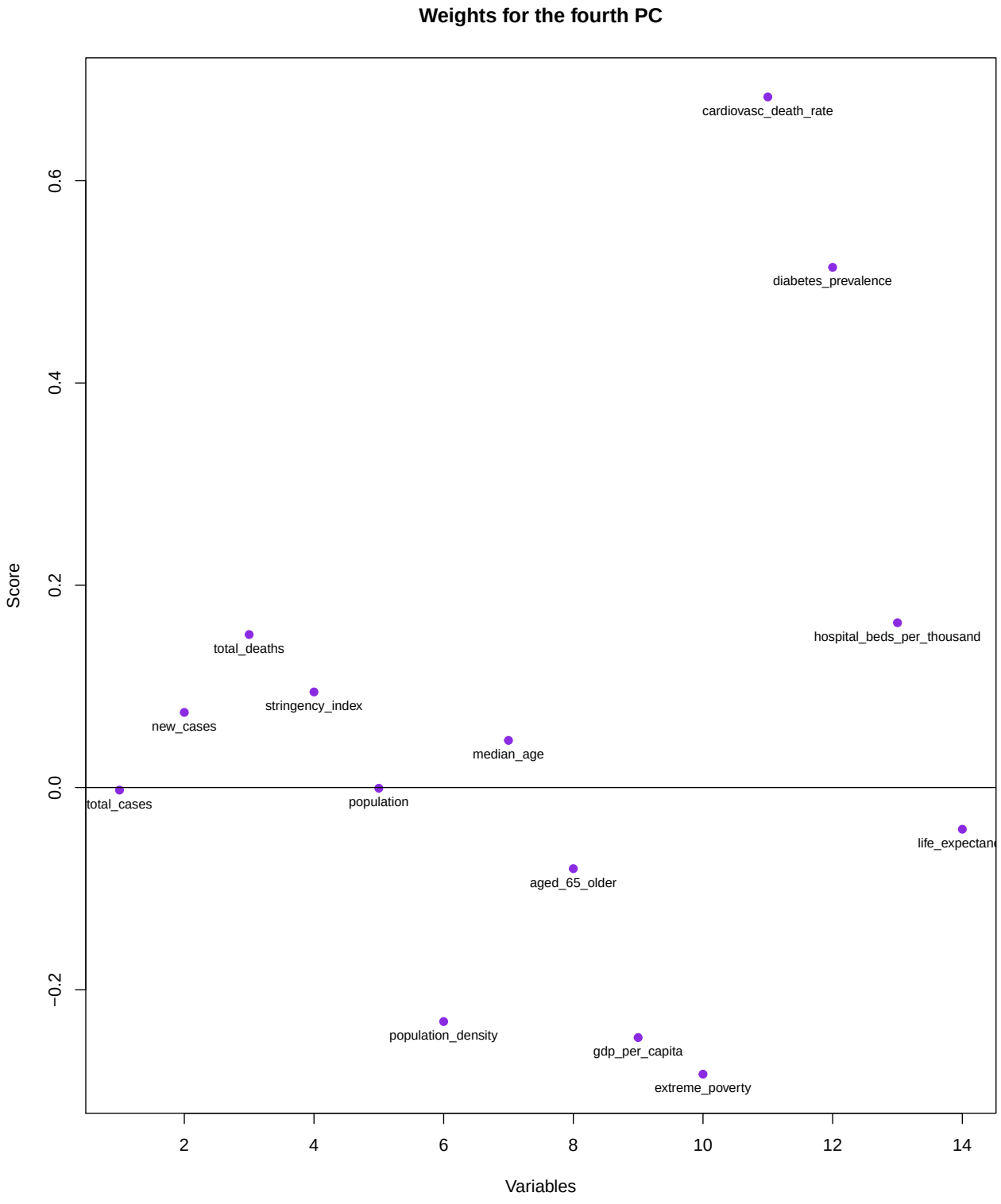
Plot of the third PC

```
plot(1:p,X_pcs$rotation[,3],pch=19,col=color_1,main="Weights for the third PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,3],labels=colnames(covid_quan_log),pos=1,col=color_5,cex=0.75)
```



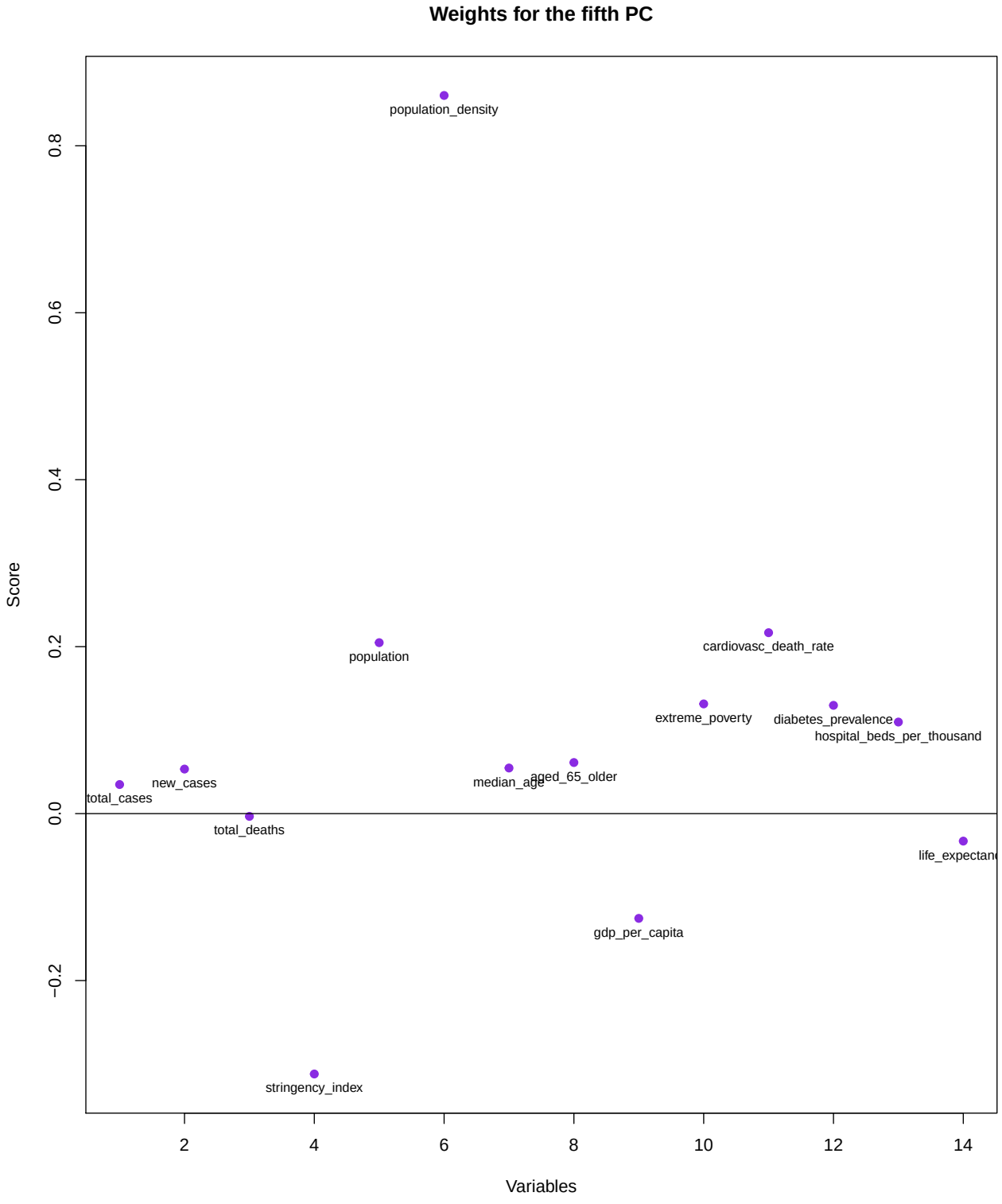
Plot of the fourth PC

```
plot(1:p,X_pcs$rotation[,4],pch=19,col=color_1,main="Weights for the fourth PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,4],labels=colnames(covid_quan_log),pos=1,col=color_5,cex=0.75)
```



Plot of the fifth PC

```
plot(1:p,X_pcs$rotation[,5],pch=19,col=color_1,main="Weights for the fifth PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,5],labels=colnames(covid_quan_log),pos=1,col=color_5,cex=0.75)
```

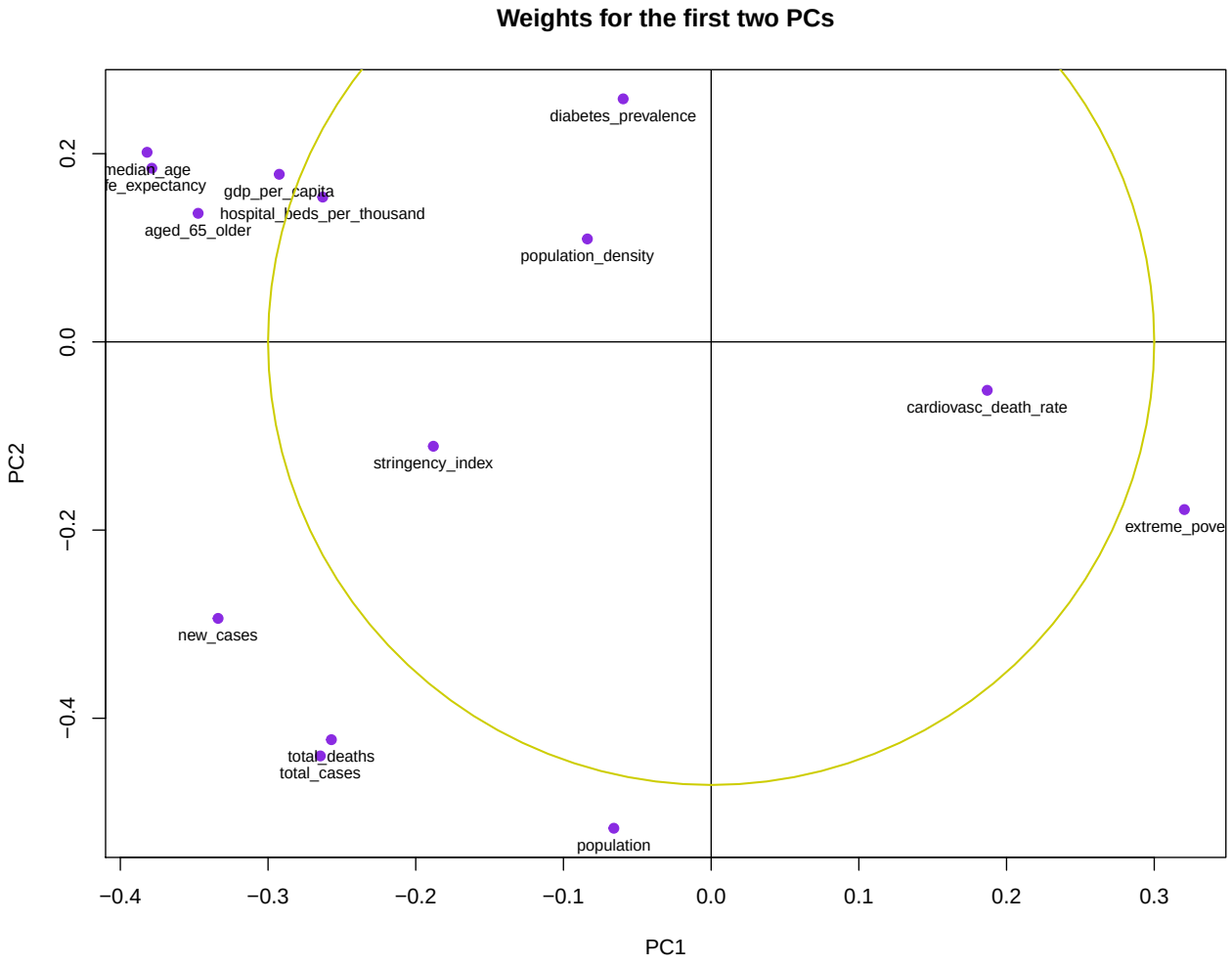


The largest values in magnitude of first principal component are associated with developing countries because of the high values of the variable extreme poverty, and high rate of cardiovascular death rate.

The largest values in magnitude of second principal component are associated with developed countries because of the high values of the variable GDP per capita, life expectancy and hospital beds per thousand, also, we have seen in the first part (vusualization of variables) that the median age of a country is related to how developed the country is, so the conclusion here fits the previous analysis.

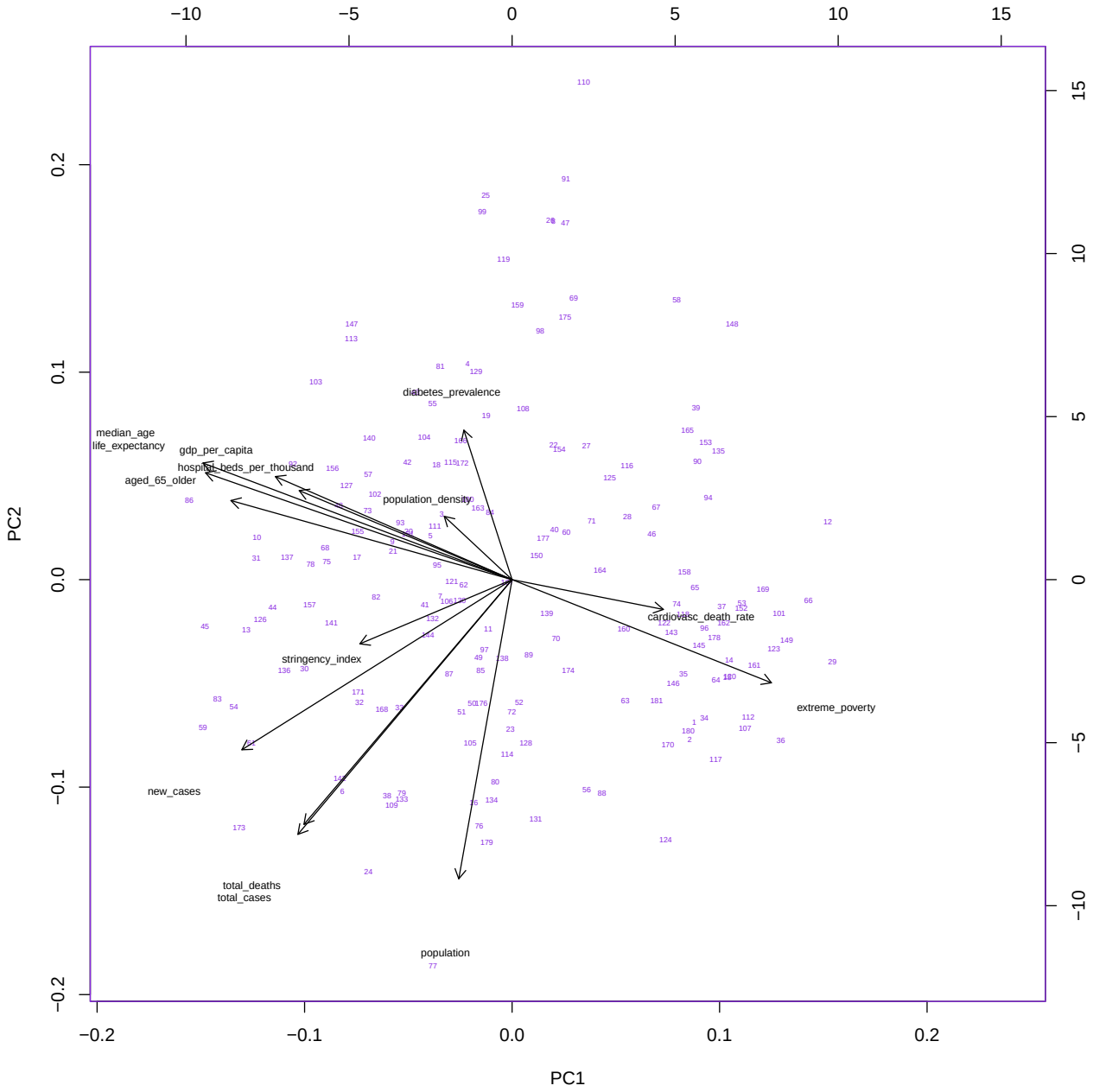
On the other hand, there is something interesting here, it may tell us that the people from highly developed countries may be more prone to develop diabetes. (The radius is arbitrary)

```
plot(X_pcs$rotation[,1:2],pch=19,col=color_1,main="Weights for the first two PCs")
abline(h=0,v=0)
text(X_pcs$rotation[,1:2],labels=colnames(covid_quan),pos=1,col=color_5,cex=0.75)
draw.circle(0,0,0.3,border=color_4,lwd=1.5)
```



The biplot is an alternative way to plot points and the first two PCs together, this is useful for us because our dataset is not very large.

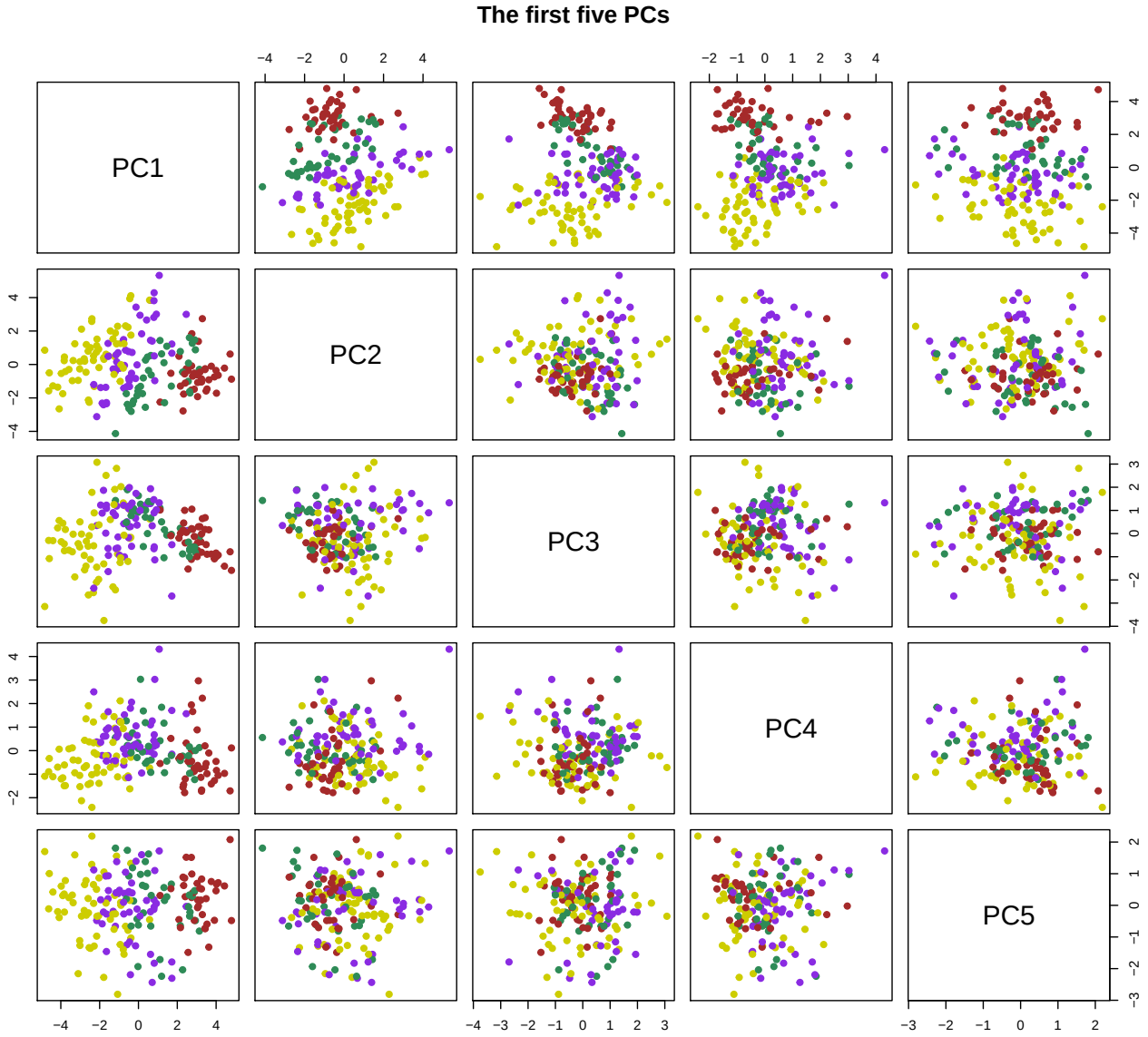
```
biplot(X_pcs,col=c(color_1,color_5),cex=c(0.4,0.6))
```



Plotting PC scores

Now we plot the scores of the first five principal components, and it shows that the first principal component is the key to show the 4 groups of development.

```
pairs(X_pcs$x[,1:5],col=development_colors,pch=19,main="The first five PCs")
```



The larger the value of the first PC is, then the least developed the country in question might be. Our purpose here is to check the largest values of the first PC.

```
sort(X_pcs$x[,1],decreasing=TRUE)[1:10]
#>   [1] 4.791435 4.721829 4.431973 4.107863 4.020891 3.995233 3.914876 3.752949
#>   [9] 3.624458 3.532041
```

Plotting the first two PCs grouped by development

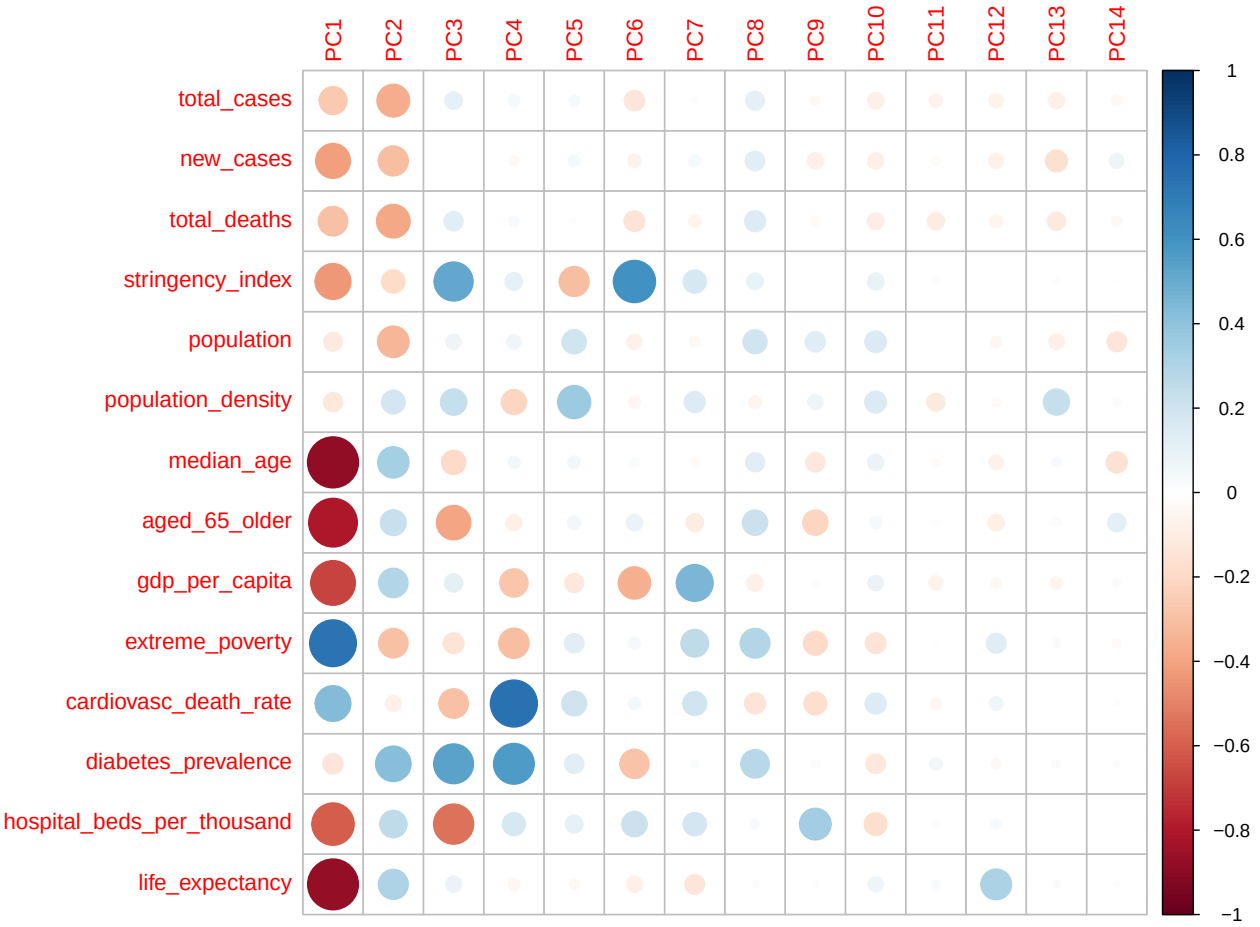
Now we plot the correlations between the original data set and the principal components, it is used to understand what are the most important variables in our dataset in terms of variability.

We can observe from the following plot that the first principal component explains the most the variability, then the second, and so on.

As we said before, the variables that are important for the first principal component are median age, life expectancy, percentage of aged 65 or older, and with a negative correlation, extreme poverty and cardiovascular death rate with a positive correlation. The higher the first principal component is, the lower the HDI of country will most likely be, and other way around.

On the contrary, the second principal component is positively correlated to GDP per capita, and negatively correlated to total deaths and total cases of COVID-19. Which probably means that, the higher the value of PC2, the more developed a country is (higher HDI).

```
corrplot(cor(covid_quan,X_pcs$x),is.corr=T)
```

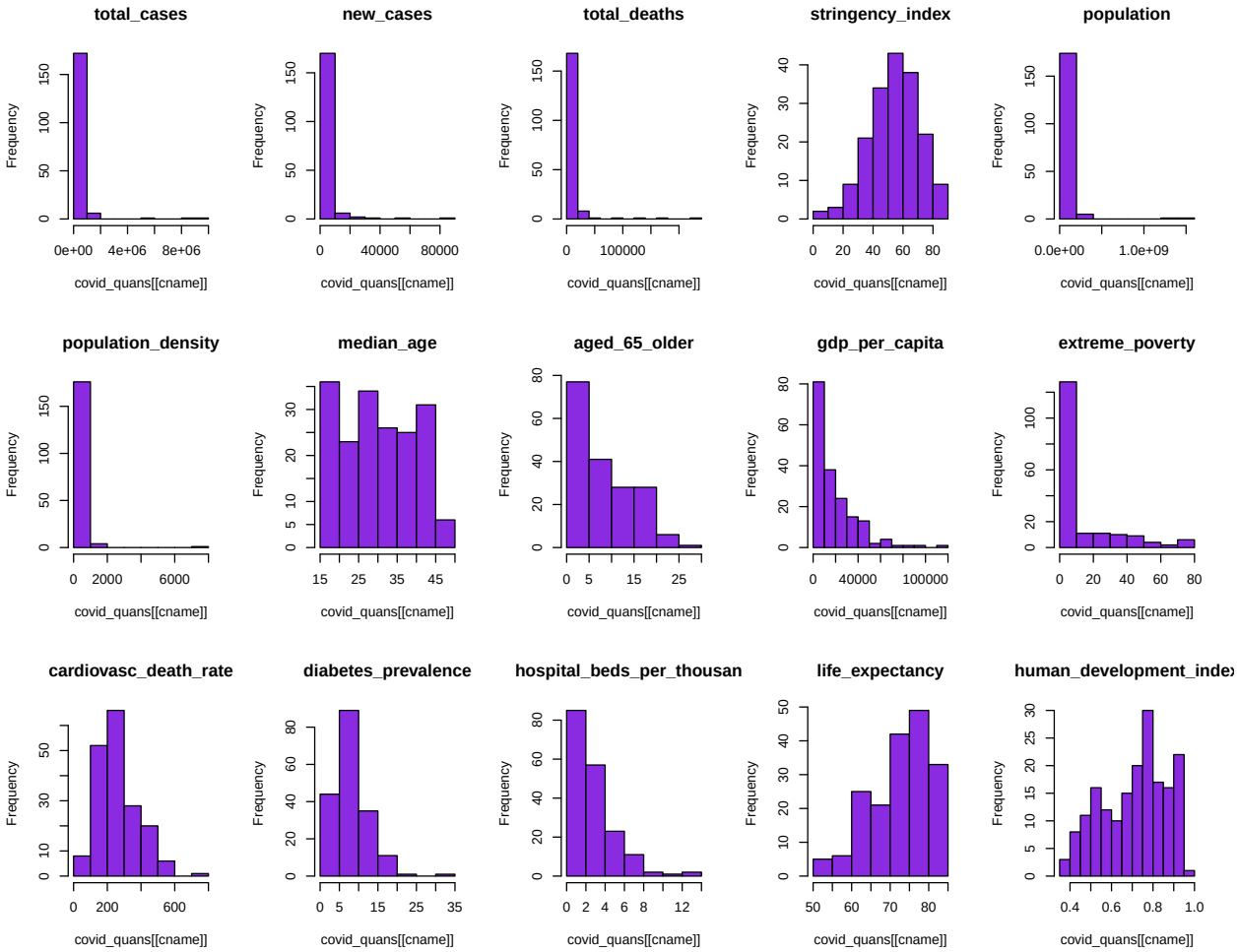


PCA segregating by continent

```
n <- nrow(covid_quans)
p <- ncol(covid_quans)
```

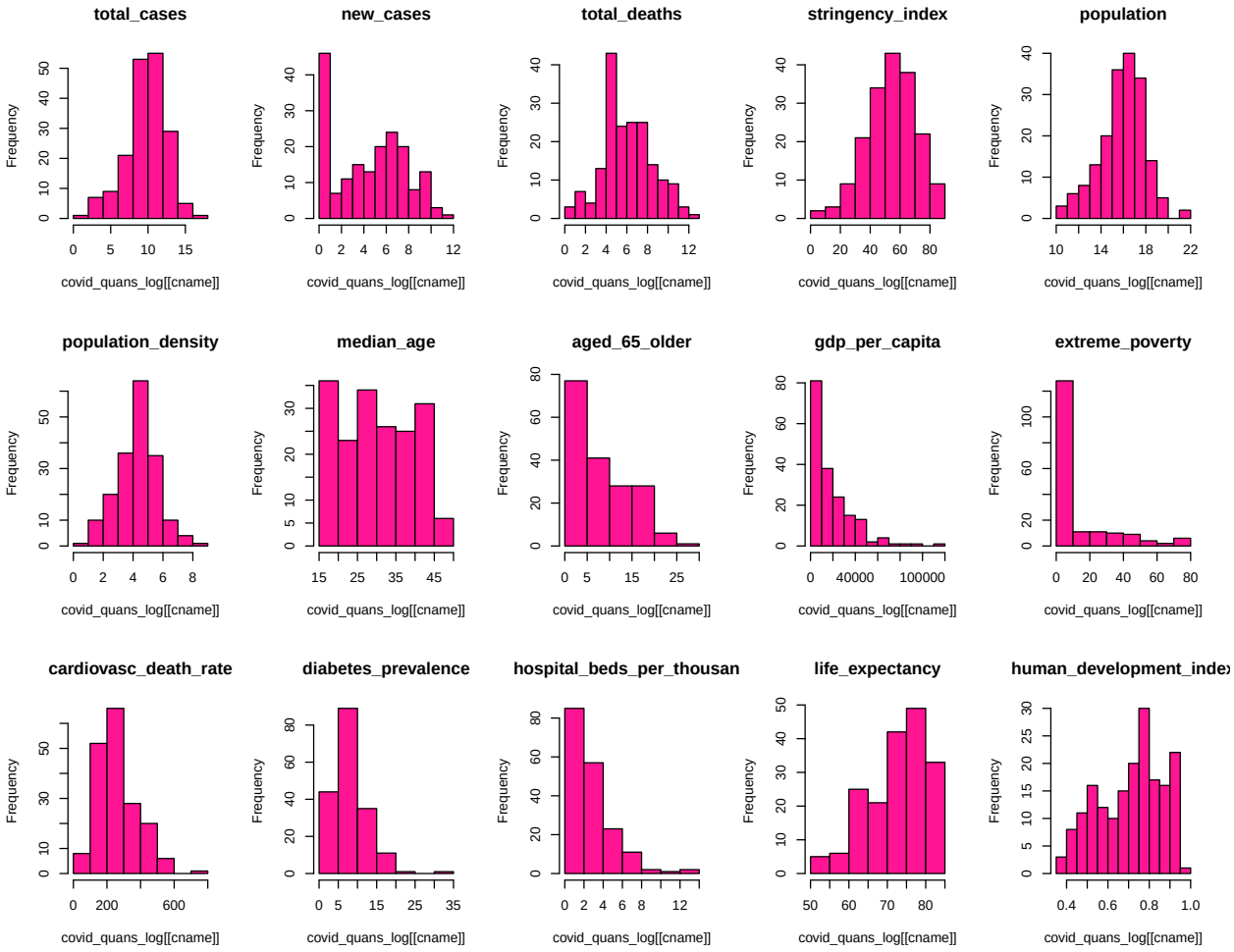
Checking the distribution of each variable using histograms

```
par(mfrow=c(3,5))
sapply(names(covid_quans),function(cname){hist(covid_quans[[cname]],main=cname,col = color_1)})
```



Transforming some of the variables using log because some of the variables are highly positively skewed

```
covid_quans_log=covid_quans
for (i in 1:6) {
    if (i != 4) {
        covid_quans_log[,i]=log(covid_quans[,i])
    }
}
covid_quans_log$new_cases[which(covid_quans_log$new_cases== -Inf )] = 0
par(mfrow=c(3,5))
sapply(names(covid_quans_log),function(cname){hist(covid_quans_log[[cname]],main=cname,col = color_6)})
```



PC Scores:

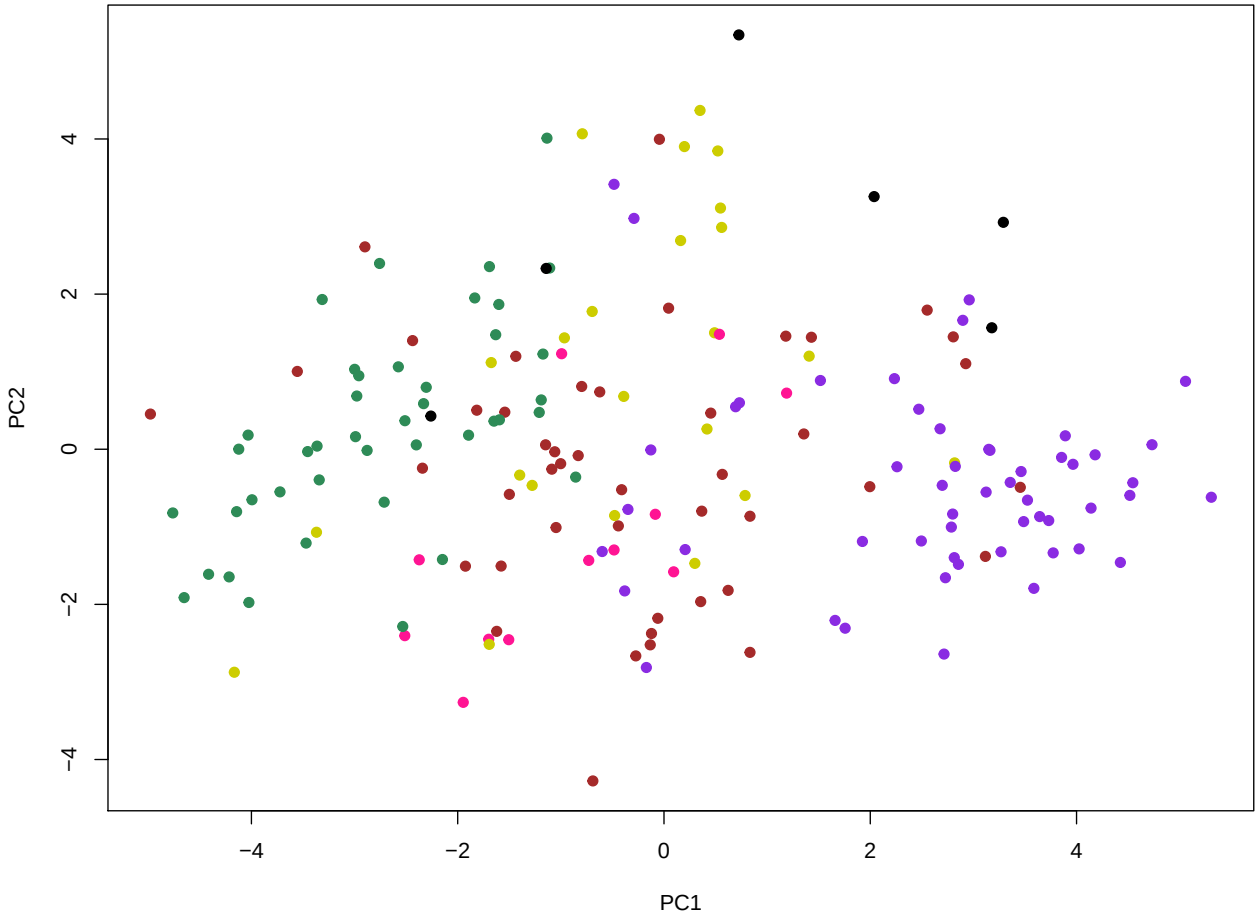
```
X_pcs <- prcomp(covid_quans_log, scale=TRUE)
```

Plotting the first two PCs grouped by continent

We can see here the principal components don't differentiate perfectly all 6 continents, but it classifies very well between Europe (*seagreen*) and Africa (*blueviolet*), just as what we had hypothesised on the first part (we can separate very clearly these two continents by using some important features like GDP or extreme poverty).

And we cannot comfortably conclude anything about Oceania as we only have 6 observations.

```
plot(X_pcs$x[,1:2],pch=19,col=continent_colors)
```



Checking eigenvalues and explained variance

Now we check the eigenvalues and the cumulative percentage of explained variance, and we have decided to select the first 5 principal components.

We have used the 33% of the variables and will be keeping the 82.6% of the total information.

By using the same amount of principal components, we explain a higher percentage of the total variance because of the variable HDI.

```
get_eigenvalue(X_pcs) #maybe 5
#>        eigenvalue variance.percent cumulative.variance.percent
#> Dim.1  6.16956410       41.1304273                   41.13043
#> Dim.2  2.78973598       18.5982398                   59.72867
#> Dim.3  1.31350536        8.7567024                   68.48537
#> Dim.4  1.19308362        7.9538908                   76.43926
#> Dim.5  0.94826535        6.3217690                   82.76103
#> Dim.6  0.72928726        4.8619151                   87.62294
#> Dim.7  0.42189163        2.8126109                   90.43556
#> Dim.8  0.37284471        2.4856314                   92.92119
#> Dim.9  0.32313905        2.1542603                   95.07545
#> Dim.10 0.23261294        1.5507530                   96.62620
#> Dim.11 0.17872448        1.1914966                   97.81770
#> Dim.12 0.14418062        0.9612042                   98.77890
#> Dim.13 0.09242524        0.6161683                   99.39507
#> Dim.14 0.05355892        0.3570595                   99.75213
#> Dim.15 0.03718072        0.2478715                  100.00000
```

We can examine the importance of each variable for the different principal components by using the loadings.

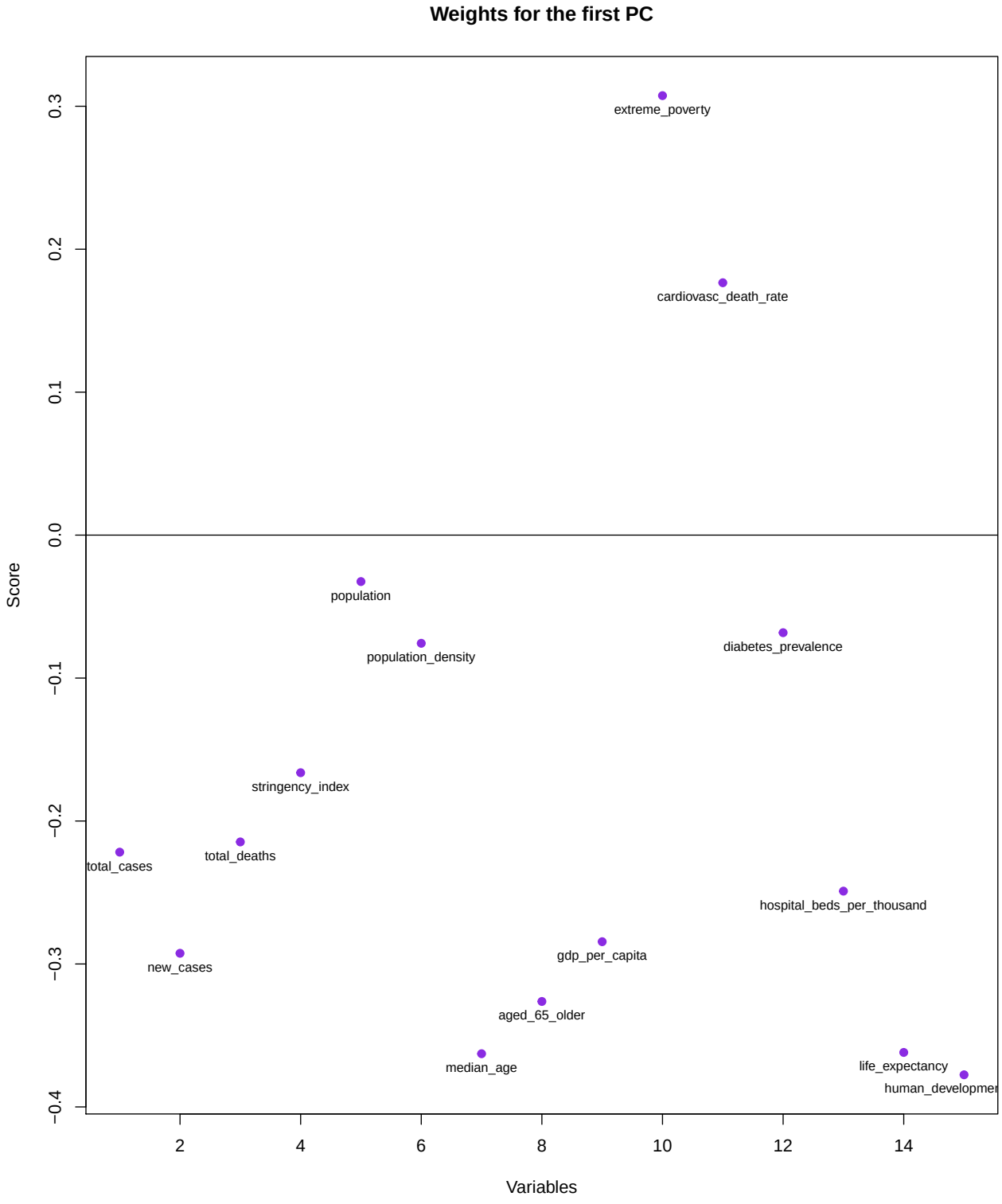
In this case, we see that the variable HDI is a very important variable as it is highly correlated with the first principal component (-0.37720), and other variables that are relevant for the first principal component are exactly the same as what we did before.

And the variables that are important for the second principal component are population (weight of -0.52200), total_cases of COVID19 (weight of -0.45787), etc, which correspond to the first PCA that we have performed.

```
X_pcs$rotation[,1:5]
#>                                  PC1         PC2          PC3          PC4
#> total_cases                -0.22176863 -0.46222779  0.044384568 -0.002692137
#> new_cases                  -0.29252261 -0.32738487  0.006153229  0.074673623
#> total_deaths               -0.21461888 -0.44545235  0.072495325  0.151204126
#> stringency_index           -0.16620742 -0.13202052  0.449440352  0.094686016
#> population                 -0.03249565 -0.51669922 -0.069275393 -0.001071483
#> population_density         -0.07567331  0.09020006  0.305131109 -0.229266152
#> median_age                 -0.36277932  0.15375621 -0.172271354  0.047964023
#> aged_65_older              -0.32622702  0.09291685 -0.347809936 -0.078843976
#> gdp_per_capita             -0.28442337  0.14634413  0.098909631 -0.246978635
#> extreme_poverty             0.30743282 -0.14048124 -0.127022158 -0.284075393
#> cardiovasc_death_rate       0.17655335 -0.03061868 -0.256374818  0.682710230
#> diabetes_prevalence        -0.06826167  0.24664700  0.468549621  0.514875348
#> hospital_beds_per_thousand -0.24904579  0.11918315 -0.474471881  0.164231237
#> life_expectancy            -0.36188079  0.14021567  0.076644059 -0.040340036
#> human_development_index    -0.37752332  0.13831076 -0.002070034 -0.004589218
#>                                  PC5
#> total_cases                0.028619928
#> new_cases                  0.060081971
#> total_deaths              -0.003338583
#> stringency_index          -0.277580848
#> population                 0.184864105
#> population_density         0.869048213
#> median_age                 0.081293663
#> aged_65_older              0.089301601
#> gdp_per_capita            -0.123470255
#> extreme_poverty            0.113994183
#> cardiovasc_death_rate      0.201781716
#> diabetes_prevalence        0.127656857
#> hospital_beds_per_thousand 0.135769658
#> life_expectancy           -0.014722442
#> human_development_index   -0.102975495
```

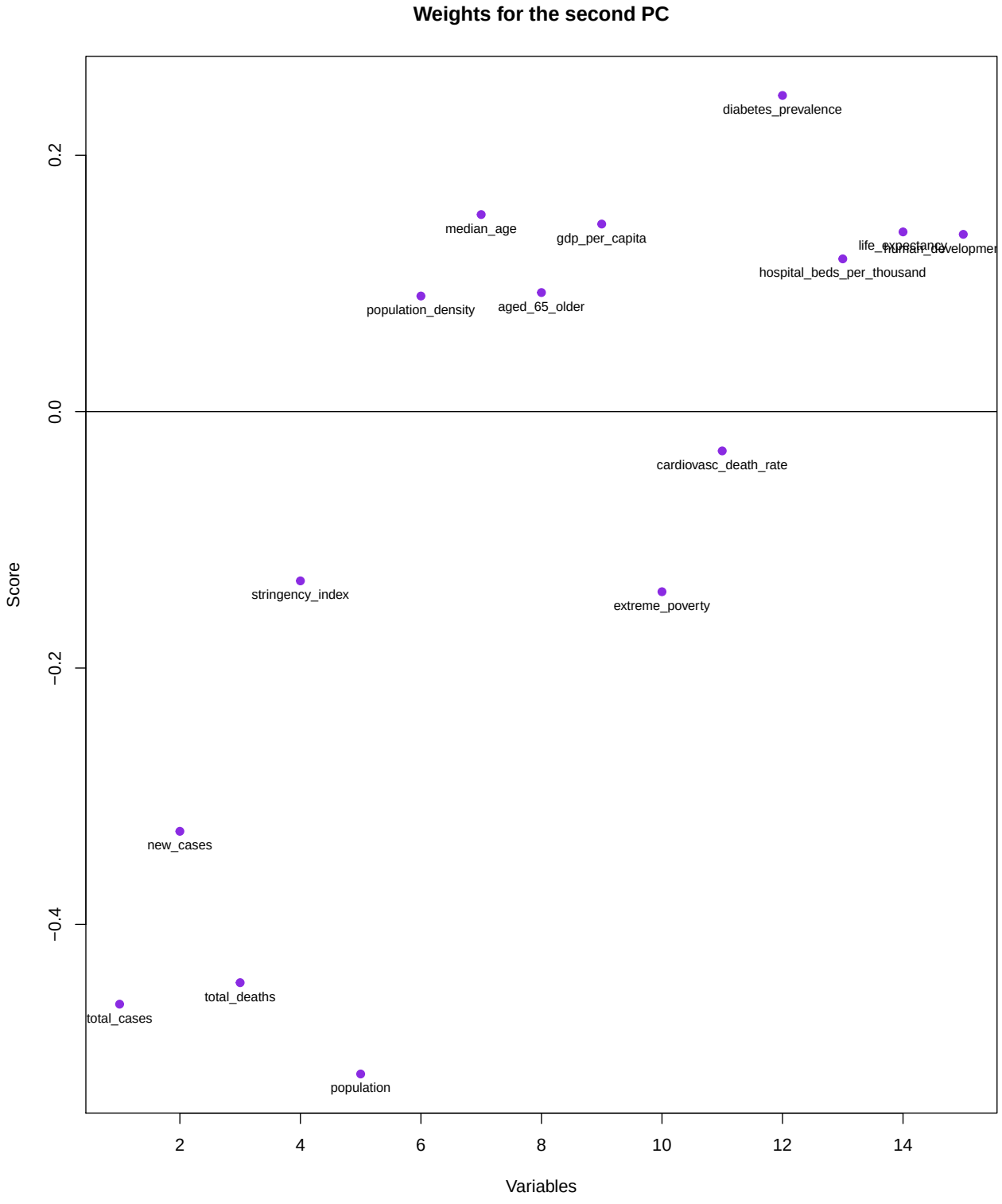
Plot of the first PC

```
plot(1:p,X_pcs$rotation[,1],pch=19,col=color_1,main="Weights for the first PC", xlab="Variables",ylab="S
abline(h=0)
text(1:p,X_pcs$rotation[,1],labels=colnames(covid_quans_log),pos=1,col=color_5,cex=0.75)
```



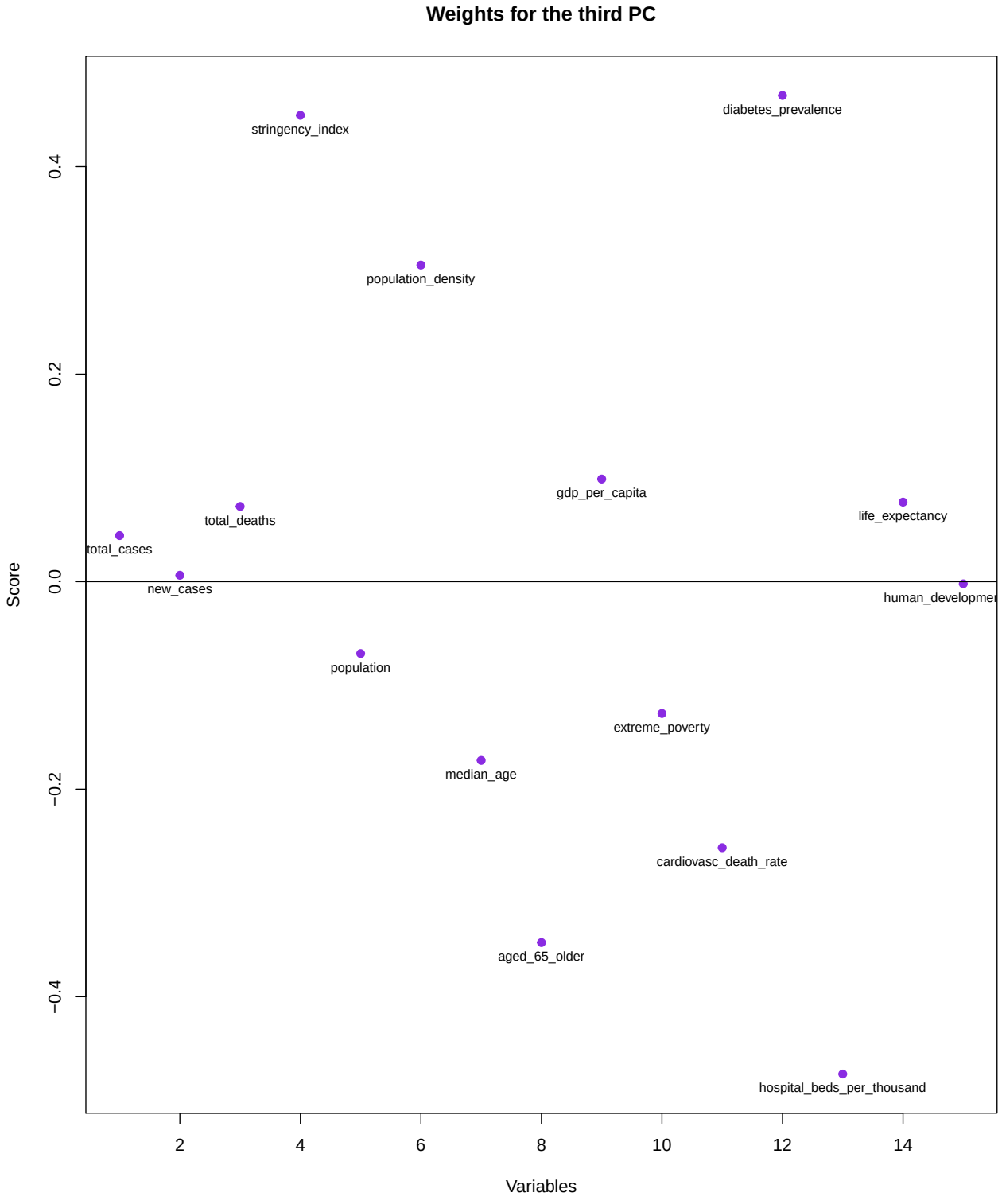
Plot of the second PC

```
plot(1:p,X_pcs$rotation[,2],pch=19,col=color_1,main="Weights for the second PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,2],labels=colnames(covid_quans_log),pos=1,col=color_5,cex=0.75)
```



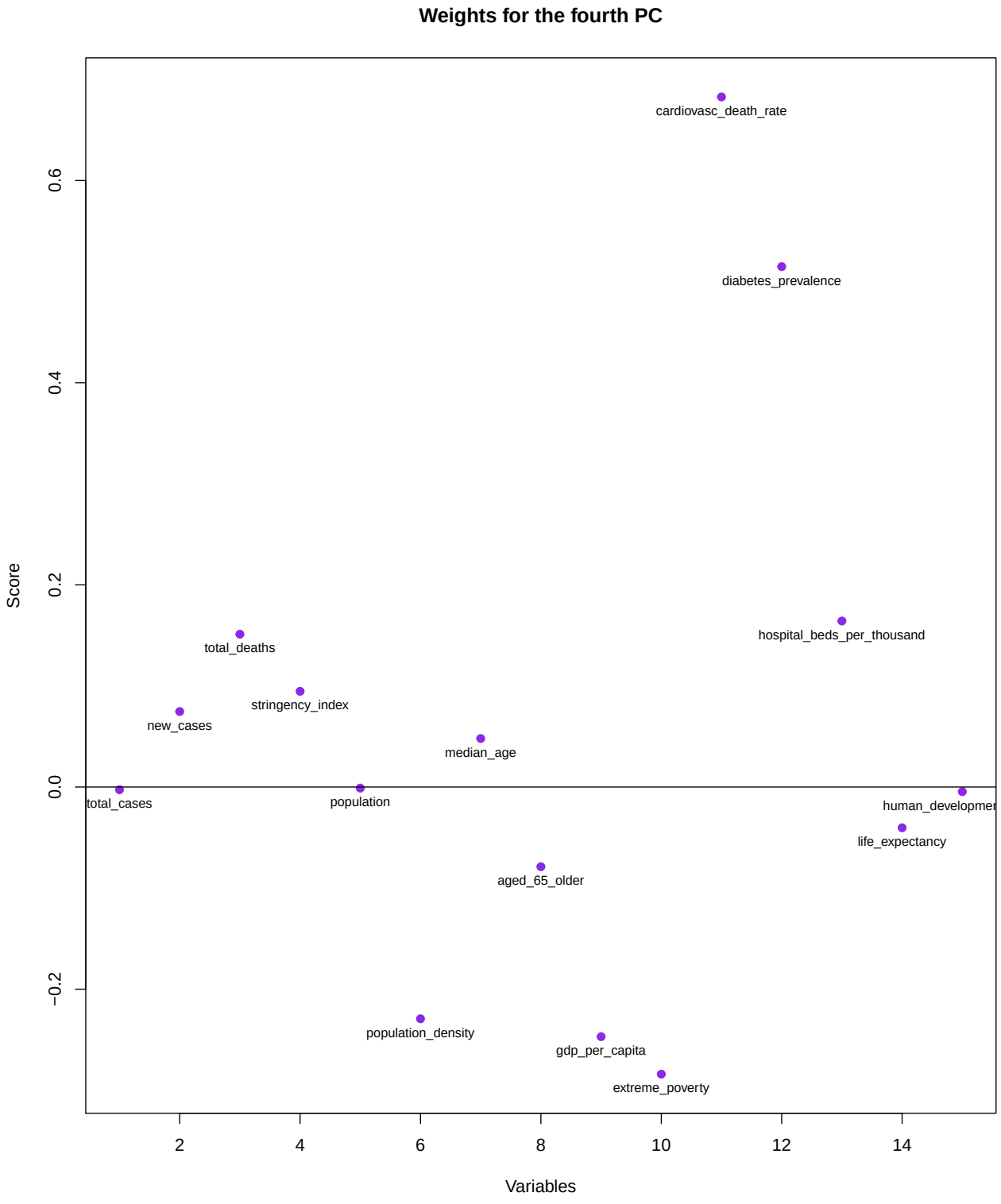
Plot of the third PC

```
plot(1:p,X_pcs$rotation[,3],pch=19,col=color_1,main="Weights for the third PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,3],labels=colnames(covid_quans_log),pos=1,col=color_5,cex=0.75)
```



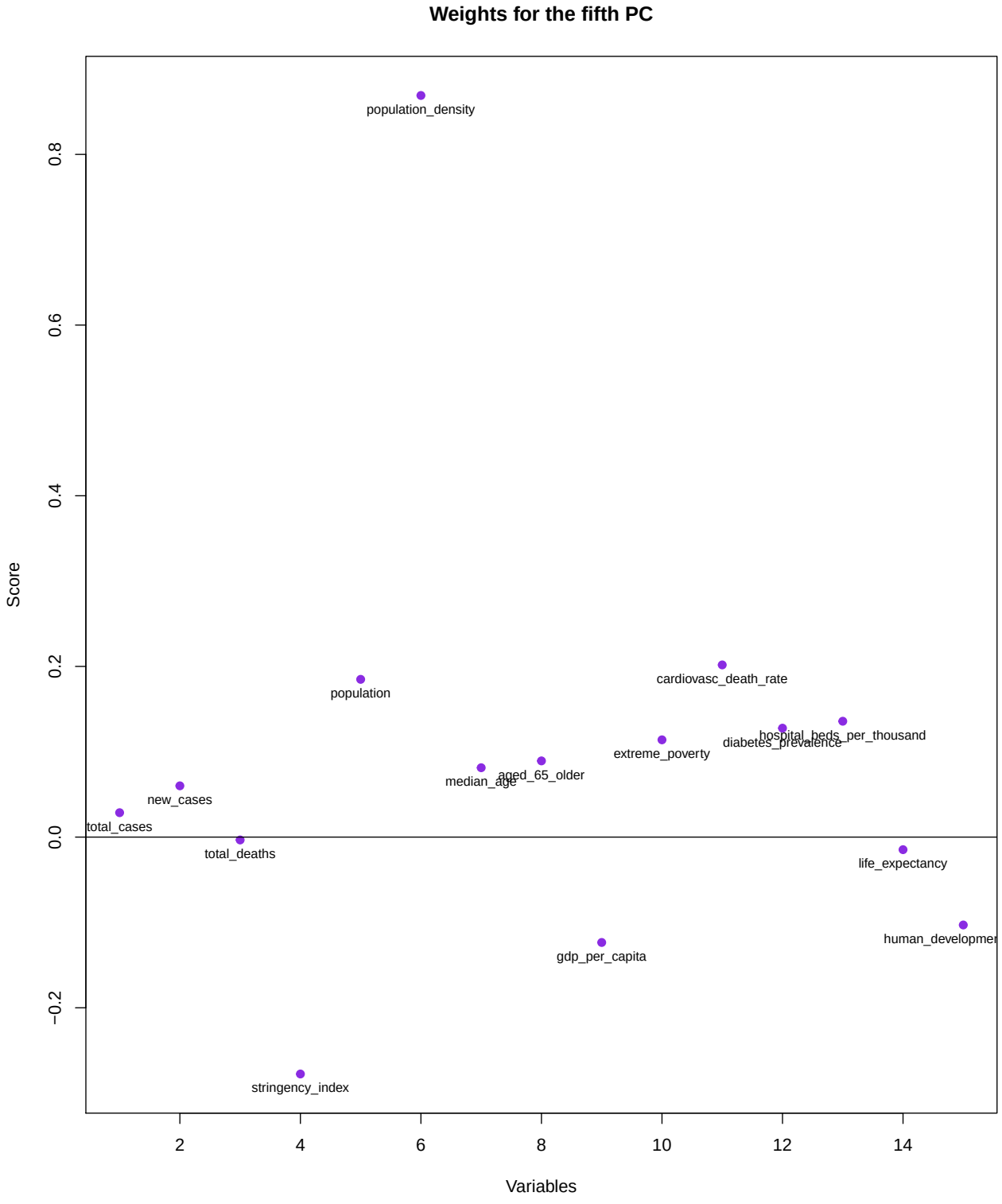
Plot of the fourth PC

```
plot(1:p,X_pcs$rotation[,4],pch=19,col=color_1,main="Weights for the fourth PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,4],labels=colnames(covid_quans_log),pos=1,col=color_5,cex=0.75)
```



Plot of the fifth PC

```
plot(1:p,X_pcs$rotation[,5],pch=19,col=color_1,main="Weights for the fifth PC",
     xlab="Variables",ylab="Score")
abline(h=0)
text(1:p,X_pcs$rotation[,5],labels=colnames(covid_quans_log),pos=1,col=color_5,cex=0.75)
```

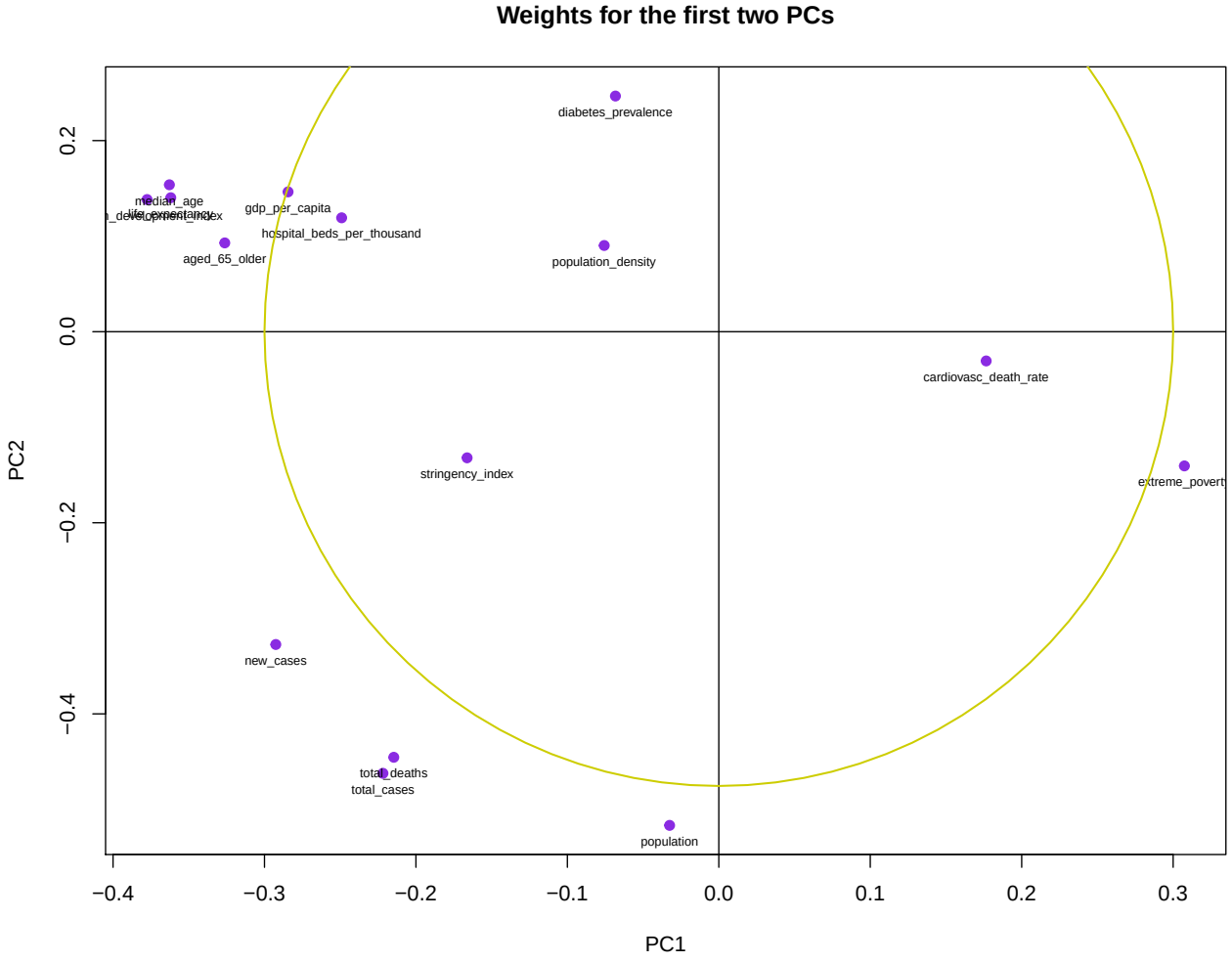


The conclusions that we can get from this plot are basically the same as before.

Largest values in magnitude of first principal component are associated with less developed countries because of the high values of the variable extreme poverty, and high rate of cardiovascular death rate.

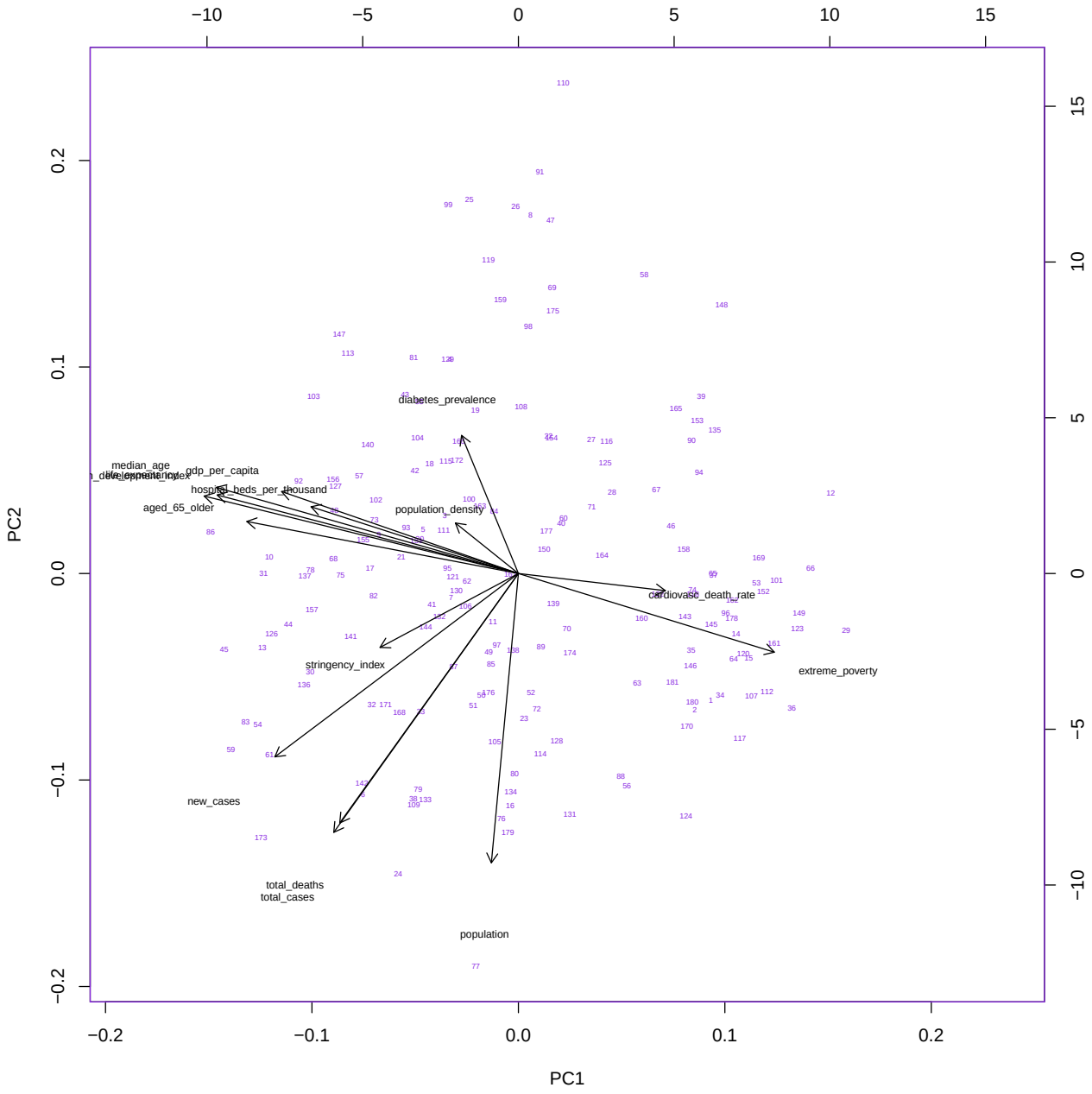
Largest values in magnitude of second principal component are associated with more developed countries because of the high values of the HDI, variable GDP, life expectancy and hospital beds per thousand. (The radius is arbitrary)

```
plot(X_pcs$rotation[,1:2],pch=19,col=color_1,main="Weights for the first two PCs")
abline(h=0,v=0)
text(X_pcs$rotation[,1:2],labels=colnames(covid_quans_log),pos=1,col=color_5,cex=0.6)
library(plotrix)
draw.circle(0,0,0.3,border=color_4,lwd=1.5)
```



The biplot is an alternative way to plot points and the first two PCs together.

```
biplot(X_pcs,col=c(color_1,color_5),cex=c(0.4,0.6))
```

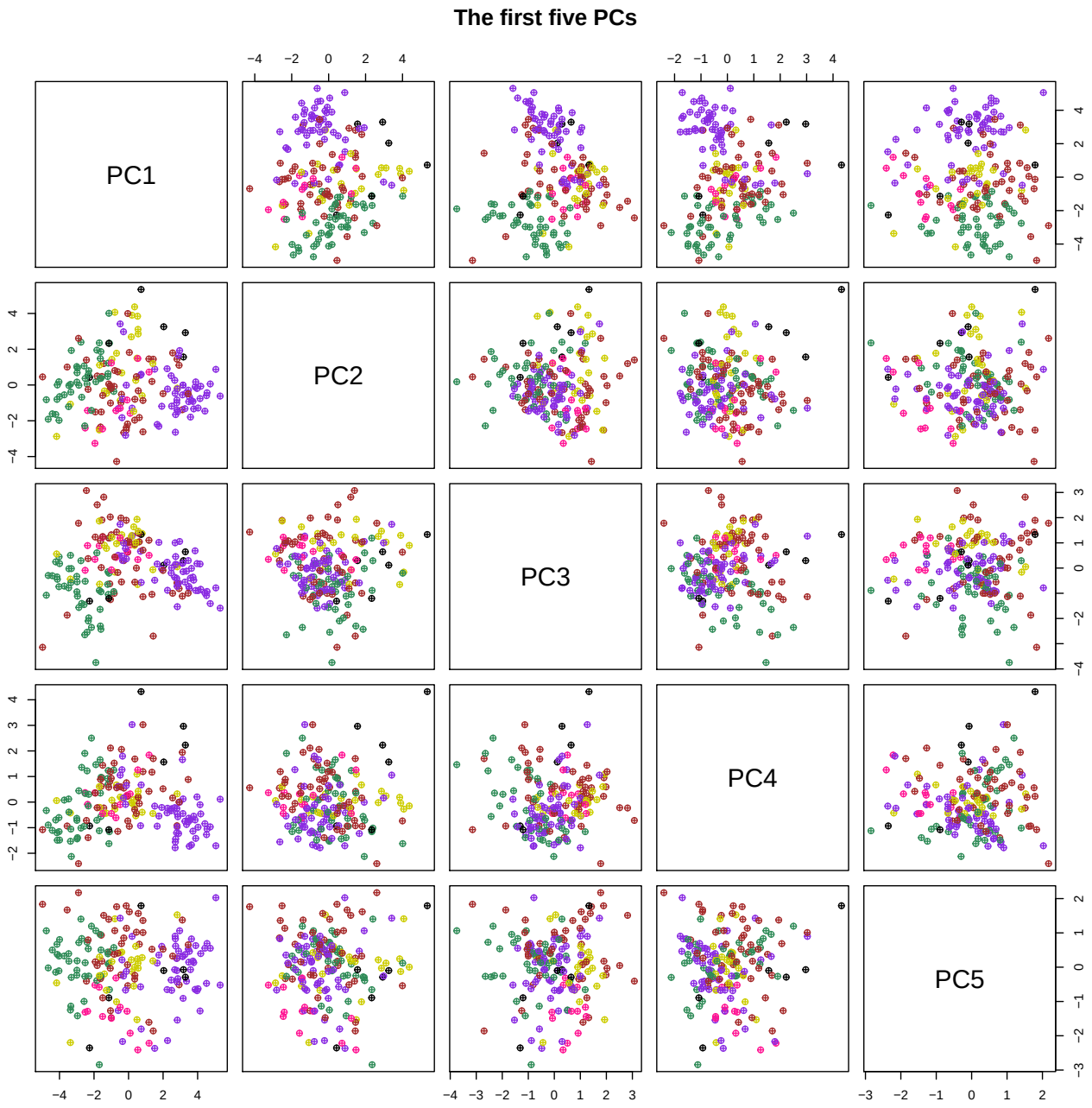


Plotting PC scores

Now we plot the scores of the first five principal components, and it shows that the first principal component is the key to show the 4 groups of development.

And we can only differ Africa from Europe very clearly.

```
pairs(X_pcs$x[,1:5],col=continent_colors,pch=10,main="The first five PCs")
```



The larger the value of the first PC is, then the least developed the country in question might be. Our purpose here is to check the largest values of the first PC.

```
sort(X_pcs$x[,1],decreasing=TRUE)[1:10]
#>  [1] 5.306764 5.056603 4.731922 4.545968 4.516960 4.425111 4.180303 4.140803
#>  [9] 4.025157 3.965852
```

Plotting the first two PCs grouped by continent

Now we plot the correlations between the original dataset and the principal components.

The variables that are important for the first principal component are the same with the addition of HDI with a negative correlation. The higher the first principal component is, the higher the probability that the country belongs to Africa and that it isn't in Europe, but we cannot say anything about the other continents, there are not enough differences between them for us to classify.

On the other hand, the second principal component is also very similar to our first PCA. Which probably means that, the higher the value of PC2, the more likely it becomes that it is located in Europe.

```
corrplot(cor(covid_quans,X_pcs$x),is.corr=T)
```