Final project: Step 1

Danyu Zhang, Limingrui Wan, Daniel Alonso

December 9th, 2020

Importing libraries

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

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_ca
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

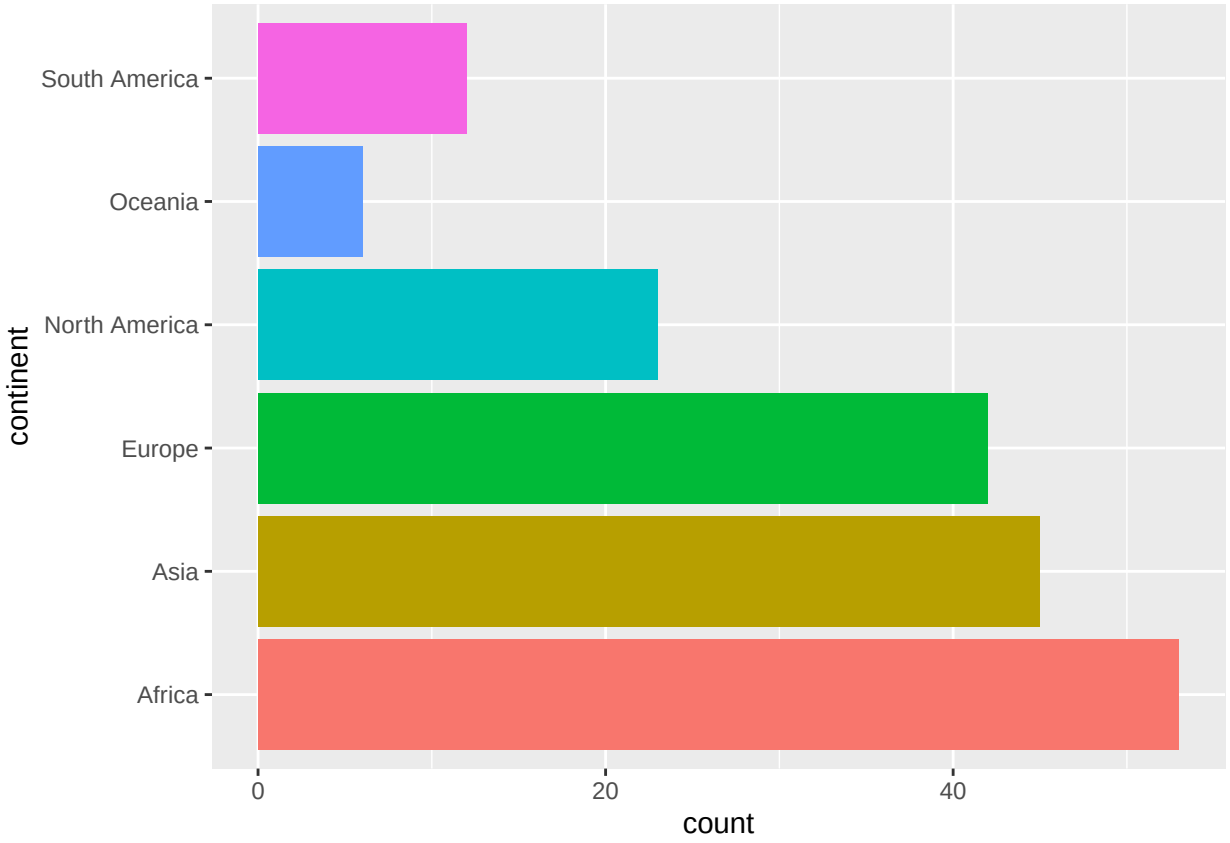
We select variables that we consider interesting to visualize, as the ones we haven't selected might be ralated to these or even ratios of them (in the case of total cases per million)

```
categorical <- c('location','continent','development')
interesting_vars <- c('total_cases','new_cases','total_deaths','stringency_index','population','populat
```

Plots with categorical variables

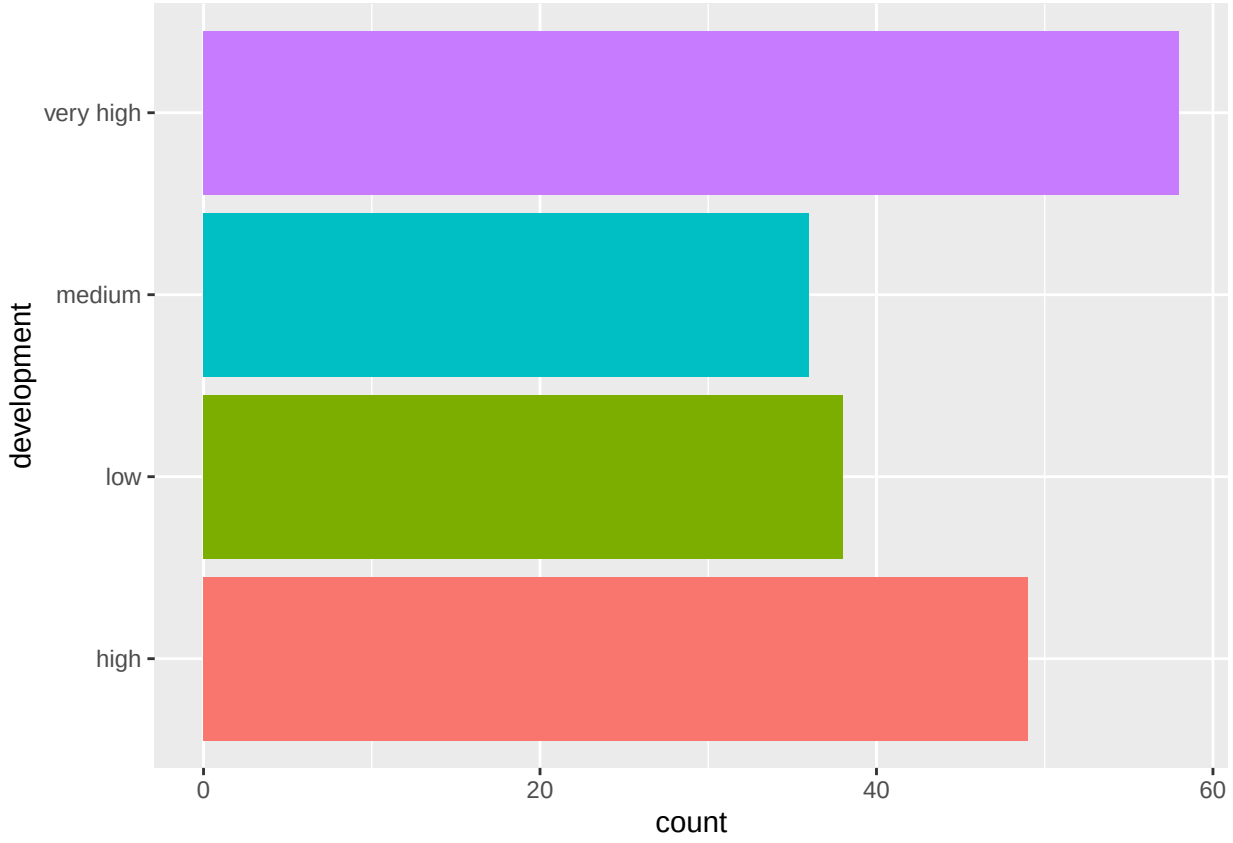
Countries per continent in the dataset

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



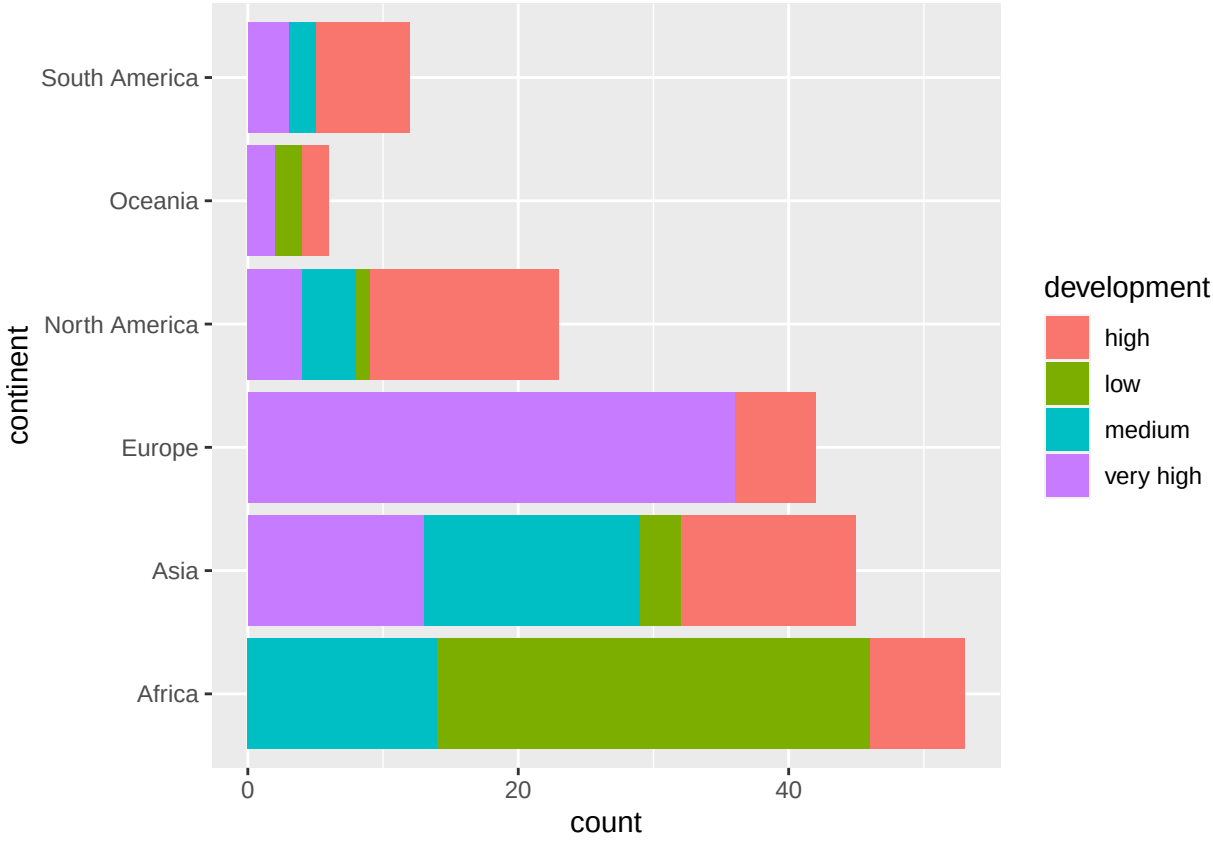
Amount of countries per HDI

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



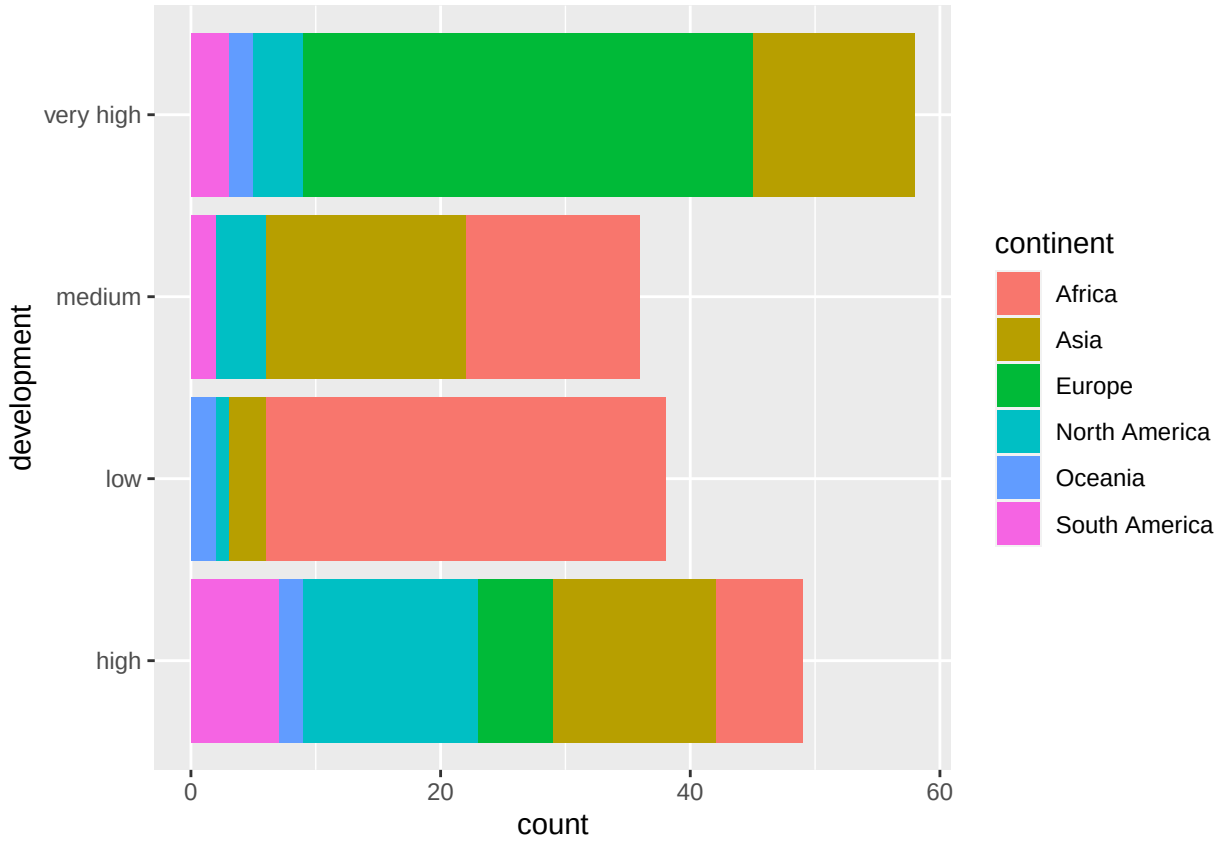
Countries per continent per HDI

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



Proportions of HDI per continent

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



Plots with numerical variables

Defining Colors:

```
color_1 <- "khaki"
color_2 <- "lightseagreen"
color_3 <- "lightpink2"
color_4 <- "gold"
```

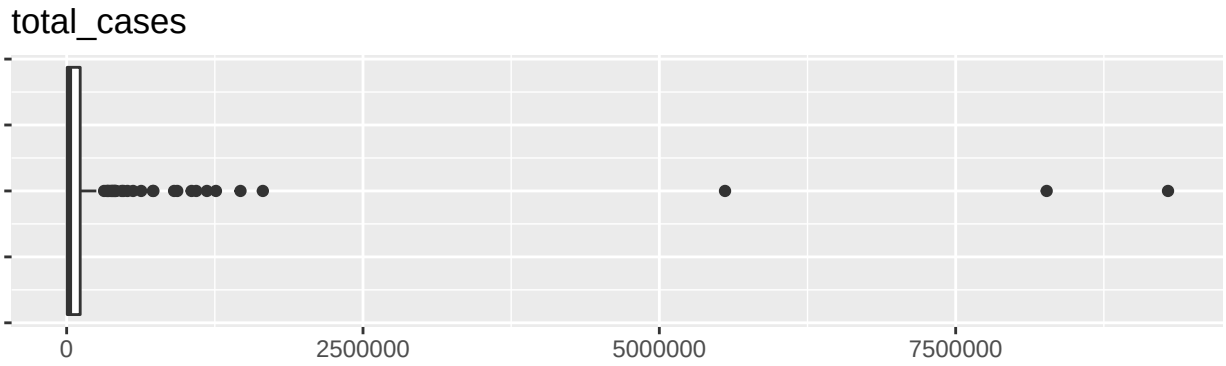
Function to plot quantitative variables

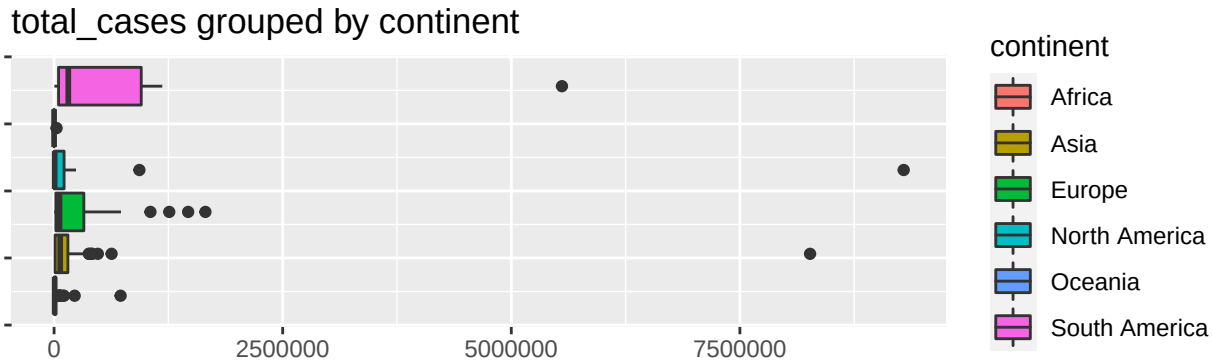
```
plots <- function(dataset ,col, type, density=TRUE, bins='default', xtick_angles='default') {
    var <- dataset %>% dplyr::select(col)
    if (bins == 'default') {bins = rep(10,3)}
    if (xtick_angles == 'default') {xtick_angles = rep(90,3)}
    if (type == 'boxplot') {
        p1 <- dataset %>% 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() +
```

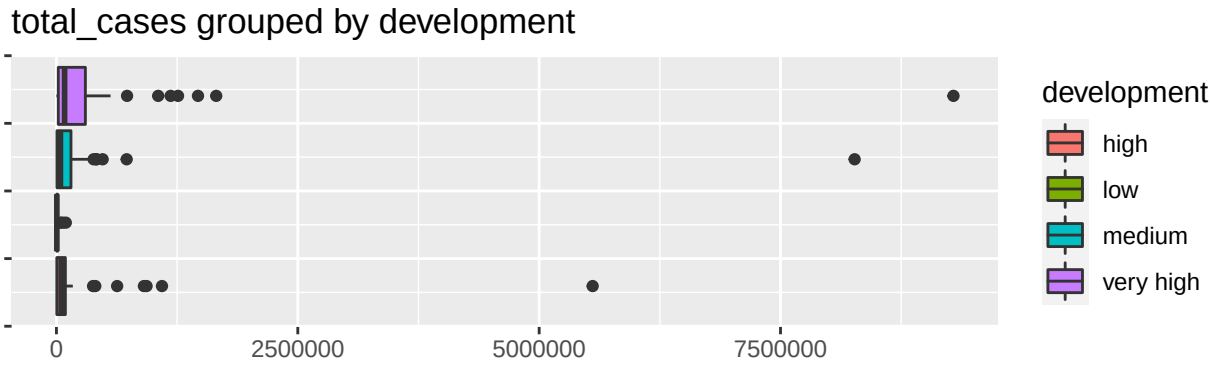
```
            ggtitle(str_interp("${col} grouped by continent")) +
            theme(axis.title.x=element_blank(),
                  axis.text.y=element_blank())
        p3 <- dataset %>% ggplot(aes(x=var[,1], fill=development)) +
            geom_boxplot() +
            ggtitle(str_interp("${col} 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')
```







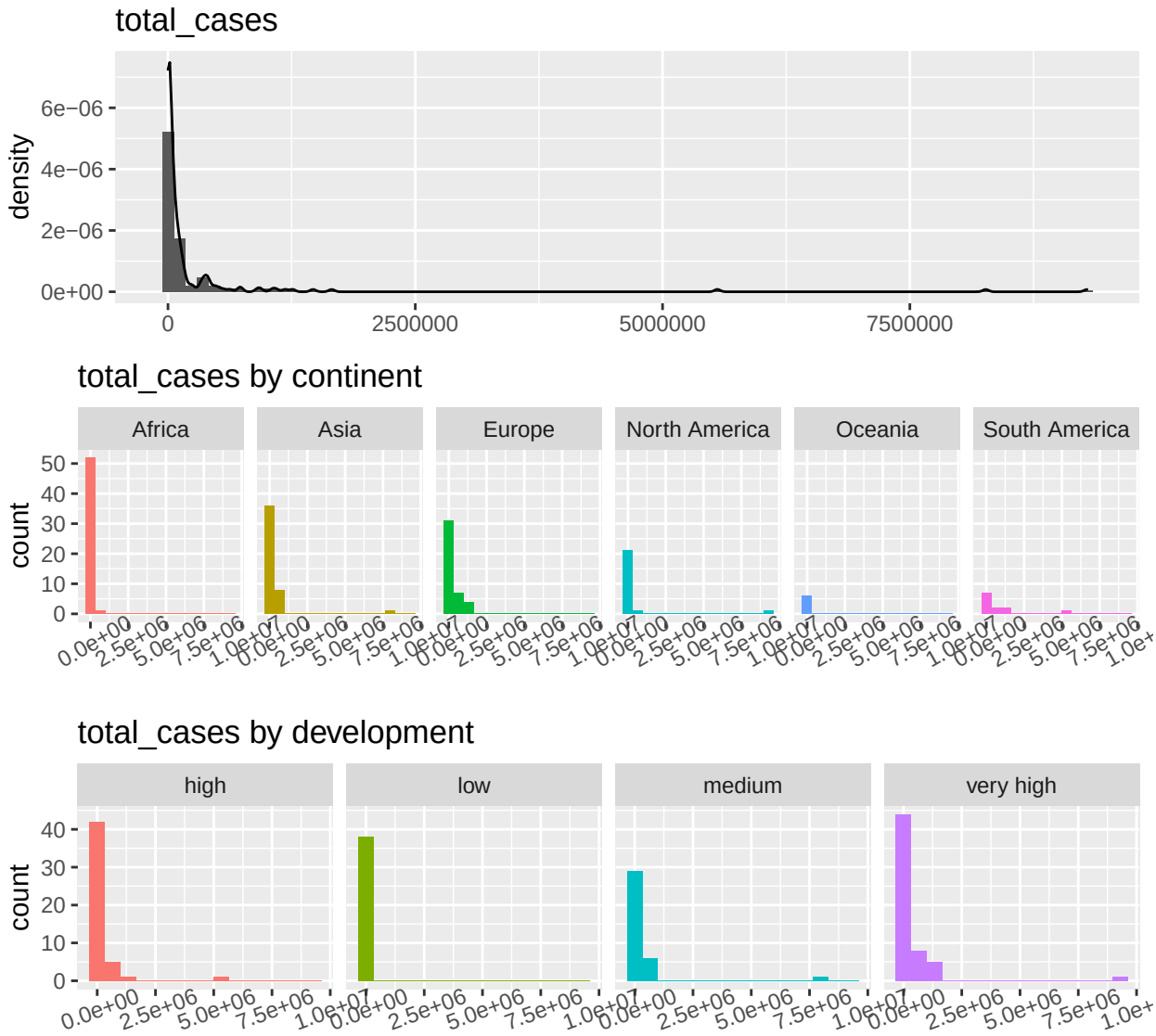

```
# By continent
data$location[data$total_cases==max(data$total_cases)]
#> [1] "United States"
data$location[data$total_cases==max(data$total_cases[data$continent=='South America'])]
#> [1] "Brazil"
data$location[data$total_cases==max(data$total_cases[data$continent=='Asia'])]
#> [1] "India"

# By development
data$location[data$total_cases==max(data$total_cases[data$development=='very high'])]
#> [1] "United States"
data$location[data$total_cases==max(data$total_cases[data$development=='high'])]
#> [1] "Brazil"
data$location[data$total_cases==max(data$total_cases[data$development=='medium'])]
```

```
#> [1] "India"
```

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

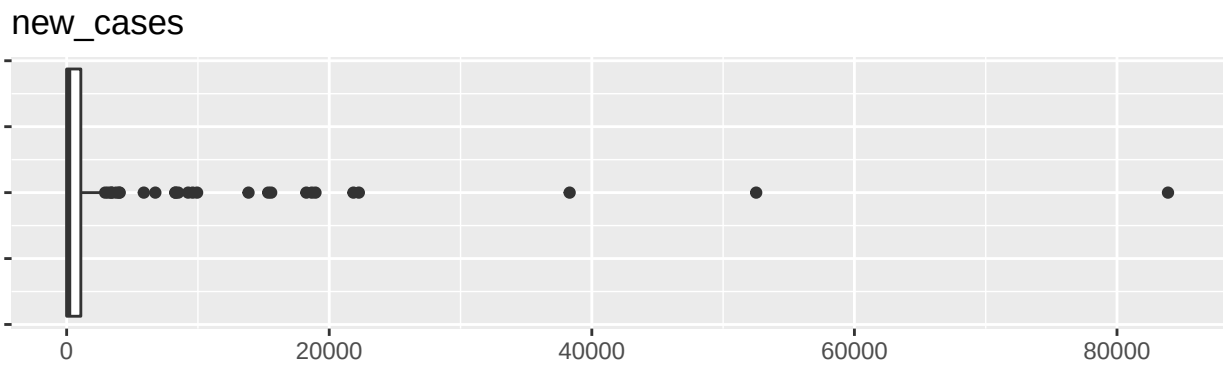
```
plots(dataset=data, col='total_cases',type='hist', density=FALSE, bins = c(80,15,15),xtick_angles=c(0,3
```

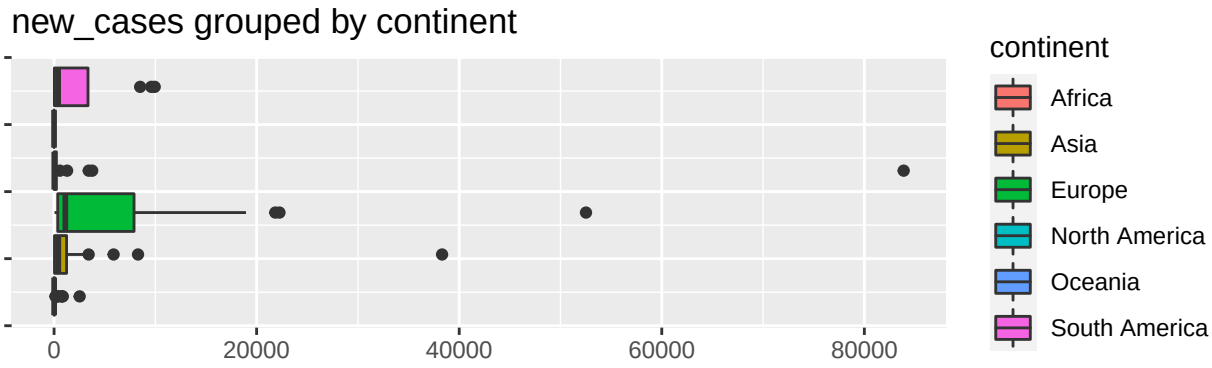


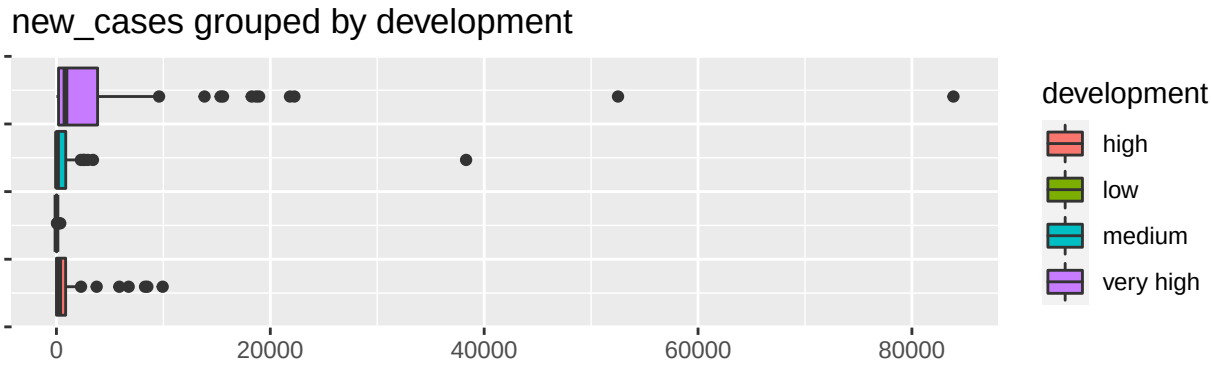
Observing this graph, we can confirm that the distribution is very right-skewed. Only Africa, Europe and Oceania don't have outliers. But it is probably because we don't have the dataset updated yet (we have the data-set updated on 3rd of November, 2020). About the development of different countries,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







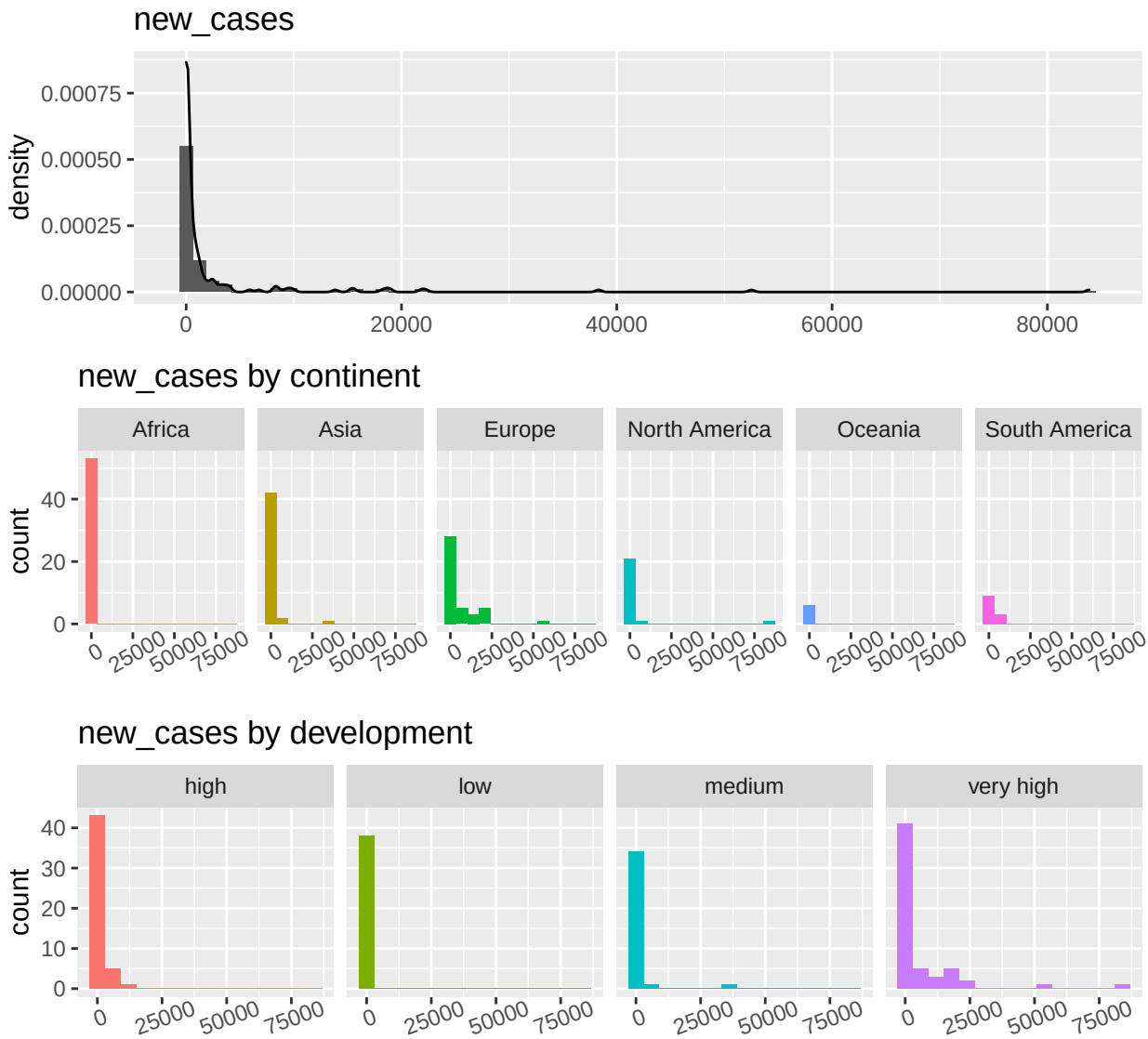

```
# By continent
data$location[data$new_cases==max(data$new_cases)]
#> [1] "United States"
data$location[data$new_cases==max(data$new_cases[data$continent=='Europe'])]
#> [1] "France"
data$location[data$new_cases==max(data$new_cases[data$continent=='Asia'])]
#> [1] "India"

# By development
data$location[data$new_cases==max(data$new_cases[data$development=='medium'])]
#> [1] "India"
head(data[order(data$new_cases,decreasing = TRUE),])
#>       X     continent       location total_cases new_cases new_cases_smoothed
#> 173 172 North America  United States     9291245     83883          83817.286
#> 59   58        Europe         France     1466433     52518          43022.143
```

```
#> 77   76          Asia          India     8267623     38310          45884.857
#> 83   82        Europe          Italy      731588     22253          26971.286
#> 31   30        Europe    Switzerland      175570     21842           7841.429
#> 61   60        Europe United Kingdom     1053864     18950          22739.143
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 173       231551        555              830.857                28070.002
#> 59         37435        416              345.286                22465.974
#> 77        123097        490              513.571                 5991.012
#> 83         39059        233              225.714                12099.998
#> 31          2128         93               30.714                20286.280
#> 61         46853        136              265.000                15524.025
#>     new_cases_per_million new_cases_smoothed_per_million
#> 173               253.421                        253.222
#> 59                804.584                        659.106
#> 77                 27.761                         33.250
#> 83                368.050                        446.088
#> 31               2523.739                        906.040
#> 61                279.144                        334.961
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 173                  699.544                  1.677            62.50  331002647
#> 59                   573.510                  6.373            78.70   65273512
#> 77                    89.200                  0.355            61.57 1380004385
#> 83                   646.011                  3.854            66.67   60461828
#> 31                   245.880                 10.746            50.93    8654618
#> 61                   690.172                  2.003            75.00   67886004
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 173             35.608       38.3        15.413         9.732       54225.446
#> 59             122.578       42.0        19.718        13.079       38605.671
#> 77             450.419       28.2         5.989         3.414        6426.674
#> 83             205.859       47.9        23.021        16.240       35220.084
#> 31             214.243       43.1        18.436        12.644       57410.166
#> 61             272.898       40.8        18.517        12.527       39753.244
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 173             1.2               151.089               10.79
#> 59               NA                86.060                4.77
#> 77             21.2               282.280               10.39
#> 83              2.0               113.151                4.78
#> 31               NA                99.739                5.59
#> 61              0.2               122.137                4.28
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 173                       2.77           78.86                   0.924
#> 59                        5.98           82.66                   0.901
#> 77                        0.53           69.66                   0.640
#> 83                        3.18           83.51                   0.880
#> 31                        4.53           83.78                   0.944
#> 61                        2.54           81.32                   0.922
#>     development
#> 173   very high
#> 59    very high
#> 77       medium
#> 83    very high
#> 31    very high
#> 61    very high
```

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

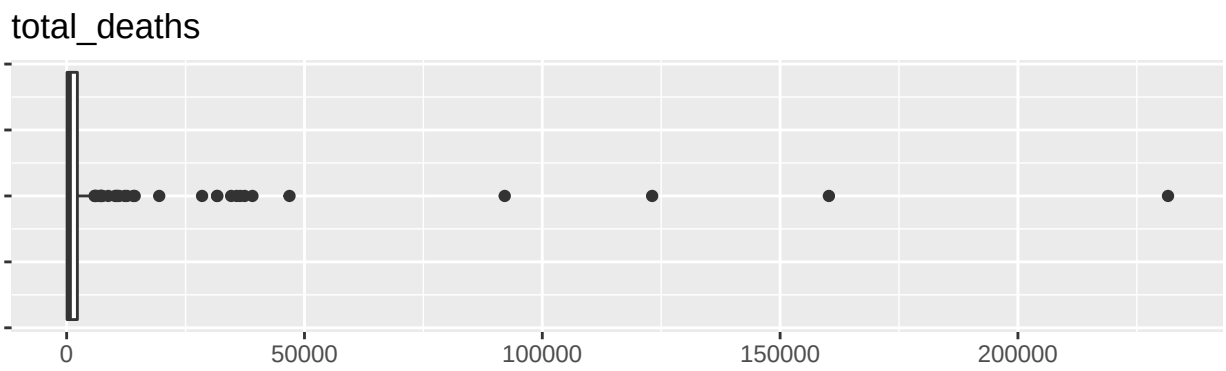
```
plots(dataset=data, col='new_cases',type='hist', density=FALSE, bins = c(70,13,15),xtick_angles=c(0,30,
```

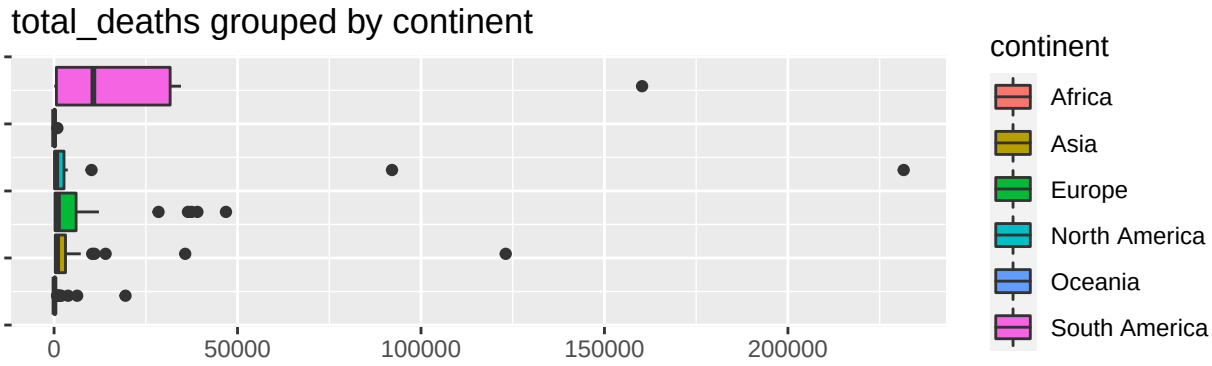


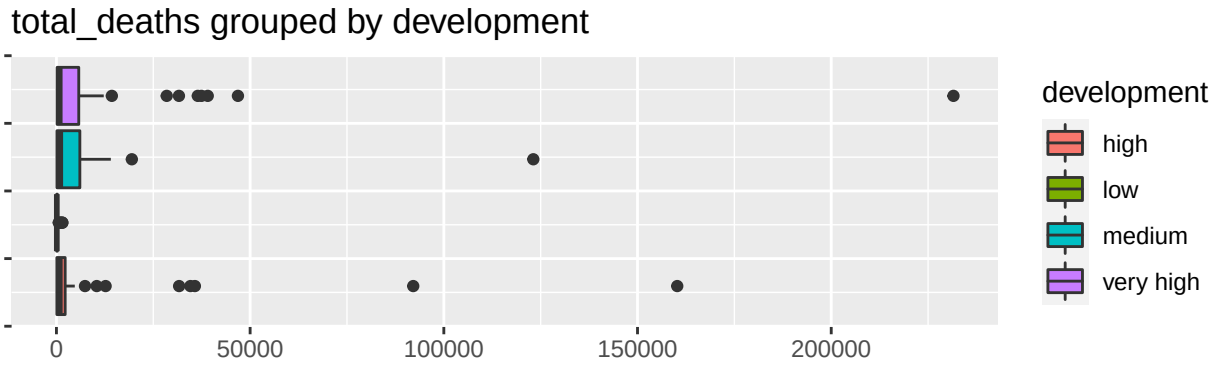
Observing this graph, we can confirm that the distribution is very right-skewed with some outliers (especially the US, France and India). In the histograms of new cases by continent, we can see that Europe has the most dispersed distribution comparing to other continents, which means that the countries 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')
```







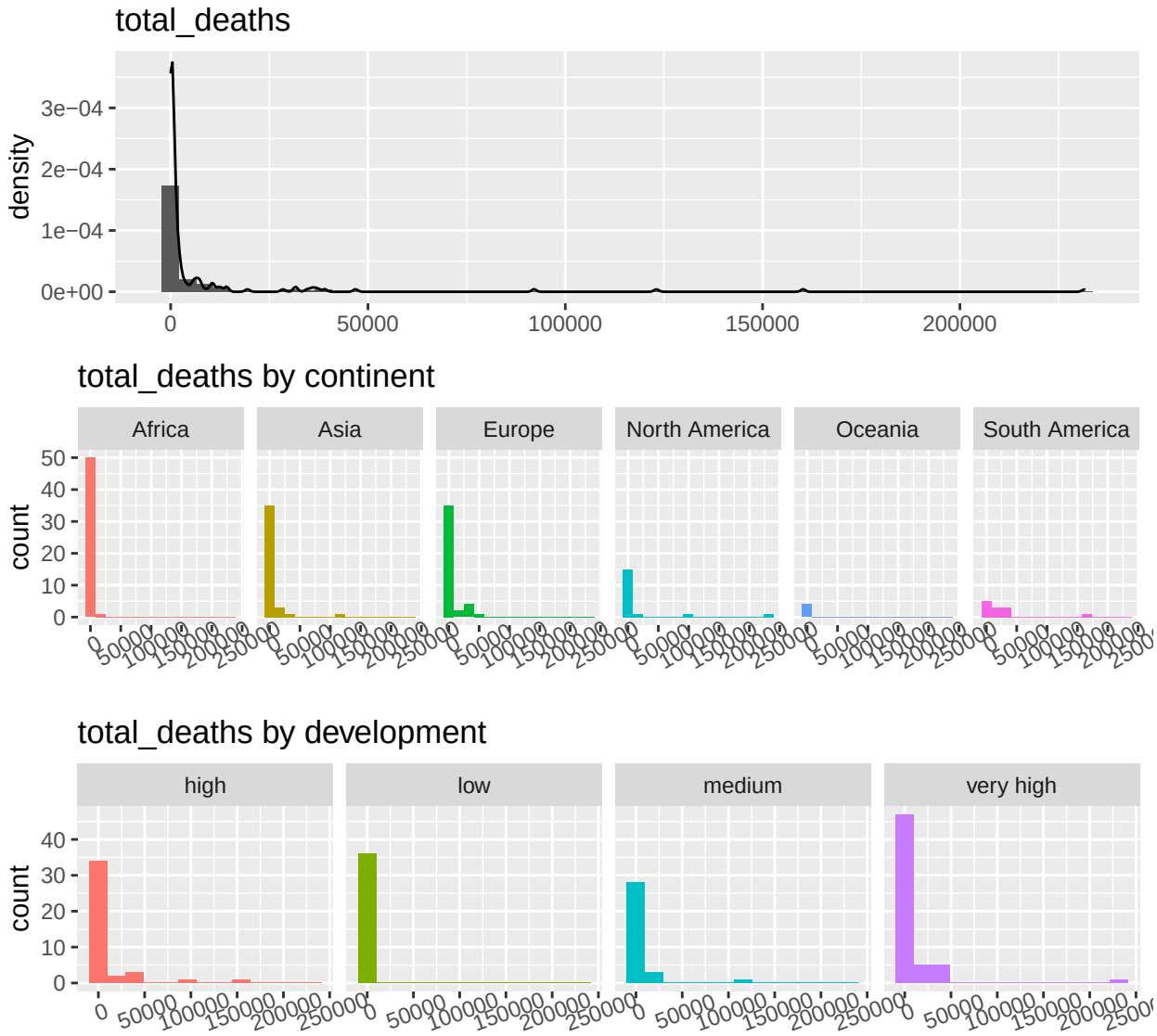

```
head(data[order(data$total_deaths,decreasing = TRUE),])
#>       X     continent       location total_cases new_cases new_cases_smoothed
#> 173 172 North America  United States     9291245     83883          83817.286
#> 24   23 South America         Brazil     5554206      8501          20621.714
#> 77   76          Asia          India     8267623     38310          45884.857
#> 109 108 North America         Mexico      933155      3763           5404.143
#> 61   60        Europe United Kingdom     1053864     18950          22739.143
#> 83   82        Europe          Italy      731588     22253          26971.286
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 173       231551        555             830.857                28070.002
#> 24        160253        179             408.000                26130.135
#> 77        123097        490             513.571                 5991.012
#> 109        92100        205             418.429                 7237.533
#> 61         46853        136             265.000                15524.025
```

```
#> 83         39059        233              225.714               12099.998
#>     new_cases_per_million new_cases_smoothed_per_million
#> 173               253.421                        253.222
#> 24                 39.994                         97.016
#> 77                 27.761                         33.250
#> 109                29.186                         41.914
#> 61                279.144                        334.961
#> 83                368.050                        446.088
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 173                  699.544                 1.677            62.50  331002647
#> 24                   753.921                 0.842            57.87  212559409
#> 77                    89.200                 0.355            61.57 1380004385
#> 109                  714.326                 1.590            71.76  128932753
#> 61                   690.172                 2.003            75.00   67886004
#> 83                   646.011                 3.854            66.67   60461828
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 173             35.608       38.3         15.413          9.732      54225.446
#> 24              25.040       33.5          8.552          5.060      14103.452
#> 77             450.419       28.2          5.989          3.414       6426.674
#> 109             66.444       29.3          6.857          4.321      17336.469
#> 61             272.898       40.8         18.517         12.527      39753.244
#> 83             205.859       47.9         23.021         16.240      35220.084
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 173             1.2                151.089               10.79
#> 24              3.4                177.961                8.11
#> 77             21.2                282.280               10.39
#> 109             2.5                152.783               13.06
#> 61              0.2                122.137                4.28
#> 83              2.0                113.151                4.78
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 173                       2.77           78.86                   0.924
#> 24                        2.20           75.88                   0.759
#> 77                        0.53           69.66                   0.640
#> 109                       1.38           75.05                   0.774
#> 61                        2.54           81.32                   0.922
#> 83                        3.18           83.51                   0.880
#>     development
#> 173   very high
#> 24         high
#> 77       medium
#> 109        high
#> 61    very high
#> 83    very high
```

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

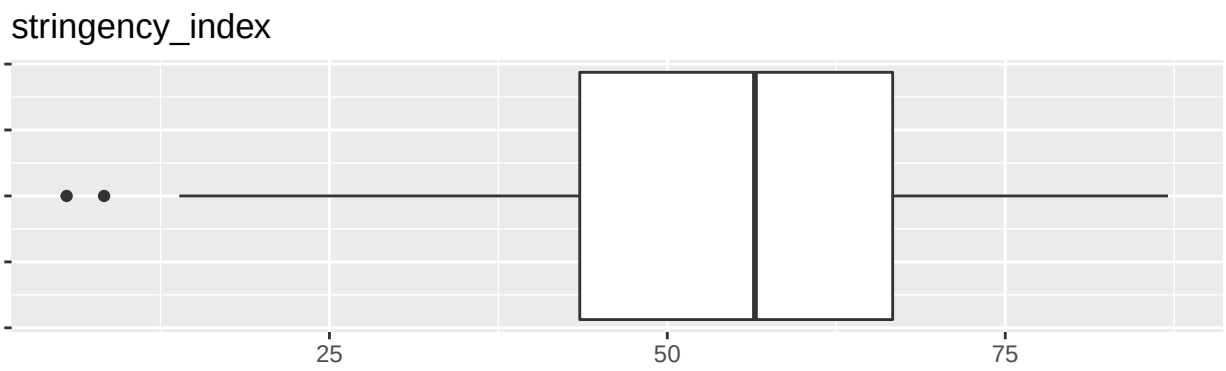
```
plots(dataset=data, col='total_deaths',type='hist', density=FALSE, bins = c(55,15,13),xtick_angles=c(0,
```

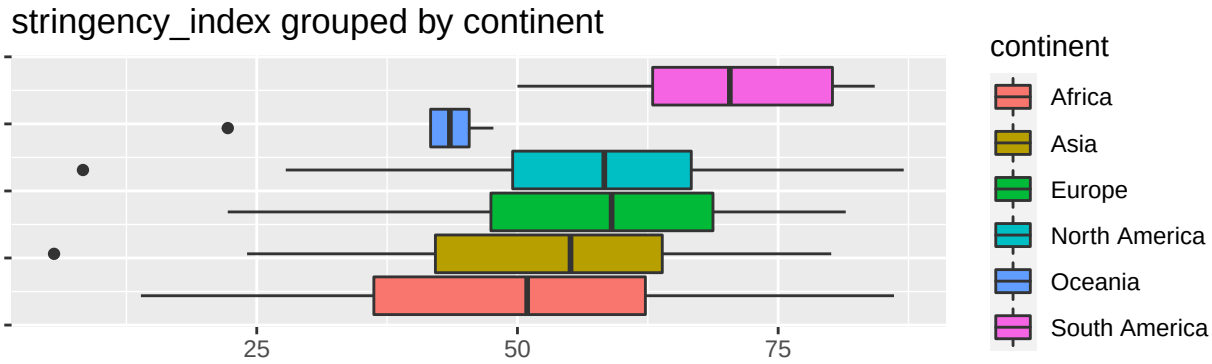


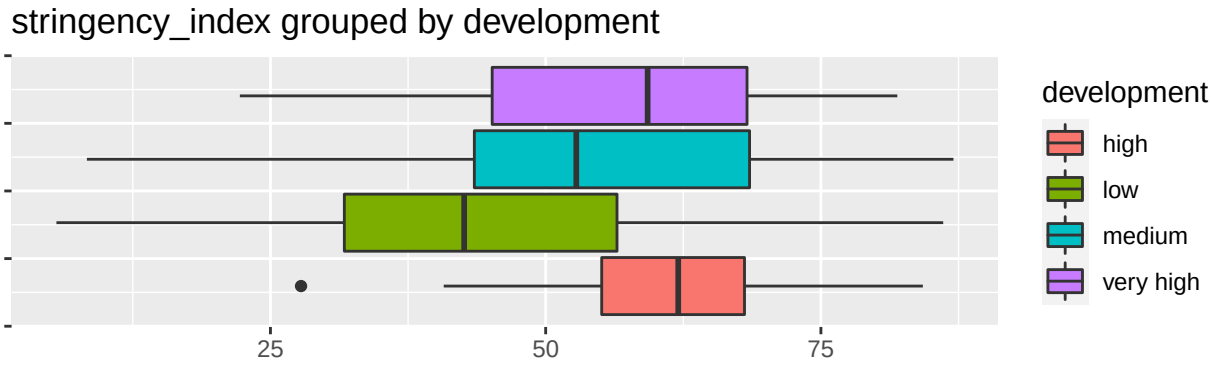
Again, the distribution is very right-skewed with some outliers (the US, Brazil and India). In the histograms of new cases by continent and histograms by development, the distributions are also very right-skewed, some of them have outliers. We still can't group the coutries in terms of how they have 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')
```








```
head(data[order(data$stringency_index),])
#>       X     continent    location total_cases new_cases new_cases_smoothed
#> 1     0          Asia Afghanistan       41728        95             99.429
#> 125 124 North America   Nicaragua        5514         0             11.429
#> 118 117        Africa  Mauritania        7725         8              8.857
#> 169 168        Africa    Tanzania         509         0              0.000
#> 12   11        Africa     Burundi         589         0              4.571
#> 21   20        Europe     Belarus      100400       941            956.143
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 1           1544          3                3.143                1071.918
#> 125          156          0                0.143                 832.358
#> 118          163          0                0.000                1661.412
#> 169           21          0                0.000                   8.521
#> 12             1          0                0.000                  49.534
```

```
#> 21           989          4               4.000               10625.102
#>     new_cases_per_million new_cases_smoothed_per_million
#> 1                   2.440                          2.554
#> 125                 0.000                          1.725
#> 118                 1.721                          1.905
#> 169                 0.000                          0.000
#> 12                  0.000                          0.384
#> 21                 99.584                        101.186
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 1                     39.663                  0.077             5.56   38928341
#> 125                   23.549                  0.000             8.33    6624554
#> 118                   35.056                  0.000            13.89    4649660
#> 169                    0.352                  0.000            13.89   59734213
#> 12                     0.084                  0.000            14.81   11890781
#> 21                   104.664                  0.423            22.22    9449321
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 1               54.422       18.6         2.581         1.337        1803.987
#> 125             51.667       27.3         5.445         3.519        5321.444
#> 118              4.289       20.3         3.138         1.792        3597.633
#> 169             64.699       17.7         3.108         1.874        2683.304
#> 12             423.062       17.5         2.562         1.504         702.225
#> 21              46.858       40.3        14.799         9.788       17167.967
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 1                NA               597.029                9.59
#> 125             3.2               137.016               11.47
#> 118             6.0               232.347                2.42
#> 169            49.1               217.288                5.75
#> 12             71.7               293.068                6.05
#> 21               NA               443.129                5.18
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 1                          0.5           64.83                   0.498
#> 125                        0.9           74.48                   0.658
#> 118                         NA           64.92                   0.520
#> 169                        0.7           65.46                   0.538
#> 12                         0.8           61.58                   0.417
#> 21                        11.0           74.79                   0.808
#>     development
#> 1           low
#> 125      medium
#> 118         low
#> 169         low
#> 12          low
#> 21    very high
```

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).

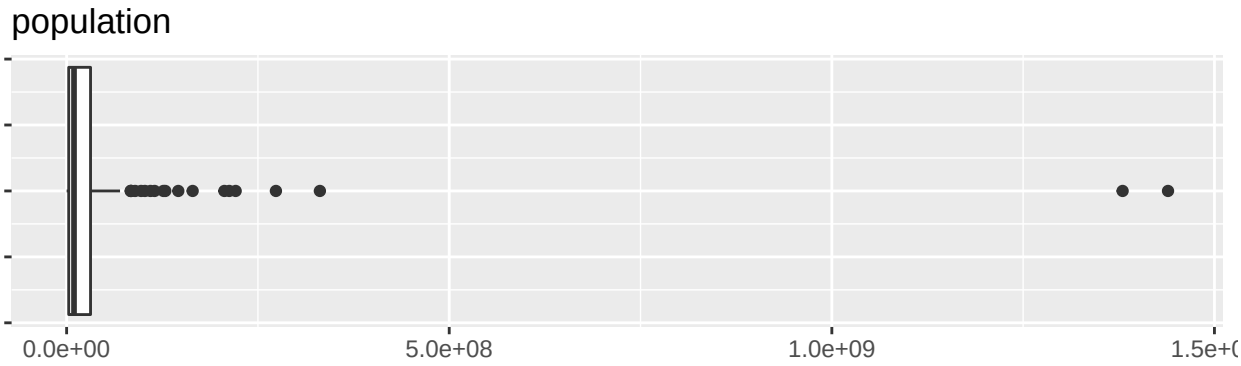
Histogram and kernel density for stringency index

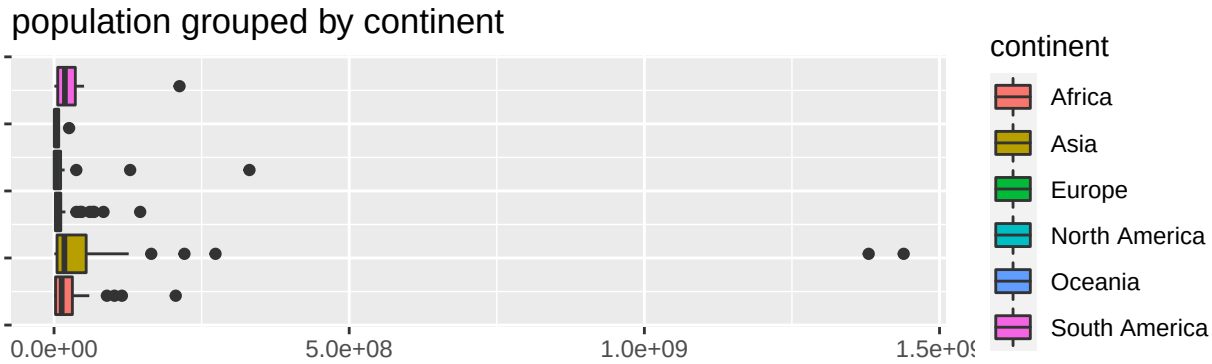
```
plots(dataset=data, col='stringency_index',type='hist', density=TRUE,xtick_angles=c(0,30,20))
```

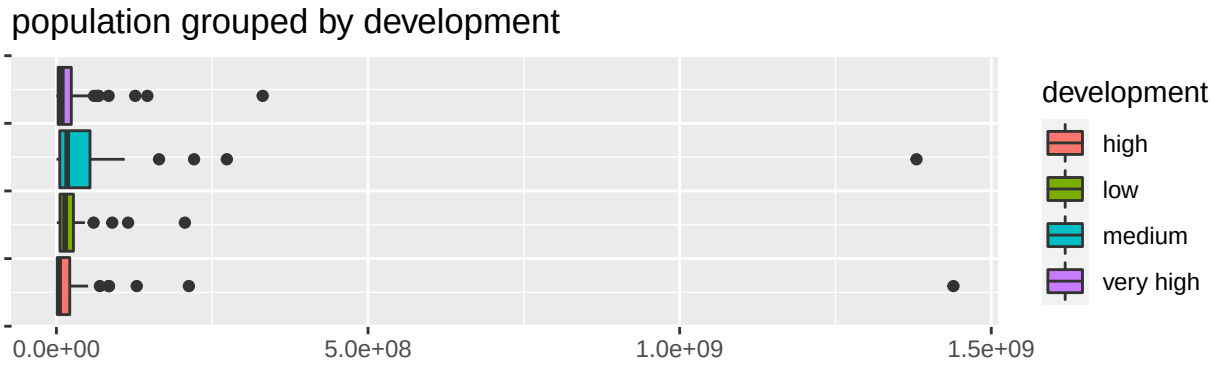
The distribution is quite symmetric distributed for different continents. Except for Oceania, there are some countries have really low stringency index; and South America, the distribution of this variable is quite left skewed. We can probably distinguish the countries with low HDI from others, these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

Boxplots for population

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







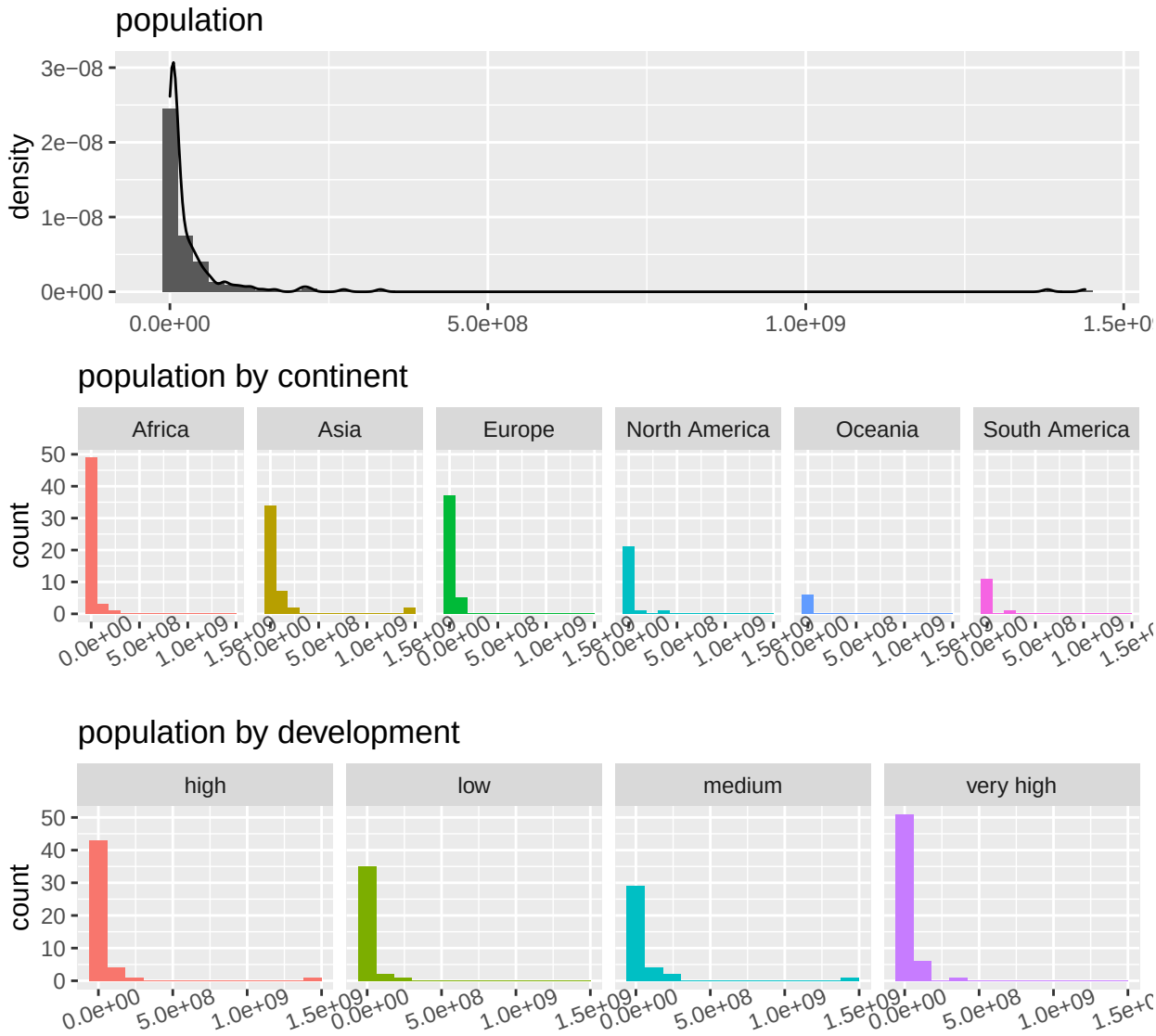

```
head(data[order(data$population,decreasing=TRUE), ])
#>       X     continent      location total_cases new_cases new_cases_smoothed
#> 33   32          Asia         China       91396        49             31.571
#> 77   76          Asia         India     8267623     38310          45884.857
#> 173 172 North America United States     9291245     83883          83817.286
#> 76   75          Asia     Indonesia      415402      2618           3209.714
#> 131 130          Asia      Pakistan      336260      1167            983.571
#> 24   23 South America        Brazil     5554206      8501          20621.714
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 33          4739          0                0.000                  63.499
#> 77        123097        490              513.571                5991.012
#> 173       231551        555              830.857               28070.002
#> 76         14044        101               90.429                1518.706
#> 131         6849         14               14.857                1522.280
```

```
#> 24        160253        179             408.000               26130.135
#>     new_cases_per_million new_cases_smoothed_per_million
#> 33                  0.034                         0.022
#> 77                 27.761                        33.250
#> 173               253.421                       253.222
#> 76                  9.571                        11.735
#> 131                 5.283                         4.453
#> 24                 39.994                        97.016
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 33                     3.293                  0.000            63.43 1439323774
#> 77                    89.200                  0.355            61.57 1380004385
#> 173                  699.544                  1.677            62.50  331002647
#> 76                    51.345                  0.369            50.46  273523621
#> 131                   31.006                  0.063            53.24  220892331
#> 24                   753.921                  0.842            57.87  212559409
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 33             147.674       38.7        10.641         5.929       15308.712
#> 77             450.419       28.2         5.989         3.414        6426.674
#> 173             35.608       38.3        15.413         9.732       54225.446
#> 76             145.725       29.3         5.319         3.053       11188.744
#> 131            255.573       23.5         4.495         2.780        5034.708
#> 24              25.040       33.5         8.552         5.060       14103.452
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 33              0.7               261.899                9.74
#> 77             21.2               282.280               10.39
#> 173             1.2               151.089               10.79
#> 76              5.7               342.864                6.32
#> 131             4.0               423.031                8.35
#> 24              3.4               177.961                8.11
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 33                        4.34           76.91                   0.752
#> 77                        0.53           69.66                   0.640
#> 173                       2.77           78.86                   0.924
#> 76                        1.04           71.72                   0.694
#> 131                       0.60           67.27                   0.562
#> 24                        2.20           75.88                   0.759
#>     development
#> 33         high
#> 77       medium
#> 173   very high
#> 76       medium
#> 131      medium
#> 24         high
```

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

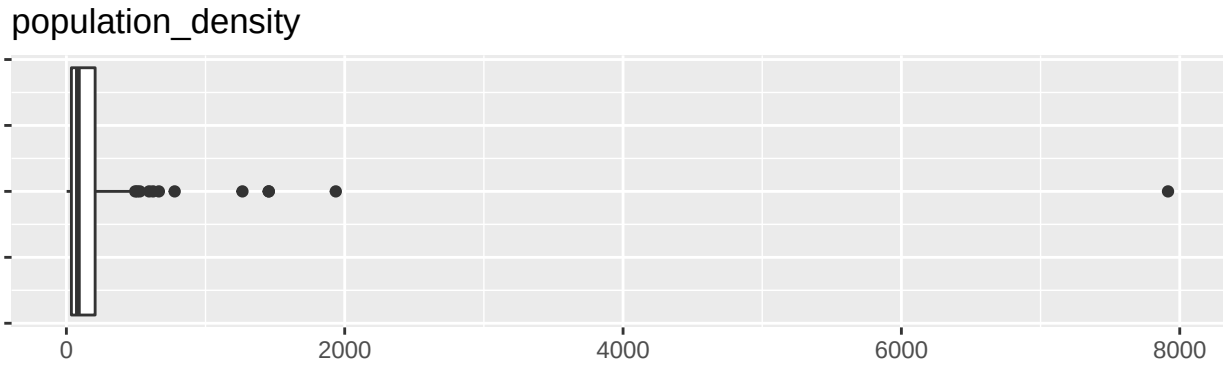
```
plots(dataset=data, col='population',type='hist', density=FALSE, bins = c(60,13,13),xtick_angles=c(0,30
```

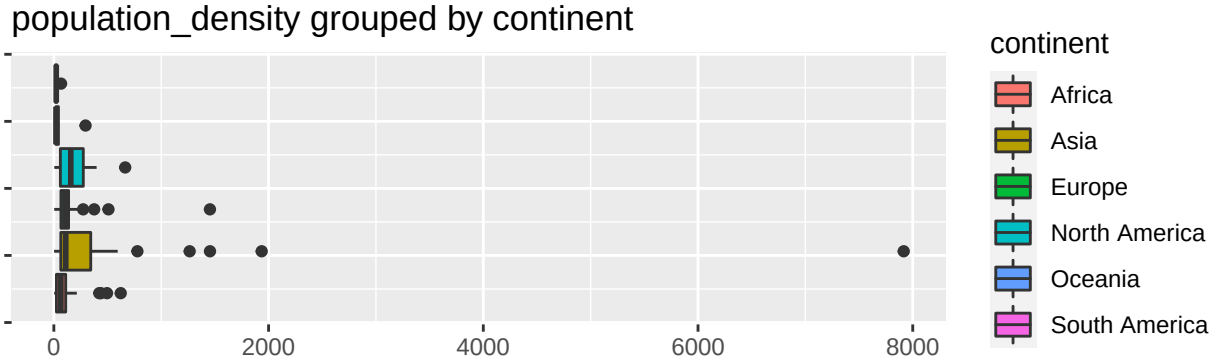


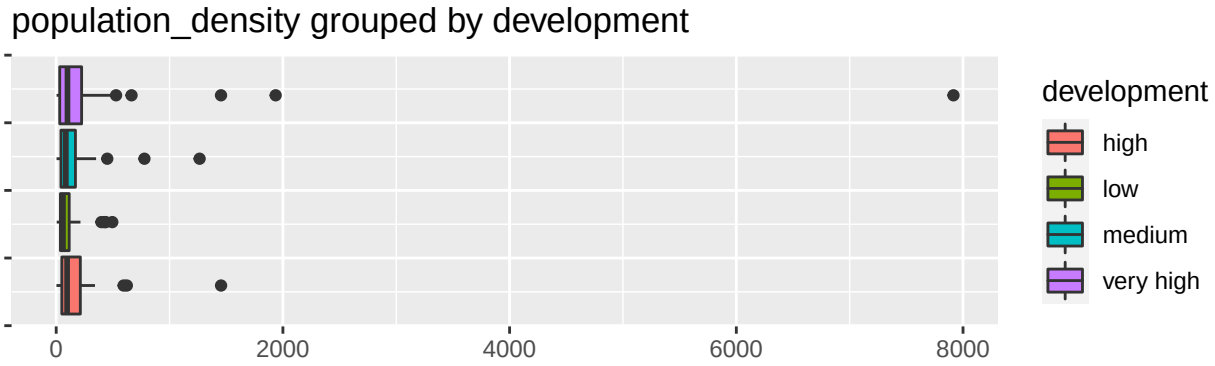
The distribution is very right-skewed, the population of each country is very different from others, but the variable population does not provide any information of whether the country has high HDI or not.

Boxplots for population density

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







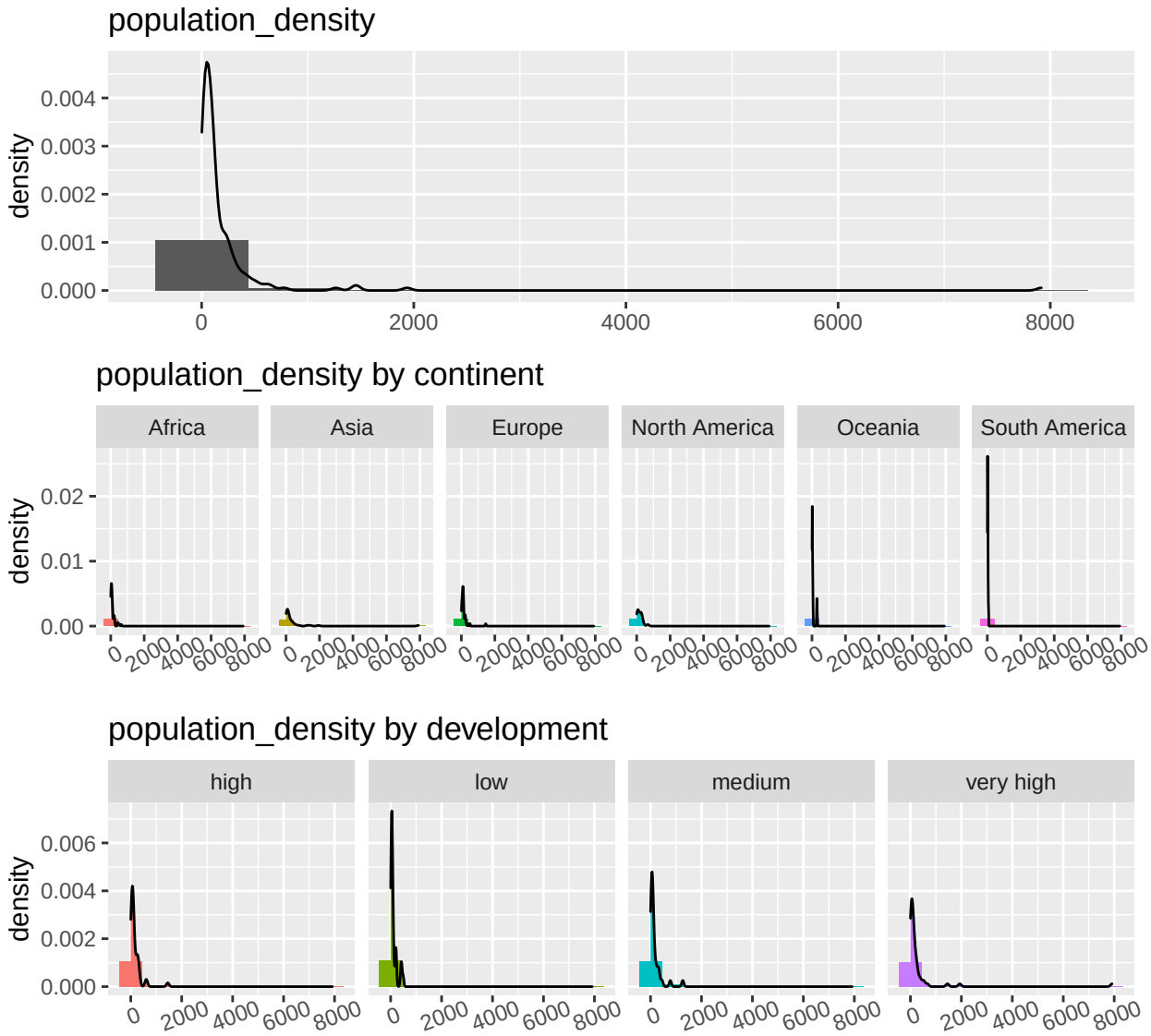

```r
head(data[order(data$population_density,decreasing=TRUE), ])
#>       X continent   location total_cases new_cases new_cases_smoothed
#> 147 146      Asia  Singapore       58020         1              6.714
#> 18   17      Asia    Bahrain       82133       210            228.571
#> 108 107      Asia   Maldives       11737        36             29.286
#> 113 112    Europe      Malta        6400       218            117.571
#> 16   15      Asia Bangladesh      410988      1736           1533.857
#> 139 138      Asia  Palestine       66551       749            643.571
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 147           28          0                0.000                9917.367
#> 18           323          2                1.000               48268.583
#> 108           38          0                0.143               21713.391
#> 113           64          0                2.000               14494.756
#> 16          5966         25               21.143                2495.534
#> 139          565          4                6.714               13045.594
#>     new_cases_per_million new_cases_smoothed_per_million
#> 147                 0.171                         1.148
#> 18                123.414                       134.329
```

```
#> 108                66.600                          54.178
#> 113               493.728                         266.276
#> 16                 10.541                           9.314
#> 139               146.822                         126.155
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 147                    4.786                   0.000            52.78    5850343
#> 18                   189.823                   1.175            58.33    1701583
#> 108                   70.300                   0.000               NA     540542
#> 113                  144.948                   0.000               NA     441539
#> 16                    36.226                   0.152            80.09  164689383
#> 139                  110.754                   0.784            40.74    5101416
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 147          7915.731       42.4        12.922         7.049      85535.383
#> 18           1935.907       32.4         2.372         1.387      43290.705
#> 108          1454.433       30.6         4.120         2.875      15183.616
#> 113          1454.037       42.4        19.426        11.324      36513.323
#> 16           1265.036       27.5         5.098         3.262       3523.984
#> 139           778.202       20.4         3.043         1.726       4449.898
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 147              NA                92.243               10.99
#> 18               NA               151.689               16.52
#> 108              NA               164.905                9.19
#> 113             0.2               168.711                8.83
#> 16             14.8               298.003                8.38
#> 139             1.0               265.910               10.59
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 147                      2.400           83.62                   0.932
#> 18                       2.000           77.29                   0.846
#> 108                         NA           78.92                   0.717
#> 113                      4.485           82.53                   0.878
#> 16                       0.800           72.59                   0.608
#> 139                         NA           74.05                   0.686
#>     development
#> 147   very high
#> 18    very high
#> 108       high
#> 113   very high
#> 16       medium
#> 139      medium
```

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

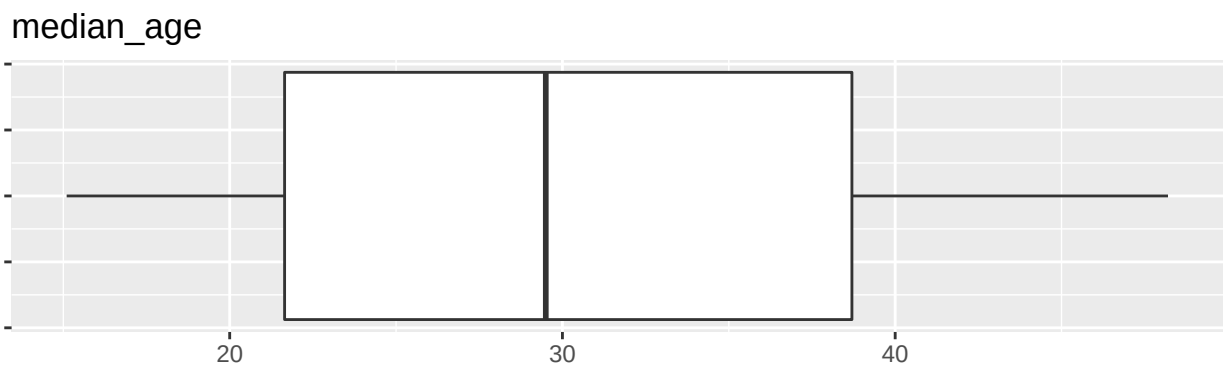
```
plots(dataset=data, col='population_density',type='hist', density=TRUE, xtick_angles=c(0,30,20))
```

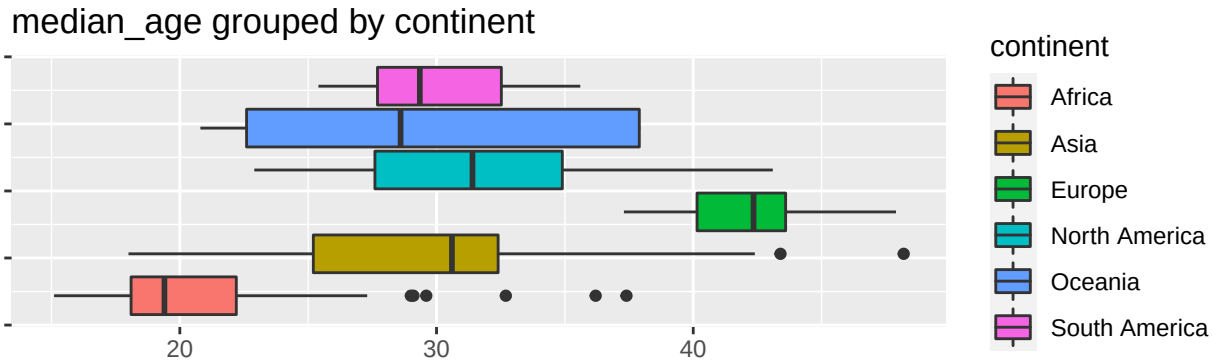


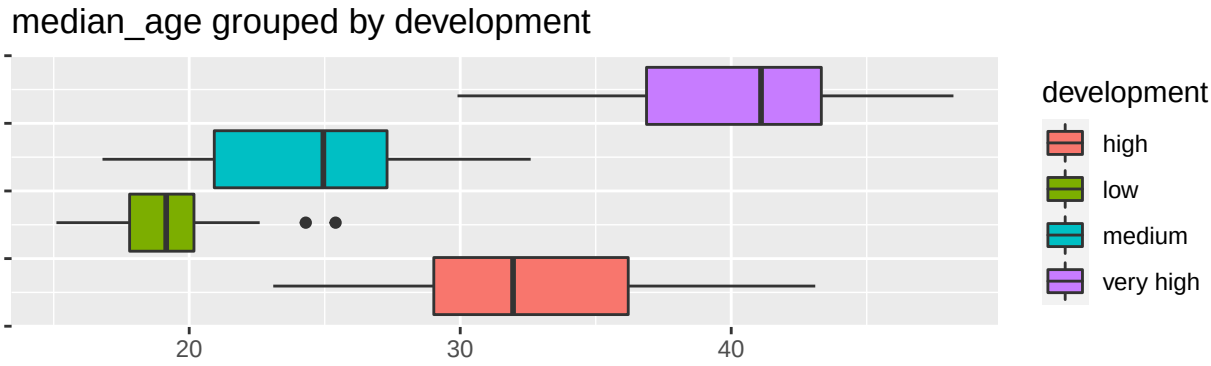
The distribution is very right-skewed, the population density of each country is very different from others. And the variable does not provide any information of whether the country has high HDI or not.

Boxplots for median age

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







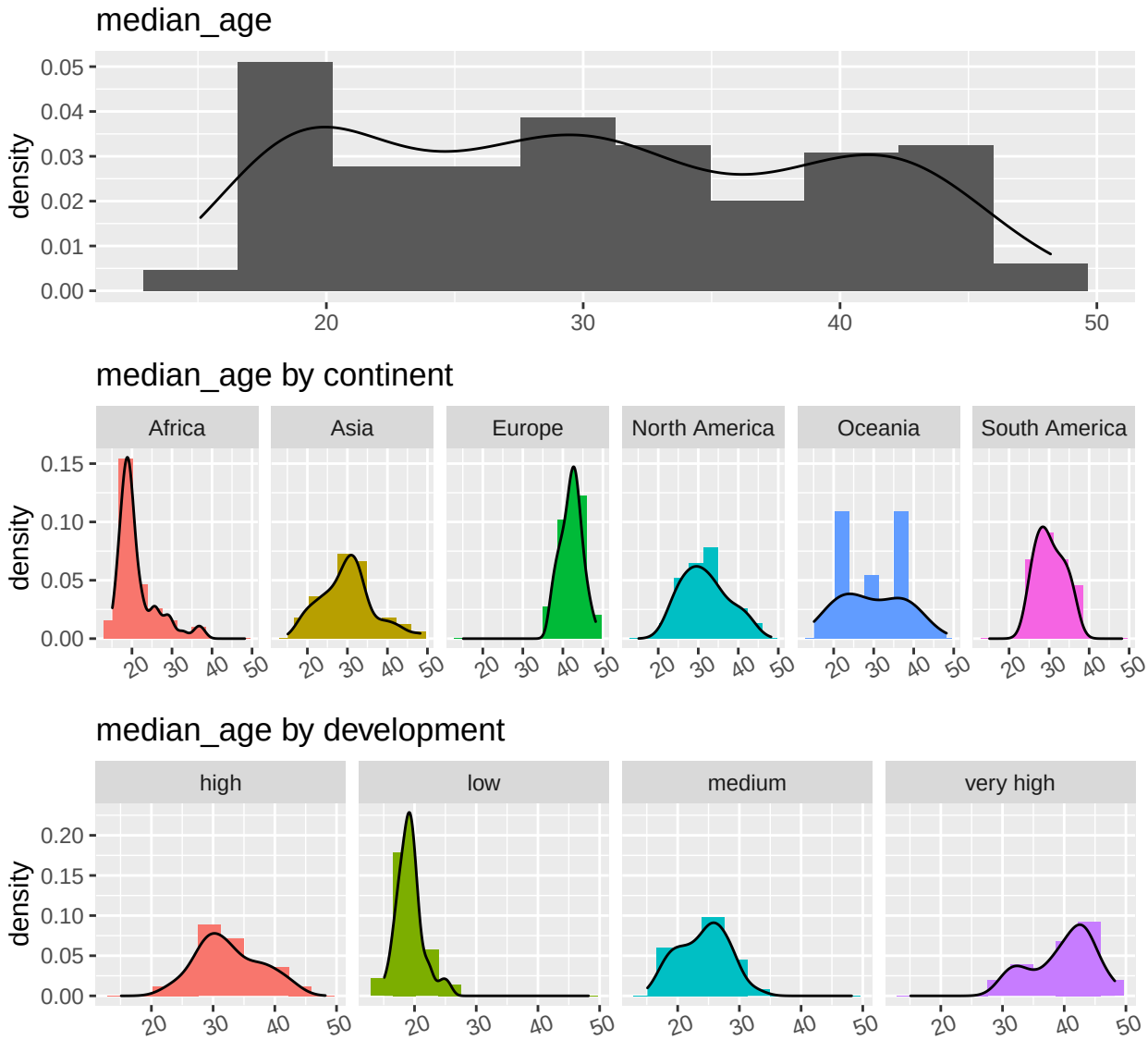

```
head(data[order(data$median_age,decreasing=TRUE), ])
#>       X continent location total_cases new_cases new_cases_smoothed
#> 86   85      Asia    Japan      102281       468            683.286
#> 83   82    Europe    Italy      731588     22253          26971.286
#> 45   44    Europe  Germany      560379     15352          15872.000
#> 137 136    Europe Portugal      146847      2506           3673.429
#> 54   53    Europe    Spain     1259366     18669          20375.429
#> 68   67    Europe   Greece       42080      1151           1512.000
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 86          1780          6                7.857                808.696
#> 83         39059        233              225.714              12099.998
#> 45         10661        131               80.429               6688.382
#> 137         2590         46               35.286              14401.414
#> 54         36495        238              171.000              26935.554
```

```
#> 68           642          7               8.714                4037.204
#>     new_cases_per_million new_cases_smoothed_per_million
#> 86                  3.700                         5.402
#> 83                368.050                       446.088
#> 45                183.233                       189.440
#> 137               245.766                       360.256
#> 54                399.296                       435.793
#> 68                110.428                       145.063
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 86                     14.074                 0.047            38.89  126476458
#> 83                    646.011                 3.854            66.67   60461828
#> 45                    127.244                 1.564            59.26   83783945
#> 137                   254.004                 4.511            62.96   10196707
#> 54                    780.562                 5.090            71.30   46754783
#> 68                     61.594                 0.672            63.43   10423056
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 86             347.778       48.2        27.049        18.493       39002.22
#> 83             205.859       47.9        23.021        16.240       35220.08
#> 45             237.016       46.6        21.453        15.957       45229.25
#> 137            112.371       46.2        21.502        14.924       27936.90
#> 54              93.105       45.5        19.436        13.799       34272.36
#> 68              83.479       45.3        20.396        14.524       24574.38
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 86               NA                79.370                5.72
#> 83              2.0               113.151                4.78
#> 45               NA               156.139                8.31
#> 137             0.5               127.842                9.85
#> 54              1.0                99.403                7.17
#> 68              1.5               175.695                4.55
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 86                       13.05           84.63                   0.909
#> 83                        3.18           83.51                   0.880
#> 45                        8.00           81.33                   0.936
#> 137                       3.39           82.05                   0.847
#> 54                        2.97           83.56                   0.891
#> 68                        4.21           82.24                   0.870
#>     development
#> 86    very high
#> 83    very high
#> 45    very high
#> 137   very high
#> 54    very high
#> 68    very high
```

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

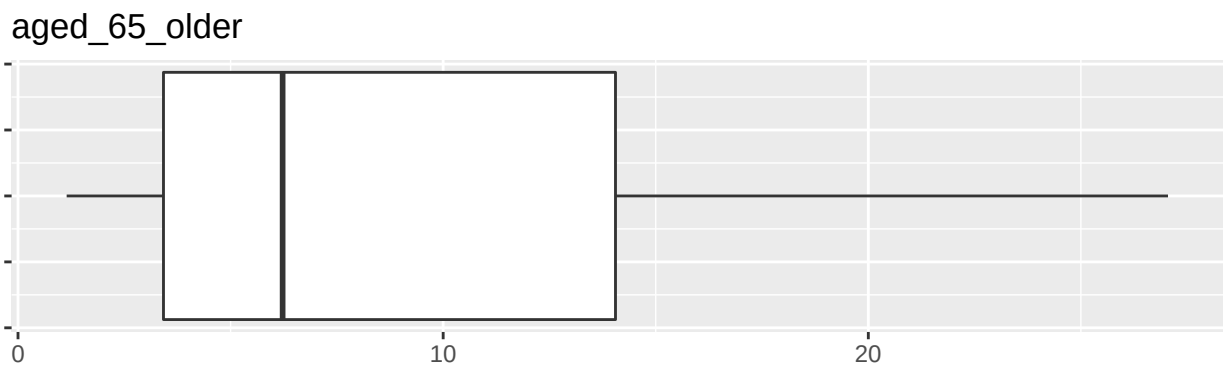
```
plots(dataset=data, col='median_age',type='hist', density=TRUE, xtick_angles=c(0,30,20))
```

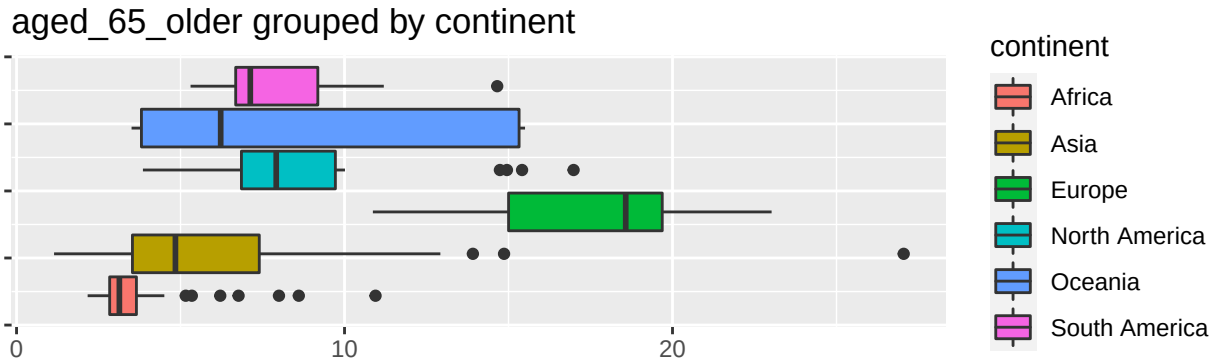


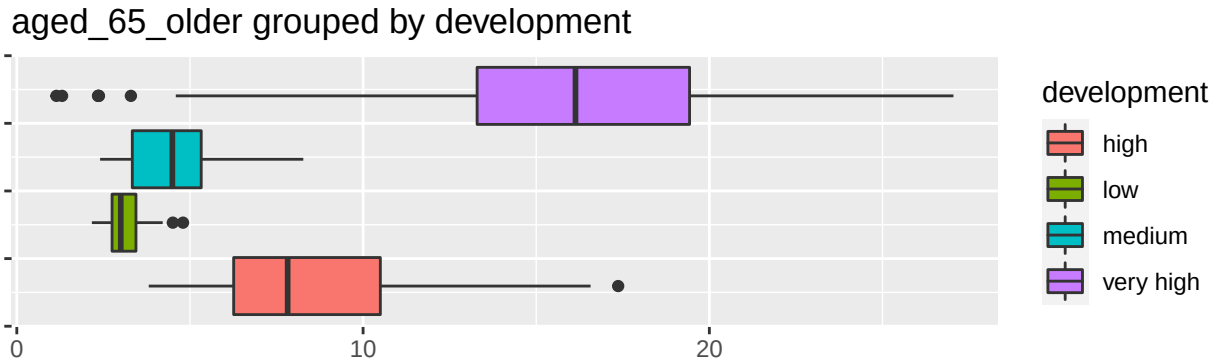
The distributions of median age for different continent are very different. The distribution of Africa is right skewed while others are symmetric. The distributions of Asia, North America, Oceania and South America are more flat (platykurtic), and the distributions of Africa and Europe are more concentrated (leptokurtic).

Boxplots for the percentage of population aged 65 or older

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







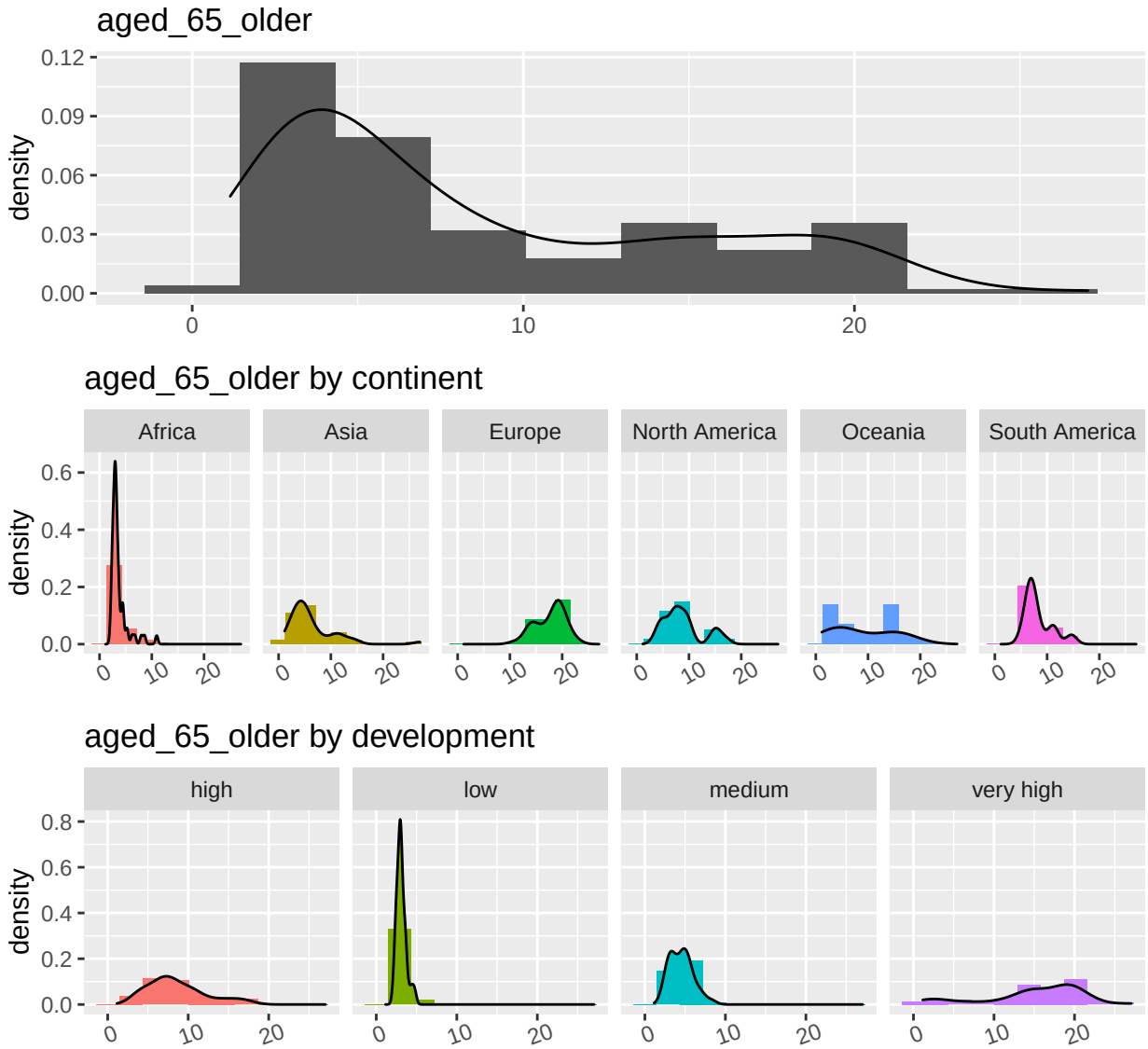

```
head(data[order(data$aged_65_older,decreasing=TRUE), ])
#>       X continent location total_cases new_cases new_cases_smoothed
#> 86   85      Asia    Japan      102281       468            683.286
#> 83   82    Europe    Italy      731588     22253          26971.286
#> 137 136    Europe Portugal      146847      2506           3673.429
#> 45   44    Europe  Germany      560379     15352          15872.000
#> 57   56    Europe  Finland       16400       109            204.286
#> 17   16    Europe Bulgaria       56496      2427           2337.714
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 86          1780          6                7.857                808.696
#> 83         39059        233              225.714              12099.998
#> 137         2590         46               35.286              14401.414
#> 45         10661        131               80.429               6688.382
#> 57           359          1                0.714               2959.905
```

```
#> 17          1349         51              30.429                8130.740
#>     new_cases_per_million new_cases_smoothed_per_million
#> 86                  3.700                         5.402
#> 83                368.050                       446.088
#> 137               245.766                       360.256
#> 45                183.233                       189.440
#> 57                 19.673                        36.870
#> 17                349.287                       336.437
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 86                     14.074                 0.047            38.89  126476458
#> 83                    646.011                 3.854            66.67   60461828
#> 137                   254.004                 4.511            62.96   10196707
#> 45                    127.244                 1.564            59.26   83783945
#> 57                     64.793                 0.180            40.74    5540718
#> 17                    194.144                 7.340            48.15    6948445
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 86             347.778       48.2        27.049        18.493       39002.22
#> 83             205.859       47.9        23.021        16.240       35220.08
#> 137            112.371       46.2        21.502        14.924       27936.90
#> 45             237.016       46.6        21.453        15.957       45229.25
#> 57              18.136       42.8        21.228        13.264       40585.72
#> 17              65.180       44.7        20.801        13.272       18563.31
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 86               NA                79.370                5.72
#> 83              2.0               113.151                4.78
#> 137             0.5               127.842                9.85
#> 45               NA               156.139                8.31
#> 57               NA               153.507                5.76
#> 17              1.5               424.688                5.81
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 86                      13.050          84.63                   0.909
#> 83                       3.180          83.51                   0.880
#> 137                      3.390          82.05                   0.847
#> 45                       8.000          81.33                   0.936
#> 57                       3.280          81.91                   0.920
#> 17                       7.454          75.05                   0.813
#>     development
#> 86    very high
#> 83    very high
#> 137   very high
#> 45    very high
#> 57    very high
#> 17    very high
```

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

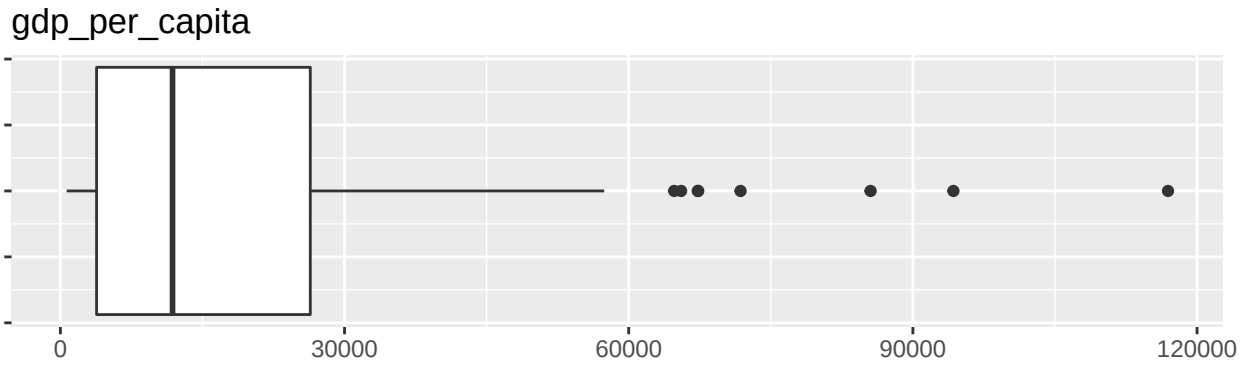
```
plots(dataset=data, col='aged_65_older',type='hist', density=TRUE, xtick_angles=c(0,30,20))
```

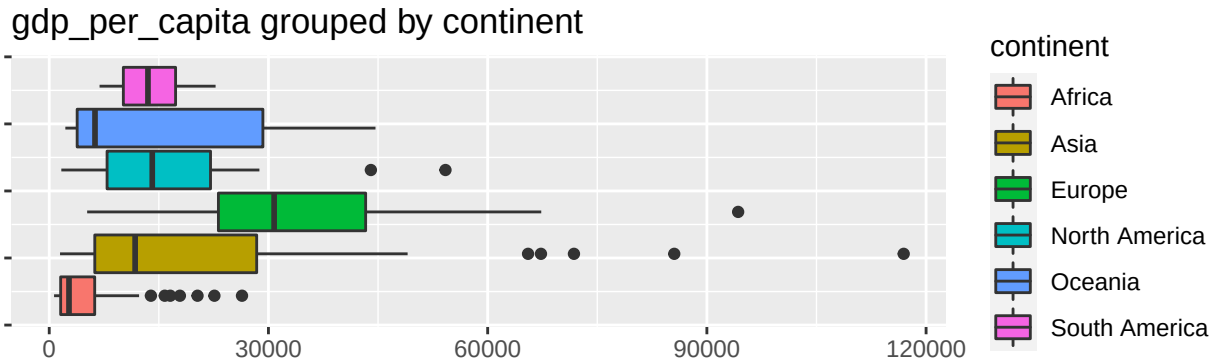


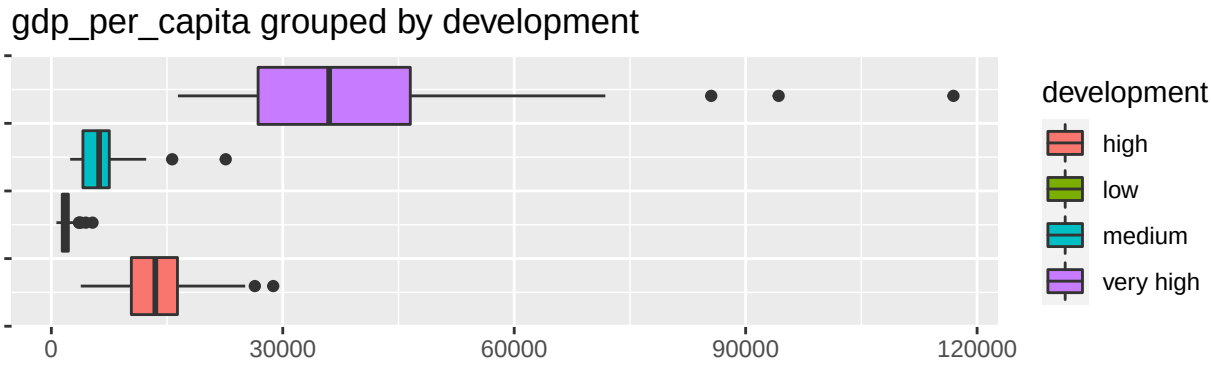
The distributions of this variable for different continent are very different. The distributions of Africa, Asia, North America and South America are right skewed while the distribution of Europe is left skewed.
The distribution of Africa is more concentrated (leptokurtic) while others are more flat. 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')
```







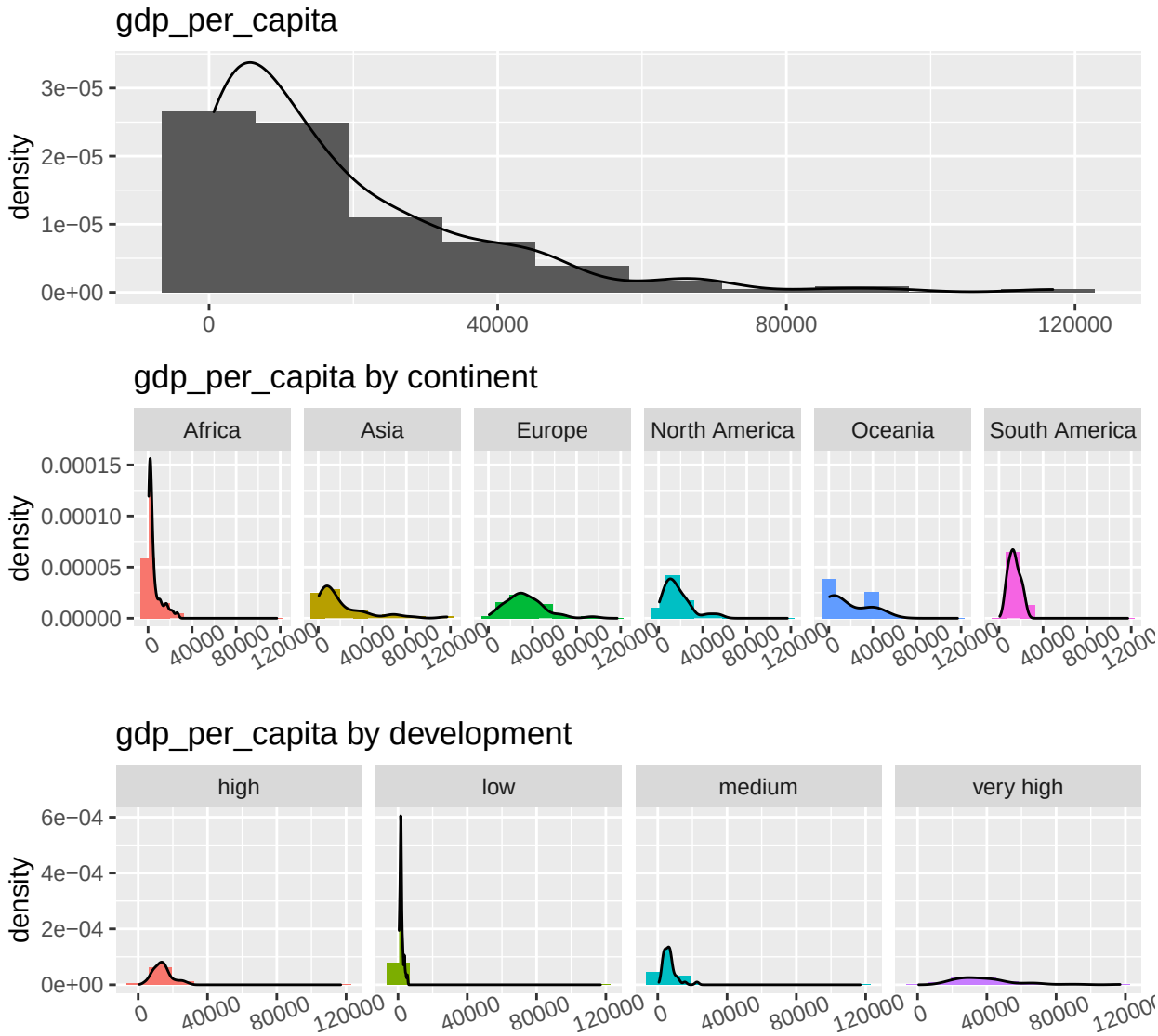

```
head(data[order(data$gdp_per_capita,decreasing=TRUE), ])
#>       X continent             location total_cases new_cases new_cases_smoothed
#> 140 139      Asia                Qatar      132917       197            212.143
#> 103 102    Europe           Luxembourg       19101       319            671.714
#> 147 146      Asia            Singapore       58020         1              6.714
#> 26   25      Asia               Brunei         148         0              0.000
#> 78   77    Europe              Ireland       62750       748            669.000
#> 5     4      Asia United Arab Emirates      135141      1234           1272.429
#>     total_deaths new_deaths new_deaths_smoothed total_cases_per_million
#> 140          232          0                0.286               46134.756
#> 103          160          0                1.857               30513.949
#> 147           28          0                0.000                9917.367
#> 26             3          0                0.000                 338.299
#> 78          1917          2                4.571               12708.099
```

```
#> 5            497         1               2.429              13663.856
#>     new_cases_per_million new_cases_smoothed_per_million
#> 140                68.378                         73.634
#> 103               509.604                       1073.067
#> 147                 0.171                          1.148
#> 26                  0.000                          0.000
#> 78                151.485                        135.486
#> 5                 124.767                        128.653
#>     total_deaths_per_million new_deaths_per_million stringency_index population
#> 140                   80.526                 0.000            64.81    2881060
#> 103                  255.601                 0.000            56.48     625976
#> 147                    4.786                 0.000            52.78    5850343
#> 26                     6.857                 0.000            35.19     437483
#> 78                   388.230                 0.405            81.48    4937796
#> 5                     50.251                 0.101            47.22    9890400
#>     population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 140            227.322       31.9         1.307         0.617      116935.60
#> 103            231.447       39.7        14.312         9.842       94277.96
#> 147           7915.731       42.4        12.922         7.049       85535.38
#> 26              81.347       32.4         4.591         2.382       71809.25
#> 78              69.874       38.7        13.928         8.678       67335.29
#> 5              112.442       34.0         1.144         0.526       67293.48
#>     extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 140              NA                176.690               16.52
#> 103             0.2                128.275                4.42
#> 147              NA                 92.243               10.99
#> 26               NA                201.285               12.79
#> 78              0.2                126.459                3.28
#> 5                NA                317.840               17.26
#>     hospital_beds_per_thousand life_expectancy human_development_index
#> 140                       1.20           80.23                   0.856
#> 103                       4.51           82.25                   0.904
#> 147                       2.40           83.62                   0.932
#> 26                        2.70           75.86                   0.853
#> 78                        2.96           82.30                   0.938
#> 5                         1.20           77.97                   0.863
#>     development
#> 140   very high
#> 103   very high
#> 147   very high
#> 26    very high
#> 78    very high
#> 5     very high
```

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 relevent 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

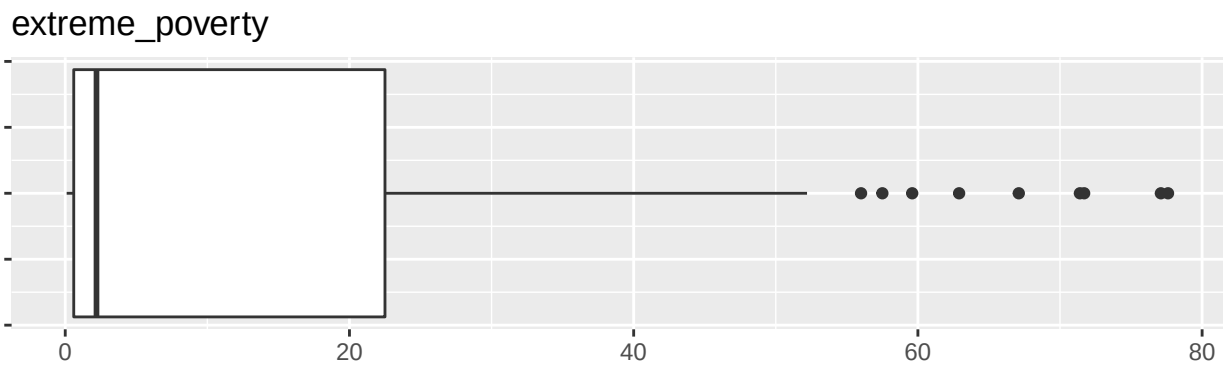
```
plots(dataset=data, col='gdp_per_capita',type='hist', density=TRUE, xtick_angles=c(0,30,20))
```

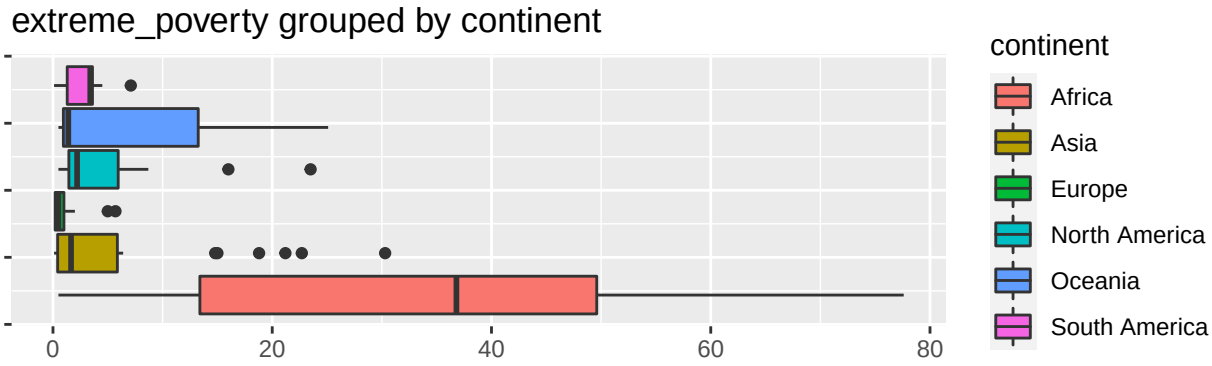


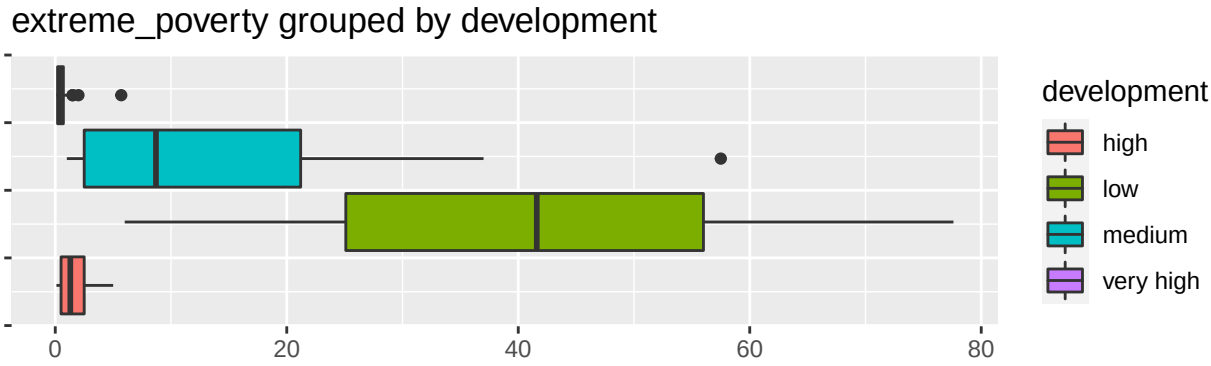
The global distribution of this variable is very right skewed. The distributions of Africa is more concentrated (leptokurtic) while others are more flat (platykurtic). The distributions of countries that have low HDI are more concentrated, and they usually have less GDP per capita.

Boxplots for percentage of population in extreme proverty

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







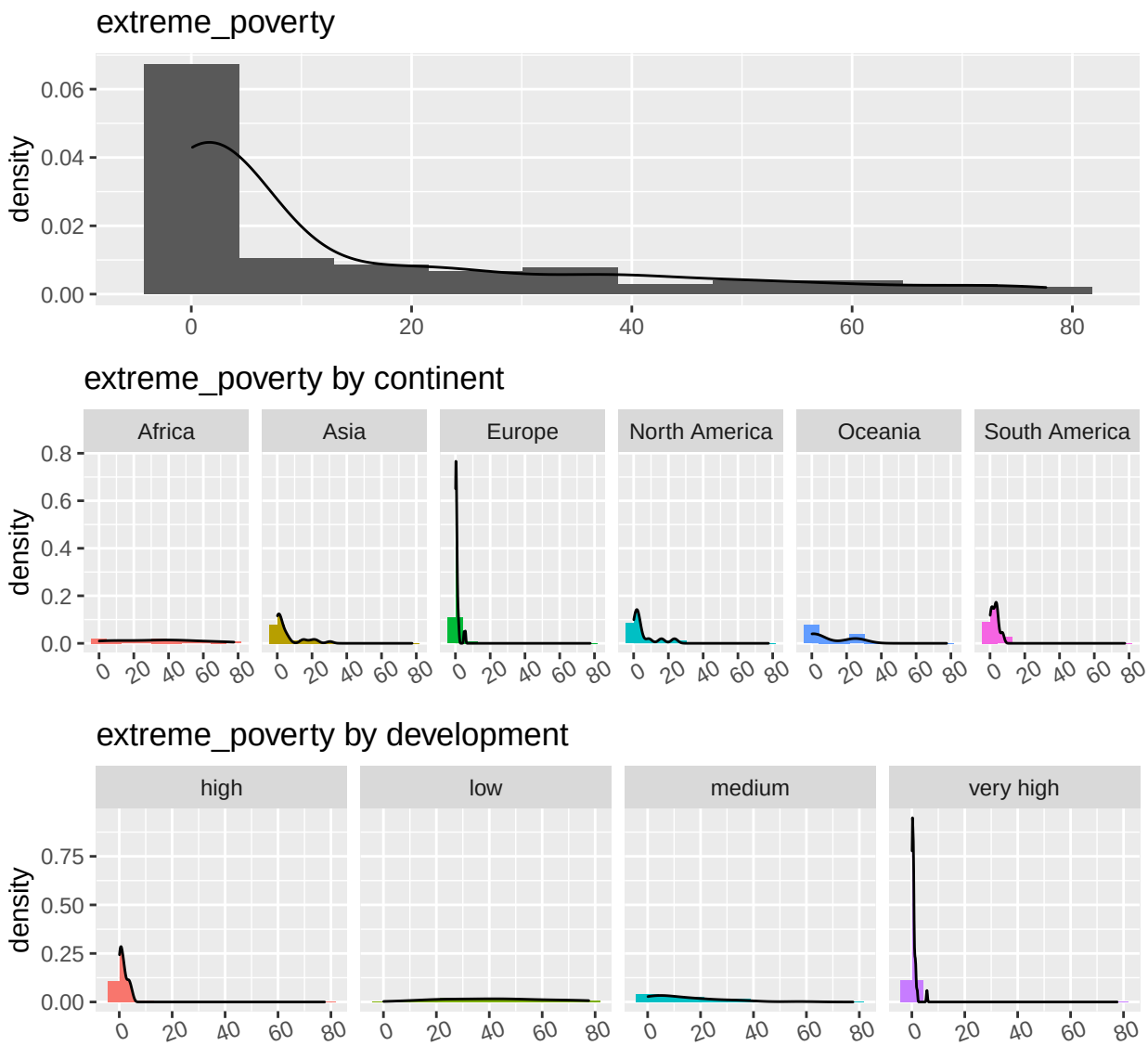

```
head(data[order(data$extreme_poverty,decreasing=TRUE), ])
#>       X continent                    location total_cases new_cases
#> 107 106    Africa                  Madagascar       17111         0
#> 36   35    Africa Democratic Republic of Congo       11372         1
#> 12   11    Africa                    Burundi         589         0
#> 120 119    Africa                     Malawi        5933         1
#> 66   65    Africa              Guinea-Bissau        2413         0
#> 117 116    Africa                 Mozambique       13130       142
#>     new_cases_smoothed total_deaths new_deaths new_deaths_smoothed
#> 107             20.429          244          0                0.000
#> 36              28.429          307          0                0.429
#> 12               4.571            1          0                0.000
#> 120              5.571          184          0                0.143
#> 66               1.429           41          0                0.000
```

```
#> 117            138.429          94          1               0.857
#>     total_cases_per_million new_cases_per_million
#> 107                 617.926                 0.000
#> 36                  126.974                 0.011
#> 12                   49.534                 0.000
#> 120                 310.142                 0.052
#> 66                 1226.119                 0.000
#> 117                 420.087                 4.543
#>     new_cases_smoothed_per_million total_deaths_per_million
#> 107                          0.738                    8.812
#> 36                           0.317                    3.428
#> 12                           0.384                    0.084
#> 120                          0.291                    9.618
#> 66                           0.726                   20.833
#> 117                          4.429                    3.007
#>     new_deaths_per_million stringency_index population population_density
#> 107                  0.000            52.78   27691019              43.951
#> 36                   0.000               NA   89561404              35.879
#> 12                   0.000            14.81   11890781             423.062
#> 120                  0.000            50.93   19129955             197.519
#> 66                   0.000               NA    1967998              66.191
#> 117                  0.032            56.48   31255435              37.728
#>     median_age aged_65_older aged_70_older gdp_per_capita extreme_poverty
#> 107       19.6         2.929         1.686       1416.440            77.6
#> 36        17.0         3.020         1.745        808.133            77.1
#> 12        17.5         2.562         1.504        702.225            71.7
#> 120       18.1         2.979         1.783       1095.042            71.4
#> 66        19.4         3.002         1.565       1548.675            67.1
#> 117       17.7         3.158         1.870       1136.103            62.9
#>     cardiovasc_death_rate diabetes_prevalence hospital_beds_per_thousand
#> 107               405.994                3.94                        0.2
#> 36                318.949                6.10                         NA
#> 12                293.068                6.05                        0.8
#> 120               227.349                3.94                        1.3
#> 66                382.474                2.42                         NA
#> 117               329.942                3.30                        0.7
#>     life_expectancy human_development_index development
#> 107           67.04                   0.519         low
#> 36            60.68                   0.457         low
#> 12            61.58                   0.417         low
#> 120           64.26                   0.477         low
#> 66            58.32                   0.455         low
#> 117           60.85                   0.437         low
```

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

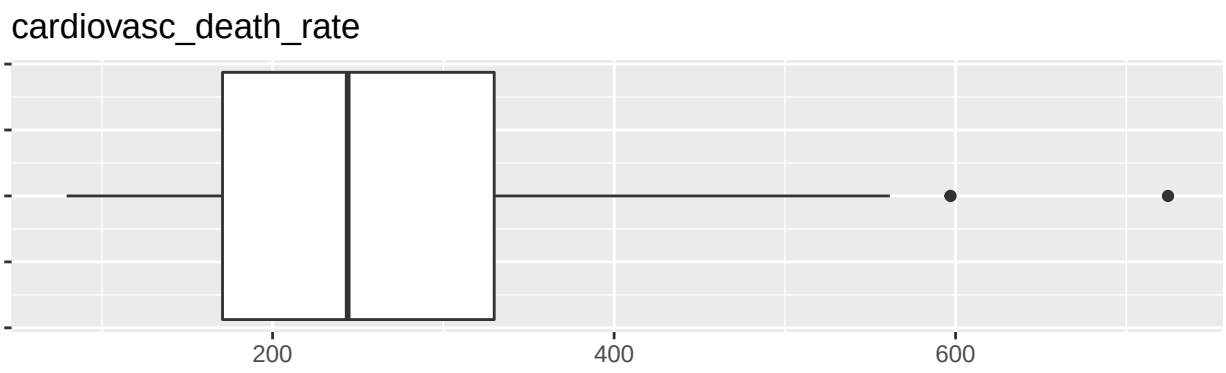
```
plots(dataset=data, col='extreme_poverty',type='hist', density=TRUE, xtick_angles=c(0,30,20))
```

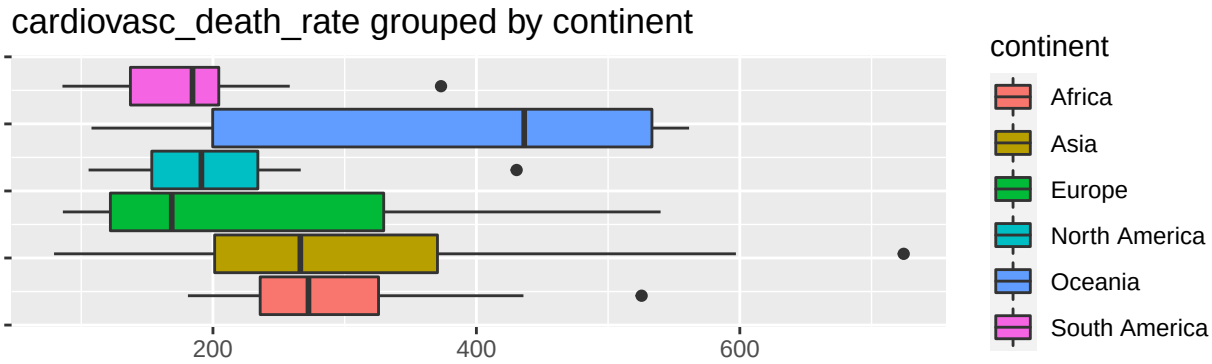


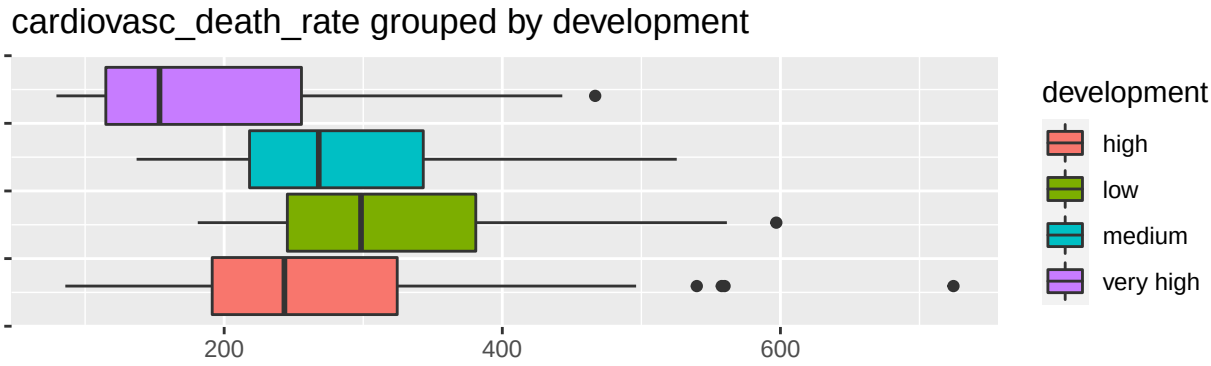
The global distribution of this variable is very right skewed. The distributions of Europe is more concentrated (leptokurtic) in low values while others are more flat (platykurtic). The distributions of countries that have very high HDI are more concentrated, and they usually have lower extreme poverty.

Boxplots for cardiovascular death rate

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







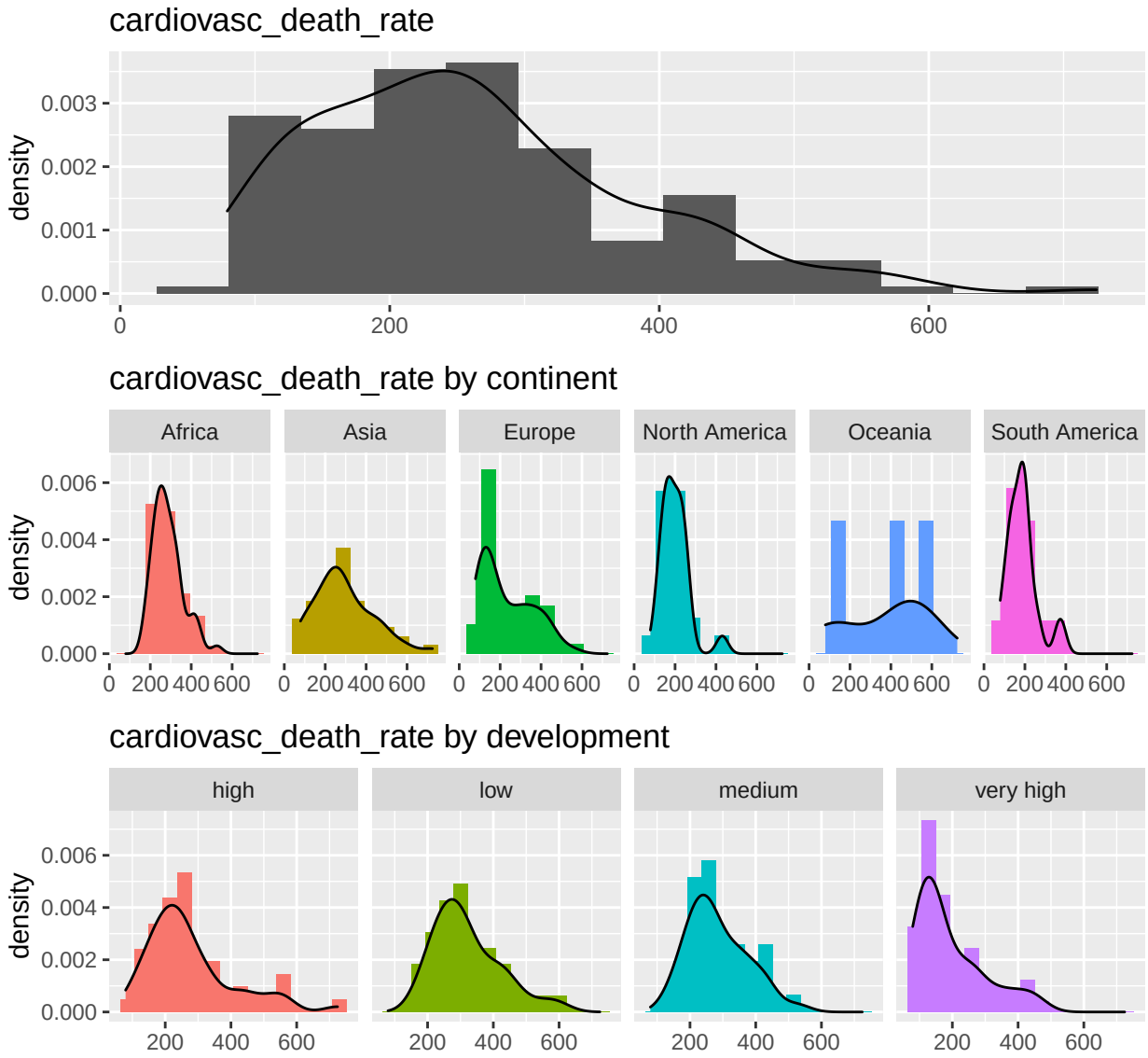
It appears that most countries seem to have a cardiovascular death rate between 170.67 and 329.79 deaths per 100,000 inhabitants. With Uzbekistan being in the absolute extreme, with about 724 deaths by cardiovascular disease per 100,000 inhabitants.

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_angl
```



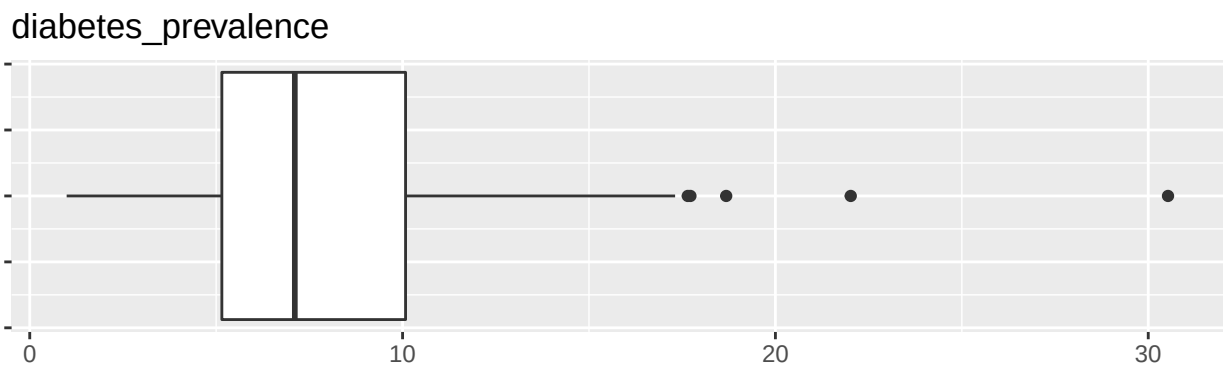
For cardiovascular death rate we see a similar story here than with the general boxplot. The larger concentration of countries clumps around the previously mentioned interval, and the distribution of teh variable as is is somewhat normal-like with a relatively long left tail.

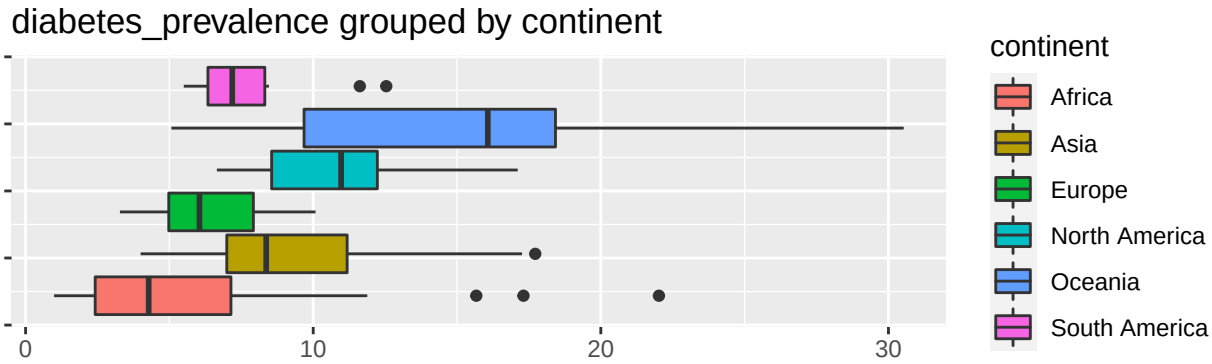
Per continent we see that Europe has a significant concentration of countries below 200, along with South america, which, on average, is the continent with the lowest death rate from cardiovascular disease according to our data. For oceania we see a flat distribution with some high numbers and low numbers, 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.

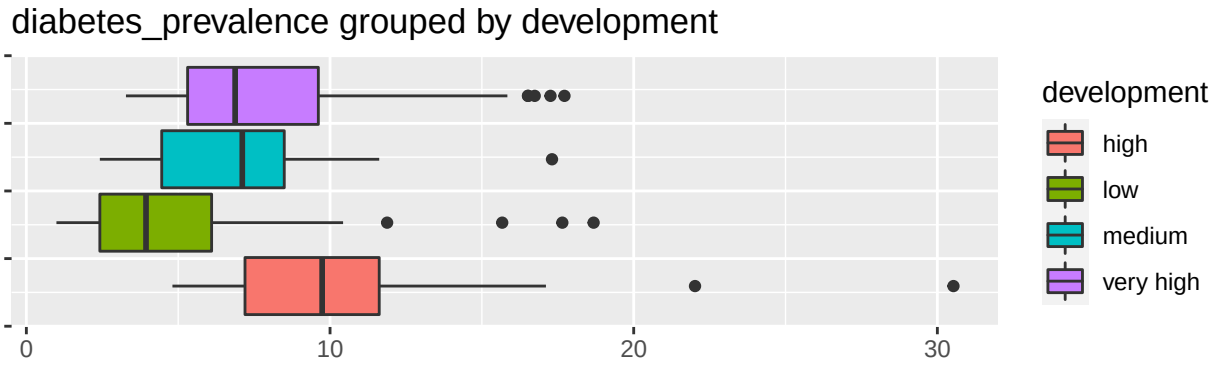
Looking at development we see the much higher concentration of low cardiovascular death rates for very high development countries. Which in general tend to have better healthcare. However while lower for low and medium development countries, we don't see too much of a difference between the two in terms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

Boxplots for diabetes prevalence

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







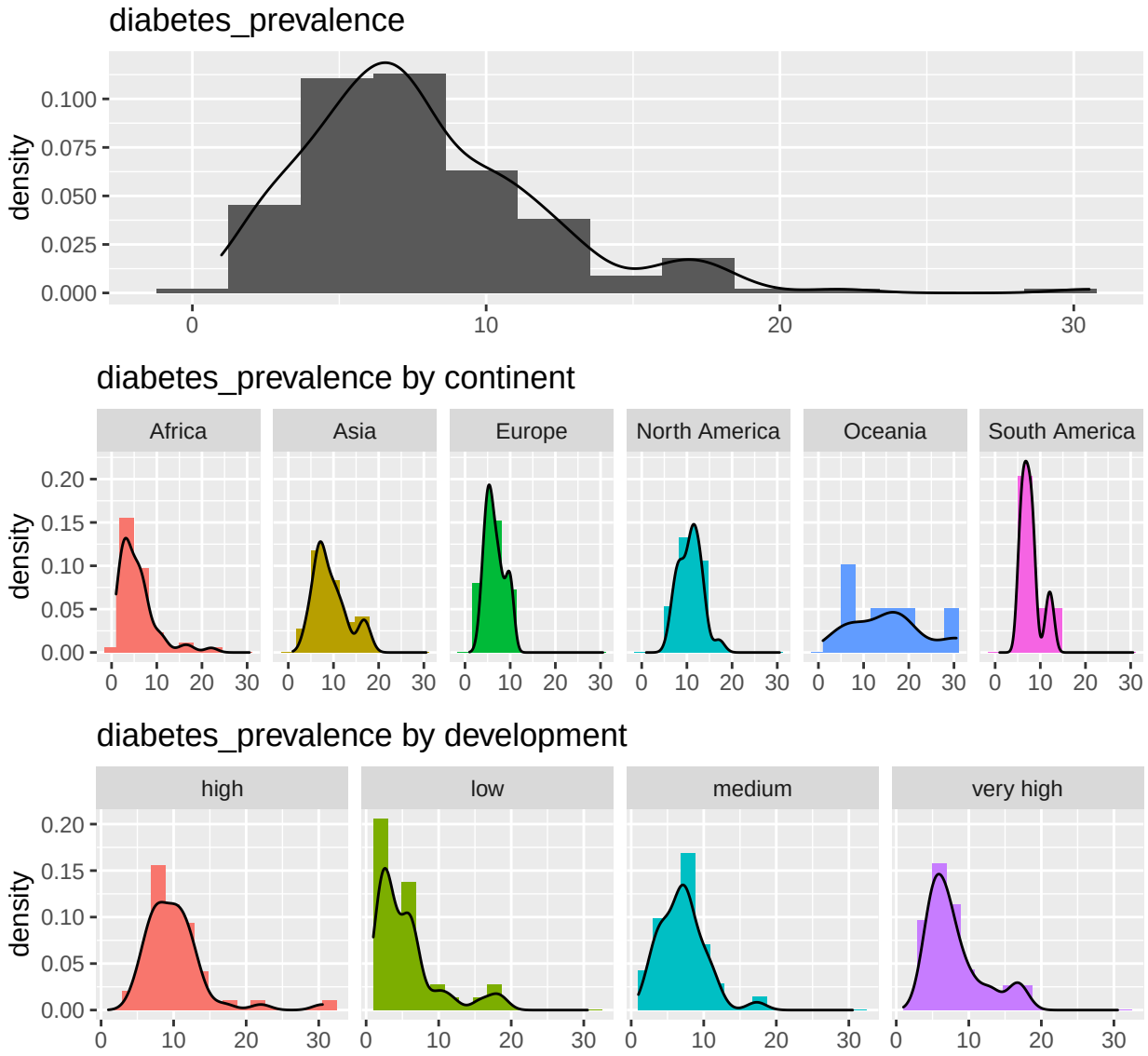
For diabetes prevalence we can see that most countries sit at a value of less than 10, but higher than 5. Some countries surpassing even 30%. These extreme values correspond to a few countries in Oceania and Africa.

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=
```



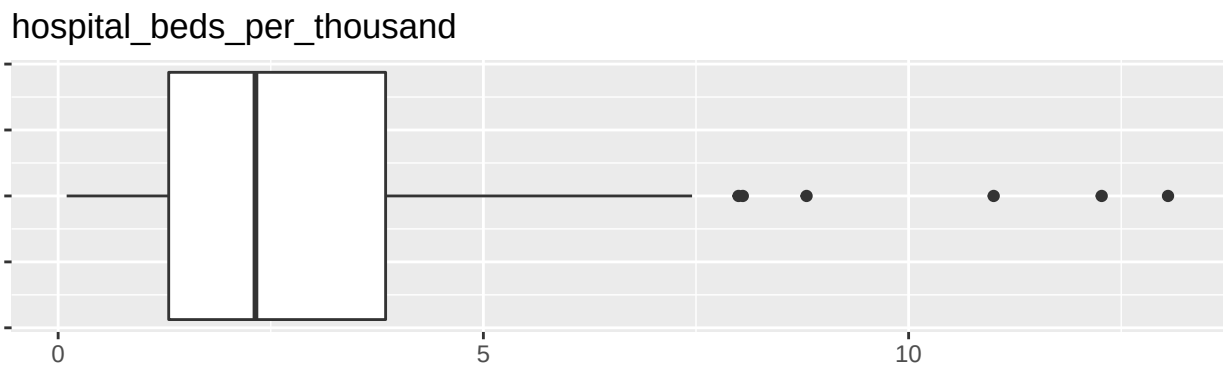
For the distribution of the data we see that it is resembles a normal distribution with a long left tail and the most countries clumped around the mean of ~7.9%.

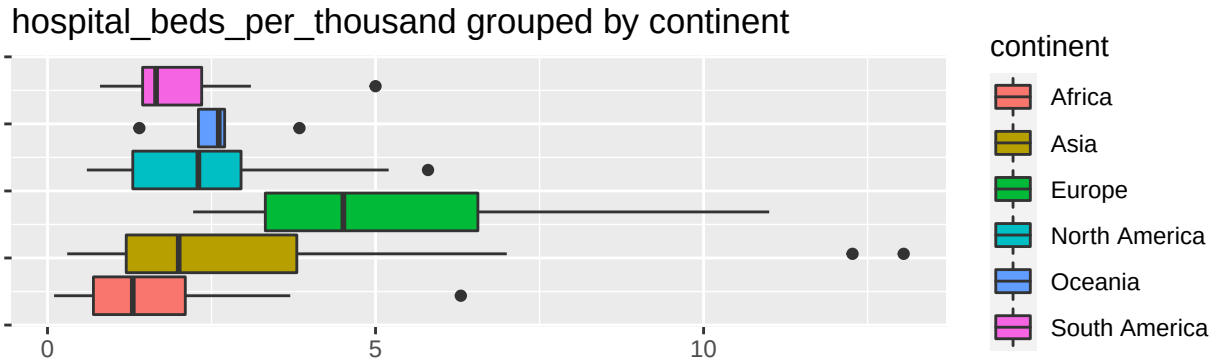
For each continent the incidence seems to be quite different, with some continents having a much higher incidence than others (for example Oceania vs Africa), however they all seem to clump around similar values.

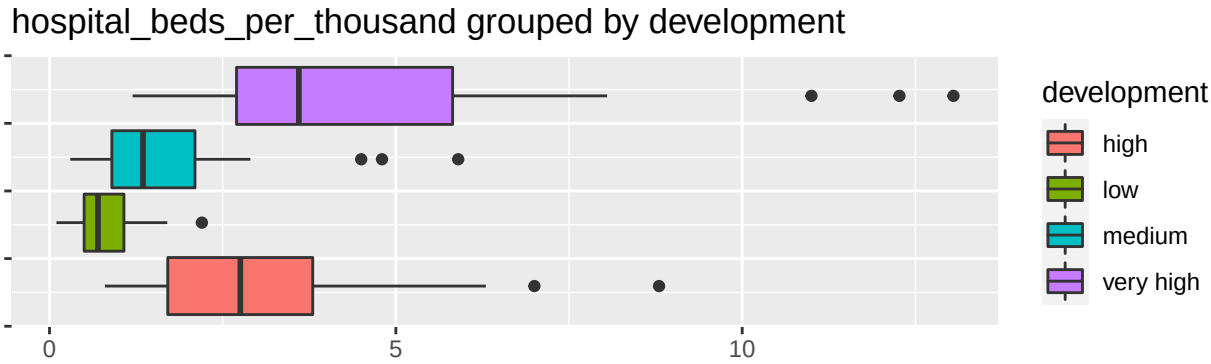
For the development we see the same we saw in the boxplots. Not much of a pattern or indication that there's any specific relationship between HDI and diabetes prevalence.

Boxplots for hospital beds per thousand inhabitants

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







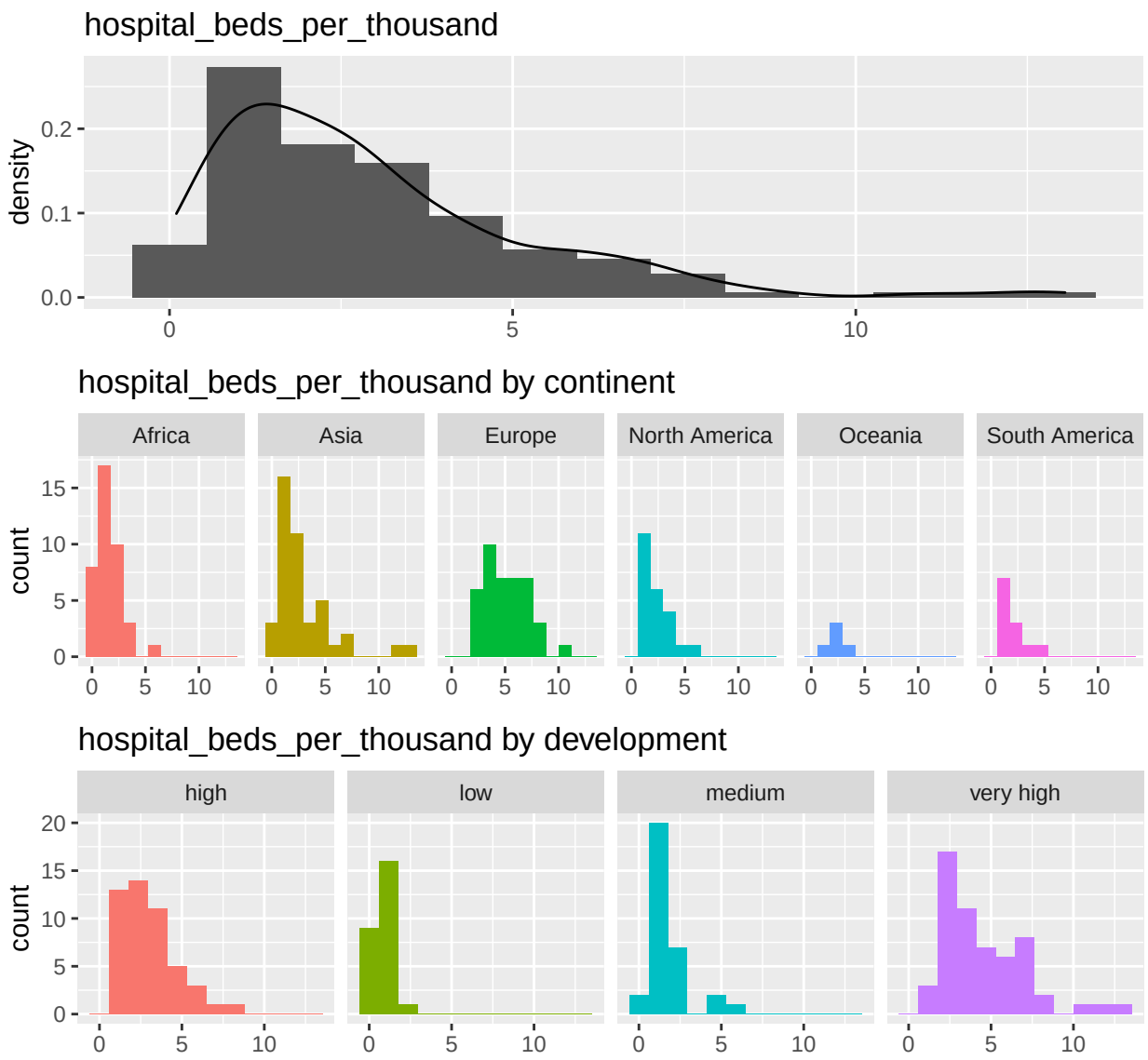
Looking at the hospital beds per thousand inhabitants variable boxplots we can see a few interesting things. We could use this variable as a measure of the quality of a healthcare system of a country. Where the higher the bed availability in hospitals is, the better the health system can cope with the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

We can see that some extremely poor 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=FALSE, bins=c(13,12,12), xtic
```



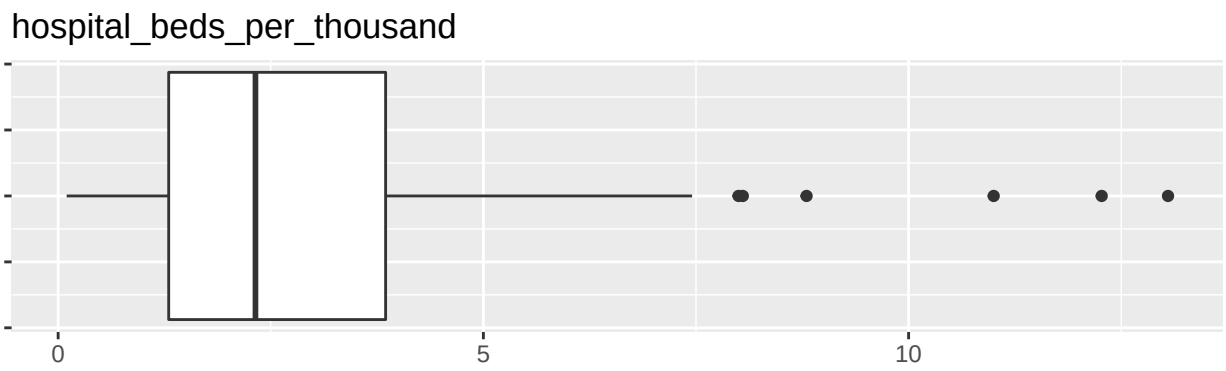
These plots tell a little bit of a different story to the boxplots. Where the largest concentration of countries is between 0 and 5 hospital beds per thousand inhabitants with an extremely scarce amount of countries with more than 10 beds per thousand inhabitants.

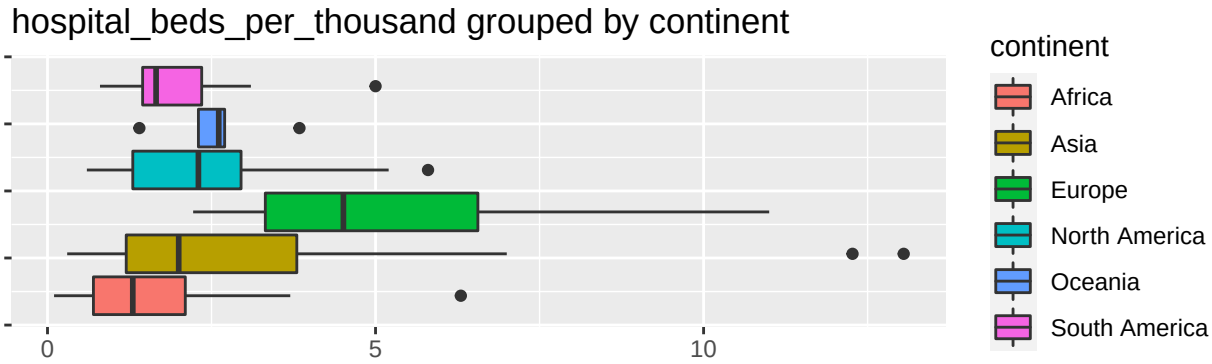
By segregating the data by continent we see that development does not necesarily mean greater healthcare capacity, with most continents boasting very similar numbers in this aspect whil esome like Asia, Africa, Europe and North america possessing some exceptions with extremely high numbers compred to 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.

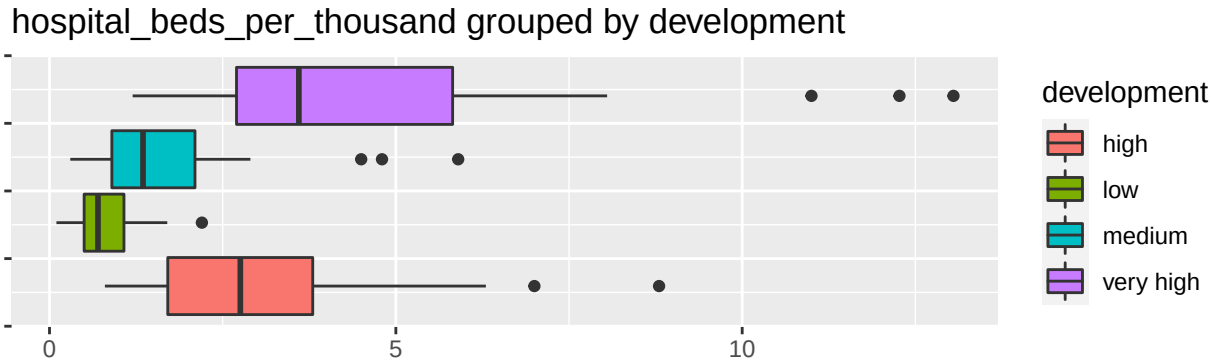
Finally, looking at development we see that it is rare for much less developed countries to have high bed capacity, while it is much easier for high to very high developed countries to have greater capacity. However, we can't confidently say that there's lots of exceptions to this 'rule'.

Boxplots for life expectancy

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







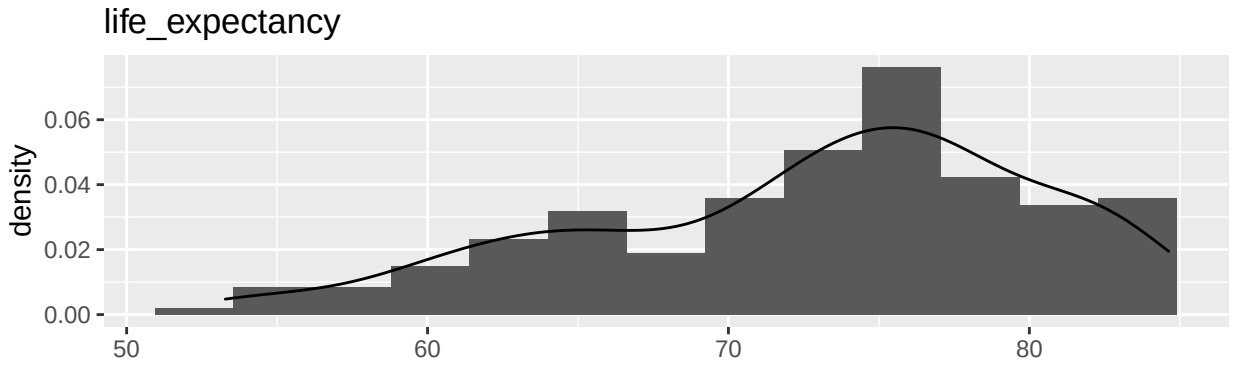
For life expectancy we can see moost countries sitting amove 66 years of age, with values going as low as 53.24 and as high as 84.63.

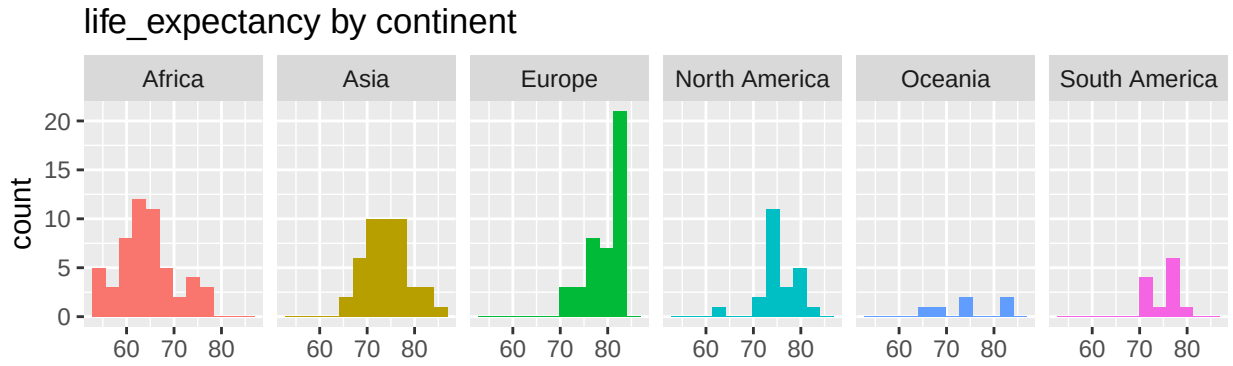
Africa has the lowest life expectancy while europe has the highest. The rest of the continents sit at roughly similar ranges.

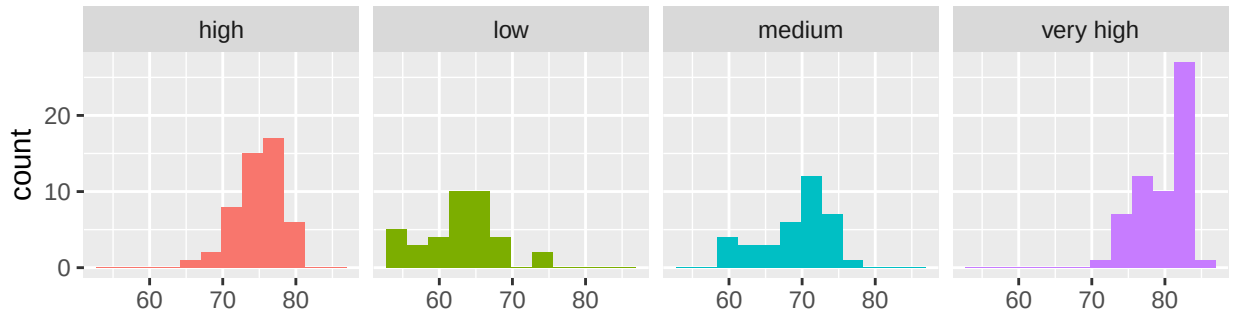
Grouping by HDI, we can see that the most developed countries have a significantly higher life expectancy than those with low HDI. It clearly shows a strong positive correlation between them. Where the higher the life expectancy the higher the HDI. With very few exceptions.

Histogram and kernel density for life expectancy

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







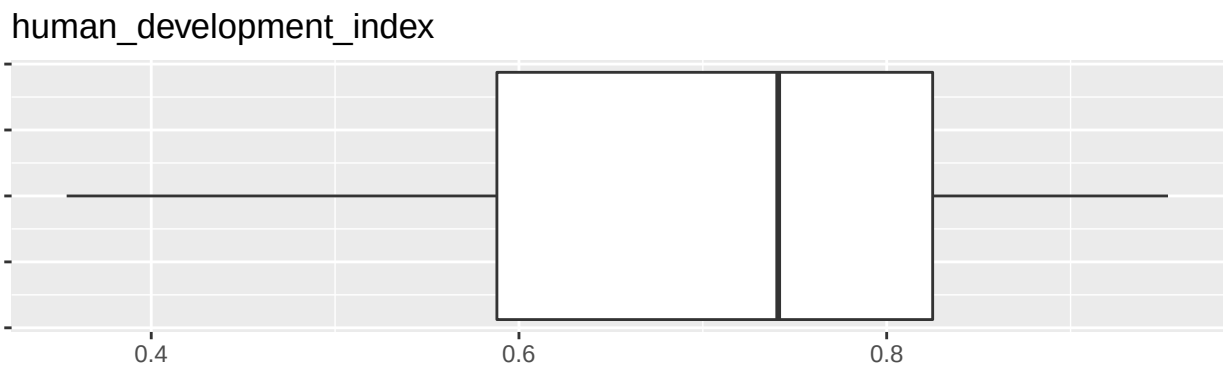
The general plot is somewhat left skewed, as most countries (about 80%) have a life expectancy higher than 65 years of age. Our density plot shows a strong concentration between 70 and 80 years of age, as this range covers the most nations.

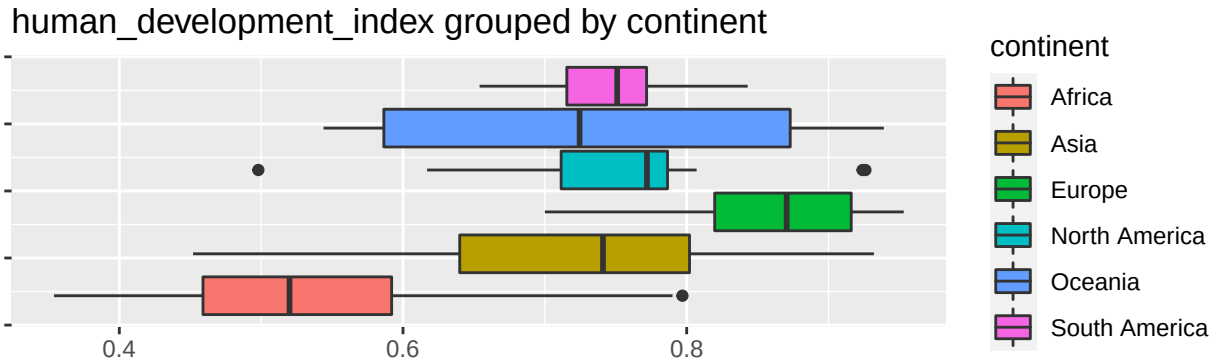
For each continent we see that europe shows a typically very high life expectancy while Africa shows a typically lower-than-average life expectancy for most countries with some exceptions. The rest of the continents sit at about the average life expectancy with some countries in Asia and North America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

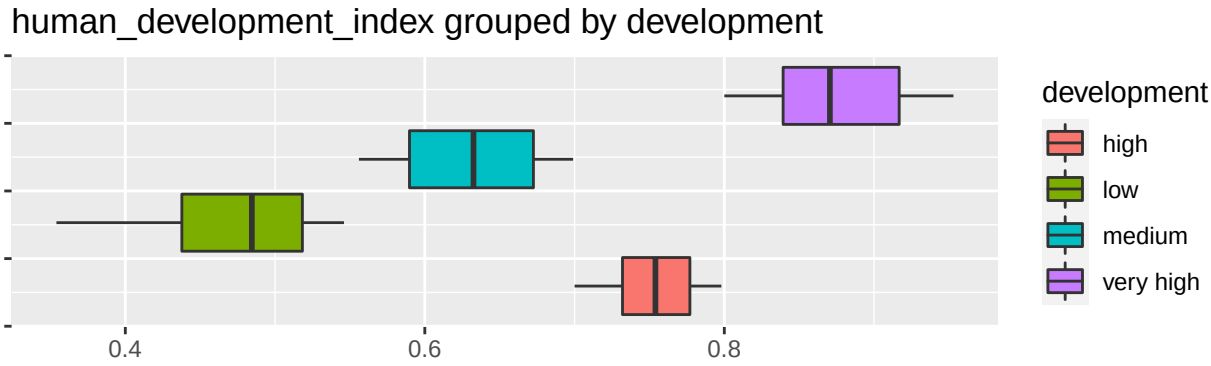
For HDI we can again see some of the strong correlation, where life expectancy for very highly developed nations seems to be also quite high and the same happens with less developed nations.

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.

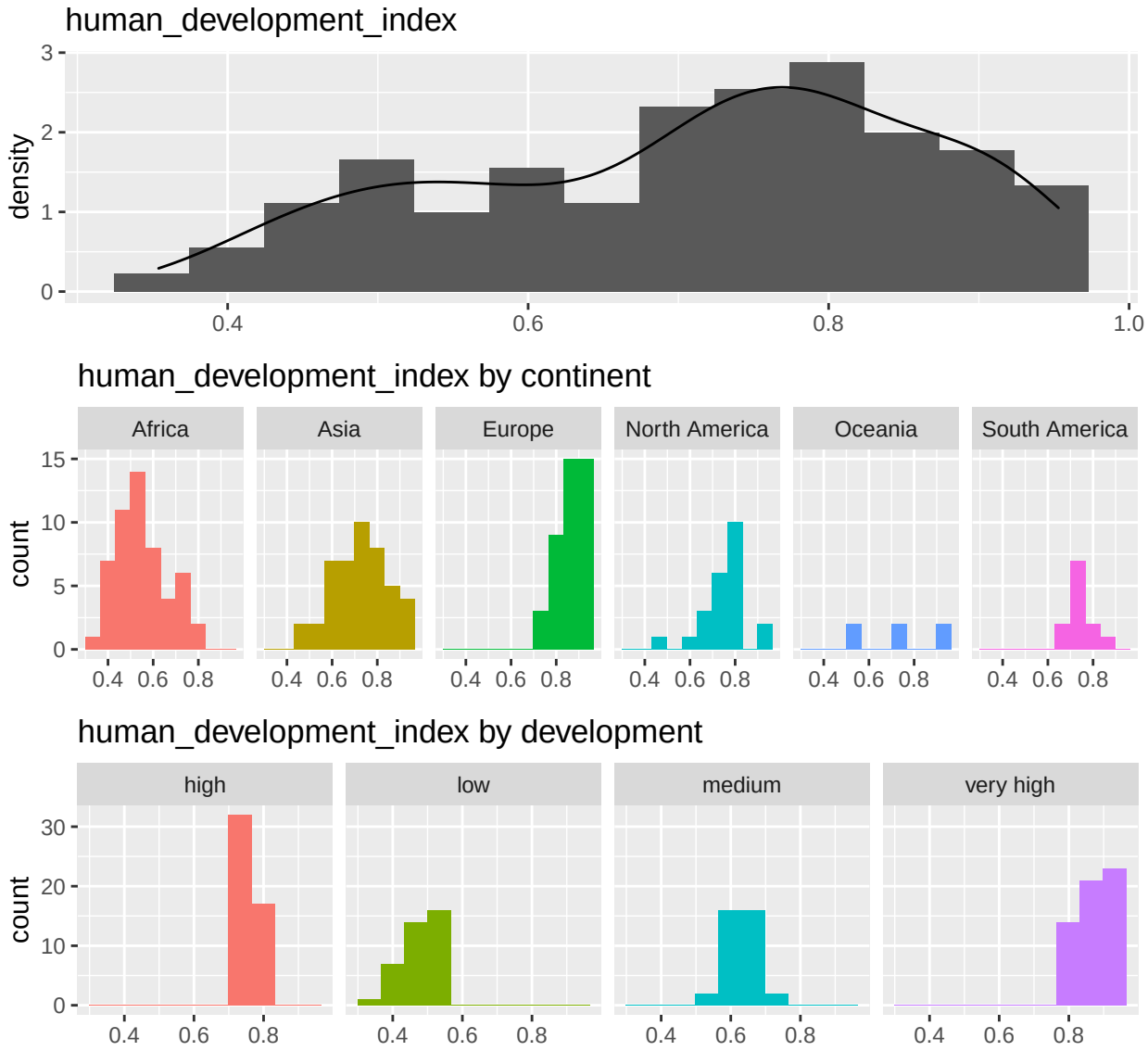
For continents we can see Africa lagging behind with most of its countries between 0.4 and 0.6 HDI, probably given the porverty situation in the continent.

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

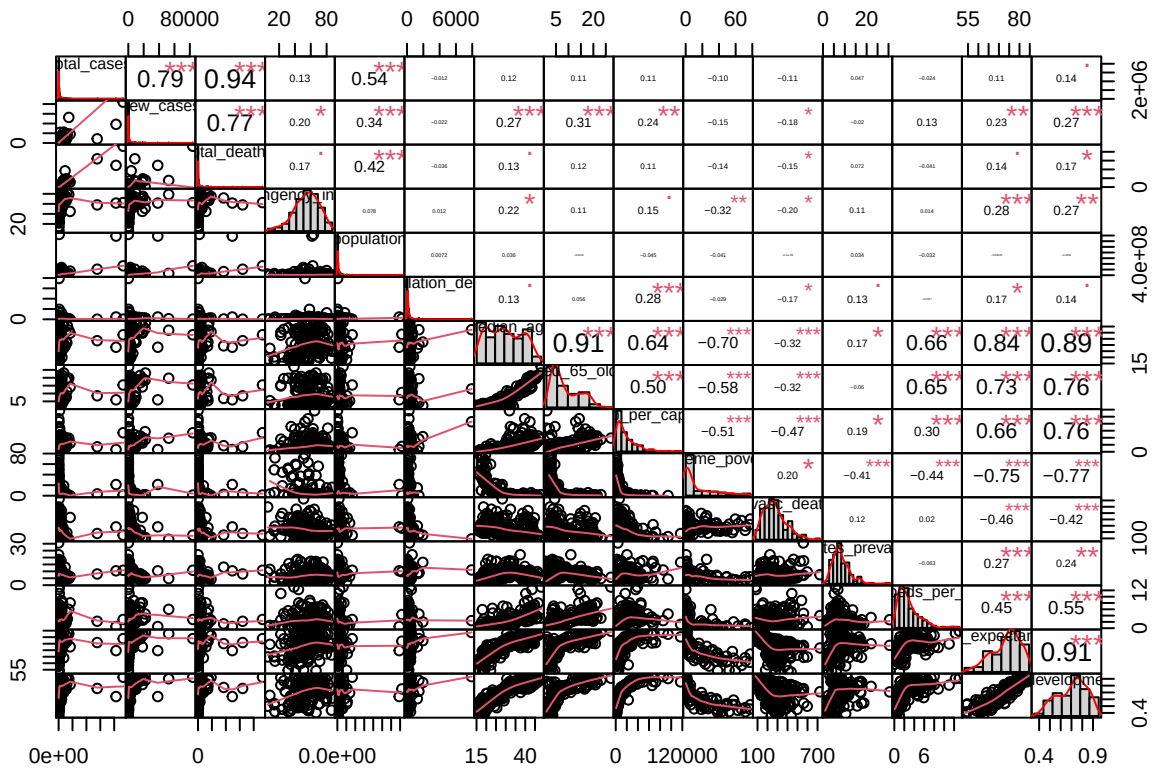
```
plots(dataset=data, col='human_development_index',type='hist', density=FALSE, bins=c(13,10,10), xtick_a
```



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.

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



PCP Plot

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)
```

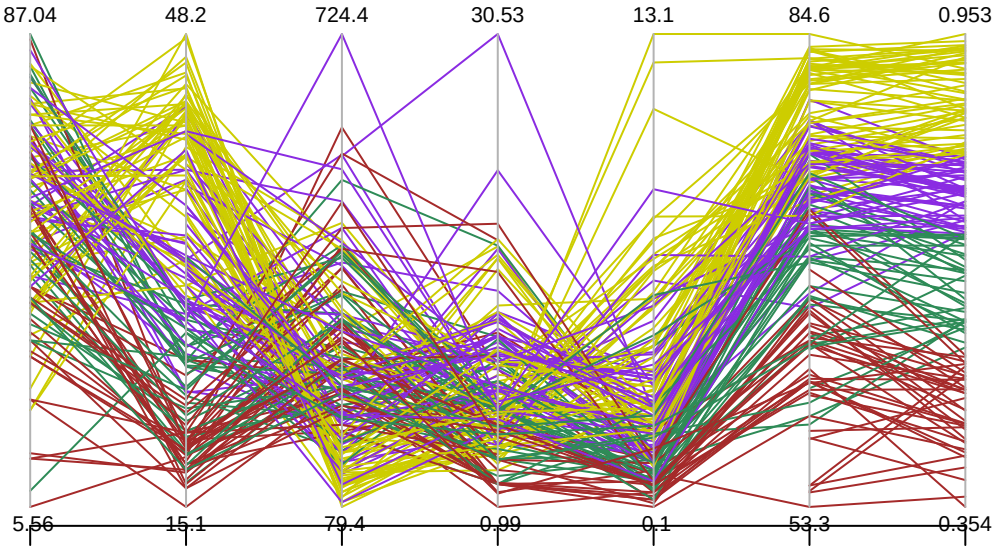
```
development_colors <- colors(data$development,palette1)
continent_colors <- colors(data$continent,palette2)
```

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.

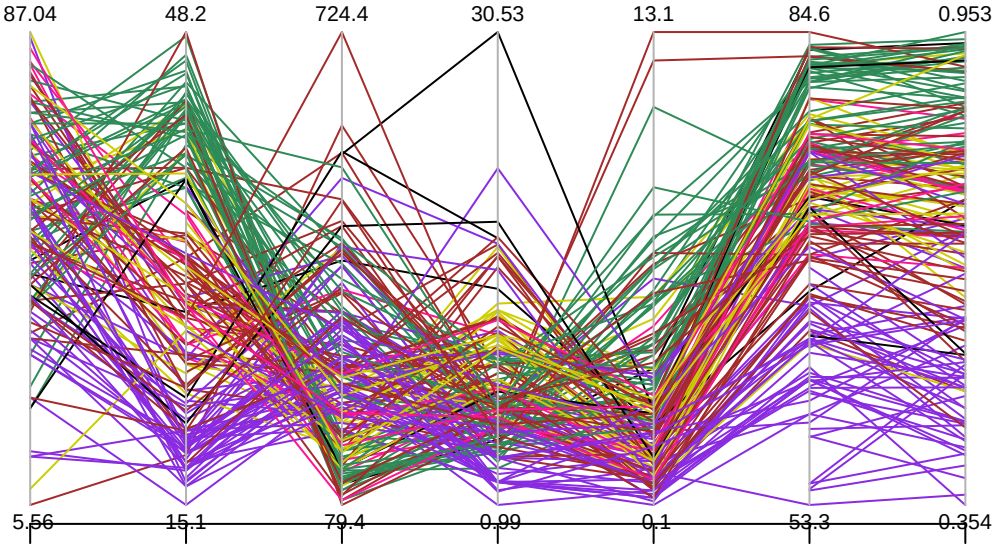
```
right_skewed <- c('total_cases','new_cases','population',
                  'population_density','aged_65_older',
                  'gdp_per_capita','extreme_poverty')
right_skewed <- data_n %>% dplyr::select(right_skewed)

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

Right skewed variables PCP

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

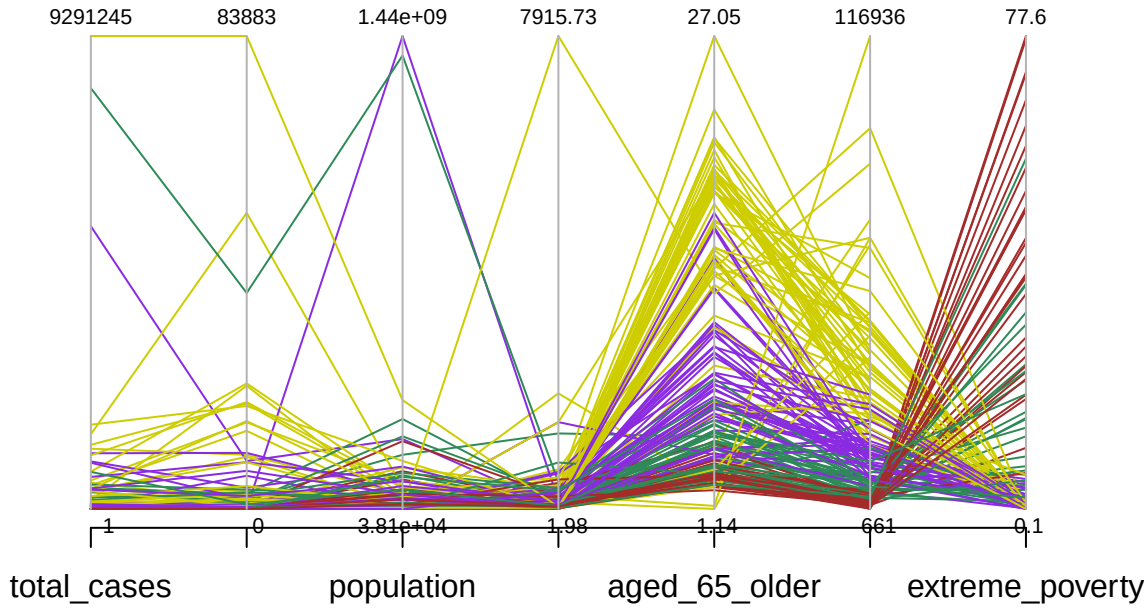



```
parcoord(others,var.label=TRUE, col=continent_colors)
```

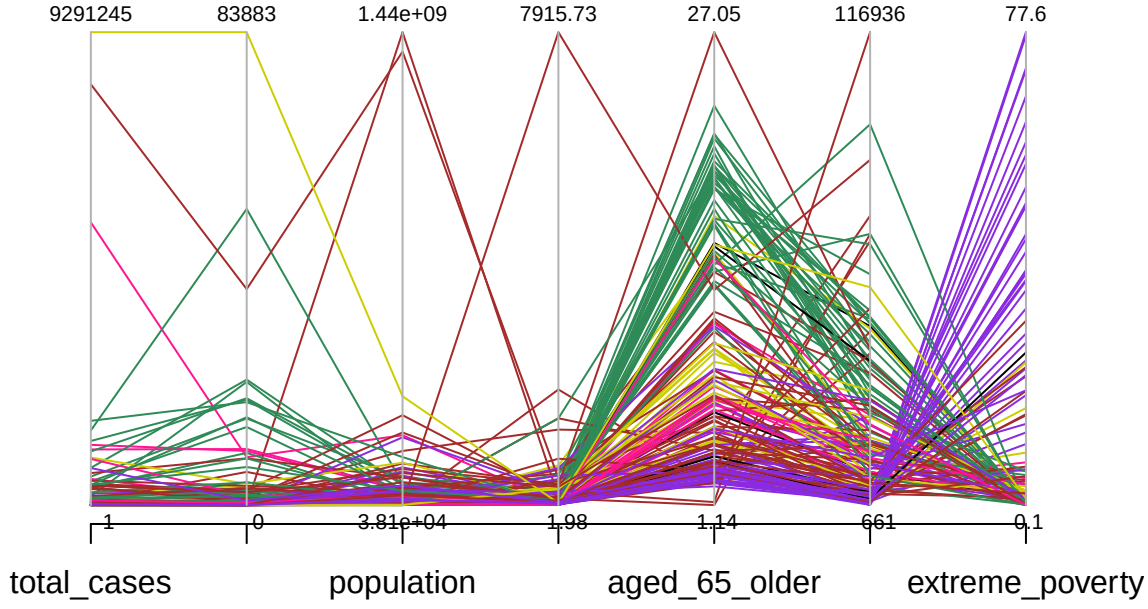


Other variables PCP

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




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




```
library(mice)
covid = data[c('continent','location','development',
               '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')]
covid$continent=factor(covid$continent)
covid$development=factor(covid$development)
covid_imp=mice(covid,m=5,method = "cart")
covid_imp=complete(covid_imp)
```