# 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)
```

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_cas
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

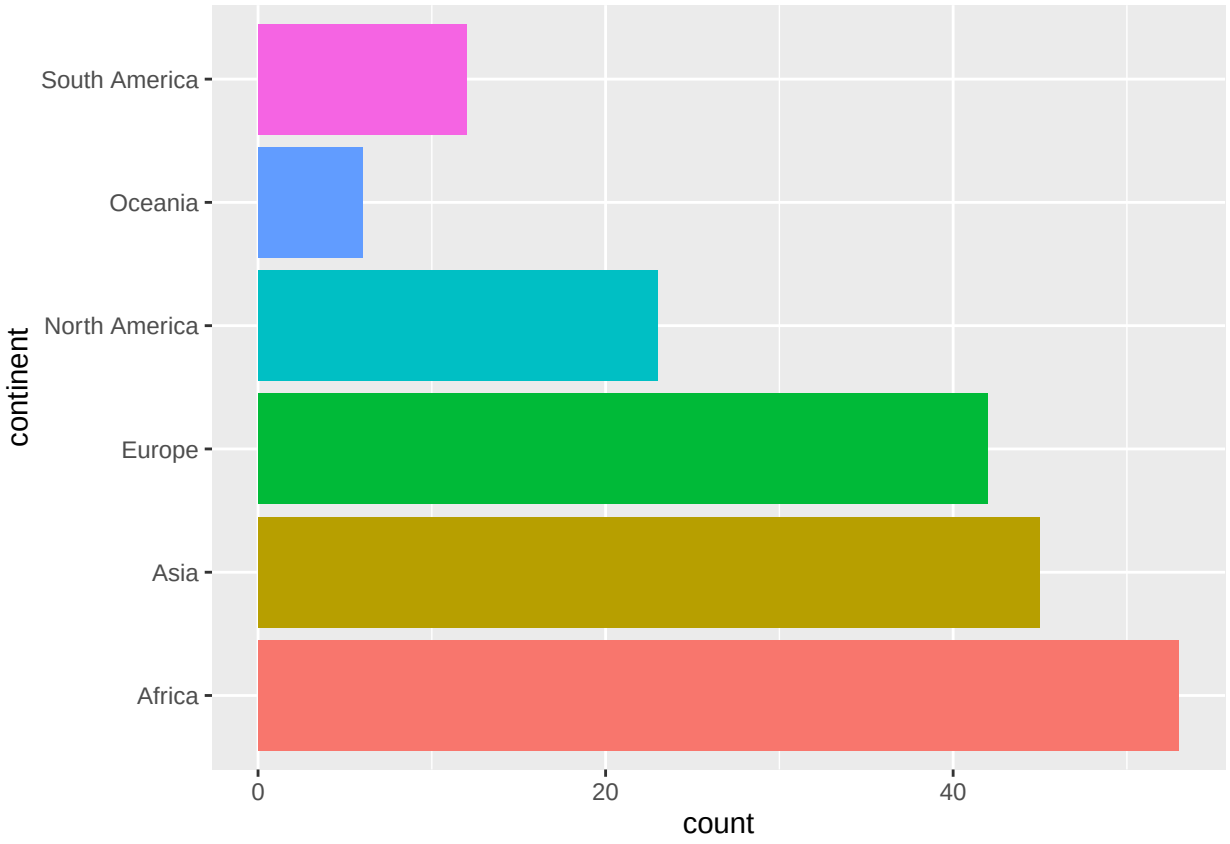
We select variables that we consider interesting to visualize, as the ones we haven't selected might be ralated to these or even ratios of them (in the case of total cases per million)

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

### Plots with categorical variables

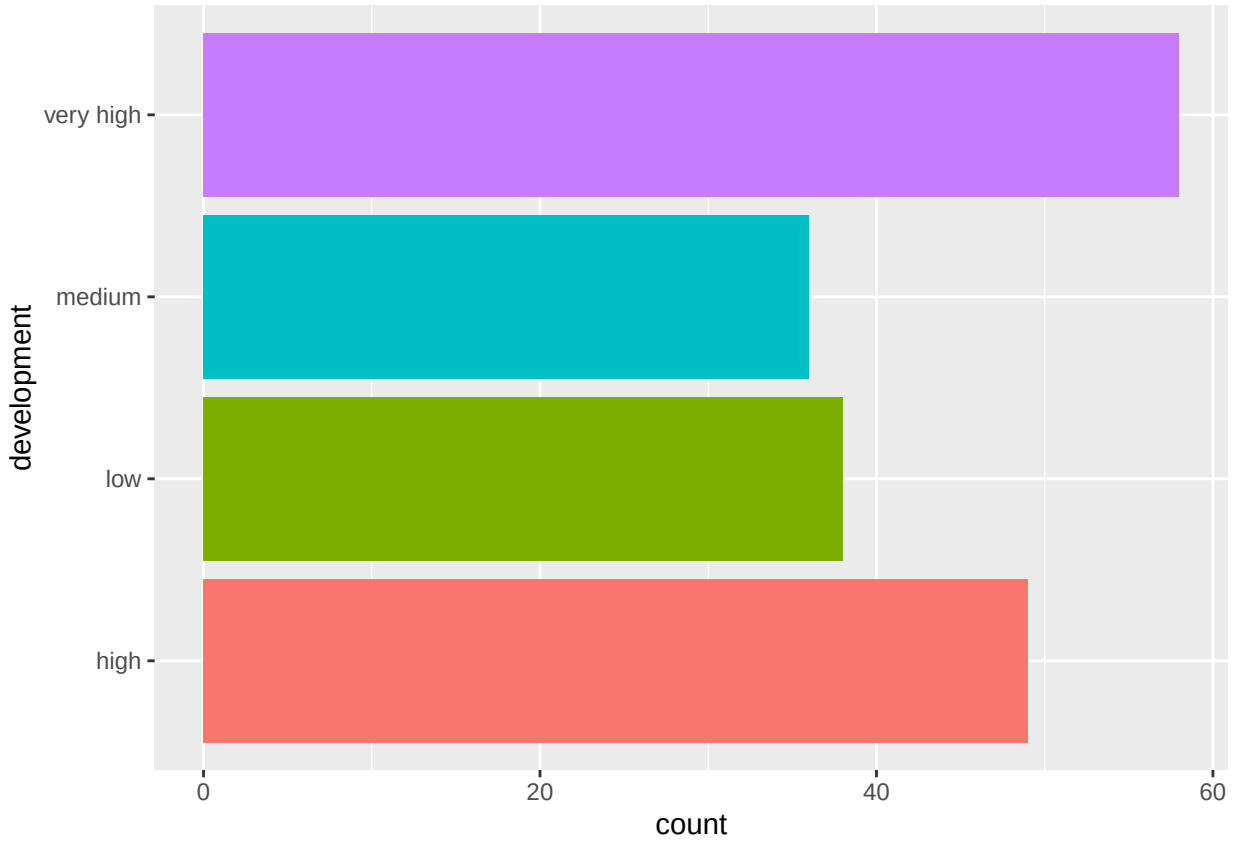
#### Countries per continent in the dataset

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



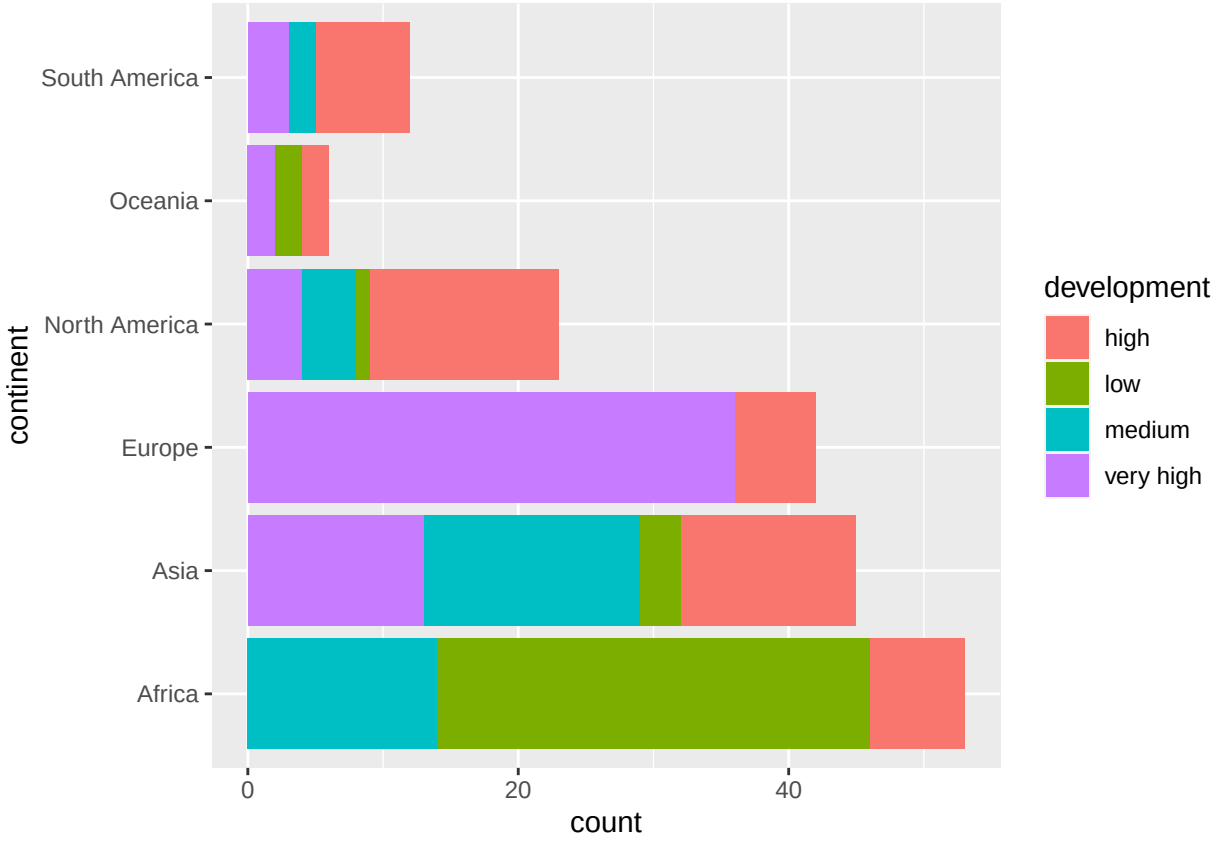

#### Amount of countries per HDI

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



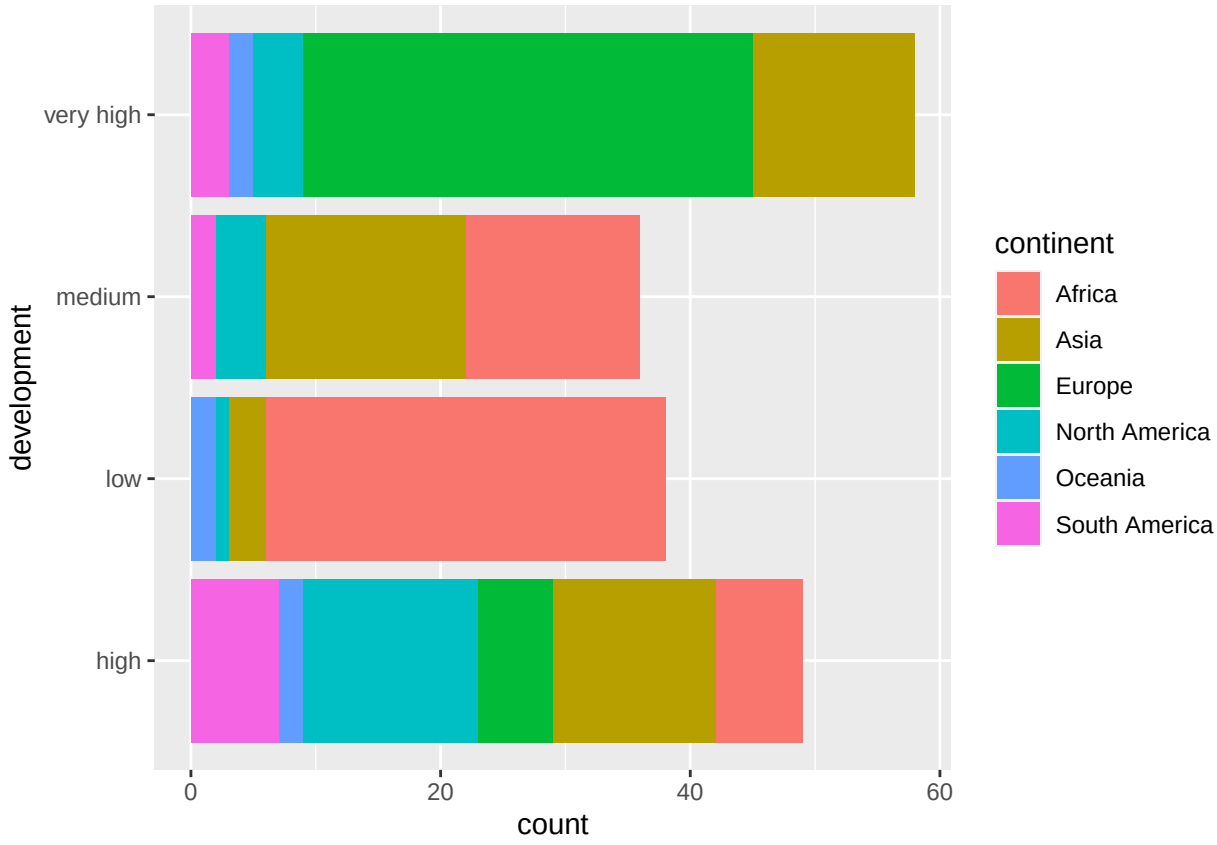
**Countries per continent per HDI**

```
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"
```

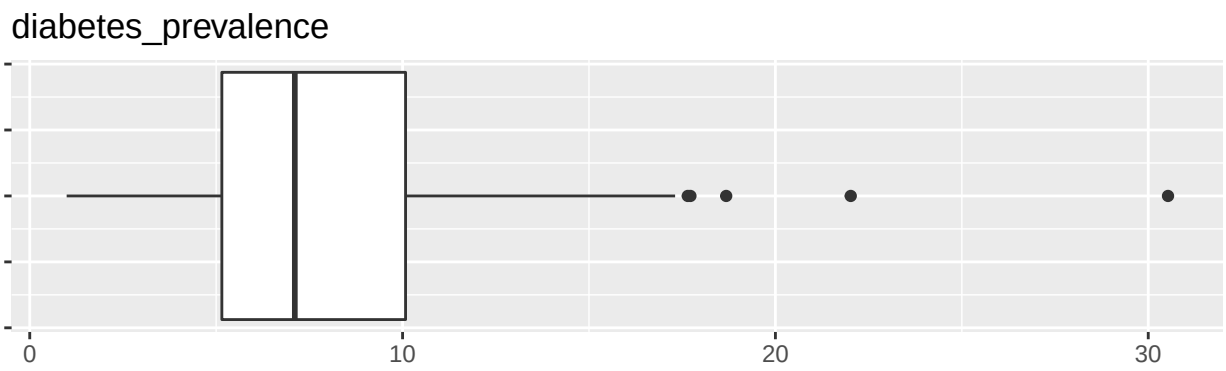
### Function to plot quantitative variables

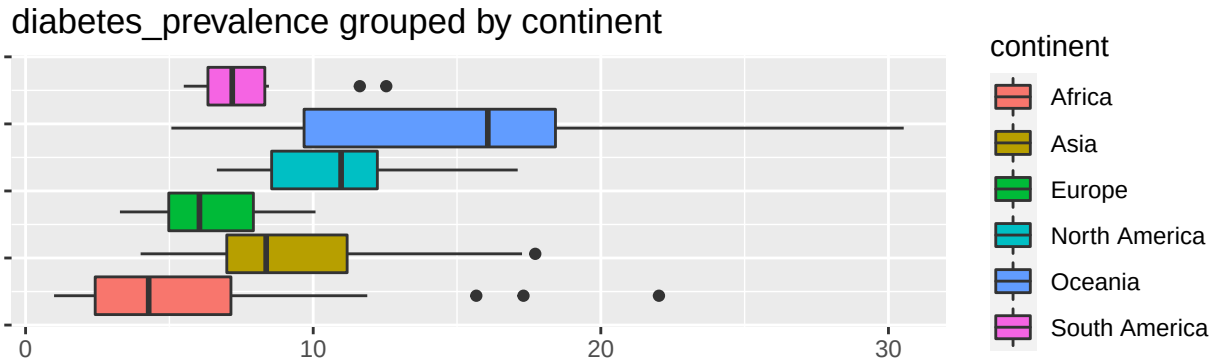
```
plots <- function(dataset ,col, type, density=TRUE, bins='default', xtick_angles='default') {
    var <- dataset %>% select(col)
    if (bins == 'default') {bins = rep(10,3)}
    if (xtick_angles == 'default') {xtick_angles = rep(90,3)}
    if (type == 'boxplot') {
        p1 <- dataset %>% ggplot(aes(x=var[,1])) +
            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() +
```

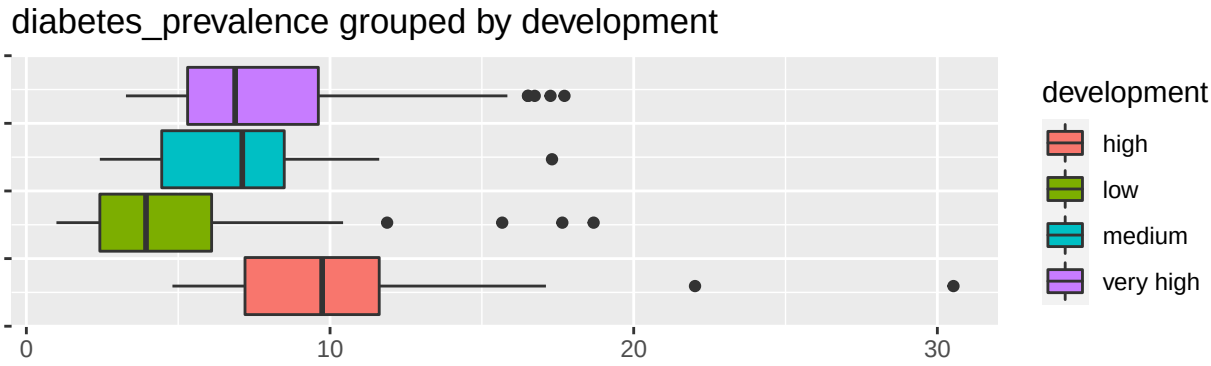
```
            ggtitle(str_interp("${col} grouped by continent")) +
            theme(axis.title.x=element_blank(),
                  axis.text.y=element_blank())
        p3 <- dataset %>% ggplot(aes(x=var[,1], fill=development)) +
            geom_boxplot() +
            ggtitle(str_interp("${col}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

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

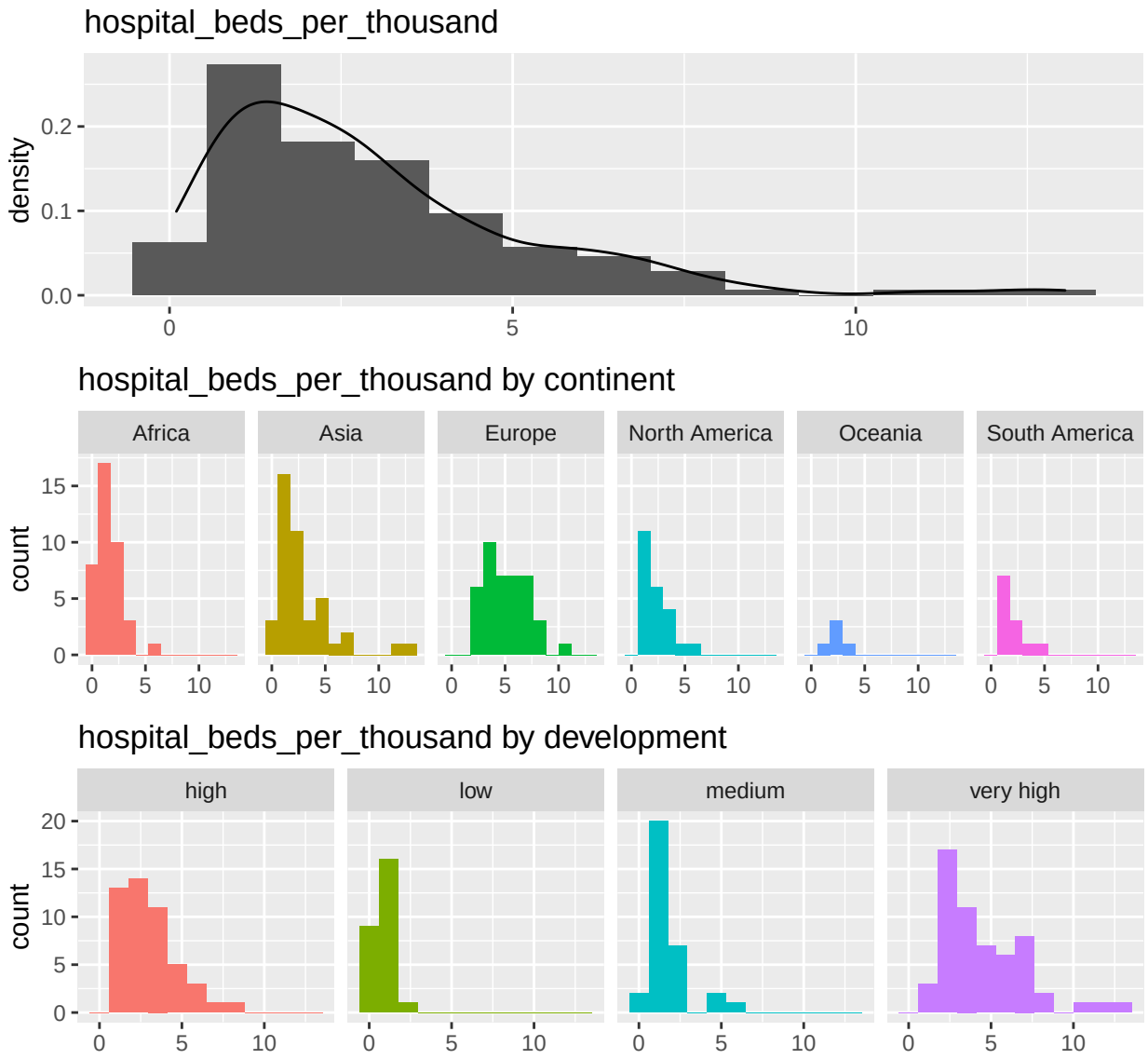


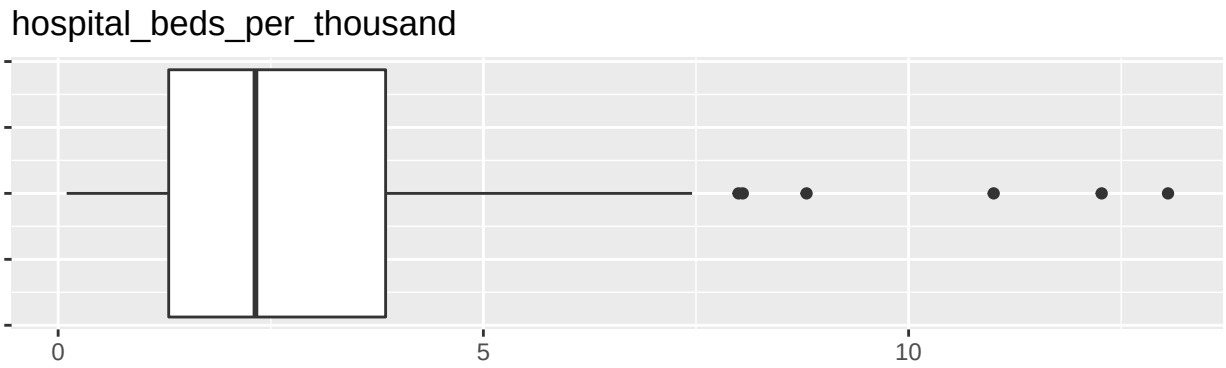

**Histogram and kernel density for diabetes prevalence**

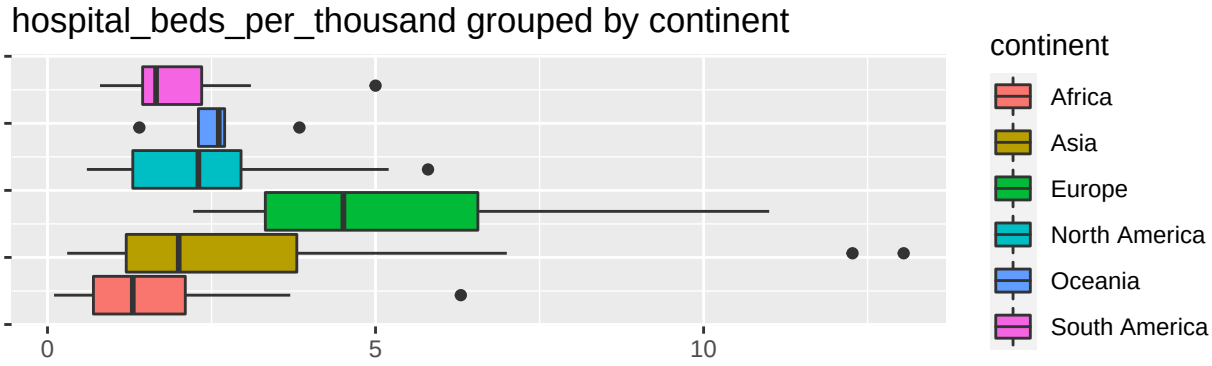
```
plots(dataset=data, col='hospital_beds_per_thousand',type='hist', density=FALSE, bins=c(13,12,12), xtic
```

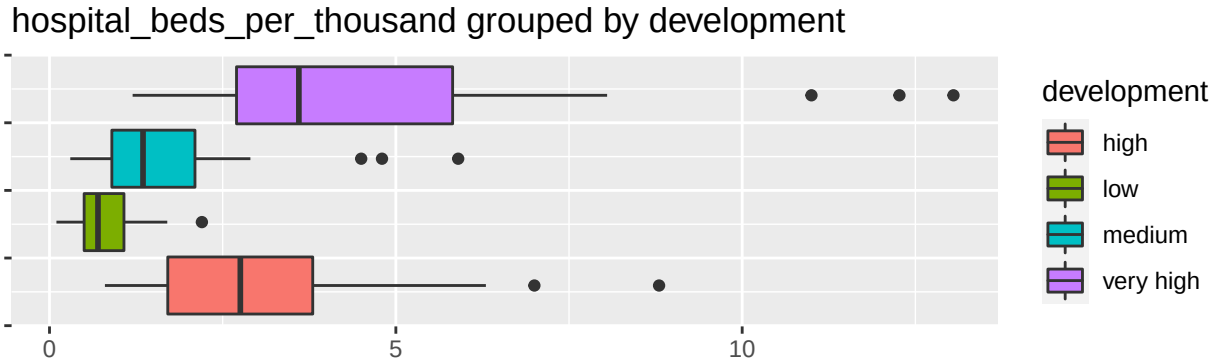



**Boxplots for hospital beds per thousand inhabitants**

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



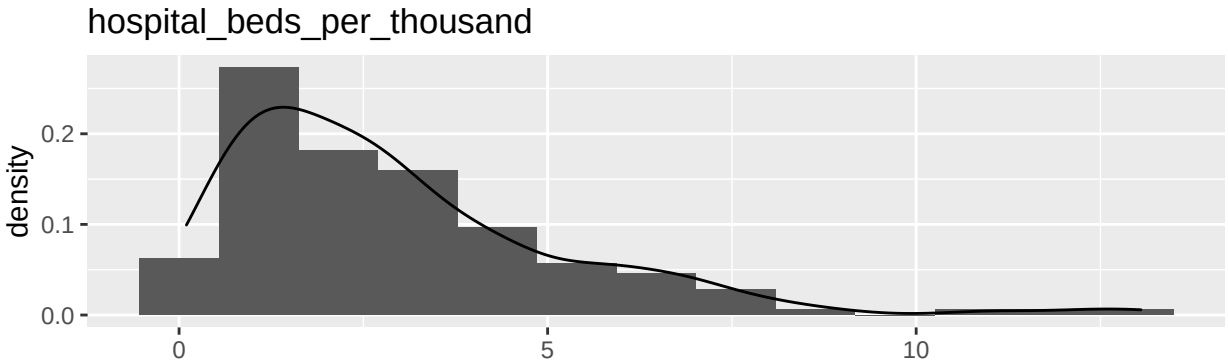

Looking at the hospital beds per thousand inhabitants variable boxplots we can see a few interesting things. We could use this variable as a measure of the quality of a healthcare system of a country. Where the higher the bed availability in hospitals is, the better the health system can cope with the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

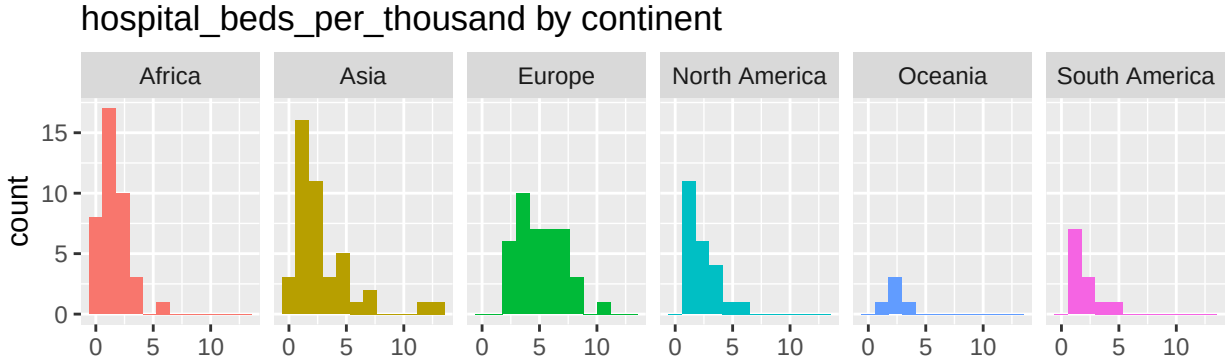
We can see that some extremely poor countries have about 0.1-0.3 beds per thousand inhabitants, like it is the case with Mali and Niger. Some other countries like South Korea or Belarus have an extremely high capacity, with around 12 and 11 beds per thousand inhabitants respectively. However, even if 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.

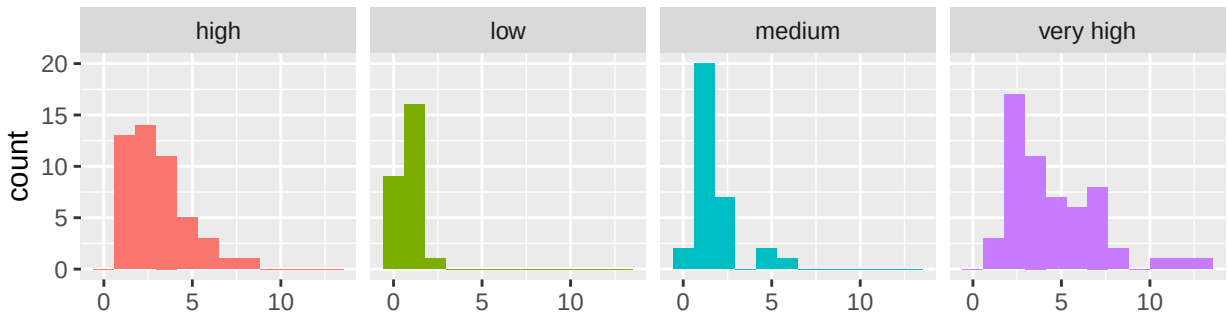
For countries with high and very high HDI, there's a clear bias towards having greater bed capacity, however, this is not the case for all countries with that quality as there's clearly some countries with medium HDI that have a quite formidable bed capacity as well.

**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
```



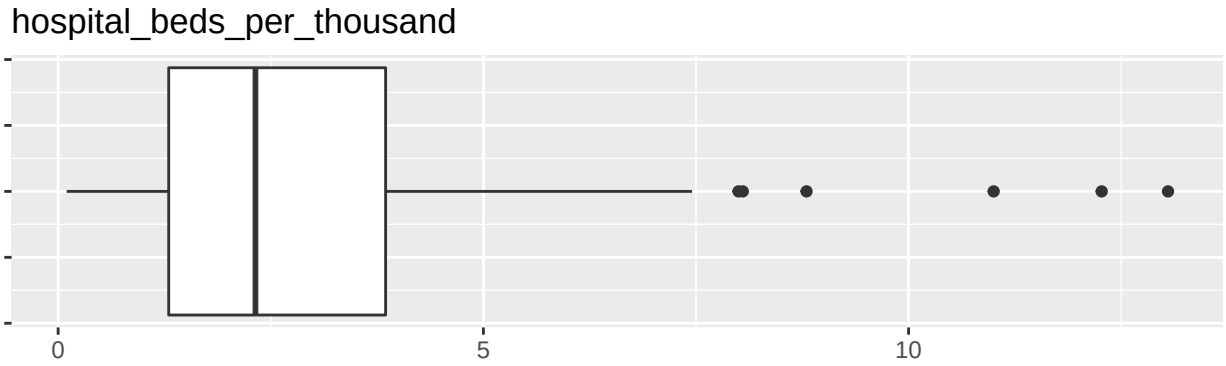

These plots tell a little bit of a different story to the boxplots. Where the largest concentration of countries is between 0 and 5 hospital beds per thousand inhabitants with an extremely scarce amount of countries with more than 10 beds per thousand inhabitants.

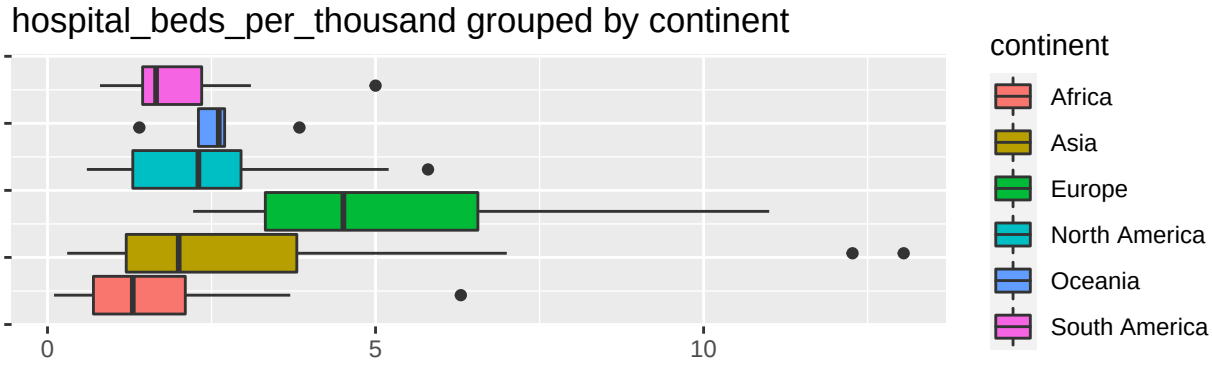
By segregating the data by continent we see that development does not necesarily mean greater healthcare capacity, with most continents boasting very similar numbers in this aspect whil esome like Asia, Africa, Europe and North america possessing some exceptions with extremely high numbers compred to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

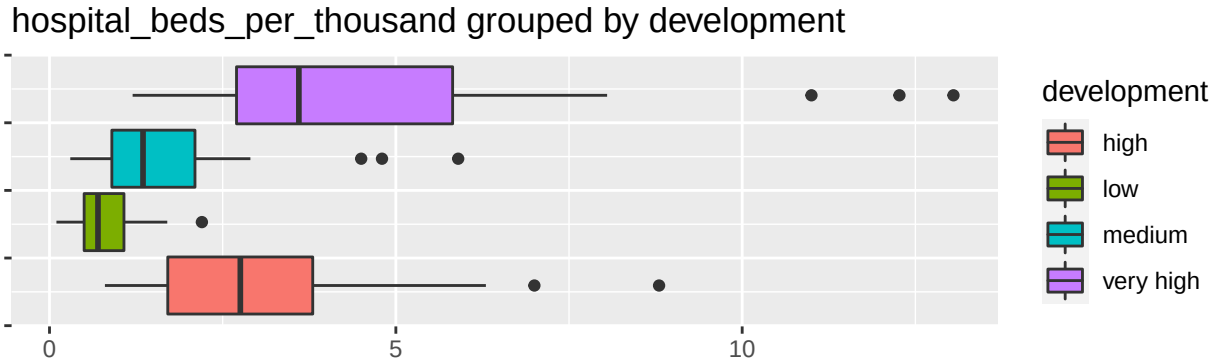
Finally, looking at development we see that it is rare for much less developed countries to have high bed capacity, while it is much easier for high to very high developed countries to have greater capacity. However, we can't confidently say that there's lots of exceptions to this 'rule'.

#### Boxplots for life expectancy

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



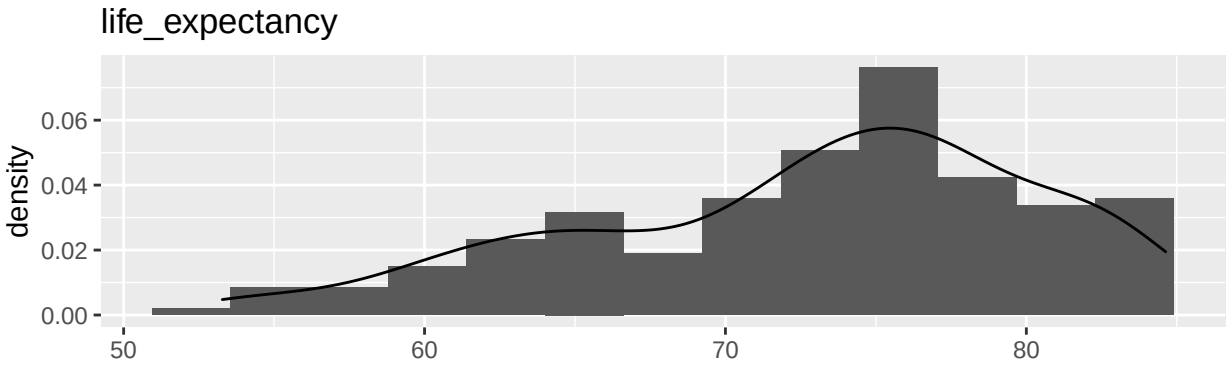

For life expectancy we can see moost countries sitting amove 66 years of age, with values going as low as 53.24 and as high as 84.63.

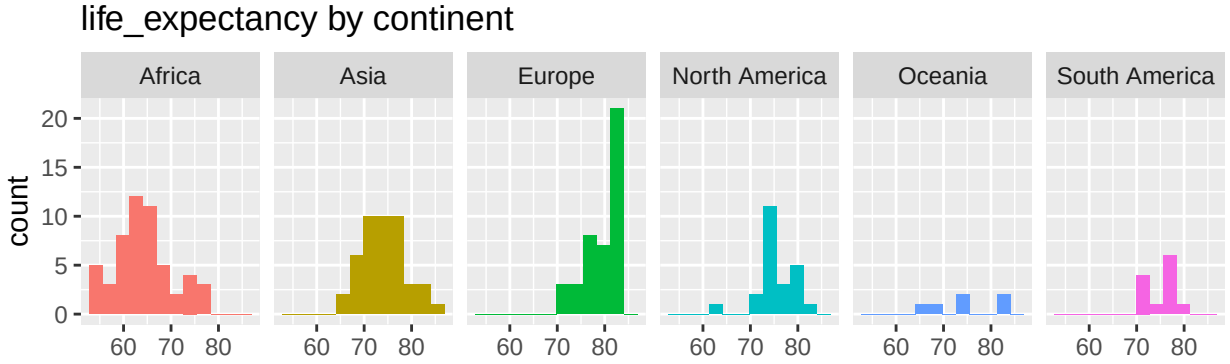
Africa has the lowest life expectancy while europe has the highest. The rest of the continents sit at roughly similar ranges.

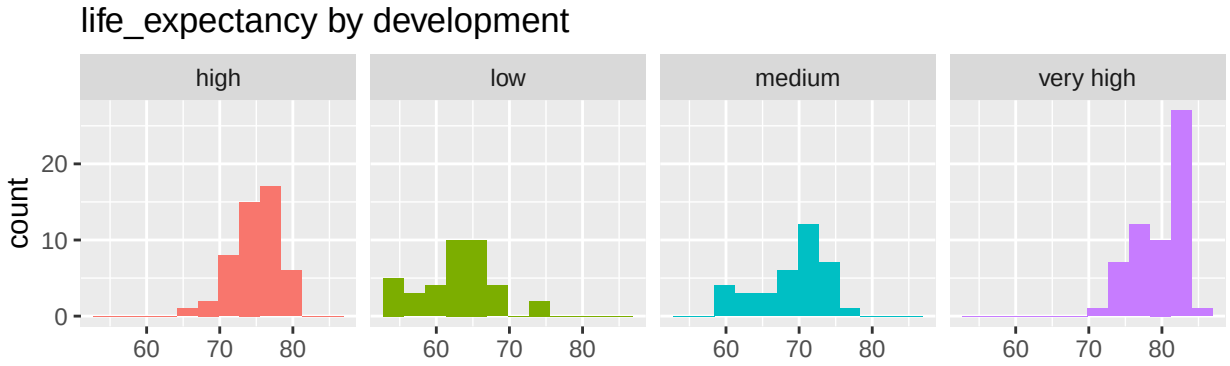
Grouping by HDI, we can see that the most developed countries have a significantly higher life expectancy than those with low HDI. It clearly shows a strong positive correlation between them. Where the higher the life expectancy the higher the HDI. With very few exceptions.

### 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(
```



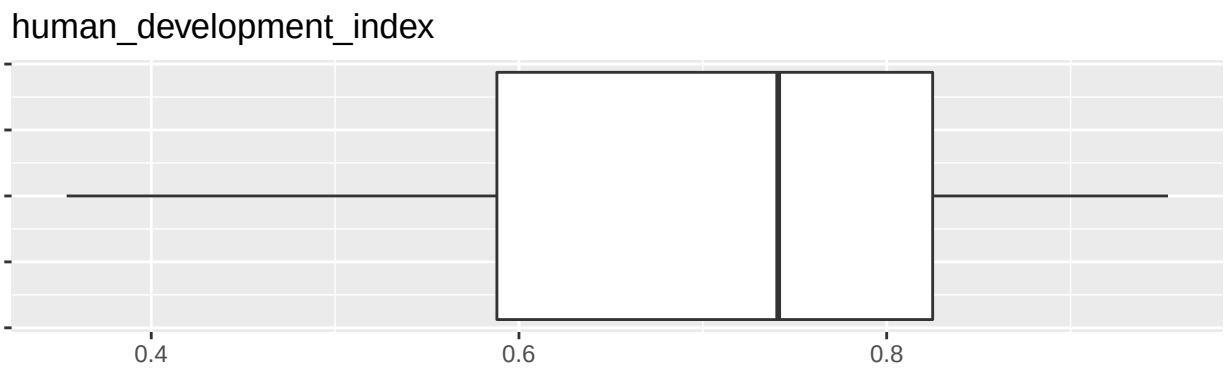

The general plot is somewhat left skewed, as most countries (about 80%) have a life expectancy higher than 65 years of age. Our density plot shows a strong concentration between 70 and 80 years of age, as this range covers the most nations.

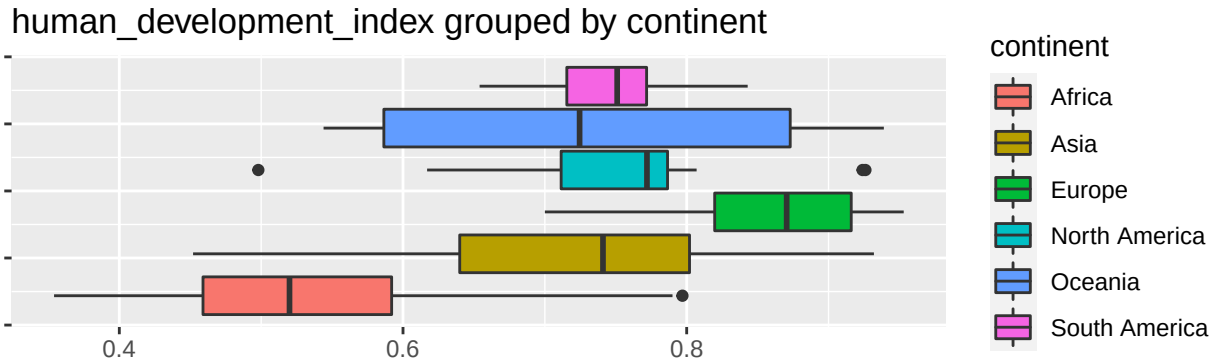
For each continent we see that europe shows a typically very high life expectancy while Africa shows a typically lower-than-average life expectancy for most countries with some exceptions. The rest of the continents sit at about the average life expectancy with some countries in Asia and North America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

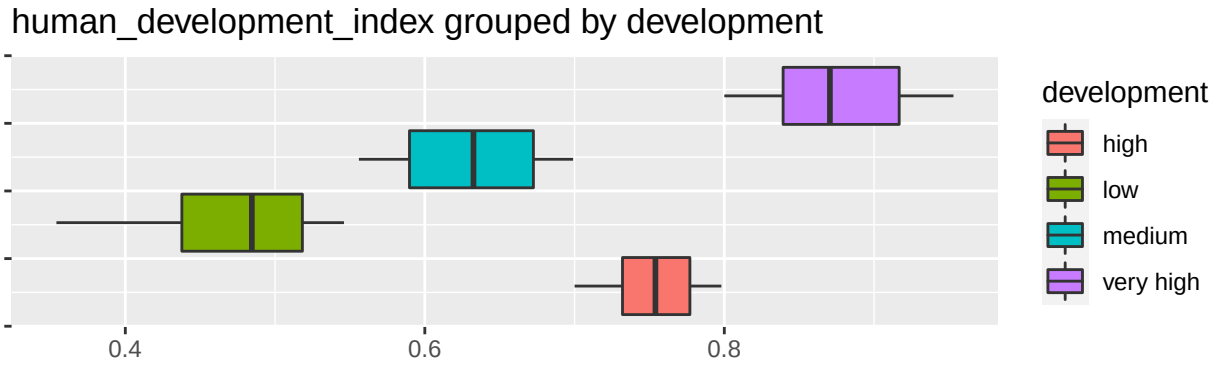
For HDI we can again see some of the strong correlation, where life expectancy for very highly developed nations seems to be also quite high and the same happens with less developed nations.

**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.

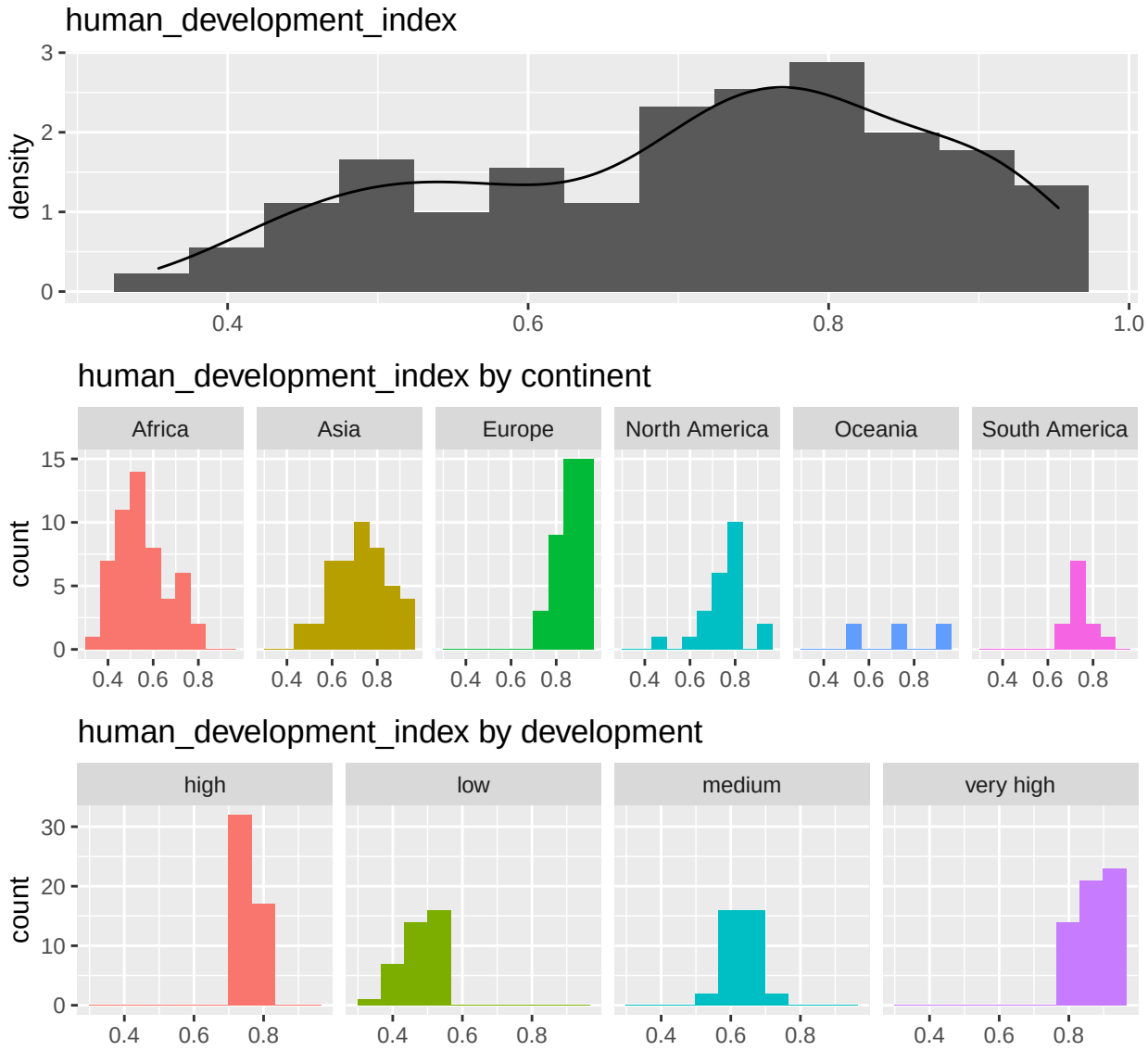
For continents we can see Africa lagging behind with most of its countries between 0.4 and 0.6 HDI, probably given the porverty situation in the continent.

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")
```