Final project: Step 1

Danyu Zhang, Limingrui Wan, Daniel Alonso

December 9th, 2020

Importing libraries

```
library(dplyr)
library(ggplot2)
library(reshape2)
library(PerformanceAnalytics)
```

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

```
columns_selected <- names(data)[names(data) != 'new_deaths_smoothed' & names(data) != 'new_cases_smooth
data_n <- data %>% select(all_of(columns_selected))
```

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

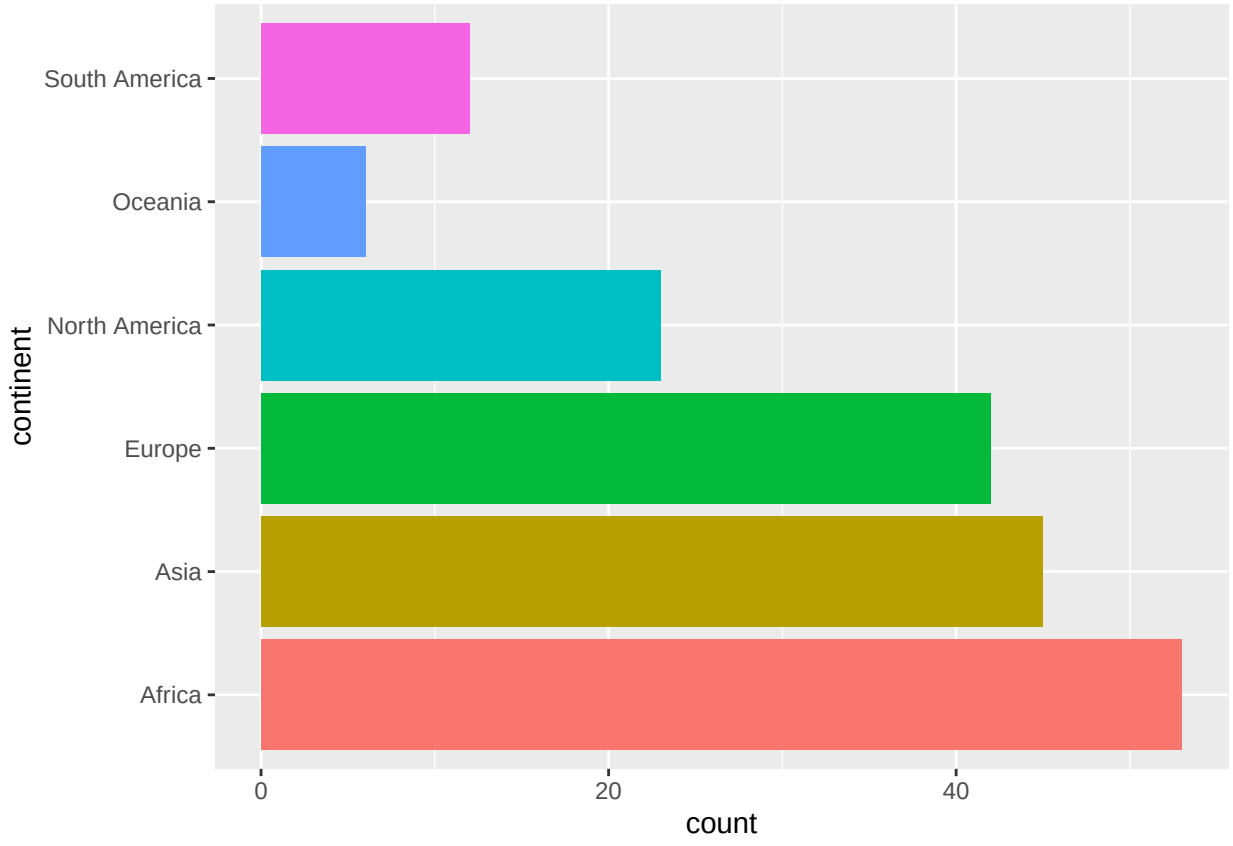
We select variables that we consider interesting to visualize, as the ones we haven't selected might be ralated to these or even ratios of them (in the case of total cases per million)

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

Plots with categorical variables

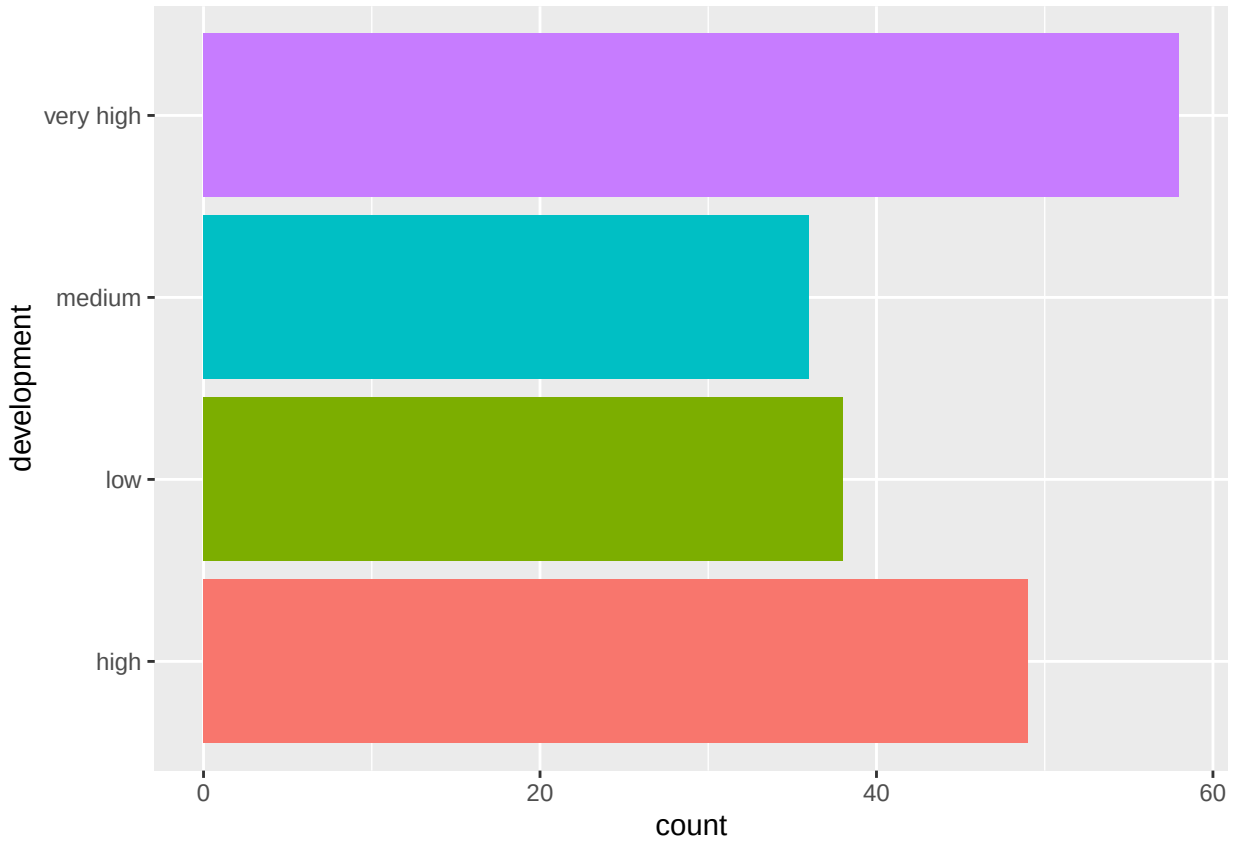
Countries per continent in the dataset

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



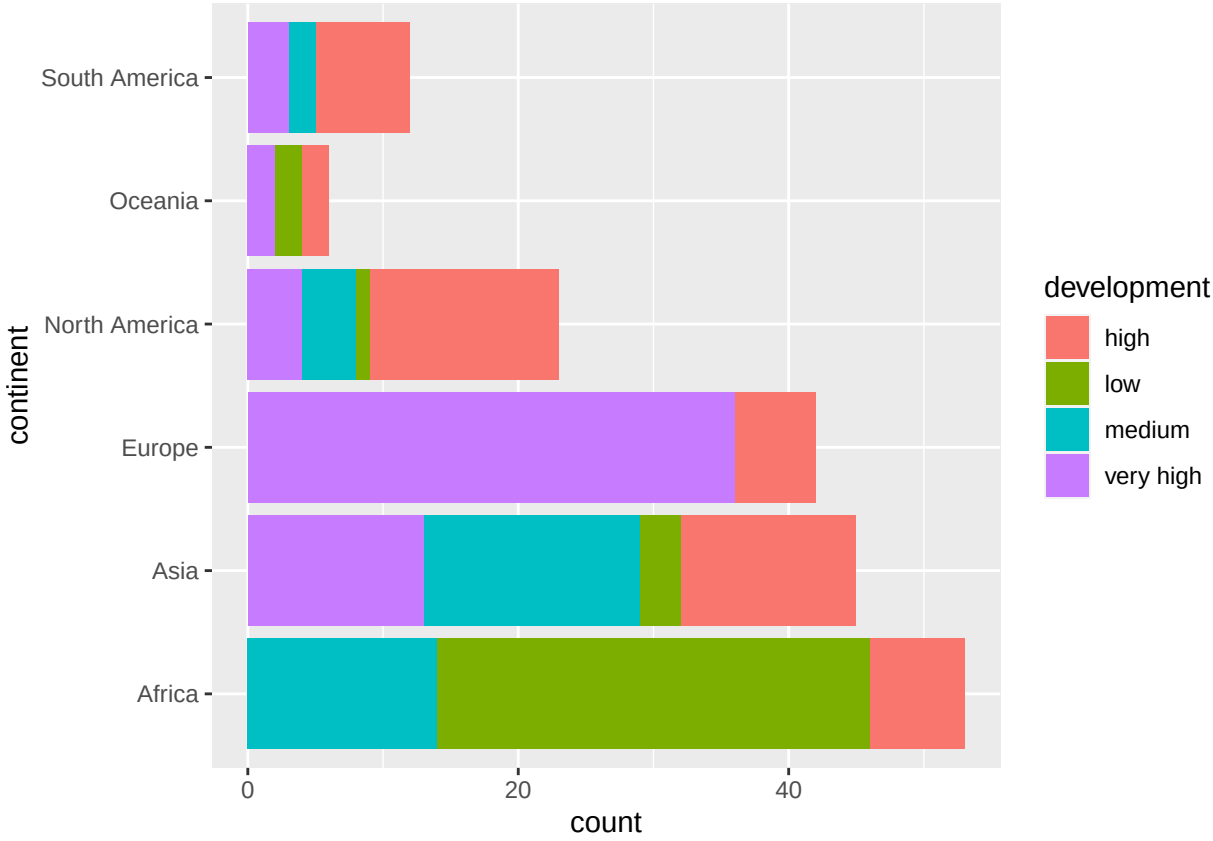
Amount of countries per HDI

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


very high
medium
low
high
development
0
20
40
60
count

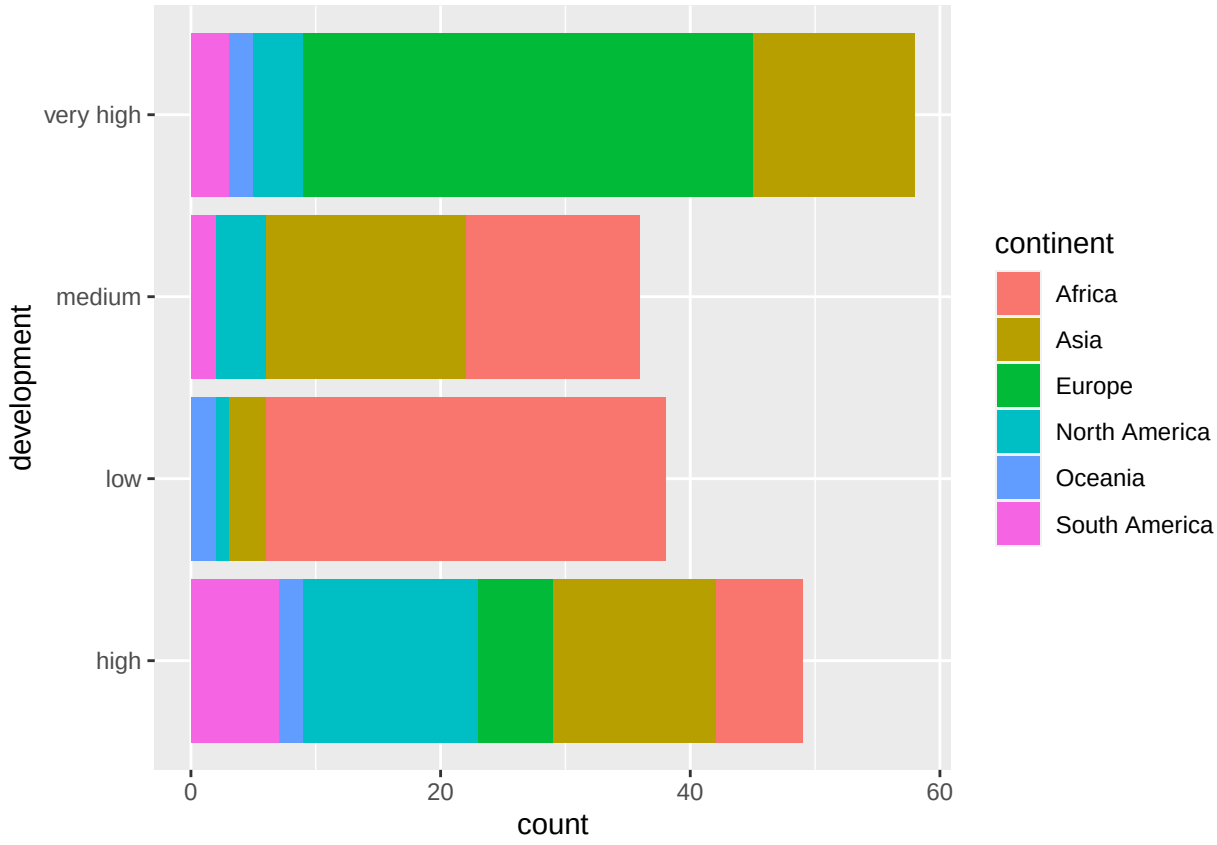
Countries per continent per HDI

```
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:

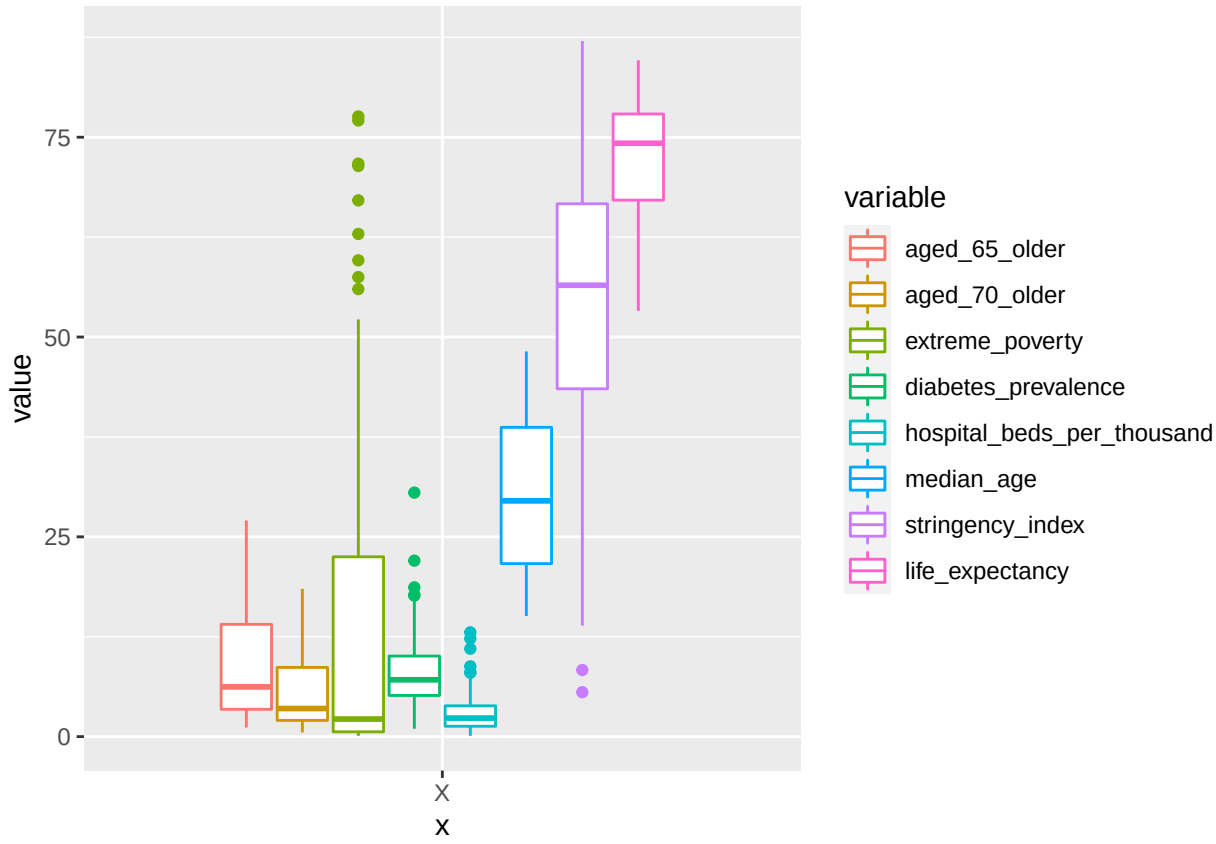
```
color_1 <- "khaki"
color_2 <- "lightseagreen"
color_3 <- "lightpink2"
```

Boxplots

Grouping variables with a max. value below 87 to show all in a single plot:

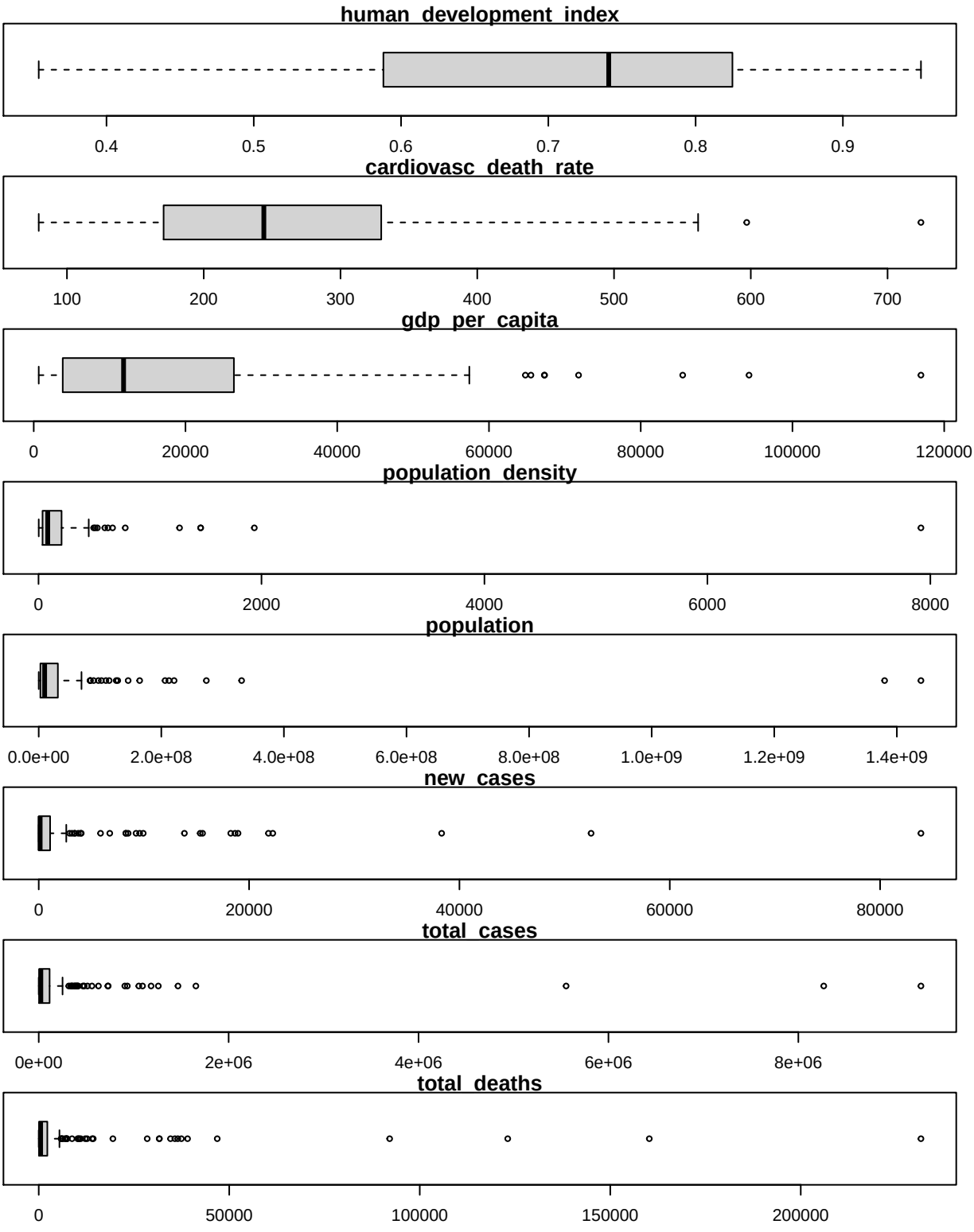
```
lower_than_87 <- c('aged_65_older','aged_70_older','extreme_poverty','diabetes_prevalence','hospital_bed
lower_than_87 = melt(data, id.vars='X', measure.vars=lower_than_87)

lower_than_87 %>%
    ggplot(aes(x="X", y=value)) +
    geom_boxplot(aes(color=variable))
```



Plotting the rest of the variables:

```
other_vars <- c('human_development_index','cardiovasc_death_rate','gdp_per_capita','population_density'
par(mfrow=c(length(other_vars),1), mar=c(2,1,1,1))
for (i in 1:length(other_vars)) {
    boxplot(data %>% select(other_vars[i]), horizontal=TRUE, main=other_vars[i])
}
```


human_development_index
0.4
0.5
0.6
0.7
0.8
0.9
cardiovasc_death_rate
100
200
300
400
500
600
700
gdp_per_capita
0
20000
40000
60000
80000
100000
120000
population_density
0
2000
4000
6000
8000
population
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
1.4e+09
new_cases
0
20000
40000
60000
80000
total_cases
0e+00
2e+06
4e+06
6e+06
8e+06
total_deaths
0
50000
100000
150000
200000

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