Final project: Step 1

Danyu Zhang, Limingrui Wan, Daniel Alonso

December 9th, 2020

Importing libraries

```
library(dplyr)
library(ggplot2)
```

Importing data

```
data <- read.csv('./data/data.csv')
head(data)
#>   X      continent                          location total_cases new_cases
#> 1 0 Asia           Afghanistan                            41728        95
#> 2 1 Africa         Angola                                 11035       230
#> 3 2 Europe         Albania                                21523       321
#> 4 3 Europe         Andorra                                 4888        63
#> 5 4 Asia           United Arab Emirates                  135141      1234
#> 6 5 South America  Argentina                            1183118      9598
#>   new_cases_smoothed total_deaths new_deaths new_deaths_smoothed
#> 1             99.429         1544          3               3.143
#> 2            236.286          286          2               2.571
#> 3            296.857          527          9               6.714
#> 4             80.429           75          0               0.429
#> 5           1272.429          497          1               2.429
#> 6          11547.143        31623        483             331.714
#>   total_cases_per_million new_cases_per_million new_cases_smoothed_per_million
#> 1                1071.918                 2.440                          2.554
#> 2                 335.755                 6.998                          7.189
#> 3                7478.977               111.544                        103.154
#> 4               63262.797               815.376                       1040.944
#> 5               13663.856               124.767                        128.653
#> 6               26177.623               212.365                        255.492
#>   total_deaths_per_million new_deaths_per_million stringency_index population
#> 1                   39.663                  0.077             5.56   38928341
#> 2                    8.702                  0.061               NA   32866268
#> 3                  183.126                  3.127            50.93    2877800
#> 4                  970.685                  0.000            59.26      77265
#> 5                   50.251                  0.101            47.22    9890400
#> 6                  699.689                 10.687            81.94   45195777
#>   population_density median_age aged_65_older aged_70_older gdp_per_capita
#> 1             54.422       18.6         2.581         1.337       1803.987
#> 2             23.890       16.8         2.405         1.362       5819.495
#> 3            104.871       38.0        13.188         8.643      11803.431
#> 4            163.755         NA            NA            NA             NA
#> 5            112.442       34.0         1.144         0.526      67293.483
#> 6             16.177       31.9        11.198         7.441      18933.907
#>   extreme_poverty cardiovasc_death_rate diabetes_prevalence
#> 1              NA               597.029                9.59
#> 2              NA               276.045                3.94
#> 3             1.1               304.195               10.08
#> 4              NA               109.135                7.97
#> 5              NA               317.840               17.26
#> 6             0.6               191.032                5.50
#>   hospital_beds_per_thousand life_expectancy human_development_index
#> 1                       0.50           64.83                   0.498
#> 2                         NA           61.15                   0.581
#> 3                       2.89           78.57                   0.785
#> 4                         NA           83.73                   0.858
#> 5                       1.20           77.97                   0.863
#> 6                       5.00           76.67                   0.825
#>   development
#> 1         low
#> 2      medium
#> 3        high
#> 4   very high
#> 5   very high
#> 6   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_smoothe
data <- 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

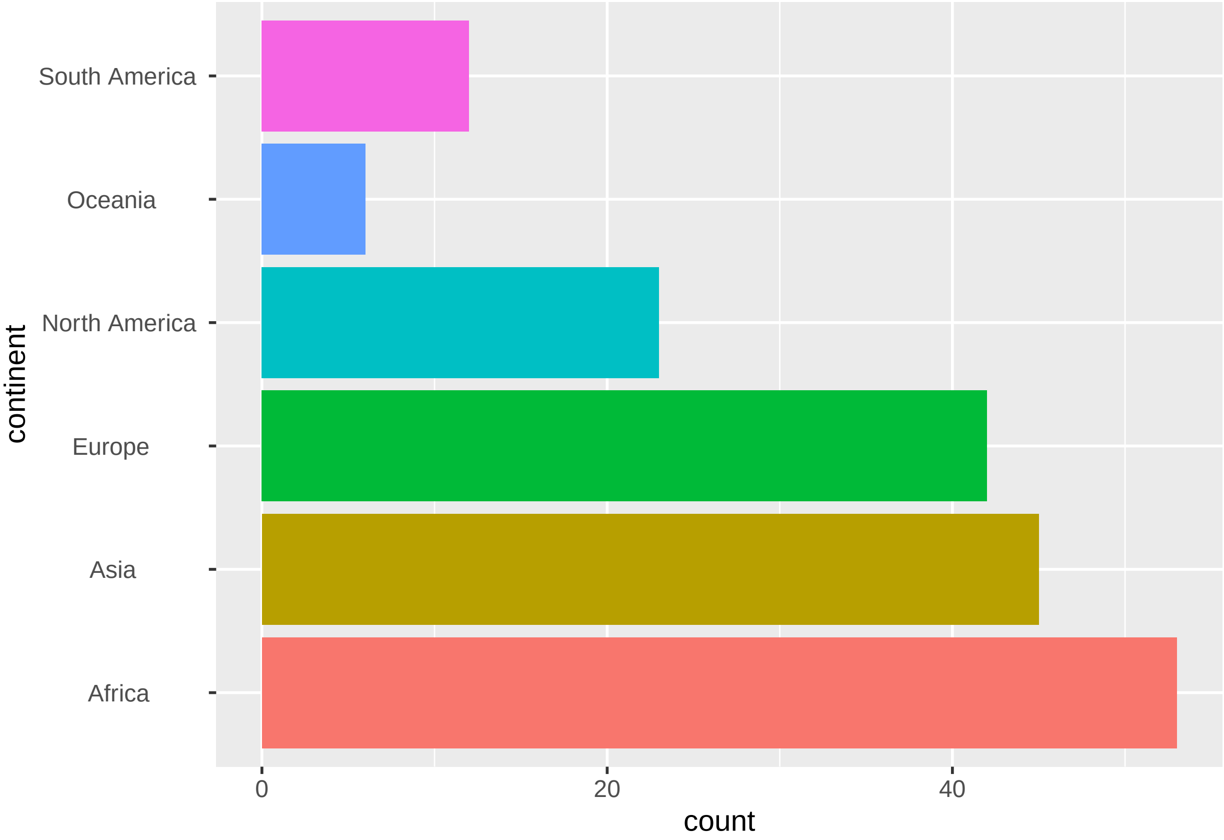
We select variables that we consider interesting to visualize, as the ones we haven't selected might be ralated to these or even ratios of them (in the case of total cases per million)

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

Plots with categorical variables

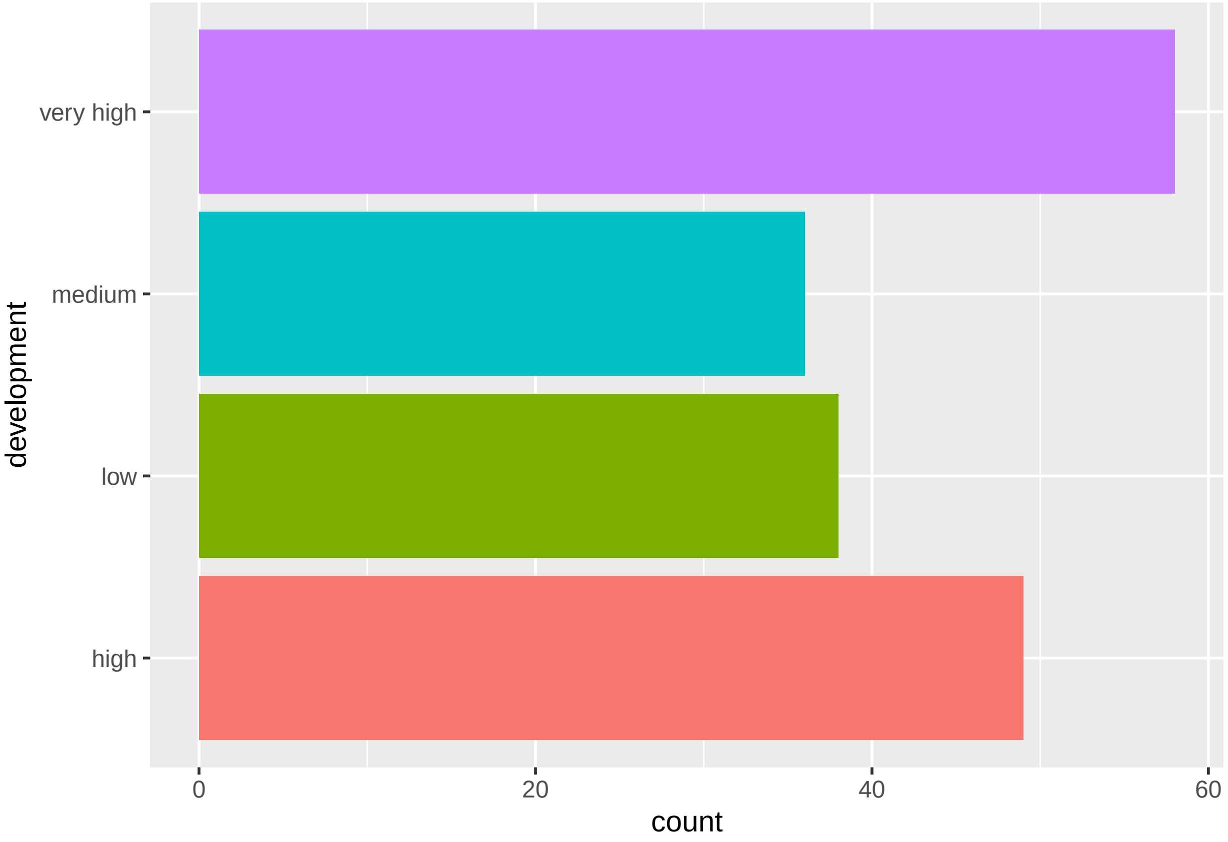
Countries per continent in the dataset

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



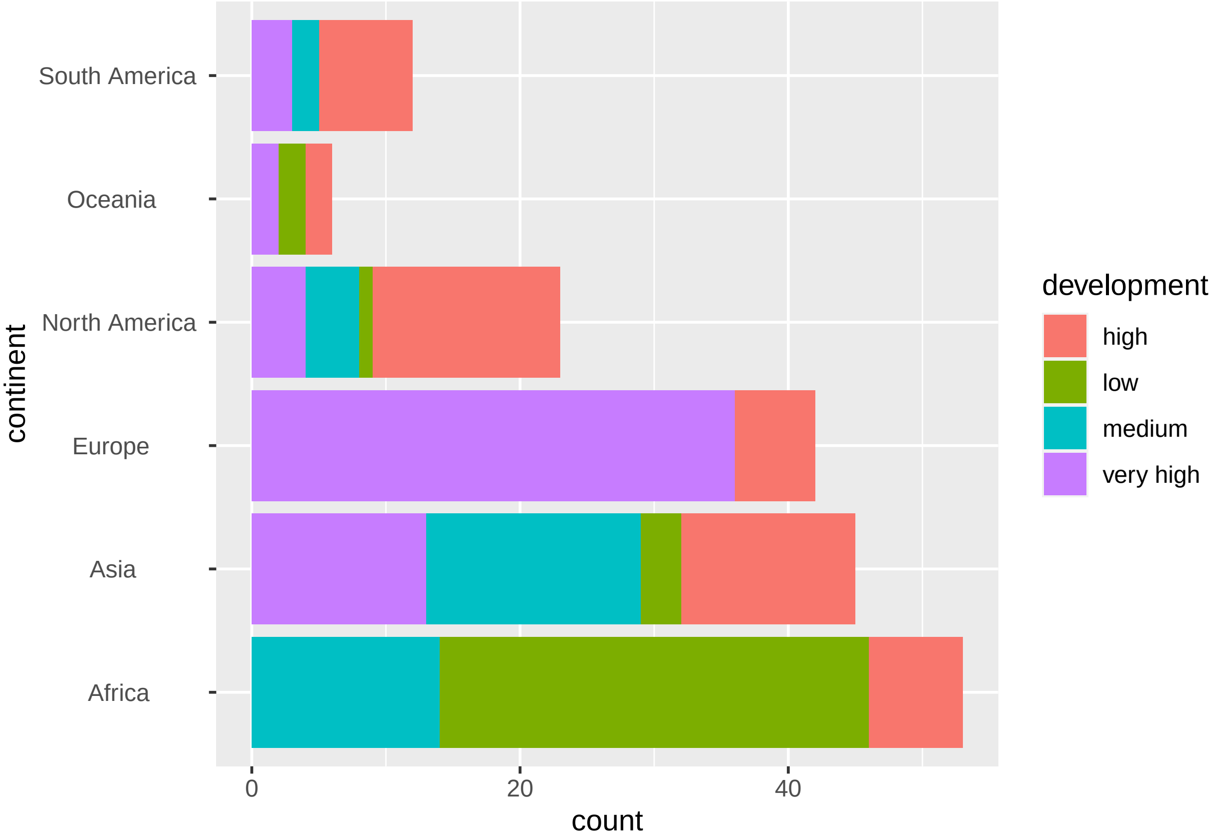
Amount of countries per HDI

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



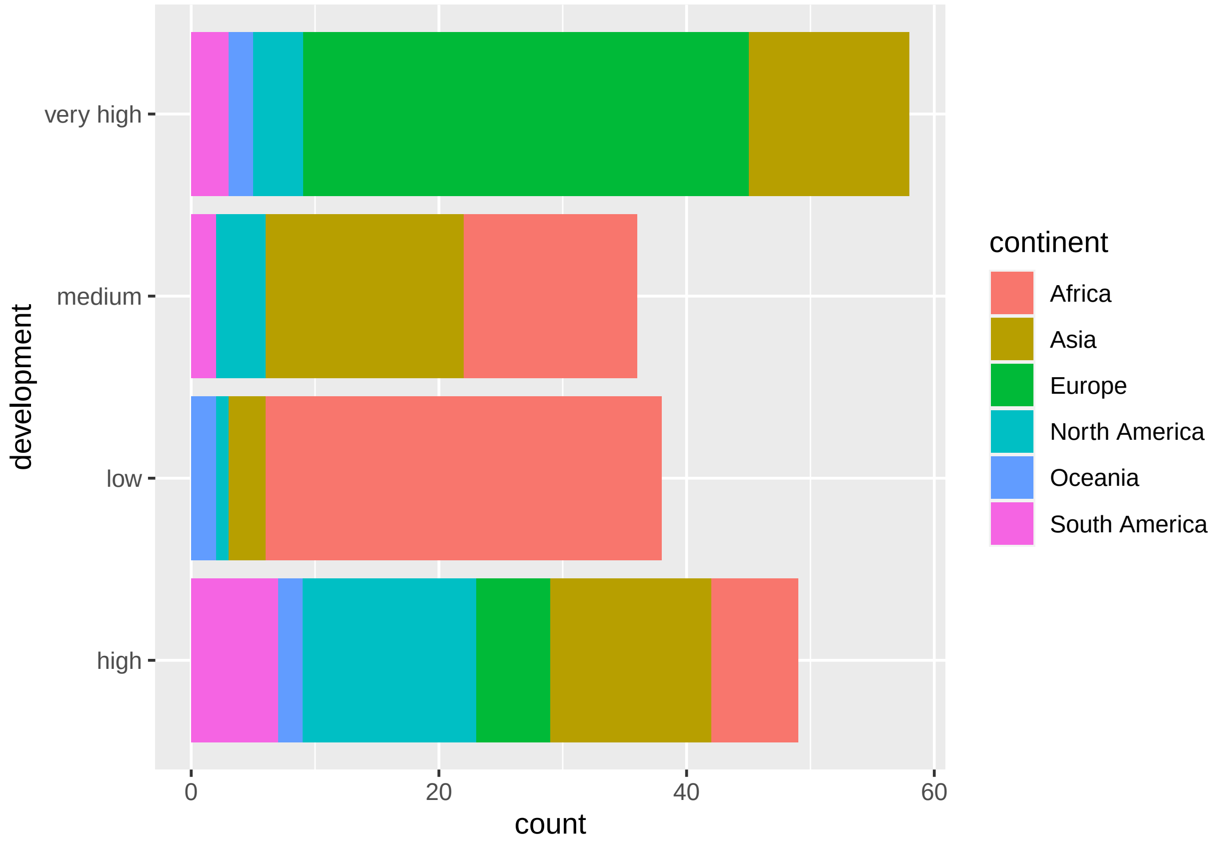
Countries per continent per HDI

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



Proportions of HDI per continent

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


very high
medium
low
high
development
0
20
40
60
count
continent
Africa
Asia
Europe
North America
Oceania
South America