explanation

September 21, 2021

MA615 FALL 2021

HW#2

- Wrangling
- Visualization

Wrangling This is the first part of this homework, called wrangling. The data sets are from the Gapminder website (https://www.gapminder.org/data/). Pick two individual indicators and download each of their data sets into .cvs data files. Then, wrangle these two data files into one tibble that is "tidy".

I choose the adult literacy rate and adult employment rate to find the relationship between this two indicators.

Literacy: Adult literacy rate is the percentage of people ages 15 and above who can, with understanding, read and write a short, simple statement on their everyday life.

Employment: Percentage of total population, age group 15+, that has been employed during the given year.

```
## upload two individual indicators

Literacy_org <- read.csv("/Users/odd/Desktop/FALL2021/MA615/MA615-HW-2/data/literacy_rate_adult_total_pe

employment_org <- read.csv("/Users/odd/Desktop/FALL2021/MA615/MA615-HW-2/data/aged_15plus_employment_rat
```

```
## select the adult literacy rate and adult employment rate
Literacy1 <- Literacy_org[, c(1,28)]
sum(is.na(Literacy1$X2000))       # see how many NAs are on the dataset
```

Import datasets

```
## [1] 126
```

```
Literacy_final <- na.omit(Literacy1)
Literacy_final$literacy <- Literacy_final$X2000
# list the literacy rate in 2000 for each country

literacy <- Literacy_final[, c(1,3)]    # divide
head(literacy)      # look the first 6 rows of the data frame
```

```
##                country literacy
## 3               Angola     67.4
## 5              Albania     98.7
## 8            Argentina     97.2
## 9              Armenia     99.4
## 10 Antigua and Barbuda     99.0
## 15          Bangladesh     47.5
```

```
employment_final <- employment_org[, c(1,12)]
sum(is.na(employment_final$X2000))       # see how many NAs are on the dataset
```

```
## [1] 0
```

```
employment_final$employment <- employment_final$X2000
# list the employment rate in 2000 for each country

Employment <- employment_final[, c(1,3)]
head(Employment)       # look the first 6 rows of the data frame
```

```
##                country employment
## 1          Afghanistan       45.9
## 2               Angola       59.3
## 3              Albania       48.0
## 4 United Arab Emirates       73.5
## 5            Argentina       51.0
## 6              Armenia       49.1
```

```

## create tibbles
as_tibble(literacy)
```

Create tibbles

```
## # A tibble: 30 x 2
##    country             literacy
##    <fct>                  <dbl>
##  1 Angola                  67.4
##  2 Albania                 98.7
##  3 Argentina               97.2
##  4 Armenia                 99.4
##  5 Antigua and Barbuda     99
##  6 Bangladesh              47.5
##  7 Bulgaria                98.2
##  8 Bahrain                 86.5
##  9 Bolivia                 86.7
## 10 Brunei                  92.7
## # ... with 20 more rows
```

```
as_tibble(Employment)
```

```
## # A tibble: 189 x 2
##    country              employment
##    <fct>                     <dbl>
##  1 Afghanistan                45.9
##  2 Angola                     59.3
##  3 Albania                    48
##  4 United Arab Emirates       73.5
##  5 Argentina                  51
##  6 Armenia                    49.1
##  7 Australia                  59.3
##  8 Austria                    55.6
##  9 Azerbaijan                 56.4
## 10 Burundi                    81.7
## # ... with 179 more rows
## combine two tibbles
trend_org <- left_join(literacy, Employment)
```

Combine two tibbles

```
## Joining, by = "country"
```

```
trend <- na.omit(trend_org)
trend       # look at the tibble "trend"
```

```
##                    country literacy employment
## 1                   Angola    67.40       59.3
## 2                  Albania    98.70       48.0
## 3                Argentina    97.20       51.0
## 4                  Armenia    99.40       49.1
## 6               Bangladesh    47.50       55.6
## 7                 Bulgaria    98.20       41.4
## 8                  Bahrain    86.50       65.0
## 9                  Bolivia    86.70       67.0
## 10                  Brunei    92.70       63.9
## 11       Congo, Dem. Rep.    67.20       69.6
## 12                  Cyprus    96.80       59.6
## 13                 Ecuador    91.00       60.5
## 14                  Greece    96.00       46.9
## 15                Honduras    80.00       61.7
## 16                 Croatia    98.20       44.9
## 17                   India    61.00       56.7
## 18                   Italy    98.40       43.4
## 19                     Lao    68.70       78.5
## 20               Sri Lanka    90.70       52.3
## 21               Lithuania    99.70       48.8
## 22            Macao, China    91.30       61.7
## 23                 Namibia    85.00       45.2
## 24                   Niger     9.39       77.6
## 25               Nicaragua    76.70       57.9
## 26                   Nepal    48.60       84.0
## 27 Sao Tome and Principe    84.90       47.0
## 28             Timor-Leste    37.60       52.0
## 29                 Ukraine    99.40       49.8
## 30               Venezuela    93.00       55.9
```

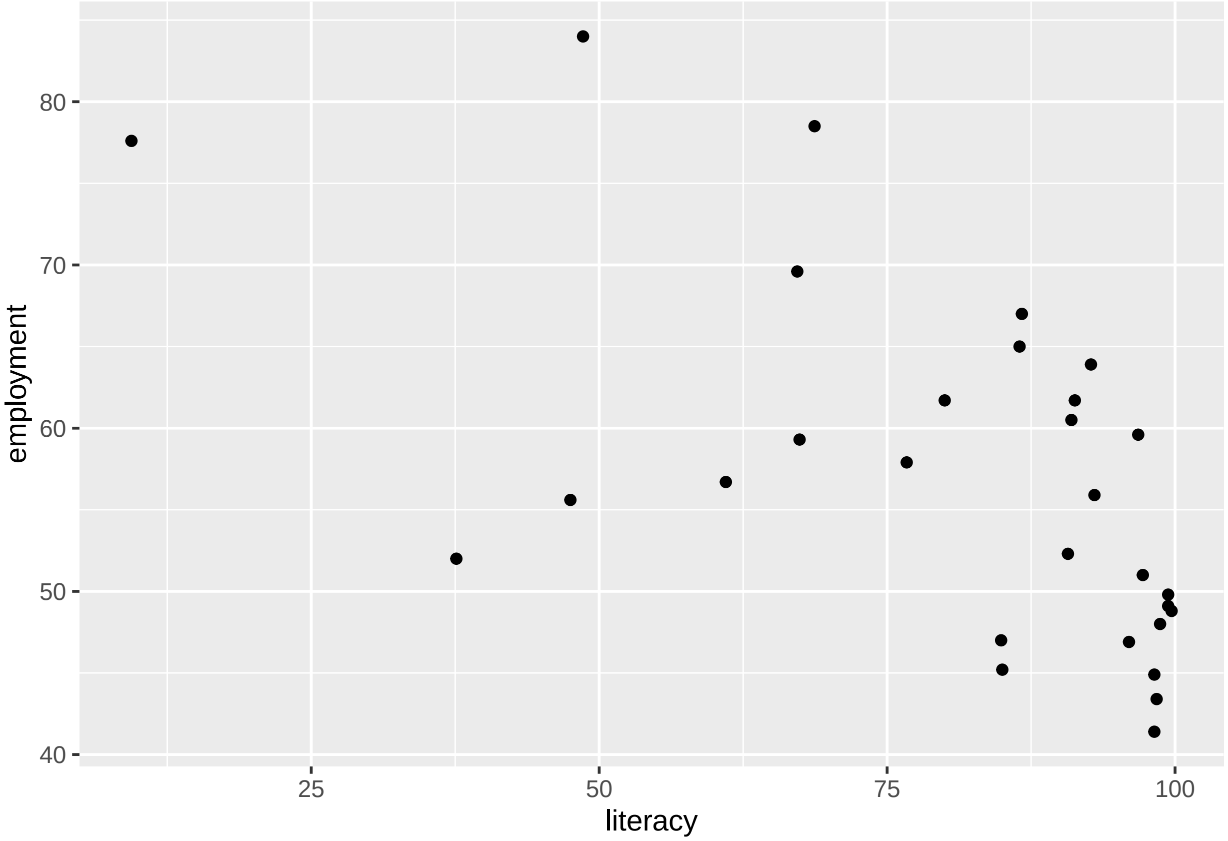
Creating plot Before we do the visualization, we can guess whether the higher the adult literacy rate is, the higher the adult employment rate will be. Since the majority of jobs require people understanding, read and write a short, simple statement, which is the basic requirement, this guess is reasonable.

Now, Create a ggplot visualization of this data.

```
# library packages
library(ggplot2)
library(tidyverse)
library(gapminder)

# library(printr)
library(RColorBrewer) ## to chose different colors for the graph
```

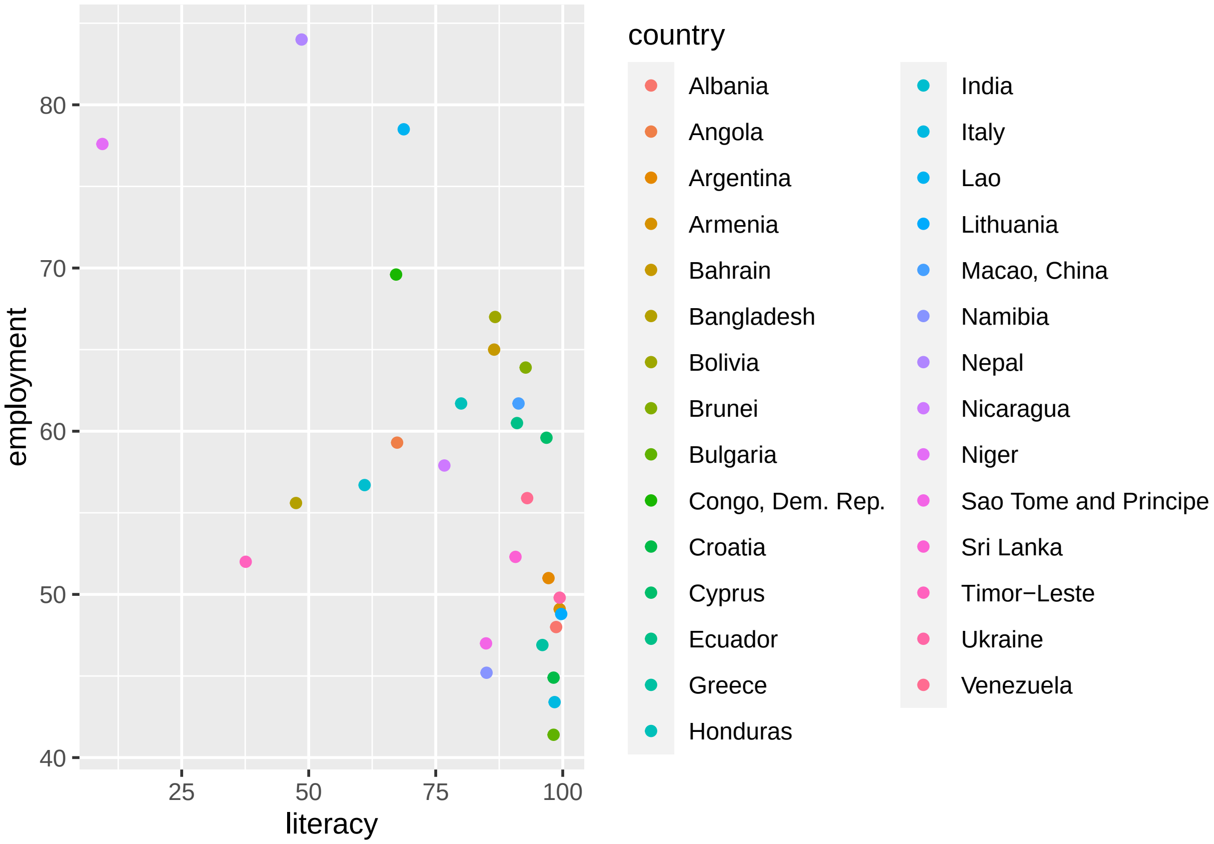
ggplot Let's turn this code into a reusable template for making graphs with ggplot2. ggplot(data =) + (mapping = aes())

```
## To plot trend, run this code to put literacy on the x-axis and
## Employment on the y-axis
ggplot(data = trend) +
  geom_point(mapping = aes(x = literacy, y = employment))
```



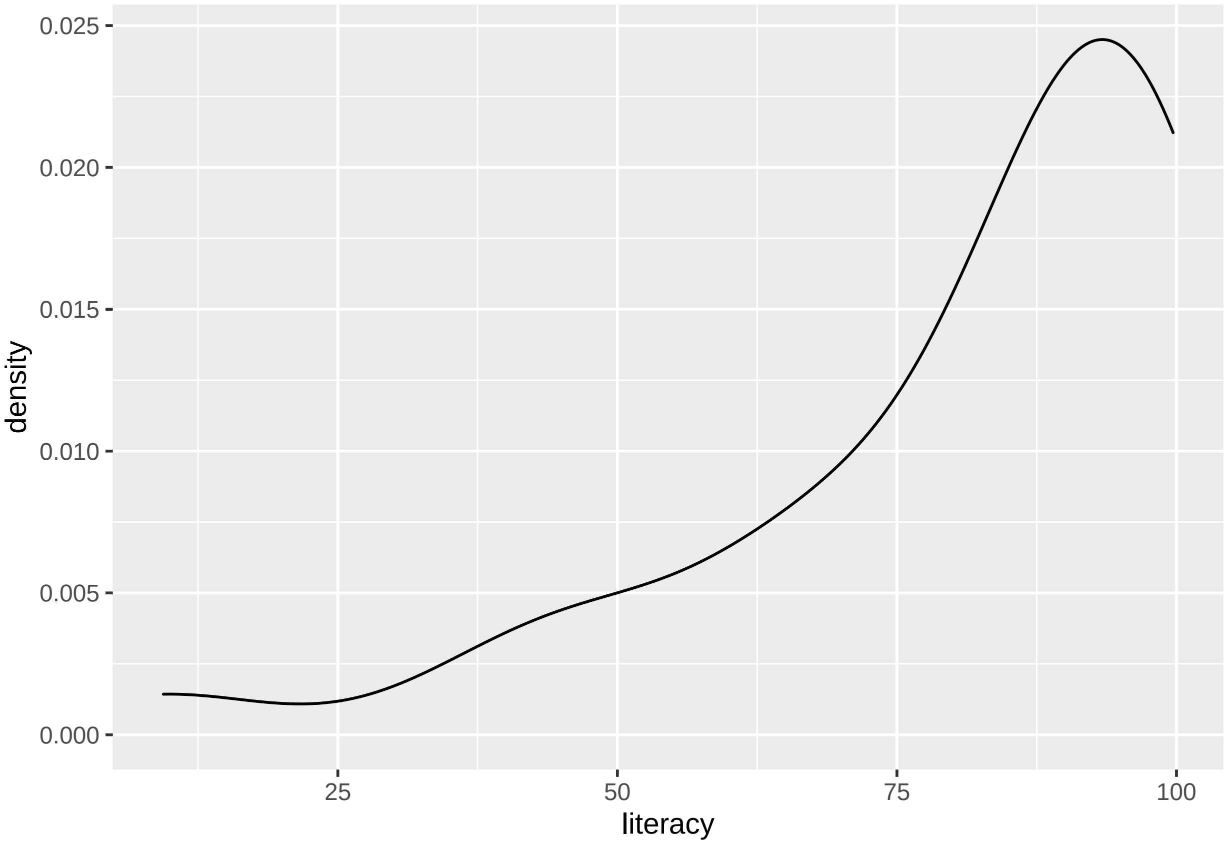
```
## The plot shows a positive relationship between literacy rate (literacy) and
## employment rate (employment) with each country. In other words, high literacy
## rate leads high employment rate.
```

```
ggplot(data = trend) +
  geom_point(mapping = aes(x = literacy, y = employment, color = country))
```

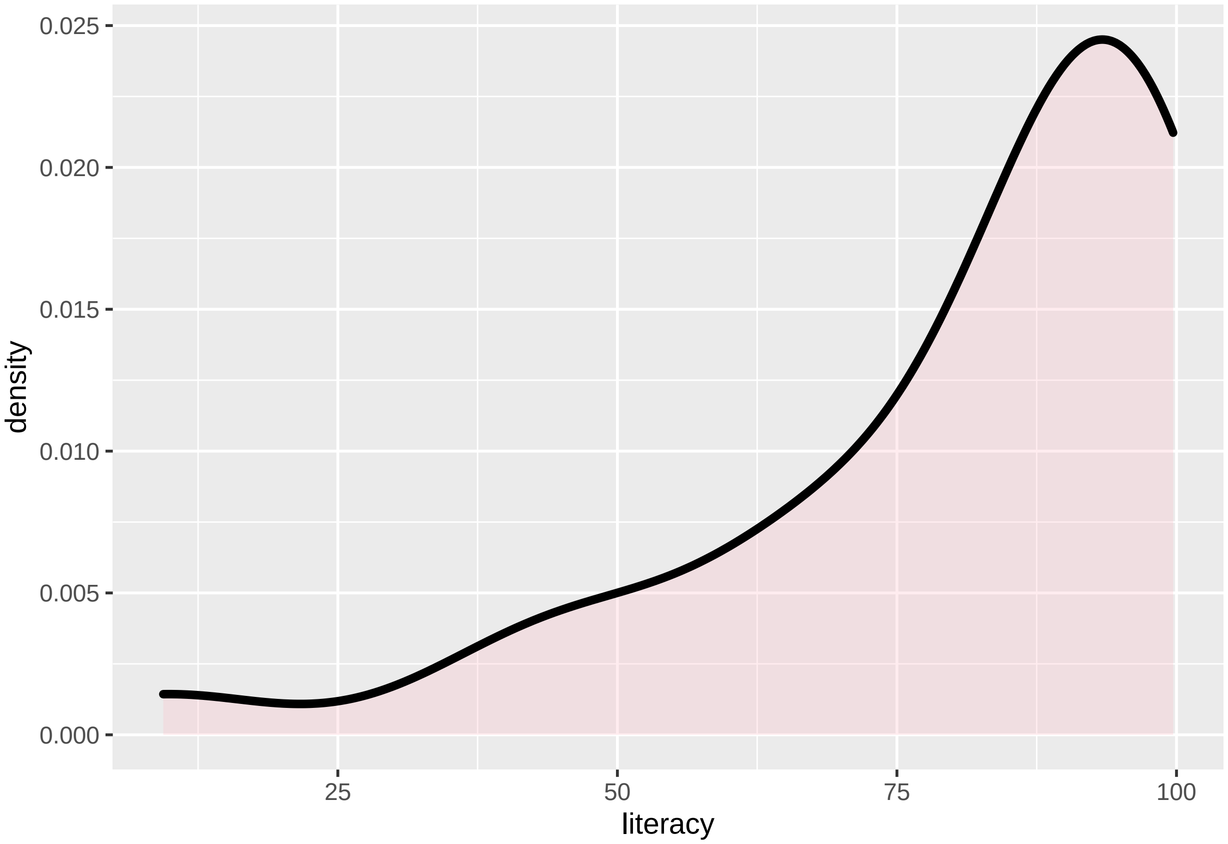


```
## map the colors of points to the class variable to reveal the relationship
## between literacy rate (literacy) and employment rate (employment)
## for each country.
```

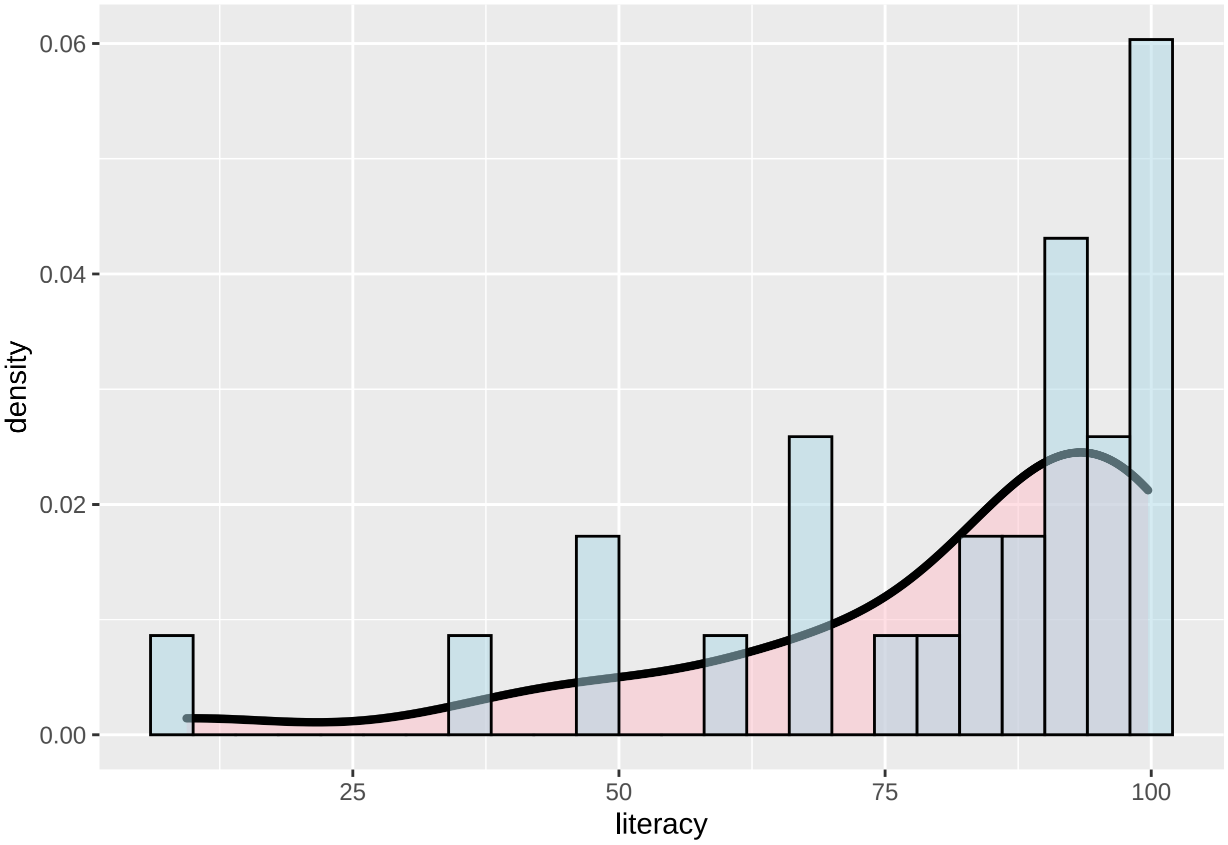
```
## exploring continuous variales  -- distributions
ggplot(data=trend, aes(x=literacy)) +
  geom_density()
```



```
ggplot(data=trend, aes(x=literacy)) +
  geom_density(size=1.5, fill="pink", alpha=0.3)
```



```
ggplot(data=trend, aes(x=literacy)) +
  geom_density(size=1.5, fill="pink", alpha=0.5) +
  geom_histogram(aes(y=..density..), binwidth=4, color="black", fill="lightblue", alpha=0.5)
```

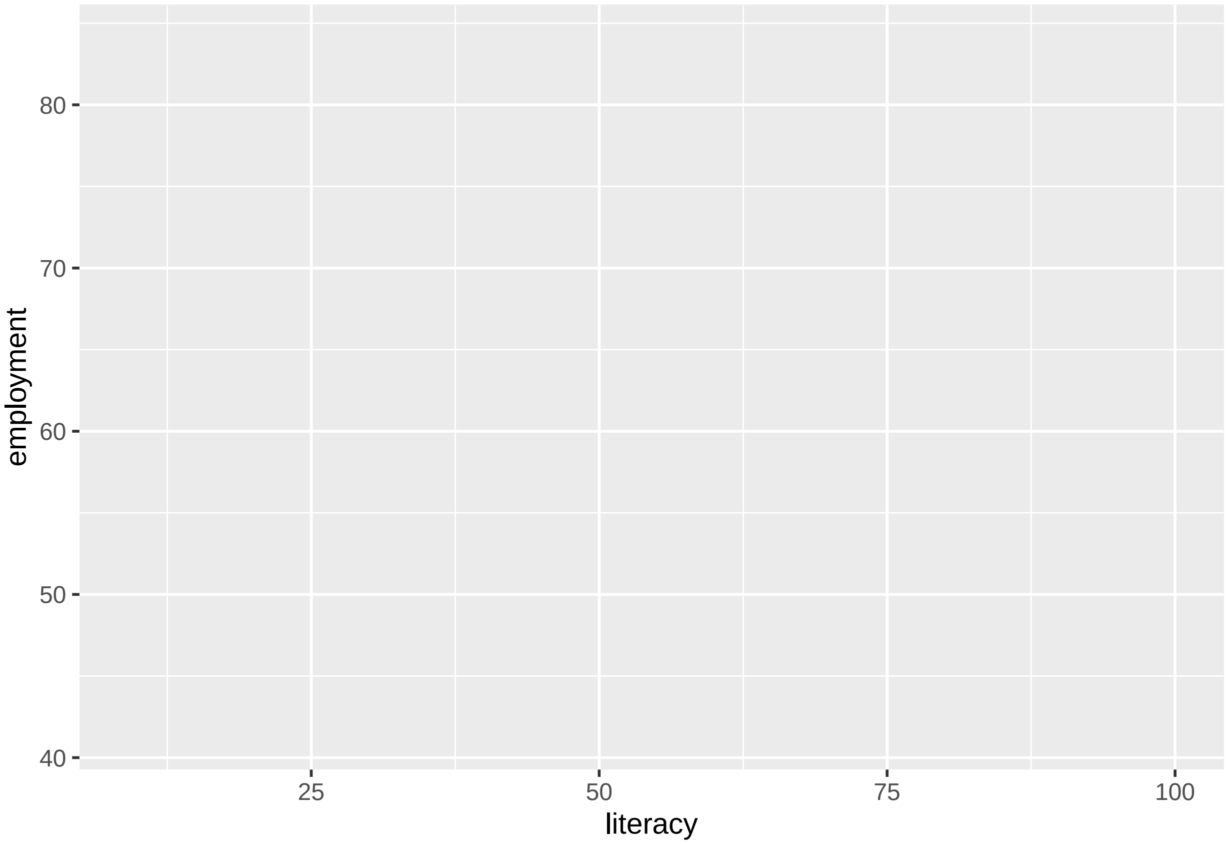


```
geom_histogram(aes(y=..density..), binwidth=4, color="black", fill="lightblue", alpha=0.5)
```

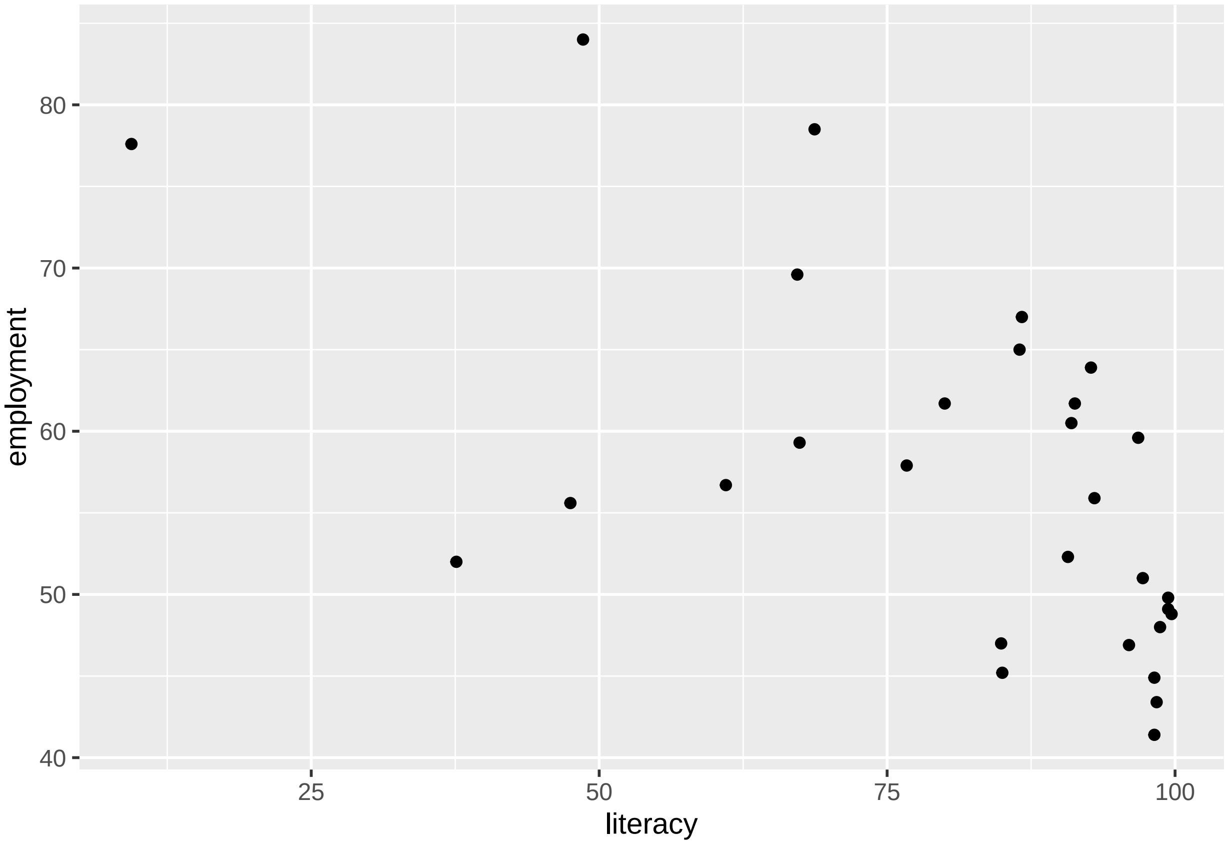
```
## mapping: y = ~..density..
## geom_bar: na.rm = FALSE, orientation = NA
## stat_bin: binwidth = 4, bins = NULL, na.rm = FALSE, orientation = NA, pad = FALSE
## position_stack
```

```
## In the majority of those countries, the employment rate has a positive
## relationship with the literacy rate. We should consider that the first
## requirment to get a job is recognize the words.
```

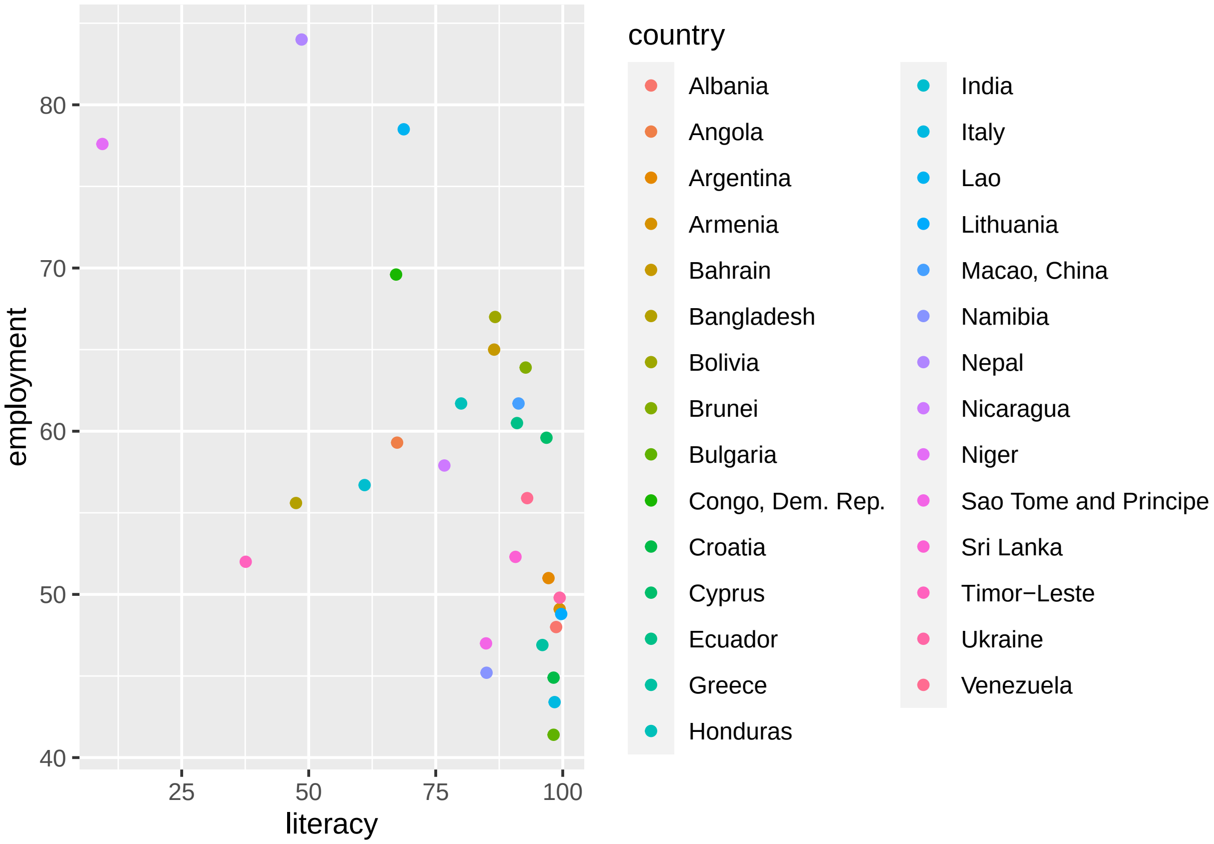
```
##  using layers

plt <- ggplot(data=trend,
              aes(x=literacy, y=employment))
plt
```



```r
plt + geom_point()
```

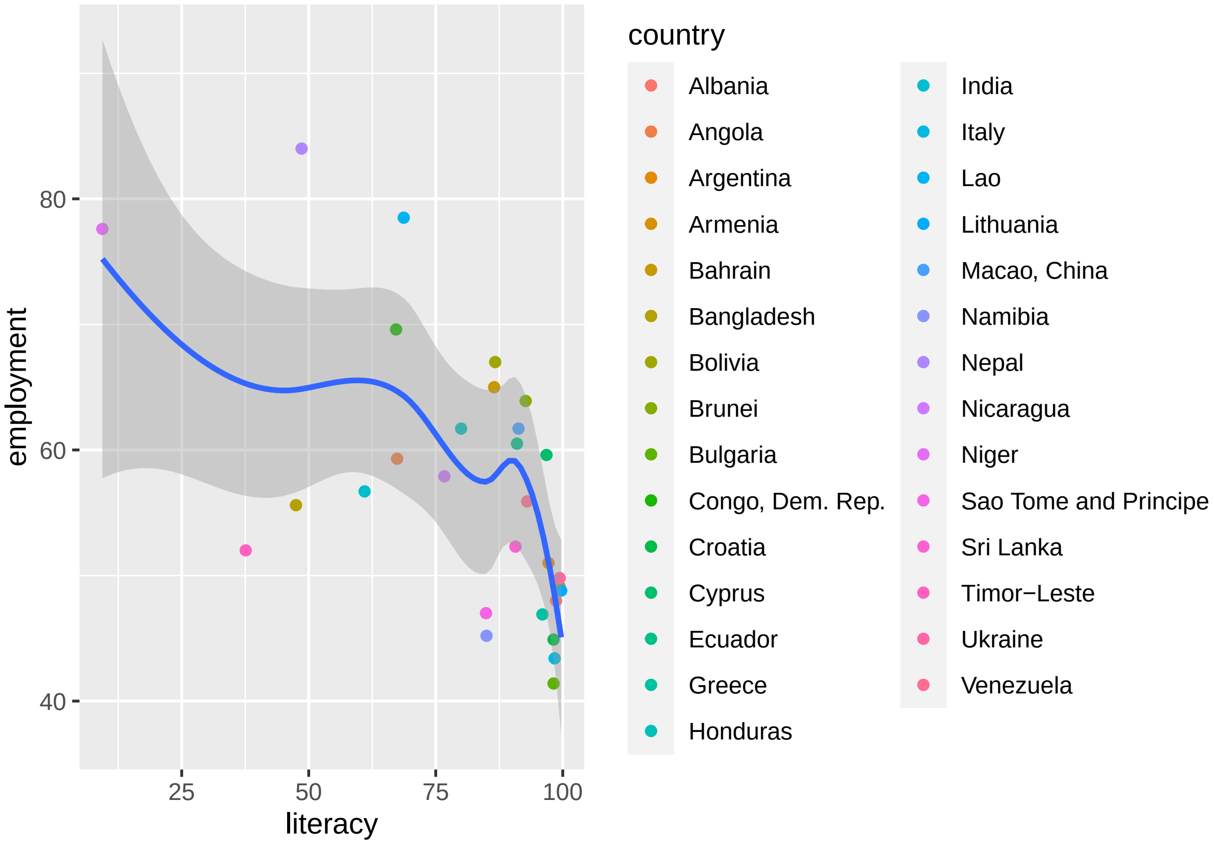


```
plt + geom_point(aes(color=country))
```



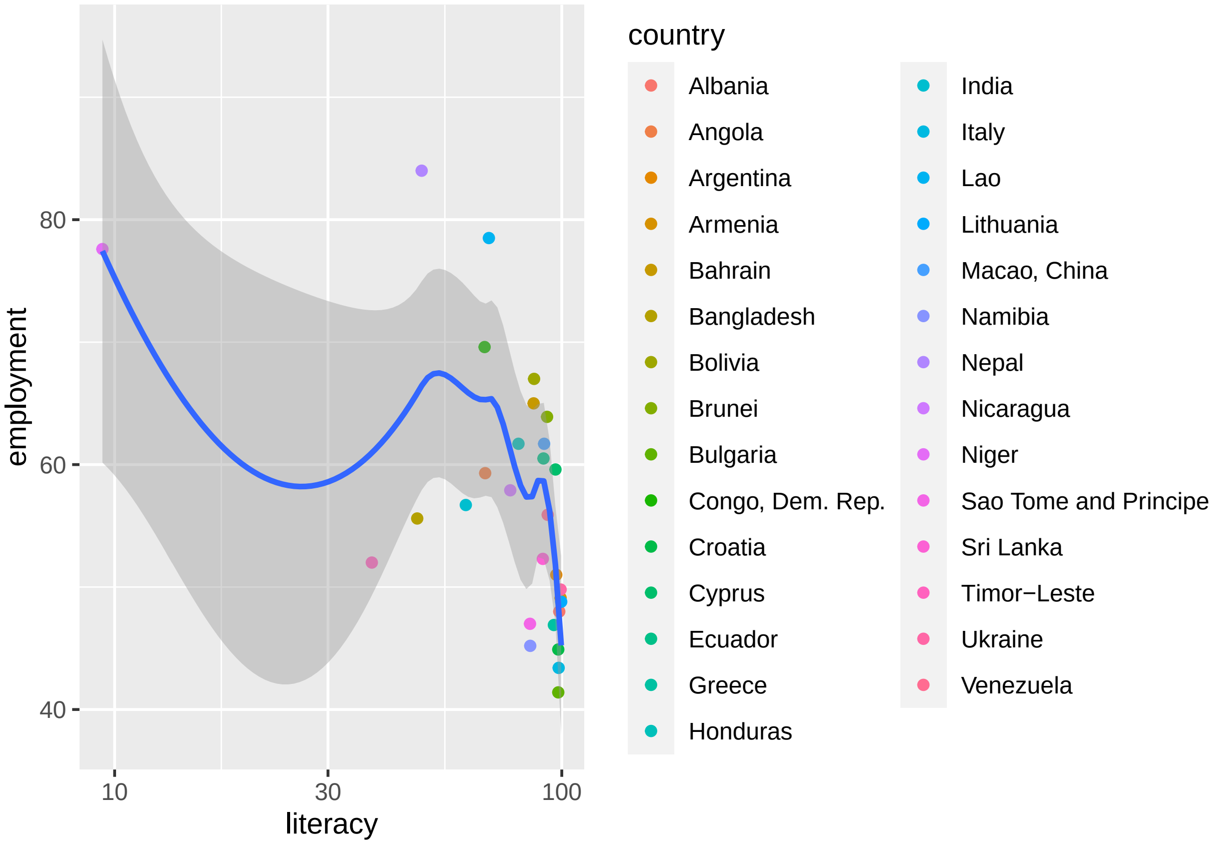
```r
plt + geom_point(mapping = aes(color=country)) +
  geom_smooth(method="loess")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
plt + geom_point(aes(color=country)) +
  geom_smooth(mapping = aes(x=literacy, y=employment), method="loess") +
  scale_x_log10()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
## use these mappings to extend or overwrite the global mappings
```

```

## Loading required package: sp
library(sp)
library(lattice)
library(survival)
library(Formula)

library(dplyr)
library(rworldmap) ## plotting the data on World Map
```

World Map

```
## ### Welcome to rworldmap ###
```

```
## For a short introduction type :   vignette('rworldmap')
```

```
library(countrycode) ## Converting the country name to Country code
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
```

```
##     format.pval, units
```

```
## view trend
dim(trend)
```

```
## [1] 29  3
```

```
colnames(trend)
```

```
## [1] "country"    "literacy"   "employment"
```

```
sum(complete.cases(trend)) ## No missing values found
```

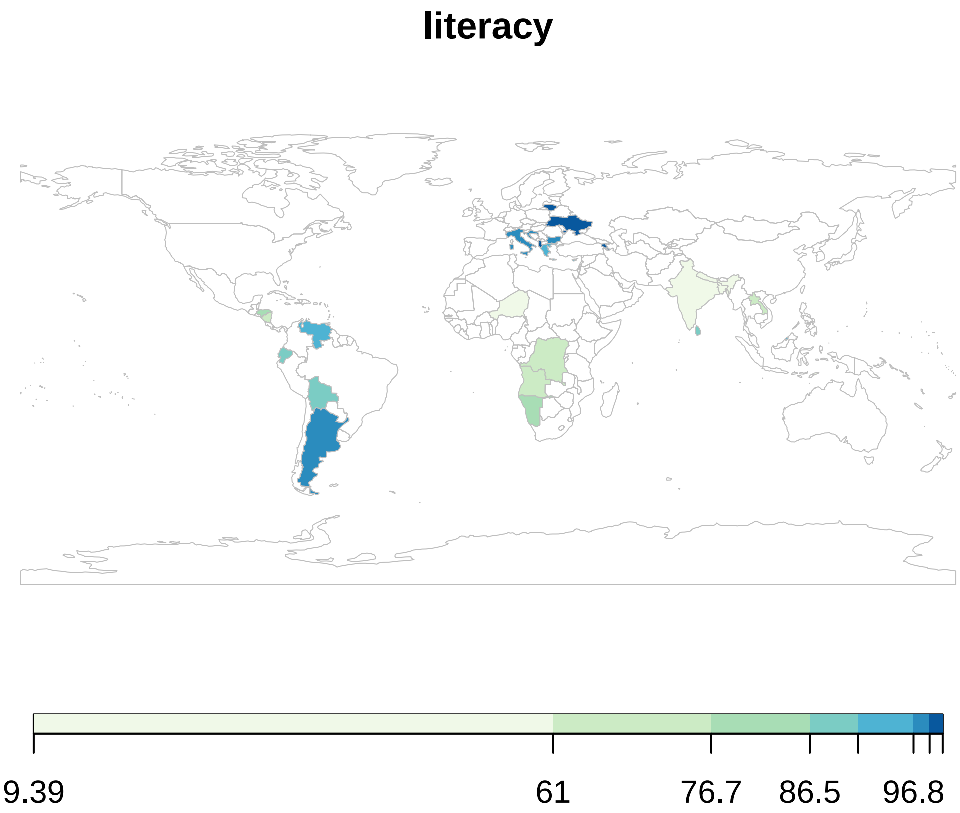
```
## [1] 29
```

```
describe(trend)  ## see Hmisc
```

```
## trend
##
##  3  Variables      29  Observations
## --------------------------------------------------------------------------------
## country
##        n  missing distinct
##       29        0       29
##
## lowest : Albania               Angola                Argentina             Armenia               Bah
## highest: Sao Tome and Principe Sri Lanka             Timor-Leste           Ukraine               Ven
## --------------------------------------------------------------------------------
## literacy
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       29        0       27        1    80.96    22.71    41.56    48.38
##      .25      .50      .75      .90      .95
##    68.70    90.70    97.20    98.84    99.40
##
## lowest :  9.39 37.60 47.50 48.60 61.00, highest: 98.20 98.40 98.70 99.40 99.70
## --------------------------------------------------------------------------------
## employment
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       29        0       28        1    57.04    12.27    44.00    45.14
##      .25      .50      .75      .90      .95
##    48.80    55.90    61.70    71.20    78.14
##
## lowest : 41.4 43.4 44.9 45.2 46.9, highest: 67.0 69.6 77.6 78.5 84.0
## --------------------------------------------------------------------------------
```

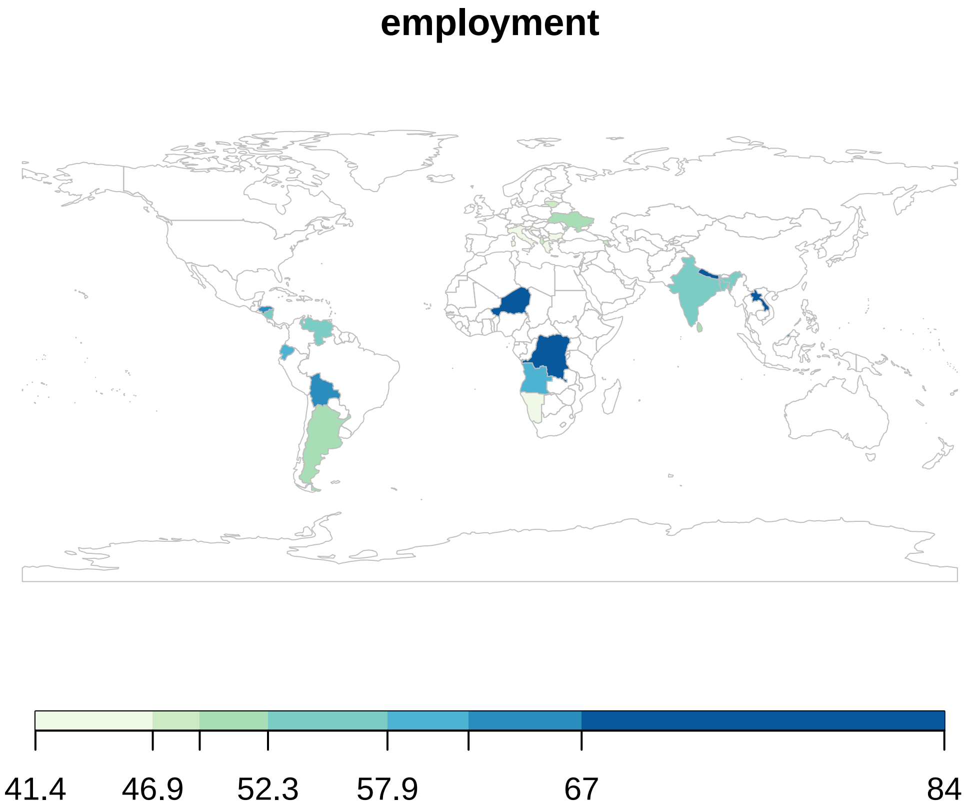
```
trend$countrycode <- countrycode(trend$country, 'country.name', 'iso3c')

sPDF <- joinCountryData2Map(trend
                            ,joinCode = "ISO3"
                            ,nameJoinColumn = "countrycode"
                            ,suggestForFailedCodes = FALSE
                            , verbose = T)
```

```
## 29 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
##      failedCodes failedCountries
## 214 codes from the map weren't represented in your data
## Finally, we build the "literacy rate" map over the 29 countries.
colourPalette <- brewer.pal(7,'GnBu')

mapParams <- mapCountryData(sPDF,
                            nameColumnToPlot="literacy",
                            addLegend=FALSE,
                            colourPalette=colourPalette )

## draw a color standard line to displays the literacy rate values
## corresponding to different colors

do.call(addMapLegend
        ,c(mapParams
           ,legendLabels="all"
           ,legendWidth=0.5
           ,legendIntervals="data"
           ,legendMar = 2))
```



```
## Finally, we build the "employment rate" map over the 29 countries.
colourPalette <- brewer.pal(7,'GnBu')

mapParams1 <- mapCountryData(sPDF,
                             nameColumnToPlot="employment",
                             addLegend=FALSE,
                             colourPalette=colourPalette )

## draw a color standard line to displays the employment rate values
## corresponding to different colors
```

```
do.call(addMapLegend
        ,c(mapParams1
           ,legendLabels="all"
           ,legendWidth=0.5
           ,legendIntervals="data"
           ,legendMar = 2))
```


employment
41.4
46.9
52.3
57.9
67
84