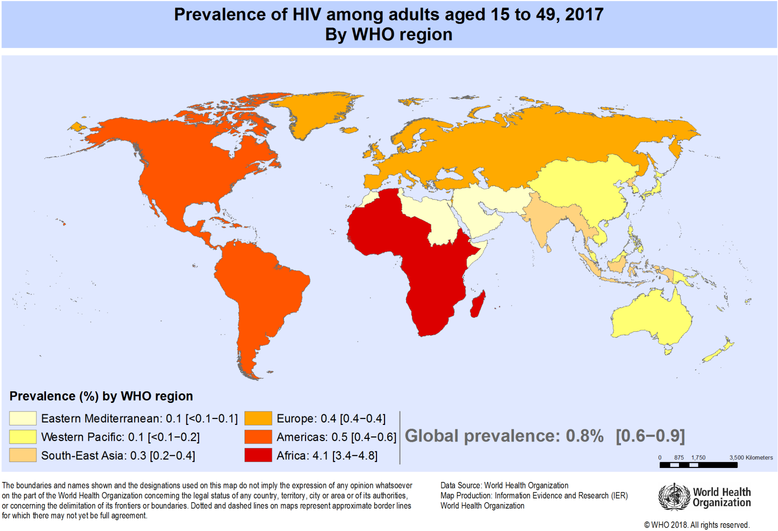
Case Study: HIV prevalence in the world

Sonia Mazzi

28 November, 2019

Data Science Campus

CASE STUDY HIV prevalence in the world - excel data from Gapminder



Prevalence of HIV among adults aged 15 to 49, 2017
By WHO region

Prevalence (%) by WHO region
Eastern Mediterranean: 0.1 [<0.1−0.1]
Western Pacific: 0.1 [<0.1−0.2]
South-East Asia: 0.3 [0.2−0.4]
Europe: 0.4 [0.4−0.4]
Americas: 0.5 [0.4−0.6]
Africa: 4.1 [3.4−4.8]
0
875
1,750
3,500 Kilometers

Global prevalence: 0.8% [0.6−0.9]

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: World Health Organization
Map Production: Information Evidence and Research (IER)
World Health Organization

World Health Organization

- Gapminder, https://www.gapminder.org, has a large collection of data sets, mostly in excel format.
- Retrieve the data about adults with HIV (estimated prevalence of HIV in percentage, ages 15-49) from Gapminder.
- The url is https://docs.google.com/spreadsheet/pub?key=pyj6tScZqmEfbZyl0qjbiRQ&output=xlsx
- The data is already in the DataFiles folder and it is named HIV.xlsx.

EXERCISE

- Read in the file "DataFiles/HIV.xlsx" into a tibble named `HIV`.
- Inspect `HIV`. How many rows and columns does the tibble have?
- What are the observational units? What are the variables, fixed and measured?
- What are the names of the columns of `HIV`? Do you notice anything peculiar about the names?
- Is `HIV` tidy data? If not, which manipulations are needed to make the data tidy?

Read in the “HIV.xlsx” file

```
HIV <- read_excel("DataFiles/HIV.xlsx")
dim(HIV)
```

```
## [1] 275  34
```

```
HIV
```

```
## # A tibble: 275 x 34
##    `Estimated HIV ~ `1979.0` `1980.0` `1981.0` `1982.0` `1983.0` `1984.0`
##    <chr>                <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 Abkhazia          NA             NA       NA       NA       NA       NA
##  2 Afghanistan       NA             NA       NA       NA       NA       NA
##  3 Akrotiri and Dh~  NA             NA       NA       NA       NA       NA
##  4 Albania           NA             NA       NA       NA       NA       NA
##  5 Algeria           NA             NA       NA       NA       NA       NA
##  6 American Samoa    NA             NA       NA       NA       NA       NA
##  7 Andorra           NA             NA       NA       NA       NA       NA
##  8 Angola             0.0265        NA       NA       NA       NA       NA
##  9 Anguilla          NA             NA       NA       NA       NA       NA
## 10 Antigua and Bar~  NA             NA       NA       NA       NA       NA
## # ... with 265 more rows, and 27 more variables: `1985.0` <dbl>,
## #   `1986.0` <dbl>, `1987.0` <dbl>, `1988.0` <lgl>, `1989.0` <lgl>,
## #   `1990.0` <dbl>, `1991.0` <dbl>, `1992.0` <dbl>, `1993.0` <dbl>,
## #   `1994.0` <dbl>, `1995.0` <dbl>, `1996.0` <dbl>, `1997.0` <dbl>,
## #   `1998.0` <dbl>, `1999.0` <dbl>, `2000.0` <dbl>, `2001.0` <dbl>,
## #   `2002.0` <dbl>, `2003.0` <dbl>, `2004.0` <dbl>, `2005.0` <dbl>,
## #   `2006.0` <dbl>, `2007.0` <dbl>, `2008.0` <dbl>, `2009` <chr>,
```

- Observational units are countries, variables are year and estimated prevalence.
- The column with countries is named with the title of the worksheet.
- Other columns contain the prevalence of HIV by year but some names are numerical.
- Skip the first row containing column names and assign these in R.

```r
HIV <- read_excel("DataFiles/HIV.xlsx", skip = 1, col_names = F)
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ... and 29 more problems
```

```r
HIV
```

```
## # A tibble: 275 x 34
##    ...1     ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9 ...10 ...11
##    <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl>
##  1 Abkh~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  2 Afgh~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  3 Akro~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  4 Alba~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  5 Alge~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  6 Amer~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  7 Ando~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
##  8 Ango~  0.0265    NA    NA    NA    NA    NA    NA    NA    NA NA
##  9 Angu~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
## 10 Anti~ NA         NA    NA    NA    NA    NA    NA    NA    NA NA
## # ... with 265 more rows, and 23 more variables: ...12 <lgl>, ...13 <dbl>,
## #   ...14 <dbl>, ...15 <dbl>, ...16 <dbl>, ...17 <dbl>, ...18 <dbl>,
## #   ...19 <dbl>, ...20 <dbl>, ...21 <dbl>, ...22 <dbl>, ...23 <dbl>,
```

```r
#print the column names of HIV
names(HIV)
```

```
##  [1] "...1"  "...2"  "...3"  "...4"  "...5"  "...6"  "...7"  "...8" 
##  [9] "...9"  "...10" "...11" "...12" "...13" "...14" "...15" "...16"
## [17] "...17" "...18" "...19" "...20" "...21" "...22" "...23" "...24"
## [25] "...25" "...26" "...27" "...28" "...29" "...30" "...31" "...32"
## [33] "...33" "...34"
```

```r
aux <- seq(1979, 2011, 1)
#rename the columns of HIV
names(HIV) <- c("Country", as.character(aux))
```

```r
names(HIV)
```

```
##  [1] "Country" "1979"    "1980"    "1981"    "1982"    "1983"    "1984"   
##  [8] "1985"    "1986"    "1987"    "1988"    "1989"    "1990"    "1991"   
## [15] "1992"    "1993"    "1994"    "1995"    "1996"    "1997"    "1998"   
## [22] "1999"    "2000"    "2001"    "2002"    "2003"    "2004"    "2005"   
## [29] "2006"    "2007"    "2008"    "2009"    "2010"    "2011"
```

```r
head(HIV)
```

```
## # A tibble: 6 x 34
##   Country `1979` `1980` `1981` `1982` `1983` `1984` `1985` `1986` `1987`
##   <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Abkhaz~     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 2 Afghan~     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 3 Akroti~     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 4 Albania     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 5 Algeria     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 6 Americ~     NA     NA     NA     NA     NA     NA     NA     NA     NA
## # ... with 24 more variables: `1988` <lgl>, `1989` <lgl>, `1990` <dbl>,
## #   `1991` <dbl>, `1992` <dbl>, `1993` <dbl>, `1994` <dbl>, `1995` <dbl>,
## #   `1996` <dbl>, `1997` <dbl>, `1998` <dbl>, `1999` <dbl>, `2000` <dbl>,
## #   `2001` <dbl>, `2002` <dbl>, `2003` <dbl>, `2004` <dbl>, `2005` <dbl>,
## #   `2006` <dbl>, `2007` <dbl>, `2008` <dbl>, `2009` <chr>, `2010` <chr>,
## #   `2011` <chr>
```

- The last three columns have been read as character.
- The columns corresponding to 1988 and 1989 are of class logical because all their entries are NAs.

- Coerce the last three columns to numeric mode.

```
HIV <- HIV %>%
  mutate_at(32:34, as.numeric)
```

```
HIV
```

```
## # A tibble: 275 x 34
##    Country  `1979` `1980` `1981` `1982` `1983` `1984` `1985` `1986` `1987`
##    <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Abkhaz~ NA            NA     NA     NA     NA     NA     NA     NA     NA
##  2 Afghan~ NA            NA     NA     NA     NA     NA     NA     NA     NA
##  3 Akroti~ NA            NA     NA     NA     NA     NA     NA     NA     NA
##  4 Albania NA            NA     NA     NA     NA     NA     NA     NA     NA
##  5 Algeria NA            NA     NA     NA     NA     NA     NA     NA     NA
##  6 Americ~ NA            NA     NA     NA     NA     NA     NA     NA     NA
##  7 Andorra NA            NA     NA     NA     NA     NA     NA     NA     NA
##  8 Angola   0.0265       NA     NA     NA     NA     NA     NA     NA     NA
##  9 Anguil~ NA            NA     NA     NA     NA     NA     NA     NA     NA
## 10 Antigu~ NA            NA     NA     NA     NA     NA     NA     NA     NA
## # ... with 265 more rows, and 24 more variables: `1988` <lgl>,
## #   `1989` <lgl>, `1990` <dbl>, `1991` <dbl>, `1992` <dbl>, `1993` <dbl>,
## #   `1994` <dbl>, `1995` <dbl>, `1996` <dbl>, `1997` <dbl>, `1998` <dbl>,
## #   `1999` <dbl>, `2000` <dbl>, `2001` <dbl>, `2002` <dbl>, `2003` <dbl>,
## #   `2004` <dbl>, `2005` <dbl>, `2006` <dbl>, `2007` <dbl>, `2008` <dbl>,
## #   `2009` <dbl>, `2010` <dbl>, `2011` <dbl>
```

- The columns up to 1990 are mostly NAs. Remove them from the data set.

```
HIV <- select(HIV, c(1,13:34))
```

```
HIV
```

```
## # A tibble: 275 x 23
##    Country `1990` `1991` `1992` `1993` `1994` `1995` `1996` `1997` `1998`
##    <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Abkhaz~  NA     NA     NA     NA     NA     NA     NA     NA     NA   
##  2 Afghan~  NA     NA     NA     NA     NA     NA     NA     NA     NA   
##  3 Akroti~  NA     NA     NA     NA     NA     NA     NA     NA     NA   
##  4 Albania  NA     NA     NA     NA     NA     NA     NA     NA     NA   
##  5 Algeria   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06
##  6 Americ~  NA     NA     NA     NA     NA     NA     NA     NA     NA   
##  7 Andorra  NA     NA     NA     NA     NA     NA     NA     NA     NA   
##  8 Angola    0.5    0.8    1      1.2    1.4    1.6    1.7    1.8    1.8 
##  9 Anguil~  NA     NA     NA     NA     NA     NA     NA     NA     NA   
## 10 Antigu~  NA     NA     NA     NA     NA     NA     NA     NA     NA   
## # ... with 265 more rows, and 13 more variables: `1999` <dbl>,
## #   `2000` <dbl>, `2001` <dbl>, `2002` <dbl>, `2003` <dbl>, `2004` <dbl>,
## #   `2005` <dbl>, `2006` <dbl>, `2007` <dbl>, `2008` <dbl>, `2009` <dbl>,
## #   `2010` <dbl>, `2011` <dbl>
```

EXERCISE

Tidy the `HIV` data. Name the tidy object `HIV2`.

Let us tidy the data ready for analysis.

- An observational unit is a country.
- The variables are year and prevalence of HIV.
- The tidy version of the data has three columns: country, year and prevalence.

```r
head(HIV)
```

```
## # A tibble: 6 x 23
##   Country `1990` `1991` `1992` `1993` `1994` `1995` `1996` `1997` `1998`
##   <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Abkhaz~  NA     NA     NA     NA     NA     NA     NA     NA     NA
## 2 Afghan~  NA     NA     NA     NA     NA     NA     NA     NA     NA
## 3 Akroti~  NA     NA     NA     NA     NA     NA     NA     NA     NA
## 4 Albania  NA     NA     NA     NA     NA     NA     NA     NA     NA
## 5 Algeria   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06
## 6 Americ~  NA     NA     NA     NA     NA     NA     NA     NA     NA
## # ... with 13 more variables: `1999` <dbl>, `2000` <dbl>, `2001` <dbl>,
## #   `2002` <dbl>, `2003` <dbl>, `2004` <dbl>, `2005` <dbl>, `2006` <dbl>,
## #   `2007` <dbl>, `2008` <dbl>, `2009` <dbl>, `2010` <dbl>, `2011` <dbl>
```

- Create a new column named `Year` with columns names which are a year number and put the corresponding values of prevalence under a new column named `PrevalenceHIV`

```r
HIV2 <- HIV %>%
  pivot_longer(-Country, names_to = "Year",
               values_to = "PrevalenceHIV")
```

```
HIV2
```

```
## # A tibble: 6,050 x 3
##    Country  Year  PrevalenceHIV
##    <chr>    <chr>         <dbl>
##  1 Abkhazia 1990             NA
##  2 Abkhazia 1991             NA
##  3 Abkhazia 1992             NA
##  4 Abkhazia 1993             NA
##  5 Abkhazia 1994             NA
##  6 Abkhazia 1995             NA
##  7 Abkhazia 1996             NA
##  8 Abkhazia 1997             NA
##  9 Abkhazia 1998             NA
## 10 Abkhazia 1999             NA
## # ... with 6,040 more rows
```

- The data is tidy.
- Let us visualise the data

- Visualise using the concepts and additional data in Gapminder.org

- The HIV prevalence data will be plotted vs. Income (GDP per capita, PPP$ inflation-adjusted).

- Income data in Gapminder is in excel format in the url https://docs.google.com/spreadsheets/d/1PybxH399kK6OjJI4T2M33UsLqgutwj3SuYbk7Yt6sxE/pub.

- The data has already been downloaded and is in the file "gdp_per_capita_ppp.xlsx" in the current working directory.

- Note how we are going back to the beginning of the data analysis process in order to make our data exploration more meaningful.

EXERCISE

- Read the data in "DataFiles/gdp_per_capita_ppp.xlsx" into an object named `income`. This file containds gdp per capita from 1800 until 2015.

- Rename all the columns of `income` to `Country` and `1800 - 2015`.

- Is the data in tidy format? If not, create an object called `income2` with the tidy version of `income` (think about observational units and variables).

- Join the information of `HIV2` and `income2` into one single tibble named `HIV_Inc`.

- Now, add region (continent, sub-continent) information from "DataFiles/DataGeographiesGapminder.xlsx". This is a workbook with many sheets. The second sheet is the one that contains the list of country names and different region denominations and other geographical information. Rename the column with country names to `Country`.

```r
income <- read_excel("DataFiles/gdp_per_capita_ppp.xlsx")
```

```r
head(income)
```

```
## # A tibble: 6 x 217
##   `GDP per capita` `1800.0` `1801.0` `1802.0` `1803.0` `1804.0` `1805.0`
##   <chr>               <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 Abkhazia               NA       NA       NA       NA       NA       NA
## 2 Afghanistan           603      603      603      603      603      603
## 3 Akrotiri and Dh~       NA       NA       NA       NA       NA       NA
## 4 Albania               667      667      668      668      668      668
## 5 Algeria               716      716      717      718      719      720
## 6 American Samoa         NA       NA       NA       NA       NA       NA
## # ... with 210 more variables: `1806.0` <dbl>, `1807.0` <dbl>,
## #   `1808.0` <dbl>, `1809.0` <dbl>, `1810.0` <dbl>, `1811.0` <dbl>,
## #   `1812.0` <dbl>, `1813.0` <dbl>, `1814.0` <dbl>, `1815.0` <dbl>,
## #   `1816.0` <dbl>, `1817.0` <dbl>, `1818.0` <dbl>, `1819.0` <dbl>,
## #   `1820.0` <dbl>, `1821.0` <dbl>, `1822.0` <dbl>, `1823.0` <dbl>,
## #   `1824.0` <dbl>, `1825.0` <dbl>, `1826.0` <dbl>, `1827.0` <dbl>,
## #   `1828.0` <dbl>, `1829.0` <dbl>, `1830.0` <dbl>, `1831.0` <dbl>,
## #   `1832.0` <dbl>, `1833.0` <dbl>, `1834.0` <dbl>, `1835.0` <dbl>,
## #   `1836.0` <dbl>, `1837.0` <dbl>, `1838.0` <dbl>, `1839.0` <dbl>,
## #   `1840.0` <dbl>, `1841.0` <dbl>, `1842.0` <dbl>, `1843.0` <dbl>,
## #   `1844.0` <dbl>, `1845.0` <dbl>, `1846.0` <dbl>, `1847.0` <dbl>,
## #   `1848.0` <dbl>, `1849.0` <dbl>, `1850.0` <dbl>, `1851.0` <dbl>,
## #   `1852.0` <dbl>, `1853.0` <dbl>, `1854.0` <dbl>, `1855.0` <dbl>,
## #   `1856.0` <dbl>, `1857.0` <dbl>, `1858.0` <dbl>, `1859.0` <dbl>,
## #   `1860.0` <dbl>, `1861.0` <dbl>, `1862.0` <dbl>, `1863.0` <dbl>,
names(income) <- c("Country", as.character(seq(1800, 2015, 1)))
income2 <- pivot_longer(income, -Country, names_to =  "Year",
                        values_to = "Income")
```

```
income2
```

```
## # A tibble: 56,592 x 3
##    Country  Year  Income
##    <chr>    <chr>  <dbl>
##  1 Abkhazia 1800      NA
##  2 Abkhazia 1801      NA
##  3 Abkhazia 1802      NA
##  4 Abkhazia 1803      NA
##  5 Abkhazia 1804      NA
##  6 Abkhazia 1805      NA
##  7 Abkhazia 1806      NA
##  8 Abkhazia 1807      NA
##  9 Abkhazia 1808      NA
## 10 Abkhazia 1809      NA
## # ... with 56,582 more rows
```

This looks better!

Do `HIV2` and `income2` contain the same countries?

```r
nrow(HIV2)
```

```
## [1] 6050
```

```r
nrow(income2)
```

```
## [1] 56592
```

- Need to combine the `HIV2` and `income2` data sets
- They have a different number of rows (countries).
- Merge them leaving out the data for countries which are not common to both data sets.

- Use inner_join() (package dplyr)

```
#merging HIV2 and income2
HIV_Inc <- inner_join(HIV2, income2)
```

```
## Joining, by = c("Country", "Year")
```

```
HIV_Inc
```

```
## # A tibble: 5,720 x 4
##    Country  Year  PrevalenceHIV Income
##    <chr>    <chr>         <dbl>  <dbl>
##  1 Abkhazia 1990             NA     NA
##  2 Abkhazia 1991             NA     NA
##  3 Abkhazia 1992             NA     NA
##  4 Abkhazia 1993             NA     NA
##  5 Abkhazia 1994             NA     NA
##  6 Abkhazia 1995             NA     NA
##  7 Abkhazia 1996             NA     NA
##  8 Abkhazia 1997             NA     NA
##  9 Abkhazia 1998             NA     NA
## 10 Abkhazia 1999             NA     NA
## # ... with 5,710 more rows
```

Data Science Campus

- Add region (continent, sub-continent) information.
- Downloaded from https://www.gapminder.org/data/geo/ into the file “DataGeographiesGapminder.xlsx”.
- This is a workbook with many sheets.
- The second sheet is the one that contains the list of country names and different region denominations and other geographical information.

```r
# read only the second sheet
continent <- read_excel(
    "DataFiles/DataGeographiesGapminder.xlsx", sheet = 2)
```

```r
head(continent)
```

```
## # A tibble: 6 x 11
##   geo   name  four_regions eight_regions six_regions members_oecd_g77
##   <chr> <chr> <chr>        <chr>         <chr>       <chr>
## 1 afg   Afgh~ asia         asia_west     south_asia  g77
## 2 alb   Alba~ europe       europe_east   europe_cen~ others
## 3 dza   Alge~ africa       africa_north  middle_eas~ g77
## 4 and   Ando~ europe       europe_west   europe_cen~ others
## 5 ago   Ango~ africa       africa_sub_s~ sub_sahara~ g77
## 6 atg   Anti~ americas     america_north america     g77
## # ... with 5 more variables: Latitude <dbl>, Longitude <dbl>, `UN member
## #   since` <dttm>, `World bank region` <chr>, `World bank income group
## #   2017` <chr>
```

```r
#rename the second column
continent <- rename(continent, Country = name)
```

```r
glimpse(continent)
```

```
## Observations: 197
## Variables: 11
## $ geo                            <chr> "afg", "alb", "dza", "and", "ag...
## $ Country                        <chr> "Afghanistan", "Albania", "Alge...
## $ four_regions                   <chr> "asia", "europe", "africa", "eu...
## $ eight_regions                  <chr> "asia_west", "europe_east", "af...
## $ six_regions                    <chr> "south_asia", "europe_central_a...
## $ members_oecd_g77               <chr> "g77", "others", "g77", "others...
## $ Latitude                       <dbl> 33.00000, 41.00000, 28.00000, 4...
## $ Longitude                      <dbl> 66.00000, 20.00000, 3.00000, 1....
## $ `UN member since`              <dttm> 1946-11-19, 1955-12-14, 1962-1...
## $ `World bank region`            <chr> "South Asia", "Europe & Central...
## $ `World bank income group 2017` <chr> "Low income", "Upper middle inc...
```

- Next we merge this information with HIV and income data

```
HIV_Inc_Cont <- inner_join(HIV_Inc, continent)
```

```
## Joining, by = "Country"
```

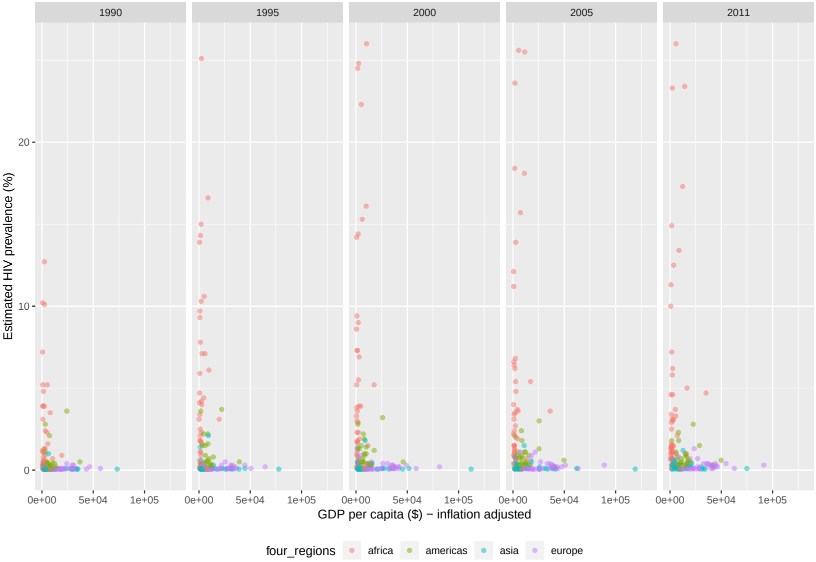
```
glimpse(HIV_Inc_Cont)
```

```
## Observations: 4,312
## Variables: 14
## $ Country                        <chr> "Afghanistan", "Afghanistan", "...
## $ Year                           <chr> "1990", "1991", "1992", "1993",...
## $ PrevalenceHIV                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA,...
## $ Income                         <dbl> 1028, 1022, 941, 810, 725, 872,...
## $ geo                            <chr> "afg", "afg", "afg", "afg", "af...
## $ four_regions                   <chr> "asia", "asia", "asia", "asia",...
## $ eight_regions                  <chr> "asia_west", "asia_west", "asia...
## $ six_regions                    <chr> "south_asia", "south_asia", "so...
## $ members_oecd_g77               <chr> "g77", "g77", "g77", "g77", "g7...
## $ Latitude                       <dbl> 33, 33, 33, 33, 33, 33, 33, 33,...
## $ Longitude                      <dbl> 66, 66, 66, 66, 66, 66, 66, 66,...
## $ `UN member since`              <dttm> 1946-11-19, 1946-11-19, 1946-1...
## $ `World bank region`            <chr> "South Asia", "South Asia", "So...
## $ `World bank income group 2017` <chr> "Low income", "Low income", "Lo...
```

- Plot prevalence vs. income distinguishing with colours by continent and having 5 parallel plots, one for each of years 1990, 1995, 2000, 2005, 2011.
- Use `filter()` to subset the data according to a logical criterion.
- Use package `ggrepel` (to avoid overlapping labels in plots)

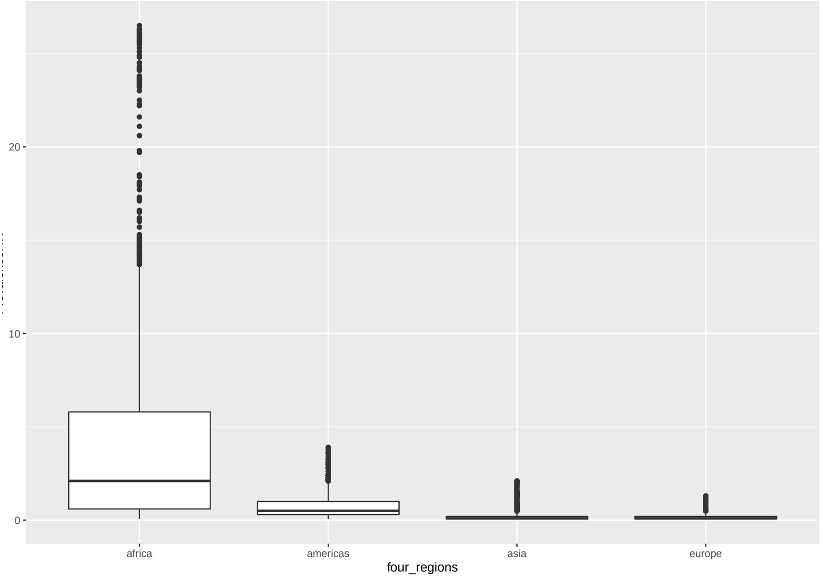
```
library(ggrepel)
```

```r
#select the rows with the desired years
HIV_Inc_Cont %>%
filter(Year %in% c("1990", "1995", "2000", "2005", "2011"))%>%
ggplot(aes(x = Income, y = PrevalenceHIV, col = four_regions) ) +
        geom_point(alpha = 0.5) +
        labs(x = "GDP per capita ($) - inflation adjusted" ) +
        labs(y = "Estimated HIV prevalence (%)" ) +
        facet_grid(.~Year) + # one plot for each of the years
        theme(legend.position = "bottom")
```


1990
1995
2000
2005
2011
Estimated HIV prevalence (%)
20
10
0
0e+00
5e+04
1e+05
GDP per capita ($) − inflation adjusted
four_regions
africa
americas
asia
europe

- Most African countries have prevalence values in a scale which is about ten times that of the rest of the world.
- This makes the visualisation difficult.
- Visualise the data for African countries separately.

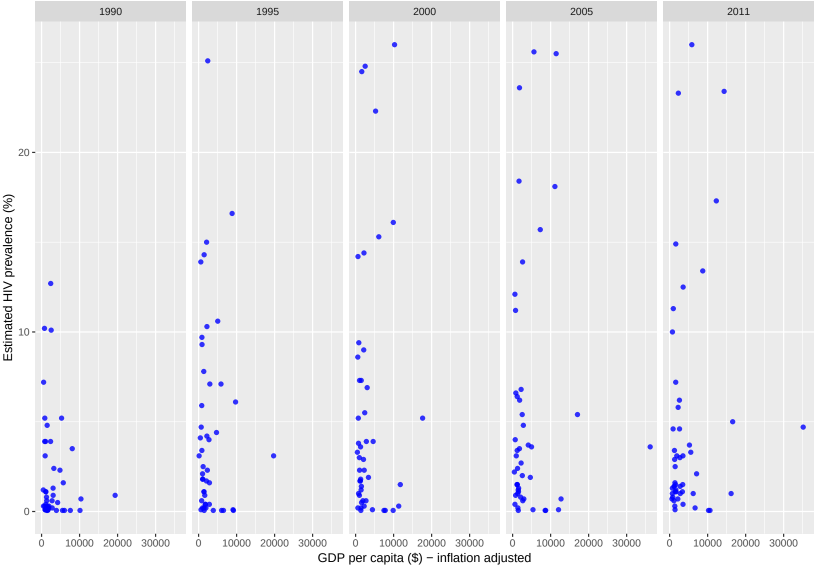
```r
ggplot(HIV_Inc_Cont, aes(x = four_regions, y = PrevalenceHIV)) +
  geom_boxplot()
```



```r
#Only Africa. Further filter the data to select only countries in Africa
#no need to define aux
aux2 <- HIV_Inc_Cont %>%
  filter(Year %in% c("1990", "1995", "2000", "2005", "2011")) %>%
  filter(four_regions == "africa")
```

```r
p_africa <-
  ggplot(aux2, aes(x = Income, y = PrevalenceHIV) ) +
    geom_point(alpha = 0.8, color = "blue") +
    labs(x = "GDP per capita ($) - inflation adjusted" ) +
    labs(y = "Estimated HIV prevalence (%)" ) +
    facet_grid(.~Year)

p_africa
```


1990
1995
2000
2005
2011
Estimated HIV prevalence (%)
0
10
20
0
10000
20000
30000
GDP per capita ($) − inflation adjusted

- Identify the African countries with HIV prevalence greater than or equal to 10%.

```
#for year 1990.
#Further filter the data selecting prevalence>=10 and year 1990
x_90 <- filter(aux2, PrevalenceHIV >= 10, Year == "1990")
select(x_90, 1:5)
```

```
## # A tibble: 3 x 5
##   Country  Year  PrevalenceHIV Income geo
##   <chr>    <chr>         <dbl>  <dbl> <chr>
## 1 Uganda   1990           10.2    767 uga
## 2 Zambia   1990           12.7   2407 zmb
## 3 Zimbabwe 1990           10.1   2532 zwe
```

```r
#for year 1995
x_95 <- filter(aux2, PrevalenceHIV >= 10, Year == "1995")
select(x_95,1:5)
```

```
## # A tibble: 7 x 5
##   Country   Year  PrevalenceHIV Income geo
##   <chr>     <chr>         <dbl>  <dbl> <chr>
## 1 Botswana  1995           16.6   8823 bwa
## 2 Kenya     1995           10.3   2199 ken
## 3 Lesotho   1995           14.3   1466 lso
## 4 Malawi    1995           13.9    593 mwi
## 5 Swaziland 1995           10.6   5043 swz
## 6 Zambia    1995           15     2106 zmb
## 7 Zimbabwe  1995           25.1   2416 zwe
```

```r
#for year 2000
x_00 <- filter(aux2, PrevalenceHIV >= 10 & Year == "2000")
select(x_00, 1:5)
```

```
## # A tibble: 8 x 5
##   Country      Year  PrevalenceHIV Income geo
##   <chr>        <chr>         <dbl>  <dbl> <chr>
## 1 Botswana     2000           26     10250 bwa
## 2 Lesotho      2000           24.5    1629 lso
## 3 Malawi       2000           14.2     632 mwi
## 4 Namibia      2000           15.3    6111 nam
## 5 South Africa 2000           16.1    9927 zaf
## 6 Swaziland    2000           22.3    5257 swz
## 7 Zambia       2000           14.4    2202 zmb
## 8 Zimbabwe     2000           24.8    2521 zwe
```

```r
#for year 2005
x_05 <- filter(aux2, PrevalenceHIV >= 10 & Year == "2005")
select(x_05, 1:5)
```

```
## # A tibble: 9 x 5
##   Country      Year  PrevalenceHIV Income geo
##   <chr>        <chr>         <dbl>  <dbl> <chr>
## 1 Botswana     2005           25.5  11460 bwa
## 2 Lesotho      2005           23.6   1810 lso
## 3 Malawi       2005           12.1    609 mwi
## 4 Mozambique   2005           11.2    774 moz
## 5 Namibia      2005           15.7   7279 nam
## 6 South Africa 2005           18.1  11133 zaf
## 7 Swaziland    2005           25.6   5618 swz
## 8 Zambia       2005           13.9   2620 zmb
## 9 Zimbabwe     2005           18.4   1689 zwe
```

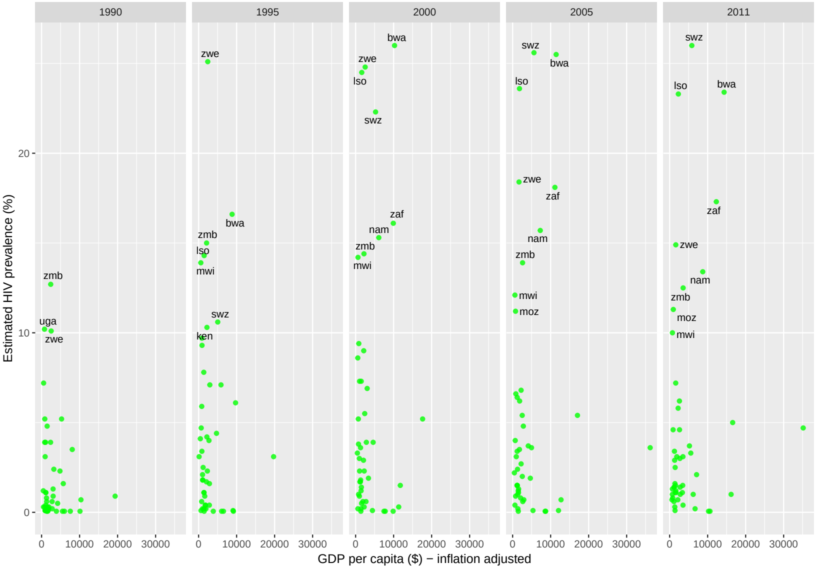
```r
#for year 2011
x_11 <- filter(aux2, PrevalenceHIV >= 10 & Year == "2011")
select(x_11, 1:5)
```

```
## # A tibble: 9 x 5
##   Country      Year  PrevalenceHIV Income geo
##   <chr>        <chr>         <dbl>  <dbl> <chr>
## 1 Botswana     2011           23.4  14341 bwa
## 2 Lesotho      2011           23.3   2301 lso
## 3 Malawi       2011           10      747 mwi
## 4 Mozambique   2011           11.3    974 moz
## 5 Namibia      2011           13.4   8715 nam
## 6 South Africa 2011           17.3  12291 zaf
## 7 Swaziland    2011           26     5846 swz
## 8 Zambia       2011           12.5   3557 zmb
## 9 Zimbabwe     2011           14.9   1626 zwe
```

```
#Add the names of the countries with high HIV prevalence to the plots.

p_africa <- p_africa +
    geom_text_repel(data = x_90, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_95, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_00, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_05, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_11, aes(label = geo) , col = "black", size = 3)

p_africa
```



Data Science Campus

Which countries are getting richer? Is that reflecting on the HIV prevalence?

```
#for year 1990
# select african countries data for year 1990 and income>=15000
x_90 <- filter(aux2, Income >= 15000 & Year == "1990")
select(x_90, 1:5)
```

```
## # A tibble: 2 x 5
##   Country Year  PrevalenceHIV Income geo  
##   <chr>   <chr>         <dbl>  <dbl> <chr>
## 1 Gabon   1990            0.9  19358 gab  
## 2 Libya   1990           NA    26928 lby  
```

```r
#for year 1995
x_95 <- filter(aux2, Income >= 15000 & Year == "1995")
select(x_95, 1:5)
```

```
## # A tibble: 3 x 5
##   Country    Year  PrevalenceHIV Income geo
##   <chr>      <chr>         <dbl>  <dbl> <chr>
## 1 Gabon      1995            3.1  19738 gab
## 2 Libya      1995           NA    23363 lby
## 3 Seychelles 1995           NA    15097 syc
```

```r
#for year 2000
x_00 <- filter(aux2, Income >= 15000 & Year == "2000")
select(x_00, 1:5)
```

```
## # A tibble: 3 x 5
##   Country    Year  PrevalenceHIV Income geo
##   <chr>      <chr>         <dbl>  <dbl> <chr>
## 1 Gabon      2000            5.2  17630 gab
## 2 Libya      2000           NA    22682 lby
## 3 Seychelles 2000           NA    18453 syc
```

```r
#for year 2005
x_05 <- filter(aux2, Income >= 15000 & Year == "2005")
select(x_05, 1:5)
```

```
## # A tibble: 4 x 5
##   Country           Year  PrevalenceHIV Income geo
##   <chr>             <chr>         <dbl>  <dbl> <chr>
## 1 Equatorial Guinea 2005            3.6  36200 gnq
## 2 Gabon             2005            5.4  17069 gab
## 3 Libya             2005           NA    26967 lby
## 4 Seychelles        2005           NA    17803 syc
```

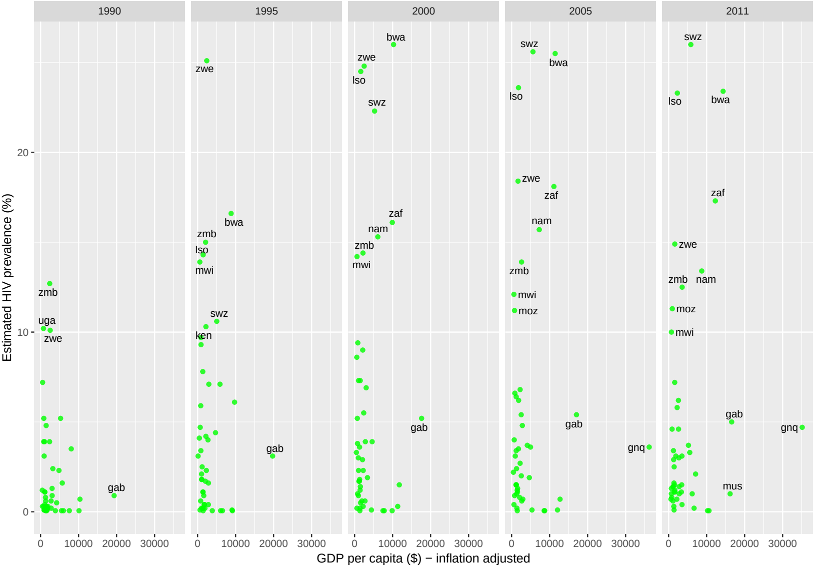
```r
#for year 2011
x_11 <- filter(aux2, Income >= 15000 & Year == "2011")
select(x_11, 1:5)
```

```
## # A tibble: 4 x 5
##   Country           Year  PrevalenceHIV Income geo
##   <chr>             <chr>         <dbl>  <dbl> <chr>
## 1 Equatorial Guinea 2011            4.7  35150 gnq
## 2 Gabon             2011            5    16590 gab
## 3 Mauritius         2011            1    16179 mus
## 4 Seychelles        2011           NA    22556 syc
```

```
#Let us add the names of the countries with high income to the plots.

p_africa <- p_africa +
    geom_text_repel(data = x_90, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_95, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_00, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_05, aes(label = geo) , col = "black", size = 3) +
    geom_text_repel(data = x_11, aes(label = geo) , col = "black", size = 3)

p_africa
```



Data Science Campus

EXERCISE

Carry out a visualisation of HIV prevalence data for rest of the world (without Africa) and the Americas, distinguishing between the sub-regions in the Americas. Identify the countries in the Americas with the highest HIV prevalence rates.