Enhanced visualization: assignment 9

ggplot2

I. ggplot2 basic features

In this assignment we will test the basic features of the ggplot2 package for R. This assignment is based on a tutorial which you can find at `http://www.personal.psu.edu/users/j/r/jre206/ggplot2.html`. We will use a data set collected from world population census of the year 2016 available at `http://www.worldpopdata.org/`.

Download the data set file *wpd.csv* from *moodle*. This data set contains population information from 210 countries. Its variables are indicated in Table 1.

Variable name	Meaning	Units
country	Country name	
pop2016	Population	Millions of inhabitants
imr	Infant mortality rate	Deaths per 1000 births
tfr	Total fertility rate	Children per woman
le	Life expectancy at birth	Years
lem	Male life expectancy	Years
lef	Female life expectancy	Years
region	Within area region name	
area	Continental area name	

Table 1: Variables and their meaning in the WPD 2016 data set.

A. ggplot and geoms: open a *rstudio* project. Load the ggplot2 package[1], load the data and check the loaded table with

```
library(ggplot2)
wpd <- read.csv('wpd.csv')
head(wpd)
```

[1] If required install the package in R with the command line `install.packages("ggplot2")`.

1. Create a ggplot2 object with the command line

   ```
   p <- ggplot(data = wpd, aes(x = le, y = tfr, color = area))
   ```

   what is already defined in the object generation?

2. Define a layer with point geometry

   ```
   p + layer(geom = 'point', stat = 'identity', position = 'identity')
   ```

   what is the type of the generated chart?

3. Try the line and step geometries. Do they make sense as a graphic?

B. stat:

1. Generate a scatter plot with an additional smooth statistic with the command lines

   ```
   p <- ggplot(data = wpd, aes(x = le, y = tfr))
   p + layer(geom = 'point', stat = 'identity', position = 'identity', params = list(shape = 1)) +
   layer(geom = 'smooth', stat = 'smooth', position = 'identity', params = list(method = 'loess'))
   ```

   what indicates the "loess" parameter?

2. Modify the previous code to obtain a linear regression statistic.

3. Generate all previous graphics using shortcuts. For doing so, you can use the R cheat sheet for ggplot2 that you can download from *moodle* (*ggplot2_cheat_sheet.pdf*).

4. Generate a bar chart with the counts of countries per area with the shortcut

   ```
   p <- ggplot(data = wpd, aes(x = area))
   p + stat_count()
   ```

   how do you change its geometry to points?

5. Generate a histogram of the countries as a function of their life expectancy with the code

   ```
   p <- ggplot(data = wpd, aes(x = le))
   p + stat_bin(binwidth = 5)
   ```

   what is the effect of the parameter binwidth?

6. Modify the previous code so that you generate an histogram with 5 bins.

C. position

1. Generate a bar chart with the following code:

   ```
   wpd$tfrGT2 <- wpd$tfr > 2
   p <- ggplot(data = wpd, aes(x = area, fill = tfrGT2))
   p + geom_bar()
   ```

   what kind of bar chart is this?

2. Change the position argument of the bar geometry to 'stack', 'dodge' and 'fill'. What is the effect of changing this argument?

3. Generate a scatter plot of the variable *le* against *tfr* and test the position argument 'jitter'? What does it do? In what situation do you think this can be useful?

D. Labels

1. Filter the data for northern Africa and generate a scatter plot of *tfr* against *le*. Indicate by labels using the command geom_text the subset of countries with *tfr*>3.

E. Coordinates and scales

1. Generate a stacked bar with

```
p <- ggplot(data = wpd, aes(x = factor(1), fill = area))
p + geom_bar()
```

   what happens if you use coord_flip()?

2. How do you generate a pie chart by changing the coordinates?

3. Generate a scatter plot of country population against *le* with the country population in logarithmic scales. Remember to filter out countries that are erroneously indicated with zero population.

F. Theme

1. In the following code indicate what each line does:

```
p <- ggplot(data = wpd, aes(x = le, y = tfr))
p + geom_jitter() + ggtitle("Life expectancy and TFR") +
xlab("Life expectancy (years)") + ylab("Total fertility rate") +
scale_x_continuous(breaks=seq(50, 80, by = 5),
labels=c(50, "fifty-five", 60, 65, 70, 75, 80)) +
theme(title = element_text(color = "red", size = 30),
axis.title = element_text(size = 14,face = "bold"),
axis.title.x = element_text(color = "green"),
axis.text = element_text(size = 14),
axis.text.y = element_text(color = "black"),
axis.text.x = element_text(color = "magenta"),
axis.ticks.y = element_blank())
```

2. What happens if you use theme_minimal()[2]?

[2] Install the package *ggthemes* to have many other options.

G. Facets

1. Generate a small multiples graphic with scatter plot of *imr* against *le* for different areas. How do you control the positioning of the facets (horizontal, vertical)?

H. Graphics challenges

1. Generate the graphics in Figures 1 and 2.

   - Hints for Figure 1:
     1) To reorder the countries as a function of a variable use reorder(factor(country),variable).
     2) Use a predefined overall theme.
     3) Delete the major grid lines on the y axis.
     4) The text annotation must be horizontally justified.
   - Hints for Figure 2:
     1) Reorder the countries as a function of *lef*.
     2) Text can be inserted with the command annotate.
     3) Use a predefined overall theme.
     4) Rotate the text on the x axes tick marks by 60 degrees and justify the text horizontally.

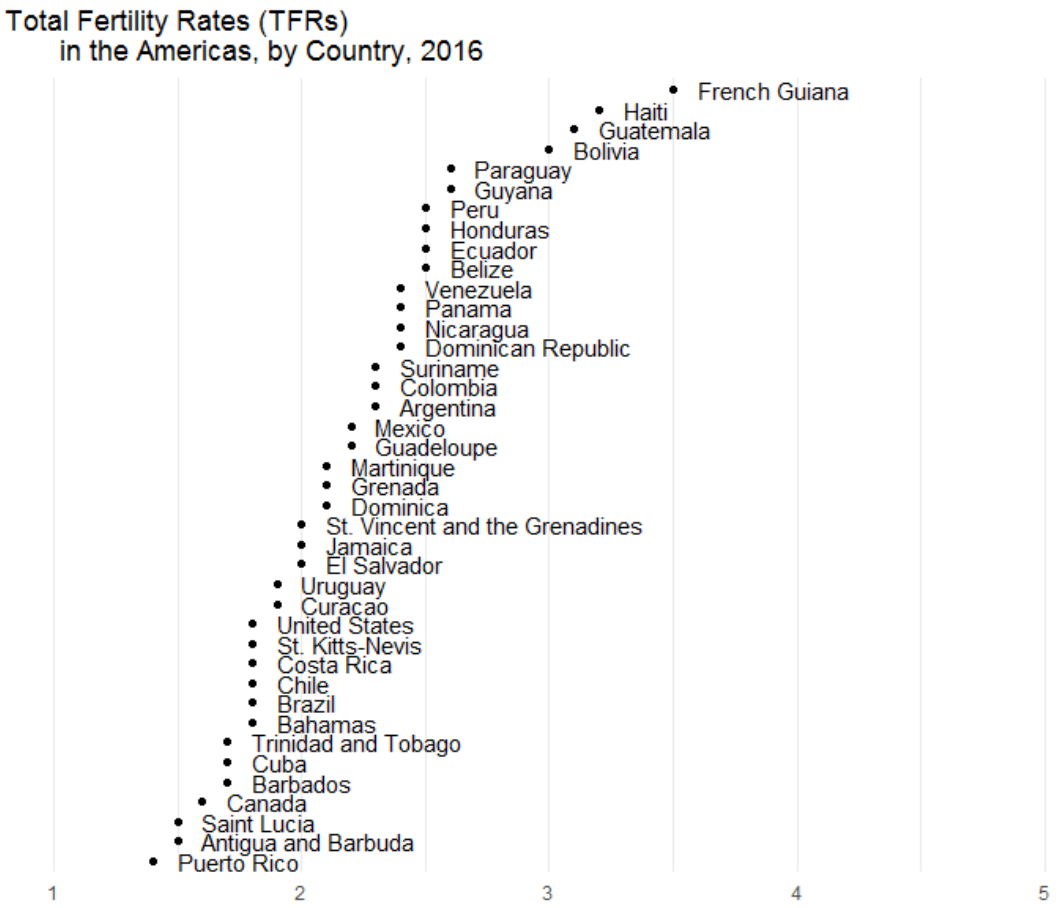

Figure 1: Graphic challenge 1. Total fertility rates for different countries in the Americas.

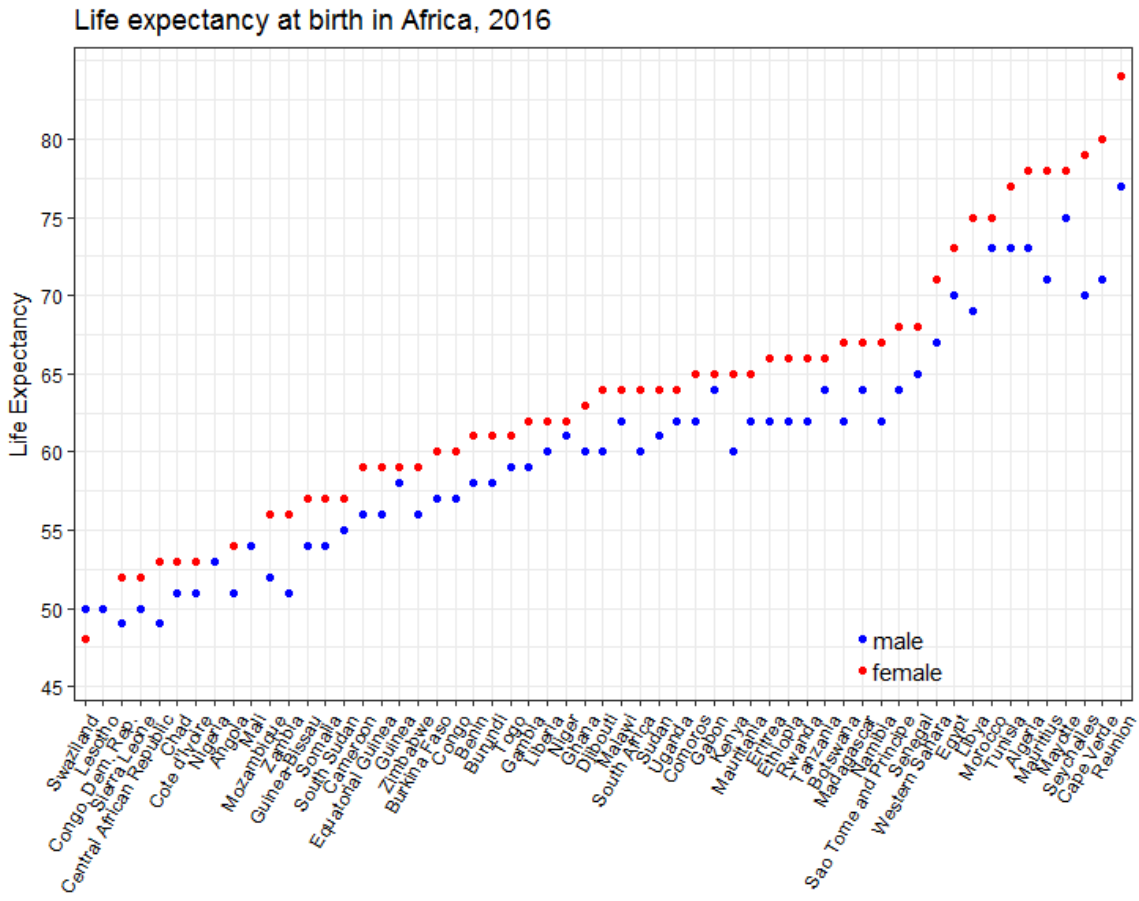

Figure 2: Graphic challenge 2. Life expectancy in years by gender for different countries of Africa.