A global analysis of the relationship between violence and life expectancy

Big Data in Social Sciences

Riccardo Omenti

University of Bologna

Introduction

Measuring **violence** in a country is a challenging task.

- **Violence** is a multifaceted concept
- It depends on many different factors
- Multiple sources are needed to capture its complexities

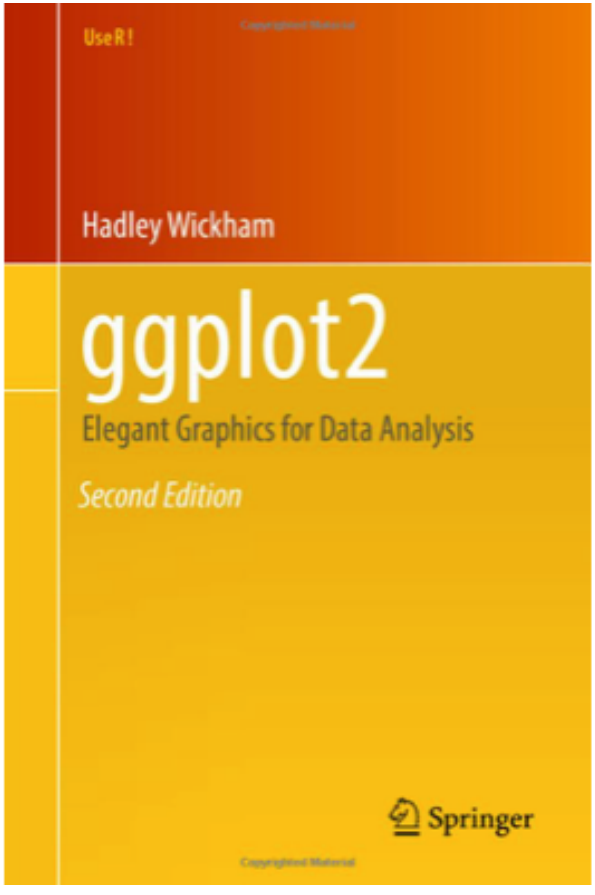
Research Question

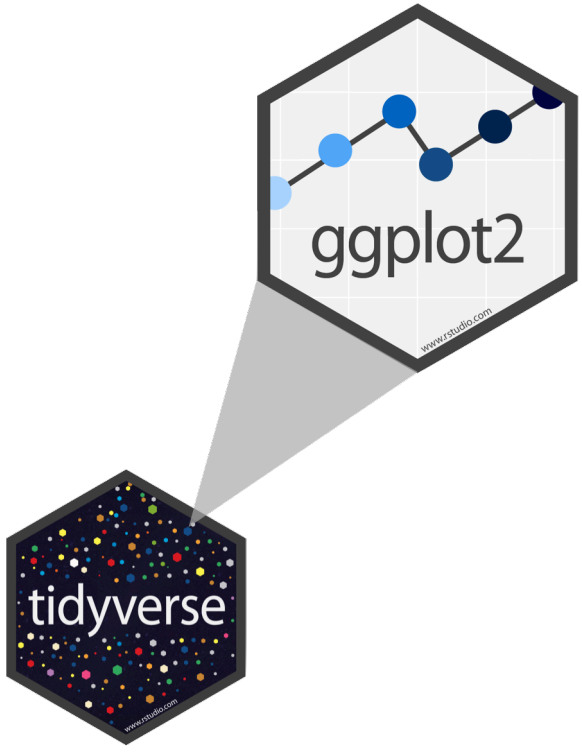
- The level of violence may affect the mortality of a country
- Today, we will rely on two major data sources to explore the relationship between violence and mortality:
  - Data on Violece from the Institute for Economics & Peace
  - Data on Mortality from the United Nations
- We will produce multiple descriptive plots with functions from **ggplot**

What about ggplot?

ggplot2 is an R package for producing statistical, or data, graphics

It is already available in *tidyverse*!





ggplot

Major advantages:

- creating graphs by combining independent components
- detailed theming system to generate nice-looking graphs
- intuitive grammar
- graphs are treated as R objects

Indicators for violence and mortality

Violence $\rightarrow$ Global Peaceful Index (GPI) that measures the violence of a country across three dimensions:

ongoing domestic and international conflict

societal safety and security

militarization

Mortality $\rightarrow$ life expectancy estimates at age 30 for men (e_{30}^{M}) and women (e_{30}^{F}) to capture adult mortality by gender

Data files

Three .csv data files on Virtuale

gpi_data_final.csv $\rightarrow$ GPI for multiple countries over 2008-2023 with regional classification and total population size included

data_male.csv $\rightarrow$ male life expectancy estimates at age 30 for multiple countries over 2008-2023

data_female.csv $\rightarrow$ female life expectancy estimates at age 30 for multiple countries over 2008-2023

Installing packages

Upload data in Rstudio

```
#install.packages('ggthemes')
#install.packages('RColorBrewer')
#install.packages('moderndive')
#install.packages('ggstats')
library('RColorBrewer')
# various qualitative color palettes
library('ggthemes')
# various themes in ggplot
library("tidyverse")
# linear regression in tidyverse
library("moderndive")
# tables
library('knitr')
# plot coefficients
library('ggstats')
library('kableExtra')
```

Upload data sets

```
gpi_data <- read.csv('Data/data_gpi_final.csv')
life_exp_male <- read.csv('Data/male_data.csv')
life_exp_female <- read.csv('Data/female_data.csv')
```

Combine the three data sets

```
data <- gpi_data |>
  inner_join(life_exp_male,by=c("iso3","Year")) |>
  inner_join(life_exp_female,by=c("iso3","Year","country"))
```

We match all records in **gpi_data**, whose iso3 and Year values have a correspondence in both **life_exp_female** and **life_exp_male**

The non-matching records in **gpi_data**, **life_exp_female** and **life_exp_male** are dropped

Question 1

Create a plot displaying the evolution of violence over the time period 2008-2023 by world region

Calculate region- and year-specific levels of violence

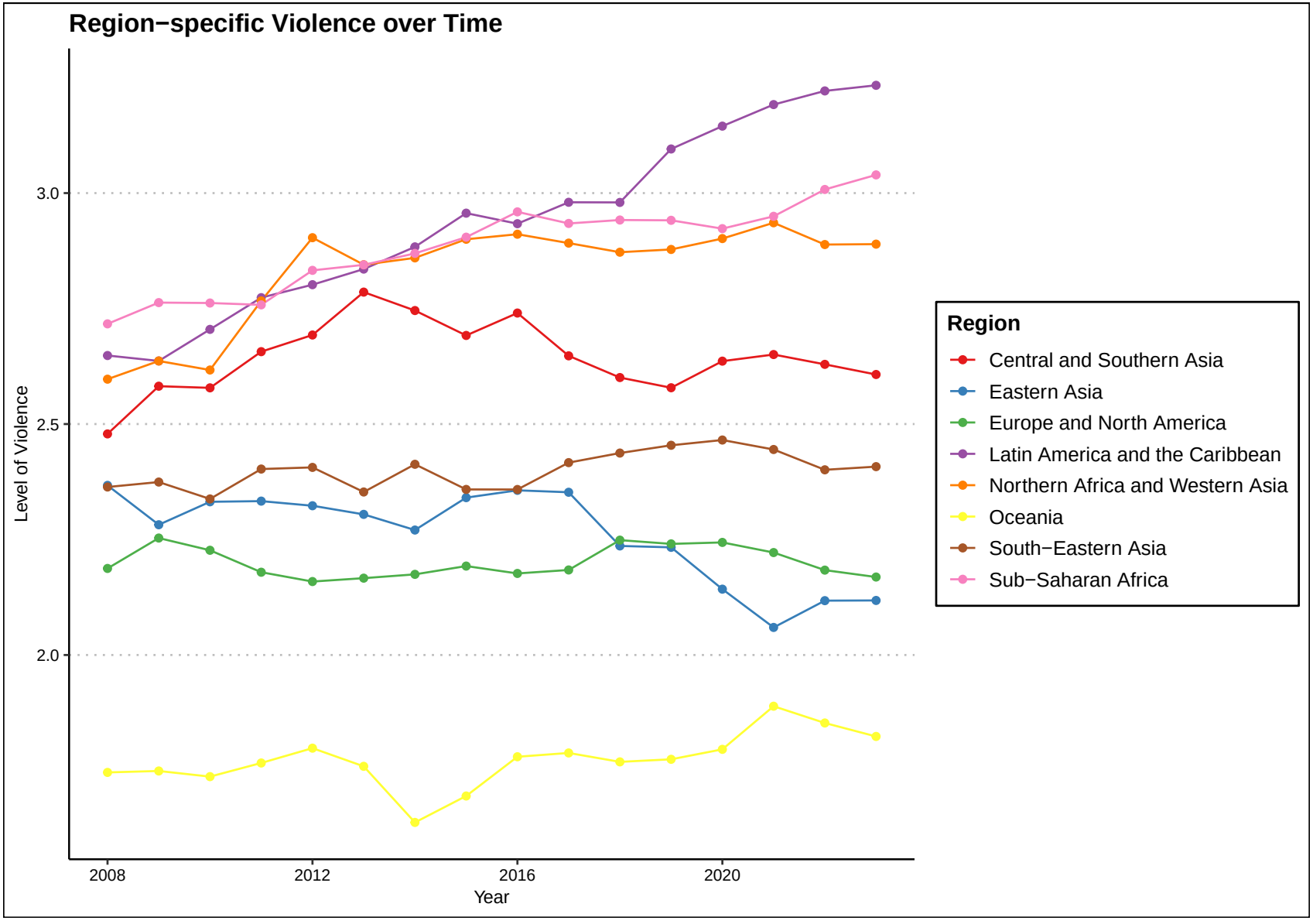
```
data_violence_region <- data |>
      group_by(Year,area) |>
      summarize(gpi=weighted.mean(x=gpi,w=pop,
                                  na.rm = FALSE))
```

Question 1

Generate the plot

```
plot1 <- ggplot(data=data_violence_region,
     mapping=aes(x=Year,y=gpi,color=area))+
geom_line()+
geom_point()+
theme_clean()+
scale_color_brewer(name = "Region", palette = "Set1")+
scale_x_continuous(breaks = seq(2008,2024,4),
                   labels = seq(2008,2024,4))+
xlab('Year')+
ylab('Level of Violence') +
ggtitle('Region-specific Violence over Time')
```

Question 1


Region−specific Violence over Time
Level of Violence
3.0
2.5
2.0
2008
2012
2016
2020
Year
Region
Central and Southern Asia
Eastern Asia
Europe and North America
Latin America and the Caribbean
Northern Africa and Western Asia
Oceania
South−Eastern Asia
Sub−Saharan Africa

Question 2

Create a similar visual aid for life expectancy

Calculate region-specific life expectancy estimates by sex

```
data_life_exp_region <- data|>
  pivot_longer(!c('Year','iso3','pop',
                  'area','gpi','country'),
               names_to = 'sex',
               values_to = 'life_exp') |>
  mutate(sex=substr(sex,10,length(sex))) |>
      group_by(Year,area,sex) |>
      summarize(life_exp=weighted.mean(x=life_exp,
                                       w=pop,
                                       na.rm = FALSE))
```

Question 2

Generate the plot

```
plot2 <- ggplot(data=data_life_exp_region,
mapping=aes(x=Year,y=life_exp,color=area))+
geom_line()+
geom_point()+
annotate("text",x=2020,y=55,label="COVID-19")+
theme_clean()+
facet_wrap(~sex) +
xlab('Year')+
ylab('Life Expectancy at Age 30') +
scale_color_brewer(name = "Region", palette = "Set1")+
ggtitle('Region-specific Life Expectancy over Time')
```

Question 2


Region−specific Life Expectancy over Time
female
male
COVID−19
COVID−19
55
50
45
40
Life Expectancy at Age 30
2008
2012
2016
2020
2008
2012
2016
2020
Year
Region
Central and Southern Asia
Eastern Asia
Europe and North America
Latin America and the Caribbean
Northern Africa and Western Asia
Oceania
South−Eastern Asia
Sub−Saharan Africa

Question 3

Produce two scatter plots to display the relationship between violence and life expectancy at age 30 in 2023 by sex. Add also a regression line for each plot.

Select records for years 2033

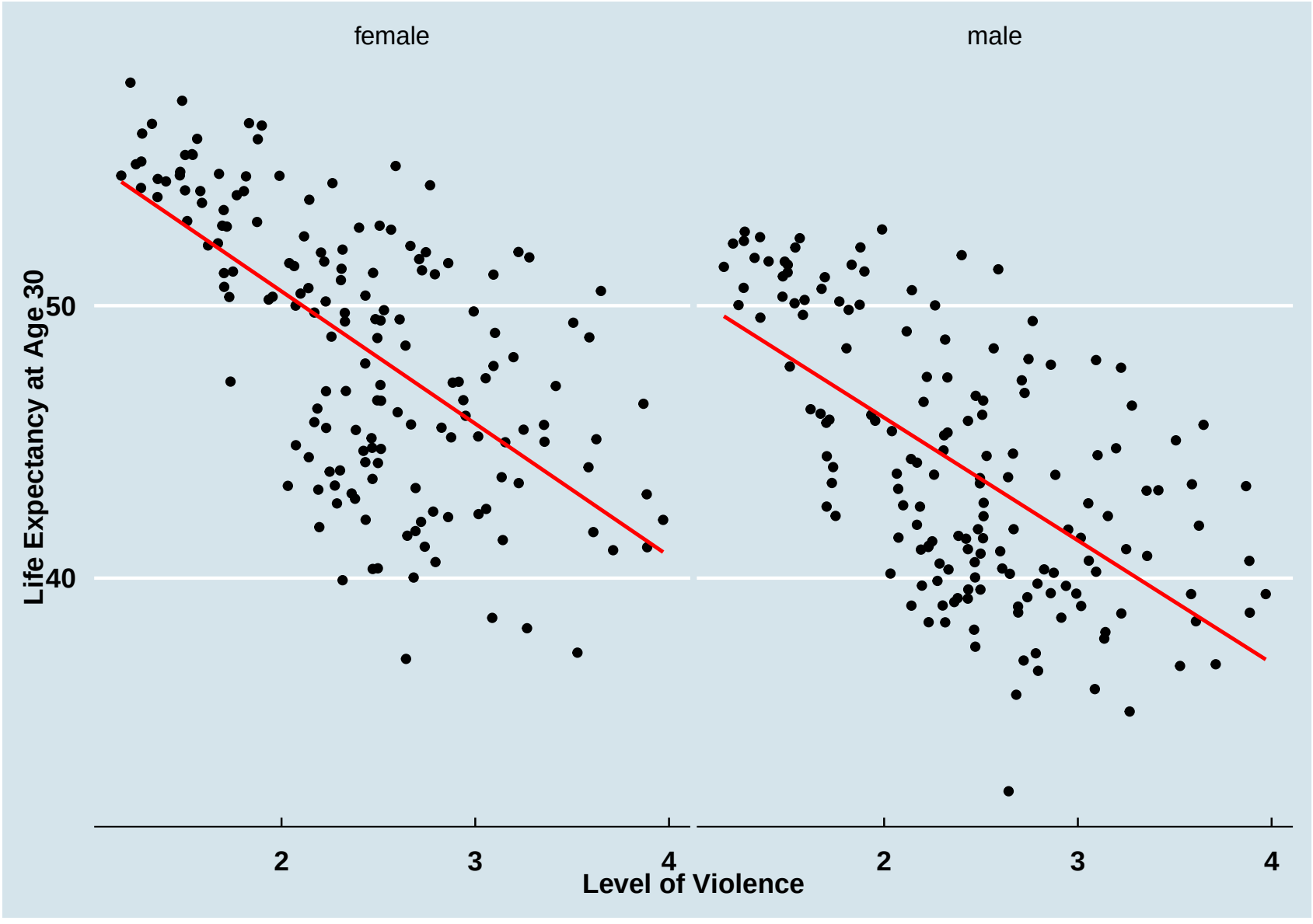
```
data_scatter <- data|>
  pivot_longer(!c('Year','iso3','pop',
                  'area','gpi','country'),
               names_to = 'sex',
               values_to = 'life_exp') |>
  mutate(sex=substr(sex,10,length(sex))) |>
  filter(Year==2023)
```

Question 3

Generate the plot

```
plot3 <- ggplot(data=data_scatter,
mapping=aes(x=gpi,y=life_exp))+
geom_point(size=2)+
xlab('Level of Violence')+
ylab('Life Expectancy at Age 30')+
theme_economist()+
geom_smooth(method = "lm", se = FALSE,color='red')+
facet_wrap(~sex) +
theme(axis.text.y = element_text(size=15,face="bold"),
axis.title.y = element_text(size=15,face="bold"),
axis.text.x = element_text(size=15,face="bold"),
axis.title.x = element_text(size=15,face="bold"))
```

Question 3


female
male
Life Expectancy at Age 30
50
40
2
3
4
Level of Violence

Question 4

Display the distribution of male life expectancy in the 20 most violent countries and in the 20 most peaceful countries in 2023

Let's create the data sets Most violent countries

```
data_most_violence <- data |>
      filter(Year==2023) |>
      slice_max(gpi,n=20) |>
      mutate(label='Most Violent')
```

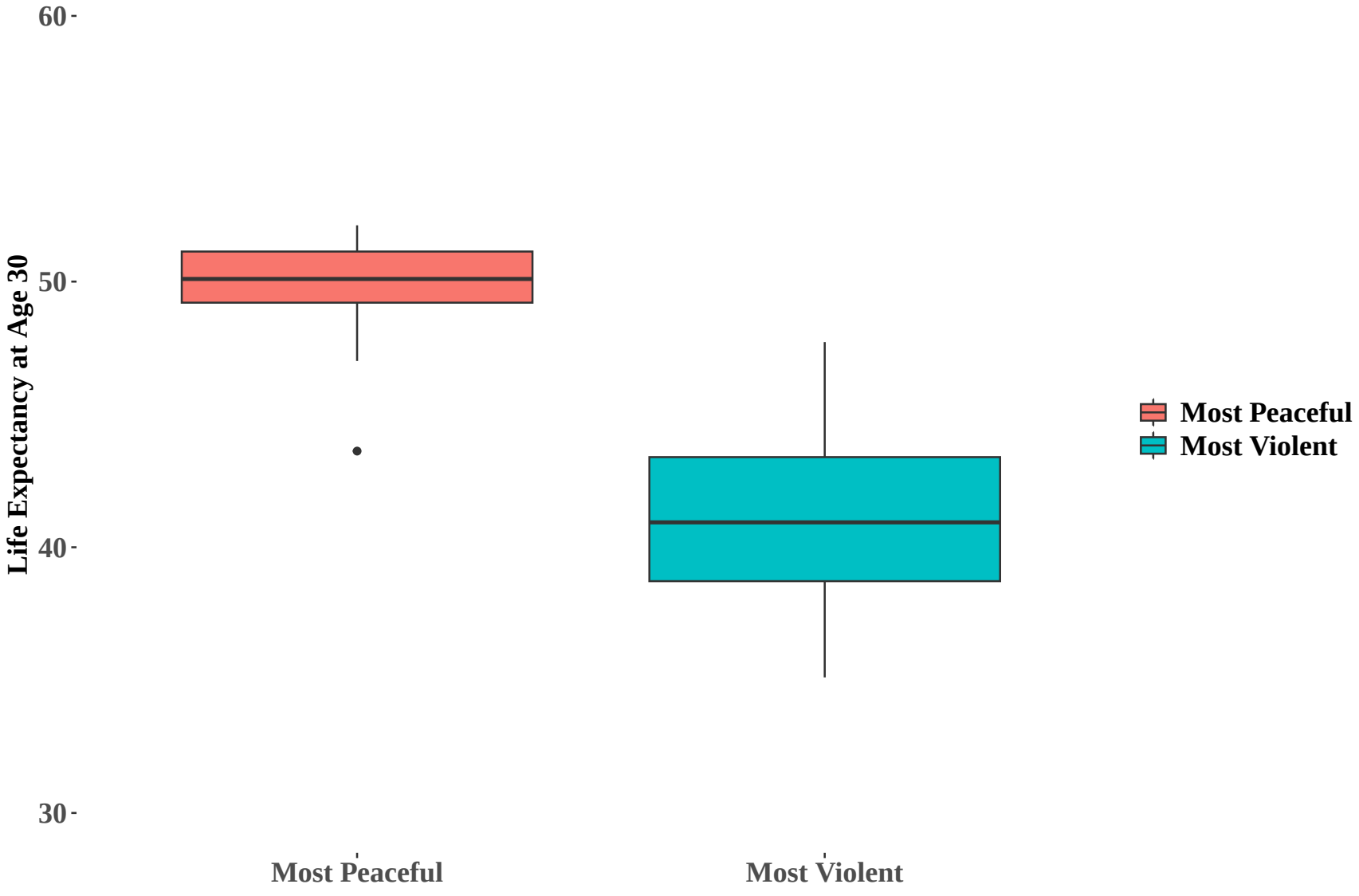
Most peaceful countries

```
data_most_peaceful <- data |>
filter(Year==2022) |>
slice_min(gpi,n=20) |>
mutate(label='Most Peaceful')
```

combine the two data sets by row

```
data_plot <- rbind(data_most_violence,data_most_peaceful)
```

Question 4

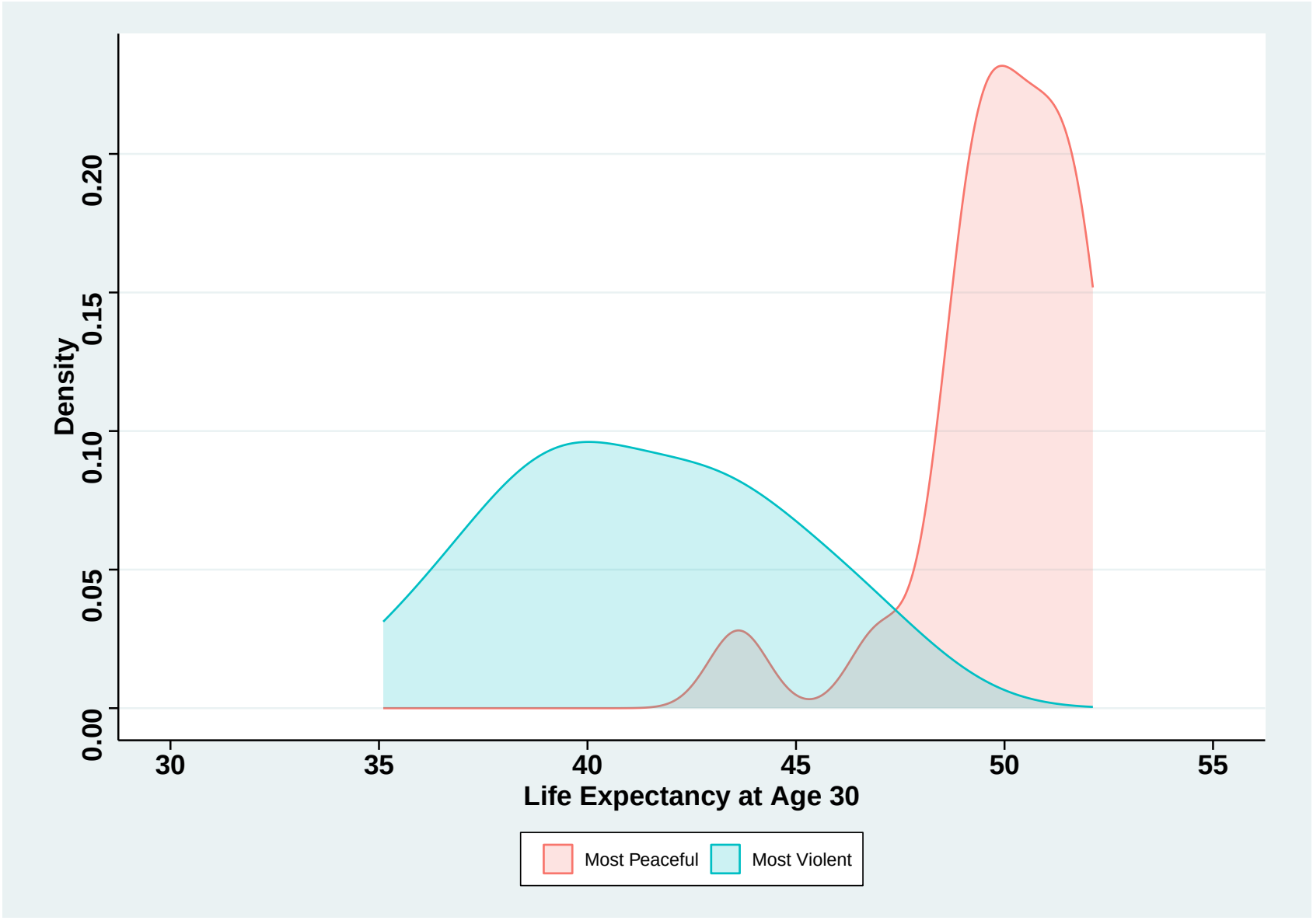

60
50
40
30
Life Expectancy at Age 30
Most Peaceful
Most Violent
Most Peaceful
Most Violent

Question 5

Perform the same task using a different visual aid.

```
plot5 <- ggplot(data=data_plot,
aes(x=life_exp_male,color=label,fill=label))+
geom_density(alpha = 0.2, na.rm = TRUE) +
theme_stata()+
scale_fill_discrete(name='')+
scale_color_discrete(name='')+
coord_cartesian(xlim=c(30,55))+
scale_x_continuous(breaks=seq(30,55,5),
                   labels=seq(30,55,5))+
ylab('Density')+
xlab('Life Expectancy at Age 30')+
theme(axis.text.y = element_text(size=15,face="bold"),
legend.position = "bottom",
axis.title.y = element_text(size=15,face="bold"),
axis.text.x = element_text(size=15,face="bold"),
axis.title.x = element_text(size=15,face="bold"))
```

Question 5


Density
0.00
0.05
0.10
0.15
0.20
30
35
40
45
50
55
Life Expectancy at Age 30
Most Peaceful
Most Violent

Question 6

Produce a scatter plot to display the relationship between violence and life expectancy in 2023 by sex.

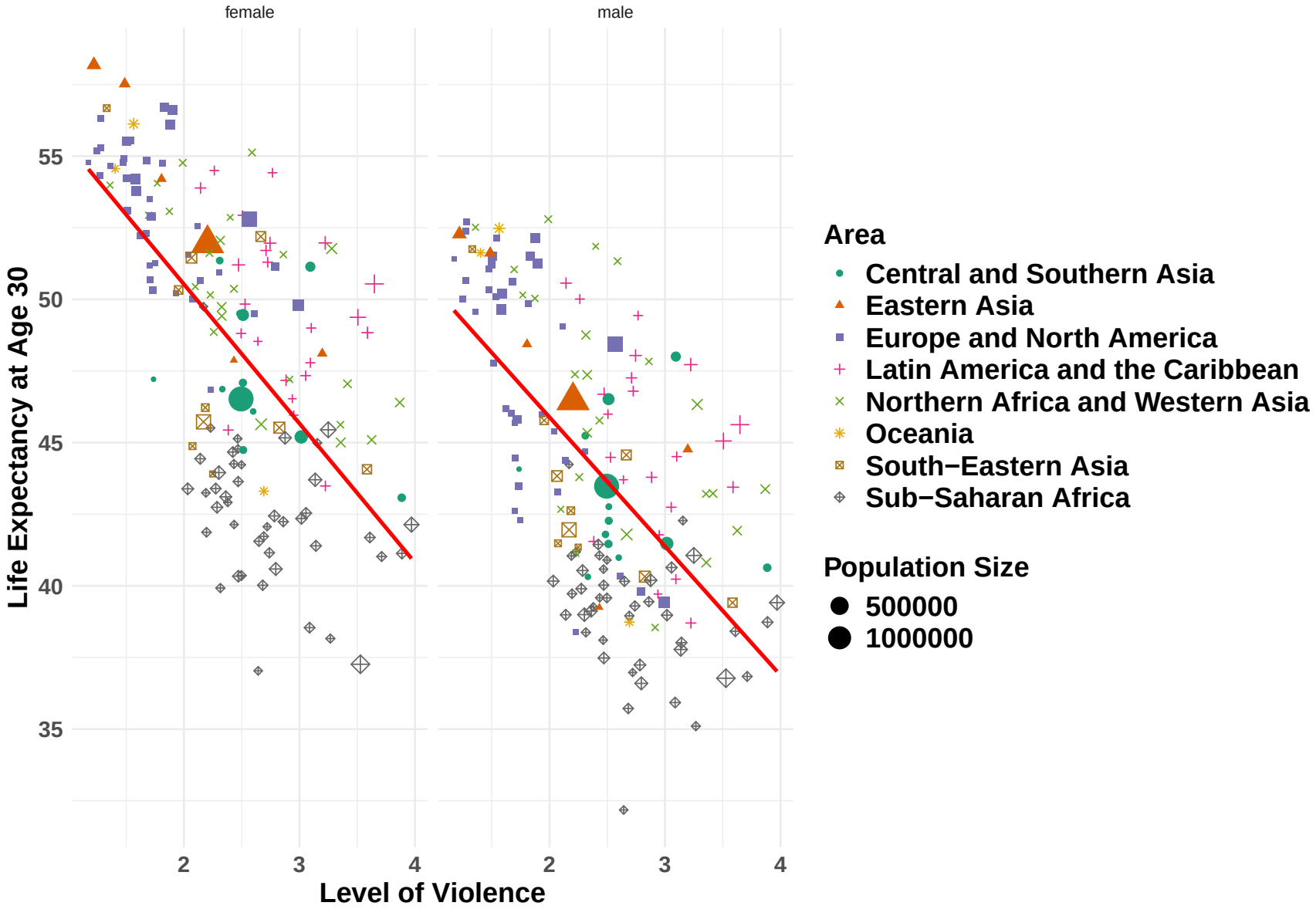
Make sure to set different shapes and colors for the points according to the region where they are located.

Fix the size of the points according to the population size of the country

Question 6

```
plot6 <- ggplot(data=data_scatter,
mapping=aes(x=gpi,y=life_exp))+
geom_point(aes(shape=area,color=area,size=pop))+
xlab('Level of Violence')+
ylab('Life Expectancy at Age 30')+
theme_minimal()+
geom_smooth(method = "lm", se = FALSE,color='red')+
labs(size = 'Population Size')+
scale_shape_manual(name   = "Area",
                   values = c(16, 17, 15,
                              3, 4, 8, 7, 9)) +
scale_color_brewer(name = "Area",palette = "Dark2") +
scale_y_continuous(breaks = seq(30,55,5),labels = seq(30,55,5))+
facet_wrap(~sex) +
theme(axis.text.y = element_text(size=12,face="bold"),
axis.title.y = element_text(size=15,face="bold"),
legend.title = element_text(size=15,face="bold"),
legend.text = element_text(size=15,face="bold"),
axis.text.x = element_text(size=12,face="bold"),
axis.title.x = element_text(size=15,face="bold"))
```

Question 6


female
male
Life Expectancy at Age 30
55
50
45
40
35
2
3
4
Level of Violence
Area
Central and Southern Asia
Eastern Asia
Europe and North America
Latin America and the Caribbean
Northern Africa and Western Asia
Oceania
South−Eastern Asia
Sub−Saharan Africa
Population Size
500000
1000000

Question 7

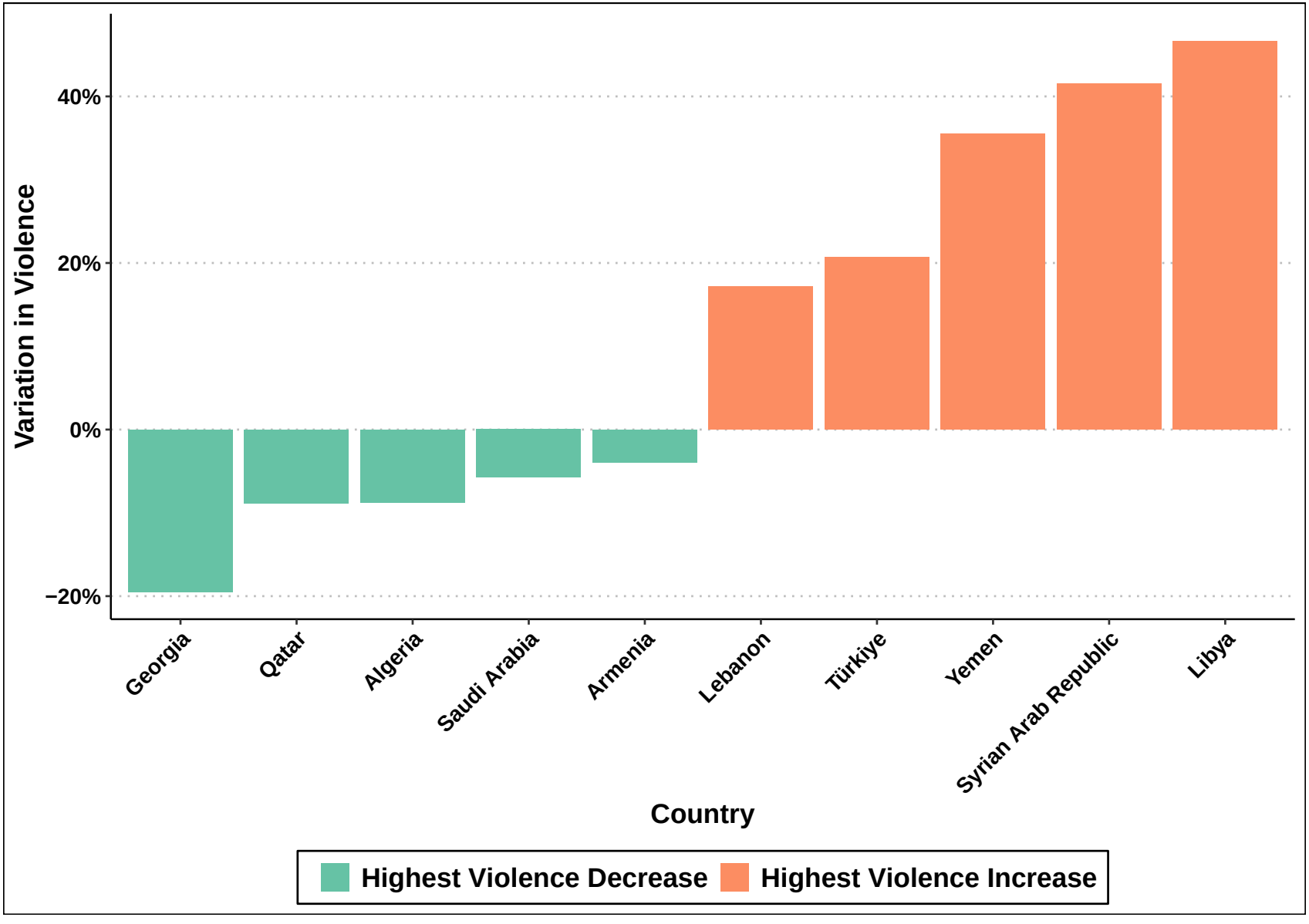
Display the variation in the level of violence between 2008 and 2023 for the top 5 countries experiencing the largest increase and the top 5 showing the largest decrease in violence in North Africa and Western Asia.

```
data_high <- data |>
filter(Year %in% c(2008,2023)) |>
select(Year,country,area,gpi) |>
pivot_wider(names_from='Year',values_from='gpi') |>
mutate(var=(`2023`-`2008`)/`2008`) |>
group_by(area)|>
slice_max(var,n=5) |>
mutate(label='Highest Violence Increase')
data_low <- data |>
filter(Year %in% c(2008,2023)) |>
select(Year,country,area,gpi) |>
pivot_wider(names_from='Year',values_from='gpi') |>
mutate(var=(`2023`-`2008`)/`2008`) |>
group_by(area)|>
slice_min(var,n=5) |>
mutate(label='Highest Violence Decrease')
```

Question 7

```
plot7 <- rbind(data_high,data_low) |>
  filter(area=='Northern Africa and Western Asia') |>
  ggplot(aes(x=reorder(country,var),y=var,fill=label))+
  geom_bar(stat='identity') +
  theme_clean()+
  xlab('Country')+
  ylab('Variation in Violence')+
  scale_fill_brewer(name='',palette='Set2')+
  scale_y_continuous(labels = scales::percent)+
  theme(axis.text.x = element_text(angle = 45,
  vjust = 1, hjust = 1, size = 12, face = "bold"),
        legend.position = 'bottom',
        axis.text.y = element_text(size = 12, face = "bold"),
        axis.title.x = element_text(size = 15, face = "bold"),
        axis.title.y = element_text(size = 15, face = "bold"),
        legend.text = element_text(size = 15, face = "bold"))
```

Question 7


40%
20%
0%
−20%
Variation in Violence
Georgia
Qatar
Algeria
Saudi Arabia
Armenia
Lebanon
Türkiye
Yemen
Syrian Arab Republic
Libya
Country
Highest Violence Decrease
Highest Violence Increase

Save a plot

ggsave() allows to save plots as images in different formats (e.g. .png, .pdf, .jpeg)

```
ggsave(filename = "Results/plot7.pdf", # name of the file
       plot = plot7,
       height = 20,
       width = 40,
       units = "cm",
       dpi = 400)
```

Question 8

Obtain a measure of the impact of the violence on life expectancy at age 30 in Year 2023.

We can fit a linear regression model

$$e^{30}_{c,s} = \beta_0 + \beta_1 \cdot GPI_c + \beta_2 \cdot Male_s + \beta_3 \cdot Area_c + \epsilon_{t,s}$$

```
data_scatter <- data_scatter |>
  mutate(
    sex  = relevel(as.factor(sex), ref = "female"),
    area = relevel(as.factor(area), ref = "Sub-Saharan Africa")
  ) |>
  filter(Year==2023)

model = lm(life_exp~gpi+sex+area,
           data=data_scatter,
           weights = pop)
```

Question 8

Show the output of the model in tabular form

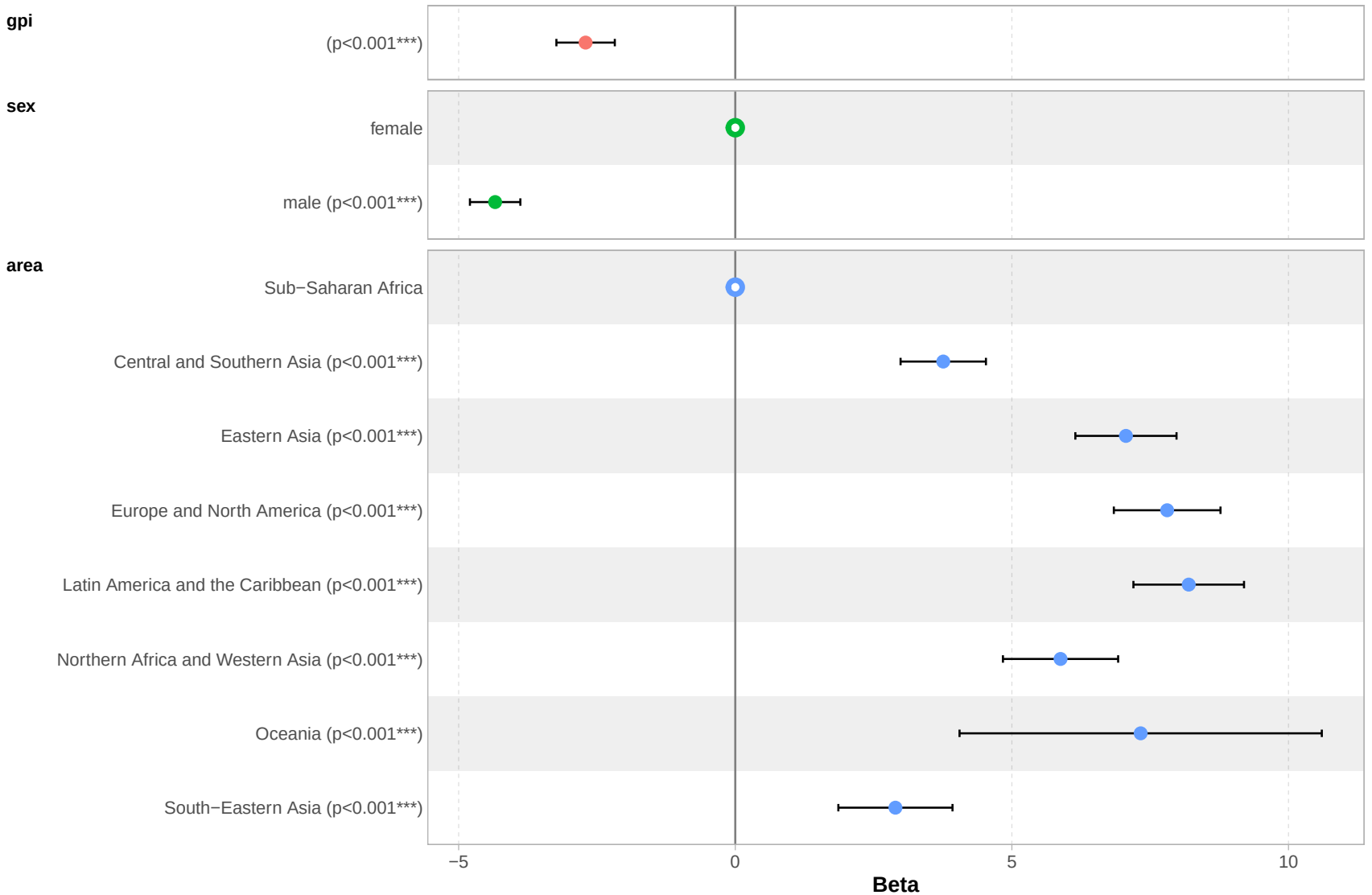
```
get_regression_table(model)
```

```
# A tibble: 10 x 7
   term                      estimate std_error statistic p_value l
   <chr>                        <dbl>     <dbl>     <dbl>   <dbl>
 1 intercept                    50.6      0.876     57.8          0
 2 gpi                          -2.70     0.268    -10.1          0
 3 sex: male                    -4.34     0.232    -18.7          0
 4 area: Central and Sou~        3.76     0.392      9.59         0
 5 area: Eastern Asia            7.06     0.465     15.2          0
 6 area: Europe and Nort~        7.81     0.49      15.9          0
 7 area: Latin America a~        8.20     0.508     16.1          0
 8 area: Northern Africa~        5.88     0.529     11.1          0
 9 area: Oceania                 7.33     1.66       4.40         0
10 area: South-Eastern A~        2.90     0.524      5.52         0
```

Question 8

Show the output of the model in graphical form

```
ggcoef_model(model)
```



Final remarks

What I presented today is just the tip of the iceberg.

For other exciting visualizations see the online book https://ggplot2-book.org