

Lab 5: Evolution of a ggplot

Elijah Russell

In this lab, we'll practice recreating some early versions of a nice looking plot that describes how student-to-teacher ratios in primary education vary across the world. **The process of creating a detailed, visually appealing final plot is iterative and additive**, as we will see here.

You should be *following a similar process as you work to create your submission for the Mini Project #1* (i.e., swapping out and layering geoms, adjusting axes to include relevant comparison points, and adding some annotations).

The data comes from the #TidyTuesday project and was shared on May 7, 2019. It contains the UNESCO Institute of Statistics' country-level data on the number of teachers and teacher-to-student ratios in primary and secondary education courses. Our goal will be to **describe how the variation in student-to-teacher ratios in primary schools across continents (and we want to show variability within each continent)**.

I've included some lines below to read in the data. Your job is to recreate the plots in the *Lab_5_Target_Plots.pdf* file (see the Files tab in the bottom right panel).

1 Data wrangling

The chunk in this section creates the dataset you'll use in all your plots. *Describe in words what this code chunk does.*

This code chunk gathers all the data from primary education schools in `df_students`, and takes the data from the latest year of each country and joins to it the data for each country from `df_world_tile`.

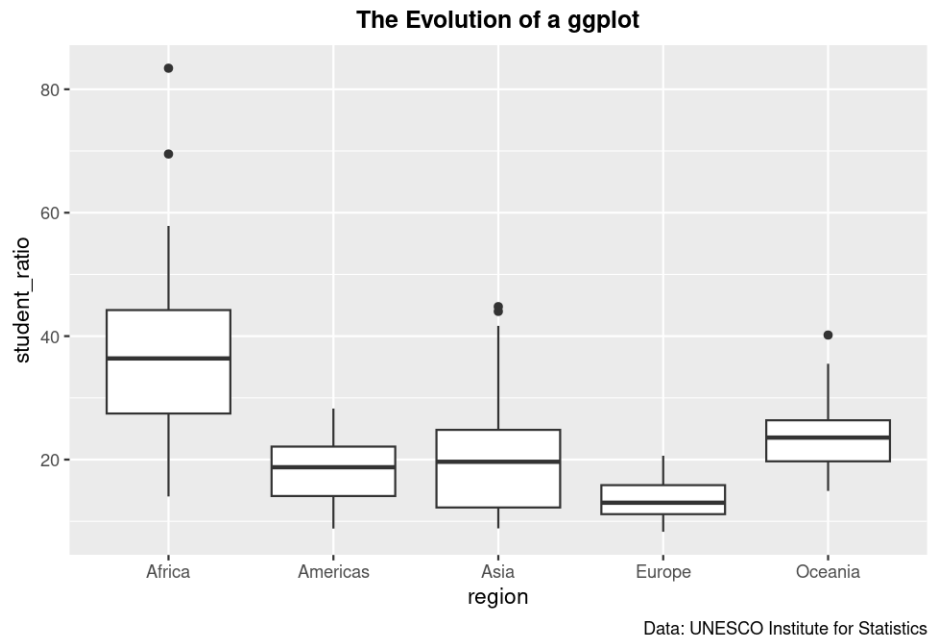
```
df_ratios <- df_students %>%  
  filter( indicator == "Primary Education" ) %>%  
  group_by(country) %>%  
  filter(year == max(year)) %>%  
  inner_join( df_world_tile,  
             by = join_by( country_code == alpha.3 ))
```

2 Basic boxplot

```
# Consider modifying the following example code:  
# ASIDE: put your modified code in its own chunk OR  
# remove the `eval = FALSE` flag on this chunk  
  
mylabs <- labs( title = "Example title",  
               caption = "Data source" )  
mythemes <- theme( plot.title = element_text(  
  face = "italic", size = 20, hjust = .5 ))  
  
ggplot(df_ratios,  
       aes( x, -y, col=region) ) +  
  geom_point()
```

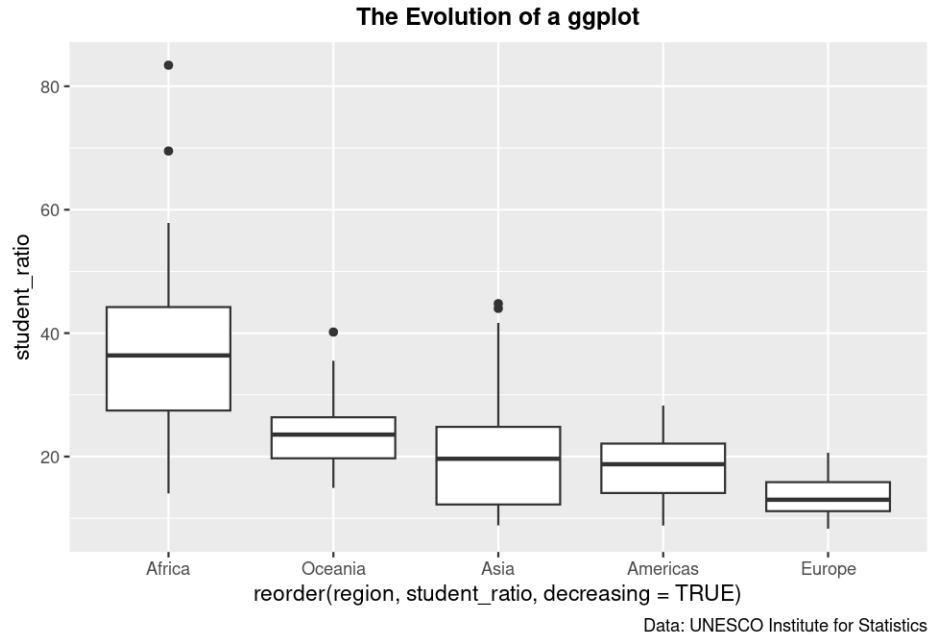
```
mylabs +  
mythemes
```

```
ggplot(df_ratios, aes(x=region, y=student_ratio)) +  
  geom_boxplot() +  
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics") +  
  theme(plot.title = element_text(  
    face = "bold", size = 12, hjust = .5 ))
```



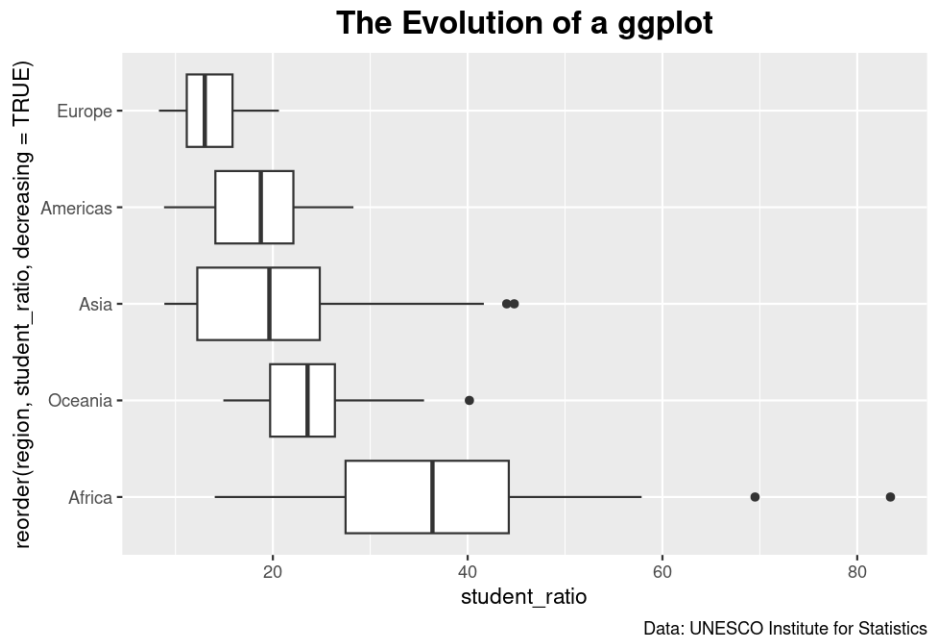
2.1 A more sensible ordering

```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio)) +  
  geom_boxplot() +  
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics") +  
  theme(plot.title = element_text(  
    face = "bold", size = 12, hjust = .5 ))
```



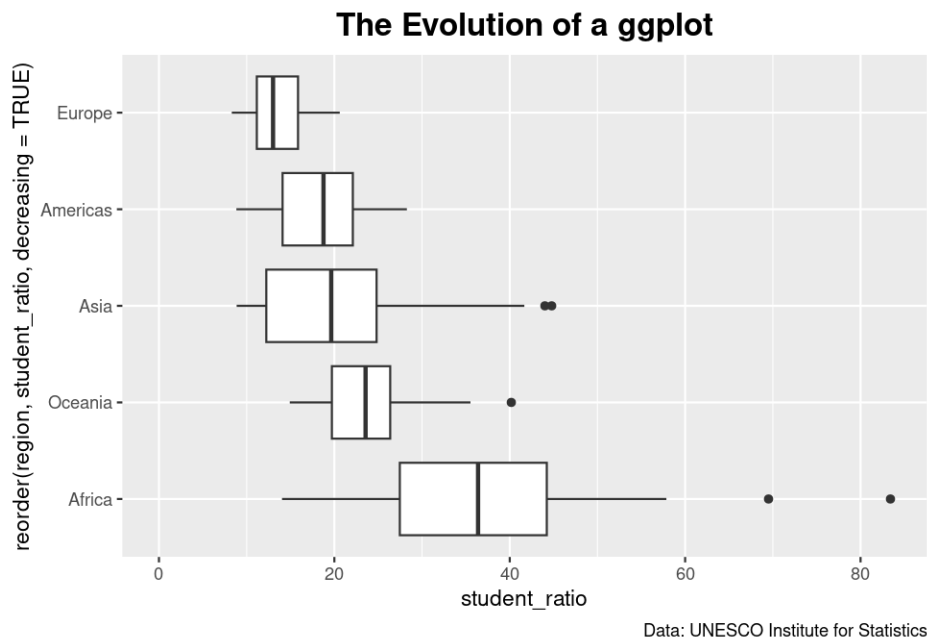
2.2 Flipped axes

```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio)) +
  geom_boxplot() +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics") +
  theme(plot.title = element_text(
    face = "bold", size = 16, hjust = .5 )) +
  coord_flip()
```



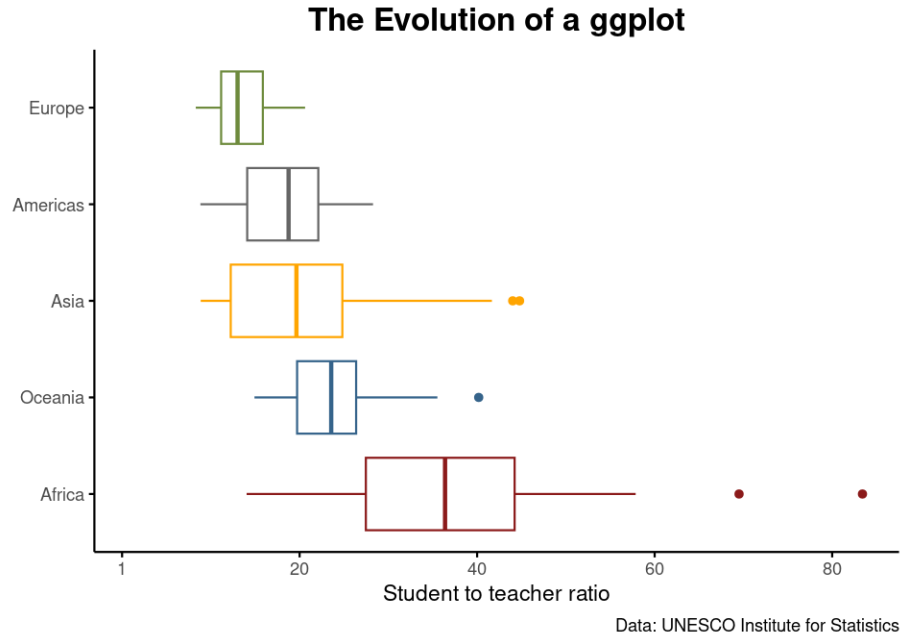
2.3 A relevant minimum

```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio)) +
  geom_boxplot() +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics") +
  theme(plot.title = element_text(
    face = "bold", size = 16, hjust = .5 )) +
  coord_flip() +
  scale_y_continuous( limits = c(0,max(df_ratios$student_ratio)) )
```



2.4 Color! Theme! Better indication of the minimum!

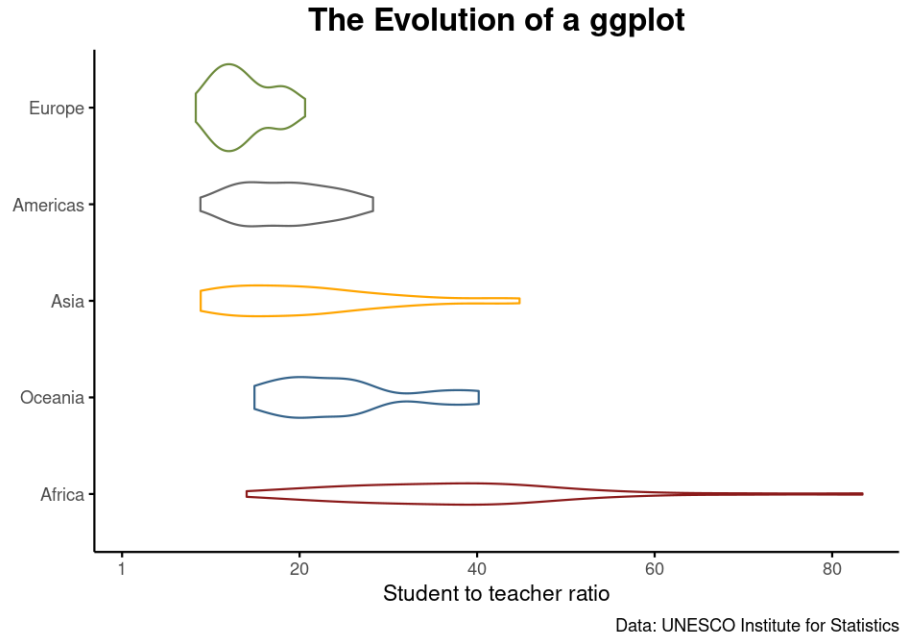
```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_boxplot() +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "student_ratio") +
  coord_flip() +
  scale_y_continuous(limits = c(1,max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80") ) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "darkblue", Africa = "darkred")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))
```



3 Beyond the boxplot

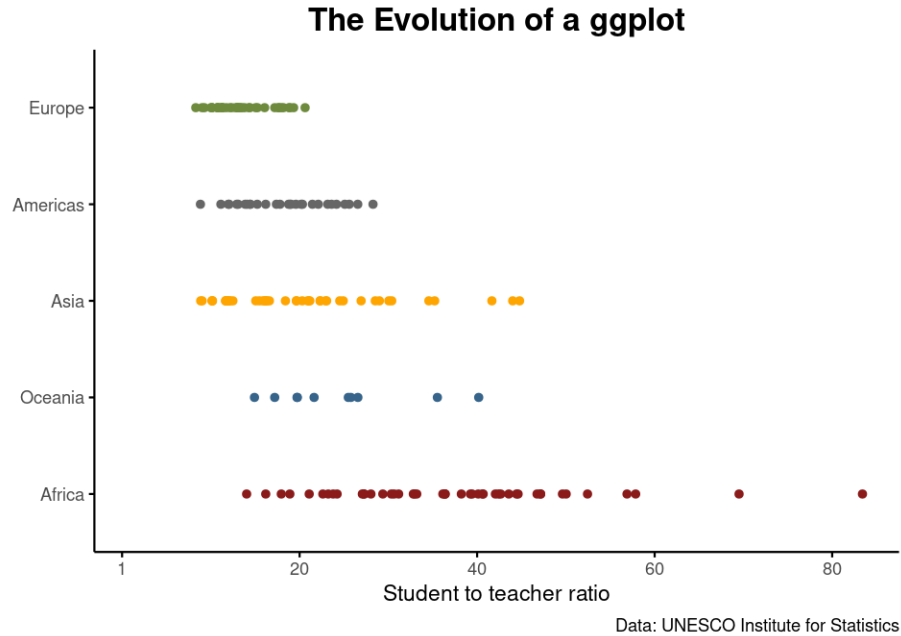
3.1 Distribution shapes

```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_violin() +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "Student to teacher ratio") +
  coord_flip() +
  scale_y_continuous(limits = c(1, max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80") ) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "steelblue", Africa = "firebrick")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))
```



3.2 Countries

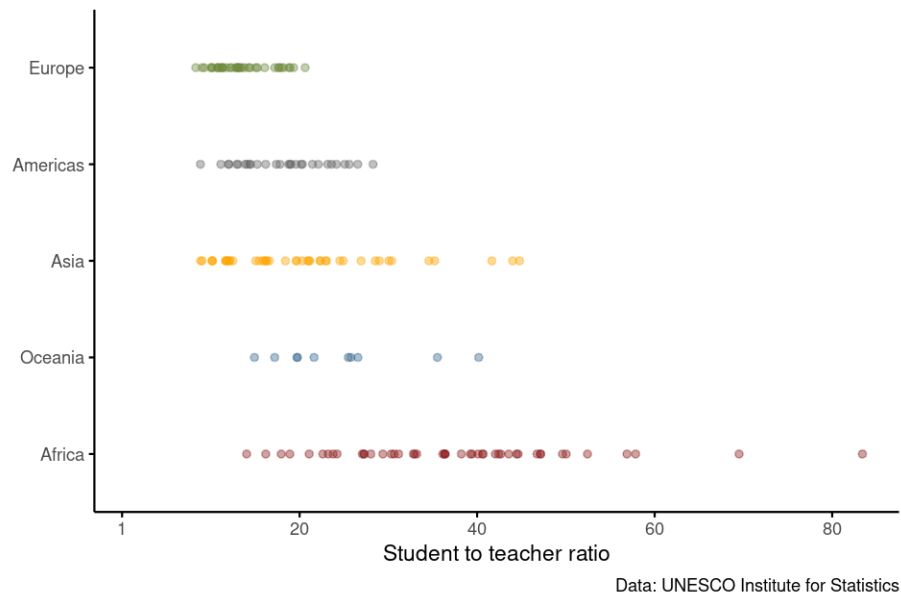
```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_point() +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "") +
  coord_flip() +
  scale_y_continuous(limits = c(1, max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80") ) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "steelblue", Africa = "firebrick")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))
```



3.3 Fix overplotting

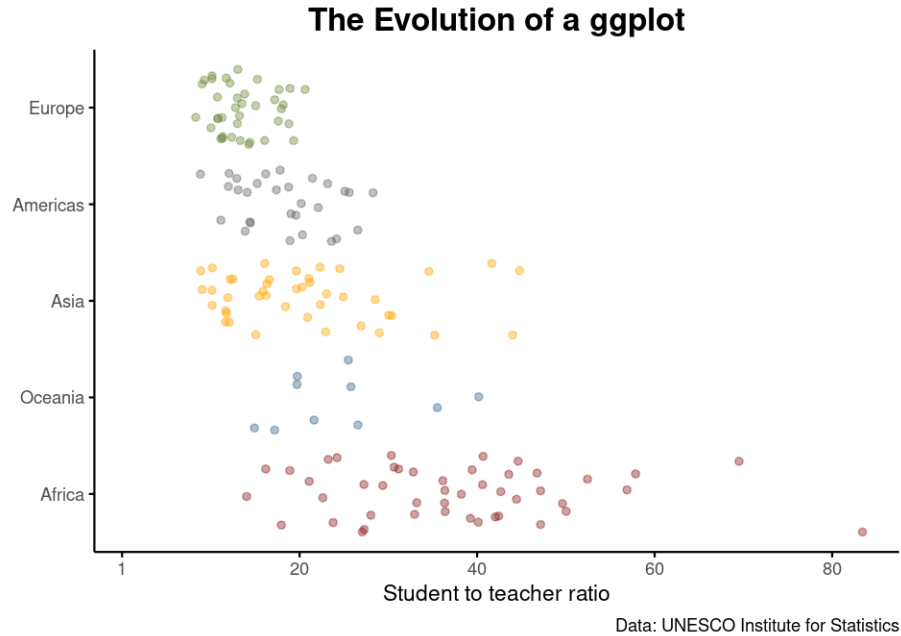
```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_point(alpha=0.4) +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "") +
  coord_flip() +
  scale_y_continuous(limits = c(1, max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80") ) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "blue", Africa = "darkred")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))
```

The Evolution of a ggplot



3.4 Better fix for overplotting

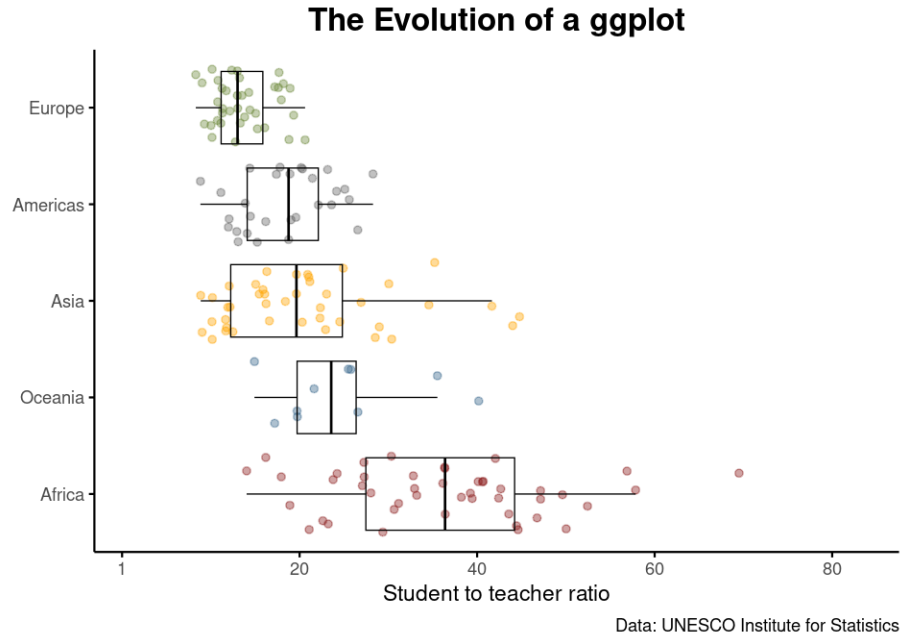
```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_jitter(alpha=0.4) +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "") +
  coord_flip() +
  scale_y_continuous(limits = c(1, max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80")) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "blue", Africa = "red")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5))
```



3.5 Return of the boxplot

```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_jitter(alpha=0.4) +
  geom_boxplot(color = "black", fill = NA, outlier.shape = NA, linewidth = 0.3) +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "Student to teacher ratio") +
  coord_flip() +
  scale_y_continuous(limits = c(1, max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80") ) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "blue", Africa = "red")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



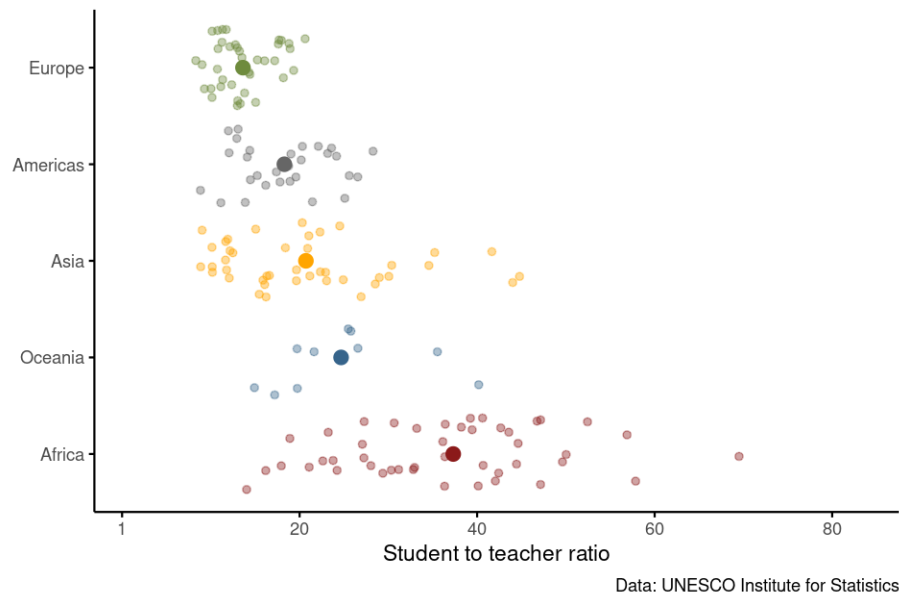
4 Final tweaks

4.1 Continental averages

```
ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_jitter(alpha=0.4) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "Student to teacher ratio") +
  coord_flip() +
  scale_y_continuous(limits = c(1, max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80")) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "blue", Africa = "red")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

The Evolution of a ggplot



4.2 World average

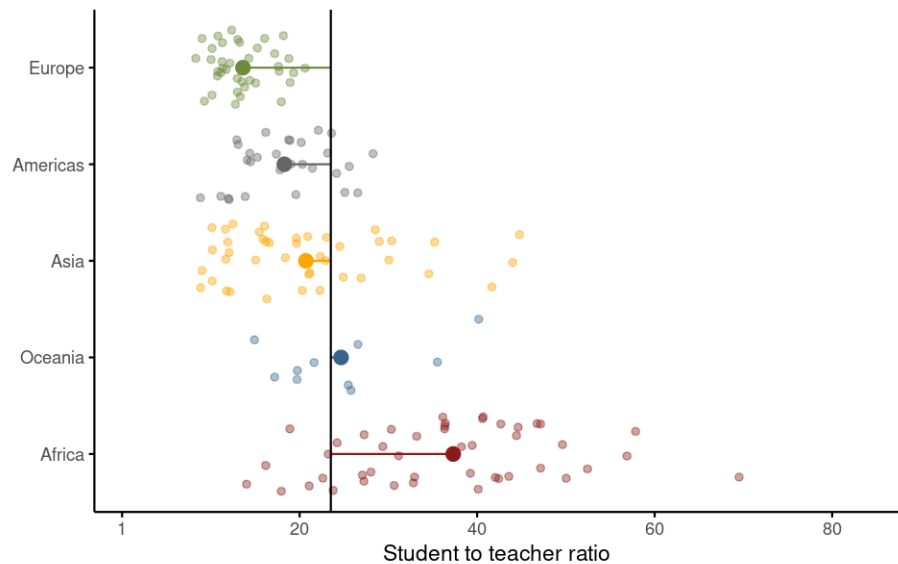
```
world_avg <- mean(df_ratios$student_ratio)

asia_avg <- mean(df_ratios$student_ratio[df_ratios$region == "Asia"], na.rm = TRUE)
africa_max <- max(df_ratios$student_ratio[df_ratios$region == "Africa"], na.rm = TRUE)
africa_min <- min(df_ratios$student_ratio[df_ratios$region == "Africa"], na.rm = TRUE)

ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio, color=region)) +
  geom_jitter(alpha=0.4) +
  geom_segment(data = df_ratios %>%
    group_by(region) %>%
    summarize( student_ratio = mean(student_ratio) ),
    yend = world_avg) +
  geom_hline(yintercept = world_avg) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "") +
  coord_flip() +
  scale_y_continuous(limits = c(1,max(df_ratios$student_ratio)),
    labels = c("1", "20", "40", "60", "80") ) +
  scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "steelblue", Africa = "firebrick")) +
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

The Evolution of a ggplot



Data: UNESCO Institute for Statistics

Some more example code:

```
world_avg <- mean(df_ratios$student_ratio)

ggplot( df_ratios,
        aes( x = region,
              y = student_ratio )) +
  geom_jitter(alpha=.5) +
  stat_summary(fun = mean, geom = "point", size = 5) +
  geom_segment( data = df_ratios %>%
                group_by(region) %>%
                summarize( student_ratio = mean(student_ratio) ),
                yend = world_avg )
```

4.3 Annotations

```
world_avg <- mean(df_ratios$student_ratio)

asia_avg <- mean(df_ratios$student_ratio[df_ratios$region == "Asia"], na.rm = TRUE)
africa_max <- max(df_ratios$student_ratio[df_ratios$region == "Africa"], na.rm = TRUE)
africa_min <- min(df_ratios$student_ratio[df_ratios$region == "Africa"], na.rm = TRUE)

plot_for_annotations <- ggplot(df_ratios, aes(x=reorder(region, student_ratio, decreasing=TRUE), y=student_ratio)) +
  geom_jitter(alpha=0.4) +
  geom_segment(data = df_ratios %>%
                group_by(region) %>%
                summarize( student_ratio = mean(student_ratio) ),
                yend = world_avg) +
  geom_hline(yintercept = world_avg) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  labs(title = "The Evolution of a ggplot", caption = "Data: UNESCO Institute for Statistics", x = "", y = "Student to teacher ratio") +
  coord_flip() +
```

```

scale_y_continuous(limits = c(1,max(df_ratios$student_ratio)),
  labels = c("1", "20", "40", "60", "80") ) +
scale_color_manual(values = c(Europe = "darkolivegreen4", Americas = "grey40", Asia = "orange", Oceania = "darkblue"),
  theme_minimal() +
  theme(legend.position = "none",
    panel.grid = element_blank(),
    axis.line = element_line(color = "black"),
    panel.border = element_blank(),
    axis.ticks = element_line(color = "black"),
    plot.title = element_text(face = "bold", size = 16, hjust = .5 ))

txt_nent <- data.frame( xstart = 2.6, ystart = 15,
  xend = 3, yend = asia_avg - .5,
  labels = "Continental average ",
  hjust = 1, vjust = .5 )

txt_tries <- data.frame( xstart = 1.3, ystart = 8,
  xend = 1, yend = 13,
  labels = "Countries \n per continent \n",
  hjust = .5, vjust = 0 )
txt_tries2 <- txt_tries %>%
  mutate( xend = 1.76, yend = 18, labels = "" )

txt_avg <- data.frame( xstart = 4.5, ystart = 30,
  xend = 4.5, yend = 25,
  labels = "Worldwide average:\n23.5 students per teacher",
  hjust = 1, vjust = .5 )

txt_max <- data.frame( xstart = 2, ystart = 77,
  xend = 1, yend = africa_max,
  labels = "The Central African Republic has by\nfar the most students per teacher",
  hjust = .5, vjust = 0 )

arrows_and_text <- rbind( txt_nent,
  txt_tries, txt_tries2,
  txt_max )

plot_for_annotations +
  geom_text( data = arrows_and_text,
    aes( x = xstart, y = ystart,
      label = labels,
      hjust = hjust, vjust = vjust ),
    col = "black",
    size = 2 ) +
  geom_text( data = txt_avg,
    aes(x = xstart, y = ystart, label = labels),
    hjust = 0, # text anchored to left
    vjust = 0.5,
    col = "black", size = 2
  ) +
  geom_segment( data = txt_avg,

```

```

aes(x = xstart, y = ystart - 0.4, xend = xend, yend = yend),
arrow = arrow(length = unit(0.08, "inch")),
color = "black",
size = 0.3

) +

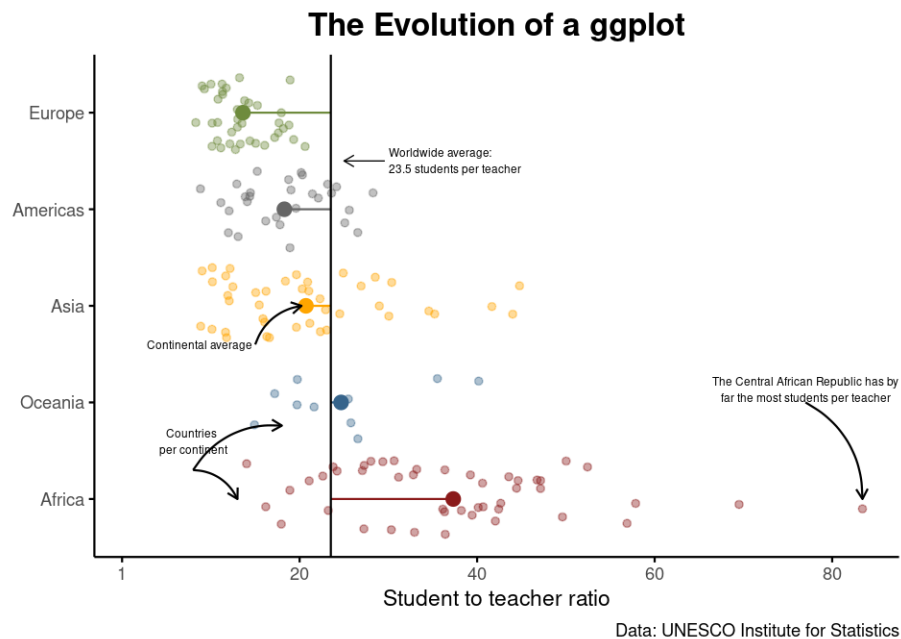
geom_curve( data = arrows_and_text,
aes(x = xstart, y = ystart,
xend = xend, yend = yend),
arrow = arrow(length = unit(0.08, "inch")),
size = 0.5, color = "black", curvature = -0.3 )

```

```

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```



Choose *ONE* of the following strategies to annotate your plot:

4.3.1 Option A: `geom_text`, `geom_curve`

```

#
# Option A: geom_text, geom_curve
#
# You'll need to add this to your plot and
# add the label for the worldwide average
#

## Code to create the details for our annotations

txt_nent <- data.frame( xstart = 2.6, ystart = 15,
xend = 3 -.1, yend = asia_avg - .5,

```

```

        labels = "Continental average ",
        hjust = 1, vjust = .5 )

txt_tries <- data.frame( xstart = 1.3, ystart = 8,
                        xend = 1, yend = 13,
                        labels = "Countries \n per continent \n",
                        hjust = .5, vjust = 0 )
txt_tries2 <- txt_tries %>%
  mutate( xend = 1.76, yend = 18, labels = "" )

txt_max <- data.frame( xstart = 1.3, ystart = 77,
                      xend = 1, yend = africa_max - 1,
                      labels = "The Central African Republic has by far\nthe most students per teacher",
                      hjust = .5, vjust = 0 )

arrows_and_text <- rbind( txt_nent,
                          txt_tries, txt_tries2,
                          txt_max )

plot_for_annotations +

  geom_text( data = arrows_and_text,
            aes( x = xstart, y = ystart,
                label = labels,
                hjust = hjust, vjust = vjust ),
            col = "black",
            size = annotate_text_size ) +

  geom_curve( data = arrows_and_text,
             aes( x = xstart, y = ystart,
                 xend = xend, yend = yend ),
             arrow = arrow(length = unit(0.08, "inch")),
             size = 0.5, color = "gray20", curvature = -0.3 )

```

4.3.2 Option B: geom_text_repel

Note: there is some randomness in this function (just like `geom_jitter`), so it's ok if you can't match the target plot exactly.

```

#
# Option B: geom_text_repel
#
# You'll need to add this to your plot and
#   add the label for the continental average
#

repel_pts <- tibble(
  x = c(4.5, 1.15, 1),
  y = c(world_avg, africa_max, africa_min),
  labels = c( paste("Worldwide average:\n",
                    round(world_avg,1),
                    "students per teacher"),
             "The Central African Republic has by far

```

```

    the most students per teacher",
    "Countries\nper continent" ))

plot_for_annotations +

  geom_text_repel( data = repel_pts,
    aes( x=x, y=y, label=labels ),
    col = "black", size = annotate_text_size,
    box.padding = 1.5, max.overlaps = Inf,
    nudge_y = c(15,0,0), nudge_x = c(0,.25,0),
    min.segment.length = 0,
    arrow = arrow(length = unit(0.008, "npc")),
    segment.curvature = -.3 )

```