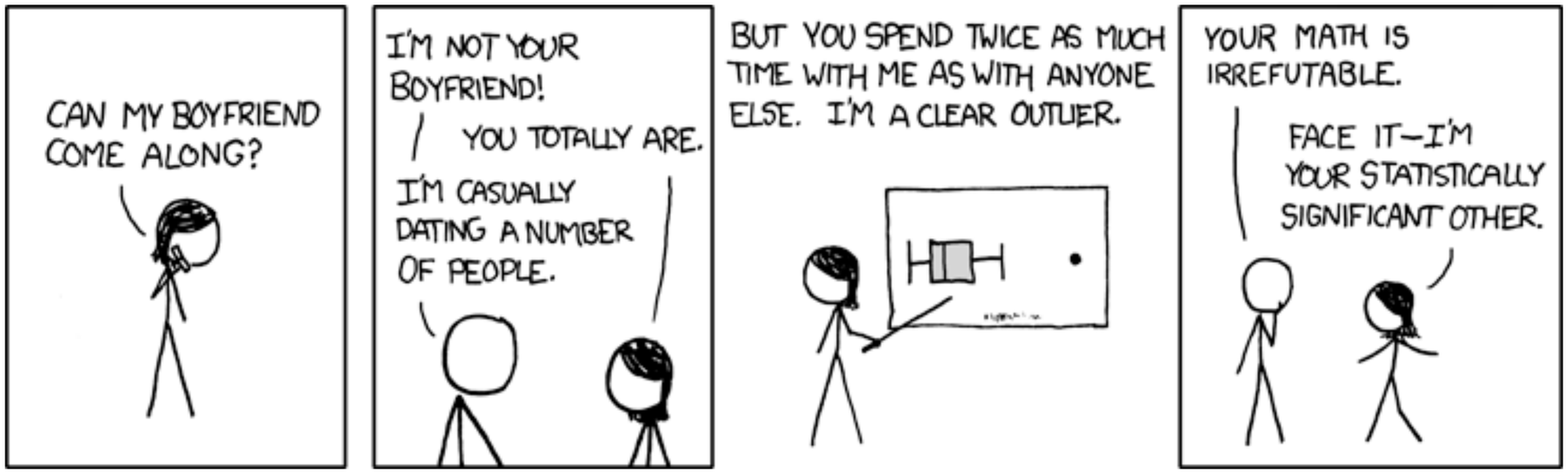
GR5702
Exploratory Data Analysis and Visualization

Prof. Joyce Robbins

Boxplot Humor



David Robinson
@drob

Following

I promise, once US isn't in a constitutional crisis with a madman dictator abusing rights I'll go straight back to tweeting base vs ggplot2

RETWEETS	LIKES
214	1,071

7:05 PM - 29 Jan 2017

18 214 1.1K

Hadley Wickham
@hadleywickham

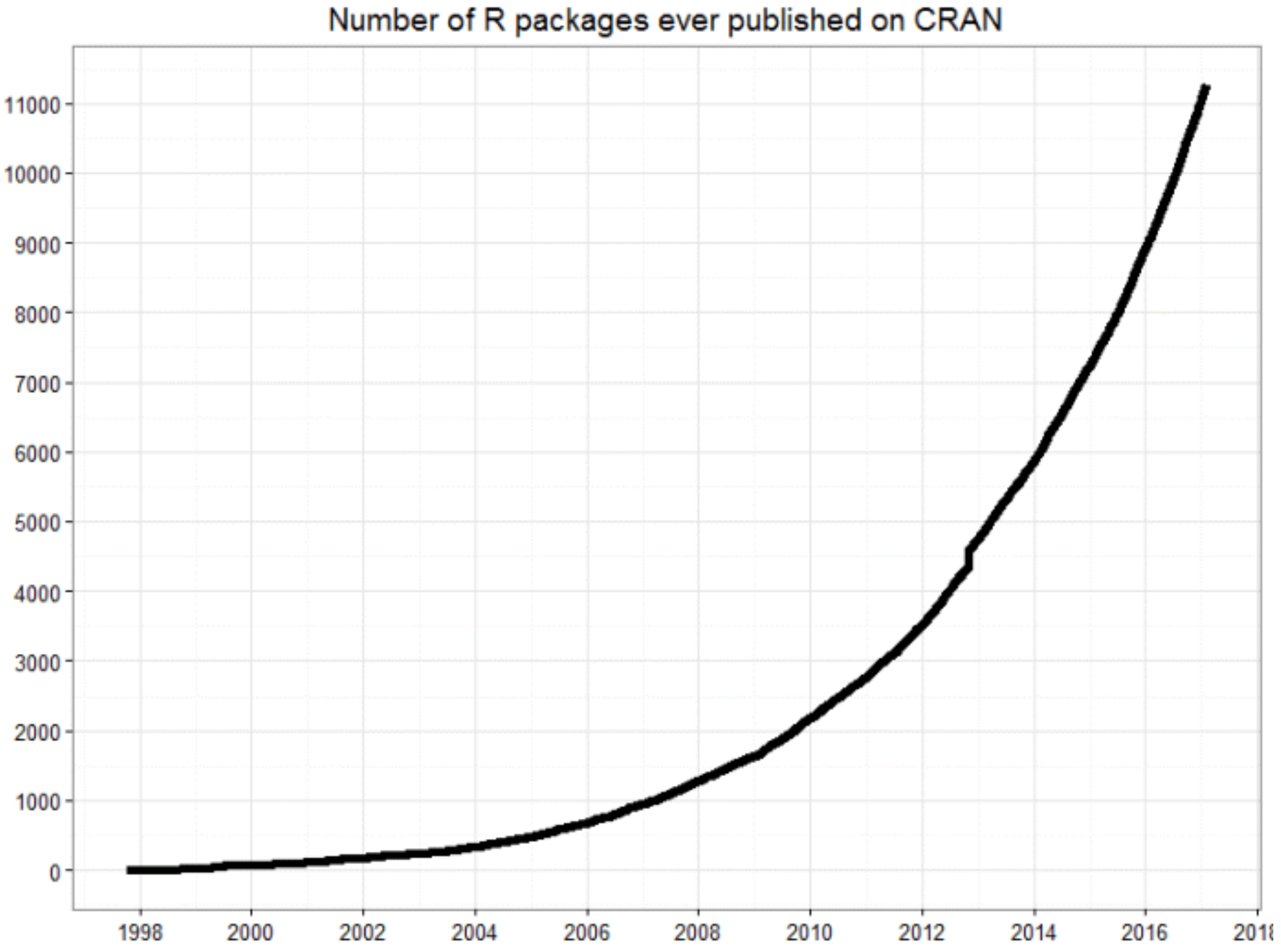
Following

I love the fact that when I don't know how to do something in ggplot2, I can google it and find out #rstats

RETWEETS 23 LIKES 18

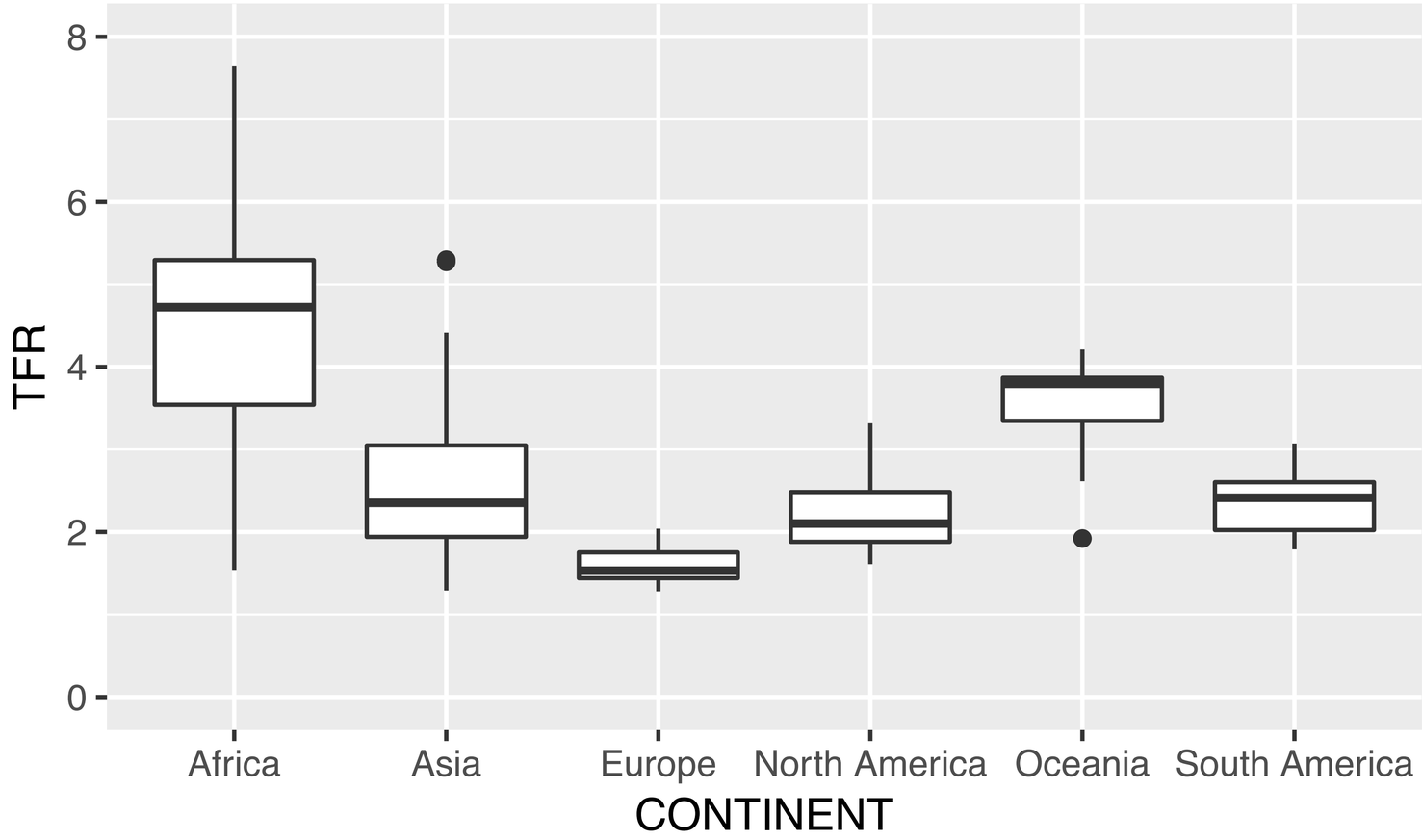
7:16 AM - 28 Mar 2013

5 23 18


Number of R packages ever published on CRAN
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016

Multiple box plots

```
g <- ggplot(world, aes(x = CONTINENT,
                       y = TFR)) + ylim(c(0, 8))
g + geom_boxplot()
```



Determine the desired order of continents

```
library(dplyr)
tfrorderdesc <- world %>% group_by(CONTINENT) %>%
    summarize(median = median(TFR), count = n()) %>%
    arrange(desc(median))
tfrorderdesc
```

```
## # A tibble: 6 × 3
##       CONTINENT median count
##          <fctr>  <dbl> <int>
## 1        Africa 4.7240    52
## 2       Oceania 3.7960     9
## 3 South America 2.4140    12
## 4          Asia 2.3530    43
## 5 North America 2.1020    21
## 6        Europe 1.5305    42
```

Reorder factor levels, calculate median

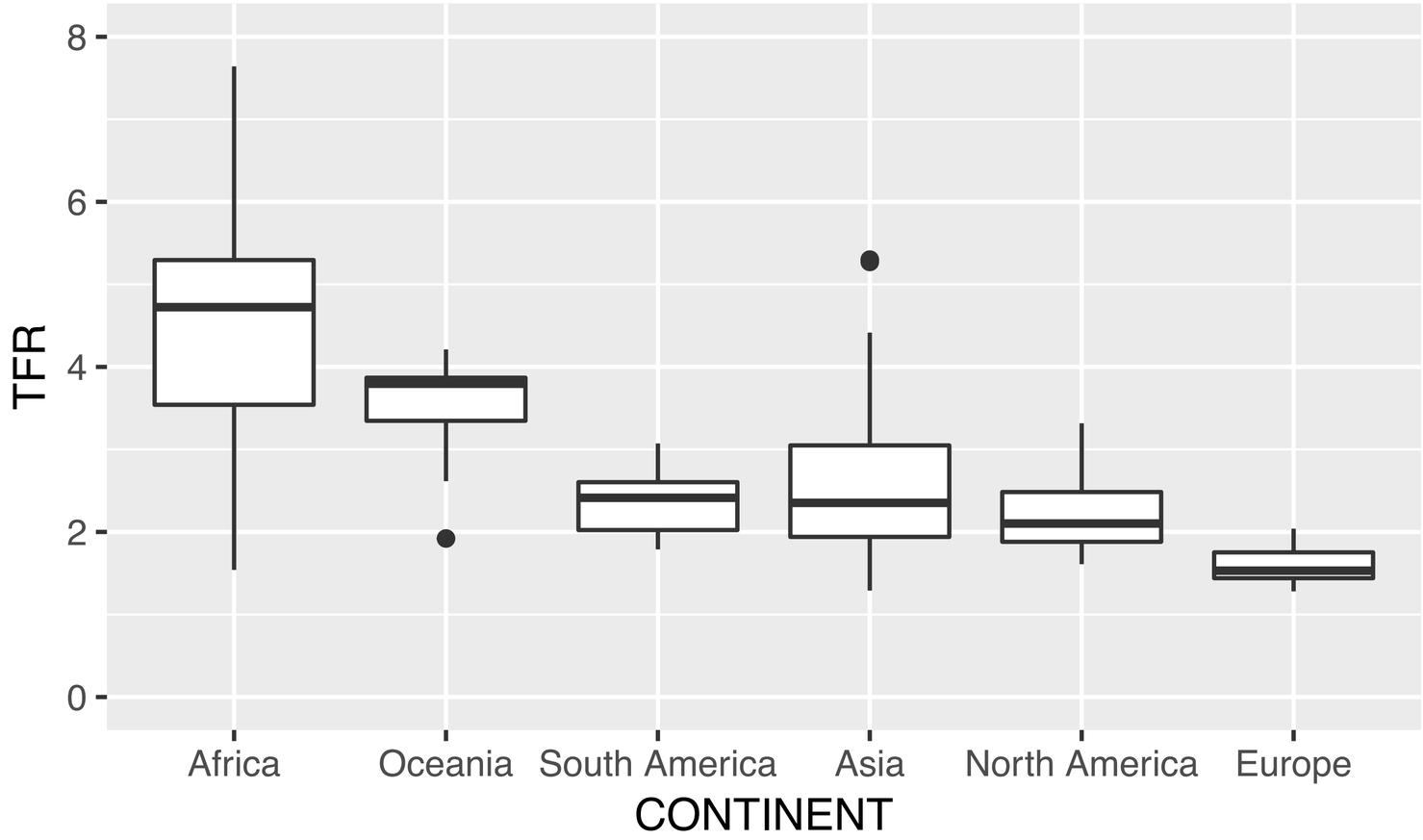
```
world$CONTINENT <- factor(world$CONTINENT,
                          levels = tfrorderdesc$CONTINENT)
cat(levels(world$CONTINENT))
```

```
## Africa Oceania South America Asia North America Europe
```

```
tfrmedian <- median(world$TFR)
```

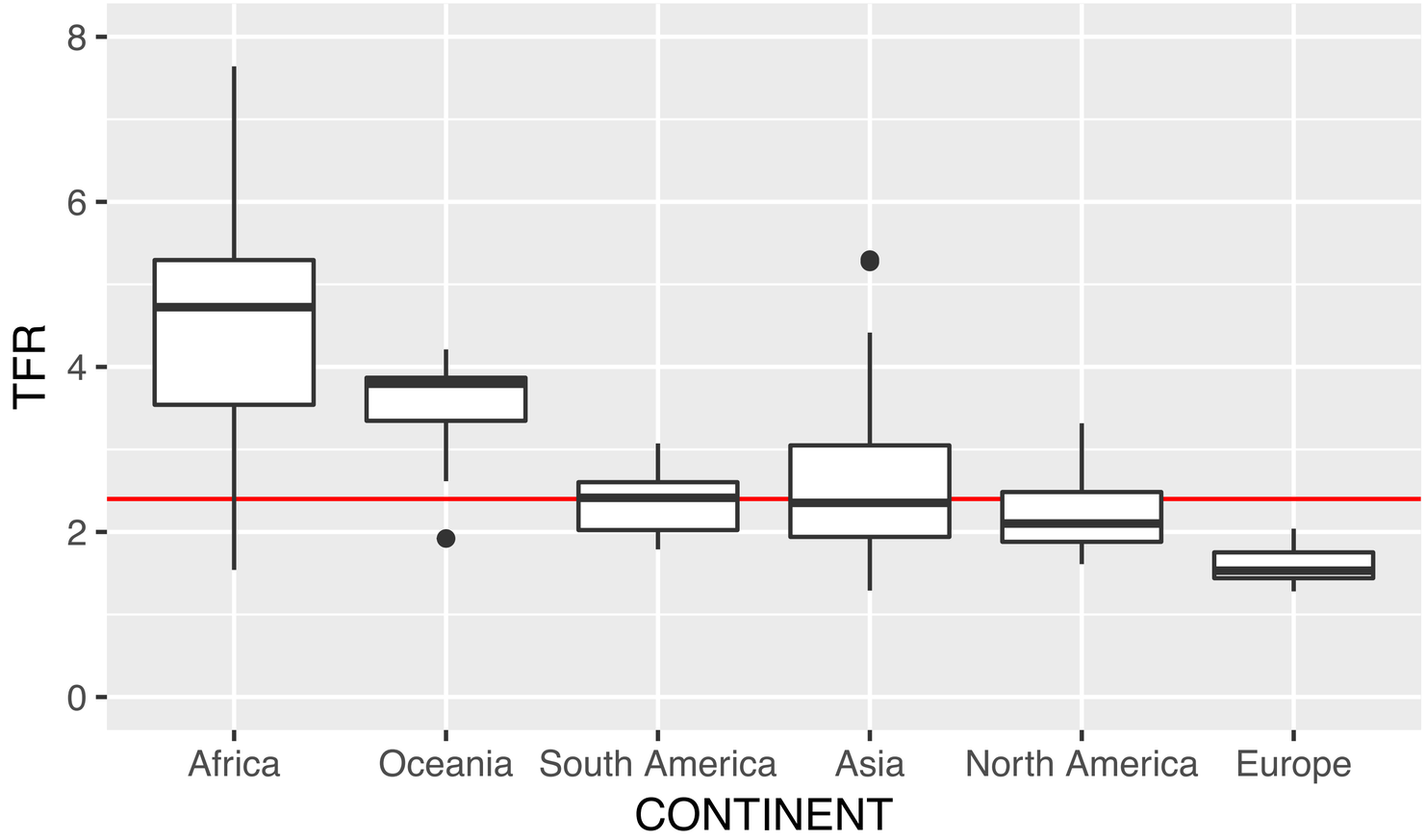
Plot again in order of descending median TFR

```
g0 <- ggplot(world, aes(x = CONTINENT,
                        y = TFR)) + ylim(c(0, 8))
g0 + geom_boxplot()
```



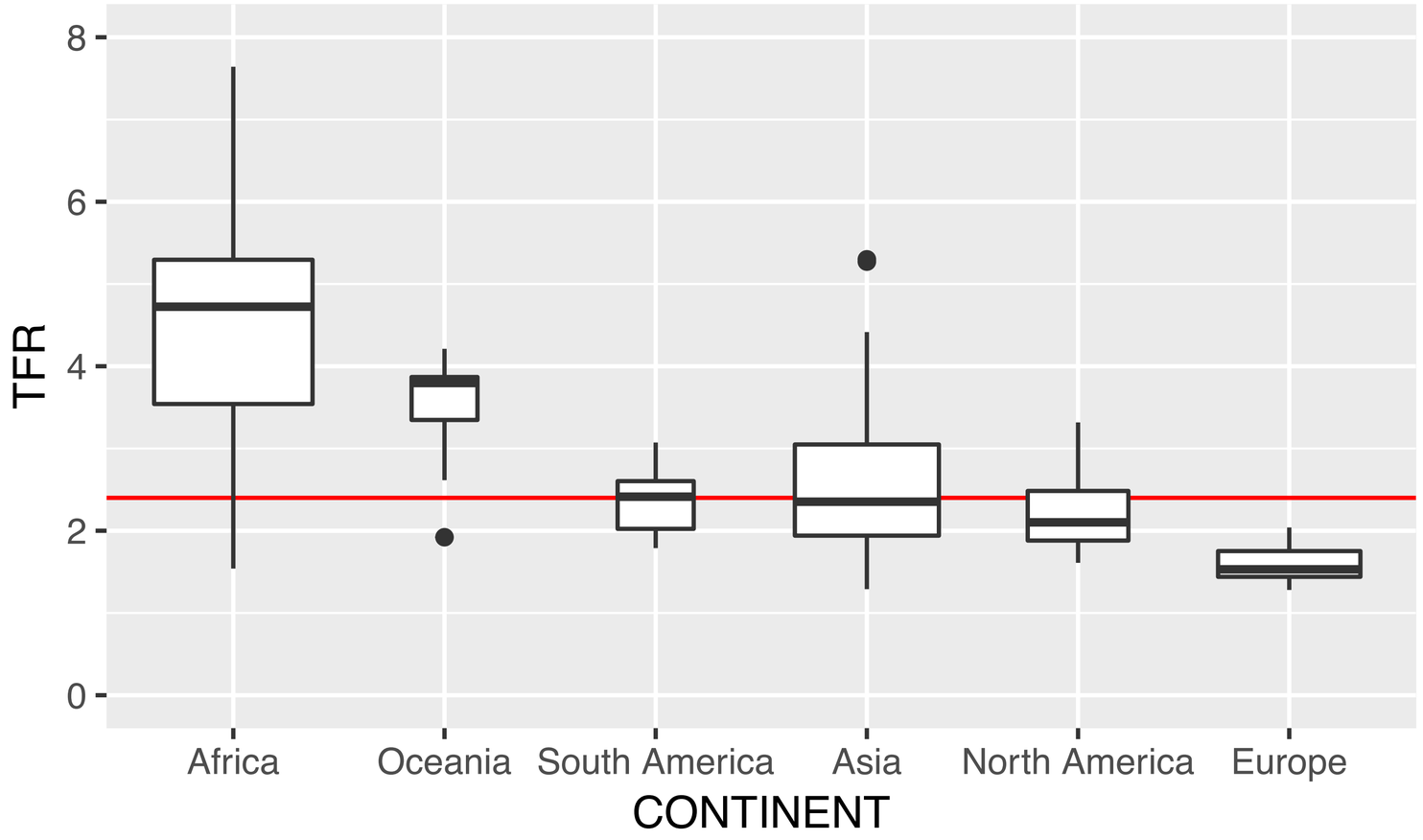
Add overall median line

```
g1 <- g0 + geom_hline(yintercept = median(world$TFR),
                      color = "red")
g1 + geom_boxplot()
```



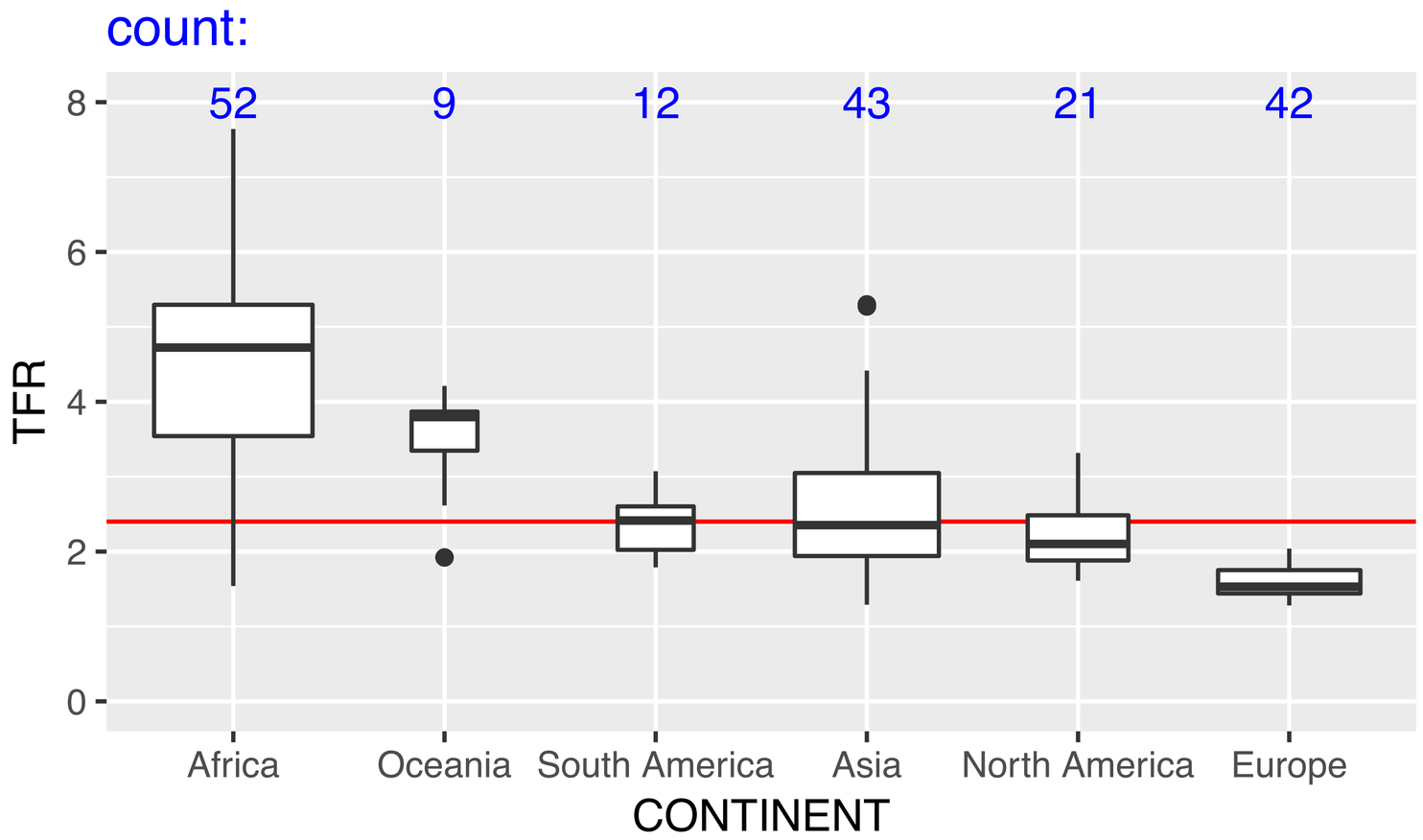
Variable width box plots

```
g2 <- g1 + geom_boxplot(varwidth = TRUE)
g2
```



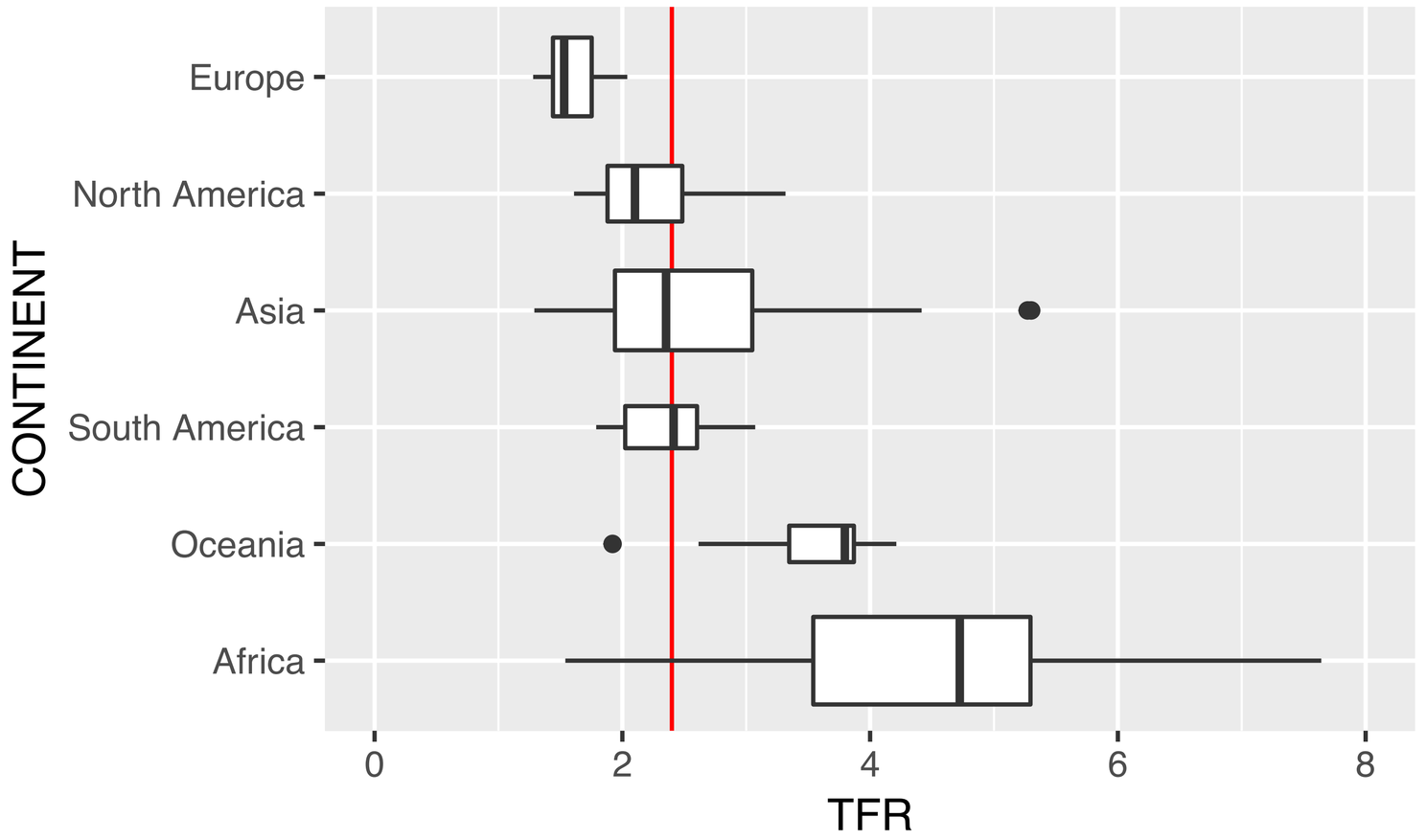
Add country count by CONTINENT

```
g2 + annotate("text", x=1:6, y = 8,
              label = tfrorderdesc$count, color = "blue") +
    ggtitle("count:") +
    theme(plot.title = element_text(color = "blue"))
```

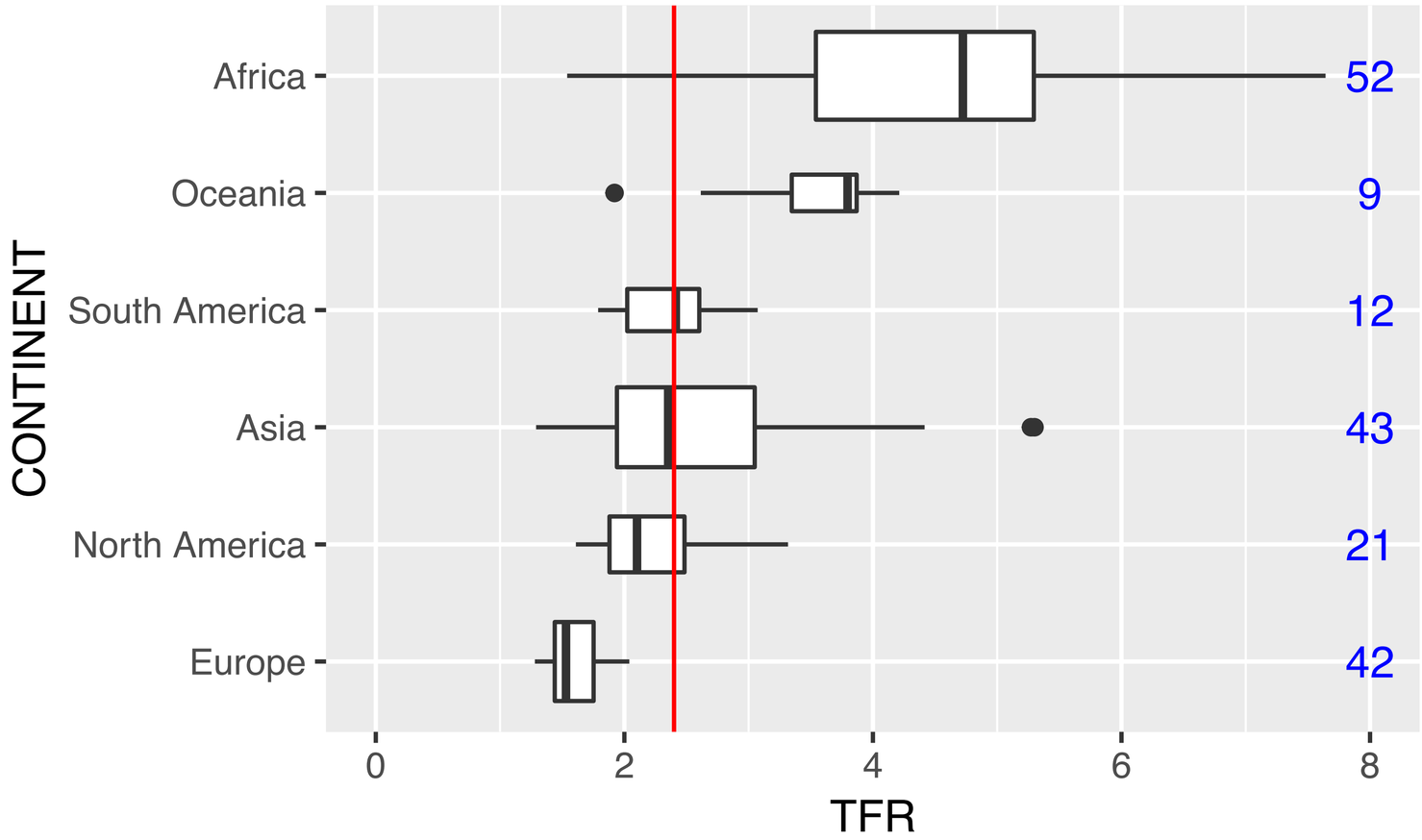


Flip the axes

```
g2 + coord_flip()
```

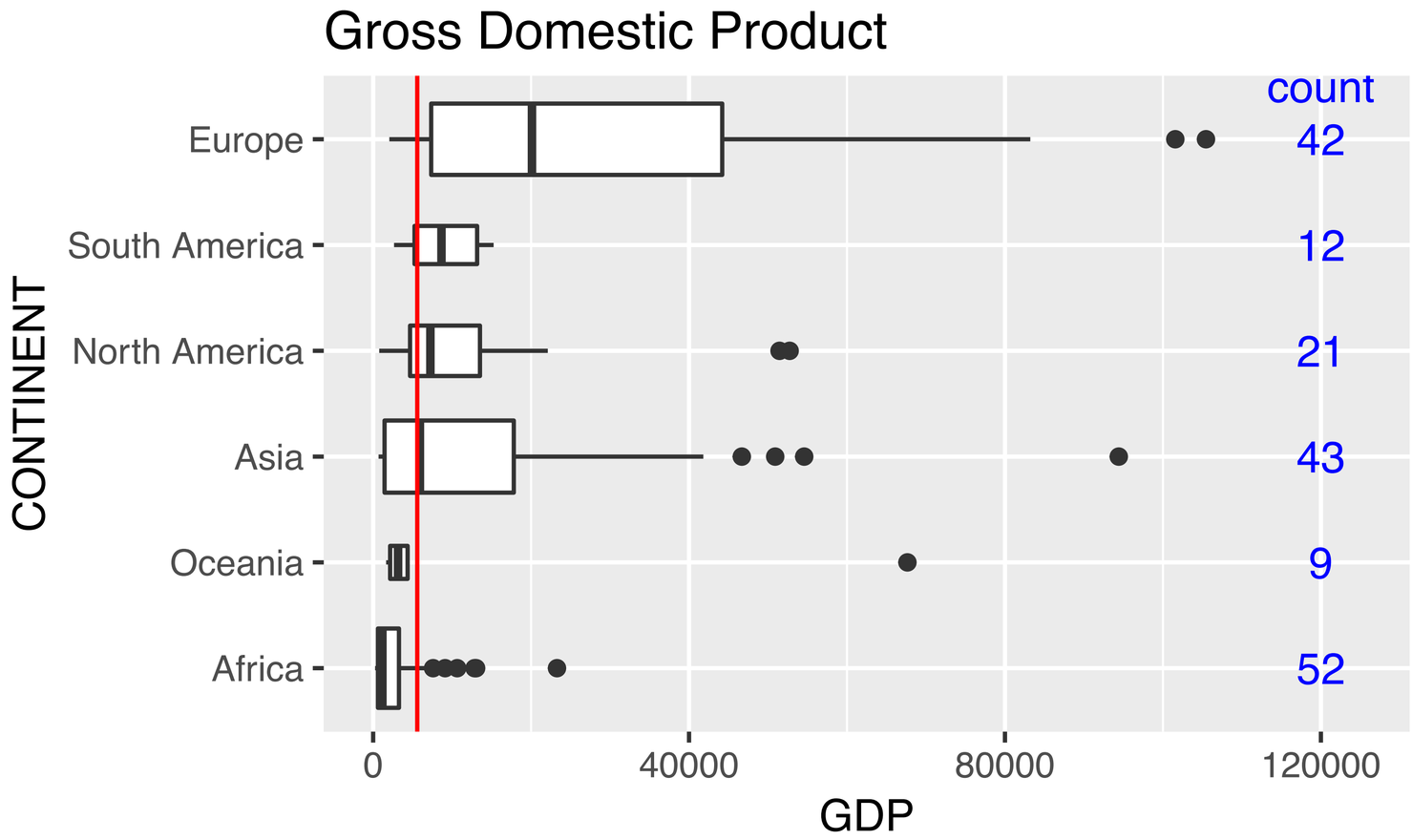


Reorder by median TFR by CONTINENT (again)

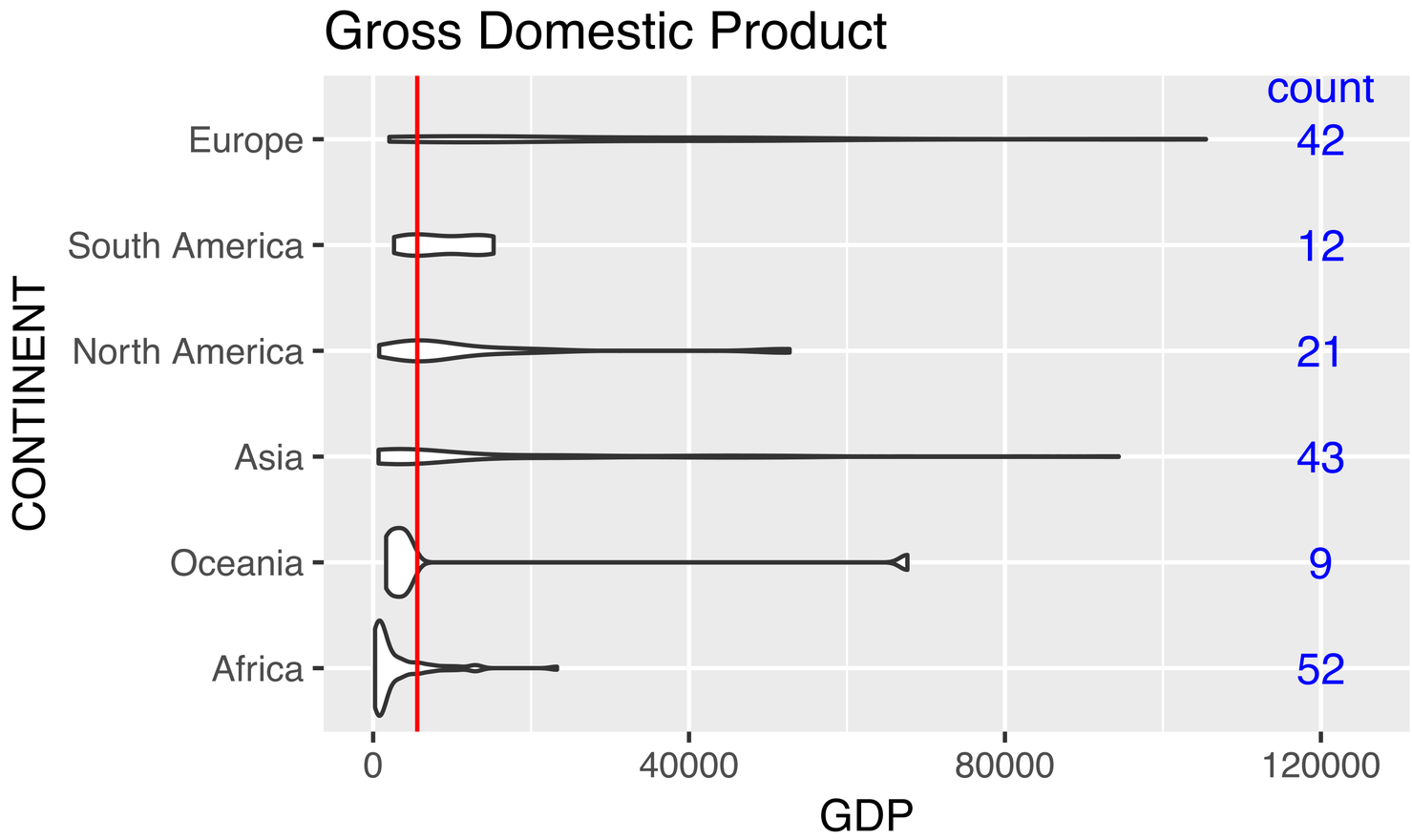


```
tfrorder <- world %>% group_by(CONTINENT) %>%
    summarize(median = median(TFR), count = n()) %>%
    arrange(median)
world$CONTINENT <- factor(world$CONTINENT,
                          levels = tfrorder$CONTINENT)
g0 + geom_boxplot(data = world, aes(x = CONTINENT,
                                    y = TFR),
                                  varwidth = TRUE) +
    geom_hline(yintercept = median(world$TFR),
               color = "red") +
    annotate("text", x=1:6, y = 8, color = "blue",
             label = tfrorder$count) +
    coord_flip()
```

Gross Domestic Product

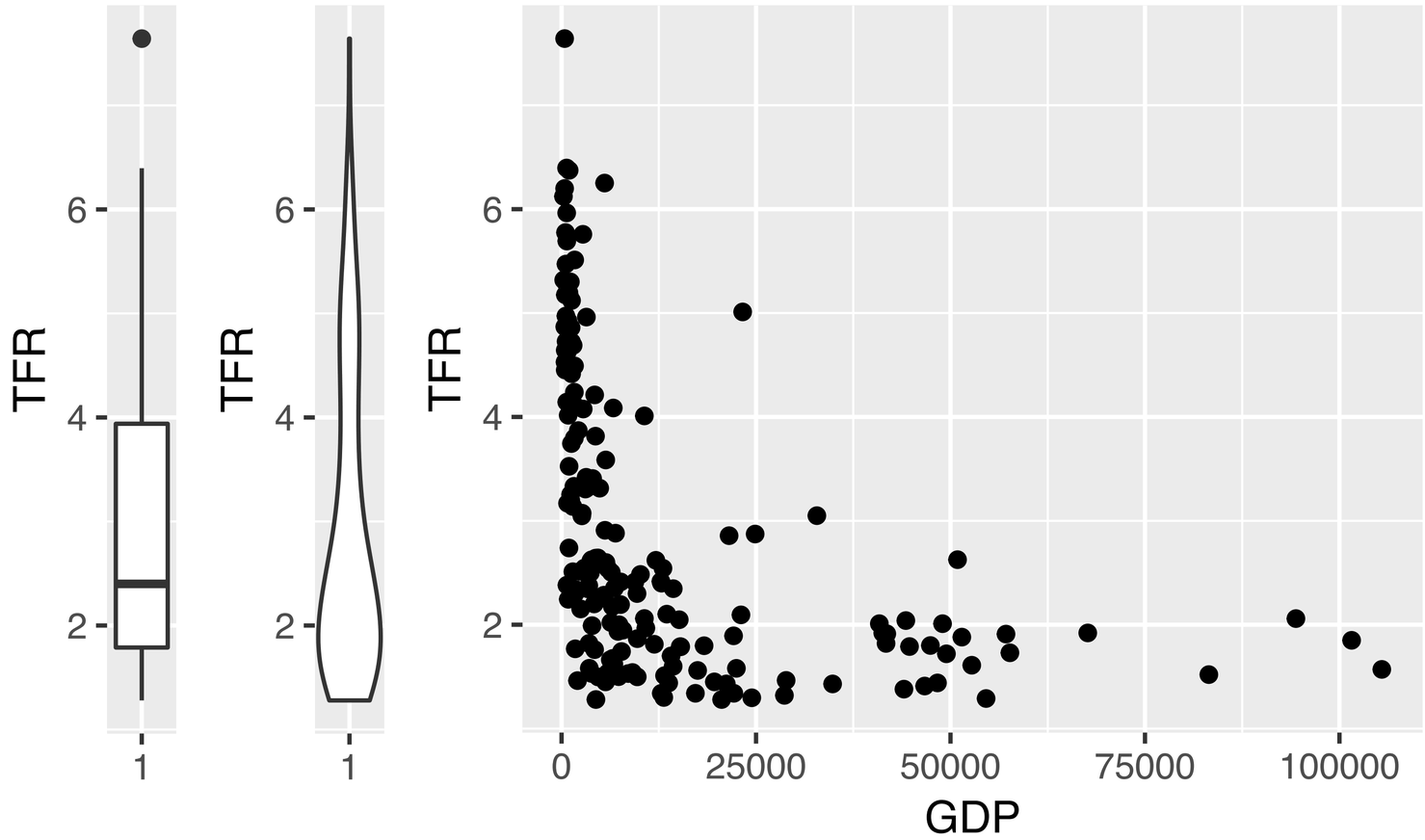


Violin plots



```r
ggplot(world, aes(x = CONTINENT, y = GDP)) +
    geom_violin() +
    ylim(c(0, 125000)) +
    geom_hline(yintercept = gdpmedian, color = "red") +
    annotate("text", x=1:6, y = 120000,
             label = gdporder$count,
             color = "blue") +
    annotate("text", x = 6.5, y = 120000, label = "count",
             color = "blue") +
    ggtitle("Gross Domestic Product") +
    coord_flip()
```

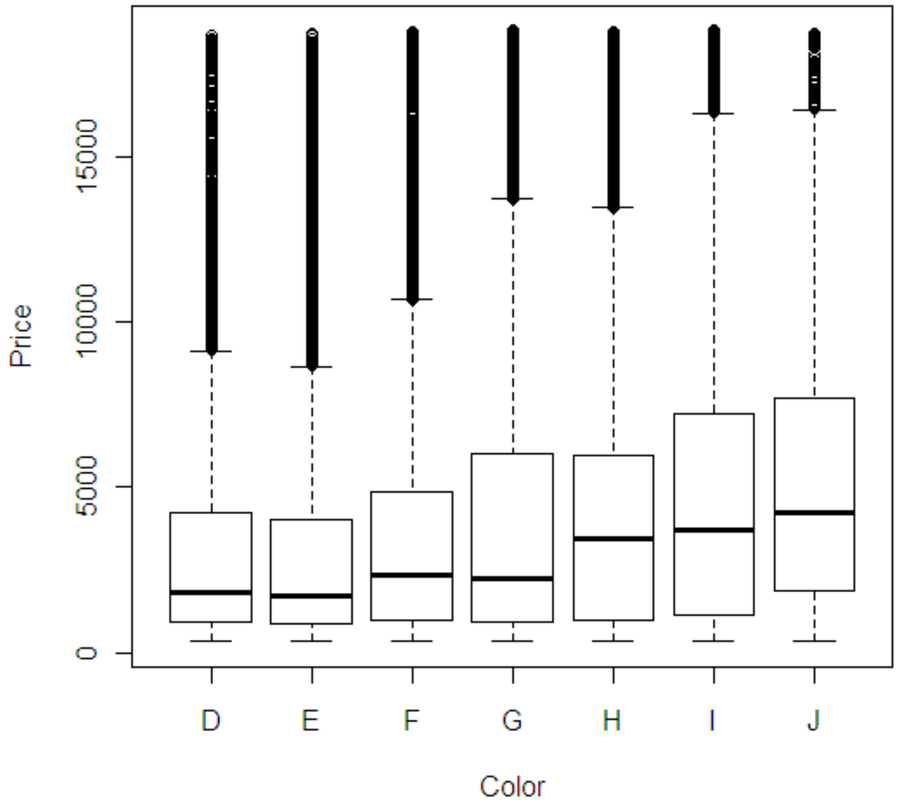
Box plot, violin plot, plus scatterplot



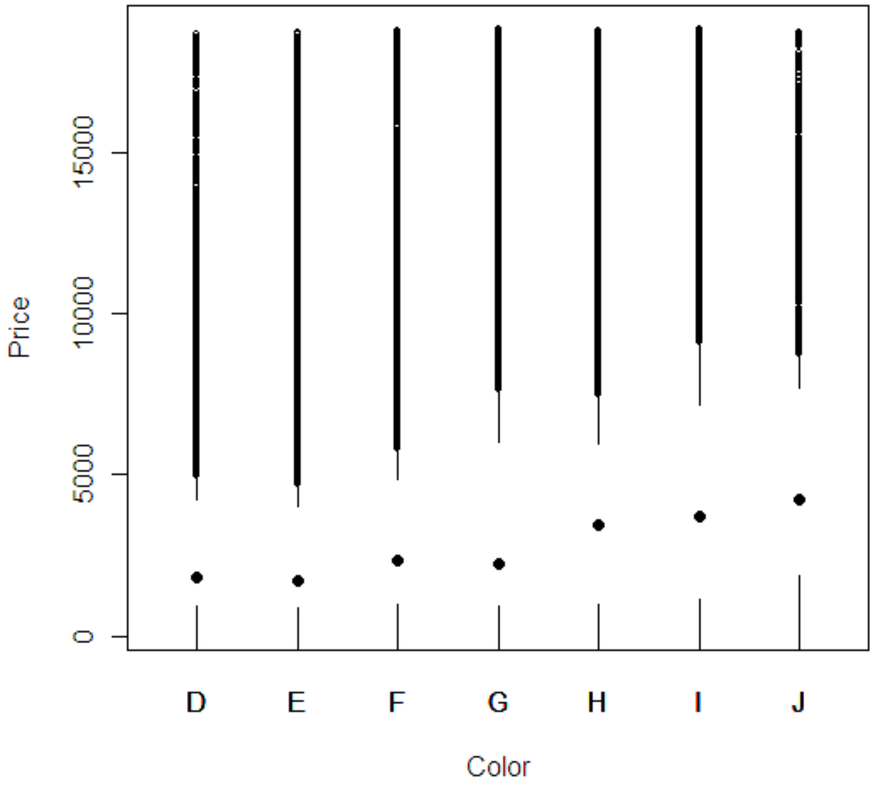
```
library(gridExtra)
g1 <- ggplot(world, aes(x = factor(1), y = TFR)) +
    geom_boxplot() + xlab("")
g2 <- ggplot(world, aes(x = factor(1), y = TFR)) +
    geom_violin() + xlab("")
g3 <- ggplot(world, aes(x = GDP, y = TFR)) + geom_point()

grid.arrange(g1, g2, g3, nrow = 1, widths = c(1, 1, 5))
```

Boxplot Alternatives

Tukey boxplots

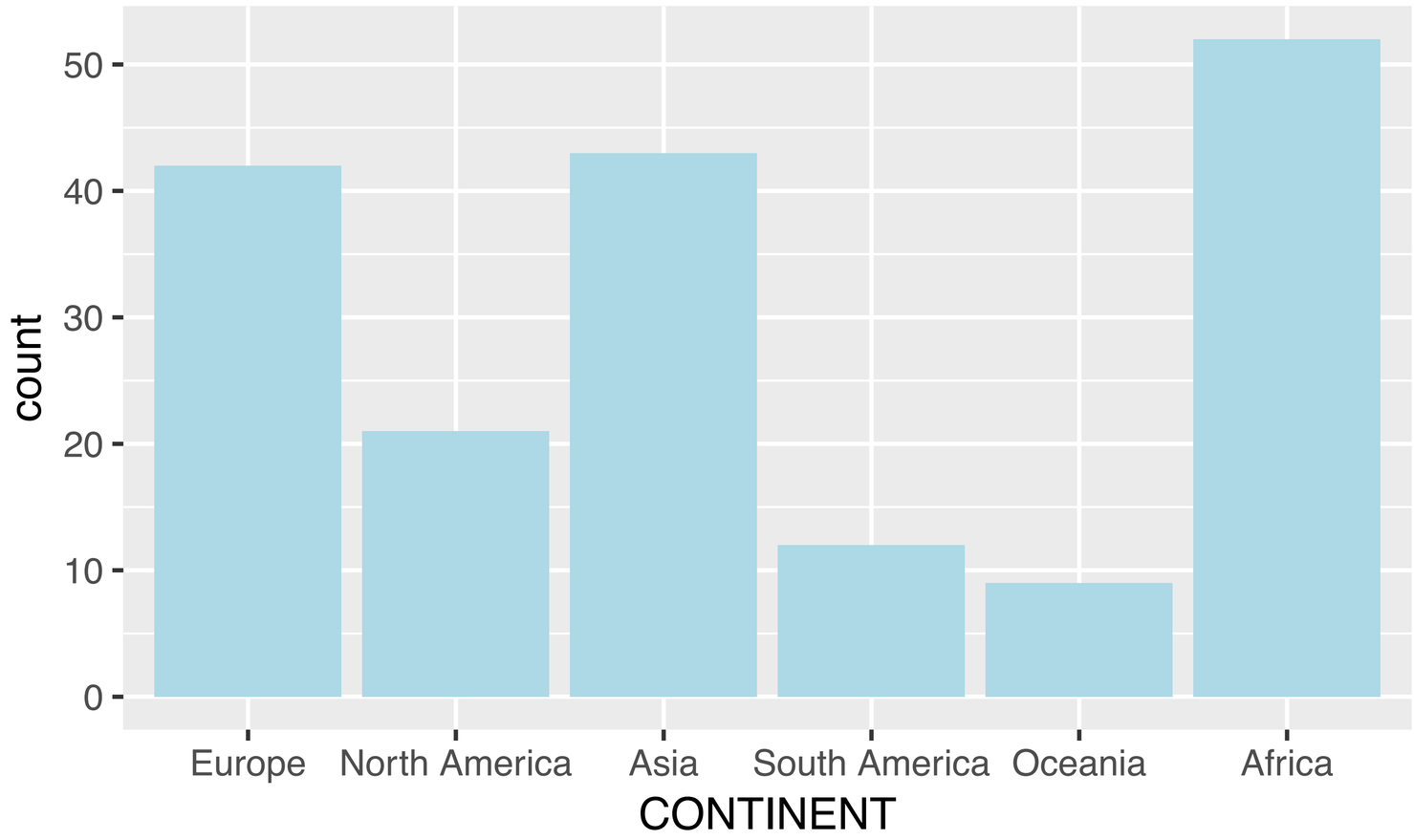


Tufte midgap plot



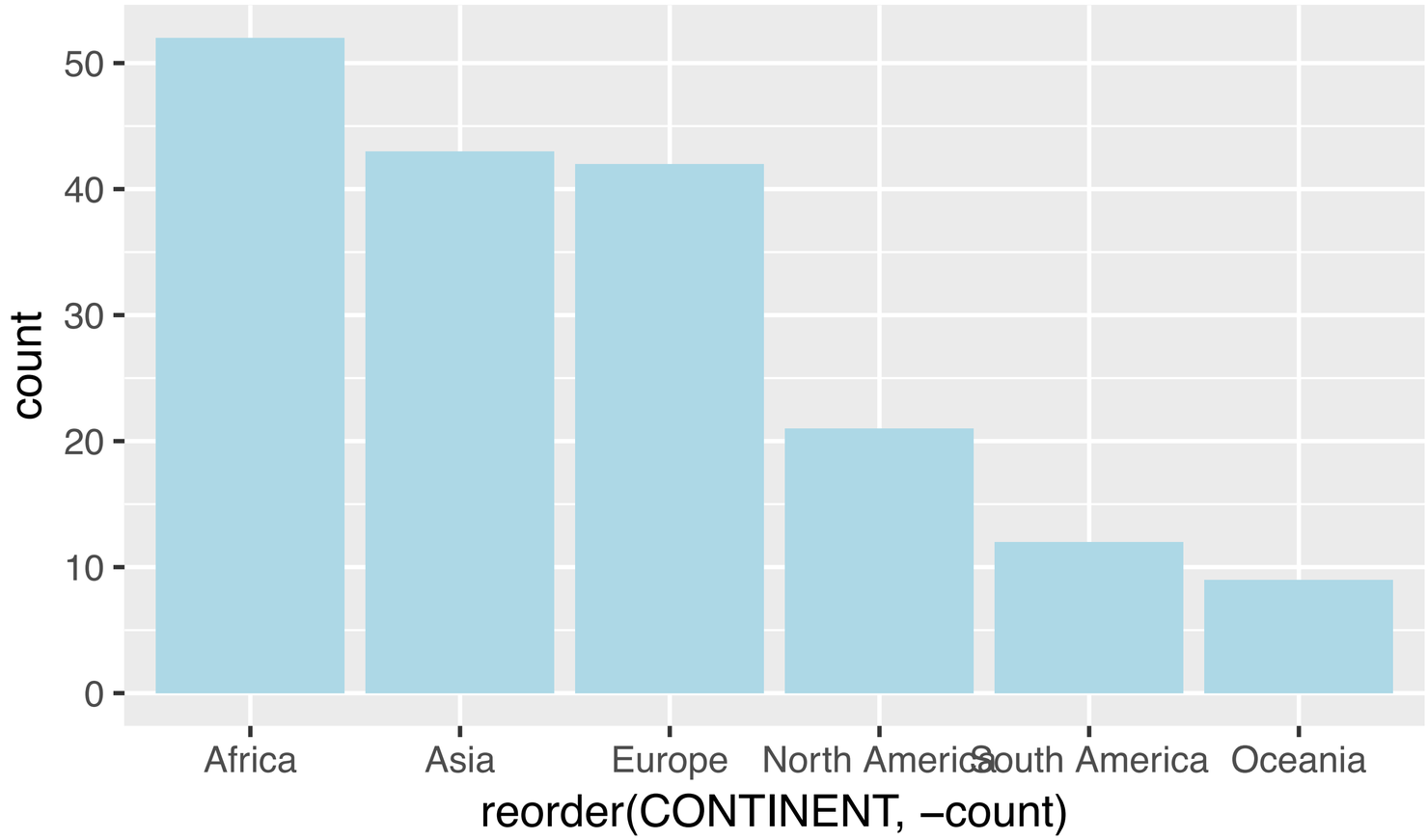
Bar charts

```
ggplot(gdporder, aes(x = CONTINENT, y = count)) +
    geom_col(fill = "lightblue")
```



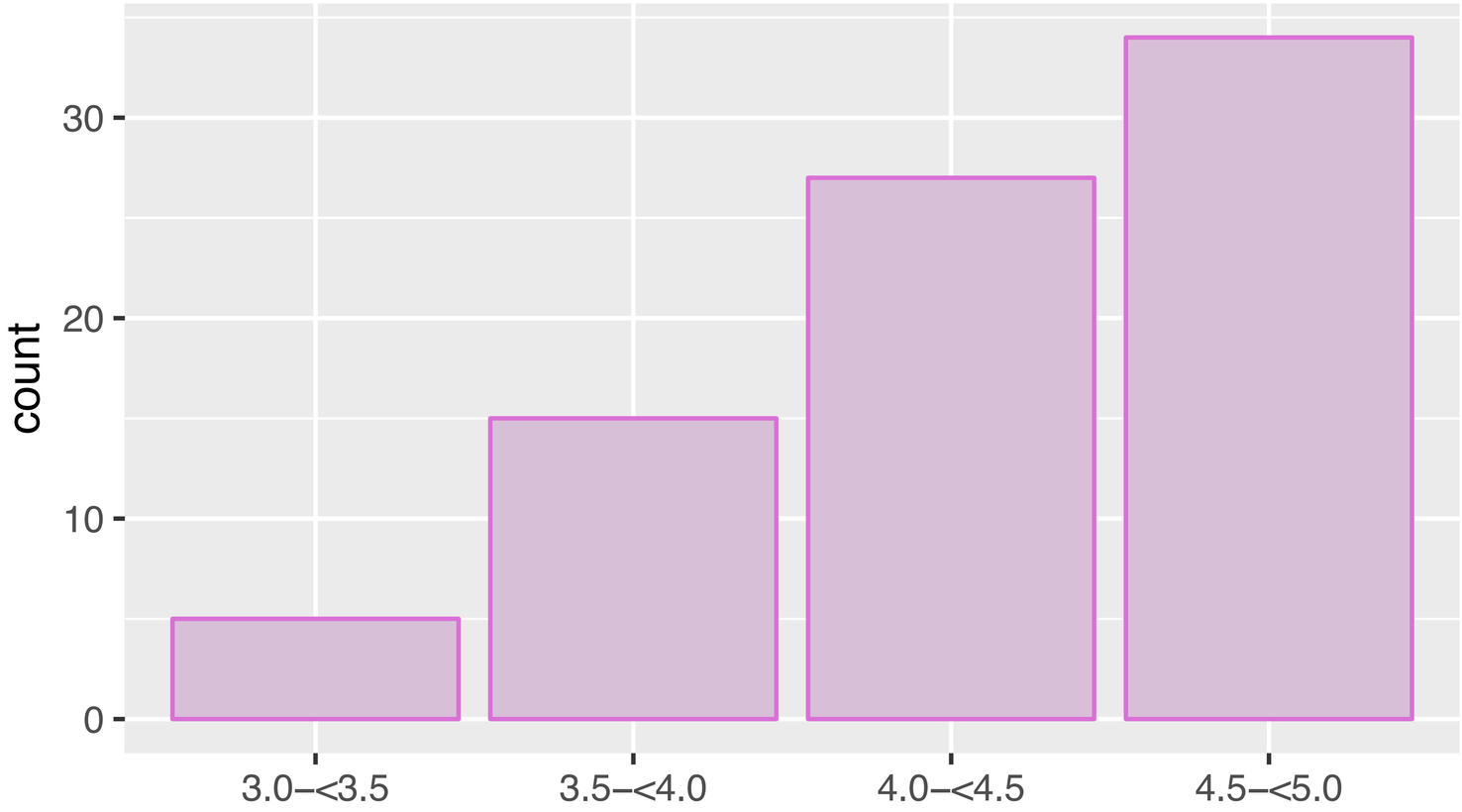
Bar chart, reordered

```
ggplot(gdporder, aes(x = reorder(CONTINENT, -count),
                     y = count)) +
    geom_col(fill = "lightblue")
```



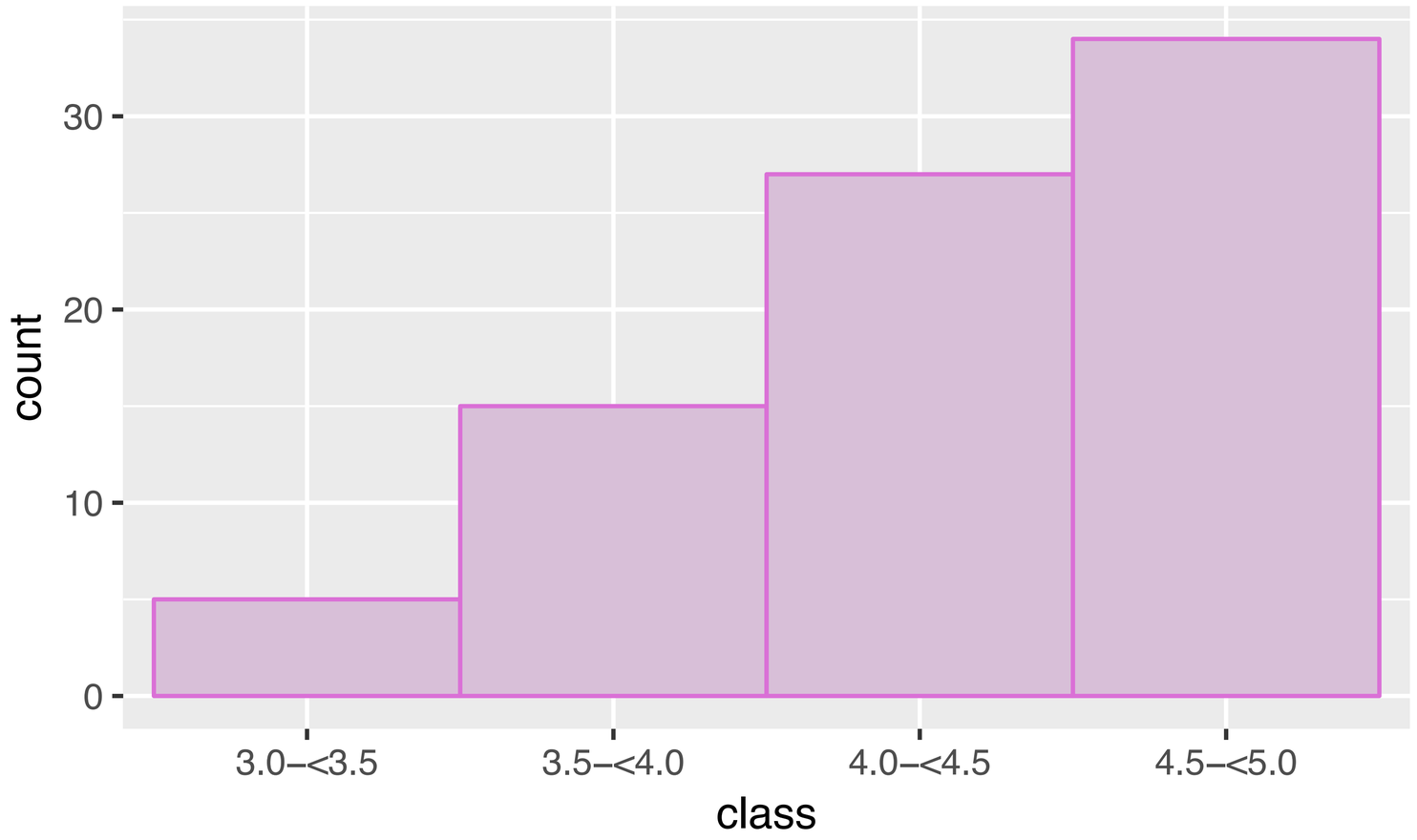
Particle counts (geom_col)

```
class <- c("3.0-<3.5", "3.5-<4.0", "4.0-<4.5", "4.5-<5.0")
count <- c(5, 15, 27, 34)
df <- data.frame(class, count)
ggplot(df, aes(class, count)) +
    geom_col(fill = "thistle", color = "orchid")
```



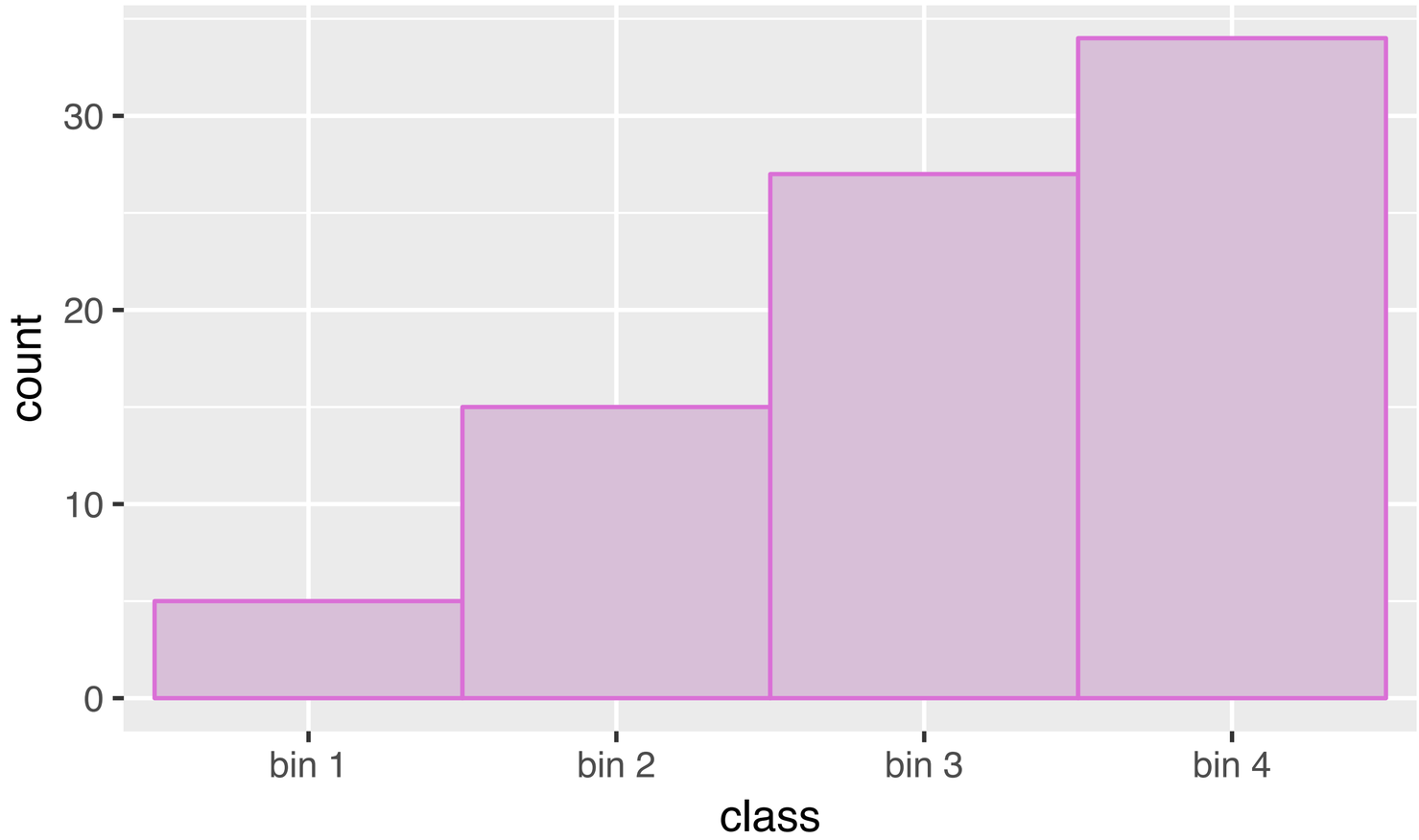
Add width = 1

```
ggplot(df, aes(class, count)) +
    geom_col(fill = "thistle", color = "orchid", width = 1)
```



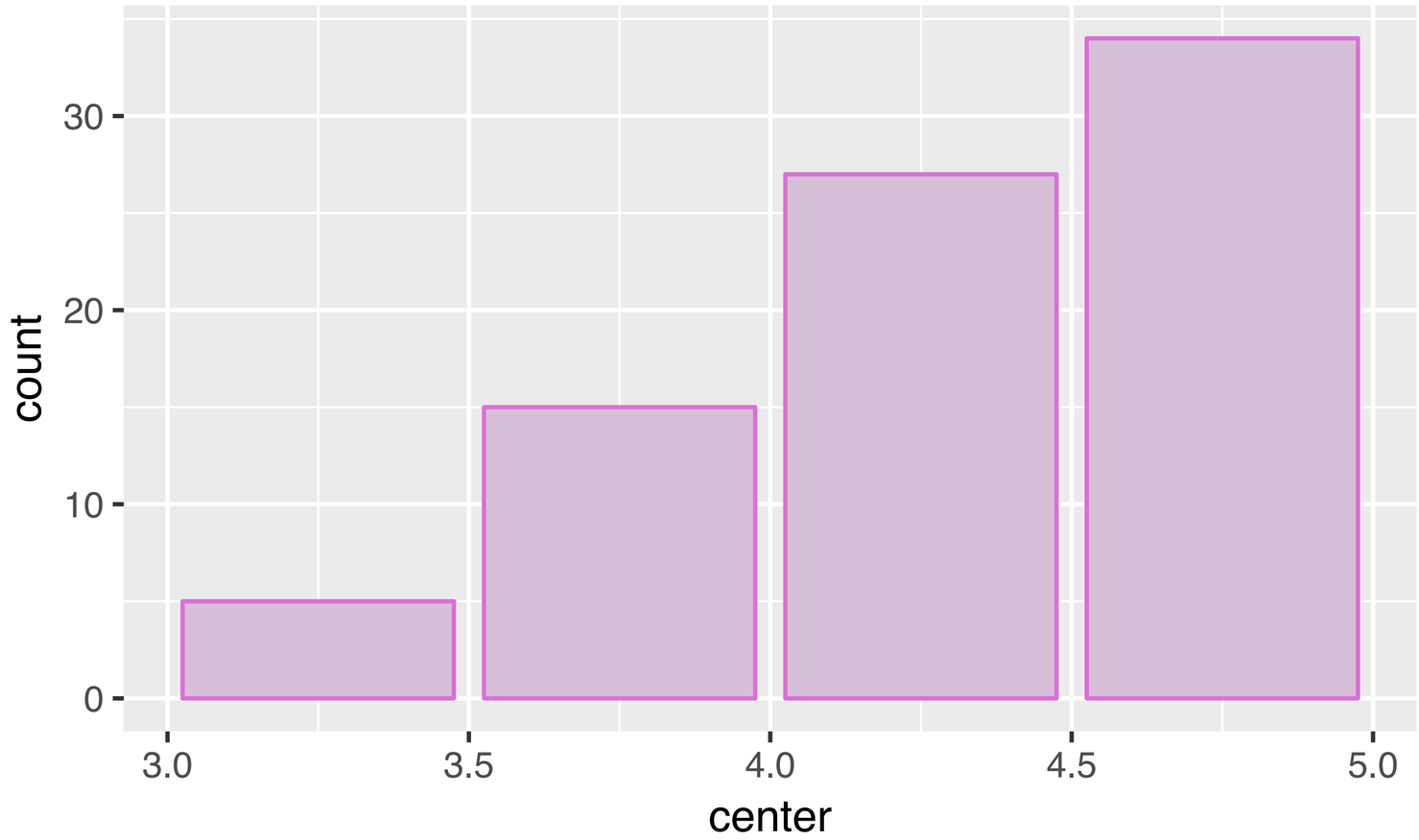
Discrete scale

```
ggplot(df, aes(class, count)) +
    geom_col(fill = "thistle", color = "orchid", width = 1) +
    scale_x_discrete(labels = c("bin 1", "bin 2", "bin 3",
                                "bin 4"))
```



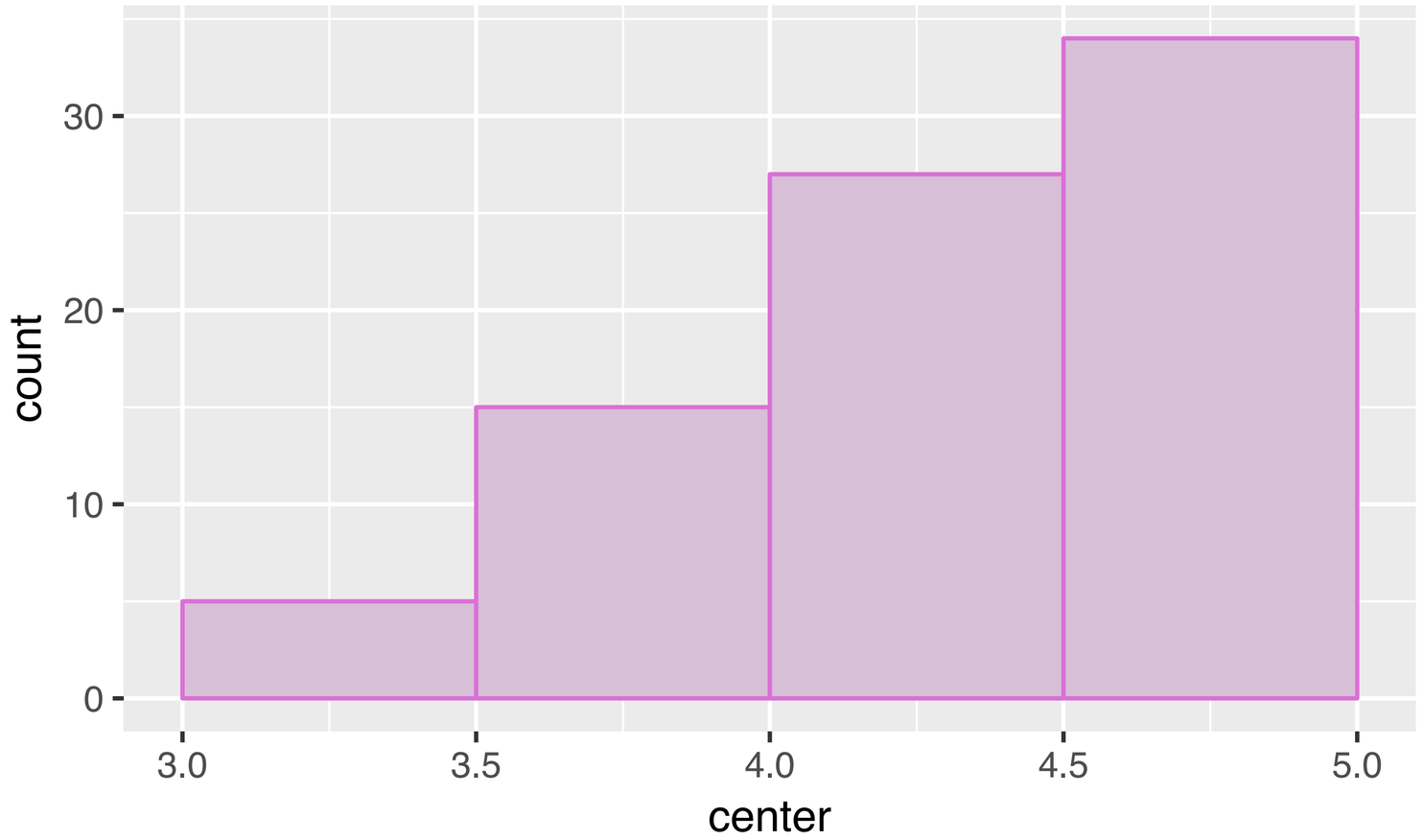
Change to continuous scale

```
# Classes: "3.0-<3.5", "3.5-<4.0", "4.0-<4.5", "4.5-<5.0"
df$center <- seq(3.25, 4.75, .5)
ggplot(df, aes(center, count)) +
    geom_col(fill = "thistle", color = "orchid")
```



Continuous scale, fix the width

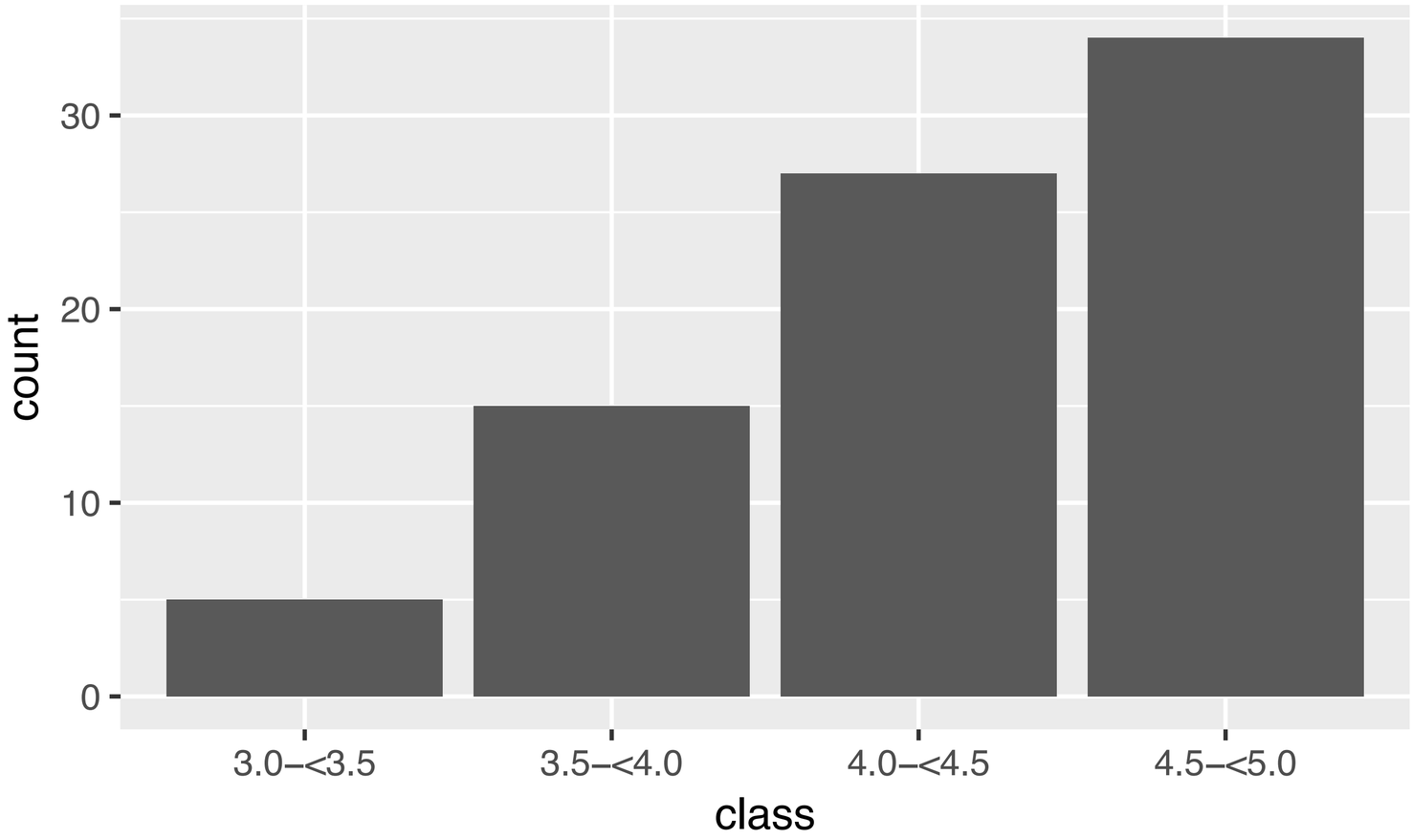
```
# Classes: "3.0-<3.5", "3.5-<4.0", "4.0-<4.5", "4.5-<5.0"
df$center <- seq(3.25, 4.75, .5)
ggplot(df, aes(center, count)) +
    geom_col(fill = "thistle", color = "orchid", width = .5)
```



Using geom_histogram

```
ggplot(df, aes(class, count)) +
    geom_histogram(stat = "identity")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Discrete distributions

```
library(vcd)
df <- data.frame(Saxony)
df
```

```
##    nMales Freq
## 1       0    3
## 2       1   24
## 3       2  104
## 4       3  286
## 5       4  670
## 6       5 1033
## 7       6 1343
## 8       7 1112
## 9       8  829
## 10      9  478
## 11     10  181
## 12     11   45
## 13     12    7
```

Plot as a bar chart (with gaps between bars)

```
ggplot(df, aes(x = nMales, y = Freq)) +
    geom_col(color = "black", fill = "thistle") +
    ggtitle("# of male children in 6115 familes of size 12")
```