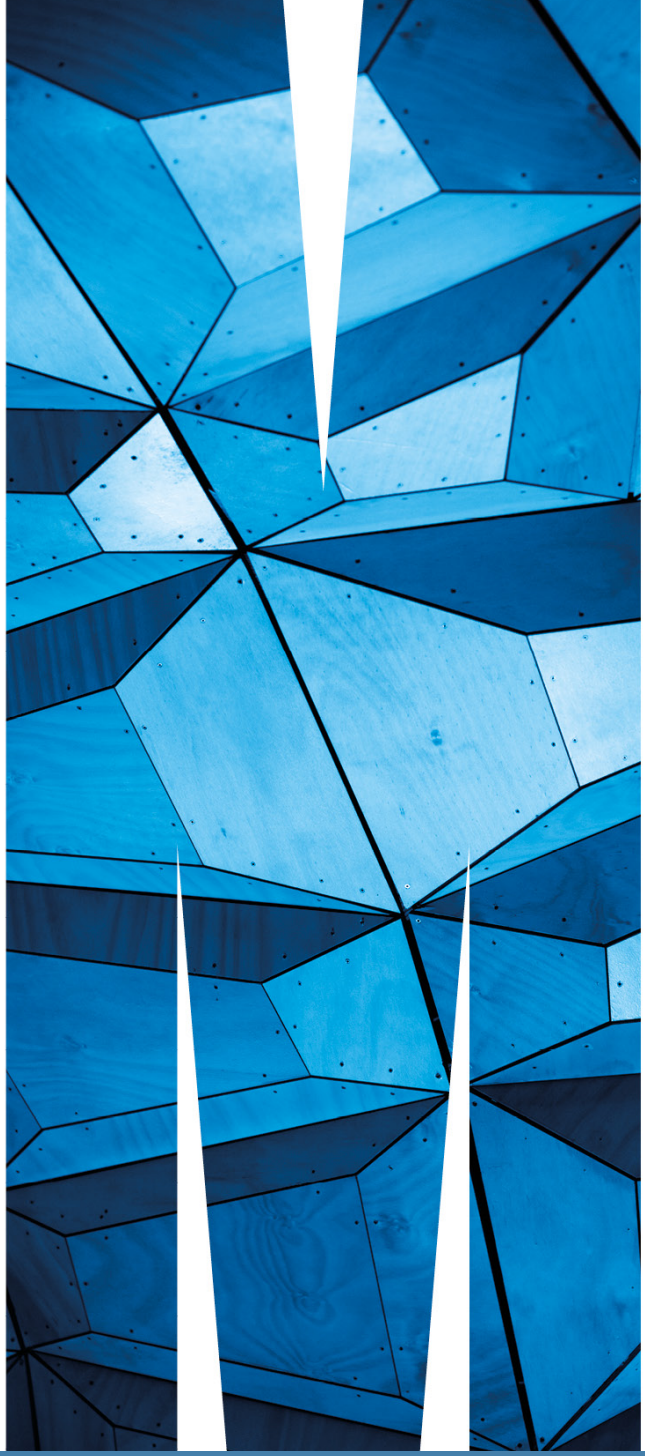
MONASH University

MONASH BUSINESS SCHOOL



Introduction to Data Analysis and Visualisation using R

Professor Di Cook, Econometrics and Business Statistics

Workshop for the Institute for Safety, Compensation and Recovery Research

GROUP OF EIGHT AUSTRALIA

AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION AMBA ACCREDITED

Session 2

Making basic plots, grammar of graphics, good practices
(If you re-started RStudio, be sure to re-open your project too.)

Using the package ggplot2

Elements of a plot

- data
- aesthetics: mapping of variables to graphical elements
- geom: type of plot structure to use
- transformations: log scale, ...

Additional components

- layers: multiple geoms, multiple data sets, annotation
- facets: show subsets in different plots
- themes: modifying style

RStudio's **cheatsheet** gives a nice, concise overview of the plotting capabilities.

Data - Currency cross rates

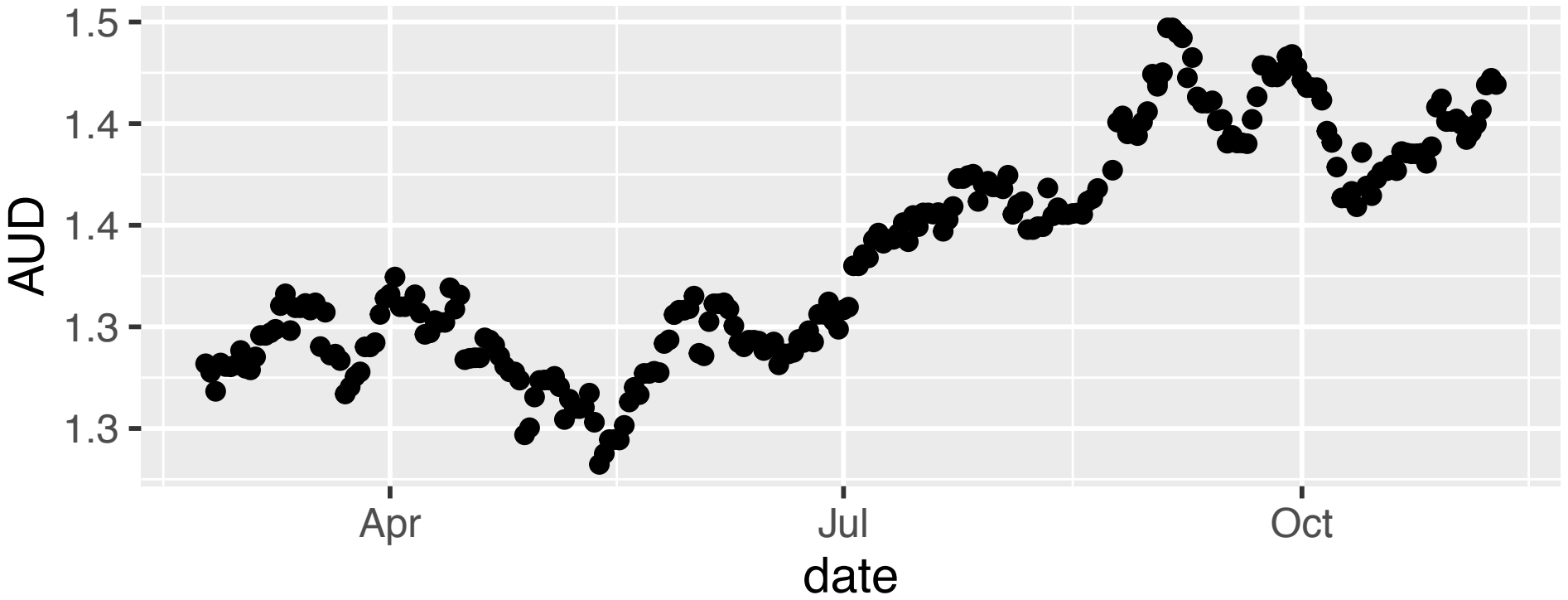
Extracted from `http://openexchangerates.org`, extracted using the json api, with the R package, jsonlite.

```
rates <- read_csv("data/rates.csv")
rates[1:5,1:8]
#> Source: local data frame [5 x 8]
#>
#>         date   AED   AFN   ALL   AMD   ANG   AOA   ARS
#>       (date) (dbl) (dbl) (dbl) (dbl) (dbl) (dbl) (dbl)
#> 1 2015-02-23   3.7    57   124   479   1.8   106   8.7
#> 2 2015-02-24   3.7    57   124   479   1.8   106   8.7
#> 3 2015-02-25   3.7    57   124   479   1.8   106   8.7
#> 4 2015-02-26   3.7    58   125   480   1.8   106   8.7
#> 5 2015-02-27   3.7    57   125   479   1.8   106   8.7
```

If you'd like to collect exchange rates yourself, see here.

Plotting points

```
ggplot(data=rates, aes(x=date, y=AUD)) + geom_point()
```



Data structure

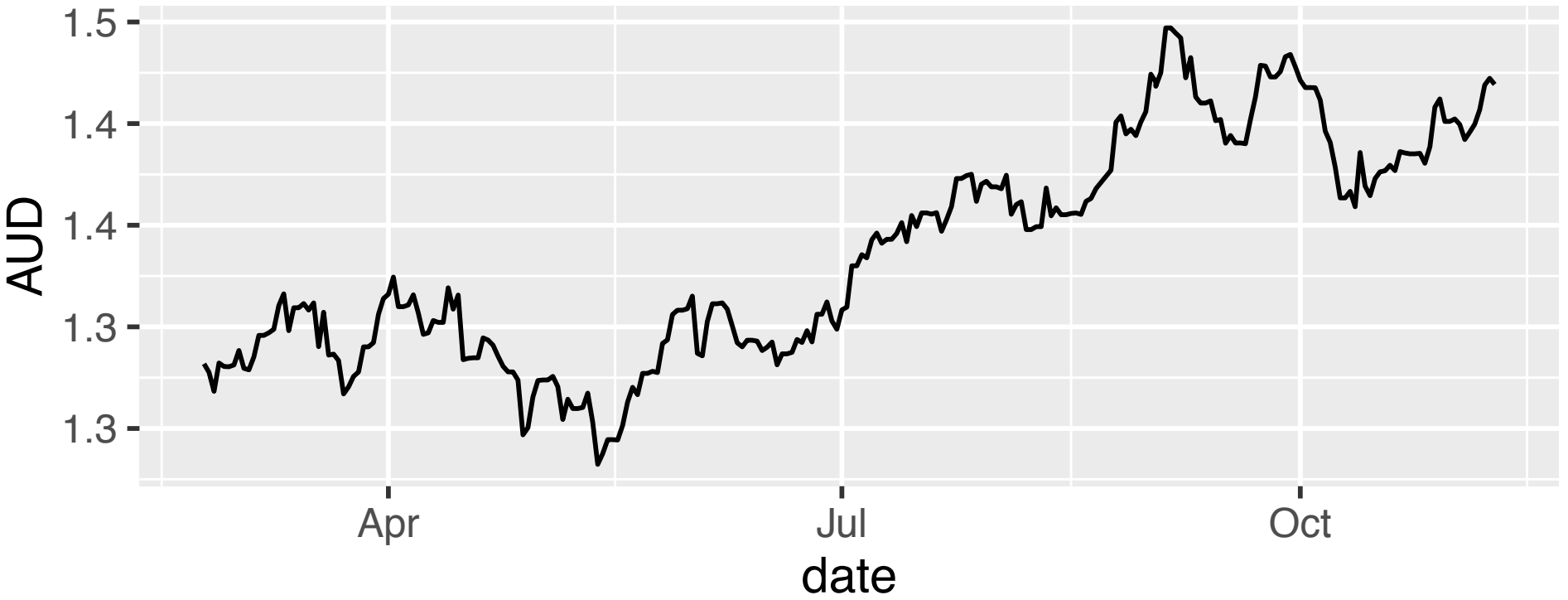
- Plots are constructed by mapping elements of data to graphical attributes.
- Having data in a tidy structure make mapping clearer
- Some ways of making mappings make it easier for the reader to perceive structure better

Plot structure

- data: rates
- aesthetics: x=date, y=AUD
- geom: point, line

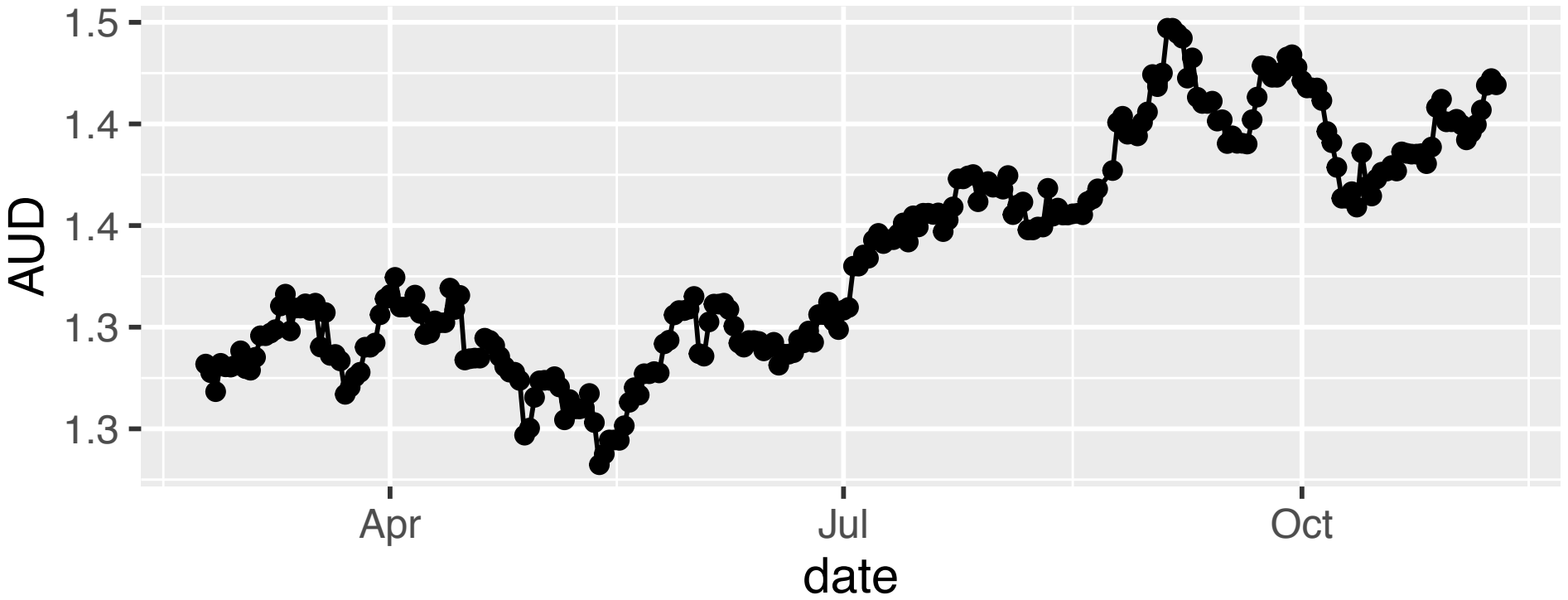
Plotting lines

```
ggplot(data=rates, aes(x=date, y=AUD)) + geom_line()
```



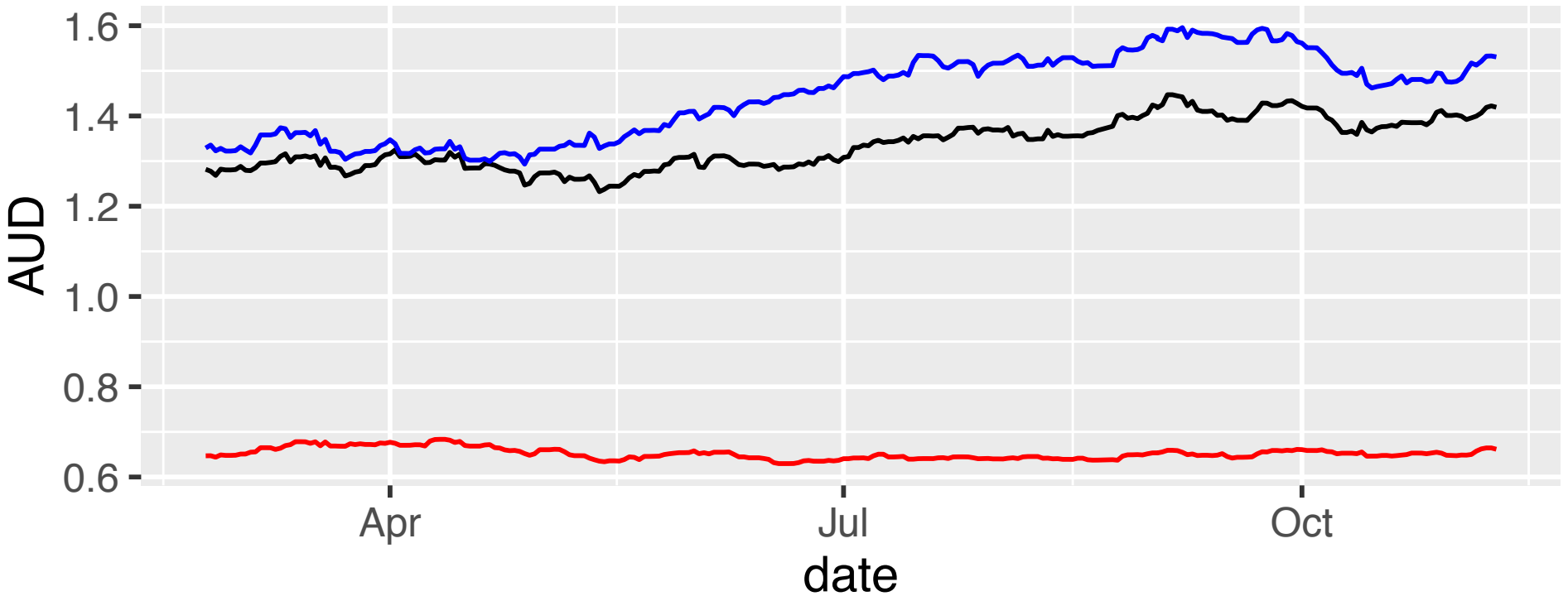
Points and lines

```
ggplot(data=rates, aes(x=date, y=AUD)) +
  geom_line() + geom_point()
```



Multiple currencies

```
ggplot(data=rates, aes(x=date, y=AUD)) + geom_line() +
  geom_line(aes(y=NZD), colour="blue") +
  geom_line(aes(y=GBP), colour="red")
```

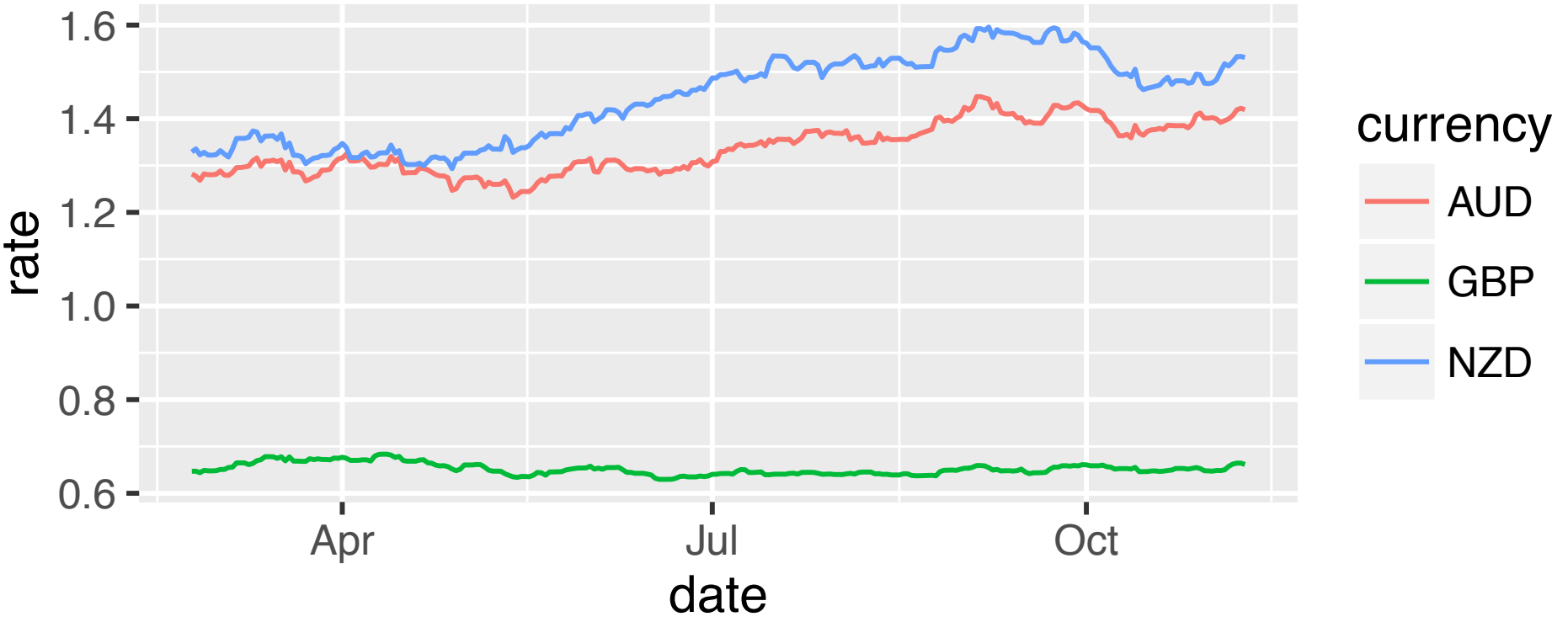


Hmmm. . .

- That code is clunky!
- Better to rearrange data, and then let ggplot2 handle the colors, legends, . . .

Better way

```
rates.sub <- select(rates, date, AUD, NZD, GBP)
rates.sub.m <- gather(rates.sub, currency, rate, -date)
ggplot(data=rates.sub.m, aes(x=date, y=rate, colour=currency))
```



GRAMMAR

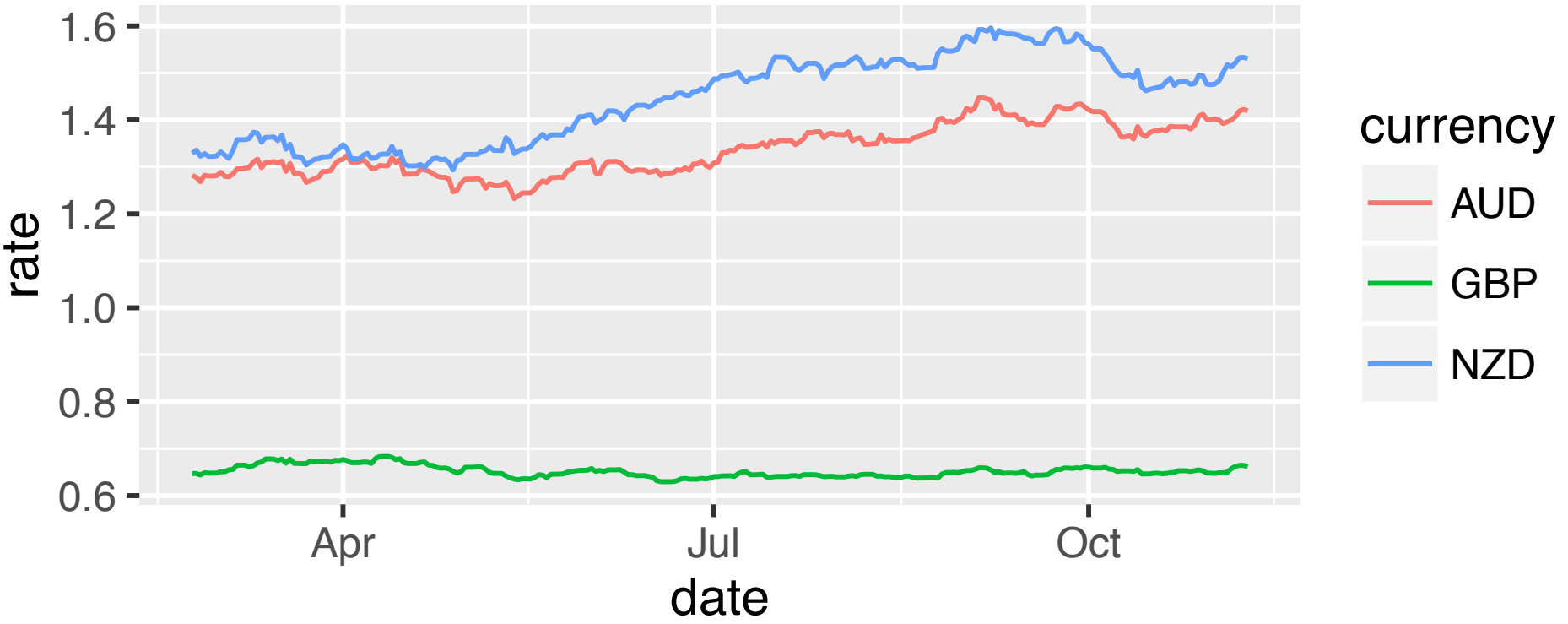
- The grammar of graphics makes the mapping of a data variable to a plot element explicit.
- This is a huge advance in data visualisation
- This provides a closer connection between data, plots and models.

Mappings

- Date is mapped to position along the x axis
- Rate is mapped to position along the y axis
- Currency is mapped to colour

Information communication

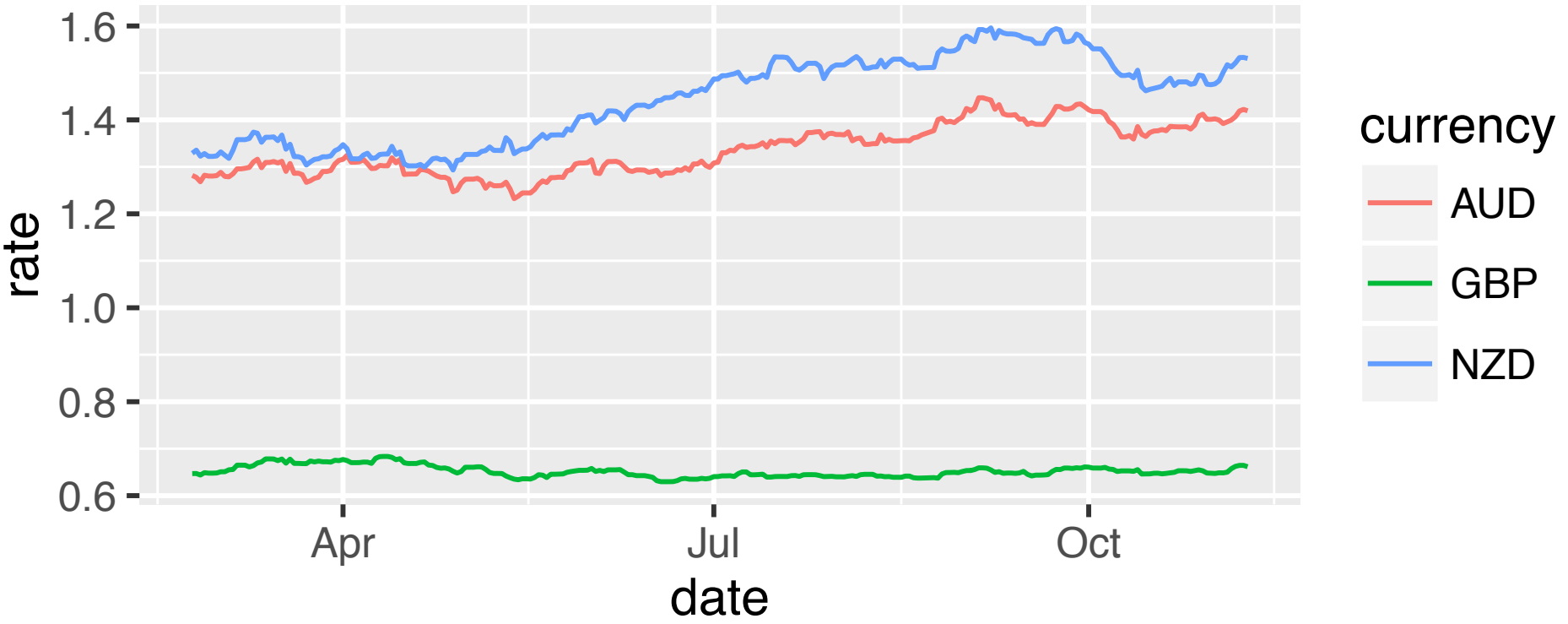
```
ggplot(data=rates.sub.m, aes(x=date, y=rate, colour=currency))
```



- What can you read from this plot? What is the main observation?
- The cross-rates for AUD and NZD with the USD are similar, $1USD can buy approximately $1.30 of both, but the GBP is lower, and $1USD only buys 2/3 of a GBP. Do we need a plot to know this?

Information communication

```
ggplot(data=rates.sub.m, aes(x=date, y=rate, colour=currency))
```

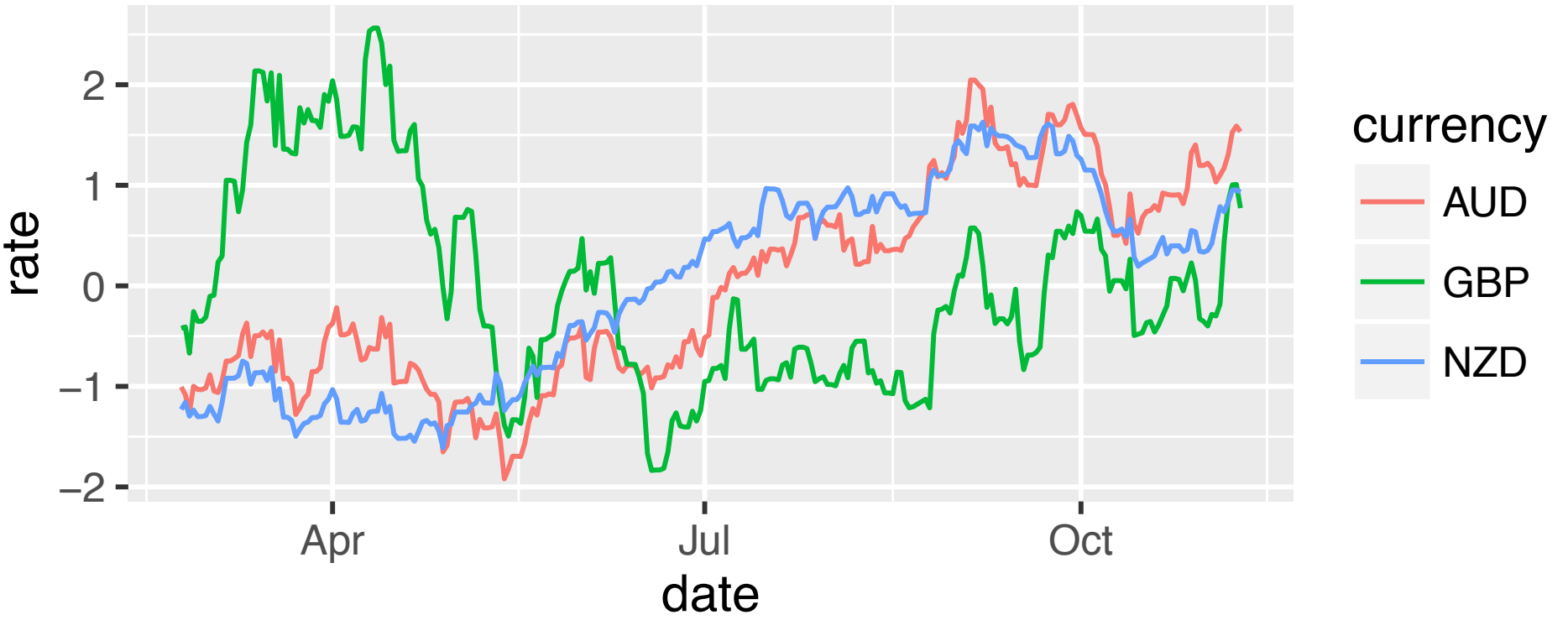


- What can you read from this plot? What is the main observation?
- The cross-rates for AUD and NZD with the USD are similar, $1USD can buy approximately $1.30 of both, but the GBP is lower, and $1USD only buys 2/3 of a GBP. Do we need a plot to know this?

Communicate trend: Scale currencies

```
rates.sub <- mutate(rates.sub, AUD=scale(AUD), NZD=scale(NZD)
rates.sub$date <- as.Date(rates.sub$date)
rates.sub.m <- gather(rates.sub, currency, rate, -date)
```

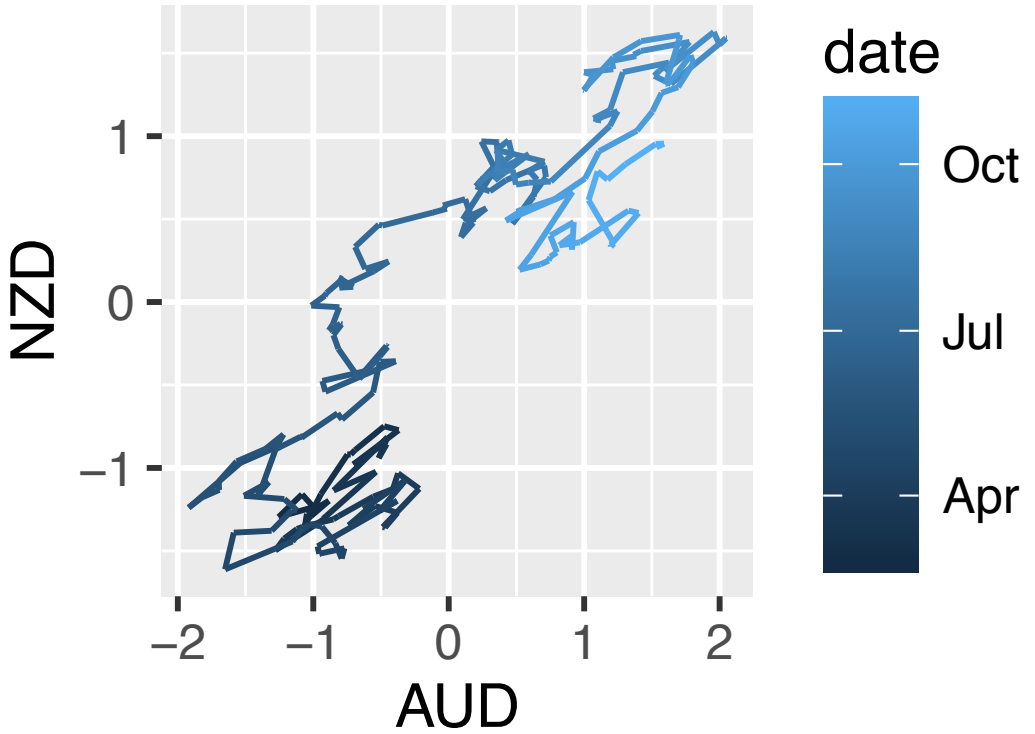
```
ggplot(data=rates.sub.m, aes(x=date, y=rate, colour=currency))
        geom_line()
```



- Now you can read off the trend: the AUD and NZD trend similarly in this time period, but the GBP is different. The GBP goes down in cross-rate, as the AUD/NZD go up.

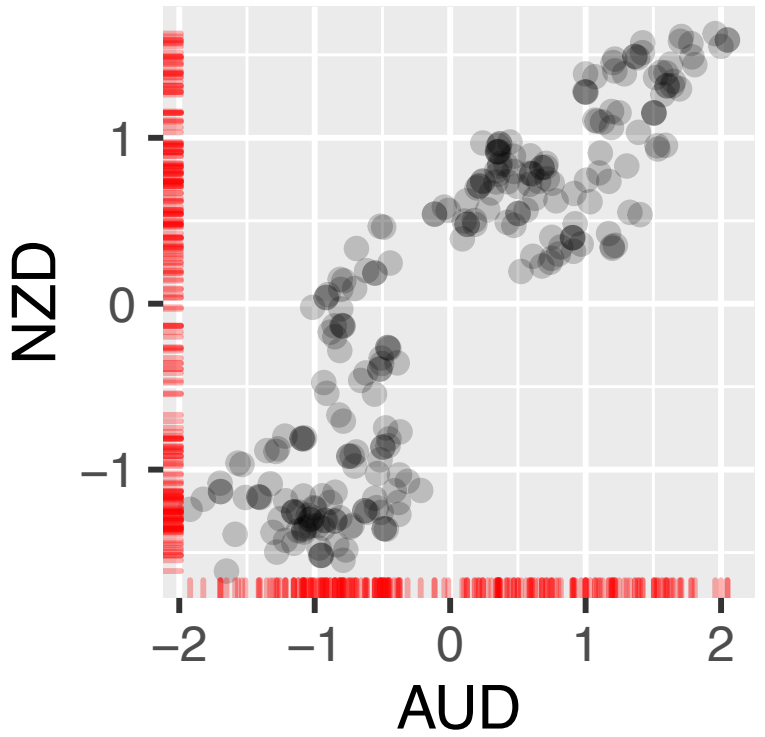
Your turn

In the plot below, how are variables mapped to plot elements?



Adding marginal rug plot

```
ggplot(data=rates.sub, aes(x=AUD, y=NZD)) +
  geom_point(alpha=0.2) + geom_rug(colour="red", alpha=0.3) +
  theme(aspect.ratio=1)
```



Other types of plots

- bar charts, pie charts
- boxplots, violins,
- histograms
- density plots
- dotplots

Look up `?geom_histogram` and choose the index for the ggplot2 package. Look at the `geom_` options. There are many! We will only cover the few main ones.

Type of variables suggests mapping

1. The values of **quantitative** variables should be mapped to **position along a line**, e.g. histogram, scatterplot. Mapping them to colour will yield only rough return of information to the reader.
2. Categorical variables could be mapped to
   - colour, if there are few categories,
   - aggregated and mapped to position along the line,
   - mapped to angle, if all categories are available.
3. Order is important, and if no natural order available then impose one e.g. using count

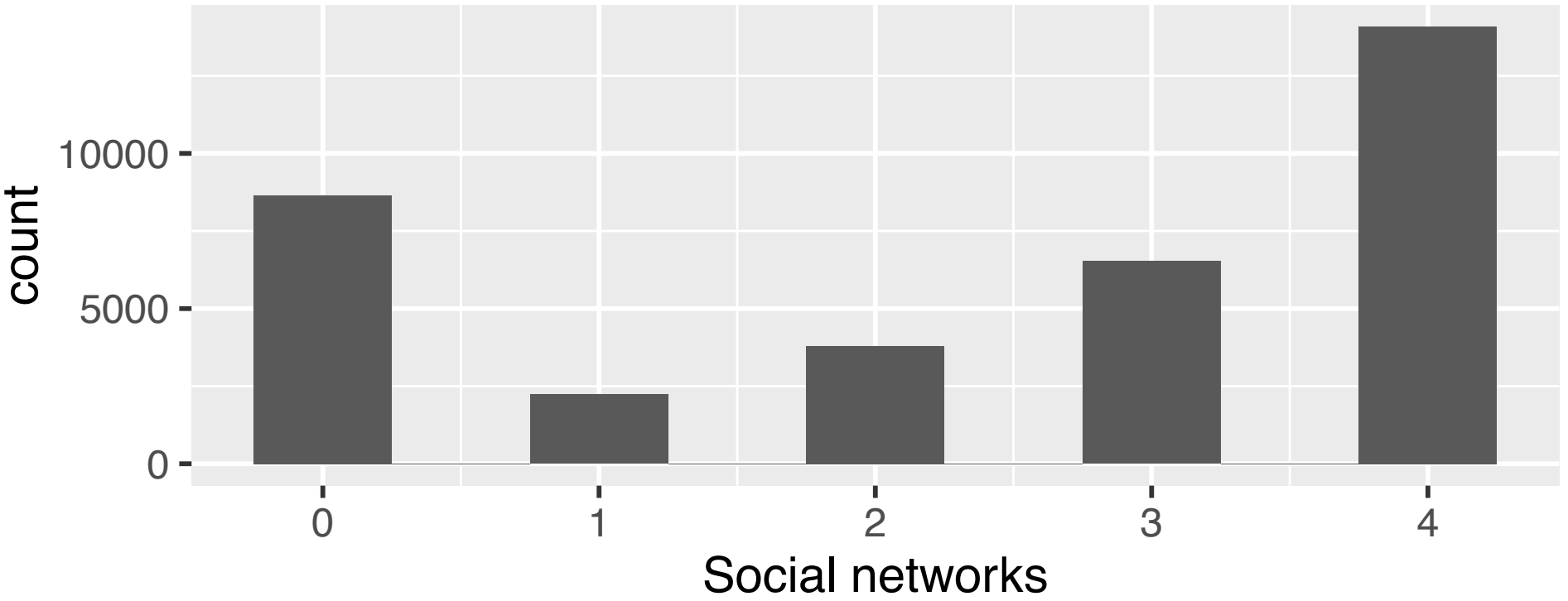
Categorical variables - barchart

The social variables of the PISA data include internet usage. This is a subset.

```
internet <- read_csv("data/internet.csv")
dim(internet)
#> [1] 37904    20
colnames(internet)
#>  [1] "name"                                "SCHOOLID"
#>  [3] "Gender"                              "One player games"
#>  [5] "Collaborative games"                 "Use email"
#>  [7] "Chat on line"                        "Social networks"
#>  [9] "Browse the Internet for fun"         "Read news"
#> [11] "Obtain practical information"        "Download music"
#> [13] "Upload content"                      "Internet for school"
#> [15] "Email students"                      "Email teachers"
#> [17] "Download from School"                "Announcements"
#> [19] "Homework"                            "Share school material
```

Categorical variables - barchart

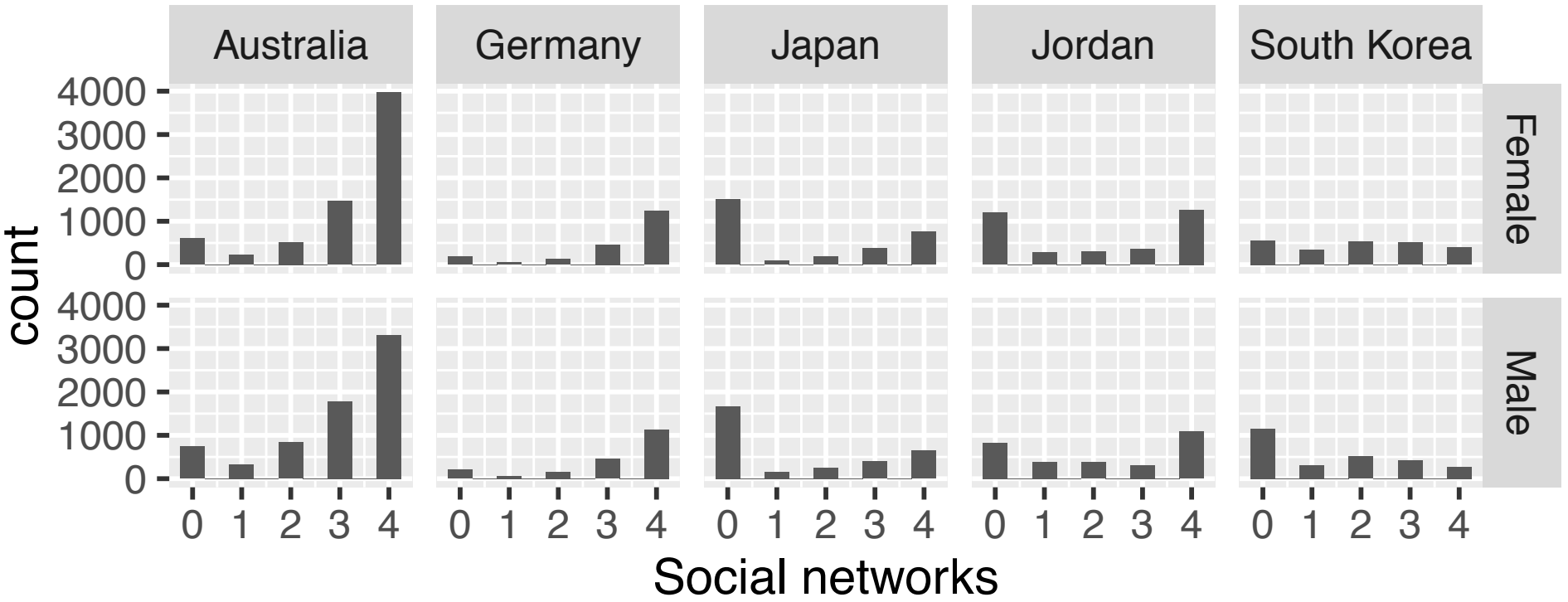
```
ggplot(data=internet, aes(x=`Social networks`)) +
  geom_bar(binwidth=0.5)
```



Categorical variables - barchart

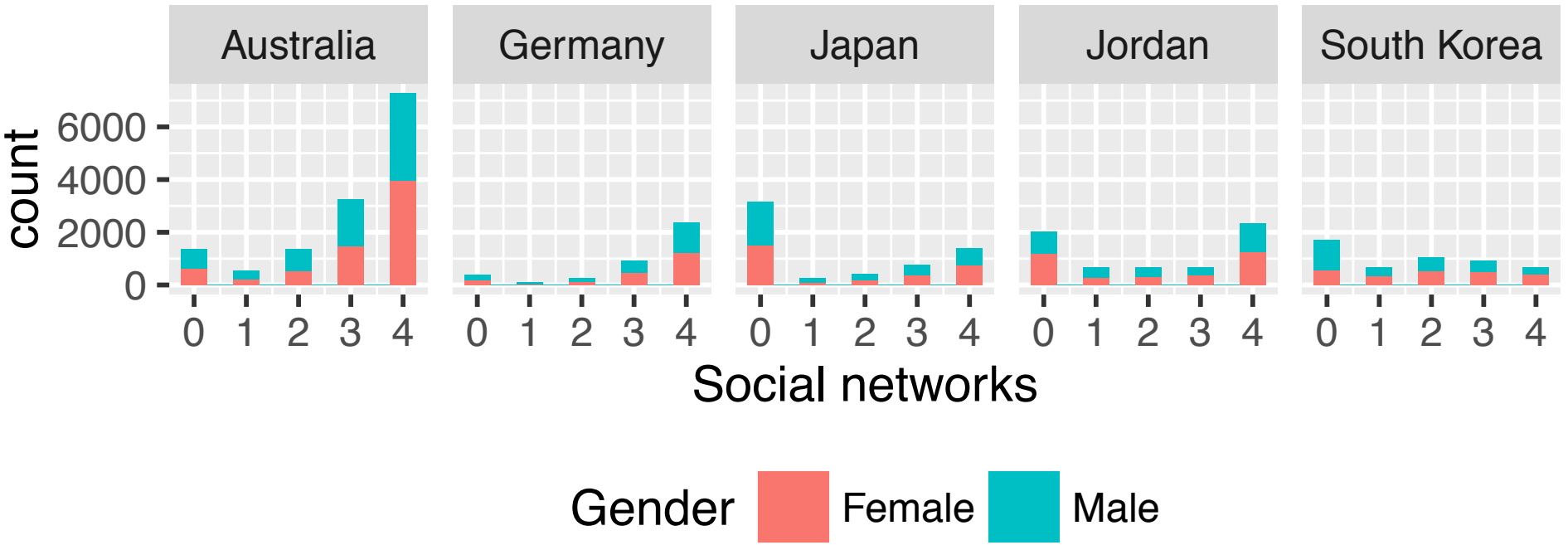
Simpson's paradox may be in play when there are multiple categorical variables. Need to divide it into basic elements.

```
ggplot(data=internet, aes(x=`Social networks`)) +
  geom_bar(binwidth=0.5) +
  facet_grid(Gender~name)
```



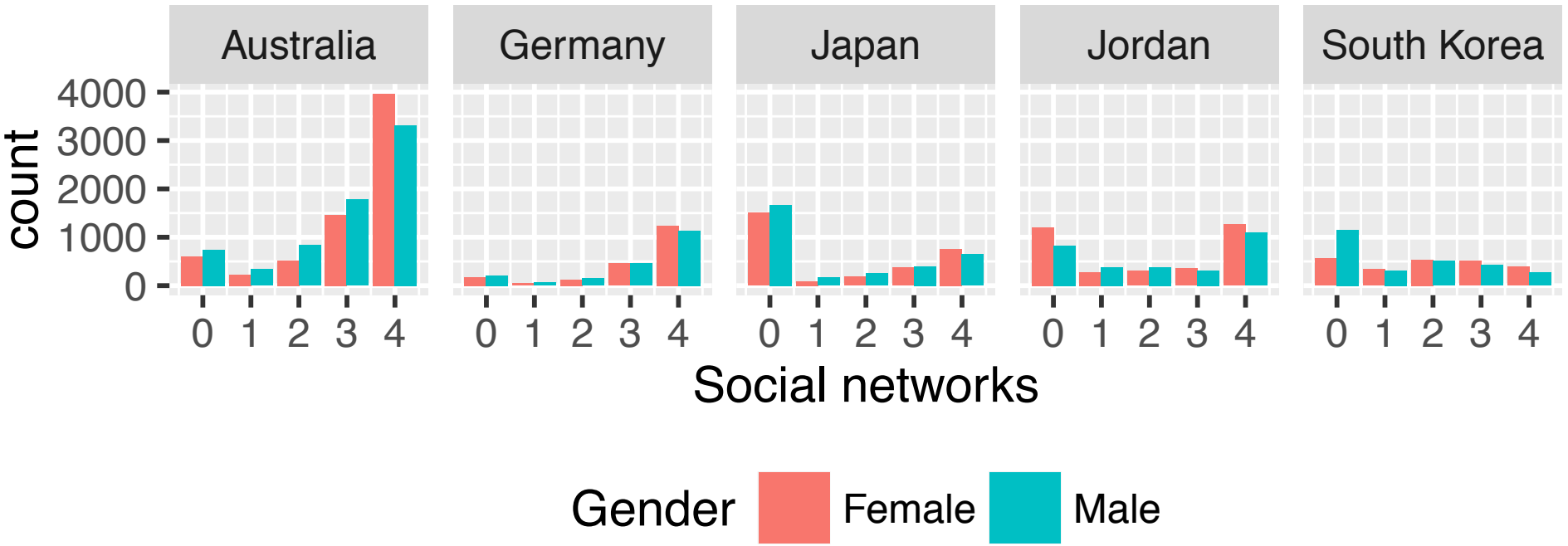
Categorical variables - stacked barchart

```
ggplot(data=internet, aes(x=`Social networks`, fill=Gender)) +
  geom_bar(binwidth=0.5) +
  facet_wrap(~name, ncol=5) + theme(legend.position="bottom")
```



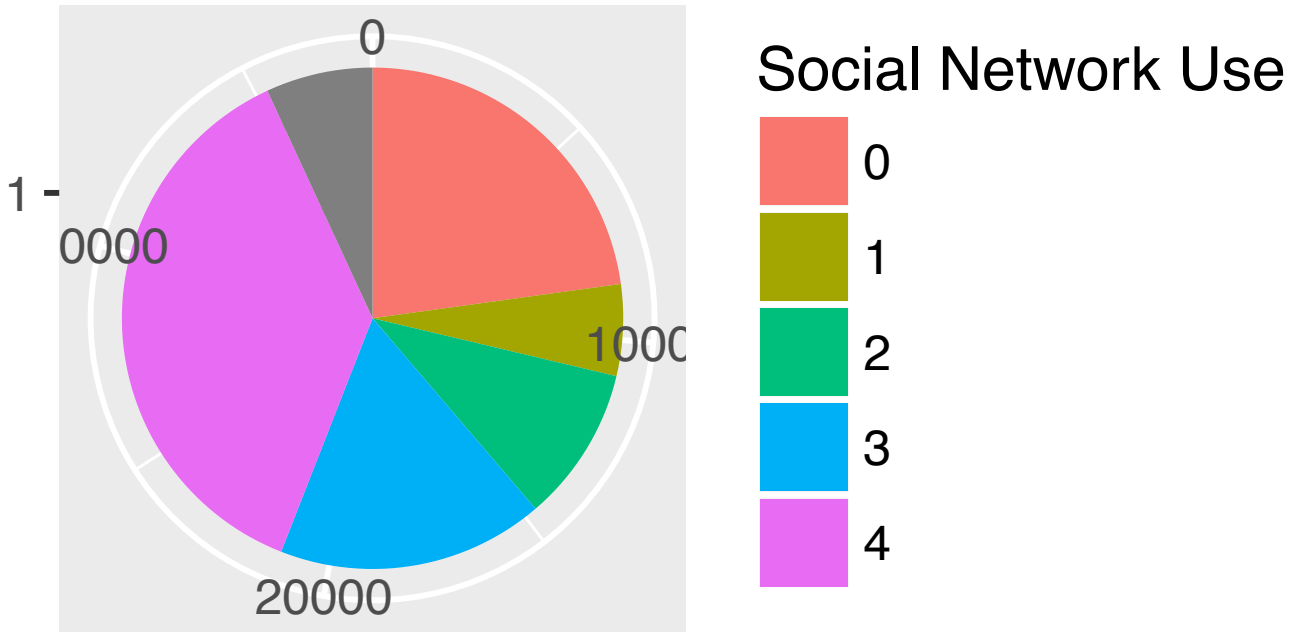
Categorical variables - dodged bars

```
ggplot(data=internet) +
  geom_bar(aes(x=`Social networks`, fill=Gender),
           position="dodge") +
  facet_wrap(~name, ncol=5) +
  theme(legend.position="bottom")
```



Categorical variables - piechart

```
ggplot(data=internet, aes(x=factor(1), fill=factor(`Social net
  geom_bar(width = 1) + scale_x_discrete("") +
  scale_y_continuous("") +
  scale_fill_hue("Social Network Use") +
  coord_polar(theta = "y")
```



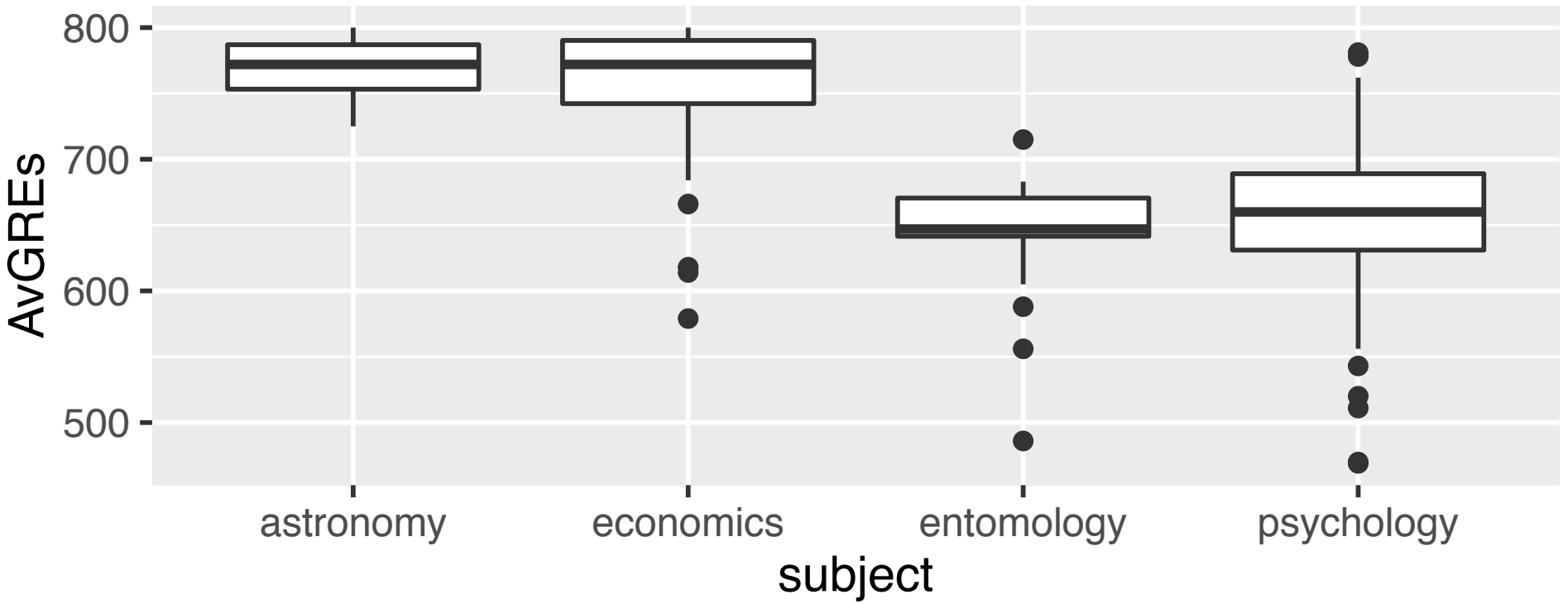
Yes, its deliberately made hard to do !

Quantitative and categorical - boxplots

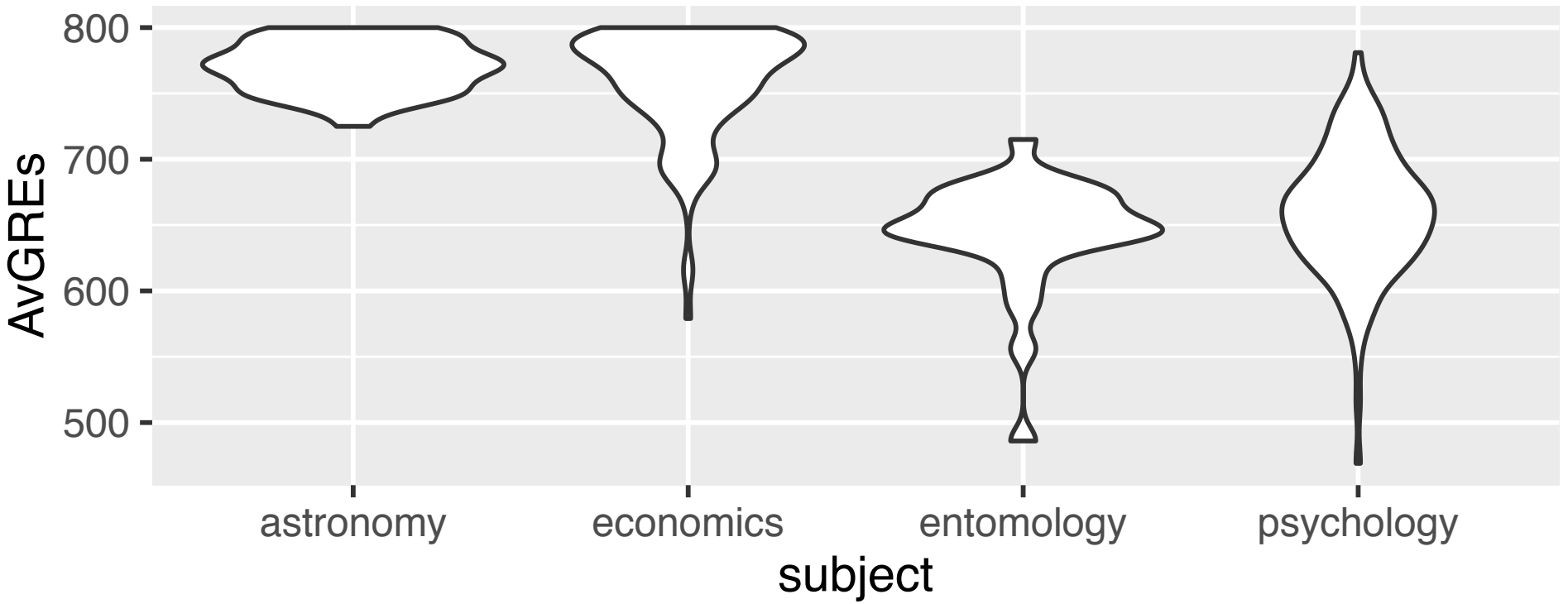
Data are measurements from the National Research Council in the USA, evaluating graduate programs in Statistics.

```
grad <- read_csv("data/graduate-programs.csv")
dim(grad)
#> [1] 412  16
colnames(grad)
#>  [1] "subject"           "Inst"              "AvNumPubs"
#>  [4] "AvNumCits"         "PctFacGrants"      "PctComplet
#>  [7] "MedianTimetoDegree" "PctMinorityFac"   "PctFemaleF
#> [10] "PctFemaleStud"     "PctIntlStud"       "AvNumPhDs"
#> [13] "AvGREs"            "TotFac"            "PctAsstPro
#> [16] "NumStud"
```

```
ggplot(data=grad, aes(x=subject, y=AvGREs)) +
  geom_boxplot()
```



```
ggplot(data=grad, aes(x=subject, y=AvGREs)) +
  geom_violin()
```



Your turn

- Create a side-by-side boxplot of average number of publications by program
- Then answer, "how do the four programs compare in terms of average number of publications?"

Cognitive principles

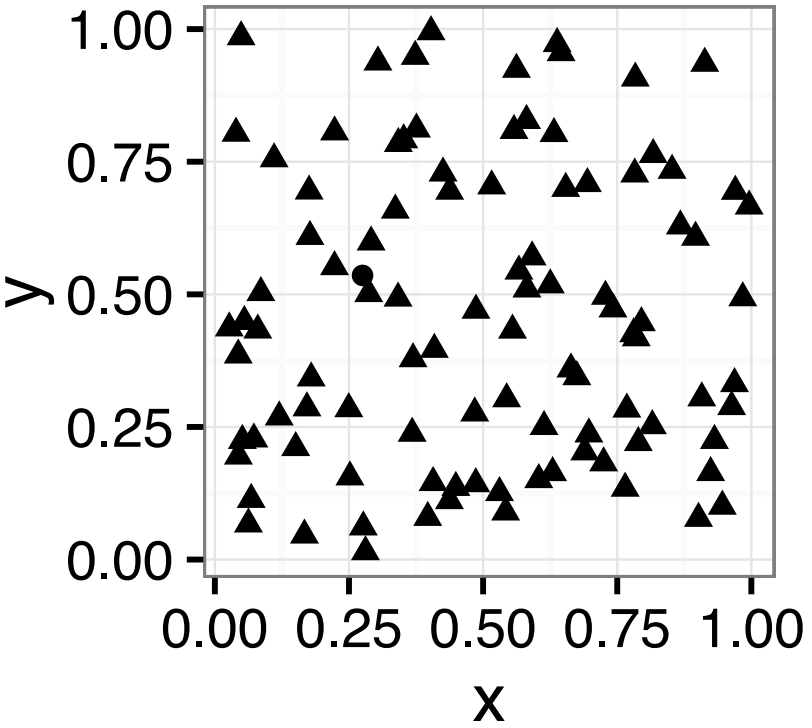
- **Hierarchy of mappings**: (first) position along an axis - (last) color (Cleveland, 1984; Heer and Bostock, 2009)
- **Pre-attentive**: Some elements like color are noticed before you even realise it. Other elements like axes are to look up information later.
- **Color palettes**: qualitative, sequential, diverging. The type of variable determines the appropriate palette.
- **Color blindness**: you can proof your plots with te dichromat package.
- **Proximity**: To compare elements, place them close together.
- **Change blindness**: When focus is interrupted differences may not be noticed, can occur when you are reading across multiple plots.

Hierarchy of mappings

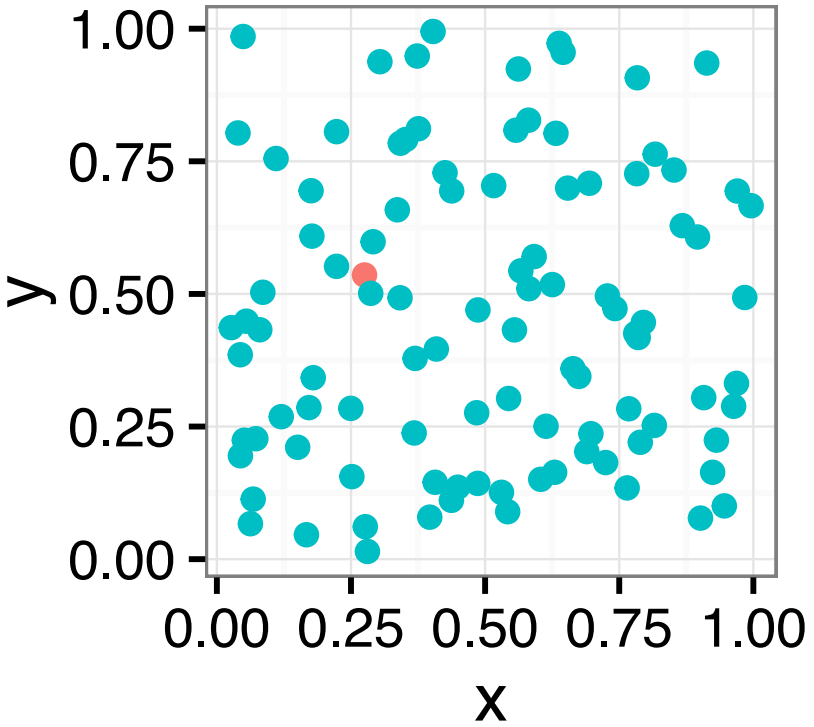
1. Position - common scale (BEST)
2. Position - nonaligned scale
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color (WORST)

Pre-attentive

Can you find the odd one out?

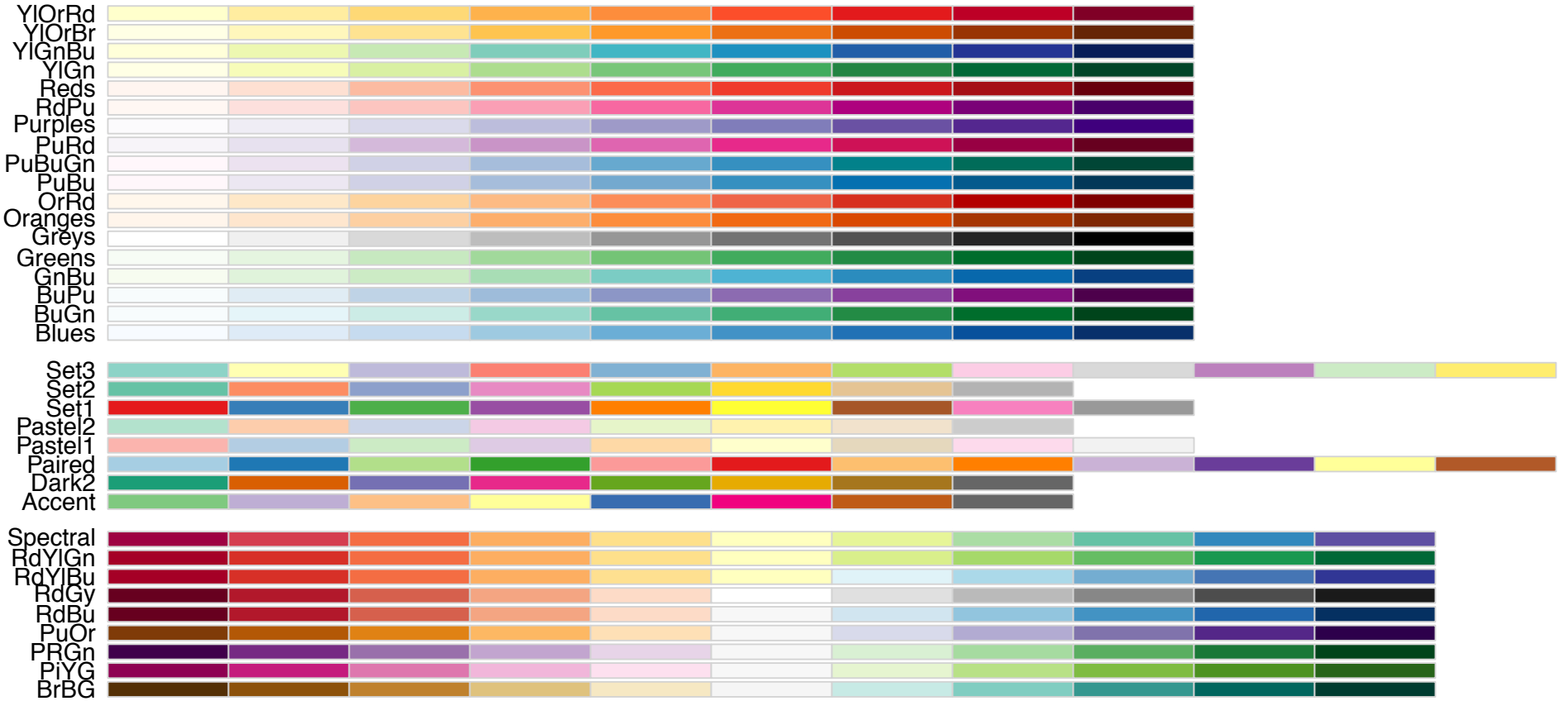


Is it easier now?



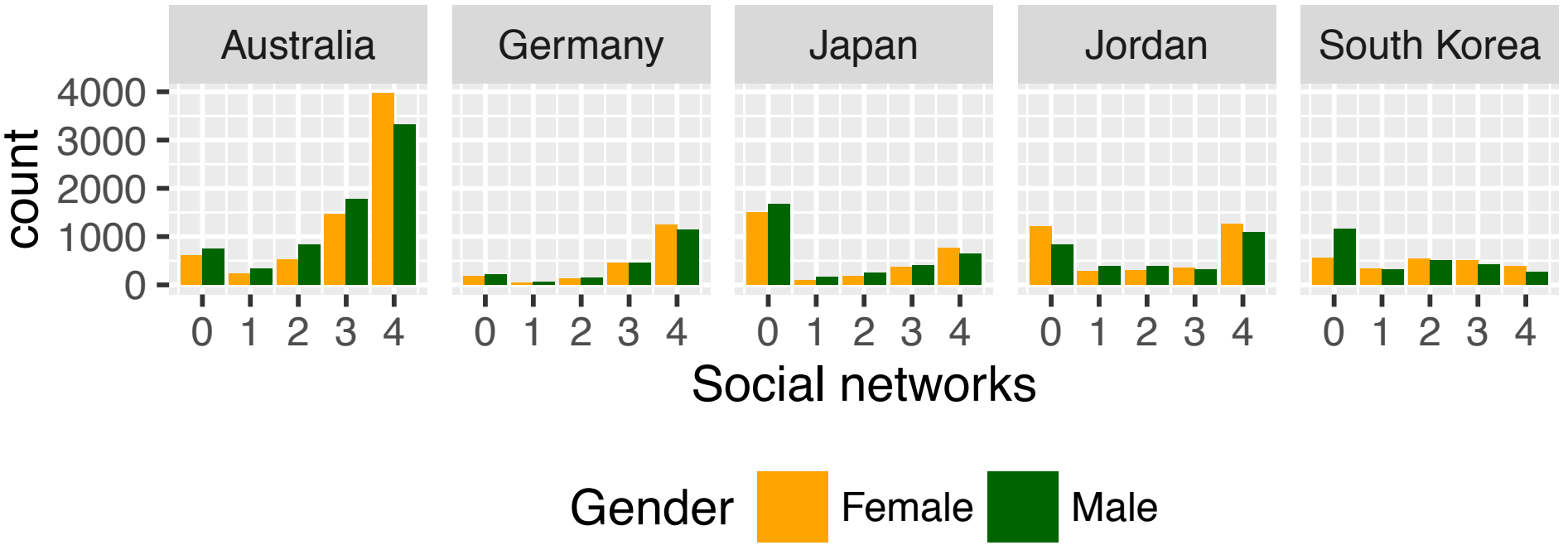
Color palettes

- Qualitative: categorical variables
- Sequential: low to high numeric values
- Diverging: negative to positive values



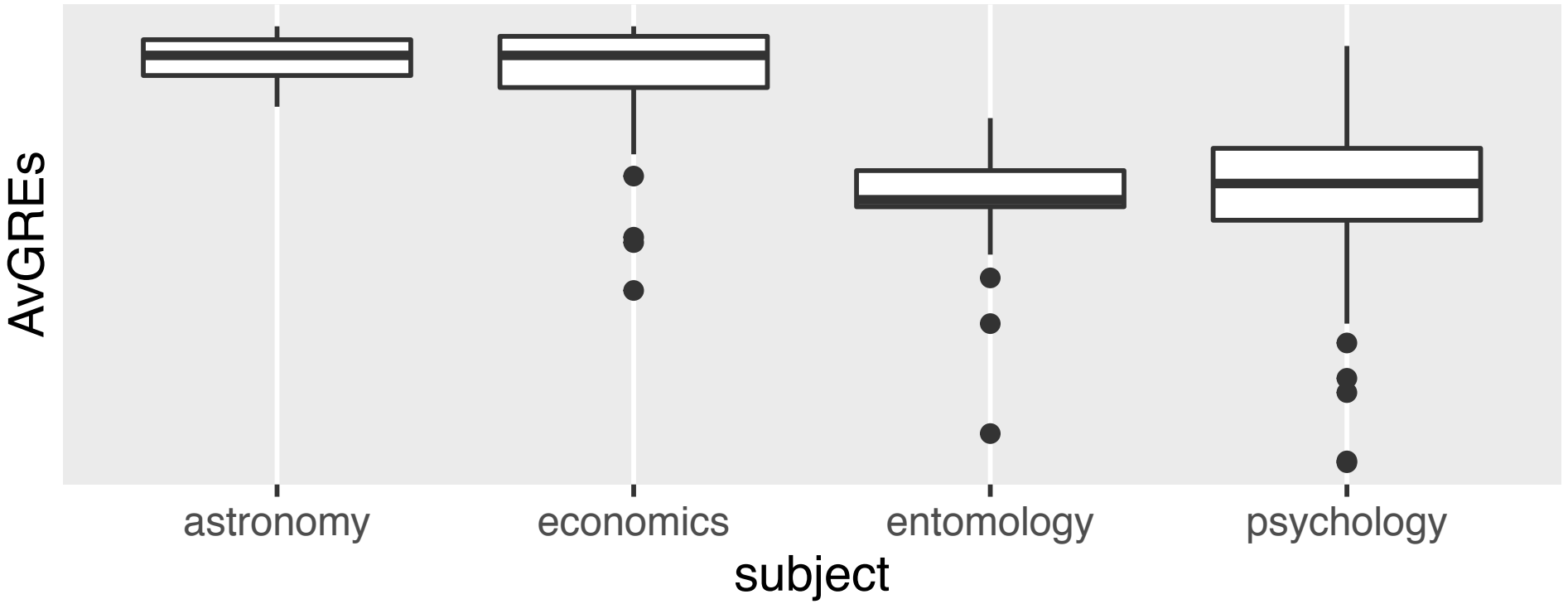
Scales

```
ggplot(data=internet, aes(x=`Social networks`, fill=Gender)) +
  geom_bar(position="dodge") +
  scale_fill_manual(values=c("Female"="orange", "Male"="darkgr
  facet_wrap(~name, ncol=5) +
  theme(legend.position="bottom")
```



Scales

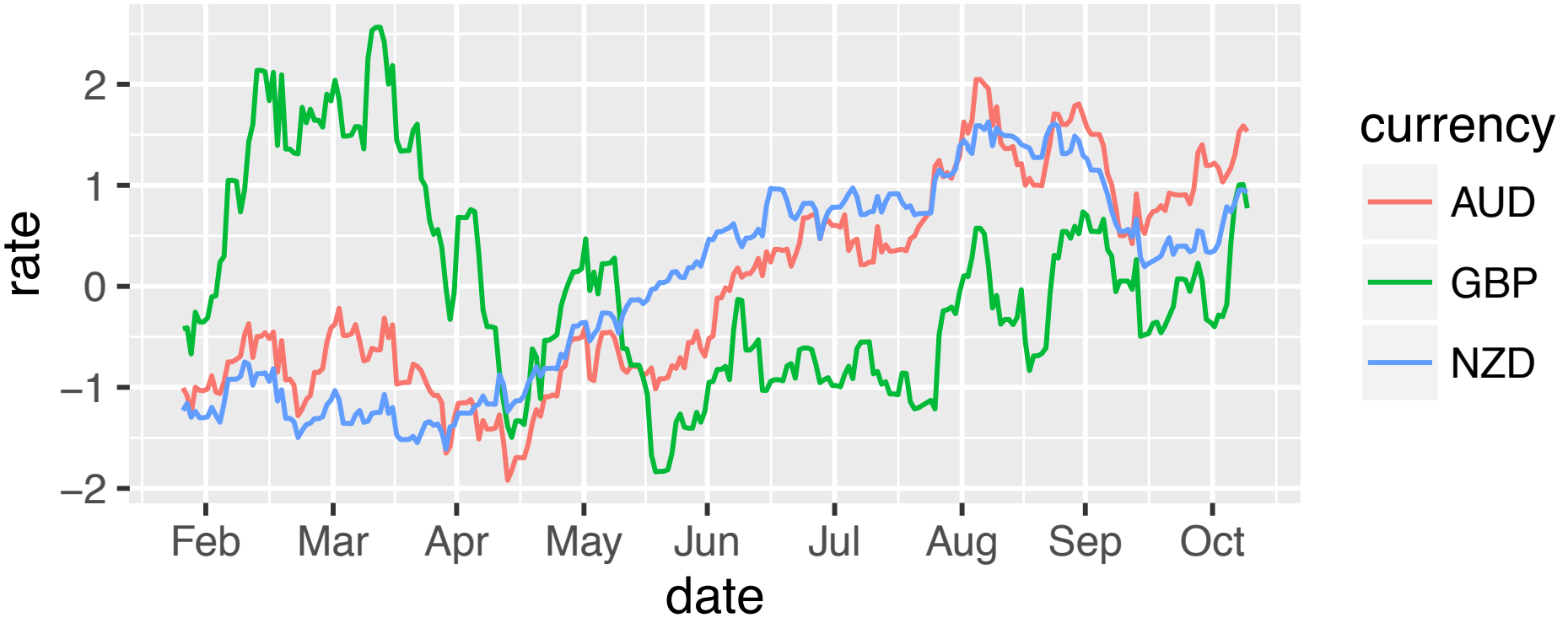
```
ggplot(data=grad, aes(x=subject, y=AvGREs)) +
  geom_boxplot() + scale_y_log10()
```



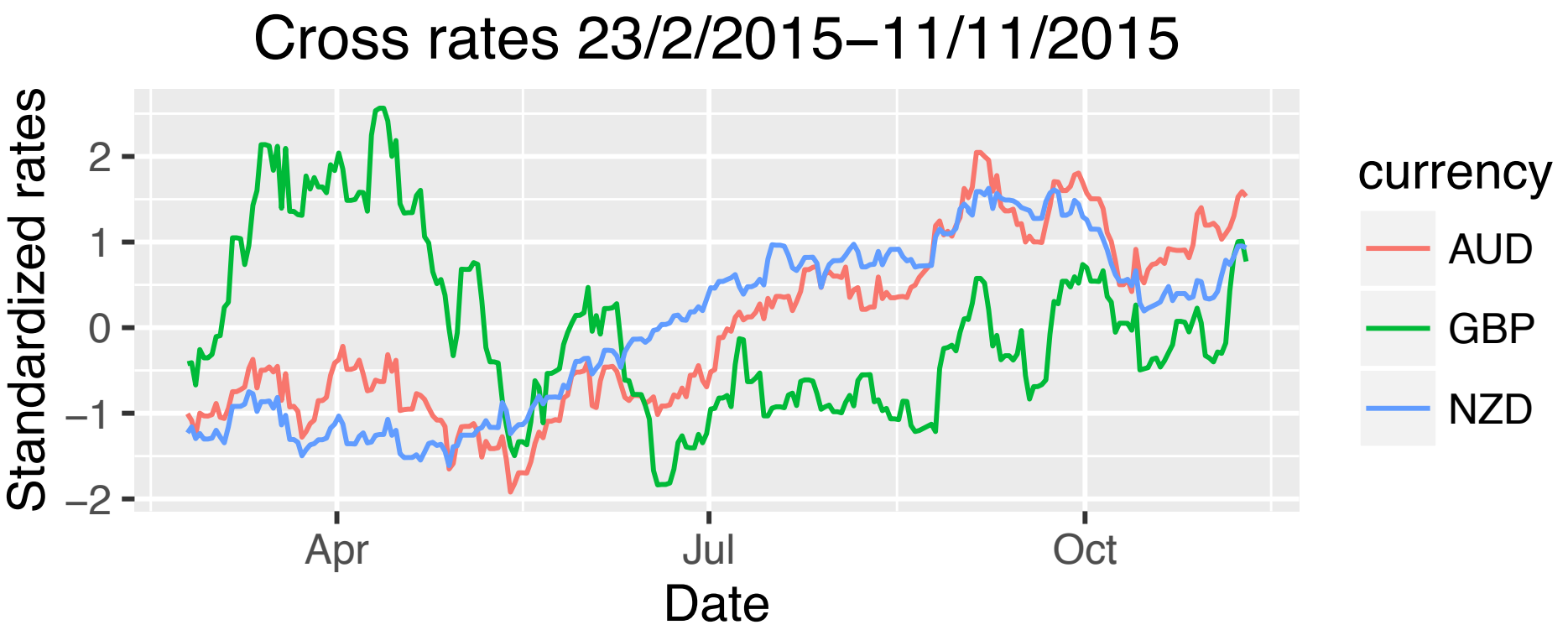
Axes

The date time axis is a little trickier to re-organise, but it can be done.

```
rates.sub.m$date <- as.POSIXct(rates.sub.m$date)
ggplot(data=rates.sub.m, aes(x=date, y=rate,
  colour=currency)) + geom_line() +
  scale_x_datetime(breaks = date_breaks("1 month"),
                   labels = date_format("%b"))
```

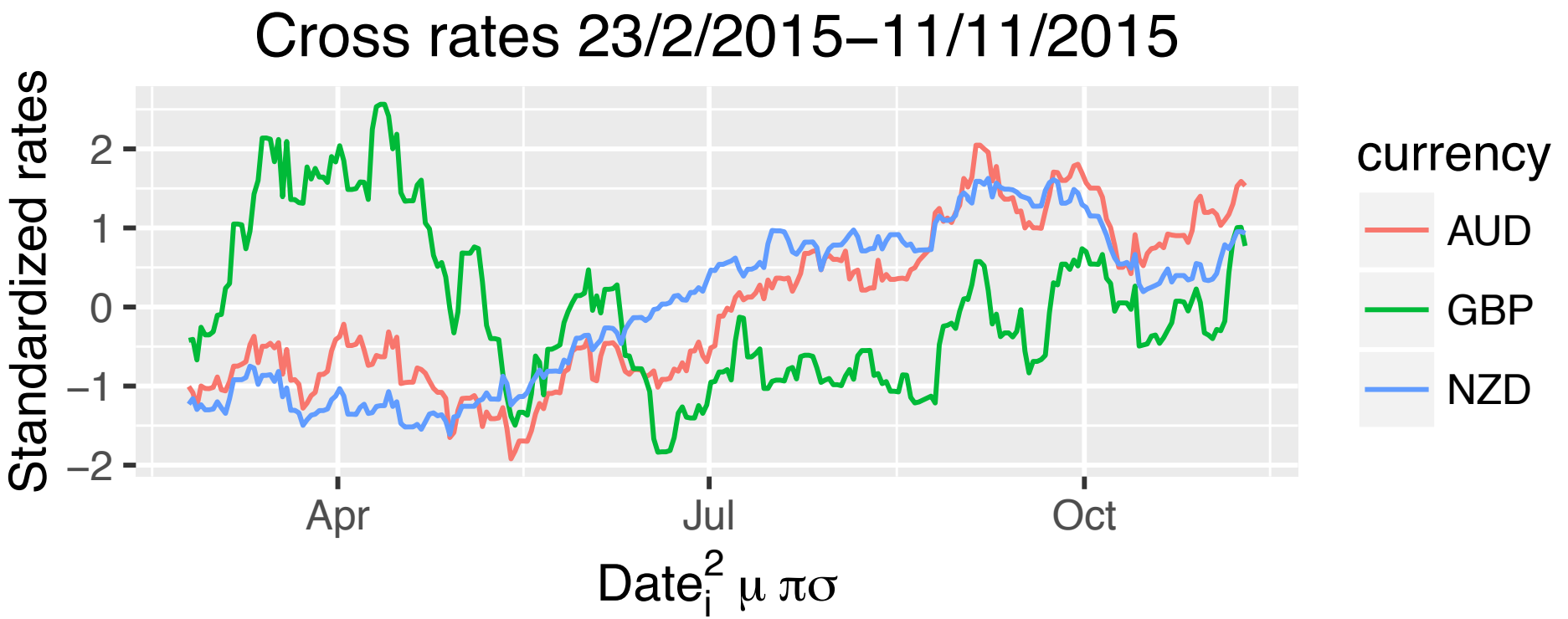


```
ggplot(data=rates.sub.m, aes(x=date, y=rate,
        colour=currency)) + geom_line() +
  xlab("Date") + ylab("Standardized rates") +
  ggtitle("Cross rates 23/2/2015-11/11/2015")
```



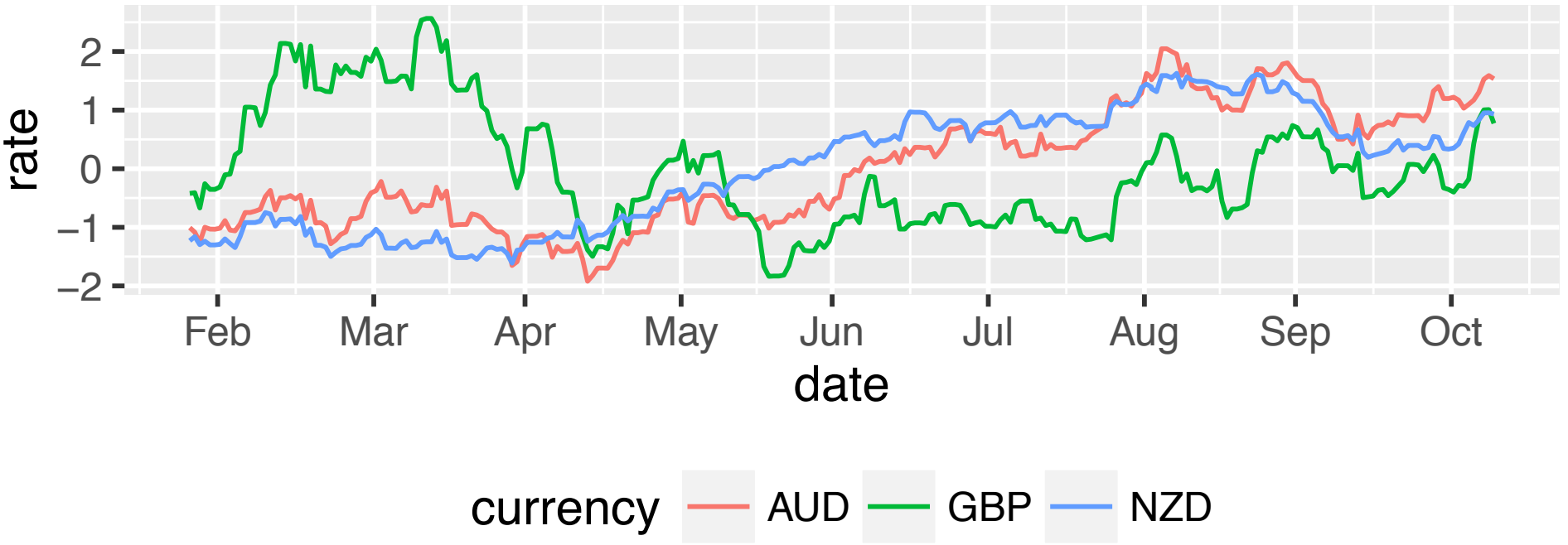
Equations in labels

```
ggplot(data=rates.sub.m, aes(x=date, y=rate, colour=currency))
  geom_line() +
  xlab(expression(Date[i]^2~ mu ~ pi * sigma)) +
  ylab("Standardized rates") +
  ggtitle("Cross rates 23/2/2015-11/11/2015")
```



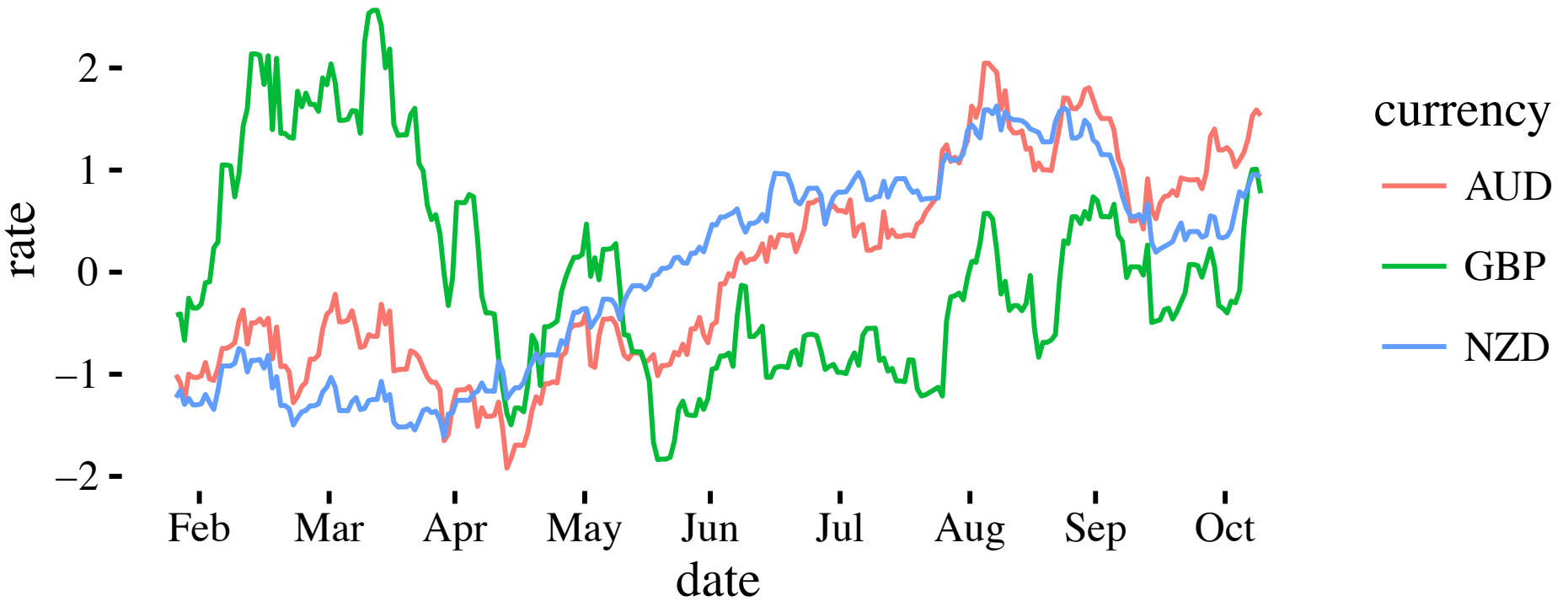
Legend Position

```
p + theme(legend.position = "bottom")
```

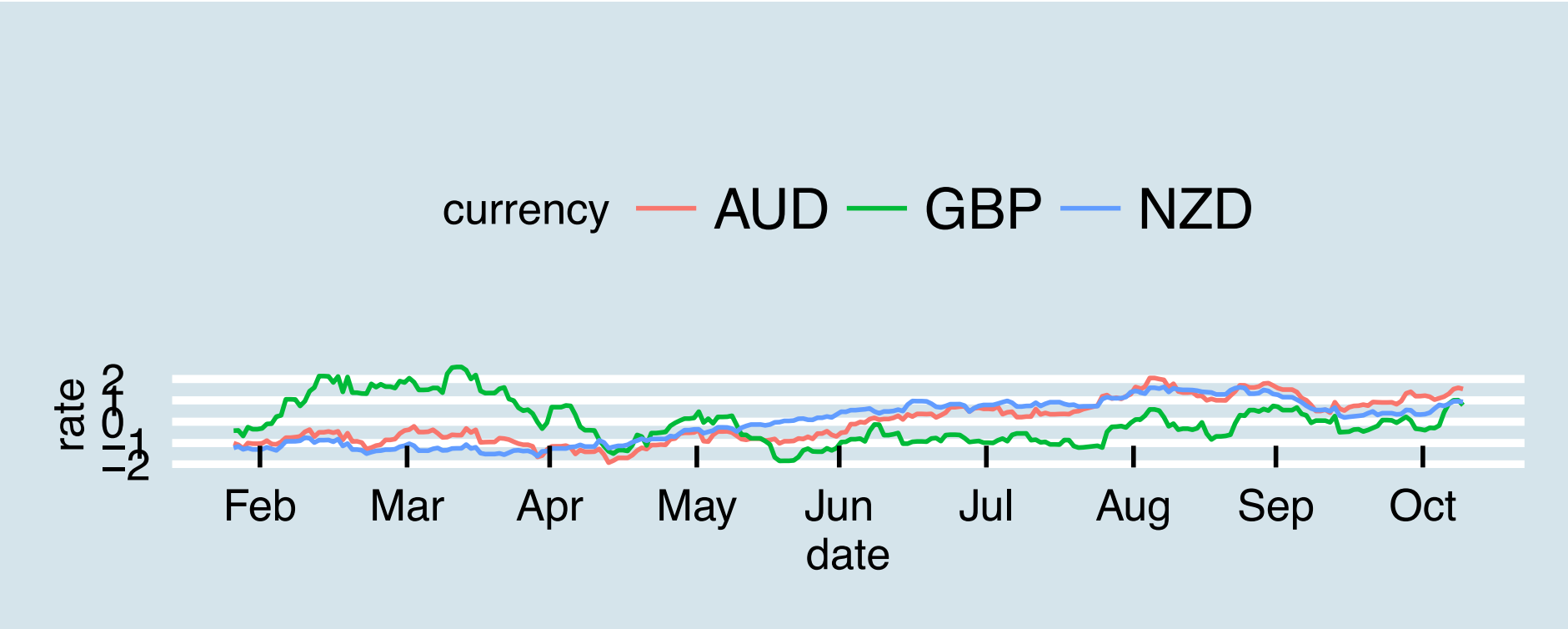


Themes

```
p + theme_tufte()
```

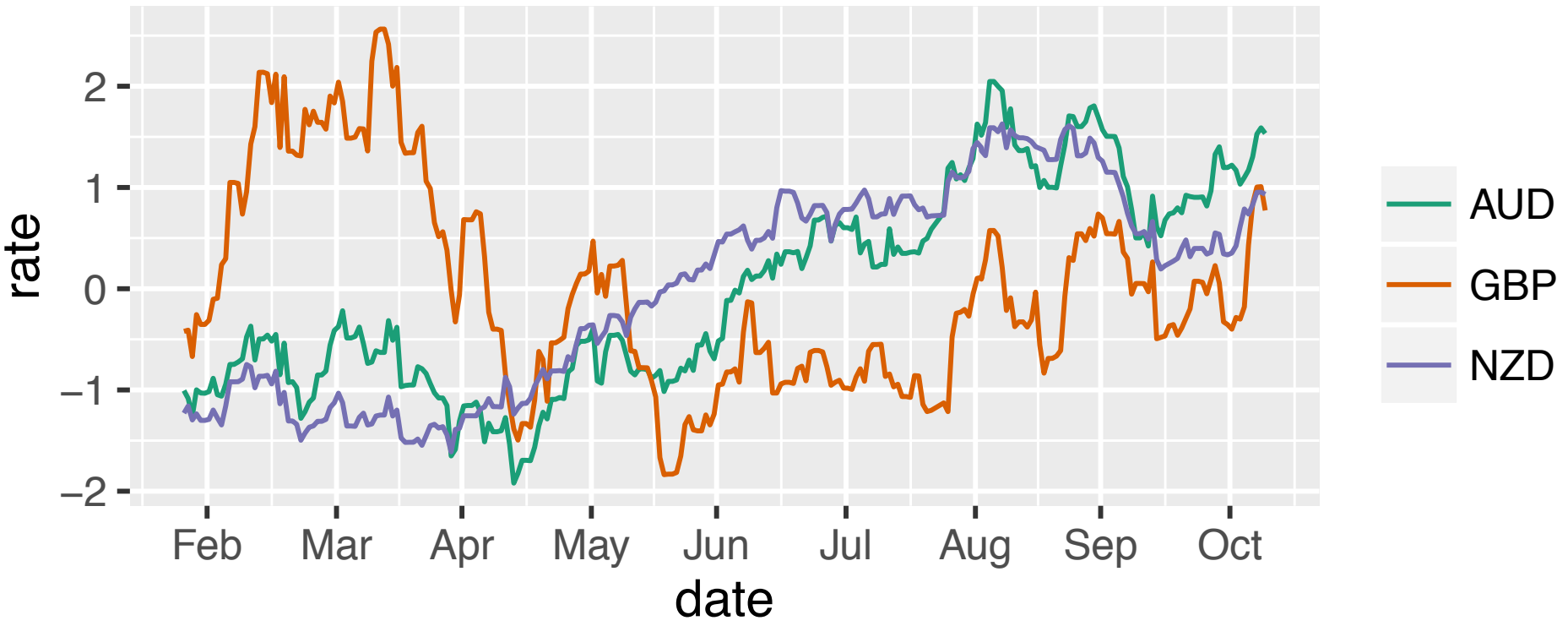


```
p + theme_economist()
```



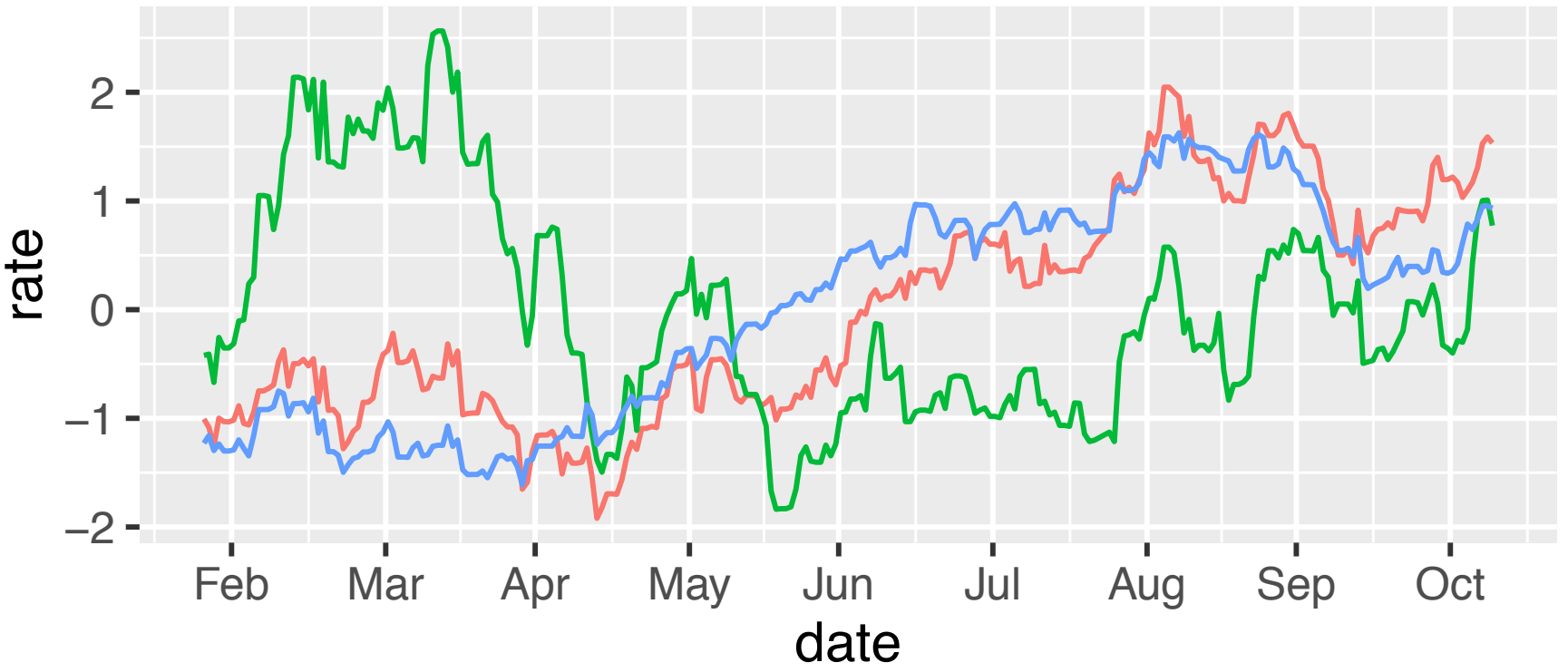
Color palettes

```
p + scale_color_brewer("", palette = "Dark2")
```

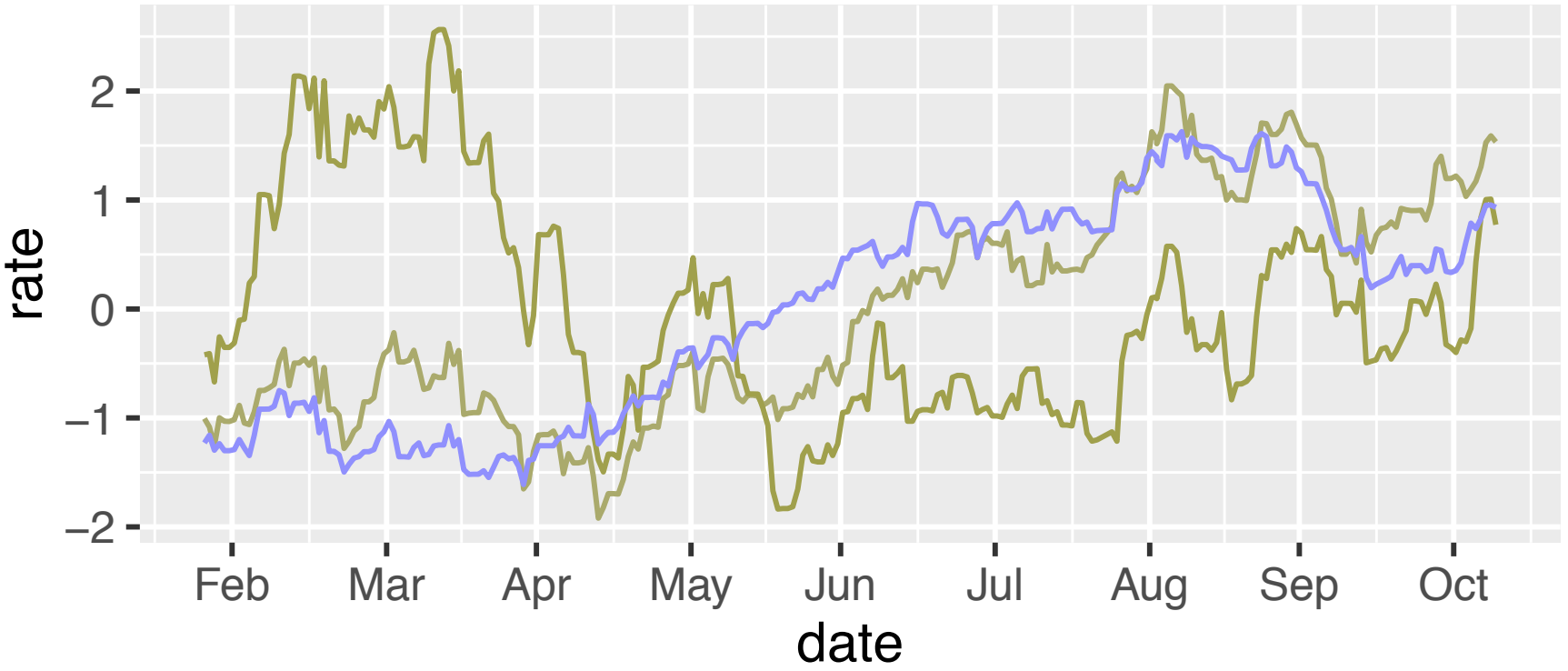


Color blind-proofing

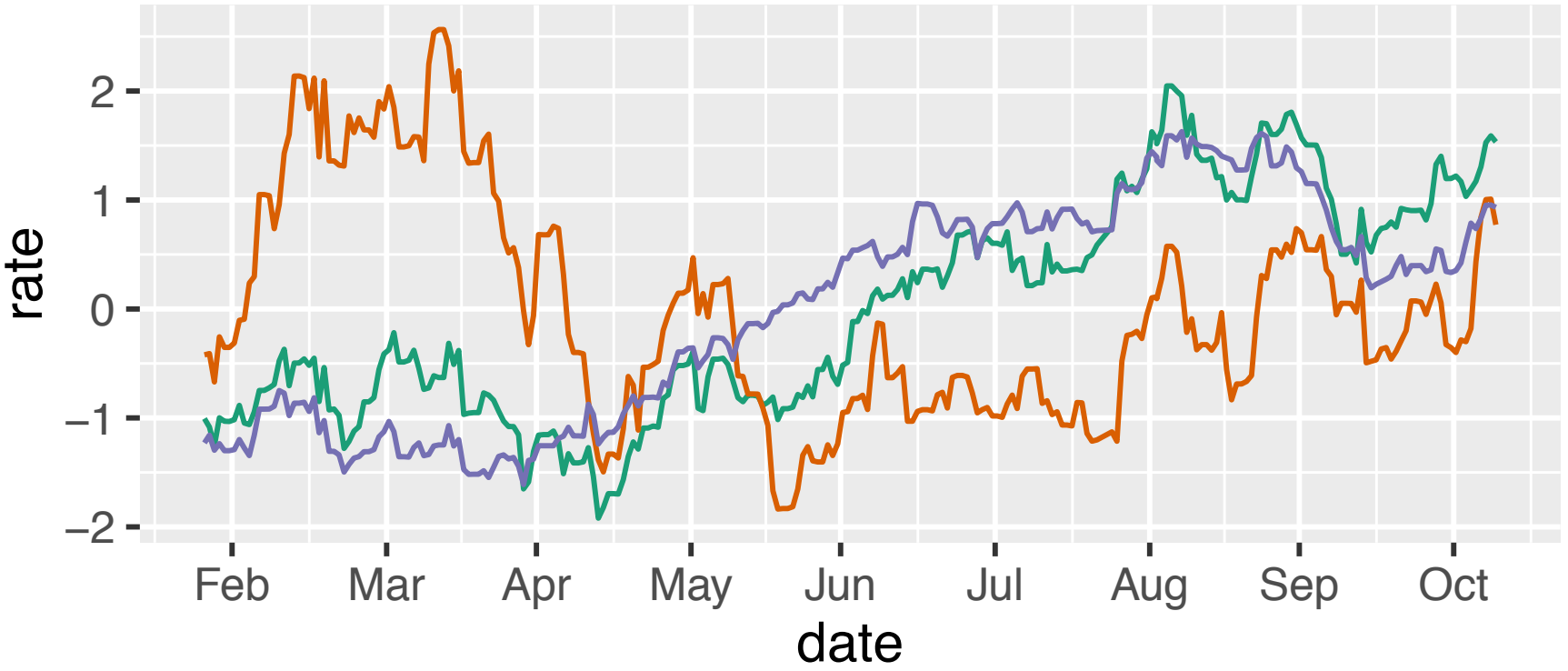
```
clrs <- hue_pal()(3)
p + scale_color_manual("", values=clrs) +
  theme(legend.position = "none")
```



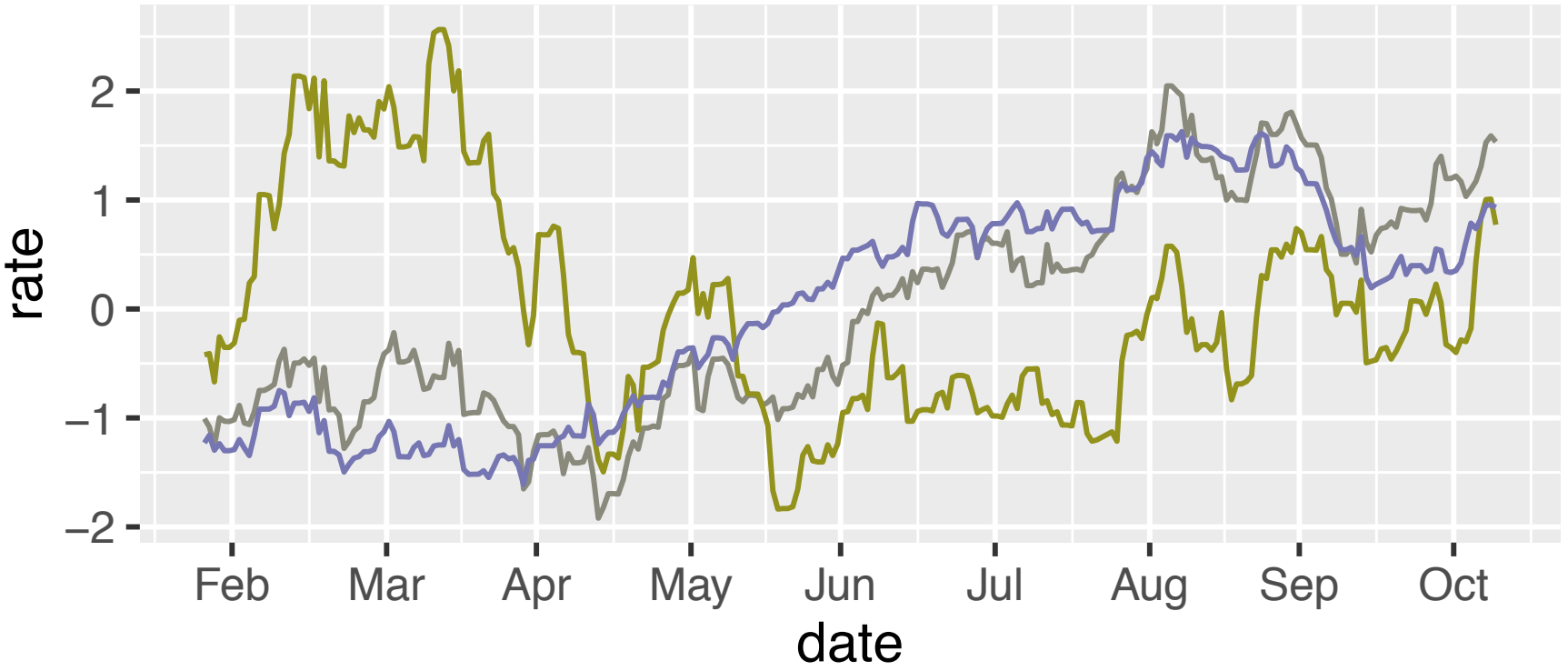

```
clrs <- dichromat(hue_pal()(3))
p + scale_color_manual("", values=clrs) +
  theme(legend.position = "none")
```




```
clrs <- brewer.pal(3, "Dark2")
p + scale_color_manual("", values=clrs) +
  theme(legend.position = "none")
```

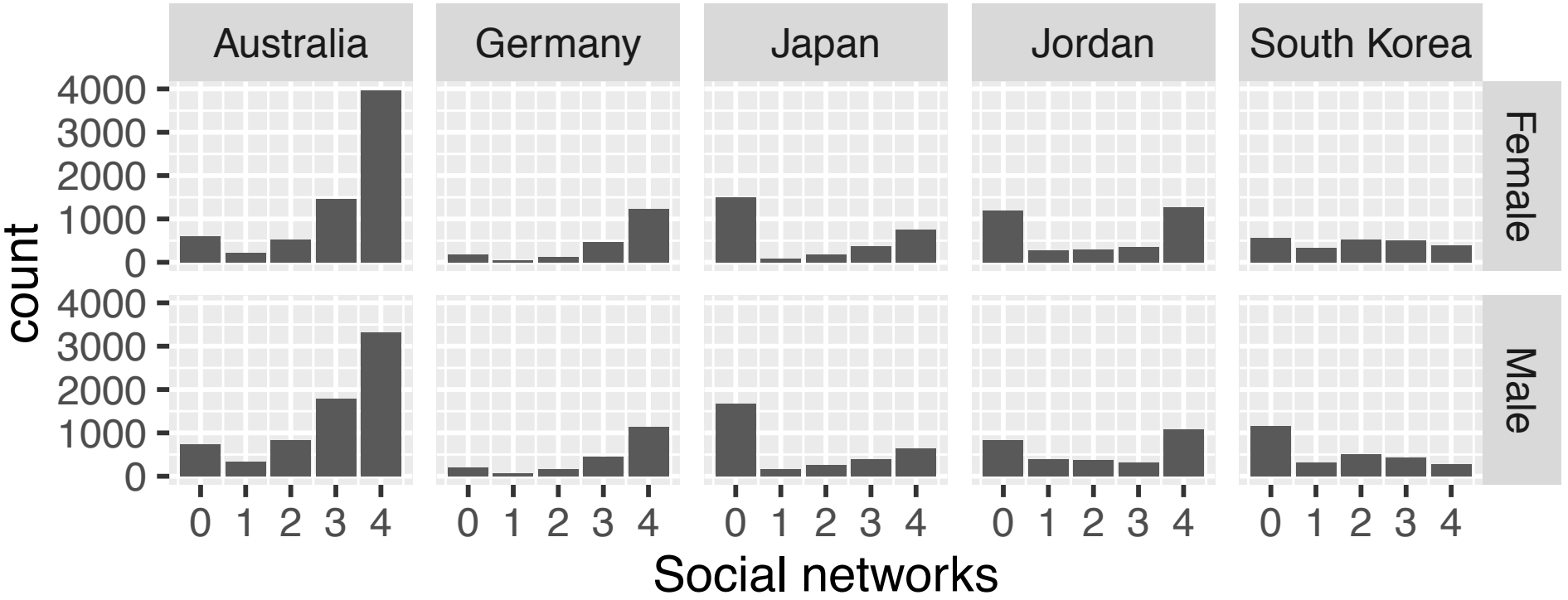



```
clrs <- dichromat(brewer.pal(3, "Dark2"))
p + scale_color_manual("", values=clrs) +
  theme(legend.position = "none")
```



Your turn

Proximity - From with plot can you answer: Is the proportion of girls who use social networks every day (4) higher than boys, in Australia? And is this different in Germany?



Your turn

- Brainstorm with your neighbour ways to rearrange this plot to answer the question.
- Then tackle this question: Are German girls more likely to report using social networks once or twice per month (1) than Japanese girls?
- What ways would you re-arrange the plot to tackle this one?

Proximity

- It is ok to make more than one plot.
- Actually it is recommended.

Your turn

For the NY workers compensation data

- Make a barchart of the district name
- Fill the barchart by Gender, but make the height of the bars equal
- Make a violin plot of age, by claim type

Your turn

How would you answer these questions?

- What is the most common district for injuries?
- Is the distribution of gender districts?
- Is the age of injury the same across the different claim types?

Credits

Notes prepared by Di Cook, building on joint workshops with Carson Sievert, Heike Hofmann, Eric Hare, Hadley Wickham.