ETC1010: Data Modelling and Computing

Lecture 5: Plotting your data

Di Cook (dicook@monash.edu, @visnut)

Week 5

Overview

- Grammar of graphics
  - Chloropleth Maps
  - Networks
  - Multivariate plots

Maps

- Maps are basically point data sets, with lines connecting dots in special order, and groups, yielding polygons defining geographic regions
- To fill polygons with colour corresponding to a variable, requires joining map data with data table
- Geographic regions often have multiple names, which may not match in the different tables

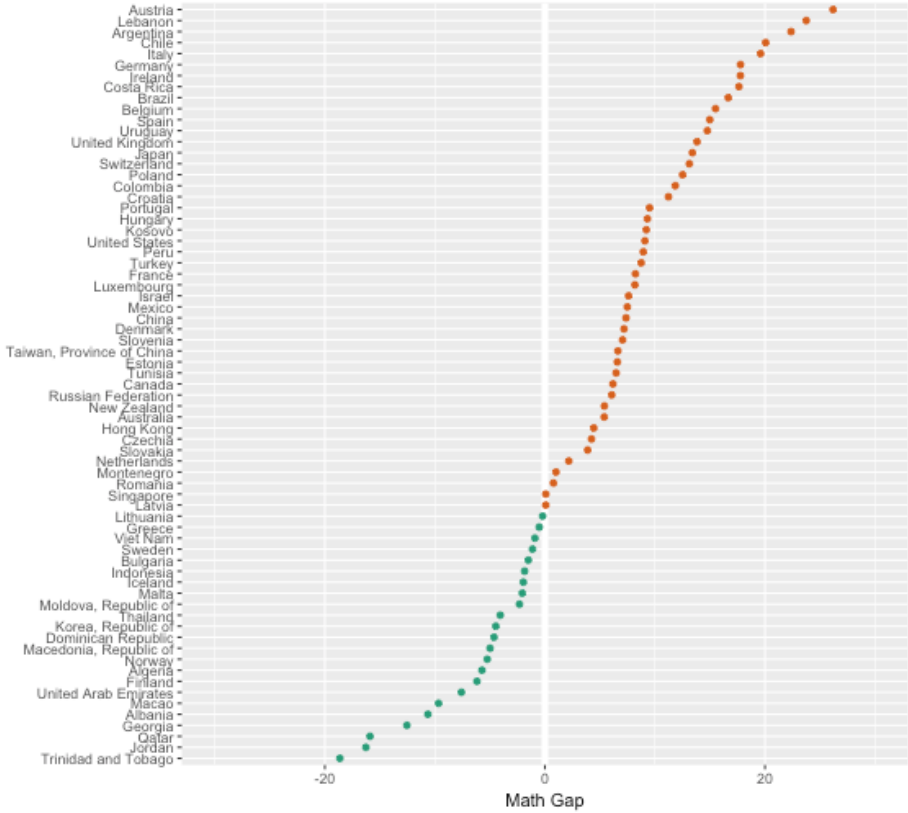
OECD PISA data

- About 500,000 students
- Approx 20,000 schools
- Around 70 countries tested every 3 years on reading, writing and science.
- Nearly 1000 variables collected on each student, and more on parents and schools

Gender gap

- We often hear in the news that boys perform better than girls in math
- Examine this by taking the PISA scores, compute the average for boys and girls, and difference these for each country. The calculations use a weighted mean, because each student in the study has a sampling weight associated with them, indicating how representative they are of their demographic in the country.

Display means



Country labels

Original labels for countries are the three letter code:

```
> scores
# A tibble: 514,397 x 6
     CNT ST004D01T PV1MATH PV1READ PV1SCIE SENWT
   <chr>     <dbl>   <dbl>   <dbl>   <dbl> <dbl>
 1   ALB         1 462.940 429.846 517.092 2.181
 2   ALB         1 430.100 462.788 479.635 2.181
 3   ALB         1 302.612 503.169 446.930 2.181
```

Map

The maps ofetn have actual country names as the labels for each geographic polygon:

```
       long      lat group order region subregion
1 -69.89912 12.45200     1     1  Aruba      <NA>
2 -69.89571 12.42300     1     2  Aruba      <NA>
3 -69.94219 12.43853     1     3  Aruba      <NA>
4 -70.00415 12.50049     1     4  Aruba      <NA>
5 -70.06612 12.54697     1     5  Aruba      <NA>
6 -70.05088 12.59707     1     6  Aruba      <NA>
```

ISO data base

```
library(ISOcodes)
data("ISO_3166_1")
ISO_3166_1 %>% select(Alpha_3, Name) %>% head()
  Alpha_3          Name
1     ABW         Aruba
2     AFG   Afghanistan
3     AGO        Angola
4     AIA      Anguilla
5     ALA Åland Islands
6     ALB       Albania
```

Join the PISA and ISO codes

```
scores <- tb %>%
  select(CNT, ST004D01T, PV1MATH, PV1READ, PV1SCIE, SENWT) %>% collect()
scores <- scores %>%
  mutate(CNT=recode(CNT, "QES"="ESP", "QCH"="CHN", "QAR"="ARG", "TAP"="TWN"))
  filter(CNT != "QUC") %>%
  filter(CNT != "QUD") %>%
  filter(CNT != "QUE")
countries <- scores %>%
  left_join(ISO_3166_1, by=c("CNT"="Alpha_3"))
countries$Name[countries$CNT == "KSV"] <- "Kosovo"
```

```
> countries
# A tibble: 514,397 x 7
     CNT ST004D01T PV1MATH PV1READ PV1SCIE SENWT    Name
   <chr>     <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <chr>
 1   ALB         1 462.940 429.846 517.092 2.181 Albania
 2   ALB         1 430.100 462.788 479.635 2.181 Albania
 3   ALB         1 302.612 503.169 446.930 2.181 Albania
 4   ALB         1 336.522 569.626 383.794 2.181 Albania
 5   ALB         1 290.929 389.138 412.304 2.181 Albania
```

Still some mismatches

In the PISA data:

```
pisa_gap <- pisa_gap %>%
  mutate(Name = recode(Name, "Czechia"="Czech Republic",
                       "Korea, Republic of"="South Korea",
                       "Macedonia, Republic of"="Macedonia",
                       "Moldova, Republic of"="Moldova",
                       "Russian Federation"="Russia",
                       "Taiwan, Province of China"="Taiwan",
                       "Trinidad and Tobago"="Trinidad",
                       "United States"="USA",
                       "United Kingdom"="UK",
                       "Viet Nam"="Vietnam"))
```

In the map data:

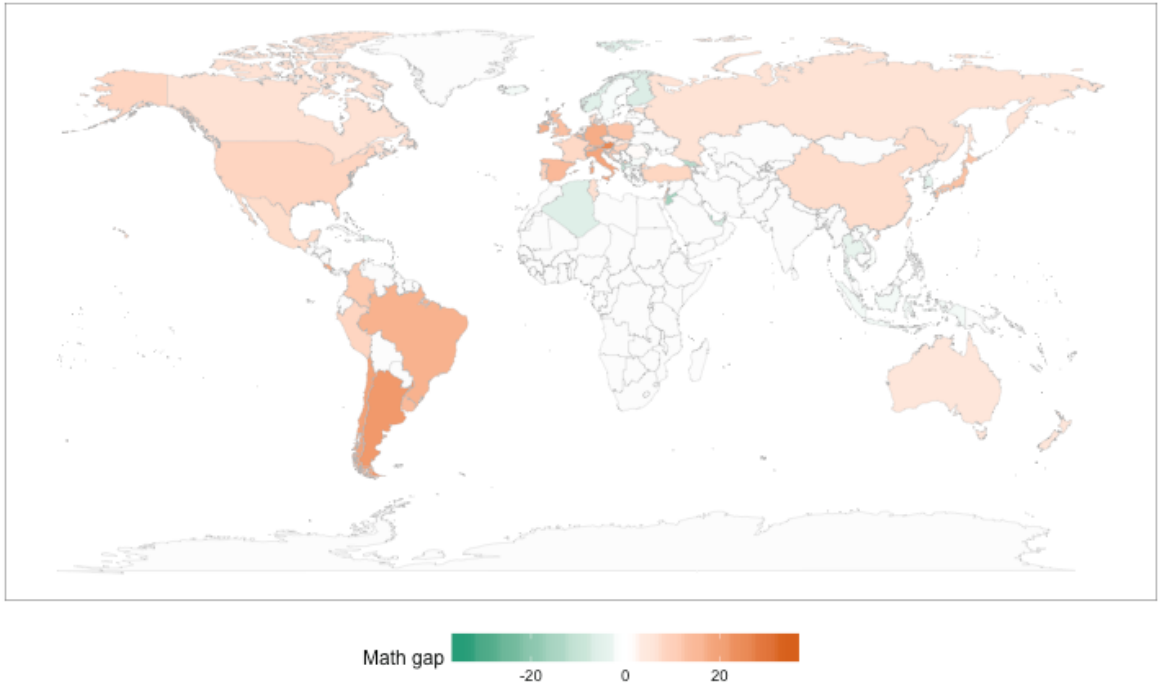
```
world_map$region[world_map$subregion == "Hong Kong"] <- "Hong Kong"
world_map$region[world_map$subregion == "Macao"] <- "Macao"
```

Now join:

```
to_map <- left_join(world_map, pisa_gap, by=c("region"="Name"))
```

Map it!

```
ggplot(to_map, aes(map_id = region)) +
    geom_map(aes(fill=wmathgap), map = world_map,
             color="grey70", size=0.1) +
    scale_fill_gradient2("Math gap", limits=c(-35, 35), na.value="grey99",
                         low="#1B9E77", high="#D95F02", mid="white") +
    expand_limits(x = world_map$long, y = world_map$lat) +
    theme_few() +
    theme(legend.position = "bottom",
          legend.key.width=unit(1.5, "cm"),
         axis.ticks = element_blank(),
         axis.title = element_blank(),
         axis.text =  element_blank())
```


Math gap
-20
0
20

Networks

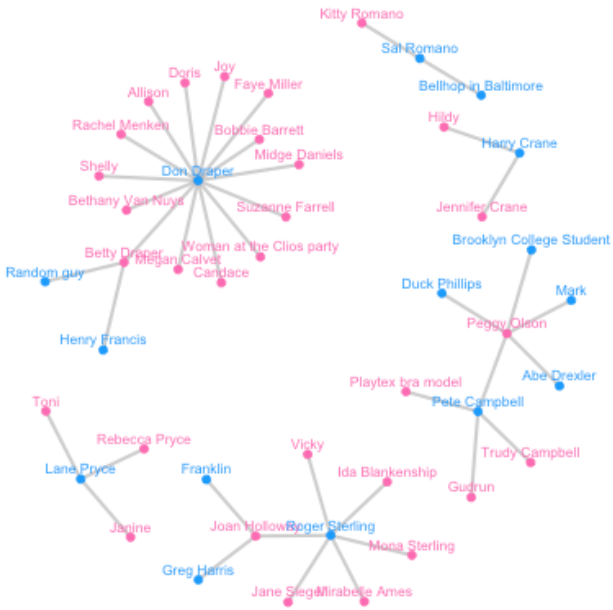
Network data arises in many settings, e.g. study of communities, biological pathways, ... Typically the data is provided in two related tables, nodes and edges. Both may have additional attributes.

Here's an example from the TV series Madmen. The nodes data contains the actors in the series, and the edges contains pairs of actors that had romantic relationships.

```
List of 2
 $ edges   :'data.frame':    39 obs. of  2 variables:
  ..$ Name1: Factor w/ 9 levels "Betty Draper",..: 1 1 2 2 2 2 2 2 2 2 ...
  ..$ Name2: Factor w/ 39 levels "Abe Drexler",..: 15 31 2 4 5 6 8 9 11 21 ...
 $ vertices:'data.frame':    45 obs. of  2 variables:
  ..$ label : Factor w/ 45 levels "Abe Drexler",..: 5 9 16 23 26 32 33 38 39 17 ...
  ..$ Gender: Factor w/ 2 levels "female","male": 1 2 2 1 2 1 2 2 2 2 ...
```

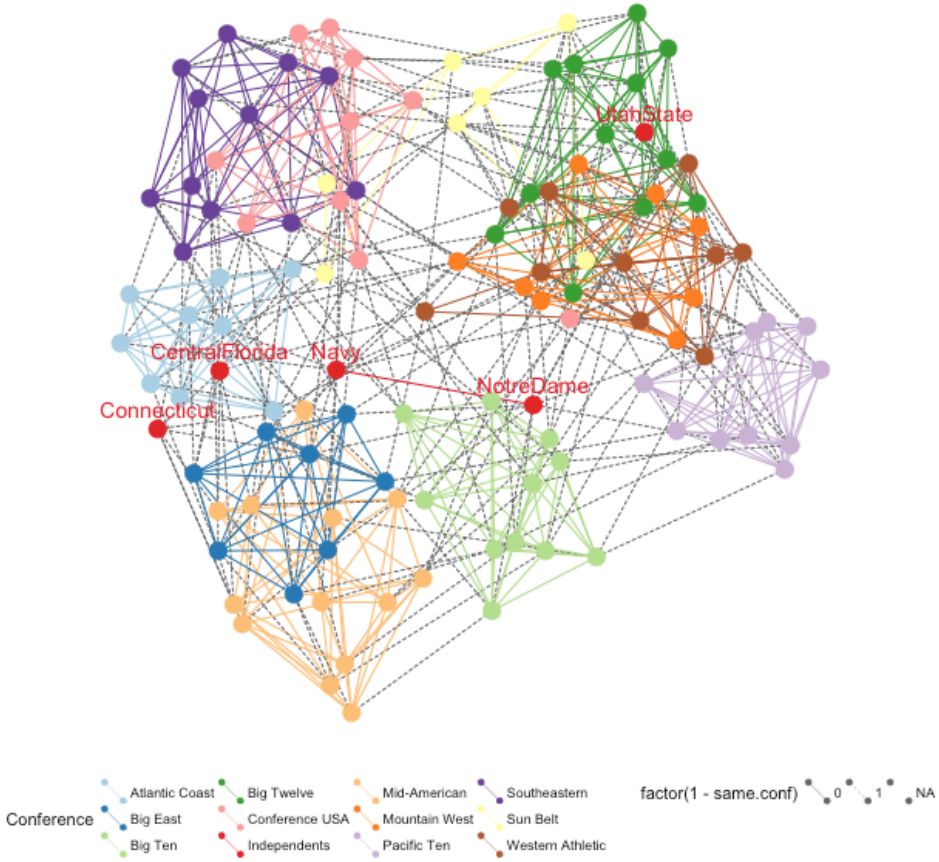
Generate a network view

- Create a layout (in 2D) which places nodes which are most related close,
- Plot the nodes as points, connect the appropriate lines
- Overlaying other aspects, e.g. gender



American college football

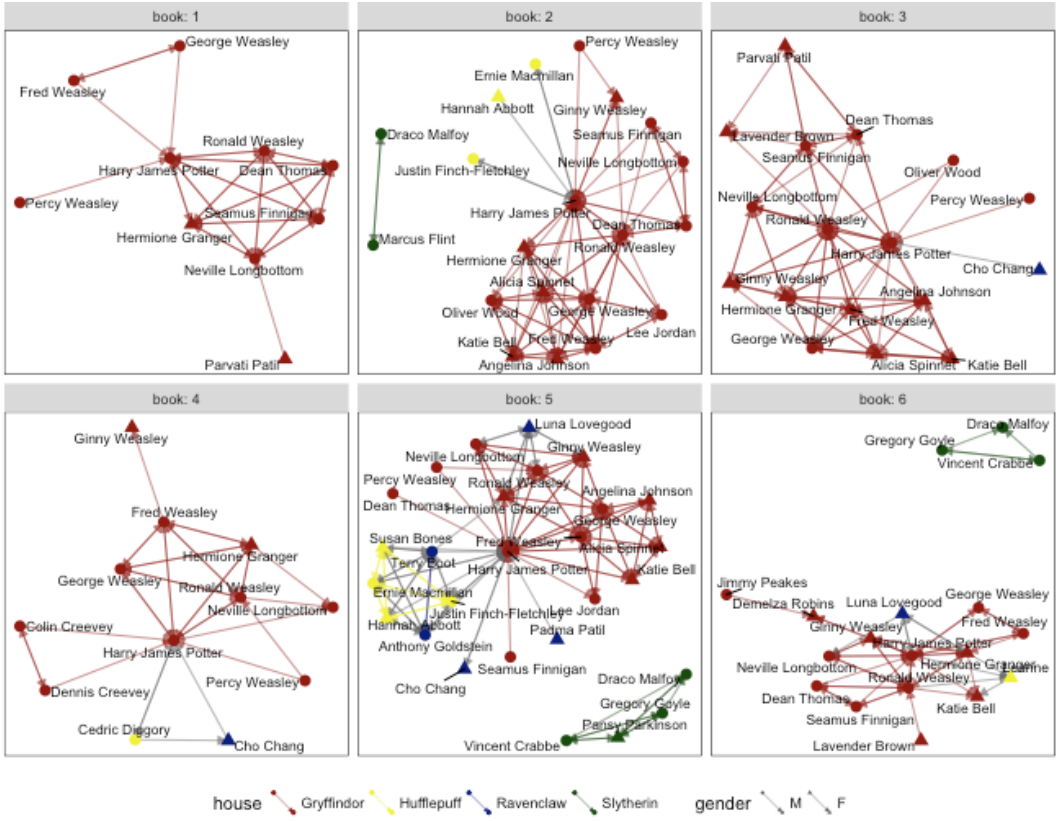
```
glimpse(football)
List of 2
 $ edges   :'data.frame':    613 obs. of  3 variables:
  ..$ from     : chr [1:613] "BrighamYoung" "Iowa" "BrighamYoung" "NewMexico"
  ..$ to       : chr [1:613] "FloridaState" "KansasState" "NewMexico" "TexasT
  ..$ same.conf: num [1:613] 0 0 1 0 1 1 0 1 0 1 ...
 $ vertices:'data.frame':    115 obs. of  2 variables:
  ..$ label: chr [1:115] "BrighamYoung" "FloridaState" "Iowa" "KansasState" .
  ..$ value: chr [1:115] "Mountain West" "Atlantic Coast" "Big Ten" "Big Twel
```


UtahState
CentralFlorida
Navy
NotreDame
Connecticut
Conference
Atlantic Coast
Big East
Big Ten
Big Twelve
Conference USA
Independents
Mid-American
Mountain West
Pacific Ten
Southeastern
Sun Belt
Western Athletic
factor(1 - same.conf)
0
1
NA

Harry Potter characters

There is a connection between two students if one provides emotional support to the other at some point in the book. Code to pull the data together is provided by Sam Tyner here.

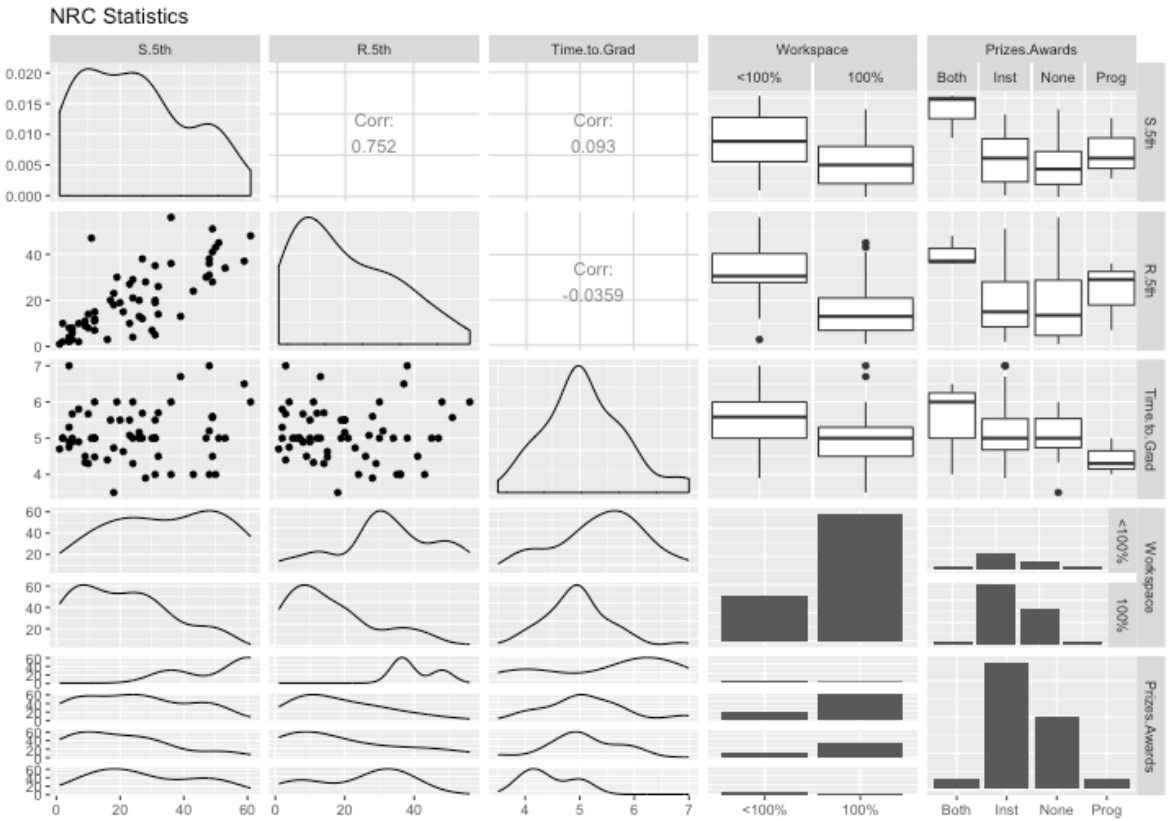
```
load("data/hpchars.rda")
load("data/hpedges.rda")
head(hp.chars)
               name schoolyear gender      house
1      Adrian Pucey       1989      M  Slytherin
2    Alicia Spinnet       1989      F Gryffindor
3  Angelina Johnson       1989      F Gryffindor
4 Anthony Goldstein       1991      M  Ravenclaw
5     Blaise Zabini       1991      M  Slytherin
6     C. Warrington       1989      M  Slytherin
head(hp.edges)
         name1              name2 book
1  Dean Thomas Harry James Potter    1
2  Dean Thomas   Hermione Granger    1
3  Dean Thomas Neville Longbottom    1
4  Dean Thomas     Ronald Weasley    1
5  Dean Thomas    Seamus Finnigan    1
6 Fred Weasley     George Weasley    1
```


book: 1
George Weasley
Fred Weasley
Ronald Weasley
Harry James Potter
Dean Thomas
Percy Weasley
Seamus Finnigan
Hermione Granger
Neville Longbottom
Parvati Patil
book: 2
Percy Weasley
Ernie Macmillan
Hannah Abbott
Ginny Weasley
Draco Malfoy
Seamus Finnigan
Justin Finch-Fletchley
Neville Longbottom
Harry James Potter
Dean Thomas
Marcus Flint
Ronald Weasley
Hermione Granger
Alicia Spinnet
Oliver Wood
George Weasley
Katie Bell
Fred Weasley
Lee Jordan
Angelina Johnson
book: 3
Parvati Patil
Dean Thomas
Lavender Brown
Seamus Finnigan
Oliver Wood
Neville Longbottom
Percy Weasley
Ronald Weasley
Harry James Potter
Cho Chang
Ginny Weasley
Angelina Johnson
Hermione Granger
Fred Weasley
George Weasley
Alicia Spinnet
Katie Bell
book: 4
Ginny Weasley
Fred Weasley
Hermione Granger
George Weasley
Ronald Weasley
Neville Longbottom
Colin Creevey
Harry James Potter
Percy Weasley
Dennis Creevey
Cedric Diggory
Cho Chang
book: 5
Luna Lovegood
Neville Longbottom
Ginny Weasley
Percy Weasley
Ronald Weasley
Angelina Johnson
Dean Thomas
Hermione Granger
George Weasley
Susan Bones
Fred Weasley
Alicia Spinnet
Terry Boot
Harry James Potter
Katie Bell
Ernie Macmillan
Lee Jordan
Justin Finch-Fletchley
Hannah Abbott
Padma Patil
Anthony Goldstein
Seamus Finnigan
Cho Chang
Draco Malfoy
Gregory Goyle
Pansy Parkinson
Vincent Crabbe
book: 6
Draco Malfoy
Gregory Goyle
Vincent Crabbe
Jimmy Peakes
Demelza Robins
Luna Lovegood
George Weasley
Ginny Weasley
Fred Weasley
Harry James Potter
Neville Longbottom
Hermione Granger
Ronald Weasley
Dean Thomas
Katie Bell
Seamus Finnigan
Lavender Brown
house
Gryffindor
Hufflepuff
Ravenclaw
Slytherin
gender
M
F

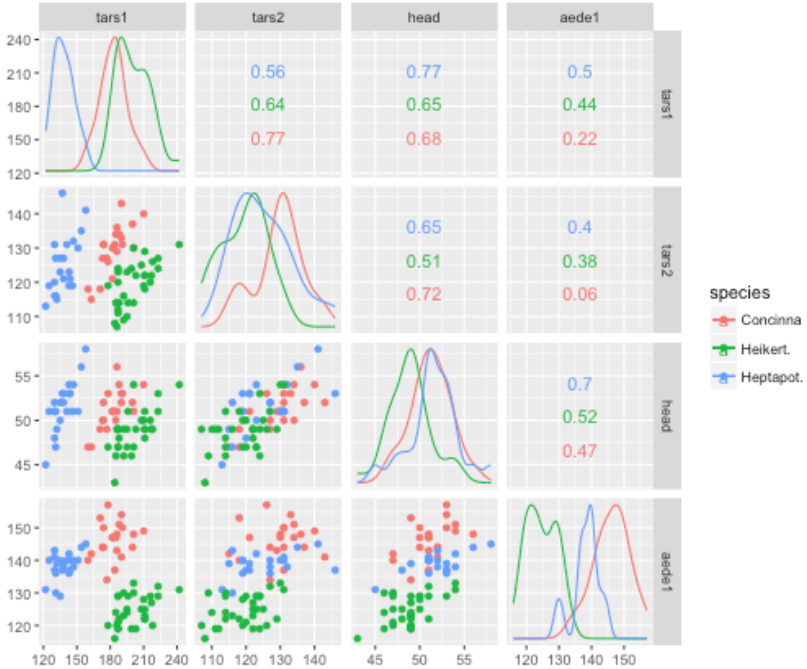
Getting beyond 2D

- Pairs of variables in a matrix layout: pairs plots or scatterplot matrix
- Parallel axes instead of orthogonal axes

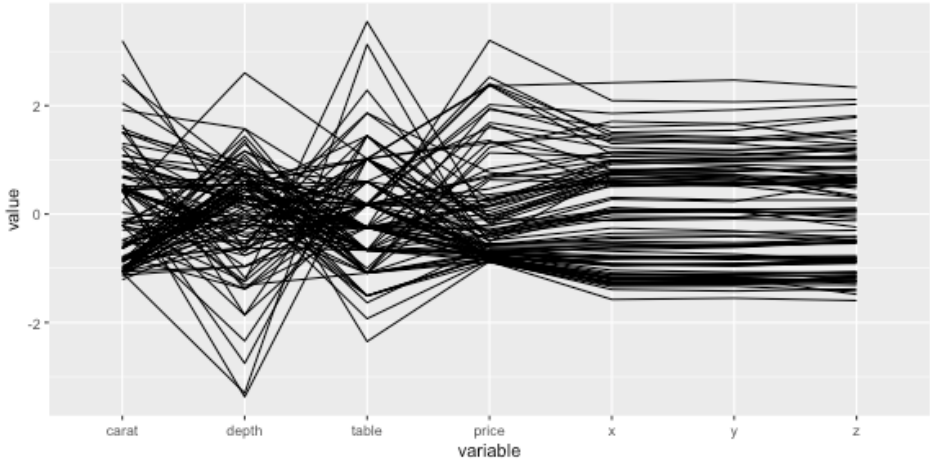
Pairs plots


NRC Statistics
S.5th
R.5th
Time.to.Grad
Workspace
Prizes.Awards
<100%
100%
Both
Inst
None
Prog
Corr:
0.752
Corr:
0.093
Corr:
-0.0359

All variables are numeric



Parallel coordinate plot



Resources

- Tyner, Briatte, Hofmann (2015) Network Visualization with ggplot2
- Pairs plots, parallel coordinate plots, and methods for high-d data

Share and share alike