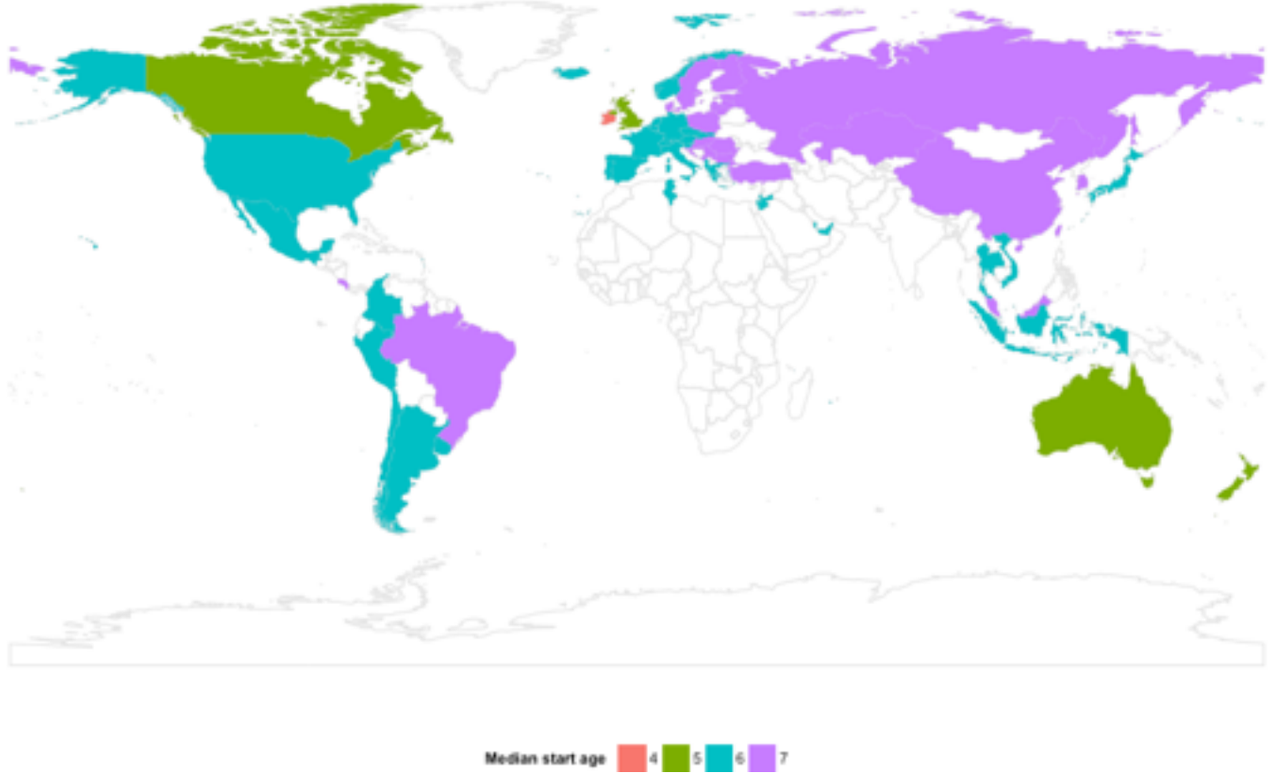
Data Visualization
Discover, Explore and be Skeptical

Di Cook
Statistics, Iowa State University
soon to be Business Analytics, Monash University

Les Diablerets, Feb 1-4, 2015

Seminar 1
Exploratory Data Analysis



Les Diablerets, Feb 1-4, 2015

DATA

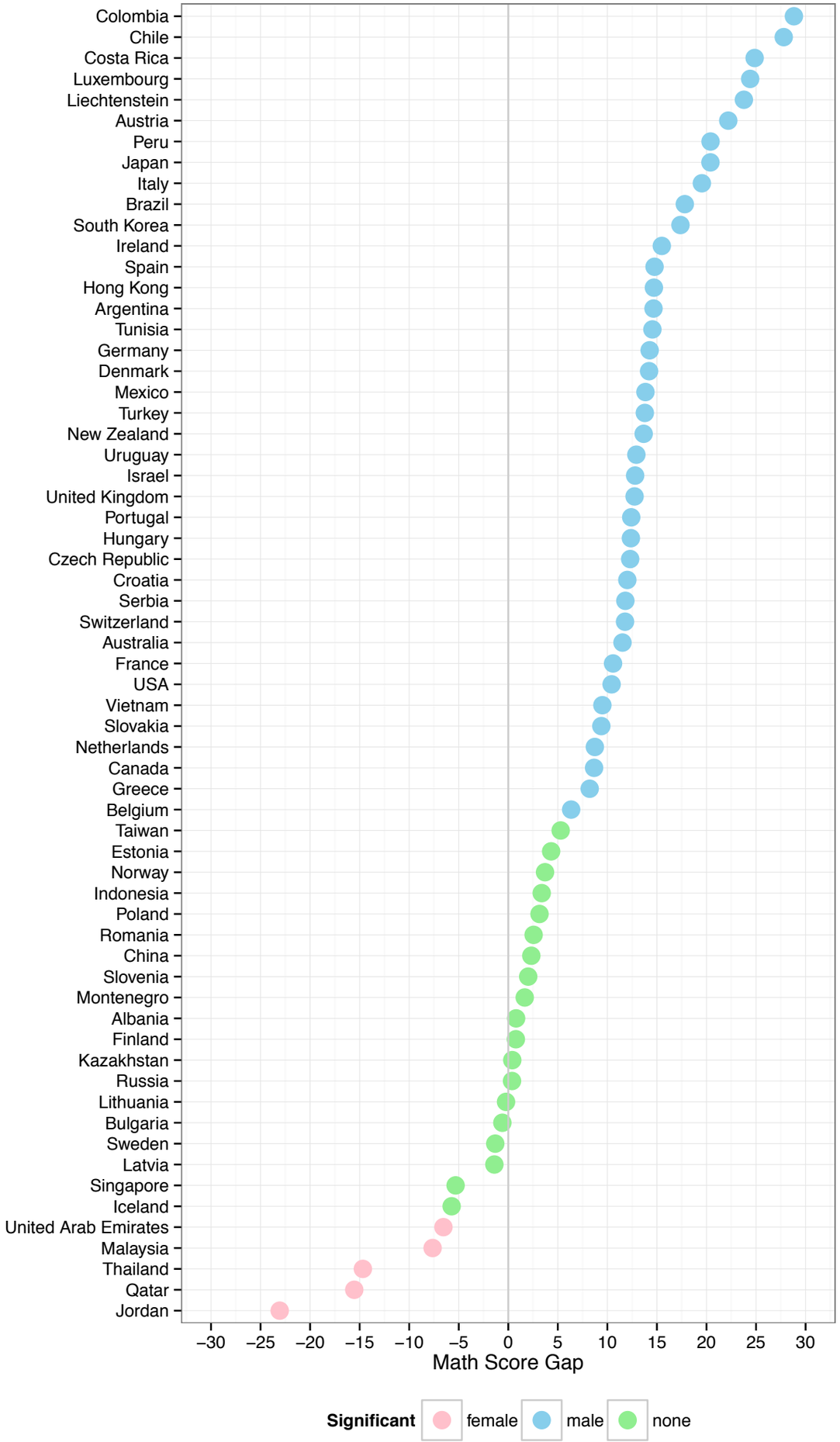
Results from 2012 OECD PISA tests: "the world's global metric for quality, equity and efficiency in school education".

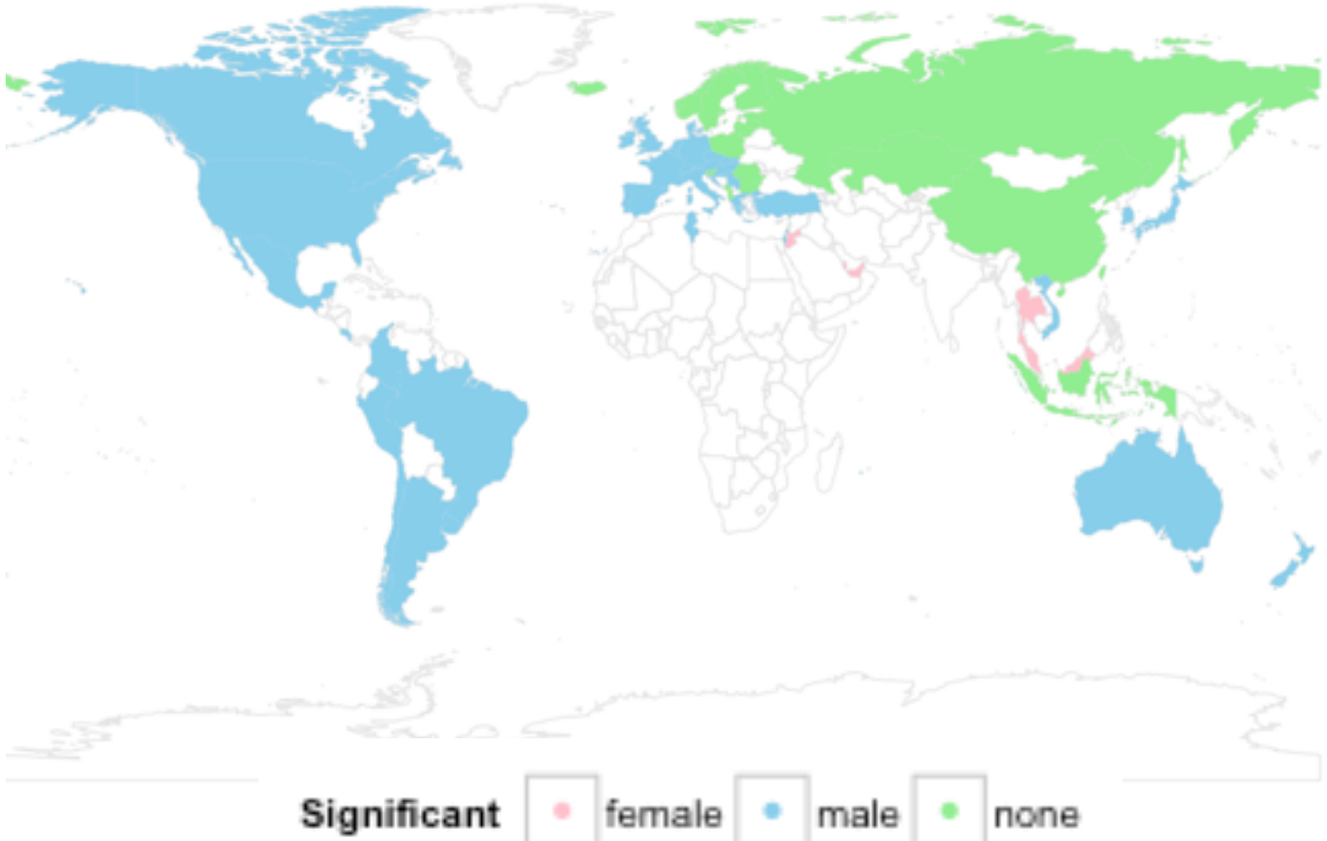
- 485,490 students math, science and reading test scores
- 65 countries, between 100-1500 schools in each
- Student questionnaires about their environment (635 vars)
- Parents surveyed on work, life, income (143 vars)
- Principals provide information about their schools (291 vars)

`http://www.oecd.org/pisa/pisaproducts/datavisualizationcontest.htm`

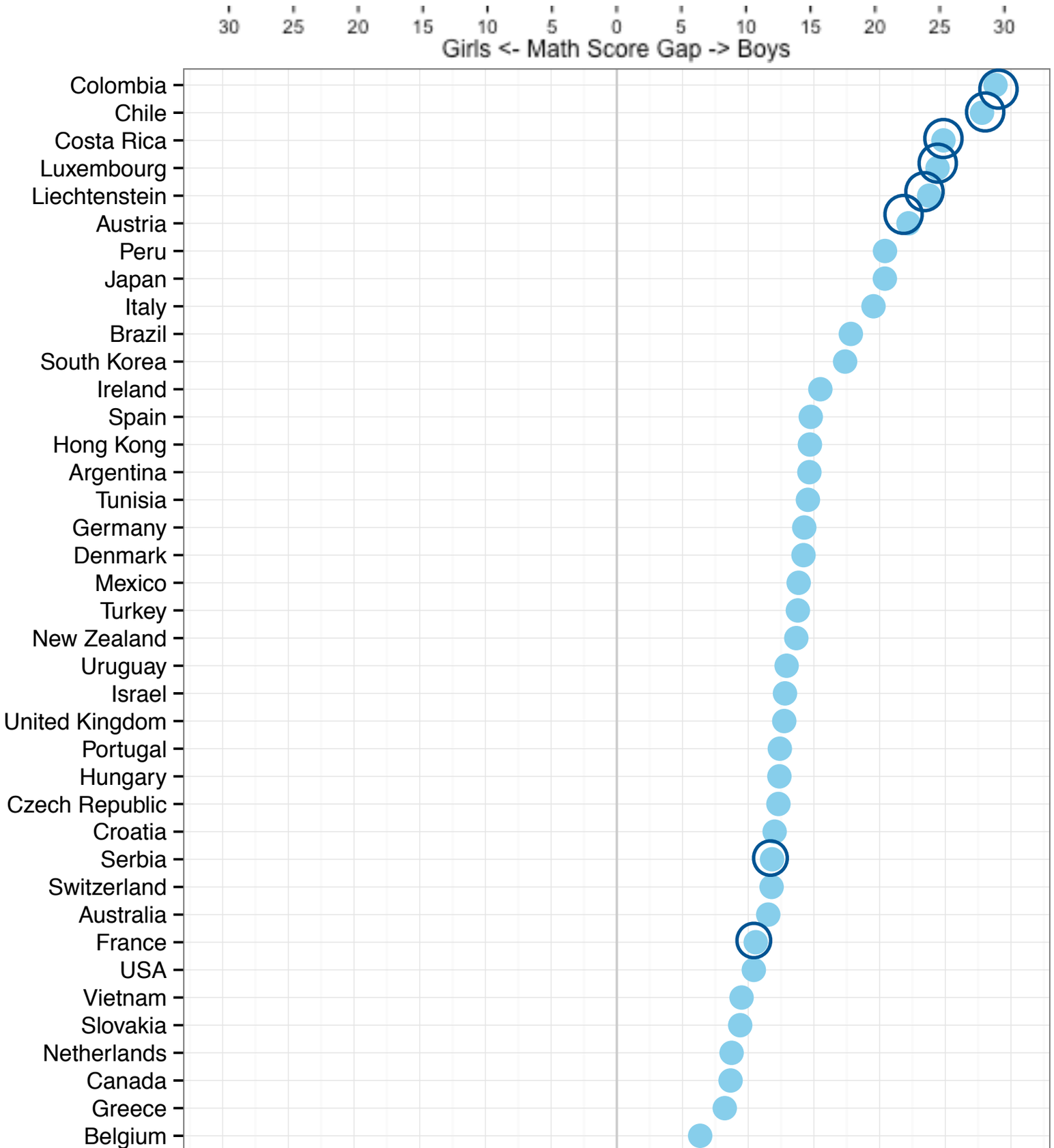
Gender & Math

1. Compute weighted means by country and by gender.
2. Show mean difference by country
3. t-test of difference (unadjusted)





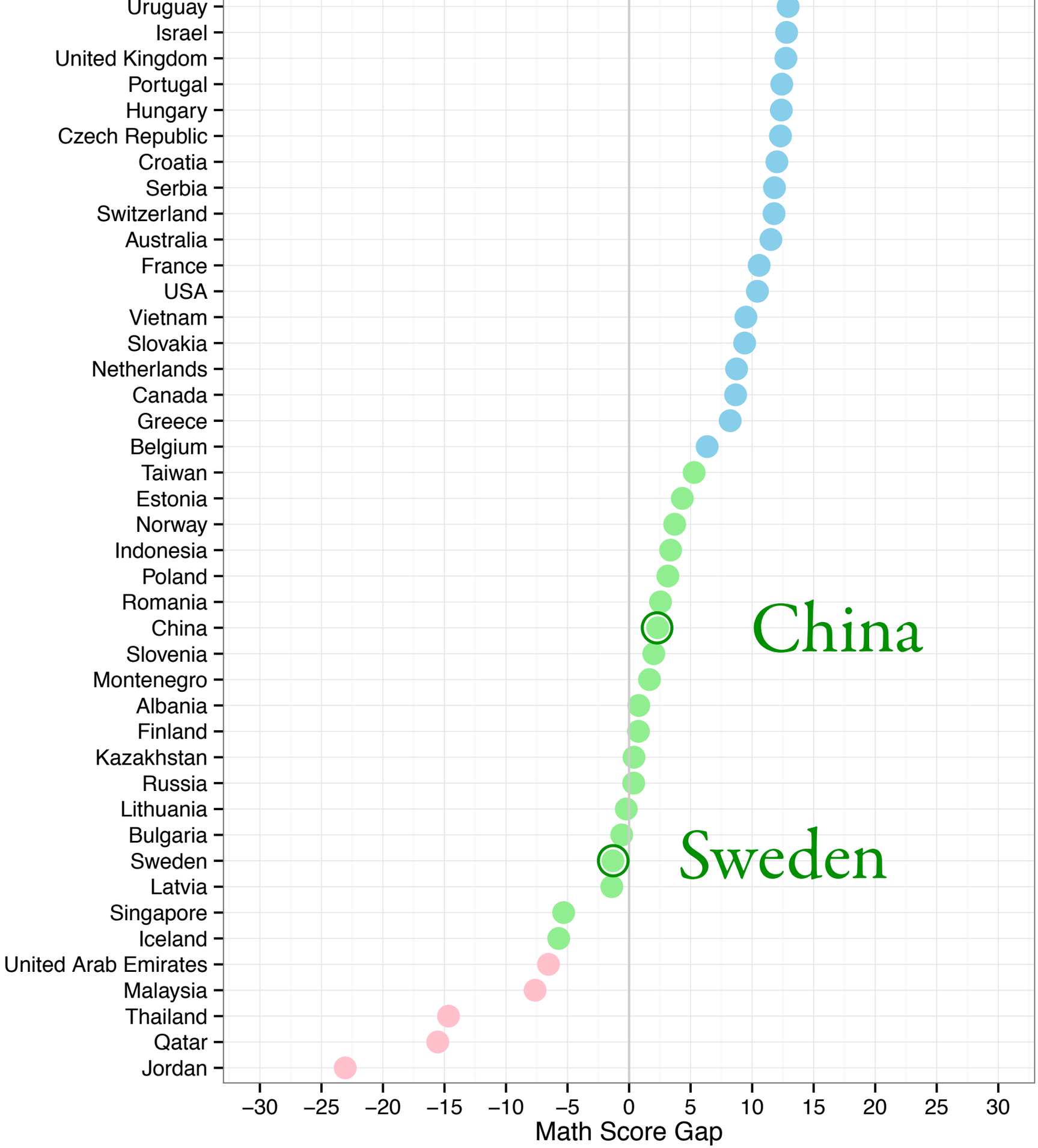
Gender & Math



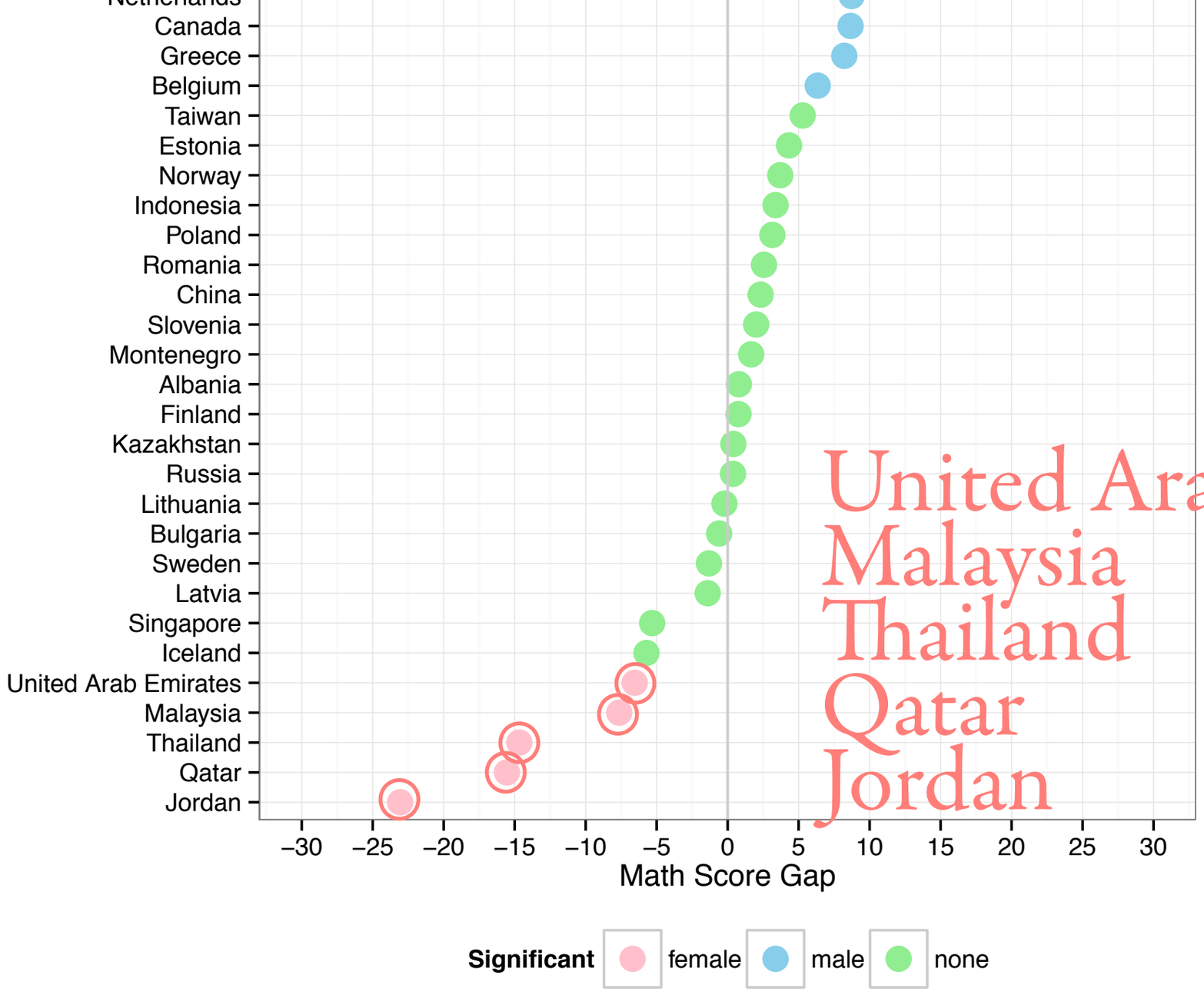
Colombia
Chile
Costa Rica
Luxembourg
Liechtenstein
Austria

Switzerland
USA

Gender & Math

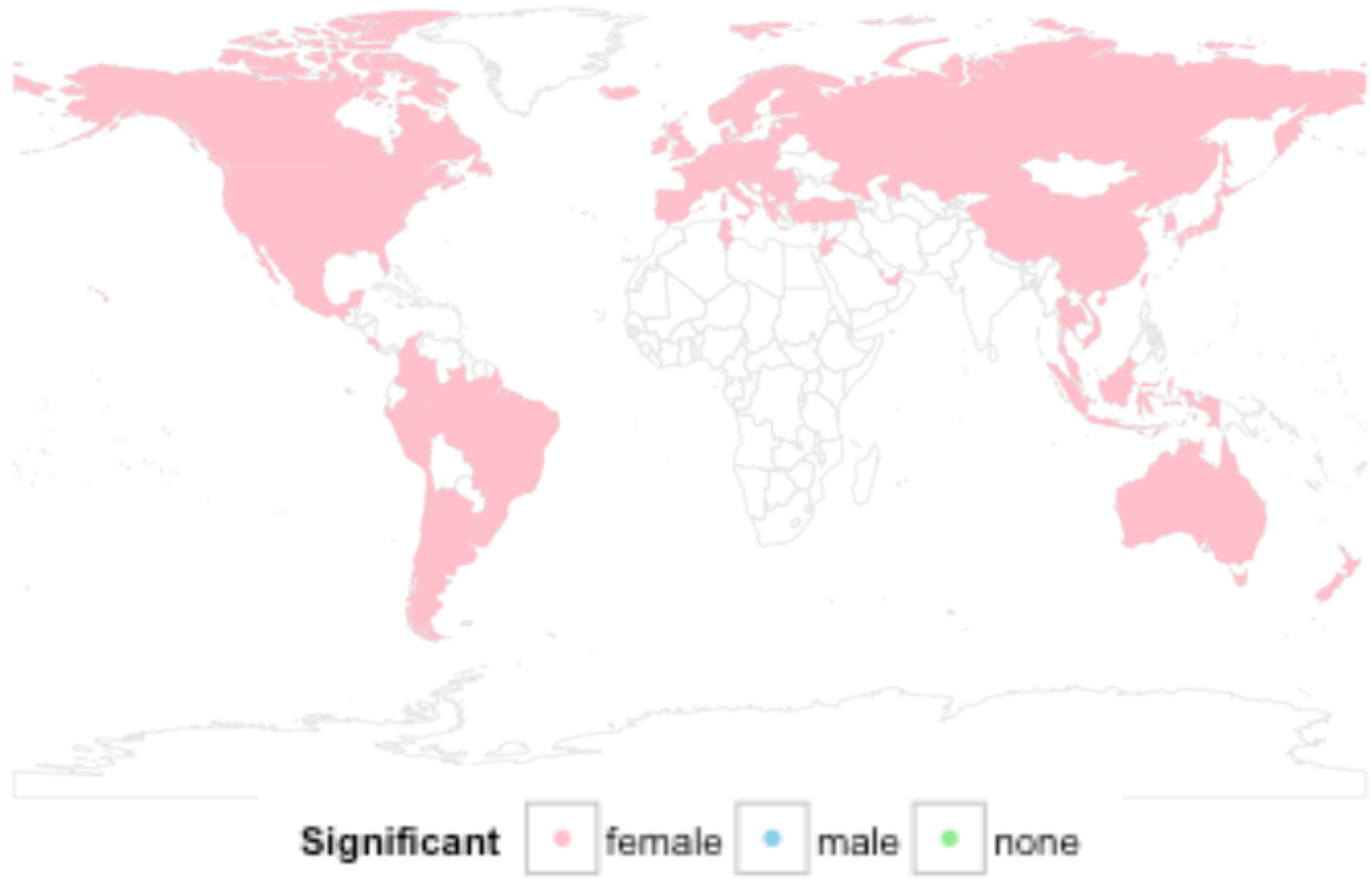

Uruguay
Israel
United Kingdom
Portugal
Hungary
Czech Republic
Croatia
Serbia
Switzerland
Australia
France
USA
Vietnam
Slovakia
Netherlands
Canada
Greece
Belgium
Taiwan
Estonia
Norway
Indonesia
Poland
Romania
China
Slovenia
Montenegro
Albania
Finland
Kazakhstan
Russia
Lithuania
Bulgaria
Sweden
Latvia
Singapore
Iceland
United Arab Emirates
Malaysia
Thailand
Qatar
Jordan
China
Sweden
−30
−25
−20
−15
−10
−5
0
5
10
15
20
25
30
Math Score Gap

Gender & Math



Surprise!

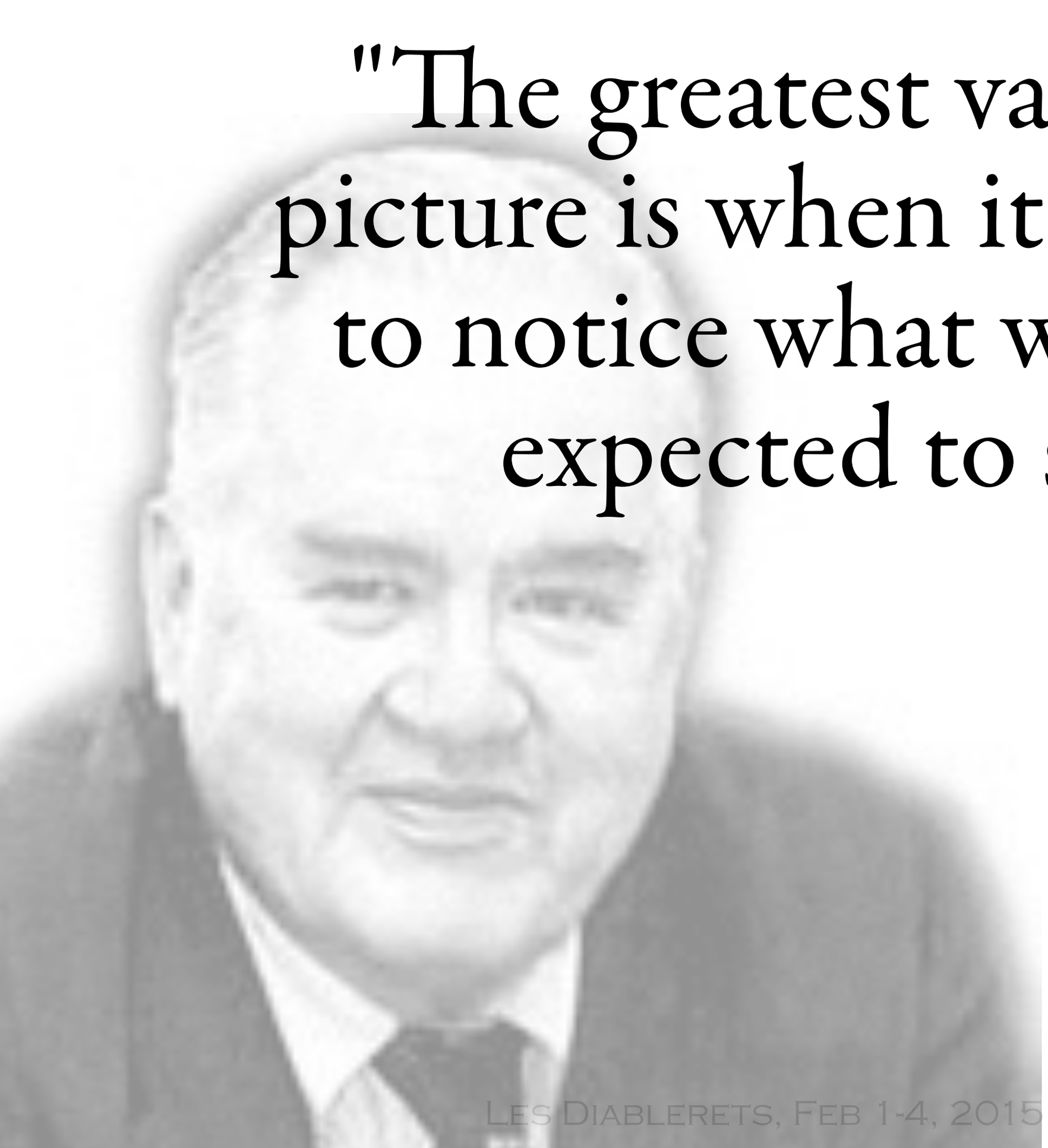
Gender & Reading



Surprised again

In reading scores, the world is PINK!
In EVERY COUNTRY tested GIRLS score significantly BETTER THAN BOYS. Why don't we talk about the gender gap in reading?

"The greatest value of a picture is when it forces us to notice what we never expected to see."

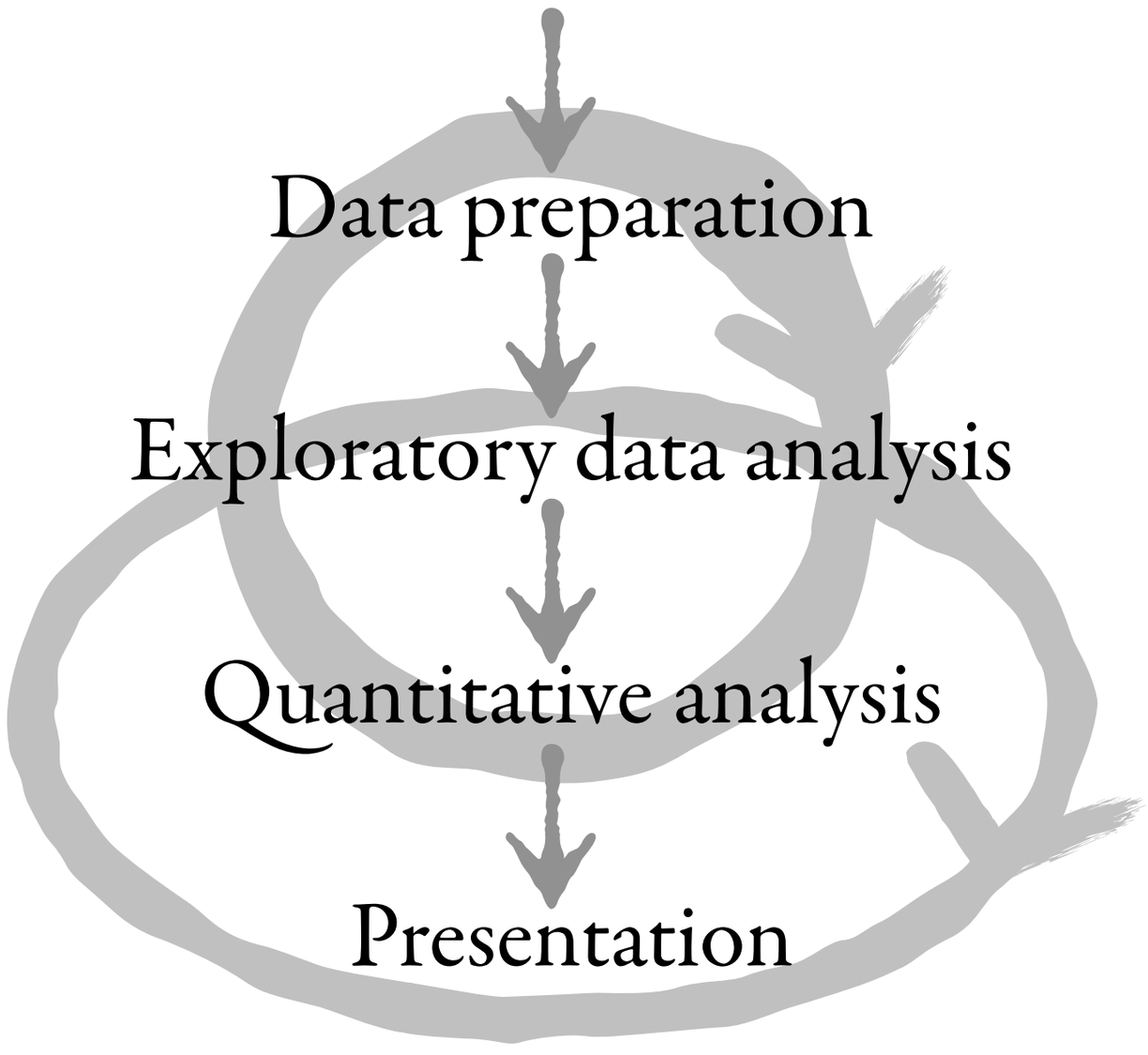


“Show the data."

Examples

- OECD PISA 2012
- CO_2 Emissions and Climate Change
- Sports: tennis

Data analysis



Exploratory data analysis

Relax the focus on the problem statement and explore broadly different aspects of the data.

Good pictures of the data provide unexpected insights, or can confirm expectations

Let the data inform = show the data

Facilitated by ~interactive~ graphics

Time to play in the sand!

EDA vs IDA

- EDA and IDA (Chatfield; Crowder & Hand), although not entirely distinct, differ in emphasis.
- Fundamental to EDA is the desire to let the data inform us, to approach the data without pre-conceived hypotheses, so that we may discover unexpected features. Of course, some of the unexpected features may be errors in the data. IDA emphasizes finding these errors by checking the quality of data prior to formal modeling.
- IDA is much more closely tied to inference than EDA: Problems with the data that violate the assumptions required for valid inference need to be discovered and fixed early.

Data snooping

- Because EDA is very graphical, it sometimes gives rise to a suspicion that patterns in the data are being detected and reported that are not really there.
- This is called data snooping.
- Validation of findings, by all means possible, is an important component of data analysis.

Data snooping

"In our experience, false discovery is the lesser danger when compared to nondiscovery. Nondiscovery is the failure to identify meaningful structure, and it may result in false or incomplete modeling. In a healthy scientific enterprise, the fear of nondiscovery should be at least as great as the fear of false discovery." (Buja's remark on discussion on Koschat & Swayne (1996))

PISA 2012 Process

- Files distributed as an external representation of an R object (`.rda` file), dictionary files=variable coding.
- Read in all files to R, browse size of files, dictionary of items, make some tables/pictures
- Make a list of questions to explore
- Read map data from `rworldmap` package, coordinate names for all countries in OECD data, and merge polygons with subset of variables

Example

```
> head(dict_student2012, 20)
   variable                                                                                description
1       CNT                                                                      Country code 3-character
2  SUBNATIO Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)
3   STRATUM                                 Stratum ID 7-character (cnt + region ID + original stratum ID)
4      OECD                                                                                OECD country
5        NC                                                            National Centre 6-digit Code
6  SCHOOLID                                 School ID 7-digit (region ID + stratum ID + 3-digit school ID)
7   STIDSTD                                                                                  Student ID
8   ST01Q01                                                                      International Grade
9   ST02Q01                                                            National Study Programme
10  ST03Q01                                                                      Birth - Month
11  ST03Q02                                                                       Birth -Year
12  ST04Q01                                                                            Gender
13  ST05Q01                                                                 Attend <ISCED 0>
14  ST06Q01                                                                 Age at <ISCED 1>
15  ST07Q01                                                              Repeat - <ISCED 1>
16  ST07Q02                                                              Repeat - <ISCED 2>
17  ST07Q03                                                              Repeat - <ISCED 3>
18  ST08Q01                                                        Truancy - Late for School
19  ST09Q01                                                   Truancy - Skip whole school day
20 ST115Q01                                          Truancy - Skip classes within school day
…
```

Example

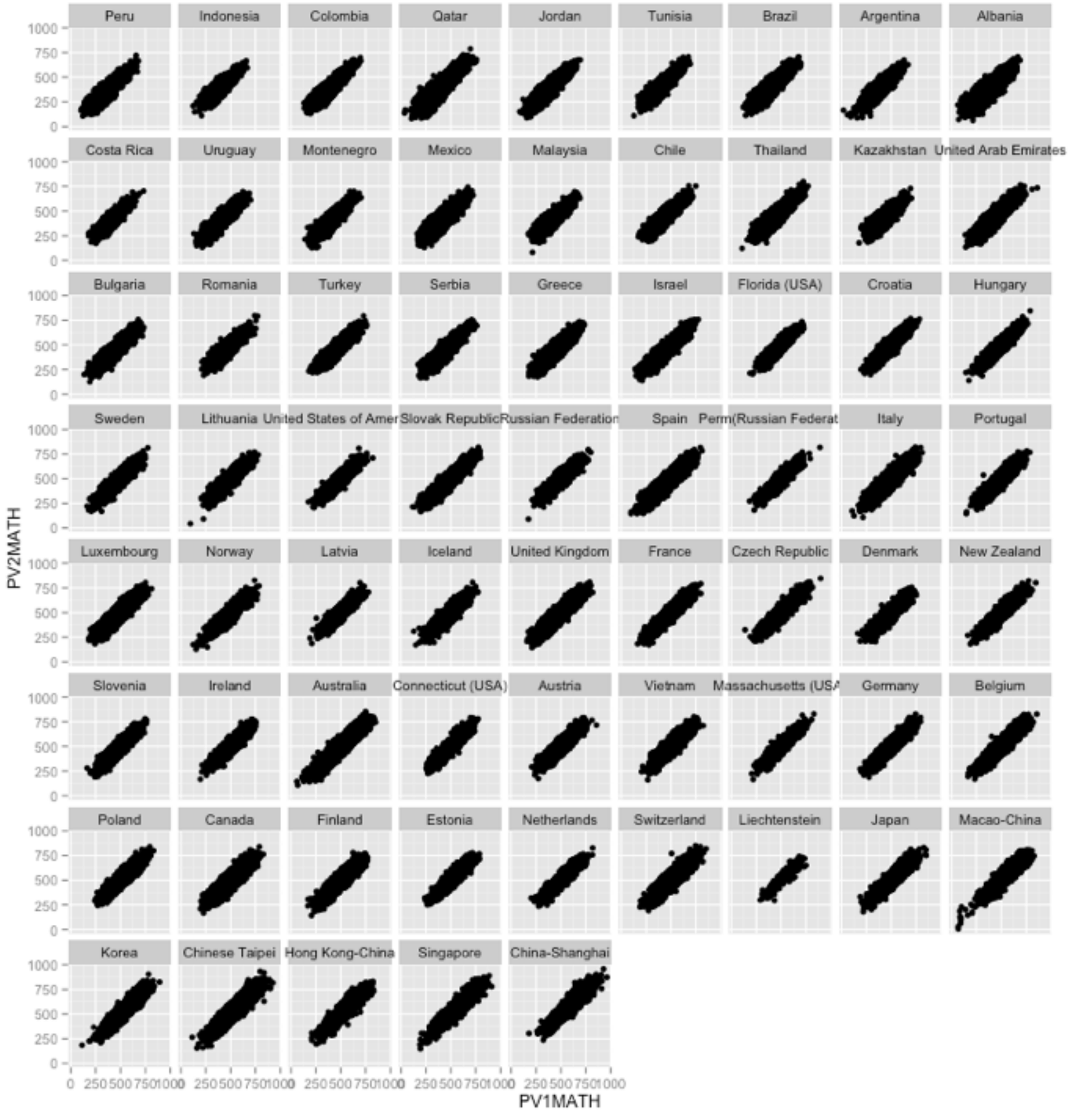
```
> tail(dict_student2012, 5)
     variable                                                  description
631   W_FSTR80                        FINAL STUDENT REPLICATE BRR-FAY WEIGHT80
632   WVARSTRR                        RANDOMIZED FINAL VARIANCE STRATUM (1-80)
633   VAR_UNIT                                RANDOMLY ASSIGNED VARIANCE UNIT
634 SENWGT_STU Senate weight - sum of weight within the country is 1000
635    VER_STU                                   Date of the database creation
```

Example

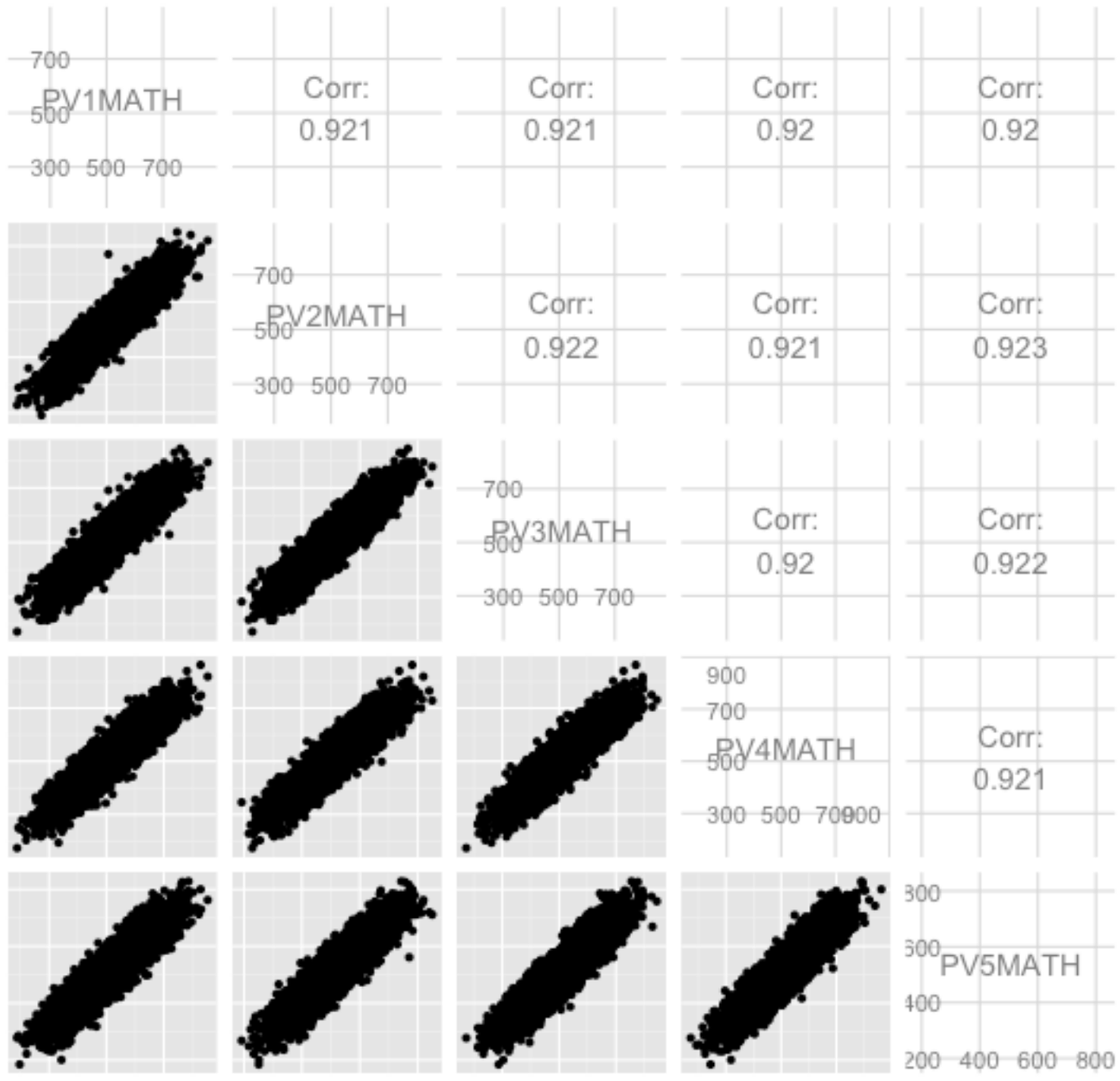
```
> sort(table(student2012$CNT))
```

Liechtenstein	Connecticut (USA)	Massachusetts (USA)	Perm(Russian Federation)
293	1697	1723	1761
Florida (USA)	Iceland	New Zealand	Latvia
1896	3508	4291	4306
Tunisia	Netherlands	Costa Rica	Poland
4407	4460	4602	4607
France	Lithuania	Hong Kong-China	Slovak Republic
4613	4618	4670	4678
...			
Russian Federation	Luxembourg	Bulgaria	Uruguay
5231	5258	5282	5315
Czech Republic	Macao-China	Singapore	Indonesia
5327	5335	5546	5622
Portugal	Kazakhstan	Argentina	Slovenia
5722	5808	5908	5911
Peru	Chinese Taipei	Japan	Thailand
6035	6046	6351	6606
Chile	Jordan	Denmark	Belgium
6856	7038	7481	8597
Finland	Colombia	Qatar	Switzerland
8829	9073	10966	11229
United Arab Emirates	United Kingdom	Australia	Brazil
11500	12659	14481	19204
Canada	Spain	Italy	Mexico
21544	25313	31073	33806

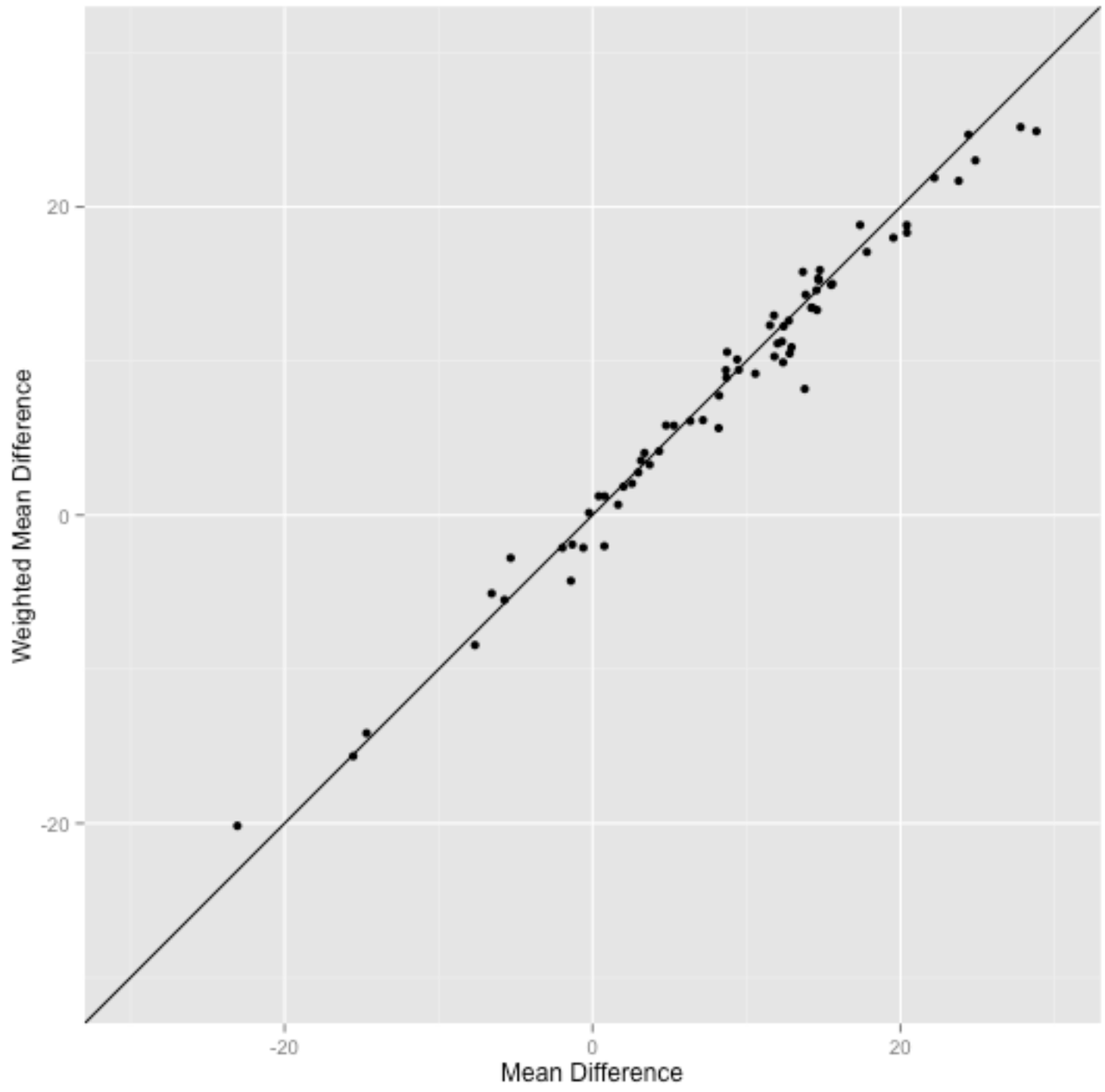
Compare plausible values



Compare plausible values for Switzerland



Do sample weights matter?



Questions

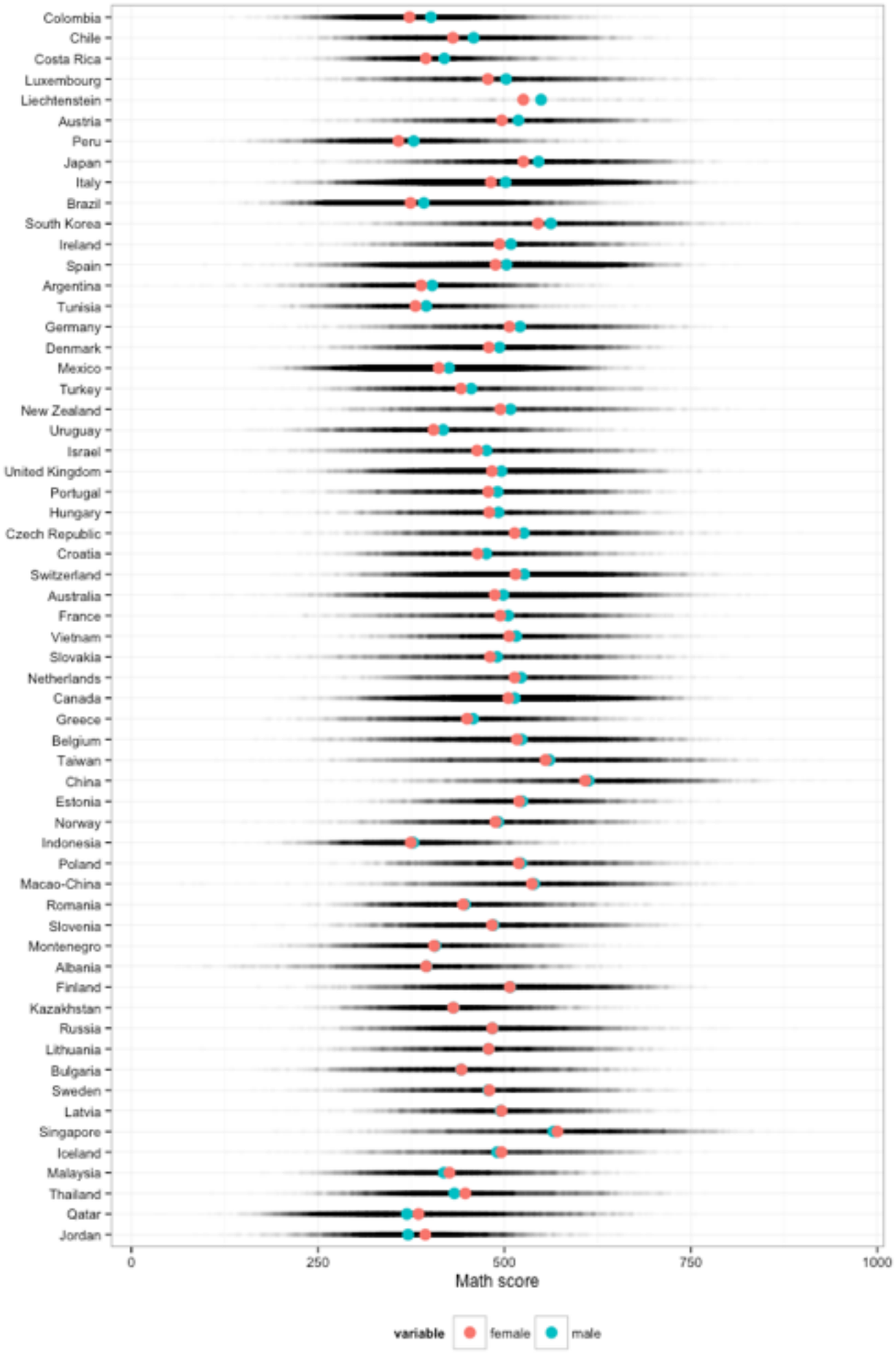
- Is the gender gap evident in the math scores?
- Do students work hard outside school hours? And does this pay off in better scores?
- Does truancy affect performance?
- Do more household possessions mean better performance?
- Do parents matter?

...

Expectations

- Gender gap: “boys are better at math”
- USA doesn’t perform well compared to other countries.
- I would expect parents, time studying and socioeconomic status to be important influences on performance.

Look at the individuals, not just summary statistics

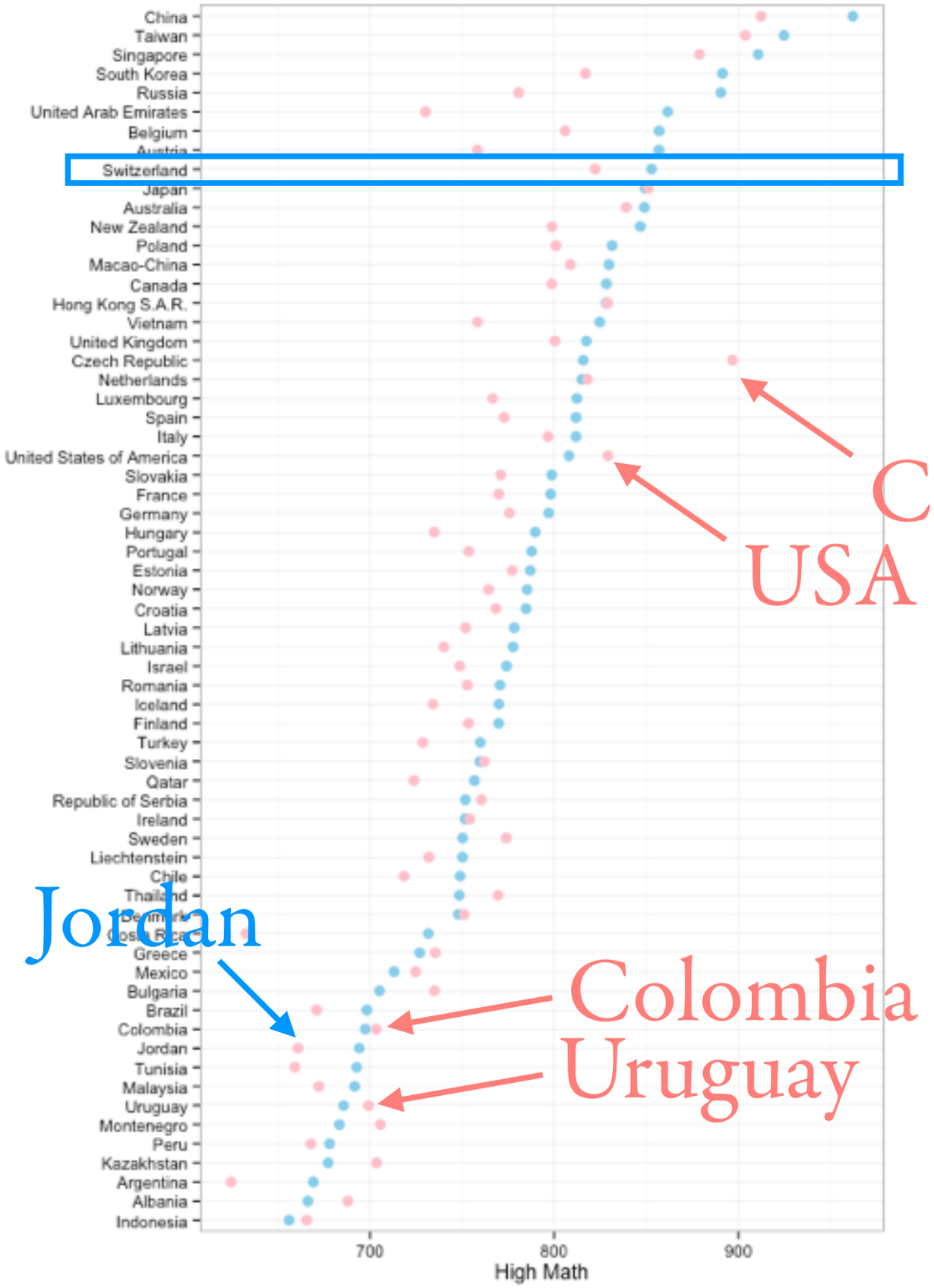


Individuals

- Test scores range from 0-1000.
- Gaps for math were at most 30 points.
- For reading at most 80 points.
- *Differences are tiny.*

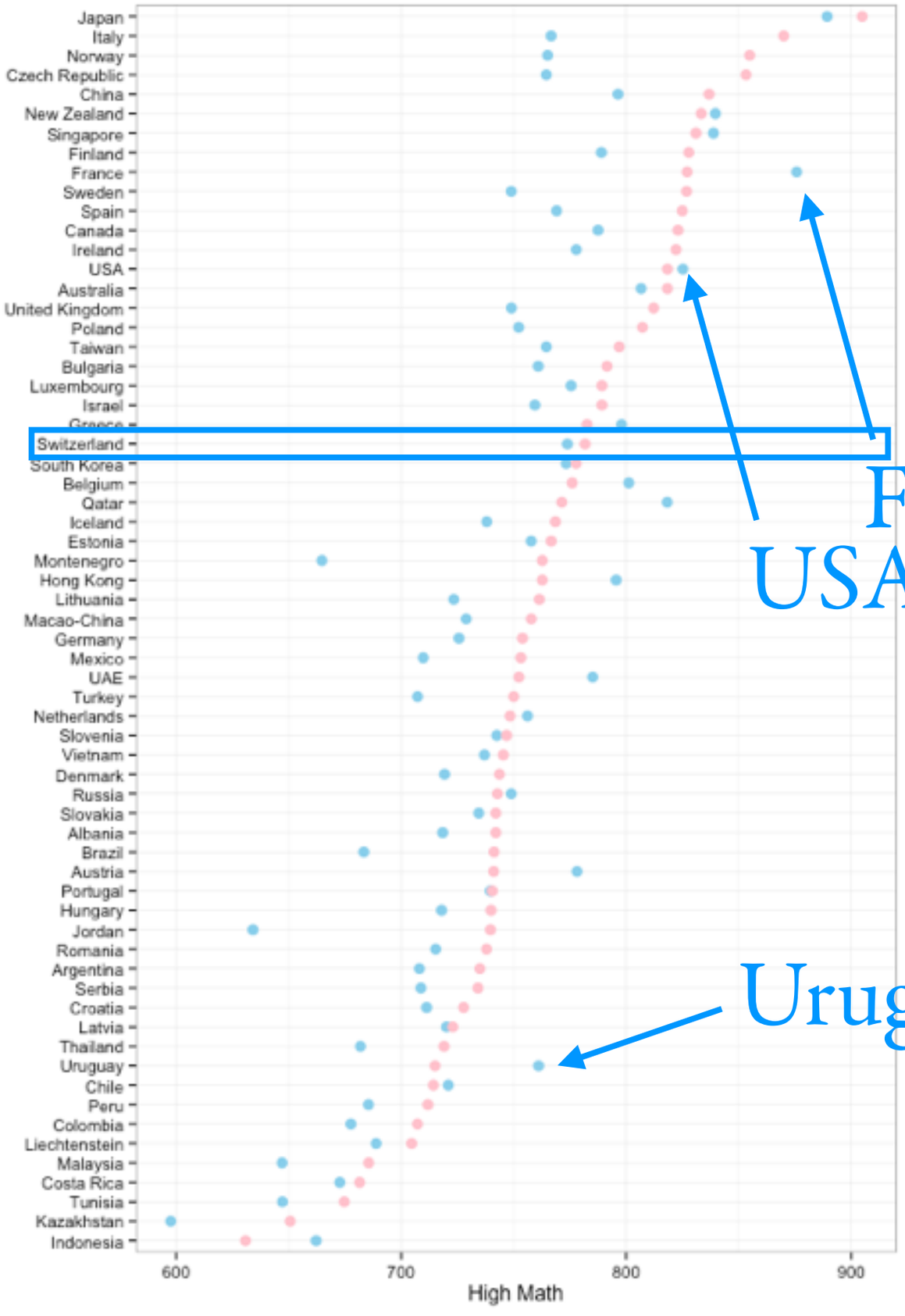
Individuals

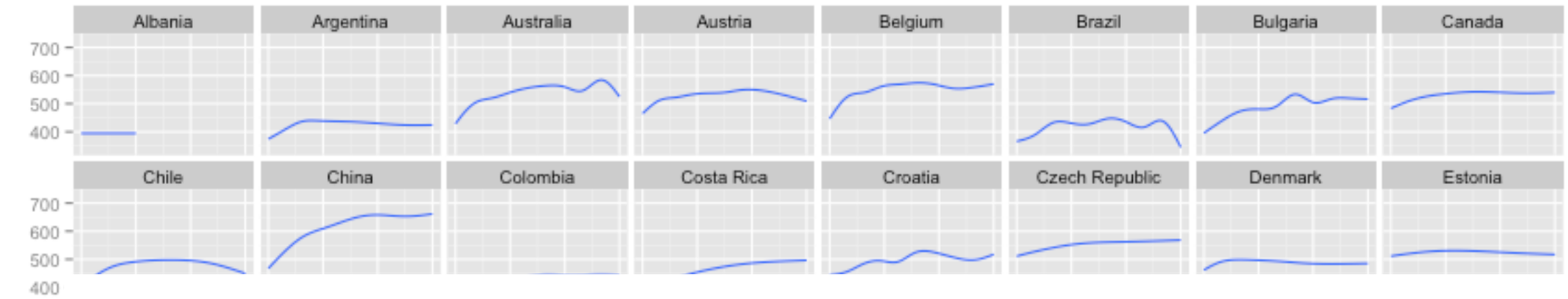
- Even countries with a math gender gap, sometimes have a girl who scored the top mark.



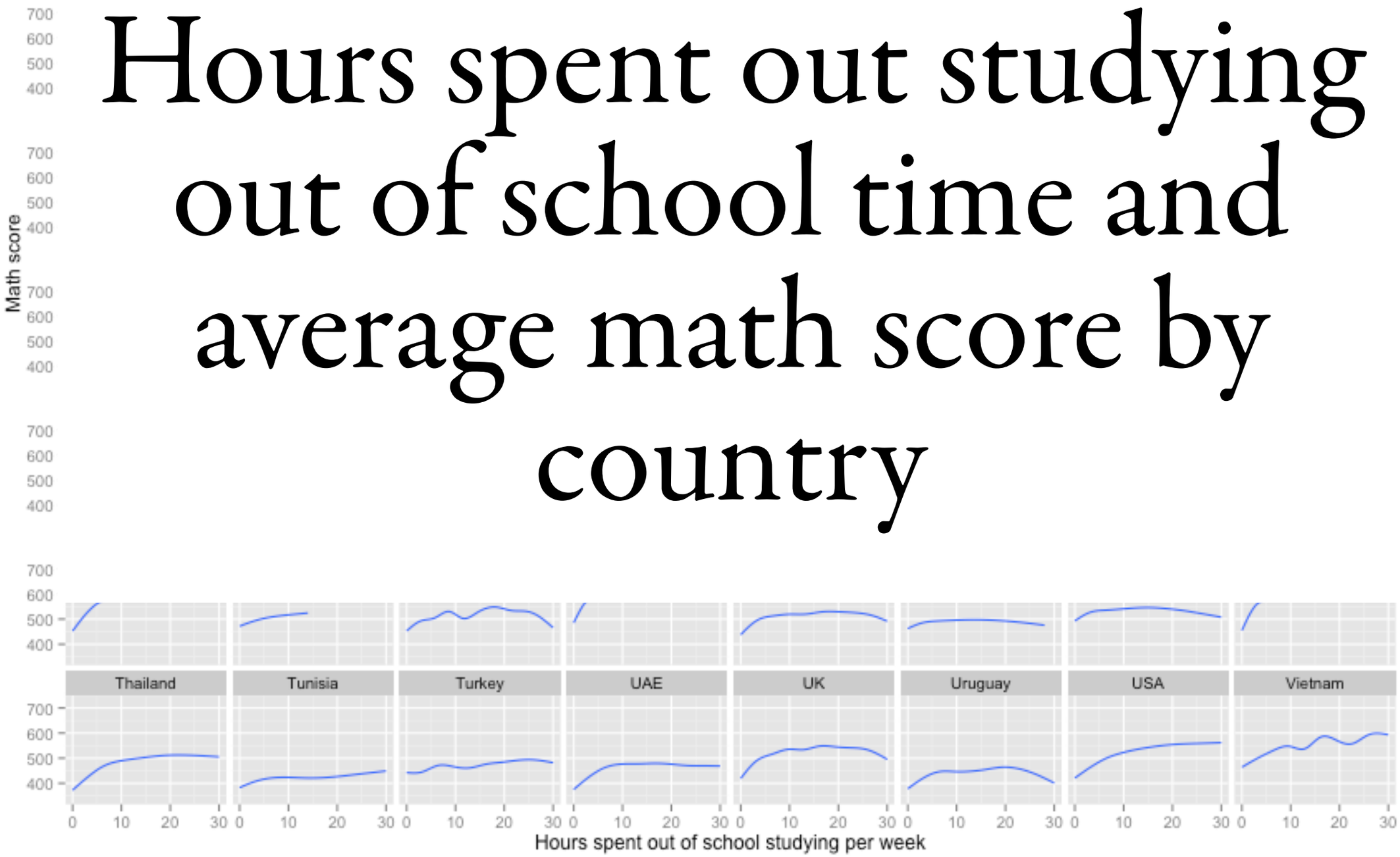
Individuals

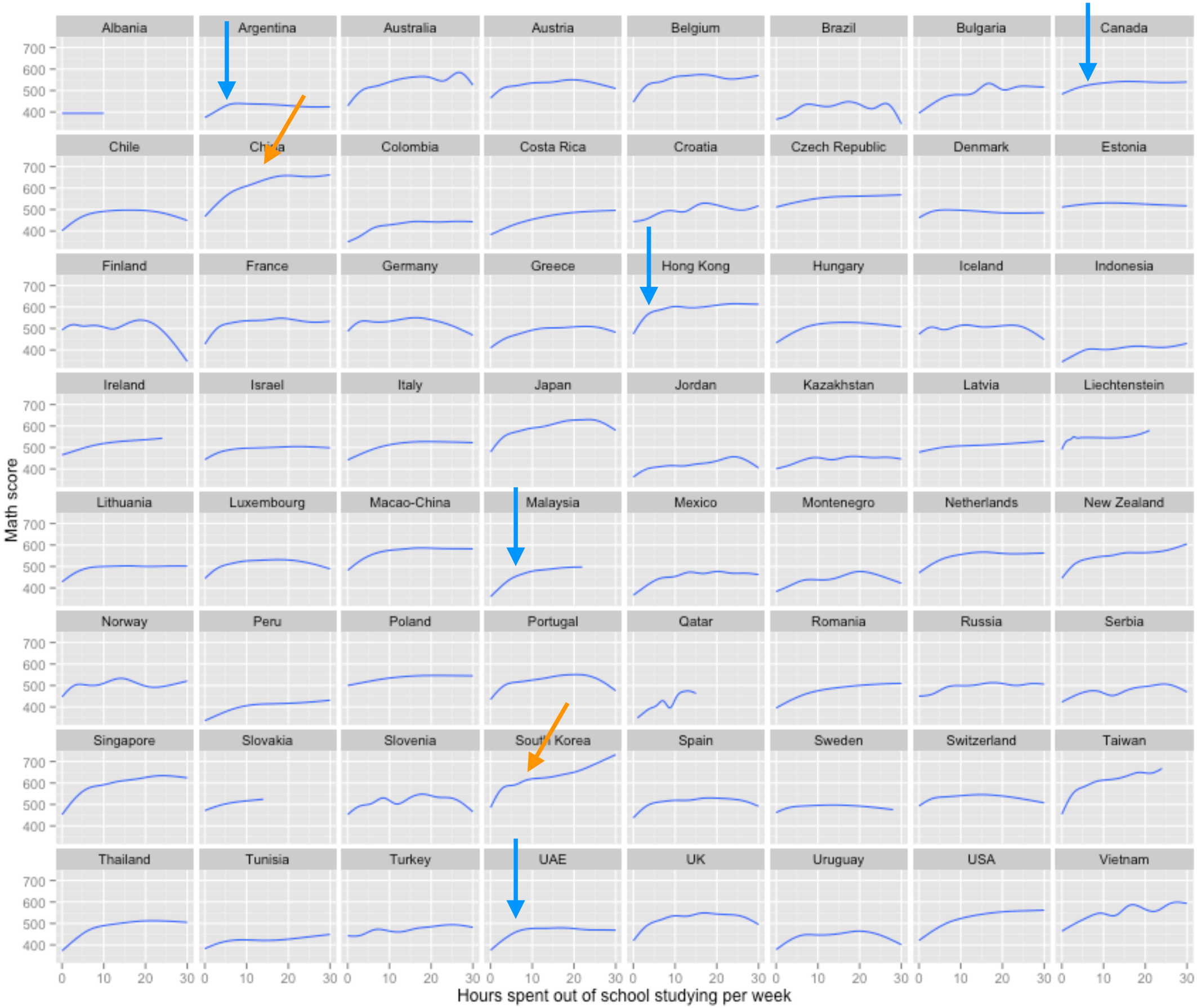
- And although there is a gender gap in reading in all countries, individually some top reading scores are obtained by boys.

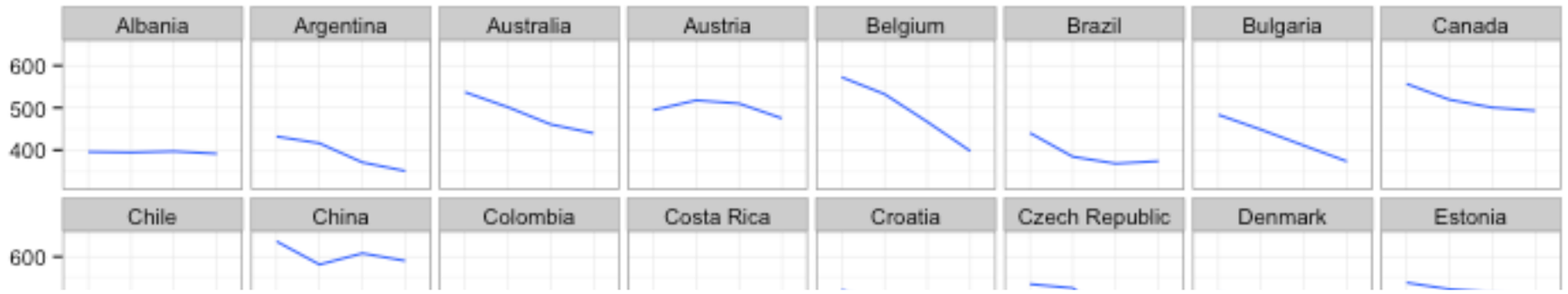




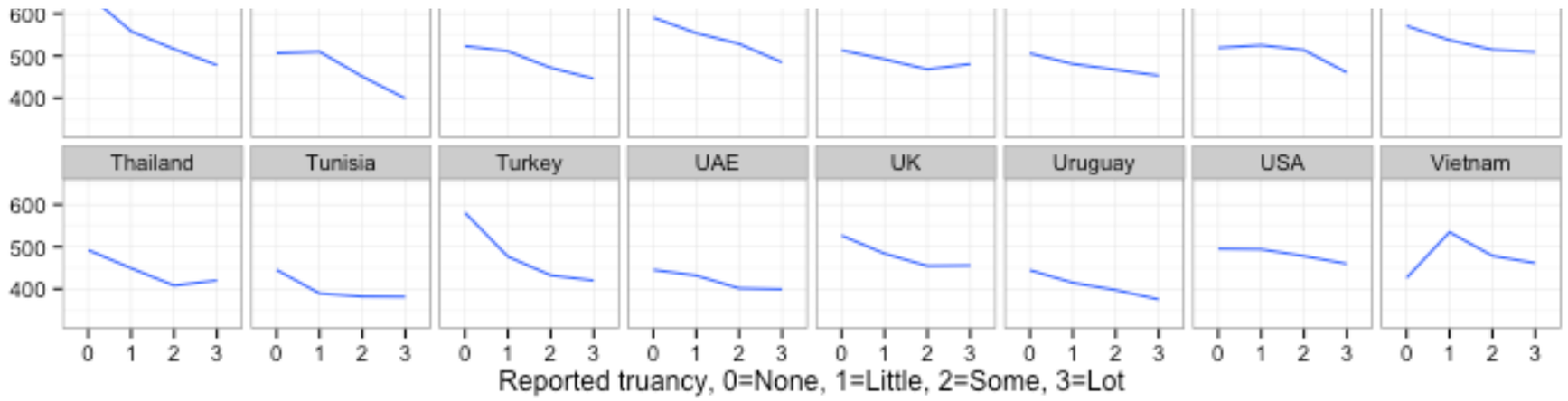
Hours spent out studying out of school time and average math score by country

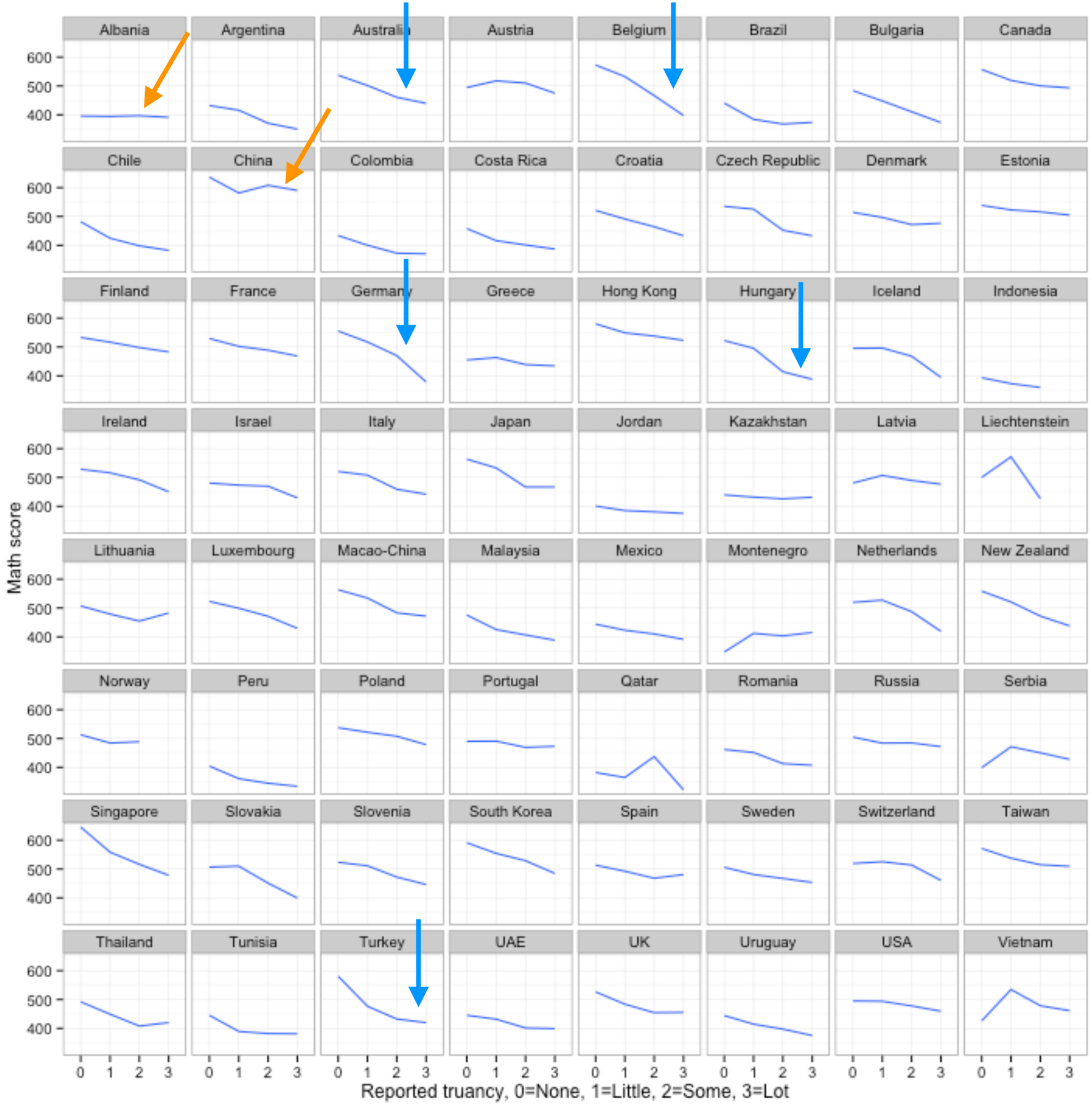


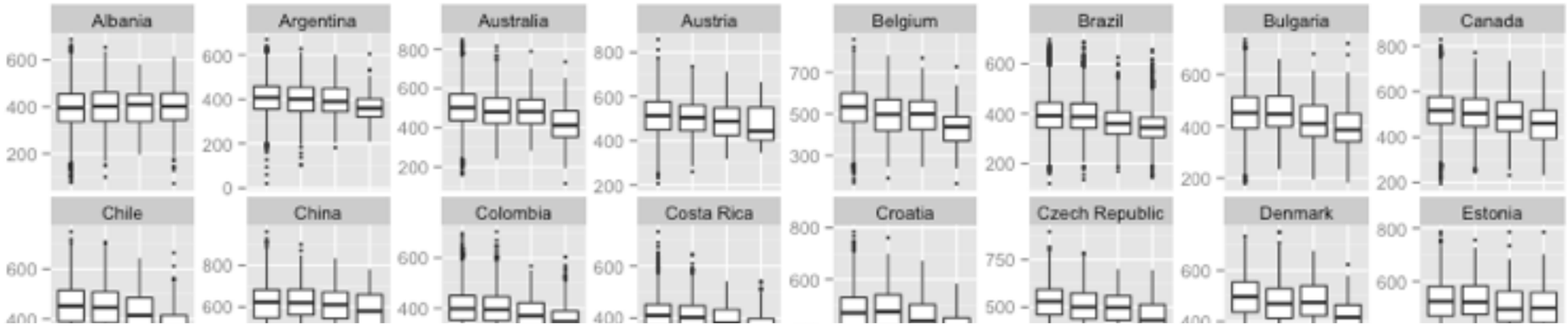

Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia
Costa Rica
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hong Kong
Hungary
Iceland
Indonesia
Ireland
Israel
Italy
Japan
Jordan
Kazakhstan
Latvia
Liechtenstein
Lithuania
Luxembourg
Macao-China
Malaysia
Mexico
Montenegro
Netherlands
New Zealand
Norway
Peru
Poland
Portugal
Qatar
Romania
Russia
Serbia
Singapore
Slovakia
Slovenia
South Korea
Spain
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
UAE
UK
Uruguay
USA
Vietnam
Math score
400
500
600
700
0
10
20
30
Hours spent out of school studying per week



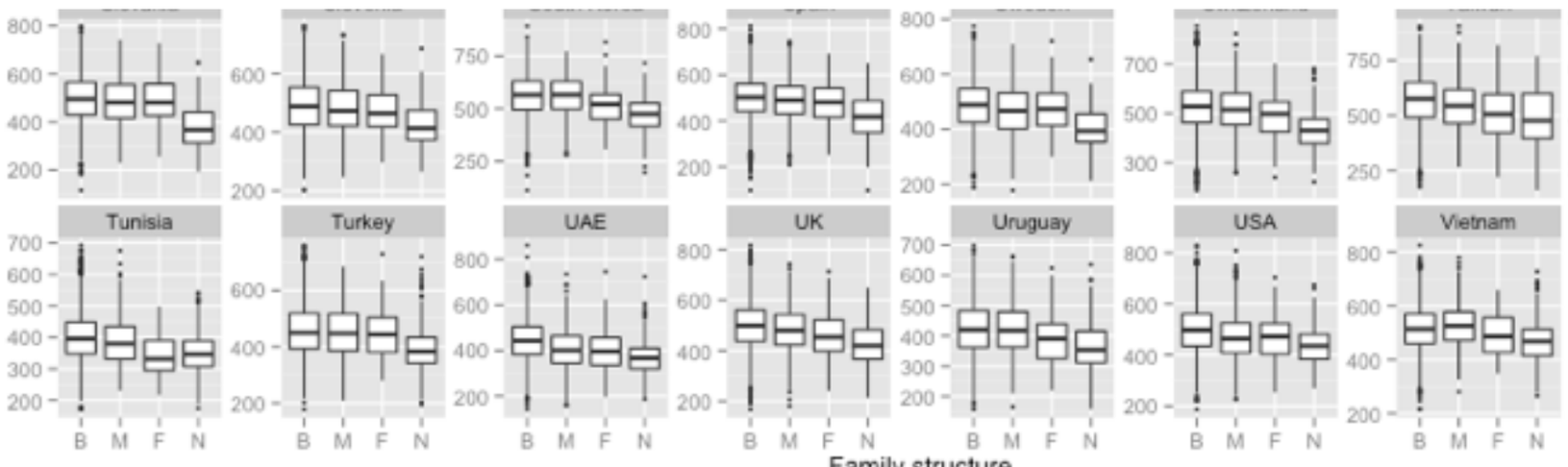
Reported truancy and average math score by country

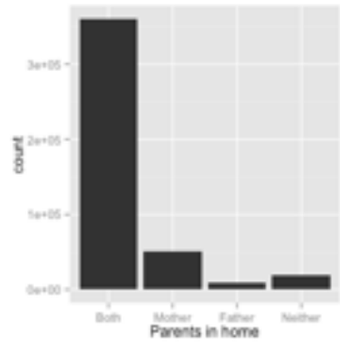


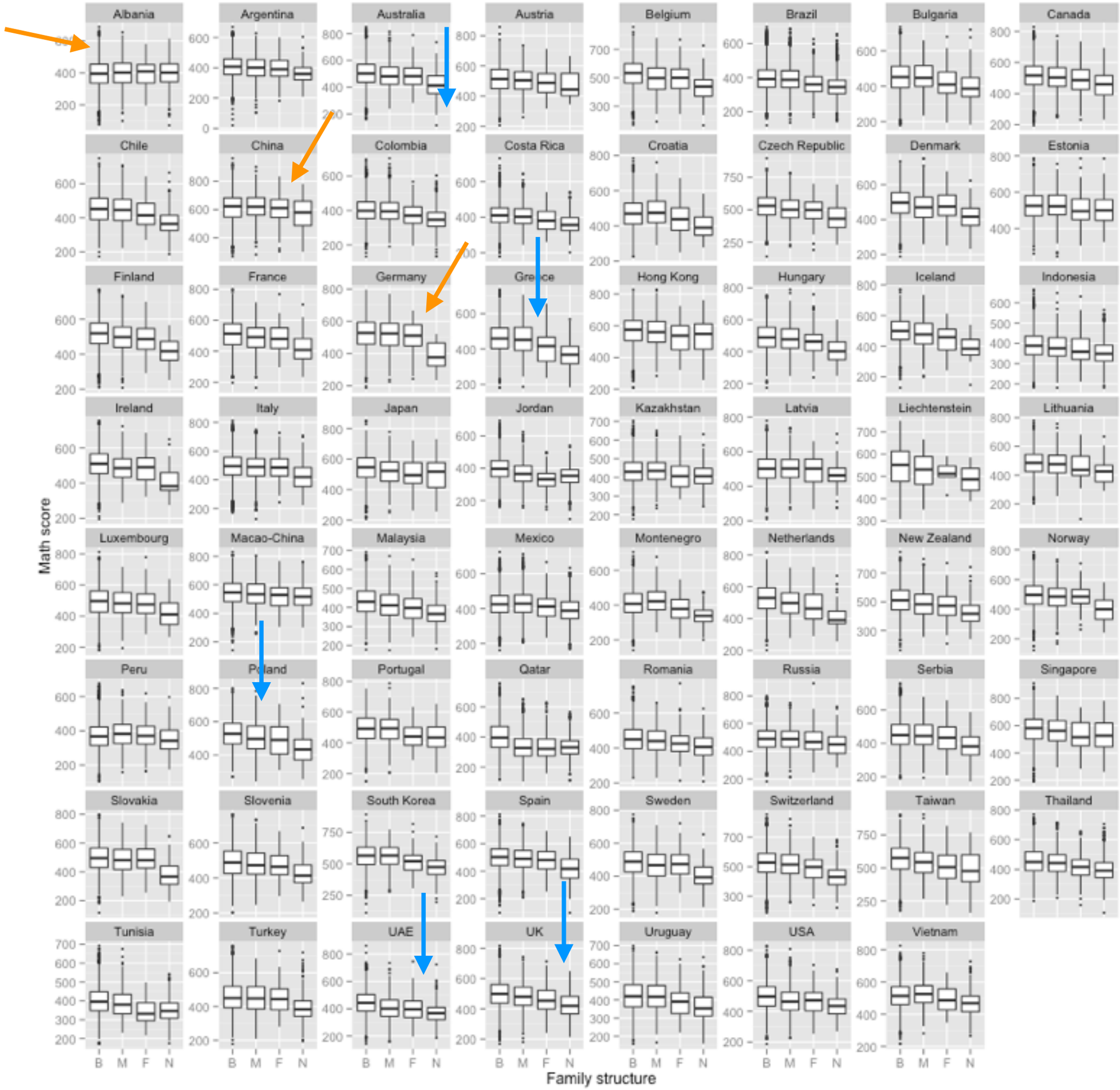

Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia
Costa Rica
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hong Kong
Hungary
Iceland
Indonesia
Ireland
Israel
Italy
Japan
Jordan
Kazakhstan
Latvia
Liechtenstein
Lithuania
Luxembourg
Macao-China
Malaysia
Mexico
Montenegro
Netherlands
New Zealand
Norway
Peru
Poland
Portugal
Qatar
Romania
Russia
Serbia
Singapore
Slovakia
Slovenia
South Korea
Spain
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
UAE
UK
Uruguay
USA
Vietnam
600
500
400
0 1 2 3
Reported truancy, 0=None, 1=Little, 2=Some, 3=Lot

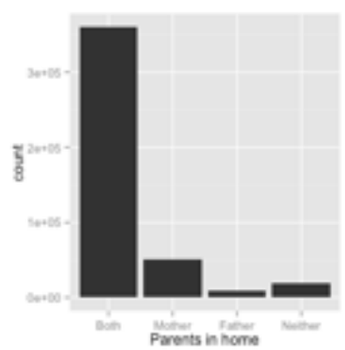


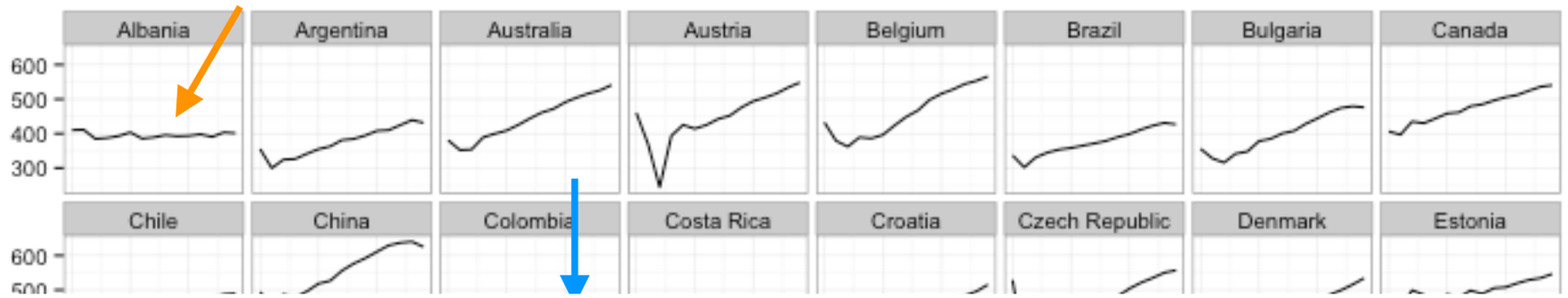
Parents in the home and average math score by country



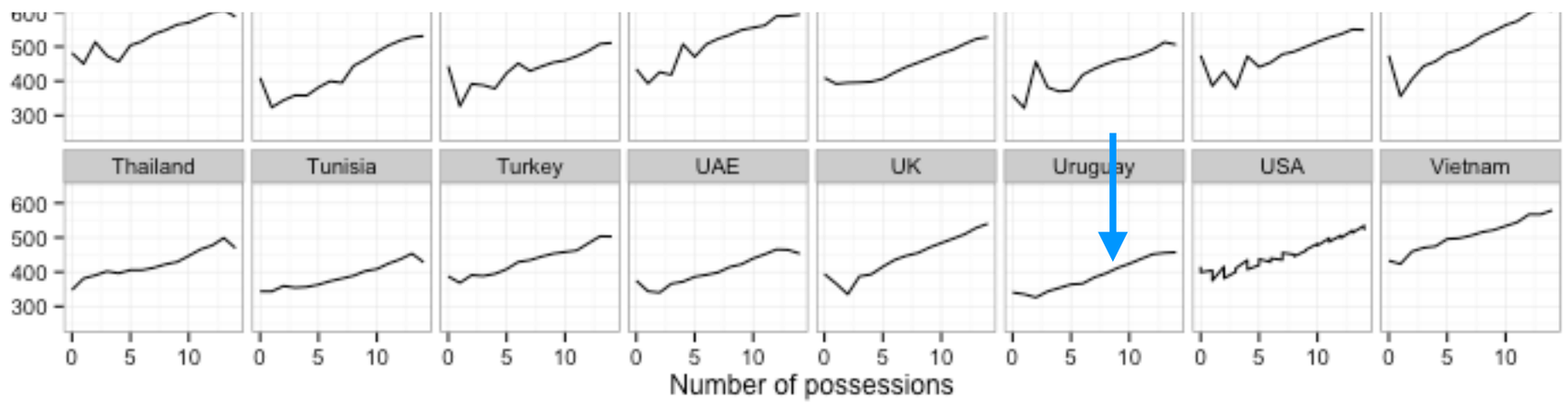


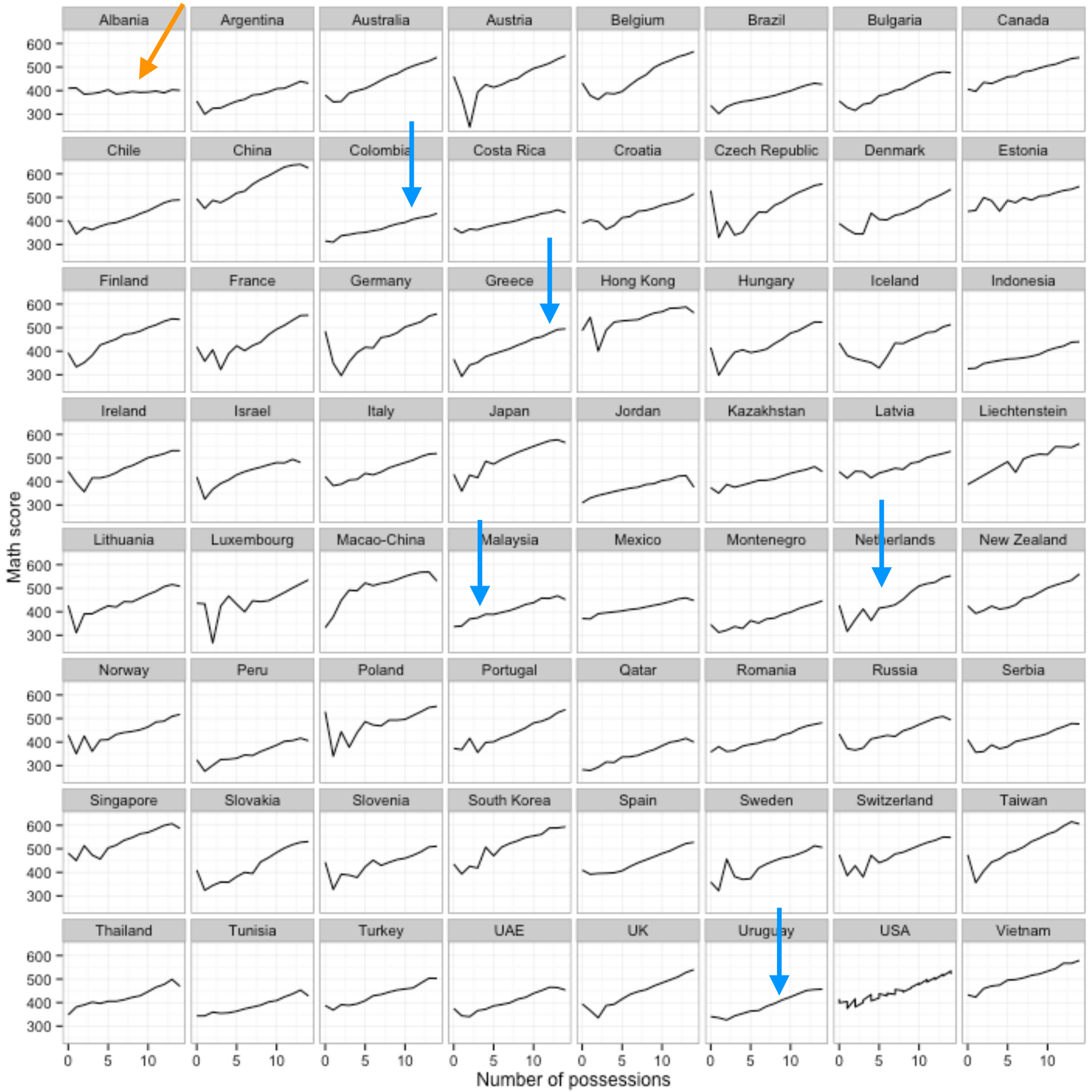

Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia
Costa Rica
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hong Kong
Hungary
Iceland
Indonesia
Ireland
Italy
Japan
Jordan
Kazakhstan
Latvia
Liechtenstein
Lithuania
Luxembourg
Macao-China
Malaysia
Mexico
Montenegro
Netherlands
New Zealand
Norway
Peru
Poland
Portugal
Qatar
Romania
Russia
Serbia
Singapore
Slovakia
Slovenia
South Korea
Spain
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
UAE
UK
Uruguay
USA
Vietnam
Math score
B M F N
Family structure

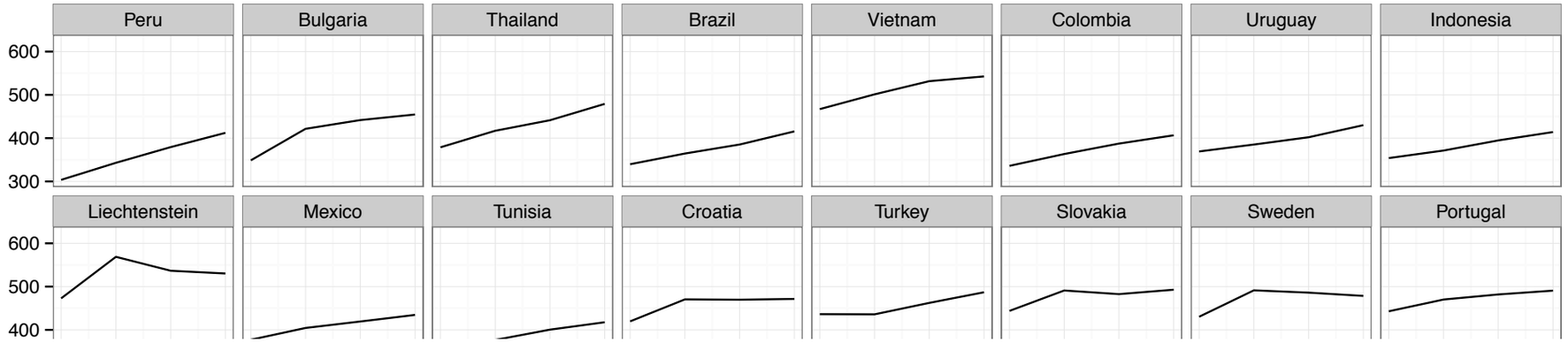

count
Both
Mother
Father
Neither
Parents in home



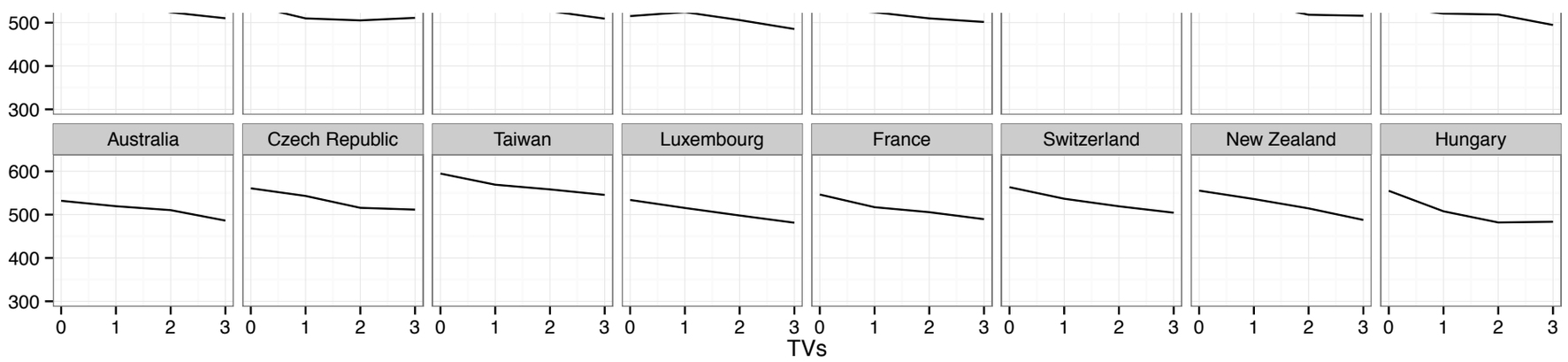
Possessions and average math score by country

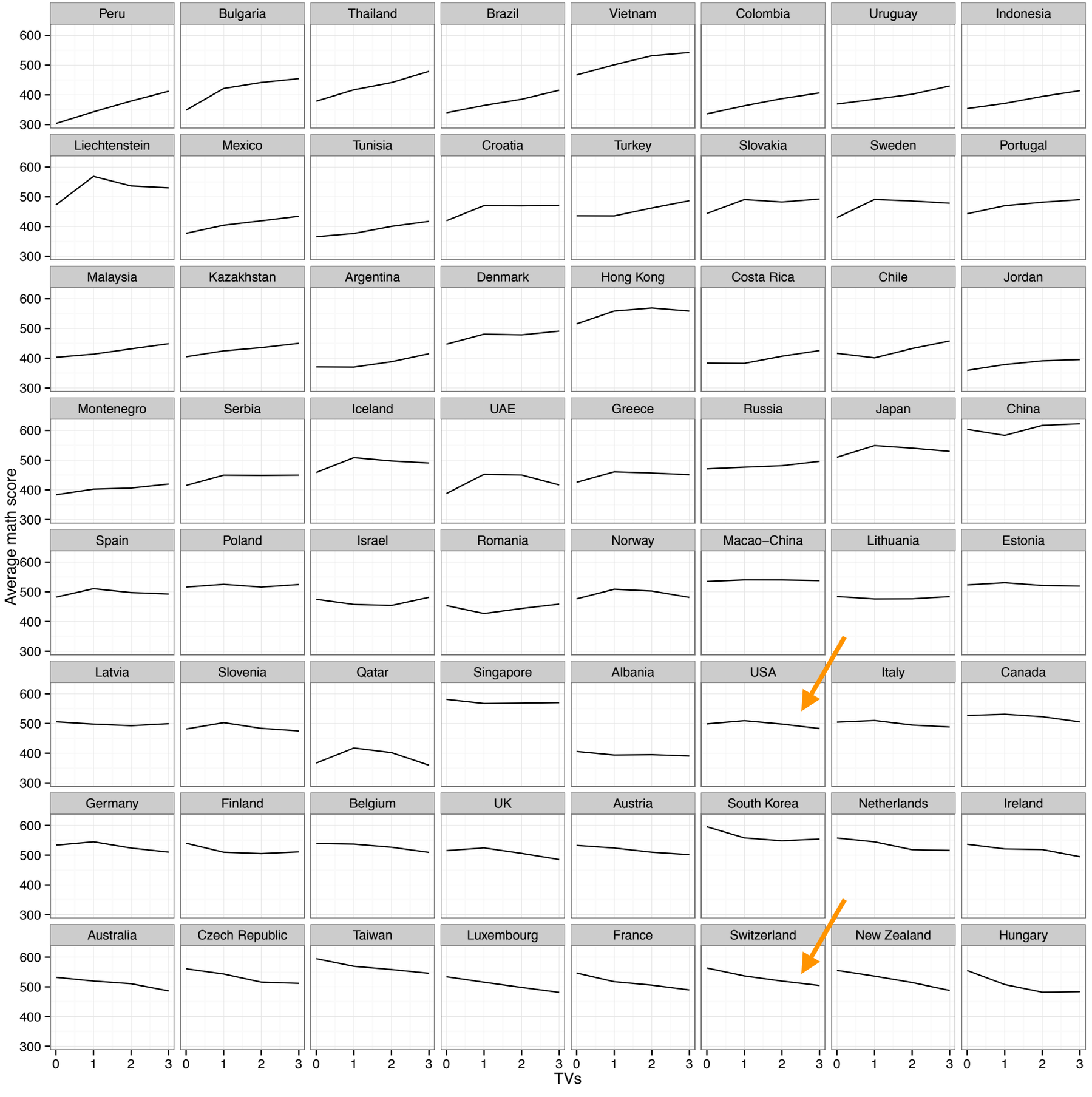



Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia
Costa Rica
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hong Kong
Hungary
Iceland
Indonesia
Ireland
Israel
Italy
Japan
Jordan
Kazakhstan
Latvia
Liechtenstein
Lithuania
Luxembourg
Macao-China
Malaysia
Mexico
Montenegro
Netherlands
New Zealand
Norway
Peru
Poland
Portugal
Qatar
Romania
Russia
Serbia
Singapore
Slovakia
Slovenia
South Korea
Spain
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
UAE
UK
Uruguay
USA
Vietnam
600
500
400
300
0
5
10
Math score
Number of possessions

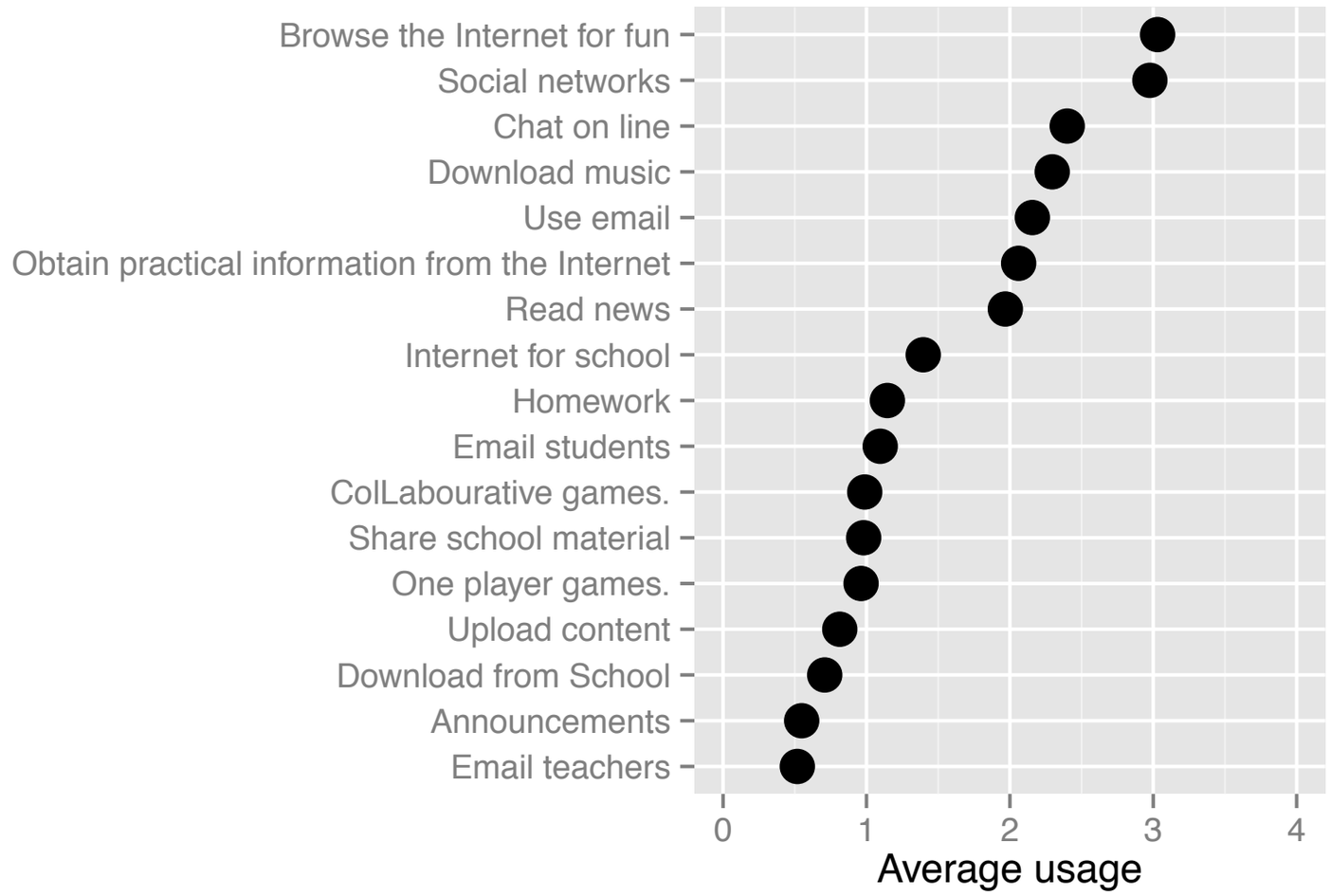


Number of TVs and average math score by country




Peru
Bulgaria
Thailand
Brazil
Vietnam
Colombia
Uruguay
Indonesia
Liechtenstein
Mexico
Tunisia
Croatia
Turkey
Slovakia
Sweden
Portugal
Malaysia
Kazakhstan
Argentina
Denmark
Hong Kong
Costa Rica
Chile
Jordan
Montenegro
Serbia
Iceland
UAE
Greece
Russia
Japan
China
Spain
Poland
Israel
Romania
Norway
Macao−China
Lithuania
Estonia
Latvia
Slovenia
Qatar
Singapore
Albania
USA
Italy
Canada
Germany
Finland
Belgium
UK
Austria
South Korea
Netherlands
Ireland
Australia
Czech Republic
Taiwan
Luxembourg
France
Switzerland
New Zealand
Hungary
Average math score
300
400
500
600
0
1
2
3
TVs

Reported internet use by Swiss teens



“Never or hardly ever” = 0
“Once or twice a month” =1
“Once or twice a week” = 2
“Almost every day” = 3
“Every day” = 4

What we learned

- Gender gap is not universal in math.
- Gender gap is universal in reading.
- Individual differences trump everything else.
- Parents matter! Mothers more than fathers!
- Socioeconomic status matters.
- Turn the TV off, if you live in a developed country!
- Something's funny about Albania's data.

Your Turn

Take two minutes to brainstorm with your neighbors.

- Based on the data, what other things would you like to investigate?
- What calculations, tables, plots would you make to tackle these questions?

Graphics choices

- Sorting!
- Ordered dot plots
- Line plots
- Histograms
- Showing stats vs all values
- Uncertainty - barcharts of category counts

Climate change

CO2 alarm bells?

"Planet's CO_2 level reaches 400 ppm for first time in human existence."

"It's not a question of if air capture technology will be adopted; it's a question of when," said Klaus Lackner

"In the last 80 years, as CO_2 emissions have most rapidly escalated, the annual rate of climate-related deaths worldwide fell by an incredible 98%. This means the incidence of death from climate is 50 times lower than it was 80 years ago," writes Epstein.

Climate change

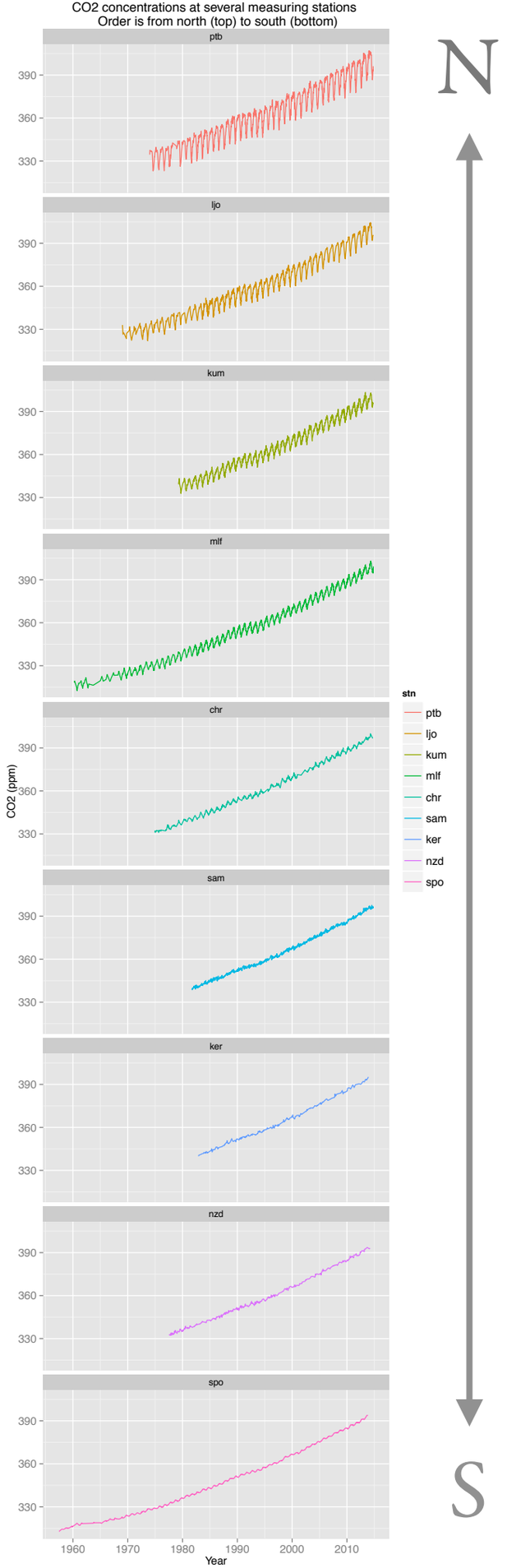
Data available at http://scrippsco2.ucsd.edu

```
> head(CO2.all)
        date   time       day  decdate n flg    co2  lat     lon stn
1 1997-01-21  13:40 35451.57 1997.056 4   0 365.82 23.3 -110.2 bcs
2 1997-02-08  15:03 35469.63 1997.106 2   0 365.57 23.3 -110.2 bcs
3 1997-02-22  15:07 35483.63 1997.144 2   0 366.28 23.3 -110.2 bcs
4 1997-03-07  14:23 35496.60 1997.180 3   0 367.85 23.3 -110.2 bcs
5 1997-03-22  14:16 35511.59 1997.221 2   0 365.28 23.3 -110.2 bcs
6 1997-04-05  13:37 35525.57 1997.259 3   0 368.96 23.3 -110.2 bcs
…
```

Climate change

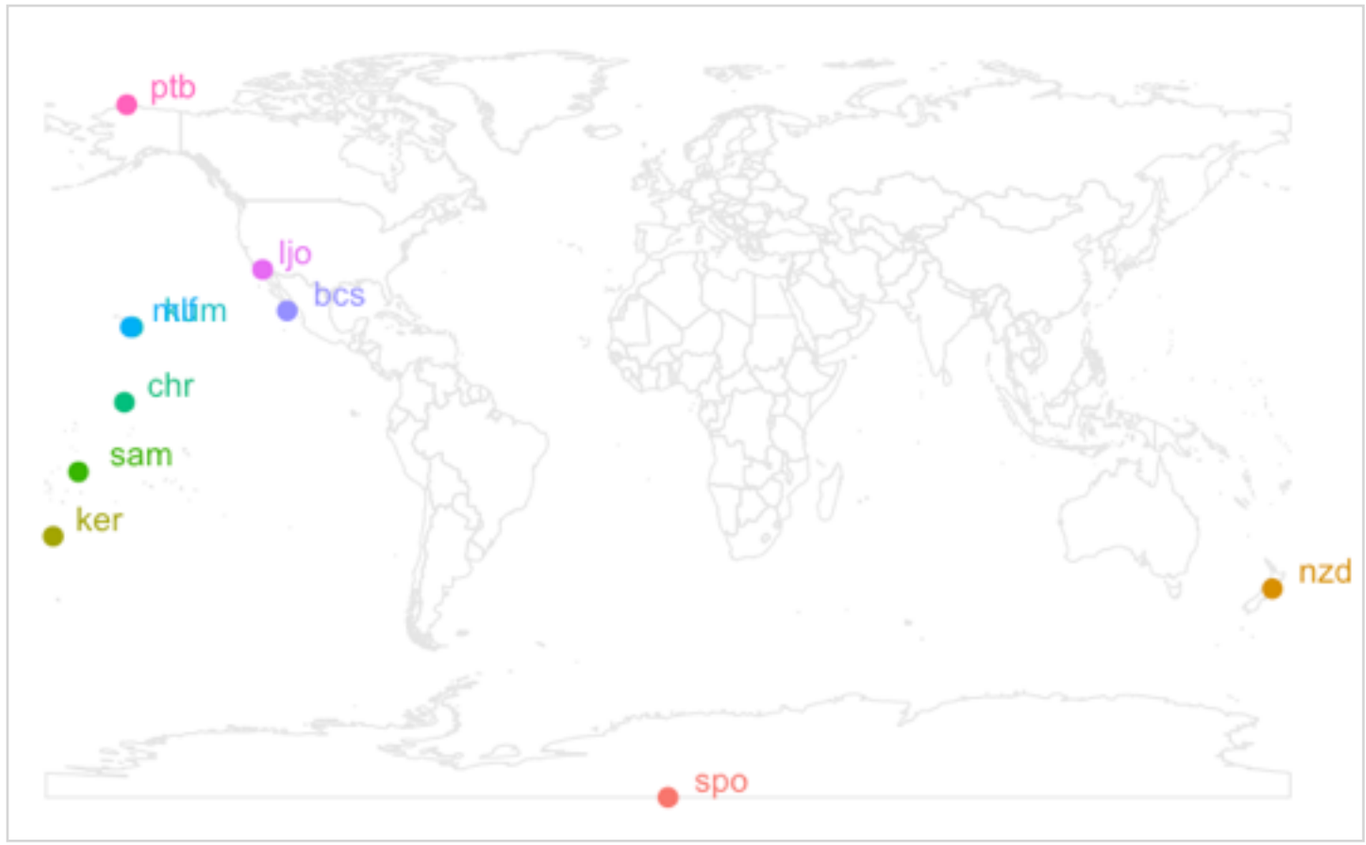
I expected:

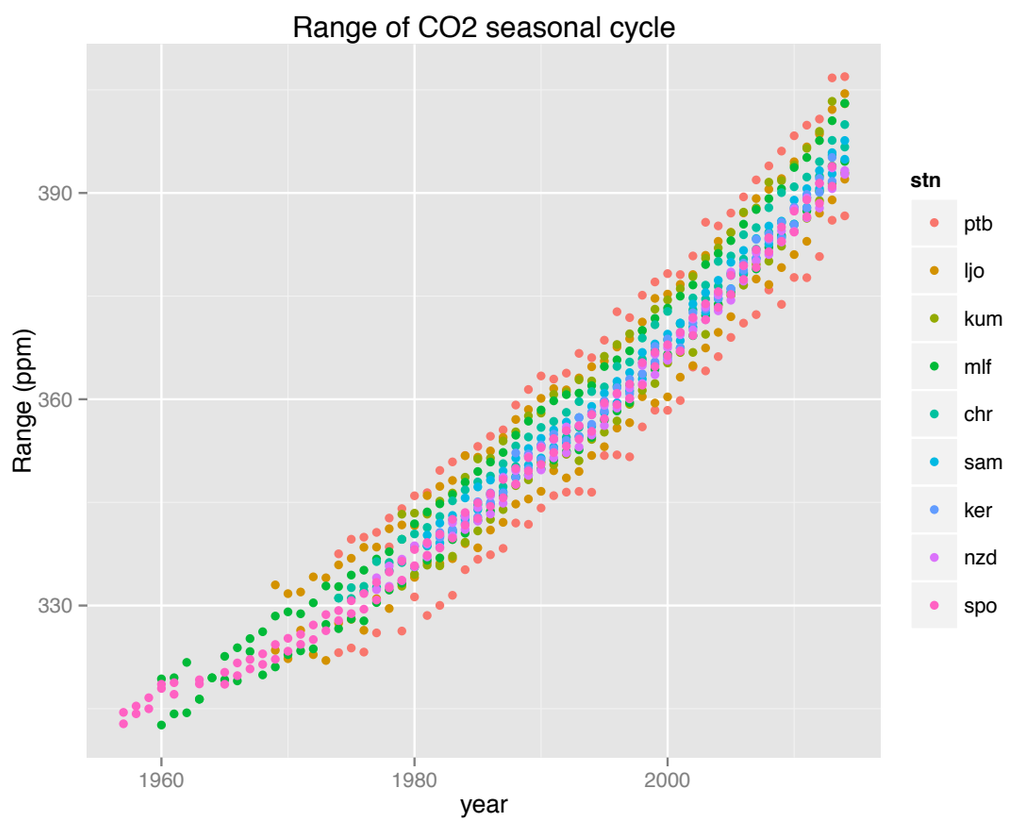
- CO2 concentrations to be flattening over time
- Raw data to look like raw data - messy

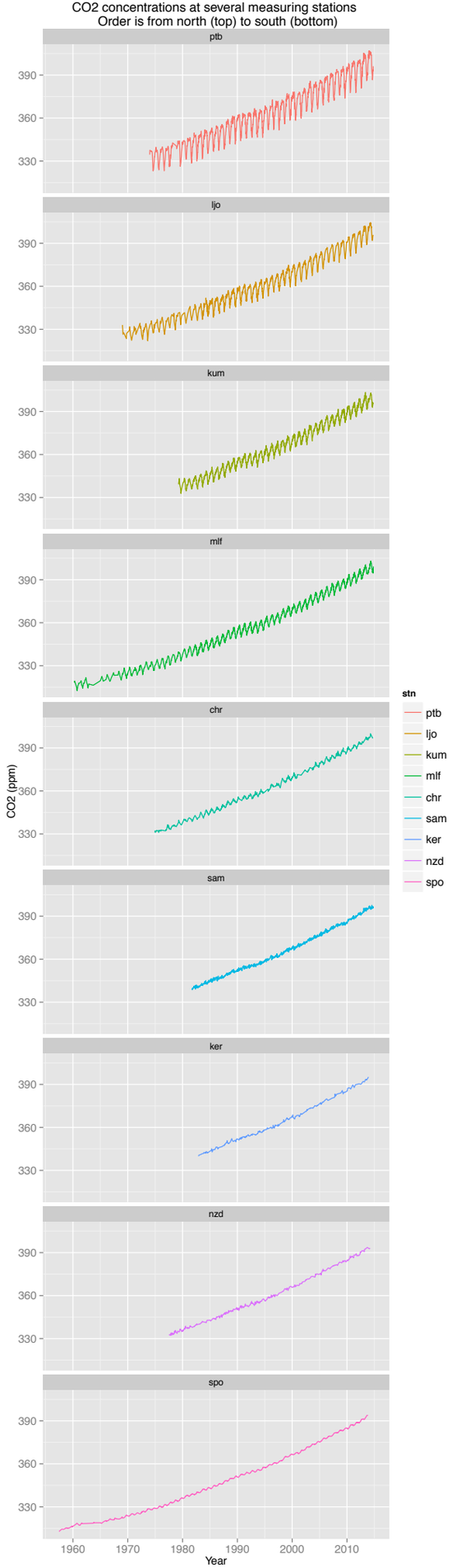

CO2 concentrations at several measuring stations
Order is from north (top) to south (bottom)
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo
CO2 (ppm)
390
360
330
1960
1970
1980
1990
2000
2010
Year
stn
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo

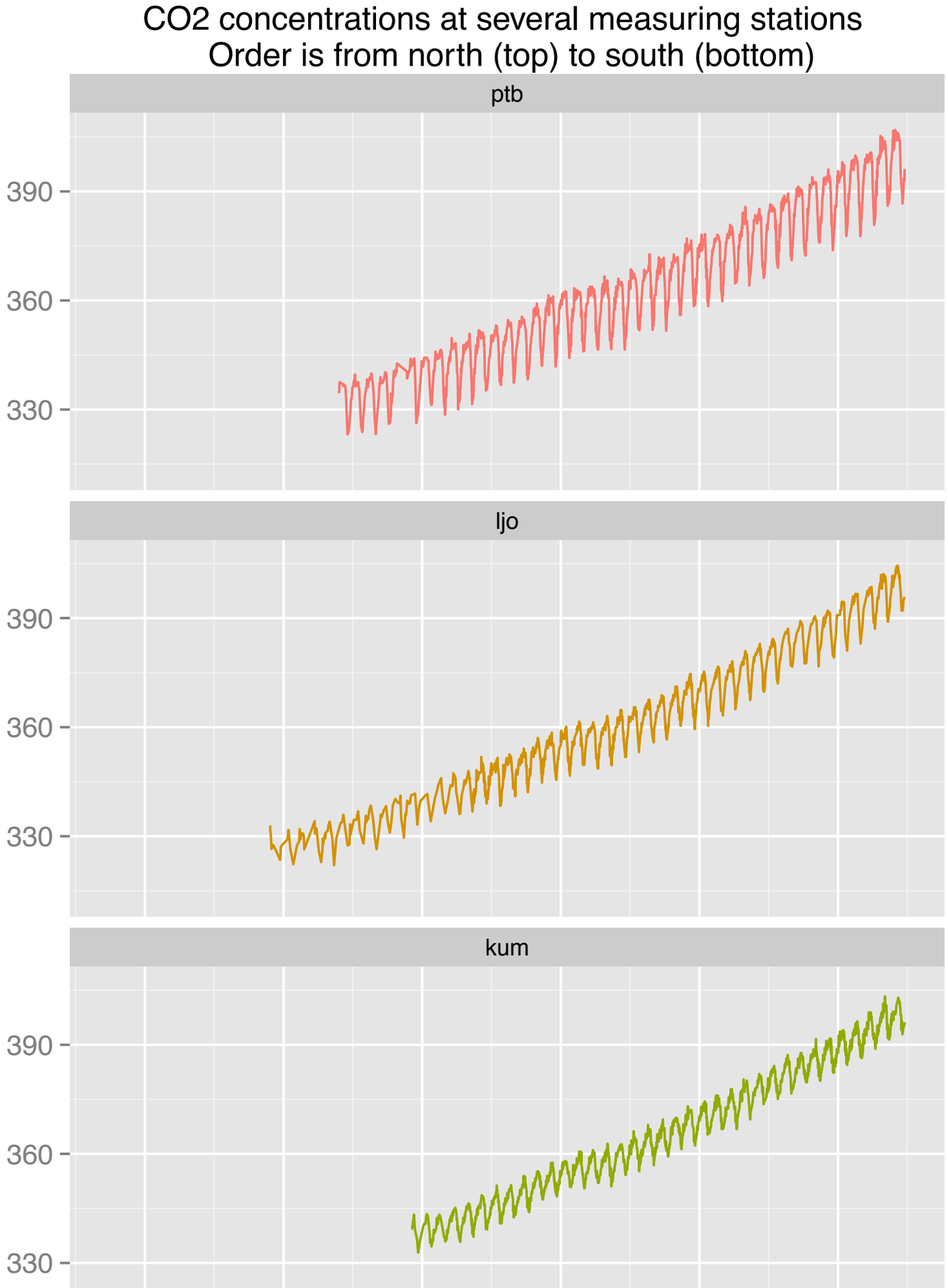
N

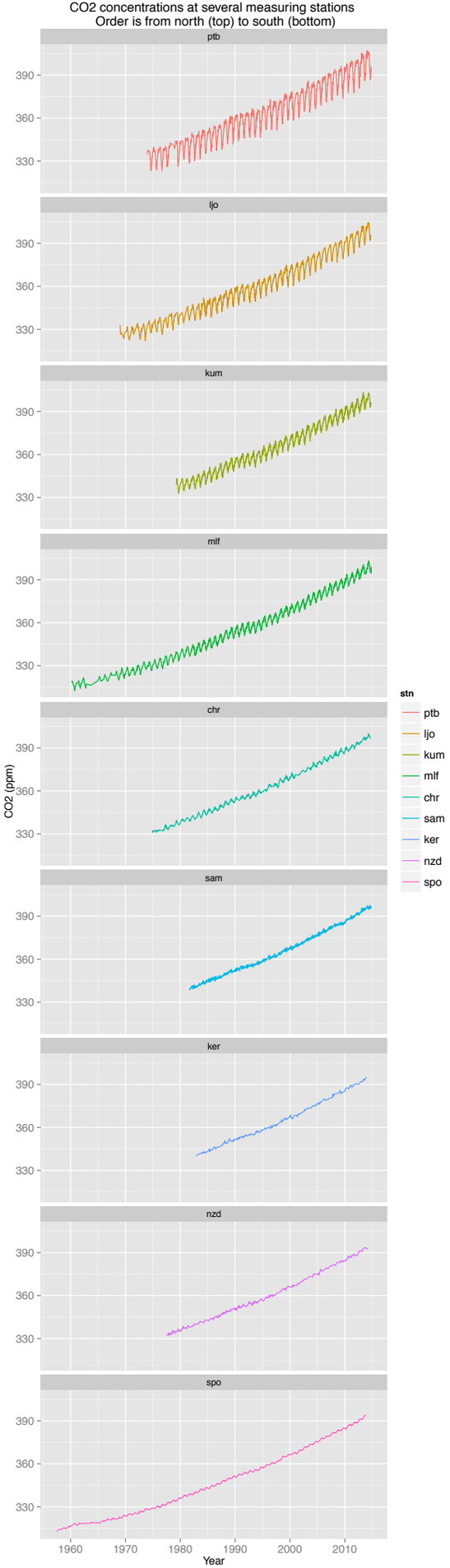
S

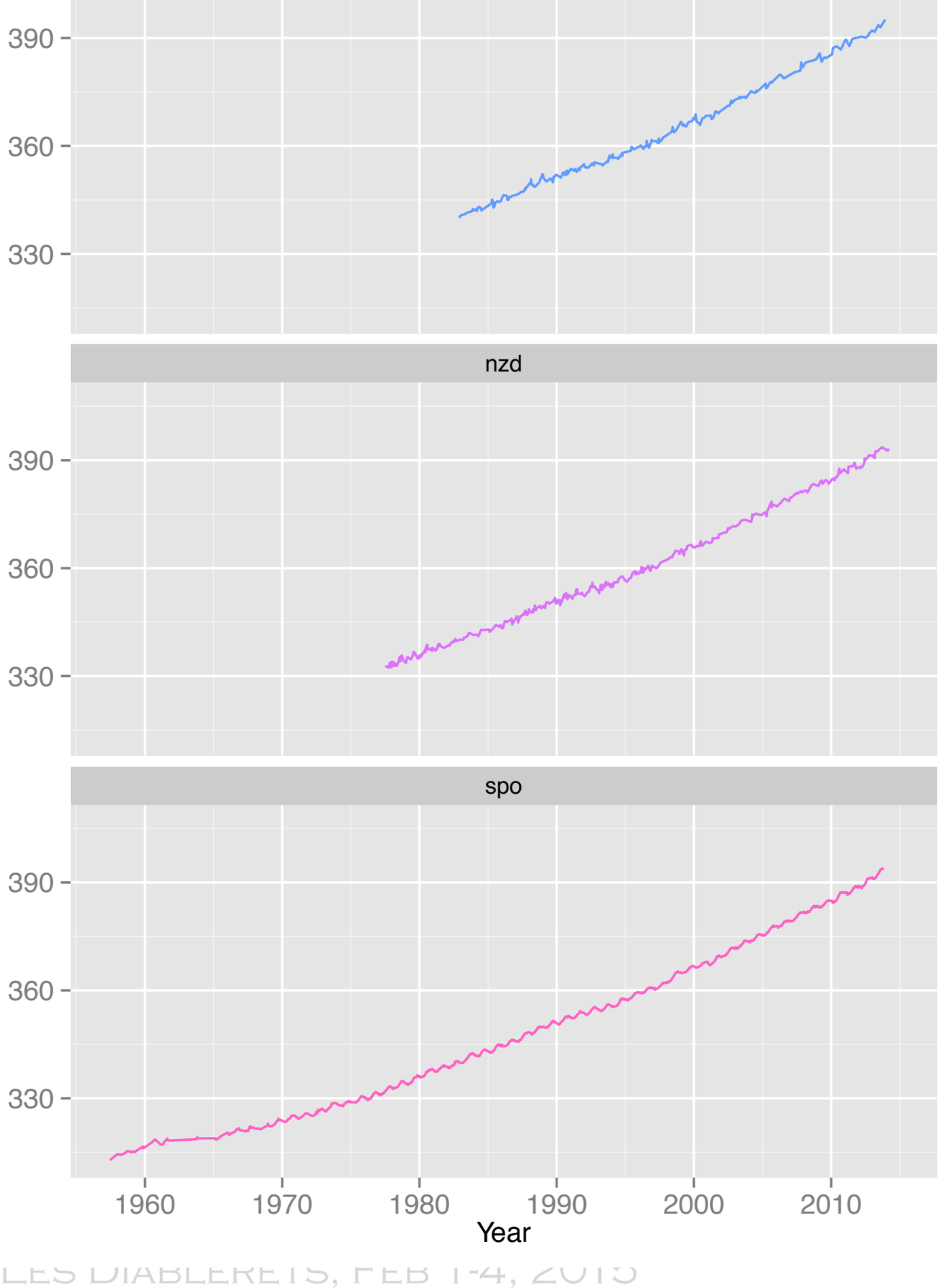

ptb
ljo
bcs
chr
sam
ker
nzd
spo

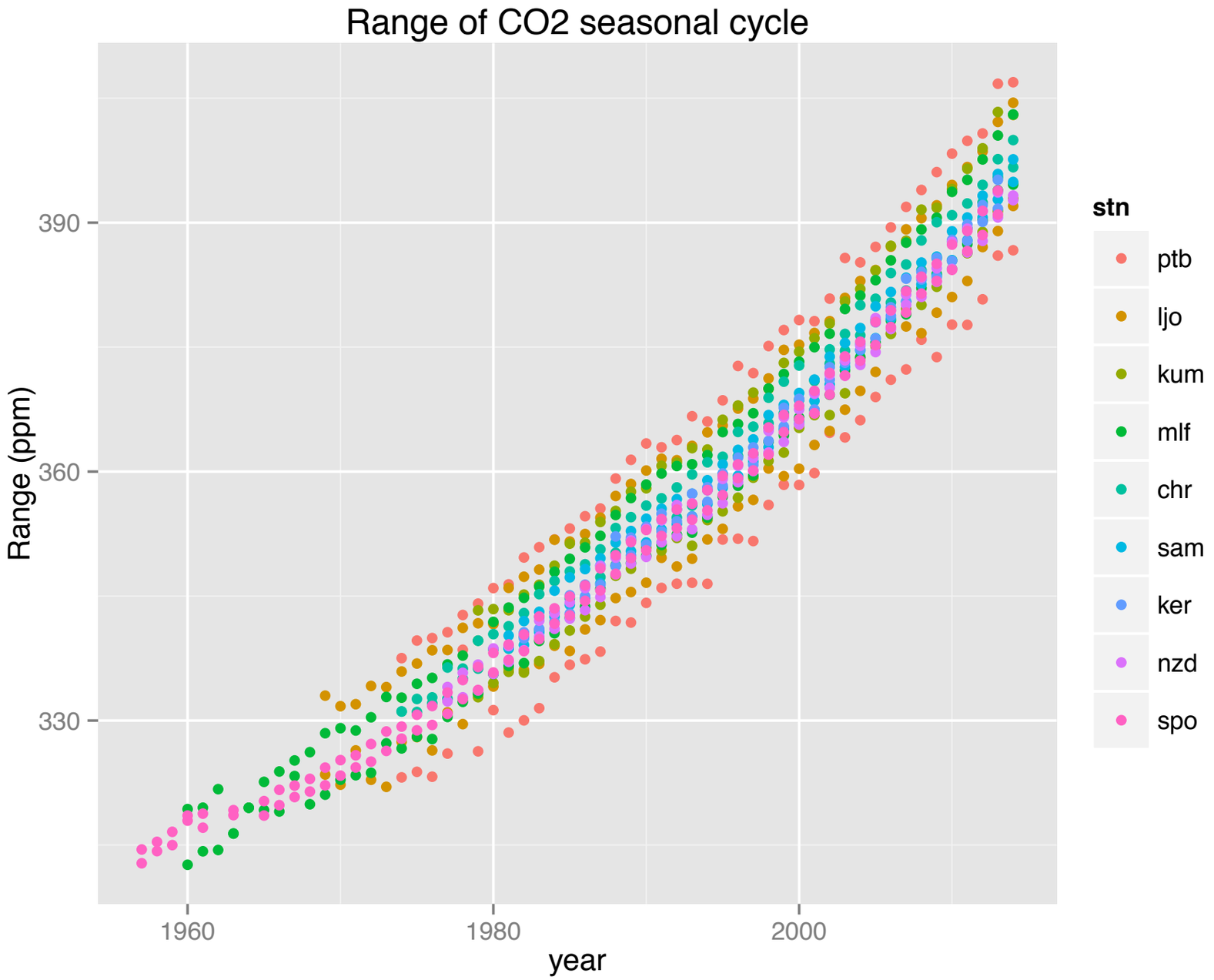

Range of CO2 seasonal cycle
Range (ppm)
390
360
330
1960
1980
2000
year
stn
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo


CO2 concentrations at several measuring stations
Order is from north (top) to south (bottom)
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo
390
360
330
CO2 (ppm)
1960
1970
1980
1990
2000
2010
Year
stn
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo

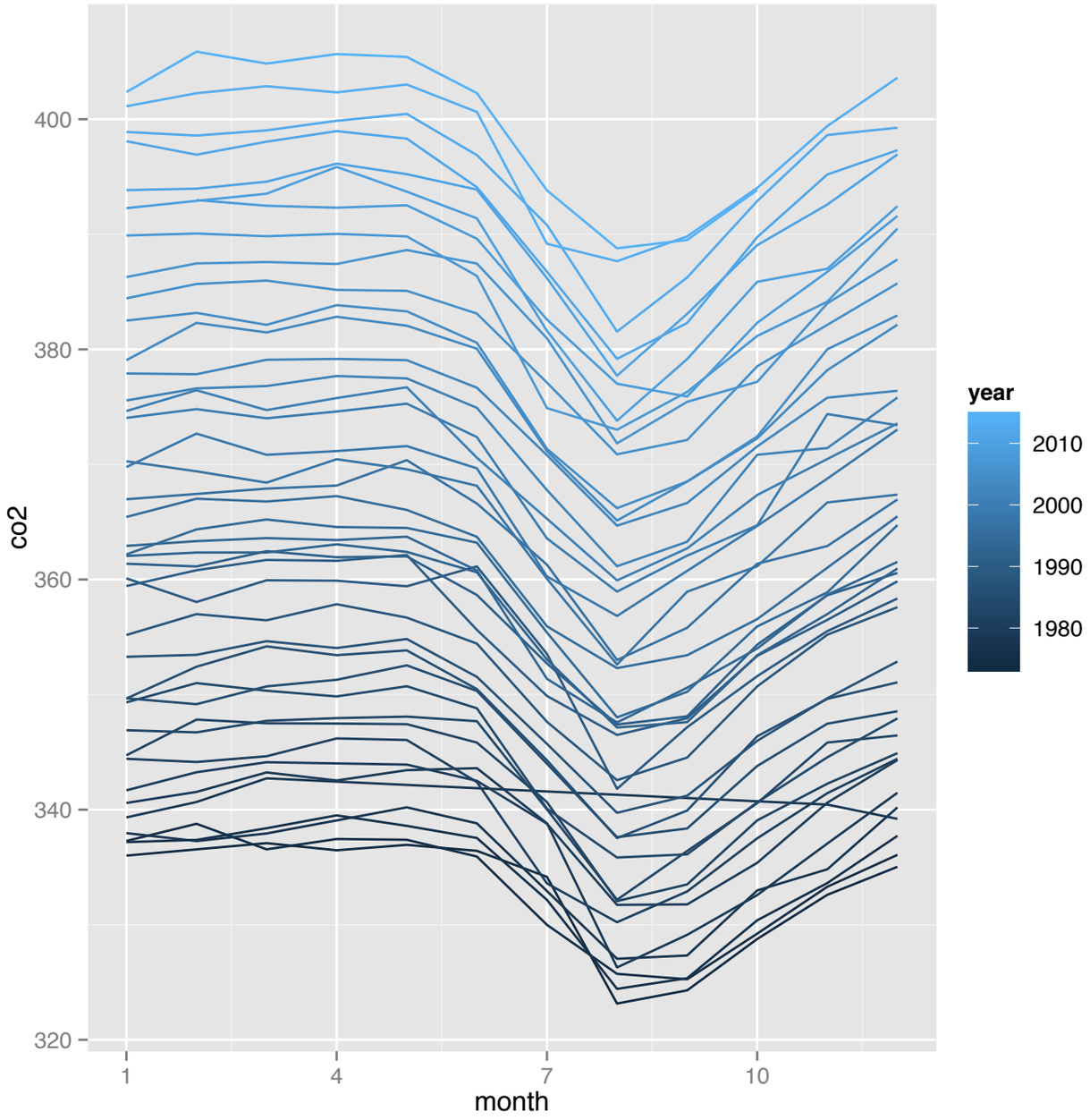

CO2 concentrations at several measuring stations
Order is from north (top) to south (bottom)
ptb
ljo
kum
390
360
330


CO2 concentrations at several measuring stations
Order is from north (top) to south (bottom)
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo
390
360
330
CO2 (ppm)
1960
1970
1980
1990
2000
2010
Year
stn
ptb
ljo
kum
mlf
chr
sam
ker
nzd
spo


nzd
spo
390
360
330
1960
1970
1980
1990
2000
2010
Year



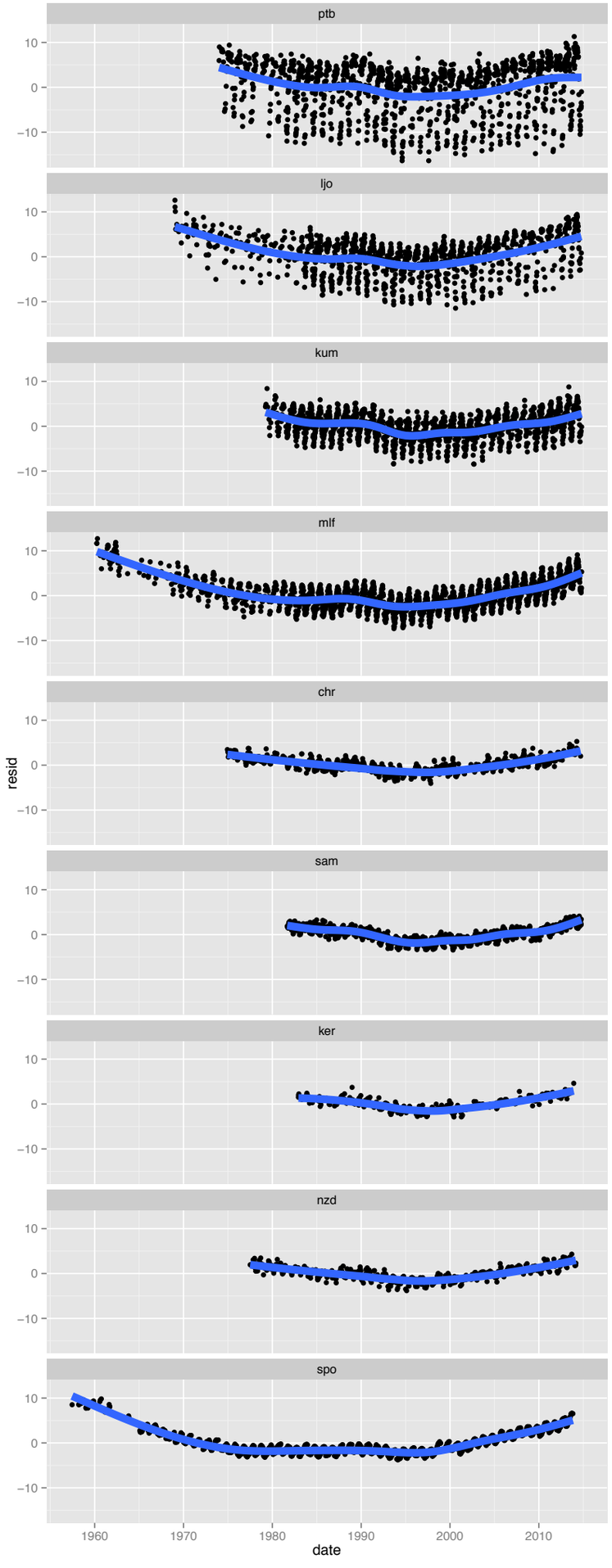
Minimum and maximum for each year and station, show that they are effectively measuring the same product

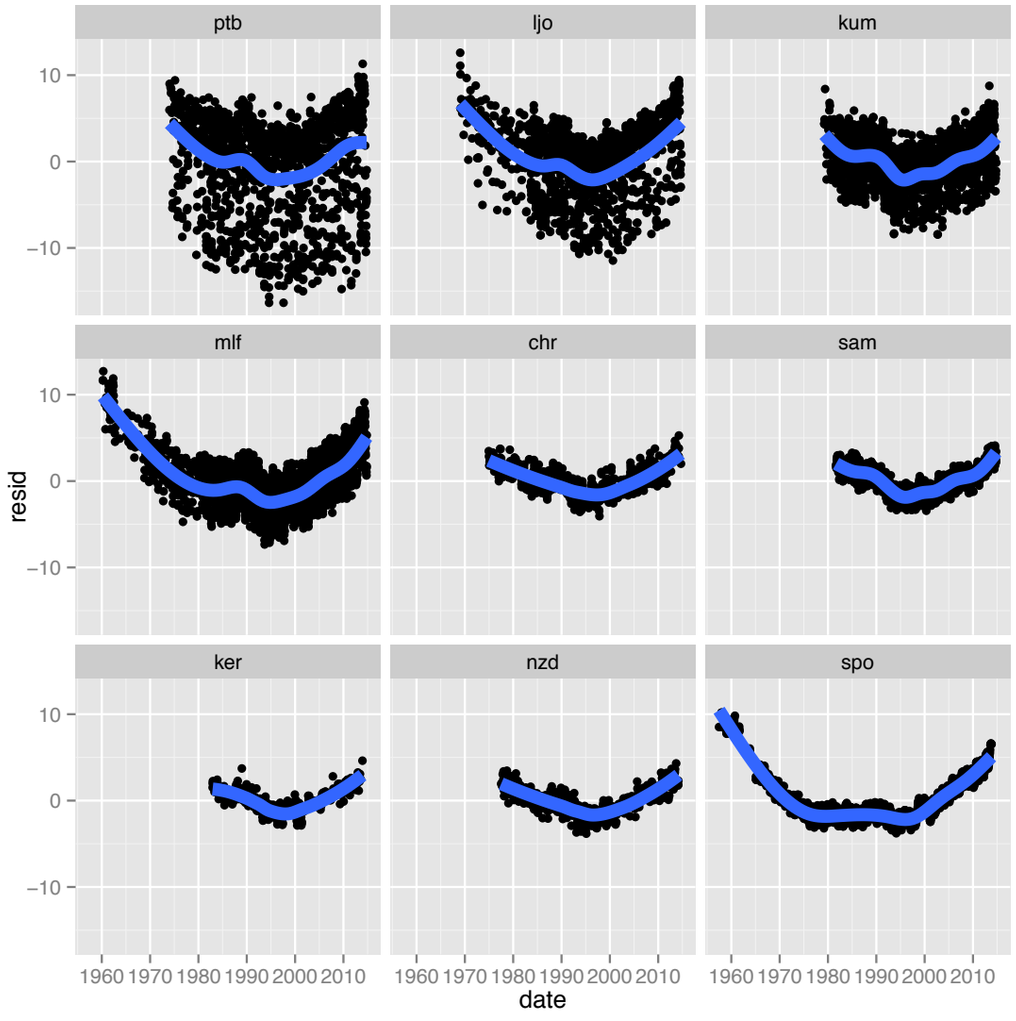


For station, ptb, most northerly station, examine monthly trend. Drop in CO_2 occurs every year in June.

Strange year is 1978, and lack measurements for summer.

Residuals from a linear model fit indicate the rate of increase is exponential.





Change aspect ratio and anomalies at different sites are visible.

Climate change

Surprises:

- CO2 concentrations increasing consistently.
- Northern stations show seasonality.
- Data looks very regular. Every recording station essentially producing same values, ignoring seasonality, with a few slight anomalies.

oco.jpl.nasa.gov

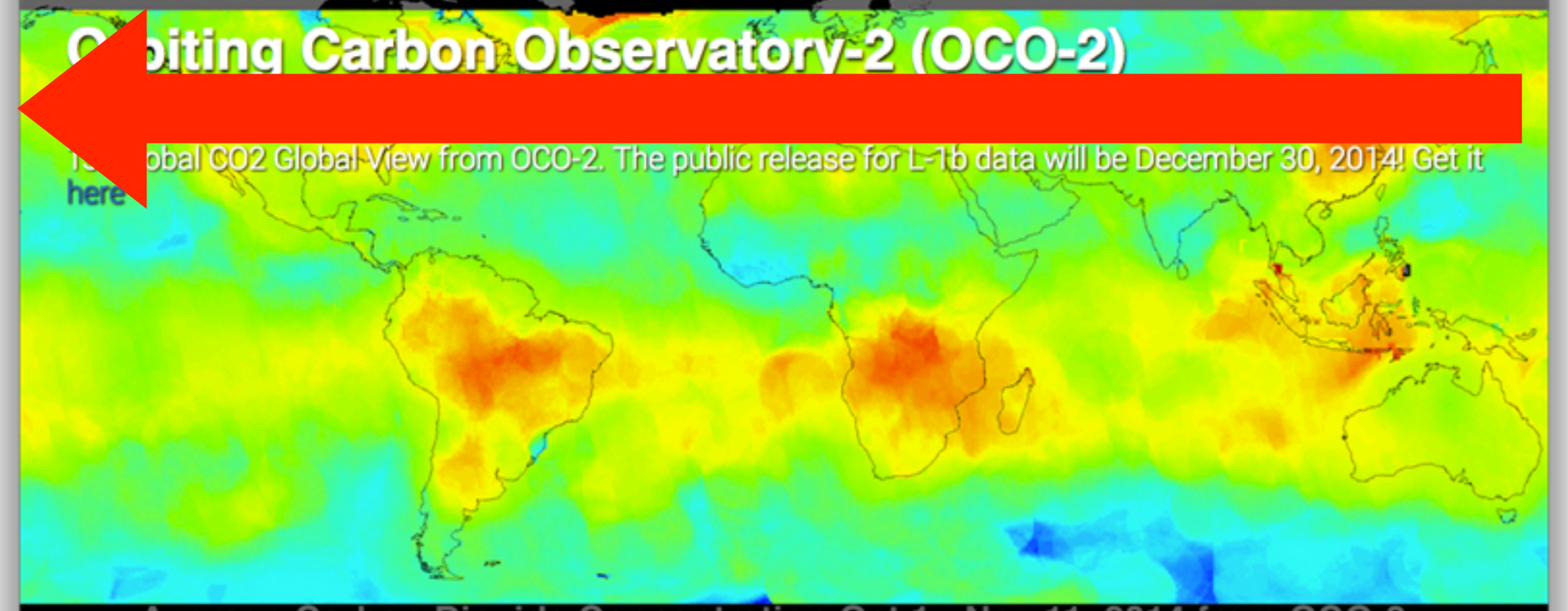
Jet Propulsion Laboratory
California Institute of Technology

JPL Home | News | Images | Videos | Missions | Social | search

Mission | Observatory | Science | Galleries | News | Glossary

Orbiting Carbon Observatory-2 (OCO-2)

Global CO2 Global View from OCO-2. The public release for L-1b data will be December 30, 2014! Get it here



Average Carbon Dioxide Concentration Oct 1 - Nov 11, 2014 from OCO-2

387 — 402.5 ppm

USA.gov | PRIVACY | FEEDBACK | CONTACTS

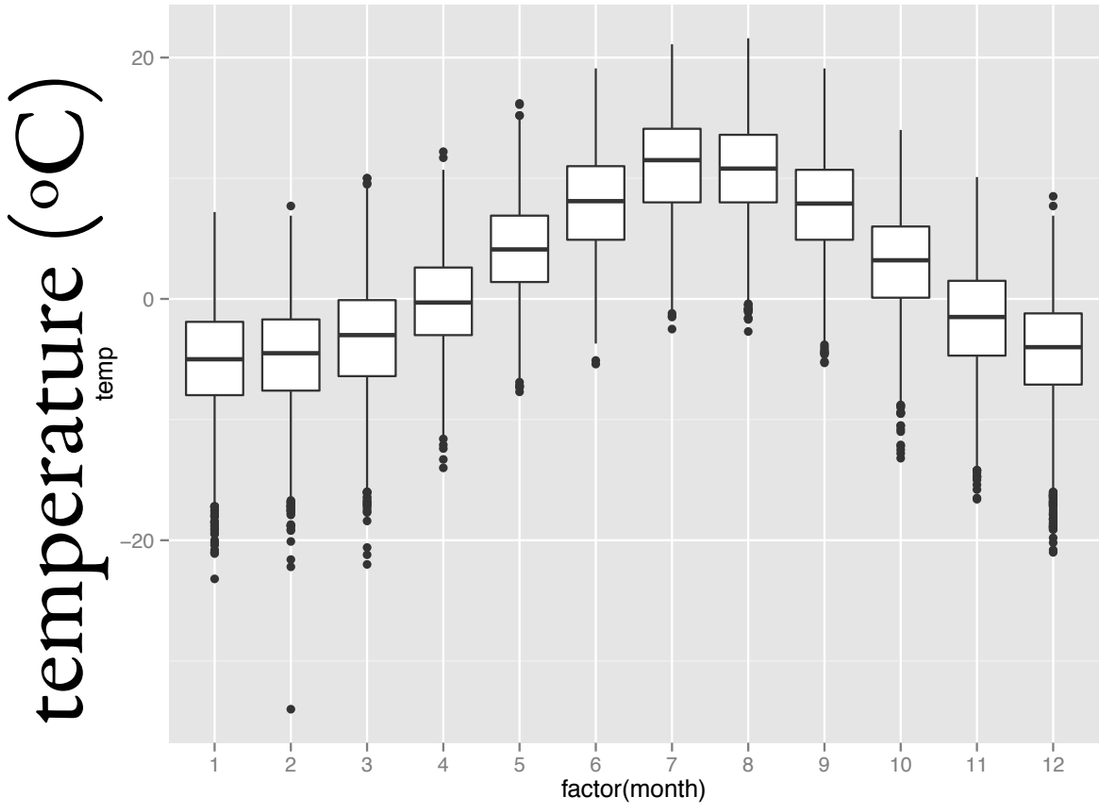
psv_suburb_unit_jun....xlsx | psv_suburb_house_ju....xlsx | 9809021e.pdf | Show All

Climate change

What's happening near Les Diablerets?

- Global Historical Climatological Network data
- Station: Col du Grand St Bernard
- Minimum/maximum temperature and precipitation

`ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt`

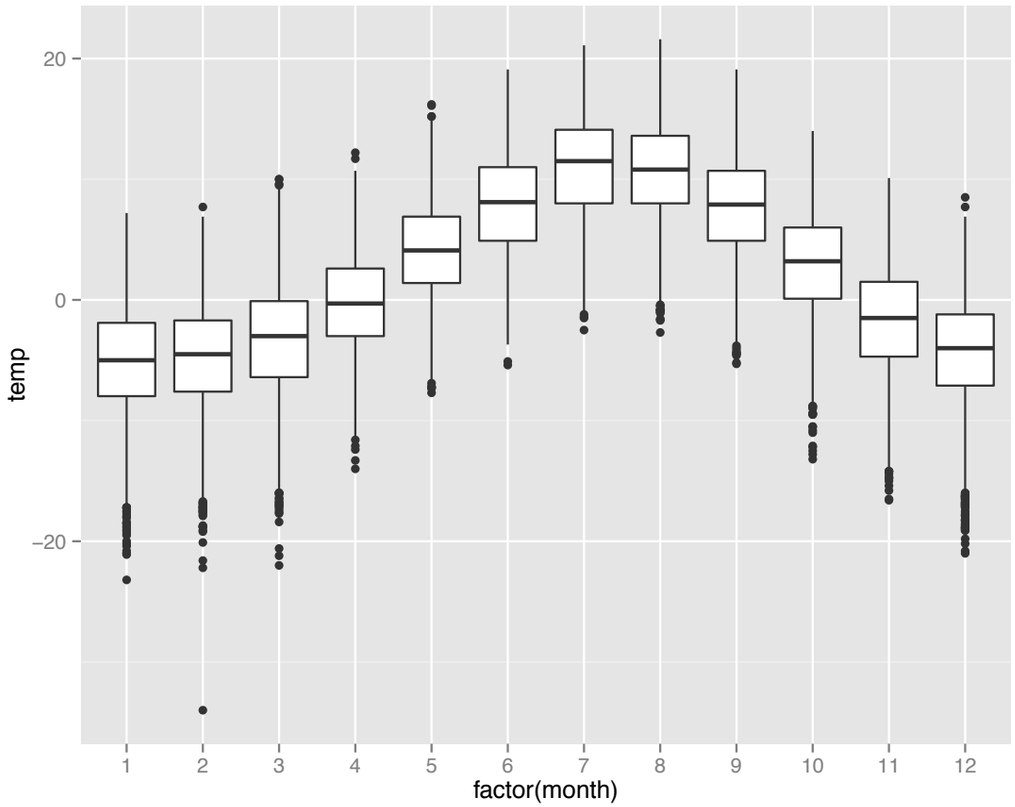


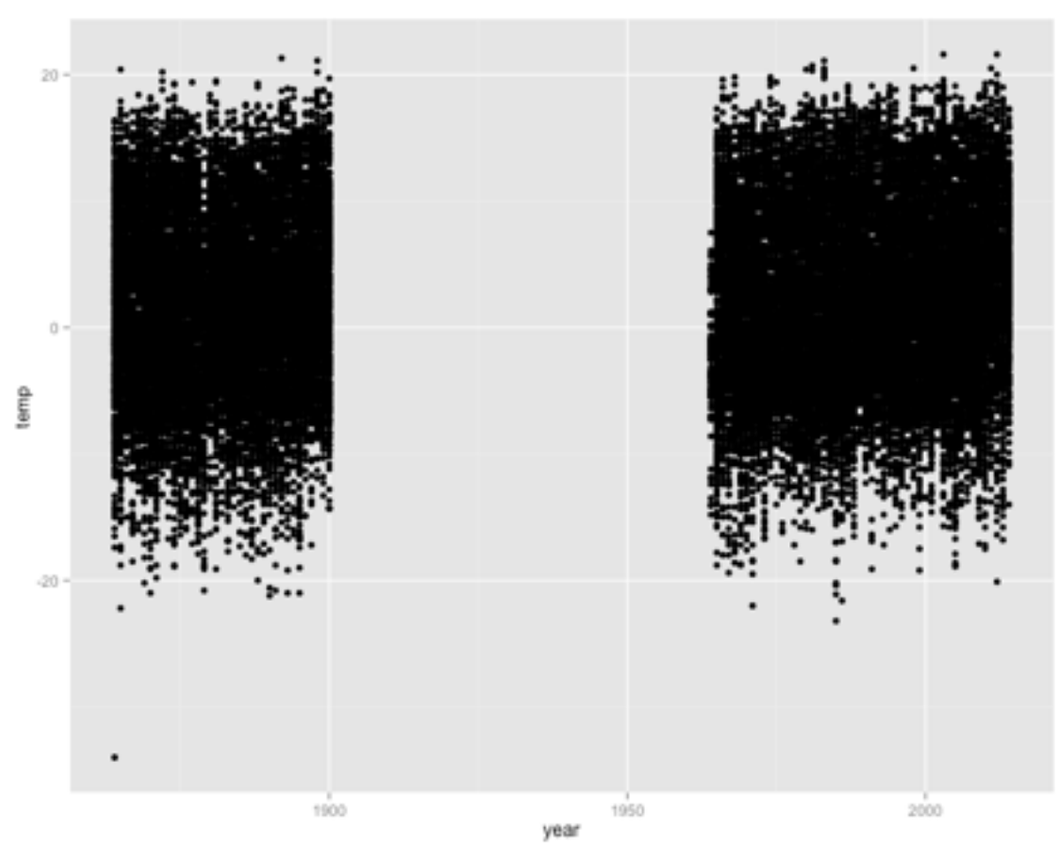
Maximum daily temperature

Seasonal trend visible from plotting temperature by month

Temperature does NOT reach freezing in SUMMER occasionally

Maximum daily temperature

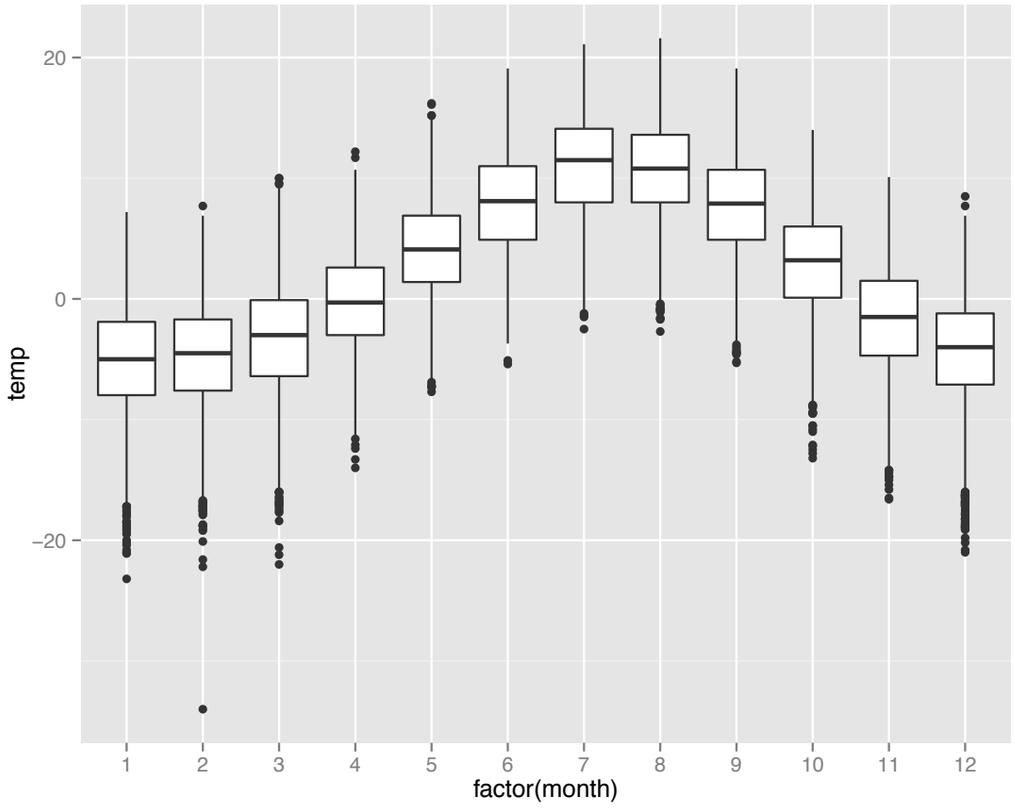


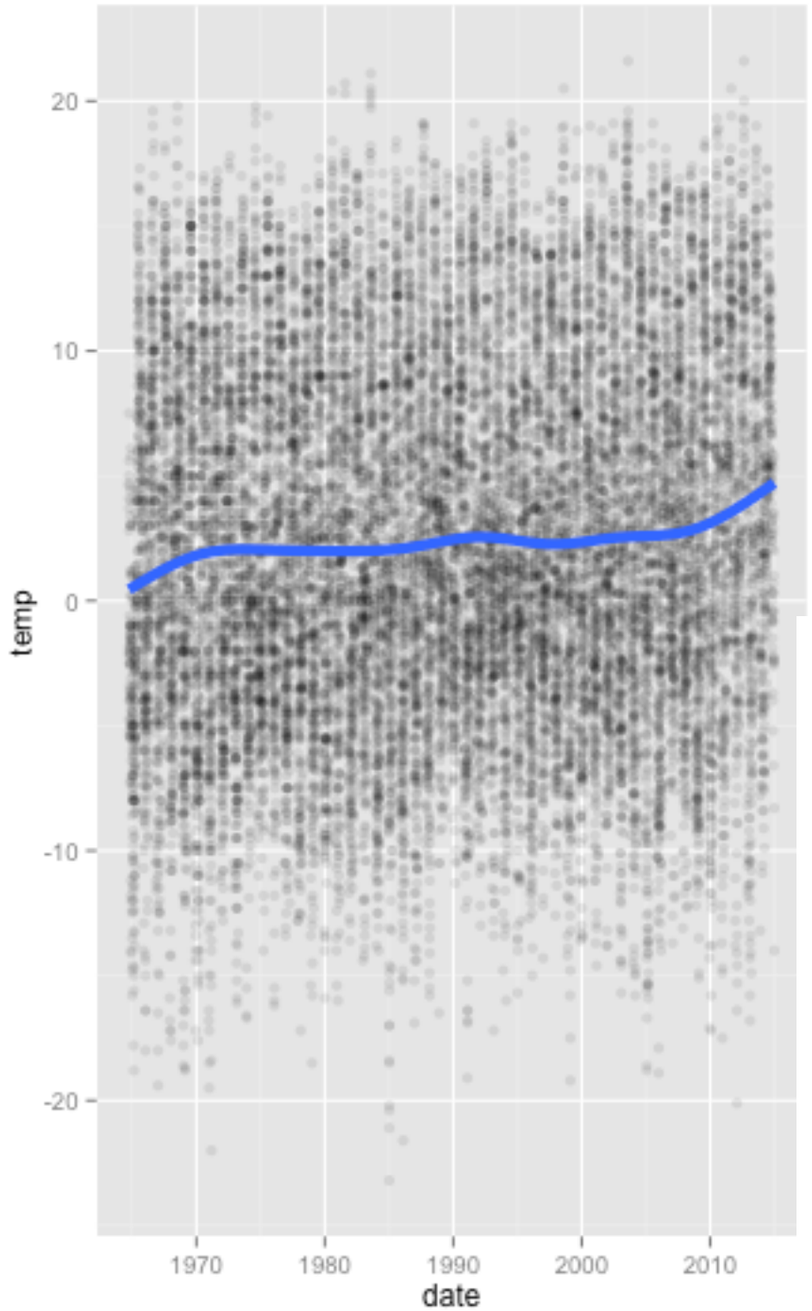


Plot of temperature by year reveals a big gap in measurements!

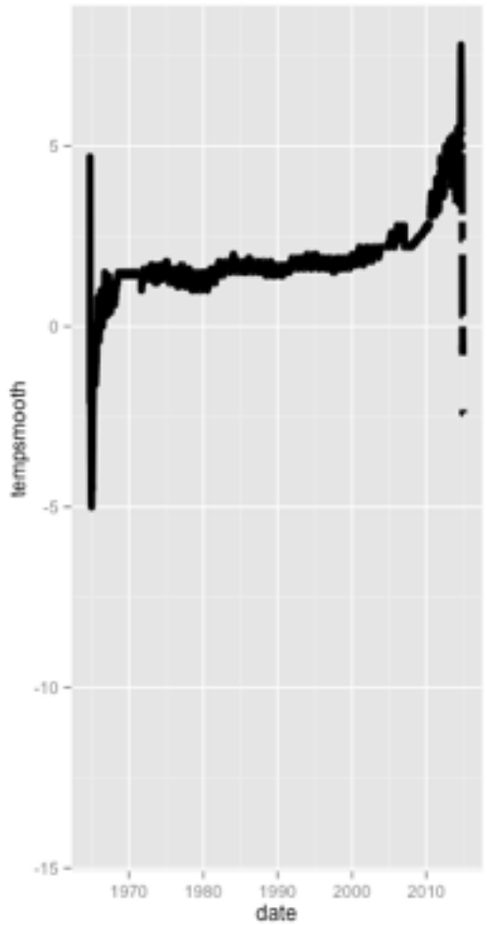
May need to focus on one period or the other.

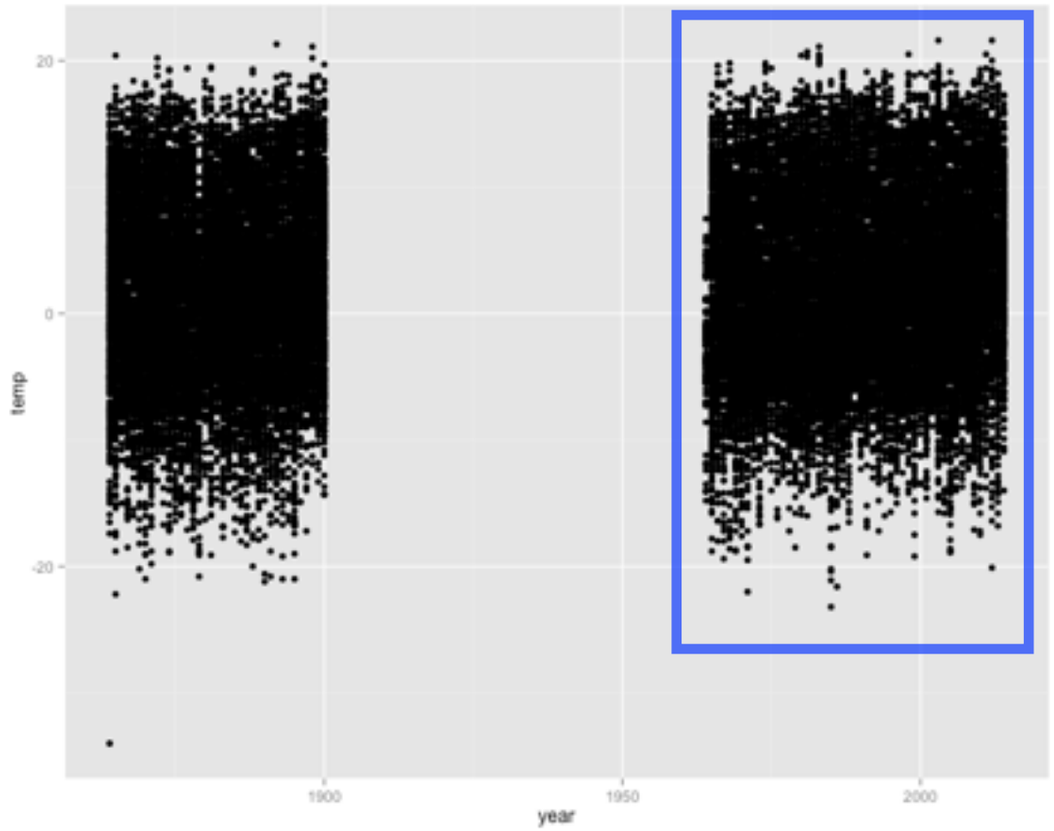
Maximum daily temperature

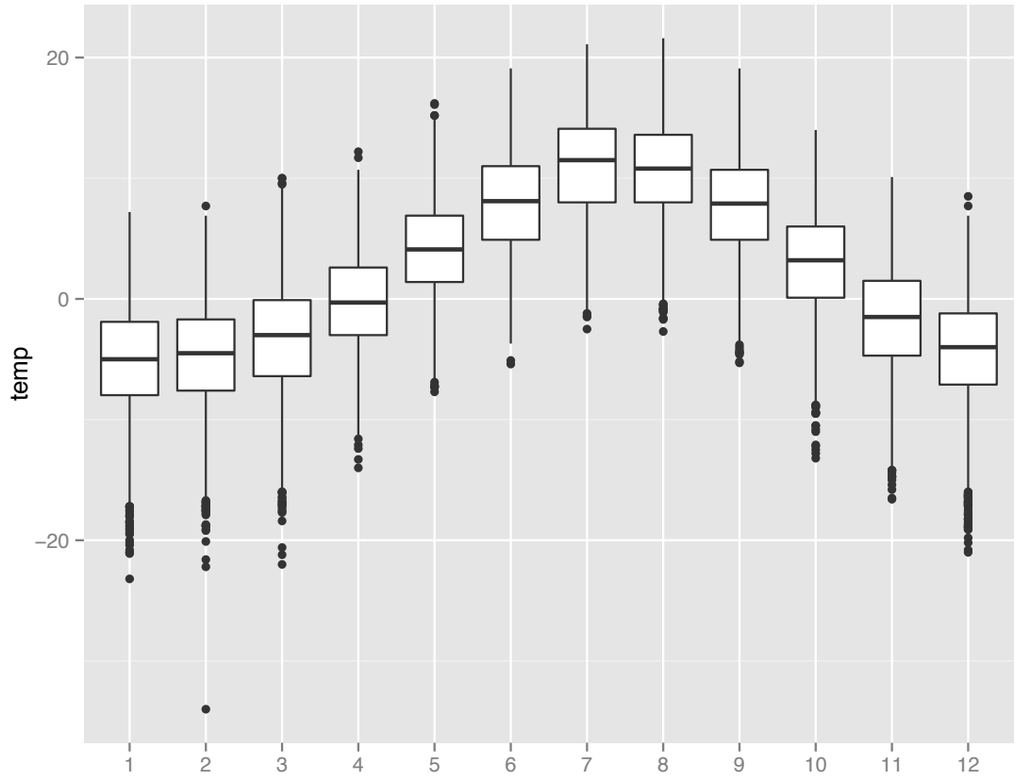




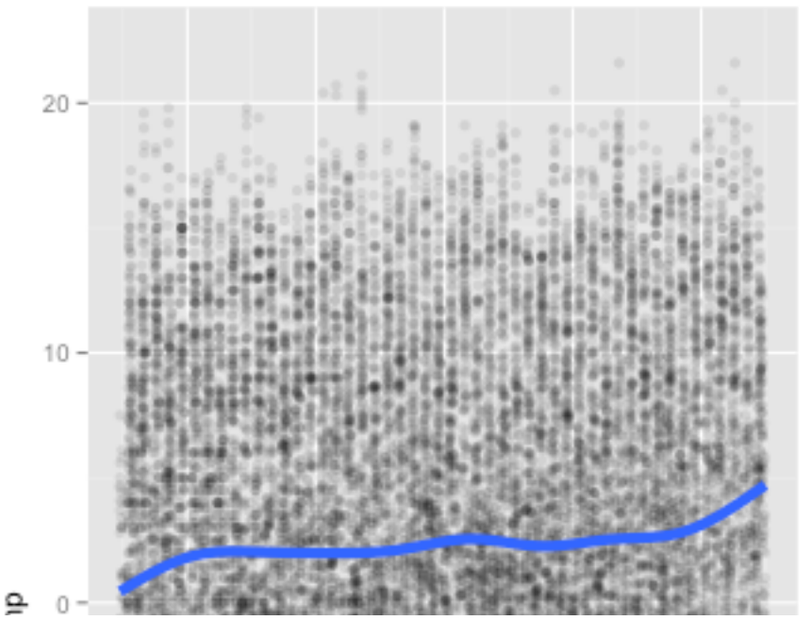
Increasing?



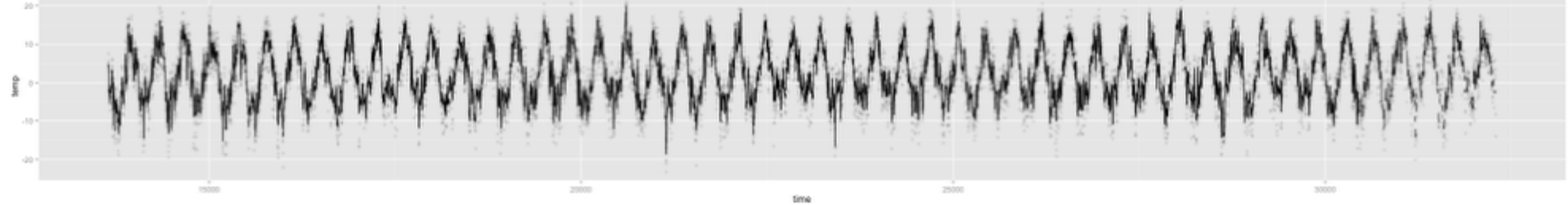


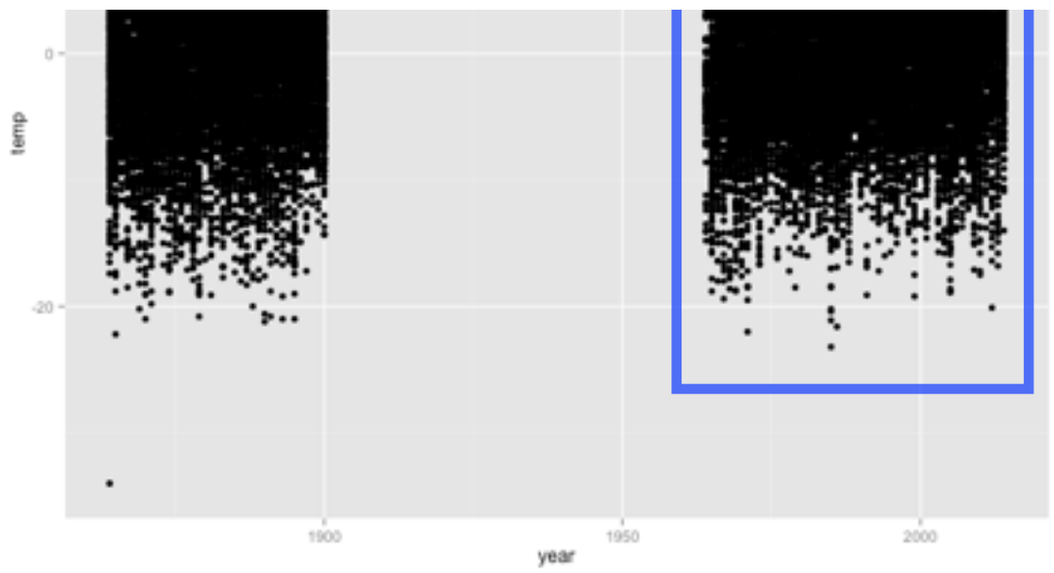


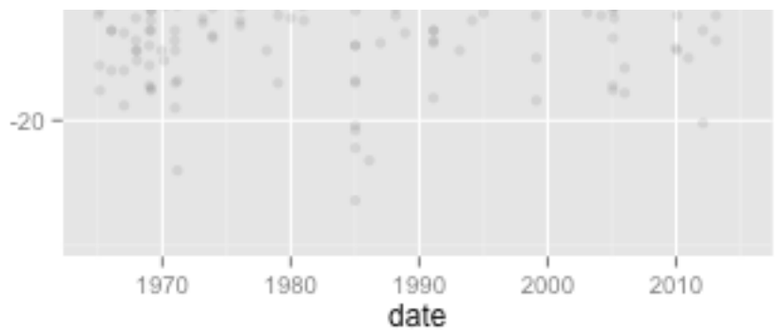
Maximum daily temperature

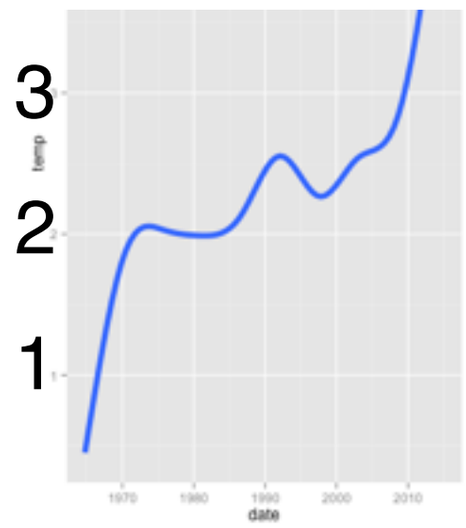


Change aspect ratio for studying season effects





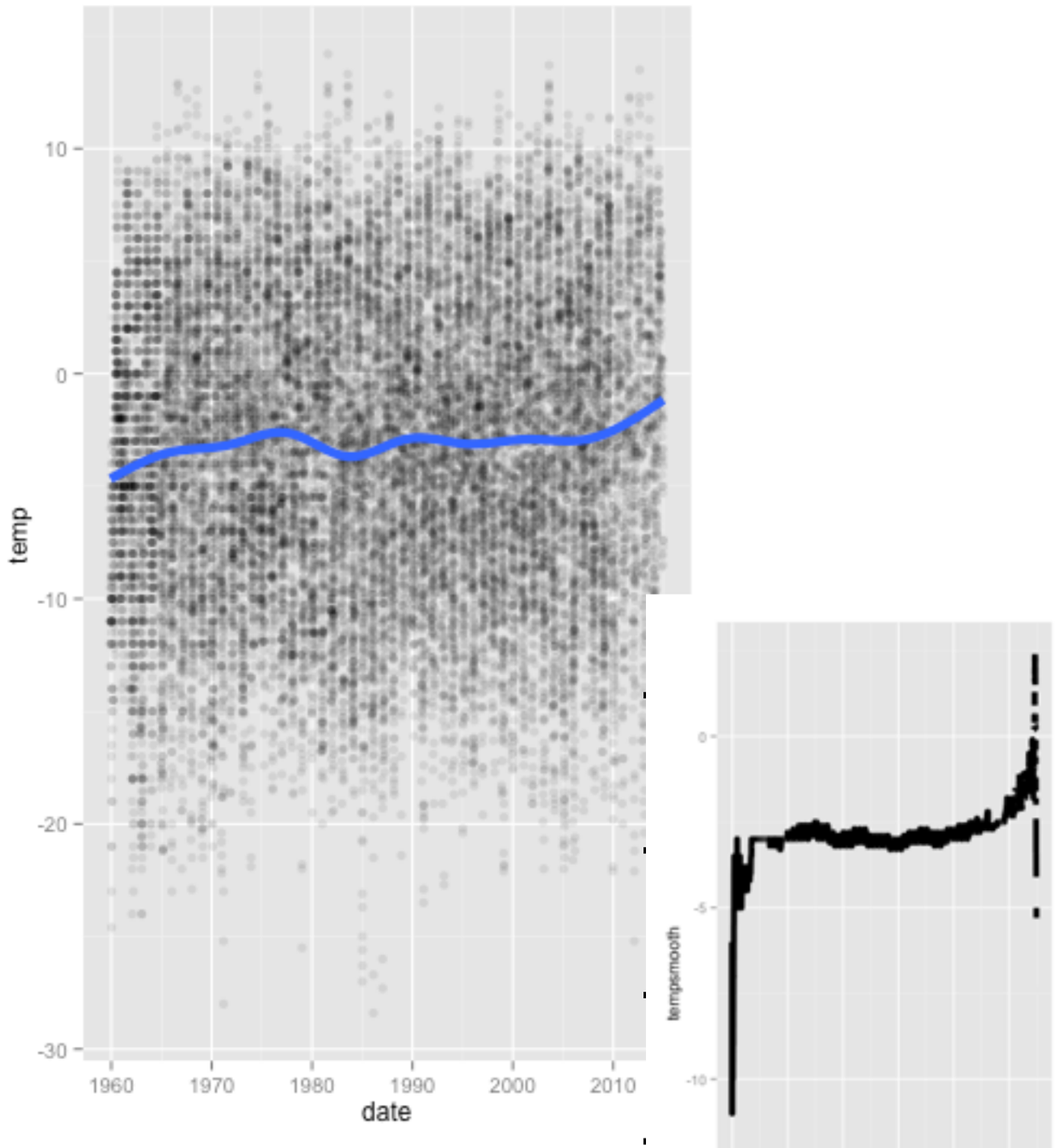




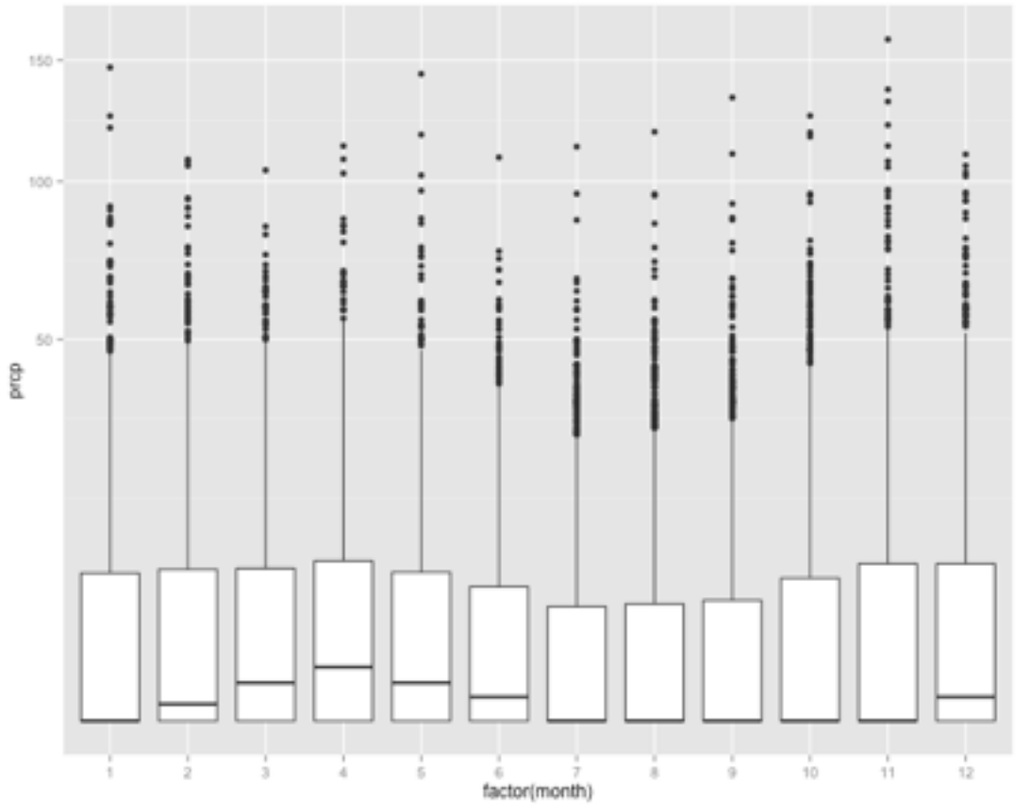
double dip winters, ...

Minimum daily temperature

Similar pattern with minimum temperature?



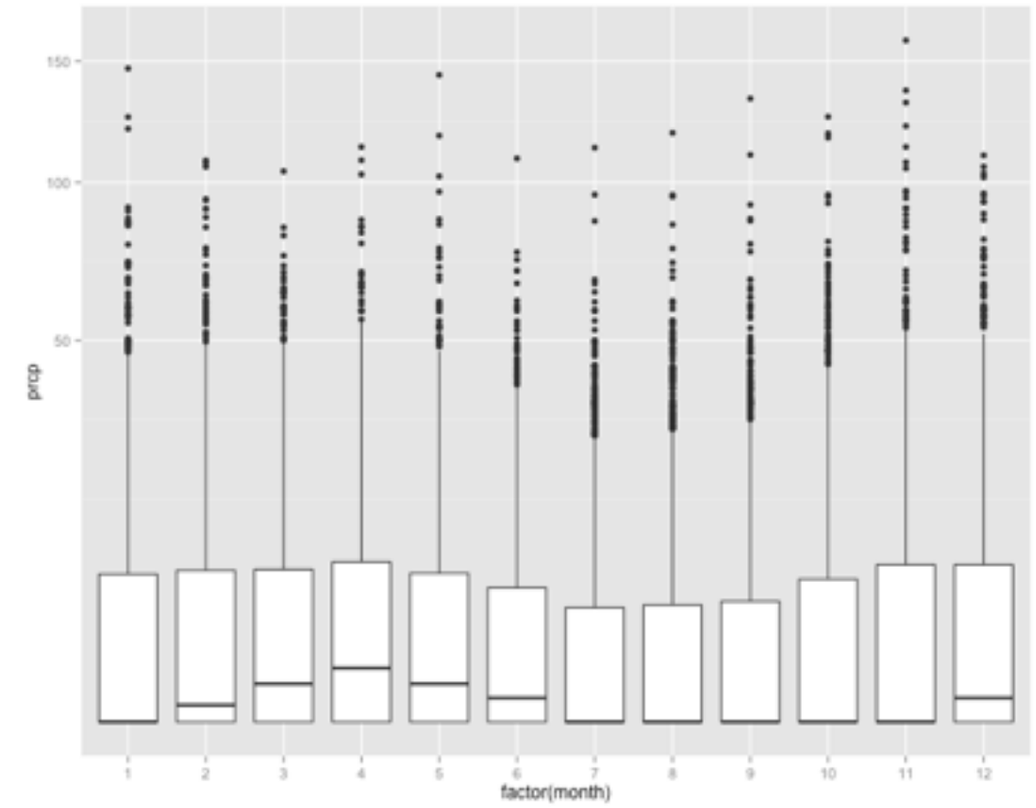
Why are there so many missing values in the last decade?

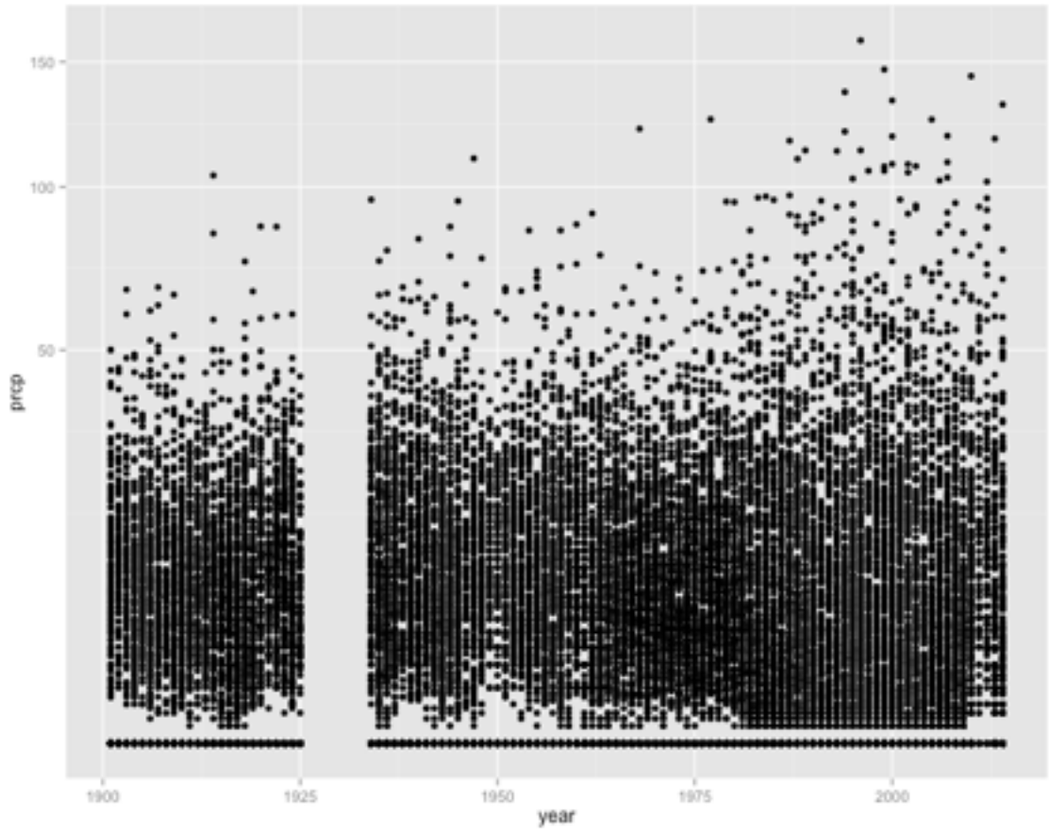


Precipitation (sqrt scale)

Small seasonal trend visible from plotting precip by month - increase Feb-Jun.

Precipitation (sqrt scale)

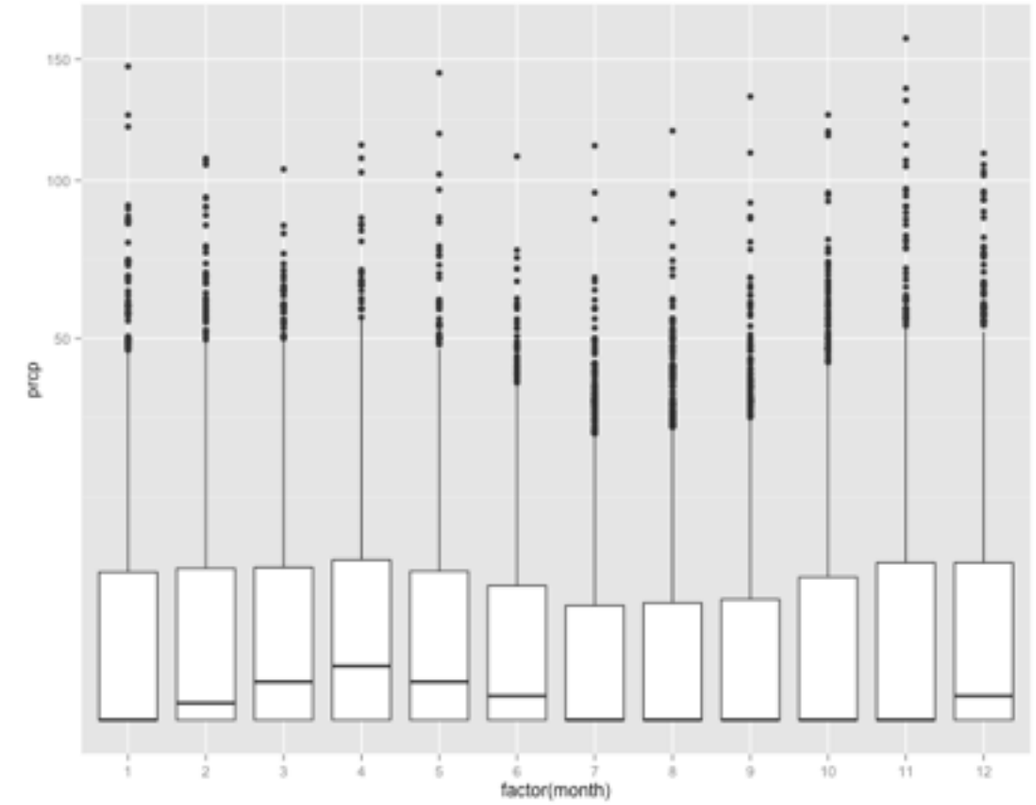


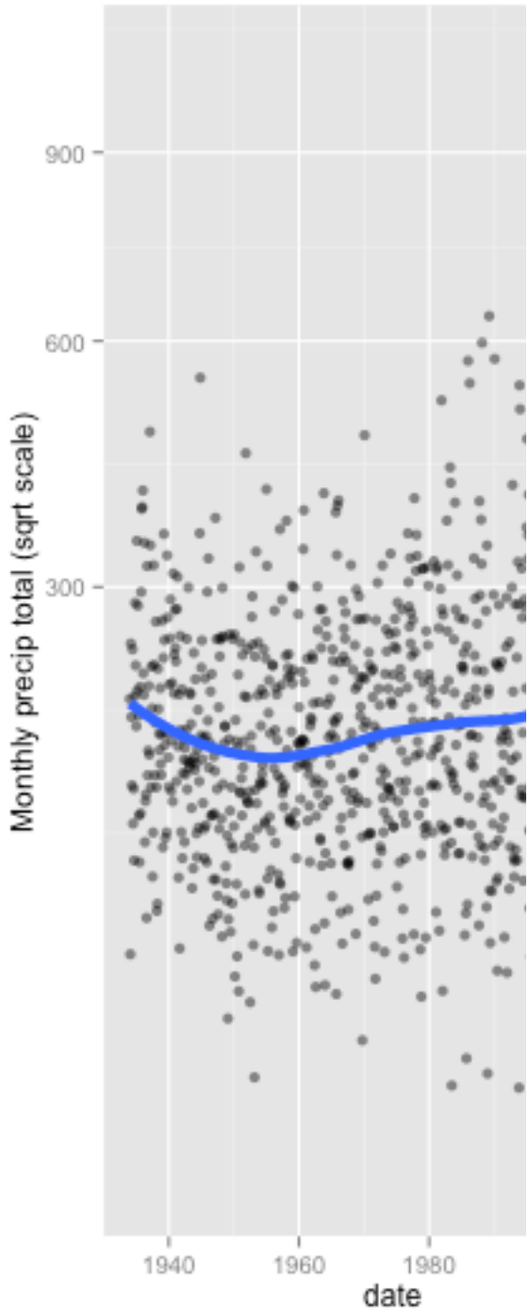


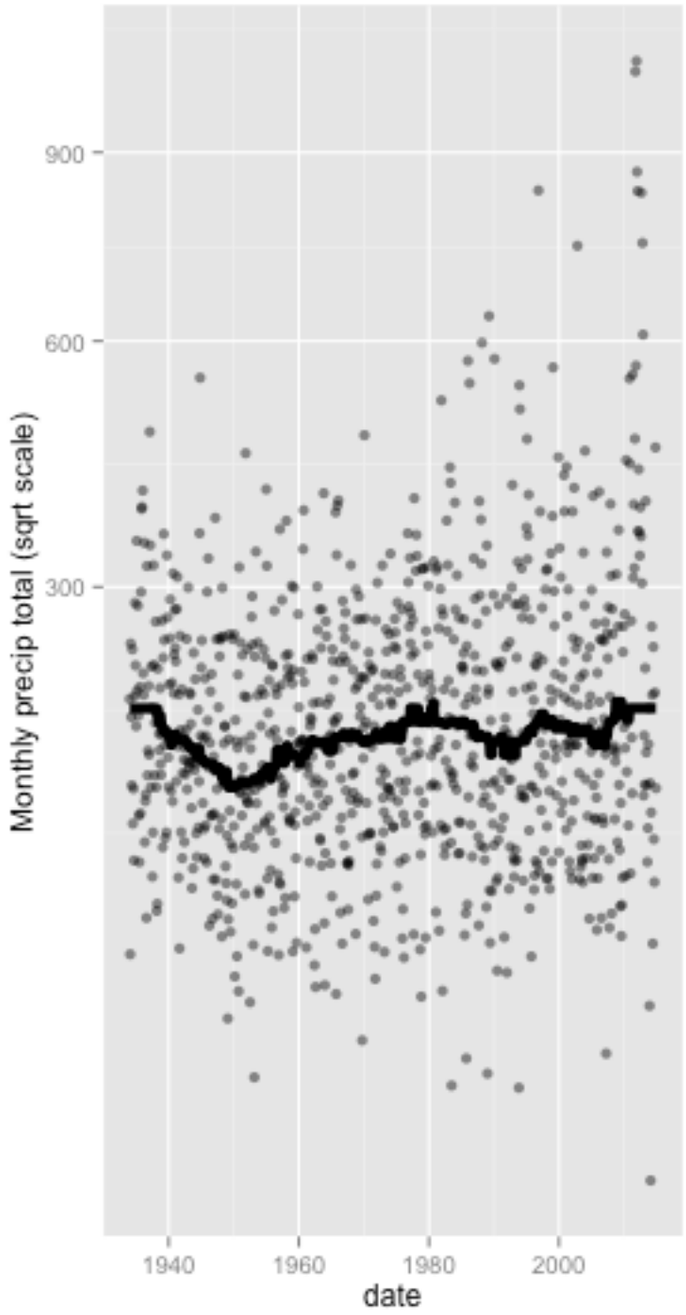
Small gap in measurement collections. Change to only look at latter measurements.

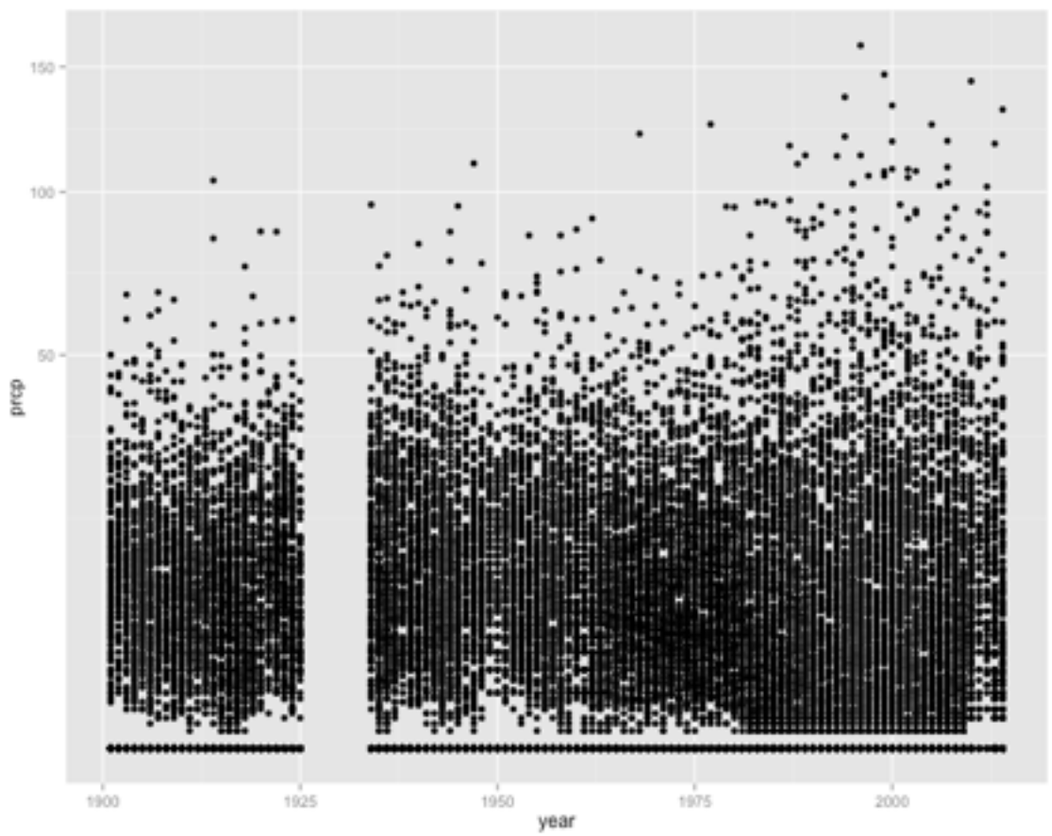
And compute monthly totals (adjusted for missing values).

Precipitation (sqrt scale)



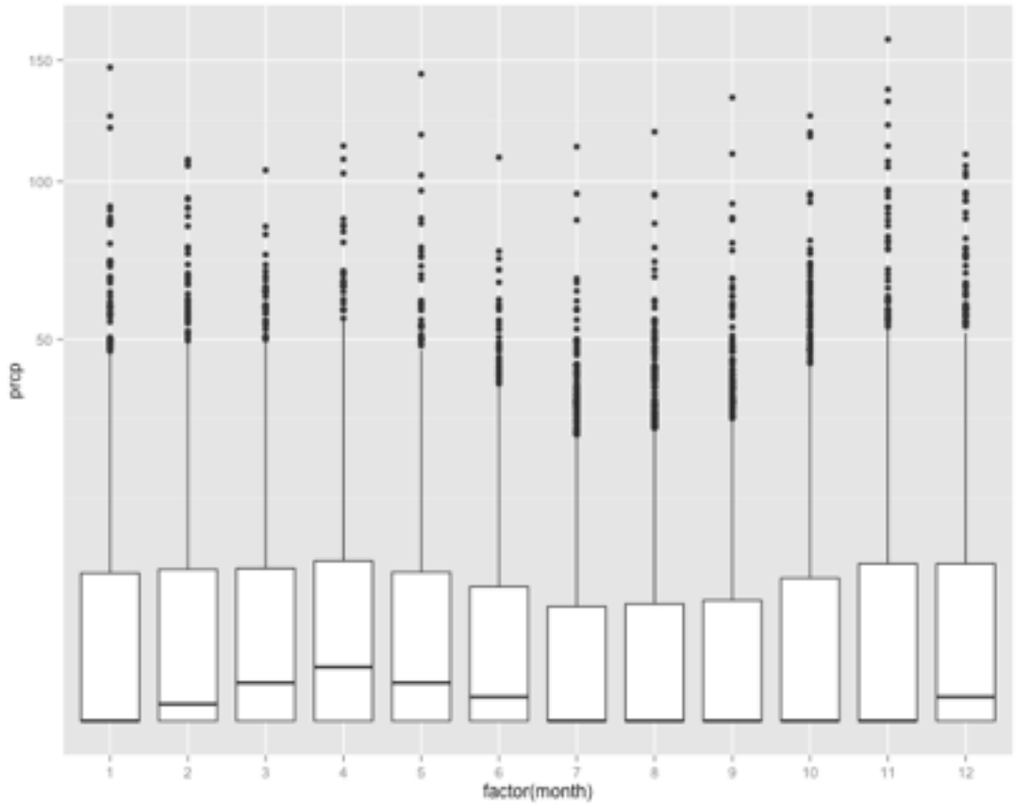


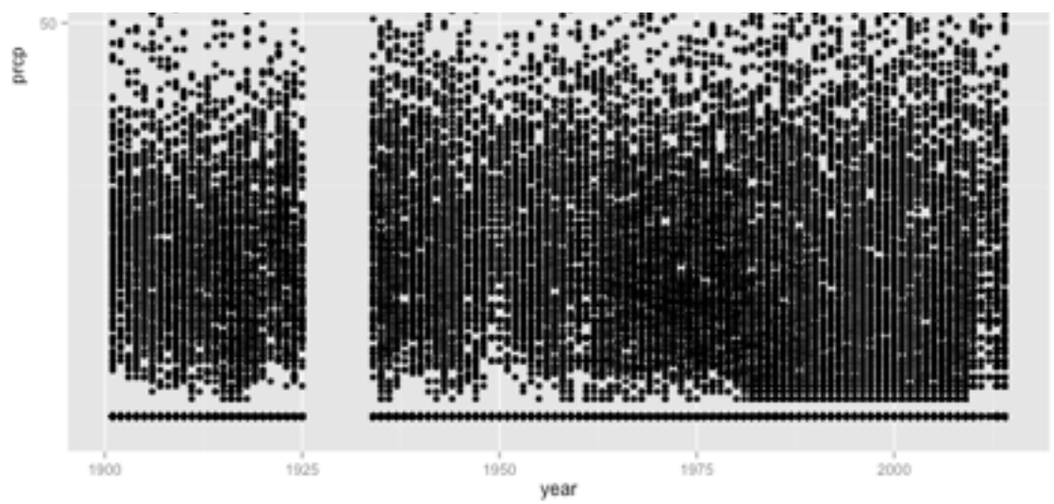




Are extremes becoming more common?

Precipitation (sqrt scale)





What happened in 2012?

Your Turn

Take two minutes to brainstorm with your neighbors.

- Based on what we have seen with min/max temperature and precip what might you do to find additional information that supports or refutes?
- What data, calculations, plots would you make?

Summary

- Pulling data to check things for yourself is so much easier today. One of the most useful skills you can attain!
- Graphics for exploratory data analysis are ephemeral. Once created, once things are learned, they evaporate, and need to be replaced with something overly produced and beautiful for general consumption.

Acknowledgements

- Code is available at:

  `https://github.com/dicook/lesdiablerets-code`

- Software used: R (http://www.R-project.org) and primarily the package ggplot2

Coming next...

- Do you like tennis?
- Data doesn't come with only two variables. How do you see in high dimensions?
- How can you quantify if what we see is "real"?