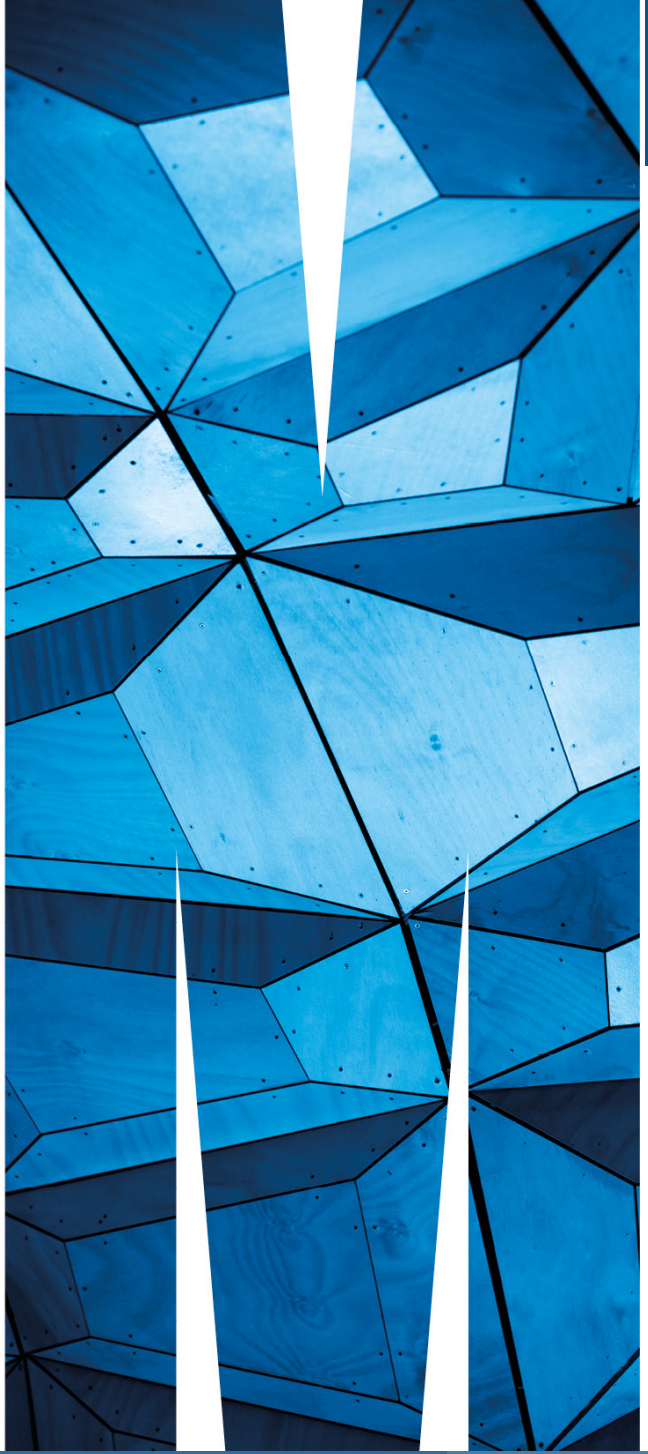
MONASH University

MONASH BUSINESS SCHOOL

ETC3250: Data visualisation

Week 6, class 1

Professor Di Cook, Econometrics and Business Statistics



GROUP OF EIGHT AUSTRALIA

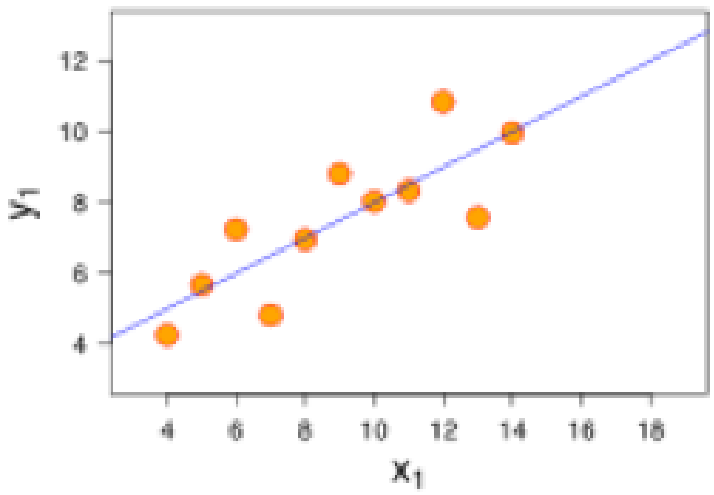
AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

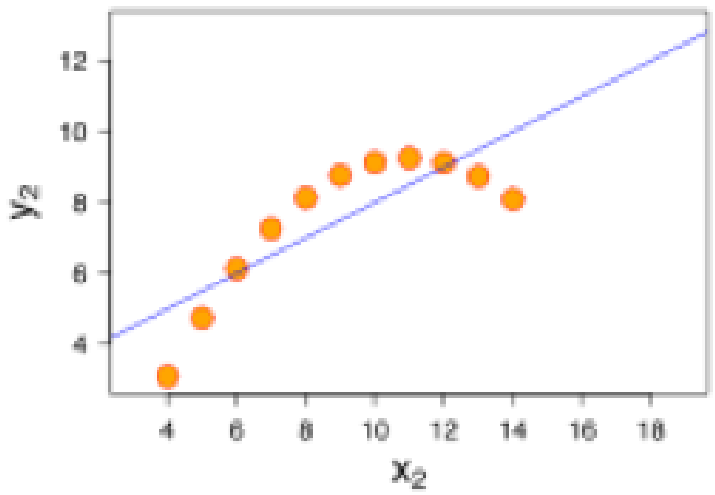
Your turn

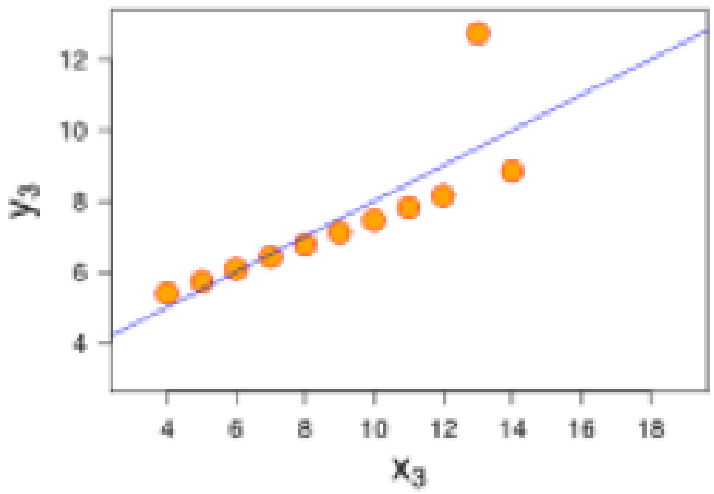
Why make a plot of data??

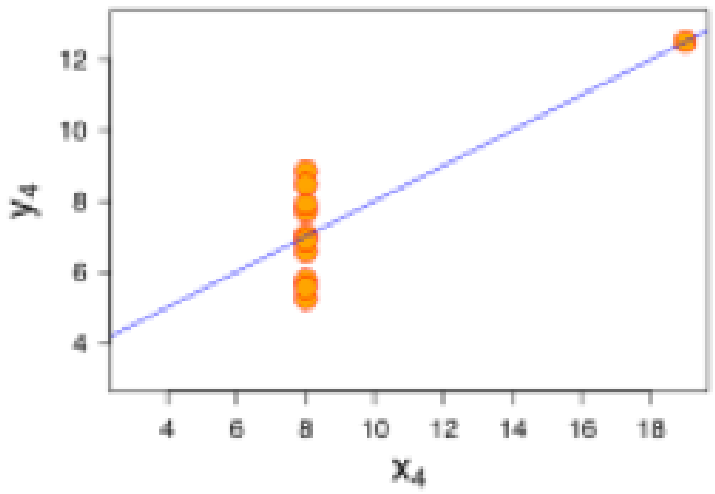
Anscombe's quartet

All of these have the same intercept, slope, correlation.







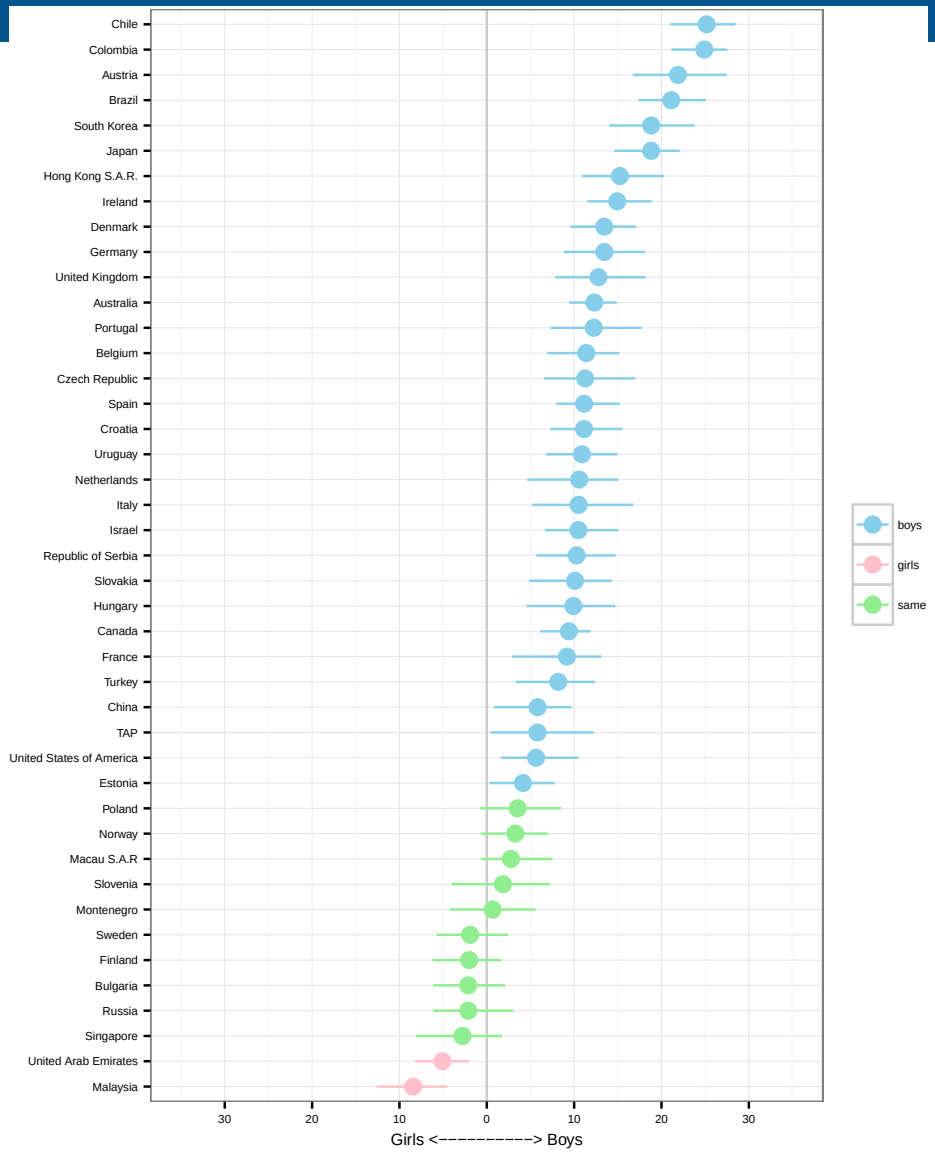


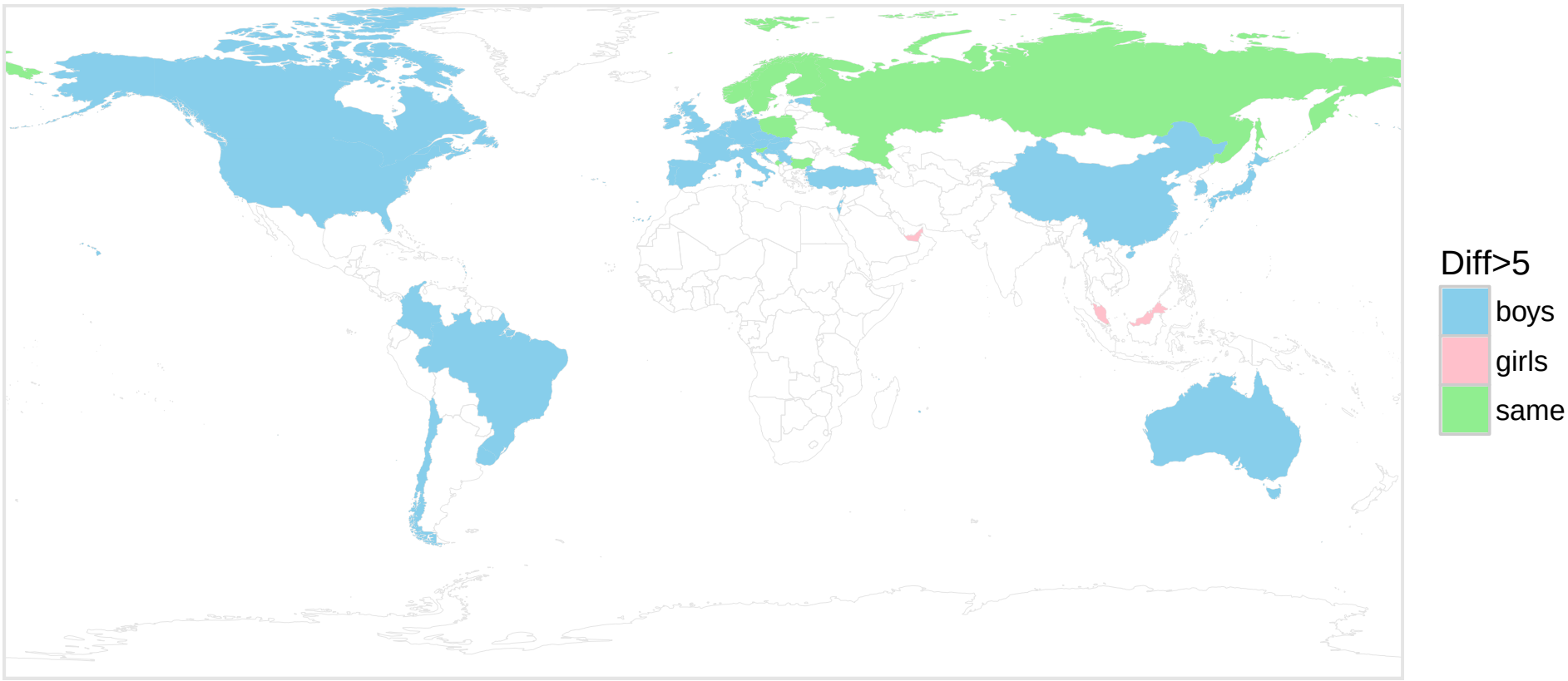
Examples from my own work

- Education
- Climate change
- US Election polls
- Airline traffic patterns
- Wages

Education: OECD PISA

- OECD PISA survey “the world’s global metric for quality, equity and efficiency in school education“.
- Workforce readiness of 15-year old students
- 500,000 students were tested across 65 countries and 18,000 schools
- Math, reading and science
- Data available from http://www.oecd.org/pisa


Chile
Colombia
Austria
Brazil
South Korea
Japan
Hong Kong S.A.R.
Ireland
Denmark
Germany
United Kingdom
Australia
Portugal
Belgium
Czech Republic
Spain
Croatia
Uruguay
Netherlands
Italy
Israel
Republic of Serbia
Slovakia
Hungary
Canada
France
Turkey
China
TAP
United States of America
Estonia
Poland
Norway
Macau S.A.R
Slovenia
Montenegro
Sweden
Finland
Bulgaria
Russia
Singapore
United Arab Emirates
Malaysia
30
20
10
0
10
20
30
Girls <----------> Boys
boys
girls
same

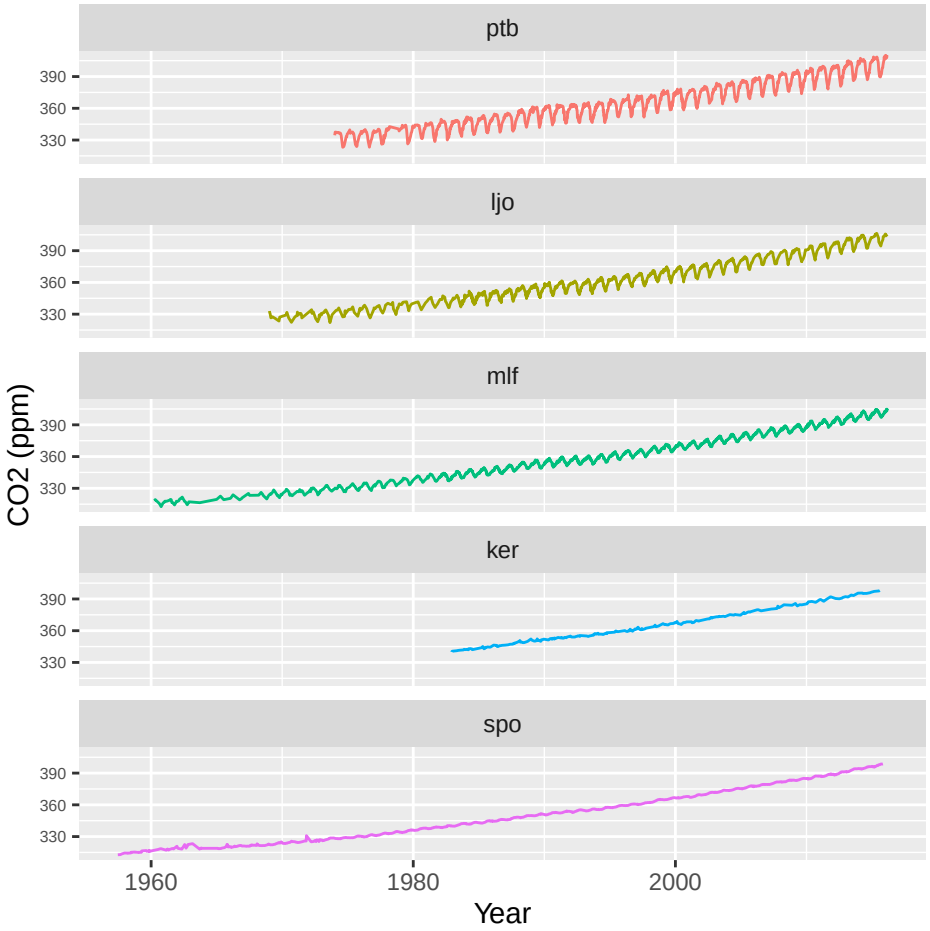

Diff>5
boys
girls
same

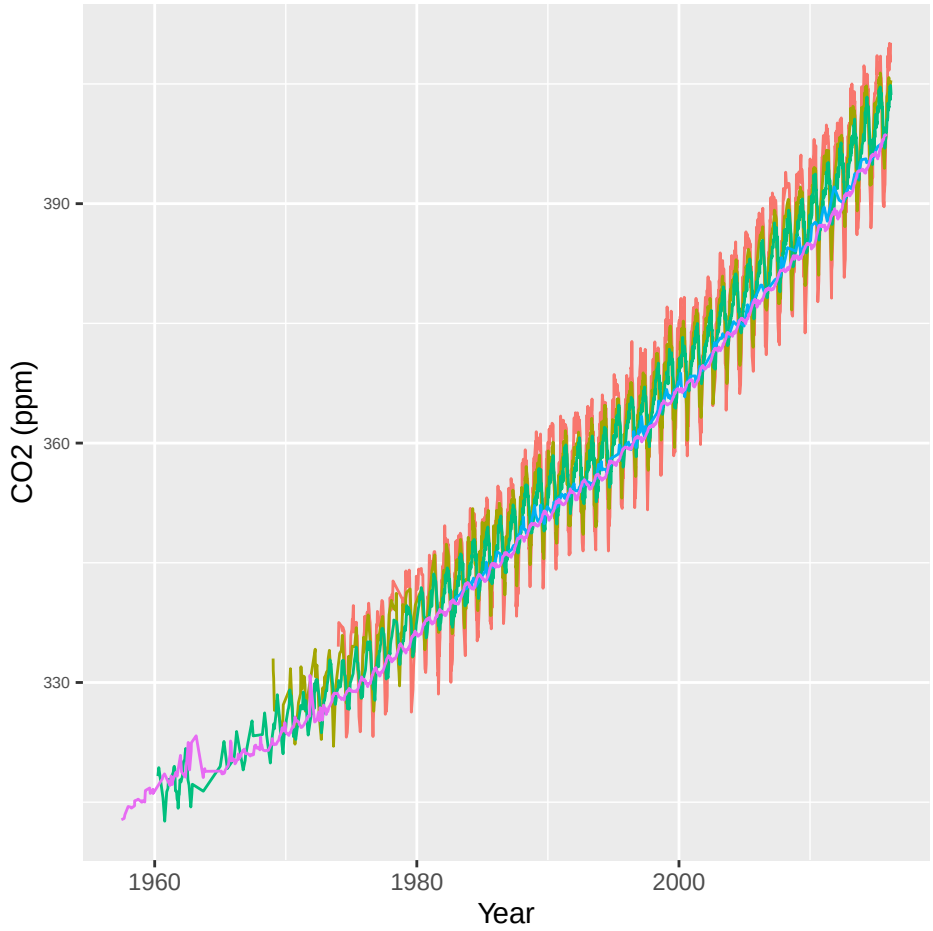
What do we learn?

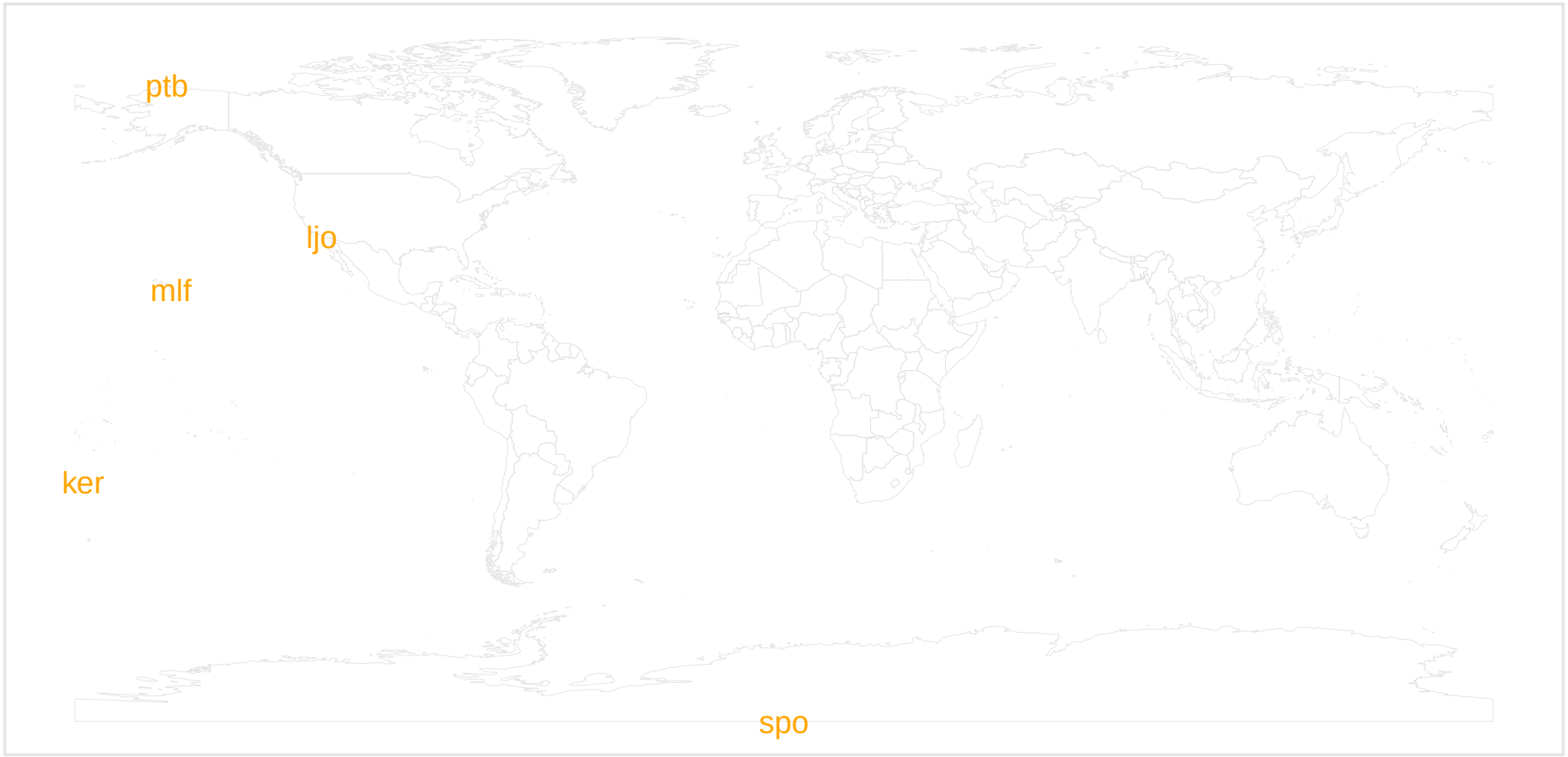
- Math gender gap is not universal
- Many countries have no substantial difference
- Reverse gap exists in surprising places
- Australia has a 10 point gap (10 points out of 1000 points)
- Individuals show different pattern, highest math score in US is by a girl
- Australia has a huge variation in scores, one of the highest countries, but also one of the lowest countries
- Reading gap is universal in favour of girls

Carbon dioxide data

- Data is collected at a number of locations world wide.
- See Scripps Inst. of Oceanography
- Let's pull the data from the web and take a look ...
-
- Recordings from South Pole (SPO), Kermadec Islands (KER), Mauna Loa Hawaii (MLF), La Jolla Pier, California (LJO), Point Barrow, Alaska (PTB).


ptb
ljo
mlf
ker
spo
390
360
330
CO2 (ppm)
1960
1980
2000
Year


390
360
330
CO2 (ppm)
1960
1980
2000
Year

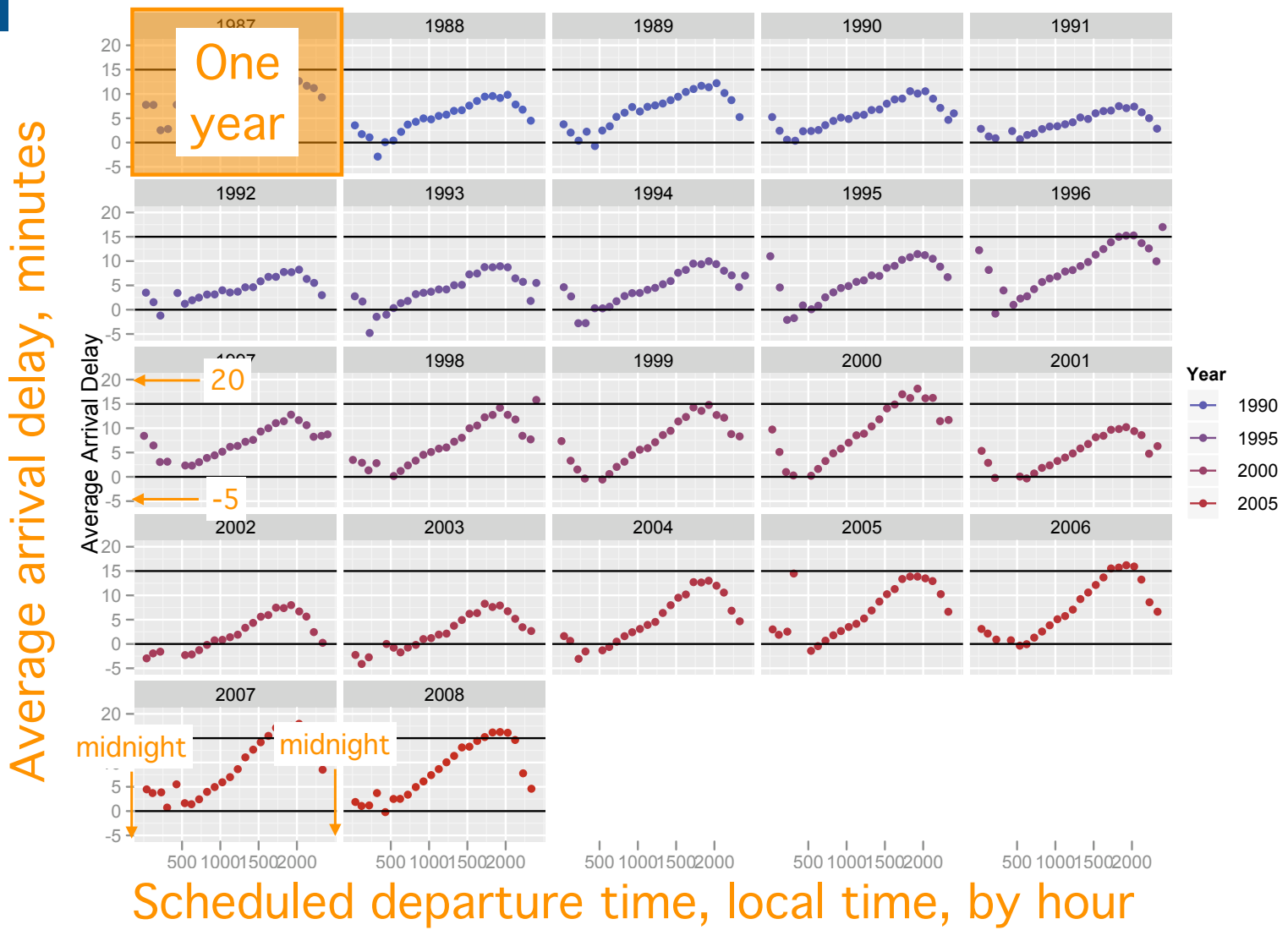

ptb
ljo
mlf
ker
spo

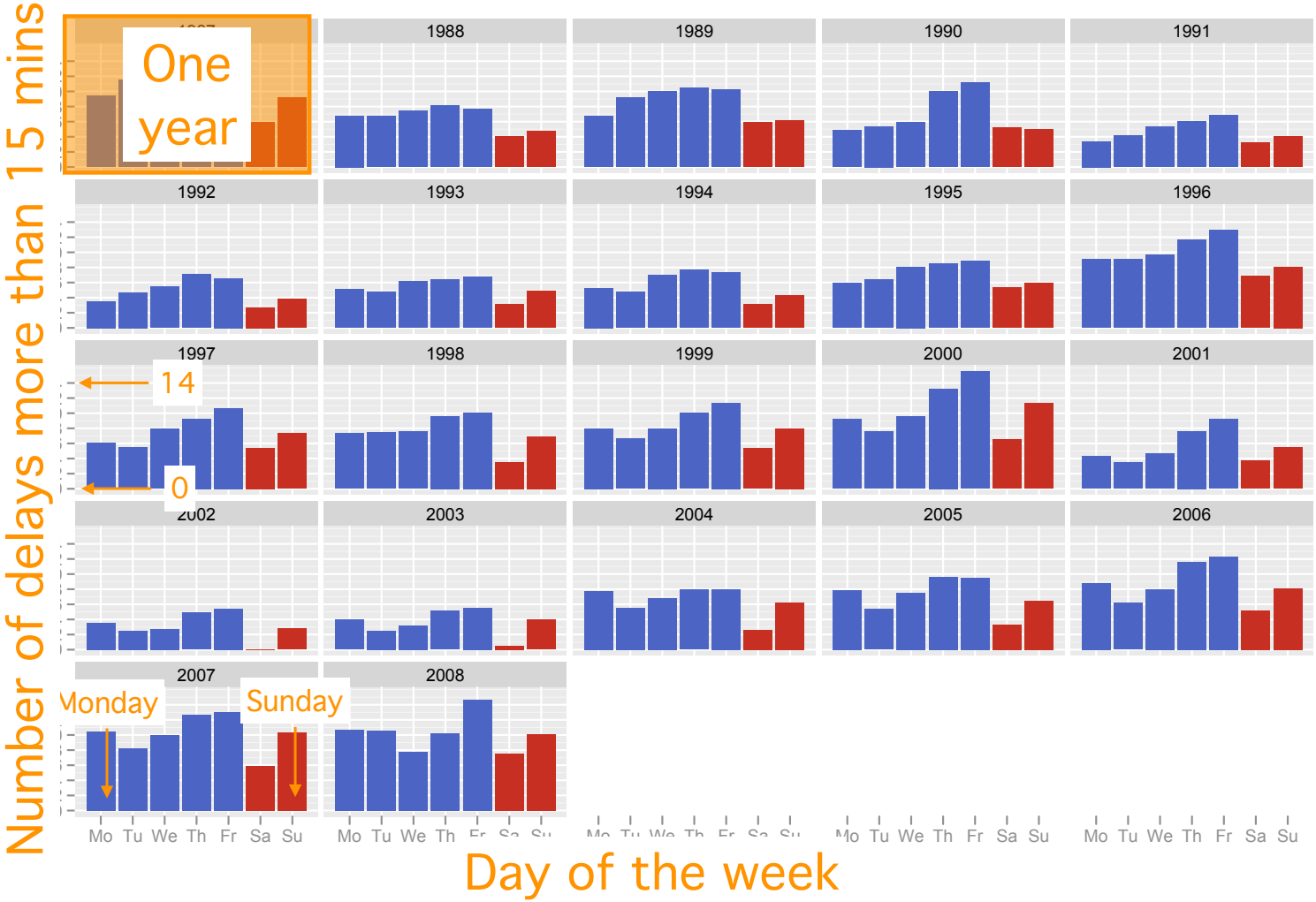
What do we learn?

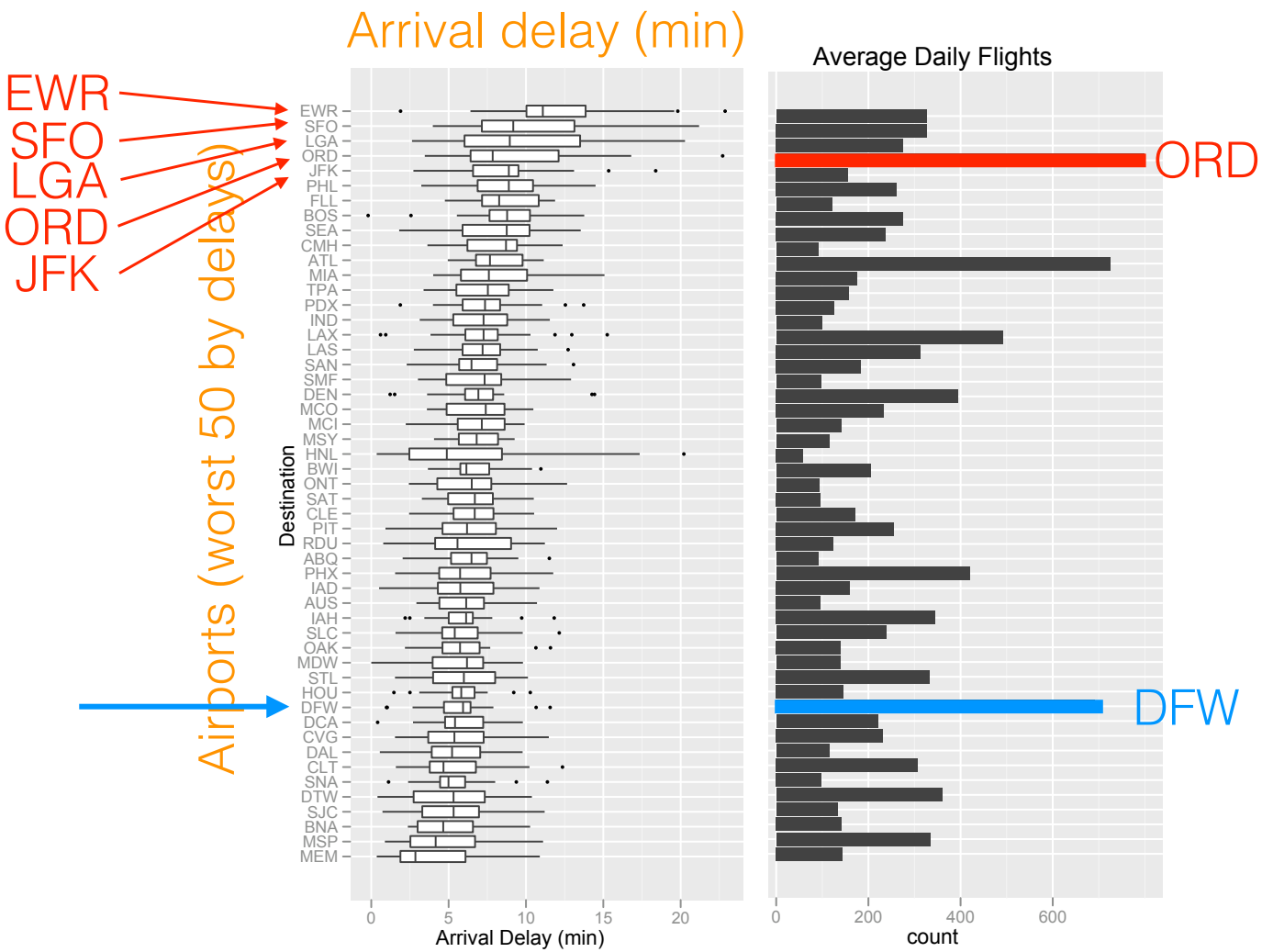
- CO_2 is increasing, and it looks like it is exponential increase. **I really expected that the concentration would have flattened out with all of the efforts to reduce carbon emissions.**
- The same trend is seen at every location - REALLY? Need some physics to understand this.
- Some stations show seasonal pattern - actually the more north the more seasonality - WHY?

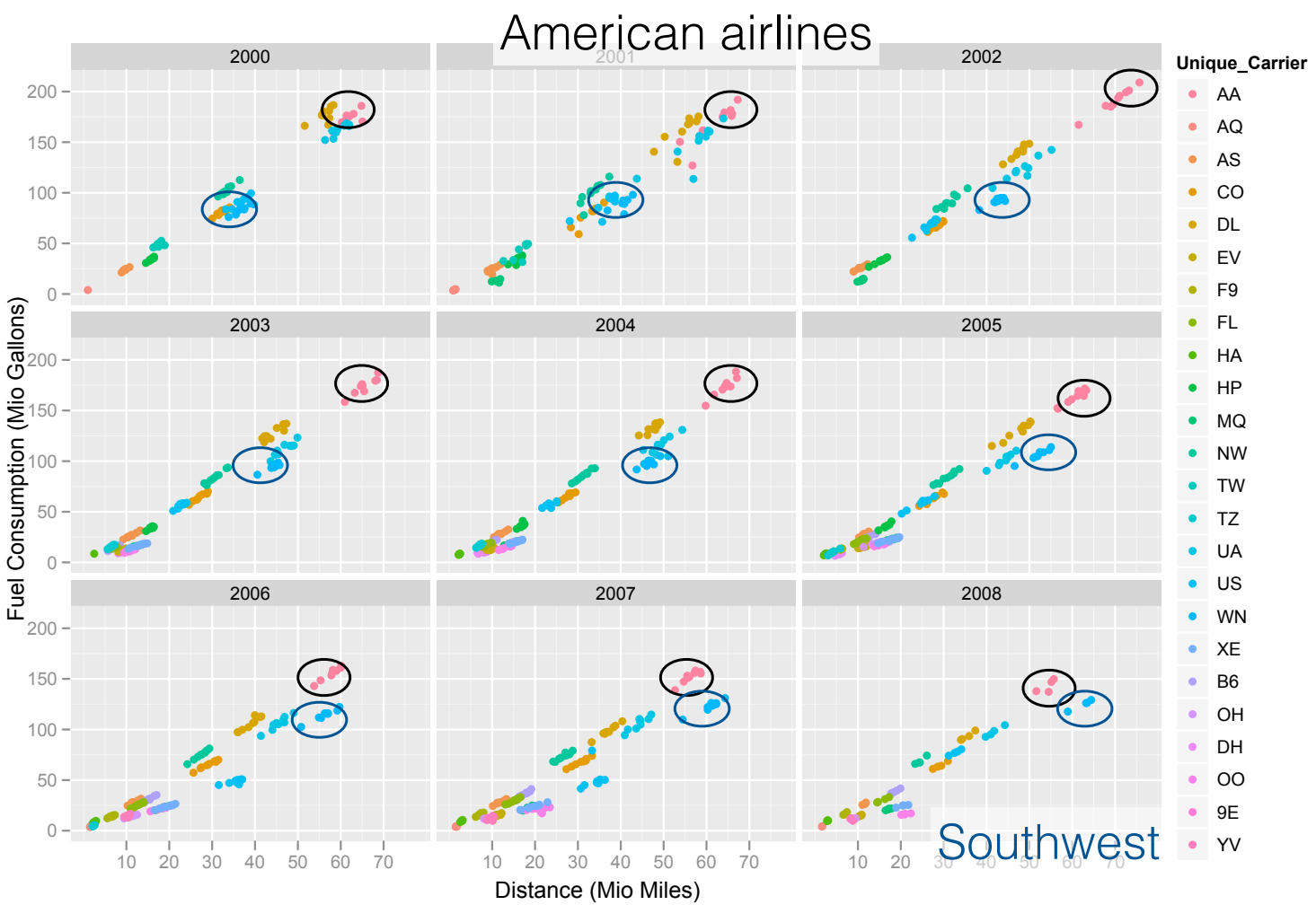
US Airline traffic

- ~15,000 flights a day
- April 1986 - present (2008)
- RITA - Research and Innovative Technology Administration (flight information, arrival delay, airline, plane id, . . .)
- On time performance database - `http://www.transtats.bts.gov/` - yes, you can download this yourself
- Analysis code examples on `https://github.com/heike/data-technologies`


Average arrival delay, minutes
One year
1988
1989
1990
1991
1992
1993
1994
1995
1996
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Average Arrival Delay
20
-5
midnight
midnight
Year
1990
1995
2000
2005
500 1000 1500 2000
Scheduled departure time, local time, by hour


Number of delays more than 15 mins
One year
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
14
0
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Monday
Sunday
2008
Mo Tu We Th Fr Sa Su
Day of the week


Arrival delay (min)
Average Daily Flights
EWR
SFO
LGA
ORD
JFK
Airports (worst 50 by delays)
Destination
EWR
SFO
LGA
ORD
JFK
PHL
FLL
BOS
SEA
CMH
ATL
MIA
TPA
PDX
IND
LAX
LAS
SAN
SMF
DEN
MCO
MCI
MSY
HNL
BWI
ONT
SAT
CLE
PIT
RDU
ABQ
PHX
IAD
AUS
IAH
SLC
OAK
MDW
STL
HOU
DFW
DCA
CVG
DAL
CLT
SNA
DTW
SJC
BNA
MSP
MEM
0
5
10
15
20
Arrival Delay (min)
0
200
400
600
count
ORD
DFW


American airlines
2000
2001
2002
2003
2004
2005
2006
2007
2008
Fuel Consumption (Mio Gallons)
Distance (Mio Miles)
Unique_Carrier
AA
AQ
AS
CO
DL
EV
F9
FL
HA
HP
MQ
NW
TW
TZ
UA
US
WN
XE
B6
OH
DH
OO
9E
YV
Southwest

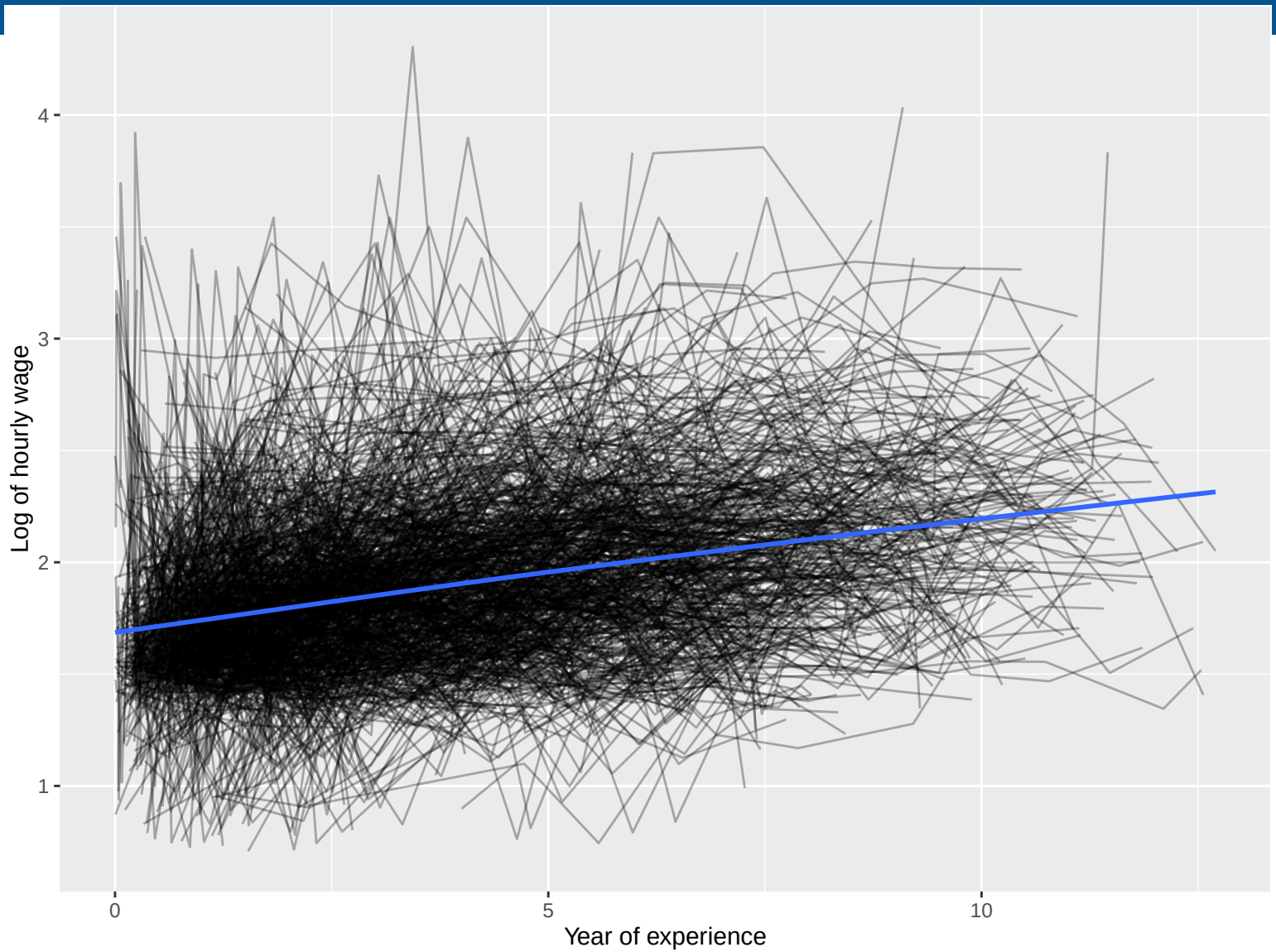
What did we learn?

- Fly early in the day, early in the week or weekends (Saturday)
- Avoid ORD, JFK, LGA, EWR
- American Airlines filed for bankruptcy Nov 29, 2013. Mining publicly available data could have sounded the alarms several years in advance

Wages

- 6402 observations on 888 high school dropouts, 1990-2002

```
# Observations: 6,402
# Variables: 15
# $ id           <fctr> 31, 31, 31, 31, 31, 31, 31, 31, 36, 3
# $ lnw          <dbl> 1.491, 1.433, 1.469, 1.749, 1.931, 1.7
# $ exper        <dbl> 0.015, 0.715, 1.734, 2.773, 3.927, 4.9
# $ ged          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
# $ postexp      <dbl> 0.015, 0.715, 1.734, 2.773, 3.927, 4.9
# $ black        <fctr> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
# $ hispanic     <fctr> 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0
# $ hgc          <int> 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9,
# $ hgc.9        <int> -1, -1, -1, -1, -1, -1, -1, -1, 0, 0,
# $ uerate       <dbl> 3.21, 3.21, 3.21, 3.30, 2.89, 2.49, 2.
# $ ue.7         <dbl> -3.785, -3.785, -3.785, -3.705, -4.105
# $ ue.centert1  <dbl> 0.000, 0.000, 0.000, 0.080, -0.320, -0
# $ ue.mean      <dbl> 3.21, 3.21, 3.21, 3.21, 3.21, 3.21, 3.
```


Log of hourly wage
4
3
2
1
0
5
10
Year of experience

Video

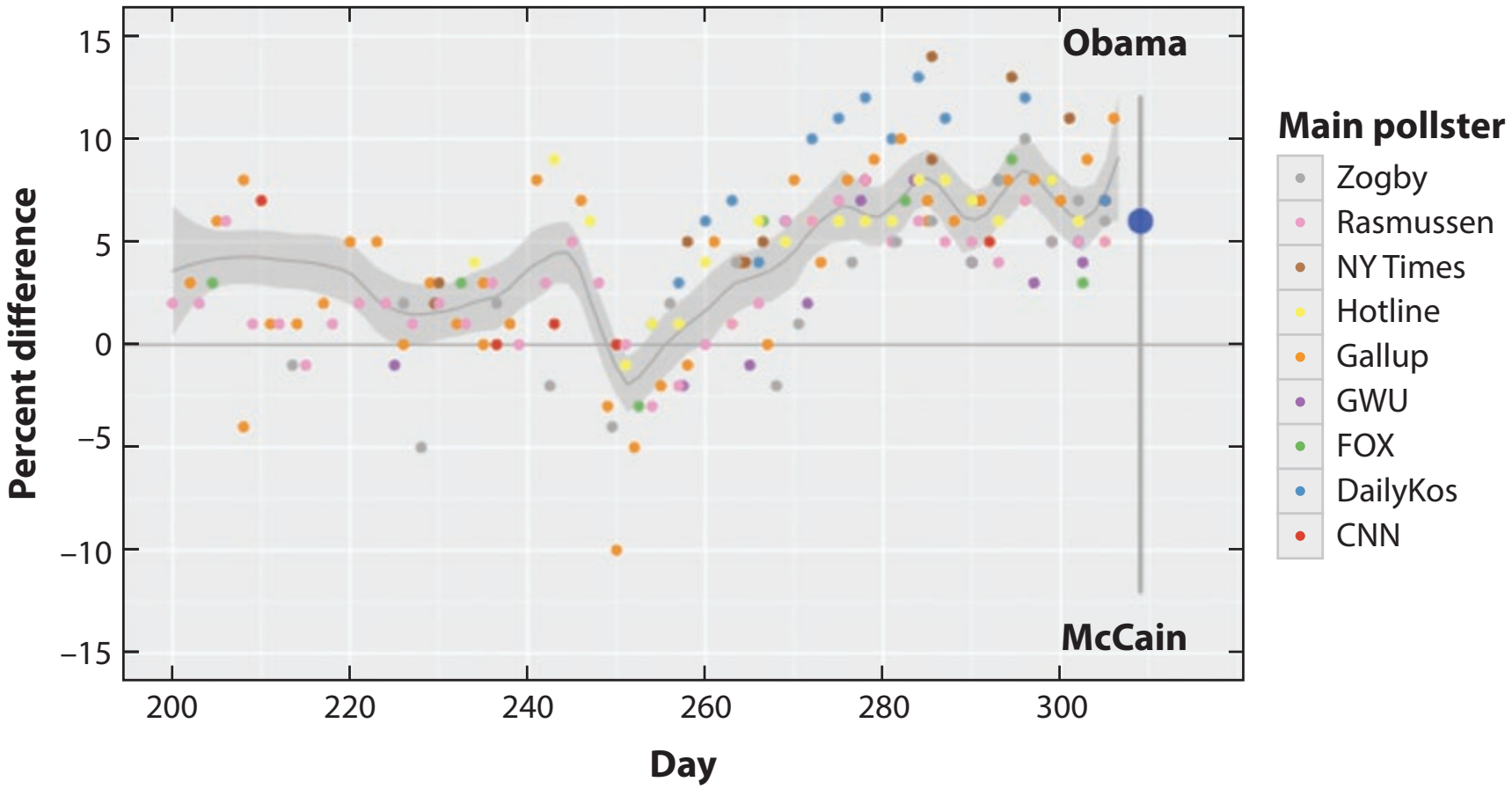
(wages-increasing.mov)

Video

(wages-decreasing.mov)

US Election

- They are in the middle of another LONG election
- Its coming to a close
- There is a lot of information about how people might vote
- We looked at how things progressed in 2008 election, in the months leading up to the vote
- We used web scrapers to pull polling data off web sites



Pollsters operating in the US are not all impartial.

Your turn

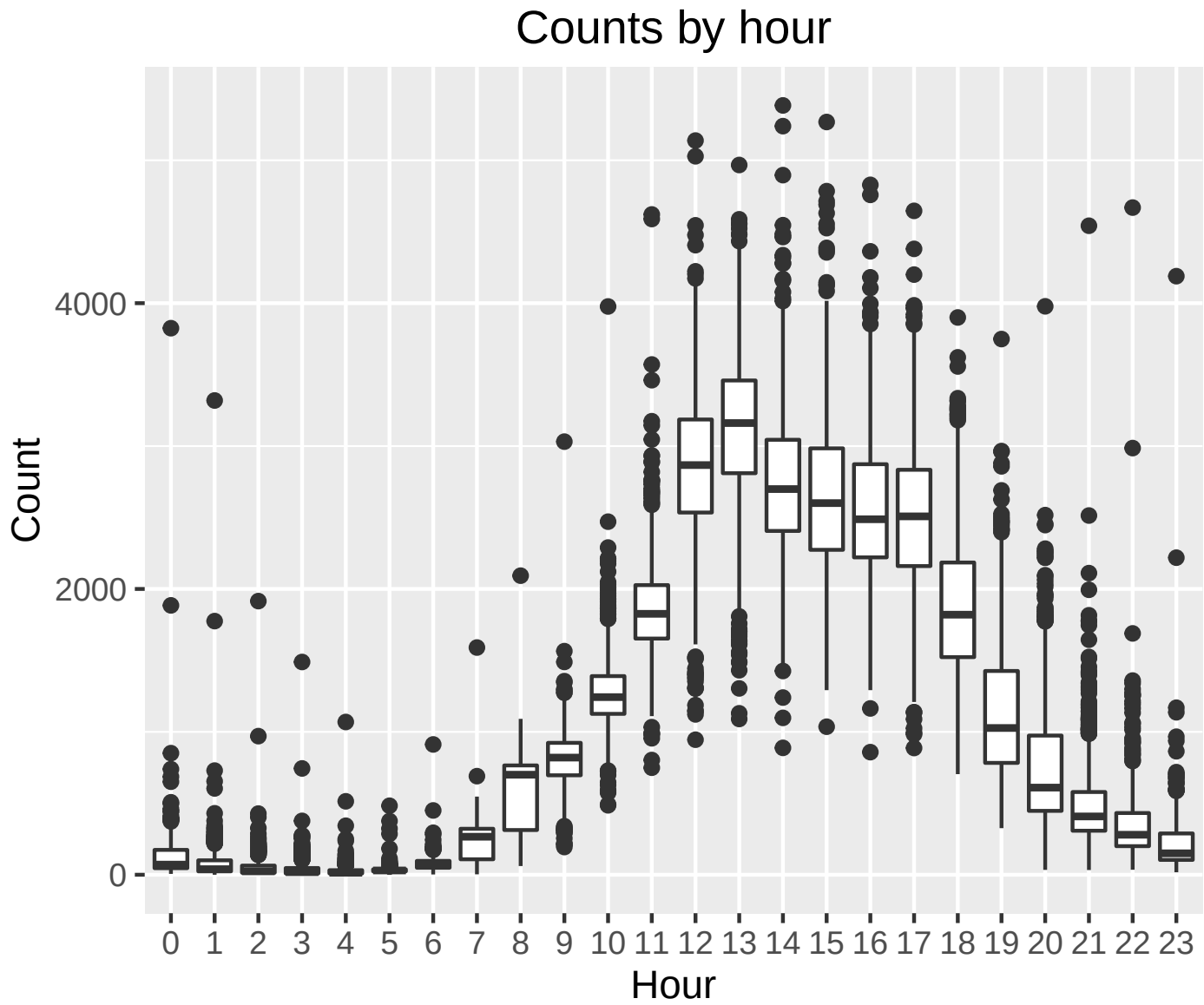
Why make a plot of data??

Your turn

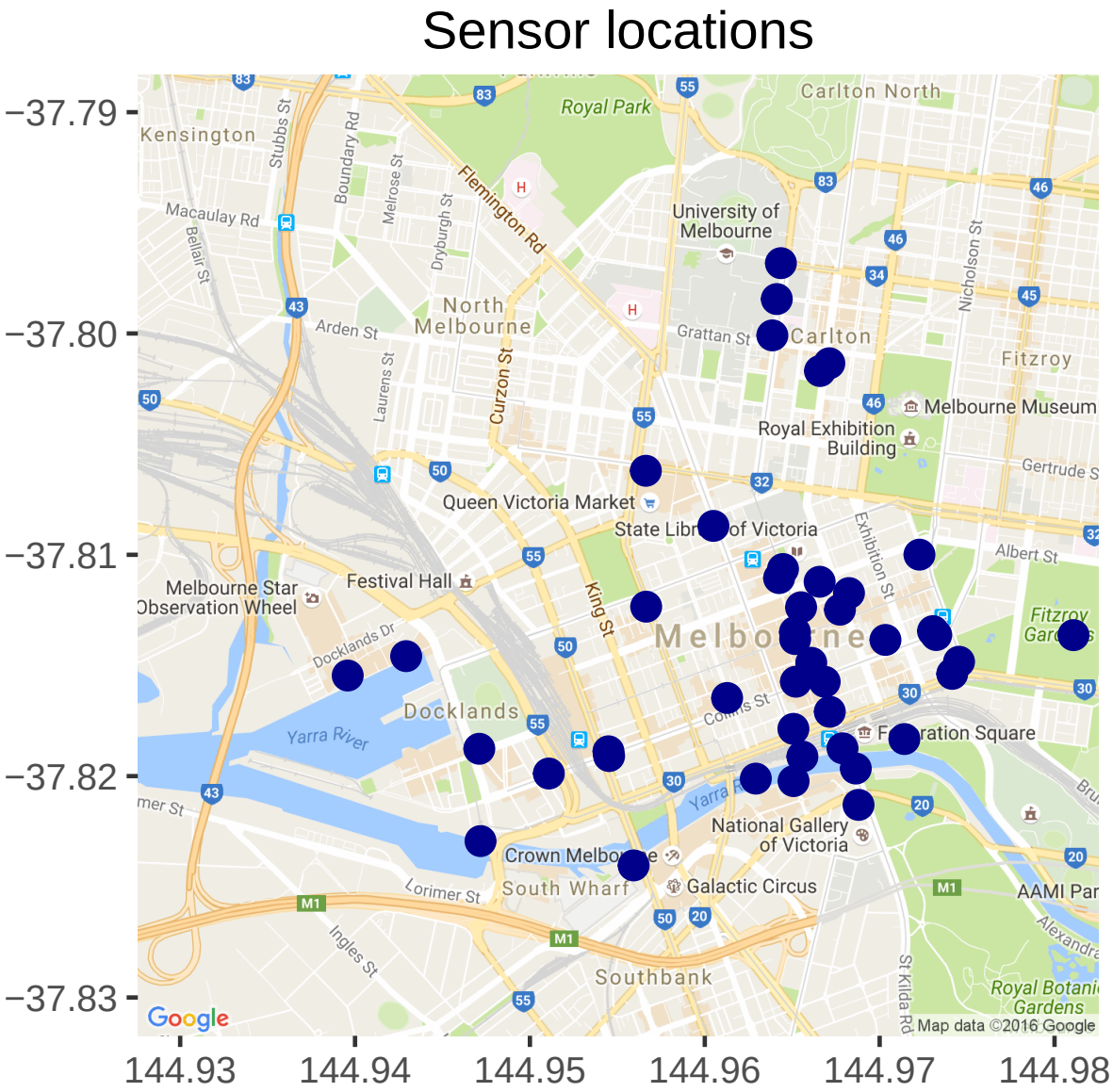
- What is a (data) plot?
- What are the three most important data plots?
- Is a plot a statistic?

Your turn

How would you describe this plot?



What about this one?



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

Why use a grammar of graphics?

Variable in the data is directly mapped to an element in the plot

Resources

- Cheat sheet
- ggplot2: Elegant Graphics for Data Analysis, Hadley Wickham, web site
- R Graphics Cookbook, Winston Chang
- Naomi Robbins, Creating More Effective Graphs
- Antony Unwin, Graphical Data Analysis with R

Share and share alike