ETC5521: Exploratory Data Analysis

Making comparisons between groups and strata

Lecturer: *Emi Tanaka*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 7 - Session 1

Case study 1 Pest resistance maize Part 1/2

- Pests, like thrips and spiders, damage maize crops. Note: Maize = Corn
- One strategy to protect crops against pests is to cultivate *genetically modified* (GM) maize that expresses a toxic protein.

Thrips	Spiders
16.6	0.8
16.4	0.8
11.0	0.6
16.8	0.4
10.6	0.6
18.4	0.8
14.2	0.0
10.2	0.6

- The species abundance on 8 *Bt* GM maize is shown.
- Is the strategy working?
- Well it didn't completely eliminate pests but *did it lower the abundance?*
- We can't tell without knowing what the typical abundance is.

> *At the heart of quantitative reasoning is a single question:*
> ***Compared to what?***
>
> *-Edward Tufte*

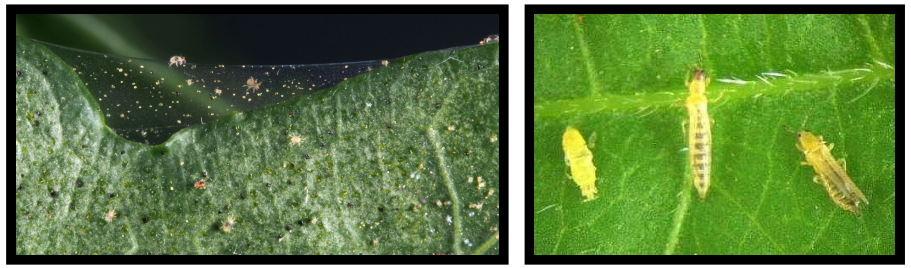
Kevin Wright (2018). agridat: Agricultural Datasets. R package version 1.16.
L. A. Hothorn, 2005. Evaluation of Bt-Maize Field Trials by a Proof of Safety. http://www.seedtest.org/upload/cms/user/presentation7Hothorn.pdf

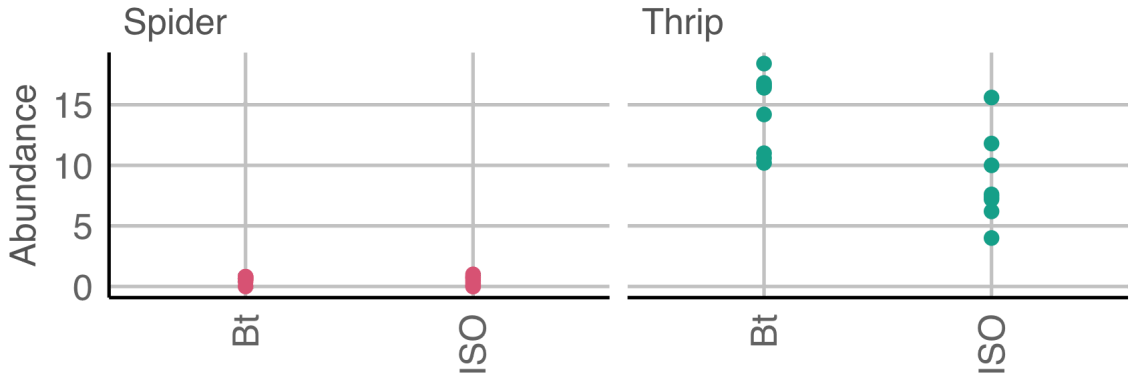
Case study 1 Pest resistance maize Part 2/2

- The actual experiment compared *Bt* variety to the **isogenic control variety**.
- How would you compare graphically?

Comparing like-with-like Part 1

Comparison should be fair - any differences should be due to the factor you wish to investigate.





Comparable populations and measurements

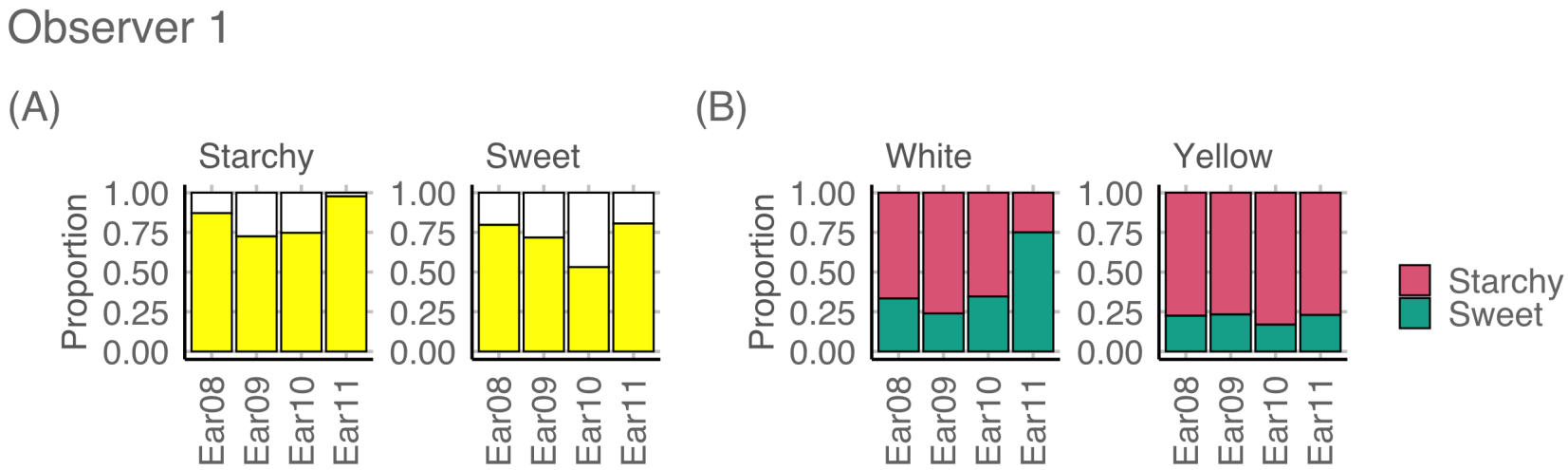
- Abundance is measured for two species: spiders and thrips.
- The abundance metric differ between species.
- Would you compare the abundance of spiders on a *Bt* variety to the abundance of thrips on a isogenic variety?

Image source: https://www.infonet-biovision.org/PlantHealth/Pests/Thrips and https://cropwatch.unl.edu/2016/managing-spider-mites-corn-and-soybean

Case study 2 Maize kernels Part 1/2

1. Plant pathologist
2. Associate plant pathologist
3. Professor of agronomy
4. Assistant professor of agronomy
5. Professor of philosophy
6. Biologist
7. Biologist (also author)
8. Assistant in biology
9. Data entry clerk (a.k.a. "Computer")
10. Farmer
11. Professor of plant physiology
12. Instructor in plant physiology
13. Assistant in plant physiology
14. Assistant in plant physiology
15. Professor of biology

- 4 maize ears selected.
- 15 observers asked to classify kernels to (i) starchy yellow, (ii) starchy white, (iii) sweet yellow or (iv) sweet white.
- Ear 11 was slightly abnormal due to a fungus attack giving some pinkish tinge to some kernels.
- Is Ear 11 different?

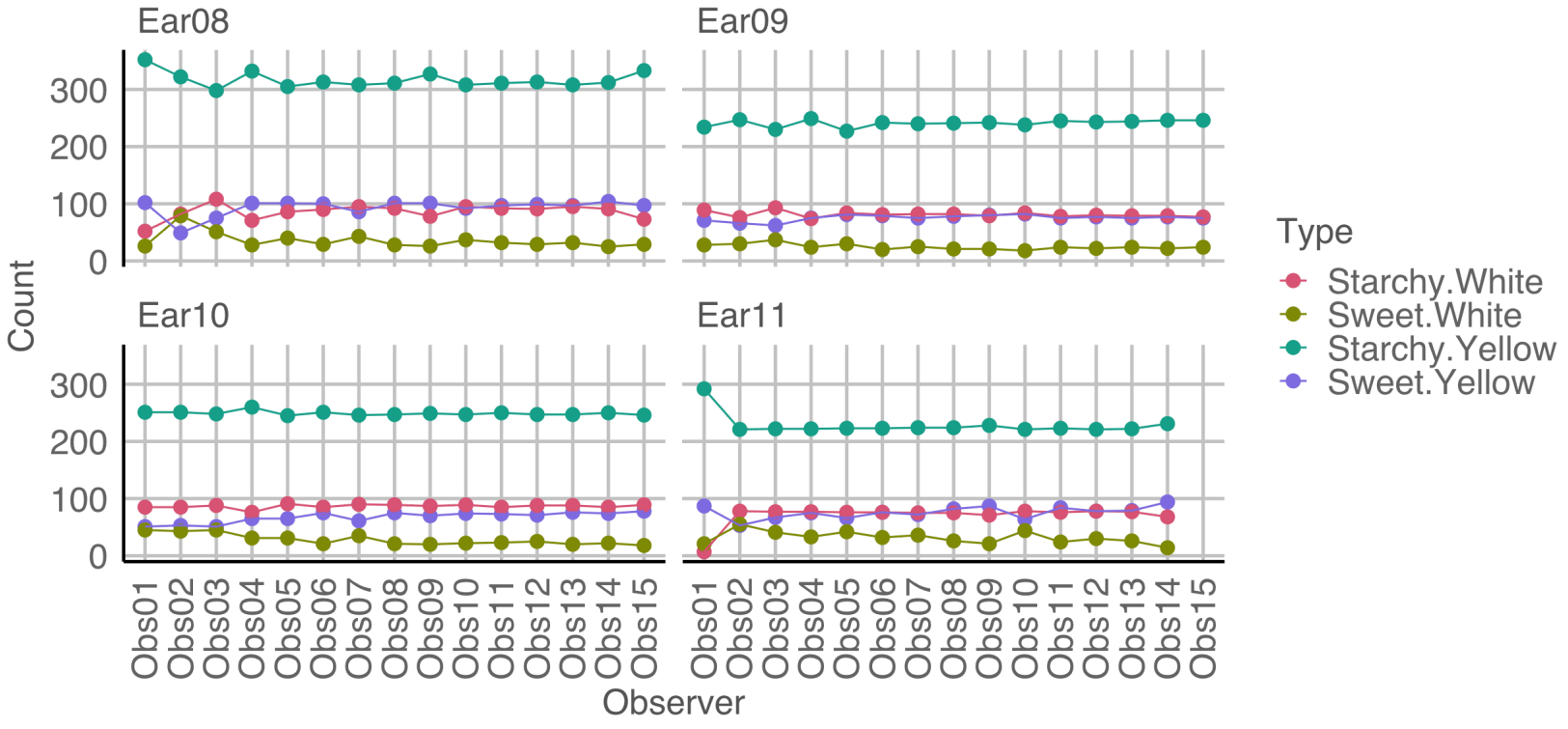


Raymond Pearl, 1911. The Personal Equation In Breeding Experiments Involving Certain Characters of Maize, Biol. Bull., 21, 339-366.

Case study 2 Maize kernels Part 2/2

1. Plant pathologist
2. Associate plant pathologist
3. Professor of agronomy
4. Assistant professor of agronomy
5. Professor of philosophy
6. Biologist
7. Biologist (also author)
8. Assistant in biology
9. Data entry clerk (a.k.a. "Computer")
10. Farmer
11. Professor of plant physiology
12. Instructor in plant physiology
13. Assistant in plant physiology
14. Assistant in plant physiology
15. Professor of biology

- All observer are counting the kernels of the same ears, however there are variations across observers.
- Notice Observer 1 classifies more kernels as yellow for Ear 11.

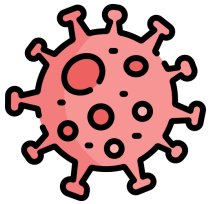


Comparing like-with-like Part 2

Comparable conditions

- The variability due to other sources need to be accounted, removed or "averaged" out for a fair comparison.

Comparing like-with-like Part 3



Comparable variables and sources

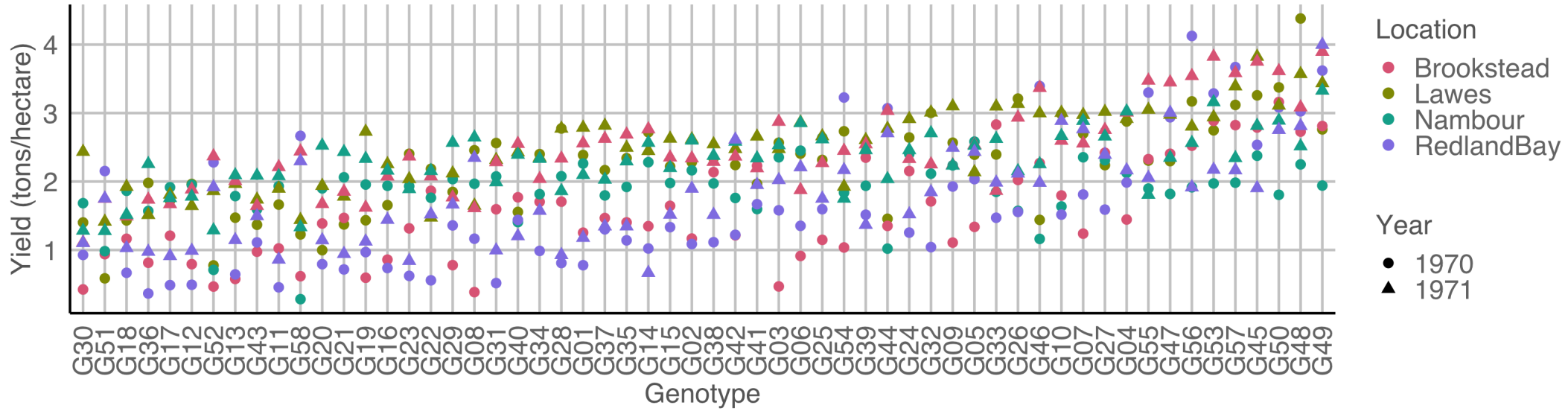
- Data collected by different sources may have different rules. E.g. in Australia, "a COVID-19 death is defined for surveillance purposes as a death in a probable or confirmed COVID-19 case, unless there is a clear alternative cause of death that cannot be related to COVID19 (e.g. trauma)"[1]
- Do other countries use the same definition?
- The COVID-19 death data often have delays in reporting and data would be updated or corrected later.

[1] https://www.health.gov.au/resources/total-covid-19-cases-and-deaths-by-states-and-territories

Case study 3 Multi-environment soybean trial

- 58 soy varieties are grown in four locations in Queensland in 1970 then 1971.
- Soy breeders are interested in finding the "best" variety.
- So which variety is the best for yield?

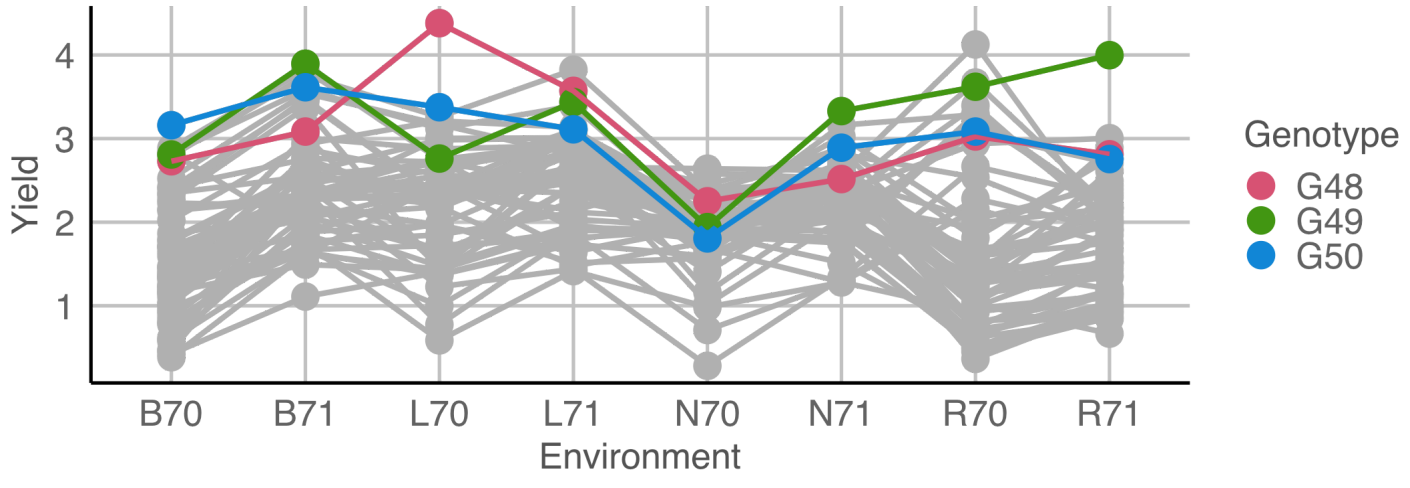
data R



Types of comparison

- Is the interest to find the best variety *for a location*?
- In that case, should the comparison be within location?
- Or is the interest to find the overall best variety *at any location*?
- Comparisons may be specific or general.
- A different type of comparison may require a different calculation or graphic for investigation.

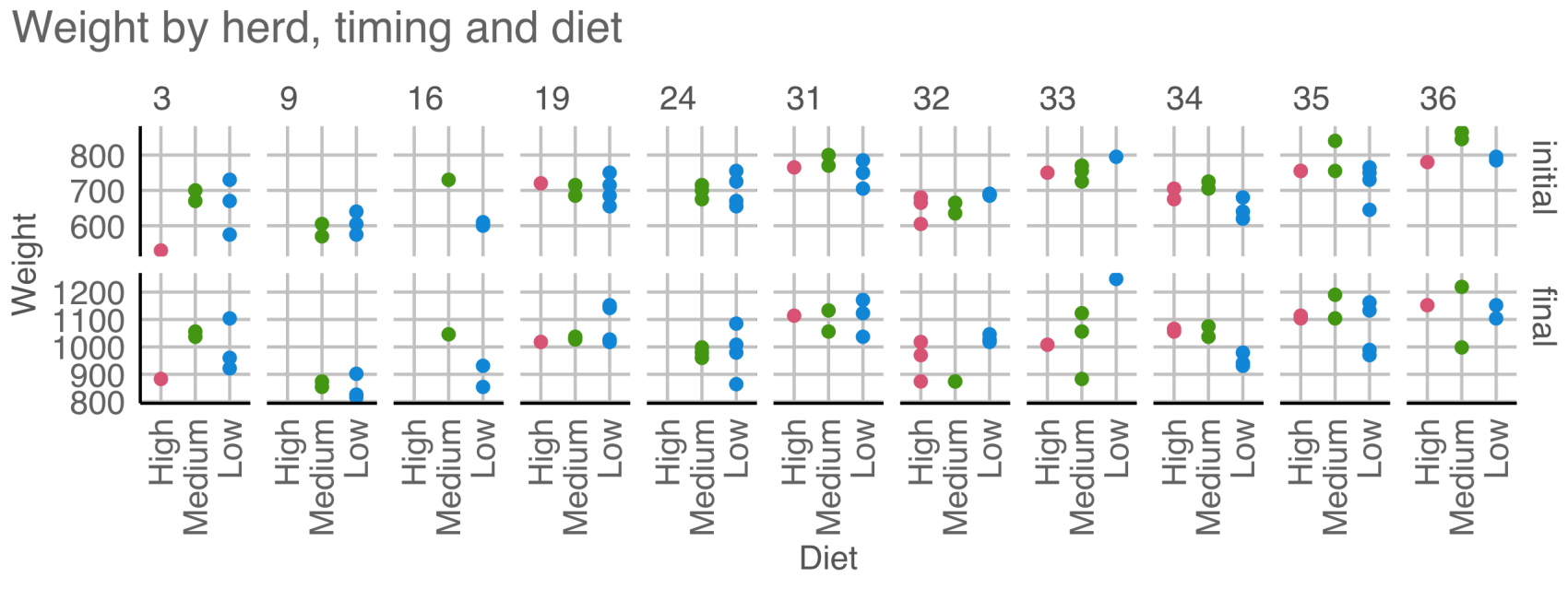
R



Case study 4 Weight of calves with different diets Part 1/2

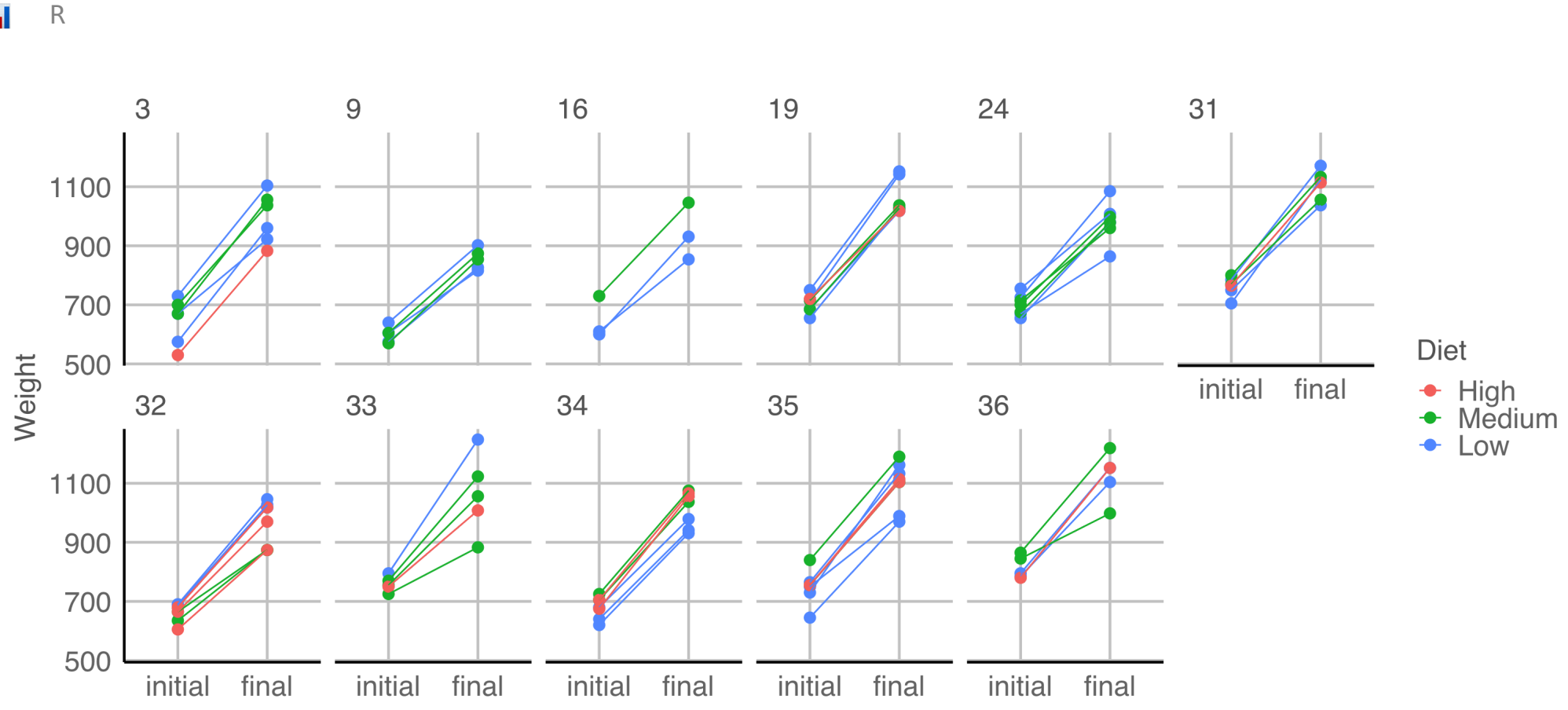
- 67 calves born in 1975 across 11 herds are fed of one of three diets with low, medium or high energy with their initial and final weights recorded.
- Different graphics and metrics will help to make comparison easier and fair.

data R



Case study 4 Weight of calves with different diets Part 2/3

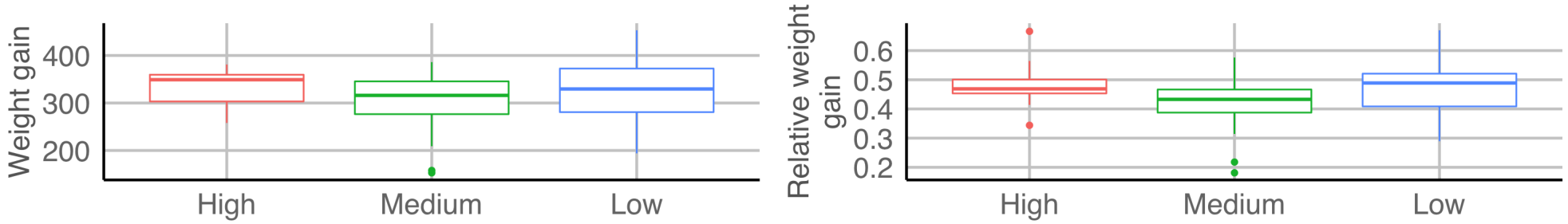
- Weight data are *paired* so comparison of initial and final weights should be matched with the animal.



Case study 4 Weight of calves with different diets Part 3/3

- People are generally better at comparing lengths on a common scale instead of angles (see Cleveland & McGill, 1985)
- We could compare the *difference in inital and final weight*.
- Weight gain doesn't take into account the initial weight though.
- We could consider computing the relative weight gain with respect to its initial weight.

R

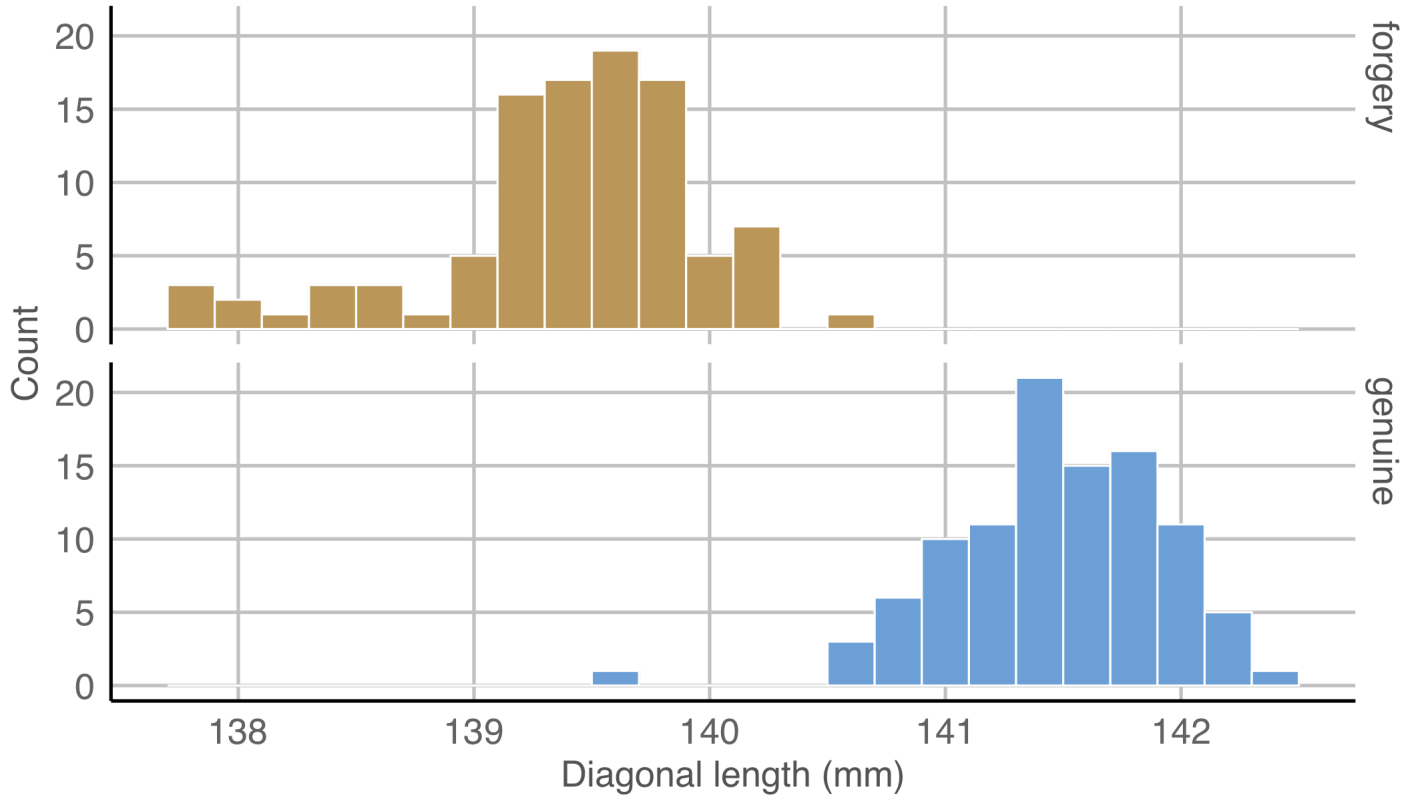


Cleveland, William S., and Robert Mc Gill. n.d. “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods.”

Case study 5 Swiss bank notes

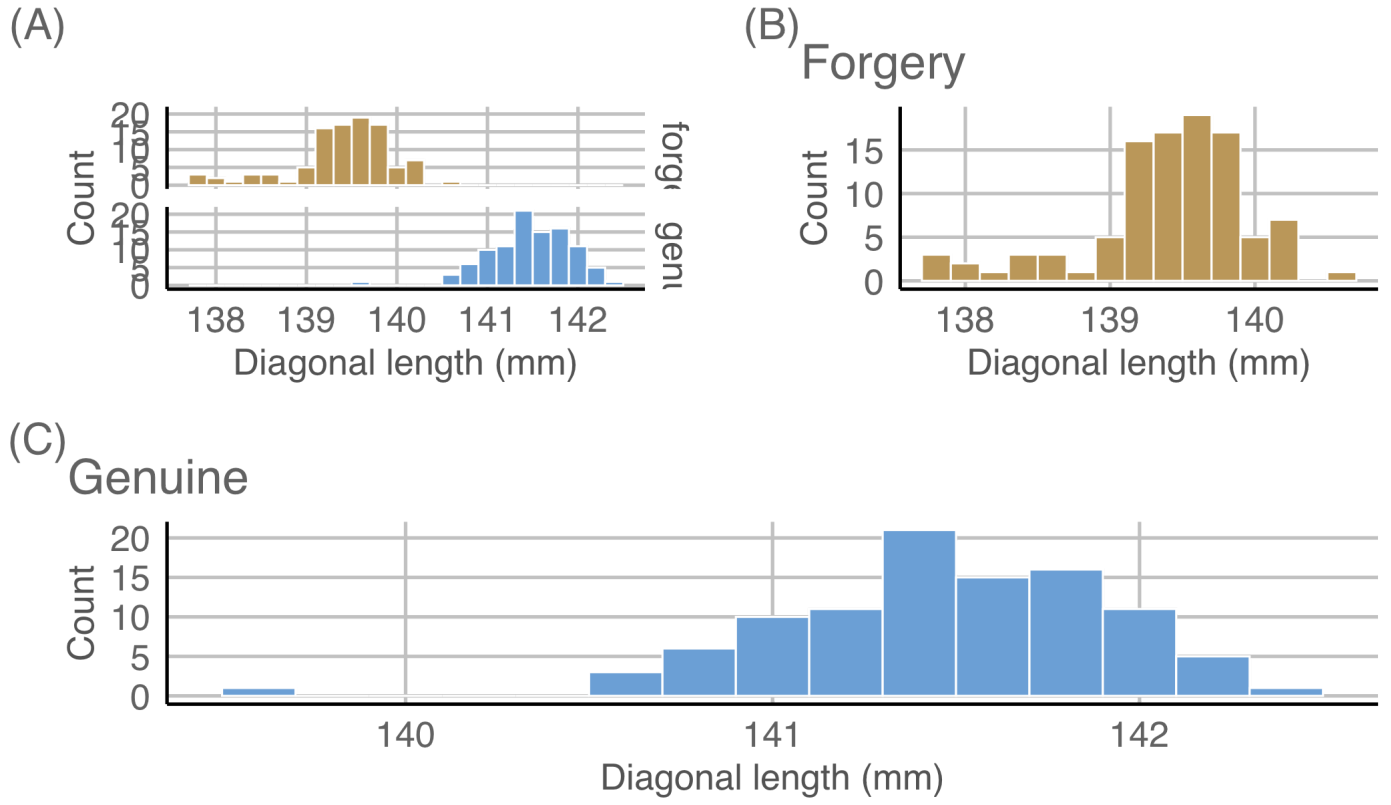
- Comparisons are not always based on point estimates.
- Below is the comparison of distribution for the diagonal length of genuine and forged Swiss bank notes.

data R



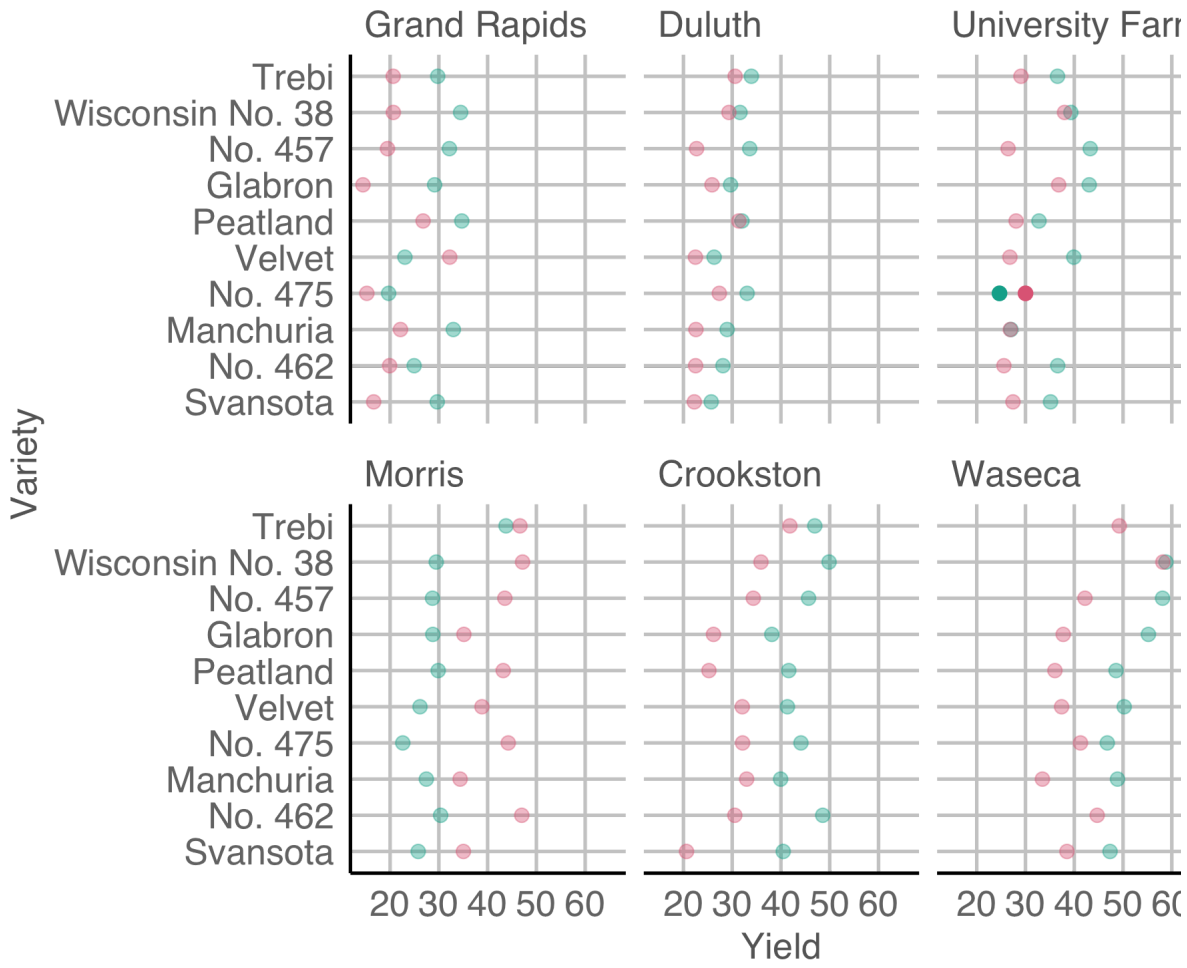
Comparing graphically Part 1

- From (A) we can see that the diagonal length distribution is quite different between forged and genuine notes.
- Comparing (B) and (C) is however difficult due to different **aspect ratio** and graphs are not in **common scale** nor **alignment**.



Case study 6 Barley Yield Part 1/2

data R

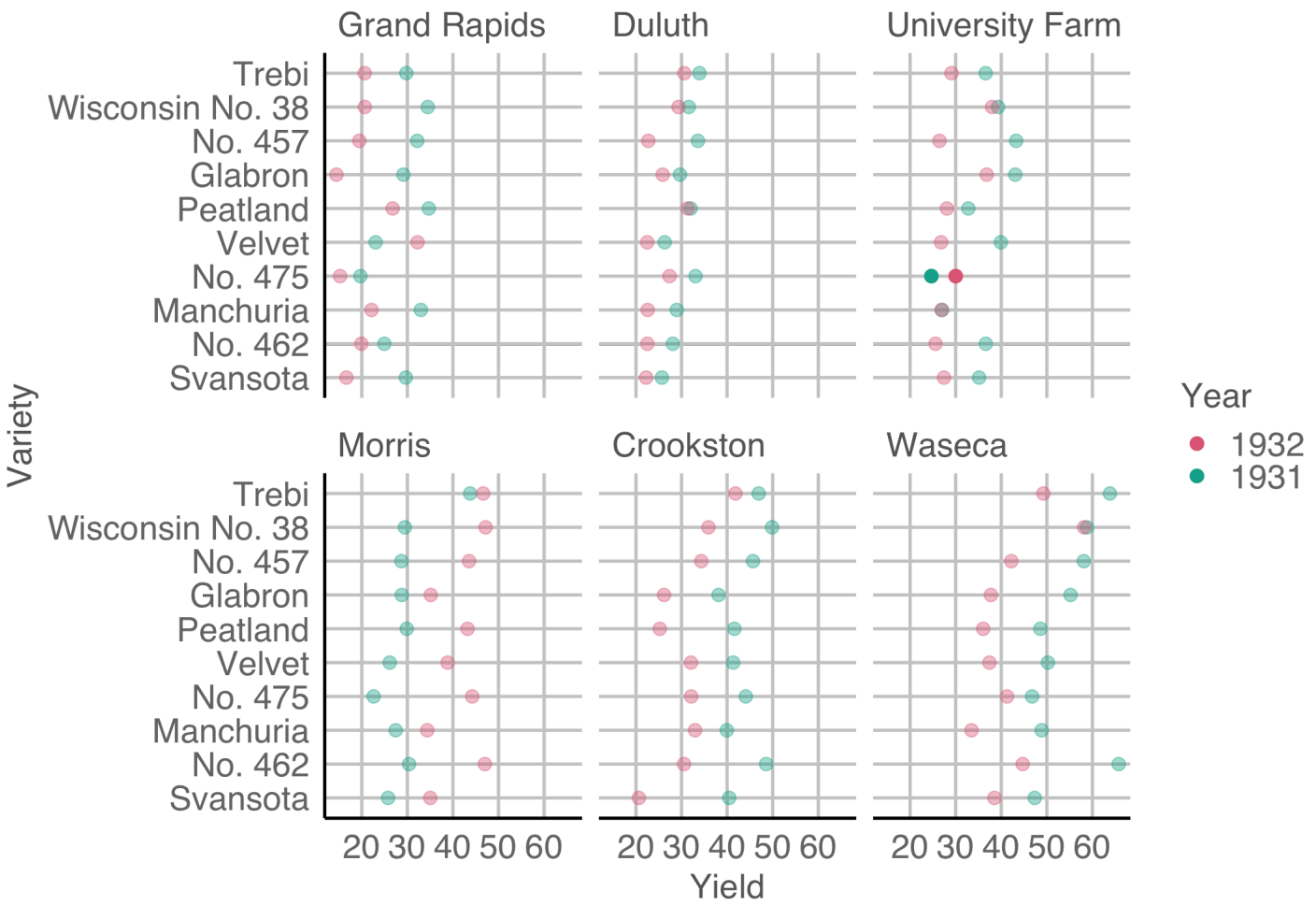


- 10 barley varieties were tested at 6 locations in 1931 and in 1932
- Do you notice anything about the yield with respect to the years?

How about now?

Immer, R. F., H. K. Hayes, and LeRoy Powers. (1934). Statistical Determination of Barley Varietal Adaptation. *Journal of the American Society of Agronomy* **26** 403–419

Case study 6 Barley Yield Part 2/2

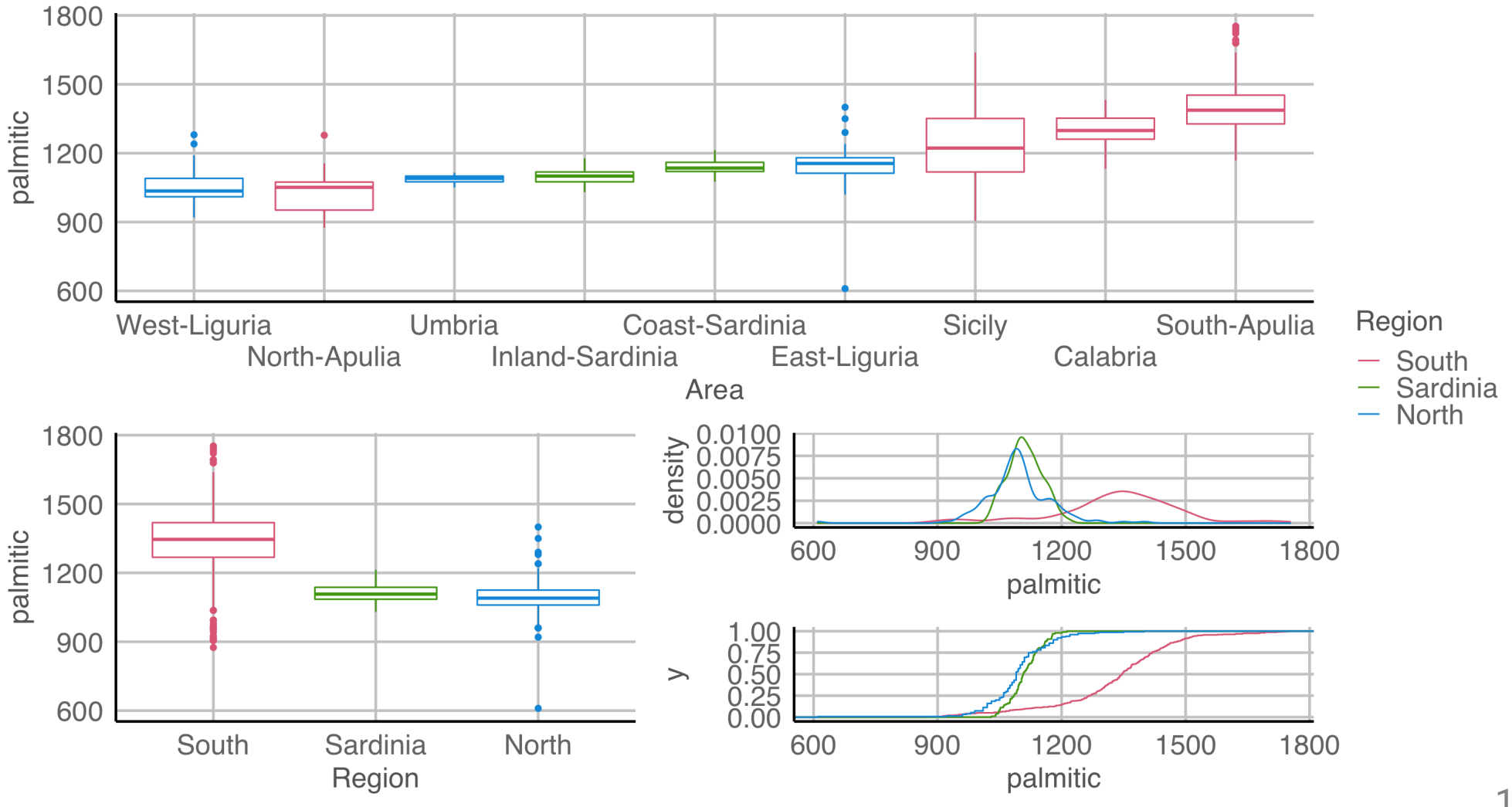


- Cleveland (1993) speculated that the year labels may have been reversed for No. 475 variety at the University Farm.
- Wright (2013) investigated this by examining extended data from 1927 to 1936, in addition to weather covariates, and found that the observations are not particularly unusual.

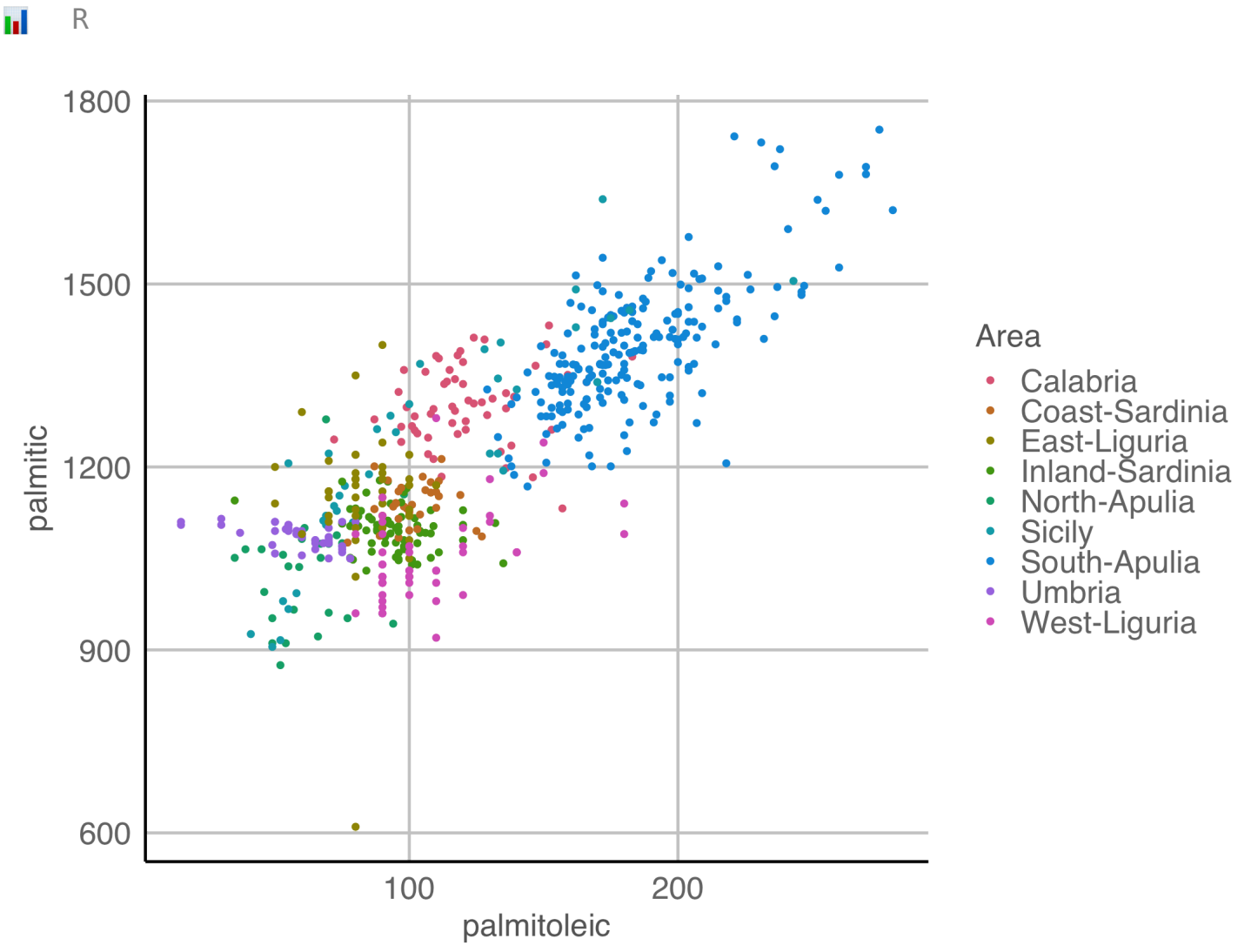
Cleveland, W. S. (1993) Visualising Data, Summit, NJ: Hobart Press.
Wright, Kevin (2013). Revisiting Immer's Barley Data. The American Statistician. **67** (3) 129–133.

Case study 7 Olive oils

data R



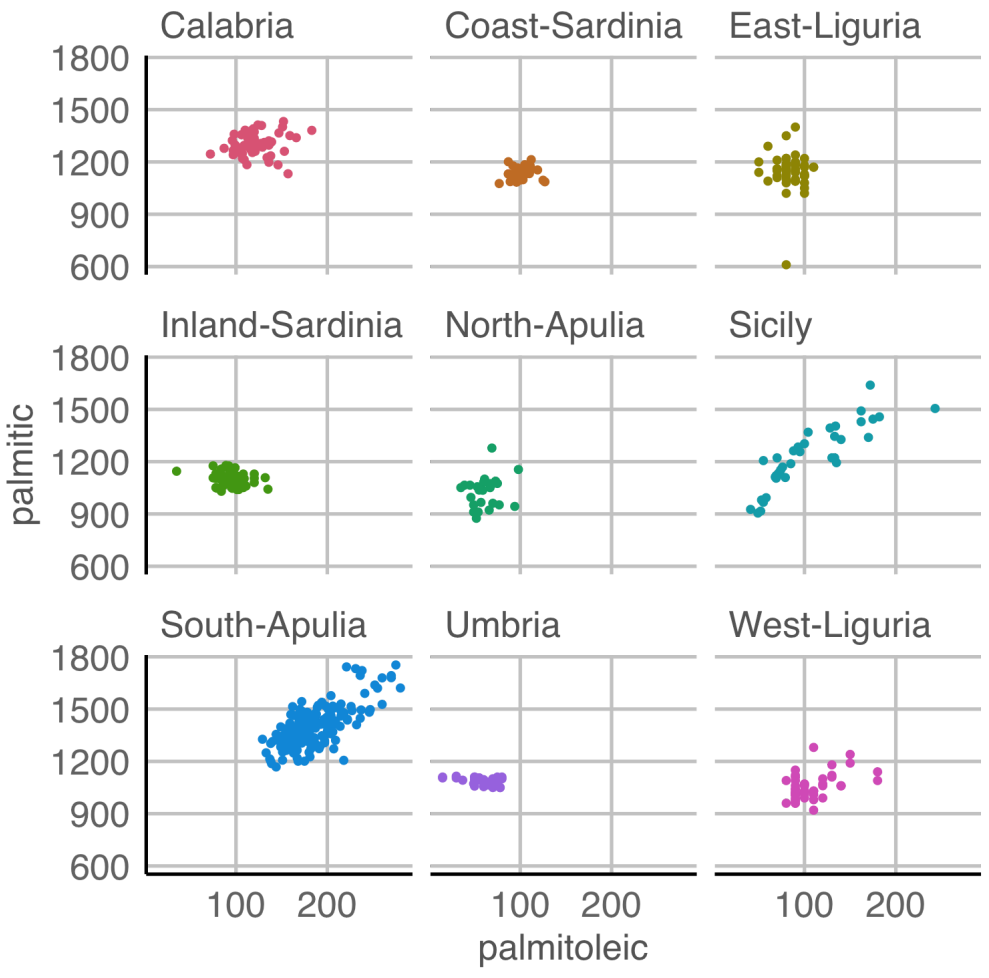
Comparing graphically Part 2



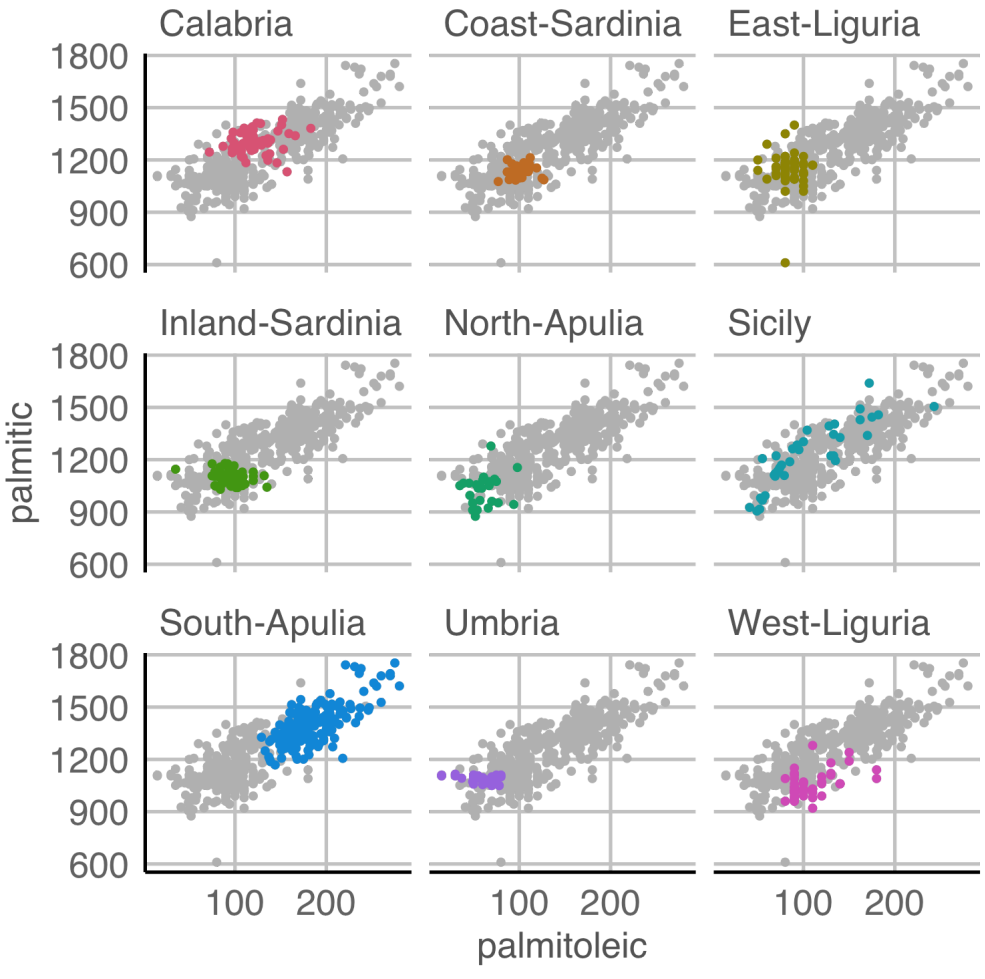
- Color is a great way to differentiate categories but if there are too many categories then it becomes hard to compare.
- In this scatter plot, there are too many overlapping points so splitting the data to **multiple windows** via facetting may make it easier to compare.

Comparing graphically Part 3

R

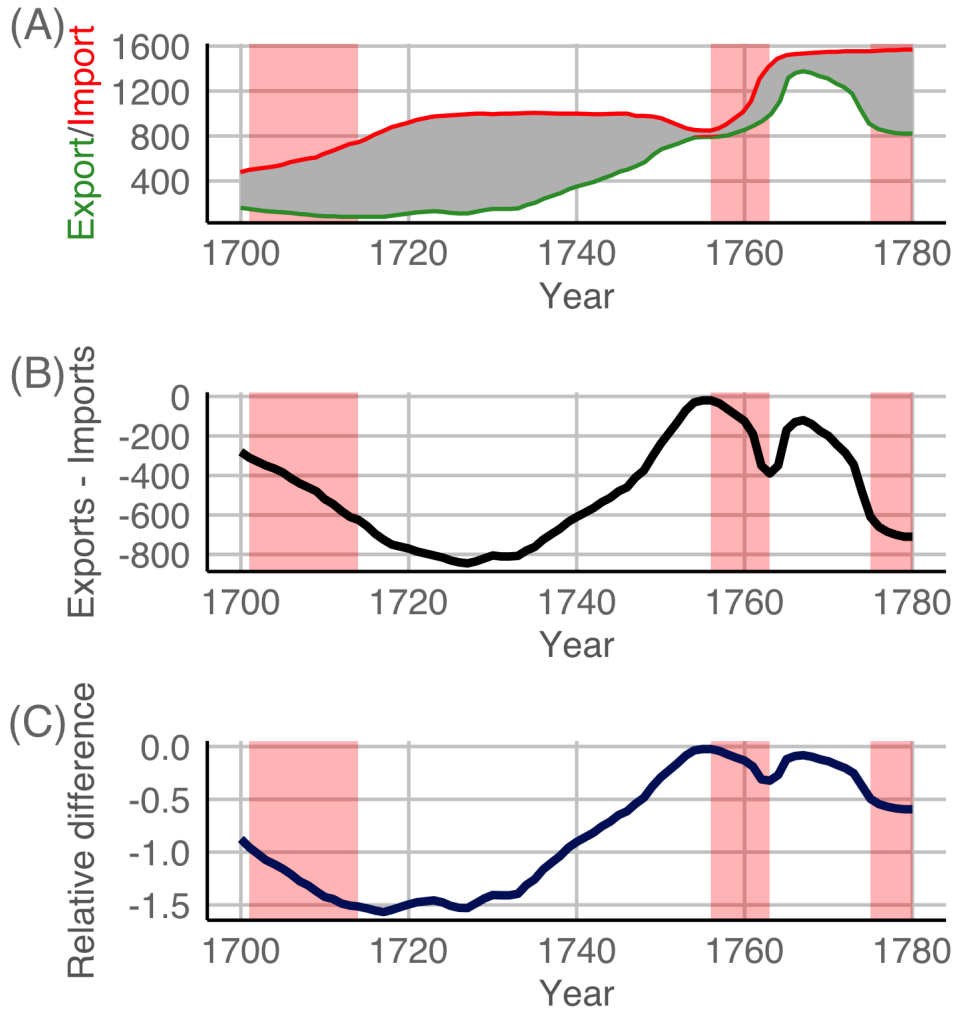


R



Case study 8 England and East Indies trade data

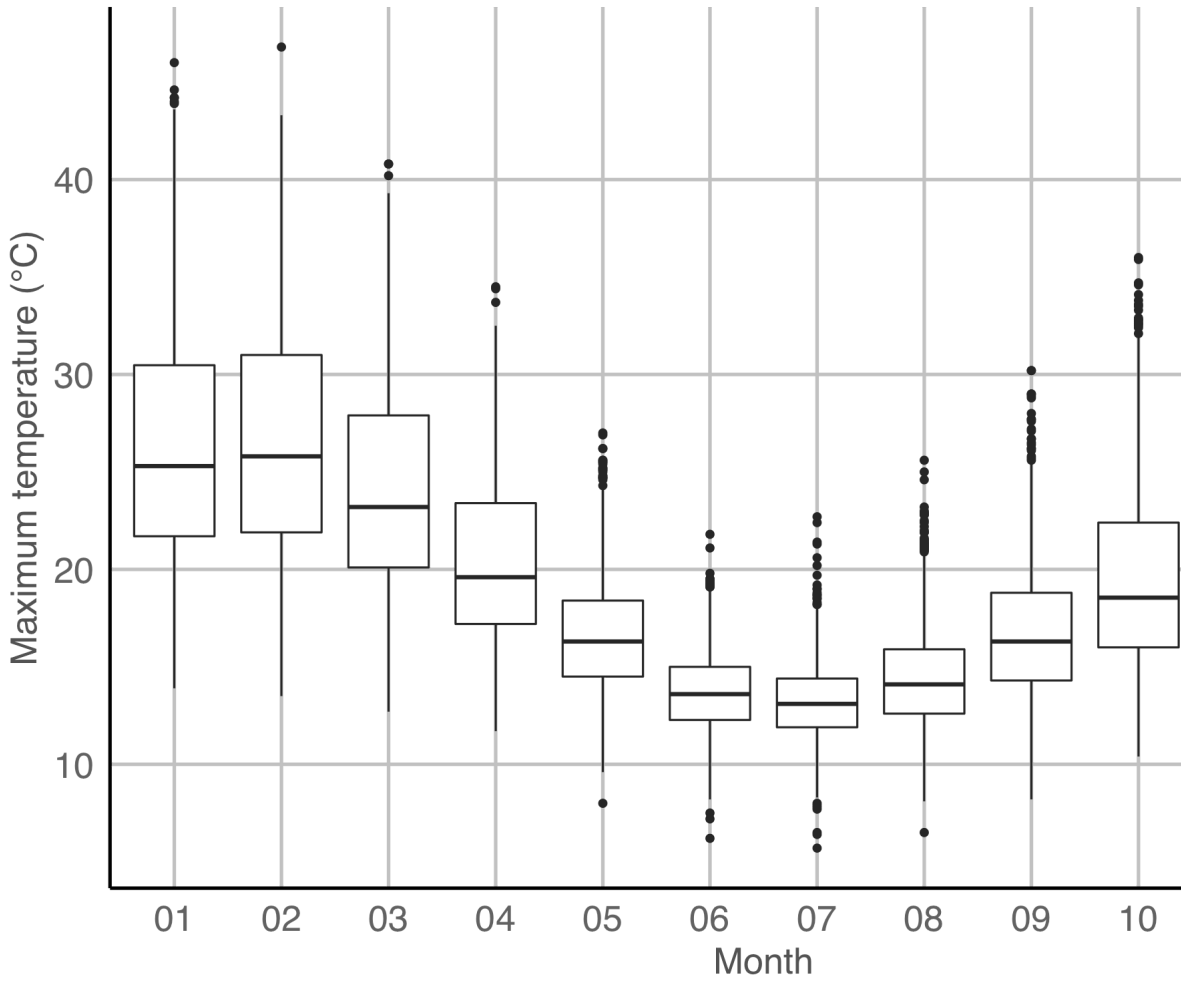
data R



- (A) shows the export from England to the East Indies and the import to England from the East Indies in millions of pounds.
- Import and export figures are easier to compare by plotting the difference like in (B).
- Relative difference may be more of an interest - (C) plots the relative difference with respect to the average of export and import values.
- The red area correspond to War of the Spanish Succession (1701-14), Seven Years' War (1756-63) and the American Revolutionary War (1775-83).

Case study 9 Melbourne's daily maximum temperature

data R



- Melbourne's daily maximum temperature from 1970 to 2020.
- How are the temperature across months?
- What about the temperature within a month?
- You'll explore this data in week 8 tutorial!

Thanks to Di for the data!

That's it, for this lecture!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu