ETC5521: Exploratory Data Analysis

Going beyond two variables, exploring high dimensions

Lecturer: *Di Cook*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 6 - Session 1

More than two variables? Use a scatterplot matrix

synonyms: splom, draughtsman plot

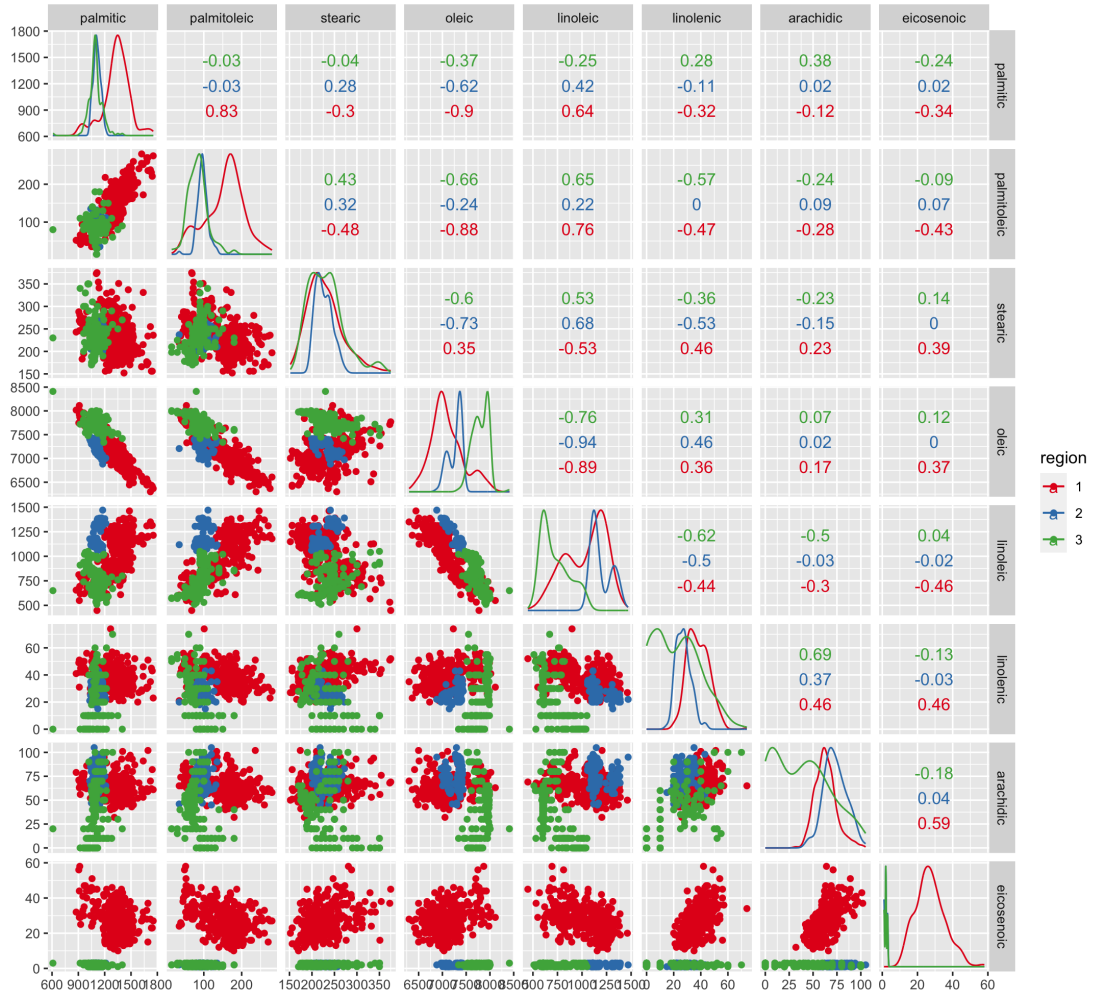
Case study 1 Olive oils

`data` `description` `R`

id	region	area	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic
1.North-Apulia	1	1	1075	75	226	7823	672	36	60
2.North-Apulia	1	1	1088	73	224	7709	781	31	61
3.North-Apulia	1	1	911	54	246	8113	549	31	63
4.North-Apulia	1	1	966	57	240	7952	619	50	78
5.North-Apulia	1	1	1051	67	259	7771	672	50	80
6.North-Apulia	1	1	911	49	268	7924	678	51	70
7.North-Apulia	1	1	922	66	264	7990	618	49	56

Case study 1 Olive oils

learn R



Read about the original book, and movie on wikipedia

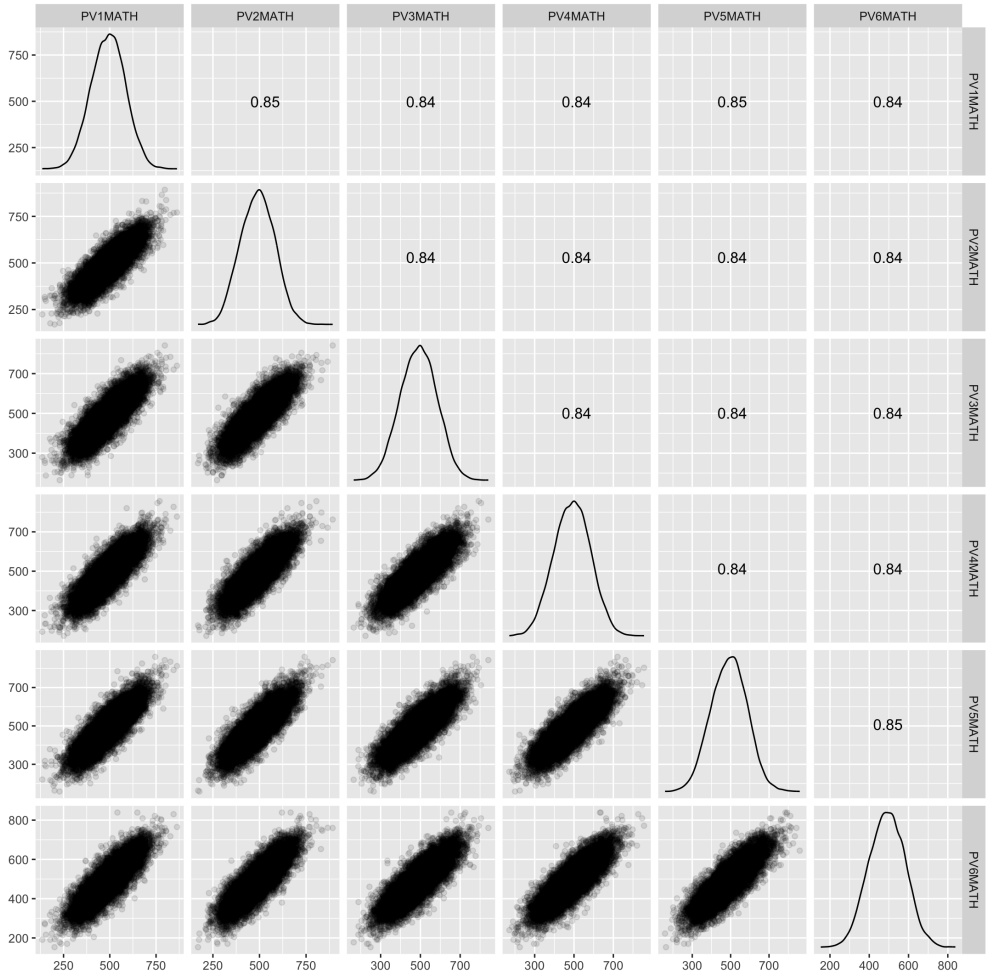
Trailer for "FLATLAND 2: SPHERELAND"

Case study 2 PISA

learn R

The Programme for International Student Assessment (PISA) is a triennial survey conducted by the Organization for Economic Cooperation and Development (OECD) on assessment measuring 15-year-old student performances in reading, mathematics and science.

Math scores for Australia for 2018. (Only 6 or the 10 shown.)



Diversion

This is an example of fraudulent synthetic data, presented in a Lancet article in May 2020 claiming hydroxychloroquinine increased risk of death.

Note: This does not mean that I support the use of HCQ.

RETRACTED: Hydroxychloroquine or chloroquine with or wit…

Download PDF [1 MB]

Fi

Background

Hydroxychloroquine or chloroquine, often in combination with a second-generation macrolide, are being widely used for treatment of COVID-19, despite no conclusive evidence of their benefit. Although generally safe when used for approved indications such as autoimmune disease or malaria, the safety and benefit of these treatment regimens are poorly evaluated in COVID-19.

Methods

We did a multinational registry analysis of the use of hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19. The registry comprised data from 671 hospitals in six continents. We included patients hospitalised between Dec 20, 2019, and April 14, 2020, with a positive laboratory finding for SARS-CoV-2. Patients who received one of the treatments of interest within 48 h of diagnosis were included in one of four treatment groups (chloroquine alone, chloroquine with a macrolide, hydroxychloroquine alone, or hydroxychloroquine with a macrolide), and patients who received none of these treatments formed the control group. Patients for whom one of the treatments of interest was initiated more than 48 h after diagnosis or while they were on mechanical ventilation, as well as patients who received remdesivir, were excluded. The main outcomes of interest were in-hospital mortality and the occurrence of de-novo ventricular arrhythmias (non-sustained or sustained ventricular tachycardia or ventricular fibrillation).

Table S3. Summary Data by Continent

Variable	North America	South America	Europe	Africa	Asia	Australia
N	63,315	3,577	16,574	4,402	7,555	609
Age (years)	54.4 +/- 17.8	53.6 +/- 17.1	52.7 +/- 17.0	53.9 +/- 16.9	51.9 +/- 17.2	55.8 +/- 17.7
BMI (Kg/m^2)	28.1 +/- 5.3	26.4 +/- 5.4	28.1 +/- 5.3	23.8 +/- 5.4	24.8 +/- 5.3	28.1 +/- 5.4
Female sex	29,288 (46.3)	1,678 (46.9)	7,730 (46.6)	1,981 (45.0)	3,486 (46.1)	263 (43.2)
Coronary artery disease	7,850 (12.4)	485 (13.6)	2,169 (13.1)	614 (13.9)	980 (13.0)	39 (6.4)
Congestive heart failure	1,639 (2.6)	73 (2.0)	366 (2.2)	105 (2.4)	179 (2.4)	6 (1.0)
History of arrhythmia	2,293 (3.6)	118 (3.3)	543 (3.3)	146 (3.3)	256 (3.4)	25 (4.1)
Diabetes mellitus	8,654 (13.7)	521 (14.6)	2,360 (14.2)	570 (12.9)	1,069 (14.1)	86 (14.1)
Hypertension	17,159 (27.1)	954 (26.7)	4,368 (26.4)	1,140 (25.9)	2,010 (26.6)	179 (29.4)
Hyperlipidemia	20,032 (31.6)	1,088 (30.4)	5,131 (31.0)	1,380 (31.3)	2,374 (31.4)	193 (31.7)
COPD	2,069 (3.3)	97 (2.7)	590 (3.6)	132 (3.0)	254 (3.4)	35 (5.7)
Current smoker	6,316 (10.0)	347 (9.7)	1,604 (9.7)	453 (10.3)	707 (9.4)	61 (10.0)
Former smoker	10,707 (16.9)	670 (18.7)	2,936 (17.7)	830 (18.9)	1,301 (17.2)	109 (17.9)
Immunocompromised	1,997 (3.2)	52 (1.5)	463 (2.8)	127 (2.9)	208 (2.8)	21 (3.4)
ACE inhibitor	5,327 (8.4)	285 (8.0)	1,341 (8.1)	325 (7.4)	605 (8.0)	66 (10.8)
Statin	6,188 (9.8)	306 (8.6)	1,552 (9.4)	436 (9.9)	674 (8.9)	89 (14.6)
ARB	3,913 (6.2)	220 (6.2)	963 (5.8)	259 (5.9)	454 (6.0)	40 (6.6)
Antiviral Therapy use	25,646 (40.5)	1,444 (40.4)	6,747 (40.7)	1,771 (40.2)	3,085 (40.8)	234 (38.4)
Chloroquine alone	1,091 (1.7)	114 (3.2)	295 (1.8)	153 (3.5)	199 (2.6)	16 (2.6)
Hydroxychloroquine alone	2,127 (3.4)	72 (2.0)	540 (3.3)	83 (1.9)	184 (2.4)	10 (1.6)
CQ + macrolide	2,324 (3.7)	217 (6.1)	562 (3.4)	256 (5.8)	391 (5.2)	33 (5.4)
HCQ + macrolide						

Table S3. Summary Data by Continent

Variable	North America	South America	Europe	Africa	Asia	Australia
N						
Coronary artery disease	7,850 (12.4)	485 (13.6)	2,169 (13.1)	614 (13.9)	980 (13.0)	39 (6.4)
Congestive heart failure	1,639 (2.6)	73 (2.0)	366 (2.2)	105 (2.4)	179 (2.4)	6 (1.0)
History of arrhythmia	2,293 (3.6)	118 (3.3)	543 (3.3)	146 (3.3)	256 (3.4)	25 (4.1)
	25,646 (40.5)	1,444 (40.4)	6,747 (40.7)	1,771 (40.2)	3,085 (40.8)	234 (38.4)
Chloroquine alone	1,091 (1.7)	114 (3.2)	295 (1.8)	153 (3.5)	199 (2.6)	16 (2.6)
Hydroxychloroquine alone	2,127 (3.4)	72 (2.0)	540 (3.3)	83 (1.9)	184 (2.4)	10 (1.6)

Another rather remarkable aspect is how beautifully uniform the aggregated data are across continents

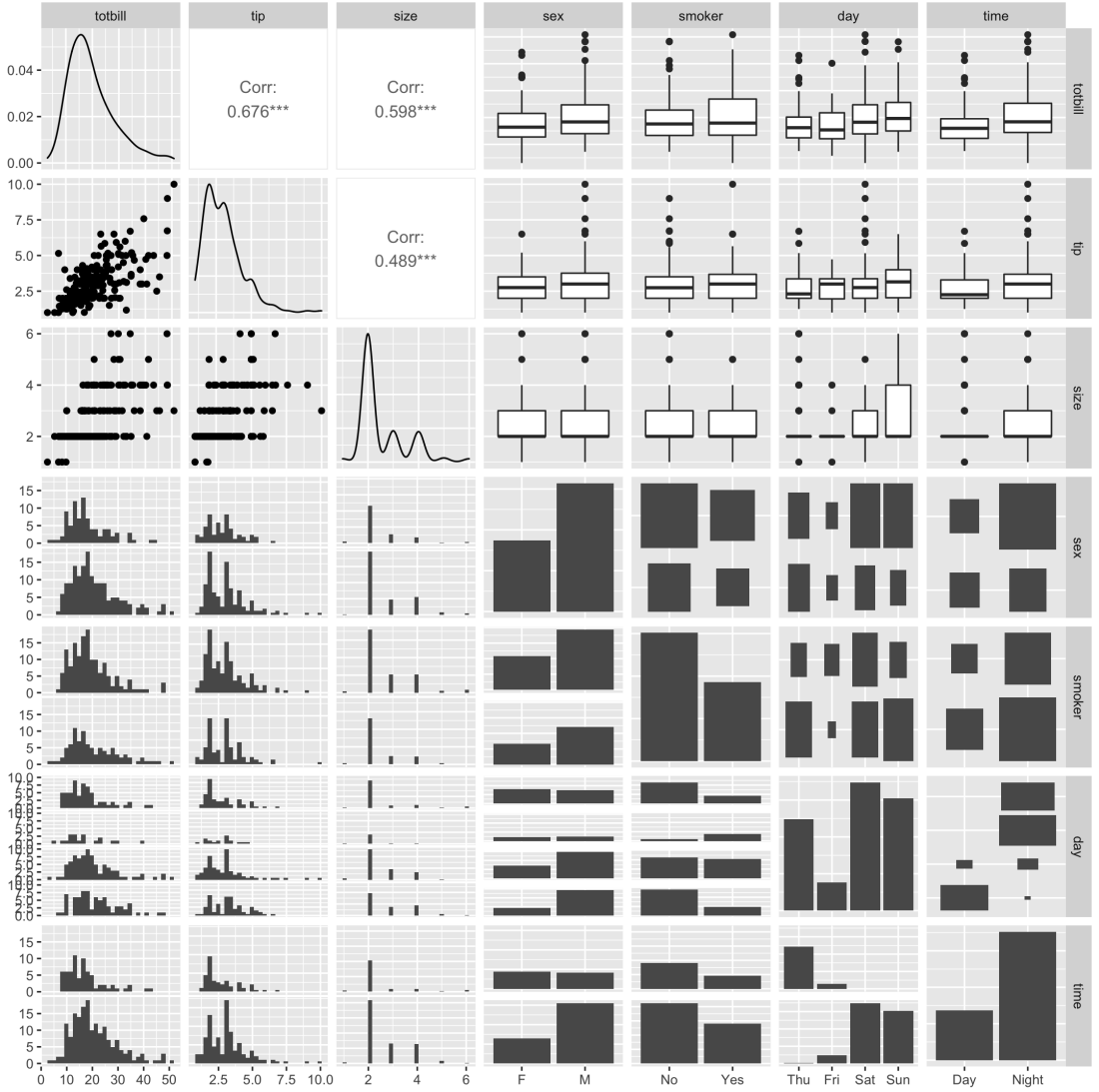
For example, smoking is almost between 9.4-10% in 6 continents. As they don't tell us which countries are involved, hard to see how this matches known smoking prevalences. Antiviral use is 40.5, 40.4, 40.7, 40.2, 40.8, 38.4%. Remarkable! I didn't realise that treatment was so well coordinated across the world. Diabetes and other co-morbidities don't vary much either.

Generalised pairs plot

If the types of variables are not both quantitative, there are some other choices of mapping

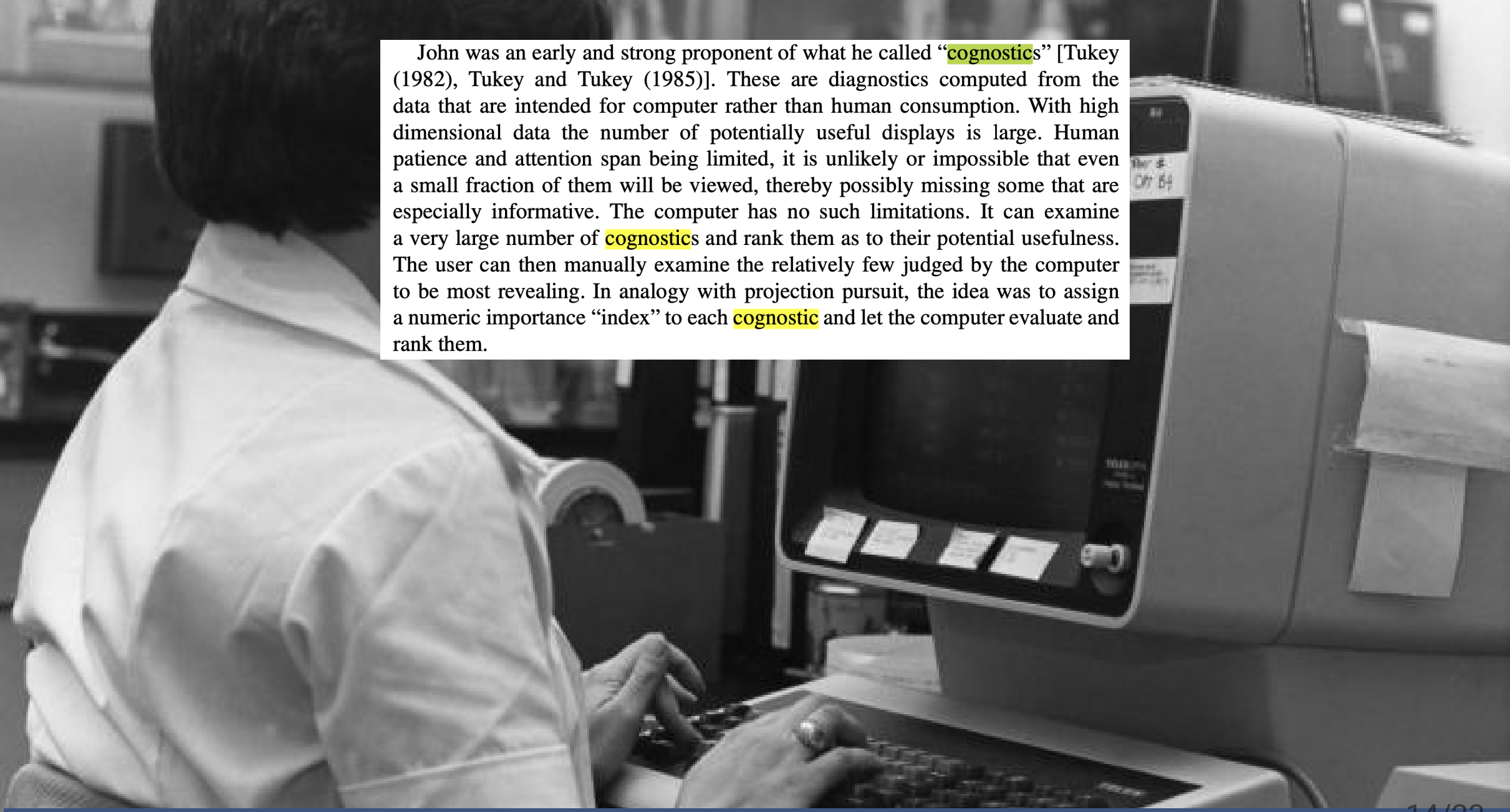
Case study 3 Tips

learn R

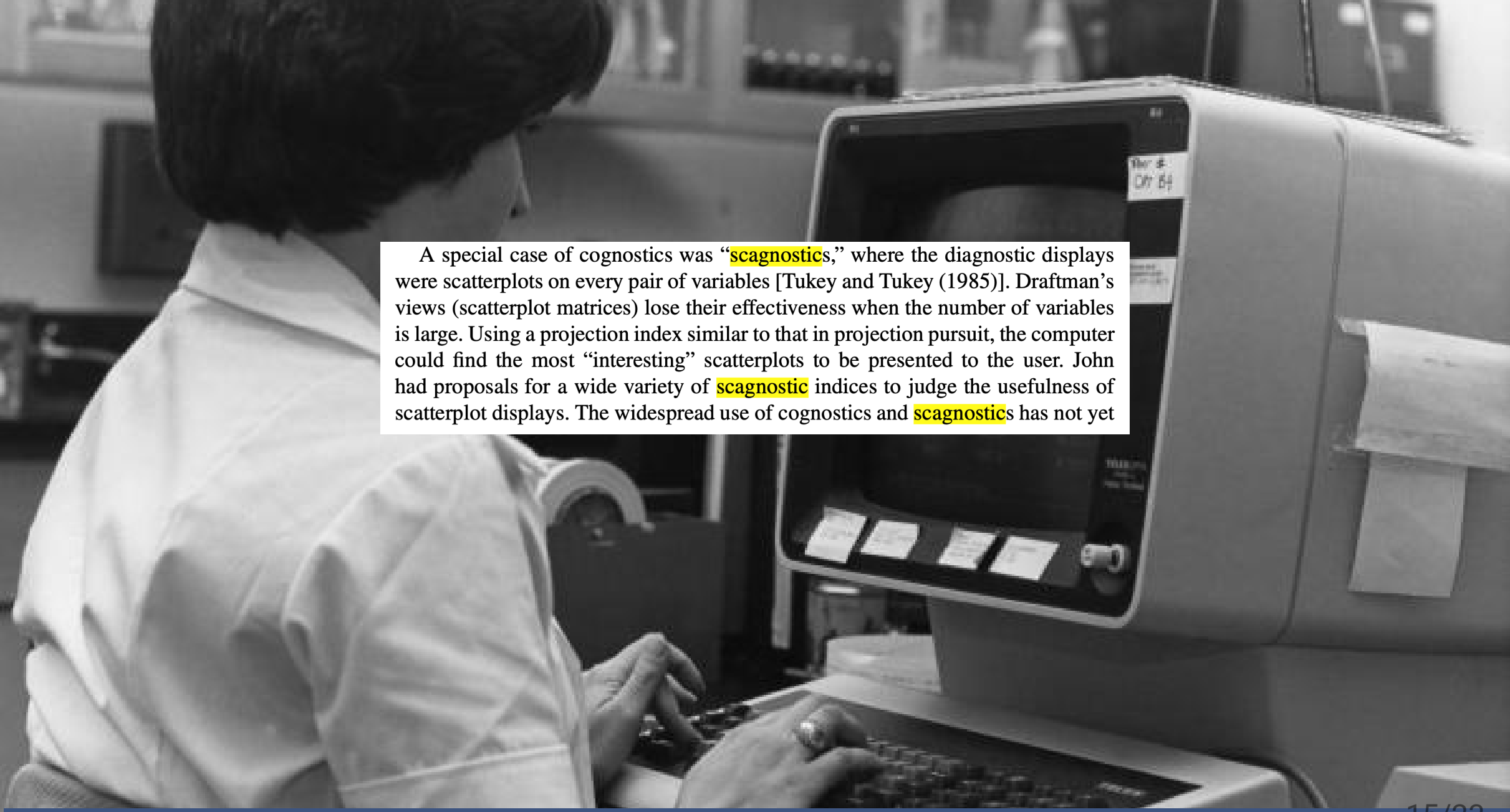


Scagnostics

Has your data got too many pairs of variables to scan easily?



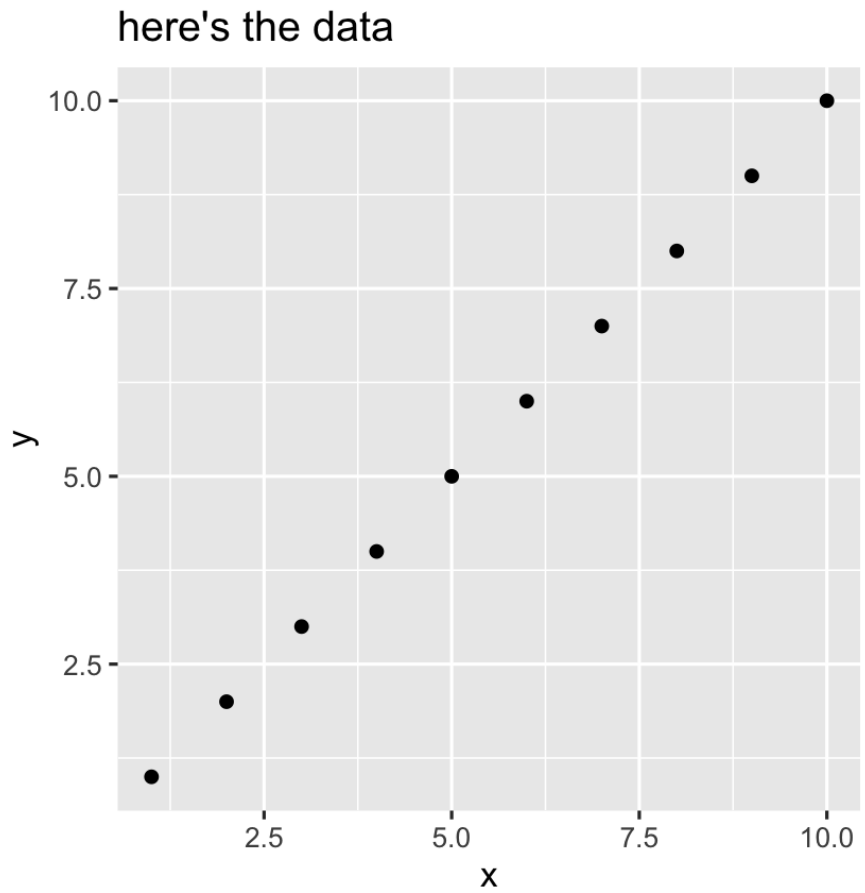
John was an early and strong proponent of what he called "cognostics" [Tukey (1982), Tukey and Tukey (1985)]. These are diagnostics computed from the data that are intended for computer rather than human consumption. With high dimensional data the number of potentially useful displays is large. Human patience and attention span being limited, it is unlikely or impossible that even a small fraction of them will be viewed, thereby possibly missing some that are especially informative. The computer has no such limitations. It can examine a very large number of cognostics and rank them as to their potential usefulness. The user can then manually examine the relatively few judged by the computer to be most revealing. In analogy with projection pursuit, the idea was to assign a numeric importance "index" to each cognostic and let the computer evaluate and rank them.

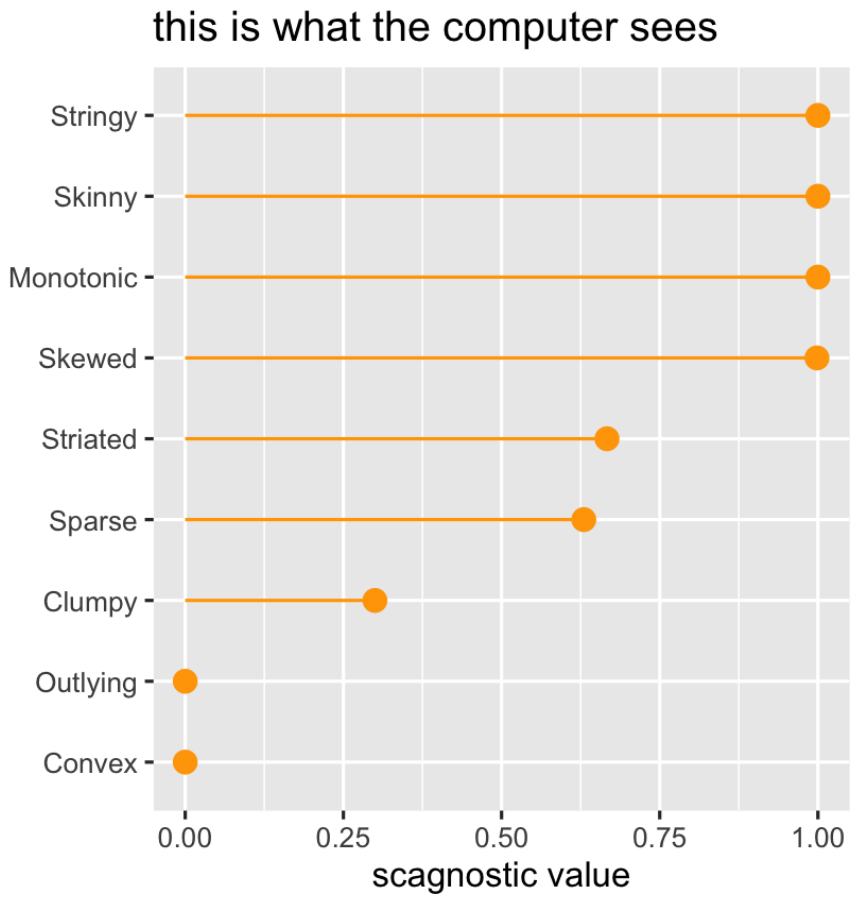


A special case of cognostics was "scagnostics," where the diagnostic displays were scatterplots on every pair of variables [Tukey and Tukey (1985)]. Draftman's views (scatterplot matrices) lose their effectiveness when the number of variables is large. Using a projection index similar to that in projection pursuit, the computer could find the most "interesting" scatterplots to be presented to the user. John had proposals for a wide variety of scagnostic indices to judge the usefulness of scatterplot displays. The widespread use of cognostics and scagnostics has not yet

Scagnostics

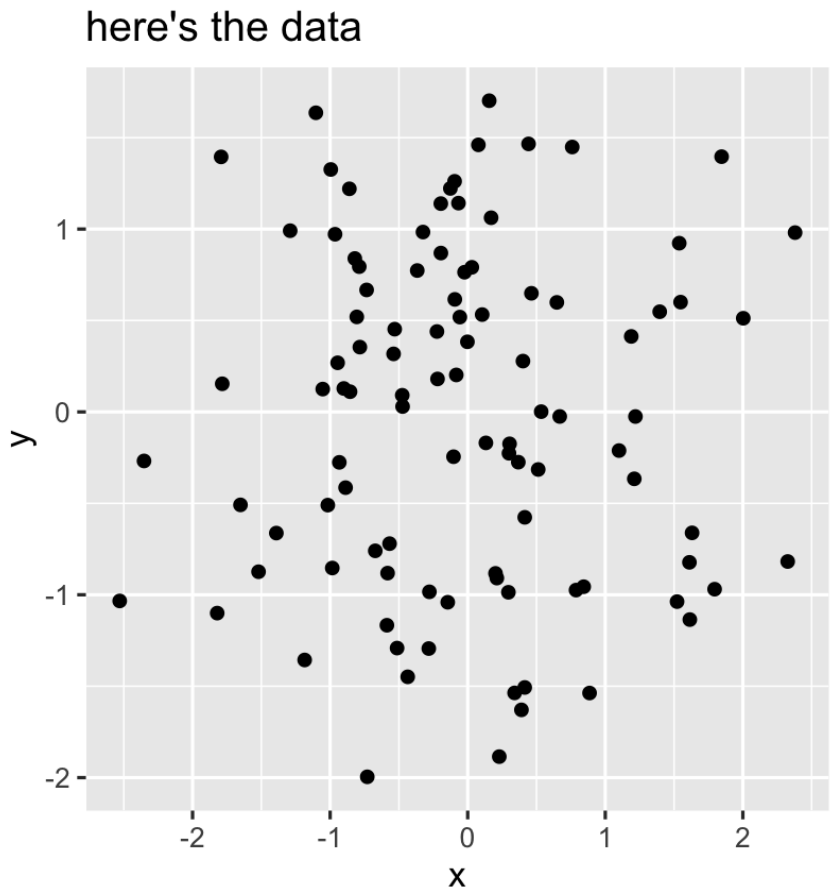
R

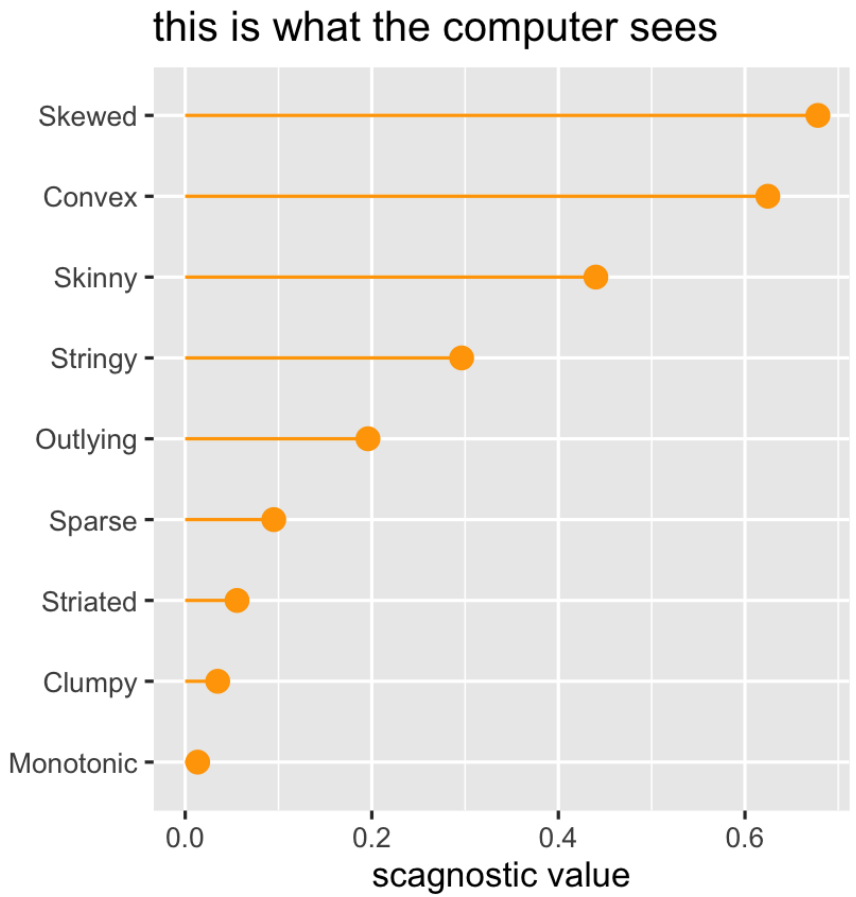




Scagnostics

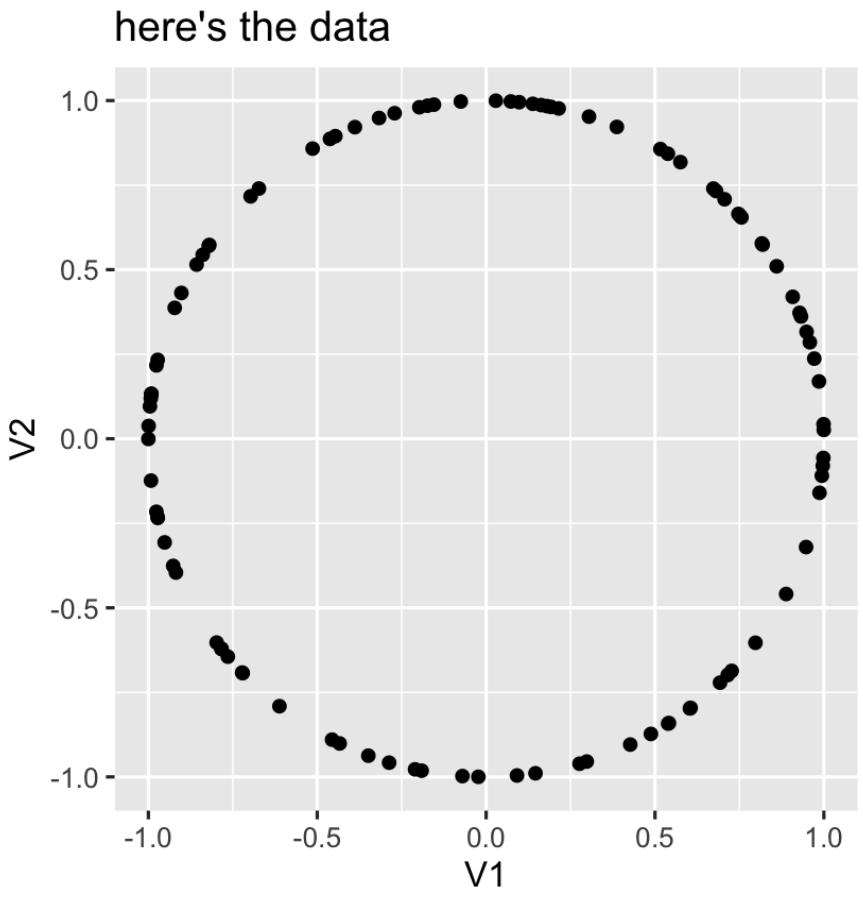
R

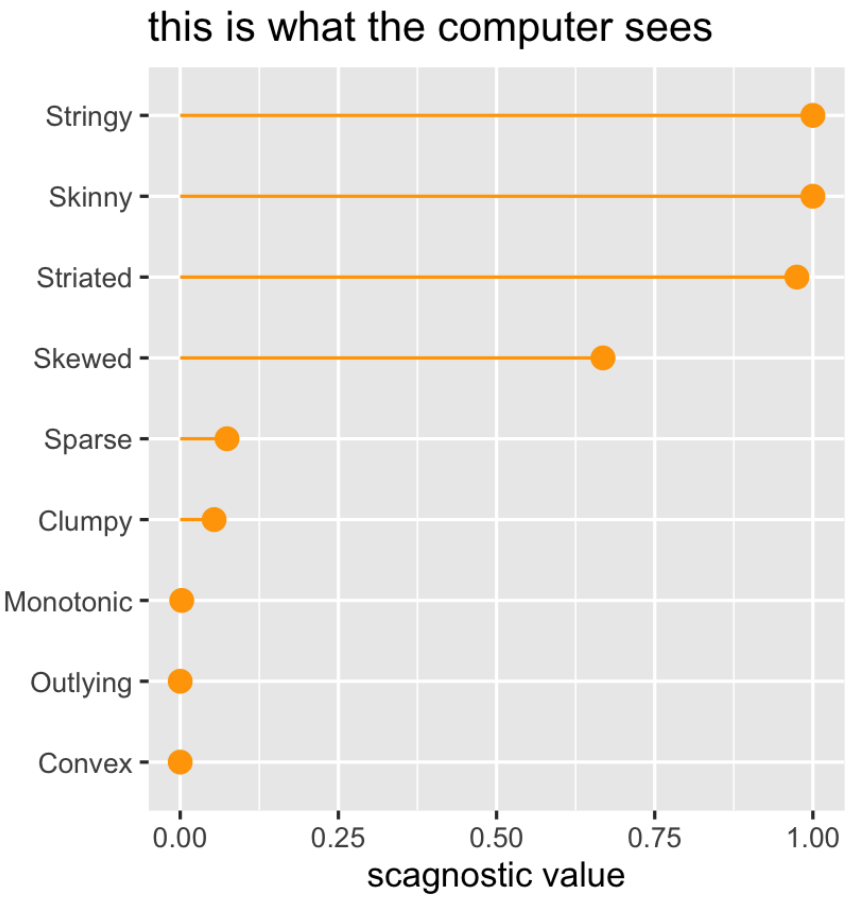




Scagnostics

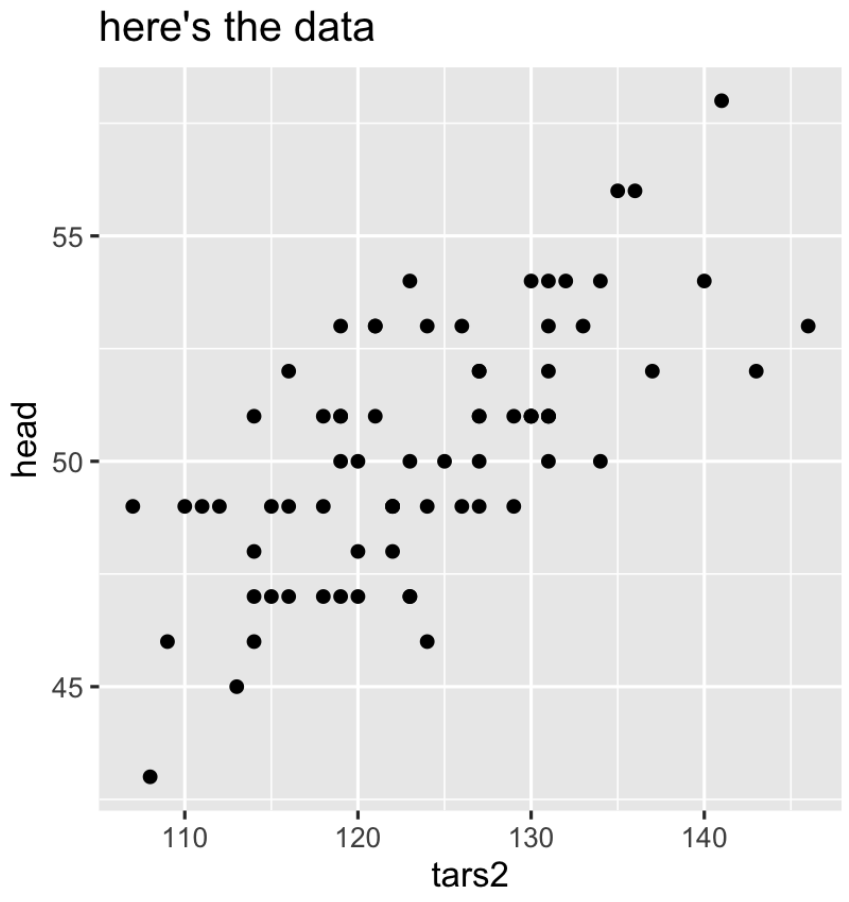
R

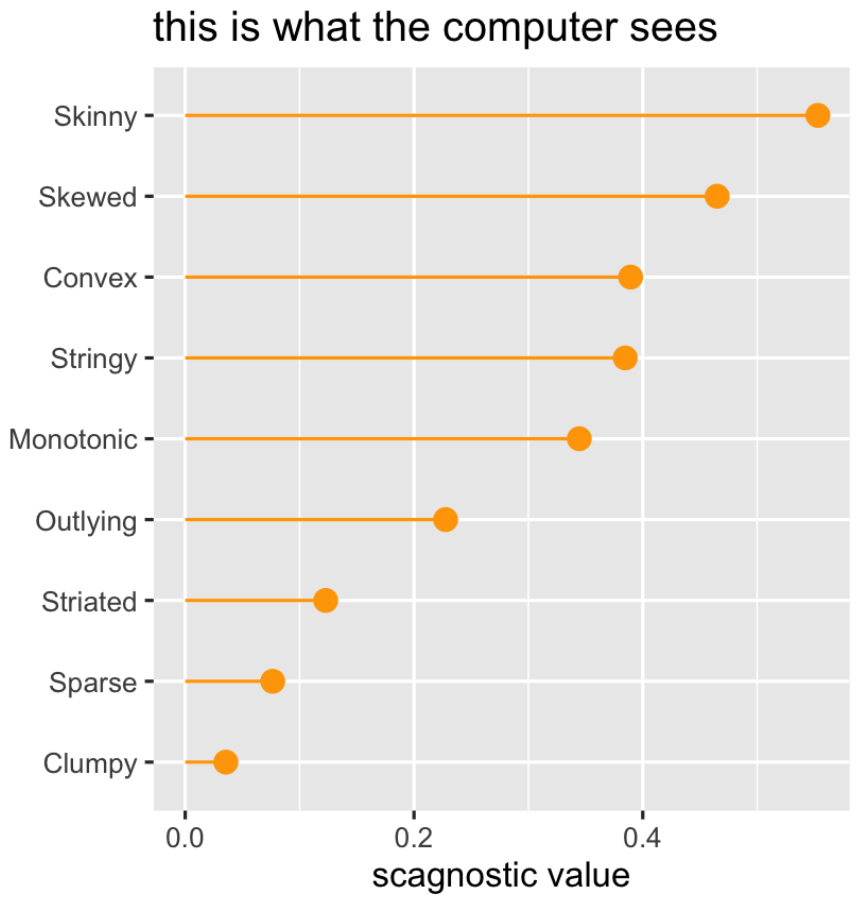




Scagnostics

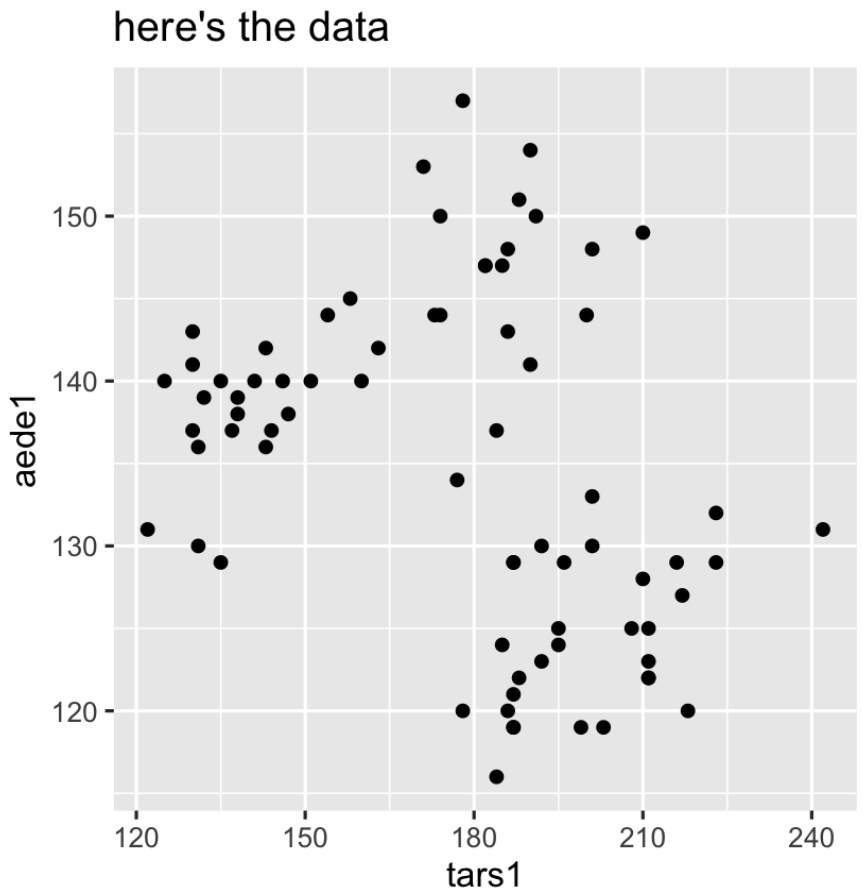
R

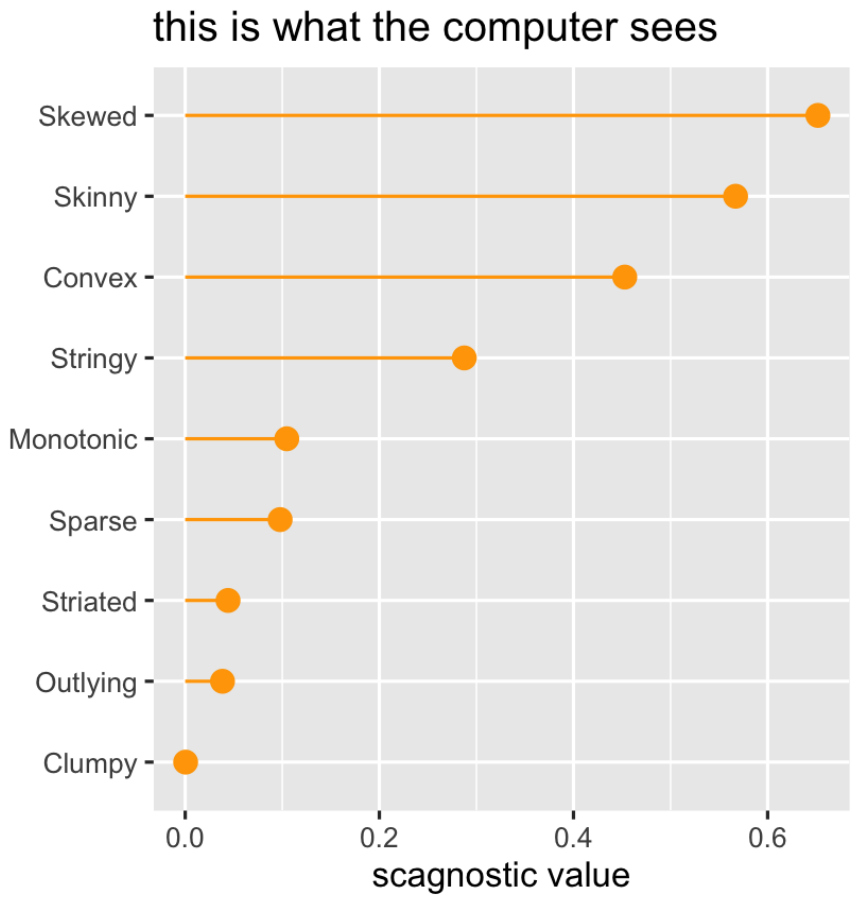




Scagnostics

R





How are scagnostics calculated?

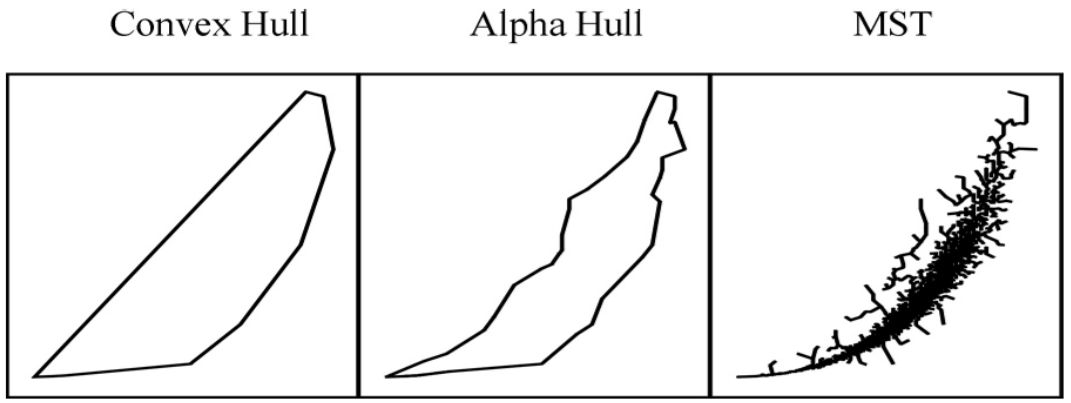


Figure 2: Graphs used as bases for computing scagnostics measures

edge lengths, e, T total number of edges

A alpha hull, H convex hull

outlying `((#e>Q_3+1.5(Q_3-Q_1))/T)`
convex `(area(A)/area(H))` (shape)
skinny `(1-\sqrt{4\pi area(A)}/perimeter(A))` (shape)
stringy `(diameter(T)/length(T))` (shape)
monotonic is `(r^2)` (trend)
skewed clumpy (density)
striated (coherence)

Wilkinson, Anand, Grossmann (1994)

Scagnostics from familiar measures

There are many more ways to numerically characterise association that can be used as scagnostics too

- Slope, intercept, and error estimate from a simple linear model
- Correlation
- Principal component analysis: first eigenvalue
- Linear discriminant analysis: Between group SS to within group SS
- Cluster metrics
- Also see
  - tignostics for time series
  - longnostics for longitudinal data

That's it, for this lecture!

Lecturer: Di Cook

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu