ETC5521: Exploratory Data Analysis

Initial data analysis: Model dependent exploration and how it differs from EDA

Lecturer: *Emi Tanaka*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 3 - Session 2

Initial Data Analysis

and

Exploratory Data Analysis

How are they different?

How are they alike?

- The statistical techniques applied may be the same or similar.
- The end result maybe similar (e.g. graphically showing characteristics of the data).
- Some definitions of EDA encompasses IDA.
- So what's the difference?
- Well let's begin first by talking about models.

Linear models in R REVIEW

```
library(tidyverse)
glimpse(cars)
```

```
## Rows: 50
## Columns: 2
## $ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10,
## $ dist  <dbl> 2, 10, 4, 22, 16, 10, 18,
```

```
ggplot(cars, aes(dist, speed)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
lm(speed ~ dist, data = cars)
```

is the same as

```
lm(speed ~ 1 + dist, data = cars)
```

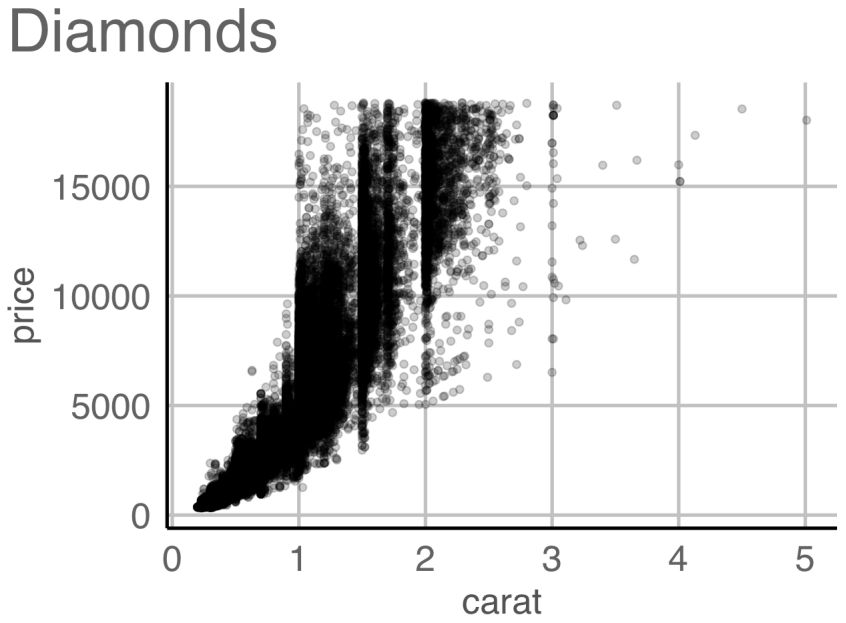
and mathematically written as

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- y_i and x_i are the speed (in mph) and stopping distance (in ft), respectively, of the i-th car;
- β_0 and β_1 are intercept and slope, respectively; and
- e_i is the random error; usually assuming $e_i \sim NID(0, \sigma^2)$.

2 Model formulation Part 1/2

- Say, we are interested in characterising the price of the diamond in terms of its carat.



- Looking at this plot, would you fit a linear model with formula

```
price ~ 1 + carat?
```

- What about

```
price ~ poly(carat, 2)?
```

which is the same as fitting:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i.$$

- The assumption for error distribution will need to be modified if so.
- Should we make some transformation before modelling?
- Are there other candidate models?

❷ Model formulation Part 2/2

- Notice that we did ***no formal statistical inference*** as we initially try to formulate the model.
- The goal of the main analysis is to characterise the price of a diamond by its carat. This may involve:
  - formal inference for model selection;
  - justification of the selected "final" model; and
  - fitting the final model.
- There may be in fact many, many models considered but discarded at the IDA stage.
- These discarded models are hardly ever reported. Consequently, majority of reported statistics give a distorted view and it's important to remind yourself what might ***not*** be reported.

Model selection

"All models are *approximate* and *tentative*; approximate in the sense that no model is exactly true and tentative in that they may be modified in the light of further data"

Chatfield (1985)

"All models are wrong but some are useful"

George Box

Case study 3 Wheat yield in South Australia Part 1/9

A wheat breeding trial to test 107 varieties (also called genotype) is conducted in a field experiment laid out in a rectangular array with 22 rows and 15 columns.

```
data("gilmour.serpentine", package = "agridat")
skimr::skim(gilmour.serpentine)
```

```
## ── Data Summary ────────────────────────────────────────────
##                            Values
## Name                       gilmour.serpentine
## Number of rows             330
## Number of columns          5
## ________________________
## Column type frequency:
##   factor                   2
##   numeric                  3
## ________________________
```

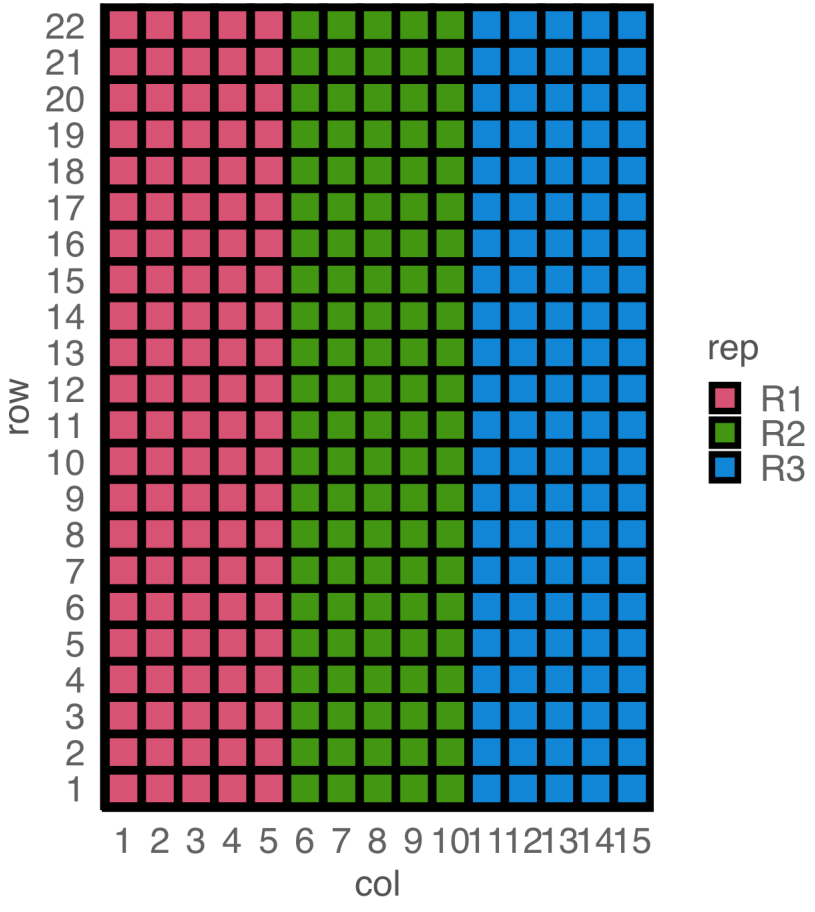
Arthur R Gilmour and Brian R Cullis and Arunas P Verbyla (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agric Biol Env Statistics* **2** 269-293

Experimental Design

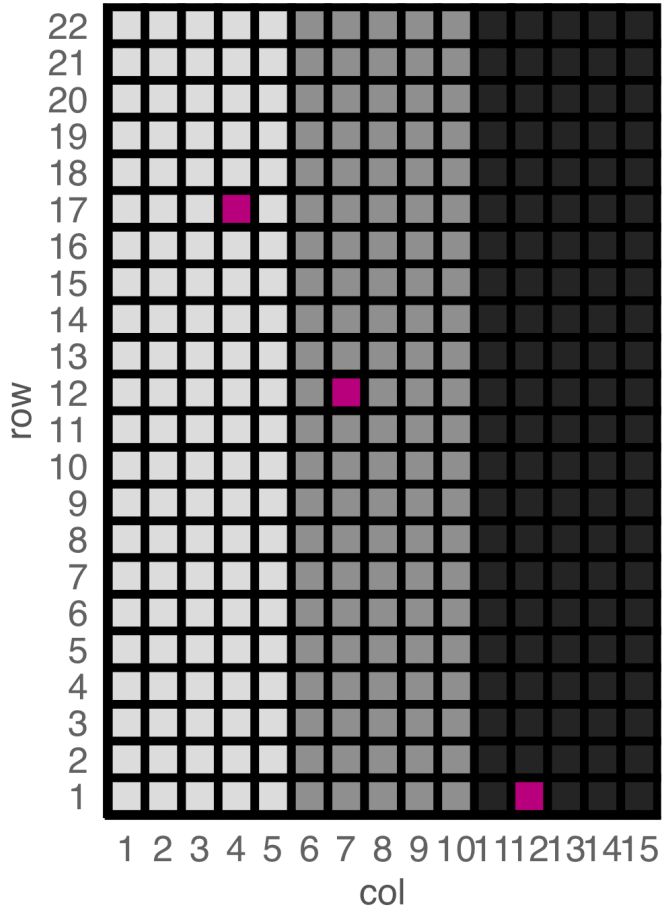
- The experiment employs what is referred to as a **randomised complete block design** (RCBD) (technically it is *near*-complete and not exactly RCBD due to check varieties have double the replicates of test varieties).
- RCBD means that
  - the there are equal number of replicates for each treatment (here it is gen);
  - each treatment appears exactly once in each block;
  - the blocks are of the same size; and
  - each treatment are randomised within block.
- In agricultural field experiments, blocks are formed spatially by grouping plots within contiguous areas (called `rep` here).
- The boundaries of blocks may be chosen arbitrary.

Experimental Design

Block structure



gen: (WqKPWmH*3Ag



Analysis

- In the main analysis, people would commonly analyse this using what is called **two-way ANOVA** model (with no interaction effect).
- The two-way ANOVA model has the form

$$\texttt{yield} = \texttt{mean} + \texttt{block} + \texttt{treatment} + \texttt{error}$$

- So for this data,

```
fit <- lm(yield ~ 1 + rep + gen,
          data = gilmour.serpentine)
```

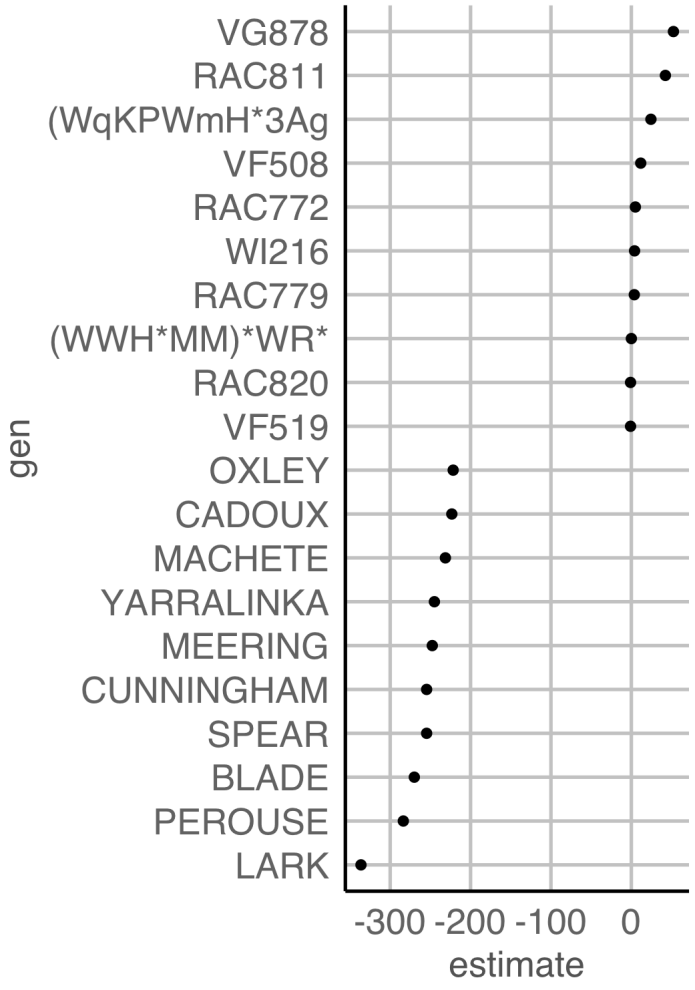
Analysis

```
summary(fit)
```

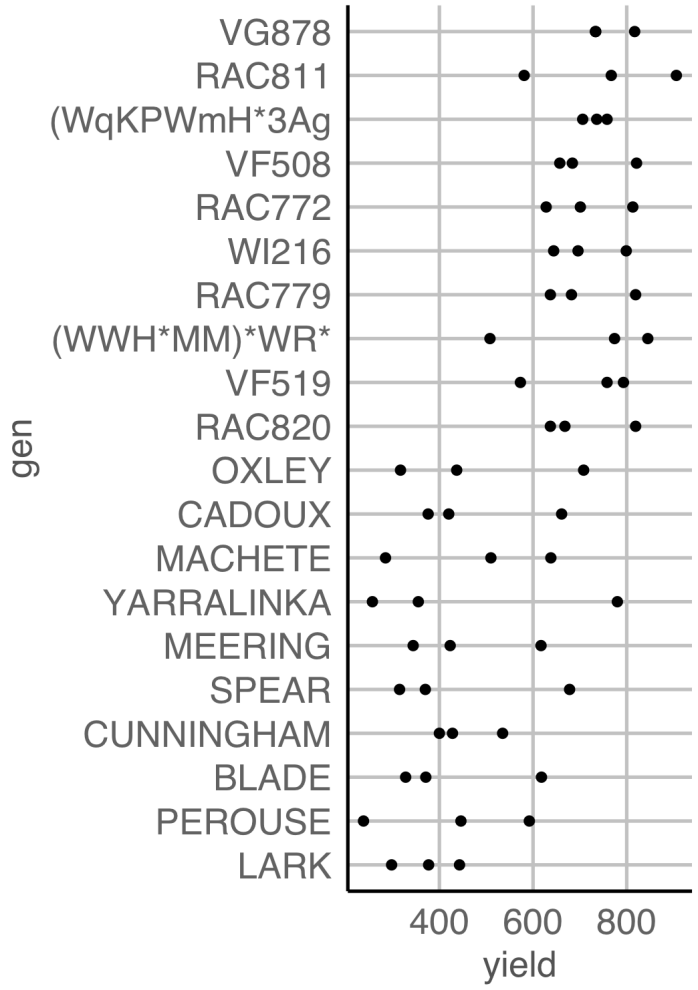
```
##
## Call:
## lm(formula = yield ~ 1 + rep + gen, data = gilmour.serpentine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -245.070  -69.695   -1.182   71.427  250.652
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      720.248     67.335  10.697  < 2e-16 ***
## repR2             96.100     15.585   6.166 3.29e-09 ***
## repR3           -129.845     15.585  -8.331 8.44e-15 ***
## gen(WqKPWmH*3Ag   24.333     94.372   0.258 0.796766
## genAMERY         -93.333     94.372  -0.989 0.323747
```

Case study 3 Wheat yield in South Australia Part 6/9

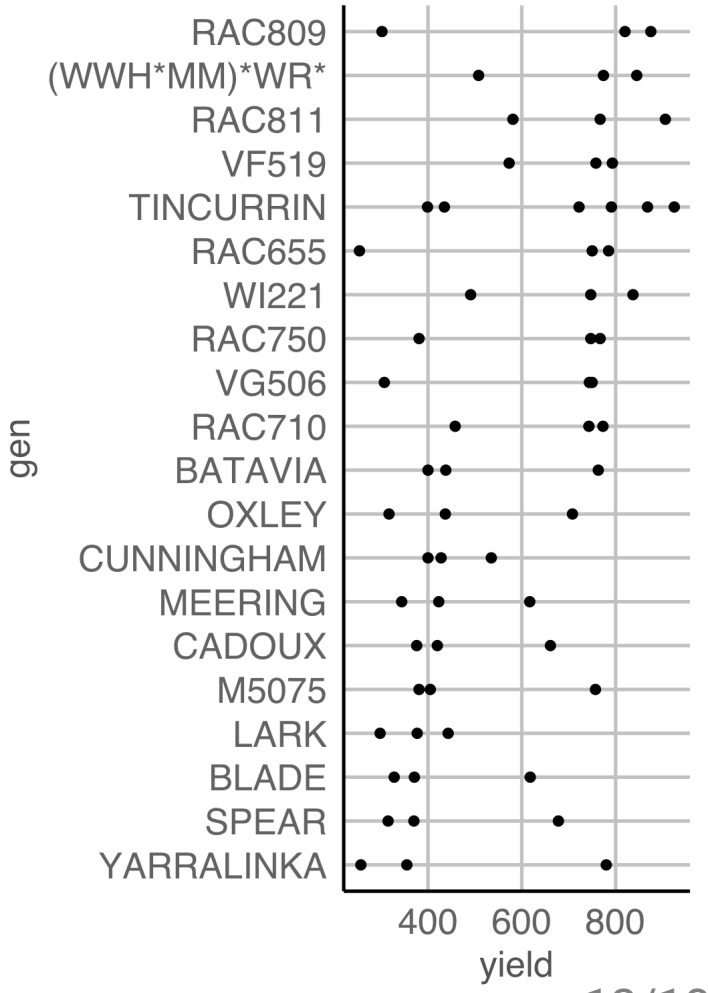
Top 10 and bottom 10 genotype by model est.



Top 10 and bottom 10 genotype by mean yield

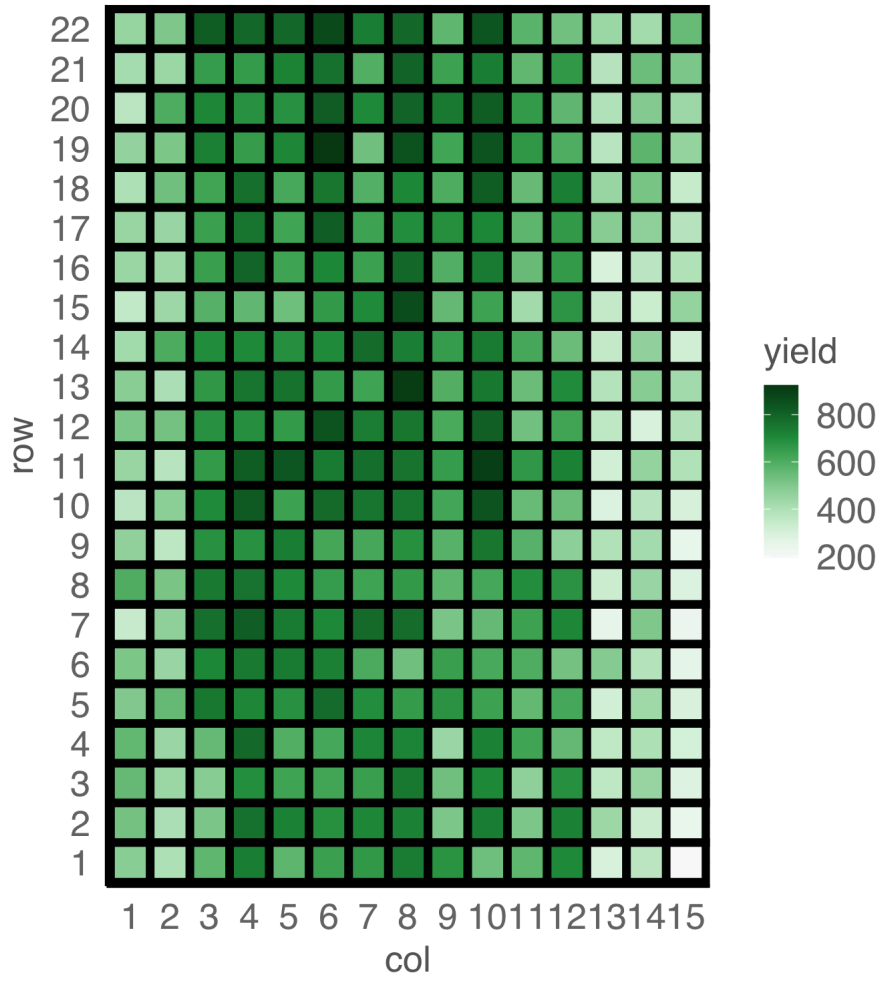


Top 10 and bottom 10 genotype by median yield

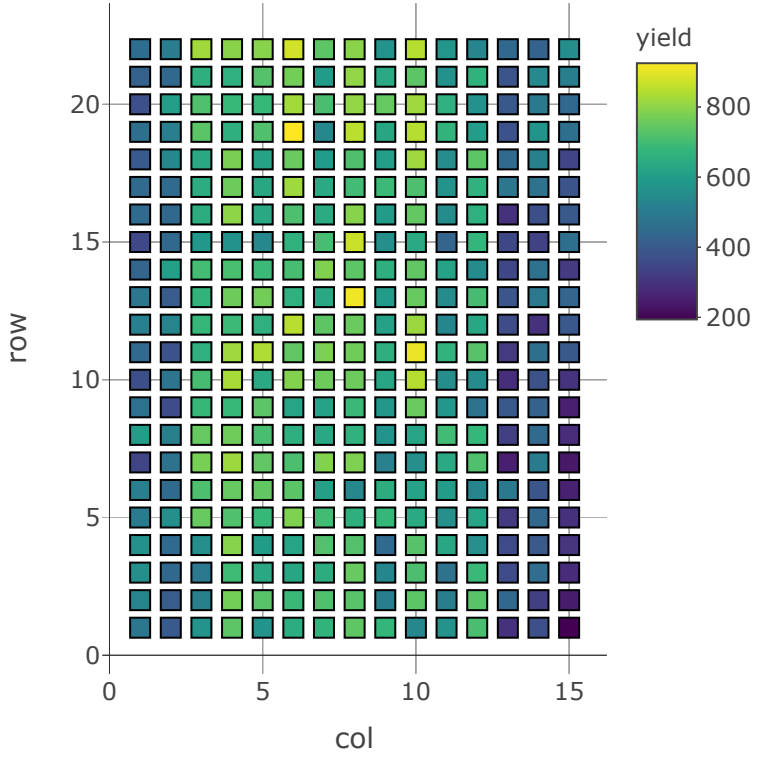


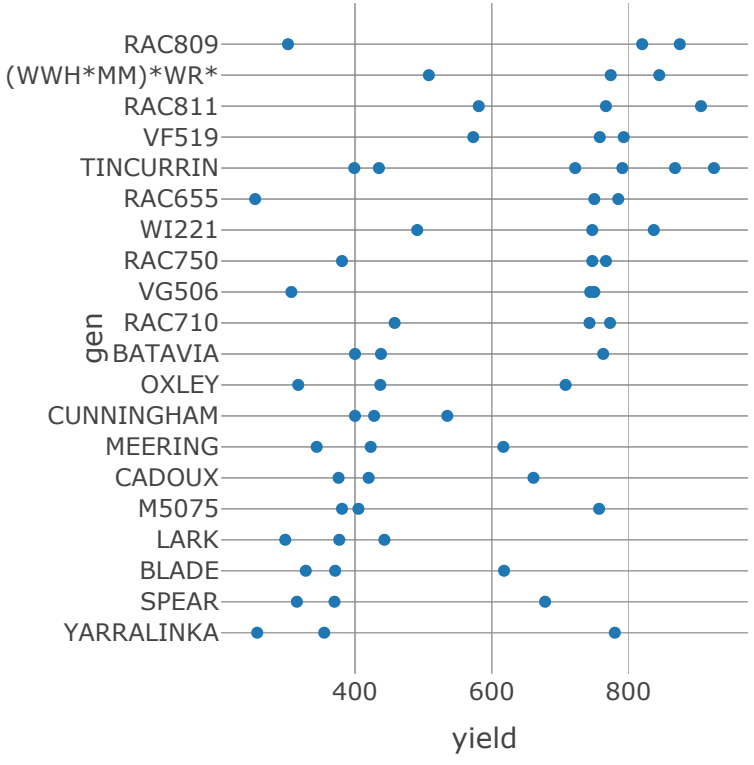
Case study 3 Wheat yield in South Australia Part 7/9

Do you notice anything from below?



Case study 3 Wheat yield in South Australia Part 8/9





Case study 3 Wheat yield in South Australia Part 9/9

- It's well known in agricultural field trials that spatial variations are introduced in traits; this could be because of the fertility trend, management practices or other reasons.
- In the IDA stage, you investigate to identify these spatial variations - you cannot just simply fit a two-way ANOVA model!

"Teaching of Statistics should provide a more balanced blend of IDA and inference"

Chatfield (1985)

Yet there is still very little emphasis of it in teaching and also at times in practice.

So don't forget to do IDA!

Summary

- ***Initial data analysis*** (IDA) is a model-focussed exploration of data with two main objectives:
  - ***data description*** including scrutinizing for data quality, and
  - ***model formulation*** without any formal statistical inference.
- IDA hardly sees the limelight even if it's the very foundation of what the main analysis is built on.

That's it, for this lecture!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu