MONASH University

ETC5521: Exploratory Data Analysis

Initial data analysis

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 3 - Session 2

Linear models in R REVIEW Part 1/3

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

```
## Rows: 50
## Columns: 2
## $ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13, 13, 14, 14, 1
## $ dist  <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34, 34, 46, 26, 3
```

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

Linear models in R REVIEW Part 2/3

- We can fit linear models in R with the `lm` function:

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

  is the same as

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

- The above model is mathematically written as

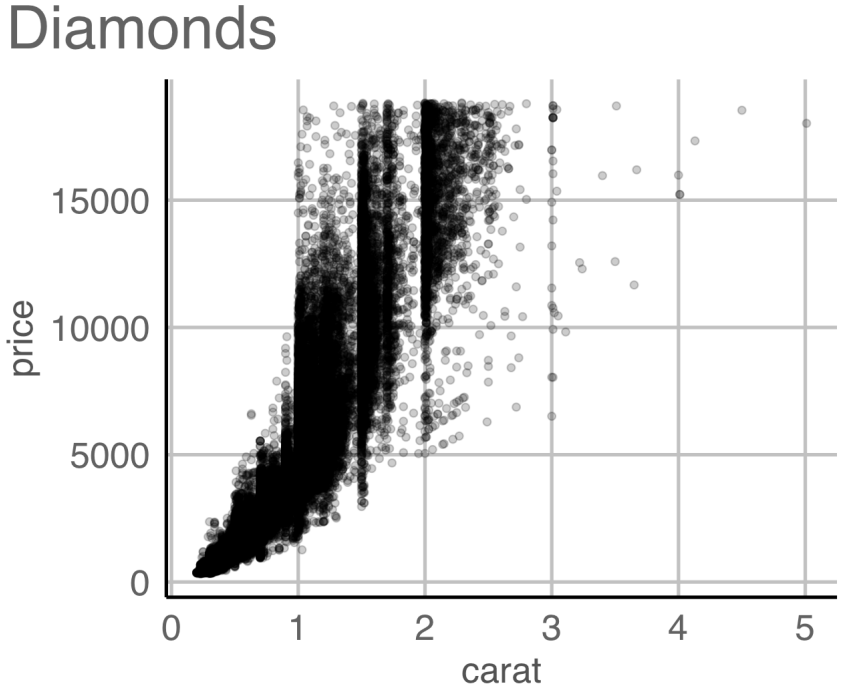
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

  where

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

❷ Model formulation Part 1/2

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

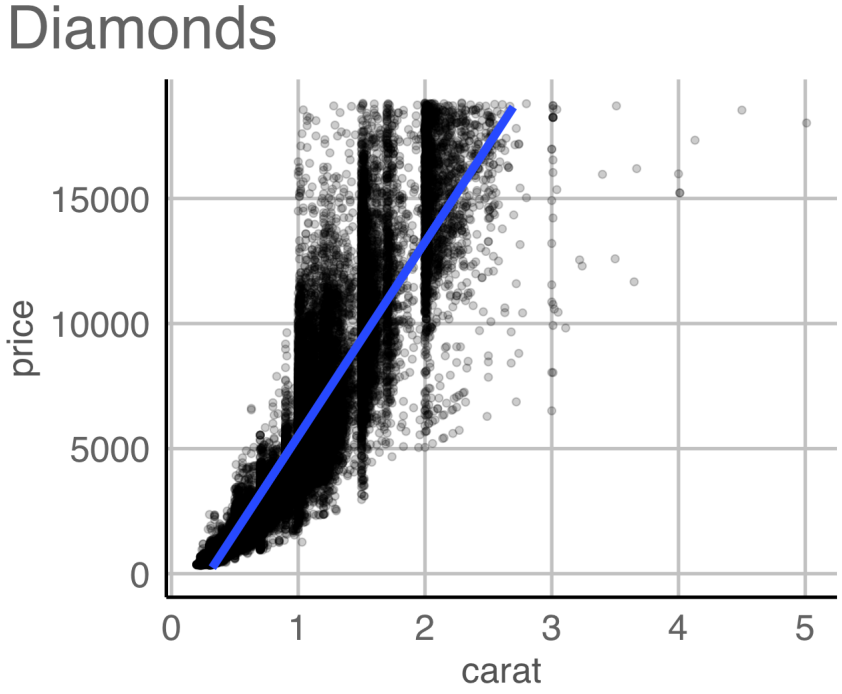


- Looking at this plot, would you fit a linear model with formula

`price ~ 1 + carat`?

2 Model formulation Part 1/2

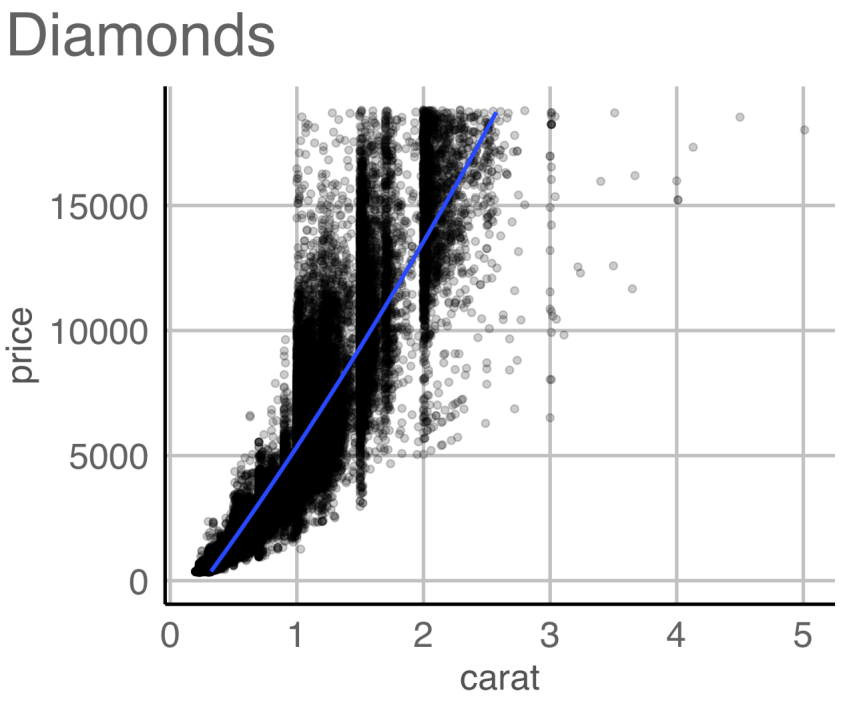
- Say, we are interested in characterising the price of the diamond in terms of its carat.



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

`price ~ 1 + carat`?

2 Model formulation Part 2/2



- What about

  `price ~ poly(carat, 2)`?

  which is the same as fitting:

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

- Should the assumption for error distribution 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 4 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
## _______________________
## Group variables            None
##
```

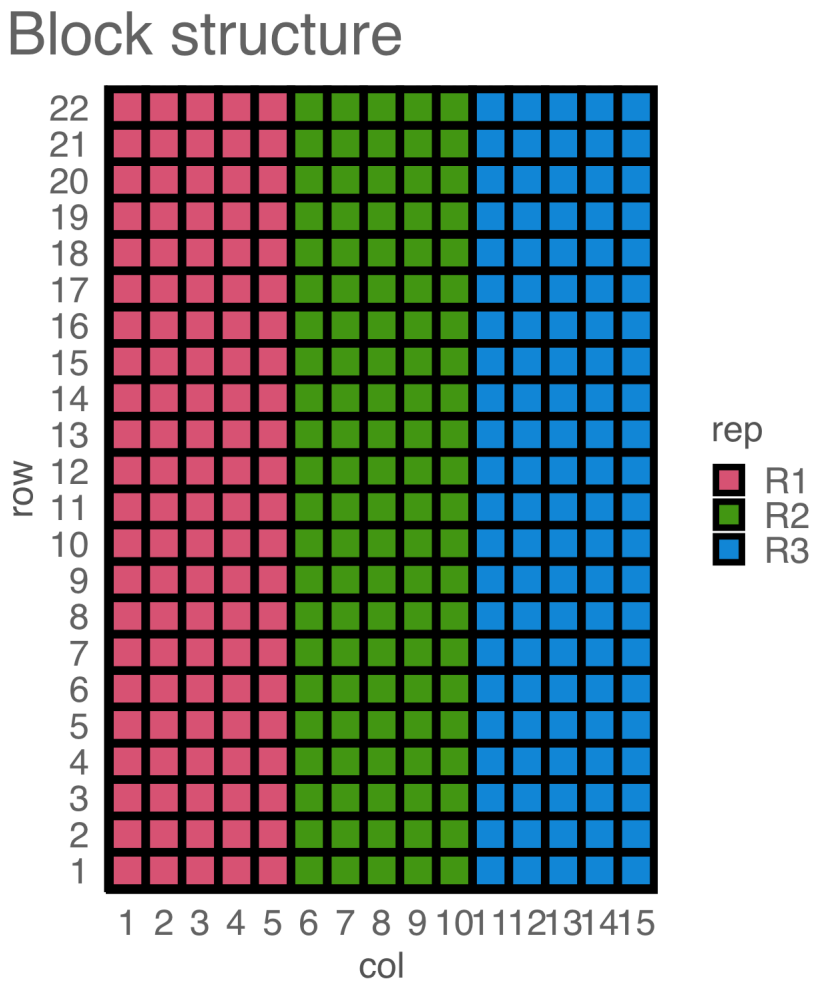
Case study 4 Wheat yield in South Australia Part 2/9

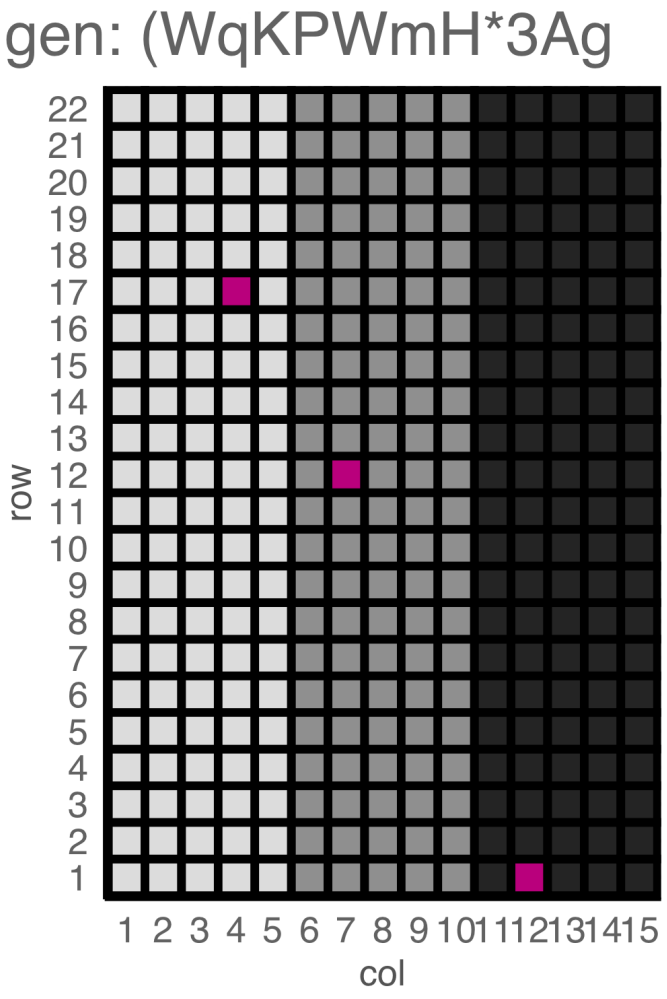
Experimental Design

- 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.

Case study 4 Wheat yield in South Australia Part 3/9

Experimental Design





Case study 4 Wheat yield in South Australia Part 4/9

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)
```

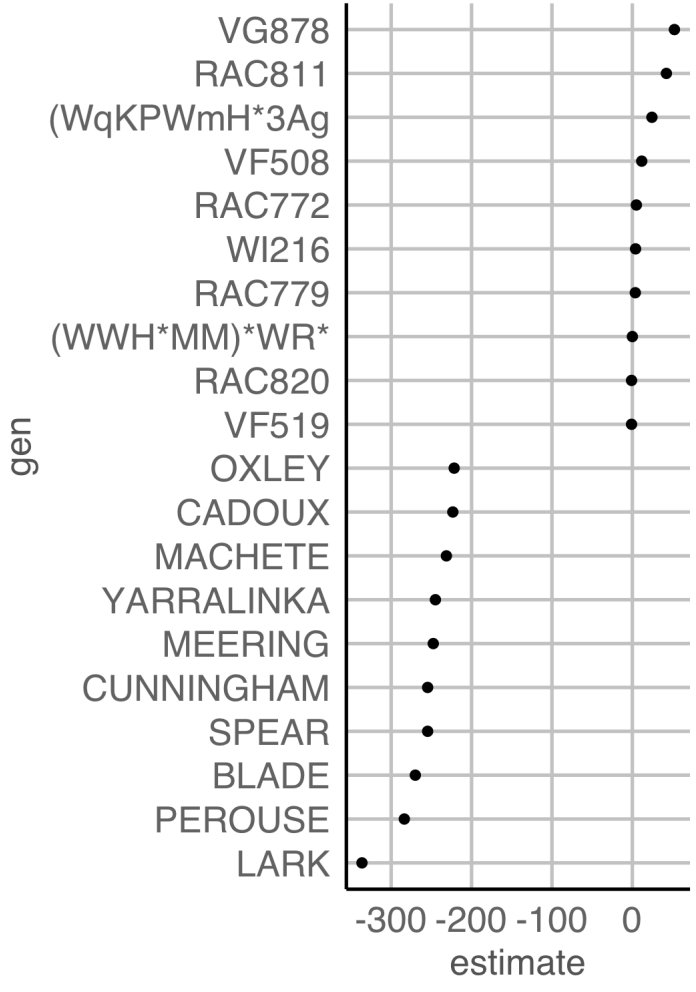
Case study 4 Wheat yield in South Australia Part 5/9

Analysis

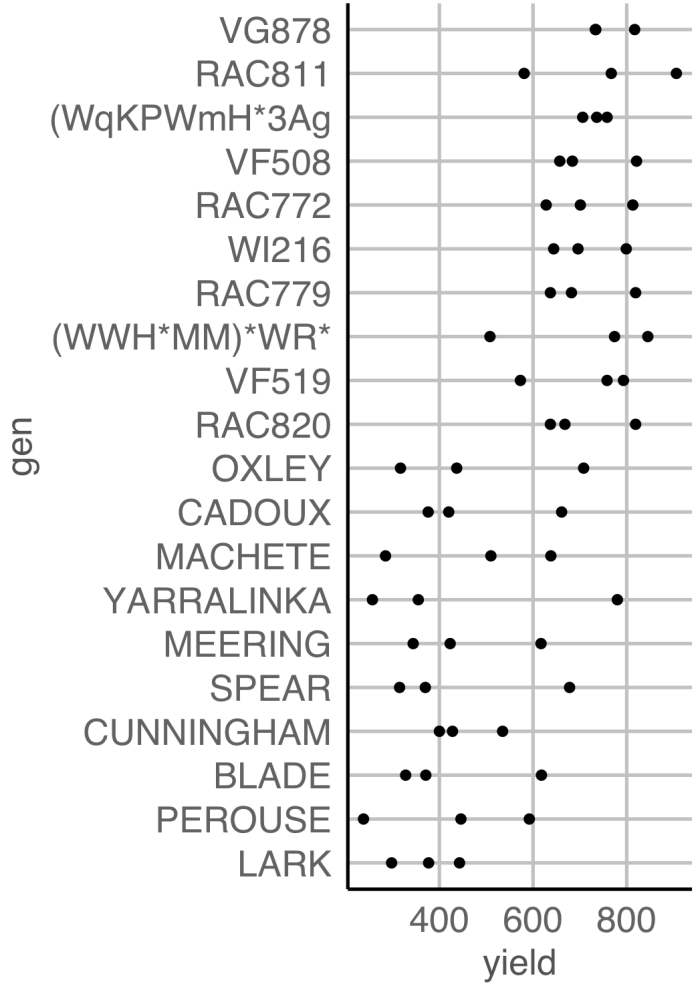
```
summary(fit)

##
## 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 4 Wheat yield in South Australia Part 6/9

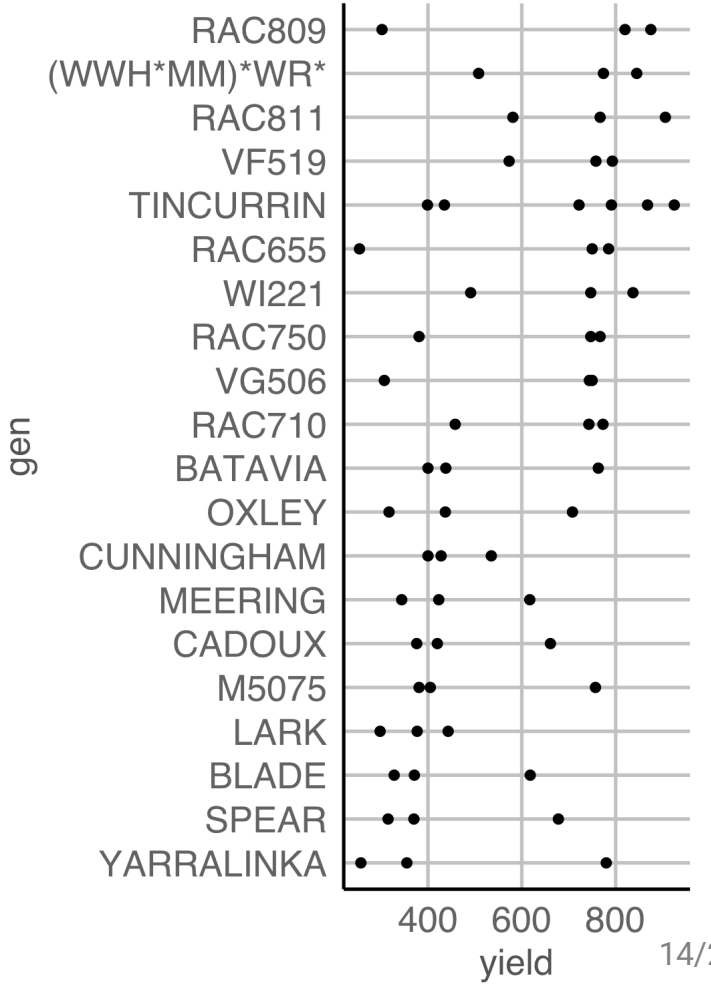
Top 10 and bottom 10 genotype by model est.



Top 10 and bottom 10 genotype by mean yield



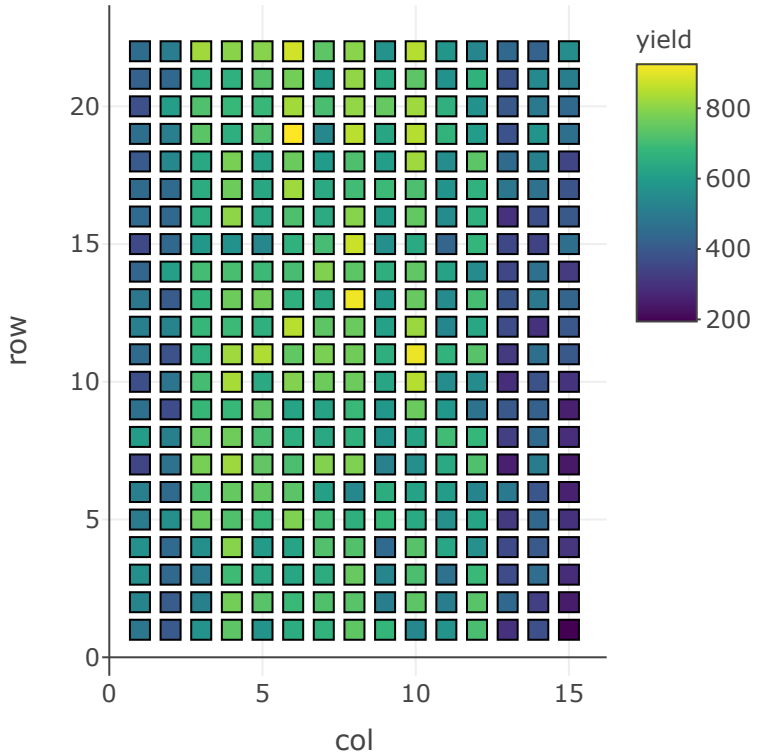
Top 10 and bottom 10 genotype by median yield

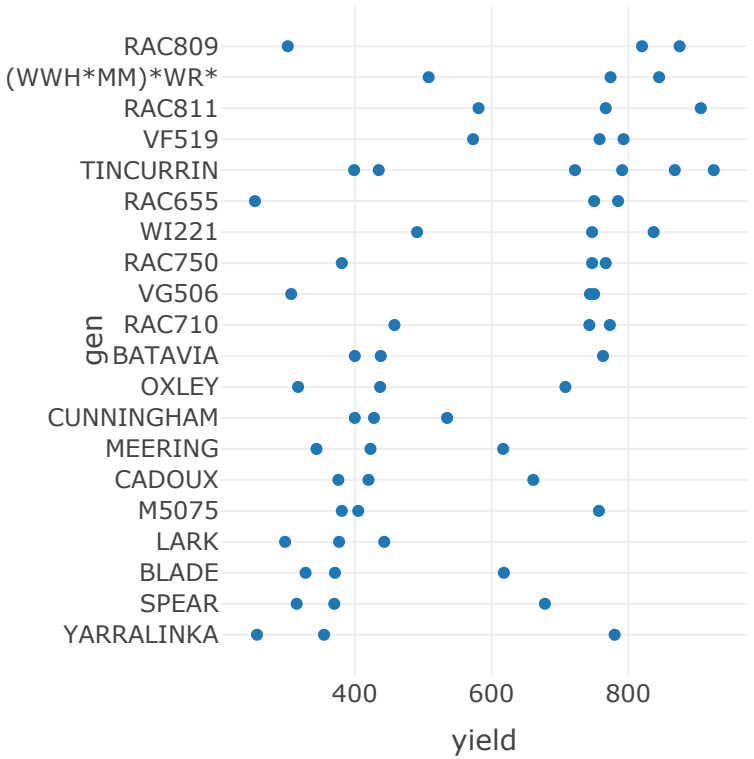


Case study 4 Wheat yield in South Australia Part 7/9

Do you notice anything from below?

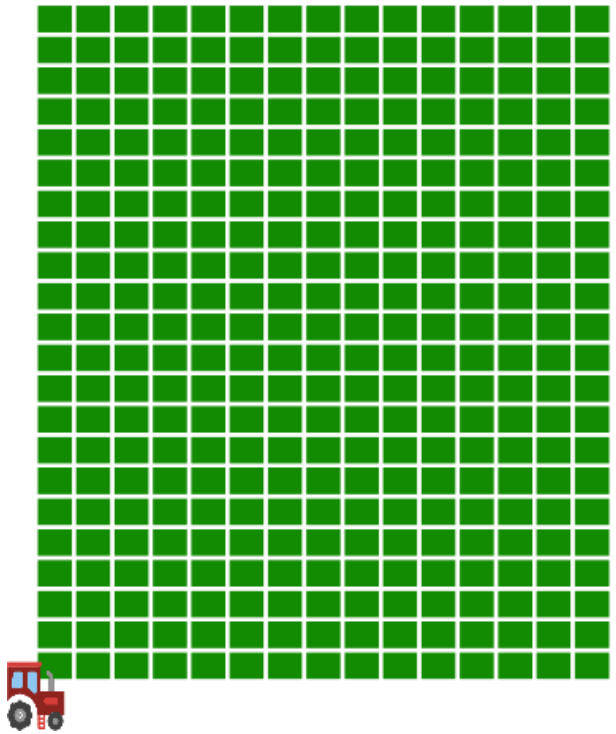
Case study 4 Wheat yield in South Australia Part 8/9

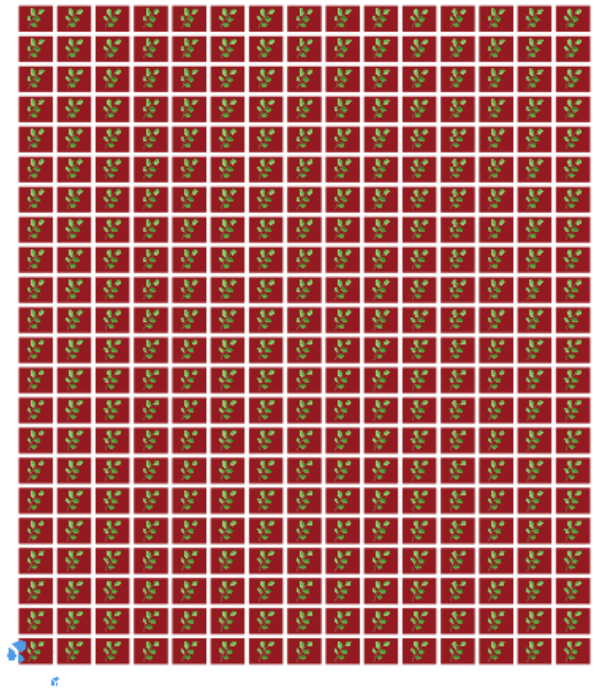




Case study 4 Wheat yield in South Australia Part 9/9

- 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!

Take away messages

- ***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.

MONASH University

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 3 - Session 2