Gaining iNZights from data

inzight

Chris Wild
University of Auckland

Slides etc at **http://bit.ly/icots10**

All links at

Abstract

inzight and web-based inzight lite have been developed to accelerate the rate at which students can experience data exploration, especially multivariate data exploration.

There are many "must-haves" in statistics education, and even more with the advent of "data science". However many of these imperatives actually conflict. Choices that make some things easy make others hard.

inzight has prioritised the ability to get useful output even when the user does not remember the names of appropriate techniques. We want people to get as quickly as possible to "Aha" moments about data without constantly being delayed by removable roadblocks.

We will show something of the capabilities of iNZight and iNZight Lite but embed this in a deeper discussion of educational priorities.

"Conflicting imperatives"

- **Imperatives are "must haves"**
  – We have lots in statistics education
- **Often contradictory** in that …
  – ***strategies*** *that are* ***good for one*** *are often* ***bad for others***

- **Leads to trade-offs & compromises**
  – True throughout statistics education
  – True in the design of statistical and educational software

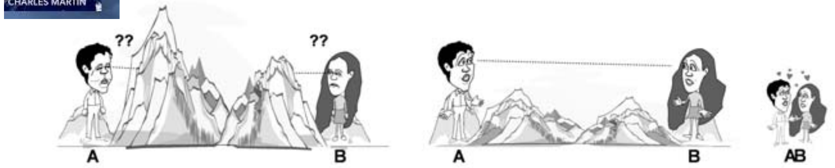
Long-term value

- **Technology is disrupting how we do things** at an increasing rate
  – Anything important that can be automated will be automated
  – ***Anything*** that is ***purely procedural can be automated***
  – So what should we most want for our students? The ability to do things machines can't do!
- **Almost none of what is "learned" in a university course sticks over the long term**
  – This makes prioritizing a small number of big-picture learnings, to be targeted for long-term retention, critically important

Minimising Cognitive Demands

- The average person ***can only hold two to six pieces of information*** in their attention ***at once***! (*Cowan, 2000*)

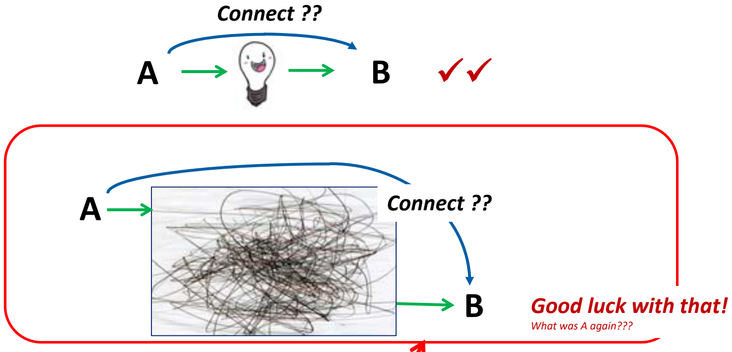
Moral: "***Short-term memory is a scarce* resource**. *Spend it wisely*!"



Wild (2007)

Making connections

What is your process ????

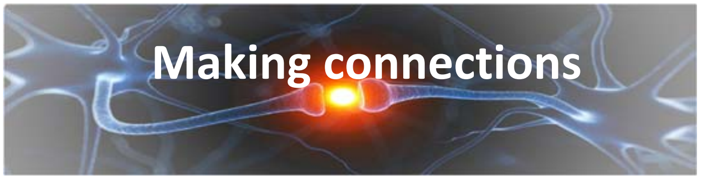


In a software enabled world … if this is your process …

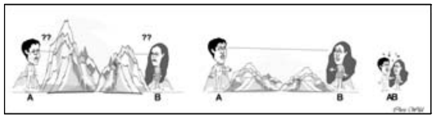
it better be because ***you value***

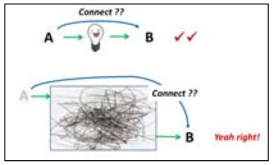
much ***more*** than ***connecting*** **A** and **B**

(because that's unlikely to happen)



Making connections





Every sequencing of ideas & experiences …

makes some things easy to connect (*even if only "Boy, this is boring!"*)

and others virtually ***impossible*** to connect

Opportunity Costs

*"**Time in teaching**/learning **is like water in the desert**"*

*"**Time spent on** learning **<u>anything</u> comes at the expense of** learning **something else**"*

*"**Time is our most precious resource**"*

*Moral: "**Time is a scarce resource** – use it wisely"*

Critical questions

- What are the *absolute must-have **fundamentals*** ?
- and what are the *nice-to-have **facilitators***?

The answers are context dependent and I won't go there today

What do you most value?

For me, for **early statistics …**

- Getting turned on to statistics/data science by experiencing …
  - the **joys of *discovery***
  - a ***feeling of empowerment***
    - I can do some cool and powerful stuff with data
  - the power of statistical ways of looking at the world
- and some big ideas for statistical-literacy

*Getting to "**Aha!!**"*

- **+** (if there'll be sufficient time) …
  - other things that'd be useful for life & further learning

Main strategy

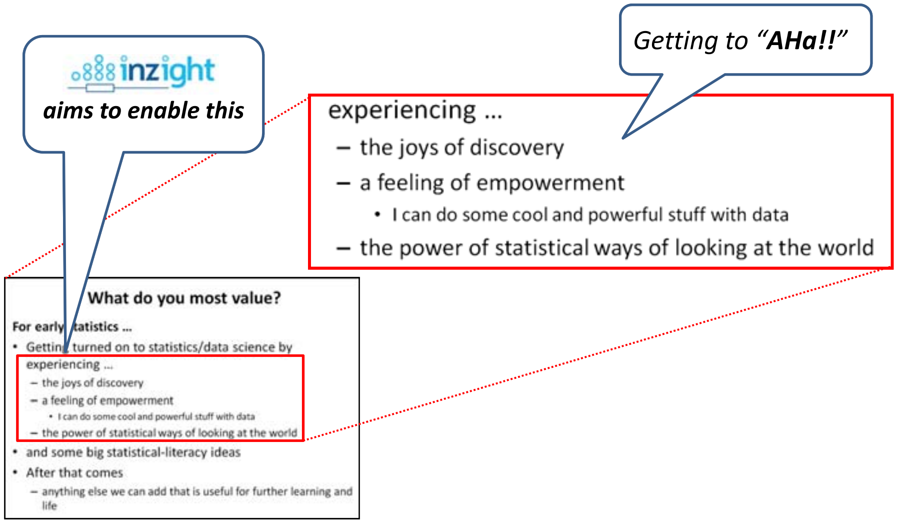
- ***First establish big picture visions and their value***
  - aiming for retention of what matters most …

"… and the vision that was planted in my brain still remains …"

– Paul Simon, "*The Sound of Silence*"

- ***Then backfill details later** (if the opportunity arises)*

What do you most value?


inzight
aims to enable this
*Getting to "**AHa!!**"*

experiencing …
- the joys of discovery
- a feeling of empowerment
  - I can do some cool and powerful stuff with data
- the power of statistical ways of looking at the world

What do you most value?

For early statistics …

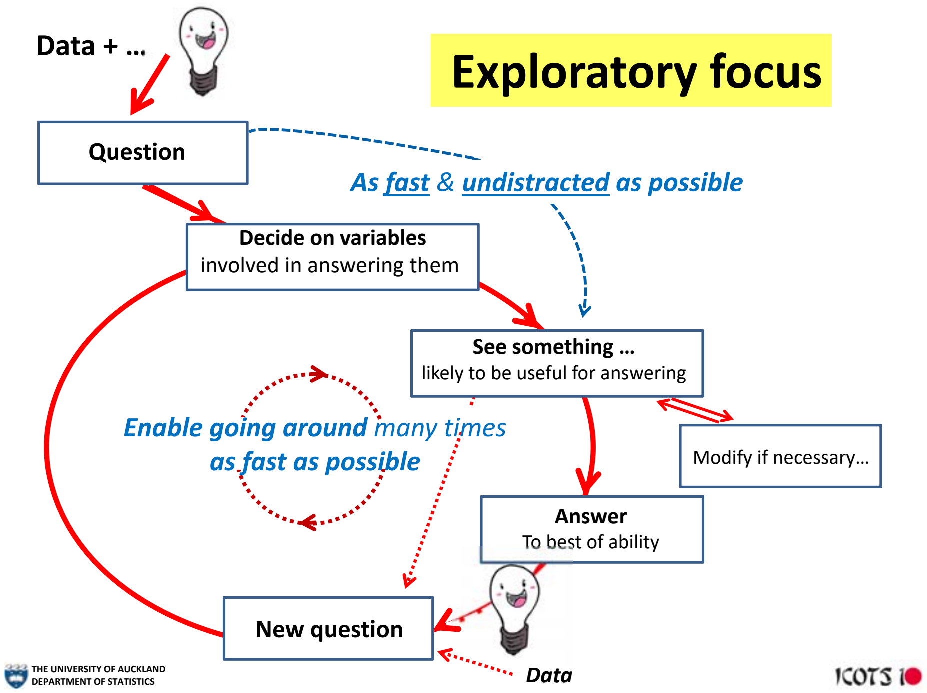
- Getting turned on to statistics/data science by experiencing …
  - the joys of discovery
  - a feeling of empowerment
    - I can do some cool and powerful stuff with data
  - the power of statistical ways of looking at the world
- and some big statistical-literacy ideas
- After that comes
  - anything else we can add that is useful for further learning and life

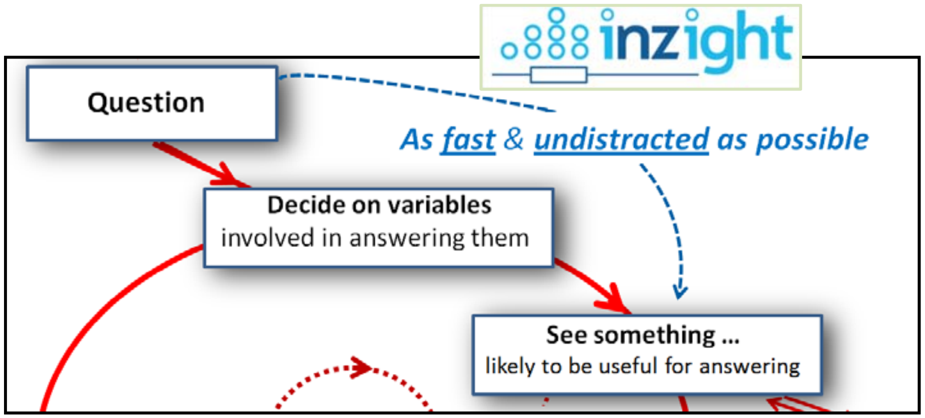
Highest priority

inzight inzight lite

- ***Enable** even beginners **to explore multivariate data** very **rapidly** and with a **minimal learning curve***
  - **Corollary** (to "minimal learning curve")**:** *can't make people learn & remember a lot of things before they can get useful displays*

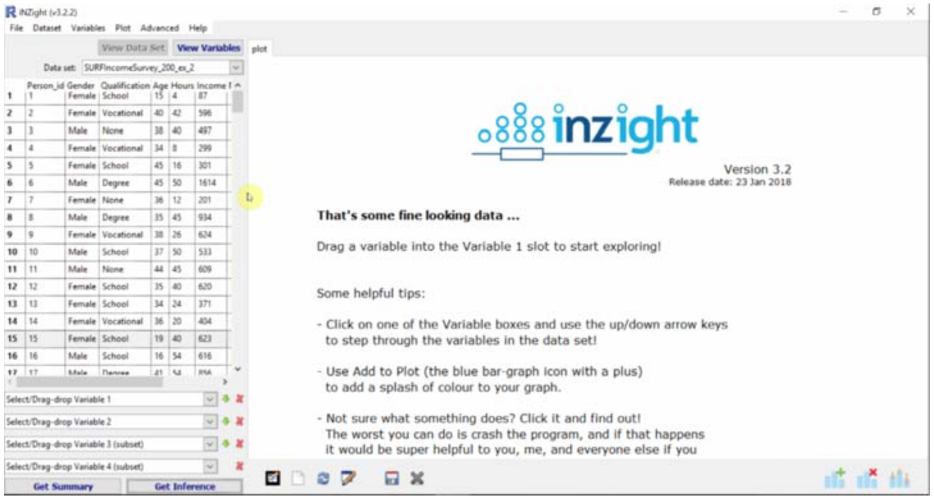
Exploratory focus




inzight
Question
As *fast* & *undistracted* as possible
Decide on variables
involved in answering them
See something …
likely to be useful for answering

instant gratification!

Show something


iNZight (v3.2.2)
File
Dataset
Variables
Plot
Advanced
Help
View Data Set
View Variables
plot
inzight
Select/Drag-drop Variable 1
Select/Drag-drop Variable 2
Select/Drag-drop Variable 3 (subset)
Select/Drag-drop Variable 4 (subset)
Get Summary
Get Inference

Data set: SURFIncomeSurvey_200_ex_2

	Person_id	Gender	Qualification	Age	Hours	Income
1	1	Female	School	15	4	87
2	2	Female	Vocational	40	42	596
3	3	Male	None	38	40	497
4	4	Female	Vocational	34	8	299
5	5	Female	School	45	16	301
6	6	Male	Degree	45	50	1614
7	7	Female	None	36	12	201
8	8	Male	Degree	35	45	934
9	9	Female	Vocational	38	26	624
10	10	Male	School	37	50	533
11	11	Male	None	44	45	609
12	12	Female	School	35	40	620
13	13	Female	School	34	24	371
14	14	Female	Vocational	36	20	404
15	15	Female	School	19	40	623
16	16	Male	School	16	54	616
17	17	Male	Degree	41	54	856

Version 3.2
Release date: 23 Jan 2018

That's some fine looking data ...

Drag a variable into the Variable 1 slot to start exploring!

Some helpful tips:

- Click on one of the Variable boxes and use the up/down arrow keys to step through the variables in the data set!

- Use Add to Plot (the blue bar-graph icon with a plus) to add a splash of colour to your graph.

- Not sure what something does? Click it and find out! The worst you can do is crash the program, and if that happens it would be super helpful to you, me, and everyone else if you

Decisions and responses

- **Human decision** – Answer …
  - "**What** combination of **variables do you want to look at?**"
    - **Necessary decisions** for <u>self-guided</u> exploration
- **System response …**
  - Give useful ***Graphics*** instantly
    - but let's you change that if you want
  - On ***Get Summary***
    - "Give me the ***summary information people usually want*** in a situation like this"
  - On ***Get Inference***
    - "Give me the ***inferential information people usually want*** in a situation like this"

For the basics that's all you have to do

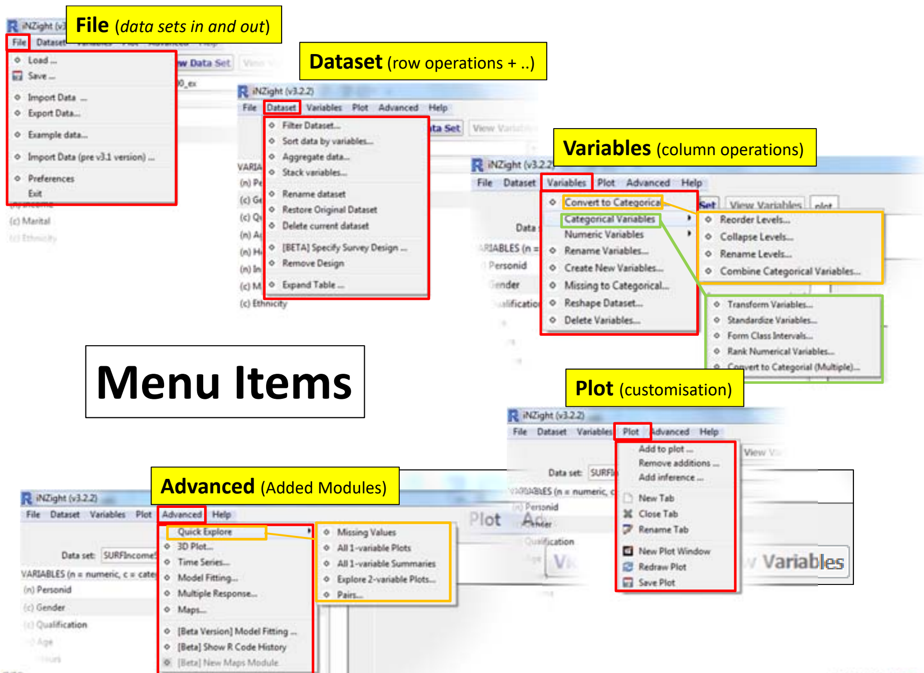
What are we prioritising?

- **Really fast for** (multivariate) **visual exploration**
- **Takes very little time to get competent**
- **Few demands on** (human) **memory**
  - Don’t have to remember lots of terms/names as a prerequisite to getting things from the software
  - Takes very **little time to get back up to speed** …
    - when you come back to it after not using it for a long period
      - unlike using most menu-driven systems or writing program code where fading memories can slow you to a crawl
- ***“I wonder what this does?” …***
  - leads people to discover new possibilities

- Emphasize the decisions that have to be human
- ***Replace*** reduced time & effort in ***“getting things” by*** increased time & effort on ***making meaning***
- Because in the long game ….

> ***“Meaning trumps mechanics”***
>
> — from *“Statistical Literacy as the earth moves”*

Menu Items



File *(data sets in and out)*

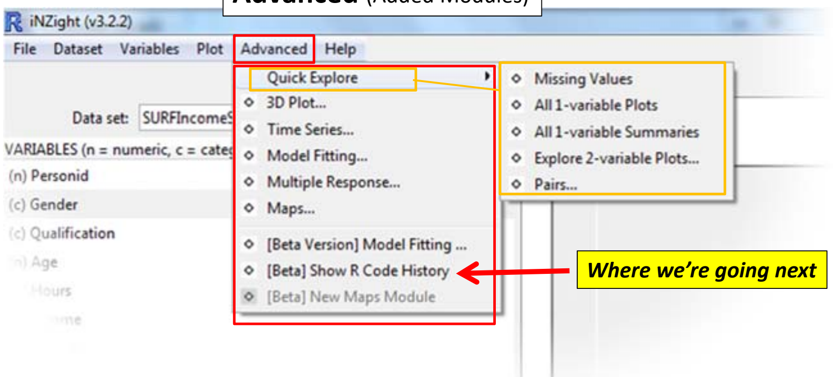
Dataset (row operations + ..)

Variables (column operations)

Plot (customisation)

Advanced (Added Modules)

Advanced (Added Modules)



Where we’re going next

Coding vs point-and-click

Ross Ihaka on point-&-click software …

Good point-and-click systems …

can allow users to …

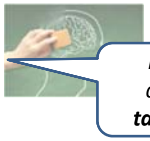
- access many capabilities with minimal learning curves

***Coding** takes big investment in time and effort*

Learning** a language **takes a long time

Menu choices …

- Answer, **"What's on offer here?"**
- act **as reminders** to counter fading memories



Forgetting** a language **takes no time

Makes point-and click systems good for …

- beginners and occasional users
- doing one-off things really fast (provided the system prioritizes them)

Can enable us to see a whole range of things we can do with our data
and do them very quickly and with very little effort

Advantages of coding

- **Audit trail** of what was done and how
  - Most importantly, any changes made to the data along the way can be seen
- **Reproducibility**
  - someone else can reproduce an analyst's results easily
- **Flexibility and extensibility**
  - With point-and-click interfaces it can be next to impossible to do anything beyond what the software explicitly provides for.
- **Long-run time-efficiency**
  - automation of repetitive tasks
    - speed advantage of a set of point-and-click choices disappears when it is realised that the data that was used should have been changed in some way so that you have to do it all over again
  - old code speeds you up whenever do something similar to something you/someone else have done before
- **Reproducible workflows** and integration in **dynamic documents**
  - expository text interspersed with blocks of code
  - documents are then compiled to produce a report/slides/thesis/book/workflow history
    - When you discover you need to change something that affects the data and will have downstream effects you do not have to do a lot rework, you just have to make a small local change and recompile the document

Moral:
Intending** statistical/data science **professionals
need to learn to code

On having our cake and eating it too

Point-and-click systems that ***not only perform actions*** but also …
make available the code that implements those actions

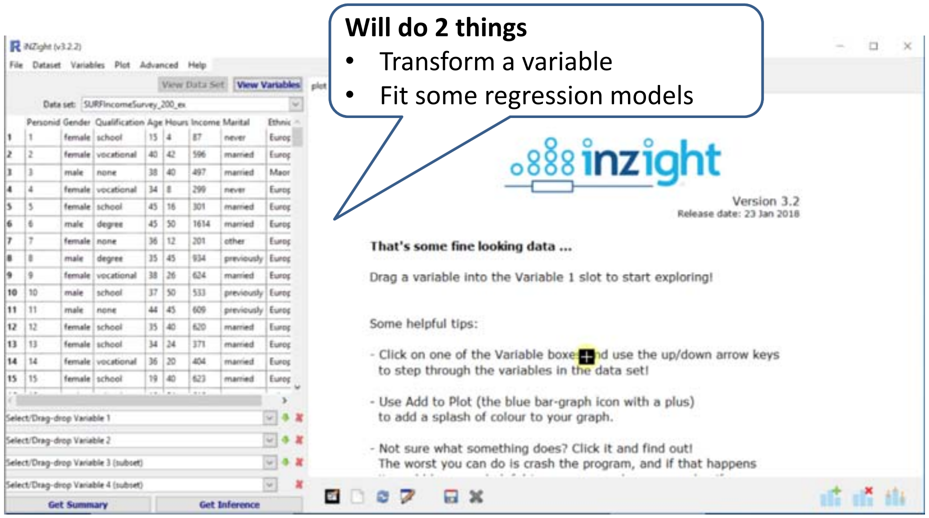
- R-Commander, which is probably the best known menu-driven interface to R, has been doing this for many years but it is a complicated system and you have to know a lot of statistics to be able to use it
- iNZight has started to do this too

- as an aid for students in learning to code
  - *Ice-breaker* role
    - modifying small pieces of code that do something obviously useful to change behaviour
  - and a *give-me-the-code* role
- and for the other code-benefits: it provides
  - code that can be modified and re-run, shared, or put into dynamic documents
  - and audit trails etc
- End aim is to be able to interact with the system through both the interface and R code
  - and even to have iNZight write R markdown documents combining commentary and outputs

Show something!!!

Will do 2 things

- Transform a variable
- Fit some regression models


iNZight (v3.2.2)
File
Dataset
Variables
Plot
Advanced
Help
View Data Set
View Variables
plot
Select/Drag-drop Variable 1
Select/Drag-drop Variable 2
Select/Drag-drop Variable 3 (subset)
Select/Drag-drop Variable 4 (subset)
Get Summary
Get Inference
inzight

Data set: SURFIncomeSurvey_200_ex

	Personid	Gender	Qualification	Age	Hours	Income	Marital	Ethnic
1	1	female	school	15	4	87	never	Europ
2	2	female	vocational	40	42	596	married	Europ
3	3	male	none	38	40	497	married	Maor
4	4	female	vocational	34	8	299	never	Europ
5	5	female	school	45	16	301	married	Europ
6	6	male	degree	45	50	1614	married	Europ
7	7	female	none	36	12	201	other	Europ
8	8	male	degree	35	45	934	previously	Europ
9	9	female	vocational	38	26	624	married	Europ
10	10	male	school	37	50	533	previously	Europ
11	11	male	none	44	45	609	previously	Europ
12	12	female	school	35	40	620	married	Europ
13	13	female	school	34	24	371	married	Europ
14	14	female	vocational	36	20	404	married	Europ
15	15	female	school	19	40	623	married	Europ

Version 3.2
Release date: 23 Jan 2018

That's some fine looking data ...

Drag a variable into the Variable 1 slot to start exploring!

Some helpful tips:

- Click on one of the Variable boxes and use the up/down arrow keys to step through the variables in the data set!
- Use Add to Plot (the blue bar-graph icon with a plus) to add a splash of colour to your graph.
- Not sure what something does? Click it and find out! The worst you can do is crash the program, and if that happens

Thank you

But I'll leave you with this …

On having our cake and eating it too

"I worry about starting too early with xxx. Yes we need to teach statistics majors to deal with xxx.

But extracting jewels from gloop is not something most people do because they love messing around in gloop.
They want the jewels. But first they have to know (i) that jewels exist, and (ii) they might be in there

So let's first have them discover jewels in places where they are easier to find. ... all these things slow down what you can see and how fast you can see it.

There should be a sniff test. Is this an enticing element of courtship? Or do I feel the skin-pricks of glass shards? So should we save it for after marriage? Or at least till after moving in?"

[*from "Further, Faster, Wider"* (2015)]