ANALYSING AND PRESENTING DATA
PART 1 – ANALYSING DATA

Professional Development Course Book

Presented by Mark Priadko

INSTITUTE OF
PUBLIC ADMINISTRATION
AUSTRALIA

IPAA
SOUTH AUSTRALIA

IPAA Personal Membership

Become part of the IPAA SA Personal Member community and access an extensive range of benefits, opportunities and support for your career.

Personal Membership Will…

Get you future ready

We will improve your career as we educate and inform you of what's to come in the future of the profession and build your skill set to ensure you are ready for it.

Build your capability

We will build your current capability to ensure you perform at the best of your ability and get you ready to face every new challenge throughout your career.

Champion you

We will champion your successes and voice them to the sector.

Build your networks

We will build your networks and connect you to like-minded people, people facing the same challenges as you and people who can help you to gain influence and penetrate the vast public sector.

Connect you

We will connect you with ideas from across the country through our National body and State jurisdictions.

Support you

We will provide you with tangible support throughout your career.

More benefits

- Extensive discounts on IPAA SA training, courses, forums and events.
- Access to exclusive member opportunities, such as our mentorship program and networking sessions.
- Receive resources and learning opportunities relevant to your career level.
- Receive a membership certificate for your CV and use of the MIPAA post nominal.
- Unlimited digital access to the Australian Journal of Public Administration.
- Receive a training voucher to the value of your membership.
- Have your say in IPAA SA governance.

How to join?

Joining is as easy as visiting our website: www.sa.ipaa.org.au/membership. Navigate to the Membership section and select the membership type most applicable to you and fill in your details online. Membership fees are generally tax-deductible.

If you have any questions about membership, contact us on 8212 7555 or email membership@sa.ipaa.org.au

Learning objectives:

- Understand where analysis fits and what purpose it serves
- Introduction to different types of analysis
- Practical application of analysis techniques

Self-evaluation

My proficiency in analysing and presenting data is:

List of what I need to know to move up one point on this scale

- .....
- .....
- .....
- .....

These course notes are designed to support the presentation of information in the module. They are based on the knowledge and experience of Mark Priadko.

Table of Contents

Quick questions

What is your best guess of the average life expectancy in Australia in 2023?

Looking back, what is your best guess of the average life expectancy in Australia in 1923 (one hundred years ago)?

What is your best guess of the average life expectancy in the UK in the early 1820s (two hundred years ago)?

Discussion notes

What purpose does our data analysis serve? Why is it important?

What is your understanding of what a narrative is?

Overview

Data and storytelling

In the public sector, we work with data:

- To monitor and evaluate services
- To understand the community – demographics, use of services, satisfaction with services, attitudes, behaviour, movement
- To understand the economy – labour markets, exports, imports, and investment
- To understand the environment – climate, rainfall, temperatures, water flows, waste management, flora and fauna
- To determine funding – hospital and school funding is based on activity data.

We analyse data to understand the past – what has happened, how much has happened, where things happen. We use the data to help understand why things happen so we can apply that understanding to future events.

We use data to monitor the past.

We use data to build predictive models and guide future decision-making.

We present data:

- For the public record
- To inform the public, leaders and managers
- To help us formulate strategies and options
- To help us make decisions.

Presenting data to others to help them make decisions requires that we can tell a story of change with the data we have.

To tell a story, we have to make choices about what of our analysis to include and what to exclude. We have to help our readers discover without them experiencing all the difficulties we have or without having to digest all the data we have.

The choice of what to include and how to include it, is our narrative.

> *"Narrative is the choice of which events to relate and in what order to relate them – so it is a representation or specific manifestation of the story, rather than the story itself."*
>
> Source: http://beemgee.com/blog/story-vs-narrative/

> *"The concept of narrative deals more with **how** the events are told. Narrative is the ordering of events into a consumable format.*
>
> *…… narrative is the method and means by which you construct the events of a story into a plot. It concerns itself with the sequence of the events, the medium on which they are told and the way these events are put together into one coherent unit."*
>
> Sourced from: http://hacktext.com/2011/09/story-vs-narrative-vs-plot-1205/

Narrative is the way the author or speaker chooses to structure events — the architecture of the story, comparable to the design of a building. Ultimately, a narrative is a way of organising the information of a story, strategy or proposal that helps it engage a reader and make sense to a reader or listener. While a story is a sequence of events, the narrative recounts those events in its own way to emphasise some aspects of the story to enhance its impact.

This workshop will consider a range of choices we can make:

- Choices of inclusion and exclusion – when confronted with too much information, we need to prioritise what is more important and include that while excluding other information.
- Choices of summarisation – we can consider numbers that summarise large volumes of data. Examples include totals, subtotals and averages.
- Choices of aggregation and consolidation – grouping items together is another way to reduce the information we provide
- Choices of sequence – how we order information influences the way we tell our stories (e.g. telling vs discovery is a way to consider sequence)
- Choices about whether to use tables or graphs
- Choices about which graph to use
- Choices about how we use space with our data visualisations.

In Government, we use data to tell the stories of our policies and of the services we deliver. They are largely advisory stories.

I find it helpful to distinguish the types of stories (documents and reports) that I present to leaders and executive groups, which will belong to one of three categories:

1. Inform (and to provide assurance)
2. Prompt discussion
3. Trigger decision-making.

The distinguishing feature between these is the nature of change and action arising from them.

No change required	Change required but how is not yet clear	Change required and understood
Information, Assurance	**Discussion**	**Decision Making**
Recommends - Noting	Recommends – direction for further work	Recommends – action to implement change

Information, Assurance & Monitoring

Stories designed for this purpose provide recipients with background and details that are important for them to know and let leaders know that previously approved projects or changes are on track.

Reports designed to provide leaders with information or assurance will usually recommend that the reader note the contents of the report and will not demand any approvals or recommend any new action.

Examples will include background briefings, status reports on finances, operations, and compliance reports where progress is within acceptable tolerances and no decision making or action is required of the recipients.

Discussion Reports/Papers

When an issue has arisen that will ultimately require change in the way a service is delivered or in policy, but the detail of the change is not yet fully developed a discussion narrative (paper) can

be provided to examine current opinion and evidence on the issue. This type of narrative should make recommendations on what further work or steps are required to get resolution or to be able to recommend the specific changes that are required.

Discussion papers may be necessary for controversial, complex or high risk issues where there is no clear solution apparent at the time or for issues that are of a large scale that involve careful consideration in the development of options and before a final decision is made.

Examples will include reports on matters of strategic importance like changes in the business model, changes in policy, adding or cutting services or major investment projects that need to be discussed amongst leaders before a final decision.

Decision-making

If an organisation is dealing with problems where research has been made into the changes necessary, executive teams will rely on advice as to how to go about dealing with problems or making changes.

In this respect, our narratives, documents and reports provide the information necessary to prompt and support decision-making.

Narratives designed to prompt or support decision-making will make recommendations that seek approval to make change or undertake new work.

Examples: Proposals for organisational initiatives or campaigns, business cases to proceed with a project, recommended changes in policy, findings and recommendations from a service or functional review.

There is a greater array of possibilities for the fiction narrative structure to invoke the imagination of the reader/listener, to create surprise, mystery and intrigue. For non-fiction narratives, the variety of structures available to us will be more limited. There will tend to be greater reliance on more straightforward and logical structures, although surprise and intrigue can still be used, their use will be more limited.

Data Analysis – Context and Principles

Data - its benefits and its limitations

Data is an abstraction of a real-world entity (person, object, or event).

Data represents an effort to measure attributes of our world so we can better understand and communicate them.

As an example, for our employees, we use data to understand things about them and their relationship with our business. We collect data on:

- Age
- Gender
- Address
- Classification
- Qualifications.

We can never understand everything about our employees and, for privacy reasons, we never want to record or understand many things about them.

Data can never fully represent the reality of our world - there is no such thing as perfect data. Data can never fully capture the complex details of a real-world entity like an employee.

We use data to move up a knowledge continuum, understanding that we can't reach the end on the right-hand side.

What we know about an aspect of our world	Nothing	Everything

We use data to move up the continuum →

We use data to support decision-making. All decision-making occurs in an environment of uncertainty and with the risk of biases. We rely on data to reduce uncertainty and reduce biases in decision-making. Just as data cannot perfectly represent our world, we cannot eliminate uncertainty. Data is used to help us make sense of our world and the decisions we make by reducing uncertainty.

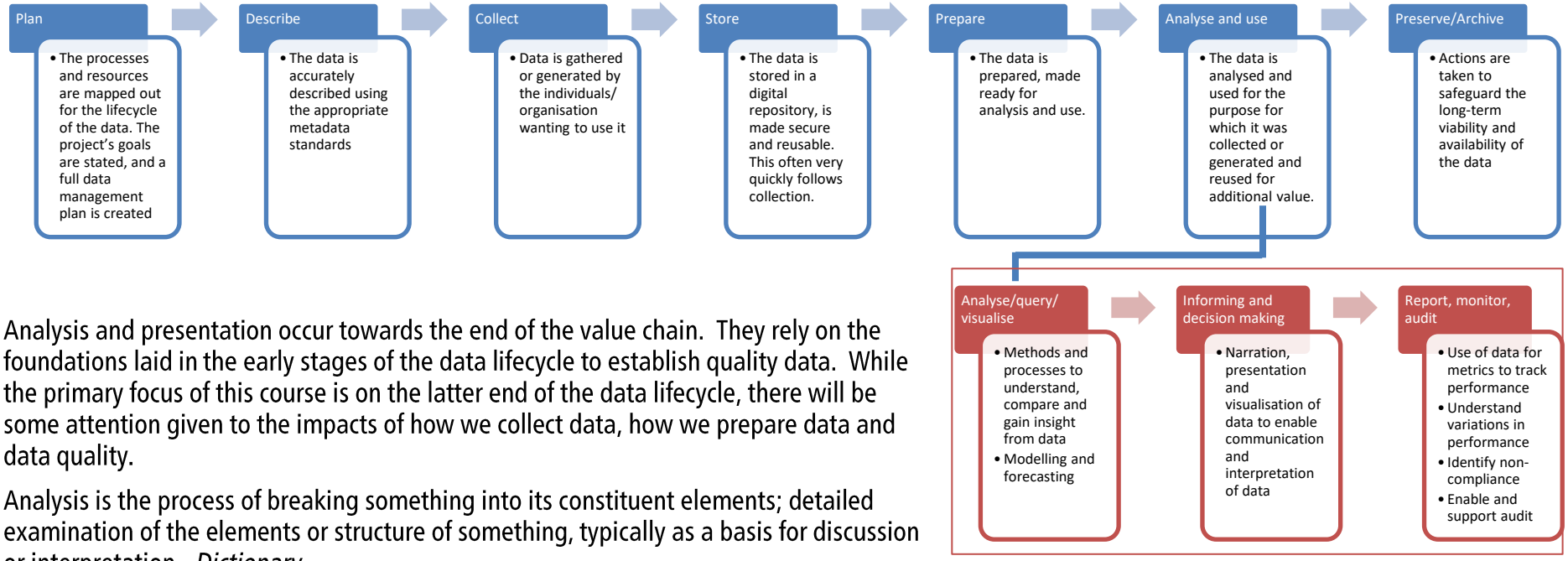
What we know about an aspect of our world	Nothing	Everything
The impact on our decisions	Uncertainty	Certainty

We use data to move up the continuums →

When data is deemed 'fit for purpose', it is fit to assist in decision-making to which the subject is related. It is fit for purpose if it moves us to the right of the continuums. It is not fit for purpose if it risks moving us to the left of the continuums.

There will always be limitations on the ability of data in decision-making. We need to be humble with data.

Data is an organisational asset. The lifecycle over which data is managed is characterised in the diagram below.



Analysis and presentation occur towards the end of the value chain. They rely on the foundations laid in the early stages of the data lifecycle to establish quality data. While the primary focus of this course is on the latter end of the data lifecycle, there will be some attention given to the impacts of how we collect data, how we prepare data and data quality.

Analysis is the process of breaking something into its constituent elements; detailed examination of the elements or structure of something, typically as a basis for discussion or interpretation - *Dictionary*

Analysis is the process of breaking a complex topic or substance into smaller parts to gain a better understanding of it. – *Wikipedia*

Discussion

What are some of the types of data we are working with?

Principles

Two key principles behind data analysis – comparisons and causality.

Comparisons

> *"A fundamental act in statistical reasoning is to answer the question "Compared with what"."*
>
> Source Edward Tufte, Beautiful Evidence

> "*Comparison is the beating heart of data analysis. What we do when we compare data really encompasses both looking for similarities and looking for differences.*" Steven Few

Analysis of data should be designed to enable meaningful and appropriate comparisons and contrasts.

- Comparisons over time enable us to understand stories of growth or decline
- Comparisons with similar examples enable us to compare our performance with our colleagues or competitors
- Comparisons between results and expectations can identify where we excel or our shortcomings
- Comparisons of before and after can enable us to reveal reasons for change or difference.

Causality

> *"Simply collecting data may provoke thoughts about cause and effect: measurements are inherently comparative, and comparisons promptly lead to reasoning about various sources of differences and variability."*
>
> Source Edward Tufte, Beautiful Evidence

A further principle of data design and presentation is that it should show causality and enable explanation as to what causes the data to be as it is, why it varies and why it changes.

In real estate, you hear the term 'location, location, location'.

For data analysis, think 'comparison, comparison, comparison'.

Types of Data

Primary (direct) and Secondary (indirect) Data

Primary data is data created as original by a researcher through direct efforts and experience, specifically for the purpose of addressing their research problem. Primary data is also known as firsthand or raw data.

Secondary data is data that has already been collected and recorded by another person(s) for a different reason or purpose than that of the current researcher. It is the readily available form of data collected from various sources like government publications, internal records of the organisation or publications for related research.

Primary data is typically more expensive and takes longer to collect but has the benefit of being specific to the question or problem being considered. The collection of primary data is necessary to explore the details behind types of behaviour, opinions or views that cannot be identified from existing data. The collection of this data will often involve the crafting of surveys and questionnaires. These methods can, of themselves, create distortions and biases in the data collected. As a result, the collection of primary data is a specialist field.

Many analysts underestimate the volumes and value of secondary data available within their organisation. Common sources of existing data include financial information, activity data, transaction volumes, data on suppliers and customers from subsidiary ledgers and data on staff.

Methods of primary (direct) data collection include:

- Surveys administered by an interviewer
- Surveys that are self-enumerated (the information written or entered directly by the respondent)
- In-depth interviews or focus groups that provide the opportunity for discussion and elaboration for collecting more detailed information about a particular issue or issues
- Observational studies in which data are gathered through the direct observation of the population or sample
- Experiments and clinical trials that involve controlled studies where researchers collect data from subset groups taken from the population of interest.

One form of secondary data is administrative data. Administrative data are collected as part of the day-to-day processes and record-keeping of organisations. Administrative data, such as historical data or public records, include: School enrolments; hospital admissions; and records of births, deaths, and marriages. The data are not collected initially for statistical purposes but can be organised to produce statistics.

Census and Sample data

A population may be studied using one of two approaches: taking a census or selecting a sample. Both provide information that can be used to draw conclusions about the whole population.

A census is a study of every unit, everyone or everything in a population. It is known as a complete count or a complete enumeration.

A sample is a subset of units in a population, selected to represent a population of interest. It is a partial count or partial enumeration. Information from the sampled units is used to estimate the characteristics of the entire population of interest.

Advantages of a census:

- Provides a true measure of the population (no sampling error)
- benchmark data may be obtained for future studies
- detailed information about small sub-groups within the population is more likely to be available.

Disadvantages of a census:

- It may be difficult to enumerate all units of the population within the available time
- higher costs, both in staff and monetary terms, than for a sample
- generally takes longer to collect, process, and release data than from a sample.

Advantages of a sample:

- Costs would generally be lower than for a census
- results may be available in less time
- if good sampling techniques are used, the results can be very representative of the actual population.

Disadvantages of a sample:

- Data may not be representative of the total population, particularly where the sample size is small
- often not suitable for producing benchmark data
- as data are collected from a subset of units and inferences made about the whole population, the data are subject to 'sampling' error
- decreased number of units will reduce the detailed information available about sub-groups within a population.

Source: ABS Statistical Language Census & Sample

Quantitative data vs Qualitative data

Quantitative data assigns numbers (quantities) to observations about a subject or population by counting or measuring aspects of the subject or population.

Qualitative data assigns labels to observations about a subject or population or assigns the subject into categories (quality attributes).

Quantitative data about a person	Qualitative data about a person
Height (Mark is 187 centimetres)	Gender (Mark is a male)
Weight (Mark is 90 kilograms)	Marital status (Mark is married)
Age (Mark is 58 years old)	Where they live (Mark lives in Adelaide)

Qualitative data can be added across a population to determine how many males there are or how many married or unmarried people there are.

Word clouds are a means to present qualitative analysis.

A 'word cloud' is a visual representation of word frequency. The more commonly the term appears within the text being analysed, the larger the word appears in the image generated.

Data about concrete nouns vs data about abstract nouns

Concrete nouns – people, places and things that we can see. Data counts and measures these nouns more exactly – number of people, money, size, weight, etc. This data is more objective. It is about objects we can see. This data is easier to independently validate.

Abstract nouns – ideas, feelings, qualities, states. Data is sought from people about their feelings, ideas, agreements and extent of satisfaction. This data is more subjective. The scores, ratings, evaluations, etc, are based on what the subject thinks or feels. This data is harder to independently validate.

Discussion - Data for business analysis

What are the different types of data and information that can be used to understand the performance of an organisation?

Data for business analysis

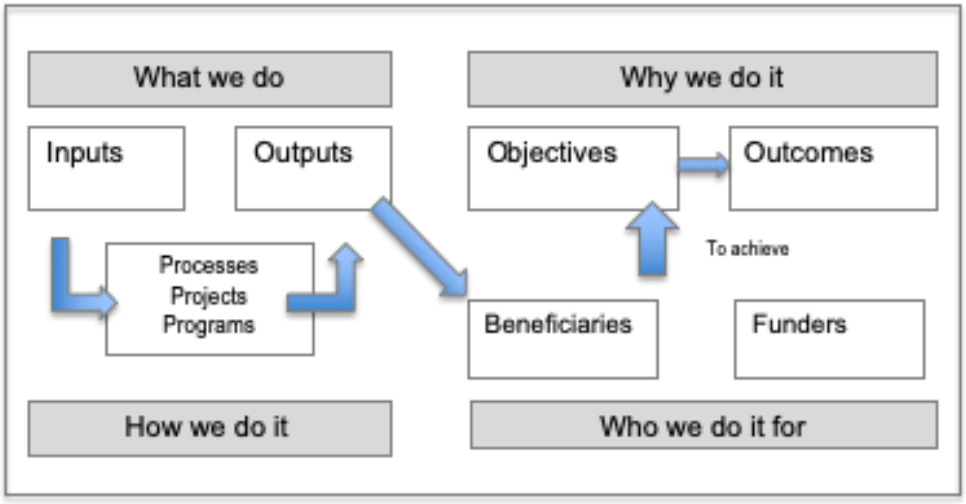
"A business model refers to the logic of the firm, the way it operates and how it creates value for its stakeholders."

From Strategy to Business Models and to Tactics, Ramon Casadesus-Masanell Joan Enric Ricart

"A business model describes the rationale of how an organisation creates, delivers and captures value."

Osterwalder, A., & Pigneur, Y. (2010). *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers.* New Jersey: John Wiley & Sons.

Consider the business model below that presents the logic of an organisation in terms of what it does, how it does it, who it does it for and why it does it.



Components of a Business Model

What the business does:

- Converts inputs (staff, goods and services and technology)
- Into outputs (goods and services)

How the organisation does it:

- Processes
- Projects
- Programs

Who is involved?

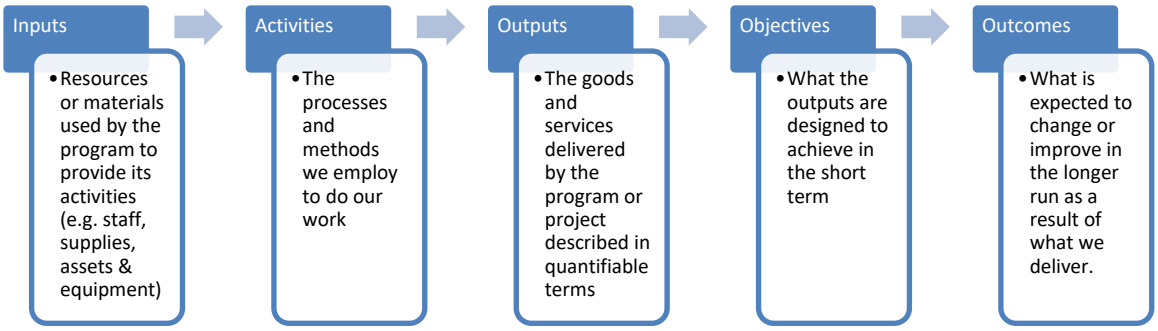
- Suppliers
- Employees
- Beneficiaries
- Funders/Financiers
- Owners

Why?

- Purpose
- Objectives
- Outcomes

The combination of answers to these questions starts to define the business model of an organisation or enterprise.

Embedded within a business model is a basic logic:



This logic can provide a framework for the types of data we might want to analyse to understand a business. It can help ensure we distinguish the following types of data:

- Input data
- Activity data
- Output data
- Objective and
- Outcome data.

In your groups, consider examples of types of data that may belong in each category.

Category	Example	Example
Input data		
Activity data		
Output data		
Objective and		
Outcome data.		

Examples of performance measures using a logic model are presented below.

	Training organisation	Fire service
Input measures	number of staff	number of firefighters
	Facilities (rooms, PCs, desks)	number of fire appliances
	annual total budget	number of fire stations
		annual total budget
Activity measures	Training sessions scheduled	average time to dispatch a fire truck
	Registrations for sessions	percentage of incidents reached by an appliance within X minutes
	Development of materials & resources	
Output measures	Materials produced	number of incidents attended
	Courses delivered	
	Assessments completed	
Objective measures	Learning goals achieved	percentage of incidences where fire is contained
	students qualified	citizen satisfaction
	attendee satisfaction	
Outcome measures	Career progression of attendees	number of deaths per thousand fires
	Manager/business satisfaction	perceptions of public safety

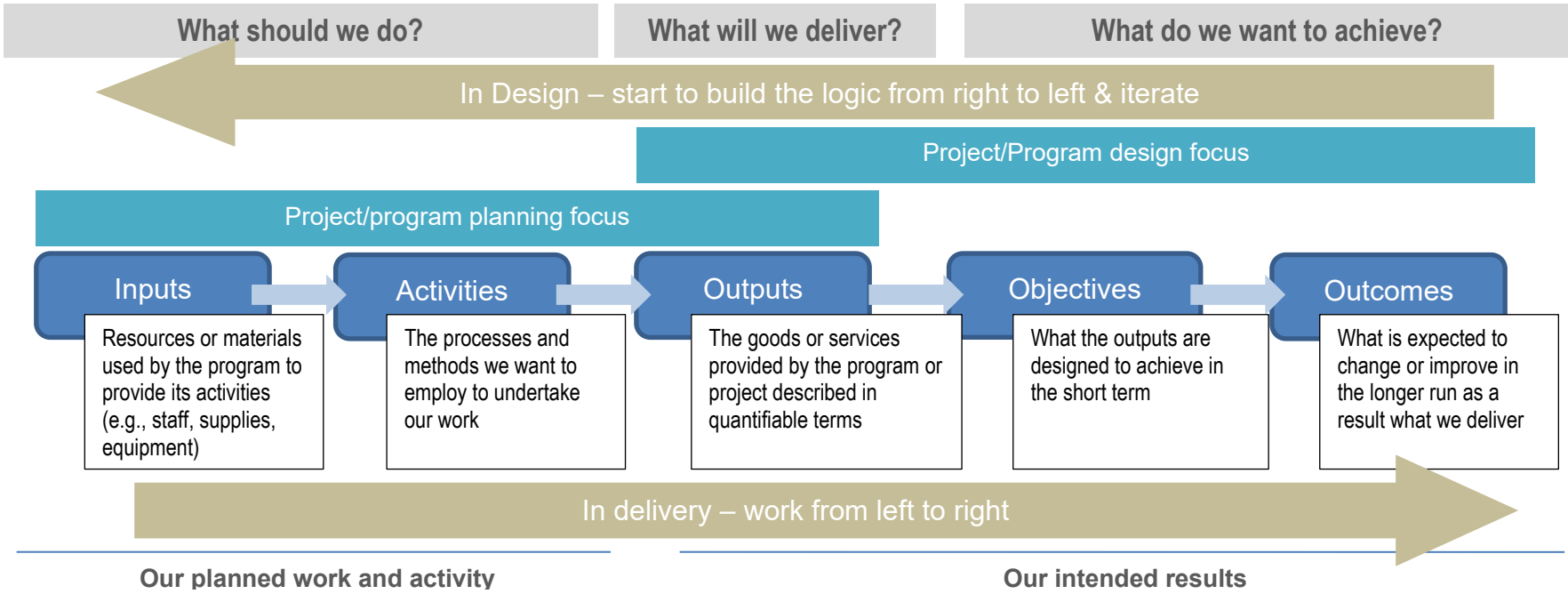
Source (fire service): *Financial Management and Accounting in the Public Sector* - Author Gary Bandy.

It is also possible to take this further, by breaking the data into:

- Quantity
- Quality
- Cost and
- Effectiveness data.

A useful tool for the design and planning of initiatives and the evaluation of initiatives is a logic model.

"A logic model is a systematic and visual way to present and share your understanding of the relationships among the resources you have to operate your program (or project), the activities you plan, and the changes or results you hope to achieve." Kellogg's Foundation.



A logic model can help create a shared understanding of, and focus on, the goals and methodology of an initiative and in relating inputs and activities to projected outcomes.

The design phase of an initiative or program is primarily focussed on the relationship between outputs, objectives and outcomes (the three elements to the right of the model) – will the work deliver what is needed to achieve goals (objectives as more immediate goals and outcomes as longer-term goals).

The planning phase examines in more detail the relationships between inputs, activities and outputs (the three elements to the left of the model).

Logic models are used as tools to support funding bids or grant applications, with applicants required to demonstrate links between inputs and funding required and outcomes sought.

The logic model is based on a series of 'if-then' statements. If we have these inputs, we can undertake these activities and produce these outputs. If we deliver these outputs, we will achieve certain objectives. If we achieve certain objectives, then we will achieve certain outcomes.

When designing a strategy or initiative, data is often sought and analysed to establish these 'if-then' relationships. For example, when designing a road safety strategy, data will be collected and analysed to understand the root causes of accidents, casualties and fatalities.

Preparing for data analysis

The preparation required for good data analysis can be compared to the preparation required to do a good job of painting a house.

To paint well is not just about slapping paint onto a brush or roller and covering a wall.

Good painting at home requires preparation – filling holes, sanding off peeled and rough old paint, cleaning the wall of stains and removing dust and grime. We have to prepare the wall so it is ready to be painted.

So too with data analysis. Before we apply different analysis techniques, we need to spend some time preparing the data so it is ready to be analysed.

Preparing data can include transforming and flattening the data, cleansing data and reconciling data.

Creating a flat data set or database

Just as we need a good base for painting, we need a good base for data – we need a flat data set or database.

We want all our data properly arranged into rows and columns such that:

- Every column is a field. A field is a single piece of information from a record.
- Each row is a record where each record has a valid entry against each field.

The main benefit of creating a database is that it enables a range of analysis techniques to be performed automatically through tools like data filters, subtotalling and pivot tables.

I have created a database from ABS population data. The data can be analysed in the form presented by the ABS but converting it into a database opens up a range of opportunities for data analysis.

For net migration data, the screen grab below details how the data is provided by the ABS.

Table 4.5 Overseas migrant arrivals and departures, by visa and citizenship groups, South Australia, 2004-05 to 2022-23(a)(b)

Direction	Visa and citizenship groups(c)		2004-05	2005-06	2006-07	2007-08	2008-09
Overseas migrant arrivals(d)	Permanent visas	Family	1,230	1,360	1,590	1,560	1,660
		Skilled (permanent)	2,550	4,530	5,220	5,320	4,710
		Special eligibility & humanitarian	1,220	1,100	1,220	840	1,050
		Other (permanent)	140	130	170	150	170
	Total permanent visas		**5,130**	**7,120**	**8,190**	**7,870**	**7,590**
	Temporary visas	Student - vocational education and training	240	240	360	860	2,650
		Student - higher education	2,380	3,050	3,860	4,570	5,120
		Student - other	1,540	1,390	1,870	1,740	1,620
		Skilled (temporary)	740	1,520	1,700	1,910	2,070
		Working holiday	250	320	470	570	690
		Visitors	1,200	1,300	1,280	1,710	1,480
		Other (temporary)	590	570	530	510	340
	Total temporary visas		**6,930**	**8,380**	**10,060**	**11,860**	**13,960**

Flattening data requires allocated categories in the first three columns across all rows:

This can be done manually by dragging the categories across all rows.

We will also need to change migrant departures to have negative signs rather than positive signs. This can be done by formula rather than by manually changing data.

For our datasets we can also combine jurisdictions by adding another field to each state showing the jurisdiction. We can then 'stack' the datasets together to arrive at a consolidated dataset for net migration by groups across all states and with arrivals with positive data and departures as negative data.

After these changes have occurred, the database appears as follows:

Jurisdiction	Direction	Visa and citizenship groups(d)	Category	Category 1	Category 2	2004-05	2005-06	2006-07	2007-08	2008-09
Australia	**Overseas migrant arrivals(e)**	Permanent visas	Family	Permanent	Family	28,430	30,400	32,480	33,100	35,060
Australia	**Overseas migrant arrivals(e)**	Permanent visas	Skilled (permanent)	Permanent	Skilled	35,540	42,750	47,540	51,600	48,360
Australia	**Overseas migrant arrivals(e)**	Permanent visas	Special eligibility & humanitarian	Permanent	Special eligibility	13,580	12,320	12,400	9,470	11,630
Australia	**Overseas migrant arrivals(e)**	Permanent visas	Other (permanent)	Permanent	Other	3,650	3,640	3,770	3,940	4,000
Australia	**Overseas migrant arrivals(e)**	**Total permanent visas**	**sub total**	**Sub total**	**Sub total**	**81,210**	**89,110**	**96,190**	**98,110**	**99,040**
Australia	**Overseas migrant arrivals(e)**	Temporary visas	Student - vocational education and training	Temporary	Student	7,990	10,500	19,970	31,420	53,570

After flattening the data, it is critical to reconcile the flattened data with the original dataset to ensure the integrity of the flattened dataset. Pivot tables can be used to calculate totals from the flattened dataset.

The benefit of this setup is that:

- Other data can be combined for earlier years or other jurisdictions
- that columns with formulas can be added for analysis purposes
- the database can be filtered and queried to provide summarised data.

The last point is most helpful as it enables the data to be queried and viewed in different ways, in particular using pivot tables so we can see the data from different perspectives.

Data analysis technologies

Improved technologies can make data analysis and data presentation more straightforward.

We are familiar with traditional spreadsheet tools like Excel that enable straightforward data analysis with small and medium-sized data sets.

More advanced tools are now available in the form of Microsoft Power BI and subscription packages like Tableau.

To use the analogy of painting, excel is like a good brush while these other technologies are like rollers and spray guns – you can cover more data more quickly with them.

However, beware, just as painting technologies like rollers and spray guns cannot overcome deficiencies in the surface for cover for lack of preparation, the same is true for data analysis technologies not overcoming deficiencies in the quality of data or the lack of preparation of the data.

Data Cleansing

"Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data." Source: Wu, S. (2013), "A review on coarse warranty data and analysis", *Reliability Engineering and System* via Wikipedia

'Dirtiness' includes:

- Duplicate records – the same item inadvertently recorded twice in a data set
- Spelling errors – items where categories have been misspelt or improperly coded (e.g. a 'Camry' recorded as a 'Carmy' in a vehicle register)
- Blank cells – records that have not had particular elements recorded
- Numbers as text – where numbers present in text format and need to be converted
- False unique identifiers – dual records with the same unique identifier
- Outliers (false or error records) – unusually large or small values attached to records
- Changes in categories and definitions over time – may require that we rely more on aggregated data or find ways to combine categories.

Basic data cleansing involves:

- Sorting data to see duplicates and multiple uses of unique identifiers. Once found, judgements are made about the integrity of the data and, if the data is OK, duplicates are removed and unique identifiers corrected.
- Conditional formatting can help identify unusual records or numbers
- Pivot tables to identify unusual categories and blanks
- Recategorising data using mapping and Vlookup can be used to further assess the quality and categorization of records.
- Reconciling the parts to totals – checking the parts are consistent with the whole and that they add to the whole to find double counting or gaps.

There are more sophisticated data cleansing methods for large data sets that are beyond the scope of this course and its presenter.

Conclusion

Despite these deficiencies and difficulties with data, I seek to work with the best data available and conclude that working with the best data available is usually better than working with no data at all.

"An approximate answer to the right problem is worth a good deal more than an exact answer to the wrong problem." John Tukey

Activity – Scenario

The Minister has seen and heard reports of significant growth in international migration and has asked us to do some analysis of data and present our findings to him/her. They are hearing different stories. They are hearing popular press about increased migration and have also seen headlines from the ABS:

- Net overseas migration was 306,000 in 2024-25, down from 429,000 a year earlier.
- Migrant arrivals decreased 14% to 568,000 from 661,000 arrivals a year earlier.
- The largest group of migrant arrivals was temporary students with 157,000 people.

What types of data analysis could we consider in developing our advice for the Minister?

Introducing types of information & analysis

We will examine the following types of analysis. They are:

1. Summary statistics
2. Ranking analysis - ordering items or individuals based on a specific criterion
3. Compositional analysis – understanding the components of our dataset.
4. Temporal analysis – understanding changes over time (trend and growth analysis)
5. Variance analysis – understanding variances from forecasts or budgets
6. Ratio analysis – (primarily used in financial analysis) analysing ratios to assess financial viability and financial health
7. Benchmarking – comparing with external figures
8. Multivariate analysis - combining different data sets (e.g. financial and non-financial data).

These methods will be applied to a range of data – survey data, migration data, and COVID data.

Summary statistics

Numbers that summarise help give us focus and help us see the parts and the whole. Data that can be developed to provide a summary include:

- Totals
- Average - an average, or central tendency of a dataset, is a measure of the "middle" value of the dataset.
- Median (middle) - a median is described as the numerical value separating the higher half of a sample, a population, or a probability distribution from the lower half.
- Mode (most common) - the mode is the value that occurs most frequently in a data set or a probability distribution.

This type of analysis is useful for surveys where we have a range of questions or topics (shown below in columns) with multiple respondents (each has their own row).

Surveys capture information by person to a range of questions.

Results are often sought using Likert scales, where respondents mark their responses on scales with a range of different ratings. Some examples include:

The training was relevant to my work

Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree

This is a five-point Likert Scale. Responses from participants can be assigned numbers for the purposes of doing numerical analysis.

Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
1	2	3	4	5

An example of survey responses for a training workshop is provided below. The top row represents the question numbers with the scores for each consistent with the Likert scale used to collect the information.

Participant	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
Participant 1	4	4	4	3	4	4	4	5	5	4	4	3	3
Participant 2	4	4	4	4	4	4	4	5	5	4	4	5	5
Participant 3	5	5	5	5	5	5	5	5	5	5	5	5	5
Participant 4	5	5	5	5	5	5	5	5	5	5	5	5	5
Participant 5	5	5	5	5	5	5	5	5	5	5	5	5	5
Participant 6	5	5	5	5	5	5	5	4	4	5	5	5	5
Participant 7	4	4	4	4	4	4	4	4	4	4	4	3	3
Participant 8	5	5	5	5	5	5	4	5	5	4	4	5	3
Participant 9	4	4	4	4	4	4	4	4	4	4	4	3	3
Participant 10	4	4	4	4	4	4	4	3	3	4	4	3	3
Participant 11	4	4	5	4	5	4	4	3	3	4	4	3	3

Summary statistics for our sample of respondents are shown below:

	Objectives explained	Objectives linked to workplace	Facilitator pace	Facilitator engagement	Faciliator knowledge	Facilitator feedback	Workshop duration	Workshop content relevance	Workshop examples	Opportuni-ties to reflect	Engagement strategies	Technical support	Overall learning experience
Average	4.5	4.5	4.5	4.4	4.5	4.5	4.4	4.4	4.4	4.4	4.4	4.1	3.9
Median	4	4	5	4	5	4	4	5	5	4	4	5	3
Mode	4	4	5	4	5	4	4	5	5	4	4	5	3
Counts													
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	2	2	0	0	5	6
4	6	6	5	5	5	6	7	3	3	7	7	0	0
5	5	5	6	5	6	5	4	6	6	4	4	6	5
	11	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**

The average score can be compared with a maximum possible of 5 and may also be compared with a target or benchmark.

Ranking analysis

It is normal for our data to initially appear in order based on the coding system used to collect it. Financial data is often initially recorded according to account code, industry data by industry code and ABS data on net migration reports data by country code (SACC code) as per the below.

SACC code(c)	Country of birth(c)	2004-05
1101	Australia	-7,430
1102	Norfolk Island(e)	0
1199	Aust E T, nec	0
1201	New Zealand	4,080
1301	New Caledonia	10
1302	PNG	70
1303	Solomon Islands	30
1304	Vanuatu	10
1401	Guam	0
1402	Kiribati	-10

Ranking analysis involves ordering data points (like items, people, or ideas) from least to greatest (or vice versa) and assigning each a numerical rank, allowing for comparison and analysis of relative positions or preferences.

Net migration by country ranked – Australia & South Australia (2024-25)

Australia net migration

Country	Net migration 2024-25
India	57,410
China	35,160
UK, CIs & IOM	25,630
New Zealand	24,320
Philippines	19,440
Nepal	16,310
Sri Lanka	12,440
Bangladesh	11,280
Pakistan	9,380
Vietnam	9,250
Afghanistan	8,490
Indonesia	8,480
South Africa	6,600
Iran	4,180
Hong Kong	3,480
Ireland	3,460
France	3,320
USA	3,230
Other Countries	43,330
Total	**305,190**

South Australia net migration

Country	Net migration 2024-25
India	4,850
China	1,770
Afghanistan	1,010
Philippines	990
Sri Lanka	960
Nepal	930
UK, CIs & IOM	890
Bangladesh	810
Pakistan	790
Indonesia	750
Vietnam	600
New Zealand	520
Iran	390
South Africa	380
Malaysia	230
Kenya	230
USA	200
Hong Kong	190
Other countries	2,230
Total	**18,720**

Source: ABS overseas net migration 2024-25

Compositional Analysis

This type of analysis is required to understand the makeup or structure of a dataset or the finances of a business or organisation. Compositional analysis is used when you want to understand or get to know a dataset or business. This analysis is a starting point for the other forms of analysis.

The table below summarises client data to show the relative size of different components of the dataset.

Overseas net migration by country compositions - Australia and South Australia (2024-25)

Australia net migration

Country	Net migration 2024-25	Share of total
India	57,410	18.8%
China	35,160	11.5%
UK, CIs & IOM	25,630	8.4%
New Zealand	24,320	8.0%
Philippines	19,440	6.4%
Nepal	16,310	5.3%
Sri Lanka	12,440	4.1%
Bangladesh	11,280	3.7%
Pakistan	9,380	3.1%
Vietnam	9,250	3.0%
Afghanistan	8,490	2.8%
Indonesia	8,480	2.8%
South Africa	6,600	2.2%
Iran	4,180	1.4%
Hong Kong	3,480	1.1%
Ireland	3,460	1.1%
France	3,320	1.1%
USA	3,230	1.1%
Other Countries	43,330	14.2%
Total	**305,190**	**100.0%**

South Australia net migration

Country	Net migration 2024-25	Share of total
India	4,850	25.9%
China	1,770	9.5%
Afghanistan	1,010	5.4%
Philippines	990	5.3%
Sri Lanka	960	5.1%
Nepal	930	5.0%
UK, CIs & IOM	890	4.8%
Bangladesh	810	4.3%
Pakistan	790	4.2%
Indonesia	750	4.0%
Vietnam	600	3.2%
New Zealand	520	2.8%
Iran	390	2.1%
South Africa	380	2.0%
Malaysia	230	1.2%
Kenya	230	1.2%
USA	200	1.1%
Hong Kong	190	1.0%
Other countries	2,230	11.9%
Total	**18,720**	**100.0%**

Source: ABS overseas net migration 2022-23.

For Australia, around 86% of migration numbers are driven by 18 countries.

For South Australia, 88% of migration numbers are driven by 18 countries.

Temporal analysis

To understand current datasets or financial information, insight can come from understanding historical and future trends. This gives us a sense of how the current data compares with the past and how they have and is projected to change. Temporal analysis involves analysing changes over time (this includes trend and growth analysis).

Overseas net migrations - Australia (2024-25 vs 2014-15)

Australia net migration

Country	2014-15	Share of total	Net migration 2024-25	Share of total	Total growth	Annual growth	change in share
India	38,400	20.8%	57,410	18.8%	49.5%	4.1%	-2.0%
China	39,100	21.2%	35,160	11.5%	-10.1%	-1.1%	-9.7%
UK, CIs & IOM	9,410	5.1%	25,630	8.4%	172.4%	10.5%	3.3%
New Zealand	940	0.5%	24,320	8.0%	2487.2%	38.4%	7.5%
Philippines	12,430	6.7%	19,440	6.4%	56.4%	4.6%	-0.4%
Nepal	6,410	3.5%	16,310	5.3%	154.4%	9.8%	1.9%
Sri Lanka	4,120	2.2%	12,440	4.1%	201.9%	11.7%	1.8%
Bangladesh	2,900	1.6%	11,280	3.7%	289.0%	14.5%	2.1%
Pakistan	7,230	3.9%	9,380	3.1%	29.7%	2.6%	-0.8%
Vietnam	7,320	4.0%	9,250	3.0%	26.4%	2.4%	-0.9%
Afghanistan	2,910	1.6%	8,490	2.8%	191.8%	11.3%	1.2%
Indonesia	2,510	1.4%	8,480	2.8%	237.8%	12.9%	1.4%
South Africa	3,470	1.9%	6,600	2.2%	90.2%	6.6%	0.3%
Iran	2,620	1.4%	4,180	1.4%	59.5%	4.8%	-0.1%
Hong Kong	1,830	1.0%	3,480	1.1%	90.2%	6.6%	0.1%
Ireland	-3,250	-1.8%	3,460	1.1%	n/a	n/a	2.9%
France	2,090	1.1%	3,320	1.1%	58.9%	4.7%	0.0%
USA	2,460	1.3%	3,230	1.1%	31.3%	2.8%	-0.3%
Other Countries	41,390	22.5%	43,330	14.2%	4.7%	0.5%	-8.3%
Total	**184,290**	**100.0%**	**305,190**	**100.0%**	**65.6%**	**5.8%**	

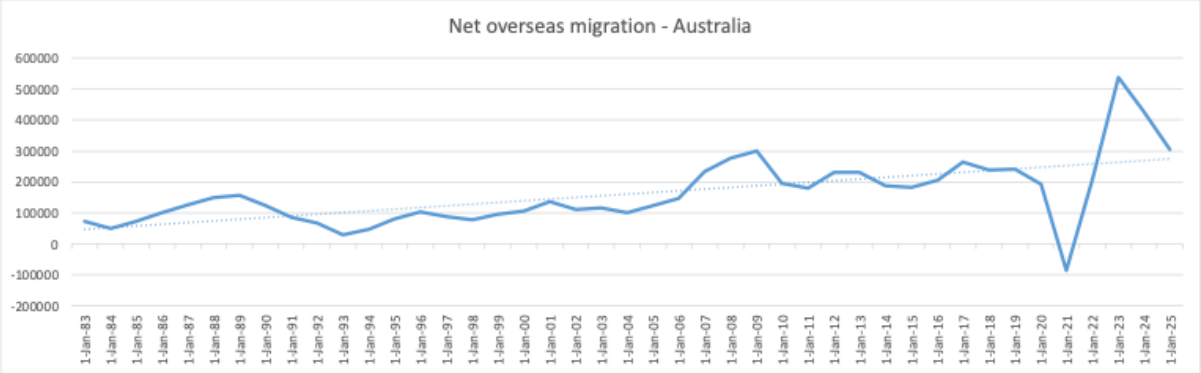
Monthly growth is a compound monthly growth rate

Some possible traps associated with temporal analysis:

- Selecting beginning and end points to find the 'right' story. This can occur by selecting a low beginning point and high-end point to exaggerate growth or vice versa (a high beginning point and low-end point to exaggerate declines).
- Growth rates are sensitive to the relative size of the initial point. When the starting point is low, growth rates can be high.
- Quoting growth rates on absolute numbers or nominal numbers when adjustments should be made for underlying changes like population or inflation.

Temporal analysis is aided by trend data presented in a graph over the page.

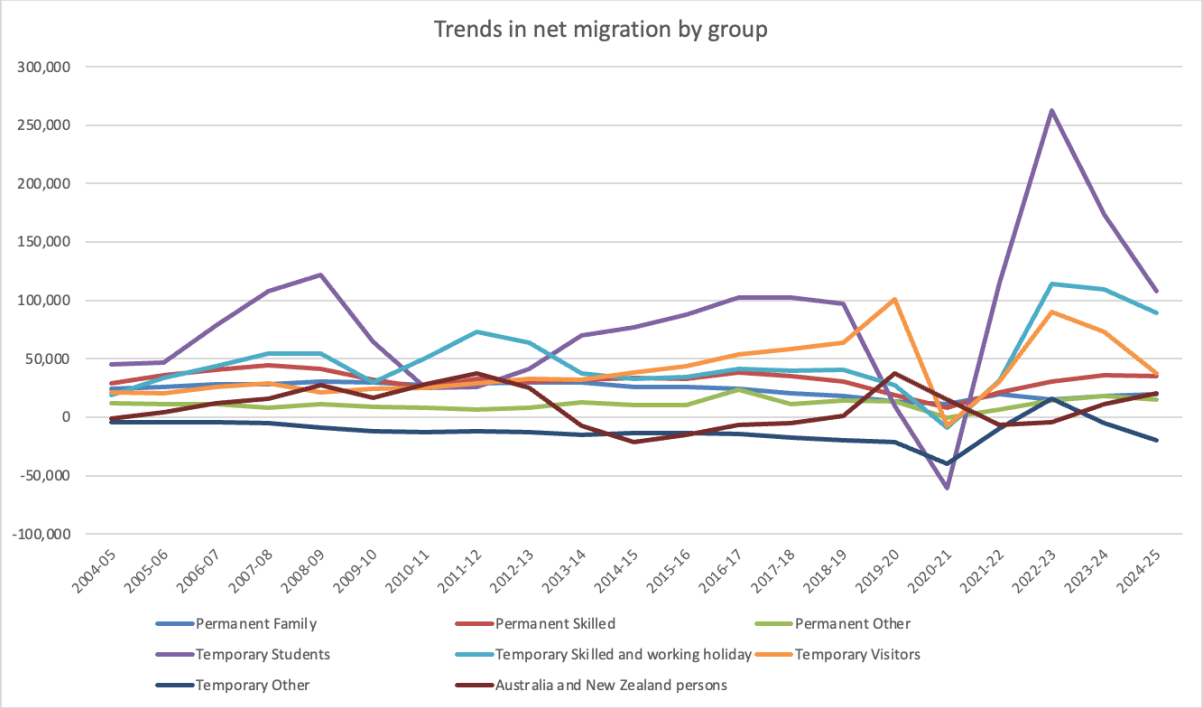
Net overseas migration for Australia – 1983 to 2025



The long-term trend shows an increase in net migration in the early 2000s.

Examples of trends are also provided for Australian migration data by groups.

Net migration trends for Australia by group (2004-05 to 2024-25)

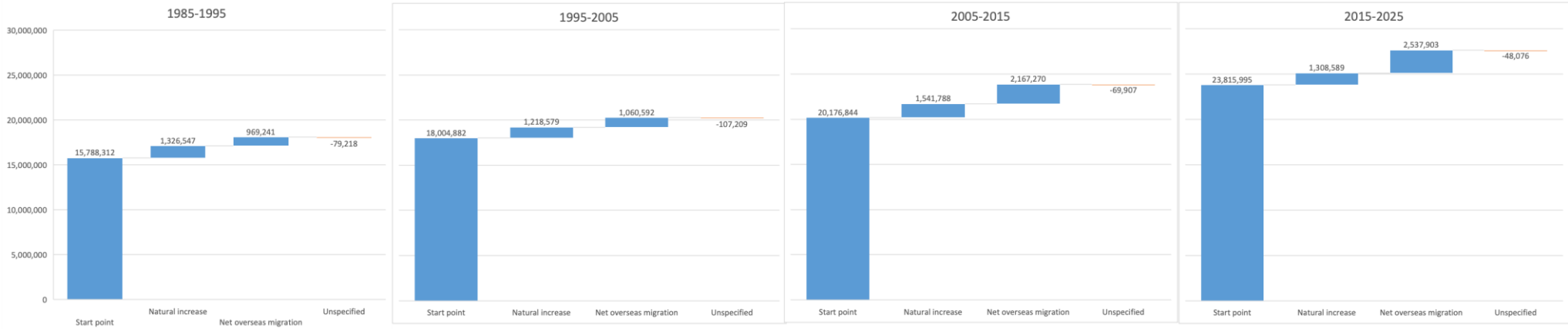


Source: ABS Migration by groups

Growth has been in temporary visa net migration, most notably students and temporary skilled and working holiday visas.

We can look at the relative scale of these figures by looking at how they compare with changes in Australia's overall population.

Changes in population – Australia



Source: ABS National, state and territory population

The changes in population show net overseas migration contributing just under 100,000 per annum to the population between the years of 1985 and 2005. Since 2005, net migration has added over 200,000 people per annum to the population.

Variance analysis

To date, we have been looking at the composition of our data and comparing changes over time. Now we need to make comparisons with other measures. For net migration data, we can see if there are variances from projections made by the ABS in 2004 and variances by state for different countries. For financial data, the most common form of variance analysis is to examine whether or not actual results vary from the budget.

Net overseas migration – actual for 2023 compared to 2023 projections made in 2004

Age	Projected in 2004	Actual in 2023	Variance	Variance %
0-4	1,328,086	1,513,072	184,986	13.9%
5-9	1,350,176	1,612,127	261,951	19.4%
10-14	1,381,488	1,638,761	257,273	18.6%
15-19	1,414,576	1,533,919	119,343	8.4%
20-24	1,458,427	1,637,698	179,271	12.3%
25-29	1,559,135	1,827,118	267,983	17.2%
30-34	1,634,479	1,915,630	281,151	17.2%
35-39	1,634,873	1,894,505	259,632	15.9%
40-44	1,615,879	1,704,808	88,929	5.5%
45-49	1,550,183	1,622,490	72,307	4.7%
50-54	1,618,336	1,653,689	35,353	2.2%
55-59	1,492,677	1,534,811	42,134	2.8%
60-64	1,507,649	1,492,429	-15,220	-1.0%
65-69	1,332,951	1,302,941	-30,010	-2.3%
70-74	1,192,154	1,144,905	-47,249	-4.0%
75-79	968,246	872,845	-95,401	-9.9%
80-84	619,402	565,473	-53,929	-8.7%
85 and over	622,604	547,178	-75,426	-12.1%
	24,281,321	**26,014,399**	**1,733,078**	**7.1%**

In this context, variance analysis is different from analysis of variance.

The first (variance analysis) is used in areas like finance where variances are used to show a difference between a budget (or forecast) and an actual. The variances are presented in dollar terms and of often presented in percentage terms. Different organisations will have views about the materiality of these variances. A materiality threshold could be expressed in dollar terms (e.g. all variances over $10,000 are deemed material and need to be explained). The materiality threshold could also be expressed in percentage terms (e.g. all variances greater than 2% or less than -2% need to be explained along with remedies).

Variances have a different meaning in statistics.

In statistics, variance measures variability from the average or mean.

Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the average.

Analysis of variance is used when testing a hypothesis. It is typically used when comparing two populations (e.g. a test group using a new drug and a control group using a placebo). For each group, calculations will include a sample size, an average result and a variation.

Ratio analysis

Ratios can be used to provide indications of numbers per….

Ratio analysis is used in financial analysis to assess liquidity (ratio of liquid assets to current liabilities, profitability (return on assets or return on equity) and levels of debt (debt/equity ratio).

Another example of ratio analysis was relevant during COVID. In addition to tracking data on case numbers and deaths, we were also interested in the fatality ratio. It was the ratio of the number of deaths per confirmed case.

Top 10 countries – COVID-19 cases from 1 January 2024 to 24 October 2024

Country	Confirmed cases	Deaths	Cases per million popn	Deaths per million popn	Fatality rate
United States of America	8,320,491	221,564	24,919	469.5	2.7%
India	7,814,682	117,956	5,624	0.7	1.5%
Brazil	5,323,630	155,900	24,928	129.5	2.9%
Russian Federation	1,497,167	25,821	10,146	303.9	1.7%
Argentina	1,053,650	27,957	22,952	651.9	2.7%
Spain	1,046,132	34,752	21,950	790.7	3.3%
France	1,010,554	34,225	14,841	152.5	3.4%
Colombia	990,270	29,636	19,293	667.0	3.0%
Peru	879,876	33,984	26,595	2,651.2	3.9%
Mexico	874,171	87,894	6,729	107.0	10.1%

Source: World Health Organisation

Ratio analysis is also used to make data more relatable. Other examples include converting annual data into daily or hourly data. For instance, in Australia in 2022, there were 190,939 deaths. This equates to an average of 523.1 per day, 21.8 per hour or one death every 2.8 minutes.

Benchmarking analysis

To date, we have been analysing a data set for one unit (Australia) on its own. An important aspect of data analysis is to compare our unit with external parties using benchmark analysis.

We can compare other aspects of the data with other jurisdictions or between jurisdictions in Australia.

A common requirement of benchmarking is that you must compare like with like or apples with apples. This is often taken too far. No two entities are alike. We necessarily need to compare different entities to identify the nature and possible causes of differences.

For our dataset, we can compare overseas migration across states like South Australia, NSW and Victoria.

Overseas migration by top 20 countries – Victoria, South Australia & NSW (2024-25)

Country	Australia	Victoria	South Australia	New South Wales
India	57,410	20,570	4,850	15,610
China	35,160	10,980	1,770	15,080
UK, CIs & IOM	25,620	5,080	890	7,680
New Zealand	24,320	5,880	520	3,870
Philippines	19,440	4,640	990	5,450
Nepal	16,310	2,870	930	8,770
Sri Lanka	12,430	5,550	960	2,710
Bangladesh	11,280	2,500	810	5,330
Vietnam	9,250	3,400	600	3,160
Pakistan	9,390	3,040	790	2,740
Afghanistan	8,490	3,820	1,010	2,100
Indonesia	8,490	2,870	750	2,770
South Africa	6,590	1,090	380	1,240
Iran	4,190	1,160	390	1,510
Hong Kong	3,480	1,250	190	1,140
USA	3,230	840	200	1,000
France	3,340	520	130	1,090
Ireland	3,480	570	70	900
Other Countries	43,610	11,120	2,490	9,420
Total	**305,490**	**87,750**	**18,720**	**91,550**

Source: ABS overseas net migration 2024-25.

What is the main problem with this comparison?

Multivariate analysis

To date, all the analysis we have done has been of specific data in isolation from other data. Data does not exist in isolation from other variables. Therefore, understanding our dataset better will involve combining our data with other variables.

In some cases, we can combine data relating to finances with output data or caseload data to determine costs per unit of output or cost per case. This is often important in doing benchmarking as we need to take into account the different sizes of states or organisations to make data more comparable.

In the case of comparing states, we can compare data figures with the population of each state in order to determine indicators per capita.

The tables on the following pages present our data on a per capita basis.

Net overseas migration per 1,000,000 of the population – Victoria, South Australia & NSW (2024-25)

Country	Australia	Victoria	South Australia	New South Wales
India	2,079	2,908	2,550	1,816
China	1,273	1,552	930	1,755
UK, CIs & IOM	928	718	468	891
New Zealand	881	831	273	450
Philippines	704	656	520	634
Nepal	591	406	489	1,020
Sri Lanka	450	785	505	315
Bangladesh	408	353	426	620
Vietnam	335	481	315	368
Pakistan	340	430	415	319
Afghanistan	307	540	531	244
Indonesia	307	406	394	322
South Africa	239	154	200	144
Iran	152	164	205	176
Hong Kong	125	177	100	133
USA	117	119	105	116
France	121	74	68	127
Ireland	126	81	37	105
Other Countries	1,579	1,572	1,309	1,096
Total	**11,063**	**12,404**	**9,841**	**10,653**

Source: ABS overseas net migration 2024-25.

States net migration by group for 2023-24 (per 1 million of population)

	Victoria	South Australia	NSW
Permanent visas	**3,037**	**3,580**	**2,617**
Family	983	761	788
Skilled (permanent)	1,213	2,157	1,294
Special eligibility & humanitarian	804	672	512
Other (permanent)	37	-11	22
Temporary Visas	**8,624**	**6,292**	**8,056**
Student - higher education	4,704	3,188	3,943
Student - vocational education and training	-157	16	-219
Student - other	552	481	487
Skilled (temporary)	1,053	835	1,188
Visitors	1,624	1,005	1,556
Working holiday	1,813	851	2,038
Other (temporary)	-966	-85	-936
Australia and New Zealand persons	**866**	**32**	**52**
Total	**12,527**	**9,904**	**10,725**

Discussion

To date, the data considered has looked at clients and services presuming each is the same (i.e. each is equal to one). For the purposes of developing policy, we may need to be more sophisticated than this. What can we do to improve the quality of our analysis?

Standardising data

When using financial data, a dollar is a dollar. Finances compare across organisations and functions because of the standard method of valuation - the dollar. Other forms of data cannot always rely on such standardisation or consistency in value.

Relevant examples of where data is not easily comparable include:

- Visitors to hospitals – counting the raw number of visitors does not consider the relative complexity of each visitor
- Number of court trials or cases– counting the number of trials or cases masks the relative complexity of each trial.
- Number of transactions in a customer service centre – this data will mask both the value and the complexity of transactions.

As a result, methods of standardising data are needed to provide data sets and summary analysis that are more sophisticated than raw counts. Using the examples above, standardising data will require the application of weights that represent the relative complexity of each transaction. In health activity measures in hospitals are weighted to enable an analysis of standard measures of activity. Sophisticated weightings have been developed for different groups of patient treatments and episodes of care. In education, weights have been developed for different student groups to determine the funding needs of schools.

For customer service transactions, the raw count of transactions can be weighted by the time for each transaction. For example:

- Renewing a motor vehicle registration takes 3 minutes
- Renewing a licence (that includes a photo) takes 6 minutes.
- The registration renewal will be deemed a standard transaction.
- Each licence renewal will therefore represent 2 standard transactions.

A simple table showing transaction volumes compares the raw transaction count with the standardised or adjusted transaction count.

Number of transactions	Raw	Standardised
Registration renewals	1,000	1,000
Licence renewals	500	1,000
Total Transactions	**1,500**	**2,000**

Based on raw data, there were 1,500 transactions. However, after considering the relative complexity of the licence renewals, the standardised data shows 2,000 transactions.

A **full-time equivalent (FTE)** is a measure of standardised data. We may have four employees working in a team, but if two of those employees are working part-time and two are full-time, they should not each be counted as one employee for budgeting purposes. If the part-time staff are working three-day weeks, they are working 60% of the hours of a full-time employee and therefore would be counted as 0.6 FTE for budgeting purposes. Therefore our team of four employees is counted as 3.2 FTE for budgeting purposes.

An example of the impact of weighting data occurs with causes of death data. For data on causes of death, the Australian Bureau of Statistics presents complementary data that aims to compare the impact of deaths based on potential years of life lost. For deaths that occur at an earlier age, more potential years of life are lost. That is, all deaths are not equal. The next tables present the original raw data along with the data standardised for years of potential lives lost.

Causes of death by chapter compositions - Australia (2022)

	Persons 2022	Males 2022	Females 2022	Persons %	Males %	Females %
Neoplasms	51,920	29,165	22,754	27.2%	29.2%	25.0%
Diseases of the circulatory system	45,005	23,112	21,893	23.6%	23.1%	24.1%
Diseases of the respiratory system	15,203	8,008	7,195	8.0%	8.0%	7.9%
External causes of morbidity and mortality	12,542	7,940	4,597	6.6%	7.9%	5.1%
Mental and behavioural disorders	12,228	4,670	7,558	6.4%	4.7%	8.3%
Diseases of the nervous system	11,908	5,614	6,294	6.2%	5.6%	6.9%
Codes for special purposes	9,860	5,484	4,376	5.2%	5.5%	4.8%
Endocrine, nutritional and metabolic diseases	8,579	4,461	4,118	4.5%	4.5%	4.5%
Diseases of the digestive system	7,672	3,853	3,819	4.0%	3.9%	4.2%
Diseases of the genitourinary system	4,726	2,199	2,527	2.5%	2.2%	2.8%
Symptoms, signs and abnormal clinical and laboratory findings, nec	4,023	2,017	2,006	2.1%	2.0%	2.2%
Certain infectious and parasitic diseases	2,918	1,446	1,471	1.5%	1.4%	1.6%
Diseases of the musculoskeletal system and connective tissue	1,756	694	1,062	0.9%	0.7%	1.2%
Diseases of the skin and subcutaneous tissue	795	338	457	0.4%	0.3%	0.5%
Congenital malformations, deformations and chromosomal abnormalities	688	374	314	0.4%	0.4%	0.3%
Diseases of the blood & blood-forming organs & certain immune mechanism disorders	561	272	289	0.3%	0.3%	0.3%
Certain conditions originating in the perinatal period	516	259	257	0.3%	0.3%	0.3%
Diseases of the ear and mastoid process	20	9	11	0.0%	0.0%	0.0%
Diseases of the eye and adnexa	13	4	9	0.0%	0.0%	0.0%
Pregnancy, childbirth and the puerperium	6	0	6	0.0%	0.0%	0.0%
Total	**190,939**	**99,919**	**91,013**	**100.0%**	**100.0%**	**100.0%**

Causes of death by chapter: years of potential life lost - Australia (2022)

	Persons 2022	Males 2022	Females 2022	Persons %	Males %	Females %
Neoplasms	313,544	171,223	143,187	30.7%	26.4%	38.3%
External causes of morbidity and mortality	246,709	181,871	64,722	24.2%	28.0%	17.3%
Diseases of the circulatory system	152,751	108,721	45,270	15.0%	16.8%	12.1%
Diseases of the digestive system	52,438	33,114	19,545	5.1%	5.1%	5.2%
Diseases of the respiratory system	46,759	26,901	20,024	4.6%	4.1%	5.4%
Endocrine, nutritional and metabolic diseases	43,791	26,919	17,037	4.3%	4.1%	4.6%
Diseases of the nervous system	41,106	24,262	16,925	4.0%	3.7%	4.5%
Symptoms, signs and abnormal clinical and laboratory findings, nec	40,729	27,346	13,462	4.0%	4.2%	3.6%
Codes for special purposes	24,411	15,505	9,054	2.4%	2.4%	2.4%
Mental and behavioural disorders	13,353	8,485	4,935	1.3%	1.3%	1.3%
Certain infectious and parasitic diseases	11,861	7,187	4,719	1.2%	1.1%	1.3%
Congenital malformations, deformations and chromosomal abnormalities	11,520	6,330	5,196	1.1%	1.0%	1.4%
Diseases of the genitourinary system	8,922	4,945	3,997	0.9%	0.8%	1.1%
Diseases of the musculoskeletal system and connective tissue	5,464	2,587	2,875	0.5%	0.4%	0.8%
Diseases of the blood & blood-forming organs & certain immune mechanism disorders	3,480	1,984	1,491	0.3%	0.3%	0.4%
Diseases of the skin and subcutaneous tissue	1,917	1,234	686	0.2%	0.2%	0.2%
Diseases of the ear and mastoid process	372	np	285	0.0%	na	0.1%
Certain conditions originating in the perinatal period	337	327	np	0.0%	0.1%	na
Pregnancy, childbirth and the puerperium	268	0	266	0.0%	0.0%	0.1%
Diseases of the eye and adnexa	np	np	np	na	na	na
Total	**1,019,732**	**648,941**	**373,676**	**100.0%**	**100.0%**	**100.0%**

Source: ABS catalogue 3301.1

The weighted data shows higher levels of years of life lost among males with external causes of morbidity and mortality being a cause where males' years lost greatly exceed females' years lost.

Reflection

What has been my most important learning and why?

Reflection/discussion

What has been my most important learning and why?

From our analysis, what will we present to our minister regarding our migration data?

How many pages are we allowed?

The analyst's mindset vs the presenter's mindset - Decomposition vs Synthesis

Analysis involves breaking down data and information into detail in order to understand it. Analysis involves decomposition.

The analyst's mindset is one of immersion into detail in order to understand. It is easy to get lost in this detail. It is also easy to fall into the trap of presenting this detail to others to show them how we reached our findings.

When we advise others, we need a presenter's mindset. This involves synthesis. Synthesis involves combining and summarising information to reveal an insight or finding. It requires us to draw conclusions from among the details.

Presentation necessarily involves making choices. Remembering the definition of a narrative – choices about what to include and exclude as well as choices about what sequence to present them in.

- We will need to prioritise some findings and data over others. We exercise judgement in determining that some findings are more important than others.
- We will need to exclude from our presentation details of the analysis we did and exclude less important points.
- Beware of the desire to show everything you did in undertaking your analysis. Our decision makers want to know what we found and which of our analysis led us to these findings.
- We will need to summarise our analysis
- We will need to combine and consolidate our work.

Another choice - to cut a long story short

The immersion into detail by the analyst can lead them to develop a long story with lots of detail.

We know many of our readers are time-poor. They don't have time for the long story. They will demand the short story.

Converting a long story into a short story requires a presenter to make choices of prioritisation, exclusion and summarisation.

A briefing choice – how we use space

It will be the case for government briefings that we are given a page or size limit for our advice. This limit will require us to make choices about how we best use space on the page. We may not have space to use expansive graphs or tables.

We will likely need to develop visualisations that do not occupy a lot of space. We can create tables that use the horizontal space on the page while not occupying a lot of vertical space on the page. This will require our exercising of choice about the prioritisation and ranking of data.

An example: A diagnostic process in medicine

A patient presents with symptoms of pain and discomfort arising from lumps in their stomach.

Analysis is needed to understand the nature of the lumps and the underlying causes. The lumps could be benign, they could be cysts, they could be hernias, they could be tumours.

Some historical information will be gathered

- When did the patient first notice the lumps – was it sudden, was it gradual
- Have there been any recent events that triggered the lumps or the pain?
- What is the patient's relevant health history
- Is there a family history of this

A range of tests could be performed:

- Touching and scanning
- Listening
- Blood tests
- Urine tests
- Biopsies
- X-rays
- CT scans
- MRIs

We can view the different tests and different types of analytical methods required to discover what the story is. Some of the tests may reveal nothing. It may be only one of the tests that ultimately reveals the problem. It could be a combination of the tests that ultimately reveals the problem

From these tests, a diagnosis is reached. A diagnosis is a finding or conclusion in medicine with its own language and implications for treatment.

When presenting the results of this process, the doctor will not go through the details of every test conducted. They will present their diagnosis and recommended treatment from only those tests that revealed their findings.

The results from all other tests will be excluded from the presentation to the patient.

The long story will include all the history and all the tests that have been performed in order to reach a diagnosis and the treatment that is prescribed for that diagnosis.

The short story is the specific history and the specific test(s) that are relevant to the diagnosis and the treatment that is prescribed. The details of the history and details of the range of tests done are relegated in importance. They could be an attachment to the short story told.

References

Australian Bureau of Statistics

Catalogue 1500.0 - A guide for using statistics for evidence-based policy, 2010

Understanding statistical language - https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language?OpenDocument

ABS Data Quality Framework

(http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1520.0Main%20Features10May%202009?opendocument&tabname=Summary&prodno=1520.0&issue=May%202009&num=&view=) – to help you assess the quality of a dataset and design data collections which are fit for purpose.

Moment of Clarity (2014), Christian Madsbjerg, Mikkel Vedby Rasmussen

Naked Statistics (2013), Charles Wheelan

Everybody Lies (2017), Seth Stephens-Davidowitz

I Think You'll Find It's a Bit More Complicated Than That (2015), Ben Goldacre

DataStory (2019), Nancy Duarte

Storytelling with Data (2015), Cole Nussbaumer Knaflic

The Data Detective (2021), Tim Harford

Style Manual, Commonwealth of Australia

Edward Tufte:

Website: www.edwardtufte.com/tufte/index

Books:

The Visual Display of Quantitative Information

Visual Explanations

Envisioning Information

Beautiful Evidence

Stephen Few

Website: www.perceptualedge.com

Books:

Show Me the Numbers

Signal

Information Dashboard Design

The Data Loom

Sally Bigwood & Melissa Spore

Website: www.plainfigures.com

Book: Presenting Numbers, Tables and Charts

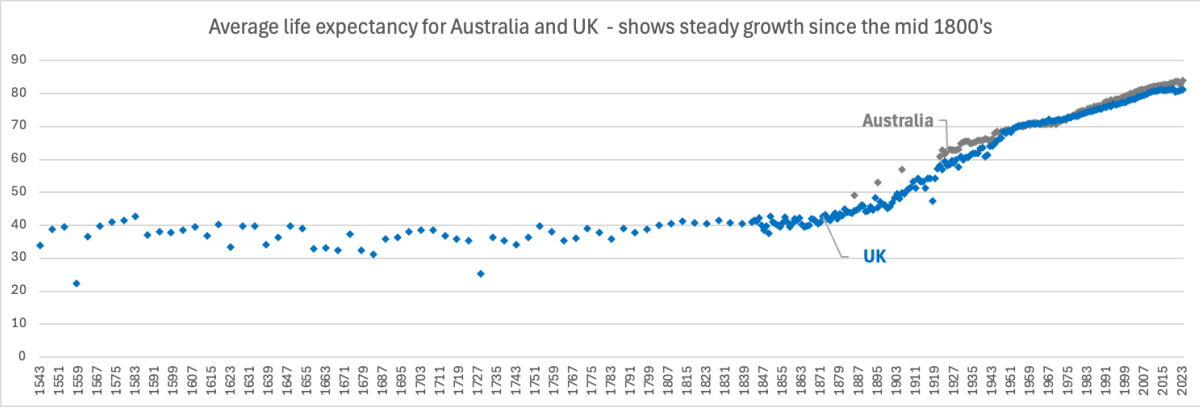
Appendix 1: A story of progress

Among the great achievements in the last two centuries has been the increase in average life expectancy. In 2023, average life expectancy in Australia was 83.9 years.

One hundred years earlier, average life expectancy in Australia was recorded as 61.7 years. While records were not kept in Australia in 1823, records from the UK and France show average life expectancy was around 40 years in the UK and France.

We have seen a doubling in average life expectancy over the last two hundred years. But the story is not just the last two hundred years. What about the years before the 1820s?

While records will not be as reliable, what records are available show that average life expectancy was consistently at levels around or just below 40 years.

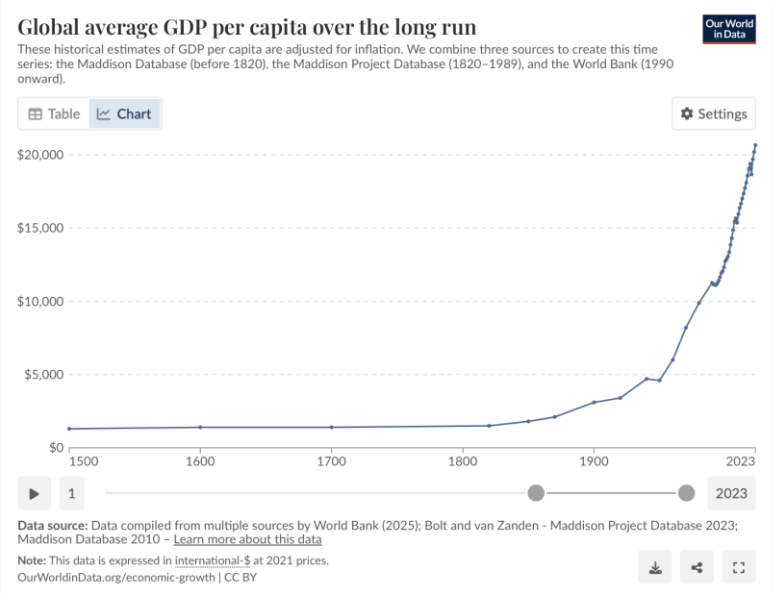


Source: Ourworldindata.org

The improvements in average life expectancy have been such that across our lifetime, our own life expectancy is growing by one year for every five years we live.

The dramatic improvement in life expectancy that started in the 19th century was not the only notable change in that period.

Records on the size of the world economy per capita shows similar dramatic growth during the 19th century.

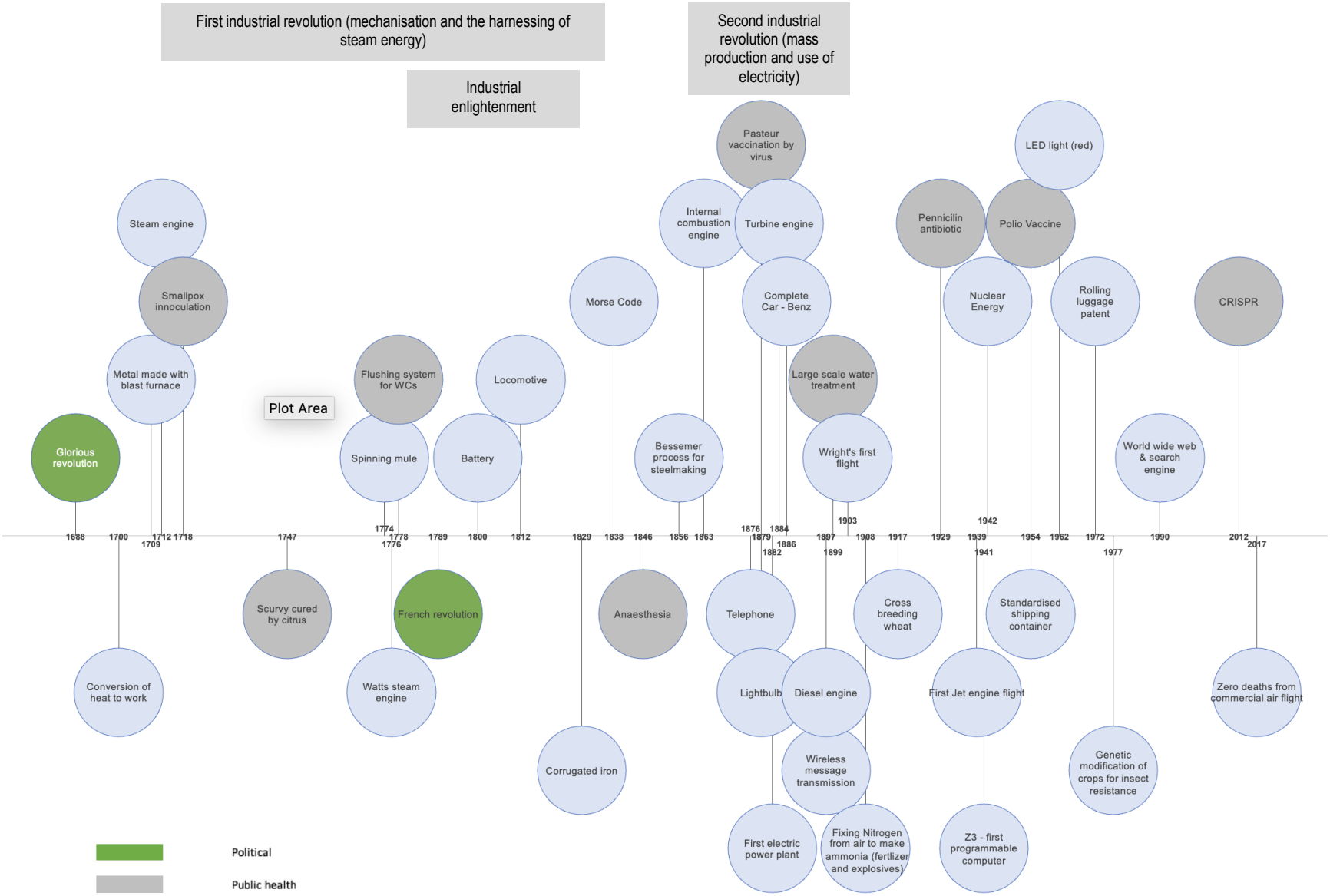


The nineteenth century has proved to be a launching pad for the huge progressions in economics and health we now benefit from.

Two formative developments preceded the nineteenth century. The first is the Glorious Revolution in the UK and the French Revolution in France. The term Glorious Revolution refers to the series of events in 1688 which culminated in the exile of King James II and the accession to the throne of William and Mary. It has also been seen as a watershed in the development of the constitution and especially of the role of Parliament. It marked the beginning of inclusive and pluralistic institutions that could provide legal frameworks that would ultimately enable economic growth and advances in health to be shared in countries like the United Kingdom, the United States and Australia.

The second development was the industrial revolution that had begun in the 1700's with the harnessing of energy for work, the invention of the steam engine, and technology like the spinning mule.

These advances were accompanied by health innovations like a smallpox inoculation and cures for scurvy. These developments, along with cultural developments that facilitated the spreading of science for industrial application, set the scene for fast progress in the 19th and 20th centuries. Technological developments, public health developments and other innovations through these centuries can be presented using a timeline.


First industrial revolution (mechanisation and the harnessing of steam energy)
Second industrial revolution (mass production and use of electricity)
Industrial enlightenment
Plot Area
Steam engine
Smallpox innoculation
Metal made with blast furnace
Glorious revolution
Conversion of heat to work
Scurvy cured by citrus
Flushing system for WCs
Spinning mule
Battery
Locomotive
French revolution
Watts steam engine
Morse Code
Corrugated iron
Internal combustion engine
Turbine engine
Bessemer process for steelmaking
Anaesthesia
Pasteur vaccination by virus
Complete Car - Benz
Large scale water treatment
Wright's first flight
Telephone
Lightbulb
Diesel engine
Wireless message transmission
First electric power plant
Fixing Nitrogen from air to make ammonia (fertlizer and explosives)
Cross breeding wheat
Pennicilin antibiotic
Nuclear Energy
Polio Vaccine
LED light (red)
Rolling luggage patent
Standardised shipping container
First Jet engine flight
Z3 - first programmable computer
Genetic modification of crops for insect resistance
World wide web & search engine
CRISPR
Zero deaths from commercial air flight
1688
1700
1709
1712
1718
1747
1774
1776
1778
1789
1800
1812
1829
1838
1846
1856
1863
1876
1879
1882
1884
1886
1897
1899
1903
1908
1917
1929
1939
1941
1942
1954
1962
1972
1977
1990
2012
2017
Political
Public health

Appendix 2: Assorted terminology and principles

Ackoff's Path

Analysis helps me navigate along a path described by Russell Ackoff (from Data to Wisdom)

Data – Information – Knowledge – Understanding - Wisdom

I do analysis to organise, discover and learn from data. I present data and analysis to help others discover and learn from data.

Data – a set of symbols or signals. In its raw form, it is unorganised and unprocessed. Census takers collect data.

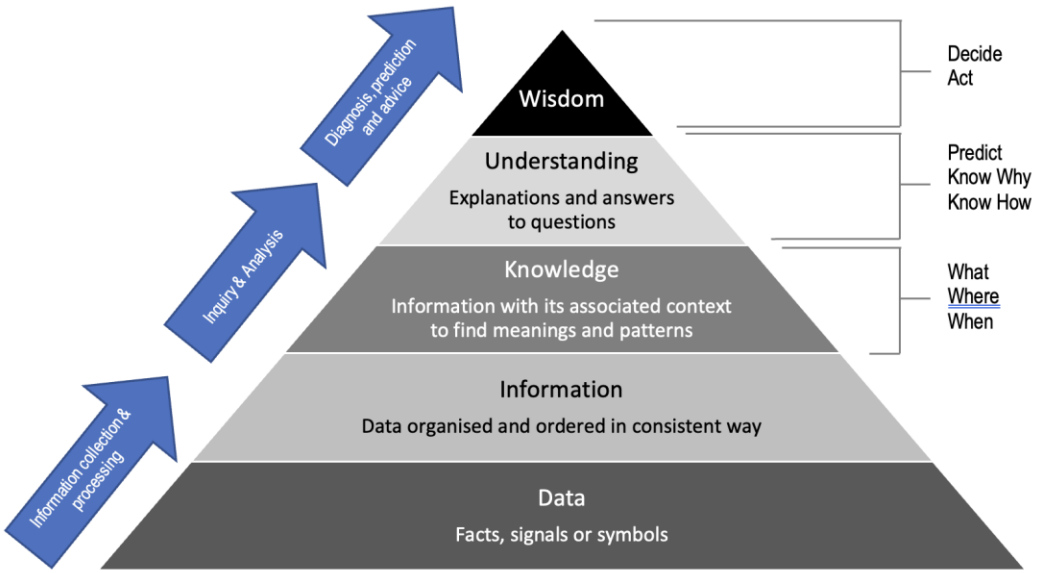
Information – consists of processed data. We apply systems to classify, organise and arrange data. Information is data with meaning. Data from the census is converted into tables that present the results from the census takers.

Knowledge – While information can help us understand relationships, knowledge is about finding meaning and patterns. Knowledge is conveyed by instructions – answers to how-to questions. From the census, we know what the average age of Australians is and how many Australians are living in different locations. We need to do some analysis to determine this.

Understanding – understanding is conveyed by explanations – answers to why questions. We need to do further analysis to answer why questions. The difference between knowledge and understanding is the difference between memorising (knowledge) and learning (understanding). Our analysis helps us understand why the average age has changed or why Australians have changed where they live.

Wisdom – apply our knowledge for future prediction and decision-making. Wisdom involves the exercising of judgement – to adapt our understanding to different circumstances requiring discernment. We can make policy decisions that improve health outcomes or improve quality of life.

Analysis is most heavily used to get from Information to Knowledge and from Knowledge to Understanding.



The MECE Principle (McKinsey Way)

"The **MECE principle**, pronounced 'me see', is a grouping principle for separating a set of items into subsets that are mutually exclusive and collectively exhaustive." Wikipedia.

The term has been popularised by McKinsey as it is deemed part of the McKinsey Way.

These terms are traditionally associated with probability theory but combine well as a principle for analysing data and setting up data sets.

Mutually Exclusive

Related in such a way that each thing makes the other thing impossible: not able to be true at the same time or to exist together

Pertains to the categorisation of data – categories should be mutually exclusive to ensure no double counting and to remove ambiguity on where items belong.

Completely Exhaustive

Does the data set capture the totality of what it is you are analysing? If so, the data set is completely exhaustive. It is a common failure of analysis to focus too narrowly on a subset of data in isolation from the total data set. This causes valuable information to be lost and can cause a change in conclusions from changes in scope.

Correlation vs Causation

The volume of data does not adequately address what is causing what. We need to look at causation vs correlation.

Correlation - Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

Causation - Connection between two events or states such that one produces or brings about the other, where one is the cause and the other its effect. Also called causality.

Correlation can often be confused with causation. Two variables can move together without it being clear which is causing which, or with the possibility that a third variable could be triggering both.

The Importance of Metadata

Imagine you have just joined a new team and you have been asked to analyse data for a monthly report.

You look at the data file and all the column headers are abbreviations, and you aren't sure what the data is in each of the columns.

You ask your colleague for written instructions on how the data is compiled and analysed so you can repeat the process, but nothing is written down and you must be shown.

You look at the last report to get further information and can see the data refers to licenses, service providers and consumers. Still, there is no further information about who the service providers and consumers are, and what sort of licenses are included in the report.

Your team members can verbally provide you with some information about the data, but in some cases, no one knows because your predecessor "managed all that".

Your job would be so much easier if information about the data had been written down…

Data is an asset, and its users need information about the data to help them make better use of it. The descriptive information about the data we use is referred to as **metadata**.

Definition

> *Metadata is the information that defines and describes data.*
>
> *Metadata is the information you need to have to have an accurate understanding of the (associated) dataset.*
>
> *Metadata helps people to correctly interpret the data. It also helps future producers of the same or similar collections of data to understand the collection or issues that might arise in a new but similar collection.*
>
> *Metadata includes details of what the data is measuring, that is, exactly how the data items are defined. Any standards of classifications used to categorise data items are also part of the metadata, as is the sort of information contained in many explanatory notes, such as giving details of the scope of the collection.* Source: ABS

For an everyday example, consider what happens when you take a photo. You are creating data in the form of a saved digital image.

While what you see is the photo, your smartphone also creates metadata accompanying that image, such as:

- Image size
- The time when the photo was taken
- Location where the photo was taken
- A thumbnail of the image.

Your phone manages this metadata, allowing the image to be easily searched and identified.

These descriptions will not only help those using the data, but they will also help those who are designing and building datasets, as the descriptions provide rules for how data should be collected and structured. They will also help people providing data as the descriptions help them to understand exactly what data we are asking for.

Big data

"Extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions." Source Dictionary.com

"Big data is a term applied to data sets whose size of type is beyond the ability of traditional relational databases to capture, manage, and process the data with low latency. The data has one or more of the following characteristics – high volume, high velocity or high variety. Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, the web and social media. Much of this data is generated in real-time and on a very large scale." Source IBM Analytics

Examples of how Government Agencies could use big data

1. To assist in Fraud Detection and Financial Market Analysis – Centrelink and Social Security
2. To assist in Health-Related Research – Food and Drug Administration for foodborne illnesses
3. To assist in Government Oversight and Education – tracking legislative changes or school results
4. To assist in Fighting Crime – Home Affairs and Homeland Security
5. To assist in Environmental Protection and Energy Exploration.

See: http://www.businessofgovernment.org/BigData3Blog.html for more details.

Big data will include data sets that are too large to view. Information about the data is sought through graphs or data summarisation. Working with big data will also require analytics to be done to test the quality of the data for things like missing values, outliers, etc.

With big data, visualisation techniques like different types of graphs are used to understand the data. With such large data sets, the size of the data set prohibits being able to draw conclusions from within the data.

The analysis of big data sets usually requires some programming skills, which are beyond the scope of this course. The skills are required to cleanse the data and to run hypothesis testing to identify relationships between different variables. Skills are also required to convert information that is not quantified data into numbers. For data sets of over 1 million rows, this work cannot be done in traditional software like Microsoft Excel and requires more specialist software like Microsoft R and Hadoop (from SAS).

Another software package called Microsoft Power BI is a data visualisation package that enables summary tables like those from pivot tables in Microsoft Excel and it enables quick graphics to be produced to understand relationships within the data.

With big datasets, Data visualisation is part of the analytical process to establish a view into the underlying data, its quality and to test possible relationships within it.

Other terms

Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data.

The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting them to predict the unknown outcome.

It uses historical data to find patterns that can be extrapolated into the future. This has been done for years in financial markets, sports and gaming. Predictive analytics aims to use correlations and causations to predict future directions of data.

Data cleansing or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Missing values: In statistics, **missing data**, or **missing values**, occur when no data value is stored for the variable in an observation. Missing data is a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

The presence of missing values means we may have to exclude the data or, in some cases, impute values.

Appendix 3: Data questions and data quality

Think about and question the data

Thoughts from Tim Harford in Data Detective:

- Step back and see the context
- Get the backstory
- Ask who and what is missing
- Demand transparency when algorithms are applied and used
- Don't take strong statistical institutions for granted
- Keep an open mind when working with data.

Questions to ask of data from Stephen Few in the Data Loom:

Required data – What data is required for the task at hand?

Semantics of the data – What do the various fields of data mean?

Relevance of the data – Are all of the data fields relevant to the task at hand?

Source of the data – What is the source of the data and is it credible?

Accuracy of the data – Is the data accurate?

Completeness of the data – Does the data set include all that's needed?

Context of the data – Have I taken all of the relevant context into account?

Representativeness of the data – Is what's revealed in the data typical?

Causes of the behaviours recorded in the data – What is causing this to happen?

Aggregation of the data - Is the level at which the data has been aggregated and the statistical method that was used to produce that aggregation appropriate for the task at hand?

Data Quality

ABS Data quality framework

Data quality is not just about accuracy. Data is generally considered high quality if it is fit for its intended uses and if it correctly represents the real-world construct to which it refers. Data is fit for its intended use if users can do what they need to do with it (analyse, report, make decisions) with reasonable assurance.

There are a number of different dimensions to data quality. We do not dismiss the use of our data because it has some quality shortcomings. We can still assist and improve decision-making with imperfect data but our decision makers need to understand its strengths and limitations.

The Australian Bureau of Statistics provides information on a data quality framework on its website.

The ABS identifies seven different dimensions of data quality.

1. Institutional environment;
2. Relevance;
3. Timeliness;
4. Accuracy;
5. Coherence;
6. Interpretability; and
7. Accessibility.

Institutional environment refers to the institutional and organisational factors that may have a significant influence on the effectiveness and credibility of the agency producing the statistics.

Relevance refers to how well the statistical product or release meets the needs of users in terms of the concept(s) measured and the population(s) represented.

Timeliness refers to the delay between the reference period (to which the data pertain) and the date at which the data become available, and the delay between the advertised date and the date at which the data become available (i.e. the actual release date).

Accuracy refers to the degree to which the data correctly describes the phenomenon they were designed to measure. This is an important component of quality as it relates to how well the data portrays reality.

Coherence refers to the internal consistency of a statistical collection, product or release, as well as its comparability with other sources of information.

Interpretability refers to the availability of information to help provide insight into the data. Information available that could assist interpretation may include the variables used and the availability of metadata, including concepts, classifications, and measures of accuracy.

Accessibility refers to the ease of access to data by users, including the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which information can be accessed.

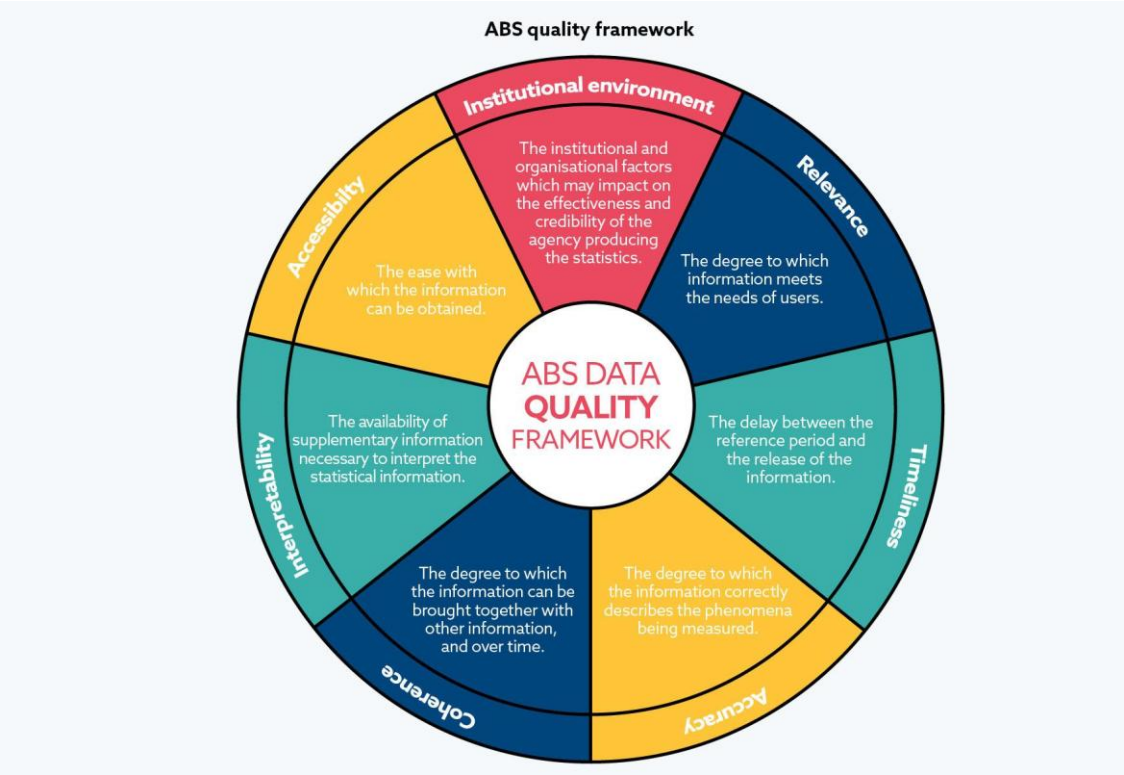
Details of the framework can be found on the ABS website by referencing:
https://www.abs.gov.au/statistics/detailed-methodology-information/concepts-sources-methods/australian-system-government-finance-statistics-concepts-sources-and-methods/2015/16-data-quality/part-b-abs-data-quality-framework

ABS data quality framework

The Australian Bureau of Statistics provides information on a data quality framework on its website. Details of the framework can be found on the ABS website by referencing:

1520.0 - ABS Data Quality Framework, May 2009

A graphic overview of the framework is below.



The framework refers to seven elements for assessing data quality.

1. Institutional Environment - Who produced or collected the data? Under what authority or legislation were the data collected? To what extent are quality guidelines documented by the agency? The agency's capability to meet the production or collection of data.
2. Relevance - Details on exactly what data was collected (e.g. Key data items and definitions, Geographies, Populations from which the data came, Relevant time periods).
3. Timeliness – how old is the data and what is the reference period (the gap between the date of reference and use), when data becomes available (publication dates) and the frequency of data collection.
4. Accuracy – considers aspects like: Sampling error, non-response error, coverage error and any revisions that have been applied to the data.
5. Coherence – the ability to link or compare with other data and has regard to whether the data has been confronted with other data sources and are the messages consistent from all data sources.
6. Interpretability - is there supporting information (e.g. concepts, sources and methods) to enable interpretation of the data?
7. Accessibility – ease or difficulty of access and use (e.g. what format does the data come in, or is it confined to particular types of software?).

Appendix 4: Establishing data sets

Everything is being converted to data

- Our names
- Our addresses
- Our movements are tracked by GPS
- The type of car we drive
- Our age
- Our marital status
- The medications we take
- Our taxation history.

Data can take many forms. I most commonly work with financial data or with transactional data.

The first starting point is to ensure our data is arranged into different fields that contain categories or values. In Excel, datasets will be presented in fields, each represented as a column.

The data I work with relies on being able to establish the following types of fields:

- Categories – e.g. teams, products/services, accounts, regions
- Time – ensuring different time periods are separated
- Quantified values – dollar values, transaction volumes, number of people
- Weights – values assigned to recognise the complexity, difficulty or time associated with a category.
- Weighted values – where original values are converted into weighted values.

This often requires that modifications be made to the original data to get it into a form possible to undertake and present analysis.

This is best demonstrated by way of example.

Working with vehicle registration data

The largest data set I have ever worked with is the register of vehicles in South Australia.

Although large (about 243,000 lines), I could work with the data in Excel. That means I could see the data and scroll through it, albeit glancing only.

The data set included:

1. The make of vehicles
2. The model of the vehicle
3. The year of manufacture
4. The number of cylinders
5. The number of vehicles in these categories

The data included no names or addresses. The only locational data was the fact that the register was of vehicles in South Australia.

The data was provided in a table format as below:

Source data	A	B	C	D	E
1	Make	Model	Year of Manufacture	Number of Cylinders	Number of vehicles
2	Toyota	Camry	1997	6	450
3	Honda	Jazz	2002	4	321
4	Ford	Mondeo	2004	6	150
5	Holden	Jackeroo	2000	6	240

A dataset like this might be used to explore the impact of different options for registering or treating vehicles. Options could include:

- Register by age
- Register by mass
- Register by length/size.

However, the original data does not come with information about the mass or length of vehicles. This is quite common in working with data, recategorising the data. In this case, cars would have to be recategorised by mass.

To recategorise data, I create data maps. Data maps will seek to assign existing categories to a new category. For example, a data map for vehicle models to a new category is:

MAP	A	B
1	**Current category**	**New Category – size**
2	Camry	Medium light vehicle
3	Jackeroo	Large light vehicle
4	Mondeo	Medium light vehicle
5	Jazz	Small light vehicle

In Excel, we can use a function called VLOOKUP to assign a new value to an existing category.

To do this, we can save a new version of our dataset and add another column for "Size". I usually insert the new category before the values or numbers columns. We can then use the VLOOKUP function to assign the sizes in our map to the vehicles in our dataset.

For the Toyota Camry, the formula would be:

=VLOOKUP(Source Data B2,MAP $A$1:$B$5,2,=FALSE)

This would result in the following table:

Make	Model	Year of Manufacture	Number of Cylinders	Size (using the VLOOKUP formula)	Number of vehicles
Toyota	Camry	1997	6	Medium light vehicle	450
Honda	Jazz	2002	4	Small light vehicle	321
Ford	Mondeo	2004	6	Medium light vehicle	150
Holden	Jackeroo	2000	6	Large light vehicle	240

This provides a new way of categorising and, therefore, analysing the vehicles.

The process can be repeated for other new categories.

Once a good dataset is established, Excel tools like pivot tables and a range of formulas can be used to summarise the data.

A ***pivot table*** is a table that summarises data in another table and is made by applying an operation such as sorting, averaging, or summing to data in the first table, typically including grouping of the data. Source: Wikipedia

A pivot table doesn't change the original spreadsheet or database itself.

For motor vehicles, a pivot table could generate a summarised table like the one below.

	Number of cylinders				
Size	**4**	**6**	**8**	**10**	**12**
Small	#	#	#	#	#
Medium	#	#	#	#	#
Large	#	#	#	#	#

Appendix 5: Insights about analysis

The analytical process

For this purpose, I will use the analogy of mining for a diamond as our goal to outline the analytical process.

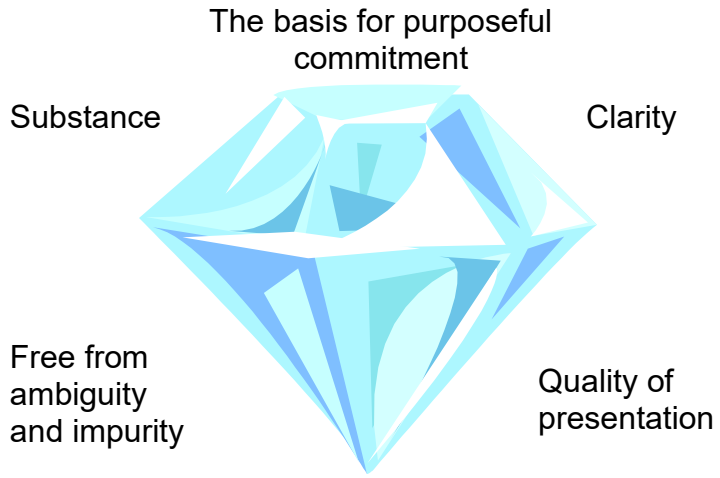
Analysis is a key step to revealing insight. Using this analogy, insight is our diamond. It is the value that we are seeking, to help others make a decision or to make a change. A diamond is also a symbol of engagement. Engagement is a critical element in getting decisions made with our analysis.

Engagement – when we want others to decide

A diamond is a symbol of engagement. A diamond also represents high quality.

When we want the engagement of our customers, our managers and our colleagues, we need to come to them with a diamond. Anything less will not cut it.

Many leaders seek the engagement of their staff and the engagement of their customers. Engagement is connecting to achieve commitment. In the workplace, it is to get a commitment to action and to change. The process of engagement is a process of seeking a decision from others to move forward with us.



When we are seeking to engage, we want the highest possible quality decisions from our work. We want:

- True value based on substance
- Value that is obvious to others
- Clarity and elimination of ambiguity.
- Incisive action and commitment

These are all the attributes of a high-quality diamond: High value, high clarity, elimination of ambiguity and a true substance that will cut through anything.

The analytical process is therefore akin to the exploration and digging processes used in mining for a diamond. In mining information, we are looking to pull it apart and find out what's inside. We are looking to identify areas that are of important value to us in revealing information that may sway decision-makers. We are looking to reveal information that may help us decide how to change the course of our project or team, to keep it viable, within budget and achieve our goals.

Following are the key steps involved in undertaking data analysis.

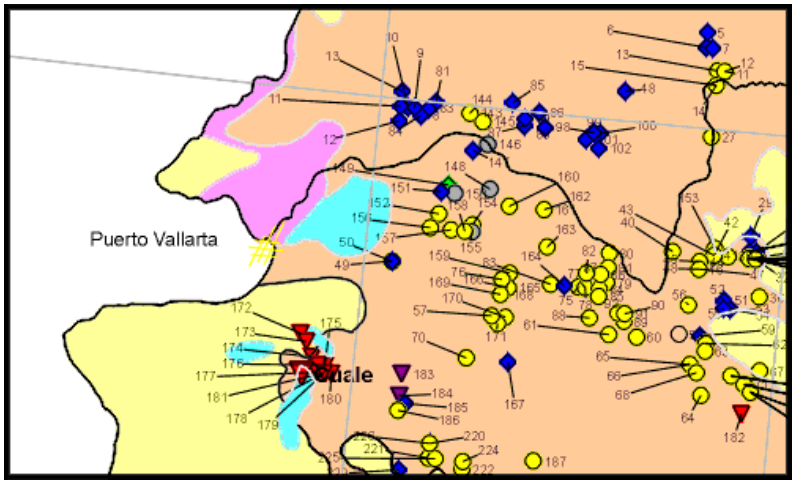
1. Be clear on the purpose of analysis – which decision(s) is it informing
2. Pre-drilling – initial exploration of data
3. Drilling into detail
4. Identify key findings
5. Start preparing for presentation – remove ambiguities
6. Polish to enhance clarity

Purpose - focus on the decision

The first step is to understand the purpose of the analysis and the key decisions that you are looking to inform or influence as a result of doing this analysis. It is important that we have clarity at this point, for our analysis to be as focused as possible.

Pre-drilling – explore overall data and identify priority areas

The second step is to source our basic data, starting from the top of the relevant reporting unit and working down. It is important with data analysis that we ensure that the story we are telling is compiled in the context of the whole organisation within which we are working. Rather than being selective, and in some cases taking a convenient snapshot, it is important that the data is seen as part of a holistic picture, if possible, to ensure that we cannot be accused of being selective or missing key elements because we are too focused on our slice of data. Initial exploration of our basic data is like exploring to identify the plot or claim that we want to set aside for further drilling and ultimately mining.



This step involves working through some basic analytical techniques to establish priority areas of focus. Initially, I will use compositional analysis to identify major elements and then undertake temporal analysis to understand significant growth or changes over time. These will start to identify where we think we need to drill further. As with mining, we are keen to ensure that our

drilling efforts are as focused as possible because it can be, like mining, time-consuming and expensive.

Start drilling (cautiously at first), some construction may be required

Having established some key areas for drilling, it is important to assess the quality of the data into which we are drilling. In some cases, data will be well formulated and consistent across time, which will enable us to quickly move to using more sophisticated analytical techniques. However, there will be situations where the quality of data is poor and effort will need to go into constructing a solid data set that is comparable over time or across units. Using the mining analogy, there may be cases where we need to reinforce areas of ground before entering them, to make sure that mining is safe. In these cases, we need to construct some mechanisms to support our analysis.

This will be particularly the case where we are comparing specific transactional information over time, where we are developing ratios, and comparisons over time or where we are constructing benchmarks to compare different entities. In these cases some basic methods of reinforcement to the data need to occur, to ensure it is robust.

Stop to identify and assess findings

By drilling into details, we can pick apart detailed information to draw some conclusions. It is at this point in the process there we are seeking to spot diamonds. It is at this stage of the process that we often need to slow down to be more careful about what we are observing.

Diamonds will not just jump out at us. We need to scratch a little, we need to manoeuvre the data in different ways, compare it in different ways and try different forms of analysis before the insight reveals itself. As with the mining process, often our diamonds are buried amongst the rubble and are not always that easy to find. This demands some reflection and careful scanning to reveal our findings.

Start preparing for the presentation – remove ambiguities

When we start to draw our conclusions, our data will often be buried amongst rubble. At this point, it is not fit for presentation. Once we have revealed key causes from within our data, we need to move to present our information to others. It is often the case that we can use many different instances of data to come up with some very simple and basic conclusions. In these cases, it is important that we leave a trail behind on how we come up with the data we use so that we can present our methodology to others.

Having gotten our hands dirty to reveal something in the data, the next step is to start working on the presentation of our information. We do not want to present to our audience a pile of rubble and let them know that amongst it is a diamond. Rather, we would prefer to present to them the diamond in all its glory and give them some background on how we arrived at it to demonstrate substance and credibility.

Final polish to enhance clarity

A key quality indicator of our data analysis is transparency. Transparency and clarity are key elements of the quality of the diamond and are also key elements in a quality decision that others make with our data analysis and information.

In this way, we see a flipside to analysis. Analysis is the deconstruction of information into its constituent elements. In contrast, the presentation of our analysis requires the reconstruction of data based on key insights. Therefore, an important aspect of analysing data is the ability to present the findings of our analysis to others in a polished and clear way.

To do anything less will compromise the extent of engagement and buy-in we get from the decision-makers receiving our analysis.

Conclusion

Analysis is the process of breaking something into its constituent elements; a detailed examination of the elements or structure of something, typically as a basis for discussion or interpretation.

We are trying to get to the bottom of an issue within an organisation to understand the key drivers behind it. By starting with basic analysis and then drilling, we are seeking to understand the cause and effect behind the data and changes in the data.

We can go through a lot of information to reach a small number of conclusions. In fact, the more thorough our analysis, we can have a situation where there is an inverse relationship between the amount of information we analyse and the number of conclusions we reach or insights we find. This is a key indicator for me of the quality of analysis.

NOTES