Australia's National
Science Agency

Educational Datasets Companion

Guides for using Data in the Classroom

Contents

1 Asking Questions of Data

When exploring data, we need to come up with a list of questions the data could provide us answers to then narrow that list of questions to the ones which are useful and relevant.

If, exploring rainfall data, it might be useful to know which months get the most average rainfall and which get the least, so we can use that information for water planning or gardening. The same data could also tell us which months have an average rainfall of less than 7mm on odd numbered days, but the answer to that question has very limited usefulness.

Quantitative Questions

Quantitative questions relate to numbers, or comparative numbers. Data is good at answering quantitative questions. Simple examples of these include: Is this value bigger than that one? What is the largest value in this list? Words that indicate quantitative questions include most, highest, largest, most frequent, most common, least, lowest, smallest, least frequent, least common.

Looking at election data, we can easily find out which parties got the most and the least votes. We can also find out which party won the lowest number of seats, and which won the highest.

Qualitative Questions

Qualitative questions look for descriptive, subjective measures. When asking someone to pick the best pair of shoes from a list, their answer will depend on how they define the term 'best' in terms of shoes. Best looking? Best fitting? Cheapest? A combination of those qualities? In answering these questions, opinion is a big factor. Two people can have very different opinions about which shoes are nicer looking or better fitting.

Questions asking which item is best, nicest, prettiest, worst, safest or most dangerous are difficult to answer with data. We can change the focus of our question or the way we collect data to try to answer similar questions. Instead of asking which pair of shoes is best looking, ask which pair gets the highest average 'attractiveness' rating from survey respondents. Instead of asking which is worst, we could ask which gets the lowest ratings across all categories on the same survey.

Time Questions

Data is also great at answering questions relating to time: When does an event take place? When posting content to social media, what time will likely give the highest number of views and interactions?

Using rainfall data as an example, we might look at weeks of highest rainfall across an entire year, or examine specific times of day, to see if rainfall is more likely in the afternoon or the morning. This would give us information that would help planning crop cycles and water tank usage.

Place Questions

Data can answer where questions: ‘where are the most car owners located?’, ‘Where are the least number of supermarkets’, ‘where does it rain at least 40 days a year?’, and complex questions like ‘which cities get rain at least 40 days a year, and also have a population of over 200,000 people?’

How & Why Questions

We often want to understand why events occur. Data can show correlation and indicate that when one thing happens, something else is likely to happen but it’s very difficult to prove that one thing caused another. Finding correlation in the data does not mean that there is a causative link.

One study found that women were 47% more likely to be severely injured in car accidents than men. It was assumed that this indicated that women were worse drivers. Later, it was noted that the crash test dummies used to test cars for safety were created only in male body shapes, resulting in safety design being optimised for males and not females and a higher injury rate among women. There was a correlation between injuries and biological sex, but the cause of the injuries was not being tracked in the data. Correlation is easy to check for, but causation can be difficult to identify.

Data can answer questions about whether things might be related, but it can be difficult to determine how they are related. A strong correlation in the data indicates that a relationship may exist, but it isn’t evidence of one thing causing another. Correlation does not imply causation.

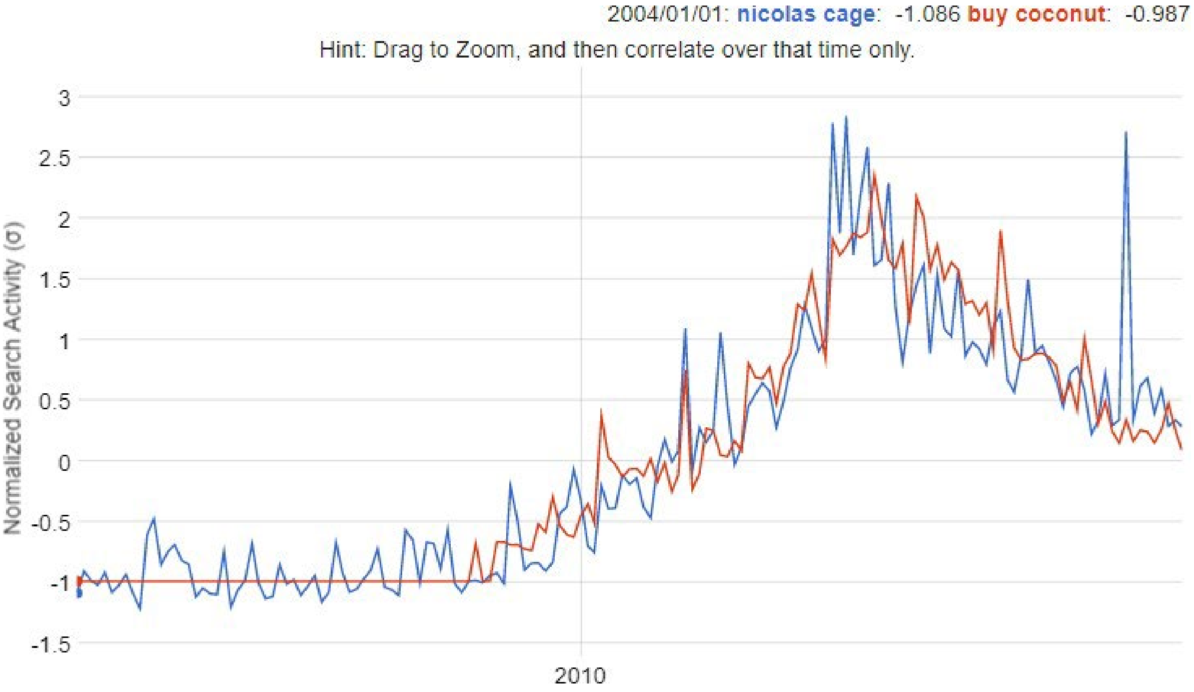


Figure 1 - Comparing Google Searches over time for ‘Nicolas Cage’ and ‘Buy Coconut’ in Australia indicates a very strong correlation. Data Source: Google Correlate

This example indicates a strong correlation between the two searches, with r=0.9484 (r=1, meaning the graphs have an identical pattern). If correlation was an indication of causation, it might mean that Nicolas Cage movies inspire Australians to buy coconut, or that buying coconut makes Australians interested in Nicolas Cage. Without further investigation to understand what the linkmight be, we cannot say that one causes the other.

2 Misrepresented Data

Misrepresented data can be a huge concern. Data can be used to tell any number of stories, and if manipulated in the right way, whether maliciously or accidentally, the story may be a little different to what the data says when taken at face value.

There are any number of ways that data can be misrepresented, but here are two of the most common methods used to misrepresent data.

Cherry Picking

Cherry picking means that a presenter is ignoring data that does not fit the narrative they are trying to support, and only using the data that fits. People can be prone to cherry picking when performing experiments with an expected outcome. Knowing what the results should be creates an inclination to bend the data to try and fit the expectations.

A key example of cherry picking would be looking at growth within a business. Sales data from a specific period of growth could be used to indicate tremendous ongoing growth and imply an unsustainable upward trend without looking at longer term patterns.

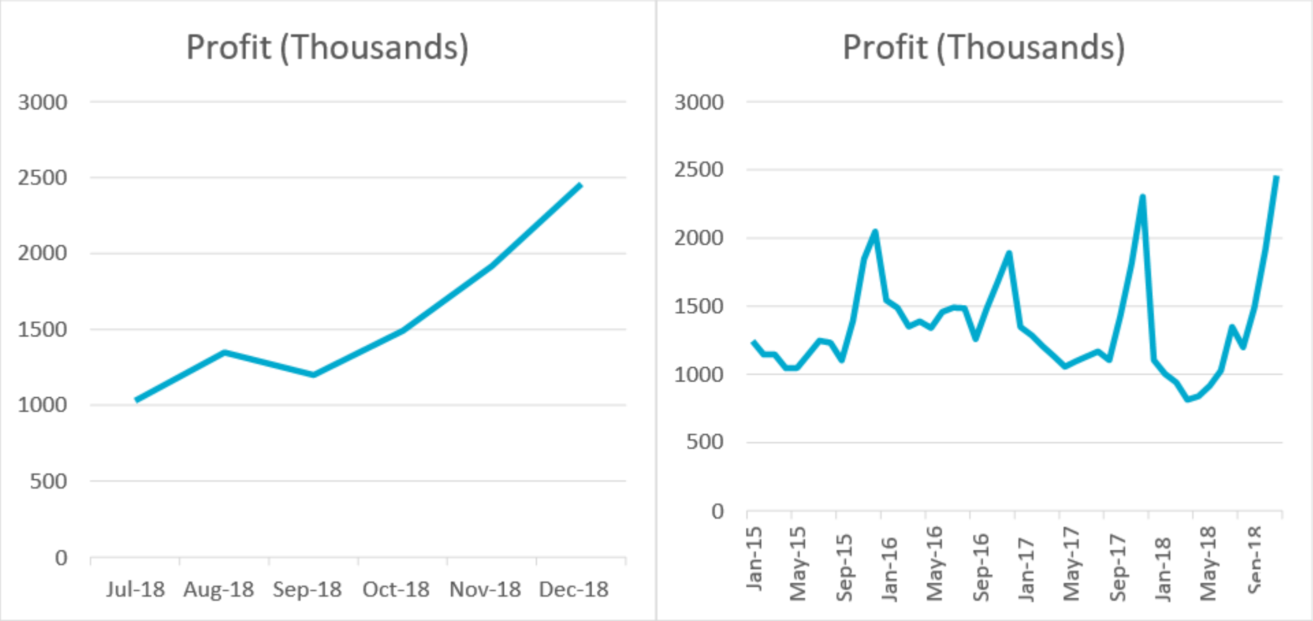


Figure 2 - These two charts show the same dataset, but in the left chart, only the last 6 months of sales are shown, indicating a high level of profit growth. The right chart indicates that there is a pattern of seasonal spikes, and profitis likely to drop off again in the next month.

Cherry picking is a regular tactic for people wanting to manipulate data to tell a specific story, as it is often easy to isolate small subsets of data that trend in the direction needed, despite the overall trend being in the opposite direction. It is often used when analysing climate data, pointing out sections of very cold temperatures in specific places, despite the overall trend indicating that the global average temperature is rising.

Scale and Axis Range Manipulation

Graphs can be manipulated in several ways. Most notably the scale and range of the axes can be changed to tell different stories. Having a smaller range on one of the axes will suggest at first viewing that there is a much higher variance in the data being displayed.

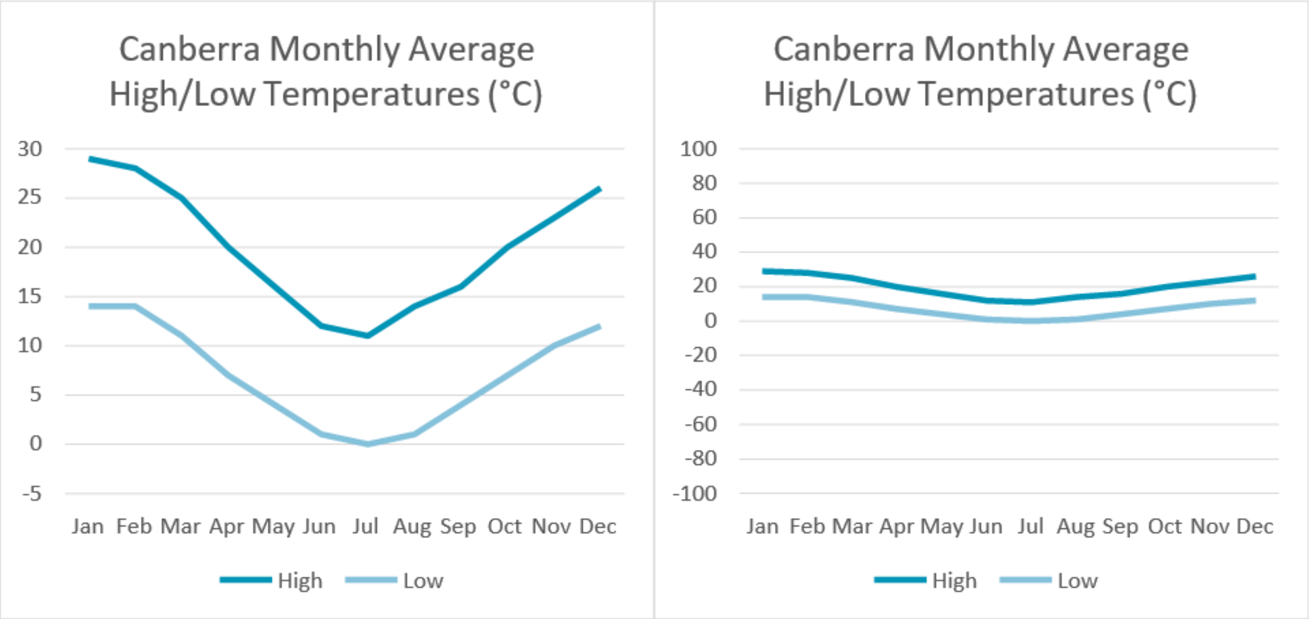


Figure 3 - Both graphs show the exact same data, but with completely different ranges and scales. Looking at the right-hand graph alone, you could be forgiven for thinking there is very little temperature variance throughout the year in Canberra. That the scale for the right-hand graph exceeds the highest and lowest temperatures ever recorded on Earth makes the scale disingenuous.

Adjusting the scale, changing the starting value or showing no scale at all are very common methods of manipulating the scale to misrepresent trend data, to make it look like trends are larger or smaller than they seem.

3 Outliers – What Are They?

An outlier is a data point that is very different to the rest of the dataset. Unfortunately, there's no easy way to decide how different a data point must be to be an outlier. There's also no easy way to work out how to treat outliers. They can be mistakes – measurement errors, entry errors, or calculation errors, but others are legitimate readings that fall outside the expected range. An example of this is a hard-working student who scores 100% on a test when the rest of the class scores in the range of 40%-70%. A result of 100% falls well outside the expected range when looking at all the other scores, but it is a valid data point. Alternately, if while recording a mark of 71% the teacher's finger slips and they enter 41%, that is a data point that falls inside the expected range, but should be identified and fixed, so it doesn't change the class average (or the student's results!).

If we knew which data points were erroneous, we could remove them from the dataset so that they don't distort our data. Unfortunately, as in the example above, it's not always easy to identify the mistakes. We could have a perfectly legitimate result that is well outside of the expected values or an erroneous value that falls within the expected values.

There is no rigid rule to identify outliers in any given data set, as it requires considering the nature of the data being collected. It can be a good idea to examine the data around the potential outlier. For example, if data shows temperature readings across the day, and the temperature suddenly goes from 26°C to 43°C then back to 26°C, it's likely that the 43°C was an error, caused in measuring, recording, or by external forces impacting on our reading like someone opening an oven door near our recording equipment. A reading of 43°C is a valid temperature recording, but to have a reading so high for a short period of time indicates that it is likely to be an anomalous outlier (or a very rare atmospheric phenomenon called a heat burst).

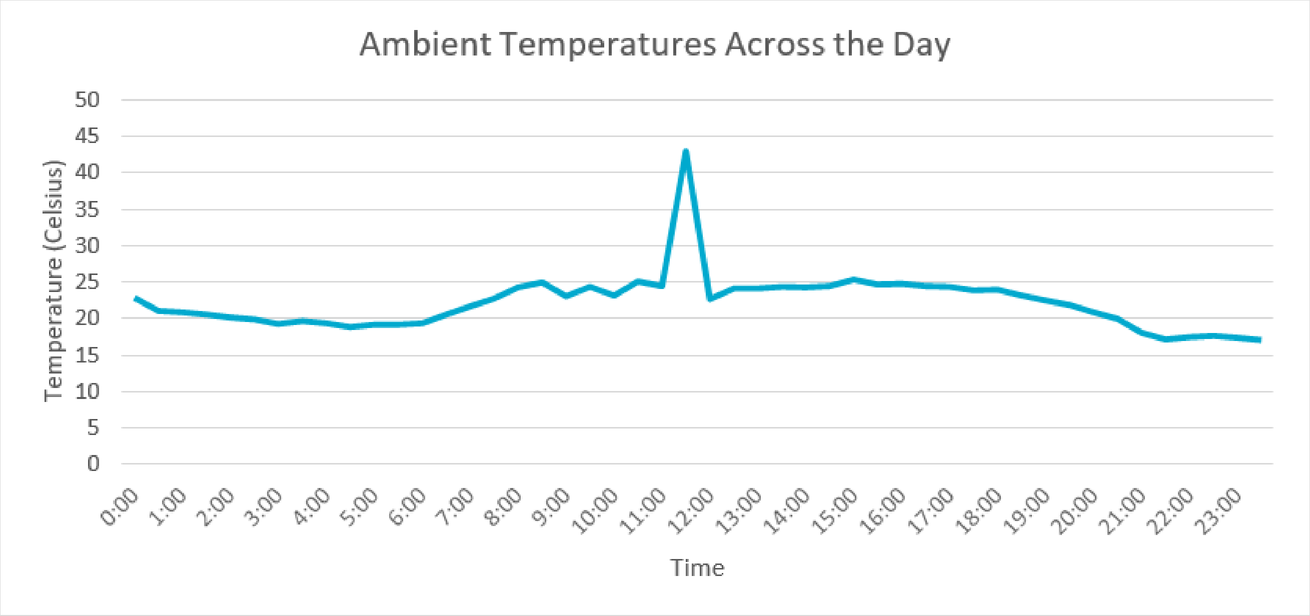


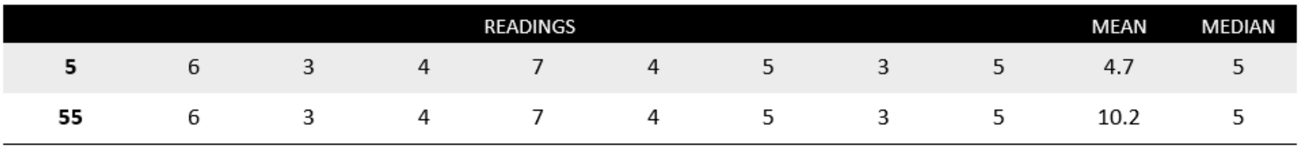
Figure 4 - A temperature graph of hourly temperature readings for a single day, with one anomalous reading

What if the above data was not measurements of temperature, but rainfall measurements from a digital rain gauge? It would be less likely that the spike indicates an obvious error. There may have been a dramatic increase in rainfall at noon as the storm intensified. In this case, what would be an obvious recording or measuring error in one dataset tells us something valuable in another.

The examples above highlight that we can only ignore outliers when we know for certain they are errors in either recording or measurement. If we ignore outliers that are legitimate readings we could miss valuable details, which could lead to incredible discoveries in areas that were not anticipated in the original research. Many stories exist in science of unintended discoveries, such as Fleming's experiments with Staphylococci becoming contaminated by Penicillium mould, which inhibited the growth of the bacteria. Many of the methods that are used to search for exoplanets (planets outside our solar system) involve examining stars and looking for unexpected data points in observation data that can be best explained by an exoplanet, including dips in the light the star gives off as planets pass between our observation point and their parent star.

Outliers can impact our analysis of data by distorting the summary statistics of a dataset – shifting the mean, standard deviation, and other measures. If the outlier is a valid value then that shift is generally also valid, but if the outlier is an error, whether due to measurement, recording, or another reason, then the distortion can interfere with our understanding of the dataset.

Consider this small dataset:



READINGS									MEAN	MEDIAN
5	6	3	4	7	4	5	3	5	4.7	5
55	6	3	4	7	4	5	3	5	10.2	5

Figure 5 – A small dataset showing an example of correct and incorrect readings, demonstrating how a single reading error can skew the mean

The mean of the first row is 4.7, but if the first entry was mistyped as 55 instead of 5, the mean becomes 10.2 and is higher than all of the data points except the first. If the mean is an important aspect of your dataset, this shift will be a problem. This is one reason that the median will be used instead of mean. The median for both datasets is 5. Medians are resistant to outliers.

Specific references to outliers can be found in the following content descriptions of the Australian Curriculum:

- Technologies – Digital Technologies
  - Acquire data from a range of sources and evaluate authenticity, accuracy and timeliness (ACTDIP025)
- Mathematics
  - Investigate and identify issues involving numerical data collected from primary and secondary sources (ACMSP169)
  - Explore the variation of means and proportions of using random samples drawn from the same population (ACMSP293)
  - Investigate the effect of individual data values, including outliers on the mean and median (ACMSP207)
- Science
  - Measure and control variables, select equipment appropriate to the task and collect data with accuracy (ACSIS141)
  - Summarise data, from students' own investigations and secondary sources, and use scientific understanding to identify relationships and draw conclusions based on evidence (ACSIS145)
  - Reflect on scientific investigations including evaluating the quality of the data collected, and identifying improvements (ACSIS146)

4 Data Visualisations

What is Data Visualisation?

Data visualisations are a common way of sharing information with a given audience. They are designed to give the audience an understanding of the information being imparted with a quick glance. Data visualisations can be found on TV news, used in advertising, shared on social media, in seminar presentations, or just about anywhere that large amounts of data are being presented that needs to be understood quickly and easily.

Why Visualise Data?

There are several reasons that we might want to present data visually. Visualising data allows us to spread knowledge quickly through social media and advertising. It allows us to decrease the amount of time and effort required to understand the data.

When we create data visualisations, we are looking to show our data in ways that allow the user to see it in a new light. It allows viewers to visually observe patterns, exceptions, and to build the possible stories that the data tells. Data visualisation is a tool for revealing information that is not apparent in the raw data.

Data visualisation is also a way of strengthening the story data is telling. It allows you to draw the viewer's attention to important trends and elements of the data and tie those trends and elements to a course of action that the viewer may pursue based on the evidence.

Creating a Visualisation

There are several steps to consider when selecting an appropriate visualisation for a given dataset. One major consideration is the purpose of your visualisation. What aspect of your data are you trying to present to the viewer? What is the story you want your audience to take away from this data? What questions would you like them to be able to answer once they've viewed your visualisation?

Another important consideration is the audience themselves. Who is your target audience? What level of numeracy skill do they have? Will the visuals need to have more weight than the numbers?

Special features of the dataset you're trying to represent are important parts of this decision-making process as well. Does the data compare changes over time? Does it display relationships and connections between entities? Does it indicate hierarchical structures and part-to whole relationships? Does the data have a geographical component that requires the use of mapping tools?

Once these questions have been answered, they can be used to select an appropriate visualisation type that can be understood by your audience and can present the story you're trying to tell successfully.

Common Types of Data Visualisation

Bar Chart

Bar charts are a very common type of visualisation and represent data through the height or length of a bar. They allow comparison of the sizes of different categories of data. When creating this type of visualisation, it is important to start the values at 0 to allow for a fair comparison of the bar's size.

Use of colour in bar charts can help to draw the viewer's attention to the values of specific categories by highlighting them, as well as using related colours to represent similar categories.

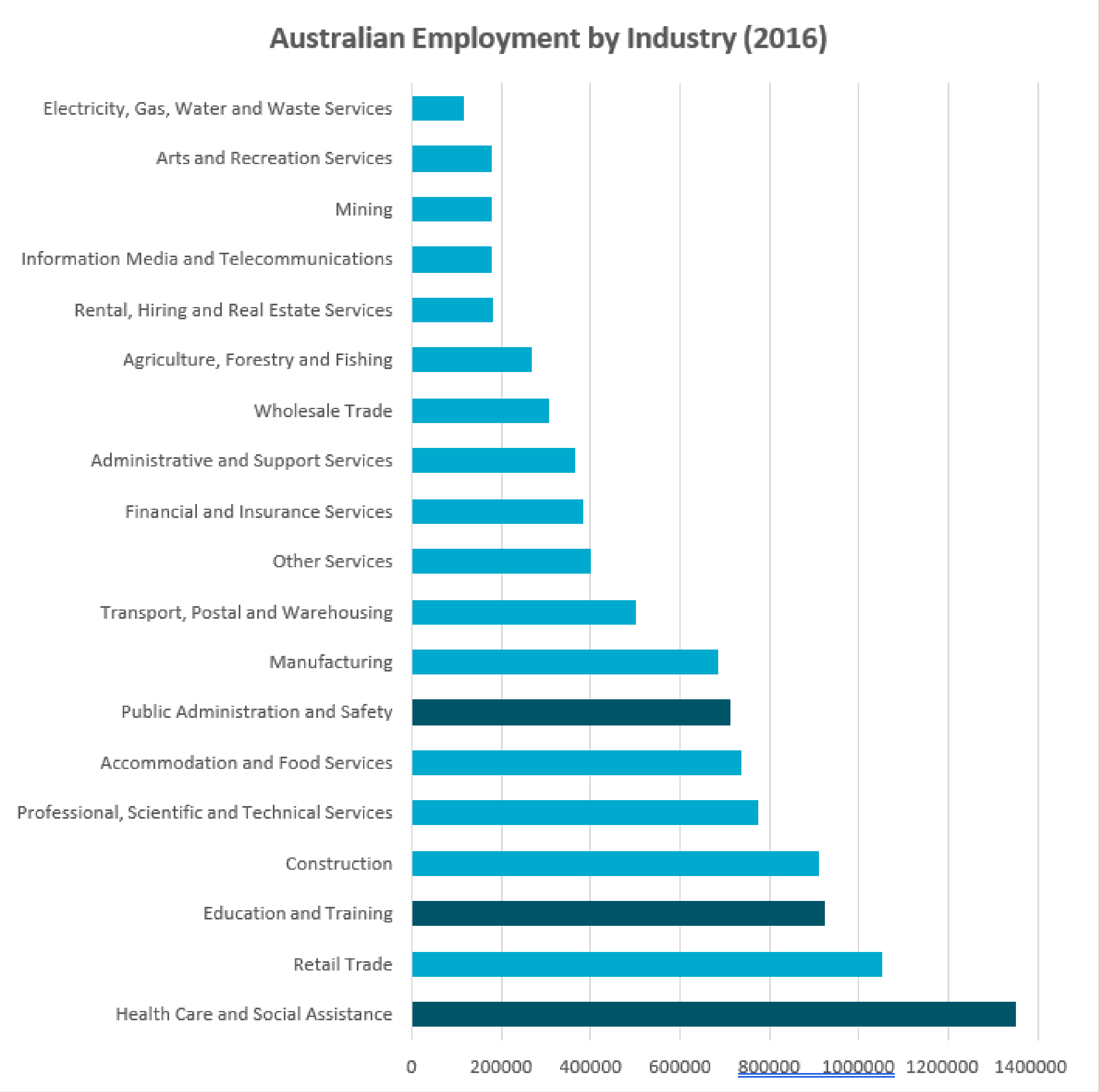


Figure 6 - Bar chart of Australia's workforce grouped by industry in 2016. The length of the bar indicates the numberof people employed in that industry, with darker bars indicating industries that hold high numbers of public service roles. Generated using data from Australian Bureau of Statistics

Dot Plot

Dot plots represent each value using an individual symbol. They group values into categories called bins and show the number of values that fall inside each bin. Colours and different symbols can be used to provide further information. Classroom sticker charts indicating reward progress are often arranged as a dot plot, with names as categories, and stickers indicating a rewardable activity.

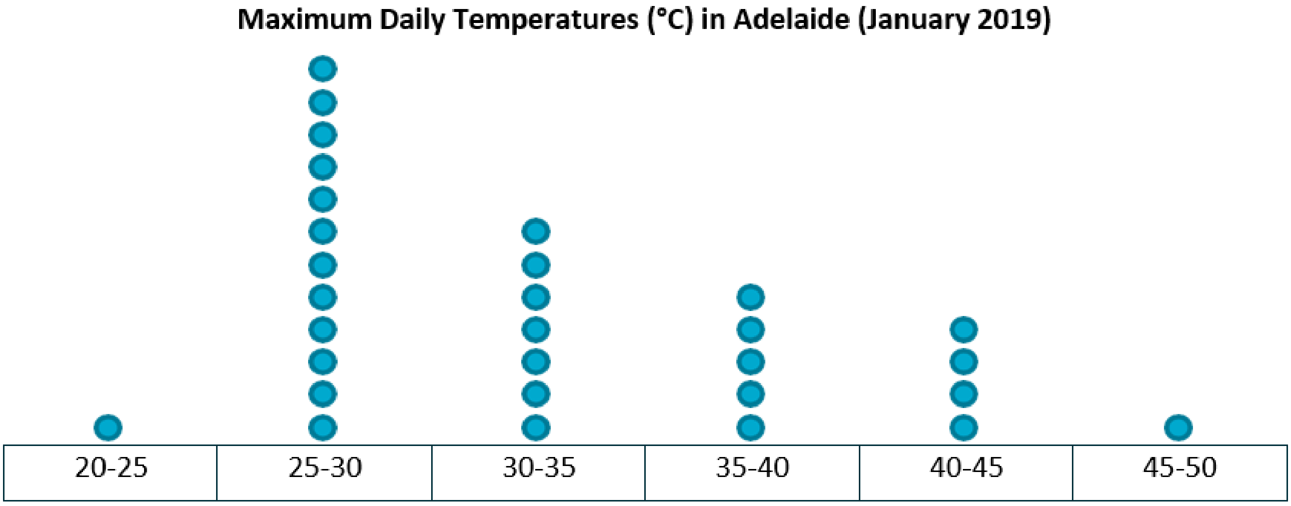


Figure 7 - Dot plot of maximum temperatures for January 2019 in Adelaide. The number of dots above each range indicates the number of days that fell into that range. Generated using data from Australian Bureau of Meteorology

Histogram

Histograms look similar to bar charts but have more in common with dot plots. A histogram groups values into categories, or bins. Bar height indicates the number of values that fall inside that bin. Histograms do not have gaps between the bars, unlike bar graphs.

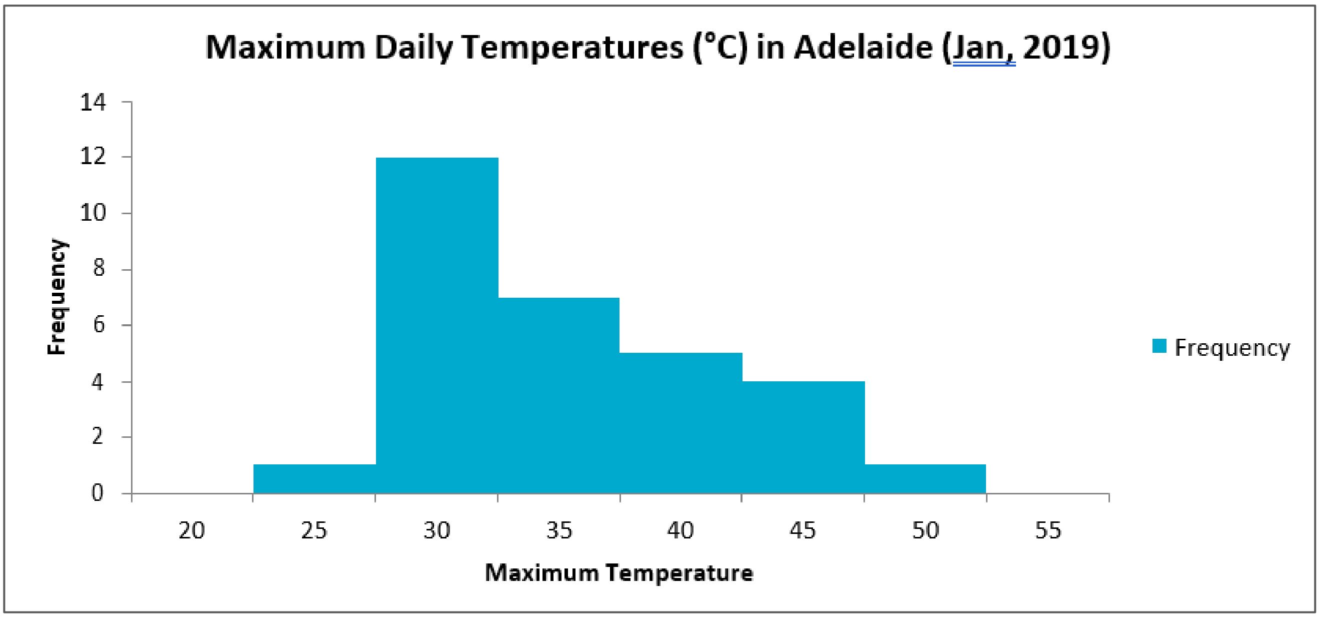


Figure 8 - Histogram of maximum temperatures for January 2019 in Adelaide. The height of each bar indicates the number of days that the maximum temperature was between the number at the bottom of the bar, and the number at the bottom of the previous bar. Generated using data from Australian Bureau of Meteorology

Pie Chart

Pie charts are a good way to demonstrate how something is split into parts. They are good for comparing relative size of values, especially when those differences can be made obvious at a glance. Where two or more of the values have only very slight differences, or there are more than 6 categories being represented, pie charts can become very difficult to read.

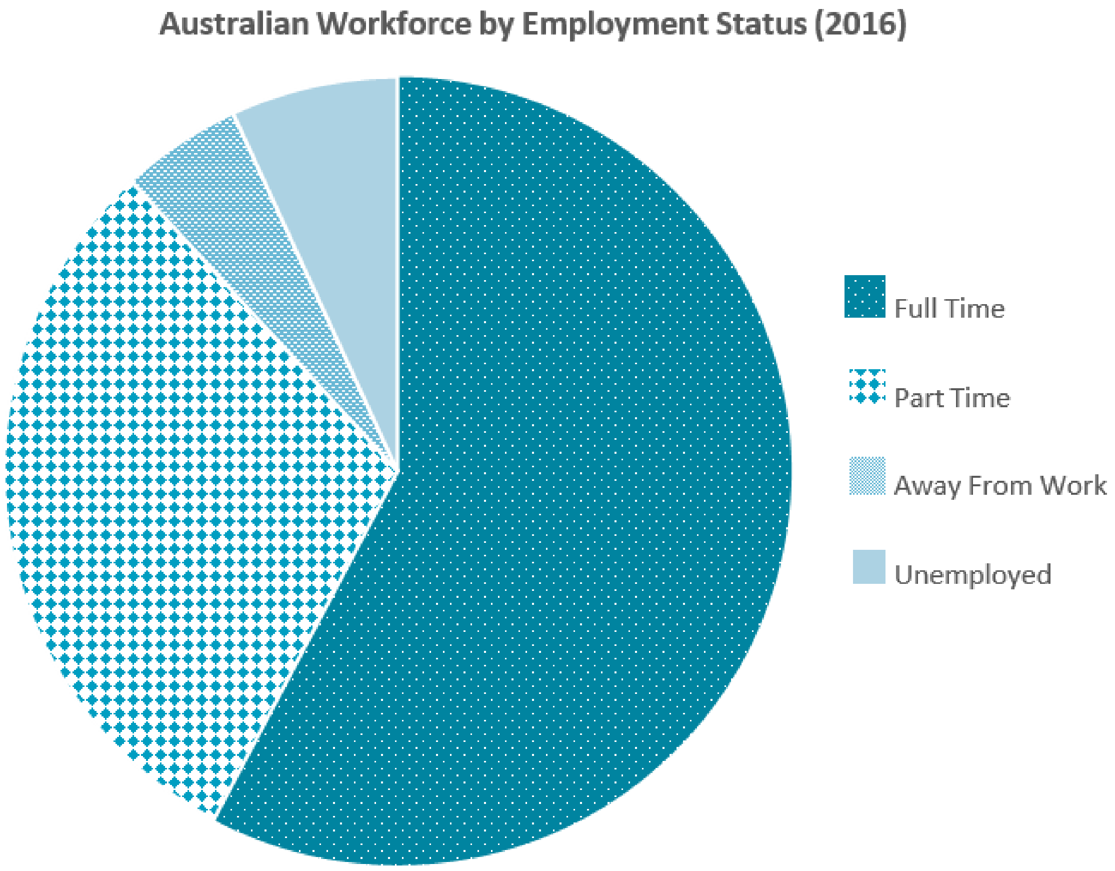


Figure 9 - Pie chart of Australia's workforce grouped by type of employment in 2016. Generated using data from Australian Bureau of Statistics

Tree Map

Tree maps use nested shapes (usually rectangles) to display data that is structured into a hierarchy. Examples of this might be sales data, organised by regions, individual offices and then individual staff members, where the size of the rectangle is the amount of sales. Tree maps are often used to represent government spending, with the highest level of rectangle representing the whole budget and smaller rectangles breaking that sum down into categories of spending.

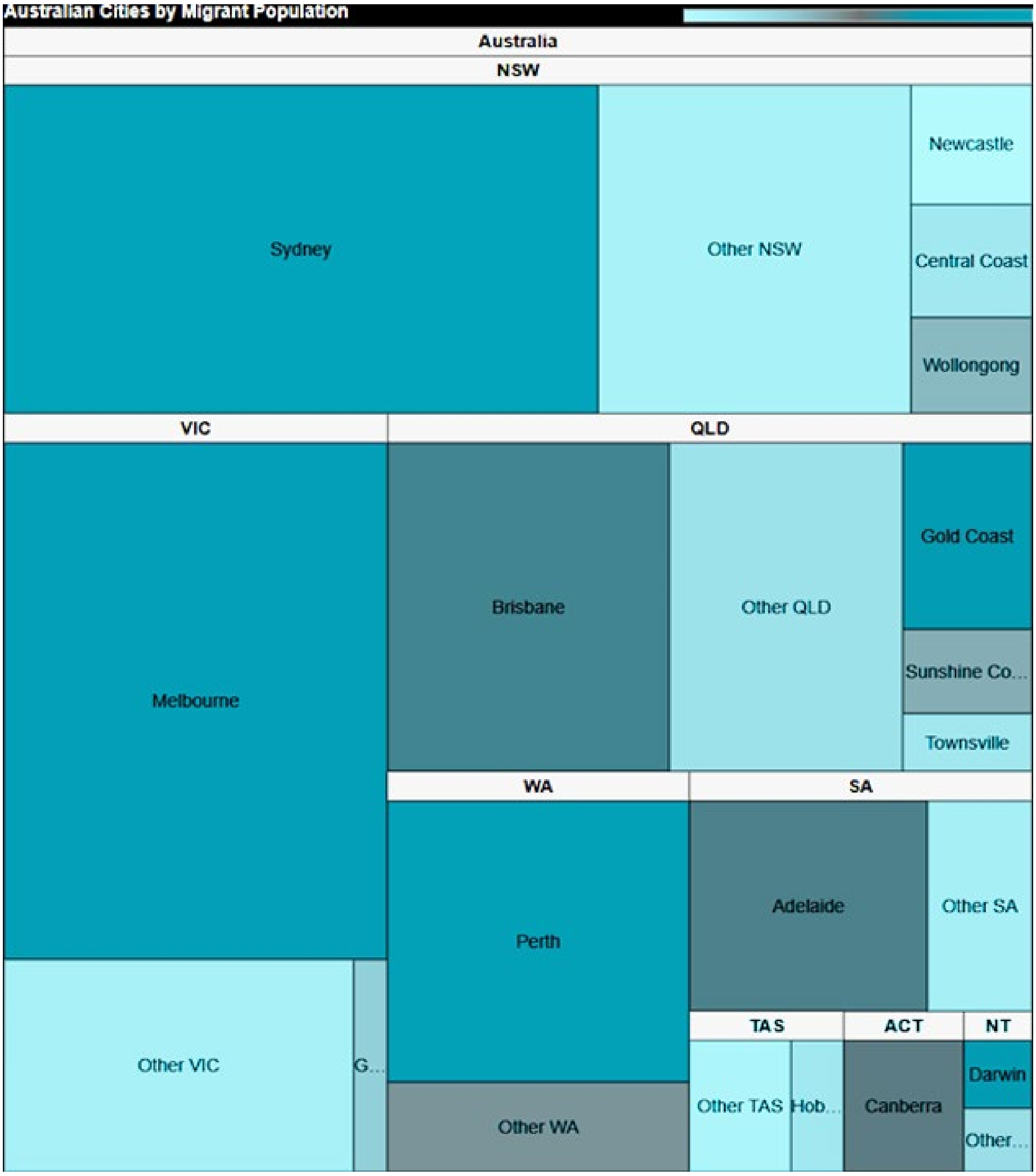


Figure 10 - Australia’s population, grouped by major urban areas (Population >100,000). Rectangle size represents population and colour represents percentage of first-generation migrants. Yellow indicates a high percentage while green indicates a low percentage. Placement of rectangles can make elements difficult to compare, such as the population of NSW and Victoria. Generated using data from Australian Bureau of Statistics

Line Chart

Line charts display data as a series of data points, joined by lines. They are useful for presenting narratives about the transitions between two data points or identifying trends over a given period.

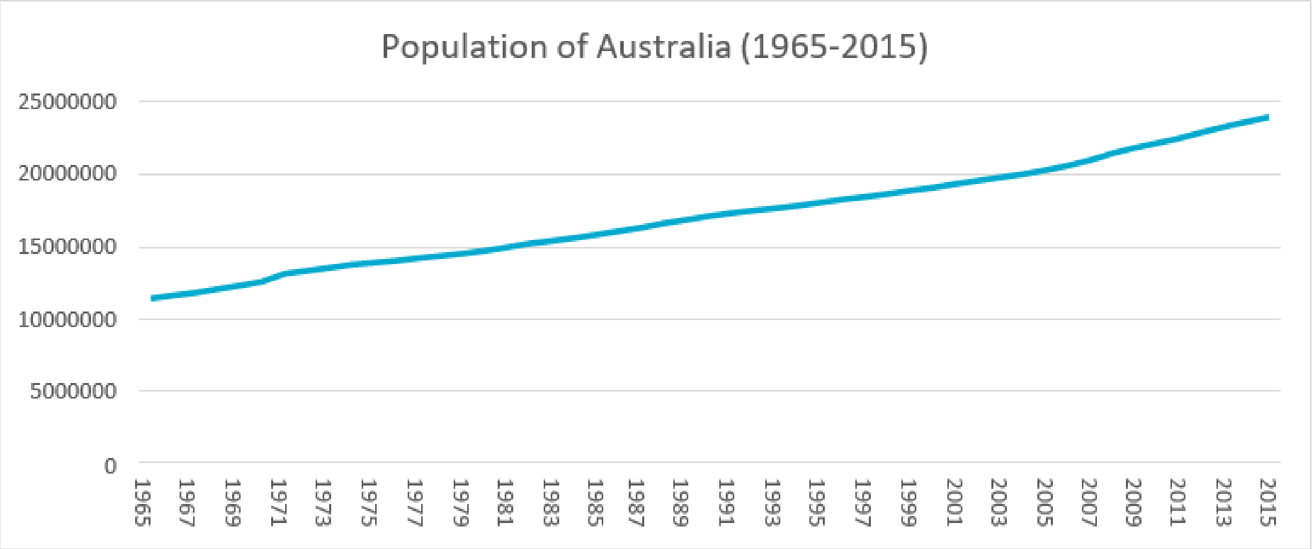


Figure 11 - Australia's population, from 1965 to 2015. Displaying the data this way makes it easy to see the continuous trend of growth over that time. Generated using data from Australian Bureau of Statistics

Area Chart

Area charts are similar to line charts but colour the area under the line. Area charts are often used to compare related sets of data, as changes in the width of the different coloured areas can allow for very quick comparisons.

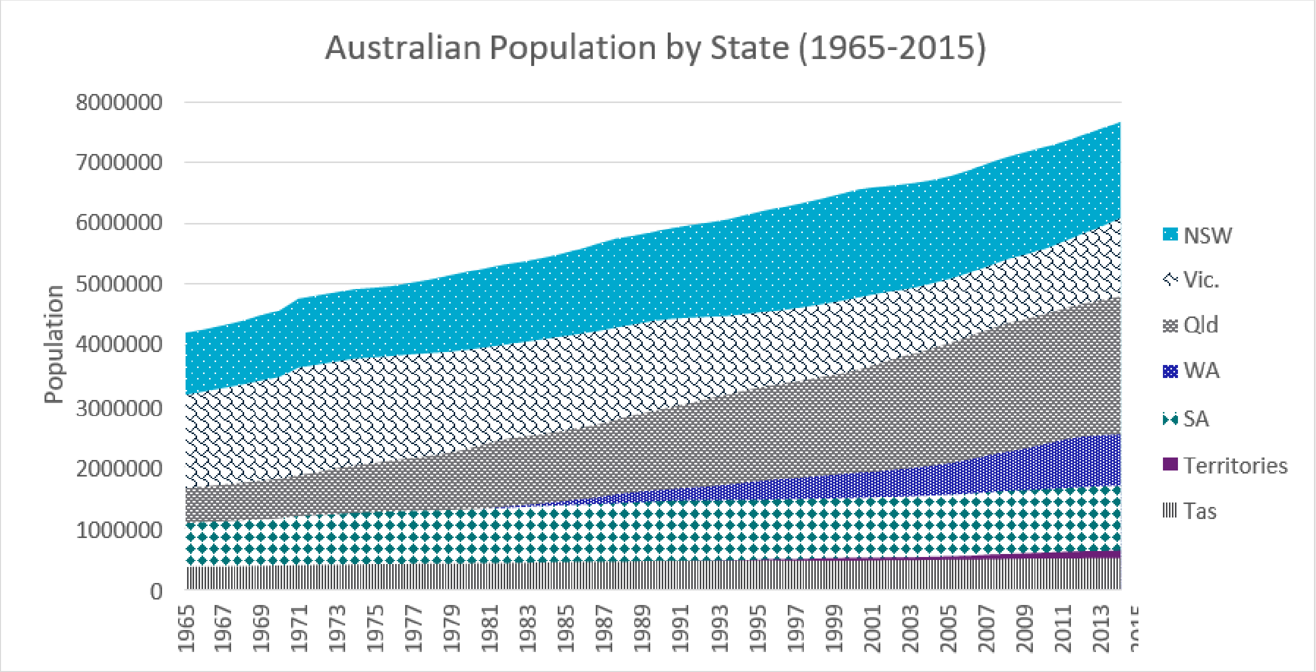


Figure 12 - Australia's population grouped by state, from 1965 to 2015. Each colour represents a different state's data. Note the absence of WA data before 1982. Before 1982, WA had a smaller population than SA and is hidden behind SA's values. This is a limitation of an area chart. Generated using data from Australian Bureau of Statistics

Box Plot

Box plots indicate the median, upper and lower quartiles, minimum and maximum of a given dataset. Box plots are also often called Box and Whisker Plots. The box encompasses the Interquartile range and contains 50% of the data points in the set, with a line inside the box indicating the median value. There are several methods for the display of the whiskers in a box plot. One particularly common one is to show the entire range using the arms. The other common method is to have the whiskers extend to the maximum and minimum values that fall within 1.5 times the Interquartile range of the upper and lower quartile. When using this method, data points that fall outside this range are usually marked using a dot.

This method of visualisation allows comparison of similar collections of data to look at the ways in which they vary. Datasets which at first glance have a similar median and range may have a very different interquartile range, which becomes apparent when visualised this way.

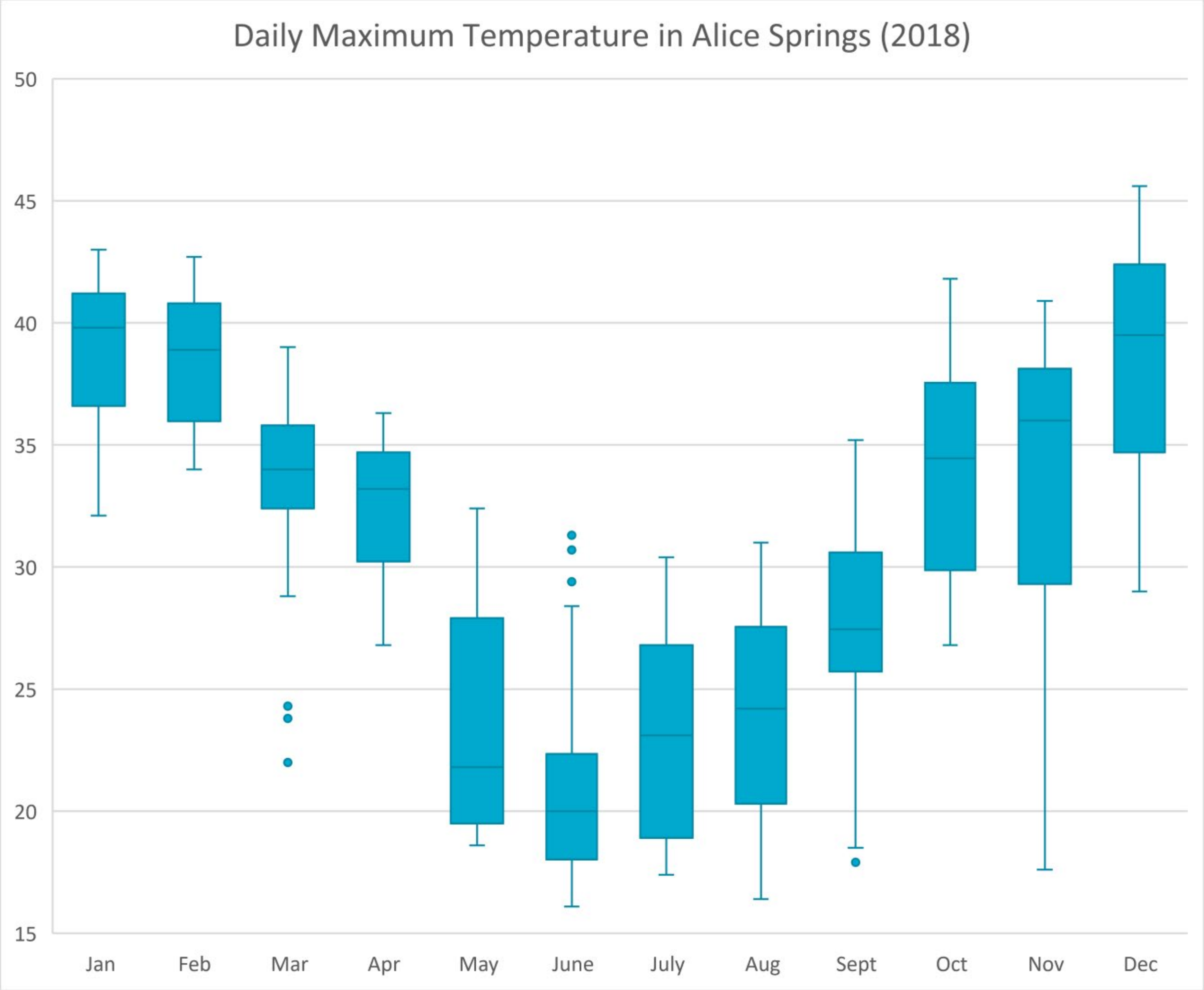


Figure 13 - Daily maximum temperatures in Alice Springs during 2018. Examining only the median, Dec and Jan appear to be very similar months, but with this visualisation we can see that December had a much higher temperature range, with far higher and lower maximum daily temperatures than January. Note that March, June and September have values that fall outside of 1.5IQR and are charted independently. Generated using data from Australian Bureau of Meteorology

Scatter Plot

A scatter plot uses two different variables on the x and y axes to show patterns of correlation, clustering and outliers. It is a very important chart type when becoming familiar with a dataset, as it makes several trends apparent, allowing for further investigation. Scatter plots are often used when assessing student results and comparing results across multiple subjects or focus areas. Scatter plots can use coloured points to differentiate between categories of data.

Scatter plots are excellent when assessing datasets that contain a large number of individual data points.

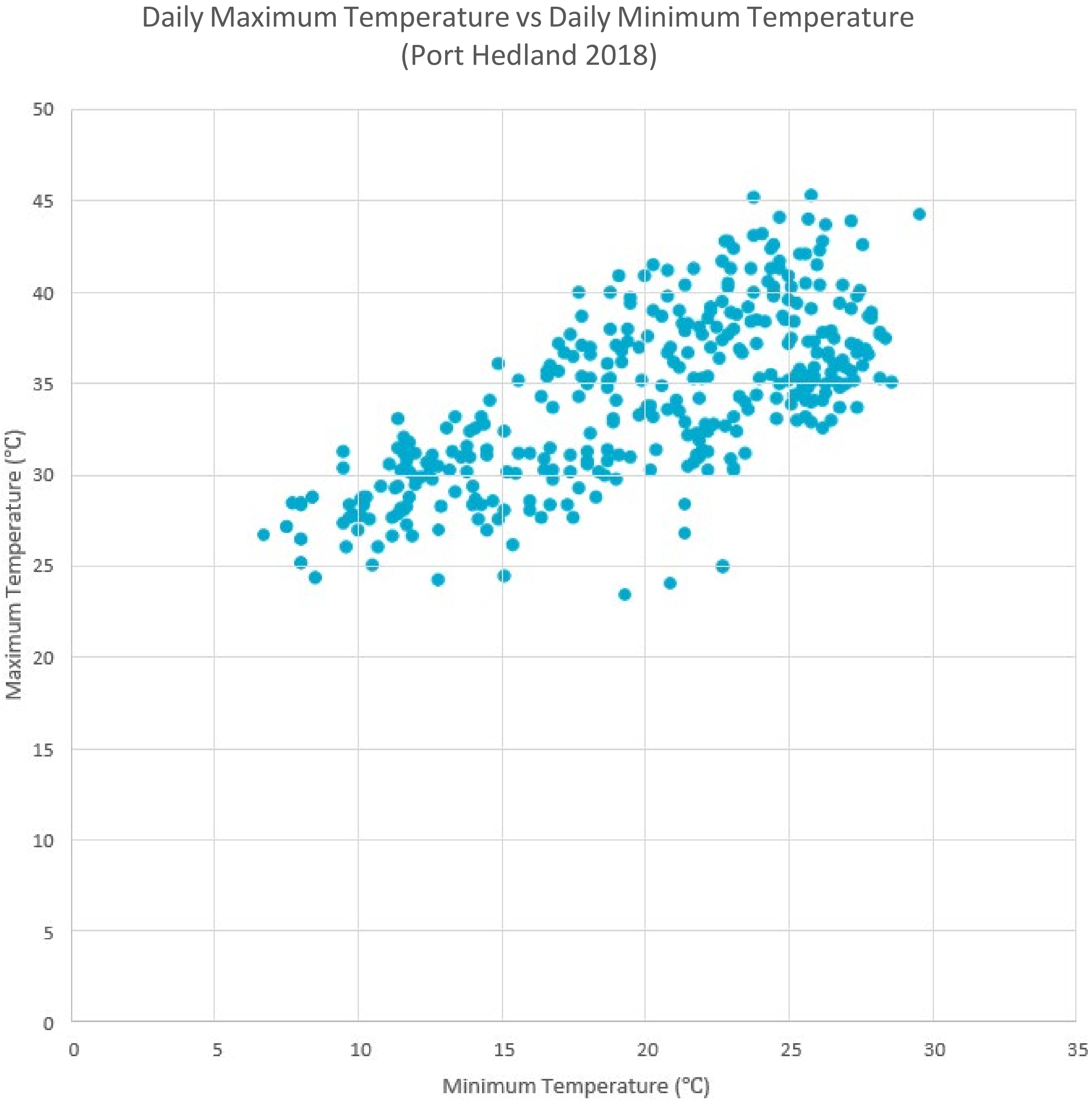


Figure 14 - Maximum and minimum daily temperatures in Port Hedland during 2018. With this scatter plot, we can see a clear cluster of values, particularly around (12, 30) and (26, 35). Using this visualisation, a trend can be identified that links daily maximum and minimum temperatures. Generated using data from Australian Bureau of Meteorology

Bubble Plot

A bubble plot is very similar to a scatter plot in that it plots points against two different variables on the x and y axis, but it adds an additional dimension to date in area. The size of the plotted point is used to represent an additional variable. Like a scatter plot, bubble plots can be used to quickly identify trends and can use colour to differentiate categories of data. Bubble plots are notable for being able to carry a large amount of data in a single visualisation but can often be difficult to interpret if they become too cluttered, or points that should be assessed by size are too distant from each other.

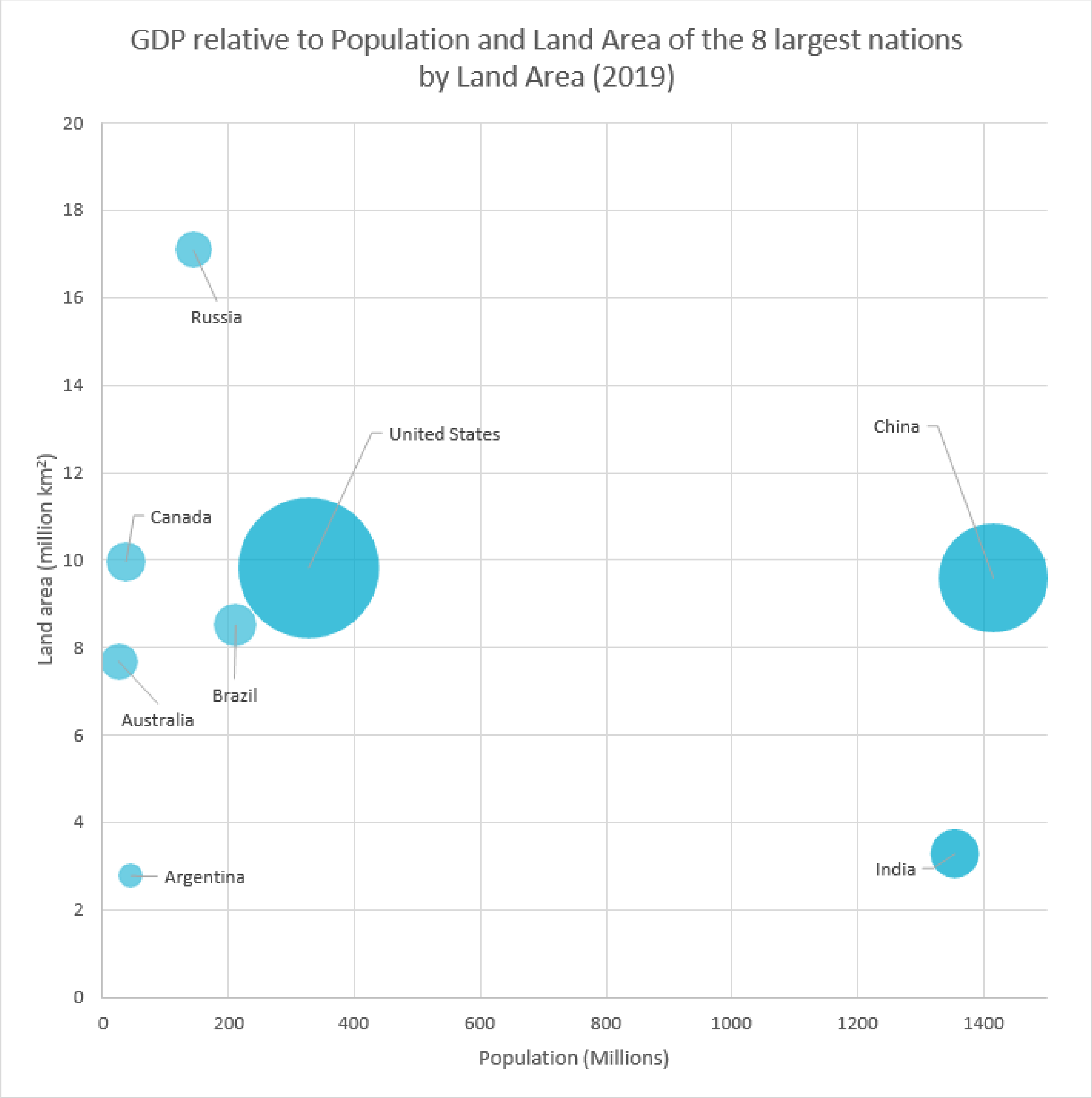


Figure 15 - Land areas of the eight largest nations (In millions of km^2) plotted against their population (in millions). The size of each bubble indicates the nation's Gross Domestic Product. Generated using data from UN Data

Choropleth Maps

Choropleth maps are used for geographical data, colouring sections of the map to represent different aspects of the data. This style of data visualisation is commonly used in reporting and visualising election data, with each electorate coloured to indicate which party has won the seat. These maps can often use gradients to indicate data with a range of values, or alternately, data that is in question. An example of this is using a choropleth map to indicate pre-election polling data. Seats that have a clear expected outcome are coloured blue or red for the appropriate party, while those in question are coloured in purple, with varying mixes of red and blue indicating the likely results.

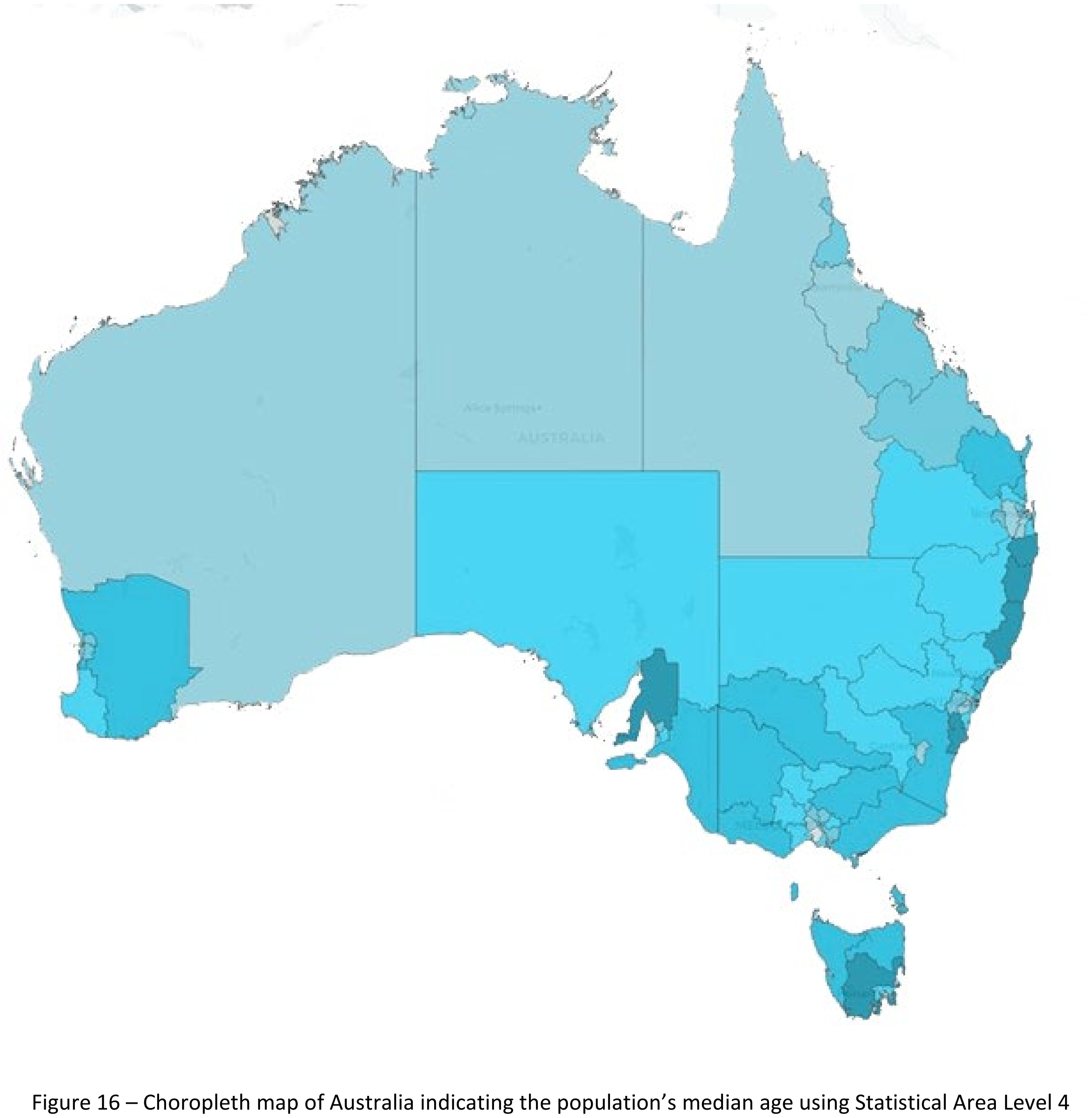
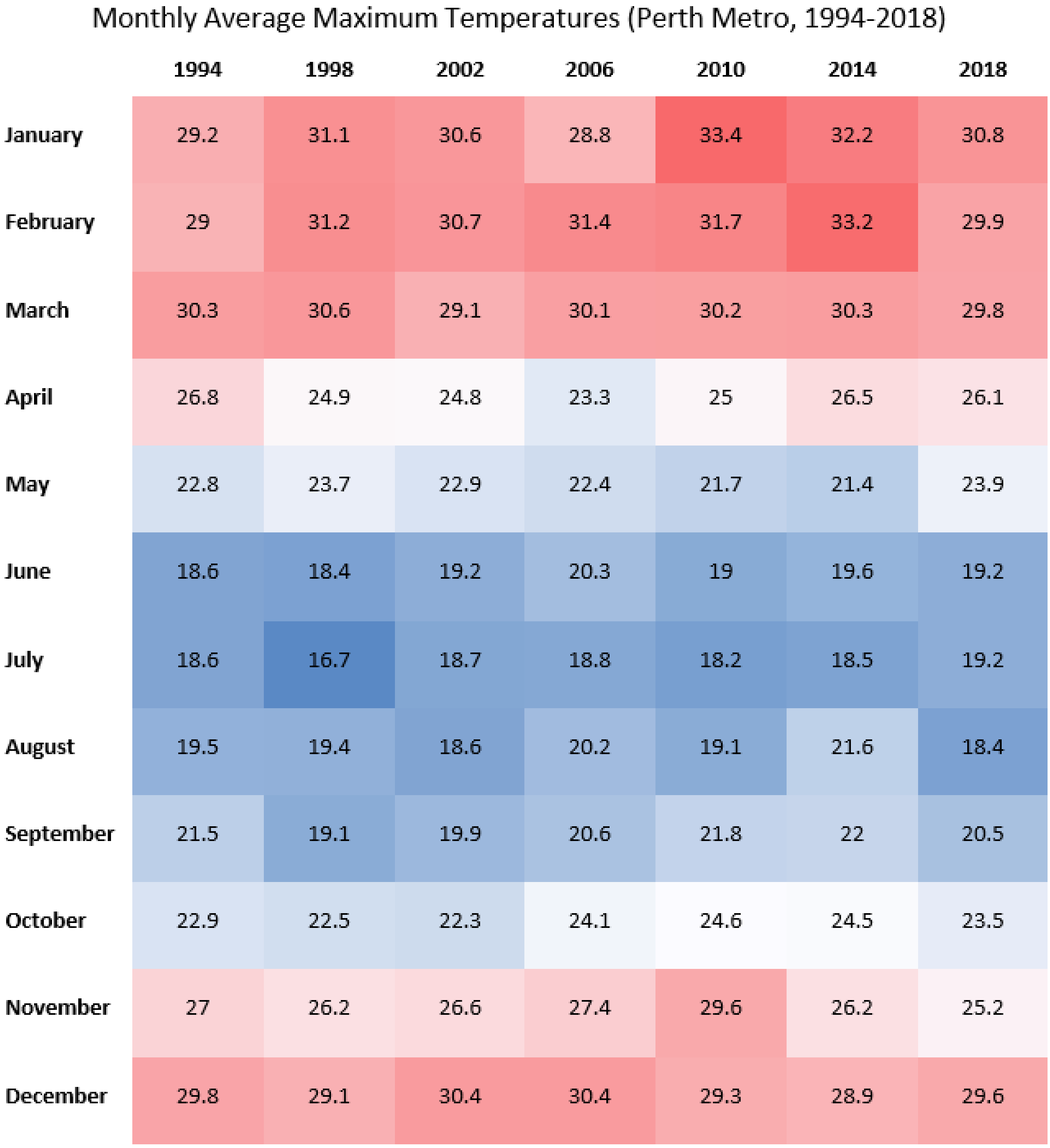


Figure 16 – Choropleth map of Australia indicating the population's median age using Statistical Area Level 4 boundaries. Regions with darker colouration indicate a higher median age. Map created using National Map with age data from Australian Bureau of Statistics

Heat Map

Heat maps use colour to indicate high and low datapoints and are commonly used to interpret large datasets, particularly datasets with two or more dimensions. Use of colour draws the eye to key clusters of data. Heatmaps can include the numerical values of the data, but this is not always the case. Basic heatmaps can be created using conditional formatting in spreadsheet packages.

Heat maps can be overlaid on other visualisations such as geographical maps to indicate key areas, such as high traffic or population density, or on websites to show key areas of the user interface.



Monthly Average Maximum Temperatures (Perth Metro, 1994-2018)

	1994	1998	2002	2006	2010	2014	2018
January	29.2	31.1	30.6	28.8	33.4	32.2	30.8
February	29	31.2	30.7	31.4	31.7	33.2	29.9
March	30.3	30.6	29.1	30.1	30.2	30.3	29.8
April	26.8	24.9	24.8	23.3	25	26.5	26.1
May	22.8	23.7	22.9	22.4	21.7	21.4	23.9
June	18.6	18.4	19.2	20.3	19	19.6	19.2
July	18.6	16.7	18.7	18.8	18.2	18.5	19.2
August	19.5	19.4	18.6	20.2	19.1	21.6	18.4
September	21.5	19.1	19.9	20.6	21.8	22	20.5
October	22.9	22.5	22.3	24.1	24.6	24.5	23.5
November	27	26.2	26.6	27.4	29.6	26.2	25.2
December	29.8	29.1	30.4	30.4	29.3	28.9	29.6

Figure 17 - Monthly average maximum temperatures for the Perth Metro weather station between 1994 and 2018. Colours vary from dark blue (coldest) to dark red (hottest), with gradients indicating values lying between. Summer and winter months can be identified quickly with this visualisation. This data has undergone sampling, taking the data from every fourth year. Generated using data from Australian Bureau of Meteorology

Sankey Diagram

Sankey diagrams use a series of bands to represent data. The width of each band indicates the size of the variable. They were developed to indicate flows within an engine and are generally used to indicate energy transfers within a system. Sankey diagrams can also be used to represent two linked concepts with several different linkage weights, but they can very easily become overwhelming for the viewer with a high number of links between concepts.

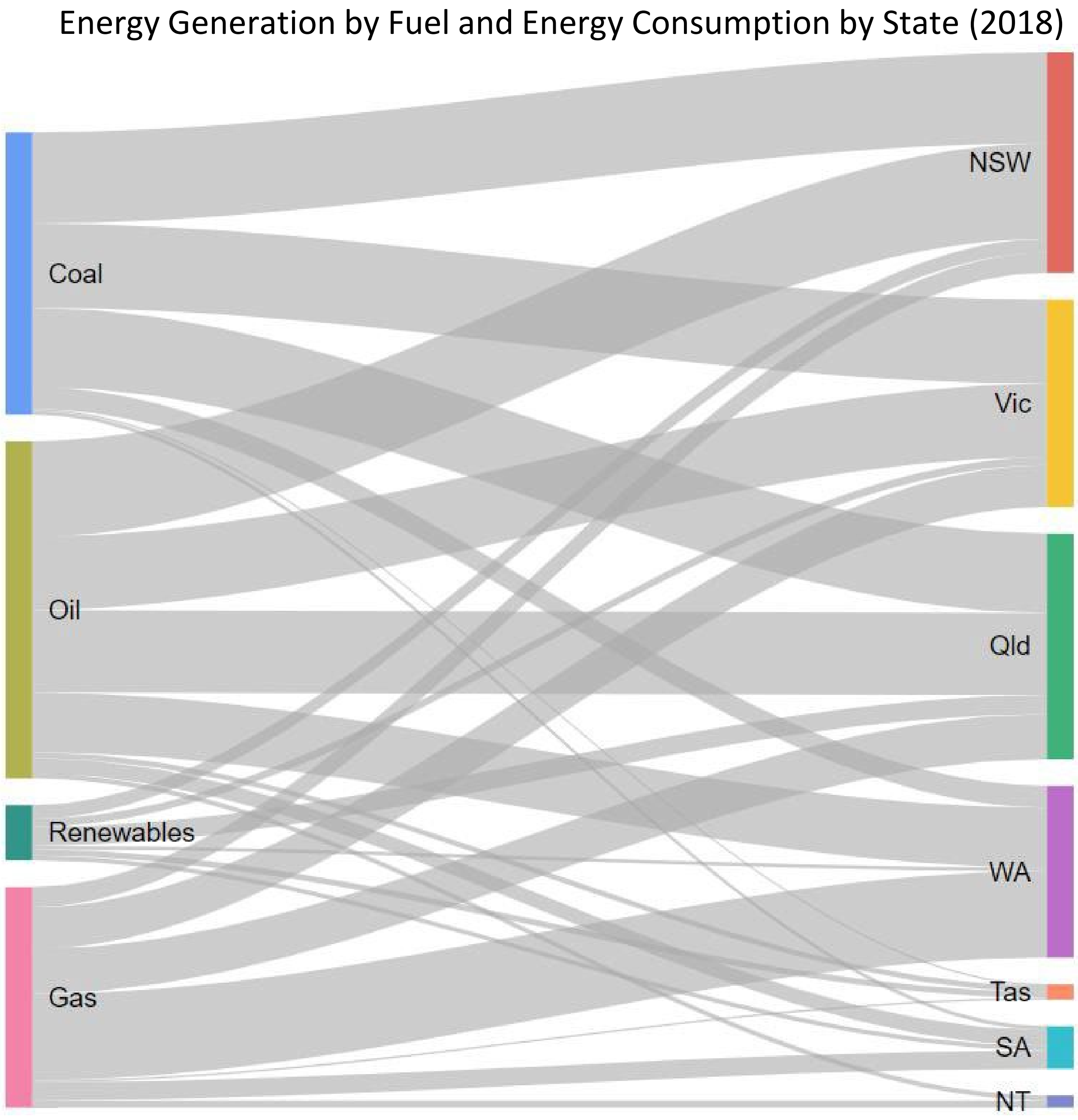


Figure 18 - Sankey diagram linking the consumption of energy in each state and territory with fuels used for generation. In the original dataset, the ACT is included in NSW data. The weight of the lines connecting the states and fuel sources indicates the amount of energy generated using that fuel source that was consumed by that state. Generated using data from Australian Department of Energy and the Environment

Software Resources for Creating Visualisations:

Spreadsheet Software – All spreadsheet packages come pre-loaded with a selection of charts and graphs that can be created using entered data.

National Map – Map of Australia with open data sets available and the capacity to add location-based datasets

Google Charts – Web-based tools for developing interactive visualisations to embed in websites. Tutorials are available, along with sandbox mode to create visualisations without creating code from scratch.

Gapminder Tools – Web-based tools that use global public datasets to allow creation visualisations comparing nations on a wide range of variables. Multiple visualisation types are available, but the easiest way to use your own data is to use the offline tools.

5 Glossary of Data Terms

Correlation: A relationship between the change in two sets of data where one changes relative to changes in the other. If two variables are positively correlated, when the value of one increases the other will also increase. If they are negatively correlated, when the value of one increases the other will decrease. Correlation is measured using r where a r of 1 indicates a perfect positive correlation, 0 indicates no relationship and -1 indicates a perfect negative correlation.

Data: A collection of facts, measurements and statistics generally used to represent a real-world object or system using numbers. Can be analysed to support decision making.

Data Point: A single piece of data, also referred to as a datum.

Data Science: The use of mathematical and computational processes to analyse data and examine trends and patterns in large datasets

Data Structure: A method of arranging and organising data, such as a table, or a list.

Data Visualisation: Representing data in a visual manner, using charts, maps, diagrams or pictures to better allow viewers to quickly gather information from a dataset and to better enable an audience to understand notable trends indicated by analysis.

Dataset: A collection of similar pieces of data that can be examined and analysed as a single unit.

Discrete Variable: A variable whose value can only be one of a set list of values. Often seen in survey questions, where a rating is given as a whole number between 1 and 5.

Information: Data given meaning through analysis or organisation and the application of context.

Outliers: Values that fall outside of the overall pattern created by the dataset. Some outliers can be explained by anomalies in recording or storage, while others are legitimate results that fall outside of expected parameters.

Qualitative Variable: Describing a quality using categories and labels. Describing a person's hair by listing the colour is an example of a qualitative variable.

Quantitative Variable: Describing a quality using numerical values. Describing a person's hair by listing the length is an example of a quantitative variable.

Variable: A logical manner of grouping data, that collects all the data points that refer to a specific attribute. When storing data on species of animal, 'number of legs' might be a variable that stores a single number for each species recorded.

As Australia’s national science agency and innovation catalyst,
CSIRO is solving the greatest challenges through innovative science and technology.
CSIRO. Unlocking a better future for everyone.

Contact us
1300 363 400
+61 3 9545 2176
csiroenquiries@csiro.au csiro.au

For further information
CSIRO Education and Outreach
1300 136 376
+61 2 9490 5588
education@csiro.au csiro.au/education

Cover Image Credit
CSIRO Marine Research