



## AUSTRALIAN DATA SCIENCE EDUCATION INSTITUTE

### Introduction

The solar energy installations dataset shows how many solar systems were installed, in each postcode, from 2001 - 2016. It is a great way to understand how to explore and characterise datasets, as well as explore the use of data in the media, as the media reporting of this dataset makes a classic mistake, doing a simple analysis that gives a misleading conclusion.

This is a beautifully rich dataset that offers a range of options for exploration.

### Links

<http://www.cleanenergyregulator.gov.au/RET/Forms-and-resources/Postcode-data-for-small-scale-installations>

### Engager

#### Activities

Open the csv file in a spreadsheet. Look at the postcodes. In most spreadsheets this data will show 1, 3, and 4 digit postcodes, when postcodes in Australia are all 4 digits. What has happened to these postcodes?

- This is an example of your spreadsheet hiding things it thinks you don't need to know about – in this case, leading 0s. Mathematically speaking, there's no difference between 0, 00, 000, and 0000. They all just mean 0. So spreadsheets (and other software) tend to remove the leading 0s, which means postcode 0 is actually 0000, 200 is actually 0200, etc.
- This is your first lesson in data cleaning. Sometimes you have to convert the data back to its original form, to fix errors that spreadsheets and other software introduce in an attempt to be “helpful”.

This work belongs to [adsei.org](http://adsei.org) and is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Now let's look at the first two solar columns. The first is historical installations from 2001 to 2016. We don't seem to have any data from before 2001, but that's not because nobody was installing solar before that. It turns out that it's because 2001 is when the government introduced the mandatory renewable energy target and began tracking renewable energy.

Next question: how much solar is actually operating now? Answer? We don't know. This data tracks installations. It doesn't track people getting rid of their solar panels, or the panels ceasing to work. Installations are a reasonable measure of how much solar we have, but not perfect.

This opens the way for a great conversation about the data we want, versus the data we have, and how many data studies work with flawed or missing data, simply because it's all we have available.

Ok, so let's look at the first column. Having it sorted by postcode is logical, but not terribly interesting. Let's look at the top 20 postcodes – to do that, we can sort the entire table by the second column (how many installations happened between 2001 and 2016), in descending order. In other words, put the largest values up the top.

1	Small Unit Installation Postcode	Previous Years (2001-2016) - Installations Quantity	Previous Years (2001-2016) - SGU Rated Output In kW	Jan 20
2	4670	10,598	34,298.313	
3	6210	10,101	25,929.687	
4	4655	9,346	27,713.172	
5	4551	8,611	23,956.862	
6	4350	7,766	25,187.085	
7	6065	7,174	22,991.619	
8	4211	6,990	23,588.384	
9	4305	6,773	19,862.104	
10	4740	6,443	26,016.072	
11	4207	6,399	20,911.231	
12	6155	6,341	18,627.154	
13	4570	6,248	20,379.991	
14	3029	6,203	18,688.427	
15	3977	6,146	18,643.949	
16	6164	5,986	17,725.288	
17	4556	5,899	17,801.277	
18	4306	5,806	19,665.363	
19	6112	5,785	17,822.82	
20	4510	5,740	18,658.14	

A quick glance shows us that the majority of the top 20 postcodes start with a 4, meaning they're in Queensland. (If you're not sure which postcode is where, as I'm not, you can check at a [postcode site](#).) The top postcode, 4670 covers 53 regions, including Bundaberg. There's a surprisingly large gap between the top postcodes and the bottom of the top twenty, which is interesting. Most of the postcodes in this list that aren't in Queensland are in Western Australia. except for 3029, which is West of Melbourne, around Hoppers Crossing, and 3977, which is South East of Melbourne, in the Cranbourne area.

There's a rich conversation to be had around why these suburbs have so much more solar than other places in Victoria. Toorak, for example, a notoriously wealthy suburb, comes in at 1701 on the list. My suspicion is that areas with a lot of new housing are more likely to have solar, as it gets put in when the house is built as a way to increase the energy rating of the house. But this is a topic worth exploring! You don't have to know all the answers, as it's an opportunity for the kids to research and explore, and come up with their own theories for why it might be the case.

Let's look at column 4: solar installations in January 2017. How different are the top 20 if you sort the whole table by this column?

1	Small Unit Installation Postcode	Previous Years	Previous Years	Jan 2017 - Installations Quantity
2	6065	7,174	22,991.619	99
3	3977	6,146	18,643.949	82
4	6210	10,101	25,929.687	77
5	6164	5,986	17,725.288	72
6	4211	6,990	23,588.384	71
7	4670	10,598	34,298.313	67
8	6112	5,785	17,822.82	65
9	6000	220	1,187.91	59
10	4655	9,346	27,713.172	55
11	6155	6,341	18,627.154	55
12	4510	5,740	18,658.14	52
13	4300	5,563	17,935.325	49
14	6069	4,296	13,534.08	46
15	4560	4,544	13,567.41	45
16	2259	4,215	12,031.366	45
17	6031	1,671	5,367.85	44
18	4680	5,472	19,982.779	42
19	6169	4,546	11,636.551	42
20	6107	3,725	10,233.827	42

Now WA scores better, and the rest is still largely over to Queensland, except for one Victorian postcode (Cranbourne area again), and this time one NSW representative.

Why do WA and Queensland do so well on both historic and recent measures? This is an opportunity to explore the politics and have your students find out what incentives there are to install solar in those states. Could it be due to solar feed in tariffs, government incentives, or home energy rating requirements?

This work belongs to [adsei.org](http://adsei.org) and is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

This is where the mistake the media made comes in - and I also made it the first time I looked at this data. Despite WA and Queensland dominating the top 20 postcodes list, when you calculate the average for each state, you get a quite different picture.

For examples of the media take:

<https://www.domain.com.au/news/australias-top-10-solarfriendly-postcodes-revealed-20180502-h0zjp0/>

<https://www.smh.com.au/environment/sustainability/queensland-city-tops-national-solar-panel-list-20180615-p4zlni.html>

“ Queensland is leading Australia’s rooftop solar boom with eight of the country’s top 10 postcodes for installations in the Sunshine State, according to the new Clean Energy Australia 2018 report.”

### **What happens if you use average stats for each state?**

You can do that in Excel or any other spreadsheeting package by sorting the data by postcode, and then just copying and pasting each state into a separate sheet, but it’s also nice and easy in Python. (I’ve been lazy and lumped the ACT in with NSW.)

```
average solar installations by state between 2001 and 2016
QLD 170.03
WA 113.0
Vic 212.13
NSW 246.05
TAS 202.39
NT 91.5
```

Interestingly this shows that the state that dominates the top 20 doesn’t perform as well when you average over all of its postcodes, so there is another rich conversation to be had about different ways of ranking data outcomes, and how you can characterise data in accurate but misleading ways. The media reported this data saying that Queensland and WA were the best for solar. But although they had the highest ranking postcodes, as a state they ranked very low.

You can keep going and explore the different columns, or you could step it up a notch and start to look at how the columns are related. For example, are postcodes with a lot of historical solar installations also likely to have a lot of recent ones? You can do that roughly by eye, simply by looking at whether the top twenty when sorted by those two columns is similar or very different, or you can use the correlation function to find out whether the columns are correlated, or even go heavy on the statistics and try to work out whether both values are equally predictive of a postcode’s place in the ranking. (I won’t go into that here, lest I scare away the non-statto’s among us!)

You can use this dataset to explore different attitudes to solar power around the country, and the possible reasons for them. You can use it to question which incentives work and which ones fall flat, or whether solar incentives actually make a difference.

Now, what if you wanted to visualise this data? Well, you could find out the names of the top 10 postcode areas and graph them. (You could just graph the postcodes but it's not terrible meaningful to anyone who hasn't memorised the postcodes of Australia!) Top 10 is a fairly arbitrary selection, aimed at not putting too many places into the one graph. It would make more sense to choose a place in the data where there's a big drop from one value to the next. In this case I might go top 5, since there's a big drop from 5 to 6. It shows you the top performers well, but doesn't show you much else.

Another technique would be to colour a map by number of solar installations. Say, bright red for >9000, and becoming paler for each drop of 1000. This would be rather time consuming given that there are 2795 postcodes listed, so this is an opportunity to consider aggregating your data. It's a great example of not needing complex technical skills to explore a dataset. Being able to program unlocks more ways of looking at the data, but to get started all you need here is the ability to sort a spreadsheet by different columns, and a wealth of information is at your fingertips!

We will publish more datasets and more explorations as we go along, but in the meantime why not find your own datasets, and explore the things it can tell you? There are no right or wrong answers in this game, just different ways to play with the data. The more you play, the greater your data literacy.