A Minimal Book Example

John Doe

2023-06-07

Contents

Chapter 1

Acknowledgements

Chapter 2

Introduction

Chapter 3

Strucure

_ Problem statement - Obtaining the data - EDA - Modelling - Conclusions

APPENDICES - HOW we waht the data obtained - Detailed EDa - TEchnical note on buulding packages - vignettes

Chapter 4

Data

4.1 Data Sources

The first in the process was to source demographic and electoral data, which has been provided from two sources:

- **The Australian Electoral Commission** (AEC) [Commission, 2023a] . The AEC contains detailed online records for every federal election held in the 21st century, through their Tally Room website [Commission, 2023b].
- The **Australian Bureau of Statistics** (ABS) [of Statistics, 2023a]. The ABS provides a wide number of national statistics and is responsible to conduct a national census of population and housing every 5 years. Comprehensive census data is provided in multiple formats, including csv files through Census Data Packs [of Statistics, 2023b], which are available for censuses from 2006 onwards.

Both organisations are the authoritative source for electoral and statistical data in Australia, and the data is provided openly. Although there are no quality issues, the way that data is provided presents other challenges, namely:

- In both cases, data are provided in large volumes and exhaustive granularity. If not done effectively, data extraction and aggregation can be time-consuming and resource intensive.
- Census data points are provided using the ABS own geographical standard - and only a small selection of census data is provided at the electoral division level. Conversion between ABS geographical structures and electoral divisions is not straightforward as there is no 1:1 correspondence. Both geographical systems change from election to election and census to census.
- Despite the best efforts of both organisations in keeping consistency, names of electorates, parties, and census attributes change over time - to compare

similar statistics manual mapping is necessary.

To address these issues and ensure repeatability, three R packages have been written to undertake this task:

- **{auspol}** [**Yáñez Santibáñez, 2023b**], which extracts and presents electoral results.
- **{auscensus}** [**Yáñez Santibáñez, 2023a**], which allows to interact with Census Data Packs to extract different statistics across geographical units, and across censuses.
- **{aussiemaps}** [**Yáñez Santibáñez, 2023c**], which assists with aggregating census data into electoral divisions, by matching and apportioning different geographical structures.

The appendix contains a vignette for each package, explaining their respective *modus operandi*. At a higher level, the extraction pipeline for this project is represented by figure 4.1.

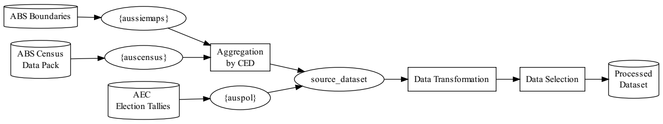


Figure 4.1: Flow of data from sources to dataset

In summary, the process followed consisted of the below steps:

1. Census data was extracted from the respective Census Data Pack using **{auscensus}**. Using the package workflow, key attributes were identified in each census, extracted from the respective files and given common names. Data were extracted for statistical areas and apportioned into Commonwealth Electoral Divisions by overlapping area, with the help of functions written into **{aussiemaps}**

2. Primary vote results for each division were extracted using the **auspol** package.
3. All the data was stored in a local database, from where was extracted and put together in a single dataset.
4. From there, the "raw" data was further processed and stored in a single "consolidated" dataset.

4.2 Data Selection

4.2.1 Census and Election Years

The first to address when extracting the data is to establish a correspondence between census and election data. Since election the census cycle (5 years) does not match the electoral cycle (determined by the incumbent government, with a 3-year term for the House of Representatives), there is a potential problem of the census data not being completely representative of the population on a given election day. Figure 4.2 presents the best matches between both events held in the 21st century.

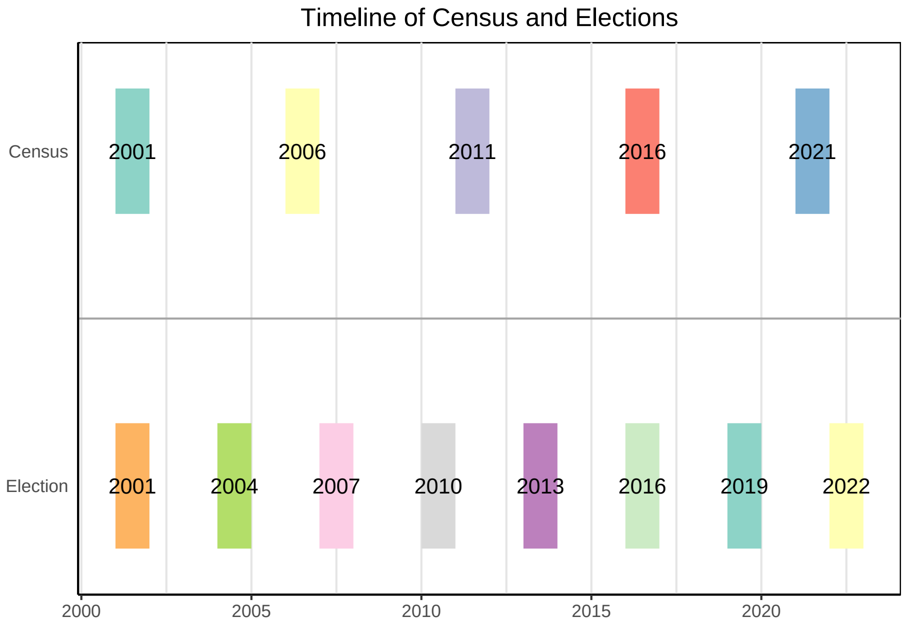
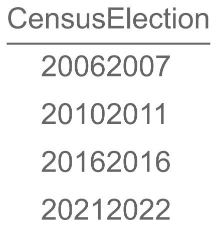


Figure 4.2: Timeline of elections and censuses in the 21st Century

Considering the census data available and selecting the elections closer to each census, four pairs of events were selected for data extraction. there are presented in table 4.1

Table 4.1: Selected Census-Election pairings



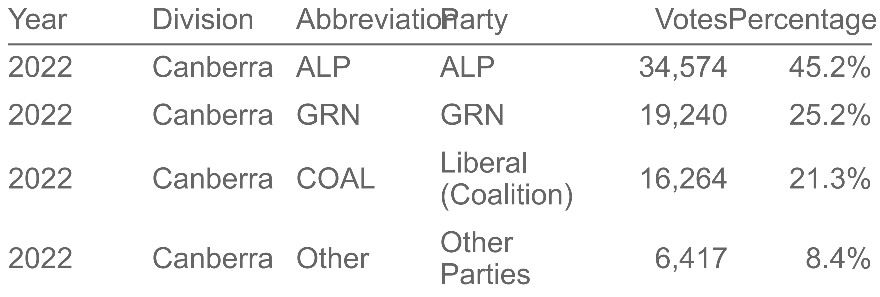
Census	Election
2006	2007
2010	2011
2016	2016
2021	2022

Please note that this selection will remove half of the elections within the period, which may have an effect on model accuracy. However, since the objective is not to obtain an accurate prediction this has been accepted as a trade-off to avoid having to interpolate demographic attributes between censuses - which is also subject to inaccuracies given the rapid demographic changes experienced in Australia's main cities.

4.2.2 Electoral Data

In the case of the electoral data. not much processing was required. The source data already contains records of primary voting for each electorate and only percentages have been calculated. In addition, the number of total votes per party at the national and state level have been calculated. A sample of the extracted data is presented in table 4.2.

Table 4.2: Sample extraction - Canberra 2022



Year	Division	Abbreviation	Party	Votes	Percentage
2022	Canberra	ALP	ALP	34,574	45.2%
2022	Canberra	GRN	GRN	19,240	25.2%
2022	Canberra	COAL	Liberal (Coalition)	16,264	21.3%
2022	Canberra	Other	Other Parties	6,417	8.4%

4.2.3 Census Data

As mentioned in section 4.1, a major challenge with respect of census data is the large volume of data points collected. For instance, the data pack for the

2022 Census contains 62 different tables, ranging from 8 [1] to 1,590 [2] attributes.

To select which variables to extract, literature and journalistic sources were consulted ([Biddle and McAllister, 2022], [Parliament], [Jakubowicz and Ho]) to inform an initial set of covariates. In total (XYZ) variables were selected, which correspond to below to the following groups:

1. **Income** : Distribution of population in pre-set income brackets.
2. **Education Level**: Distribution of educational achievement (from incomplete secondary to vocational education and academic degrees).
3. **Age**: Distribution of the population in generational cohorts. Taking into account the selected elections, the four groups of interest are Baby Boomers (1946 to 1964), Generation X (1965 to 1980), Generation Y (1981 to 1996) and Generation Z (1997 to 2021).
4. **Relationship status**: Variables describing civil status (e.g. living alone, married, in a de facto relationship).
5. **Household type:** Descriptors of type of housing , (e.g. standalone house, semi-detached, flats).
6. **Household tenure:** Descriptors of house ownership, rental or other arrangement (e.g. public housing).
7. **Country of Birth:** Selection of the XX most popular countries of birth of the Australian population ()
8. **Citizenship**: Percentage of population that
9. **Religion:**

[1] *02 -Selected Medians and Averages*

[2] *09 - Country of Birth of Person by Age by Sex*

Chapter 5

(APPENDIX) {auscensus} Vignette

Extracted from https://carlosyanez.github.io/auscensus/articles/complex_case.html on Sunday 22 January 2023

This vignette shows a more complex use case of auscensus. Let's assume we want to extract the percentage of Australian Citizens for all Commonwealth Electoral Divisions, as measured in last 4 Censuses (2006-2021).

An initial exploration shows that this data can be found in table 01 (across all four censuses) - which provided an statistical summary. However, is not published aggregated by electorate across all censuses.

Therefore, we will retrieve the data from the lowest statistical unit. However, SA1 were not available in 2006 - where the smallest area was a "CD".

The next step is to figure the attributes for the numbers of Australian citizen and total population, which are presented below:

Using *attribute_tibble_to_list*, this data frame can be converted into the required format.

Now, we can cycle through the four censuses and extract the data. Please note that CDs and SA1s are not equivalent, but they are stored together for convenience:

To aggregate the data, **aussiemaps::geo_aggregate()** can help using area to apportion on non-overalpping cases. Then, this package's *calculate_percentage()* will take the totals from the list and calculate percentages.

Chapter 6

(APPENDIX) {auspol} Vignette

Extracted from https://carlosyanez.github.io/auspol/articles/house_primary_vote.html on Sunday 22 January 2023

auspol includes two functions to interact with the preference distribution data:

- get_house_primary_vote()
- house_primary_vote_summary()
- house_primary_comparison_plot()
- house_primary_historic_plot()

6.1 What is this?

If you are unfamiliar with the Australian electoral system and preferential voting, please look at this [explainer(https://www.aec.gov.au/learn/preferential-voting.html) before proceeding.

6.2 Getting the data

get_house_primary_vote() is the basic function to retrieve primary vote data published by the AEC. Without any arguments, it will deliver all the results for all elections, but it comes with parameters to facilitate filtering. For instance, to get the results for Brisbane for 2022:

Both parameters can include more than one value, e.g.

By default, the results are presented by polling place, with the possibility to aggregate them.

It is also possible to restrict the results to selected polling places

Additionally, it is possible to select one or more states instead of a group of divisions, e.g.:

It is also possible to filter results by one or more parties:

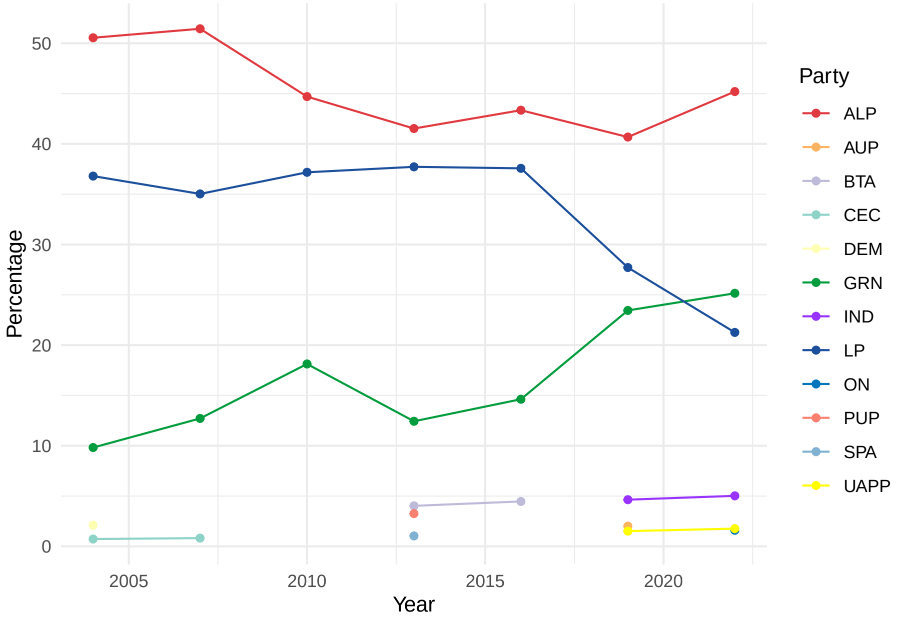
house_primary_vote_summary() builds on the basic function and summarises data .

Using the previous filters, it is possible to get ad-hoc summaries, for instance - all the ALP votes in Queensland in 2022, or the historic Liberal vote in Franklin.

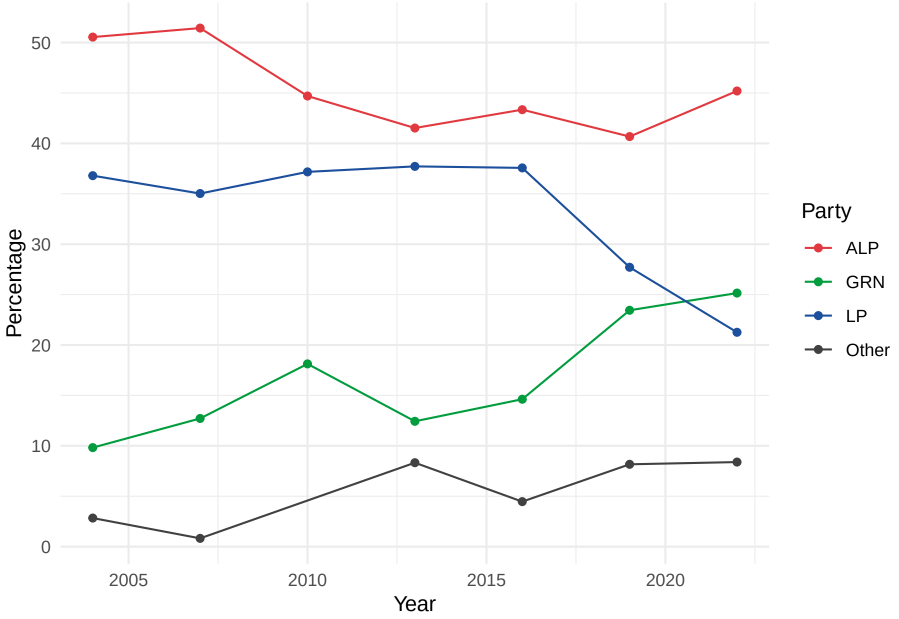
6.3 Plotting

6.3.1 Historic Trends

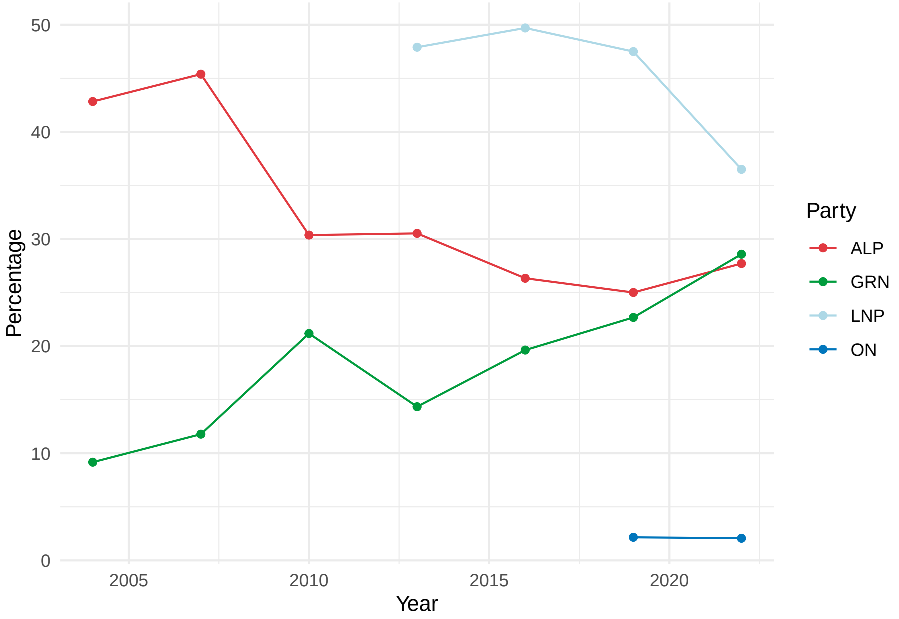
The first plotting convenience function in this package allows comparing the evolution of primary voting across time. This function relies on house_primary_summary and uses many of its options. Its first use is to represent party trends in one electorate:



As they can be many minor parties, it is sometimes useful just to focus on a number of parties. This function allows filtering by a number of parties or by filtering by the most voted in a certain year. In both cases, it is possible to consolidate others' votes.

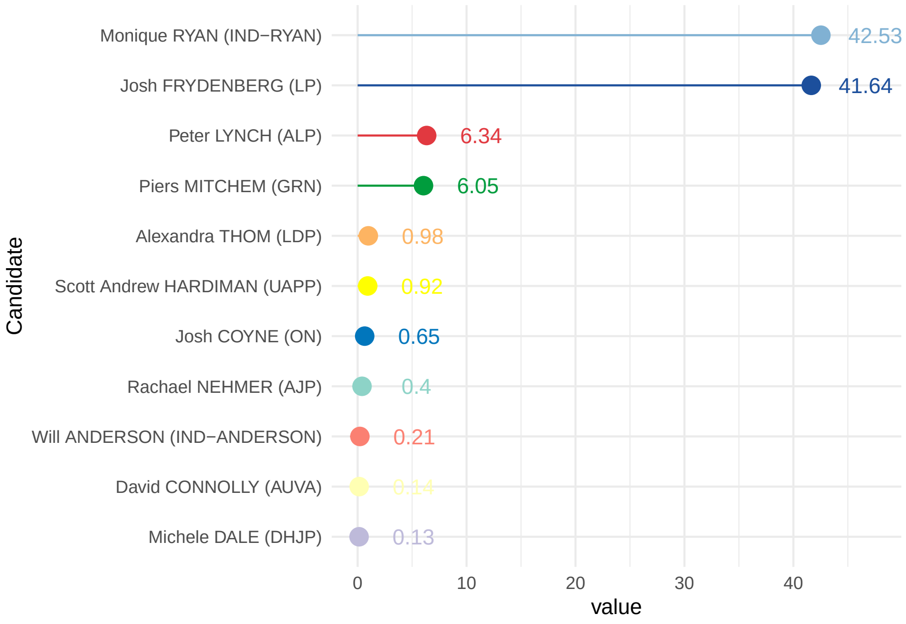


Finally, it is possible to aggregate party acronyms - sometimes the same party has changed named or registered differently

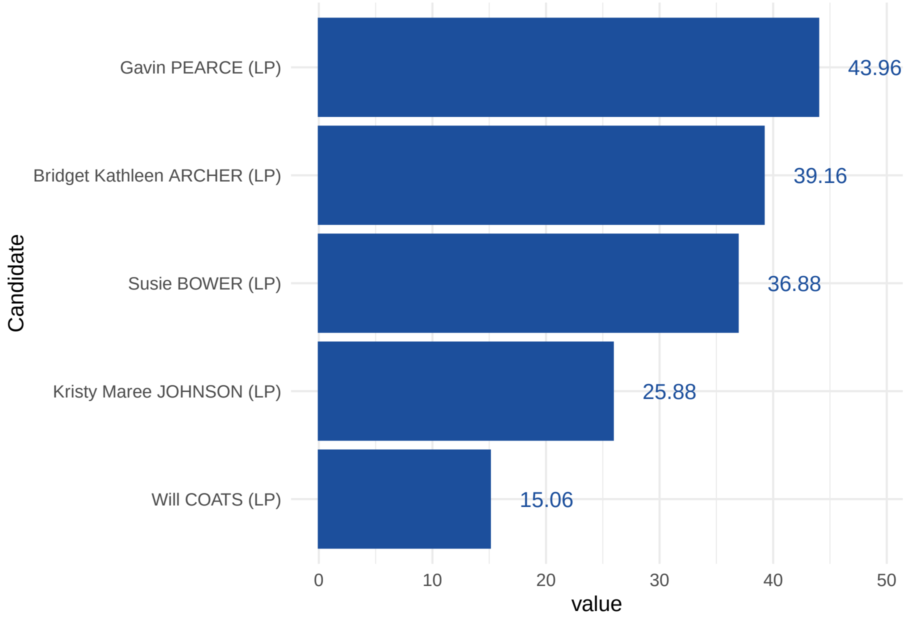


6.4 Results for one election

This package also contains a convenience function to look at the primary vote results for one division. Lile the previous function, this also inherits many of the attributes of *get_house_primary_vote*.



The plots can also be displayed using bars, as shown below


Gavin PEARCE (LP)
Bridget Kathleen ARCHER (LP)
Susie BOWER (LP)
Kristy Maree JOHNSON (LP)
Will COATS (LP)
Candidate
43.96
39.16
36.88
25.88
15.06
0
10
20
30
40
50
value

Chapter 7

(APPENDIX) {aussiemaps} Vignette

Extracted from https://carlosyanez.github.io/aussiemaps/articles/aussiemaps.html on Sunday 22 January 2023

7.1 {aussiemaps} - Yet another maps package

This package has been built to facilitate the use of the geographic boundary files published by the Australian Bureau of Statistics (ABS). The ABS has published several boundary files - i.e. the Australian Statistical Geography Standard (ASGS) from 2006 onwards and the Australian Standard Geographical Classification (ASGC) before that - covering both:

- Statistical Geographic Structures created and maintained by the ABS - and used to collect data.
- Non-ABS structure, e.g Postal Areas, Electoral Divisions, LGA boundaries.

This package has four versions of the above, aligned with Census years 2006, 2011,2016 and 2021. This makes it easy to mix use with Census data packs or the {auscensus} package.

This package provides access to a processed version of those boundaries - as sf objects, allowing it to cater for the following scenarios:

- Get the boundaries of an electoral division across time.
- Get all the S1 or S1 areas within a Council area.
- Get all postcodes in a state or territory.

This repository also contains the R script used to process the files. Although not tested, the functions could also accommodate BYO structures for other years.

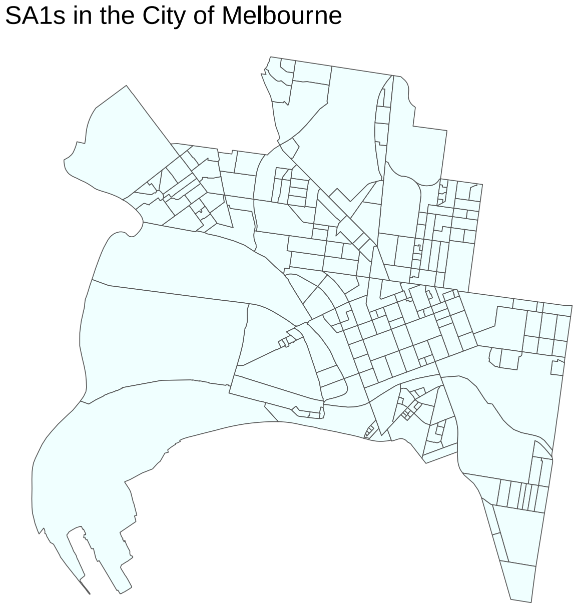
7.2 Getting started.

The core function of this package is get_map(), which retrieves the sf files. get_map provides several filters to narrow down the data retrieved and avoid getting everything unless is needed. The key parameters for this function are:

- How the data will be filtered (e.g. return only objects in a particular state, council or metro area)
- Which year/version of the data will be retrieved?
- Which aggregation will be used (e.g. which will be the resulting objects)

Filters and column names follow the same name convention used in the original ABS files. The function list_attributes(), will present them in tibble format:

Let's say we want to retrieve all SA1 in the City of Melbourne for 2016 - this can be done via:



7.3 Filtering via regular expressions

The filter arguments are intended to be regular expressions, for instance:

```
## Simple feature collection with 8 features and 3 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 115.6286 ymin: -41.3658 xmax: 152.0004 ymax: -20.34465
## Geodetic CRS:  GDA94
##                SSC_NAME_2016                       UCL_NAME_2016     STE_NAME_2016
## 1                  Prestons                                  Sydney   New South Wales POLYGON ((1
## 2  Preston (Toowoomba - Qld)  Remainder of State/Territory (Qld)        Queensland POLYGON ((1
## 3 Preston (Whitsunday - Qld)  Remainder of State/Territory (Qld)        Queensland POLYGON ((1
## 4             Preston (Tas.) Remainder of State/Territory (Tas.)          Tasmania POLYGON ((1
## 5              South Preston Remainder of State/Territory (Tas.)          Tasmania POLYGON ((1
## 6              Preston Beach   Remainder of State/Territory (WA) Western Australia POLYGON ((1
## 7         Preston Settlement   Remainder of State/Territory (WA) Western Australia POLYGON ((1
## 8             Preston (Vic.)                               Melbourne          Victoria POLYGON ((1
```

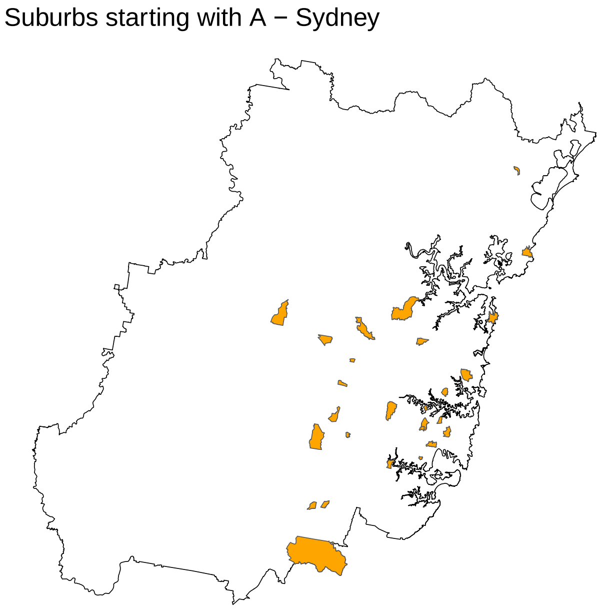
Whereas

```
## Simple feature collection with 3 features and 3 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 146.0066 ymin: -41.33851 xmax: 150.8979 ymax: -33.9263
## Geodetic CRS:  GDA94
##      SSC_NAME_2016                       UCL_NAME_2016   STE_NAME_2016
## 1         Prestons                              Sydney New South Wales POLYGON ((150.8737 -33.
## 2 Preservation Bay Remainder of State/Territory (Tas.)        Tasmania POLYGON ((146.0401 -41.
## 3   Preston (Tas.) Remainder of State/Territory (Tas.)        Tasmania POLYGON ((146.0962 -41.
```

7.4 Even more complex filtering

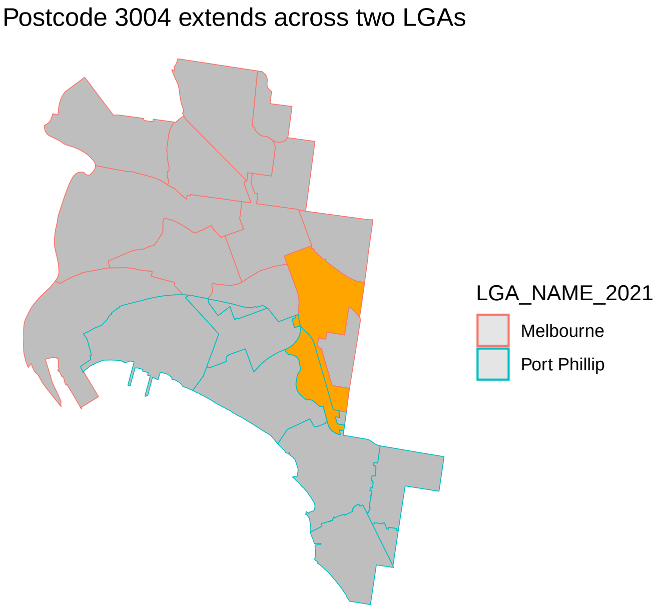
If more complex subsetting is needed, it is possible to pass a table with the elements to be selected. In order to do that, list_structure() comes to help. This function uses the same year and filters parameters than get_map() (actually this function calls the former if no table is provided). Once you have the dataset, you can use any ad-hoc filter to get the needed structures. For example

```
## Reading layer `cache_2021_6766fccc' from data source `C:\Users\carlo\OneDrive\Documents\.aussi
## Simple feature collection with 1 feature and 36 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 149.9719 ymin: -34.33116 xmax: 151.6306 ymax: -32.99606
## Geodetic CRS:  GDA2020
```



7.5 Aggregation

It is worth noticing that the *aggregation* parameter accepts more than one variable. Those parameters are passed to dplyr::group_by() before aggregation - thus more variables will impact how sf objects are aggregated. For instance, if we look at the postal areas (ABS approximation of a postcode) in the cities of Melbourne and Port Phillip:

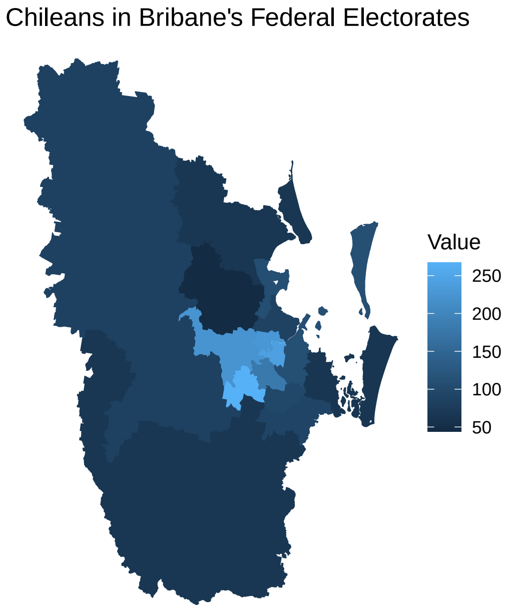


## Using external data

This package provides sf data, thus the result can be easily merged with any other data frame. Since data has been taken from the ABS and the output contains both names and **codes** of geographic structures, data can be joined using an un-ambiguous key. Furthermore, with {auscensus}, this package can be used as data filters to retrieve said data in the first place. For example:

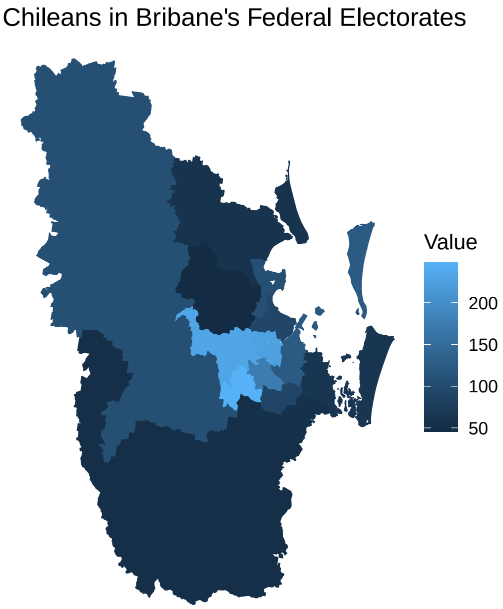
```
## Reading layer `cache_2021_4ec18365' from data source `C:\Users\carlo\OneDrive\Documents\.aussi
## Simple feature collection with 15 features and 36 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 152.0734 ymin: -28.36387 xmax: 153.5467 ymax: -26.45233
## Geodetic CRS:  GDA2020
```

```
##  [1]  85 109  64 228  44  90 241  87  66 180 267 107  96 223  66
```



7.6 Data Aggregation

As a bonus function, *geo_aggregate()* aggregates data, transforming between geographic structures. For instance, let's imagine that for the previous case, it is only possible to get data by SA2. *geo_aggregate()* can aggregate the data to obtain an approximation for each electorate. When an SA1 is not fully contained by an electorate, the function will use the overlapping area as the weighting factor.


Chileans in Bribane's Federal Electorates
Value
200
150
100
50

Bibliography

Nicholas Biddle and Ian McAllister. Explaining the 2022 australian federal election result. Technical report, 06 2022. URL https://apo.org.au/node/318286.

Australian Electoral Commission. Website, 2023a. URL https://www.aec.gov.au/.

Australian Electoral Commission. Tally room archive, 2023b. URL https://results.aec.gov.au/.

Andrew Jakubowicz and Christina Ho. Was there an 'ethnic vote' in the 2019 election and did it make a difference? URL http://theconversation.com/was-there-an-ethnic-vote-in-the-2019-election-and-did-it-make-a-difference-117911.

Australian Bureau of Statistics. Website, 2023a. URL https://abs.gov.au/.

Australian Bureau of Statistics. Census data packa, 2023b. URL https://abs.gov.au/census/find-census-data/datapacks/.

Commonwealth Parliament. Voting patterns by generation. URL https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/FlagPost/2022/April/Voting_patterns_by_generation. Archive Location: Australia Last Modified: 2022-04-29 Publisher: corporateName=Commonwealth Parliament; address=Parliament House, Canberra, ACT, 2600; contact=+61 2 6277 7111.

Carlos Yáñez Santibáñez. *auscensus: Access Australian Census Data (2006-2021)*, 2023a. URL https://carlosyanez.github.io/auscensus/. R package version 0.0.1.0008.

Carlos Yáñez Santibáñez. *auspol: Australian Federal Election Results (2004-2022)*, 2023b. URL https://carlosyanez.github.io/auspol/. R package version 0.0.1.0000.

Carlos Yáñez Santibáñez. *aussiemaps: Maps of Australia*, 2023c. URL https://carlosyanez.github.io/aussiemaps/. R package version 0.2.0.0013.