DATA ANALYTICS GLASGOW

University of Glasgow

School of Mathematics and Statistics

Data Analytics MSc (ODL)

Project

Analysing Australian Electoral Results

Carlos Yanez Santibanez

26 June 2023

To Lauren. Thanks for you constant, never-ending support, encouragement and patience.

To you and Lucas, with whom I have made Australia home.

Contents

Abstract

In politics, it is important for political actors to understand populations are inclined to specific political parties. This is a topic of interest to political movements, the media and political science. According to the latter, factors like age, wealth and education can be used as predictors of vote. However, these axioms have been challenged as changes in the political system (like the rise of populism and new political movements) seem to defy conventional knowledge of the political leanings of the electorates.

This project presents an analysis of how voting patterns in Australia are impacted by demographic factors. In particular, it examines how changes in demographic attributes [1] influence primary voting [2] for federal elections for the Commonwealth's House of Representatives.

To achieve this objective, Census results were reviewed, and key demographic statistics were selected based on a review of the existing literature ([1], [2], [3]) and media narrative ([4], [5], [6], [7]). Code in the form of R packages was written to extract census data and election results systematically and repeatedly. Once data was selected, commonalities across electoral divisions were identified, which were used to classify them into three distinct segments: 1. mostly inner-city areas in large metropolitan areas, 2. the suburban areas in middle, and 3. outer suburbs, and regional and rural constituencies. Then, multivariate elastic net regression models (for each cluster) were fitted to identify which demographic factors influence primary voting for the three largest political groups with national presence (namely the Australian Labor Party, the Coalition and the Australian Greens).

From this composite model, it was possible to establish a link between age, income, religious beliefs, cultural diversity, and primary voting paterrns across all electorates. In particular, it was possible to establish voting in inner-city electorates may be driven by age (younger people voting left, older voting right), income (higher income population favouring both the Greens and the Coalition) and religiosity (non-religious people leaning Green, different religious groups preferring either Labor or the Coalition). Those factors were also found to be relevant in the suburban fringes, where also different culturally diverse communities seem to prefer either Labor or the Coalition. However, in regional areas the same factors seem less relevant, and education levels and higher-income are positively associated with Labor vote.

After fitting this model, a naive retrospective forecast of the 2022 election was attempted. Although the model only has a discrete predicting power (as expected), it proves to be a useful analysis tool to understand the drivers behind changes in voting patterns. Some particular examples are presented, including the results of the so-called "Teal Wave". In those cases it is possible to observe that the Coalition vote was in-line with demographic expectations, concluding that teal MPs weren't elected backed by dissatisfied Coalition voters. Another case presented examines the continuous growth of the Green vote in selected inner-city electorates, driven by generational change and gentrification. Finally, a more diverse set of electorates were also analysed, looking for cases where the selected variables did not necessarily correctly interpret the results. This was deemed useful to understand potential limitations.

To conclude, a series of improvements were suggested to improve predictive and analytical power.

[1] From Census data,

[2] First choices in a preferential voting system.

1 Introduction

The start of the 21st century has come with significant disruption across the world. Although it may sound cliché, the end of the post-war order has come with many challenges: climate change, global recessions, and a the first global pandemic in hundred years. Additionally, improved quality of life has brought both longer life expectancy and decreased birth rates across the world. All these changes have shaken the political systems and conventions of the 20th century. Uncertainty about the future has proven fertile ground for populist movements. Longer lifespans mean that the post-war generation (Baby Boomers) are still active in political life as many of them enter their 8th decade, worrying about keeping the lifestyle as they age. On the other hand "young" Millenials are reaching their forties with uncertain prospects due to the growing of effects climate change, and their perceived inability to reach the same level of wealth as their parents.

In Australia, the last decade of political life has been marked by these issues, which have been contentious points both across the political divide, and within the major parties that have dominated in politics for much of the past century. Political narratives (from the media and politicians) tend to characterise this issues in terms of conflict between different population groups (stoic Baby Boomers vs. Millenials, inner-city "lefties" vs. "real Australia", southern vs. northern states). As many of these perceived issues can be described in demographic terms, it is of interest to analyse how demographic factors may or may not affect political choices.

1.1 A brief introduction to Australia's political system

The Commonwealth of Australia is a federal, constitutional monarchy. At a federal level, the political system consists of multi-party parliamentary democracy, largely based on the Westminster model. Like the UK national Parliament, the lower chamber (House of Representatives) is composed of elected members, each being the sole representative of one geographical area (a Commonwealth Electoral Division or CED, also referred to as an electorate or a division). The number of and geographical extent of each electorate is determined by the Australian Electoral Commission - an independent body - following constitutional provisions. Although redistributions are regularly conducted to aim for equal representation, in practice electorate rolls range between 72,345 and 138,836 voters per division, due to great differences in population densities across the country, and to ensure minimum representation for each state/territory, as enshrined in the Constitution.

Similarly to other Westminster systems, the leader of the party or Coalition with the largest number of elected representatives (Members of Parliament or MPs) gets invited by the Governor General (as representative of the Australian monarch) to form the executive branch of government (the Government). The second majority is formally designed "The Opposition", while all remaining MPs are known as "the crossbench".

In terms of the electoral system, voting is compulsory for all Australian citizens living in the country. Registration is automatic upon turning 18 years old or becoming a citizen, based on their place of residence. Both failing to update enrolment details and failing to vote attract fines - which are enforced. Participation rates usually range above 90%.

MPs are elected using a preferential voting system, in which voters must rank candidates in the ballot - with 1 being the preferred candidate, 2 their second choice and so on. When tallying up the results, the below process is followed:

- Votes are tallied based on first preferences - which are known as the **primary vote**.
- Candidates are sorted by their respective number of votes. If a candidate obtains an absolute

majority (i.e. one vote above half of the valid ballots), they are declared the winner.

- If no candidate reaches absolute majority, the ballots from the candidate with the lowest number of votes are taken aside and redistributed to the other candidate based on the second preference.
- Votes are tallied again, and the process is repeated (based on each ballot's highest preference still in the race) until a candidate obtains over 50% of the votes.

A graphical example of preference flows for the Division of Cooper is presented in figure 1.

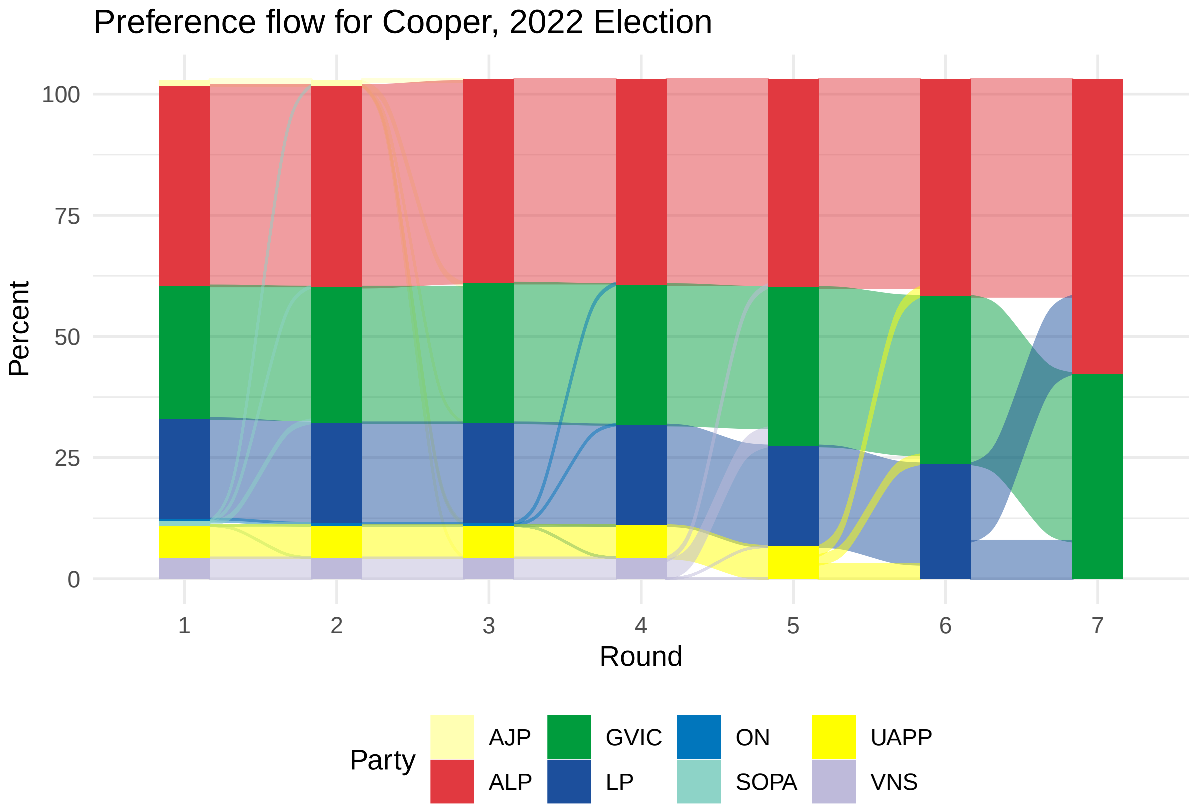


Figure 1: Preference flow for Cooper 2022

1.2 Australian Political landscape

Although a multi-party system in theory, Australian politics have operated as a de facto two-party system through most of the Federation's history - traditionally a labour-aligned party and its opposing counterpart. Nowadays the landscape is composed of two major forces ("parties of government") and a number of smaller groups, namely:

- The **Australian Labor Party (ALP)** (https://www.alp.org.au) is the oldest party in the system. As its name suggests, the party's origin is in the labour movement from the late Victorian era. This is a left-of-centre party and it is made up of two official factions: the centrist "Right Faction" (historically associated with some migrant communities and Catholic Social Teachings) and a "Left Faction" (with socialist roots). The party membership is comprised both of individuals and unions.
- The **Coalition (COAL)** is a grouping of two federal right-of-centre parties (fused in two states). The largest and leading party is the **Liberal Party (LIB/LP)** (https://www.liberal.org.au), founded in the 1940s by Australia's longest-serving Prime Minister, Sir Robert Menzies. Historically has served as a "big tent" organisation, agglutinating classically liberal and conservative groups. The second

party in the coalition is the **National Party (NAT)** (https://nationals.org.au), which is a conservative party representing rural constituencies. At a federal level, this a permanent coalition - when elected the Liberal leader becomes the Prime Minister with the National leader serving as deputy.

- The **Australian Greens (GRN)** (https://greens.org.au) is considered the largest of the "minor parties". One of the oldest green parties in the world, they are described as a left-wing, progressive party with a focus on environmental policies. On the political spectrum, they are perceived as being to the left of Labor.
- **Minor right-wing parties** like One Nation and United Australian Party (UAP), Fishers and Shooters, etc. Most of them are perceived as further right than then then Coalition. Most of the time don't have any parliamentary representation, except of representation for One Nation and the UAP. Their support base is in rural areas, especially in North South Wales and Queensland.
- **Minor left-wing parties**, like Socialist Alternative and Socialist Alliance. None of these parties has ever attained federal political representation.
- **Independents**, which traditionally do not fit into the classical political spectrum. Historically independents didn't have common positions, which changed in the last election with the arrival to Parliament of centrist "Teal" independents.

Generally speaking, small parties do obtain some moderate representation in the Senate. Senators are elected under a different set of rules, which are of the scope of this project.

1.3 Problem statement and research questions

As expected, general political knowledge has mapped distinct audiences to each group. According to this working-class people vote Labor, wealthy professionals and business owners are inclined to the Liberal Party, farmers prefer the Nationals and an inner city, young liberal population vote for the Green. Political science also postulates that people become more conservative with age and as they accumulate wealth. However, global and local events in the last 10 years have shifted some of those political narratives. In particular, both major parties have experienced internal friction when defining their policies towards climate change (an issue resonating with urban electorates and with growing effects in Australia but contentious with some businesses and unions involved in Coal and Gas exporting sector), housing (rising prices due, partly influenced by policies favouring the real estate market as an investment mechanism). According to the media, political commentary and the same parties, this has shifted traditional voting patterns where:

- There is an increasing divide between wealthy inner metropolitan areas, the suburbs, and regional and rural populations. In this divide, inner metropolitan areas seem to favour left-wing parties, and rural areas are conservative, leaving suburbs as the middle ground where economic issues are relevant.
- People are not shifting to conservative positions with age. In particular, Millennials are not voting for the Coalition due to climate policies and being locked out of the housing market. According to this, their vote is captured by either Labor or the Greens.
- Voters from historically Liberal (Party) wealthy, and classically liberal (as in liberalism) electorates (so-called "little l Liberals") are dissatisfied with the Coalition's perceived conservative turn, especially when it comes to climate change and positions on "moral issues". In this narrative, these voters have supported the rise of the "Teal" movement (environmentally progressive, socially liberal, economically conservative/moderate).

Another aspect usually considered in political analysis is the influence of ethnically diverse voters. Being a “migration nation” and with roughly half of the population being a migrant or a child of at least one migrant, culturally diverse communities are perceived as relevant political audiences.

Within this context, this project attempts to study how population make-up may influence how Australian citizens may vote for a particular party. Specifically, the project will look into the following questions:

- Is there a demographic divide between the inner city and suburbia? ([7],[5])
- What are the main demographic factors influencing political persuasion?

This then will be used to examine some common political narratives against the 2022 election results, such as:

- Was the so-called “Teal Wave” supported by discontent moderate Liberal voters? ([8])
- Are Millenials not becoming more conservative as they age? ([9],[10])
- Is culturally diverse voting relevant? ([3])

This project does **not** aim to develop a forecast tool. A naive retrospective forecast of the recent 2022 is included in this report, however it is intention is to validate primary voting drivers. On the same line, this project does not consider the problem of forecasting the actual winner of each election, which can be defined as a different topic altogether.

2 Data

2.1 Data Sources

Data used in this project comes from two sources, namely:

- The **Australian Electoral Commission** (AEC, https://www.aec.gov.au/). The national body overseeing and running federal elections, the AEC contains detailed election result records. All results for federal elections held in the 21st century are available online, through their Tally Room website [11].
- The **Australian Bureau of Statistics** (ABS, https://www.abs.gov.au/) . 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 [12], 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:

- In both cases, data are provided in large volumes and exhaustive granularity. Data extraction and aggregation can be time-consuming and resource-intensive if not done effectively.
- Census data points are provided using the ABS own geographical standard - and only a small selection of census data is provided already aggregated for each Commonwealth Electoral Division. Conversion between ABS geographical structures and electoral divisions is not straightforward as there is no 1:1 correspondence. Both geographical reference systems are modified at each election and each census.
- Despite the best efforts of both organisations in keeping consistency, names of electorates, parties, and census attributes change over time, which requires keeping track of all those changes and mapping them accordingly.

To assist in dealing with these issues and ensure repeatability, it was necessary to write code to guarantee some level of repeatability and consistency when extracting and transforming data. This resulted in three R packages being written to undertake this task:

- **{auspol}** [13] https://carlosyanez.github.io/auspol/), which extracts and presents electoral results.
- **{auscensus} [14]** https://carlosyanez.github.io/auscensus/, which allows to interact with Census Data Packs to extract different statistics across geographical units, and across censuses.
- **{aussiemaps} [15]** https://carlosyanez.github.io/aussiemaps/, which assists with aggregating census data into electoral divisions, by matching and apportioning different geographical structures.

Although this project's scope is data analysis, writing this three packages was a significant task, requiring as large time investment. Appendix B presents a more detailed view on their design principles. Additionally appendices C, D, and E contain a copy of some of their vignettes explaining their use. With their help, it was possible to build a basic data extraction and transformation pipeline, which is represented by figure 2.

In four steps, the extraction process consists of:

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

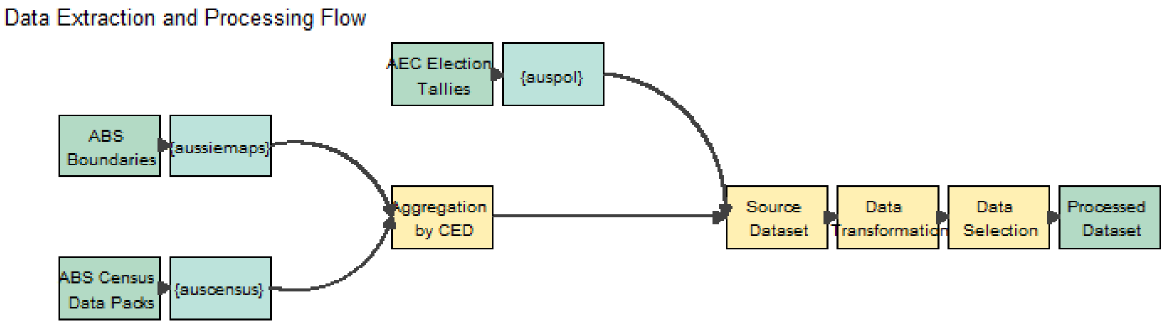


Figure 2: Flow of data from sources to dataset

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 it 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. This dataset has been further refined throughout the data exploration and modelling steps.

Processing sources into a usable dataset represented a particular arduous tasks. Although there were no issues in terms of missing or erroneous data, the previously mentioned challenges meant that intense processing was required. Getting data into the right shape consisted of:

1. Obtaining data at the most granular spatial division available in the Census packs (e.g. Statistical Area 1 (SA1) for the 2011,2016 and 2021 censuses). For reference for the 2021 Census there were 57,523 SA1 areas in Australia.
2. Mapping SA1 data into Commonwealth Electoral Divisions. For SA1s overlapping more than one electoral, data was apportioned based on overlapping areas.
3. Converting all values from absolute (e.g. number of votes, number of people in a particular demographic category) to percentages against total number of voters (elections)/ individuals (census) in the respective electorate.

2.2 Data Selection and Initial Transformation

There were a number of considerations that were taken to obtain the dataset that was eventually used, including what and how to represent the statistics and how to best align census and election data.

How to present the numbers

The first point to consider was how to represent the data in a way that is consistent across electorates and time. As mentioned in the introduction, although the aim behind the creation and geographical distribution of Commonwealth Electoral Division is to provide equal representation in Parliament for every Australian, this is not completely possible in practice, resulting in electorates varying in population. This is mainly due to the large variation in population density across Australia, combined with a constitutional mandate to guarantee a minimum number of seats per state or territory. For this reason, it is deemed necessary to represent all voting and demographic statistics as a percentage of each electorate's roll or population. This is also useful when comparing statistics across time.

The second point to address is the correspondence between census and election data. Since the 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 3 presents the best matches between both events held in the 21st century.

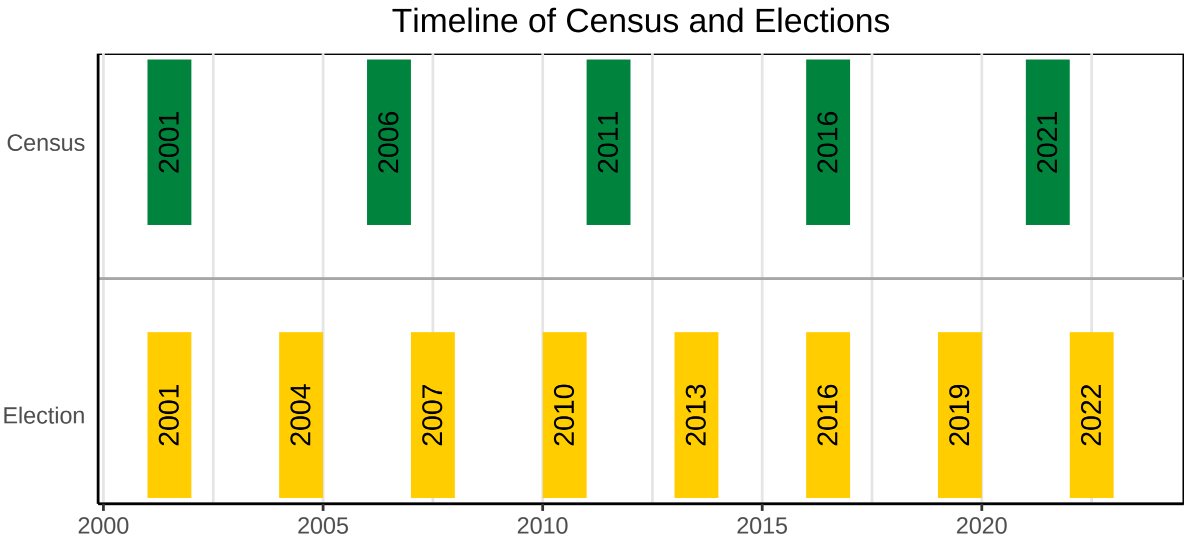


Figure 3: Census and Elections Timeline

Considering the census data available and selecting the elections closer to each census, four sets of events were selected for data extraction. They are presented in table 1.

Table 1: Census-Election pairings

Census	Election
2006	2007
2010	2011
2016	2016
2021	2022

Please note that this selection will remove half of the election events within the period, which may affect model accuracy. However, since the objective is not to obtain an accurate prediction this has been accepted as an acceptable trade-off to avoid, instead of having to interpolate demographic statistics.

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. The only adjustment was to reclassify the vote into four groups (referred to as parties in this document):

- **ALP** for the Australian Labor Party.

- **COAL** representing the Coalition, made of the Liberal Party, the National Party, the Liberal National Party of Queensland [3] and the Country Liberal Party in the Northern Territory [4] .
- **GRN** for the Australian Greens.
- **Other** to collect votes from any other candidates, including minor parties and independents.

A data sample is presented in table 2.

Table 2: Sample extraction - Canberra 2022

Year	Division	Abbreviation	Party	Votes	Percentage
2022	Canberra	ALP	Australian Labor Party	34,574	45.20
2022	Canberra	GRN	The Greens	19,240	25.15
2022	Canberra	COAL	Liberal (Coalition)	16,264	21.26
2022	Canberra	Other	Other Parties	6,417	8.39

Census Data

When it comes to Census data, a number of considerations had to be tackled during extractions, namely:

- **Large volumes of data**. Each census collected a large number of statistics. For instance, the data release for the 2022 Census contains 62 different tables, ranging from 8 [5] to 1,590 [6] attributes.

- **Data aggregated per electorate**. Although the ABS provides statistics for *non ABS* geographical structures, this only includes a subset of all data points collected. Thus, in many cases is necessary to extract data for granular-level ABS units (*SA1* in 2022) and aggregate them into electoral divisions. Without knowing the population density for each SA1, values have been approximately apportioned using areas.

- **Consistency across time**. Due to the changing nature of a Census (to better serve its purpose), there are some minor variations in how data is collected and aggregated from Census to Census.

To obtain a first selection of potentially relevant demographic variables to extract, existing literature and journalistic sources were consulted ([1], [2], [3]). Since many variables are colinear by definition (e.g. income groups) or they are closely related (e.g age and relationship status), the initial selection was inspected. After iteration, a resulting set of 55.00 attributes was chosen, which can be classed into the following categories:

1. **Income**: Distribution of the population in pre-set income brackets. The highest income bracket includes everyone earning 2,000 dollars or more each week.

2. **Education Level**: Distribution of educational achievement (from incomplete secondary to vocational education and academic degrees).

3. **Age**: Year of birth is captured in the census, which was grouped into generational cohorts. 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).

[3] In Queensland and the Northern Territory, the Liberal and National branches have merged. Elected federal MPs and senators sit with Liberals if they come from an urban area, or the Nationals when they represent a regional/rural/remote electorate.

[4] In Queensland and the Northern Territory, the Liberal and National branches have merged. Elected federal MPs and senators sit with Liberals if they come from an urban area, or the Nationals when they represent a regional/rural/remote electorate.

[5] *02 -Selected Medians and Averages*

[6] *09 - Country of Birth of Person by Age by Sex*

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 another arrangement (e.g. public housing).

7. **Citizenship**: Percentage of the population that hold Australian citizenship. Although non-citizens are not entitled to vote, this variable can be taken as a proxy for the relative integration of migrant communities into civic life.

8. **Religion:** Percentage of the population declaring to profess a religion. For this analysis, large and high-growth religious groups were selected. For practical reasons and to use as a potential community proxy, the values of Anglican, Presbyterian and Uniting followers were merged into a single statistic.

9. **Language**: Languages spoken in the community. Similar to religion, a selection of relevant languages have been included to reflect the historic and current migrant communities.

Additionally, each electorate was classified as **metropolitan** if it lies within the boundaries of Australian capital cities or **non-metropolitan**, when it is not the case. Altogether, these variables try to reflect wealth and education (cited by [1] as key factors influencing political persuasion), as well as the stage in life and belonging to a particular migrant community (sometimes cited as an influential factor, for instance in [3]).

A sample of the resulting dataset is present in table 3. A detailed list of all variables is presented in appendix A.

Table 3: Dataset sample

			Age		Household	Language		Relationship	Income
Election Year	Division	Australian Citizens	Baby Boomers	Gen Y	Rented	Chinese	Italian	Single Parent	2000 or_more
2016	Adelaide	77.31	17.77	25.73	31.68	6.88	2.25	3.82	10.20
2010	Bonner	84.58	16.24	22.48	24.25	2.74	0.67	4.52	7.19
2010	Calare	90.65	19.23	18.49	22.76	0.30	0.25	4.79	4.41
2022	Clark	80.24	20.06	27.09	28.19	4.59	0.32	4.81	10.16
2022	Dawson	84.51	21.10	21.24	26.82	0.57	0.34	4.67	11.62
2007	Fadden	81.27	25.37	21.29	31.95	1.17	0.46	4.86	2.49
2016	Forrest	84.47	20.80	19.38	22.60	0.53	0.74	4.42	8.31
2010	Gippsland	89.78	22.20	17.67	19.46	0.44	0.94	4.80	3.95
2010	Hindmarsh	85.73	19.29	21.83	24.62	2.74	3.64	4.22	4.73
2010	Lyne	91.72	25.62	14.32	22.48	0.19	0.16	5.10	2.43
2007	Lyons	91.11	29.34	19.01	16.40	0.08	0.10	4.05	1.10
2022	Makin	87.22	20.23	22.08	17.72	1.79	0.98	5.02	6.74
2016	Maranoa	88.03	23.18	17.49	28.28	0.22	0.25	4.32	4.38
2022	Maribyrnong	86.74	18.73	24.81	27.88	3.41	5.20	4.16	17.51
2022	McMahon	78.95	18.47	22.46	29.92	4.13	1.37	5.21	6.27
2010	Moncrieff	75.43	18.81	24.27	37.17	1.91	0.70	5.04	4.58
2007	North Sydney	78.80	24.81	17.71	38.21	5.70	1.03	3.05	16.10
2016	Port Adelaide	84.90	17.65	22.79	22.26	1.78	2.55	6.12	2.94
2022	Riverina	89.07	22.64	18.84	23.17	0.56	0.06	4.85	7.88
2022	Scullin	81.47	16.99	25.55	24.21	3.99	4.40	4.80	6.51
2022	Spence	86.74	17.77	23.22	25.24	0.42	0.69	6.93	3.95
2007	Tangney	83.23	29.04	25.56	17.82	7.67	1.08	3.92	5.84
2010	Wakefield	87.55	16.08	20.70	21.45	0.24	1.04	6.19	1.77
2007	Wide Bay	89.00	28.94	18.84	22.09	0.10	0.16	4.77	0.96
2022	Wills	81.73	13.99	32.64	35.44	2.26	6.00	3.79	14.96

2.3 Training, Validation and Testing Split

After obtaining the data, the election results and census statistics for the 2021/2022 cycle were set aside, since they have been used as testing dataset, in a election forecast attempt. The remaining data has been used in exploratory analysis, data mining and creating and fitting models.

2.4 Data Exploration

In total, the resulting dataset is made up of 4 response variables and 55 potential predictors, plus identification attributes like division name and election year. As expected the many covariates exhibit moderate to high collinearity. Also, it is possible to observe some loose correlation between some of the covariates and some of the responses. The complete list of variables is presented in appendix A.

As an example, figure 4 shows a somewhat weak correlation between Coalition primary vote and the percentage of the Baby Boomers. Figure 5 presents the correlation values for religion and language variables, where is possible to see:

- A positive correlation between monolingual English speakers and membership in Anglican, Presbyterian and Uniting churches. Together, they are likely proxies for Anglo-Celtic population.
- Similarly, there are somewhat expecting origins that most likely indicate concentrations of linguistically and culturally diverse pockets, e.g. Hinduism and South Asian languages, Catholicism and Italian, and Buddhism and East Asian languages.

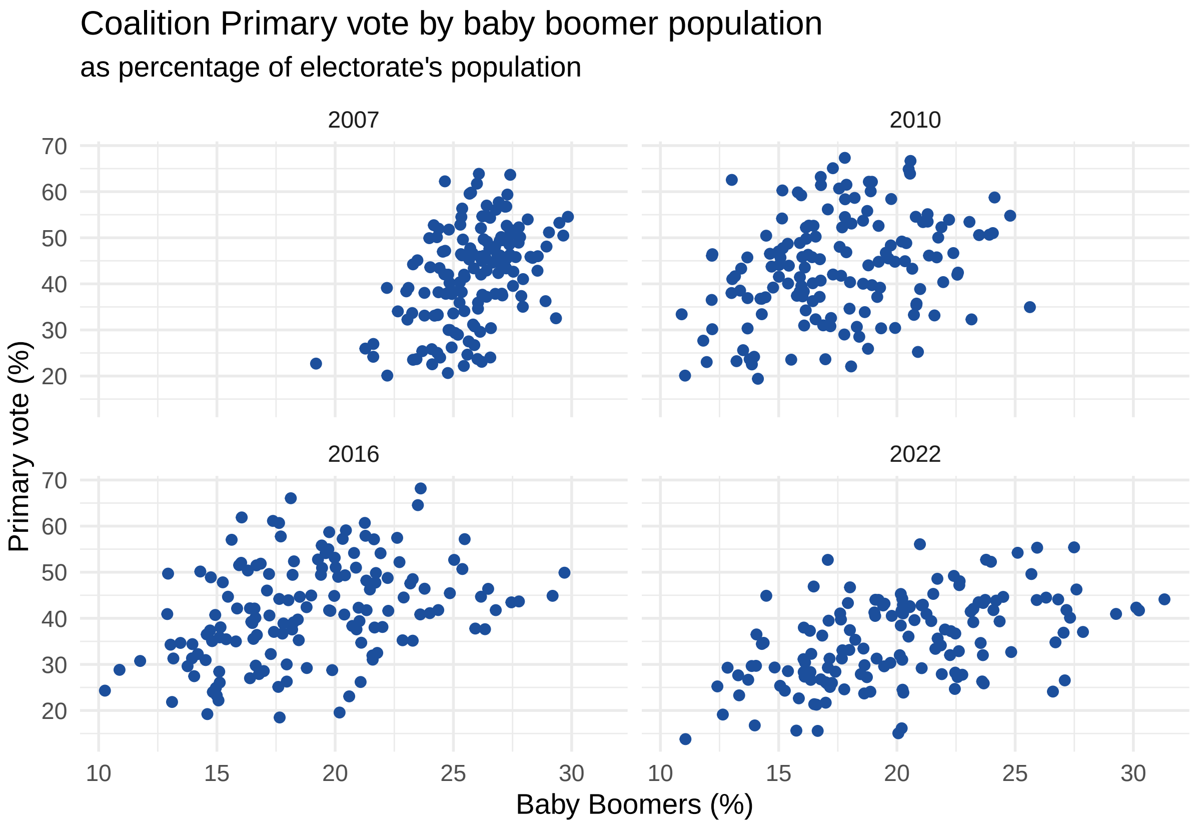


Figure 4: Correlation between Coalition vote and Baby boomer population

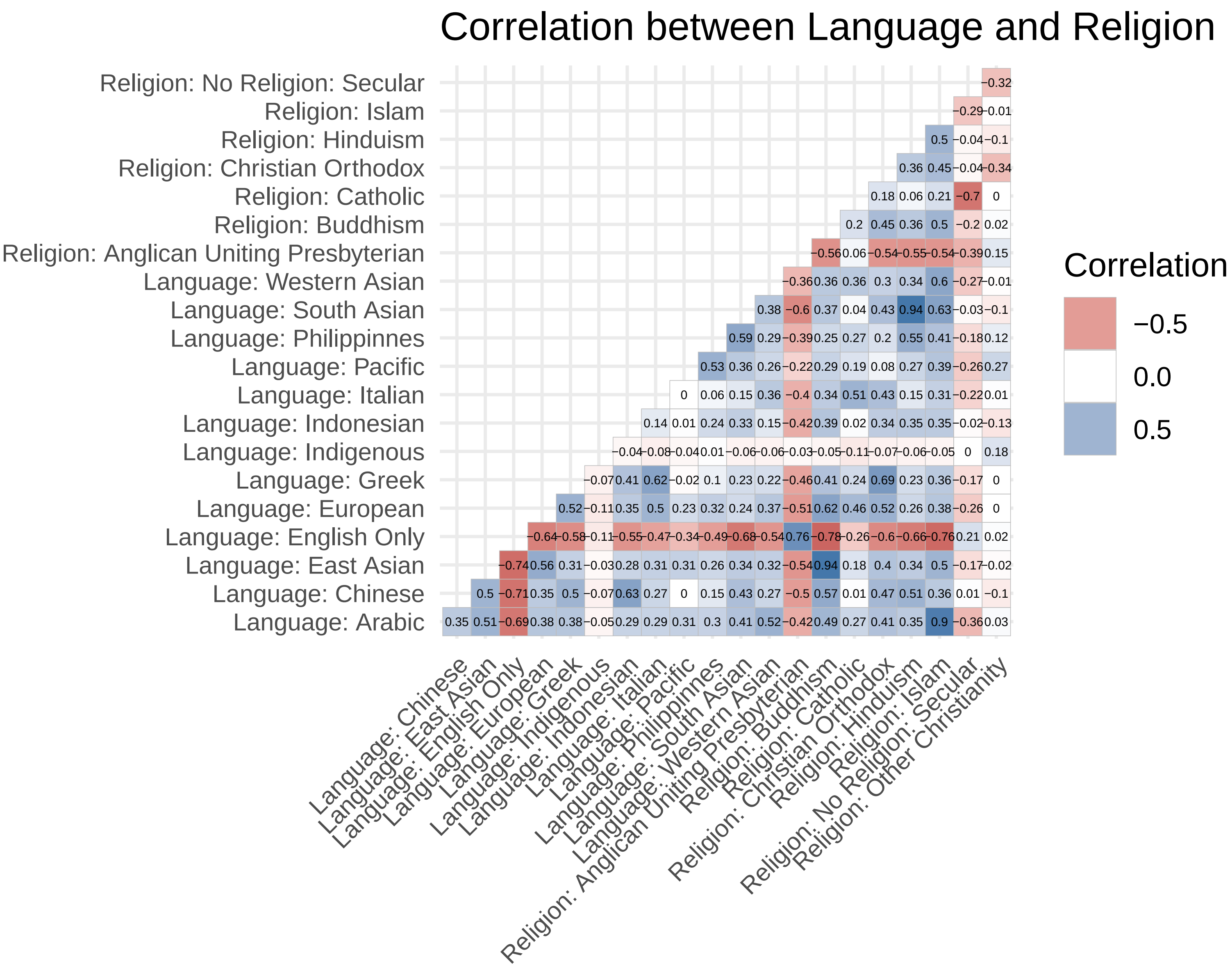


Figure 5: Correlation for selected covariates

Additionally, after a detailed inspection, it is worth noticing that :

- There is no apparent change in the relationship between a given covariate and the responses when broken down by state or capital city.
- There are no obviously distinguishable differences when splitting results by each election.

Dimensionality reduction using Multiple Factor Analysis

Given the large number of colinear covariates, it is worth exploring if a reducation of dimensionality could help to better measure variation in a meaningful way and in a more manageable number. To achieve this, **multiple factor analysis** (MFA) [16] was used as the clustering algorithm. MFA is essentially an extension of Principal Component Analysis that can deal with categorical and numerical variables, as well as variables that belong to groups (e.g in this case, Income, Religion, Household Tenure, etc.).

The resulting scree plot and cumulative variance are presented in figure 6.

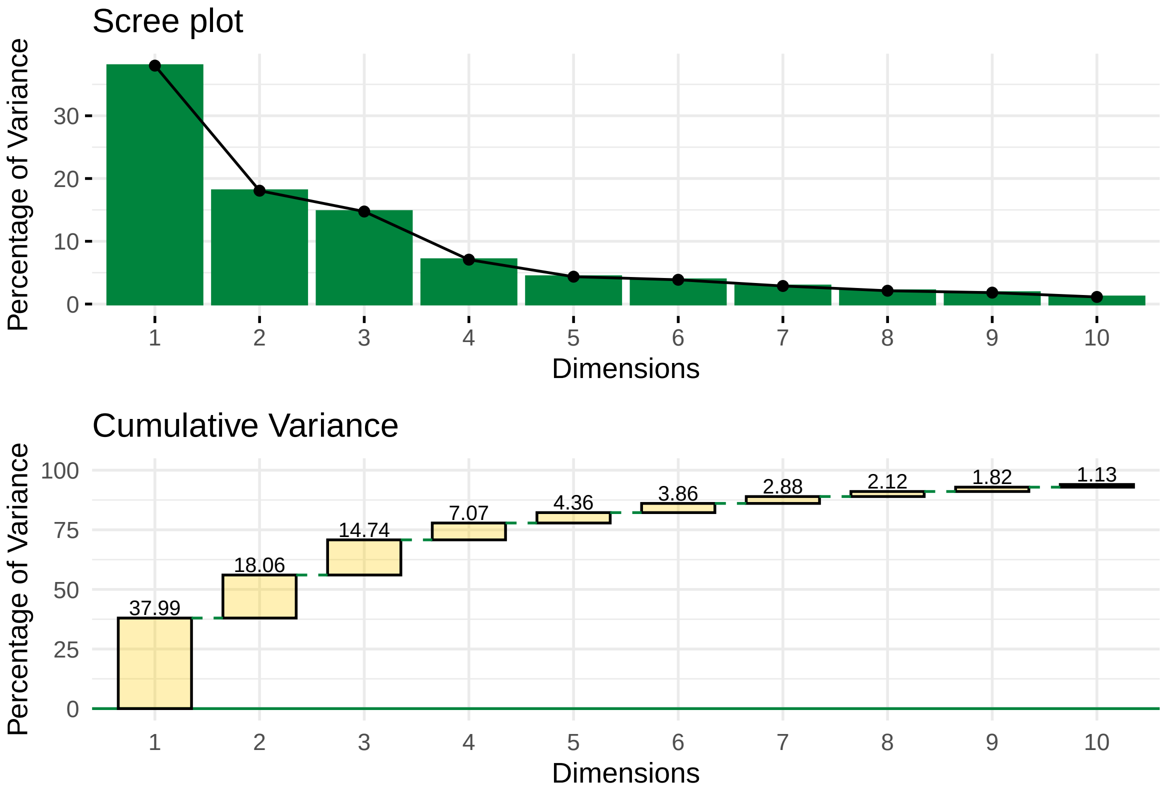


Figure 6: Scree plot and cumulative variance

Figure 7 presents group biplots for the 8 most important dimensions. Unfortunately, there is no straightforward representation except with Dimension 2 and Education variables.

Electorate segments

Normally when characterising votes, Australian politicians and political media make a distinction between inner-city voters (touted as wealthy and progressive), suburbia (“middle Australia”), and the bush and outback areas (conservative, “battlers”, “real Australia”). Therefore, it is of interest to explore if this can be substantiated by demographic attributes, as it may have an impact on primary voting.

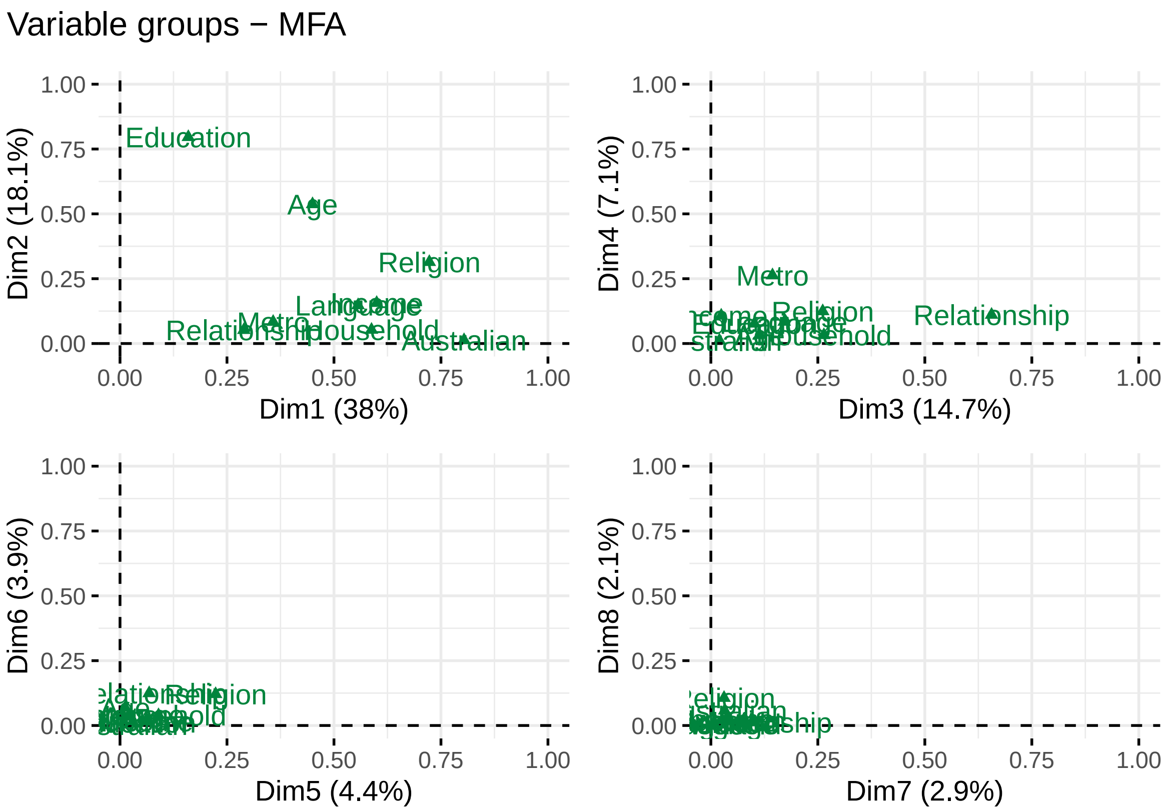


Figure 7: Group plots for first 8 dimensions

Using all demographic variables a clustering algorithm has been applied to identify those clusters. Different clustering approaches were, eventually choosing to:

- ignore Census years and pool all records in a single pool.
- transform all demographic attributes to represent the difference between each data point and their corresponding national value (in the same year).
- use *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) [17], a density-based hierarchical clustering algorithm.

HDBSCAN is essentially a density-based method to find common density areas, for which then a hierarchical analysis of the results is used to heuristically determine the number of processes. This is an appropriate method in this scenario.

This results in 3 distinct clusters of electorates. When presented in a map, it is possible to obtain figure 8 for 2016.

These three clusters are:

- **cluster 0** seems to mostly contain electorates located in the inner cities, especially in Sydney and Melbourne. These areas tend to be more affluent, either “established” or “gentrified” suburbs. Notably, it also contains the three northernmost, remote electorates.

- **cluster 1** comprises all regional areas outside state capitals (with the exception of Hobart in Tasmania).

- **cluster 2** largely represents “suburbia”. It is also more prevalent in Brisbane and Perth compared when comparing capital cities.

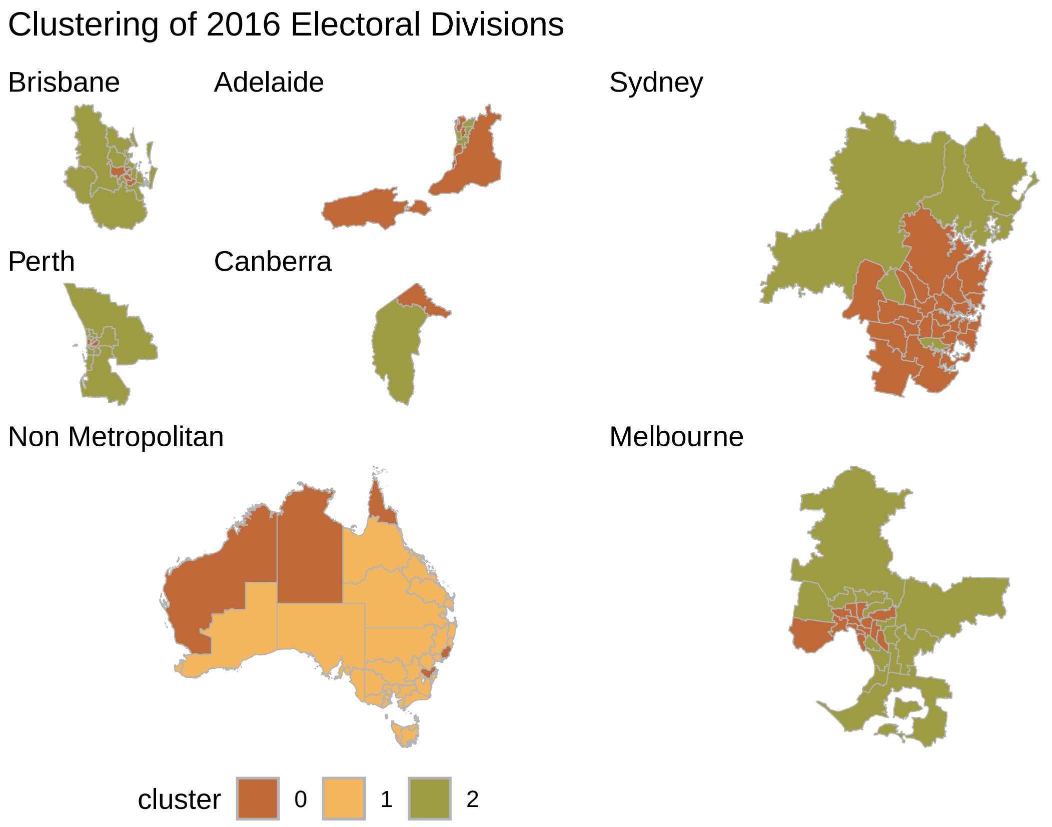


Figure 8: Clusters for 2016 Election

Revisiting the demographic attributes can help to understand how these clusters differ from each other. A selection of those variables is presented in figure 9.

Even though it is possible to find electorates from every country across the spectrum for every attribute, it is possible to observe that cluster 0 tends to concentrate areas with significant Millenial, highly educated, and relatively affluent populations. These areas also tend to attract newer migrants (lower numbers of citizens) and therefore they possess higher percentages of multicultural populations (such as Chinese speakers). Cluster 1 tends to concentrate older people, with lower percentages of tertiary and vocational education and possibly higher proportions of Anglo-Celtic Australians. Cluster 2 seems to be sitting in the middle of the other two clusters. Adding these findings to the geographical locations seems to confirm there is some element of truth in the stereotypical classification of voters.

3 Method

Going back to the introduction, the objective of the exercise is to determine if demographic attributes can influence or explain voting patterns. This can be restated into determining if demographic attributes can serve as predictors of primary voting. In mathematical terms, this can be expressed in a simple way by equation (1).

$$\mathbf{Y} = f(\mathbf{X}) \tag{1}$$

where **Y** represents a vector with primary voting for an electorate, and **X** represents the vector of respective

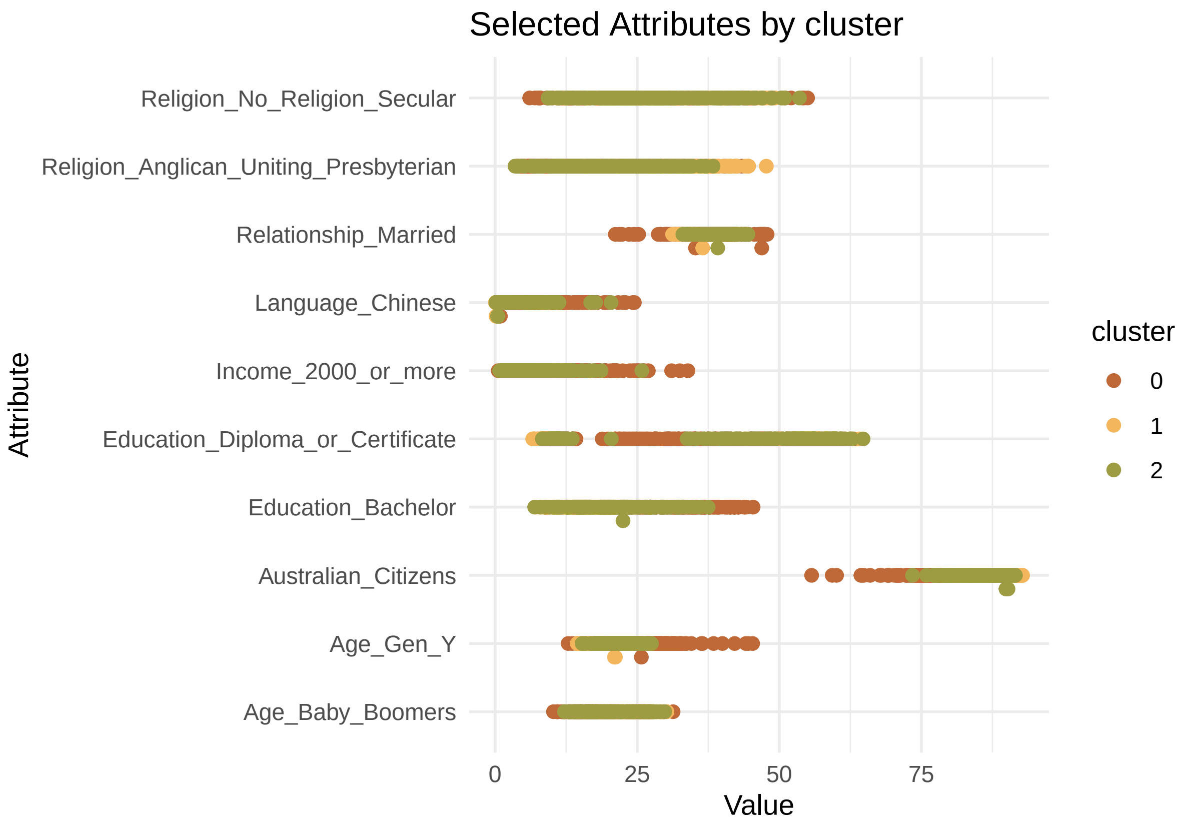


Figure 9: Selected attributes, coloured by cluster.

demographic attributes.

In this simple form, other factors that may influence voting are not explicitly shown in the equation. However, these factors may be difficult to quantify as they potentially relate to a myriad of factors including the state of the economy, foreign affairs, perceptions about the governing party or any party in the election, or the mood of the times.

To solve this challenge, it is possible to naively assume tools like polling can effectively capture the *zeitgeist*. If that is the case, it is possible to split the original function $f()$ into a poll component and a demographic component. Since it is not in the scope of this project, we can also ignore the polling component and focus on the difference between the absolute value and polling results. To further simplify things, we can temporarily assume that polling results are uniform across the country, thus demographic statistics only influence the difference between the electorates' primary vote and the respective national percentage. This is expressed by equation (2), which also accounts for general error.

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon \tag{2}$$

where
$\mathbf{Y} = \mathbf{Y} - \mathbf{Yp}$
$\mathbf{X} = \mathbf{X} - \mathbf{Xn}$
and
Yp : Primary voting polling results
Xn : Demographic values at national level

This second iteration does not take into account that in different electorates, different demographic attributes may have a different effect on the primary vote. For instance, in more progressive areas, a higher proportion of younger people may have a greater effect on left-leaning preferences when compared with similar proportions of younger people in rural electorates. Equation (3) is intended to acknowledge those differences.

$$\mathbf{Y_i} = f_i(\mathbf{X_i}) + \epsilon \tag{3}$$

where
$\mathbf{Y} = \mathbf{Y} - \mathbf{Yp_i}$
$\mathbf{X} = \mathbf{X} - \mathbf{Xn_i}$

It is worth noticing that i represents a particular grouping of electorates, and for each group predictors can be different - as different attributes may have different impacts.

In terms of choosing an appropriate f_i, it would depend on the objective of the model. Given the large number of predictors and requirements on interpretability and accuracy, this could be a complex task. In this particular case, the focus is on understanding the factors that influence voting rather than producing accurate electoral predictions. After examining possible alternatives a model in the **regularised regression** family is deemed as an appropriate choice. The selection is based on the following factors:

- Regularised models are a good fit in cases with a large number of variables. Lasso regression in particular is a good method for variable selection, helping to identify the most relevant parameters.
- Regularised regression models can also deal with predictors exhibiting a high level of collinearity.
- The results provide an straightforward interpretation on how each predictor influences the results. In the context of this projects, that is a more important priority than forecasting accuracy.

Consequently, the task at hand consists in finding the regularised regression coefficients for the set of formulas represented by equation (4)

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{in} \end{pmatrix} = \begin{pmatrix} \beta_{i11} & \beta_{i12} & \cdots & \beta_{i1m} \\ \beta_{i21} & \beta_{i22} & \cdots & \beta_{i2m} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{in1} & \beta_{in2} & \cdots & \beta_{inm} \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{im} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \cdots \\ \epsilon_{in} \end{pmatrix} \tag{4}$$

A complication of this approach is that requires separating the electorates into different segments. This requires having a method to map electorates into clusters if such an assignment is not provided. Thus, the modelling task consists of:

- A classification model map new records into clusters with similar electorates.
- A regularised regression model to determine how demographic factors influence primary voting for each party.

4 Fitting and analysing a model

As mentioned in the previous section, this exercise requires fitting both a classification and regularised regression model.

4.1 Cluster classification

Although HDBSCAN can be used to map new data points into the existing clusters, a different approach has been taken: to "reverse engineer" the clusters by training a classification model. The intent behind this is to leverage the trained model to identify the main contributors to the classification.

Different models were tried, starting with basic tree partitioning. After a couple of trials, a **random forest** model was selected. Although not directly interpretable like a simple hierarchical tree, the use of **variable importance plots** still allows to understand which variables are the most significant influencers in the classification algorithm. This is deemed an acceptable trade-off between increase accuracy and explainability. The model was trained with:

- Census data from 2007 to 2016 (mirroring elections between 2006 to 2016), which was used for training and validation.
- Values for demographic attributes, which were centred around the overall percentage for said attribute, for the respective cluster.
- Clusters previously obtained with HDBSCAN, used as the response variable.
- Since the year has been "discounted", all values will be considered as one pool. An assumption has been made that the period in question is short enough to drastically affect the clustering model. If demographic values change - cluster assignment (for instance because of re-distribution), the effect is similar to being a different electorate.

The initial fitting produces the results presented in tables 4 and 5. A variable importance plot is also presented in figure 10.

Table 4: First Model - Metrics

Metric	Estimate
Accuracy	0.8333
ROC AUC	0.9621

Table 5: Accuracy by Cluster - First Model

Cluster	Accuracy
0	0.7742
1	1.0000
2	0.7576

From the chart above, it is possible to see that only a handful of variables significantly contribute to the cluster selection. Aiming for simplification, a random forest model with reduced variables was also trained, achieving similar results in accuracy and variable importance (shown in tables 6 and 7, and figure 11.

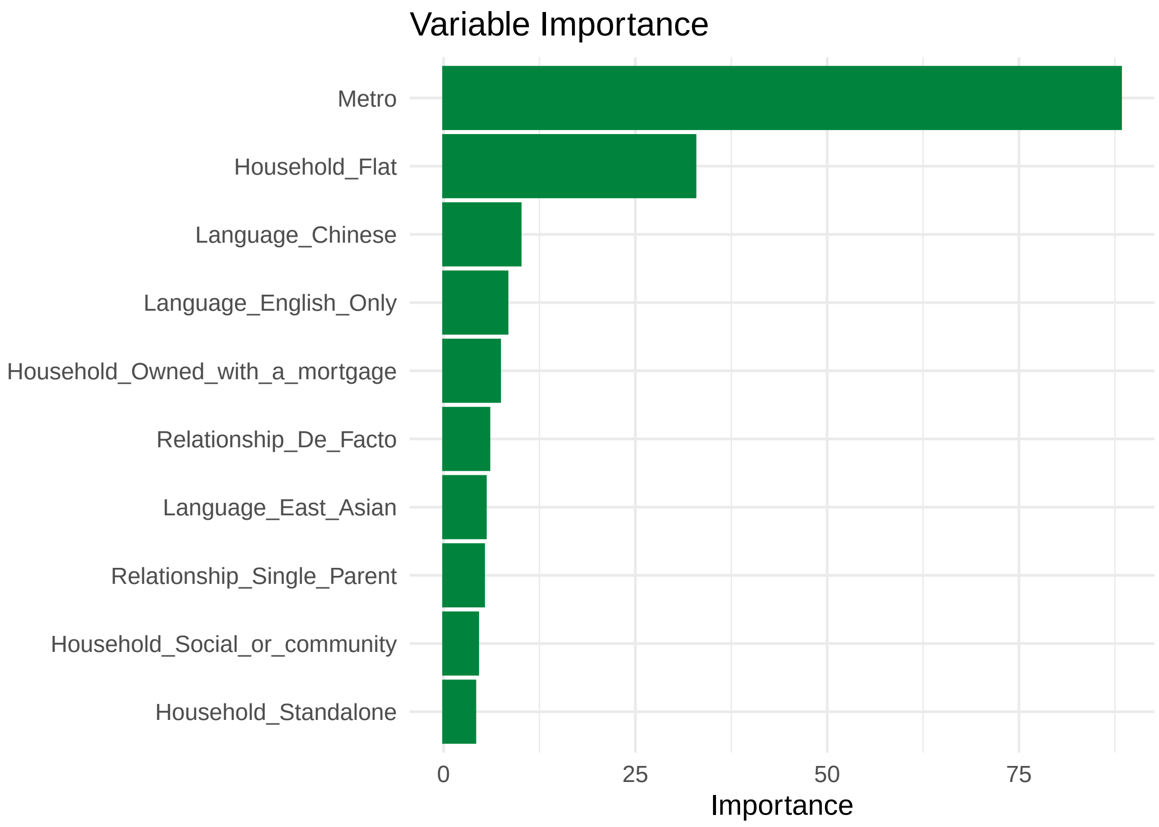


Figure 10: Variable importace - First classification model

Table 6: Improved Classification Model - Metrics

Metric	Estimate
Accuracy	0.8667
ROC AUC	0.9531

Table 7: Accuracy by Cluster - Improved Model

Cluster	Accuracy
0	0.8065
1	1.0000
2	0.8182

Looking at variable importance, it is possible to appreciate that cluster placement can be driven by :

- Location in a large metropolitan area or the regions.
- Population density, (type of household)
- Life stage (relationship) -Wealth (type of household ownership)
- Multicultural make-up of the area - first and second-generation migrants are more likely to be bilingual - thus the proportion of monolingual people is a proxy variable for this.

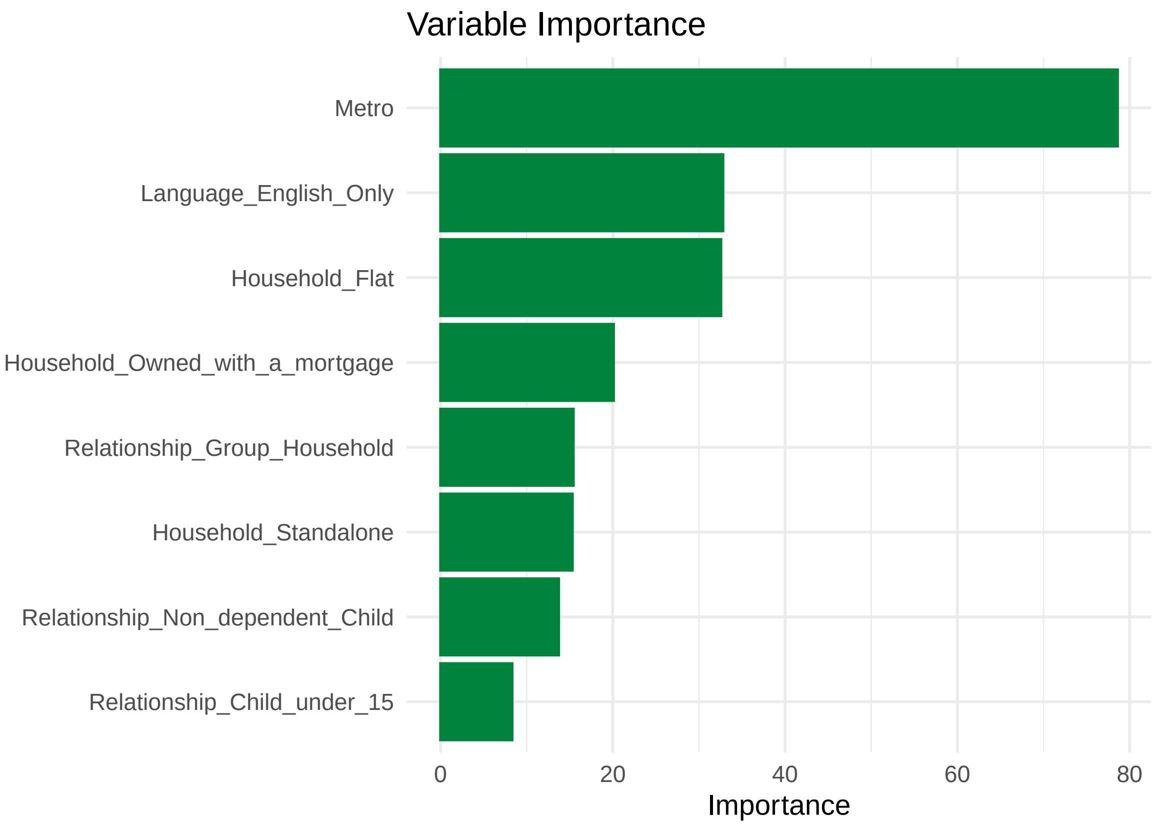


Figure 11: Variable importace - Improved classification model

This picture fits with the media narrative about differences in the electorate (quote).

4.2 Regularised regression

Due to the large number of variables, the first step is to see if it is possible to identify which factors may be of influence. For this, a Lasso regression was conducted with the sole intent of variable selection. Then an elastic net was fitted, with the goal to optimise the root square mean error (RMSE). This process was done separately for each cluster. Although precision is not a key objective of this exercise, table 8 presents the best RMSE result per cluster, alongside the selected tuning parameters.

Table 8: Best regression results by cluster

			RMSE				
Cluster	α	λ	Overall	GRN	ALP	COAL	Other
0	0.5416	0.5191	5.9363	5.7787	5.9668	7.4913	3.9802
1	0.9976	1.5452	5.9408	2.8817	6.1343	6.8756	6.9258
2	0.8408	0.5437	4.8961	1.6438	5.4578	6.0699	5.1531

However, the main objective is to understand the coefficients for each covariate, which are presented in figures 12,13 and 14.

It is worth noticing that some of the selected covariates may not be relevant in all electorates, by account of their small absolute various or being relatively uniform across the segment. For this reason, the

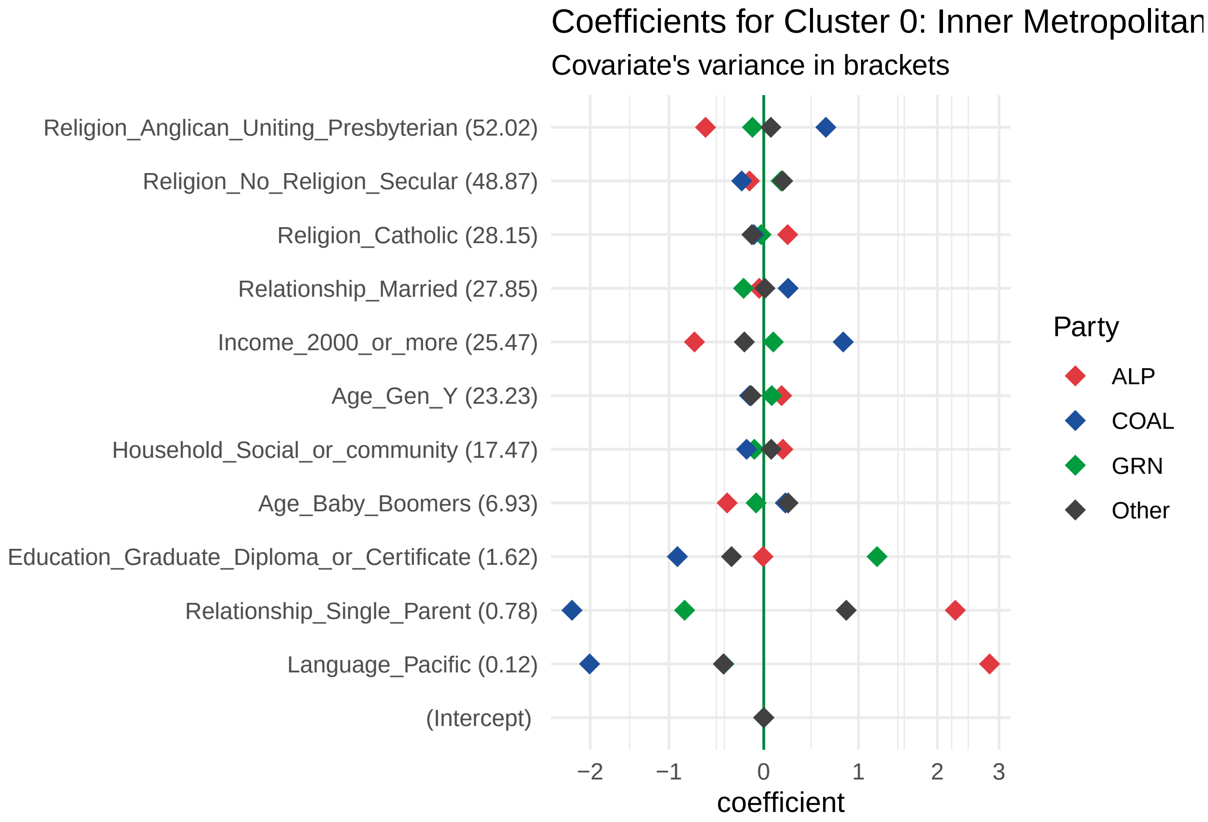


Figure 12: Resulting coefficients - cluster 0

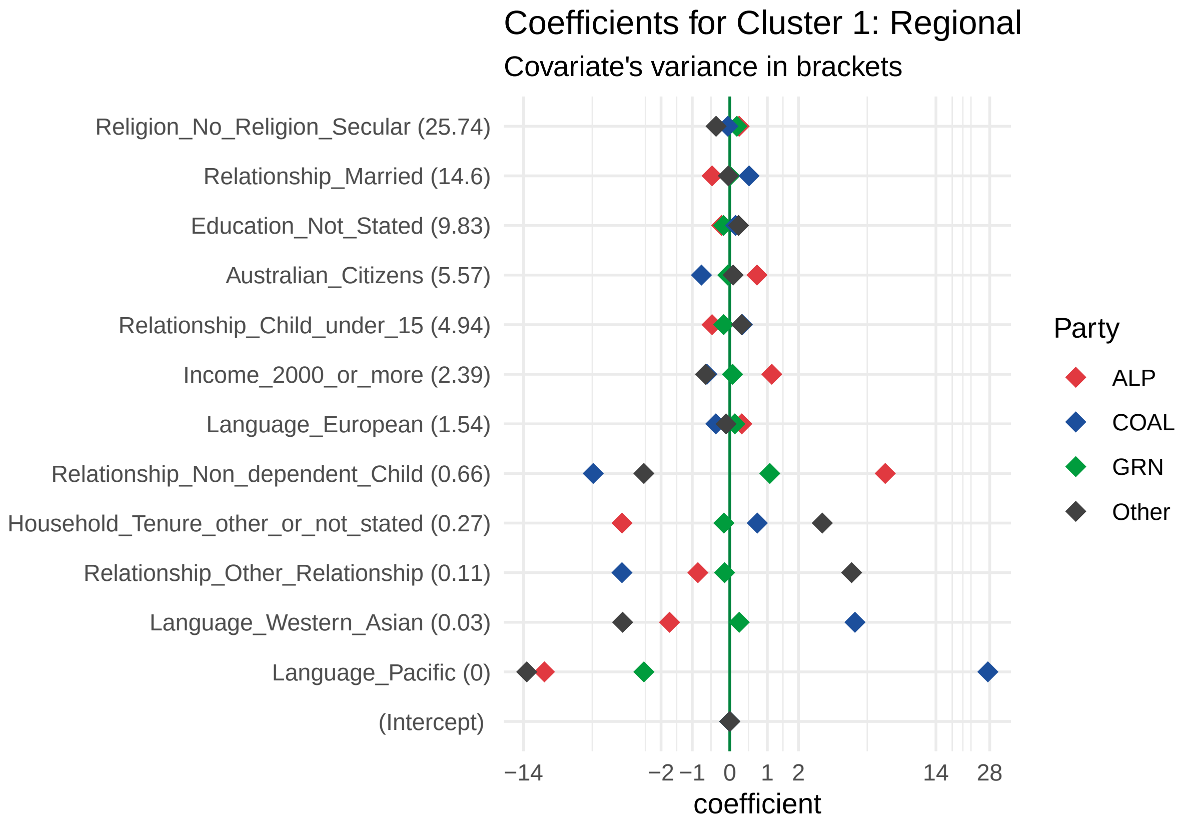


Figure 13: Resulting coefficients - cluster 0

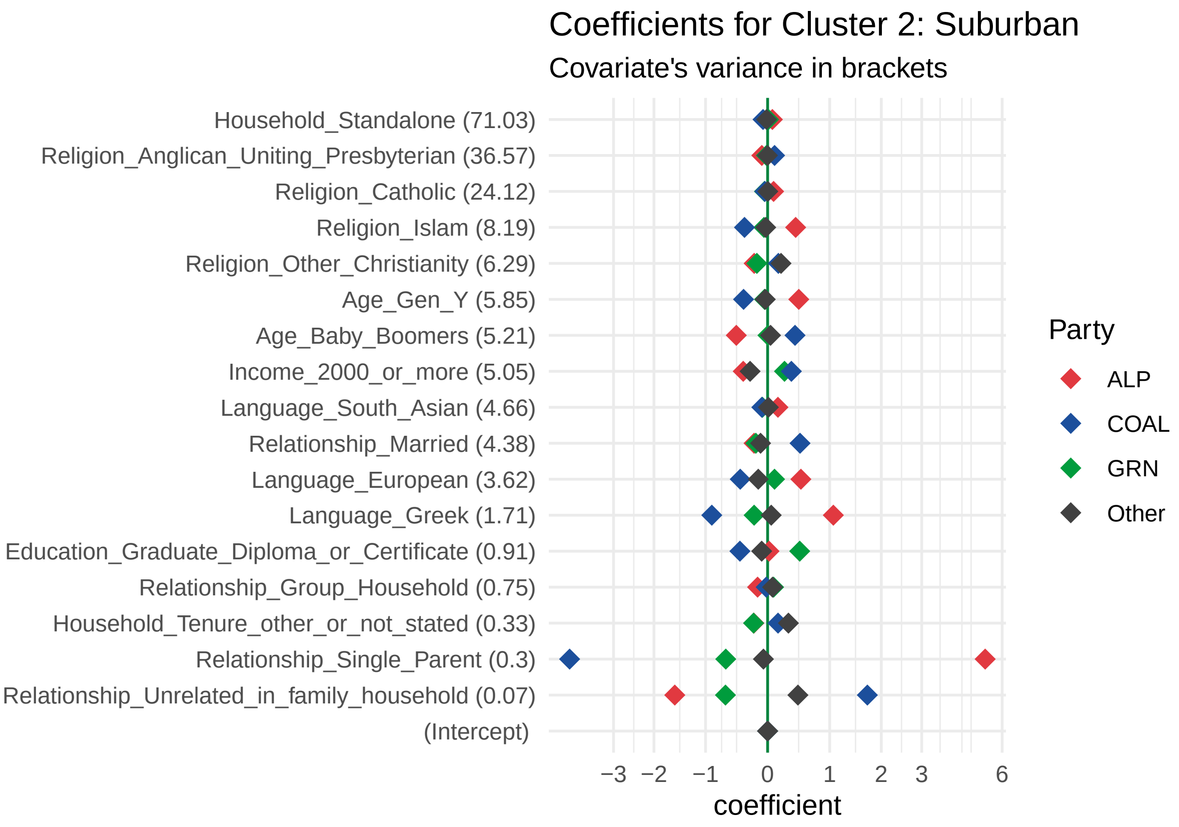


Figure 14: Resulting coefficients - cluster 0

covariates in figures 12,13 and 14 have been ordered by their respective variance - when assessing their overall effect / relevance this must also be taken into account.

When looking at each cluster, it is possible to summarise the different demographic effects as follows:

- In **cluster 0** (mostly inner metropolitan areas) political divides are drawn across wealth, religiosity (i.e. values) and generational lines.
  - In these areas, coalition vote is associated with higher percentages of followers of Anglican, Uniting and Presbyterian churches, people on higher income and Baby Boomers.
  - Labor vote is turn driven by followers of the Catholic Church (partially a reflection of the historic association between the Australian Catholic Church and the labour movement, and Irish and Italian migration) and Millennials. There is some association between less-advantaged populations and social and community housing.
  - Green vote is also driven by Millennials, but unlike Labor there is a positive association with higher income groups. Green votes are also related to the irreligiosity o secular population groups.
- In **cluster 1** (regional areas, including mid-size cities and rural areas), demographic variance is smaller. However, when it happens, it follows a different pattern from the main cities.
  - In this area, the Coalition vote has also a positive association with religiosity - this is not dissimilar to cluster 1, especially when considering that Anglicanism/Presbyterianism/Unitiarianism are the largest religious groups in the area). However, a key difference with the cities is that in case higher wealth groups have a negative association with Coalition vote.

  - Labor vote in these areas is driven by a larger proportion of Australian citizens and higher-income voters.
  - Overall, it seems there are no demographic factors influencing Green votes in these areas. Interestingly, age does not rank as a variable of importance.

- As expected, **cluster 2** (metropolitan suburbia), shares some traits with their inner-city counterparts, showing the same associations along religious, age and wealth lines. However, there are a larger number of predictors associated with the multicultural make-up of the electorates. Those covariates tend to have a positive effect on Labor vote and a negative influence on Coalition and Green voting. This difference is interesting, especially considering inner city areas are as multicultural as the suburbs.

5 Results

5.1 Forecasting the 2022 Federal Election

The previously fitted model can be used to attempt to retroactively forecast the 2022 Federal Election. Through this process, it is possible to illustrate the model's strengths and shortcomings in capturing how demographic factors succeed and fail to capture the change in voting patterns.

This exercise uses the results from the 2021 Census of Population and Housing. The base voting percentages are taken from the last Newspoll prior to the election [18] . Newspoll is usually considered a good predictor of the Australian election. The values are shown in table 9. Please note these values are national, but since there is no cluster-level data, they will be used nonetheless.

Table 9: NewsPoll primary vote forecast, 20 May 2022

Party	Forecast
COAL	35%
ALP	36%
GRN	12%
Other	17%

The first step in the forecasting process is to map the electorates into three clusters. The result is presented in figure 15.

After clustering, the regression models have been used to calculate a predicted outcome. Results have been transformed back to absolute values and then compared against actual and historical results. This is presented in figure 16, together with RMSE values in table 10.

Table 10: RMSE per cluster, overall and by parrty

cluster	Overall	GRN	ALP	COAL	Other
0	10.56	9.48	9.42	8.43	15.04
1	10.37	4.16	10.21	14.35	10.11
2	8.26	3.16	7.51	6.05	13.40

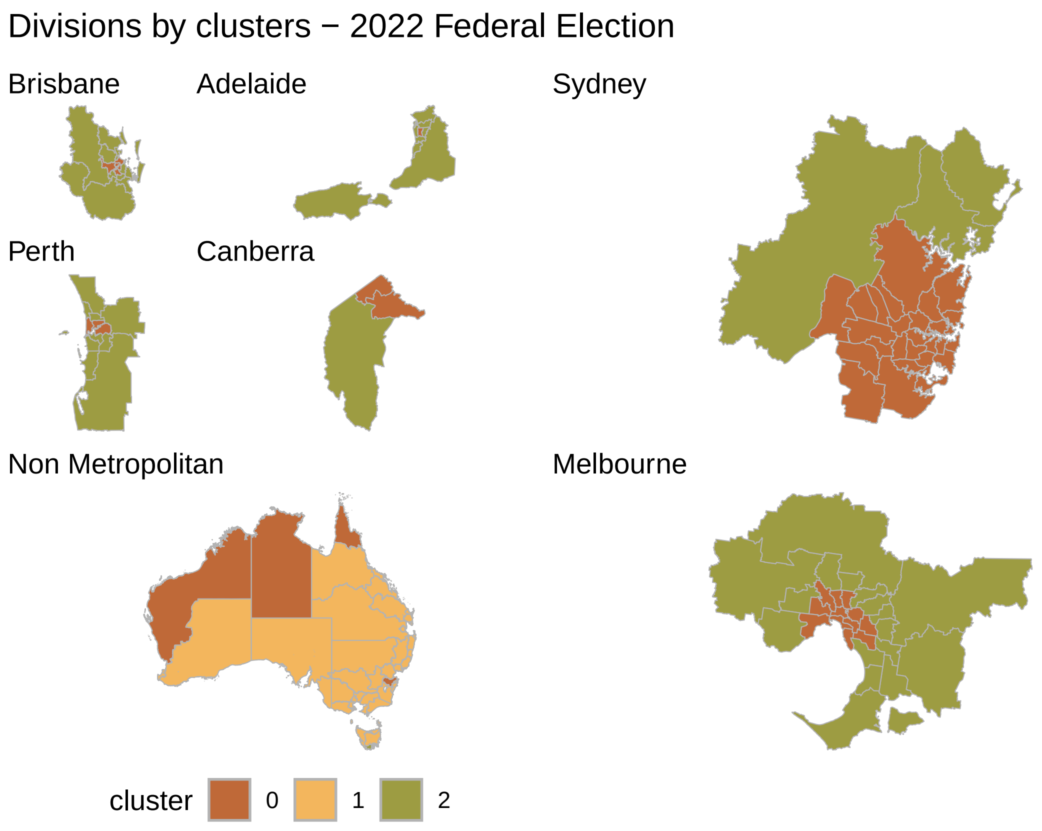


Figure 15: Clusters in 2022 Election

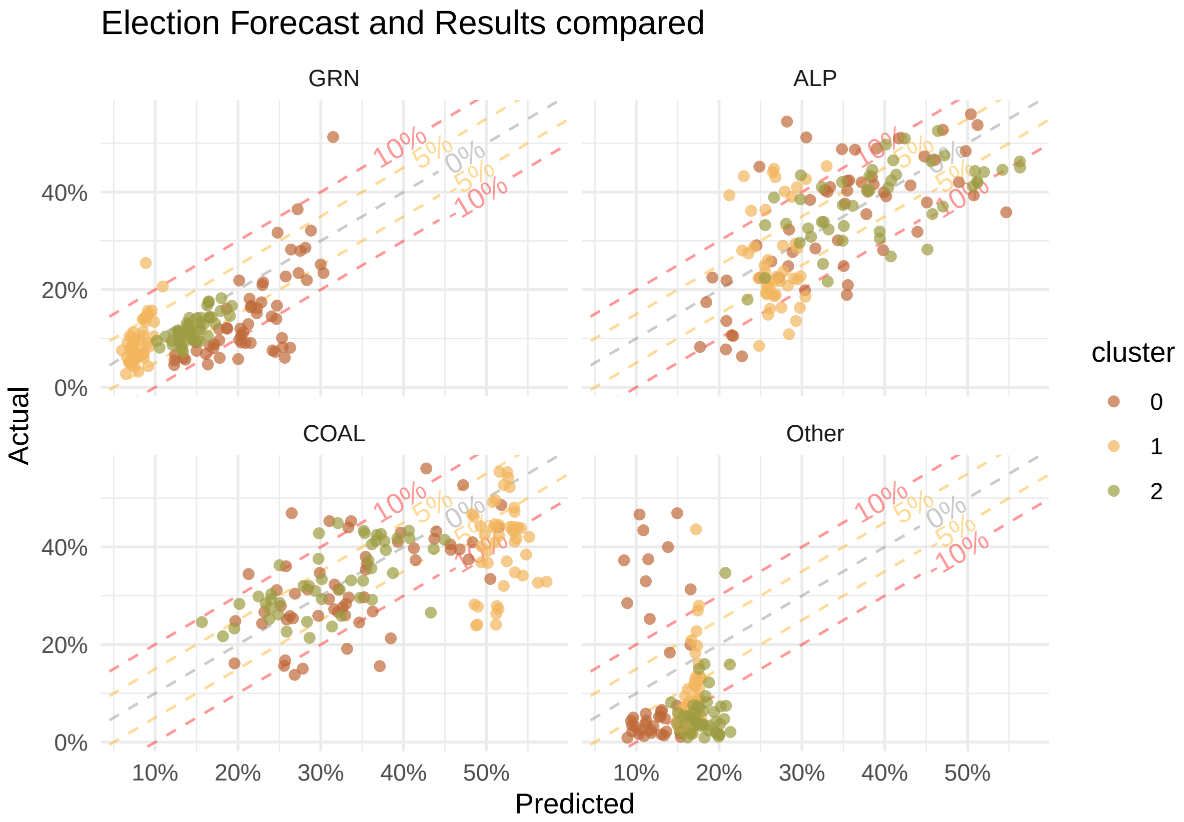


Figure 16: Comparison between prediction and election results

As expected, the results fail to adequately forecast primary voting, especially when it comes to Other parties and independents. However, it can be used as a tool to analyse the vote dynamics.

5.2 The Teal Wave

A particular phenomenon of the last election consisted in the so-called "Teal Wave", where centrist independents campaigned in traditional Coalition electorates. Most of these electorates are located in inner-city, wealthy areas of Melbourne and Sydney, where voters have consistently voted Coalition since the Australian Federation. Right-leaning voters in these areas are perceived as moderate, socially liberal ("little-l liberals") who were dissatisfied with a perceived conservative turn in Coalition politics. Teal candidates managed to unseat incumbent MPs - did they in effect capture the dissatisfied Coalition base? The results and predictions for 4 cases are presented in figure 17.

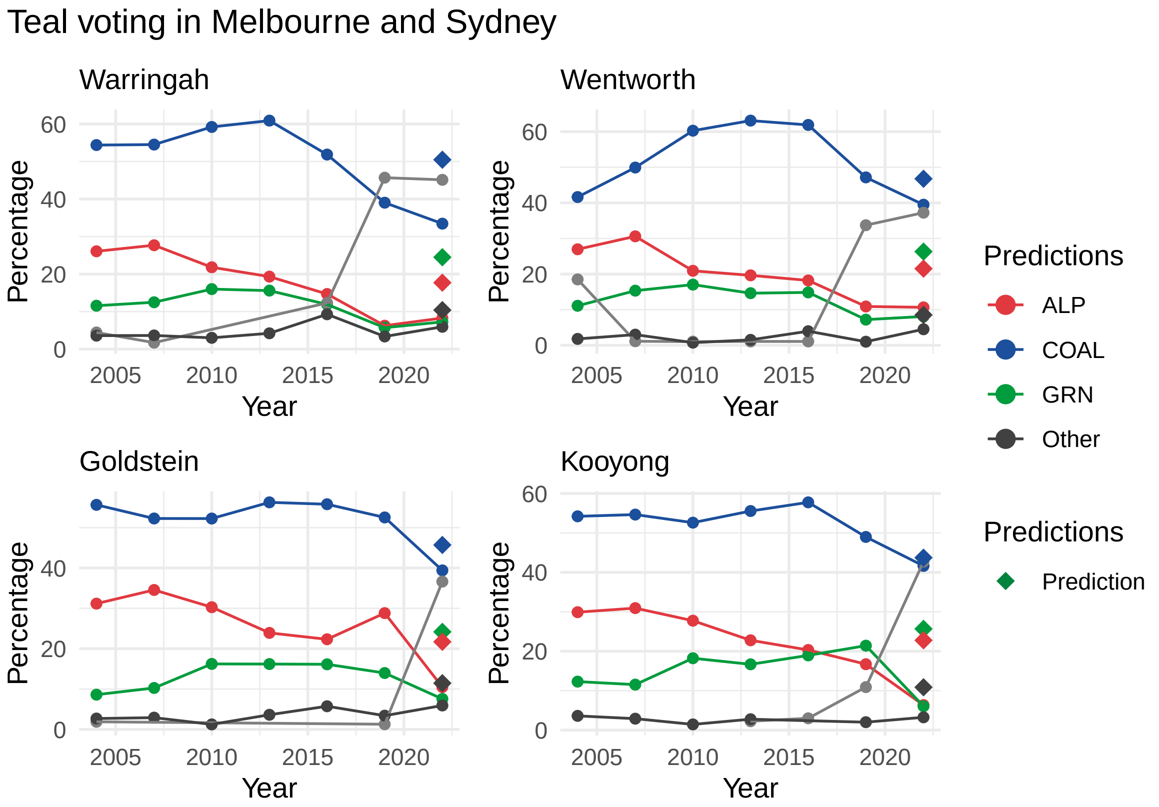


Figure 17: Example 1: Teal Wave

The answer in this case seems to indicate that the dwindling Coalition vote may not be entirely related to a new teal competitor. When comparing these results with demographic statistics from figure 18, these generational change is happening at the same pace or slower than the rest of Australia (shown by flat or growing differences in the Baby Boomer population). The same applies to the percentage of high earners. Nevertheless, the relatively low error in the prediction for the Coalition seems to indicate that the new independents managed to capture Labor and Green voters - likely of a "Labor Right" and "Blue Green" persuasion considering the areas' affluence - rather than attracting a dissatisfied Coalition base.

5.3 The Green Wave

Another feature of the past election was the increase in the number of Green Party MPs. In addition to the division of Melbourne, green candidates also won the seats of Griffith and Ryan in Brisbane. Again,

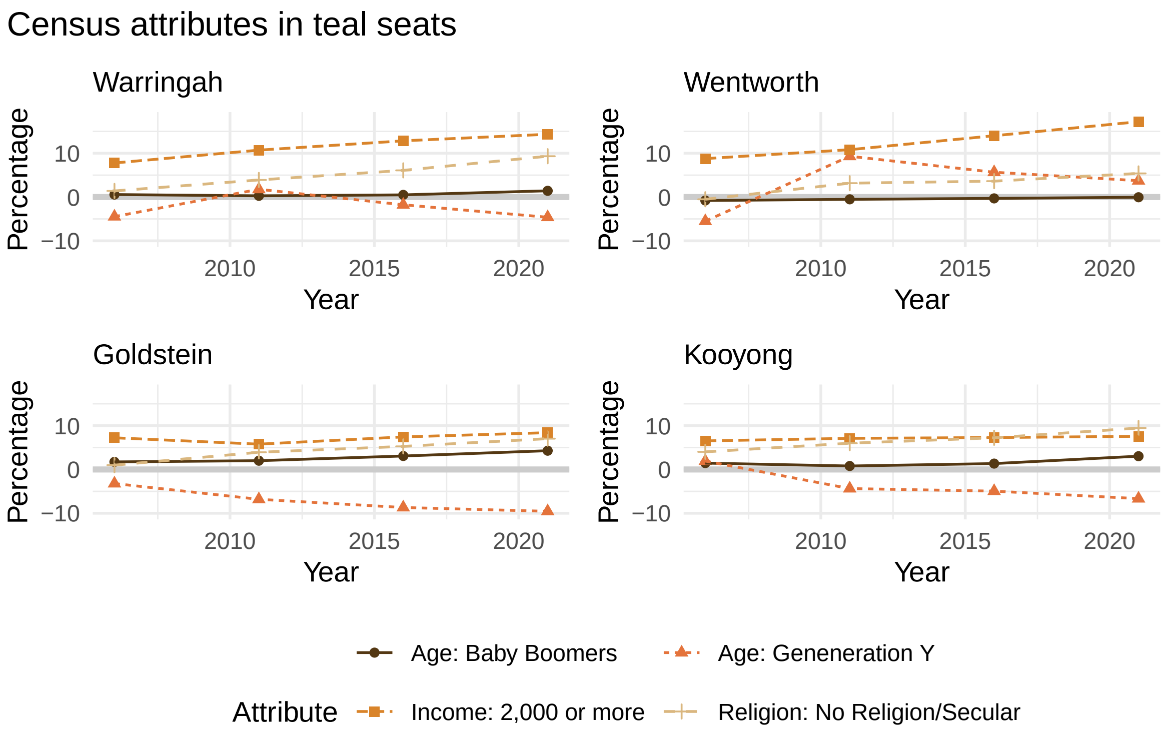


Figure 18: Selected demographics for teal seats

do these victories have a demographic driver? Are there any differences between these electorates and contiguous divisions, and between them and other electorates where the Green have been strong contenders?

Figure 19 shows the prediction of the latest and historic election results. Figure 20 presents selected demographic attributes for those areas.

All four cases show a similar story of continuous growth of the Green vote and progressive decline of Coalition and Labor polling results. However, there are two distinct dynamics at play. In three Queensland electorates (Griffith, Ryan and Brisbane) the Green's growth is sustained in a smaller percentage or older population, being replaced by a younger, wealthier, more secular electorate. In the Victorian seat of Wills, income growth is smaller, generational renewal is slower and although the rise in secularism is faster than average, the area used to have a very high concentration of Catholic followers (Northern suburbs of Melbourne being a popular area amongst post-war Italian migrants). These factors have given Labor a stronger hold in the area.

5.4 The Changing Face of Suburbia

For a comparison outside inner city areas, four suburban electorates have been chosen: Hasluck (Perth), Menzies (Melbourne), Fowler (Sydney) and Kingston (Adelaide). Their respective predictions and results are presented in figure 21. A selection of key demographic variables is presented in figure 22.

From both figures, there are perhaps four different stories in these electorates:

- In **Hasluck** (WA) [19], the changes have the top maybe be driven by generational renewal. The "Other" vote increase includes progressive independents and localist parties, which may have

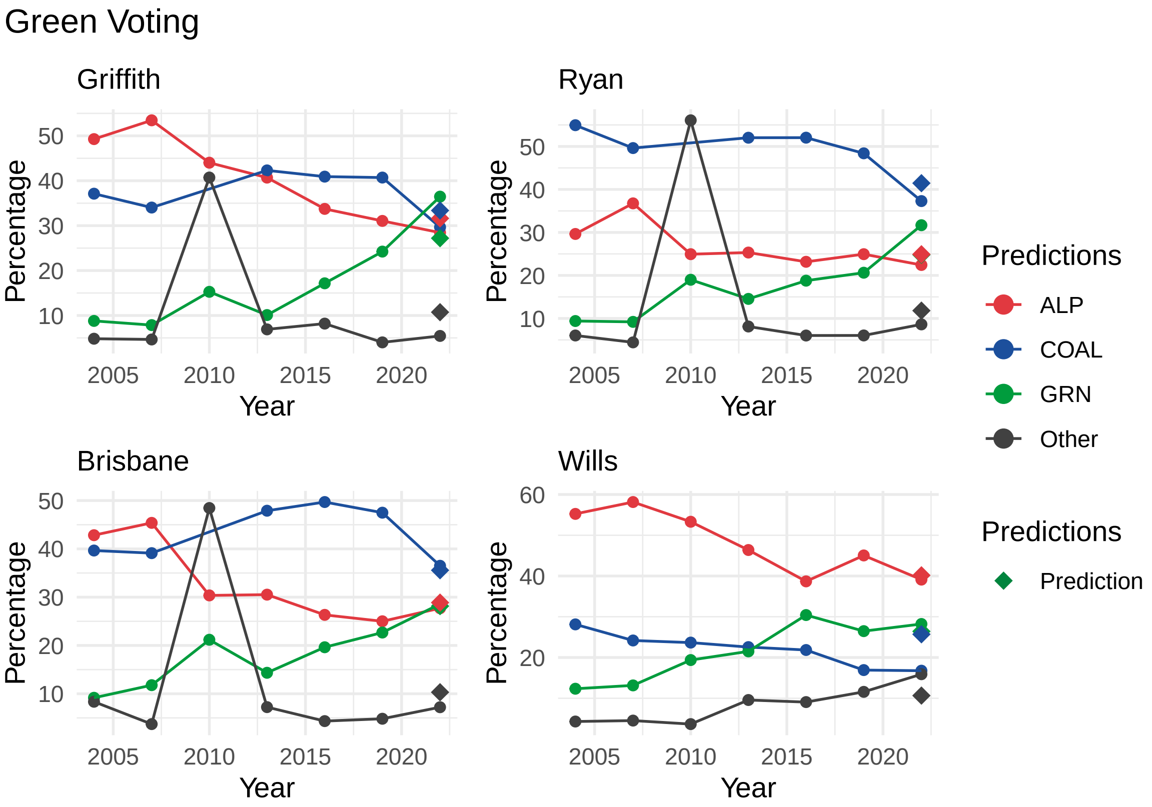


Figure 19: Green Voting

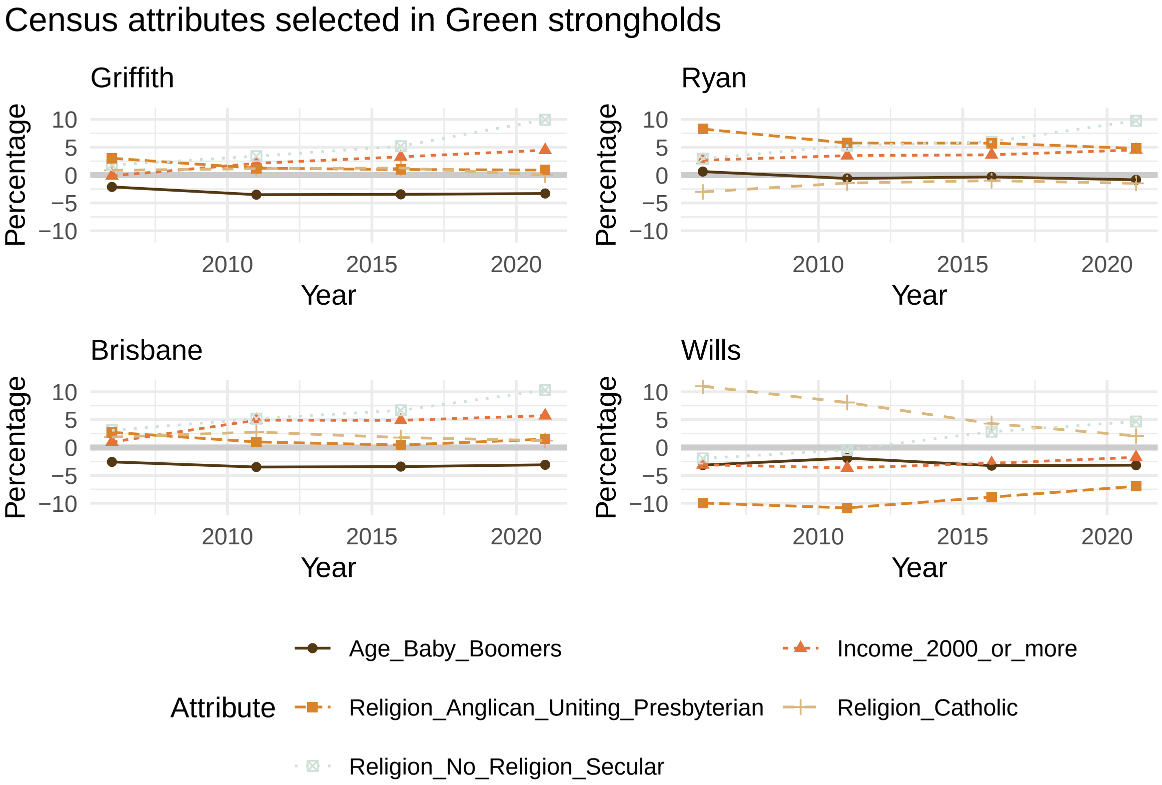


Figure 20: Demographics in Green strongholds

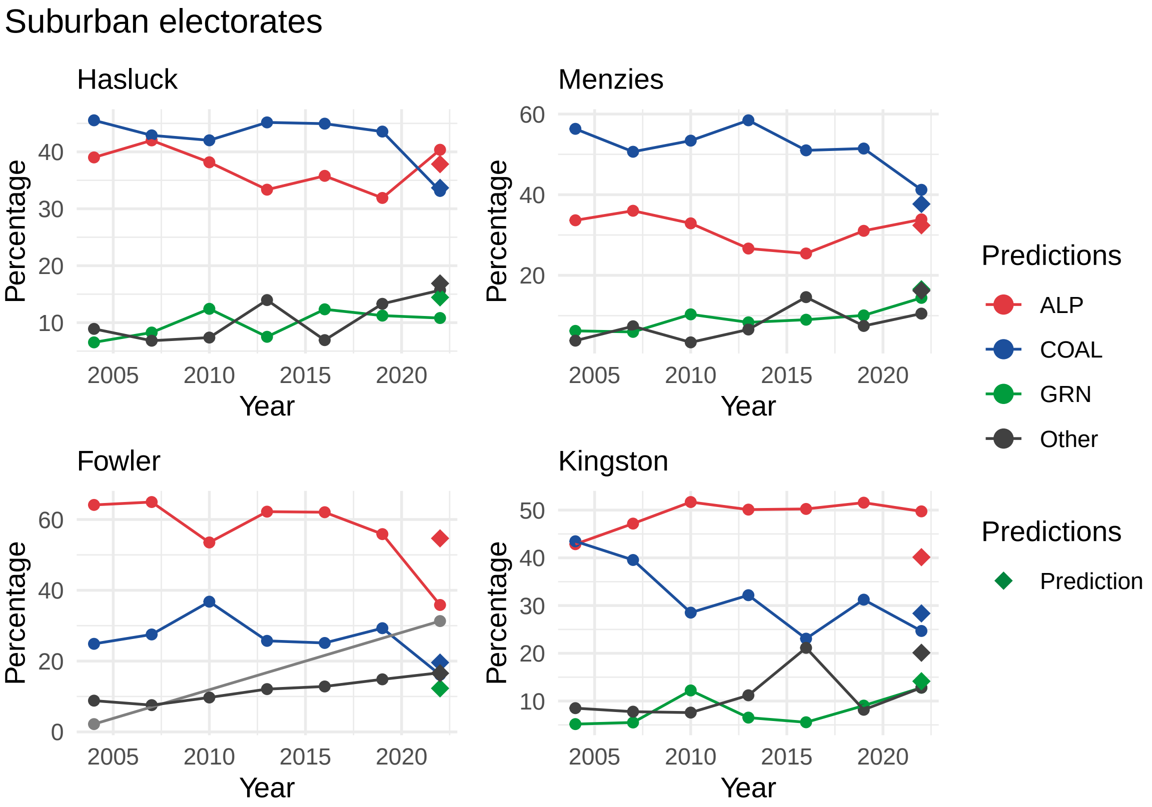


Figure 21: Suburban Voting

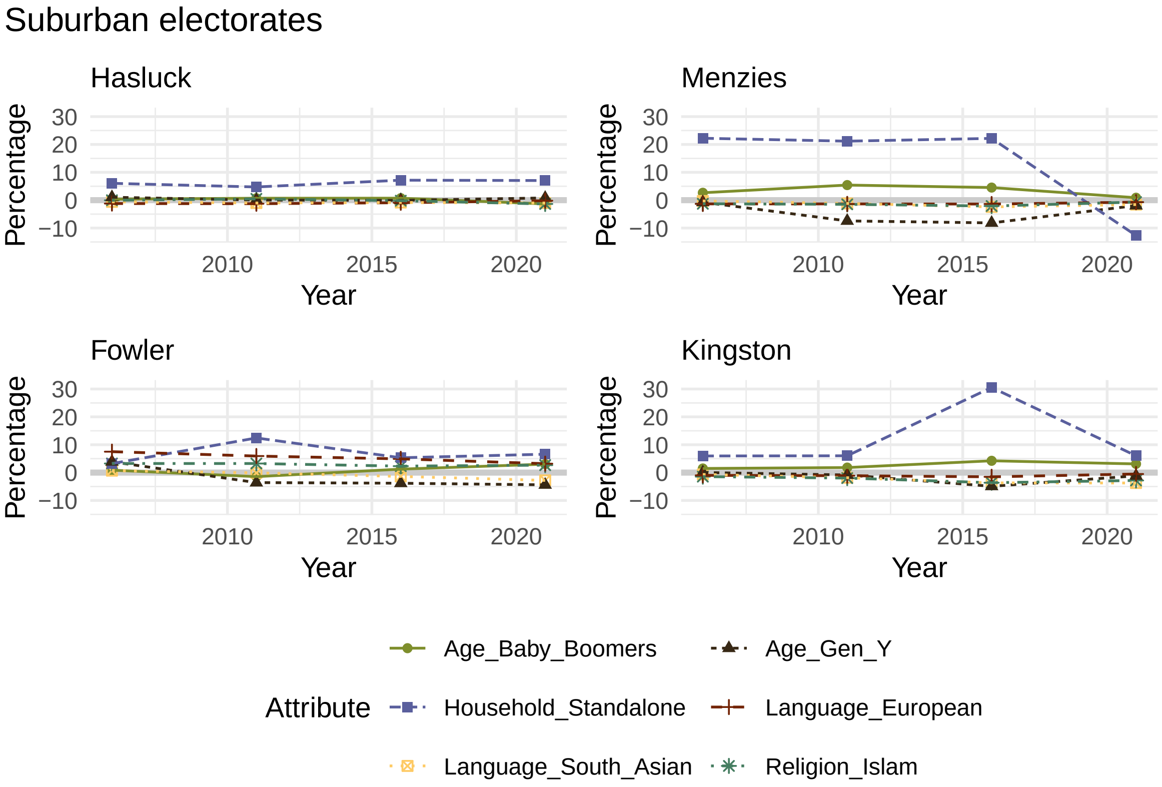


Figure 22: Demographics in suburban seats

influenced the lower-than-predicted results for the Greens.

- In **Menzies** (VIC) [20], Coalition numbers decline influenced by generational change and a large decrease in the percentage of standalone houses. This abrupt change took place as an effect of the 2021 redistribution, where semi-rural areas moved into another electorate [21]. This a good example where the existing model was able to effectively predict the primary vote based on those demographic changes.
- In **Fowler** (NSW) [22], an independent candidate altered Labor's trend. In these cases, a community-based candidate captured the multicultural vote from a "parachuted" Labor nominee - which is a variable not considered in the model. [23]
- In the case of **Kingston (SA) [24]**, it seems difficult to explain the results by effect of demographic changes - even when

6 Conclusion

In summary, this project presents an attempt to understand and explain changes in primary voting through the influence of the demographic composition of Australian federal electorates. Seeking easy interpretability, the approach involved the use of clustering to identify groups of electorates with similar composition, for which simple regularised regression models were developed with the aim of identifying the main demographic drivers of voting.

In general terms, the model presented in this document achieves the goal of identifying key demographic characteristics the affect primary voting for a particular political party.

Although accuracy was not a key consideration, the model managed to produce moderately accurate results. Nevertheless, reasonable improvements may be achieved by exploring some the following:

- Further refining the segmentation into a recommender-type model, where not only similarity clusters are refined by other factors like State and degree of rurality are considered.
- Consider the introduction of a longitudinal element to account for the electorate's history and the influence of incumbency.
- Explore how to address the mismatch between election and census cycles to use the data from all the elections.
- Explore how federal and state elections influence each other.

Taking aside issues regarding the effectiveness of this model, it is also relevant to raise a note of caution about how to interpret the model correctly. By using demographic data is important to keep in mind that certain attributes must be interpreted as proxies of attitudes and values that have an effect on how voters choose. It is very important to make this distinction and avoid statements such as "Community XYZ votes/don't vote for Party A". This is definitely not an aim behind this exercise and it should not be interpreted this way.

Going back to the problem statement presented in section [#problem-statement], it is worth reminding that this project is focused on primary voting only. Even though primary voting is useful to understand general inclinations and may serve as a good based to guess who will be elected, this project does not extend into any analysis of who is eventually elected into Parliament. A natural extension of this work would be to understand how the make-up of a particular electorate influence how voters of a certain party direct their preferences. This is now especially relevant as the Greens and Independents are becoming competitive challengers in many divisions (becoming three-way contests), and many formerly Labor and Coalition electorates have become marginal.

Finally, it is also important to recognise that "all models are wrong but some are useful". Capturing and quantifying human behaviour can be a challenging task, but in this case, having a tool for analysis can prove value for parties, the media and the voters to check the accuracy of political narratives.

Appendix

A Detailed list of variables

The below tables present all variables eventually used in this project. Please note that throughout the analysis and modelling all variables have eventually been transformed into difference in percentage against national or cluster percentages, yet the same basic definitions apply.

A.1 Identification variables

The below variables constitute identify each data point, i.e. they serve as each record's "primary key".

Table 11: ID Variables

Variable	Description
election_year	Year the election was held
DivisionNm	Name of the Commonwealth Electoral Division
StateAb	State of Territory where the Division is located
Metro_Area	Name of Greater Metropolitan Area

A.2 Response Variables

The response variables represent primary voting results.

Table 12: Response Variables

Variable	Party
ALP	Australian Labor Party
COAL	The Coalition, either Liberal or National Party
GRN	Australian Greens
Other	Consolidated results of all other minor parties and independents

A.3 Covariates

The following covariates represent data collected by the Australian Census of Population and Housing. Variable relating to personal attributes are self-declared.

Table 13: Response Variables - citizenship and Metropolitan Flag

Variable	Description
Australian_Citizens	Percentage of People holding Australian citizenship
Metro	Flag indicating if electorates lies within a metropolitan area

Table 14: Response Variables - Age

Variable	Description
Age_Baby_Boomers	Percentage of Baby Boomers (born between 1946 and 1964)
Age_Gen_X	Percentage of Gen X people (born between 1964 and 1980)
Age_Gen_Y	Percentage of Millenials (born between 1981 and 1996)
Age_Gen_Z	Percentage of Gen Z people (born between 1997 and 2012)
Age_Silent_Gen	Percentage of Silent Generation people (born between 1928 and 1945)

Table 15: Response Variables - Language

Variable	Description
Language_Arabic	Percentage of Arabic speakers
Language_Chinese	Percentage of speakers of a Chinese Language (including Mandarin, Cantonese and Others
Language_East_Asian	Percentage of speakers of an East Asian Language (excluding Chinese languages
Language_English_Only	Percentage of monolingual English speakers
Language_Greek	Percentage of Greek speakers
Language_Italian	Percentage of Italian speakers
Language_European	Percentage of speakers of other European Languages
Language_Indigenous	Percentage of speakera of an Australian indigenous language
Language_Indonesian	Percentage of Indonesian speakers
Language_Pacific	Percentage of speakers of a Pacific language
Language_Philippines	Percentage of speakers of a Philippino Language
Language_South_Asian	Percentage of speaks of a South Asian language (including Hindi, Tamil, Urdu, Nepali and others)
Language_Western_Asian	Speaker of a Western Asian Language (excluding Arabic)

Table 16: Response Variables - Religion

Variable	Description
Religion_Anglican_Uniting_Presbyterian	Percentage of followers of Anglicanism, Uniting and Presbyterian churches (combined).
Religion_Buddhism	Percentage of followers of Buddhism, of any denomination
Religion_Catholic	Percentage of followers of the Roman Catholic Church, of any rite
Religion_Christian_Orthodox	Percentage of followers of any Christian Orthodox church (excludes Eastern Rite Catholics)
Religion_Other_Christianity	Percentage of followers of other Christian Churches (other Protestant, Pentecostals, and other Christian)
Religion_Hinduism	Percentage of followers of Hinduism
Religion_Islam	Percantage of followers of Islam, of any denomination
Religion_No_Religion_Secular	Percentage of people self-declared as atheist, non-religious, agnostic or secular

Table 17: Response Variables - Household income

Variable	Description
Income_1_to_999	Percentage of households with a total weekly income between 1 and 999 (Australian) dollars
Income_1000_to_1999	Percentage of households with a total weekly income between 1,000 and 1,999 (Australian) dollars
Income_2000_or_more	Percentage of households with a total weekly income over 2,000 (Australian) dollars
Income_Negative	Percentage of households with a negative total weekly income
Income_Not_Stated	Percentage of households whose total weekly income was not stated

Table 18: Response Variables - Household type

Variable	Description
Household_Flat	Percentage of households that are flats (appartments)
Household_Standalone	Percentage of standalone households
Household_Semi_detached	Percentage of of semi-detached households (townhouses, villas)
Household_Other	Percentage of other types of housing structures

Table 19: Response Variables - Household tenure

Variable	Description
Household_Owned_outright	Percentage of households owned outright
Household_Owned_with_a_mortgage	Percentage of households whose owners hold a mortgage
Household_Rented	Percentage of households rented by their inhabitants
Household_Social_or_community	Percentage of social and community housing
Household_Tenure_other_or_not_stated	Percentage of households, for which the tenure type was not stated

Table 20: Response Variables - Highest Educational Attainment

Variable	Description
Education_Diploma_or_Certificate	Percentage of the population whose highest educational attainment is vocational education
Education_Bachelor	Percentage of the population whose highest educational attainment Bachelor degree
Education_Graduate_Diploma_or_Certificate	Percentage of the population whose highest educational attainment is a graduate diploma or graduate certificate
Education_Postgraduate	Percentage of the population whose highest educational attainment is a postgraduate degree (Master or higher degree)
Education_Not_Stated	Percentage of the population whose highest educational attainment was not stated

Table 21: Relationship in Houselhold

Variable	Description
Relationship_Child_under_15	Percentage of people 15 years of younger
Relationship_De_Facto	Percentage of people in a de facto relationship
Relationship_Group_Household	Percentage of people living in a group household (e.g. flatsharing)
Relationship_Living_Alone	Percentage of people living alone
Relationship_Married	Percentage of married people
Relationship_Single_Parent	Percentage of single parents
Relationship_Non_dependent_Child	Percentage of people with the parents but not dependant (e.g. adult children living in family home)
Relationship_Other_Relationship	Percentage of people with other type of relationship to the rest of the houselhold (e.g. senior parents in child's home)
Relationship_Unrelated_in_family_household	Percentage of people living in family home but unrelated to family groups (e..g lodger)

B R Packages

Although this project is at its core an analytical exercise, the nature of the data sources meant that significant effort was needed to select, requiring the implementation of a consistent framework to find, extract and process. For this reason, it was deemed appropriate to create software packages to separate this from the proper analytical stage of the work. All the code was organised into 3 R packages, namely:

- **{auspol}** Which provides an interface to extract and visualise electoral results for the House of Representatives (https://carlosyanez.github.io/auspol).
- **{auscensus}**, designed as a "Swiss Army knife" tool set to interact with Census Data Packs (https://carlosyanez.github.io/auscensus).
- **{aussiemaps}**, which contains granular maps of Australia and enables geographical aggregation of census data (https://carlosyanez.github.io/aussiemaps).

Although these three packages are slightly different, their basic architecture shares some common concepts:

- Retrieving and storing data.
- Processing, aggregating and presenting results.
- Coding and documentation practices.

B.1 Retrieving, distributing and storing data

For the required data there was no programmatic resource to access the relevant data, or it was deemed inadequate due to the large volumes of initial data required in the project. Most of the data is available for manual download. The large amount of data meant it could not be distributed as part of the package. Such a large amount of data cannot be stored directly on a GitHub repository either. After assessing the options, two different strategies were adopted:

- {aussiemaps} and {auspol} data is stored on a GitHub release. Each package contains functions to download data on demand.

- {auscensus} does not hold any data. Users are required to download Census Data Packs directly from the ABS, either manually or through a built-in function

A persistent cache is set up on the user's computer in all three cases. This was based on the same mechanism developed for the package **{tigris}** [25] (https://github.com/walkerke/tigris). Its code was used and adapted with the author's permission. Each package also contains utilities to manage the amount of space used. A common approach was used in all three cases, where data was download and cache only when needed.

Finally, both {auscensus} and {auspol}'s data required some prior processing. The code to achieve this has been included in their respective **data-raw** folders.

B.2 Processing, aggregating and presenting results

In all three cases, the large volumes of data meant that is not practical to libraries like **readr** to open and process files, due to the large amount of memory resources this operation required when dealing with large data sources (e.g. reading from files with more than 500 attributes for tens of thousands statistical areas). To deal with this, the following strategies were adopted:

- Use of **Arrow** framework [26] to read csv files, allowing to perform 'database-like' operations when reading, processing and aggregating data. This reduced the amount of data loaded in memory on each operation.

- Storing all cached data in **parquet** format. Parquet allows for progressive caching of census data, which then later can be retrieved together.
- Allowing for the packages to cache intermediate results, for faster retrieval of commonly used datasets.
- Use of the **GeoPackage** format [27] to store spatial data, which also allows for efficient out-of-memory selection and aggregation.

Additionally, functions on each package provide an easy way to systematically aggregate and filter data, for example grouping votes from different political parties or providing a partial consolidation of census statistics.

B.3 Coding and documentation practices

For all three packages, some best practices around creating R packages were followed, including:

1. Use functional programming principles when defining functions.
2. Use testing tools like **testthat** [28] to continuously ensure package functionality.
3. Provide adequate in-line documentation using **roxygen2**[29].
4. Write descriptive vignettes, detailing how each package work (selection of vignettes provided as appendices)
5. Provide online documentation, using GitHub Pages to deploy a **pkgdown** [30] package website.

C {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()

C.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.

C.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:

```
get_house_primary_vote(division="Brisbane",year=2022)
```

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

```
get_house_primary_vote(division="Brisbane",year=c(2019,2022))
```

```
get_house_primary_vote(division=c("Brisbane","Perth"),year=c(2019,2022))
```

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

```
get_house_primary_vote(division=c("Brisbane","Perth"),year=c(2019,2022),
    aggregation = TRUE)
```

```
get_house_primary_vote(division=c("Brisbane","Perth"),year=c(2019,2022),
    polling_places = c("Yokine North"))
```

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.:

```
get_house_primary_vote(state=c("TAS"),year=c(2019,2022),aggregation = TRUE)
```

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

```
get_house_primary_vote(state=c("NT"),year=c(2019,2022),aggregation = TRUE,
    party_abb=c("ALP","CLP"))
```

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

```
house_primary_vote_summary(division = "Brisbane", year=2022)
```

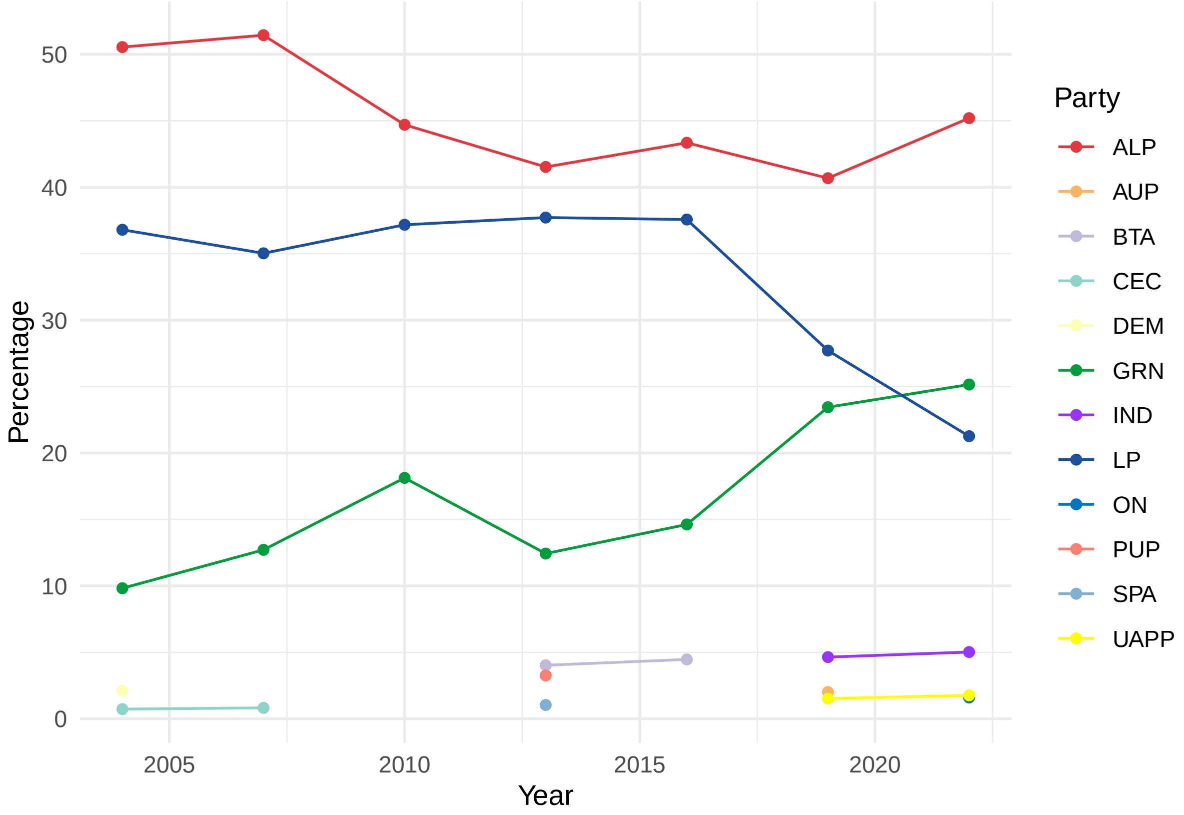
Using the previous filters, it is possible to get ad-hoc summaries, for instance - all the ALP votes in Queensland in 2022, or the historic Liberal vote in Franklin.

```
house_primary_vote_summary(state = "QLD", year=2022,parties="ALP")

house_primary_vote_summary(division="Franklin",parties="LP")
```

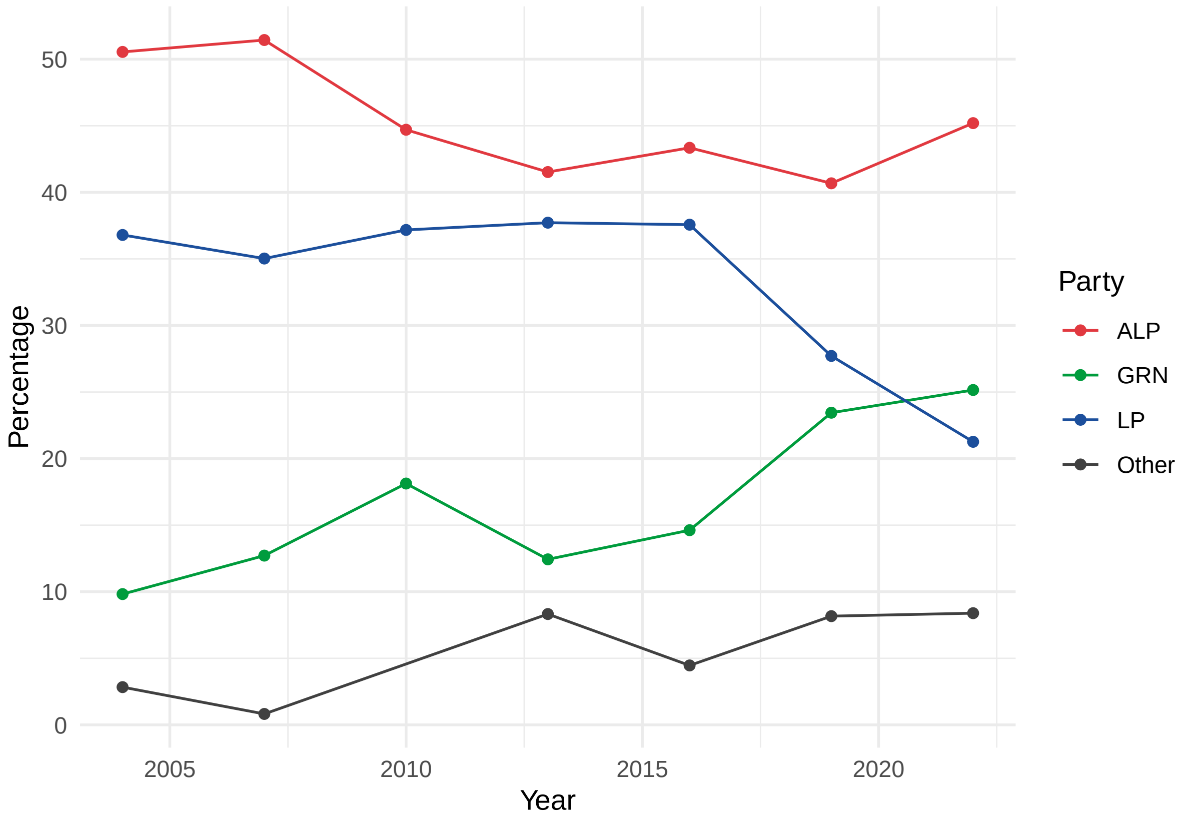
C.3 Plotting

Historic Trends

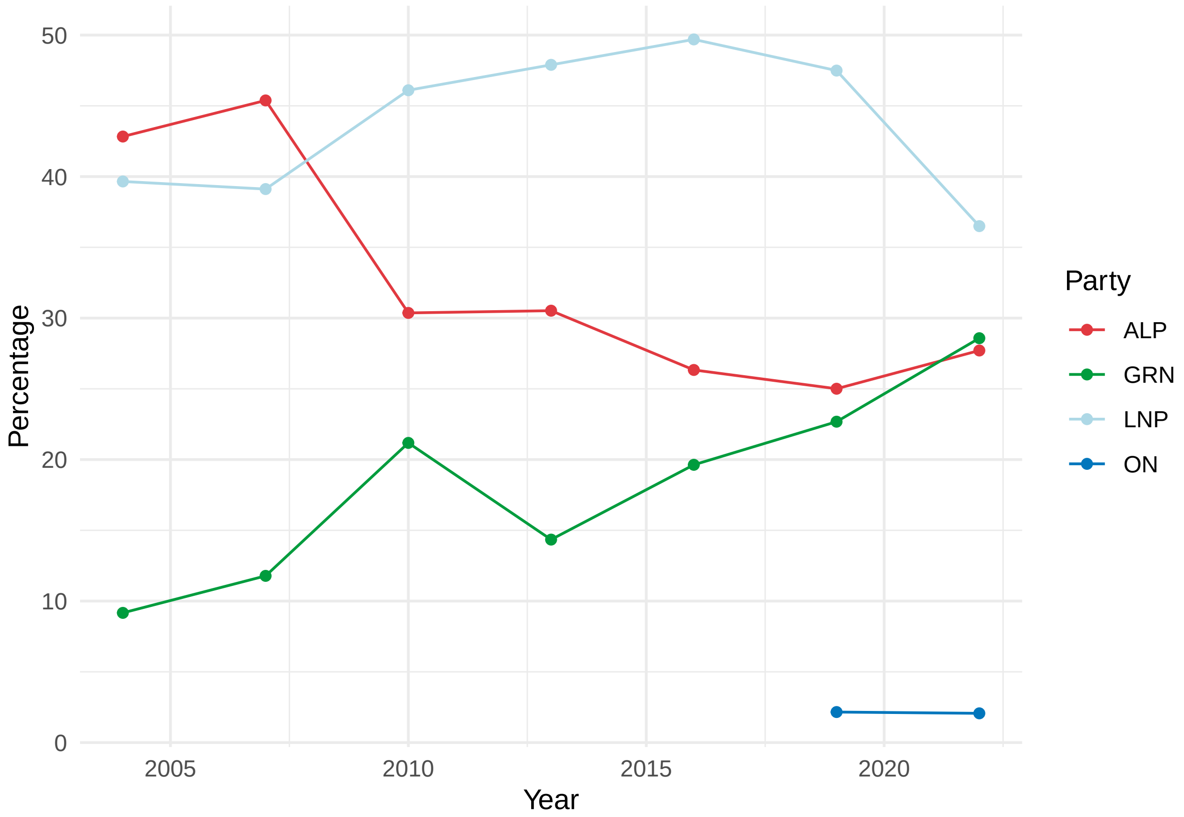
The first plotting convenience function in this package allows comparing the evolution of primary voting across time. This function relies on house_primary_summary and uses many of its options. Its first use is to represent party trends in one electorate:



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.

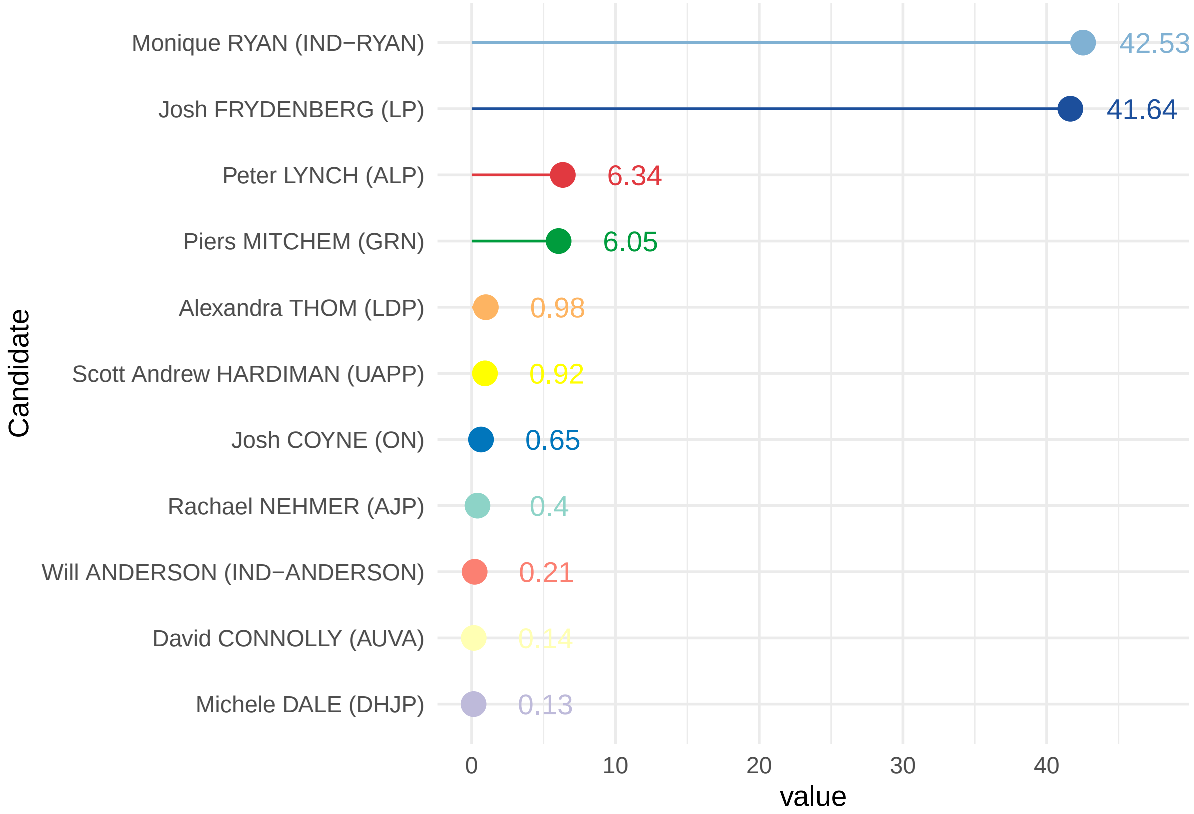


Finally, it is possible to aggregate party acronyms - sometimes the same party has changed named or registered differently

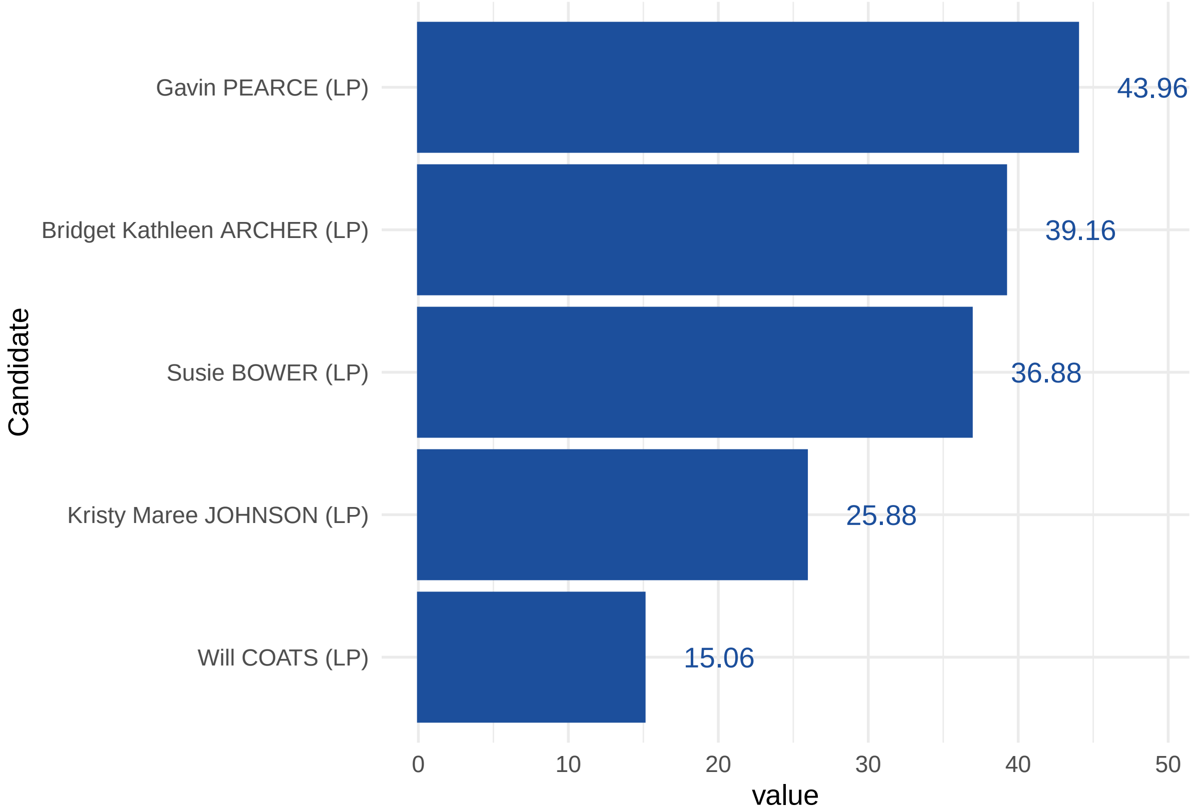


C.4 Results for one election

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)
43.96
39.16
36.88
25.88
15.06
Candidate
0
10
20
30
40
50
value

D {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.

```
census_tables <- list_census_tables()

census_tables <- census_tables |>
                 filter(if_any(any_of(c(”Number”)),~ .x %in% c(”01”)))

tables_summary <- census_tables |>
                  select(-contains(”Table”)) |>
                  pivot_longer(-Number, names_to=”Year”,values_to = ”Value”
                     ) |>
                  filter(!is.na(Value)) |>
                  select(-Value)


list_census_geo_tables(year = tables_summary$Year,geo=”CED|CED_2007|CED_
    2004”,table_number = tables_summary$Number) |>
  mutate(Year=as.character(Year)) |>
  right_join(tables_summary, by=c(”table_number”=”Number”,”Year”=”Year”))
```

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”.

```
list_census_geo_tables(year = tables_summary$Year,geo=c(”SA1|CD”),table_
    number = tables_summary$Number) |>
   mutate(Year=as.character(Year)) |>
  right_join(tables_summary, by=c(”table_number”=”Number”,”Year”=”Year”))
```

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

```
citizenship_attributes <- tibble()

for(i in 1:nrow(tables_summary)){

  table_i <- tables_summary[i,]$Number
  year_i  <- tables_summary[i,]$Year

  attr_i <- list_census_attributes(table_i) |>
              pivot_longer(-c(Table,Attribute),
                           names_to=”Year”,values_to = ”Value”) |>
               filter(Year==year_i) |>
              filter(!is.na(Value)) |>
              select(-Value)

  citizenship_attributes <- bind_rows(attr_i,citizenship_attributes)

}



citizenship_attributes |>
  head()

citizenship_attributes <- citizenship_attributes  |>
                       distinct(Attribute,Year) |>
                      filter(str_detect(Attribute,”[Aa]ustralian”)) |>
                      filter(str_detect(Attribute,”[Pp]erson”)) |>
                      mutate(dummy=TRUE) |>
                      pivot_wider(names_from = Year, values_from = dummy)
                         |>
                      bind_rows(
                         citizenship_attributes  |>
                       distinct(Attribute,Year) |>
                      filter(str_detect(Attribute,”^[Tt]otal”)) |>
                      filter(str_detect(Attribute,”[Pp]ersons$”)) |>
                      mutate(dummy=TRUE) |>
                      pivot_wider(names_from = Year, values_from = dummy)
                      )


citizenship_attributes
```

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

```
citizenship_attributes <- citizenship_attributes |>
                          select(Attribute)      |>
                          mutate(AttrNew = case_when(
                            str_detect(Attribute,”Australian”) ~ ”
                               Australian␣Citizens”,
                            str_detect(Attribute,”Total”)      ~ ”Total”
                          ))

levels <- attribute_tibble_to_list(citizenship_attributes)
```

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:

```
census_years <- c(”2006”,”2011”,”2016”,”2021”)

citizenship <- tibble()

for(year in census_years){
  if(year==”2006”){
    geo_structure_x <- ”CD”
  } else{
    geo_structure_x <- ”SA1”
  }

  citizenship_i <- get_census_summary(census_table = census_tables,
                                      selected_years = year,
                                      geo_structure=geo_structure_x,
                                      attribute = levels)


  citizenship <- bind_rows(citizenship,citizenship_i)


}

rm(citizenship_i,geo_structure_x,levels,citizenship_attributes)
```

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.

```
citizenship_ced<- tibble()
codes <- c(”CD_CODE_2006”,”SA1_7DIGITCODE_2011”,”SA1_7DIGITCODE_2016”,”SA1_CODE_2021”)
ceds  <- c(”CED_NAME_2006”,”CED_NAME_2011”,”CED_NAME_2016”,”CED_NAME_2021”)

 for(i in 1:length(census_years)){

   year <- as.double(census_years[i])

   value_i <- citizenship |>
                   filter(Year==year)       |>
                   select(-Unit)                     |>
                   rename(!!codes[i]:=”Census_Code”) |>
                   collect() |>
                   aussiemaps::geo_aggregate(
                               values_col=”Value”,
                               original_geo=codes[i],
                               new_geo=ceds[i],
                               grouping_col = c(”Year”,”Attribute”),
                               year=census_years[i]) |>
              rename(”Unit”=ceds[i])   |>
              filter(!is.na(Unit))     |>
              filter(str_detect(Unit,”[Ss]hipping”,TRUE))|>
              filter(str_detect(Unit,”[Uu]sual”,TRUE))    |>
              filter(str_detect(Unit,”[Aa]pplicable”,TRUE))
```

```
    value_i <- value_i  |>
              auscensus::calculate_percentage(key_col = "Attribute",
                                              value_col = "Value",
                                              key_value = "Total",
                                              percentage_scale = 100)

   citizenship_ced <- bind_rows(citizenship_ced,value_i)


 }


citizenship_ced
```

E {aussiemaps} Vignette

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

E.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.

E.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:

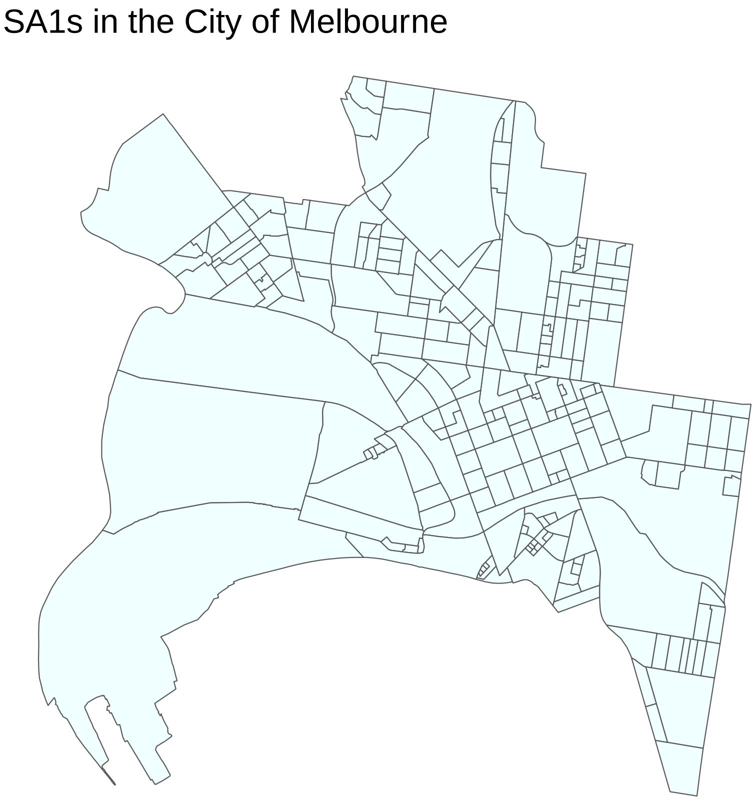
- How the data will be filtered (e.g. return only objects in a particular state, council or metro area)
- Which year/version of the data will be retrieved?
- Which aggregation will be used (e.g. which will be the resulting objects)

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:

```
list_attributes() |>
  head(10)
```

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

```
melbourne_sa1 <- get_map(filters=list(LGA_NAME_2016=c("Melbourne")),
                         year=2016,
                         aggregation = c("SA1_MAINCODE_2016"))

#just an empty plot

melbourne_sa1 |>
  ggplot()+
  geom_sf(fill="azure1") +
  theme_void() +
  labs(title="SA1s␣in␣the␣City␣of␣Melbourne")
```



E.3 Filtering via regular expressions

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

```
preston <- get_map(filters=list(SSC_NAME_2016=c("Preston")),
                        year=2016,
                        aggregation = c("SSC_NAME_2016"))

preston |>
  select(SSC_NAME_2016,UCL_NAME_2016,STE_NAME_2016)
```

```
## 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                         geom
## 1                    Prestons                               Sydney    New
    South Wales POLYGON ((150.8737 -33.9276...
## 2  Preston (Toowoomba - Qld)  Remainder of State/Territory (Qld)
    Queensland POLYGON ((151.9873 -27.6787...
## 3 Preston (Whitsunday - Qld)  Remainder of State/Territory (Qld)
   Queensland POLYGON ((148.6227 -20.3747...
## 4            Preston (Tas.) Remainder of State/Territory (Tas.)
           Tasmania POLYGON ((146.0962 -41.2507...
## 5             South Preston Remainder of State/Territory (Tas.)
           Tasmania POLYGON ((146.0302 -41.3338...
## 6             Preston Beach   Remainder of State/Territory (WA) Western
    Australia POLYGON ((115.6492 -32.8839...
## 7        Preston Settlement   Remainder of State/Territory (WA) Western
    Australia POLYGON ((116.117 -33.40392...
## 8            Preston (Vic.)                            Melbourne
           Victoria POLYGON ((144.9798 -37.7427...
```

Whereas

```
prestons <- get_map(filters=list(SSC_NAME_2016=c("^Pres"),
                                 STE_NAME_2016=c("Wales","^T")
                                 ),
                         year=2016,
                         aggregation = c("SSC_NAME_2016"))

prestons |>
  select(SSC_NAME_2016,UCL_NAME_2016,STE_NAME_2016)
```

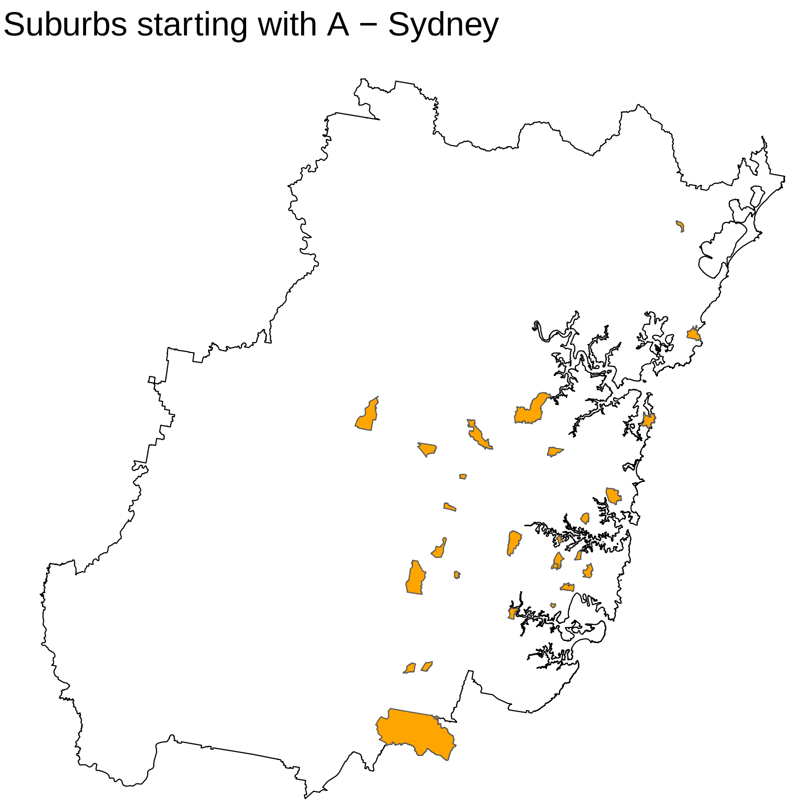
```
## Simple feature collection with 3 features and 3 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 146.0066 ymin: -41.33851 xmax: 150.8979 ymax:
   -33.9263
## Geodetic CRS:  GDA94
##       SSC_NAME_2016                       UCL_NAME_2016    STE_NAME_2016
                                        geom
## 1          Prestons                              Sydney New South Wales
   POLYGON ((150.8737 -33.9276...
## 2 Preservation Bay Remainder of State/Territory (Tas.)         Tasmania
   POLYGON ((146.0401 -41.0973...
## 3   Preston (Tas.) Remainder of State/Territory (Tas.)         Tasmania
   POLYGON ((146.0962 -41.2507...
```

E.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

```
greater_sydney <- list_structure(year=2021,filters=list(GCCSA_NAME_2021="
    Greater␣Sydney"))

#use_cache option stores the results/reuses pre-processed results
```

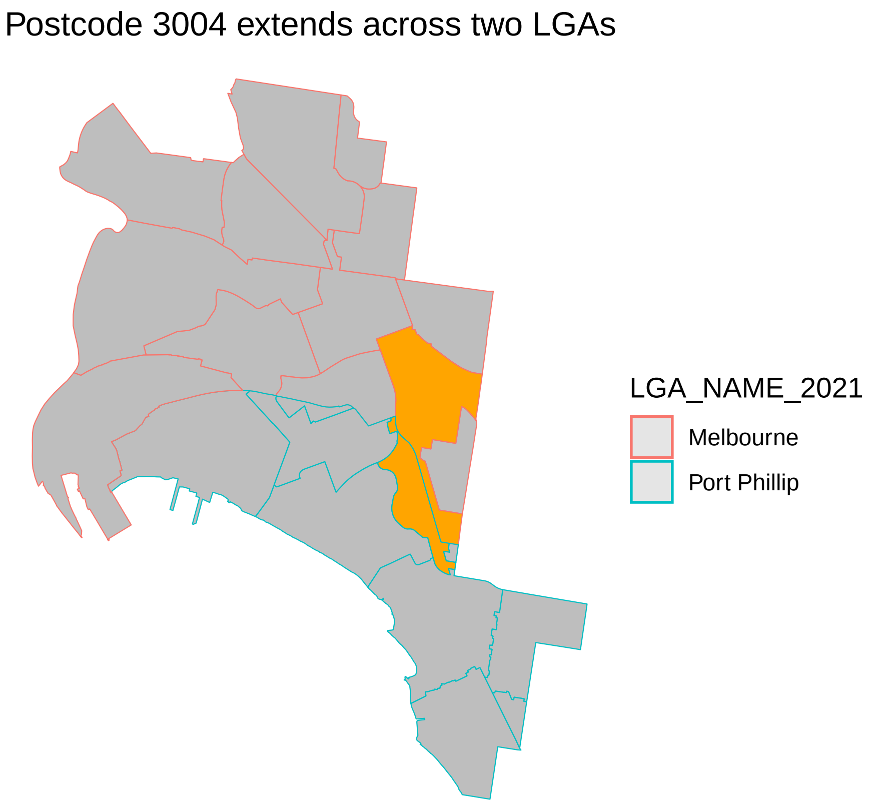
```
sydney_area <- get_map(filter_table = greater_sydney,
                       year=2021,
                       aggregation = ”GCCSA_NAME_2021”,
                       use_cache = TRUE)

## Reading layer `cache_2021_6766fccc' from data source `C:\Users\carlo\
   OneDrive\Documents\.aussiemaps_cache\cache_2021_6766fccc.gpkg' using
   driver `GPKG'
## 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

#outline

# all suburbs starting wit A
suburbs_a_filter <- greater_sydney |>
                    filter(str_detect(SAL_NAME_2021,”^A”))

suburbs_a  <- get_map(filter_table = suburbs_a_filter,
                       year=2021,
                       aggregation = ”SAL_NAME_2021”) |>
                mutate(border=”orange”,fill=”orange”)


ggplot() +
  geom_sf(data=sydney_area,fill=”white”,colour=”black”)+
  geom_sf(data=suburbs_a,fill=”orange”) +
  labs(title=”Suburbs␣starting␣with␣A␣-␣Sydney”) +
  theme_void()
```



E.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:

```
poas_inner_melbourne <- get_map(filters=list(LGA_NAME_2021=c("Melbourne","
    Phillip$")),
                                year=2021,
                                aggregation = c("POA_NAME_2021","LGA_NAME_
                                   2021"))


poas_inner_melbourne |>
  mutate(colour=case_when(
    POA_NAME_2021=="3004" ~ "orange",
    TRUE ~ "grey"
  )) |>
  ggplot()+
  geom_sf(aes(fill=colour,colour=LGA_NAME_2021)) +
  scale_fill_identity() +
  theme_void() +
  labs(title="Postcode␣3004␣extends␣across␣two␣LGAs")
```



## 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:

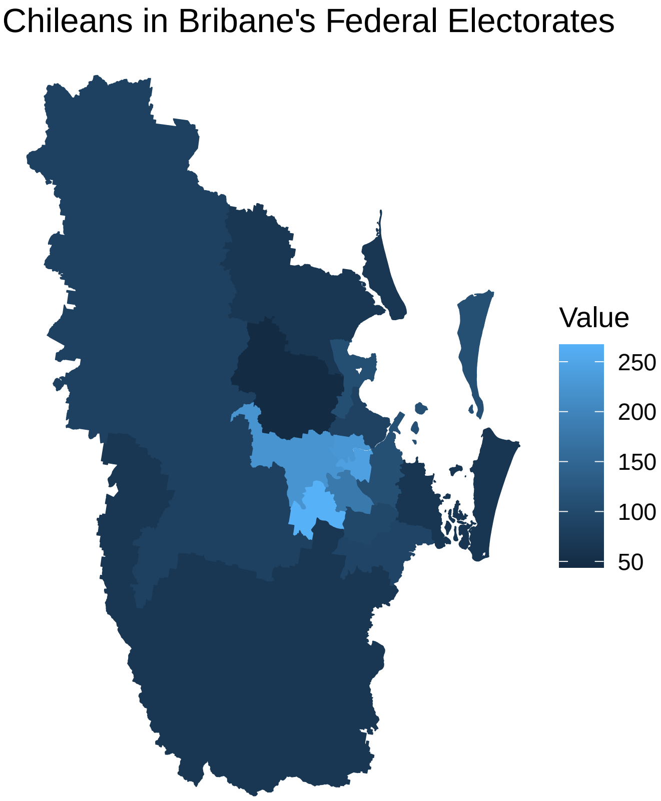
```
# Chileans by Commonwealth Electoral Divisions in Metropolitan Brisbane,
    2021

attr <- list_structure(year=2021,filters=list(GCCSA_NAME_2021=c("Brisbane")
    )) |>
          distinct(CED_NAME_2021)

chileans <- auscensus::get_census_summary(table_number= "09",
                                          selected_years = "2021",
                                          geo_structure = "CED",
                                          geo_unit_names =   attr$CED_NAME_
                                              2021,
                                          attribute = list(Chileans=c("
                                              Persons_chile_total")),
                                          reference_total = list(Total=c("
                                              Persons_total_total")),
                                          percentage_scale =100)

brisbane_ced <- get_map(filters = list(GCCSA_NAME_2021=c("Brisbane")),
                        year = 2021,
                        aggregation = c("CED_NAME_2021"),
                         use_cache = TRUE)
```

```
## Reading layer `cache_2021_4ec18365' from data source `C:\Users\carlo\
    OneDrive\Documents\.aussiemaps_cache\cache_2021_4ec18365.gpkg' using
    driver `GPKG'
## 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
```

```
chileans$Value
```

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

```
brisbane_ced |>
  left_join(chileans,by=c("CED_NAME_2021"="Unit")) |>
  ggplot(aes(fill=Value,colour=Value)) +
  geom_sf() +
  scale_fill_continuous()+
  labs(title="Chileans␣in␣Bribane's␣Federal␣Electorates") +
  theme_void()
```

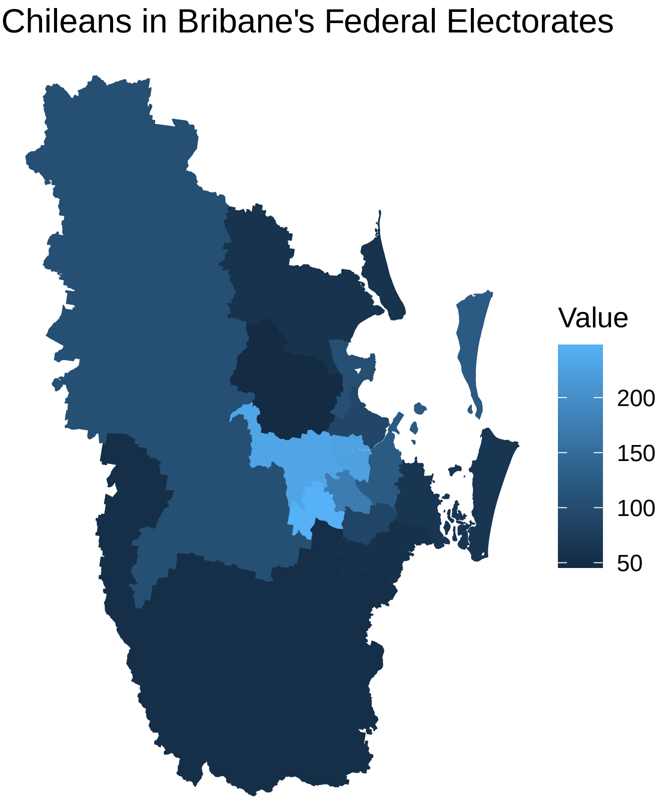


E.6 Data Aggregation

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.

```
attr <- list_structure(year=2021,filters=list(GCCSA_NAME_2021=c("Brisbane")
    )) |>
        distinct(SA2_CODE_2021)

chileans_sa2 <- auscensus::get_census_summary(table_number= "09",
                                              selected_years = "2021",
                                              geo_structure = "SA2",
                                              geo_unit_codes =     attr$SA2_
                                                 CODE_2021,
                                              attribute = list(Chileans=c("
                                                 Persons_chile_total"))) |>
                    rename("SA2_CODE_2021"="Census_Code")

# please note these Electoral divisions are not built from SA2s -
    proportional allocation will result in factional
# Therefore - This is an approximation
chileans <- geo_aggregate(original_data = chileans_sa2,
                          values_col = "Value",
                          original_geo = "SA2_CODE_2021",
                          new_geo      = "CED_NAME_2021",
                          grouping_col =   c("Year","Attribute"),
                          year=2021) |>
            rename("Unit"="CED_NAME_2021")



brisbane_ced |>
  left_join(chileans,by=c("CED_NAME_2021"="Unit")) |>
  ggplot(aes(fill=Value,colour=Value)) +
  geom_sf() +
  scale_fill_continuous()+
  labs(title="Chileans␣in␣Bribane's␣Federal␣Electorates") +
  theme_void()
```


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

Bibliography

[1] N. Biddle and I. McAllister, “Explaining the 2022 Australian federal election result,” Jun. 2022. Available: https://apo.org.au/node/318286

[2] Commonwealth Parliament, “Voting patterns by generation.” Available: https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/FlagPost/2022/April/Voting_patterns_by_generation

[3] A. Jakubowicz and C. Ho, “Was there an ’ethnic vote’ in the 2019 election and did it make a difference?” Available: http://theconversation.com/was-there-an-ethnic-vote-in-the-2019-election-and-did-it-make-a-difference-117911

[4] Australian Broadcasting Corporation, “Inner-city raving lunatics: Michael McCormack on the greens and climate change.” Nov. 11, 2019. Available: https://www.abc.net.au/news/2019-11-11/inner-city-raving-lunatics:-michael-mccormack-on-greens/11694044

[5] Sidney Morning Herald, “‘We have two australias’: Election results show a growing divide within the nation.” Available: https://www.smh.com.au/federal-election-2019/we-have-two-australias-election-results-show-a-growing-divide-within-the-nation-20190524-p51qu8.html

[6] C. Long, “‘Bandaids on bullet holes’: Why labor snubs younger australians at its peril,” *ABC News*, May 2023, Available: https://www.abc.net.au/news/2023-05-11/budget-younger-australians-housing-jobseeker-climate-change/102333220

[7] Sidney Morning Herald, “Does the inner city left-wing elite exist?” Available: https://www.smh.com.au/culture/tv-and-radio/does-the-abc-s-inner-city-left-wing-elite-exist-20201026-p568r0.html

[8] C. Wahlquist, “Teal independents: Who are they and how did they upend Australia’s election?” *The Guardian*, May 2022, Accessed: Jun. 26, 2023. [Online]. Available: https://www.theguardian.com/australia-news/2022/may/23/teal-independents-who-are-they-how-did-they-upend-australia-election

[9] Commonwealth Parliament, “Voting patterns by generation.” Accessed: Jun. 26, 2023. [Online]. Available: https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/FlagPost/2022/April/Voting_patterns_by_generation

[10] S. Jackman, “Millennials and gen Z have deserted the Coalition – this could be dire for the opposition | Simon Jackman | The Guardian.” Accessed: Jun. 26, 2023. [Online]. Available: https://www.theguardian.com/commentisfree/2022/dec/05/millennials-and-gen-z-have-deserted-the-coalition-this-could-be-dire-for-the-opposition

[11] Australian Electoral Commission, “Tally room archive,” 2023. https://results.aec.gov.au/ (accessed Jun. 01, 2021).

[12] Australian Bureau of Statistics, “Census data packages,” 2023. https://abs.gov.au/census/find-census-data/datapacks/ (accessed Jun. 01, 2021).

[13] C. Yáñez Santibáñez, *Auspol: Australian federal election results (2004-2022)*. 2023. Available: https://carlosyanez.github.io/auspol/

[14] C. Yáñez Santibáñez, *Auscensus: Access australian census data (2006-2021)*. 2023. Available: https://carlosyanez.github.io/auscensus/

[15] C. Yáñez Santibáñez, *Aussiemaps: Maps of australia*. 2023. Available: https://carlosyanez.github.io/aussiemaps/

[16] B. Escofier and J. Pagès, *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation*. in Sciences sup. Dunod, 2008, pp. 328 p. Available: https://hal.science/hal-00382085

[17] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 160–172. doi: 10.1007/978-3-642-37456-2_14.

[18] S. Benson, "Newspoll: Labor lead over coalition narrows," 2023. https://www.theaustralian.com.au/nation/politics/newspoll-labor-lead-over-coalition-narrows/news-story/937dbfe8479e9380d93da4121f63c09d (accessed May 20, 2022).

[19] Australian Broadcasting Corporation, "Hasluck (key seat) - federal election 2022," 2022. https://www.abc.net.au/news/elections/federal/2022/guide/hasl (accessed Jun. 01, 2022).

[20] Australian Broadcasting Corporation, "Menzies - federal election 2022," 2022. https://www.abc.net.au/news/elections/federal/2022/guide/menz (accessed Jun. 01, 2022).

[21] Australian Electoral Commission, "Proposed redistribution for victoria." 2021. Available: https://www.aec.gov.au/Electorates/Redistributions/2021/vic/proposed-redistribution/index.htm

[22] Australian Broadcasting Corporation, "Fowler - federal election 2022," 2022. https://www.abc.net.au/news/elections/federal/2022/guide/fowl (accessed Jun. 01, 2022).

[23] C. Hanrahan, "Labor was wiped out in this sydney seat — these charts show how it happened," *ABC News*, May 2022, Available: https://www.abc.net.au/news/2022-05-25/charting-independent-dai-le-win-over-kristina-keneally/101095794

[24] Australian Broadcasting Corporation, "Kingston - federal election 2022," 2022. https://www.abc.net.au/news/elections/federal/2022/guide/king (accessed Jun. 01, 2022).

[25] K. Walker, *Tigris: Load census TIGER/line shapefiles*. 2023. Available: https://CRAN.R-project.org/package=tigris

[26] N. Richardson *et al.*, *Arrow: Integration to 'apache' 'arrow'*. 2023.

[27] "OGC GeoPackage." Accessed: Jun. 25, 2023. [Online]. Available: https://www.geopackage.org/

[28] H. Wickham, "Testthat: Get started with testing," vol. 3, 2011, Available: https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

[29] H. Wickham, P. Danenberg, G. Csárdi, and M. Eugster, *roxygen2: In-line documentation for r*. 2022. Available: https://CRAN.R-project.org/package=roxygen2

[30] H. Wickham, J. Hesselberth, and M. Salmon, *Pkgdown: Make static HTML documentation for a package*. 2022. Available: https://CRAN.R-project.org/package=pkgdown