DATA ANALYTICS GLASGOW

University of Glasgow

School of Mathematics and Statistics

Data Analytics MSc (ODL)

# My Course

# Analysing Australian Election Results

Carlos Yanez Santibanez

'r Sys.Date()'

**Acknowledgements**

## Contents

**Abstract**

# 1 Introduction

# 2 Data

## 2.1 Data Sources

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

- The **Australian Electoral Commission** (AEC) [Commission, 2023a] . 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 [Commission, 2023b].
- The **Australian Bureau of Statistics** (ABS) [of Statistics, 2023a]. The ABS provides a wide number of national statistics and is responsible to conduct a national census of population and housing every 5 years. Comprehensive census data is provided in multiple formats, including CSV files through Census Data Packs [of Statistics, 2023b], 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} [Yáñez Santibáñez, 2023b]**, which extracts and presents electoral results.
- **{auscensus} [Yáñez Santibáñez, 2023a]**, which allows to interact with Census Data Packs to extract different statistics across geographical units, and across censuses.
- **{aussiemaps} [Yáñez Santibáñez, 2023c]**, which assists with aggregating census data into electoral divisions, by matching and apportioning different geographical structures.

The way each package operates is described on their respective websites Using them, it was possible to build a basic data extraction and transformation pipeline, which is represented by figure 1.

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 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.

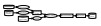

Figure 1: Flow of data from sources to dataset

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.

## 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. 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 (between 72,345 and 138,836 voters). 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 2 presents the best matches between both events held in the 21st century.

Considering the census data available and selecting the elections closer to each census, four sets of events were selected for data extraction. there 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.

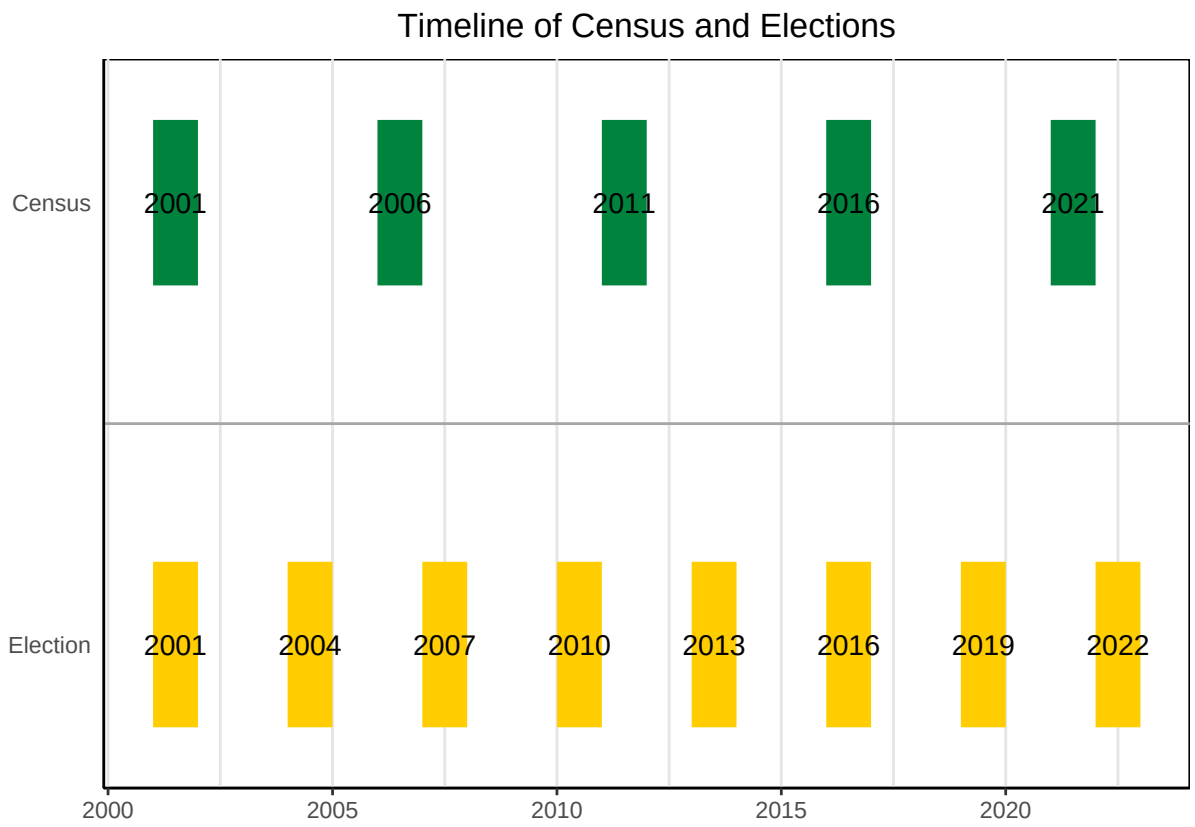


Figure 2: Census and Elections Timeline

- **COAL** representing the Coalition, made of the Liberal Party, the National Party, the Liberal National Party of Queensland [1] and the Country Liberal Party in the Northern Territory [2] .
- **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

[1] 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.

[2] 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.

release for the 2022 Census contains 62 different tables, ranging from 8 [3] to 1,590 [4] 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 ([Biddle and McAllister, 2022], [Parliament], [Jakubowicz and Ho, a]). 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).
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 [Biddle and McAllister, 2022] 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 [Jakubowicz and Ho, b]).

A sample of the resulting dataset is present in table 3.

[3] *02 -Selected Medians and Averages*
[4] *09 - Country of Birth of Person by Age by Sex*

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 |
| 2010 | Bonner | 84.58 | 16.24 | 22.48 | 24.25 | 2.74 | 0.67 | 4.52 | 7.19 |
| 2007 | Canberra | 90.88 | 27.94 | 23.34 | 17.45 | 1.40 | 0.93 | 4.55 | 6.45 |
| 2010 | Cunningham | 87.77 | 17.21 | 21.00 | 22.61 | 2.11 | 1.90 | 4.53 | 6.32 |
| 2016 | Deakin | 82.43 | 18.20 | 22.41 | 22.55 | 10.59 | 1.05 | 4.18 | 8.55 |
| 2016 | Dickson | 88.33 | 17.64 | 20.46 | 20.69 | 0.54 | 0.31 | 4.27 | 8.40 |
| 2007 | Fowler | 87.37 | 25.65 | 26.02 | 24.54 | 10.30 | 2.60 | 6.64 | 0.55 |
| 2022 | Fowler | 78.18 | 20.20 | 22.19 | 33.77 | 7.46 | 1.62 | 6.25 | 4.42 |
| 2016 | Fremantle | 80.23 | 16.69 | 24.99 | 21.81 | 2.68 | 2.80 | 4.35 | 11.79 |
| 2010 | Gilmore | 90.55 | 23.91 | 15.83 | 20.18 | 0.26 | 0.58 | 4.93 | 3.17 |
| 2016 | Gorton | 81.89 | 14.04 | 25.61 | 19.93 | 3.02 | 1.47 | 5.07 | 3.91 |
| 2007 | Greenway | 86.67 | 24.30 | 23.64 | 21.00 | 2.74 | 1.21 | 4.71 | 2.00 |
| 2007 | Higgins | 81.17 | 24.17 | 21.03 | 33.59 | 4.32 | 1.44 | 3.05 | 11.72 |
| 2016 | Indi | 89.29 | 24.85 | 16.45 | 20.53 | 0.30 | 0.87 | 4.49 | 3.88 |
| 2010 | Lilley | 85.16 | 15.90 | 24.83 | 27.59 | 1.69 | 1.31 | 4.72 | 6.27 |
| 2007 | Lindsay | 87.94 | 24.68 | 24.05 | 21.98 | 0.64 | 0.70 | 4.96 | 2.32 |
| 2022 | Macnamara | 74.68 | 16.78 | 35.50 | 47.99 | 5.14 | 1.04 | 3.00 | 24.70 |
| 2007 | Mallee | 92.21 | 26.08 | 20.41 | 19.13 | 0.18 | 1.64 | 3.98 | 1.21 |
| 2007 | Moore | 83.01 | 29.84 | 24.99 | 17.60 | 1.03 | 0.77 | 3.91 | 5.20 |
| 2010 | Moreton | 77.50 | 14.70 | 25.77 | 28.85 | 12.27 | 0.46 | 4.15 | 6.19 |
| 2022 | Moreton | 78.74 | 16.38 | 26.76 | 31.51 | 13.95 | 0.26 | 4.07 | 13.48 |
| 2010 | Reid | 72.48 | 13.42 | 29.83 | 34.02 | 15.34 | 4.66 | 3.85 | 8.60 |
| 2016 | Richmond | 86.50 | 25.92 | 15.90 | 24.60 | 0.29 | 0.27 | 5.40 | 4.41 |
| 2007 | Robertson | 87.34 | 26.20 | 20.02 | 22.51 | 0.39 | 0.36 | 5.17 | 2.59 |
| 2022 | Ryan | 85.57 | 16.32 | 20.77 | 29.11 | 4.78 | 0.36 | 3.51 | 21.24 |
| 2010 | Shortland | 92.31 | 21.59 | 17.32 | 18.75 | 0.35 | 0.34 | 5.48 | 4.58 |

### 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 55.00 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

As an example, figure 3 shows a somewhat weak correlation between Coalition primary vote and the percentage of the Baby Boomers. Figure 4 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.

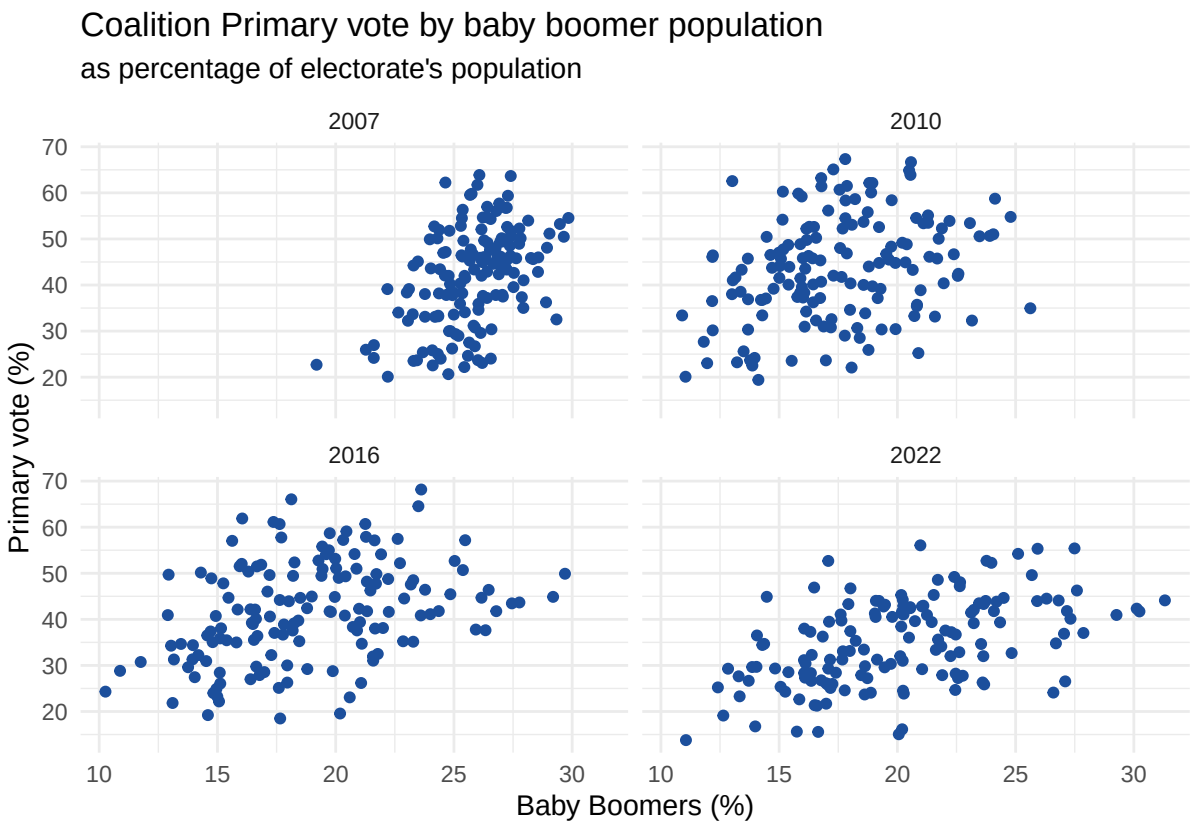


Figure 3: Correlation between Coalition vote and Baby boomer population

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.

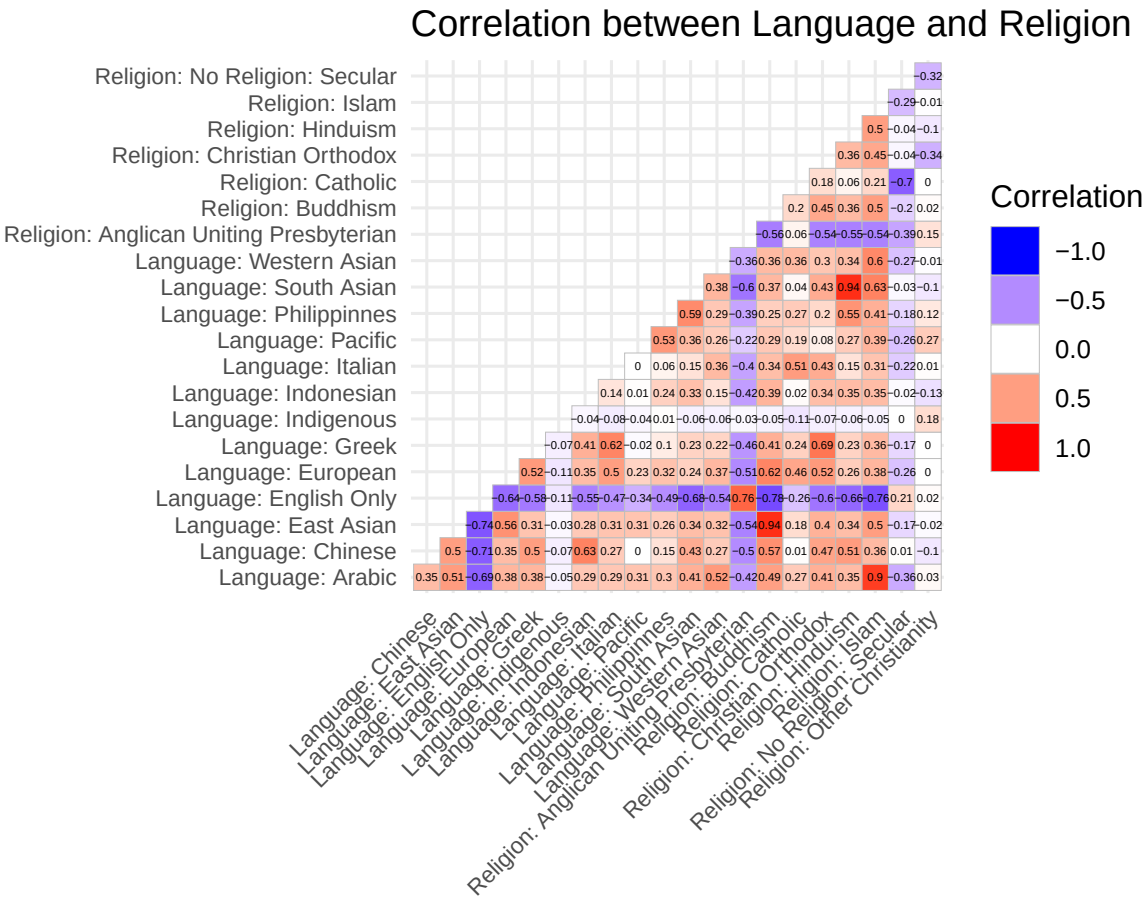


Figure 4: Correlation for selected covariates

**Dimensionality reduction using Multiple Factor Analysis**

Given the large number of colinear covariates, it is worth exploring if a change of space could help to better measure variation in a meaningful way and in a more manageable number. To achieve this, **multiple factor analysis* (MFA) [Escofier and Pagès, 2008] was used as the clustering algorithm. MFA is essentially an extension of Principal Component Analysis that can deal with variables that belong to groups (like this case). It can also combine quantitative and qualitative variables (such as belonging to a metropolitan area).

The resulting scree plot and cumulative variance are presented in figure 5.

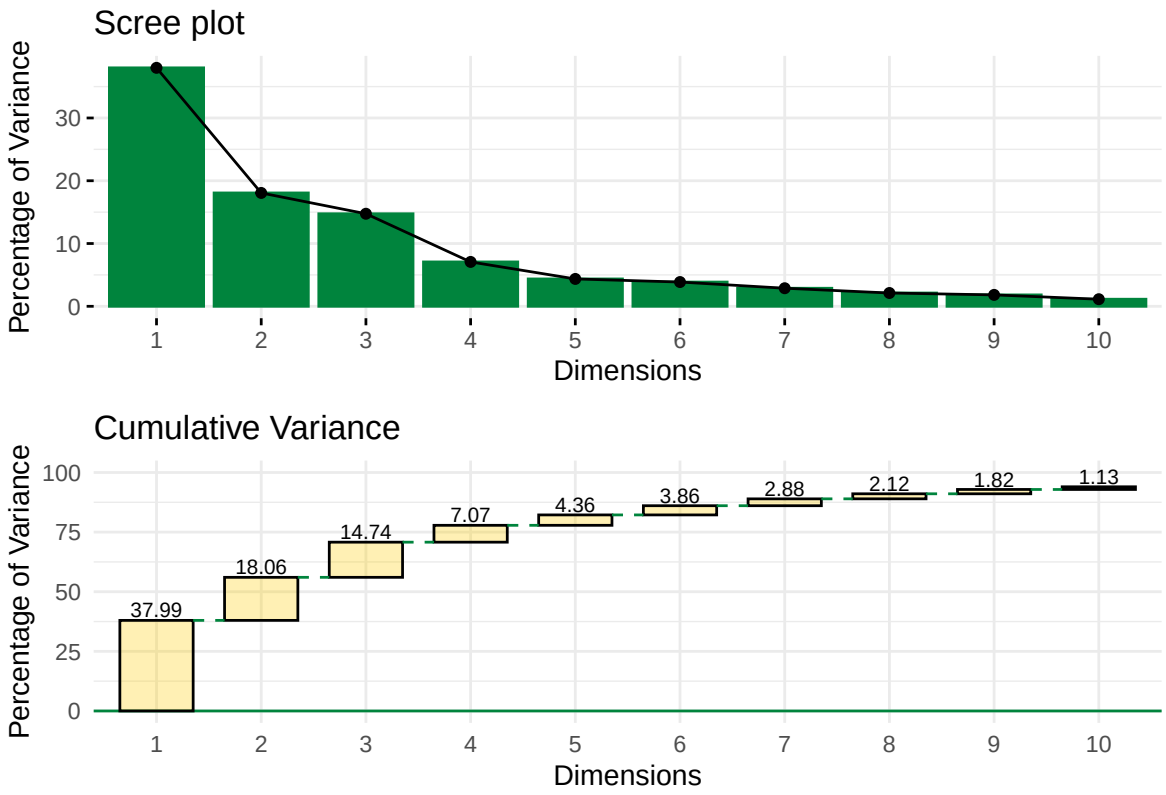


Figure 5: Scree plot and cumulative variance

Figure 6 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.

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.

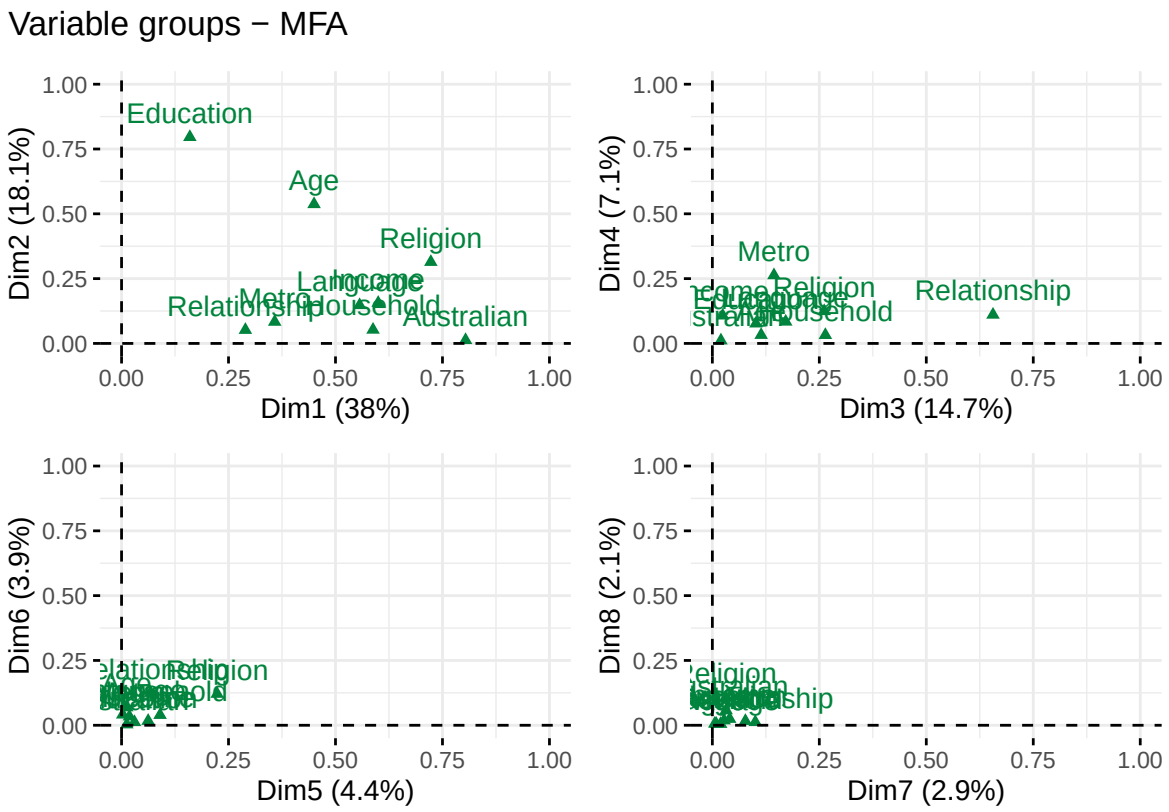


Figure 6: Group plots for first 8 dimensions

- transform all demographic attributes to represent the difference between each data point and their corresponding national value (in the same year).
- use HDBSCAN [Campello et al., 2013], a density-based hierarchical clustering algorithm. Instead of pre-setting a target number of clusters, HDBSCAN determines the optimal number of clusters based on its tuning parameters.

This results in 3 distinct clusters of electorates. When presented in a map, it is possible to obtain figure 7 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.

Revisiting the demographic attributes can help to understand how these clusters differ from each other. A selection of those variables is presented in figure 8.

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

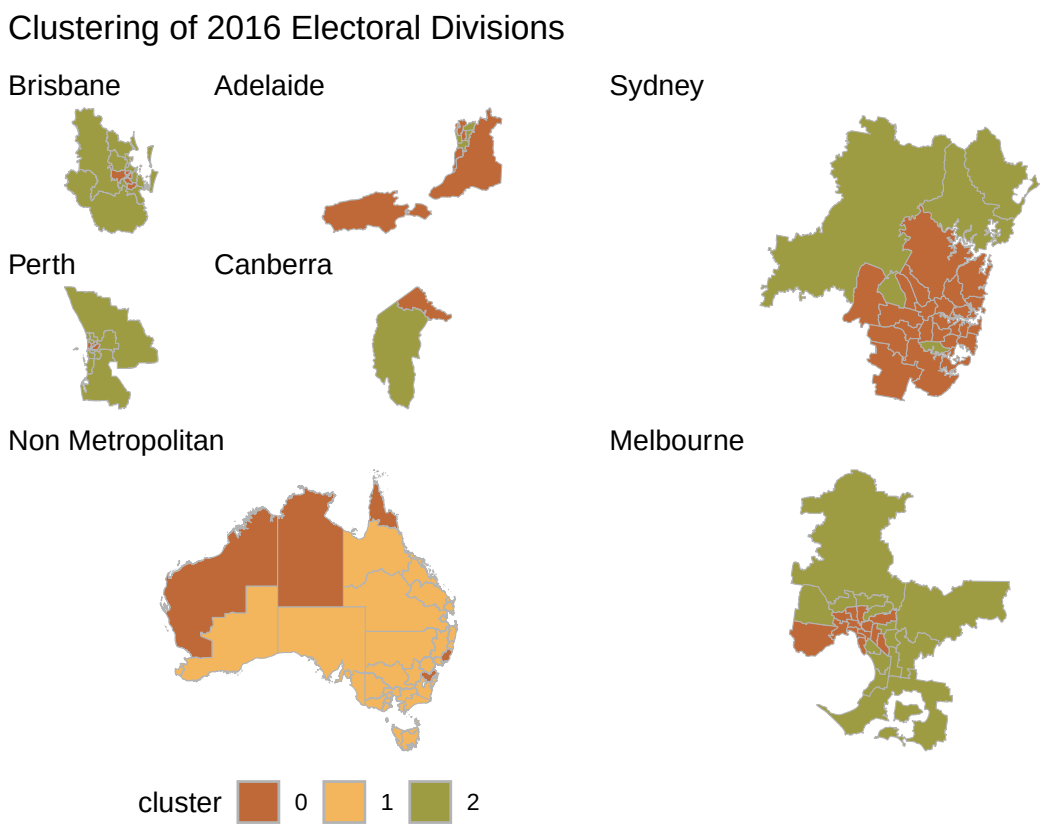


Figure 7: Clusters for 2016 Election

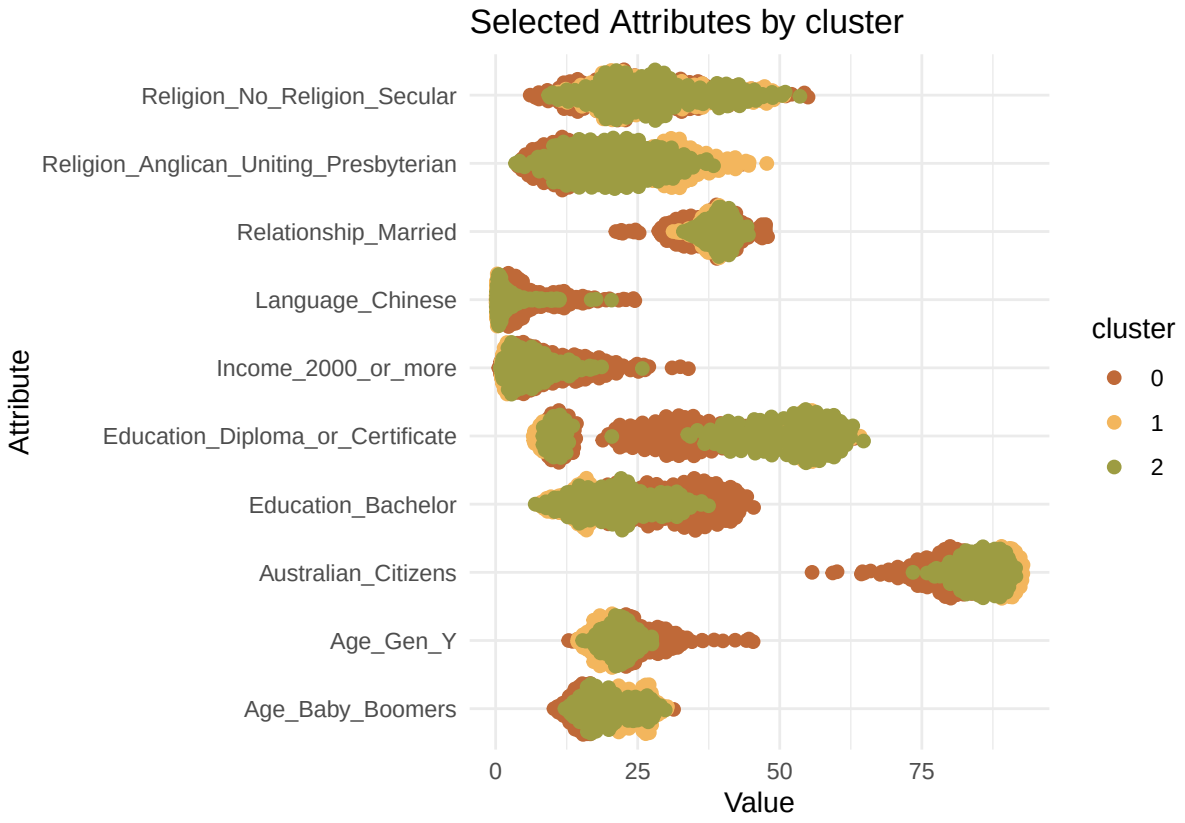


Figure 8: Selected attributes, coloured by cluster.

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 S 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 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 (10), which also accounts for general error.

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

$$\tag{3}$$

where (4)

$$\mathbf{Y} = \mathbf{Y} - \mathbf{Yp} \tag{5}$$

$$\mathbf{X} = \mathbf{X} - \mathbf{Xn} \tag{6}$$

$$\tag{7}$$

and (8)

**Yp** : Primary voting polling results (9)

**Xn** : Demographic values at national level (10)

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 (16) is intended to acknowledge those differences.

$$\mathbf{Y_i} = f_i(\mathbf{X_i}) + \epsilon \tag{11}$$

where (12)

$$\mathbf{Y} = \mathbf{Y} - \mathbf{Yp_i} \tag{13}$$

$$\mathbf{X} = \mathbf{X} - \mathbf{Xn_i} \tag{14}$$

(15)

(16)

Please note 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 (which is attempted nevertheless), for which the use of **regularised regression** models is an appropriate choice.

Consequently, the task at hand consists in finding the regularised regression coefficients for the set of formulas represented by equation (17)

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{in} \end{pmatrix} = \begin{pmatrix} \beta_{i11} & \beta_{i12} & \dots & \beta_{i1m} \\ \beta_{i21} & \beta_{i22} & \dots & \beta_{i2m} \\ \dots & \dots & \dots & \dots \\ \beta_{in1} & \beta_{in2} & \dots & \beta_{inm} \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{im} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \dots \\ \epsilon_{in} \end{pmatrix} \tag{17}$$

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. The algorithm 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 9

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 10.

Table 6: Improved Classification Model - Metrics

| Metric | Estimate |
|---|---|
| Accuracy | 0.8667 |
| ROC AUC | 0.9531 |

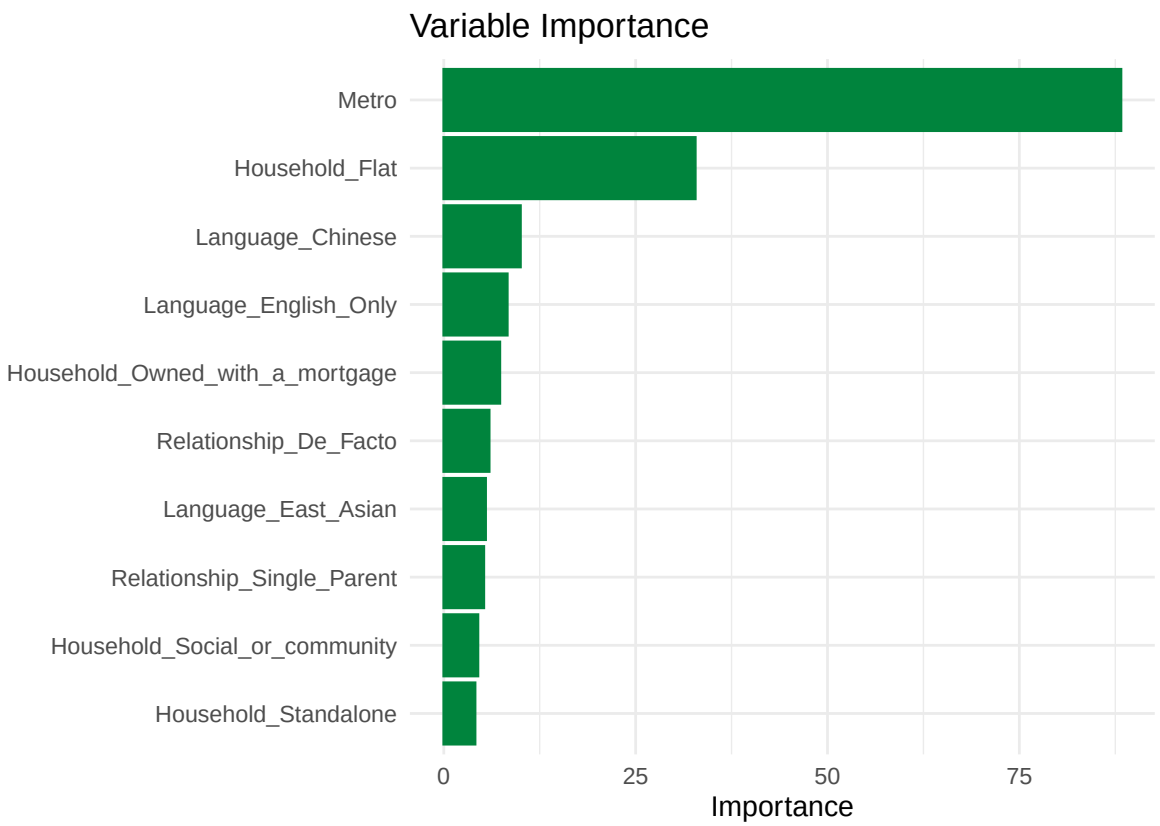


Figure 9: Variable importace - First classification model

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.

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

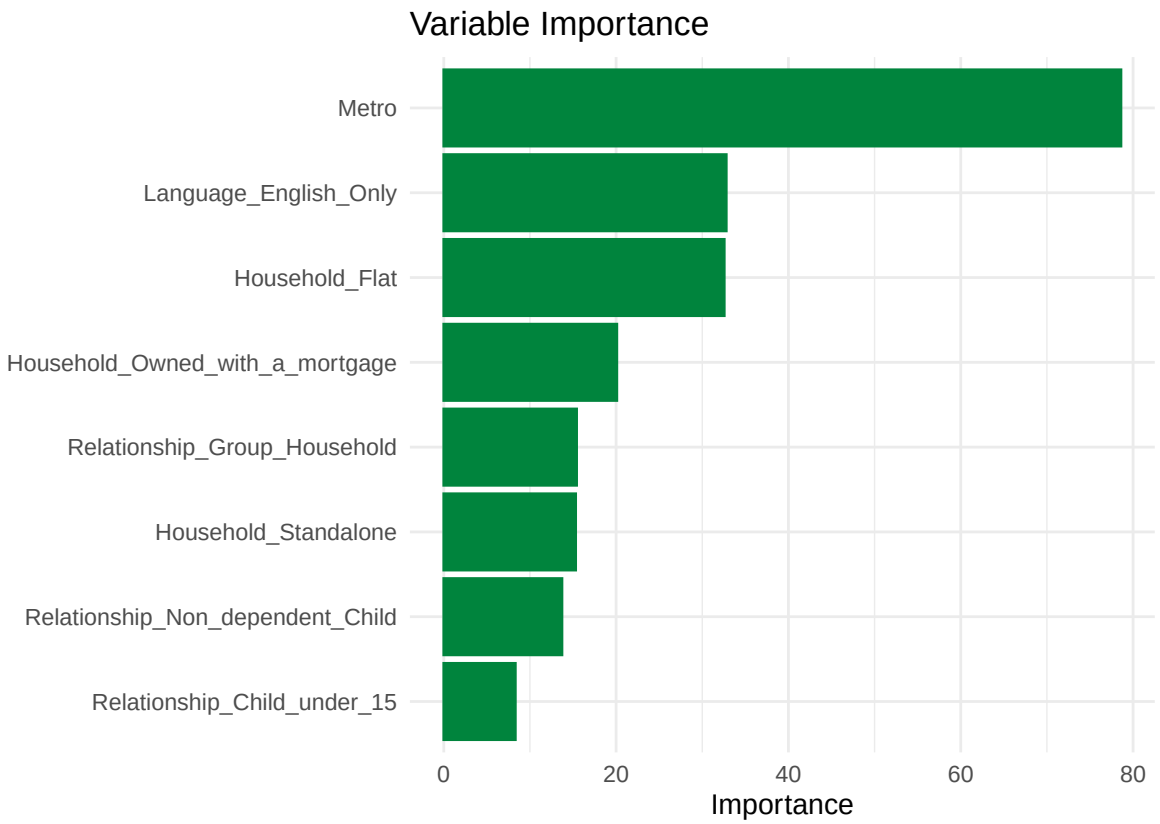


Figure 10: Variable importace - Improved classification model

presents the best RMSE result per cluster, alongside the selected tuning parameters.

Table 8: Best regression results by cluster

| | | | RMSE | | | | |
|---|---|---|---|---|---|---|---|
| Cluster | □ | □ | 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 **??**,**??** and **??**.

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 covariates in figures **??**,**??** and **??** 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

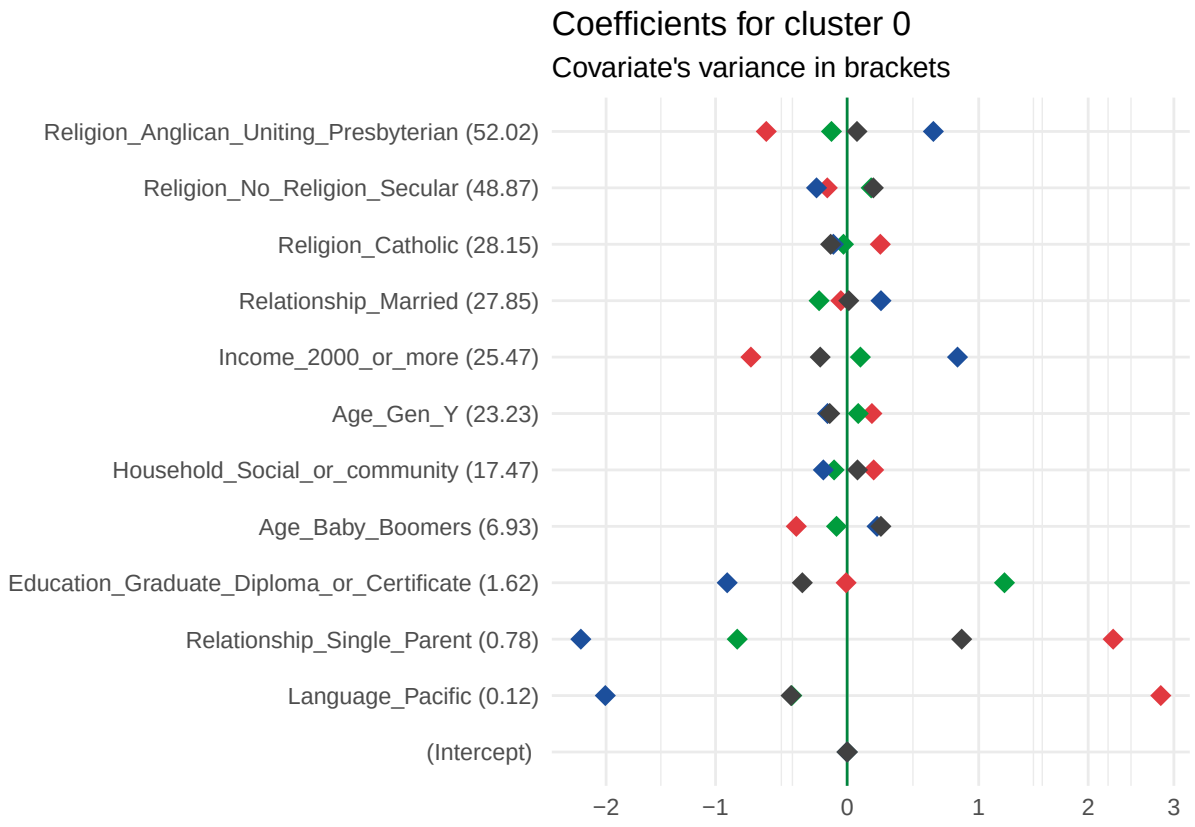


Figure 11: Resulting coefficients - cluster 0

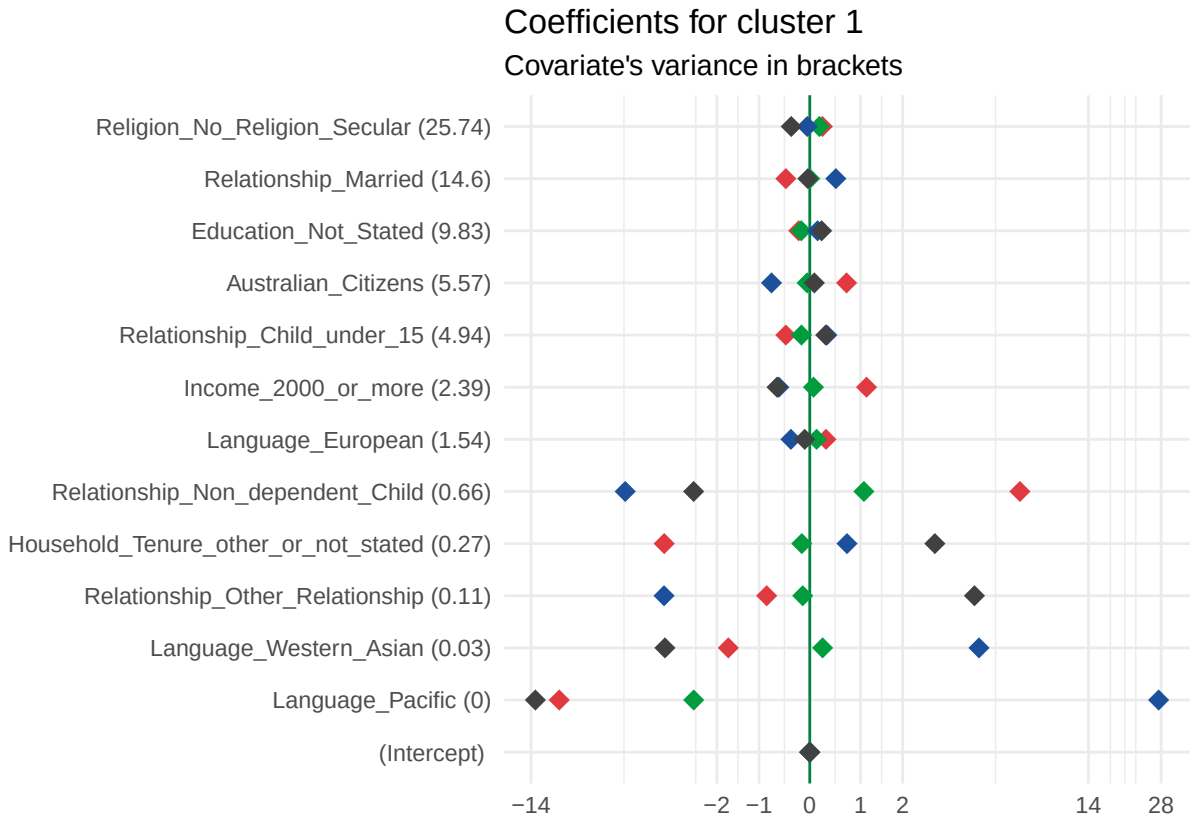


Figure 12: Resulting coefficients - cluster 0

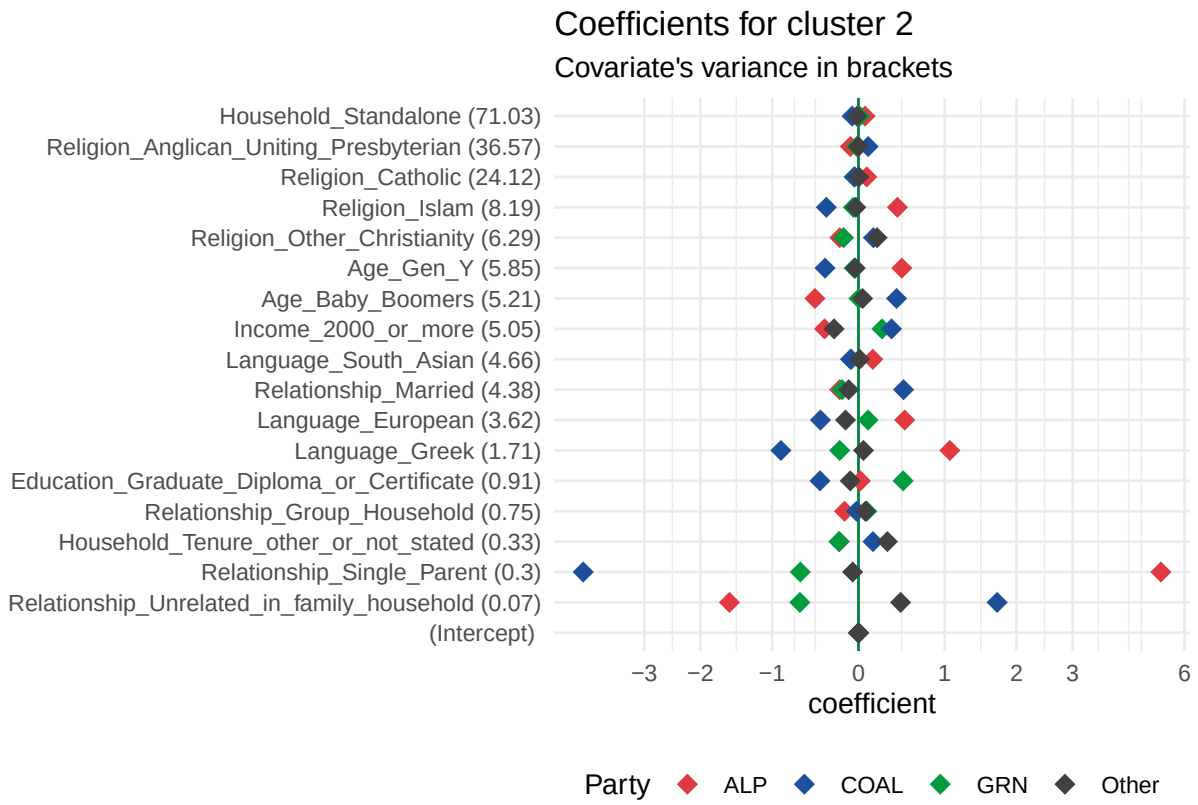


Figure 13: Resulting coefficients - cluster 0

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 [Benson, 2023] . 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 14.

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 15, 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 |

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

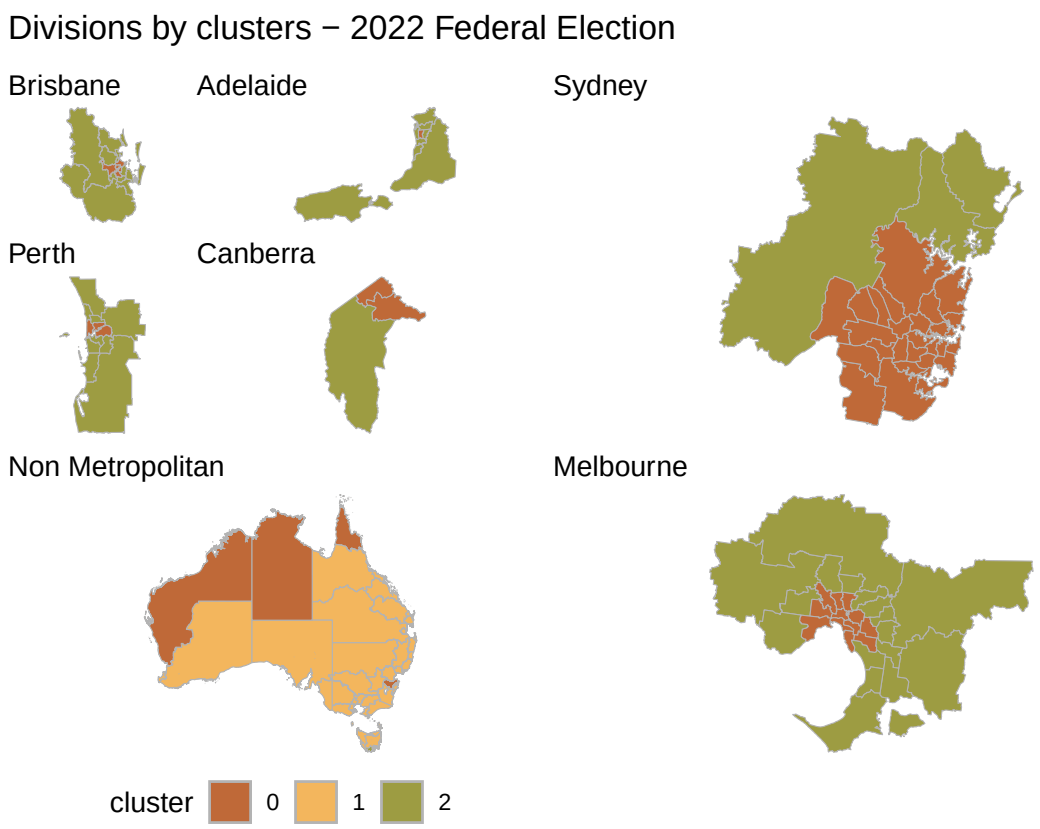


Figure 14: Clusters in 2022 Election

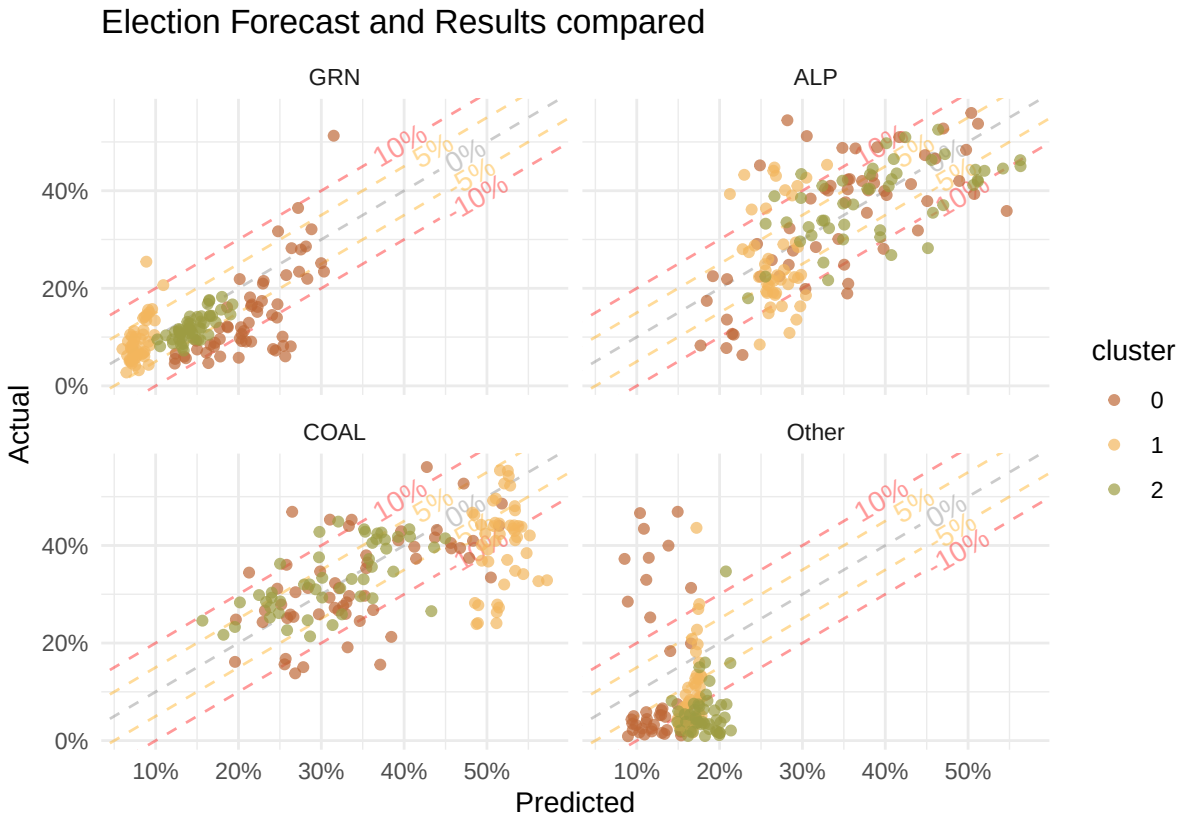


Figure 15: Comparison between prediction and election results

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 16.

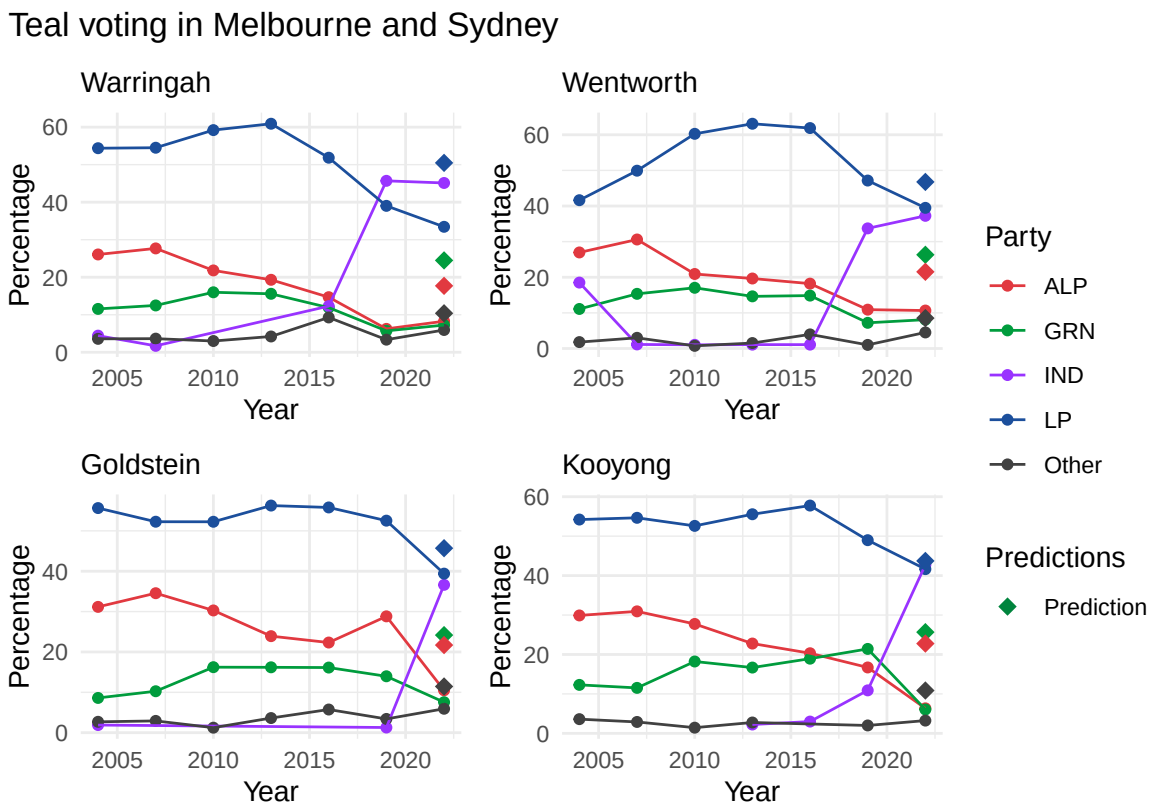


Figure 16: Example 1: Teal Wave

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 17, 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, 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? 18 shows the prediction of the latest and historic election results. figure 19 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

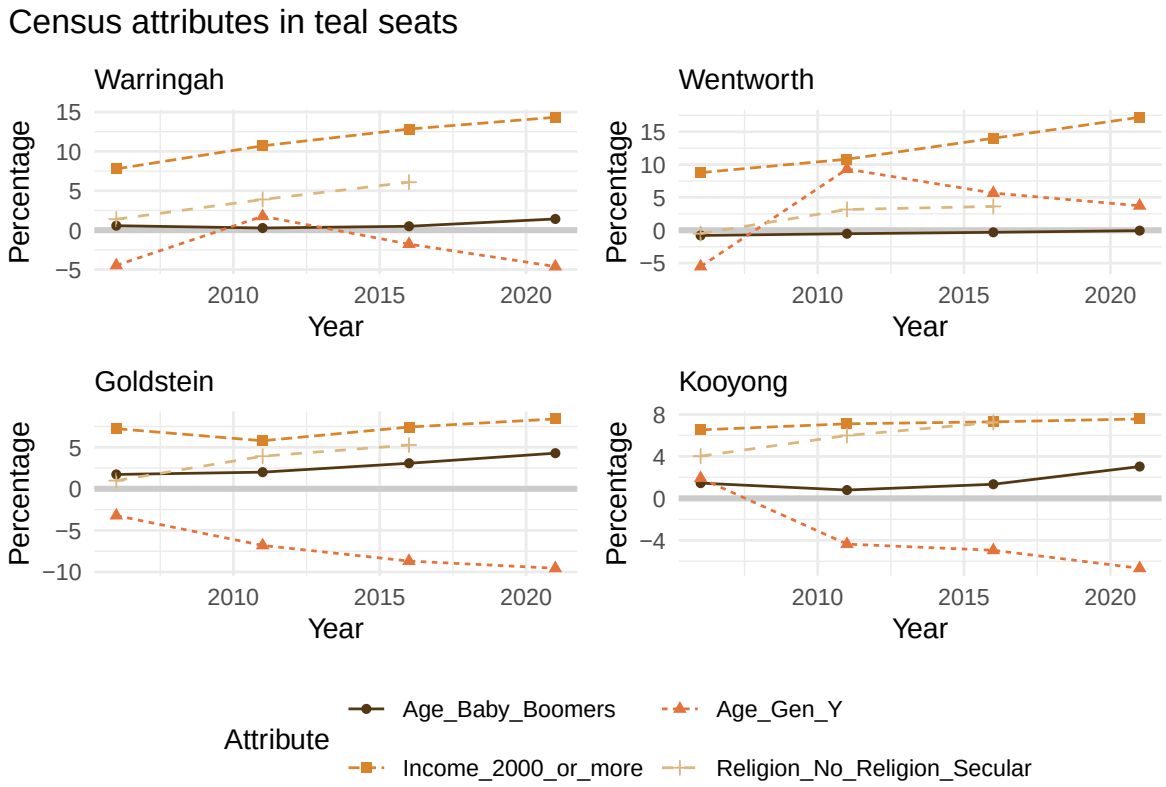


Figure 17: Selected demographics for teal seats

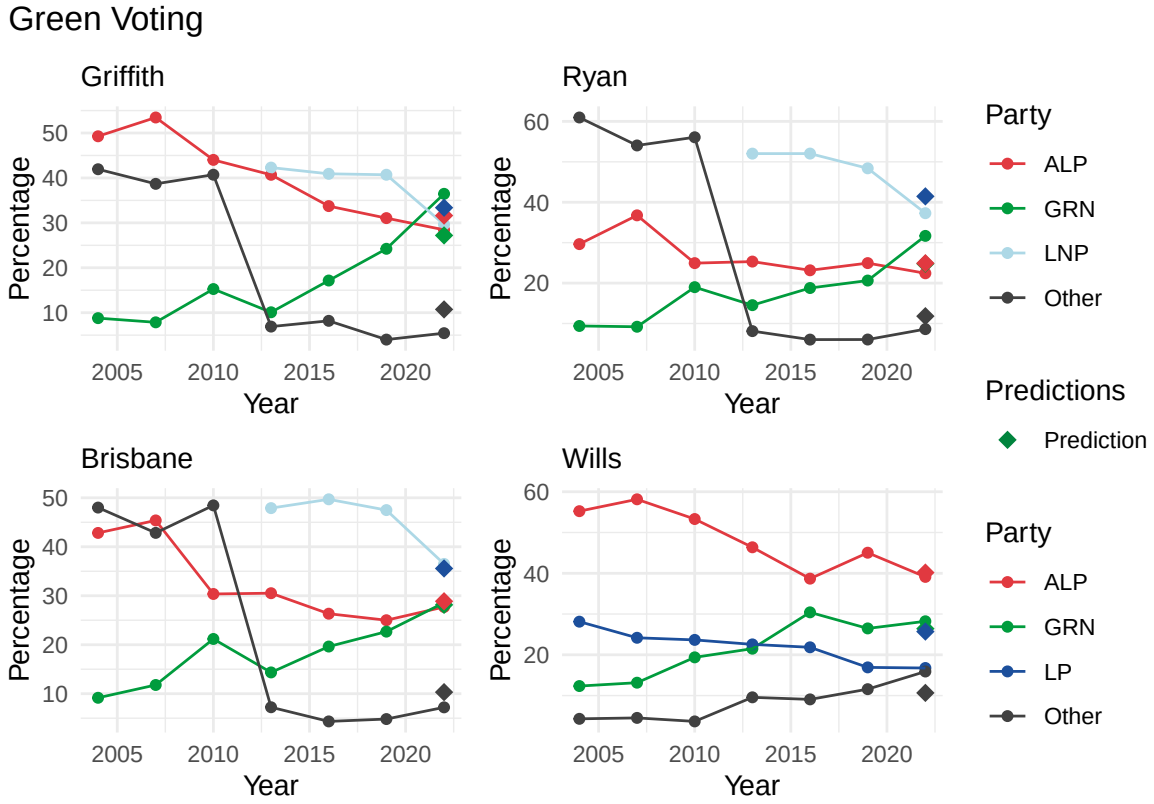


Figure 18: Green Voting

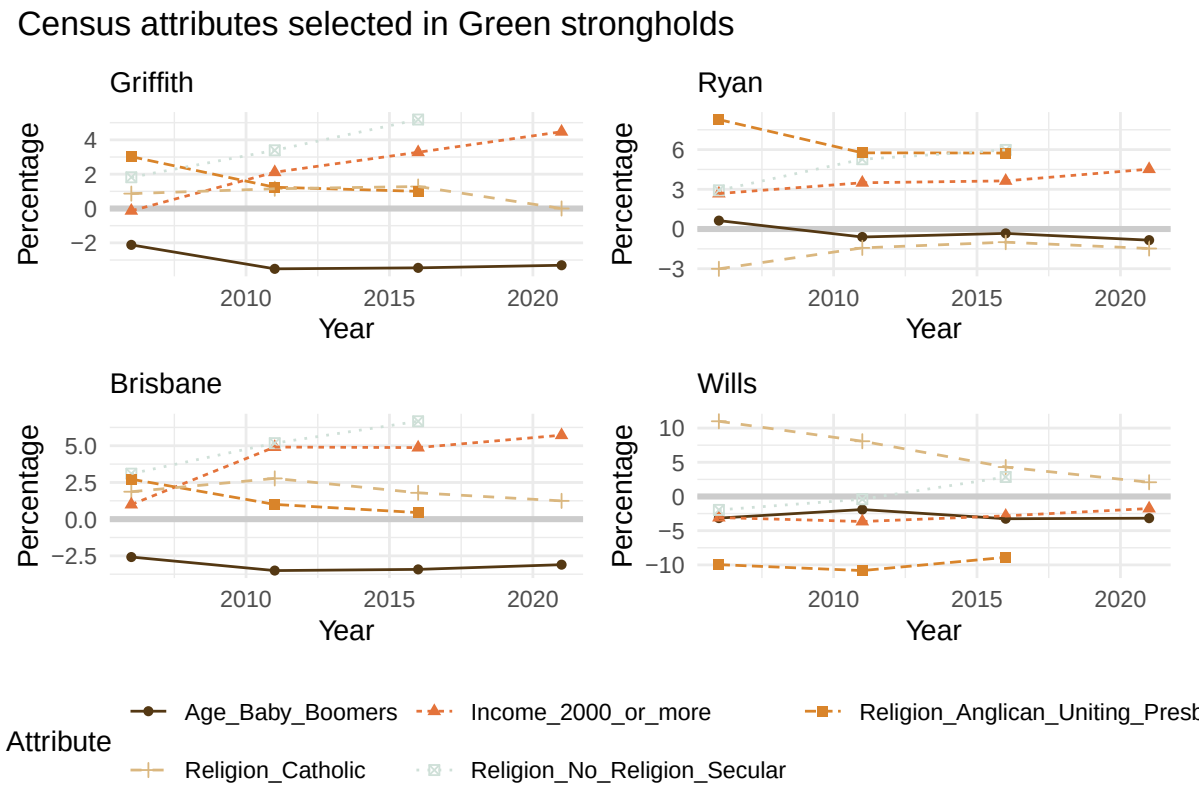


Figure 19: Demographics in Green strongholds

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, let's compare four suburban electorates: Hasluck (Perth), Menzies (Melbourne), Fowler (Sydney) and Kingston (Adelaide). Their respective predictions and results are presented in figure 20. A selection of key demographic variables is presented in figure 21.

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

- In **Hasluck** (WA) [Corporation, 2022b], the changes have the top maybe be driven by generational renewal. The "Other" vote increase includes progressive independents and localist parties, which may have influenced the lower-than-predicted results for the Greens.
- In **Menzies** (VIC) [corporate=Australian Broadcasting Corporation, 2022], 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 [corporate=Australian Electoral Commission, 2021]. This a good example where the existing model was able to effectively predict the primary vote based on those demographic changes.
- In **Fowler** (NSW) [Corporation, 2022a], an independent candidate altered Labor's trend. In these cases, a community-based candidate captured the multicultural vote from a "parachuted" Labor

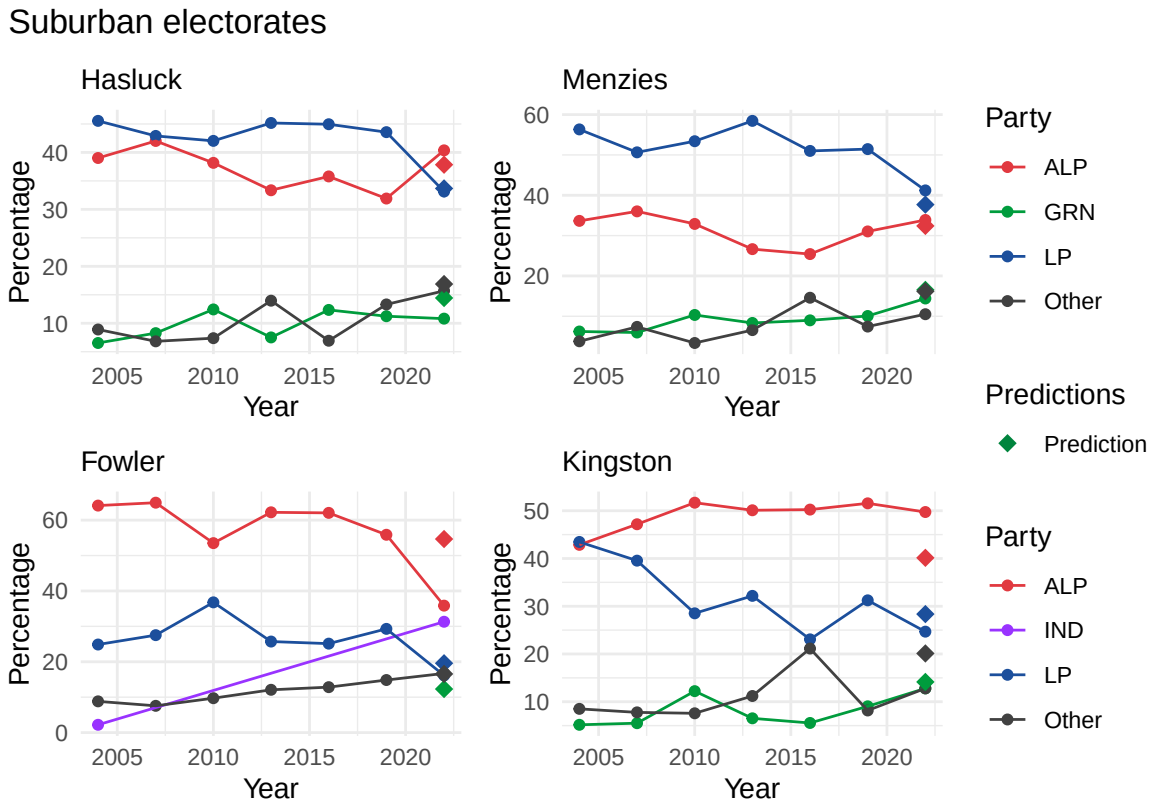


Figure 20: Suburban Voting

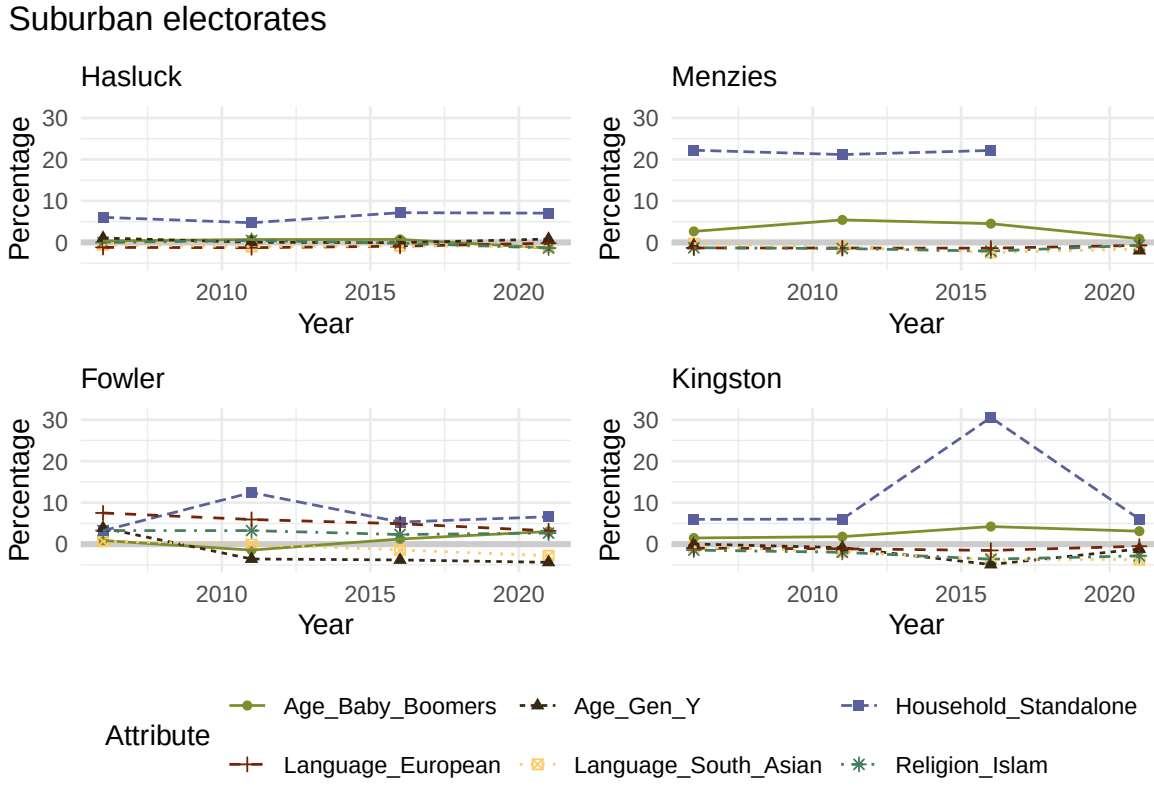


Figure 21: Demographics in suburban seats

nominee - which is a variable not considered in the model. [Hanrahan, 2022]
- Although the predictions for **Kingston (SA) [Corporation, 2022c]**, are no far off the actual results, it seems more difficult to explain them in terms of demographic changes - it is worth noticing that the model for cluster 2 uses 17 different variables.

## 6 Conclusion

In summary, this document 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, this could be improved by exploring 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 longitudinality 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.

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.

## References

Simon Benson. Newspoll: Labor lead over coalition narrows, 2023. URL https://www.theaustralian.com.au/nation/politics/newspoll-labor-lead-over-coalition-narrows/news-story/937dbfe8479e9380d93da4121f63c09d.

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

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. Lecture Notes in Computer Science, pages 160–172, Berlin, Heidelberg, 2013. Springer. doi: 10.1007/978-3-642-37456-2_14.

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

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

corporate=Australian Broadcasting Corporation. Menzies - federal election 2022, 2022. URL https://www.abc.net.au/news/elections/federal/2022/guide/menz.

corporate=Australian Electoral Commission. Proposed redistribution for victoria, 2021. URL https://www.aec.gov.au/Electorates/Redistributions/2021/vic/proposed-redistribution/index.htm. Last Modified: 2021-03-19.

Australian Broadcasting Corporation. Fowler - federal election 2022, 2022a. URL https://www.abc.net.au/news/elections/federal/2022/guide/fowl.

Australian Broadcasting Corporation. Hasluck (key seat) - federal election 2022, 2022b. URL https://www.abc.net.au/news/elections/federal/2022/guide/hasl.

Australian Broadcasting Corporation. Kingston - federal election 2022, 2022c. URL https://www.abc.net.au/news/elections/federal/2022/guide/king.

Brigitte Escofier and Jérome Pagès. *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation*. Sciences Sup. Dunod, 2008. URL https://hal.science/hal-00382085.

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

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

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

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

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

Commonwealth Parliament. Voting patterns by generation. URL https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/FlagPost/2022/April/Voting_patterns_by_generation. Archive Location: Australia Last Modified: 2022-04-29 Publisher:

corporateName=Commonwealth Parliament; address=Parliament House, Canberra, ACT, 2600; contact=+61 2 6277 7111.

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

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

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