Analysing Australian Election Results

Carlos Yanez Santibanez

2023-06-09

Contents

Chapter 1

Acknowledgements

Chapter 2

Introduction

Chapter 3

Strucure

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

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

Chapter 4

Data

4.1 Data Sources

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

- **The Australian Electoral Commission** (AEC) (**?**) . The AEC contains detailed online records for every federal election held in the 21st century, through their Tally Room website (**?**).
- The **Australian Bureau of Statistics** (ABS) (**?**). 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 (**?**), which are available for censuses from 2006 onwards.

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

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

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

- **{auspol} (?)**, which extracts and presents electoral results.
- **{auscensus} (?)**, which allows to interact with Census Data Packs to extract different statistics across geographical units, and across censuses.
- **{aussiemaps} (?)**, which assists with aggregating census data into electoral divisions, by matching and apportioning different geographical structures.

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

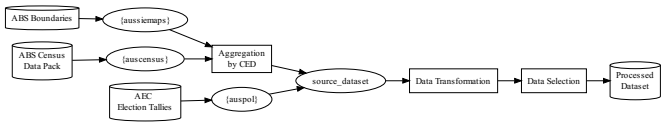

Figure 4.1: Flow of data from sources to dataset

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

1. Census data was extracted from the respective Census Data Pack using **{auscensus}**. Using the package workflow, key attributes were identified in each census, extracted from the respective files and given common names. Data were extracted for statistical areas and apportioned into Commonwealth Electoral Divisions by overlapping area, with the help of functions written into **{aussiemaps}**
2. Primary vote results for each division were extracted using the **auspol** package.
3. All the data was stored in a local database, from where was extracted and put together in a single dataset.

4. From there, the "raw" data was further processed and stored in a single "consolidated" dataset.

4.2 Data Selection

4.2.1 Census and Election Years

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

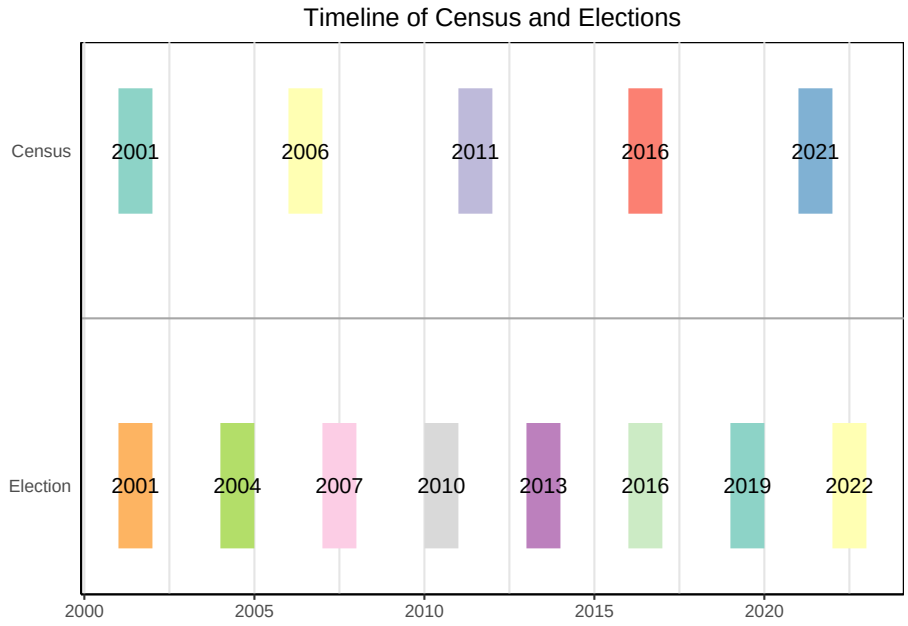

Figure 4.2: Census and Elections Timeline

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

Table 4.1: Selected Census-Election pairings

Census	Election
2006	2007
2010	2011
2016	2016
2021	2022

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

4.2.2 Electoral Data

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

Table 4.2: Sample extraction - Canberra 2022

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

4.2.3 Census Data

As mentioned in section **??**, a major challenge with respect of census data is the large volume of data points collected. For instance, the data pack for the 2022 Census contains 62 different tables, ranging from 8 [1] to 1,590 [2] attributes.

[1] *02 -Selected Medians and Averages*
[2] *09 - Country of Birth of Person by Age by Sex*

To select which variables to extract, literature and journalistic sources were consulted ((**?**), (**?**), (**?**)) to inform an initial set of covariates. In total (XYZ) variables were selected, which correspond to below to the following groups:

1. **Income** : Distribution of population in pre-set income brackets.
2. **Education Level**: Distribution of educational achievement (from incomplete secondary to vocational education and academic degrees).
3. **Age**: Distribution of the population in generational cohorts. Taking into account the selected elections, the four groups of interest are Baby Boomers (1946 to 1964), Generation X (1965 to 1980), Generation Y (1981 to 1996) and Generation Z (1997 to 2021).
4. **Relationship status**: Variables describing civil status (e.g. living alone, married, in a de facto relationship).
5. **Household type:** Descriptors of type of housing , (e.g. standalone house, semi-detached, flats).
6. **Household tenure:** Descriptors of house ownership, rental or other arrangement (e.g. public housing).
7. **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 relative integration of migrant communities into civic life.
8. **Religion:** Percentage of the population declaring to profess a religion. For this analysis, largest and high growth religious groups were selected (No religion/secular, Roman Catholic, Anglican-Presbyterian-Uniting, Christian Orthodox, Other Christianity, Islam, Hinduism, Buddhism).
9. **Language**: Languages spoken in the community. Similar to religion, a selection of relevant language have been included to reflect the historic and current migrant communities.

Apart from those, each electorate has been classified as **metropolitan** if it lies within the boundaries of Australian capital cities or **non-metropolitan** if not. Altogether, these variables try to reflect wealth and education (cited by (**?**) as key factors in deciding political persuasion), as well as stage in life and belonging to a particular migrant community (sometimes cited as an influential factor, for instance in (**?**)).

A sample of the resulting dataset is present in table **??**.

Table 4.3: Dataset sample

election_year	DivisionNm	ALP	COAL	GRN	Other	Australian_Citizens	Age_Baby_Boomers	Age_Gen_X	Language_Chinese
2022	Spence	11.70	-10.76	-0.61	-6.58	86.74	17.77	18.91	0.42
2022	Forde	-4.63	0.91	-2.13	-2.01	80.19	16.84	19.61	1.25
2022	Berowra	-10.37	13.23	4.01	-3.32	87.76	21.71	23.14	12.02
2022	Riverina	-12.27	11.85	-5.91	0.72	89.07	22.64	18.15	0.56
2007	Perth	3.63	-5.76	2.58	-0.22	81.40	25.26	26.05	3.01
2010	Kooyong	-10.75	9.10	6.78		83.82	16.47	22.63	9.41
2016	Bass	5.84	-3.84	0.82	0.91	88.15	21.66	21.33	0.94
2022	McEwen	4.33	-2.28	2.06	-6.33	88.79	17.70	21.33	0.68

4.3 Data Exploration

In total, the resulting dataset is made up of 4 response variables and 55 potential predictors, plus identificatory attributes like division mane and election year. As expected, an initial inspection shows that some of the covariates are loosely correlated with primary vote. Also expected, many of the covariates exhibit medium to high correlation levels amongst themselves, e.g. negative correlation between high and low level income groups, and certain age brackets with household type and tenure.

As examples, figure **??** show a somewhat weak correlation between Coalition primary vote and percentage of baby boomer population. Figure **??** presents the correlation values for religion and language attributes that aside from expected pairings (e.g. Hinduism and South Asian languages or Italian speakers and percentage of declared catholics), there is a almost exclusive positive correlation between membership to Anglican, Presbyterian and Uniting churches and percentage of monolingual English speakers. The percentage of monolingual English speakers is also negative correlated to all other language groups.

Besides from this, 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.

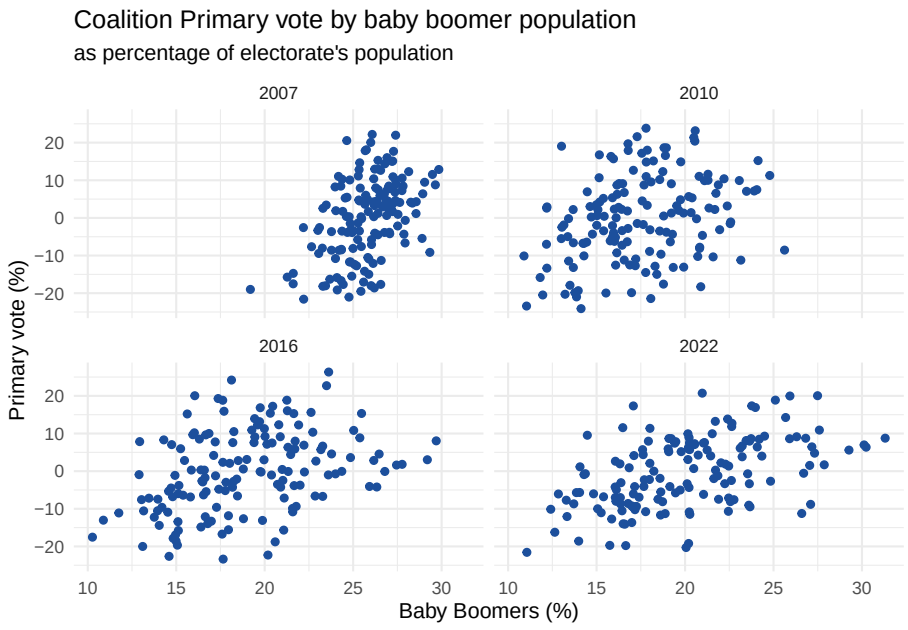

Figure 4.3: Correlation between Coalition vote and Baby boomer population

- There are also no obviously distinguishable differences when splitting results by each election.

4.3.1 Dimensionality reduction using Multiple Factor Analysis

Given the large number of variables and considering their correlation, it is worth exploring if a change of space could help to better identify variation, and whether the number of covariates can be reduced in a meaningful way. For this **multiple factor analysis** (MFA) (**?**) was used, given that:

- MFA allows to use variables that belong to groups.
- Allows to combine quantitative and qualitative variables.

The resulting scree plot and cumulative variance is presented in figure **??**.

In terms of interpretability of the new dimensions, figure **??** present group biplots for the 8 most important dimensions. From there, it is possible show that there is not straightforward representation expect with Dimension 2 and Education variables.

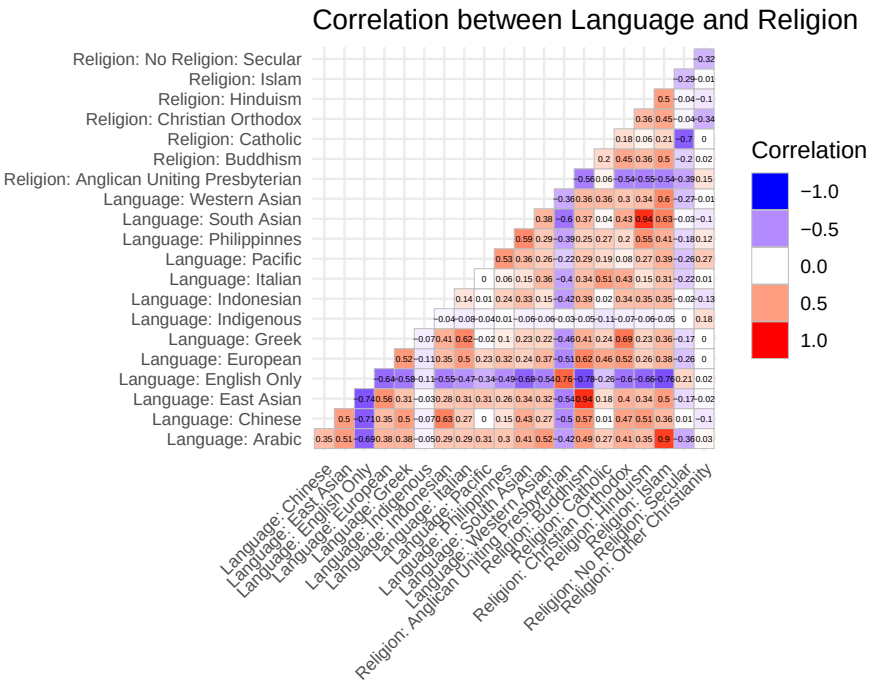

Figure 4.4: Correlation for selected covariates

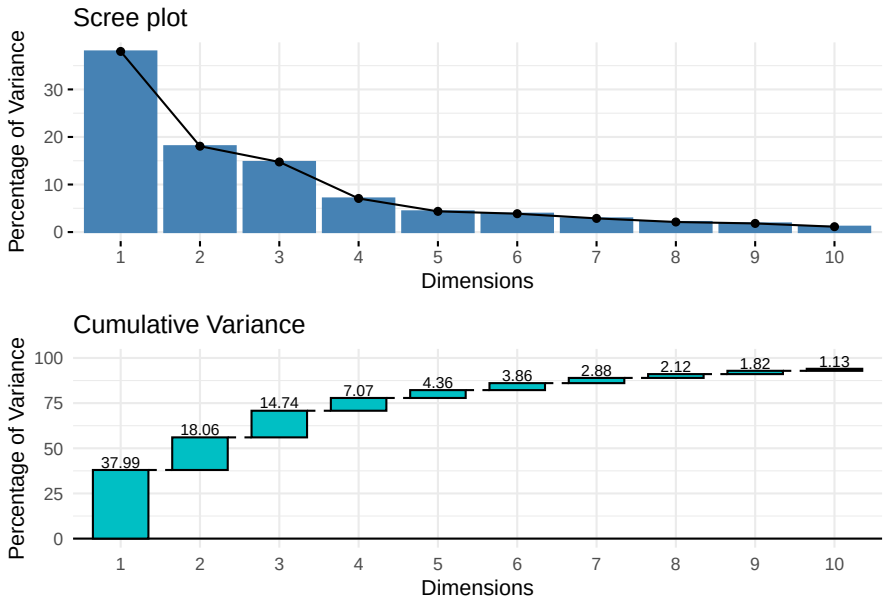

Figure 4.5: Scree plot and cumulative variance

4.3.2 Electorate segments

A common trope of Australian politics is to map voters's political persuasion to whether they live in the inner cities, suburbia or regional areas. Therefore, it is relevant to explore if this can be substantiated by demographic attributes, or if there any other grouping of voters that may influence primary voting.

To explore this, a clustering algorithm has been applied using all electorates for election from 2006 up to 2016. After trial and error, the clustering procedure consists in:

- Transforming all demographic attributes to represent the difference of each data point and their corresponding national value (for the relevant year).
- Using the HDBSCAN (**?**), a density-based hierarchical clustering algorithm optimises the number of clusters and assignment

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

This mapping seems to align with popular political knowledge, where:

- **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

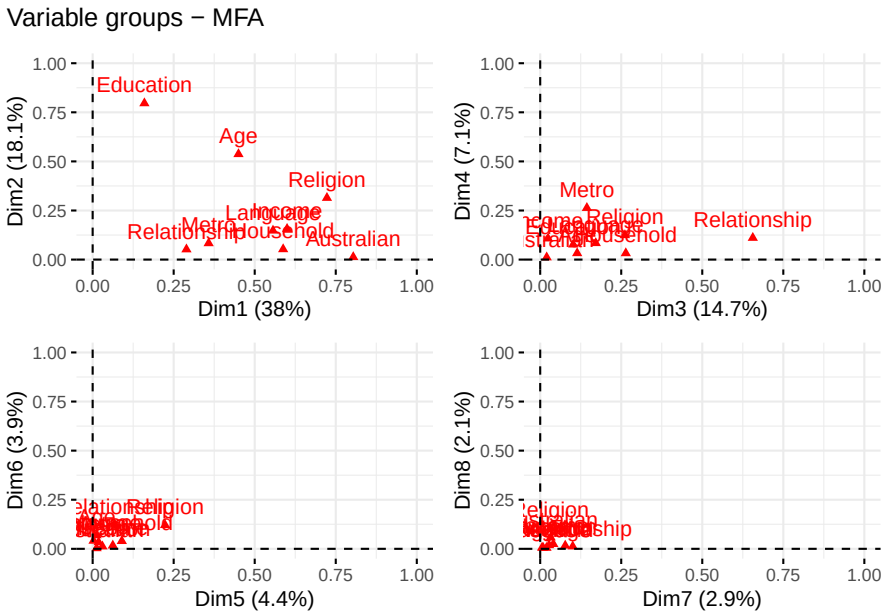

Figure 4.6: Group plots for first 8 dimensions

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 it also more prevalent in Brisbane and Perth compared when comparing capital cities.

Revisiting the demographic attributescan help to understand how this clusters differ from each other. A selection of those variables is presented in figure **??**.

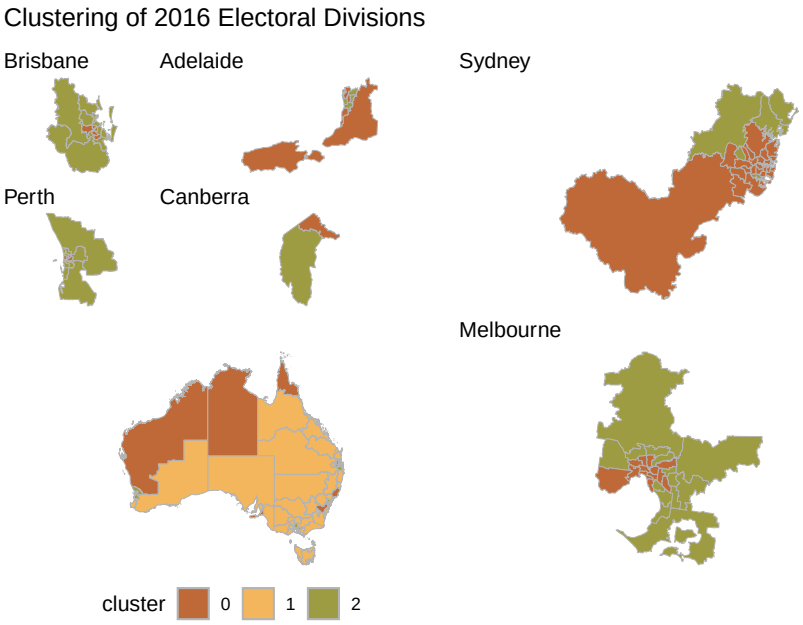

Figure 4.7: Clusters for 2016 Election

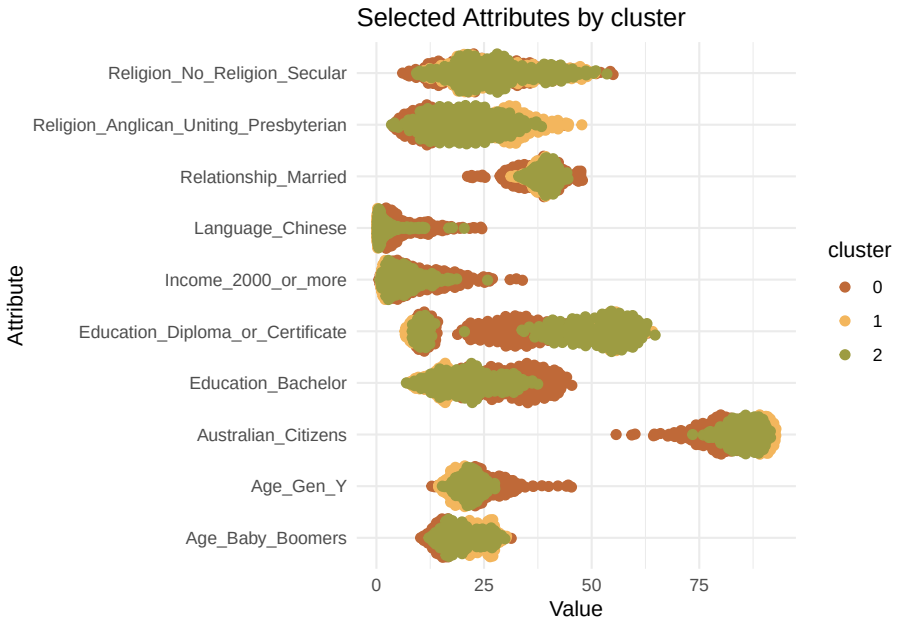

Figure 4.8: Selected attributes, coloured by cluster.

Chapter 5

Method

Having obtained the data, it modelling time! Starting from a simple point, research in question can be expressed by equation (**??**).

$$\mathbf{Y} = \mathrm{f}(\mathbf{X}) \tag{5.1}$$

where $\mathbf{Y}$ represents a vector with primary voting for an electorate, and $\mathbf{X}$ represents the vector of respective demographic attributes.

Considering the problem in question, this model does not consider other factors that may influence voting. These factors may difficult to quantify as they may related 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. However, it is possible to make the naive assumption that electoral polling aims to capture all those factors - thus it is possible to restate the objective : instead of calculating the primary vote itself, the aim is to determine how demographics influences the voting when compared to a the national values (as captured by polls). Furthermore, it is possible to redefine the problem once more, to aiming to determine how differences in demographic attributes, when compared against a national baseline makes primary voting in an electoral to differ from the national measure. This is states in (**??**), which allow accounts to error attributable to polling error and demographic drift between census and election.

$$\mathbf{Y} = \mathrm{f}(\mathbf{X}) +$$