

ETC5512: Wild Caught Data

Week 7

Census and Election Data

Lecturer: *Emi Tanaka*

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu

6th May 2020



What is wild-caught data?

- 👾 data can be freely used
- 👾 data can be modified
- 👾 data can be shared by anyone for any purpose
- 👾 data source is traceable
- 👾 data collection is transparent
- 👾 data are updated as new measurements arrive
- 👾 if any data processing, the process is clearly described and reproducible

Observational data

i

Population is the whole set of units (such as people, animal, place etc) to which the question or experiment pertains to.

Sample is a subset of population that (hopefully) represents the population.

Sample or population?

i

Vò is a small town in northern Italy with 3,300 inhabitants. All inhabitants of the town were tested and retested for COVID-19. On 6th March, there were 89 infected in Vò. There are 4,636 known cases of infection out of about 60 million people in Italy on 6th March. As of 31st March, there are no longer new cases of infection in Vò and 101,739 known cases in all of Italy.

- _DEPENDS Depends on the question of interest!
- DEPENDS We have the population data for Vò but for the whole of Italy, the number of known infection cases would be a sample.

Aim

This week we are interested in extracting and studying the **personal income data** from the 2016 Australian census and the **election data** from the 2019 Australian federal election.

You'll learn about **tidy data**.

Australian Bureau of Statistics

Census Data 2016

Australian Bureau of Statistics (ABS)

- ✿ ABS is the independent statistical agency of the Government of Australia.
- ✿ If you are from outside Australia, find the statistical government agency in your country 🔧, e.g. in Japan, this is the [Statistics Bureau of Japan](#).
- ✿ ABS provides key statistics on a wide range of economic, population, environmental and social issues, to assist and encourage informed decision making, research and discussion within governments and the community.



ABS Census Data

- The first Australian census was held in 1911.
- Since 1961, the census occurs every 5 years in Australia.
- The last census was in 2016 at a cost of \$440 million.
- The next census will be held in 2021.
- The ABS is legislated to collect and disseminate census data under the ABS Act 1975 and Census and Statistics Act 1905.
- Similar legislation are in place in many countries.

Please use CAPITAL letters only.		12	Person 1	Person 2
29	What is the level of the <i>highest</i> qualification the person has <i>completed</i> ? For example: TRADE CERTIFICATE, BACHELOR DEGREE, ASSOCIATE DIPLOMA, CERTIFICATE II, ADVANCED DIPLOMA.	Level of qualification		Level of qualification
30	What is the main field of study for the person's <i>highest</i> qualification <i>completed</i> ? For example: PLUMBING, HISTORY, PRIMARY SCHOOL TEACHING, HAIRDRESSING, GREENKEEPING.	Field of study		Field of study
31	Did the person <i>complete</i> this qualification before 1998? Remember to mark the box like this: ■■■■■	Yes, before 1998 No, 1998 or later		Yes, before 1998 No, 1998 or later
32	For each female, how many babies has she ever given birth to? Exclude adopted, foster and step children. <small>(1) Go to census.abs.gov.au for more information.</small>	Number of babies None		Number of babies None
33	What is the <i>total</i> of all income the person <i>usually receives</i> ? Mark one box only. Do not deduct: tax, superannuation contributions, amounts salary sacrificed, or any other automatic deductions. Include: <ul style="list-style-type: none">- Wages and salaries<ul style="list-style-type: none">- Regular overtime- Commissions and bonuses- Government pensions, benefits and allowances<ul style="list-style-type: none">- Age pension- Youth and student allowances- Family tax benefit- Parenting payment- Disability support pension- Newstart allowance- Any other government pension/allowance- Profit or loss from<ul style="list-style-type: none">- Unincorporated business/farm (e.g. sole traders, partnerships)- Rental properties- Other income<ul style="list-style-type: none">- Income from superannuation- Private pensions- Child support- Interest- Dividends from shares- Workers' compensation- Any other income <small>Information from this question provides an indication of living standards in different areas. (1) Go to census.abs.gov.au for more information.</small>	\$3,000 or more per week \$166,000 or more per year \$2,000 - \$2,999 per week \$104,000 - \$155,999 per year \$1,750 - \$1,999 per week \$91,000 - \$103,999 per year \$1,500 - \$1,749 per week \$78,000 - \$90,999 per year \$1,250 - \$1,499 per week \$65,000 - \$77,999 per year \$1,000 - \$1,249 per week \$52,000 - \$64,999 per year \$800 - \$999 per week \$41,600 - \$51,999 per year \$650 - \$799 per week \$33,800 - \$41,599 per year \$500 - \$649 per week \$26,000 - \$33,799 per year \$400 - \$499 per week \$20,800 - \$25,999 per year \$300 - \$399 per week \$15,600 - \$20,799 per year \$150 - \$299 per week \$7,800 - \$15,599 per year \$100 - \$149 per week \$5,100 - \$7,799 per year \$Nil income \$Negative income		\$3,000 or more per week \$156,000 or more per year \$2,000 - \$2,999 per week \$104,000 - \$155,999 per year \$1,750 - \$1,999 per week \$91,000 - \$103,999 per year \$1,500 - \$1,749 per week \$78,000 - \$90,999 per year \$1,250 - \$1,499 per week \$65,000 - \$77,999 per year \$1,000 - \$1,249 per week \$52,000 - \$64,999 per year \$800 - \$999 per week \$41,600 - \$51,999 per year \$650 - \$799 per week \$33,800 - \$41,599 per year \$500 - \$649 per week \$26,000 - \$33,799 per year \$400 - \$499 per week \$20,800 - \$25,999 per year \$300 - \$399 per week \$15,600 - \$20,799 per year \$150 - \$299 per week \$7,800 - \$15,599 per year \$100 - \$149 per week \$5,100 - \$7,799 per year \$Nil income \$Negative income

Get the ABS 2016 Census Data



<https://datapacks.censusdata.abs.gov.au/datapacks/>

- > 2016 Census Datapacks
- > General Community Profile
 - > All geographies
 - > Vic

Census DataPacks

Step 1: Select Census year

2016 Census Datapacks



Step 2: Select DataPacks type

General Community Profile



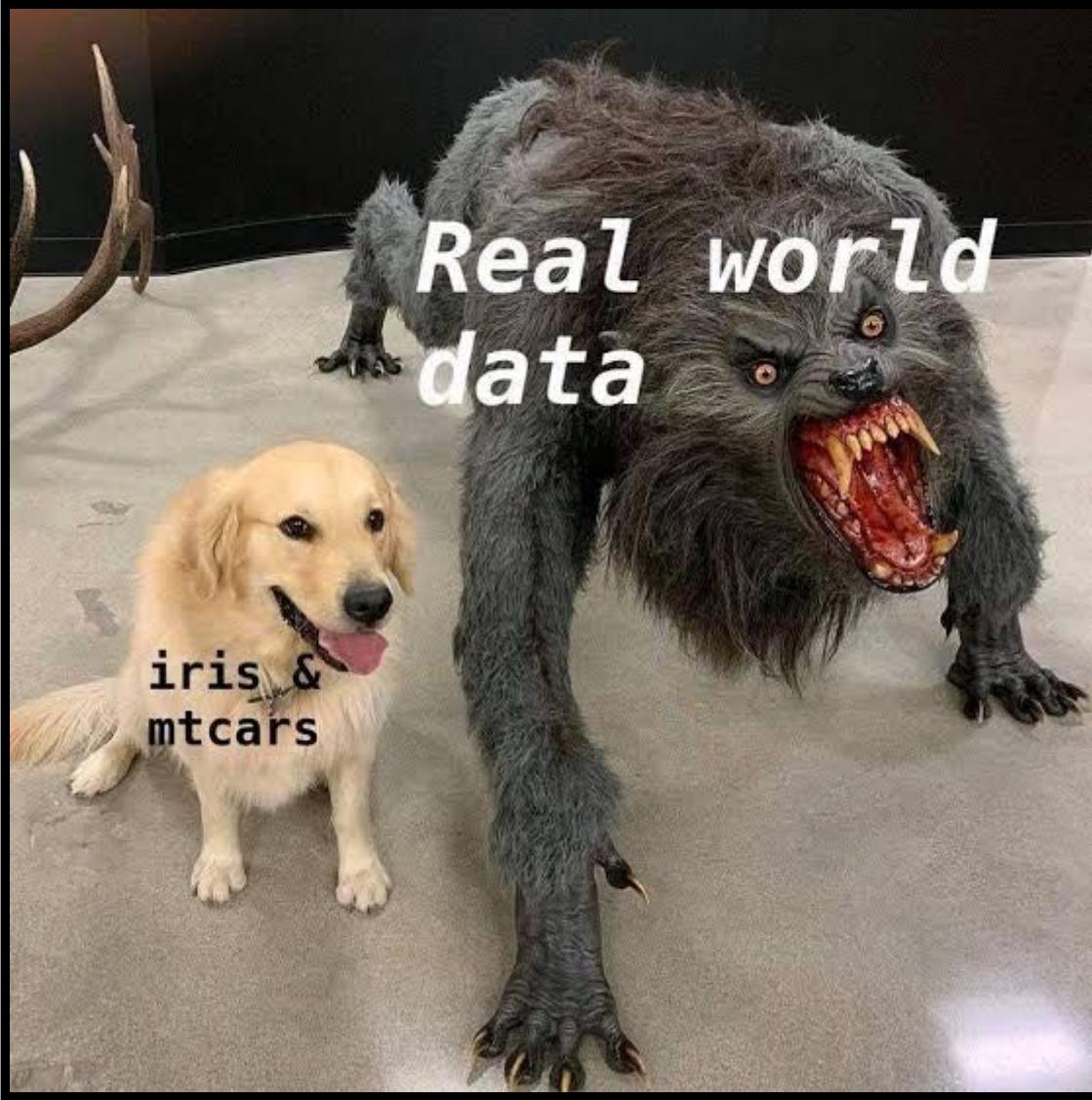
Step 3: Select Geography

All geographies



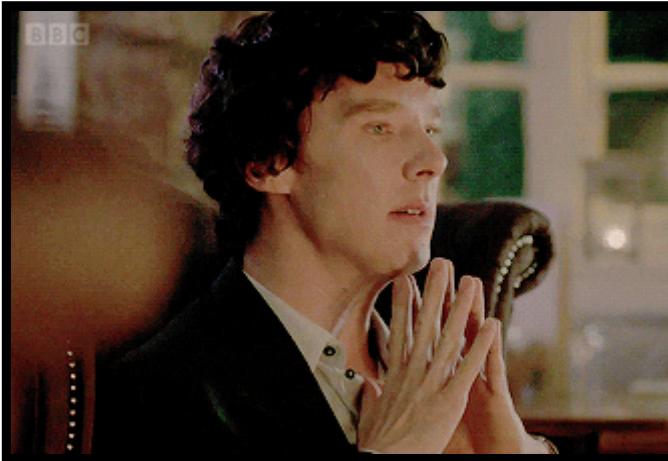
Need help with DataPacks? Download the help file : [Zip File](#)

Geographies	Aust	NSW	Vic	Qld	SA	WA	Tas	NT	ACT	OT	Geography boundary: ESRI Shapefile	Geography boundary: Map Interchange
All geographies												



Navigating ABS Census data

- ✿ First, pray hard that there is some description!



- ✿ Without some description or understanding of the variables, it will be near impossible to extract meaningful information from the data.

Navigating ABS Census data

2016_GCP_ALL_for_Vic_short-header

- └── 2016 Census GCP All Geographies **for** VIC
- └── Metadata
- └── Readme

✿ Readme is a good place to start here (phew!)

*"About DataPacks_readme.md - "Read Me"
documentation containing helpful information for
users about the data and how it is structured
.md)"*

✿ But there is no DataPacks_readme.md??

Navigating ABS Census data

```
2016_GCP_ALL_for_Vic_short-header/Readme
    ├── 2016POA_readme.txt
    ├── AboutDatapacks_readme.txt
    ├── CreativeCommons_Licensing_readme.txt
    ├── Formats_readme.txt
    ├── Summary_of_Changes.txt
    ├── esri_arcmapper_readme.txt
    └── mapinfo_readme.txt
```

- ✿ There is no DataPacks_readme.md but there is [AboutDatapacks_readme.txt](#).
- ✿ But it's not helpful in locating the income data...

Navigating ABS Census data

We could also try going through the meta-data.

2016_GCP_ALL_for_Vic_short-header/Metadata

- └── 2016Census_geog_desc_1st_2nd_3rd_release.xlsx
- └── 2016_GCP_Sequential_Template.xlsx
- └── Metadata_2016_GCP_DataPack.xlsx

[Metadata_2016_GCP_DataPack.xlsx](#)

Table number	Table name	Table population
G17	Total Personal Income (Weekly) by Age by Sex	Persons aged 15 years and over
	Total Family Income (Weekly) by Family	Families in family

Navigating ABS Census data

Where is Table G17?

2016_GCP_ALL_for_Vic_short-header/2016 Census GCP All Geographies **for** VIC/

- └── CED
- └── GCCSA
- └── LGA
- └── POA
- └── RA
- └── SA1
- └── SA2
- └── SA3
- └── SA4
- └── SED
- └── SOS
- └── SOSR
- └── SSC
- └── STE
- └── SUA
- └── UCL

Navigating ABS Census data

Back to metadata

2016_GCP_ALL_for_Vic_short-header/Metadata

- └── 2016Census_geog_desc_1st_2nd_3rd_release.xlsx
- └── 2016_GCP_Sequential_Template.xlsx
- └── Metadata_2016_GCP_DataPack.xlsx

Let's open

[2016Census_geog_desc_1st_2nd_3rd_release.xlsx](#)

... and there are the region names of each geographical code.

Let's go with the easy one: [STE Victoria](#).

Navigating ABS Census data

STE/VIC/

- |—— ...
- |—— 2016Census_G17A_VIC_STE.csv
- |—— 2016Census_G17B_VIC_STE.csv
- |—— 2016Census_G17C_VIC_STE.csv
- |—— 2016Census_G18_VIC_STE.csv
- |—— ...

✿ G17A, G17B, G17C?

Why is the table organised like this?

✿ Examine the files 2016Census_G17A_VIC_STE.csv,
2016Census_G17B_VIC_STE.csv and

Tables G17A-G17C

2016Census_G17A_VIC_STE.csv

STE_CODE_2016♦	M_Neg_Nil_income_15_19_yrs♦	M_Neg_Nil_income_20_24_yrs♦
2	88338	31685

2016Census_G17B_VIC_STE.csv

STE_CODE_2016♦	F_400_499_15_19_yrs♦	F_400_499_20_24_yrs♦	F_400_499_25_34_yrs♦
2	4020	17474	

2016Census_G17C_VIC_STE.csv

STE_CODE_2016♦	P_1000_1249_15_19_yrs♦	P_1000_1249_20_24_yrs♦	P_1000_1249_25_34_yrs♦
2	1061	25642	

Table G17

There are few things to note:

- ✿ There are 201 columns in G17A and G17B and 81 columns in G17C.
- ✿ Perhaps there is an export limitation for a data that contains more than 200 columns, thus it is broken up into different csv files.
- ✿ Which means that you have to join the tables G17A, G17B and G17C as one (you'll do this in the tutorial .

But what does the data show?

What is Tidy Data?

i

Tidy Data Principles

1. Each variable must have its own column
2. Each observation must have its own row
3. Each value must have its own cell

So what about the ABS 2016 Census Data?

- ✿ The table header in fact contains information!
- ✿ E.g. `F_400_499_15_19_yrs` is female aged 15-19 years old who earn \$400-499 per week (in Victoria).
- ✿ The number in the cells are the **counts**.
- ✿ Is the data tidy?

Tidying the ABS 2016 Census Data

- ✿ Ideally we want the data to look like:

age_min	age_max	gender	income_min	income_max	count
15	19	female	400	499	4020

- ✿ You can include other information, e.g. geography code (useful if combining with other geographical area) or average age/income.
- ✿ Note that some don't have upper bounds, e.g. [M_3000_more_85ov](#). In R, `-Inf` and `Inf` are used to represent $-\infty$ and ∞ , respectively.
- ✿ You'll wrangle the data into the tidy form in tutorial 

Raw Data vs. Aggregated Data

- ✿ Although the data collected was from individual households surveying each person in the household (see sample form [here](#)), the downloaded data are **aggregated**.
- ✿ Aggregated data presents summary statistics from the *raw data*. When the only summary statistics are counts then it is generally called *frequency data*.
- ✿ The raw data collected would be similar to the form

household_id	person	gender	age	marital_status	income_per_week
1	John Smith	F	40	Married	400-499
1	Jane Smith	M	39	Married	300-399
1	David Smith	M	10	Never married	Nil
1	Mary Smith	F	8	Never married	Nil
2	John Citizen	M	32	Never married	400-499

What you lose in aggregate data

- ✿✿ For aggregate data, there are less scope for you to draw insights conditioned on other variables.
- ✿✿ E.g. based on frequency data alone, you cannot answer questions like: how many middle income families with 2 children?
- ✿✿ Raw data are desirable if you can get hold of it!

Trust and skepticism

- ✿✿ By the way, did you notice anything odd about the dummy data presented in the last slide?
- ✿✿ John Smith was recorded as female and Jane Smith as male. Data may have been incorrectly recorded.
- ✿✿ How much do you trust the aggregate data?
- ✿✿ Have some healthy dose of skepticism in your data.

Data Confidentiality

- ✿ The data is not just aggregated, but it is also anonymised
- ✿ E.g. in [2016_GCP_Sequential_Template.xlsx](#), Sheet "G 17a", footnote says "*Please note that there are **small random adjustments** made to all cell values to protect the confidentiality of data. These adjustments may cause the sum of rows or columns to differ by small amounts from table totals.*"
- ✿ Why is confidentiality of data important?
- ✿ 2013 New York City taxi data :
 - ⌚ ~20GB of data on over 170 million taxi trips
 - ⌚ anonymised taxi license numbers were easily decoded
 - ⌚ the taxi trips were matched with celebrities that have photos taken with the taxi license plate number and reveals how they tip

Australian Federal Election 2019

Get the distribution of preferences by candidate by division for the 2019 Australian Federal Election



<https://results.aec.gov.au/>

- > 2019 federal election
 - > Downloads
- > Distribution of preferences by candidate by div

2019 Australian Federal Election

✿✿ Parliament of Australia comprises two houses:

- Senate (upper house) comprising 76 senators
- House of Representatives (lower house) comprising 151 members

✿✿ Government is formed by the party or coalition with majority of the seats in the lower house

✿✿ The 2019 Australian Federal Election was held on Sat 18th May 2019

✿✿ Voting is compulsory if you are an Australian citizen

✿✿ Major parties in Australia:

- Coalition:



LIBERAL



National

- Labor



✿✿ Some minor parties in Australia:

- The Greens
- One Nation



The Greens

PAULINE HANSON'S
One NATION

Ballots

- 👾 House of Representatives uses the instant-runoff voting system
 - 👾 Senate uses the single transferable voting system

	House of Representatives Ballot Paper	
<p>State</p> <p>Electoral Division of Division Name</p> <hr/> <p>Number the boxes from 1 to 8 in the order of your choice</p> <hr/>		
2 	SURNAME, Given Names <small>INDEPENDENT</small>	
3 	SURNAME, Given Names <small>PARTY</small>	
7 	SURNAME, Given Names <small>PARTY</small>	
4 	SURNAME, Given Names <small>PARTY</small>	
1 	SURNAME, Given Names <small>PARTY</small>	
5 	SURNAME, Given Names <small>PARTY</small>	
6 	SURNAME, Given Names <small>PARTY</small>	
8 	SURNAME, Given Names <small>PARTY</small>	

Remember... number every box to make your vote count!

SAMPLE

Senate Ballot Paper State – Election of 6 Senators						
<p>You may vote in one of two ways</p> <p>Either</p> <p>Above the line</p> <p>By numbering at least 6 of these boxes in the order of your choice (with number 1 as your first choice).</p>						
A <input type="checkbox"/> LOGO 5 PARTY	B <input type="checkbox"/> LOGO 2 PARTY	C <input type="checkbox"/> LOGO <input type="checkbox"/> LOGO 1 PARTY	D <input type="checkbox"/> LOGO <input type="checkbox"/> PARTY	E <input type="checkbox"/> LOGO <input type="checkbox"/> LOGO 3 PARTY	F <input type="checkbox"/> LOGO <input type="checkbox"/> LOGO 6 PARTY	G <input type="checkbox"/> 4
<p>Or</p> <p>Below the line</p> <p>By numbering at least 12 of these boxes in the order of your choice (with number 1 as your first choice).</p>						
PARTY	PARTY	PARTY	PARTY	PARTY	PARTY	UNGROUPED
<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names NO PDP/PNT
<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names NO PDP/PNT
<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names NO PDP/PNT
<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names PARTY	<input type="checkbox"/> SURNAME Given Names NO PDP/PNT
						
<p style="text-align: right;">SAMPLE</p>						

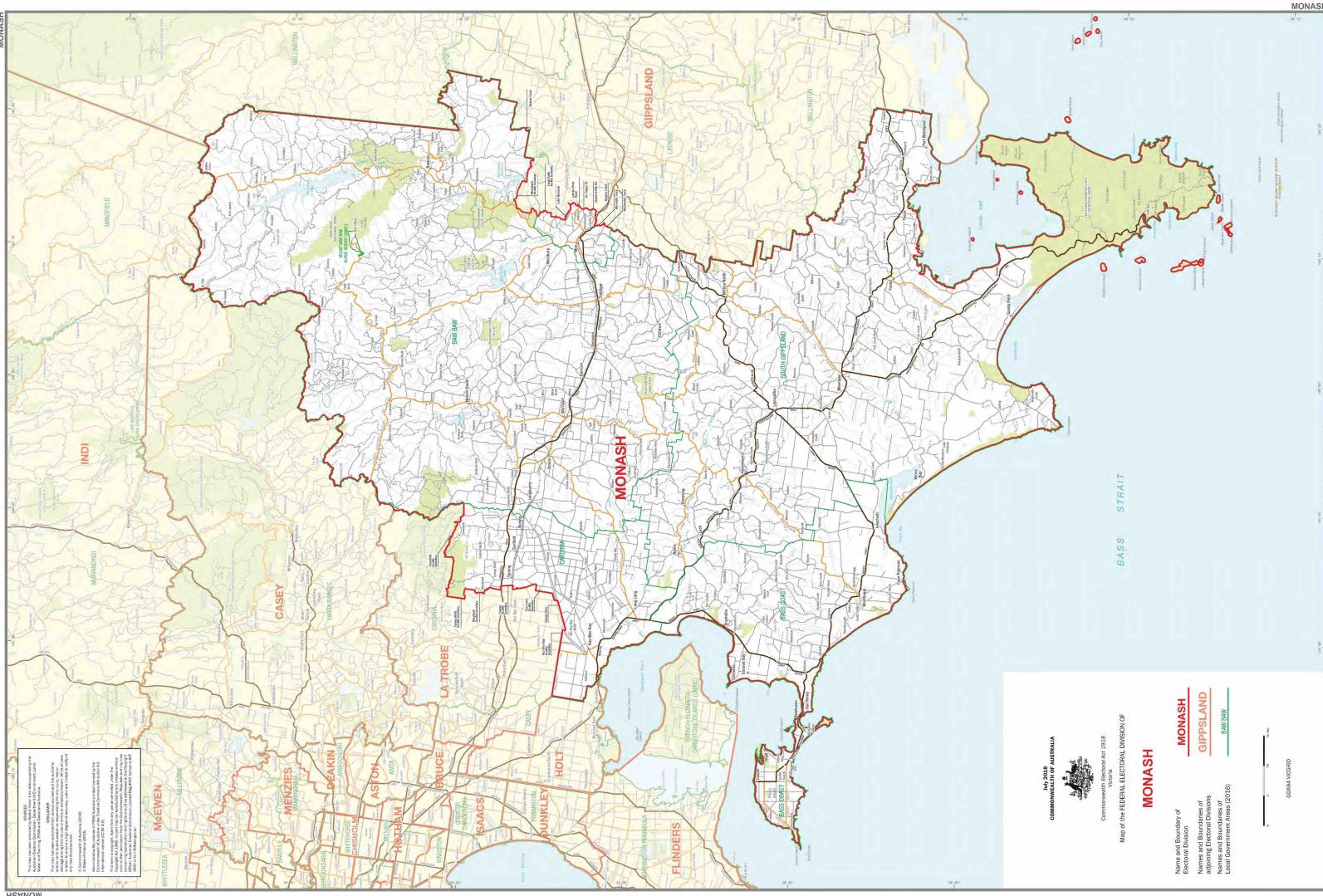
House of Representative Voting Data

```
library(tidyverse)
library(gganimate)
dat <- read_csv("https://results.aec.gov.au/24310/Website/Downloads/HouseD
glimpse(dat)

## Rows: 26,632
## Columns: 14
## $ StateAb          <chr> "ACT", "ACT", "ACT", "ACT", "ACT", "ACT", "ACT",
## $ DivisionID        <dbl> 318, 318, 318, 318, 318, 318, 318, 318, 318,
## $ DivisionNm        <chr> "Bean", "Bean", "Bean", "Bean", "Bean", "Bean",
## $ CountNumber       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ BallotPosition    <dbl> 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4,
## $ CandidateID       <dbl> 33426, 33426, 33426, 33426, 32130, 32130,
## $ Surname           <chr> "FAULKNER", "FAULKNER", "FAULKNER", "FAULKNER",
## $ GivenNm           <chr> "Therese", "Therese", "Therese", "Therese", "Jan
## $ PartyAb           <chr> "AUP", "AUP", "AUP", "AUP", "IND", "IND", "IND",
## $ PartyNm           <chr> "Australian Progressives", "Australian Progress:
## $ Elected            <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"
## $ HistoricElected   <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N"
```

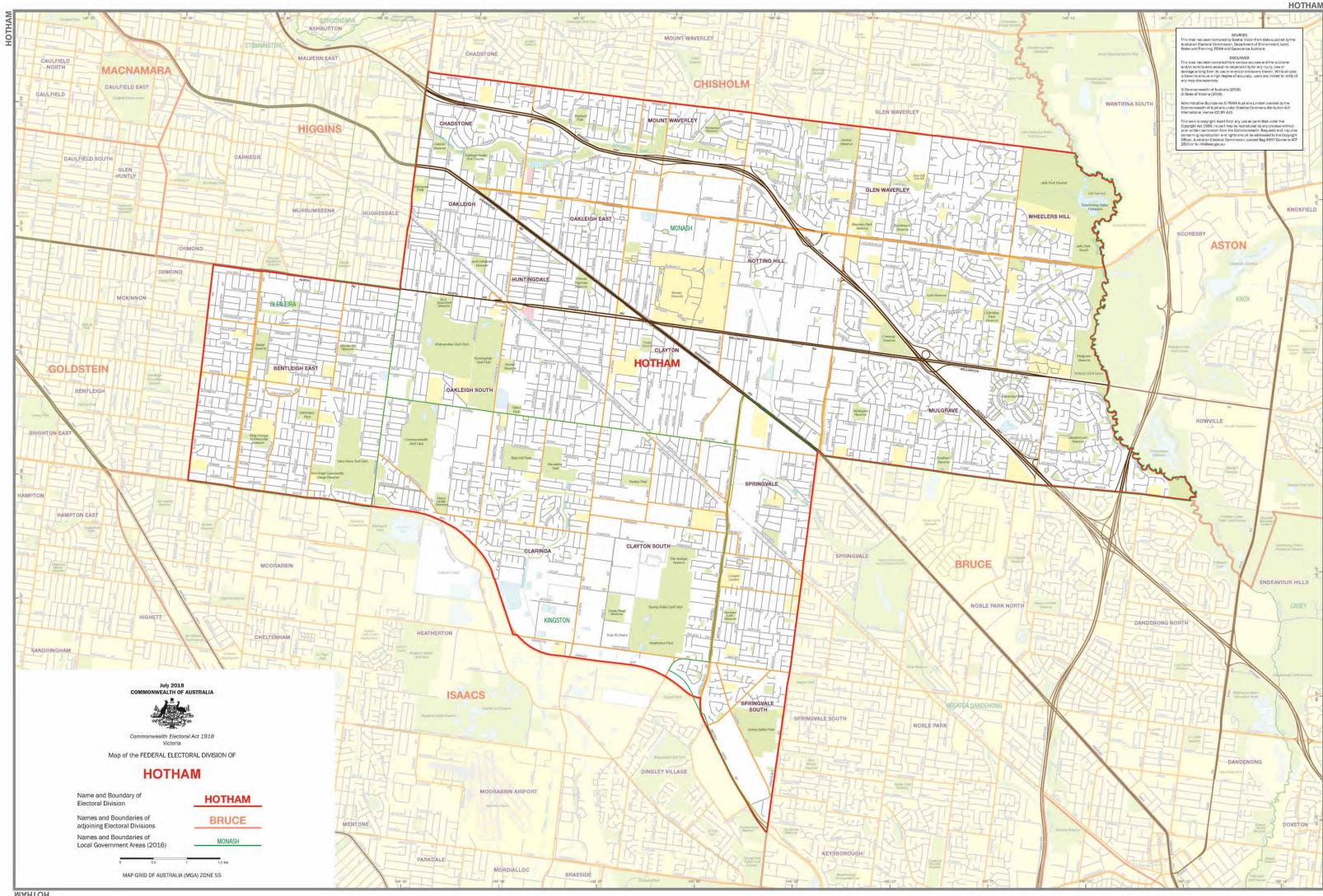
Electoral district of Monash

 ...doesn't include Monash Clayton campus



Electoral district of Hotham

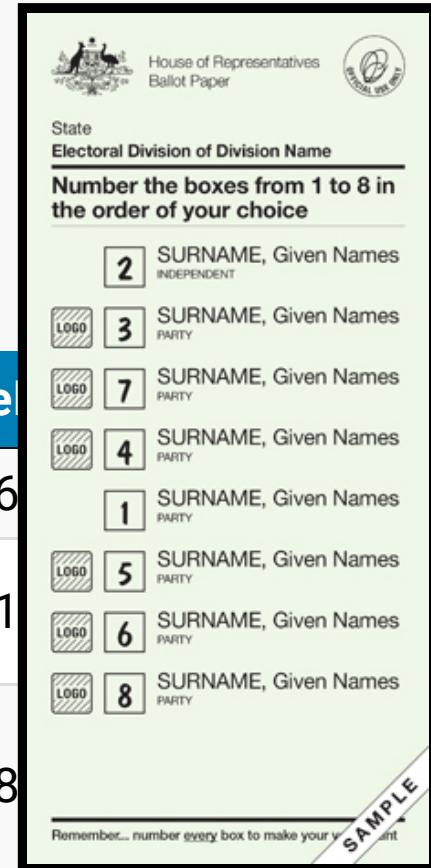
📍 Does include Monash Clayton campus



District: Monash

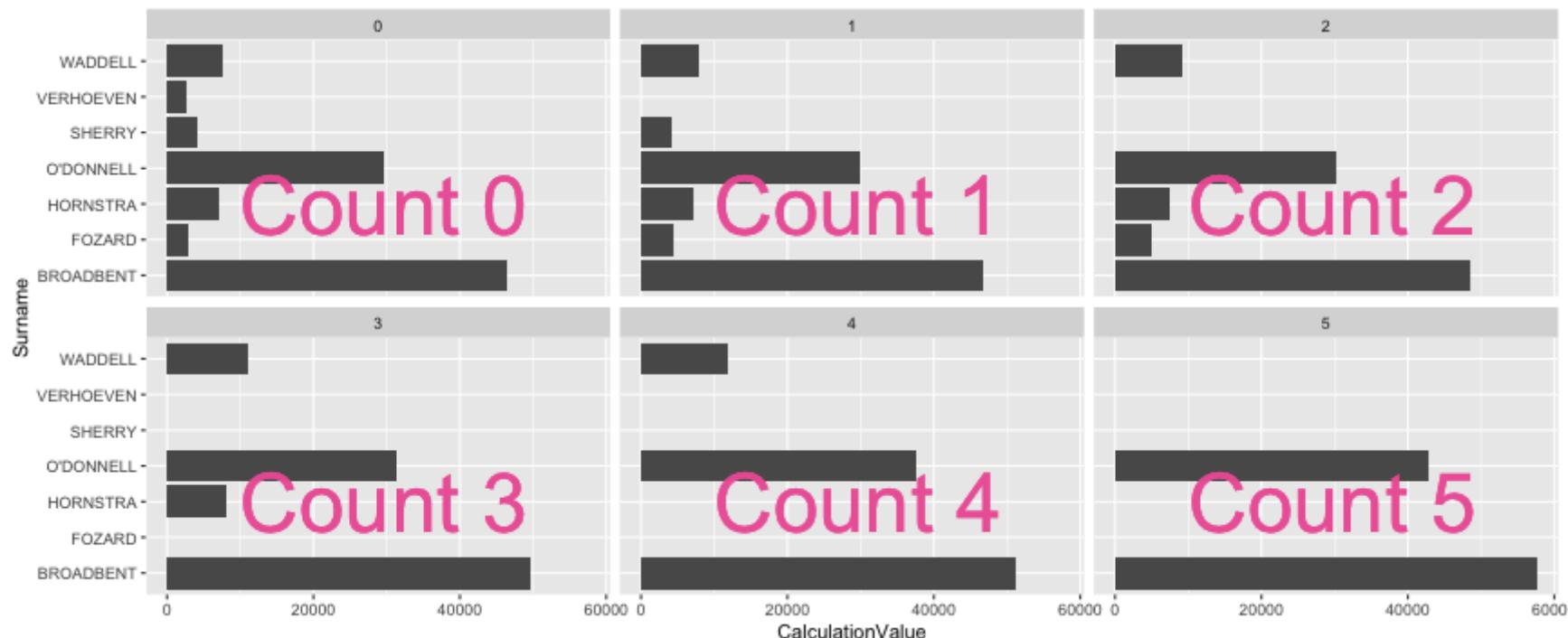
```
dat_monash <- dat %>%
  # get the preference count only
  filter(CalculationType == "Preference Count") %>%
  # get the Monash division
  filter(DivisionNm == "Monash")
```

	StateAb	DivisionID	DivisionNm	CountNumber	BallotPosition	Candidate
1	VIC	323	Monash	0	1	326
2	VIC	323	Monash	0	2	321
3	VIC	323	Monash	0	3	328
4	VIC	323	Monash	0	4	3229



Visualising the counts

```
dat_monash %>%  
  ggplot() +  
  geom_col(aes(x = CalculationValue, y = Surname)) +  
  geom_text(aes(label = paste("Count", CountNumber)),  
            x = 10000, y = 3, size = 16, color = "#ee64a4", alpha = 0.4, h  
  facet_wrap(~CountNumber)
```



... but better to order candidates by counts

```
dat_monash %>%  
  mutate(Surname = fct_reorder(Surname, CalculationValue)) %>%  
  ggplot() +  
  geom_col(aes(x = CalculationValue, y = Surname)) +  
  geom_text(aes(label = paste("Count", CountNumber + 1)),  
            x = 10000, y = 3, size = 16, color = "#ee64a4", alpha = 0.4, h  
  facet_wrap(~CountNumber)
```

Winner:
Russel
Broadbent



House of Representative Voting Animation

Division: La Trobe

VIC

WOOD



CURTIS



REGOROVICH

BAKER

ELL-BROWNE

DEAN

BAKER

Coun

0

20,000

40,000

60,0

Number of Votes

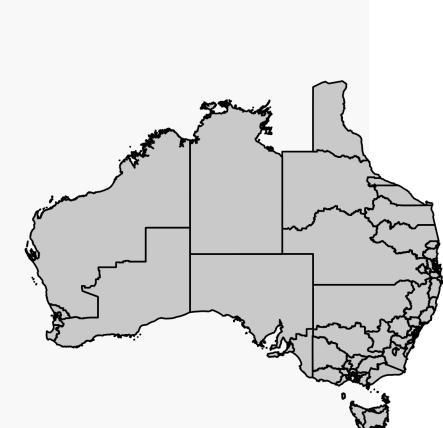


Combining Australian Election and Census Data

eechidna

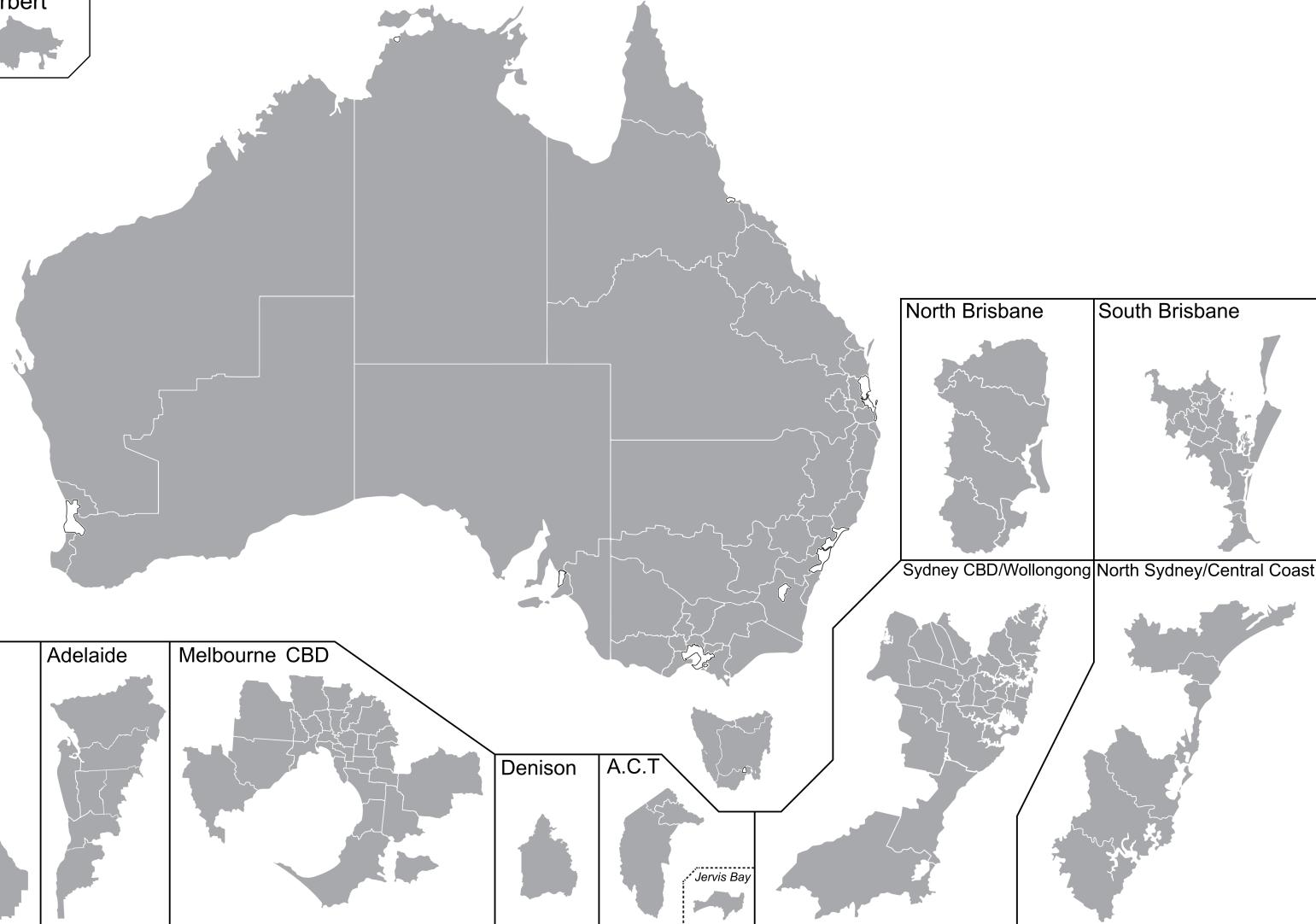
- ✿ eechidna (Exploring Election and Census Highly Informative Data Nationally for Australia) provides data from the Australian Federal elections from 2001-2019 and census information from 2001-2016.
- ✿ It also includes the map data! Read more about getting the shape files [here](#).

```
library(eechidna)
nat_map19 <- nat_map_download(2019)
ggplot(data=nat_map19) +
  geom_polygon(aes(x = long, y = lat, group = group),
               color = "black") +
  theme_void() +
  coord_equal() +
  theme(legend.position="bottom")
```



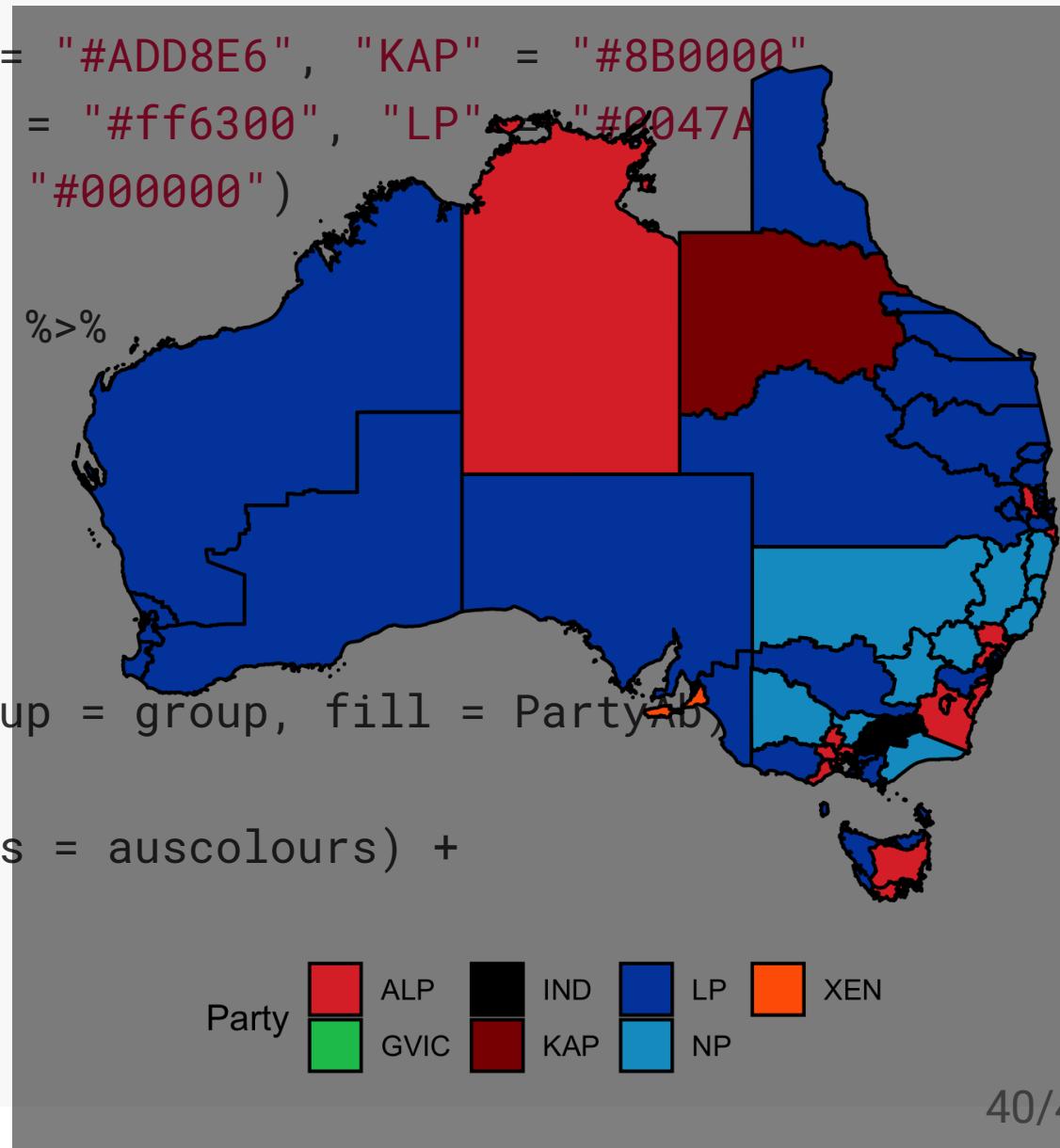
Australian Electorates Divisions

There are 151 electorates in 2019.



Drawing Chloropleth Map with R

```
auscolours <- c("ALP" = "#DE3533", "LNP" = "#ADD8E6", "KAP" = "#8B0000"  
  "GVIC" = "#10C25B", "XEN" = "#ff6300", "LP" = "#2047A1",  
  "NP" = "#0a9cca", "IND" = "#000000")  
  
map_winners <- fp19 %>%  
  mutate(elect_div = toupper(DivisionNm)) %>%  
  filter(Elected == "Y") %>%  
  select(elect_div, PartyAb, PartyNm) %>%  
  left_join(nat_map19, by = "elect_div")  
  
ggplot(data = map_winners) +  
  geom_polygon(aes(x = long, y = lat, group = group, fill = PartyAb,  
                    color = "black")) +  
  scale_fill_manual(name = "Party", values = auscolours) +  
  theme_void() +  
  coord_equal() +  
  theme(legend.position="bottom")
```

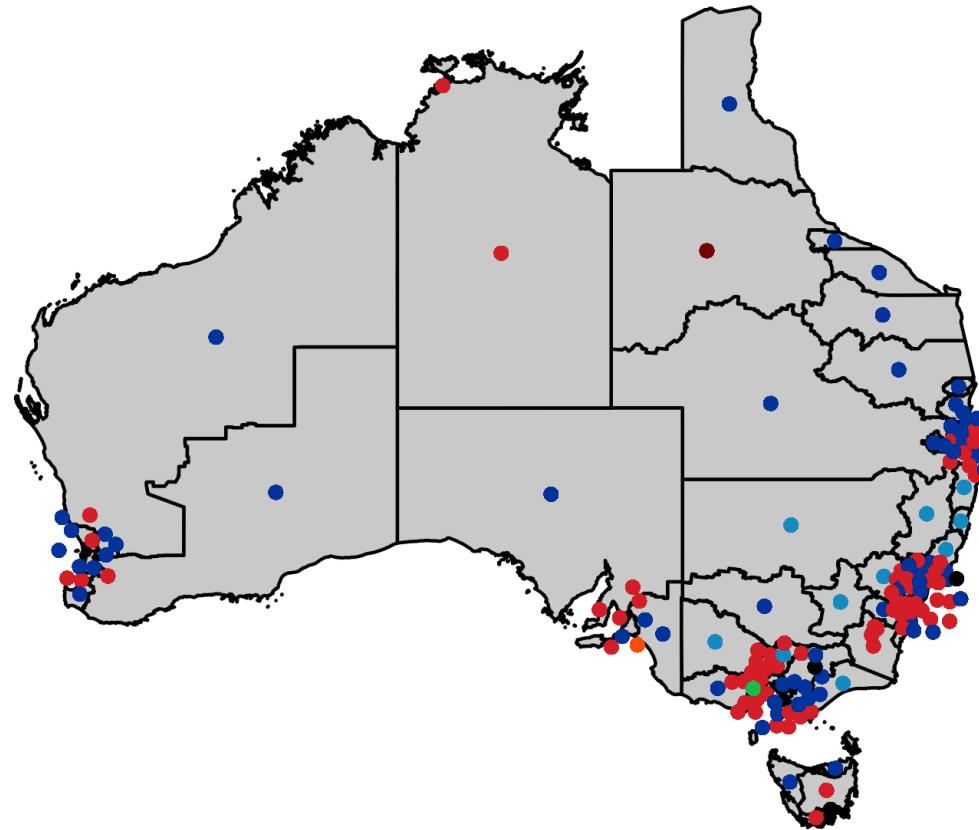


Choloropleth Map

Which party won from looking at this map and by how much?

Liberal/National Coalition: 77
Labor: 68
Greens: 1
Katter's Australian: 1
Centre Alliance: 1
Independents: 3

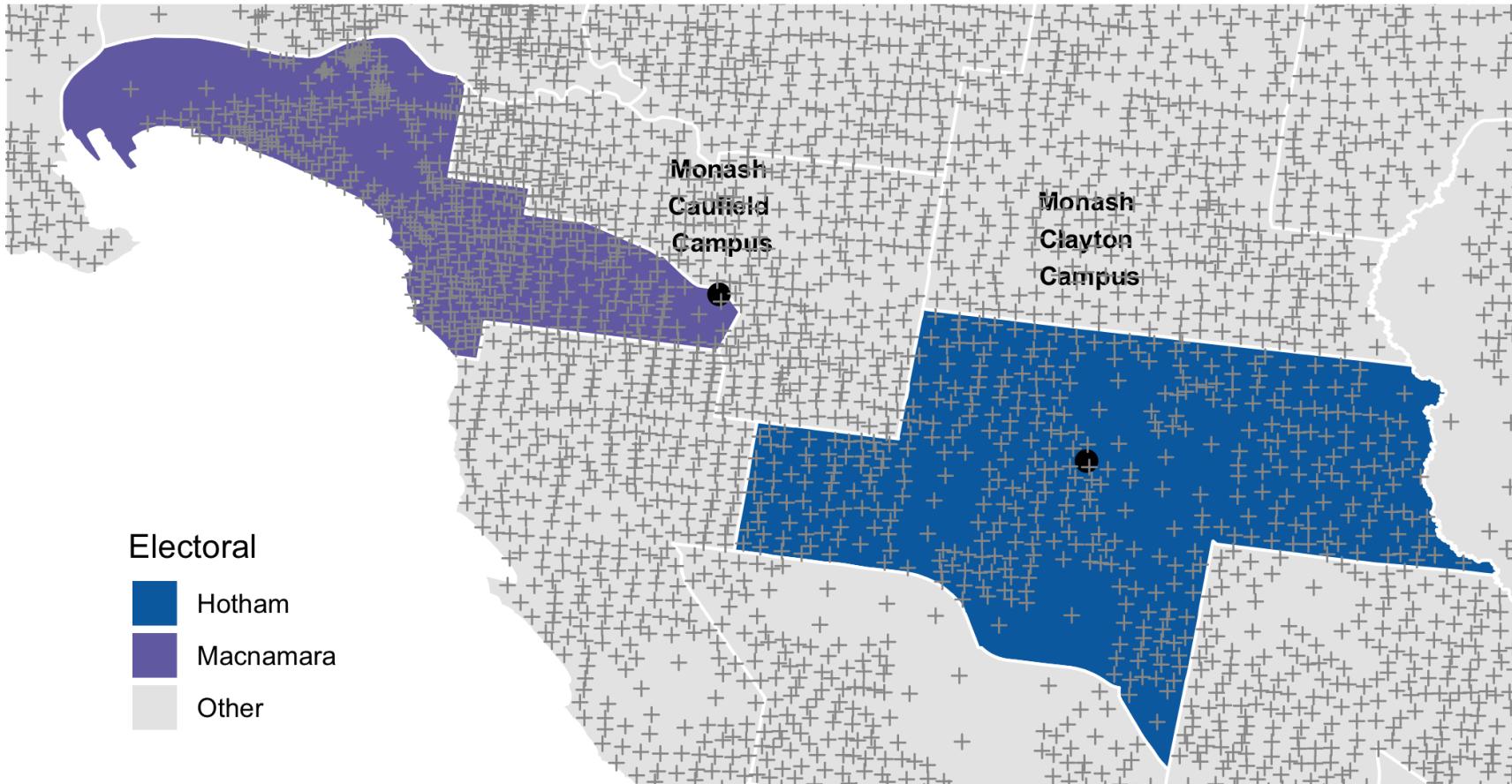
Non-Contiguous, Dorling Cartogram



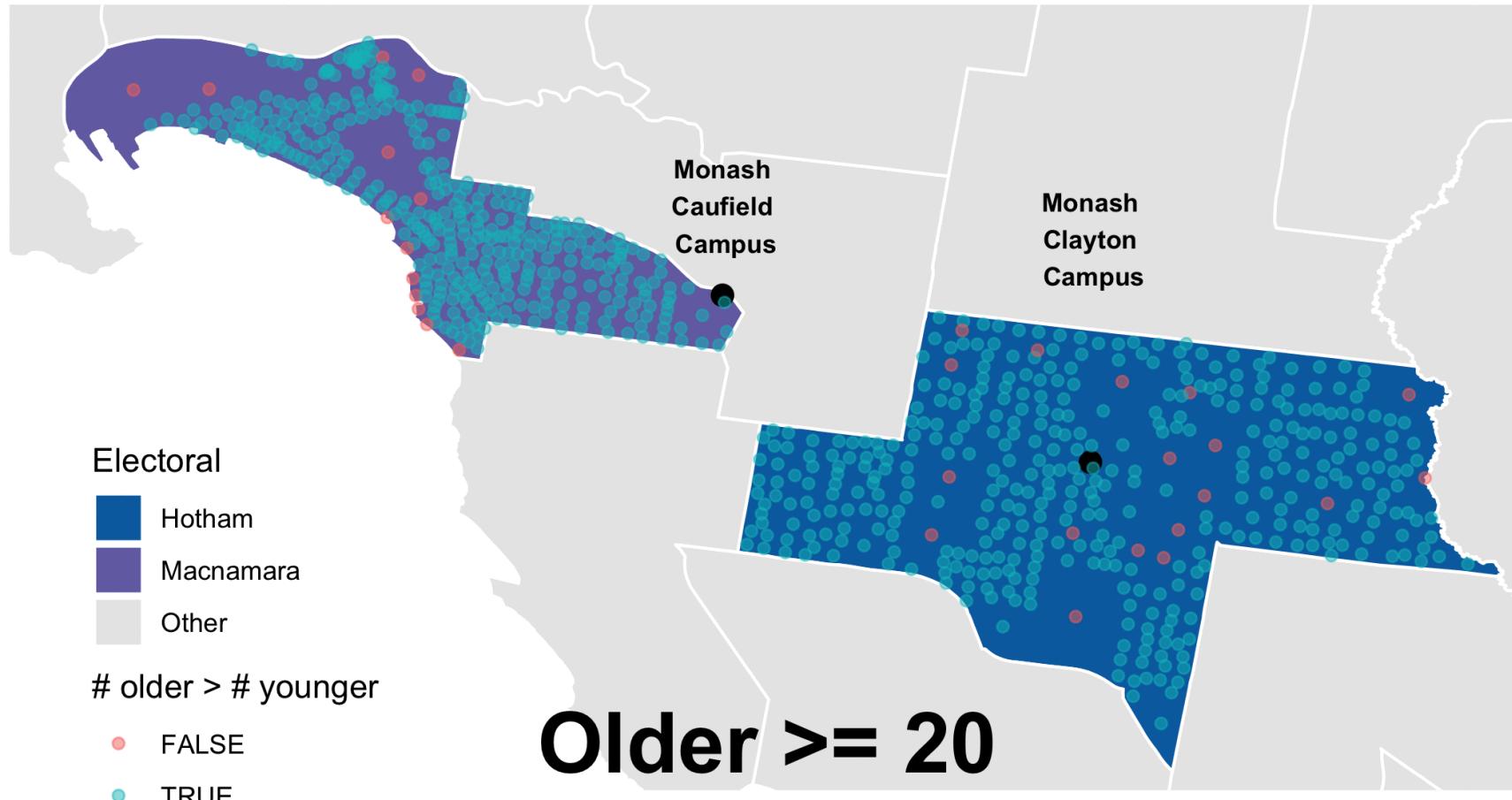
Party	Color
ALP	Red
IND	Black
LP	Blue
XEN	Orange
GVIC	Green
KAP	Maroon
NP	Cyan

Electorate Map

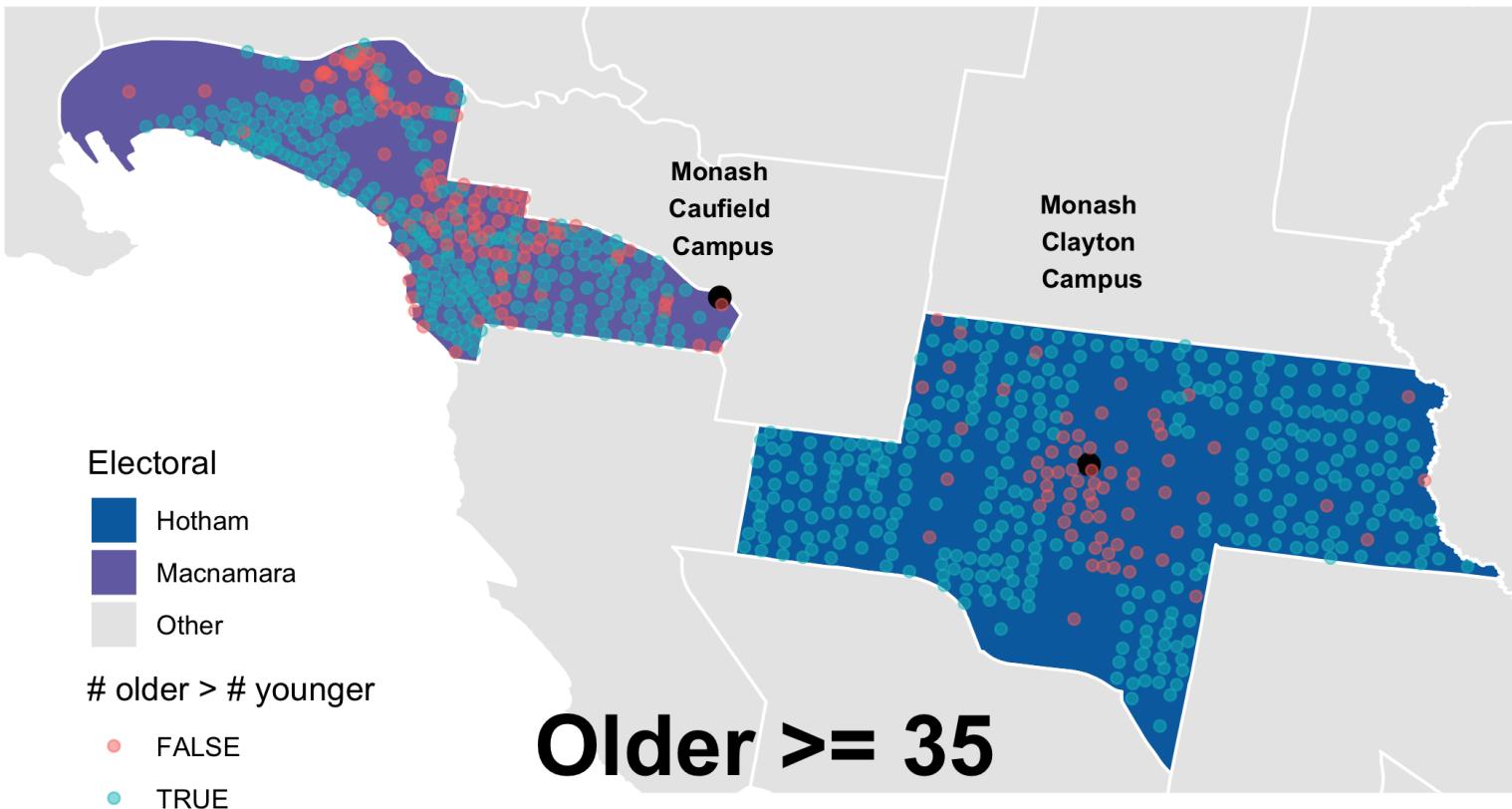
Now superimposed with the centroids of the Statistical Area 1 (SA1) from 2016 Census



Spatial distribution of age group



Spatial distribution of age group



There is a clear cluster of younger people (< 35 years old) centered around Monash clayton campus.

References

- ✿ To install eechinda R-packages use

```
devtools::install_github("emitanaka/eechidna")
```

Normally you should replace emitanaka with ropenscilabs, but current forked repo @emitanaka contains some (untested) bug fixes.

- ✿ Check out the [vignettes](#) for eechidna for more details.
- ✿ Also check out the paper by [Forbes, Cook & Hyndman \(2020\)](#) [Spatial modelling of the two-party preferred vote in Australian federal elections: 2001–2016. *Australian & New Zealand Journal of Statistics* \(to appear\).](#) .
- ✿ The RStudio Cloud Project containing the code for the maps and animations can be found [here](#).

That's it!



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu

