ETC5512: Wild Caught Data

Combining Australian census and election data

Lecturer: Emi Tanaka
Department of Econometrics and Business Statistics
✉ ETC5512.Clayton-x@monash.edu
📅 Week 6
Today you will:

- look at the ABS geographical boundaries for the 2016 census
- integrate data from different sources (census and election) to make exploratory inferences
Recall 2019 Federal Election Data

```r
library(tidyverse)
library(sf)
aec_map <- read_sf(here::here("data/vic-july-2018-esri/E_AUGFN3_region.shp"))

winners_fix <- votes %>%
  mutate(DivisionNm = ifelse(DivisionNm == "McEwen", "Mcewen", DivisionNm)) %>%
  filter(Elected == "Y" & CountNumber == 0 & CalculationType == "Preference Count") %>%
  # get the winner
  right_join(aec_map, by = c("DivisionNm" = "Elect_div")) # join the data

auscolours <- c("ALP" = "#DE3533", "LNP" = "#ADD8E6", "KAP" = "#800000", "GVIC" = "#10C25B", "XEN" = "#ff6300",
                 "LP" = "#0047AB", "NP" = "#0a9cca", "IND" = "#000000")

ggplot(winners_fix) +
  geom_sf(aes(fill = PartyAb, geometry = geometry)) +
  scale_fill_manual(values = auscolours)
```

There are two sources of data:

1. Electoral boundary
2. The votes for candidates in each electorate
Recall 2016 ABS Census Data

ABS Census 2016
GeoPackages
A GeoPackage (GPKG) is an open, non-proprietary, platform-independent and standards-based data format for geographic information system implemented as a SQLite database container. Defined by the Open Geospatial Consortium (OGC) with the backing of the US military and published in 2014, GeoPackage has seen widespread support from various government, commercial, and open source organizations.

— Wikipedia

Recall: OGC also defines the WKT
ABS GeoPackage

- Or use the `strayr` package! We'll use the one from the ABS website instead.

```r
ggeopath <- here::here("data/Geopackage_2016_EIUWA_for_VIC/census2016_eiuwa_vic_short.gpkg")
st_layers(ggeopath)
```

## Driver: GPKG
## Available layers:

<table>
<thead>
<tr>
<th>layer_name</th>
<th>geometry_type</th>
<th>features</th>
<th>fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>census2016_eiuwa_vic_ced_short</td>
<td>39</td>
<td>489</td>
<td></td>
</tr>
<tr>
<td>census2016_eiuwa_vic_gccsa_short</td>
<td>4</td>
<td>489</td>
<td></td>
</tr>
</tbody>
</table>
The Australian Statistical Geography Standard (ASGS)

Non-ABS Structures
Structures not defined by the ABS. They are updated annually.

Data Availability for Non-ABS Regions
Data availability for non-ABS Regions is provided for the following:
- Urban pop. & persons
- Marriages
- Births
- Deaths
- Electorates
- Schools
- Hospitals
- Economic Industries
- Transport

Data Availability for Mesh Blocks
Mesh Blocks are available as building blocks for other data availability. Details on data availability can be found on the Australian Bureau of Statistics website.
The number of regions for each layer

```r
st_layers(geopath) %>%
  # make it into a data.frame first
tibble(!!!!.) %>%
  # then you can the dplyr operations
dplyr::arrange(features)

## A tibble: 16 × 5
##  name                             geomtype       driver features fields
##  <chr>                            <list>         <chr>     <dbl>  <dbl>
## 1 census2016_eiuwa_vic_ste_short <chr [1]>     GPKG          1    489
## 2 census2016_eiuwa_vic_gccsa_short <chr [1]>   GPKG          4    489
## 3 census2016_eiuwa_vic_ra_short  <chr [1]>     GPKG          6    489
## 4 census2016_eiuwa_vic_sos_short <chr [1]>     GPKG         12    489
## 5 census2016_eiuwa_vic_sosr_short <chr [1]>    GPKG         19    489
## 6 census2016_eiuwa_vic_sa4_short <chr [1]>     GPKG         22    489
## 7 census2016_eiuwa_vic_sua_short <chr [1]>     GPKG         39    489
## 8 census2016_eiuwa_vic_sed_short <chr [1]>     GPKG         68    489
## 9 census2016_eiuwa_vic_sa3_short <chr [1]>     GPKG         82    489
## 10 census2016_eiuwa_vic_lga_short <chr [1]>    GPKG         90    489
## 11 census2016_eiuwa_vic_sed_short <chr [1]>    GPKG        353    489
## 12 census2016_eiuwa_vic_ucl_short <chr [1]>    GPKG        464    489
## 13 census2016_eiuwa_vic_poa_short <chr [1]>    GPKG        698    489
## 14 census2016_eiuwa_vic_ssc_short <chr [1]>    GPKG        2931    489
## 15 census2016_eiuwa_vic_sa1_short <chr [1]>    GPKG      14073    489
```
Data in the layer

```r
# Load necessary library
library(sf)

# Read shapefile from geoserver
vicmap_ste <- read_sf(geopath, layer = "census2016_eiuwa_vic_sa1_short")

# Print geometry information
vicmap_ste$geom

## Geometry set for 14073 features  (with 4 geometries empty)
## Geometry type: MULTIPOLYGON
## Dimension:      XY
## Geodetic CRS:  GDA94
## First 5 geometries:

# Print a summary of the dataset
str(vicmap_ste)

## sf [14073 × 490] (S3: sf.tbl_df/tbl/data.frame)
## $ sa1_7digitcode_2016 : chr [1:14073] "2145523" "2111727" "2104305" "2128614" ...
## $ Median_age_persons : num [1:14073] 35 26 45 43 43 48 38 48 35 54 ...
## $ Median_mortgage_repay_monthly: num [1:14073] 1419 2134 2167 1517 2600 ...
## $ Median_tot_prsnl_inc_weekly : num [1:14073] 659 403 672 671 763 477 595 586 521 445 ...
## $ Median_rent_weekly           : num [1:14073] 350 462 340 250 400 312 418 215 280 150 ...
## $ Median_tot_fam_inc_weekly    : num [1:14073] 1640 1624 1906 1279 2437 ...
## $ Average_num_psns_per_bedroom : num [1:14073] 0.8 1.0 0.8 0.8 0.8 0.8 0.7 0.8 0.8 ...
## $ Median_tot_hhd_inc_weekly    : num [1:14073] 1525 1031 1805 1279 1906 ...
## $ Average_household_size       : num [1:14073] 2.7 2.1 2.8 2.5 2.7 2.7 2.1 2.4 1.8 ...
## $ M_Neg_Nil_income_15_19_yrs   : num [1:14073] 9 7 8 6 6 0 3 0 3 3 ...
## $ M_Neg_Nil_income_20_24_yrs   : num [1:14073] 0 6 0 0 0 0 0 4 0 4 0 ...
## $ M_Neg_Nil_income_25_34_yrs   : num [1:14073] 0 5 0 0 0 0 0 0 3 3 ...
```
State or Territory (STE)

vicmap_ste <- read_sf(geopath, layer = "census2016_eiuwa_vic_ste_short")
ggplot(vicmap_ste) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))

nrow(vicmap_ste)

## [1] 1
Greater Capital City Statistical Areas (GCCSA)

- Each region with variable population

```r
vicmap_gccsa <- read_sf(geopath, layer = "census2016_eiuwa_vic_gccsa_short")
ggplot(vicmap_gccsa) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_gccsa)
## [1] 4
```
Section of State (SOS)

- Major urban, other urban, bounded locally & rural balance

```r
vicmap_sos <- read_sf(geopath, layer = "census2016_eiuwa_vic_sos_short")
ggplot(vicmap_sos) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_sos)
## [1] 6
```
Remoteness Areas (RA)

```r
vicmap_ra <- read_sf(geopath, layer = "census2016_eiuwa_vic_ra_short")
ggplot(vicmap_ra) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```
nrow(vicmap_ra)
## [1] 6
```
Section of State Ranges (SOSR)

```r
vicmap_sosr <- read_sf(geopath, layer = "census2016_eiuwa_vic_sosr_short")
ggplot(vicmap_sosr) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```
nrow(vicmap_sosr)
```

```
## [1] 12
```
Statistical Area Level 4 (SA4)

- Each region with population of 100,000 - 500,000

```r
vicmap_sa4 <- read_sf(geopath, layer = "census2016_eiuwa_vic_sa4_short")
ggplot(vicmap_sa4) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```
nrow(vicmap_sa4)
## [1] 19
```
Significant Urban Areas (SUA)

```r
vicmap_sua <- read_sf(geopath, layer = "census2016_eiuwa_vic_sua_short")
ggplot(vicmap_sua) + geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_sua)
## [1] 22
```
Commonwealth Electoral Division (CED)

```r
vicmap_ced <- read_sf(geopath, layer = "census2016_eiuwa_vic_ced_short")
ggplot(vicmap_ced) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_ced)
## [1] 39
```
Statistical Area Level 3 (SA3)

- Each region with population of 30,000 - 130,000

```r
vicmap_sa3 <- read_sf(geopath, layer = "census2016_eiuwa_vic_sa3_short")
ggplot(vicmap_sa3) + geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_sa3)
## [1] 68
```
Local Government Area (LGA)

```r
vicmap_lga <- read_sf(geopath, layer = "census2016_eiuwa_vic_lga_short")
ggplot(vicmap_lga) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```
> nrow(vicmap_lga)
## [1] 82
```
State Electoral Division (SED)

```
> vicmap_sed <- read_sf(geopath, layer = "census2016_eiuwa_vic_sed_short")
> ggplot(vicmap_sed) + geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```
> nrow(vicmap_sed)
## [1] 90
```
Urban Centres and Localities (UCL)

```
vicmap_ucl <- read_sf(geopath, layer = "census2016_eiuwa_vic_ucl_short")
ggplot(vicmap_ucl) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```
nrow(vicmap_ucl)
## [1] 353
```
Statistical Area Level 2 (SA2)

- Each region with populations in the range of 3,000-25,000

```r
vicmap_sa2 <- read_sf(geopath, layer = "census2016_eiuwa_vic_sa2_short")
ggplot(vicmap_sa2) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_sa2)
## [1] 464
```
Postal Areas (POA)

```r
census2016_eiuwa_vic_poa_short"

# [1] 698
State Suburbs (SSC)

```r
df <- read_sf(geopath, layer = "census2016_eiuwa_vic_ssc_short")
ggplot(df) + geom_sf(aes(geometry = geom, fill = Median_age_persons))

nrow(df)
## [1] 2931
```
Statistical Area Level 1 (SA1)

- Each region with a population of range 200-800

```r
vicmap_sa1 <- read_sf(geopath, layer = "census2016_eiuwa_vic_sa1_short")
ggplot(vicmap_sa1) +
  geom_sf(aes(geometry = geom, fill = Median_age_persons))
```

```r
nrow(vicmap_sa1)
## [1] 14073
```
Electorate boundary vs Census boundary

Estimate a median age for an electorate
Comparing SED 2016 and electorates divisions 2019

See here for `winners_fix` data was.

```r
ggplot(winners_fix) + geom_sf(data = vicmap_sed, aes(geometry = geom, fill = Median_age_persons), alpha = 1, color = "white", size = 2) + geom_sf(aes(geometry = geometry), fill = "transparent", color = "red") + coord_sf(xlim = c(144.95, 145.24), ylim = c(-38.05, -37.85))
```
There are 10 SED regions that intersect with Hotham electorate.
There are 5 SED areas with at least 5% intersection with the electoral area.

How would you characterise the median age for Hotham?
Closer look 🕵️ Hotham electorate 📊

Strategy 1

\[
\text{sort}(\text{sed_intersect2}\$\text{Median}\_\text{age}\_\text{persons})
\]

## [1] 32 35 39 40 40

Strategy 2

\[
\text{mean}(\text{sed_intersect2}\$\text{Median}\_\text{age}\_\text{persons})
\]

## [1] 37.2

Strategy 3

\[
\text{weighted.mean}(\text{sed_intersect2}\$\text{Median}\_\text{age}\_\text{persons}, \\
\text{sed_intersect2}\$\text{perc}\_\text{area})
\]

## [1] 36.35205
Closer look 🎫 Hotham electorate 📈

```r
sa1_intersect <- vicmap_sa1 %>%
  filter(st_intersects(geom,
                      electorate$geometry,
                      sparse = FALSE)[,1])

sa1_intersect2 <- sa1_intersect %>%
  mutate(geometry = st_intersection(geom, electorate$geometry),
         perc_area = 100 * st_area(geometry) / st_area(geom),
         perc_area = as.numeric(perc_area)) %>%
  filter(perc_area > 5)

ggplot(sa1_intersect) +
  geom_sf(color = "red",
          aes(fill = Median_age_persons,
              geometry = geom)) +
  geom_sf(data = electorate, color = "white", size = 2,
          fill = "transparent",
          aes(geometry = geometry)) +
  theme(legend.position = "bottom")
```
Closer look 🧐 Hotham electorate

**Strategy 1**

\[
\text{fivenum}(\text{sa1_intersect2}$\text{Median_age_persons})
\]

## [1] 0 34 38 42 82

**Strategy 2**

\[
\text{mean}(\text{sa1_intersect2}$\text{Median_age_persons})
\]

## [1] 37.38235

**Strategy 3**

\[
\text{weighted.mean}(\text{sa1_intersect2}$\text{Median_age_persons}, \text{sa1_intersect2}$\text{perc_are})
\]

## [1] 37.35034

**Strategy 4**

\[
\text{ggplot}(\text{sa1_intersect2, aes(x = Median_age_persons)) + geom_histogram(binwidth = 1)}
\]
Closer look 🧐 Zero median age

```r
sa1_intersect2 %>%
  filter(Median_age_persons==0) %>%
  ggplot() +
  geom_sf() +
  geom_sf(data = electorate, color = "red",
           fill = "transparent",
           aes(geometry = geometry))
```
### Closer look 🕵️‍♂️ Hotham electorate 6

#### Before

<table>
<thead>
<tr>
<th>Strategy 1</th>
<th>fivenum(sa1_intersect2$Median_age_persons)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1] 0 34 38 42 82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy 2</th>
<th>mean(sa1_intersect2$Median_age_persons)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1] 37.38235</td>
</tr>
</tbody>
</table>

| Strategy 3  | weighted.mean(sa1_intersect2$Median_age_persons,                        |
|             |                                                          |
|             | [1] 37.35034                                  |

#### After

```r
sa1_intersect3 <- sa1_intersect2 %>%
  filter(Median_age_persons != 0)
```

<table>
<thead>
<tr>
<th>Strategy 1</th>
<th>fivenum(sa1_intersect3$Median_age_persons)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1] 20 34 38 42 82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy 2</th>
<th>mean(sa1_intersect3$Median_age_persons)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1] 38.61266</td>
</tr>
</tbody>
</table>

| Strategy 3  | weighted.mean(sa1_intersect3$Median_age_persons, sa1_intersect3$perc_a                        |
|             |                                                                 |
|             | [1] 38.58491                                |
```r
sa1_intersect4 <- sa1_intersect %>%
  mutate(centroid = st_centroid(geom))

ggplot(sa1_intersect4) +
  geom_sf(data = electorate,
    aes(geometry = geometry), size = 2, fill = "grey60") +
  geom_sf(aes(geometry = centroid, color = Median_age_persons),
    size = 0.5, shape = 3) +
  scale_color_viridis_c(name = "Median age", option = "magma")
```
Closer look 🕵️‍♂️ Hotham electorate

```r
sa1_intersect5 <- sa1_intersect4 %>%
  filter(st_intersects(centroid,
    electorate$geometry,
    sparse = FALSE)[,1],
    Median_age_persons!=0)
```

**Strategy 1**

```r
fivenum(sa1_intersect5$Median_age_persons)
## [1] 20 34 38 42 82
```

**Strategy 2**

```r
mean(sa1_intersect5$Median_age_persons)
## [1] 38.58015
```

**Strategy 4**

```r
ggplot(sa1_intersect5, aes(x = Median_age_persons)) +
geom_histogram(binwidth = 1)
```

- There are many ways to characterise an electorate.
- Estimates of median age of an electorate is more consistent using SA1 map data than SED map data.
Summary

- We looked at mapping the 2016 census boundaries and projected a summary of the census variable (i.e. median age) onto a 2019 electoral district.
