Analysis of Census 2016

Hugh Parsonage

2017-07-17

Census 2016

This package contains selected totals from the Community Profiles time series data, released by the Australian Bureau of Statistics as part of the Census 2016 release.

```r
library(knitr)
pkgs_suggested <- c("magrittr",
                    "ggplot2",
                    "scales",
                    "ggrepel")
suggested_packages_usable <-
  all(vapply(pkgs_suggested, requireNamespace, logical(1), quietly = TRUE))

knitr::opts_chunk$set(eval = suggested_packages_usable,
                      # dev = "png",
                      fig.width = 8,
                      fig.height = 6)
```

```r
library(magrittr)
library(data.table)
library(Census2016)
```

```
## For additional data packages for the 2016 Census, visit https://github.com/HughParsonage/Census2016.
```

```r
library(ggplot2)
library(scales)
library(ggrepel)
```

```r
data.kable <- function(DT) {
  current_knitr.kable.NA <- options("knitr.kable.NA")
  options(knitr.kable.NA = '...')
  on.exit(options(knitr.kable.NA = current_knitr.kable.NA))
  if (nrow(DT) > 50) {
    middle_row <- as.data.table(matrix(nrow = 1, ncol = ncol(DT)))
    setnames(middle_row, seq_along(middle_row), names(DT))
    DT_topn <- rbind(head(DT),
                     middle_row,
                     tail(DT))
    kable(DT_topn, format.args = list(big.mark = ","))
  } else {
    kable(DT, format.args = list(big.mark = ","))
  }
}
```

There is one function `see_question()` and 6 data sets.

Census_wide_by_SA2_year

This is a simple `data.table` of multiple variables for each statistical area 2 (`SA2`)-census year combination. The columns are ordered roughly by the order of the questions on the Census form. Not all values are available to satisfy CRAN's limits on package size.

Both the `sa2_code` and `sa2_name` are provided for convenience.

Median mortgage vs median income

```
Census2016_wide_by_SA2_year %>%
  .[year == 2016] %>%
  .[, .(sa2_name, persons, median_household_income, median_annual_mortgage)] %>%
  .[order(median_annual_mortgage)] %>%
  data.kable
```

sa2_name	persons	median_household_income	median_annual_mortgage
Deua - Wadbilliga	25	44,148	0
Port Kembla Industrial	10	0	0
Illawarra Catchment Reserve	8	0	0
Prospect Reservoir	40	0	0
Banksmeadow	18	0	0
Port Botany Industrial	10	0	0
. . .	. . .	. . .	. . .
Cottesloe	7,375	138,788	39,000
Lilli Pilli - Port Hacking - Dolans Bay	3,148	155,220	39,300
Nedlands - Dalkeith - Crawley	18,534	118,976	39,600
Rose Bay - Vaucluse - Watsons Bay	11,840	150,228	41,604
Balgowlah - Clontarf - Seaforth	20,186	145,496	41,604
Hunters Hill - Woolwich	10,345	142,064	42,000

```
Census2016_wide_by_SA2_year %>%
  .[year == 2016] %>%
  .[, .(sa2_name, persons, median_household_income, median_annual_mortgage)] %>%
  .[median_annual_mortgage > 0] %>%
  .[, mortgage_less_income := median_annual_mortgage - median_household_income] %>%
  .[, text := NA_character_] %>%
  .[, color := "black"] %>%
  .[order(mortgage_less_income)] %>%
  .[.N:1 <= 5, text := sa2_name] %>%
  .[.N:1 <= 5, color := "red"] %>%
  .[1:.N <= 5, text := sa2_name] %>%
  .[1:.N <= 5, color := "blue"] %>%
  ggplot(aes(x = median_household_income,
           y = median_annual_mortgage,
           size = persons,
           alpha = persons,
           color = color)) +
  geom_point() +
  scale_color_identity() +
  scale_size(labels = comma) +
  scale_alpha_continuous(labels = comma,
```

```
                    range = c(0, 0.5)) +
scale_x_continuous("Median annual household income", labels = dollar) +
scale_y_continuous("Median annual mortgage", labels = dollar) +
geom_label_repel(aes(label = text),
                 alpha = 1,
                 na.rm = TRUE) +
ggtitle("High-income households live alongside high-mortgage households",
        subtitle = paste0("SA2s, 2016 with 5 highest (red) or lowest (blue)",
                          " nonzero mortgage relative to income"))
```


High−income households live alongside high−mortgage households
SA2s, 2016 with 5 highest (red) or lowest (blue) nonzero mortgage relative to income

Changes from previous years

In addition to the 2016 data, the package also includes 2006 and 2011 census data as part of the time series. The ABS has released these data series to be comparable; even though the SA2 boundaries have changed you may assume that they refer to the same geographic area.

We can see that Mandarin has become much more common

```
languages_spoken_by_year <-
  Census2016_languages %>%
  .[, .(persons = sum(persons)), keyby = .(language, year)] %>%
  setorder(-year, -persons) %>%
  .[]
```

```
languages_spoken_by_year %>%
  # Examine the top six languages,
  # leave the others unlabelled and grey
  .[language %in% languages_spoken_by_year$language[1:6],
    Language := language] %>%
  .[year == 2016, text := Language] %>%
  .[, Language := reorder(Language, -persons)] %>%
  ggplot(aes(x = year,
             y = persons,
             group = language,
             color = Language,
             label = text)) +
  geom_line() +
  scale_y_continuous(label = comma) +
  geom_text_repel(na.rm = TRUE,
                  fontface = "bold",
                  force = 1.5,
                  nudge_x = 0.5)
```
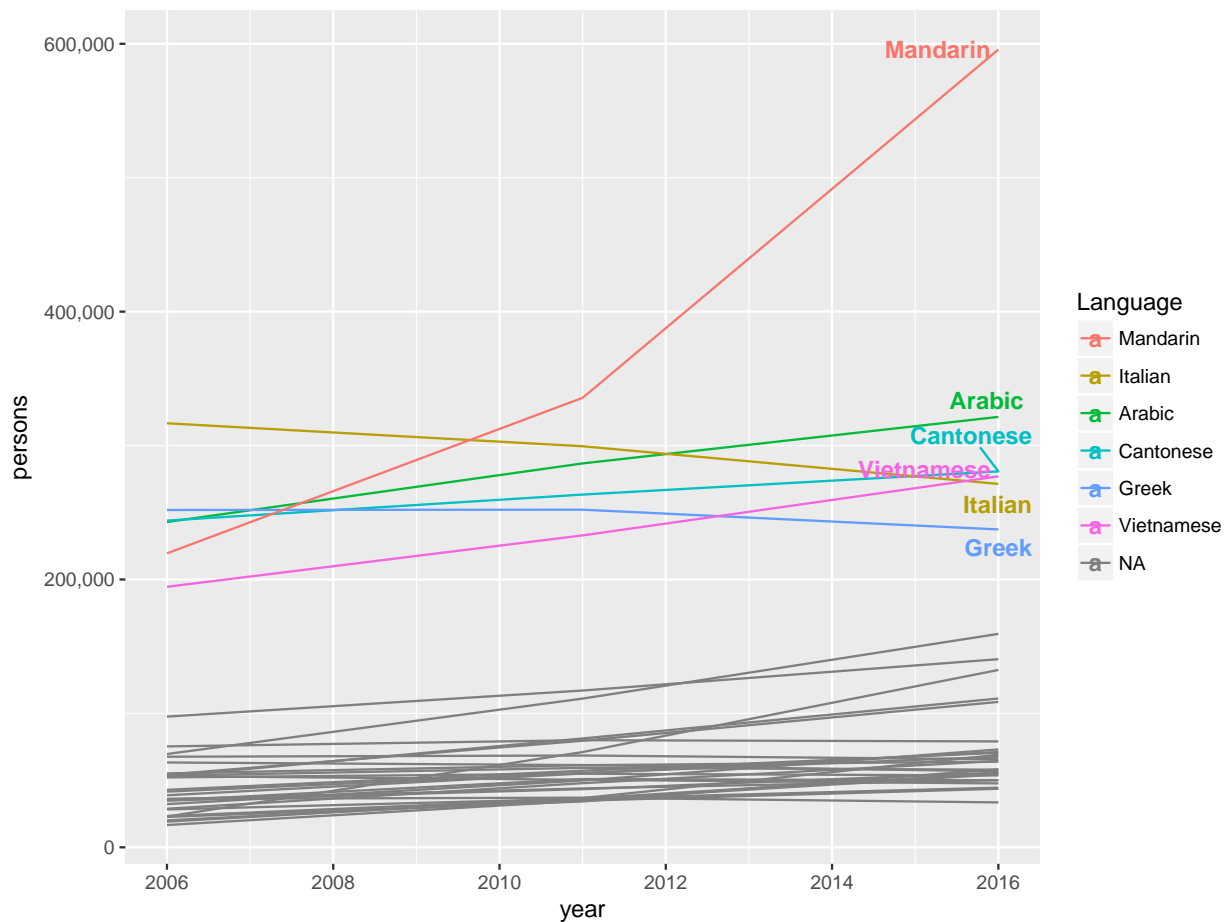

see_question

Although `Census2016` is intended as a data-only package, there is one function, `see_question`.

It is frequently useful to view the actual question that was asked when looking at survey data. `see_question` provides a convenient way to do this without leaving RStudio (or even your keyboard). There are two methods: `see_question.numeric` takes a question number and prints it.

```
see_question(3)
```


The other method is dispatched when one of the two-dimensional tables is supplied. This method returns the relevant question to the data set. For example,

```
see_question(Census2016_ancestories)
```

(The data input is returned invisibly.)