Census Data Cleaning

Scott White

2019-04-18

```
install.packages(c("Hmisc", "tidyverse", "here"))
```

```
library(Hmisc)
library(tidyverse)
library(here)
```

```
Data <- read_csv(here("Final Project", "data", "Census_-_Census_Families.csv"),
                 col_types = cols(
                   `Census Year` = col_factor(),
                   `Boundary Name` = col_character(),
                   `Boundary Type` = col_character(),
                   .default = col_double()),
                 na = "-")
```

Remove the variables: Boundary Number

```
Data <- Data %>% select(-`Boundary Number`)
```

```
Data
```

```
## # A tibble: 477 x 76
##    `Census Year` `Boundary Type` `Boundary Name` `Size 2 Person`
##    <fct>         <chr>           <chr>                    <dbl>
##  1 2001          City            City of Winnip~          78770
##  2 2001          City            Downtown Winni~           1620
##  3 2001          City            Inner City               14830
##  4 2001          City            Non-Inner City           63940
##  5 2001          City            Winnipeg Censu~          84720
##  6 2001          Community Area  Assiniboine So~           4270
##  7 2001          Community Area  Downtown Commu~           7150
##  8 2001          Community Area  Fort Garry Com~           7670
##  9 2001          Community Area  Inkster Commun~           2805
## 10 2001          Community Area  Point Douglas ~           4405
## # ... with 467 more rows, and 72 more variables: `Size 3 Person` <dbl>,
## #   `Size 4 Person` <dbl>, `Size Over 4 Person` <dbl>, `Size Total
## #   Families` <dbl>, `Size Average Person` <dbl>, `Size Average Children
## #   Home` <dbl>, `Structure All Families` <dbl>, `Structure All
## #   Couples` <dbl>, `Structure Married Couples` <dbl>, `Structure Married
## #   Couple No Children Home` <dbl>, `Structure Married Couple Children
## #   Home` <dbl>, `Structure Married Couple 1 Child` <dbl>, `Structure
## #   Married Couple 2 Children` <dbl>, `Structure Married Couple 3+
## #   Children` <dbl>, `Structure Common Law Couple` <dbl>, `Structure
## #   Common Law Couple No Children Home` <dbl>, `Structure Common Law
## #   Couple Children Home` <dbl>, `Structure Common Law Couple 1
## #   Child` <dbl>, `Structure Common Law Couple 2 Children` <dbl>,
## #   `Structure Common Law Couple 3+ Children` <dbl>, `Structure One Parent
## #   Total` <dbl>, `Structure One Parent Female` <dbl>, `Structure One
## #   Parent Female 1 Child` <dbl>, `Structure One Parent Female 2
## #   Children` <dbl>, `Structure One Parent Female 3+ Children` <dbl>,
```

```
## #   `Structure One Parent Male` <dbl>, `Structure One Parent Male 1
## #   Child` <dbl>, `Structure One Parent Male 2 Children` <dbl>, `Structure
## #   One Parent Male 3+ Children` <dbl>, `Census Families With Children 1
## #   Child` <dbl>, `Census Families With Children 2 Children` <dbl>,
## #   `Census Families With Children 3+ Children` <dbl>, `Census Families
## #   With Children` <dbl>, `Income Under 10000` <dbl>, `Income
## #   10000-19999` <dbl>, `Income 20000-29999` <dbl>, `Income
## #   30000-39999` <dbl>, `Income 40000-49999` <dbl>, `Income
## #   50000-59999` <dbl>, `Income 60000-69999` <dbl>, `Income
## #   70000-79999` <dbl>, `Income 80000-89999` <dbl>, `Income
## #   90000-99999` <dbl>, `Income 100000+` <dbl>, `Income All` <dbl>,
## #   `Income Average` <dbl>, `Income Median` <dbl>, `Income Standard
## #   Error` <dbl>, `Income All Families` <dbl>, `Income All Families
## #   Average` <dbl>, `Income All Couples` <dbl>, `Income All Couples
## #   Average` <dbl>, `Income All Couples Median` <dbl>, `Income All Couples
## #   Standard Error` <dbl>, `Income Married Couples` <dbl>, `Income Married
## #   Couples Average` <dbl>, `Income Married Couples Median` <dbl>, `Income
## #   Married Couples Standard Error` <dbl>, `Income Common Law` <dbl>,
## #   `Income Common Law Average` <dbl>, `Income Common Law Median` <dbl>,
## #   `Income Common Law Standard Error` <dbl>, `Income One Parent
## #   Female` <dbl>, `Income One Parent Female Average` <dbl>, `Income One
## #   Parent Female Median` <dbl>, `Income One Parent Female Standard
## #   Error` <dbl>, `Income One Parent Male` <dbl>, `Income One Parent Male
## #   Average` <dbl>, `Income One Parent Male Median` <dbl>, `Income One
## #   Parent Male Standard Error` <dbl>, `Income All Families Median` <dbl>,
## #   `Income All Families Standard Error` <dbl>
```

Get a list of all areas that do not have census data for both 2001 and 2006

```
single_records <- Data %>% group_by(`Boundary Name`) %>% summarise(Count = n()) %>%
  filter(Count != 2) %>%
  select(`Boundary Name`) %>%
  as_vector()
```

Remove the entries that are indicated by the single_records variable.

```
Data <- Data %>% filter(`Boundary Name` %nin% single_records)
```

Save the cleaned data

```
write_csv(Data, here("Final Project", "data_output", "cleaned.csv"))
```