Unveiling Vulnerability: Data Cleaning For the Social Vulnerability Index, and the Enviornmental Justice Index

Thesis for a Master of Public Health, Epidemiology

Nathan Garcia-Diaz

Brown University, School of Public Health

Last Modified on August 1, 2024

Contents

Note: the table of contents acts as in-document hyperlinks

Person Statement

The code will import and clean data for the Center for Disease Control 2022 Version of the Social Vulnerability Index (Data & Documentation link), the Environmental Justice Index (Data link; Documentation link), and the Rhode Island Health Equity Index (link provided if and when available). The website inputs include Year = 2022 (when applicable), Geography = Rhode Island, and File Type = CSV File.

Please note that while data can be downloaded at state or national geographies, rankings within both geographic levels of these data represent comparisons to all other census tracts in the nation. For example, an EJI ranking of 0.85 signifies that 85% of tracts in the nation likely experience less severe cumulative impacts from environmental burden than the tract of interest, and that 15% of tracts in the nation likely experience more severe cumulative impacts from environmental burden.

Inputs: raw Social Vulnerability Index (SVI) and Environmental Justice Index (EJI) files downloaded from the provided links Outputs: two csv file that will be inputs for the `SVI_EDA.rmd` and `EJI_EDA.rmd` files, respectively.

SVI Suggested Citation: Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index [2022] Database [Rhode Island]. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html. Accessed on [07/2024].

EJI Suggested Citation: Centers for Disease Control and Prevention and Agency for Toxic Substances Disease Registry. 2022 Environmental Justice Index. Accessed [07/2024]. https://www.atsdr.cdc.gov/placeandhealth/eji/index.html

Preparation

The following two outputs were generated by the `str(df)` function, and they provide an initial peak into the raw data files. Note that the that “raw” files from the were changed given their initial similarity. SVI initially was called “RhodeIsland.csv”, but in the project directory it is called “SVI_RI_Data.csv’. Meanwhile, the EJI was called”Rhode Island.csv”, and it is now called “EJI_RI_Data.csv” The renaming of files was preformed in the File Explorer.

```r
# Importing Packages
library(tidyverse) # general data manipulation
library(knitr)     # Rmarkdown interactions
library(janitor)   # cleans the column names
library(here)      # define top level of project folder
                        # this allows for specification of where
                        # things live in relation to the top level
library(readr)     # imports csv files
library(naniar)    # missing data visualization
library(tigris)    # obtaining shp files
library(distill)   # markdown settings

# Importing Data
svi = read_csv(here::here("01_Data", "SVI_RI_Data.csv"))
eji = read_csv(here::here("01_Data", "EJI_RI_Data.csv"))
```

Data Cleaning

The following sections will be applied to the Social Vulnerability Index, the Environmental Justice Index, and the Rhode Island Health Equity Index:

- Trimming, Renaming, and Changing Data Types of Columns
- Determining Missing Data Entries and Peculiar Values
- Investigation of Missing Data

Social Vulnerability Index

Trimming, Renaming, and Changing Data Types of Columns

This section is completely about personal preference. I do not like to code with capital letters, therefore I like to use the the `janitor::clean_names(df)`, which will create columns names that are in snake case and contain letters in lower case. Lastly, `geoid` will be changed to a chr variable type to ensure ease in future data joins. The data initially has 247 observations, however there are 244 cesus tracts according to the United States Census Bureau

```
svi_cleaned = svi %>%
  clean_names() %>%
  select(
    # administrative data
    fips, location,
    # summary value
    rpl_themes,
    # predictor values
    starts_with("e_"),
    # removing undesirable columns
    -e_totpop, -e_hu, -e_hh, -(e_daypop:e_otherrace)) %>%
  mutate(fips = as.character(fips))
```

Determining Missing Data Entries and Peculiar Values

Census tracts with NA for could be missing values for multiple reasons. Please refer to the CDC/ASTR Technical Documentation for the 2022 Version of the Environmental Justice Index. Given that the missing data represents 0.4% of the entire data, permutations will not be implemented, and an examination of the missing data will be preformed in the following section. `eji_df` is defined in this section, and it serves as the working file for the following `.rmd` files.

```
## examination of variables
svi_cleaned %>% select(-fips, -location) %>% summary()
```

```
##    rpl_themes          e_pov150         e_unemp         e_hburd
##  Min.   :-999.000   Min.   :   0.0   Min.   :  0.0   Min.   :   0.0
##  1st Qu.:   0.247   1st Qu.: 344.5   1st Qu.: 70.5   1st Qu.: 317.5
##  Median :   0.498   Median : 587.0   Median :119.0   Median : 459.0
##  Mean   :  -3.547   Mean   : 760.1   Mean   :140.1   Mean   : 496.6
##  3rd Qu.:   0.749   3rd Qu.:1034.0   3rd Qu.:189.0   3rd Qu.: 638.0
##  Max.   :   1.000   Max.   :3345.0   Max.   :477.0   Max.   :1373.0
##     e_nohsdp        e_uninsur         e_age65        e_age17
##  Min.   :   0.0   Min.   :   0.0   Min.   :   0   Min.   :   0.0
##  1st Qu.: 124.5   1st Qu.:  77.0   1st Qu.: 477   1st Qu.: 525.5
##  Median : 257.0   Median : 137.0   Median : 768   Median : 815.0
##  Mean   : 331.7   Mean   : 188.0   Mean   : 791   Mean   : 843.4
##  3rd Qu.: 434.0   3rd Qu.: 225.5   3rd Qu.:1057   3rd Qu.:1090.5
##  Max.   :1837.0   Max.   :1017.0   Max.   :1983   Max.   :2672.0
##     e_disabl         e_sngpnt        e_limeng         e_minrty
##  Min.   :   0.0   Min.   :  0.0   Min.   :   0.0   Min.   :   0.0
##  1st Qu.: 417.0   1st Qu.: 42.5   1st Qu.:  16.0   1st Qu.: 462.5
##  Median : 543.0   Median : 80.0   Median :  65.0   Median : 881.0
##  Mean   : 587.2   Mean   :113.7   Mean   : 175.3   Mean   :1336.4
##  3rd Qu.: 710.0   3rd Qu.:152.5   3rd Qu.: 197.0   3rd Qu.:1695.0
##  Max.   :1596.0   Max.   :638.0   Max.   :1901.0   Max.   :6287.0
##     e_munit          e_mobile         e_crowd          e_noveh
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  43.5   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  44.5
##  Median : 163.0   Median :  0.00   Median : 15.00   Median :  97.0
##  Mean   : 243.4   Mean   : 16.98   Mean   : 34.99   Mean   : 158.2
##  3rd Qu.: 344.5   3rd Qu.:  0.00   3rd Qu.: 45.00   3rd Qu.: 224.0
##  Max.   :2563.0   Max.   :366.00   Max.   :300.00   Max.   :1053.0
##     e_groupq
##  Min.   :   0.0
##  1st Qu.:   8.0
##  Median :  32.0
##  Mean   : 182.2
##  3rd Qu.: 111.5
##  Max.   :5073.0
```

```
# conclusion: there seems to be a few particular entries

## examination of variables that do not contain NAs
svi_cleaned %>%
  select(-fips, -location) %>%
  filter(!(rpl_themes == -999)) %>%
  summary()
```

```
##    rpl_themes        e_pov150         e_unemp          e_hburd
##  Min.   :0.0000   Min.   :  84.0   Min.   :  0.00   Min.   :  28.0
##  1st Qu.:0.2500   1st Qu.: 345.2   1st Qu.: 71.25   1st Qu.: 318.0
##  Median :0.5000   Median : 591.5   Median :119.50   Median : 459.0
##  Mean   :0.4999   Mean   : 763.2   Mean   :140.72   Mean   : 498.6
##  3rd Qu.:0.7500   3rd Qu.:1042.5   3rd Qu.:189.00   3rd Qu.: 639.5
##  Max.   :1.0000   Max.   :3345.0   Max.   :477.00   Max.   :1373.0
##     e_nohsdp        e_uninsur         e_age65          e_age17
##  Min.   :   6.0   Min.   :   0.0   Min.   :  49.0   Min.   :  60.0
##  1st Qu.: 125.8   1st Qu.:  77.0   1st Qu.: 478.5   1st Qu.: 527.2
##  Median : 257.0   Median : 137.5   Median : 770.5   Median : 815.5
##  Mean   : 333.1   Mean   : 188.8   Mean   : 794.2   Mean   : 846.8
##  3rd Qu.: 434.0   3rd Qu.: 225.8   3rd Qu.:1062.0   3rd Qu.:1090.8
##  Max.   :1837.0   Max.   :1017.0   Max.   :1983.0   Max.   :2672.0
##     e_disabl         e_sngpnt        e_limeng          e_minrty
##  Min.   :  83.0   Min.   :  0.0   Min.   :   0.00   Min.   :  42.0
##  1st Qu.: 417.5   1st Qu.: 43.0   1st Qu.:  16.25   1st Qu.: 466.2
##  Median : 544.5   Median : 80.0   Median :  65.00   Median : 895.5
##  Mean   : 589.6   Mean   :114.2   Mean   : 176.03   Mean   :1341.8
##  3rd Qu.: 710.0   3rd Qu.:152.8   3rd Qu.: 197.50   3rd Qu.:1695.0
##  Max.   :1596.0   Max.   :638.0   Max.   :1901.00   Max.   :6287.0
##     e_munit          e_mobile         e_crowd          e_noveh
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  44.0   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  45.0
##  Median : 163.0   Median :  0.00   Median : 15.50   Median :  98.0
##  Mean   : 244.4   Mean   : 17.05   Mean   : 35.13   Mean   : 158.9
##  3rd Qu.: 345.2   3rd Qu.:  0.00   3rd Qu.: 45.00   3rd Qu.: 224.5
##  Max.   :2563.0   Max.   :366.00   Max.   :300.00   Max.   :1053.0
##     e_groupq
##  Min.   :   0.0
##  1st Qu.:   8.0
##  Median :  32.0
##  Mean   : 182.9
##  3rd Qu.: 112.2
##  Max.   :5073.0
```

```r
# conclusion: the removal of the NA values seems to have corrected the data
# continue by dropping these values

## redefining the eji_df
svi_df = svi_cleaned %>% filter(!(rpl_themes == -999))
```

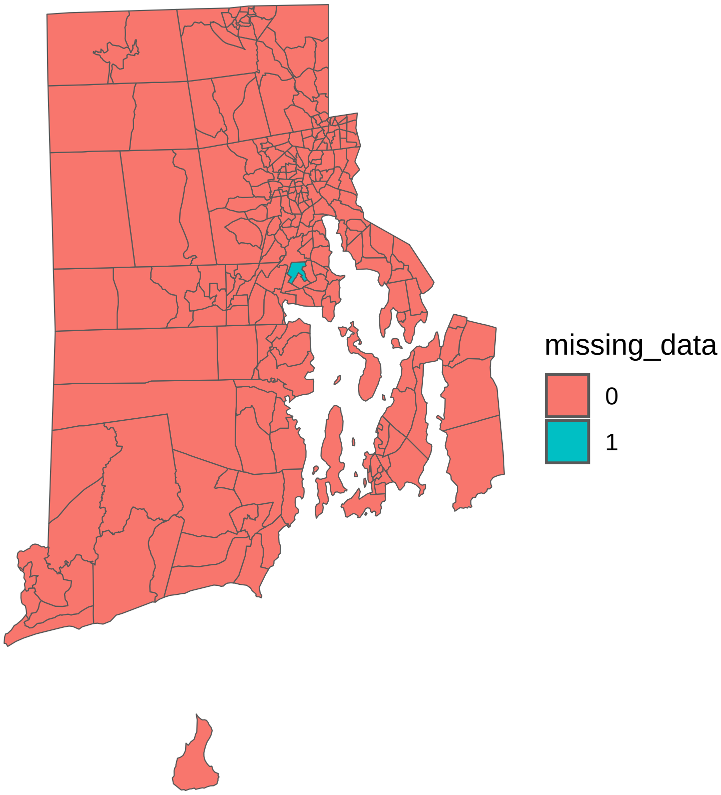
Investigation of Missing Data

A summary of the missing data is illustrated below. The graph illustrates combinations of missingness and intersections of missingess amongst variables. Despite 3 census tracts being listed as having NA columns, it seems that the airport is the only census tract to be missing from the map. DataCommon.org states that census tract 44005990000 contains no a small island off the coast of Gooseberry Beach, Newport. CensusReporter.org illustrates census tract 44009990100 is primarily coast line with no real population. All three census tracts contain no permanent population, and therefore some themes could not be calculated, which prevent a final summary index from being calculated.

```r
temp = svi_cleaned %>% filter(rpl_themes == -999)

# visualization of missing data
print(temp)
```

```
## # A tibble: 1 x 19
##   fips   location rpl_themes e_pov150 e_unemp e_hburd e_nohsdp e_uninsur e_age65
##   <chr> <chr>          <dbl>    <dbl>   <dbl>   <dbl>    <dbl>     <dbl>   <dbl>
## 1 44003~ Census ~       -999        0       0       0        0         0       0
## # i 10 more variables: e_age17 <dbl>, e_disabl <dbl>, e_sngpnt <dbl>,
## #   e_limeng <dbl>, e_minrty <dbl>, e_munit <dbl>, e_mobile <dbl>,
## #   e_crowd <dbl>, e_noveh <dbl>, e_groupq <dbl>
```

```
# filtering for missing census tracts
lst_missing = c("44003980000") # airport: population 0

# obtaining census tracts
ri_tracts = tracts(state = "RI", cb = TRUE)

# creating temporary data frame
temp_tracts = ri_tracts %>%
  mutate(missing_data = as.character(ifelse(GEOID %in% lst_missing, 1,0)))

# making the data
ggplot(temp_tracts, aes(fill = missing_data)) + geom_sf() + theme_void()
```



The following code exports the `eji_df` as a csv file. This file will be used in the subsequent `.rmd` files.

```
write.csv(svi_df,
          "/Users/diazg/Documents/GitHub/MPH-Thesis_GeographicalRandomForest/01_Data/svi_df.cs
```

Enviormental Justice Index

Trimming, Renaming, and Changing Data Types of Columns

This section is completely about personal preference. I do not like to code with capital letters, therefore I like to use the the `janitor::clean_names(df)`, which will create columns names that are in snake case and contain letters in lower case. Lastly, `geoid` will be changed to a chr variable type to ensure ease in future data joins. The data initially has 243 observations, however there are 244 census tracts according to the United States Census Bureau

```r
eji_cleaned = eji %>%
  clean_names() %>%
  select(
    # administrative Information
    geoid, location,
    # summary index value
    rpl_eji,
    # environmental burden
    starts_with("e_"),
    # social vulnerability
    starts_with("ep_"),
    # health vulnerability
    starts_with("f_'"),
    # undesirable columns
    -e_totpop, -e_daypop) %>%
  mutate(geoid = as.character(geoid))
```

Determining Missing Data Entries and Peculiar Values

Census tracts with NA for could be missing values for multiple reasons. Please refer to the CDC/ASTR Technical Documentation for the 2022 Version of the Environmental Justice Index. Given that the missing data represents 1.2% of the entire, permutations will not be implemented, and an examination of the missing data will be preformed in the following section. `eji_df` is defined in this section, and it serves as the working file for the following `.rmd` files.

```r
## examination of variables
eji_cleaned %>%
  select(-geoid, -location) %>%
  summary()
```

```
##     rpl_eji          e_ozone          e_pm           e_dslpm
##  Min.   :0.0407   Min.   :1.33   Min.   :5.750   Min.   :0.1700
##  1st Qu.:0.3331   1st Qu.:2.00   1st Qu.:6.430   1st Qu.:0.3300
##  Median :0.5233   Median :2.33   Median :6.990   Median :0.4400
##  Mean   :0.5326   Mean   :2.41   Mean   :6.817   Mean   :0.5107
##  3rd Qu.:0.7253   3rd Qu.:2.67   3rd Qu.:7.250   3rd Qu.:0.6400
##  Max.   :0.9949   Max.   :5.33   Max.   :7.380   Max.   :1.1600
##  NA's   :3        NA's   :2      NA's   :2       NA's   :2
##     e_totcr          e_npl            e_tri            e_tsd
##  Min.   :16.73   Min.   : 0.000   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:21.93   1st Qu.: 0.000   1st Qu.: 39.47   1st Qu.:  0.000
##  Median :25.33   Median : 0.000   Median : 92.12   Median :  0.000
##  Mean   :24.81   Mean   : 3.138   Mean   : 69.72   Mean   :  4.501
##  3rd Qu.:27.45   3rd Qu.: 0.000   3rd Qu.:100.00   3rd Qu.:  0.000
##  Max.   :34.51   Max.   :92.620   Max.   :100.00   Max.   :100.000
##  NA's   :2       NA's   :2        NA's   :2        NA's   :2
##      e_rmp            e_coal      e_lead      e_park          e_houage
##  Min.   :  0.00   Min.   :0   Min.   :0   Min.   :  0.00   Min.   :22.47
##  1st Qu.:  0.00   1st Qu.:0   1st Qu.:0   1st Qu.: 49.54   1st Qu.:61.92
##  Median :  0.00   Median :0   Median :0   Median : 98.25   Median :80.15
##  Mean   : 19.92   Mean   :0   Mean   :0   Mean   : 72.68   Mean   :74.93
##  3rd Qu.: 29.13   3rd Qu.:0   3rd Qu.:0   3rd Qu.:100.00   3rd Qu.:87.84
##  Max.   :100.00   Max.   :0   Max.   :0   Max.   :100.00   Max.   :98.67
##                                            NA's   :2        NA's   :3
##     e_wlkind          e_rail           e_road          e_airprt
##  Min.   : 1.000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.: 9.405   1st Qu.:  0.00   1st Qu.: 37.33   1st Qu.:  0.000
##  Median :13.670   Median : 44.22   Median : 90.45   Median :  0.000
##  Mean   :12.363   Mean   : 47.72   Mean   : 68.61   Mean   :  5.393
##  3rd Qu.:15.275   3rd Qu.:100.00   3rd Qu.:100.00   3rd Qu.:  0.000
##  Max.   :18.780   Max.   :100.00   Max.   :100.00   Max.   :100.000
##                   NA's   :2        NA's   :2        NA's   :2
##     e_impwtr        ep_minrty       ep_pov200       ep_nohsdp
##  Min.   :  1.53   Min.   : 0.00   Min.   : 0.00   Min.   : 0.60
##  1st Qu.: 58.28   1st Qu.: 8.55   1st Qu.:14.00   1st Qu.: 5.40
##  Median : 80.66   Median :14.90   Median :22.68   Median : 9.00
##  Mean   : 71.59   Mean   :27.62   Mean   :27.37   Mean   :11.55
##  3rd Qu.: 87.02   3rd Qu.:39.40   3rd Qu.:37.15   3rd Qu.:15.03
##  Max.   :100.00   Max.   :98.70   Max.   :76.38   Max.   :46.20
##  NA's   :2                                         NA's   :3
##     ep_unemp        ep_renter        ep_houbdn        ep_uninsur
##  Min.   : 0.000   Min.   :  2.50   Min.   : 5.371   Min.   : 0.00
##  1st Qu.: 2.900   1st Qu.: 19.73   1st Qu.:23.702   1st Qu.: 2.10
##  Median : 4.650   Median : 35.30   Median :30.352   Median : 3.70
##  Mean   : 5.603   Mean   : 40.66   Mean   :31.313   Mean   : 4.71
##  3rd Qu.: 7.200   3rd Qu.: 60.23   3rd Qu.:38.186   3rd Qu.: 6.40
##  Max.   :25.400   Max.   :100.00   Max.   :57.202   Max.   :18.90
##  NA's   :3        NA's   :3        NA's   :3        NA's   :3
##     ep_noint        ep_age65        ep_age17       ep_disabl
##  Min.   : 2.50   Min.   : 0.70   Min.   : 0.00   Min.   :-666666666
##  1st Qu.:10.00   1st Qu.:12.28   1st Qu.:16.15   1st Qu.:        10
##  Median :13.85   Median :16.85   Median :19.00   Median :        13
##  Mean   :15.70   Mean   :17.10   Mean   :19.02   Mean   :  -8230439
##  3rd Qu.:20.07   3rd Qu.:21.02   3rd Qu.:22.15   3rd Qu.:        16
##  Max.   :51.10   Max.   :38.90   Max.   :35.10   Max.   :        40
##  NA's   :3       NA's   :3
##    ep_limeng        ep_mobile         ep_groupq        ep_bphigh
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000   Min.   :11.20
##  1st Qu.: 0.500   1st Qu.: 0.0000   1st Qu.: 0.100   1st Qu.:30.80
##  Median : 1.400   Median : 0.0000   Median : 0.400   Median :32.60
##  Mean   : 4.007   Mean   : 0.8421   Mean   : 3.718   Mean   :32.13
##  3rd Qu.: 4.950   3rd Qu.: 0.5000   3rd Qu.: 2.150   3rd Qu.:34.42
##  Max.   :26.500   Max.   :21.0000   Max.   :93.700   Max.   :48.30
##                   NA's   :3                          NA's   :3
##    ep_asthma        ep_cancer         ep_mhlth      ep_diabetes
##  Min.   : 9.20   Min.   : 1.100   Min.   : 7.60   Min.   : 2.20
##  1st Qu.:10.90   1st Qu.: 6.175   1st Qu.:11.97   1st Qu.: 9.10
##  Median :11.50   Median : 7.400   Median :13.45   Median : 9.90
##  Mean   :11.69   Mean   : 7.132   Mean   :13.91   Mean   :10.21
##  3rd Qu.:12.30   3rd Qu.: 8.200   3rd Qu.:15.50   3rd Qu.:11.12
##  Max.   :15.60   Max.   :12.600   Max.   :24.90   Max.   :24.20
##  NA's   :3       NA's   :3        NA's   :3       NA's   :3
```

```
# conclusion: there seems to be a few particular entries

## examination of variables that do not contain NAs
eji_cleaned %>%
  select(-geoid, -location) %>%
  filter(!is.na(rpl_eji)) %>%
  summary()
```

```
##     rpl_eji          e_ozone          e_pm          e_dslpm
##  Min.   :0.0407   Min.   :1.33   Min.   :5.750   Min.   :0.1700
##  1st Qu.:0.3331   1st Qu.:2.00   1st Qu.:6.428   1st Qu.:0.3300
##  Median :0.5233   Median :2.33   Median :6.990   Median :0.4400
##  Mean   :0.5326   Mean   :2.41   Mean   :6.817   Mean   :0.5105
##  3rd Qu.:0.7253   3rd Qu.:2.67   3rd Qu.:7.253   3rd Qu.:0.6425
##  Max.   :0.9949   Max.   :5.33   Max.   :7.380   Max.   :1.1600
##     e_totcr          e_npl            e_tri            e_tsd
##  Min.   :16.73   Min.   : 0.000   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:21.91   1st Qu.: 0.000   1st Qu.: 38.99   1st Qu.:  0.00
##  Median :25.34   Median : 0.000   Median : 91.98   Median :  0.00
##  Mean   :24.82   Mean   : 3.151   Mean   : 69.59   Mean   :  4.52
##  3rd Qu.:27.45   3rd Qu.: 0.000   3rd Qu.:100.00   3rd Qu.:  0.00
##  Max.   :34.51   Max.   :92.620   Max.   :100.00   Max.   :100.00
##      e_rmp            e_coal      e_lead      e_park          e_houage
##  Min.   :  0.00   Min.   :0   Min.   :0   Min.   :  0.00   Min.   :22.47
##  1st Qu.:  0.00   1st Qu.:0   1st Qu.:0   1st Qu.: 50.14   1st Qu.:61.92
##  Median :  0.00   Median :0   Median :0   Median : 98.42   Median :80.15
##  Mean   : 20.15   Mean   :0   Mean   :0   Mean   : 72.98   Mean   :74.93
##  3rd Qu.: 29.45   3rd Qu.:0   3rd Qu.:0   3rd Qu.:100.00   3rd Qu.:87.84
##  Max.   :100.00   Max.   :0   Max.   :0   Max.   :100.00   Max.   :98.67
##     e_wlkind          e_rail           e_road           e_airprt
##  Min.   : 3.280   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.: 9.435   1st Qu.:  0.00   1st Qu.: 37.16   1st Qu.:  0.000
##  Median :13.715   Median : 43.96   Median : 90.68   Median :  0.000
##  Mean   :12.453   Mean   : 47.61   Mean   : 68.53   Mean   :  4.999
##  3rd Qu.:15.342   3rd Qu.:100.00   3rd Qu.:100.00   3rd Qu.:  0.000
##  Max.   :18.780   Max.   :100.00   Max.   :100.00   Max.   :100.000
##     e_impwtr         ep_minrty         ep_pov200         ep_nohsdp
##  Min.   :  1.53   Min.   : 1.200   Min.   : 4.319   Min.   : 0.60
##  1st Qu.: 58.26   1st Qu.: 9.075   1st Qu.:14.293   1st Qu.: 5.40
##  Median : 80.72   Median :15.200   Median :22.837   Median : 9.00
##  Mean   : 71.56   Mean   :27.966   Mean   :27.716   Mean   :11.55
##  3rd Qu.: 87.05   3rd Qu.:39.650   3rd Qu.:37.531   3rd Qu.:15.03
##  Max.   :100.00   Max.   :98.700   Max.   :76.381   Max.   :46.20
##     ep_unemp         ep_renter         ep_houbdn         ep_uninsur
##  Min.   : 0.000   Min.   :  2.50   Min.   : 5.371   Min.   : 0.00
##  1st Qu.: 2.900   1st Qu.: 19.73   1st Qu.:23.702   1st Qu.: 2.10
##  Median : 4.650   Median : 35.30   Median :30.352   Median : 3.70
##  Mean   : 5.603   Mean   : 40.66   Mean   :31.313   Mean   : 4.71
##  3rd Qu.: 7.200   3rd Qu.: 60.23   3rd Qu.:38.186   3rd Qu.: 6.40
##  Max.   :25.400   Max.   :100.00   Max.   :57.202   Max.   :18.90
##     ep_noint         ep_age65         ep_age17        ep_disabl
##  Min.   : 2.50   Min.   : 0.70   Min.   : 1.90   Min.   : 4.50
##  1st Qu.:10.00   1st Qu.:12.28   1st Qu.:16.30   1st Qu.:10.55
##  Median :13.85   Median :16.85   Median :19.15   Median :13.00
##  Mean   :15.70   Mean   :17.10   Mean   :19.26   Mean   :13.68
##  3rd Qu.:20.07   3rd Qu.:21.02   3rd Qu.:22.20   3rd Qu.:15.82
##  Max.   :51.10   Max.   :38.90   Max.   :35.10   Max.   :39.50
##    ep_limeng         ep_mobile          ep_groupq         ep_bphigh
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000   Min.   :11.20
##  1st Qu.: 0.575   1st Qu.: 0.0000   1st Qu.: 0.200   1st Qu.:30.80
##  Median : 1.500   Median : 0.0000   Median : 0.450   Median :32.60
##  Mean   : 4.057   Mean   : 0.8421   Mean   : 3.764   Mean   :32.13
##  3rd Qu.: 5.000   3rd Qu.: 0.5000   3rd Qu.: 2.200   3rd Qu.:34.42
##  Max.   :26.500   Max.   :21.0000   Max.   :93.700   Max.   :48.30
##    ep_asthma        ep_cancer          ep_mhlth       ep_diabetes
##  Min.   : 9.20   Min.   : 1.100   Min.   : 7.60   Min.   : 2.20
##  1st Qu.:10.90   1st Qu.: 6.175   1st Qu.:11.97   1st Qu.: 9.10
##  Median :11.50   Median : 7.400   Median :13.45   Median : 9.90
##  Mean   :11.69   Mean   : 7.132   Mean   :13.91   Mean   :10.21
##  3rd Qu.:12.30   3rd Qu.: 8.200   3rd Qu.:15.50   3rd Qu.:11.12
##  Max.   :15.60   Max.   :12.600   Max.   :24.90   Max.   :24.20
```

```r
# conclusion: the removal of the NA values seems to have corrected the data
# continue by dropping these values

## redefining the eji_df
eji_df = eji_cleaned %>% filter(!is.na(rpl_eji))
```

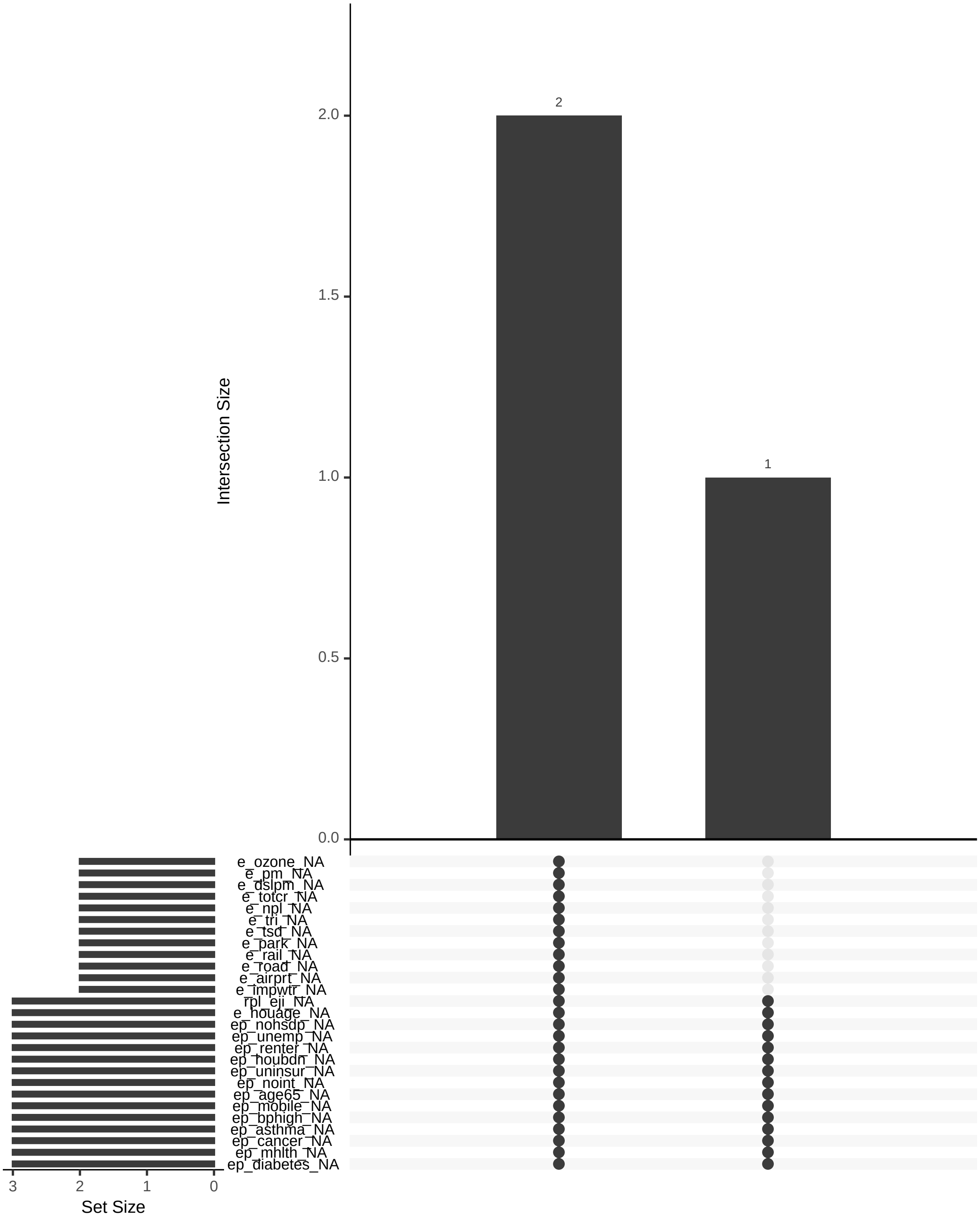
Investigation of Missing Data

A summary of the missing data is illustrated below. The graph illustrates combinations of missingness and intersections of missingess amongst variables. Despite 3 census tracts being listed as having NA columns, it seems that the airport is the only census tract to be missing from the map. DataCommon.org states that census tract 44005990000 contains no a small island off the coast of Gooseberry Beach, Newport. CensusReporter.org illustrates census tract 44009990100 is primarily coast line with no real population. All three census tracts contain no permanent population, and therefore some themes could not be calculated, which prevent a final summary index from being calculated.

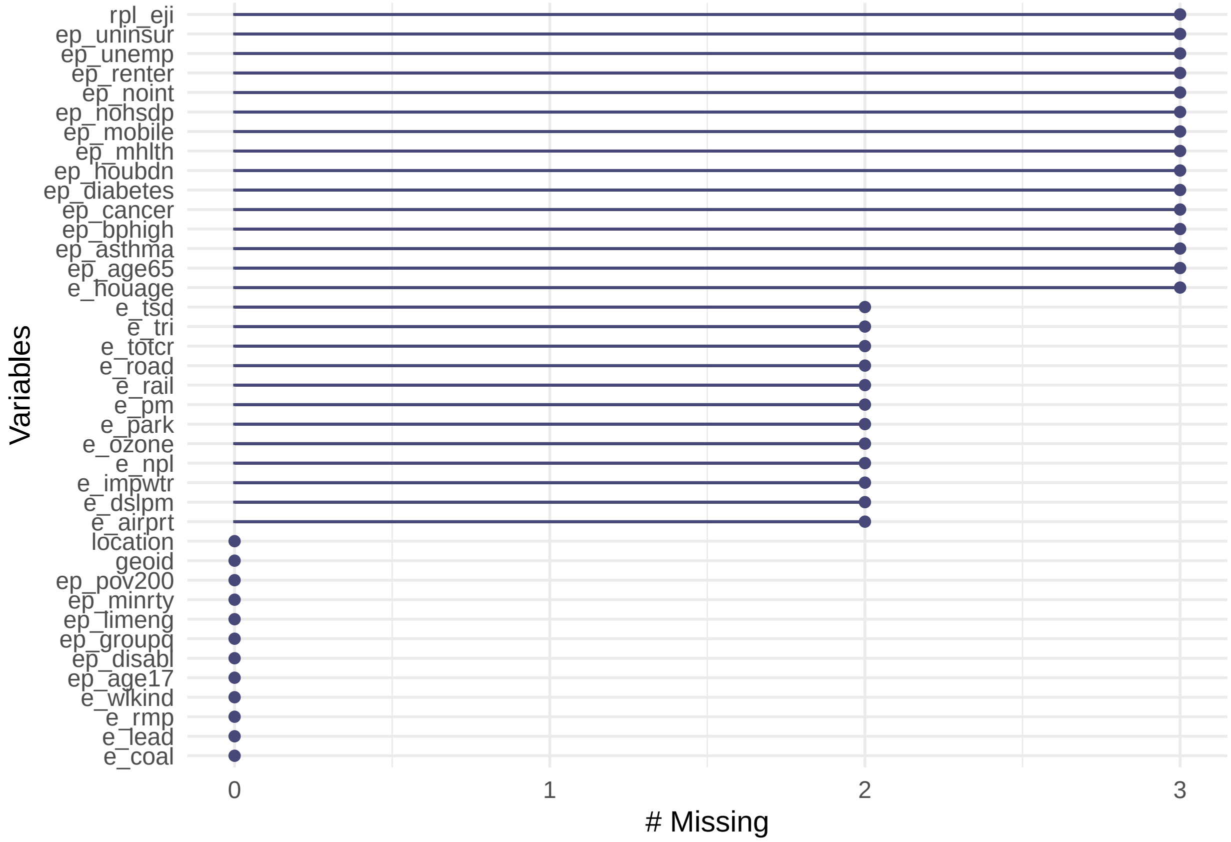
```
temp = eji_cleaned %>% filter(is.na(rpl_eji))

# visualization of missing data
print(temp)
```

```
## # A tibble: 3 x 39
##   geoid   location rpl_eji e_ozone  e_pm e_dslpm e_totcr e_npl e_tri e_tsd e_rmp
##   <chr>   <chr>      <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 440039~ Census ~      NA    2.33  6.79    0.56    22.1     0   100     0  4.66
## 2 440059~ Census ~      NA   NA    NA      NA       NA      NA    NA    NA  0
## 3 440099~ Census ~      NA   NA    NA      NA       NA      NA    NA    NA  0
## # i 28 more variables: e_coal <dbl>, e_lead <dbl>, e_park <dbl>,
## #   e_houage <dbl>, e_wlkind <dbl>, e_rail <dbl>, e_road <dbl>, e_airprt <dbl>,
## #   e_impwtr <dbl>, ep_minrty <dbl>, ep_pov200 <dbl>, ep_nohsdp <dbl>,
## #   ep_unemp <dbl>, ep_renter <dbl>, ep_houbdn <dbl>, ep_uninsur <dbl>,
## #   ep_noint <dbl>, ep_age65 <dbl>, ep_age17 <dbl>, ep_disabl <dbl>,
## #   ep_limeng <dbl>, ep_mobile <dbl>, ep_groupq <dbl>, ep_bphigh <dbl>,
## #   ep_asthma <dbl>, ep_cancer <dbl>, ep_mhlth <dbl>, ep_diabetes <dbl>
```

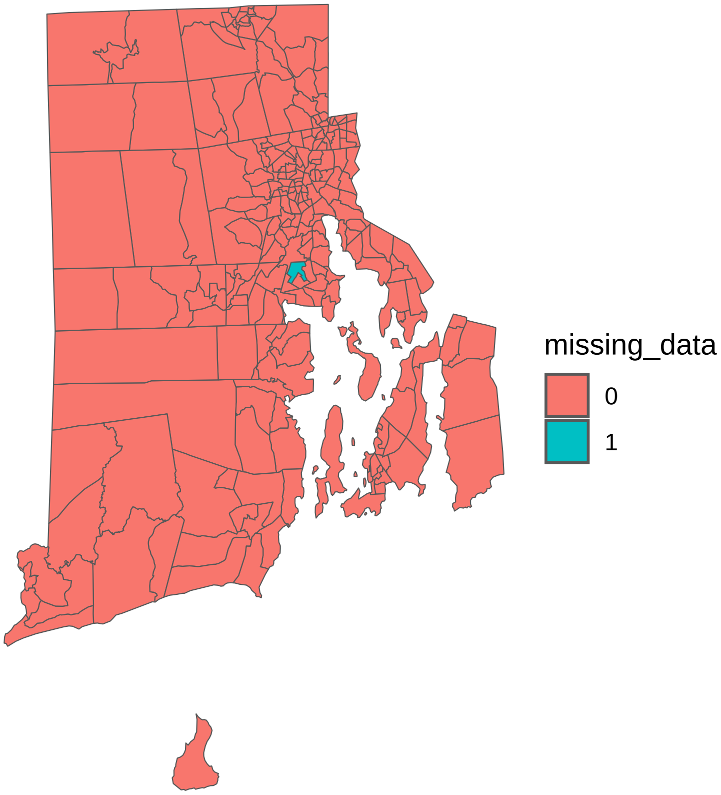
```
gg_miss_upset(temp, nsets = n_var_miss(temp))
```



```
gg_miss_var(temp)
```



```
# filtering for missing census tracts
lst_missing = c("44003980000", # airport: population 0
                "44005990000", # new port county: population 0
                "44009990100") # washington county Rhode island: population 0

# creating temprary data frame
temp_tracts = ri_tracts %>%
  mutate(missing_data = as.character(ifelse(GEOID %in% lst_missing, 1,0)))

# making the
ggplot(temp_tracts, aes(fill = missing_data)) + geom_sf() + theme_void()
```



The following code exports the `eji_df` as a csv file. This file will be used in the subsequent `.rmd` files.

```
write.csv(eji_df,
          "/Users/diazg/Documents/GitHub/MPH-Thesis_GeographicalRandomForest/01_Data/eji_df.cs
```