Preliminary Analysis

Jon Geiger

2022-04-01

Importing Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
```

Importing Data

```
url <- "https://urban-data-catalog.s3.amazonaws.com/drupal-root-live/2020/06/08/NHGIS_District_data.xlsx
destfile <- "NHGIS_District_data.xlsx"
curl::curl_download(url, destfile)
district_data <- read_excel(destfile)
```

Data Cleaning

Fortunately, there isn't much initial restructuring required of this data.

```
names(district_data)
```

```
##  [1] "School ID"
##  [2] "State"
##  [3] "Geographic School District"
```

```
##  [4] "Children 5-17 (SAIPE Estimate)"
##  [5] "% Poverty (SAIPE Estimate)"
##  [6] "% Single Parent Estimate"
##  [7] "Single Parent Margin of Error"
##  [8] "% HHs With Vulnerable Job Estimate"
##  [9] "Vulnerable Job Margin of Error"
## [10] "% Crowded Conditions Estimate"
## [11] "HH With Crowded Conditions Margin of Error"
## [12] "% No Computer or Internet Estimate"
## [13] "No Computer or Internet Margin of Error"
## [14] "% Children with Disability"
## [15] "Children with Disability Margin of Error"
## [16] "% Linguistically Isolated Children"
## [17] "Linguistically Isolated Children Margin of Error"
```

Let's work on renaming our variables to more R-friendly names.

```
names_list <- c(
    "school_ID",
    "state",
    "school_district",
    "children_5_to_17",
    "pct_poverty",
    "pct_singleParent",
    "singleParent_MOE",
    "pct_HHvulnerableJob",
    "HHvulnerableJob_MOE",
    "pct_crowdedConditions",
    "crowdedConditions_MOE",
    "pct_noComputerInternet",
    "NoComputerInternet_MOE",
    "pct_childDisability",
    "childDisability_MOE",
    "pct_childLinguisticallyIsolated",
    "childLinguisticallyIsolated_MOE"
)
names(district_data) <- names_list
district_pcts <- district_data %>%
    select(-ends_with("MOE"))
```

We make a new dataset, `dsistrict_pcts`, which contains all the values without any of the margins of error.

We can now do some preliminary analysis on this data.

```
View(district_pcts)
```