How to Use This Package

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'igraph'

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

This package has 4 main parts:

1. Scraping
2. Harmonizing
3. Linking
4. Merging

Each of these packages makes use of the package `here`. `here` uses a project's directory to navigate to other files. Therefore, in order to run these scripts, you need to open project "ffsgData" in RStudio. Opening the project "ffsg" will *not* work.

Each of these have their own directory which hold both the source code and the datasets which result from running that source code. In general, you will want to run these scripts in order. When calling Harmonizing or Linking, the code will automatically run the prior scripts.

The linking process generally takes around 45 minutes on my 2 GHz, 8 GB RAM computer. Because of this long run time, the Merging script does *not* automatically run the prior scripts.

Scraping

Scraping is composed of five scraping scripts and one master script. The master script sources the content of all five scraping scripts. Four of the scraping scripts scrape the police shooting datasets, and the fifth scraping script scrapes the US census. Upon running, each scripts saves its dataset in `R/Scraping/ScrapedFiles`. In the case of the census script, it saves both county and state datasets in `R/Scraping/ScrapedFiles/Populations`. To run all 5 scripts, simply run `MasterScraper.R`

Harmonizing

Harmonizing is organized slightly differently than Scraping. Harmonizer only has one script, Harmonizer.R, which Harmonizes all 4 datasets (it does nothing to the census datasets). Upon completion, it saves all 4 datasets in a single file `HarmonizedDataSets.RData` in the folder `R/Harmonizing/HarmonizedFiles`. If a new dataset is ever added, it should be relatively straight forward to follow the example of existing datasets in `Harmonizer.R`.

Linking

Linkage consists of two scripts: `ClericalReview.R` and `Linker.R`.

- `ClericalReview.R` should be fun first if you are just getting started – it is a script which lets a user estimate a good threshold for a weight cuttoff. You can run the script, follow the instructions, and get an estimate of how to set the cutoff. Through my trials, I've found 6 to be reasonable.
- `Linker.R` is the script that runs the Fellegi and Sunter algorithm to actually find links between the records. It produced two files: `full_classification.RData` and `full_combined_harmonized.RData`. `full_classification` contains the output of running record linkage, which you can read about on page 11 here.

Output

`full_classification.RData` is an RData object with

1. a list object named *classification* that contains 10 elements output from the linkage algorithm:

- data, a data frame with 16 columns (date, name, aka, age, race, state, year, month, day, firstname, lastname, middlename, str_age, source, uid)
- pairs, a data frame with each possible pair assessed for linkage. it consists of 13 col (id1, id2, age, sex, race, state, year, month, day, firstname, middlename, is_math)
- frequencies, a vector consisting of the frequencies of(age, sex, race, state, year, month, day, firstname, lastname, middlename)
- type, deduplication
- M and U are estimated m- and uprobabilities for the present comparison pattern. Estimation of M and U is done by an EM algorithm, implemented by mygllm. For every comparison pattern, the estimated numbers of matches and non-matches are used to compute the corresponding probabilities.
- W,
- Wdata,
- prediction a vector consisting of 3 levels: N, P, L
- threshold: the value to set the cutoff, in this code 6.

2. `full_combined_harmonized` is a dataframe in which each of the harmonized datasets is stacked row-wise. The columns each dataset has in common are not duplicated; all other columns represent variables that are in a subset of the datasets (where subset size can be 1). For the columns that are present in some datasets but not others `NA` is filled in for the rows of the datasets that are missing that column.

Merging

Merging consists of one file: `Merging.R`. It takes the contents of `full_classification`, which is a set of links between rows in `full_combined_harmonized`, and turns those links into a graph. It then uses that graph to find which sets of records are all linked together, indicating we think they are the same person. It then collapses `full_combined_harmonized` by the sets of linked records by naively choosing the first non-null value it finds to be the representive for that person. It also adds four columsn which indicate which datasets that person was originally found in. It outputs this new, collapsed file in `R/Merging/Merged`.

Output

a data frame named **final_merged** consisting of 63 columns: (person, date, name, aka, age, sex, race, state, X., Source, year, month, day, firstname, lastname, middlename, str_age, source, URLpic, address, city, zip, county, fullAddress, latitude, longitude, agency, causeOfDeath, circumstances. . .)

Common columns in 4 datasets : age, aka, date, day, firstname, lastname, middlename, month, name, race, sex, source, state, str_age, year.

Shared columns:

- Fatal encounter & Washington Post : "city"
- Fatal encounter & Mapping Police Violence: "zip"

Unique in each dataset :

- Fatal encounters: address, agency, causeOfDeath, circumstances, city, county, Description, fullAddress, kbp.filter, latitude, longitude, officialDisposition, URLarticle, URLpic, zip
- Killed by Police: Source, X.
- Washington Post: armed, body_camera, city, flee, id, manner_of_death, signs_of_mental_illness, threat_level
- Mapping Police Violence: ..25, A brief description of the circumstances surrounding the death, Agency responsible for death, Alleged Threat Level (Source: WaPo), Alleged Weapon (Source: WaPo), Body Camera (Source: WaPo), Cause of death, City, County, Criminal Charges?, Fleeing (Source: WaPo), Link to news article or photo of official document, Official disposition of death (justified or other), Street Address of Incident, Symptoms of mental illness?, Unarmed, URL of image of victim, WaPo ID (If included in WaPo database), zip

Existing Work

http://v-neck.github.io http://v-neck.github.io/final/ https://docs.google.com/presentation/d/19XFzorND9YIxosJm74sgS-V
qJSvCoGWwsGXubt-N0/edit?usp=sharing

Old Tabulations

https://github.com/statnet/ffsg/tree/master/Data/ffsgData/vignettes/Autumn%202018%20Reports