BUNMD Supplementary Geography Variables File Vignette

2023-03-09

Below are instructions on how to merge these files and what to look out for. First, we will read in both the Birth_Death_Features and BUNDMD files with data table. For this vignette, I'll select only key variables of interest from the BUNDMD in order to make it more manageable.

```
bunmd_geography_file <- fread("birth_death_features.csv")
bunmd <- fread("bunmd_v2.csv", select=c("ssn","byear","dyear","death_age"))
```

BUNDMD has the best death coverage for people born between 1910 and 1920 who died between the years 1988 and 2005. Here, we will subset the BUNMD dataset to exclude anyone who does not meet this criteria.

```
bunmd <- filter(bunmd, byear %in% 1910:1920 & dyear %in% 1988:2005)
```

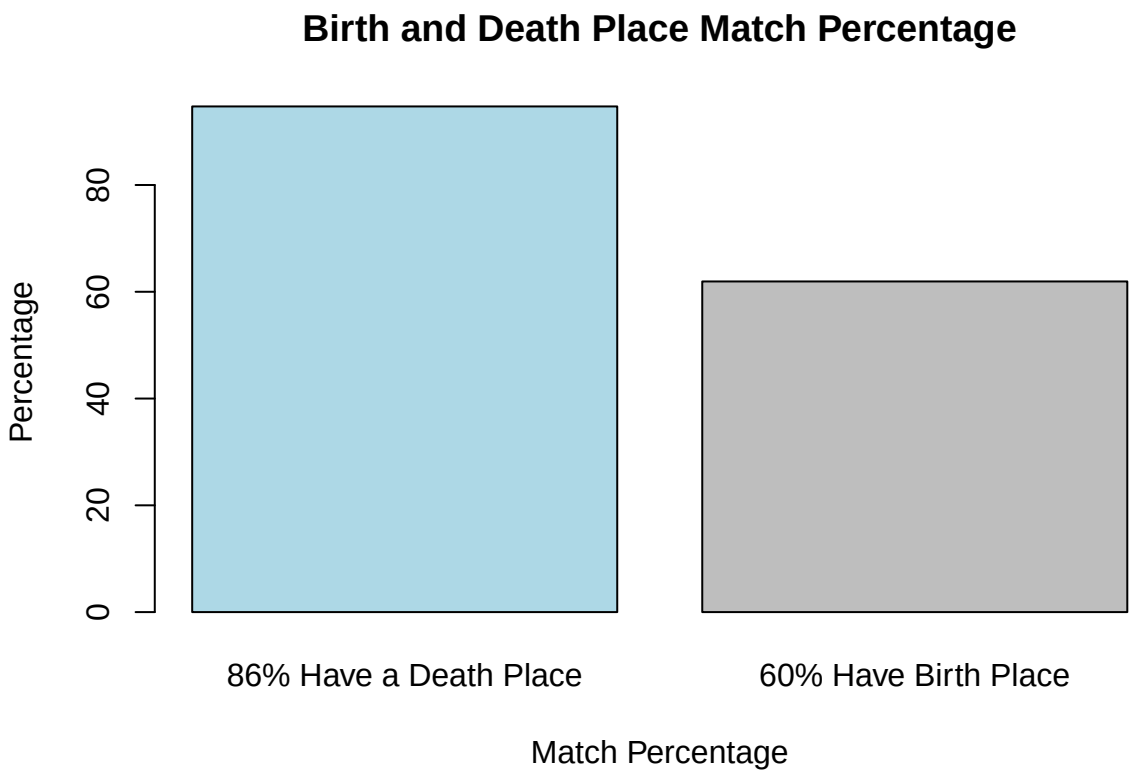
Now that we have read in both files, we can start merging them. Since we will be merging based on the Social Security Number, which is unique to each individual, we can use a "left join" where the BUNDMD file will be the left file and the Birth_Death_Features file will be joined to it.

```
bunmd_merged <- merge(bunmd, bunmd_geography_file, by = "ssn", all.x = TRUE)
rm(bunmd)
rm(bunmd_geography_file)
```

Now that we've merged, we can see how what percentage of people have a birth city or a valid death ZIP.

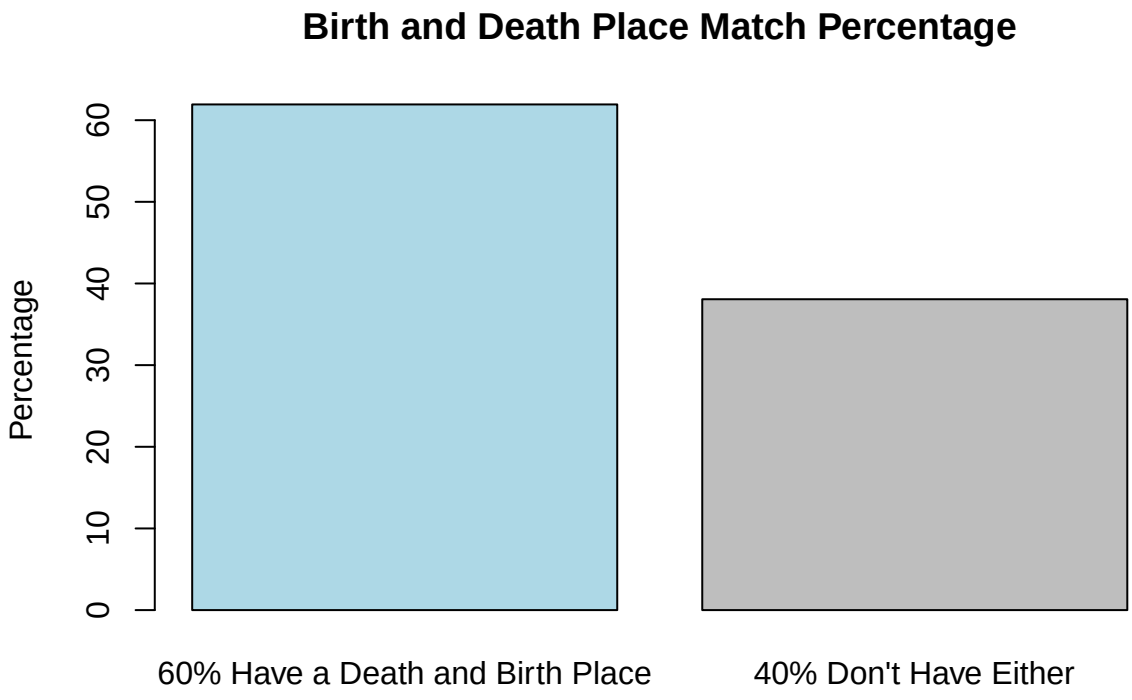
About 60% of people in the BUNMD data set are matched with a birth city. Much more coverage is available for death place. Around 86% of people in the BUNMD have a death place.

```
non_missing_birth_percent <- 100 - (mean(is.na(bunmd_merged$birth_city)) * 100)
non_missing_death_percent <- 100 - ((mean(is.na(bunmd_merged$death_zip)) * 100))
non_missing_values <- c(non_missing_death_percent, non_missing_birth_percent)
rm(non_missing_birth_percent,non_missing_death_percent)

barplot(non_missing_values, names.arg = c("86% Have a Death Place", "60% Have Birth Place"),
        xlab = "Match Percentage", ylab = "Percentage",
        main = "Birth and Death Place Match Percentage",
        col = c("lightblue", "grey"))
```



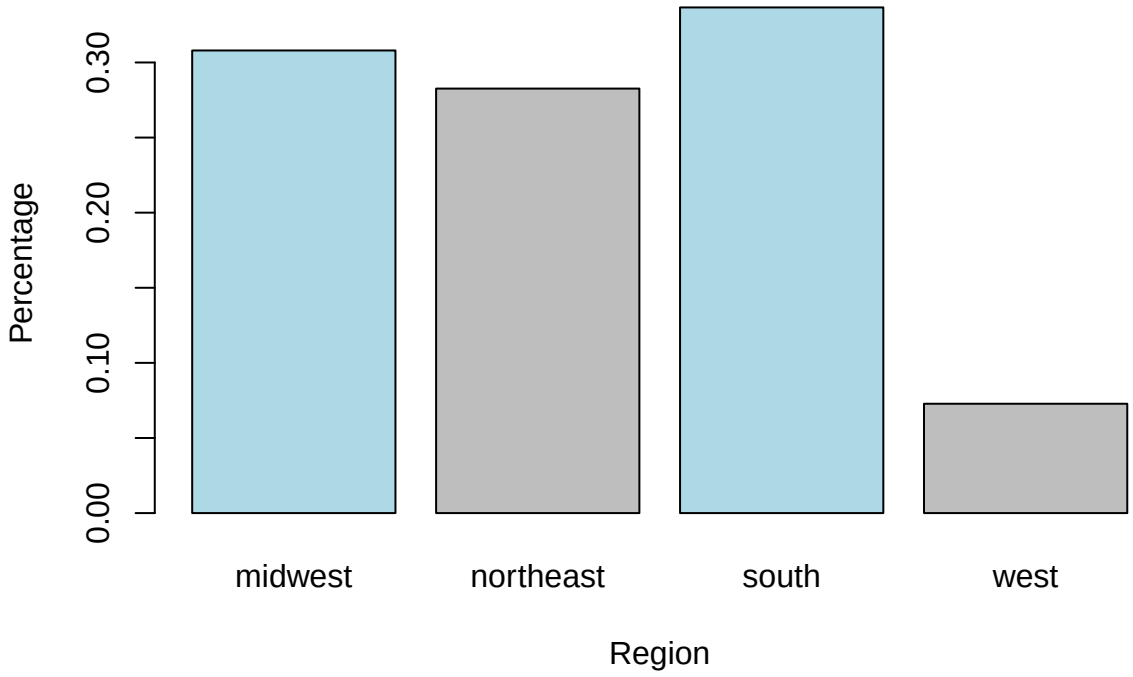
We can also see what percentage of people have both

```
bunmd_merged$missing_both <- ifelse(!is.na(bunmd_merged$death_zip) & !is.na(bunmd_merged$birth_city), 1
missing_both_percent <- mean(is.na(bunmd_merged$missing_both)) * 100
non_missing_both_percent <- 100 - missing_both_percent
both_values <- c(non_missing_both_percent, missing_both_percent)

rm(missing_both_percent)
rm(non_missing_both_percent)

barplot(both_values, names.arg = c("60% Have a Death and Birth Place", "40% Don't Have Either"),
        xlab = "Missing vs Not Missing", ylab = "Percentage",
        main = "Birth and Death Place Match Percentage",
        col = c("lightblue", "grey"))
```



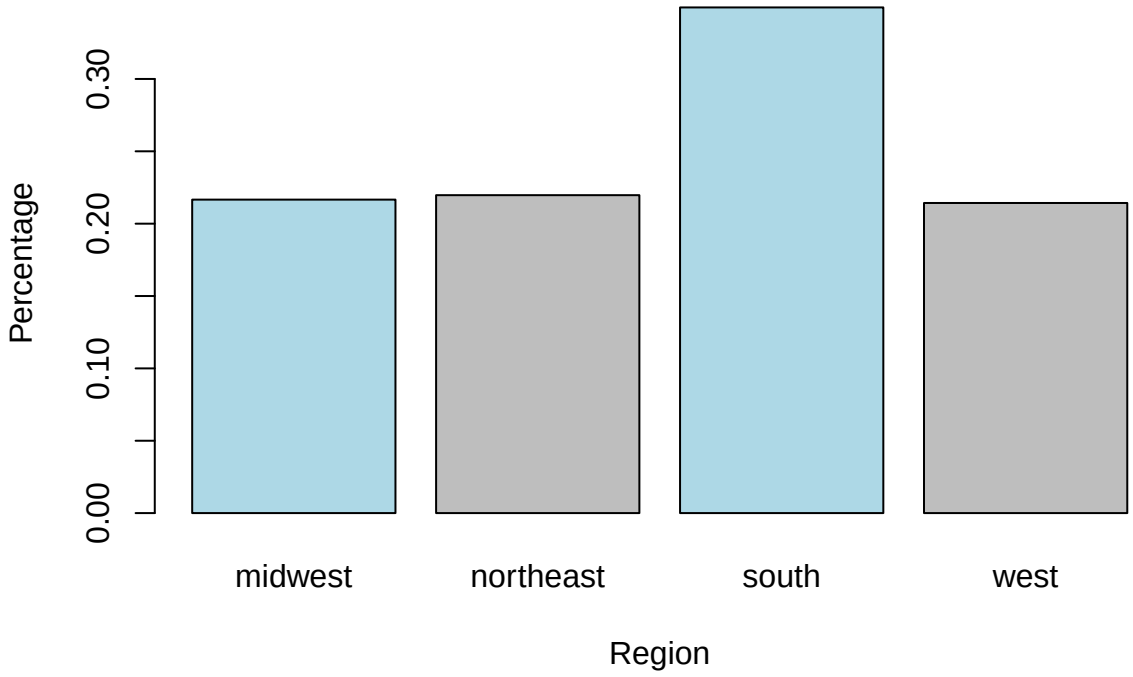
Next, let's take a look at which census regions people are born in

```
prop_region_birth <- table(bunmd_merged$birth_region)

barplot(prop_region_birth/sum(prop_region_birth), names.arg = c("midwest", "northeast","south","west"),
        xlab = "Region", ylab = "Percentage",
        main = "Percentage of matched individuals by birth region",
        col = c("lightblue", "grey"))
```



And where they die

```
prop_region_death <- table(bunmd_merged$death_region)

barplot(prop_region_death/sum(prop_region_death), names.arg = c("midwest", "northeast","south","west"),
        xlab = "Region", ylab = "Percentage",
        main = "Percentage of matched individuals by death region",
        col = c("lightblue", "grey"))
```


Percentage of matched individuals by death region
Percentage
0.00
0.10
0.20
0.30
midwest
northeast
south
west
Region