Data Cleaning 2-11-2025

Jorge R. Soldevila Irizarry

2025-02-18

This document details data cleaning processes for Family Composition Variables

First we will read in our libraries and packages.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stringr)
library(fedmatch)
```

We can now read in our data. It contains at the moment of typing this data_v9 and data_v10.

```
load("data.Rdata")
```

Place of birth

```
#First we will select the variables that correspond to Place of Birth. This
#includes the place of birth for all members of the family.
plc_birth <- data_v9 %>% #We use data_v9 becuase we want to make changes on all.
  select(ID_long,p01_plc,p02_plc,p03_plc,p04_plc,p05_plc,
         p06_plc,p07_plc,p08_plc,p09_plc,p010_plc,p011_plc,p012_plc)

#We can now convert the data set into a long format that allows us to look at
#the unique responses to our variable
plc_birth_long <- pivot_longer(plc_birth,
                           cols = starts_with("p0"),
```

```
                          names_to = "var",
                          values_to = "Place",
                          values_drop_na = F)
```

```
#Let's see unique responses to our variable.
unique(plc_birth_long$Place)
```

```
##   [1] "PR"                          "USA"
##   [3] NA                            ""
##   [5] "Italy"                       "Japan"
##   [7] "China"                       "US"
##   [9] "Cuba"                        "Yugoslavia"
##  [11] "Morta"                       "Ireland"
##  [13] "Poland"                      "Belo"
##  [15] "Spain"                       "Greece"
##  [17] "Ukraine"                     "Germany"
##  [19] "38"                          "Scotland"
##  [21] "NY"                          "Dom. Rep."
##  [23] "Sweden"                      "Hungary"
##  [25] "Pr"                          "italy"
##  [27] "Russia"                      "South America"
##  [29] "greece"                      "cuba"
##  [31] "ireland"                     "england"
##  [33] "france"                      "germany"
##  [35] "norway"                      "cuban"
##  [37] "Canada"                      "hawaii"
##  [39] "peru"                        "Armenia"
##  [41] "Austria"                     "England"
##  [43] "Europe"                      "Puerto Rico"
##  [45] "46"                          "France"
##  [47] "Ukrainian"                   "Estonia"
##  [49] "Morocco"                     "Gibraltar"
##  [51] "I"                           "Dominican Republic"
##  [53] "russia"                      "British West Indies"
##  [55] "austria"                     "Jamaica, British West Indies "
##  [57] "cyprus"                      "Cyprus"
##  [59] "canada"                      "russian"
##  [61] "spain"                       "Phillipines"
##  [63] "B.W.I."                      "Romania"
##  [65] "Holland"                     "Trinidad"
##  [67] "malta"                       "1950"
##  [69] "1955"                        "2,000.00"
##  [71] "Belgium"                     "1925"
##  [73] "1956"                        "Mexico"
##  [75] "turkey"                      "ecuador"
##  [77] "1943"                        "C. Amer."
##  [79] "Turkey"                      "50"
##  [81] "U.S.A."                      "Denmark"
##  [83] "Costa Rica"                  "CUBA"
##  [85] "52"                          "1923"
##  [87] "IRELAND"                     "ENGLAND"
##  [89] "Luxembourg"                  "USa"
##  [91] "Nicaragua"                   "West Ind"
```

```
##  [93] "Hong Kong"                    "1926"
##  [95] "Colombia"                    "1954"
##  [97] "1933"                        "Philipines"
##  [99] "Haiti"                       "virgin islands"
## [101] "Virgin Islands"              "St. Thomas"
## [103] "Virgin islands"              "Switzerland"
## [105] "1900"                        "1898"
## [107] "1952"                        "Philippines"
## [109] "BW"                          "1887"
## [111] "British Guiana"              "NORWAY"
## [113] "cananda"                     "1906"
## [115] "Phillipine Islands"          "Bali"
## [117] "BWI"                         "South Africa"
## [119] "INDIA"                       "1922"
## [121] "pakistan"                    "68"
## [123] "Martinique"                  "grecce"
## [125] "Columbia"                    "PRR"
## [127] "1957"                        "Panama"
## [129] "Ecuador"                     "1928"
## [131] "1929"                        "1958"
## [133] "bulgaria"                    "1941"
## [135] "1949"                        "Latvia"
## [137] "Rog. Lat."                   "Malta"
## [139] "Brazil"                      "PA"
## [141] "So H"                        "Dom Rep"
## [143] "dom rep"                     "phillipines"
## [145] "DR"                          "60"
```

```
#We see that a number of cases have numeric responses that are closer to income or
#year of birth.
```

The responses given to place of birth are sporadic. They include misspellings, abbreviation, and numerical responses. We will begin addressing these errors by focusing on answers that are given using digits because they indicate both an error in one or more of the entries.

```
#Let's filter for cases where the response to place of birth was numeric or
#included numbers.
plc_birth_num <- plc_birth_long[str_detect(plc_birth_long$Place,("([0-9])")),]
plc_birth_num <- plc_birth_num %>% filter(!is.na(Place))
#There are 33 cases where the response to the Place of Birth was numeric.
```

```
#Now we will compare responses to the ones given in the original record.
#Create a new data set where we will alter mistakes in the original data entry.
data_v11 <- data_v9

#For record with ID 3105082501-60956-001_3105082501-60956-002 we have that the
#place of birth for person #2 was 38.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-001_3105082501-60956-002")

data_v11$p02_plc[which(data_v11$ID_long == "3105082501-60956-001_3105082501-60956-002")] <- "Poland"
data_v11$p02_incwk[which(data_v11$ID_long == "3105082501-60956-001_3105082501-60956-002")] <- "38.00"

#For record with ID 3105221470-60967-005_3105221470-60967-006 we see that the
```

```r
#place of birth of person #2 is given as 46.
t <- data_v11 %>% filter(ID_long == "3105221470-60967-005_3105221470-60967-006")

data_v11$p02_plc[which(data_v11$ID_long == "3105221470-60967-005_3105221470-60967-006")] <- "PR"
data_v11$p02_incwk[which(data_v11$ID_long == "3105221470-60967-005_3105221470-60967-006")] <- "46.00"

#For record with ID 3105296384-60979-025_3105296384-60979-026 we see that the
#place of birth for person #4 was given as 1950.
t <- data_v11 %>% filter(ID_long == "3105296384-60979-025_3105296384-60979-026")

data_v11$p04_plc[which(data_v11$ID_long == "3105296384-60979-025_3105296384-60979-026")] <- "Malta"
data_v11$p04_incwk[which(data_v11$ID_long == "3105296384-60979-025_3105296384-60979-026")] <- NA

#For record with ID 3105302993-60982-005_3105302993-60982-006 we see that the
#place of birth of for person #10 was given as 1955.
t <- data_v11 %>% filter(ID_long == "3105302993-60982-005_3105302993-60982-006")

data_v11$p010_plc[which(data_v11$ID_long == "3105302993-60982-005_3105302993-60982-006")] <- "PR"
data_v11$p010_rel[which(data_v11$ID_long == "3105302993-60982-005_3105302993-60982-006")] <- "Son"
data_v11$p010_year[which(data_v11$ID_long == "3105302993-60982-005_3105302993-60982-006")] <- "1955"

#For record with ID 3105302993-60982-013_3105302993-60982-014 we see that the
#place of birth for person #2 was 2000.00.
t <- data_v11 %>% filter(ID_long == "3105302993-60982-013_3105302993-60982-014")

data_v11$p02_plc[which(data_v11$ID_long == "3105302993-60982-013_3105302993-60982-014")] <- "USA"
data_v11$p02_incan[which(data_v11$ID_long == "3105302993-60982-013_3105302993-60982-014")] <- "2000.00"


#For record with ID 3195364380-61052-021_3195364380-61052-022 we see that the
#place of birth for person #1 was 1925
t <- data_v11 %>% filter(ID_long == "3195364380-61052-021_3195364380-61052-022")

data_v11$p01_plc[which(data_v11$ID_long == "3195364380-61052-021_3195364380-61052-022")] <- "USA"
data_v11$p01_incwk[which(data_v11$ID_long == "3195364380-61052-021_3195364380-61052-022")] <- "60.00"

#For record with ID 3195368322-61081-015_3195368322-61081-016 we see that the
#place of birth for person #4 was 1956.
t <- data_v11 %>% filter(ID_long == "3195368322-61081-015_3195368322-61081-016")

data_v11$p04_plc[which(data_v11$ID_long == "3195368322-61081-015_3195368322-61081-016")] <- "USA"
data_v11$p04_incwk[which(data_v11$ID_long == "3195368322-61081-015_3195368322-61081-016")] <- NA

#For record with ID 3262901700-60987-007_3262901700-60987-008 we see that the
#place of birth for person #4 was 1943
t <- data_v11 %>% filter(ID_long == "3262901700-60987-007_3262901700-60987-008")

data_v11$p04_plc[which(data_v11$ID_long == "3262901700-60987-007_3262901700-60987-008")] <- "PR"

#For record with ID 3262901700-60987-033_3262901700-60987-034 we see that the
#place of birth for person #3 was 50.
t <- data_v11 %>% filter(ID_long == "3262901700-60987-033_3262901700-60987-034")
```

```r
data_v11$p03_plc[which(data_v11$ID_long == "3262901700-60987-033_3262901700-60987-034")] <- "PR"
data_v11$p03_incwk[which(data_v11$ID_long == "3262901700-60987-033_3262901700-60987-034")] <- "50.00"

#For record with ID 3262902338-60989-019_3262902338-60989-020 we see that the
#place of birth for person #1 was 46.
t <- data_v11 %>% filter(ID_long == "3262902338-60989-019_3262902338-60989-020")

data_v11$p01_rel[which(data_v11$ID_long == "3262902338-60989-019_3262902338-60989-020")] <- "H"
data_v11$p01_year[which(data_v11$ID_long == "3262902338-60989-019_3262902338-60989-020")] <- "1940"
data_v11$p01_plc[which(data_v11$ID_long == "3262902338-60989-019_3262902338-60989-020")] <- "PR"
data_v11$p01_incwk[which(data_v11$ID_long == "3262902338-60989-019_3262902338-60989-020")] <- "46.00"

#For record with ID 3262902617-60990-007_3262902617-60990-008 we see that the
#place of birth for person #3 was 52.
t <- data_v11 %>% filter(ID_long == "3262902617-60990-007_3262902617-60990-008")

data_v11$p03_plc[which(data_v11$ID_long == "3262902617-60990-007_3262902617-60990-008")] <- "USA"
data_v11$p03_incwk[which(data_v11$ID_long == "3262902617-60990-007_3262902617-60990-008")] <- "52.00"

#For record with ID 3262902861-60991-001_3262902861-60991-002 we see that the
#place of birth for person #1 was 1923.
t <- data_v11 %>% filter(ID_long == "3262902861-60991-001_3262902861-60991-002")

data_v11$p01_year[which(data_v11$ID_long == "3262902861-60991-001_3262902861-60991-002")] <- "1923"
data_v11$p01_plc[which(data_v11$ID_long == "3262902861-60991-001_3262902861-60991-002")] <- "USA"

#For record  with ID 3262908762-60101-005_3262908762-60101-006 we see that the
#place of birth for person #2 was 1926.
t <- data_v11 %>% filter(ID_long == "3262908762-60101-005_3262908762-60101-006")

data_v11$p02_year[which(data_v11$ID_long == "3262908762-60101-005_3262908762-60101-006")] <- "1926"
data_v11$p02_plc[which(data_v11$ID_long == "3262908762-60101-005_3262908762-60101-006")] <- "PR"

#For record with ID 3262911388-61003-007_3262911388-61003-008 we see that the
#place of birth for person #3 was 1954.
t <- data_v11 %>% filter(ID_long == "3262911388-61003-007_3262911388-61003-008")

data_v11$p02_incwk[which(data_v11$ID_long == "3262911388-61003-007_3262911388-61003-008")] <- "welfare"
data_v11$p03_rel[which(data_v11$ID_long == "3262911388-61003-007_3262911388-61003-008")] <- "D"
data_v11$p03_year[which(data_v11$ID_long == "3262911388-61003-007_3262911388-61003-008")] <- "1954"
data_v11$p03_plc[which(data_v11$ID_long == "3262911388-61003-007_3262911388-61003-008")] <- "USA"
data_v11$p03_incwk[which(data_v11$ID_long == "3262911388-61003-007_3262911388-61003-008")] <- NA

#For record with ID 3262913528-61008-005_3262913528-61008-006 we see that the
#place of birth for person #2 was 1933.
t <- data_v11 %>% filter(ID_long == "3262913528-61008-005_3262913528-61008-006")

data_v11$p02_year[which(data_v11$ID_long == "3262913528-61008-005_3262913528-61008-006")] <- "1933"
data_v11$p02_plc[which(data_v11$ID_long == "3262913528-61008-005_3262913528-61008-006")] <- "PR"

#For record with ID 3262921422-61017-023_3262921422-61017-024 we see that the
#place of birth for person #1 was 1900.
t <- data_v11 %>% filter(ID_long == "3262921422-61017-023_3262921422-61017-024")
```

```
data_v11$p01_year[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "1900"
data_v11$p01_plc[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "PR"

#For record with ID 3262921422-61017-023_3262921422-61017-024 we see that the
#place of birth for person #2 was 1898.
t <- data_v11 %>% filter(ID_long == "3262921422-61017-023_3262921422-61017-024")

data_v11$p02_year[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "1898"
data_v11$p02_plc[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "PR"

#For record with ID 3262921422-61017-023_3262921422-61017-024 we see that the
#place of birth for person #3 was 1952.
t <- data_v11 %>% filter(ID_long == "3262921422-61017-023_3262921422-61017-024")

data_v11$p03_year[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "1952"
data_v11$p03_plc[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "USA"

#For record with ID 3262921422-61017-023_3262921422-61017-024 we see that the
#place of birth for person #4 was 1952.
t <- data_v11 %>% filter(ID_long == "3262921422-61017-023_3262921422-61017-024")

data_v11$p04_year[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "1956"
data_v11$p04_plc[which(data_v11$ID_long == "3262921422-61017-023_3262921422-61017-024")] <- "USA"

#For record with ID 3458104431-61035-011_3458104431-61035-012 we see that the
#place of birth for person #1 was 1887.
t <- data_v11 %>% filter(ID_long == "3458104431-61035-011_3458104431-61035-012")

data_v11$p01_year[which(data_v11$ID_long == "3458104431-61035-011_3458104431-61035-012")] <- "1887"
data_v11$p01_plc[which(data_v11$ID_long == "3458104431-61035-011_3458104431-61035-012")] <- "Germany"

#For record with ID 3458105308-61032-013_3458105308-61032-014 we see that the
#place of birth for person #1 was 1906.
t <- data_v11 %>% filter(ID_long == "3458105308-61032-013_3458105308-61032-014")

data_v11$p01_rel[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "H"
data_v11$p01_year[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "1906"
data_v11$p01_plc[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "USA"
data_v11$p01_incwk[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "65.00"
data_v11$p01_incmo[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- NA

#For record with ID 3458109563-61046-019_3458109563-61046-020 we see that the
#place of birth for person #1 was 1922.
t <- data_v11 %>% filter(ID_long == "3458109563-61046-019_3458109563-61046-020")

data_v11$p01_rel[which(data_v11$ID_long == "3458109563-61046-019_3458109563-61046-020")] <- "Head"
data_v11$p01_year[which(data_v11$ID_long == "3458109563-61046-019_3458109563-61046-020")] <- "1922"
data_v11$p01_plc[which(data_v11$ID_long == "3458109563-61046-019_3458109563-61046-020")] <- "PR"
data_v11$p01_incwk[which(data_v11$ID_long == "3458109563-61046-019_3458109563-61046-020")] <- "98.00 bi
data_v11$p01_incmo[which(data_v11$ID_long == "3458109563-61046-019_3458109563-61046-020")] <- NA

#For record with ID 3458110635-61049-079_3458110635-61049-080 we see that the
#place of birth for person #7 was 1956. ***
```

```r
t <- data_v11 %>% filter(ID_long == "3458110635-61049-079_3458110635-61049-080")

data_v11$p07_rel[which(data_v11$ID_long == "3458110635-61049-079_3458110635-61049-080")] <- "Son"
data_v11$p07_year[which(data_v11$ID_long == "3458110635-61049-079_3458110635-61049-080")] <- "1956"
data_v11$p07_plc[which(data_v11$ID_long == "3458110635-61049-079_3458110635-61049-080")] <- "USA"
data_v11$p07_incwk[which(data_v11$ID_long == "3458110635-61049-079_3458110635-61049-080")] <- NA

#For record with ID 3458129566-61050-047_3458129566-61050-048 we see that the
#place of birth for person #1 was 68.
t <- data_v11 %>% filter(ID_long == "3458129566-61050-047_3458129566-61050-048")

data_v11$p01_rel[which(data_v11$ID_long == "3458129566-61050-047_3458129566-61050-048")] <- "H"
data_v11$p01_year[which(data_v11$ID_long == "3458129566-61050-047_3458129566-61050-048")] <- "1936"
data_v11$p01_plc[which(data_v11$ID_long == "3458129566-61050-047_3458129566-61050-048")] <- "USA"
data_v11$p01_incwk[which(data_v11$ID_long == "3458129566-61050-047_3458129566-61050-048")] <- "68.00"

#For record with ID 3458129566-61050-057_3458129566-61050-058 we see that the
#place of birth for person #5 was 1954.
t <- data_v11 %>% filter(ID_long == "3458129566-61050-057_3458129566-61050-058")

data_v11$p04_rel[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "D of 2"
data_v11$p04_year[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "1952"
data_v11$p04_plc[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "USA"
data_v11$p05_rel[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "D of 2"
data_v11$p05_year[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "1954"
data_v11$p05_plc[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "USA"

#For record with ID 3568902147-61059-019_3568902147-61059-020 we see that the
#place of birth for person #5 was 1957.
t <- data_v11 %>% filter(ID_long == "3568902147-61059-019_3568902147-61059-020")

data_v11$p05_rel[which(data_v11$ID_long == "3568902147-61059-019_3568902147-61059-020")] <- "D"
data_v11$p05_year[which(data_v11$ID_long == "3568902147-61059-019_3568902147-61059-020")] <- "1955"
data_v11$p05_plc[which(data_v11$ID_long == "3568902147-61059-019_3568902147-61059-020")] <- "USA"
data_v11$p05_incwk[which(data_v11$ID_long == "3568902147-61059-019_3568902147-61059-020")] <- NA

#For record with ID 3568905752-61063-023_3568905752-61063-024 we see that the
#place of birth for person #2 was 1928.
t <- data_v11 %>% filter(ID_long == "3568905752-61063-023_3568905752-61063-024")

data_v11$p02_year[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- "1928"
data_v11$p02_plc[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- NA

#For record with ID 3568905752-61063-023_3568905752-61063-024 we see that the
#place of birth for person #3 was 1928.
t <- data_v11 %>% filter(ID_long == "3568905752-61063-023_3568905752-61063-024")

data_v11$p03_year[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- "1928"
data_v11$p03_plc[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- NA

#For record with ID 3568905752-61063-023_3568905752-61063-024 we see that the
#place of birth for person #4 was 1929.
t <- data_v11 %>% filter(ID_long == "3568905752-61063-023_3568905752-61063-024")
```

```
data_v11$p04_year[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- "1929"
data_v11$p04_plc[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- NA

#For record with ID 3568905752-61063-023_3568905752-61063-024 we see that the
#place of birth for person #5 was 1928.
t <- data_v11 %>% filter(ID_long == "3568905752-61063-023_3568905752-61063-024")

data_v11$p05_year[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- "1958"
data_v11$p05_plc[which(data_v11$ID_long == "3568905752-61063-023_3568905752-61063-024")] <- NA

#For record with ID 3568915429-61077-001_3568915429-61077-002 we see that the
#place of birth for person #6 was 1941.
t <- data_v11 %>% filter(ID_long == "3568915429-61077-001_3568915429-61077-002")

data_v11$p06_rel[which(data_v11$ID_long == "3568915429-61077-001_3568915429-61077-002")] <- "S"
data_v11$p06_year[which(data_v11$ID_long == "3568915429-61077-001_3568915429-61077-002")] <- "1941"
data_v11$p06_plc[which(data_v11$ID_long == "3568915429-61077-001_3568915429-61077-002")] <- "PR"
data_v11$p06_incwk[which(data_v11$ID_long == "3568915429-61077-001_3568915429-61077-002")] <- NA

#For record with ID 3687346643-61085-001_3687346643-61085-002 we see that the
#place of birth for person #3 was 1949.
t <- data_v11 %>% filter(ID_long == "3687346643-61085-001_3687346643-61085-002")

data_v11$p03_rel[which(data_v11$ID_long == "3687346643-61085-001_3687346643-61085-002")] <- "D"
data_v11$p03_year[which(data_v11$ID_long == "3687346643-61085-001_3687346643-61085-002")] <- "1949"
data_v11$p03_plc[which(data_v11$ID_long == "3687346643-61085-001_3687346643-61085-002")] <- "PR"
data_v11$p03_incwk[which(data_v11$ID_long == "3687346643-61085-001_3687346643-61085-002")] <- NA

#For record with ID 3687371744-61112-021_3687371744-61112-022 we see that the
#place of birth for person #1 was 60.00.
t <- data_v11 %>% filter(ID_long == "3687371744-61112-021_3687371744-61112-022")

data_v11$p01_rel[which(data_v11$ID_long == "3687371744-61112-021_3687371744-61112-022")] <- "H"
data_v11$p01_year[which(data_v11$ID_long == "3687371744-61112-021_3687371744-61112-022")] <- "1926"
data_v11$p01_plc[which(data_v11$ID_long == "3687371744-61112-021_3687371744-61112-022")] <- "PR"
data_v11$p01_incwk[which(data_v11$ID_long == "3687371744-61112-021_3687371744-61112-022")] <- "60"
```

We can now check that there are no missing responses with digits.

```
#We can create a similar data set as we had initially done so to clean the data but using data_v11.
plc_birth <- data_v11 %>% #We use data_v9 becuase we want to make changes on all.
  select(ID_long,p01_plc,p02_plc,p03_plc,p04_plc,p05_plc,
         p06_plc,p07_plc,p08_plc,p09_plc,p010_plc,p011_plc,p012_plc)

#We can now convert the data set into a long format that allows us to look at
#the unique responses to our variable
plc_birth_long <- pivot_longer(plc_birth,
                           cols = starts_with("p0"),
                           names_to = "var",
                           values_to = "Place",
                           values_drop_na = F)

plc_birth_num <- plc_birth_long[str_detect(plc_birth_long$Place,("([0-9])")),]
```

```r
unique(plc_birth_long$Place)
```

```
##   [1] "PR"                          "USA"
##   [3] NA                            ""
##   [5] "Italy"                       "Japan"
##   [7] "China"                       "US"
##   [9] "Cuba"                        "Yugoslavia"
##  [11] "Morta"                       "Ireland"
##  [13] "Poland"                      "Belo"
##  [15] "Spain"                       "Greece"
##  [17] "Ukraine"                     "Germany"
##  [19] "Scotland"                    "NY"
##  [21] "Dom. Rep."                   "Sweden"
##  [23] "Hungary"                     "Pr"
##  [25] "italy"                       "Russia"
##  [27] "South America"               "greece"
##  [29] "cuba"                        "ireland"
##  [31] "england"                     "france"
##  [33] "germany"                     "norway"
##  [35] "cuban"                       "Canada"
##  [37] "hawaii"                      "peru"
##  [39] "Armenia"                     "Austria"
##  [41] "England"                     "Europe"
##  [43] "Puerto Rico"                 "France"
##  [45] "Ukrainian"                   "Estonia"
##  [47] "Morocco"                     "Gibraltar"
##  [49] "I"                           "Dominican Republic"
##  [51] "russia"                      "British West Indies"
##  [53] "austria"                     "Jamaica, British West Indies "
##  [55] "cyprus"                      "Cyprus"
##  [57] "canada"                      "russian"
##  [59] "spain"                       "Phillipines"
##  [61] "B.W.I."                      "Romania"
##  [63] "Holland"                     "Trinidad"
##  [65] "malta"                       "Malta"
##  [67] "Belgium"                     "Mexico"
##  [69] "turkey"                      "ecuador"
##  [71] "C. Amer."                    "Turkey"
##  [73] "U.S.A."                      "Denmark"
##  [75] "Costa Rica"                  "CUBA"
##  [77] "IRELAND"                     "ENGLAND"
##  [79] "Luxembourg"                  "USa"
##  [81] "Nicaragua"                   "West Ind"
##  [83] "Hong Kong"                   "Colombia"
##  [85] "Philipines"                  "Haiti"
##  [87] "virgin islands"              "Virgin Islands"
##  [89] "St. Thomas"                  "Virgin islands"
##  [91] "Switzerland"                 "Philippines"
##  [93] "BW"                          "British Guiana"
##  [95] "NORWAY"                      "cananda"
##  [97] "Phillipine Islands"          "Bali"
##  [99] "BWI"                         "South Africa"
## [101] "INDIA"                       "pakistan"
```

```
## [103] "Martinique"                  "grecce"
## [105] "Columbia"                    "PRR"
## [107] "Panama"                      "Ecuador"
## [109] "bulgaria"                    "Latvia"
## [111] "Rog. Lat."                   "Brazil"
## [113] "PA"                          "So H"
## [115] "Dom Rep"                     "dom rep"
## [117] "phillipines"                 "DR"
```

We see that no cases are left with digits in their answer. We can now focus on grammatical errors and other mistakes.

```
#Here we will be evaluating and replacing values that are nor written correctly.
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place," ","")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"IRELAND","Ireland")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"P.R.","PR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"PRR","PR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"U.S.A.","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"NY","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"usa","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"pr","PR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Pr","PR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"ireland","Ireland")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"cuba","Cuba")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Cuban","Cuba")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"PuertoRico","PR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"USa","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"NY","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"hawaii","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"US","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Dom.Rep.","DR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"italy","Italy")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"greece","Greece")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"england","England")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"france","France")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"germany","Germany")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"norway","Norway")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"cuban","Cuba")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"peru","Peru")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"hawaii","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"cananda","Canada")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Ukranian","Ukraine")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"DominicanRepublic","DR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"russia","Russia")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"austria","Austria")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"USAA","USA")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"cyPRus","Cyprus")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"CyPRus","Cyprus")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"canada","Canada")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"spain","Spain")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"malta","Malta")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"turkey","Turkey")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"ecuador","Ecuador")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"CUBA","Cuba")
```

```
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"ENGLAND","England")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"virginislands","Virgin Islands")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"VirginIslands","Virgin Islands")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Virginislands","Virgin Islands")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"NORWAY","Norway")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"PhillipinesIslands","Phillipines")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"PhillipineIslands","Phillipines")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Philippines","Phillipines")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"INDIA","India")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"pakistan","Pakistan")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"grecce","Greece")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"bulgaria","Bulgaria")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"DomRep","DR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"domrep","DR")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"phillipines","Phillipines")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"BritishWestIndies","WestInd")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Jamaica,WestInd","WestInd")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"BWI","WestInd")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"B.W.I.","WestInd")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Ukrainian","Ukraine")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Russian","Russia")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"PA","Panama")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"SoH","")
plc_birth_long$Place <- str_replace_all(plc_birth_long$Place,"Rog.Lat.","")
```

We can now check for unique answers and look for any remaining mistakes.

```
unique(plc_birth_long$Place)
```

```
##  [1] "PR"            "USA"           NA              ""
##  [5] "Italy"         "Japan"         "China"         "Cuba"
##  [9] "Yugoslavia"    "Morta"         "Ireland"       "Poland"
## [13] "Belo"          "Spain"         "Greece"        "Ukraine"
## [17] "Germany"       "Scotland"      "DR"            "Sweden"
## [21] "Hungary"       "Russia"        "SouthAmerica"  "England"
## [25] "France"        "Norway"        "Canada"        "Peru"
## [29] "Armenia"       "Austria"       "Europe"        "Estonia"
## [33] "Morocco"       "Gibraltar"     "I"             "WestInd"
## [37] "Cyprus"        "Phillipines"   "Romania"       "Holland"
## [41] "Trinidad"      "Malta"         "Belgium"       "Mexico"
## [45] "Turkey"        "Ecuador"       "C.Amer."       "Denmark"
## [49] "CostaRica"     "Luxembourg"    "Nicaragua"     "HongKong"
## [53] "Colombia"      "Philipines"    "Haiti"         "Virgin Islands"
## [57] "St.Thomas"     "Switzerland"   "BW"            "BritishGuiana"
## [61] "Bali"          "SouthAfrica"   "India"         "Pakistan"
## [65] "Martinique"    "Columbia"      "Panama"        "Bulgaria"
## [69] "Latvia"        "Brazil"
```

We now have a standardized set of answers to place of birth. To append this set to the original data set we will need to create a new set of place of birth variables.

```
#Because some of the data analysis being carried out here has been done before, place #of birth was alr
#created and re create them.

data_v11 <- data_v11 %>% select(-c(p01_plc_2:race2))

#Now we can create a new set of variables for place of birth that corresponds to the
#updated changes.

#We can convert thhe our long data set to a wide data set.
plc_birth_wide <- plc_birth_long %>%
                    pivot_wider(names_from = "var",
                                values_from = "Place")

#Now we have a data frame that is wide and can be joined one to one with our
#original data frame.
view(plc_birth_wide)

#We will change the names of our variables to include a 2 at the end as to
#indicate that they are duplicates when we join the wide data frame to the original.
colnames(plc_birth_wide) <- paste(colnames(plc_birth_wide),"2",sep="_")
names(plc_birth_wide)
```

```
##  [1] "ID_long_2"  "p01_plc_2"  "p02_plc_2"  "p03_plc_2"  "p04_plc_2"
##  [6] "p05_plc_2"  "p06_plc_2"  "p07_plc_2"  "p08_plc_2"  "p09_plc_2"
## [11] "p010_plc_2" "p011_plc_2" "p012_plc_2"
```

```
#Now we will perform our join and clean some of the unnecessary variables.
data_v11 <- left_join(data_v11,plc_birth_wide, by = c("ID_long"="ID_long_2"))
```

Now we have a data set with an extra set of variables that contained a cleaned version of the place of birth.

Year of birth

```
#We will select the variables that correspond to year of birth. This
#includes the year of birth for all members of the family.
yr_birth <- data_v11 %>%
  select(ID_long,p01_year,p02_year,p03_year,p04_year,p05_year,
         p06_year,p07_year,p08_year,p09_year,p010_year,p011_year,p012_year)

#We know that there are variables related to year of birth that have different
#types.
summary(yr_birth)
```

```
##    ID_long            p01_year           p02_year           p03_year
##  Length:2130        Length:2130        Length:2130        Length:2130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##    p04_year           p05_year           p06_year           p07_year
##  Length:2130        Length:2130        Length:2130        Length:2130
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     p08_year       p09_year     p010_year           p011_year      p012_year
##  Min.   :1057   Min.   :1888   Length:2130        Min.   :1956   Min.   :1957
##  1st Qu.:1952   1st Qu.:1952   Class :character   1st Qu.:1956   1st Qu.:1957
##  Median :1954   Median :1955   Mode  :character   Median :1957   Median :1957
##  Mean   :1937   Mean   :1949                      Mean   :1957   Mean   :1957
##  3rd Qu.:1957   3rd Qu.:1956                      3rd Qu.:1958   3rd Qu.:1957
##  Max.   :1958   Max.   :1959                      Max.   :1958   Max.   :1957
##  NA's   :2070   NA's   :2100                      NA's   :2128   NA's   :2129
```

Since some of the variables related to place of birth are character and others numeric, we will first focus on cleaning the variables that are type character.

```
#We will select year of birth variables where the type is character.
yr_birth_ch <- yr_birth %>% select_if(is.character)
yr_birth_ch %>% head(2)
```

```
##                                      ID_long p01_year p02_year p03_year p04_year
## 1 3104718669-60949-001_3104718669-60949-002     1931     1920     1941     1946
## 2 3104718669-60949-003_3104718669-60949-004     1931     1931     1951     1954
##   p05_year p06_year p07_year p010_year
## 1     <NA>     <NA>     <NA>      <NA>
## 2     1955     <NA>     <NA>      <NA>
```

```
#We can now convert the data set into a long format that allows us to look at
#the unique responses to our variable
yr_birth_ch_long <- pivot_longer(yr_birth_ch,
                                 cols = starts_with("p0"),
                                 names_to = "var",
                                 values_to = "year",
                                 values_drop_na = F)

#Let's see unique responses to our variable.
unique(yr_birth_ch_long$year)
```

```
##   [1] "1931"          "1920"          "1941"          "1946"
##   [5] NA              "1951"          "1954"          "1955"
##   [9] "1900"          "1902"          "1921"          "1914"
##  [13] "1899"          "1904"          "1933"          "1875"
##  [17] "1887"          "1883"          "92"            "1925"
##  [21] "1927"          "1945"          "1953"          "1912"
##  [25] "1913"          "1874"          "1928"          "1956"
##  [29] "1923"          "1937"          "1903"          "1895"
##  [33] "1924"          "1948"          "1950"          "1908"
##  [37] "1926"          "1952"          "1934"          "1940"
```

```
##  [41] "1942"          "1932"          "1957"          "1930"
##  [45] "1935"          "1929"          "1898"          "1936"
##  [49] "1894"          "1877"          "1922"          "1 -"
##  [53] "1889"          "1943"          "1944"          "1916"
##  [57] "1901"          "1882"          "1909"          "1897"
##  [61] "1947"          "1949"          "1886"          "61"
##  [65] "65"            "1958"          "1919"          "1938"
##  [69] "1911"          "1915"          "1939"          "1907"
##  [73] "1896"          "1905"          "1906"          "1910"
##  [77] "1918"          "(unclear)"     "21+"           "6"
##  [81] "4"             "1893"          "29"            "66"
##  [85] "21"            "20"            "1917"          "1879"
##  [89] "1881"          "1629"          "Son"           "1878"
##  [93] "1884"          "1891"          "1890"          "1885"
##  [97] "1892"          "1888"          "Spain"         "1870"
## [101] "USA"           "1925=6"        "1880"          "1047"
## [105] "1876"          "1872"          "1866"          "1832"
## [109] "1818"          "S"             "PR"            "152"
## [113] "190"           "19448"         "1959"          "1991"
## [117] "9153"          "W"             "7/27/1930"     "11/6/1953"
## [121] "4/8/1957"      "1053"          "deceased 1956" "1052"
## [125] "1962"          "1873"          "1934PR"        "1988"
## [129] "1848"          "1863"          "1968"
```

```
length(unique(yr_birth_ch_long$year))
```

```
## [1] 131
```

```
#We see that there are 133 unique responses.
```

Now we can filter for values that are not years of birth. Because of the #nature of the data, and after visual evaluation,we will filter for cases where #the answer given does not start with 18 or 19, which represent year dates that #would correspond with the time frame.

```
yr_birth_wrng <- yr_birth_ch_long %>% filter(!str_starts(year,"19|18"))
```

There are 52 responses that do not begin with 18 or 19. We have determined these #responses are "wrong". Besides these, we can identify some cases that do begin #with either 18 or 19 that have incorrect information. We will deal with these later.

```
#For record with ID 3104734180-60950-001_3104734180-60950-002 the year of
#birth for person #1 was 92.
t <- data_v11 %>% filter(ID_long == "3104734180-60950-001_3104734180-60950-002")
#The record indicates that the year of birth was 92. Based on the response to
#the "time in city" question, which was also 92, it can be assumed that the
#answer given reflects the age of the person. Hence we can calculate their age.

data_v11$p01_year[which(data_v11$ID_long == "3104734180-60950-001_3104734180-60950-002")] <- "1866"

#For record with ID 3104744331-60951-017_3104744331-60951-018 the year of birth
#for person 1 was 1 -.
t <- data_v11 %>% filter(ID_long == "3104744331-60951-017_3104744331-60951-018")
```

```r
#The information record during data entry corresponds with the information provided
#in the relocation record. There is no supplemental information that points to
#a specific year of birth. Changing the answer given to NA.

data_v11$p01_year[which(data_v11$ID_long == "3104744331-60951-017_3104744331-60951-018")] <- NA

#For record with ID 3105040525-60953-009_3105040525-60953-010 the year of birth
#for person #1 was 61.
t <- data_v11 %>% filter(ID_long == "3105040525-60953-009_3105040525-60953-010")
#The information recorded in the data entry is consistent with the information
#in the relocation record. Person #2 in the same record also has a similar
#situation as Person #1. We will presume the answer given is their age.

data_v11$p01_year[which(data_v11$ID_long == "3105040525-60953-009_3105040525-60953-010")] <- "1897"
data_v11$p02_year[which(data_v11$ID_long == "3105040525-60953-009_3105040525-60953-010")] <- "1893"

#For record with ID 3105082501-60956-001_3105082501-60956-002 the year of birth
#for person #2 was indicated to be unclear.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-001_3105082501-60956-002")
#In reviewing the relocation record, the information given on year of birth for
#person #2 is unclear.

data_v11$p02_year[which(data_v11$ID_long == "3105082501-60956-001_3105082501-60956-002")] <- NA

#For record with ID 3105082501-60956-003_3105082501-60956-004 the year of birth
#of person #2 was 21+.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-003_3105082501-60956-004")
#The information imputed for person #2 is consistent with available information
#in the relocation record. Because this is not sufficient information and no other
#information is provided in record to clarify we are changing the answer to NA.

data_v11$p02_year[which(data_v11$ID_long == "3105082501-60956-003_3105082501-60956-004")] <- NA

#For record with ID 3105082501-60956-003_3105082501-60956-004 the year of birth
#for person #3 was 65.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-003_3105082501-60956-004")
#We will assume that the number given refers to the age of the person and change
#accordingly.

data_v11$p03_year[which(data_v11$ID_long == "3105082501-60956-003_3105082501-60956-004")] <- "1893"

#For record with ID 3105082501-60956-003_3105082501-60956-004 the year of birth
#for person #4 was 6.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-003_3105082501-60956-004")
#We will assume that the number given refers to the age of the person and change
#accordingly.

data_v11$p04_year[which(data_v11$ID_long == "3105082501-60956-003_3105082501-60956-004")] <- "1952"

#For record with ID 3105082501-60956-003_3105082501-60956-004 the year of birth
#for person #5 was 4.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-003_3105082501-60956-004")
#We will assume that the number given refers to the age of the person and change
```

```
#accordingly.

data_v11$p05_year[which(data_v11$ID_long == "3105082501-60956-003_3105082501-60956-004")] <- "1954"

#For record with ID 3105082501-60956-013_3105082501-60956-014 the year of birth
#for person #1 was 29
t <- data_v11 %>% filter(ID_long == "3105082501-60956-013_3105082501-60956-014")
#The information introduced during data entry reflects the information included
#in the relocation card. We will assume that the number given is the actual age
#and change accordingly.

data_v11$p01_year[which(data_v11$ID_long == "3105082501-60956-013_3105082501-60956-014")] <- "1929"

#For record with ID 3105082501-60956-013_3105082501-60956-014 the year of birth
#for person #2 was 66
t <- data_v11 %>% filter(ID_long == "3105082501-60956-013_3105082501-60956-014")
#The information introduced during data entry reflects the information included
#in the relocation card. We will assume that the number given is the actual age
#and change accordingly.

data_v11$p02_year[which(data_v11$ID_long == "3105082501-60956-013_3105082501-60956-014")] <- "1892"

#For record with ID 3105082501-60956-015_3105082501-60956-016 the year of birth
#for person #3 was 21.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-015_3105082501-60956-016")
#The information introduced during data entry reflects the information included
#in the relocation card. We will assume that the number given is the actual age
#and change accordingly.

data_v11$p03_year[which(data_v11$ID_long == "3105082501-60956-015_3105082501-60956-016")] <- "1937"

#For record with ID 3105082501-60956-015_3105082501-60956-016 the year of birth
#for person #4 was 20.
t <- data_v11 %>% filter(ID_long == "3105082501-60956-015_3105082501-60956-016")
#The information introduced during data entry reflects the information included
#in the relocation card. We will assume that the number given is the actual age
#and change accordingly.

data_v11$p04_year[which(data_v11$ID_long == "3105082501-60956-015_3105082501-60956-016")] <- "1938"

#For record with ID 3105154754-60958-001_3105154754-60958-002 the year of birth
#for person #2 was 1629.
t <- data_v11 %>% filter(ID_long == "3105154754-60958-001_3105154754-60958-002")

data_v11$p02_year[which(data_v11$ID_long == "3105154754-60958-001_3105154754-60958-002")] <- "1929"

#For record with ID 3105154754-60958-003_3105154754-60958-004 the year of birth
#for person #1 was 21+.
t <- data_v11 %>% filter(ID_long == "3105154754-60958-003_3105154754-60958-004")
#The information introduced during data entry reflects the information included
#in the relocation card. There is no information that can clarify a specific
#year of birth. We will change the response to NA.
```

```r
data_v11$p01_year[which(data_v11$ID_long == "3105154754-60958-003_3105154754-60958-004")] <- NA

#For record with ID 3105154754-60958-019_3105154754-60958-020 the year of birth
#for person #3 was Son.
t <- data_v11 %>% filter(ID_long == "3105154754-60958-019_3105154754-60958-020")

data_v11$p03_year[which(data_v11$ID_long == "3105154754-60958-019_3105154754-60958-020")] <- "1914"
data_v11$p03_rel[which(data_v11$ID_long == "3105154754-60958-019_3105154754-60958-020")] <- "Son"

#For record with ID 3105228825-60968-001_3105228825-60968-002 the year of birth
#for person #1 was Spain.
t <- data_v11 %>% filter(ID_long == "3105228825-60968-001_3105228825-60968-002")

data_v11$p01_year[which(data_v11$ID_long == "3105228825-60968-001_3105228825-60968-002")] <- "1899"
data_v11$p01_rel[which(data_v11$ID_long == "3105228825-60968-001_3105228825-60968-002")] <- "Head"

#For record with ID 3105247721-60970-025_3105247721-60970-026 the year of birth
#for person #1 was 21+.
t <- data_v11 %>% filter(ID_long == "3105247721-60970-025_3105247721-60970-026")
#The information introduced during data entry reflects the information included
#in the relocation card. There is no information that can clarify a specific
#year of birth. We will change the response to NA.

data_v11$p01_year[which(data_v11$ID_long == "3105247721-60970-025_3105247721-60970-026")] <- NA

#For record with ID 3105247721-60970-025_3105247721-60970-026 the year of birth
#for person #2 was 21+.
t <- data_v11 %>% filter(ID_long == "3105247721-60970-025_3105247721-60970-026")
#The information introduced during data entry reflects the information included
#in the relocation card. There is no information that can clarify a specific
#year of birth. We will change the response to NA.

data_v11$p02_year[which(data_v11$ID_long == "3105247721-60970-025_3105247721-60970-026")] <- NA

#For record with ID 3105255100-60971-003_3105255100-60971-004 the year of birth
#for person #1 was 21+.
t <- data_v11 %>% filter(ID_long == "3105255100-60971-003_3105255100-60971-004")
#The information introduced during data entry reflects the information included
#in the relocation card. There is no information that can clarify a specific
#year of birth. We will change the response to NA.

data_v11$p01_year[which(data_v11$ID_long == "3105255100-60971-003_3105255100-60971-004")] <- NA

#For record with ID 3105255100-60971-003_3105255100-60971-004 the year of birth
#for person #2 was 21+.
t <- data_v11 %>% filter(ID_long == "3105255100-60971-003_3105255100-60971-004")
#The information introduced during data entry reflects the information included
#in the relocation card. There is no information that can clarify a specific
#year of birth. We will change the response to NA.

data_v11$p02_year[which(data_v11$ID_long == "3105255100-60971-003_3105255100-60971-004")] <- NA

#For record with ID 3105288010-60976-011_3105288010-60976-012 the year of birth
```

```
#for person #1 was 21+.
t <- data_v11 %>% filter(ID_long == "3105288010-60976-011_3105288010-60976-012")

data_v11$p01_year[which(data_v11$ID_long == "3105288010-60976-011_3105288010-60976-012")] <- NA

#For record with ID 3105288010-60976-011_3105288010-60976-012 the year of birth
#for person #2 was 21+.
t <- data_v11 %>% filter(ID_long == "3105288010-60976-011_3105288010-60976-012")

data_v11$p02_year[which(data_v11$ID_long == "3105288010-60976-011_3105288010-60976-012")] <- NA

#For record with ID 3105288010-60976-011_3105288010-60976-012 the year of birth
#for person #3 was USA.
t <- data_v11 %>% filter(ID_long == "3105288010-60976-011_3105288010-60976-012")

data_v11$p03_year[which(data_v11$ID_long == "3105288010-60976-011_3105288010-60976-012")] <- NA
data_v11$p03_plc[which(data_v11$ID_long == "3105288010-60976-011_3105288010-60976-012")] <- "USA"

#For record with ID 3105300183-60981-005_3105300183-60981-006 the year of birth
#for person #4 was 1047.
t <- data_v11 %>% filter(ID_long == "3105300183-60981-005_3105300183-60981-006")

data_v11$p04_year[which(data_v11$ID_long == "3105300183-60981-005_3105300183-60981-006")] <- "1947"

#For record with ID 3195368322-61081-015_3195368322-61081-016 the year of birth
#for person #4 was USA.
t <- data_v11 %>% filter(ID_long == "3195368322-61081-015_3195368322-61081-016")

data_v11$p04_year[which(data_v11$ID_long == "3195368322-61081-015_3195368322-61081-016")] <- "1956"

#For record with ID 3262901700-60987-007_3262901700-60987-008 the year of birth
#for person #4 was S.
t <- data_v11 %>% filter(ID_long == "3262901700-60987-007_3262901700-60987-008")

data_v11$p04_year[which(data_v11$ID_long == "3262901700-60987-007_3262901700-60987-008")] <- "1943"
data_v11$p04_rel[which(data_v11$ID_long == "3262901700-60987-007_3262901700-60987-008")] <- "S"

#For record with ID 3262901700-60987-033_3262901700-60987-034 the year of birth
#for person #3 was PR.
t <- data_v11 %>% filter(ID_long == "3262901700-60987-033_3262901700-60987-034")

data_v11$p03_year[which(data_v11$ID_long == "3262901700-60987-033_3262901700-60987-034")] <- "1939"

#For record with ID 3262905663-60994-009_3262905663-60994-010 the year of birth
#for person #3 was 152.
t <- data_v11 %>% filter(ID_long == "3262905663-60994-009_3262905663-60994-010")

data_v11$p03_year[which(data_v11$ID_long == "3262905663-60994-009_3262905663-60994-010")] <- "1952"

#For record with ID 3262906641-60995-015_3262906641-60995-016 the year of birth
#for person #1 was USA.
t <- data_v11 %>% filter(ID_long == "3262906641-60995-015_3262906641-60995-016")
```

```
data_v11$p01_year[which(data_v11$ID_long == "3262906641-60995-015_3262906641-60995-016")] <- "1929"
data_v11$p01_rel[which(data_v11$ID_long == "3262906641-60995-015_3262906641-60995-016")] <- "H"

#For record with ID 3262911388-61003-009_3262911388-61003-010 the year of birth
#for person #6 was 9153.
t <- data_v11 %>% filter(ID_long == "3262911388-61003-009_3262911388-61003-010")

data_v11$p06_year[which(data_v11$ID_long == "3262911388-61003-009_3262911388-61003-010")] <- "1953"

#For record with ID 3262913099-61007-015_3262913099-61007-016 the year of birth
#for person #1 was W.
t <- data_v11 %>% filter(ID_long == "3262913099-61007-015_3262913099-61007-016")

data_v11$p01_year[which(data_v11$ID_long == "3262913099-61007-015_3262913099-61007-016")] <- NA
data_v11$p01_rel[which(data_v11$ID_long == "3262913099-61007-015_3262913099-61007-016")] <- "H"
data_v11$p02_rel[which(data_v11$ID_long == "3262913099-61007-015_3262913099-61007-016")] <- "W"

#For record with ID 3262913967-61009-013_3262913967-61009-014 the year of birth
#for person #1 was 7/27/1930.
t <- data_v11 %>% filter(ID_long == "3262913967-61009-013_3262913967-61009-014")
#The data imputed for year of birth includes day and month. This is the case also
# for persons 4 and 5 in the family. We will keep only the year.

data_v11$p01_year[which(data_v11$ID_long == "3262913967-61009-013_3262913967-61009-014")] <- "1930"
data_v11$p04_year[which(data_v11$ID_long == "3262913967-61009-013_3262913967-61009-014")] <- "1953"
data_v11$p05_year[which(data_v11$ID_long == "3262913967-61009-013_3262913967-61009-014")] <- "1957"

#For record with ID 3458104222-61036-001_3458104222-61036-002 the year of birth
#for person #4 was 1053.
t <- data_v11 %>% filter(ID_long == "3458104222-61036-001_3458104222-61036-002")

data_v11$p04_year[which(data_v11$ID_long == "3458104222-61036-001_3458104222-61036-002")] <- "1953"

#For record with ID 3458105308-61032-013_3458105308-61032-014 the year of birth
#for person #1 was H.
t <- data_v11 %>% filter(ID_long == "3458105308-61032-013_3458105308-61032-014")

data_v11$p01_year[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "1906"
data_v11$p01_rel[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "H"
data_v11$p01_plc[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "USA"
data_v11$p01_incwk[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "65.00"
data_v11$p01_incmo[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- NA

#For record with ID 3458105308-61032-013_3458105308-61032-014 the year of birth
#for person #1 was H.
t <- data_v11 %>% filter(ID_long == "3458105308-61032-013_3458105308-61032-014")

data_v11$p01_year[which(data_v11$ID_long == "3458105308-61032-013_3458105308-61032-014")] <- "1906"

#For record with ID 3458107841-61025-031_3458107841-61025-032 the year of birth
#for person #2 was deceased 1956.
t <- data_v11 %>% filter(ID_long == "3458107841-61025-031_3458107841-61025-032")
#The information introduced during data entry reflects the information included
```

```r
#in the relocation card. Because, as the record indicates, the person was deceased
#we will proceed with changing any information related to person 2 to NA.

data_v11$p02_year[which(data_v11$ID_long == "3458107841-61025-031_3458107841-61025-032")] <- NA
data_v11$p02_rel[which(data_v11$ID_long == "3458107841-61025-031_3458107841-61025-032")] <- NA

#For record with ID 3458109952-61047-017_3458109952-61047-018 the year of birth
#for person #3 was 1052.
t <- data_v11 %>% filter(ID_long == "3458109952-61047-017_3458109952-61047-018")

data_v11$p03_year[which(data_v11$ID_long == "3458109952-61047-017_3458109952-61047-018")] <- "1952"

#For record with ID 3458110635-61049-067_3458110635-61049-068 the year of birth
#for person #2 was USA.
t <- data_v11 %>% filter(ID_long == "3458110635-61049-067_3458110635-61049-068")

data_v11$p01_incan[which(data_v11$ID_long == "3458110635-61049-067_3458110635-61049-068")] <- NA
data_v11$p02_rel[which(data_v11$ID_long == "3458110635-61049-067_3458110635-61049-068")] <- "W"
data_v11$p02_year[which(data_v11$ID_long == "3458110635-61049-067_3458110635-61049-068")] <- "1912"
data_v11$p02_plc[which(data_v11$ID_long == "3458110635-61049-067_3458110635-61049-068")] <- "USA"

#For record with ID 3458129566-61050-001_3458129566-61050-002 the year of birth
#for person #1 was USA.
t <- data_v11 %>% filter(ID_long == "3458129566-61050-001_3458129566-61050-002")

data_v11$p01_year[which(data_v11$ID_long == "3458129566-61050-001_3458129566-61050-002")] <- NA
data_v11$p01_plc[which(data_v11$ID_long == "3458129566-61050-001_3458129566-61050-002")] <- "USA"

#For record with ID 3458129566-61050-001_3458129566-61050-002 the year of birth
#for person #2 was USA.
t <- data_v11 %>% filter(ID_long == "3458129566-61050-001_3458129566-61050-002")

data_v11$p02_year[which(data_v11$ID_long == "3458129566-61050-001_3458129566-61050-002")] <- NA
data_v11$p02_plc[which(data_v11$ID_long == "3458129566-61050-001_3458129566-61050-002")] <- "USA"

#For record with ID 3458129566-61050-057_3458129566-61050-058 the year of birth
#for person #4 was D of 2.
t <- data_v11 %>% filter(ID_long == "3458129566-61050-057_3458129566-61050-058")

data_v11$p04_year[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "1952"
data_v11$p04_plc[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "USA"

#For record with ID 3458129566-61050-057_3458129566-61050-058 the year of birth
#for person #5 was USA.
t <- data_v11 %>% filter(ID_long == "3458129566-61050-057_3458129566-61050-058")

data_v11$p05_year[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "1954"
data_v11$p05_plc[which(data_v11$ID_long == "3458129566-61050-057_3458129566-61050-058")] <- "USA"

#For record with ID 3568895288-61053-025_3568895288-61053-026 the year of birth
#for person #3 was USA.
t <- data_v11 %>% filter(ID_long == "3568895288-61053-025_3568895288-61053-026")
```

```
data_v11$p03_year[which(data_v11$ID_long == "3568895288-61053-025_3568895288-61053-026")] <- NA
data_v11$p03_plc[which(data_v11$ID_long == "3568895288-61053-025_3568895288-61053-026")] <- "USA"

#For record with ID 3568900074-61057-013_3568900074-61057-014 the year of birth
#for person #1 was PR.
t <- data_v11 %>% filter(ID_long == "3568900074-61057-013_3568900074-61057-014")

data_v11$p01_year[which(data_v11$ID_long == "3568900074-61057-013_3568900074-61057-014")] <- "1925"
data_v11$p01_rel[which(data_v11$ID_long == "3568900074-61057-013_3568900074-61057-014")] <- "Head"
data_v11$p01_plc[which(data_v11$ID_long == "3568900074-61057-013_3568900074-61057-014")] <- "PR"

#For record with ID 3568913303-61072-003_3568913303-61072-004 the year of birth
#for person #1 was PR.
t <- data_v11 %>% filter(ID_long == "3568913303-61072-003_3568913303-61072-004")
#We see that in this record Places of birth were indicated in the Year of birth
#question for all 4 members of the household. We will fix them all bellow.
data_v11$p01_year[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- NA
data_v11$p01_plc[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- "PR"
data_v11$p02_year[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- NA
data_v11$p02_plc[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- "PR"
data_v11$p03_year[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- NA
data_v11$p03_plc[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- "USA"
data_v11$p04_year[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- NA
data_v11$p04_plc[which(data_v11$ID_long == "3568913303-61072-003_3568913303-61072-004")] <- "USA"

#For record with ID 3687359409-61098-013_3687359409-61098-014 the year of birth
#for person #4 was USA.
t <- data_v11 %>% filter(ID_long == "3687359409-61098-013_3687359409-61098-014")

data_v11$p04_year[which(data_v11$ID_long == "3687359409-61098-013_3687359409-61098-014")] <- "1951"
data_v11$p04_rel[which(data_v11$ID_long == "3687359409-61098-013_3687359409-61098-014")] <- "S"
data_v11$p04_plc[which(data_v11$ID_long == "3687359409-61098-013_3687359409-61098-014")] <- "USA"

#For record with ID 3687361245-61102-017_3687361245-61102-018 the year of birth
#for person #1 was UPR.
t <- data_v11 %>% filter(ID_long == "3687361245-61102-017_3687361245-61102-018")

data_v11$p01_year[which(data_v11$ID_long == "3687361245-61102-017_3687361245-61102-018")] <- NA
data_v11$p01_plc[which(data_v11$ID_long == "3687361245-61102-017_3687361245-61102-018")] <- "PR"
```

Having dealt with records that either start with a string that is not 18 or 19 and those that are not numeric characters, we will now focus on records that #begin with 18 or 19 but may still be considered errors.

```
unique(yr_birth_ch_long$year)
```

```
##   [1] "1931"          "1920"          "1941"          "1946"
##   [5] NA              "1951"          "1954"          "1955"
##   [9] "1900"          "1902"          "1921"          "1914"
##  [13] "1899"          "1904"          "1933"          "1875"
##  [17] "1887"          "1883"          "92"            "1925"
##  [21] "1927"          "1945"          "1953"          "1912"
##  [25] "1913"          "1874"          "1928"          "1956"
##  [29] "1923"          "1937"          "1903"          "1895"
```

```
##  [33] "1924"          "1948"          "1950"          "1908"
##  [37] "1926"          "1952"          "1934"          "1940"
##  [41] "1942"          "1932"          "1957"          "1930"
##  [45] "1935"          "1929"          "1898"          "1936"
##  [49] "1894"          "1877"          "1922"          "1 -"
##  [53] "1889"          "1943"          "1944"          "1916"
##  [57] "1901"          "1882"          "1909"          "1897"
##  [61] "1947"          "1949"          "1886"          "61"
##  [65] "65"            "1958"          "1919"          "1938"
##  [69] "1911"          "1915"          "1939"          "1907"
##  [73] "1896"          "1905"          "1906"          "1910"
##  [77] "1918"          "(unclear)"     "21+"           "6"
##  [81] "4"             "1893"          "29"            "66"
##  [85] "21"            "20"            "1917"          "1879"
##  [89] "1881"          "1629"          "Son"           "1878"
##  [93] "1884"          "1891"          "1890"          "1885"
##  [97] "1892"          "1888"          "Spain"         "1870"
## [101] "USA"           "1925=6"        "1880"          "1047"
## [105] "1876"          "1872"          "1866"          "1832"
## [109] "1818"          "S"             "PR"            "152"
## [113] "190"           "19448"         "1959"          "1991"
## [117] "9153"          "W"             "7/27/1930"     "11/6/1953"
## [121] "4/8/1957"      "1053"          "deceased 1956" "1052"
## [125] "1962"          "1873"          "1934PR"        "1988"
## [129] "1848"          "1863"          "1968"
```

By looking at unique values for the year variable in yr_birth_ch_long we can see cases that begin with 19 or 18 but can be considered mistakes or outside of the realm of possibility. We will check and make changes when necessary.

```
#All these cases were typos in the process of data entry.
t <- yr_birth_ch_long[yr_birth_ch_long$year == "1925=6",]
data_v11$p02_year[which(data_v11$ID_long == "3105291390-60977-013_3105291390-60977-014")] <- "1926"

t <- yr_birth_ch_long[yr_birth_ch_long$year == "190",]
data_v11$p01_year[which(data_v11$ID_long == "3262906641-60995-009_3262906641-60995-010")] <- "1920"

t <- yr_birth_ch_long[yr_birth_ch_long$year == "19448",]
data_v11$p03_year[which(data_v11$ID_long == "3262906930-60996-007_3262906930-60996-008")] <- "1948"

t <- yr_birth_ch_long[yr_birth_ch_long$year == "1832",]
data_v11$p03_year[which(data_v11$ID_long == "3195236545-60983-005_3195236545-60983-006")] <- "1932"

t <- yr_birth_ch_long[yr_birth_ch_long$year == "1934PR",]
data_v11$p02_year[which(data_v11$ID_long == "3568906814-61064-013_3568906814-61064-014")] <- "1934"
t <- data_v11 %>% filter(ID_long == "3568906814-61064-013_3568906814-61064-014")
data_v11$p02_plc[which(data_v11$ID_long == "3568906814-61064-013_3568906814-61064-014")] <- "PR"
```

We can now create a data set with a combined all year of birth variables.

```
yr_birth <- data_v11 %>%
  select(ID_long,p01_year,p02_year,p03_year,p04_year,p05_year,
         p06_year,p07_year,p08_year,p09_year,p010_year,p011_year,p012_year)
```

```
yr_birth <- yr_birth %>%
  mutate_at(c("p01_year","p02_year","p03_year","p04_year","p05_year",
              "p06_year","p07_year","p08_year","p09_year","p010_year",
              "p011_year","p012_year"), as.numeric)
#There are no "NAs introduced by coercion" into the data set when converting our
#character variables into numeric variables. This indicates that all responses
#were indeed of numeric strings.
```

```
#We can use the summary function to look at the composition of the variables and
#look for outlier or values outside the realm of possibility.
summary(yr_birth)
```

```
##    ID_long             p01_year       p02_year       p03_year       p04_year
##  Length:2130        Min.   :1832   Min.   :1818   Min.   :1870   Min.   :1878
##  Class :character   1st Qu.:1904   1st Qu.:1915   1st Qu.:1940   1st Qu.:1944
##  Mode  :character   Median :1918   Median :1927   Median :1947   Median :1950
##                     Mean   :1915   Mean   :1925   Mean   :1944   Mean   :1947
##                     3rd Qu.:1927   3rd Qu.:1936   3rd Qu.:1952   3rd Qu.:1954
##                     Max.   :1991   Max.   :1962   Max.   :1958   Max.   :1959
##                     NA's   :513    NA's   :786    NA's   :1115   NA's   :1422
##     p05_year       p06_year       p07_year       p08_year       p09_year
##  Min.   :1887   Min.   :1866   Min.   :1886   Min.   :1057   Min.   :1888
##  1st Qu.:1946   1st Qu.:1948   1st Qu.:1950   1st Qu.:1952   1st Qu.:1952
##  Median :1952   Median :1953   Median :1953   Median :1954   Median :1955
##  Mean   :1949   Mean   :1949   Mean   :1950   Mean   :1937   Mean   :1949
##  3rd Qu.:1955   3rd Qu.:1956   3rd Qu.:1956   3rd Qu.:1957   3rd Qu.:1956
##  Max.   :1959   Max.   :1958   Max.   :1959   Max.   :1958   Max.   :1959
##  NA's   :1685   NA's   :1892   NA's   :2011   NA's   :2070   NA's   :2100
##    p010_year      p011_year      p012_year
##  Min.   :1949   Min.   :1956   Min.   :1957
##  1st Qu.:1954   1st Qu.:1956   1st Qu.:1957
##  Median :1955   Median :1957   Median :1957
##  Mean   :1955   Mean   :1957   Mean   :1957
##  3rd Qu.:1957   3rd Qu.:1958   3rd Qu.:1957
##  Max.   :1958   Max.   :1958   Max.   :1957
##  NA's   :2114   NA's   :2128   NA's   :2129
```

```
#The summary breakdown suggests a number of cases where values may be outside the
#realm of possibility.
```

To deal with these cases we will create a new data set in long format and filter for cases where responses indicate the year of birth was after 1958, the year the interviews were carried out, and before 1878, which would make individuals older than 80 years, over 10 years older than the life expectancy of the population at the time.

```
yr_birth_long <- pivot_longer(yr_birth,
                              cols = starts_with("p0"),
                              names_to = "var",
                              values_to = "year",
                              values_drop_na = F)

yr_birth_long_wrng <- yr_birth_long %>% filter(year > 1958 | year < 1878)
```

There are 30 cases where values introduced are higher than 1958, the year people #were being interviewed, or lower than 1878, which would make these individuals, over 80 years old. These records might be correct but we will check them because #they are significantly higher than the life expectancy of individuals in 1958.

We will use the IDs in yr_birth_long_wrng to compare to the original records if the values introduced are incorrect or not. Cases where a mistake was made will be corrected bellow.

```
#For record with ID 3195236545-60983-005_3195236545-60983-006 the year of birth
#for person #1 was 1832
t <- data_v11 %>% filter(ID_long == "3195236545-60983-005_3195236545-60983-006")
data_v11$p01_year[which(data_v11$ID_long == "3195236545-60983-005_3195236545-60983-006")] <- "1932"

#For record with ID 3262901378-60986-027_3262901378-60986-028 the year of birth
#for person #2 was 1818.
t <- data_v11 %>% filter(ID_long == "3262901378-60986-027_3262901378-60986-028")
data_v11$p02_year[which(data_v11$ID_long == "3262901378-60986-027_3262901378-60986-028")] <- "1918"

#For record with ID 3262911388-61003-005_3262911388-61003-006 the year of birth
#for person #1 was 1991.
t <- data_v11 %>% filter(ID_long == "3262911388-61003-005_3262911388-61003-006")
data_v11$p01_year[which(data_v11$ID_long == "3262911388-61003-005_3262911388-61003-006")] <- "1911"

#For record with ID 3458110635-61049-039_3458110635-61049-040 the year of birth
#for person #2 was 1959.
t <- data_v11 %>% filter(ID_long == "3458110635-61049-039_3458110635-61049-040")
data_v11$p02_year[which(data_v11$ID_long == "3458110635-61049-039_3458110635-61049-040")] <- "1957"
data_v11$rent_os[which(data_v11$ID_long == "3458110635-61049-039_3458110635-61049-040")] <- "18 wk" #ad

#For record with ID 3568902147-61059-035_3568902147-61059-036 the year of birth
#for person 2 was 1962.
t <- data_v11 %>% filter(ID_long == "3568902147-61059-035_3568902147-61059-036")
data_v11$p02_year[which(data_v11$ID_long == "3568902147-61059-035_3568902147-61059-036")] <- "1896"

#For record with ID 3568902741-61060-021_3568902741-61060-022 the year of birth
#for person #3 was 1873. There is no person #3 in the record. All data for person
#3 will be deleted.
t <- data_v11 %>% filter(ID_long == "3568902741-61060-021_3568902741-61060-022")
data_v11$p03_year[which(data_v11$ID_long == "3568902741-61060-021_3568902741-61060-022")] <- NA
data_v11$p03_rel[which(data_v11$ID_long == "3568902741-61060-021_3568902741-61060-022")] <- NA
data_v11$p03_plc[which(data_v11$ID_long == "3568902741-61060-021_3568902741-61060-022")] <- NA
data_v11$p03_incwk[which(data_v11$ID_long == "3568902741-61060-021_3568902741-61060-022")] <- NA

#For record with ID 3568913707-61073-001_3568913707-61073-002 the year of birth
#for person #9 was 1959.
t <- data_v11 %>% filter(ID_long == "3568913707-61073-001_3568913707-61073-002")
data_v11$p09_year[which(data_v11$ID_long == "3568913707-61073-001_3568913707-61073-002")] <- "1957"

#For record with ID 3568915830-61078-017_3568915830-61078-018 the year of birth
#for person #1 was 1988. No sufficient information to clarify.
t <- data_v11 %>% filter(ID_long == "3568915830-61078-017_3568915830-61078-018")
data_v11$p01_year[which(data_v11$ID_long == "3568915830-61078-017_3568915830-61078-018")] <- NA

#For record with ID 3687346103-61084-021_3687346103-61084-022 the year of birth
#for person 2 was 1848.
```

```
t <- data_v11 %>% filter(ID_long == "3687346103-61084-021_3687346103-61084-022")
data_v11$p02_year[which(data_v11$ID_long == "3687346103-61084-021_3687346103-61084-022")] <- "1898"

#For record with ID 3687352813-61090-027_3687352813-61090-028 the year of birth
#for person #8 was 1057.
t <- data_v11 %>% filter(ID_long == "3687352813-61090-027_3687352813-61090-028")
data_v11$p02_year[which(data_v11$ID_long == "3687352813-61090-027_3687352813-61090-028")] <- "1957"

#For record with ID 3687356601-61093-009_3687356601-61093-010 the year of birth
#for person #1 was 1968.
t <- data_v11 %>% filter(ID_long == "3687356601-61093-009_3687356601-61093-010")
data_v11$p01_year[which(data_v11$ID_long == "3687356601-61093-009_3687356601-61093-010")] <- "1890"
#In reviewing the data imputed we see that the year of birth is also incorrect
#although it falls within the margins we had deemed ok.
data_v11$p02_year[which(data_v11$ID_long == "3687356601-61093-009_3687356601-61093-010")] <- "1900"

#For record with ID 3687357055-61094-009_3687357055-61094-010 the year of birth
#for person #4 was 1959.
t <- data_v11 %>% filter(ID_long == "3687357055-61094-009_3687357055-61094-010")
data_v11$p04_year[which(data_v11$ID_long == "3687357055-61094-009_3687357055-61094-010")] <- "1957"

#For record with ID 3687361616-61103-035_3687361616-61103-036 the year of birth
#for person #5 was 1959.
t <- data_v11 %>% filter(ID_long == "3687361616-61103-035_3687361616-61103-036")
data_v11$p05_year[which(data_v11$ID_long == "3687361616-61103-035_3687361616-61103-036")] <- "1957"
```

Because Year of Birth is a numeric variable, we will now convert them all to numeric.

```
#We will now convert the year of birth variables to numeric.
data_v11 <- data_v11 %>%
  mutate_at(c("p01_year","p02_year","p03_year","p04_year","p05_year",
              "p06_year","p07_year","p08_year","p09_year","p010_year",
              "p011_year","p012_year"), as.numeric)
```

Weekly income

We will now look at the responses to the weekly income question. Responses containing alphabetical characters may be indicative of mistakes in the data collection, i.e. typing place of birth in the income section, or may be indicative of a distinct response given for income such as h/wife or school.

```
inc_wk <- data_v11 %>%
  select(ID_long,p01_incwk,p02_incwk,p03_incwk,p04_incwk,p05_incwk,
         p06_incwk,p07_incwk,p08_incwk,p09_incwk,p010_incwk,p011_incwk,p012_incwk)
#Convert data frame to long format to facilitate analysis.
inc_wk_long <- pivot_longer(inc_wk,
                            cols = starts_with("p0"),
                            names_to = "var",
                            values_to = "Inc",
                            values_drop_na = T)
```

Income should be, for the most part, numeric values. We will subset our data frame for cases that contain letters. In some instances the imputed response that contains letters corresponds with the information in the original record and requires no change.

```
inc_wk_wrng <- inc_wk_long %>%
  filter(grepl("[A-Za-z]",Inc))
```

```
#For record with ID 3105195350-60963-003_3105195350-60963-004 the weekly income
#was listed as 98.00 from social security but social security is a monthly income.
data_v11$p01_incwk[which(data_v11$ID_long == "3105195350-60963-003_3105195350-60963-004")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3105195350-60963-003_3105195350-60963-004")] <- "98.00 s.s

#For record with ID 3105260901-60973-021_3105260901-60973-022 the weekly income
#was listed as 52.50 from social security but social security is a monthly income.
data_v11$p01_incwk[which(data_v11$ID_long == "3105260901-60973-021_3105260901-60973-022")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3105260901-60973-021_3105260901-60973-022")] <- "52.50 so

#For record with ID 3105298148-60980-009_3105298148-60980-010 the weekly income
#was listed as PR.
data_v11$p02_plc[which(data_v11$ID_long == "3105298148-60980-009_3105298148-60980-010")] <- "PR"
data_v11$p02_incwk[which(data_v11$ID_long == "3105298148-60980-009_3105298148-60980-010")] <- NA

#For record with ID 3105300183-60981-011_3105300183-60981-012 the weekly income
#was listed as 103.95 from social security but social security is a monthly income.
data_v11$p01_incwk[which(data_v11$ID_long == "3105300183-60981-011_3105300183-60981-012")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3105300183-60981-011_3105300183-60981-012")] <- "103.95 s

#For record with ID 3262908762-60101-011_3262908762-60101-012 information for
#person #4 was imputed in income weekly and monthly for person #3.
t <- data_v11 %>% filter(ID_long == "3262908762-60101-011_3262908762-60101-012")
data_v11$p04_year[which(data_v11$ID_long == "3262908762-60101-011_3262908762-60101-012")] <- "1952"
data_v11$p04_rel[which(data_v11$ID_long == "3262908762-60101-011_3262908762-60101-012")] <- "D"
data_v11$p04_plc[which(data_v11$ID_long == "3262908762-60101-011_3262908762-60101-012")] <- "PR"
data_v11$p03_incwk[which(data_v11$ID_long == "3262908762-60101-011_3262908762-60101-012")] <- NA
data_v11$p03_incmo[which(data_v11$ID_long == "3262908762-60101-011_3262908762-60101-012")] <- NA
```

From observing some of the responses given, we have also noticed cases where multiple sources of income are listed or cases where income is listed as semi or bi weekly. These are fixed bellow to represent weekly income.

```
data_v11$p01_incwk[which(data_v11$ID_long == "3458105558-61030-033_3458105558-61030-034")] <- "37.50"
data_v11$p01_incwk[which(data_v11$ID_long == "3458109563-61046-019_3458109563-61046-020")] <- "49.00"
data_v11$p01_incwk[which(data_v11$ID_long == "3568895561-61054-031_3568895561-61054-032")] <- "48.00"
data_v11$p04_incwk[which(data_v11$ID_long == "3568904271-61062-013_3568904271-61062-014")] <- "20.67"
```

Something else we can do to check if the ranges of weekly income and determine if there are any anomalies, is convert the variables to numeric and compare. The median monthly income for a household in the US in 1958 was $425.00. Let's call inc_wk again, with the changes introduced before.

```
inc_wk <- data_v11 %>%
  select(ID_long,p01_incwk,p02_incwk,p03_incwk,p04_incwk,p05_incwk,
         p06_incwk,p07_incwk,p08_incwk,p09_incwk,p010_incwk,p011_incwk,p012_incwk)
#Convert data frame to long format to facilitate analysis.
inc_wk_long <- pivot_longer(inc_wk,
                            cols = starts_with("p0"),
                            names_to = "var",
```

```
                            values_to = "Inc",
                            values_drop_na = T)
inc_wk_long$Inc <- as.numeric(inc_wk_long$Inc)
```

```
## Warning: NAs introduced by coercion
```

```
summary(inc_wk_long)
```

```
##    ID_long              var                 Inc
##  Length:1394        Length:1394        Min.   :   5.00
##  Class :character   Class :character   1st Qu.:  45.00
##  Mode  :character   Mode  :character   Median :  51.00
##                                        Mean   :  60.44
##                                        3rd Qu.:  65.00
##                                        Max.   :2000.00
##                                        NA's   :155
```

The summary suggests there are extremely low values ($5) and extremely high #values ($2000).

```
#Some responses where missing digits.

#For record with ID 3687362500-61105-019_3687362500-61105-020 weekly income for
#person #2 was 5.
data_v11$p02_incwk[which(data_v11$ID_long == "3687362500-61105-019_3687362500-61105-020")] <- "55.00"
#Some were information recorded in the wrong place.
data_v11$p04_incwk[which(data_v11$ID_long == "3262914572-61012-005_3262914572-61012-006")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3568912231-61070-003_3568912231-61070-004")] <- "20.00"
t <- data_v11 %>% filter(ID_long == "3105154754-60958-019_3105154754-60958-020")
data_v11$p03_incwk[which(data_v11$ID_long == "3105154754-60958-019_3105154754-60958-020")] <- NA
t <- data_v11 %>% filter(ID_long == "3105302993-60982-009_3105302993-60982-010")
data_v11$p02_incwk[which(data_v11$ID_long == "3105302993-60982-009_3105302993-60982-010")] <- NA

#Some cases had monthly information recorded in the income weekly box.
#All these cases where checked with original record and confirmed that the income #should have been lis
data_v11$p01_incwk[which(data_v11$ID_long == "3104718669-60949-001_3104718669-60949-002")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3104718669-60949-001_3104718669-60949-002")] <- "248.00"

data_v11$p04_incwk[which(data_v11$ID_long == "3568906814-61064-017_3568906814-61064-018")] <- NA
data_v11$p04_incmo[which(data_v11$ID_long == "3568906814-61064-017_3568906814-61064-018")] <- "200.00"

data_v11$p01_incwk[which(data_v11$ID_long == "3262913099-61007-019_3262913099-61007-020")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3262913099-61007-019_3262913099-61007-020")] <- "162.50"

data_v11$p01_incwk[which(data_v11$ID_long == "3262913099-61007-011_3262913099-61007-012")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3262913099-61007-011_3262913099-61007-012")] <- "145.00"

data_v11$p01_incwk[which(data_v11$ID_long == "3568900074-61057-001_3568900074-61057-002")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3568900074-61057-001_3568900074-61057-002")] <- "130.00"

data_v11$p01_incwk[which(data_v11$ID_long == "3105211839-60965-023_3105211839-60965-024")] <- NA
data_v11$p01_incmo[which(data_v11$ID_long == "3105211839-60965-023_3105211839-60965-024")] <- "124.00"
```

Monthly Income

We will now look at monthly income. Like weekly income, monthly income should be for the most part numeric values. Responses containing alphabetical characters may contain similar responses to the weekly income or may be suggesting that income came from social security or welfare.

```
inc_mo <- data_v11 %>%
  select(ID_long,p01_incmo,p02_incmo,p03_incmo,p04_incmo,p05_incmo,
         p06_incmo,p07_incmo,p08_incmo,p09_incmo,p010_incmo,p011_incmo,p012_incmo)
#Convert data frame to long format to facilitate analysis.
inc_mo_long <- pivot_longer(inc_mo,
                            cols = starts_with("p0"),
                            names_to = "var",
                            values_to = "Inc",
                            values_drop_na = T)
```

An analysis of all responses to monthly income suggests that a number of cases were imputed as monthly income when they should have been weekly. Others were corrected to include welfare or social security if it was not initially stipulated that income was derived from welfare or social security.

```
data_v11$p01_incmo[which(data_v11$ID_long == "3104718669-60949-009_3104718669-60949-010")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3104718669-60949-009_3104718669-60949-010")] <- "110.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3104734180-60950-003_3104734180-60950-004")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3104734180-60950-003_3104734180-60950-004")] <- "100.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3104744331-60951-005_3104744331-60951-006")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3104744331-60951-005_3104744331-60951-006")] <- "80.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3104750306-60952-005_3104750306-60952-006")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3104750306-60952-005_3104750306-60952-006")] <- "96.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3104750306-60952-007_3104750306-60952-008")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3104750306-60952-007_3104750306-60952-008")] <- "90.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3104750306-60952-015_3104750306-60952-016")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3104750306-60952-015_3104750306-60952-016")] <- "87.61"

data_v11$p01_incmo[which(data_v11$ID_long == "3105040525-60953-005_3105040525-60953-006")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105040525-60953-005_3105040525-60953-006")] <- "70.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3105040525-60953-009_3105040525-60953-010")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105040525-60953-009_3105040525-60953-010")] <- "70.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3105078105-60955-017_3105078105-60955-018")] <- NA
data_v11$p07_incmo[which(data_v11$ID_long == "3105078105-60955-017_3105078105-60955-018")] <- NA
data_v11$p07_incwk[which(data_v11$ID_long == "3105078105-60955-017_3105078105-60955-018")] <- "80.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3105169069-60961-005_3105169069-60961-006")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105169069-60961-005_3105169069-60961-006")] <- "141.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3105195350-60963-023_3105195350-60963-024")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105195350-60963-023_3105195350-60963-024")] <- "146.00"
```

```
data_v11$p01_incmo[which(data_v11$ID_long == "3105200949-60964-007_3105200949-60964-008")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105200949-60964-007_3105200949-60964-008")] <- "91.90"

data_v11$p01_incmo[which(data_v11$ID_long == "3105221470-60967-007_3105221470-60967-008")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105221470-60967-007_3105221470-60967-008")] <- "78.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3105221470-60967-009_3105221470-60967-010")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105221470-60967-009_3105221470-60967-010")] <- "68.00"

data_v11$p03_incmo[which(data_v11$ID_long == "3105221470-60967-009_3105221470-60967-010")] <- NA
data_v11$p03_incwk[which(data_v11$ID_long == "3105221470-60967-009_3105221470-60967-010")] <- "80.00"

data_v11$p03_incmo[which(data_v11$ID_long == "3105298148-60980-007_3105298148-60980-008")] <- NA

data_v11$p01_incmo[which(data_v11$ID_long == "3105221470-60967-009_3105221470-60967-010")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3105221470-60967-009_3105221470-60967-010")] <- "60.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3262897985-60984-011_3262897985-60984-012")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3262897985-60984-011_3262897985-60984-012")] <- "90.00"

data_v11$p02_incmo[which(data_v11$ID_long == "3262897985-60984-015_3262897985-60984-016")] <- NA
data_v11$p02_incwk[which(data_v11$ID_long == "3262897985-60984-015_3262897985-60984-016")] <- "40.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3262897985-60984-017_3262897985-60984-018")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3262897985-60984-017_3262897985-60984-018")] <- "49.00"

data_v11$p04_incmo[which(data_v11$ID_long == "3262901080-60985-011_3262901080-60985-012")] <- NA
data_v11$p04_incwk[which(data_v11$ID_long == "3262901080-60985-011_3262901080-60985-012")] <- "65.00"

data_v11$p05_incmo[which(data_v11$ID_long == "3262901080-60985-011_3262901080-60985-012")] <- NA
data_v11$p05_incwk[which(data_v11$ID_long == "3262901080-60985-011_3262901080-60985-012")] <- "45.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3262901080-60985-017_3262901080-60985-018")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3262901080-60985-017_3262901080-60985-018")] <- "49.00"

data_v11$p02_incmo[which(data_v11$ID_long == "3262901080-60985-021_3262901080-60985-022")] <- "255.5 wel

data_v11$p01_incmo[which(data_v11$ID_long == "3262901080-60985-029_3262901080-60985-030")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3262901080-60985-029_3262901080-60985-030")] <- "49.00"
data_v11$p06_incmo[which(data_v11$ID_long == "3262901080-60985-029_3262901080-60985-030")] <- NA
data_v11$p06_incwk[which(data_v11$ID_long == "3262901080-60985-029_3262901080-60985-030")] <- "40.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3262901378-60986-019_3262901378-60986-020")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3262901378-60986-019_3262901378-60986-020")] <- "52.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3262902338-60989-007_3262902338-60989-008")] <- "197.60"

data_v11$p01_incmo[which(data_v11$ID_long == "3262902338-60989-011_3262902338-60989-012")] <- "181.20"

data_v11$p01_incmo[which(data_v11$ID_long == "3262902861-60991-021_3262902861-60991-022")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3262902861-60991-021_3262902861-60991-022")] <- "82.00"

data_v11$p04_incmo[which(data_v11$ID_long == "3262902861-60991-021_3262902861-60991-022")] <- NA
```

```
data_v11$p04_incwk[which(data_v11$ID_long == "3262902861-60991-021_3262902861-60991-022")] <- "47.00"

data_v11$p04_incmo[which(data_v11$ID_long == "3262921422-61017-015_3262921422-61017-016")] <- NA
data_v11$p04_incwk[which(data_v11$ID_long == "3262921422-61017-015_3262921422-61017-016")] <- "125.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3458098649-61037-009_3458098649-61037-010")] <- "181.80 w

data_v11$p01_incmo[which(data_v11$ID_long == "3458102587-61042-007_3458102587-61042-008")] <- "113.00 w

data_v11$p01_incmo[which(data_v11$ID_long == "3458103921-61031-035_3458103921-61031-036")] <- "80.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3458109121-61045-003_3458109121-61045-004")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3458109121-61045-003_3458109121-61045-004")] <- "50.00"
data_v11$p02_incmo[which(data_v11$ID_long == "3458109121-61045-003_3458109121-61045-004")] <- NA
data_v11$p02_incwk[which(data_v11$ID_long == "3458109121-61045-003_3458109121-61045-004")] <- "40.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3458109121-61045-011_3458109121-61045-012")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3458109121-61045-011_3458109121-61045-012")] <- "56.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3458109563-61046-013_3458109563-61046-014")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3458109563-61046-013_3458109563-61046-014")] <- "75.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3458110245-61048-007_3458110245-61048-008")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3458110245-61048-007_3458110245-61048-008")] <- "70.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3458110245-61048-009_3458110245-61048-010")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3458110245-61048-009_3458110245-61048-010")] <- "70.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3568901018-61058-017_3568901018-61058-018")] <- NA

t <- data_v11 %>% filter(ID_long == "3568902147-61059-019_3568902147-61059-020")
data_v11$p04_incmo[which(data_v11$ID_long == "3568902147-61059-019_3568902147-61059-020")] <- NA

data_v11$p01_incmo[which(data_v11$ID_long == "3568902147-61059-027_3568902147-61059-028")] <- "213.50 d

t <- data_v11 %>% filter(ID_long == "3568902147-61059-037_3568902147-61059-038")
data_v11$p01_incmo[which(data_v11$ID_long == "3568902147-61059-037_3568902147-61059-038")] <- NA

data_v11$p01_incmo[which(data_v11$ID_long == "3568905752-61063-017_3568905752-61063-018")] <- NA

data_v11$p01_incmo[which(data_v11$ID_long == "3568909924-61068-001_3568909924-61068-002")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3568909924-61068-001_3568909924-61068-002")] <- "40.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3687350085-61088-015_3687350085-61088-016")] <- "54.00"

data_v11$p01_incmo[which(data_v11$ID_long == "3687354473-61091-001_3687354473-61091-002")] <- "60.00 d/

data_v11$p01_incmo[which(data_v11$ID_long == "3687361245-61102-029_3687361245-61102-030")] <- NA
data_v11$p01_incwk[which(data_v11$ID_long == "3687361245-61102-029_3687361245-61102-030")] <- "110.00"
```

Annual Income

We can now look at the responses to annual income.

```
inc_an <- data_v11 %>%
  select(ID_long,p01_incan,p02_incan,p03_incan,p04_incan,p05_incan,
         p06_incan,p07_incan,p08_incan,p09_incan,p010_incan,p011_incan,p012_incan)
#Convert data frame to long format to facilitate analysis.
inc_an_long <- pivot_longer(inc_an,
                            cols = starts_with("p0"),
                            names_to = "var",
                            values_to = "Inc",
                            values_drop_na = T)
```

Responses to annual income were visually reviewed. A large number of responses contain alphabetical characters because they indicate welfare, social security, or other responses that, although they are not numeric, are in fact representative of the information provided in the relocation cards. Only three cases were identified where changes needed to be made. These are all in respect to the information provided not matching a general perception of what annual income should be. All cases where the response for annual income was bellow 2,000 were also checked with the original records. Save for one, all others were correct.

```
t <- data_v11 %>% filter(ID_long == "3105260901-60973-021_3105260901-60973-022")
data_v11$p01_incan[which(data_v11$ID_long == "3105260901-60973-021_3105260901-60973-022")] <- NA
data_v11$p02_rel[which(data_v11$ID_long == "3105260901-60973-021_3105260901-60973-022")] <- "Roomer"
data_v11$p01_year[which(data_v11$ID_long == "3105260901-60973-021_3105260901-60973-022")] <- "1887"

data_v11$p01_incan[which(data_v11$ID_long == "3150219515-60966-007_3150219515-60966-008")] <- "4000.00"

data_v11$p02_incan[which(data_v11$ID_long == "3568897909-61056-019_3568897909-61056-020")] <- NA
data_v11$p02_incmo[which(data_v11$ID_long == "3568897909-61056-019_3568897909-61056-020")] <- "75.90 s.
```

This concludes the data cleaning portion for the Family Composition variables.

Following we have a set of scripts for cleaning and recoding other variables.

Recoding place of birth

We will now re-code the Place of Birth category to one of 9 major groups. These groups are meant to represent important locations (US or PR) or have continental boundaries.

```
#Let's see all unique values for Place of Birth
plc_birth <- data_v11 %>%
  select(ID_long,p01_plc_2:p012_plc_2)

plc_birth_long <- pivot_longer(plc_birth,
                           cols = starts_with("p0"),
                           names_to = "var",
                           values_to = "Place",
                           values_drop_na = F)

unique(plc_birth_long$Place)
```

```
##  [1] "PR"            "USA"           NA              ""
##  [5] "Italy"         "Japan"         "China"         "Cuba"
##  [9] "Yugoslavia"    "Morta"         "Ireland"       "Poland"
## [13] "Belo"          "Spain"         "Greece"        "Ukraine"
## [17] "Germany"       "Scotland"      "DR"            "Sweden"
## [21] "Hungary"       "Russia"        "SouthAmerica"  "England"
## [25] "France"        "Norway"        "Canada"        "Peru"
## [29] "Armenia"       "Austria"       "Europe"        "Estonia"
## [33] "Morocco"       "Gibraltar"     "I"             "WestInd"
## [37] "Cyprus"        "Phillipines"   "Romania"       "Holland"
## [41] "Trinidad"      "Malta"         "Belgium"       "Mexico"
## [45] "Turkey"        "Ecuador"       "C.Amer."       "Denmark"
## [49] "CostaRica"     "Luxembourg"    "Nicaragua"     "HongKong"
## [53] "Colombia"      "Philipines"    "Haiti"         "Virgin Islands"
## [57] "St.Thomas"     "Switzerland"   "BW"            "BritishGuiana"
## [61] "Bali"          "SouthAfrica"   "India"         "Pakistan"
## [65] "Martinique"    "Columbia"      "Panama"        "Bulgaria"
## [69] "Latvia"        "Brazil"
```

We see that the records have been standardized with a few exceptions were it was not clear what the response referred too. We will now recode these answers to provide a more coherent set of places of birth.

```
#We are creating a new variable in our long data set that will contain the recoded places of birth. The
plc_birth_long$Plc_rec <- NA
plc_birth_long$Plc_rec[plc_birth_long$Place == "USA"] <- "USA"
plc_birth_long$Plc_rec[plc_birth_long$Place == "PR"] <- "PR"
plc_birth_long$Plc_rec[plc_birth_long$Place == "USA"] <- "USA"
plc_birth_long$Plc_rec[plc_birth_long$Place == "Italy" |
                         plc_birth_long$Place == "Yugoslavia" |
                       plc_birth_long$Place == "Ireland" |
                       plc_birth_long$Place == "Poland" |
                       plc_birth_long$Place == "Belo" |
                       plc_birth_long$Place == "Spain" |
                       plc_birth_long$Place == "Greece" |
                       plc_birth_long$Place == "Ukraine" |
                       plc_birth_long$Place == "Germany" |
                       plc_birth_long$Place == "Scotland" |
                       plc_birth_long$Place == "Scotland" |
                       plc_birth_long$Place == "Scotland" |
                       plc_birth_long$Place == "Scotland" |
                       plc_birth_long$Place == "Armenia" |
                       plc_birth_long$Place == "Austria" |
                       plc_birth_long$Place == "Europe" |
                       plc_birth_long$Place == "Sweeden" |
                       plc_birth_long$Place == "Hungary" |
                       plc_birth_long$Place == "Russia" |
                       plc_birth_long$Place == "England"|
                       plc_birth_long$Place == "France" |
                       plc_birth_long$Place == "Norway" |
                       plc_birth_long$Place == "Estonia"|
                       plc_birth_long$Place == "Gibraltar"|
                       plc_birth_long$Place == "Cyprus"|
                       plc_birth_long$Place == "Romania"|
                       plc_birth_long$Place == "Holland"|
```

```
                     plc_birth_long$Place == "Malta"|
                     plc_birth_long$Place == "Belgium"|
                     plc_birth_long$Place == "Turkey"|
                     plc_birth_long$Place == "Denmark"|
                     plc_birth_long$Place == "Luxemburg"|
                     plc_birth_long$Place == "Switzerland"|
                     plc_birth_long$Place == "Bulgaria"|
                     plc_birth_long$Place == "Latvia"] <- "Europe"

plc_birth_long$Plc_rec[plc_birth_long$Place == "Peru" |
                          plc_birth_long$Place == "Ecuador" |
                     plc_birth_long$Place == "C.Amer." |
                     plc_birth_long$Place == "CostaRica" |
                     plc_birth_long$Place == "Nicaragua" |
                     plc_birth_long$Place == "Colombia" |
                     plc_birth_long$Place == "Panama" |
                     plc_birth_long$Place == "Brazil" |
                     plc_birth_long$Place == "Columbia" |
                     plc_birth_long$Place == "SouthAmerica"] <- "Lat. Am."

plc_birth_long$Plc_rec[plc_birth_long$Place == "Cuba" |
                          plc_birth_long$Place == "DR" |
                     plc_birth_long$Place == "Trinidad" |
                     plc_birth_long$Place == "WestInd" |
                     plc_birth_long$Place == "Haiti" |
                     plc_birth_long$Place == "Virgin Islands" |
                     plc_birth_long$Place == "St.Thomas" |
                     plc_birth_long$Place == "BritishGuiana" |
                     plc_birth_long$Place == "BW" |
                     plc_birth_long$Place == "Martinique"] <- "Caribbean"

plc_birth_long$Plc_rec[plc_birth_long$Place == "Canada" |
                          plc_birth_long$Place == "Mexico"] <- "Nrth. Am."

plc_birth_long$Plc_rec[plc_birth_long$Place == "Japan" |
                          plc_birth_long$Place == "China" |
                     plc_birth_long$Place == "HongKong" |
                     plc_birth_long$Place == "Bali" |
                     plc_birth_long$Place == "India" |
                     plc_birth_long$Place == "Pakistan" |
                     plc_birth_long$Place == "Phillipines"] <- "Asia"

plc_birth_long$Plc_rec[plc_birth_long$Place == "SouthAfrica"] <- "Africa"
```

Now that we have recoded the Place of birth variable we will rename the var variable to highlight that these are recoded.

```
plc_birth_long$var <- gsub('plc','plcrec',plc_birth_long$var)
plc_birth_long <- plc_birth_long %>% select(-Place)
```

We can now convert place of birth from long format to wide format and join to data_v11

```
plc_birth_wide <- plc_birth_long %>%
                    pivot_wider(names_from = "var",
                                values_from = "Plc_rec")

data_v11 <- left_join(data_v11,plc_birth_wide,by = "ID_long")
```

We now have three sets of Place of birth Variables. These include the original place of birth as imputed, a standardized version with grammatical errors and others fixed, and a third version with a recode classifying places of birth in to brad categories.

Family Size

Family size is an important variable for our analysis. We will look into it's composition and determine if changes need to be made to the variable.

```
#First let's generate a data frame with only the relationship answers to ease
#our analysis.
#We will select from our data_v11 all family composition variables
fam_vars <- data_v11 %>% select(ID_long,fam_size,p01_plc_2:p012_plc_2,p01_rel,p02_rel,
                                p03_rel,p04_rel,p05_rel,p06_rel,p07_rel,p08_rel,
                                p09_rel,p010_rel,p011_rel,p012_rel,p01_year,p02_year,
                                p03_year,p04_year,p05_year,p06_year,p07_year,p08_year,
                                p09_year,p010_year,p011_year,p012_year,p01_incwk,
                                p02_incwk,p03_incwk,p04_incwk,p05_incwk,p06_incwk,
                                p07_incwk,p08_incwk,p09_incwk,p010_incwk,p011_incwk,
                                p012_incwk,p01_incmo,p02_incmo,p03_incmo,p04_incmo,
                                p05_incmo,p06_incmo,p07_incmo,p08_incmo,
                                p09_incmo,p010_incmo,p011_incmo,p012_incmo,p01_incan,
                                p02_incan,p03_incan,p04_incan,p05_incan,p06_incan,
                                p07_incan,p08_incan,p09_incan,p010_incan,p011_incan,
                                p012_incan)
#We will convert all cases where their is blank answer to NA.
fam_vars[fam_vars == ''] <- NA

fam_vars <- fam_vars %>% replace(is.na(.),999)
```

We will check that all instances where fam_size is Null (999), are actually instances where no persons where reported in the household.

```
fam_vars <- fam_vars %>%
  mutate(fam_na = if_all(p01_rel:p012_incan, ~.x == 999))
#fam_na is a binary variable indicating cases where our analysis suggests there are no #family members

#We are changing the position of the variable we just created so that we can visually compare to
#the family size variable
fam_vars <- fam_vars %>% relocate(fam_na, .after=fam_size)
```

Let's see how many instances there are of records where the fam_na is FALSE, meaning that at least one of the considered family variables has a response, and where fam_size is 999, which suggests that a mistake may have been made in data entry.

```
nrow(subset(fam_vars, fam_na == FALSE & fam_size == 999))
```

```
## [1] 13
```

We see that there are 13 instances where this occurred. By manually analyzing these instances we have determined 11 cases where we can #definitively say that at least one person was living in the household based on responses to the Family Relationship Variables.

```
fam_vars$fam_size2 <- fam_vars$fam_size
fam_vars <- arrange(fam_vars,fam_size)

fam_vars[1864,"fam_size2"] <- 1
fam_vars[1868,"fam_size2"] <- 2
fam_vars[1875,"fam_size2"] <- 1
fam_vars[1892,"fam_size2"] <- 4
fam_vars[1913,"fam_size2"] <- 1
fam_vars[1944,"fam_size2"] <- 1
fam_vars[1945,"fam_size2"] <- 1
fam_vars[1950,"fam_size2"] <- 1
fam_vars[1951,"fam_size2"] <- 1
fam_vars[1961,"fam_size2"] <- 1
fam_vars[2096,"fam_size2"] <- 1
```

Here we are joining the new family size variable to the original data frame.

```
fam_vars2 <- fam_vars %>% select(ID_long,fam_size2)
data_v11 <- data_v11 %>%
  left_join(fam_vars2, by = "ID_long")
```

Race

Race is a key variable in our analysis of the population living in the area. Experience from data entry suggests that race can be recoded to simplify analysis.

```
race <- data_v11 %>% select(ID_long,race)
race$race <- clean_strings(race$race)

unique(race$race)
```

```
##  [1] "puerto rican"                    NA
##  [3] "white"                           "oriental"
##  [5] "white puerto rican"              "cuba"
##  [7] "hawaiian"                        "cuban"
##  [9] "negro"                           "dominican"
## [11] "white spanish"                   "phillipines"
## [13] "puerto rican dominican"          "white negro oriental puerto rican"
## [15] "costa rica"                      "colombian"
## [17] "haiti"                           "negro puerto rican"
## [19] "hawaii"                          "dom rep"
## [21] "white oriental"                  "mexico"
```

```
## [23] "negro oriental"                    "ecuador"
## [25] "greece"                            "puerto rican columbia"
## [27] "white negro"                       "spain"
## [29] "negro spanish"                     "panama"
```

Besides the 4 categories included in the relocation cards, we have responses that include combined racial categories and some indicate specific national origins. In cases where multiple racial categories are selected we will categorize them as "Multi", except when the racial mic includes Puerto Ricans. In that case they will be categorized as "Puerto Rican mix". Cases that include all racial categories are considered NA and cases that include specific origin are considered Others.

```
#We will create a new variable where race is recoded.
race$race2 <- case_when(
  race$race == "puerto rican" ~ "Puerto Rican",
  race$race == "white" ~ "White",
  race$race == "negro" ~ "Negro",
  race$race == "oriental" ~ "Oriental",
  race$race == "white puerto rican" ~ "Puerto Rican mix",
  race$race == "puerto rican dominican" ~ "Puerto Rican mix",
  race$race == "negro puerto rican" ~ "Puerto Rican mix",
  race$race == "white oriental" ~ "Multi",
  race$race == "negro oriental" ~ "Multi",
  race$race == "white negro" ~ "Multi",
  race$race == "puerto rican columbia" ~ "Puerto Rican mix",
  race$race == "negro spanish" ~ "Multi",
  race$race == "white spanish" ~ "White",
  race$race == "white negro oriental puerto rican" ~ NA,
  is.na(race$race) ~ NA,
  .default = "other"
)
```

We can now join this race2 variable to data_v11.

```
race <- race %>% select(-race)
data_v11 <- left_join(data_v11,race,by="ID_long")
```