processRawData.R

Wed Mar 7 14:29:55 2018

```
# Introduction ------------------------------------------------------------

# This script will read in the raw data from the NLSY97 based on extracts from
# NLS Explorer. Part of this process will be linking the roster data which
# contains information on parental race with individual records. The final data
# will be saved as a CSV/RData.


# Read in the Data --------------------------------------------------------
demog <- read.csv("input/demographic/demographic.csv")
roster <- read.csv("input/roster/roster.csv")

#how many cases are missing due to non-interview in 2002?
sum(demog$S1531300==-5)
```

```
## [1] 1088
```

```
#remove all of these cases
demog <- subset(demog, S1531300!=-5)

# Code Demographic Variables ----------------------------------------------
demog$id <- demog$R0000100

# GENDER
demog$gender <- factor(demog$R0536300, levels=c(1,2),
                       labels=c("Male","Female"))
table(demog$R0536300, demog$gender, exclude=NULL)
```

```
##
##     Male Female
##   1 3997      0
##   2    0   3899
```

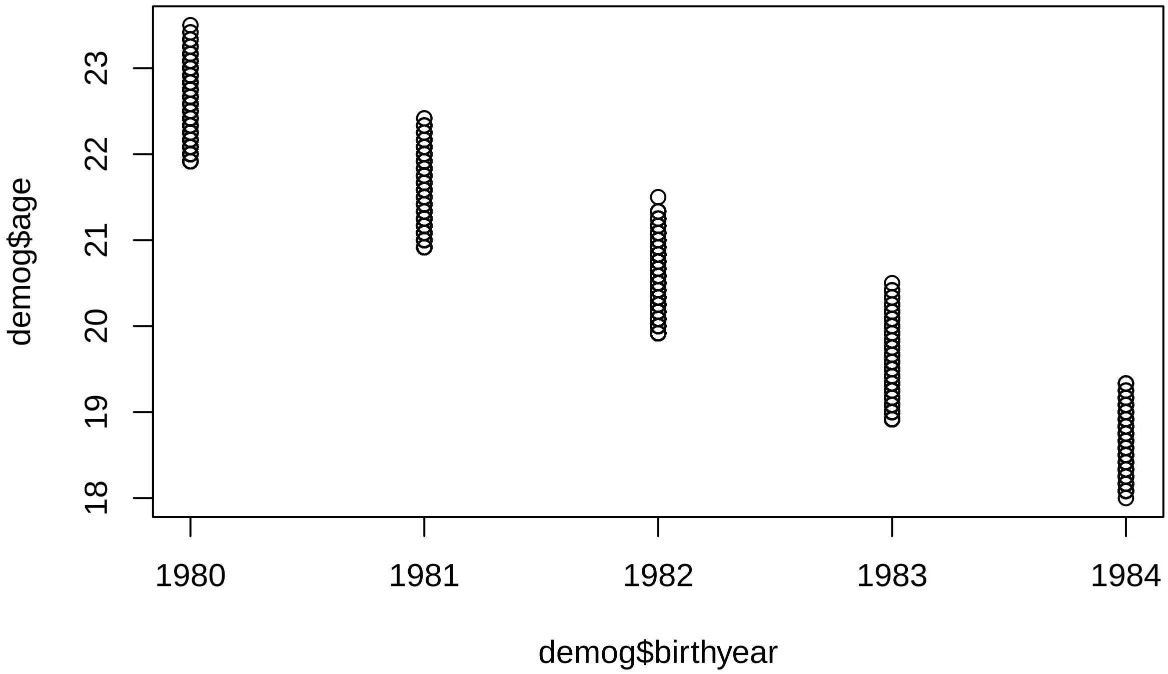
```
# BIRTH COHORT/AGE
demog$birthyear <- demog$R0536402
summary(demog$birthyear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1980    1981    1982    1982    1983    1984
```

```
#age is recorded in months
demog$age <- ifelse(demog$S1531300==-5, NA, demog$S1531300/12)
summary(demog$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   19.25   20.42   20.46   21.67   23.50
```

```
plot(demog$birthyear, demog$age)
```



```
# HOUSEHOLD STRUCTURE
demog$family <- factor(demog$R1205300,
                       levels=1:10,
                       labels=c("Two bio p","Two p, bio mom","Two p, bio dad",
                                "Bio mom", "Bio dad","Adoptive","Foster",
                                "Grandparents","Other relatives",
                                "Something else"))
table(demog$R1205300, demog$family, exclude=NULL)
```

```
##
##      Two bio p Two p, bio mom Two p, bio dad Bio mom Bio dad Adoptive
##   -3         0              0              0       0       0        0
##   1       3866              0              0       0       0        0
##   2          0            881              0       0       0        0
##   3          0              0            183       0       0        0
##   4          0              0              0    2243       0        0
##   5          0              0              0       0     256        0
##   6          0              0              0       0       0       89
##   7          0              0              0       0       0        0
##   8          0              0              0       0       0        0
##   9          0              0              0       0       0        0
##   10         0              0              0       0       0        0
##
##      Foster Grandparents Other relatives Something else <NA>
##   -3      0            0               0              0   27
##   1       0            0               0              0    0
##   2       0            0               0              0    0
##   3       0            0               0              0    0
##   4       0            0               0              0    0
##   5       0            0               0              0    0
##   6       0            0               0              0    0
##   7      33            0               0              0    0
##   8       0          170               0              0    0
##   9       0            0              93              0    0
##   10      0            0               0             55    0
```

```
#compare this to HH structure in 2002
demog$hh2002 <- factor(demog$S1542000,
                       levels=1:10,
                       labels=c("Two bio p","Two p, bio mom","Two p, bio dad",
                                "Bio mom", "Bio dad","Adoptive","Foster",
                                "Grandparents","Other relatives",
                                "Something else"))
table(demog$S1542000, demog$hh2002, exclude=NULL)
```

```
##
##      Two bio p Two p, bio mom Two p, bio dad Bio mom Bio dad Adoptive
##   1       2677              0              0       0       0        0
##   2          0            593              0       0       0        0
##   3          0              0            120       0       0        0
##   4          0              0              0    1553       0        0
##   5          0              0              0       0     246        0
##   6          0              0              0       0       0       32
##   7          0              0              0       0       0        0
##   8          0              0              0       0       0        0
##   9          0              0              0       0       0        0
##   10         0              0              0       0       0        0
##
##      Foster Grandparents Other relatives Something else
##   1       0            0               0              0
##   2       0            0               0              0
##   3       0            0               0              0
##   4       0            0               0              0
##   5       0            0               0              0
##   6       0            0               0              0
##   7       9            0               0              0
##   8       0           50               0              0
##   9       0            0              15              0
##   10      0            0               0           2601
```

```
table(demog$family, demog$hh2002)
```

```
##
##                   Two bio p Two p, bio mom Two p, bio dad Bio mom Bio dad
##   Two bio p            2520             26              9     259      90
##   Two p, bio mom         16            361             15     115      17
##   Two p, bio dad          5             18             45      21      17
##   Bio mom                91            163             18    1071      43
##   Bio dad                 9              8             26      29      65
##   Adoptive               16              0              1       7       0
##   Foster                  1              0              0       3       0
##   Grandparents            4              6              1      24       3
##   Other relatives         4              5              0      11       6
##   Something else          4              4              4       6       3
##
##                   Adoptive Foster Grandparents Other relatives
##   Two bio p              0      0            0               0
##   Two p, bio mom         2      0            0               0
##   Two p, bio dad         0      0            0               0
##   Bio mom                0      1            0               0
##   Bio dad                0      0            0               0
##   Adoptive              29      0            0               0
##   Foster                 0      7            0               0
##   Grandparents           1      1           50               0
##   Other relatives        0      0            0              14
##   Something else         0      0            0               0
##
##                    Something else
##   Two bio p                   962
##   Two p, bio mom              355
##   Two p, bio dad               77
##   Bio mom                     856
##   Bio dad                     119
##   Adoptive                     36
##   Foster                       22
##   Grandparents                 80
##   Other relatives              53
##   Something else               34
```

```
demog$moved_out <- demog$family!="Something else" & demog$hh2002=="Something else"
summary(demog$moved_out)
```

```
##    Mode   FALSE    TRUE    NA's
## logical    5329    2560       7
```

```
# PARENTAL EDUCATION
demog$biodaded <- ifelse(demog$R1302400<0 | demog$R1302400>20,NA,demog$R1302400)
demog$biomomed <- ifelse(demog$R1302500<0 | demog$R1302500>20,NA,demog$R1302500)
demog$resdaded <- ifelse(demog$R1302600<0 | demog$R1302600>20,NA,demog$R1302600)
demog$resmomed <- ifelse(demog$R1302700<0 | demog$R1302700>20,NA,demog$R1302700)
summary(demog[,c("biodaded","biomomed","resdaded","resmomed")])
```

```
##     biodaded        biomomed        resdaded        resmomed
##  Min.   : 2.00   Min.   : 1.00   Min.   : 2.00   Min.   : 1.00
##  1st Qu.:12.00   1st Qu.:12.00   1st Qu.:12.00   1st Qu.:12.00
##  Median :12.00   Median :12.00   Median :12.00   Median :12.00
##  Mean   :12.57   Mean   :12.47   Mean   :12.89   Mean   :12.55
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:15.00   3rd Qu.:14.00
##  Max.   :20.00   Max.   :20.00   Max.   :20.00   Max.   :20.00
##  NA's   :1624    NA's   :597     NA's   :2887    NA's   :839
```

```
#get highest parental education level - add a column of -4 values so I don't
#get warning when all are missing
demog$highparented <- apply(cbind(rep(-4,nrow(demog)),
                                  demog[,c("biodaded","biomomed",
                                           "resmomed","resdaded")]),
                            1,max,na.rm=TRUE)
demog$highparented[demog$highparented<0] <- NA
summary(demog$highparented)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   12.00   13.00   13.29   16.00   20.00     310
```

```
# HOUSEHOLD INCOME
#For household income we are going to make valid
```

```r
#negative values and zero the smallest non-neg number (5)
demog$hhinc <- ifelse(demog$R1204500>=-4 & demog$R1204500<0, NA,demog$R1204500)
demog$hhinc <- ifelse(demog$hhinc<=0, min(demog$hhinc[demog$hhinc>0]),
                      demog$hhinc)
summary(demog$hhinc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       5   18700   38000   46871   61750  246474    2115
```

```r
demog$hhnetworth <- ifelse(demog$R1204700>=-4 & demog$R1204700<0, NA,
                           demog$R1204700)
summary(demog$hhnetworth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -935251    5550   34500   89144  114080  600000    1992
```

```r
# URBANICITY
demog$urban97 <- factor(demog$R1217500, levels=0:1,
                        labels=c("Rural","Urban"))
table(demog$R1217500, demog$urban97, exclude=NULL)
```

```
## 
##      Rural Urban <NA>
##   0   1810     0    0
##   1      0  5758    0
##   2      0     0  328
```

```r
# REGION
demog$region97 <- factor(demog$R1200300, levels=1:4,
                         labels=c("Northeast","North Central","South", "West"))
table(demog$R1200300, demog$region97, exclude=NULL)
```

```
## 
##      Northeast North Central South West
##   1       1380             0     0    0
##   2          0          1794     0    0
##   3          0             0  2979    0
##   4          0             0     0 1743
```

```r
# MIGRATION BETWEEN 1997 and 2002
demog$migration <- factor(demog$S1530100, levels=c(-4, 1:4),
                         labels=c("Non-movers","Within county",
                                  "Different county","Different state",
                                  "Different country"))
table(demog$S1530100, demog$migration, exclude=NULL)
```

```
## 
##       Non-movers Within county Different county Different state
##   -4        6009             0                0               0
##   -3           0             0                0               0
##   1            0           341                0               0
##   2            0             0              846               0
##   3            0             0                0             598
##   4            0             0                0               0
## 
##       Different country <NA>
##   -4                  0    0
##   -3                 0   41
##   1                  0    0
##   2                  0    0
##   3                  0    0
##   4                 61    0
```

```
# ASVAB score - lets standardize it
demog$asvab <- scale(ifelse(demog$R9829600<0, NA, demog$R9829600))
summary(demog$asvab)
```

```
##        V1
##  Min.   :-1.5532
##  1st Qu.:-0.8905
##  Median :-0.0806
##  Mean   : 0.0000
##  3rd Qu.: 0.8522
##  Max.   : 1.8561
##  NA's   :1503
```

```
demog$gpa_overall <- ifelse(demog$R9871900<0, NA, demog$R9871900/100)
summary(demog$gpa_overall)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.100   2.440   2.870   2.829   3.270   4.110    2391
```

```
#ENROLLMENT
demog$enrollment02 <- factor(demog$S1538001, levels=1:11,
               labels=c("Not enrolled, no HS degree",
                        "Not enrolled, GED",
                        "Not enrolled, HS Degree",
                        "Not enrolled, some college",
                        "Not enrolled, 2-yr college grad",
                        "Not enrolled, 4-yr college grad",
                        "Not enrolled, grad degree",
                        "Enrolled in HS",
                        "Enrolled in 2-yr college",
                        "Enrolled in 4-yr college",
                        "Enrolled in grad program"))
table(demog$enrollment02, demog$S1538001, exclude=NULL)
```

```
##
##                                      -3    1    2    3    4    5    6    7
##   Not enrolled, no HS degree          0 1134    0    0    0    0    0    0
##   Not enrolled, GED                   0    0  467    0    0    0    0    0
##   Not enrolled, HS Degree             0    0    0 1805    0    0    0    0
##   Not enrolled, some college          0    0    0    0  912    0    0    0
##   Not enrolled, 2-yr college grad     0    0    0    0    0   56    0    0
##   Not enrolled, 4-yr college grad     0    0    0    0    0    0   74    0
##   Not enrolled, grad degree           0    0    0    0    0    0    0    1
##   Enrolled in HS                      0    0    0    0    0    0    0    0
##   Enrolled in 2-yr college            0    0    0    0    0    0    0    0
##   Enrolled in 4-yr college            0    0    0    0    0    0    0    0
##   Enrolled in grad program            0    0    0    0    0    0    0    0
##   <NA>                               18    0    0    0    0    0    0    0
##
##                                       8    9   10   11
##   Not enrolled, no HS degree         0    0    0    0
##   Not enrolled, GED                  0    0    0    0
##   Not enrolled, HS Degree            0    0    0    0
##   Not enrolled, some college         0    0    0    0
##   Not enrolled, 2-yr college grad    0    0    0    0
##   Not enrolled, 4-yr college grad    0    0    0    0
##   Not enrolled, grad degree          0    0    0    0
##   Enrolled in HS                   712    0    0    0
##   Enrolled in 2-yr college           0  825    0    0
##   Enrolled in 4-yr college           0    0 1866    0
##   Enrolled in grad program           0    0    0   26
##   <NA>                               0    0    0    0
```

```
#how many respondents are still in HS by age?
table(demog$enrollment02, floor(demog$age))
```

```
##
##                                     18  19  20  21  22  23
##   Not enrolled, no HS degree       205 248 264 221 174  22
##   Not enrolled, GED                 45  93 103 117  97  12
##   Not enrolled, HS Degree          266 400 402 367 332  38
##   Not enrolled, some college        25  92 192 280 285  38
##   Not enrolled, 2-yr college grad    0   1  13  16  24   2
##   Not enrolled, 4-yr college grad    0   0   0   4  57  13
##   Not enrolled, grad degree          0   0   0   0   1   0
##   Enrolled in HS                   569  95  27  12   9   0
##   Enrolled in 2-yr college         143 225 179 168 102   8
##   Enrolled in 4-yr college         295 451 438 407 257  18
##   Enrolled in grad program           0   0   0   1  24   1
```

```
# Code Multiple Race Responses --------------------------------------------

#Separate Yes/No variables for each race option. Turn into binaries and
#code missing values
white02 <- ifelse(demog$S1224900<0,NA,demog$S1224900)==1
black02 <- ifelse(demog$S1224901<0,NA,demog$S1224901)==1
#we can't distinguish Asians and PI on parental ID, so collapse them.
asian02 <- ifelse(demog$S1224902<0,NA,demog$S1224902)==1 |
  ifelse(demog$S1224903<0,NA,demog$S1224903)==1
indian02 <- ifelse(demog$S1224904<0,NA,demog$S1224904)==1
other02 <- ifelse(demog$S1224905<0,NA,demog$S1224905)==1
hispanic02 <- ifelse(demog$S1224906<0,NA,demog$S1224906)==1
summary(cbind(white02,black02,asian02,indian02,other02,hispanic02))
```

```
##   white02         black02         asian02         indian02
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:3225      FALSE:5737      FALSE:7690      FALSE:7776
##  TRUE :4661      TRUE :2149      TRUE :196       TRUE :110
##  NA's :10        NA's :10        NA's :10        NA's :10
##   other02        hispanic02
##  Mode :logical   Mode :logical
##  FALSE:7498      FALSE:6987
##  TRUE :388       TRUE :899
##  NA's :10        NA's :10
```

```
#create multiracial categories from these responses
#ignore non-hispanic other
demog$multirace02 <- paste(ifelse(white02,"W",""),
                           ifelse(black02,"B",""),
                           ifelse(indian02,"I",""),
                           ifelse(asian02,"A",""),
                           ifelse(hispanic02,"H",""), sep="")
demog$multirace02 <- factor(demog$multirace02,
                            levels=c("W","B","I","A","H",
                                     "WB","WI","WA","WH","BI","BA","BH","IA","IH","AH",
                                     "WBI","WBA","WBH","WIA","WIH","WAH","BIA","BIH","BAH","IAH",
                                     "WBIA","WBIH","WBAH","WIAH","BIAH",
                                     "WBIAH"))
table(demog$multirace02)
```

```
##
##     W     B     I     A     H    WB    WI    WA    WH    BI    BA    BH
##  4527  2078    55   170   866    49    45    18    19     7     4     9
##    IA    IH    AH   WBI   WBA   WBH   WIA   WIH   WAH   BIA   BIH   BAH
##     0     0     4     2     0     0     0     1     0     0     0     0
##   IAH  WBIA  WBIH  WBAH  WIAH  BIAH WBIAH
##     0     0     0     0     0     0     0
```

```
# how does this look if add in the other category?
multirace.other <- paste(ifelse(white02,"W",""),
                         ifelse(black02,"B",""),
                         ifelse(indian02,"I",""),
                         ifelse(asian02,"A",""),
                         ifelse(hispanic02,"H",""),
                         ifelse(other02,"O",""), sep="")
multirace.other <- factor(multirace.other,
                            levels=c("W","B","I","A","H","O",
                                     "WB","WI","WA","WH","WO","BI","BA","BH","BO","IA","IH","IO","AH","
                                     "WBI","WBA","WBH","WBO","WIA","WIH","WIO","WAH","WAO","BIA","BIH",
                                     "WBIA","WBIH","WBIO","WBAH","WBAO","WIAH","WIAO","BIAH","BIAO","IA
                                     "WBIAH","WBIAO","WBIAHO"))
table(multirace.other, droplevels(demog$multirace02), exclude=NULL)
```

```
##
## multirace.other    W    B    I    A    H   WB   WI   WA   WH   BI   BA
##          W      4525    0    0    0    0    0    0    0    0    0    0
##          B         0 2076    0    0    0    0    0    0    0    0    0
##          I         0    0   54    0    0    0    0    0    0    0    0
##          A         0    0    0  156    0    0    0    0    0    0    0
##          H         0    0    0    0  574    0    0    0    0    0    0
##          O         0    0    0    0    0    0    0    0    0    0    0
##          WB        0    0    0    0    0   37    0    0    0    0    0
##          WI        0    0    0    0    0    0   42    0    0    0    0
##          WA        0    0    0    0    0    0    0   13    0    0    0
##          WH        0    0    0    0    0    0    0    0    8    0    0
##          WO        2    0    0    0    0    0    0    0    0    0    0
##          BI        0    0    0    0    0    0    0    0    0    7    0
##          BA        0    0    0    0    0    0    0    0    0    0    3
##          BH        0    0    0    0    0    0    0    0    0    0    0
##          BO        0    2    0    0    0    0    0    0    0    0    0
##          IA        0    0    0    0    0    0    0    0    0    0    0
##          IH        0    0    0    0    0    0    0    0    0    0    0
##          IO        0    0    1    0    0    0    0    0    0    0    0
##          AH        0    0    0    0    0    0    0    0    0    0    0
##          AO        0    0    0   14    0    0    0    0    0    0    0
##          HO        0    0    0    0  292    0    0    0    0    0    0
##          WBI       0    0    0    0    0    0    0    0    0    0    0
##          WBA       0    0    0    0    0    0    0    0    0    0    0
##          WBH       0    0    0    0    0    0    0    0    0    0    0
##          WBO       0    0    0    0    0   12    0    0    0    0    0
##          WIA       0    0    0    0    0    0    0    0    0    0    0
##          WIH       0    0    0    0    0    0    0    0    0    0    0
##          WIO       0    0    0    0    0    0    3    0    0    0    0
##          WAH       0    0    0    0    0    0    0    0    0    0    0
##          WAO       0    0    0    0    0    0    0    5    0    0    0
##          BIA       0    0    0    0    0    0    0    0    0    0    0
##          BIH       0    0    0    0    0    0    0    0    0    0    0
##          BIO       0    0    0    0    0    0    0    0    0    0    0
##          BAH       0    0    0    0    0    0    0    0    0    0    0
##          BAO       0    0    0    0    0    0    0    0    0    0    1
##          IAH       0    0    0    0    0    0    0    0    0    0    0
##          IAO       0    0    0    0    0    0    0    0    0    0    0
##          AHO       0    0    0    0    0    0    0    0    0    0    0
##          WBIA      0    0    0    0    0    0    0    0    0    0    0
##          WBIH      0    0    0    0    0    0    0    0    0    0    0
##          WBIO      0    0    0    0    0    0    0    0    0    0    0
##          WBAH      0    0    0    0    0    0    0    0    0    0    0
##          WBAO      0    0    0    0    0    0    0    0    0    0    0
##          WIAH      0    0    0    0    0    0    0    0    0    0    0
##          WIAO      0    0    0    0    0    0    0    0    0    0    0
##          BIAH      0    0    0    0    0    0    0    0    0    0    0
##          BIAO      0    0    0    0    0    0    0    0    0    0    0
##          IAHO      0    0    0    0    0    0    0    0    0    0    0
##          WBIAH     0    0    0    0    0    0    0    0    0    0    0
##          WBIAO     0    0    0    0    0    0    0    0    0    0    0
##          WBIAHO    0    0    0    0    0    0    0    0    0    0    0
##          <NA>      0    0    0    0    0    0    0    0   11    0    0
##
## multirace.other   BH   AH  WBI  WIH <NA>
##          W         0    0    0    0    0
##          B         0    0    0    0    0
##          I         0    0    0    0    0
##          A         0    0    0    0    0
##          H         0    0    0    0    0
##          O         0    0    0    0   32
##          WB        0    0    0    0    0
##          WI        0    0    0    0    0
##          WA        0    0    0    0    0
##          WH        0    0    0    0    0
##          WO        0    0    0    0    0
##          BI        0    0    0    0    0
##          BA        0    0    0    0    0
##          BH        0    0    0    0    0
##          BO        0    0    0    0    0
##          IA        0    0    0    0    0
##          IH        0    0    0    0    0
##          IO        0    0    0    0    0
##          AH        0    1    0    0    0
##          AO        0    0    0    0    0
##          HO        0    0    0    0    0
##          WBI       0    0    2    0    0
##          WBA       0    0    0    0    0
##          WBH       0    0    0    0    0
##          WBO       0    0    0    0    0
##          WIA       0    0    0    0    0
##          WIH       0    0    0    0    0
##          WIO       0    0    0    0    0
##          WAH       0    0    0    0    0
##          WAO       0    0    0    0    0
##          BIA       0    0    0    0    0
##          BIH       0    0    0    0    0
##          BIO       0    0    0    0    0
##          BAH       0    0    0    0    0
##          BAO       0    0    0    0    0
##          IAH       0    0    0    0    0
##          IAO       0    0    0    0    0
##          AHO       0    3    0    0    0
##          WBIA      0    0    0    0    0
##          WBIH      0    0    0    0    0
##          WBIO      0    0    0    0    0
##          WBAH      0    0    0    0    0
##          WBAO      0    0    0    0    0
##          WIAH      0    0    0    0    0
##          WIAO      0    0    0    0    0
##          BIAH      0    0    0    0    0
##          BIAO      0    0    0    0    0
##          IAHO      0    0    0    0    0
##          WBIAH     0    0    0    0    0
##          WBIAO     0    0    0    0    0
##          WBIAHO    0    0    0    0    0
##          <NA>      9    0    0    1   10
```

```
# Collect Roster Data -----------------------------------------------------

# The demographic information is listed in columns for each household member and
# then each non-HH member. First I need to collect these arrays for specific
# demographic characteristics, then loop through and pull out the bio mom and
# dad based on the indicated relationship to the respondent of that column.

#how deep to go in the household roster. Some dads as deep as #16.
hhdepth <- 16
#how deep to go in the non-HH roster. Deepest parent is at #8
nhhdepth <- 8

#add an NA column to each roster in order to easily include missing parents
age <- cbind(roster[,c(paste("R",seq(from=1080300, by=100, length=hhdepth), sep=""))],
             roster[,c(paste("R",seq(from=1163700, by=100, length=nhhdepth), sep=""))],
             NA)
```

```
ethnic <- cbind(roster[,c(paste("R",seq(from=1094600, by=100, length=hhdepth), sep=""))],
                roster[,c(paste("R",seq(from=1172500, by=100, length=nhhdepth), sep=""))],
                NA)
grade <- cbind(roster[,c(paste("R",seq(from=1099400, by=100, length=hhdepth), sep=""))],
               roster[,c(paste("R",seq(from=1176900, by=100, length=nhhdepth), sep=""))],
               NA)
race <- cbind(roster[,c(paste("R",seq(from=1115400, by=100, length=hhdepth), sep=""))],
              roster[,c(paste("R",seq(from=1184500, by=100, length=nhhdepth), sep=""))],
               NA)
relate <- cbind(roster[,c(paste("R",seq(from=1315800, by=100, length=hhdepth), sep=""))],
              roster[,c(paste("R",seq(from=1186600, by=100, length=nhhdepth), sep=""))])
informant <- roster[,c(paste("R",seq(from=1102600, by=100, length=hhdepth), sep=""))]==1

#get informant relationship
#check to make sure there is always one and only one informant
summary(apply(informant,1,sum))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.9994  1.0000  2.0000
```

```
#damn it!
sum(apply(informant,1,sum)==0)
```

```
## [1] 8
```

```
sum(apply(informant,1,sum)>1)
```

```
## [1] 3
```

```
#8 cases of no informant, 3 cases of multiple informants.

#ok, so lets figure out informant. note that 0 is a real value for respondent,
#so need to add one to relationship values and then subtract in final product
informant_relationship <- (relate[,1:hhdepth]+1)*informant
informant_relationship[informant_relationship<=0] <- 100
#for cases of double values, take the relationship that has a smaller value
temp <- apply(informant_relationship, 1, min)-1
informant_relationship <- cut(temp, c(0,1,3,5,88,100), right=FALSE,
                              labels=c("Self","Spouse","Bio Parent",
                                       "Other","Unknown"))
summary(informant_relationship)
```

```
##       Self     Spouse Bio Parent      Other    Unknown
##        415          5       7116       1432         16
```

```
round(prop.table(table(informant_relationship))*100,1)
```

```
## informant_relationship
##       Self     Spouse Bio Parent      Other    Unknown
##        4.6        0.1       79.2       15.9        0.2
```

```
#lets combine spouse, other, unknown together for parsimony
temp <- factor(ifelse(informant_relationship=="Self","Self",
                      ifelse(informant_relationship=="Bio Parent",
                             "Bio Parent",
                             "Other")),
               levels=c("Self","Bio Parent","Other"))
```

```
table(informant_relationship, temp, exclude=NULL)
```

```
##                       temp
## informant_relationship Self Bio Parent Other
##             Self        415          0     0
##             Spouse        0          0     5
##             Bio Parent    0       7116     0
##             Other         0          0  1432
##             Unknown       0          0    16
```

```
informant_relationship <- temp

#use which to quickly extract the index of all bio parents. Fathers are 4, mothers are 3.
dadid <- momid <- rep(ncol(relate)+1, nrow(relate))
temp <- which(relate==4, arr.ind=TRUE)
dadid[temp[,1]] <- temp[,2]
temp <- which(relate==3, arr.ind=TRUE)
momid[temp[,1]] <- temp[,2]

#How do I pull out variable columns of a matrix? Thank you StackOverflow!
#https://stackoverflow.com/questions/25584039/how-to-extract-different-columns-from-each-row-of-a-data-

#one addition to this is that I need to deal with the zero dadid and momid
#because they will be dropped by routine making my vectors too small. So I
#replace zeros by the last column of the matrices which is just an NA column.
dadid[dadid==0] <- ncol(age)
momid[momid==0] <- ncol(age)

parents <- data.frame(id=roster$R0000100,
                      informant=informant_relationship,
                      fage=age[cbind(seq_along(dadid), dadid)],
                      feduc=grade[cbind(seq_along(dadid), dadid)],
                      fethnic=ethnic[cbind(seq_along(dadid), dadid)],
                      frace=race[cbind(seq_along(dadid), dadid)],
                      mage=age[cbind(seq_along(momid), momid)],
                      meduc=grade[cbind(seq_along(momid), momid)],
                      methnic=ethnic[cbind(seq_along(momid), momid)],
                      mrace=race[cbind(seq_along(momid), momid)])

#now recode variables, fix missing values, etc.
parents$fage <- ifelse(!is.na(parents$fage) & parents$fage<0, NA, parents$fage)
parents$mage <- ifelse(!is.na(parents$mage) & parents$mage<0, NA, parents$mage)
parents$feduc <- ifelse(!is.na(parents$feduc) & (parents$feduc<0 | parents$feduc>20),
                        NA, parents$feduc)
parents$meduc <- ifelse(!is.na(parents$meduc) & (parents$meduc<0 | parents$meduc>20),
                        NA, parents$meduc)
parents$fethnic <- factor(parents$fethnic, levels=0:1, labels=c("Not Hispanic","Hispanic"))
parents$methnic <- factor(parents$methnic, levels=0:1, labels=c("Not Hispanic","Hispanic"))
parents$frace <- factor(parents$frace, levels=1:7,
                        labels=c("White","Black","AmIndian","Asian","Other","Hispanic","Mixed"))
parents$mrace <- factor(parents$mrace, levels=1:7,
                        labels=c("White","Black","AmIndian","Asian","Other","Hispanic","Mixed"))
#replace race with hispanic if ethnicity variable hispanic
temp <- factor(ifelse(!is.na(parents$fethnic) & parents$fethnic=="Hispanic",
                      "Hispanic", as.character(parents$frace)),
               levels=c("White","Black","AmIndian","Asian","Other","Hispanic","Mixed"))
table(parents$frace, temp, exclude=NULL)
```

```
##           temp
##            White Black AmIndian Asian Other Hispanic Mixed <NA>
##   White     4242     0        0     0     0      663     0    0
##   Black        0  2006        0     0     0       46     0    0
##   AmIndian     0     0       49     0     0       16     0    0
##   Asian        0     0        0   142     0        6     0    0
##   Other        0     0        0     0   114      324     0    0
##   Hispanic     0     0        0     0     0      431     0    0
##   Mixed        0     0        0     0     0        6     6    0
##   <NA>         0     0        0     0     0       49     0  884
```

```
parents$frace <- temp
temp <- factor(ifelse(!is.na(parents$methnic) & parents$methnic=="Hispanic",
                      "Hispanic", as.character(parents$mrace)),
               levels=c("White","Black","AmIndian","Asian","Other","Hispanic","Mixed"))
table(parents$mrace, temp, exclude=NULL)
```

```
##           temp
##            White Black AmIndian Asian Other Hispanic Mixed <NA>
##   White     4526     0        0     0     0      729     0    0
##   Black        0  2197        0     0     0       43     0    0
##   AmIndian     0     0       65     0     0       19     0    0
##   Asian        0     0        0   173     0        5     0    0
##   Other        0     0        0     0    50      156     0    0
##   Hispanic     0     0        0     0     0      698     0    0
##   Mixed        0     0        0     0     0        9    10    0
##   <NA>         0     0        0     0     0       56     0  248
```

```
parents$mrace <- temp

with(parents, table(frace, mrace, exclude=NULL))
```

```
##           mrace
## frace      White Black AmIndian Asian Other Hispanic Mixed <NA>
##   White     3978     8       30    28     4      153     5   36
##   Black       74  1835        1     5     5       41     3   42
##   AmIndian    24     4       18     0     1        2     0    0
##   Asian       16     0        0   120     0        0     0    6
##   Other       21     6        1     2    30       52     1    1
##   Hispanic   160    40        8     6     5     1303     0   19
##   Mixed        4     1        0     0     0        0     1    0
##   <NA>       249   303        7    12     5      164     0  144
```

```
#create parent mixed race variable. Ignore gender of parent because DF
temp <- paste(parents$frace, parents$mrace, sep=".")
TF <- !is.na(parents$frace) & !is.na(parents$mrace) & parents$frace==parents$mrace
temp[TF] <- as.character(parents$frace)[TF]
TF <- !is.na(parents$frace) & !is.na(parents$mrace) &
  as.numeric(parents$frace)>as.numeric(parents$mrace)
temp[TF] <- paste(parents$mrace, parents$frace, sep=".")[TF]
temp <- gsub("White","W",temp)
temp <- gsub("Black","B",temp)
```

```r
temp <- gsub("AmIndian","I",temp)
temp <- gsub("Asian","A",temp)
temp <- gsub("Hispanic","H",temp)
#mixed, other, or missing parents are NA for our purposes
temp[grepl("NA|Mixed|Other", temp)] <- NA
parents$mixedrace_parent <- factor(gsub("\\.","", temp),
                                   levels=c("W","B","I","A","H",
                                            "WB","WI","WA","WH",
                                            "BI","BA","BH",
                                            "IA","IH","AH"))
table(parents$mixedrace_parent, exclude=NULL)
```

```
##
##     W    B    I    A    H   WB   WI   WA   WH   BI   BA   BH   IA   IH   AH
## 3978 1835   18  120 1303   82   54   44  313    5    5   81    0   10    6
## <NA>
## 1130
```

```r
#code residential parents
#parents$dadres <- dadid<=hhdepth
#parents$momres <- momid<=hhdepth

summary(parents)
```

```
##        id             informant         fage           feduc
##  Min.   :   1   Self      : 415   Min.   : 0.0   Min.   : 0.00
##  1st Qu.:2249   Bio Parent:7116   1st Qu.:38.0   1st Qu.:12.00
##  Median :4502   Other     :1453   Median :42.0   Median :12.00
##  Mean   :4504                     Mean   :42.2   Mean   :12.45
##  3rd Qu.:6758                     3rd Qu.:46.0   3rd Qu.:14.00
##  Max.   :9022                     Max.   :81.0   Max.   :20.00
##                                   NA's   :1245   NA's   :1816
##          fethnic           frace           mage            meduc
##  Not Hispanic:6566   White   :4242   Min.   :  0.00   Min.   : 0.00
##  Hispanic    :1541   Black   :2006   1st Qu.: 36.00   1st Qu.:12.00
##  NA's        : 877   Hispanic:1541   Median : 39.00   Median :12.00
##                      Asian   : 142   Mean   : 39.69   Mean   :12.38
##                      Other   : 114   3rd Qu.: 43.00   3rd Qu.:14.00
##                      (Other) :  55   Max.   :117.00   Max.   :20.00
##                      NA's    : 884   NA's   :406      NA's   :705
##          methnic           mrace      mixedrace_parent
##  Not Hispanic:7013   White   :4526   W      :3978
##  Hispanic    :1710   Black   :2197   B      :1835
##  NA's        : 261   Hispanic:1715   H      :1303
##                      Asian   : 173   WH     : 313
##                      AmIndian:  65   A      : 120
##                      (Other) :  60   (Other): 305
##                      NA's    : 248   NA's   :1130
```

```r
#remove cases that do not have a result for parentally based combined race

# Merge Datasets and Save -------------------------------------------------------

# First limit variables to just the key ones for analysis
demog <- subset(demog,
                select = c("id","gender","age","urban97","region97","migration",
                           "moved_out","family","hhinc","highparented","asvab",
                           "gpa_overall","enrollment02","multirace02"))

parents <- subset(parents,
                  select=c("id","informant","fage","mage","mixedrace_parent",
                           "frace","mrace"))

#now merge by id
nlsy <- merge(demog, parents, by="id", all.x=FALSE, all.y=FALSE)

#how many cases are missing on each race variable and combined
sum(is.na(nlsy$multirace02))
```

```
## [1] 42
```

```
sum(is.na(nlsy$mixedrace_parent))
```

```
## [1] 992
```

```
sum(is.na(nlsy$multirace02) | is.na(nlsy$mixedrace_parent))
```

```
## [1] 1023
```

```
#remove cases missing on either race variable
nlsy <- subset(nlsy, !is.na(multirace02) & !is.na(mixedrace_parent))
nrow(nlsy)
```

```
## [1] 6873
```

```
#drop any unused factor levels to simplify multiple imputation later
nlsy <- droplevels(nlsy)

#summary to check everything
summary(nlsy)
```

```
##        id          gender          age         urban97
##  Min.   :   1   Male  :3484   Min.   :18.00   Rural:1632
##  1st Qu.:2296   Female:3389   1st Qu.:19.25   Urban:4951
##  Median :4502                 Median :20.42   NA's : 290
##  Mean   :4506                 Mean   :20.45
##  3rd Qu.:6696                 3rd Qu.:21.67
##  Max.   :9020                 Max.   :23.50
##
##           region97                migration    moved_out
##  Northeast    :1175   Non-movers       :5213   Mode :logical
##  North Central:1624   Within county    : 299   FALSE:4676
##  South        :2571   Different county : 737   TRUE :2193
##  West         :1503   Different state  : 533   NA's :4
##                       Different country:  55
##                       NA's             :  36
##
##              family         hhinc        highparented       asvab.V1
##  Two bio p     :3801   Min.   :     5   Min.   : 1.00   Min.   :-1.5532
##  Bio mom       :1771   1st Qu.: 20000   1st Qu.:12.00   1st Qu.:-0.8585
##  Two p, bio mom: 691   Median : 40000   Median :13.00   Median :-0.0421
##  Bio dad       : 203   Mean   : 48234   Mean   :13.36   Mean   : 0.0246
##  Two p, bio dad: 151   3rd Qu.: 63500   3rd Qu.:16.00   3rd Qu.: 0.8816
##  (Other)       : 244   Max.   :246474   Max.   :20.00   Max.   : 1.8561
##  NA's          :  12   NA's   :1704     NA's   :164     NA's   :1242
##   gpa_overall                         enrollment02   multirace02
##  Min.   :0.420   Enrolled in 4-yr college  :1688   W      :4115
##  1st Qu.:2.460   Not enrolled, HS Degree   :1577   B      :1711
##  Median :2.885   Not enrolled, no HS degree: 938   H      : 727
##  Mean   :2.844   Not enrolled, some college: 803   A      : 138
##  3rd Qu.:3.280   Enrolled in 2-yr college  : 726   I      :  51
##  Max.   :4.110   (Other)                   :1124   WI     :  39
##  NA's   :2033    NA's                      :  17   (Other):  92
##       informant         fage            mage        mixedrace_parent
##  Self      : 307   Min.   : 0.00   Min.   :  0.00   W      :3441
##  Bio Parent:5669   1st Qu.:38.00   1st Qu.: 36.00   B      :1630
##  Other     : 897   Median :42.00   Median : 39.00   H      :1143
##                    Mean   :42.24   Mean   : 39.76   WH     : 285
##                    3rd Qu.:46.00   3rd Qu.: 43.00   A      : 106
##                    Max.   :81.00   Max.   :117.00   WB     :  72
##                    NA's   :329     NA's   :122      (Other): 196
##       frace           mrace
##  White   :3641   White   :3686
##  Black   :1731   Black   :1676
##  AmIndian:  44   AmIndian:  54
##  Asian   : 119   Asian   : 142
##  Hispanic:1338   Hispanic:1315
##
##
```

```
#table of multiple race response by parental race response
with(nlsy, table(multirace02, mixedrace_parent, exclude=NULL))
```

```
##            mixedrace_parent
## multirace02    W    B    I    A    H   WB   WI   WA   WH   BI   BA   BH
##         W   3385   10    2    2  447    3   30   16  212    0    0    5
##         B      4 1604    0    0    9   42    0    0    0    3    3   46
##         I      8    0   11    0   16    0   10    0    1    0    0    0
##         A      5    1    3  103    6    0    0   16    2    0    1    0
##         H     12    0    0    0  651    1    0    0   53    0    0    9
##         WB     5    6    0    0    0   24    0    0    0    0    0    2
##         WI    16    0    2    0    8    0    8    0    5    0    0    0
##         WA     3    0    0    0    0    0    0    8    0    0    0    0
##         WH     3    0    0    0    5    0    0    0   11    0    0    0
##         BI     0    3    0    0    0    1    0    0    0    1    0    0
##         BA     0    1    0    0    0    1    0    0    0    0    0    0
##         BH     0    3    0    0    0    0    0    0    0    0    0    5
##         AH     0    0    0    1    1    0    0    0    0    0    0    0
##         WBI    0    2    0    0    0    0    0    0    0    0    0    0
##         WIH    0    0    0    0    0    0    0    0    1    0    0    0
##            mixedrace_parent
## multirace02   IH   AH
##         W      3    0
##         B      0    0
##         I      5    0
##         A      0    1
##         H      1    0
##          WB     0    0
##          WI     0    0
##          WA     1    3
##          WH     0    0
##          BI     0    0
##          BA     0    0
##          BH     0    0
##          AH     0    1
##          WBI    0    0
##          WIH    0    0
```

```
#table of parents race
with(nlsy, table(frace, mrace, exclude=NULL))
```

```
##            mrace
## frace       White Black AmIndian Asian Hispanic
##   White      3441     8       27    27      138
##   Black        64  1630        1     4       32
##   AmIndian     21     3       18     0        2
##   Asian        13     0        0   106        0
##   Hispanic    147    35        8     5     1143
```

```
save(nlsy, file="output/nlsy_processed.RData")
```