**Overview**

This document describes how we've translated data from the National Longitudinal Survey of Youth 1979 and 1997 cohorts for machine learning.

The "key" listed under each question explains the values that we've chosen to use in our final data. (In other words, it may not exactly match the keys used in the original 1979 or 1997 cohort data.)

The "exception handling" sections under each question describe how we've handled exception codes specific to that variable. Some general exception codes also exist, and where applicable, they've been preserved in our data:

-1 REFUSAL

-2 DON'T KNOW

-3 INVALID SKIP

-4 VALID SKIP

-5 NON-INTERVIEW

**Race**

**Explanation:** This represents the respondent's race, as reported in their initial interview.  
 **Questions:** In the 1979 cohort, SAMPLE\_RACE. In the 1997 cohort, KEY!RACE\_ETHNICITY.

**Key:**

1 Black

2 Hispanic

3 Non-Black / Non-Hispanic

**Exception handling:** In the 1997 cohort, a separate category existed for "Mixed Race (Non-Hispanic)". This included only 83 respondents and is translated in our data into the "Non-Black / Non-Hispanic" category.

**Gender**

**Explanation:** This represents the respondent's gender, as reported in their initial interview.

**Questions:** In the 1997 cohort, SAMPLE\_SEX. In the 1997 cohort, KEY!SEX.

**Key:**

1 Male

2 Female

**Age**

**Explanation:** This is a continuous variable representing the respondent's age as of the interview date.

**Questions:** In the 1979 cohort, AGEATINT. In the 1997 cohort, CV\_AGE\_INT\_DATE.

**Note:** This was a constructed variable in the original data set.

**Highest Grade Completed**

**Explanation:** This is a continuous variable representing years of schooling, ranging from 0 (no education) to 20.

**Questions:** In the 1979 cohort, HGCREV79 to HGCREV16. In the 1997 cohort, YSCH-3112.

**Key:**

0 NONE

1 1ST GRADE

2 2ND GRADE

3 3RD GRADE

4 4TH GRADE

5 5TH GRADE

6 6TH GRADE

7 7TH GRADE

8 8TH GRADE

9 9TH GRADE

10 10TH GRADE

11 11TH GRADE

12 12TH GRADE

13 1ST YEAR COLLEGE

14 2ND YEAR COLLEGE

15 3RD YEAR COLLEGE

16 4TH YEAR COLLEGE

17 5TH YEAR COLLEGE

18 6TH YEAR COLLEGE

19 7TH YEAR COLLEGE

20 8TH YEAR COLLEGE OR MORE

**Exception handling:** Across the 1979 and 1997 cohorts, several exception codes were utilized:

93 PRE-KINDERGARTEN

94 KINDERGARTEN

95 UNGRADED

Our translator converts any "9X" code to 0.

**Hours Worked in Last Calendar Year**

**Explanation:** This is a continuous variable representing the total number of hours the respondent worked in the past calendar year.

**Variables:** In the 1997 cohort, HRSWK-PCY. In the 1997 cohort, CVC\_HOURS\_WK\_YR\_ALL.97 to CVC\_HOURS\_WK\_YR\_ALL.16.

**Weeks Worked in Last Calendar Year**

**Explanation:** This is a continuous variable representing the total number of weeks in which the respondent worked at any job in the past calendar year.

**Variables:** In the 1979 cohort, WKSWK-PCY. In the 1997 cohort, CVC\_WKSWK\_YR\_ALL.97 to CVC\_WKSWK\_YR\_ALL.16.

**Limited in Kind of Work Because of Health**

**Explanation:** This is a binary variable representing whether the respondent was limited in the kind of work they could perform in the past calendar year due to health conditions.

**Variables:** In the 1979 cohort, Q11-4. In the 1997 cohort, YHEA-1005.

**Key:**

1 YES

0 NO

**Data limitations:** For the 1997 cohort, this question was asked only from 2007 on. (At that point, only about 5% of respondents reported a health limitation.)

**Limited in Amount of Work Because of Health**

**Explanation:** This is a binary variable representing whether the respondent was limited in the amount of work they could perform in the past calendar year due to health conditions.

**Variables:** In the 1979 cohort, Q11-5. In the 1997 cohort, YHEA-1006.

**Key:**

1 YES

0 NO

**Data limitations:** For the 1997 cohort, this question was asked only from 2007 on. (At that point, only about 5% of respondents reported a health limitation.)

**Household Size**

**Explanation:** This is a continuous variable representing the respondent's household size.

**Variables:** In the 1979 cohort, FAMSIZE. In the 1997 cohort, CV\_HH\_SIZE.

**Data limitations:** This represents a subtly different question in each cohort. In 1979, it represents family size; in 1997, it represents household size.

**Region**

**Explanation:** This is a categorical variable representing the region of the respondent's household.

**Variables:** In the 1979 cohort, REGION. In the 1997 cohort, CV\_CENSUS\_REGION.

**Key:**

1 NORTHEAST

2 NORTH CENTRAL

3 SOUTH

4 WEST

**Urban-Rural**

**Explanation:** This is a binary variable representing whether the respondent lives in an urban or rural area.

**Variables:** In the 1979 cohort, URBAN-RURAL. In the 1997 cohort, CV\_URBAN-RURAL.

**Key:**

0 Rural

1 Urban

**Exception handling:** In the 1997 data, a "2" value is included for "UNKNOWN." Our translator converts those responses to the broader "-2" exception code (i.e., the standard "DON'T KNOW" exception).

**Marital Status**

**Explanation:** This is a categorical variable representing the respondent's marital status as of the interview date.

**Variables:** In the 1979 cohort, MARSTAT-KEY. In the 1997 cohort, CV\_MARSTAT\_COLLAPSED.

**Key:**

0 NEVER MARRIED

1 MARRIED

2 SEPARATED

3 DIVORCED

4 WIDOWED

**Issues:** In the 1979 cohort, the code "6" indicated widowed. That has been translated to "4" in our data to match the 1997 key.

In the 1979 cohort, a "5" code indicated "REMARRIED", but the 1997 data has no comparable code. In our data, these responses have been translated to the broader "1" key for "MARRIED".

**Current Pregnancy**

**Explanation:** This is a binary variable indicating whether the respondent is currently pregnant.

**Variables:** In the 1979 cohort, FFER-13 through 1990 and Q11-5C afterward. In the 1997 cohort, YSAQ-315 through 2001 and YSAQ2-315 afterward.

**Key:**

1 Yes

0 No

**Issues:** This data exists only for female respondents.

**Number of Children Ever Born**

**Explanation:** This is a continuous variable indicating the number of biological children ever born to the respondent.

**Variables:** In the 1979 cohort, NUMKID80 through NUMKID16. In the 1997 cohort, YSAQ-324 through 2001 and YSAQ2-324 afterward.

**Issues:** There's a subtle difference between the 1979 question, which asks about number of children ever born, and the 1997 question, which asks specifically about live births.

**Number of Children in the Household**

**Explanation:** This is a continuous variable indicating the number of children currently living in the respondent's household.

**Variables:** In the 1979 cohort, NUMCH79 through NUMCH16. In the 1997 cohort, CV\_HH\_UNDER\_18.

**Issues:** The 1997 question specifically asks about biological children, adopted children, and stepchildren in the household. The 1997 question, by contrast, asks about all household residents under the age of 18.

**Total Income from Wages and Salaries**

**Explanation:** This is a continuous variable representing the respondent's total wage and salary income from the prior year.

**Variables:** In the 1979 cohort, Q13-5 through 1981, then Q13-5\_TRUNC\_REVISED through 2000, then Q13-5\_TRUNC afterward. In the 1997 cohort, YINC-1700.

**Issues:** Very high salary values are topcoded in the NLSY data in various ways over time. Our translator returns "topcoded" for a topcoded value.

To enable meaningful comparisons of dollar values over time, our translator converts all figures to 2015 dollars.