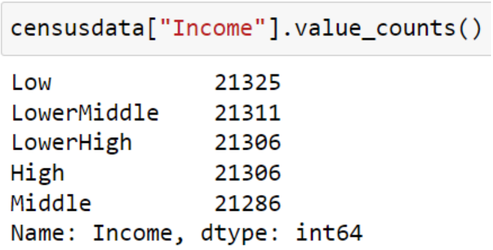
# Methodology

1) Collect data
   Main data source: https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html
   Selected only “Person” dataset
   Initial attributes: 40, Initial Observations: 152732
2) Initial Data Analysis
   Data was factorized already except “Income” attribute
   cpsmar22.pdf (census.gov) – page 391 –according to the “Public Use Benchmarks” excluding people without income; where PTOTVAL(Total persons income) =0. Also removing negative incomes.
   Secondary Instances: 106534, attributes: 40

Preparing target variable “Income”—taking 5 quantiles and labelling them "1", "2", "3", "4", "5"
1= Low, 2= LowerMiddle, 3= Middle, 4= LowerHigh, 5= High



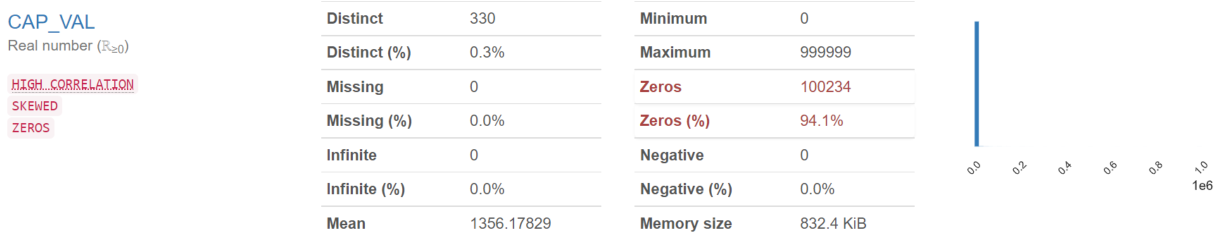
```
censusdata["Income"].value_counts()
```

```
Low            21325
LowerMiddle    21311
LowerHigh      21306
High           21306
Middle         21286
Name: Income, dtype: int64
```

**[This part done with R]**

Feature selection

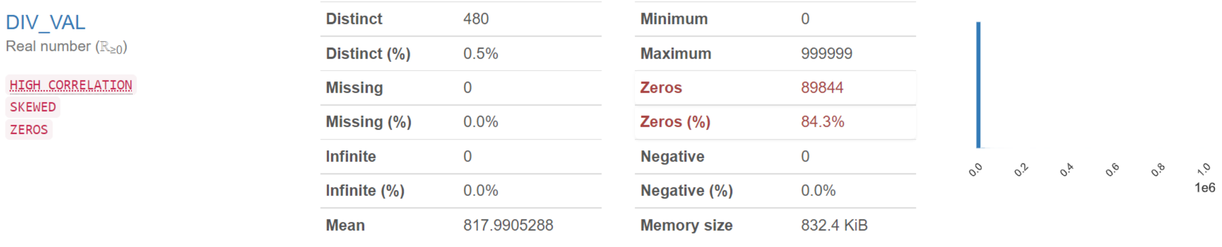
a) Filter method
   Missingness: CAP_VAL has 94.1% and DIV_VAL has 84.3% missing values. In initial “Filter” better to remove these two attributes. STATETAX_A has 55.4% and STATETAX_B has 57.6% missing values. We will keep it for now. PRUNTYPE has 98% missing data. In initial “Filter” better to remove it. (37 Attributes)



CAP_VAL
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
SKEWED
ZEROS

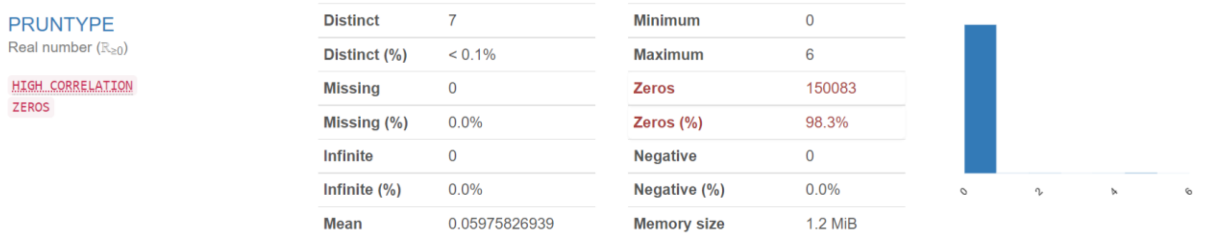
| | | | |
|---|---|---|---|
| Distinct | 330 | Minimum | 0 |
| Distinct (%) | 0.3% | Maximum | 999999 |
| Missing | 0 | Zeros | 100234 |
| Missing (%) | 0.0% | Zeros (%) | 94.1% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 1356.17829 | Memory size | 832.4 KiB |



DIV_VAL
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
SKEWED
ZEROS

| | | | |
|---|---|---|---|
| Distinct | 480 | Minimum | 0 |
| Distinct (%) | 0.5% | Maximum | 999999 |
| Missing | 0 | Zeros | 89844 |
| Missing (%) | 0.0% | Zeros (%) | 84.3% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 817.9905288 | Memory size | 832.4 KiB |



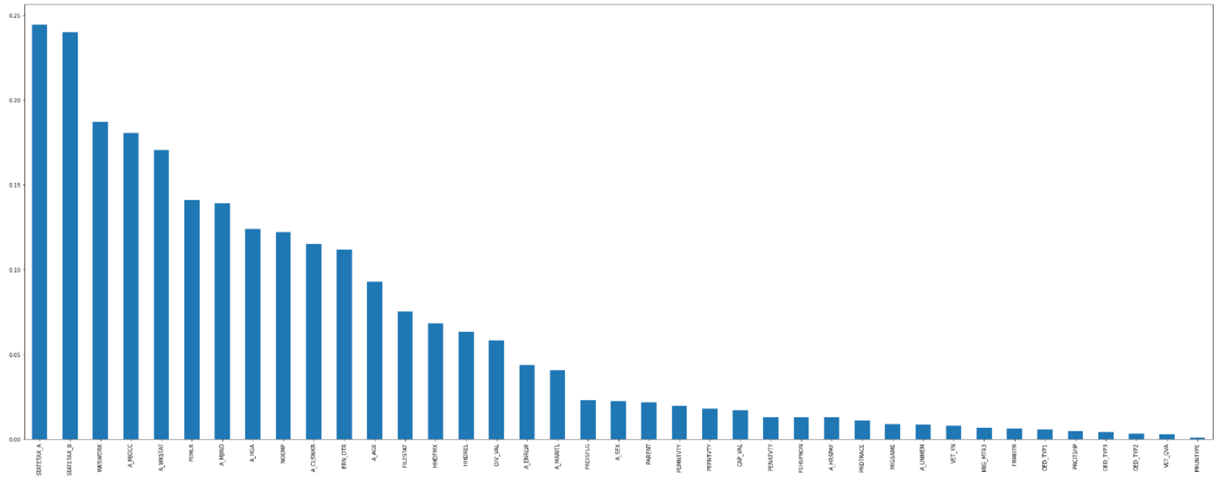
PRUNTYPE
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

| | | | |
|---|---|---|---|
| Distinct | 7 | Minimum | 0 |
| Distinct (%) | < 0.1% | Maximum | 6 |
| Missing | 0 | Zeros | 150083 |
| Missing (%) | 0.0% | Zeros (%) | 98.3% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 0.05975826939 | Memory size | 1.2 MiB |

Mutual Information Gain: Using whole dataset –Selecting k=28 best, where MI >= 0.01

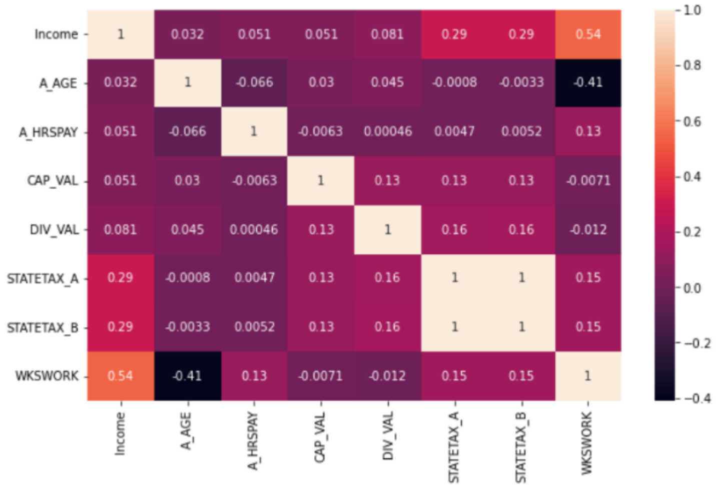
['PEHSPNON', 'PENATVTY', 'PEMNTVTY', 'PEFNTVTY', 'PRDTRACE', 'PEMLR', 'PRDISFLG', 'A_SEX', 'A_ENRLW', 'A_MARITL', 'A_HGA', 'A_AGE', 'A_MJIND', 'A_MJOCC', 'A_HRSPAY', 'ERN_OTR', 'CAP_VAL', 'DIV_VAL', 'FILESTAT', 'STATETAX_A', 'STATETAX_B', 'A_CLSWKR', 'A_WKSTAT', 'PARENT', 'HHDFMX', 'HHDREL', 'WKSWORK', 'NOEMP']



Variance Threshold: Threshold=0.5, Selected = 25 attributes

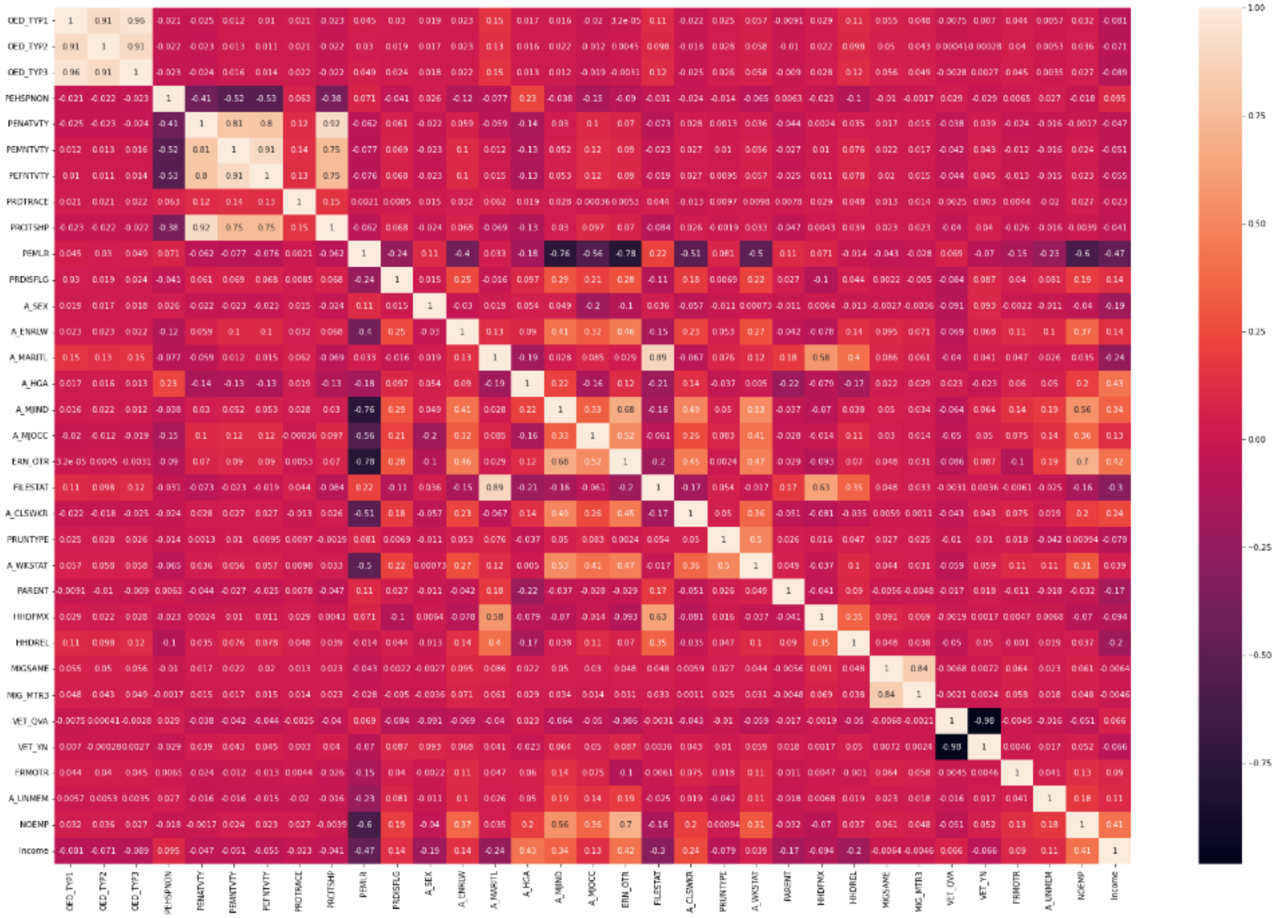
'PENATVTY', 'PEMNTVTY', 'PEFNTVTY', 'PRDTRACE', 'PRCITSHP', 'PEMLR', 'A_ENRLW', 'A_MARITL', 'A_HGA', 'A_AGE', 'A_MJIND', 'A_MJOCC', 'A_HRSPAY', 'ERN_OTR', 'CAP_VAL', 'DIV_VAL', 'FILESTAT', 'STATETAX_A', 'STATETAX_B', 'A_CLSWKR', 'A_WKSTAT', 'HHDFMX', 'HHDREL', 'WKSWORK', 'NOEMP'

Correlation Coefficients: Creating subsets of nominal and numeric attributes.
Numeric= 7+ ”Income”, Nominal = 32+”Income”



| Nominal: | Numeric: |
|---|---|
| 1)Very strong positive correlation (1.0 to 0.8)<br>OED_TYP1 & OED_TYP2<br>OED_TYP1 & OED_TYP3<br>OED_TYP2 & OED_TYP3<br>PENATVTY & PEFNTVTY<br>PENATVTY & PEMNTVTY<br>PEFNTVTY & PEMNTVTY<br>PENATVTY & PRCITSHP<br>FILESTAT & A_MARITL<br>MIGSAME & MIG_MTR3 | 1)Very strong positive correlation (1.0 to 0.8)<br>STATETAX_A & STATETAX_B |
| 2)Strong positive correlation (0.8 to 0.6)<br>PEMNTVTY & PRCITSHP<br>PEFNTVTY & PRCITSHP<br>ERN_OTR & A_MJIND<br>ERN_OTR & NOEMP<br>HHDFMX & FILESTAT | 2)Moderate positive correlation (0.6 to 0.4)<br>Income & WKSWORK |
| 3)Moderate positive correlation (0.6 to 0.4)<br>HHDFMX & A_MARITL<br>NOEMP & A_MJIND<br>ERN_OTR & A_WKSTAT<br>ERN_OTR & A_CLSWKR<br>A_MJIND & A_WKSTST<br>A_MJIND & A_CLSWKR<br>A_MJOCC & A_WKSTAT<br>ERN_OTR & A_MJIND<br>ERN_OTR & A_MJOCC<br>HHDREL & A_MARITL | 3)Moderate negative correlation (-0.4 to -0.6)<br>A_AGE & WKSWORK |

| | |
|---|---|
| 4)Moderate negative correlation (-0.4 to -0.6)<br>PEMLR & Income<br>PEMLR & NOEMP<br>PEMLR & A_WKSTST<br>PEMLR & A_CLSWKR<br>PEMLR & A_MJOCC<br>PEMLR & A_ENRLW<br>PENATVTY & PEHSPNON<br>PEMNTVTY & PEHSPNON<br>PEFNTVTY & PEHSPNON | |
| 5)Strong negative correlation (-0.6 to -0.8)<br>PEMLR & A_MJIND<br>PEMLR & ERN_OTR | |
| 6)Very strong negative correlation (-0.8 to -1.0)<br>VET_YN & VET_QVA | |



### Attributes that are highly corrected _40Attributes(ideally >0.75)

Highly Correlated Attributes = [34 10 20 17 16 23 14  6  7  5 24  3  2 36 19 33]
[“WKSWORK”, “PEMLR”, “ERN_OTR”, “A_MJIND”, “A_AGE”, “FILESTAT”, “A_MARITL”, “PEMNTVTY”, “PEFNTVTY”, “PENATVTY”, “STATETAX_A”, “OED_TYP3”, “OED_TYP2”, “VET_YN”, “A_HIRSPAY”, “MIG_MTR3”]

b) Wrapper Method
Forward Selection / Backward Elimination

Optimal 32 features are:

[Income ~ OED_TYP1 + OED_TYP2 + PEHSPNON + PENATVTY + PEMNTVTY + PRDTRACE
+ PEMLR + PRDISFLG + A_SEX + A_ENRLW + A_MARITL +  A_HGA + A_AGE + A_MJIND
+ A_MJOCC + ERN_OTR + CAP_VAL + DIV_VAL + FILESTAT + STATETAX_A + A_CLSWKR
+ PRUNTYPE + A_WKSTAT + PARENT + HHDFMX + HHDREL + MIGSAME + WKSWORK
+ VET_YN + FRMOTR + A_UNMEM + NOEMP]

# Removing "MIFSAME" since coefficient is not significant, adding the target variable

[OED_TYP1, OED_TYP2, PEHSPNON, PENATVTY, PEMNTVTY, PRDTRACE, PEMLR, PRDISFLG, A_SEX, A_ENRLW, A_MARITL, A_HGA, A_AGE, A_MJIND, A_MJOCC, ERN_OTR, CAP_VAL, DIV_VAL, FILESTAT, STATETAX_A, A_CLSWKR, PRUNTYPE, A_WKSTAT, PARENT, HHDFMX, HHDREL, WKSWORK, VET_YN, FRMOTR, A_UNMEM, NOEMP, Income]

c) Embedded Method
Sequential Forward Selection: (Using RandomForestClassifier, cv=10)
Selected 29 attributes +Target variable “Income”

['OED_TYP1', 'OED_TYP2', 'OED_TYP3', 'PEHSPNON', 'PENATVTY', 'PRDTRACE', 'PEMLR', 'A_SEX', 'A_HGA', 'A_AGE', 'A_MJIND', 'A_MJOCC', 'A_HRSPAY', 'ERN_OTR', 'CAP_VAL', 'DIV_VAL', 'FILESTAT', 'STATETAX_A', 'STATETAX_B', 'A_CLSWKR', 'PRUNTYPE', 'A_WKSTAT', 'PARENT', 'HHDFMX', 'WKSWORK', 'VET_QVA', 'VET_YN', 'A_UNMEM', 'NOEMP', 'Income']

## Detail Data Dictionary

| Number | Data Name | Description | Type | Details |
|---|---|---|---|---|
| 1 | OED_TYP1 | Other government assistance | Nominal | Values: 0 = niu 1 = yes 2 = no |
| 2 | OED_TYP2 | Scholarships, grants etc. from the school | Nominal | Values: 0 = niu 1 = yes 2 = no |
| 3 | OED_TYP3 | Other assistance (employers friends, etc.) | Nominal | Values: 0 = niu 1 = yes 2 = no |
| 4 | PEHSPNON | Spanish, Hispanic, or Latino? | Nominal | Values: 1 = Yes 2 = No |
| 5 | PENATVTY | Country of birth | Nominal | Number of Country:162 |
| 6 | PEMNTVTY | Mother's country of birth | Nominal | Number of Country:162 |
| 7 | PEFNTVTY | Father's country of birth | Nominal | Number of Country:162 |
| 8 | PRDTRACE | Race | Nominal | Values:<br>01 = White only<br>02 = Black only<br>03 = American Indian, Alaskan Native only (AI) 04 = Asian only<br>05 = Hawaiian/Pacific Islander only (HP)<br>06 = White-Black<br>07 = White-AI<br>08 = White-Asian<br>09 = White-HP<br>10 = Black-AI<br>11 = Black-Asian<br>12 = Black-HP<br>13 = AI-Asian<br>14 = AI-HP<br>15 = Asian-HP<br>16 = White-Black-AI<br>17 = White-Black-Asian<br>18 = White-Black-HP<br>19 = White-AI-Asian<br>20 = White-AI-HP<br>21 = White-Asian-HP<br>22 = Black-AI-Asian<br>23 = White-Black-AI-Asian<br>24 = White-AI-Asian-HP<br>25 = Other 3 race comb.<br>26 = Other 4 or 5 race comb. |
| 9 | PRCITSHP | Citizenship Group | Nominal | Values:<br>1 = Native, born in US<br>12 = Native, born in PR or US outlying area |

| | | | | |
|---|---|---|---|---|
| | | | | 3 = Native, born abroad of US parent(s)<br>4 = Foreign born, US cit by naturalization<br>5 = Foreign born, not a US citizen |
| 10 | NOEMP | Total number of persons who work for employers in all locations | Nominal | Values:<br>0 = niu<br>1 = under 10<br>2 = 10 - 24<br>3 = 25 - 99<br>4 = 100 - 499<br>5 = 500 - 999<br>6 = 1000+ |
| 11 | A_UNMEM | Member of a labor union or of an employee association | Nominal | Values:<br>0 = Not in universe or children and Armed Forces<br>1 = Yes<br>2 = No |
| 12 | FRMOTR | Receiving farm self-employment from secondary source | Nominal | Values:<br>0 = niu<br>1 = yes<br>2 = no |
| 13 | PTOTVAL | Total persons income | Numeric | Values:<br>0 = none<br>negative amt = income (loss)<br>positive amt = income<br>Universe: All Persons aged 15+ |
| 14 | VET_YN | Receive veterans' payments? | Nominal | Values:<br>0 = niu<br>1 = yes<br>2 = no |
| 15 | VET_QVA | Fill out an annual income questionnaire for the veteran's administration? | Nominal | Values:<br>0 = niu<br>1 = yes<br>2 = no |
| 16 | WKSWORK | Number of weeks worked | Numeric | Values:<br>0 = niu<br>1 = 1 week ...<br>52 = 52 weeks |
| 17 | MIG_MTR3 | Within area moved | Nominal | Values:<br>1 = Nonmover<br>2 = Same county<br>3 = Different county, same state<br>4 = Different state, same division<br>5 = Different division, same region<br>6 = Different region<br>7 = Abroad |

| | | | | |
|---|---|---|---|---|
| | | | | 8 = Not in universe (children under 1 yr old) |
| 18 | MIGSAME | Living in same place last 1 year | Nominal | Values:<br>0 = niu<br>1 = yes (nonmover)<br>2 = no, different house in u.s. (mover)<br>3 = no, outside the u.s. (mover) |
| 19 | HHDREL | Detailed household summary | Nominal | Values:<br>In household:<br>1 = Householder<br>2 = Spouse of householder<br>Child of householder:<br>3 = Under 18 years, single (never married) 4 = Under 18 years, ever married<br>5 = 18 years and over Other household members:<br>6 = Other relative of householder<br>7 = Nonrelative of householder<br>In group quarters:<br>8 = Secondary individual |
| 20 | HHDFMX | Detailed household and family status | Nominal | Values:<br>In primary family:<br>01 = Householder<br>02 = Spouse of householder<br>Child of householder: Under 18, single (never married):<br>03 = Reference person of subfamily<br>04 = Not in a subfamily Under 18, ever-married:<br>05 = Reference person of subfamily<br>06 = Spouse of subfamily reference person<br>07 = Not in a subfamily 18 years and over, single (never married):<br>08 = Head of a subfamily<br>09 = Not in a subfamily 18 years and over, ever-married:<br>10 = Reference person of subfamily<br>11 = Spouse of subfamily reference person<br>12 = Not in a subfamily Grandchild of householder:<br>Under 18, single (never married):<br>23 = Reference person of subfamily<br>24 = Child of a subfamily |

| | | | | |
|---|---|---|---|---|
| | | | | 25 = Not in a subfamily Under 18, ever-married:<br>26 = Reference person of subfamily<br>27 = Spouse of subfamily reference person<br>28 = Not used<br>29 = Not in a subfamily 18 years and over, single (never married):<br>30 = Reference person of a subfamily<br>31 = Not in a subfamily 18 years and over, ever-married:<br>32 = Reference person of subfamily<br>33 = Spouse of subfamily reference person<br>34 = Not in a subfamily Other relative of householder:<br>Under 18, single (never married):<br>35 = Reference person of subfamily<br>36 = Child of subfamily reference person<br>37 = Not in a subfamily Under 18, ever-married:<br>38 = Reference person of subfamily<br>39 = Spouse of subfamily reference person<br>40 = Not in a subfamily 18 years and over, single (never married):<br>41 = Reference person of a subfamily<br>42 = Not in a subfamily 18 years and over, ever-married:<br>43 = Reference person of subfamily<br>44 = Spouse of subfamily reference person<br>45 = Not in a subfamily In unrelated subfamily: 46 = Reference person of unrelated subfamily 47 = Spouse of unrelated subfamily reference person<br>48 = Child < 18, single (never married) of unrelated subfamily reference person Not in a family:<br>49 = Nonfamily householder<br>50 = Secondary individual<br>51 = In group quarters |
| 21 | PARENT | Presence of parents | Nominal | Values:<br>0 = Not in universe |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  |  |  | 1 = Both parents present<br>2 = Mother only present<br>3 = Father only present<br>4 = Neither parent present<br>Universe: Family members under 18 |
| 22 | A_WKSTAT | Full/part-time status | Nominal | Values:<br>0 = Children or Armed Forces<br>1 = Not in labor force<br>2 = Full-time schedules<br>3 = Part-time for economic reasons, usually FT<br>4 = Part-time for non-economic reasons, usually PT<br>5 = Part-time for economic reasons, usually PT<br>6 = Unemployed FT<br>7 = Unemployed PT |
| 23 | PRUNTYPE | Reason for unemployment | Nominal | Values:<br>0 = NIU<br>1 = Job loser/on layoff<br>2 = Other job loser<br>3 = Temporary job ended<br>4 = Job leaver<br>5 = Re-entrant<br>6 = New-entrant |
| 24 | A_CLSWKR | Class of worker | Nominal | Values:<br>0 = Not in universe or children and Armed Forces<br>1 = Private<br>2 = Federal government<br>3 = State government<br>4 = Local government<br>5 = Self-employed-incorporated<br>6 = Self-employed-not incorporated<br>7 = Without pay<br>8 = Never worked |
| 25 | STATETAX_B | State income tax liability, before credits | Numeric | Values: 0 = none; dollar amount |
| 26 | STATETAX_A | State income tax liability, after all credits | Numeric | Values: 0 = none; dollar amount |
| 27 | FILESTAT | Tax filer status | Nominal | Values:<br>1 = joint, both<br>2 = joint, one ><65 & one<br>3 = joint, both 65+<br>4 = head of household |

| | | | | |
|---|---|---|---|---|
| | | | | 5 = single<br>6 = non-filer |
| 28 | DIV_VAL | Dividends from stocks or mutual funds | Numeric | Values:<br>0 = none or niu<br>1-999999 = dividends |
| 29 | CAP_VAL | Capital gains value | Numeric | Values:<br>0 = none or niu<br>1-999999 = captial gains amount |
| 30 | ERN_OTR | Wage & salary from other work | Nominal | Values:<br>0 = niu<br>1 = yes<br>2 = no |
| 31 | A_HRSPAY | Income per hour | Numeric | Values:<br>0000 = Not in universe or children and Armed Forces<br>0001-9999 = Entry (2 implied decimal places) |
| 32 | A_MJOCC | Major occupation recode | Nominal | Values:<br>0 = Not in universe or children<br>1 = Management, business, and financial occupations<br>2 = Professional and related occupations<br>3 = Service occupations<br>4 = Sales and related occupations<br>5 = Office and administrative support occupations<br>6 = Farming, fishing, and forestry occupations<br>7 = Construction and extraction occupations<br>8 = Installation, maintenance, and repair occupations<br>9 = Production occupations<br>10 = Transportation and material moving occupations<br>11 = Military specific occupations |
| 33 | A_MJIND | Major industry code | Nominal | Values:<br>0 = Not in universe, or children<br>1 = Agriculture, forestry,fishing, and hunting<br>2 = Mining, quarrying, and oil and gas extraction<br>3 = Construction<br>4 = Manufacturing<br>5 = Wholesale and retail trade |

| | | | | |
|---|---|---|---|---|
| | | | | 6 = Transportation, warehousing and utilities<br>7 = Information<br>8 = Finance and insurance, and real estate and rental and leasing<br>9 = Professional, scientific, management and adminstrative, and waste mangement services<br>10 = Educational services, and health care and social assistance<br>11 = Arts, entertainment, recreation and accomodation, and food services<br>12 = Other services, except public adminstration<br>13 = Public administration<br>14 = Military |
| 34 | A_HGA | Educational attainment | Nominal | Values:<br>0 = Children<br>31 = Less than 1st grade<br>32 = 1st,2nd,3rd,or 4th grade<br>33 = 5th or 6th grade<br>34 = 7th and 8th grade<br>35 = 9th grade 36 = 10th grade<br>37 = 11th grade<br>38 = 12th grade no diploma<br>39 = High school graduate - high school diploma or equivalent<br>40 = Some college but no degree<br>41 = Associate degree in college - occupation/vocation program<br>42 = Associate degree in college - academic program<br>43 = Bachelor's degree (for example: BA,AB,BS)<br>44 = Master's degree (for example:MA,MS,MENG,MED,MSW, MBA)<br>45 = Professional school degree (for example: MD,DDS,DVM,LLB,JD)<br>46 = Doctorate degree (for example: PHD,EDD) |
| 35 | A_MARITL | Marital status | Nominal | Values:<br>1 = Married - civilian spouse present<br>2 = Married - AF spouse present<br>3 = Married - spouse absent (exc.separated) |

| | | | | |
|---|---|---|---|---|
| | | | | 4 = Widowed<br>5 = Divorced<br>6 = Separated<br>7 = Never married |
| 36 | A_ENRLW | Last week was attending or enrolled in a high school, college or university | Nominal | Values:<br>0 = Not in universe or children and Armed Forces<br>1 = Yes<br>2 = No |
| 37 | A_SEX | Sex | Nominal | Values:<br>1 = Male<br>2 = Female |
| 38 | PRDISFLG | Any disability conditions? | Nominal | Values:<br>-1 = NIU<br>1 = Yes<br>2 = No |
| 39 | PEMLR | Major labor force recode | Nominal | Values:<br>0 = NIU<br>1 = Employed - at work<br>2 = Employed - absent<br>3 = Unemployed - on layoff<br>4 = Unemployed - looking<br>5 = Not in labor force - retired<br>6 = Not in labor force - disabled<br>7 = Not in labor force - other |
| 40 | A_AGE | Age | Numeric | Values:<br>00-79 = 0-79 years of age<br>80 = 80-84 years of age<br>85 = 85+ years of age |