PUD (13 vars): Correct the k-anon violations (2)

Summarize existing utility

```
quick <- quick_summary(data, label="all_fields", qis=quasi_identifiers)
```

```
## [1] "Quick summary:"
##                    all_fields  quasi_fields
## total_fields       13          4
## total_records      23,940,221  23,940,221
## total_cells        311,222,873 95,760,884
## missing_fields     62,283,866  4,516,843
## missing_pct        20%         5%
## complete_fields    248,939,007 91,244,041
## complete_pct       80%         95%
## unknown_fields     22,387,406  10,956,426
## unknown_pct        7%          11%
## suppressed_fields  444         444
## suppressed_pct     0%          0%
## available_fields   226,551,157 80,287,171
## available_pct      73%         84%
```

Summarize existing utility

```
utility <- summarize_utility(data, quasi_identifiers)
```

```
## Utility summary:
## Total records in dataset: 23,940,221
##                         suppressed suppressed_percent   missing missing_percent
## sex                             72               0.0%    52,971             0.2%
## age_group                      326               0.0%   170,203             0.7%
## race                            44               0.0% 2,197,145             9.2%
## ethnicity                        2               0.0% 2,096,524             8.8%
## records_with_any_field         326               0.0% 3,794,263            15.8%
```

Recoding all the "NA" (already suppressed), Missings and Unknowns to NA for purposes of k-anonymity

```
data_na <- recode_to_na(data,quasi_identifiers,BLANK_CATEGORIES)
```

Set up sdcMicro object, using data_na and change alpha=c(0)) to alpha=c(1)) to correct a k-anon violations (2) problem:

Output from using data before recode to na and alpha=c(0)): a k-anon violations (2) for k=(5) and quasi-identifiers (race ethnicity sex age_group). *If greater than zero violations, then here's 5 violations. race ethnicity sex age_group fk* 15600394 NA NA NA NA 2 *15600393 NA NA NA NA 2

```
sdcObj <- createSdcObj(dat=data_na,
                       keyVars=quasi_identifiers,
                       numVars=NULL,
                       weightVar=NULL,
                       hhId=NULL,
                       strataVar=NULL,
                       pramVars=NULL,
                       excludeVars=NULL,
                       seed=0,
                       randomizeRecords=FALSE,
                       alpha=c(1))

# print to confirm observations, num variales, quasis, quasi describes, and risk info
sdc_print(sdcObj, KANON_LEVEL)
```

```
## SDC summary for k-anon-level( 5 ).

## The input dataset consists of 23940221 rows and 13 variables.
##   --> Categorical key variables: sex, age_group, race, ethnicity
## ----------------------------------------------------------------------

## Information on categorical key variables:
##
## Reported is the number, mean size and size of the smallest category >0 for recoded variables.
## In parenthesis, the same statistics are shown for the unmodified data.
## Note: NA (missings) are counted as seperate categories!

##  Key Variable Number of categories        Mean size
##           sex                    4  (4) 7891235.667 (7891235.667)
##     age_group                   10 (10) 2641076.889 (2641076.889)
##          race                    7  (7) 2909931.000 (2909931.000)
##     ethnicity                    3  (3) 7692093.000 (7692093.000)
##  Size of smallest (>0)
##                     443     (443)
##                  947997  (947997)
##                   58302   (58302)
##                 4442351 (4442351)
```

```
## ------------------------------------------------------------------------
```

```
## Risk measures:
##
## Number of observations with higher risk than the main part of the data: 0
## Expected number of re-identifications: 30.23 (0.00 %)
```

Print out the number of violations and a sample: k-anon violations should be zero

```
fk = summarize_violations(data_na, sdcObj, KANON_LEVEL, quasi_identifiers)
```

```
## k-anon violations ( 0 ) for k=( 5 ) and quasi-identifiers ( sex age_group race ethnicity ). If greater than zero violations, then here's 5
```