Imputation & Weighting example

Imputation

This is an example code of missing data imputation using MICE package.

```
## Import example data
data <- airquality
data_sub <- data[,c(1,2)]
summary(data_sub)
```

```
##      Ozone           Solar.R
##  Min.   :  1.00   Min.   :  7.0
##  1st Qu.: 18.00   1st Qu.:115.8
##  Median : 31.50   Median :205.0
##  Mean   : 42.13   Mean   :185.9
##  3rd Qu.: 63.25   3rd Qu.:258.8
##  Max.   :168.00   Max.   :334.0
##  NA's   :37       NA's   :7
```

```
## Quick view of missing data
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(data_sub, 2, pMiss)
```

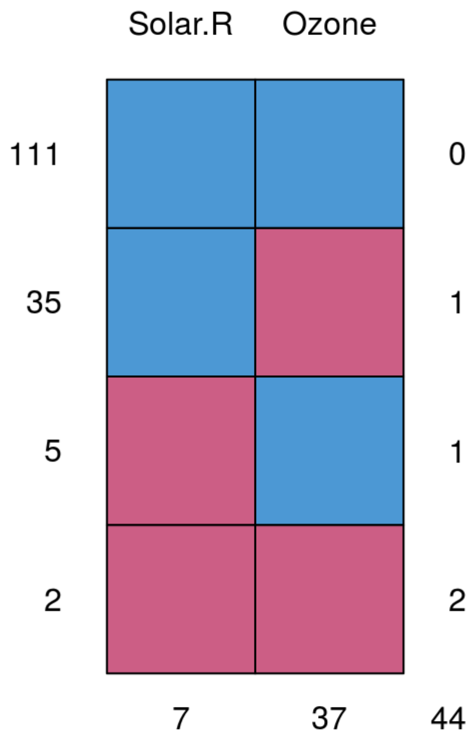
```
##     Ozone   Solar.R
## 24.183007  4.575163
```

```
## Look at missing data pattern
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
md.pattern(data_sub)
```



```
##     Solar.R Ozone
## 111       1     1  0
## 35        1     0  1
## 5         0     1  1
## 2         0     0  2
##           7    37 44
```

Results indicated that 111 observations have complete data, 35 observations have missing data in variable Ozone, 5 observations have missing data in variable Solar.R, and 2 observations have missing data on both variables.

```
## Look at missing data pattern with more plots
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

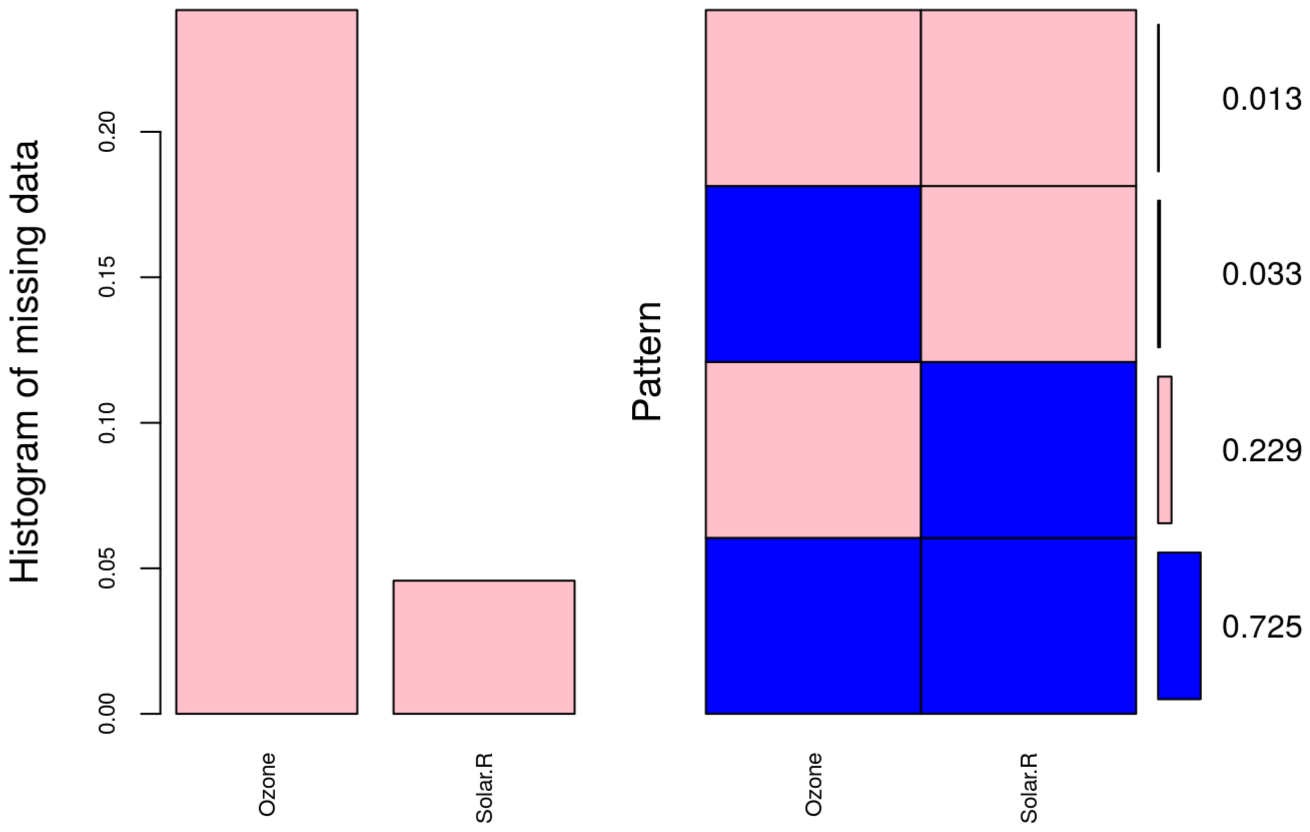
```
## Loading required package: data.table
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issu
es
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
plot1 <- aggr(data_sub, col=c('blue','pink'), number=TRUE,
              sortVars=TRUE, labels=names(data_sub), cex.axis=.7,
              gap=3, ylab=c("Histogram of missing data",
                            "Pattern"))
```



```
##
##  Variables sorted by number of missings:
##  Variable      Count
##     Ozone 0.24183007
##   Solar.R 0.04575163
```

The plots showed that 72.50% of data are complete, 22.90% of data has missing value on Solar.R variable, 3.30% of data has missing value on Ozone variable, and 1.30% of data has missing values on both variables.

```
## Imputing missing data
# using predictive mean matching imputation
data_impute <- mice(data_sub, m=5, meth='pmm', maxit=50, seed=500)
```

```
##
##  iter imp variable
##   1   1  Ozone  Solar.R
##   1   2  Ozone  Solar.R
##   1   3  Ozone  Solar.R
##   1   4  Ozone  Solar.R
##   1   5  Ozone  Solar.R
##   2   1  Ozone  Solar.R
##   2   2  Ozone  Solar.R
##   2   3  Ozone  Solar.R
##   2   4  Ozone  Solar.R
##   2   5  Ozone  Solar.R
##   3   1  Ozone  Solar.R
##   3   2  Ozone  Solar.R
##   3   3  Ozone  Solar.R
##   3   4  Ozone  Solar.R
##   3   5  Ozone  Solar.R
##   4   1  Ozone  Solar.R
##   4   2  Ozone  Solar.R
##   4   3  Ozone  Solar.R
##   4   4  Ozone  Solar.R
##   4   5  Ozone  Solar.R
##   5   1  Ozone  Solar.R
##   5   2  Ozone  Solar.R
##   5   3  Ozone  Solar.R
##   5   4  Ozone  Solar.R
##   5   5  Ozone  Solar.R
##   6   1  Ozone  Solar.R
##   6   2  Ozone  Solar.R
##   6   3  Ozone  Solar.R
##   6   4  Ozone  Solar.R
##   6   5  Ozone  Solar.R
##   7   1  Ozone  Solar.R
##   7   2  Ozone  Solar.R
##   7   3  Ozone  Solar.R
##   7   4  Ozone  Solar.R
##   7   5  Ozone  Solar.R
##   8   1  Ozone  Solar.R
##   8   2  Ozone  Solar.R
##   8   3  Ozone  Solar.R
##   8   4  Ozone  Solar.R
##   8   5  Ozone  Solar.R
##   9   1  Ozone  Solar.R
##   9   2  Ozone  Solar.R
##   9   3  Ozone  Solar.R
##   9   4  Ozone  Solar.R
##   9   5  Ozone  Solar.R
##   10   1  Ozone  Solar.R
##   10   2  Ozone  Solar.R
##   10   3  Ozone  Solar.R
##   10   4  Ozone  Solar.R
##   10   5  Ozone  Solar.R
##   11   1  Ozone  Solar.R
##   11   2  Ozone  Solar.R
##   11   3  Ozone  Solar.R
##   11   4  Ozone  Solar.R
##   11   5  Ozone  Solar.R
##   12   1  Ozone  Solar.R
##   12   2  Ozone  Solar.R
##   12   3  Ozone  Solar.R
##   12   4  Ozone  Solar.R
##   12   5  Ozone  Solar.R
##   13   1  Ozone  Solar.R
##   13   2  Ozone  Solar.R
##   13   3  Ozone  Solar.R
##   13   4  Ozone  Solar.R
##   13   5  Ozone  Solar.R
##   14   1  Ozone  Solar.R
##   14   2  Ozone  Solar.R
##   14   3  Ozone  Solar.R
##   14   4  Ozone  Solar.R
##   14   5  Ozone  Solar.R
##   15   1  Ozone  Solar.R
##   15   2  Ozone  Solar.R
##   15   3  Ozone  Solar.R
##   15   4  Ozone  Solar.R
##   15   5  Ozone  Solar.R
##   16   1  Ozone  Solar.R
##   16   2  Ozone  Solar.R
##   16   3  Ozone  Solar.R
##   16   4  Ozone  Solar.R
##   16   5  Ozone  Solar.R
##   17   1  Ozone  Solar.R
##   17   2  Ozone  Solar.R
##   17   3  Ozone  Solar.R
##   17   4  Ozone  Solar.R
##   17   5  Ozone  Solar.R
##   18   1  Ozone  Solar.R
##   18   2  Ozone  Solar.R
##   18   3  Ozone  Solar.R
##   18   4  Ozone  Solar.R
##   18   5  Ozone  Solar.R
##   19   1  Ozone  Solar.R
##   19   2  Ozone  Solar.R
##   19   3  Ozone  Solar.R
##   19   4  Ozone  Solar.R
##   19   5  Ozone  Solar.R
##   20   1  Ozone  Solar.R
##   20   2  Ozone  Solar.R
##   20   3  Ozone  Solar.R
##   20   4  Ozone  Solar.R
##   20   5  Ozone  Solar.R
##   21   1  Ozone  Solar.R
##   21   2  Ozone  Solar.R
##   21   3  Ozone  Solar.R
##   21   4  Ozone  Solar.R
##   21   5  Ozone  Solar.R
##   22   1  Ozone  Solar.R
##   22   2  Ozone  Solar.R
##   22   3  Ozone  Solar.R
##   22   4  Ozone  Solar.R
##   22   5  Ozone  Solar.R
##   23   1  Ozone  Solar.R
##   23   2  Ozone  Solar.R
##   23   3  Ozone  Solar.R
##   23   4  Ozone  Solar.R
##   23   5  Ozone  Solar.R
##   24   1  Ozone  Solar.R
##   24   2  Ozone  Solar.R
##   24   3  Ozone  Solar.R
##   24   4  Ozone  Solar.R
##   24   5  Ozone  Solar.R
##   25   1  Ozone  Solar.R
##   25   2  Ozone  Solar.R
##   25   3  Ozone  Solar.R
##   25   4  Ozone  Solar.R
##   25   5  Ozone  Solar.R
##   26   1  Ozone  Solar.R
##   26   2  Ozone  Solar.R
##   26   3  Ozone  Solar.R
##   26   4  Ozone  Solar.R
##   26   5  Ozone  Solar.R
##   27   1  Ozone  Solar.R
##   27   2  Ozone  Solar.R
##   27   3  Ozone  Solar.R
##   27   4  Ozone  Solar.R
##   27   5  Ozone  Solar.R
##   28   1  Ozone  Solar.R
##   28   2  Ozone  Solar.R
##   28   3  Ozone  Solar.R
##   28   4  Ozone  Solar.R
##   28   5  Ozone  Solar.R
##   29   1  Ozone  Solar.R
##   29   2  Ozone  Solar.R
##   29   3  Ozone  Solar.R
##   29   4  Ozone  Solar.R
##   29   5  Ozone  Solar.R
##   30   1  Ozone  Solar.R
##   30   2  Ozone  Solar.R
##   30   3  Ozone  Solar.R
##   30   4  Ozone  Solar.R
##   30   5  Ozone  Solar.R
##   31   1  Ozone  Solar.R
##   31   2  Ozone  Solar.R
##   31   3  Ozone  Solar.R
##   31   4  Ozone  Solar.R
##   31   5  Ozone  Solar.R
##   32   1  Ozone  Solar.R
##   32   2  Ozone  Solar.R
##   32   3  Ozone  Solar.R
##   32   4  Ozone  Solar.R
##    32   5  Ozone  Solar.R
##    33   1  Ozone  Solar.R
##    33   2  Ozone  Solar.R
##    33   3  Ozone  Solar.R
##    33   4  Ozone  Solar.R
##    33   5  Ozone  Solar.R
##    34   1  Ozone  Solar.R
##    34   2  Ozone  Solar.R
##    34   3  Ozone  Solar.R
##    34   4  Ozone  Solar.R
##    34   5  Ozone  Solar.R
##    35   1  Ozone  Solar.R
##    35   2  Ozone  Solar.R
##    35   3  Ozone  Solar.R
##    35   4  Ozone  Solar.R
##    35   5  Ozone  Solar.R
##    36   1  Ozone  Solar.R
##    36   2  Ozone  Solar.R
##    36   3  Ozone  Solar.R
##    36   4  Ozone  Solar.R
##    36   5  Ozone  Solar.R
##    37   1  Ozone  Solar.R
##    37   2  Ozone  Solar.R
##    37   3  Ozone  Solar.R
##    37   4  Ozone  Solar.R
##    37   5  Ozone  Solar.R
##    38   1  Ozone  Solar.R
##    38   2  Ozone  Solar.R
##    38   3  Ozone  Solar.R
##    38   4  Ozone  Solar.R
##    38   5  Ozone  Solar.R
##    39   1  Ozone  Solar.R
##    39   2  Ozone  Solar.R
##    39   3  Ozone  Solar.R
##    39   4  Ozone  Solar.R
##    39   5  Ozone  Solar.R
##    40   1  Ozone  Solar.R
##    40   2  Ozone  Solar.R
##    40   3  Ozone  Solar.R
##    40   4  Ozone  Solar.R
##    40   5  Ozone  Solar.R
##    41   1  Ozone  Solar.R
##    41   2  Ozone  Solar.R
##    41   3  Ozone  Solar.R
##    41   4  Ozone  Solar.R
##    41   5  Ozone  Solar.R
##    42   1  Ozone  Solar.R
##    42   2  Ozone  Solar.R
##    42   3  Ozone  Solar.R
##    42   4  Ozone  Solar.R
##    42   5  Ozone  Solar.R
##    43   1  Ozone  Solar.R
##    43   2  Ozone  Solar.R
##    43   3  Ozone  Solar.R
##   43   4  Ozone  Solar.R
##   43   5  Ozone  Solar.R
##   44   1  Ozone  Solar.R
##   44   2  Ozone  Solar.R
##   44   3  Ozone  Solar.R
##   44   4  Ozone  Solar.R
##   44   5  Ozone  Solar.R
##   45   1  Ozone  Solar.R
##   45   2  Ozone  Solar.R
##   45   3  Ozone  Solar.R
##   45   4  Ozone  Solar.R
##   45   5  Ozone  Solar.R
##   46   1  Ozone  Solar.R
##   46   2  Ozone  Solar.R
##   46   3  Ozone  Solar.R
##   46   4  Ozone  Solar.R
##   46   5  Ozone  Solar.R
##   47   1  Ozone  Solar.R
##   47   2  Ozone  Solar.R
##   47   3  Ozone  Solar.R
##   47   4  Ozone  Solar.R
##   47   5  Ozone  Solar.R
##   48   1  Ozone  Solar.R
##   48   2  Ozone  Solar.R
##   48   3  Ozone  Solar.R
##   48   4  Ozone  Solar.R
##   48   5  Ozone  Solar.R
##   49   1  Ozone  Solar.R
##   49   2  Ozone  Solar.R
##   49   3  Ozone  Solar.R
##   49   4  Ozone  Solar.R
##   49   5  Ozone  Solar.R
##   50   1  Ozone  Solar.R
##   50   2  Ozone  Solar.R
##   50   3  Ozone  Solar.R
##   50   4  Ozone  Solar.R
##   50   5  Ozone  Solar.R
```

```
summary(data_impute)
```

```
## Class: mids
## Number of multiple imputations:  5 
## Imputation methods:
##   Ozone Solar.R 
##   "pmm"   "pmm" 
## PredictorMatrix:
##         Ozone Solar.R
## Ozone       0       1
## Solar.R     1       0
## Check the imputed data for variables
data_impute$imp$Ozone
```

	1	2	3	4	5
	<int>	<int>	<int>	<int>	<int>
5	6	118	52	39	23
10	27	85	82	73	78
25	22	35	18	7	8
26	76	108	37	24	40
27	23	32	80	45	20
32	115	89	18	37	11
33	66	89	18	37	11
34	14	110	24	32	28
35	12	91	76	73	44
36	85	76	24	110	46

1-10 of 37 rows Previous **1** 2 3 4 Next

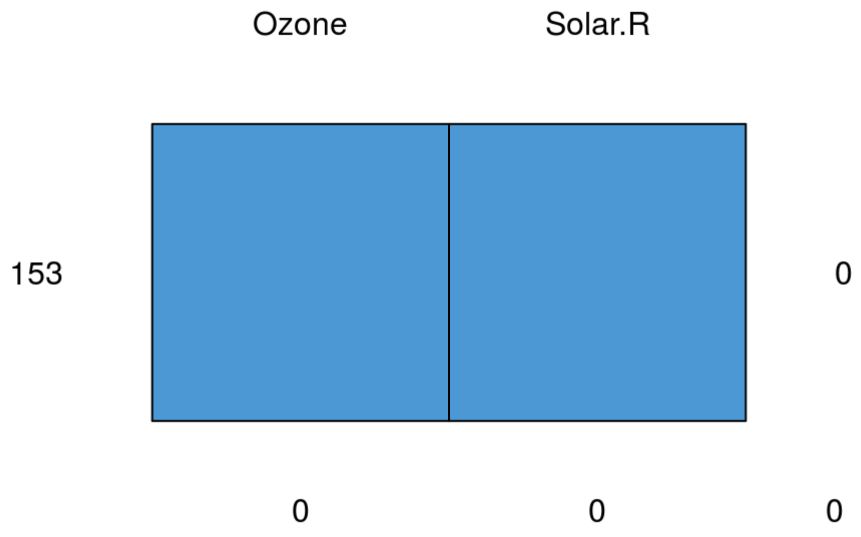
```
data_impute$imp$Solar.R
```

	1	2	3	4	5
	<int>	<int>	<int>	<int>	<int>
5	78	188	258	322	252
6	13	259	139	153	273
11	25	24	259	8	8
27	99	148	253	323	59
96	213	51	157	203	167
97	92	273	314	238	250
98	291	237	285	253	213

7 rows

```
## Get the completed dataset after imputing missing data
data_complete <- complete(data_impute, 1)
md.pattern(data_complete)
```

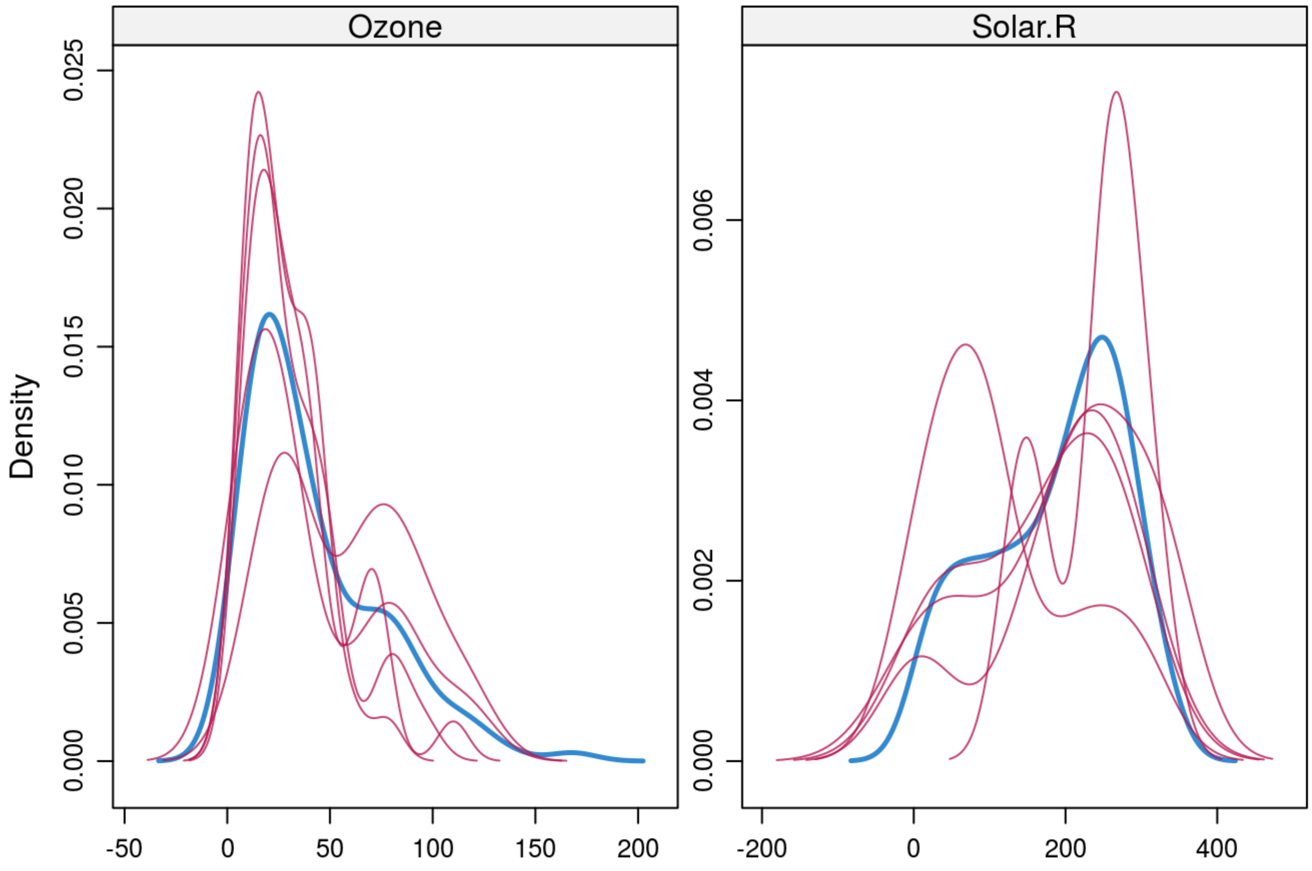
```
##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```



```
##     Ozone Solar.R  
## 153     1       1 0
##         0       0 0
```

Results indicated that now there is no missing value in the dataset.

```
## Inspecting the distribution of original and imputed data
densityplot(data_impute)
```



We expect the distributions of original and imputed dataset to be similar. The distribution of original dataset is plotted in blue and the imputed dataset is plotted in red. As the plots showed, for both variables, the distributions of original and imputed dataset are similar.

Weighting

The following code of survey weights will be demonstrated using a fake dataset. This weighting process will be mimic the situation where we have certain demongraphic variables (gender in the following analysis) may not be representative of the population. The assumption is that the population variable is available.

```
## Generate variables and created dataset
set.seed(12345)
gender = c("Female", "Male")
gender = sample(gender, 100, replace = TRUE)
gender = as.numeric(factor(gender))
ethnicity = c("White", "African_American", "Other")
ethnicity = sample(ethnicity, 100, replace = TRUE)
ethnicity = as.numeric(factor(ethnicity))
income = c(0:100000)
income = sample(income, 100, replace = TRUE)

data = cbind(gender,ethnicity,income)
data = as.data.frame(data)

## Create unweighted dataset with survey package
library(survey)
```

```
## Loading required package: Matrix
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
data.svy.unweighted <- svydesign(ids=~1, data=data)
```

Next, the weighting will be performed based on the population gender probabilities. I assume the population values for female (1) and male(2) are .45 and .55.

```
gender.dist <- data.frame(gender = c("1","2"),
                          Freq =nrow(data)*c(0.45, 0.55))
```

rake function in survey package will be used to weight the data by population gender values.

```
data.svy.rake <- rake(design = data.svy.unweighted,
                      sample.margins = list(~gender),
                      population.margins = list(gender.dist))
```

In case the weights is too large or too small, I put limits on the weights using the trimWeights function.

```
data.svy.rake.trim <- trimWeights(data.svy.rake,
                                  lower=0.3, upper=3,
                                  strict=TRUE)
```

Next, I’m going to compare mean of variables in the weighted dataset and the original dataset.

```
svymean(data, data.svy.rake.trim) # weighted dataset
```

```
##                 mean        SE
## gender         1.550    0.0000
## ethnicity      2.018    0.0796
## income     47276.462 2785.5155
```

```
apply(data, 2, mean) # original dataset
```

```
##    gender ethnicity    income
##      1.54      2.02  47201.65
```