These exercises use data from a study of mathematics skill levels in students within classrooms within schools. The data are available in a comma-separated value (csv) file as

`http://www-personal.umich.edu/~bwest/classroom.csv`

1. Start R and create a data frame called `classroom` using the `read.csv` function with the (quoted) file name as the first and only argument. Notice that you can do this in two ways: either read directly from the web site by giving the quoted URL as the argument or first download the file then enter the file name. Try it both ways. Remember that `file.choose()` brings up a chooser panel to help you navigate to the file name.

2. Check the structure of the data frame using `str(classroom)`. Are any of the variables in the data frame stored as factors? Should any of these variables be stored as factors? The first few lines should look like

   ```
   > str(classroom)
   'data.frame':        1190 obs. of  12 variables:
    $ sex     : int  1 0 1 0 0 1 0 0 1 0 ...
    $ minority: int  1 1 1 1 1 1 1 1 1 1 ...
    $ mathkind: int  448 460 511 449 425 450 452 443 422 480 ...
    $ mathgain: int  32 109 56 83 53 65 51 66 88 -7 ...
   ```

3. Check the `summary` of this data frame. Is the summary of the `sex` variable meaningful? It happens that the coding for `sex` is

   0 Male
   1 Female

   Convert this variable to a `factor` with labels `"M"` and `"F"`. Check the summary for this variable after conversion. You can either ask for a summary of the whole data frame again or for a summary of this variable only using

   ```
   > summary(classroom$sex)
     M   F
   588 602
   ```

4. Convert the `minority` variable to a factor with levels `"N"` and `"Y"`. Convert the `childid`, `classid` and `schoolid` variables to factors (it is easiest to retain the numeric values for the levels). Check `str` and the `summary` again. The summary of the factor variables should be like

   ```
   > summary(subset(classroom, select = c(sex, minority, childid,
   +     classid, schoolid)))
    sex     minority    childid         classid         schoolid
    M:588   N:384    1      :   1    26     :  10    11     :  31
    F:602   Y:806    2      :   1    42     :  10    12     :  27
                     3      :   1    13     :   9    71     :  27
                     4      :   1    189    :   9    76     :  27
                     5      :   1    205    :   9    77     :  24
                     6      :   1    253    :   9    31     :  22
                     (Other):1184    (Other):1134    (Other):1032
   ```

5. Save the modified data frame as a file named `"classroom.rda"`. Remove the data frame. Load the file and check that the `classroom` data frame has the expected structure.

   ```
   > str(classroom)
   ```

```
'data.frame':       1190 obs. of  12 variables:
 $ sex     : Factor w/ 2 levels "M","F": 2 1 2 1 1 2 1 1 2 1 ...
 $ minority: Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ mathkind: int  448 460 511 449 425 450 452 443 422 480 ...
 $ mathgain: int  32 109 56 83 53 65 51 66 88 -7 ...
 $ ses     : num  0.46 -0.27 -0.03 -0.38 -0.03 0.76 -0.03 0.2 0.64 0.13 ...
 $ yearstea: num  1 1 1 2 2 2 2 2 2 2 ...
 $ mathknow: num  NA NA NA -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 ...
 $ housepov: num  0.082 0.082 0.082 0.082 0.082 0.082 0.082 0.082 0.082 0.082 ...
 $ mathprep: num  2 2 2 3.25 3.25 3.25 3.25 3.25 3.25 3.25 ...
 $ classid : Factor w/ 312 levels "1","2","3","4",..: 160 160 160 217 217 217 217 217 2
 $ schoolid: Factor w/ 107 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ childid : Factor w/ 1190 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

You might also consider these questions.

1. I claim that the `childid` variable serves no purpose because it is the same as the row number. Check this.
2. You can remove a variable from a data frame by assigning it the special value `NULL`. Try this

   ```
   > classroom$childid <- NULL
   > names(classroom)
    [1] "sex"      "minority" "mathkind" "mathgain" "ses"      "yearstea"
    [7] "mathknow" "housepov" "mathprep" "classid"  "schoolid"
   ```

3. The variable `housepov` should be associated with the `schoolid` factor. Check that there is only one value of `housepov` for each school.
4. The variables `yearstea`, `mathknow`, `mathprep` and `schoolid` should be a property of the `classid`. Check that they are.
5. Create two new data frames, `school` and `class`, containing one row for each level of `schoolid` (resp. one row for each level of `classid`) and the corresponding variable(s) that are associated with the school or class.
6. Look up the documentation of the `merge` function and see if you can merge the two data frames you created in the previous question into one frame that contains all the variables associated with the class plus the `housepov` associated with the class.