Lab 2: Data cleaning

Julieth Santamaria

February 1, 2018

1. Import nscg17. Remember that you need to change the working directory!

```
setwd("G:/My Drive/U of M/TA/TA APEC3003/APEC 3003 - 2019/APEC 3003 R work/labs/")
load("../data/nscg17.rdata")
```

   - How many observations does the dataset have? How many variables does it have? There are 83672 observations and 519 variables.

   - List the first 5 observations of the variable *age*

```
    nscg17[1:5,"age"]
```

```
## [1] 37 37 45 52 45
```

   - List the first 5 observations of the variables *age* and *gender*

```
    nscg17[1:5,c("age","gender")]
```

```
##   age gender
## 1  37      F
## 2  37      F
## 3  45      M
## 4  52      F
## 5  45      M
```

2. Look at the codebook and look for invalid values for the variable for earnings (`EARN`). Then, fix it. (Notice that *earnings* are different from *salary*).

Steps to answer this question:

- Step 1: Look at the codebook (or questionnaire). Codebooks and questionnaires and in the "codebooks" folder. Find the variable of interest. Note the type of variable you are dealing with and which are valid/invalide values for that variable.

```
# Look at variable EARN in the codebook. Notice that it is a numerical variable.
# The maximum value, according to the codebook, is 9999998.
class(nscg17$earn)   # This confirms it is a numerical variable
```

```
## [1] "numeric"
```

```
summary(nscg17$earn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   49000   82000 1290450  140000 9999998
```

Values that are not in the codebook should be transformed to NAs. So, all values higher than 9999998 should be transformed to NAs.

- Step 2: Clean the variable or create variables if requested

```
nscg17 <- within(nscg17, {
  earn[earn == 9999998] <- NA      # This set values larger than 9999998 as NAs
})
```

- Step 3: Check that your cleaning was done correctly

```r
summary(nscg17$earn)   # There should be some NAs now
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   45000   74000   96549  110000 1988018   10087
```

3. In the codebook find the variable `JOBSATIS`. What type of variable is it? Why is there a logical skip?

4. Generate a dummy for gender

```r
nscg17 <- within(nscg17, {
  female=NULL # This creates a new empty column
  female[gender=="F"]<-1 # female takes the value of 1 if gender is "F"
  female[gender=="M"]<-0 # female is zero if gender is "M"
})
```