Recoding Variables

The R interface is a just that - a basic interface. In order to obtain R, you must first download the software from the link provided here: https://www.r-project.org/. Then, in order to use the software, you must load a dataset and use packages and commands in order to tell the software how to evaluate and handle the data. Each package is open-sourced and contains different tools in order to implement all sorts of data manipulation techniques.

Note: The datasets used in this tutorial and the R Script are on Moodle:

Loading the 2016 CCES dataset

```
install.packages("foreign", dependencies=TRUE)
library(foreign)

dat <- read.dta(file.choose(), convert.factors=FALSE)
```

Understanding the Structure of Variables and Recoding

1. Basic Exploration of Variables

Variables are almost never in a readily useable format. Usually, there are missing values and people that did not answer questions. Therefore, the first step to any analysis is exploring the structure of the variables you want to use and making sure that they are accurately depicting the measurement scales intended.

EXAMPLE: Let us pretend that the thing I ultimately want to explore is how Americans view themselves ideologically (Very liberal - 0 to middle of the road - 4 to very conservative - 7). In the survey, the codebook indicates that the variable "CC16_340a" provides this information. I can view a summary of this information by telling R that I want a summary of the CC16_340a variable in the "dat" dataset that I loaded with the command below. Note: The dataset I am loading is titled "dat" and the variable in the dataset is "CC16_340a".

```
summary(dat$CC16_340a)
```

Discussion: The output indicates that the minimum value for the variable is 1, median is 4, mean is 4.267 and max is 8. Does this make sense? No. Remember, the ideology variable is measured from 0-7. Therefore, we need to continue our exploration.

Example: You could also make a table of the variable in order to get a better view at the categories of the variable.

```
table(dat$CC16_340a)
```

Discussion: As you can see, there is the category 8 that do not fit in our 1-7 scale. If we look in the codebook, 8 indicates "not sure." In order to successfully move on and perform an analysis, we will need to recode this variable in order to remove the categories.

2. Recoding Variables - Numerical

In most instances, we might want to recode data so that these people do not exist in our analysis or "Not Applicable (NA)". For recoding variables, the "car" package provides us tool to do this type of recoding.

```
install.packages("car", dependencies=TRUE)
library(car)
```

Example: As indicated before, it is necessary to recode the categories/number "8" as NA so that we can perform meaningful statistical tests on the data.

```
dat$ideology <- recode(dat$CC16_340a, "8=NA")
summary(dat$ideology)
table(dat$ideology)
```

Discussion: After recoding the variable, the summary of the variable indicates that we now have a minimum value of 1, max of 7, mean of 4.054, and 3,735 observations have been set to NA. As you can see, the new mean makes much more sense theoretically as it is closer to the middle.

Note: Sometimes, you need to tell R that a variable is a continuous variable rather than a variable is made up of characters after recoding. Meaning, R does not know that characters are numbers until you tell it. This is an easy process.

```
dat$ideology <- as.numeric(as.character(dat$ideology))
```

3. Recoding Variables - Categorical (nominal)

Another situation that might arise is that R might treat a categorical variable as numeric in ways that we do not wish. For example, a variable that explores profession might have several categories that cannot be ranked in a meaningful way numerically. Therefore, the variable should be recoded as a factor.

Example: For simplicity sake, here we are recoding the gender variable ("gender") to be categorical instead of numeric. In addition, if respondents refused to answer, we could do as we did above and recode other values to NA. The codebook indicates that for the variable a 1 indicates male and a 2 indicates female.

```
table(dat$gender)
dat$gender1 <- recode(dat$gender, "1='Man'; 2='Woman'")
dat$gender1 <- as.factor(dat$gender1)
table(dat$gender1)
```

Or, we could code it as numeric in a more meaningful way.

```
dat$gender2 <- recode(dat$gender, "1='0'; 2='1'")
dat$gender2 <- as.factor(dat$gender2)
table(dat$gender2)
```

4. Recoding Variables - Categorical (ordinal level)

With ordinal level variables you can leave them as numeric or recode them into factor variables. What is important is that you check the variable in order to make sure it makes sense. Here, we use the variables that asks how religious a respondent sees themselves as being (i.e. variables "pew_religimp"). The variable is coded as 1 = very important to 4 = not at all important. We want to make sure to order the variables in a more meaningful way (i.e. switch the number ordering) and make sure that people skipping the question are coded as NA.

```
table(dat$pew_religimp)

dat$religiosity <- recode(dat$pew_religimp, "4=0; 3=1; 2=2; 1=3; else=NA")
table(dat$religiosity
```

5. Recoding Variables - Special Situations and Issues

Here we load the 2020 European Social Survey Finland dataset for this example.
Note: This dataset is a more recent version of STATA (STATA 13), which requires a different code to load it.

```
library(readstata13)

dat1 <- read.dta13(file.choose(), convert.factors=TRUE)
```

Example: We make a table for the party voted for in the last national election in Finland. Here, we can basically see two major issues. The first is that there are special characters in the names of parties. The second is that there are empty categories like "don't know". There are numerous ways of dealing with these issues. I provide a few solutions below.

```
table(dat1$prtvtefi)
```

Getting rid of special characters requires changing the code used to recode the variables slightly. In particular, we swap the single quotes for double quotes, and double quotes for single quotes - as well as indicating to R that a character is special and not an aspect of the code. Let's say we wanted to remove the apostrophe from The Swedish People's Party category in the vote choice variable nad replace it with the abbreviation to make it more useable.

```
dat1$vote1 <- recode(dat1$prtvtefi, '"The Swedish People\'s Party (SPP)" = "SPP"')
table(dat1$vote1)
```

Let us say that we wanted to recode all of the party vote choice labels to be abbreviations so that we can more easily make readable tables. Here, you will note there is a subtle issue. In particular, there is an extra space after the "Christian Democrats" label, which we must account for in the recode commands. Remember, these datasets are constructed by humans and it is easy to include extra spaces or other issues. Therefore, a keen eye is necessary.

```
dat1$vote2 <- recode(dat1$vote1, "'Christian Democrats ' = 'CD'; 'Green League' =
'GL'; 'Left Alliance' = 'LA'; 'Social Democratic Party' = 'SDP'; 'SPP' = 'SPP';
'The Centre Party' = 'CP'; 'The National Coalition Party' = 'NCP'; 'True Finns' =
'TF'; 'Other' = 'Other'; 'Movement Now'= 'Other'; 'Communist Party' = 'Other';
'Pirate Party' = 'Other'; else=NA")

table(dat1$vote2)

dat1$vote2 <- factor(dat1$vote2, levels=c("LA", "SDP", "GL", "CP", "CD", "TF",
"SPP", "NCP", "Other"))
table(dat1$vote2)
summary(dat1$vote2)
```

There are other issues that could arise when recoding variables. What is important to remember is that the first step is almost always to look at the code and the labels of the variables. One small error can create issues. It is also important to note that for variables with long character labels that you can convert the labels to numeric in order to make the labels easier to deal with using the "as.numeric" command. In other words, keep in mind whether it is easier to work with labels with names or numbers.

Lab Activity

In the 2020 Finland European Social Survey dataset, you are to find the variable labels for the three variables provided below. Then, explore the three variables and recode them so that they are in a usable format. Explain what each variable conveys substantively.

Variable 1 - A measure for the religiosity of the respondent.

Variable 2 - Gender of the respondent.

Variable 3 - Standardized measure of the respondent's highest level of education.