

# Lecture 6 problem set

*INSERT YOUR NAME HERE*

*INSERT DATE*

## Contents

|                                   |   |
|-----------------------------------|---|
| Required reading and instructions | 1 |
| Load library and data             | 1 |
| Step 1: Investigate Variables     | 2 |

## Required reading and instructions

### Required reading

- GW 15.1-15.2
- Optional: GW 15.3-15.5
- Optional: GW 10, GW 20.6-20.7,

**General instructions** The purpose of this problem set is to familiarize yourself with a new dataset, the National Longitudinal Study of 1972 (NLS-72). NLS is a nationally representative, longitudinal study of 12th graders in 1972 with follow-up surveys throughout their postsecondary years. You will be using the Postsecondary Education Transcript File of the NLS-72, which contains information on transcripts from NLS-72 senior cohort members who reported attending a postsecondary institution after high school.

## Load library and data

You'll need to load the `tidyverse`, `haven` and `labelled` libraries in order to load and work with the NLS data. If these packages are not yet installed, then you must install before you load. Install in “console” rather than .Rmd file

- Generic syntax: `install.packages("package_name")`
- Install “haven”: `install.packages("haven")`

Note: when we **load** package, name of package is not in quotes; but when we **install** package, name of package is in quotes:

- `install.packages("tidyverse")`
- `library(tidyverse)`

```
library(tidyverse)
library(haven)
library(labelled)
```

```
rm(list = ls()) # remove all objects
```

```
nls_crs<- read_dta(file="https://github.com/ozanj/rclass/raw/master/data/nls72/nls72petscrs_v2.dta", en
```

## Step 1: Investigate Variables

1. Use `typeof`, `class`, `str`, and `attributes` functions to investigate the following variables: `crsgrada`, `crsgradb`, `gradtype`, `crscred`.

```
typeof(nls_crs$crsgrada)
```

```
## [1] "character"
```

```
class(nls_crs$crsgrada)
```

```
## [1] "character"
```

```
attributes(nls_crs$crsgrada)
```

```
## $label
```

```
## [1] "COURSE GRADE ALPHA"
```

```
##
```

```
## $format.stata
```

```
## [1] "%2s"
```

```
typeof(nls_crs$crsgradb)
```

```
## [1] "double"
```

```
class(nls_crs$crsgradb)
```

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

```
attributes(nls_crs$crsgradb)
```

```
## $label
```

```
## [1] "COURSE GRADE NUMERIC"
```

```
##
```

```
## $format.stata
```

```
## [1] "%12.0g"
```

```
typeof(nls_crs$gradtype)
```

```
## [1] "double"
```

```
class(nls_crs$gradtype)
```

```
## [1] "haven_labelled"
```

```
attributes(nls_crs$gradtype)
```

```
## $label
```

```
## [1] "TYPE OF GRADE"
```

```
##
```

```
## $format.stata
```

```
## [1] "%12.0g"
```

```
##
```

```
## $class
```

```
## [1] "haven_labelled"
```

```
##
```

```
## $labels
```

```
##      1. letter      2. numeric  9. {MISSING}
```

```
##           1           2           9
```

```
typeof(nls_crs$crsecured)
```

```
## [1] "double"
```

```
class(nls_crs$crsecured)
```

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

```
attributes(nls_crs$crsecured)
```

```
## $label
```

```
## [1] "COURSE CREDITS POSSIBLE"
```

```
##
```

```
## $format.stata
```

```
## [1] "%12.0g"
```

2. `crsgrada` is the variable for letter course grades. Create a factor version of the `crsgrada` variable. Hint: knowing what class the variable is currently and investigating the variable using `table()` will be helpful to creating the new factor version. Retain the new factor version variable in the `nls_crs` dataframe using the variable name `crsgrad_fac`. Check that this new variable is a factor class.

```
table(nls_crs$crsgrada_fac)
```

```
## Warning: Unknown or uninitialised column: 'crsgrada_fac'.
```

```
## < table of extent 0 >
```

```
nls_crs$crsgrada_fac <- factor(nls_crs$crsgrada, levels = c("99", "A", "A-", "A+", "AU", "B", "B-", "B+",  
"C", "C-", "C+", "CR", "D", "D-", "D+",  
"E", "F", "I", "NO", "P", "S", "U", "W", "WF"
```

```
typeof(nls_crs$crsgrada_fac)
```

```
## [1] "integer"
```

```
nls_crs %>% count(crsgrada_fac)
```

```
## # A tibble: 25 x 2  
##   crsgrada_fac      n  
##   <fct>          <int>  
## 1 99             24814  
## 2 A             113200  
## 3 A-             5221  
## 4 A+             523  
## 5 AU             598  
## 6 B             126003  
## 7 B-             3813  
## 8 B+             6639  
## 9 C             89782  
## 10 C-            1841  
## # ... with 15 more rows
```

```
class(nls_crs$crsgrada_fac)
```

```
## [1] "factor"
```

```
attributes(nls_crs$crsgrada_fac)
```

```
## $levels
```

```
## [1] "99" "A" "A-" "A+" "AU" "B" "B-" "B+" "C" "C-" "C+" "CR" "D" "D-"
```

```
## [15] "D+" "E" "F" "I" "NO" "P" "S" "U" "W" "WF" "WP"
```

```
##
## $class
## [1] "factor"
```

3. Create a numeric course grade version of the `crsgrada_fac` variable named `numgrade` with the following numeric values based on attribute levels from `crsgrada_fac` Hint: use `mutate()` and `recode()`. Retain this new `numgrade` variable.
  - A+= 4; A=4; A-=3.7; B+=3.3; B=3; B-=2.7; C+=2.3; C=2; C-=1.7; D+=1.3; D=1; D-=.7; F=0; E=0; WF=0
  - All other letter grades should have missing values for `numgrade`. Hint: use the `.default`
  - When recoded to missing `NA_real_` rather than `NA` due to `recode()` needing a double type/numeric class value to recode and `NA` is a logical)

```
nls_crs <- nls_crs %>%
  mutate(numgrade =
    recode(crsgrada_fac,
      "A+" = 4,
      "A" = 4,
      "A-" = 3.7,
      "B+" = 3.3,
      "B" = 3,
      "B-" = 2.7,
      "C+" = 2.3,
      "C" = 2,
      "C-" = 1.7,
      "D+" = 1.3,
      "D" = 1,
      "D-" = 0.7,
      "F" = 0,
      "E" = 0,
      "WF" = 0,
      .default = NA_real_
    )
  )

nls_crs %>% count(numgrade)
```

```
## # A tibble: 13 x 2
##   numgrade      n
##   <dbl> <int>
## 1      0  14838
## 2    0.7   286
## 3      1  22883
## 4    1.3   610
## 5    1.7  1841
## 6      2  89782
## 7    2.3  4285
## 8    2.7  3813
## 9      3 126003
## 10   3.3  6639
## 11   3.7  5221
## 12    4 113723
## 13   NA  94598
```

4. `gradtype` is a labelled class variable for the type of grade given for each course. Retrieve the variable and value labels for `gradtype`. Get a count of `gradtype` showing the values and the value labels. Now,

get another count by filtering for observations associated with "{MISSING}".

```
nls_crs %>% select(gradtype) %>% var_label()
```

```
## $gradtype
## [1] "TYPE OF GRADE"
```

```
nls_crs %>% select(gradtype) %>% val_labels()
```

```
## $gradtype
## 1. letter 2. numeric 9. {MISSING}
## 1 2 9
```

```
nls_crs %>% count(gradtype) %>% as_factor()
```

```
## # A tibble: 3 x 2
## gradtype n
## <fct> <int>
## 1 1. letter 459348
## 2 2. numeric 10517
## 3 9. {MISSING} 14657
```

```
nls_crs %>% filter(gradtype==9) %>% count()
```

```
## # A tibble: 1 x 1
## n
## <int>
## 1 14657
```

5. `coursecred` is the variable for numerical course grades. There are several issues with this variable.. First, missing observations for `crsgradb` are currently 999 and 999.999. The variable also has values greater than 4 (problematic when the highest possible grade  $A+ = 4$ ). Create and retain a new `crsgradb_v2` variable that replaces all values greater than 4 for `crsgradb` to NA (Hint: you can use the `mutate` and `if_else()` functions to either replace the value to NA or keep the current value of the variable based on whether the expression you specify evaluates to TRUE or FALSE. See below...

## ANSWER PROVIDED FOR YOU

```
table(nls_crs$crsgradb)
```

```
##
## 0 0.1 0.2 0.25 0.3 0.325 0.4 0.5 0.571
## 13972 2 5 1 2 1 4 12 1
## 0.6 0.657 0.7 0.769 0.775 0.8 0.9 0.914 1
## 3 1 296 3 1 12 6 2 22075
## 1.075 1.086 1.1 1.115 1.12 1.171 1.2 1.225 1.231
## 2 1 17 1 1 1 24 7 2
## 1.24 1.257 1.3 1.34 1.343 1.346 1.36 1.375 1.4
## 4 4 646 1 1 3 5 5 33
## 1.429 1.45 1.462 1.5 1.514 1.525 1.53 1.577 1.6
## 2 5 4 256 1 4 1 5 60
## 1.64 1.67 1.675 1.69 1.692 1.7 1.72 1.75 1.8
## 1 2 12 1 2 1868 4 33 52
## 1.808 1.825 1.83 1.84 1.857 1.86 1.87 1.9 1.923
## 1 8 1 2 3 1 1 72 4
## 1.93 1.94 1.943 1.95 1.96 1.975 2 2.02 2.038
## 1 1 2 1 5 17 86113 2 1
## 2.04 2.05 2.06 2.07 2.08 2.1 2.114 2.125 2.13
## 2 21 1 1 8 57 1 25 1
```

```
## 2.133 2.15 2.154 2.16 2.17 2.2 2.22 2.23 2.237
## 1 1 1 1 3 94 1 2 1
## 2.267 2.275 2.28 2.29 2.3 2.32 2.33 2.333 2.35
## 1 16 1 1 4344 14 3 2 21
## 2.357 2.371 2.385 2.4 2.425 2.44 2.467 2.5 2.51
## 1 1 1 91 24 10 2 3296 1
## 2.522 2.53 2.533 2.543 2.55 2.56 2.575 2.58 2.6
## 1 1 3 1 1 10 21 1 82
## 2.629 2.64 2.65 2.667 2.68 2.7 2.725 2.733 2.75
## 1 1 22 5 5 3891 19 1 2
## 2.777 2.78 2.8 2.83 2.84 2.86 2.867 2.875 2.882
## 1 1 145 1 1 2 3 18 1
## 2.9 2.905 2.91 2.92 2.933 2.94 2.95 2.97 2.99
## 88 1 1 4 3 1 16 1 1
## 3 3.01 3.025 3.04 3.06 3.067 3.1 3.11 3.13
## 122913 1 17 4 1 5 109 1 2
## 3.133 3.136 3.143 3.15 3.16 3.175 3.2 3.22 3.25
## 5 1 1 1 2 7 113 2 31
## 3.267 3.27 3.28 3.29 3.295 3.3 3.314 3.325 3.33
## 1 1 8 1 1 6714 1 3 3
## 3.333 3.34 3.35 3.36 3.362 3.4 3.43 3.45 3.467
## 3 4 1 1 1 140 2 1 1
## 3.47 3.475 3.5 3.52 3.533 3.54 3.57 3.571 3.58
## 1 2 974 5 2 1 1 1 2
## 3.595 3.6 3.61 3.62 3.625 3.63 3.64 3.662 3.667
## 1 87 1 2 4 1 5 1 1
## 3.7 3.745 3.75 3.76 3.8 3.81 3.83 3.86 3.87
## 5298 1 2 5 71 1 2 1 1
## 3.9 3.91 3.93 4 4.5 6 7 8 9
## 31 2 1 109753 1 3 1 1 15
## 12 14 17 19 20 21 23 25 25.5
## 2 1 2 1 1 1 2 2 1
## 26 27.5 35 38 39 43 45 46 50
## 2 1 3 1 1 1 1 2 4
## 51 51.5 52.5 54 55 56 57 58 59.5
## 1 1 1 2 3 1 1 2 1
## 60 60.5 61 62 63 64 65 66 67
## 2 1 1 1 1 1 1 4 2
## 67.5 68 68.5 69 70 71 72 74 75
## 1 1 1 3 5 1 2 3 1
## 76 77 78 79 79.5 80 83 84 85
## 3 2 2 2 1 2 1 3 4
## 88 89 90 91 93 94 97 97.5 100
## 3 1 1 1 1 1 2 1 1
## 999 999.999
## 146 99770
```

```
nls_crs<- nls_crs %>%
  mutate(crsgradb_v2= ifelse(crsgradb>4, NA, crsgradb))
```

5. `crscred` is the variable for how many total credits were possible for each course. Missing observations for `crscred` are currently 999 and 999.999. Using code similar to Question 5, create and retain a new `crscred_v2` variable that replaces values of 999 and 999.999 to NA.

```
nls_crs <- nls_crs %>%
  mutate(crsecredv2= ifelse(crsecred>=900, NA, crsecred))
```

6. Create a “final” numerical grade variable named `numgrade_v2` that has the numeric values below based on attribute levels from `crsgrada_fac` **when** `gradtype` indicates letter grades were used and `crsecred_v2` is not missing. Additionally, `num_grade_v2` should equal the value of `crsgradb_v2` **when** `gradtype` indicates numeric grades were used and `crsecred_v2` is not missing. Hint: use `mutate()` and `case_when()`.

- A+= 4; A=4; A-=3.7; B+=3.3; B=3; B-=2.7; C+=2.3; C=2; C-=1.7; D+=1.3; D=1; D-=.7; F=0; E=0; WF=0

```
nls_crs_v2<- nls_crs %>%
  mutate(
    numgrade_v2=case_when(
      crsgrada_fac %in% c("A+", "A") & gradtype==1 & (!is.na(crsecredv2)) ~ 4,
      crsgrada_fac=="A-" & gradtype==1 & (!is.na(crsecredv2)) ~ 3.7,
      crsgrada_fac=="B+" & gradtype==1 & (!is.na(crsecredv2)) ~ 3.3,
      crsgrada_fac=="B" & gradtype==1 & (!is.na(crsecredv2)) ~ 3,
      crsgrada_fac=="B-" & gradtype==1 & (!is.na(crsecredv2)) ~ 2.7,
      crsgrada_fac=="C+" & gradtype==1 & (!is.na(crsecredv2)) ~ 2.3,
      crsgrada_fac=="C" & gradtype==1 & (!is.na(crsecredv2)) ~ 2,
      crsgrada_fac=="C-" & gradtype==1 & (!is.na(crsecredv2)) ~ 1.7,
      crsgrada_fac=="D+" & gradtype==1 & (!is.na(crsecredv2)) ~ 1.3,
      crsgrada_fac=="D" & gradtype==1 & (!is.na(crsecredv2)) ~ 1,
      crsgrada_fac=="D-" & gradtype==1 & (!is.na(crsecredv2)) ~ 0.7,
      crsgrada_fac %in% c("F", "E", "WF") & gradtype==1 & (!is.na(crsecredv2)) ~ 0,
      crsgradb<=4 & gradtype==2 & (!is.na(crsecredv2)) ~ crsgradb # use values of numeric var crsgradb
    )
  )

nls_crs_v2 %>% count(numgrade_v2)
```

```
## # A tibble: 201 x 2
##   numgrade_v2     n
##   <dbl> <int>
## 1         0 13866
## 2       0.1      1
## 3       0.2      5
## 4       0.3      2
## 5       0.4      4
## 6       0.5      9
## 7       0.6      1
## 8       0.7     290
## 9     0.769      3
## 10    0.775      1
## # ... with 191 more rows
```