Chapter 4: Data Management

Exercises

Brian Fogarty

15 May 2018

Contents

EXERCISE I

Perform the following exercises:

1. Read in the Stata version of the 2015 UK Millennium Cohort survey dataset (`mcs.dta`).
2. Read in the csv version of the 2015 UK Millennium Cohort survey dataset (`mcs.csv`).

3. Do the two versions have the same number of variables and observations?

ANSWERS TO EXERCISE I

Question 1.1

```
setwd("C:/QSSD/Exercises/Chapter 4 - Exercises")
getwd()
```

```
[1] "C:/QSSD/Exercises/Chapter 4 - Exercises"
```

```
library(haven)
mcs <- read_dta("mcs.dta")
```

You need to use the `haven` package to read in this .dta file; and not the `foreign` package.

Question 1.2

```
mcs1 <- read.csv("mcs.csv")
```

Question 1.3

Yes, both versions of the dataset have the same number of variables (52) and observations (11872).

EXERCISE II

Using either version of the data you read in,

1. Examine the data using the `View()` function.
2. Examine the variable names using the `names()` function.

ANSWERS TO EXERCISE II

Question 2.1

```
View(mcs)
```

Question 2.2

```
names(mcs)
```

```
 [1] "mcsid"     "cnum"      "sex"       "tv"        "games"
 [6] "compu"     "internet"  "social"    "engl"      "mths"
[11] "scien"     "phyed"     "hmwk"      "help"      "place"
[16] "best"      "inter"     "hand"      "ethn6"     "countr"
```

```
[21] "sibl_fl"   "grand"     "poor"      "edu_par"   "class"
[26] "class1"    "class2"    "class3"    "class4"    "class5"
[31] "class6"    "class7"    "ethn_dy1"  "ethn_dy2"  "ethn_dy3"
[36] "ethn_dy4"  "ethn_dy5"  "ethn_dy6"  "countr1"   "countr2"
[41] "countr3"   "countr4"   "edu_par1"  "edu_par2"  "edu_par3"
[46] "edu_par4"  "edu_par5"  "econstat"  "lang_home" "lang1"
[51] "lang2"     "lang3"
```

EXERCISE III

Using the .csv version of the data you read in,

1. Examine the structure of the first 10 variables using the `str()` function.
2. What is the class of the variable `games`?
3. What does the summary of the variable `games` tell you?

ANSWERS TO EXERCISE III

Question 3.1

```
str(mcs1,list.len=10)
```

```
'data.frame':   11872 obs. of  52 variables:
 $ mcsid    : Factor w/ 11714 levels "M10002P","M10007U",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ cnum     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ sex      : Factor w/ 2 levels "boy","girl": 1 1 2 1 2 2 2 1 1 1 ...
 $ tv       : Factor w/ 9 levels "","1 hour t",..: 3 4 8 3 4 4 3 2 1 7 ...
 $ games    : Factor w/ 9 levels "","1 hour t",..: 8 6 8 2 9 2 8 3 1 4 ...
 $ compu    : Factor w/ 3 levels "","no","yes": 3 2 3 3 2 3 2 3 1 3 ...
 $ internet : Factor w/ 9 levels "","1 hour t",..: 2 6 8 2 5 3 2 7 1 3 ...
 $ social   : Factor w/ 9 levels "","1 hour t",..: 7 5 8 8 2 3 2 6 1 7 ...
 $ engl     : Factor w/ 4 levels "","Agree","Disagree",..: 2 4 2 2 2 2 2 2 1 3 ...
 $ mths     : Factor w/ 4 levels "","Agree","Disagree",..: 4 3 3 2 2 2 2 3 1 2 ...
  [list output truncated]
```

Question 3.2

```
class(mcs1$games)
```

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

The class is “factor”

Question 3.3

```
summary(mcs1$games)
```

```
    1 hour t  2 hours  3 hours  5 hours  7 hours  Half an Less tha
 360     1751     1583     1478      770      862     1343     1565
None
2160
```

The summary provides the number of observations at each category of the variable `games`.

EXERCISE IV

Using the .csv version of the data you read in,

1. Convert the variable `games` to a numeric variable.
2. Convert the numeric version of `games` to a character variable.
3. Convert the character version of `games` to a factor variable.

ANSWERS TO EXERCISE IV

Question 4.1

```
mcs1$games.num <- as.numeric(mcs1$games)
class(mcs1$games.num)
```

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

Question 4.2

```
mcs1$games.char <- as.character(mcs1$games.num)
class(mcs1$games.char)
```

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

Question 4.3

```
mcs1$games.factor <- as.factor(mcs1$games.char)
class(mcs1$games.factor)
```

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

EXERCISE V

Using the .csv version of the data you read in,

1. Subset the data to remove all missing values.
   (a) How many observations does the subsetted data have now?
   (b) Why is there a difference between the number of observations in the original and the subsetted data?

2. Using the original version of the data, subset the data so that it only contains `mths` and `scien`. Use the `head()` function to check if you were successful.
3. Using the original version of the data, subset the data so that it only contains the first seven variables. Use the `head()` function to check if you were successful.
4. Using the original version of the data, subset the data so that it only contains the first seven variables and `mths` and `scien`. Use the `head()` function to check if you were successful.

ANSWERS TO EXERCISE V

Question 5.1

```
mcs.omit <- na.omit(mcs1)
```

Question 5.1.a

The new data has 7756 observations.

Question 5.1.b

The reason it is less is because we removed all observations that had a missing value for at least one variable.

Question 5.2

```
mcs1.small <- subset(mcs1, select=c(mths,scien))
head(mcs1.small)
```

```
      mths    scien
1 Strongly Strongly
2 Disagree    Agree
3 Disagree Strongly
4    Agree    Agree
5    Agree Strongly
6    Agree Disagree
```

Question 5.3

```
mcs1.small.2 <- subset(mcs1, select=c(mcsid:internet))
head(mcs1.small.2)
```

```
    mcsid cnum  sex       tv    games compu internet
1 M10002P    1  boy  2 hours Less tha   yes 1 hour t
2 M10007U    1  boy  3 hours  7 hours    no  7 hours
3 M10015U    1 girl Less tha Less tha   yes Less tha
4 M10016V    1  boy  2 hours 1 hour t   yes 1 hour t
5 M10018X    1 girl  3 hours     None    no  5 hours
6 M10020R    1 girl  3 hours 1 hour t   yes  2 hours
```

Question 5.4

```
mcs1.small.3 <- subset(mcs1, select=c(mcsid:internet,mths,scien))
head(mcs1.small.3)
```

```
    mcsid cnum  sex       tv    games compu internet     mths    scien
1 M10002P    1  boy  2 hours Less tha   yes 1 hour t Strongly Strongly
2 M10007U    1  boy  3 hours  7 hours    no  7 hours Disagree    Agree
3 M10015U    1 girl Less tha Less tha   yes Less tha Disagree Strongly
4 M10016V    1  boy  2 hours 1 hour t   yes 1 hour t    Agree    Agree
5 M10018X    1 girl  3 hours     None    no  5 hours    Agree Strongly
6 M10020R    1 girl  3 hours 1 hour t   yes  2 hours    Agree Disagree
```