The Variable Key

Paul Johnson¹

¹Center for Research Methods and Data Analysis

2018



Outline

What is the Problem?

Variable Key Solution

Oetails worth mentioning

4 Where We Are Now



This talk is about . . .

- The kutils package for R (R Core Team, 2017)
 - release version available on CRAN
 - test versions on KRAN: http://rweb.crmda.ku.edu/kran
- Vignette, "The Variable Key Data Management Framework"



Outline

- 1 What is the Problem?
- Variable Key Solution
- Details worth mentioning
- Where We Are Now



Clients, Data Managers and Data Analysts

- Clients give us data structures
- Perhaps there are
 - many files, various variable names in different files
 - data entry errors that need to be corrected
- Data Analysts "fall in a hole".
 - They think they make something that works, but it is difficult to be entirely confident



After a few days, they have 1000 lines of code like this

```
## Read in data
dat<-read.csv (file="fulldata.csv", header = TRUE, na.string =
    \texttt{c} \, (\, "-980" \, , \quad "-981" \, , \quad "-982" \, , \quad "-983" \, , \quad "-984" \, , \quad "-985" \, , \quad "-986" \, ,
    "-987", "-988", "-989", "-990", "-991", "-992", "-993".
    "-994", "-995", "-996", "-997", "-998", "-999"))
##head(dat)
summary (as.factor (dat $w2_Dis12))
summary(as.factor(dat$np1Dis_Recod))
##family predictors, home independence
dat$Rnp1G1e<-recode(dat$np1G1e, "0=2;1=1;2=0")
dat$Rnp1G1h<-recode(dat$np1G1h, "0=2;1=1;2=0")
dat$Rnp1G5a<-recode(dat$np1G5a, "4=1;3=2;2=3;1=4")
dat Rnp1G5b < -recode(dat np1G5b, "4=1;3=2;2=3;1=4")
dat $Rnp1G5c<-recode (dat $np1G5c, "4=1;3=2;2=3;1=4")
dat$Rnp1G5d<-recode(dat$np1G5d, "4=1;3=2;2=3;1=4")
dat Rnp1F1d < -recode(dat np1F1d, "1=6;2=5;3=4;4=3;5=2;6=1")
##family predictors, parent perception of school exp
dat$Rnp1D12a<-recode(dat$np1D12a, "4=1;3=2;2=3;1=4")
dat$Rnp1D12b<-recode(dat$np1D12b, "4=1;3=2;2=3;1=4")
dat$Rnp1D12c<-recode(dat$np1D12c, "4=1;3=2;2=3;1=4")
```



After a few days, they have 1000 lines of code like this ...

```
dat$Rnp1D12d<-recode(dat$np1D12d, "4=1;3=2;2=3;1=4")
dat$Rnp1D12e<-recode(dat$np1D12e, "4=1;3=2;2=3;1=4")
dat$Rnp1H4<-recode(dat$np1H4, "4=1;3=2;2=3;1=4")

### student predictors, cpmmunication skills
dat$Rnp1B5a<-recode(dat$np1B5a, "4=1;3=2;2=3;1=4")
dat$Rnp1B5b<-recode(dat$np1B5b, "4=1;3=2;2=3;1=4")
dat$Rnp1B5d<-recode(dat$np1B5d, "4=1;3=2;2=3;1=4")
dat$Rnp1B5d<-recode(dat$np1B5d, "4=1;3=2;2=3;1=4")
dat$Rnp1B5e<-recode(dat$np1B5e, "4=1;3=2;2=3;1=4")</pre>
```



What's wrong here?

- Hard-to-catch user errors
 - The blending of *project-specific values* with *programming idioms* requires an expert in both to review the work
- Difficult to report back to client about everything that was done
- Difficult to coordinate efforts of teammates



Outline

- What is the Problem?
- Variable Key Solution
- Details worth mentioning
- 4 Where We Are Now



All Involved Parties have a Shared Variable Key

Coordinate by creating a Variable Key file, a rectangular worksheet

name_old	name_new	class_old	class_new	value_old	value_new
V23419	sex	integer	factor	1 2 3	male female neither
V32422	education	integer	ordered	1 2 3 4 5	elem hs somecoll ba post
V54532	income	numeric	numeric		

We focus on the most common variable types:
 logical , character , integer , double (aka numeric), factor ,
 ordered



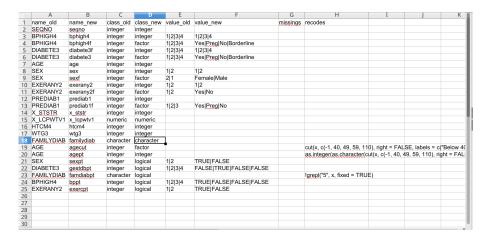
The Key Is A Programmable Codebook

- Team leader—or client—can revise the variable key file without digging into a lot of programming details.
- The implied recodes are automagically implemented (by functions in kutils)



Paul Johnson (CRMDA)

Example: Johnson County Basic Risk Factors





Workflow Step 1. Create a key template

The function keyTemplate can scan an existing data frame and create a template variable key

name_old	name_new	class_old	class_new	value_old	value_new
V23419	V23419	integer	integer	1 2 3	1 2 3
V32422	V32422	integer	integer	1 2 3 4 5	1 2 3 4 5
V54532	V54532	numeric	numeric		

- Researchers/clients name_new , value_new , class_new
- Discrete variables can have an enumerated list of values
- Numeric variables are treated differently (recodes mentioned below)



Client or Worker fills in key

name_old	name_new	class_old	class_new	value_old	value_new
V23419	sex	integer	factor	1 2 3	male female neither
V32422	education	integer	ordered	1 2 3 4 5	elem hs somecoll ba post
V54532	income	numeric	numeric		

kutils includes a function keylmport. Does data integrity checks



And the most important step is...

- The analyst "applies" the key to the data with keyApply
- The variables are renamed, the values are re-aligned
- Profuse diagnostic output

```
brf2 <- keyApply(brfss, key)</pre>
[1] "Variable segno has 20 unique values. Too large for
   a table."
      BPHIGH4 (old var)
bphigh4
      1 1090 0 0
          0 307 0
           0
               0 558
                         45
            BPHIGH4 (old
                         var)
bphigh4f
                         3
            1090
  Yes
  Preg
                  307
```



And the most important step is.......

```
No
                      558
 Borderline 0
                           45
        DIABETE3 (old var)
diabete3f
        239 0
              16
                0 1709
                        36
           DIABETE3 (old var)
diabete3
             239
 Yes
 Preg
                 16
 No
                  0 1709
 Borderline
                            36
[1] "Variable age has 20 unique values. Too large for a
   table."
  SEX (old var)
sex
   974
```



And the most important step is.......

```
2 0 1026
     SEX (old var)
sexf
 Female 0 1026
Male 974 0
     EXERANY2 (old var)
exerany2 1 2
     1 1154 0
     2 0 846
      EXERANY2 (old var)
exerany2f 1 2
    Yes 1154 0
    No 0 846
     PREDIAB1 (old var)
prediab1 1 2 3 < NA >
    190 0 0
    0 167 0 0
      0 0 1404 0
  <NA> 0
                0 239
```



And the most important step is......

```
PREDIAB1 (old var)
prediab1f 1 2 3 <NA>
Yes 190 0 0 0
Preg 0 167 0 0
No 0 0 1404 0
<NA> 0 0 0 239
```



Outline

- 1 What is the Problem?
- Variable Key Solution
- 3 Details worth mentioning
- 4 Where We Are Now



The 'long' key

• In the Wide Key, editing values may be difficult in a spreadsheet

```
"low | moderate | medium | warm | hot | boiling "
```

 The Long Key in an equivalent representation, but with one row per value



The ''long'' key ...

name_old	name_new	class_old	class_new	values_old	values_new
V22012	water	character	factor	low	low
V22012	water	character	factor	moderate	moderate
V22012	water	character	factor	medium	medium
V22012	water	character	factor	warm	warm
V22012	water	character	factor	hot	hot
V22012	water	character	factor	boiling	boiling
V23419	sex	integer	factor	1	male
V23419	sex	integer	factor	2	female
V23419	sex	integer	factor	3	neither



The 'long' key ...

kutils provides functions wide2long and long2wide for conversion



Partial Variable Keys

- keyApply argument drop = c("vars", "vals")
- If drop = "vars", then variables that are not mentioned in the key are removed from the new data frame
 - Use Case: We want to use 20 variables from data set that includes 1000s of columns
 - Otherwise, all columns remain in data
- If drop = "vals", then key omission of scores from "value_old" will cause those observations to be changed to missing
 - Otherwise, values omitted from key pass through to output data unaltered



Outline

- What is the Problem?
- Variable Key Solution
- Oetails worth mentioning
- Where We Are Now



It Works!

- keyTemplate , keyImport , and keyApply work as intended
- Tested in several projects



Checking that it REALLY Does Work

- Validated variable type conversion among the 6 variable types described above.
- Formalized "unit tests" implemented in 2017 to validate code updates



New Feature: Key for SPSS & Stata Data

- Variable Key can summarize the existing coding nomenclature
 - "value" and
 - "labels"



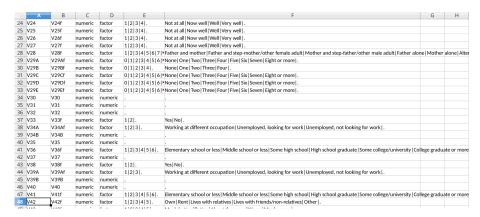
Key depiction of an SPSS data set

Children of Immigrants Study in wide key format

	A	В	С	D	E	F
1	name_ ▼	name_r	class_o 🔻	class_n ▼	value_old ▼	value_new 🔽 ŗ
2	CASEID	CASEID	numeric	numeric		
3	V1	V1	numeric	numeric		
4	V2	V2f	numeric	factor	1 3 4	Miami Ft. Lauderdale San Diego
5	V4	V4	numeric	numeric		
6	V5	V5f	numeric	factor	7 8 9 10	Seventh grade Eighth grade Ninth grade Tenth grade
7	V7	V7f	numeric	factor	1 2 3 .	Yes No Dead/unknown .
8	V8	V8f	numeric	factor	1 2 3 4 5 6 7	Same city Another city in Miami smsa Another city in Florida Another US state Abroad San Diego/neighboring city 5
9	V9	V9	numeric	numeric		
10	V10	V10	numeric	numeric		
11	V11	V11f	numeric	factor	1 2 .	Yes No .
12	V13	V13f	numeric	factor	1 2 .	Yes No .
13	V14	V14f	numeric	factor	1 2 3 4 5 6 7	Same city Another city in Miami smsa Another city in Florida Another US state Abroad San Diego/neighboring city 5
14	V15	V15	numeric	numeric		
15	V16	V16	numeric	numeric		
16	V17	V17f	numeric	factor	1 2 .	Yes No .
17	V18	V18f	numeric	factor	1 2	Male Female
18	V19	V19	numeric	numeric		
19	V20	V20	numeric	numeric		
20	V21	V21	numeric	numeric		
21	V21A	V21A	numeric	numeric		
22	V22	V22f	numeric	factor	1 2 3 4 .	All my life Ten years or more Five to nine years Less than five years .
23	V23	V23f	numeric	factor	1 2 .	Yes No .
24	V24	V24f	numeric	factor	1 2 3 4 .	Not at all Now well Well Very well .
25	V25	V25f	numeric	factor		Not at all Now well Well Very well .
26	1/24	11244		fantas		Not at all November 1990 (1990)



Key depiction of an SPSS data set ...



[... snip many rows]



Working on Codebook Generator

- We want output that integrates the key information with observed data frequencies
- Existing code can generate nice reports for discrete variables



References

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

