The Variable Key

Paul Johnson[1]

[1]Center for Research Methods and Data Analysis

2018

Outline

KU

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"

KU

Outline

KU

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 =
    c("-980", "-981", "-982", "-983", "-984", "-985", "-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")
```

KU

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$Rnp1B5e<-recode(dat$np1B5e, "4=1;3=2;2=3;1=4")
```

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

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`)

Example: Johnson County Basic Risk Factors

	A	B	C	D	E	F	G	H	I	J	K
1	name_old	name_new	class_old	class_new	value_old	value_new	missings	recodes			
2	SEQNO	seqno	integer	integer							
3	BPHIGH4	bphigh4	integer	integer	1|2|3|4	1|2|3|4					
4	BPHIGH4	bphigh4f	integer	factor	1|2|3|4	Yes|Preg|No|Borderline					
5	DIABETE3	diabete3f	integer	integer	1|2|3|4	1|2|3|4					
6	DIABETE3	diabete3	integer	factor	1|2|3|4	Yes|Preg|No|Borderline					
7	AGE	age	integer	integer							
8	SEX	sex	integer	integer	1|2	1|2					
9	SEX	sexf	integer	factor	2|1	Female|Male					
10	EXERANY2	exerany2	integer	integer	1|2	1|2					
11	EXERANY2	exerany2f	integer	factor	1|2	Yes|No					
12	PREDIAB1	prediab1	integer	integer							
13	PREDIAB1	prediab1f	integer	factor	1|2|3	Yes|Preg|No					
14	X_STSTR	x_ststr	integer	integer							
15	X_LCPWTV1	x_lcpwtv1	numeric	numeric							
16	HTCM4	htcm4	integer	integer							
17	WTG3	wtg3	integer	integer							
18	FAMILYDIAB	familydiab	character	character							
19	AGE	agecut	integer	factor				cut(x, c(-1, 40, 49, 59, 110), right = FALSE, labels = c("Below 4(			
20	AGE	agept	integer	integer				as.integer(as.character(cut(x, c(-1, 40, 49, 59, 110), right = FAL:			
21	SEX	sexpt	integer	logical	1|2	TRUE|FALSE					
22	DIABETE3	gestdbpt	integer	logical	1|2|3|4	FALSE|TRUE|FALSE|FALSE					
23	FAMILYDIAB	famdiabpt	character	logical				!grepl("5", x, fixed = TRUE)			
24	BPHIGH4	bppt	integer	logical	1|2|3|4	TRUE|FALSE|FALSE|FALSE					
25	EXERANY2	exercpt	integer	logical	1|2	TRUE|FALSE					
26											
27											
28											
29											
30											

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 `keyImport`. 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)
[1] "Variable seqno has 20 unique values. Too large for
    a table."
        BPHIGH4 (old var)
bphigh4    1    2    3    4
      1 1090    0    0    0
      2    0  307    0    0
      3    0    0  558    0
      4    0    0    0   45
              BPHIGH4 (old var)
bphigh4f         1    2    3    4
  Yes         1090    0    0    0
  Preg           0  307    0    0
```

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

```
  No                0     0   558     0
  Borderline        0     0     0    45
          DIABETE3 (old var)
diabete3f     1     2     3     4
        1   239     0     0     0
        2     0    16     0     0
        3     0     0  1709     0
        4     0     0     0    36
              DIABETE3 (old var)
diabete3          1     2     3     4
  Yes           239     0     0     0
  Preg            0    16     0     0
  No              0     0  1709     0
  Borderline      0     0     0    36
[1] "Variable age has 20 unique values. Too large for a
    table."
   SEX (old var)
sex     1     2
  1   974     0
```

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

```
  2     0 1026
        SEX (old var)
sexf          1     2
  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>
    1       190     0     0     0
    2         0   167     0     0
    3         0     0  1404     0
    <NA>      0     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

KU

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

KU

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

KU

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	B	C	D	E	F
1	name_d	name_f	class_o	class_n	value_old	value_new
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|S
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|S
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	1|2|3|4|.	Not at all|Now well|Well|Very well|.
26	V26	V26f	numeric	factor	1|2|3|4|	Not at all|Now well|Well|Very well|

Key depiction of an SPSS data set ...

	A	B	C	D	E	F	G	H
24	V24	V24f	numeric	factor	1|2|3|4|.	Not at all|Now well|Well|Very well|.		
25	V25	V25f	numeric	factor	1|2|3|4|.	Not at all|Now well|Well|Very well|.		
26	V26	V26f	numeric	factor	1|2|3|4|.	Not at all|Now well|Well|Very well|.		
27	V27	V27f	numeric	factor	1|2|3|4|.	Not at all|Now well|Well|Very well|.		
28	V28	V28f	numeric	factor	1|2|3|4|5|6|7|	Father and mother|Father and step-mother/other female adult|Mother and step-father/other male adult|Father alone|Mother alone|Alter		
29	V29A	V29Af	numeric	factor	0|1|2|3|4|5|6|	None|One|Two|Three|Four|Five|Six|Seven|Eight or more|.		
30	V29B	V29Bf	numeric	factor	0|1|2|3|4|.	None|One|Two|Three|Four|.		
31	V29C	V29Cf	numeric	factor	0|1|2|3|4|5|6|	None|One|Two|Three|Four|Five|Six|Seven|Eight or more|.		
32	V29D	V29Df	numeric	factor	0|1|2|3|4|5|6|	None|One|Two|Three|Four|Five|Six|Seven|Eight or more|.		
33	V29E	V29Ef	numeric	factor	0|1|2|3|4|5|6|	None|One|Two|Three|Four|Five|Six|Seven|Eight or more|.		
34	V30	V30	numeric	numeric	.	.		
35	V31	V31	numeric	numeric	.	.		
36	V32	V32	numeric	numeric	.	.		
37	V33	V33f	numeric	factor	1|2|.	Yes|No|.		
38	V34A	V34Af	numeric	factor	1|2|3|.	Working at different occupation|Unemployed, looking for work|Unemployed, not looking for work|.		
39	V34B	V34B	numeric	numeric	.	.		
40	V35	V35	numeric	numeric	.	.		
41	V36	V36f	numeric	factor	1|2|3|4|5|6|.	Elementary school or less|Middle school or less|Some high school|High school graduate|Some college/university|College graduate or more		
42	V37	V37	numeric	numeric	.	.		
43	V38	V38f	numeric	factor	1|2|.	Yes|No|.		
44	V39A	V39Af	numeric	factor	1|2|3|.	Working at different occupation|Unemployed, looking for work|Unemployed, not looking for work|.		
45	V39B	V39B	numeric	numeric	.	.		
46	V40	V40	numeric	numeric	.	.		
47	V41	V41f	numeric	factor	1|2|3|4|5|6|.	Elementary school or less|Middle school or less|Some high school|High school graduate|Some college/university|College graduate or more		
48	V42	V42f	numeric	factor	1|2|3|4|5|.	Own|Rent|Lives with relatives|Lives with friends/non-relatives|Other|.		

[. . . 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.