Groningen Exercise Report

This reports narrates the results of modeling in the Groningen Harmonization Exercise

```r
# Attach these packages so their functions don't need to be qualified: http://r-pkgs.had.co.nz/namespac
library(magrittr) # enables piping : %>%
# library(ggplot2)
# library(glmulti)
# library(rJava)
# require(MASS)

# Verify these packages are available on the machine, but their functions need to be qualified: http://
requireNamespace("testit")# For asserting conditions meet expected patterns.
requireNamespace("ggplot2") # graphing
requireNamespace("tidyr") # data manipulation
requireNamespace("dplyr") # Avoid attaching dplyr, b/c its function names conflict with a lot of packag
requireNamespace("plyr")
```

```r
# Call `base::source()` on any repo file that defines functions needed below.  Ideally, no real operati
source("./scripts/common-functions.R") # used in multiple reports
source("./scripts/graph-presets.R") # fonts, colors, themes
source("./scripts/graph-logistic.R")
```

```r
#Put code in here.  It doesn't call a chunk in the codebehind file.
```

> This report is a record of interaction with a data transfer object (dto) produced by `./manipulation/0-ellis-island.R`.

The next section recaps this script, exposes the architecture of the DTO, and demonstrates the language of interacting with it.

Exposition

Ellis Island

> All data land on Ellis Island.

The script `0-ellis-island.R` is the first script in the analytic workflow. It accomplished the following:

- (1) Reads in raw data files from the candidate studies

- (2) Extract, combines, and exports their metadata (specifically, variable names and labels, if provided) into `./data/shared/derived/meta-data-live.csv`, which is updated every time Ellis Island script is executed.

- (3) Augments raw metadata with instructions for renaming and classifying variables. The instructions are provided as manually entered values in `./data/shared/meta-data-map.csv`. They are used by automatic scripts in later harmonization and analysis.

- (4) Combines unit and metadata into a single DTO to serve as a starting point to all subsequent analyses.

```
# load the product of 0-ellis-island.R,  a list object containing data and metadata
dto <- readRDS("./data/unshared/derived/dto_h.rds")
```

```
# the list is composed of the following elements
names(dto)
```

```
[1] "studyName" "filePath"  "unitData"  "metaData"
```

```
# 1st element - names of the studies as character vector
dto[["studyName"]]
```

```
[1] "alsa"  "lbsl"  "satsa" "share" "tilda"
```

```
# 2nd element - file paths of the data files for each study as character vector
dto[["filePath"]]
```

```
[1] "./data/unshared/raw/ALSA-Wave1.Final.sav"         "./data/unshared/raw/LBSL-Panel2-Wave1.Final.sav
[3] "./data/unshared/raw/SATSA-Q3.Final.sav"           "./data/unshared/raw/SHARE-Israel-Wave1.Final.sa
[5] "./data/unshared/raw/TILDA-Wave1.Final.sav"
```

```
# 3rd element - is a list object containing the following elements
names(dto[["unitData"]])
```

```
[1] "alsa"  "lbsl"  "satsa" "share" "tilda"
```

```
# each of these elements is a raw data set of a corresponding study, for example
dplyr::tbl_df(dto[["unitData"]][["lbsl"]])
```

```
Source: local data frame [656 x 39]

        id AGE94 SEX94  MSTAT94 EDUC94     NOWRK94  SMK94                                         SMOKE
     (int) (int) (int)   (fctr)  (int)      (fctr) (fctr)                                        (fctr)
1  4001026    68     1 divorced     16 no, retired     no                                  never smoked
2  4012015    94     2  widowed     12 no, retired     no                                  never smoked
3  4012032    94     2  widowed     20 no, retired     no don't smoke at present but smoked in the past
4  4022004    93     2       NA     NA          NA     NA                                  never smoked
5  4022026    93     2  widowed     12 no, retired     no                                  never smoked
6  4031031    92     1  married      8 no, retired     no don't smoke at present but smoked in the past
7  4031035    92     1  widowed     13 no, retired     no don't smoke at present but smoked in the past
8  4032201    92     2       NA     NA          NA     NA don't smoke at present but smoked in the past
9  4041062    91     1  widowed      7          NA     no don't smoke at present but smoked in the past
10 4042057    91     2       NA     NA          NA     NA                                            NA
..     ...   ...   ...      ...    ...         ...    ...                                           ...
Variables not shown: ALCOHOL (fctr), WINE (int), BEER (int), HARDLIQ (int), SPORT94 (int), FIT94 (int),
  SPEC94 (int), DANCE94 (int), CHORE94 (int), EXCERTOT (int), EXCERWK (int), HEIGHT94 (int), WEIGHT94 (
  (int), HHEIGHT (int), SRHEALTH (fctr), smoke_now (lgl), smoked_ever (lgl), year_of_wave (dbl), age_in
  year_born (dbl), female (lgl), marital (chr), single (lgl), educ3 (chr), current_work_2 (lgl), curren
  sedentary (lgl), poor_health (lgl), bmi (dbl)
```

Meta

```
# 4th element - a dataset names and labels of raw variables + added metadata for all studies
dto[["metaData"]] %>%
  dplyr::select(study_name, name, item, construct, type, categories, label_short, label) %>%
  DT::datatable(
    class   = 'cell-border stripe',
    caption = "This is the primary metadata file. Edit at `./data/shared/meta-data-map.csv",
    filter  = "top",
    options = list(pageLength = 6, autoWidth = TRUE)
  )
```

```
# t <- table(ds$smoke_now, ds$study_name, useNA="always");t[t==0]<-".";t
```

Assembly

The dto containing harmonized operationalizations is queried to assemble analysis-ready dataset.

```
assemble_dto <- function(dto, get_these_variables){

  lsh <- list() #  list object with harmonized data
  for(s in dto[["studyName"]]){
    ds <- dto[["unitData"]][[s]] # get study data from dto
    variables_present <- colnames(ds) %in% get_these_variables # variables on the list
    lsh[[s]] <- ds[, variables_present] # keep only them
  }
  return(lsh)
}
lsh <- assemble_dto(
  dto=dto,
  get_these_variables <- c(
    "id",
    "year_of_wave","age_in_years","year_born",
    "female",
    "educ3",
    "marital", "single",
    "smoke_now","smoked_ever",
    "poor_health",
    "sedentary",
    "current_work_2",
    "current_drink"
  )
)
lapply(lsh, names) # view the contents of the list object
```

```
$alsa
 [1] "id"             "smoke_now"      "smoked_ever"    "year_of_wave"   "age_in_years"   "year_born"
 [7] "female"         "marital"        "single"         "educ3"          "current_work_2" "current_drin
[13] "sedentary"      "poor_health"

$lbsl
 [1] "id"             "smoke_now"      "smoked_ever"    "year_of_wave"   "age_in_years"   "year_born"
 [7] "female"         "marital"        "single"         "educ3"          "current_work_2" "current_drin
[13] "sedentary"      "poor_health"

$satsa
 [1] "id"             "smoke_now"      "smoked_ever"    "year_of_wave"   "age_in_years"   "year_born"
 [7] "female"         "marital"        "single"         "educ3"          "current_work_2" "current_drin
[13] "sedentary"      "poor_health"
```

```
$share
 [1] "id"             "smoke_now"      "smoked_ever"    "year_of_wave"   "year_born"      "age_in_years
 [7] "female"         "marital"        "single"         "educ3"          "current_work_2" "current_drin
[13] "sedentary"      "poor_health"

$tilda
 [1] "id"             "smoke_now"      "smoked_ever"    "year_of_wave"   "age_in_years"   "year_born"
 [7] "female"         "marital"        "single"         "educ3"          "current_work_2" "current_drin
[13] "sedentary"      "poor_health"
```

```
ds <- plyr::ldply(lsh,data.frame, .id = "study_name")
ds$id <- 1:nrow(ds) # some ids values might be identical, replace
ds %>% names()
```

```
 [1] "study_name"     "id"             "smoke_now"      "smoked_ever"    "year_of_wave"   "age_in_years
 [7] "year_born"      "female"         "marital"        "single"         "educ3"          "current_work_
[13] "current_drink"  "sedentary"      "poor_health"
```

This dataset, which includes harmonized variables, will be used to fit the models.

According to the rules of the exercise,however, only the participants over the age of 50 were kept for the analysis:

```
# restrict analysis to respondents age 50+
ds <- ds %>%
  dplyr::filter(age_in_years >= 50)
```

Harmonization rules

This section narrates the harmonization rules applied to candidate variables from each study and provides the descriptives of harmonized variables

Harmonization has been carried out by sequential execution of the follwoing scripts:

- `./reports/harmonize-smoking/harmonize-smoking.R`
- `./reports/harmonize-age/harmonize-age.R`
- `./reports/harmonize-sex/harmonize-sex.R`
- `./reports/harmonize-marital/harmonize-marital.R`
- `./reports/harmonize-education/harmonize-education.R`
- `./reports/harmonize-work/harmonize-work.R`
- `./reports/harmonize-alcohol/harmonize-alcohol.R`
- `./reports/harmonize-physact/harmonize-physact.R`
- `./reports/harmonize-health/harmonize-health.R`
- `./reports/harmonize-physique/harmonize-physique.R`

the reports are produced by knitting their respective `.Rmd` files, located in corresponding folders.

The following subsections provide summary details on the harmonization implemented to produce each target variable. Please refer to full reports listed above for further details.

Smoking

View descriptives : smoking for closer examination of each variable that contributed to the computation of the harmonized variable.

Are you a smoker presently? - `smoke_now`

- `0` - `FALSE` - *healthy* - Reference group

- `1` - `TRUE` - *unhealthy* - Risk factor

```
t <- table(ds$smoke_now, ds$study_name, useNA="always");t[t==0]<-".";t
```

```
        alsa lbsl satsa share tilda <NA>
  FALSE 1851 430  934   2113  6674  .   
  TRUE  217  60   246   390   1488  .   
  <NA>  19   92   60    4     1     .   
```

Have you ever smoked? - `smoked_ever`

- `0` - `FALSE` - *healthy* - Reference group

- `1` - `TRUE` - *unhealthy* - Risk factor

The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

```
t <- table( ds$smoked_ever,ds$study_name, useNA="always");t[t==0]<-".";t
```

```
        alsa lbsl satsa share tilda <NA>
  FALSE 1851 173  621   1485  3561  .   
  TRUE  217  324  530   1018  4601  .   
  <NA>  19   85   89    4     1     .   
```

Age

View descriptives : age for closer examination of raw variables. For each study, three variables have been formulated and computed:

- `year_of_wave` - Calendar year in which the measurement wave occured. These data values are added manually, after consulting respective study's documentation.

- `year_born` - Calendar year in which the respondent was born

- `age_in_years` - Age of respondent in years

```
lsh_age <- assemble_dto(dto, c("id","year_of_wave","age_in_years","year_born"))
lapply(lsh_age, head) # view the contents of the list object
```

```
$alsa
   id year_of_wave age_in_years year_born
1  41         1992           86      1906
2  42         1992           78      1914
3  61         1992           89      1903
4  71         1992           78      1914
```

```
5  91          1992            85      1907
6 121          1992            92      1900

$lbsl
       id year_of_wave age_in_years year_born
1 4001026         1994           68      1926
2 4012015         1994           94      1900
3 4012032         1994           94      1900
4 4022004         1994           93      1901
5 4022026         1994           93      1901
6 4031031         1994           92      1902

$satsa
     id year_of_wave age_in_years year_born
1  2321         1991     64.81331      1926
2  2322         1991     64.81331      1926
3  2501         1991     64.80330      1926
4  2502         1991     64.80330      1926
5  2621         1991     64.75332      1926
6 11301         1991     90.20333      1900

$share
          id year_of_wave year_born age_in_years
1 2.5052e+12         2006      1942           64
2 2.5052e+12         2006      1945           61
3 2.5052e+12         2006      1947           59
4 2.5052e+12         2006      1946           60
5 2.5052e+12         2006      1937           69
6 2.5052e+12         2006      1940           66

$tilda
                  id year_of_wave age_in_years year_born
1 1091                       2009           80      1929
2 1111                       2009           51      1958
3 1112                       2009           51      1958
4 1151                       2009           60      1949
5 1281                       2009           72      1937
6 1411                       2009           66      1943
```

```r
rm(lsh_age)



# age summary across studies
ds %>%
  dplyr::group_by(study_name) %>%
  na.omit() %>%
  dplyr::summarize(
    mean_age     = round(mean(age_in_years),1),
    sd_age       = round(sd(age_in_years),2),
    observed     = n(),
    min_born     = min(year_born),
    med_born     = median(year_born),
    max_born     = max(year_born)
  ) %>%
  dplyr::ungroup()
```

```
Source: local data frame [5 x 7]

  study_name mean_age sd_age observed min_born med_born max_born
      (fctr)    (dbl)  (dbl)    (int)    (dbl)    (dbl)    (dbl)
1       alsa     78.1   6.65     2053     1889     1915     1927
2       lbsl     71.3   9.92      463     1900     1923     1944
3      satsa     67.5   9.31     1087     1900     1922     1998
4      share     64.7   9.67     2467     1911     1943     1956
5      tilda     63.6   9.08     5632     1929     1946     1959
```

```
# see counts across age groups and studies
t <- table(
  cut(ds$age_in_years,breaks = c(49,seq(from=45,to=100,by=5), Inf)),
  ds$study_name,
  useNA="always"
);t[t==0]<-".";t
```

```
            alsa lbsl satsa share tilda <NA>
  (45,49]   .    .    .     .     .     .
  (49,50]   .    6    .     26    334   .
  (50,55]   .    45   162   475   1637  .
  (55,60]   .    28   126   543   1590  .
  (60,65]   13   87   168   361   1388  .
  (65,70]   258  101  222   415   1138  .
  (70,75]   552  81   235   274   884   .
  (75,80]   513  67   198   221   1192  .
  (80,85]   425  110  96    130   .     .
  (85,90]   254  43   28    43    .     .
  (90,95]   58   13   4     19    .     .
  (95,100]  12   1    1     .     .     .
  (100,Inf] 2    .    .     .     .     .
  <NA>      .    .    .     .     .     .
```

```
# now after centering
ds$age_in_years_70 <- ds$age_in_years - 70
t <- table(
  cut(ds$age_in_years_70,breaks = c(-Inf,seq(from=-25,to=30,by=5), Inf)),
  ds$study_name,
  useNA = "always"
); t[t==0] <- "."; t
```

```
             alsa lbsl satsa share tilda <NA>
  (-Inf,-25] .    .    .     .     .     .
  (-25,-20]  .    6    .     26    334   .
  (-20,-15]  .    45   162   475   1637  .
  (-15,-10]  .    28   126   543   1590  .
  (-10,-5]   13   87   168   361   1388  .
  (-5,0]     258  101  222   415   1138  .
  (0,5]      552  81   235   274   884   .
  (5,10]     513  67   198   221   1192  .
  (10,15]    425  110  96    130   .     .
  (15,20]    254  43   28    43    .     .
  (20,25]    58   13   4     19    .     .
  (25,30]    12   1    1     .     .     .
  (30, Inf]  2    .    .     .     .     .
  <NA>       .    .    .     .     .     .
```

Sex

View descriptives : sex for closer examination of each variable that contributed to the computation of the harmonized variable. f unique response vectors.

Is respondent female? - `female`

- `0` - `FALSE` - male - Reference group

- `1` - `TRUE` - female

The specific harmonization rules have been encoded over the observed frequencies

```
t <- table( ds$female, ds$study_name, useNA="always");t[t==0]<-".";t
```

```
      alsa lbsl satsa share tilda <NA>
FALSE 1056 292  506   1136  3740  .
TRUE  1031 290  734   1371  4423  .
<NA>  .    .    .     .     .     .
```

Education

View descriptives : education for closer examination of each variable that contributed to the computation of the harmonized variable.

Highest level of education achieved - `educ3`

- `-1` - `less then high school`
- `0` - `high school` - Reference group
- `1` - `more than high school`

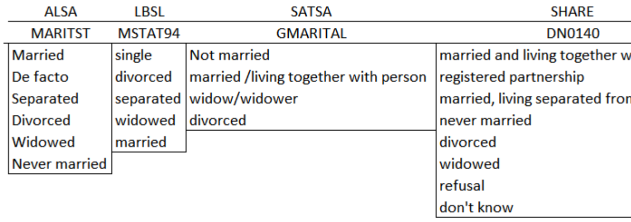
The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

```
t <- table( ds$educ3,ds$study_name, useNA="always");t[t==0]<-".";t
```

```
                      alsa lbsl satsa share tilda <NA>
high school           819  157  119   853   2607  .
less than high school 337  73   999   935   5092  .
more than high school 905  263  106   693   460   .
<NA>                  26   89   16    26    4     .
```

Marital status

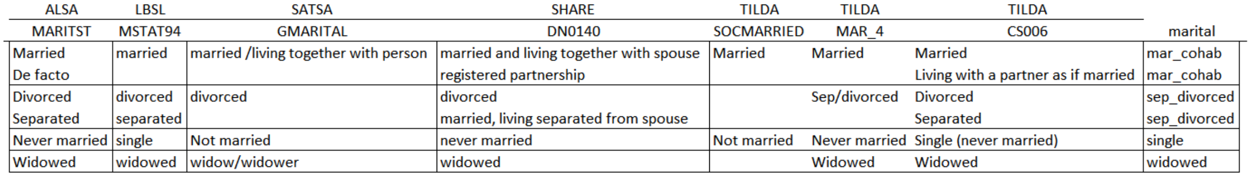
View descriptives : marital for closer examination of each variable that contributed to the computation of the harmonized variable.



ALSA	LBSL	SATSA	SHARE
MARITST	MSTAT94	GMARITAL	DN0140
Married	single	Not married	married and living together w
De facto	divorced	married /living together with person	registered partnership
Separated	separated	widow/widower	married, living separated fron
Divorced	widowed	divorced	never married
Widowed	married		divorced
Never married			widowed
			refusal
			don't know

The responses to variables loading on the construct `marital` are as such:

After reorganizing the possible repsonses, the following clustering has emerged



ALSA	LBSL	SATSA	SHARE	TILDA	TILDA	TILDA	
MARITST	MSTAT94	GMARITAL	DN0140	SOCMARRIED	MAR_4	CS006	marital
Married	married	married /living together with person	married and living together with spouse	Married	Married	Married	mar_cohab
De facto			registered partnership			Living with a partner as if married	mar_cohab
Divorced	divorced	divorced	divorced		Sep/divorced	Divorced	sep_divorced
Separated	separated		married, living separated from spouse			Separated	sep_divorced
Never married	single	Not married	never married	Not married	Never married	Single (never married)	single
Widowed	widowed	widow/widower	widowed		Widowed	Widowed	widowed

Figure 1: marital harmonized

After reviewing descriptives and relevant codebooks, the following operationalization of the harmonized variables have been adopted:

Current marital status - `marital`

- `-1` - `mar_cohab` - married or cohabiting
- `0` - `single`- not married - REFERENCE group
- `1` - `widowed` - widowed
- `2` - `sep_divorced` - separated or divorced

However, such categorization resulted in data sparseness: some categories were not populated heavily enough to allow for convergence during estimation. To address this, a simpler harmonization rule has been adopted :

Current marital status - `single`

- `0` - `FALSE` - married / living together - Reference group
- `1` - `TRUE` - single / lining along

The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

```
t <- table( ds$single,ds$study_name, useNA="always");t[t==0]<-".";t
```

```
        alsa lbsl satsa share tilda <NA>
  FALSE 1367 295  771   1961  5631  .
  TRUE  719  203  454   543   2532  .
  <NA>  1    84   15    3     .     .
```

Health (SR)

View descriptives : health for closer examination of each variable that contributed to the computation of the harmonized variable.

Does respondent report poor health? - `poor_health`

- `0` - `FALSE` - Reference group
- `1` - `TRUE` - Risk factor

The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

```
t <- table( ds$poor_health, ds$study_name, useNA="always");t[t==0]<-".";t
```

```
        alsa lbsl satsa share tilda <NA>
  FALSE 1423 306  676   1336  6263  .   
  TRUE  658  197  550   1168  1899  .   
  <NA>  6    79   14    3     1     .   
```

Physical activity

View descriptives : physact for closer examination of each variable that contributed to the computation of the harmonized variable.

Does respondent lead a sendentary lifestyle? - `sedentary`

- `0` - `FALSE` - Reference group
- `1` - `TRUE` - Risk factor

The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

The operationalization of this variable ***is not*** sensitive to the intensity of exercise. Any reponses indicating an activity at least as vigorous as ***walking*** generated values `TRUE` on the harmonized variable.

```
t <- table( ds$sedentary, ds$study_name, useNA="always");t[t==0]<-".";t
```

```
        alsa lbsl satsa share tilda <NA>
  FALSE 1250 422  465   1975  6643  .   
  TRUE  814  73   752   528   1515  .   
  <NA>  23   87   23    4     5     .   
```

Employment status

View descriptives : work for closer examination of each variable that contributed to the computation of the harmonized variable.

Is respondent currently in the work force? - `current_work_2`

- `0` - `FALSE` - Reference group
- `1` - `TRUE` - Risk factor

The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

The operationalization of this variable ***does not*** distinguish between retired and unemployed statuses.

```
t <- table( ds$current_work_2,ds$study_name,useNA="always");t[t==0]<-".";t
```

```
        alsa lbsl satsa share tilda <NA>
  FALSE 2038 372  916   1617  5094  .   
```

```
TRUE  31   118  303   882   3067  .
<NA>  18   92   21    8     2     .
```

Alcohol

View descriptives : alcohol for closer examination of each variable that contributed to the computation of the harmonized variable.

Does respondent currently consume alcohol? - `current_drink`

- `0` - `FALSE`- Reference group
- `1` - `TRUE`- Risk factor

The specific harmonization rules have been encoded over the observed frequencies of unique response vectors.

The operationalization of this variable ***is not*** sensitive to the intensity of consumption: any indications of non-abstaining generated `TRUE` values on the harmonizaed variable. It also doesn't account for the history of consumption, reflecting only the present habits.

```
t <- table( ds$current_drink,ds$study_name, useNA="always");t[t==0]<-".";t
```

```
      alsa lbsl satsa share tilda <NA>
FALSE 774  168  515   1785  1779  .
TRUE  1293 334  699   718   3859  .
<NA>  20   80   26    4     2525  .
```

Harmonized dataset

Guide to Models

Each of the following models (`A`, `B`, `AA`, and `BB`) are fitted to the data from each study separately. When fitted to the pooled data, an additional predictor, `study_name` is added after the intercept.

predictors/model	A	B	AA	BB	best
age	age_in_years	age_in_years	age_in_years	age_in_years	?
sex	female	female	female	female	?
education	educ3	educ3	educ3	educ3	?
marital status	single	single	single	single	?
health		poor_health		poor_health	?
physical activity		sedentary		sedentary	?
employment		current_work		current_work	?
alcohol use		current_drink_2		current_drink_2	?
interactions	*none*	*none*	all pairwise	all pairwise	?

Odds-ratios with 95% confidence intervals are reported. The model labeled "best" represents the solution suggested by the top ranked model from the best subset search propelled by genetic algorithm with AICC as the guiding selection criteria.

Between models

The following table reports comparison across five model types (A, B, AA, BB, best) and six datasets (alsa, lbsl, satsa, share, tilda, pooled). You can think of this as multiple tables of various heights stacked on top of each other. You select a single table by choosing the value for `study_name`. (you may need to adjust the number of entries to view, at the top left of the dynamic table)

Within models

The following table reports estimates and odds from every model that has been fit during the exercise. You can think of this as multiple tables of various heights stacked on top of each other. You select a single table by chosing the values for `study_name` and `model_type`. (you may need to adjust the number of entries to view, at the top left of the dynamic table)

Static tables

You can examine individual static table from the dynamic tables above in a stand-alone appendix report

session

```r
sessionInfo()
```

```
R version 3.2.5 (2016-04-14)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

locale:
[1] LC_COLLATE=English_United States.1252  LC_CTYPE=English_United States.1252    LC_MONETARY=English_Un
[4] LC_NUMERIC=C                           LC_TIME=English_United States.1252    

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base     

other attached packages:
[1] knitr_1.12.3  MASS_7.3-45   glmulti_1.0.7 rJava_0.9-8   ggplot2_2.1.0 magrittr_1.5 

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.5        RColorBrewer_1.1-2 formatR_1.3        plyr_1.8.3         highr_0.5.1        tool
 [7] extrafont_0.17     digest_0.6.9       jsonlite_0.9.20    evaluate_0.9       gtable_0.2.0       DBI_
[13] yaml_2.1.13        parallel_3.2.5     Rttf2pt1_1.3.3     dplyr_0.4.3        stringr_1.0.0      html
[19] grid_3.2.5         DT_0.1.40          R6_2.1.2           rmarkdown_0.9.6    tidyr_0.4.1        ext
[25] scales_0.4.0       htmltools_0.3.5    rsconnect_0.4.2.1  assertthat_0.1     dichromat_2.0-0    tes
[31] colorspace_1.2-6   stringi_1.0-1      lazyeval_0.1.10    munsell_0.4.3     
```