# Missing Value Analysis

*Yufei Gui*

*4/4/2017*

```r
#load data
dat<- read.csv('imdb_raw.csv')

# change all the empty cell to NA form
dat[dat==""]<-NA

dat<- data.frame(dat)

# number of the total observations and columns
nrow(dat)
```

```
## [1] 100
```

```r
ncol(dat)
```

```
## [1] 87
```

```r
# explore the type of the input variables
variables<-split(names(dat),sapply(dat, function(x) paste(class(x), collapse=" ")))
variables
```

```
## $factor
##  [1] "actors"
##  [2] "actresses"
##  [3] "aka"
##  [4] "also.known.as"
##  [5] "art.direction.by"
##  [6] "casting"
##  [7] "casting.by"
##  [8] "certificate"
##  [9] "certification"
## [10] "certifications"
## [11] "cinematography"
## [12] "cinematography.by"
## [13] "color"
## [14] "costume.and.wardrobe.department"
## [15] "costume.design"
## [16] "costume.design.by"
## [17] "country"
## [18] "cover"
## [19] "crew.members"
## [20] "crewmembers"
## [21] "directed.by"
## [22] "distribution"
## [23] "distribution.companies"
## [24] "distribution.company"
## [25] "distributor"
## [26] "editing"
## [27] "film.editing"
## [28] "film.editing.by"
## [29] "genre"
## [30] "lang"
## [31] "language"
## [32] "make.up"
## [33] "makeup"
## [34] "makeup.department"
## [35] "misc.companies"
## [36] "misc.company"
## [37] "misc.crew"
## [38] "miscellaneous.company"
## [39] "miscellaneouscrew"
## [40] "music"
## [41] "original.music.by"
## [42] "other.companies"
## [43] "other.company"
## [44] "other.crew"
## [45] "plot.summaries"
## [46] "plot.summary"
## [47] "produced.by"
## [48] "production.company"
## [49] "production.countries"
## [50] "production.country"
## [51] "production.management"
## [52] "runtime"
## [53] "second.unit.director"
## [54] "second.unit.director.or.assistant.director"
## [55] "set.decoration.by"
## [56] "sound.department"
## [57] "special.effects.company"
## [58] "stunts"
## [59] "visual.effects.by"
## [60] "writing.credits"
##
## $integer
## [1] "X"       "imdb_id"
##
## $logical
##  [1] "amazon.review"                "created.by"
##  [3] "episodes.cast"                "episodes.number"
##  [5] "faq"                          "frequently.asked.questions"
##  [7] "full.size.cover"              "guest"
##  [9] "guest.appearances"            "merchandise"
## [11] "merchandising"                "miscellaneous"
## [13] "miscellaneous.links"          "non.original.music.by"
## [15] "notable.tv.guest.appearances" "parental.guide"
## [17] "photographs"                  "sales"
## [19] "seasons"                      "soundclips"
## [21] "special.effects.by"           "tv.guests"
## [23] "tv.schedule"                  "videoclips"
##
## $numeric
## [1] "user.rating"
```

```r
summary(variables)
```

```
##         Length Class  Mode     
## factor  60     -none- character
## integer  2     -none- character
## logical 24     -none- character
## numeric  1     -none- character
```

```r
# explore missing rate of each variables

# define a function of explore the missing rate
propmiss <- function(dataframe) {
    m <- sapply(dataframe, function(x) {
        data.frame(
            nmiss=sum(is.na(x)),
            n=length(x),
            propmiss=sum(is.na(x))/length(x)
        )
    })
    d <- data.frame(t(m))
    d <- sapply(d, unlist)
    d <- as.data.frame(d)
    d$variable <- row.names(d)
    row.names(d) <- NULL
    d <- cbind(d[ncol(d)],d[-ncol(d)])
    return(d[order(d$propmiss), ])
}

# missing rate for the train set
propmiss(dat)
```

```
##                                       variable nmiss   n propmiss
## 1                                            X     0 100     0.00
## 2                                       actors     0 100     0.00
## 3                                    actresses     0 100     0.00
## 4                                          aka     0 100     0.00
## 5                                also.known.as     0 100     0.00
## 10                                 certificate     0 100     0.00
## 11                               certification     0 100     0.00
## 12                              certifications     0 100     0.00
## 15                                       color     0 100     0.00
## 19                                     country     0 100     0.00
## 20                                       cover     0 100     0.00
## 24                                 directed.by     0 100     0.00
## 37                                       genre     0 100     0.00
## 40                                     imdb_id     0 100     0.00
## 41                                        lang     0 100     0.00
## 42                                    language     0 100     0.00
## 64                              plot.summaries     0 100     0.00
## 65                                plot.summary     0 100     0.00
## 67                          production.company     0 100     0.00
## 68                        production.countries     0 100     0.00
## 69                          production.country     0 100     0.00
## 71                                     runtime     0 100     0.00
## 84                                user.rating   0 100      0.00
## 87                            writing.credits   0 100      0.00
## 66                                produced.by   4 100      0.04
## 13                             cinematography   8 100      0.08
## 14                          cinematography.by   8 100      0.08
## 25                               distribution   8 100      0.08
## 26                     distribution.companies   8 100      0.08
## 27                       distribution.company   8 100      0.08
## 28                                distributor   8 100      0.08
## 29                                    editing   8 100      0.08
## 33                               film.editing   8 100      0.08
## 34                            film.editing.by   8 100      0.08
## 22                               crew.members  12 100      0.12
## 23                                crewmembers  12 100      0.12
## 48                             misc.companies  12 100      0.12
## 49                               misc.company  12 100      0.12
## 50                                  misc.crew  12 100      0.12
## 52                      miscellaneous.company  12 100      0.12
## 54                          miscellaneouscrew  12 100      0.12
## 59                            other.companies  12 100      0.12
## 60                              other.company  12 100      0.12
## 61                                 other.crew  12 100      0.12
## 77                           sound.department  12 100      0.12
## 55                                      music  16 100      0.16
## 58                          original.music.by  16 100      0.16
## 70                      production.management  16 100      0.16
## 74                       second.unit.director  20 100      0.20
## 75 second.unit.director.or.assistant.director  20 100      0.20
## 43                                    make.up  24 100      0.24
## 44                                     makeup  24 100      0.24
## 45                          makeup.department  24 100      0.24
## 8                                     casting  28 100      0.28
## 9                                  casting.by  28 100      0.28
## 16            costume.and.wardrobe.department  28 100      0.28
## 17                             costume.design  28 100      0.28
## 18                          costume.design.by  28 100      0.28
## 7                            art.direction.by  32 100      0.32
## 86                          visual.effects.by  32 100      0.32
## 81                                     stunts  36 100      0.36
## 76                          set.decoration.by  40 100      0.40
## 80                    special.effects.company  48 100      0.48
## 6                               amazon.review 100 100      1.00
## 21                                 created.by 100 100      1.00
## 30                              episodes.cast 100 100      1.00
## 31                            episodes.number 100 100      1.00
## 32                                        faq 100 100      1.00
## 35                 frequently.asked.questions 100 100      1.00
## 36                            full.size.cover 100 100      1.00
## 38                                      guest 100 100      1.00
## 39                          guest.appearances 100 100      1.00
## 46                                merchandise 100 100      1.00
## 47                              merchandising 100 100      1.00
## 51                              miscellaneous 100 100      1.00
## 53                        miscellaneous.links 100 100      1.00
## 56               non.original.music.by   100 100      1.00
## 57       notable.tv.guest.appearances   100 100      1.00
## 62                      parental.guide   100 100      1.00
## 63                         photographs   100 100      1.00
## 72                               sales   100 100      1.00
## 73                             seasons   100 100      1.00
## 78                          soundclips   100 100      1.00
## 79                  special.effects.by   100 100      1.00
## 82                           tv.guests   100 100      1.00
## 83                         tv.schedule   100 100      1.00
## 85                          videoclips   100 100      1.00
```