Anly501 data cleaning

Yangyi LI

```
dataset_Name <- "marvel.csv"

# read the csv
marvel_df <- read.csv(dataset_Name)
head(marvel_df, n = 15)
```

```
##    page_id                                  name
## 1     1678             spider-man (peter parker)
## 2     7139       captain america (steven rogers)
## 3    64786 wolverine (james \\"logan\\" howlett)
## 4     1868   iron man (anthony \\"tony\\" stark)
## 5     2460                   thor (thor odinson)
## 6     2458            benjamin grimm (earth-616)
## 7     2166             reed richards (earth-616)
## 8     1833            hulk (robert bruce banner)
## 9    29481             scott summers (earth-616)
## 10    1837            jonathan storm (earth-616)
## 11   15725               henry mccoy (earth-616)
## 12    1863               susan storm (earth-616)
## 13    7823            namor mckenzie (earth-616)
## 14    2614              ororo munroe (earth-616)
## 15    1803            clinton barton (earth-616)
##                                    urlslug               id              align
## 1              \\/spider-man_(peter_parker)  secret identity    good characters
## 2         \\/captain_america_(steven_rogers)  public identity    good characters
## 3  \\/wolverine_(james_%22logan%22_howlett)  public identity neutral characters
## 4    \\/iron_man_(anthony_%22tony%22_stark)  public identity    good characters
## 5                     \\/thor_(thor_odinson) no dual identity    good characters
## 6             \\/benjamin_grimm_(earth-616)  public identity    good characters
## 7              \\/reed_richards_(earth-616)  public identity    good characters
## 8             \\/hulk_(robert_bruce_banner)  public identity    good characters
## 9              \\/scott_summers_(earth-616)  public identity neutral characters
## 10            \\/jonathan_storm_(earth-616)  public identity    good characters
## 11               \\/henry_mccoy_(earth-616)  public identity    good characters
## 12               \\/susan_storm_(earth-616)  public identity    good characters
## 13            \\/namor_mckenzie_(earth-616) no dual identity neutral characters
## 14              \\/ororo_munroe_(earth-616)  public identity    good characters
## 15            \\/clinton_barton_(earth-616)  public identity    good characters
##           eye       hair               sex gsm             alive appearances
## 1  hazel eyes brown hair   male characters     living characters        4043
## 2   blue eyes white hair   male characters     living characters        3360
## 3   blue eyes black hair   male characters     living characters        3061
## 4   blue eyes black hair   male characters     living characters        2961
## 5   blue eyes blond hair   male characters     living characters        2258
## 6   blue eyes    no hair   male characters     living characters        2255
## 7  brown eyes brown hair   male characters     living characters        2072
## 8  brown eyes brown hair   male characters     living characters        2017
## 9  brown eyes brown hair   male characters     living characters        1955
## 10  blue eyes blond hair   male characters     living characters        1934
## 11  blue eyes  blue hair   male characters     living characters        1825
## 12  blue eyes blond hair female characters     living characters        1713
## 13 green eyes black hair   male characters     living characters        1528
## 14  blue eyes white hair female characters     living characters        1512
## 15  blue eyes blond hair   male characters     living characters        1394
##    first.appearance year
## 1            aug-62 1962
## 2            mar-41 1941
## 3            oct-74 1974
## 4            mar-63 1963
```

```
## 5            nov-50 1950
## 6            nov-61 1961
## 7            nov-61 1961
## 8            may-62 1962
## 9            sep-63 1963
## 10           nov-61 1961
## 11           sep-63 1963
## 12           nov-61 1961
## 13                  NA
## 14           may-75 1975
## 15           sep-64 1964
```

```
# check data types
str(marvel_df)
```

```
## 'data.frame':    16376 obs. of  13 variables:
##  $ page_id         : int  1678 7139 64786 1868 2460 2458 2166 1833 29481 1837 ...
##  $ name            : chr  "spider-man (peter parker)" "captain america (steven roger
s)" "wolverine (james \\\"logan\\\" howlett)" "iron man (anthony \\\"tony\\\" stark)"
...
##  $ urlslug         : chr  "\\/spider-man_(peter_parker)" "\\/captain_america_(steven_
rogers)" "\\/wolverine_(james_%22logan%22_howlett)" "\\/iron_man_(anthony_%22tony%22_sta
rk)" ...
##  $ id              : chr  "secret identity" "public identity" "public identity" "publ
ic identity" ...
##  $ align           : chr  "good characters" "good characters" "neutral characters" "g
ood characters" ...
##  $ eye             : chr  "hazel eyes" "blue eyes" "blue eyes" "blue eyes" ...
##  $ hair            : chr  "brown hair" "white hair" "black hair" "black hair" ...
##  $ sex             : chr  "male characters" "male characters" "male characters" "male
characters" ...
##  $ gsm             : chr  "" "" "" "" ...
##  $ alive           : chr  "living characters" "living characters" "living characters"
"living characters" ...
##  $ appearances     : int  4043 3360 3061 2961 2258 2255 2072 2017 1955 1934 ...
##  $ first.appearance: chr  "aug-62" "mar-41" "oct-74" "mar-63" ...
##  $ year            : int  1962 1941 1974 1963 1950 1961 1961 1962 1963 1961 ...
```

```
# change variables to lower case
names(marvel_df)[1:13] <- tolower(names(marvel_df)[1:13])

# check colnames
(ColNames<-names(marvel_df))
```

```
##  [1] "page_id"          "name"             "urlslug"          "id"
##  [5] "align"            "eye"              "hair"             "sex"
##  [9] "gsm"              "alive"            "appearances"      "first.appearance"
## [13] "year"
```

```
# There are some data I may not want have right now
# drop the column "page_id", "urlslug", first.appearance
marvel_df <- marvel_df %>% select(name, id, align, eye, hair, sex, gsm, alive, appearanc
es, year)
head(marvel_df)
```

```
##                                   name               id              align
## 1             spider-man (peter parker)  secret identity    good characters
## 2       captain america (steven rogers)  public identity    good characters
## 3 wolverine (james \\"logan\\" howlett)  public identity neutral characters
## 4   iron man (anthony \\"tony\\" stark)  public identity    good characters
## 5                   thor (thor odinson) no dual identity    good characters
## 6            benjamin grimm (earth-616)  public identity    good characters
##          eye       hair             sex gsm             alive appearances year
## 1 hazel eyes brown hair male characters     living characters        4043 1962
## 2  blue eyes white hair male characters     living characters        3360 1941
## 3  blue eyes black hair male characters     living characters        3061 1974
## 4  blue eyes black hair male characters     living characters        2961 1963
## 5  blue eyes blond hair male characters     living characters        2258 1950
## 6  blue eyes    no hair male characters     living characters        2255 1961
```

```
# when I check colnames and datatype, gsm seems blank
# check gsm
marvel_df$gsm[1:100]
```

```
##   [1] ""                    ""                    ""
##   [4] ""                    ""                    ""
##   [7] ""                    ""                    ""
##  [10] ""                    ""                    ""
##  [13] ""                    ""                    ""
##  [16] ""                    ""                    ""
##  [19] ""                    ""                    ""
##  [22] ""                    ""                    ""
##  [25] ""                    ""                    ""
##  [28] ""                    "bisexual characters" ""
##  [31] ""                    ""                    ""
##  [34] ""                    ""                    ""
##  [37] ""                    ""                    ""
##  [40] ""                    ""                    ""
##  [43] "bisexual characters" ""                    ""
##  [46] "bisexual characters" ""                    ""
##  [49] ""                    ""                    ""
##  [52] ""                    ""                    ""
##  [55] ""                    ""                    ""
##  [58] ""                    ""                    ""
##  [61] ""                    "bisexual characters" ""
##  [64] ""                    ""                    ""
##  [67] ""                    ""                    ""
##  [70] ""                    ""                    "transvestites"
##  [73] ""                    ""                    ""
##  [76] ""                    ""                    ""
##  [79] ""                    ""                    ""
##  [82] ""                    ""                    ""
##  [85] ""                    ""                    ""
##  [88] ""                    ""                    ""
##  [91] ""                    ""                    ""
##  [94] ""                    ""                    ""
##  [97] ""                    ""                    ""
## [100] ""
```

```
# There are some specific case like bisexual characters and homosexual characters
# let's set Heterosexual characters for default
marvel_df$gsm <- ifelse(marvel_df$gsm == "",
                        "heterosexual characters", marvel_df$gsm)

# check it
marvel_df$gsm[1:100]
```

```
##   [1] "heterosexual characters" "heterosexual characters"
##   [3] "heterosexual characters" "heterosexual characters"
##   [5] "heterosexual characters" "heterosexual characters"
##   [7] "heterosexual characters" "heterosexual characters"
##   [9] "heterosexual characters" "heterosexual characters"
##  [11] "heterosexual characters" "heterosexual characters"
##  [13] "heterosexual characters" "heterosexual characters"
##  [15] "heterosexual characters" "heterosexual characters"
##  [17] "heterosexual characters" "heterosexual characters"
##  [19] "heterosexual characters" "heterosexual characters"
##  [21] "heterosexual characters" "heterosexual characters"
##  [23] "heterosexual characters" "heterosexual characters"
##  [25] "heterosexual characters" "heterosexual characters"
##  [27] "heterosexual characters" "heterosexual characters"
##  [29] "bisexual characters"     "heterosexual characters"
##  [31] "heterosexual characters" "heterosexual characters"
##  [33] "heterosexual characters" "heterosexual characters"
##  [35] "heterosexual characters" "heterosexual characters"
##  [37] "heterosexual characters" "heterosexual characters"
##  [39] "heterosexual characters" "heterosexual characters"
##  [41] "heterosexual characters" "heterosexual characters"
##  [43] "bisexual characters"     "heterosexual characters"
##  [45] "heterosexual characters" "bisexual characters"
##  [47] "heterosexual characters" "heterosexual characters"
##  [49] "heterosexual characters" "heterosexual characters"
##  [51] "heterosexual characters" "heterosexual characters"
##  [53] "heterosexual characters" "heterosexual characters"
##  [55] "heterosexual characters" "heterosexual characters"
##  [57] "heterosexual characters" "heterosexual characters"
##  [59] "heterosexual characters" "heterosexual characters"
##  [61] "heterosexual characters" "bisexual characters"
##  [63] "heterosexual characters" "heterosexual characters"
##  [65] "heterosexual characters" "heterosexual characters"
##  [67] "heterosexual characters" "heterosexual characters"
##  [69] "heterosexual characters" "heterosexual characters"
##  [71] "heterosexual characters" "transvestites"
##  [73] "heterosexual characters" "heterosexual characters"
##  [75] "heterosexual characters" "heterosexual characters"
##  [77] "heterosexual characters" "heterosexual characters"
##  [79] "heterosexual characters" "heterosexual characters"
##  [81] "heterosexual characters" "heterosexual characters"
##  [83] "heterosexual characters" "heterosexual characters"
##  [85] "heterosexual characters" "heterosexual characters"
##  [87] "heterosexual characters" "heterosexual characters"
##  [89] "heterosexual characters" "heterosexual characters"
##  [91] "heterosexual characters" "heterosexual characters"
##  [93] "heterosexual characters" "heterosexual characters"
##  [95] "heterosexual characters" "heterosexual characters"
##  [97] "heterosexual characters" "heterosexual characters"
##  [99] "heterosexual characters" "heterosexual characters"
```

```
head(marvel_df)
```

```
##                                    name               id              align
## 1             spider-man (peter parker)  secret identity    good characters
## 2       captain america (steven rogers)  public identity    good characters
## 3 wolverine (james \\"logan\\" howlett)  public identity neutral characters
## 4   iron man (anthony \\"tony\\" stark)  public identity    good characters
## 5                   thor (thor odinson) no dual identity    good characters
## 6            benjamin grimm (earth-616)  public identity    good characters
##          eye       hair             sex                    gsm
## 1 hazel eyes brown hair male characters heterosexual characters
## 2  blue eyes white hair male characters heterosexual characters
## 3  blue eyes black hair male characters heterosexual characters
## 4  blue eyes black hair male characters heterosexual characters
## 5  blue eyes blond hair male characters heterosexual characters
## 6  blue eyes    no hair male characters heterosexual characters
##               alive appearances year
## 1 living characters        4043 1962
## 2 living characters        3360 1941
## 3 living characters        3061 1974
## 4 living characters        2961 1963
## 5 living characters        2258 1950
## 6 living characters        2255 1961
```

```
# Let's make summary of all the columns
lapply(marvel_df,summary)
```

```
## $name
##    Length     Class      Mode
##     16376 character character
##
## $id
##    Length     Class      Mode
##     16376 character character
##
## $align
##    Length     Class      Mode
##     16376 character character
##
## $eye
##    Length     Class      Mode
##     16376 character character
##
## $hair
##    Length     Class      Mode
##     16376 character character
##
## $sex
##    Length     Class      Mode
##     16376 character character
##
## $gsm
##    Length     Class      Mode
##     16376 character character
##
## $alive
##    Length     Class      Mode
##     16376 character character
##
## $appearances
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    1.00    3.00   17.03    8.00 4043.00    1096
##
## $year
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1939    1974    1990    1985    2000    2013     815
```

```
# we can see some missing values in our datasets
sum(is.na(marvel_df$name))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$id))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$align))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$eye))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$hair))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$sex))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$gsm))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$alive))
```

```
## [1] 0
```

```
sum(is.na(marvel_df$appearances)) # exist missing value
```

```
## [1] 1096
```

```
sum(is.na(marvel_df$year))  #exist missing value
```

```
## [1] 815
```

```
# since we know this marvel characters. set the appearances for 0
# the character may never show up just exit in conversation
marvel_df$appearances <- ifelse(is.na(marvel_df$appearances),
                         0, marvel_df$appearances)

sum(is.na(marvel_df$appearances))
```

```
## [1] 0
```

```
# I do not want to set the default year since I do not know when did this character firs
t show up
# Let's set it "need to drop" so I will remember the missing values while analyze the ti
meline
# I do not drop it right now because the whole row I may want to analyze in the future
marvel_df$year <- ifelse(is.na(marvel_df$year),
                         "need to drop", marvel_df$year)
sum(is.na(marvel_df$year))
```

```
## [1] 0
```

```
head(marvel_df)
```

```
##                                    name               id              align
## 1             spider-man (peter parker)  secret identity    good characters
## 2       captain america (steven rogers)  public identity    good characters
## 3 wolverine (james \\"logan\\" howlett)  public identity neutral characters
## 4   iron man (anthony \\"tony\\" stark)  public identity    good characters
## 5                   thor (thor odinson) no dual identity    good characters
## 6            benjamin grimm (earth-616)  public identity    good characters
##          eye       hair             sex                       gsm
## 1 hazel eyes brown hair male characters heterosexual characters
## 2  blue eyes white hair male characters heterosexual characters
## 3  blue eyes black hair male characters heterosexual characters
## 4  blue eyes black hair male characters heterosexual characters
## 5  blue eyes blond hair male characters heterosexual characters
## 6  blue eyes    no hair male characters heterosexual characters
##               alive appearances year
## 1 living characters        4043 1962
## 2 living characters        3360 1941
## 3 living characters        3061 1974
## 4 living characters        2961 1963
## 5 living characters        2258 1950
## 6 living characters        2255 1961
```

```
# write to a csv file
write.csv(marvel_df, "marvel_df.csv")
```

```
# The mean of average time that marvel character appeared
mean(marvel_df$appearances)
```

```
## [1] 15.89338
```

```
# The standard deviation of appearances
sd(marvel_df$appearances)
```

```
## [1] 93.18919
```

```
# How many characters
nrow(marvel_df)
```

```
## [1] 16376
```

```
# When did marvel universe created
min(marvel_df$year)
```

```
## [1] "1939"
```