Data Cleaning

LANGAT ERICK

2023-12-28

## Data Cleaning

#Explore  
#Claen  
#Manipulate  
#Describe & Summarize  
#Visualiaze  
#Analyze

#install libraries  
suppressPackageStartupMessages(library(tidyverse))  
View(starwars)  
data1=starwars  
# write.csv(data1, file = "starwars.csv")

#variables types  
glimpse(starwars)

## Rows: 87  
## Columns: 14  
## $ name <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or…  
## $ height <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2…  
## $ mass <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.…  
## $ hair\_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N…  
## $ skin\_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "…  
## $ eye\_color <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",…  
## $ birth\_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, …  
## $ sex <chr> "male", "none", "none", "male", "female", "male", "female",…  
## $ gender <chr> "masculine", "masculine", "masculine", "masculine", "femini…  
## $ homeworld <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T…  
## $ species <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma…  
## $ films <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return…  
## $ vehicles <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp…  
## $ starships <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",…

class(starwars$gender)

## [1] "character"

unique(starwars$gender)

## [1] "masculine" "feminine" NA

starwars$gender <- as.factor(starwars$gender)  
class(starwars$gender)

## [1] "factor"

levels(starwars$gender)

## [1] "feminine" "masculine"

starwars$gender <- factor((starwars$gender), levels = c("masculine", 'feminine'))#changinging level arrangement  
   
  
#select variables  
starwars %>% select(name, height, ends\_with("color")) %>%   
 names()

## [1] "name" "height" "hair\_color" "skin\_color" "eye\_color"

#filter observation  
unique(starwars$hair\_color)

## [1] "blond" NA "none" "brown"   
## [5] "brown, grey" "black" "auburn, white" "auburn, grey"   
## [9] "white" "grey" "auburn" "blonde"   
## [13] "unknown"

starwars %>% select(name, height, ends\_with("color")) %>%   
 filter(hair\_color %in% c("blond","brown") & height<180)

## # A tibble: 9 × 5  
## name height hair\_color skin\_color eye\_color  
## <chr> <int> <chr> <chr> <chr>   
## 1 Luke Skywalker 172 blond fair blue   
## 2 Leia Organa 150 brown light brown   
## 3 Beru Whitesun lars 165 brown light blue   
## 4 Wedge Antilles 170 brown fair hazel   
## 5 Wicket Systri Warrick 88 brown brown brown   
## 6 Finis Valorum 170 blond fair blue   
## 7 Cordé 157 brown light brown   
## 8 Dormé 165 brown light brown   
## 9 Padmé Amidala 165 brown light brown

#missing data  
mean(starwars$height, na.rm = TRUE)

## [1] 174.358

starwars %>% select(name, gender, hair\_color, height)

## # A tibble: 87 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 Luke Skywalker masculine blond 172  
## 2 C-3PO masculine <NA> 167  
## 3 R2-D2 masculine <NA> 96  
## 4 Darth Vader masculine none 202  
## 5 Leia Organa feminine brown 150  
## 6 Owen Lars masculine brown, grey 178  
## 7 Beru Whitesun lars feminine brown 165  
## 8 R5-D4 masculine <NA> 97  
## 9 Biggs Darklighter masculine black 183  
## 10 Obi-Wan Kenobi masculine auburn, white 182  
## # ℹ 77 more rows

starwars %>% select(name, gender, hair\_color, height) %>%   
 na.omit()

## # A tibble: 73 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 Luke Skywalker masculine blond 172  
## 2 Darth Vader masculine none 202  
## 3 Leia Organa feminine brown 150  
## 4 Owen Lars masculine brown, grey 178  
## 5 Beru Whitesun lars feminine brown 165  
## 6 Biggs Darklighter masculine black 183  
## 7 Obi-Wan Kenobi masculine auburn, white 182  
## 8 Anakin Skywalker masculine blond 188  
## 9 Wilhuff Tarkin masculine auburn, grey 180  
## 10 Chewbacca masculine brown 228  
## # ℹ 63 more rows

starwars %>% select(name, gender, hair\_color, height) %>%   
 filter(complete.cases(.))

## # A tibble: 73 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 Luke Skywalker masculine blond 172  
## 2 Darth Vader masculine none 202  
## 3 Leia Organa feminine brown 150  
## 4 Owen Lars masculine brown, grey 178  
## 5 Beru Whitesun lars feminine brown 165  
## 6 Biggs Darklighter masculine black 183  
## 7 Obi-Wan Kenobi masculine auburn, white 182  
## 8 Anakin Skywalker masculine blond 188  
## 9 Wilhuff Tarkin masculine auburn, grey 180  
## 10 Chewbacca masculine brown 228  
## # ℹ 63 more rows

starwars %>% select(name, gender, hair\_color, height) %>%   
 filter(!complete.cases(.))

## # A tibble: 14 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 C-3PO masculine <NA> 167  
## 2 R2-D2 masculine <NA> 96  
## 3 R5-D4 masculine <NA> 97  
## 4 Greedo masculine <NA> 173  
## 5 Jabba Desilijic Tiure masculine <NA> 175  
## 6 Arvel Crynyd masculine brown NA  
## 7 Ric Olié <NA> brown 183  
## 8 Quarsh Panaka <NA> black 183  
## 9 Sly Moore <NA> none 178  
## 10 Finn masculine black NA  
## 11 Rey feminine brown NA  
## 12 Poe Dameron masculine brown NA  
## 13 BB8 masculine none NA  
## 14 Captain Phasma <NA> unknown NA

starwars %>% select(name, gender, hair\_color, height) %>%   
 filter(!complete.cases(.)) %>% drop\_na(height)

## # A tibble: 8 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 C-3PO masculine <NA> 167  
## 2 R2-D2 masculine <NA> 96  
## 3 R5-D4 masculine <NA> 97  
## 4 Greedo masculine <NA> 173  
## 5 Jabba Desilijic Tiure masculine <NA> 175  
## 6 Ric Olié <NA> brown 183  
## 7 Quarsh Panaka <NA> black 183  
## 8 Sly Moore <NA> none 178

starwars %>% select(name, gender, hair\_color, height) %>%   
 drop\_na(height)

## # A tibble: 81 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 Luke Skywalker masculine blond 172  
## 2 C-3PO masculine <NA> 167  
## 3 R2-D2 masculine <NA> 96  
## 4 Darth Vader masculine none 202  
## 5 Leia Organa feminine brown 150  
## 6 Owen Lars masculine brown, grey 178  
## 7 Beru Whitesun lars feminine brown 165  
## 8 R5-D4 masculine <NA> 97  
## 9 Biggs Darklighter masculine black 183  
## 10 Obi-Wan Kenobi masculine auburn, white 182  
## # ℹ 71 more rows

starwars %>% select(name, gender, hair\_color, height) %>%   
 filter(!complete.cases(.)) %>% mutate(hair\_color= replace\_na(hair\_color, 'none'))

## # A tibble: 14 × 4  
## name gender hair\_color height  
## <chr> <fct> <chr> <int>  
## 1 C-3PO masculine none 167  
## 2 R2-D2 masculine none 96  
## 3 R5-D4 masculine none 97  
## 4 Greedo masculine none 173  
## 5 Jabba Desilijic Tiure masculine none 175  
## 6 Arvel Crynyd masculine brown NA  
## 7 Ric Olié <NA> brown 183  
## 8 Quarsh Panaka <NA> black 183  
## 9 Sly Moore <NA> none 178  
## 10 Finn masculine black NA  
## 11 Rey feminine brown NA  
## 12 Poe Dameron masculine brown NA  
## 13 BB8 masculine none NA  
## 14 Captain Phasma <NA> unknown NA

#Recoding Variables  
  
starwars %>% select(name, gender) %>%   
 mutate(gender= recode(gender, 'masculine'=1, 'feminine'=2))

## # A tibble: 87 × 2  
## name gender  
## <chr> <dbl>  
## 1 Luke Skywalker 1  
## 2 C-3PO 1  
## 3 R2-D2 1  
## 4 Darth Vader 1  
## 5 Leia Organa 2  
## 6 Owen Lars 1  
## 7 Beru Whitesun lars 2  
## 8 R5-D4 1  
## 9 Biggs Darklighter 1  
## 10 Obi-Wan Kenobi 1  
## # ℹ 77 more rows