Cleaning Data With R

Similoluwa Soremekun

In this document, I will use different methods to clean the starwars dataset. This dataset comes with the tidyverse library as does hundreds of others.

Let's load the tidyverse library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Next let's load the data

```
head(starwars)
```

```
## # A tibble: 6 x 14
##   name         height  mass hair_~1 skin_~2 eye_c~3 birth~4 sex   gender homew~5
##   <chr>         <int> <dbl> <chr>   <chr>   <chr>     <dbl> <chr> <chr>  <chr>
## 1 Luke Skywal~    172    77 blond   fair    blue       19   male  mascu~ Tatooi~
## 2 C-3PO           167    75 <NA>    gold    yellow    112   none  mascu~ Tatooi~
## 3 R2-D2            96    32 <NA>    white,~ red        33   none  mascu~ Naboo
## 4 Darth Vader     202   136 none    white   yellow     41.9 male  mascu~ Tatooi~
## 5 Leia Organa     150    49 brown   light   brown      19   fema~ femin~ Aldera~
## 6 Owen Lars       178   120 brown,~ light   blue       52   male  mascu~ Tatooi~
## # ... with 4 more variables: species <chr>, films <list>, vehicles <list>,
## #   starships <list>, and abbreviated variable names 1: hair_color,
## #   2: skin_color, 3: eye_color, 4: birth_year, 5: homeworld
```

I will check the variable types to better understand the data

```
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",~
```

The data shows that variables like name, hair_color, skin_color, eye_color, sex, gender, homeworld and species are characters.

Height, is the only integer, while films, vehicles and starships are all list items. However, we have variables like gender which i would like to make into ordered categorical data (factor). First let's find the unique types in gender.

```
unique(starwars$gender)
```

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

We observe masculine, feminine and empty data (NA). Now let's change gender into a factor variable.

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

```
## [1] "factor"
```

Note that starwars$gender has been replaced in the entire dataset.If you want to use the gender category as a character type later, you might want to give the as.factor code a distinct name. Now the reason I made gender a factor is that I am interested in levels. Let's check it out.

```
levels(starwars$gender)
```

```
## [1] "feminine"  "masculine"
```

We have feminine as level 1 and masculine as level 2. We can always change the levels as needed

```
starwars$gender <- factor((starwars$gender), levels = c('masculine', 'feminine'))
levels(starwars$gender)
```

```
## [1] "masculine" "feminine"
```

The dataset is large so I will filter for relevant data. I'll go with name, height and any column that ends with color

```
starwars %>%
  select(name, height, ends_with('color'))
```

```
## # A tibble: 87 x 5
##    name               height hair_color    skin_color  eye_color
##    <chr>               <int> <chr>         <chr>       <chr>
##  1 Luke Skywalker        172 blond         fair        blue
##  2 C-3PO                 167 <NA>          gold        yellow
##  3 R2-D2                  96 <NA>          white, blue red
##  4 Darth Vader           202 none          white       yellow
##  5 Leia Organa           150 brown         light       brown
##  6 Owen Lars             178 brown, grey   light       blue
##  7 Beru Whitesun lars    165 brown         light       blue
##  8 R5-D4                  97 <NA>          white, red  red
##  9 Biggs Darklighter     183 black         light       brown
## 10 Obi-Wan Kenobi        182 auburn, white fair        blue-gray
## # ... with 77 more rows
```

I can further filter this data. Let's use hair color

```
unique(starwars$hair_color)
```

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

There's quite a lot of different hair colors. I'll trim it down to blond and brown

```
starwars %>%
  select(name, height, ends_with('color')) %>%
  filter(hair_color %in% c('blond', 'brown', 'blonde') & height <180)
```

```
## # A tibble: 10 x 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 Zam Wesell               168 blonde     fair, green, yellow yellow
## 10 Padmé Amidala            165 brown      light               brown
```

I also filtered for height less than 180cm.

Dealing With Missing Data

Let's try to look for the mean of all heights

```
mean(starwars$height)
```

```
## [1] NA
```

As expected it does not work, because we have missing data. We have to modify the code to deal with the empty values

```
mean(starwars$height, na.rm = TRUE)
```

```
## [1] 174.358
```

The mean becomes **174.358** because the missing values have been removed. While this can work in some cases, it is often not advised. A better way is done below

```
starwars %>%
  select(name, gender, hair_color, height)
```

```
## # A tibble: 87 x 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
## # ... with 77 more rows
```

To understand the mising data above, it's always good to filter fot it

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.))
```

```
## # A tibble: 14 x 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
```

The next step is where the domain knowledge comes in. From the data, there is missing gender, hair color and height. Intuitively, everyone and everything has a height. So the NAs are data that have not been captured, and as such can be removed. Hair color however is different. The observations with missing hair color are droids and just don't have hair. We can always make this **none** and preserve other details of the observation.

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.)) %>%
  drop_na(height)
```

```
## # A tibble: 8 x 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
```

Replacing hair color

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.)) %>%
  drop_na(height) %>%
  mutate(hair_color= replace_na(hair_color, 'none'))
```

```
## # A tibble: 8 x 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 Ric Olié              <NA>      brown         183
## 7 Quarsh Panaka         <NA>      black         183
## 8 Sly Moore             <NA>      none          178
```

We can also make it a new variable

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.)) %>%
  drop_na(height) %>%
  mutate(hair_color2= replace_na(hair_color, 'none'))
```

```
## # A tibble: 8 x 5
##   name                  gender    hair_color height hair_color2
##   <chr>                 <fct>     <chr>       <int> <chr>
## 1 C-3PO                 masculine <NA>          167 none
## 2 R2-D2                 masculine <NA>           96 none
## 3 R5-D4                 masculine <NA>           97 none
## 4 Greedo                masculine <NA>          173 none
## 5 Jabba Desilijic Tiure masculine <NA>          175 none
## 6 Ric Olié              <NA>      brown         183 brown
## 7 Quarsh Panaka         <NA>      black         183 black
## 8 Sly Moore             <NA>      none          178 none
```

Finally just a little re-coding

```
starwars %>%
  select(name, gender) %>%
  mutate(gender_code = recode(gender,
                          'masculine' = 1,
                          'feminine' = 2 ))
```

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

I have tried to make this as explanatory as possible so that anyone that comes accross this can follow it. Thank you.

[1]

[1] This document was made with R markdown