Data cleaning in R

vasanth

2/17/2022

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

R Markdown

```
view(starwars)
```

```
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",~
```

Rename the colum name

```
starwars %>%
  rename("movies"= "films") %>%
  glimpse()
```

```
## 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~
## $ movies     <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",~
```

Reordering the dataframe

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

```
## # A tibble: 87 x 14
##    name    gender  height  mass hair_color skin_color eye_color birth_year sex
##    <chr>   <chr>    <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
##  1 Luke S~ mascul~    172    77 blond      fair       blue            19   male
##  2 C-3PO   mascul~    167    75 <NA>       gold       yellow         112   none
##  3 R2-D2   mascul~     96    32 <NA>       white, bl~ red             33   none
##  4 Darth ~ mascul~    202   136 none       white      yellow          41.9 male
##  5 Leia O~ femini~    150    49 brown      light      brown           19   fema~
##  6 Owen L~ mascul~    178   120 brown, gr~ light      blue            52   male
##  7 Beru W~ femini~    165    75 brown      light      blue            47   fema~
##  8 R5-D4   mascul~     97    32 <NA>       white, red red             NA   none
##  9 Biggs ~ mascul~    183    84 black      light      brown           24   male
## 10 Obi-Wa~ mascul~    182    77 auburn, w~ fair       blue-gray       57   male
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

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

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

##Because of catergorical value we change into datatype (factor)

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

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

Filter values

```
starwars %>%
  select(name,height, ends_with("color")) %>%
  filter(hair_color %in% c("blonde", "brown"),height <175)
```

```
## # A tibble: 8 x 5
##   name                  height hair_color skin_color          eye_color
##   <chr>                  <int> <chr>      <chr>               <chr>
## 1 Leia Organa              150 brown      light               brown
## 2 Beru Whitesun lars       165 brown      light               blue
## 3 Wedge Antilles           170 brown      fair                hazel
## 4 Wicket Systri Warrick     88 brown      brown               brown
## 5 Cordé                    157 brown      light               brown
## 6 Dormé                    165 brown      light               brown
## 7 Zam Wesell               168 blonde     fair, green, yellow yellow
## 8 Padmé Amidala            165 brown      light               brown
```

Mean of Height for drop values

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

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

Filter NA Values

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

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

Delete the NA values in height coloum

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

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

Replace NA into none

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

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

Recoding variable

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

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