Data Wrangling

Prof. Eric A. Suess

August 26, 2020

Data Wrangling

Today we will get started with Data Wrangling.

Data Wrangling is the process of tidying into usable forms.

The R package that will be using from the tidyverse is the dplyr package.

The grammar of data wrangling

The 5 verbs of data wrangling

select() # take a subset of columns

filter() # take a subset of rows

mutate() # add or modify existing columns

arrange() # sort the rows

summarize() # aggregate the data across rows

RStudio Cheatsheet for dplyr

The RStudio dplyr cheatsheet is very useful.

Star Wars examples

```
library(tidyverse)
data("starwars")
glimpse(starwars)
```

```
## Observations: 87
## Variables: 13
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97,
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0,
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "br
## $ skin_color <chr> "fair", "gold", "white, blue", "white
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "b
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0,
## $ gender     <chr> "male", NA, NA, "male", "female", "ma
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tat
## $ species    <chr> "Human", "Droid", "Droid", "Human", "
## $ films      <list> [<"Revenge of the Sith", "Return of
## $ vehicles   <list> [<"Snowspeeder", "Imperial Speeder B
```

Star Wars

```
starwars %>% select(name, species)
```

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

Star Wars

```
starwars %>%
  filter(species == "Droid")
```

```
## # A tibble: 5 x 13
##   name  height  mass hair_color skin_color eye_color bi
##   <chr>  <int> <dbl> <chr>      <chr>       <chr>
## 1 C-3PO    167    75 <NA>       gold        yellow
## 2 R2-D2     96    32 <NA>       white, bl~ red
## 3 R5-D4     97    32 <NA>       white, red red
## 4 IG-88    200   140 none       metal       red
## 5 BB8       NA    NA none       none        black
## # ... with 4 more variables: species <chr>, films <list>
## #   starships <list>
```

Star Wars

```
starwars %>%
  select(name, ends_with("color"))
```

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

Star Wars

```
starwars %>%
  mutate(name, bmi = mass / ((height / 100)  ^ 2)) %>%
  select(name:mass, bmi)
```

```
## # A tibble: 87 x 4
##    name                height  mass   bmi
##    <chr>                <int> <dbl> <dbl>
##  1 Luke Skywalker         172    77  26.0
##  2 C-3PO                  167    75  26.9
##  3 R2-D2                   96    32  34.7
##  4 Darth Vader            202   136  33.3
##  5 Leia Organa            150    49  21.8
##  6 Owen Lars              178   120  37.9
##  7 Beru Whitesun lars     165    75  27.5
##  8 R5-D4                   97    32  34.0
##  9 Biggs Darklighter      183    84  25.1
## 10 Obi-Wan Kenobi         182    77  23.2
## # ... with 77 more rows
```

Star Wars

```
starwars %>%
  arrange(desc(mass))
```

```
## # A tibble: 87 x 13
##    name  height  mass hair_color skin_color eye_color bi
##    <chr>  <int> <dbl> <chr>      <chr>      <chr>
##  1 Jabb~    175  1358 <NA>       green-tan~ orange
##  2 Grie~    216   159 none       brown, wh~ green, y~
##  3 IG-88    200   140 none       metal      red
##  4 Dart~    202   136 none       white      yellow
##  5 Tarf~    234   136 brown      brown      blue
##  6 Owen~    178   120 brown, gr~ light      blue
##  7 Bossk    190   113 none       green      red
##  8 Chew~    228   112 brown      unknown    blue
##  9 Jek ~    180   110 brown      fair       blue
## 10 Dext~    198   102 none       brown      yellow
## # ... with 77 more rows, and 5 more variables: homeworld
## #   films <list>, vehicles <list>, starships <list>
```

Star Wars

```
starwars %>%
  group_by(species) %>%
  summarise(
    n = n(),
    mass = mean(mass, na.rm = TRUE)
  ) %>%
  filter(n > 1)
```

```
## # A tibble: 9 x 3
##   species      n  mass
##   <chr>    <int> <dbl>
## 1 Droid        5  69.8
## 2 Gungan       3  74
## 3 Human       35  82.8
## 4 Kaminoan     2  88
## 5 Mirialan     2  53.1
## 6 Twi'lek      2  55
## 7 Wookiee      2 124
```

Presidential examples

Try the code from the book in Section 4.1

```
presidential
```

```
## # A tibble: 11 x 4
##    name       start      end        party
##    <chr>      <date>     <date>     <chr>
##  1 Eisenhower 1953-01-20 1961-01-20 Republican
##  2 Kennedy    1961-01-20 1963-11-22 Democratic
##  3 Johnson    1963-11-22 1969-01-20 Democratic
##  4 Nixon      1969-01-20 1974-08-09 Republican
##  5 Ford       1974-08-09 1977-01-20 Republican
##  6 Carter     1977-01-20 1981-01-20 Democratic
##  7 Reagan     1981-01-20 1989-01-20 Republican
##  8 Bush       1989-01-20 1993-01-20 Republican
##  9 Clinton    1993-01-20 2001-01-20 Democratic
## 10 Bush       2001-01-20 2009-01-20 Republican
## 11 Obama      2009-01-20 2017-01-20 Democratic
```