Summary Data

AUTHOR
Jeremy Mikecz

Get Summary Data (Practice with preloaded datasets)

1. Retrieve Pre-loaded Datasets

Base R and some of its core data science packages (like tidyverse and its dependencies) come with preloaded datasets. We can use these preloaded to practice our data science skills.

First, we must install and run **dplyr**:

```
#install.packages("tidyverse")
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.4.2


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

to view the available preloaded datasets, run the following cell:

```
#data()
```

In this lesson we will work with the **Star Wars** dataset (`starwars`) that loads with the **dplyr** library (part of the tidyverse package).

To learn more about this dataset run:

```
help(starwars)
```

```
starting httpd help server ... done
```

```
starwars
```

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

2. Get Summary Data about a Dataframe

When working with an unfamiliar dataset, it helps to first extract some basic summary data from that dataset. Run the following commands and examine what each did.

```
head(starwars)
```

```
# A tibble: 6 × 14
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Luke Sky…    172    77 blond      fair       blue            19   male  mascu…
2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu…
3 R2-D2         96    32 <NA>       white, bl… red             33   none  mascu…
4 Darth Va…    202   136 none       white      yellow          41.9 male  mascu…
5 Leia Org…    150    49 brown      light      brown           19   fema… femin…
6 Owen Lars    178   120 brown, gr… light      blue            52   male  mascu…
# ℹ 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

```
#also try:
head(starwars, 3)
```

```
# A tibble: 3 × 14
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Luke Sky…    172    77 blond      fair       blue            19   male  mascu…
2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu…
3 R2-D2         96    32 <NA>       white, bl… red             33   none  mascu…
# ℹ 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

```
dim(starwars)
```

```
[1] 87 14
```

```
colnames(starwars)
```

```
 [1] "name"       "height"     "mass"       "hair_color" "skin_color"
 [6] "eye_color"  "birth_year" "sex"        "gender"     "homeworld"
[11] "species"    "films"      "vehicles"   "starships"
```

```
table(starwars$species)
```

```
       Aleena       Besalisk         Cerean       Chagrian       Clawdite
           1              1              1              1              1
       Droid            Dug           Ewok      Geonosian         Gungan
           6              1              1              1              3
       Human           Hutt       Iktotchi        Kaleesh       Kaminoan
          35              1              1              1              2
     Kel Dor       Mirialan   Mon Calamari           Muun       Nautolan
           1              2              1              1              1
    Neimodian         Pau'an       Quermian         Rodian        Skakoan
           1              1              1              1              1
    Sullustan     Tholothian        Togruta          Toong      Toydarian
           1              1              1              1              1
   Trandoshan        Twi'lek     Vulptereen        Wookiee          Xexto
           1              2              1              2              1
Yoda's species         Zabrak
           1              2
```

```
xtabs(data=starwars, ~species + homeworld) |>
    as_tibble() |>
    filter(n > 0)
```

```
# A tibble: 51 × 3
   species   homeworld          n
   <chr>     <chr>          <int>
 1 Human     Alderaan           3
 2 Aleena    Aleen Minor        1
 3 Human     Bespin             1
 4 Neimodian Cato Neimoidia     1
 5 Cerean    Cerea              1
 6 Chagrian  Champala           1
 7 Human     Chandrila          1
 8 Human     Concord Dawn       1
 9 Human     Corellia           2
10 Human     Coruscant          2
# ℹ 41 more rows
```

```
#summary(starwars)
```

```
#str(starwars)
```

3. Retrieve Data

Get one row

```
starwars[1,]
```

```
# A tibble: 1 × 14
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Luke Sky…    172    77 blond      fair       blue              19 male  mascu…
# ℹ 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Get one cell

```
starwars$name[4]
```

```
[1] "Darth Vader"
```

Get item in list

```
print(starwars$starships[1])
```

```
[[1]]
[1] "X-wing"           "Imperial shuttle"
```

```
starwars$starships[[1]][1]
```

```
[1] "X-wing"
```

Exercise

Choose another built-in dataset visible when you run `data()`. Use the commands above to explore the size, shape, type, and nature of the data stored in that dataset.