Code ▾

Week-5 Tutorial

Bibek Sapkota

Data Wrangling in R with dplyr tutorial

task 1: Install and Load the packages

Hide

```
packages_to_install <- c("tidyverse")

for (package_name in packages_to_install) {
  if (!requireNamespace(package_name, quietly = TRUE)) {
    install.packages(package_name)
  }
}

library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────────────────────────── t
idyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.0     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2     ── Conflicts ──────────────────────────────────────────────
────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the ⍰]8;;http://conflicted.r-lib.org/⍰conflicted package⍰]8;;⍰ to force all conflicts t
o become errors
```

task 2:Displaying the dataset

Hide

```
starwars
```

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 14 columns Previous 1 2 3 4 5 6 ... 9 Next

dplyr syntax

task 1: Filter the data which have species Human.

Hide

```
filter(starwars, species=="Human")
```

name <chr>	height <int>	ma… <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr
Luke Skywalker	172	77.0	blond	fair	blue	19.0	male
Darth Vader	202	136.0	none	white	yellow	41.9	male
Leia Organa	150	49.0	brown	light	brown	19.0	fema
Owen Lars	178	120.0	brown, grey	light	blue	52.0	male
Beru Whitesun Lars	165	75.0	brown	light	blue	47.0	fema
Biggs Darklighter	183	84.0	black	light	brown	24.0	male
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male
Anakin Skywalker	188	84.0	blond	fair	blue	41.9	male
Wilhuff Tarkin	180	NA	auburn, grey	fair	blue	64.0	male
Han Solo	180	80.0	brown	fair	brown	29.0	male

1-10 of 35 rows | 1-9 of 14 columns Previous 1 2 3 4 Next

Hide

```
starwars[starwars$species=="Human"&!is.na(starwars$species),]
```

name <chr>	height <int>	ma… <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr
Luke Skywalker	172	77.0	blond	fair	blue	19.0	male
Darth Vader	202	136.0	none	white	yellow	41.9	male
Leia Organa	150	49.0	brown	light	brown	19.0	fema
Owen Lars	178	120.0	brown, grey	light	blue	52.0	male

name	height	ma…	hair_color	skin_color	eye_color	birth_year	sex
<chr>	<int>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<chr
Beru Whitesun Lars	165	75.0	brown	light	blue	47.0	fema
Biggs Darklighter	183	84.0	black	light	brown	24.0	male
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male
Anakin Skywalker	188	84.0	blond	fair	blue	41.9	male
Wilhuff Tarkin	180	NA	auburn, grey	fair	blue	64.0	male
Han Solo	180	80.0	brown	fair	brown	29.0	male

1-10 of 35 rows | 1-9 of 14 columns Previous 1 2 3 4 Next

The pipe %>% operator

task 1:filtering the data by species Droid and then arrange it by its height

Hide

```
filter(starwars,species=="Droid")%>%
  arrange(height)
```

name	height	m…	hair_color	skin_color	eye_color	birth_year	sex	gender	homewo
<chr>	<int>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>
R2-D2	96	32	NA	white, blue	red	33	none	masculine	Naboo
R4-P17	96	NA	none	silver, red	red, blue	NA	none	feminine	NA
R5-D4	97	32	NA	white, red	red	NA	none	masculine	Tatooine
C-3PO	167	75	NA	gold	yellow	112	none	masculine	Tatooine
IG-88	200	140	none	metal	red	15	none	masculine	NA
BB8	NA	NA	none	none	black	NA	none	masculine	NA

6 rows | 1-10 of 14 columns

task 2:Filtering the data by species and displaying it

Hide

```
#Both are same but different ways of writing
filter(starwars, species=="Human")
```

name	height	ma…	hair_color	skin_color	eye_color	birth_year	sex
<chr>	<int>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<chr
Luke Skywalker	172	77.0	blond	fair	blue	19.0	male
Darth Vader	202	136.0	none	white	yellow	41.9	male
Leia Organa	150	49.0	brown	light	brown	19.0	fema

name <chr>	height <int>	ma... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr
Owen Lars	178	120.0	brown, grey	light	blue	52.0	male
Beru Whitesun Lars	165	75.0	brown	light	blue	47.0	fema
Biggs Darklighter	183	84.0	black	light	brown	24.0	male
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male
Anakin Skywalker	188	84.0	blond	fair	blue	41.9	male
Wilhuff Tarkin	180	*NA*	auburn, grey	fair	blue	64.0	male
Han Solo	180	80.0	brown	fair	brown	29.0	male

1-10 of 35 rows | 1-9 of 14 columns Previous 1 2 3 4 Next

Hide

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

name <chr>	height <int>	ma... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr
Luke Skywalker	172	77.0	blond	fair	blue	19.0	male
Darth Vader	202	136.0	none	white	yellow	41.9	male
Leia Organa	150	49.0	brown	light	brown	19.0	fema
Owen Lars	178	120.0	brown, grey	light	blue	52.0	male
Beru Whitesun Lars	165	75.0	brown	light	blue	47.0	fema
Biggs Darklighter	183	84.0	black	light	brown	24.0	male
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male
Anakin Skywalker	188	84.0	blond	fair	blue	41.9	male
Wilhuff Tarkin	180	*NA*	auburn, grey	fair	blue	64.0	male
Han Solo	180	80.0	brown	fair	brown	29.0	male

1-10 of 35 rows | 1-9 of 14 columns Previous **1** 2 3 4 Next

Hide

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

name <chr>	height <int>	ma... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr
Luke Skywalker	172	77.0	blond	fair	blue	19.0	male
Darth Vader	202	136.0	none	white	yellow	41.9	male
Leia Organa	150	49.0	brown	light	brown	19.0	fema

name <chr>	height <int>	ma... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr
Owen Lars	178	120.0	brown, grey	light	blue	52.0	male
Beru Whitesun Lars	165	75.0	brown	light	blue	47.0	fema
Biggs Darklighter	183	84.0	black	light	brown	24.0	male
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	57.0	male
Anakin Skywalker	188	84.0	blond	fair	blue	41.9	male
Wilhuff Tarkin	180	NA	auburn, grey	fair	blue	64.0	male
Han Solo	180	80.0	brown	fair	brown	29.0	male

1-10 of 25 rows | 1-9 of 14 columns   Previous 1 2 3 4 Next

Selecting columns with select()

task 1:Using dplyr’s glimpse() function to get a quick look at our data

Hide

```
glimpse(starwars)
```

```
Rows: 87
Columns: 14
$ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Organa", "Owen La
rs", "Beru Whit…
$ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 228, 180, 173, 1
75, 170, 180, 6…
$ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.0, 84.0, NA, 11
2.0, 80.0, 74.0…
$ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", NA, "black", "au
burn, white", "…
$ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "light", "white,
red", "light",…
$ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue", "red", "brow
n", "blue-gray", …
$ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, 41.9, 64.0, 20
0.0, 29.0, 44.0,…
$ sex        <chr> "male", "none", "none", "male", "female", "male", "female", "none", "mal
e", "male", "male…
$ gender     <chr> "masculine", "masculine", "masculine", "masculine", "feminine", "masculin
e", "feminine", …
$ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "Tatooine", "Tato
oine", "Tatooin…
$ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Human", "Droid", "H
uman", "Human",…
$ films      <list> <"A New Hope", "The Empire Strikes Back", "Return of the Jedi", "Revenge
of the Sith", "…
$ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imperial Speeder B
ike", <>, <>, <…
$ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1", <>, <>, <>, <
>, "X-wing", <"J…
```

task 2:Displaying only selected columns

Hide

```
starwars %>%
    select(name,    height, mass,   hair_color, skin_color,
               eye_color,   birth_year, sex, gender,    homeworld,  species)
```

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 ... 9 Next

task 3: Removing the selected columns

Hide

```
starwars %>% select(-films, -vehicles, -starships)
```

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 11 columns Previous 1 2 3 4 5 6 ... 9 Next

task 4:Displaying the table from name to species

Hide

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

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 ... 9 Next

Rename your columns with rename()

task 1:Renaming name columns to username

Hide

```
starwars%>%
  rename(username=name)
```

username <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red

username <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 14 columns Previous 1 2 3 4 5 6 ... 9 Next

task 2:Displaying the data and then renaming the height column

Hide

```
starwars%>%
  select(name:species)%>%
  rename(height_cm=height)
```

name <chr>	height_cm <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 11 columns Previous 1 2 3 4 5 6 ... 9 Next

Sort a data with arrange()

task 1:Loading the data

Hide

```
starwars
```

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red

name <chr>	height <int>	mass <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 14 columns Previous **1** 2 3 4 5 6 ... 9 Next

task 2:Diaplaying the data from name to mass and arrange it by height

Hide

```
starwars %>%
  select(name:mass) %>%
  arrange(height)
```

name <chr>	height <int>	mass <dbl>
Yoda	66	17.0
Ratts Tyerel	79	15.0
Wicket Systri Warrick	88	20.0
Dud Bolt	94	45.0
R2-D2	96	32.0
R4-P17	96	NA
R5-D4	97	32.0
Sebulba	112	40.0
Gasgano	122	NA
Watto	137	NA

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 3:Arranging the height columns in descending order

Hide

```
starwars%>%
  select(name:mass)%>%
  arrange(desc(height))
```

name <chr>	height <int>	mass <dbl>
Yarael Poof	264	*NA*
Tarfful	234	136.0
Lama Su	229	88.0
Chewbacca	228	112.0
Roos Tarpals	224	82.0
Grievous	216	159.0
Taun We	213	*NA*
Rugor Nass	206	*NA*
Tion Medon	206	80.0
Darth Vader	202	136.0

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 4:Arranging the columns first by homeworld columns and then by height

Hide

```
starwars %>%
  select(name, homeworld, height) %>%
  arrange(homeworld, desc(height))
```

name <chr>	homeworld <chr>	height <int>
Bail Prestor Organa	Alderaan	191
Raymus Antilles	Alderaan	188
Leia Organa	Alderaan	150
Ratts Tyerel	Aleen Minor	79
Lobot	Bespin	175
Jek Tono Porkins	Bestine IV	180
Nute Gunray	Cato Neimoidia	191
Ki-Adi-Mundi	Cerea	198
Mas Amedda	Champala	196
Mon Mothma	Chandrila	150

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

Select rows based on their position using slice()

task 1:Displaying first 3 rows of the data from name to mass

Hide

```
starwars%>%
  slice(1:3)%>%
  select(name:mass)
```

name	height	mass
<chr>	<int>	<dbl>
Luke Skywalker	172	77
C-3PO	167	75
R2-D2	96	32

3 rows

task 2: Arraning the data by height then displaying 1st 3 rows from name to mass

Hide

```
starwars%>%
  arrange(height)%>%
  slice(1:3)%>%
  select(name:mass)
```

name	height	mass
<chr>	<int>	<dbl>
Yoda	66	17
Ratts Tyerel	79	15
Wicket Systri Warrick	88	20

3 rows

Subset your data with filter()

task 1:using filter() to filter species and homeworld

Hide

```
starwars %>%
  select(name, species, homeworld, height, mass, birth_year) %>%
  filter(species=="Human" & homeworld=="Tatooine")
```

name	species	homeworld	height	mass	birth_year
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>
Luke Skywalker	Human	Tatooine	172	77	19.0
Darth Vader	Human	Tatooine	202	136	41.9
Owen Lars	Human	Tatooine	178	120	52.0
Beru Whitesun Lars	Human	Tatooine	165	75	47.0

name	species	homeworld	height	mass	birth_year
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>
Biggs Darklighter	Human	Tatooine	183	84	24.0
Anakin Skywalker	Human	Tatooine	188	84	41.9
Shmi Skywalker	Human	Tatooine	163	*NA*	72.0
Cliegg Lars	Human	Tatooine	183	*NA*	82.0

8 rows

task 2:Using comma(,) in & operator

Hide

```
starwars %>%
  select(name, species, homeworld, height, mass, birth_year) %>%
  filter(species=="Human", homeworld=="Tatooine")
```

name	species	homeworld	height	mass	birth_year
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>
Luke Skywalker	Human	Tatooine	172	77	19.0
Darth Vader	Human	Tatooine	202	136	41.9
Owen Lars	Human	Tatooine	178	120	52.0
Beru Whitesun Lars	Human	Tatooine	165	75	47.0
Biggs Darklighter	Human	Tatooine	183	84	24.0
Anakin Skywalker	Human	Tatooine	188	84	41.9
Shmi Skywalker	Human	Tatooine	163	*NA*	72.0
Cliegg Lars	Human	Tatooine	183	*NA*	82.0

8 rows

task 3:Using '|' if we wanted characters that are either Humans or Droids to represent 'or'

Hide

```
starwars %>%
  select(name, species, homeworld, height, mass, birth_year) %>%
  filter(species=="Human" | species=="Droid")
```

name	species	homeworld	height	mass	birth_year
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>
Luke Skywalker	Human	Tatooine	172	77.0	19.0
C-3PO	Droid	Tatooine	167	75.0	112.0
R2-D2	Droid	Naboo	96	32.0	33.0
Darth Vader	Human	Tatooine	202	136.0	41.9

name <chr>	species <chr>	homeworld <chr>	height <int>	mass <dbl>	birth_year <dbl>
Leia Organa	Human	Alderaan	150	49.0	19.0
Owen Lars	Human	Tatooine	178	120.0	52.0
Beru Whitesun Lars	Human	Tatooine	165	75.0	47.0
R5-D4	Droid	Tatooine	97	32.0	*NA*
Biggs Darklighter	Human	Tatooine	183	84.0	24.0
Obi-Wan Kenobi	Human	Stewjon	182	77.0	57.0

1-10 of 41 rows Previous **1** 2 3 4 5 Next

task 4: rewriting above code using %in% in place of |

Hide

```
starwars %>%
  select(name, species, homeworld, height, mass, birth_year) %>%
  filter(species %in% c("Human", "Droid"))
```

name <chr>	species <chr>	homeworld <chr>	height <int>	mass <dbl>	birth_year <dbl>
Luke Skywalker	Human	Tatooine	172	77.0	19.0
C-3PO	Droid	Tatooine	167	75.0	112.0
R2-D2	Droid	Naboo	96	32.0	33.0
Darth Vader	Human	Tatooine	202	136.0	41.9
Leia Organa	Human	Alderaan	150	49.0	19.0
Owen Lars	Human	Tatooine	178	120.0	52.0
Beru Whitesun Lars	Human	Tatooine	165	75.0	47.0
R5-D4	Droid	Tatooine	97	32.0	*NA*
Biggs Darklighter	Human	Tatooine	183	84.0	24.0
Obi-Wan Kenobi	Human	Stewjon	182	77.0	57.0

1-10 of 41 rows Previous **1** 2 3 4 5 Next

task 5:Printing birthyear of 15,19 and 21 using %in% operator

Hide

```
starwars %>%
  select(name, species, homeworld, height,  mass, birth_year) %>%
  filter(birth_year %in% c(15, 19, 21))
```

name <chr>	species <chr>	homeworld <chr>	height <int>	mass <dbl>	birth_year <dbl>
Luke Skywalker	Human	Tatooine	172	77	19
Leia Organa	Human	Alderaan	150	49	19
Wedge Antilles	Human	Corellia	170	77	21
IG-88	Droid	*NA*	200	140	15

4 rows

Making robust filters with pull()

task 1:Filtering using name

Hide

```
starwars %>%
  filter(name %in% c("C-3PO", "R2-D2", "R5-D4"))
```

na… <chr>	height <int>	m… <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr>	gender <chr>	homewo… <chr>
C-3PO	167	75	*NA*	gold	yellow	112	none	masculine	Tatooine
R2-D2	96	32	*NA*	white, blue	red	33	none	masculine	Naboo
R5-D4	97	32	*NA*	white, red	red	*NA*	none	masculine	Tatooine

3 rows | 1-10 of 14 columns

Hide

```
starwars %>%
  filter(name %in% c("C-3PO", "R2-D2", "R5-D4", "IG-88"))
```

na… <chr>	height <int>	m… <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr>	gender <chr>	homewo… <chr>
C-3PO	167	75	*NA*	gold	yellow	112	none	masculine	Tatooine
R2-D2	96	32	*NA*	white, blue	red	33	none	masculine	Naboo
R5-D4	97	32	*NA*	white, red	red	*NA*	none	masculine	Tatooine
IG-88	200	140	none	metal	red	15	none	masculine	*NA*

4 rows | 1-10 of 14 columns

task 2:Filtering species

Hide

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

name <chr>	height <int>	m… <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr>	gender <chr>	homewo <chr>
C-3PO	167	75	NA	gold	yellow	112	none	masculine	Tatooine
R2-D2	96	32	NA	white, blue	red	33	none	masculine	Naboo
R5-D4	97	32	NA	white, red	red	NA	none	masculine	Tatooine
IG-88	200	140	none	metal	red	15	none	masculine	NA
R4-P17	96	NA	none	silver, red	red, blue	NA	none	feminine	NA
BB8	NA	NA	none	none	black	NA	none	masculine	NA

6 rows | 1-10 of 14 columns

task 3:Using pull() by passing it a data set and a column=name and it ‘pulls’ all the values into a vector

Hide

```
starwars %>%
  pull(name)
```

```
 [1] "Luke Skywalker"        "C-3PO"                 "R2-D2"                 "Darth Vader"
 [5] "Leia Organa"           "Owen Lars"             "Beru Whitesun Lars"    "R5-D4"
 [9] "Biggs Darklighter"     "Obi-Wan Kenobi"        "Anakin Skywalker"      "Wilhuff Tarkin"
[13] "Chewbacca"             "Han Solo"              "Greedo"                "Jabba Desilijic
Tiure"
[17] "Wedge Antilles"        "Jek Tono Porkins"      "Yoda"                  "Palpatine"
[21] "Boba Fett"             "IG-88"                 "Bossk"                 "Lando Calrissia
n"
[25] "Lobot"                 "Ackbar"                "Mon Mothma"            "Arvel Crynyd"
[29] "Wicket Systri Warrick" "Nien Nunb"             "Qui-Gon Jinn"          "Nute Gunray"
[33] "Finis Valorum"         "Padmé Amidala"         "Jar Jar Binks"         "Roos Tarpals"
[37] "Rugor Nass"            "Ric Olié"              "Watto"                 "Sebulba"
[41] "Quarsh Panaka"         "Shmi Skywalker"        "Darth Maul"            "Bib Fortuna"
[45] "Ayla Secura"           "Ratts Tyerel"          "Dud Bolt"              "Gasgano"
[49] "Ben Quadinaros"        "Mace Windu"            "Ki-Adi-Mundi"          "Kit Fisto"
[53] "Eeth Koth"             "Adi Gallia"            "Saesee Tiin"           "Yarael Poof"
[57] "Plo Koon"              "Mas Amedda"            "Gregar Typho"          "Cordé"
[61] "Cliegg Lars"           "Poggle the Lesser"     "Luminara Unduli"       "Barriss Offee"
[65] "Dormé"                 "Dooku"                 "Bail Prestor Organa"   "Jango Fett"
[69] "Zam Wesell"            "Dexter Jettster"       "Lama Su"               "Taun We"
[73] "Jocasta Nu"            "R4-P17"                "Wat Tambor"            "San Hill"
[77] "Shaak Ti"              "Grievous"              "Tarfful"               "Raymus Antille
s"
[81] "Sly Moore"             "Tion Medon"            "Finn"                  "Rey"
[85] "Poe Dameron"           "BB8"                   "Captain Phasma"
```

task 4:pulling(data set, column) so it knows where to pull the names from

Hide

```
trending <- sample_n(starwars, 10) %>%
                    select(name)

starwars %>%
  select(name:species) %>%
  filter(name %in% pull(trending, name))
```

name <chr>	height <int>	m... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_
Owen Lars	178	120	brown, grey	light	blue	
Jabba Desilijic Tiure	175	1358	NA	green-tan, brown	orange	
IG-88	200	140	none	metal	red	
Watto	137	NA	black	blue, grey	yellow	
Sebulba	112	40	none	grey, red	orange	
Jango Fett	183	79	black	tan	brown	
Zam Wesell	168	55	blonde	fair, green, yellow	yellow	
Dexter Jettster	198	102	none	brown	yellow	
Lama Su	229	88	none	grey	black	

name <chr>	height <int>	m... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth
Captain Phasma	NA	NA	none	none	unknown	

1-10 of 10 rows | 1-7 of 11 columns

task 5:If we run the above code again we'll get a different 10 random characters

Hide

```
trending <- sample_n(starwars, 10) %>%
                    select(name)

starwars %>%
  select(name:species) %>%
  filter(name %in% pull(trending, name))
```

name <chr>	height <int>	m... <dbl>	hair_color <chr>	skin_color <chr>	eye_color <chr>	birth_year <dbl>	sex <chr>	gen <chr
Biggs Darklighter	183	84	black	light	brown	24	male	mas
Jek Tono Porkins	180	110	brown	fair	blue	NA	NA	NA
Yoda	66	17	white	green	brown	896	male	mas
Bossk	190	113	none	green	red	53	male	mas
Ackbar	180	83	none	brown mottle	orange	41	male	mas
Saesee Tiin	188	NA	none	pale	orange	NA	male	mas
Gregar Typho	185	85	black	dark	brown	NA	NA	NA
Cordé	157	NA	brown	light	brown	NA	NA	NA
Sly Moore	178	48	none	pale	white	NA	NA	NA
BB8	NA	NA	none	none	black	NA	none	mas

1-10 of 10 rows | 1-9 of 11 columns

Create new columns with mutate()

task 1:Creating new column usingmutate(new_column = something)

Hide

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

name <chr>	height <int>	mass <dbl>	BMI <dbl>
Luke Skywalker	172	77.0	26.02758

name	height	mass	BMI
<chr>	<int>	<dbl>	<dbl>
C-3PO	167	75.0	26.89232
R2-D2	96	32.0	34.72222
Darth Vader	202	136.0	33.33007
Leia Organa	150	49.0	21.77778
Owen Lars	178	120.0	37.87401
Beru Whitesun Lars	165	75.0	27.54821
R5-D4	97	32.0	34.00999
Biggs Darklighter	183	84.0	25.08286
Obi-Wan Kenobi	182	77.0	23.24598

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 2:Converting height from cm into feet and weight from kg into pounds

Hide

```
starwars %>%
  select(name:mass) %>%
  mutate(height_ft = height * 0.0328084,
            weight_pounds = mass*2.20462)
```

name	height	mass	height_ft	weight_pounds
<chr>	<int>	<dbl>	<dbl>	<dbl>
Luke Skywalker	172	77.0	5.643045	169.75574
C-3PO	167	75.0	5.479003	165.34650
R2-D2	96	32.0	3.149606	70.54784
Darth Vader	202	136.0	6.627297	299.82832
Leia Organa	150	49.0	4.921260	108.02638
Owen Lars	178	120.0	5.839895	264.55440
Beru Whitesun Lars	165	75.0	5.413386	165.34650
R5-D4	97	32.0	3.182415	70.54784
Biggs Darklighter	183	84.0	6.003937	185.18808
Obi-Wan Kenobi	182	77.0	5.971129	169.75574

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 3:Using a simple yes/no test i.e. taller than 6ft or not we can do this with ifelse():

Hide

```
starwars %>%
  select(name:mass ) %>%
  mutate(height_ft = height * 0.0328084,
         over_6ft = ifelse(height_ft>6, 1, 0))
```

name <chr>	height <int>	mass <dbl>	height_ft <dbl>	over_6ft <dbl>
Luke Skywalker	172	77.0	5.643045	0
C-3PO	167	75.0	5.479003	0
R2-D2	96	32.0	3.149606	0
Darth Vader	202	136.0	6.627297	1
Leia Organa	150	49.0	4.921260	0
Owen Lars	178	120.0	5.839895	0
Beru Whitesun Lars	165	75.0	5.413386	0
R5-D4	97	32.0	3.182415	0
Biggs Darklighter	183	84.0	6.003937	1
Obi-Wan Kenobi	182	77.0	5.971129	0

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 4:Using multiple logical conditions in our mutate

Hide

```
starwars %>%
  select(name:mass ) %>%
  mutate(height_ft = height * 0.0328084,
               height_group = case_when(    is.na(height) ~ "Missing",
                                                height_ft<5   ~ "Under 5ft",
                                                height_ft>6   ~ "Over  6ft",
                                                TRUE  ~ "Between 5-6ft"))
```

name <chr>	height <int>	mass <dbl>	height_ft <dbl>	height_group <chr>
Luke Skywalker	172	77.0	5.643045	Between 5-6ft
C-3PO	167	75.0	5.479003	Between 5-6ft
R2-D2	96	32.0	3.149606	Under 5ft
Darth Vader	202	136.0	6.627297	Over 6ft
Leia Organa	150	49.0	4.921260	Under 5ft
Owen Lars	178	120.0	5.839895	Between 5-6ft
Beru Whitesun Lars	165	75.0	5.413386	Between 5-6ft
R5-D4	97	32.0	3.182415	Under 5ft

name <chr>	height <int>	mass <dbl>	height_ft <dbl>	height_group <chr>
Biggs Darklighter	183	84.0	6.003937	Over 6ft
Obi-Wan Kenobi	182	77.0	5.971129	Between 5-6ft

1-10 of 87 rows Previous 1 2 3 4 5 6 ... 9 Next

task 5:Displaying result what will come if we don’t use TRUE at last

Hide

```
starwars %>%
  select(name:mass ) %>%
  mutate(height_ft = height * 0.0328084,
              height_group = case_when(is.na(height) ~ "Missing",
                                               height_ft<5   ~ "Under 5ft",
                                               height_ft>6   ~ "Over  6ft"))
```

name <chr>	height <int>	mass <dbl>	height_ft <dbl>	height_group <chr>
Luke Skywalker	172	77.0	5.643045	NA
C-3PO	167	75.0	5.479003	NA
R2-D2	96	32.0	3.149606	Under 5ft
Darth Vader	202	136.0	6.627297	Over 6ft
Leia Organa	150	49.0	4.921260	Under 5ft
Owen Lars	178	120.0	5.839895	NA
Beru Whitesun Lars	165	75.0	5.413386	NA
R5-D4	97	32.0	3.182415	Under 5ft
Biggs Darklighter	183	84.0	6.003937	Over 6ft
Obi-Wan Kenobi	182	77.0	5.971129	NA

1-10 of 87 rows Previous 1 2 3 4 5 6 ... 9 Next

Window functions with mutate()

task 1:creating new columns that contain summaries of data from within the table

Hide

```
starwars %>%
  select(name:mass ) %>%
  mutate(avg_height = mean(height, na.rm=T))
```

name <chr>	height <int>	mass <dbl>	avg_height <dbl>
Luke Skywalker	172	77.0	174.6049

name	height	mass	avg_height
<chr>	<int>	<dbl>	<dbl>
C-3PO	167	75.0	174.6049
R2-D2	96	32.0	174.6049
Darth Vader	202	136.0	174.6049
Leia Organa	150	49.0	174.6049
Owen Lars	178	120.0	174.6049
Beru Whitesun Lars	165	75.0	174.6049
R5-D4	97	32.0	174.6049
Biggs Darklighter	183	84.0	174.6049
Obi-Wan Kenobi	182	77.0	174.6049

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 2:Creating a new column that compares the individual height

Hide

```
starwars %>%
  select(name:mass) %>%
  mutate( avg_height = mean(height, na.rm=T),
            height_index = height/avg_height,
             height_group=case_when(height_index<=.8  ~ "short",
                                                          height_index>=1.2 ~ "tall",
                                                    TRUE  ~ "average"))
```

name	height	mass	avg_height	height_index	height_group
<chr>	<int>	<dbl>	<dbl>	<dbl>	<chr>
Luke Skywalker	172	77.0	174.6049	0.9850810	average
C-3PO	167	75.0	174.6049	0.9564449	average
R2-D2	96	32.0	174.6049	0.5498126	short
Darth Vader	202	136.0	174.6049	1.1568974	average
Leia Organa	150	49.0	174.6049	0.8590822	average
Owen Lars	178	120.0	174.6049	1.0194442	average
Beru Whitesun Lars	165	75.0	174.6049	0.9449905	average
R5-D4	97	32.0	174.6049	0.5555398	short
Biggs Darklighter	183	84.0	174.6049	1.0480803	average
Obi-Wan Kenobi	182	77.0	174.6049	1.0423531	average

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 3:Just using the calculation directly as part of our height_index formula

Hide

```
starwars %>%
  select(name:mass) %>%
  mutate(height_group=case_when(height / mean(height, na.rm=T)<=.8  ~ "short",
                                                        height / mean(height, na.rm=T)>=1.2
~ "tall",
                                                        TRUE  ~ "average"))
```

name <chr>	height <int>	mass <dbl>	height_group <chr>
Luke Skywalker	172	77.0	average
C-3PO	167	75.0	average
R2-D2	96	32.0	short
Darth Vader	202	136.0	average
Leia Organa	150	49.0	average
Owen Lars	178	120.0	average
Beru Whitesun Lars	165	75.0	average
R5-D4	97	32.0	short
Biggs Darklighter	183	84.0	average
Obi-Wan Kenobi	182	77.0	average

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

Aggregating data using summarise()

task 1:Using summarise to get average height of the data set:

Hide

```
starwars %>%
  summarise(avg_height=mean(height, na.rm=T))
```

avg_height <dbl>
174.6049

1 row

task 2:Summarizing the data

Hide

```
starwars %>%
  summarise(num_records=n(),                              # Number of records in the table
                 distinct_species=n_distinct(species), # Number of unique values of "species"
                     avg_mass=mean(mass, na.rm=T),             # Average mass excluding any m
issing values
                 median_mass=median(mass, na.rm=T),    # Median mass excluding any missing va
lues
                 IQR_mass=IQR(mass, na.rm=T),     # The interquartile range for mass excludin
g any missing values
                 shortest=min(height, na.rm=T),        # Min value of height excluding any mi
ssing values
                     tallest=max(height, na.rm=T))          # Max value of height excluding a
ny missing values
```

num_records	distinct_species	avg_mass	median_mass	IQR_ma...	shortest	tallest
<int>	<int>	<dbl>	<dbl>	<dbl>	<int>	<int>
87	38	97.31186	79	28.9	66	264

1 row

dplyr + base R = conditional sums

task 1:Displaying species in dataset

Hide

```
starwars$species
```

```
 [1] "Human"           "Droid"           "Droid"           "Human"           "Human"           "Hu
man"
 [7] "Human"           "Droid"           "Human"           "Human"           "Human"           "Hu
man"
[13] "Wookiee"         "Human"           "Rodian"          "Hutt"            "Human"           NA
[19] "Yoda's species" "Human"           "Human"           "Droid"           "Trandoshan"      "Hu
man"
[25] "Human"           "Mon Calamari"    "Human"           "Human"           "Ewok"            "Su
llustan"
[31] "Human"           "Neimodian"       "Human"           "Human"           "Gungan"          "Gu
ngan"
[37] "Gungan"          "Human"           "Toydarian"       "Dug"             "Human"           "Hu
man"
[43] "Zabrak"          "Twi'lek"         "Twi'lek"         "Aleena"          "Vulptereen"      "Xe
xto"
[49] "Toong"           "Human"           "Cerean"          "Nautolan"        "Zabrak"          "Th
olothian"
[55] "Iktotchi"        "Quermian"        "Kel Dor"         "Chagrian"        NA                NA
[61] "Human"           "Geonosian"       "Mirialan"        "Mirialan"        "Human"           "Hu
man"
[67] "Human"           "Human"           "Clawdite"        "Besalisk"        "Kaminoan"        "Ka
minoan"
[73] "Human"           "Droid"           "Skakoan"         "Muun"            "Togruta"         "Ka
leesh"
[79] "Wookiee"         "Human"           NA                "Pau'an"          "Human"           "Hu
man"
[85] "Human"           "Droid"           "Human"
```

task 2:Applying a logical condition to it which turns our vector into a logical one

Hide

```
starwars$species=="Human"
```

```
 [1]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALS
E FALSE  TRUE
[18]    NA FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALS
E  TRUE  TRUE
[35] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALS
E  TRUE FALSE
[52] FALSE FALSE FALSE FALSE FALSE FALSE FALSE    NA    NA  TRUE FALSE FALSE FALSE  TRUE  TRU
E  TRUE  TRUE
[69] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE    NA FALSE  TRU
E  TRUE  TRUE
[86] FALSE  TRUE
```

task 3:Doing same case in another way

Hide

```
starwars$species=="Human" & !is.na(starwars$species)
```

```
 [1]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALS
E FALSE  TRUE
[18] FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALS
E  TRUE  TRUE
[35] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALS
E  TRUE FALSE
[52] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRU
E  TRUE  TRUE
[69] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRU
E  TRUE  TRUE
[86] FALSE  TRUE
```

task 4:Identifing entries labeled as “Human” and not missing, and then extract these “Human” entries for display.

Hide

```
starwars$species
```

```
 [1] "Human"          "Droid"          "Droid"          "Human"          "Human"          "Hu
man"
 [7] "Human"          "Droid"          "Human"          "Human"          "Human"          "Hu
man"
[13] "Wookiee"        "Human"          "Rodian"         "Hutt"           "Human"          NA
[19] "Yoda's species" "Human"          "Human"          "Droid"          "Trandoshan"     "Hu
man"
[25] "Human"          "Mon Calamari"   "Human"          "Human"          "Ewok"           "Su
llustan"
[31] "Human"          "Neimodian"      "Human"          "Human"          "Gungan"         "Gu
ngan"
[37] "Gungan"         "Human"          "Toydarian"      "Dug"            "Human"          "Hu
man"
[43] "Zabrak"         "Twi'lek"        "Twi'lek"        "Aleena"         "Vulptereen"     "Xe
xto"
[49] "Toong"          "Human"          "Cerean"         "Nautolan"       "Zabrak"         "Th
olothian"
[55] "Iktotchi"       "Quermian"       "Kel Dor"        "Chagrian"       NA               NA
[61] "Human"          "Geonosian"      "Mirialan"       "Mirialan"       "Human"          "Hu
man"
[67] "Human"          "Human"          "Clawdite"       "Besalisk"       "Kaminoan"       "Ka
minoan"
[73] "Human"          "Droid"          "Skakoan"        "Muun"           "Togruta"        "Ka
leesh"
[79] "Wookiee"        "Human"          NA               "Pau'an"         "Human"          "Hu
man"
[85] "Human"          "Droid"          "Human"
```

Hide

```
starwars$species=="Human"& !is.na(starwars$species)
```

```
 [1]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALS
E FALSE  TRUE
[18] FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALS
E  TRUE  TRUE
[35] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALS
E  TRUE FALSE
[52] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRU
E  TRUE  TRUE
[69] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRU
E  TRUE  TRUE
[86] FALSE  TRUE
```

Hide

```
starwars$species[starwars$species=="Human" & !is.na(starwars$species)]
```

```
 [1] "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human"
"Human" "Human"
[14] "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human"
"Human" "Human"
[27] "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human" "Human"
```

task 5:Changing our filter to bring back all the names of the human characters rather instead of their species

Hide

```
starwars$name[starwars$species=="Human"& !is.na(starwars$species)]
```

```
 [1] "Luke Skywalker"      "Darth Vader"         "Leia Organa"         "Owen Lars"          
 [5] "Beru Whitesun Lars"  "Biggs Darklighter"   "Obi-Wan Kenobi"      "Anakin Skywalker"   
 [9] "Wilhuff Tarkin"      "Han Solo"            "Wedge Antilles"      "Palpatine"          
[13] "Boba Fett"           "Lando Calrissian"    "Lobot"               "Mon Mothma"         
[17] "Arvel Crynyd"        "Qui-Gon Jinn"        "Finis Valorum"       "Padmé Amidala"      
[21] "Ric Olié"            "Quarsh Panaka"       "Shmi Skywalker"      "Mace Windu"         
[25] "Cliegg Lars"         "Dormé"               "Dooku"               "Bail Prestor Organa"
[29] "Jango Fett"          "Jocasta Nu"          "Raymus Antilles"     "Finn"               
[33] "Rey"                 "Poe Dameron"         "Captain Phasma"     
```

task 6: Using the length() function from base R which counts how many elements are in our vector.

Hide

```
length(starwars$name[starwars$species=="Human" & !is.na(starwars$species)])
```

```
[1] 35
```

task 7:Using it to filter for character’s heights and then average these as normal.

Hide

```
mean(starwars$height[starwars$species=="Human" & !is.na(starwars$species)],na.rm=T)
```

```
[1] 178
```

task 8:Using pipe %>% to the subsequent functions.

Hide

```
starwars %>%
  summarise(number_humans= length(name[species=="Human" & !is.na(species)]),
            number_droids= length(name[species=="Droid" & !is.na(species)]),
            avg_height_humans= mean(height[species=="Human"& !is.na(species)],na.rm = T),
            avg_height_droid= mean(height[species=="Droid" & !is.na(species)],na.rm = T))
```

number_humans	number_droids	avg_height_humans	avg_height_droid
<int>	<int>	<dbl>	<dbl>
35	6	178	131.2

1 row

Aggregate by groups with group_by()

task 1:Using group by to group.

Hide

```
starwars%>%
  group_by(species,gender)
```

name	height	mass	hair_color	skin_color	eye_color
<chr>	<int>	<dbl>	<chr>	<chr>	<chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 14 columns Previous 1 2 3 4 5 6 ... 9 Next

task 2: Using ungroup to remove this group.

Hide

```
starwars%>%
  group_by(species,gender) %>%
  ungroup()
```

name	height	mass	hair_color	skin_color	eye_color
<chr>	<int>	<dbl>	<chr>	<chr>	<chr>
Luke Skywalker	172	77.0	blond	fair	blue
C-3PO	167	75.0	NA	gold	yellow
R2-D2	96	32.0	NA	white, blue	red
Darth Vader	202	136.0	none	white	yellow
Leia Organa	150	49.0	brown	light	brown
Owen Lars	178	120.0	brown, grey	light	blue
Beru Whitesun Lars	165	75.0	brown	light	blue
R5-D4	97	32.0	NA	white, red	red
Biggs Darklighter	183	84.0	black	light	brown
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray

1-10 of 87 rows | 1-7 of 14 columns Previous **1** 2 3 4 5 6 ... 9 Next

task 3:Checking if a table has a group_by() already applied to it we can use the group_vars() function.

Hide

```
starwars%>%
  group_by(species,gender)%>%
  group_vars()
```

```
[1] "species" "gender"
```

task 4:calculating the average height for each of the different species in our data set.

Hide

```
starwars%>%
  group_by(species)%>%
  summarise(avg_height=mean(height, na.rm = T))
```

species	avg_height
<chr>	<dbl>
Aleena	79.0000
Besalisk	198.0000
Cerean	198.0000
Chagrian	196.0000
Clawdite	168.0000
Droid	131.2000
Dug	112.0000

species <chr>	avg_height <dbl>
Ewok	88.0000
Geonosian	183.0000
Gungan	208.6667

1-10 of 38 rows Previous **1** 2 3 4 Next

task 5:If we want to use multiple groups at the same time we can just specify them all with a comma in between.

Hide

```
starwars%>%
  group_by(species,gender)%>%
    summarise(avg_height=mean(height,na.rm = T))%>%
  filter(n()>1)
```

```
`summarise()` has grouped output by 'species'. You can override using the `.groups` argument.
```

species <chr>	gender <chr>	avg_height <dbl>
Droid	feminine	96.0000
Droid	masculine	140.0000
Human	feminine	163.5714
Human	masculine	182.3913
Kaminoan	feminine	213.0000
Kaminoan	masculine	229.0000
Twi'lek	feminine	178.0000
Twi'lek	masculine	180.0000

8 rows

task 6:It selects name, height, and species from the starwars dataset, groups by species, and adds columns for the average species height and height categorization relative to this average.

Hide

```
starwars%>%
  select(name, height, species) %>%
  group_by(species)%>%
  mutate(avg_species_height=mean(height, na.rm = T),
         height_group=case_when(height/avg_species_height<=.8~"short",
                                height/avg_species_height>=1.2~"tall",
                                TRUE ~ "average"))
```

name <chr>	height <int>	species <chr>	avg_species_height <dbl>	height_group <chr>
Luke Skywalker	172	Human	178.0000	average
C-3PO	167	Droid	131.2000	tall
R2-D2	96	Droid	131.2000	short
Darth Vader	202	Human	178.0000	average
Leia Organa	150	Human	178.0000	average
Owen Lars	178	Human	178.0000	average
Beru Whitesun Lars	165	Human	178.0000	average
R5-D4	97	Droid	131.2000	short
Biggs Darklighter	183	Human	178.0000	average
Obi-Wan Kenobi	182	Human	178.0000	average

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next

task 7:calculateing average height per species and per homeworld, and adds these as new columns.

Hide

```
starwars%>%
  select(name, height, species, homeworld)%>%
  group_by(species)%>%
  mutate(avg_height_species=mean(height,na.rm = T))%>%
  group_by(homeworld)%>%
  mutate(avg_height_homeworld=mean(height, na.rm = T))
```

name <chr>	height <int>	species <chr>	homeworld <chr>	avg_height_species <dbl>	a
Luke Skywalker	172	Human	Tatooine	178.0000	
C-3PO	167	Droid	Tatooine	131.2000	
R2-D2	96	Droid	Naboo	131.2000	
Darth Vader	202	Human	Tatooine	178.0000	
Leia Organa	150	Human	Alderaan	178.0000	
Owen Lars	178	Human	Tatooine	178.0000	
Beru Whitesun Lars	165	Human	Tatooine	178.0000	
R5-D4	97	Droid	Tatooine	131.2000	
Biggs Darklighter	183	Human	Tatooine	178.0000	
Obi-Wan Kenobi	182	Human	Stewjon	178.0000	

1-10 of 87 rows Previous **1** 2 3 4 5 6 ... 9 Next task 8:It sorts height by descending order, selects the top 3 tallest individuals from each homeworld, and filters for homeworlds with at least 3 entries.

Hide

```
starwars%>%
  select(name, height, homeworld)%>%
  arrange(desc(height))%>%
  group_by(homeworld)%>%
  slice(1:3)%>%
  filter(n()>=3)
```

name <chr>	height <int>	homeworld <chr>
Bail Prestor Organa	191	Alderaan
Raymus Antilles	188	Alderaan
Leia Organa	150	Alderaan
Adi Gallia	184	Coruscant
Finis Valorum	170	Coruscant
Jocasta Nu	167	Coruscant
Lama Su	229	Kamino
Taun We	213	Kamino
Boba Fett	183	Kamino
Roos Tarpals	224	Naboo

1-10 of 18 rows Previous **1** 2 Next

Apply functions across multiple columns using across()

task 1:calculates and summarizes the number of missing values for the species, name, and homeworld columns in the starwars dataset.

Hide

```
starwars%>%
  summarise(species = sum(is.na(species)),
            name = sum(is.na(name)),
            homeworld = sum(is.na(homeworld)))
```

species <int>	name <int>	homeworld <int>
4	0	10

1 row

task 2: Across using a vector of column names

Hide

```
starwars %>%
  summarise(across(c(species, name, homeworld), ~sum(is.na(.x))))
```

species	name	homeworld
<int>	<int>	<int>
4	0	10

1 row

task 3: The across function in R applies a specified function to multiple columns in a data frame.

Hide

```
verb(across(columns_to_go_across, ~ function_to_apply(.x)))
```

```
Error in verb(across(columns_to_go_across, ~function_to_apply(.x))) :
  could not find function "verb"
```

task 4:Using Across everything()

Hide

```
starwars %>%
  summarise(across(everything(), ~sum((is.na(.x)))))
```

n…	height	m…	hair_color	skin_color	eye_color	birth_year	s..	gen…	homewo…	▶
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	
0	6	28	5	0	0	44	4	4	10	

1 row | 1-10 of 14 columns

task 5:Using the ‘starts_with()’ helper to pick all the columns starting with ‘s’:

Hide

```
starwars %>%
  summarise(across(starts_with('s'), ~sum((is.na(.x)))))
```

skin_color	sex	species	starships
<int>	<int>	<int>	<int>
0	4	4	0

1 row

task 6:Using and combine different selection helpers to pick which columns we want to call our function on

Hide

```
starwars %>%
  summarise(across(!c(species, name, homeworld), ~sum((is.na(.x)))))
```

height	m…	hair_color	skin_color	eye_color	birth_year	s..	gen…	films	vehicles ▸
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
6	28	5	0	0	44	4	4	0	0

1 row | 1-10 of 11 columns

Hide

```
starwars %>%
  summarise(across(where(is.numeric), ~sum((is.na(.x)))))
```

height	mass	birth_year
<int>	<int>	<int>
6	28	44

1 row

Hide

```
starwars %>%
  summarise(across(where(is.numeric) & !c(height), ~sum((is.na(.x)))))
```

mass	birth_year
<int>	<int>
28	44

1 row

task 7:Using the dplyr package in R to calculate the count of missing values across numeric columns (excluding the height column) in the starwars dataset, generating column-wise summaries with customized names.

Hide

```
starwars %>%
  summarise(across(where(is.numeric) & !c(height), ~sum((is.na(.x))), .names = "num_missing_
{.col}"))
```

num_missing_mass	num_missing_birth_year
<int>	<int>
28	44

1 row

Hide

```
starwars %>%
  summarise(across(where(is.numeric) & !c(height), ~sum((is.na(.x))), .names = "{.col}_num_mi
ssing"))
```

mass_num_missing <int>	birth_year_num_missing <int>
28	44

1 row

task 8:Using dplyr to compute mean and standard deviation for numeric columns (excluding height) in the dataset, summarizing these statistics in a structured format.

Hide

```
starwars %>%
  summarise(across(where(is.numeric) & !c(height), list(mean =  ~mean(.x, na.rm=T),
                                                                sd   =  ~sd(.x, na.rm=
T))))
```

mass_mean <dbl>	mass_sd <dbl>	birth_year_mean <dbl>	birth_year_sd <dbl>
97.31186	169.4572	87.56512	154.6914

1 row

task 9:It selects columns from name to mass in the dataset, imputes missing numeric values with the mean of each respective column, and then filters rows where either height or mass is missing.

Hide

```
starwars %>%
  select(name:mass) %>%
  mutate(across(where(is.numeric), list(imputed =  ~ifelse(is.na(.x), mean(.x, na.rm=T), .
x)))) %>%
  filter(is.na(height) | is.na(mass))
```

name <chr>	height <int>	mass <dbl>	height_imputed <dbl>	mass_imputed <dbl>
Wilhuff Tarkin	180	NA	180.0000	97.31186
Mon Mothma	150	NA	150.0000	97.31186
Arvel Crynyd	NA	NA	174.6049	97.31186
Finis Valorum	170	NA	170.0000	97.31186
Rugor Nass	206	NA	206.0000	97.31186
Ric Olié	183	NA	183.0000	97.31186
Watto	137	NA	137.0000	97.31186
Quarsh Panaka	183	NA	183.0000	97.31186
Shmi Skywalker	163	NA	163.0000	97.31186
Bib Fortuna	180	NA	180.0000	97.31186

1-10 of 28 rows Previous **1** 2 3 Next

Case Study: picking players for a basketball tournament

task 1:It filters characters with known heights, categorizes them into teams based on species.

Hide

```
starwars %>%
  select(name, species, height, mass) %>%
  filter(!is.na(height)) %>%     # remove any characters with missing height
  mutate(team=case_when(    species == "Human" ~ "Human",   # create the 3 teams
                                    species == "Droid" ~ "Droid",
                                    T ~ "All-Star 5")) %>%
  group_by(species) %>%
  mutate(mass=ifelse(    is.na(mass),  # for any characters missing mass use their species average
                                mean(mass, na.rm=T),
                                mass)) %>%
  ungroup() %>%
  mutate(mass=ifelse(is.na(mass),  # for any characters still missing mass use total avg
                               mean(mass, na.rm=T),
                               mass)) %>%
  arrange(team, desc(height), mass) %>% # sort by team, tallest - shortest and split any ties by lightest first
  group_by(team) %>%
  mutate(rank=row_number()) %>% # rank by tallest-shortest by team
  filter(rank<=5) # keep the top 5 as they will form our team
```

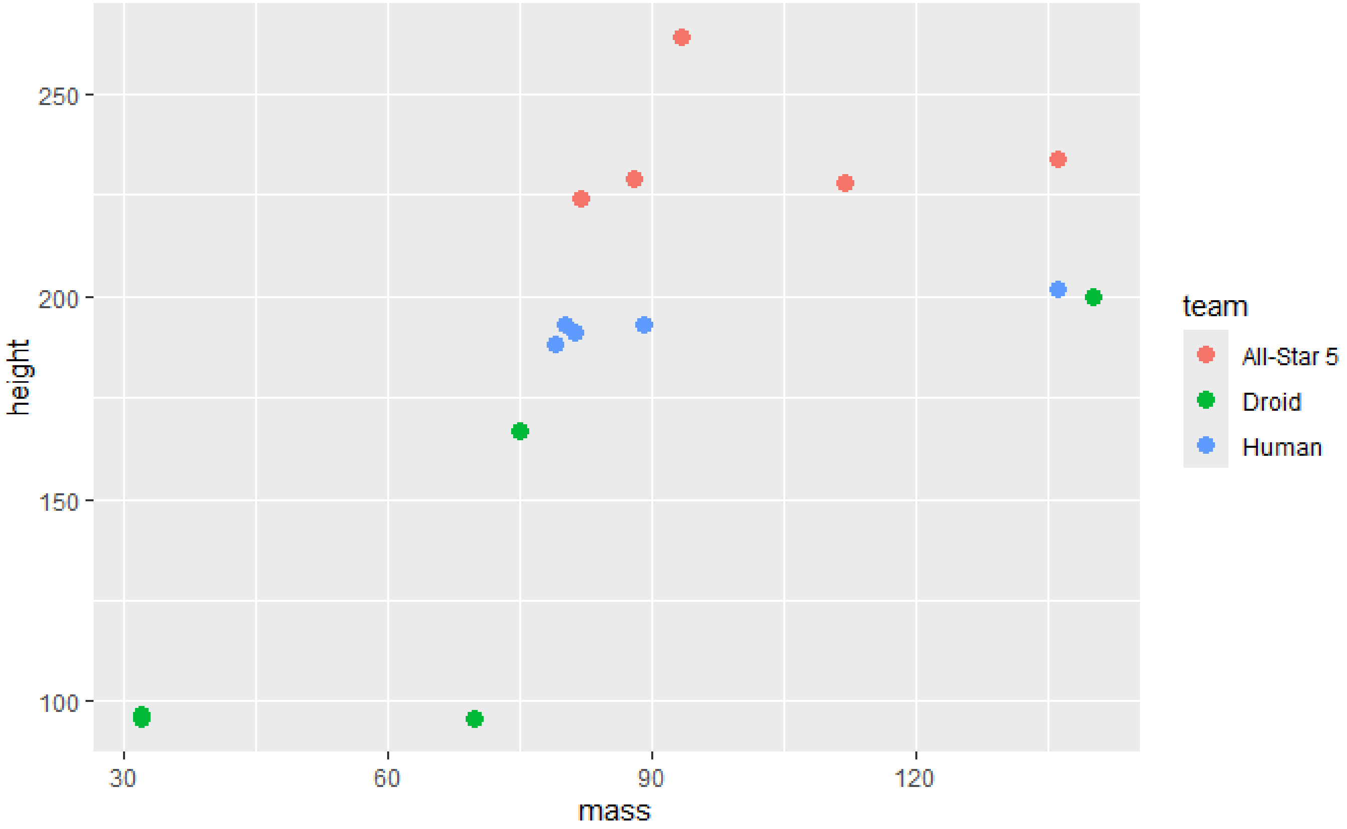
name <chr>	species <chr>	height <int>	mass <dbl>	team <chr>	rank <int>
Yarael Poof	Quermian	264	93.36333	All-Star 5	1
Tarfful	Wookiee	234	136.00000	All-Star 5	2
Lama Su	Kaminoan	229	88.00000	All-Star 5	3
Chewbacca	Wookiee	228	112.00000	All-Star 5	4
Roos Tarpals	Gungan	224	82.00000	All-Star 5	5
IG-88	Droid	200	140.00000	Droid	1
C-3PO	Droid	167	75.00000	Droid	2
R5-D4	Droid	97	32.00000	Droid	3
R2-D2	Droid	96	32.00000	Droid	4
R4-P17	Droid	96	69.75000	Droid	5

1-10 of 15 rows   Previous **1** 2 Next

task 2:Selecting the top 5 by height within each team, and creates a scatterplot to visualize mass against height using ggplot2.

Hide

```
starwars %>%
  select(name, species, height, mass) %>%
  filter(!is.na(height)) %>%     # remove any characters with missing height
  mutate(team=case_when(species == "Human" ~ "Human",    # create the 3 teams
                        species == "Droid" ~ "Droid",
                        T ~ "All-Star 5")) %>%
  group_by(species) %>%
  mutate(mass=ifelse(is.na(mass),  # for any characters missing mass use their species average
                     mean(mass, na.rm=T),
                     mass)) %>%
  ungroup() %>%
  mutate(mass=ifelse(is.na(mass),  # for any characters still missing mass use total avg
                     mean(mass, na.rm=T),
                     mass)) %>%
  arrange(team, desc(height), mass) %>% # sort by team, tallest - shortest and split any ties by lightest first
  group_by(team) %>%
  mutate(rank=row_number()) %>% # rank by tallest-shortest by team
  filter(rank<=5) %>%  # keep the top 5 as they will form our team
    ggplot(aes(x=mass, y=height, colour=team)) +     # Pass the pipeline to tidyverse plotting package ggplot2
      geom_point(size=3)  # Create a scatterplot of height and weight to get an idea of who might win
```



Case Study: finding the outlier

task 1:It calculates the average and median mass of characters in the dataset, ignoring missing values.

Hide

```
starwars %>%
  summarise(avg_mass=mean(mass, na.rm=T),
                     median_mass=median(mass, na.rm=T))
```

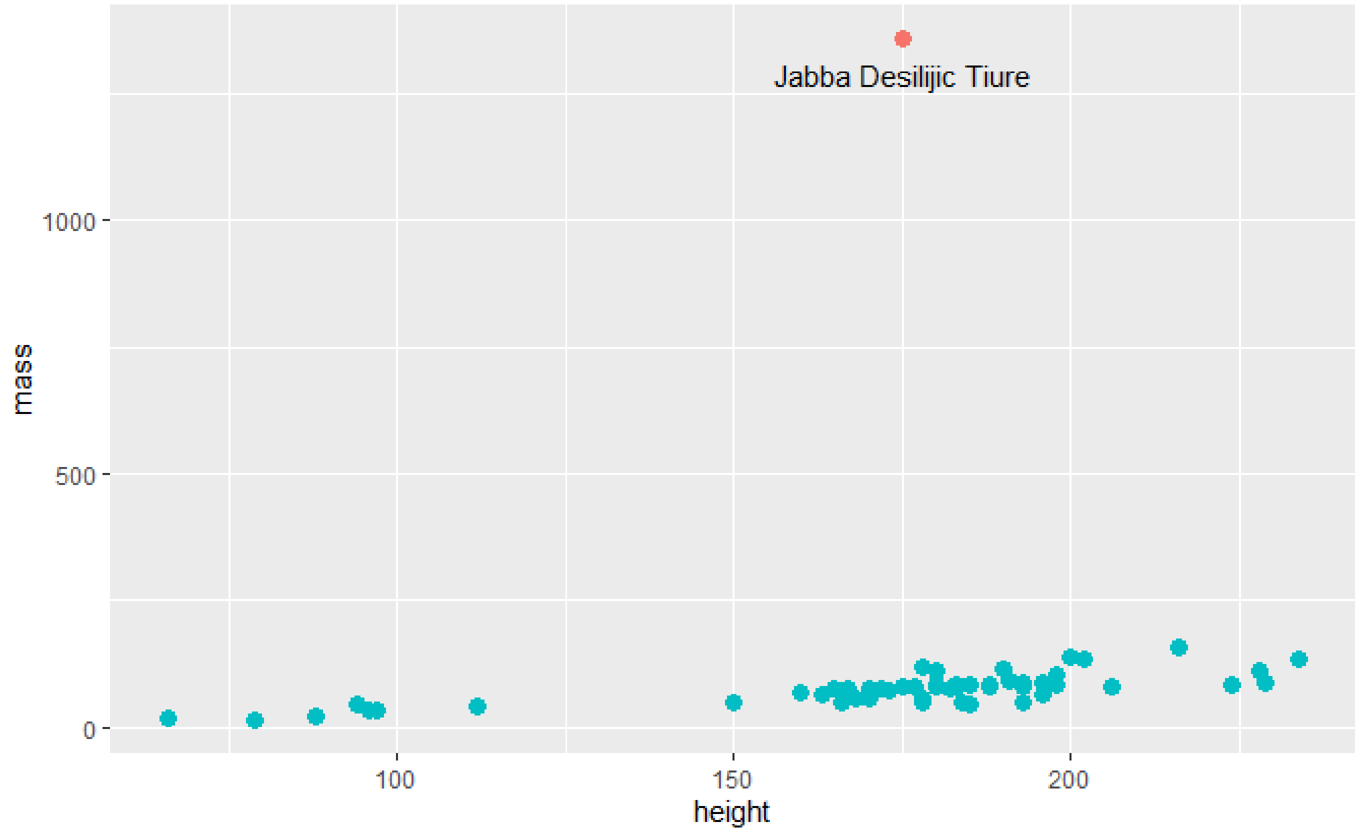
avg_mass	median_mass
<dbl>	<dbl>
97.31186	79

1 row

task 2:Displaying the outlier

Hide

```
starwars %>%
  select(name, height, mass) %>%     # keep name, height and mass columns
  filter(!is.na(mass)) %>%                   # remove any rows with missing mass
  mutate(avg_mass=mean(mass),      # calculate the average mass for the data set
              SD_mass=sd(mass),           # calculate the standard deviation of mass for the
data set
              outlier=ifelse(mass>(avg_mass+(3*SD_mass)),1,0)) %>%  # flag rows where mass >
mean + 3xSD
  arrange(desc(outlier)) %>%             # sort the data set so the outlier is at the top
  ggplot(aes(x=height, y=mass)) +    # pass it to ggplot() to visualise the data
  geom_point(aes(colour = as.factor(-outlier)), size=3) +
  geom_text(aes(label=ifelse(outlier==1, as.character(name),'')), hjust=0.5, vjust=2) +
  theme(legend.position="none")
```



Reshaping data with pivot from tidyr

task 1:Creating the data

Hide

```
# Create some long data
starwars %>%
  filter(species %in% c("Human", "Kaminoan", "Twi'lek")) %>%
  group_by(species, sex) %>%
  summarise(avg_height= mean(height, na.rm=T))
```

```
`summarise()` has grouped output by 'species'. You can override using the `.groups` argument.
```

species <chr>	sex <chr>	avg_height <dbl>
Human	female	163.5714
Human	male	182.3913
Kaminoan	female	213.0000
Kaminoan	male	229.0000
Twi'lek	female	178.0000
Twi'lek	male	180.0000

6 rows

Hide

```
# Add pivot_wider() to reshape the data
starwars %>%
  filter(species %in% c("Human", "Kaminoan", "Twi'lek")) %>%
  group_by(species, sex) %>%
  summarise(avg_height= mean(height, na.rm=T)) %>%
  pivot_wider(id_cols=species,
                    names_from=sex,
                    values_from=avg_height)
```

```
`summarise()` has grouped output by 'species'. You can override using the `.groups` argument.
```

species <chr>	female <dbl>	male <dbl>
Human	163.5714	182.3913
Kaminoan	213.0000	229.0000
Twi'lek	178.0000	180.0000

3 rows

task 2:sorting the results by the number of missing values in descending order

Hide

```
# Create some wide data
starwars %>%
  summarise(across(everything(), ~sum((is.na(.x)))))
```

n…	height	m…	hair_color	skin_color	eye_color	birth_year	s..	gen…	homewo…
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
0	6	28	5	0	0	44	4	4	10

1 row | 1-10 of 14 columns

Hide

```
# Add pivot_longer()
starwars %>%
  summarise(across(everything(), ~sum((is.na(.x))))) %>%
  pivot_longer( cols=everything(),
                   names_to = "variable",
                   values_to = "number_of_missing") %>%
  arrange(desc(number_of_missing))
```

variable	number_of_missing
<chr>	<int>
birth_year	44
mass	28
homeworld	10
height	6
hair_color	5
sex	4
gender	4
species	4
name	0
skin_color	0

1-10 of 14 rows

Previous **1** 2 Next