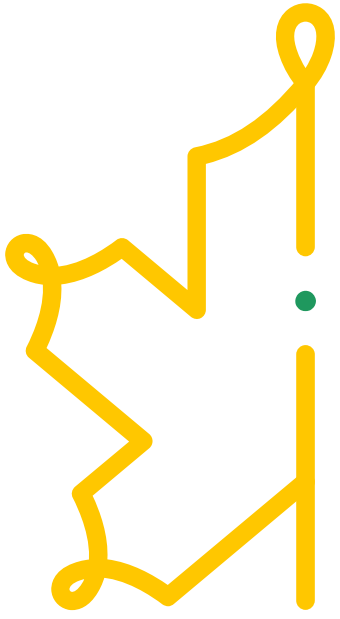
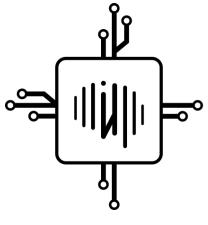


Razвед анализ данных

Введение в Tidyverse

tibbles, dplyr, readr



Фонд интеллект

Анастасия Жарикова

Лекция 3 - 2022

Пакет пакетов

~~Забудьте все, чему вас учили...~~

Пакеты

Репозитории

- CRAN - https://cran.r-project.org/: install.packages()
- Bioconductor - https://www.bioconductor.org/
- Github - https://github.com/

Пакеты

Полезные команды

- install.packages("packagename") - установить пакет
- remove.packages("packagename") - удалить пакет
- update.packages() - обновить все пакеты
- library() - список доступных пакетов
- library("packagename") - загрузить установленный пакет в текущую R сессию
- vignette("packagename") - посмотреть "красивый" мануал по пакету, есть не для всех пакетов

Пакеты

Пакет нужно установить один раз из соответствующего репозитория

Затем нужно только подгрузить пакет с помощью library(packagename) при каждом запуске рабочего сеанса

tidyverse

install.packages("tidyverse")

library(tidyverse)

https://www.tidyverse.org/

```
tidyverse_packages()
```

```
 [1] "broom"         "cli"           "crayon"        "dbplyr"
 [5] "dplyr"         "dtplyr"        "forcats"       "googledrive"
 [9] "googlesheets4" "ggplot2"       "haven"         "hms"
[13] "httr"          "jsonlite"      "lubridate"     "magrittr"
[17] "modelr"        "pillar"        "purrr"         "readr"
[21] "readxl"        "reprex"        "rlang"         "rstudioapi"
[25] "rvest"         "stringr"       "tibble"        "tidyr"
[29] "xml2"          "tidyverse"
```

tibbles

Похоже на data.frame

Ведут себя более предсказуемо и удобно

При работе с данными с помощью коллекции пакетов tidyverse в большинстве случаев на выходе получаются tibble-фреймы

tibbles

Создание

```
tibble(
  x = 1:5,
  y = 1,
  `1z b` = x^2 + y
)
```

```
# A tibble: 5 × 3
      x     y `1z b`
  <int> <dbl>  <dbl>
1     1     1      2
2     2     1      5
3     3     1     10
4     4     1     17
5     5     1     26
```

tibbles

Не преобразует строки в факторы

Не изменяет имена переменных

Можно (но нужно ли?) использовать "недопустимые" имена столбцов

Выводит информацию о размере выводимого фрейма и типе данных в столбцах

Позволяет ссылаться на только что созданные переменные

tibbles

Создание

```
tribble(
  ~x,~y,~z,
  "a",2,3.6,
  "b",5,1.0
)
```

```
# A tibble: 2 × 3
  x         y     z
  <chr> <dbl> <dbl>
1 a         2   3.6
2 b         5   1
```

tibbles

Создание

```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

tibbles

Создание

```
as_tibble(iris)
```

```
# A tibble: 150 × 5
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
          <dbl>       <dbl>        <dbl>       <dbl> <fct>
 1          5.1         3.5          1.4         0.2 setosa
 2          4.9         3            1.4         0.2 setosa
 3          4.7         3.2          1.3         0.2 setosa
 4          4.6         3.1          1.5         0.2 setosa
 5          5           3.6          1.4         0.2 setosa
 6          5.4         3.9          1.7         0.4 setosa
 7          4.6         3.4          1.4         0.3 setosa
 8          5           3.4          1.5         0.2 setosa
 9          4.4         2.9          1.4         0.2 setosa
10          4.9         3.1          1.5         0.1 setosa
# … with 140 more rows
```

Выведет только 10 первых строк и помещающиеся столбцы, а не "простыню"

dplyr

https://dplyr.tidyverse.org/reference/index.html

Очень много разных функций

У каждой функции очень много разных опций

Разберем только некоторые наиболее употребимые

starwars

```
head(starwars, 4)
```

```
# A tibble: 4 × 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, blue red             33   none  mascu…
4 Darth V…    202   136 none       white       yellow          41.9 male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Фильтрация строк по условию - filter()

```
nrow(starwars)
```

```
[1] 87
```

```
a <- filter(starwars, height > 150)
nrow(a)
```

```
[1] 69
```

dplyr

Фильтрация строк по условию - filter()

Подключаем конвейер %>%

```
starwars %>%
  filter(height > 150)
```

```
# A tibble: 69 × 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 S…    172    77 blond       fair       blue            19   male  mascu…
 2 C-3PO      167    75 <NA>        gold       yellow         112   none  mascu…
 3 Darth …    202   136 none        white      yellow          41.9 male  mascu…
 4 Owen L…    178   120 brown, grey light      blue            52   male  mascu…
 5 Beru W…    165    75 brown       light      blue            47   fema… femin…
 6 Biggs …    183    84 black       light      brown           24   male  mascu…
 7 Obi-Wa…    182    77 auburn, wh… fair       blue-gray       57   male  mascu…
 8 Anakin…    188    84 blond       fair       blue            41.9 male  mascu…
 9 Wilhuf…    180    NA auburn, gr… fair       blue            64   male  mascu…
10 Chewba…    228   112 brown       unknown    blue           200   male  mascu…
# … with 59 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

dplyr

Фильтрация строк по условию - filter()

```
starwars %>%
  filter(height > 150, mass < 100, hair_color == "blond")
```

```
# A tibble: 2 × 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 Anakin S…    188    84 blond      fair       blue            41.9 male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Фильтрация строк по условию - filter()

```
starwars %>%
  filter(between(height, 100, 150))
```

```
# A tibble: 5 × 14
  name     height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Leia Or…    150    49 brown      light      brown             19 fema… femini…
2 Mon Mot…    150    NA auburn     fair       blue              48 fema… femini…
3 Watto       137    NA black      blue, grey yellow            NA male  mascul…
4 Sebulba     112    40 none       grey, red  orange            NA male  mascul…
5 Gasgano     122    NA none       white, bl… black             NA male  mascul…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор строк по позициям - slice()

```
starwars %>%
  slice(10:13)
```

```
# A tibble: 4 × 14
  name     height  mass hair_color  skin_color eye_color birth_year sex   gender
  <chr>     <int> <dbl> <chr>       <chr>      <chr>          <dbl> <chr> <chr>
1 Obi-Wan…    182    77 auburn, wh… fair       blue-gray       57   male  mascu…
2 Anakin …    188    84 blond       fair       blue            41.9 male  mascu…
3 Wilhuff…    180    NA auburn, gr… fair       blue            64   male  mascu…
4 Chewbac…    228   112 brown       unknown    blue           200   male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Конвейерный подход - можно комбинировать все!

```
starwars %>%
  filter(height > 150, hair_color == "blond") %>%
  slice(1:3)
```

```
# 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 Anakin S…    188    84 blond      fair       blue            41.9 male  mascu…
3 Finis Va…    170    NA blond      fair       blue            91   male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор строк по позициям - slice()

n() - число записей

```
nrow(starwars)
```

```
[1] 87
```

```
starwars %>%
  slice(-(11:n())) %>%
  nrow()
```

```
[1] 10
```

dplyr

Выбор строк по позициям - slice()

```
starwars %>%
  slice_head(n=2)
```

```
# A tibble: 2 × 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  mascul…
2 C-3PO       167    75 <NA>       gold       yellow           112 none  mascul…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

```
starwars %>%
  head(2)
```

```
# A tibble: 2 × 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  mascul…
2 C-3PO       167    75 <NA>       gold       yellow           112 none  mascul…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор строк по позициям - slice()

```
set.seed(123)
starwars %>%
  slice_sample(n=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 Qui-Gon …    193    89 brown      fair       blue               92 male  mascu…
2 Raymus A…    188    79 brown      light      brown              NA male  mascu…
3 Eeth Koth    171    NA black      brown      brown              NA male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор строк по позициям - slice()

```
starwars %>%
  slice_max(mass)
```

```
# 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 Jabba …    175  1358 <NA>       green-tan,… orange           600 herma… mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор строк по позициям - slice()

```
starwars %>%
  slice_max(mass, n=4)
```

```
# A tibble: 5 × 14
  name    height  mass hair_color skin_color  eye_color  birth_year sex   gender
  <chr>    <int> <dbl> <chr>      <chr>       <chr>           <dbl> <chr> <chr>
1 Jabba …    175  1358 <NA>       green-tan,… orange          600   herm… mascu…
2 Grievo…    216   159 none       brown, whi… green, ye…       NA   male  mascu…
3 IG-88      200   140 none       metal       red              15   none  mascu…
4 Darth …    202   136 none       white       yellow           41.9 male  mascu…
5 Tarfful    234   136 brown      brown       blue             NA   male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор строк по позициям - slice()

```
starwars %>%
  slice_max(mass, n=4, with_ties=F)
```

```
# A tibble: 4 × 14
  name    height  mass hair_color skin_color  eye_color  birth_year sex   gender
  <chr>    <int> <dbl> <chr>      <chr>       <chr>           <dbl> <chr> <chr>
1 Jabba …    175  1358 <NA>       green-tan,… orange          600   herm… mascu…
2 Grievo…    216   159 none       brown, whi… green, ye…       NA   male  mascu…
3 IG-88      200   140 none       metal       red              15   none  mascu…
4 Darth …    202   136 none       white       yellow           41.9 male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

dplyr

Выбор колонок по имени - select()

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

```
# A tibble: 87 × 4
   name               height  mass hair_color
   <chr>               <int> <dbl> <chr>
 1 Luke Skywalker        172    77 blond
 2 C-3PO                 167    75 <NA>
 3 R2-D2                  96    32 <NA>
 4 Darth Vader           202   136 none
 5 Leia Organa           150    49 brown
 6 Owen Lars             178   120 brown, grey
 7 Beru Whitesun lars    165    75 brown
 8 R5-D4                  97    32 <NA>
 9 Biggs Darklighter     183    84 black
10 Obi-Wan Kenobi        182    77 auburn, white
# … with 77 more rows
```

dplyr

Выбор колонок по имени - select()

```
starwars %>%
  select(name:hair_color) %>%
  head(3)
```

```
# A tibble: 3 × 4
  name           height  mass hair_color
  <chr>           <int> <dbl> <chr>
1 Luke Skywalker    172    77 blond
2 C-3PO             167    75 <NA>
3 R2-D2              96    32 <NA>
```

dplyr

Выбор колонок по имени - select()

```
starwars %>%
  select(-(name:hair_color)) %>%
  head(3)
```

```
# A tibble: 3 × 10
  skin_color  eye_color birth_year sex   gender homeworld species films vehicles
  <chr>       <chr>          <dbl> <chr> <chr>  <chr>     <chr>   <lis> <list>
1 fair        blue              19 male  mascu… Tatooine  Human   <chr… <chr [2…
2 gold        yellow           112 none  mascu… Tatooine  Droid   <chr… <chr [0…
3 white, blue red               33 none  mascu… Naboo     Droid   <chr… <chr [0…
# … with 1 more variable: starships <list>
```

dplyr

Выбор колонок по имени - select()

```
starwars %>%
  select(!c(name, hair_color)) %>%
  head(3)
```

```
# A tibble: 3 × 12
  height  mass skin_color  eye_color birth_year sex   gender   homeworld species
   <int> <dbl> <chr>       <chr>          <dbl> <chr> <chr>    <chr>     <chr>
1    172    77 fair        blue              19 male  masculi… Tatooine  Human
2    167    75 gold        yellow           112 none  masculi… Tatooine  Droid
3     96    32 white, blue red               33 none  masculi… Naboo     Droid
# … with 3 more variables: films <list>, vehicles <list>, starships <list>
```

dplyr

Выбор колонок по имени - select ()

```
as_tibble(iris) %>%
  select(starts_with('Sepal'))
```

```
# A tibble: 150 × 2
   Sepal.Length Sepal.Width
          <dbl>       <dbl>
 1          5.1         3.5
 2          4.9         3
 3          4.7         3.2
 4          4.6         3.1
 5          5           3.6
 6          5.4         3.9
 7          4.6         3.4
 8          5           3.4
 9          4.4         2.9
10          4.9         3.1
# … with 140 more rows
```

Аналогично:

starts with(), ends with(), matches() и contains()

dplyr

Выбор колонок по имени - select ()

```
starwars %>%
  select(where(is.numeric))
```

```
# A tibble: 87 × 3
   height  mass birth_year
    <int> <dbl>      <dbl>
 1    172    77       19
 2    167    75      112
 3     96    32       33
 4    202   136       41.9
 5    150    49       19
 6    178   120       52
 7    165    75       47
 8     97    32       NA
 9    183    84       24
10    182    77       57
# … with 77 more rows
```

dplyr

Выбор колонок по имени - select()

```
starwars %>%
  select(height, everything()) %>%
  head(3)
```

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

dplyr

Изменение порядка колонок - relocate()

```
starwars %>%
  relocate(height)
```

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

dplyr

Изменение порядка колонок - relocate()

```
starwars %>%
  relocate(height, .before = name)
```

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

dplyr

Выбрать 1 столбец и взять как вектор - pull()

```
starwars %>%
  select(name) %>% head(3)
```

```
# A tibble: 3 × 1
  name
  <chr>
1 Luke Skywalker
2 C-3PO
3 R2-D2
```

```
starwars %>%
  pull(name) %>% head(3)
```

```
[1] "Luke Skywalker" "C-3PO"          "R2-D2"
```

dplyr

Сортировка по колонкам - arrange()

```
starwars %>%
  arrange(name,desc(mass),hair_color) %>%
  slice_head(n = 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 Ackbar    180    83 none       brown mottle orange             41 male  mascul…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Сначала сортируем по столбцу name по возрастанию.

Потом сортируем по столбцу mass по убыванию.

Потом сортируем по столбцу hair color по возрастанию.

Отсутствующие значения ВСЕГДА в конце

dplyr

Подсчитать новую колонку - mutate()

```
starwars %>%
  select(name:hair_color) %>%
  mutate(HM = height*mass)
```

```
# A tibble: 87 × 5
   name               height  mass hair_color       HM
   <chr>               <int> <dbl> <chr>         <dbl>
 1 Luke Skywalker        172    77 blond         13244
 2 C-3PO                 167    75 <NA>          12525
 3 R2-D2                  96    32 <NA>           3072
 4 Darth Vader           202   136 none          27472
 5 Leia Organa           150    49 brown          7350
 6 Owen Lars             178   120 brown, grey   21360
 7 Beru Whitesun lars    165    75 brown         12375
 8 R5-D4                  97    32 <NA>           3104
 9 Biggs Darklighter     183    84 black         15372
10 Obi-Wan Kenobi        182    77 auburn, white 14014
# … with 77 more rows
```

dplyr

Подсчитать новую колонку с условием - mutate() + case_when()

```
starwars %>%
  select(name:eye_color, species) %>%
  mutate(
    height_type = case_when(height > median(height, na.rm = T) ~ 'tall',
                            height <= median(height, na.rm = T) ~ 'short',
                            is.na(height) ~ "nope"))
```

```
# A tibble: 87 × 8
   name        height  mass hair_color  skin_color eye_color species height_type
   <chr>        <int> <dbl> <chr>       <chr>      <chr>     <chr>   <chr>
 1 Luke Skywa…    172    77 blond       fair       blue      Human   short
 2 C-3PO          167    75 <NA>        gold       yellow    Droid   short
 3 R2-D2           96    32 <NA>        white, bl… red       Droid   short
 4 Darth Vader    202   136 none        white      yellow    Human   tall
 5 Leia Organa    150    49 brown       light      brown     Human   short
 6 Owen Lars      178   120 brown, grey light      blue      Human   short
 7 Beru White…    165    75 brown       light      blue      Human   short
 8 R5-D4           97    32 <NA>        white, red red       Droid   short
 9 Biggs Dark…    183    84 black       light      brown     Human   tall
10 Obi-Wan Ke…    182    77 auburn, wh… fair       blue-gray Human   tall
# … with 77 more rows
```

dplyr

Подсчитать новую колонку и убрать остальные - transmute()

```
starwars %>%
  transmute(inv_mass = 1/mass,
            inv_height = 1/height) %>% head(3)
```

```
# A tibble: 3 × 2
  inv_mass inv_height
     <dbl>      <dbl>
1   0.0130    0.00581
2   0.0133    0.00599
3   0.0312    0.0104
```

dplyr

Переименовать колонку - rename()

```
starwars %>%
  select(name:hair_color) %>%
  rename(NAME=name)
```

```
# A tibble: 87 × 4
   NAME               height  mass hair_color
   <chr>               <int> <dbl> <chr>
 1 Luke Skywalker        172    77 blond
 2 C-3PO                 167    75 <NA>
 3 R2-D2                  96    32 <NA>
 4 Darth Vader           202   136 none
 5 Leia Organa           150    49 brown
 6 Owen Lars             178   120 brown, grey
 7 Beru Whitesun lars    165    75 brown
 8 R5-D4                  97    32 <NA>
 9 Biggs Darklighter     183    84 black
10 Obi-Wan Kenobi        182    77 auburn, white
# … with 77 more rows
```

dplyr

Переименовать колонку с помощью функции - rename_with()

```
starwars %>%
  select(name:hair_color) %>%
  rename_with(toupper) %>% head(3)
```

```
# A tibble: 3 × 4
  NAME           HEIGHT  MASS HAIR_COLOR
  <chr>           <int> <dbl> <chr>
1 Luke Skywalker    172    77 blond
2 C-3PO             167    75 <NA>
3 R2-D2              96    32 <NA>
```

dplyr

Взять уникальные значения - distinct()

```
df <- tribble(
  ~a,~b,
  "a",1,
  "a",1,
  "a",2,
  "b",3,
  "b",3
)

distinct(df)
```

```
# A tibble: 3 × 2
  a         b
  <chr> <dbl>
1 a         1
2 a         2
3 b         3
```

dplyr

Получение групповых итогов - summarize()

```
starwars %>%
  summarise(mass_mean_noNA = mean(mass, na.rm=T),
            mass_mean_withNA = mean(mass),
            heihgt_max = max(height,na.rm = T),
            count = n(),
            sp_n = n_distinct(species))
```

```
# A tibble: 1 × 5
  mass_mean_noNA mass_mean_withNA heihgt_max count  sp_n
           <dbl>            <dbl>      <int> <int> <int>
1           97.3               NA        264    87    38
```

dplyr

Работа с группами - group_by() + summarise()

```
starwars %>%
  select(name:eye_color, species) %>%
  drop_na() %>%
  group_by(eye_color) %>%
  summarise(count = n(),
            height_max = max(height),
            sp_n = n_distinct(species)) %>% head(5)
```

```
# A tibble: 5 × 4
  eye_color     count height_max  sp_n
  <chr>         <int>      <int> <int>
1 black             6        229     6
2 blue             12        234     4
3 blue-gray         1        182     1
4 brown            13        193     3
5 green, yellow     1        216     1
```

Число строк, максимальный вес и количество уникальных видов подсчитаны для каждого цвета глаз отдельно

dplyr

Работа с группами

После применения group by() все манипуляции будут проходить для каждой группы отдельно

Если далее нужно вернуться к работе с полным набором данных, нужно разгруппировать tibble-фрейм - ungroup()

dplyr

Работа с группами

```
iris %>%
  group_by(Species) %>%
  slice_max(Sepal.Length) %>%
  mutate(meanSL = mean(Sepal.Length))
```

```
# A tibble: 3 × 6
# Groups:   Species [3]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species    meanSL
         <dbl>       <dbl>        <dbl>       <dbl> <fct>       <dbl>
1          5.8         4            1.2         0.2 setosa        5.8
2          7           3.2          4.7         1.4 versicolor    7
3          7.9         3.8          6.4         2   virginica     7.9
```

dplyr

Работа с группами

```
iris %>%
  group_by(Species) %>%
  slice_max(Sepal.Length) %>%
  ungroup() %>%
  mutate(meanSL = mean(Sepal.Length))
```

```
# A tibble: 3 × 6
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species    meanSL
         <dbl>       <dbl>        <dbl>       <dbl> <fct>       <dbl>
1          5.8         4            1.2         0.2 setosa        6.9
2          7           3.2          4.7         1.4 versicolor    6.9
3          7.9         3.8          6.4         2   virginica     6.9
```

dplyr

Работа с группами

```
starwars %>%
  group_by(species, skin_color, sex) %>%
  summarise(n = n()) %>% head(2)
```

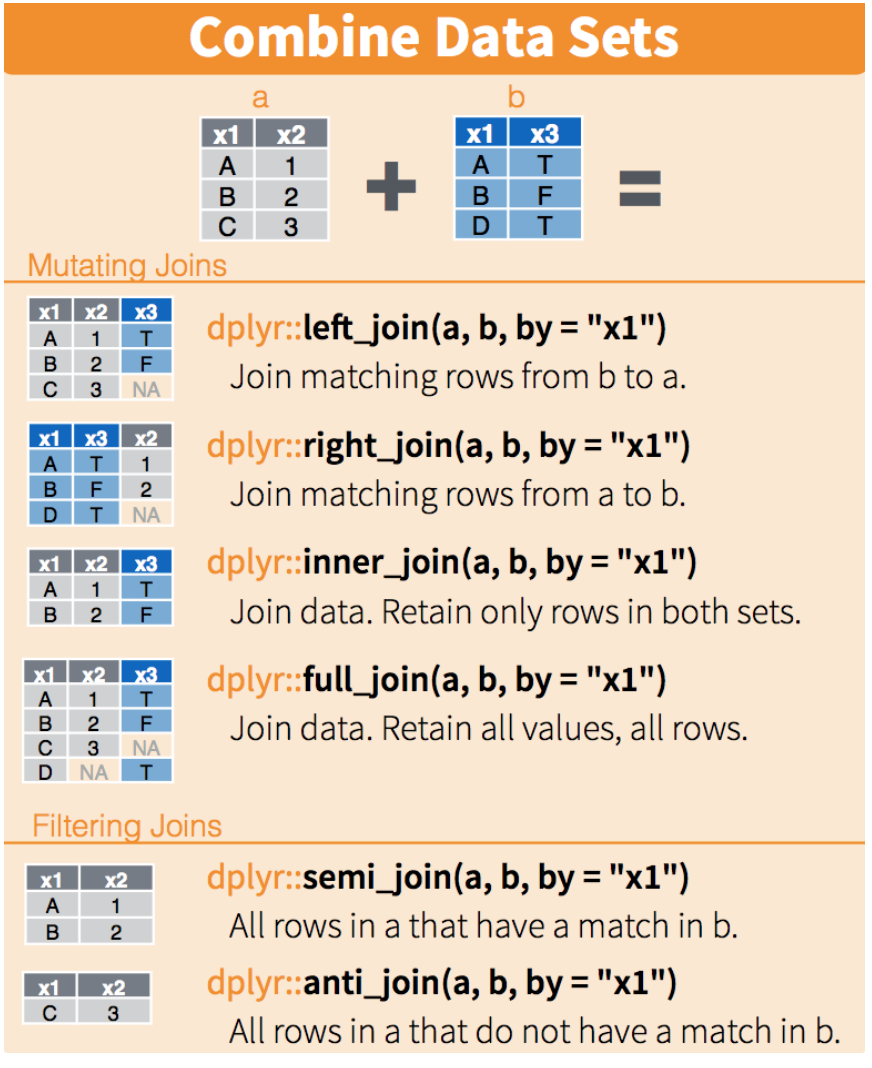
```
# A tibble: 2 × 4
# Groups:   species, skin_color [2]
  species  skin_color sex       n
  <chr>    <chr>      <chr> <int>
1 Aleena   grey, blue male      1
2 Besalisk brown      male      1
```

```
starwars %>%
  count(species, skin_color, sex) %>% head(2)
```

```
# A tibble: 2 × 4
  species  skin_color sex       n
  <chr>    <chr>      <chr> <int>
1 Aleena   grey, blue male      1
2 Besalisk brown      male      1
```

dplyr

Объединение двух таблиц - join



dplyr

Объединение двух таблиц - join

```
band_members
```

```
# A tibble: 3 × 2
  name  band
  <chr> <chr>
1 Mick  Stones
2 John  Beatles
3 Paul  Beatles
```

```
band_instruments
```

```
# A tibble: 3 × 2
  name  plays
  <chr> <chr>
1 John  guitar
2 Paul  bass
3 Keith guitar
```

dplyr

Объединение двух таблиц - join

```
band_members %>% inner_join(band_instruments)
```

```
# A tibble: 2 × 3
  name  band    plays
  <chr> <chr>   <chr>
1 John  Beatles guitar
2 Paul  Beatles bass
```

```
band_members %>% full_join(band_instruments)
```

```
# A tibble: 4 × 3
  name  band    plays
  <chr> <chr>   <chr>
1 Mick  Stones  <NA>
2 John  Beatles guitar
3 Paul  Beatles bass
4 Keith <NA>    guitar
```

dplyr

Объединение двух таблиц - join

```
band_members %>% left_join(band_instruments)
```

```
# A tibble: 3 × 3
  name  band    plays
  <chr> <chr>   <chr>
1 Mick  Stones  <NA>
2 John  Beatles guitar
3 Paul  Beatles bass
```

```
band_members %>% right_join(band_instruments)
```

```
# A tibble: 3 × 3
  name  band    plays
  <chr> <chr>   <chr>
1 John  Beatles guitar
2 Paul  Beatles bass
3 Keith <NA>    guitar
```

dplyr

Объединение двух таблиц - join

```
band_members %>% semi_join(band_instruments)
```

```
# A tibble: 2 × 2
  name  band
  <chr> <chr>
1 John  Beatles
2 Paul  Beatles
```

```
band_members %>% anti_join(band_instruments)
```

```
# A tibble: 1 × 2
  name  band
  <chr> <chr>
1 Mick  Stones
```

Чтение файла `readr`

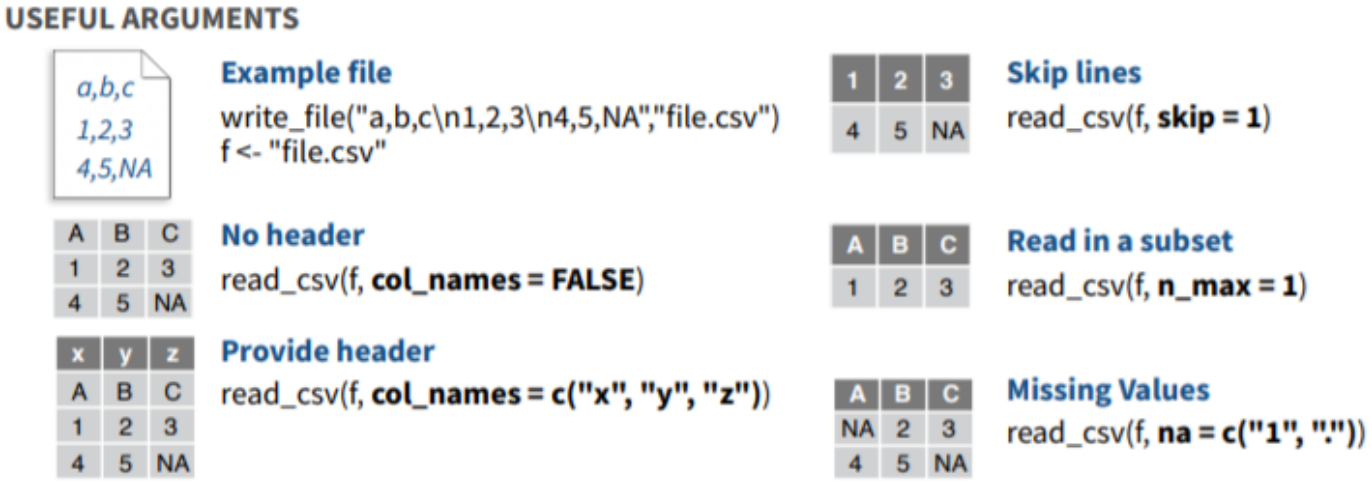
```
library(readr) # или library(tidyverse)

penguins <- read_csv("penguins.csv")
penguins
```

```
# A tibble: 344 × 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
   <chr>   <chr>              <dbl>         <dbl>             <dbl>       <dbl>
 1 Adelie  Torgersen           39.1          18.7               181        3750
 2 Adelie  Torgersen           39.5          17.4               186        3800
 3 Adelie  Torgersen           40.3          18                 195        3250
 4 Adelie  Torgersen           NA            NA                  NA          NA
 5 Adelie  Torgersen           36.7          19.3               193        3450
 6 Adelie  Torgersen           39.3          20.6               190        3650
 7 Adelie  Torgersen           38.9          17.8               181        3625
 8 Adelie  Torgersen           39.2          19.6               195        4675
 9 Adelie  Torgersen           34.1          18.1               193        3475
10 Adelie  Torgersen           42            20.2               190        4250
# … with 334 more rows, and 2 more variables: sex <chr>, year <dbl>
```

Чтение файла `readr`

Посмотреть все параметры: `?read_csv`.



Запись файла

Посмотреть все параметры: `?write_csv`.

```
write_csv(penguins, "more_penguins.csv", append = TRUE)
```

Чтение xlsx файла `readxl`

```
library(readxl)

my_file <- read_excel("data/my_file.xlsx", sheet = "Best sheet ever")
```