Основы обработки данных с помощью R

vmarinin0

2022-10-19

Цель работы

1. Развить практические навыки использования языка программирования R для обработки данных
2. Закрепить знания базовых типов данных языка R
3. Развить пркатические навыки использования функций обработки данных пакета dplyr – функции select(), filter(), mutate(), arrange(), group_by()

Задание

Проанализировать встроенный в пакет dplyr набор данных starwars с помощью языка R и ответить на вопросы

Подготовка

```
library(dplyr)
```

```
##
## Присоединяю пакет: 'dplyr'
```

```
## Следующие объекты скрыты от 'package:stats':
##
##     filter, lag
```

```
## Следующие объекты скрыты от 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
starwars
```

```
## # A tibble: 87 × 14
##    name        height  mass hair_…¹ skin_…² eye_c…³ birth…⁴ sex   gender homew…⁵
##    <chr>        <int> <dbl> <chr>   <chr>   <chr>     <dbl> <chr> <chr>  <chr>
##  1 Luke Skywa…    172    77 blond   fair    blue       19   male  mascu… Tatooi…
##  2 C-3PO          167    75 <NA>    gold    yellow    112   none  mascu… Tatooi…
##  3 R2-D2           96    32 <NA>    white,… red        33   none  mascu… Naboo
##  4 Darth Vader    202   136 none    white   yellow     41.9 male  mascu… Tatooi…
##  5 Leia Organa    150    49 brown   light   brown      19   fema… femin… Aldera…
##  6 Owen Lars      178   120 brown,… light   blue       52   male  mascu… Tatooi…
##  7 Beru White…    165    75 brown   light   blue       47   fema… femin… Tatooi…
##  8 R5-D4           97    32 <NA>    white,… red        NA   none  mascu… Tatooi…
##  9 Biggs Dark…    183    84 black   light   brown      24   male  mascu… Tatooi…
## 10 Obi-Wan Ke…    182    77 auburn… fair    blue-g…    57   male  mascu… Stewjon
## # … with 77 more rows, 4 more variables: species <chr>, films <list>,
## #   vehicles <list>, starships <list>, and abbreviated variable names
## #   ¹hair_color, ²skin_color, ³eye_color, ⁴birth_year, ⁵homeworld
```

```
starwars <- starwars
```

Задание 1

1. Сколько строк в датафрейме?

```
starwars %>% nrow()
```

```
## [1] 87
```

Задание 2

2. Сколько столбцов в датафрейме?

```
starwars %>% ncol()
```

```
## [1] 14
```

Задание 3

3. Как просмотреть примерный вид датафрейма?

```
starwars %>% glimpse()
```

```
## Rows: 87
## Columns: 14
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or…
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2…
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.…
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N…
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "…
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",…
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, …
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female",…
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini…
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T…
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma…
## $ films      <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return…
## $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp…
## $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",…
```

Задание 4

4. Сколько уникальных рас персонажей (species) представлено в данных?

```
x <- is.na(starwars$species)
length(unique(starwars$species[!x]))
```

```
## [1] 37
```

Задание 5

5. Найти самого высокого персонажа.

```
starwars[which.max(starwars$height),]$name
```

```
## [1] "Yarael Poof"
```

Задание 6

6. Найти всех персонажей ниже 170

```
s <- is.na(starwars$height)
k <- starwars$height[!s]
starwars[starwars$height %in% k & starwars$height <170,]$name
```

```
##  [1] "C-3PO"                 "R2-D2"                 "Leia Organa"
##  [4] "Beru Whitesun lars"    "R5-D4"                 "Yoda"
##  [7] "Mon Mothma"            "Wicket Systri Warrick" "Nien Nunb"
## [10] "Watto"                 "Sebulba"               "Shmi Skywalker"
## [13] "Dud Bolt"              "Gasgano"               "Ben Quadinaros"
## [16] "Cordé"                 "Barriss Offee"         "Dormé"
## [19] "Zam Wesell"            "Jocasta Nu"            "Ratts Tyerell"
## [22] "R4-P17"                "Padmé Amidala"
```

Задание 7

7. Подсчитать ИМТ (индекс массы тела) для всех персонажей. ИМТ подсчитать по формуле I = hm2, где m – масса (weight), а h – рост (height).

```
imt <- starwars %>% filter(!is.na(mass)) %>% filter(!is.na(height))%>%   group_by(name)  %>% summarise(IMT=mass/
(height/100)^2)
knitr::kable(imt, "pipe")
```

name	**IMT**
Ackbar	25.61728
Adi Gallia	14.76843
Anakin Skywalker	23.76641
Ayla Secura	17.35892
Barriss Offee	18.14487
Ben Quadinaros	24.46460
Beru Whitesun lars	27.54821
Biggs Darklighter	25.08286
Boba Fett	23.35095
Bossk	31.30194
C-3PO	26.89232
Chewbacca	21.54509
Darth Maul	26.12245
Darth Vader	33.33007
Dexter Jettster	26.01775
Dooku	21.47709
Dud Bolt	50.92802
Greedo	24.72518
Gregar Typho	24.83565
Grievous	34.07922
Han Solo	24.69136
IG-88	35.00000
Jabba Desilijic Tiure	443.42857
Jango Fett	23.58984
Jar Jar Binks	17.18034
Jek Tono Porkins	33.95062
Ki-Adi-Mundi	20.91623
Kit Fisto	22.64681
Lama Su	16.78076
Lando Calrissian	25.21625
Leia Organa	21.77778
Lobot	25.79592
Luke Skywalker	26.02758
Luminara Unduli	19.44637
Mace Windu	23.76641
Nien Nunb	26.56250
Nute Gunray	24.67038
Obi-Wan Kenobi	23.24598
Owen Lars	37.87401
Padmé Amidala	16.52893
Palpatine	25.95156
Plo Koon	22.63468
Poggle the Lesser	23.88844
Qui-Gon Jinn	23.89326
R2-D2	34.72222
R5-D4	34.00999
Ratts Tyerell	24.03461
Raymus Antilles	22.35174
Roos Tarpals	16.34247
Sebulba	31.88776
Shaak Ti	17.99015
Sly Moore	15.14960
Tarfful	24.83746
Tion Medon	18.85192
Wat Tambor	12.88625
Wedge Antilles	26.64360
Wicket Systri Warrick	25.82645
Yoda	39.02663
Zam Wesell	19.48696

Задание 8

8. Найти 10 самых “вытянутых” персонажей. “Вытянутость” оценить по отношению массы (mass) к росту (height) персонажей.

```
dat <- starwars %>% group_by(name)  %>% summarise(Elongation=mass/height)
head(arrange(dat,desc(Elongation)),10)
```

```
## # A tibble: 10 × 2
##    name                  Elongation
##    <chr>                      <dbl>
##  1 Jabba Desilijic Tiure      7.76
##  2 Grievous                   0.736
##  3 IG-88                      0.7
##  4 Owen Lars                  0.674
##  5 Darth Vader                0.673
##  6 Jek Tono Porkins           0.611
##  7 Bossk                      0.595
##  8 Tarfful                    0.581
##  9 Dexter Jettster            0.515
## 10 Chewbacca                  0.491
```

Задание 9

9. Найти средний возраст персонажей каждой расы вселенной Звездных войн.

```
starwars %>% filter(!is.na(birth_year))%>% filter(!is.na(species)) %>% group_by(species)  %>% summarise(age= mean
(birth_year))
```

```
## # A tibble: 15 × 2
##    species          age
##    <chr>          <dbl>
##  1 Cerean          92
##  2 Droid           53.3
##  3 Ewok             8
##  4 Gungan          52
##  5 Human           53.4
##  6 Hutt           600
##  7 Kel Dor         22
##  8 Mirialan        49
##  9 Mon Calamari    41
## 10 Rodian          44
## 11 Trandoshan      53
## 12 Twi'lek         48
## 13 Wookiee        200
## 14 Yoda's species 896
## 15 Zabrak          54
```

Задание 10

10. Найти самый распространенный цвет глаз персонажей вселенной Звездных войн.

```
eye <- starwars %>% group_by(eye_color)  %>% summarise(count=n())
head(arrange(eye,desc(count)),1)
```

```
## # A tibble: 1 × 2
##   eye_color count
##   <chr>     <int>
## 1 brown        21
```

Задание 11

11. Подсчитать среднюю длину имени в каждой расе вселенной Звездных войн.

```
sr <- starwars %>% filter(!is.na(species)) %>% group_by(species)  %>% summarise(length=mean(nchar(name)))
knitr::kable(sr, "pipe")
```

species	**length**
Aleena	13.000000
Besalisk	15.000000
Cerean	12.000000
Chagrian	10.000000
Clawdite	10.000000
Droid	4.833333
Dug	7.000000
Ewok	21.000000
Geonosian	17.000000
Gungan	11.666667
Human	11.285714
Hutt	21.000000
Iktotchi	11.000000
Kaleesh	8.000000
Kaminoan	7.000000
Kel Dor	8.000000
Mirialan	14.000000
Mon Calamari	6.000000
Muun	8.000000
Nautolan	9.000000
Neimodian	11.000000
Pau’an	10.000000
Quermian	11.000000
Rodian	6.000000
Skakoan	10.000000
Sullustan	9.000000
Tholothian	10.000000
Togruta	8.000000
Toong	14.000000
Toydarian	5.000000
Trandoshan	5.000000
Twi’lek	11.000000
Vulptereen	8.000000
Wookiee	8.000000
Xexto	7.000000
Yoda’s species	4.000000
Zabrak	9.500000