```
library(dplyr)

rladies_global %>%
  filter(city == 'Bucharest')
```

EXPLORE YOUR DATA IN R USING dplyr

26.11.2019

Maria Romanescu

Studentă Master

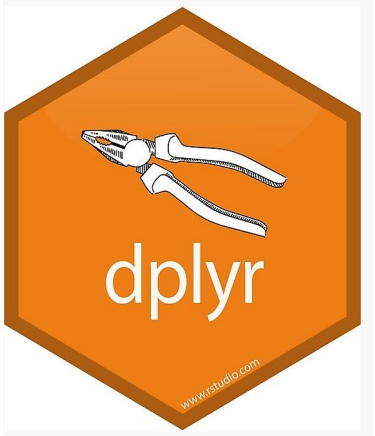
Statistică Aplicată și Data Science, ASE

Andra Garoi

Studentă Master

Statistică Aplicată și Data Science, ASE

Ce este dplyr?



- un pachet din sistemul **tidyverse**
- un pachet specific etapei de manipulare și curățare a datelor,
- include funcții de selectare, filtrare , ordonare, sumarizare și creare de noi variabile.

Cum începem cu dplyr

#Începem să folosim R

RStudio, VIM, Sublime,
Revolution R, etc...

```
# Instalăm pachetul complet tidyverse:
install.packages("tidyverse")

# Alternativ, instalăm doar dplyr:
install.packages("dplyr")

# Pentru utilizare, încărcăm pachetul în memorie
library(dplyr)
```

we.tl/t-2xwvEfRyP0

Women in workforce

Setul de date

Obiecte din memorie

jobs_gender.csv
(2088 rânduri, 12 coloane)

year
occupation
major_category
minor_category
workers_male
workers_female
total_earnings_male
total_earnings_female

Earnings_female.csv
(264 rânduri, 3 coloane)

year
group
percent

employed_gender.csv
(64 rânduri, 7 coloane)

year
full_time_female
part_time_female
full_time_male
part_time_male

Glimpse()

glimpse(jobs_gender)

```
Observations: 2,088
Variables: 12
$ year                  <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 20...
$ occupation            <chr> "Chief executives", "General and operations managers", "...
$ major_category        <chr> "Management, Business, and Financial", "Management, Busi...
$ minor_category        <chr> "Management", "Management", "Management", "Management", ...
$ total_workers         <dbl> 1024259, 977284, 14815, 43015, 754514, 44198, 109703, 48...
$ workers_male          <dbl> 782400, 681627, 8375, 17775, 440078, 16141, 72873, 35436...
$ workers_female        <dbl> 241859, 295657, 6440, 25240, 314436, 28057, 36830, 13467...
$ percent_female        <dbl> 23.6, 30.3, 43.5, 58.7, 41.7, 63.5, 33.6, 27.5, 53.5, 76...
$ total_earnings        <dbl> 120254, 73557, 67155, 61371, 78455, 74114, 62187, 99167,...
$ total_earnings_male   <dbl> 126142, 81041, 71530, 75190, 91998, 90071, 66579, 101318...
$ total_earnings_female <dbl> 95921, 60759, 65325, 55860, 65040, 66052, 55079, 90940, ...
$ wage_percent_of_male  <dbl> 76.04208, 74.97316, 91.32532, 74.29179, 70.69719, 73.333...
```

* Afișarea se face în funcție de mărimea consolei

Select()

select(tabel,col1,col2, …)

select(jobs_gender, year, occupation, total_earnings_female)

```
# A tibble: 2,088 x 3
    year occupation                                 total_earnings_female
   <dbl> <chr>                                                      <dbl>
 1  2013 Chief executives                                          95921
 2  2013 General and operations managers                           60759
 3  2013 Legislators                                               65325
 4  2013 Advertising and promotions managers                       55860
 5  2013 Marketing and sales managers                              65040
 6  2013 Public relations and fundraising managers                 66052
 7  2013 Administrative services managers                          55079
 8  2013 Computer and information systems managers                 90940
 9  2013 Financial managers                                        57406
10  2013 Compensation and benefits managers                        68207
# ... with 2,078 more rows
```

- Este folosit pentru selecția **coloanelor** menționate

Filter()

filter(tabel, numecol==”nume observatie”)

```
filter(jobs_gender, occupation=="Cost estimators")
```

```
# A tibble: 4 x 12
   year occupation major_category minor_category total_workers workers_male workers_female
  <dbl> <chr>      <chr>          <chr>                  <dbl>        <dbl>          <dbl>
1  2013 Cost esti~ Management, B~ Business and ~        105744        91484          14260
2  2014 Cost esti~ Management, B~ Business and ~        115808       103325          12483
3  2015 Cost esti~ Management, B~ Business and ~        117961       102374          15587
4  2016 Cost esti~ Management, B~ Business and ~        122804       105522          17282
# ... with 5 more variables: percent_female <dbl>, total_earnings <dbl>,
#   total_earnings_male <dbl>, total_earnings_female <dbl>, wage_percent_of_male <dbl>
```

- Returnează datele în funcție de filtrul adaugat
- Se folosește pe **randuri**

Arrange()

arrange(tabel,col1); arrange(tabel,desc(col2))

arrange(jobs_gender, desc(total_earnings_female))

```
# A tibble: 2,088 x 6
    year occupation          major_category     workers_male workers_female total_earnings_~
   <dbl> <chr>               <chr>                     <dbl>          <dbl>            <dbl>
 1  2016 Physicians and su~ Healthcare Practi~       489748         253635           166388
 2  2016 Derrick, rotary d~ Natural Resources~        15545             90           158929
 3  2016 Nurse anesthetists Healthcare Practi~        10941          15312           151667
 4  2015 Physicians and su~ Healthcare Practi~       477655         237706           150975
 5  2014 Physicians and su~ Healthcare Practi~       461288         225539           150053
 6  2015 Nurse anesthetists Healthcare Practi~        10381          12452           148873
 7  2014 Nurse anesthetists Healthcare Practi~         8875          11151           142372
 8  2013 Nurse anesthetists Healthcare Practi~         8259          13405           142185
 9  2013 Physicians and su~ Healthcare Practi~       495061         242485           140036
10  2015 Architectural and~ Management, Busin~       130504          12188           131780
# ... with 2,078 more rows
```

- Ordonează setul de date crescator sau descrescator, în funcție de o coloana anume

Summarise()

summarise(tabel, x=fct(var1), y=fct(var2), ...)

```
summarise(jobs_gender,
medie_femei_angajate=mean(workers_female),
medie_barbati_angajati=mean(workers_male))
```

```
# A tibble: 1 x 2
  medie_femei_angajate medie_barbati_angajati
                 <dbl>                  <dbl>
1                84540.                111515.
```

- Reduce observațiile la un singur rând

Mutate()

mutate(tabel, coloana noua=var1*/-+var2)

mutate(jobs_gender, diferenta = total_earnings_male - total_earnings_female)

```
# A tibble: 2,088 x 5
    year occupation                                 total_earnings_male total_earnings_female diferenta
   <dbl> <chr>                                                    <dbl>                 <dbl>     <dbl>
 1  2013 Chief executives                                       126142                 95921     30221
 2  2013 General and operations managers                         81041                 60759     20282
 3  2013 Legislators                                             71530                 65325      6205
 4  2013 Advertising and promotions managers                     75190                 55860     19330
 5  2013 Marketing and sales managers                            91998                 65040     26958
 6  2013 Public relations and fundraising managers               90071                 66052     24019
 7  2013 Administrative services managers                        66579                 55079     11500
 8  2013 Computer and information systems managers              101318                 90940     10378
 9  2013 Financial managers                                      90278                 57406     32872
10  2013 Compensation and benefits managers                      97552                 68207     29345
# ... with 2,078 more rows
```

- Creeaza variabile noi într-un tabel

Group_by()

group_by(tabel, var1, var2, ...)

group_by(jobs_gender, year)

```
# A tibble: 2,088 x 12
# Groups:   year [4]
    year occupation major_category minor_category total_workers
   <dbl> <chr>      <chr>          <chr>                  <dbl>
 1  2013 Chief exe~ Management, B~ Management           1024259
 2  2013 General a~ Management, B~ Management            977284
 3  2013 Legislato~ Management, B~ Management             14815
 4  2013 Advertisi~ Management, B~ Management             43015
 5  2013 Marketing~ Management, B~ Management            754514
 6  2013 Public re~ Management, B~ Management             44198
 7  2013 Administr~ Management, B~ Management            109703
 8  2013 Computer ~ Management, B~ Management            489048
 9  2013 Financial~ Management, B~ Management            990611
10  2013 Compensat~ Management, B~ Management             14656
# ... with 2,078 more rows, and 7 more variables:
#   workers_male <dbl>, workers_female <dbl>, percent_female <dbl>,
#   total_earnings <dbl>, total_earnings_male <dbl>,
#   total_earnings_female <dbl>, wage_percent_of_male <dbl>
```

- Grupează datele după una sau mai multe variabile

%>%
operatorul “pipe”

```
select(jobs_gender, total_earnings_female)

jobs_gender %>% # din tabelul jobs_gender
  select(total_earnings_female) # selectam coloana total_earnings_female
```

- Facilitează rularea mai multor funcții
- Creează o legatură între funcții

O legătură între group_by()+filter()+summarise()

%>%

```
group_by(jobs_gender, major_category)
filter(jobs_gender, !is.na(total_earnings_female))
summarise(jobs_gender, media_veniturilor_fem=mean(total_earnings_female))

jobs_gender %>% #in tabelul jobs_gender
  group_by(major_category) %>% #grupam după categorii majore
   filter(!is.na(total_earnings_female)) %>% #selectam doar randurile fără N/A
   summarise(media_veniturilor_fem=mean(total_earnings_female)) %>%
   arrange(desc(media_veniturilor_fem)) # calc media veniturilor pe categ majore
   # ordonam rezultatele descrescător după medie venituri femei
```

Care este media veniturilor femeilor pe categorii mari?

Output: group_by()+filter()+summarise()+arrange()

```
# A tibble: 8 x 2
  major_category                                media_veniturilor_fem
  <chr>                                                         <dbl>
1 Computer, Engineering, and Science                           69427.
2 Healthcare Practitioners and Technical                       68887.
3 Management, Business, and Financial                          59070.
4 Education, Legal, Community Service, Arts, and Media         46258.
5 Natural Resources, Construction, and Maintenance             38549.
6 Sales and Office                                             37106.
7 Production, Transportation, and Material Moving              32438.
8 Service                                                      31988.
```

inner_join()

inner_join(tabel1, tabel2, by=”nume coloana comuna”)

```
jobs_gender %>%
  group_by(year) %>%
  summarise(medie_angajati_total=mean(total_workers),
                femei=  mean(workers_female), barbati=mean(workers_male))
```

```
# A tibble: 4 x 4
   year medie_angajati_total  femei barbati
  <dbl>                <dbl>  <dbl>   <dbl>
1  2013              189364. 81723. 107640.
2  2014              193872. 83428. 110445.
3  2015              198685. 85578. 113107.
4  2016              202299  87429. 114870.
```

- Unește două sau mai multe tabele în funcție de o coloană comună, adăugând variabilele din ambele tabele, însă doar pentru observațiile comune

inner_join()

```
employed_gender %>%
  select(year, full_time_female, part_time_female) %>%
  arrange(desc(year))
```

```
# A tibble: 49 x 3
    year full_time_female part_time_female
   <dbl>            <dbl>            <dbl>
 1  2016             75.1             24.9
 2  2015             74.8             25.2
 3  2014             74.2             25.8
 4  2013             74               26
 5  2012             73.7             26.3
 6  2011             73.5             26.5
 7  2010             73.4             26.6
 8  2009             73.5             26.5
 9  2008             75.4             24.6
10  2007             75.3             24.7
# ... with 39 more rows
```

inner_join()

inner_join(employed_gender, jobs_gender, by="year")

```
jobs_gender %>% group_by(year) %>%
  summarise(medie_angajati_total=mean(total_workers),
                femei=mean(workers_female), barbati=mean(workers_male)) %>%
  inner_join(employed_gender, jobs_gender, by="year") %>%
  select(year, medie_angajati_total, femei, full_time_female, part_time_female)
```

```
# A tibble: 4 x 5
   year medie_angajati_total  femei full_time_female part_time_female
  <dbl>                <dbl>  <dbl>            <dbl>            <dbl>
1  2013              189364. 81723.             74               26
2  2014              193872. 83428.             74.2             25.8
3  2015              198685. 85578.             74.8             25.2
4  2016              202299  87429.             75.1             24.9
```

9 verbe de reținut despre dplyr

- Glimpse()
- Select()
- Filter()
- Arrange()
- Summarise()
- Mutate()
- Group_by()
- %<%
- Inner_join()

Mai multe continuați independent: https://dplyr.tidyverse.org/
Sau împreună la un alt R-Ladies Bucharest meetup:-)

VĂ MULȚUMIM