AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

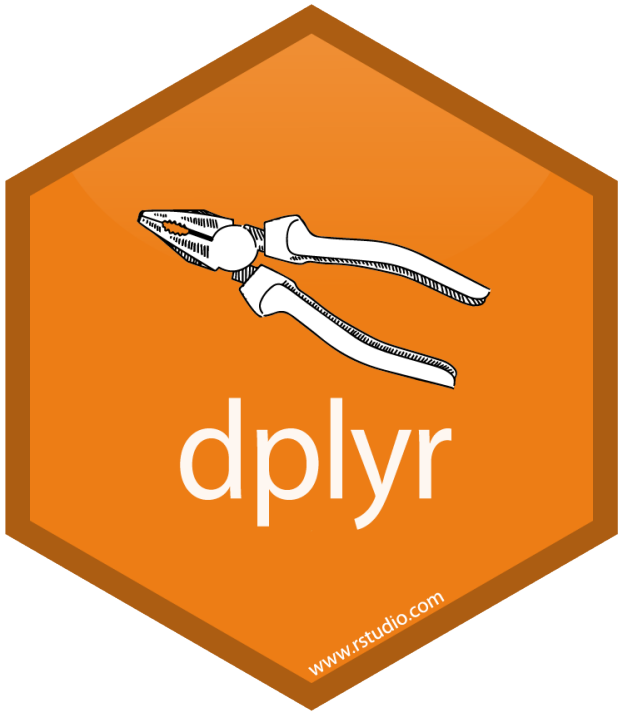
Biblioteka dplyr

proste operacje na zbiorach danych

Kornelia Moskal
Kraków, 12.01.2018

Czym jest dplyr?

dplyr – biblioteka działająca w języku R, wchodząca w skład pakietu tidyverse. Ułatwia ona manipulację danymi poprzez wykorzystanie poleceń analogicznych do zapytań SQL. Biblioteka działa na zasadach wolnego oprogramowania



Źródło: www.tidyverse.org

Dane wejściowe

źródło: System Monitorowania jakości powietrza
okres: 11.2017
miejsce: Zakopane
ilość wierszy: 30
ilość kolumn: 10

	CZAS	SO2	NO2	NOx	NO	O3	O3_8h	CO	CO_8h	PM10
1	1.11	7,2	9	10	1	56	61	169	362	12
2	2.11	8,6	19	37	12	39	54	293	543	22
3	3.11	6,8	21	27	4	37	54	305	591	22
4	4.11	13,8	31	89	38	18	47	925	1889	61
5	5.11	10,3	17	27	6	46	82	420	1967	29
6	6.11	12,9	26	46	13	31	67	591	994	46
7	7.11	7,3	14	17	2	39	54	329	1077	42
8	8.11	15,4	28	71	28	19	41	870	1654	84
9	9.11	16,2	34	75	27	8	12	849	2112	74
10	10.11	9,9	24	45	14	31	43	549	1150	40
11	11.11	9,9	23	30	5	43	63	463	1135	32
12	12.11	10,2	14	16	1	42	48	389	1025	21
13	13.11	10,6	22	30	5	28	42	482	549	20
14	14.11	12,8	27	37	7	32	47	593	985	35
15	15.11	19,6	43	80	24	19	39	922	1290	54
16	16.11	19,2	45	121	49	16	35	1078	1736	66
17	17.11	19	44	95	33	3	4	1185	1669	76
18	18.11	14,5	28	42	10	20	42	701	1530	41
19	19.11	8,3	13	15	2	46	51	312	404	18
20	20.11	7,7	13	16	2	55	63	289	464	19
21	21.11	11,4	21	31	6	28	43	442	568	28
22	22.11	12,1	26	50	15	15	22	557	684	32
23	23.11	22,2	49	157	71	5	10	1076	1521	68
24	24.11	22,7	54	153	64	10	20	1158	1900	70
25	25.11	13,5	25	41	11	37	76	552	1807	33
26	26.11	8,4	12	14	1	43	45	327	509	15
27	27.11	15,9	31	47	10	32	46	683	1199	44
28	28.11	33,3	60	153	60	12	20	1603	2712	119
29	29.11	30,5	40	98	38	17	38	1439	2870	111
30	30.11	23,4	35	50	10	15	24	1045	2105	57

Filtrowanie i selekcja

Filter

```
> filtr=filter(dane,PM10>50,CO<1600)
> filtr
    CZAS  SO2 NO2 NOx NO O3 O3_8h   CO CO_8h PM10
1   4.11 13,8  31  89 38 18    47  925  1889   61
2   8.11 15,4  28  71 28 19    41  870  1654   84
3   9.11 16,2  34  75 27  8    12  849  2112   74
4  15.11 19,6  43  80 24 19    39  922  1290   54
5  16.11 19,2  45 121 49 16    35 1078  1736   66
6  17.11   19  44  95 33  3     4 1185  1669   76
7  23.11 22,2  49 157 71  5    10 1076  1521   68
8  24.11 22,7  54 153 64 10    20 1158  1900   70
9  29.11 30,5  40  98 38 17    38 1439  2870  111
10 30.11 23,4  35  50 10 15    24 1045  2105   57
```

Select

```
> selekcja = select(filtr,CZAS,SO2,NO)
> selekcja
    CZAS  SO2 NO
1   4.11 13,8 38
2   8.11 15,4 28
3   9.11 16,2 27
4  15.11 19,6 24
5  16.11 19,2 49
6  17.11   19 33
7  23.11 22,2 71
8  24.11 22,7 64
9  29.11 30,5 38
10 30.11 23,4 10
```

Porządkowanie informacji

Arrange

```
> a = arrange(selekcja,NO)
> a
     CZAS  SO2 NO
1   30.11 23,4 10
2   15.11 19,6 24
3    9.11 16,2 27
4    8.11 15,4 28
5   17.11   19 33
6    4.11 13,8 38
7   29.11 30,5 38
8   16.11 19,2 49
9   24.11 22,7 64
10 23.11 22,2 71
```

Group_by

```
> gr = group_by(selekcja,NO)
> gr
# A tibble: 10 x 3
# Groups:   NO [9]
     CZAS    SO2    NO
    <dbl> <fctr> <int>
 1   4.11   13,8    38
 2   8.11   15,4    28
 3   9.11   16,2    27
 4 15.11   19,6    24
 5 16.11   19,2    49
 6 17.11     19    33
 7 23.11   22,2    71
 8 24.11   22,7    64
 9 29.11   30,5    38
10 30.11   23,4    10
```

Dodawanie nowych danych

Mutate

```
> mut = mutate(gr,prop = NO/10)
> mut
# A tibble: 10 x 4
# Groups:   NO [9]
     CZAS    SO2     NO  prop
    <dbl> <fctr> <int> <dbl>
 1  4.11   13,8     38   3.8
 2  8.11   15,4     28   2.8
 3  9.11   16,2     27   2.7
 4 15.11   19,6     24   2.4
 5 16.11   19,2     49   4.9
 6 17.11     19     33   3.3
 7 23.11   22,2     71   7.1
 8 24.11   22,7     64   6.4
 9 29.11   30,5     38   3.8
10 30.11   23,4     10   1.0
```

Transmute

```
> tr = transmute(gr,prop = NO/10)
Adding missing grouping variables: `NO`
> tr
# A tibble: 10 x 2
# Groups:   NO [9]
      NO  prop
   <int> <dbl>
 1    38   3.8
 2    28   2.8
 3    27   2.7
 4    24   2.4
 5    49   4.9
 6    33   3.3
 7    71   7.1
 8    64   6.4
 9    38   3.8
10    10   1.0
```

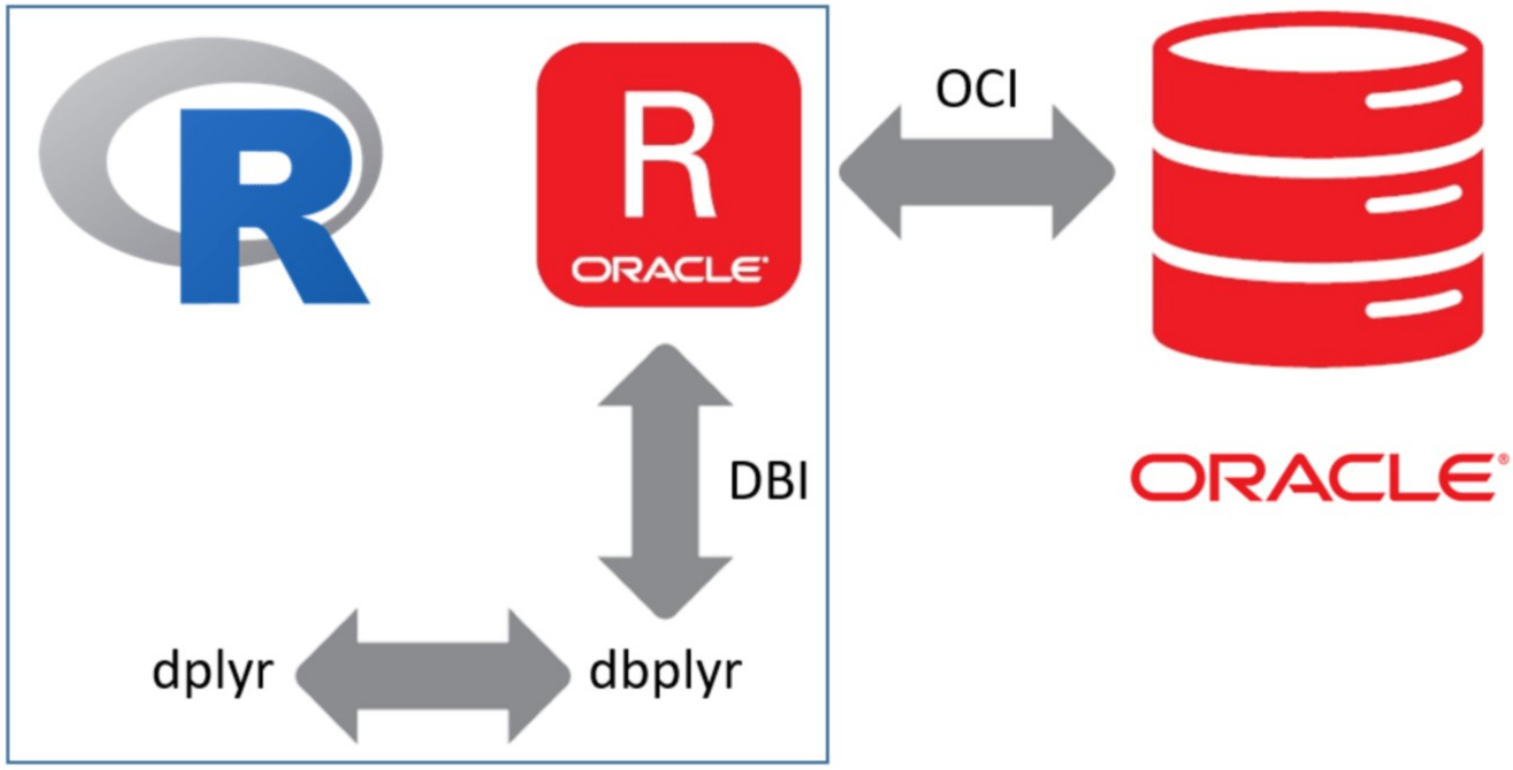
Łączenie tabel - join

```
> s1
    CZAS   CO
1   4.11  925
2   8.11  870
3   9.11  849
4  15.11  922
5  16.11 1078
6  17.11 1185
7  23.11 1076
8  24.11 1158
9  29.11 1439
10 30.11 1045
```

```
> s2
    CZAS NO
1   4.11 38
2   8.11 28
3   9.11 27
4  15.11 24
5  16.11 49
6  17.11 33
7  23.11 71
8  24.11 64
9  29.11 38
10 30.11 10
```

```
> laczenie = inner_join(s1,s2)
Joining, by = "CZAS"
> laczenie
    CZAS   CO NO
1   4.11  925 38
2   8.11  870 28
3   9.11  849 27
4  15.11  922 24
5  16.11 1078 49
6  17.11 1185 33
7  23.11 1076 71
8  24.11 1158 64
9  29.11 1439 38
10 30.11 1045 10
```

Współpraca z biblioteką dbplyr



Źródło: www. technology.amis.nl

Zalety dplyr

- Posiada proste i instynktowne funkcje, a jego wykorzystanie sprawia że kod pozostaje przejrzysty.
- Jest bardzo wydajny co skraca czas oczekiwania na odpowiedź komputera
- Pozwala na wykorzystanie funkcjonalności, zbliżonej do zapytań SQL, bez konieczności tworzenia bazy danych
- Znacznie usprawnia manipulowanie danymi

Dziękuję za uwagę