Machine Learning - 1100-ML0ENG (Ćwiczenia informatyczne Z-23/24)

The dplyr package

It is simply the most useful package in R for data manipulation. One of the greatest advantages of this package is you can use the pipe function **%>%** to combine different functions in R. From filtering to grouping the data, this package does it all.

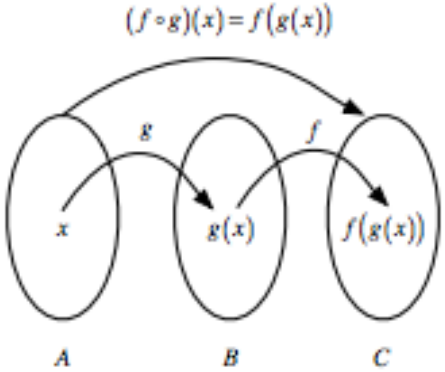
Here is the list of functions **dplyr**:

- **select()**: Select columns from your dataset;
- **filter()**: Filter out certain rows that meet your criteria(s);
- **group_by()**: Group different observations together such that the original dataset does not change. Only the way it is represented is changed in the form of a list;
- **summarise()**: Summarise any of the above functions;
- **mutate()**: Create new columns by preserving the existing variables

```
install.packages("dplyr")
library(dplyr)
starwars
select(starwars, species, name, height, mass, homeworld)
filter(starwars, species == "Human")
filter(starwars, mass > 70)
filter(starwars, hair_color == "none" & eye_color == "black")
```

Pipe %>% or new version |>

Pipes are structures that allow you to take the output of one function and send it directly to the next. Most often, they are used when you use multiple functions on a single dataset.




```
pi %>% sin %>% cos
cos(sin(pi))
```

examples

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

```
sw.droid<-starwars %>%
 filter(species == "Droid") %>%
 select(name,height, mass, homeworld)
sw.droid
```

```
t4<-read.csv("http://imul.math.uni.lodz.pl/~bartkiew/ml/data/titanic.csv")
str(t4)
summary(t4)
t4%>%
 filter(Sex=="male" & Survived ==1)%>%
 select(Name,Age)%>%
 head()
```

mutate

```
x = mutate(starwars, h.inch = height / 2.5) #1 inch=2.5 cm
```

```
starwars %>%
 mutate(h.inch = height / 2.5) %>%
 head()
```

```
starwars %>%
 mutate(h.inch = height / 2.5) %>%
 select(name, h.inch, homeworld, mass) %>%
 filter(!is.na(mass)) %>%
 head(n = 10)
```

summarise and group_by

Many data analysis tasks can be approached using the “split-apply-combine” paradigm: split the data into groups, apply some analysis to each group, and then combine the results.

In dplyr package we can use the **group_by()** function, which splits the data into groups and the **summarise()** function to produce a summary statistic.

summarise() and **summarize()** are synonyms.

Useful functions:

- Center: `mean(), median()`

- Spread: `sd(), IQR(), mad()`
- Range: `min(), max(), quantile()`
- Position: `first(), last(), nth()`
- Count: `n(), n_distinct()`
- Logical: `any(), all()`

```
starwars %>%
 group_by(species) %>%
 summarise(
 number = n(),
 w = mean(mass, na.rm = TRUE)
 )
```

airquality ex

```
summary(airquality)
airquality
str(airquality)
airquality%>%
 filter(Month==5)%>%
 mutate(temp.C=(Temp-32)/2)
```

```
airquality%>%
 mutate(temp.C=(Temp-32)/2)%>%
 select(Ozone,Solar.R,Temp, temp.C)%>%
 filter(Month==5)
```

```
airquality%>%
 group_by(Month)%>%
 summarise(
 num.obser=n(),
 temp.min=min(Temp),
 ozon.average=mean(Ozone, na.rm=T)
 )
airquality%>%
 mutate(temp.C=(Temp-32)/2)%>%
 group_by(Month)%>%
 summarise(
 num.obser=n(),
 temp.min=min(Temp),
 temp.min.C=min(temp.C),
 temp.max.C=max(temp.C),
 ozon.average=mean(Ozone, na.rm=T),
 ozon.sd=sd(Ozone, na.rm=T)
 )
```

Last modified: sobota, 29 października 2022, 1:29

Accessibility settings

Przetwarzanie danych osobowych

Platformą administruje Komisja ds. Doskonalenia Dydaktyki wraz z Centrum Informatyki Uniwersytetu Łódzkiego Więcej

Informacje na temat logowania

Na platformie jest wykorzystywana metoda logowania za pośrednictwem Centralnego Systemu Logowania.

Studentów i pracowników Uniwersytetu Łódzkiego obowiązuje nazwa użytkownika i hasło wykorzystywane podczas logowania się do systemu USOSweb.

Deklaracja dostępności