Podstawy statystyki matematycznej w programie R

Piotr Ćwiakowski
Wydział Fizyki Uniwersytetu Warszawskiego

1 marca 2017 r.

Program zajęć

Wprowadzenie do R i badań statystycznych

- podstawowe operacje w R
- eksploracyjna analiza danych
- podstawowe metody wizualizacji

Statystyka opisowa

- miary tendencji centralnej
- miary rozproszenia
- histogram
- wykres pudełkowy
- wykres słupkowy
- rozkład empiryczny zmiennej
- rozkłady statystyczne zmiennej ciągłej i dyskretnej

Elementy rachunku prawdopodobieństwa

- dystrybuanta i funkcja gęstości
- Prawo Wielkich Liczb
- Centralne Twierdzenie Graniczne
- schematy losowań próby losowej – losowanie proste, warstwowe, grupowe
- metodologia badań statystycznych – wybrane zagadnienia

Program zajęć (2)

Budowa testu statystycznego:

- błąd standardowy (pojęcie i oszacowanie),
- budowa hipotez statystycznych, weryfikacja hipotez – błędy I i II rodzaju,
- określanie i znaczenie poziomu istotności, konstrukcja statystyki testowej
- weryfikacja i interpretacja wyniku testu statystycznego
- przedział ufności – budowa i interpretacja
- badanie mocy testu, określanie wielkości próby do badania

Przegląd testów parametrycznych

- testy średniej w jednej próbie
- test wariancji w jednej próbie
- testy równości średnich w dwóch próbach,
- testy homogeniczności wariancji w dwóch próbach

Badanie normalności rozkładu

- znaczenie założenia o normalności rozkładu,
- testowanie hipotezy o normalności rozkładu (przegląd testów)

Przegląd testów nieparametrycznych

- test znaków
- test Manna-Witneya
- testy Wicoxona
- testy odsetka

Program zajęć (3)

Analiza korelacji

- korelacja Pearsona
- korelacja Spearmana
- korelacja Tau Kendalla
- korelogram – wizualizacja macierzy korelacji
- testy parametryczne i nieparametryczne istotności korelacji

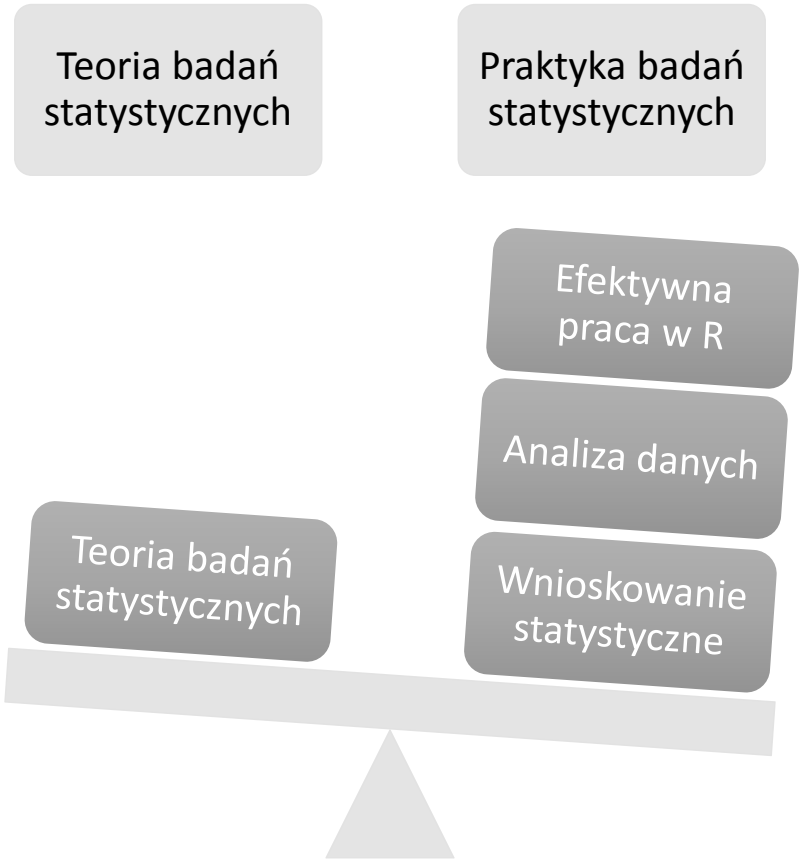
Analiza tablicy kontyngencji

- test zgodności i niezależności chi-kwadrat,
- poprawka Yatesa
- test Fishera
- statystyka V-Cramera
- współczynnik Phi
- wizualizacja tablicy kontyngencji – (np. wykres mozaikowy, *balloon plot*)

Wstęp do regresji liniowej

- teoria budowy modelu regresji liniowej
- algorytm wyznaczania parametrów
- regresja z jedną zmienną
- regresja wielu zmiennych
- oszacowanie i interpretacja wyników

Charakter kursu



Terminy i tematyka spotkań

Lp.	Data	Temat zajęć
1	1 marca	Wprowadzenie do programu
2	8 marca	Przetwarzanie danych w programie R
3	15 marca	Wizualizacja danych w R
4	22 marca	Statystyka opisowa, wprowadzenie do badań statystycznych
5	29 marca	Elementy rachunku prawdopodobieństwa
6	5 kwietnia	Budowa testu statystycznego testy parametryczne
7	12 kwietnia	Testy parametryczne - ćwiczenia

Terminy i tematyka spotkań

Lp.	Data	Temat zajęć
8	19 kwietnia	Testy na normalność rozkładu, testy nieparametryczne
9	26 kwietnia	Moc testu statystycznego, wyznaczanie optymalnej wielkości próby
10	10 maja	Model ANOVA, testy post-hoc
11	17 maja	Analiza korelacji
12	24 maja	Analiza tablicy kontyngencji
13	31 maja	Analiza regresji liniowej cz. 1
14	7 czerwca	Analiza regresji liniowej cz. 2
15	14 czerwca	Wprowadzenie do statystyki bayesowskiej

Zaliczenie kursu

- obecność (możliwe trzy nieobecności nieusprawiedliwione)
- praca na zajęciach
  - Za aktywność dodatkowe punkty
- prace domowe
  - Prace domowe są po każdych zajęciach
  - Łączna liczba punktów z prac domowych wynosi 100 pkt. (100% zaliczenia)
  - Prac domowych będzie 14 (różnie punktowane)
- praca dodatkowe – pojawiają się nieregularnie, można zdobyć dodatkowe punkty

Materiały

- **Materiały dostępne są:**
  - na platformie edukacyjnej Moodle (mikroekonomia.wne.uw.edu.pl)

- **W skład materiałów wchodzą:**
  - Prezentacja (pdf)
  - Kod R-owy z komentarzem (rozszerzenie *.R)
  - Bazy danych
  - Treści zadań (pdf)
  - Rozwiązania zadań (dostępne po zajęciach)

Organizacja pracy na zajęciach

Tryb pracy na zajęciach

- wykład (część teoretyczna zajęć),
- ćwiczenia (praktyczna implementacja w programie R)
- warsztat (praca samodzielna pod kierunkiem prowadzącego),
- praca samodzielna w domu (powtarzanie materiału z zajęć)

Kontakt mailowy

- W sprawach organizacyjnych oraz merytorycznych
- W czasie trwania kursu oraz po jego zakończeniu
- Adres: pcwiakowski@wne.uw.edu.pl

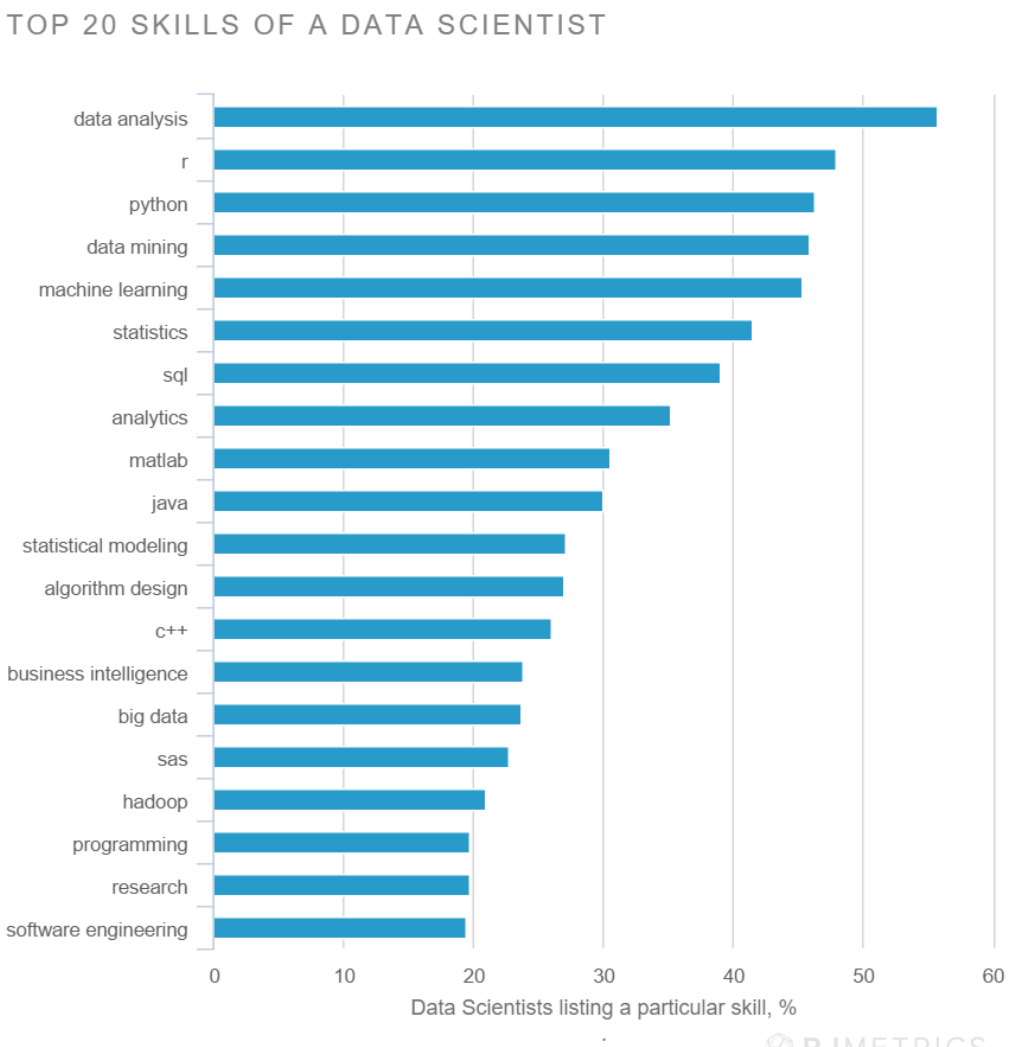
Środowisko pracy

- Program R
- Nakładka R Studio

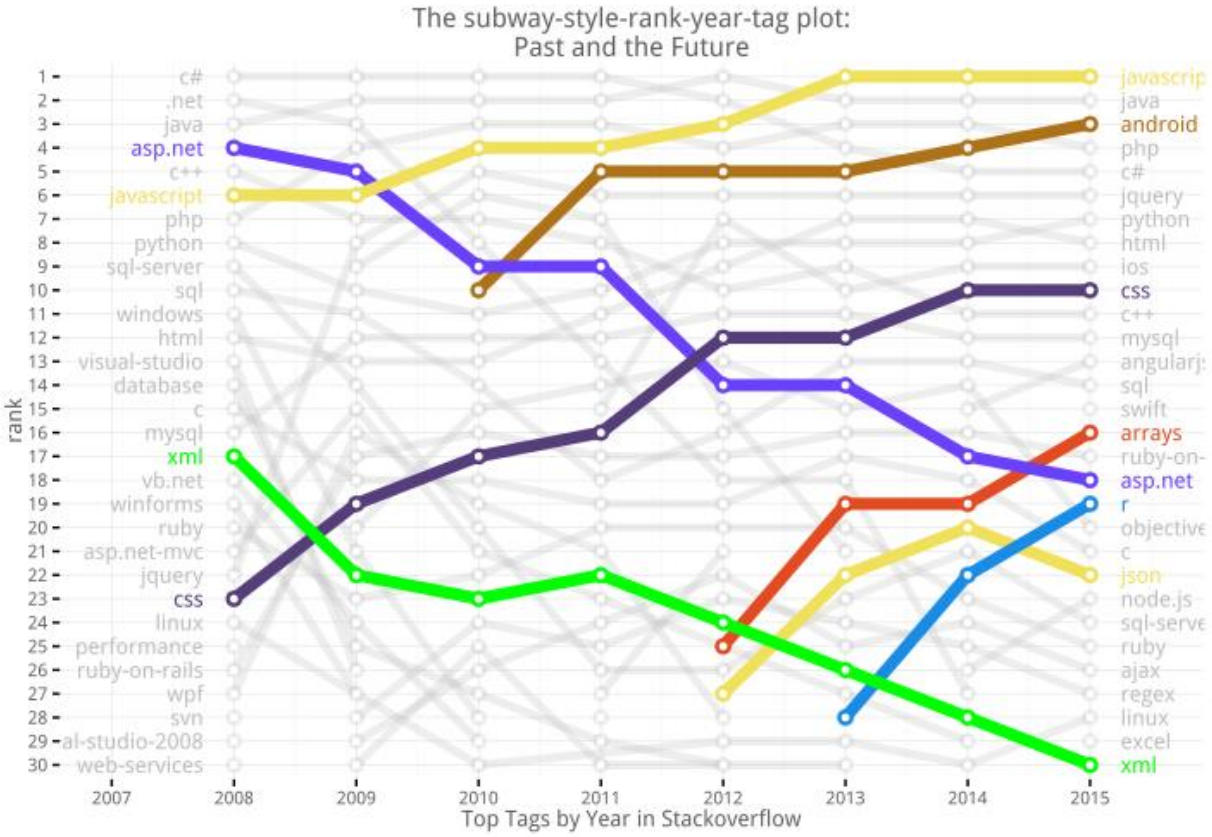
O programie R

- Licencja freeware
- Olbrzymia funkcjonalność
- Łatwa rozbudowa pakietu
- Duża społeczność użytkowników
- Dynamiczny rozwój w ostatnich latach
- Podstawowe narzędzie w Data Science
- język funkcyjny i obiektowy jednocześnie
- Bezkonkurencyjny pod względem możliwości graficznych

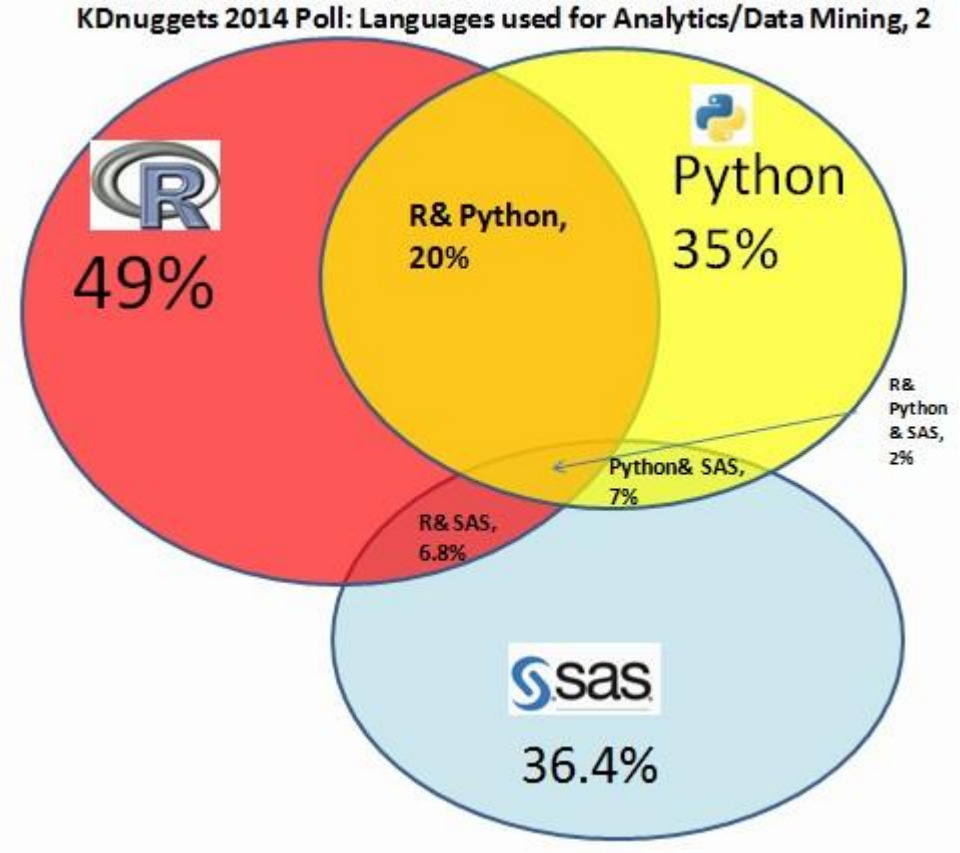
Dlaczego R? (1)


TOP 20 SKILLS OF A DATA SCIENTIST
data analysis
r
python
data mining
machine learning
statistics
sql
analytics
matlab
java
statistical modeling
algorithm design
c++
business intelligence
big data
sas
hadoop
programming
research
software engineering
0
10
20
30
40
50
60
Data Scientists listing a particular skill, %
RJMETRICS

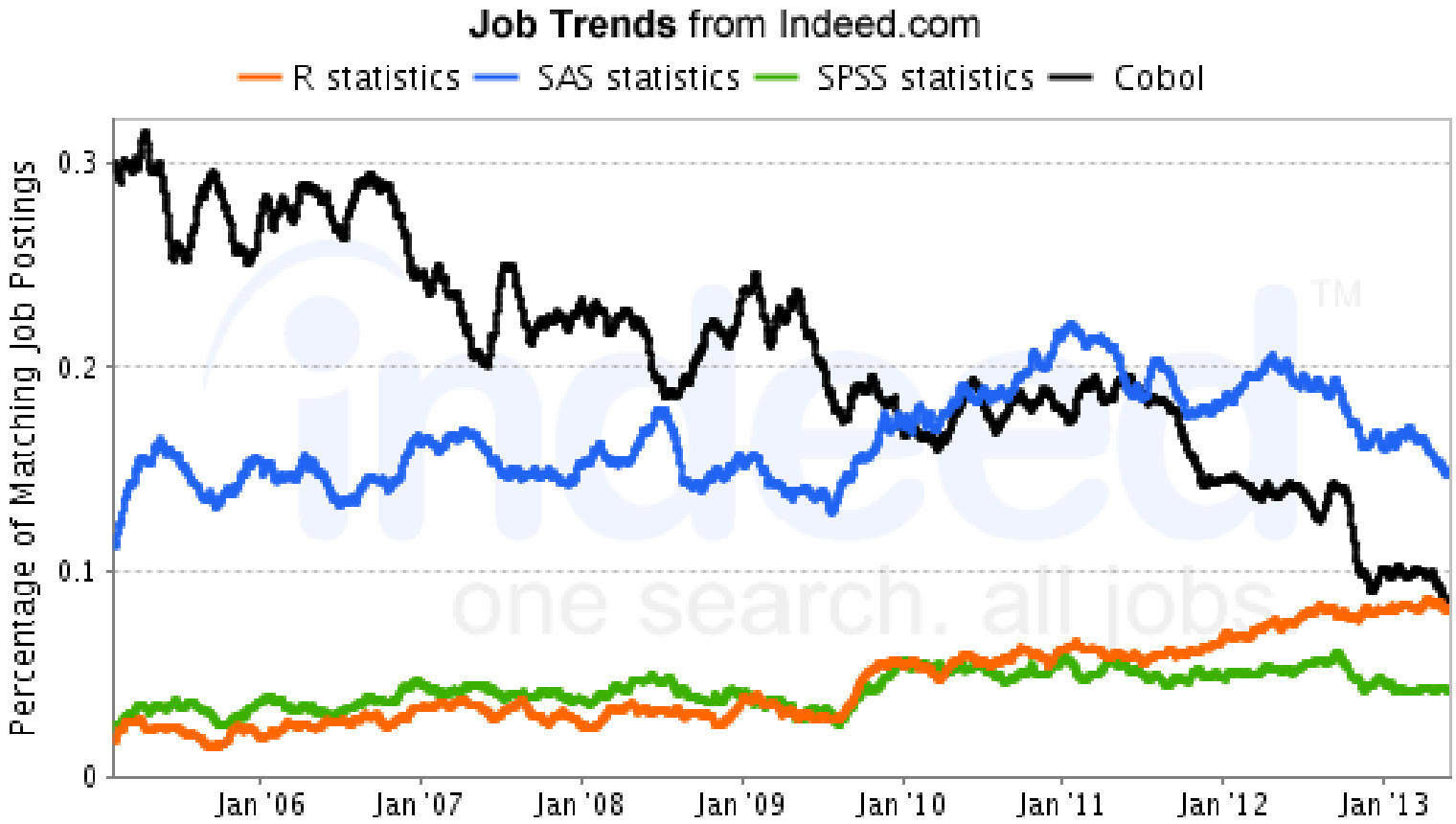
Dlaczego R? (2)


The subway-style-rank-year-tag plot:
Past and the Future
rank
1 c#
2 .net
3 java
4 asp.net
5 c++
6 javascript
7 php
8 python
9 sql-server
10 sql
11 windows
12 html
13 visual-studio
14 database
15 c
16 mysql
17 xml
18 vb.net
19 winforms
20 ruby
21 asp.net-mvc
22 jquery
23 css
24 linux
25 performance
26 ruby-on-rails
27 wpf
28 svn
29 al-studio-2008
30 web-services
javascrip
java
android
php
c#
jquery
python
html
ios
css
c++
mysql
angularj
sql
swift
arrays
ruby-on-
asp.net
r
objective
c
json
node.js
sql-serve
ruby
ajax
regex
linux
excel
xml
2007
2008
2009
2010
2011
2012
2013
2014
2015
Top Tags by Year in Stackoverflow

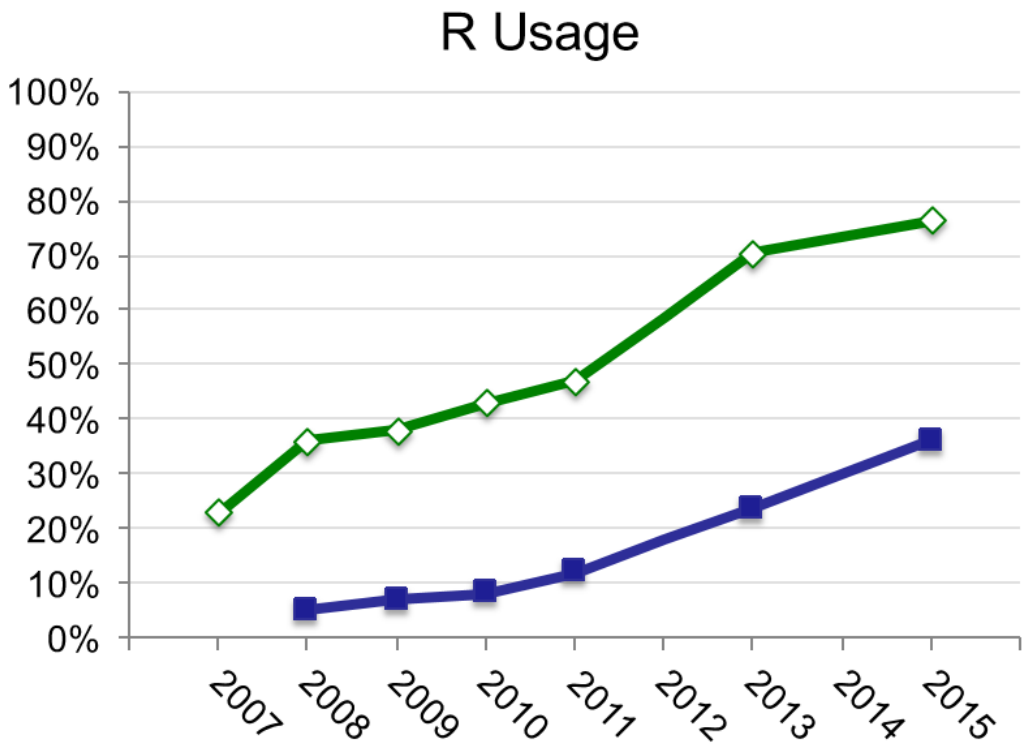
Dlaczego R? (3)


KDnuggets 2014 Poll: Languages used for Analytics/Data Mining, 2
R
49%
R& Python,
20%
Python
35%
R&
Python
& SAS,
2%
Python& SAS,
7%
R& SAS,
6.8%
sas
36.4%

Dlaczego R? (4)


Job Trends from Indeed.com
R statistics
SAS statistics
SPSS statistics
Cobol
Percentage of Matching Job Postings
0.3
0.2
0.1
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
Jan '13
indeed
one search. all jobs.

Dlaczego R? (5)



76% of analytic professionals report using R

36% select R as their primary tool

W chwilach zagubienia…

Search functions, lists, and more

R-Studio

- Licencja freeware
- Możliwość optymalizacji pracy w R (tworzenie projektom etc.)
- Dużo udogodnień:
  - Podpowiedzi w pisaniu kodu i sprawdzanie części składni,
  - Notatnik powiązany z konsolą
- Możliwość przeglądania obiektów
- Możliwość samodzielnego rozbudowania poprzez umieszczanie własnych wtyczek
- R Markdown
- Shiny
- Wiele innych…

Gdzie szukać informacji o R?

Oficjalna strona projektu: cran.r-project.org (do ściągnięcia R)

Oficjalna strona R-Studio: www.rstudio.com (do ściągnięcia R-Studio)

- **Wyszukiwarka R:** rseek.org
- **Wyszukiwarka Google:** zapytanie z dopiskiem „*in R*”

- **Podręczniki do R:** www.statmethods.net/about/books.html
- **Blog użytkowników R:** www.r-bloggers.com
- **Forum użytkowników R (i nie tylko):** stackoverflow.com

R – pierwsze kroki

Podstawowe zasady pracy w R

- Wielkość liter ma znaczenie – R rozróżnia: „A” i „a”
- Poszczególne komendy mogą być rozdzielane albo enterem (kolejna linijka) albo znakiem „0”;
- Komentarze można wprowadzać po znaku # (całą treść po tym znaku R ignoruje do końca linijki).
- W przypadku niekompletnej komendy R informuje znakiem „+”, najczęściej chodzi o:
  - brakujący nawias (lub o jeden za dużo),
  - brakujący cudzysłów (lub o jeden za dużo)
- Przerwanie uzupełniania komendy za pomocą klawisza „Esc”
- Nazwy zmiennych mogą zawierać: litery, cyfry, „_” i „.”
- Polskie znaki – obsługiwane, ale lepiej unikać

Rozróżniamy w R:

- Wyrażenie – obliczone, pokazane i wartość jest ostatecznie nigdzie niezapisana.
- Przypisanie – wartość wyliczona i zapisana, ale nie jest automatycznie pokazana.

Dziękuję za uwagę!