

Spis treści

1. Wprowadzenie	1
1.1. Data science, czyli dlaczego warto poznać R	1
1.2. Jak wygląda praca z programem R	4
1.2.1. Przykład: Pozycja Polski w rankingu FIFA	5
1.3. Jak przygotować środowisko pracy	7
1.3.1. Instalacja podstawowego środowiska R	8
1.3.2. Instalacja edytora RStudio	9
1.3.3. Instalacja dodatkowych pakietów	11
1.4. Gdzie szukać dalszych informacji	13
2. Podstawy pracy z R	16
2.1. Wczytywanie danych	16
2.2. Struktury danych	18
2.2.1. Wektory	18
2.2.2. Ramki danych	20
2.2.3. Listy	22
2.3. Statystyki opisowe	23
2.3.1. Zmienne ilościowe	25
2.3.2. Zmienne jakościowe	26
2.4. Statystyki graficzne	28
2.4.1. Wykres słupkowy	28
2.4.2. Histogram	29
2.4.3. Wykres pudełkowy	31
2.4.4. Jądrowy estymator gęstości	32
2.4.5. Wykres kropkowy	35
2.4.6. Wykres mozaikowy	36
2.5. Jak przetwarzać dane z pakietem dplyr	38
2.5.1. Jak filtrować wiersze	39
2.5.2. Jak wybierać kolumny	40
2.5.3. Jak tworzyć i transformować zmienne	40
2.5.4. Jak sortować wiersze	41
2.5.5. Jak pracować z potokami	42

2.5.6.	Jak wyznaczać agregaty/statystyki w grupach	45
2.5.7.	Postać szeroka/postać wąska	47
2.5.8.	Sklejanie/rozcinanie kolumn	49
2.6.	Jak wczytać i zapisać dane w różnych formatach	50
2.6.1.	Wczytywanie danych z pakietów	50
2.6.2.	Wczytywanie danych z plików tekstowych	51
2.6.3.	Zapisywanie danych do pliku tekstowego	57
2.6.4.	Wczytywanie i zapisywanie z zastosowaniem formatu JSON	58
2.6.5.	Wczytywanie danych w formatach HTML i XML	60
2.6.6.	Inne formaty plików tekstowych	60
2.6.7.	Wczytywanie danych z plików Excela	62
2.6.8.	Wczytywanie danych z SPSS'a	64
2.6.9.	Wczytywanie danych z programu Matlab	65
2.6.10.	Wczytywanie danych z SAS	66
2.6.11.	Inne funkcje do importu danych	66
2.7.	Automatyczne raporty, powtarzalne badania	67
2.7.1.	Pakiet knitr, markdown a raporty w HTML	68
2.7.2.	Prezentacje w HTML5	72
2.7.3.	Pakiet Sweave a raporty w języku \LaTeX	74
3.	Niezbędnik programisty	79
3.1.	Instrukcje sterujące	80
3.1.1.	Jak tworzyć funkcje	80
3.1.2.	Jak stosować instrukcje warunkowe	85
3.1.3.	Jak budować pętle	88
3.2.	Jak pracować z bazami danych	92
3.2.1.	Jak pracować z bazą danych SQLite	93
3.2.2.	Jak pracować z większymi relacyjnymi bazami danych	94
3.2.3.	Jak używać pakietu dplyr w pracy z bazami danych	96
3.3.	Budowa aplikacji WWW z pakietem shiny	98
3.3.1.	Jak wygląda model akcja–reakcja	99
3.3.2.	Jak opisać interfejs użytkownika	99
3.3.3.	Jak opisać przetwarzanie na serwerze	100
3.3.4.	Jak dodawać kontrolki sterujące	102
3.4.	Budowanie własnych pakietów	105
3.4.1.	Niezbędne oprogramowanie	106
3.4.2.	Jak wygląda struktura pakietu	106
3.4.3.	Jak stworzyć nowy pakiet	107
3.4.4.	Plik DESCRIPTION	108
3.4.5.	Jak dodawać funkcje do pakietu	109
3.4.6.	Jak dodawać zbiory danych do pakietu	111

3.4.7.	Jak dodawać testy do pakietu	113
3.4.8.	Jak budować stworzony pakiet	115
3.5.	Git, GitHub i kontrola wersji	117
3.5.1.	Jak kopiować repozytorium – clone	117
3.5.2.	Jak dodawać zmiany – commit	119
3.6.	Debugger	119
3.6.1.	Co mogłoby pójść źle?	119
3.6.2.	Błędy i ostrzeżenia	120
3.6.3.	Co można zrobić post-mortem – funkcja traceback()	120
3.6.4.	Jak zastawić pułapkę – funkcja recover()	121
3.6.5.	Jak śledzić krok po kroku – funkcja debug()	122
3.6.6.	Jak ignorować błędy – funkcja try()	123
3.6.7.	Zaokrąglenia numeryczne – studium przypadku	124
3.7.	Profiler	125
3.7.1.	Jak mierzyć czas działania bloku instrukcji	125
3.7.2.	Jak mierzyć czas co do milisekundy	126
3.7.3.	Jak szukać wąskich gardeł	127
3.7.4.	Jak przedstawiać graficznie wyniki profilowania	130
3.7.5.	Jak zrównoleglać obliczenia	130
3.7.6.	Jak zwiększać wydajność R	135
3.8.	Więcej o obiektach w R	138
3.8.1.	Funkcje polimorficzne i klasy S3	138
3.8.2.	Tworzenie własnych operatorów	140
3.8.3.	Obiekty wywołań funkcji	140
3.8.4.	Leniwa ewaluacja	141
3.8.5.	Zasięg symboli w przestrzeniach nazw	143
3.8.6.	Współdzielona przestrzeń nazw	145
3.8.7.	Obiekty	146
3.8.8.	Klasy S4	148
3.8.9.	Formuły	151
3.9.	Inne przydatne funkcje	153
3.9.1.	Rodzina funkcji *apply	154
3.9.2.	Pakiety plyr i reshape2	157
3.9.3.	Funkcje systemowe	161
3.9.4.	Operacje na plikach i katalogach	162
4.	Niezbędnik statystyka	165
4.1.	Generatory liczb losowych	166
4.1.1.	Wprowadzenie do generatorów liczb pseudolosowych	166
4.1.2.	Popularne rozkłady zmiennych losowych	168
4.1.3.	Wybrane metody generowania zmiennych losowych	175

4.1.4.	Estymacja parametrów rozkładu	186
4.2.	Wstępne przetwarzanie danych	186
4.2.1.	Brakujące obserwacje	187
4.2.2.	Normalizacja, skalowanie i transformacje nieliniowe	191
4.3.	Analiza wariancji, regresja liniowa i logistyczna	195
4.3.1.	Wprowadzenie do analizy wariancji	196
4.3.2.	Analiza jednoczynnikowa	197
4.3.3.	Analiza wieloczynnikowa	206
4.3.4.	Regresja	211
4.3.5.	Wprowadzenie do regresji logistycznej	226
4.4.	Testowanie	243
4.4.1.	Testowanie zgodności	244
4.4.2.	Testowanie hipotezy o równości parametrów położenia	251
4.4.3.	Testowanie hipotezy o równości parametrów skali	256
4.4.4.	Testowanie hipotez dotyczących wskaźnika struktury	258
4.4.5.	Testy istotności zależności pomiędzy dwoma zmiennymi	260
4.4.6.	Testowanie zbioru hipotez	269
4.4.7.	Rozkład statystyki testowej	272
4.5.	Bootstrap	276
4.5.1.	Rozkład i obciążenie estymatorów	277
4.5.2.	Testy bootstrapowe	280
4.6.	Analiza przeżycia	282
4.6.1.	Krzywa przeżycia	282
4.6.2.	Model Coxa	284
4.7.	Wybrane funkcje matematyczne	287
4.7.1.	Operacje na zbiorach	287
4.7.2.	Operacje arytmetyczne	288
4.7.3.	Wielomiany	290
4.7.4.	Bazy wielomianów ortogonalnych	291
4.7.5.	Szukanie maksimum/minimum/zer funkcji	293
4.7.6.	Rachunek różniczkowo-całkowy	294
5.	Graficzna prezentacja danych	296
5.1.	Znajdź siedem różnic	297
5.2.	Jak zapisać wykres do pliku	298
5.3.	Pakiet lattice	301
5.3.1.	Wprowadzenie	301
5.3.2.	Szablony dla wykresów	301
5.3.3.	Formuły i wybór zmiennych	305
5.3.4.	Panele i mechanizm warunkowania	305
5.3.5.	Mechanizm grupowania	306

5.3.6.	Legenda wykresu	306
5.3.7.	Atlas funkcji graficznych z pakietu lattice	308
5.3.8.	Więcej o panelach	318
5.3.9.	Motywy i parametry graficzne	321
5.3.10.	Zaawansowane opcje	321
5.4.	Pakiet ggplot2	326
5.4.1.	Wprowadzenie	326
5.4.2.	Warstwy wykresu	327
5.4.3.	Mapowanie zmiennych na atrybuty wykresu	329
5.4.4.	Geometria warstwy	332
5.4.5.	Statystyki i agregacje	334
5.4.6.	Mechanizm warunkowania	335
5.4.7.	Kontrola skal	337
5.4.8.	Układ współrzędnych i osie wykresu	339
5.4.9.	Motywy i kompozycje graficzne	340
5.4.10.	Pozycjonowanie wykresów na rysunku	341
5.4.11.	Obiekt klasy gg	342
5.5.	Pakiet graphics	344
5.5.1.	Wprowadzenie	344
5.5.2.	Funkcja plot()	347
5.5.3.	Funkcja matplot()	348
5.5.4.	Osie wykresu	348
5.5.5.	Legenda wykresu	349
5.5.6.	Funkcja image()	350
5.5.7.	Wyrażenia matematyczne	351
5.5.8.	Kolory	353
5.5.9.	Właściwości linii	354
5.5.10.	Właściwości punktów/symboli	355
5.5.11.	Atomowe, niskopoziomowe funkcje graficzne	355
5.5.12.	Tytuł, podtytuł i opisy osi wykresu	356
5.5.13.	Pozycjonowanie wykresu, wiele wykresów na rysunku	360
5.5.14.	Wykres słonecznikowy	361
5.5.15.	Wykresy kropkowe, dwu- i wielowymiarowe	362
5.5.16.	Wykres macierzy korelacji	364
5.5.17.	Wykres konturowy	364
5.5.18.	Wykres koniczyny	366
5.5.19.	Wielowymiarowy, jądrowy estymator gęstości	366
5.5.20.	Wykresy konturowe	368
5.5.21.	Wykres mapa ciepła	368
5.5.22.	Wykres profili obserwacji	369
5.5.23.	Parametry graficzne	370

5.6. Pakiet rCharts	374
5.6.1. Wprowadzenie	374
5.6.2. Biblioteka NVD3	374
5.6.3. Biblioteka Leaflet	376
5.6.4. Inne szablony	378
Opis zbiorów danych	379
Skorowidz	384