

INFORMATYKA II

ANALIZA DANYCH W JĘZYKU R

CEL LABORATORIUM

W trakcie zajęć zaprezentowane zostaną podstawy operacji na plikach w formacie JSON w języku R. Dane zostaną poddane analizie przeglądowej, w tym ekstrakcji informacji takich jak średnia, mediana, wartość minimalna, maksymalna, wariancja, odchylenie standardowe czy IQR. Pokazane zostaną także metody generacji wykresów punktowych, pudełkowych i histogramów.

MATERIAŁY POMOCNICZE

Wykorzystywany w trakcie zajęć plik JSON można znaleźć pod następującym adresem: <http://www.stawarz.edu.pl/informatyka2/katalog.json>.

Środowisko R można pobrać na stronie <https://www.r-project.org>.

Podstawowa dokumentacja języka R jest dostępna pod adresem <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>.

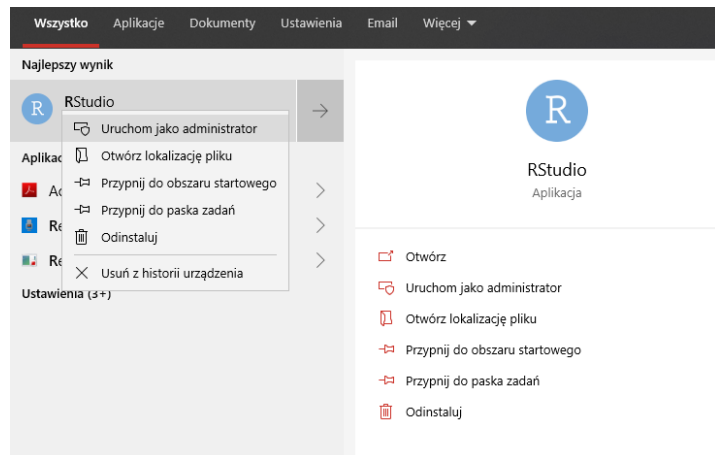
Wykorzystywany podczas zajęć program RStudio Desktop jest dostępny pod adresem: <https://rstudio.com>.

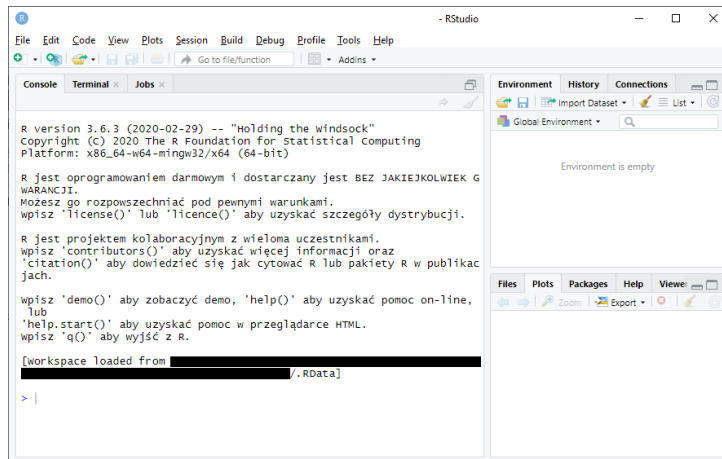
Do kompilacji pakietu *jsonlite* w RStudio na systemach z rodziny Windows, wymagane jest ściągnięcie narzędzi Rtools. Są one dostępne do pobrania pod adresem: <https://cran.r-project.org/bin/windows/Rtools>. Na systemach z rodziny Mac istnieje oprogramowanie XCode. Różne dystrybucje systemu Linux, posiadają własne zestawy narzędzi i należy się skonsultować z dokumentacją danej dystrybucji.

TWORZENIE NOWEGO PROJEKTU

Aby uniknąć problemów, pracę należy rozpocząć od uruchomienia aplikacji R **jako administrator**. W przeciwnym razie, pobranie i skompilowanie pakietu *jsonlite* nie będzie możliwe.

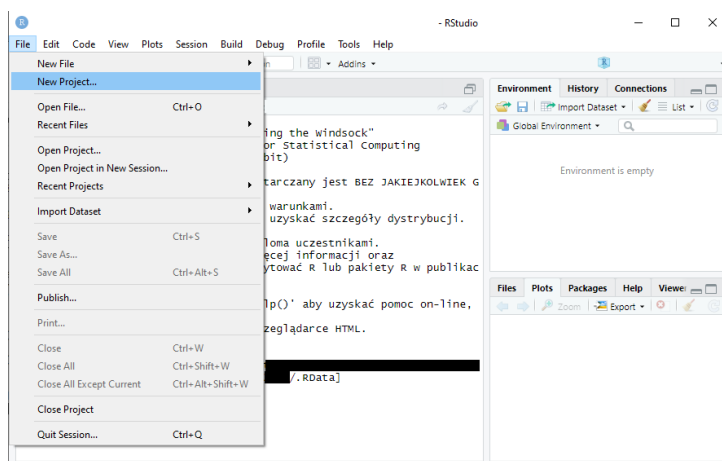
Po uruchomieniu programu, wyświetla się okno główne. Powinno ono przypominać wyglądem to, pokazane na zdjęciu znajdującym się na górze następnej strony.



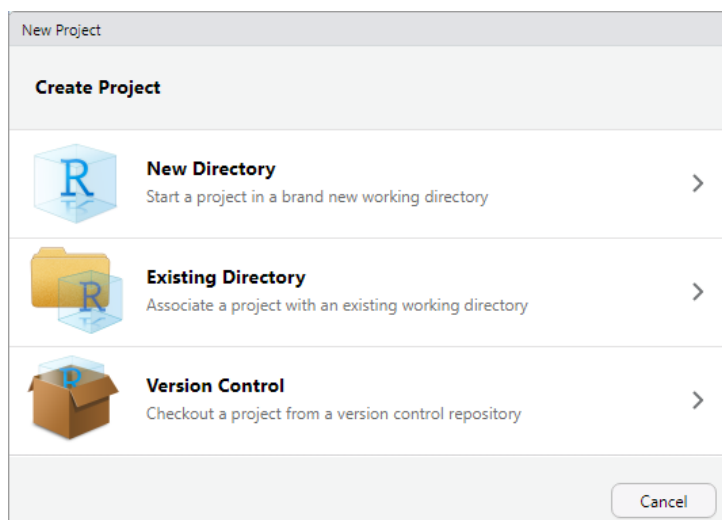


Następnie należy utworzyć nowy projekt. W tym celu należy wykonać następujące kroki:

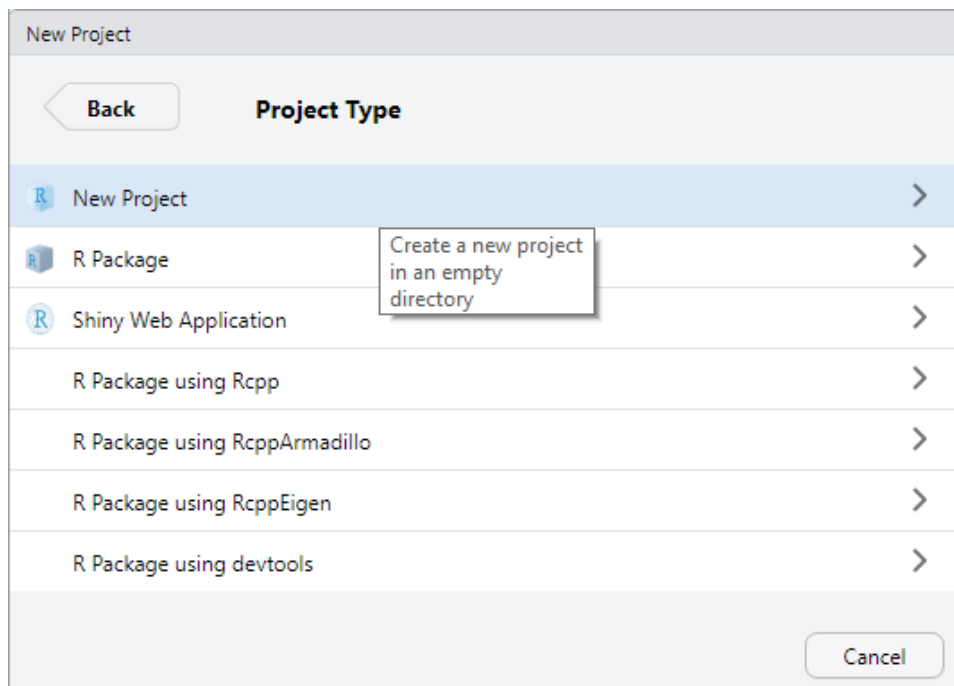
Krok 1. Z menu aplikacji wybrać pozycję „File” i z menu kontekstowego wybrać pozycję „New Project...”:



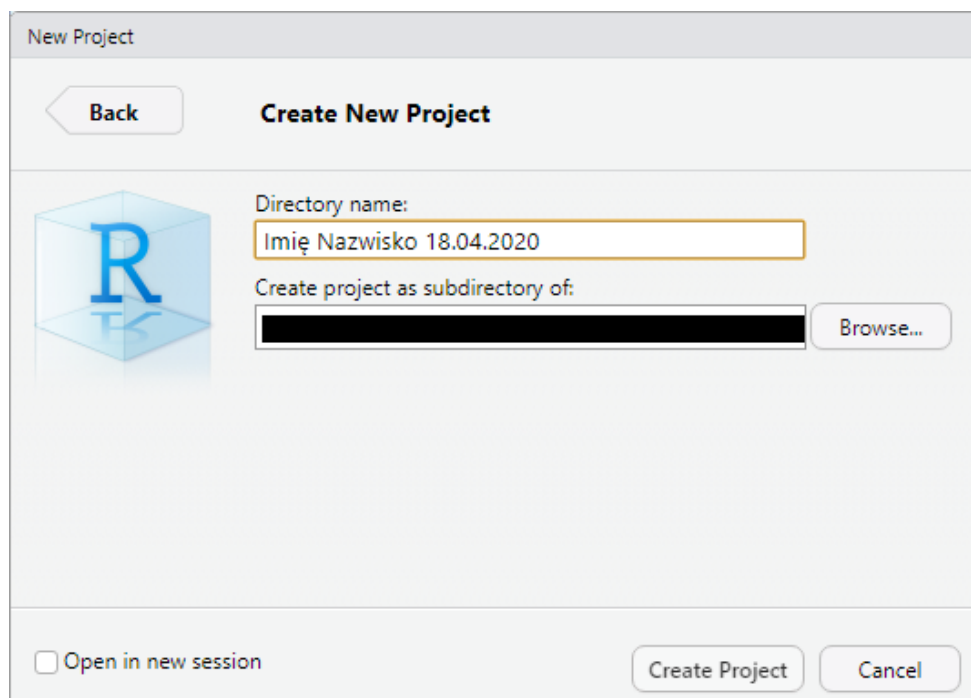
Krok 2. Jeżeli praca odbywa się w laboratorium, należy wybrać opcję „New Directory”:



Krok 3. W następnym oknie należy zdecydować się na „New Project”:

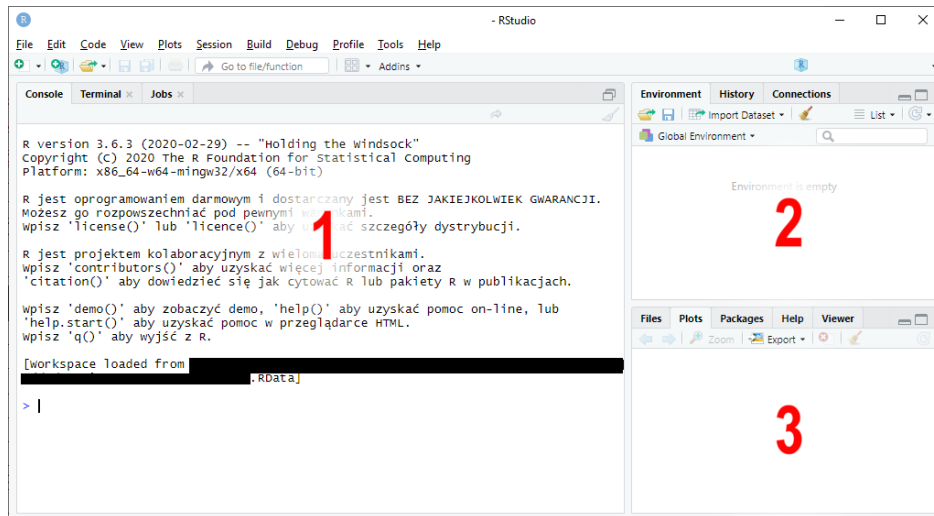


Krok 4. Nowemu projektowi należy nadać stosowną nazwę. Tak jak w przypadku projektów w środowisku PyCharm, tak i projektom tworzonym w RStudio, należy nadać nazwę według schematu „Imię nazwisko data”. Katalog należy pozostawić bez zmian. Decyzję trzeba zatwierdzić poprzez wciśnięcie przycisku „Create Project”:




PRACA Z PROGRAMEM RSTUDIO

W oknie głównym programu RStudio można wyróżnić kilka elementów:



1. Okno terminala. Tutaj należy wpisywać komendy języka, które będą wykonywane na bieżąco.
2. Sekcja listująca wartości zmiennych. Dwukliknięcie na nazwę zmiennej spowoduje rozwinięcie listy wartości lub otwarcie nowego okna.
3. Sekcja dająca dostęp do wszystkich wygenerowanych grafów, dołączonych pakietów i plików.

Nowy skrypt można utworzyć wybierając z menu pozycję „File”, a następnie „New File” i klikając na pozycję „R Script” menu kontekstowego. Okno konsoli zostanie wtedy **podzielone** na dwie części. Górna część zawierać będzie okno skryptu, zaś dolna konsolę.

W sekcjach 2 i 3 znajduje się pomocny przycisk z ikonką miotły („”). Pozwala on wyczyścić zawartość pamięci programu – w tym obliczone zmienne i nakreślone podczas pracy wykresy. Jeżeli realizacja instrukcji odbywa się w laboratorium, z opcji tej **trzeba** skorzystać tuż przed zakończeniem pracy z programem.

Podczas pisania skryptu w osobnym pliku, warto również zwrócić uwagę na opcję „Source on Save”, która znajduje się w części okna konsoli dotyczącej otwartego skryptu. Warto **zaznaczyć** tę opcję. Sprawi to, że program RStudio będzie wykonywał skrypt po każdym zapisaniu zmian.

INSTALACJA PAKIETU JSONLITE

Wczytywanie pliku w formacie JSON odbywać się będzie za pomocą pakietu jsonlite. Domyślnie pakiet ten nie jest dostarczony ani ze środowiskiem R, ani przez program RStudio. Aby go pobrać, należy w konsoli programu wpisać następującą komendę:

```
install.packages("jsonlite")
```

Zainstalowanie pakietu zakończy się powodzeniem wyłącznie wtedy, gdy program RStudio był uruchomiony z uprawnieniami administratora oraz na komputerze został zainstalowany zestaw narzędzi Rtools. Po poprawnym zakończeniu operacji, konsola programu powinna zawierać komunikat podobny do następującego:

```
próbowanie adresu URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/jsonlite_1.6.1.zip'  
Content type 'application/zip' length 1165728 bytes (1.1 MB)  
downloaded 1.1 MB
```

```
package 'jsonlite' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in  
x:\Jakaś ścieżka\downloaded_packages
```

ANALIZA PRZEGLĄDOWA CEN LAMP

Plik JSON należy umieścić w katalogu projektu. Dostęp do katalogu projektu najłatwiej uzyskać z zakładki „files” (jej umiejscowienie w oknie programu zostało wyjaśnione w poprzedniej sekcji instrukcji). W dalszej części instrukcji poczyniono założenie, że nazwa pliku z danymi to „katalog.json”.

Aby możliwe było wykorzystanie pakietu *jsonlite*, należy wywołać polecenie *library*:

```
library("jsonlite")
```

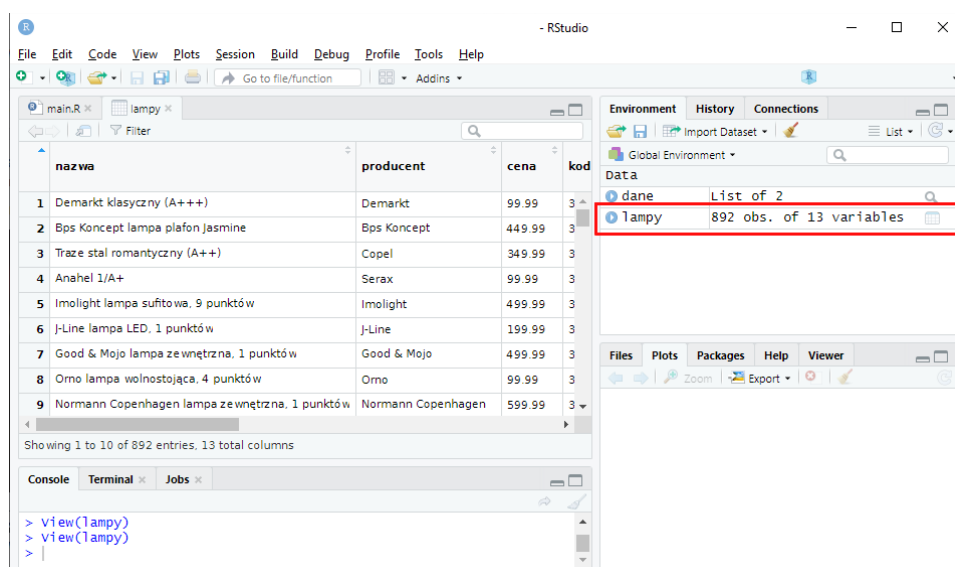
Następnie dane można pobrać wykorzystując metodę *fromJSON*. Pierwszym argumentem powinna być nazwa pliku:

```
dane <- fromJSON("katalog.json")
```

Zmienna *dane* powinna zawierać teraz wszystkie informacje zawarte w pliku *katalog.json*. Aby wydobyć z nich tylko te, które dotyczą lamp, można użyć następującej instrukcji:

```
lampy <- dane[["lampy"]]
```

W sekcji okna programu listującego dane, powinna pojawić się teraz pozycja „lampy”. Podwójne kliknięcie jej nazwy, otworzy w oknie głównym podgląd danych. Ten sam efekt można uzyskać za pomocą komendy „view”:



Aby ze zmiennej *lampy* wydobyć informacje dotyczące wyłącznie cen tychże, można przeprowadzić bezpośrednią indeksację:

```
cenyLamp = lampy$scena
```

Analizę przeglądowną można wykonać na zmiennej `cenyLamp`. Przykładowa analiza wykonana była podczas **wykładu**. Interesujące są te same właściwości danych, jakie obliczano na poprzednich zajęciach laboratoryjnych w języku Python: minimum, maksimum, średnia, mediana, wartość pierwszego i trzeciego kwartylu, przedział międzykwartylowy, odchylenie standardowe i wariancja. Zwracane są one przez funkcje języka R: `summary`, `IQR`, `var` i `sd`. Do każdej z tych funkcji wystarczy jako argument podać zmienną `cenyLamp`, jak pokazano poniżej:

```
summary(cenyLamp)
```

W odpowiedzi, konsola powinna wyświetlić żądane informacje:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
49.99	149.99	249.99	470.62	499.99	11999.99

ANALIZA KORELACJI I RYSOWANIE WYKRESÓW W ŚRODOWISKU R

Podczas poprzednich zajęć zaprezentowano analizę korelacji pomiędzy ceną lampy, a siłą strumienia świetlnego. Aby zrealizować to zagadnienie w języku R, warto utworzyć zmienną pomocniczą przechowującą wyłącznie dane o sile strumienia lamp. W celu odwołania się do atrybutu, którego nazwa zawiera białe znaki, należy otoczyć go grawisami, tak jak miało to miejsce w języku SQL:

```
siłaStrumienia = lampy$`strumień świetlny`
```

Do obliczenia współczynnika korelacji służy funkcja **cor**. Jej argumentami są wartości, pomiędzy którymi obliczana ma być korelacja. W analizowanym przypadku wartości te są zapisane w zmiennych `cenyLamp` i `siłaStrumienia`, dlatego też wywołanie funkcji przyjmie następującą postać:

```
cor(cenyLamp, siłaStrumienia)
```

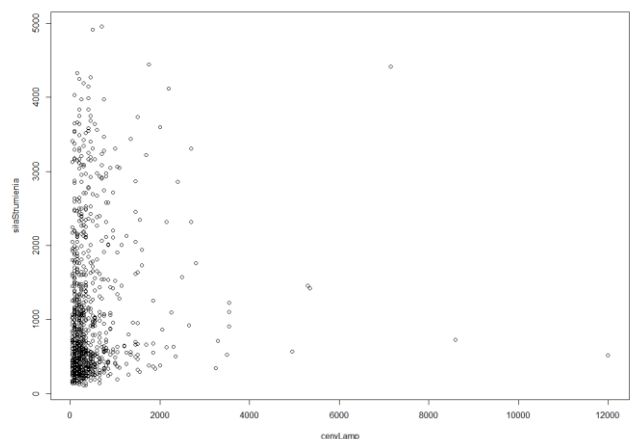
Możliwe jest także narysowanie wykresu punktowego za pomocą funkcji `plot`. Funkcja ta wymaga podania parametrów dla osi x i y. Przykładowe wywołanie funkcji `plot` może więc wyglądać następująco:

```
plot(x = cenyLamp, y = siłaStrumienia)
```

Dla omawianych danych, wykres przyjmie postać zaprezentowaną obok.

Od razu widać, że dane, pomijając kilka odstających punktów, są dość skupione i niezbyt skorelowane. W przeciwnym wypadku, wykres byłby bardziej podobny do linii ciągłej.

Warto spojrzeć na wykres cen lamp. Obraz bywa łatwiejszy w interpretacji niż surowe liczby. Aby wyświetlić histogram wartości, można się posłużyć funkcją `hist`:



```
hist(cenyLamp)
```

Dostępna jest również funkcja `boxplot`, która wykreśla wykres pudełkowy:

```
boxplot(cenyLamp)
```

W analizowanym przypadku, tego typu wykres nie będzie jednak zbyt pomocny. Bardziej pomocny może okazać się wykres słupkowy, który jest efektem wywołania komendy `barplot`:

```
barplot(cenyLamp)
```

ZADANIA DO SAMODZIELNEGO WYKONANIA

ZADANIE 1

Wykonaj analizę przeglądową cen telewizorów.

ZADANIE 2

Wykonaj analizę korelacji pomiędzy siłą strumienia świetlnego lampy, a ilością punktów światła.

ZADANIE 3

Wyświetl wykres słupkowy przedstawiający ilość lamp z daną oprawką.

Podpowiedź: możesz wykorzystać funkcję `table`. Pozwala zliczyć elementy listy mające identyczne wartości.

ZADANIE 4

Wyświetl wykres słupkowy przedstawiający ilość lamp wytworzonych z danego materiału.

Podpowiedź: możesz wykorzystać funkcję `table`. Pozwala zliczyć elementy listy mające identyczne wartości.

ZADANIE 5

Wykonaj analizę przeglądową przekątnej telewizorów.

ZADANIE 6

Wykonaj analizę korelacji pomiędzy ceną telewizora, a jego przekątną

ZADANIE 7

Wykonaj analizę korelacji pomiędzy ceną telewizora, a jego typem matrycy. Matryca może być następującego typu: TN, VA, IPS, Plazma lub OLED.

Podpowiedź: w celu obliczenia korelacji, wartości tekstowe należy skonwertować na numeryczne.

ZADANIE 8 (NAGRADZANE 2 PLUSAMI)

Odpowiedz na pytanie: jaka właściwość lampy ma największy wpływ na jej cenę? Wynik zaprezentuj na wykresie przedstawiającym korelacje pomiędzy ceną, a wszystkimi innymi parametrami opisującymi lampę.

ZADANIE 9 (NAGRADZANE 2 PLUSAMI)

Odpowiedz na pytanie: jaka właściwość telewizora ma największy wpływ na jej cenę? Wynik zaprezentuj na wykresie przedstawiającym korelacje pomiędzy ceną, a wszystkimi innymi parametrami opisującymi telewizor.

Autor:	Mgr inż. Paweł Stawarz, 20.04.2020
Korekta:	Mgr inż. Michał Madera, SoftSystem Sp. z o.o.