# 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: <u>http://www.stawarz.edu.pl/informatyka2/katalog.json</u>.

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

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

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

Do kompilacji pakietu *jsonlite* w RStudio na systemach z rodziny Windows, wymagane jest ściągnięcie narzędzi Rtools. Są one dostępne do pobrania pod adresem: <u>https://cran.r-project.org/bin/windows/Rtools</u>. 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.



B - RStudio			-		×
<u>File Edit Code View Plots Session Build Debug Profile Tools Help</u>					
💿 🔹 🧐 💣 🖌 📄 📄 🌧 Go to file/function 🔤 🗄 👻 Addins 👻					
Console Terminal × Jobs ×	Environm	nent History	Connect	tions	
2 4	😭 🔒	Import Dat	aset 🔹 🥑	🖉 🗉 List	•   C
Profession 2 6 2 (2020 02 20) "Helding the Windcock"		al Environment 👻	Q,		
Copyright (C) 2020 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit)	Environment is empty				
R jest oprogramowaniem darmowym i dostarczany jest BEZ JAKIEJKOLWIEK G					
Możesz go rozpowszechniać pod pewnymi warunkami. wpisz 'license()' lub 'licence()' aby uzyskać szczegóły dystrybucji.					
R jest projektem kolaboracyjnym z wieloma uczestnikami. Wpisz 'contributors()' aby uzyskać więcej informacji oraz 'citation()' aby dowiedzieć się jak cytować R lub pakiety R w publikac jach.					
		Plots Package	es Help	Viewe	
<pre>Wpisz 'demo()' aby zobaczyć demo, 'help()' aby uzyskać pomoc on-line, lub</pre>		🔎 Zoom   🖥	Export 🔹	0	(C
'help.start()' aby uzyskać pomoc w przeglądarce HTML. wpisz 'q()' aby wyjść z R.					
[workspace loaded from					
/.RData]					
>					

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":

New Project				
Back Project Type				
R New Project		>		
🔋 R Package	Create a new project in an empty	>		
R Shiny Web Application	directory	>		
R Package using Rcpp	>			
R Package using RcppArmadillo	>			
R Package using RcppEigen	>			
R Package using devtools		>		
		Cancel		

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":

New Project		
Back	Create New Project	
R	Directory name: Imię Nazwisko 18.04.2020 Create project as subdirectory of:	Browse
Open in new ses	sion	Create Project Cancel

## PRACA Z PROGRAMEM RSTUDIO

#### W oknie głównym programu RStudio można wyróżnić kilka elementów:

	RStudio – 🗆 🗙
Elle Edit Code View Plots Session Build Debug Profile Tools Help ♥ • ♥	
Console Terminal × Jobs × R version 3.6.3 (2020-02-29) "Holding the Windsock" Copyright (C) 2020 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit)	Environment History Connections Environment History Connections Global Environment *
R jest oprogramowaniem darmowym i dostarczany jest BEZ JAKIEJKOLI Mozesz go rozpowszechniać pod pewnymi wyrokami. Wpisz 'license() lub 'licence()' aby u dać szczegóły dystrybu R jest projektem kolaboracyjnym z wielom uczestnikami. Wpisz 'contributors()' aby uzyskać więcej 'informacji oraz 'citation()' aby dowiedzieć się jak cytować R lub pakiety R w pu	cji. Environment is empty blikacjach.
wpisz 'demo()' aby zobaczyć demo, 'help()' aby uzyskać pomoc on- 'help.start()' aby uzyskać pomoc w przeglądarce HTML. wpisz 'q()' aby wyjść z R.	11ne, 1ub Files Plots Packages Help Viewer □ ↓ ♪ Zoom 2 Export • 0 √
[Workspace loaded from .RData]	
>1	3

- 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/jso nlite\_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")</pre>
```

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"]]</pre>
```

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":

R			- R	Studio	– 🗆 X
Eile	Edit Code View Plots Session Build Debug   (𝔅) (𝔅) (𝔅) (𝔅) (𝔅) (𝔅) (𝔅)	Profile Tools Help			. ا
	main.R × lampy ×	0	c	-	Environment History Connections
-	nazwa	producent	cena	kod	Global Environment * Q. Data
1	Demarkt klasyczny (A+++)	Demarkt	99.99	3 ^	💽 dane List of 2 Q
2	Bps Koncept lampa plafon Jasmine	Bps Koncept	449.99	3	Iampy 892 obs. of 13 variables
3	Traze stal romantyczny (A++)	Copel	349.99	3	
4	Anahel 1/A+	Serax	99.99	3	
5	Imolight lampa sufitowa, 9 punktów	Imolight	499.99	3	
6	J-Line lampa LED, 1 punktów	J-Line	199.99	3	
7	Good & Mojo lampa ze wnętrzna, 1 punktó w	Good & Mojo	499.99	3	Files Plots Packages Help Viewer
8	Orno lampa wolnostojąca, 4 punktów	Orno	99.99	3	🖕 🔿 🎤 Zoom 🖓 📶 Export 🔹 🥝 🇹 🛞
9	Normann Copenhagen lampa ze wnętrzna, 1 punktów	Normann Copenhagen	599.99	3 🗸	
4				+	
Sho	wing 1 to 10 of 892 entries, 13 total columns				
Cor	sole Terminal × Jobs ×			_	
> \ > \ >	iew(lampy) iew(lampy)			* 	

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

cenyLamp = lampy\$cena

Analizę przeglądową 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, silaStrumienia)
```

Możliwe jest także nakreślenie 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 P

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 P

Wykonaj analizę przeglądową przekątnej telewizorów.

# ZADANIE 6 🖻

Wykonaj analizę korelacji pomiędzy ceną telewizora, a jego przekątną

# ZADANIE 7 P

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.