

Projek RPIS (Sprawozdanie/Manual)

autor: Oleh Volosnik
nr indeksu: 140244
mail: olegvolosnik@gmail.com

Opis projektu: celem projektu jest napisanie narzędzia do analizy danych z plików „CSV”. Program ma być prosty w obsłudze i przyjazny dla każdego użytkownika. Program działa na w trybie wsadowym, gdzie jako argument podajemy nazwę pliku.

1. Tryb wsadowy.

1.1. Otwieramy Command Prompt (wpisując CMD w wyszukiwarce Windows).

1.2. Przechodzimy do folderu w którym znajduje się skrypt/narzędzie („1.R”), w moim przypadku jest to w lokalizacji: „D:\Bioinformatics\4 semester\Statistics\LAB\Project”. Żeby zmienić dysk C na D wpisujemy „:d + ENTER”.

```
C:\Users\xc>d:
```

Teraz za pomocą komendy „cd + ścieżka_do_folderu” zmieniamy katalog.

```
D:\>  
D:\>cd "Bioinformatics\4 semester\Statistics\LAB\Project"  
D:\Bioinformatics\4 semester\Statistics\LAB\Project>
```

1.3. Wykonywanie skryptu „1.R”. Mam w folderze „D:\Bioinformatics\4 semester\Statistics\LAB\Project” zarówno skrypt „1.R” jak i plik „CSV”.

Jest to wersja zalecana.

Żeby wykonać skrypt mamy podać:

- ścieżkę do R.exe („C:\Program Files\R\R-3.6.3\bin\R.exe”)
- CMD BATCH --vanilla „--args nazwa_pliku.csv” 1.R

```
D:\Bioinformatics\4 semester\Statistics\LAB\Project>"C:\Program Files\R\R-3.6.3\bin\R.exe"  
CMD BATCH --vanilla "--args PrzykladoweDane-Projekt.csv" 1.R
```

Jako wynik, w folderze z skryptem oraz „CSV”, pojawią się pliki z raportami oraz wykresy.

2. Wymogi.

2.1. Wymogi techniczne.

Na komputerze ma być zainstalowany „R”. Konieczne biblioteki dla narzędzia:

```
library(magrittr)
library(dplyr)
library(xlsx)
library(car)
library(dunn.test)
library(FSA)
library(ggpubr)
library(ggplot2).
```

Instrukcja do instalowania pakietów:

<https://www.r-bloggers.com/how-to-install-packages-on-r-screenshots/>

2.2. Wymogi do pliku „CSV”.

Plik ma posiadać dwie nazwy zdefiniowane „na sztywno”.

Jedna z kolumn – która reprezentuje grupy badane ma nazywać się: **„grupa”** (istotna jest wielkość liter!).


Jedna z grup badanych, reprezentujących grupę kontrolną ma nazywać się: **„KONTROLA”** (istotna jest wielkość liter!).

Komórki mają być separowane przez = ";", a liczby po przecinku mają być w postaci **„9,14”** a nie „9.14”.

Plik nie może zawierać inne znaki niż litery alfabetu oraz liczby.

3. Działanie narzędzia.

3.1 Pierwszy raport „report_NA_replace.txt” zawiera informacje o zamianie komórek „NA – not available” na średnią wartość danego parametru w poszczególnej grupie.

 report_NA_replace - Notepad

File Edit Format View Help

HGB 13 replaced with avg in group/column: 12.4114125

HGB 68 replaced with avg in group/column: 11.263575

MON 5 replaced with avg in group/column: 0.8579166666666667

3.2. Drugim raportem są dwa pliki: „summary1.1.xlsx” oraz „summary1.2.csv”.

M19	A	B	C	D
1	Var1	Var2	Freq	
2		grupa	CHOR1 :25	
3		grupa	CHOR2 :0	
4		grupa	KONTROLA: 0	
5		grupa	#N/A	
6		grupa	#N/A	
7		grupa	#N/A	
8		plec	k:14	
9		plec	m:11	
10		plec	#N/A	
11		plec	#N/A	
12		plec	#N/A	
13		plec	#N/A	
14		wiek	Min. :17.00	
15		wiek	1st Qu.:26.00	
16		wiek	Median :29.00	
17		wiek	Mean :29.56	
18		wiek	3rd Qu.:32.00	
19		wiek	Max. :43.00	
20		hsCRP	Min. :0.4876	
21		hsCRP	1st Qu.:2.3227	
22		hsCRP	Median :3.9665	

A1		\bar{x}	Σ	Var1
	A	B	C	D
1	Var1	Var2	Freq	
2		grupa	CHOR1 : 0	
3		grupa	CHOR2 :25	
4		grupa	KONTROLA: 0	
5		grupa	#N/A	
6		grupa	#N/A	
7		grupa	#N/A	
8		plec	k:12	
9		plec	m:13	
10		plec	#N/A	
11		plec	#N/A	
12		plec	#N/A	
13		plec	#N/A	
14		wiek	Min. :21.00	
15		wiek	1st Qu.:25.00	
16		wiek	Median :30.00	
17		wiek	Mean :30.04	
18		wiek	3rd Qu.:33.00	
19		wiek	Max. :42.00	
20		hsCRP	Min. : 0.3351	
21		hsCRP	1st Qu.: 2.0781	
22		hsCRP	Median : 3.4455	

A1	\sum	\bar{x}	\sum	\bar{x}	Var1
	A	B	C	D	
1	Var1	Var2	Freq		
2		grupa	CHOR1 : 0		
3		grupa	CHOR2 : 0		
4		grupa	KONTROLA:25		
5		grupa	#N/A		
6		grupa	#N/A		
7		grupa	#N/A		
8		plec	k:14		
9		plec	m:11		
10		plec	#N/A		
11		plec	#N/A		
12		plec	#N/A		
13		plec	#N/A		
14		wiek	Min. :23.00		
15		wiek	1st Qu. :29.00		
16		wiek	Median :32.00		
17		wiek	Mean :32.32		
18		wiek	3rd Qu. :35.00		
19		wiek	Max. :48.00		
20		hsCRP	Min. : 0.7584		
21		hsCRP	1st Qu. : 2.3022		
22		hsCRP	Median : 4.2204		

„summary1.2.csv“

	A	B	C	D	E	F	G	H	I	J	K
1	grupa	plec	wiek	hsCRP	ERY	PLT	HGB	HCT	MCHC	MON	LEU
2	CHOR1 :25	k:14	Min.: 17.00	Min.: 0.4876	Min.: 3.530	Min.: 128.0	Min.: 9.505	Min.: 0.2800	Min.: 32.56	Min.: 0.4800	Min.: 6.79
3	CHOR2 :0	m:11	1st Qu.:26.00	1st Qu.: 2.3227	1st Qu.: 4.070	1st Qu.:179.0	1st Qu.:11.921	1st Qu.:0.3500	1st Qu.:34.71	1st Qu.:0.6100	1st Qu.:10.11
4	KONTROLA:0	NA	Median :29.00	Median : 3.9665	Median : 4.200	Median :217.0	Median :12.405	Median :0.3630	Median :35.05	Median :0.7600	Median :11.66
5	NA	NA	Mean :29.56	Mean : 6.1030	Mean : 5.363	Mean :225.3	Mean :12.411	Mean :0.3636	Mean :35.13	Mean :0.8759	Mean :12.02
6	NA	NA	3rd Qu.:32.00	3rd Qu.: 4.9935	3rd Qu.: 4.510	3rd Qu.:266.0	3rd Qu.:13.210	3rd Qu.:0.3860	3rd Qu.:35.60	3rd Qu.:1.0700	3rd Qu.:14.48
7	NA	NA	Max.: 34.00	Max.: 42.6499	Max.: 33.000	Max.: 336.0	Max.: 14.499	Max.: 0.4050	Max.: 36.87	Max.: 1.5200	Max.: 16.81
8	grupa	plec	wiek	hsCRP	ERY	PLT	HGB	HCT	MCHC	MON	LEU
9	CHOR1 :0	k:12	Min.: 21.00	Min.: 0.3351	Min.: 3.250	Min.: 91.0	Min.: 9.827	Min.: 0.0423	Min.: 32.89	Min.: 0.1400	Min.: 7.95
10	CHOR2 :25	m:13	1st Qu.:25.00	1st Qu.: 2.0781	1st Qu.:3.850	1st Qu.:172.0	1st Qu.:11.760	1st Qu.:0.3300	1st Qu.:34.88	1st Qu.:0.5500	1st Qu.:10.70
11	KONTROLA:0	NA	Median :30.00	Median : 3.4455	Median : 4.270	Median :195.0	Median :12.566	Median :0.3600	Median :35.55	Median :0.6600	Median :12.00
12	NA	NA	Mean :30.04	Mean : 5.5360	Mean : 4.198	Mean :209.1	Mean :12.806	Mean :0.3460	Mean :35.55	Mean :0.9528	Mean :12.04
13	NA	NA	3rd Qu.:33.00	3rd Qu.: 8.6093	3rd Qu.:4.430	3rd Qu.:223.0	3rd Qu.:13.694	3rd Qu.:0.3900	3rd Qu.:36.04	3rd Qu.:0.8800	3rd Qu.:13.34
14	NA	NA	Max.: 42.00	Max.: 19.2124	Max.: 5.040	Max.: 456.0	Max.: 22.232	Max.: 0.4120	Max.: 38.87	Max.: 7.0000	Max.: 16.59
15	grupa	plec	wiek	hsCRP	ERY	PLT	HGB	HCT	MCHC	MON	LEU
16	CHOR1 :0	k:14	Min.: 23.00	Min.: 0.7584	Min.: 3.090	Min.: 147.0	Min.: 9.505	Min.: 0.2790	Min.: 32.06	Min.: 0.3500	Min.: 4.83
17	CHOR2 :0	m:11	1st Qu.:29.00	1st Qu.: 2.3022	1st Qu.:3.820	1st Qu.:188.0	1st Qu.:10.472	1st Qu.:0.3200	1st Qu.:33.72	1st Qu.:0.6500	1st Qu.: 9.22
18	KONTROLA:25	NA	Median :32.00	Median : 4.2204	Median :3.980	Median :214.0	Median :11.438	Median :0.3390	Median :34.55	Median :0.7600	Median :10.68
19	NA	NA	Mean :32.32	Mean : 5.2951	Mean :4.013	Mean :225.9	Mean :11.264	Mean :0.3376	Mean :34.40	Mean :0.7604	Mean :11.36
20	NA	NA	3rd Qu.:35.00	3rd Qu.: 6.8521	3rd Qu.:4.330	3rd Qu.:254.0	3rd Qu.:11.760	3rd Qu.:0.3530	3rd Qu.:35.21	3rd Qu.:0.8600	3rd Qu.:13.59
21	NA	NA	Max.: 48.00	Max.: 14.3951	Max.: 5.050	Max.: 454.0	Max.: 13.210	Max.: 0.3890	Max.: 36.04	Max.: 1.2500	Max.: 17.46

Ten plik posiada jeden wspólny arkusz. Przedstawia dokładnie te same dane co i „summary1.1.xlsx”. Poznać która grupa jest obserwowana można po nie zerowej wartości w kolumnie „A”, jak zostało opisane wyżej.

3.3. Raport „Rplots.pdf”

Zawiera histogramy jeżeli dane są mierzalne(1, 3.14, 0.99) lub ploty jeżeli dane są jakościowe („niebieski”, „kobieta”).

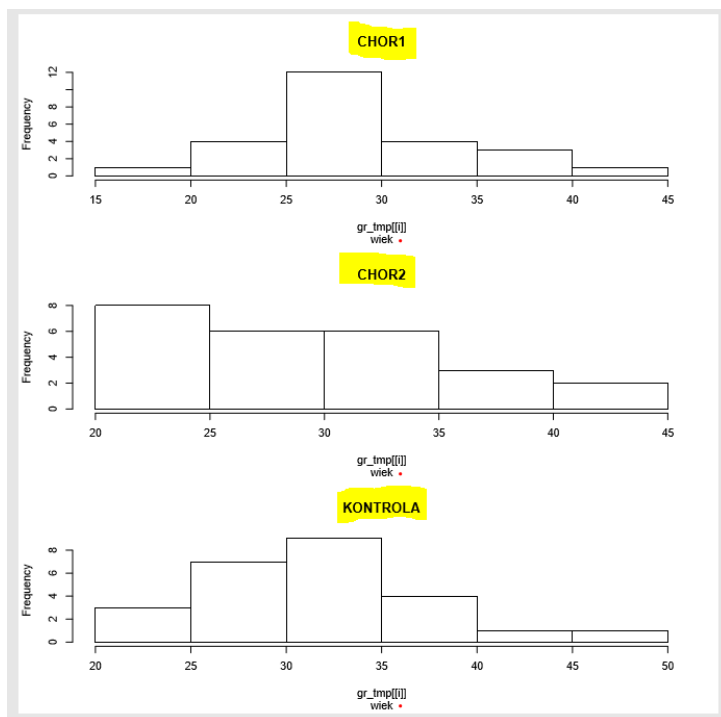


Figure 1: Histogram wieku w poszczególnych grupach.

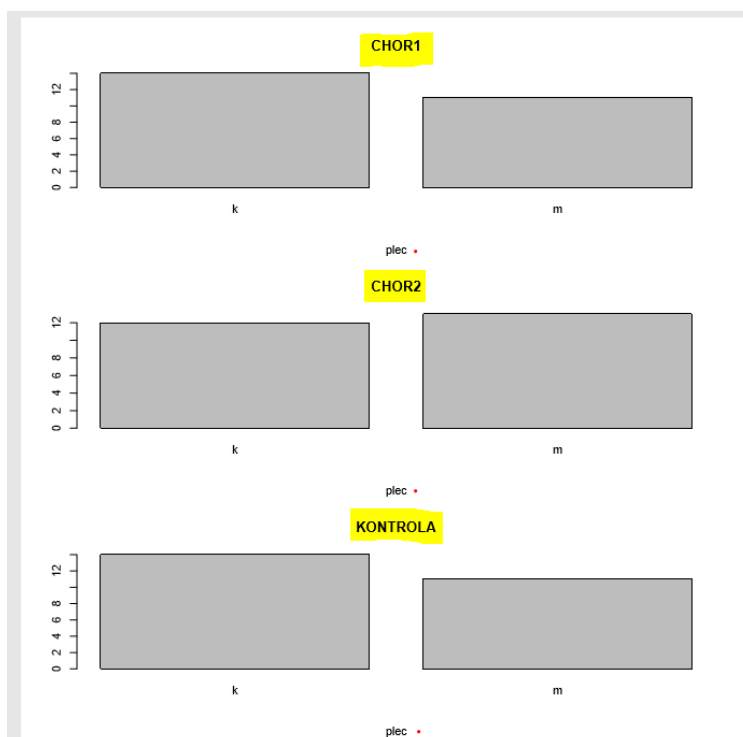


Figure 2: Plot stosunku płciowego.

3.4. Raport „groups_dif_report.txt”.

```
groups_dif_report - Notepad
File Edit Format View Help
plec is not numeric - so you can fund plot in Rplots.pdf as Red and Blue plots, after histograms.

wiek 0.205627845266204 > 0.05 - brak roznic pomiedzy grupami
rsCRP 0.880721219282965 > 0.05 - brak roznic pomiedzy grupami
ERY 0.154351363489479 > 0.05 - brak roznic pomiedzy grupami
PLT 0.32403069707862 > 0.05 - brak roznic pomiedzy grupami

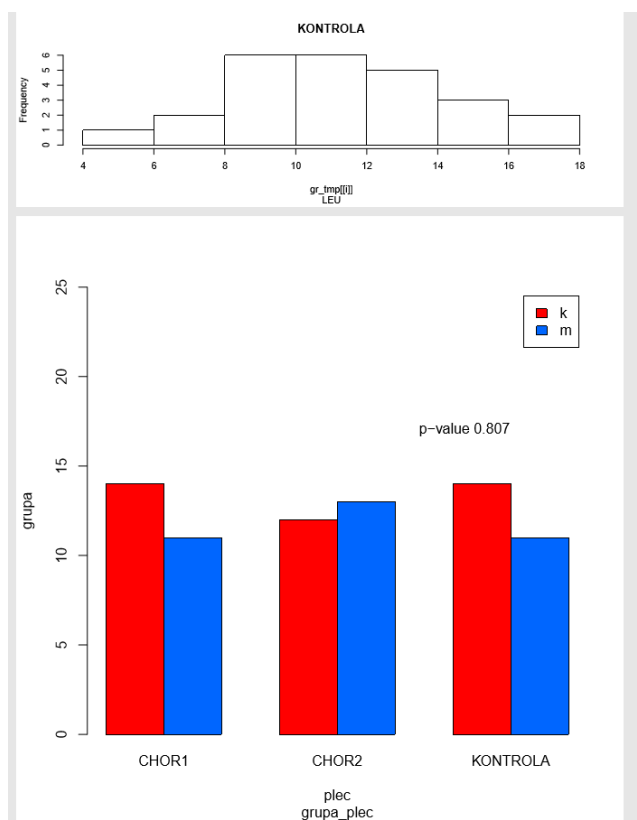
fGB 0.000767331503178641 < 0.05 - sa roznice pomiedzy grupami
post hoc Dunna:
Dunn test outputs are in file dunnTest7.csv.
Order of columns: Comparison Z P.unadj P.adj

fCT 0.0189609134312322 < 0.05 - sa roznice pomiedzy grupami
post hoc Dunna:
Dunn test outputs are in file dunnTest8.csv.
Order of columns: Comparison Z P.unadj P.adj

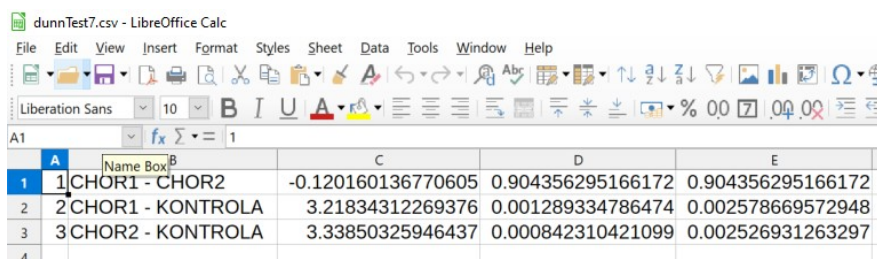
fCHC 0.00185981027219031 < 0.05 - sa roznice pomiedzy grupami
post hoc Tukeya:
TukeyHSD test outputs are in file TukeyHSD9.csv.
Order of columns: Comparison diff lwr p adj

fON 0.25421349117813 > 0.05 - brak roznic pomiedzy grupami
fEU 0.596500941384286 > 0.05 - brak roznic pomiedzy grupami
```

„plec is not numeric - so you can fund plot in Rplots.pdf as Red and Blue plots, after histograms.” - niżej pokazany przykładowy plot dotyczący tego komentarza.



„HGB 0.000767331503178641 < 0.05 - są różnice pomiędzy grupami post hoc Dunna:
Dunn test outputs are in file dunnTest7.csv.
Order of columns: Comparison Z P.unadj P.adj” - takiego typu komentarze oznaczają że skutek wykonanych testów istnieją istotne różnice pomiędzy grupami które są w osobnym pliku: „dunnTest7.csv”. Kolejność kolumn od „B” do „E” – została opisana odpowiednio:
„Comparison Z P.unadj P.adj”.



	A	B	C	D	E
1	1	CHOR1 - CHOR2	-0.120160136770605	0.904356295166172	0.904356295166172
2	2	CHOR1 - KONTROLA	3.21834312269376	0.001289334786474	0.002578669572948
3	3	CHOR2 - KONTROLA	3.33850325946437	0.000842310421099	0.002526931263297

„MON 0.25421349117813 > 0.05 - brak różnic pomiędzy grupami” – oznacza brak różnic.

Porównanie grup niezależnych			
Ilość porównywanych grup	Zgodność z rozkładem normalnym	Jednorodność wariancji	Wybrany test
2	TAK	TAK	test t-Studenta (dla gr. niezależnych)
		NIE	test Welcha
	NIE	-	test Wilcoxona (Manna-Whitneya)
> 2	TAK	TAK	test ANOVA (post hoc Tukeya)
		NIE	test Kruskala-Wallisa (post hoc Dunna)
	NIE	-	

Figure 3: Tabela z schematem wykonanych testów.

***Metodologia wyboru testu

Warto wspomnieć że została badana jednorodność wariancji oraz rozkład normalny dla wyboru odpowiedniego testu końcowego. Dane przykładowe zawierali 3 grupy: „CHOR1”, „CHOR2”, „KONTROLA”. Badamy parametr „ERY”. Wszystkie 3 grupy mają zgodność z rozkładem normalnym, natomiast tylko 2 z 3 mają jednorodność wariancji → wybieramy test ANOVA(post hoc Tukeya).

3.5. Raport „report_correlation.txt”.

```
report_correlation - Notepad
File Edit Format View Help
CHOR1 hsCRP MON 0.6303 r > 0 korelacja dodatnia
(silna korelacja dodatnia)

CHOR1 hsCRP LEU 0.513 r > 0 korelacja dodatnia
(silna korelacja dodatnia)

CHOR1 ERY PLT 0.4149 r > 0 korelacja dodatnia
(korelacja dodatnia o srednim natezeniu)

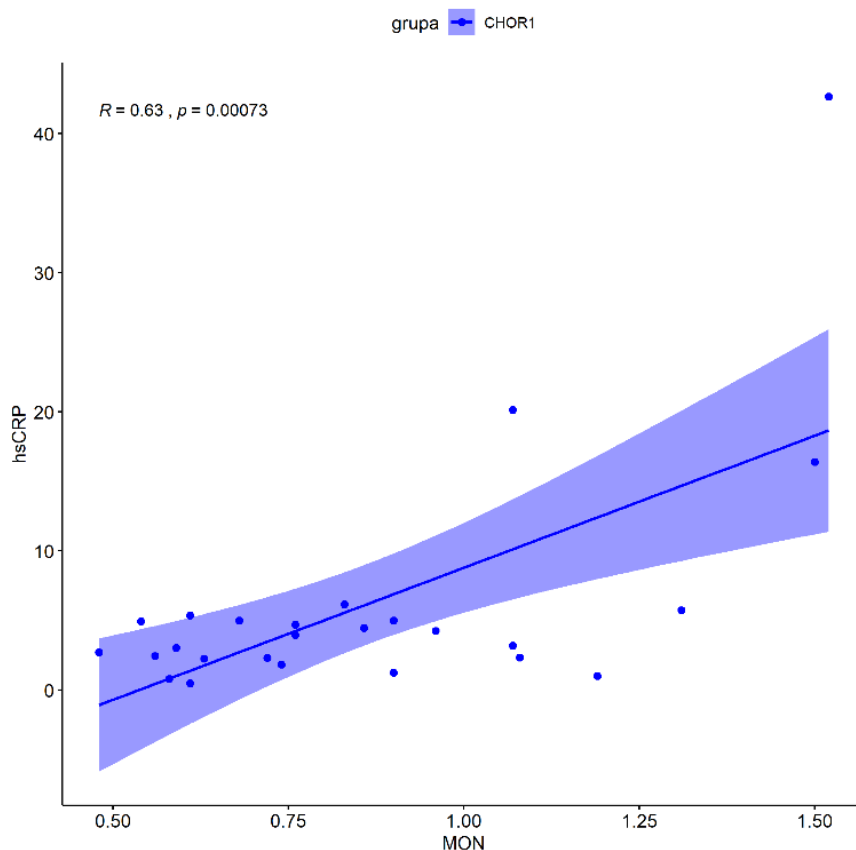
CHOR1 ERY HGB -0.4642 r < 0 korelacja ujemna
(korelacja ujemna o srednim natezeniu)

CHOR1 ERY HCT -0.5249 r < 0 korelacja ujemna
(silna korelacja ujemna)

CHOR1 HGB HCT 0.9533 r > 0 korelacja dodatnia
(bardzo silna korelacja dodatnia)
```

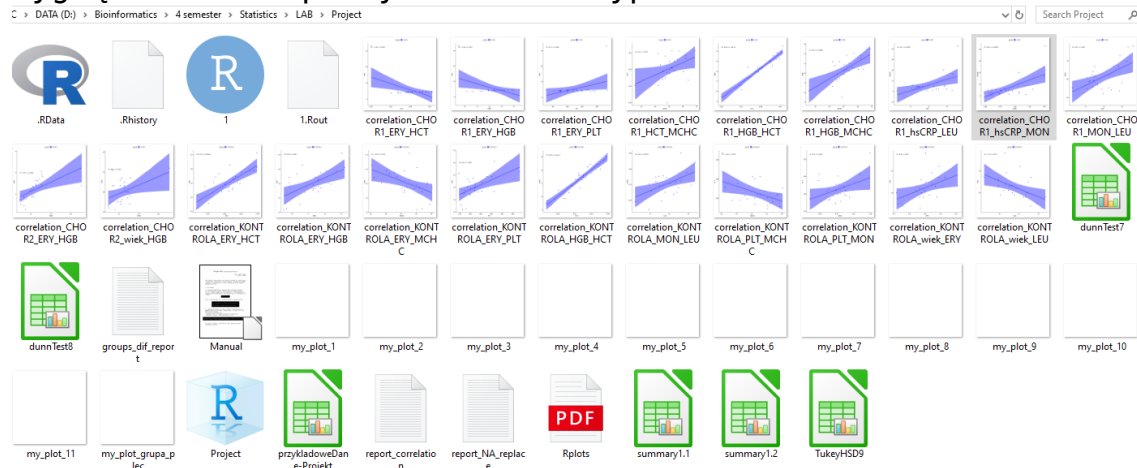
„CHOR1 hsCRP MON 0.6303 r > 0 korelacja dodatnia (silna korelacja dodatnia)” – oznacza to że korelacja między parametrami: „hsCRP” i „MON” w grupie „CHOR1” jest dodatnia i silna.

Do każdego komentarza dotyczącego korelacji z tego raportu jest dodany plik „PNG” z ilustracją korelacji. Plik nazywa się w sposób „correlation_grupa_parametr1_parametr2.png”, czyli dla danego komentarza jest to plik: „correlation_CHOR1_hsCRP_MON.png”.



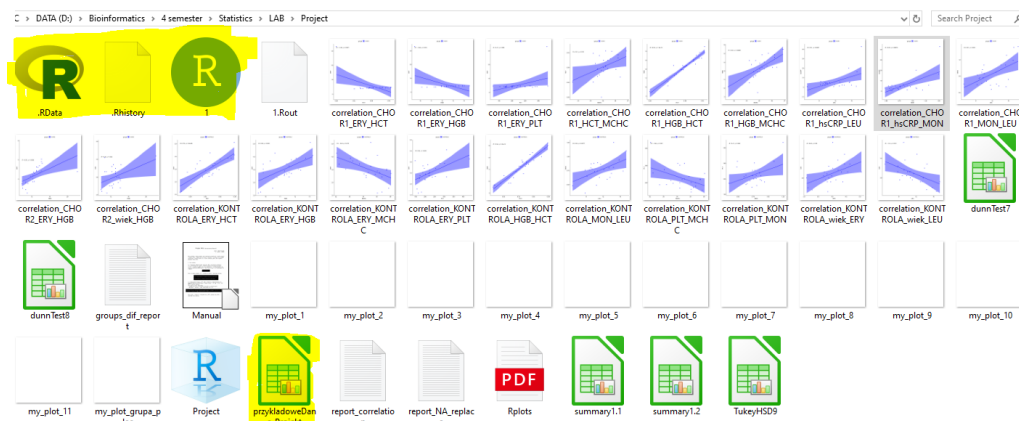
4. Uwagi ogólne.

Tak wygląda folder po wykonaniu skryptu.



Pliki o nazwie: „my_plot_1”, „my_plot_2”... śmiało proszę usuwać. Wszystkie te wykresy znajdują się w „Rplots.pdf” i są oni po prostu produktem ubocznym.

Poniżej na żółto zaznaczam pliki konieczne do działania programu które nie można usuwać z tym że „przykładoweDane-Projekt” w waszym przypadku może nazywać się inaczej, w zależności od pliku na którym przeprowadza się analiza.



Po przeprowadzaniu analizy trzeba wszystkie pliki raportowe przenieść do osobnego katalogu, inaczej przy wykonaniu ponownej analizy dostaniemy pomieszane wyniki z 2 analiz.

Powodzenia!