RPiS II - projekt zaliczeniowy

Sprawozdanie (manual)

Cel projektu

Utworzenie prostego w obsłudze narzędzia do analizy danych zapisanych w formacie .csv, które ma być przyjazny dla użytkownika nieznającego się na programowaniu.

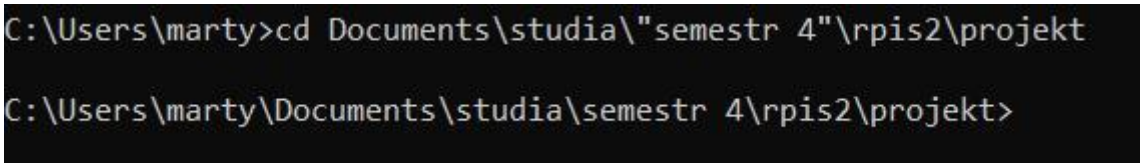
Obsługa programu

1. Uruchamianie skryptu z trybu wsadowego

Aby uruchomić skrypt należy otworzyć Command Prompt poprzez wyszukanie frazy “CMD” w Menu startowym. Następnie należy ustawić ścieżkę do folderu, w którym znajduje się skrypt oraz plik z danymi do wczytania. Jest to możliwe za pomocą komendy:

```
cd [ścieżka do folderu]
```

Przykład:

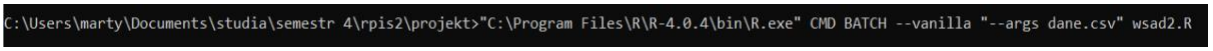


W celu zapewnienia łatwego dostępu do plików, które powstaną po wykonaniu skryptu, zalecane jest umieszczenie skryptu .R oraz pliku .csv z danymi w osobnym folderze.

Aby uruchomić skrypt należy wpisać komendę:

```
"C:\Program Files\R\R-4.0.4\bin\R.exe" CMD BATCH --vanilla "--
        args [nazwa pliku].csv" [nazwa skryptu].R
```

Przykład:



Wynikiem działania skryptu są pliki .png z wykresami oraz jeden plik `output.txt` zawierający m.in. charakterystyki grup, wyniki przeprowadzanych testów czy analizę korelacji.

2. Wymogi dla pliku .csv

W pierwszej kolumnie danych musi zostać zawarta informacja o badanych grupach. Konieczne jest nazwanie tej kolumny “grupa”. Jeżeli dane zawierają płeć badanych osób, należy umieścić ją w drugiej kolumnie. Dane w pozostałych kolumnach muszą być numeryczne, przy czym aby wyrazić wartość dziesiętną należy skorzystać z przecinka (“,”), a nie kropki (“.”). Dane muszą być oddzielone znakiem średnika (“;”). Przykład prawidłowo wyglądających danych:

	A	B	C	D	E	F	G	H
1	grupa;plec;wiek;hsCRP;ERY;PLT;HGB;HCT;MCHC;MON;LEU							
2	CHOR1;k;3	711;4	19;201;13	2102;0	392;34	7149;0	48;11	86
3	CHOR1;m;	69938;4	48;222;13	0491;0	38;35	3793;0	76;10	32
4	CHOR1;k;3	35354;3	59;278;10	1493;0	321;32	5556;1	08;13	6
5	CHOR1;m;	27161;3	66;200;11	277;0	336;34	5488;0	63;10	11
6	CHOR1;m;	46519;4	41;128;12	4047;0	363;35	2132;;10	55	
7	CHOR1;m;	16269;3	68;176;11	4381;0	34;34	7149;0	83;9	28
8	CHOR1;k;2	98836;4	12;288;12	2436;0	357;35	3793;0	9;10	7
9	CHOR1;k;2	84938;4	44;231;13	2102;0	398;34	2166;0	74;9	56
10	CHOR1;m;	1548;4	13;153;12	5658;0	384;35	59523;1	07;14	48
11	CHOR1;m;	20405;4	02;249;11	9214;0	353;34	881;1	07;10	51
12	CHOR1;m;	487607;4	07;177;11	9214;0	35;35	0471;0	61;6	79
13	CHOR1;k;3	32268;4	11;295;12	2436;0	36;35	0471;0	72;14	97
14	CHOR1;m;	4069;4	18;174;;0	334;36	3759;1	5;16		
15	CHOR1;k;2	04427;4	59;207;13	8546;0	394;36	2098;0	59;9	23
16	CHOR1;m;	6499;4	2;170;12	4047;0	364;35	2132;1	52;16	81
17	CHOR1;k;1	803675;4	47;179;13	5324;0	386;36	2098;0	58;11	66
18	CHOR1;k;3	96646;3	53;217;10	6326;0	321;34	2166;0	76;9	72
19	CHOR1;k;3	99354;4	62;266;12	888;0	38;35	0471;0	68;10	98
20	CHOR1;k;2	25615;4	58;230;14	499;0	405;36	8742;0	96;10	21
21	CHOR1;m;	46103;4	51;202;12	7269;0	371;34	3827;0	56;13	23
22	CHOR1;k;3	01799;3	98;250;12	4047;0	355;36	0437;1	19;15	84
23	CHOR1;k;3	93164;4	56;271;14	0157;0	404;35	7115;0	54;14	57
24	CHOR1;k;2	36056;33;3	5049;0	28;34	881;0	61;12	81	
25	CHOR1;k;2	2655;4	51;199;13	8546;0	405;35	2132;0	9;12	72

Działanie programu

1. Plik output.txt

W momencie, w którym skrypt skończy się wykonywać, utworzona zostaje ostateczna wersja pliku output.txt. Zawiera on przedstawienie wyników działania konkretnych metod, które zostały użyte w kodzie. Na początku widnieje informacja, w których miejscach wartości NA (Not Available - wartości, które nie zostały uwzględnione w pliku .csv, przez co dana komórka była pusta) zostały zamienione na wartość średnią dla danej kolumny oraz grupy, z której pochodzi rekord. Podana zostaje także wartość, którą wstawiono w pustą komórkę.

Przykład:

```
Zaimputowano dane w 13 wierszu i 7 kolumnie: 12.41141
Zaimputowano dane w 68 wierszu i 7 kolumnie: 11.26357
Zaimputowano dane w 5 wierszu i 10 kolumnie: 0.8579167
```

Następnie dla wartości liczbowych sporządzono podstawową charakterystykę na podstawie danych dotyczących konkretnych atrybutów oraz podzieloną na odpowiednie grupy. Charakterystyka zawiera:

- Table: - nagłówek informujący o nazwie charakteryzowanego atrybutu
- count - liczbę rekordów w danej grupie
- max - maksymalną wartość dla danego atrybutu
- min - minimalną wartość dla danego atrybutu
- IQR - rozstęp międzykwartylowy, czyli różnica między górnym, a dolnym kwartylem
- mean - średnią wartość dla danego atrybutu
- sd - odchylenie standardowe dla danego atrybutu
- median - medianę dla danego atrybutu

Przykład:

Table: PLT

grupa	count	max	min	IQR	mean	sd	median
KONTROLA	25	434.00	147.00	66.00	225.88	63.81	214.00
CHOR1	25	336.00	128.00	87.00	225.28	54.22	217.00
CHOR2	25	456.00	91.00	51.00	209.12	75.22	195.00

Kolejną informacją zawartą w pliku output.txt jest zestaw tabel będących podsumowaniem testu Shapiro-Wilka, czyli testu mającego za zadanie sprawdzenie, czy dane są zgodne z rozkładem normalnym. W tabelach widnieje nazwa grupy, kolumna *statistic*, która przechowuje wartość statystyczną testu Shapiro-Wilka oraz kolumna *p-value*, która określa prawdopodobieństwo testowe, czyli prawdopodobieństwo, że zależność jaką zaobserwowano w losowej próbie z populacji mogła wystąpić przypadkowo, wskutek losowej zmienności prób, choć w populacji wcale nie występuje.

Przykład:

Table: MCHC

grupa	statistic	p.value
KONTROLA	0.95	0.27
CHOR1	0.95	0.23
CHOR2	0.95	0.28

W dalszych testach przeprowadzanych przez program istotne będą tylko wartości p-value, dlatego zostały one zapisane w jednej tabeli.

Przykład:

pValues dla testu Shapiro

	grupa	wiek	hsCRP	ERY	PLT	HGB	HCT	MCHC	MON	LEU
1	KONTROLA	0.33	0.01	0.63	0.00	0.39	0.61	0.27	0.40	0.88
2	CHOR1	0.62	0.00	0.00	0.55	0.51	0.21	0.23	0.03	0.33
3	CHOR2	0.36	0.00	0.91	0.00	0.00	0.00	0.28	0.00	0.62

Na podstawie wartości w danej kolumnie program określa, czy rozkład dla danego atrybutu jest zgodny z rozkładem normalnym, czy od niego odbiega: jeżeli jakakolwiek wartość p-value w kolumnie jest mniejsza od 0.05, to oznacza, że dla danego atrybutu rozkład odbiega od rozkładu normalnego. Jest to istotne w kolejnych przeprowadzanych testach.

Przy wykonywaniu testu Shapiro-Wilka dla każdego atrybutu tworzony jest wykres gęstości, który wizualizuje rozkład dla danego atrybutu.

Następnie mają miejsce testy jednorodności wariancji, czyli testy Levene'a. Ich wynikiem jest wartość p-value dla konkretnego atrybutu numerycznego.

Przykład:

```
pValues dla testu jednorodnosci wariancji

grupa : NA
wiek : 0.7330851
hsCRP : 0.7907312
ERY : 0.4303854
PLT : 0.9375149
HGB : 0.1593675
HCT : 0.1237021
MCHC : 0.2688996
MON : 0.263275
LEU : 0.3592121
```

Na podstawie wartości p-value dla konkretnego atrybutu określa się, czy zachowana jest jednorodność wariancji: jeżeli p-value jest mniejsze od 0.05, to jednorodność wariancji nie jest zachowana. Ta własność również ma wpływ na dobór kolejnych testów.

Następnie ma miejsce analiza porównawcza grup. Testy dobierane są według poniższego schematu:

Tablica 1: Wyboru testu statystycznego dla 2 i > 2 grup niezależnych.

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	-	

W programie uwzględnione są przypadki, gdy porównywane są tylko dwie grupy, jednak na przykładach tutaj zawartych pokazywane są dane, które mają w sobie więcej niż dwie grupy.

W przypadku testu ANOVA oraz testu Kruskala-Wallisa, gdy wynikiem będzie wartość p-value mniejsza od 0.05 oznacza to, że pomiędzy porównywanymi atrybutami istnieją znaczące różnice, które należy zbadać. Wykonywane są zatem dodatkowe testy: *post hoc* Tukeya w przypadku testu ANOVA oraz *post hoc* Dunna w przypadku testu Kruskala-Wallisa.

Przykłady, gdy p-value jest większe od 0.05:

```
-----------------
>>> hsCRP <<<

[Test Kruskala-Wallisa]
Dla hsCRP pValue > 0.05, NIE istnieja znaczace roznice miedzy grupami


-----------------
>>> ERY <<<

[Test ANOVA]
Dla ERY pValue > 0.05, NIE istnieja znaczace roznice miedzy grupami


-----------------
>>> PLT <<<

[Test Kruskala-Wallisa]
Dla PLT pValue > 0.05, NIE istnieja znaczace roznice miedzy grupami
```

Przykład, gdy p-value jest mniejsze od 0.05 - test Tukeya:

```
[Test ANOVA]
Dla MCHC pValue < 0.05 ISTNIEJA znaczace roznice miedzy grupami

[Test Tukeya]

=======================================
              diff   lwr    upr  p adj
---------------------------------------
CHOR1-KONTROLA 0.726 -0.024 1.477 0.060
CHOR2-KONTROLA 1.149 0.399  1.900 0.001
CHOR2-CHOR1    0.423 -0.327 1.174 0.373
---------------------------------------
```

Oznaczenia:

- **diff** - różnica w średnich
- **lwr** - dolny próg przedziału
- **upr** - górny próg przedziału
- **p adj** - p-value po dostosowaniu dla różnych porównań; wysoka wartość (>0.05) oznacza, że wykonane porównania nie są istotne statystycznie

Test Tukeya wykonuje się, aby porównać pary średnich.
Dla każdego testu Tukeya tworzony jest odpowiedni wykres.

Przykład, gdy p-value jest mniejsze od 0.05 - test Dunna:

```
>>> HCT <<<

[Test Kruskala-Wallisa]
Dla HCT pValue < 0.05, ISTNIEJA znaczace roznice miedzy grupami

[Test Dunna]
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 7.9308, df = 2, p-value = 0.02


                          Comparison of x by group
                               (No adjustment)
Col Mean-|
Row Mean |      CHOR1      CHOR2
---------+----------------------
   CHOR2 |   0.788547
         |     0.2152
         |
KONTROLA |   2.735578   1.947030
         |    0.0031*     0.0258

alpha = 0.05
Reject Ho if p <= alpha/2
            pair mean.rank.diff       pval
1 CHOR1-KONTROLA          16.86 0.01868122
2 CHOR2-KONTROLA          12.00 0.15459306
3    CHOR2-CHOR1          -4.86 1.00000000
```

W tabeli 2x2 widoczne są wyniki testu Dunna. Wartość oznaczona gwiazdką (*) oznacza wartość dostosowaną. Największe różnice widać między średnimi w parze KONTROLA-CHOR1.

W tabeli poniżej widać różnice w średnich pomiędzy parami grup oraz p-value dla tych par.

Test Dunna również wykonuje się, aby porównać pary średnich.

Dla każdego testu Dunna tworzony jest odpowiedni wykres.

Kolejnym krokiem, który wykonuje program, jest analiza korelacji pomiędzy badanymi atrybutami dla konkretnych grup. W tym celu wykorzystuje się dwa testy:

- **Pearsona** - współczynnik korelacji liniowej, czyli parametryczny test korelacji; jeżeli dla danej grupy wartości p-value z testu Shapiro oraz p-value z testu Levene'a były większe od 0.05
- **Spearmana** - współczynnik korelacji rangowej, czyli nieparametryczny test korelacji; wykonuje się go w każdym innym przypadku.

Wynikiem każdego z tych testów są dwie wartości: **p-value** wykonanego testu oraz **estimate**, czyli współczynnik korelacji.

Na podstawie wartości p-value z każdego testu określa się korelację: jeżeli wartość p-value jest większa od 0.05, to dwa atrybuty najprawdopodobniej nie są ze sobą skorelowane. Natomiast gdy wartość p-value jest większa od 0.05, to określa się siłę korelacji:

- $-1 < r < -0.7$ bardzo silna korelacja ujemna,
- $-0.7 < r < -0.5$ silna korelacja ujemna,
- $-0.5 < r < -0.3$ korelacja ujemna o średnim natężeniu,
- $-0.3 < r < -0.2$ słaba korelacja ujemna,
- $-0.2 < r < 0.2$ brak korelacji,
- $0.2 < r < 0.3$ słaba korelacja dodatnia,
- $0.3 < r < 0.5$ korelacja dodatnia o średnim natężeniu,
- $0.5 < r < 0.7$ silna korelacja dodatnia,
- $0.7 < r < 1$ bardzo silna korelacja dodatnia,

gdzie *r* to współczynnik korelacji.

Jeżeli korelacja jest uznana za istotną (bardzo silna korelacja ujemna, silna korelacja ujemna, silna korelacja dodatnia oraz bardzo silna korelacja dodatnia), to zostają wykonane oraz zapisane wykresy przedstawiające te korelacje.

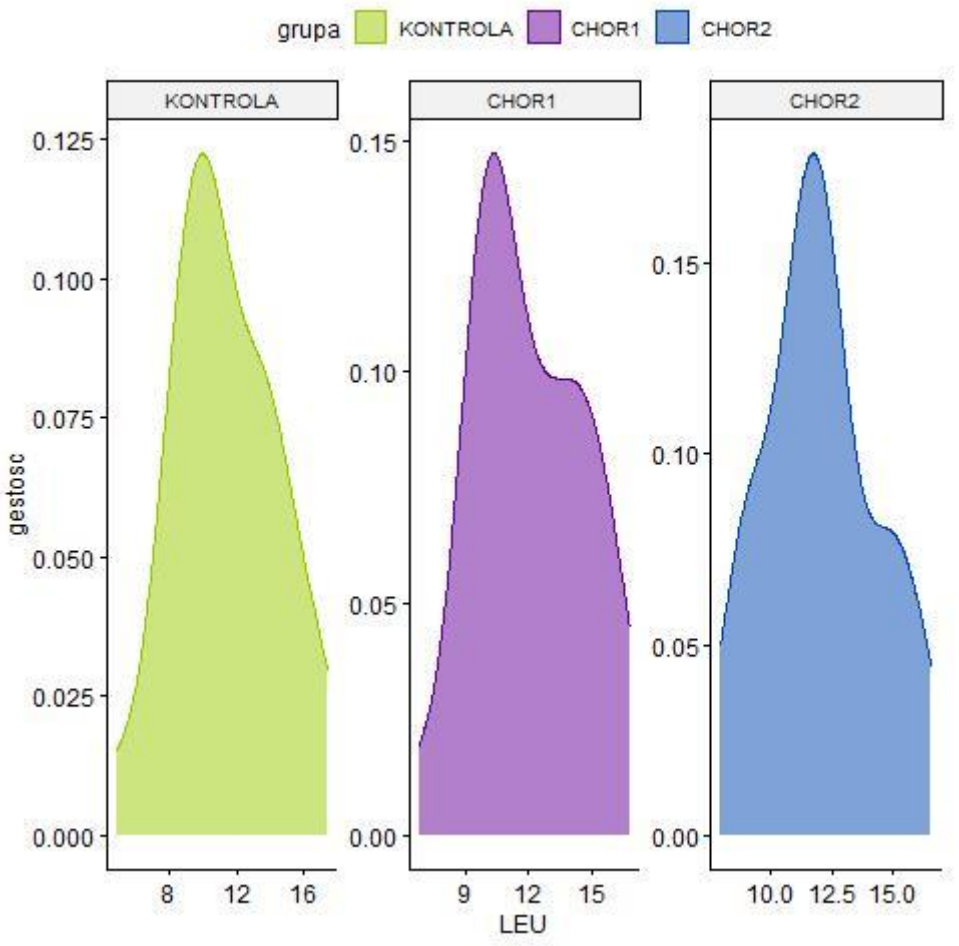
Przykładowe informacje o korelacji z pliku output.txt:

```
==================
[Spearman test]
Atrybuty ERY oraz hsCRP sa prawdopodobnie nieskorelowane


==================
[Spearman test]
Korelacja dodatnia o srednim natezeniu atrybutu ERY oraz PLT (wspolczynnik korelacji 0.454283 )


==================
[Pearson test]
Silna korelacja dodatnia atrybutu ERY oraz HGB (wspolczynnik korelacji 0.5871174 )
Utworzono wykres o nazwie ERY-and-HGB-correlationPearson.png
```

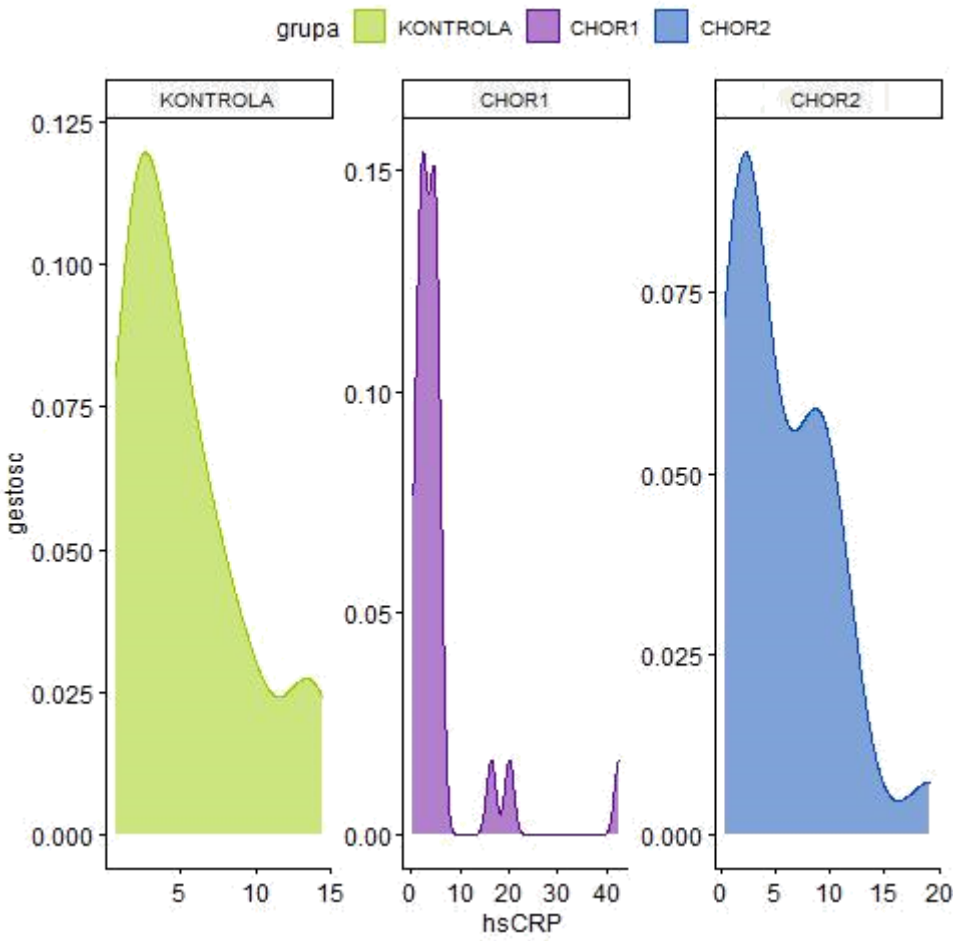
2. Wykresy zapisane w plikach .png

Wykres gęstości dla rozkładu - test Shapiro-Wilka

Rozkład normalny charakteryzuje się dzwonowatym kształtem. Przykład:



Na powyższych wykresach można zauważyć, że są one zbliżone kształtem do dzwonu, stąd można wywnioskować, że dane dla tego konkretnego atrybutu charakteryzują się rozkładem normalnym.

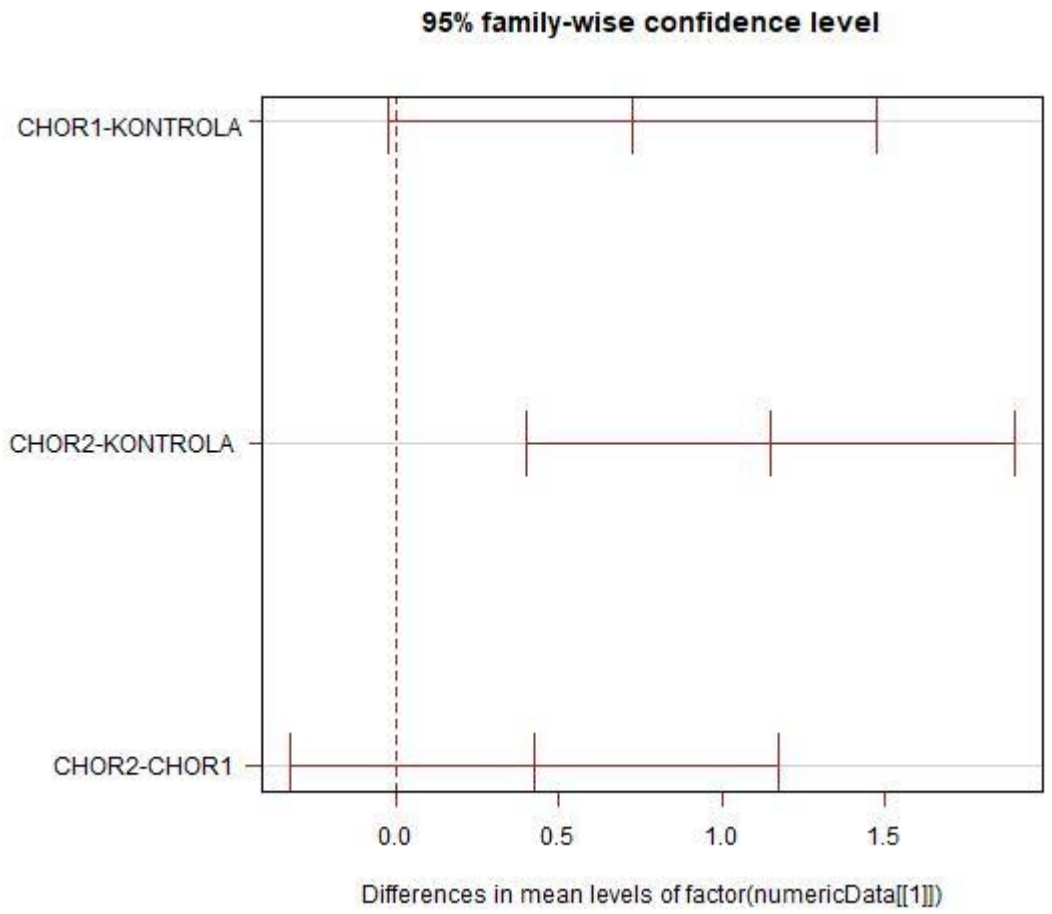


W przypadku tych wykresów nie można jednak powiedzieć, że mają one kształt zbliżony do dzwonu, stąd można wywnioskować, że dane dla tego atrybutu nie są zgodne z rozkładem normalnym.

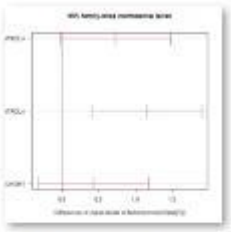
Test Tukeya

Test Tukeya pozwala na znalezienie dla danego atrybutu takich średnich z różnych par grup, które są od siebie znacząco różne. Robi się to poprzez porównanie wszystkich średnich od każdej możliwej do utworzenia pary grup.

Przykład:



Powyższy wykres przedstawia różnice w średnich dla atrybutu MCHC z przykładowo wczytanych danych. Z wykresu można wywnioskować, że średnie pomiędzy parami badanych grup nie różnią się od siebie w znaczącym stopniu. Nazwa atrybutu nie widnieje na wykresie, jednak znajduje się w nazwie pliku:

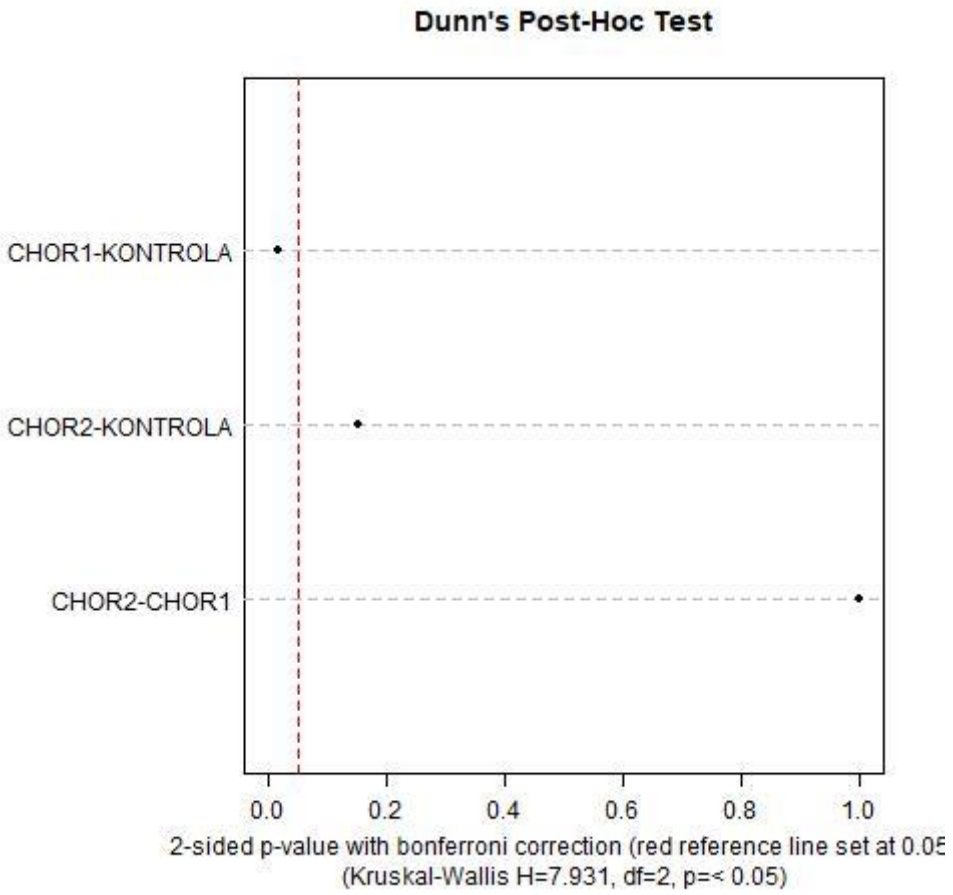


MCHC-tukeyPlot

Test Dunna

Test Dunna również jest testem porównującym średnie między parami grup, jednak przeprowadzany jest dla danych nieparametrycznych. Na wykresie wizualizującym wyniki takiego testu nie będzie podanych zakresów średnich jak to miało miejsce w przypadku testu Tukeya. Zamiast tego oznaczone są wartości p-value dla konkretnych par grup oraz wartość progowa zaznaczona czerwoną przerywaną linią.

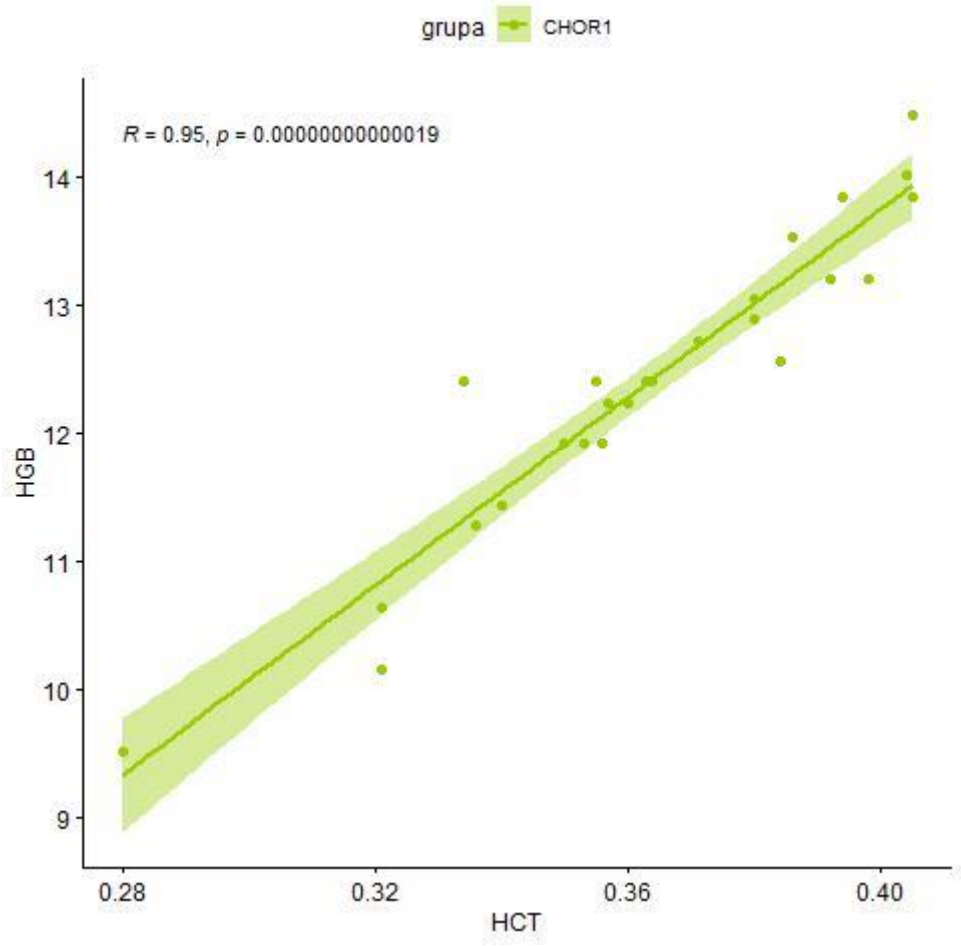
Przykład:



Powyższy wykres przedstawia wynik testu *post hoc* Dunna dla atrybutu HCT (nazwa atrybutu, tak samo jak w poprzednim przypadku, podana jest w nazwie pliku). Można z niego wywnioskować, że pomiędzy parami grup CHOR1-KONTROLA oraz CHOR2-KONTROLA istnieją różnice pomiędzy wartościami średnich, jednak są one niewielkie w porównaniu do różnic, które istnieją w średnich pomiędzy tymi dwoma parami a parą CHOR2-CHOR1.

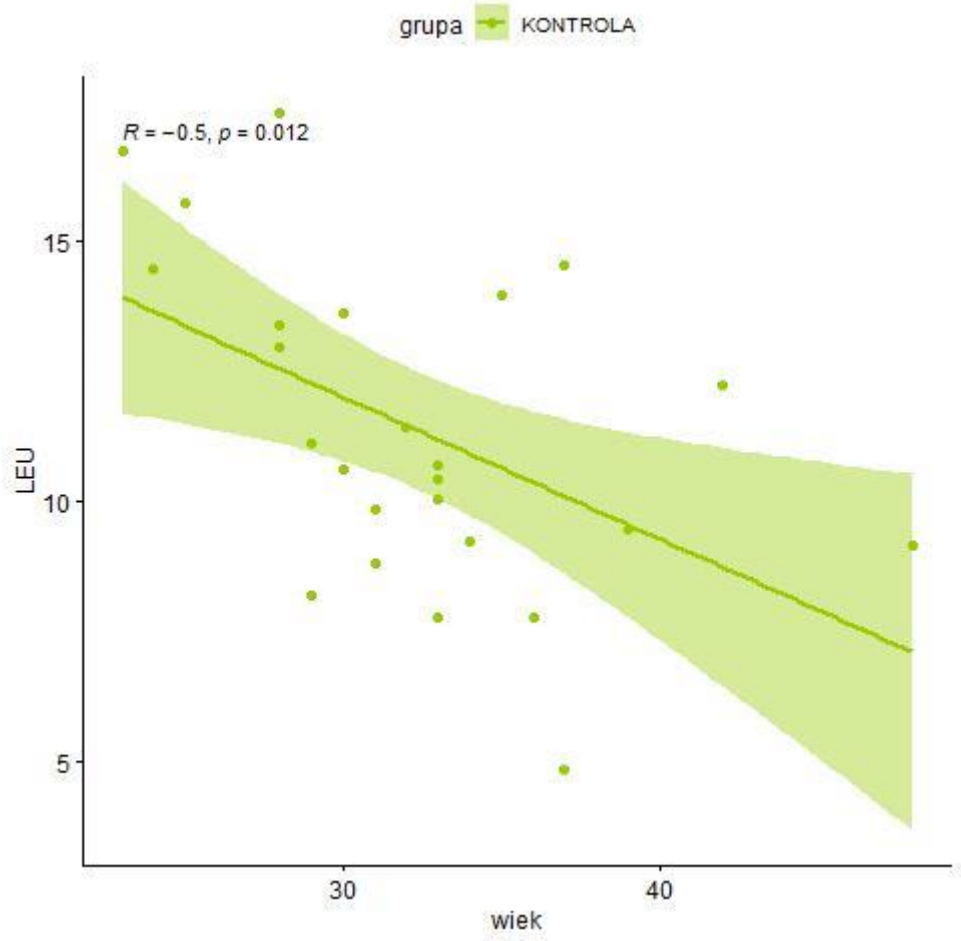
Wykresy korelacji - testy Pearsona i Spearmana

Przykład:



Na wykresie widać bardzo niewielką liczbę wartości odstających, łatwo też zauważyć, że wartości układają się mniej więcej w jedną linię. Można zatem wywnioskować, że pomiędzy tymi dwoma atrybutami istnieje bardzo silna korelacja, a kierunek wykresu (rosnący) wskazuje na korelację dodatnią.

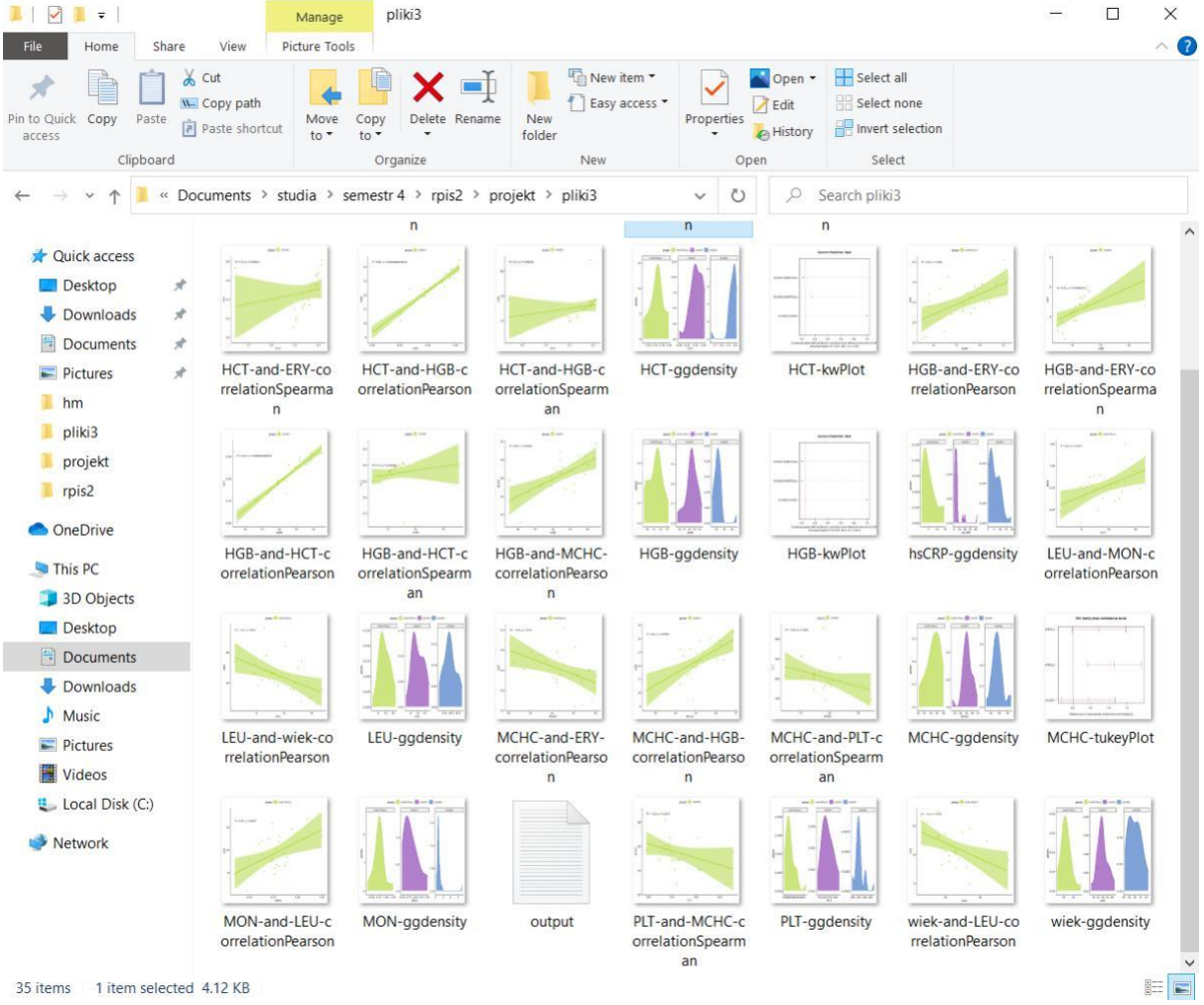
Przykład:



Na powyższym wykresie widać znacznie więcej wartości odstających, które nie tworzą jednej linii. Można zatem wywnioskować, że pomiędzy dwoma badanymi atrybutami istnieje korelacja, jednak nie jest ona bardzo silna. Po kierunku wykresu (malejący) można określić, że jest to korelacja ujemna.

Uwagi ogólne

Po wykonaniu się skryptu folder będzie zawierał m.in. takie pliki:



Przed wykonaniem kolejnej analizy sugerowane jest przeniesienie powstałych plików do osobnego folderu, w innym przypadku wykresy mogą się pomieszać, a plik output.txt zostanie nadpisany.