



**AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE**

Biblioteka rattle pakietu R

mgr inż. Beata Śpiewak

spiewak@agh.edu.pl

**Wydział Geodezji Górniczej i Inżynierii Środowiska
Katedra Geomatyki**

Kraków, 21.12.2015 r.

The R Analytical Tool To Learn Easily

Pakiet Rattle to narzędzie do eksploracji danych,
które zbiera w jednym miejscu większość
najważniejszych funkcji z obszaru data mining
dzięki czemu może konkurować z
oprogramowaniami komercyjnymi, tj. Statistica czy
Matlab.

Author Graham Williams [aut, cph, cre],

- Mark Vere Culp [cph],
- Ed Cox [ctb],
- Anthony Nolan [ctb],
- Denis White [cph],
- Daniele Medri [ctb],
- Akbar Waljee [ctb] (OOB AUC for Random Forest),
- Brian Ripley [cph] (Author of original print.summary.nnet),
- Jose Magana [ctb] (Contributed the ggpairs plots).

- Eksploracja danych to „proces automatycznego odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców, schematów, podobieństw lub trendów w dużych repozytoriach danych”[Ważniak].
- Korzenie eksploracji danych wywodzą się ze statystyki.



Data mining w Rattle

- Wczytanie danych
- Eksploracja danych
- Wykonanie testów statystycznych
- Transformacja danych w celu dopasowania do modelu
- Budowanie modelu
- Przegląd otrzymanych wyników

RGui

File Edit View Misc Packages Windows Help

R Console

	age	sex	chest_pain	trestbps	chol	fbs	restecg	thalach	exang	disease
1	31	male	asympt	120	270	f	normal	153	yes	positive
2	33	female	asympt	100	246	f	normal	150	yes	positive
3	34	male	typ_angina	140	156	f	normal	180	no	positive
4	35	male	atyp_angina	110	257	f	normal	140	no	positive
5	36	male	atyp_angina	120	267	f	normal	160	no	positive
6	37	male	asympt	140	207	f	normal	130	yes	positive
7	38	male	asympt	110	196	f	normal	166	no	positive
8	38	male	asympt	120	282	f	normal	170	no	positive
9	38	male	asympt	92	117	f	normal	134	yes	positive
10	41	male	asympt	110	289	f	normal	170	no	positive
11	43	male	asympt	150	247	f	normal	130	yes	positive
12	46	male	asympt	110	202	f	normal	150	yes	positive
13	46	male	asympt	118	186	f	normal	124	no	positive
14	46	male	asympt	120	277	f	normal	125	yes	positive
15	47	male	non_anginal	140	193	f	normal	145	yes	positive
16	47	male	asympt	150	226	f	normal	98	yes	positive
17	48	male	asympt	120	260	f	normal	115	no	positive
18	48	male	asympt	160	268	f	normal	103	yes	positive
19	49	female	non_anginal	160	180	f	normal	156	no	positive
20	49	male	non_anginal	115	265	f	normal	175	no	positive
21	49	male	asympt	130	206	f	normal	170	no	positive
22	50	female	non_anginal	140	288	f	normal	140	yes	positive
23	50	male	asympt	145	264	f	normal	150	no	positive
24	51	female	asympt	160	303	f	normal	150	yes	positive

- Age – wiek
- Sex – płeć
- Chest pain – ból zamostkowy
- Fbs – poziom cukru
- Rest ecg – wynik EKG
- Thalach – puls
- Exang – objawy w trakcie wysiłku
- Disease - choroba

Nakładka graficzna Rattle

R Data Miner - [Rattle (dane)]

Project Tools Settings Help Rattle Version 2.6.26 togaware.com

Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakńcz

Date Explore Test Transform Cluster Associate Model Evaluate Log

Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script

Data Name: dane

Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: Auto Categorical Numeric Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	age	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 38
2	sex	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
3	chest_pain	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 4
4	trestbps	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 31
5	chol	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 150
6	fbs	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
7	restecg	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 3
8	thalach	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 69
9	exang	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
10	disease	Categorical	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2



Summary

R Data Miner - [Rattle (dane)]

Project Tools Settings Help

Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakończ

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: **Summary** Distributions Correlation Principal Components Interactive

Summary Describe Basics Kurtosis Skewness Show Missing Cross Tab

Below we summarise the dataset.

The data is limited to the training dataset.

Data frame: crs\$dataset[crs\$sample, c(crs\$input, crs\$risk, crs\$target)] 200 observations and 10 variables Maximum # NAs:0

	Levels	Storage
age		integer
sex	2	integer
chest_pain	4	integer
trestbps		integer
chol		integer
fbs	2	integer
restecg	3	integer
thalach		integer
exang	2	integer
disease	2	integer

```
+-----+-----+
|Variable|Levels|
+-----+-----+
|sex     |female,male|
+-----+-----+
|chest_pain|asympt,atyp_angina,non_anginal,typ_angina|
+-----+-----+
|fbs     |f,t|
+-----+-----+
|restecg |left_vent_hyper,normal,st_t_wave_abnormality|
+-----+-----+
|exang   |no,yes|
+-----+-----+
|disease |negative,positive|
+-----+-----+
```



R Data Miner - [Rattle (dane)]

Project Tools Settings Help

Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakończ

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: Summary Distributions Correlation Principal Components Interactive

Summary Describe Basics Kurtosis Skewness Show Missing Cross Tab

```
age          2 integer
disease      2 integer

+-----+-----+
|Variable |Levels |
+-----+-----+
|sex      |female,male |
+-----+-----+
|chest_pain|asympt,atyp_angina,non_anginal,typ_angina |
+-----+-----+
|fbs      |f,t |
+-----+-----+
|restecg  |left_vent_hyper,normal,st_t_wave_abnormality|
+-----+-----+
|exang    |no,yes |
+-----+-----+
|disease  |negative,positive |
+-----+-----+

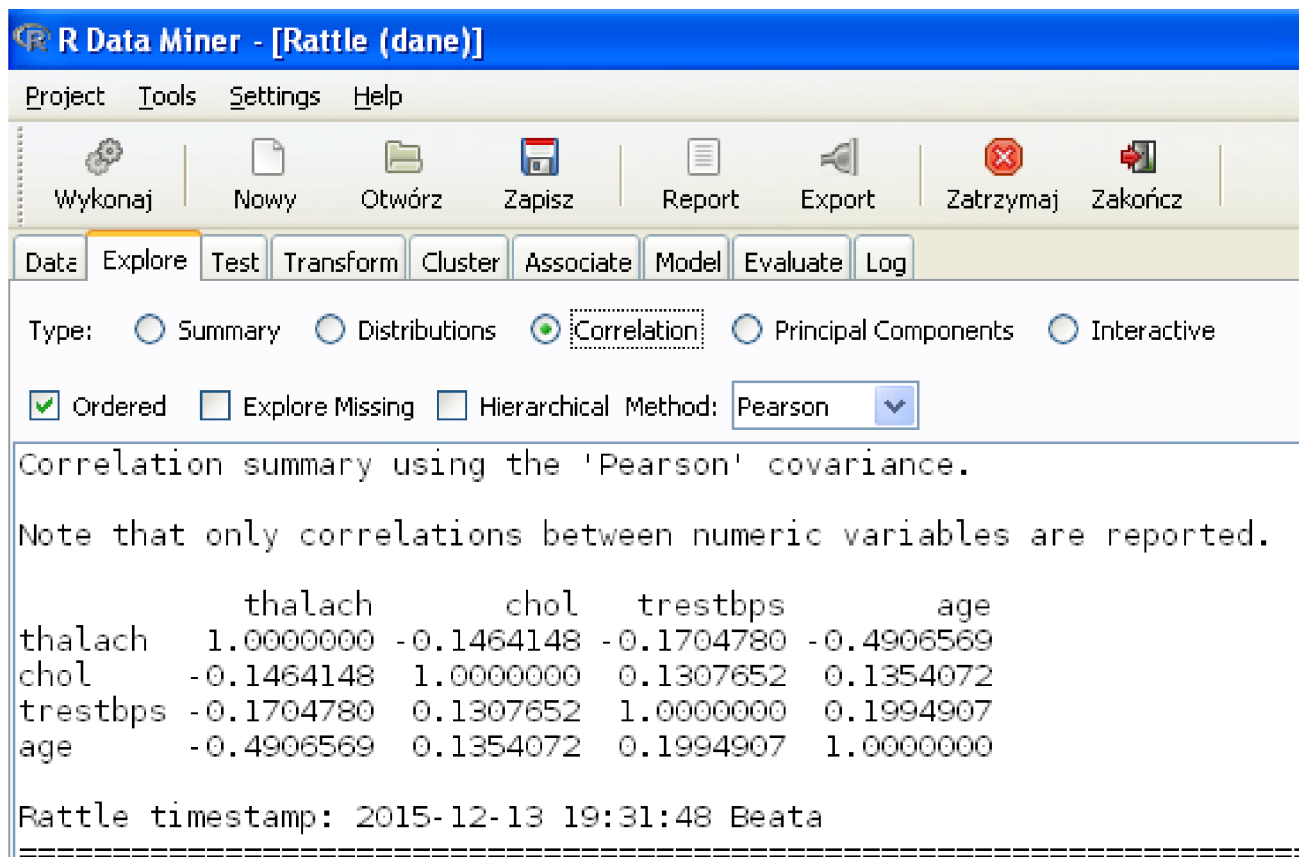
For the simple distribution tables below the 1st and 3rd Qu.
refer to the first and third quartiles, indicating that 25%
of the observations have values of that variable which are
less than or greater than (respectively) the value listed.

      age          sex          chest_pain      trestbps          chol
Min.   :28.00  female: 53  asympt      :81  Min.   : 98.0  Min.   :132.0
1st Qu.:42.00  male  :147  atyp_angina:77  1st Qu.:120.0  1st Qu.:211.0
Median :49.00                non_anginal:34  Median :130.0  Median :250.0
Mean   :48.27                typ_angina : 8  Mean   :133.7  Mean   :252.0
3rd Qu.:54.00                Max.   :190.0  3rd Qu.:140.0  3rd Qu.:277.5
Max.   :65.00                Max.   :190.0  Max.   :603.0

fbs      restecg          thalach          exang          disease
f:184   left_vent_hyper      : 5  Min.   : 82.0  no :142  negative:135
t: 16   normal              :157 1st Qu.:122.0  yes: 58  positive: 65
        st_t_wave_abnormality: 38  Median :140.0
        Mean   :139.6
        3rd Qu.:155.2
        Max.   :190.0

Rattle timestamp: 2015-12-13 19:29:52 Beata
=====
```


Korelacja pomiędzy zmiennymi



R Data Miner - [Rattle (dane)]

Project Tools Settings Help

Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakończ

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: Summary Distributions Correlation Principal Components Interactive

Ordered Explore Missing Hierarchical Method: Pearson

Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

	thalach	chol	trestbps	age
thalach	1.0000000	-0.1464148	-0.1704780	-0.4906569
chol	-0.1464148	1.0000000	0.1307652	0.1354072
trestbps	-0.1704780	0.1307652	1.0000000	0.1994907
age	-0.4906569	0.1354072	0.1994907	1.0000000

Rattle timestamp: 2015-12-13 19:31:48 Beata

=====

Explore - > Distribution

R Data Miner - [Rattle (dane)]

Project: Tools Settings Help

Rattle Version 2.6.26 ogamare.com

Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakończ

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: Summary Distributions Correlation Principal Components Interactive

Numeric: Wyczyść Plots per Page: 4 Annotate Target: disease

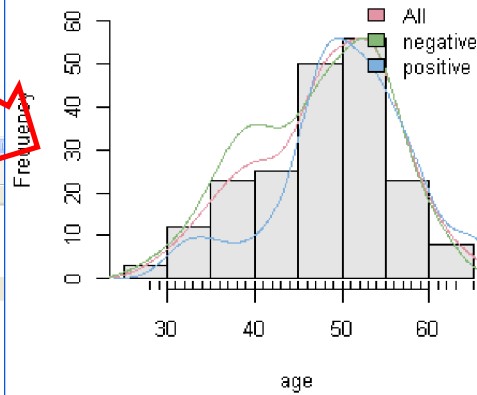
Benford Bars Benford Digit: 1 abs +ve -ve

No. Variable	Box Plot	Histogram	Cumulative	Benford	Min; Median/Mean; Max
1 age	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	28.00; 49.00/47.85; 66.00
4 trestbps	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	92.00; 130.00/132.75; 200.00
5 chol	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	65.00; 248.00/249.37; 603.00
8 thalach	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	62.00; 140.00/139.16; 190.00

Categoric: Wyczyść

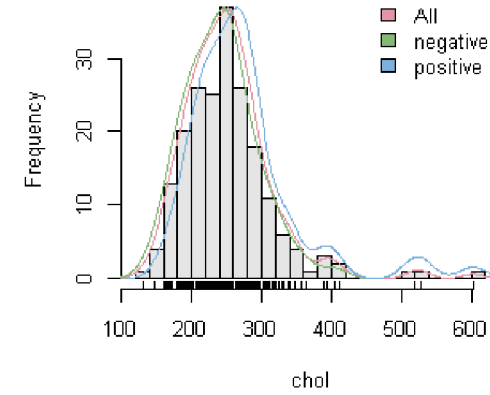
No. Variable	Bar Plot	Dot Plot	Mosaic	Levels
2 sex	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	2
3 chest_pain	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4
6 fbs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2
7 restecg	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	3
9 exang	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2
10 disease	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2

Distribution of age (sample) by disease



Rattle 2015-gru-13 22:35:48 Beata

Distribution of chol (sample) by disease



Rattle 2015-gru-13 22:35:48 Beata

Explore - > Distribution

R Data Miner - [Rattle (dane)]

Project Tools Settings Help Rattle Version 2.6.26 toqaware.com

Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakończ

Dane Explore Test Transform Cluster Associate Model Evaluate Log

Type: Summary Distributions Correlation Principal Components Interactive

Numeric: Wyczyść Plots per Page: 4 Annotate

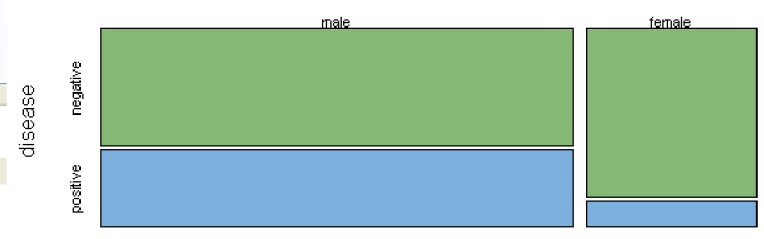
Benford Bars Benford Digit: 1 abs +ve -ve

No. Variable	Box Plot	Histogram	Cumulative	Benford	Min; Median/Mean; Max
1 age	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	28.00; 49.00/47.85; 66.00
4 trestbps	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	92.00; 130.00/132.75; 200.00
5 chol	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	85.00; 248.00/249.37; 603.00
8 thalach	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	82.00; 140.00/139.18; 190.00

Category: Wyczyść

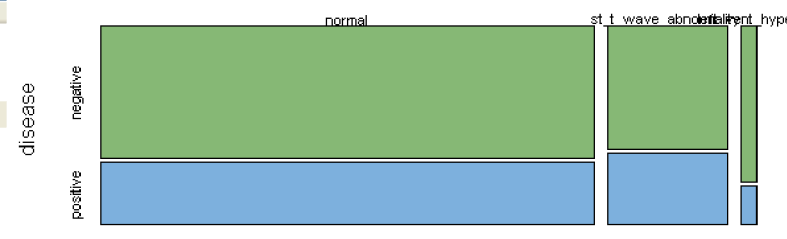
No. Variable	Bar Plot	Dot Plot	Mosaic	Levels
2 sex	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	2
3 chest_pain	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4
6 fbs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2
7 restecg	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	3
9 exang	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2
10 disease	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2

Mosaic of sex (sample) by disease



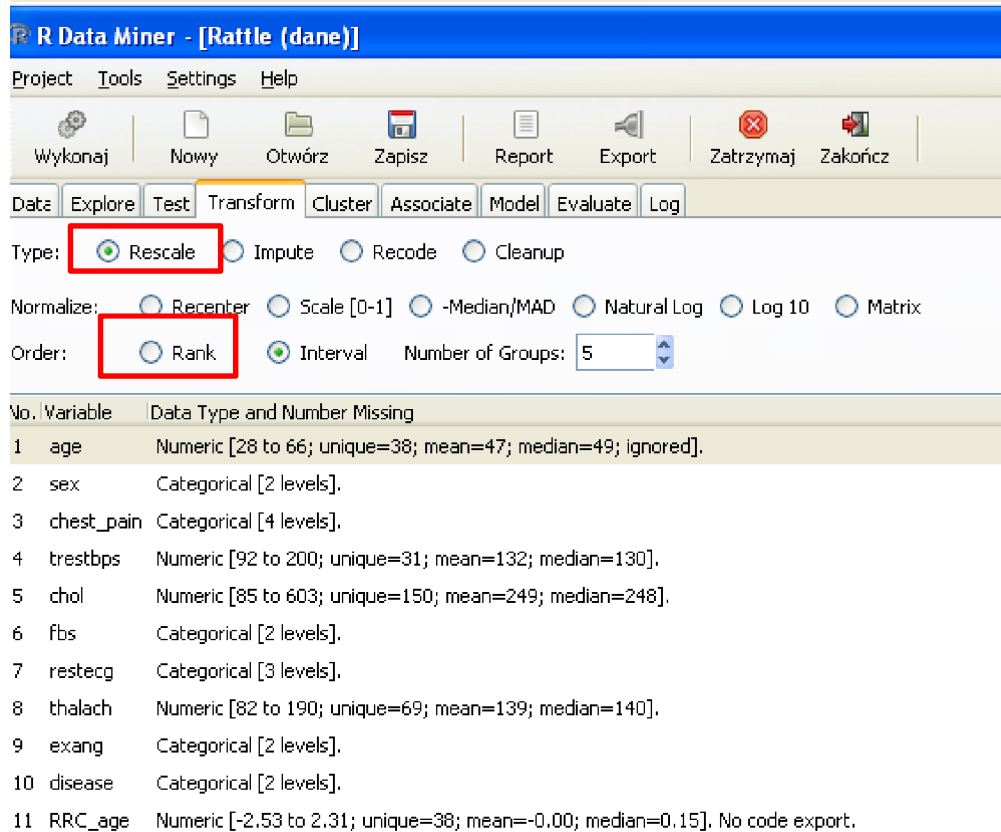
sex
Rattle 2015-gru-13 22:37:05 Beata

Mosaic of restecg (sample) by disease



restecg
Rattle 2015-gru-13 22:37:05 Beata

Transform - > Rescale/Order



The screenshot shows the R Data Miner interface with the Transform menu open. The 'Rescale' option is selected under the 'Type' section, and the 'Rank' option is selected under the 'Order' section. The 'Number of Groups' is set to 5.

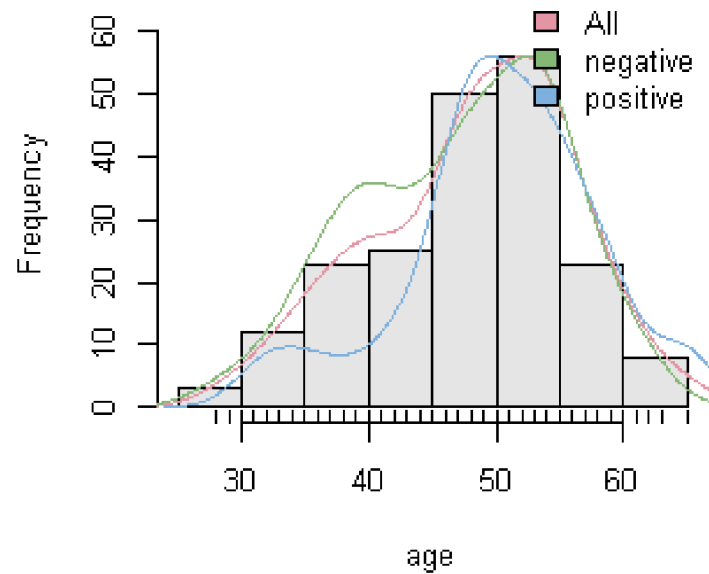
No.	Variable	Data Type and Number Missing
1	age	Numeric [28 to 66; unique=38; mean=47; median=49; ignored].
2	sex	Categorical [2 levels].
3	chest_pain	Categorical [4 levels].
4	trestbps	Numeric [92 to 200; unique=31; mean=132; median=130].
5	chol	Numeric [85 to 603; unique=150; mean=249; median=248].
6	fbs	Categorical [2 levels].
7	restecg	Categorical [3 levels].
8	thalach	Numeric [82 to 190; unique=69; mean=139; median=140].
9	exang	Categorical [2 levels].
10	disease	Categorical [2 levels].
11	RRC_age	Numeric [-2.53 to 2.31; unique=38; mean=-0.00; median=0.15]. No code export.

- Przeskalowanie
- Redukcja skośności
- Uzupełnienie brakujących danych
- Zmiana danych ilościowych na jakościowe



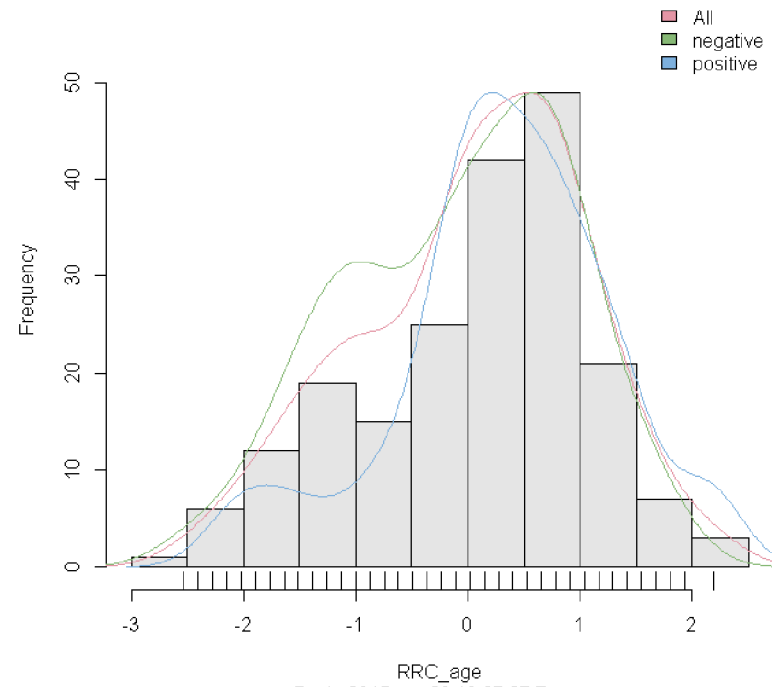
Transform - > Rescale

Distribution of age (sample) by disease



Rattle 2015-gru-13 22:35:48 Beata

Distribution of RRC_age (sample) by disease



Rattle 2015-gru-20 18:27:27 Beata



Cluster – GRUPOWANIE

forma automatycznego łączenia (grupowania) obiektów, które są do siebie podobne według określonych kryteriów

```
R Data Miner - [Rattle (dane)]
Project Tools Settings Help
Wykonaj Nowy Otwórz Zapisz Report Export Zatrzymaj Zakończ
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type:  KMeans  Ewkm  Hierarchical  BiCluster
Number of clusters: 10 Seed: 42 Runs: 1
 Use HClust Centers  Iterate Clusters  Plots:   
Cluster sizes:
[1] "13 8 37 26 13 30 3 23 36 11"
Data means:
      trestbps      chol      thalach      RRC_age
133.65500000 252.00500000 139.61000000 0.05466617
Cluster centers:
      trestbps      chol      thalach      RRC_age
1 144.6154 174.0000 154.3077 -0.10786151
2 121.0000 163.1250 179.8750 -1.20731348
3 122.2973 210.9189 147.1892 -0.20088212
4 145.3462 211.0000 114.9615 0.55401593
5 150.7692 364.6923 128.2308 0.52950195
6 135.0000 270.5000 111.6667 0.66122374
7 131.0000 550.0000 128.3333 -0.19284330
8 135.7391 304.5217 153.9565 -0.13003067
9 122.8889 256.4167 144.8333 -0.01933881
10 148.1818 254.8182 171.3636 -0.67569441
within cluster sum of squares:
[1] 7740.733 2601.825 16769.370 19519.001 21442.779 24632.884 4661.682
[8] 16915.788 13562.898 4762.522
Time taken: 0.06 secs
Rattle timestamp: 2015-12-20 21:43:06 Beata
=====
```



Podsumowanie

- Powszechnie dostępne narzędzie
- Nie wymaga umiejętności programowania
- Graficzny interfejs ułatwiający korzystanie z pakietu
- Pozwala na ulepszanie swoich możliwości przez zaawansowanych użytkowników



Bibliografia

- <https://cran.r-project.org/web/packages/rattle/index.html>
- <https://cran.r-project.org/web/packages/rattle/rattle.pdf>
- http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/heart_for_rattle.txt
- <http://rattle.togaware.com/>



Dziękuję za uwagę!