

SPRAWOZDANIE

Zajęcia: Zbiory Big Data i Eksploracja Danych

Prowadząca: dr inż. Ruslana Ziubina

Laboratorium nr 4 Data rozpoczęcia: 1.12.2023 Temat: Rattle – początki pracy	Rafał Klinowski Informatyka II stopień, stacjonarne, Semestr 2, gr. a
--	--

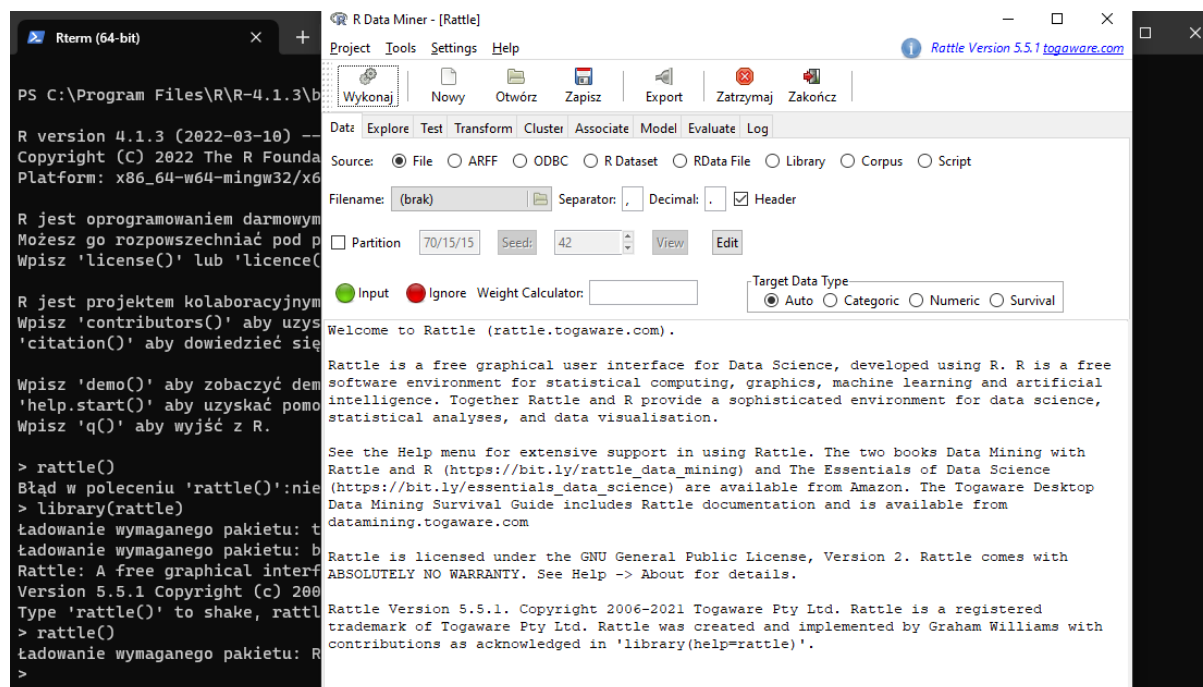
Poszczególne ćwiczenia będą wykonywane w pliku źródłowym edytowanym przy pomocy środowiska RStudio oraz Rattle, opisanego w poprzednich laboratoriach.

Ćw. 1.

Pierwszym zadaniem była instalacja pakietu Rattle wraz z jego zależnościami. Podobnie jak w ramach Laboratorium 1, wymagane było skorzystanie z wcześniejszej wersji R (4.1.3) oraz pobranie pakietów z CRAN. W szczególności problemy występowały z pakietem RGtk2 oraz stringi.

Ćw. 2.

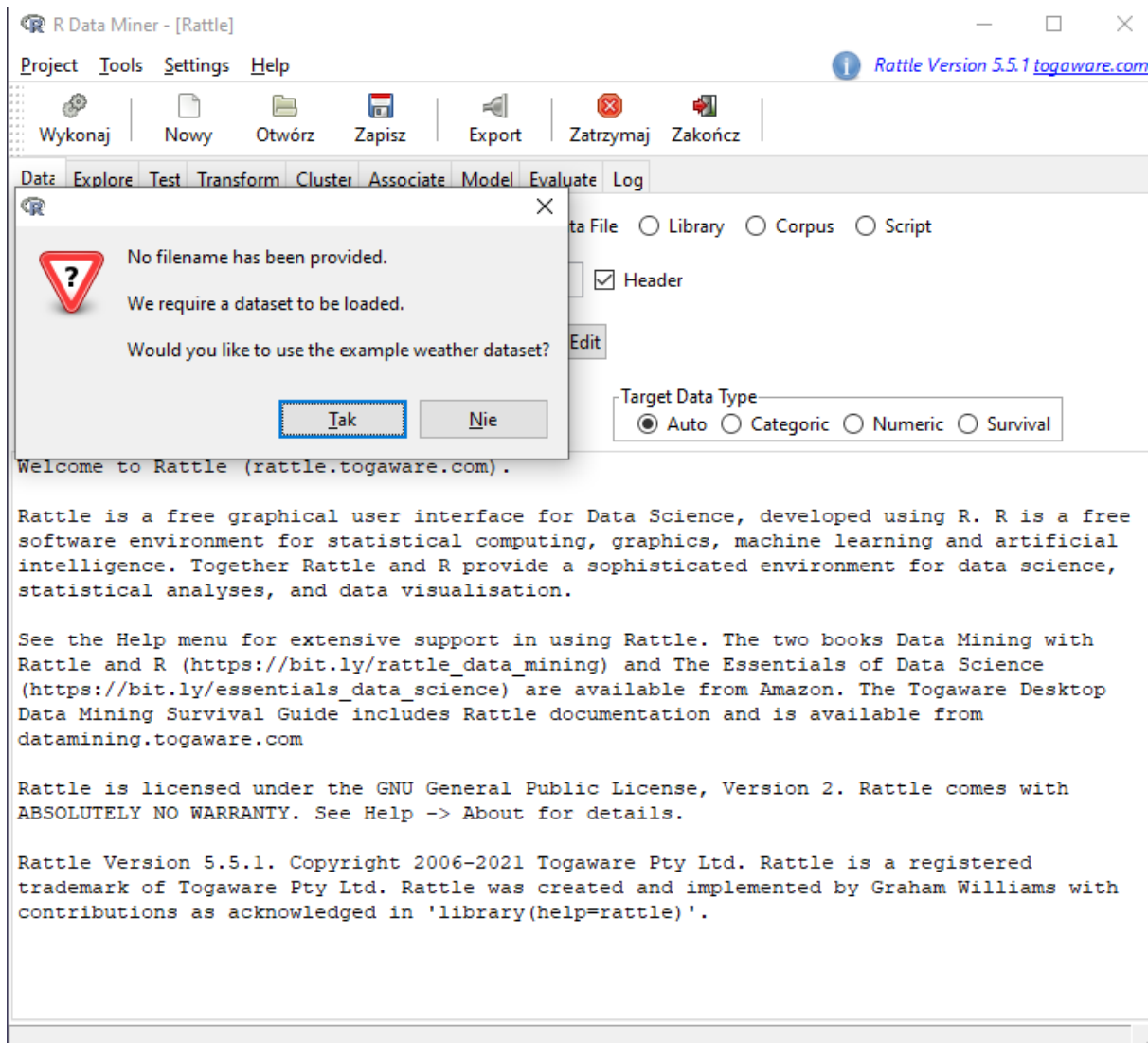
Po zainstalowaniu pakietów uruchomiono środowisko Rattle.



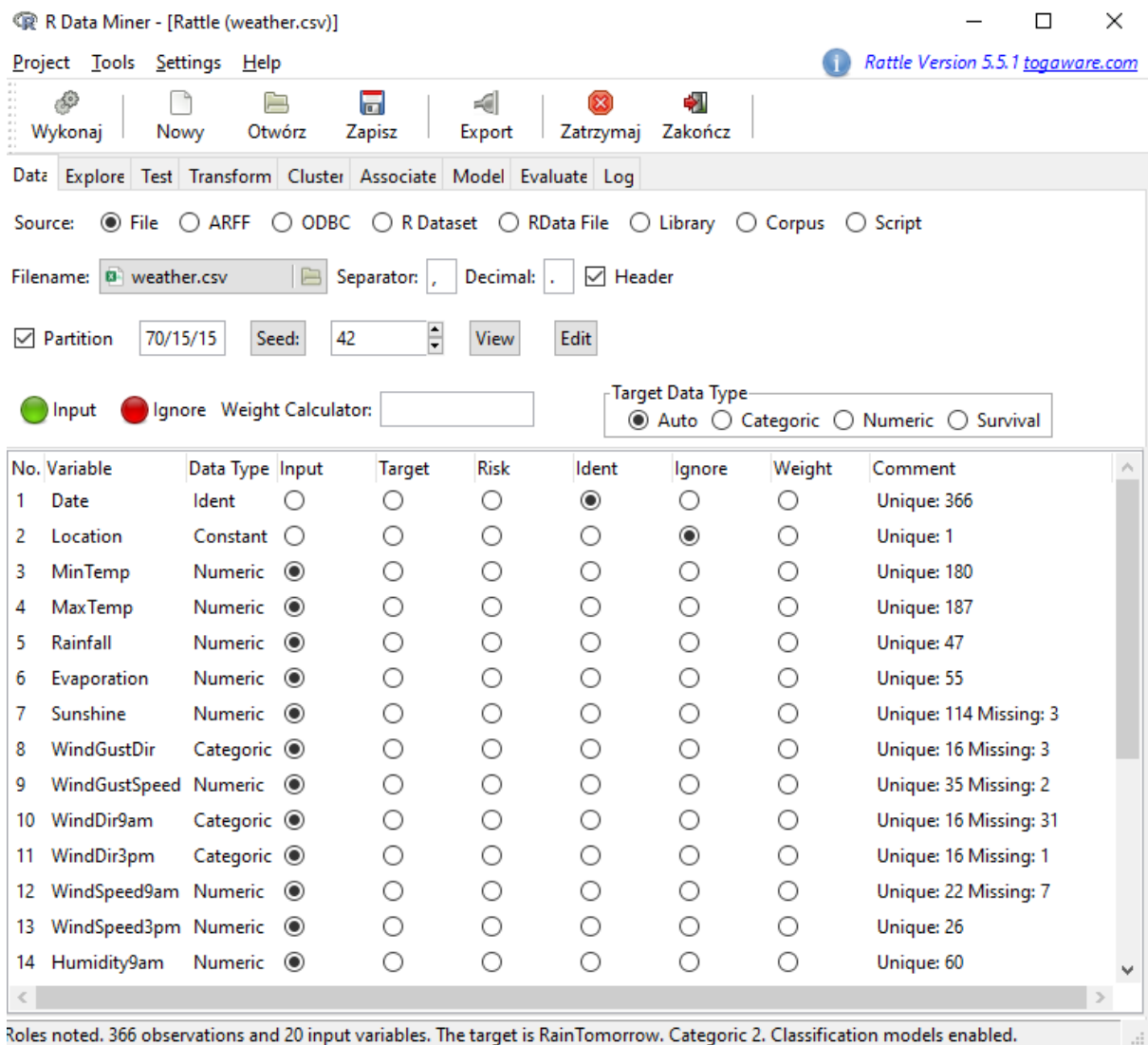
Rysunek 1. Środowisko Rattle.

Ćw. 3.

Wczytanie przykładowego zbioru danych w Rattle.



Rysunek 2. Ekran po kliknięciu Execute.



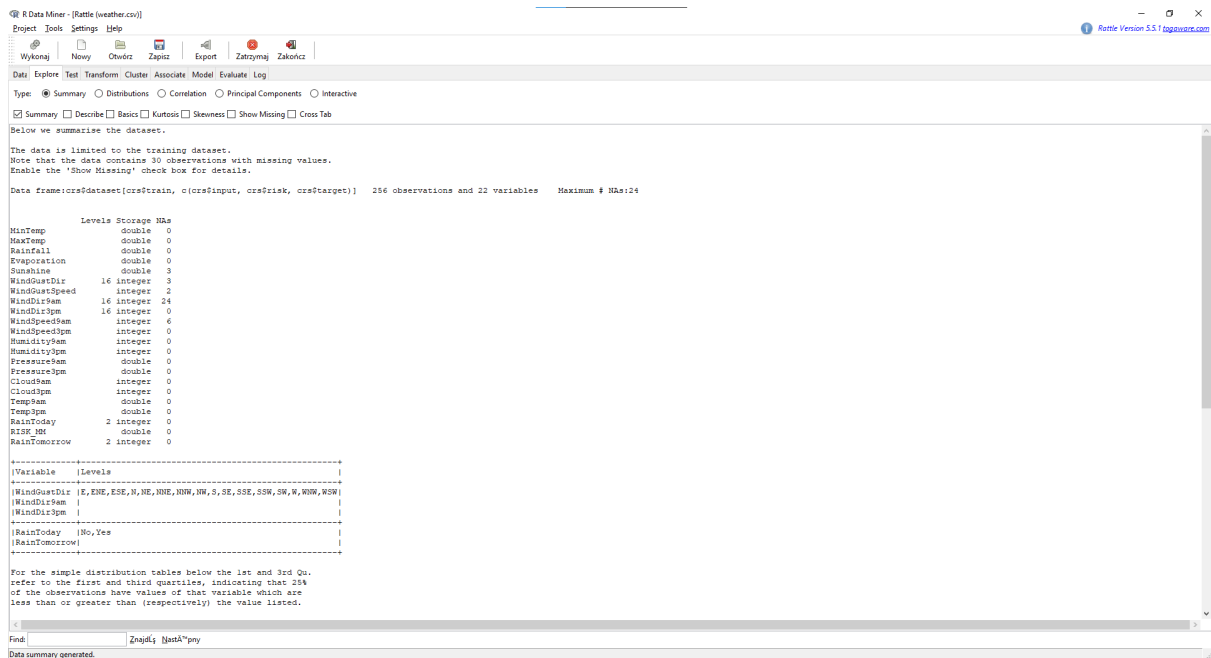
Rysunek 3. Wygląd wczytanego zbioru danych.

Ćw. 4.

Proste podsumowanie tekstowe zbioru danych Weather.

```
> data("weather")
> summary(weather[7:9])
      sunshine      windGustDir      windGustSpeed
Min.   : 0.000      NW       : 73      Min.   :13.00
1st Qu.: 5.950      NNW      : 44      1st Qu.:31.00
Median : 8.600      E       : 37      Median :39.00
Mean   : 7.909      WNW     : 35      Mean   :39.84
3rd Qu.:10.500     ENE     : 30      3rd Qu.:46.00
Max.   :13.600     (Other):144     Max.   :98.00
NA's   :3          NA's    : 3      NA's   :2
```

Rysunek 4. Podsumowanie zbioru danych w R.



Rysunek 5. Podsumowanie zbioru danych w Rattle.

Ćw. 5.

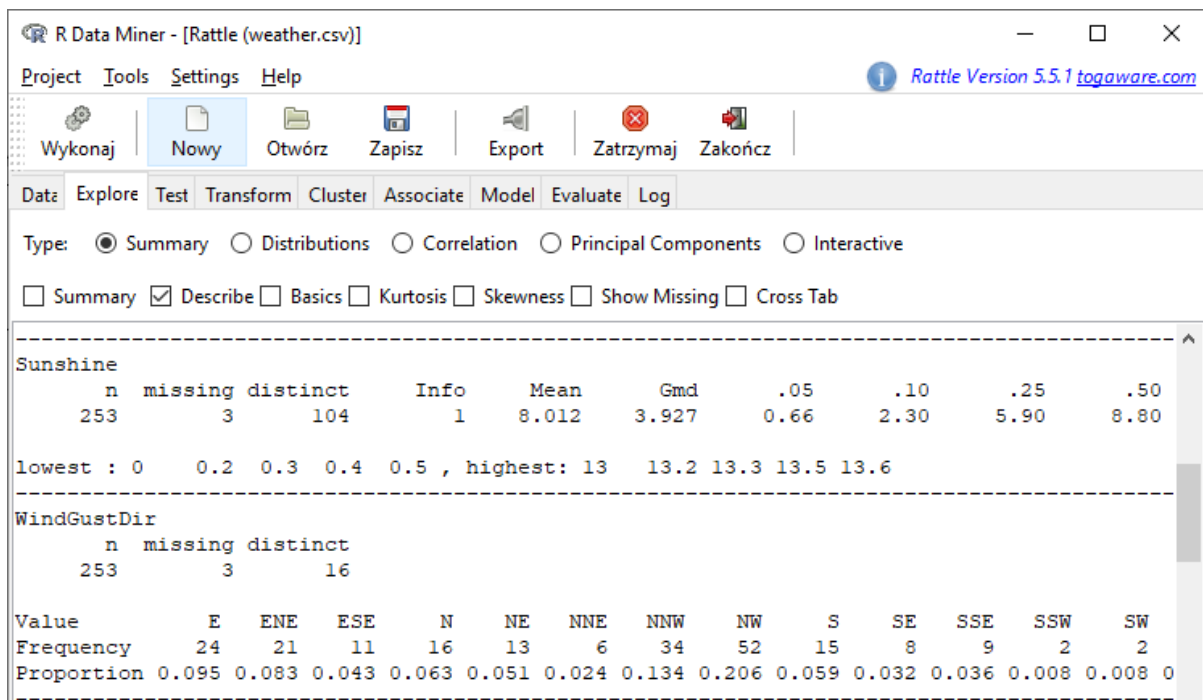
Podsumowanie danych za pomocą Hmisc. Konieczne było najpierw zainstalowanie pakietu.

```
> describe(weather[7])
weather[7]

1 variables      366 observations
-----
Sunshine
  n missing distinct    Info    Mean     Gmd   .05   .10   .25   .50   .75   .90   .95
363      3       114      1  7.909   3.875  0.60  2.04  5.95  8.60 10.50 11.80 12.60

lowest : 0    0.1 0.2 0.3 0.4 , highest: 13.1 13.2 13.3 13.5 13.6
-----
> weather[7]
# A tibble: 366 x 1
  sunshine
  <dbl>
1     6.3
2     9.7
3     3.3
4     9.1
5    10.6
6     8.2
7     8.4
8     4.6
9     4.1
10    7.7
# i 356 more rows
# i Use `print(n = ...)` to see more rows
```

Rysunek 6. Podsumowanie danych przy pomocy Hmisc w R.



Rysunek 7. Podsumowanie w Rattle.

Ćw. 6.

Podsumowanie numeryczne z wykorzystaniem pakietu fBasics.

```
> basicStats(weather$Sunshine)
X..weather.Sunshine
nobs          366.000000
NAS           3.000000
Minimum       0.000000
Maximum      13.600000
1. Quartile   5.950000
3. Quartile  10.500000
Mean          7.909366
Median        8.600000
Sum          2871.100000
SE Mean       0.182732
LCL Mean      7.550016
UCL Mean      8.268716
Variance      12.120962
Stdev         3.481517
Skewness      -0.723454
Kurtosis      -0.270625
```

Rysunek 8. Podstawowe statystyki w R.

```

> skewness(weather[,c(7,9,12,13)], na.rm=TRUE)
      Sunshine windGustSpeed windSpeed9am windSpeed3pm
-0.7234543      0.8361055      1.3601713      0.5912721
attr(,"method")
[1] "moment"

> kurtosis(weather[,c(7,9,12,13)], na.rm=TRUE)
      Sunshine windGustSpeed windSpeed9am windSpeed3pm
-0.2706248      1.4761027      1.4758254      0.1963276
attr(,"method")
[1] "excess"

```

Rysunek 9. Skośność i kurtoza dla przykładowych danych w R.

R Data Miner - [Rattle (weather.csv)]

Project

Tools

Settings

Help

Wykonaj

Nowy

Otwórz

Zapisz

Export

Zatrzymaj

Zakończ

Rattle Version 5.5 / [help@rattle.org](#)

Date

Explore

Test

Transform

Cluster

Associate

Model

Evaluate

Log

Type:

☒ Summary

☐ Distributions

☐ Correlation

☐ Principal Components

☐ Interactive

☐ Summary

☐ Describe

☒ Basics

☐ Kurtosis

☐ Skewness

☐ Show Missing

☐ Cross Tab

Basic statistics for each numeric variable of the dataset.

\$MinTemp

nbobs	256.000000
NA's	0.000000
Minimum	-5.300000
Maximum	18.000000
1. Quartile	2.100000
3. Quartile	12.500000
Mean	7.165645
Median	7.450000
Sum	1834.400000
SE Mean	0.376593
LCL Mean	6.423996
UCL Mean	7.907254
Variance	36.306500
Stdev	6.025488
Skewness	-0.033284
Kurtosis	-1.229643

\$MaxTemp

nbobs	256.000000
NA's	0.000000
Minimum	7.600000
Maximum	35.800000
1. Quartile	14.800000
3. Quartile	28.550000
Mean	20.419922
Median	19.750000
Sum	5227.500000
SE Mean	0.418966
LCL Mean	19.594786
UCL Mean	21.245058
Variance	48.943052
Stdev	6.703961
Skewness	0.233501
Kurtosis	-0.800890

\$Rainfall

nbobs	256.000000
NA's	0.000000
Minimum	0.000000
Maximum	39.200000
1. Quartile	0.000000
3. Quartile	0.200000
Mean	1.535937

Find

Znajdź

WzrostA"pny

Rysunek 10. Podsumowanie numeryczne w Rattle.

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help

Wykonaj Nowy Otwórz Zapisz 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

Kurtosis for each numeric variable of the dataset.
Larger values mean sharper peaks and flatter tails.
Positive values indicate an acute peak around the mean.
Negative values indicate a smaller peak around the mean.

MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm
-1.22964310	-0.80088975	26.77250290	-0.46167304	-0.35765926	1.03874723	1.45080067	-0.05047269
Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
-0.12025080	-0.05784140	-0.32166065	-0.20762695	-1.72551766	-1.62760983	-1.00241317	-0.72522759

RISK_MM
44.42728670
attr(,"method")
[1] "excess"

Rattle timestamp: 2023-12-01 21:21:10 Rafał Klinowski

Skewness for each numeric variable of the dataset.
Positive means the right tail is longer.

MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm
-0.03328628	0.29390080	4.52615980	0.60345452	-0.69741629	0.74972599	1.36020603	0.54486643
Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
-0.12539639	0.57503042	-0.26378688	-0.18833373	0.14043694	0.11595010	-0.01195378	0.25833694

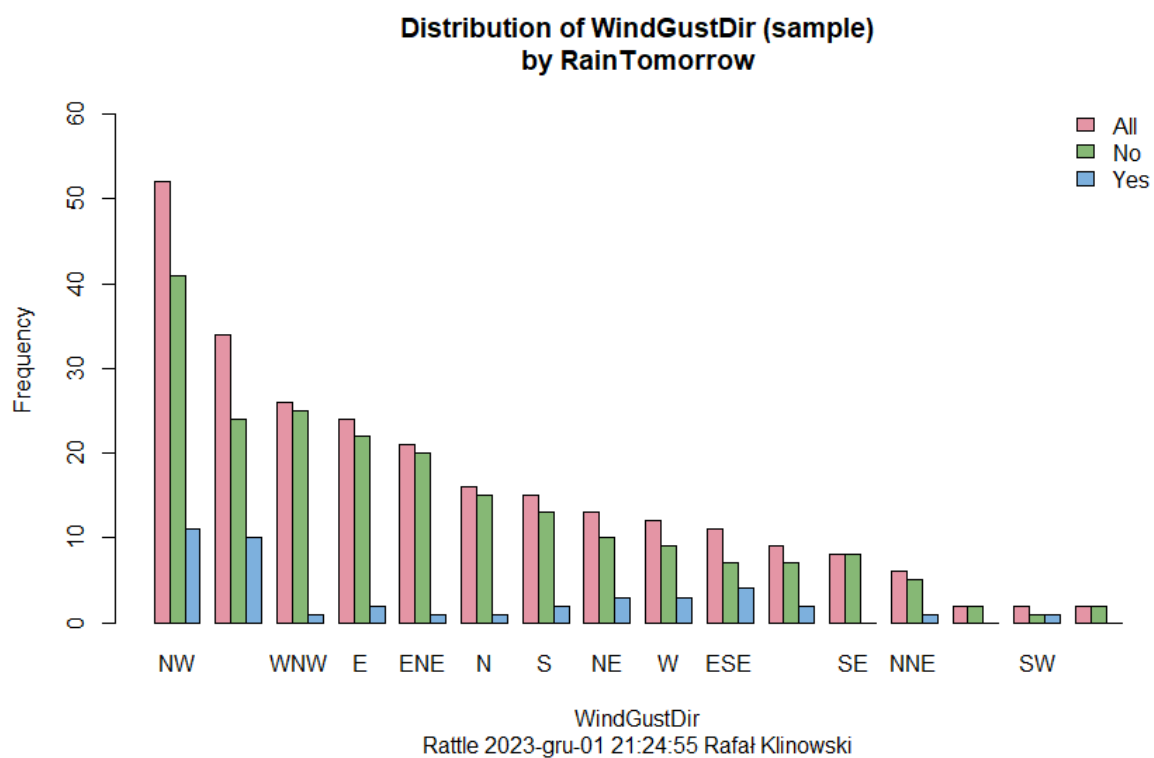
RISK_MM
5.94245103
attr(,"method")
[1] "moment"

Rattle timestamp: 2023-12-01 21:21:10 Rafał Klinowski

Rysunek 11. Skośność i kurtoza w Rattle.

Ćw. 7.

Wykres słupkowy w Rattle.



Rysunek 12. Wykres słupkowy w Rattle.

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 5.5.1 togaware.com

Wykonaj Nowy Otwórz Zapisz Export Zatrzymaj Zakończ

Datę Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☐ Summary ☒ Distributions ☐ Correlation ☐ Principal Components ☐ Interactive

Numeric: ☐ Annotate Group By: RainTomorrow

Benfords: ☐ Bars Starting Digit: 1 Digits: 1 ☒ abs ☐ +ve ☐ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Pairs	Min; Median/Mean; Max
3	MinTemp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	-5.30; 7.45/7.27; 20.90
4	MaxTemp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	7.60; 19.65/20.55; 35.80
5	Rainfall	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.00; 0.00/1.43; 39.80
6	Evaporation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.20; 4.20/4.52; 13.80
7	Sunshine	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.00; 8.60/7.91; 13.60
9	WindGustSpeed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	13.00; 39.00/39.84; 98.00
12	WindSpeed9am	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.00; 7.00/9.65; 41.00

Category: Wyyczyść

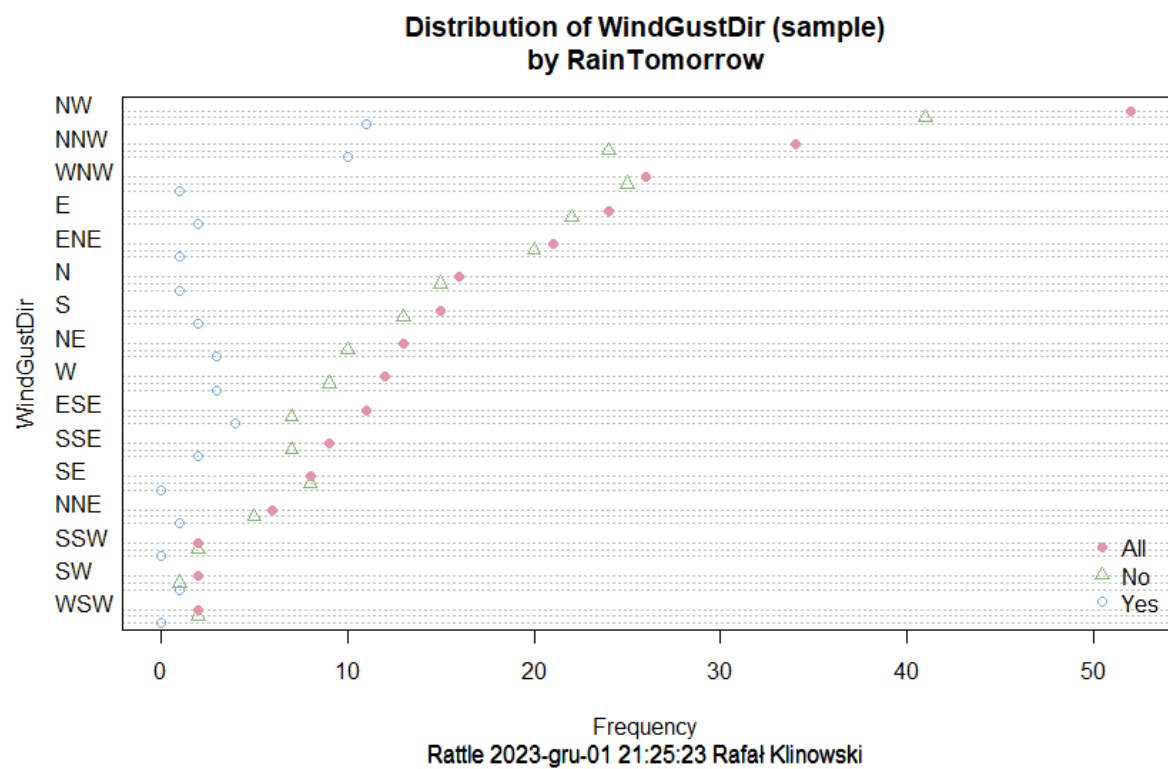
No.	Variable	Bar Plot	Dot Plot	Mosaic	Pairs	Levels
8	WindGustDir	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16
10	WindDir9am	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16
11	WindDir3pm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16
22	RainToday	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2
24	RainTomorrow	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2

One plot has been generated.

Rysunek 13. Parametry do utworzenia wykresu w Rattle.

Ćw. 8.

Wykres kropkowy w Rattle.



Rysunek 14. Wykres kropkowy utworzony w Rattle.

R Data Miner - [Rattle (weather.csv)]

Project Tools Settings Help Rattle Version 5.5.1 togaware.com

Wykonaj Nowy Otwórz Zapisz Export Zatrzymaj Zakończ

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☐ Summary ☒ Distributions ☐ Correlation ☐ Principal Components ☐ Interactive

Numeric: ☐ Annotate Group By: RainTomorrow

Benfords: ☐ Bars Starting Digit: 1 Digits: 1 ☒ abs ☐ +ve ☐ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Pairs	Min; Median/Mean; Max
3	MinTemp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	-5.30; 7.45/7.27; 20.90
4	MaxTemp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	7.60; 19.65/20.55; 35.80
5	Rainfall	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.00; 0.00/1.43; 39.80
6	Evaporation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.20; 4.20/4.52; 13.80
7	Sunshine	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.00; 8.60/7.91; 13.60
9	WindGustSpeed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	13.00; 39.00/39.84; 98.00
12	WindSpeed9am	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.00; 7.00/9.65; 41.00

Categoric: Wyżżyć

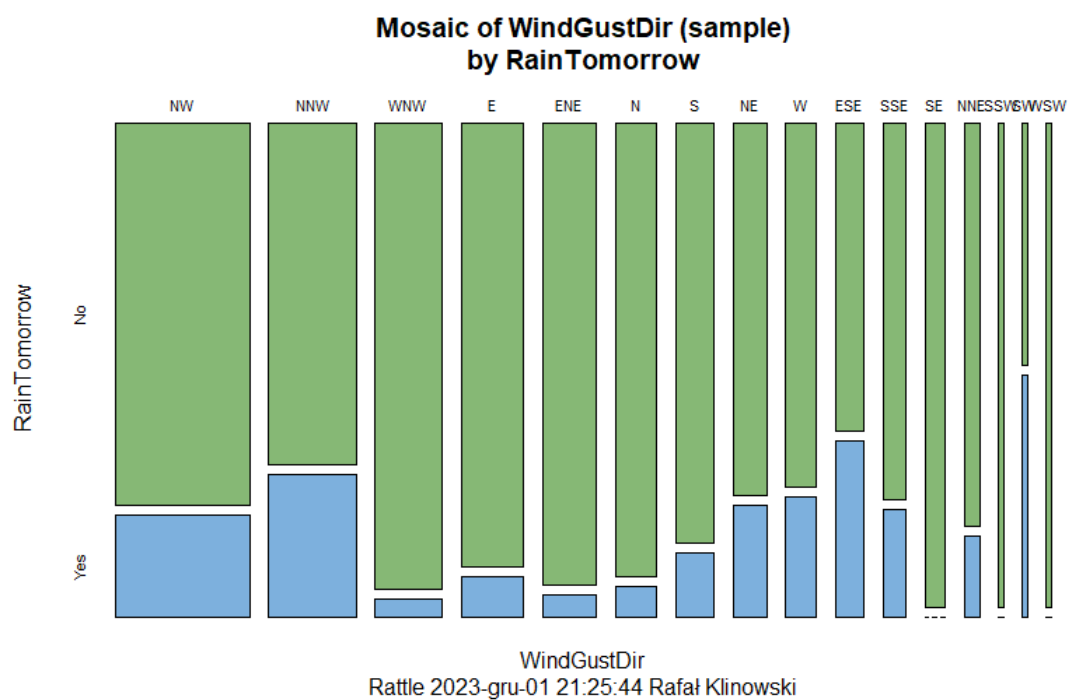
No.	Variable	Bar Plot	Dot Plot	Mosaic	Pairs	Levels
8	WindGustDir	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16
10	WindDir9am	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16
11	WindDir3pm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16
22	RainToday	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2
24	RainTomorrow	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2

One plot has been generated.

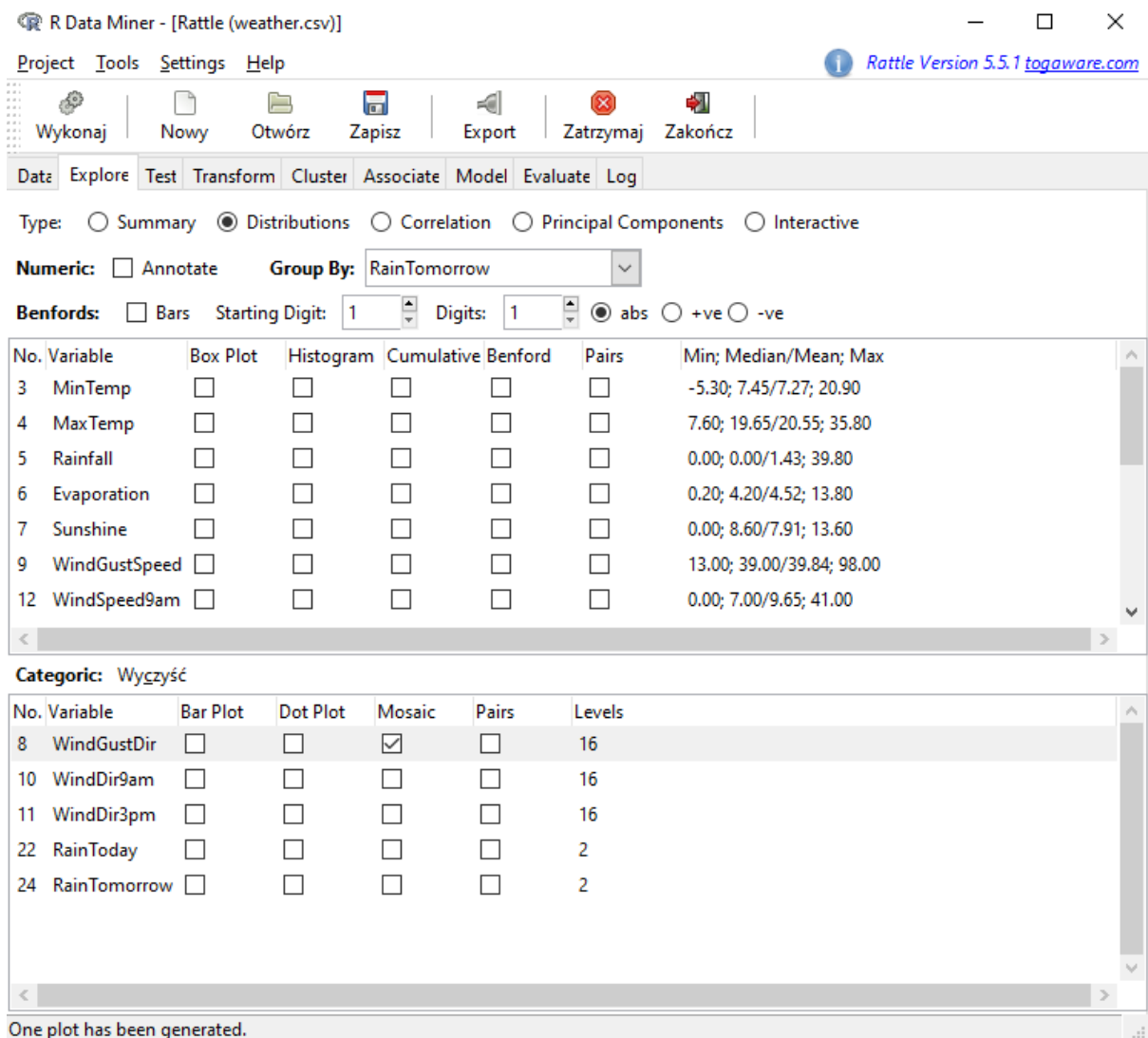
Rysunek 15. Ustawienia w Rattle do wygenerowania wykresu.

Ćw. 9.

Wykres mozaikowy w Rattle.



Rysunek 16. Wykres mozaikowy utworzony w Rattle.



Rysunek 17. Ustawienia w Rattle do wygenerowania wykresu mozaikowego.

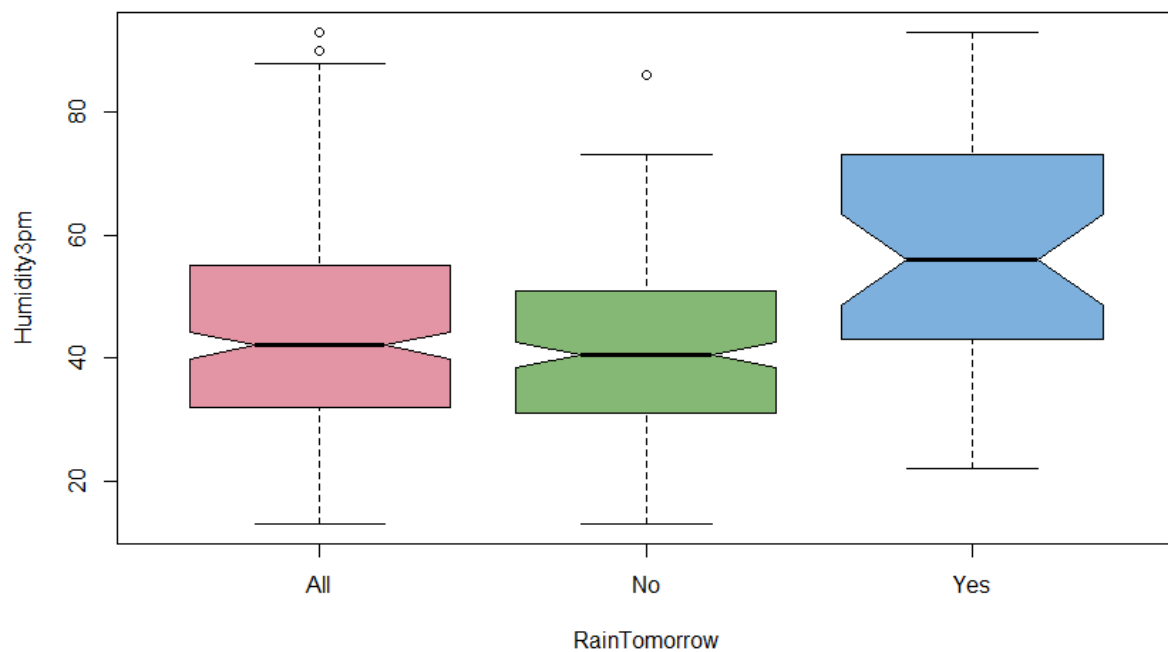
Ćw. 10.

Utworzenie i edycja wykresu w R przy pomocy domyślnego zestawu danych Weather oraz doBy.

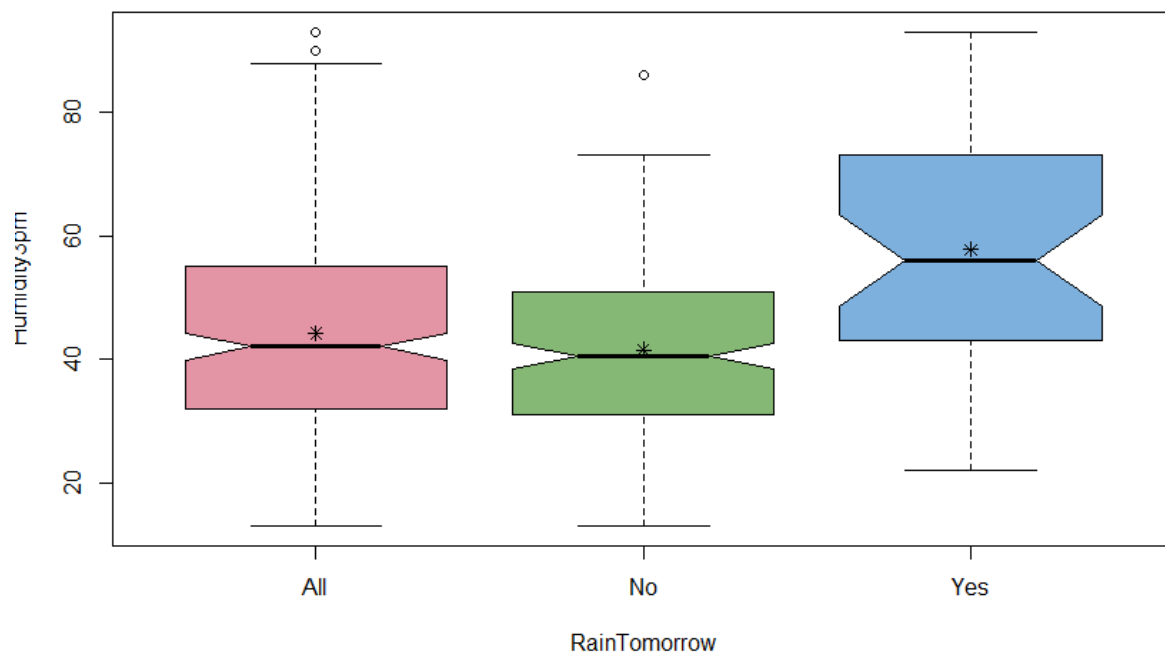
Na początku zgodnie z instrukcjami utworzono wykres dla kolumny Humidity3pm.

```
> ds <- with(crs$dataset[crs$train,], rbind(data.frame(dat=Humidity3pm,grp="All"), data.frame(dat=Humidity3pm[RainTomorrow==
="No"],grp="No"), data.frame(dat=Humidity3pm[RainTomorrow=="Yes"],grp="Yes")))
> bp <- boxplot(formula=dat ~ grp, data=ds, col=rainbow_hcl(3), xlab="RainTomorrow", ylab="Humidity3pm", notch=TRUE)
```

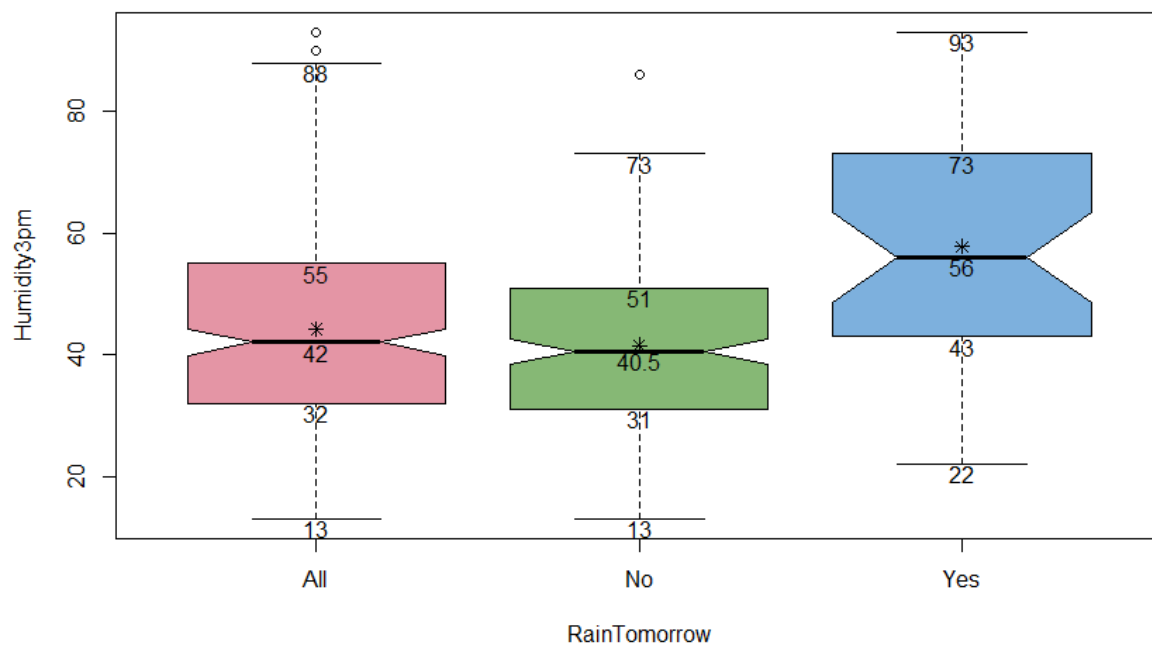
Rysunek 18. Utworzenie danych oraz prostego wykresu.



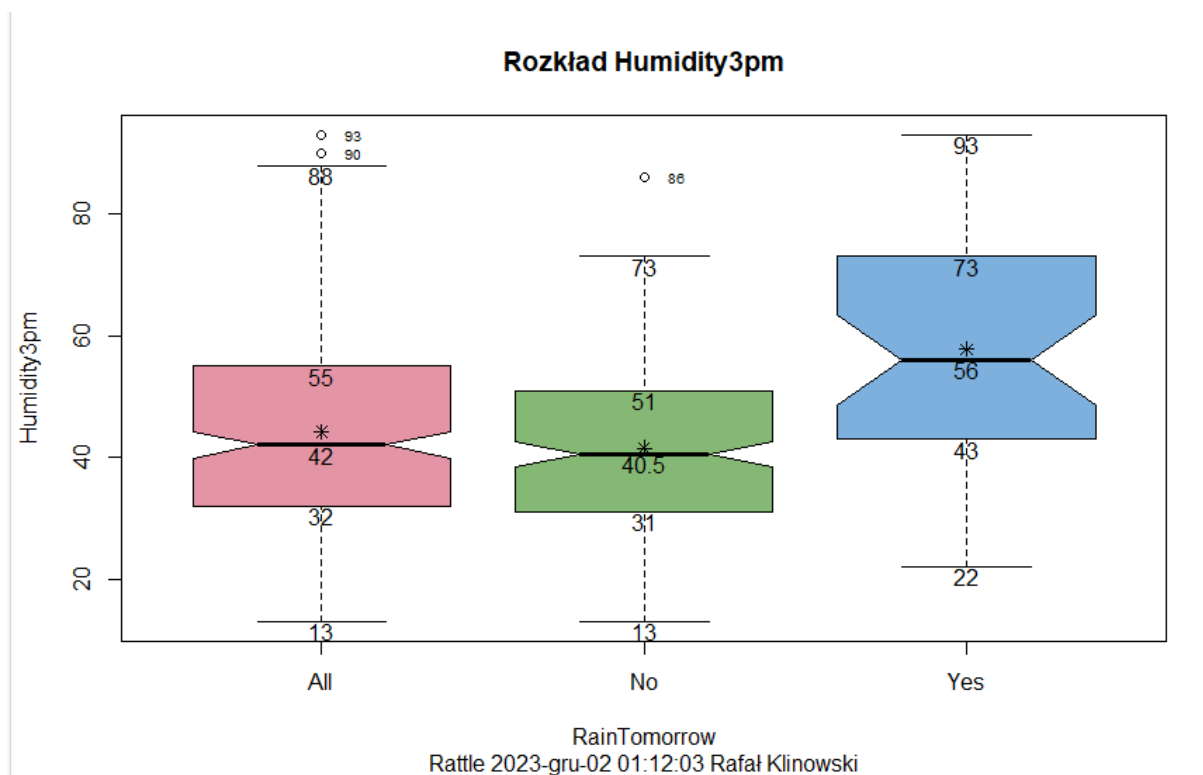
Rysunek 19. Wygląd podstawowego wykresu.



Rysunek 20. Wykres po dodaniu dodatkowych punktów danych.

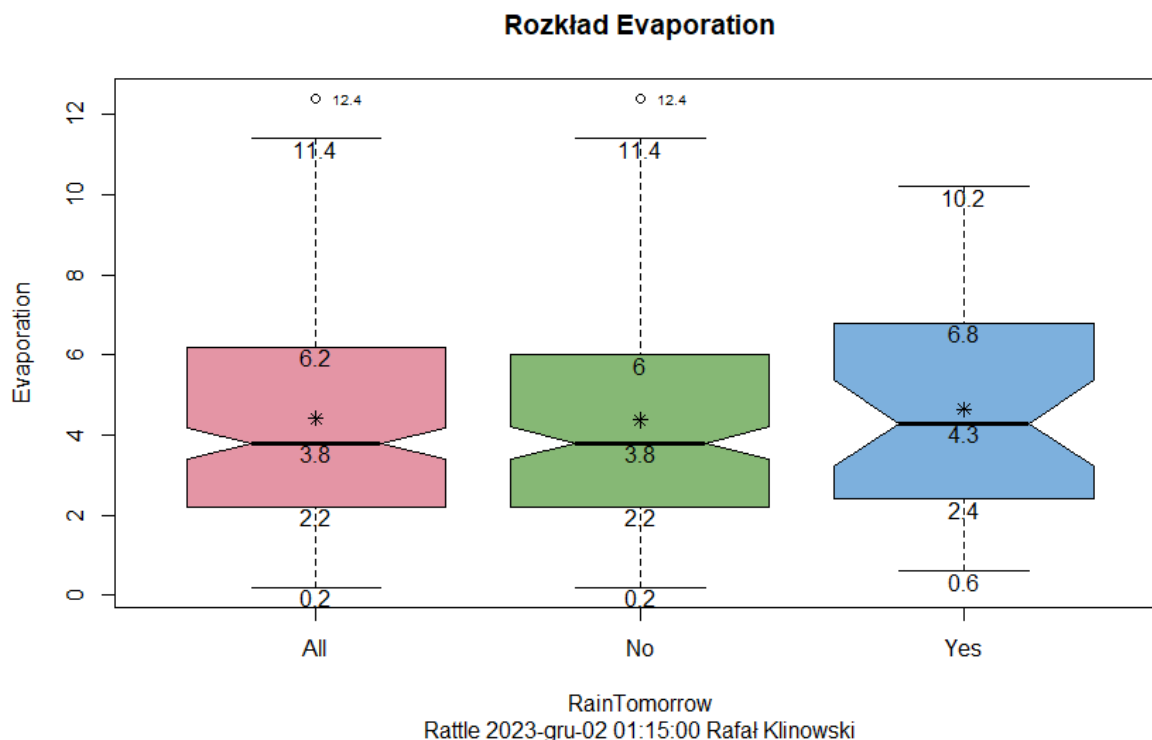


Rysunek 21. Wykres po dodaniu dodatkowego tekstu reprezentującego wartości.



Rysunek 22. Gotowy wykres po dodaniu tytułu oraz wszystkich poprzednich elementów.

Proces powtórzono dla kolumny Evaporation.



Rysunek 23. Powtórzenie powyższych instrukcji analogicznie dla kolumny Evaporation.

```
> ds <- with(crs$dataset[crs$train,], rbind(data.frame(dat=Evaporation,grp="All"), data.frame(dat=Evaporation[RainTomorrow=
="No"],grp="No"), data.frame(dat=Evaporation[RainTomorrow=="Yes"],grp="Yes")))
> bp <- boxplot(formula=dat ~ grp, data=ds, col=rainbow_hcl(3), xlab="RainTomorrow", ylab="Evaporation", notch=TRUE)
> points(x=1:3, y=summaryBy(formula=dat ~ grp, data=ds, FUN=mean, na.rm=TRUE)$dat.mean, pch=8)
+ for (i in seq(ncol(bp$stats))) {
+   text(x=i, y=bp$stats[,i] - 0.02 * (max(ds$dat, na.rm=TRUE) - min(ds$dat, na.rm=TRUE)), labels=bp$stats[,i])
+ }
> text(x=bp$group+0.1, y=bp$out, labels=bp$out, cex=0.6)
> title(main="Rozkład Evaporation", sub=paste("Rattle", format(sys.time(), "%Y-%b-%d %H:%M:%S"), sys.info()["user"]))
```

Rysunek 24. Polecenia wykorzystane do utworzenia powyższego wykresu.

Wnioski.

Największą trudnością podczas realizacji laboratorium było zainstalowanie pakietu Rattle oraz jego zależności. Wymagało to skorzystanie z systemu Windows (poprzez problemy z kompatybilnością na innych systemach, w tym korzystając z kontenera Docker) oraz ręczną instalację wielu z zależności (między innymi RGtk2 czy stringi), co zajęło znacznie więcej czasu niż sama realizacja laboratorium.

Środowisko Rattle umożliwia w prosty sposób pracę nad danymi, w tym tworzenie wykresów, eksplorację danych czy uzyskiwanie podsumowań. Rattle zawiera również sporą ilość narzędzi wycelowanych w uczenie maszynowe czy grupowanie danych. Środowisko współpracuje z wieloma dodatkowymi pakietami i zawiera przejrzysty interfejs, w którym dość łatwo znaleźć wszystkie interesujące nas opcje.