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

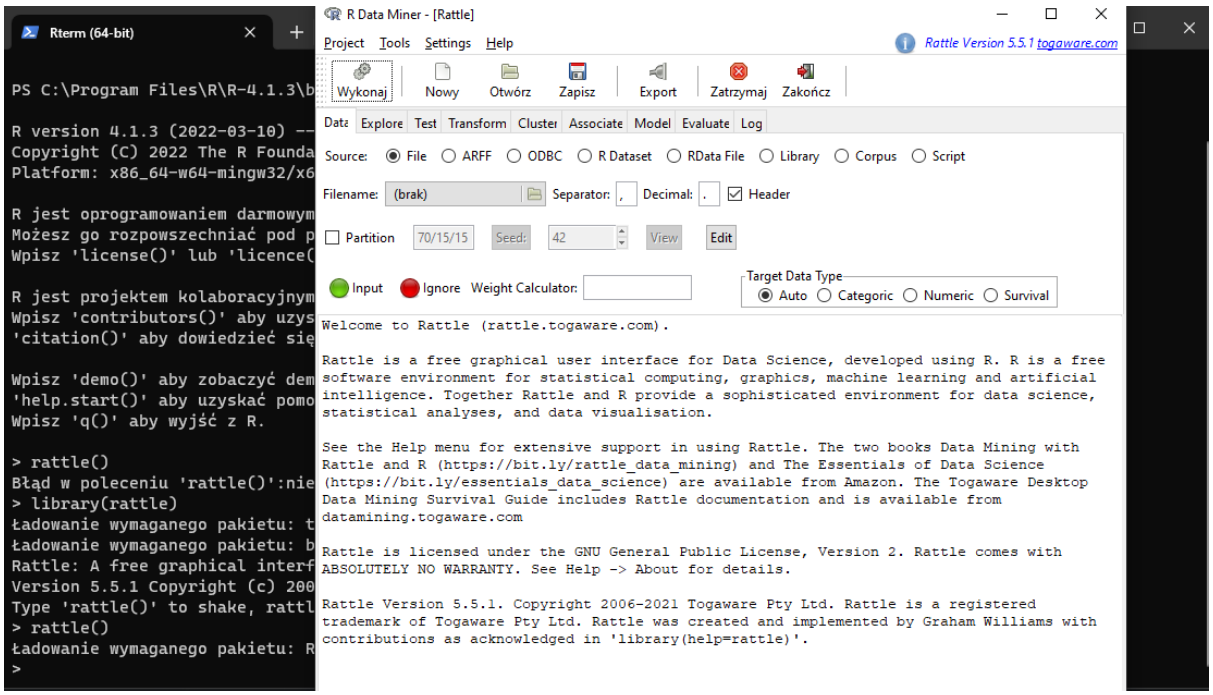
Poszczególne ćwiczenia będą wykonywane w pliku źródłowym edytowanym przy pomocy środowiska RStudio oraz Rattle, opisanego w poprzednich laboratoriach.

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.

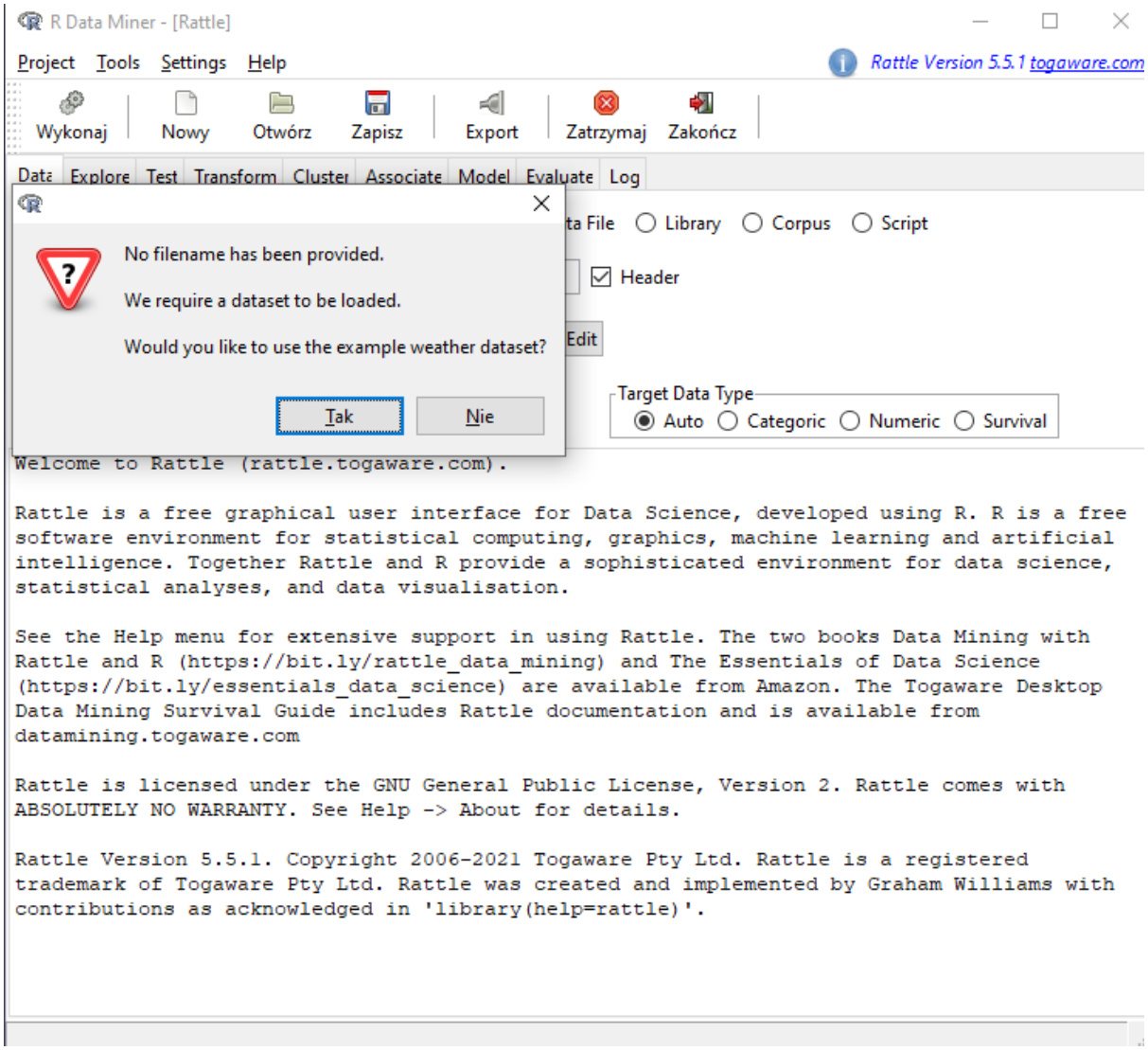
Po zainstalowaniu pakietów uruchomiono środowisko Rattle.



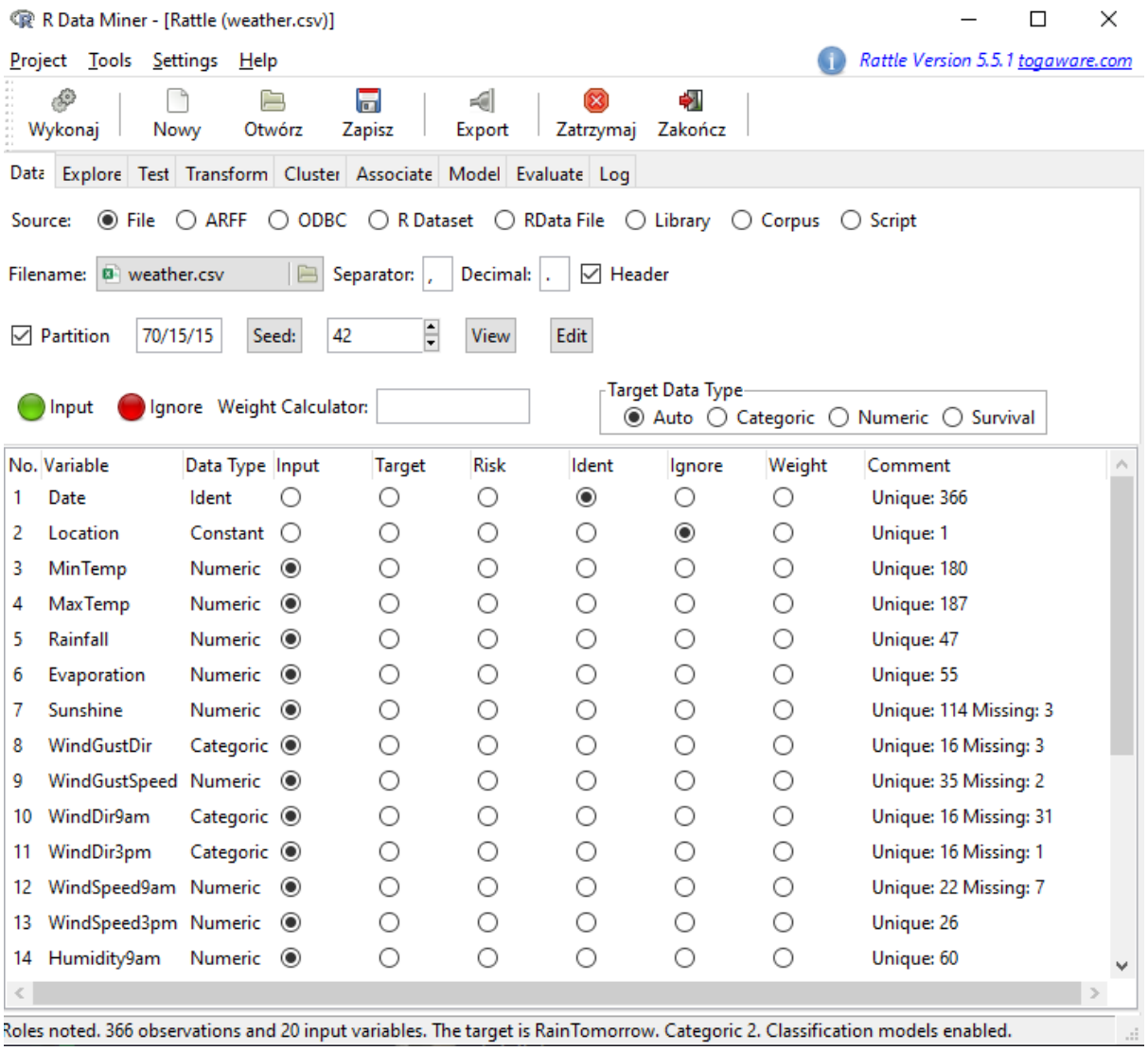
Rysunek 1. Środowisko Rattle.

Ćw. 3.

Wczytanie przykładowego zbioru danych w Rattle.



Rysunek 2. Ekran po kliknięciu Execute.



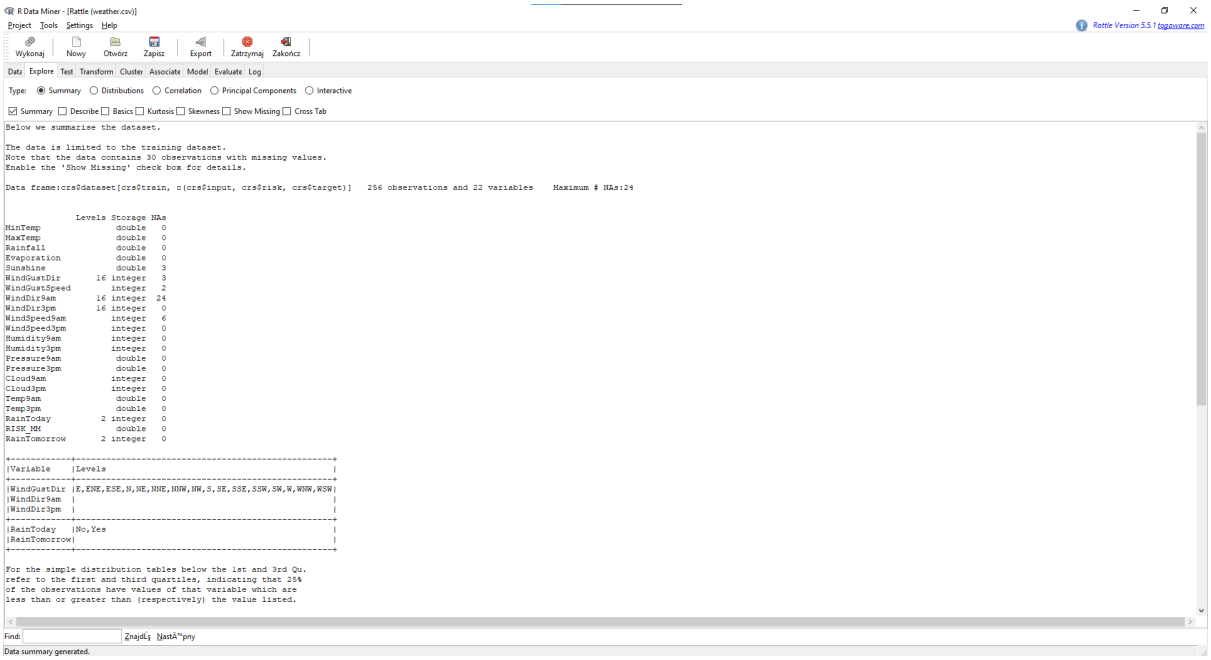
Rysunek 3. Wygląd wczytanego zbioru danych.

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.

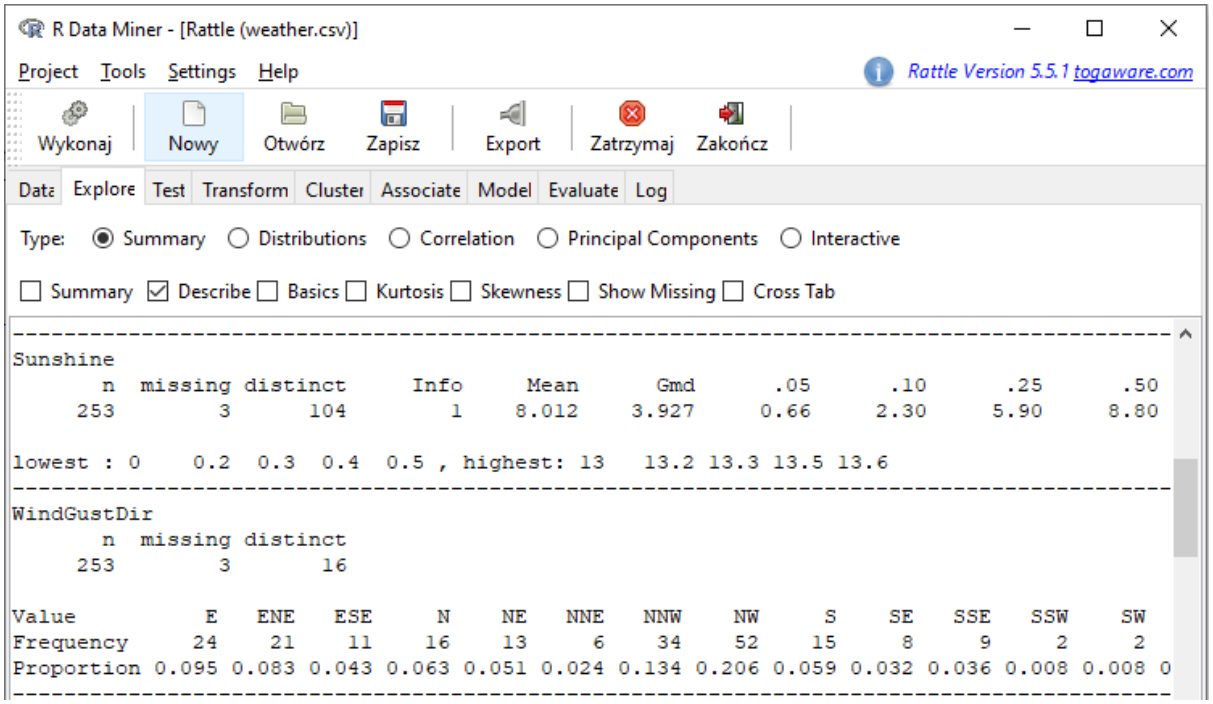
Podsumowanie danych za pomocą Hmisc. Konieczne było najpierw zainstalowanie pakietu.

```
> 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.

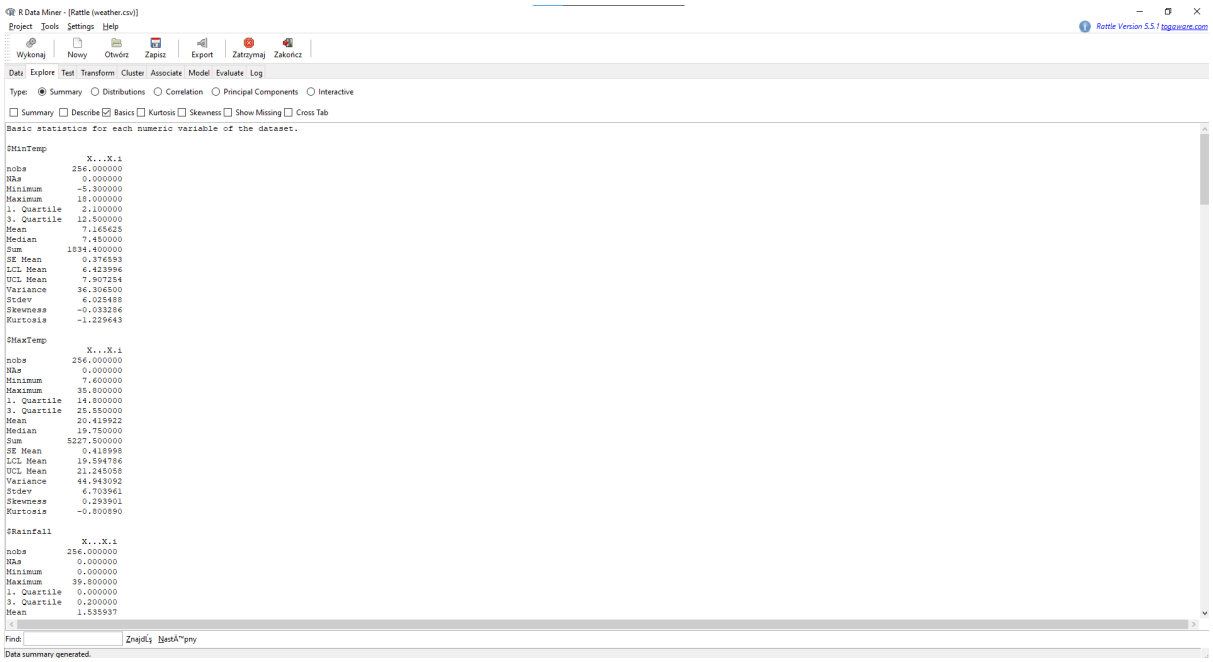
Podsumowanie numeryczne z wykorzystaniem pakietu fBasics.

```
> 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
```

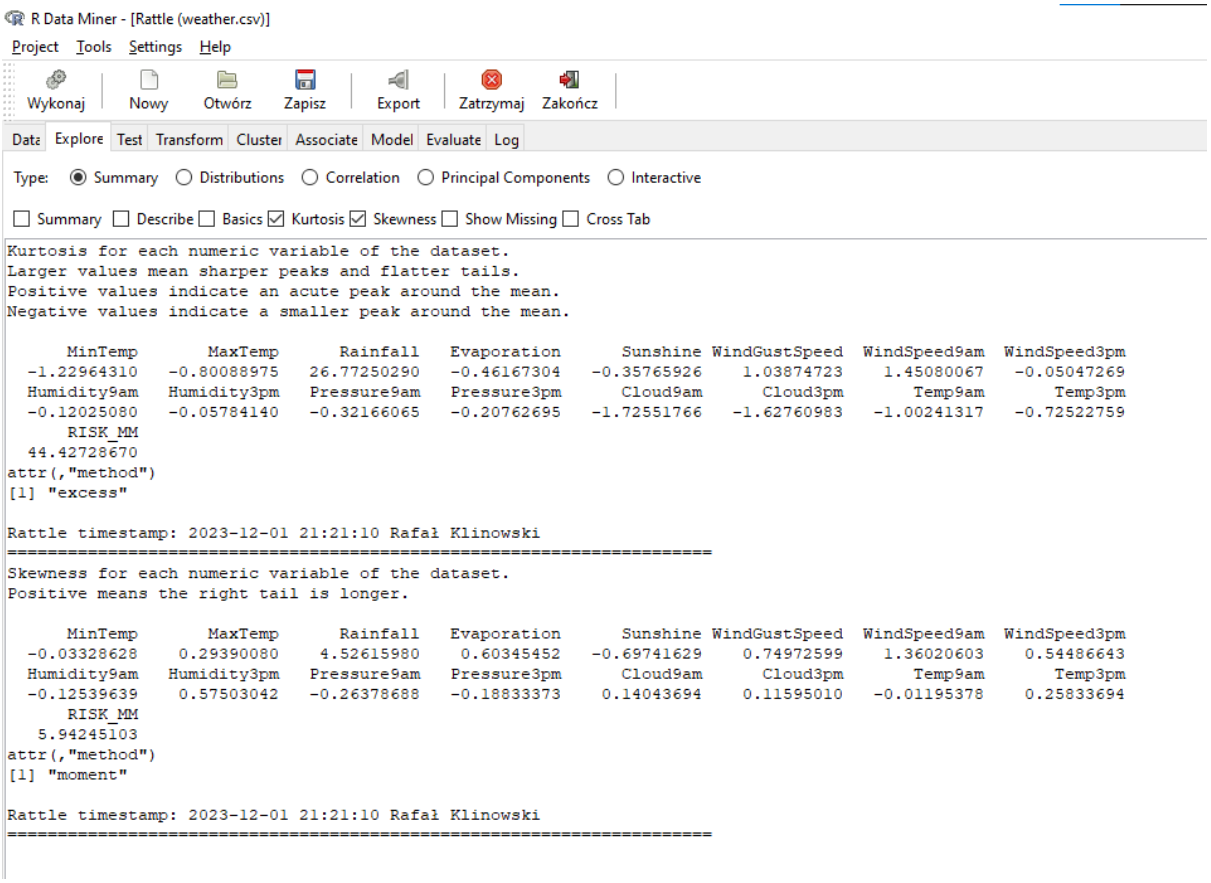
Rysunek 8. Podstawowe statystyki w R.

```
> 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.



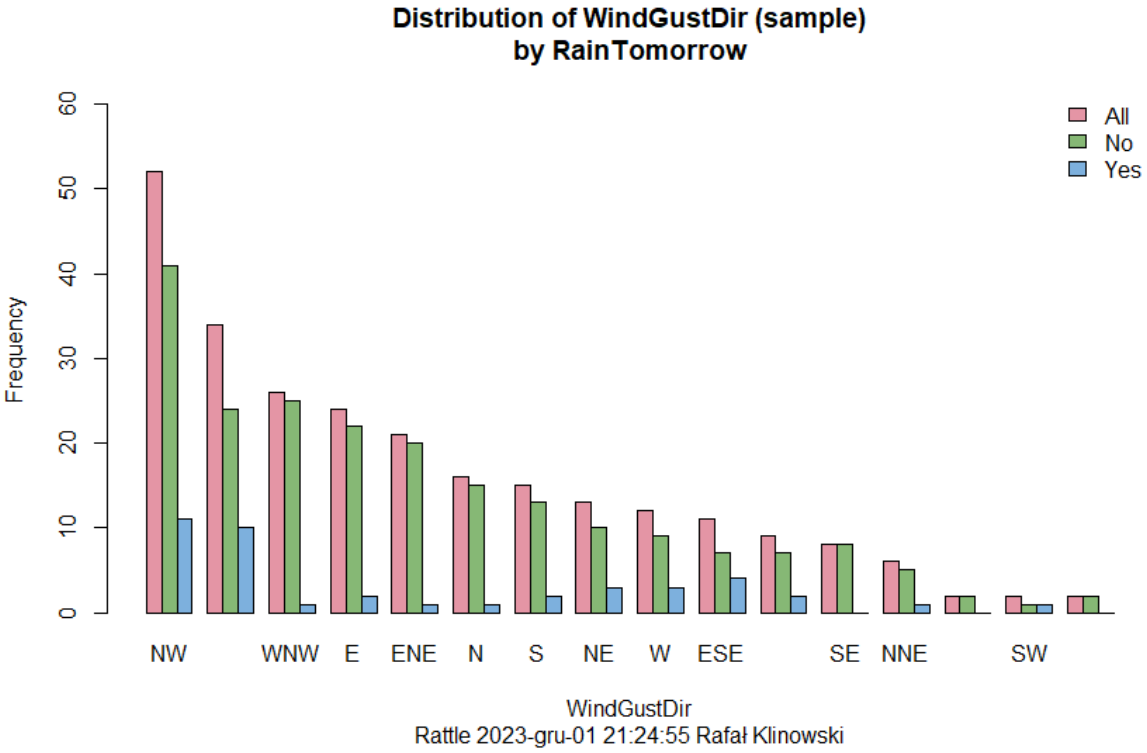
Rysunek 10. Podsumowanie numeryczne w Rattle.



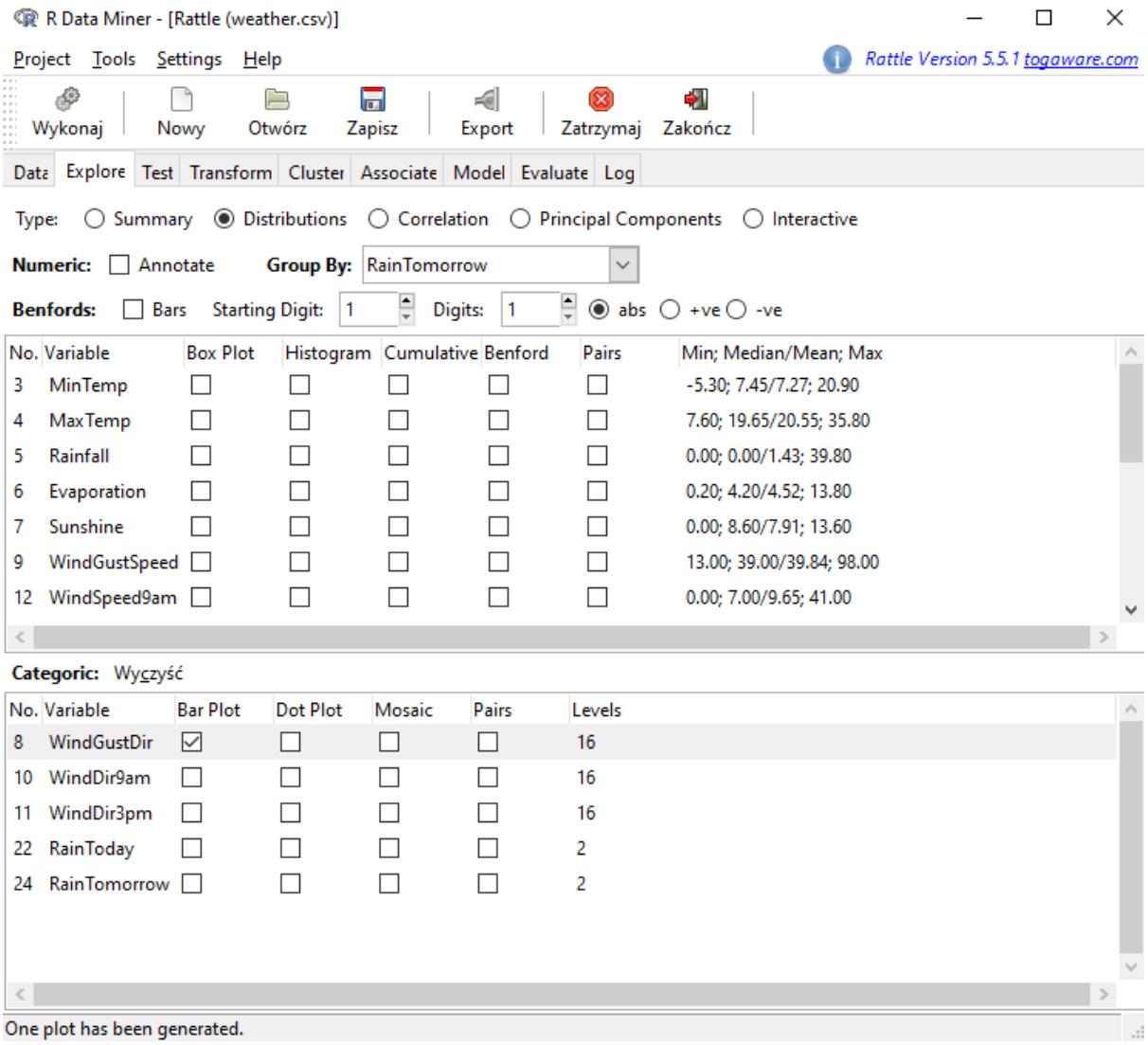
Rysunek 11. Skośność i kurtoza w Rattle.

Ćw. 7.

Wykres słupkowy w Rattle.



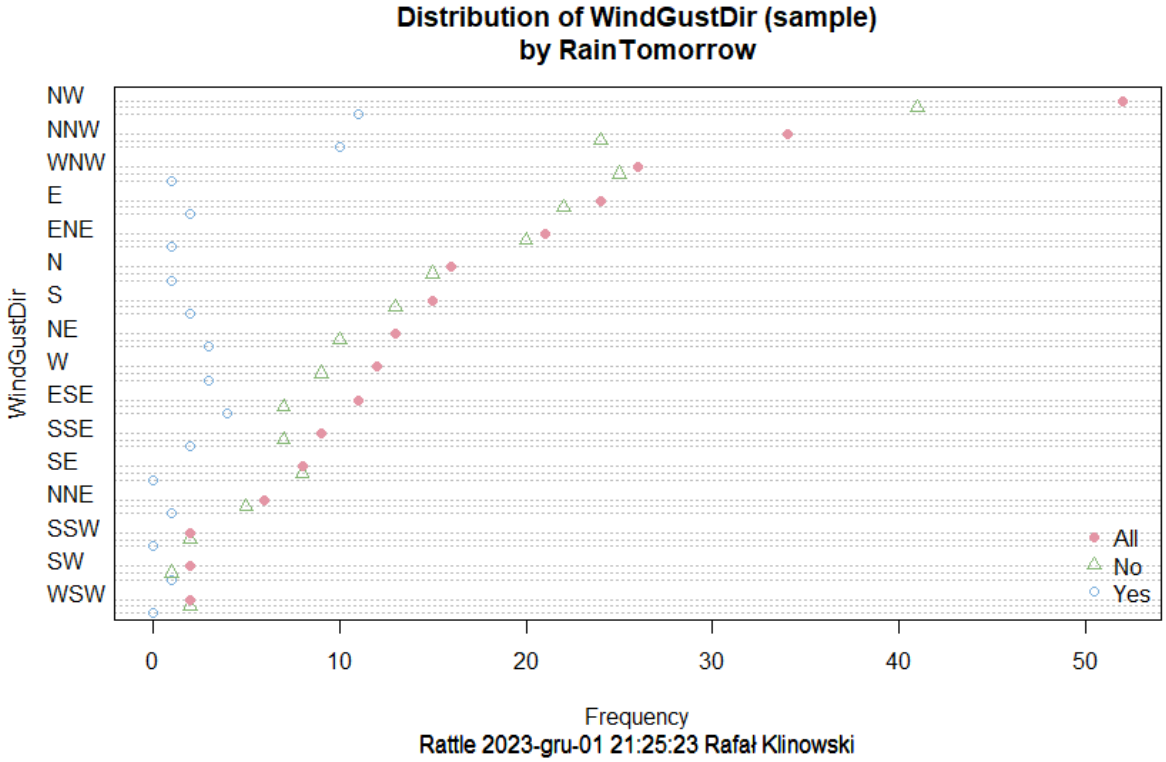
Rysunek 12. Wykres słupkowy w Rattle.



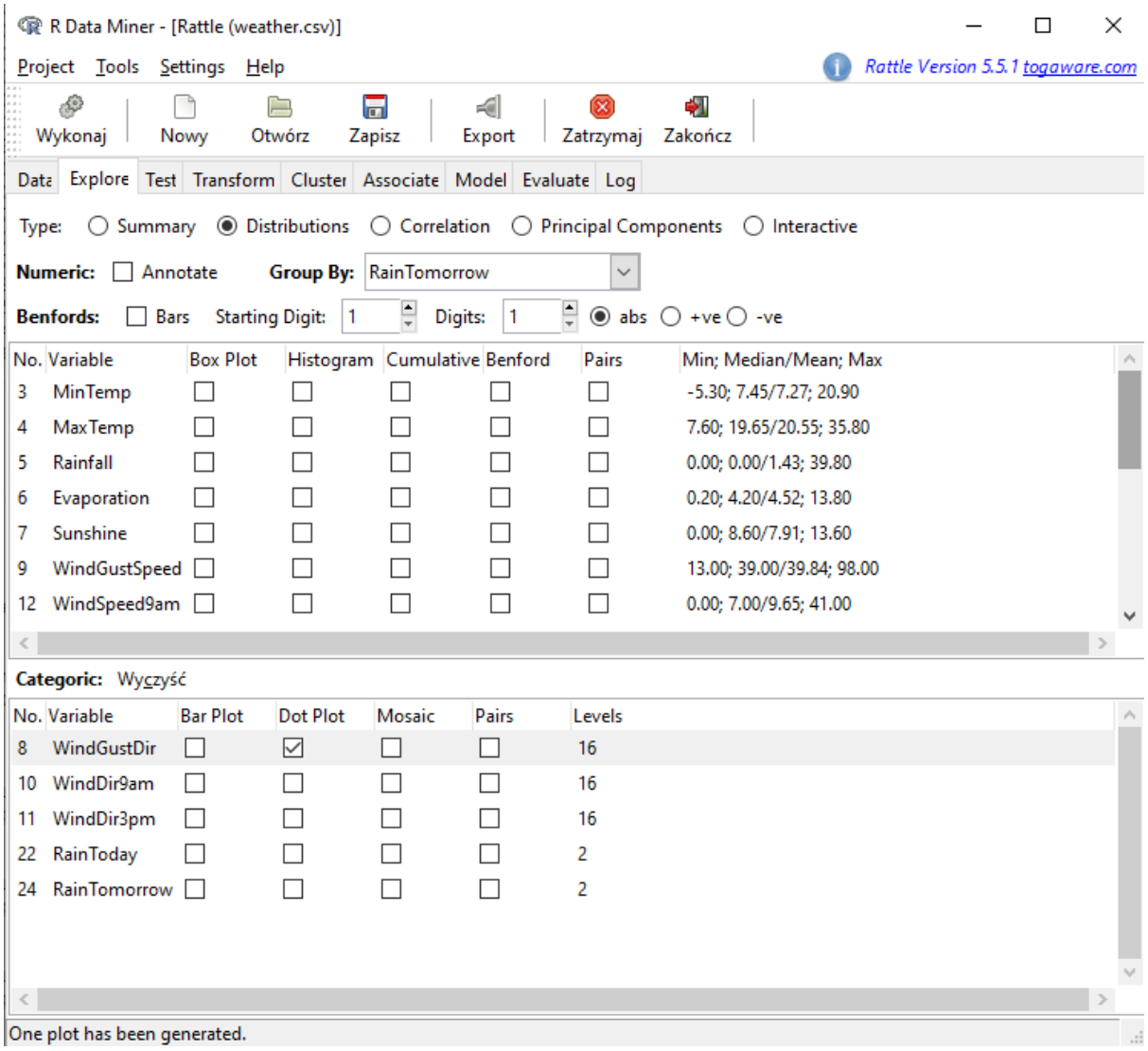
Rysunek 13. Parametry do utworzenia wykresu w Rattle.

Ćw. 8.

Wykres kropkowy w Rattle.



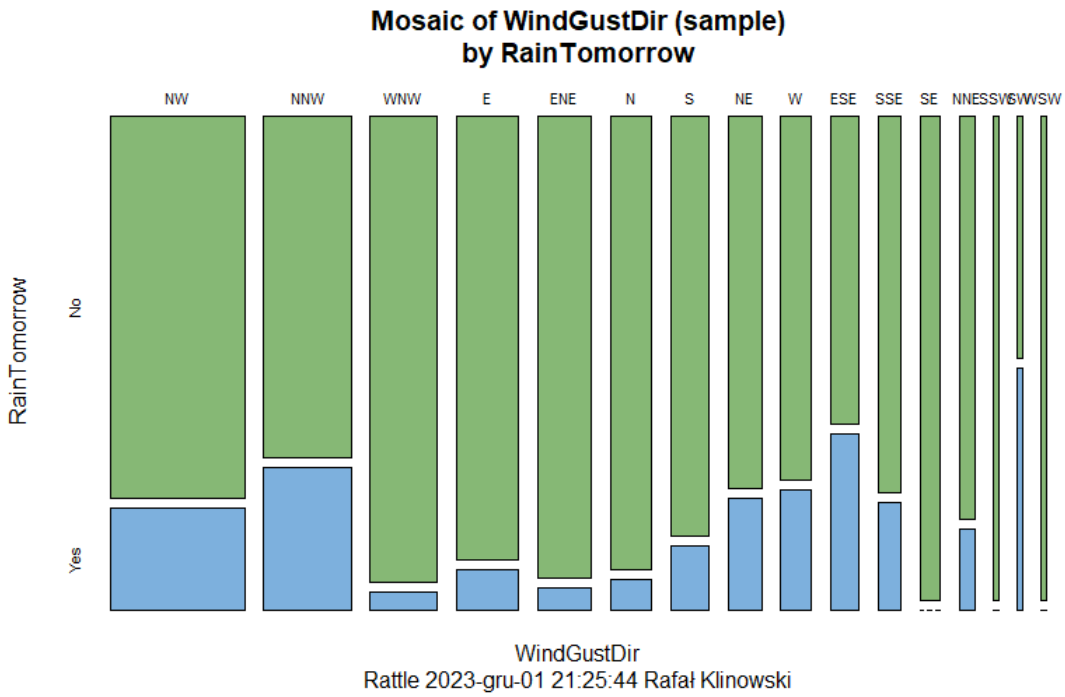
Rysunek 14. Wykres kropkowy utworzony w Rattle.



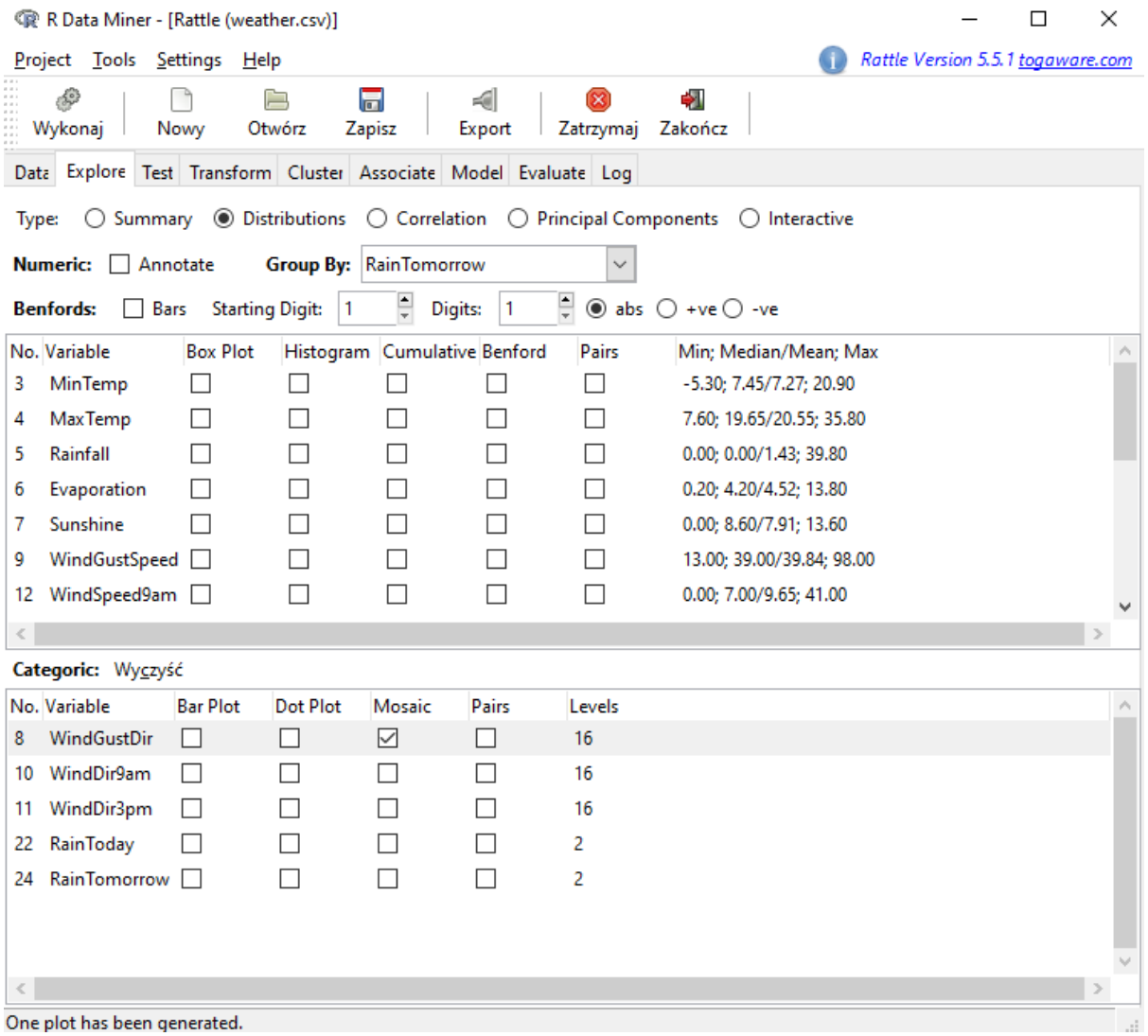
Rysunek 15. Ustawienia w Rattle do wygenerowania wykresu.

Ćw. 9.

Wykres mozaikowy w Rattle.



Rysunek 16. Wykres mozaikowy utworzony w Rattle.



Rysunek 17. Ustawienia w Rattle do wygenerowania wykresu mozaikowego.

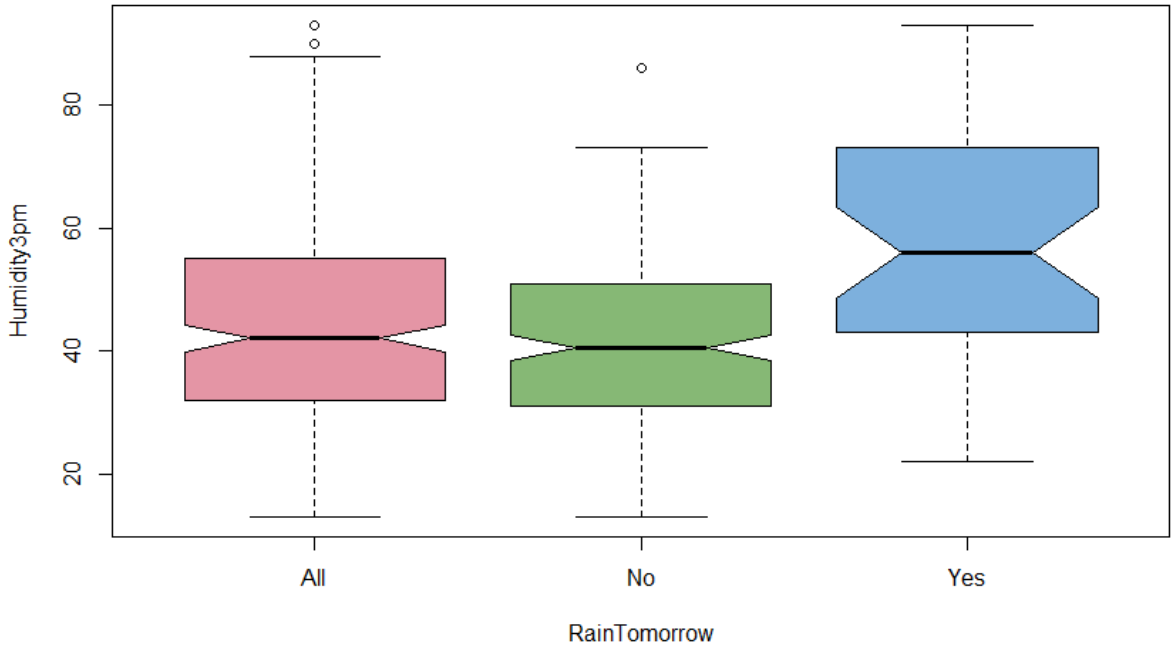
Ćw. 10.

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

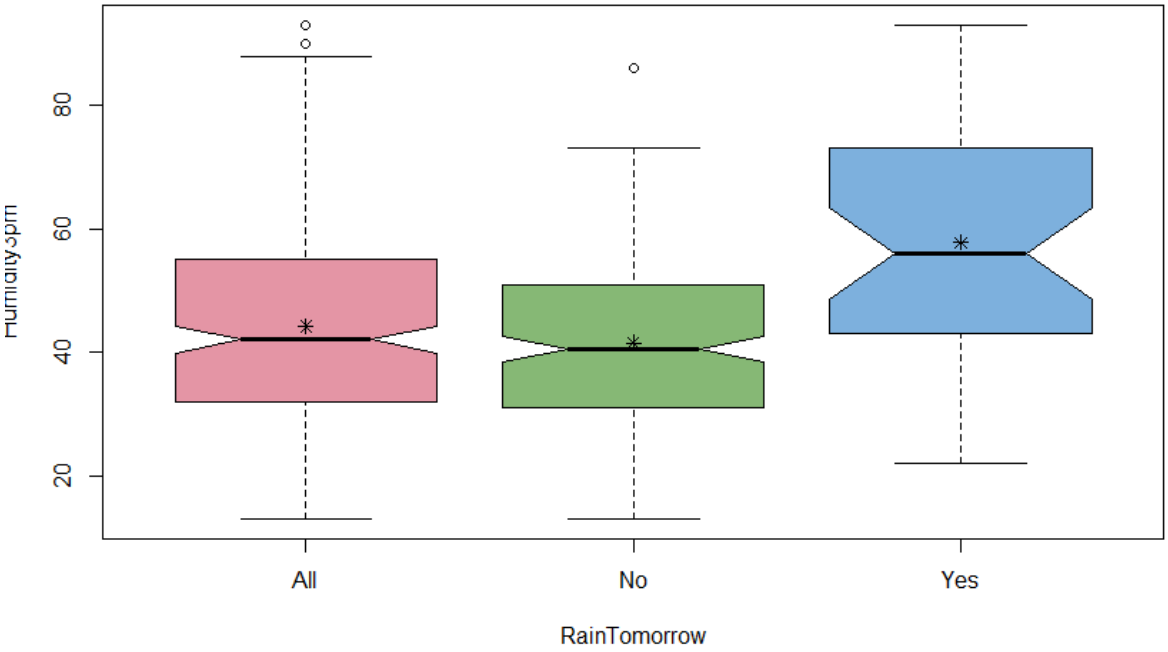
Na początku zgodnie z instrukcjami utworzono wykres dla kolumny Humidity3pm.

```
> 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)
```

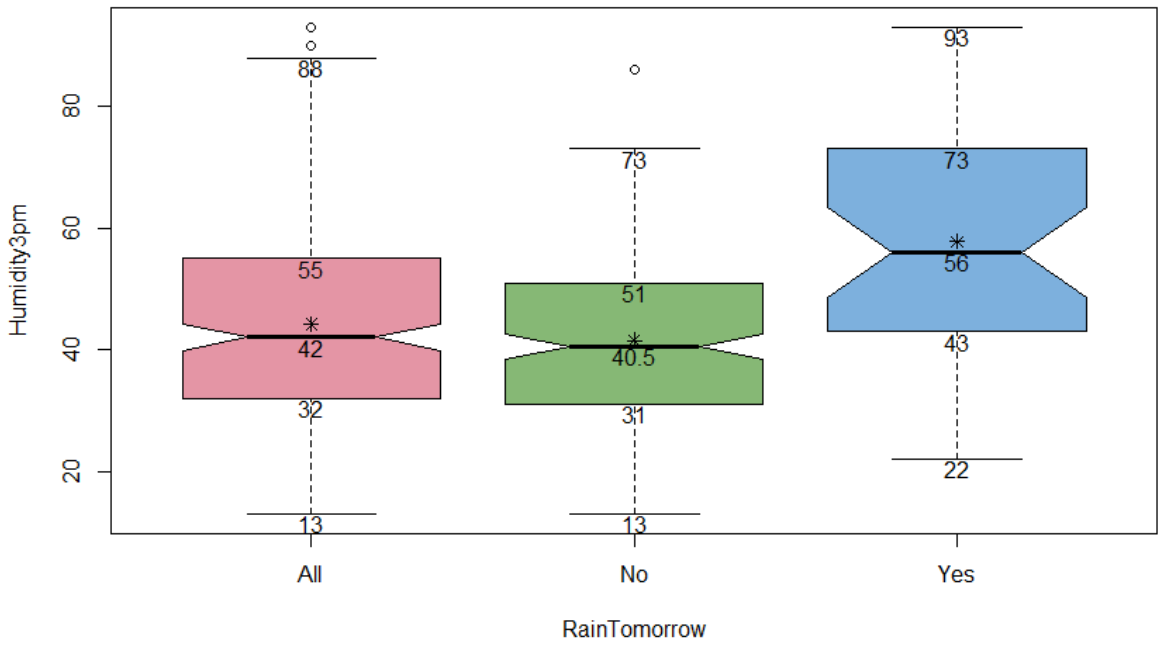
Rysunek 18. Utworzenie danych oraz prostego wykresu.



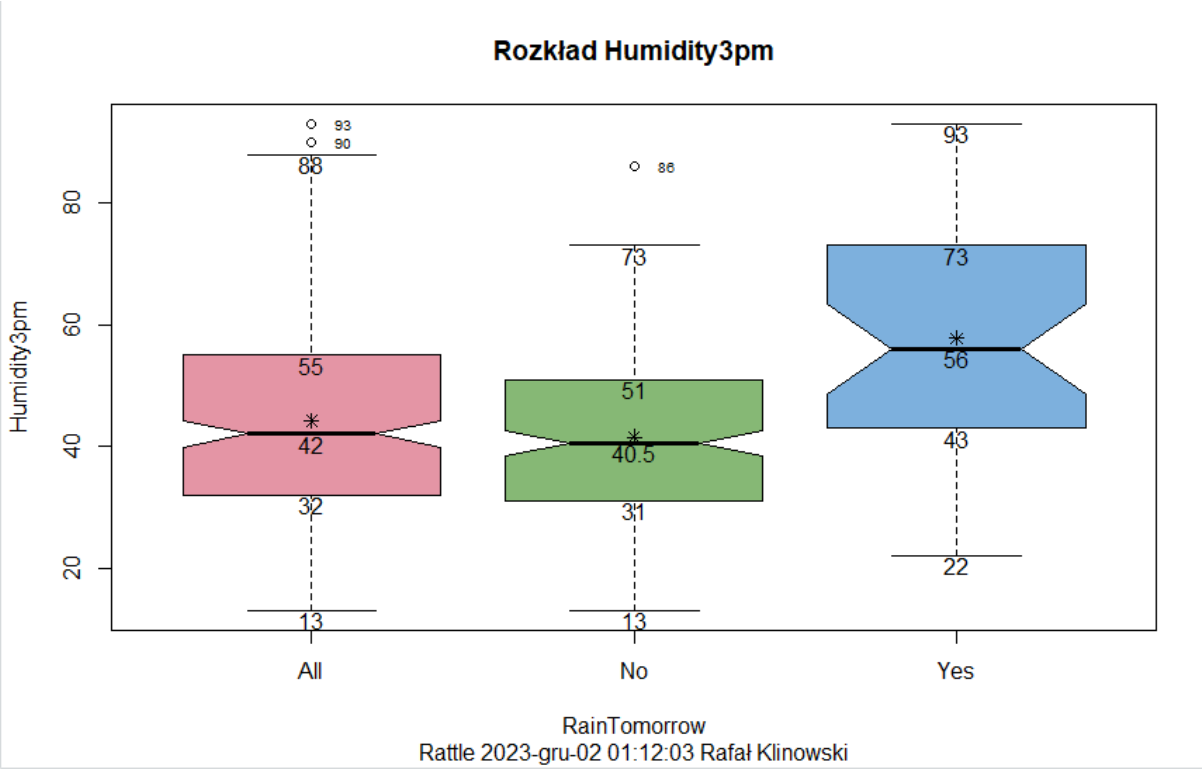
Rysunek 19. Wygląd podstawowego wykresu.



Rysunek 20. Wykres po dodaniu dodatkowych punktów danych.

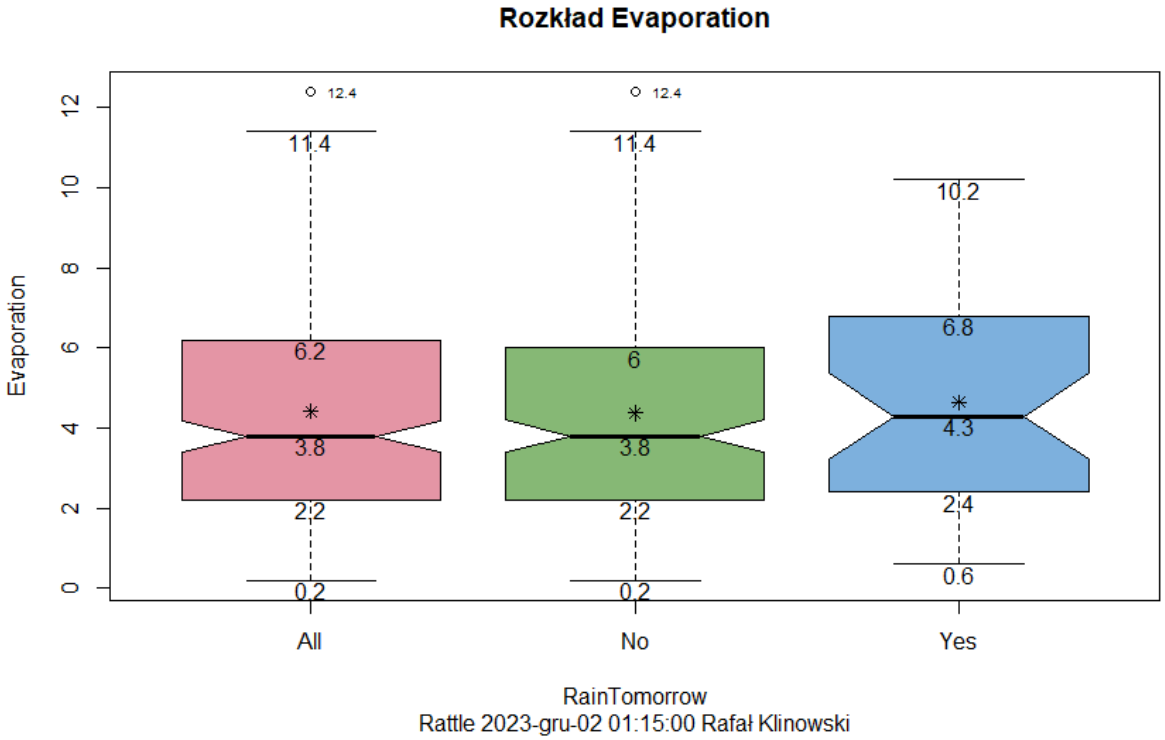


Rysunek 21. Wykres po dodaniu dodatkowego tekstu reprezentującego wartości.



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.