SPRAWOZDANIE

Zajęcia: Zbiory Big Data i Eksploracja Danych

Prowadząca: dr inż. Ruslana Ziubina

Laboratorium nr 5 Data rozpoczęcia: 15.12.2023 Temat: Reguły asocjacyjne	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 było „ręczne” policzenie wsparcia, pewności i liftu dla przykładowego zestawu danych związanych z zakupami kawy i herbaty.

```
> data <- data.frame(
Warning message:
In normalizePath(path.expand(path), winslash, mustWork) :
  path[1]="C:/Users/Rafa? Klinowski/Documents": Nazwa pliku, nazwa katalogu lub składnia etykiety woluminu jest niepoprawna
+     produkt = c("herbata", "nie_herbata", "suma"),
+     kawa = c(20, 70, 90),
+     nie_kawa = c(5, 5, 10),
+     suma = c(25, 75, 100)
+ )
> data
      produkt kawa nie_kawa suma
1     herbata   20        5   25
2 nie_herbata   70        5   75
3       suma   90       10  100
```

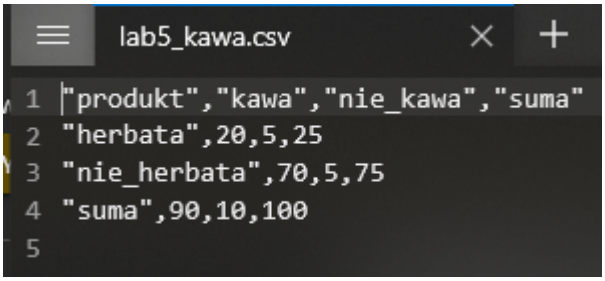
Rysunek 1. Utworzenie zestawu danych.

```
> kawa_count <- data[1, "kawa"]
> nie_kawa_count <- data[2, "nie_kawa"]
> suma_count <- data[3, "suma"]
> support_kawa <- kawa_count / suma_count
>
> support_nie_kawa <- nie_kawa_count / suma_count
>
> confidence_kawa_nie_kawa <- kawa_count / suma_count
```

Rysunek 2. Obliczenie wsparcia i pewności dla kawy.

values	
confidence_kawa_nie_kawa	0.2
kawa_count	20
nie_kawa_count	5
suma_count	100
support_kawa	0.2
support_nie_kawa	0.05

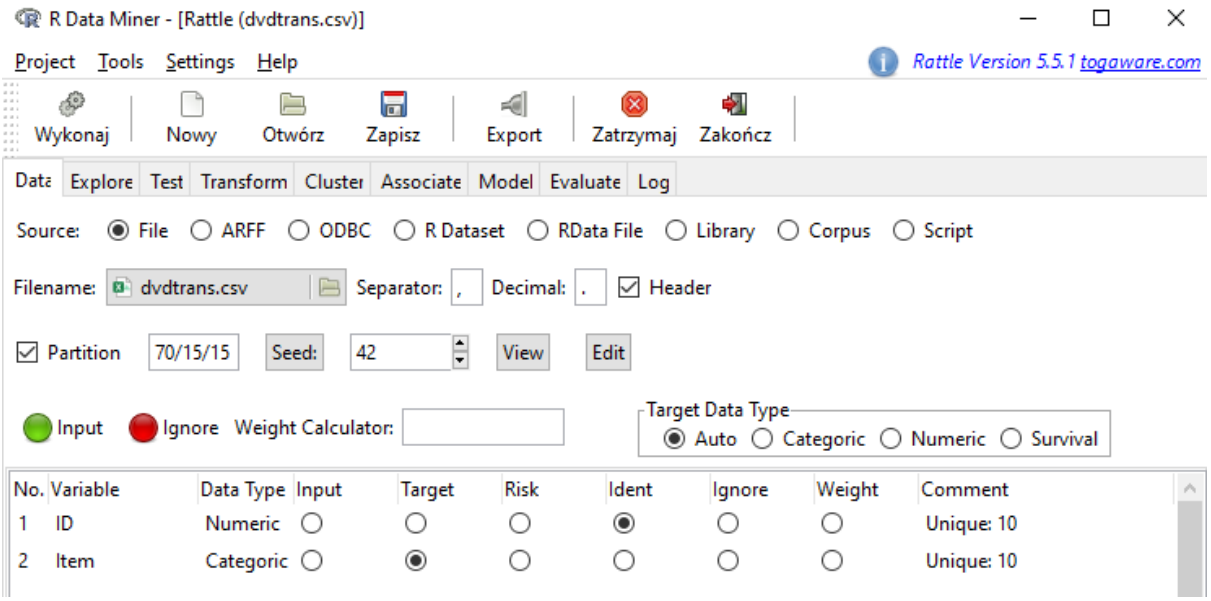
Rysunek 3. Uzyskane wyniki.



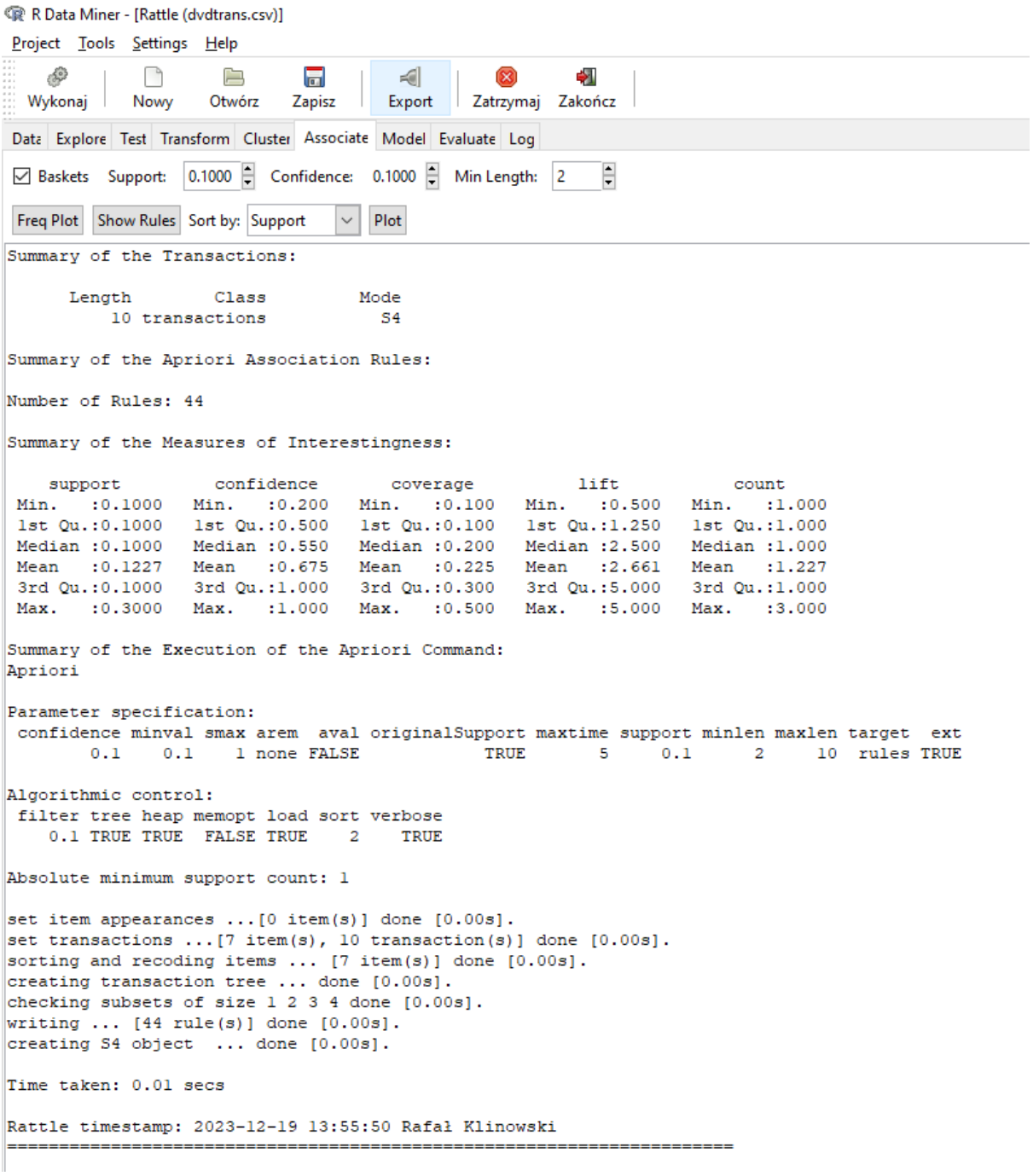
Rysunek 4. Zbiór danych zapisany do pliku CSV.

Ćw. 2.

W tym ćwiczeniu przeprowadzono analizę reguł asocjacyjnych w Rattle na podstawie zbioru danych DVDtrans.



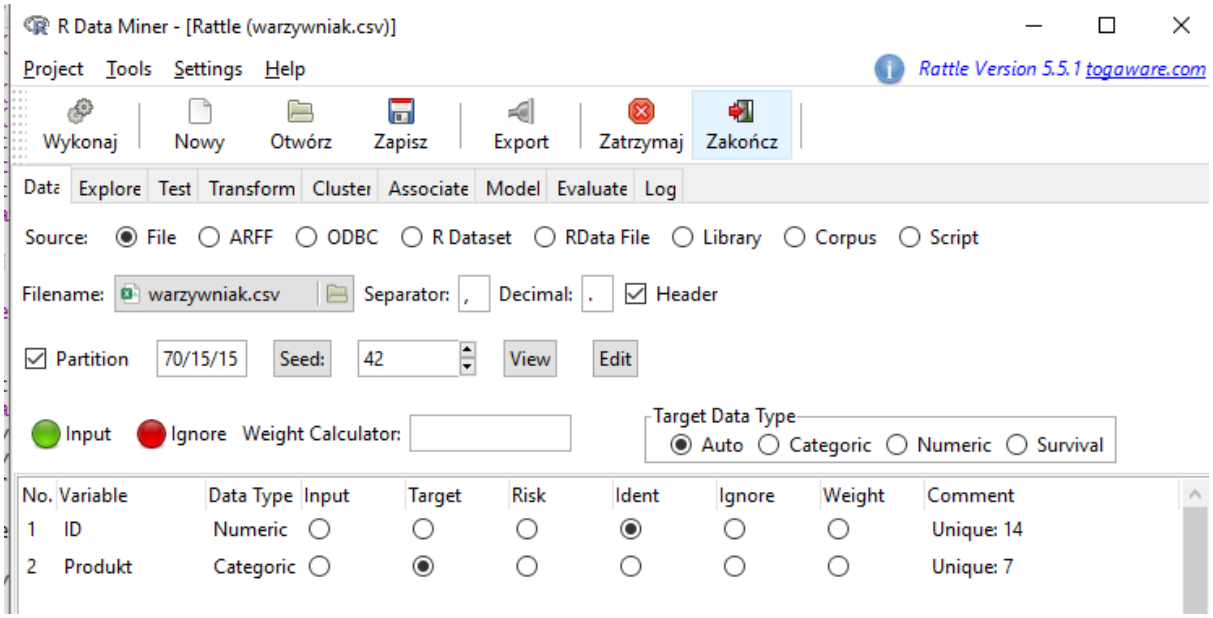
Rysunek 5. Załadowanie zbioru danych DVDtrans.



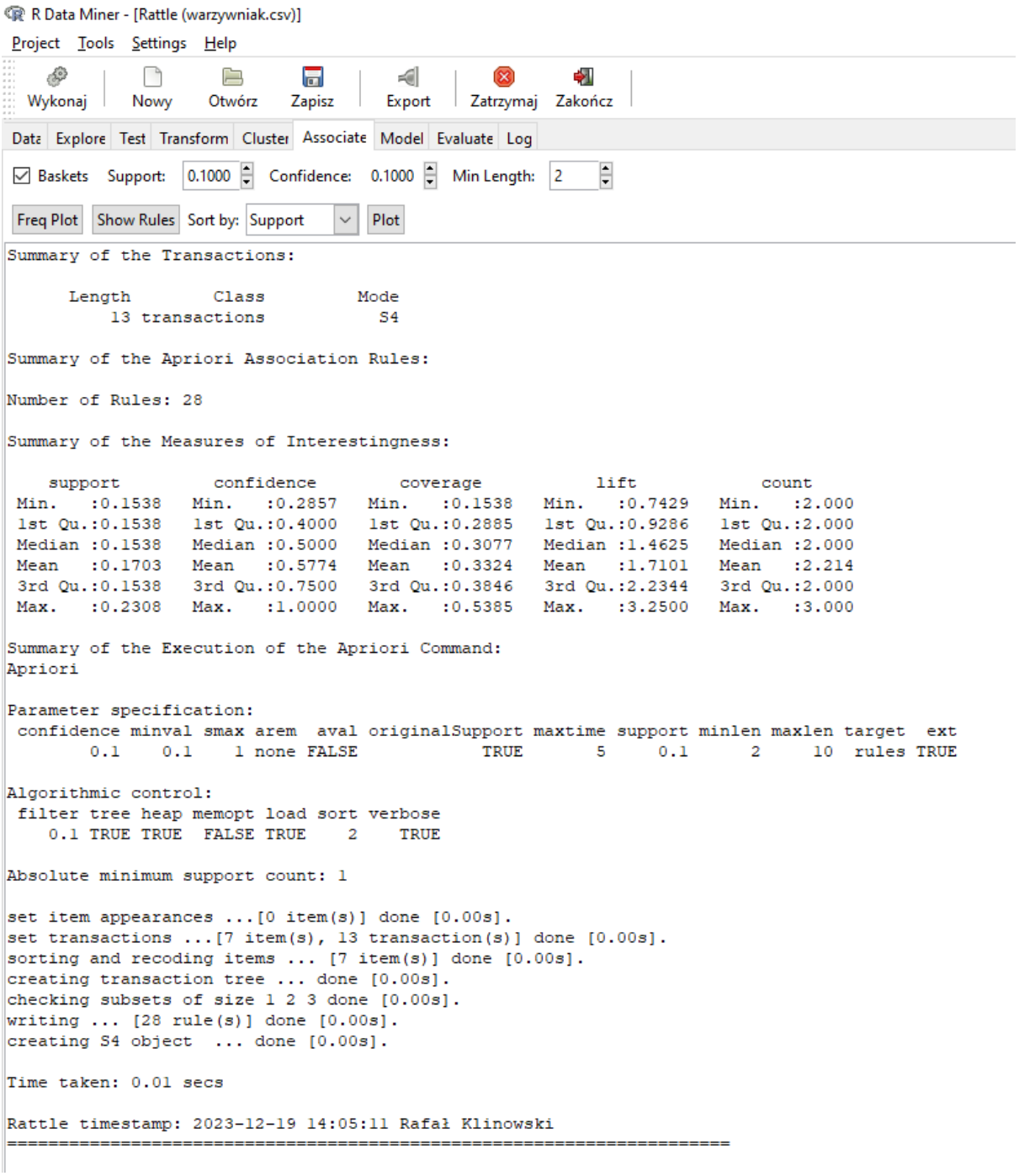
Rysunek 6. Reguły asocjacyjne dla tego zbioru danych w Rattle.

Ćw. 3.

Przeanalizowano reguły asocjacyjne dla sklepu warzywnego przy pomocy Rattle.



Rysunek 7. Wczytanie danych z pliku warzywniak.csv.



Rysunek 8. Reguły asocjacyjne wygenerowane dla tego zbioru danych.

Ćw. 4.

Reguły asocjacyjne w R – wykorzystując zbiór DVDtrans wygenerowano reguły bezpośrednio w R przy pomocy algortymu „apriori”.

```
> library(arules)
> library(rattle)
> dvdtrans <- read.csv(system.file("csv","dvdtrans.csv",package="rattle"))
> dvdDS <- new.env()
> dvdDS$data <- as(split(dvdtrans$Item, dvdtrans$ID), "transactions")
> dvdDS$data
transactions in sparse format with
 10 transactions (rows) and
 10 items (columns)
```

Rysunek 9. Wczytanie danych DVDtrans.

```
> evalq({model <- apriori(data, parameter=list(support=0.2, confidence=0.1))}, dvdAPRIORI)
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.1    0.1    1 none FALSE            TRUE       5     0.2      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10 item(s), 10 transaction(s)] done [0.00s].
sorting and recoding items ... [7 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [20 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Rysunek 10. Wygenerowanie reguł asocjacyjnych przy pomocy apriori.

```
> inspect(sort(dvdAPRIORI$model, by="confidence")[1:5])
    lhs                          rhs           support confidence coverage lift     count
[1] {LOTR1}                   => {LOTR2}       0.2     1          0.2      5.000000 2
[2] {LOTR2}                   => {LOTR1}       0.2     1          0.2      5.000000 2
[3] {Green Mile}              => {Sixth Sense} 0.2     1          0.2      1.666667 2
[4] {Patriot}                 => {Gladiator}   0.6     1          0.6      1.428571 6
[5] {Patriot, Sixth Sense} => {Gladiator}   0.4     1          0.4      1.428571 4
```

Rysunek 11. Uzyskane reguły.

Ćw. 5.

Reguły asocjacyjne w R dla zbioru danych „groceries”. Zapisanie reguł do pliku.

```
> g <- read.transactions("E:\\Zdalna Edukacja\\Magisterskie\\Semestr 2\\ZBDIED\\groceries.csv", sep=",")
Warning message:
In readLines(file, encoding = encoding) :
  incomplete final line found on 'E:\Zdalna Edukacja\Magisterskie\Semestr 2\ZBDIED\groceries.csv'
> reguly.zakupy <- apriori(g, parameter=list(supp=0.001, conf=0.8))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       5   0.001      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> inspect(reguly.zakupy[1:7])
    lhs                                                rhs            support     confidence coverage    lift      count
[1] {liquor, red/blush wine}                        => {bottled beer} 0.001931876 0.9047619  0.002135231 11.235269 19
[2] {cereals, curd}                                 => {whole milk}   0.001016777 0.9090909  0.001118454  3.557863 10
[3] {cereals, yogurt}                               => {whole milk}   0.001728521 0.8095238  0.002135231  3.168192 17
[4] {butter, jam}                                   => {whole milk}   0.001016777 0.8333333  0.001220132  3.261374 10
[5] {bottled beer, soups}                           => {whole milk}   0.001118454 0.9166667  0.001220132  3.587512 11
[6] {house keeping products, napkins}               => {whole milk}   0.001321810 0.8125000  0.001626843  3.179840 13
[7] {house keeping products, whipped/sour cream} => {whole milk}   0.001220132 0.9230769  0.001321810  3.612599 12
> summary(reguly.zakupy)
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 29 229 140  12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   4.329   5.000   6.000

summary of quality measures:
    support           confidence        coverage             lift            count
 Min.   :0.001017   Min.   :0.8000   Min.   :0.001017   Min.   : 3.131   Min.   :10.00
 1st Qu.:0.001017   1st Qu.:0.8333   1st Qu.:0.001220   1st Qu.: 3.312   1st Qu.:10.00
 Median :0.001220   Median :0.8462   Median :0.001322   Median : 3.588   Median :12.00
 Mean   :0.001247   Mean   :0.8663   Mean   :0.001449   Mean   : 3.951   Mean   :12.27
 3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.:0.001627   3rd Qu.: 4.341   3rd Qu.:13.00
 Max.   :0.003152   Max.   :1.0000   Max.   :0.003559   Max.   :11.235   Max.   :31.00

mining info:
 data ntransactions support confidence                                                call
    g          9835   0.001        0.8 apriori(data = g, parameter = list(supp = 0.001, conf = 0.8))
```

Rysunek 12. Wczytanie danych i przeprowadzenie analizy reguł asocjacyjnych dla zbioru groceries.csv.

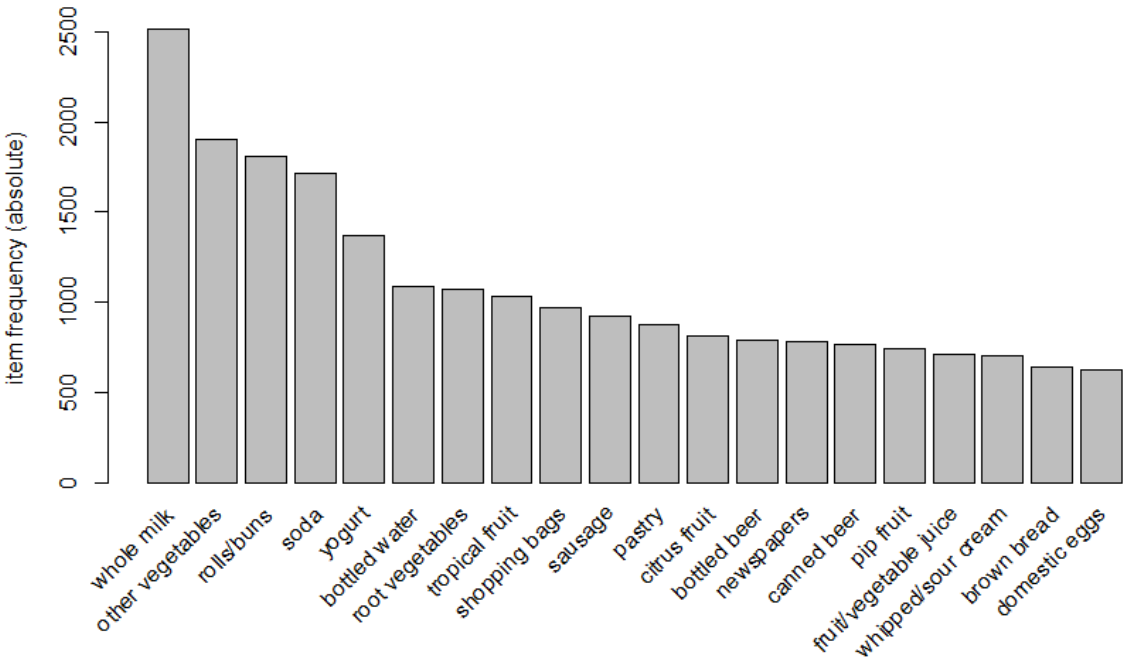
```
> r <- apriori(data=g, parameter=list(supp=0.001, conf=0.15, minlen=2), appearance = list(default="rhs", lhs="whole milk"), c
ontrol = list(verbose=F))
> r <- sort(r, decreasing=TRUE, by="confidence")
> inspect(r[1:5])
    lhs             rhs                support    confidence coverage lift     count
[1] {whole milk} => {other vegetables} 0.07483477 0.2928770  0.255516 1.513634 736
[2] {whole milk} => {rolls/buns}       0.05663447 0.2216474  0.255516 1.205032 557
[3] {whole milk} => {yogurt}           0.05602440 0.2192598  0.255516 1.571735 551
[4] {whole milk} => {root vegetables}  0.04890696 0.1914047  0.255516 1.756031 481
[5] {whole milk} => {tropical fruit}   0.04229792 0.1655392  0.255516 1.577595 416
```

Rysunek 13. Wyświetlenie posortowanych reguł asocjacyjnych.

Ćw. 6. – Praca Domowa

W ramach pracy domowej konieczne było dokonanie analizy reguł dla zbioru groceries, a następnie utworzenia dwóch plików zawierających konkretne reguły.

Najpierw przeprowadzono analizę zgodnie ze stroną „Basket Analysis with R”.



Rysunek 14. Wykres zawartości koszyka dla danych groceries.

Przeprowadzono analizę apriori dla tego zbioru danych.

```
    lhs                          rhs            support confidence coverage lift count
[1] {liquor, red/blush wine} => {bottled beer} 0.0019  0.90       0.0021   11.2 19
[2] {curd, cereals}          => {whole milk}   0.0010  0.91       0.0011    3.6 10
[3] {yogurt, cereals}        => {whole milk}   0.0017  0.81       0.0021    3.2 17
[4] {butter, jam}            => {whole milk}   0.0010  0.83       0.0012    3.3 10
[5] {soups, bottled beer}    => {whole milk}   0.0011  0.92       0.0012    3.6 11
```

Posortowano wyniki po wartości confidence (ufność).

```
    lhs                                   rhs          support confidence coverage lift count
[1] {rice, sugar}                      => {whole milk} 0.0012  1          0.0012   3.9  12
[2] {canned fish, hygiene articles}    => {whole milk} 0.0011  1          0.0011   3.9  11
[3] {root vegetables, butter, rice}    => {whole milk} 0.0010  1          0.0010   3.9  10
[4] {root vegetables, whipped/sour cream, flour} => {whole milk} 0.0017  1          0.0017   3.9  17
[5] {butter, soft cheese, domestic eggs} => {whole milk} 0.0010  1          0.0010   3.9  10
```

Najpierw należało wskazać reguły mówiące o tym, co kupują klienci, którzy kupili masło. W związku z tym posortowano reguły w taki sposób, aby po lewej stronie (lhs) występowało masło.

```
> trans_matrix <- as(Groceries, "ngCMatrix")
>
> transactions <- as(trans_matrix, "transactions")
>
> rules <- apriori(transactions, parameter = list(support = 0.001, confidence = 0.8))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       5   0.001      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
> butter_rules <- subset(rules, lhs %in% "butter")
```

Rysunek 15. Uzyskanie reguł zawierających produkty, w tym masło.

Drugim przypadkiem były reguły wskazujące, które towary kupują klienci wraz z masłem. Wobec tego wskazano, by po prawej stronie (rhs) występowało masło. Przy poziomie ufności 0.8 wskazanym w algorytmie apriori nie uzyskano żadnych wyników – wobec tego powtórzono algorytm z poziomem ufności 0.5.

```
> rules <- apriori(transactions, parameter = list(support = 0.001, confidence = 0.5))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.5    0.1    1 none FALSE            TRUE       5   0.001      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [5668 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
> together_with_butter_rules <- subset(rules, rhs %in% "butter")
```

Rysunek 16. Uzyskanie reguł zawierających masło po stronie prawej, aby uzyskać inne produkty kupowane wraz z nim.

Oba pliki z regułami zostały załączone wraz z niniejszym sprawozdaniem.

Wnioski.

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. W szczególności pozwala ono w łatwy sposób przeprowadzać analizę reguł asocjacyjnych bez konieczności ręcznego wpisywania poszczególnych poleceń z odpowiednimi parametrami.

Podczas analizy reguł asocjacyjnych w praktyczny sposób dowiedziałem się, jak taką analizę przeprowadzać oraz jaka jest jej użyteczność. W szczególności dobrym przykładem okazała się analiza zbioru Groceries, gdzie w praktyczny sposób udało się zapoznać z sensem i interpretacją takich reguł.