Raport 4

Grupowanie, wykonany za pomocą programu R

	Type	Calories	Protein	Fat	Sodium	Fiber	Carbo	Sugars	Potass
Type	1.00	-0.04	0.16	0.00	-0.23	-0.11	0.04	-0.11	-0.01
Calories	-0.04	1.00	0.03	0.51	0.30	-0.30	0.27	0.57	-0.07
Protein	0.16	0.03	1.00	0.20	0.01	0.51	-0.04	-0.29	0.58
Fat	0.00	0.51	0.20	1.00	0.00	0.01	-0.28	0.29	0.20
Sodium	-0.23	0.30	0.01	0.00	1.00	-0.07	0.33	0.04	-0.04
Fiber	-0.11	-0.30	0.51	0.01	-0.07	1.00	-0.38	-0.15	0.91
Carbo	0.04	0.27	-0.04	-0.28	0.33	-0.38	1.00	-0.45	-0.37
Sugars	-0.11	0.57	-0.29	0.29	0.04	-0.15	-0.45	1.00	0.00
Potass	-0.01	-0.07	0.58	0.20	-0.04	0.91	-0.37	0.00	1.00

Po usunięciu braku danych w zmiennej Type mamy jedynie jedną wartość (0), dlatego nie będziemy ją brali pod uwagę.

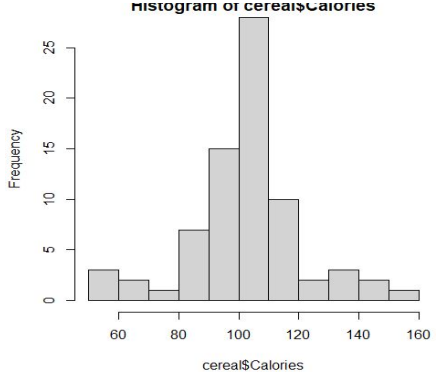
Można zauważyć korelację pomiędzy Potass i Fiber (0,91).

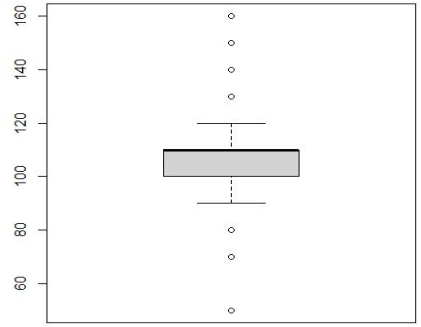
Zmienne z których będziemy korzystać to: Calories, Protein, Fat, Sodium, Carbo, Sugars oraz Potass.

Zbadano rozkłady zmiennych numerycznych, poprzez narysowanie ich histogramów i wykresów skrzynkowych. Na kolejnych slajdach można zobaczyć wyniki. Można zauważyć że większość zmiennych ma rozkład prawostronnie skośny i występują w nich obserwacje odstające. Może to pogorszyć jakość grupowania, więc zastosowano przekształcenie za pomocą funkcji logarytmu, a następnie zestandaryzowano. Po przekształceniach również narysowano histogramy i wykresy skrzynkowe. Wykonano również test Shapiro-Wilka.

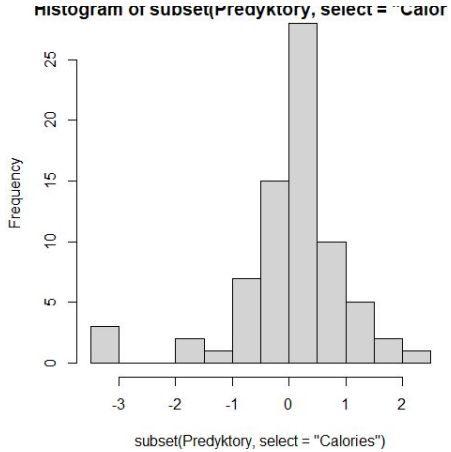
Histogramy i wykresy skrzynkowe dla zmiennej Calories

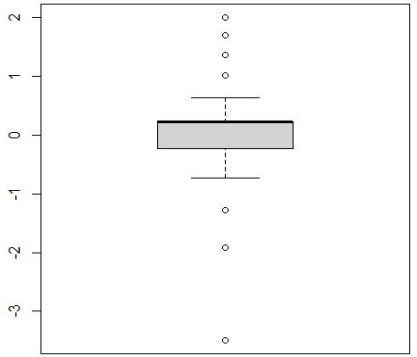
Przed przekształceniami





Po przekształceniach

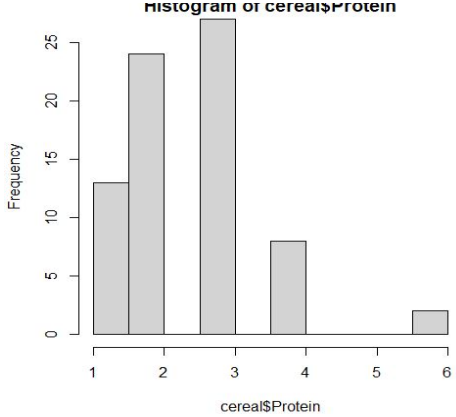


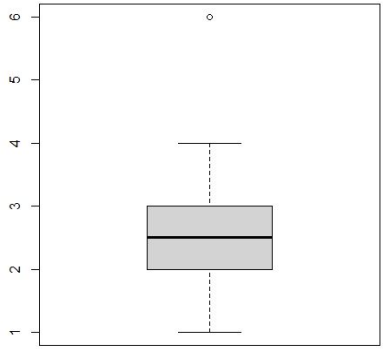



```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Calories")
W = 0.81723, p-value = 3.754e-08
```

Histogramy i wykresy skrzynkowe dla zmiennej Protein

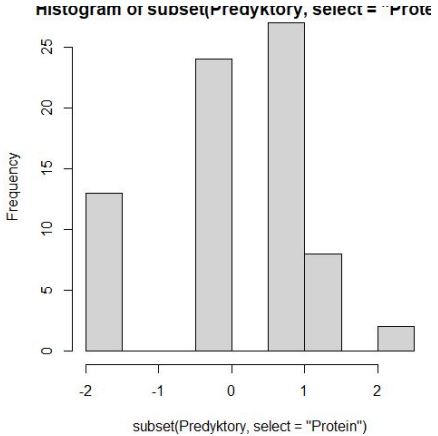
Przed przekształceniami

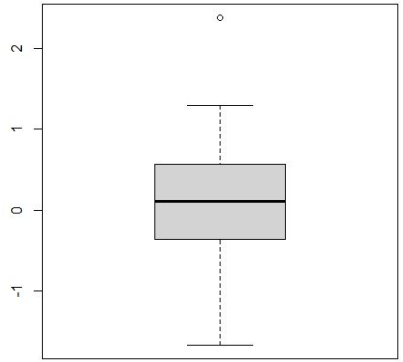




Po przekształceniach

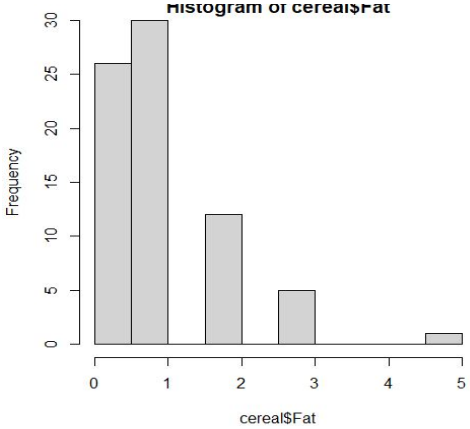
```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Protein")
W = 0.88858, p-value = 8.581e-06
```

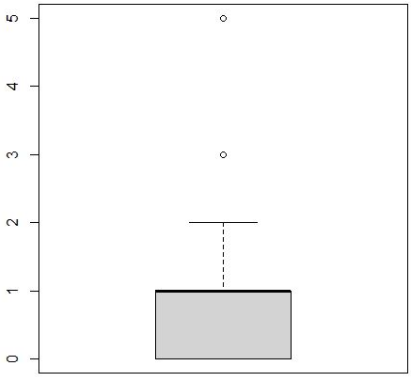




Histogramy i wykresy skrzynkowe dla zmiennej Fat

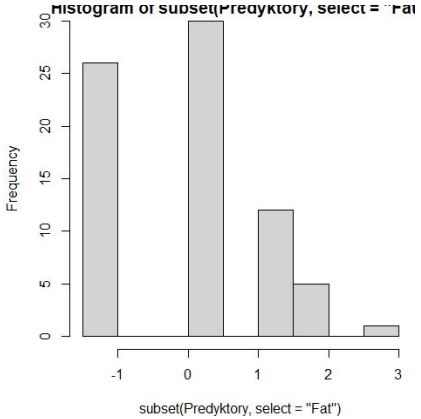
Przed przekształceniami

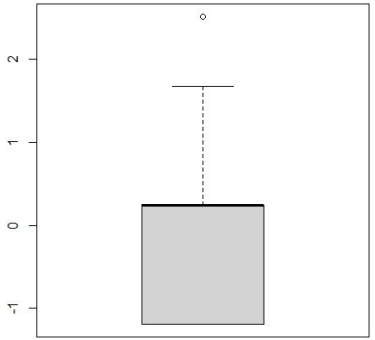




Po przekształceniach

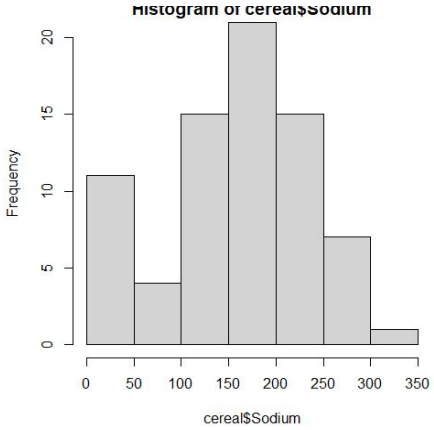
```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Fat")
W = 0.84534, p-value = 2.66e-07
```

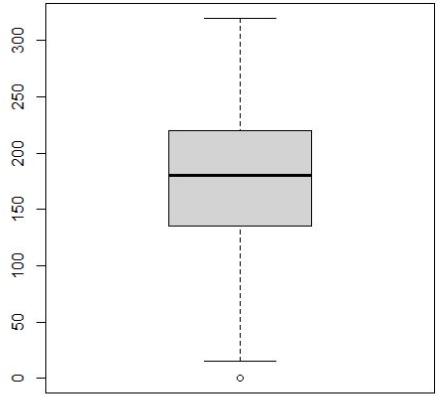




Histogramy i wykresy skrzynkowe dla zmiennej Sodium

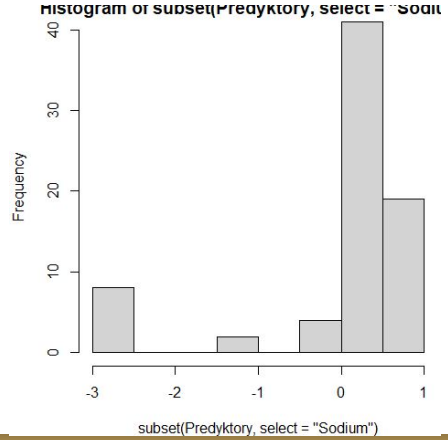
Przed przekształceniami

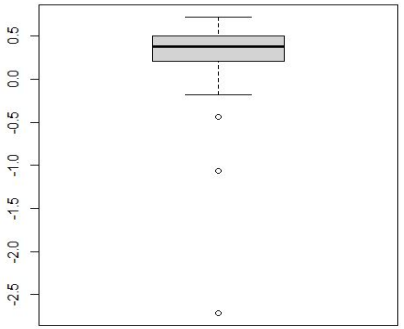




Po przekształceniach

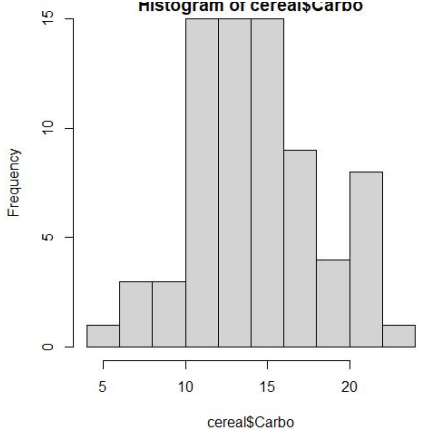
```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Sodium")
W = 0.57338, p-value = 2.455e-13
```

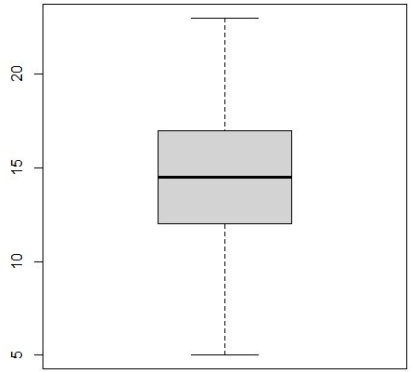




Histogramy i wykresy skrzynkowe dla zmiennej Carbo

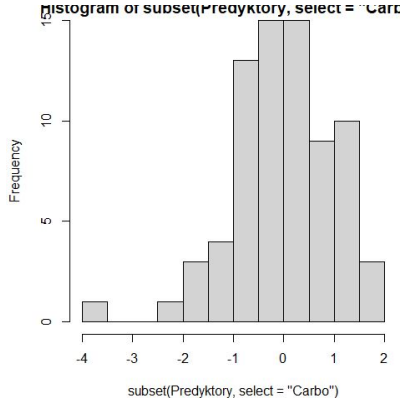
Przed przekształceniami

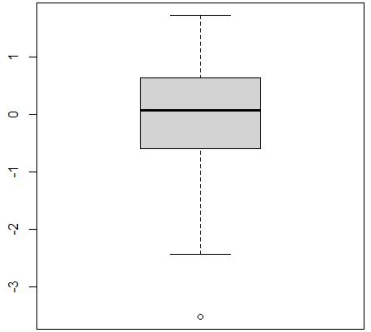




Po przekształceniach

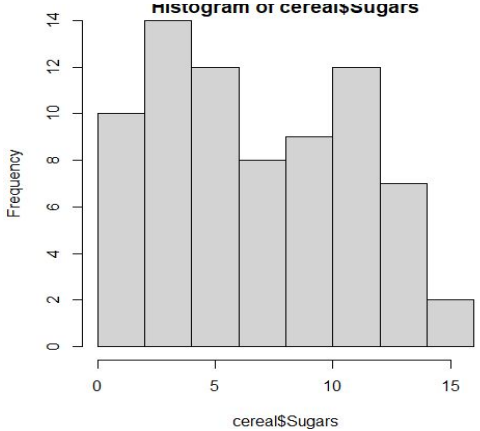
```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Carbo")
W = 0.95911, p-value = 0.01733
```

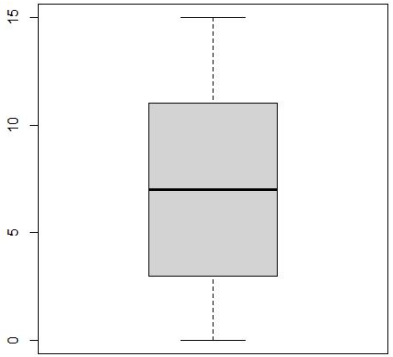




Histogramy i wykresy skrzynkowe dla zmiennej Sugars

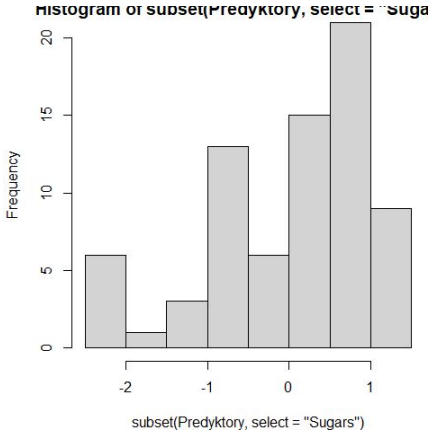
Przed przekształceniami

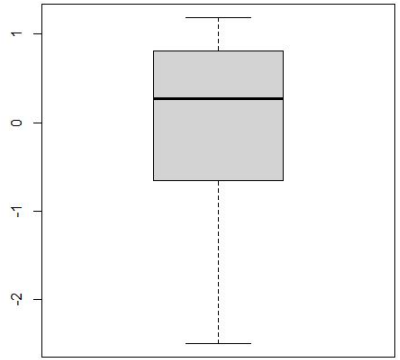




Po przekształceniach

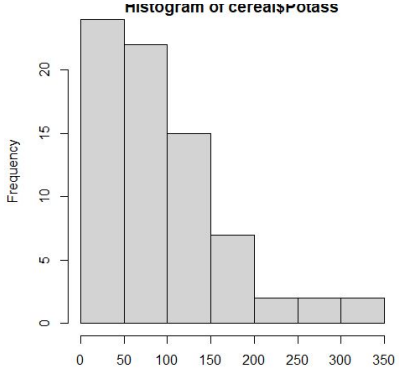
```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Sugars")
W = 0.86479, p-value = 1.173e-06
```

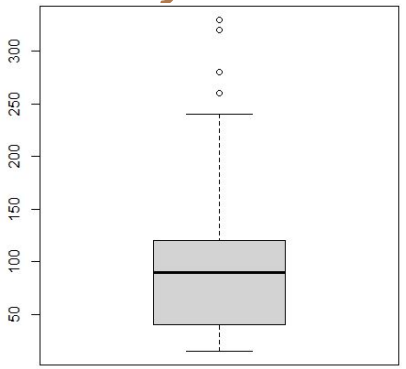




Histogramy i wykresy skrzynkowe dla zmiennej Potass

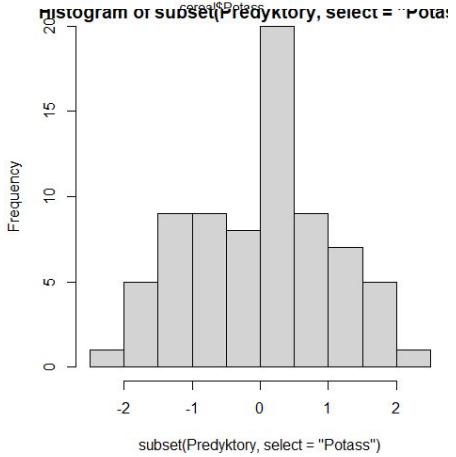
Przed przekształceniami

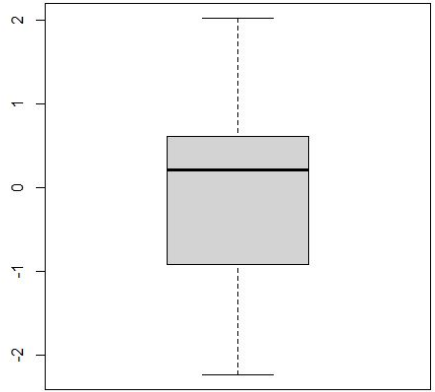




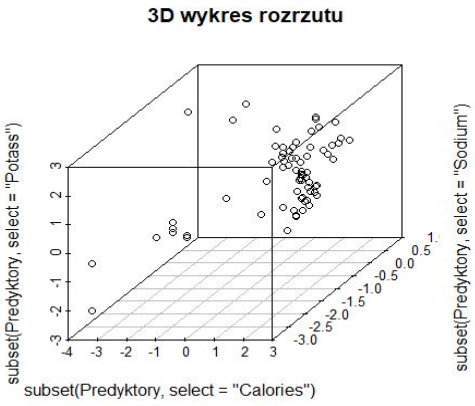
Po przekształceniach

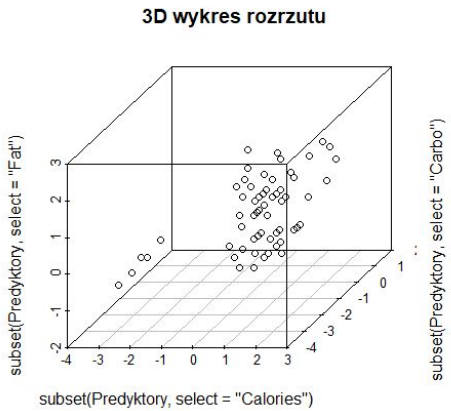
```
        Shapiro-Wilk normality test

data:  subset(Predyktory, select = "Potass")
W = 0.97735, p-value = 0.205
```

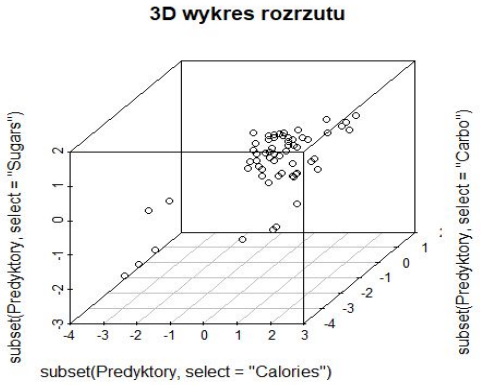




Przykładowe wykresy rozrzutu 3d

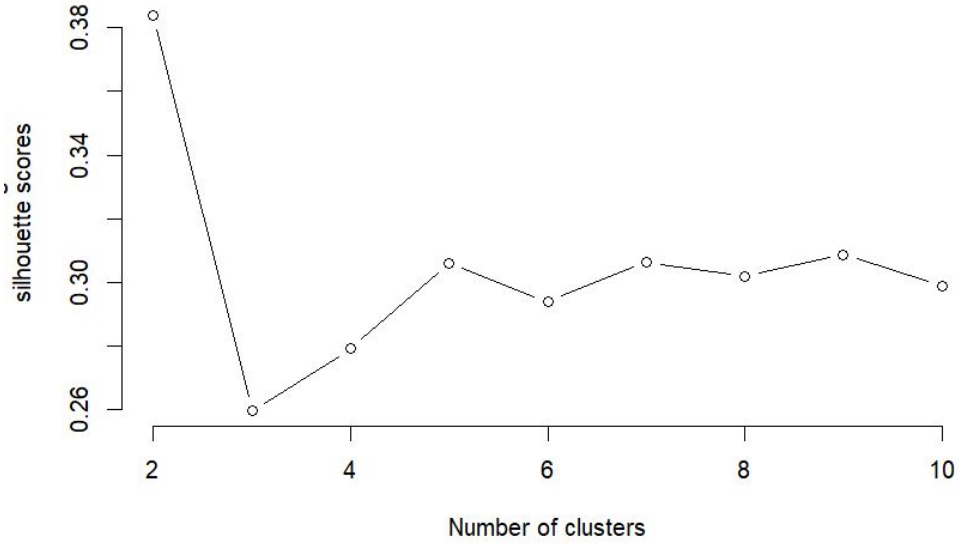






Z wykresów rozrzutu 3d nie można jednoznacznie powiedzieć na ile grup należy podzielić, ale przypuszczać można że mogą być to 3 grupy.

Miara Silhouette



Przeprowadzona miara Silhouette wykazuje że najlepsza podział byłby na 2 grupy.

Podział na 2 grupy

```
    Calories     Protein       Fat    Sodium       Carbo     Sugars      Potass
1  0.2217444 -0.02352657  0.143148  0.330674  0.004975023  0.2349598 -0.03142226
2 -1.6014874  0.16991409 -1.033847 -2.388201 -0.035930722 -1.6969316  0.22693856
> km$size
[1] 65  9
```

W przypadku podziału na 2 grupy powstała jedna liczna grupa i druga niewielka grupka. Ten podział jest nieoptymalny.

Podział na 3 grupy

```
    Calories    Protein        Fat     Sodium       Carbo     Sugars      Potass
1 -1.5140654  0.1167200 -1.0497708 -2.2557534 -0.01357912 -1.5389346  0.2252691
2  0.1742953 -0.7668397 -0.2897584  0.3880012  0.25861969  0.2267928 -0.8436647
3  0.2988501  0.7303646  0.6178117  0.3169217 -0.25437621  0.2541243  0.7732682
> km_new$size
[1] 10 32 32
```

W przypadku podziału na 3 grupy, jedna z grup jest mniejsza od dwóch pozostałych, ale wszystkie są dosyć liczne. Można stwierdzić, że podział jest lepszy niż w przypadku 2 grup.

1 grupa: Odchudzające

2 grupa: Niskopotasowe

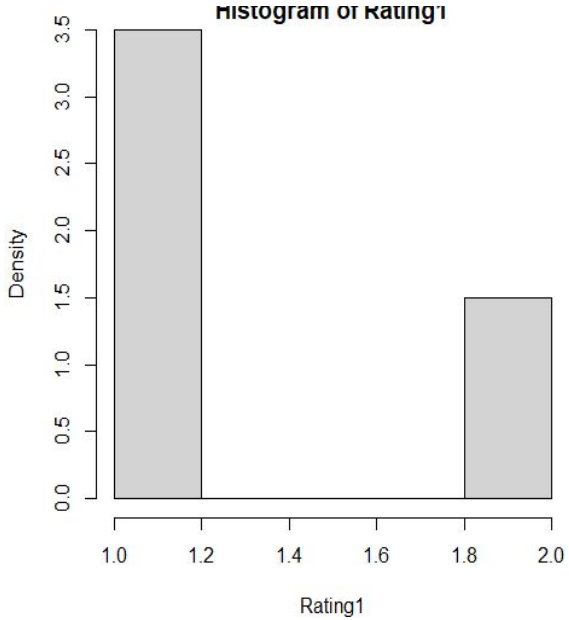
3 grupa: Wysokobiałkowe

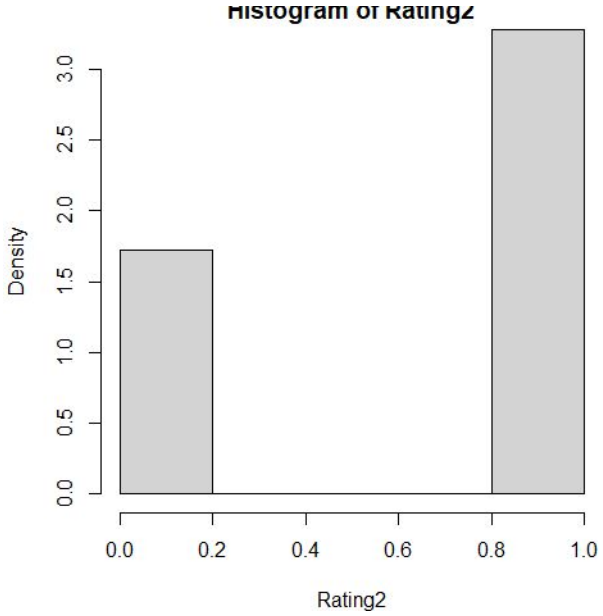
Odchudzające - płatki z mniejszą liczbą kalorii oraz mniejszą zawartością tłuszczu i cukru, mogą być to płatki dla osób które dbają o swoją wagę lub dla diabetyków

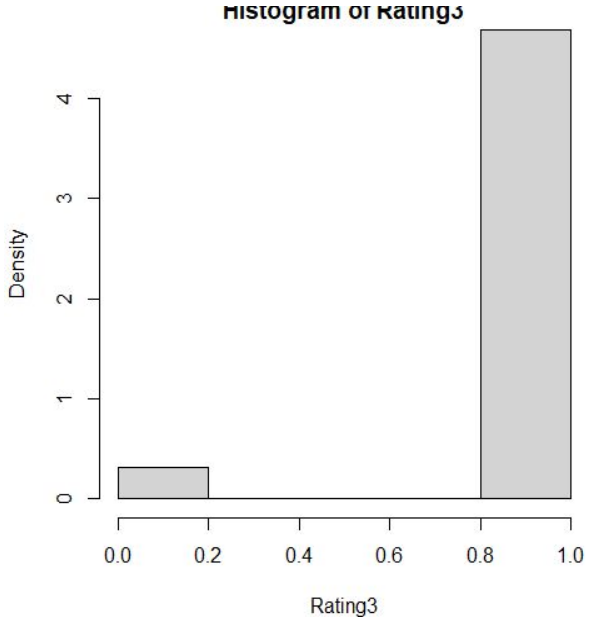
Niskopotasowe - płatki z mniejszą zawartością białka i potasu, mogą być to płatki dla osób chorujących na hiperkaliemie

Wysokobiałkowe - płatki z wyższą zawartością białka i tłuszczu, mogą być to płatki dla osób które chcą zwiększyć zawartość białka w swojej diecie

Wpływ na ocenę klienta







Do pierwszej grupy należą płatki które zostały ocenione przez klientów wysoko i średnio, w pozostałych grupach średnio i nisko.

Raport wykonała: Anna Cabaj