Laboratorium 16.03 – Techniki eksploracji danych wielowymiarowych

Maja Jarmoc, grupa: 1, nr indeksu: 274263

Cel ćwiczenia: podstawowa analiza danych znajdujących się w przygotowanym zestawie: 31 rodzajów płatków śniadaniowych wraz z informacjami o kaloriach, ilości białka, tłuszczu, błonnika, węglowodanów, cukrów (w gramach) i sodu (w miligramach); (porcja: 30g.)

1) Test Shapiro - Wilka

Test Shapiro-Wilka jest to test służący do oceny, czy zebrane przez nas wyniki posiadają rozkład normalny.

Hipoteza zerowa (H0) – próba badawcza pochodzi z populacji o normalnym rozkładzie.

Hipoteza alternatywna (HA) – próba badawcza ma rozkład odbiegający od krzywej Gaussa.

Gdy p > alfa (poziom istotności) – przyjmujemy hipotezę zerową, a odrzucamy hipotezę alternatywną.

Test Shapiro – Wilka (metoda z biblioteki scipy.stats) przeprowadziłam dla każdej zmiennej z następującymi wynikami (alfa = 0.05):

Zmienna *„calories”* ma rozkład odbiegający od normalnego z wartością p = 0.0006

Zmienna *„protein”* ma rozkład odbiegający od normalnego z wartością p = 0.00004

Zmienna *„fat”* ma rozkład odbiegający od normalnego z wartością p = 0.000109

Zmienna *„sodium”* ma rozkład normalny z wartością p = 0.5132

Zmienna *„fiber”* ma rozkład odbiegający od normalnego z wartością p = 0.0006177

Zmienna *„carbo”* ma rozkład odbiegający od normalnego z wartością p = 0.00175

Zmienna *„sugars”* ma rozkład normalny z wartością p = 0.16411

Nie była to jedyna metoda, której użyłam do określenia rozkładu zmiennej. Dodatkowo przeprowadzone testy:

1) czy wartość MIN/MAX > 0,1 ?
2) czy |d-m| < s ?
3) czy wartość r/s należy do przedziału <3;5> ?
4) czy |q| < 2 ?

MIN – wartość minimalna, MAX – wartość maksymalna, d = (MAX+MIN) / 2, m – wartość średnia, s – odchylenie standardowe, r – rozstęp (MAX – MIN), q – indeks skośności rozkładu

---

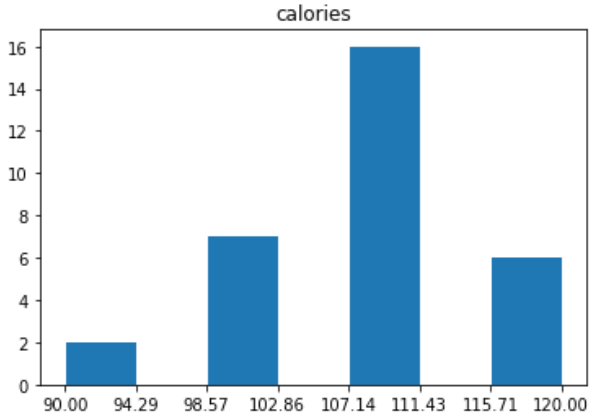
Wyniki z 4 przeprowadzonych testów odbiegające wynikiem od wyżej przeprowadzonego testu Shapiro-Wilka:

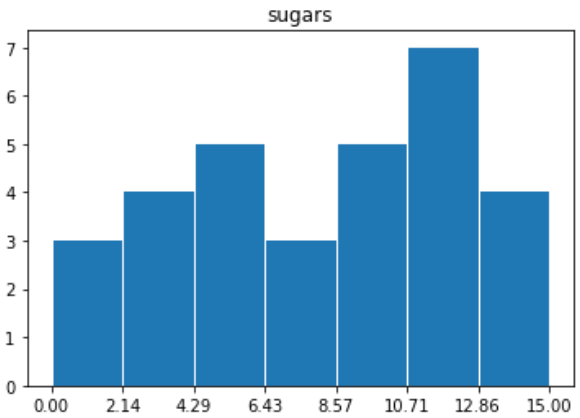
Zmienna calories ma rozkład normalny, po przyjrzeniu się histogramowi zmiennej tak też uznałam.

Zmienna *„carbo”* ma rozkład normalny, po przyjrzeniu się histogramowi uznaję wynik testu Shapiro-Wilka – nie ma rozkładu normalnego.

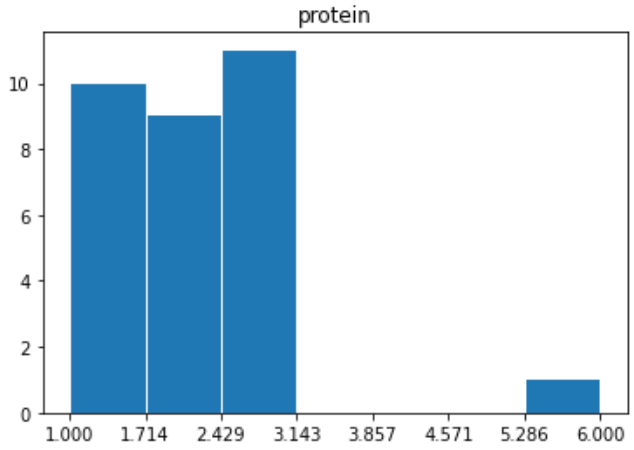
Zmienna *„sodium”* nie ma rozkładu normalnego, po przyjrzeniu się histogramowi nie uznaję rozkładu normalnego.

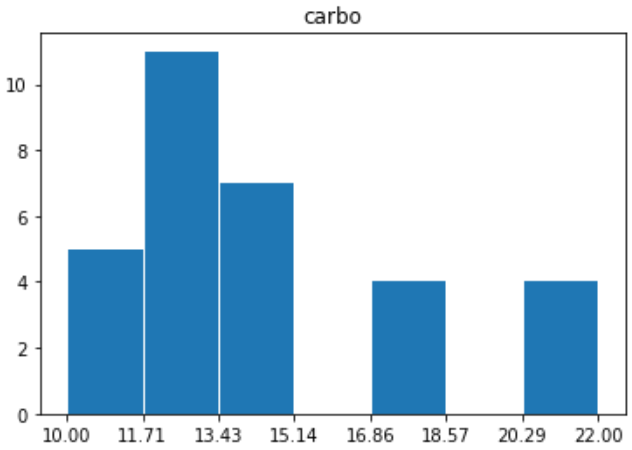
2) Histogramy dla zmiennych o rozkładzie normalnym:

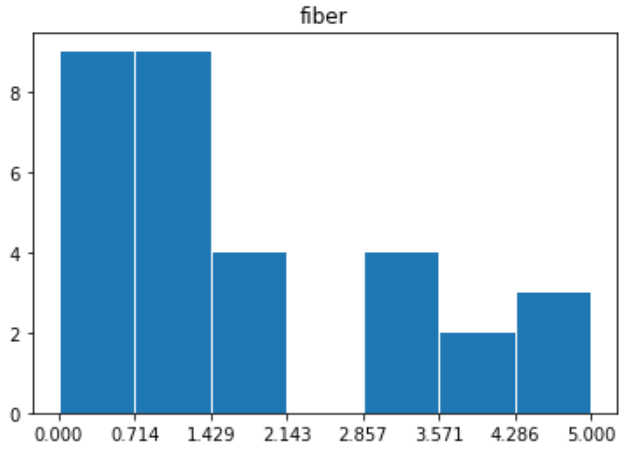


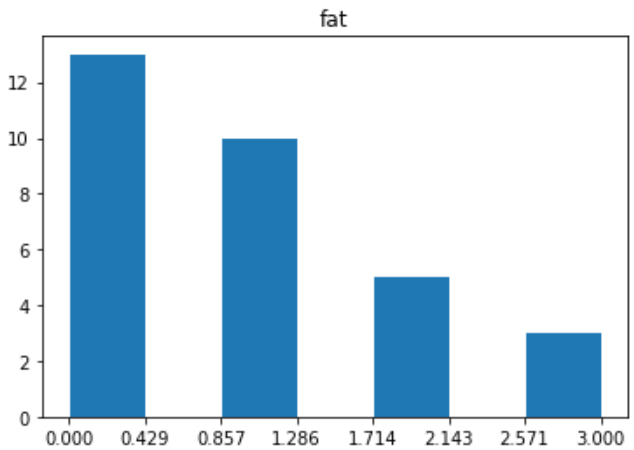


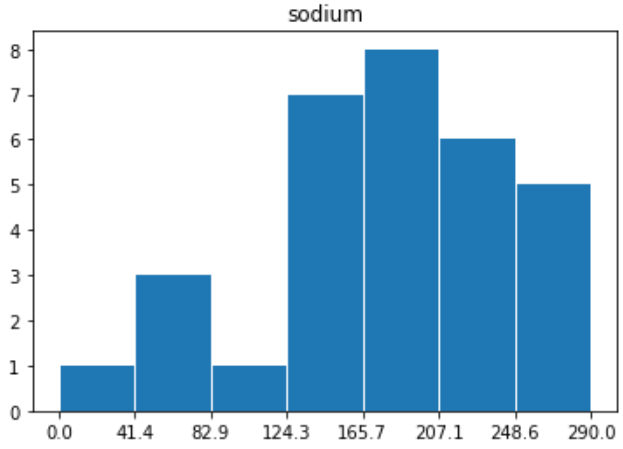
3) Histogramy dla zmiennych odbiegających od rozkładu normalnego:











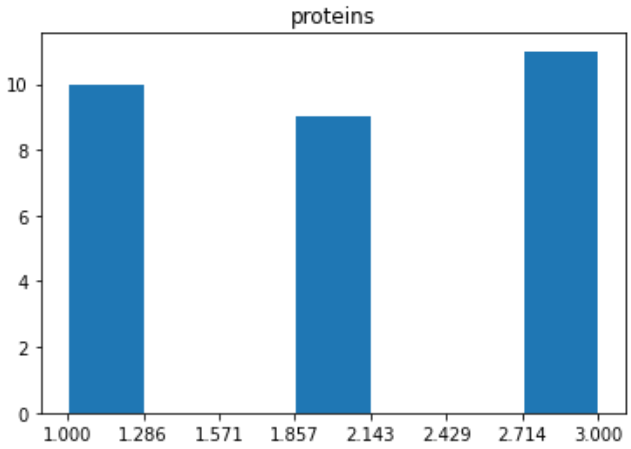
4) PUNKTY ODSTAJĄCE

Punkt odbiegający pojawił się w zmiennej „protein” i wynosił 6. Aby uzyskać rozkład symetryczny zdecydowałam się usunąć punkt odstający. Punkt odbiegający pojawił się w zestawie danych ze względu na płatki „Cheerios”, które posiadają 6g białka w porcji 30g.

Aby ustalić czy mogę usunąć punkt odstający zastosowałam metodę przedziału ufności z następującymi wynikami:

Przedział ufności: (0.2943257314913035, 3.7723409351753627)

Wartość 6 nie mieści się w podanym przedziale, mogę usunąć punkt odstający:



Histogram po usunięciu punktu odstającego

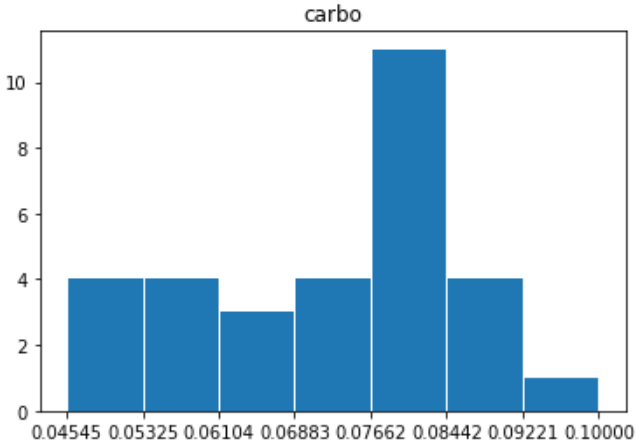
5) TRANSFORMACJA

Aby uzyskać rozkład normalny (bądź chociażby symetryczny) podjęłam się transformacji danych funkcjami matematycznymi.

Zmiennej „protein” po usunięciu punktu odstającego zdecydowałam się nie transformować, powodem jest brak rezultatu transformacji. Histogram jest symetryczny.

Zmienna *„fat”* oraz *„fiber”* także nie poddały się transformacji. Rozkład normalny lub symetryczny nie został uzyskany.

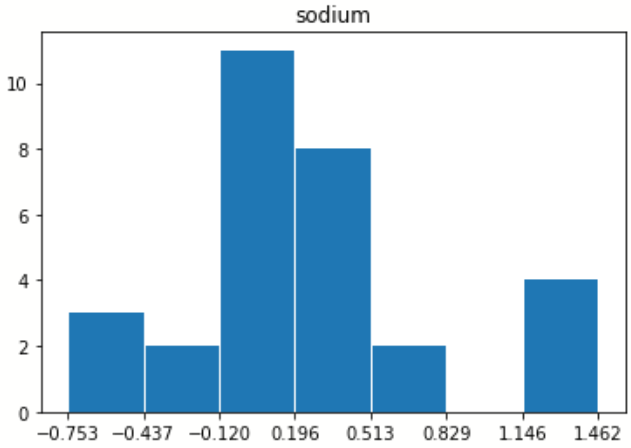
Zmienna *„carbo”* po transformacji funkcją $\frac{1}{x}$:



Test Shapiro-Wilka przeprowadzony po transformacji: p = 0.085 (większe od alfa = 0.05)

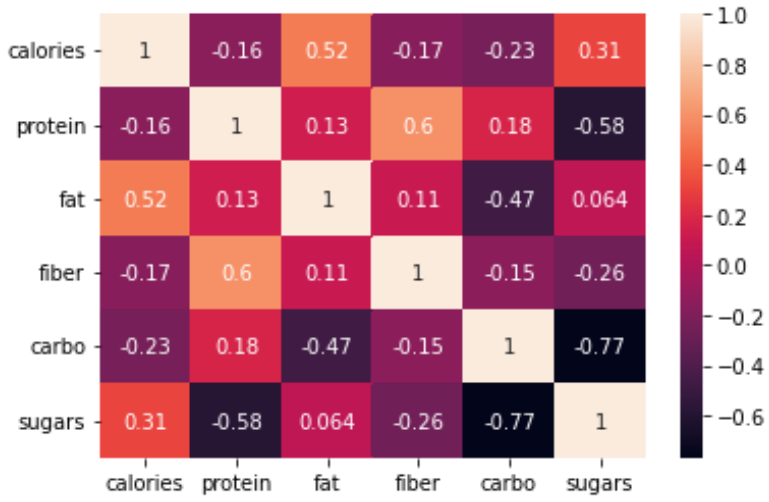
Zmienna *„sodium”* po transformacji funkcją $log_{10}(x\ /\ a - x)$, a = 300. Test Shapiro-Wilka przeprowadzony po transformacji:

P = 0.20 (większe od alfa = 0.05)



6) MACIERZ KORELACJI

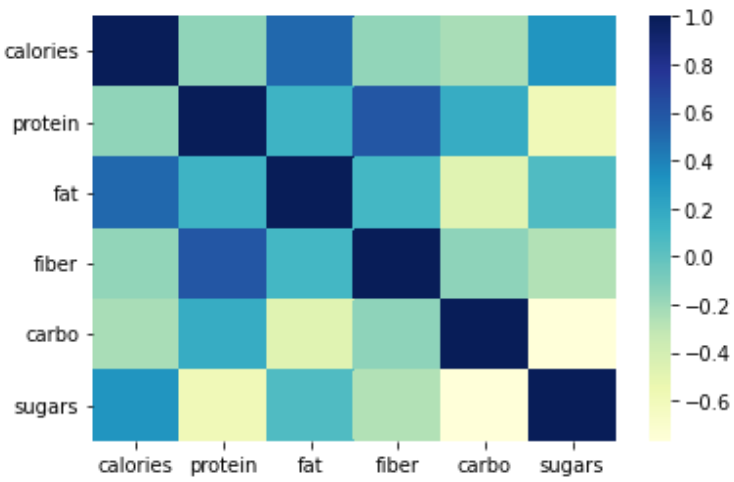
Wersja 1.



Wersja 2.

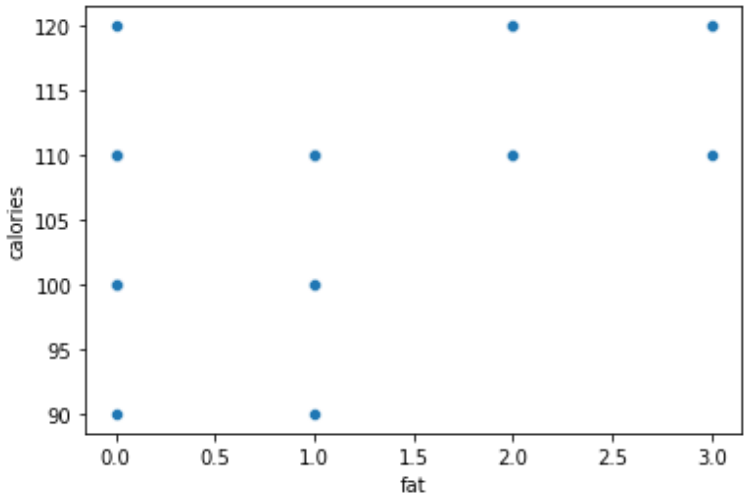
	calories	protein	fat	fiber	carbo	sugars
calories	1.000000	-0.155117	0.516178	-0.165818	-0.231855	0.314673
protein	-0.155117	1.000000	0.131463	0.595709	0.177784	-0.582014
fat	0.516178	0.131463	1.000000	0.108412	-0.467190	0.064485
fiber	-0.165818	0.595709	0.108412	1.000000	-0.148982	-0.258770
carbo	-0.231855	0.177784	-0.467190	-0.148982	1.000000	-0.768411
sugars	0.314673	-0.582014	0.064485	-0.258770	-0.768411	1.000000

Heatmap

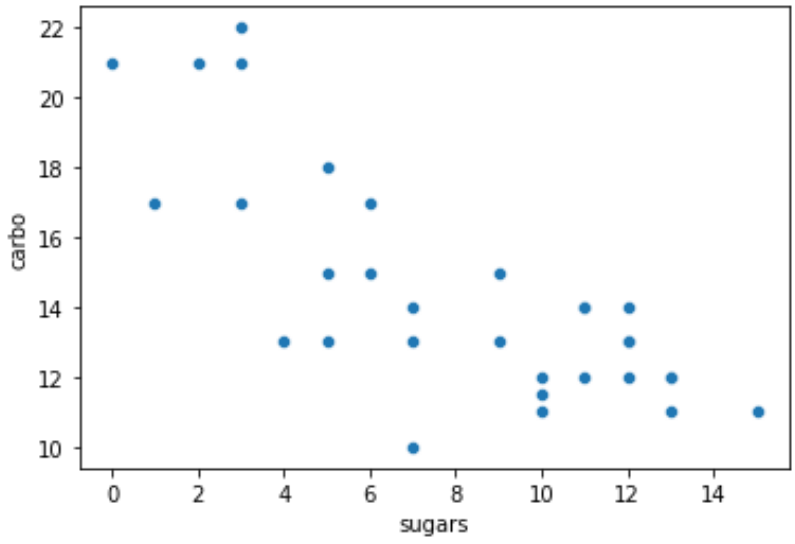


7) WYKRESY KORELACJI WYBRANYCH ZMIENNYCH

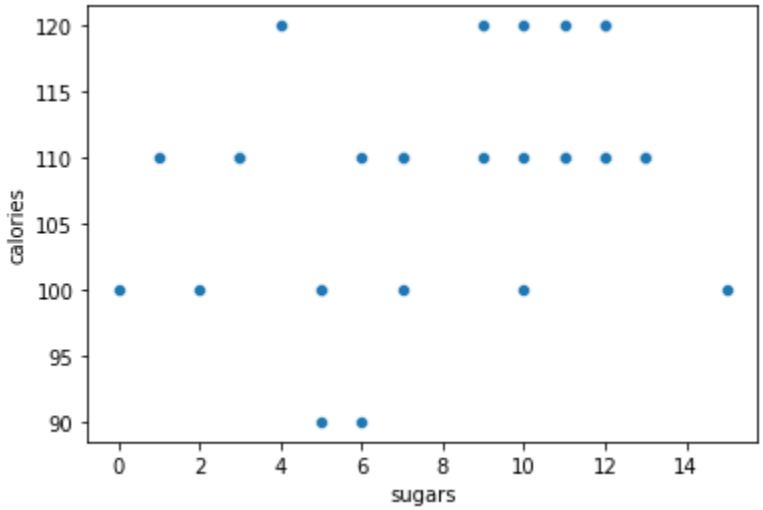
Korelacja kalorii i tłuszczu w porcji 30g. Wzrost kalorii przy wzroście gramów tłuszczu.



Korelacja cukrów i węglowodanów w porcji 30g. Korelacja jest widoczna, jednak ważna jest jej interpretacja. Teoretycznie w składzie płatków cukry zawierają się w węglowodanach, także wraz ze wzrostem ilości cukru węglowodany także powinny rosnąć. Dzieje się jednak na odwrót. Wykres wygląda tak ponieważ płatki zawierające dużo cukru są „mniej zdrowe" i tym samym posiadają mniej tych „zdrowszych węglowodanów" niż same cukry proste. W płatkach o dużej ilości węglowodanów cukry proste najczęściej zajmują mały procent węglowodanów.



Korelacja kalorii i cukrów w porcji 30g. Brak znaczącej korelacji.



Korelacja kalorii i węglowodanów w porcji 30g. Brak znaczącej korelacji.