ZAJĘCIA WPROWADZAJĄCE

1. Zasady zaliczenia:

Samodzielne wykonanie wszystkich zadanych ćwiczeń w pracowni komputerowej. Nieobecność można odrobić na zajęciach z drugą grupą lub na konsultacjach u prowadzącego.

- 4 kolokwia pisemne/wejściówki: hierarchiczna analiza skupień, analiza głównych składowych, grupowanie metodą k-średnich, samoorganizujące się mapy Kohonena.
- Analogicznie 4 sprawozdania.

Ocena: 40% kolokwia; **60%** sprawozdania.

Ocena może być podwyższona o połowę za aktywność.

Trzeba zaliczyć wszystkie wejściówki i sprawozdania!

Sprawozdania można poprawiać dowolną ilość razy. Wejściówki pisemnie, poprawa ustnie, maksymalna ocena na poprawie - 3. Poprawy wejściówek odbywają się poza zajęciami, na konsultacjach.

Konsultacje do ustalenia z prowadzącym zajęcia.
Wytyczne ogólne dotyczące sprawozdań zostaną przedstawione na zajęciach.

Literatura: A.1. wykorzystywana podczas zajęć

Skrypt do ćwiczeń laboratoryjnych przygotowywany przez pracowników Zespołu Chemometrii Środowiska

A.2. studiowana samodzielnie przez studenta

J. Mazerski: Podstawy chemometrii. Gdańsk: Wydawnictwo Politechniki Gdańskiej, 2000
M. Lutz: Python. Wprowadzenie. Helion, 2002
S. Raschka: Python. Uczenie maszynowe. Helion, 2016

Praca domowa: zebrać dane (6-10 zmiennych), => 30 obiektów

1) Dane powinny składać się z >30 obiektów o dowolnym charakterze, opisywanych przez 6-10 cech.
2) Cechy, opisujące obiekty, powinny być możliwe do przedstawienia, w sposób jednoznaczny, w postaci liczb. W związku z powyższym, cechy takie jak: kolor farby, smak owocu, przystojność aktora i funkcjonalność telefonu będą eliminowane na starcie przez Prowadzącego. Możliwe jest, co prawda, uwzględnienie zmiennych o charakterze zero-jedynkowym (0 = telefon nie posiada Bluetooth, 1 = telefon posiada Bluetooth), nie polecamy ich jednak z uwagi na potencjalnie niekorzystny wpływ na wyniki późniejszych analiz.
3) Wartości wszystkich cech, opisujących obiekty, muszą być sprecyzowane dla każdego obiektu. Oznacza to, iż niedopuszczalna jest nieznajomość nawet jednej wartości cechy dla pojedynczego obiektu.

ĆWICZENIE 1

Biblioteki: numpy, pandas, random.

1. Przygotować wektor wertykalny (A) zawierający 10 argumentów z przedziału 1:20 (array(10,0)).
2. Przygotować wektor wertykalny (B) zawierający 10 argumentów z przedziału 0:1 (array(10,0)).
3. Wykonać mnożenie dwóch wektorów A i B.
4. Następnie wykonać mnożenie skalarne wektora A oraz c = 2.
5. Stworzyć macierz 4x4 (W), w których argumenty powinny być losowymi liczbami całkowitymi o wartościach z przedziału 1:100. Czy typ ndarray i matrix w pythonie to to samo?
6. Wyznaczyć przekątną macierzy.
7. Przeprowadzić transponowanie macierzy ‘na piechotę’.

ĆWICZENIE 2 – kontrola pojedynczych zmiennych

Biblioteki: numpy, pandas, matplotlib

1. Import danych z excela.
2. Kontrola danych.

   Kontrola danych pomoże odpowiedzieć na następujące pytania:
   - Jaki jest charakter rozkładu poszczególnych zmiennych?
   - Czy istnieją przesłanki o konieczności dokonania transformacji?
   - Czy wśród zestawu obiektów znajdują się punkty odbiegające?

Aby znaleźć odpowiedź na te pytania należy obliczyć szereg prostych statystyk:

- Wartość minimalną zmiennej MIN
- Wartość maksymalną zmiennej MAX
- Stosunek MIN/MAX
- Rozstęp rozkładu zmiennej (r = MAX-MIN)
- Środek rozkładu zmiennej (d =(MAX+MIN)/2)
- Wartość średnią zmiennej (m)
- Odchylenie standardowe z populacji zmiennej ($s = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$)
- Indeks skośności rozkładu (q = $3\frac{m-w}{s}$), w – mediana, porównać z wartością obliczoną przy użyciu biblioteki *scipy.stats.skew*

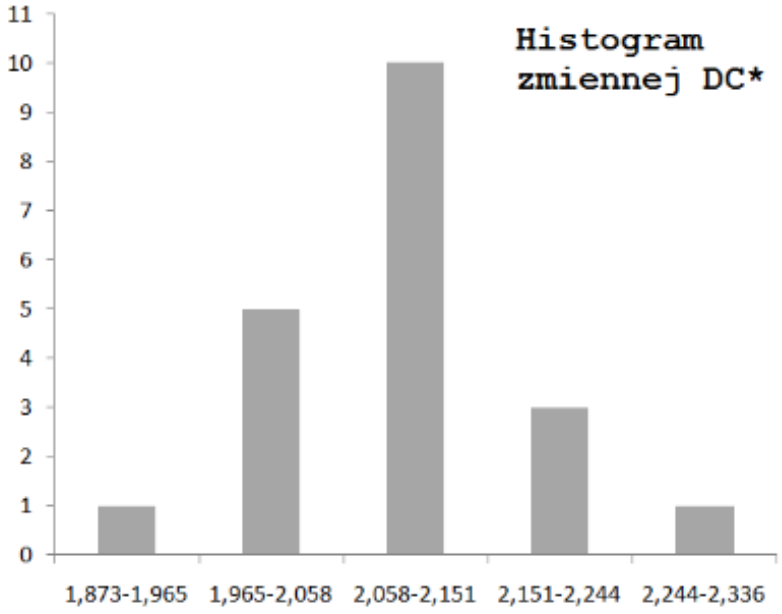
Otrzymane dla każdej zmiennej charakterystyki należy teraz poddać następującym testom:
1) czy wartość MIN/MAX > 0,1 ?
2) czy |d-m| < s ?
3) czy wartość r/s należy do przedziału <3;5> ?
4) czy |q| < 2 ?

Jeżeli dla danej zmiennej odpowiedzi na cztery powyższe pytania brzmią TAK, zmienna ma prawdopodobnie rozkład zbliżony do normalnego i - przynajmniej do czasu następnego ćwiczenia - przestaje być "interesująca".

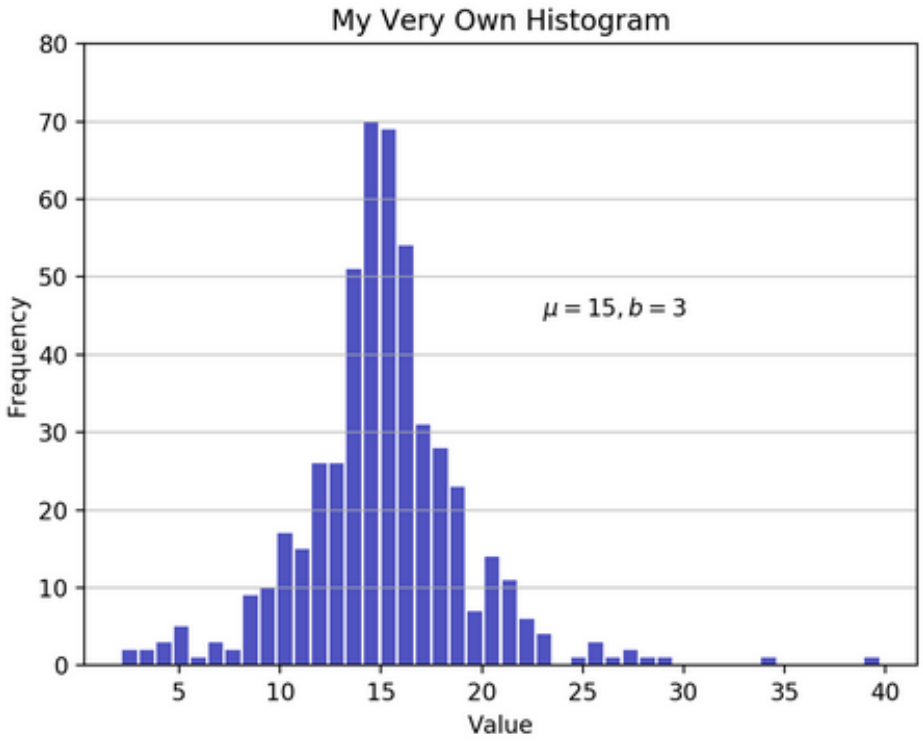
Jeżeli zaś, dla danej zmiennej, odpowiedź na przynajmniej jedno powyższe pytanie brzmi NIE, zmienna staje się "podejrzana". Przyczyny takiego stanu rzeczy mogą być dwie: i) wśród wartości zmiennej występuje punkt lub punkty odbiegające; ii) rozkład zmiennej jest silnie asymetryczny lub wielomodalny.

Aby ustalić, dlaczego rozkład danej zmiennej odbiega od rozkładu normalnego, należy wykonać **histogram** wartości tej zmiennej.

Histogram powinien mieć wyraźnie zdefiniowane granice przedziału dla każdego zakresu.



Rys.1. Prawidłowo wykonany histogram.



Rys.2. Nieprawidłowo wykonany histogram. Źródło: https://realpython.com/python-histograms/ ; dostpęp: 07.03.2022

Dobranie odpowiedniej ilości przedziałów w histogramie może stanowić pewien problem. Jednak wykorzystanie prostej reguły empirycznej zwyczajowo wystarcza, mianowicie:

Liczba przedziałów, k, nie powinna być większa niż ¼ liczby wartości zmiennej, n, czyli $k \leq n/4$.

Oprócz tego warto przestrzegać kilku poniższych reguł w trakcie tworzenia histogramu:

i) przedziały muszą mieć jednakową szerokość
ii) krańce przedziałów powinny być liczbami możliwie "okrągłymi", np. 1,5 zamiast 1,48
iii) skrajne punkty rozkładu powinny przypadać w pobliżu środka, a nie na krawędzi skrajnych przedziałów histogramu
iv) przy nieparzystej liczbie przedziałów wartość średnia powinna znaleźć się w pobliżu środka środkowego przedziału, a w przypadku parzystej liczby przedziałów jak najbliżej granicy pomiędzy środkowymi przedziałami.

Jeżeli histogram zmiennej odbiega od rozkładu normalnego, a ściśle mówiąc jest silnie prawo lub lewo skośny należy wykonać transformację wszystkich wartości zmiennej przy użyciu odpowiednich funkcji matematycznych.

Charakter zmiennej	Przykłady funkcji transformujących
stosunek MIN/MAX<0,1; jest silnie prawoskośna	$x^* = \log_{10}(x)$, $x^* = \log_{10}(x+a)$; $x+a > 0$
zmienna jest silnie lewoskośna	$x^* = \log_{10}(a-x)$; $a > x_{MAX}$
zmienna ma postać % i x < 15%	$x^* = \log_{10}(x)$
zmienna ma postać % i x > 85%	$x^* = \log_{10}(a-x)$, $a = 100$
inne	$x^* = \log_{10}(x/(a-x))$, $x^* = 1/x$, inne

Tabela 1. Chemometria w praktyce. Ćwiczenia laboratoryjne; Jan Mazerski, Tomasz Laskowski

Po dokonaniu transformacji należy za każdym razem sprawdzić czy histogram dla danej zmiennej przybiera formę rozkładu zbliżonego do normlanego.