Celem ćwiczenia będzie podstawowa analiza danych znajdujących się w przygotowanym przez Studenta zestawie. Część zadań jest powtórzona względem pierwszych laboratoriów. Optymalne zatem będzie wykorzystanie elementów kodu z poprzednich zajęć.

**Biblioteki:** numpy, pandas, matplotlib, seaborn, …

1. **Charakter rozkładu zmiennych**
2. Zaimportować dane.
3. Obliczyć podstawowe statystyki zawarte w skrypcie do pierwszych zajęć. Ustalić (ponownie ze skryptu z zajęć pierwszych) czy któraś ze zmiennych może odbiegać od rozkładu normalnego. Co do zmiennych, które wydają się niepokojące należy przeprowadzić dalsze obliczenia.
4. Następnie należy sprawdzić, czy rozkład ma charakter normalny przy użyciu wzoru Shapiro-Wilka.

Text, letter

Description automatically generated

1. Porównać wartość obliczonego W oraz wartość W obliczoną przy użyciu *scipy.stats.*
2. Następnie wykonać histogramy dla zmiennych, które mogą odbiegać od rozkładu normalnego.
3. Po wykonaniu histogramów rozkładu wszystkich "podejrzanych" zmiennych, należy   
   przyjrzeć im się i odpowiedzieć na następujące pytania (dla każdej ze zmiennych):   
   1) czy rozkład zmiennej jest wielomodalny?;   
   2) jeżeli rozkład zmiennej jest jednomodalny - czy jest symetryczny lub zbliżony do   
   symetrycznego?;   
   3) jeżeli rozkład zmiennej jest jednomodalny - czy jest silnie lewo- lub prawoskośny?;   
   4) czy na histogramie widoczny jest punkt odbiegający?   
     
   Jeżeli odpowiedź na pytanie 1) brzmi TAK - należy zostawić zmienną w spokoju.   
   Zmienna taka może odegrać dużą rolę w analizie podobieństwa obiektów lub w analizie   
   skupień.   
     
   Jeżeli odpowiedź na pytanie 2) brzmi TAK - należy zostawić zmienną w spokoju.   
   Pomimo, iż jej rozkład nie jest normalny, możną ją z powodzeniem stosować praktycznie we   
   wszystkich metodach chemometrycznych.   
     
   Jeżeli odpowiedź na pytanie 4) brzmi TAK - należy przejść do sekcji związanej z określaniem punktów odbiegających.   
     
   Jeżeli odpowiedź na pytanie 3) brzmi TAK (rozkład jest silnie lewo- lub   
   prawoskośny) - należy dokonać transformacji zmiennej. Transformacja zmiennej polega na   
   przekształceniu wszystkich wartości danej zmiennej za pomocą odpowiedniej funkcji   
   matematycznej. Po dokonaniu transformacji należy ponownie wykonać histogram z   
   otrzymanych wartości danej zmiennej i ocenić, czy jej rozkład stał się przynajmniej   
   symetryczny.
4. **Po dokonaniu transformacji zmiennych należy przygotować nową tabelę danych, w**   
   **której wartości zmiennych transformowanych zastąpią wartości "oryginalne". Należy**   
   **również zaznaczyć, które zmienne zostały poddane transformacji (najczęściej czyni się to**   
   **poprzez dodanie \* do etykiet zmiennych), a także odnotować - blisko tabeli - postaci**   
   **funkcji transformujących.**
5. **Punkty odbiegające**

Należy teraz podjąć decyzję, czy obiekt, który jest charakteryzowany przez   
odbiegającą wartość danej zmiennej, powinien pozostać w tabeli danych, czy też należy go   
usunąć. Decyzję o ewentualnym usunięciu obiektu należy podjąć w oparciu o podany poniżej algorytm postępowania:   
1) Należy tymczasowo usunąć wartość odbiegającą zmiennej i wykonać nowy   
histogram tej zmiennej.   
2) Jeżeli rozkład zmiennej (po usunięciu wartości odbiegającej) stał się zbliżony do   
normalnego bądź przynajmniej symetryczny, metodą przedziału ufności (o   
niej za chwilę) należy ocenić, czy obiekt opisywany przez tę wartość usunąć z   
tabeli, czy też nie.   
3) Jeżeli po usunięciu wartości odbiegającej rozkład zmiennej nie uległ   
"poprawie", należy przywrócić usuniętą wartość i dokonać transformacji   
zmiennej.   
4) Jeżeli po dokonaniu transformacji zmiennej jej rozkład stał się symetryczny,   
nie należy usuwać "podejrzanego" obiektu z tabeli.   
5) Jeżeli po dokonaniu transformacji zmiennej na histogramie w dalszym ciągu   
widoczny jest punkt odbiegający, należy tymczasowo usunąć wartość   
odbiegającą transformowanej zmiennej i wykonać nowy histogram   
transformowanej zmiennej.   
6) Jeżeli rozkład transformowanej zmiennej (po usunięciu wartości odbiegającej)   
stał się symetryczny, metodą przedziału ufności należy ocenić, czy usunąć   
"podejrzany" obiekt, czy też nie.

1. **Metoda przedziału ufności.**

Załóżmy, iż nasza "podejrzana" zmienna przyjmuje 25 wartości, przy czym jedna z nich   
jawi się na histogramie jako wartość odbiegająca. Tymczasowo usuwamy ją z zestawu danych -   
pozostaną 24 wartości. Dla tych 24 wartości obliczamy wartość średnią (m) i odchylenie   
standardowe średniej (s), oraz odczytujemy z tabeli wartość testu t-Studenta dla poziomu   
istotności 0,05 oraz n-1 stopni swobody (w tym przypadku n = 24 - jest to liczba wartości   
po odrzuceniu "podejrzanego" obiektu; zatem n-1 = 23). Następnie, obliczamy krańce   
przedziału ufności:   
  
xmin = m-t∙s;   
xmax = m+t∙s.   
  
Jeżeli "podejrzana" wartość mieści się w przedziale wyznaczonym przez te granice   
- nie należy usuwać z tabeli obiektu przez nią opisywanego; jeżeli zaś nie mieści się -   
obiekt ten można usunąć z zestawu danych.

1. **Operacje na macierzach**

Używając poleceń z biblioteki *numpy*, wykonaj operacje:

* Dodania macierzy **X +** **Y**,
* Odejmowania **X - Y**,
* (standardowego) mnożenia macierzy („dot product”) **X · Y**,
* dla każdej macierzy osobno:
  + zsumowania elementów z wszystkich kolumn,
  + a następnie zsumowania elementów z wszystkich wierszy,
* znalezienie macierzy odwrotnej do macierzy **X**, czyli **X-1.**

1. **Analiza przygotowanego zestawu danych -**

Używając **przygotowany przez Ciebie zestaw danych**:

* Dla każdej zmiennej w zestawie danych, znajdź wartość minimalną, maksymalną, medianę (tj. 2. kwartyl) i średnią, 1. i 3. kwartyl; uzyskane wyniki zapisz do pliku w formacie **.csv.**
* Wykonaj wykres rozkładu dla każdej zmiennej (w formie histogramu lub kernel density estimator (KDE)/ wykres estymatora jądrowego gęstości).
* Oblicz korelację między zmiennymi w zestawie danych; uzyskaną macierz korelacji przedstaw w postaci mapy cieplnej (“heatmap”).
* Sporządź wykresy korelacyjne dla wszystkich par zmiennych w postaci wykresów punktowych.