Ćwiczenie 3

Praca ze zbiorem danych. Korelacja.

1. Wprowadzenie

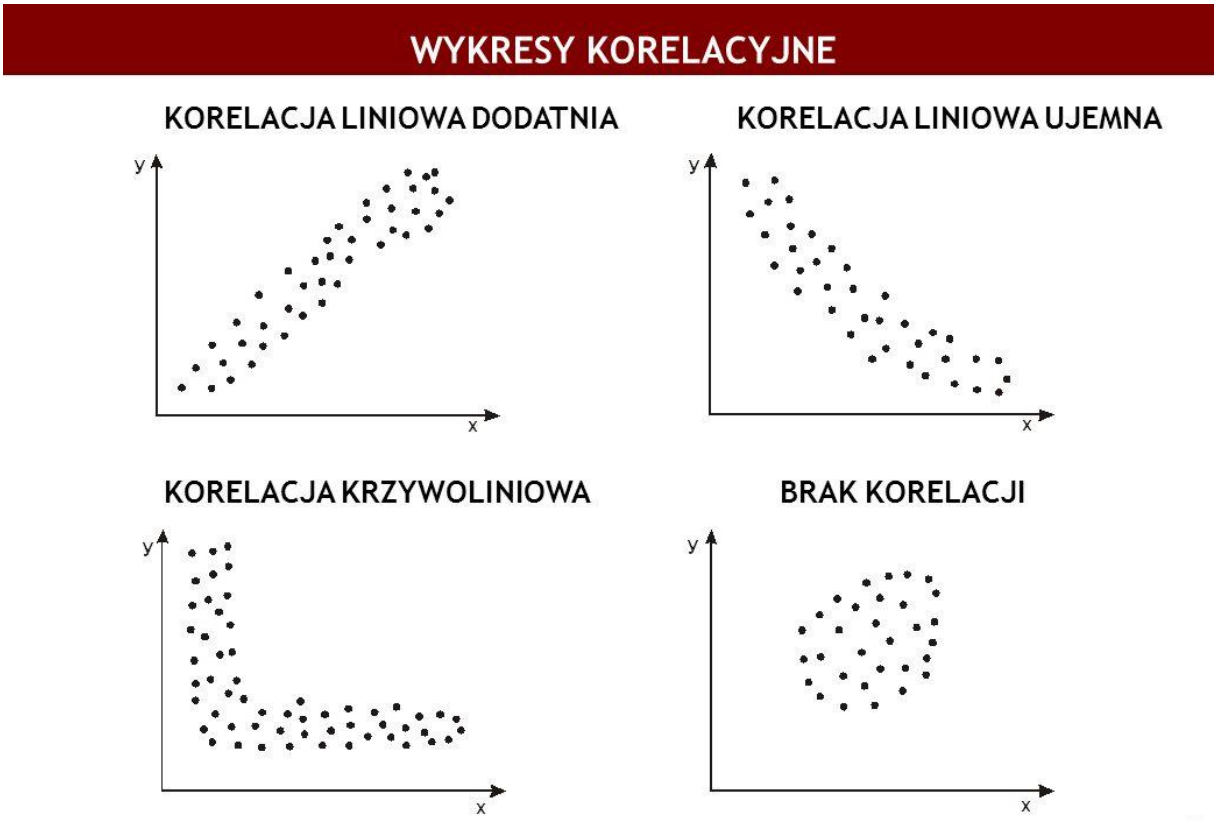
Celem ćwiczenia jest zapoznanie się z pracą ze zbiorem danych o charakterze losowym w którym występują związki o charakterze korelacyjnym między wartościami.

Korelacja jest to pewien rodzaj zależności pomiędzy zmiennymi losowymi, z których każda wyznaczona jest przez pewną cechę, ze względu na którą bada się daną populację.

Siłę związku pomiędzy zmiennymi losowymi wyraża się za pomocą współczynnika korelacji. Wartość współczynnika korelacji jest normowana do przedziału domkniętego [-1,1].

Dla współczynnika korelacji liniowej (Pearsona), wartość:

- r_{xy}=1 oznacza dokładną liniową dodatnią korelację pomiędzy cechami x i y;
- r_{xy}=0 oznacza całkowity brak korelacji pomiędzy cechami x i y;
- r_{xy}=-1 oznacza dokładną liniową ujemną korelację pomiędzy cechami x i y.



Gdzie:
C(X,Y) – kowariancja pomiędzy cechami X i Y;
S_x(S_y) – odchylenie standardowe cechy X (Y);

$$C(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\backslash(x_i - \overline{x}) \cdot (y_i - \overline{y})$$

Korzystając z odpowiedniej skali (Stanisza, Guillforda), można wyrazić stopień zależności między badanymi cechami.

Przedział	Zależność	Współczynnik
0,00±0,2	Słaba	Prawie nic nieznaczący
±0,20±0,40	Niska	Wyraźna, ale słaba
±0,40±0,70	Umiarkowana	Rzeczywisty
±0,70±0,90	Wysoka	Znaczny
±0,90±1,00	Bardzo wysoka	Pewny

Znak współczynnika r_{xy} informuje o kierunku zależności:

- **+(dodatni)** - zależność liniowa dodatnia, wraz ze wzrostem wartości jednej cechy rosną średnie wartości cechy drugiej;
- **-(ujemny)** - zależność liniowa ujemna, wraz ze wzrostem wartości jednej cechy maleją średnie wartości cechy drugiej;

2. Pobieranie danych

W tym ćwiczeniu należy wygenerować dane statystyczne wykorzystując w tym celu dane zawarte na stronie Głównego Urzędu Statystycznego (http://stat.gov.pl). Na głównej stronie GUS należy skorzystać z ikony „Bank Danych Lokalnych". Dane może pozyskać na dwa sposoby:

- W okienku „szukaj" należy wpisać nas słowo. Przypuśćmy, że chcemy sprawdzić przeciętne wynagrodzenie pracowników w poszczególnych województwach, w tym celu w polu „szukaj" wpisujemy np. wynagrodzenie. Wówczas z wyszukanej listy wybieramy najbardziej odpowiednią opcję i w tym celu klikamy na „stosik monet" .

- Z ikonki „Dane" wybieramy „Dane wg dziedziny" , szukamy dziedziny w której mogą znaleźć się interesujące na dane. W naszym przykładzie jest to kategoria „Podmioty gospodarcze", Grupa " podmioty niefinansowane", podgrupa „przychody i koszty".

Po odnalezieniu danych i wybraniu „stosika monet" w tabelach zaznaczamy lata, które nas interesują i zatwierdzamy białą strzałką .

Kolejny krok polega na wyeksportowaniu bazy danych do Excela. W tym celu wybieramy ikonkę „Export" > „XLS tablica przestawna".

3. Zadania do wykonania

Na podstawie wygenerowanych danych należy obliczyć kowariancję i współczynnik korelacji Pearsona dla następujących grup danych:

- liczba ludności – liczba rozwodów;
- liczba ludności – zawarte małżeństwa;

- liczba rozwodów – liczba urodzeń;
- wydatki na jedną osobę – liczba rozwodów;
- wydatki na jedną osobę – liczba urodzeń;