DATA MINING 5

Zadanie 1

Poniższa tabela przedstawia macierz danych dla dwóch osób A oraz B.

Wzrost (cm)	Waga (kg)	Staz (lata)	Zarobki (tys.)	Ocena (pkt.)	Piętro	Dzieci	Odległosc (km)	Ubezp.	Wzrost (cm)
A	190	88	3	3.5	7	6	1	25	TAK
B	172	70	12	4.3	5	1	4	12	NIE

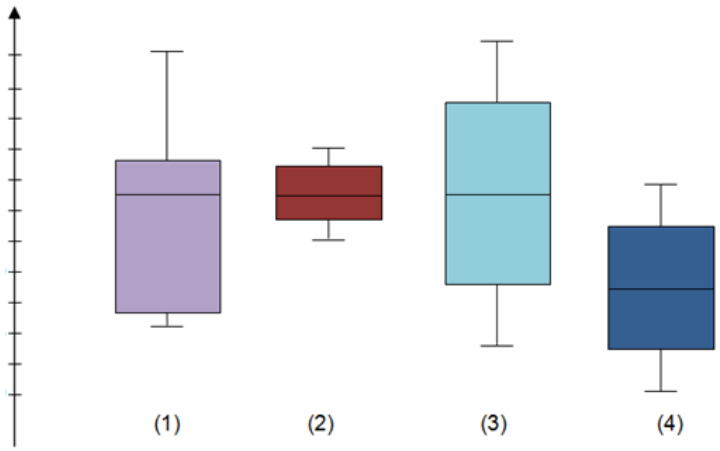
1. Pobierz powyższe dane z pliku **PersonsData.csv**.
2. Oblicz **odległość Euklidesową** d(A,B)
3. Co się dzieje z tą odległością jeżeli zmienimy skalę w przypadku zarobków (tyś → zł)?
4. Korzystając z **odchylenia standardowego** wyeliminuj wpływ skali (podziel wartość każdego atrybutu przez odchylenie standardowe).
5. Dla A i B oblicz **odległość Minkowskiego** i **miejską**.

Zadanie 2

Dla wielowymiarowych danych binarnych z pliku **BinaryData.csv** wyszukaj obiekty o najmniejszych odległościach dla odległości obliczanych według formuł **Jacard'a** i **Dice'a**.

Zadanie 3

Wśród studentów przeprowadzono badanie na temat oceny poziomu studiów (zajęć, infrastruktury etc.). Badanie sprowadzało się do odpowiedzi na kilkanaście pytań i wypełnieniu pewnego formularza. Pytania podzielone były na 4 kategorie. Wyniki zobrazowano na poniższym wykresie.



1. Jak można zinterpretować małą wysokość pudełka w przypadku kategorii 2?
2. Jak można zinterpretować dużą wysokość pudełek w przypadku kategorii 1 i 3?
3. Jak można zinterpretować dużą różnicę wysokości pudełek 3 i 4?
4. Jak można zinterpretować tę samą wartość mediany w przypadku kategorii 1, 2 i 3?
5. Jak można zinterpretować przesunięcie w pionie pudełek 2 i 4?
6. Jaka jest różnica między wynikami w kategorii 1 i 2?

Zadanie 4

Poniżej wypisane są graniczne wartości **naprężenia powodujących pękanie asfaltu** (w MPa), wyznaczone w oparciu o szereg niezależnych testów.

30, 75, 79, 80, 85, 105, 126, 130, 138, 140, 149, 149, 152, 156, 161, 166, 173, 179, 182, 184, 198, 223, 240, 242, 245, 247, 254, 274, 291, 384, 470

1. Narysuj **histogram**. Przetestuj różne ilości przedziałów.
2. Policz **modę**, **medianę** i **średnią**.
3. Jaka jest **skośność** rozkładu?
4. Oblicz **kwartyle** i wartość **rozstępu kwartylowego**.
5. Narysuj **wykres pudełkowy** dla tych danych wykorzystując bibliotekę **matplotlib**:

   https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.boxplot.html

Zadanie 5 (http://endrju.ovh.org/statystyka/statystyka_opisowa.doc)

Oto wyniki pierwszej serii konkursu Pucharu Świata w skokach narciarskich w Titisee-Neustadt 3 lutego 2007. Odległości w metrach są pogrupowane w przedziały długości 5 metrów.

i	x_{0i}	x_{1i}	n_i
1	105	110	1
2	110	115	6
3	115	120	10
4	120	125	10
5	125	130	15
6	130	135	2
7	135	140	6

Dane możesz pobrać z pliku **SkokiNarciarskie.csv**.

1. Narysuj **histogram**.
2. Oblicz **średnią długość** skoku.
3. Policz **modę**, **medianę** i **średnią**.
4. Jaka jest **skośność** rozkładu?
5. Oblicz **kwartyle**. Wyznacz rozstęp **międzykwartylowy.**
6. Sporządź **wykres pudełkowy**.