Ćwiczenie SOW

Statystyki opisowe i wizualizacja danych

Część teoretyczna

Materiał z wykładu dot. statystyk opisowych i wizualizacji danych.

Zadania do wykonania

Dokonaj opisu statystycznego i wizualizacji danych charakteryzujących samochody.

1. Wczytaj i zmodyfikuj zbiór danych:

```
data = pd.read_csv('carbig.csv')

np.random.seed(int(nr_gr * r_k))

# Replace NaN
columns = ['Acceleration', 'Displacement', 'Horsepower', 'Weight', 'MPG']
for col in columns:
    data[col] = data[col].fillna(data[col].mode()[0])

# Select columns
x = data[columns].values

# Add random noise (1% variation)
noise_multiplier = 1 + (np.random.rand(*x.shape) * 2 - 1) / 100
x = x * noise_multiplier
```

gdzie za `nr_gr` wstaw numer swojej sekcji a za `r_k` aktualny rok kalendarzowy.

Objaśnij powyższe polecenia. Podejrzyj i opisz zmienne.

2. Wyznacz wartości minimalne i maksymalne oraz miary średnie (slajdy 22 i 23) dla każdej zmiennej numerycznej (kolumny macierzy x).

3. Wyznacz rozstęp, kwartyle, rozstęp międzykwartylowy, odchylenie ćwiartkowe i typowy obszar zmienności (slajd 24) dla każdej zmiennej numerycznej. Utwórz wykresy pudełkowe dla wszystkich zmiennych numerycznych (`boxplot`).

4. Wyznacz miary rozrzutu (slajd 26 i 27) dla każdej zmiennej numerycznej.

5. Utwórz histogramy dla każdej zmiennej numerycznej (`hist`; slajd 32).

6. Wyznacz współczynniki asymetrii (`skew`) oraz kurtozy (`kurtosis`) dla każdej zmiennej numerycznej (slajdy 30 i 31).

7. Utwórz wykres `probplot` dla każdej zmiennej numerycznej. Wyjaśnij co on obrazuje.

8. Utwórz i zinterpretuj wykres `parallel_coordinates`. Uwaga, przed wykonaniem wykresu dokonaj standaryzacji danych (`zscore`) i przekształć dane do postaci DataFrame. Dodaj do DataFrame jeszcze jedną kolumnę 'Cylinders'. Potraktuj tę kolumnę jako `class_column` (patrz opis funkcji `parallel_coordinates`).

9. Wyznacz współczynniki korelacji liniowej dla każdej pary zmiennych numerycznych (slajd 33).

10. Utwórz wykres obrazujący współzależności pomiędzy każdą parą zmiennych numerycznych (`scatter_matrix`; slajd 47).

11. Wyznacz dodatkową zmienną (`Zuzycie`) reprezentującą zużycie paliwa w litrach na 100 km (na podstawie zmiennej MPG). Przyjmij 1 mila = 1,6 km, 1 galon = 3,79 litra. Sporządź wykres zależności zużycia paliwa od MPG. Utwórz wykresy zależności `Zuzycie od Acceleration, Displacement, Horsepower, Weight`.

12. Jakie jest średnie zużycie paliwa w dla samochodów z 3, 4, 5, 6 i 8 cylindrami?

Co powinno znaleźć się w sprawozdaniu

A) Cel ćwiczenia.
B) Treść zadania.
C) Raport z wykonania ćwiczenia. Dla każdego punktu: polecenia Pythona, wyniki (tabele, wykresy), komentarz.
D) Wnioski końcowe.

Zadania dodatkowe dla ambitnych

1. Zaimplementuj ćwiczenie w innym środowisku, np. Matlab, R, C#, ...

Przykładowe zagadnienia i pytania zaliczeniowe

1. Statystyki opisowe i ich implementacje w Pythonie
2. Metody wizualizacji danych i ich implementacje w Pythonie
3. Materiał ze sprawozdania.

Do przygotowania na następne zajęcia

1. Zapoznać się z instrukcją do kolejnego ćwiczenia.
2. Zapoznać się z częścią teoretyczną do kolejnego ćwiczenia.