


Informatyka Stosowana		
Laboratorium 6	Pandas Profiling - zautomatyzowana analiza danych	 <b>POLITECHNIKA BYDGOSKA</b> Wydział Telekomunikacji, Informatyki i Elektrotechniki
Przedmiot	Eksploracyjna analiza danych	
Prowadzący	mgr inż. Gracjan Kątek	

## 1. Wprowadzenie

W dzisiejszym świecie analiza danych jest kluczowym elementem podejmowania decyzji w niemal każdej dziedzinie – od biznesu, przez naukę, aż po technologię. Praca z dużymi zbiorami danych może być jednak czasochłonna i wymagająca, zwłaszcza w fazie eksploracyjnej, gdzie celem jest zrozumienie struktury danych, rozkładów poszczególnych zmiennych, identyfikacja braków i potencjalnych problemów w danych.

Tutaj na scenę wchodzi pandas-profiling – potężne narzędzie do automatycznego generowania raportów eksploracyjnych, które pozwala szybko i efektywnie zdobyć kluczowe informacje o danych.

### Czym jest pandas-profiling?

pandas-profiling to biblioteka w Pythonie, która automatyzuje proces wstępnej analizy danych. Umożliwia wygenerowanie szczegółowego raportu o:

- ➔ Podstawowych statystykach: liczba rekordów, liczba brakujących danych, typy danych.
- ➔ Rozkładach zmiennych: średnia, mediana, odchylenie standardowe, rozkład wartości.
- ➔ Brakujących danych: identyfikacja kolumn z brakującymi wartościami oraz ich procentowy udział.
- ➔ Relacjach między zmiennymi: korelacje, wykrywanie potencjalnych zależności.
- ➔ Wartościach odstających: identyfikacja anomalii i nietypowych wartości.

### Dlaczego warto używać pandas-profiling?

- Oszczędność czasu: Zamiast pisać dziesiątki linii kodu do eksploracji danych, można uzyskać kompleksowy raport jednym poleceniem.
- Wszechstronność: Narzędzie działa na różnych typach danych (liczbowych, kategoriowych, tekstowych).
- Wizualizacja: Raporty zawierają intuicyjne wykresy i diagramy, które ułatwiają zrozumienie danych.
- Identyfikacja problemów: Automatyczne wykrywanie problemów, takich jak brakujące wartości, wartości odstające czy błędne typy danych.

## 2. Zadania do samodzielnego wykonania

### Zadanie 1: Wygenerowanie podstawowego raportu

Wczytaj zestaw danych o samochodach (plik CSV). Skorzystaj z biblioteki **pandas-profiling**, aby wygenerować podstawowy raport opisujący dane. Następnie odpowiedz na pytania:

- Które kolumny mają brakujące dane?
- Czy w danych znajdują się wartości odstające?

### Zadanie 2: Analiza korelacji

Korzystając z wyników raportu wygenerowanego w zadaniu 1, zidentyfikuj kolumny, które są silnie skorelowane (wartość korelacji powyżej 0.8 lub poniżej -0.8). Wyświetl dane tylko dla tych kolumn i sprawdź, czy ich usunięcie wpłynęłoby na analizy.

### Zadanie 3: Rozbudowany raport

Wygeneruj bardziej zaawansowany raport, zmieniając ustawienia **pandas-profiling**, aby:

- Włączyć analizę wartości odstających (outliers).
- Zwiększyć dokładność analizy liczbowej poprzez dostosowanie liczby próbek do większego zestawu danych.

Wyświetl najciekawsze wykryte cechy danych (np. nietypowe rozkłady, kolumny z małą ilością unikalnych wartości).

#### **Zadanie 4: Usuwanie lub imputacja brakujących danych**

Na podstawie raportu zdecyduj, jak poradzić sobie z brakującymi wartościami:

- Usuń kolumny z dużą liczbą braków (powyżej 50%).
- Uzupełnij brakujące wartości w pozostałych kolumnach za pomocą średniej, mediany lub najczęściej występującej wartości.

#### **Zadanie 5: Optymalizacja typów danych**

Na podstawie raportu zoptymalizuj typy danych w zestawie, np.:

- Zamień kolumny liczbowo-kategoryczne na typ `category`.
- Przekonwertuj liczby całkowite na bardziej efektywny typ (np. `int8` zamiast `int64`, jeśli dane to umożliwiają). Następnie sprawdź, jak zmienił się rozmiar DataFrame w pamięci.

#### **Zadanie 6: Własna eksploracja danych**

Wybierz jedną z kolumn o nietypowym rozkładzie (zgodnie z raportem). Zwizualizuj jej dane (histogram, wykres pudełkowy) i odpowiedz na pytania:

- Jakie cechy można zauważyć?
- Czy kolumna wymaga dodatkowych przekształceń (np. logarytmowania, normalizacji)?