


Informatyka Stosowana		
Laboratorium 2	<i>Problemy przy pracy na prawdziwych danych</i>	 POLITECHNIKA BYDGOSKA Wydział Telekomunikacji, Informatyki i Elektrotechniki
Przedmiot	Eksploracyjna analiza danych - laboratorium	
Prowadzący	mgr inż. Gracjan Kątek	

1. Wprowadzenie

W pracy z rzeczywistymi danymi często napotykamy na różnorodne wyzwania, które nie występują w danych sztucznie generowanych lub starannie przygotowanych do celów dydaktycznych. Takie dane mogą pochodzić z rozmaitych źródeł (takich jak systemy CRM, pliki Excel, internetowe API, bazy danych SQL) i nie zawsze są w pełni kompletnie lub dokładnie opisane. Mogą zawierać błędy, brakujące wartości, wartości odstające, duplikaty, niezgodności w formatach danych, a także złożone struktury.

Typowe problemy, na które natrafiamy w praktyce, obejmują:

- ☑ **Brakujące wartości** – wartości, które zostały pominięte lub usunięte, co może powodować problemy podczas analizy i modelowania danych.
- ☑ **Wartości odstające** – wartości, które są nieoczekiwanie duże lub małe w porównaniu do pozostałych danych, mogą zakłócać analizy, szczególnie statystyczne.
- ☑ **Błędy w formacie** – dane, które mają nieprawidłowy typ lub format (np. liczby zapisywane jako teksty, niezgodność w formatach daty), co może wymagać konwersji lub korekty.
- ☑ **Duplikaty** – powtarzające się wiersze lub rekordy, które mogą zniekształcać wyniki analizy i prowadzić do błędnych wniosków.
- ☑ **Niespójności i błędy typograficzne** – różne zapisy tych samych wartości (np. „PLN” vs. „pln”) lub błędy w nazwach.

W niniejszym laboratorium nauczysz się identyfikować oraz radzić sobie z powyższymi problemami, a także stosować metody obróbki, które pomogą w standaryzacji i oczyszczaniu danych.



2. Zadania do wykonania

Zadanie 1. Znajdowanie brakujących wartości i ich obsługa

Wczytaj dane z pliku CSV i przeanalizuj, czy w zestawie danych występują brakujące wartości. Wykorzystaj metodę `isnull()` oraz `sum()`, aby określić, które kolumny zawierają najwięcej brakujących danych. Następnie wyświetl procent brakujących wartości dla każdej kolumny. Kolejnym krokiem jest wybór odpowiedniej metody obsługi braków – uzupełnij wartości średnią (dla danych liczbowych), najczęściej występującą wartością (dla danych kategorycznych), a także usuń wszystkie wiersze zawierające więcej niż 50% braków.

Zadanie 2. Usuwanie duplikatów

Zidentyfikuj i usuń duplikaty w danych. Najpierw użyj metody `duplicated()` oraz `sum()`, aby policzyć liczbę powtarzających się wierszy. Następnie usuń duplikaty za pomocą `drop_duplicates()`. Po wykonaniu tej operacji sprawdź, czy liczba wierszy w DataFrame zmniejszyła się zgodnie z oczekiwaniami.

Zadanie 3. Identyfikacja i obsługa wartości odstających

Przeanalizuj dane pod kątem wartości odstających. Skorzystaj z metody `describe()`, aby sprawdzić, czy w kolumnach liczbowych występują wartości, które znacznie odbiegają od reszty danych (np. wartości maksymalne). Wybierz jedną z kolumn, w której odchylenia są najbardziej widoczne, a następnie zastosuj z-score lub metodę międzykwartylową (IQR), aby wykryć wartości odstające. Usuń wiersze z wartościami odstającymi z wybranej kolumny.

Zadanie 4. Korekcja błędów typograficznych i formatowania

Sprawdź, czy w kolumnach tekstowych występują różne formaty lub błędy typograficzne. Na przykład w kolumnie kategoria mogą występować różne zapisy tego samego słowa, np. „Spożywcze” i „spozywcze”. Skorzystaj z metod takich jak `str.lower()` lub `str.capitalize()` oraz funkcji `replace()` w pandas, aby znormalizować format tekstowy.

Zadanie 5. Konwersja typów danych

Sprawdź typy danych w każdej kolumnie za pomocą metody `dtypes`. Zidentyfikuj kolumny, w których typy danych nie są zgodne z oczekiwanymi – na przykład, jeśli kolumna cena została wczytana jako typ tekstowy (object), skonwertuj ją na typ liczbowy (float). Dodatkowo, zmień format daty w kolumnie `data_sprzedaży` na standardowy format `datetime`, korzystając z `pd.to_datetime()`.

Zadanie 6. Uzupelnianie brakujacych danych przy uzyciu interpolacji

Wybierz kolumnę liczbową z brakującymi wartościami i uzupełnij te braki, stosując interpolację (np. metodę liniową). Przeanalizuj, jak zmieniają się wyniki analizy po zastosowaniu interpolacji, porównując je z poprzednimi wynikami.

Zadanie 7. Obsługa danych niezgodnych z formatem lub zakresem

Zidentyfikuj wiersze, w których wartości w kolumnie cena przekraczają określony zakres – na przykład większe niż 500 lub mniejsze niż 5. Zastosuj funkcję clip(), aby ograniczyć wartości do zakresu 5–500. Następnie przeanalizuj, w jaki sposób taka korekta wpływa na wyniki obliczeń statystycznych.

Zadanie 8. Zapis oczyszczonych danych

Na koniec zapisz oczyszczone i przekształcone dane do nowego pliku CSV. Upewnij się, że nowe dane nie zawierają braków, duplikatów ani nieprawidłowych formatów. Zapisz plik z nowymi nazwami kolumn lub ze zaktualizowanymi formatami danych, co ułatwi dalszą analizę i pracę na tych danych.

3. Sprawozdanie

Sprawozdanie powinno zawierać:

- ☒ Treść zadania
- ☒ Kod napisanego programu
- ☒ Wynik działania napisanego programu
- ☒ Opis działania programu
- ☒ Wnioski końcowe

Sprawozdanie musi być przesłane w formacie notatnika jupyter. Sprawozdanie należy dostarczyć najpóźniej do północy dnia poprzedzającego dzień kolejnych laboratoriów. W przypadku spóźnienia przysługują 2 terminy poprawkowe wskazane przez prowadzącego.