|  |  |  |
| --- | --- | --- |
| ***Informatyka Stosowana*** | | |
| Laboratorium 3 | *Zaawansowane operacje na danych w Pandas* | Obraz zawierający tekst  Opis wygenerowany automatycznie |
| Przedmiot | Eksploracyjna analiza danych - laboratorium |
| Prowadzący | mgr inż. Gracjan Kątek |

## Wprowadzenie

Biblioteka Pandas to narzędzie kluczowe dla analizy danych, szczególnie dzięki możliwości wykonywania zaawansowanych operacji na dużych zbiorach danych. W tym laboratorium skupisz się na zaawansowanych technikach przekształcania i manipulacji danymi, takich jak łączenie danych, tworzenie tabel przestawnych oraz agregacja. Są to operacje niezbędne w pracy analityka, gdy mamy do czynienia z wieloma źródłami danych, które wymagają integracji i przeorganizowania w celu wyciągnięcia użytecznych informacji.

Podczas dzisiejszego laboratorium omówimy:

* merge(), concat(), join() – używane do łączenia danych po różnych kluczach i w różnych układach.
* pivot\_table() i melt() – do reorganizowania danych w różnych formatach tabelarycznych.
* apply() – do zaawansowanego przetwarzania i transformacji danych na poziomie wierszy lub kolumn.
* Tworzenie nowych kolumn na podstawie istniejących za pomocą operacji matematycznych, logicznych i tekstowych.

Dzięki tym technikom będziesz w stanie tworzyć czytelne raporty, złożone analizy oraz przygotowywać dane do modeli predykcyjnych.

## Zadania do wykonania

1. Łączenie zestawów danych

Zaimportuj dwa dostarczone zestawy danych – dane klientów oraz dane zamówień. Wykonaj następujące kroki:

* 1. Połącz dane za pomocą merge() na wspólnym kluczu, np. customer\_id.
  2. Wykorzystaj różne typy połączeń (inner, outer, left, right) i porównaj wyniki.
  3. W połączonym zestawie danych znajdź klientów, którzy nie złożyli żadnych zamówień.
  4. Użyj concat() do dodania nowych rekordów do tabeli zamówień i pokaż różnicę.

1. Tworzenie tabel przestawnych

Wykorzystaj zestaw zamówień i przeprowadź następujące operacje:

* 1. Stwórz tabelę przestawną (pivot\_table) pokazującą sumę wartości zamówień (order\_value) w podziale na klientów oraz kategorie produktów.
  2. Dla każdego klienta oblicz średnią wartość zamówienia oraz liczbę zamówień.
  3. Za pomocą melt() przekształć tabelę przestawną do formatu długiego (ang. long format).

1. Tworzenie nowych kolumn

Na zestawie zamówień wykonaj następujące operacje:

1. Dodaj nową kolumnę vat, obliczając 23% podatku VAT od wartości każdego zamówienia.
2. Stwórz kolumnę order\_date\_month, która zawiera tylko nazwę miesiąca z daty zamówienia.
3. Zdefiniuj nową kolumnę high\_value\_order, która przyjmuje wartość True, jeśli zamówienie przekracza wartość 1000, i False w przeciwnym razie.
4. Przekształcanie danych za pomocą apply()

Przeprowadź zaawansowane operacje na danych:

1. Użyj apply() na kolumnie product\_name, aby policzyć długość nazwy każdego produktu.
2. Stwórz funkcję, która oznacza zamówienia jako „małe”, „średnie” lub „duże” w zależności od wartości zamówienia (order\_value), i zaimplementuj ją przy użyciu apply().
3. Agregacja danych

Skorzystaj z zestawu danych połączonych w Zadaniu 1 i wykonaj:

1. Grupowanie po kolumnie customer\_id, aby obliczyć:
   * 1. łączną wartość zamówień każdego klienta,
     2. liczbę zamówień,
     3. maksymalną wartość zamówienia.
2. Dla każdej kategorii produktu oblicz sumaryczną wartość sprzedaży oraz liczbę zamówień.
3. Przekształć wyniki agregacji na formę tabeli przestawnej.
4. Zapisanie i załadowanie danych
5. Zapisz wyniki swojej analizy (np. tabelę przestawną) do pliku CSV.
6. Wczytaj zapisany plik i potwierdź, że dane zostały poprawnie zapisane i wczytane.
7. Zadanie otwarte

Wykorzystując dwa zestawy danych, które łączą informacje o klientach i ich zamówieniach:

1. Stwórz raport podsumowujący w formie tabeli, który zawiera:
2. Identyfikator klienta.
3. Całkowitą wartość zamówień.
4. Najczęściej kupowaną kategorię produktów.
5. Procentową wartość dużych zamówień (powyżej 1000 zł) w stosunku do wszystkich zamówień.
6. Zapisz raport do pliku CSV i przygotuj opis dla osoby analizującej dane.

## 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.