SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

|  |  |
| --- | --- |
| Laboratorium Nr 2  Data 19.10.2024 Temat: Wykorzystanie pakietu Pandas do manipulacji i przetwarzania danych w Pythonie Wariant: 5 | Mateusz Łysoń  Informatyka  II stopień, niestacjonarne,  1 semestr, gr. b |

**Polecenie**

Wykonaj zestaw zadań związanych z operacjami na danych z wykorzystaniem pakietu Pandas w Pythonie do manipulacji, analizy i przetwarzania danych.

• Wczytywanie danych z plików CSV

• Obliczanie podstawowych statystyk

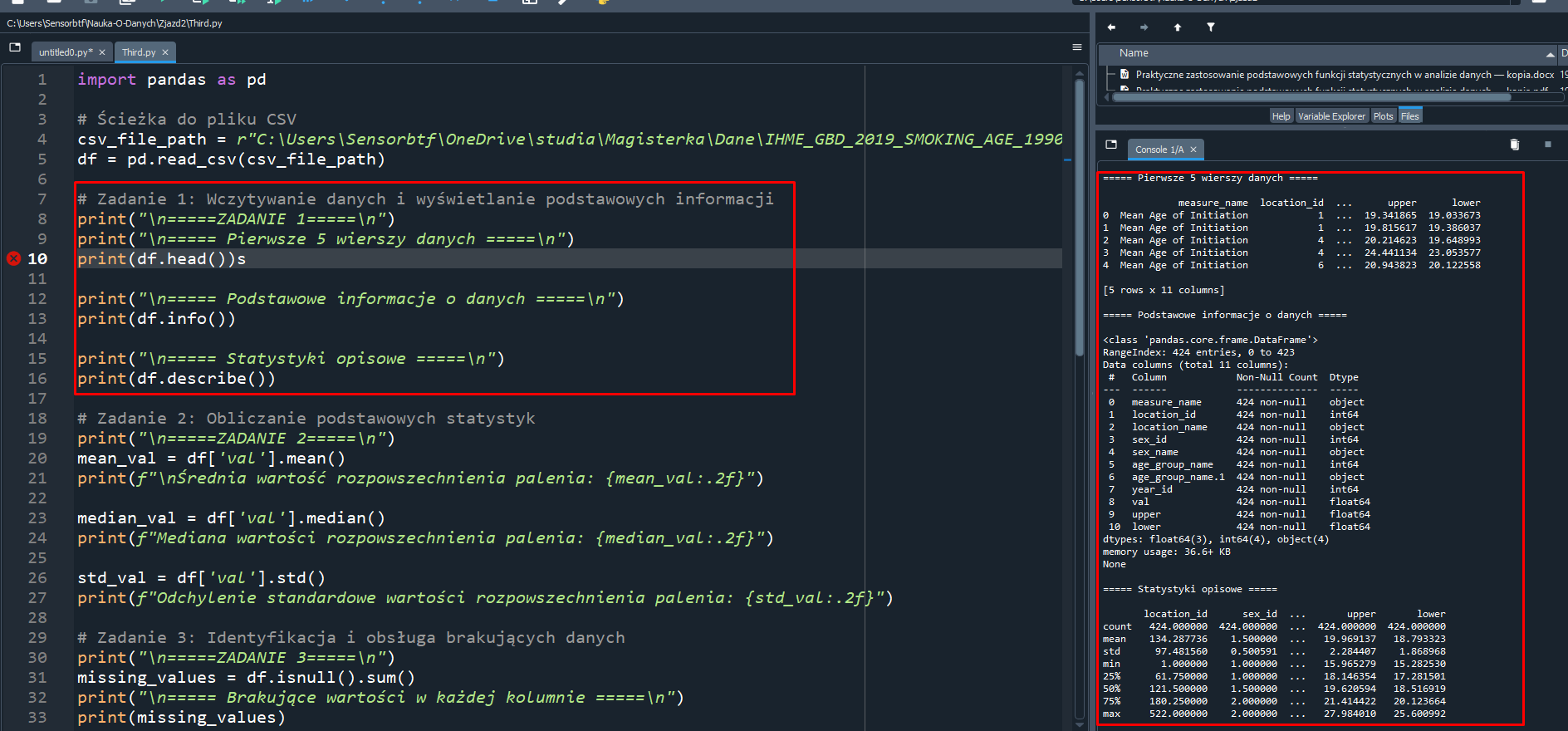
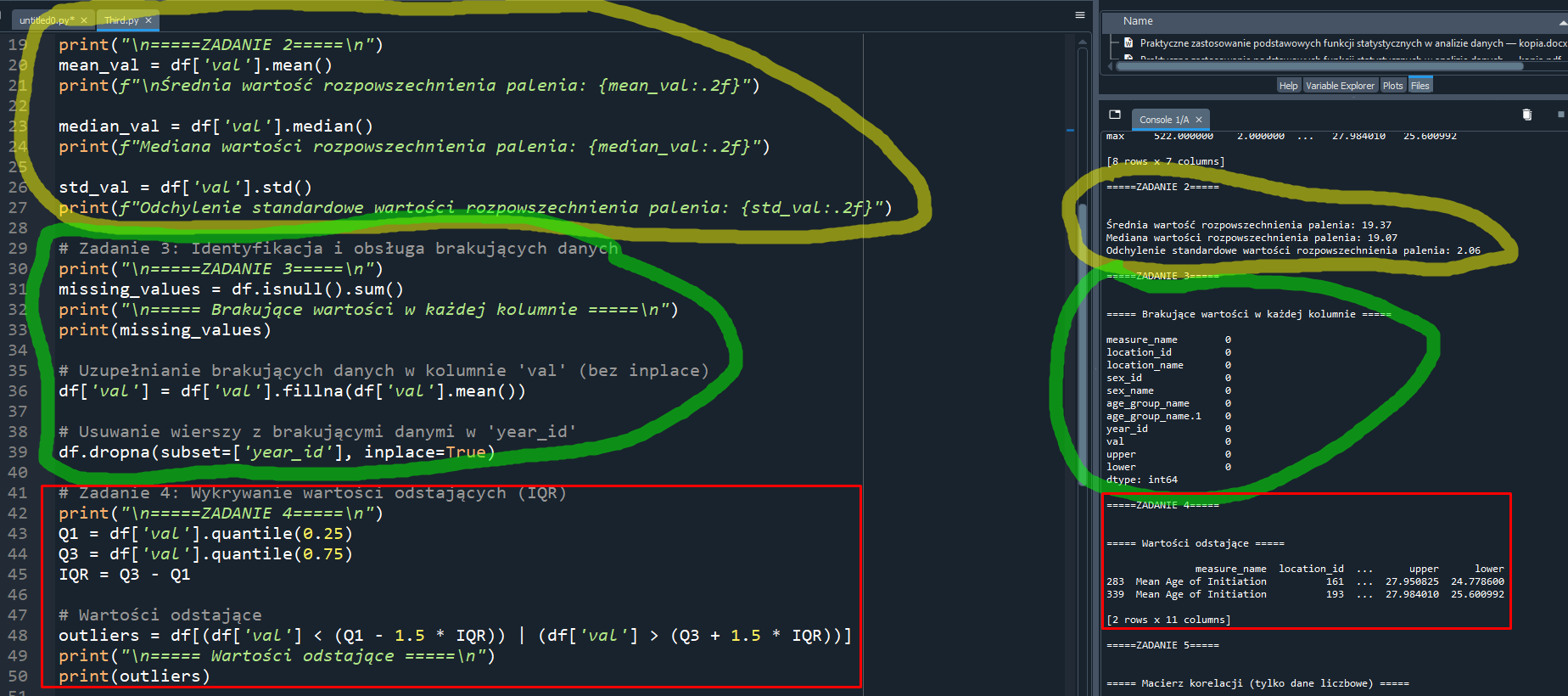
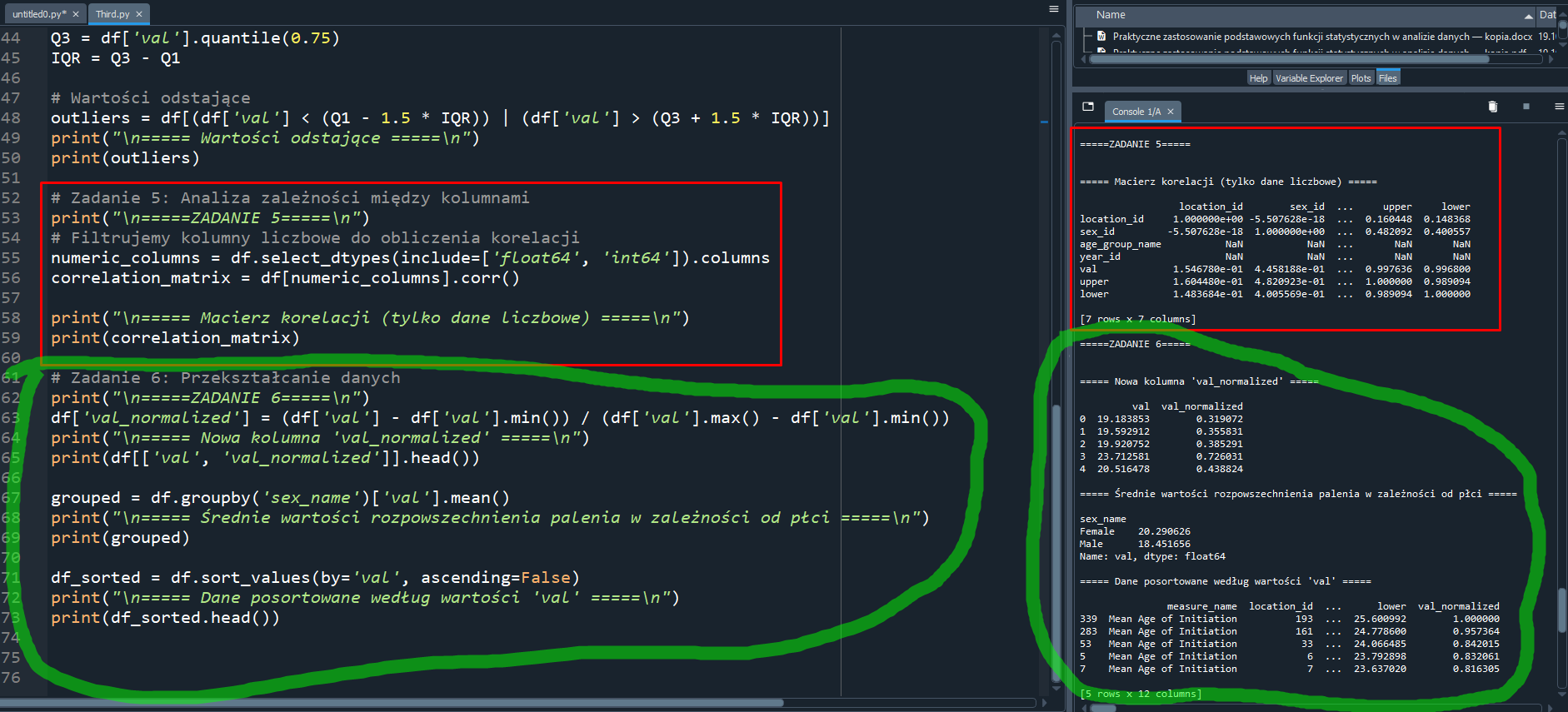
• Radzenie sobie z brakującymi danymi

• Wykrywanie wartości odstających

• Analiza zależności między zmiennymi

• Manipulacja i przekształcanie danych

**Opis programu opracowanego**

1. 
2. 
3. 

**Wnioski**

### Zadanie 1: Wczytywanie danych i wyświetlanie podstawowych informacji

* W zbiorze danych znajdują się **424 wiersze i 11 kolumny**, zawierające informacje o nazwie miary, lokalizacji, płci, grupie wiekowej, roku, wartości inicjacji palenia oraz granicach przedziału ufności
* Brak brakujących wartości we wszystkich kolumnach. To oznacza, że zbiór danych jest kompletny

### Zadanie 2: Obliczanie podstawowych statystyk

* **Średnia wartość inicjacji palenia** wynosi **19,37**, co sugeruje, że średni wiek rozpoczęcia palenia wynosi około 19 lat.
* **Mediana wartości** wynosi **19,07**, co oznacza, że połowa osób zaczyna palić przed ukończeniem 19 lat, a druga połowa po 19. roku życia.
* **Odchylenie standardowe** wynosi **2,06**, co wskazuje na dość dużą zmienność w wieku rozpoczęcia palenia w zależności od lokalizacji i innych zmiennych.

### Zadanie 3: Identyfikacja i obsługa brakujących danych

* W zbiorze danych nie występują brakujące wartości. Dzięki temu dane mogą być bezpośrednio analizowane, bez potrzeby ich uzupełniania lub usuwania wierszy.

### Zadanie 4: Wykrywanie wartości odstających (IQR)

* Zidentyfikowano dwie wartości odstające, dotyczące lokalizacji o identyfikatorach **161** i **193**. Wartości te charakteryzują się najwyższymi wartościami średniego wieku rozpoczęcia palenia, które wynoszą odpowiednio **27,95** oraz **27,98** lat

### Zadanie 5: Analiza zależności między kolumnami

* Macierz korelacji pokazuje pewne zależności:
  + Istnieje umiarkowana pozytywna korelacja (**0,4458**) między płcią (sex\_id) a wiekiem inicjacji palenia (val), co może sugerować, że płeć ma pewien wpływ na wiek, w którym osoby zaczynają palić.

### Zadanie 6: Przekształcanie danych

* Znormalizowano wartości inicjacji palenia (val\_normalized) w celu ułatwienia dalszej analizy.
* Średnia wartość inicjacji palenia w zależności od płci wskazuje, że kobiety mają średnio wyższy wiek rozpoczęcia palenia (**20,29 lat**) w porównaniu do mężczyzn (**18,45 lat**). Sugeruje to, że kobiety zaczynają palić później niż mężczyźni.

GIthub: https://github.com/sensorbtf/Nauka-O-Danych/tree/main