SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

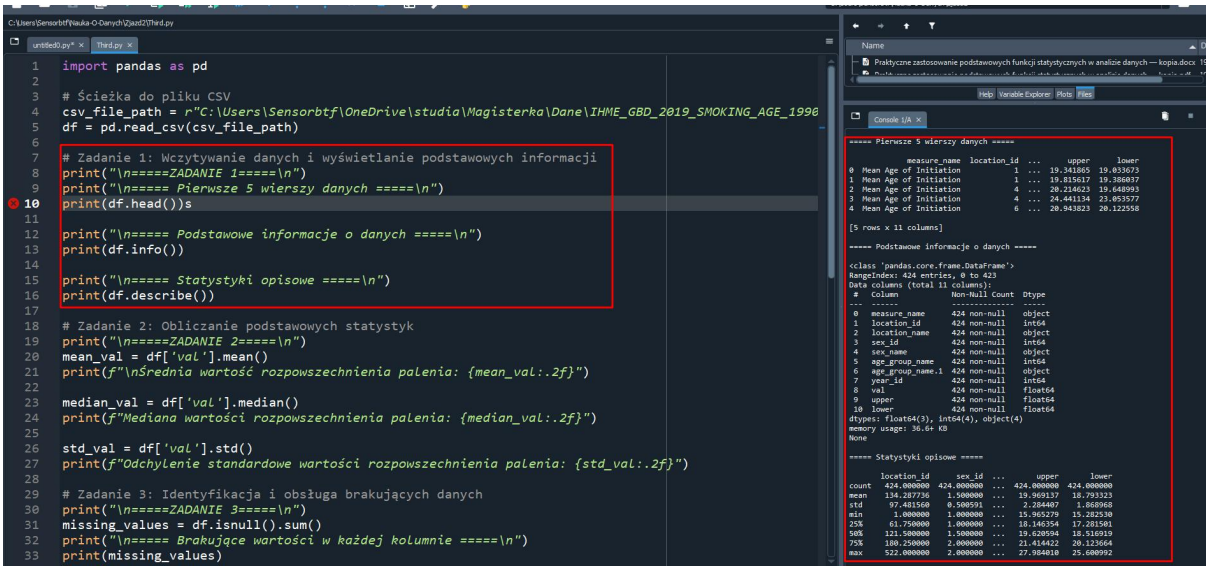
Laboratorium Nr 2 Data 19.10.2024 Temat: Wykorzystanie pakietu Pandas do manipulacji i przetwarzania danych w Pythonie Wariant: 5	Mateusz Łysoń Informatyka II stopień, niestacjonarne, 1 semestr, gr. b

Polecenie

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

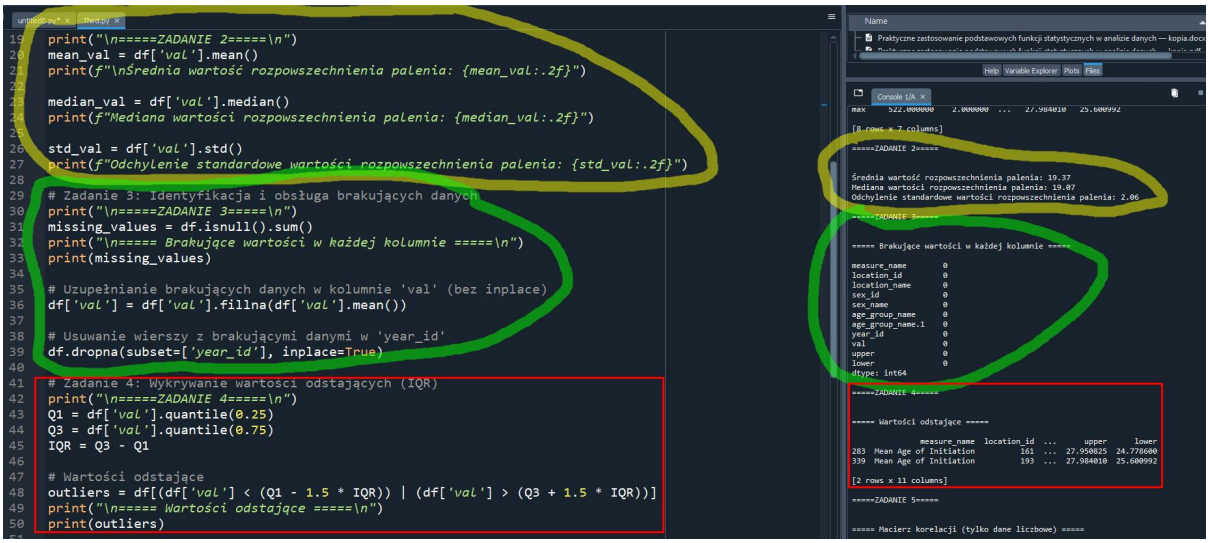
- Wczytywanie danych z plików CSV
- Obliczanie podstawowych statystyk
- Radzenie sobie z brakującymi danymi
- Wykrywanie wartości odstających
- Analiza zależności między zmiennymi
- Manipulacja i przekształcanie danych

Opis programu opracowanego

1. 

```
import pandas as pd

# Ścieżka do pliku CSV
csv_file_path = r"C:\Users\Sensorbtf\OneDrive\studia\Magisterka\Dane\IHME_GBD_2019_SMOKING_AGE_1996
df = pd.read_csv(csv_file_path)

# Zadanie 1: Wczytywanie danych i wyświetlanie podstawowych informacji
print("\n======ZADANIE 1=====\n")
print("\n====== Pierwsze 5 wierszy danych =====\n")
print(df.head())s

print("\n====== Podstawowe informacje o danych =====\n")
print(df.info())

print("\n====== Statystyki opisowe =====\n")
print(df.describe())

# Zadanie 2: Obliczanie podstawowych statystyk
print("\n======ZADANIE 2=====\n")
mean_val = df['val'].mean()
print(f"\nŚrednia wartość rozpowszechnienia palenia: {mean_val:.2f}")

median_val = df['val'].median()
print(f"Mediana wartości rozpowszechnienia palenia: {median_val:.2f}")

std_val = df['val'].std()
print(f"Odchylenie standardowe wartości rozpowszechnienia palenia: {std_val:.2f}")

# Zadanie 3: Identyfikacja i obsługa brakujących danych
print("\n======ZADANIE 3=====\n")
missing_values = df.isnull().sum()
print("\n====== Brakujące wartości w każdej kolumnie =====\n")
print(missing_values)
```

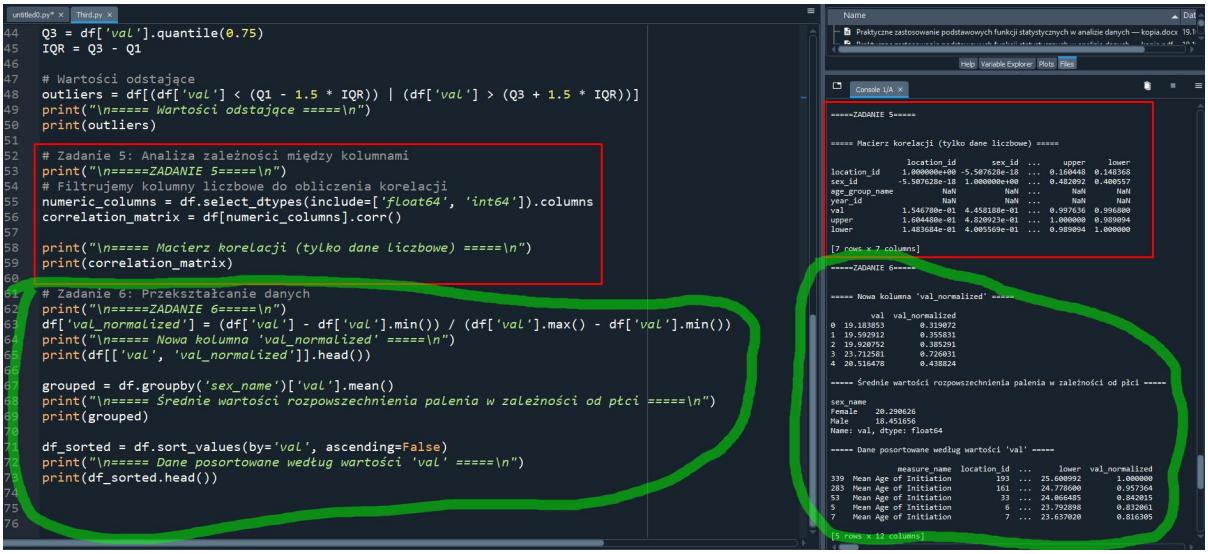
```
===== Pierwsze 5 wierszy danych =====

          measure_name  location_id  ...      upper      lower
0  Mean Age of Initiation            1  ...  19.341865  19.033673
1  Mean Age of Initiation            1  ...  19.815617  19.386037
2  Mean Age of Initiation            4  ...  20.214623  19.640893
3  Mean Age of Initiation            4  ...  24.441134  23.053577
4  Mean Age of Initiation            6  ...  20.943823  20.122558

[5 rows x 11 columns]

===== Podstawowe informacje o danych =====

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 424 entries, 0 to 423
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   measure_name     424 non-null    object
 1   location_id      424 non-null    int64
 2   location_name    424 non-null    object
 3   sex_id           424 non-null    int64
 4   sex_name         424 non-null    object
 5   age_group_name   424 non-null    int64
 6   age_group_name.1 424 non-null    object
 7   year_id          424 non-null    int64
 8   val              424 non-null    float64
 9   upper            424 non-null    float64
 10  lower            424 non-null    float64
dtypes: float64(3), int64(4), object(4)
memory usage: 36.6+ KB
None

===== Statystyki opisowe =====

       location_id      sex_id  ...      upper      lower
count   424.000000  424.000000  ...  424.000000  424.000000
mean    134.287736    1.500000  ...   19.969137   18.793323
std      97.481560    0.500591  ...    2.284407    1.868968
min       1.000000    1.000000  ...   15.965279   15.282530
25%      61.750000    1.000000  ...   18.146354   17.281501
50%     121.500000    1.500000  ...   19.620594   18.516919
75%     180.250000    2.000000  ...   21.414422   20.123664
max     522.000000    2.000000  ...   27.984010   25.600992
```

2. 

```
print("\n======ZADANIE 2=====\n")
mean_val = df['val'].mean()
print(f"\nŚrednia wartość rozpowszechnienia palenia: {mean_val:.2f}")

median_val = df['val'].median()
print(f"Mediana wartości rozpowszechnienia palenia: {median_val:.2f}")

std_val = df['val'].std()
print(f"Odchylenie standardowe wartości rozpowszechnienia palenia: {std_val:.2f}")

# Zadanie 3: Identyfikacja i obsługa brakujących danych
print("\n======ZADANIE 3=====\n")
missing_values = df.isnull().sum()
print("\n====== Brakujące wartości w każdej kolumnie =====\n")
print(missing_values)

# Uzupełnianie brakujących danych w kolumnie 'val' (bez inplace)
df['val'] = df['val'].fillna(df['val'].mean())

# Usuwanie wierszy z brakującymi danymi w 'year_id'
df.dropna(subset=['year_id'], inplace=True)

# Zadanie 4: Wykrywanie wartosci odstających (IQR)
print("\n======ZADANIE 4=====\n")
Q1 = df['val'].quantile(0.25)
Q3 = df['val'].quantile(0.75)
IQR = Q3 - Q1

# Wartości odstające
outliers = df[(df['val'] < (Q1 - 1.5 * IQR)) | (df['val'] > (Q3 + 1.5 * IQR))]
print("\n====== Wartości odstające =====\n")
print(outliers)
```

```
[8 rows x 7 columns]
======ZADANIE 2=====


Średnia wartość rozpowszechnienia palenia: 19.37
Mediana wartości rozpowszechnienia palenia: 19.07
Odchylenie standardowe wartości rozpowszechnienia palenia: 2.06

======ZADANIE 3=====


====== Brakujące wartości w każdej kolumnie =====

measure_name        0
location_id         0
location_name       0
sex_id              0
sex_name            0
age_group_name      0
age_group_name.1    0
year_id             0
val                 0
upper               0
lower               0
dtype: int64

======ZADANIE 4=====


====== Wartości odstające =====

          measure_name  location_id  ...      upper      lower
283  Mean Age of Initiation          161  ...  27.958025  24.778600
339  Mean Age of Initiation          193  ...  27.984010  25.600992

[2 rows x 11 columns]

======ZADANIE 5=====


====== Macierz korelacji (tylko dane liczbowe) =====
```

3.



```
Q3 = df['val'].quantile(0.75)
IQR = Q3 - Q1

# Wartości odstające
outliers = df[(df['val'] < (Q1 - 1.5 * IQR)) | (df['val'] > (Q3 + 1.5 * IQR))]
print("\n===== Wartości odstające =====\n")
print(outliers)

# Zadanie 5: Analiza zależności między kolumnami
print("\n=====ZADANIE 5=====\n")
# Filtrujemy kolumny liczbowe do obliczenia korelacji
numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns
correlation_matrix = df[numeric_columns].corr()

print("\n===== Macierz korelacji (tylko dane liczbowe) =====\n")
print(correlation_matrix)

# Zadanie 6: Przekształcanie danych
print("\n=====ZADANIE 6=====\n")
df['val_normalized'] = (df['val'] - df['val'].min()) / (df['val'].max() - df['val'].min())
print("\n===== Nowa kolumna 'val_normalized' =====\n")
print(df[['val', 'val_normalized']].head())

grouped = df.groupby('sex_name')['val'].mean()
print("\n===== Średnie wartości rozpowszechnienia palenia w zależności od płci =====\n")
print(grouped)

df_sorted = df.sort_values(by='val', ascending=False)
print("\n===== Dane posortowane według wartości 'val' =====\n")
print(df_sorted.head())
```

```
=====ZADANIE 5=====


===== Macierz korelacji (tylko dane liczbowe) =====

                location_id        sex_id  ...     upper     lower
location_id    1.000000e+00 -5.507628e-18  ...  0.169448  0.148368
sex_id        -5.507628e-18  1.000000e+00  ...  0.482092  0.400557
age_group_name          NaN           NaN  ...       NaN       NaN
year_id                 NaN           NaN  ...       NaN       NaN
val            1.546780e-01  4.458188e-01  ...  0.997636  0.996800
upper          1.604480e-01  4.820923e-01  ...  1.000000  0.989094
lower          1.483684e-01  4.005569e-01  ...  0.989094  1.000000

[7 rows x 7 columns]

=====ZADANIE 6=====


===== Nowa kolumna 'val_normalized' =====

        val  val_normalized
0 19.183853        0.319072
1 19.592912        0.355831
2 19.928752        0.385291
3 23.712581        0.726031
4 20.516478        0.438824

===== Średnie wartości rozpowszechnienia palenia w zależności od płci =====

sex_name
Female    20.290626
Male      18.451656
Name: val, dtype: float64

===== Dane posortowane według wartości 'val' =====

             measure_name  location_id  ...      lower  val_normalized
339  Mean Age of Initiation          193  ...  25.600992        1.000000
283  Mean Age of Initiation          161  ...  24.778600        0.957364
53   Mean Age of Initiation           33  ...  24.065485        0.843015
5    Mean Age of Initiation            6  ...  23.792898        0.832061
7    Mean Age of Initiation            7  ...  23.637020        0.816305

[5 rows x 12 columns]
```

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