Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій

Лабораторна робота №3

з дисципліни

Аналіз даних з використанням мови Python

Виконав:

студент групи ІА-24:
Криворучек В.С.

Перевірила:

ст. викладач
Тимофєєва Ю.С.

Київ 2025

Тема: Структури даних Pandas

Мета роботи: Ознайомитись з основними структурами даних бібліотеки Pandas: Series DataFrame, операціями над ними. Навчитись використовувати групування.

Хід роботи

Завдання:

Файл insurance.csv

1. Вивести інформацію про набір даних, типи ознак. Які ознаки є категоріальними, а які – кількісними?

2. Використовуючи заданий набір даних:

а) зберегти назви стовпців у окрему змінну і вивести її;

б) вивести кількість курців і не курців;

в) вивести дані випадкового чоловіка-курця з витратами більше 30000;

г) додати новий рядок до DataFrame з довільними даними;

3. Робота із групованими даними. Для виконання кожного з наступних підзавдань достатньо одного рядка коду:

а) знайти медіанний вік за регіоном;

б) додати новий стовпець, який містить середній індекс маси тіла за регіоном;

в) вивести дані клієнтів лише того віку, для якого середня кількість дітей менше 0,5.

4. За допомогою pivot_table створити нову таблицю, що буде містити середні витрати та індекс маси тіла для людей різної статі та з різних регіонів. Зберегти у окрему змінну середній індекс маси тіла для чоловіків з північно-західного регіону.

Код програми:

```
import pandas as pd

# Зчитування даних з файлу
df = pd.read_csv("insurance.csv")

# 1. Інформація про набір даних та типи ознак
print("=" * 50)
print("ІНФОРМАЦІЯ ПРО НАБІР ДАНИХ")
```

```
print("=" * 50)
print(df.info())

print("\nТипи ознак:")
print(df.dtypes)

# Категоріальні ознаки: sex, smoker, region
# Кількісні ознаки: age, bmi, children, expenses

# 2. Операції над заданим набором даних
print("\n" + "=" * 50)
print("СПИСОК НАЗВ СТОВПЦІВ")
print("=" * 50)
columns_list = df.columns.tolist()
print(", ".join(columns_list))

print("\n" + "=" * 50)
print("КІЛЬКІСТЬ КУРЦІВ І НЕ КУРЦІВ")
print("=" * 50)
smoker_counts = df['smoker'].value_counts()
for category, count in smoker_counts.items():
    print(f"{category.capitalize()}: {count}")

print("\n" + "=" * 50)
print("ВИПАДКОВИЙ ЧОЛОВІК-КУРЕЦЬ З ВИТРАТАМИ > 30 000")
print("=" * 50)
random_male_smoker = df[(df['sex'] == 'male') & (df['smoker'] == 'yes')
& (df['expenses'] > 30000)].sample(n=1, random_state=42)
print(random_male_smoker.to_string(index=False))

print("\n" + "=" * 50)
print("ДОДАНО НОВИЙ РЯДОК")
print("=" * 50)
new_row = {
    'age': 40,
    'sex': 'female',
    'bmi': 25.5,
    'children': 2,
    'smoker': 'no',
    'region': 'southwest',
    'expenses': 20000
}
df.loc[len(df)] = new_row
print(df.tail(1).to_string(index=False))

# 3. Робота із груповаими даними
print("\n" + "=" * 50)
print("МЕДІАННИЙ ВІК ЗА РЕГІОНОМ")
print("=" * 50)
median_age_by_region = df.groupby('region')['age'].median()
print(median_age_by_region.to_string())

print("\n" + "=" * 50)
print("СЕРЕДНІЙ BMI ЗА РЕГІОНОМ")
print("=" * 50)
df['avg_bmi'] = df.groupby('region')['bmi'].transform('mean')
print(df[['region', 'bmi',
'avg_bmi']].drop_duplicates().head(10).to_string(index=False))

print("\n" + "=" * 50)
```

```
print("КЛІЄНТИ, ДЕ СЕРЕДНЯ КІЛЬКІСТЬ ДІТЕЙ < 0.5")
print("=" * 50)
ages_with_few_children = df.groupby('age')['children'].mean()
ages_selected = ages_with_few_children[ages_with_few_children <
0.5].index
clients_selected = df[df['age'].isin(ages_selected)]
print("\nДані клієнтів віку, де середня кількість дітей < 0.5:\n",
clients_selected.head(10).to_string(index=False))

# 4. Pivot-таблиця
print("\n" + "=" * 50)
print("PIVOT-ТАБЛИЦЯ (СЕРЕДНІ ВИТРАТИ ТА BMI)")
print("=" * 50)
pivot_df = df.pivot_table(index=['sex', 'region'], values=['expenses',
'bmi'], aggfunc='mean')
print(pivot_df.round(2).to_string())

# 4.1 Отримати середній BMI для чоловіків з північно-західного регіону
male_nw_avg_bmi = pivot_df.loc[('male', 'northwest'), 'bmi']
print("\n" + "=" * 50)
print(f"Середній BMI для чоловіків з північно-західного регіону:
{male_nw_avg_bmi:.2f}")
print("=" * 50)
```

Результат виконання:

```
==================================================
ІНФОРМАЦІЯ ПРО НАБІР ДАНИХ
==================================================
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
None
```

```
Типи ознак:
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
expenses    float64
dtype: object

==================================================
СПИСОК НАЗВ СТОВПЦІВ
==================================================
age, sex, bmi, children, smoker, region, expenses

==================================================
КІЛЬКІСТЬ КУРЦІВ І НЕ КУРЦІВ
==================================================
No: 1064
Yes: 274
```

```
==================================================
ВИПАДКОВИЙ ЧОЛОВІК-КУРЕЦЬ З ВИТРАТАМИ > 30 000
==================================================
 age  sex  bmi  children smoker    region  expenses
  60 male 40.9         0    yes southeast  48673.56

==================================================
ДОДАНО НОВИЙ РЯДОК
==================================================
 age    sex  bmi  children smoker    region  expenses
  40 female 25.5         2     no southwest   20000.0

==================================================
МЕДІАННИЙ ВІК ЗА РЕГІОНОМ
==================================================
region
northeast    39.5
northwest    39.0
southeast    39.0
southwest    39.5
```

```
================================================
СЕРЕДНІЙ BMI ЗА РЕГІОНОМ
================================================
   region  bmi   avg_bmi
southwest 27.9 30.580982
southeast 33.8 33.359341
southeast 33.0 33.359341
northwest 22.7 29.201846
northwest 28.9 29.201846
southeast 25.7 33.359341
southeast 33.4 33.359341
northwest 27.7 29.201846
northeast 29.8 29.176235
northwest 25.8 29.201846

================================================
КЛІЄНТИ, ДЕ СЕРЕДНЯ КІЛЬКІСТЬ ДІТЕЙ < 0.5
================================================

Дані клієнтів віку, де середня кількість дітей < 0.5:
  age    sex  bmi  children smoker    region  expenses   avg_bmi
  19 female 27.9         0    yes southwest  16884.92 30.580982
  18   male 33.8         1     no southeast   1725.55 33.359341
  60 female 25.8         0     no northwest  28923.14 29.201846
  19   male 24.6         1     no southwest   1837.24 30.580982
  60 female 36.0         0     no northeast  13228.85 29.176235
  18   male 34.1         0     no southeast   1137.01 33.359341
  18 female 26.3         0     no northeast   2198.19 29.176235
  19 female 28.6         5     no southwest   4687.80 30.580982
  19   male 20.4         0     no northwest   1625.43 29.201846
  60   male 39.9         0    yes southwest  48173.36 30.580982
```

```
==================================================
PIVOT-ТАБЛИЦЯ (СЕРЕДНІ ВИТРАТИ ТА BMI)
==================================================
                     bmi  expenses
sex    region
female northeast  29.33  12953.20
       northwest  29.28  12479.87
       southeast  32.67  13499.67
       southwest  30.03  11327.94
male   northeast  29.02  13854.01
       northwest  29.12  12354.12
       southeast  33.99  15879.62
       southwest  31.13  13412.88

==================================================
Середній BMI для чоловіків з північно-західного регіону: 29.12
==================================================

Process finished with exit code 0
```

Висновок: У ході виконання даної лабораторної роботи я ознайомився з основними структурами даних бібліотеки Pandas: Series DataFrame та операціями над ними. Також я навчився використовувати групування.