Зображення, що містить текст

Автоматично згенерований опис

Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

**Лабораторна робота №3**

Аналіз даних з використанням мови Python

**Тема:** Структури даних Pandas

**Варіант:** 1

Виконав Перевірив:

студент групи ІП-11: Тимофєєва Ю. С

Панченко С. В.

Київ 2023

Зміст

[1 Мета лабораторної роботи 6](#__RefHeading___Toc242_4224234036)

[2 Завдання 7](#__RefHeading___Toc244_4224234036)

[3 Виконання 8](#__RefHeading___Toc246_4224234036)

[3.1 Виділити один зі стовпців (на вибір) з файлу як об’єкт Series, виділити з нього підмасив. Задати назви індексів цього об’єкту. Виділити підмасиви за допомогою прямої та непрямої індексацій 8](#__RefHeading___Toc248_4224234036)

[3.2 До об’єкту DataFrame, в який записано вміст файлу, додати новий стовпець, що є результатом операцій над іншими стовпцями. Також продемонструвати додавання та видалення рядків, видалення стовпців 10](#__RefHeading___Toc250_4224234036)

[3.3 Встановити один зі стовпців індексом. Визначити основні статистичні характеристики та типи даних всіх стовпців. Змінити тип даних для одного з стовпців. Згрупувати дані за одним зі стовпців, застосувати кілька агрегуючих функцій, виділити підмасив за певними ознаками. 13](#__RefHeading___Toc252_4224234036)

[3.4 Створити декілька власних об’єктів DataFrame за такою ж тематикою, що й файл. Наприклад, якщо тема файлу – жаби, можна створити об’єкти, що містять розміри жаб, вагу, стать, кількість особин в популяції і т.д. Використати описані в теоретичних відомостях параметри методів merge та concat для різних видів злиття та об’єднання даних цих об’єктів 15](#__RefHeading___Toc254_4224234036)

[4 Висновок 17](#__RefHeading___Toc256_4224234036)

# Мета лабораторної роботи

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

# Завдання

Створити програму, яка за даними файлу відповідно до варіантів

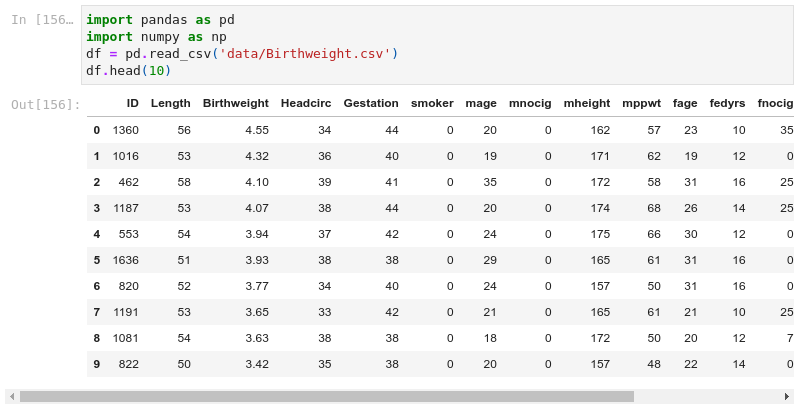
лабораторної №2, виконує наступні завдання:

1. Виділити один зі стовпців (на вибір) з файлу як об’єкт Series, виділити з нього підмасив. Задати назви індексів цього об’єкту. Виділити підмасиви за допомогою прямої та непрямої індексацій.
2. До об’єкту DataFrame, в який записано вміст файлу, додати новий стовпець, що є результатом операцій над іншими стовпцями. Також продемонструвати додавання та видалення рядків, видалення стовпців.
3. Встановити один зі стовпців індексом. Визначити основні статистичні характеристики та типи даних всіх стовпців. Змінити тип даних для одного з стовпців. Згрупувати дані за одним зі стовпців, застосувати кілька агрегуючих функцій, виділити підмасив за певними ознаками.
4. Створити декілька власних об’єктів DataFrame за такою ж тематикою, що й файл. Наприклад, якщо тема файлу – жаби, можна створити об’єкти, що містять розміри жаб, вагу, стать, кількість особин в популяції і т.д. Використати описані в теоретичних відомостях параметри методів merge та concat для різних видів злиття та об’єднання даних цих об’єктів.

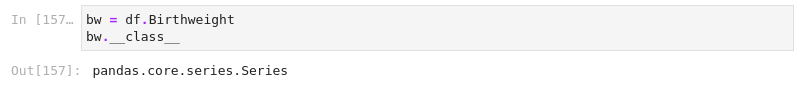
# Виконання

## Виділити один зі стовпців (на вибір) з файлу як об’єкт Series, виділити з нього підмасив. Задати назви індексів цього об’єкту. Виділити підмасиви за допомогою прямої та непрямої індексацій

Імпортуємо модуль pandas та завантажимо датасет Birthweight.csv. Виведемо перші десять рядків за допомогою методу head.

  
  
Рисунок 3.1 - Завантаження стовпця та виведення перших десяти рядків

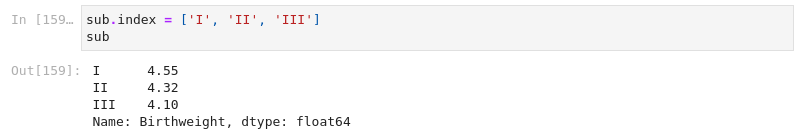
Виділимо стовпець як об'єкт типу Series.

  
  
Рисунок 3.2 - Виділення стовпця як об'єкта типу Series

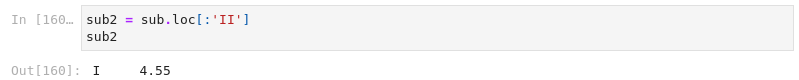
Виділимо підмасив з даного стовпця.

  
  
Рисунок 3.3 - Виділення підмасиву

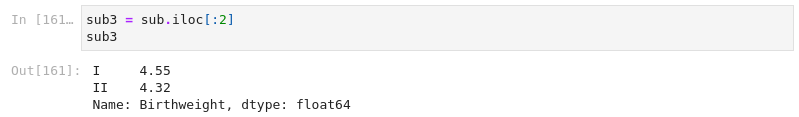
Задамо назви індексів даному підмасиву.

  
  
Рисунок 3.4 - Задання назви індексів

Виділимо підмасиви за допомогою прямої індексації. Використаємо атрибут loc та візьмемо елементи, індекси яких знаходяться до елемента з назвою індексу римського позначення двійки.

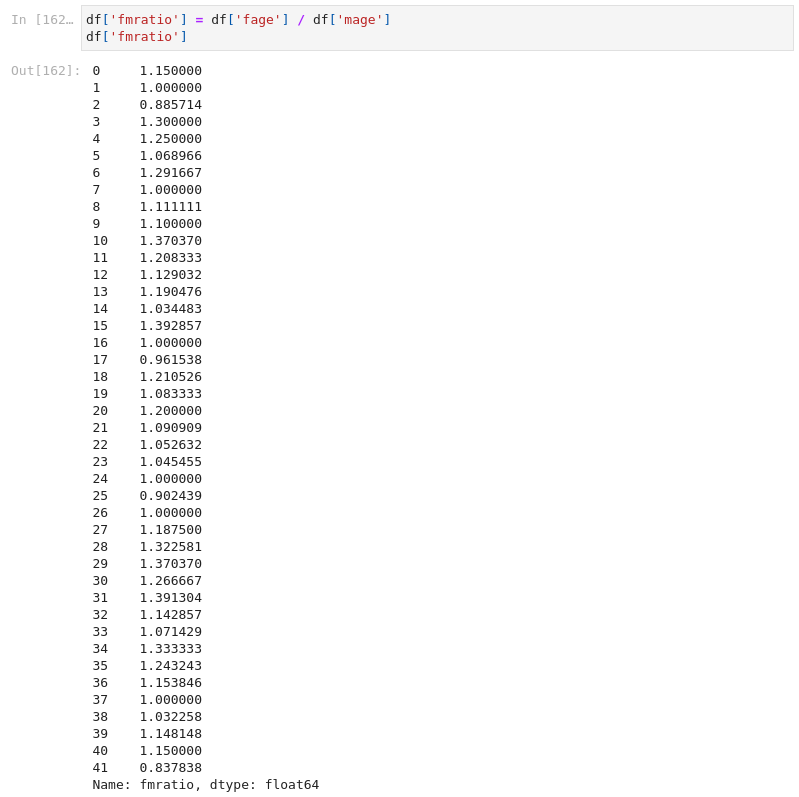
  
  
Рисунок 3.5 - Виділення підмасиву за допомогою прямої індексації

Виділимо підмасиви за допомогою непрямої індексації. Використаємо атрибут iloc та візьмемо елементи, індекси яких знаходяться до елементу індекс якого має порядковий номер 2.

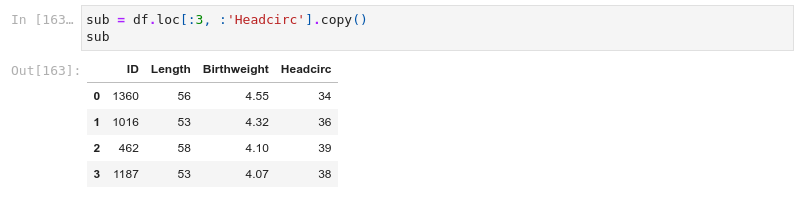
  
  
Рисунок 3.6 - Виділення підмасиву за допомогою непрямої індексації

## До об’єкту DataFrame, в який записано вміст файлу, додати новий стовпець, що є результатом операцій над іншими стовпцями. Також продемонструвати додавання та видалення рядків, видалення стовпців

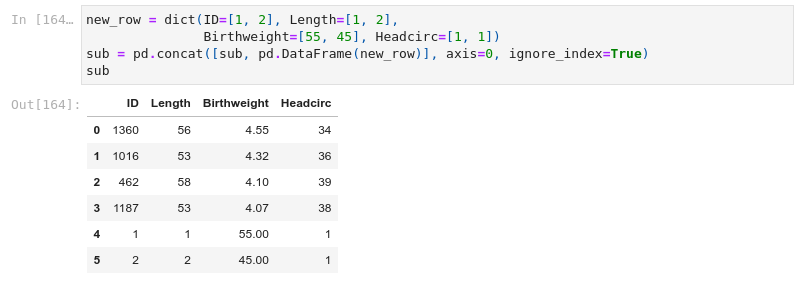
Додамо новий стовпець, що покаже відношення віку батька до віку матері.

  
  
Рисунок 3.7 - Стовпець, що позначає відношення віку батька до віку матері

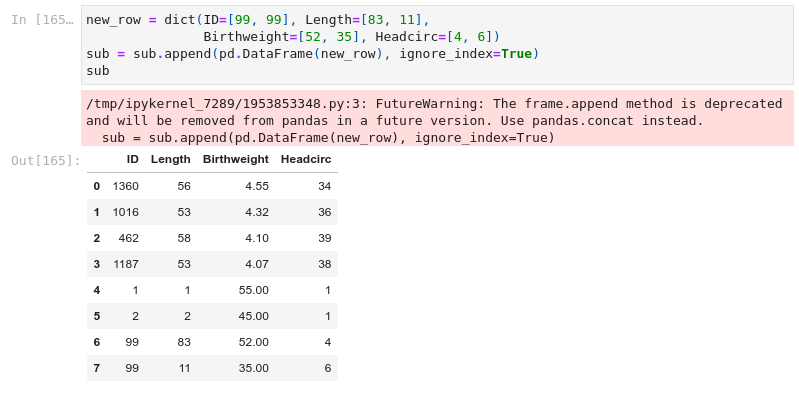
Для зручності виділимо частину датафрейму та будемо працювати з ним.

  
  
Рисунок 3.8 - Виділення чатини датафрейму

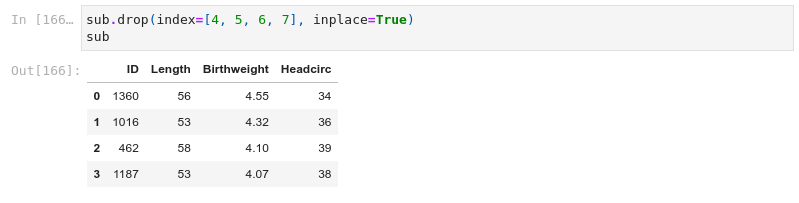
Додамо рядки до датафрейму за допомогою метода concat. За допомогою параметру за замовчуванням з'єднаємо два датафрейми, переіндексувавши новоутворений.

  
  
Рисунок 3.9 - Додавання рядків функцією concat

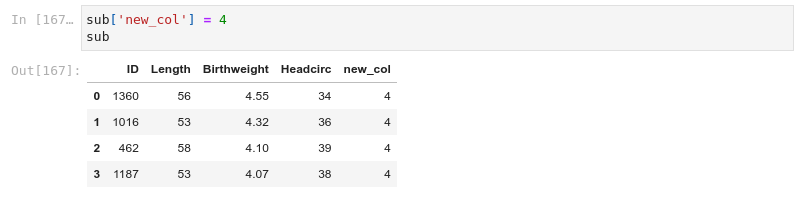
Додамо рядки до датафрейму за допомогою методу append.

  
  
Рисунок 3.10 - Додавання рядків методом append

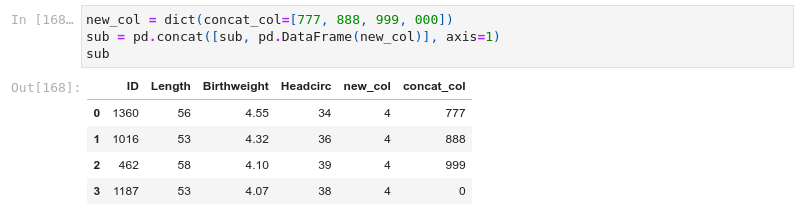
Видалимо чотири рядки з датафрейму за допомогою методу drop.

  
  
Рисунок 3.11 - Видалення рядків

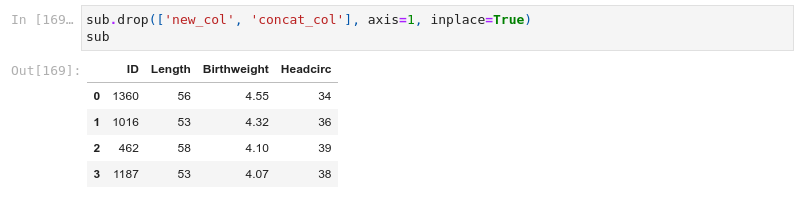
Додамо новий стовпець до датафрейму та проініціалізуємо його.

  
  
Рисунок 3.12 - Додавання стовпця за допомогою передачі індексу

Додамо новий стовпець за допомогою методу concat.

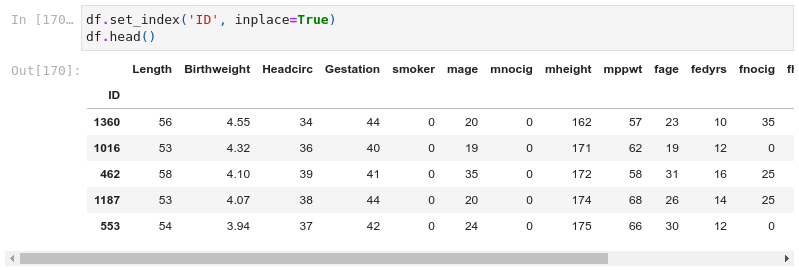
  
  
Рисунок 3.13 - Додавання стовпця за допомогою методу concat

Видалимо новостворені стовпці за допомогою методу drop.

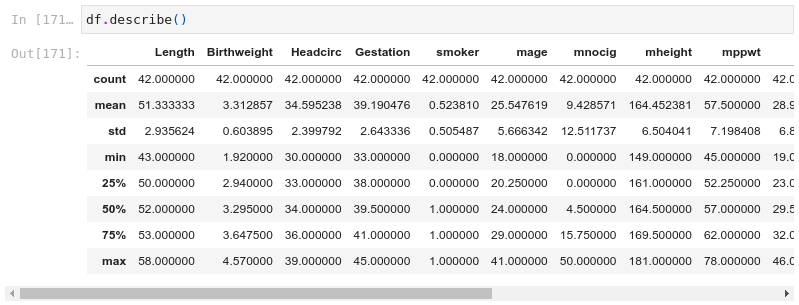
  
  
Рисунок 3.14 - Видалення стовпців методом drop

## Встановити один зі стовпців індексом. Визначити основні статистичні характеристики та типи даних всіх стовпців. Змінити тип даних для одного з стовпців. Згрупувати дані за одним зі стовпців, застосувати кілька агрегуючих функцій, виділити підмасив за певними ознаками.

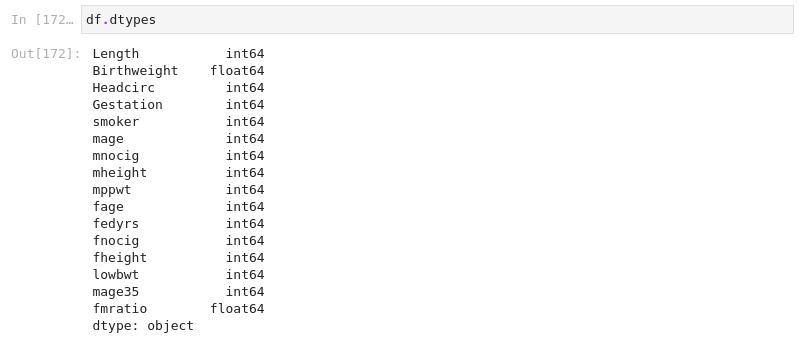
Встановимо стовпець ID індексом датафрейму за допомогою методу set\_index.

  
  
Рисунок 3.15 - Встановлення індексу методом set\_index

Виведемо основні статистичні характеристики методом describe.

  
  
Рисунок 3.16 - Визначення основних статистичних характеристик

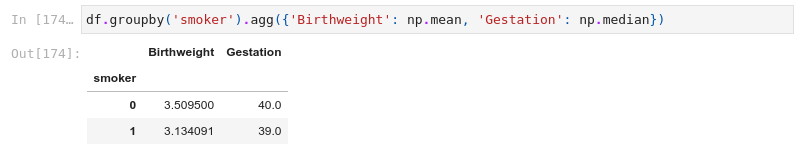
Виведемо типи даних всіх стовпців за допомогою атрибуту dtypes.

  
  
Рисунок 3.17 - Виведення даних стовпців за допомогою атрибуту dtypes

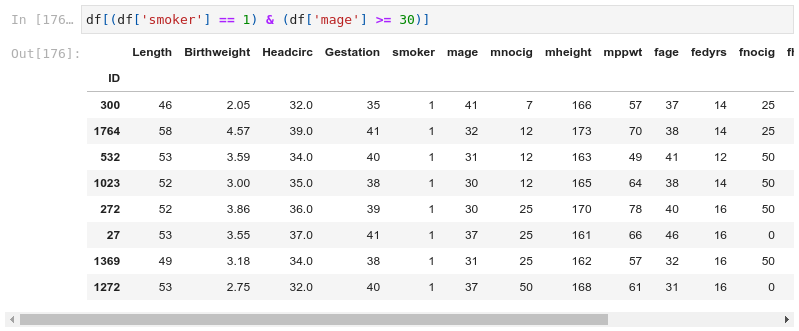
Змінимо тип даних для стовпця Headcirc з int64 до float64 за допомогою методу astype.

  
  
Рисунок 3.18 - Зміна типу даних стовпця методом astype

Згрупуємо дані за стовпцем 'smoker' та застосуємо на стовпці 'Birthweight' функцію np.mean, а на 'Gestation' np.median.

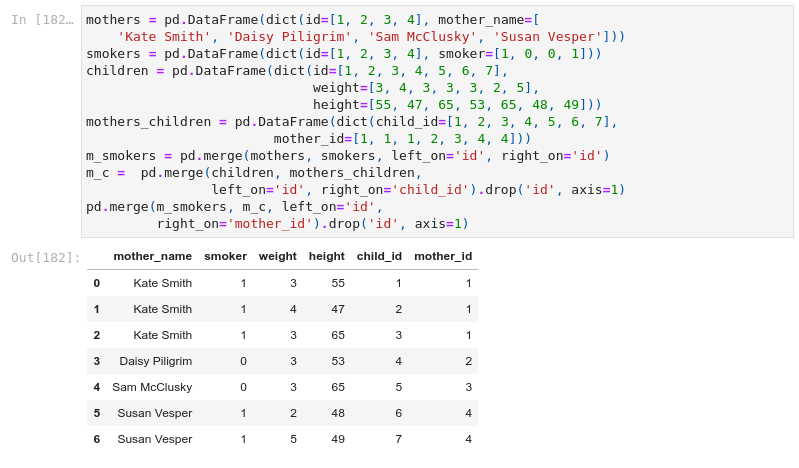
  
  
Рисунок 3.19 - Групування даних та застосування агрегуючих функцій

Виділимо підмасив за певними ознаками, наприклад, матері, що палять, та страші 30.

  
  
Рисунок 3.20 - Виділення підмасиву за ознаками

## Створити декілька власних об’єктів DataFrame за такою ж тематикою, що й файл. Наприклад, якщо тема файлу – жаби, можна створити об’єкти, що містять розміри жаб, вагу, стать, кількість особин в популяції і т.д. Використати описані в теоретичних відомостях параметри методів merge та concat для різних видів злиття та об’єднання даних цих об’єктів

Створимо декілька датафреймів: mothers - містить id матері, її ім'я; smokers - містить id матері та поле smoker, чи палить вона; children - містить id дитини, її вагу, зріст; mothers\_children - містить поле child\_id та mother\_id, що позначає родинний зв'язок.

  
  
Рисунок 3.21 - Застосування функції merge для join-операцій

# Висновок

Під час виконання даної лабораторної роботи я ознайомився з основними структурами даних бібліотеки Pandas: Series DataFrame, операціями над ними.

У першому завданні виділив підмасиви за допомогою прямої та непрямої індексації — атрибутами loc та iloc відповідно. До того ж задав індекси методом set\_index.

У другому завданні продемонстрував роботу з рядками та стовпцями: додавання(concat, append) і видалення(drop).

У третьому завданні встановив за індекс стовпець “ID”, визначив основні характеристики методом describe, застосував агрегуючи функції(agg).

У четвертому завданні створив декілька датафреймів і застосував на них аналог SQL-join метод merge.