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 автоматики та управління в технічних системах

**Лабораторна робота №1**

Прикладні задачі машинного навчання

**Тема:** Введення в data science

Виконав Перевірив:

студент групи ІА-11: Нестерук А. О

Свіржевський Н.Л.

Київ 2023

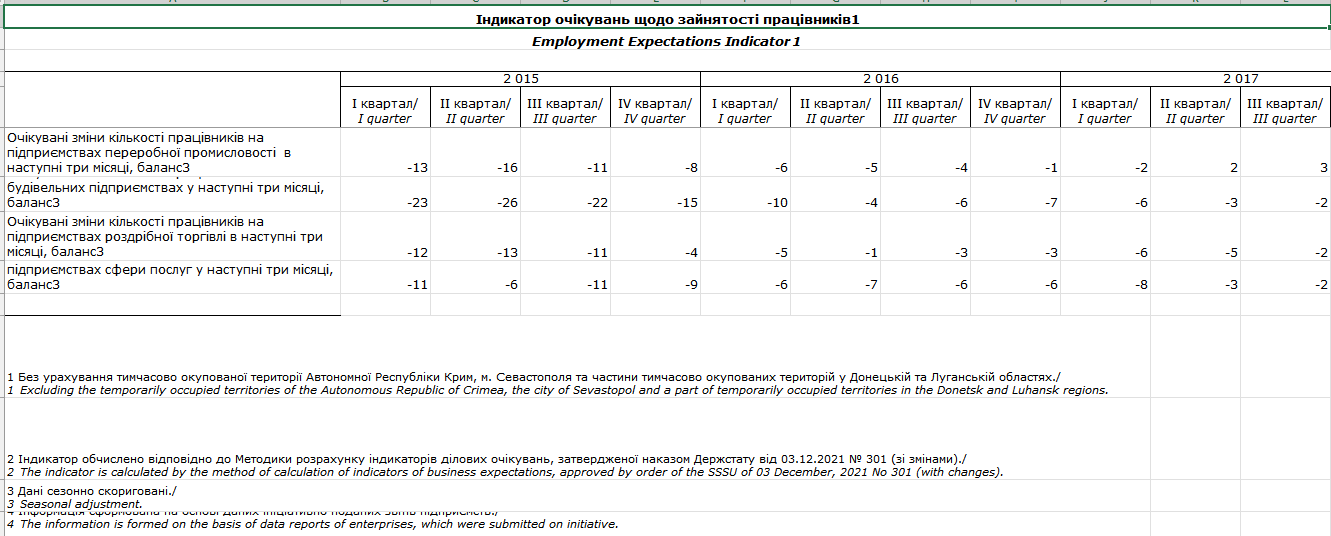
**Мета:** ознайомитися з бібліотеками мови Python: mathplotlib, pandas, numpy, дізнатися як працюють такі структури даних як Dataframe, Series, навчитися обробляти xslx та csv файли.

**Хід роботи**

1. **На сайті** [**http://www.ukrstat.gov.ua/**](http://www.ukrstat.gov.ua/) **обрати дані які для Вас є цікавими, можна використати будь-який ресурс з відкритими даними, та завантажте дані**

Посилання на обрані дані: <https://ukrstat.gov.ua/operativ/operativ2023/fin/io_zp/io_zp_15-23.xlsx>

Вигляд таблиці:



1. **Знайти математичне сподівання, медіану, моду, дисперсію, середньоквадратичне відхилення (поясніть їх зміст)**

Дістаємо дані з Excel документу

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

Path: шлях до документу

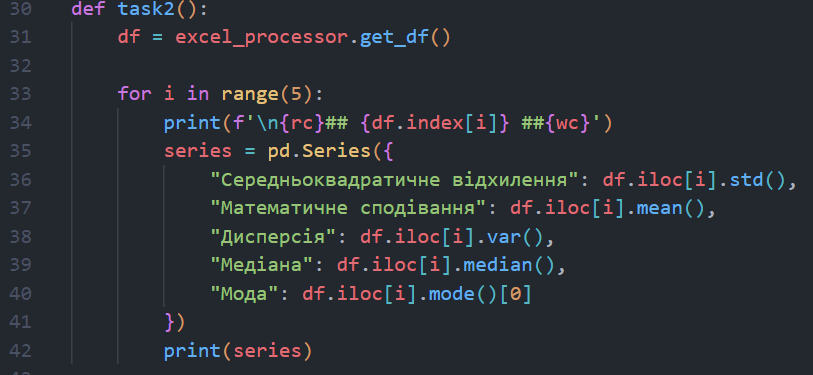
Index\_col: Певні значення, які будуть застосовуватися замість числових індексів (аналог – map)

Header: Назви колонок (стовпців), масив значень формує мультиіндекс (тобто до стовпців (групи стовпців) можна звертатися за різними індексами)

Skiprows: Рядки, які треба пропустити (тут пропускається рядок без значень)

Skipfooter: Рядки, які треба пропустити з кінця (останні 5 рядків)

**Знаходимо статистичні характеристики:**



Series застосовується для кращого форматування (текст зліва – величини справа)

Df.iloc – отримуємо рядок таблиці у вигляді Series

Відповідні методи для характеристик підписані на малюнку

Маємо таблицю з 5 рядків, для кожного відображаємо його характеристики:

Зображення, що містить текст

Автоматично згенерований опис

Математичне сподівання – середнє арифметичне значень множини

Медіана – центральне значення відсортованої множини даних (якщо значень два (кількість значень множини парна), то можна взяти середнє арифметичне цих двох значень)

Мода – значення, яке найчастіше зустрічається в множині

Дисперсія – міра розкиду значень відносно середнього арифметичного (наскільки далеко від середнього значення розташовані інші)

Знаходиться як середнє арифметичне квадратів різниці значень та середнього арифметичного множини.

Середньоквадратичне відхилення - квадратний корінь з дисперсії, ця метрика трохи компенсує ефект викидів

Чим менше дисперсія і стандартне відхилення, тим ближче значення даних до математичного сподівання і тим менше загальне «розпорошення» між значеннями і математичним сподіванням.

1. **Візуалізувати завантажені дані за допомогою гістограми**

Гістограма показує кількість значень, які належать певному проміжку

Зображення, що містить текст

Автоматично згенерований опис

Fig, ax – параметри графіків, plt.subplots(3, 2) робить так, щоб графіки відображались в таблиці 3х2

.axis(‘off’) – видалення графіка (у нас 5 рядків зі значеннями)

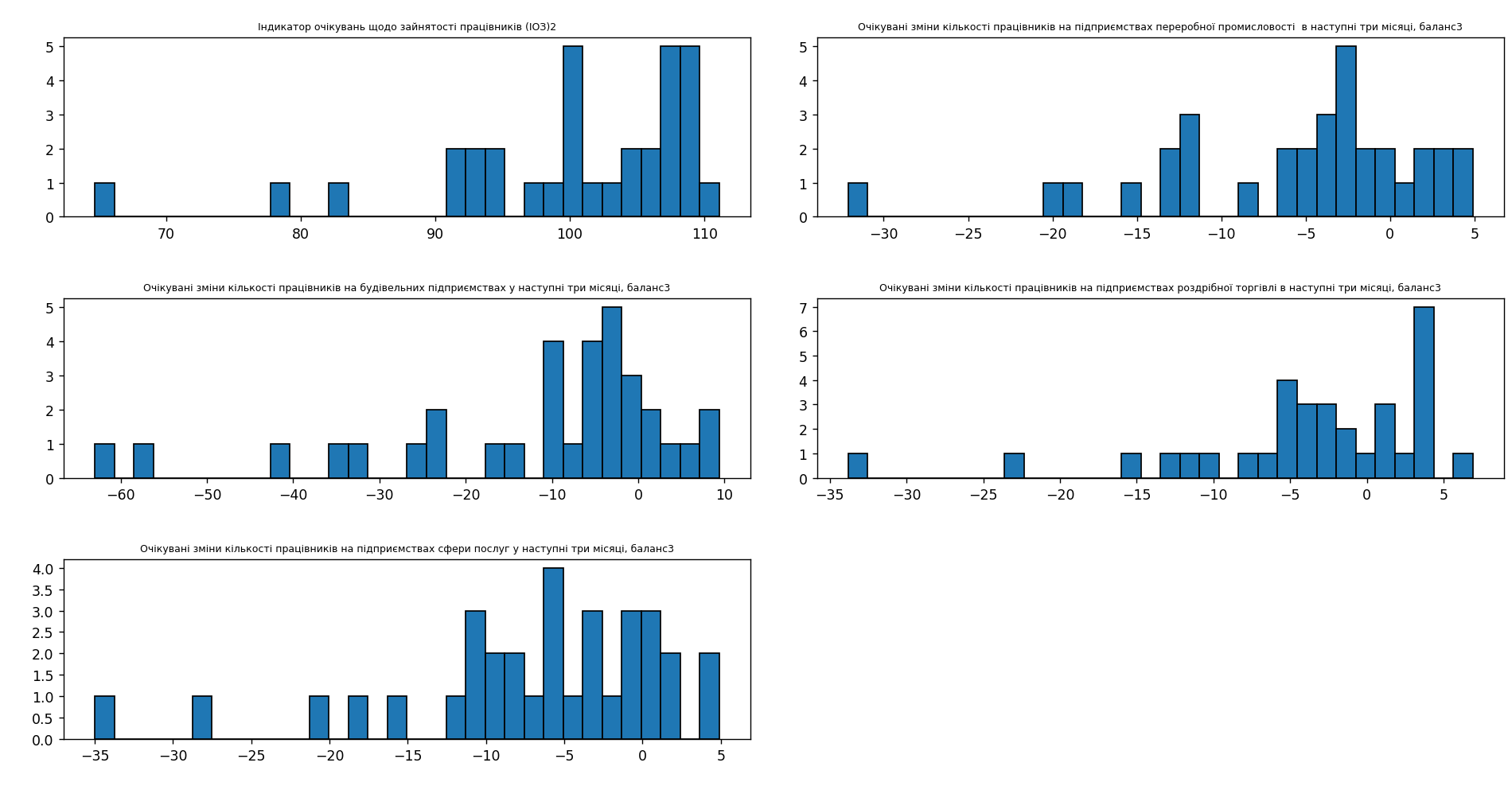
.hist() – встановлення параметрів для гістограми

X – значення, bins – максимальна кількість стовпців, edgecolor – колір обведення стовпців

Plt.show – відобразити графіки

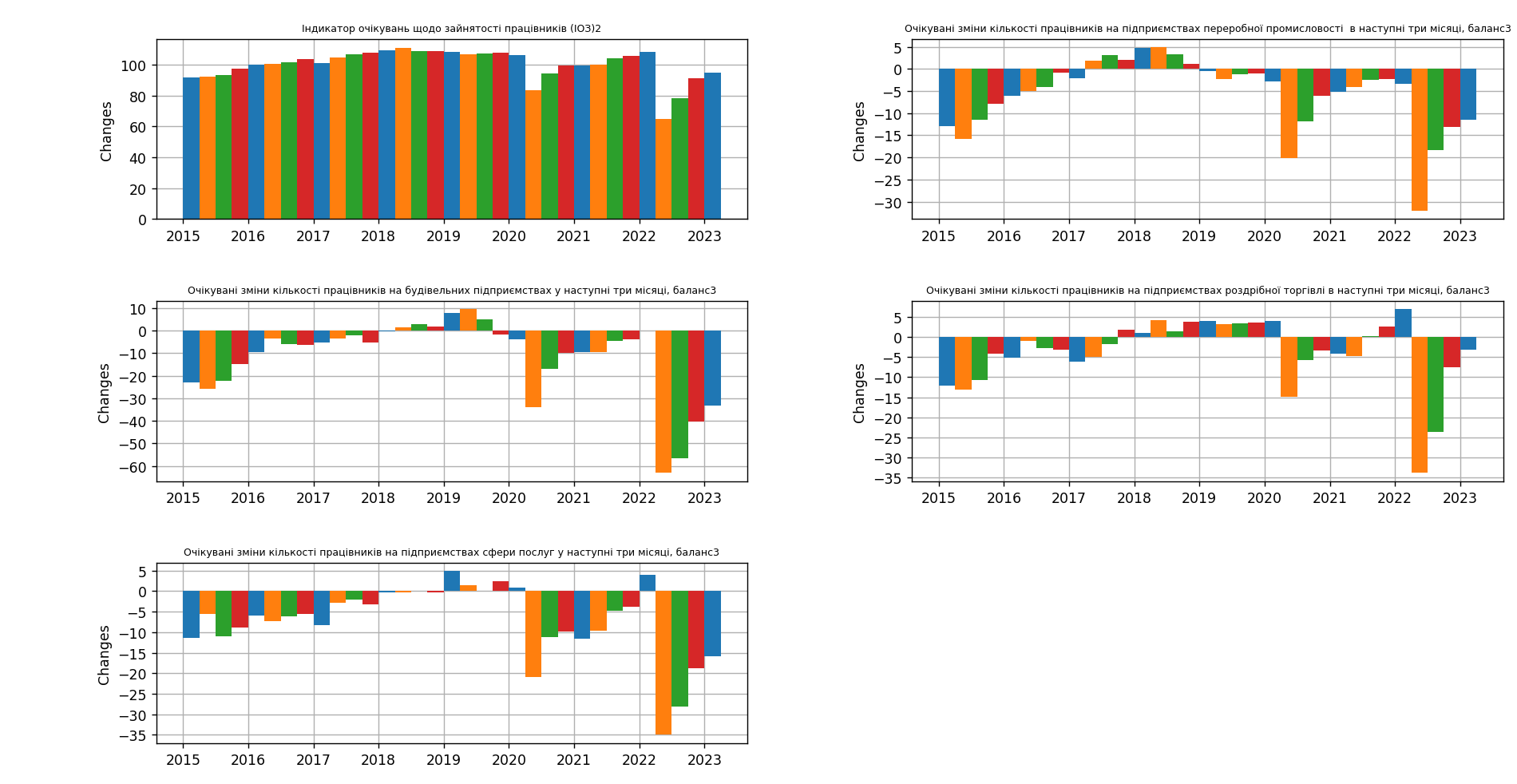
Зображення, що містить текст

Автоматично згенерований опис



Зображено гістограми для кожного з рядків таблиці.

Барчарти для наглядності результатів:



Побудова барчарту:

Зображення, що містить текст

Автоматично згенерований опис

*Короткий опис:* на одному графіку будуються 4 барчарти (для кожного кварталу)

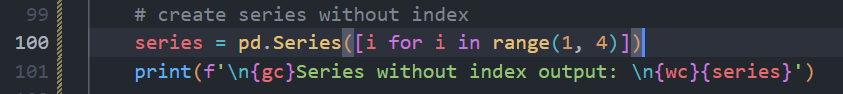
1. **Для цих даних проробити всі дії з пункту колекції Series і DataFrame бібліотеки pandas**

*Нотатка: перейменуємо назви стовпців для скорочення розмірів:*



*I квартал / I quarter -> І*

***Series – 1) Створення Series***

Створити Series без індексів можна так: 

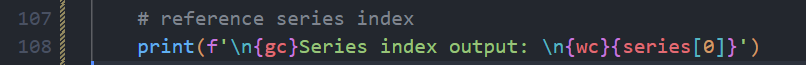
Зображення, що містить текст

Автоматично згенерований опис

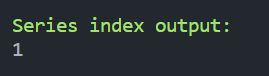
В даному випадку індексами виступають значення 0, 1, 2…

***Series – 2) Звернення до елементів Series***

Звернутися до елементу можна так:

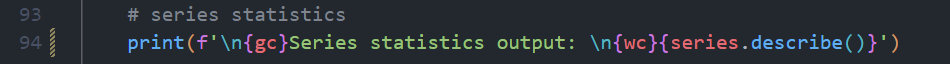


Де в квадратних дужках вводимо значення індексу



Отримали значення першого рядка

***Series – 3) Описові статистики Series***

******

Ця функція виводить всі описові статистики

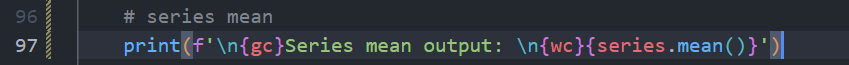
Зображення, що містить текст

Автоматично згенерований опис

25%, 75%, 50% - медіани відповідних частин (перша половина масиву, друга, весь масив) відсортованих значень (у випадку двох медіан береться середнє значення)

Std – стандартне відхилення

Вивести одну характеристику можна викликавши відповідну функцію:



Зображення, що містить текст

Автоматично згенерований опис

***Series – 4) Створення колекції Series з нестандартними індексами***

Створювати такі колекції можна двома способами:

***Зображення, що містить текст, монітор, екран, відображений

Автоматично згенерований опис***

***Зображення, що містить текст

Автоматично згенерований опис***

В першому випадку ми передаємо два масиви (другий з параметром index=[])

В другому випадку ми передаємо map

Також можна створити мультиіндекс:

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

***Series – 5) Звернення до елементів Series з використанням нестандартних індексів***

Можна звернутися до нестандартного індекса, просто вписавши його в квадратних дужках:



Зображення, що містить текст

Автоматично згенерований опис

Series також має багато вбудованих атрибутів, наприклад –

Зображення, що містить текст

Автоматично згенерований опис

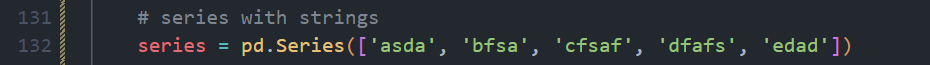
Зображення, що містить текст

Автоматично згенерований опис

Перший виводить тип значень колекції, другий значення колекції.

***Series – 6) Створення колекції Series із строковими елементами***

Створюємо колекцію:



Існує багато функцій для роботи з стрінгами, наприклад:

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

Перша ф-я показує, чи містять слова певну букву, друга перетворює слова в верхній регістр.

***Dataframe – 1) Створення DataFrame***

Використовуємо Dataframe з ексель-таблиці:

Зображення, що містить текст

Автоматично згенерований опис

Дана функція створює Dataframe з ексель-таблиці

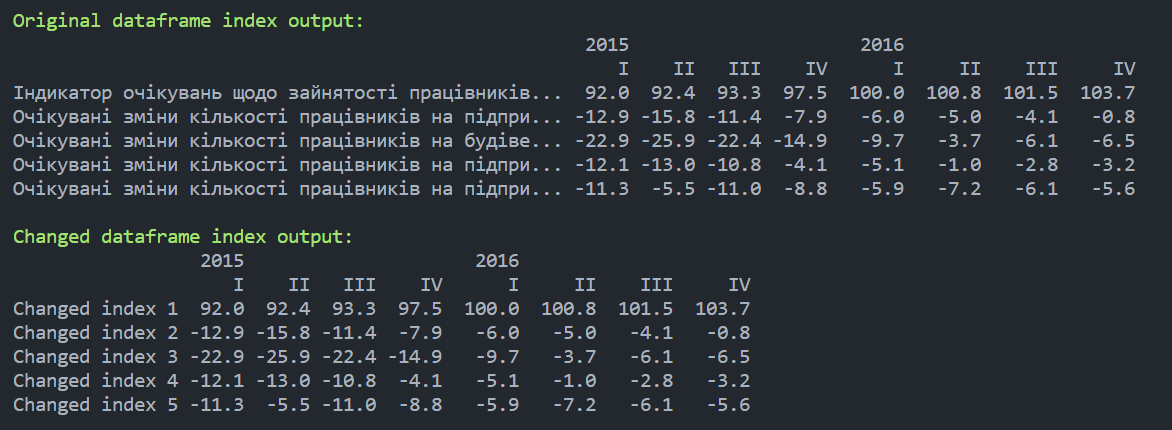
***Dataframe – 2) Налаштування індексів DataFrame з використанням атрибута index***

Змінимо індекси нашої таблиці:

Зображення, що містить текст, екран

Автоматично згенерований опис

(iloc буде пояснено пізніше, тут застосуємо його щоб зменшити розмір таблиці)



Як бачимо, індекси змінено

***Dataframe – 2) Звернення до стовпців DataFrame***

Звертаємося, просто ввівши назву колонки в квадратні дужки



Зображення, що містить текст

Автоматично згенерований опис

Можна звернутися за мультиіндексом:



Зображення, що містить текст

Автоматично згенерований опис

Бачимо, що повернувся тільки один стовпчик

***Dataframe – 3) Вибір рядків з використанням атрибутів loc і iloc***

Loc – звернення за створеним індексом

Iloc – звернення за числовим індексом

