НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

**Лабораторна робота № 1**

з дисципліни «Прикладні задачі машинного навчання»

Тема: «Введення в data science»

|  |  |
| --- | --- |
| **Прийняв:** | **Виконав:**  **студент групи ІП-13**  **Недельчев Є.О.** |

Київ – 2023

**Завдання:**

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.**
2. **Знайти математичне сподівання, медіану, моду, дисперсію, середньоквадратичне відхилення (поясніть їх зміст)**
3. **Візуалізувати завантажені дані за допомогою гістограми**
4. **Для цих даних проробити всі дії з пункту колекції Series і DataFrame бібліотеки pandas**
5. **Прочитати набір даних катастрофи «Титаніка»**
6. **Завантажити набір даних катастрофи «Титаніка» за URL-адресою**
7. **Переглянути рядки набору даних катастрофи «Титаніка»**
8. **Налаштувати назви стовпців**
9. **Провести простий аналіз даних**
10. **Побудувати гістограму віку пасажирів**

**Виконання**

**1.**

При виконанні роботи було використано набір даних, присвячений комп’ютерним іграм, які було випущено станом на грудень 2022 року. Спочатку імпортуємо всі необхідні для роботи бібліотеки:

Изображение выглядит как текст

Автоматически созданное описание

Далі завантажимо сам датасет:

****

Виконаємо первинну обробку даних (видалимо зайвий стовбець sort\_no та перейменуємо стовпець metascore на більш зрозумілий rating):



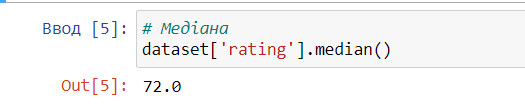
**2.**

Знайдемо математичне сподівання рейтингу ігор (те, що очікується в середньому в результаті багатьох випробувань):

Изображение выглядит как текст

Автоматически созданное описание

Знайдемо медіану (значення, яке розділяє масив даних на дві рівні частини: половина значень масиву більша за медіану, а інша половина – менша):



Знайдемо моду (це значення, яке зустрічається найбільшу кількість разів у масиві даних):

Изображение выглядит как текст

Автоматически созданное описание

Знайдемо дисперсію (наскільки далеко значення в масиві даних розташовані від середнього значення):

Изображение выглядит как текст

Автоматически созданное описание

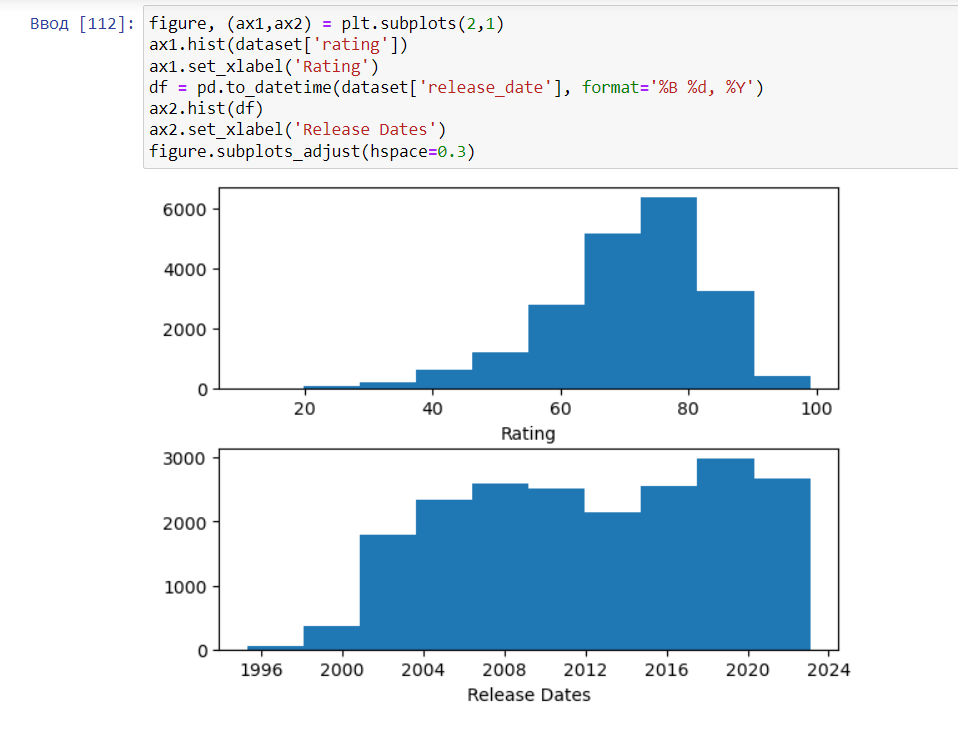
Знайдемо середньоквадратичне відхилення(наскільки сильно значення розкидані відносно середнього):

Изображение выглядит как текст

Автоматически созданное описание

**3.**

Візуалізуємо завантажені дані за допомогою гістограми:



**4.** **Виконаємо усі дії** **з пункту колекції Series і DataFrame бібліотеки pandas:**

Створимо Series, використавши дані зі стовпця ‘rating’ попереднього датафрейму:

Изображение выглядит как стол

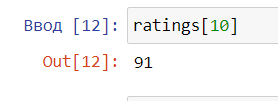
Автоматически созданное описание

А зараз створимо Series за допомогою функції range():

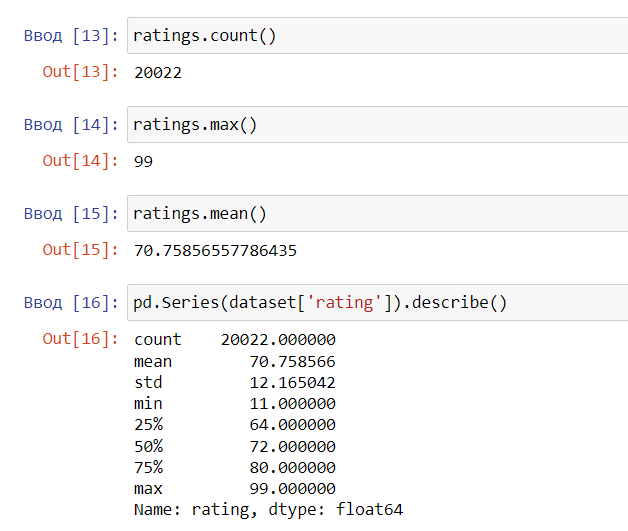
Изображение выглядит как текст

Автоматически созданное описание

Звернемося до десятого елементу:



Обчислимо описові статистики:



Створимо колекцію з нестандартними індексами:

Изображение выглядит как текст

Автоматически созданное описание

Використаємо в якості ініціалізатора словник:

Изображение выглядит как текст

Автоматически созданное описание

Звернемося до елементу з використанням нестандартного індексу:

Изображение выглядит как текст

Автоматически созданное описание

Створимо колекцію DataFrame на базі словника:

Изображение выглядит как текст

Автоматически созданное описание

Налаштуємо індекси з використанням атрибуту index:

Изображение выглядит как текст

Автоматически созданное описание

Звернемося до стовпця ‘name’:

Изображение выглядит как текст, стол

Автоматически созданное описание

А зараз до стовпця country:

Изображение выглядит как текст

Автоматически созданное описание

Виберемо рядки за допомогою loc та iloc:

Изображение выглядит как текст

Автоматически созданное описание

Індексом також може бути цілий сегмент:

Изображение выглядит как текст

Автоматически созданное описание

А для вибору конкретних рядків використаємо синтаксис списку:

Изображение выглядит как текст

Автоматически созданное описание

Виберемо підмножину рядків і стовпців:

Изображение выглядит как текст

Автоматически созданное описание

Створимо новий датафрейм для більш зручної демонстрації логічного індексування:

Изображение выглядит как текст

Автоматически созданное описание

Виберемо усі значення, які більші за 7:Изображение выглядит как стол

Автоматически созданное описание

Або значення, які одночасно більші за 10 та менші за 20:

Изображение выглядит как стол

Автоматически созданное описание

Звернемося до конкретного осередку DataFrame по рядку і стовпцю: Изображение выглядит как текст

Автоматически созданное описание

Використаємо метод describe(), щоб отримати описову статистику:

Изображение выглядит как стол

Автоматически созданное описание

Транспонуємо наш датафрейм, скориставшись атрибутом Т:

Изображение выглядит как стол

Автоматически созданное описание

Та отримаємо його описову статистику:

Изображение выглядит как стол

Автоматически созданное описание

Відсортуємо рядки датафрейму за індексами:

Изображение выглядит как стол

Автоматически созданное описание

Або за стовпцями:

Изображение выглядит как стол

Автоматически созданное описание

Відсортуємо датафрейм за значенням стовпця ‘age’

Изображение выглядит как текст

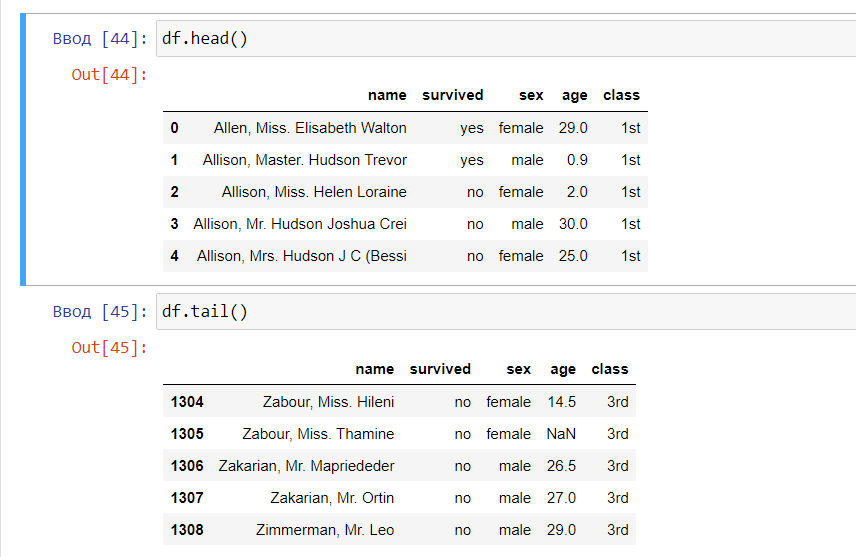
Автоматически созданное описание

**5. Прочитаємо та 6. Завантажимо набір даних катастрофи Титаніка:**

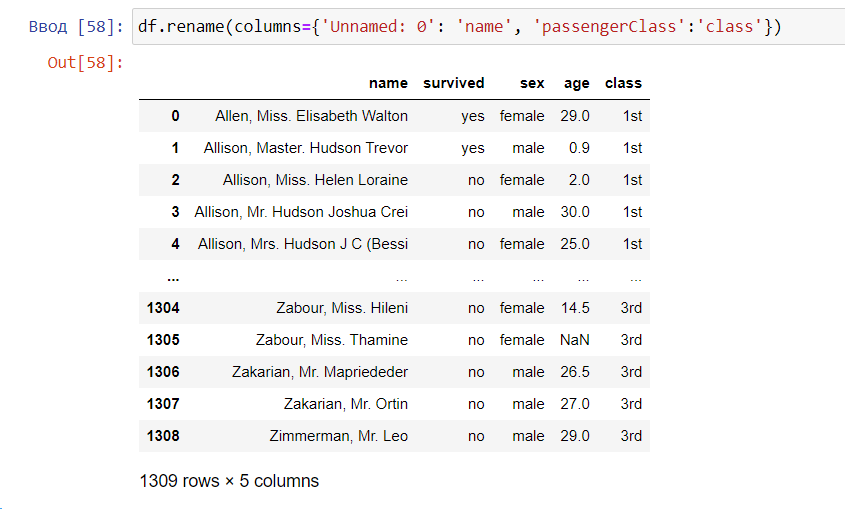
**Изображение выглядит как стол

Автоматически созданное описание**

**7. Переглянемо рядки набору даних, використавши функції head() і tail():**

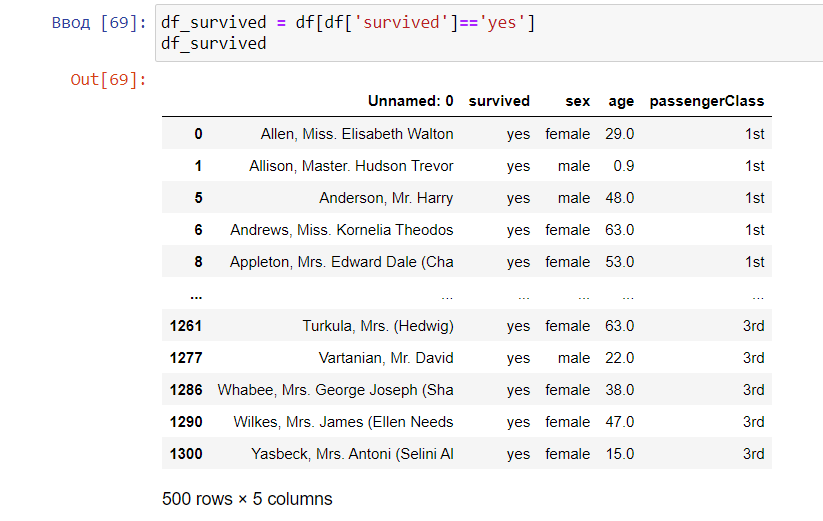
****

**8. Налаштуємо назви стовпців:**

****

**9. Проведемо простий аналіз даних:**

Спочатку виокремимо пасажирів, яким вдалося вижити:



* Визначимо наймолодшого пасажира серед них:

Изображение выглядит как текст

Автоматически созданное описание

* Найстаршого:

Изображение выглядит как текст

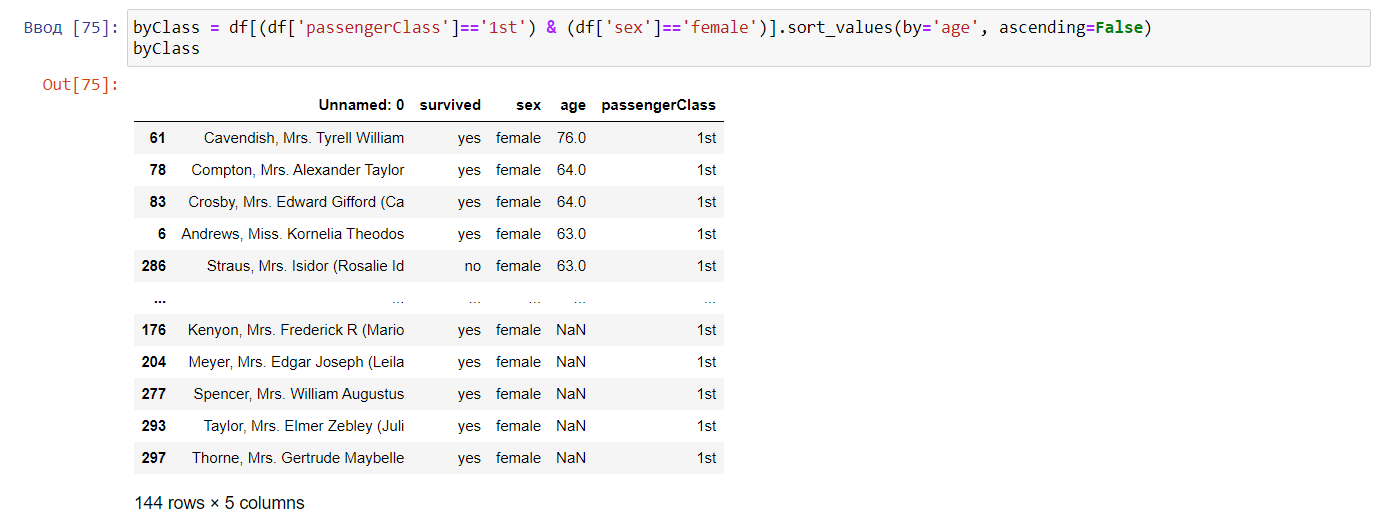
Автоматически созданное описание

* Середній вік:

Изображение выглядит как текст

Автоматически созданное описание

Знайдемо жінок-пасажирів першого класу, яким вдалося вижити, та відсортуємо їх за віком:

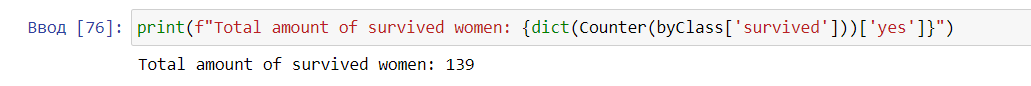


Знайдемо наймолодшу та найстаршу серед них:

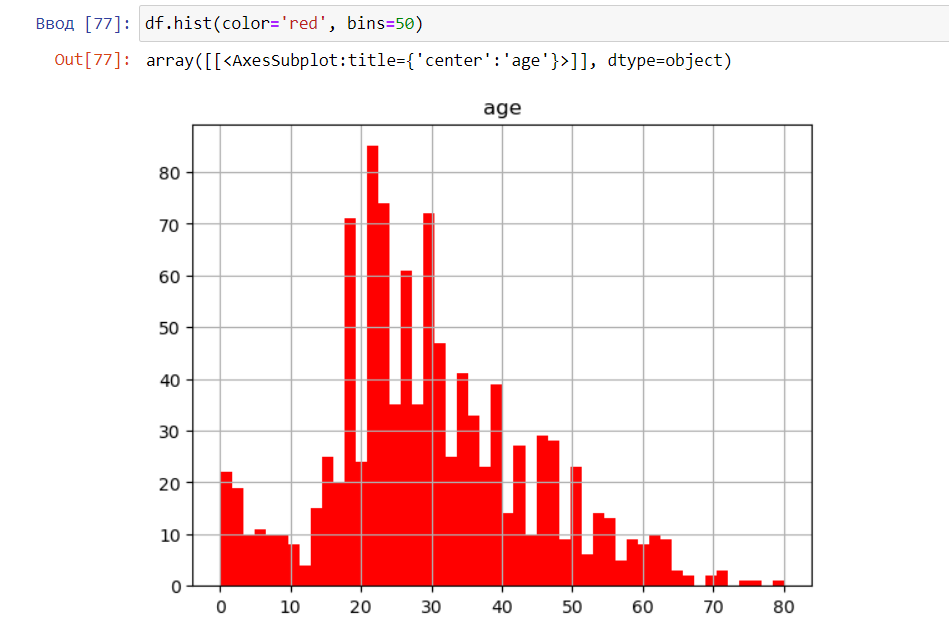
Изображение выглядит как текст

Автоматически созданное описание

Загальна кількість виживших жінок, які були пасажирками першого класу:



1. **Побудуємо гістограму віку пасажирок:**



**Висновок**

Виконуючи цю лабораторну роботу я ознайомився з бібліотеками pandas та matplotlib. Також було проведено невеликий аналіз даних, присвячених катастрофі лайнера «Титанік».