Державний торговельно-економічний університет  
Факультет інформаційних технологій  
Кафедра інженерії програмного забезпечення та кібербезпеки

**ЛАБОРАТОРНА РОБОТА №1**

**З ДИСЦИПЛІНИ «Моделі і структури даних»**

**НА ТЕМУ: «Математичні основи машинного навчання»**

**Виконавець**:

Студентка групи 4-4

Авєріна Наталія Ігорівна

Київ 2024

**Лабораторна робота №1**

**Тема**: «Математичні основи машинного навчання»

**Цілі**: навчитися виконувати первинний аналіз даних, розрахувати основні статистичні показники для числових стовбчиків, виконати різні типи візуалізацій для аналізу даних, сформувати висновки на основі статистичних показників і візуалізацій.

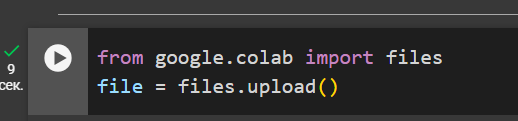
**Хід виконання:**

Завдання. Варіант 2

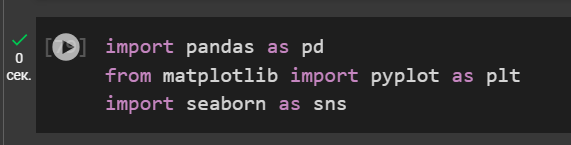
Оберіть власний датасет, де було б не менше сотні рядків, а також як числові, так і категоріальні стовпчики. Зробіть EDA (Exploratory data analysis), подібний до того, який ви могли бачити на другому практичному занятті. Зробіть не менше трьох різних візуалізацій (не однотипних). Для кожного числового стовпчика порахуйте середнє, стандартне відхилення, а також квартилі. Спробуйте зробити якісь висновки про дані. Ви можете користуватись будь-якими технологіями та мовою програмування, але ваша робота буде оцінена лише після невеличкого демо (захисту) під час одного з практичних занять. Дані можна взяти, наприклад, тут: <https://www.kaggle.com/datasets>

**Виконана робота:** <https://colab.research.google.com/drive/1oWo1EV8Zao5cpAbkY7aG0tVX_8otFJ-7?usp=sharing>

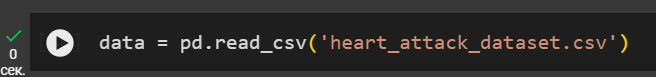
1. Завантажуємо датасет на ПК. Використаємо наступний код для завантаження файлу в ноутбук Google Colab.



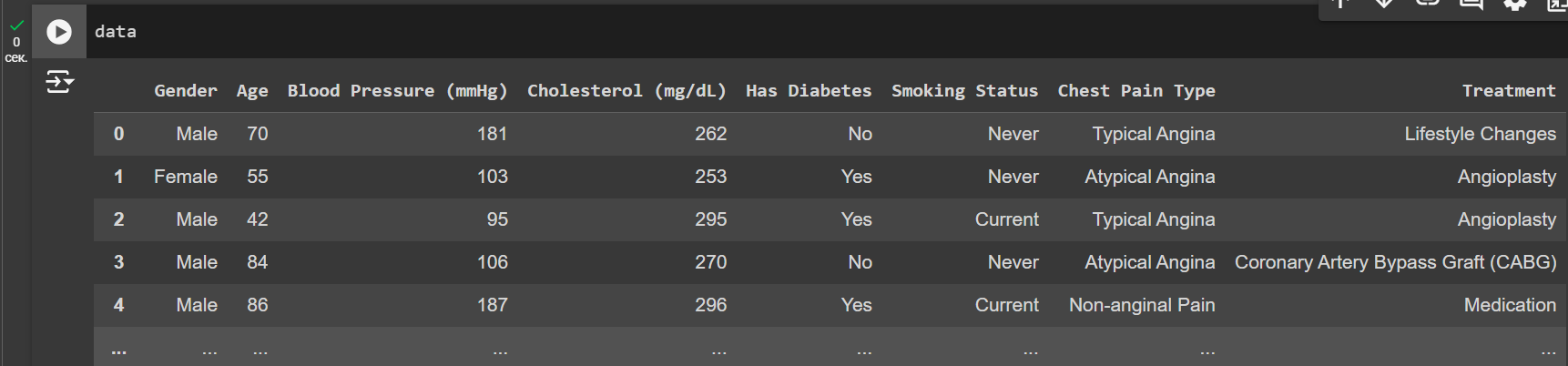
1. Імпортуємо потрібні бібліотеки для роботи з даними

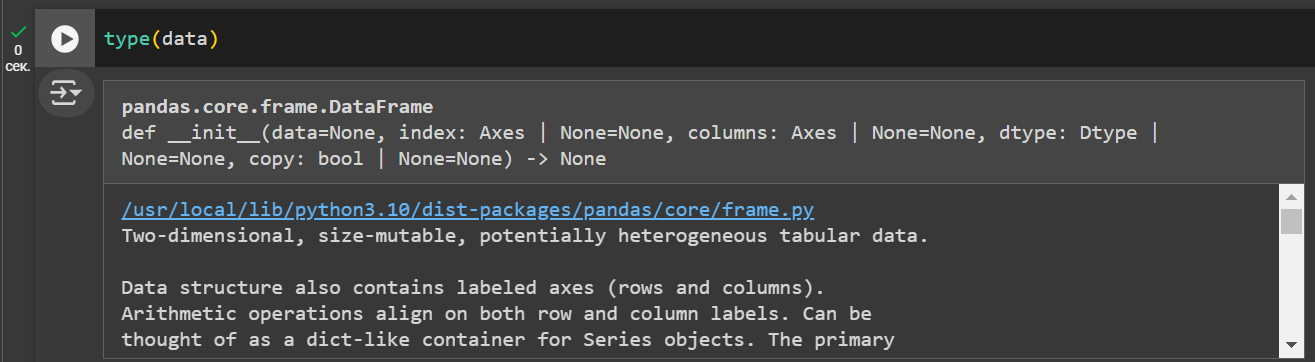


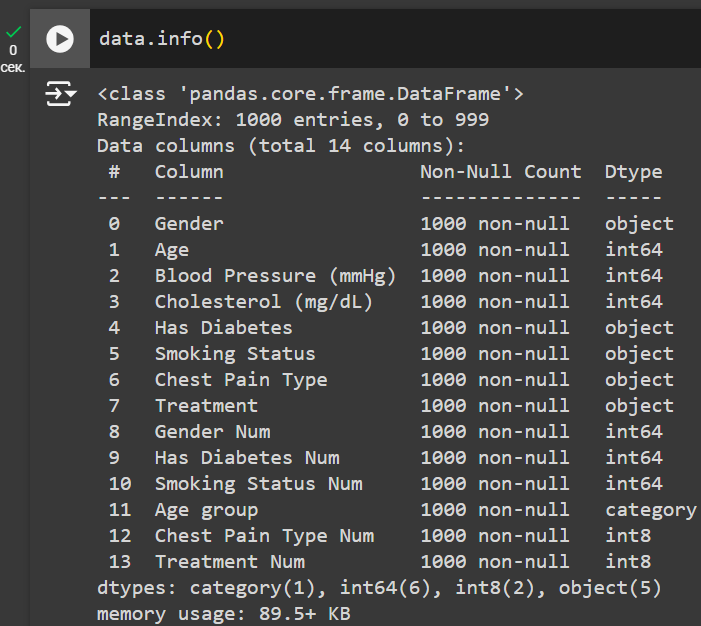
1. Завантажуємо файл у датафрейм за допомогою бібліотеки Pandas.

****

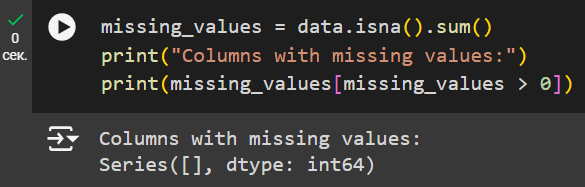
1. Переглядаємо дані, тип даних та загальну інформацію про створений датафрейм.



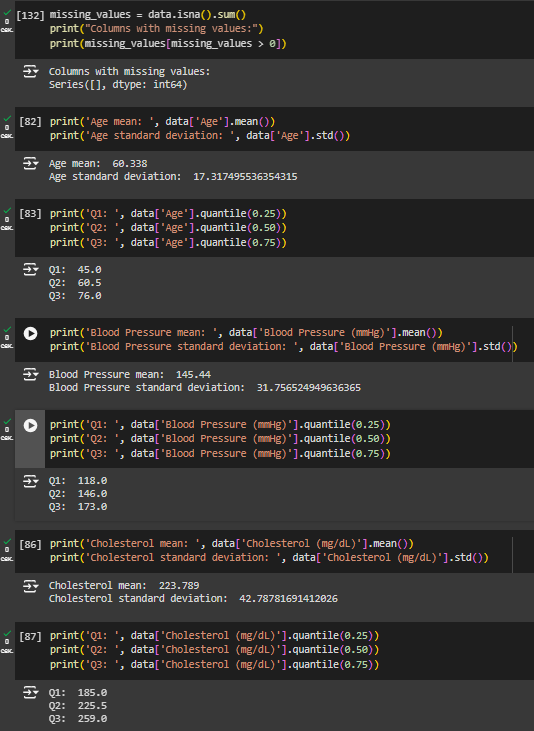




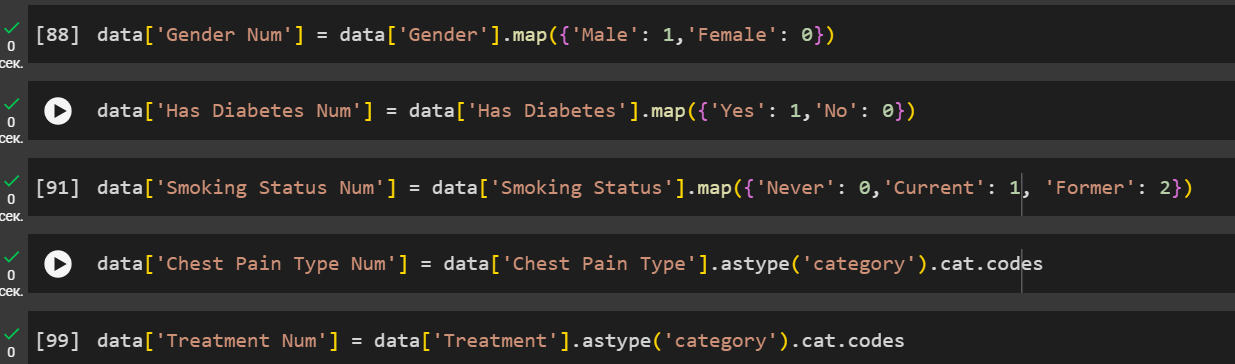
1. Перевіряємо чи є пропущені дані у стовбчиках. Наш датасет таких не має.



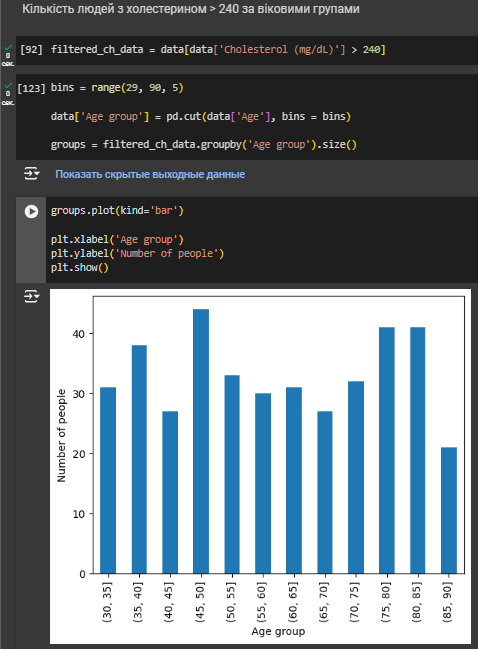
1. Для кожного числового стовпчика розрахуємо середнє, стандартне відхилення, а також квартилі.



1. Перетворюємо категорійні стовбці на числові для подальшого аналізу.



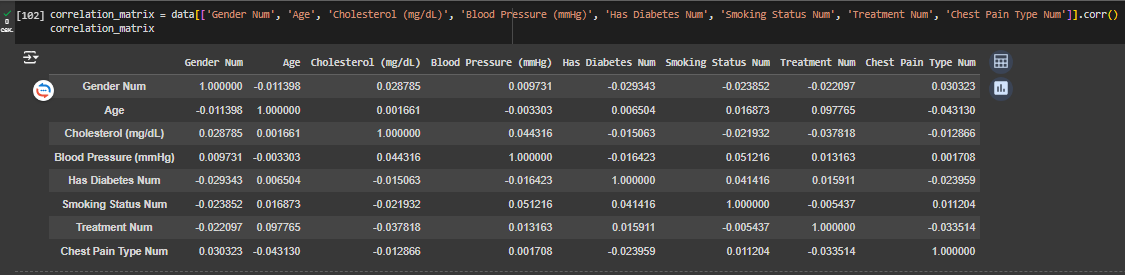
1. Створюємо стовбчасту діаграму, яка показує кількість людей з холестерином більшим за 240 (критичний рівень) за віковими проміжками по 5 років.

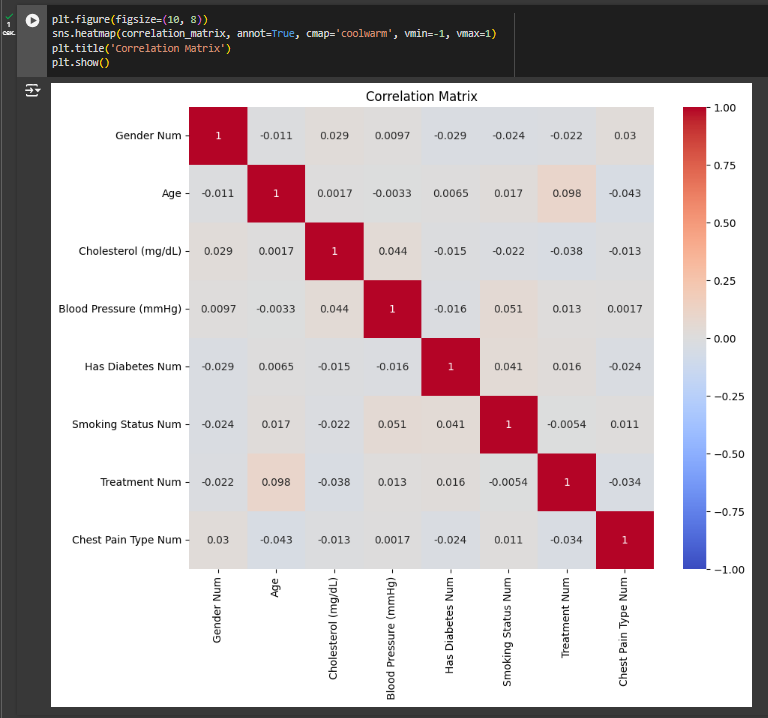


Можна побачити, що найбільшу кількість людей з критичним рівнем холестерину в крові має діапазон віку від 45 до 50 років. Це може бути пов’язане з віковими змінами в організмі, які починаються в середньому віці. Після, бачимо зниження рівня. Це може бути пов’язане з більшим піклуванням про здоров’я після початку проблем із ним. В діапазоні від 75 до 85 бачимо знову дуже велику кількість людей з критичним рівнем холестерину. В цьому випадку може бути знижена ефективність лікування або менша увага до контролю показника.

Також такі дані можуть бути не точними, адже багато факторів не враховані в аналіз.

1. Створимо кореляційну матрицю для оцінки взаємозв’язків між змінними.

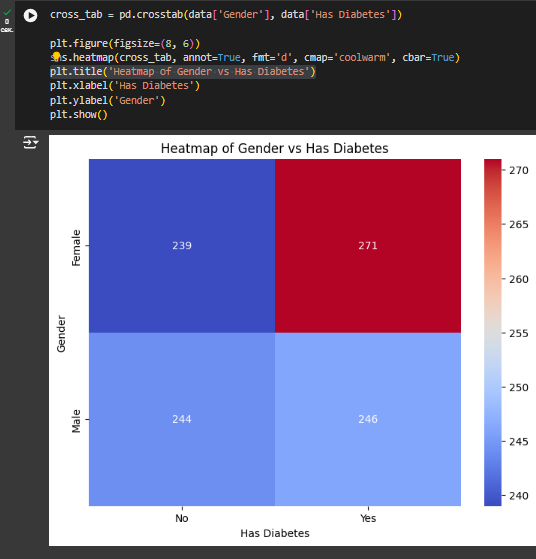




Матриця показує, що між змінними майже не корелюють одна з одною (наближені до 0). Найбільша кореляція є між віком пацієнта та способом лікуванням. Можемо допустити, що відповідно до віку лікарі можуть назначати різні види лікування. Наприклад, молодим людям – медикаментозне лікування або зміну способу життя на більш здоровий. Людям середнього віку, коли проблема може бути більш серйозною – хірургічні втручання та препарати, і звісно змінення способу життя. Людям дуже похилого віку лікарі можуть не запропонувати хірургічне втручання, адже це може бути небезпечно для життя.

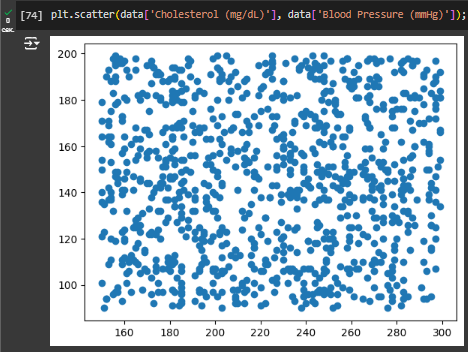
Загалом можна сказати, що цей датасет має незалежні одна від одної змінні. Це ускладнює виявлення чітких закономірностей.

1. Створюємо теплову карту залежності між статтю та наявністю діабету.



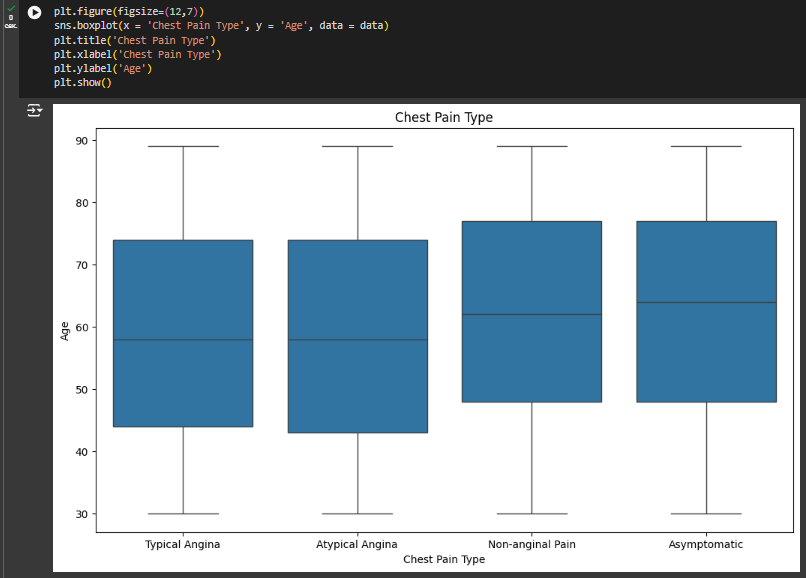
Можна чітко побачити, що страждають на діабет більше жінки, ніж чоловіки.

1. Створюємо діаграму розсіювання, яка візуалізує зв'язок між значеннями рівня холестерину та артеріального тиску



Точки розкидані випадковим чином, без чіткої форми або напряму. Це вказує на відсутність лінійного зв'язку між змінними.

1. Створюємо коробкову діаграму для візуалізації розподілу віку серед різних видів болю.



Лінія всередині коробки вказує середнє значення віку. Бачимо, що у діапазоні від 43 до 74 років люди частіше стикаються з типічним чи нетипічним болем. А від 47 до 77 років з неангінальним болем або навіть безсимптомно.

**Висновки**:

Проведений EDA дозволив виявити певні закономірності та аномалії у даних. Деякі змінні виявилися слабо взаємопов'язаними, що ускладнює виявлення чітких залежностей. Це свідчить про те, що подальший аналіз може вимагати використання більш складних моделей для розуміння взаємодії між змінними.

Виявлені аномалії та тенденції можуть бути корисними для розробки подальших рекомендацій щодо лікування пацієнтів різних вікових груп та визначення ризиків розвитку серцево-судинних захворювань.

**Відповіді на контрольні питання:**

1. Нехай A, B -- квадратні матриці. Чи може бути так, що AB=BA? Поясніть.   
     
   Так, може бути. Це означає, що матриці комутують одна з одною. Це можливо, коли матриці є діагональними, або А = В, або одна з матриць є скалярною, або А і В є степенями однієї і тієї ж матриці.
2. Дайте визначення наступним поняттям: локальний мінімум та локальний максимум функції, глобальний мінімум та глобальний максимум функції. **Локальний мінімум** функції — це така точка x0, що в деякому околі цієї точки значення функції f(x0) є меншим або рівним, ніж у всіх інших точках цього околу. Тобто f(x0)≤f(x) для всіх x в околі точки x0​.

**Локальний максимум** функції — це така точка x0​, що в деякому околі цієї точки значення функції f(x0) є більшим або рівним, ніж у всіх інших точках цього околу. Тобто f(x0)≥f(x) для всіх x в околі точки x0​​.

**Глобальний мінімум** функції — це така точка x0​, що значення функції f(x0) є меншим або рівним, ніж у всіх інших точках області визначення функції. Тобто f(x0)≤f(x) для всіх x в області визначення функції.

**Глобальний максимум** функції — це така точка x0​, що значення функції f(x0) є більшим або рівним, ніж у всіх інших точках області визначення функції. Тобто f(x0)≥f(x) для всіх x в області визначення функції.

1. Опишіть метод, за допомогою якого ви шукали мінімуми і максимуми функцій у школі.
   1. Спочатку знаходили похідну функції f′(x).
   2. Визначали критичні точки, де f′(x)=0 або де f′(x) не існує.
   3. Далі перевіряли знак похідної f′(x) до і після кожної критичної точки. Якщо похідна змінювала знак з позитивного на негативний, то точка була локальним максимумом. Якщо похідна змінювала знак з негативного на позитивний, то точка була локальним мінімумом.
   4. Для підтвердження глобальних екстремумів розглядали поведінку функції на межах області визначення.
2. Припустимо, дані містять стовпчик “class” та має значення “A”, “B”, “C”. Як, на вашу думку, доцільно перетворити цей стовпчик у числовий?

Є 2 варіанти:

* 1. Кожному унікальному значенню присвоюється ціле число. Наприклад: "A" = 0, "B" = 1, "C" = 2
  2. Створюються окремі бінарні стовпчики для кожної категорії. Наприклад:

Для "A" створюється стовпчик Class\_A зі значенням 1, якщо рядок має значення "A", і 0, якщо ні.

Для "B" створюється стовпчик Class\_B зі значенням 1, якщо рядок має значення "B", і 0, якщо ні.

Для "C" створюється стовпчик Class\_C зі значенням 1, якщо рядок має значення "C", і 0, якщо ні.

1. Припустимо, дані містять стовпчик “colour” та має значення “red”, “green”, “blue”, “yellow”, “black”. Як, на вашу думку, доцільно перетворити ці дані у числові?

Можна використати ці ж 2 підходи:

* 1. Присвоїти кожному кольору числовий індекс, наприклад, red = 0, green = 1, blue = 2, yellow = 3, black = 4.
  2. Створити окремий бінарний стовпчик для кожного кольору (Colour\_red, Colour\_green, Colour\_blue, Colour\_yellow, Colour\_black), де для кожного рядка в одному зі стовпчиків буде 1, а в інших — 0.