Міністерство освіти і науки України

Національний технічний університет України  
«Київський політехнічний інститут

імені Ігоря Сікорського»

Факультет БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ

Кафедра БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

ЗВІТ

про виконання розрахунково-графічної роботи

з кредитного модуля «Обробка та аналіз біомедичних даних»

Виконала: студента гр. БС-83 Милейко Костянтина

Прийняв: викладач 2020 р.

(Прізвище, Ім’я) (Оцінка, Підпис) (Дата)

Київ

КПІ ім. Ігоря Сікорського

2020

ЗМІСТ

[**РОЗДІЛ 1. ОЗНАЙОМЛЕННЯ З JUPYTER NOTEBOOK ТА БІБЛІОТЕКОЮ PANDAS** 4](#_Toc41278692)

[1.1. Теоретичні відомості 4](#_Toc41278693)

[1.2. Практична реалізація 6](#_Toc41278694)

[1.3 Висновки 10](#_Toc41278695)

[**РОЗДІЛ 2. ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ БІБЛІОТЕКИ MATPLOTLIB** 11](#_Toc41278696)

[2.1. Теоретичні відомості 11](#_Toc41278697)

[2.2. Практична реалізація 13](#_Toc41278698)

[2.3. Висновки 19](#_Toc41278699)

[**РОЗДІЛ 3. ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ БІБЛІОТЕКИ SEABORN** 20](#_Toc41278700)

[3.1. Теоретичні відомості 20](#_Toc41278701)

[3.2. Практична реалізація 21](#_Toc41278702)

[3.3. Висновки 26](#_Toc41278703)

[**РОЗДІЛ 4. ПЕРЕВІРКА ДАНИХ НА НОРМАЛЬНІСТЬ ЗА ДОПОМОГОЮ КРИТЕРІЮ ШАПІРА-УІЛКА** 27](#_Toc41278704)

[4.1. Теоретичні відомості 27](#_Toc41278705)

[4.2. Практична реалізація 28](#_Toc41278706)

[4.3. Висновки 31](#_Toc41278707)

[**РОЗДІЛ 5. КРИТЕРІЙ ПІРСОНА ДЛЯ ПОРІВНЯННЯ ДЕКІЛЬКОХ ГРУП ЗА РОЗПОДІЛЕННЯМ ОЗНАКИ** 32](#_Toc41278708)

[5.1. Теоретичні відомості 32](#_Toc41278709)

[5.2. Практична реалізація 34](#_Toc41278710)

[5.3. Висновки 37](#_Toc41278711)

[**РОЗДІЛ 6. ПАРНИЙ T-КРИТЕРІЙ СТЬЮДЕНТА** 38](#_Toc41278712)

[6.1 Теоретичні відомості 38](#_Toc41278713)

[6.2. Практична реалізація 39](#_Toc41278714)

[6.3 Висновки 42](#_Toc41278715)

[**РОЗДІЛ 7. T – КРИТЕРІЙ СТЬЮДЕНТА ДЛЯ НЕЗАЛЕЖНИХ ВИБІРОК** 43](#_Toc41278716)

[7.1. Теоретичні відомості 43](#_Toc41278717)

[7.2. Практична реалізація 44](#_Toc41278718)

[7.3. Висновки 47](#_Toc41278719)

[**РОЗДІЛ 8. ДИСПЕРСІЙНИЙ АНАЛІЗ** 48](#_Toc41278720)

[8.1 Теоретичні відомості 48](#_Toc41278721)

[8.2. Практична реалізація 50](#_Toc41278722)

[8.3. Висновки 53](#_Toc41278723)

[**СПИСОК ЛІТЕРАТУРИ** 54](#_Toc41278724)

**РОЗДІЛ 1. ОЗНАЙОМЛЕННЯ З JUPYTER NOTEBOOK ТА БІБЛІОТЕКОЮ PANDAS**

* 1. Теоретичні відомості

Jupyter Notebook – це командна оболонка для інтерактивних обчислень. Цей інструмент може використовуватися не лише на Python, але й на інших мовах програмування: Julia, R, Haskell і Ruby. Він часто використовується для роботи з даними, статистичним моделюванням і машинним навчанням. Jupyter notebook є графічної веб-оболонкою для IPython, яка розширює ідею консольного підходу до інтерактивних обчислень.

Pandas — програмна бібліотека, написана для мови програмування Python для маніпулювання даними та їхнього аналізу. Вона, зокрема, пропонує структури даних та операції для маніпулювання чисельними таблицями та часовими рядами. pandas є вільним програмним забезпеченням, що випускається за трипунктовою ліцензією BSD. Ця назва походить від терміну «панельні дані», який в економетрії позначає багатовимірні структуровані набори даних.

Можливості бібліотеки

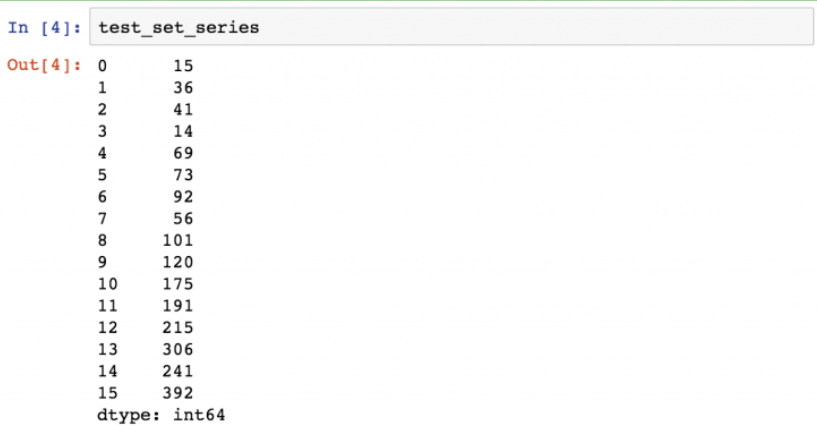
* Об'єкт DataFrame із вбудованим індексуванням для маніпулювання даними.
* Інструменти для зчитування та записування даних між структурами даних у пам'яті та різними форматами файлів.
* Вирівнювання даних та вбудована підтримка пропущених даних.
* Переформатовування для отримання зведених наборів даних.
* Отримання зрізів за мітками, індексування з розширеними можливостями та отримання піднаборів з великих наборів даних.
* Вставляння та вилучення стовпчиків у структурах даних.
* Рушій групування, що дозволяє робити з наборами даних операції розділення-зміни-об'єднання
* Злиття та з'єднання наборів даних.
* Ієрархічне індексування осей для роботи з даними високої вимірності в структурі даних нижчої вимірності.
* Функціональність для часових рядів: породження діапазонів дат та перетворення частоти, [статистики](https://uk.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)) рухливого вікна, лінійні регресії рухливого вікна, зсування дат та запізнювання.

Імпортувати pandas в Jupyter Notebook можна за допомогою такого рядка:

import pandas

У pandas є два види структур даних: Series і DataFrame

Series в pandas - це одномірна структура даних, яка зберігає дані. Для кожного значення в ній є унікальний індекс.

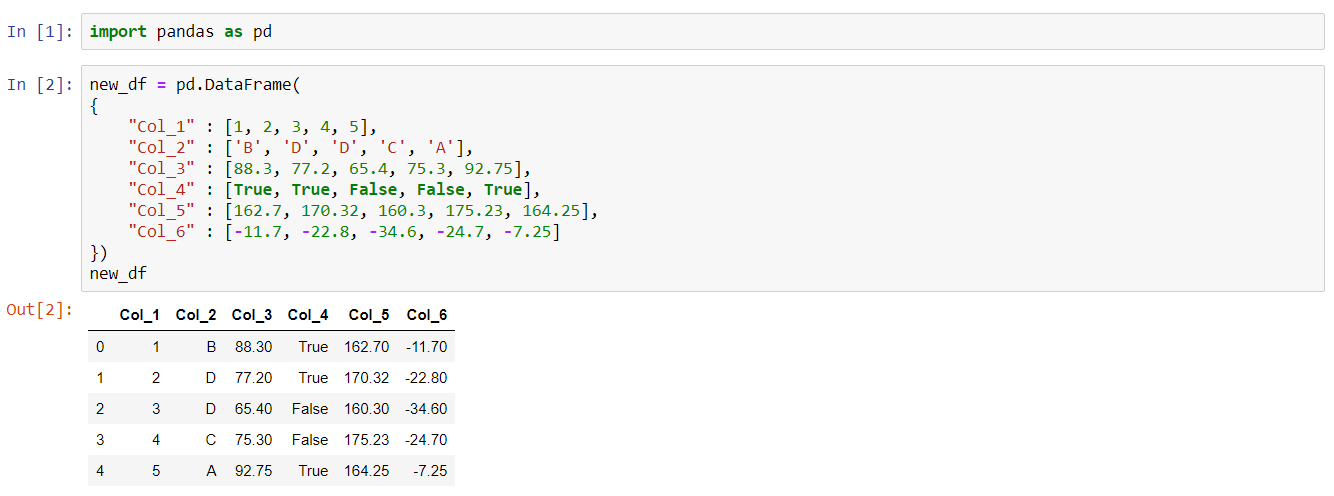


DataFrame - двохвимірна структура, що складається з колонок і рядків. У колонок є імена, а у рядків - індекси.

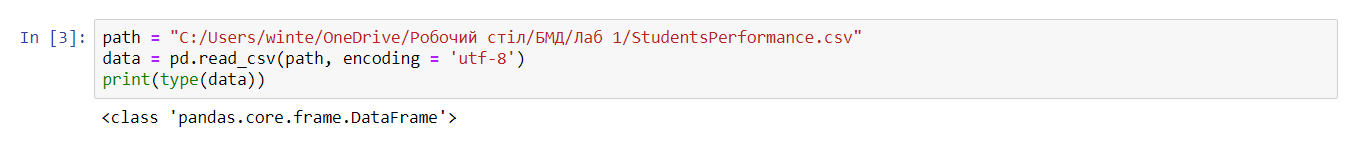
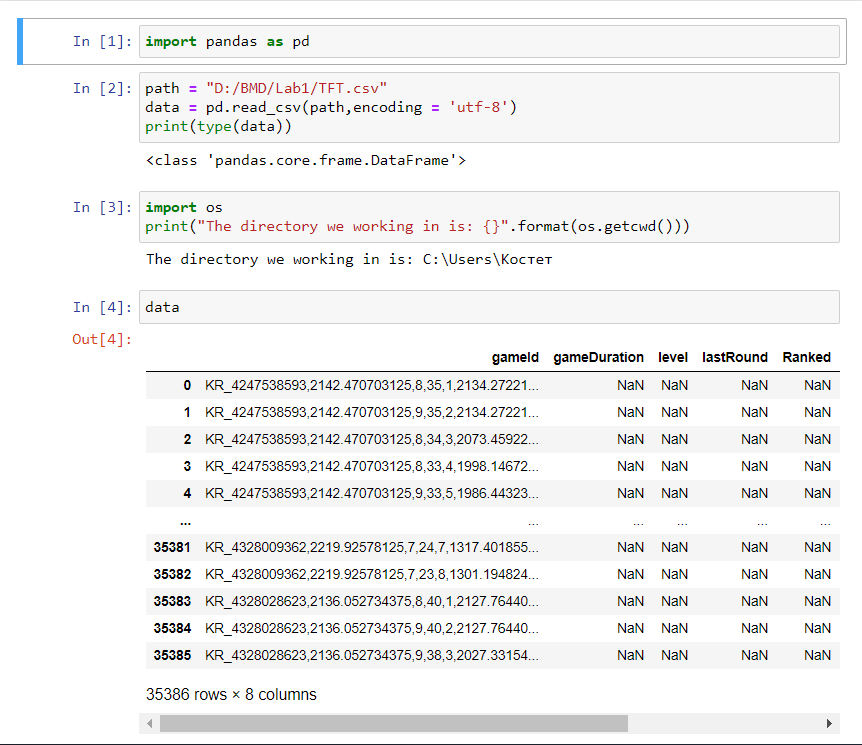


В загальному більший акцент роблять на DataFrames бо збільшою частиною аналітичних методів логічніше працювати в двохвимірній структурі.

* 1. Практична реалізація
     1. Введення даних вручну



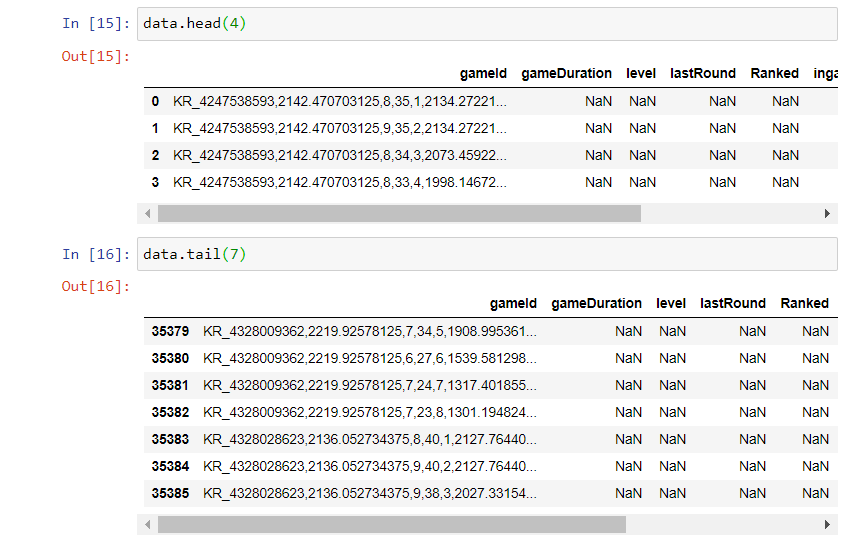
* + 1. Завантаження даних з CSV файлу
       1. Завантаження даних з вказанням повного шляху до файлу



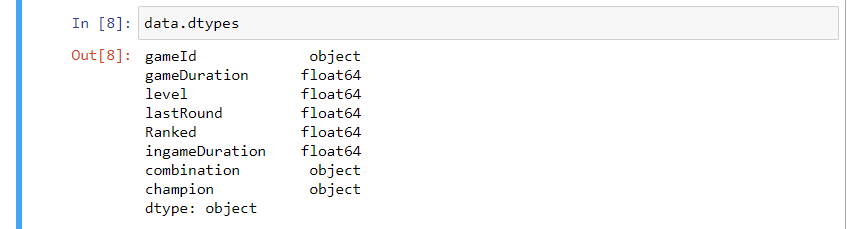
* + - 1. Для завантаження з робочої дерикторії достатньо ввести назву таблиці
    1. Базова статистика про завантажені дані



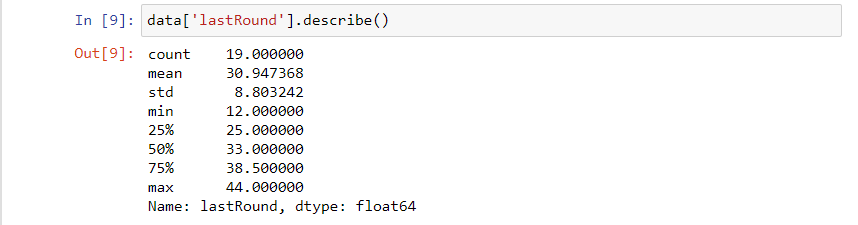
* + 1. Використання функцій head і tail



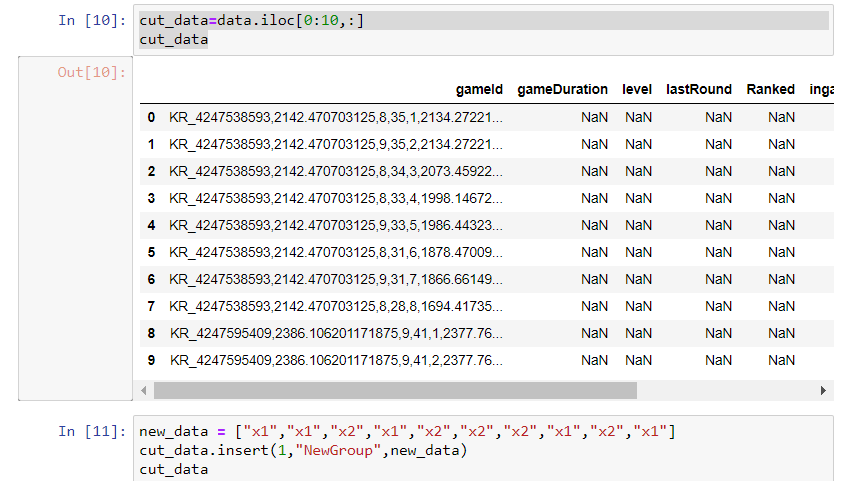
* + 1. Використання функції dtypes - для визначення типу даних в рядках



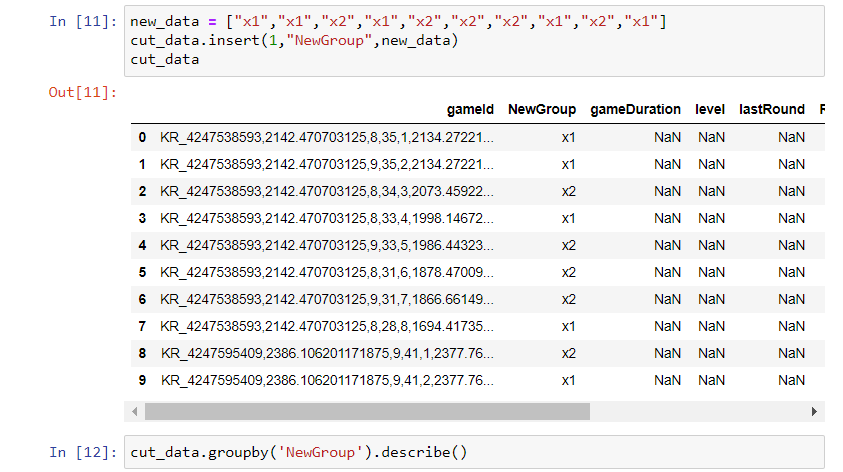
* + 1. Використання функції 'describe' для опису різних типів даних



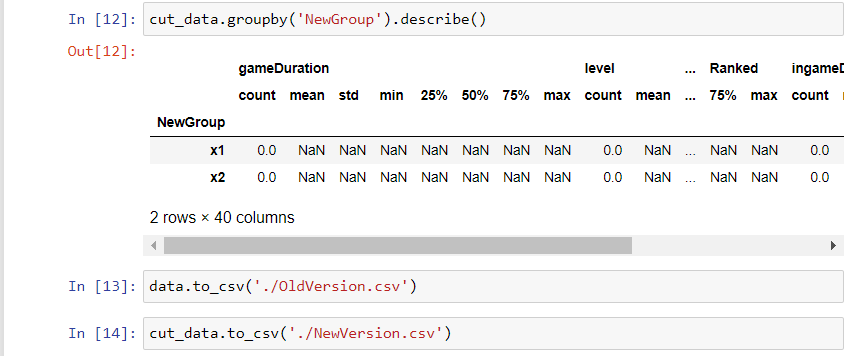
* + 1. Створення нового масиву даних, що містить лише перші 10 рядків з завантаженого масиву



* + 1. Додаємо новий стовпчик, та заповнюємо його даними, що відносять кожен з рядків до однієї з двох груп
    2. Виведення статистики з розділенням на групи



* + 1. Зберігаємо два файли даних
       1. Файл з обрізаним масивом даних, та з новою змінною
       2. Файл з повною статистикою по початковому масиву даних



* 1. Висновки

Jupyter Notebook – є зручним інструментом для створення гарних аналітичних звітів, так як він дозволяє зберігати разом код, зображення, коментарі, формули, графіки.

Бібліотека Pandas дозволяє працювати з табличними даними: фільтрувати, групувати дані, будувати зведені таблиці та будувати графіки – тому є дуже зручною у використанні.

**РОЗДІЛ 2. ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ БІБЛІОТЕКИ MATPLOTLIB**

* 1. Теоретичні відомості

Бібліотека matplotlib - це бібіліотека двовимірної графіки для мови програмування python за допомогою якої можна створювати високоякісні малюнки різних форматів.

Matplotlib поширюється на умовах BSD-подібної ліцензії. Генеруючі в різних форматах зображення можуть бути використані в інтерактивній графіці, в наукових публікаціях, графічному інтерфейсі користувача, веб-додатках, де потрібна побудова діаграм.

Для установки використовують пакетний менеджер:

conda install matplotlib

або

Імпортувати matplotlib в Jupyter Notebook можна за допомогою такого рядка:

import matplotlib

Примітка: до matplotlib можна звертатися за допомогою абревіатури для цього необхідно додати as /ім’я скорочення/. Якщо в кінці інструкції з import є as mpl, Jupyter Notebook розуміє, що в майбутньому, при введенні mpl мається на увазі саме бібліотека matplotlib.

Matplotlib підтримує багато видів графіків і діаграм:

* Графіки (line plot)
* Діаграми розкиду (scatter plot)
* Стовпчасті діаграми (bar chart) і гістограми (histogram)
* Кругові діаграми (pie chart)
* Стовбур-лист діаграми (stem plot)
* Контурні графіки (contour plot)
* Поля градієнтів (quiver)
* Спектральні діаграми (spectrogram)

Для побудови графіків потрібно підключити бібліотеку matplotlib.pyplot, додавши наступний рядок на початку програми:

import matplotlib.pyplot as plt

Стандартні функції:

*plt.plot (x, y, linewidth = 2.0)* – для побудови простих двовимірних графіків, х координати по горизонтальній осі, у – координати вертикальної осі, linewidth дозволяє визначити зовнішній вигляд маркера (точки) - в лапках прописується перша буква кольору, а далі зовнішній вигляд маркера (Наприклад: plt.plot([1,2,3],[1,2,3], 'r--') – червона перериваюча лінія)

*plt.ylabel('Назва вертикальної осі')*

*plt.xlabel('Назва горизонтальної осі')*

*plt.title('Назва графіка')*

*plt.axis([x\_min, x\_max, y\_min, y\_max])* - цією функцією можна обмежити область відображення графіка

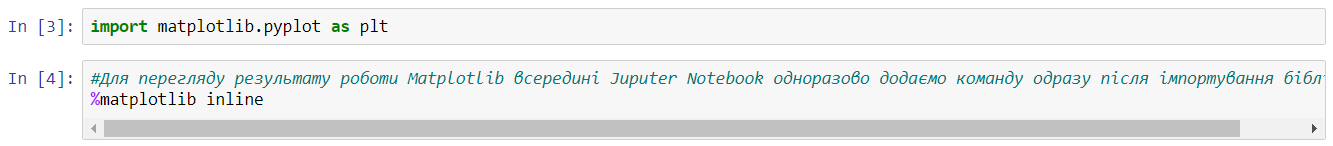
*plt.show()* – відображення графіка в окремому вікні

*plt.scatter()* – точковий графік (без ліній)

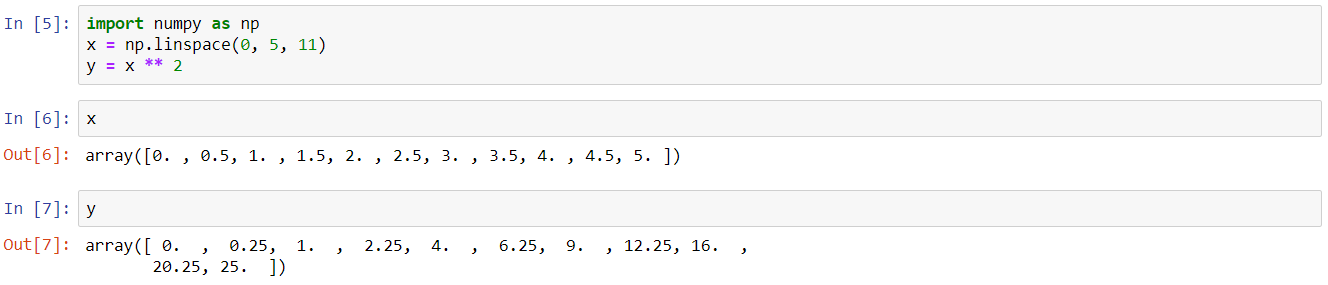
*plt.figure()* – для відображення декількох графіків поруч або один під одним, де в якості параметра прописується порядковий номер форми (вікна)

*plt.subplot(ijk)* – для гарного компонування графіків, i – кількість рядків, j- кількість стовпців, порядковий номер графіка в поточному вікні.

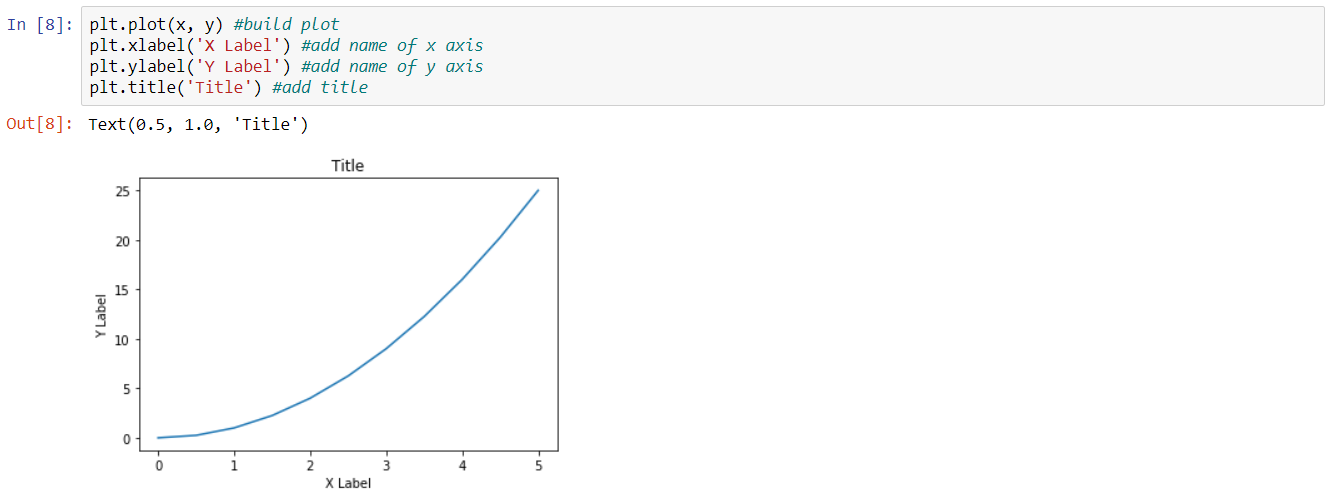
* 1. Практична реалізація
     1. Імпорт бібліотеки

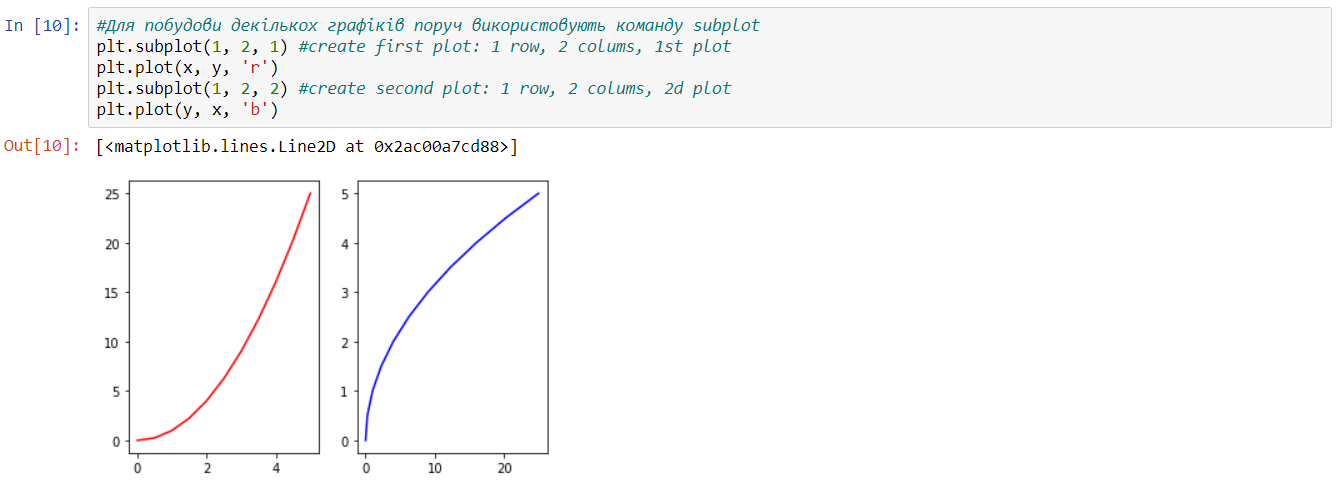


* + 1. Основи побудови графіків

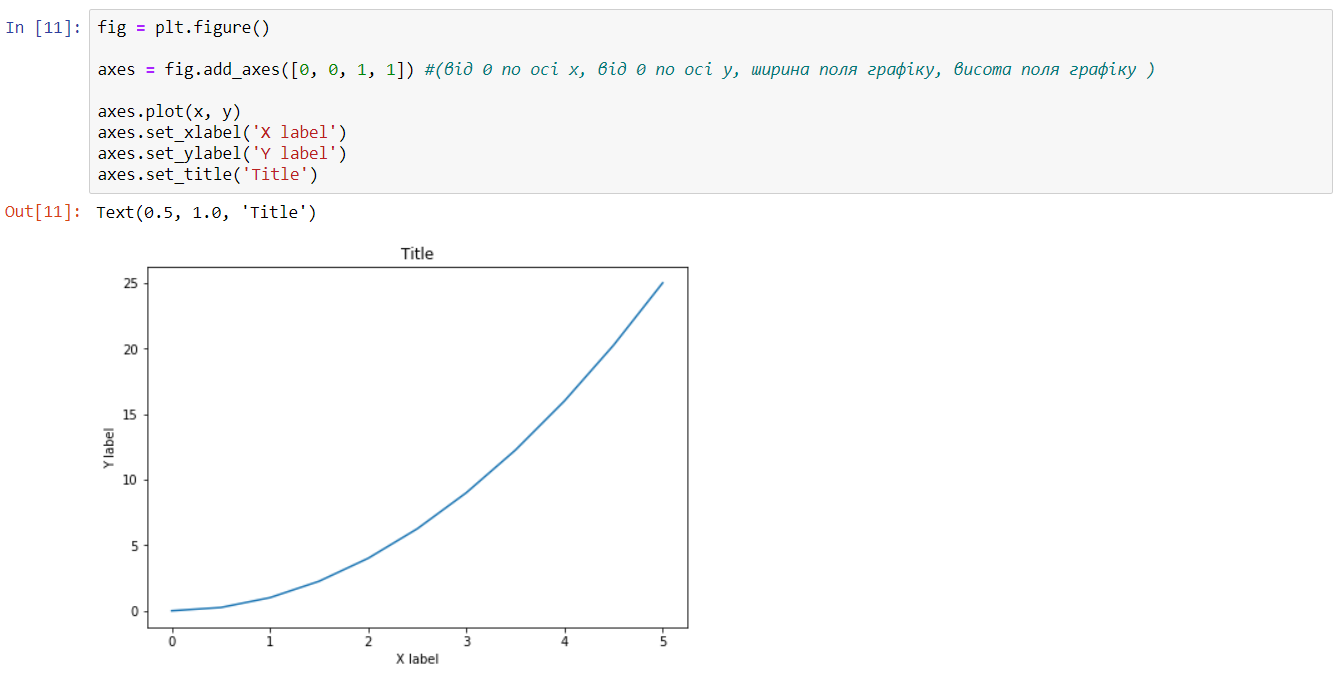


* + 1. Побудова графіків
       1. Функціональний метод

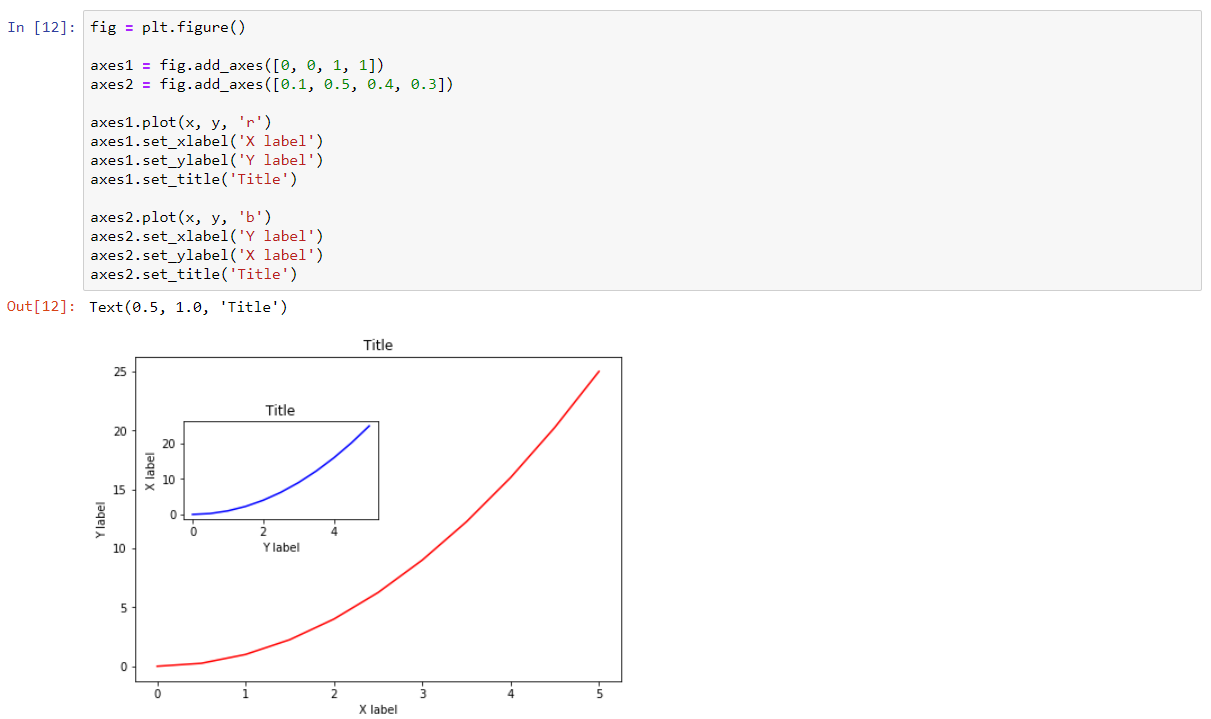




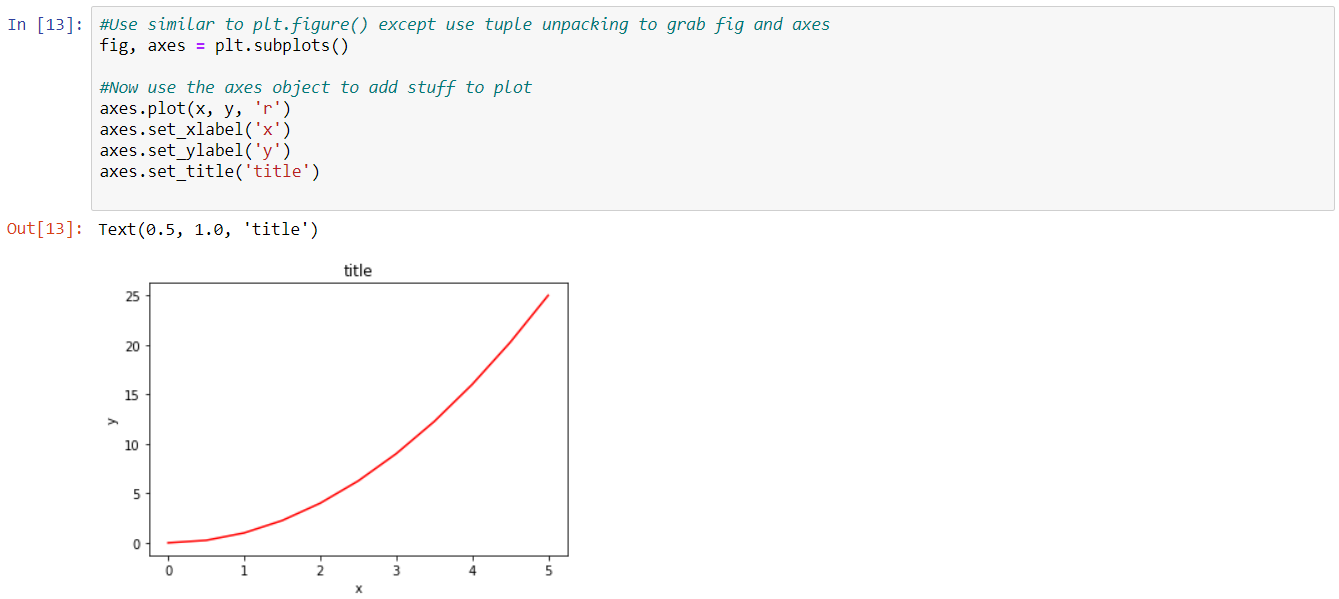
* + - 1. Об'єктно-орієнтований метод

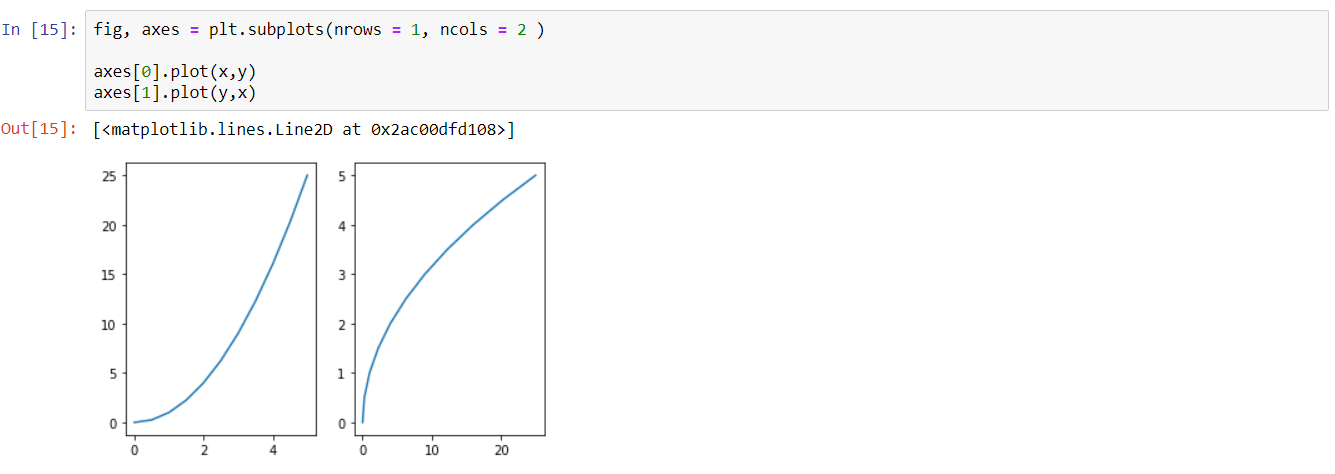


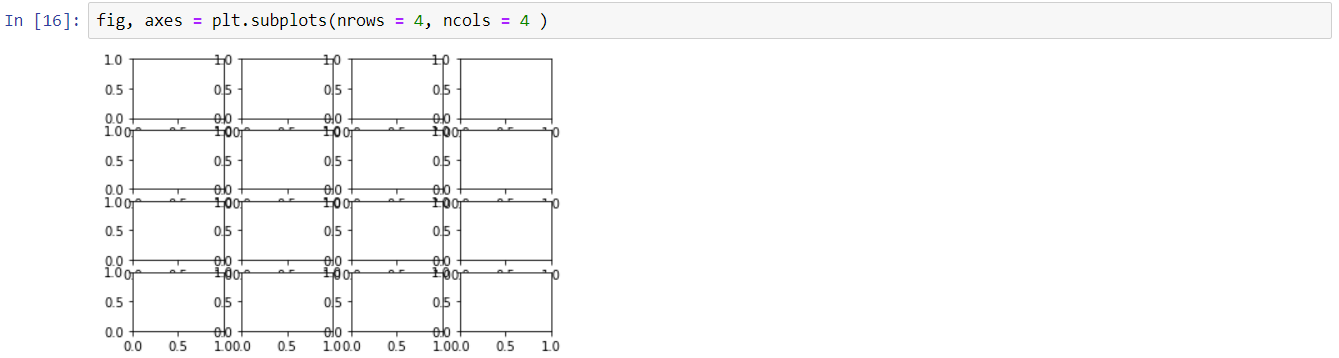
* + 1. Одночасна побудова декількох графіків

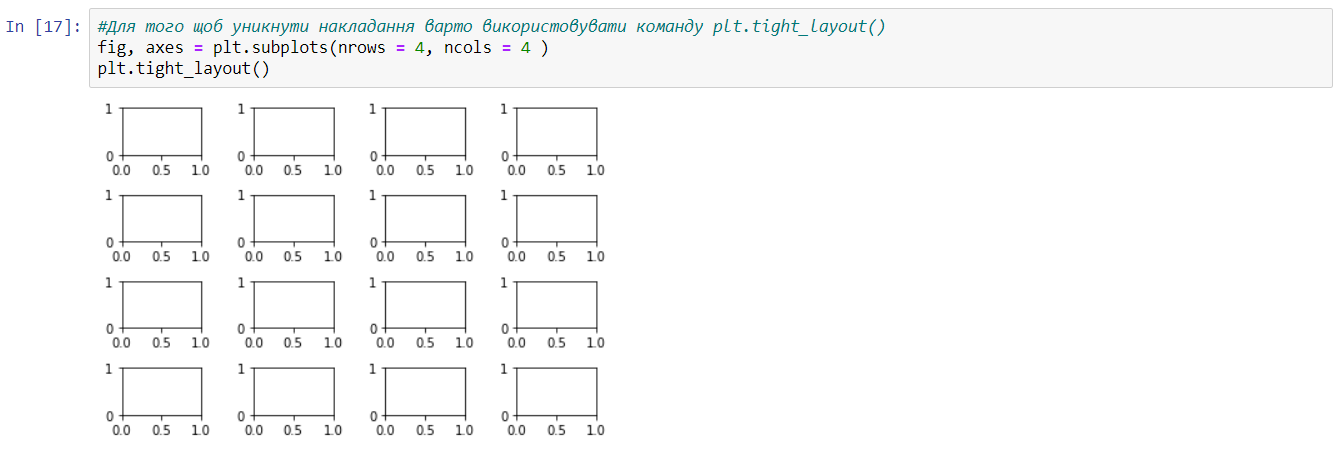


* + 1. Використання іншого функціоналу бібліотеки

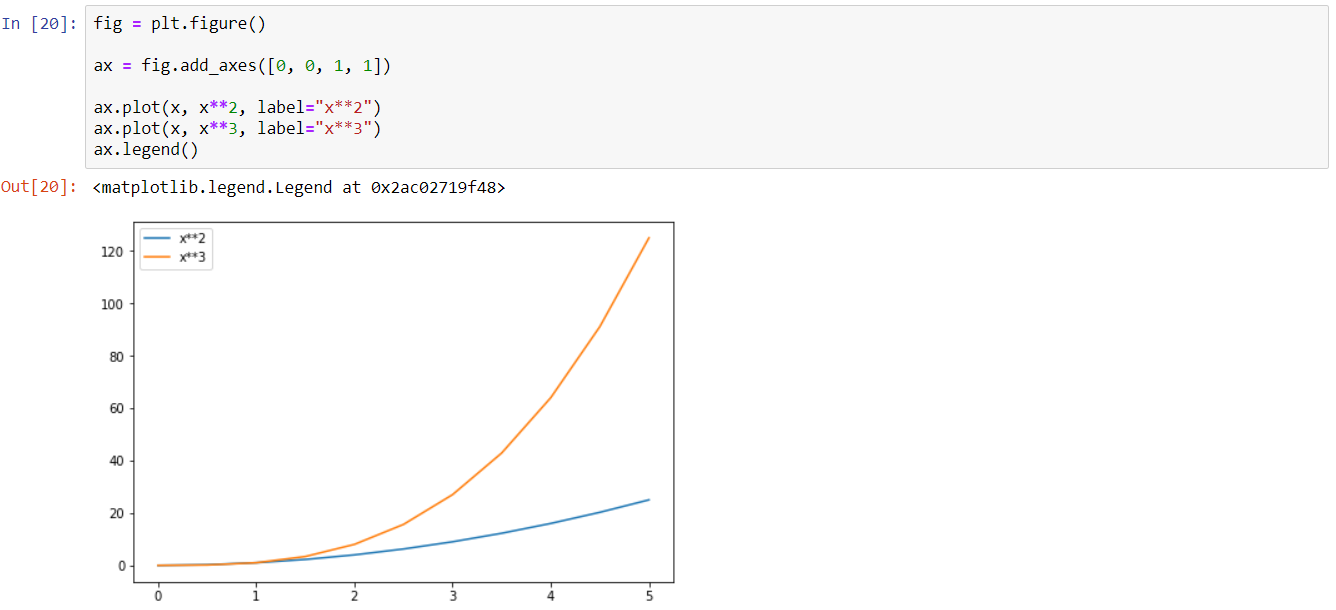




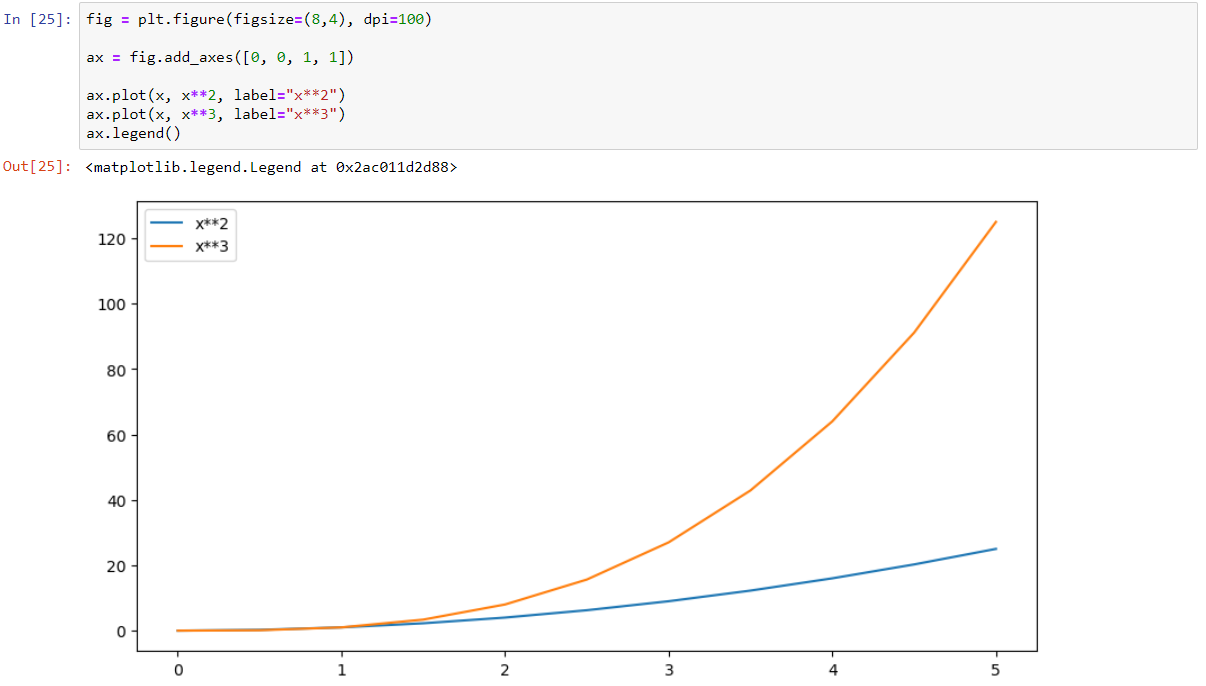




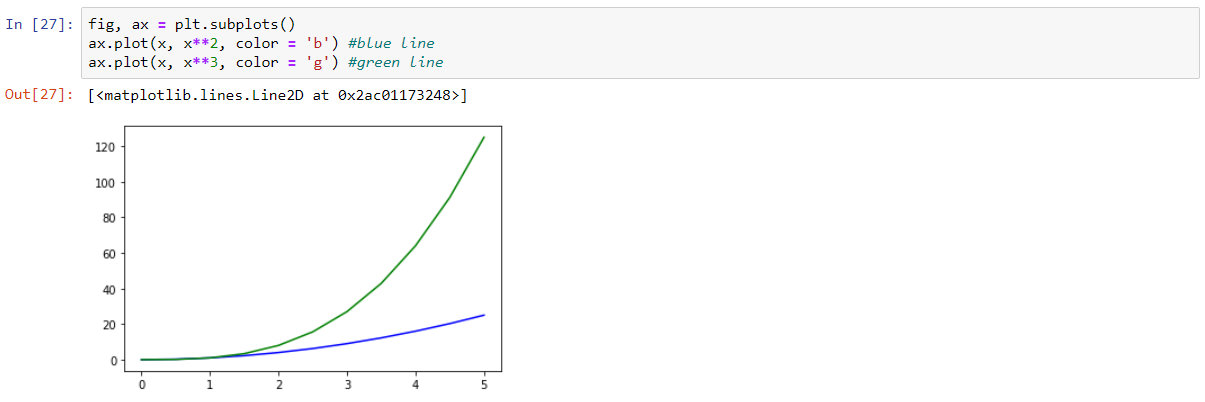
* + 1. Додавання легенди до графіку з використанням команди .legend()



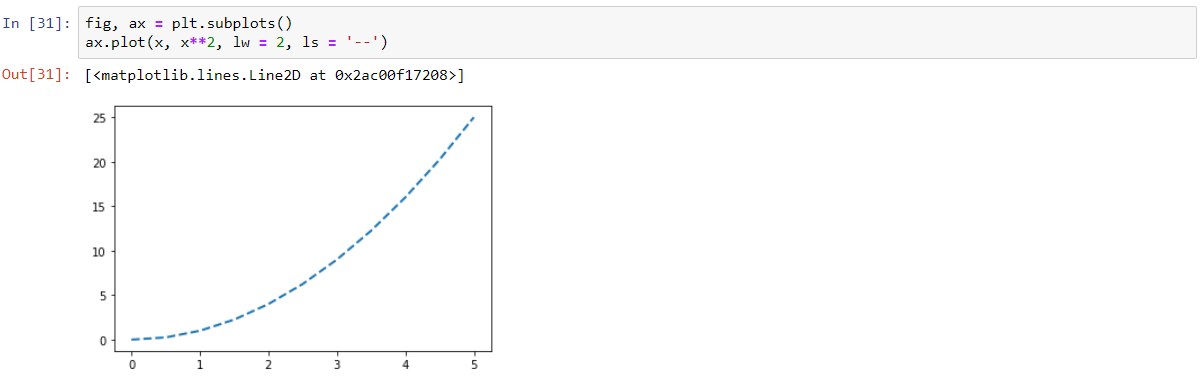
* + 1. Змінення розміру графіка та DPI



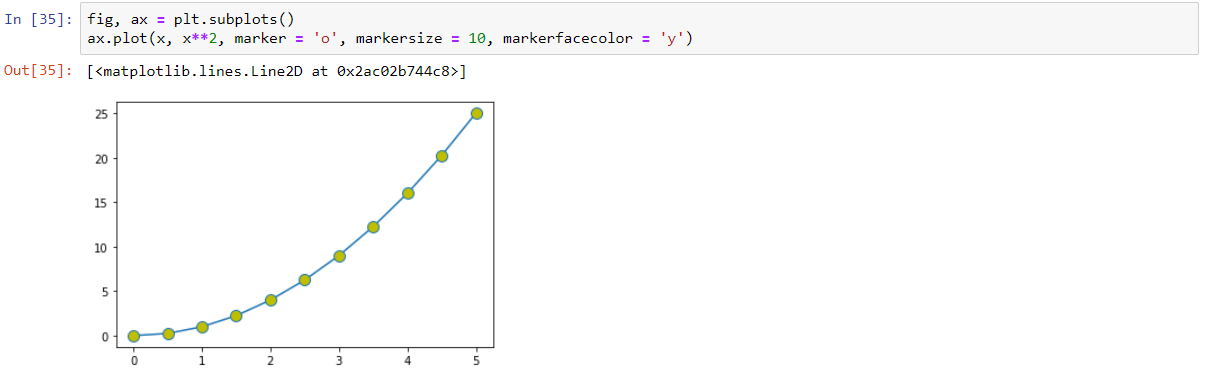
* + 1. Надання характеристик графіку
       1. Колір графіку



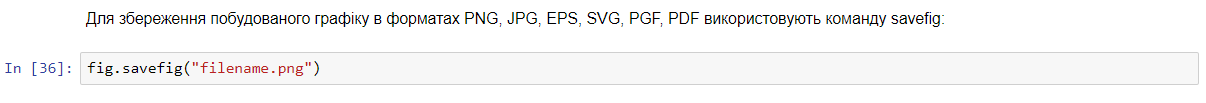
* + - 1. Тип та ширина лінії



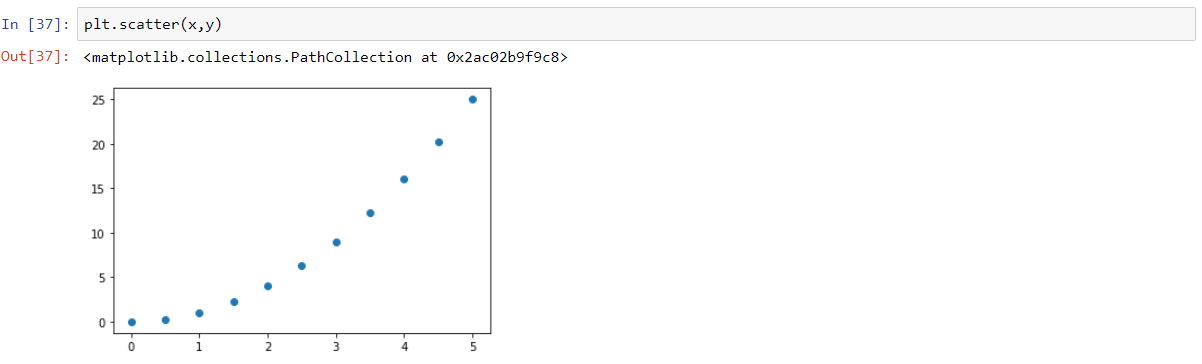
* + - 1. Тип, розмір та колір маркерів



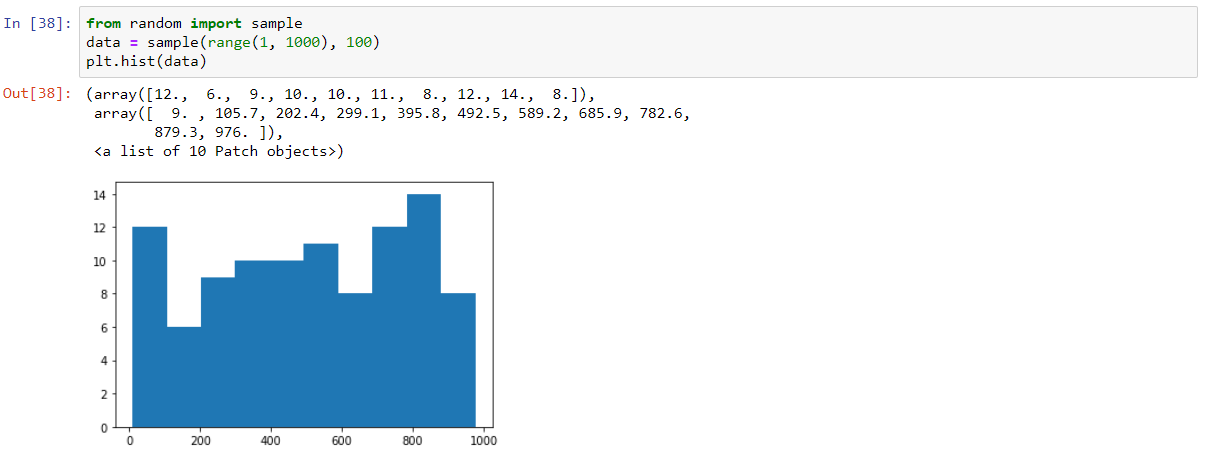
* + 1. Збереження побудованого графіку



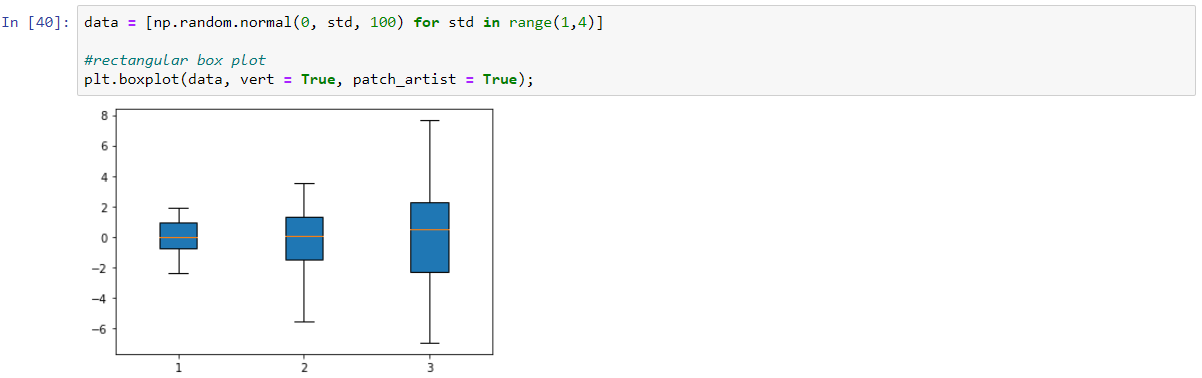
* + 1. Спеціальні графіки в Matplotlib
       1. Точкова діаграма



* + - 1. Гістограма



* + - 1. Box-plot діаграма



* 1. Висновки

Бібліотека Matplotlib - є одиним з найпопулярніших пакетів Python, який використовується для візуалізації даних. Це кросплатформена бібліотека для створення 2D та 3D графіків з даних в массивах. Ця бібліотека дає можливість будувати графіки двома методами: функціональним і об'єктно-орієнтованим. Стандартні функції бібліотеки matplotlib – дають можливість для підпису осей графіка, надання йому назви, додавання легенд, також дають можливість розміщувати графіки у вікні так як того бажає користувач – тому ця бібліотека є доволі зручною у використанні.

**РОЗДІЛ 3. ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ БІБЛІОТЕКИ SEABORN**

* 1. Теоретичні відомості

Seaborn - це більш високорівневе API на базі бібліотеки matplotlib. За допомогою Seaborn зручніше оформлювати графіки. Якщо просто додати в код import seaborn, то картинки стануть набагато симпатичніші. Також в бібліотеці є досить складні типи візуалізації, які в matplotlib зажадали б великого кількість коду.

Для візуалізації розподілу метричних змінних використовуються наступні типи графіків:

* distplot
* jointplot
* rugplot

Можна таже візуалізувати відносно розподілу між парами змінних за допомогою методів:

* PairGrid
* pairplot
* FaceGrid

*distplot* - одночасно показує гістограму і графік щільності розподілу

*jointplot()* - показує спільний розподіл за двома змінним

*rugplot -* показує те, що і графік щільності розподілу, тільки в одновимірній формі. Чим щільніше розташовані лінії, тим вище щільність. Краще використовувати його спільно з іншими видами графіків.

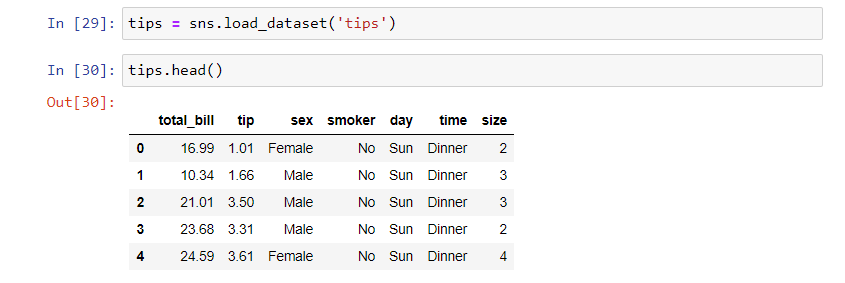
*pairplot* - показує відносини між усіма парами змінних (це спрощена версія іншої функції, яка називається *PairGrid*)

*FaceGrid* - дозволяє візуалізувати спільний розподіл окремих ознак деяких змінних

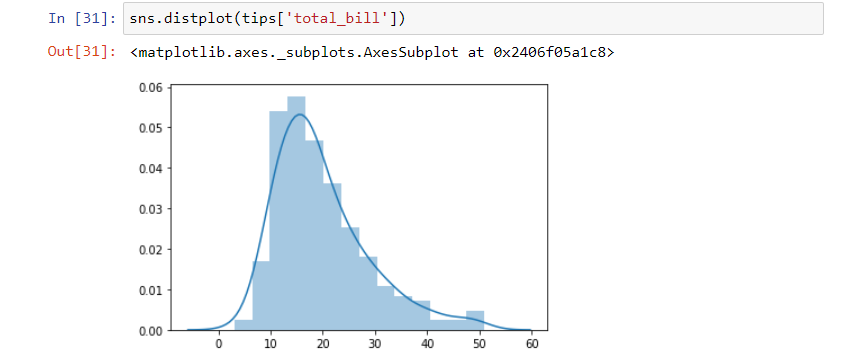
* 1. Практична реалізація
     1. Імпортування бібліотеки

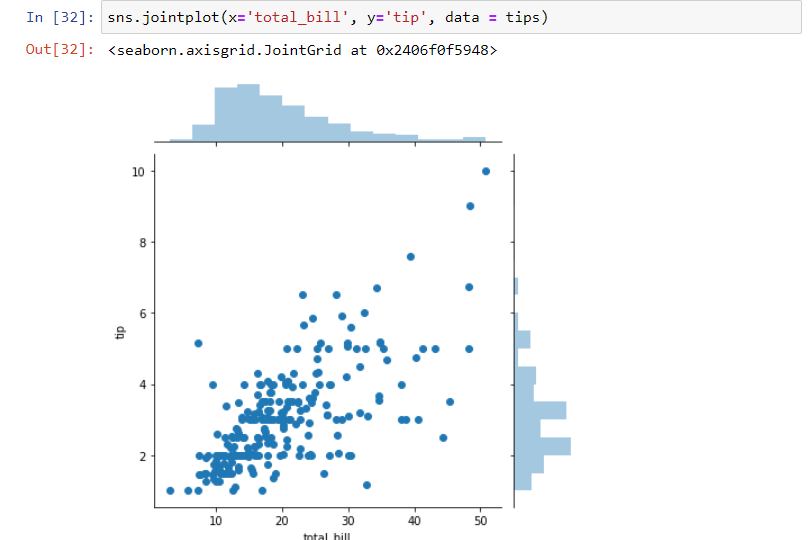


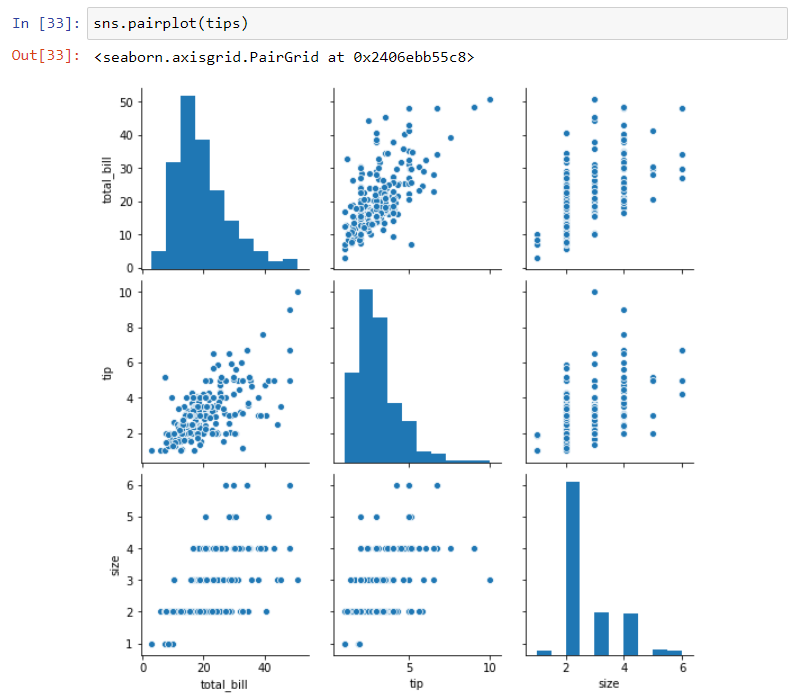
* + 1. Збереження даних

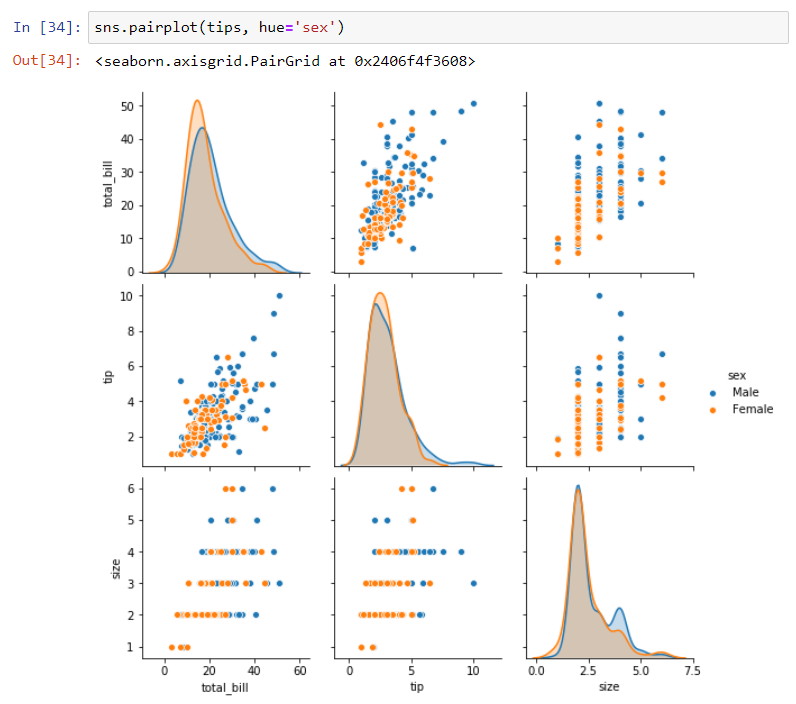


* + 1. Візуалізація даних

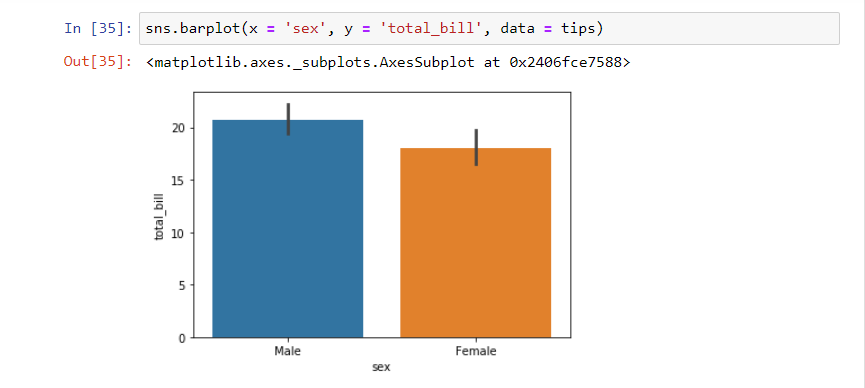




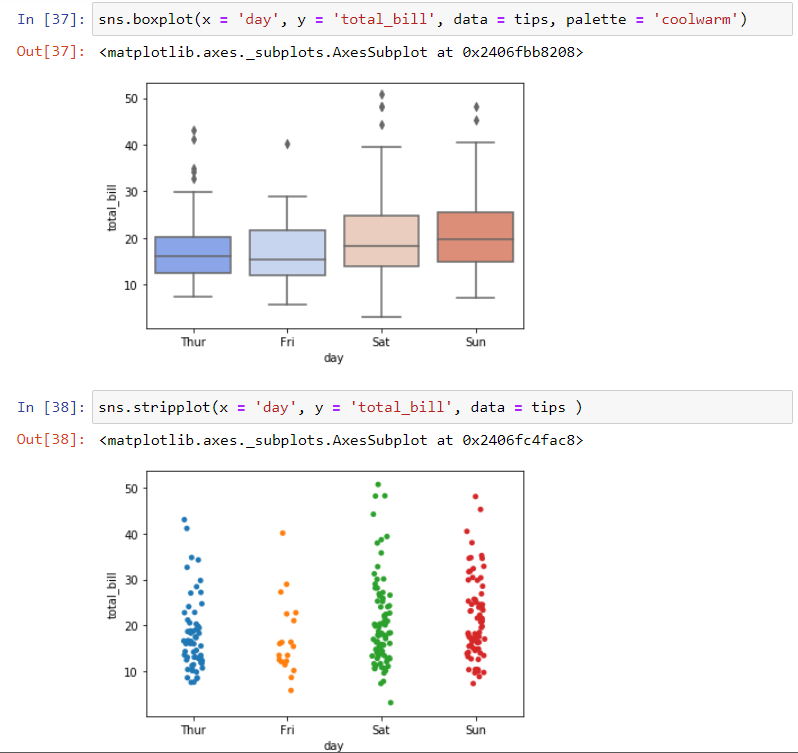




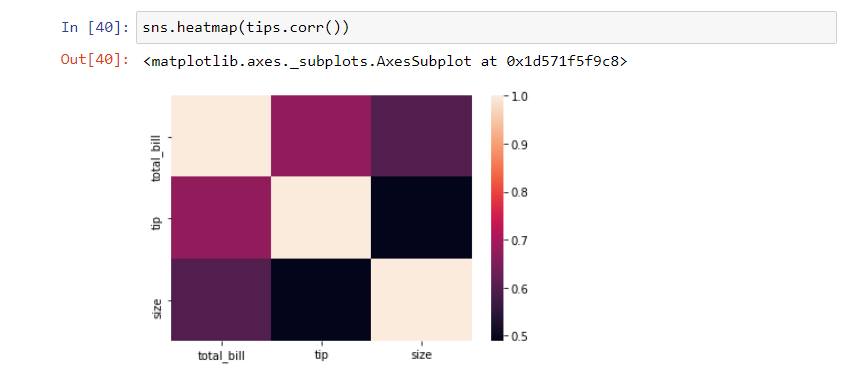
* + 1. Графіки для аналізу категоріальних даних

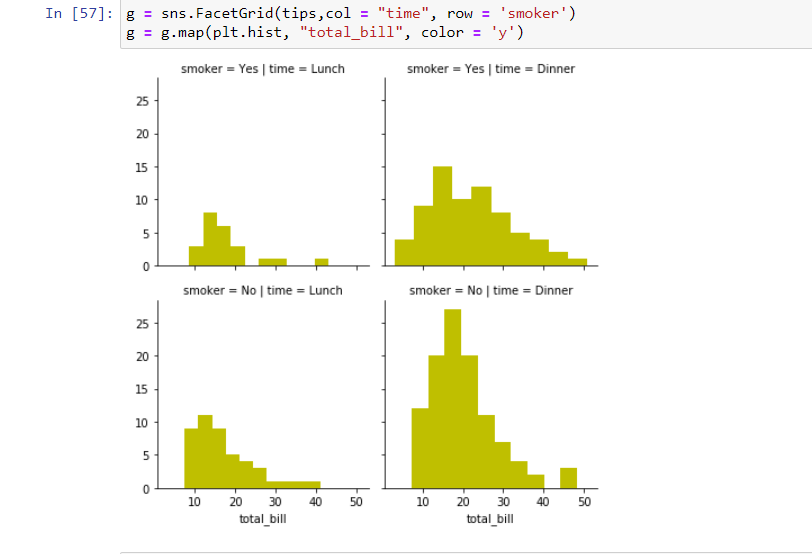






* + 1. Графіки для аналізу взаємозв'язку даних





* 1. Висновки

Бібліотека Seaborn надає високорівневий інтерфейс для побудови графіків для аналізу зв’язку даних, діаграм розсіювання. Містить гарні стилі за замовчуванням.

**РОЗДІЛ 4. ПЕРЕВІРКА ДАНИХ НА НОРМАЛЬНІСТЬ ЗА ДОПОМОГОЮ КРИТЕРІЮ ШАПІРА-УІЛКА**

* 1. Теоретичні відомості

Критерій Шапіро-Уілка використовується для перевірки гіпотез:

: випадкова величина X розподілена нормально

:: випадкова величина X розподілена не нормально

Критерій Шапіро-Уілкі заснований на оптимальній лінійно незміщеній оцінці дисперсії до її звичайної оцінки методом максимальної правдоподібності. Статистика критерію має вигляд:

де

Чисельник є квадратом оцінки середньоквадратичного відхилення Ллойда

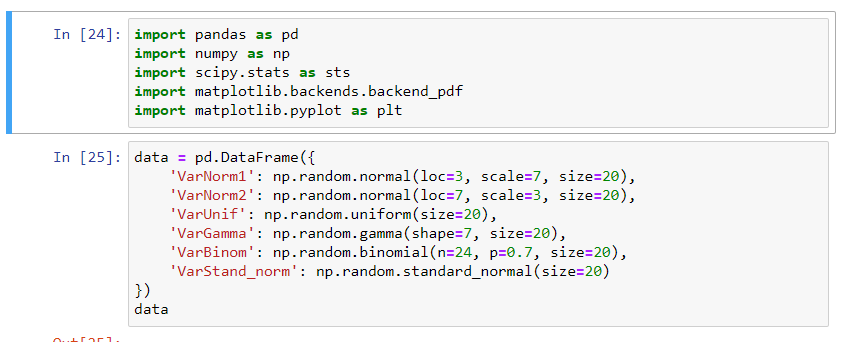
Коефіцієнти беруться з таблиць.

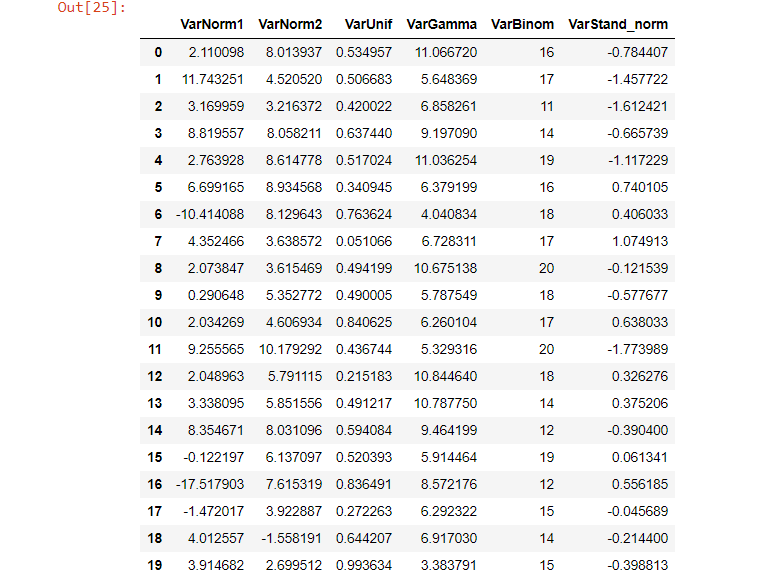
Критичне значення також знаходяться таблично

Якщо W < то про нормальність розподілу відхиляється і приймається

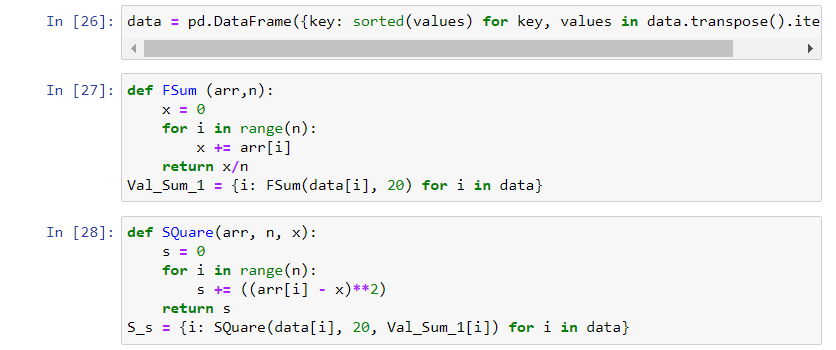
Критерій Шапіро-Уілкі є дуже потужним критерієм для перевірки нормальності, але, на жаль, має обмежену придатність. При великих значеннях n (n>100) таблиці коефіцієнтів стають незручними.

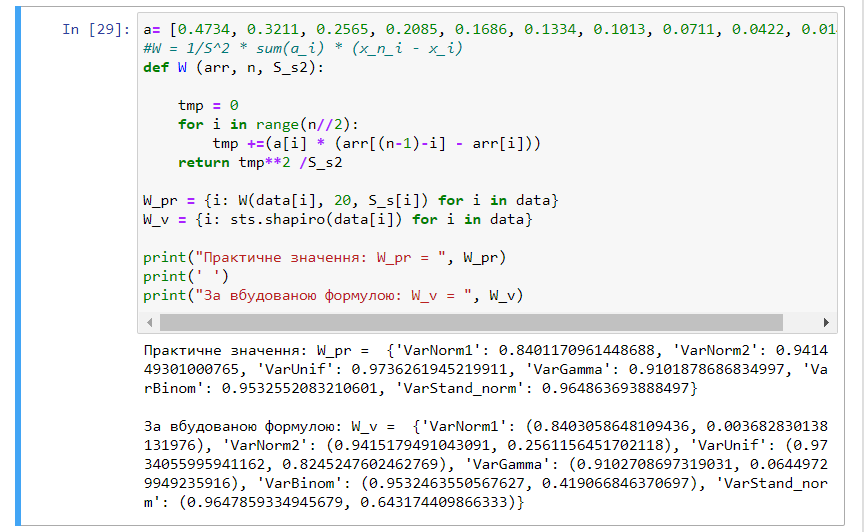
* 1. Практична реалізація
     1. Генерування набору даних, що містить не менше 6 змінних. Частину змінних згенеровано за нормальним розподілом, частину за будь-яким іншим розподілом. Кількість спостережень – 20



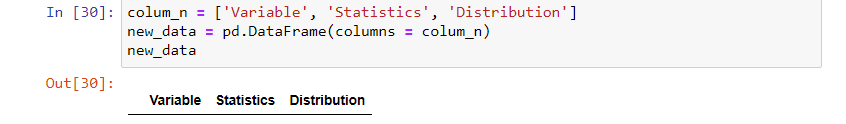


* + 1. Створення функції, що реалізує критерій Шапіро-Уілка та застосування його до кожної змінної набору даних

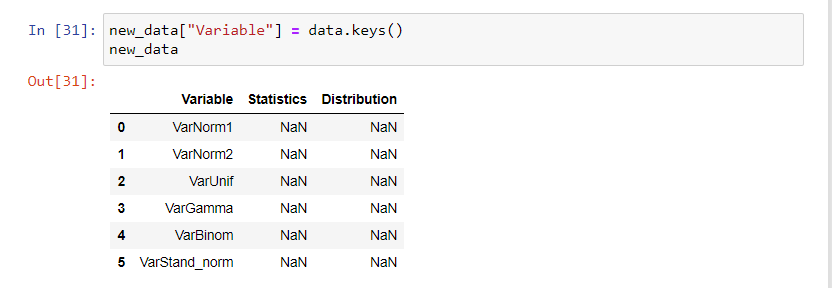




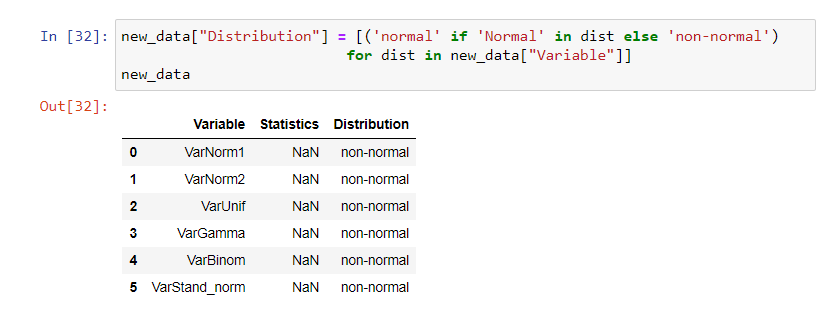
* + 1. Генерування пустої таблиці в якій кількість рядків – це кількість змінних+1 Кількість стопчиків – 3. Назва стопчиків – Variable, Statistics, Distribution.



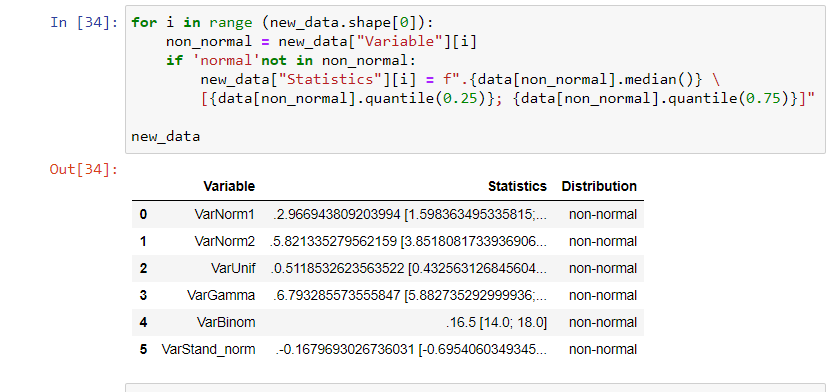
* + 1. Занесення в стопчик Variable назви змінних



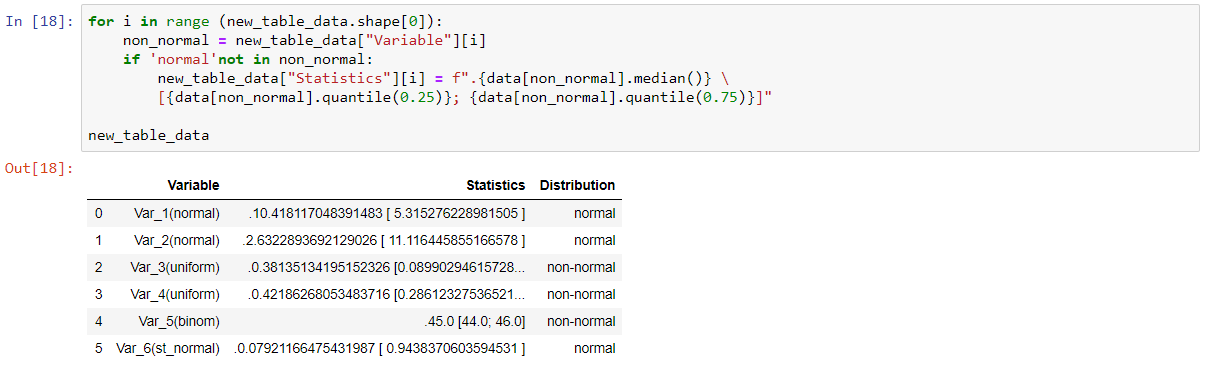
* + 1. Занесення в стопчик Distribution результатів роботи критерію Шапіро-Уілка. Для змінних, які розподілені за нормальним розподілом вносимо “normal”, в протилежному вигляді - “non-normal”



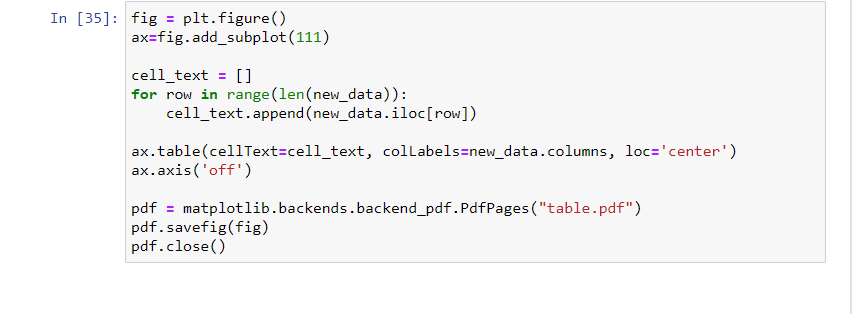
* + 1. Розрахувати для всіх змінних, що розподілені за нормальним законом середнього значення та стандартне відхилення



* + 1. Розрахувати для всіх змінних, що розподілені за розподілом відсутнім від нормального розрахувати медіану та інтерквартильний розмах



* + 1. Зберегти таблицю в PDF форматі



* 1. Висновки

Користуючись критерієм Шапіра-Уілка, можна визначати чи має вибірка нормальний розподіл. Аналізуючи результати даного комп’ютерного практикуму можна помітити, що за допомогою критерія Шапіро-Уілка можна безпомилково визначити. які дані маю нормальний розподіл, а отже практика підтверджує теорію.

**РОЗДІЛ 5. КРИТЕРІЙ ПІРСОНА ДЛЯ ПОРІВНЯННЯ ДЕКІЛЬКОХ ГРУП ЗА РОЗПОДІЛЕННЯМ ОЗНАКИ**

* 1. Теоретичні відомості

Хі-квадрат Пірсона один з найпопулярніших статистичних критеріїв для аналізу якісних даних (номінальних, порядкових, рангових), аналізу частот.

Хі-квадрат використовується перш за все для аналізу таблиць спряженості (вид таблиці, яка враховує спільне вплив фактора на результат, дані в таблиці спряженості повинні бути представлені у вигляді частоти номінальних даних або інтервалами, але не безперервними кількісними величинами). Варто відзначити, що при роботі з сполученими таблицями хі-квадрат часто є підтримкою для аналізу впливу факторів ризику за допомогою розрахунку ризиків (абсолютний і відносний ризики) і відношення шансів. Таблиці пов'язаності можуть приймати різні форми.

Умови застосування статистичного критерію хі-квадрата Пірсона:

1. Тип даних: параметри повинні бути якісними цілочисельних частотами, вимірюваними в номінальній шкалі
2. Бажано, щоб загальна кількість спостережень була більше 20
3. Очікувана частота, відповідна нульовій гіпотезі повинна бути більше 5, якщо очікуване явище приймає значення менше 5, то необхідно використовувати точний Критерій Фішера.
4. Порівнянні частоти повинні бути приблизно одного розміру
5. Зіставляючі групи повинні бути незалежними

Забороняється: використовувати хі-квадрат для аналізу безперервних абсолютних даних, відсотків.

Статистичні гіпотези:

: Ознаки не пов’язані

: Ознаки пов’язані

Алгоритм розрахунку:

1. Побудувати таблицю спряженості

|  |  |  |
| --- | --- | --- |
| Ознака | Група 1 | Група 2 |
| Градація1 | n1 | p1 |
| Градація 2 | n2 | p2 |

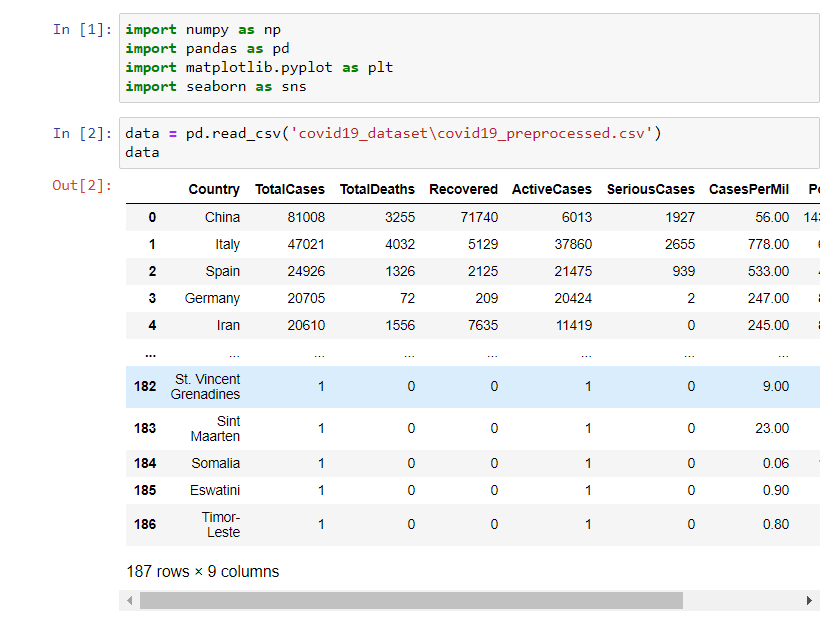
1. Перевіряємо чи n1+n2 = p1+p2, якщо суми не рівні необхідно їх вирівняти
2. Шукаємо різницю між емпіричними і теоретичними частотами для кожної градації

|  |  |  |  |
| --- | --- | --- | --- |
| Ознака | Група 1 | Група 2 |  |
| Градація1 | n1 | p1 |  |
| Градація 2 | n2 | p2 |  |

1. Розділяємо отриману різницю на теоретичні частоти

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ознака | Група 1 | Група 2 |  |  |
| Градація1 | n1 | p1 |  |  |
| Градація 2 | n2 | p2 |  |  |

1. Знаходимо суму отриманих значень та позначаємо її як
2. Визначаємо ступінь критерію за формулою r = m – 1 , де m – це кількість градацій ознаки (рядків в таблиці спряженості). За допомогою ступня критерію знаходимо за таблицею
3. Якщо то розбіжністі між розподіленнями між розподіленями статистично значимі на даному рівні значимості, тобто приймаємо альтернативну гіпотезу , в протилежному випадку приймаємо (ститистично незначимі).
   1. Практична реалізація
      1. Підключення данних

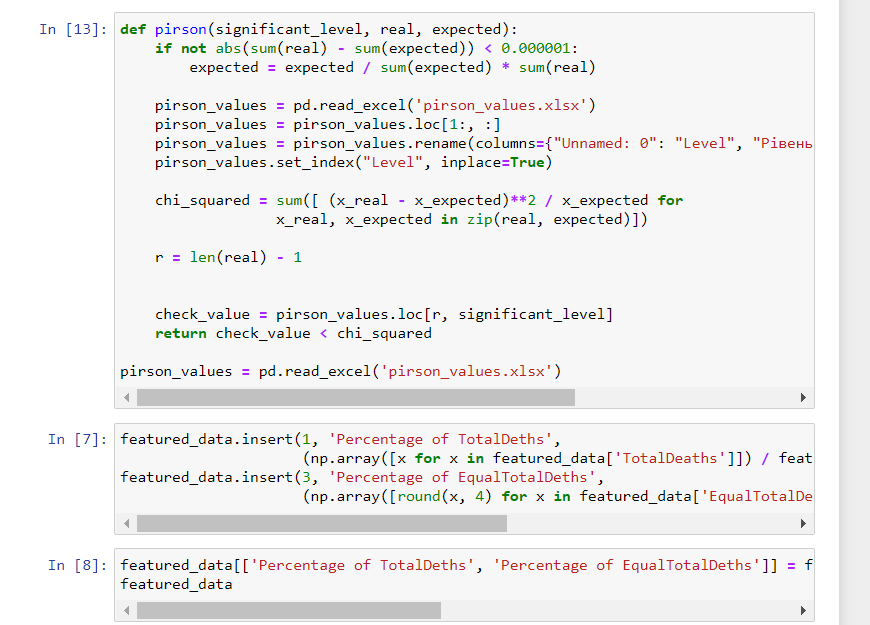


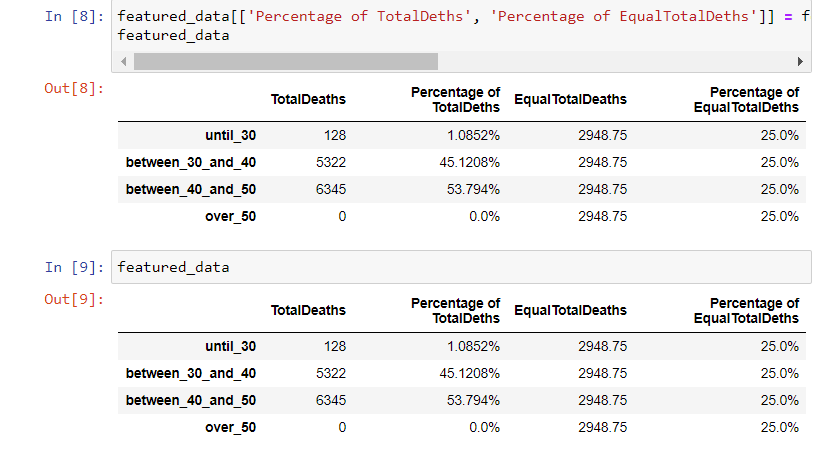


* + 1. Побудова графіка, що відображає розподіл ознак в групах

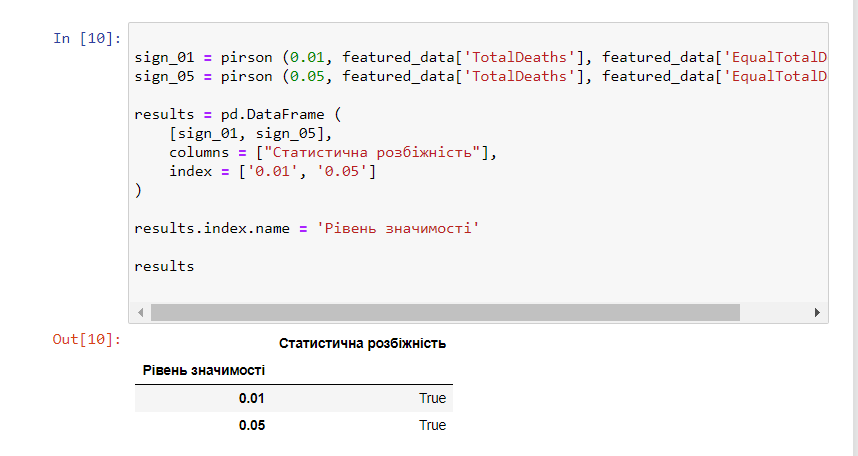


* + 1. Функція, що оцінює зв’язок між ознакою та групою за критерієм Пірсона. Функція оцінює розбіжність на двох рівнях 0.01 та 0.05 в залежності від параметру significsnt\_level, що передається в неї. Врахований випадок, що сума частот ознаки у групах може бути відмінною





* + 1. Додавання до функції стопчика з відсотковим розподілом по градаціям всередині груп. Окремо виведені результати роботи критерію (чи є статистична розбіжність між групами)



* + 1. Висновки

Під час порівняння декількох груп за розподіленням ознаки за допомогою Критерію Пірсона, було отримано результат, що статична розбіжність між групами на рівнях значимості 0.05 і 0.01 є статистично незначимою, так як в двох випадках χ^2 < χ^2(критичне)

**РОЗДІЛ 6. ПАРНИЙ T-КРИТЕРІЙ СТЬЮДЕНТА**

* 1. Теоретичні відомості

Критерій Стьюдента t відноситься до одного з найбільш давно розроблених і широко використовуваних методів статистики. Найчастіше він застосовується для перевірки нульової гіпотези про рівність середніх значень двох сукупностей, хоча існує також і одновиборочна модифікація цього методу.

Метод дозволяє перевіряти гіпотезу про те, що середнє значення двох генеральних сукупностей із яких витягнуті дві порівнювальні залежні вибірки, відрізняються один від одного. Припущення залежності частіше за все означає, що ознака виміряна на одній і тій же виборці двічі, наприклад до впливу і після нього.

Гіпотези:

: парні групи рівні

: парні групи відмінні

Обмеження:

* Розподіл ознаки в обох вибірках повинні суттєво не відрізняються від нормального розподілу
* Дисперсії вибірок рівні
* Ознака виміряна в метричній шкалі

Перед визначенням відмінності за парним критерієм, потрібно перевірити зв’язок між групами це можна перевірити за допомогою критерія Пірсона.

Формула для парного t-критерія Стьюдента:

*-* середнє значення різниці між значеннями

*–* стандартне відхилення різниці між значеннями

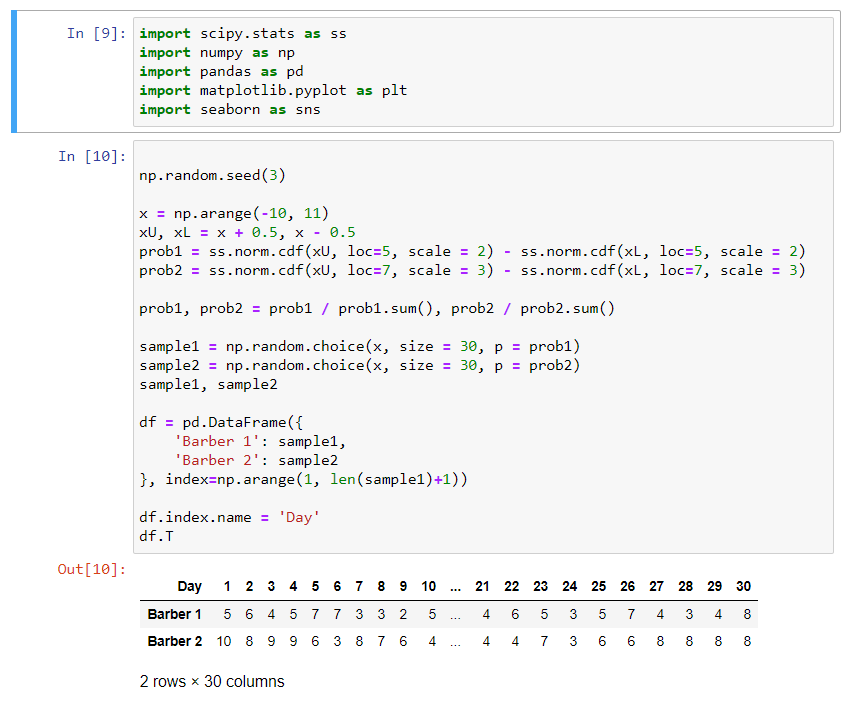
N – кількість досліджень в вибірці

df - число степенів свободи

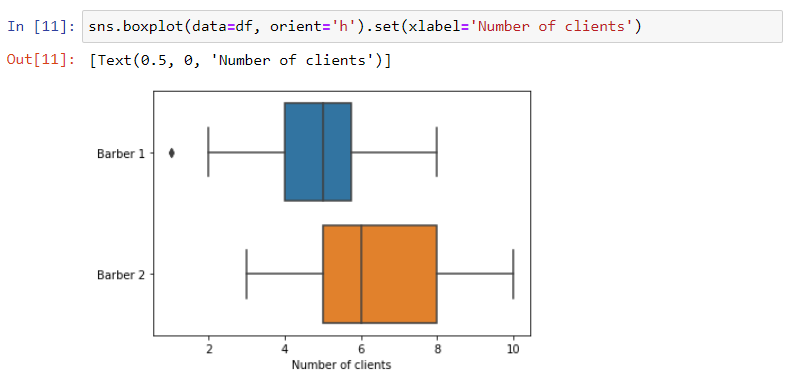
За допомогою df та таблиці критичних значень t-Стьюдента знаходимо

Якщо > приймаємо гіпотезу про відмінність середніх в протилежному випадку приймаємо – про рівність середніх

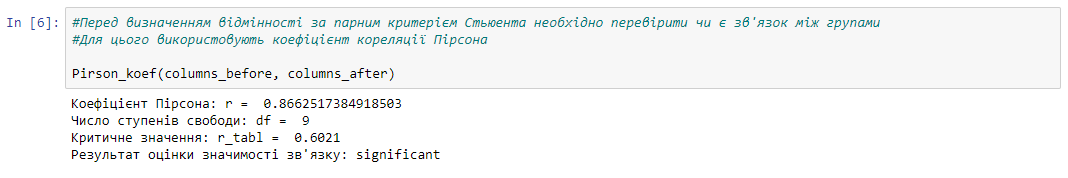
* 1. Практична реалізація
     1. Генерування випадкових даних. Дані відповідають вимогам



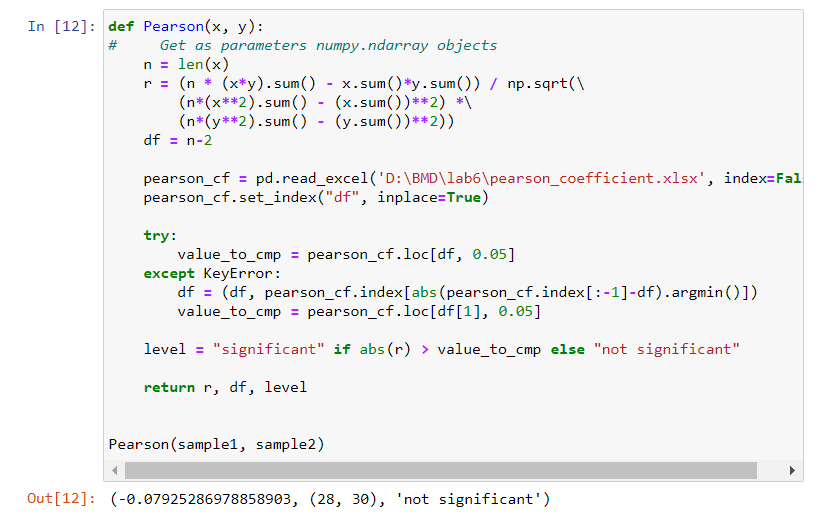
* + 1. Побудова графіка, що відображає розкид даних в групах



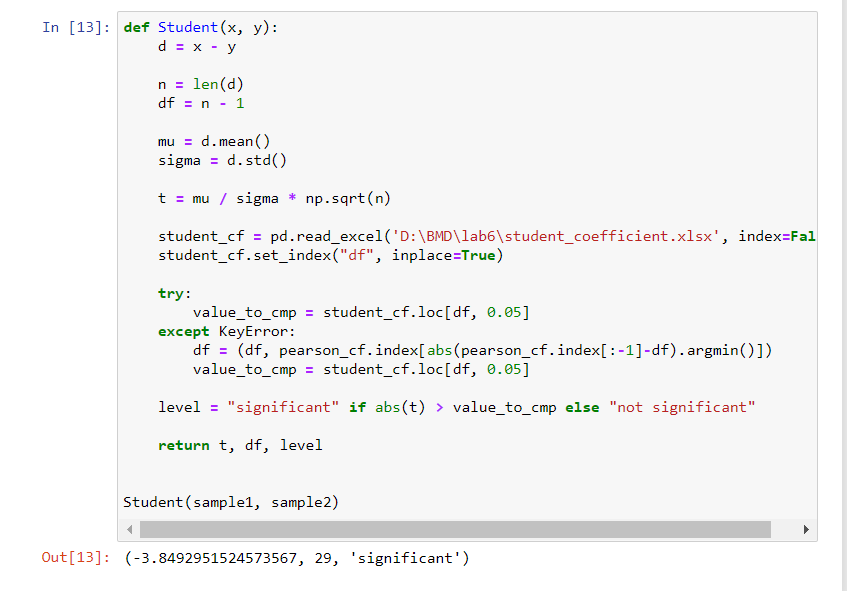
* + 1. Функція для оцінки зв’язку між групами за коефіцієнтом кореляції Пірсона. Функція повертає значення коефіцієнту кореляції, значення ступеня свободи та результат оцінки значимості зв’язку (significant або not significant)



* + 1. Функція для визначення відмінності між групами за критерієм Стьюдента. Функція повертає значення критерію Стьюдента, значення ступеня свободи та результат оцінки значимості зв’язку (significant або not significant)



Повернуто кортеж, що має на першому місці показник ступеня вільності, а на другому значення, що було взято з таблиці для отримання критичного значення.



* 1. Висновки

Аналізуючи швидкість розв'язування студентами логічного завдання до та після курсу можна зробити висновок що:

1. За коефіцієнтом кореляції Пірсона було доведено, що існує статистично значима залежность між групами. Так як коефіцієнт кореляції менший за табличний ( r\_tabl < r ), величина цього зв'язку: "висока" так як r = 0.87
2. Присутня наявність статистично значущих відмінностей швидкості вирішення логічної задачі студентами до та після курсу (так як t\_tabl < t\_st), звітси слідує, що була прийнята альтернативна гіпотеза про відмінність середніх

**РОЗДІЛ 7. T – КРИТЕРІЙ СТЬЮДЕНТА ДЛЯ НЕЗАЛЕЖНИХ ВИБІРОК**

* 1. Теоретичні відомості

t-критерій для двох незалежних вибірок перевіряє гіпотезу про рівність середніх в двох вибірках. Критерій застосовується, наприклад, якщо необхідно порівняти результати балів ЗНО в двох різних школах.

Гіпотези:

: середні двох незалежних груп рівні

: середні двох незалежних груп відмінні

Обмеження:

* Розподіл ознаки в обох вибірках повинні суттєво не відрізняються від нормального розподілу
* Дисперсії вибірок рівні
* Ознака вимірюється в метричній шкалі

Для перевірки рівності дисперсій варто застосувати тест Лівеля і в залежності від його результатів вибрати формулу t-Стьюдента для незалежних вибірок: якщо рівність дисперсій доведено (1), якщо не доведено (2).

Формула t –Стьюдента для незалежних вибірок

і – середнє значення в відповідних вибірках

і – стандартне відхилення в відповідних вибірках

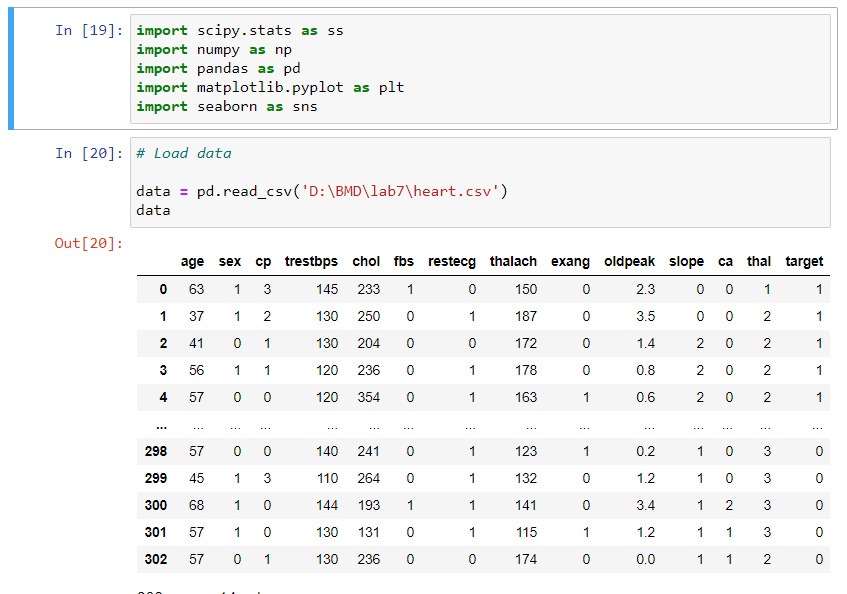
і – кількість досліджень в відповідних вибірках

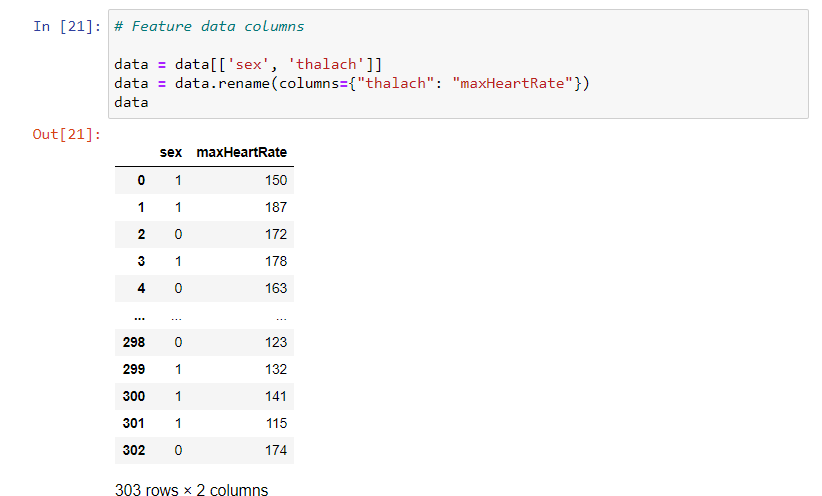
df – число степенів свободи

За допомогою числа ступенів свободи та таблиці критичних значень t-Стьюдента визначаємо критичне значення

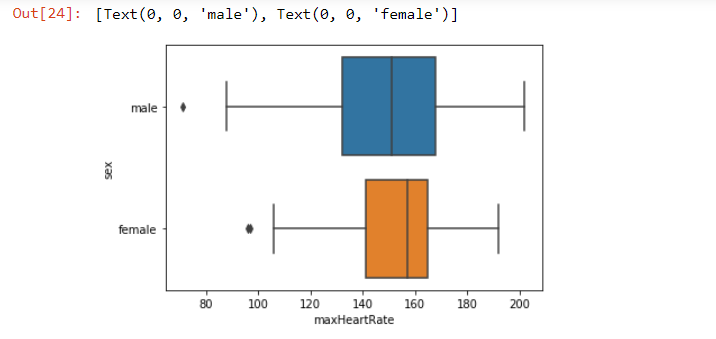
Якщо > - приймають та роблять висновок про статичні відмінність середніх, в протилежному випадку приймають – та роблять висновок про рівність середніх.

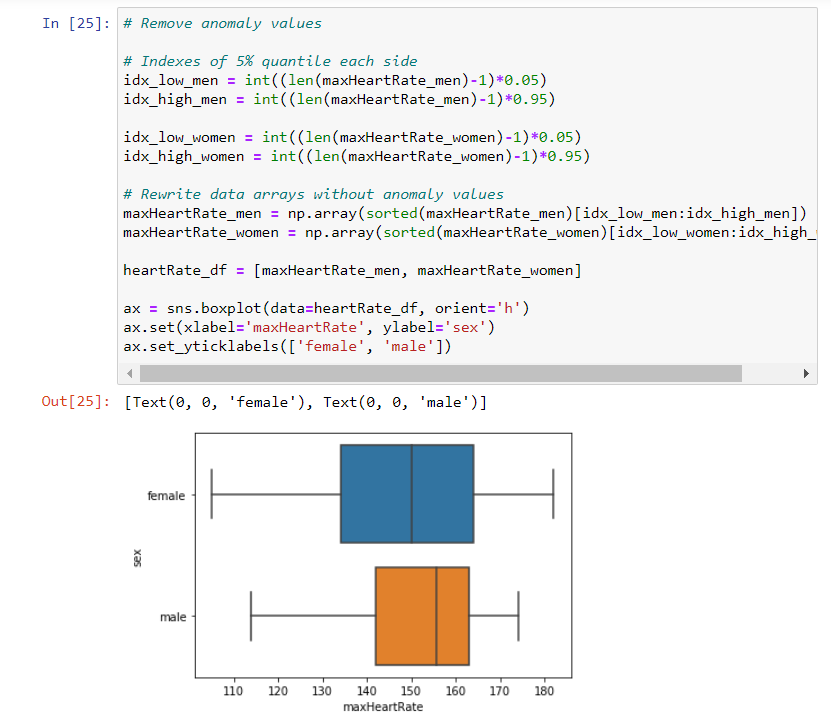
* 1. Практична реалізація
     1. Додавання данних. Дані відповідають вимогам



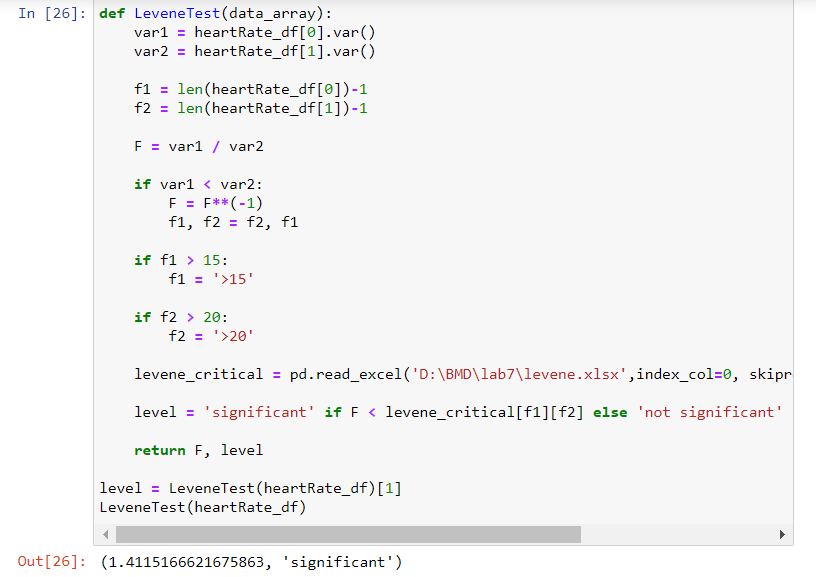
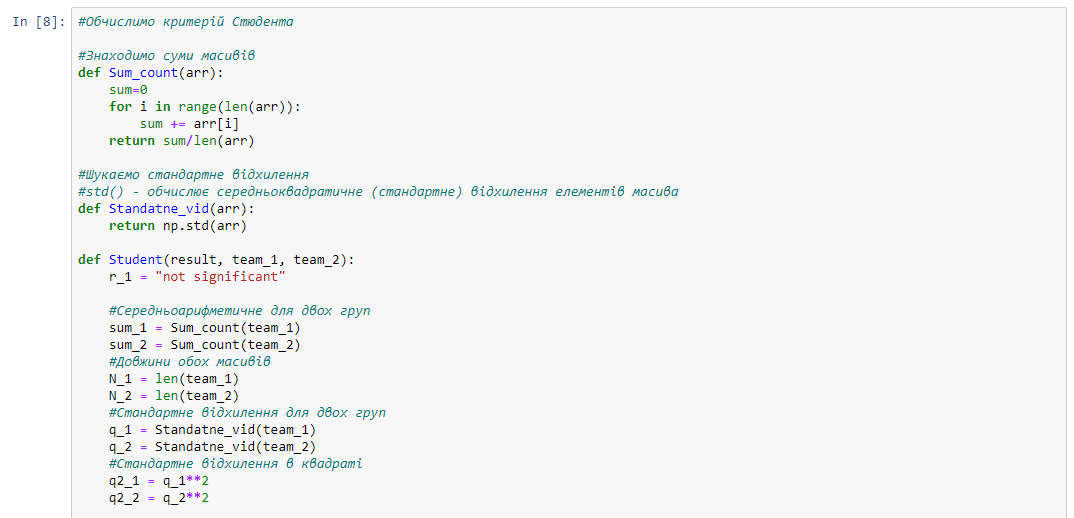








* + 1. Функція для визначення відмінності між групами за критерієм Стьюдента в залежності від рівності /відмінності дисперсій. Функція повертає значення критерію Стьюдента, значення ступеня свободи та результат оцінки значимості зв’язку (significant або not significant)



* 1. Висновки

В результаті перевірки даних на рівність дисперсії за тестом Лівеня було, отримано результат F < F\_tabl, що означає що ми повинні прийняти нульову гіпотезу, тобто зробити висновок про статичну рівність дисперсій.

В результаті перевірки відмінності між групами за допомогою критерію Стюдента для незалежних вибірок, було отримано результат t > t\_tabl, що свідчить про те, що ми відхиляємо нульову гіпотезу (тобто приймаємо альтернативну), та робимо висновок про статичну відмінність середніх

Стосовно наведеного прикладу, можна стверджувати, що між групами чоловіків і жінок є різниця в швидкості виконання тесту.

**РОЗДІЛ 8. ДИСПЕРСІЙНИЙ АНАЛІЗ**

* 1. Теоретичні відомості

Дисперсійний аналіз є сукупністю статистичних методів, призначених для перевірки гіпотез про зв’язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії. У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англомовної назви Analysis of Variations). Вперше цей метод було розроблено Р. Фішером в 1925 р.

Дисперсійний аналіз є одним з основних методів статистичної оцінки результатів експерименту. Все більш широке застосування отримує він і в аналізі економічної інформації. Дисперсійний аналіз дає змогу встановити, наскільки вибіркові показники зв'язку результативного і факторних ознак достатні для поширення одержаних за вибіркою даних на генеральну сукупність. Достоїнством цього методу є те, що він дає досить надійні висновки по вибірках невеликої чисельності.

Гіпотези:

: середні незалежні групи рівні

: середні незалежні групи відмінні

Відповідно до принципової схеми дисперсійний аналіз можна подати у вигляді п'яти послідовно виконуваних етапів:

1) визначення і розкладання варіації;

2) визначення числа ступенів свободи варіації;

3) обчислення дисперсій та їх співвідношень;

4) аналіз дисперсій та їх співвідношень;

5) оцінка вірогідності різниці між середніми і формулювання висновків з перевірки нульової гіпотези.

Алгоритм:

1. Підрахувати

де

– індивідуальні значення

- суми індивідуальних значень

n – кількість випробуваних у кожній групі

N – загальна кількість індивідуальних значень

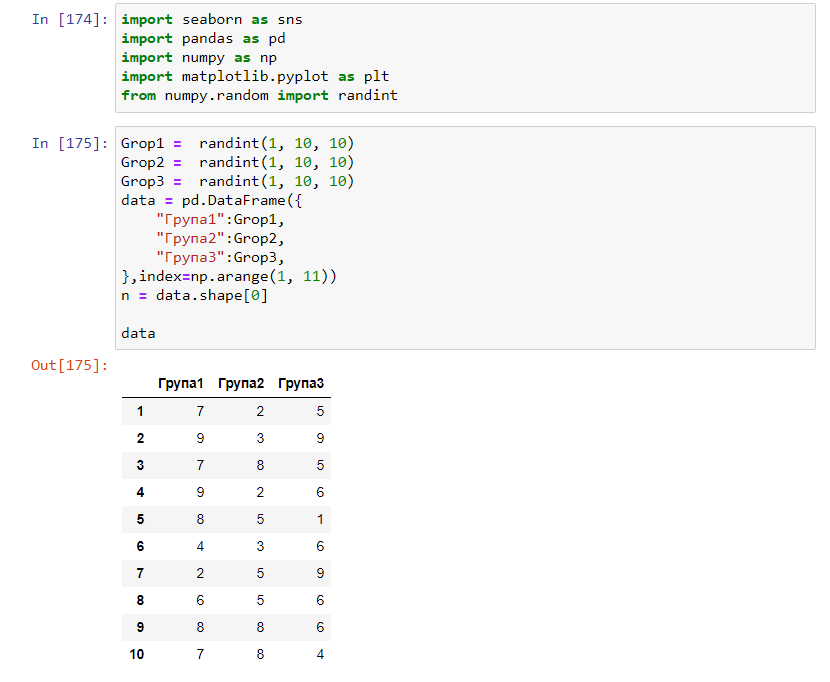
1. Підрахувати (сума квадратів загальна)
2. Підрахувати випадкову залишкову величину (сума квадратів випадкова)
3. Визначити число ступенів волі

с – кількість умов (градацій фактора/груп)

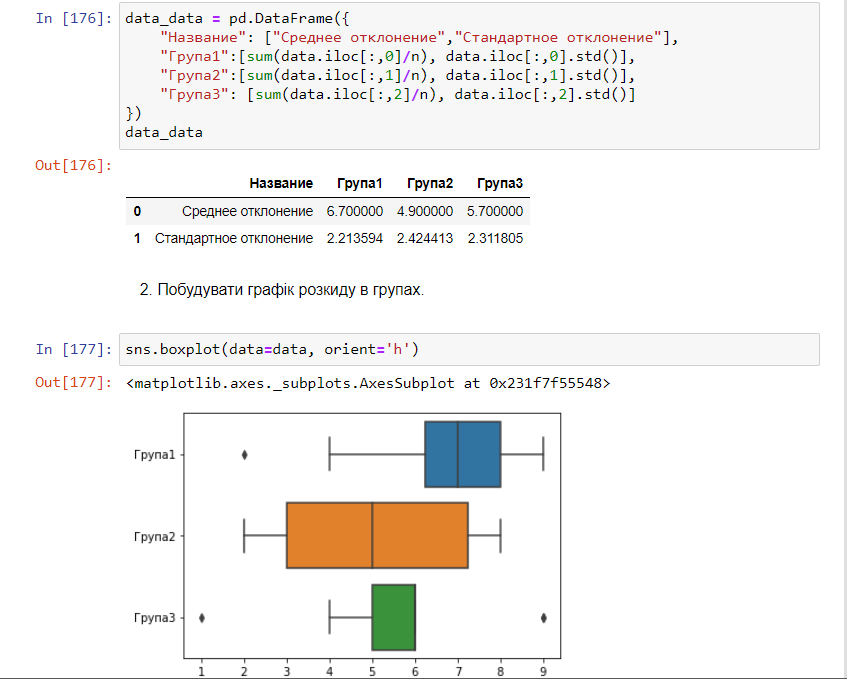
1. Розділити кожну суму квадратів на відповідне число ступенів волі
2. Підрахувати значення
3. Визначити з таблиці критичні значення
4. Зіставити емпіричне і критичне значення

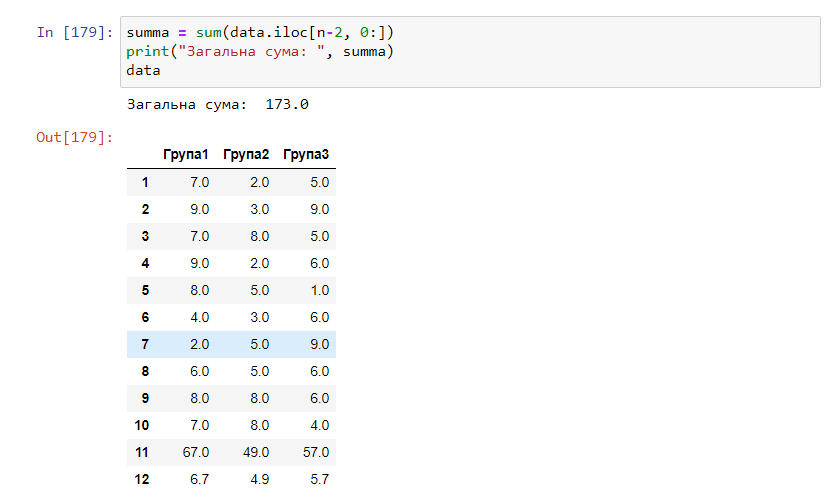
Якщо > H0 про рівність середніх відхиляється і приймається H1, в протилежному випадку навпаки H1 про відмінність середніх відхиляється приймається H0.

* 1. Практична реалізація
     1. Генеруємо данні з приблизно однаковою дисперсією

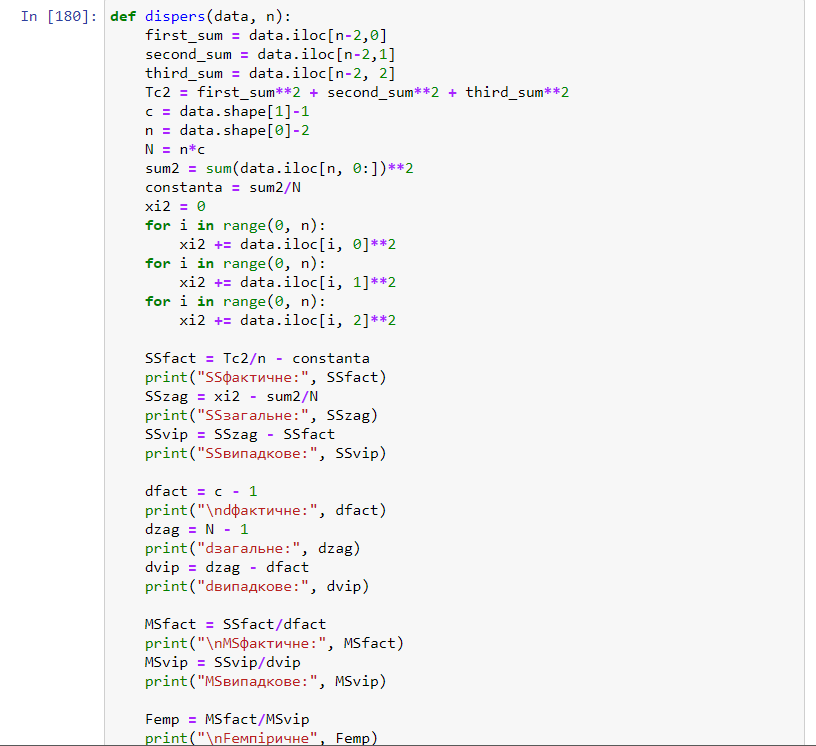


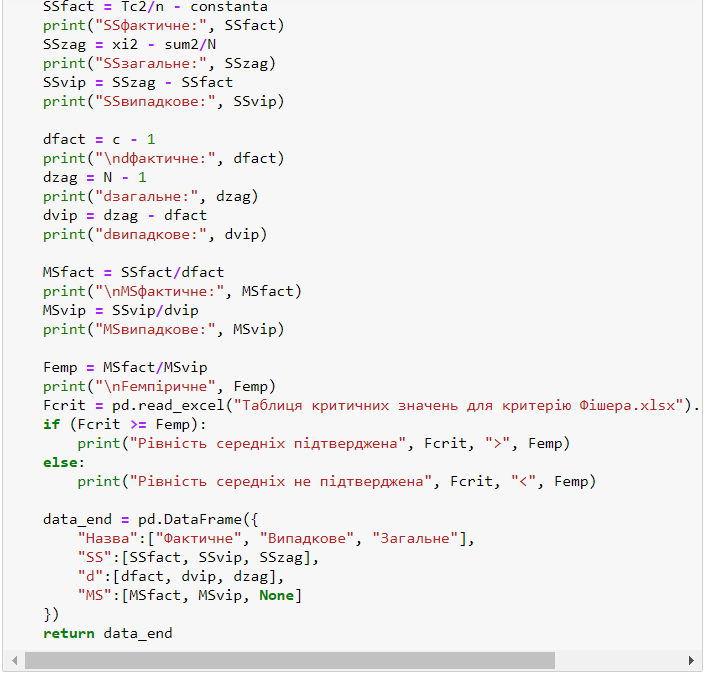
* + 1. Розрахувати та вивести у вигляді таблиці статистичні характиристики даних (середні та стандартні відхилення для кожної з груп).



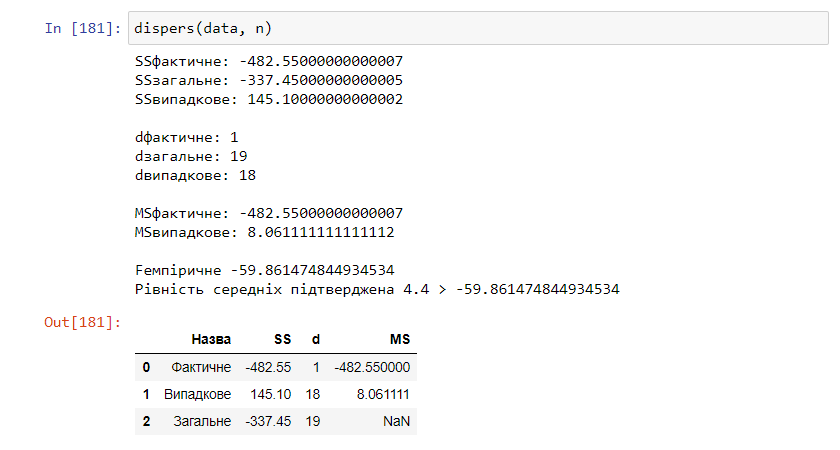


* + 1. Функція що реалізовує одно факторний дисперсійний аналіз





* + 1. Виведення результатів розрахунку кожного з етапів у вигляді таблиці



* 1. Висновки

Під час дисперсійного аналізу був отриманий результат 𝐹\_емпір > 𝐹\_крит, що свідчить про те що нульову гіпотезу про рівність середніх ми відхиляємо і приймаємо альтернативну: середні незалежних груп - відмінні. Що в нашому прикладі свідчить про те, що розпушування грунту впливає на річний приріст рослин.

**СПИСОК ЛІТЕРАТУРИ**

1. Введение в визуализацию данных с Matplotlib [Електронний ресурс] – Режим доступу до ресурсу: <https://nagornyy.me/courses/data-science/intro-to-matplotlib/#api-matplotlib-v-funktsionalnom-stilie>.
2. Введение в визуализацию данных с Seaborn [Електронний ресурс] – Режим доступу до ресурсу: <https://nagornyy.me/courses/data-science/intro-to-seaborn/>
3. Визуализация данных с использованием Seaborn [Електронний ресурс] – Режим доступу до ресурсу: <https://www.machinelearningmastery.ru/data-visualization-using-seaborn-fc24db95a850/>.
4. ДИСПЕРСІЙНИЙ АНАЛІЗ [Електронний ресурс] – Режим доступу до ресурсу: <http://virtuni.education.zp.ua/edu_cpu/file.php/2753/Analiz_danikh_3.pdf>.
5. Дисперсійний аналіз. Теоретичні основи і принципова схема дисперсійного аналізу [Електронний ресурс] – Режим доступу до ресурсу: <https://pidruchniki.com/1929100153035/statistika/dispersiyniy_analiz>.
6. Классические методы статистики: t-критерий Стьюдента [Електронний ресурс] – Режим доступу до ресурсу: <https://r-analytics.blogspot.com/2012/03/t.html>.
7. Критерий Шапиро-Уилка [Електронний ресурс] – Режим доступу до ресурсу: <http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%A8%D0%B0%D0%BF%D0%B8%D1%80%D0%BE-%D0%A3%D0%B8%D0%BB%D0%BA%D0%B0>.
8. Математические методы в психологии [Електронний ресурс] – Режим доступу до ресурсу: <https://ppt-online.org/37630>.
9. Математические методы в психологии [Електронний ресурс] – Режим доступу до ресурсу: <https://mypresentation.ru/presentation/1569694093_matematicheskie-metody-v-psixologii>.
10. Matplotlib [Електронний ресурс] – Режим доступу до ресурсу: <https://ru.wikipedia.org/wiki/Matplotlib>.
11. Научная графика в Python [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: <https://nbviewer.jupyter.org/github/whitehorn/Scientific_graphics_in_python/blob/master/P1%20Chapter%201%20Pyplot.ipynb>.
12. Однофакторний дисперсійний аналіз / Шеховцов С. Б., Гнусов Ю. В.. – 2018. – С. 8.
13. Основы Pandas №1 Чтение файлов, DataFrame, отбор данных [Електронний ресурс] – Режим доступу до ресурсу: <https://pythonru.com/uroki/osnovy-pandas-1-chtenie-fajlov-dataframe-otbor-dannyh>.
14. Pandas [Електронний ресурс] – Режим доступу до ресурсу: <https://uk.wikipedia.org/wiki/Pandas>.
15. Работа с IPython и Jupyter Notebook [Електронний ресурс] – Режим доступу до ресурсу: <https://devpractice.ru/python-lesson-6-work-in-jupyter-notebook/>.
16. T-критерии [Електронний ресурс] – Режим доступу до ресурсу: <http://statistica.ru/theory/t-kriterii/#t-two-ind>.
17. Хи-квадрат (χ2) Пирсона: условия применения метода и интерпретация [Електронний ресурс] – Режим доступу до ресурсу: <https://lit-review.ru/biostatistika/kriterijj-khi-kvadrat-pirsona/#%D0%BF%D1%80%D0%B8%D0%BC%D0%B5%D1%80>.
18. Як налаштувати Jupyter Notebook для Python 3 [Електронний ресурс] – Режим доступу до ресурсу: <https://echo.lviv.ua/dev/5922>.