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»

Виконав: студент гр. БС-83 Шарган Павло

Прийняв: викладач 2020 р.

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

Київ

КПІ ім. Ігоря Сікорського

2020

ЗМІСТ

[РОЗДІЛ 1. КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1 4](#_30j0zll)

[РОЗДІЛ 2. ВІЗУАЛІЗАЦІЯ ДАНИХ. БІБЛІОТЕКf MATPLOTLIB](#_2et92p0) 8

[РОЗДІЛ 3. ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ БІБЛІОТЕКИ SEABORN](#_1t3h5sf) 15

[РОЗДІЛ 4. ПЕРЕВІРКА ДАНИХ НА НОРМАЛЬНІСТЬ ЗА ДОПОМОГОЮ КРИТЕРІЮ ШАПІРА-УІЛКА](#_17dp8vu) 20

[РОЗДІЛ 5. КРИТЕРІЙ ПІРСОНА ДЛЯ ПОРІВНЯННЯ ДЕКІЛЬКО ГРУП ЗА РОЗПОДІЛЕННЯМ ОЗНАКИ](#_lnxbz9) 26

[РОЗДІЛ 6. ПАРНИЙ T-КРИТЕРІЙ СТЬЮДЕНТА](#_2jxsxqh) 34

[РОЗДІЛ 7. T-КРИТЕРІЙ СТЬЮДЕНТА ДЛЯ НЕЗАЛЕЖНИХ ВИБІРОК](#_4i7ojhp) 38

[РОЗДІЛ 8. ДИСПЕРСІЙНИЙ АНАЛІЗ](#_2bn6wsx) 43

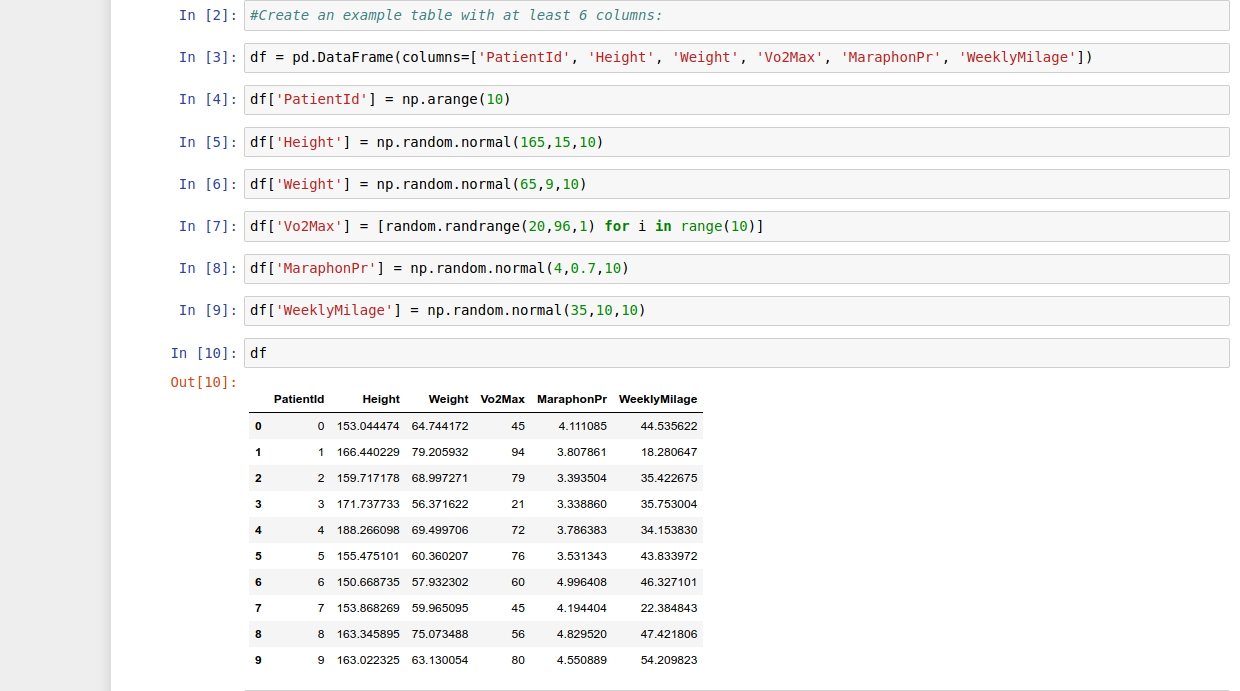
# **РОЗДІЛ 1. КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1**

## *Теоретичні відомості*

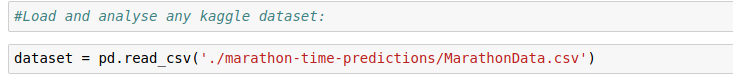
## Jupyter Notebook - це командна оболонка для інтерактивних обчислень. Цей інструмент може використовуватися не тільки з Python, але і іншими мовами програмування: Julia, R, Haskell і Ruby. Він часто використовується для роботи з даними, статистичним моделюванням і машинним навчанням.

## *1.2 Практична реалізація*

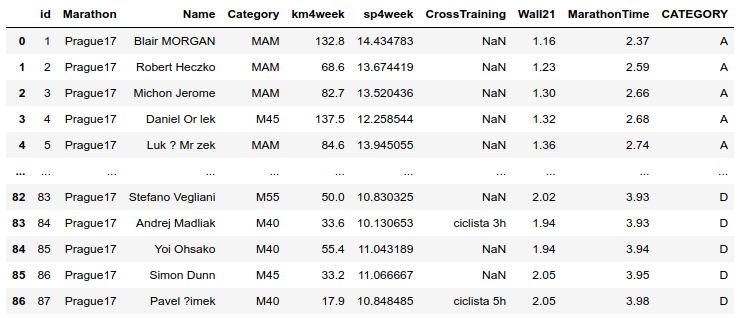
1. Генерація даних



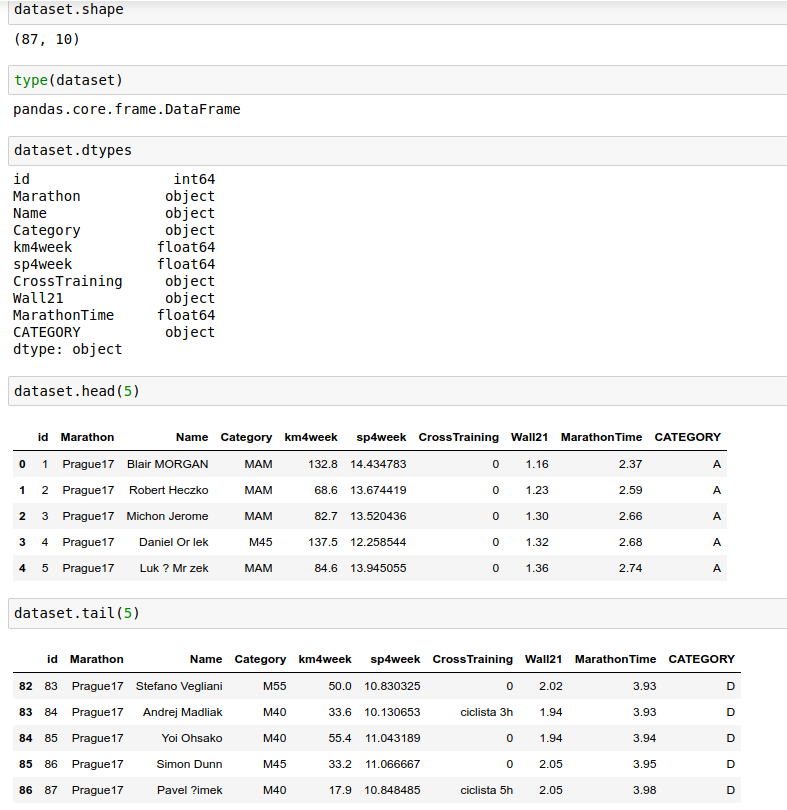
2. Завантаження даних з CSV файлу



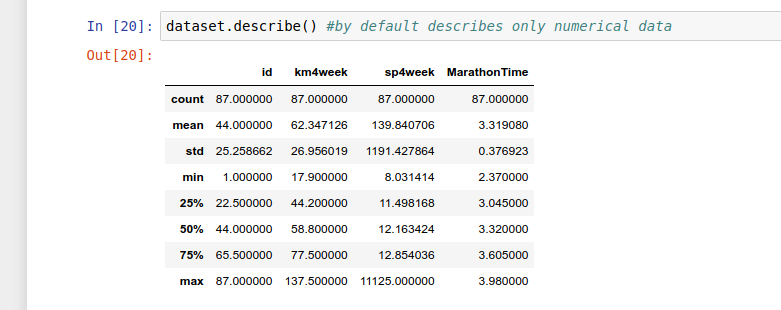
3. Відображення завантажених даних



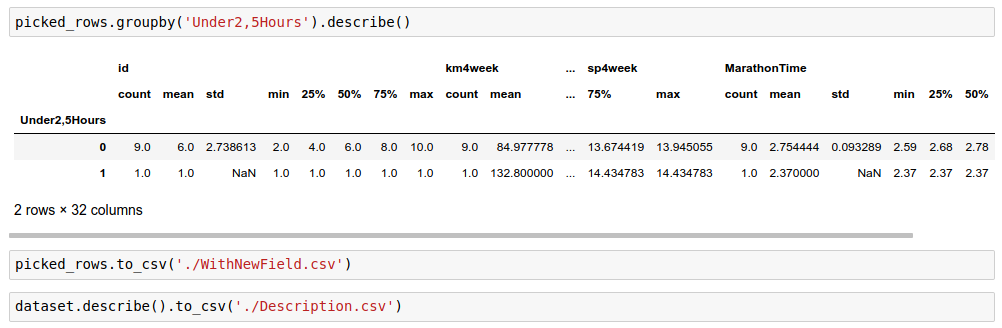
4. Базова статистика про завантажені дані



5. Функція describe()



6. Статистика з розділенням на групи та збереження даних



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

## *Теоретичні відомості*

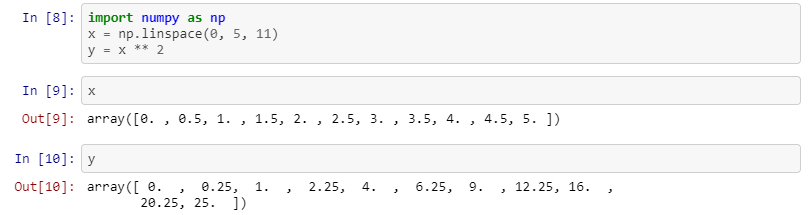
Бібліотека matplotlib - це бібіліотека двовимірної графіки для мови програмування python, за допомогою якої можна створювати високоякісні малюнки різних форматів. Matplotlib - це модуль-пакет для python.

Matplotlib є гнучким, легко конфігурованим пакетом, який разом з NumPy, SciPy і IPython надає можливості, подібні MATLAB. В даний час пакет працює з декількома графічними бібліотеками, включаючи wxWindows і PyGTK.

Завантажити та встановити matplotlib можна з офіційного сайту бібліотеки.

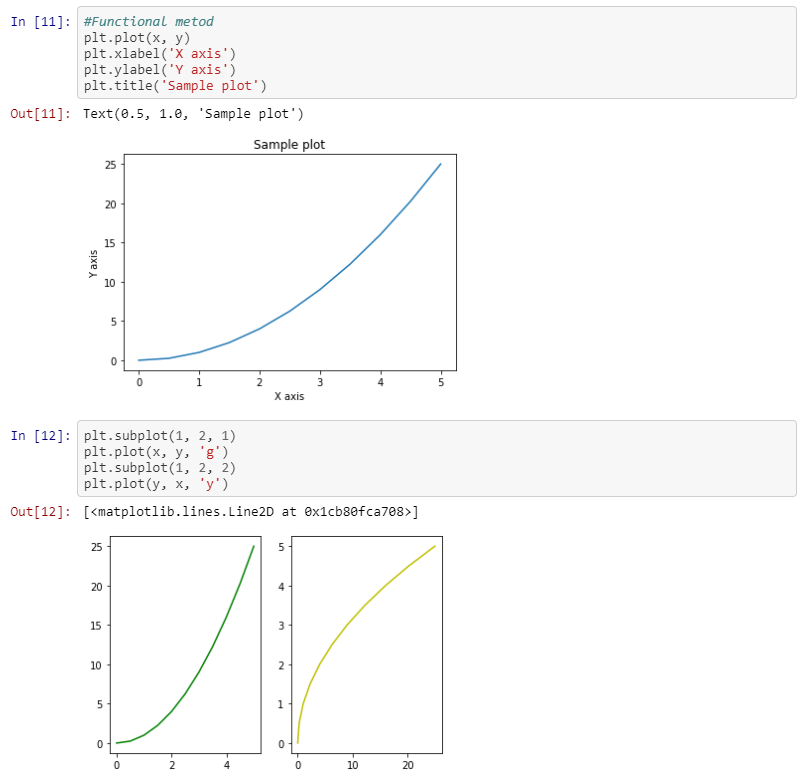
## *Практична реалізація*

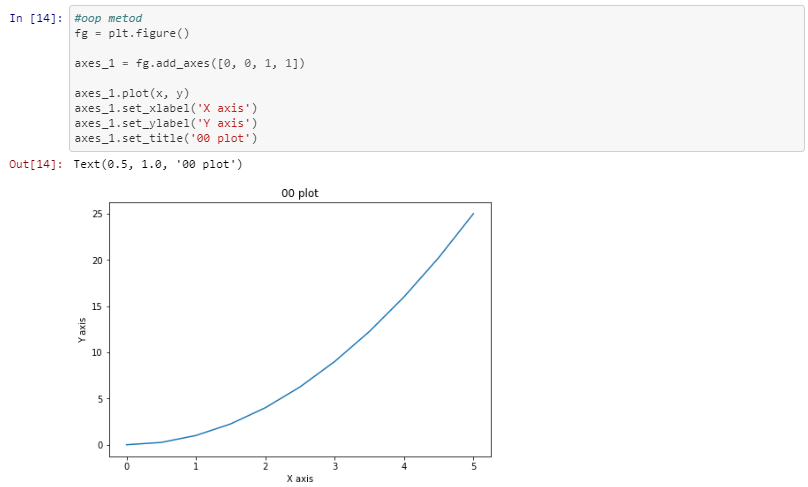
1. Основи побудови графіків



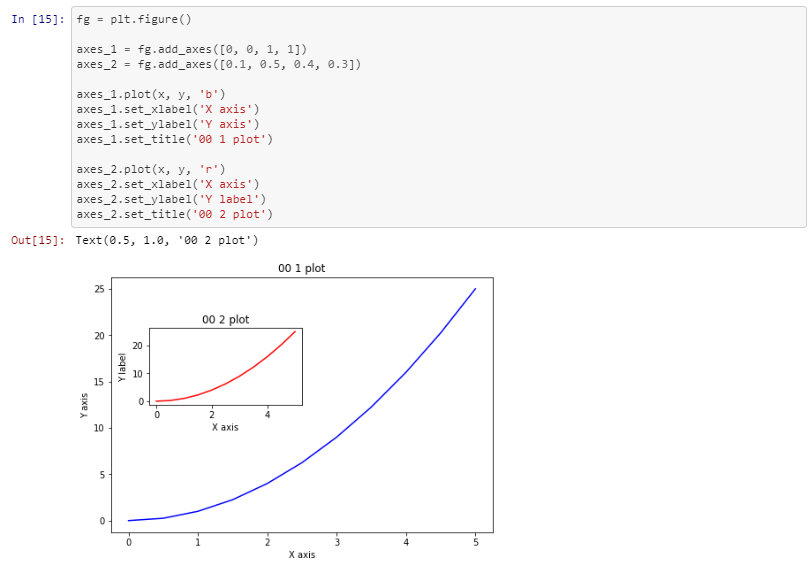
1.1 Побудова графіків:

1.1.1 Функціональний метод

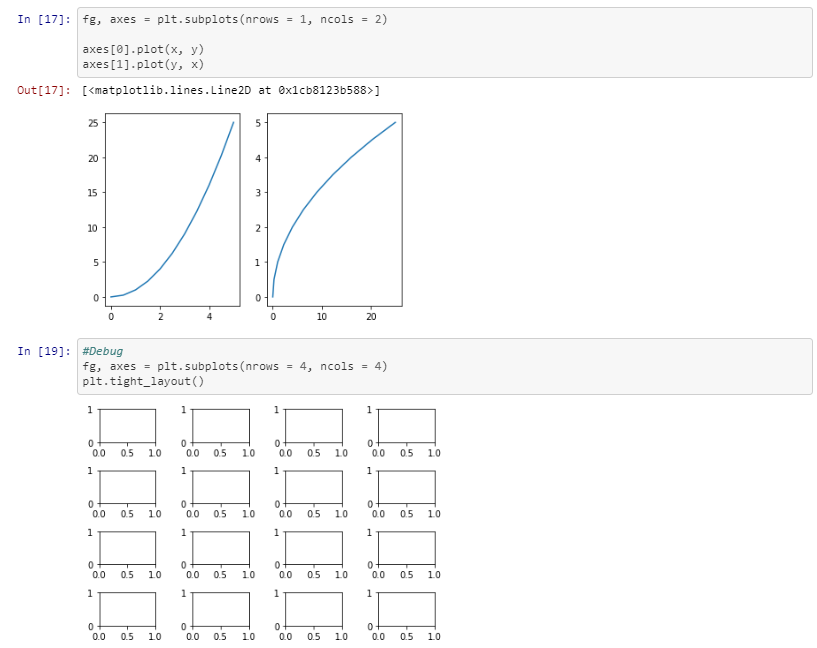


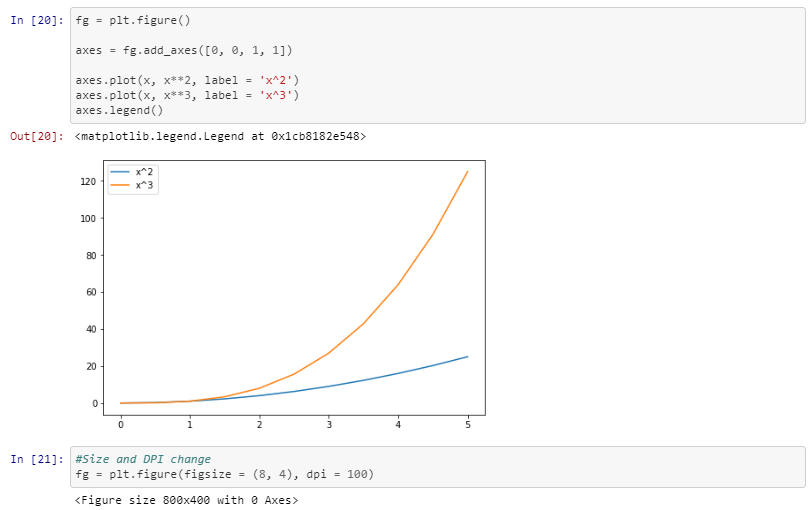
1.1.2. Об’єктно-орієнтований методд

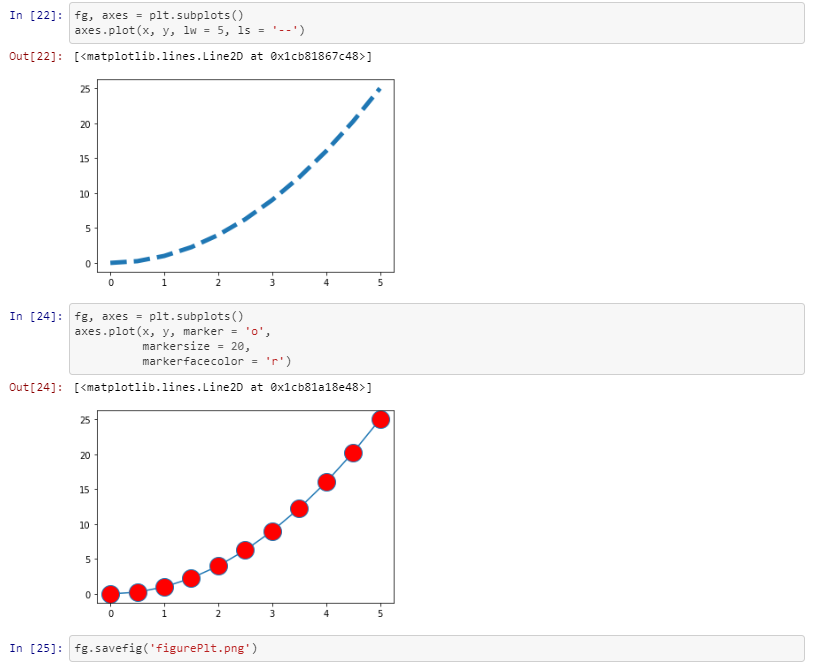
1.1.3. Одночасна побудова декількох графіків

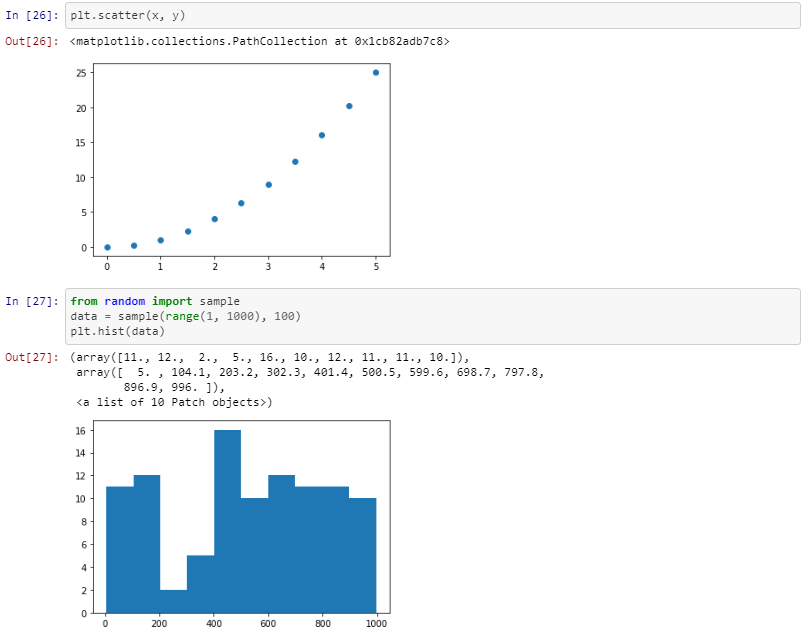


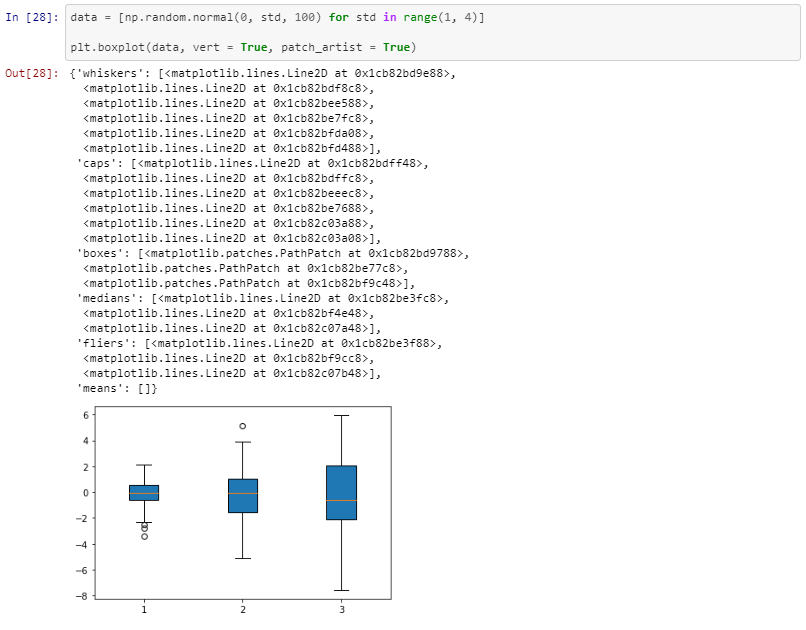
1.1.4. Використання іншого функціоналу бібліотеки











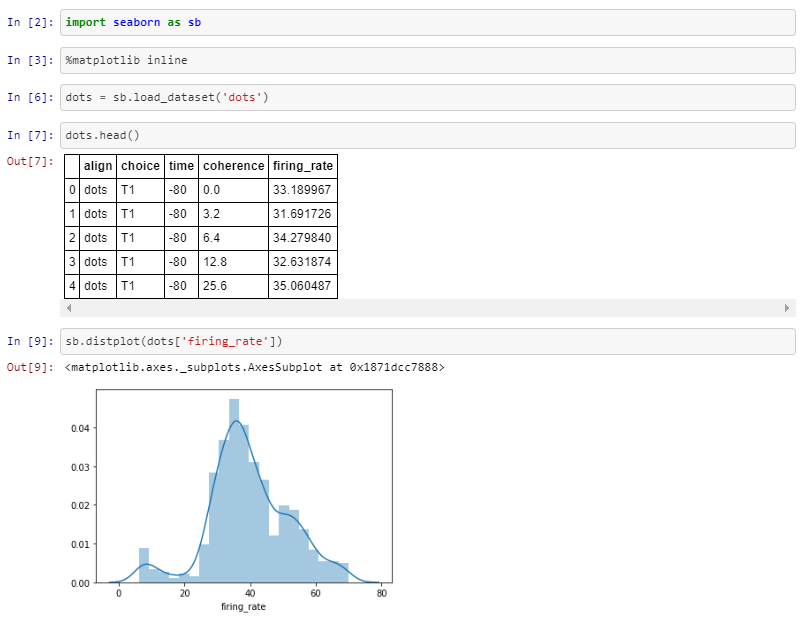
# **РОЗДІЛ 3. ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ БІБЛІОТЕКИ SEABORN**

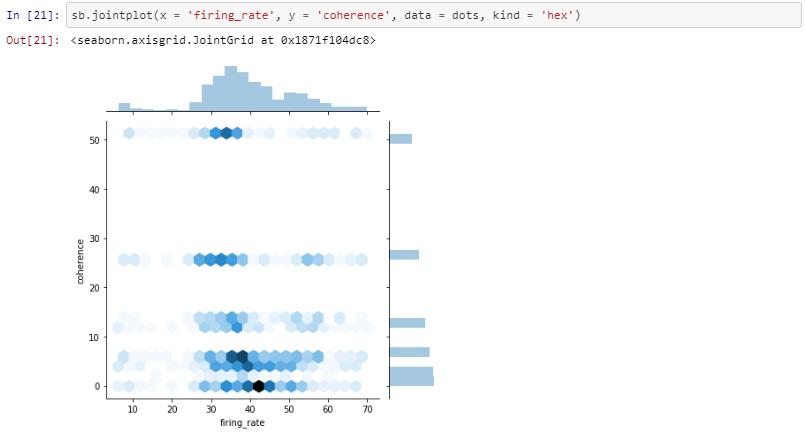
## *Теоретичні відомості*

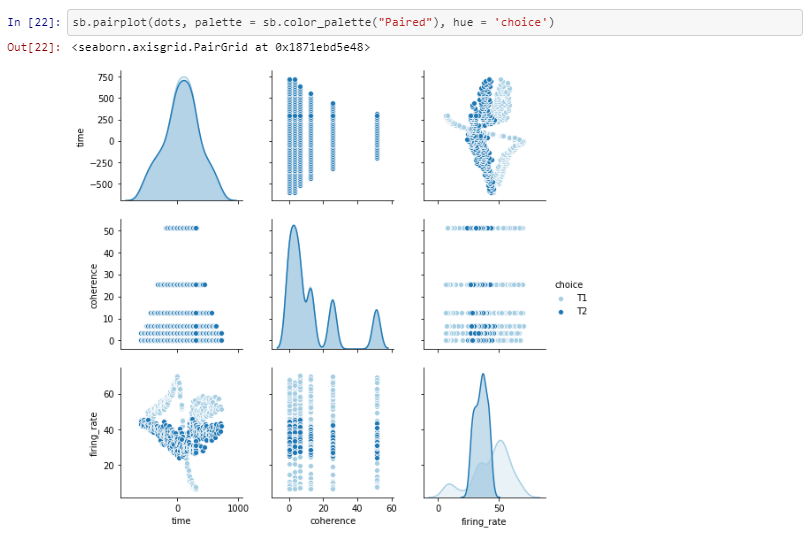
Seaborn - бібліотека візуалізації даних Python, що побудована на matplotlib, але більш високого рівня. Seaborn містить більш адекватні дефолтні настройки оформлення графіків. В бібліотеці є досить складні типи візуалізації, які в matplotlib зажадали б великого кількість коду.

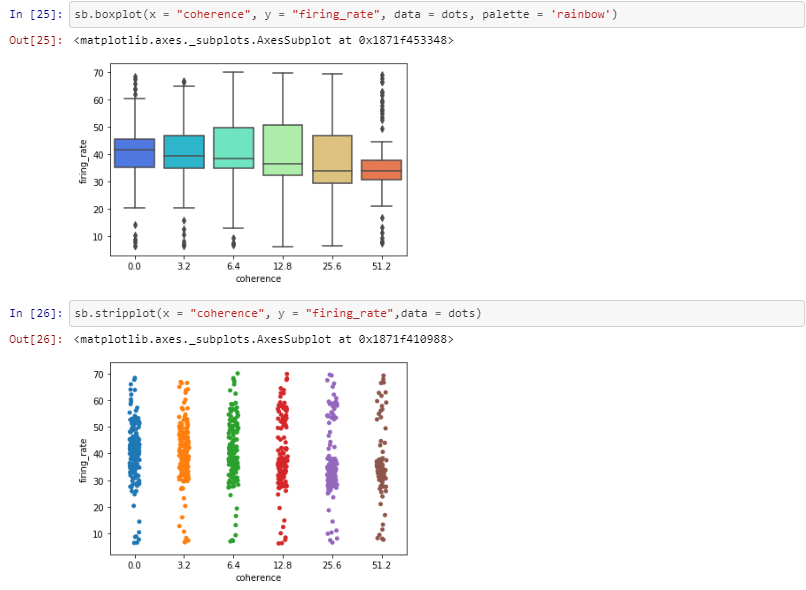
## *3.2 Практична реалізація*

1. Тестові дані, їх візуалізація за допомогою графіку для аналізу кількісних даних

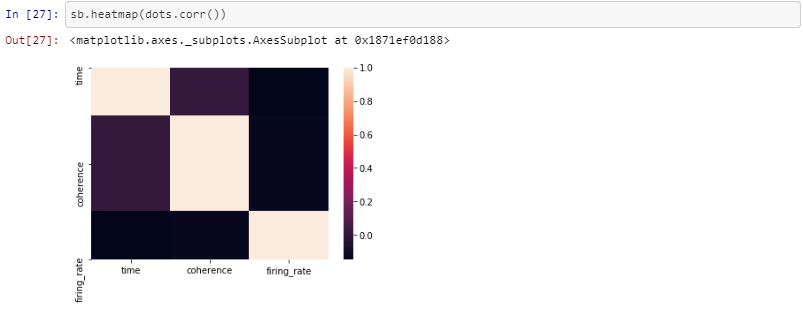




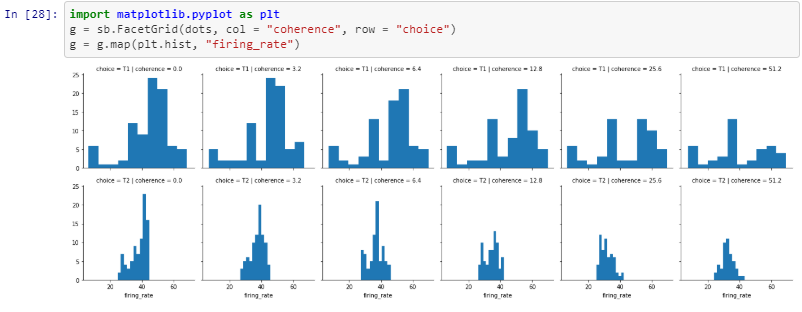


1. Візуалізація даних за допомогою графіків для аналізу категоріальних даних.

3. Візуалізація даних за допомогою графіків для аналізу взаємозв’язку даних



4. Візуалізація даних за допомогою графіків для відображення даних у вигляді сітки



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

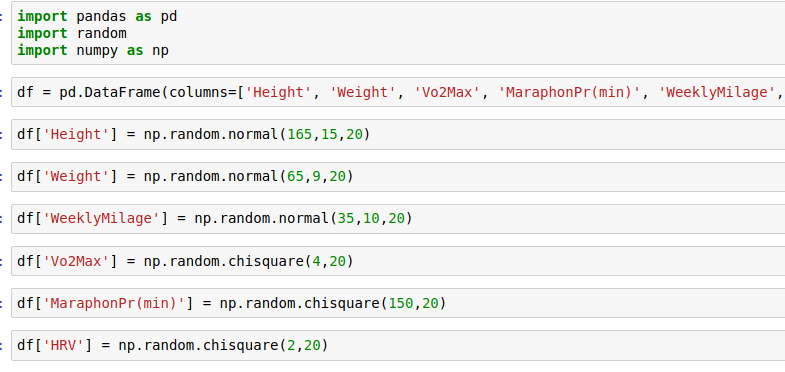
## *Теоретичні відомості*

Критерій Шапіро-Уілка використовується для перевірки гіпотези H0: «випадкова величина X розподілена нормально» і є одним найбільш ефективних критеріїв перевірки нормальності. Критерії, перевіряючі нормальність вибірки, є окремим випадком критеріїв згоди. Якщо вибірка нормальна, можна далі застосовувати потужні параметричні критерії, наприклад, критерій Фішера.

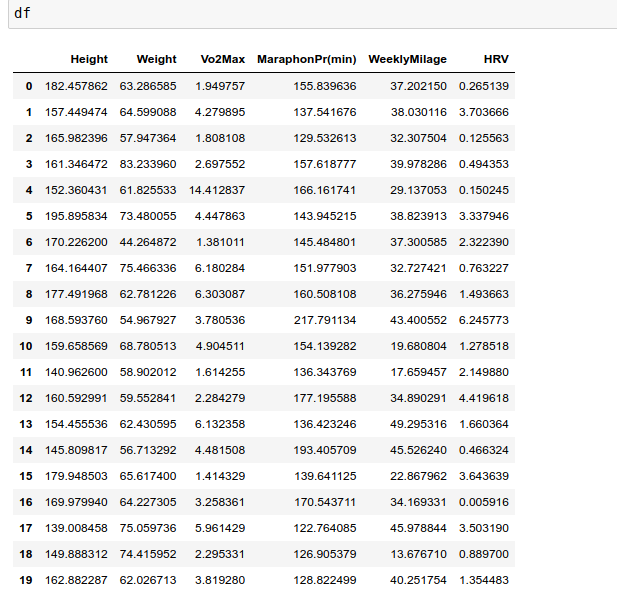
Критерій Шапіро-Уілка є найбільш ефективним, так як він має більшу потужність в порівнянні з альтернативними критеріями перевірки нормальності.

## *Практична реалізація*

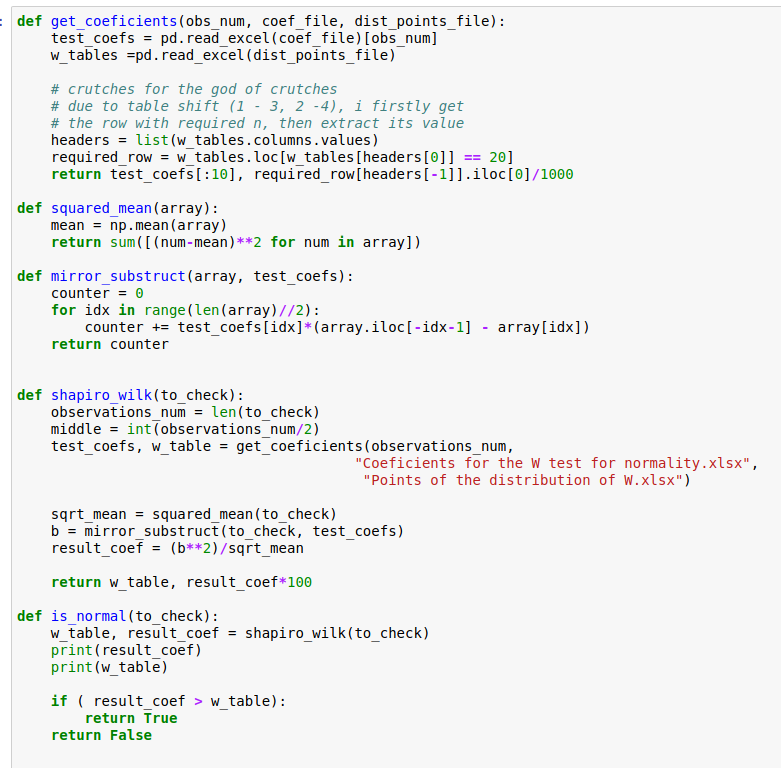
1. Генерування набору даних



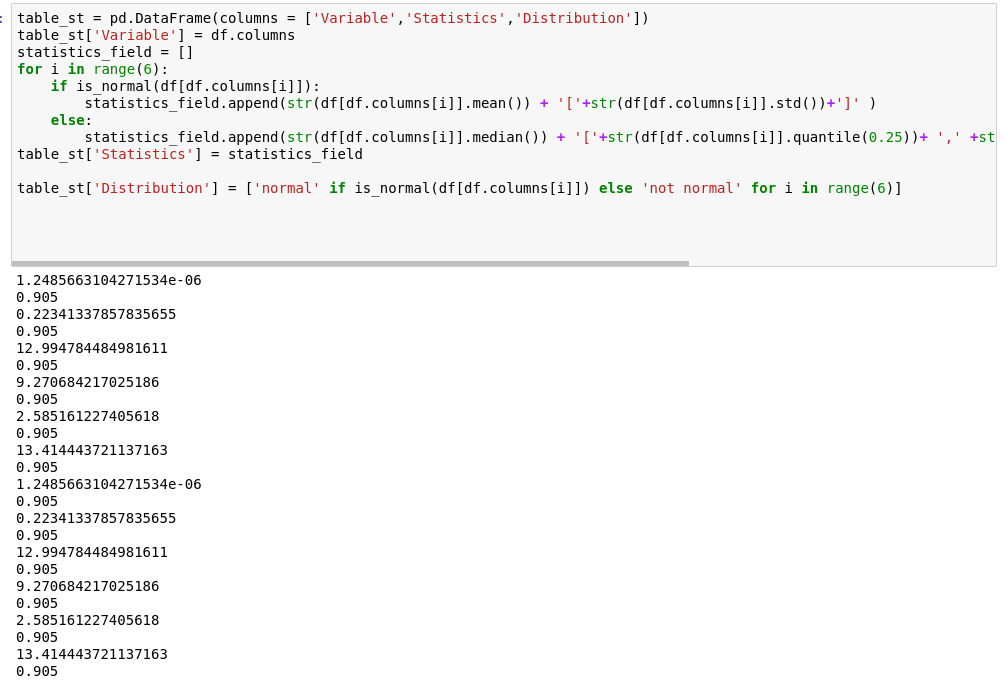
1. Вивід даних



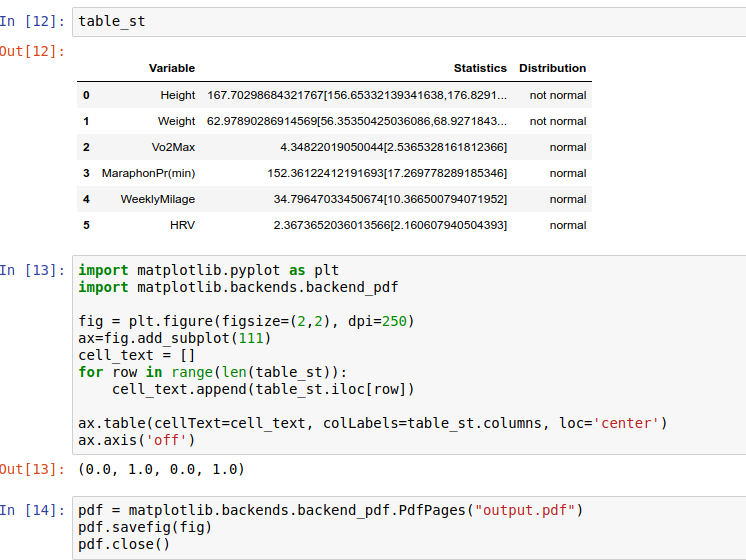
1. Створення функції, що реалізує критерій Шапіро-Уілка



1. Генерування пустої таблиці, в якій кількість рядків – це кількість змінних + 1. Назва стовпчиків – Variable, Statistics, Distribution.   
   Заповнення таблиці



1. Розрахунок для змінних медіани та інтерквартильного розмаху. Запис в таблицю та збереження таблиці у пдф форматі



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

## *Теоретичні відомості*

Критерій хі-квадрат для аналізу таблиць спряженості був розроблений і запропонований в 1900 році англійським математиком, статистиком, біологом і філософом, засновником математичної статистики і одним з основоположників біометрії Карлом Пірсоном (1857-1936).

*Таблиці спряженості* - це візуальне (табличне) відображення залежності між двома ***якісними*** ознаками. Інтрерпретувати цей зв'язок можна як взаємозвя'зок між розподілом за однією ознакою в залежності від градації іншої ознаки. Прикладом може бути залежність частоти летальності в групах, що приймали різні лікувальні препарати (залежність летальності від схеми лікування), або залежність кількості осіб, що займаються різними видами спорту та частоти травматичних випадків (оцінка травматичності зайнять різними видами спорту).

Рядки таблиці спряженості відповідають значенням однієї змінної, стовпці - значенням іншої змінної. *Для побудови таблиць спряженості кількісні шкали попередньо повинні бути згруповані в інтервали.* Область визначення випадкової величини розбивають на 𝑘k інтервалів, що не перетинаються:

𝑥0<𝑥1<...<𝑥𝑘−1<𝑥𝑘, x0<x1<...<xk−1<xk,

де 𝑥0 - нижня границя області визначення випадкової величини, 𝑥𝑘 - верхня границя.

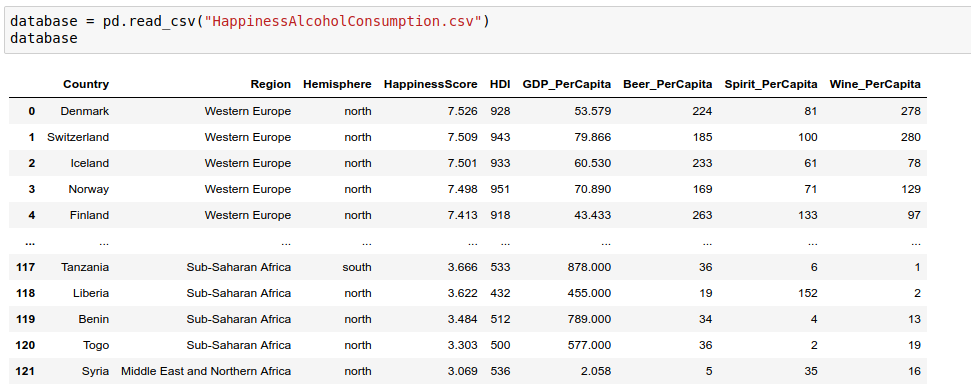
На перетині рядка і стовпця вказується частота спільної появи відповідних значень двох ознак. Сума частот по рядку називається маргінальною частотою рядка; сума частот по стовпцю - маргінальною частотою стовпчика. У таблиці спряженості можуть бути представлені як абсолютні, так і відносні частоти (в частках або відсотках). Відносні частоти можуть розраховуватися по відношенню: а) до маргінальної частоти по рядку; б) до маргінальної частоти по стовпцю; в) до обсягу вибірки.

Статистична гіпотеза для даного критерію має наступний вигляд:

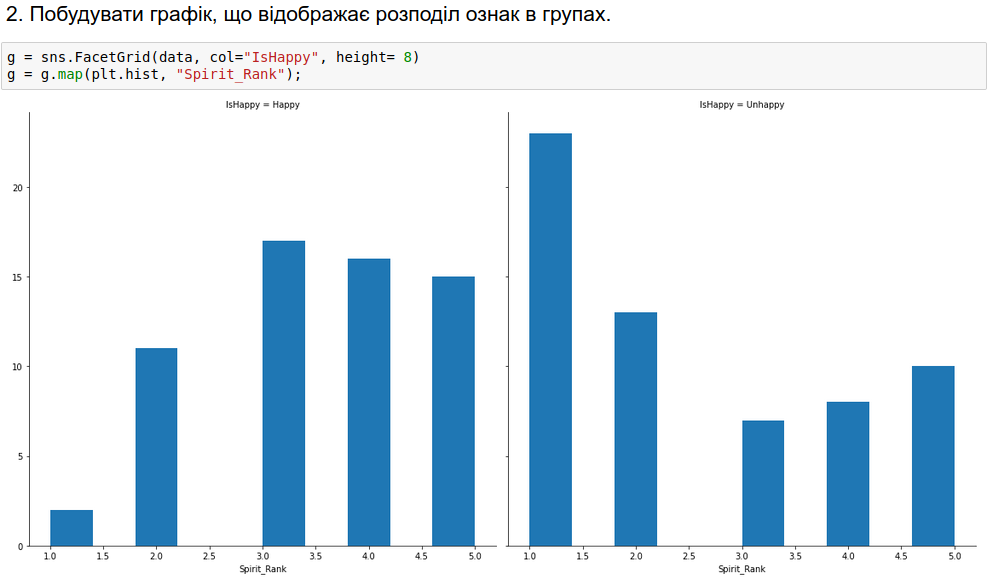
* Основна (нульова). Ознаки не пов'язані.
* Конкуруюча (альтернативна). Ознаки пов'язані.

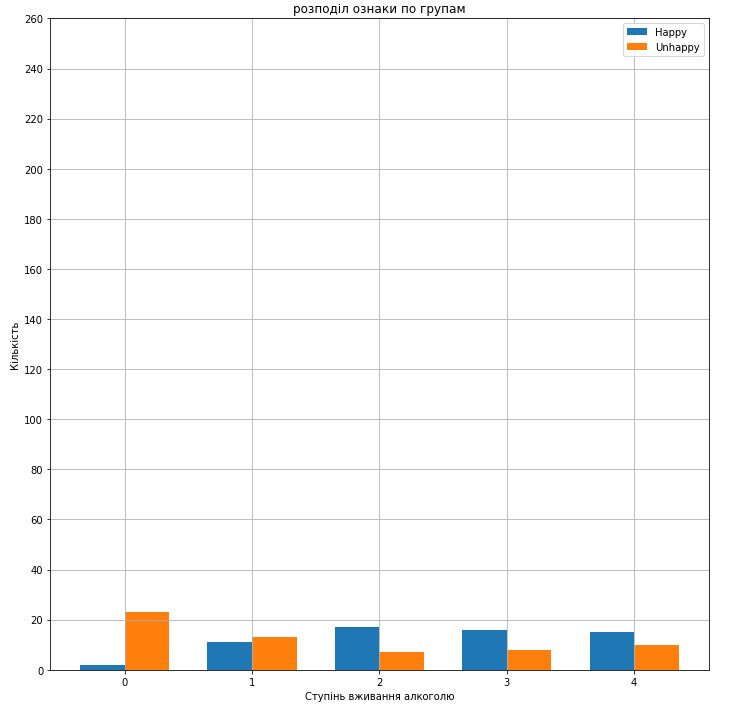
## *Практична реалізація*

1. Взяття реальних даних з kaggle – успішність студентів



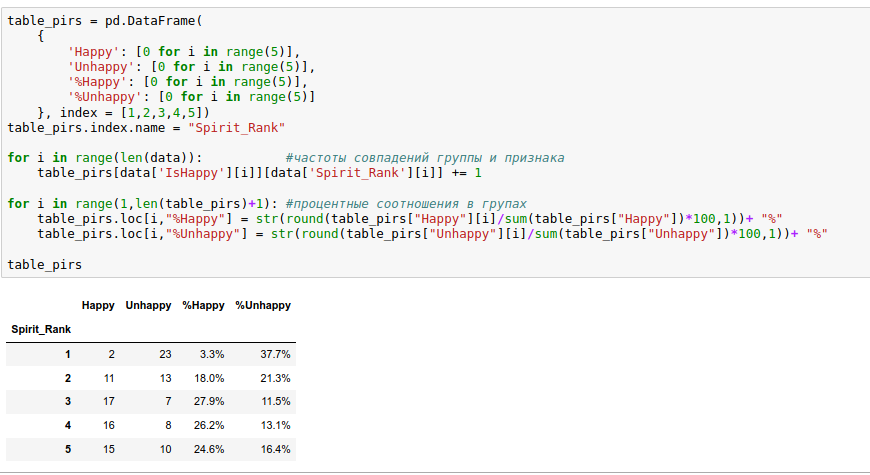
1. Будування графіку, що відображає розподіл ознак в групах



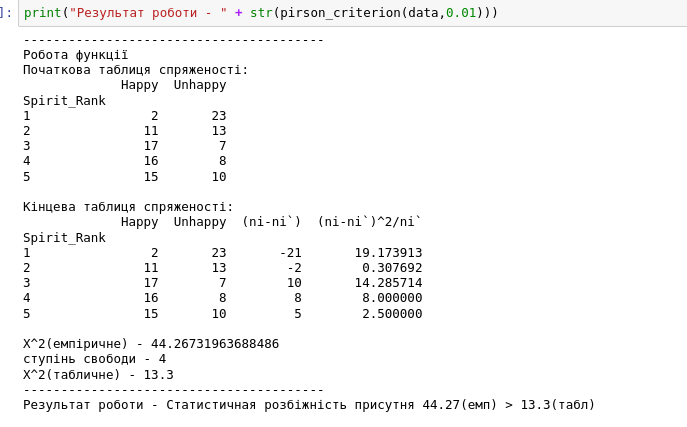


1. Створення функції для оцінки зв’зку між ознакою та групою за критерієм Пірсона





1. Виведення результатів



## *Висновки*

#### Під час порівняння двох груп на декількох груп за розподіленням ознаки з допомогою розрахунку критерію Пірсона було отримано результат, що **не** довзволяє припустити статистично значиму розбіжність між групами як на рівні значимості 0.05, так і 0.01.

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

## *Теоретичні відомості*

Т-критерій Стьюдента - загальна назва для класу методів статистичної перевірки гіпотез (статистичних критеріїв), заснованих на розподілі Стьюдента. Найбільш часті випадки застосування t-критерію пов'язані з перевіркою рівності середніх значень в двох виборках.

Критерій Стьюдента t для парних вибірок застосовується для порівняння двох середніх значень отриманих від одного і того об'єкта.

Гіпотеза, яка перевіряєть даним тестом може бути виражена наступним чином:

𝐻0:μ1=μ2 («парні групи рівні»)

𝐻1:μ1≠μ2 («парні групи відмінні»)

або:

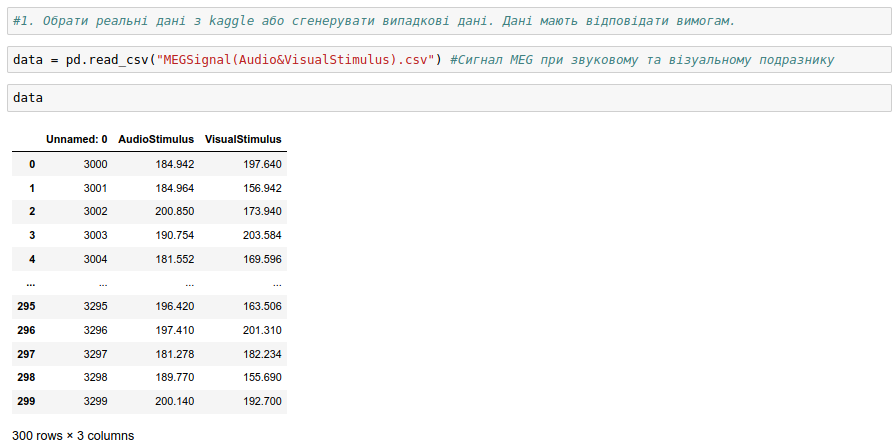
𝐻0:μ1−μ2=0 («різниця між середніми у парних групах за показником, що досліджується рівна 0»)

𝐻1:μ1−μ2≠0 («різниця між середніми у парних групах за показником, що досліджується не рівна 0»)

Тут μ1 - середнє значення змінної в першій групі; μ2 - середнє значення змінної в першій групі.

## *Практична реалізація*

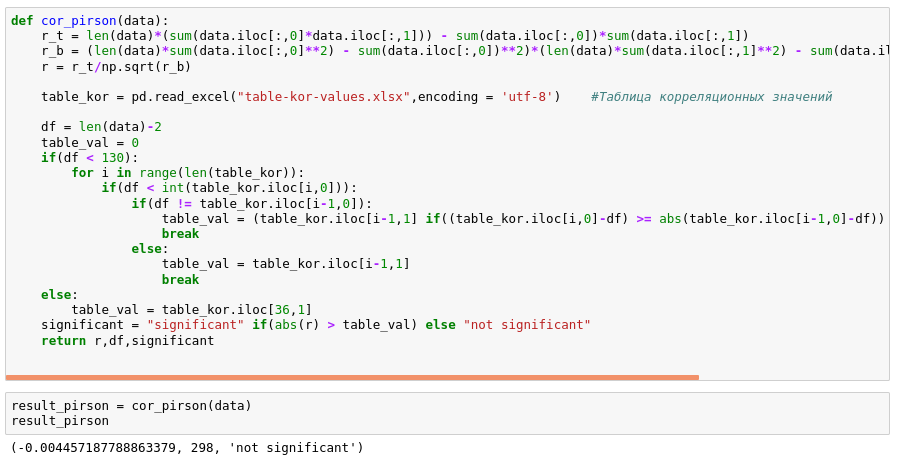
1. Генерування випадкових даних за вимогами

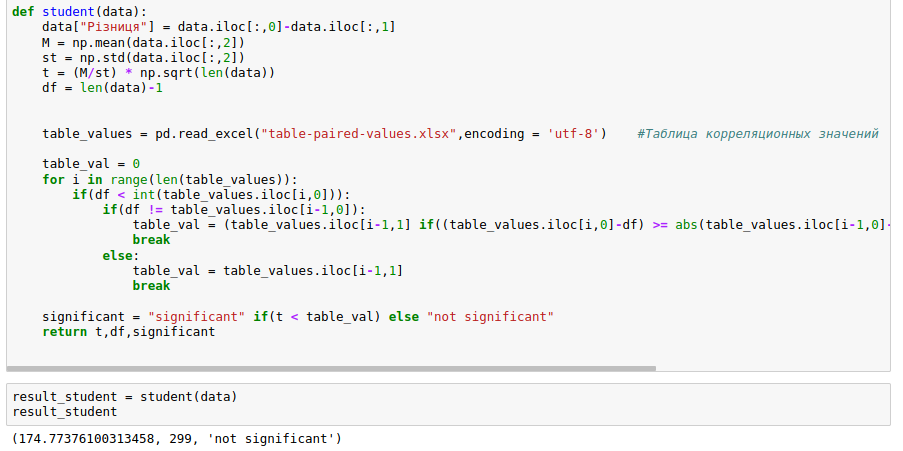


1. Будування графіку для відображення розкиду даних в групах



1. Написання функцій:  
   - для оцінки зв’язку між групами за коефіцієнтом кореляції Пірсона, має повертати значення коефіцієнта кореляції, значення ступеня свободи та результат оцінки значимості зв’язку;  
   - для визначення відмінності між групами за критерієм Стьюдента, має повертати значення критерію, значення ступеня свободи та результат оцінки значимості зв’язку;







## *Висновки*

За коефіцієнтом кореляції Пірсона робимо висновок, що статистично значимої залежності між групами не існує, так як коефіцієнт кореляції Пірсона дорівнює ~0.

Під час аналізу відмінності між групами за допомогою критерія t-Стьюдента, було прийнято нульову гіпотезу про рівність середніх, оскільки значення критерію менше за критичне.

# **РОЗДІЛ 7. T-КРИТЕРІЙ СТЬЮДЕНТА ДЛЯ НЕЗАЛЕЖНИХ ВИБІРОК**

## *Теоретичні відомості*

t-критерій Стьюдента для незалежних вибірок застосовується для порівняння середніх значень двох незалежних між собою вибірок.

Даний тест може порівнювати середні лише для двох (і лише двох) груп. Він не може проводити порівняння більш ніж двох груп.

Гіпотеза, яка перевіряєть даним тестом може бути виражена наступним чином:

𝐻0:μ1= μ2 («середні двох незалежних груп рівні»)

𝐻1:μ1≠μ2 («середні двох незалежних груп відмінні»)

або:

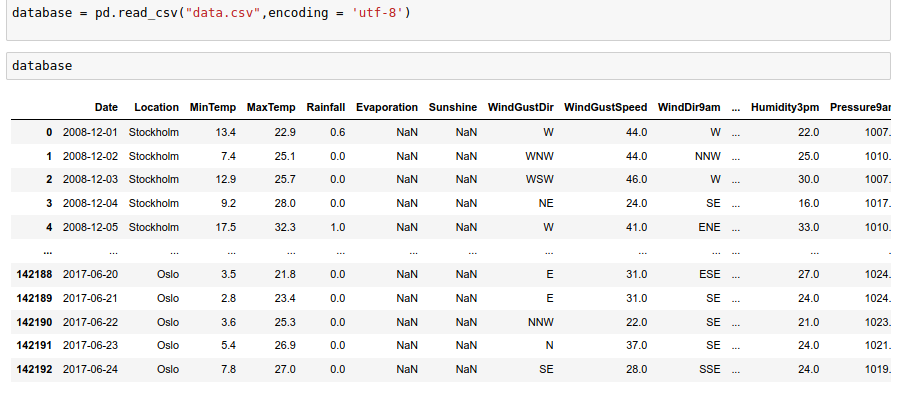
𝐻0:μ1−μ2 =0 («різниця між середніми у незалежних групах за показником, що досліджується рівна 0»)

𝐻1:μ1−μ2 ≠ 0 («різниця між середніми у незалежних групах за показником, що досліджується не рівна 0»)

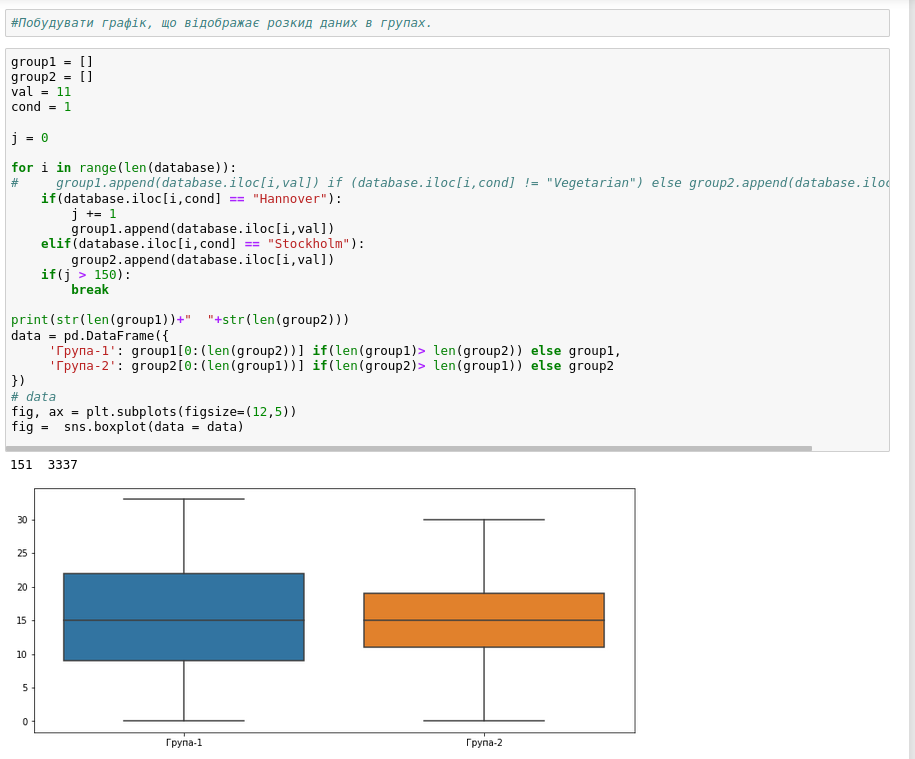
Тут μ1- середнє значення змінної в першій групі; μ2- середнє значення змінної в першій групі.

## *Практична реалізація*

1. Генерування випадкових даних



1. Будування графіку для відображення розкиду даних в групах



1. Написання функцій:  
   - для оцінки відмінностей дисперсій між групами за критерієм Лівеня, що має повертати значення тесту та результат оцінки рівності дисперсій (significant або not significant);

- для визначення відмінності між групами за критерієм Стьюдента в залежності від рівності/відмінності дисперсій, має повертати значення критерію Стьюдента, значення ступеня свободи та результат оцінки значимості зв’зку (significant або not significant);







## *Висновки*

За результатом перевірки даних на рівність дисперсії тестом Лівеня був отриманий результат F < F табличне, тому приймаємо нульову гіпотезу. В результаті перевірки відмінності між двома групами за допомогою критерію Стюдента для незалежних вибірок був отриманий результат t > t табличне.

Отже, відхиляємо нульову гіпотезу та робимо висновок про статичну відмінність середніх.

# **РОЗДІЛ 8. ДИСПЕРСІЙНИЙ АНАЛІЗ**

## *Теоретичні відомості*

Дисперсійний аналіз - це метод статистичної оцінки надійності проявлення залежності результативної ознаки від одного або кількох факторів. За допомогою методу дисперсійного аналізу проводиться перевірка статистичних гіпотез відносно середніх в кількох генеральних сукупностях, які мають нормальний розподіл. Дисперсійний аналіз - параметричний метод.

За допомогою дисперсійного аналізу вирішуються три основних завдання:

* загальна оцінка істотності відмінностей між груповими середніми;
* оцінка вірогідності взаємодії факторів;
* оцінка істотності відмінностей між парами середніх.

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;

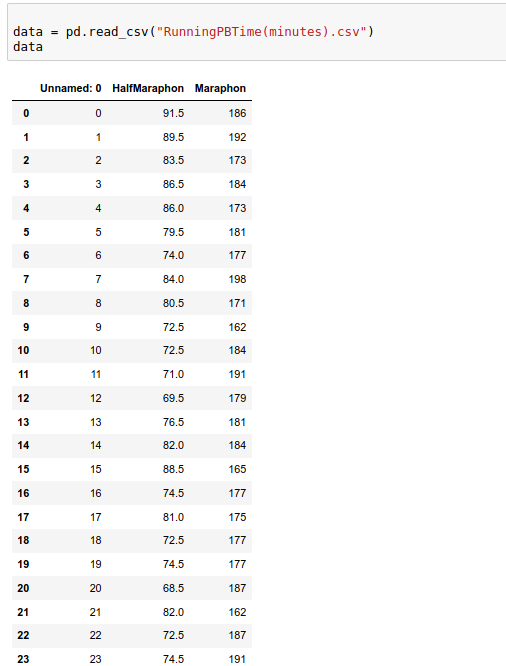
5) оцінка вірогідності різниці між середніми і формулювання висновків з перевірки нульової гіпотези.

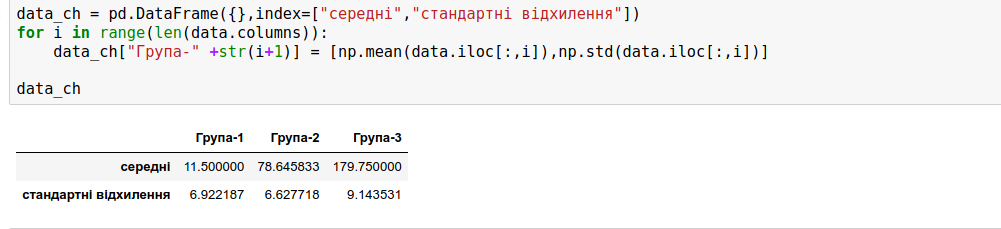
Гіпотеза, яка перевіряєть даним тестом, може бути виражена наступним чином:

𝐻0:μ1=μ2=...=μn («середні незалежних груп рівні»)

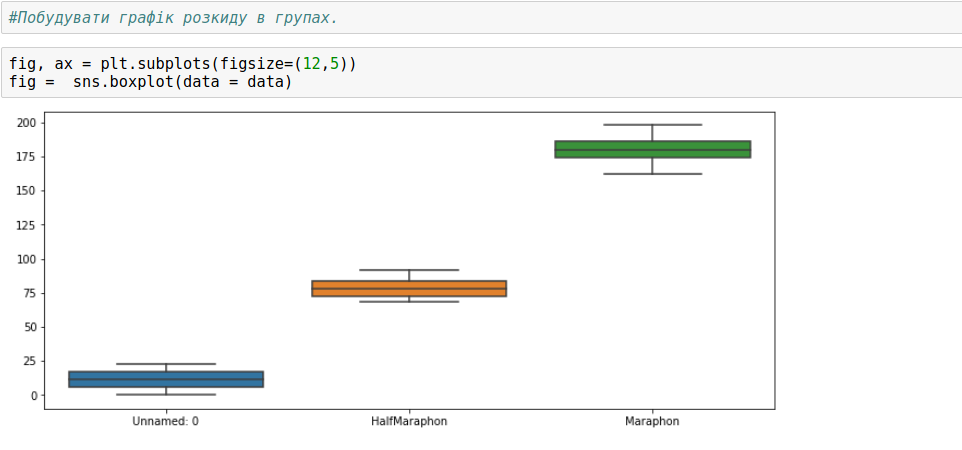
𝐻1:μ1≠μ2≠...≠μn («середні незалежних груп відмінні»)

## *Практична реалізація*

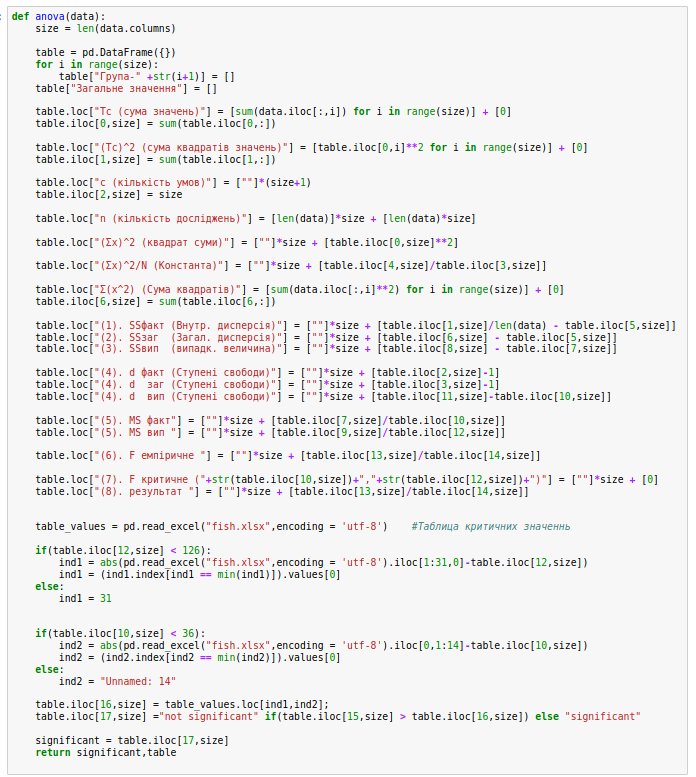
1. Обрати реальні дані з kaggle або сгенерувати випадкові дані. Дані повинні відповідати вимогам
2. 
3. Розрахунок та виведення у вигляді таблиці статистичних характеристик даних (середні та стандартні відхилення для кожної з груп)



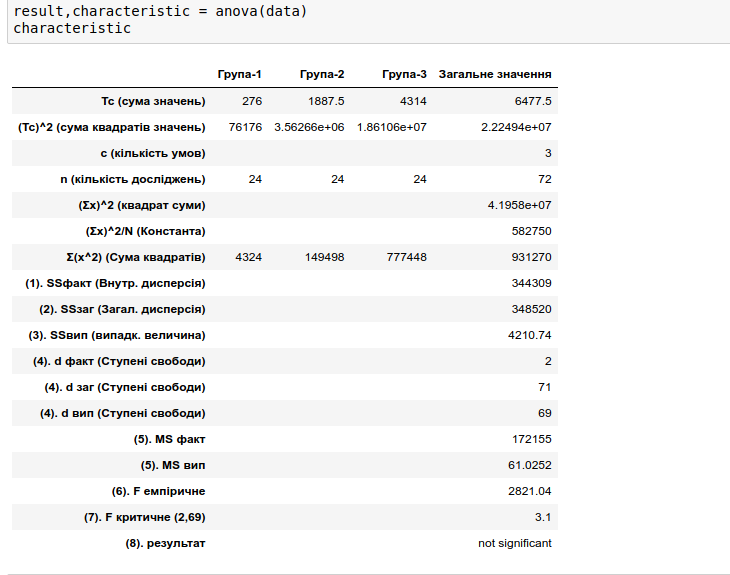
1. Будування графіку розкиду в групах



1. Будування функції, що реалізовуює однофакторний дисперсійний аналіз



1. Виведення результату розрахунку кожного с етапів вказаних вище у вигляді таблиці, перевірка результату роботи функції на реальних даних



## *Висновки*

При 𝐹 емпіричне ≥ 𝐹 критичне нульова гіпотеза про рівність середніх відхиляється. 28.21 > 3.1 - рівність середніх не підтверджена. Приймається альтернативна гіпотеза - про не рівність середніх незалежних груп