МІНІСТЕРСТВО ОСВІТИ I НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ   
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ

КАФЕДРА БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

**Розрахунково - графічна робота**

з дисципліни «Обробка біомедичних даних»

Варіант №14

**Виконав:**

студент гр. БС-83

Розмариця О.А.

Зараховано від \_\_\_.\_\_\_.\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис викладача)

Київ-2020

# 1. Основи роботи з Jupyter Notebook

## *1.1. Теоретичні відомості*

Jupyter Notebook - неймовірно потужний інструмент для інтерактивної розробки та подання проектів в області наук про дані. Найпростіше почати роботу з Jupyter Notebooks, встановивши дистрибутив Anaconda. Anaconda є найбільш широко використовуваним дистрибутивом Python для роботи з даними і поставляється з попередньо встановленими найбільш популярними бібліотеками та інструментами.

Кожен файл .ipynb є текстовим файлом, який описує вміст записної книжки у форматі JSON. Кожна ланка й її вміст, включаючи вкладення зображень, які були перетворені в рядки тексту, перераховані в ньому разом з деякими метаданими.

Cells (комірки) і kernels (ядра) є ключем як до розуміння Jupyter, так і до того, що робить його не просто текстовим процесором. Kernel (Ядро) - це «обчислювальний движок», який виконує код, що міститься в документі ноутбука. Cell (комірка) - це контейнер для тексту, який буде відображатися в записнику, або код, який буде виконуватися ядром записної книжки. Розрізняють такі типи комірок : комірка коду, що містить код, який повинен бути виконаний в ядрі, і відображає його висновок нижче та комірка Markdown, що містить текст, відформатований з використанням Markdown, і відображає його висновок на місці під час запуску.

Зазвичай починають з комірки коду, спеціально призначеної для імпорту і налаштування, тому, якщо потрібно додати або змінити що-небудь, можна просто відредагувати і повторно запустити комірку, не викликаючи побічних ефектів.

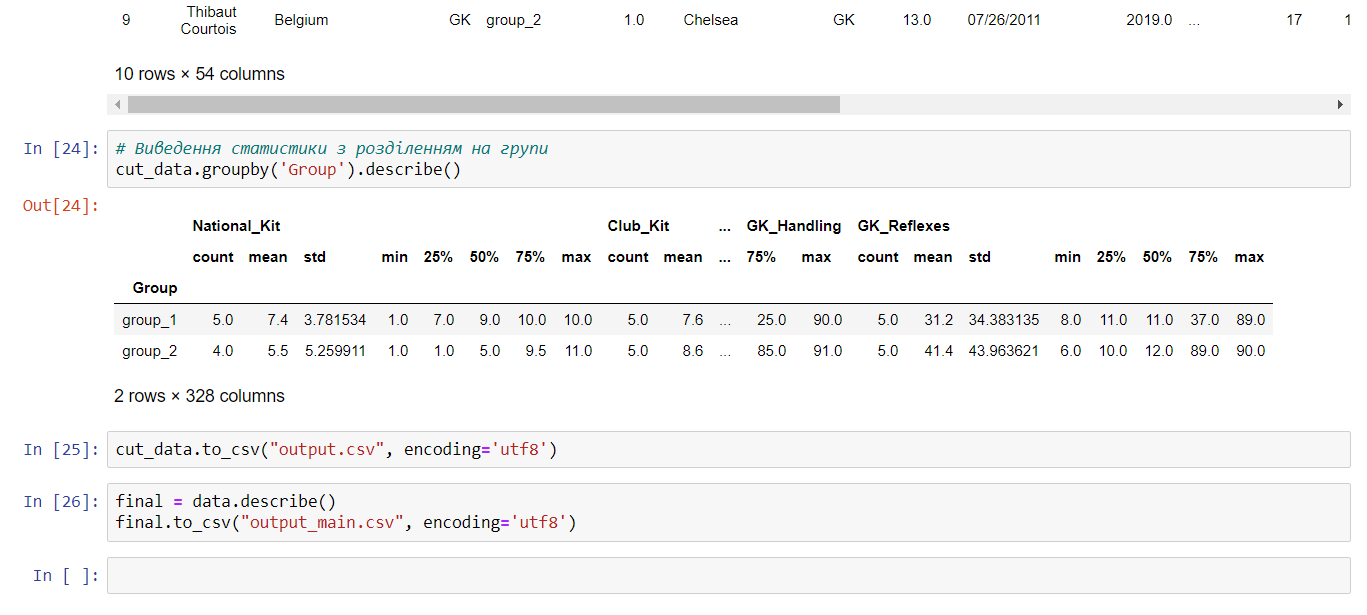
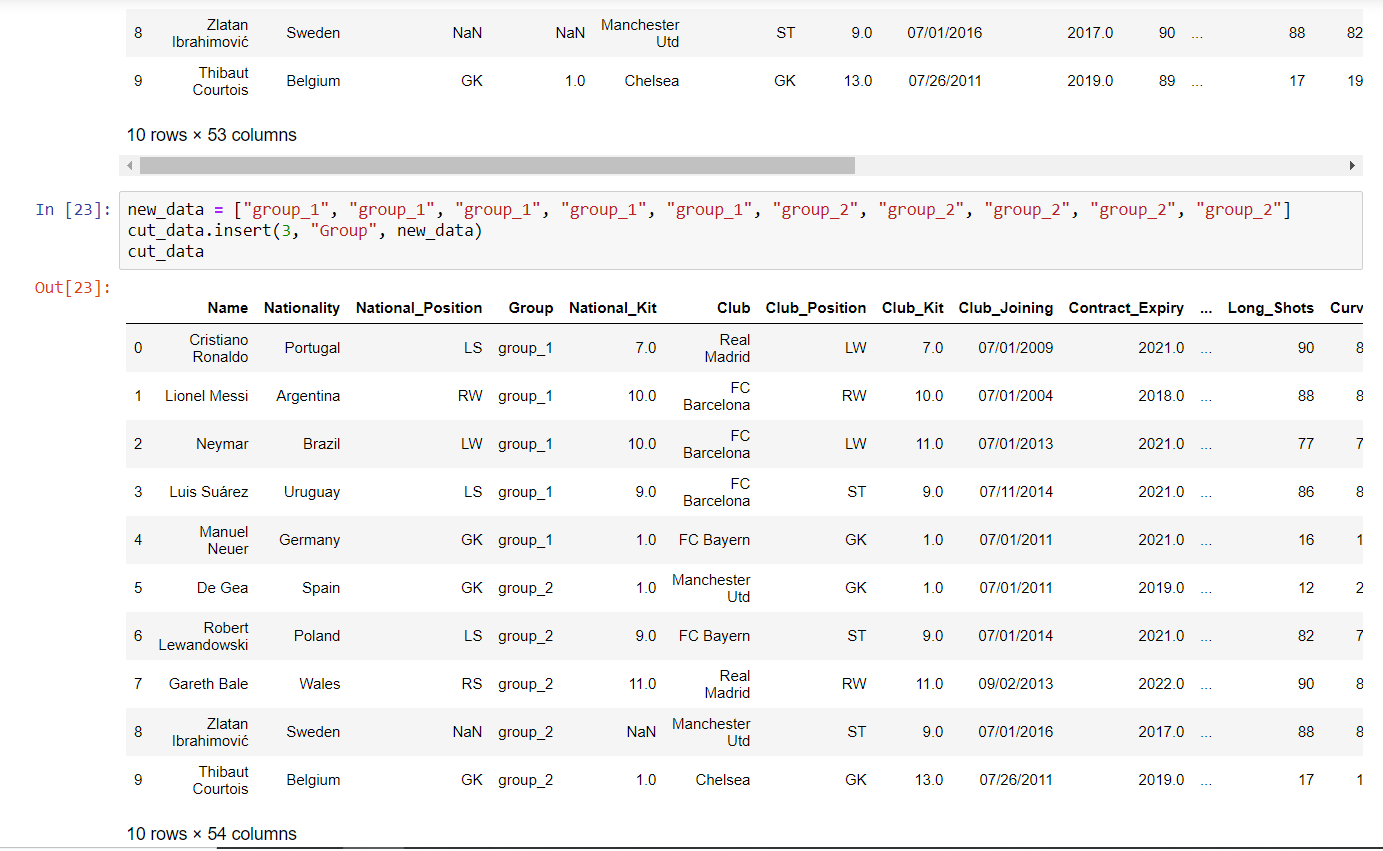
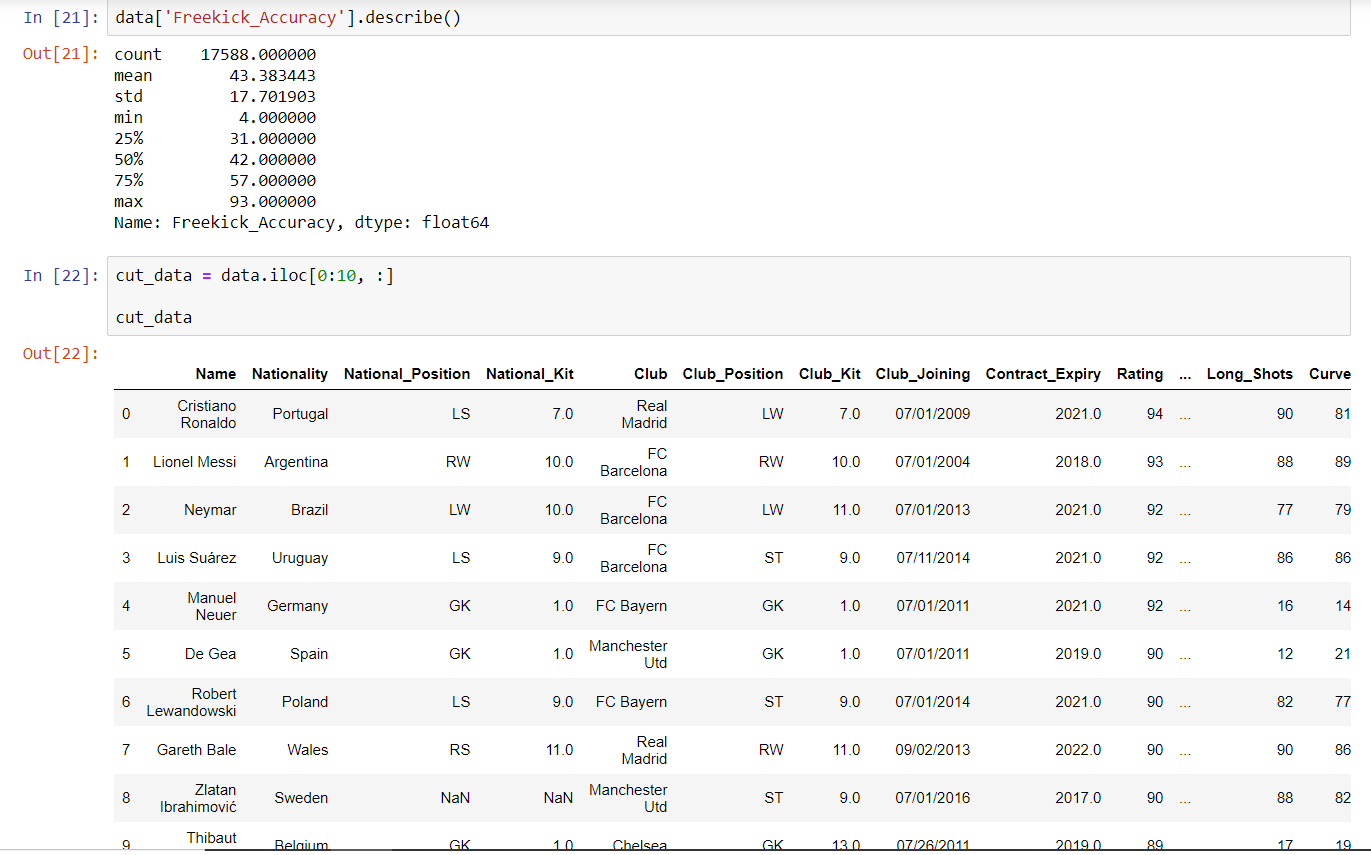
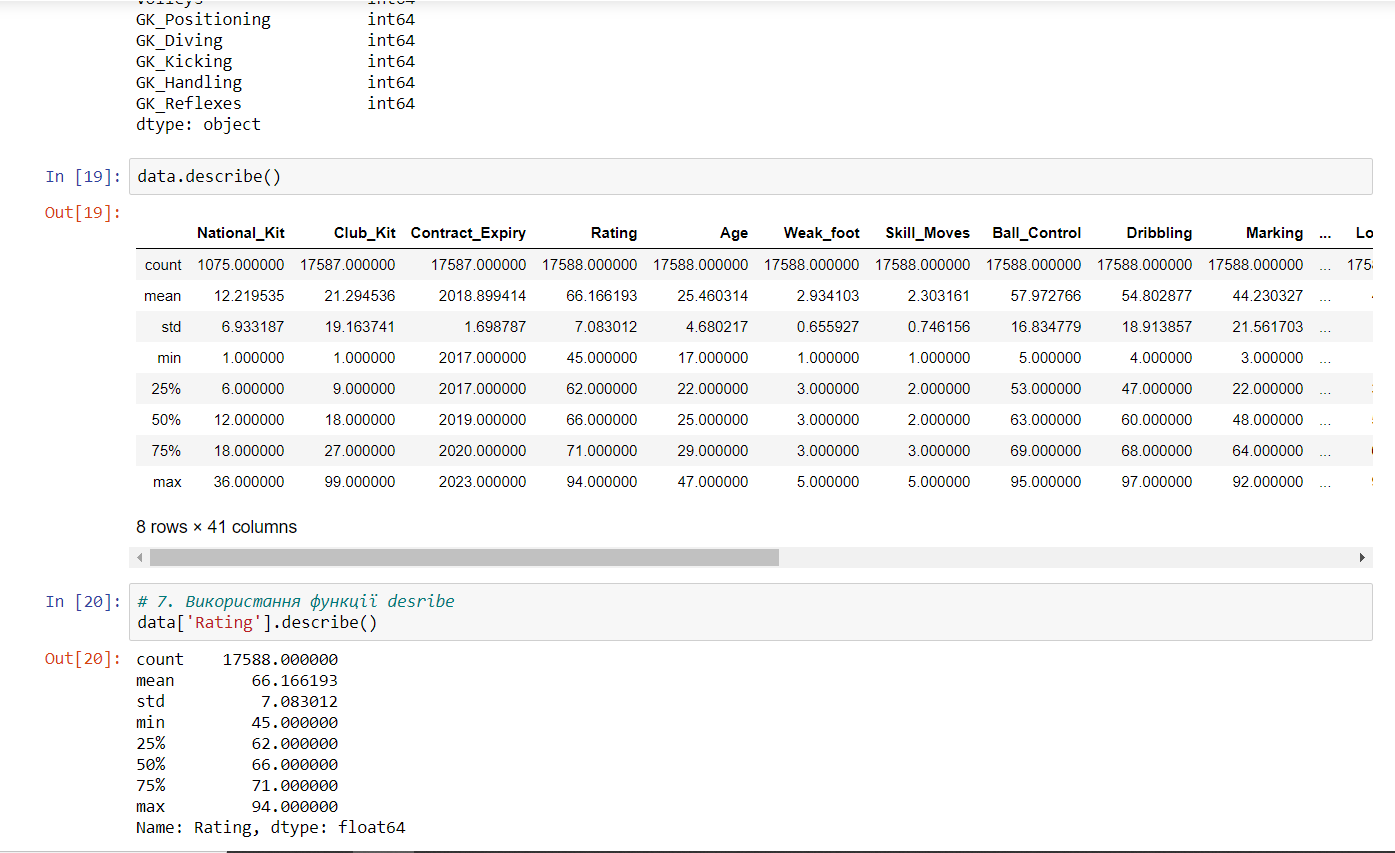
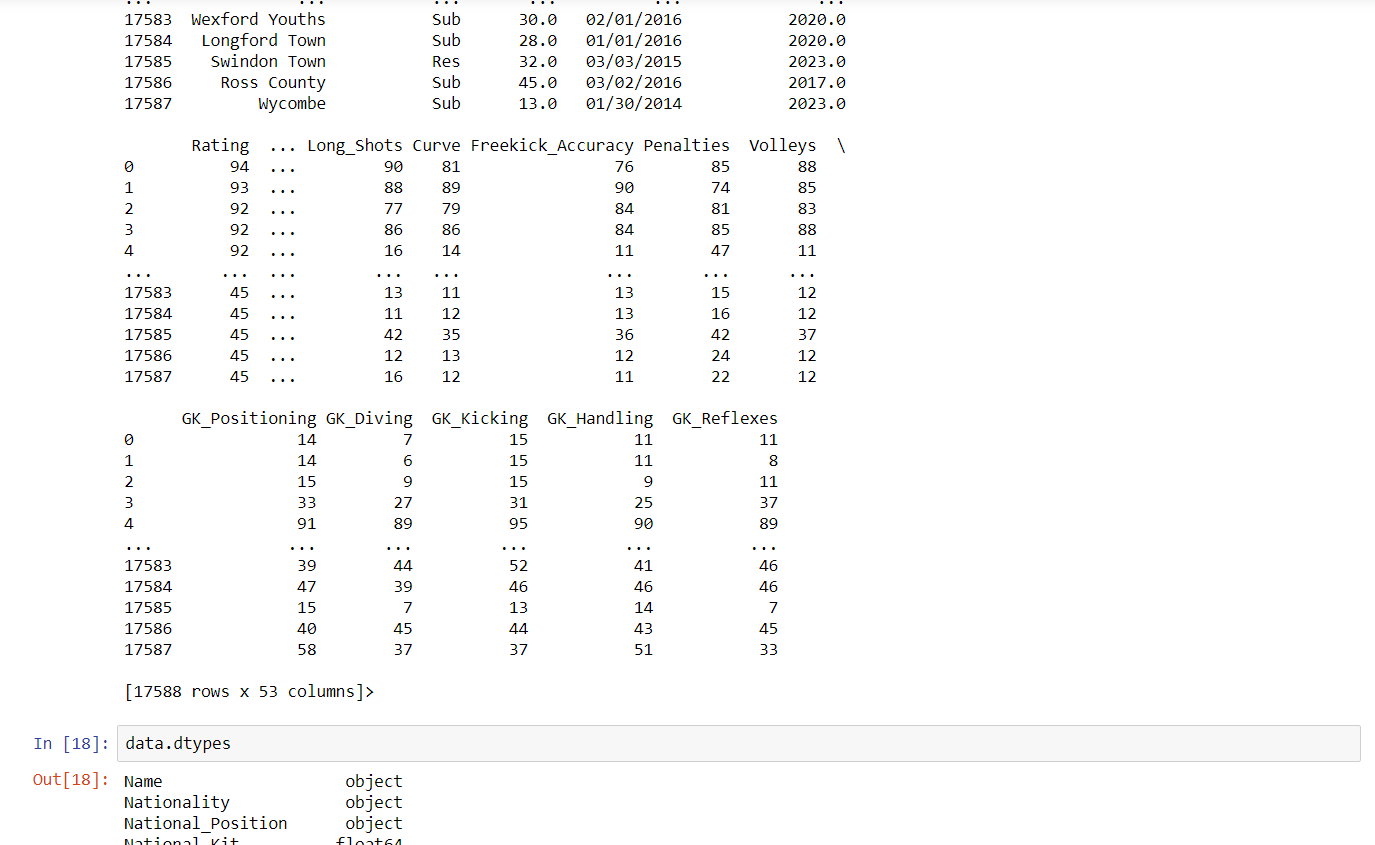
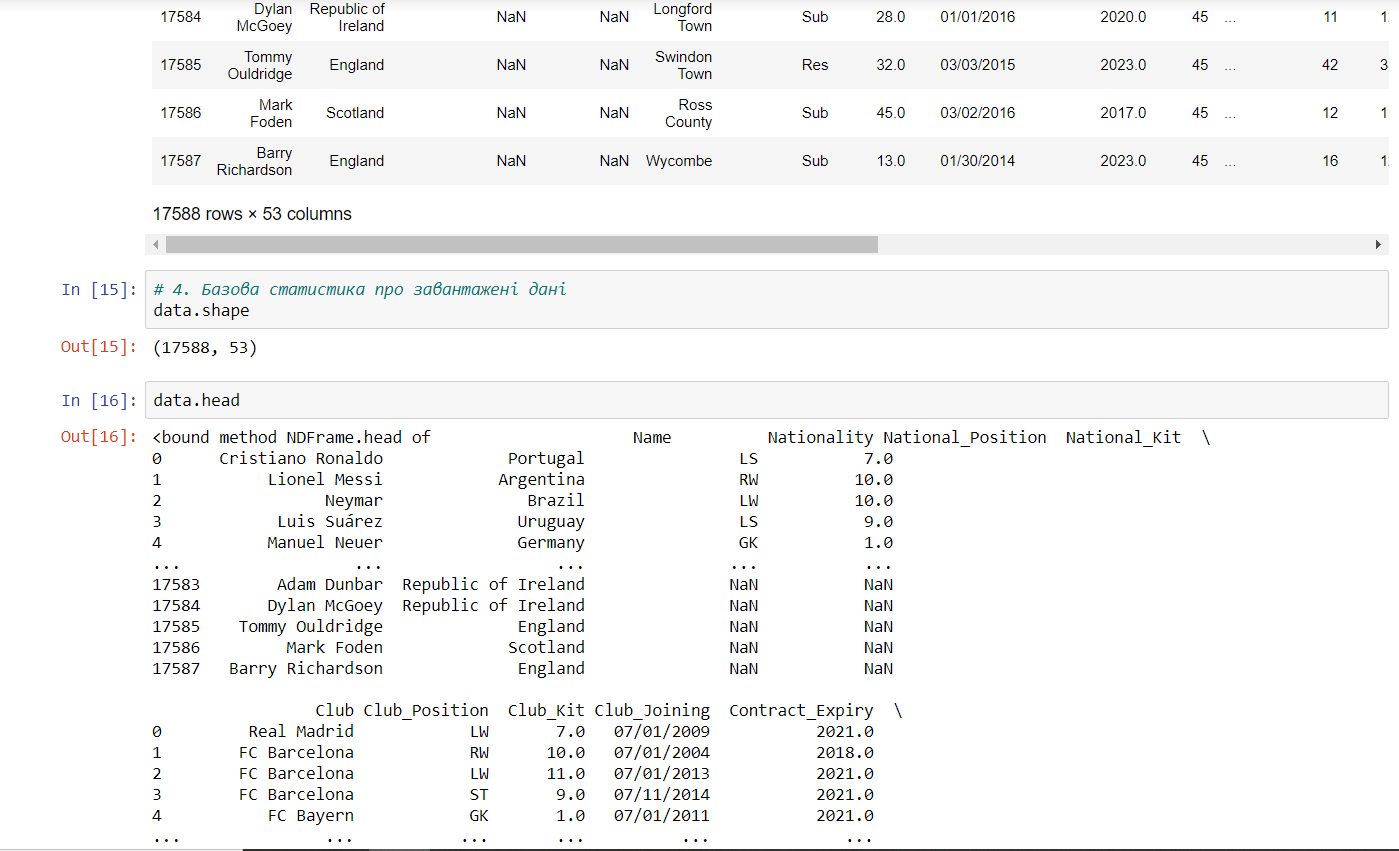
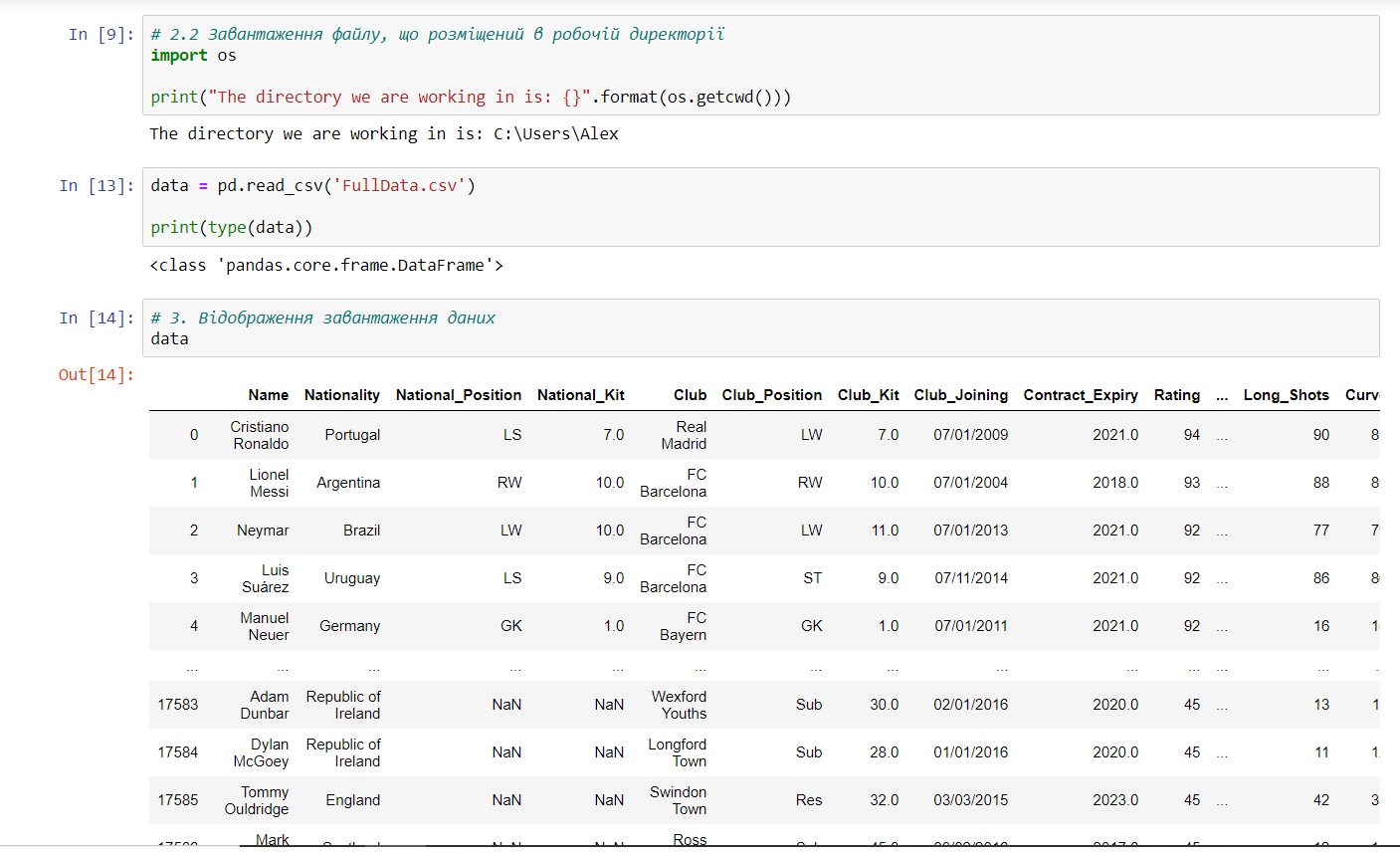
Pandas - одна з найпопулярніших бібліотек Python для аналітики і роботи з Data Science. Це як SQL для Python. Все тому, що pandas дозволяє працювати з двомірними таблицями даних в Python. DataFrame - двомірна структура, що складається з колонок і рядків. У колонок є імена, а у рядків - індекси. Для завантаження .csv файлу з даними в pandas використовується функція read\_csv (). Базовий метод для виведення всіх даних з dataframe на екран. Для цього не доведеться запускати функцію pd.read\_csv () знову і знову. Потрібно просто зберегти дані в змінну при читанні. Іноді зручно вивести не цілий dataframe, заповнивши екран даними, а вибрати кілька рядків. Наприклад, перші 5 рядків можна вивести, набравши: article\_read.head ()-перші 5 рядків dataframe, або останні 5 рядків: article\_read.tail ().

Функція .shape показує, скільки в датафреймі рядків і колонок. Він повертає пару значень (n\_rows, n\_columns). Спочатку йдуть рядки, потім колонки.

Пакет pandas дозволяє збирати загальні статистичні дані за допомогою функції DataFrame.describe (). Без додаткових параметрів функція DataFrame.describe () надасть наступну інформацію для числових типів даних:

|  |  |
| --- | --- |
| count | Підрахунок частоти тієї чи іншої події (скільки разів відбулася подія). |
| mean | Середнє значення. |
| std | Стандартне відхилення (числове значення, яке відображає зміну меж даних). |
| min | Найменше число в наборі даних. |
| 25% | 25-й процентиль. |
| 50% | 50-й процентиль. |
| 75% | 75-й процентиль. |
| max | Максимальна кількість в наборі даних. |

## *1.2. Практична реалізація*



## *1.3. Висновки до розділу*

В ході виконання даної лабораторної роботи було освоєно основи роботи програми Jupyter Notebook. А саме ознайомлення зі структурою та інтерфейсом середовища, вивчення основних методів обробки баз даних та бібліотеки pandas.

**2. Комп'ютерний практикум №2 "Візуалізація даних за допомогою бібліотеки Matplotlib"**

## *2.1. Теоретичні відомості*

Бібліотека matplotlib - це бібіліотека двовимірної графіки для мови

програмування python за допомогою якої можна створювати високоякісні

малюнки різних форматів. Matplotlib є модуль-пакет для python.

Бібліотека matplotlib підтримує два методи побудови графіків:

- об’єктно-орієнтований метод, який полягає у створенні графіків як об'єктів та

роботу з ними як з об'єктами

-функціональний метод, який полягає у простій побудові графіків та їх

елементарному перетворенні.

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.

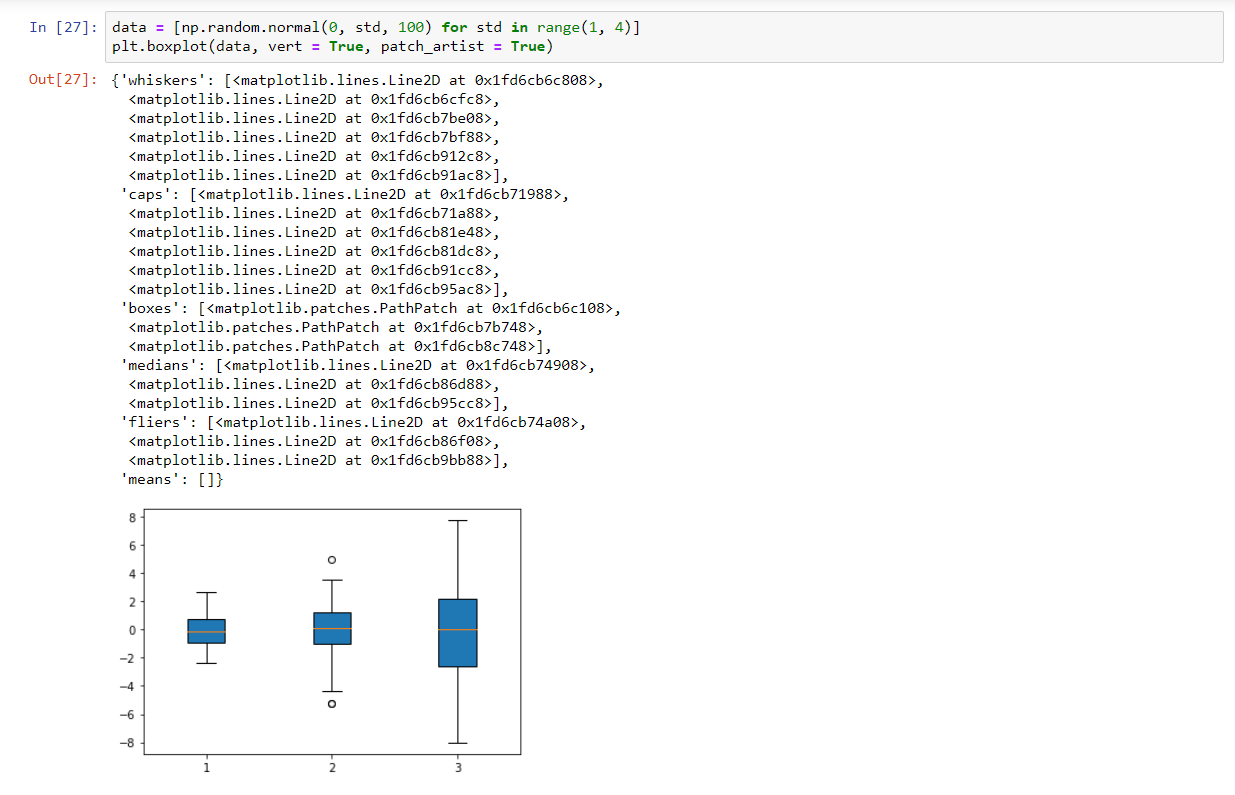
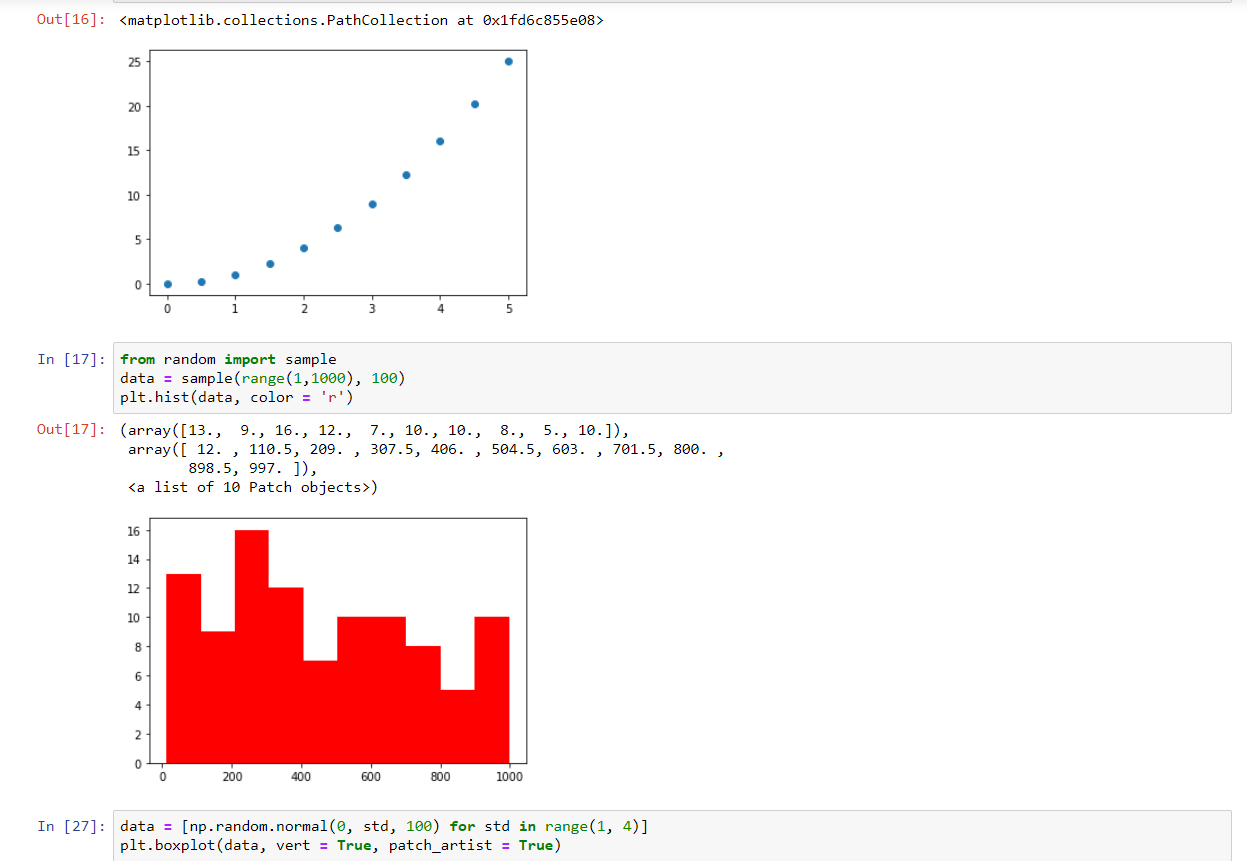
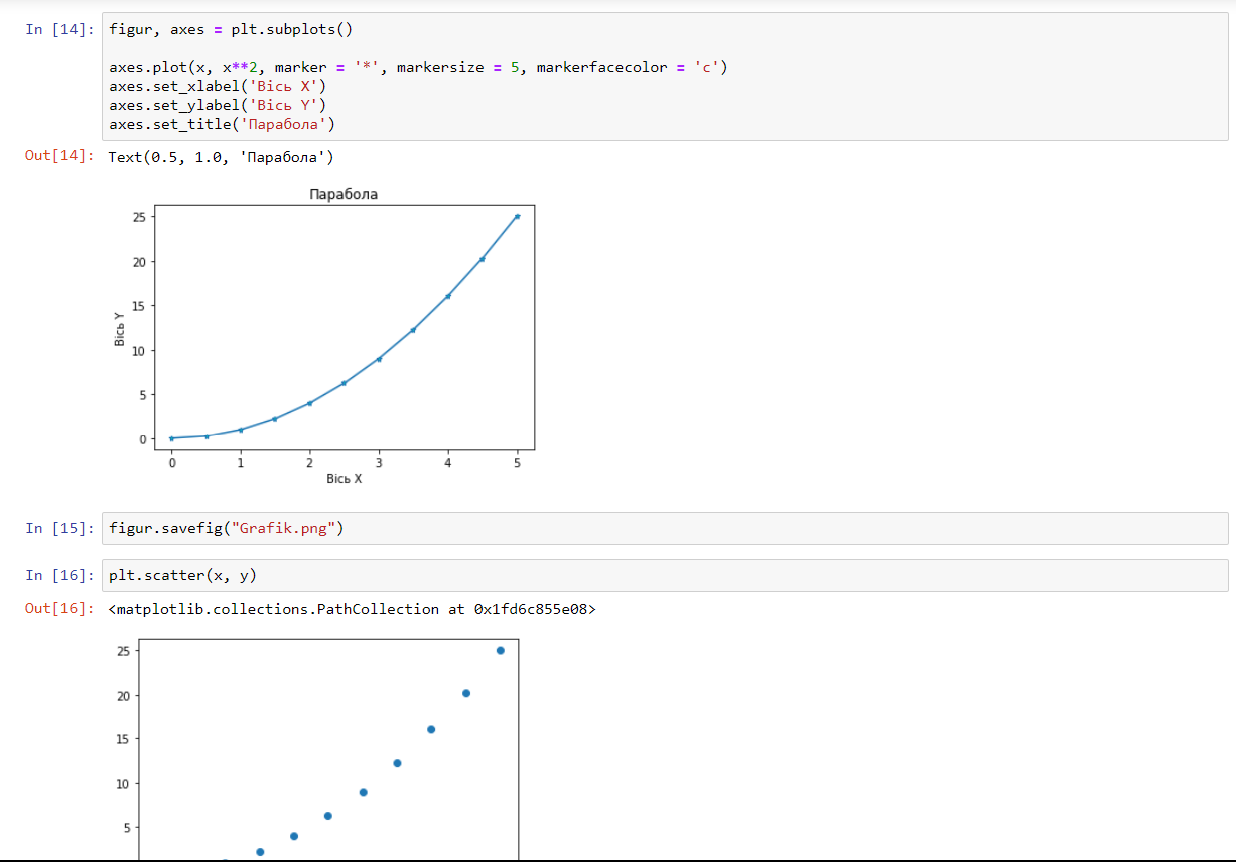
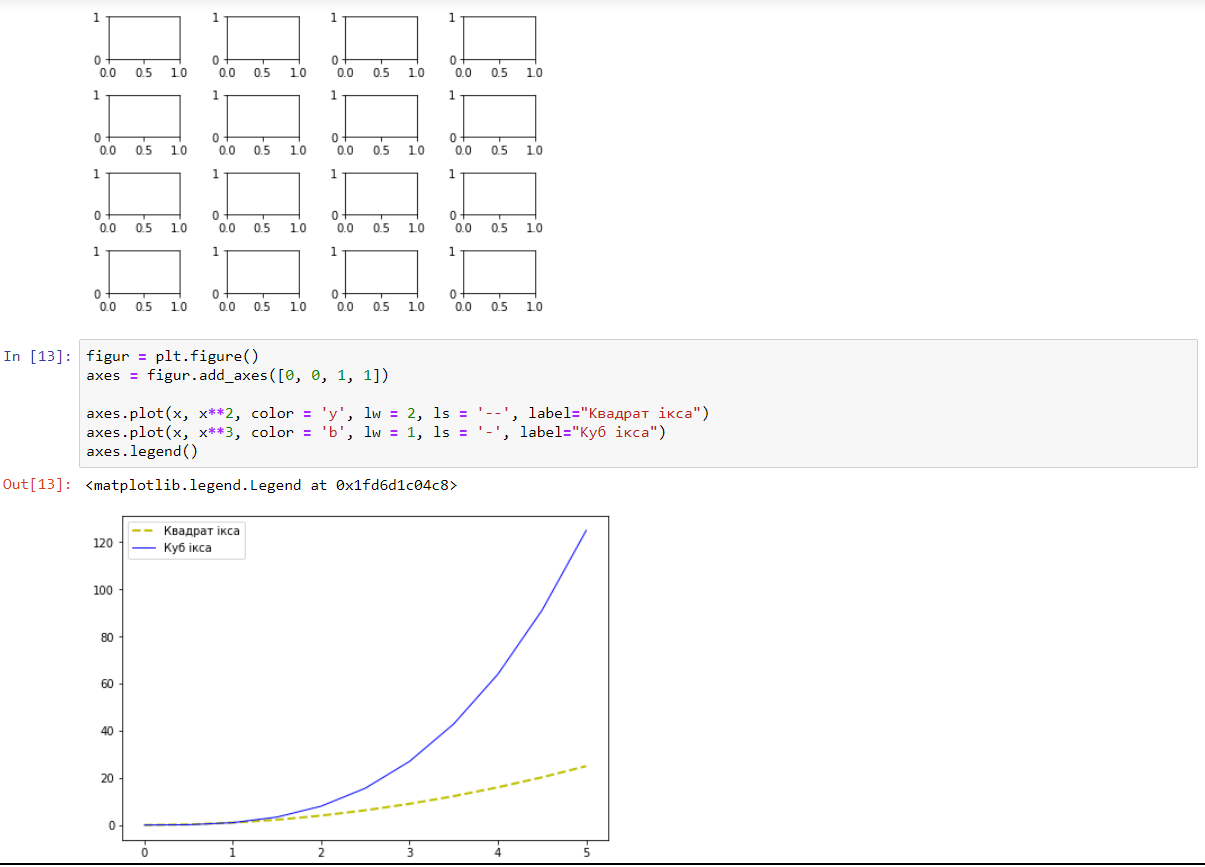
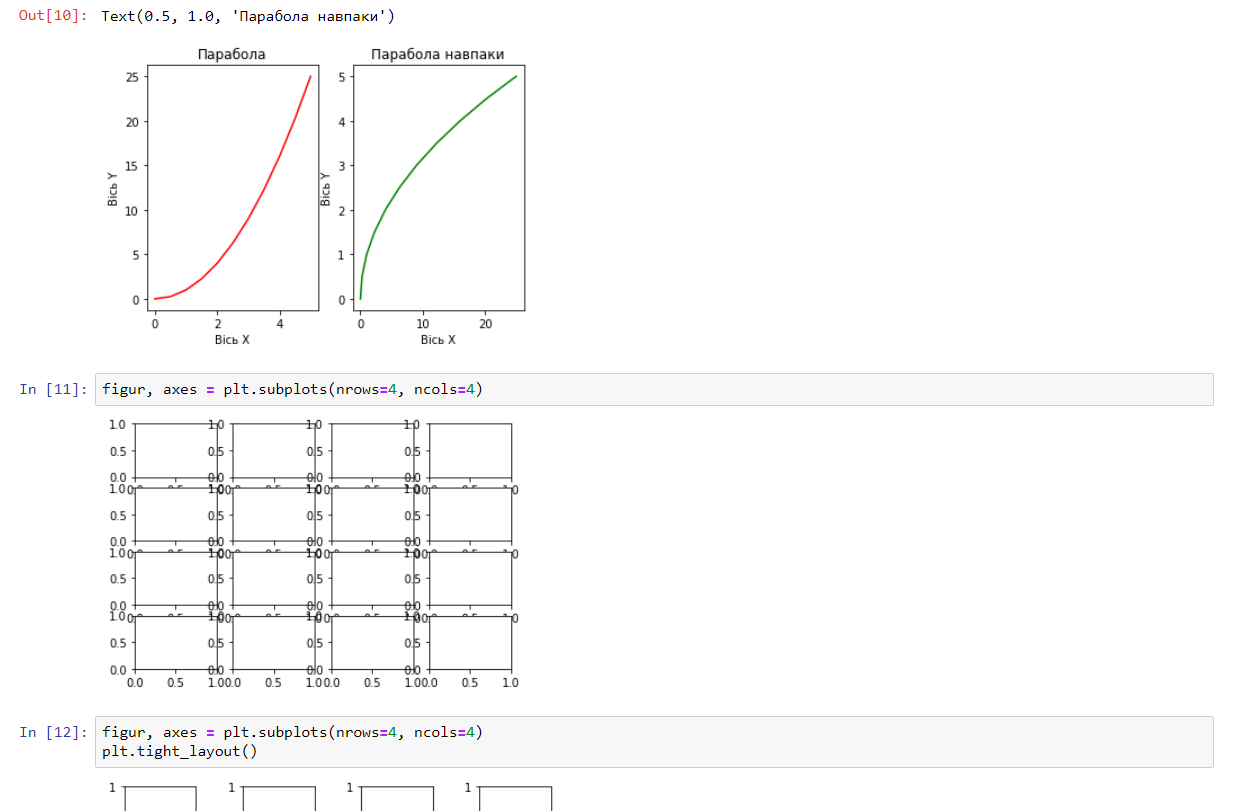
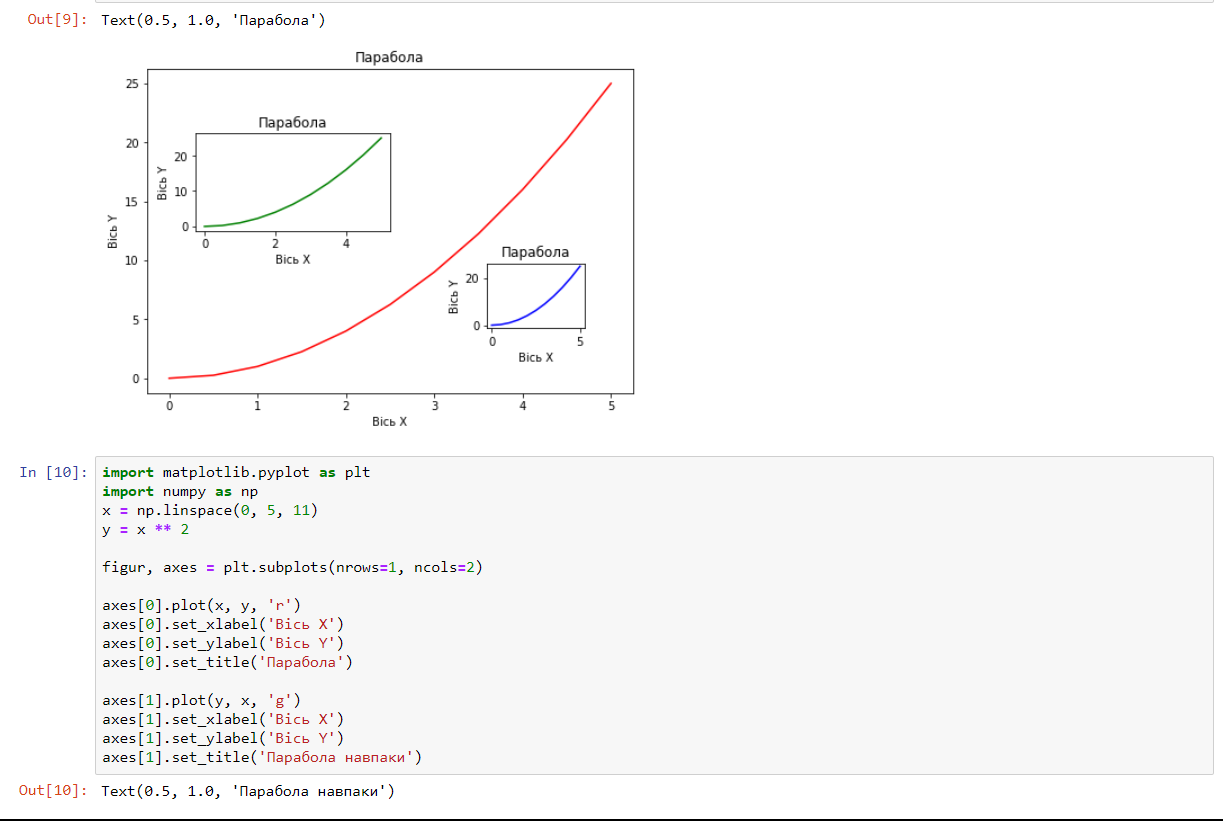
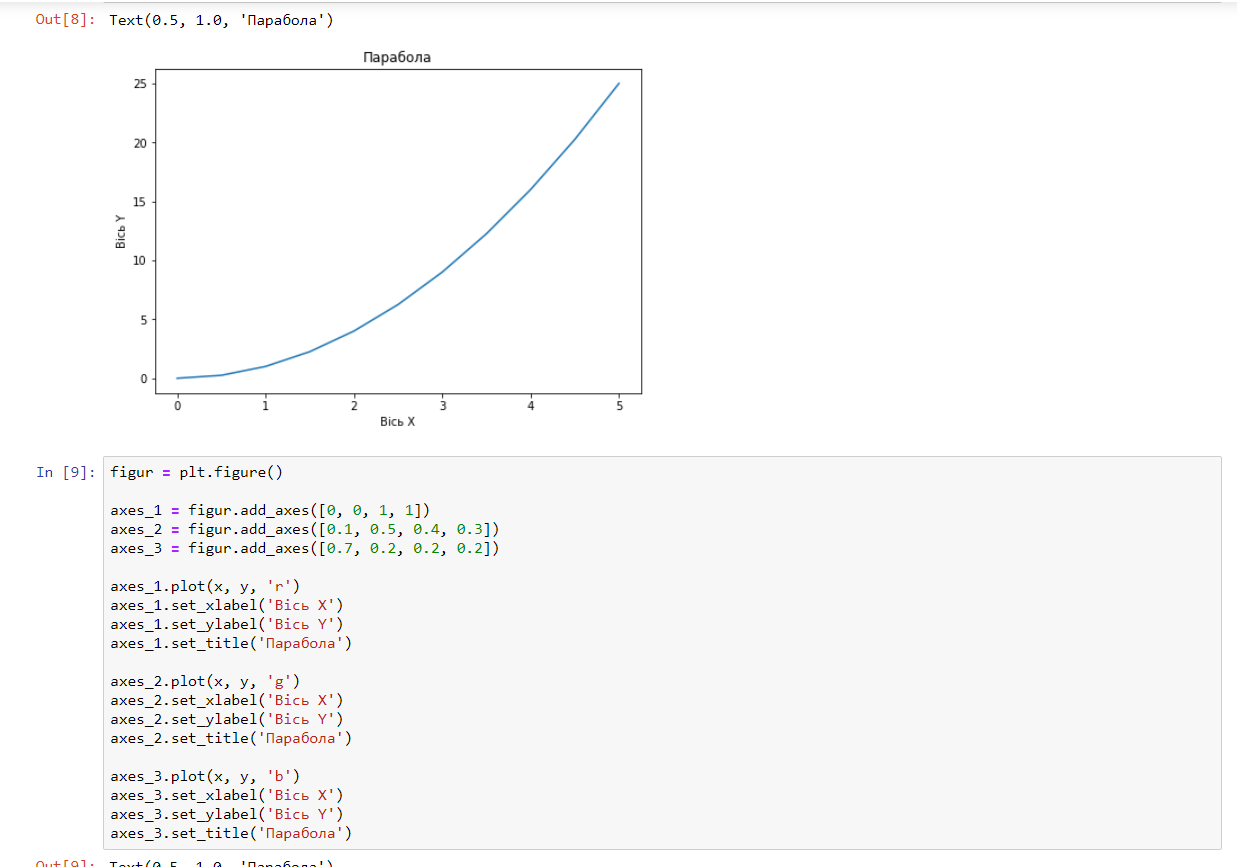
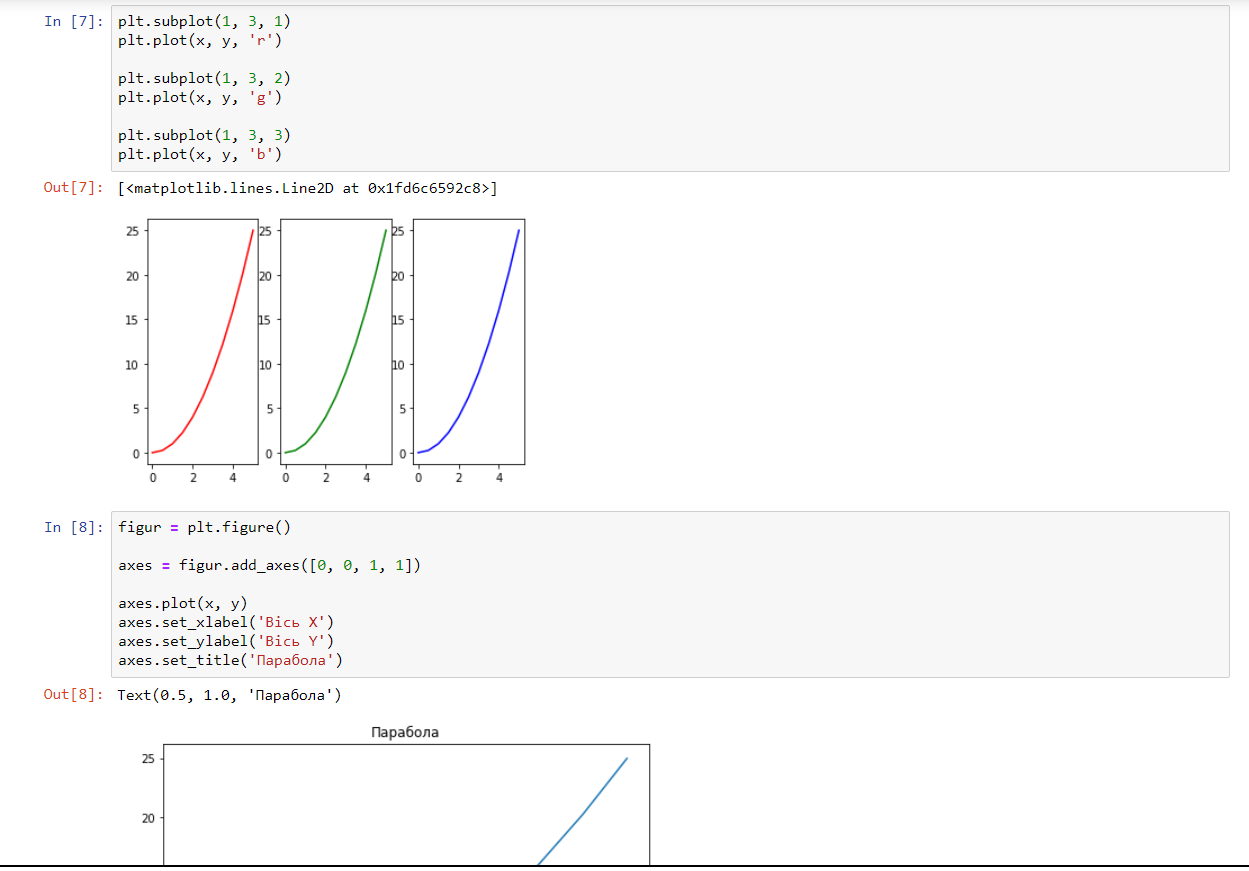
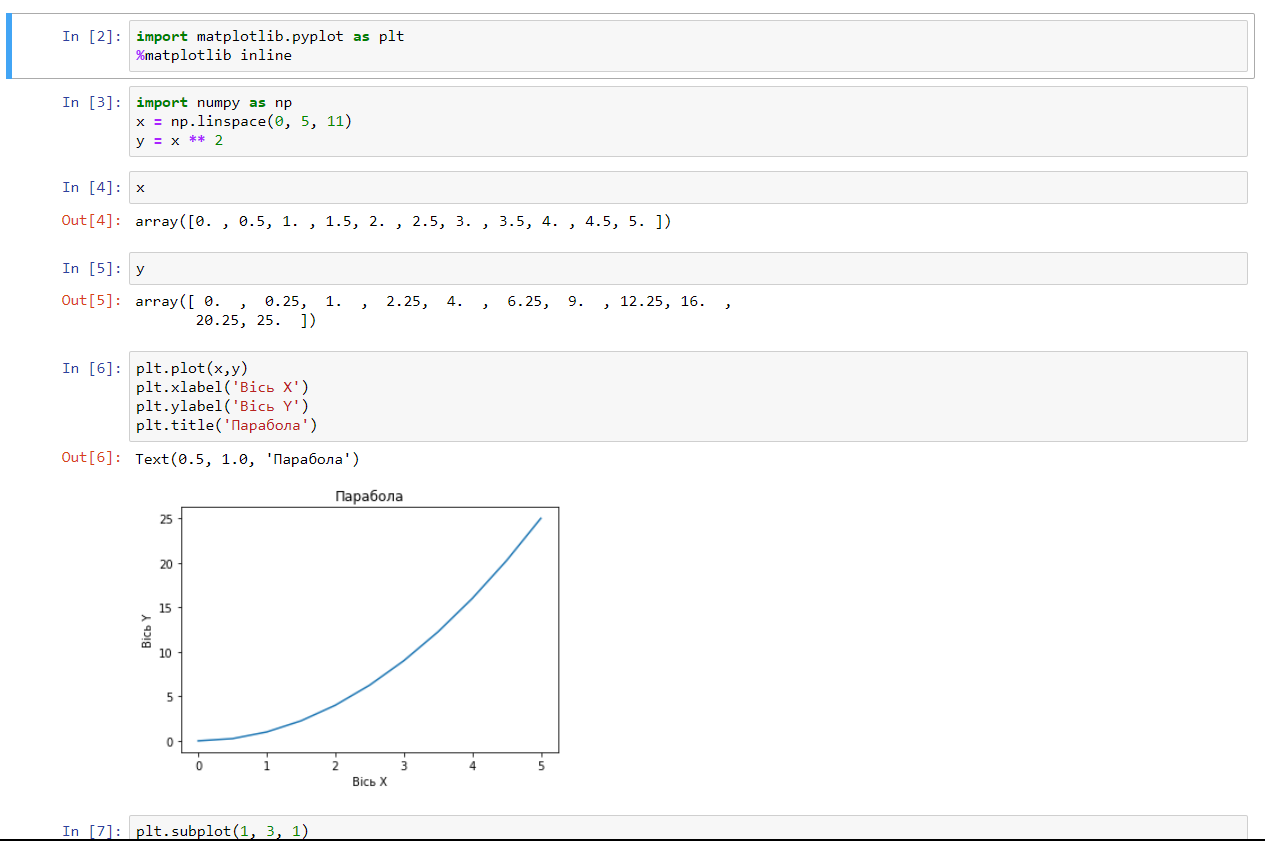
Пакет підтримує багато видів графіків і діаграм:

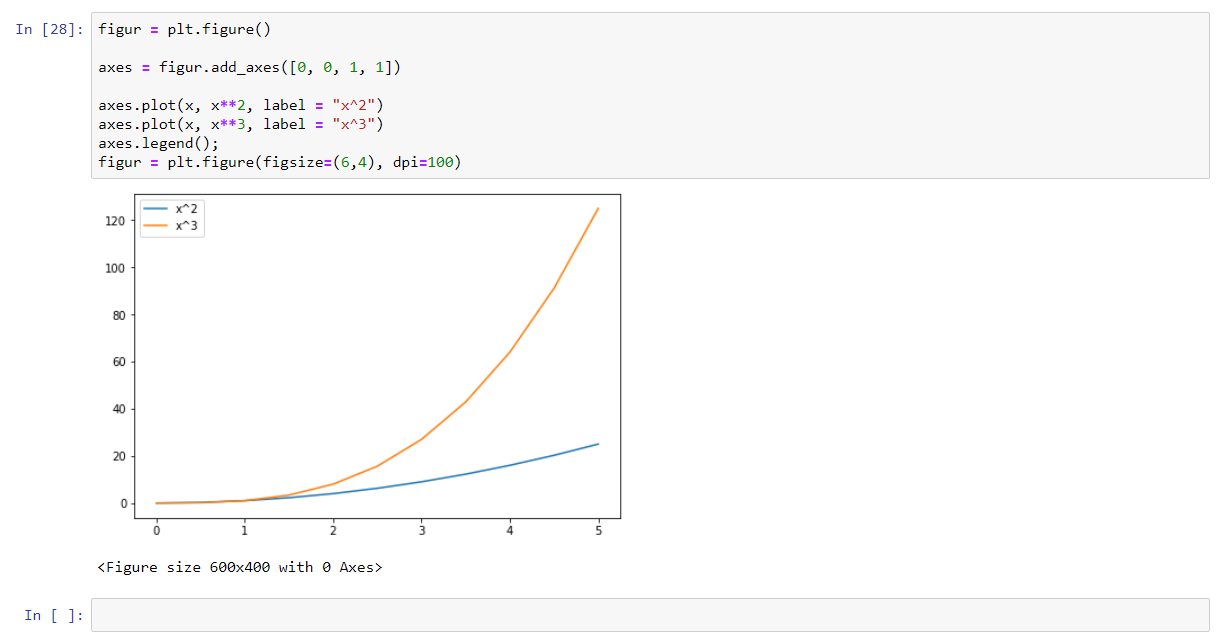
* Графіки (line plot)
* Діаграми розкиду (scatter plot)
* Стовпчасті діаграми (bar chart) і гістограми (histogram)
* Кругові діаграми (pie chart)
* Стовбур-лист діаграми (stem plot)
* Контурні графіки (contour plot)
* Поля градієнтів (quiver)
* Спектральні діаграми (spectrogram)

Користувач може вказати осі координат, грати, додати написи і пояснення,

використовувати логарифмічну шкалу або полярні координати.

## *2.2. Практична реалізація*





## *2.3. Висновки до розділу*

У даному розділі було розглянуто бібліотеку Matplotlib. Вивчено основи роботи з нею, основні види графіків та правила їх побудови. Вивчено деякі методи функціонального та об’єктно-орієнтовного підходу для роботи з графіками за допомогою даної бібліотеки.

# 3. Візуалізація даних за допомогою бібліотеки SEABORN

## *3.1. Теоретичні відомості*

Бібліотека візуалізації даних Seaborn також заснована на Matplotlib.

Вона поставляється з повнофункціональним інтерфейсом для малювання

привабливою і інформативною статистичної графіки. Його API для створення

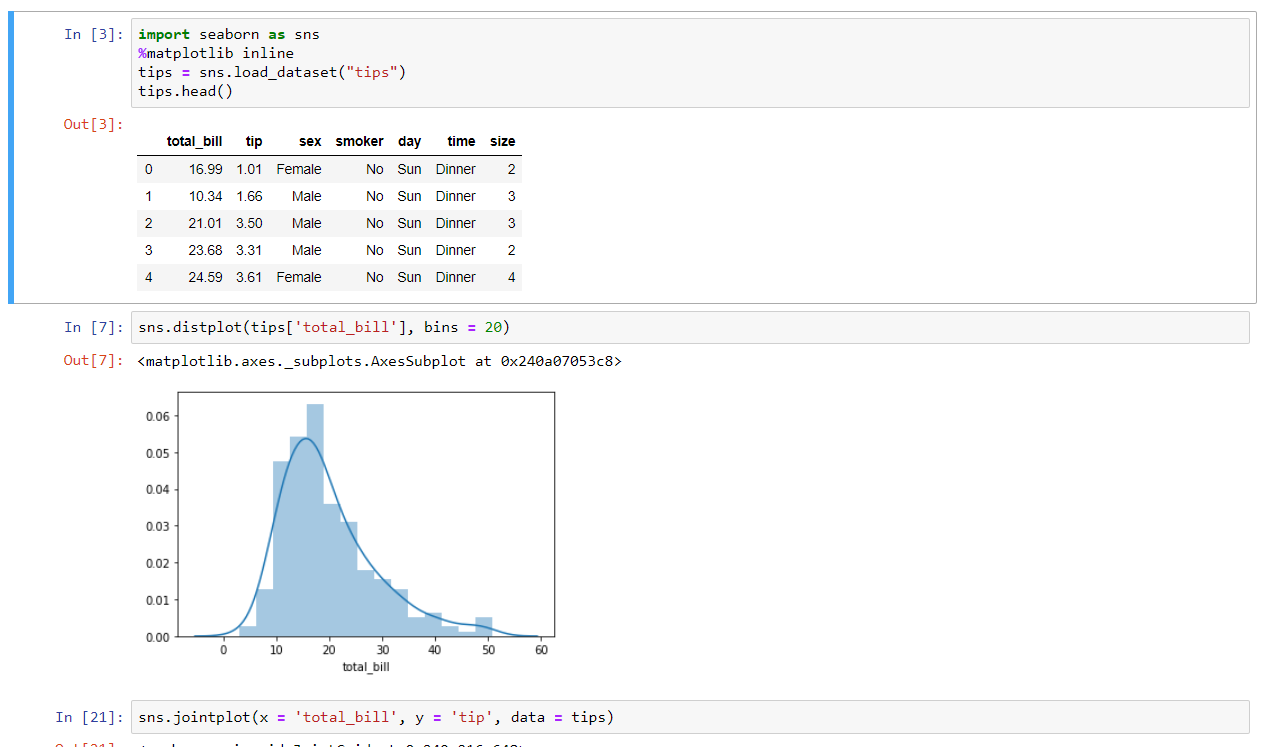
візуалізацій на основі KDE більш лаконічний, ніж у інших схожих бібліотек.

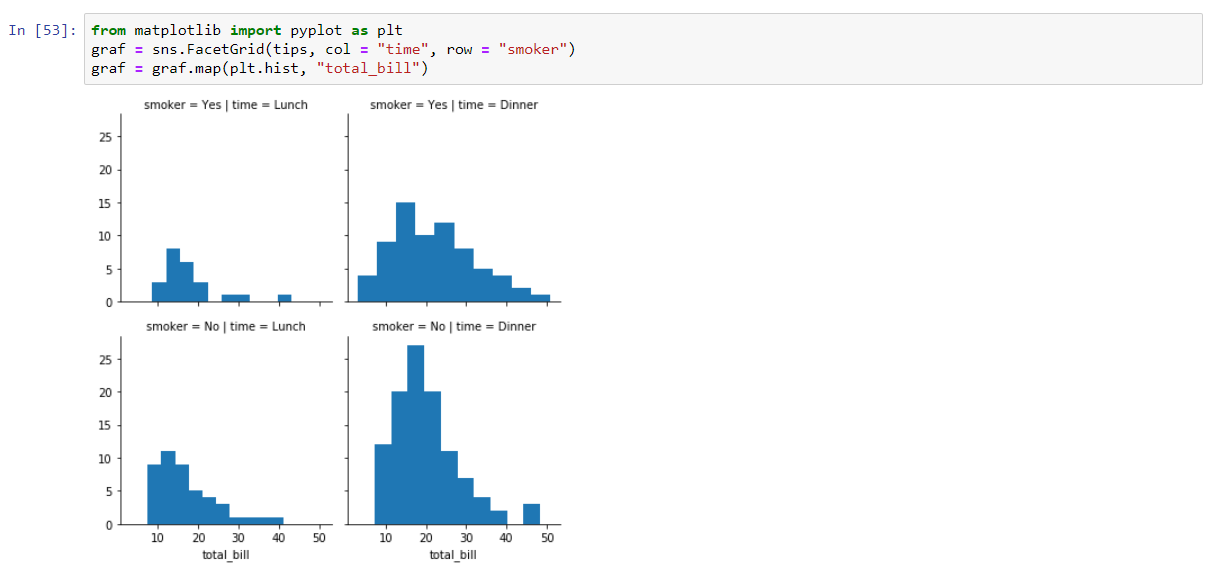
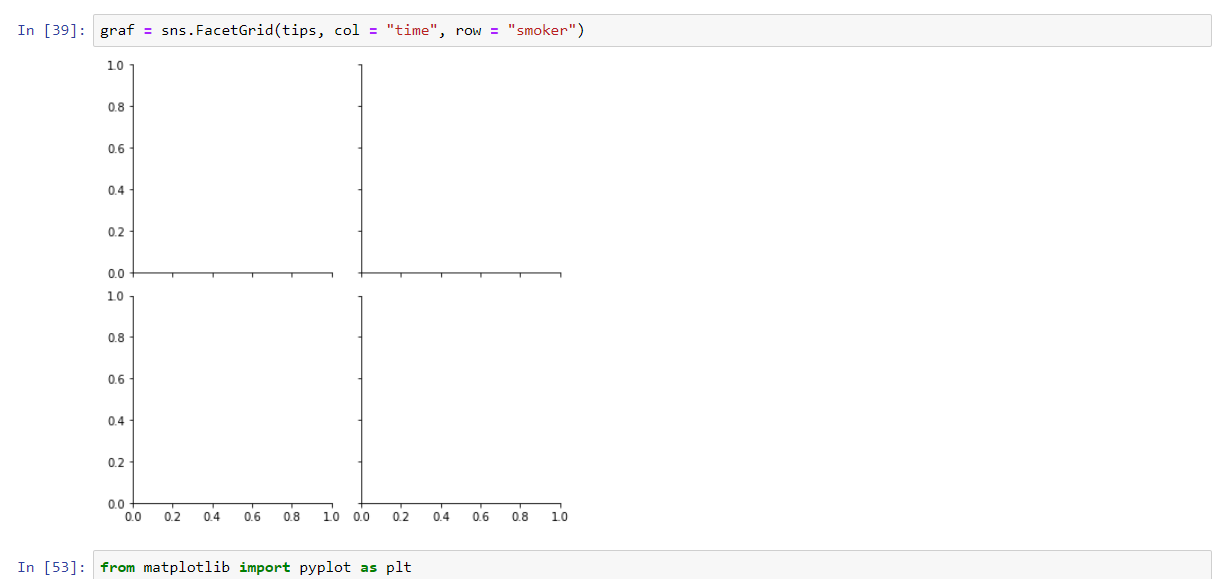
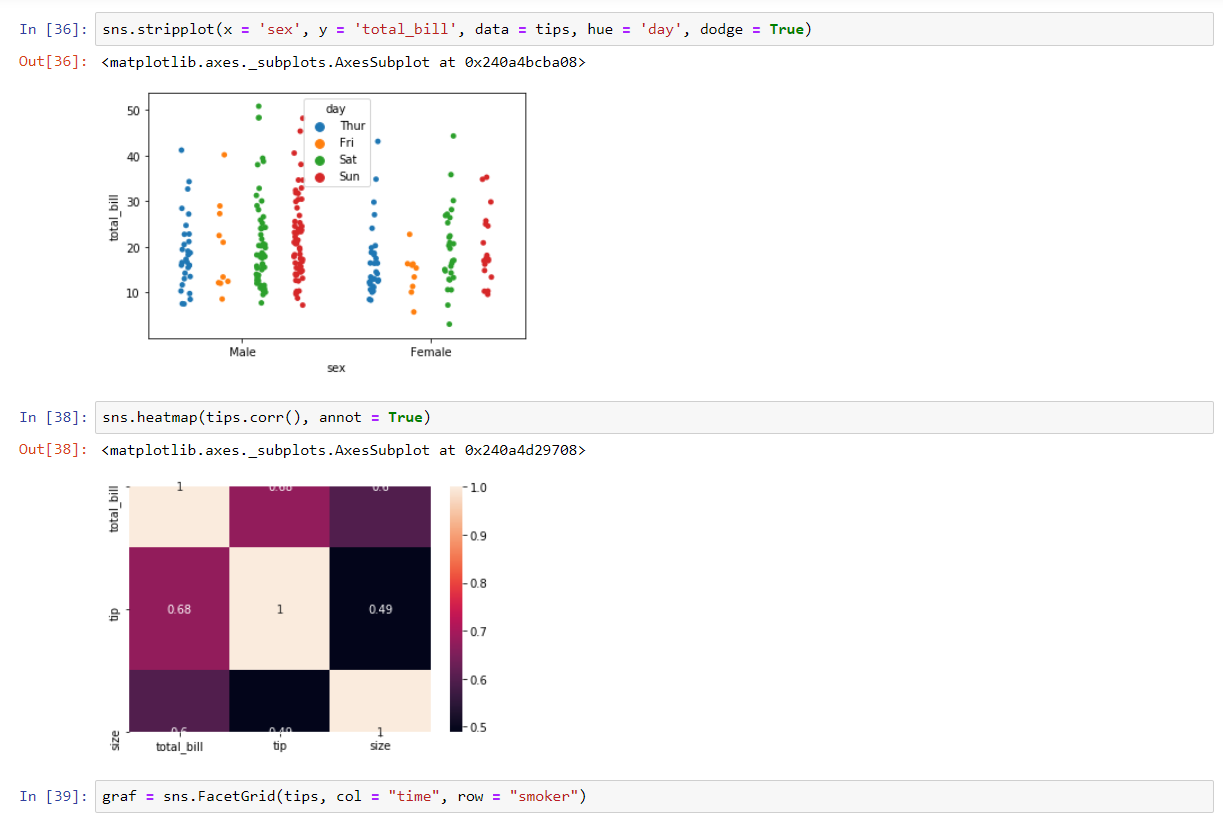
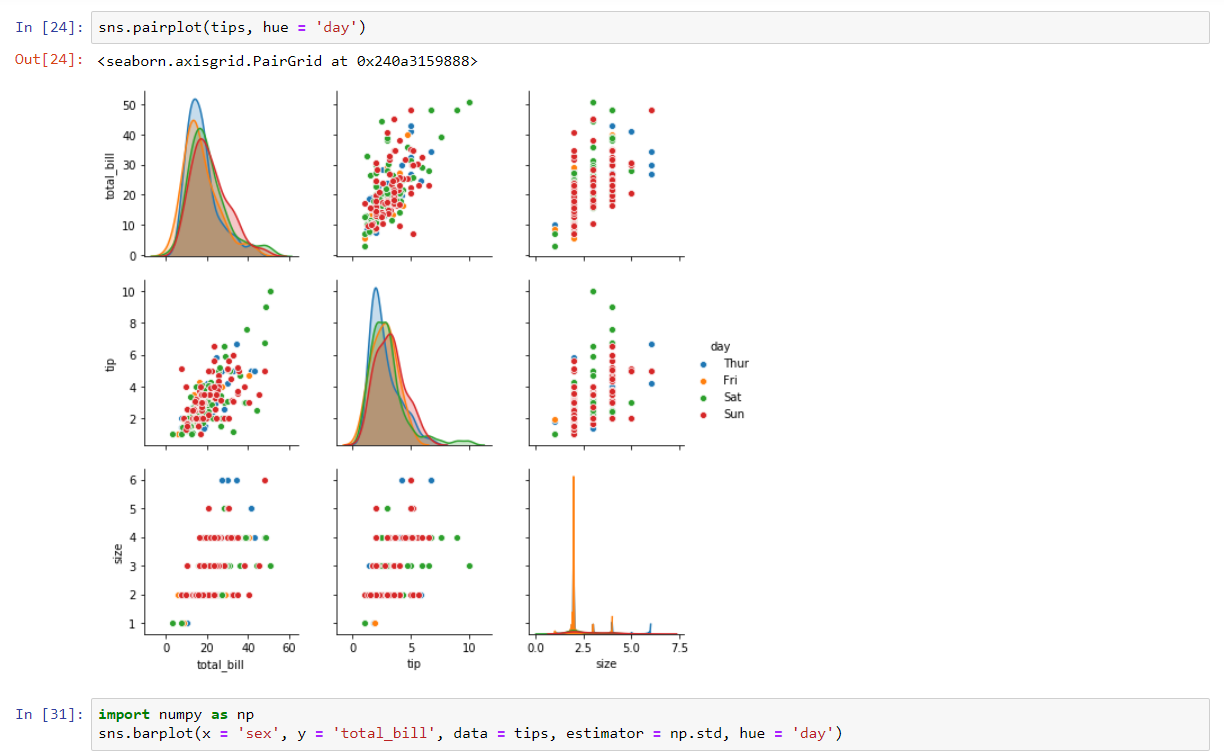
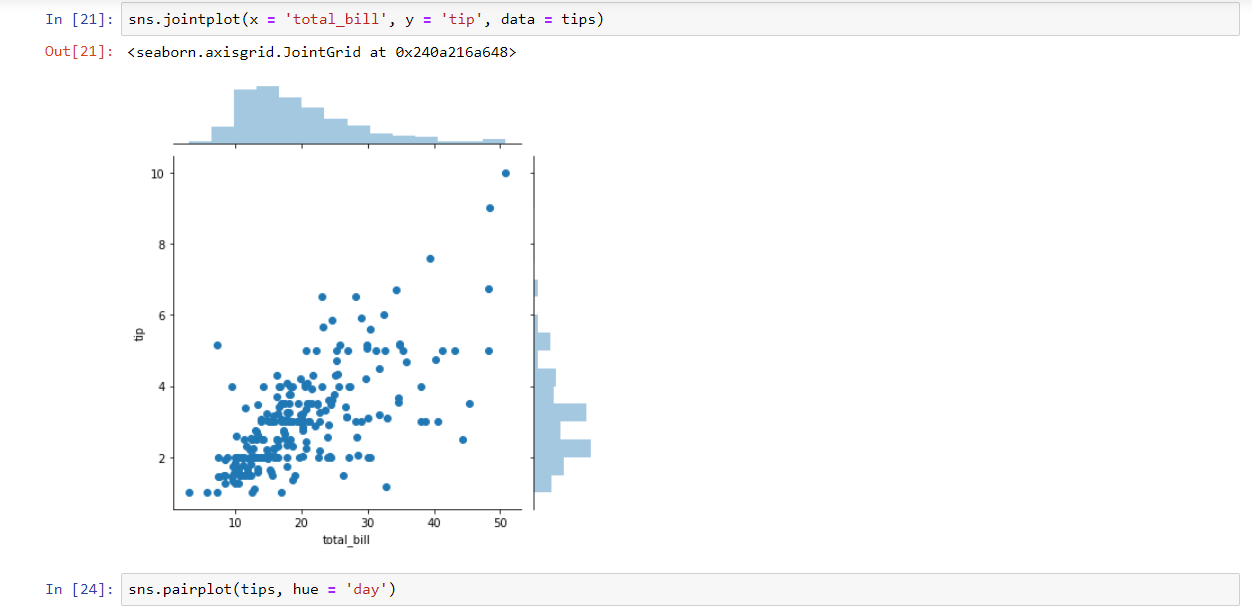
Seaborn старанно працює над тим, щоб візуалізація стала центральною

частиною розуміння і вивчення даних. Seaborn повністю інтегрований зі стеком

PyData, включаючи підтримку структур даних numpy і pandas.

## *3.2. Практична реалізація*





## *3.3. Висновки до розділу*

У даному розділі було розглянуто бібліотеку Seaborn. Вивчено основи роботи з нею, відмінності від бібліотеки Matplotlib. Вивчено деякі методи для роботи з графіками за допомогою даної бібліотеки, та параметри кольорової палітри що використовується бібліотекою Seaborn.

# 4. Перевірка даних на нормальність за допомогою критерію Шапіра-Уілка

## *4.1. Теоретичні відомості*

Критерій Шапіро-Уілка використовується для перевірки гіпотези:

«випадкова величина X розподілена нормально» і є одним найбільш

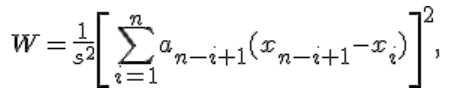
ефективних критеріїв перевірки нормальності. Критерії, перевіряючі

нормальність вибірки, є окремим випадком критеріїв згоди. Якщо вибірка

нормальна, можна далі застосовувати потужні параметричні критерії,

наприклад, критерій Фішера.

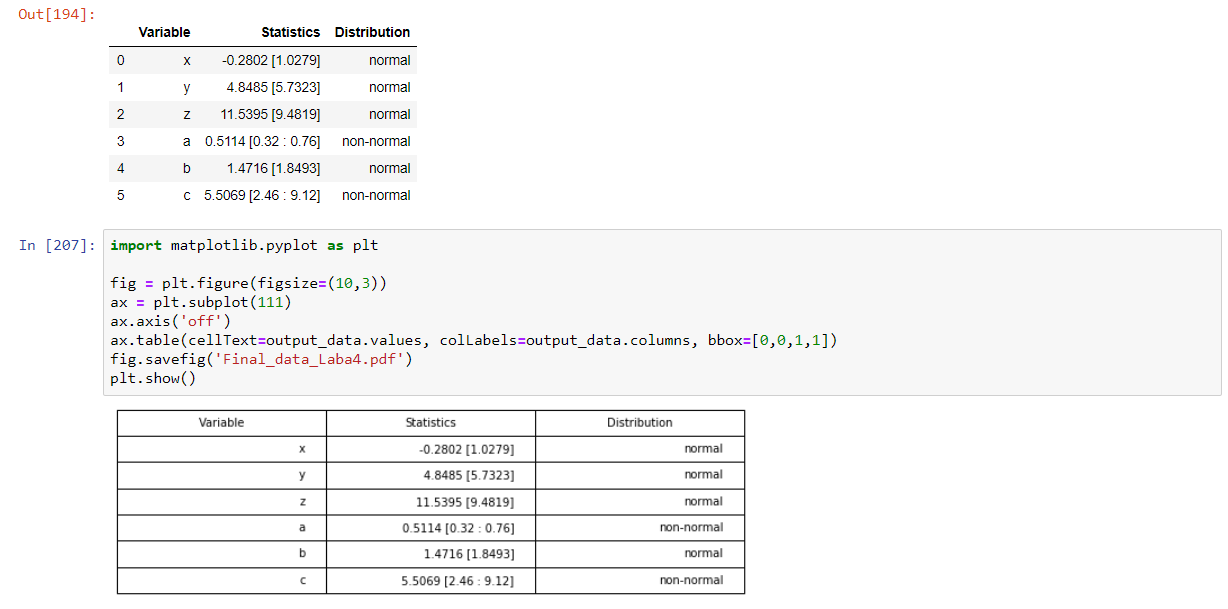
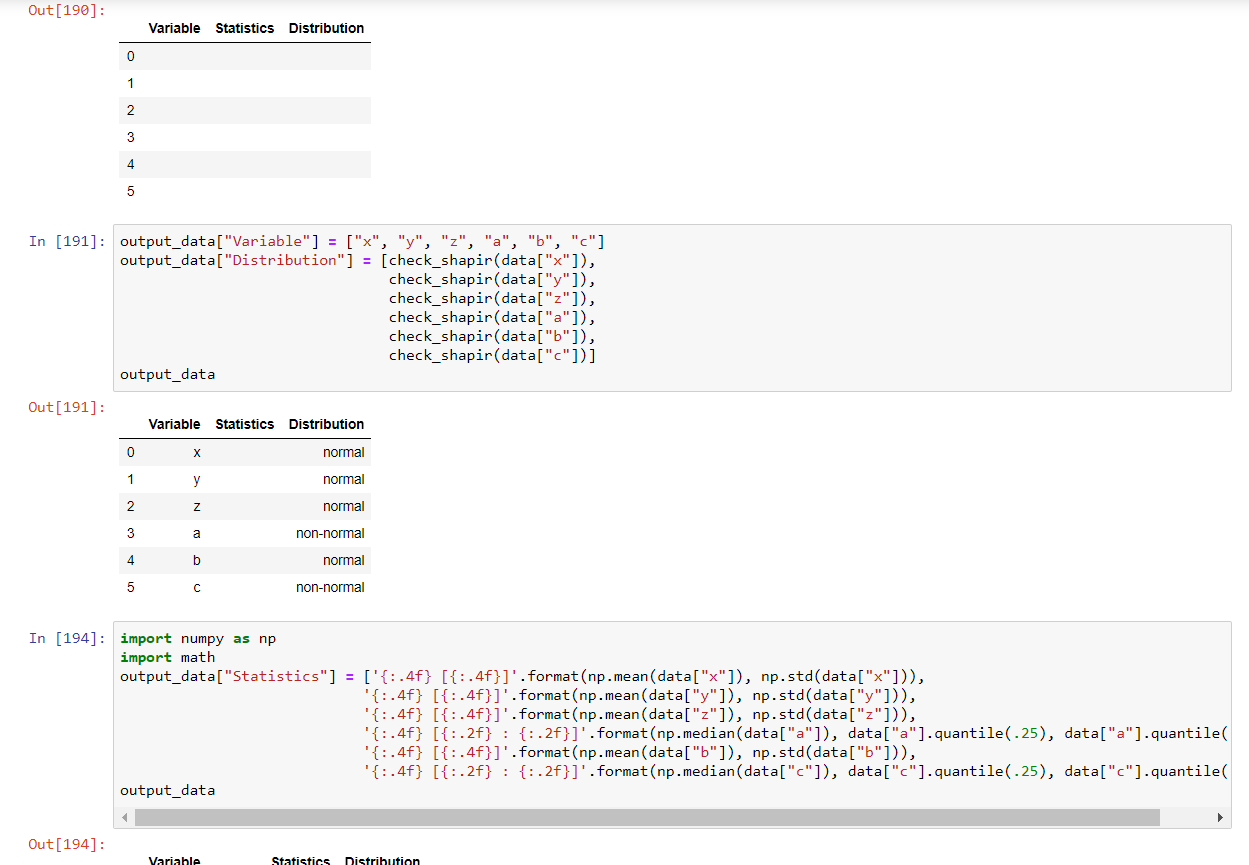
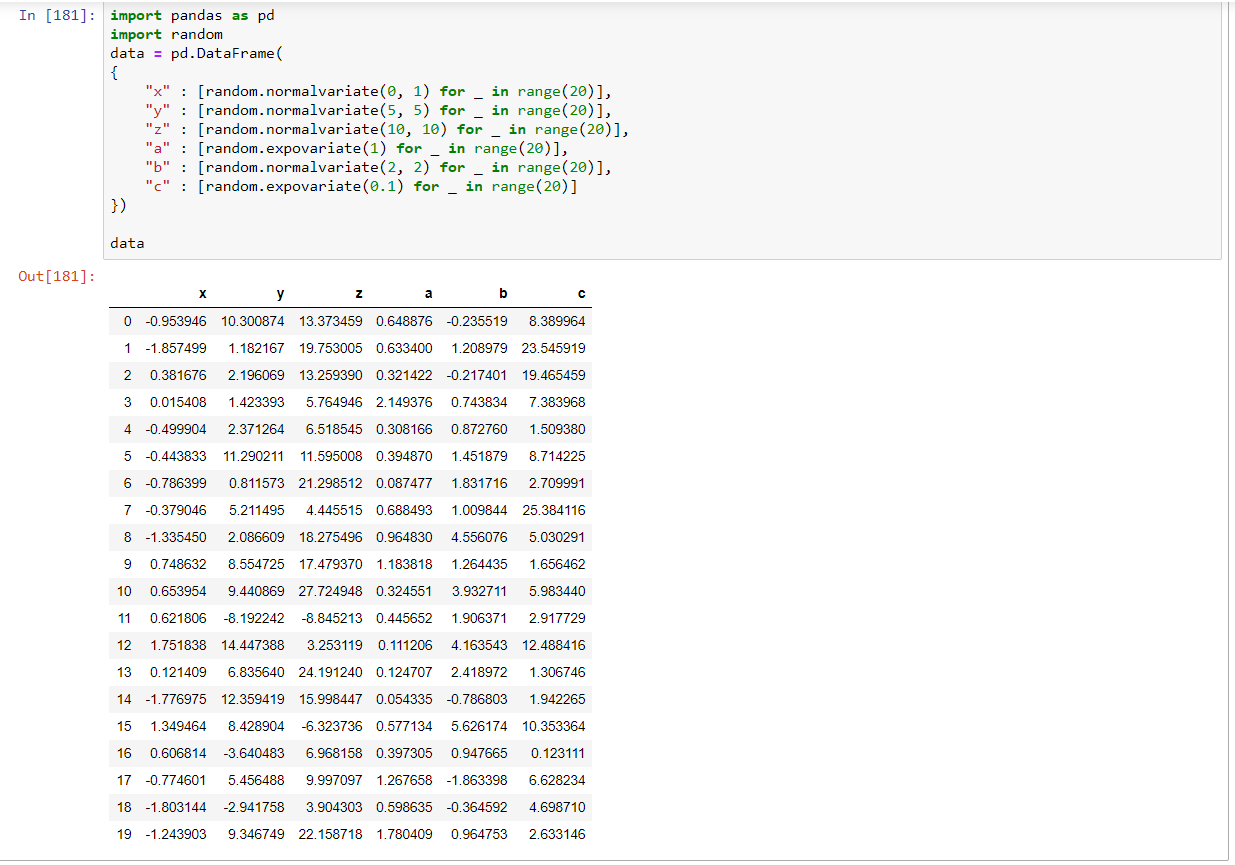
Критерій Шапіро-Уілкі заснований на оптимальній лінійної несмещённой оцінці дисперсії до її звичайної оцінці методом максимальної правдоподібності. Статистика критерію має вигляд:

s^2=\sum_{i=1}^n (x_i -\overline{x})^2, \overline{x}=\frac{1}{n}\sum_{i=1}^n x_i.

Чисельник є квадратом оцінки середньоквадратичного відхилення Ллойда.

Коефіцієнти беруться з таблиць

## *4.2. Практична реалізація*



## *4.3. Висновки до розділу*

У даному розділі було розглянуто критерій Шапіро-Уілка для перевірти нульової гіпотези. Було розраховано для всіх змінних, що розподілені за нормальним законом середне значення та стандартне відхилення, для всіх змінних, що розподілені за розподілом відсутнім від нормального медіана та інтерквартильний розмах. Результат роботи практичного коду було збережено у pdf форматі

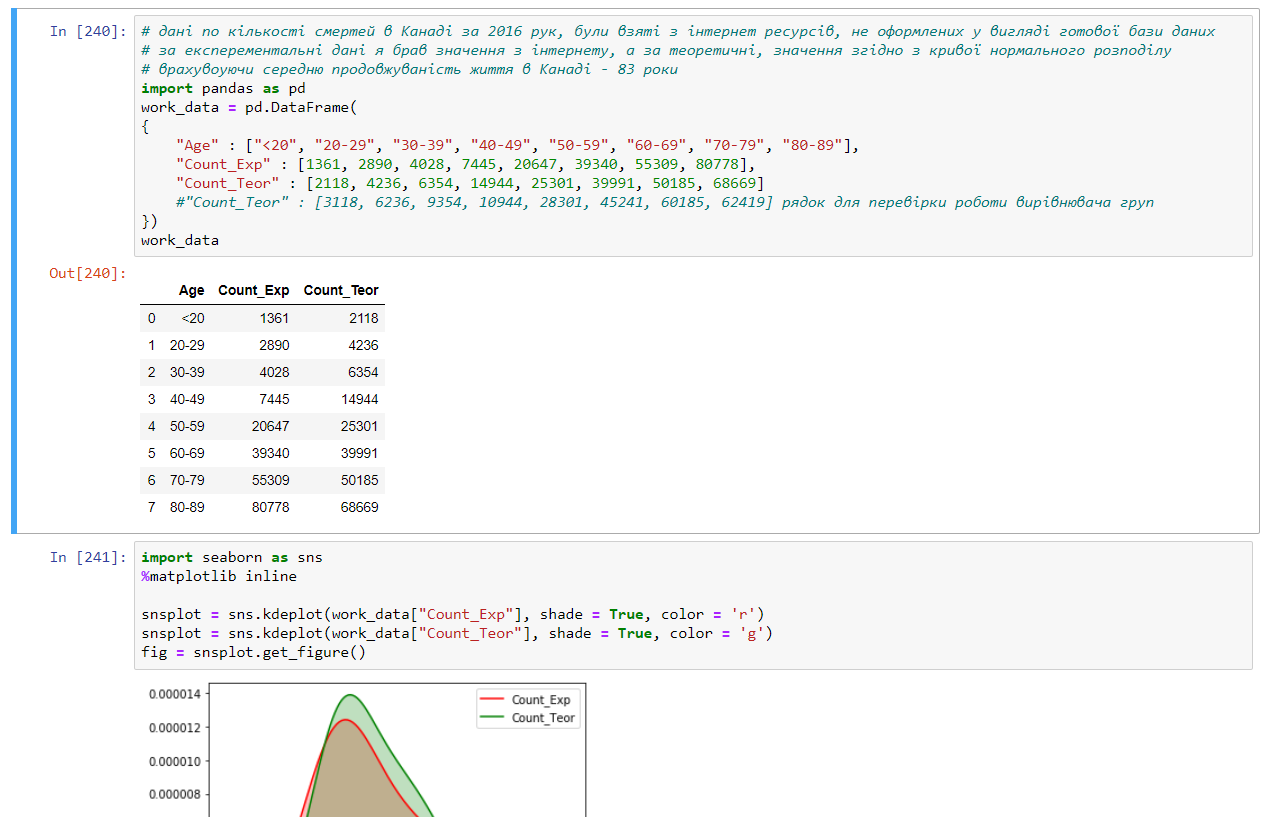
# 5. КРИТЕРІЙ ПІРСОНА ДЛЯ ПОРІВНЯННЯ ДЕКІЛЬКОЇ ГРУП ЗА РОЗПОДІЛЕННЯМ ОЗНАКИ

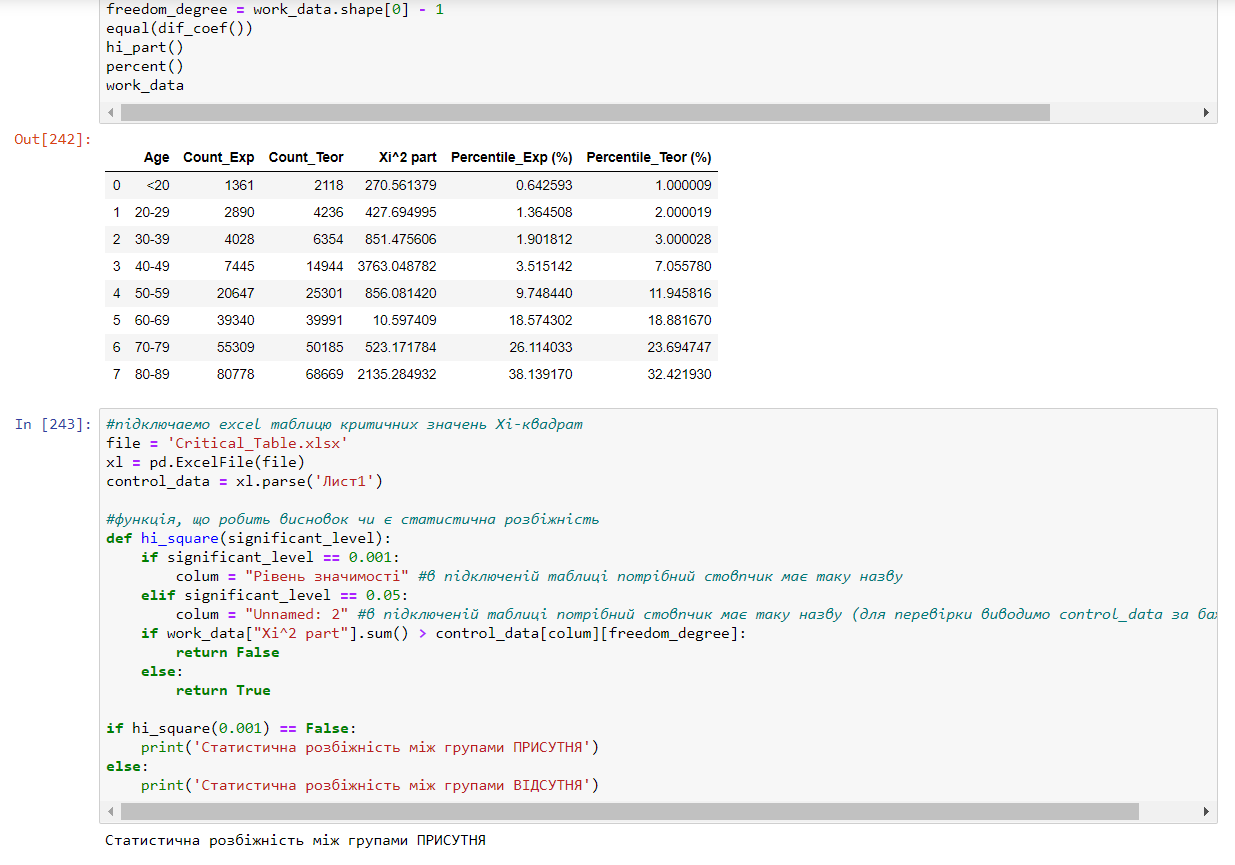
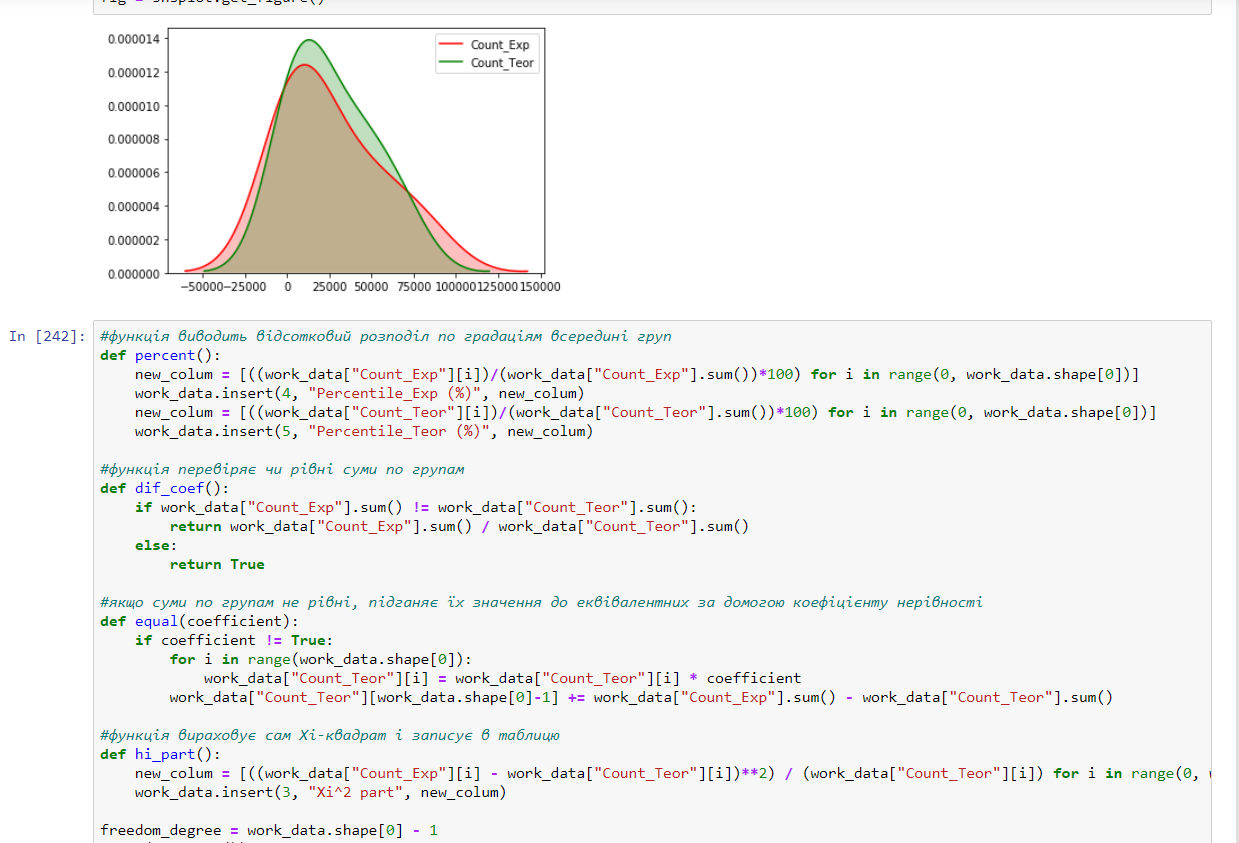
## *5.1. Теоретичні відомості*

Таблиця спряженості, або таблиця контингентні, факторна таблиця в статистиці - засіб представлення спільного розподілу двох змінних, призначене для дослідження зв'язку між ними. Таблиця спряженості є найбільш універсальним засобом вивчення статистичних зв'язків, так як в ній можуть бути представлені змінні з будь-яким рівнем вимірювання. Таблиці спряженості часто використовуються для перевірки гіпотези про наявність зв'язку між двома ознаками з використанням точного тесту Фішера або критерію згоди Пірсона.

Критерій згоди Пірсона або критерій згоди (Хі-квадрат) - це непараметричний метод, який дозволяє оцінити значимість відмінностей між фактичним (виявленим в результаті дослідження) кількістю результатів або якісних характеристик вибірки, що потрапляють в кожну категорію, і теоретичним кількістю, яке можна очікувати в досліджуваних групах при справедливості нульової гіпотези. Висловлюючись простіше, метод дозволяє оцінити статистичну значущість відмінностей двох або декількох відносних показників (частот, часткою). Найбільш часто вживається критерій для перевірки гіпотези про приналежність спостережуваної вибірки  об'ємом n деякого теоретичного закону розподілу 

## *5.2. Практична реалізація*





## 

## *5.3. Висновки до розділу*

У даному розділі було розглянуто критерій Пірсона для перевірки гіпотези про відповідність емпіричного розподілу передбачуваного теоретичного розподілу F (x) при великому обсязі вибірки (n ≥ 100). Під час порівняння груп за розподіленням ознаки за допомогою критерію Пірсона, було отримано результат і можна зробити висновок, що статична розбіжність між групами на рівнях значимості 0.05 і 0.01 є статистично значимою, так як в двох випадках χ^2 > χ^2(критичне)

# 6. Парний t-критерій Стьюдента

## *6.1. Теоретичні відомості*

Парний t-критерій Стьюдента - одна з модифікацій методу Стьюдента, використовувана для визначення статистичної значущості відмінностей парних (повторних) вимірювань.

t-критерій був розроблений Вільямом Госсетом для оцінки якості пива в компанії Гіннес. У зв'язку із зобов'язаннями перед компанією щодо нерозголошення комерційної таємниці, стаття Госсета вийшла в 1908 році в журналі «Біометрика» під псевдонімом «Student» (Студент).

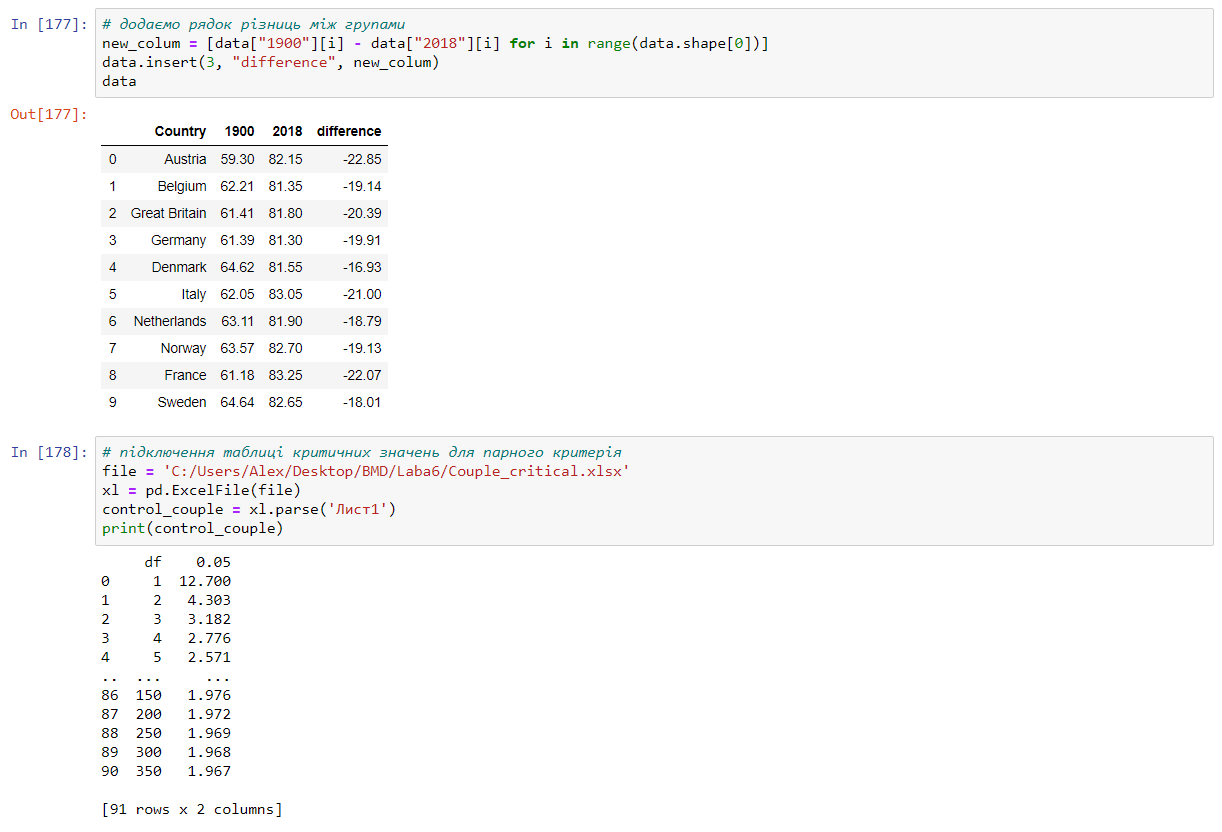
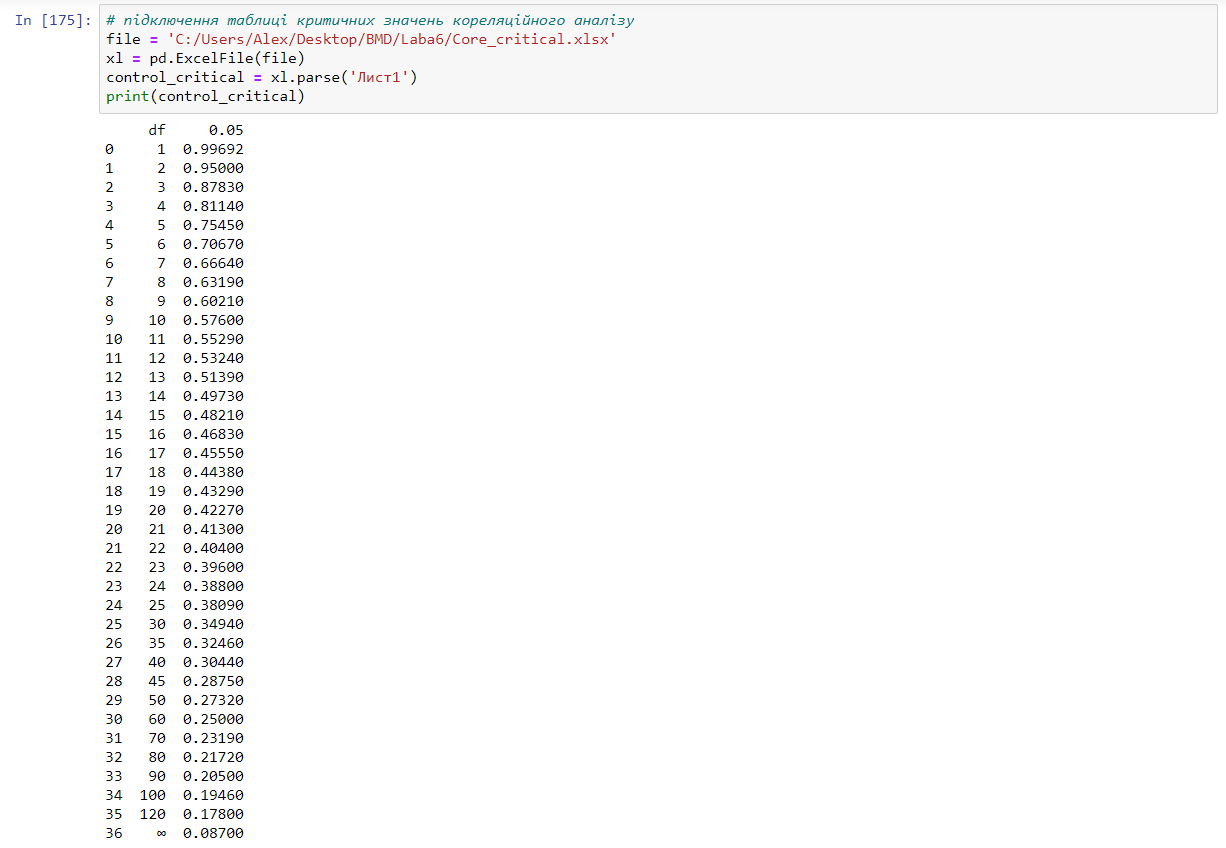
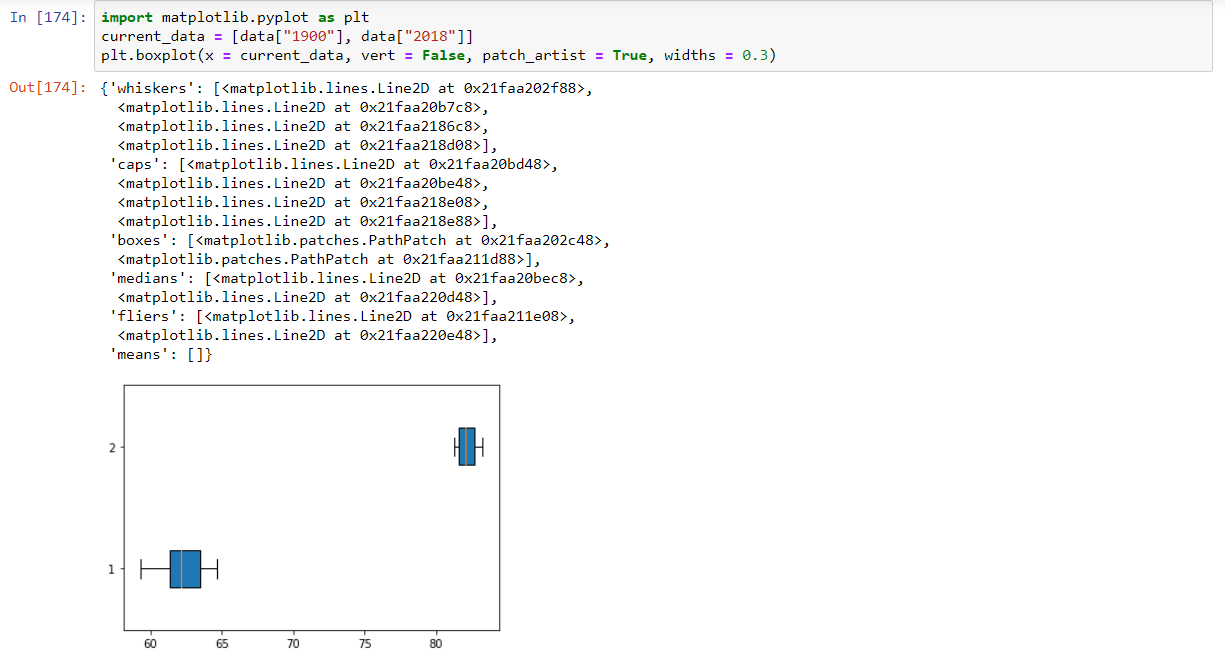
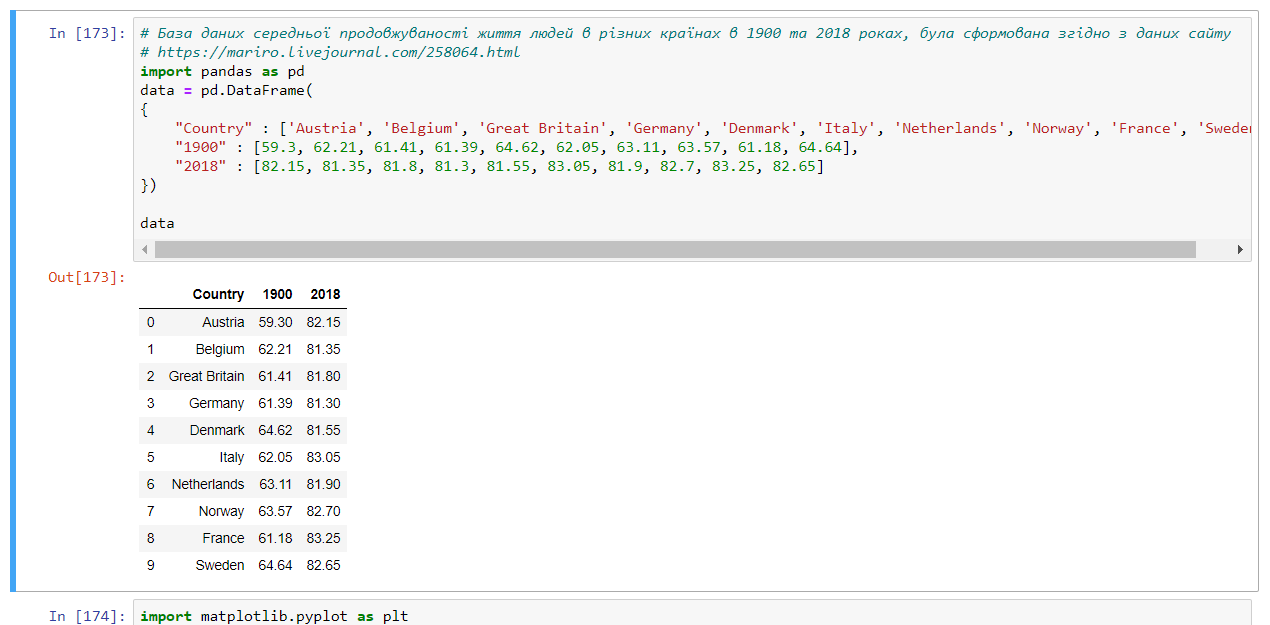
Парний t-критерій Стьюдента використовується для порівняння двох залежних (парних) вибірок. Залежними є вимірювання, виконані у одних і тих же пацієнтів, але в різний час, наприклад, артеріальний тиск у хворих на гіпертонічну хворобу до і після прийому антигіпертензивного препарату. Нульова гіпотеза свідчить про відсутність відмінностей між порівнюваними вибірками, альтернативна - про наявність статистично значущих відмінностей.

Парний t-критерій Стьюдента розраховується за такою формулою:



де Мd - середня арифметична різниць показників, виміряних до і після, σd - середньоквадратичне відхилення різниць показників, n - число досліджуваних.

## *6.2. Практична реалізація*



## *6.3. Висновки до розділу*

У даному розділі було розглянуто парний t-критерій Стьюдента для класу методів статистичної перевірки гіпотез (статистичних критеріїв). Проаналізувавши дані, можна зробити висновок що: за коефіцієнтом кореляції Пірсона було доведено, що немає статистичної залежності між групами ; Виявлена наявність статистично значущих відмінностей між двома таблицями таних, адже отримане значення критерію Стьюдента є більшим за критичне.

# 7. t-критерій Стьюдента для незалежних вибірок

## 7.1. Теоретичні відомості

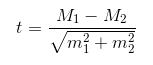
t-критерій Стьюдента - загальна назва для класу методів статистичної перевірки гіпотез (статистичних критеріїв), заснованих на розподілі Стьюдента. Найбільш часті випадки застосування t-критерію пов'язані з перевіркою рівності середніх значень у двох вибірках.

t-критерій Стьюдента використовується для визначення статистичної значущості відмінностей середніх величин. Може застосовуватися як у випадках порівняння незалежних вибірок (наприклад, групи хворих на цукровий діабет і групи здорових), так і при порівнянні пов'язаних сукупностей (наприклад, середня частота пульсу у одних і тих же пацієнтів до і після прийому антиаритмічного препарату). В останньому випадку розраховується парний t-критерій Стьюдента

Для застосування t-критерію Стьюдента необхідно, щоб вихідні дані мали нормальний розподіл. Також має значення рівність дисперсій (розподілу) порівнюваних груп (гомоскедастичність). При нерівних дисперсіях застосовується t-критерій в модифікації Уелча (Welch's t).

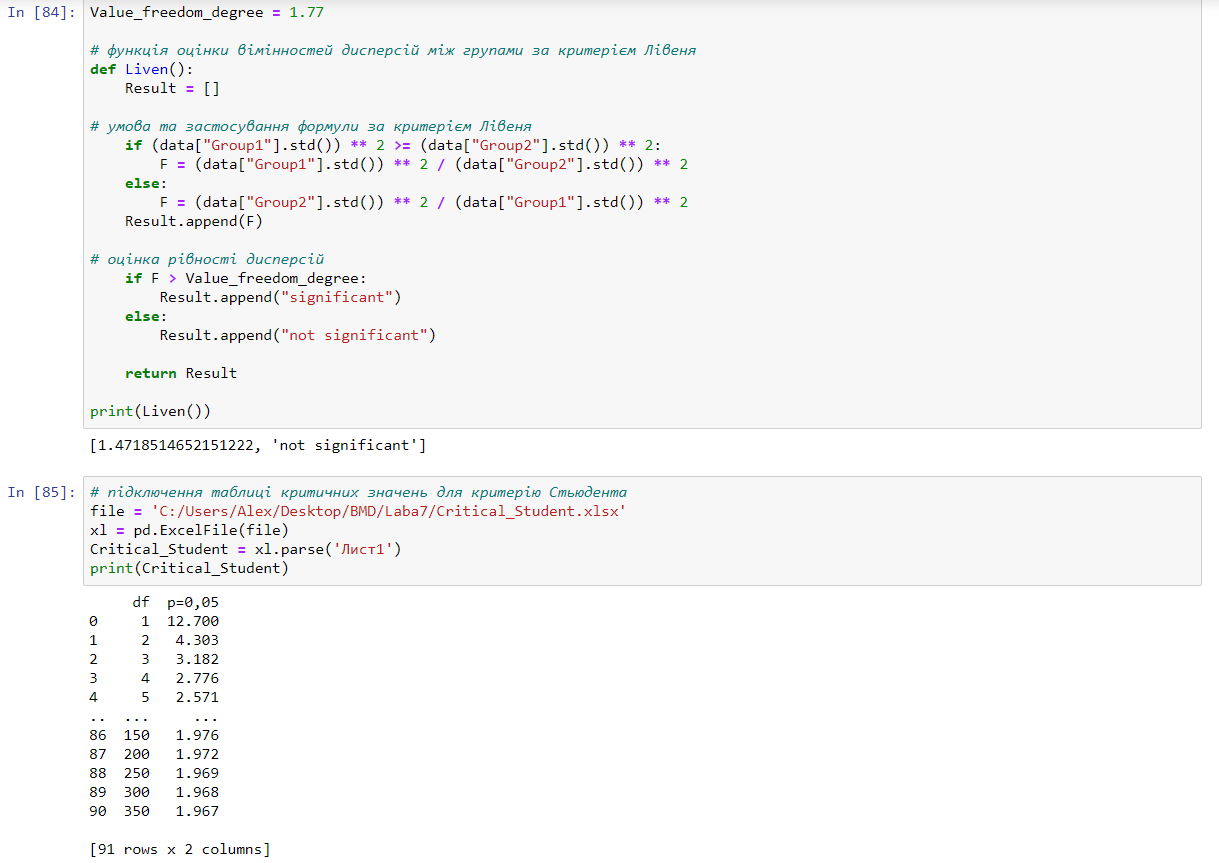
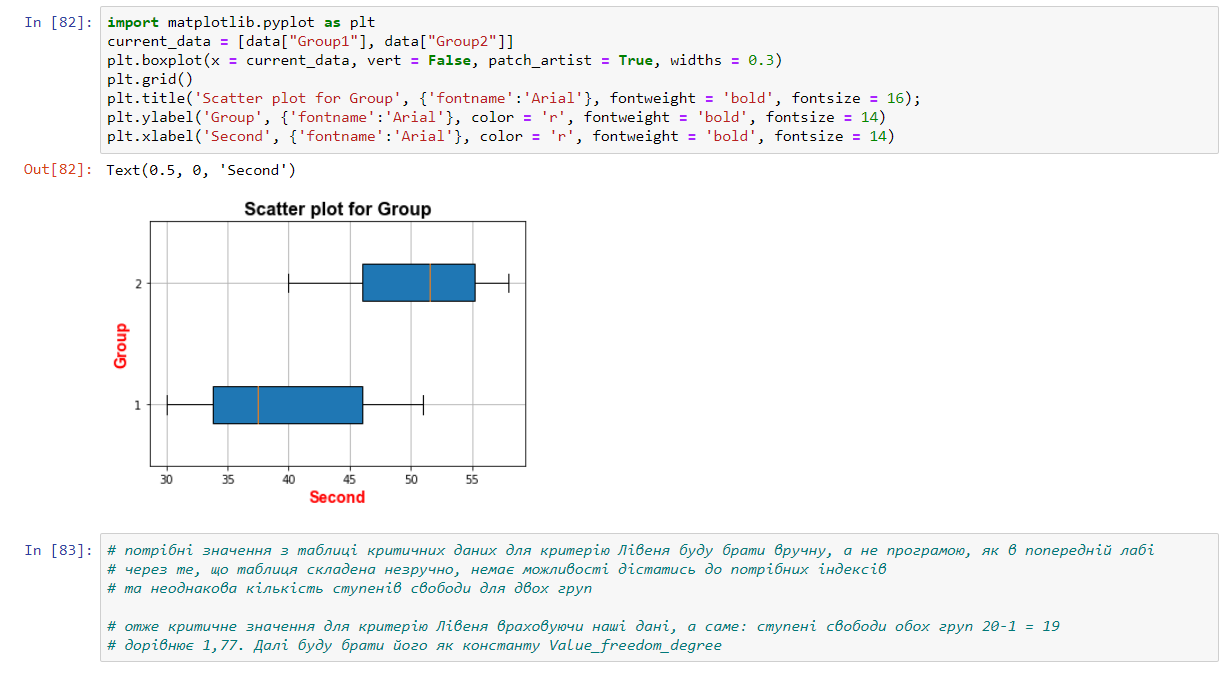
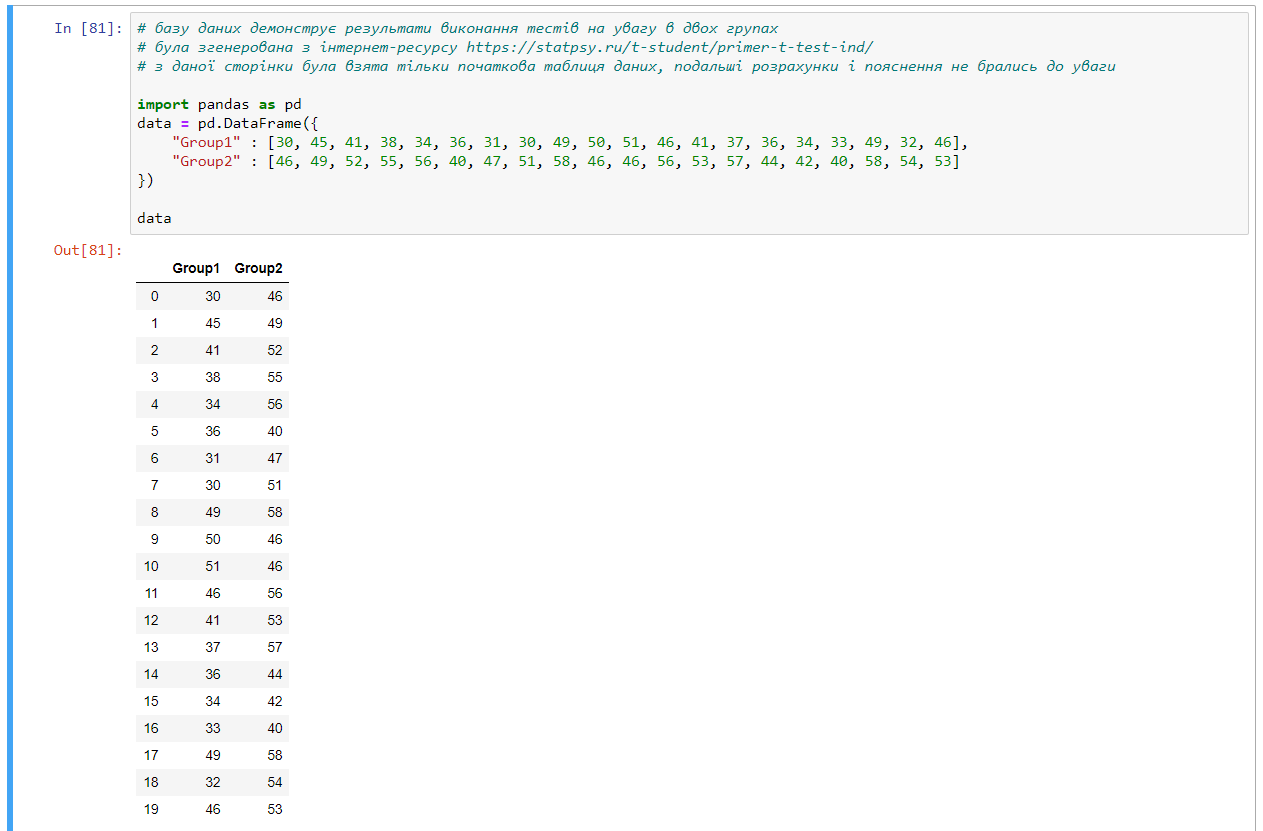
При відсутності нормального розподілу порівнюваних вибірок замість t-критерію Стьюдента використовуються аналогічні методи непараметричної статистики, серед яких найбільш відомими є U-критерій Манна - Уїтні.

Для порівняння середніх величин t-критерій Стьюдента розраховується за такою формулою:



де М1 - середня арифметична першої порівнюєш сукупності (групи), М2 - середня арифметична другий порівнюєш сукупності (групи), m1 - середня помилка першої середньої арифметичної, m2 - середня помилка другий середньої арифметичної.

## *7.2. Практична реалізація*



## *7.3. Висновки до розділу*

У даному розділі було розглянуто t-критерій Стьюдента для незалежних вибірок для порівняння середніх значень двох незалежних між собою вибірок. В результаті перевірки даних на рівність дисперсії за тестом Лівеня було, отримано результат F < F\_tabl. Отже потрібно прийняти нульову гіпотезу, тобто зробити висновок про статичну рівність дисперсій. В результаті перевірки відмінності між двома групами за допомогою критерію Стюдента для незалежних вибірок, було отримано результат t > t\_tabl. Отже, відхиляємо нульову гіпотезу (тобто приймаємо альтернативну), та робимо висновок про статичну відмінність середніх.

# 8. Дисперсійний аналіз

## *8.1. Теоретичні відомості*

Дисперсійний аналіз - (ANOVA) застосовується для дослідження впливу однієї або декількох якісних змінних (факторів) на одну залежну кількісну змінну (відгук).

В основі дисперсійного аналізу лежить припущення про те, що одні змінні можуть розглядатися як причини (фактори, незалежні змінні): , а інші як слідства (залежні змінні). Незалежні змінні називають іноді регульованими чинниками саме тому, що в експерименті дослідник має можливість варіювати ними і аналізувати виходить результат.

Основною метою дисперсійного аналізу (ANOVA) є дослідження значущості відмінності між середніми за допомогою порівняння (аналізу) дисперсій. Поділ загальної дисперсії на кілька джерел, дозволяє порівняти дисперсію, викликану відмінностями між групами, з дисперсією, викликаної внутрішньогрупової мінливістю. При істинності нульової гіпотези (про рівність середніх в декількох групах спостережень, вибраних з генеральної сукупності), оцінка дисперсії, пов'язаної з внутрішньогрупової мінливістю, повинна бути близькою до оцінки груповий дисперсії. Якщо ви просто порівнюєте середні в двох вибірках, дисперсійний аналіз дасть той же результат, що і звичайний t-критерій для незалежних вибірок (якщо порівнюються дві незалежні групи об'єктів або спостережень) або t-критерій для залежних вибірок (якщо порівнюються дві змінні на одному і тій же множині об'єктів або спостережень).

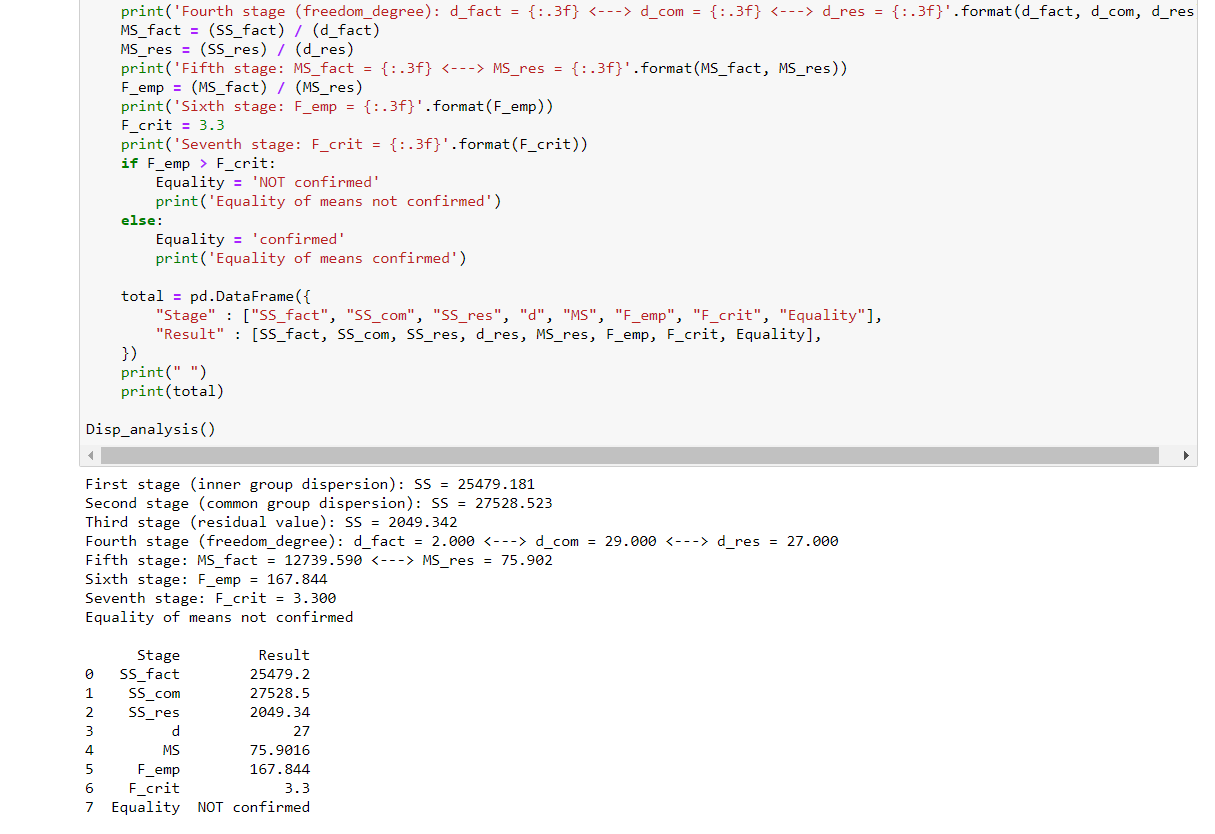
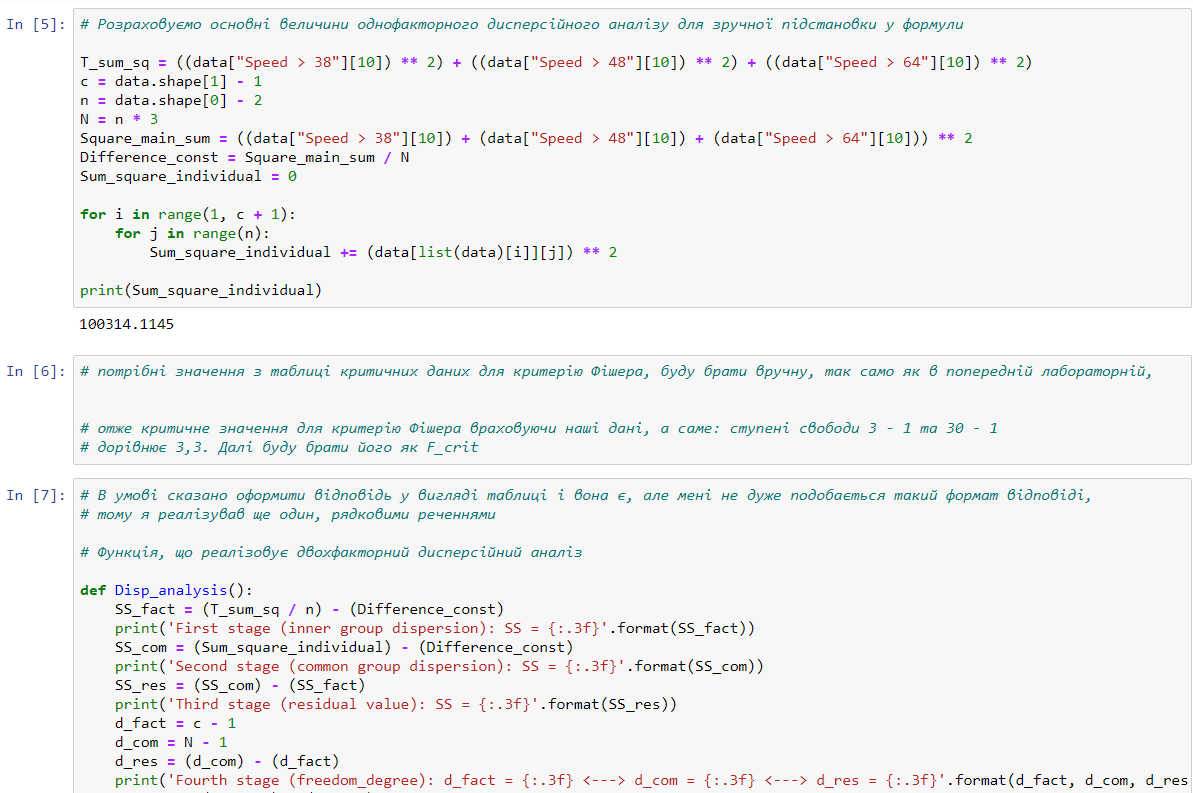
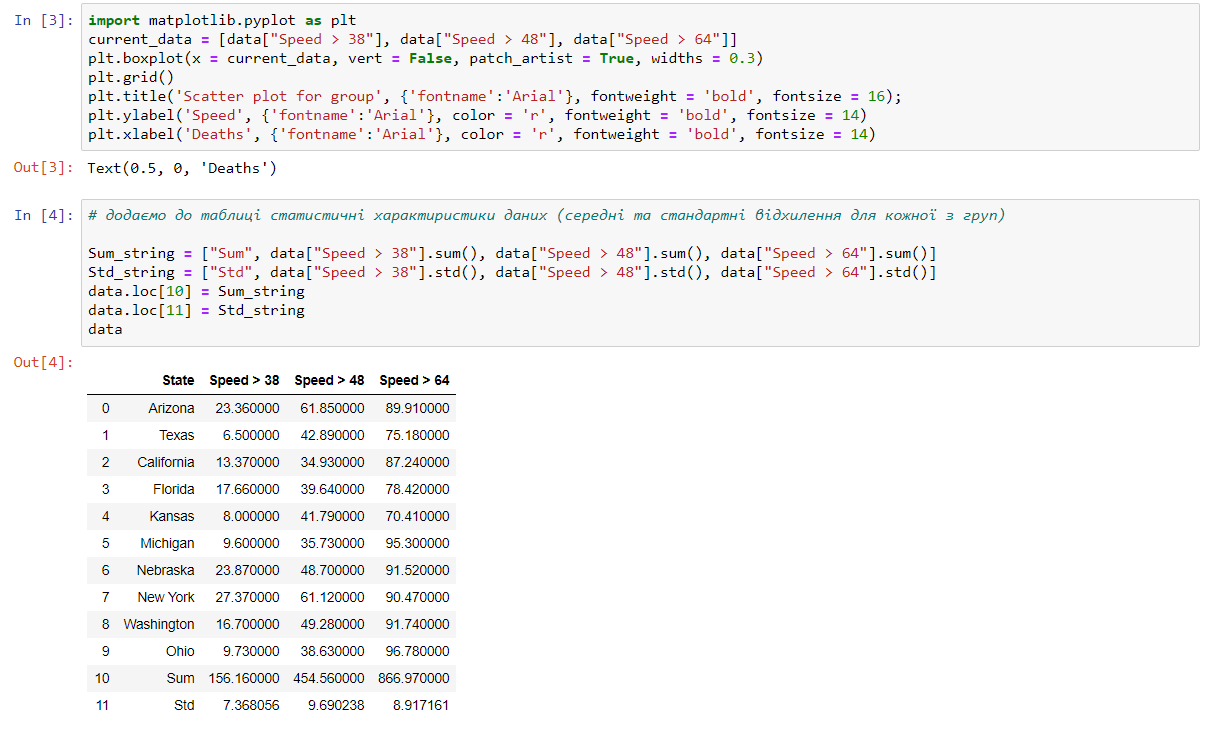
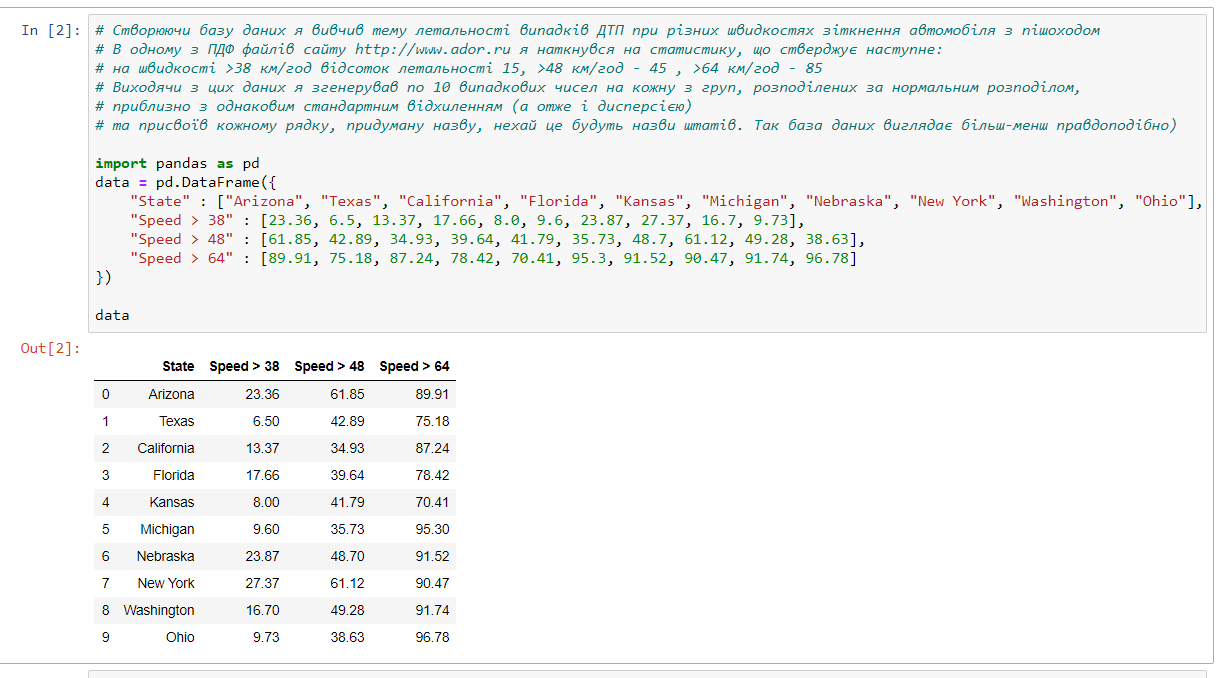
Сутність дисперсійного аналізу полягає в розчленуванні загальної дисперсії досліджуваного ознаки на окремі компоненти, зумовлені впливом конкретних факторів, і перевірці гіпотез про значущість впливу цих факторів на досліджуваний ознака. Порівнюючи компоненти дисперсії один з одним за допомогою F-критерію Фішера, можна визначити, яка частка загальної варіативності результативної ознаки зумовлена ​​дією регульованих факторів.

Вихідним матеріалом для дисперсійного аналізу є дані дослідження трьох і більше вибірок: , які можуть бути як рівними, так і нерівними за чисельністю, як зв'язковими, так і незв'язними. За кількістю виявлених регульованих факторів дисперсійний аналіз може бути однофакторний (при цьому вивчається вплив одного фактора на результати експерименту), двофакторна (при вивченні впливу двох чинників) і багатофакторним (дозволяє оцінити не тільки вплив кожного з факторів окремо, але і їх взаємодія).

Дисперсійний аналіз відноситься до групи параметричних методів і тому його слід застосовувати тільки тоді, коли доведено, що розподіл є нормальним.

Дисперсійний аналіз використовують, якщо залежна змінна вимірюється в шкалі відносин, інтервалів або порядку, а що впливають змінні мають нечислову природу (шкала найменувань).

## *8.2. Практична реалізація*



## *8.3. Висновки до розділу*

Була створена функція, що реалізовує однофакторний дисперсійний аналіз. Дані були пропущені через функцію. Рівність середніх підтверджена не була.