**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**ФАКУЛЬТЕТ КОМП’ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ**

**ЗВІТ**

**Лабораторній роботі №1**

|  |
| --- |
| Студента групи ПІ-22 |
| Кіщук Ярослава Ярославовича |
|  |
|  |
|  |
|  |
|  |
|  |

Київ 2023

**ЗМІСТ**

[**1. ОПИС ВХІДНОЇ ІНФОРМАЦІЇ 2**](#_kxg4hvcy4z8c)

[● *Дані* 2](#_vcugx0wbm3tj)

[*● Використанні змінні 2*](#_agsmmhpr2gkq)

[**2. ПОСТАНОВКА ЗАДАЧІ 3**](#_2678q09a8yd5)

[**3. ОПИС ДОДАТКОВОГО ТЕОРЕТИЧНОГО МАТЕРІАЛУ 4**](#_ni7pmf7wxoyr)

[*1. QQ-Plot (Quantile-Quantile Plot) 5*](#_2et92p0)

[*2. Trimean 6*](#_4g0jlyj4cijt)

[*3. Критерій узгодженості Колмогорова-Смірнова (KS) 6*](#_rm073aaeoer7)

[**4. ПОКРОКОВИЙ ОПИС ОБРОБКИ ДАНИХ 6**](#_qgfdv2da1rjp)

[**1. Графічне представлення змінних 7**](#_yeb781use15g)

[*Емпірична функція щільності 7*](#_xoereu906g21)

[Фактична ціна товару 7](#_jkwgd0oqagsl)

[Рейтинг товару 7](#_x93b1yt8hbqs)

[Відсоток знижки на товар 7](#_giglaxi0cyn5)

[● Висновок: 7](#_ne0cyjutg5bz)

[*Полігони частот 8*](#_o5u0y4smqcpo)

[Фактична ціна товару 8](#_1mta3oi5rolg)

[Рейтинг товару 8](#_7v70g5jk7grs)

[Відсоток знижки на товар 8](#_fag3hycfj3pr)

[● Висновок: 8](#_qpuyhnbc779x)

[**2. Зображення «Скринька з вусами» 9**](#_m9q7uqh6jffu)

[Фактична ціна товару 11](#_cwm71v9k13qr)

[Рейтинг товару 12](#_ka8wzqlzjje7)

[Відсоток знижки на товар 14](#_pfw76i4hr044)

[● Висновок: 14](#_2pqz13biphk)

[**3. Мінімальне, максимальне значення, квартилі, децилі, медіана 16**](#_l2sec967exqw)

[Фактична ціна товару 19](#_4q3gv6jc7kxq)

[Рейтинг товару 19](#_spr7so3wmnet)

[Відсоток знижки на товар 19](#_ck5xm1ermc4u)

[● Висновок: 20](#_56g0c8pjeelz)

[**4. Характеристики положення центра значень змінної 20**](#_wpoyz2feu2wr)

[Фактична ціна товару 20](#_v8vbisgxytwh)

[Рейтинг товару 21](#_3s7iq1w87eoa)

[Відсоток знижки на товар 22](#_kbkp84sslp2q)

[● Висновок: 23](#_mkb5qi7vgk2)

[**5. Характеристики розсіювання значень змінної 25**](#_pl841csutf8l)

[Фактична ціна товару 26](#_5adb34clsz0f)

[Рейтинг товару 26](#_8idwvbtaruem)

[Відсоток знижки на товар 27](#_jcjvuu2d1nw9)

[● Висновок: 27](#_q6a76n1yjxzj)

[**6. Аналіз скошеності та гостроверхості розподілу 28**](#_t9fn662al304)

[Фактична ціна товару 30](#_hib1p2sy7oe0)

[Рейтинг товару 30](#_6tieou9u65qt)

[Відсоток знижки на товар 30](#_gy3ybswic3zd)

[● Висновок: 30](#_jjzcv546yv5r)

[**7. Перевірка нормальності 30**](#_nw47ggmi9ut0)

[Фактична ціна товару 31](#_jqi9in94pgt1)

[Графік Q-Q 31](#_25u9je2whhc9)

[Критерій узгодженості Колмогорова-Смірнова 31](#_6de875zcsxjo)

[Рейтинг товару 31](#_1bcx77ve0eu1)

[Графік Q-Q 31](#_wtffs8of5hxt)

[Критерій узгодженості Колмогорова-Смірнова 31](#_pjgch5wm88r0)

[Відсоток знижки на товар 31](#_13uce17ejvxe)

[Графік Q-Q 31](#_y701nbf3ainb)

[Критерій узгодженості Колмогорова-Смірнова 31](#_ouyk10f9buk0)

[● Висновок: 31](#_a7df1livzpfb)

[**8. Інші процедури попереднього аналізу 32**](#_s3h4d8jggxil)

[Порівняння розподілу Фактичної Ціни Товару з Логнормальним розподілом 32](#_wvhyebewa3nv)

[Графік Q-Q 32](#_ck109mi8nwvp)

[Критерій узгодженості Колмогорова-Смірнова 32](#_ncdzvqdnzpwj)

[● Висновок: 32](#_5ndwgfseuma5)

[**5. ВИСНОВКИ 32**](#_yk1x9dtj9iwd)

[**6. ДОДАТОК. ПРОГРАМНА РЕАЛІЗАЦІЯ 36**](#_11dv7qxybtne)

[**7. СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 36**](#_6zxrhaa5c97l)

### ОПИС ВХІДНОЇ ІНФОРМАЦІЇ

Приводиться опис вхідної інформації (URL джерела даних, опис усіх

змінних).

## *Дані*

Набір даних містить рейтинги, деталі та відгуки на такі товари як побутові електроприлади, телевізори, смартфони та різні аксесуари до них, які продаються на Amazon.

Посилання на ресурс звідки було взято дані: [Amazon Sales Dataset | Kaggle](https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset)

## *Використанні змінні*

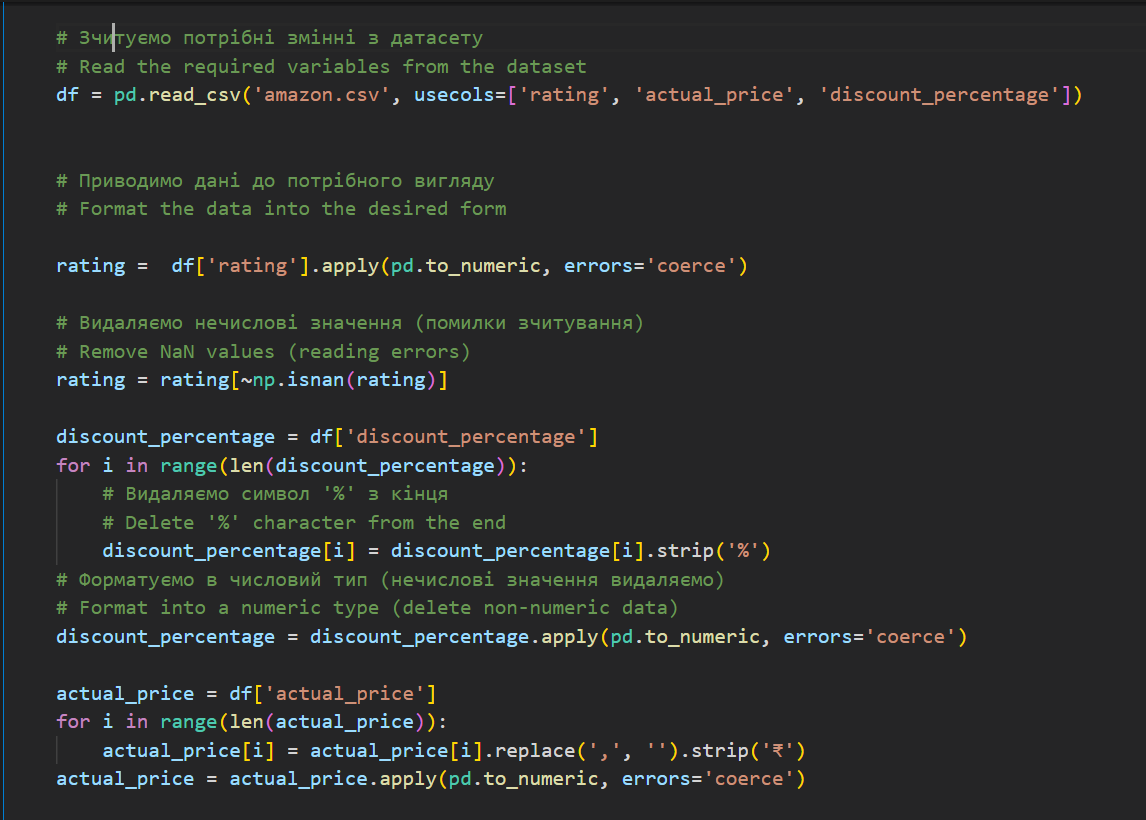
* *actual\_price* - фактична ціна товару (індійська рупія ₹), **скалярна кількісна величина**
* *rating* - рейтинг товару, **скалярна кількісна величина**
* *discount\_percentage* - відсоток знижки на товар, **скалярна кількісна величина**

*Ціна*: дані про ціни дозволяють визначити конкурентоспроможність продукту на ринку та порівняти ціни на схожі продукти.

*Рейтинг товарів -* це кількість зірочок (або оцінка від 1 до 5), яку користувачі залишають на Amazon після покупки товару. Ця змінна вказує на те, наскільки задоволені покупці продуктом та яку думку вони мають про якість товару. Це може допомогти продавцям знайти популярні продукти, щоб покращити продажі.

*Процент знижки -* це знижка від оригінальної ціни товару, яку надає Amazon. Дані про знижки дозволяють визначити рівень знижки на продукти та порівняти їх з конкурентами на ринку.

***Зчитування даних***

****

### ПОСТАНОВКА ЗАДАЧІ

Наводиться постановка задачі та її формалізація для поточної ЛР.

Попередній аналіз даних - це процес дослідження та розуміння даних перед застосуванням будь-яких моделей аналізу даних або машинного навчання. Загальна постановка задачі попереднього аналізу даних полягає в тому, щоб з'ясувати, яка інформація міститься в наборі даних, як вона зберігається та організована, які залежності та взаємозв'язки існують між різними змінними, чи є аномальні значення чи пропущені дані, які методи можуть бути застосовані для очищення даних від шуму та видалення аномалій.

Конкретна постановка задачі попереднього аналізу включає в себе:

* Вивчення загальної структури даних, такої як розмір, формат та типи змінних;
* Візуалізація розподілів змінних та взаємозв'язків між ними;
* Виявлення аномальних значень та пропущених даних;
* Перевірка гіпотез та встановлення залежностей між змінними;
* Вибір методів для очищення даних від шуму та видалення аномалій.

До цих етапів відносяться: підрахунок характеристик положення центра значень змінної (математичне сподівання, середнє геометричне, середнє гармонічне, мода, медіана), підрахунок характеристик розсіювання значень змінної (дисперсія, середнє квадратичне, коефіцієнт варіації, ймовірнісне відхилення, розмах вибірки, інтервал концентрації), дослідження скошеності та гостроверхості розподілів (коефіцієнт асиметрії та коефіцієнт ексцесу) та такі характеристики як квантилі, децилі тощо

Загалом, мета попереднього аналізу даних полягає в розумінні основних характеристик та закономірностей в наборі даних, які можуть бути корисні для подальшого дослідження та аналізу.

### ОПИС ДОДАТКОВОГО ТЕОРЕТИЧНОГО МАТЕРІАЛУ

Надається опис додаткового теоретичного матеріалу, який був використаний у даній ЛР.

## QQ-Plot (Quantile-Quantile Plot)

**QQ-Plot** - це графічний метод порівняння розподілу двох вибірок. Його використовують для перевірки гіпотези про нормальний розподіл даних та для порівняння двох розподілів даних.

QQ-Plot базується на порівнянні квантилів (percentiles) двох вибірок. Для побудови QQ-Plot, вісь x та вісь y відповідають квантилям вихідної вибірки та тестової відповідно. Якщо обидві вибірки мають однаковий розподіл, то точки QQ-Plot розташовані приблизно по діагоналі. Якщо ж вибірки мають різні розподіли, то точки QQ-Plot будуть відхилятись від діагоналі.

## Trimean

**Trimean** - це одна з мір центральної тенденції, яка представляє собою середнє арифметичне між медіаною та середнім значенням (mean) вибірки.

Trimean є більш стійкою мірою центральної тенденції, ніж середнє значення (mean), оскільки він не чутливий до великих викидів (outliers) вибірки, які можуть суттєво змінити значення середнього значення. Однак, в порівнянні з медіаною, trimean є менш чутливим до змін у краях розподілу вибірки та не дозволяє виявляти зміну форми розподілу даних.

## Критерій узгодженості Колмогорова-Смірнова (KS)

**Критерій узгодженості Колмогорова-Смірнова**- це статистичний тест, який використовується для перевірки припущення про те, що зразок даних має певний розподіл. Тест KS порівнює кумулятивний розподіл зразка даних з теоретичним розподілом,

який вважається ідеальним для порівняння.

Ілюстрація статистики Колмогорова-Смірнова (вище). Червона лінія - модельна CDF, синя лінія - емпірична CDF, а чорна стрілка - статистика KS.

### ПОКРОКОВИЙ ОПИС ОБРОБКИ ДАНИХ

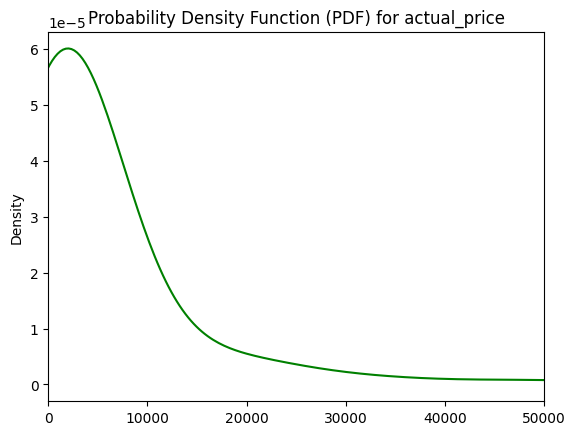
Наводиться покроковий опис обробки даних у ЛР, причому кожен крок ОБОВ'ЯЗКОВО супроводжується відповідними скріншотами фрагментів програмної реалізації та отриманих результатів, а також відповідним висновком для цього кроку.

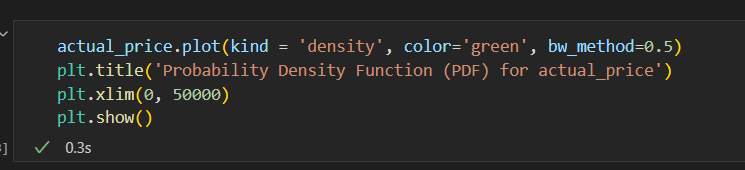
#### Графічне представлення змінних

##### *Емпірична функція щільності*

Емпірична функція щільності - це невипадкова функція, що описує розподіл вибірки даних. Вона є аналогом функції щільності ймовірності випадкової величини для емпіричного розподілу, тобто вона дозволяє оцінити ймовірність того, що випадкова величина знаходиться в певному інтервалі.

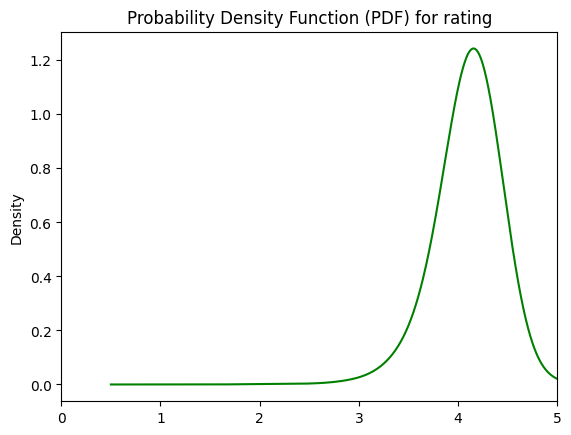
###### Фактична ціна товару

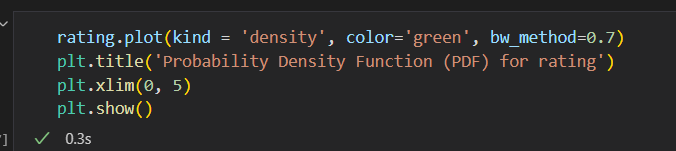




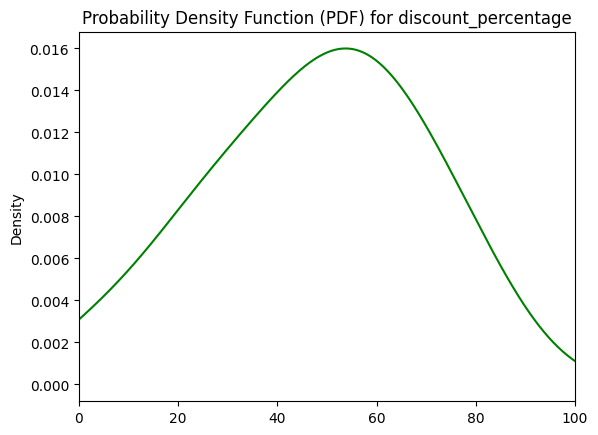
#### 

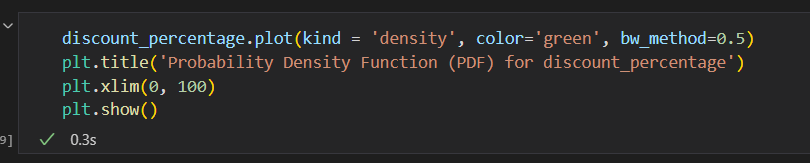
###### Рейтинг товару





###### Відсоток знижки на товар





###### Висновок:

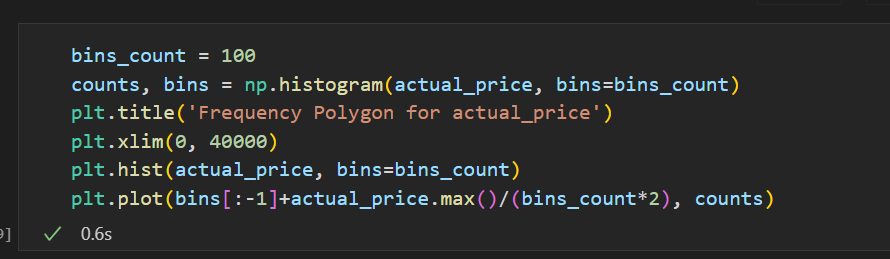
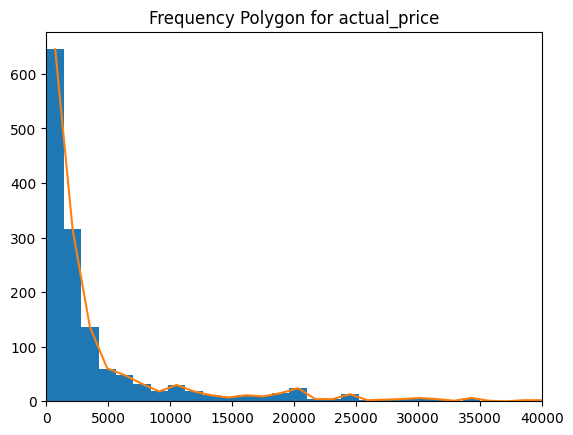
Емпірична функція щільності може допомогти визначити форму розподілу даних, наприклад, чи є вони нормально розподілені, рівномірно розподілені чи мають інший тип розподілу.

На графіку емпіричної функції щільності змінної *actual\_price* можна помітити графік логнормального розподілу. Також бачимо, що розподіл змінної *rating* є більш гостроверхий, а для змінної *discount\_percentage* скошений праворуч відносно нормального розподілу чим нагадує бета розподіл із відповідними параметрами.

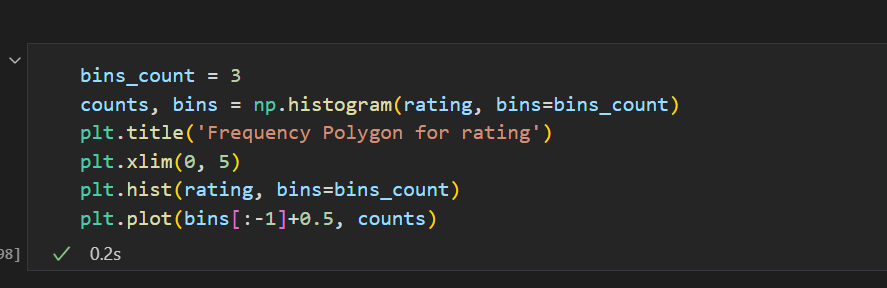
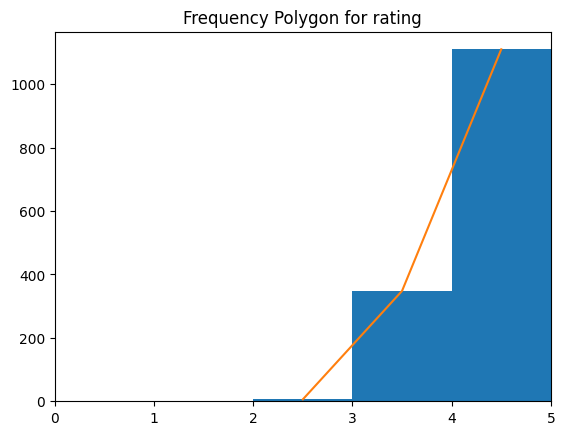
##### *Полігони частот*

Полігон частот - це графік, який показує частоти згрупованих даних. Це тип частотної діаграми, на якій середні точки інтервалів класів наносяться на частоти, а потім з'єднуються прямими лініями.

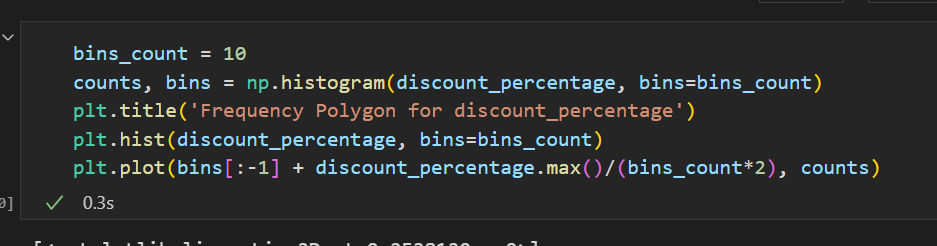
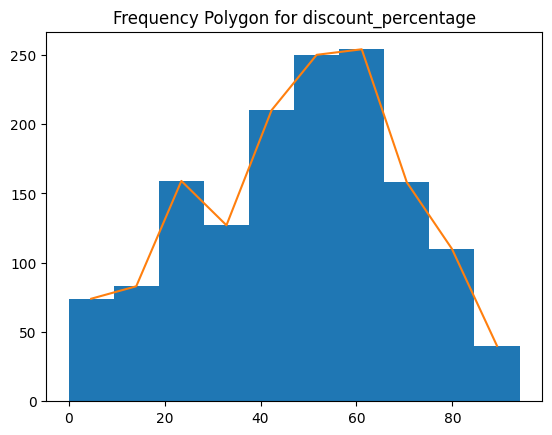
###### Фактична ціна товару



###### Рейтинг товару



###### Відсоток знижки на товар



###### Висновок:

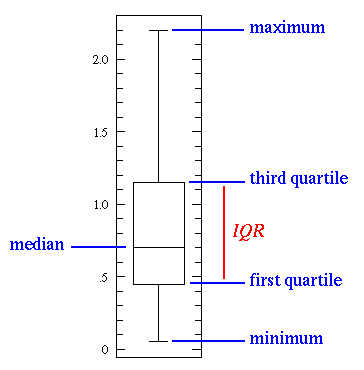
Полігон частот показує частоту або кількість спостережень для кожного значення змінної, що полегшує визначення точної кількості спостережень у певній точці на осі х. Також він може допомогти нам виявити прогалини в даних або області, де немає спостережень. Можемо побачити кластери точок даних, де певні значення зустрічаються частіше, ніж інші.

На графіку полігону частот змінної *actual\_price* бачимо, що левова частка товарів має ціну в проміжку ~ 0-1200₹.

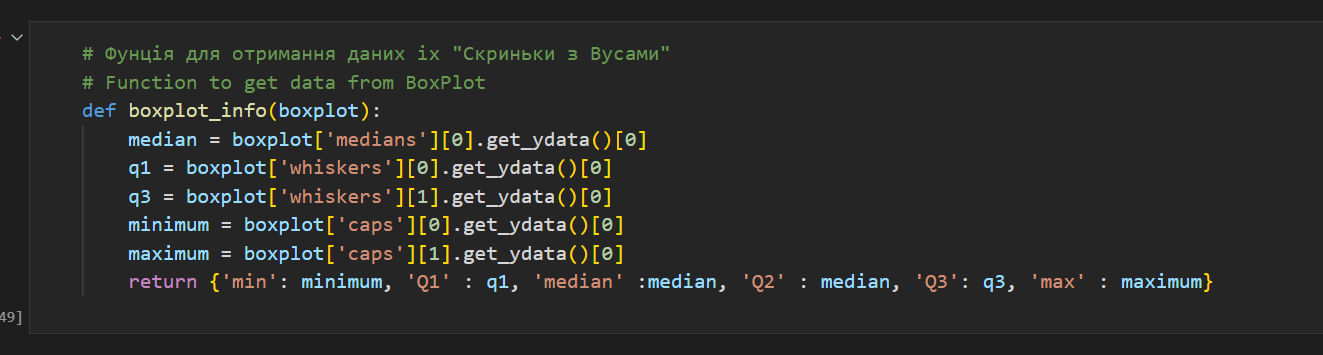
Проаналізувавши графік змінної *rating* можна припустити, що більшість покупців виявилась задоволеною від покупки.

Для змінної *discount\_percentage* можемо зробити висновок, що найбільша кількість товарів порівняно з іншими проміжками припадає на проміжки зі знижкою 45-55 та 55-65%.

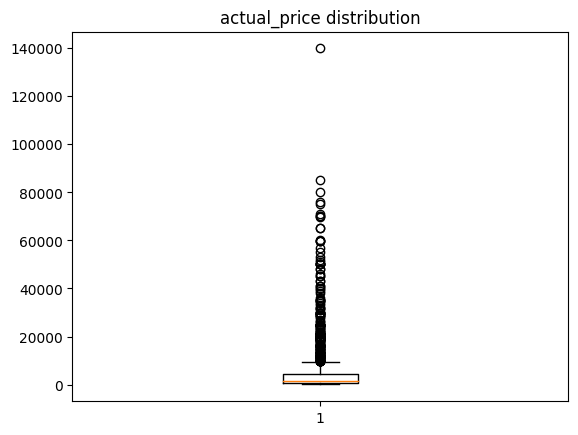
#### Зображення «Скринька з вусами»

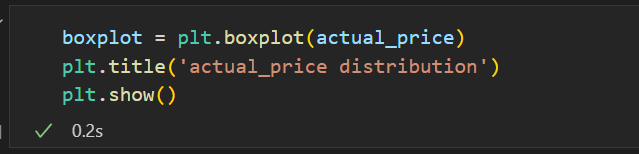
Графік "Скринька з вусами" (англ. box and whisker plot) є одним з інструментів візуалізації даних і використовується для відображення розподілу числових даних через п'ять числових показників: мінімум, нижній квартиль, медіана, верхній квартиль та максимум.

Застосування графіку "Скринька з вусами" дозволяє швидко порівнювати розподіл числових даних між різними групами, виявляти відмінності в цих розподілах та виявляти викиди даних. Це допомагає приймати рішення на основі даних та зробити висновки з підвищеною достовірністю.



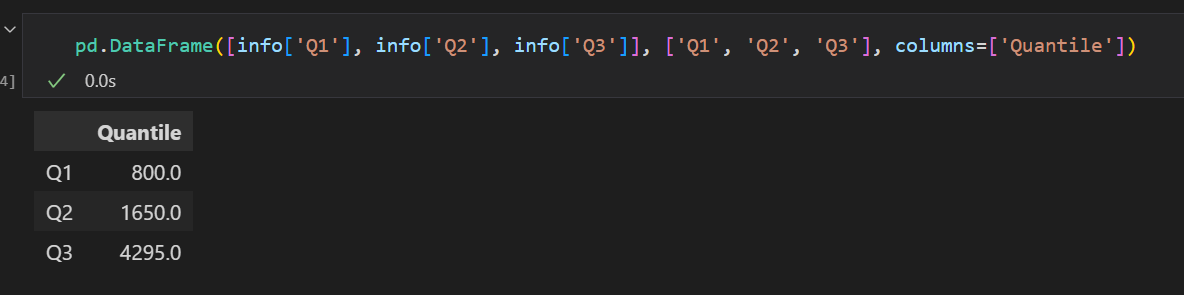
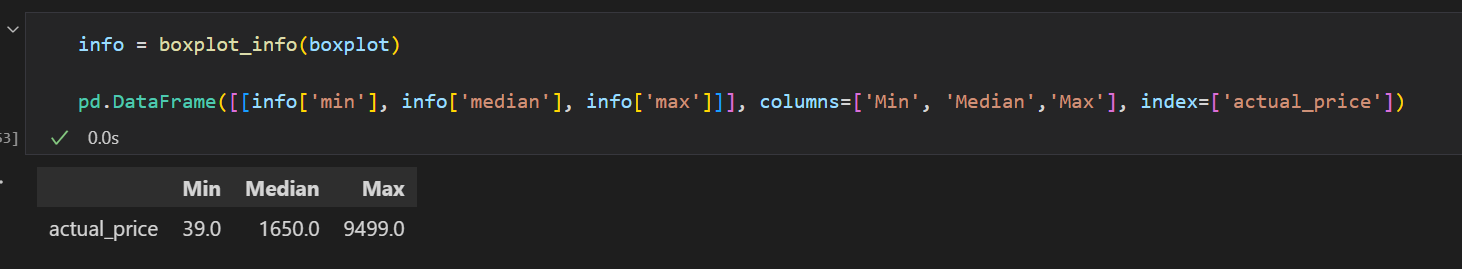
##### Фактична ціна товару



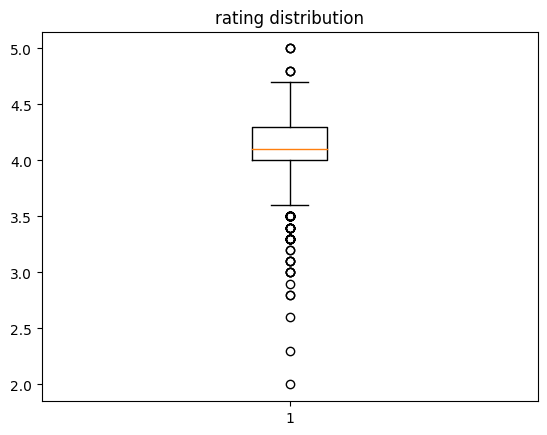


На *«скриньці з вусами»* для змінної *actual\_price* маємо:

* Мінімальне значення: 39.0
* Медіана (Q2): 1650.0
* Максимальне значення: 9499.0
* Нижній квартиль (Q1): 800.0
* Верхній квартиль: (Q3): 4295.0
* Присутні аномальні значення (викиди)

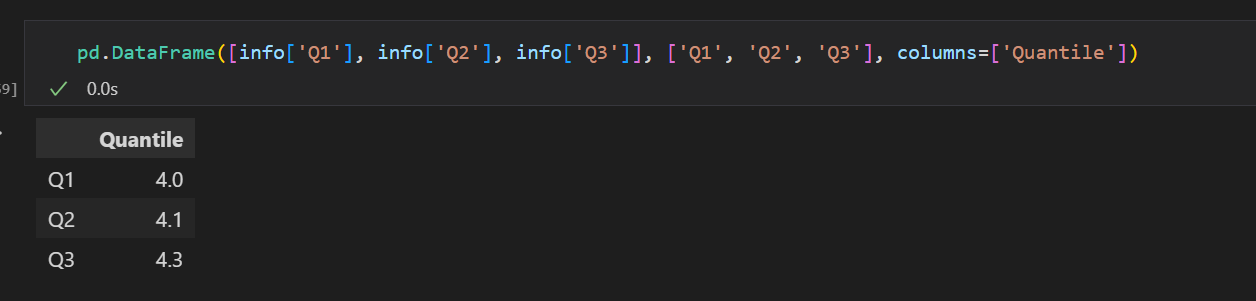
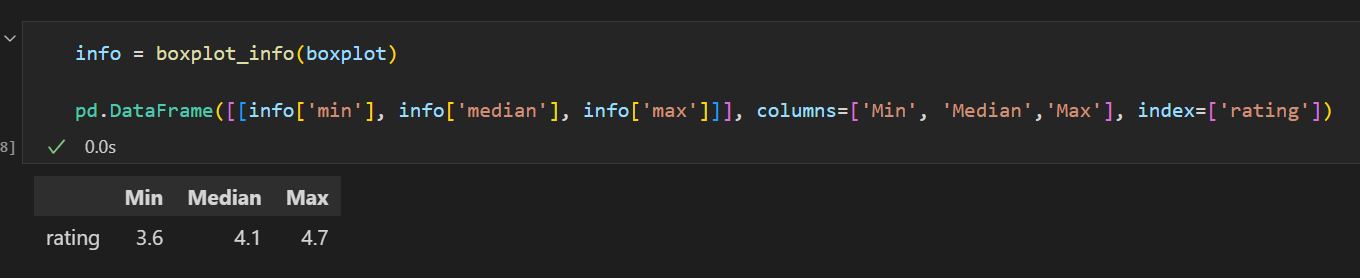


##### Рейтинг товару

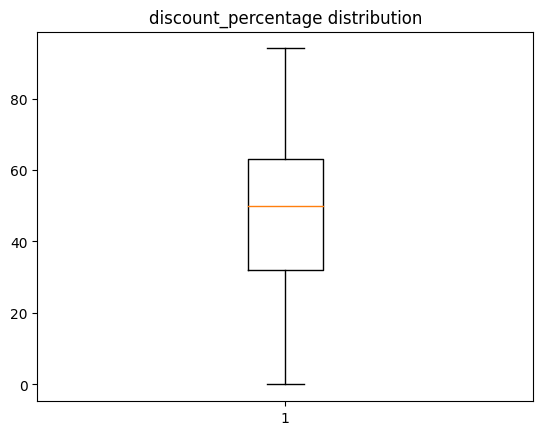


На *«скриньці з вусами»* для змінної *rating* маємо:

* Мінімальне значення: 3.6
* Медіана (Q2): 4.1
* Максимальне значення: 4.7
* Нижній квартиль (Q1): 4.0
* Верхній квартиль: (Q3): 4.3
* Присутні аномальні значення (викиди)

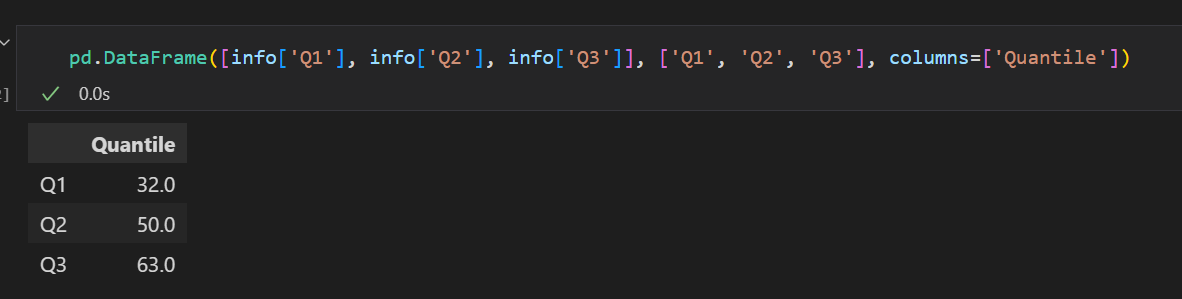
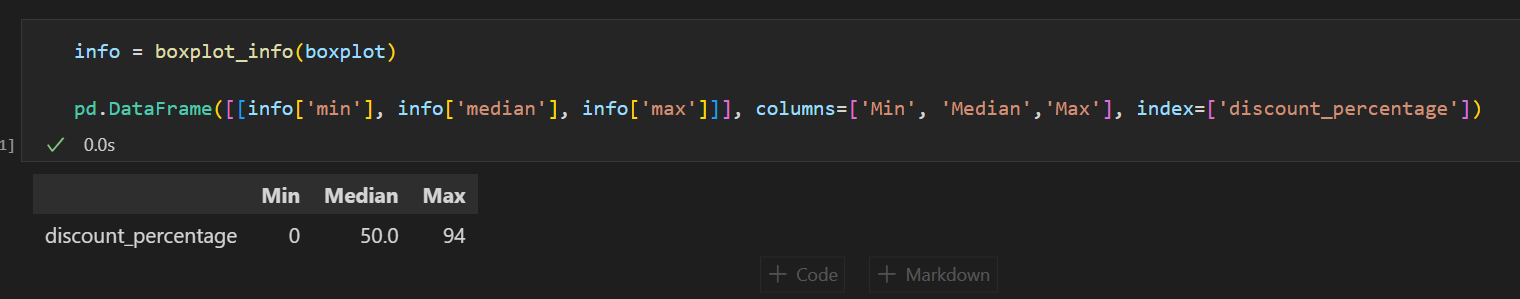


##### Відсоток знижки на товар



На *«скриньці з вусами»* для змінної *discount\_percentage* маємо:

* Мінімальне значення: 0
* Медіана (Q2): 50.0
* Максимальне значення: 94
* Нижній квартиль (Q1): 32.0
* Верхній квартиль: (Q3): 63.0
* Відсутні аномальні значення (викиди)



##### Висновок:

На графіку "Скринька з вусами" для змінної *actual\_price* можна побачити, що половина товарів мають ціну, меншу за 1650.0 рупій, чверть - меншу за 600 рупій та ¾ всіх мають меншу за 4295.0. Також бачимо, що за даними графіка,

у вибірці присутні аномальні спостереження, що не є правдою. Для підтвердження візьмемо найбільше значення - 139900 рупій (~1700$). Цим товаром є телевізор [Sony Bravia 164 cm (65 inches) 4K Ultra HD Smart LED Google TV KD-65X74K (Black)](https://www.amazon.in/Sony-Bravia-inches-Google-KD-65X74K/dp/B09WN3SRC7/ref=sr_1_291?qid=1672909138&s=electronics&sr=1-291&th=1), який належить до VIP сегменту телевізорів та має велику кількість відгуків, тому аргументовано є частиною вибірки.

Для змінної *rating* можемо зробити висновок, що більшість покупців були задоволені покупкою, 25% товарів мають оцінку нижчу за 4.0, 50% - нижчу за 4.1 і 75% нижчу за 4.7. Так само можна помітити аномальні спостереження, які насправді є нормальними. Для прикладу наведемо товар з мінімальною оцінкою 2.0, ним є обігрівач [Khaitan ORFin heater Home kitchen K0](https://www.amazon.in/Khaitan-ORFin-heater-Home-kitchen-K0/dp/B0BPJBTB3F/), який отримав таку оцінку через невелику кількість відгуків. Аналогічно для товару з найбільшою оцінкою 5.0 [Syncwire LTG to USB Cable for Fast Charging](https://www.amazon.in/Syncwire-Cable-Charging-Compatible-Devices/dp/B0BP7XLX48/ref=sr_1_194?qid=1672909134&s=electronics&sr=1-194).

Змінна *discount\_percentage* містить такі дані: 25% товарів мають знижку меншу за 32%, 50% - нижчу за 50% і 75% нижчу за 63%.

#### Мінімальне, максимальне значення, квартилі, децилі, медіана

Мінімальне значення - це найменше значення, яке може приймати даний набір даних.

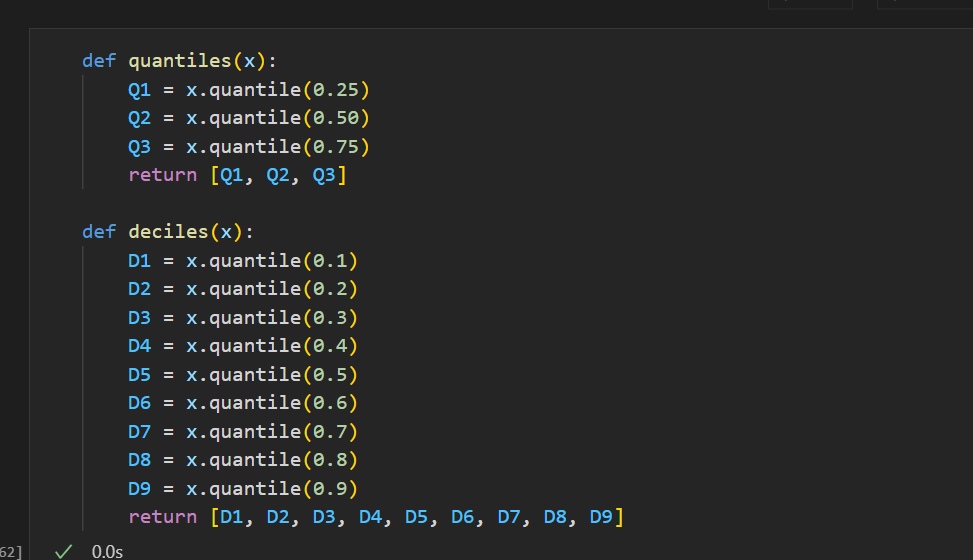
Максимальне значення - це найбільше значення, яке може приймати даний набір даних.

Квартилі - це значення, які розділяють впорядкований набір даних на чотири рівні, відповідно до їх позиції у розподілі. Перший квартиль (Q1) відповідає значенню, яке розділяє набір даних на нижні 25%, другий квартиль (Q2) - на нижні 50% (це також медіана), а третій квартиль (Q3) - на нижні 75%.

Децилі - це значення, які розділяють впорядкований набір даних на десять рівнів, відповідно до їх позиції у розподілі. Наприклад, перший дециль (D1) відповідає значенню, яке розділяє набір даних на нижні 10%, другий дециль (D2) - на нижні 20%, і так далі.

Медіана - це значення, яке розділяє впорядкований набір даних на дві рівні частини: нижню і верхню. Тобто, медіана відповідає значенню, яке стоїть посередині впорядкованого набору даних. Якщо кількість елементів у наборі непарна, то медіана - це значення серединного елемента, а якщо кількість елементів парна, то медіана - це середнє арифметичне значень двох серединних елементів.

***Функції для спрощення знаходження характеристик***



##### Фактична ціна товару

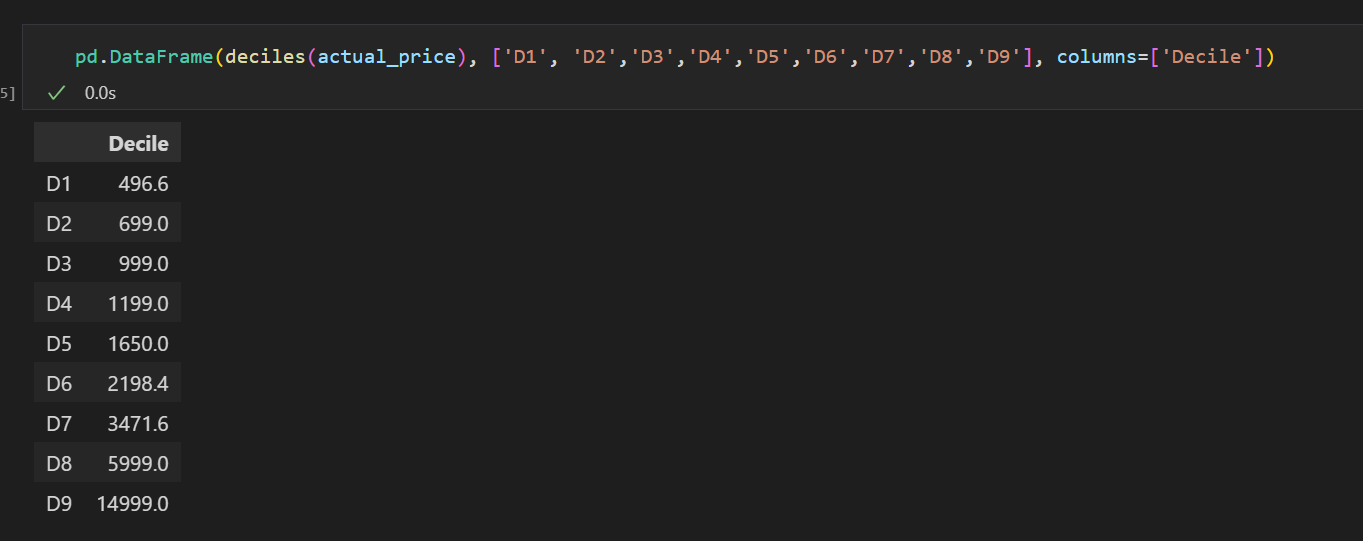
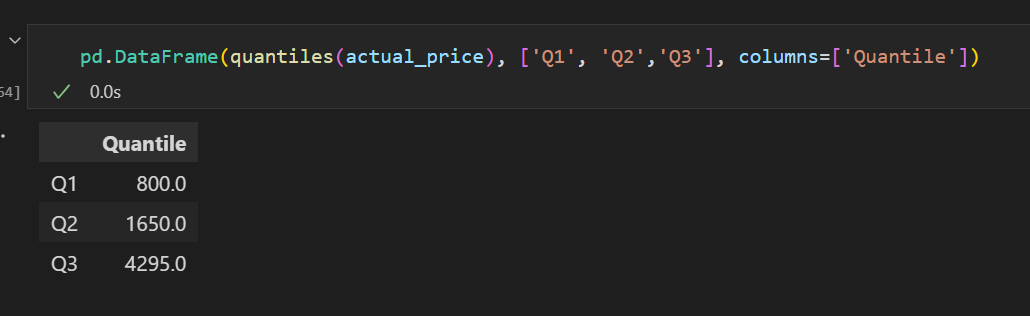
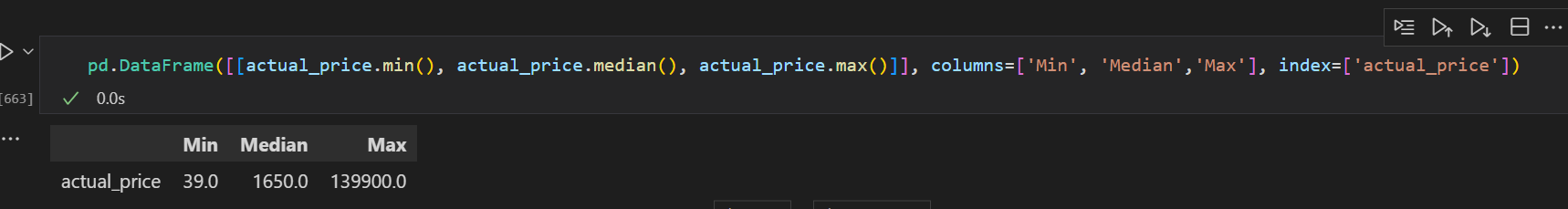
|  |  |  |
| --- | --- | --- |
| Min | Median | Max |
| 39.0 | 1650.0 | 139900.0 |

*Квартилі*

|  |  |  |
| --- | --- | --- |
| Q1 | Q2 | Q3 |
| 800.0 | 1650.0 | 4295.0 |

*Децилі*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 |
| 496.6 | 699.0 | 999.0 | 1199.0 | 1650.0 | 2198.4 | 3471.6 | 5999.0 | 14999 |



##### Рейтинг товару

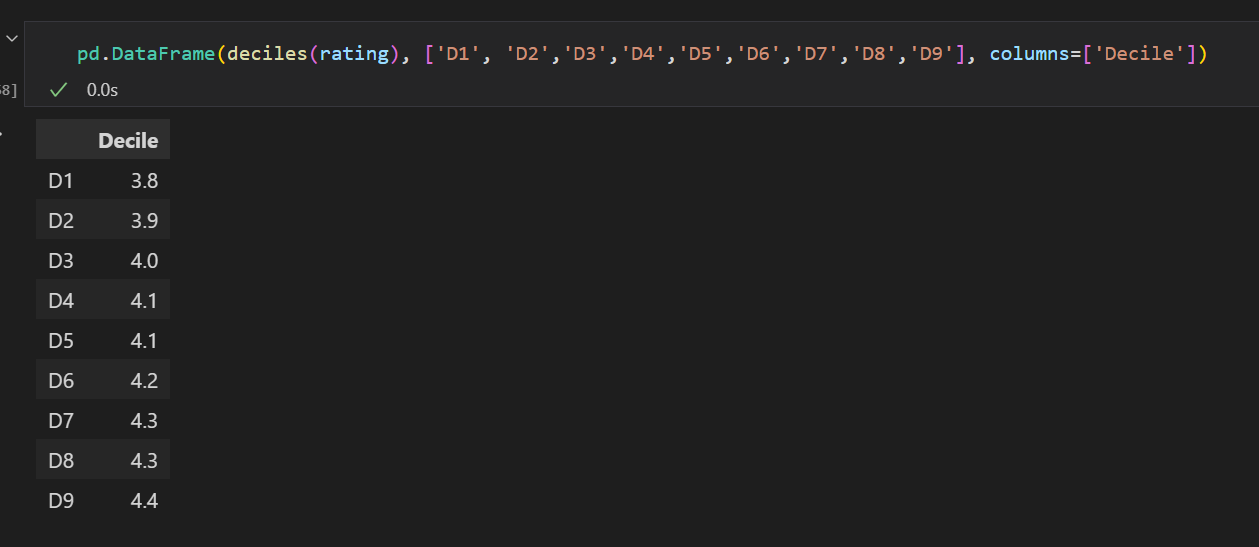
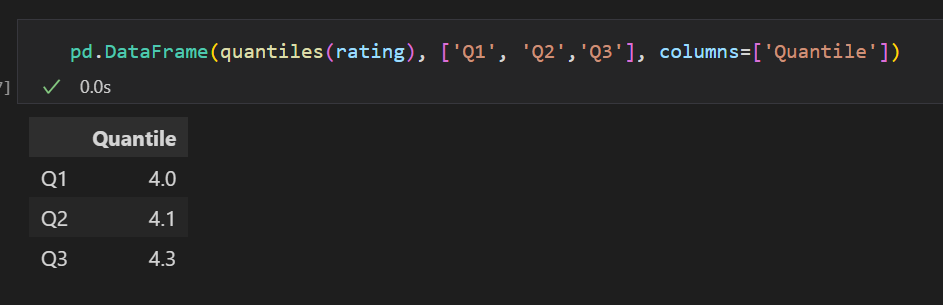
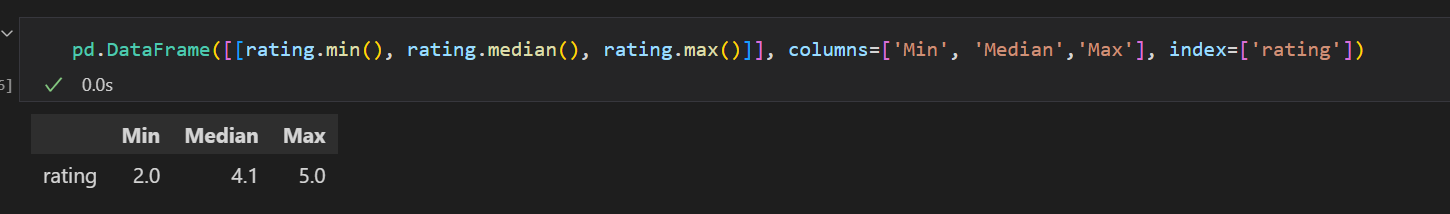
|  |  |  |
| --- | --- | --- |
| Min | Median | Max |
| 2.0 | 4.1 | 5.0 |

*Квартилі*

|  |  |  |
| --- | --- | --- |
| Q1 | Q2 | Q3 |
| 4.0 | 4.1 | 4.3 |

*Децилі*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 |
| 3.8 | 3.9 | 4.0 | 4.1 | 4.1 | 4.2 | 4.3 | 4.3 | 4.4 |



##### Відсоток знижки на товар

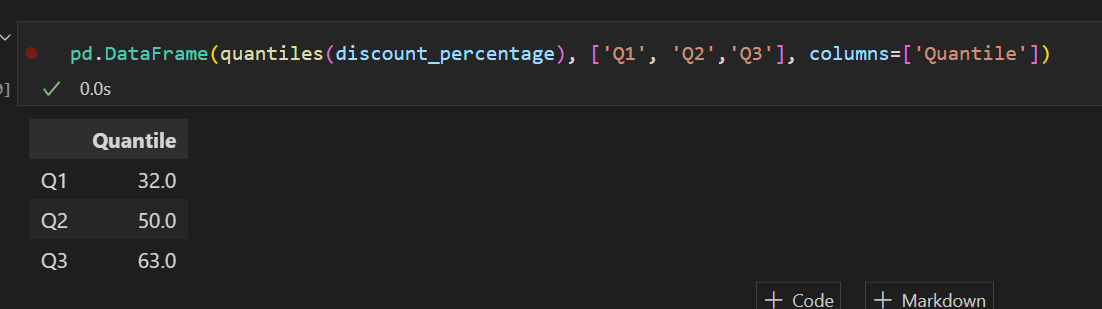
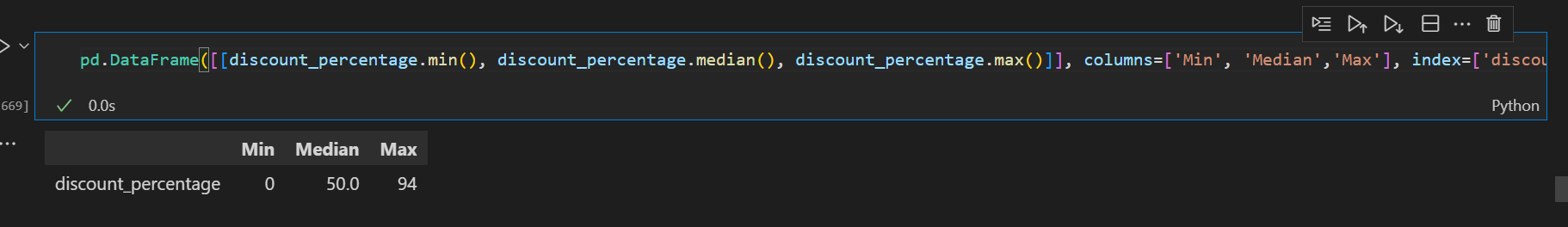
|  |  |  |
| --- | --- | --- |
| Min | Median | Max |
| 0 | 50.0 | 94 |

*Квартилі*

|  |  |  |
| --- | --- | --- |
| Q1 | Q2 | Q3 |
| 32.0 | 50.0 | 63.0 |

*Децилі*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 |
| 18.0 | 28.0 | 37.0 | 43.0 | 50.0 | 55.0 | 60.0 | 66.0 | 76.0 |



##### Висновок:

Було обчислено мінімальне та максимальне значення, нижній та верхній квантилі, медіану та децилі, які можуть свідчити про велику варіативність елементів у вибірці, наприклад, якщо очевидна велика різниця між мінімальним та максимальним значеннями в усіх змінних.

Оскільки для змінних *actual\_price* та *rating* не було відкинуто аномальні спостереження за версію графіку “Скринька з вусами” то інформація про мінімальне та максимальне значення змінилась.

Співставимо фактичну інформацію з інформацією з “Скриньки з вусами”

*actual\_price*

|  |  |  |
| --- | --- | --- |
|  | *Фактично* | *Скринька з вусами* |
| *min* | 39.0 | 39.0 |
| *max* | 139900.0 | 9499.0 |

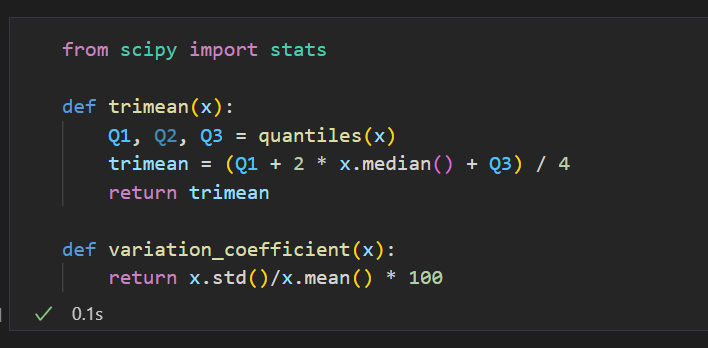
*rating*

|  |  |  |
| --- | --- | --- |
|  | *Фактично* | *Скринька з вусами* |
| *min* | 2.0 | 3.6 |
| *max* | 5.0 | 4.7 |

#### Характеристики положення центра значень змінної

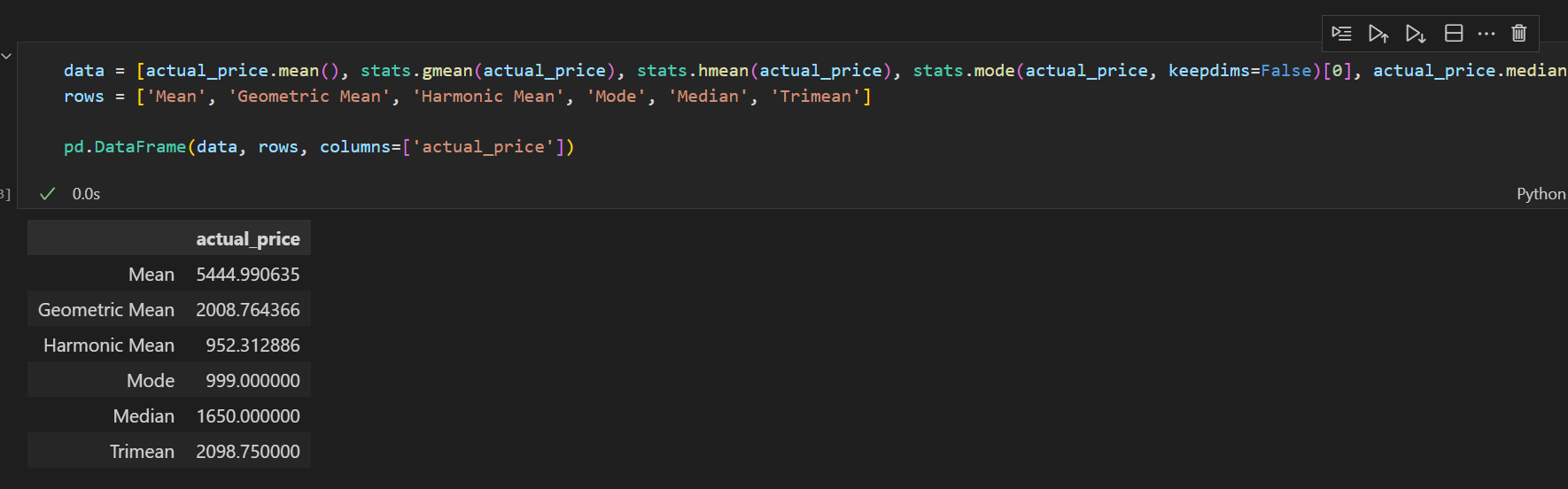
*Характеристики положення центра значень змінної* - це числові показники, які відображають центральність розподілу даних, тобто де більшість даних зосереджена в наборі. Вони є важливими характеристиками статистичного аналізу, оскільки вони дають загальний огляд про те, де знаходяться значення змінної.

***Функція для Trimean***



##### Фактична ціна товару

1. Математичне сподівання: 5444.990635
2. Середнє геометричне: 2008.764366
3. Середнє гармонічне: 952.312886
4. Moдa: 999.000000
5. Медіана: 1650.000000
6. Тримеана: 2098.750000



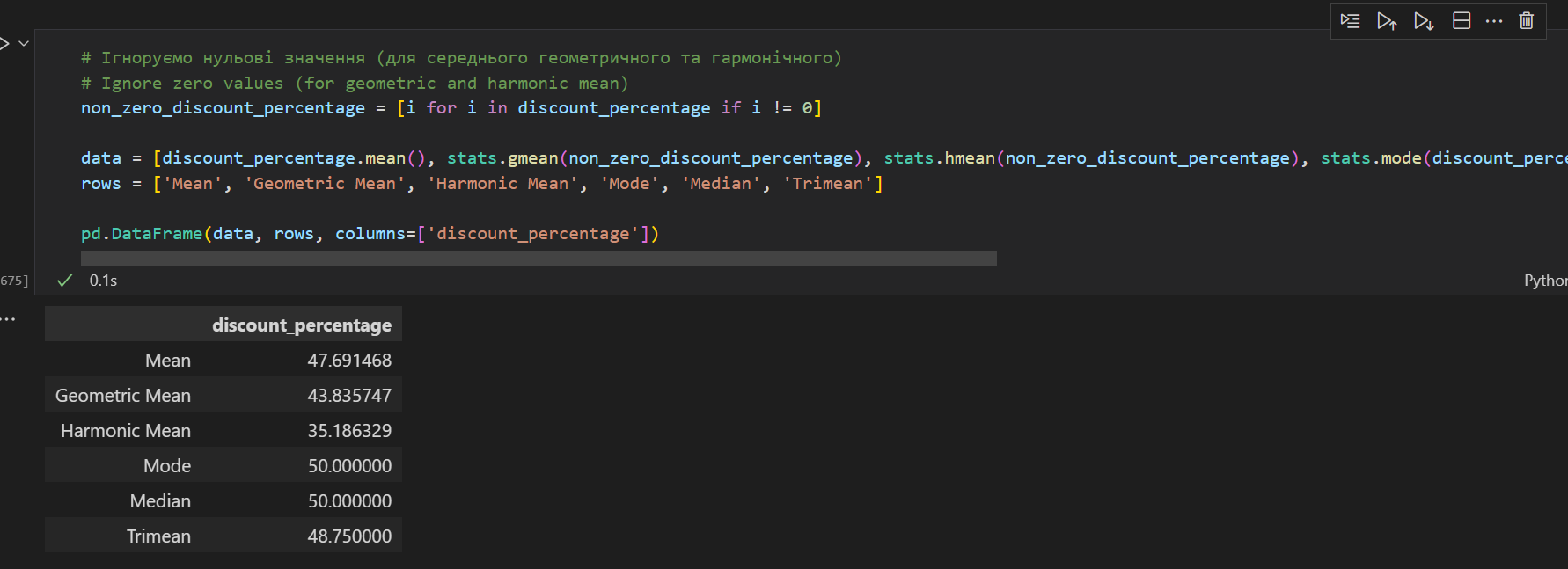
##### Рейтинг товару

1. Математичне сподівання: 4.096585
2. Середнє геометричне: 4.085348
3. Середнє гармонічне: 4.072960
4. Moдa: 4.100000
5. Медіана: 4.100000
6. Тримеана: 4.125000



##### Відсоток знижки на товар

1. Математичне сподівання: 47.691468
2. Середнє геометричне: 43.835747
3. Середнє гармонічне: 35.186329
4. Moдa: 50.000000
5. Медіана: 50.000000
6. Тримеана: 48.750000



##### Висновок:

Для змінної *actual\_price* можна зробити висновок, що характеристики положення центра значень змінної мають різне значення. Зокрема, середнє арифметичне та математичне сподівання досить близькі між собою, що може свідчити про відсутність виразної асиметрії у розподілі. У той же час, мода та медіана мають відмінні значення, що може вказувати на те, що розподіл має виразну асиметрію.

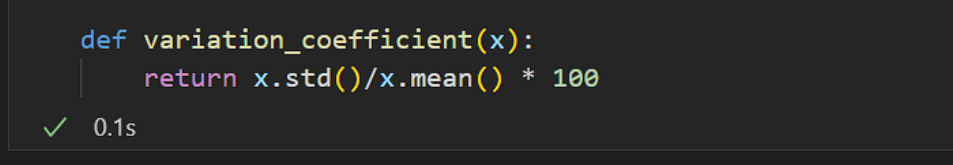
Мода та медіана розподілу змінної *rating*  майже ідентичні та дорівнюють 4.1, що може вказувати на те, що розподіл має незначну асиметрію. Середнє гармонічне та тримеана мають дещо менші значення, але все ж є близькими до інших характеристик центру розподілу.

Для розподілу змінної *discount\_percentage* можна зробити висновок, що різні характеристики центру розподілу мають значну відмінність між собою. Наприклад, середнє геометричне та середнє гармонічне мають значно менші значення в порівнянні з іншими характеристиками. Медіана та мода мають однакові значення та дорівнюють 50, що може свідчити про симетричний розподіл.

#### 5. Характеристики розсіювання значень змінної

*Характеристики розсіювання значень змінної* - це числові показники, які відображають ступінь розкиду даних відносно центрального значення (наприклад, середнього арифметичного) в наборі даних. Вони є важливими характеристиками статистичного аналізу, оскільки вони дозволяють оцінити, наскільки великий розмах змінності можливий в наборі даних.

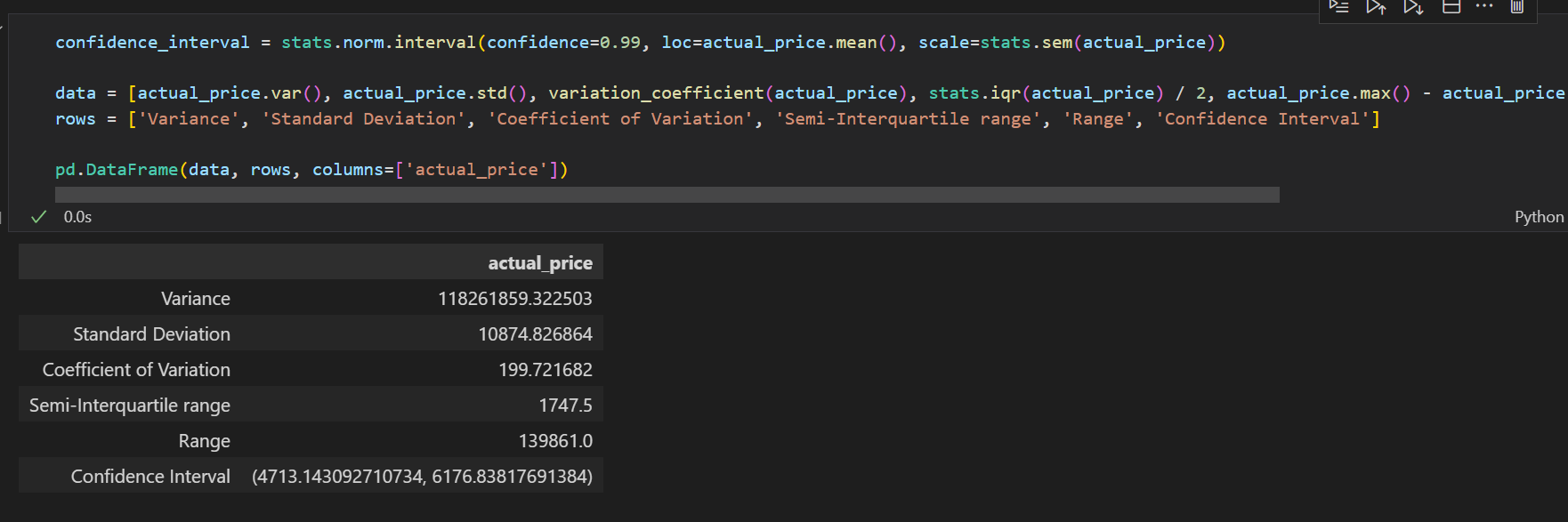
***Функція пошуку коефіцієнта варіації***



##### 

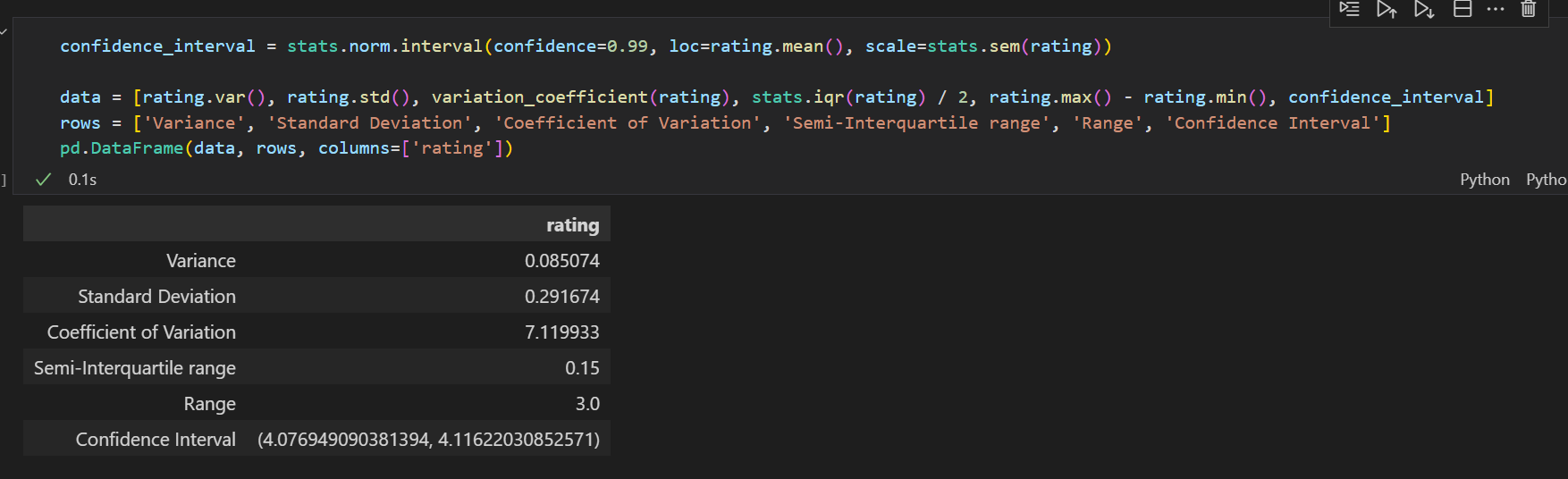
##### Фактична ціна товару

1. Дисперсія: 118261859.322503
2. Стандартне відхилення: 10874.826864
3. Коефіцієнт варіації: 199.721682
4. Ймовірнісне відхилення : 1747.5
5. Розмах вибірки: 139861.0
6. Інтервал концентрації: (4713.143092710734, 6176.83817691384)



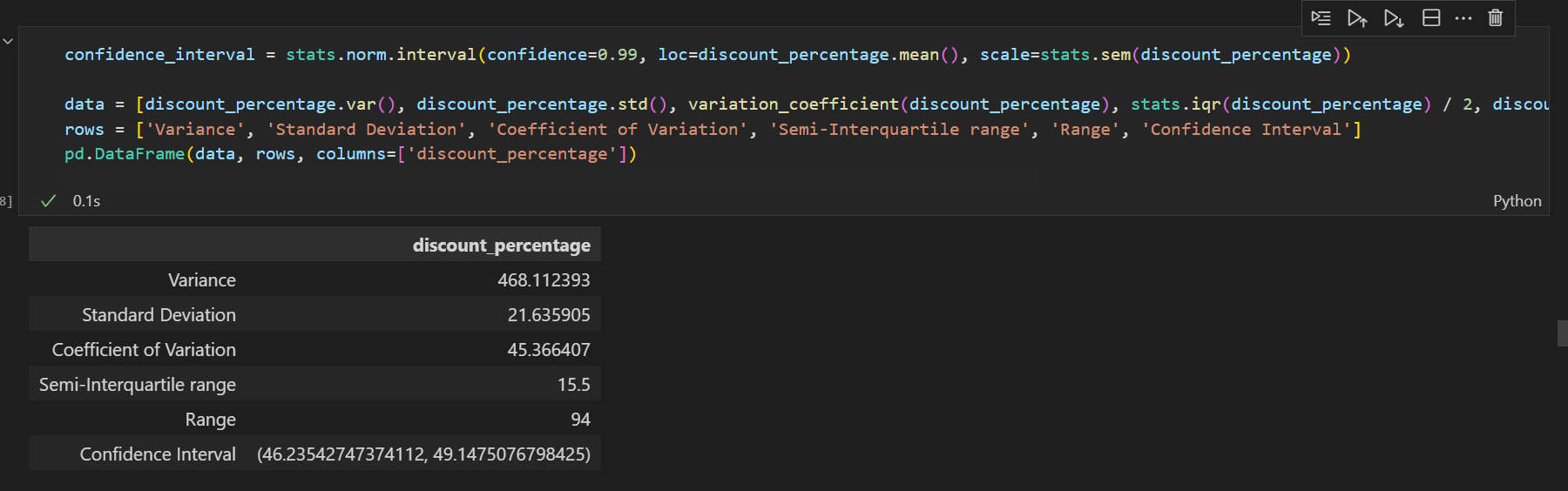
##### Рейтинг товару

1. Дисперсія: 0.085074
2. Стандартне відхилення: 0.291674
3. Коефіцієнт варіації: 7.119933
4. Ймовірнісне відхилення: 0.15
5. Розмах вибірки: 3.0
6. Інтервал концентрації: (4.076949090381394, 4.11622030852571)



##### Відсоток знижки на товар

1. Дисперсія: 468.112393
2. Стандартне відхилення: 21.635905
3. Коефіцієнт варіації: 45.366407
4. Ймовірнісне відхилення: 15.5
5. Розмах вибірки: 94
6. Інтервал концентрації: (46.23542747374112, 49.1475076798425)



##### Висновок:

Для розподілу змінної *actual\_price* маємо значну варіацію, велику дисперсію та стандартне відхилення. Коефіцієнт варіації досить великий, що також свідчить про значну розбіжність в даних. Ймовірнісне відхилення показує, що значення вибірки можуть відрізнятися на 1747.5 від стандартного відхилення, що є достатньо великим значенням. Інтервал концентрації вказує на те, що більшість значень знаходиться в цьому діапазоні, що може бути корисною інформацією при аналізі даних.

За даними характеристиками можна сказати, що розподіл змінної *rating* має досить великий коефіцієнт варіації, що свідчить про значну розбіжність значень у вибірці. Розмах вибірки складає 3 одиниці, що означає, що значення знаходяться в дуже вузькому діапазоні.

Змінна *discount\_percentage:* Цей розподіл має велику дисперсію та стандартне відхилення, що свідчить про значну варіативність даних. Коефіцієнт варіації дуже високий, що означає, що вибірка містить значну кількість випадкових відхилень від середнього значення.

#### 6. Аналіз скошеності та гостроверхості розподілу

* *Коефіцієнт асиметрії*

Додатні значення коефіцієнта свідчать про те, що розподіл має додатню асиметрію (тобто він нахилений вправо), від'ємні значення коефіцієнта свідчать про те, що розподіл має від'ємну асиметрію (тобто він нахилений вліво), а значення близьке до нуля означає, що розподіл має симетричний характер.

* *Коефіцієнт ексцесу (гостроверхості)*

Значення коефіцієнта ексцесу може бути додатним, від'ємним або дорівнювати нулю. Додатні значення коефіцієнта свідчать про те, що розподіл є більш гостроверхим (або збільшення кількості викидів вздовж головної моди) ніж нормальний розподіл, а від'ємні значення коефіцієнта свідчать про те, що розподіл є менш гостроверхим (або менше кількість викидів вздовж головної моди) ніж нормальний розподіл. Значення близьке до нуля означає, що розподіл має приблизно нормальний характер заокругленості.

##### Фактична ціна товару

*Коефіцієнт асиметрії:* 4.555209

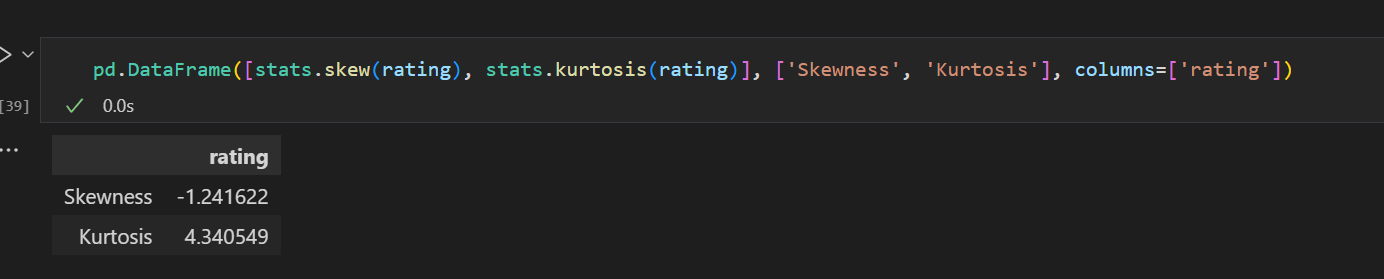
*Коефіцієнт ексцесу:* 29.617642

##### 

##### Рейтинг товару

*Коефіцієнт асиметрії:* -1.241622

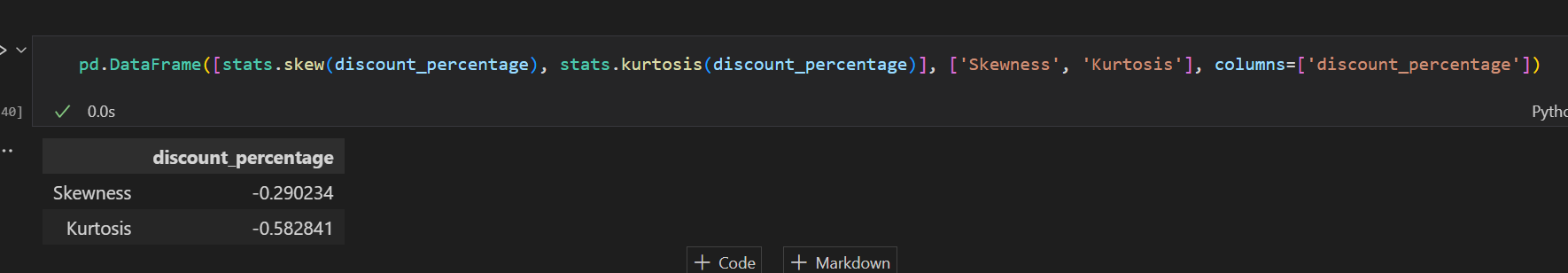
*Коефіцієнт ексцесу:* 4.340549

****

##### Відсоток знижки на товар

*Коефіцієнт асиметрії:* -0.290234

*Коефіцієнт ексцесу:* -0.582841



##### Висновок:

Розподіл змінної *actual\_priсe* можна зробити висновок про те, що розподіл є дуже асиметричним та має високий пік.

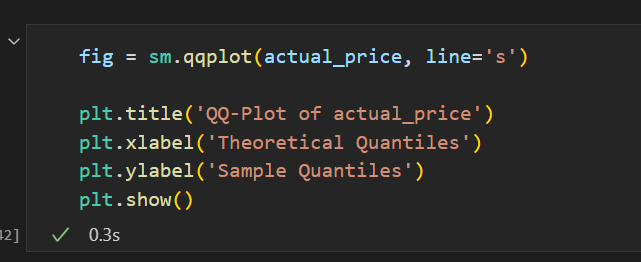
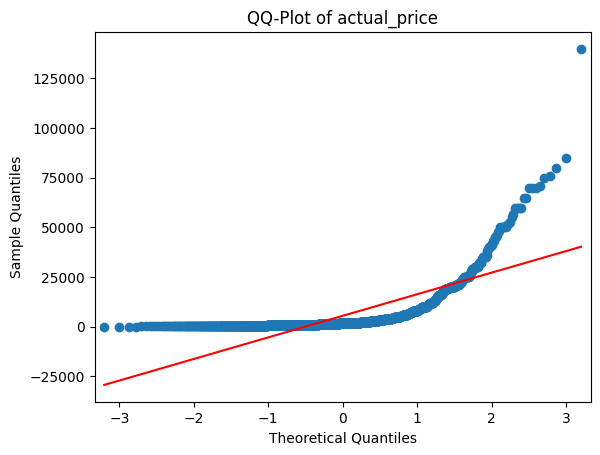
За коефіцієнтом асиметрії вважаємо, що розподіл змінної *rating* має легку ліву асиметрію. За коефіцієнтом ексцесу можна сказати, що розподіл має деяку виразність вершини.

Для змінної *discount\_percentage* можна зробити висновок, що розподіл є приблизно симетричним та має менше виразний пік, ніж нормальний розподіл. Також, коефіцієнт ексцесу від'ємний, що свідчить про те, що розподіл є менш загостреним в порівнянні з нормальним розподілом.

#### 7. Перевірка нормальності

Перевірка розподілу на нормальність є важливою процедурою в багатьох областях науки, таких як статистика, економіка, біологія, інженерія та інші. Ця перевірка дозволяє досліднику встановити, чи можна застосовувати статистичні методи, що передбачають нормальний розподіл даних, до конкретного набору даних.Фактична ціна товару

###### Графік Q-Q

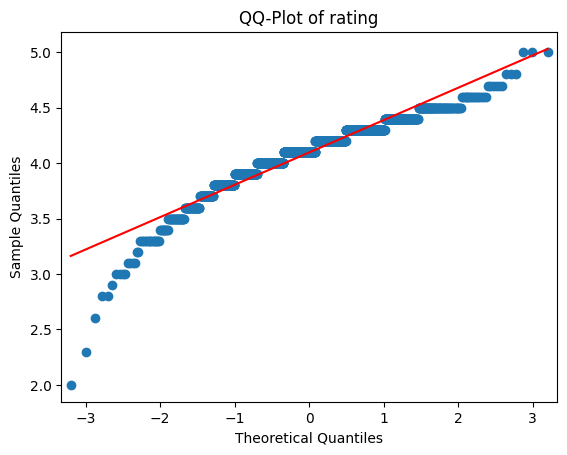


###### Критерій узгодженості Колмогорова-Смірнова

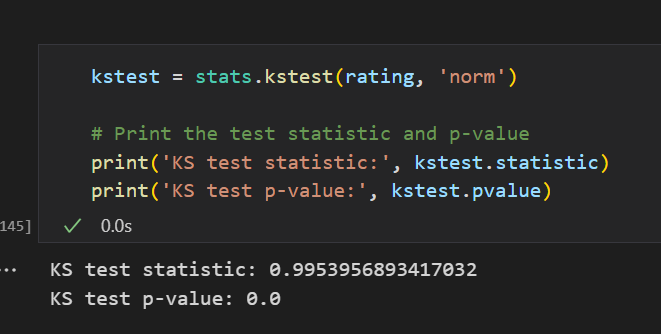
###### 

##### Рейтинг товару

###### Графік Q-Q

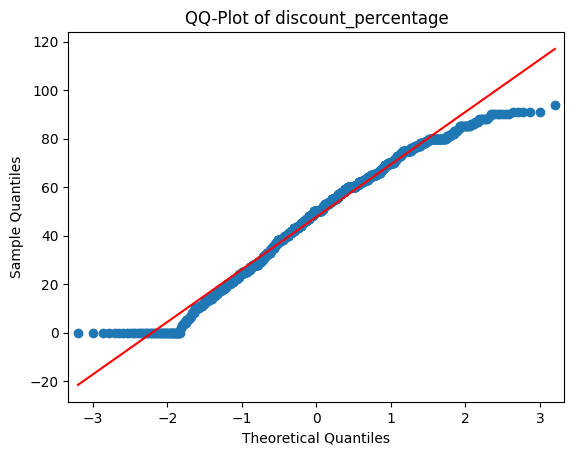


###### Критерій узгодженості Колмогорова-Смірнова

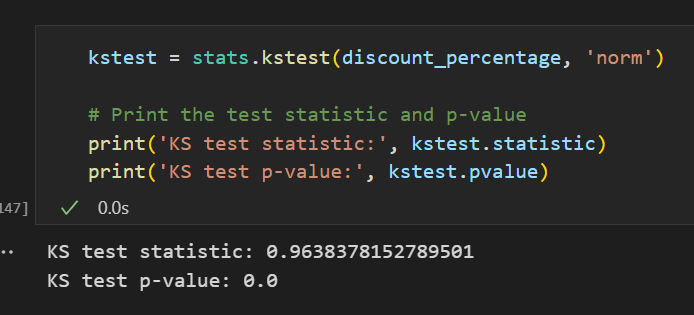


##### Відсоток знижки на товар

###### Графік Q-Q



###### Критерій узгодженості Колмогорова-Смірнова



##### Висновок:

Для розподілу змінної *actual\_price* можемо зробити висновок, що він точно не нормальний і не подібний до нього, оскільки ні qq-plot ні тест Колмогорова-Смірнова не дав позитивного результату.

Розподіл змінної *rating* не є нормальним за тими ж самими критеріями.

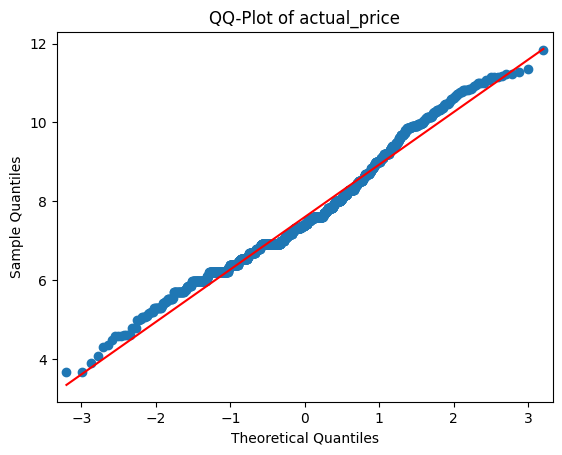
Змінна *discount\_percentage* дуже схожа на нормальний розподіл, зокрема містить аномалії на краях графіку qq-plot. Розподіл змінної не проходить тест Колмогорова-Смірнова, хоча дуже близький до нормального розподілу. Також можна взяти до уваги коефіцієнт асиметрії та коефіцієнт ексцесу, які показують близькість до характеристик нормального розподілу.

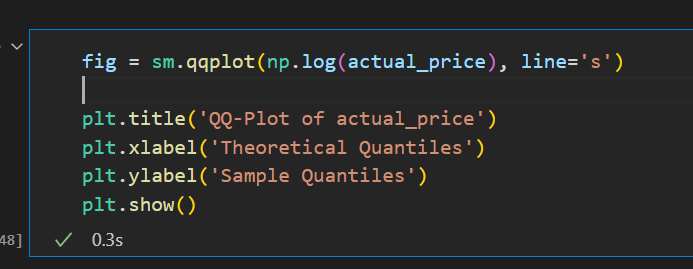
#### 8. Інші процедури попереднього аналізу

##### *Порівняння розподілу Фактичної Ціни Товару з Логнормальним розподілом*

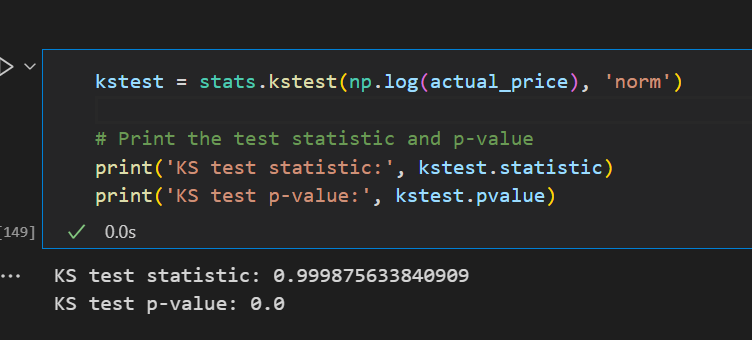
***Логнормальний розподіл*** - це розподіл, у якому натуральний логарифм змінної має нормальний розподіл.

###### Графік Q-Q





###### Критерій узгодженості Колмогорова-Смірнова



##### Висновок:

Зважаючи на результат QQ-Plot можна стверджувати, що розподіл змінної actual\_price дуже близький до Логнормального розподілу, але не є повністю Логнормальним розподілом оскільки не проходить тест Колмогорова-Смірнова.

### 5. ВИСНОВКИ

Сформулюйте основні висновки по всій виконаній ЛР. Вкажіть переваги та недоліки використаного апарату для аналізу даних.

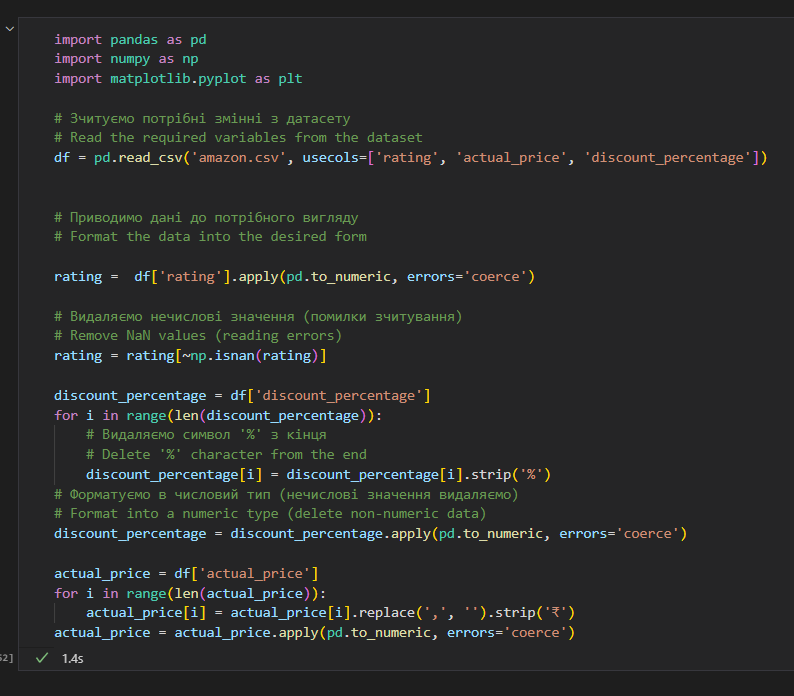
В даній лабораторній роботі було виконано попередній аналіз даних Amazon Sales Dataset. Застосовуючи Python, були проаналізовані особливості змінних actual\_price, rating та discount\_percentage.

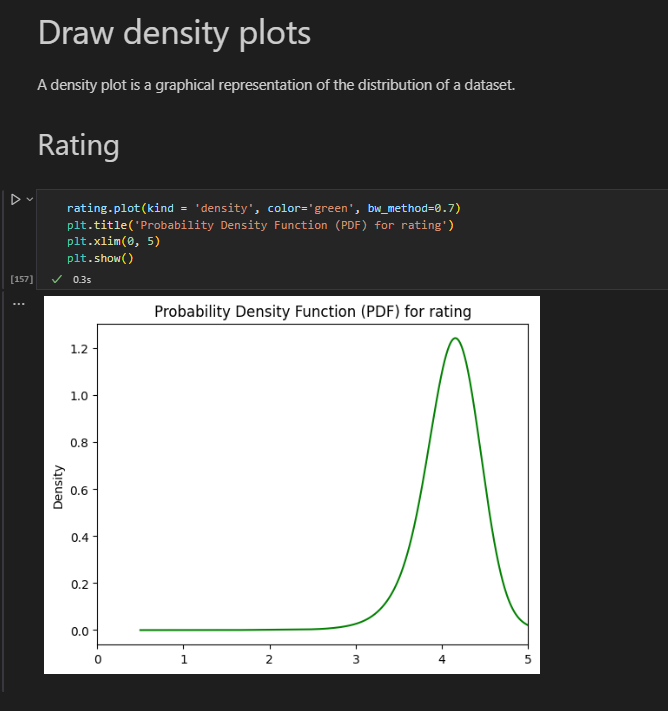
Результатом аналізу були визначені характеристики вибірок, такі як значення положення центру та розкиду змінних, мінімальні та максимальні значення, медіани та квартилі.

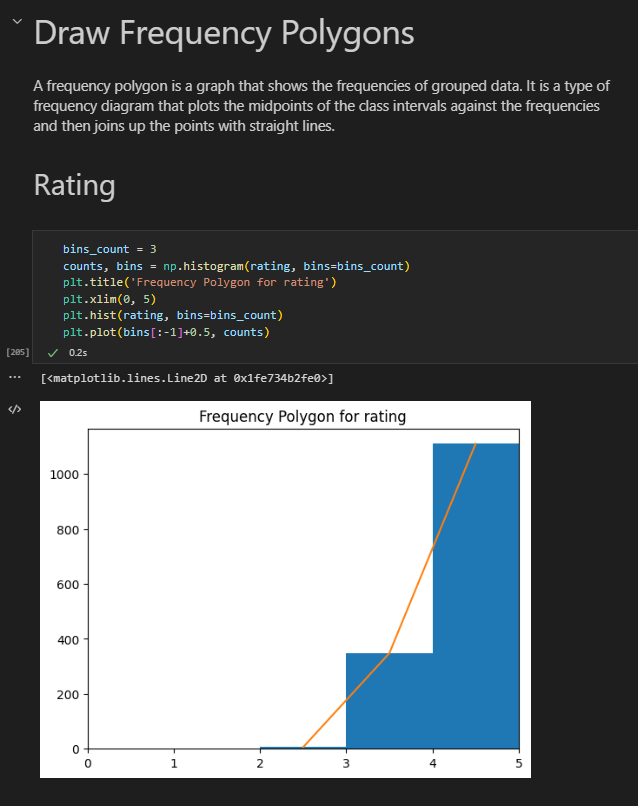
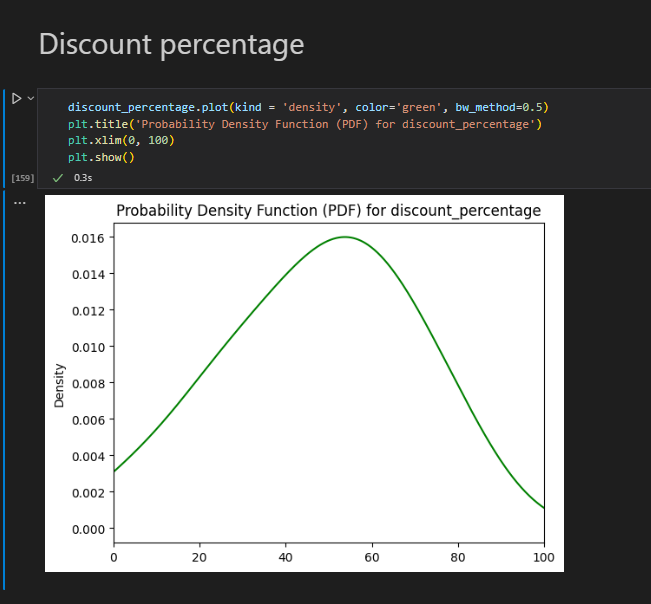
Серед характеристик центру були обчислені значення математичного сподівання, середнього геометричного, середнього гармонічного та моди. Серед характеристик розкиду були обчислені дисперсія, стандартне відхилення, коефіцієнт варіації, ймовірнісне відхилення, розмах вибірки та інтервал концентрації. Ці значення допомогли визначити скошеність та гостроту розподілів, що було досягнуто знаходженням коефіцієнтів асиметрії та ексцесу та їх інтерпретацією.

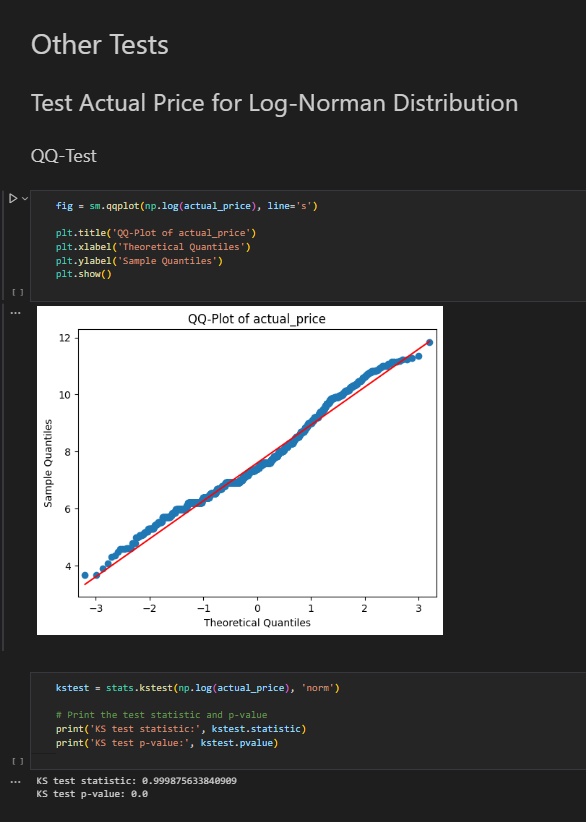
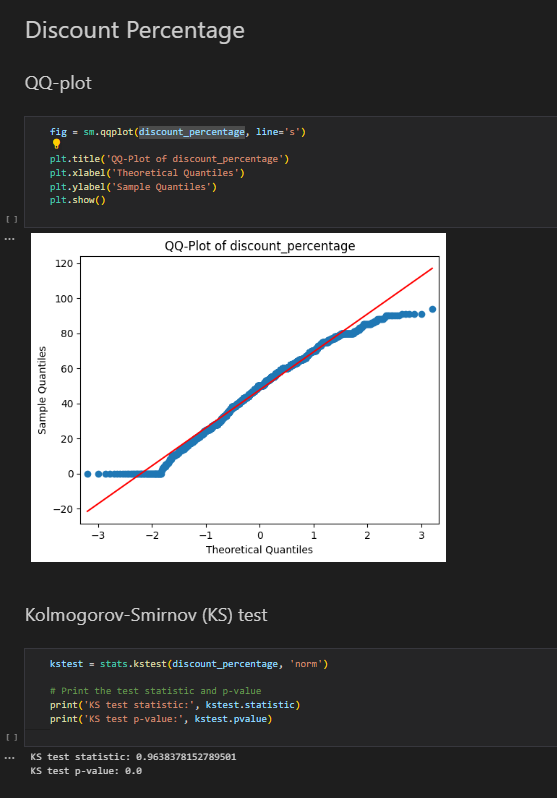
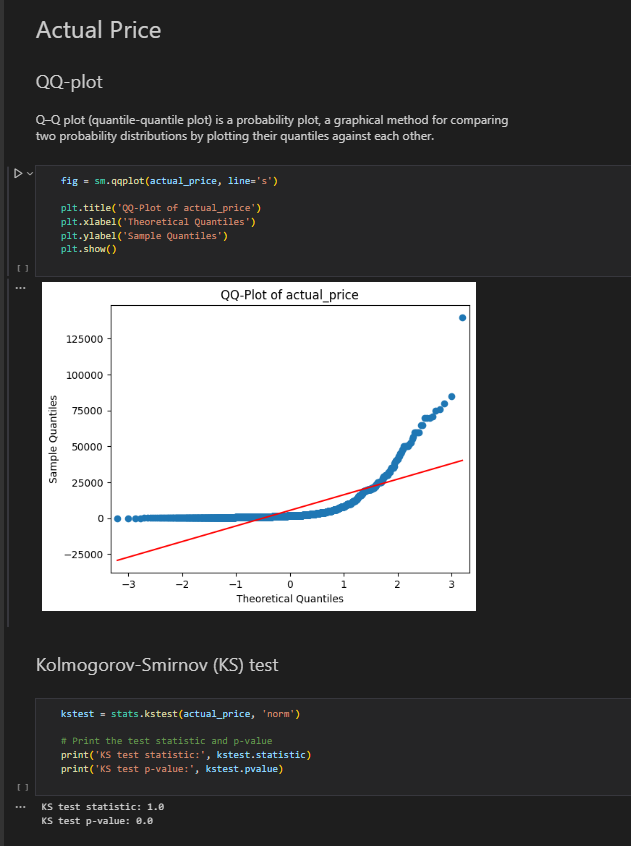
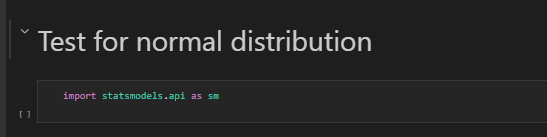
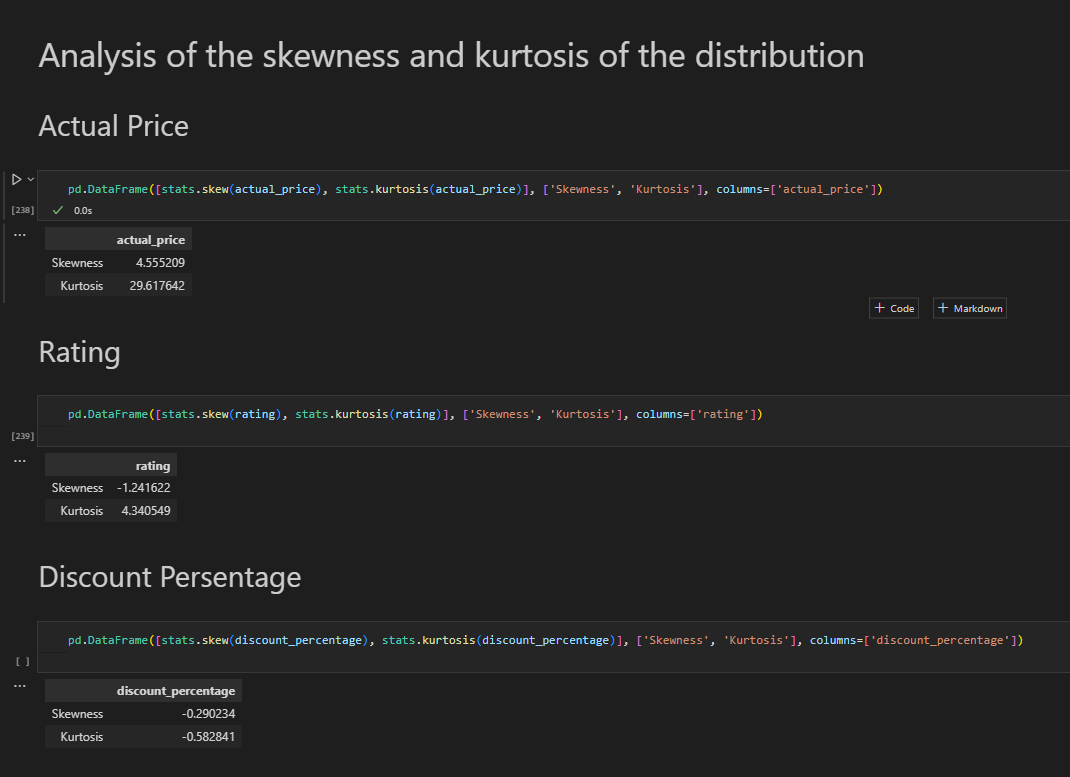
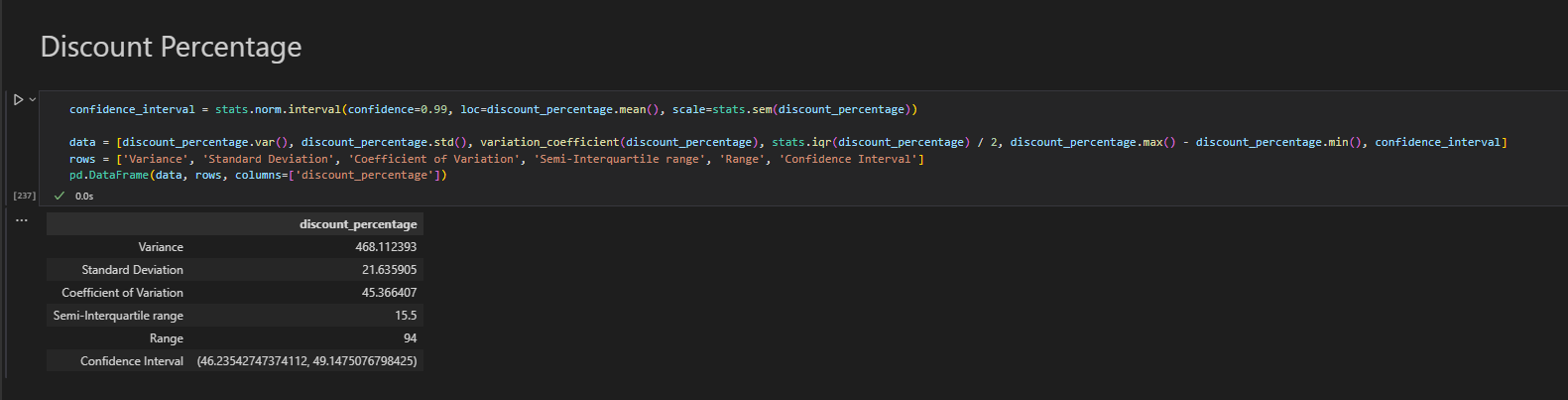
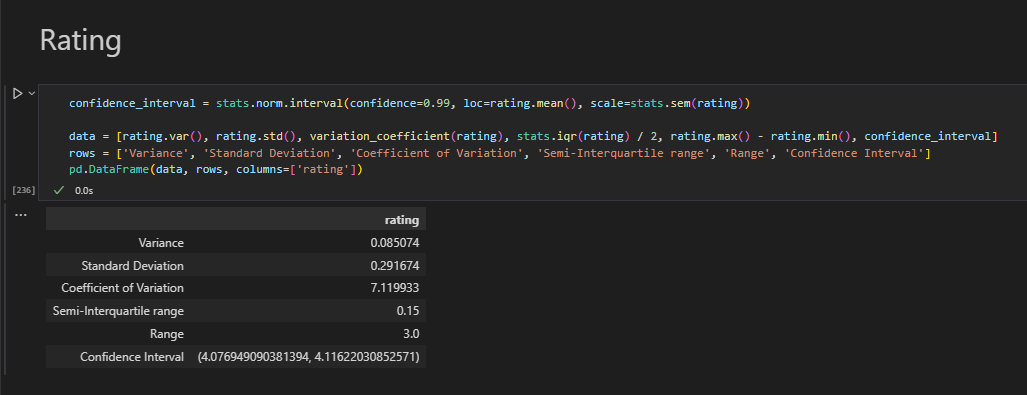
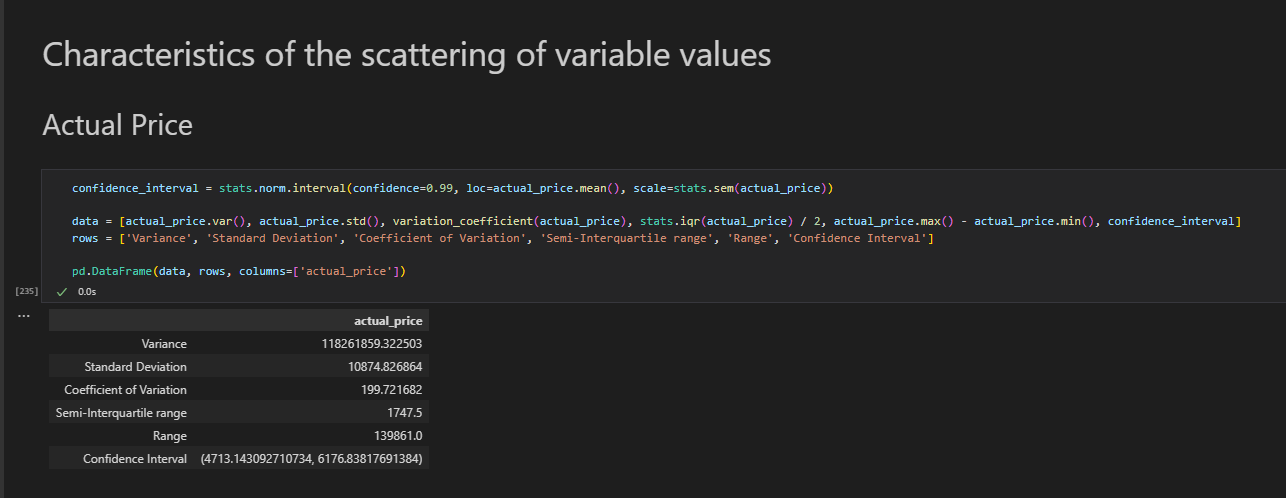
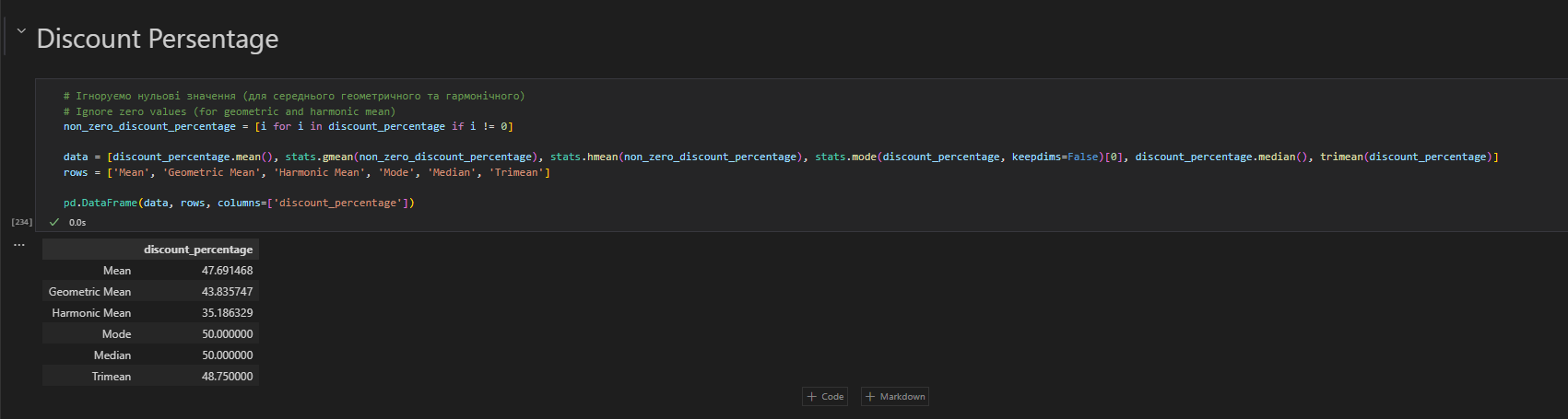
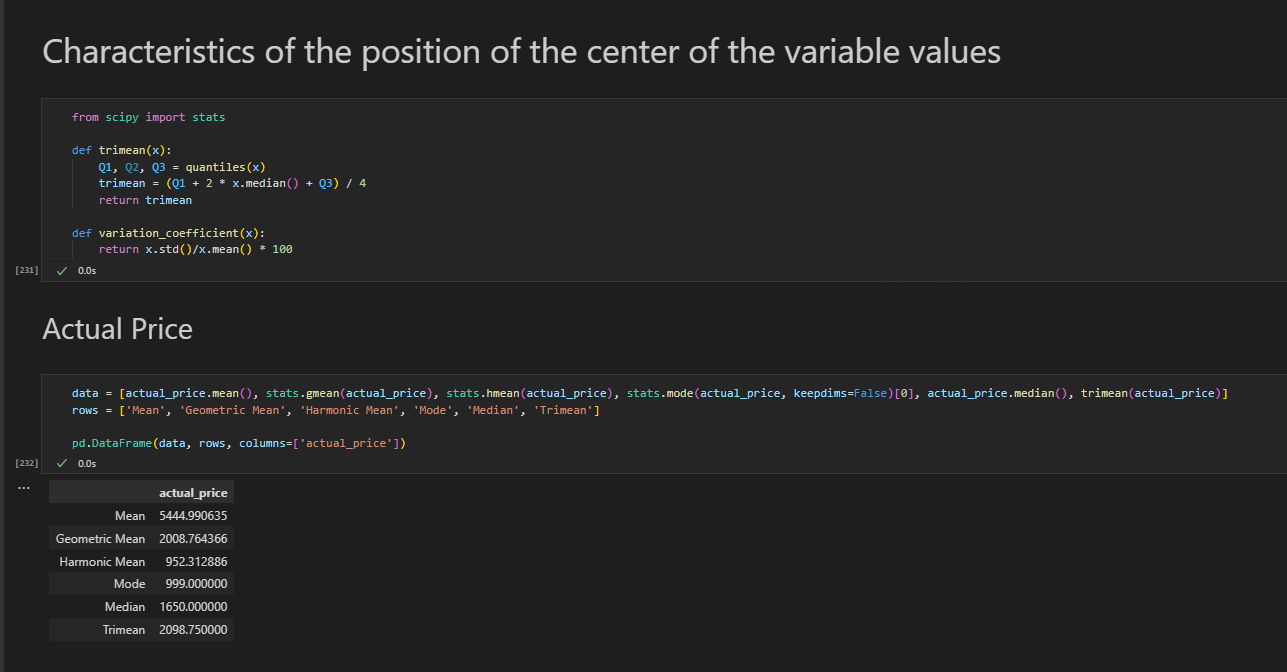
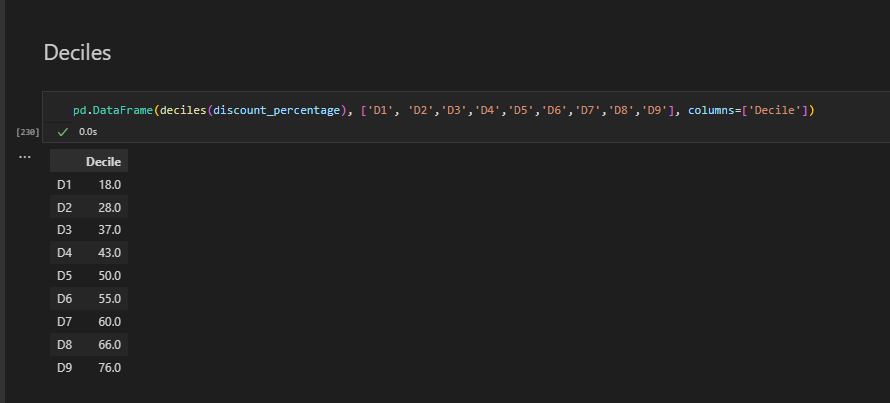
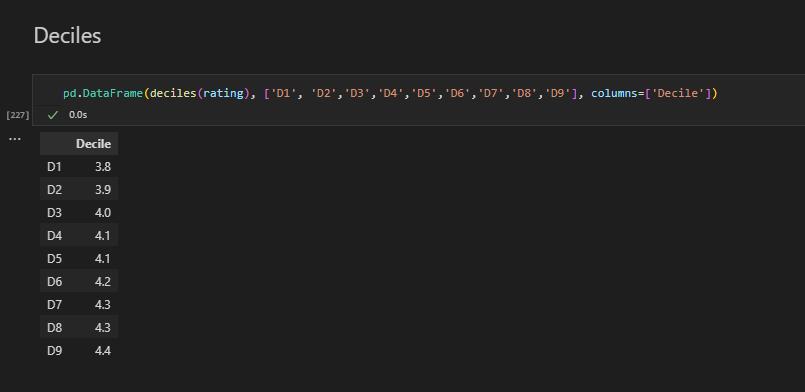
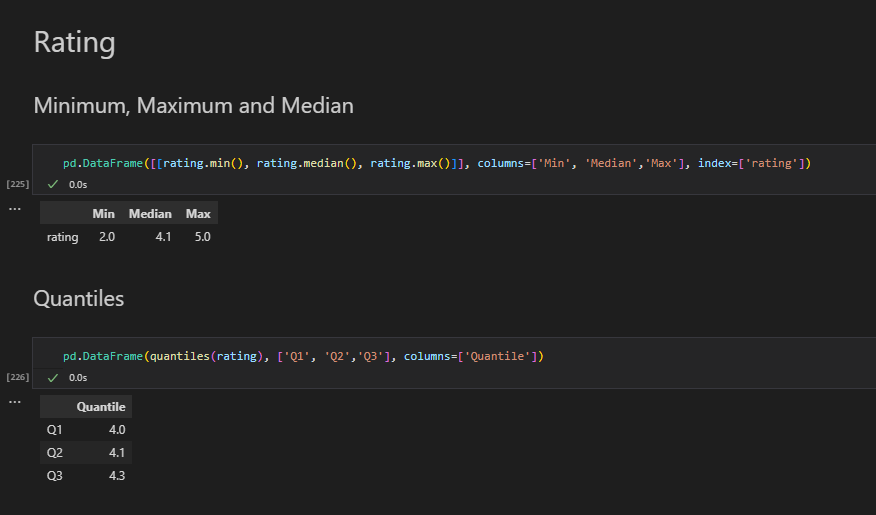
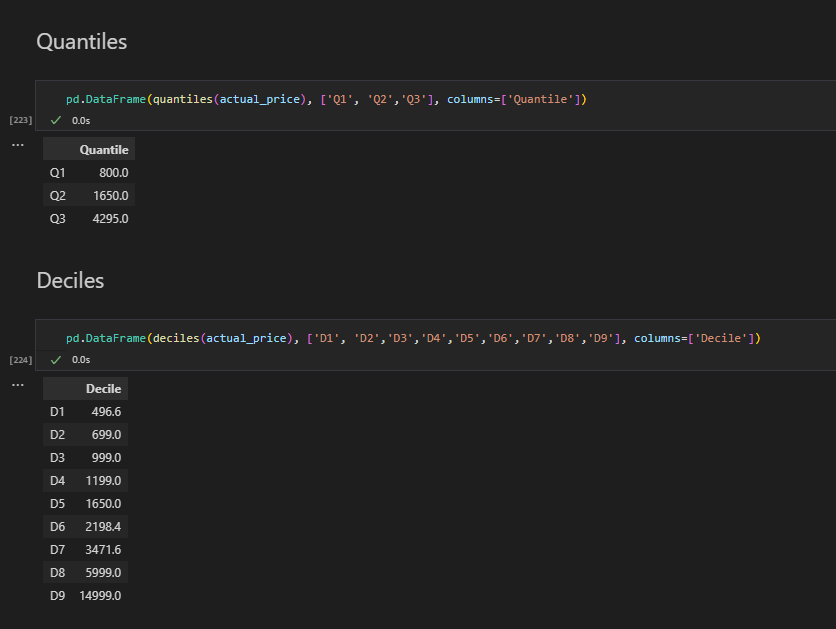
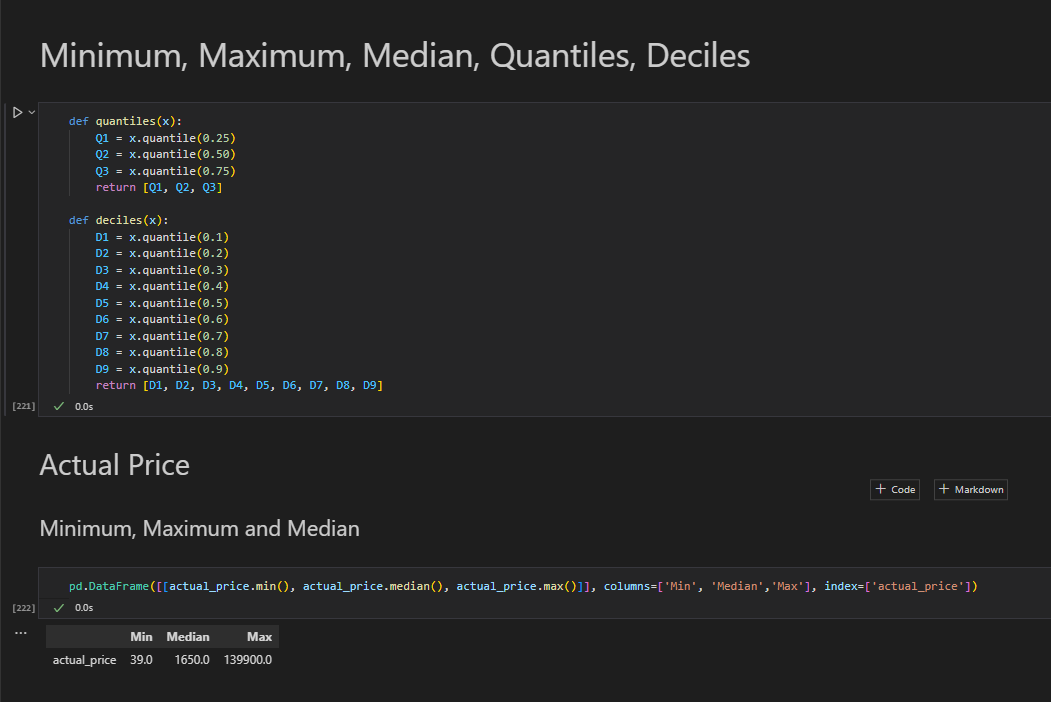
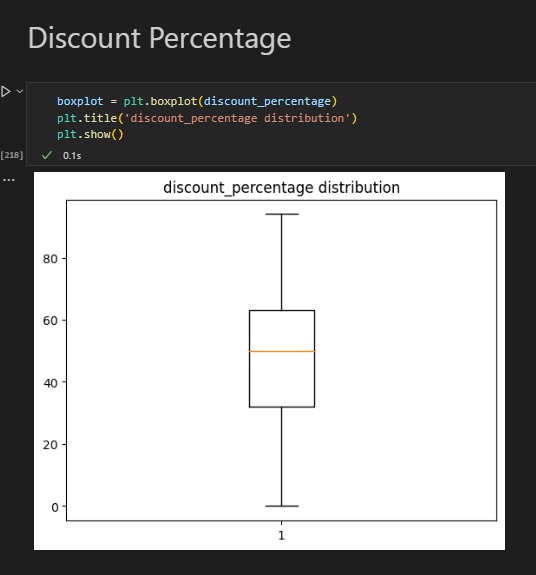
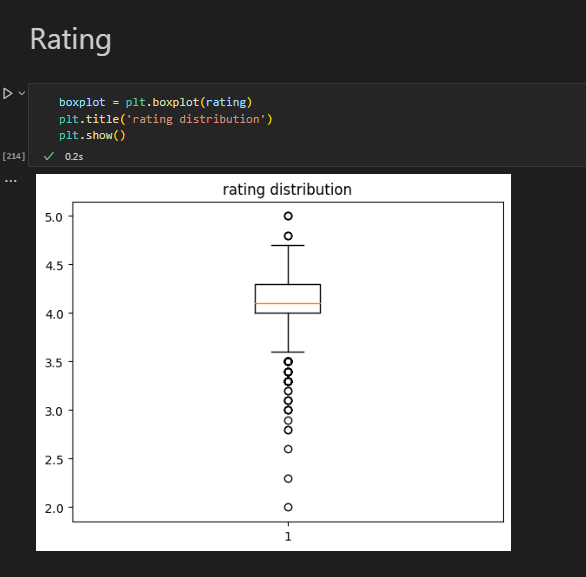
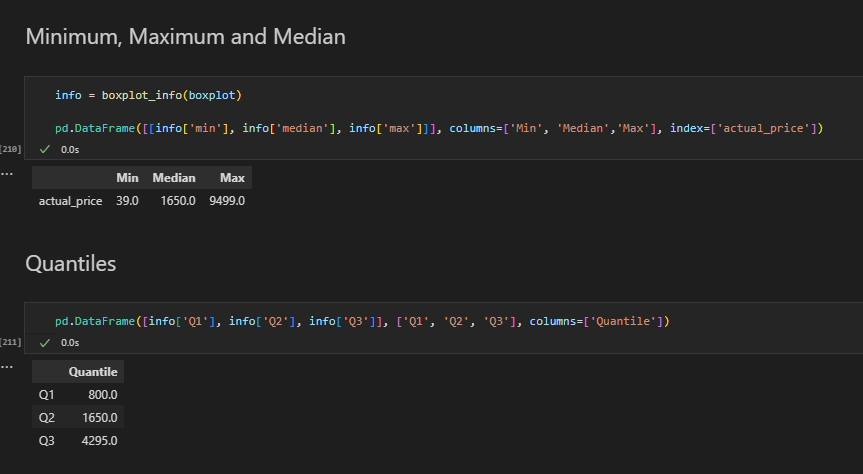
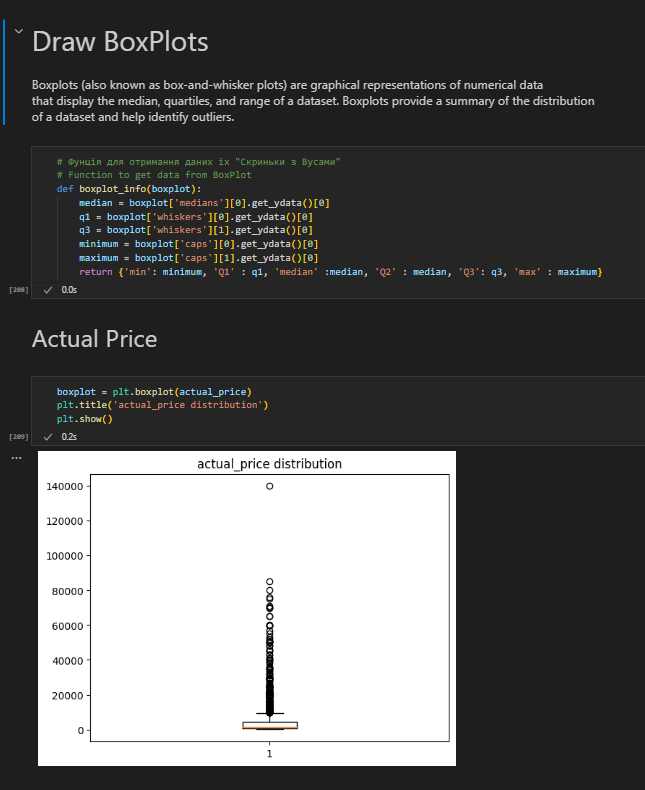
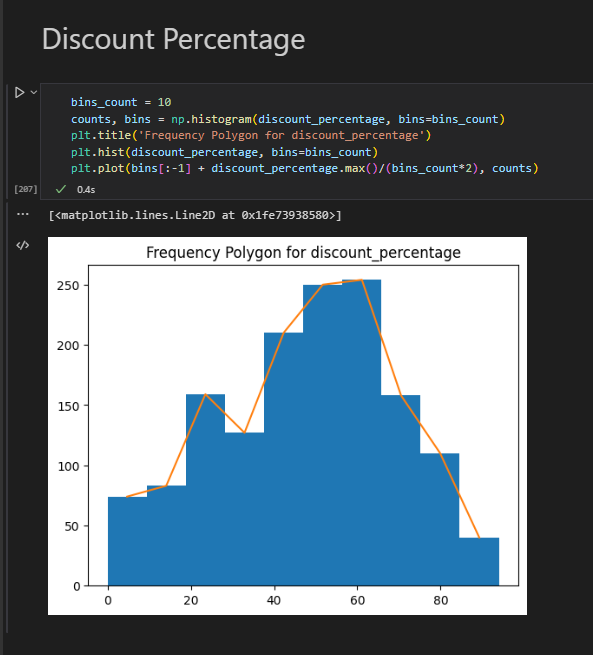
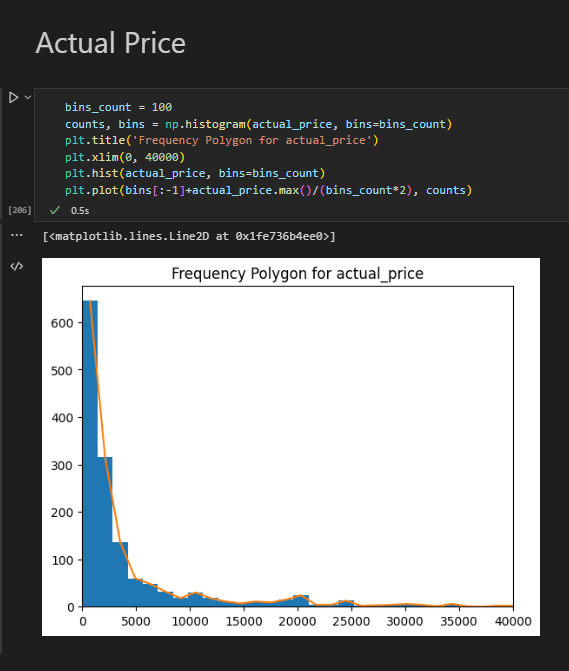
### 6. ДОДАТОК. ПРОГРАМНА РЕАЛІЗАЦІЯ

Вставляється скріншот **усієї** програмної реалізації виконаної ЛР.









### 7. СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. User Guide — pandas 1.5.3 documentation [<https://pandas.pydata.org/docs/user_guide/index.html>]
2. NumPy User Guide [<https://numpy.org/doc/1.23/numpy-user.pdf>]
3. Pyplot tutorial — Matplotlib 3.7.0 documentation [<https://matplotlib.org/stable/tutorials/introductory/pyplot.html>]
4. Statistics (scipy.stats) — SciPy v1.10.1 Manual [<https://docs.scipy.org/doc/scipy/tutorial/stats.html>]
5. Отримання інформації з “Скриньки з вусами” у Python [<https://stackoverflow.com/questions/23349626/getting-data-of-a-box-plot-matplotlib>]
6. Semi Interquartile Range / Quartile Deviation [<https://www.statisticshowto.com/semi-interquartile-range-quartile-deviation/>]
7. QQ plot distribution check [<https://stackoverflow.com/questions/59692757/qq-plot-distribution-check>]
8. Kolmogorov–Smirnov test [<https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test>]