Факультет Біомедичної інженерії

Національного технічного університету України

«Київський політехнічний інститут»

Кафедра біомедичної кібернетики

Дисципліна

«Інтелектуальний аналіз даних»

***Лабораторна робота № 1***

**Тема:** Дослідження застосування різних форм критерію роздільності для визначення оптимальної кількості кластерів в задачах аналізу функціонального стану серцево-судинної системи людини

Виконав:

студент групи БС-81

Сєров О. В.

Перевірив:

викладач

доц. каф. БМК

Павлов В.А.

дата \_\_\_\_\_\_\_\_\_\_

підпис \_\_\_\_\_\_\_\_\_

Київ – 2020

**Мета роботи:** ознайомитись зі знаннями кластеризації методом к-середніх та визначити оптимальну кількість кластерів за допомогою критерія роздільності.

**Завдання до роботи**

***Частина 1***. Застосування методу к-середніх для одержання різних варіантів кількості кластерів даних функціонального стану ССС людини

***Завдання.***

Використувавши дані функціонального стану серцево-судинної системи людини - вимірювання тиск-пульс модифікованої проби Мартине, задайте в алгоритмі к-середніх різні значення кількості кластерів: одержіть розбиття на 8, 12(10) та 14(13) кластерів.

***Частина 2***. Дослідження застосування різних форм критерію роздільності для визначення оптимальної кількості кластерів функціонального стану серцево-судинної системи людини

***Завдання.*** Розрахувати значення критериїв роздільності відповідно до додатку 1. здійснити дослідження розбиття на кластери. Зробити висновки що до якості розбиття та визначити найкращий варіант кластеризації.

Набір даних, що додається (файл додається) складається з вимірів тиск-пульс у 6-ти режимах: спокій, зразу після навантаження, та 4 виміри послідовно через хвилину відпочинку.

Задачу виконати на двох наборах даних:

1. Веріфувати розрахунок на повному наборі ознак.
2. Індивідуальний варіант розрахунку визначити наступним чином: у множину ознак включаються режим спокою та режим зразу після навантаження. Крім того випадковим чином долучаються дані 2-х чи 3-х режимів (з чотирьох, що залишилися).

Результати розрахунку представити у вигляді таблиці значень критеріїв для кожного варіанту розбиття на кластери.

**Теоретичні відомості**

**Кластериза́ція ме́тодом k-сере́дніх**  — популярний метод кластеризації, — впорядкування множини об'єктів в порівняно однорідні групи.

Мета методу — розділити *n* спостережень на *k* кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням.

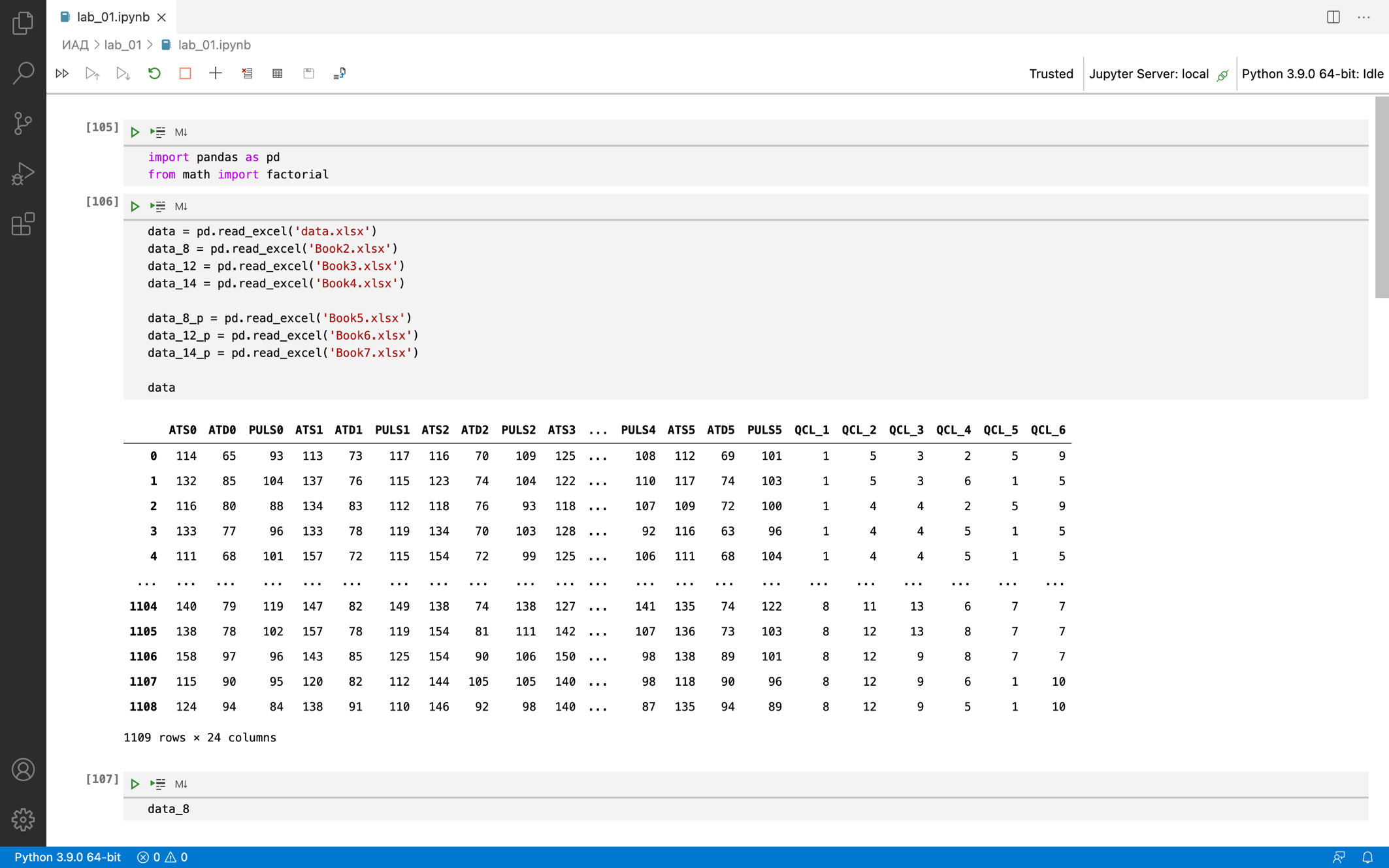
Крітерій роздільності на основі середньоквадратичного відхилення:

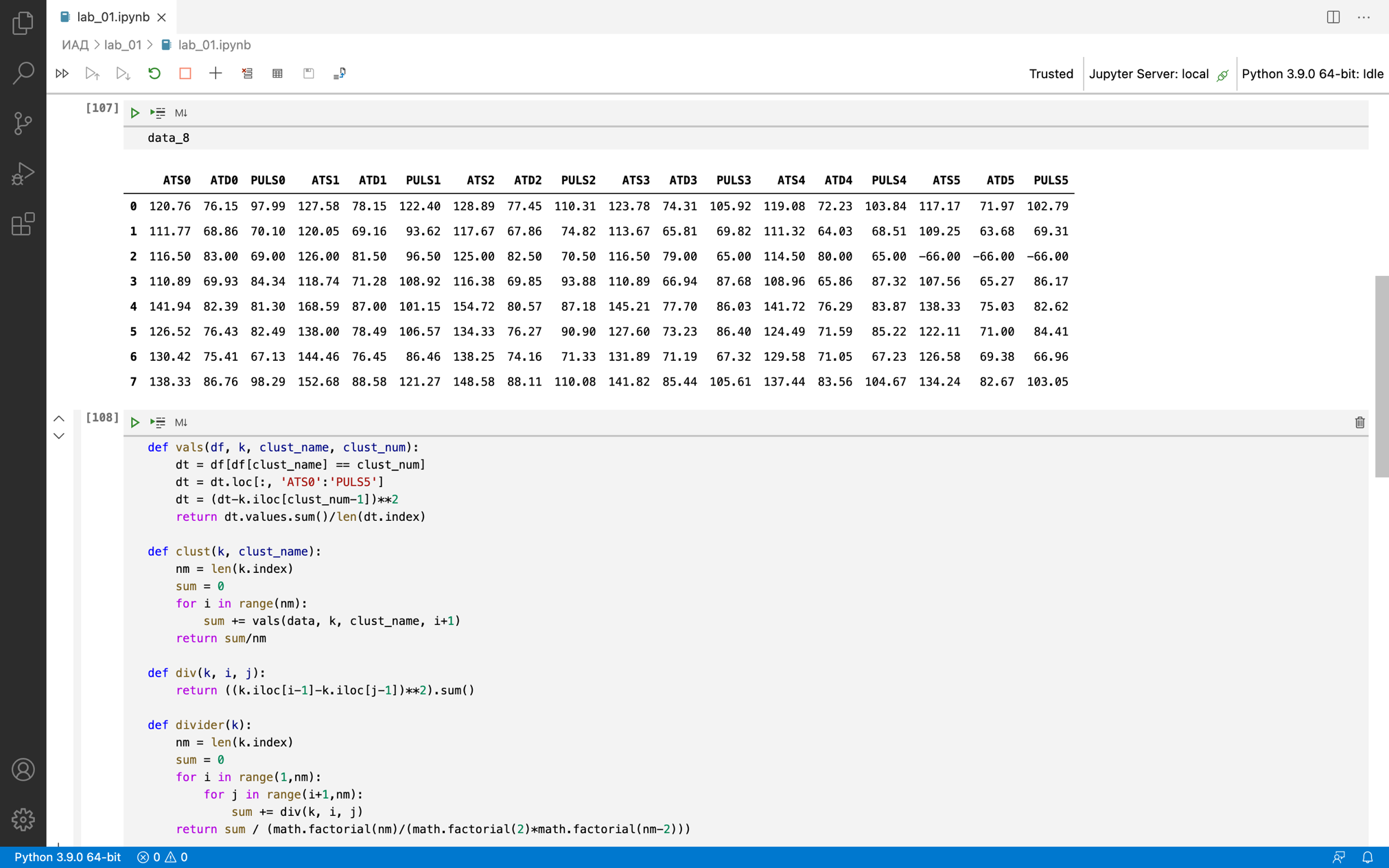
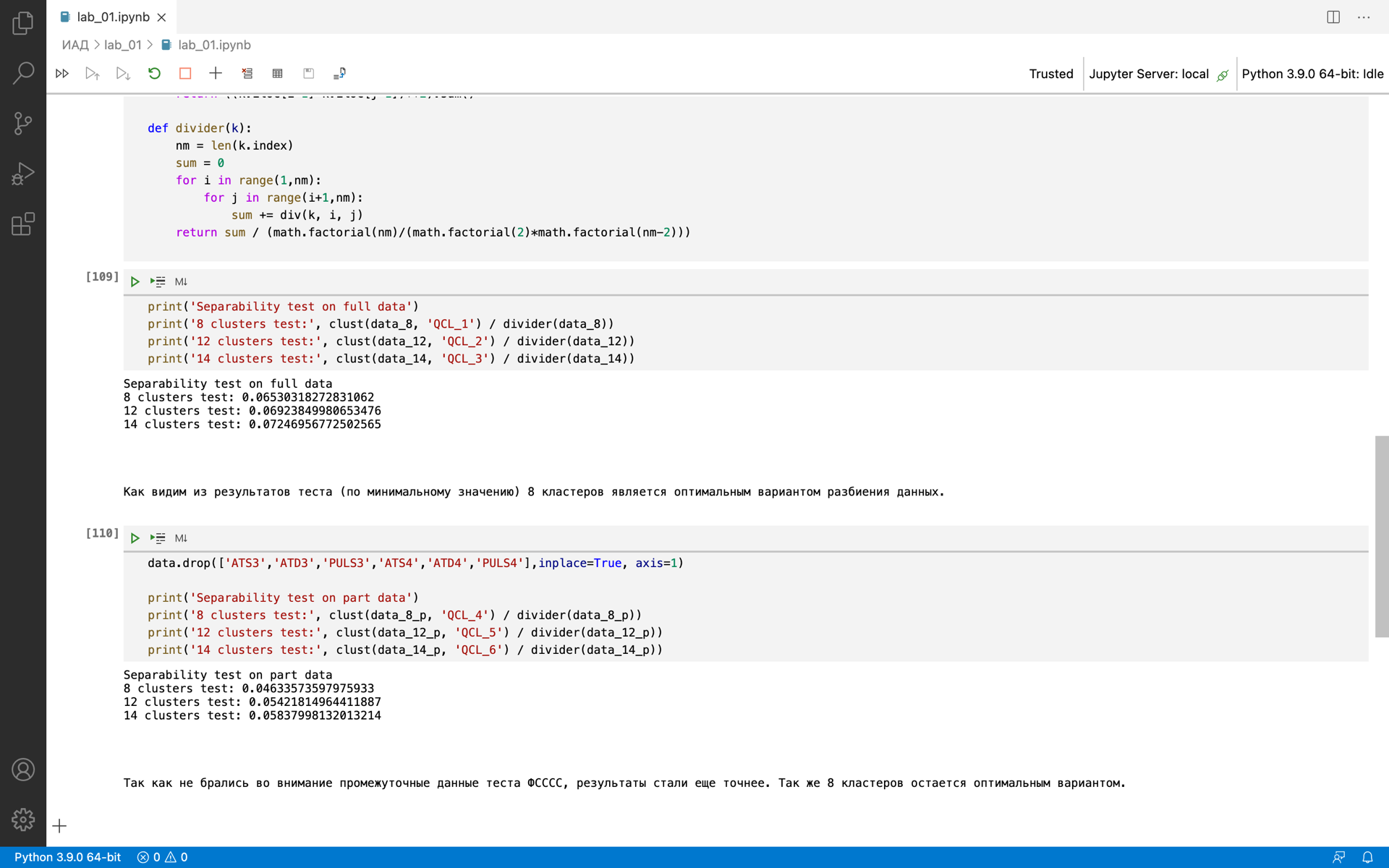


де *xip,* *i*=*1,...,nq* , *p=1,...,18* це *р -****та координата*** *i-****тої точки*** *(*їх всього *nq* одиниць *) в q-***тому кластері** (всього кластерів К штук) а ,,- **середні точки** *q*-того, *i*-того и *j*-того кластерів по р-тій координаті.

Необхідно застосувавши дані по m змінним (в прикладі m=18 ) у К кластерах (в прикладі К=8 К=12 (10) и К=14(13) кластерів ) и по nq q=1,...K точкам в кожному кластері.

**Розрахунки та результати**



**Висновки**

Кластеризація К-середніх виявилась стабільним методом, який попри зміни у даних залишив високий рівень точності кластеризації.

*Переваги*

Головні переваги методу k-середніх — його простота та швидкість виконання. Метод k-середніх більш зручний для кластеризації великої кількості спостережень, ніж метод ієрархічного кластерного аналізу (у якому дендограми стають перевантаженими і втрачають наочність).

*Недоліки*

Одним із недоліків простого методу є порушення умови зв'язності елементів одного кластера, тому розвиваються різні модифікації методу, а також його нечіткі аналоги, у яких на першій стадії алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем приналежності).

Незважаючи на очевидні переваги методу, він має суттєві недоліки:

1. Результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів
2. Алгоритм чутливий до викидів, які можуть викривлювати середнє
3. Кількість кластерів повинна бути заздалегідь визначена дослідником