**Міністерство освіти і науки України**

**Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"**

**Факультет інформатики та обчислювальної техніки**

**Кафедра інформатики та програмної інженерії**

**Звіт**

з лабораторної роботи №5 з дисципліни

«Аналіз даних в інформаційних системах»

„**Регресійні моделі**”

**Виконав(ла)**

(шифр, прізвище, ім'я, по батькові)

*ІП-11 Тарасьонок Дмитро Євгенович*

**Перевірила**

(прізвище, ім'я, по батькові)

*Ліхоузова Т. А.*

Київ 2023

ЗМІСТ

[1 Мета лабораторної роботи 3](#_Toc131630688)

[2 Завдання 4](#_Toc131630689)

[2.1 Основне завдання 4](#_Toc131630690)

[2.2 Додаткове завдання 4](#_Toc131630691)

[3 Виконання основного завдання 5](#_Toc131630692)

[3.1 Дослідити дані, підготувати їх для побудови регресійної моделі 5](#_Toc131630693)

[3.2 Розділити дані на навчальну та тестову вибірки 7](#_Toc131630694)

[3.3 Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну регресію та поліноміальну регресію обраного вами виду 8](#_Toc131630695)

[3.4 Використовуючи тестову вибірку, з'ясувати яка з моделей краща 10](#_Toc131630696)

[4 Виконання додаткового завдання 11](#_Toc131630697)

[4.1 Дослідити дані, сказати чи є мультиколінеарність, побудувати діаграми розсіювання 11](#_Toc131630698)

[4.2 Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду) 14](#_Toc131630699)

[4.3 Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати яка з моделей краща 17](#_Toc131630700)

[5 Висновок 18](#_Toc131630701)

# Мета лабораторної роботи

Мета роботи – ознайомитись з різновидами регресійних моделей.

# Завдання

## Основне завдання

Скачати дані із файлу Data2.csv

1. Дослідити дані, підготувати їх для побудови регресійної моделі
2. Розділити дані на навчальну та тестову вибірки
3. Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну регресію та поліноміальну регресію обраного вами виду
4. Використовуючи тестову вибірку, з'ясувати яка з моделей краща

## Додаткове завдання

Завантажити дані файлу Data4.csv

1. Дослідити дані, сказати чи є мультиколінеарність, побудувати діаграми розсіювання
2. Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду)
3. Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати яка з моделей краща

# Виконання основного завдання

## Дослідити дані, підготувати їх для побудови регресійної моделі

Для початку завантажимо дані, замінимо назви стовпців, щоб надалі до них було простіше звертатися.

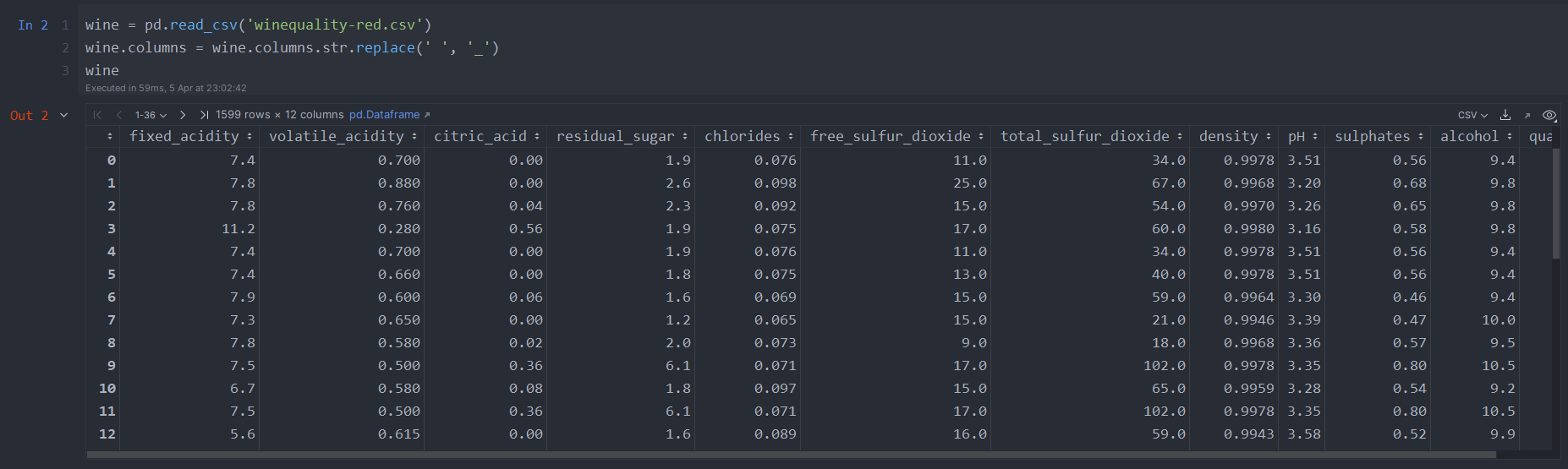


Рис 3.1 – Завантаження даних, перейменовування стовпців

Після цього виведемо теплову карту кореляції між різними стовпцями.

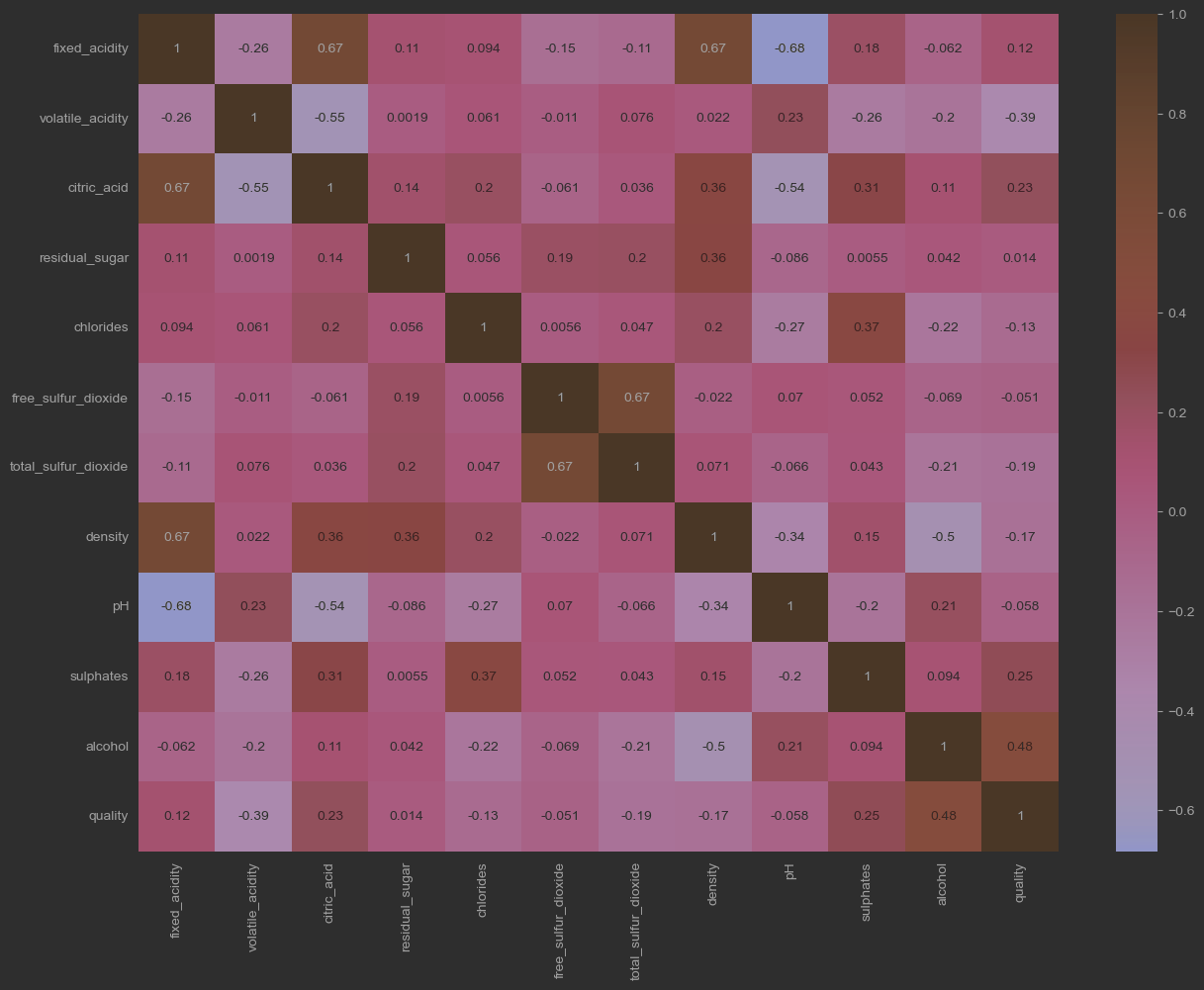


Рис 3.2 – Теплова карта кореляції

Бачимо, що на якість деякі стовпці впливають більше, а деякі – менше. Для збільшення точності регресії відкинемо стовпці pH (коефіцієнт кореляції = -0.058), free\_sulfur\_dioxide (-0.051) та residual\_sugar (0.014).

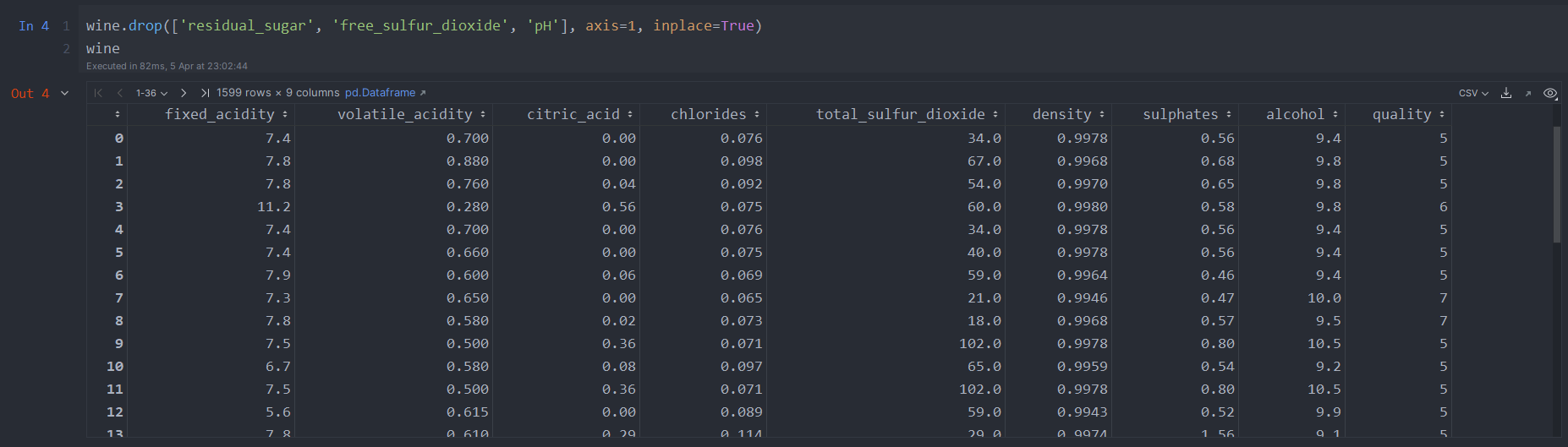


Рис 3.3 – Відкидання малозначущих стовпців

Також можемо вивести матрицю розсіювання.

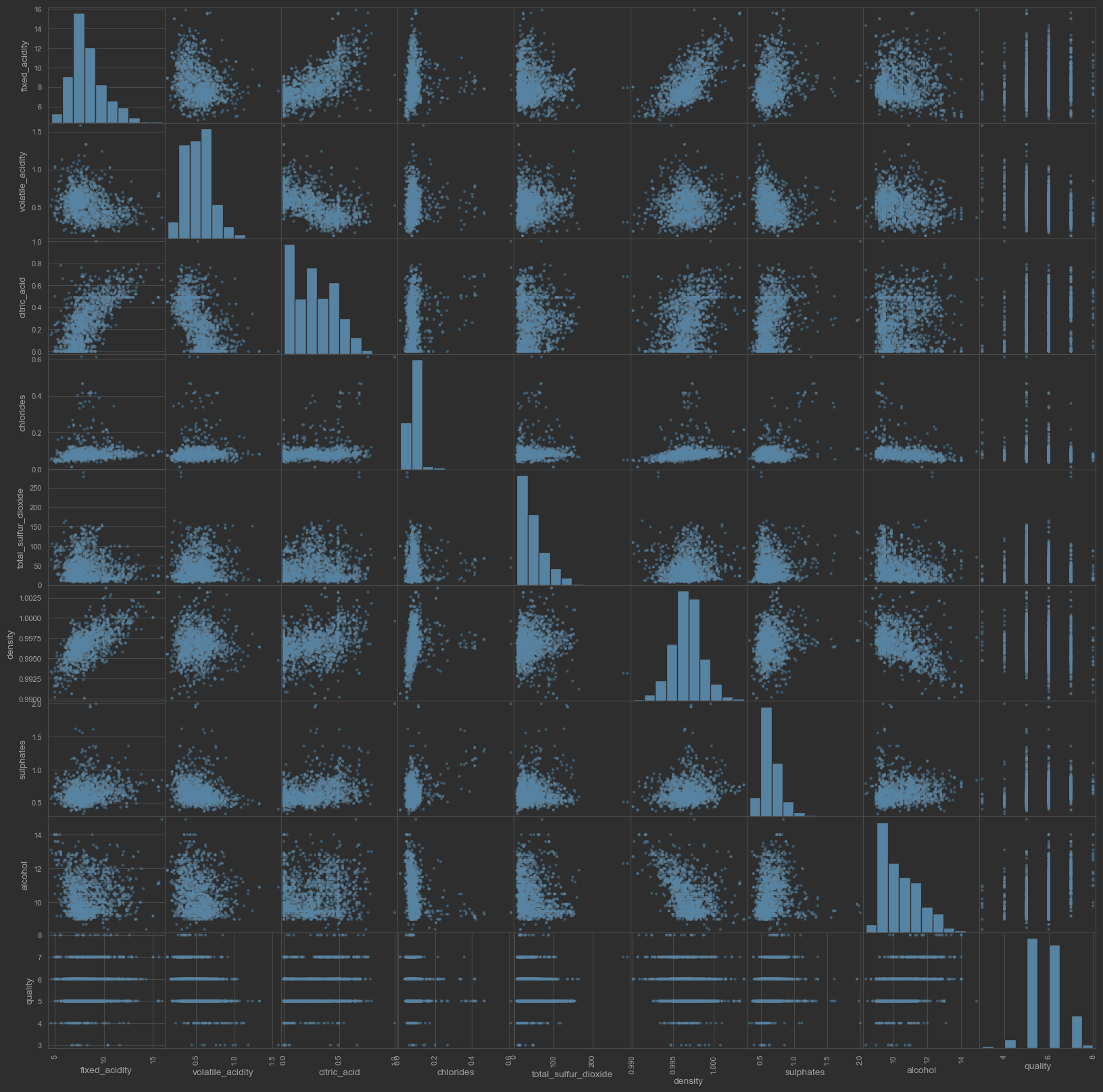


Рис 3.4 – Матриця розсіювання

## Розділити дані на навчальну та тестову вибірки

Для того, щоб розділити дані на навчальну та тестову вибірки, можемо використати функцію .train\_test\_split() пакету scikit-learn. За замовчуванням, ця функція поділяє дані по відношенню 75% та 25%. Виведемо розмір кожної з вибірок.

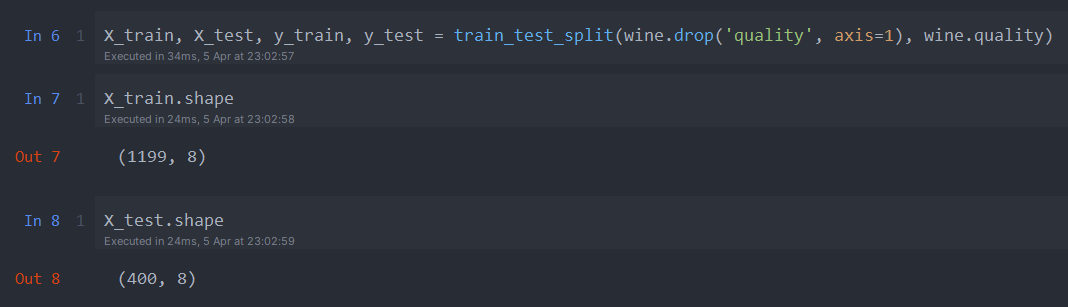


Рис 3.5 – Розділення даних на навчальну та тестову вибірки

## Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну регресію та поліноміальну регресію обраного вами виду

Для початку спробуємо побудувати звичайну лінійну регресію, навчимо її та виведемо діаграму, що відобразити результати.

Як бачимо, регресія не дала високого результату: для всіх вин, що мають якість 7, було спрогнозовано якість від 5.5 до 6.5. Обрахуємо точність моделі за допомогою середньоквадратичної помилки.

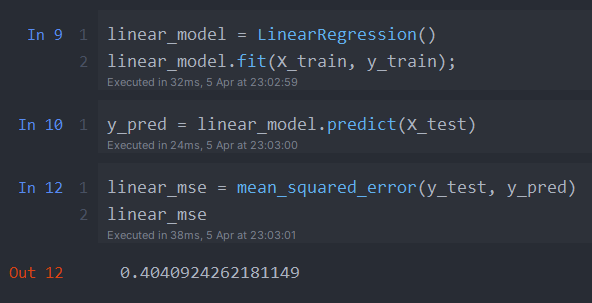


Рис 3.6 – Лінійна регресія

Отримали значення 0.4.

Далі спробуємо використати поліноміальну регресію 1, 2, 3, 4 порядків. Для кожної моделі обрахуємо середньоквадратичну помилку.

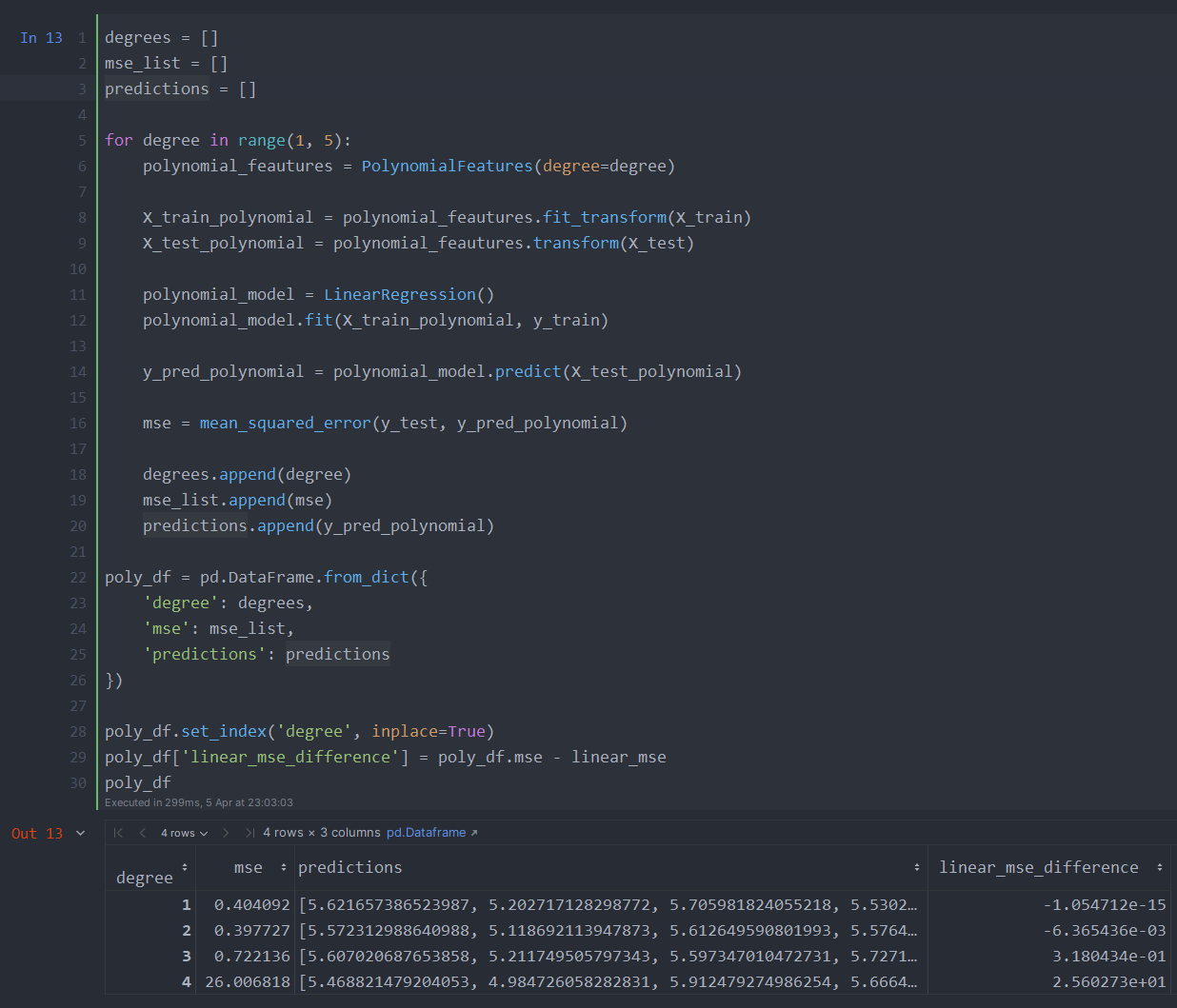


Рис 3.7 – Поліноміальна регресія

Як бачимо, найменша помилка була в поліноміальній регресії першого порядку, яка являє собою просто лінійну регресію. Можемо відобразити отримані дані на діаграмі.

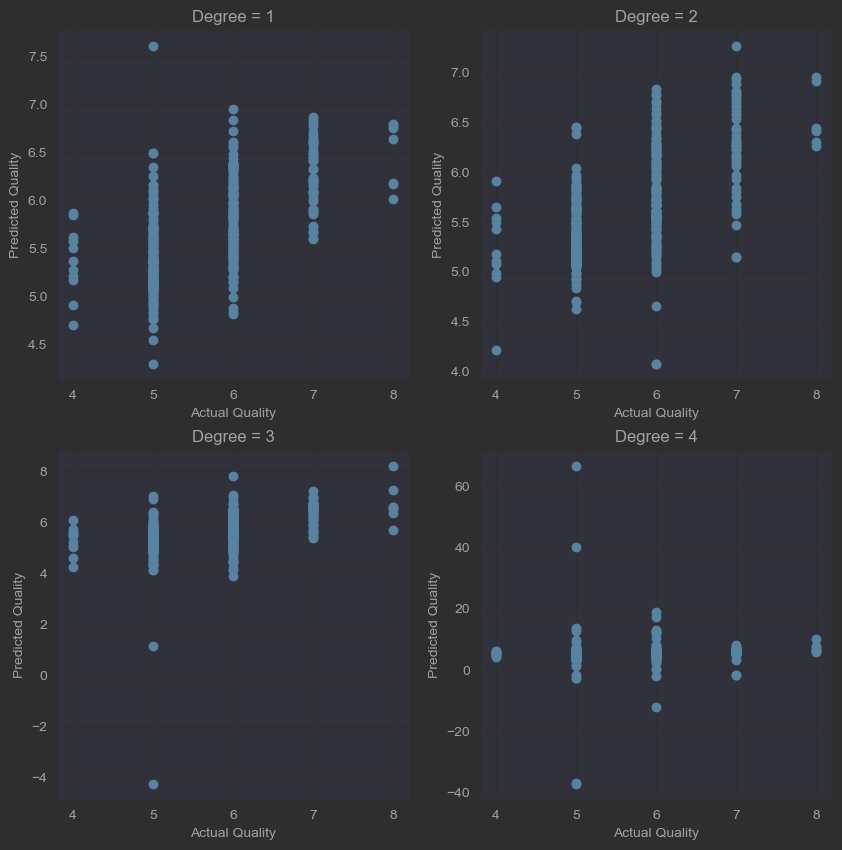


Рис 3.8 – Результати роботи поліноміальної регресії

## Використовуючи тестову вибірку, з'ясувати яка з моделей краща

Як було зазначено в попередньому розділі, найвища точність моделі спостерігалася при використанні лінійної регресії.

# Виконання додаткового завдання

## Дослідити дані, сказати чи є мультиколінеарність, побудувати діаграми розсіювання

Завантаживши дані, було помічено стандартну проблему: всі дійсні числа записані через кому. Виправимо це.

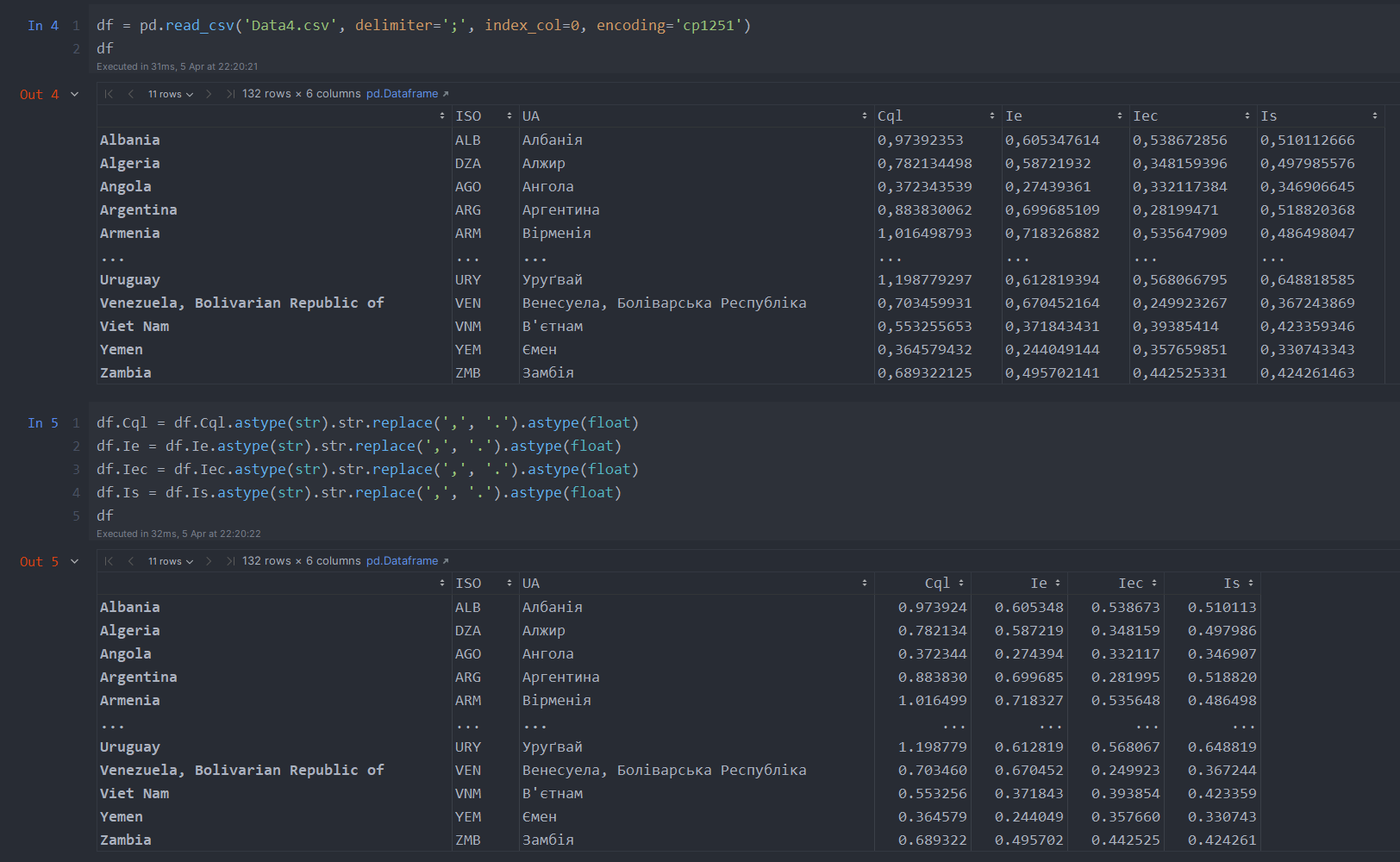


Рис 4.1 – Завантаження навчальних даних, виправлення помилок

Такі ж дії проведемо й для тестового файлу.

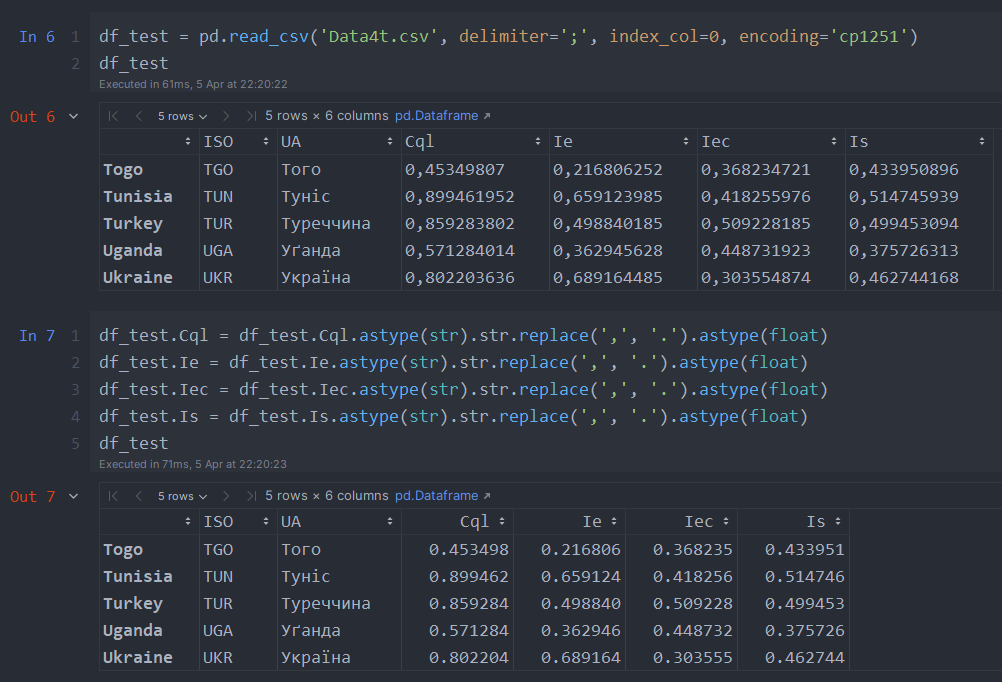


Рис 4.2 – Завантаження тестових даних, виправлення помилок

Для дослідження мультиколінеарності можемо вивести теплову карту з коефіцієнтами кореляції.

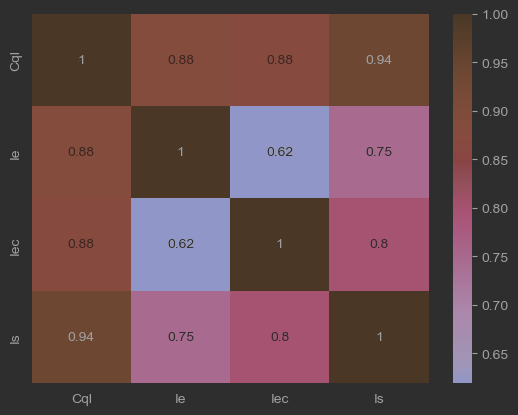


Рис 4.3 – Теплова карта кореляції

Як бачимо, у даному наборі даних є мультиколінеарність: стовпець Cql впливає на стовпці Ie та Iec – коефіцієнти кореляції аж 0.88. Виведемо діаграми розсіювання

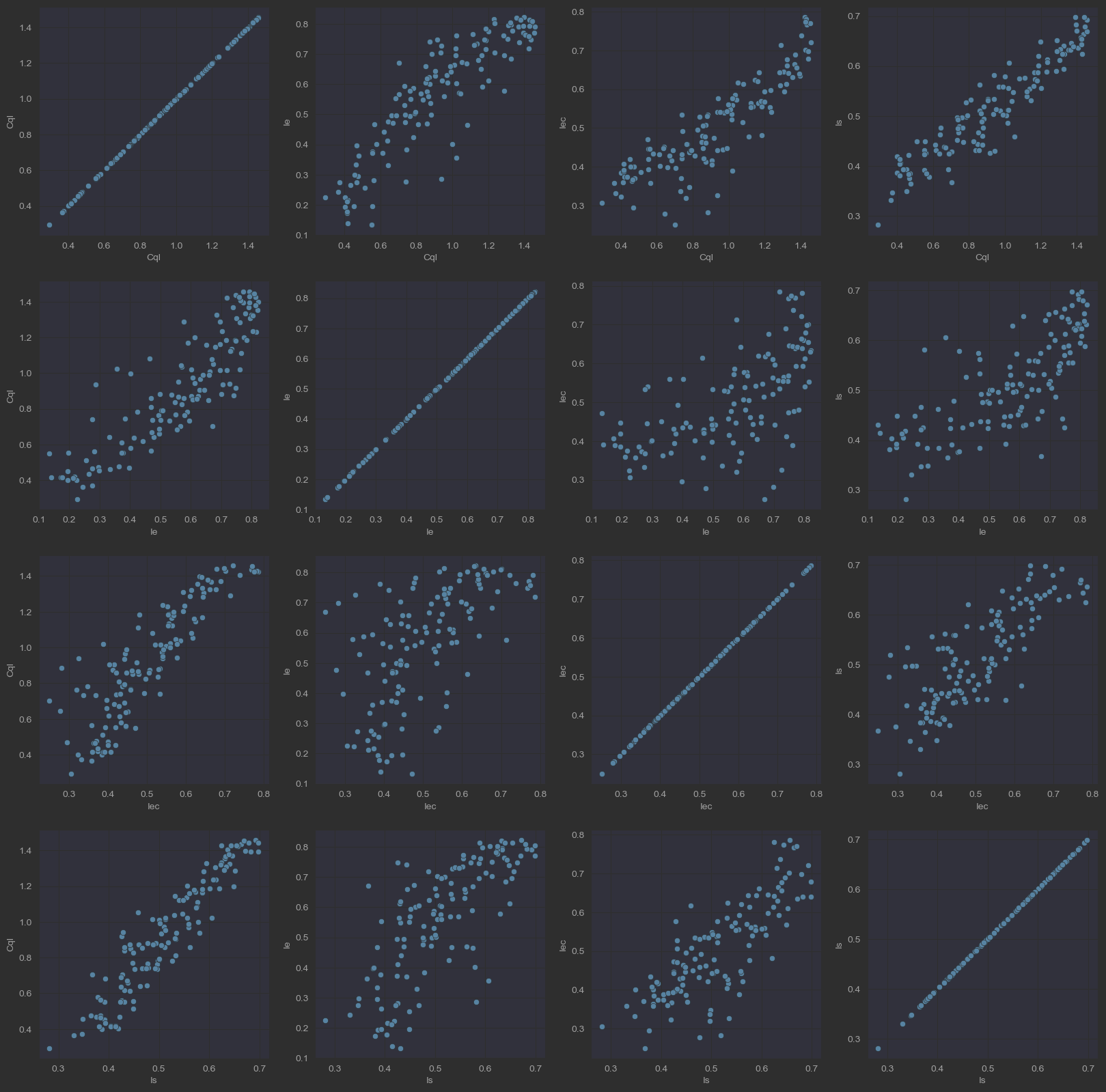


Рис 4.4 – Діаграми розсіювання

## Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду)

Я вирішив не будувати лінійну регресію окремо, а просто побудувати поліноміальну регресію порядку від 1 до 13, оскільки поліноміальна регресія першого порядку, як уже зазначалося, і є лінійною регресією. Для кожного порядку було записано точність, обчислену за допомогою методу .score() регресійної моделі, що використовує коефіцієнт детермінації (R2).



Рис 4.5 – Поліноміальна регресія

Також для наочності можемо відобразити отримані дані на графіку.



Рис 4.6 – Залежність точності регресійної моделі від порядку

Бачимо, що найвищу точність було отримано при використанні поліноміальної регресії шостого порядку – 99.73%. Можемо також побудувати діаграму розсіювання для отриманих даних.

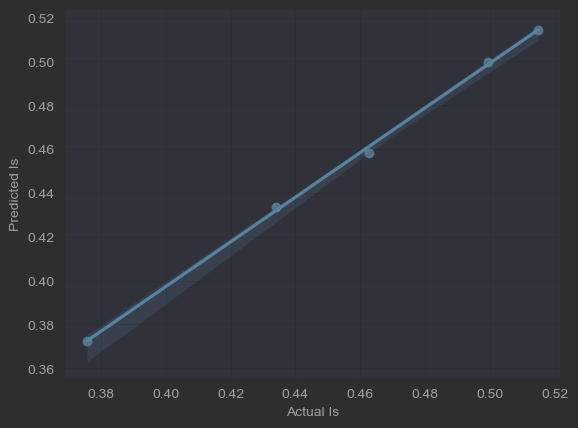


Рис 4.7 – Передбачення поліноміальної регресійної моделі шостого порядку

## Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати яка з моделей краща

Використовуючи тестову вибірку, було встановлено, що найвищу точність для заданого набору даних має поліноміальна регресія шостого порядку – її оцінка становить 99.73%.

# Висновок

У ході даної лабораторної роботи було досліджено різні види регресії: лінійна та поліноміальна. Було досліджено набір даних якості вина, побудовано різні регресійні моделі для визначення якості вина за іншими показниками. Для тестування моделей набір даних було поділено на навчальну та тестову вибірку. При тестуванні було встановлено, що найкращі передбачення робить звичайна лінійна регресія: її середньоквадратична помилка має значення 0.4.

Далі було досліджено деякий набір даних за країнами на мультиколінеарність, встановлено, що вона присутня – коефіцієнт кореляції для двох пар стовпців дорівнює 0.88 – такий зв’язок вважається сильним. Було побудовано поліноміальні регресії порядку від 1 до 13 та порівняно їхню ефективність. Найкраще себе показала лінійна регресія шостого порядку – R2 оцінка для неї становить 99.73%.