МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

**WEB-Аналітика**

ЛАБОРАТОРНА РОБОТА №1

|  |  |  |
| --- | --- | --- |
|  |  | Виконав  студент групи ФІ-42мн  Бєш Радомир Андрійович |

Київ 2025

**Завдання:**

На основі будь-якого access.log сформувати датасет, що надав би інформацію

про користувачів веб-ресурсу, а потім виконати наступні кроки:

a. Визначити кількість користувачів за днями

b. Ранжувати користувачів за User-Agent

c. Ранжувати користувачів за операційними системами

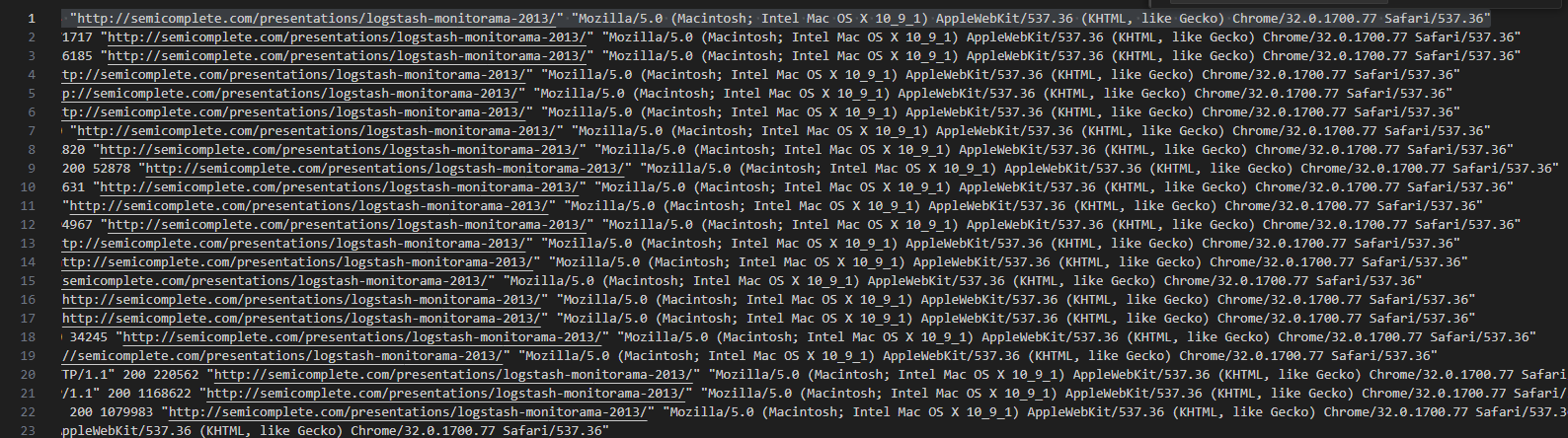
d. Ранжувати користувачів за країною запиту

e. Виокремити пошукових ботів

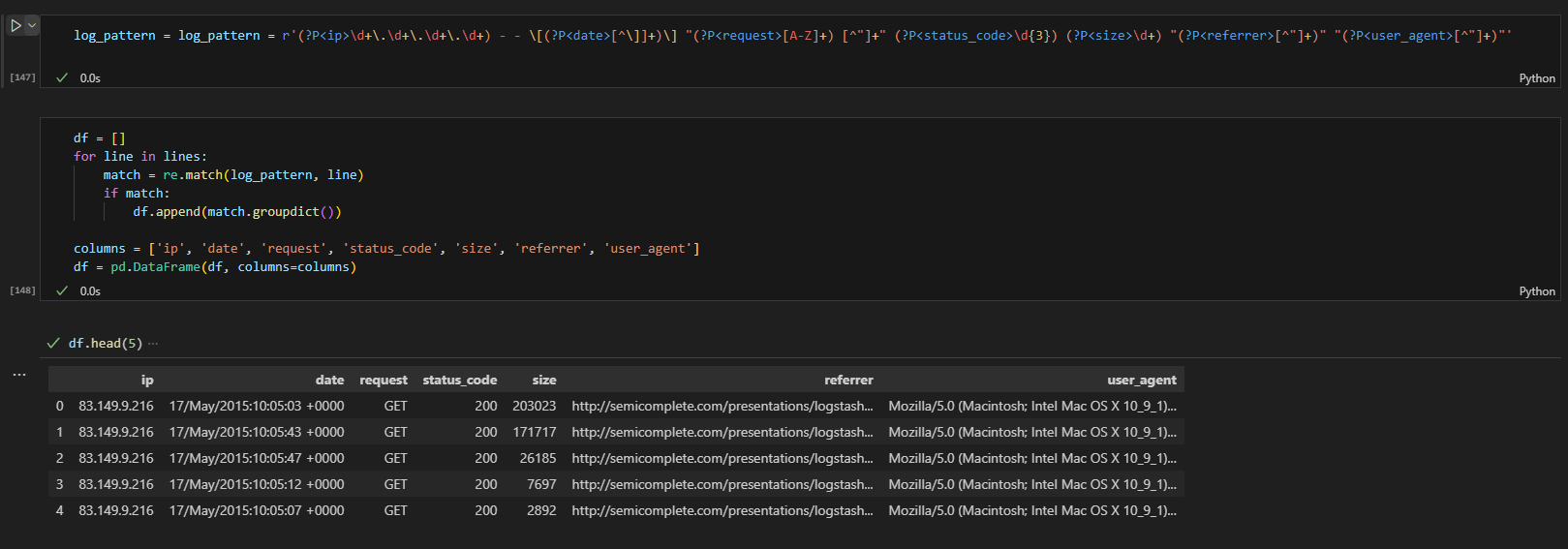
f. Детектувати аномалії (якщо такі є)

**Хід роботи**

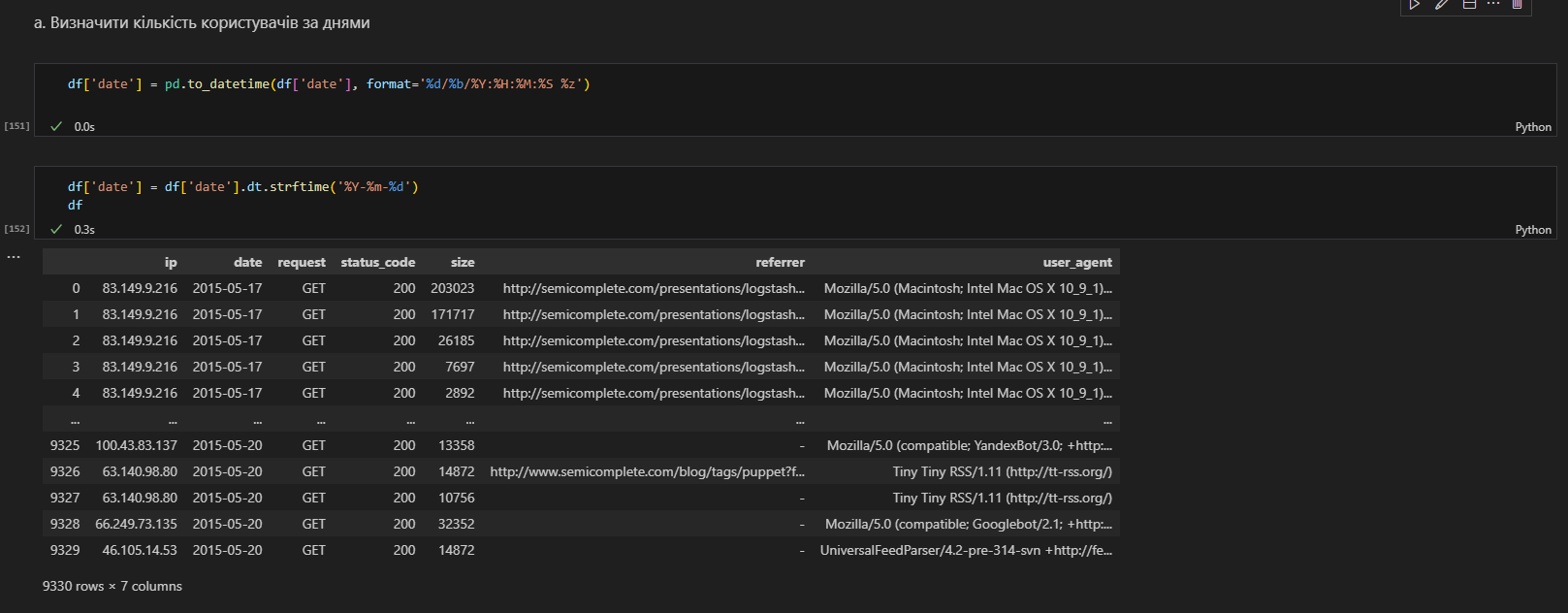
Було обрано датасет apache\_logs.txt, знайшов на Github але нажаль я загубив лінк, де я його знайшов.

Датасет виглядає так:  


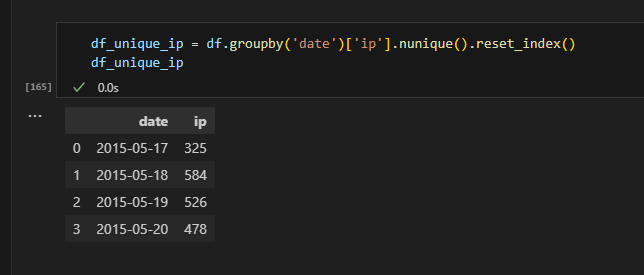
Тому приведемо його в “нормальний вид”.



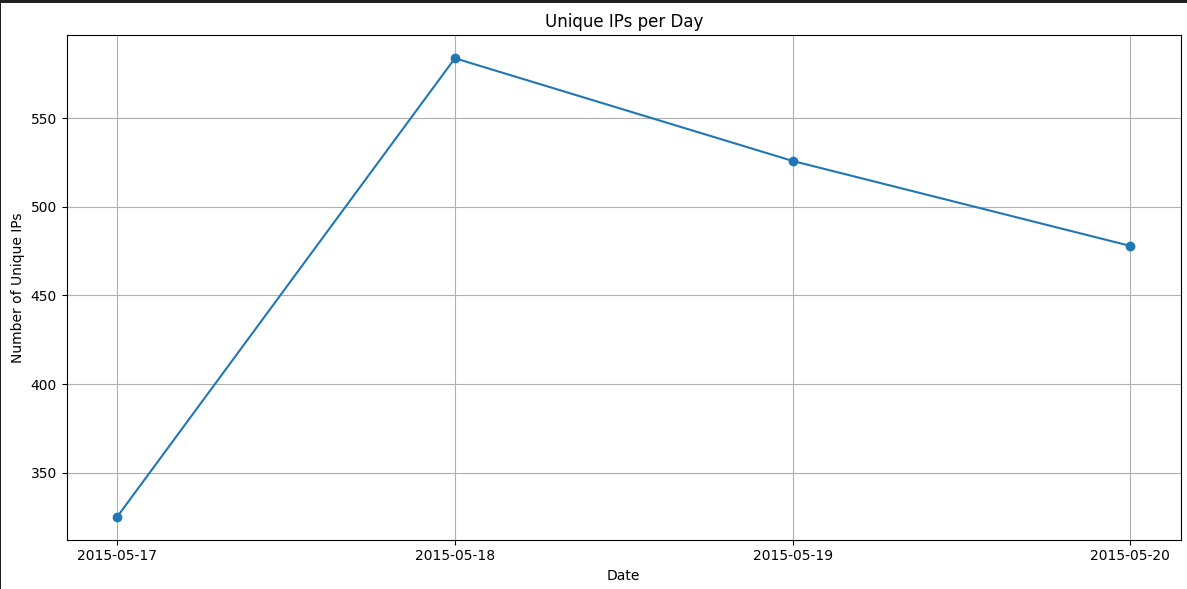
1. Визначити кількість користувачів за днями

Приведемо дату до більш прийнятого формату  


Отже клькість унікальних користувачів становить:

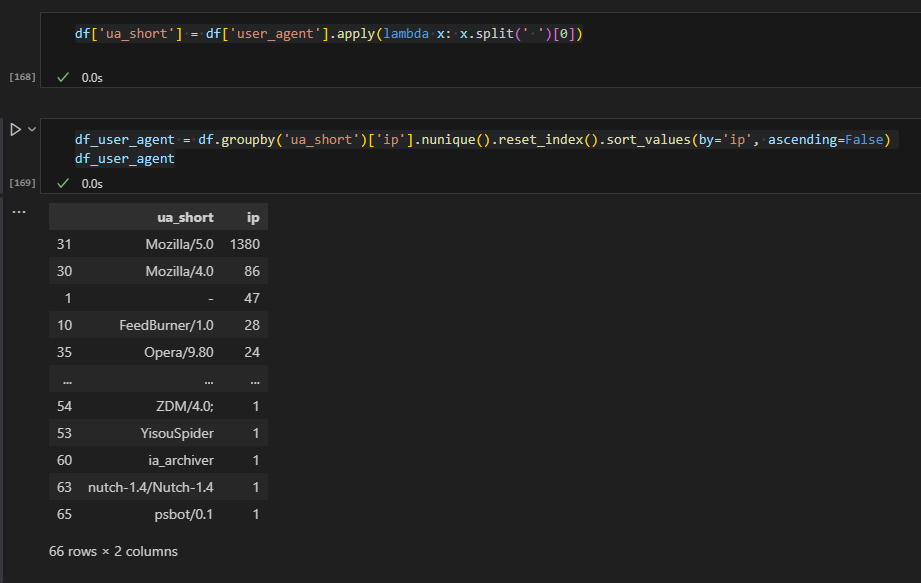


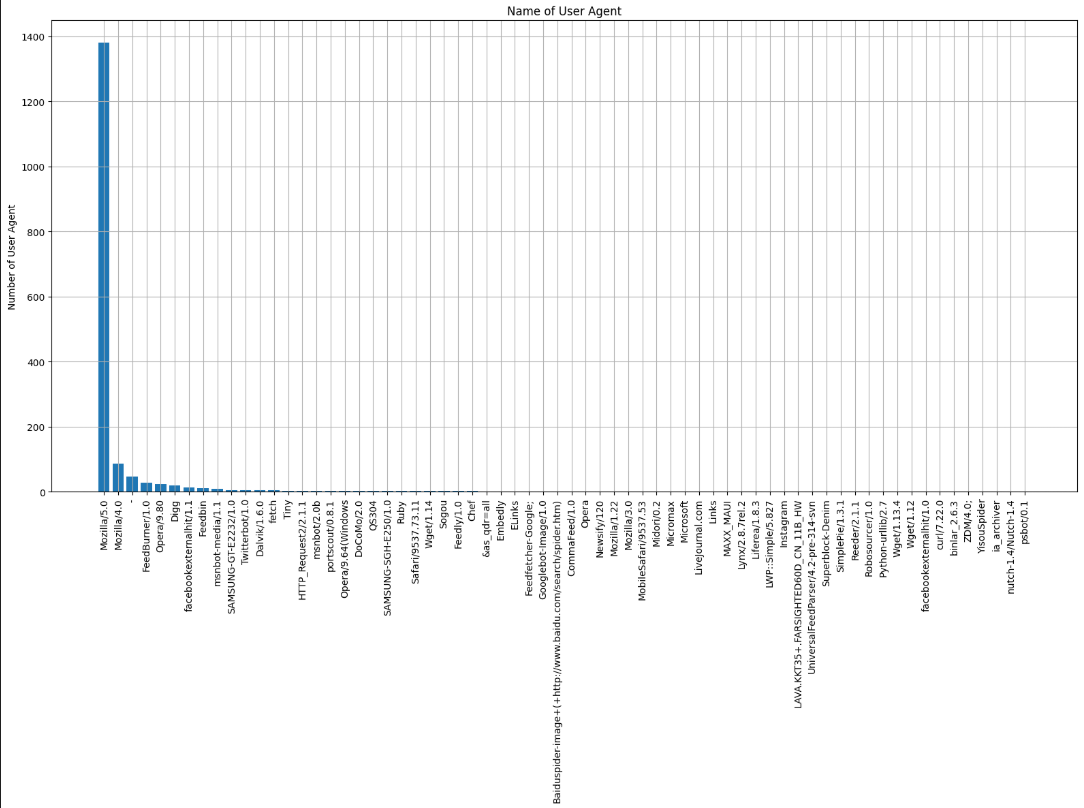
Візуалізуємо:



1. Ранжувати користувачів за User-Agent

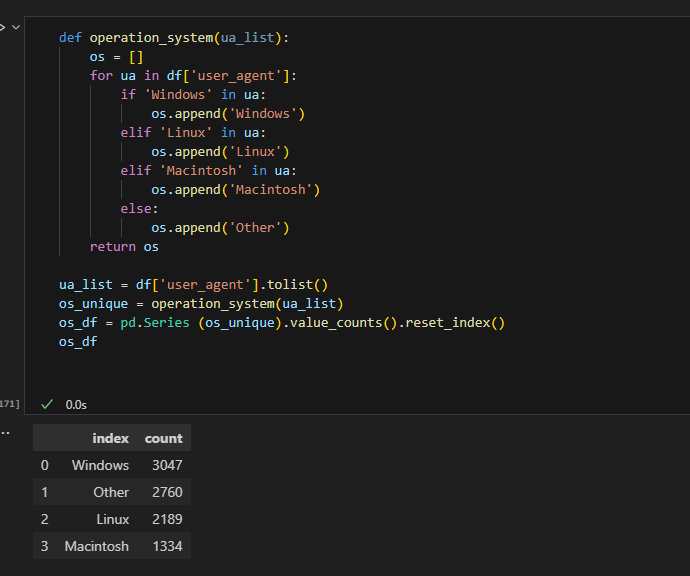
Просто відокремлемо першу частину User-Agent



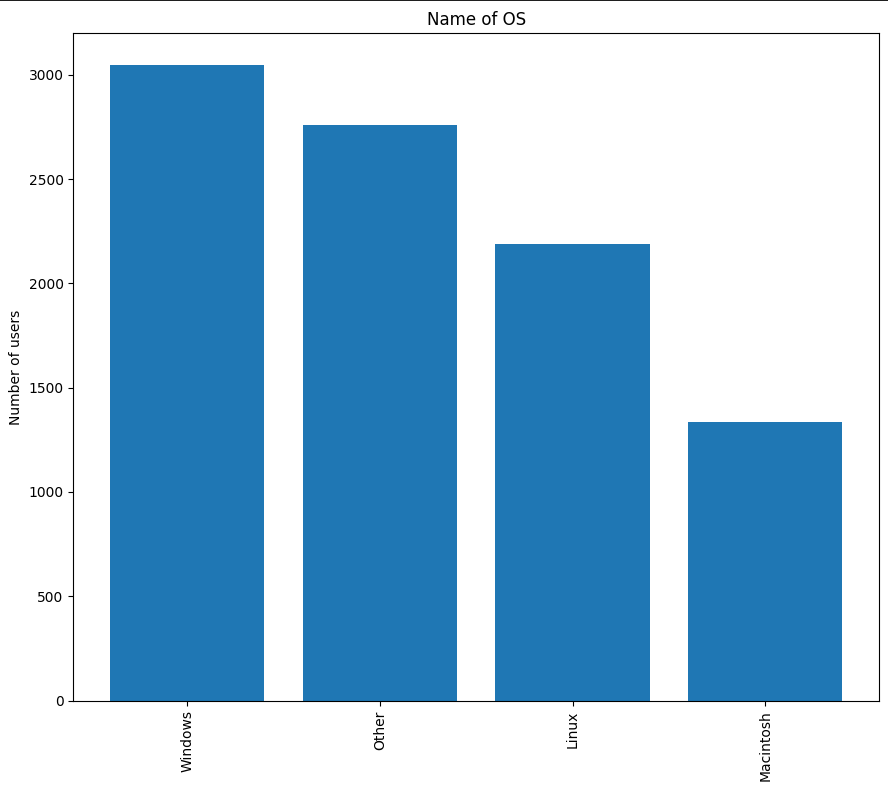
Візуалізація:

1. Ранжувати користувачів за операційними системами

Оберемо найпопулярніші системи:

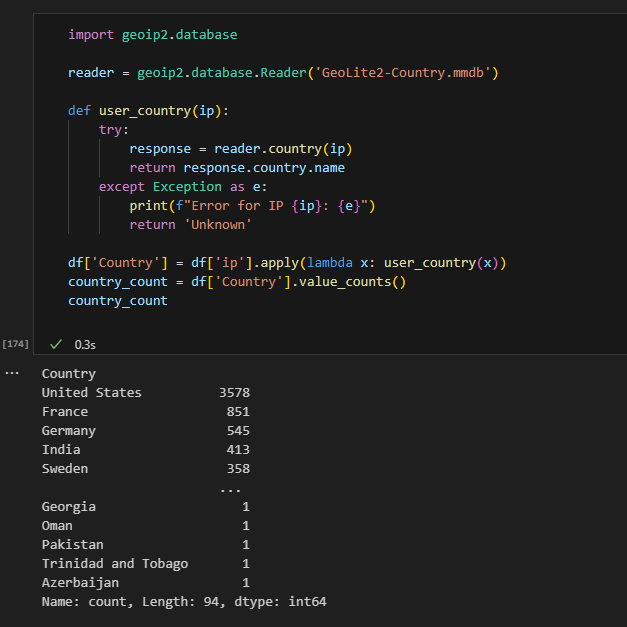


Візуалізація:

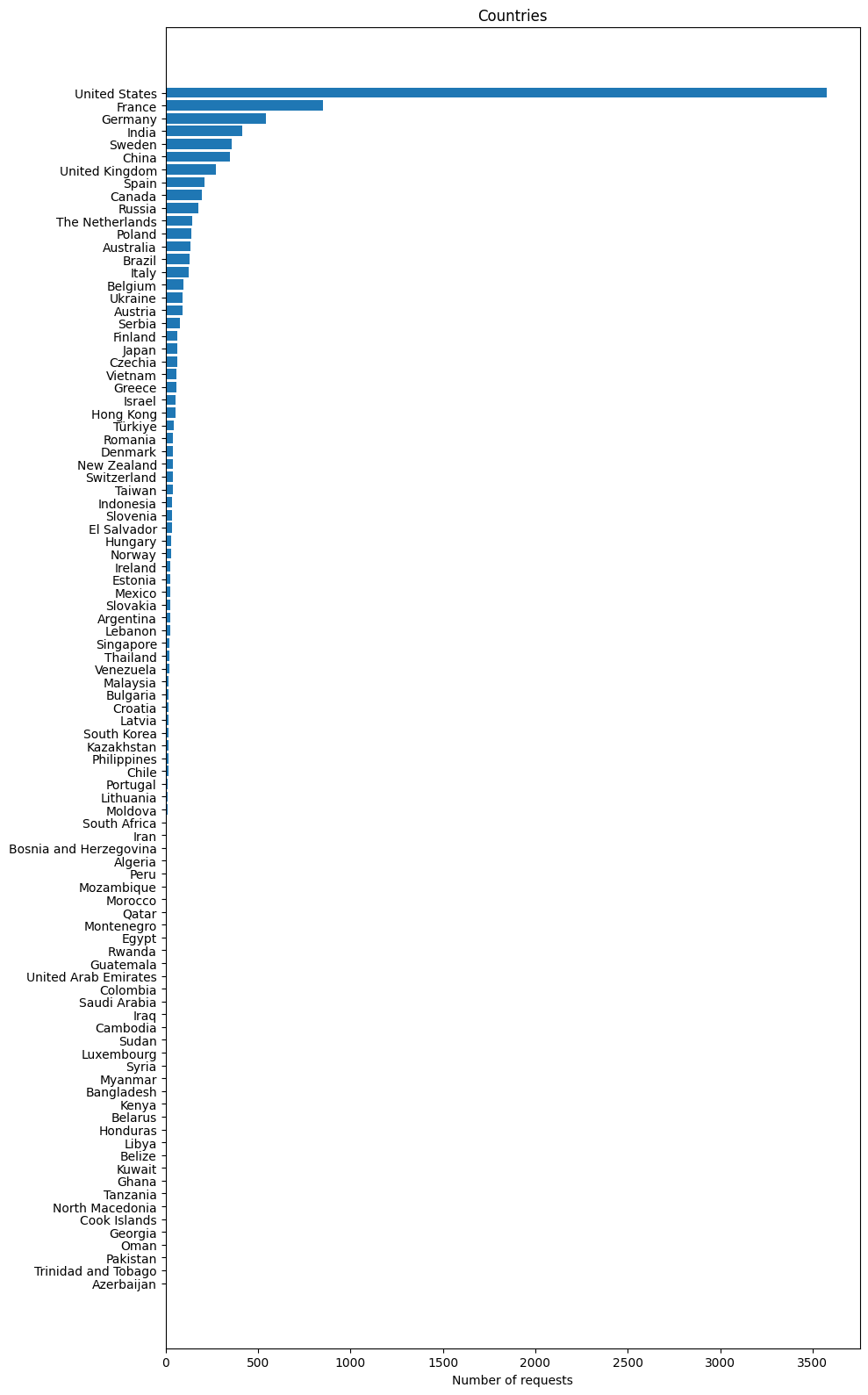


1. Ранжувати користувачів за країною запиту

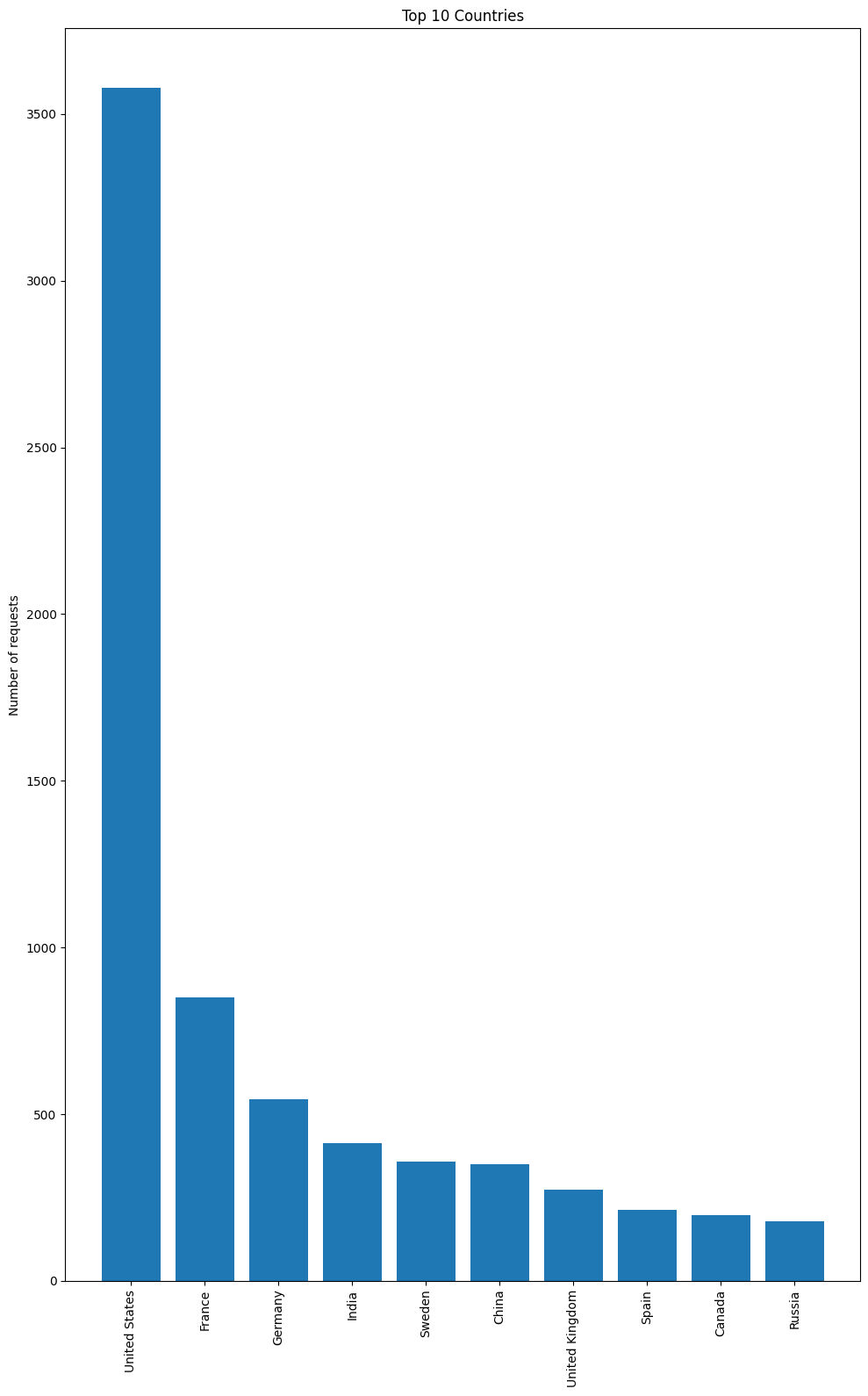
Для цього встановимо модуль geoip2 та завантажемо файл GeoLite2-Country.mmdb.



Візуалізація:

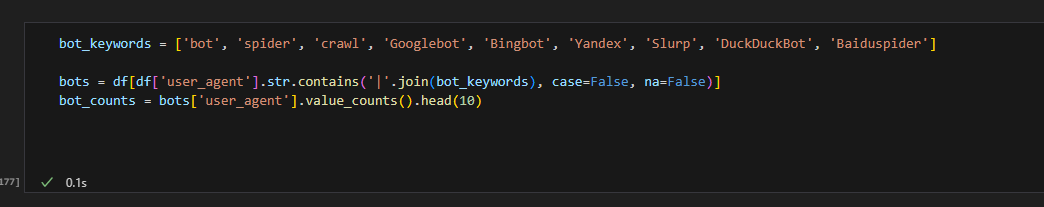


Топ 10 країн:

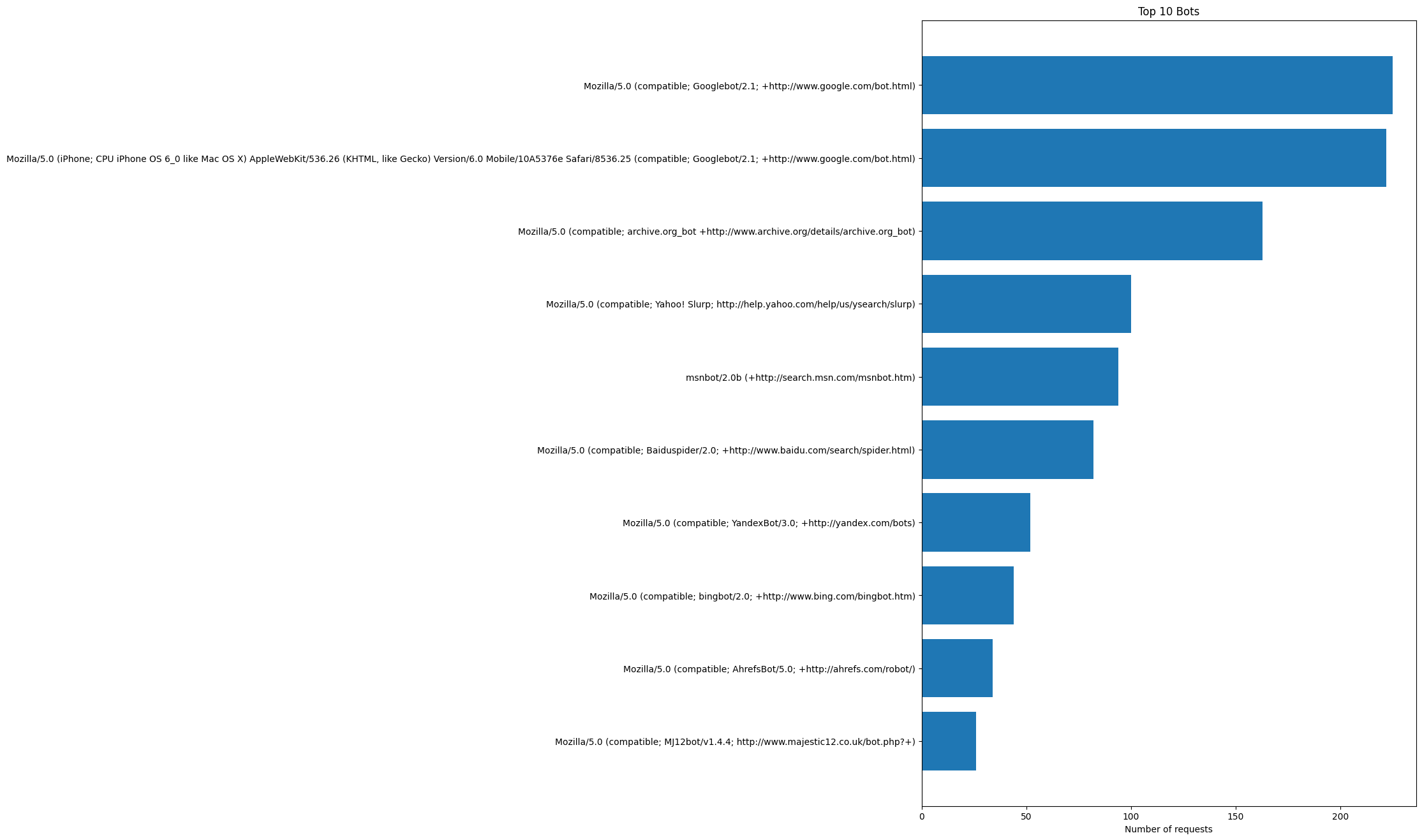


1. Виокремити пошукових ботів

Будемо шукати найпоширеніших ботів.



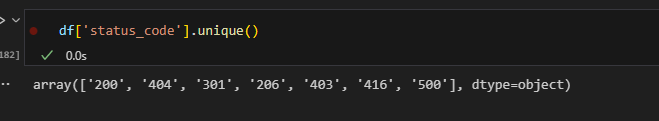
Візуалізація:



1. Детектувати аномалії (якщо такі є)

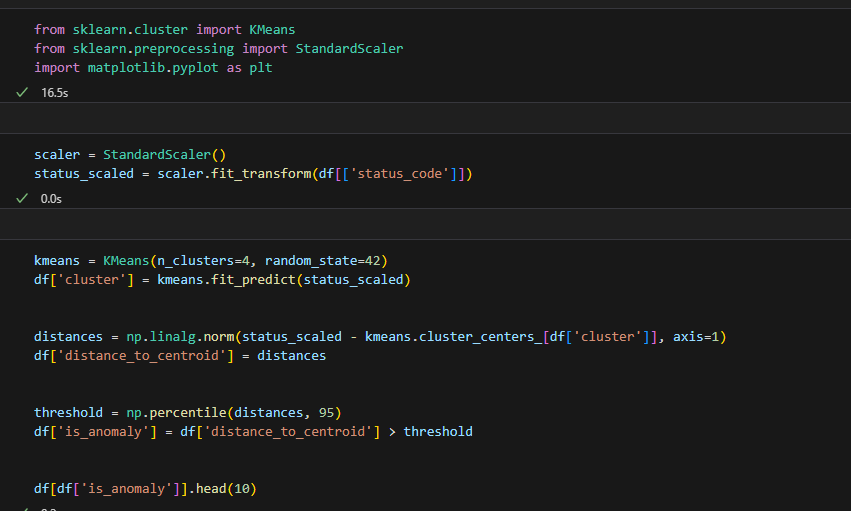
Аномалії будемо шукати за статус кодами і розмірами запитів.

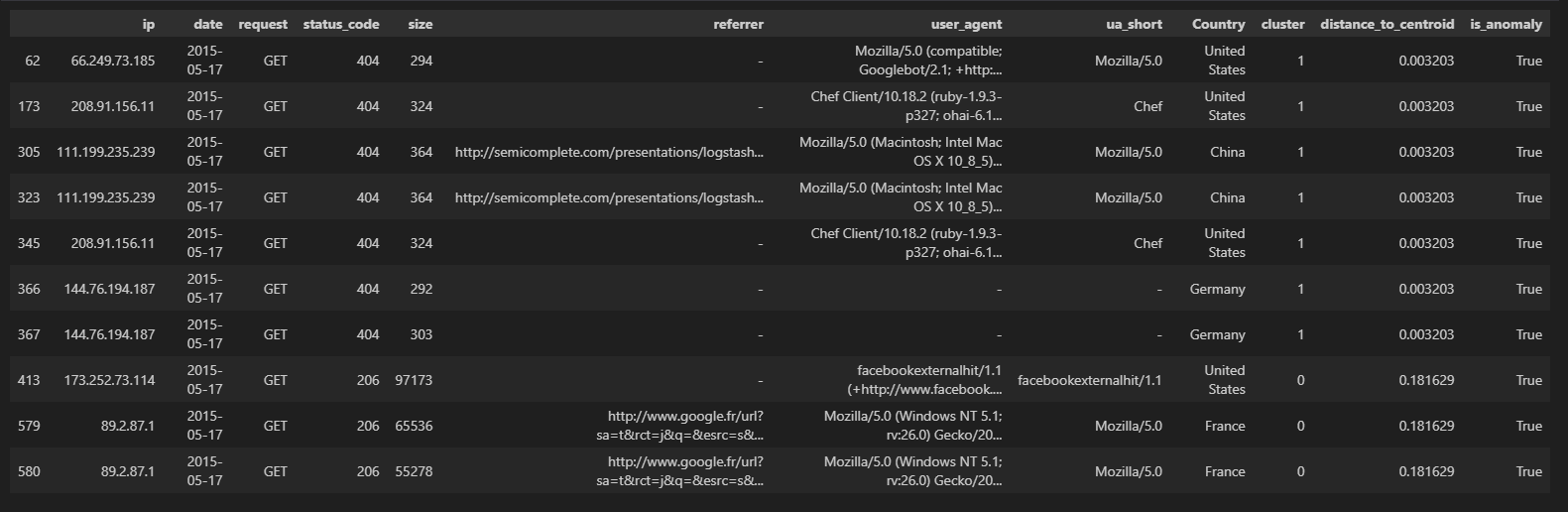
Для цього виведемо, які статус коди маємо:



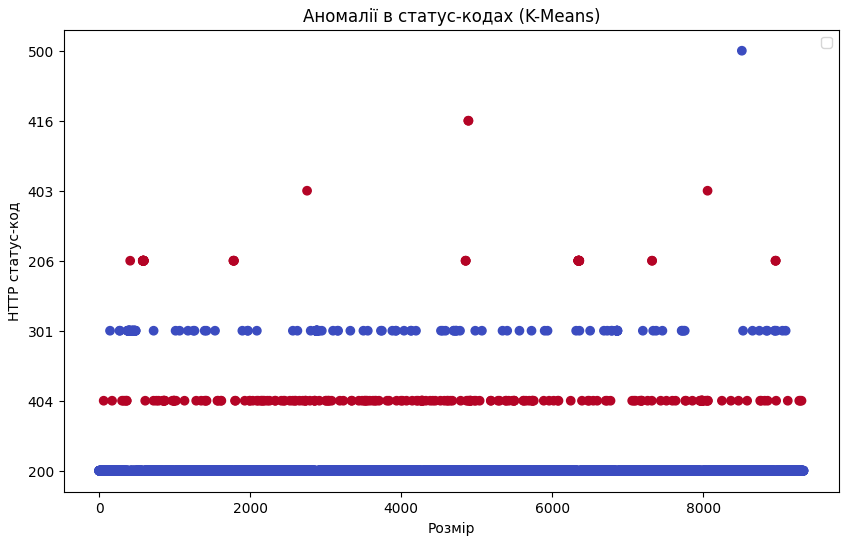
Далі за допомогою методу кластеризації K-Means спробуємо кластеризувати статус-коди з використанням K-Means і виявлення аномалій на основі відстані до центроїд.

Спочатку стандартизуємо ознаку.



На виході маємо такі дані:  


Візуалізація:



Більшість точок на графіку мають статус-код 200 і позначені синім кольором - це означає, що запити були успішними і вважаються нормальними. Червоні точки вказують на аномальні значення, які відрізняються від загальної маси. Серед них зустрічаються такі коди, як 404 (не знайдено), 403 (заборонено), 301 (перенаправлення), 206 (частковий вміст), 416 (некоректний діапазон) і навіть 500 (внутрішня помилка сервера). Наявність цих кодів серед аномалій вказує на потенційні проблеми - наприклад, помилки в роботі серверу, неправильні або підозрілі запити, можливу активність ботів чи сканерів. Також деякі аномалії мають великі розміри відповіді, що може бути нехарактерним для відповідних кодів і свідчити про спроби отримати заборонену або неіснуючу інформацію.

**Висновки:** в ході виконання лабораторної роботи було проаналізовано лог-файл веб-сервера, сформовано датасет із ключовою інформацією про користувачів, здійснено аналіз за User-Agent, операційними системами, країнами запитів та виявлено пошукових ботів. Особливу увагу було приділено виявленню аномалій за допомогою кластеризації методом K-Means. Аналіз показав, що більшість запитів були успішними (код 200), однак також були виявлені аномальні статус-коди, серед яких 404, 403, 301, 206, 416, 500, що може свідчити про наявність технічних збоїв, помилкових запитів або потенційно підозрілу активність. Отримані результати підтверджують доцільність використання методів аналізу логів для покращення безпеки та стабільності веб-ресурсів.