МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

WEB-Аналітика

ЛАБОРАТОРНА РОБОТА №4

Виконав

студент групи ФІ-42мн

Бєш Радомир Андрійович

Київ 2025

Завдання:

На основі даних з 3 лабораторної чи знайти достатньо великий датасет(від 10000 записів) на Кегель (або будь якому іншому) ресурсі використати будь-який метод машинного навчання для дослідження інформації та створення моделі даних.

Хід роботи

1) Обраний датасет:

Посилання: https://www.kaggle.com/datasets/usernam3/shopify-app-store/data

2) Завантаження датасету та первинна обробка

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[3] import kagglehub

    # Download latest version
    path = kagglehub.dataset_download("usernam3/shopify-app-store")

    print("Path to dataset files:", path)
```

```
Downloading from https://www.kaggle.com/api/v1/datasets/download/usernam3/shopify-app-store?dataset_version_number=5...
100%|██████████| 123M/123M [00:06<00:00, 21.3MB/s]Extracting files...

Path to dataset files: /root/.cache/kagglehub/datasets/usernam3/shopify-app-store/versions/5
```

```
[89] import pandas as pd
     import os
     from sklearn.ensemble import IsolationForest
     import pandas as pd
     import matplotlib.pyplot as plt
     from sklearn.cluster import KMeans
     from sklearn.preprocessing import StandardScaler
     import seaborn as sns
     for file in os.listdir(path):
         print(file)
```

```
categories.csv
pricing_plans.csv
apps.csv
key_benefits.csv
reviews.csv
apps_categories.csv
pricing_plan_features.csv
```

Будемо використовувати датасет з додатками та їх відгуками.

```
csv_path = os.path.join(path, 'apps.csv')
df = pd.read_csv(csv_path)
df.info()
```

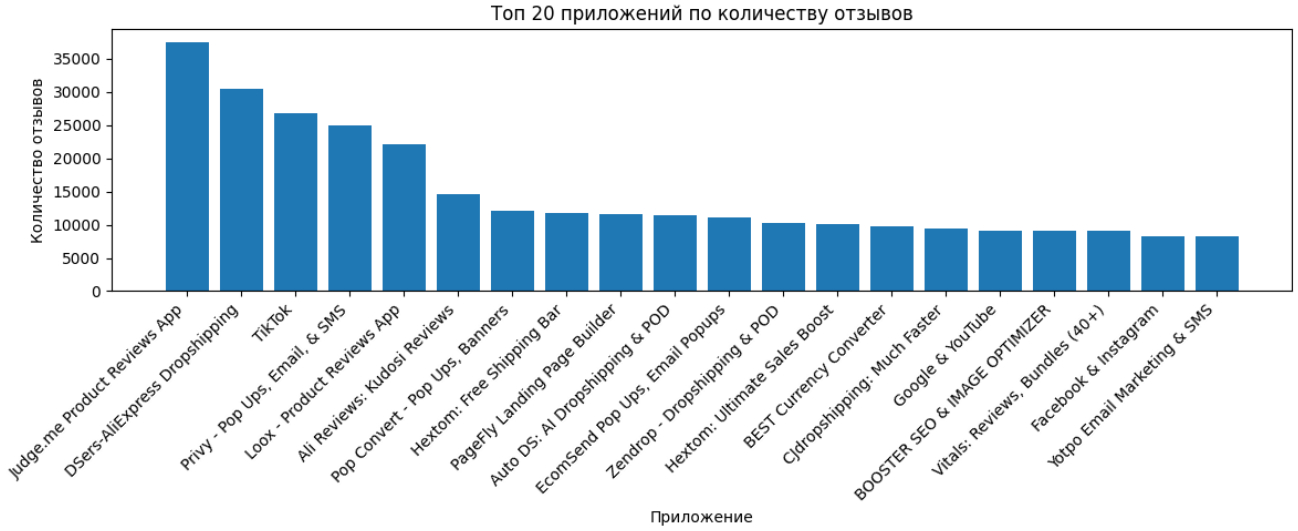
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11951 entries, 0 to 11950
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   id               11951 non-null  object 
 1   url              11951 non-null  object 
 2   title            11951 non-null  object 
 3   developer        11951 non-null  object 
 4   developer_link   11951 non-null  object 
 5   icon             11951 non-null  object 
 6   rating           11951 non-null  float64
 7   reviews_count    11951 non-null  int64  
 8   description_raw  11951 non-null  object 
 9   description      11951 non-null  object 
 10  tagline          0 non-null      float64
 11  pricing_hint     11951 non-null  object 
 12  lastmod          11951 non-null  object 
dtypes: float64(2), int64(1), object(10)
memory usage: 1.2+ MB
```

Після обробки, датасет має вигляд:

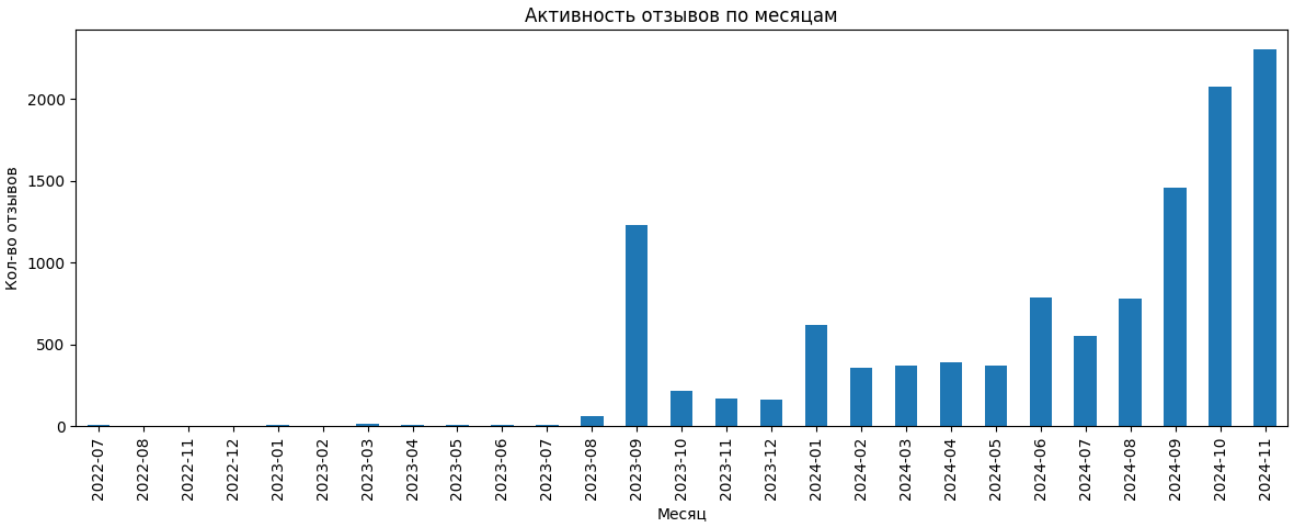
```
df.drop(columns=['id','url','developer_link','icon','description_raw','description', 'pricing_hint', 'tagline'], inplace=True)
df.head(10)
```

	title	developer	rating	reviews_count	lastmod
0	WISO MeinBüro	Buhl Data Service GmbH	5.0	2	2023-08-25
1	Iconic: Product Features	CartBoosters	4.7	70	2024-04-12
2	Checkbox RRO	Web-Systems Solutions	5.0	1	2024-08-29
3	CreditsYard — Store Credit	MerchantYard	4.4	22	2024-10-22
4	BSS Product Options, Variant	Tech Essence (by BSS Commerce)	5.0	641	2024-11-21
5	Jeluxpox	Jeluxpox	5.0	1	2023-12-06
6	Boost Shop: Countdown Timer	boostshoplimited	5.0	11	2024-01-28
7	Zeniva AI	Exarta	5.0	4	2024-10-30
8	Punch Kakao Social Login	Punch Digital	5.0	3	2023-11-26
9	Token of Trust Verification	Token of Trust, Inc	5.0	4	2023-10-26

3) Візуалізуємо топ 20 додатків за кількістю відгуків



4) Знайдемо місяці і візуалізуємо, в яких було створено найбільша кількість відгуків



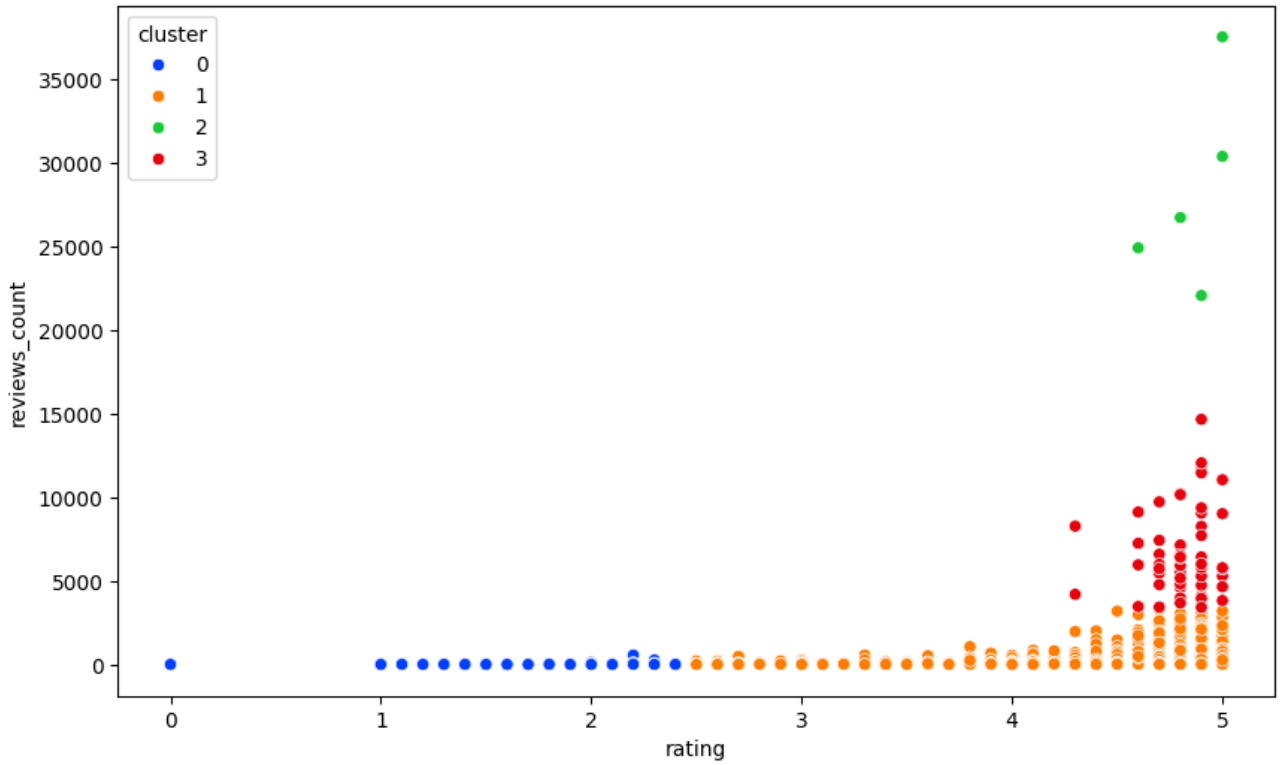
5) Кластеризація

```
X = df[['rating', 'reviews_count']]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

model = KMeans(n_clusters=4, random_state=42)
df['cluster'] = model.fit_predict(X_scaled)


plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='rating', y='reviews_count', hue='cluster', palette='bright')
plt.show()
```



Можна зробити такі висновки, що кореляція присутня між rating та reviews_count. Бачимо ,що додатки з високим рейтингом частіше мають більше відгуків, а додатки з низьким навпаки – менше. Це цілком логічно, так як популярні додатки найчастіше бувать якісними і продуманими та приваблюють більше юзерів, які і залишають відгуки. На графіку

Кластери чітко групують додатки за їхньою популярностю (кількість відгуків) та якістю (рейтинг). Це зозволяє легко обрати лідируючі додатки (кластер 2) та “проблемні” додатки (кластеры 0 и 1).

```
green_cluster_apps = df[df['cluster'] == 2]
green_cluster_apps[['title', 'developer', 'rating', 'reviews_count']]
```

	title	developer	rating	reviews_count
473	Loox - Product Reviews App	Loox	4.9	22066
2467	Judge.me Product Reviews App	Judge.me	5.0	37534
4773	DSers-AliExpress Dropshipping	DSers	5.0	30387
5085	TikTok	TikTok Inc.	4.8	26729
10214	Privy - Pop Ups, Email, & SMS	Privy Operations	4.6	24917

Додакти з низьким рейтингом і низькмими оцінками. Такі результати свідчать, що або додатки ще не набрали аудиторію (тобто про них ніхто не знає), або вони знаходяться на етапі розробки.

```
blue_cluster_apps = df[df['cluster'] == 0]
blue_cluster_apps[['title', 'developer', 'rating', 'reviews_count']]
```

	title	developer	rating	reviews_count
11	Konvera: A/B Testing	Konvera LLC	0.0	0
17	HelloConvo AI	HelloConvo	0.0	0
18	Adelfi	MBG	0.0	0
19	Foxdeli - Tracking & Up-sell	Foxdeli s.r.o.	0.0	0
20	Zeno Announcement Bar	Zenonian	0.0	0
...	...	...	...	...
11939	GetResponse	GetResponse S.A.	1.0	1
11940	Reescribir Textos	Enzipe Apps	0.0	0
11941	Peregrine Ship	Peregrine Ship	0.0	0
11943	Brand It! Calendar	JadePuma	0.0	0
11949	SwissID	tpm solutions	0.0	0

4589 rows × 4 columns

```
orange_cluster_apps = df[df['cluster'] == 1]
orange_cluster_apps[['title', 'developer', 'rating', 'reviews_count']]
```

	title	developer	rating	reviews_count
0	WISO MeinBüro	Buhl Data Service GmbH	5.0	2
1	Iconic: Product Features	CartBoosters	4.7	70
2	Checkbox RRO	Web-Systems Solutions	5.0	1
3	CreditsYard — Store Credit	MerchantYard	4.4	22
4	BSS Product Options, Variant	Tech Essence (by BSS Commerce)	5.0	641
...	...	...	...	...
11945	Octo Sales: AI Chatbot Popup	One Raino Tech	4.9	8
11946	Pickware	Pickware GmbH	4.8	10
11947	Australia Post EZ Label	Bitnext	5.0	260
11948	Endear CRM and Clienteling	Endear	4.5	21
11950	3cliques	3cliques	4.5	20

7298 rows × 4 columns

Загалом можна зробити висновки, додатки кластера 0 та 1 потребують у покращенні якості або маркетингової підтримки, щоб залучити більше користувачів та підвищити рейтинг. Додатки кластера 2 - це успішні продукти, які можна використовувати як приклад для інших розробників. Додатки кластера 3 показують гарний потенціал і можуть бути поліпшені для досягнення рівня популярності кластера 2.