

Име на проекта: *Анализ на потребителски данни за
препоръчителна система*



Автор: Николай Вещов – Data Analysis & AI Junior Freelancer

Задача: Анализ на потребителски данни за препоръчителна система

🎯 Цел

Представи си, че работиш в екип, който изгражда препоръчителна система за онлайн платформа. Твоята задача е да анализираш данните за потребителска активност и да извлечеш основни статистики и зависимости.

📊 Данни

Ще работиш със следната примерна таблица:

	User_ID	Product_ID	Rating	Timestamp
	101	A234	4	06/01/2025 14:25
	102	B678	5	06/01/2025 14:27
	101	B678	3	06/01/2025 14:30
▶	103	A234	2	06/01/2025 14:33

Таблицата може да бъде записана като CSV файл: `user_ratings.csv`

🔧 Задачи

- С SQL:**
 - Извлечи уникалните потребители и броя оценки, които всеки е дал.
 - Намери средната оценка за всеки продукт.
- С Excel:**
 - Зареди файла и използвай Pivot таблица, за да покажеш средната оценка по продукт.
 - Филтрирай по оценки ≥ 4 .
- С Python (с Pandas):**

python

```
import pandas as pd
```

```
df = pd.read_csv("user_ratings.csv")
print(df.groupby("Product_ID")["Rating"].mean())
print(df.groupby("User_ID").size())
```

Така, аз, Николай Вецов съм на ход:

1. В SQL качих файла, първоначално файл XLSX (Екселски файл), после конвертиран в SQL, след качването в системата на SQL workbench, установих следното:

	User_ID	Product_ID	Rating	Timestamp
	101	A234	4	06/01/2025 14:25
	102	B678	5	06/01/2025 14:27
	101	B678	3	06/01/2025 14:30
▶	103	A234	2	06/01/2025 14:33

При установяване на средната оценка, установих следното:

Имаме общо 3 уникални потребители, които са 101 (гласувал 2 пъти), 102 и 103. Продуктите са A234 и B678.

За продукт A234, средната оценка е 3. Защо? Защото имаме две оценки до тук и те са 4 и 2 и цифрата между тях е 3 и затова.

За продукт B678, средната оценка е 4. Тъй като има 5 и 3 и така средната оценка е 4.

Извод:

- Средната оценка на продукт A234 е 3.
- Средната оценка на продукт B678 е 4.

2. След качване на данните в Microsoft Excel, след надстойване се появи средната оценка и филтрирах по оценка ≥ 4 .

Row Labels	Average of Rating	Row Labels	Average of Rating
A234	3	B678	4
B678	4	Grand Total	4
Grand Total	3,5		

Based on what we see for the product with ID A234, the average rating is 3, because there are two ratings: 4 (from unique user 101) and 2 (from unique user 103). Therefore, by this logic, the average rating is 3. The same applies to product B678, which has two ratings: 5 (from user 102) and 3 (from user 101). Thus, the average rating is 4.

In the upper-left part of the table, the average rating for both products is displayed. In the table to the right, the average rating for product B678 is shown, because the task I was given requires filtering the data by ratings ≥ 4 (greater than or equal to 4).

User_ID	Product_ID	Rating	Timestamp
101	A234	4	1.6.2025 14:25
102	B678	5	1.6.2025 14:27
101	B678	3	1.6.2025 14:30
103	A234	2	1.6.2025 14:33

Според това, което виждаме за продукт, който е с номер A234, средната оценка е 3, поради причината, че има общо две оценки, които са 4 (от уникален потребител 101) и 2 (от уникален потребител 103). Следователно по тази логика, имаме средна оценка = 3. Същото се отнася и за продукт с номер B678. Общо две оценки - 5 (потребител с номер 102) и 3 (потребител с номер 101). Следователно средната оценка е 4. Горевляво на таблицата е показано за всеки един от двата продукта средната оценка. А в таблицата отдясно на нея се показва средната оценка на продукт B678, тъй като задачата, която ми е дадена, изисква да филтрирам ситуацията по оценки ≥ 4 (те по-голямо или равно на 4).

3. След стартиране на кода на Python, получих следния резултат:

A	B
Product_ID	Average_Rating
B678	4
A234	3

Среден рейтинг по продукт. (4 и 3)

A	B
User_ID	Ratings_Count
101	2
102	1
103	1

Потребителска активност. (потребител 101 – 2 пъти гласувал, потребител 102 – 1 път и потребител 103 – 1 път.

A	B	C	D
User_ID	Product_ID	Rating	Timestamp
101	A234	4	1.6.2025 14:25
102	B678	5	1.6.2025 14:27

Продукти с рейтинг ≥ 4 – т.е да показва продукти с рейтинг по-голям или = на 4. Оказва се, че и двата продукта имат такъв рейтинг.

A	B	C
User_ID	A234	B678
101	4	3
102		5
103	2	

Рейтинг матрица (таблица, която показва **оценките на потребители към различни обекти**).

Според нея – потребител 101 е оценил продукт A234 и продукт B678. Потребител 102 е оценил само и единствено продукт B678, а потребител 103 е оценил само и единствено продукт A234.

ВСИЧКО ТОВА, КОЕТО ПОКАЗАХ В ТЕЗИ 4 ФИГУРИ, ПОКАЗВА АНАЛИЗА НА ПОТРЕБИТЕЛИТЕ ПРИ РАЗЛИЧНИТЕ ПРОДУКТИ. ВСЕКИ ОТ ТЯХ ОЦЕНЯВА РАЗЛИЧНИТЕ ПРОДУКТИ ПО РАЗЛИЧЕН НАЧИН. ОКАЗВА СЕ, ЧЕ ПРОДУКТ B678 Е ПОВЕЧЕ ПРЕДПОЧИТАН, ЗА РАЗЛИКЕ ОТ ПРОДУКТ A234. И ВСИЧКО ТОВА Е КРАЙНИЯТ РЕЗУЛТАТ НА ТОЗИ ПОТРЕБИТЕЛСКИ АНАЛИЗ. ПО-НАДОЛУ ЩЕ ВИДИТЕ И КОДА, КОЙТО ИЗПОЛЗВАХ НА PYTHON.

Кодът, който използвах (Python):

```
import pandas as pd

# Зареждане на CSV с правилен разделител
df = pd.read_csv("user_ratings.csv", sep=";", encoding="utf-8")

# Анализ 1: Средна оценка по продукт
product_avg =
df.groupby("Product_ID")["Rating"].mean().sort_values(ascending=False)

# Анализ 2: Активност на потребители
user_activity = df["User_ID"].value_counts()

# Анализ 3: Продукти с рейтинг ≥ 4
top_rated = df[df["Rating"] >= 4]

# Анализ 4: User-Product рейтинг матрица
rating_matrix = df.pivot_table(index="User_ID", columns="Product_ID",
values="Rating")

# Записване във Excel файл
with pd.ExcelWriter("user_analysis.xlsx") as writer:
    product_avg.to_frame(name="Average_Rating").to_excel(writer,
sheet_name="Среден рейтинг по продукт")
    user_activity.to_frame(name="Ratings_Count").to_excel(writer,
sheet_name="Активност на потребители")
    top_rated.to_excel(writer, sheet_name="Продукти с рейтинг ≥4",
index=False)
    rating_matrix.to_excel(writer, sheet_name="Рейтинг матрица")
```

С този код, генерирах самия краен резултат в Excel файл, които в момента е user_analysis.xlsx. Точно там е целият анализ.