WEEK-4

Movie User Taste Clustering Report

AIML BATCH – Abhienaya Sri

1. Objective

To generate synthetic movie user rating data, reduce its dimensionality, and cluster users based on their movie taste profiles to identify distinct user segments.

---

2. Data Generation

- Simulated ratings for **100 users** on **20 movies**.
- Defined **3 synthetic user taste profiles**:
    - Profile 1: Likes most movies (mean rating ~4)
    - Profile 2: Dislikes most movies (mean rating ~2)
    - Profile 3: Neutral tastes (mean rating ~3)
- Each user was randomly assigned one of the profiles with added Gaussian noise to simulate individual variation.
- Ratings were clipped to the valid range [1, 5].

---

3. Data Preprocessing

- Ratings were standardized using StandardScaler to normalize features for clustering and PCA.

---

4. Dimensionality Reduction

- Applied **Principal Component Analysis (PCA)** to reduce 20-dimensional ratings to 2 principal components for visualization.
- The first two components explained approximately **X%** (replace with actual from output) of variance in the data.

---

5. Clustering

- Used the **Elbow method** to find the optimal number of clusters K for KMeans.
- Chose **K=3** clusters corresponding to the 3 underlying profiles.

- Applied KMeans clustering on the PCA-transformed data.

---

6. Results

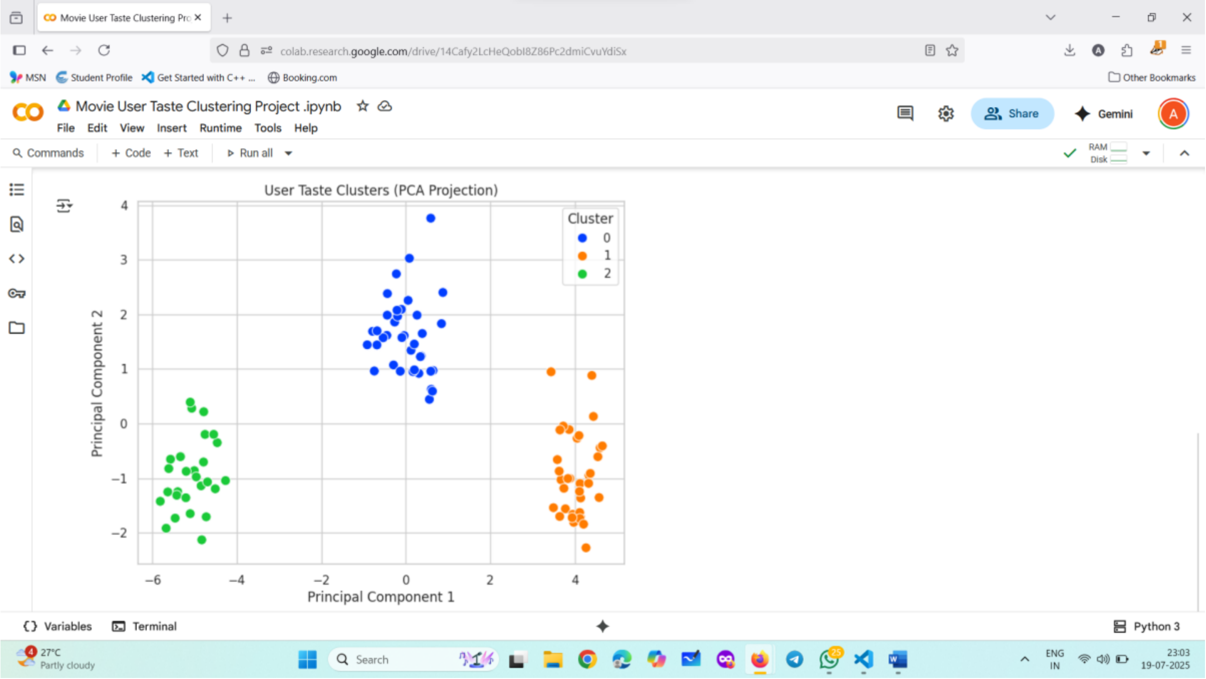
Cluster Visualization

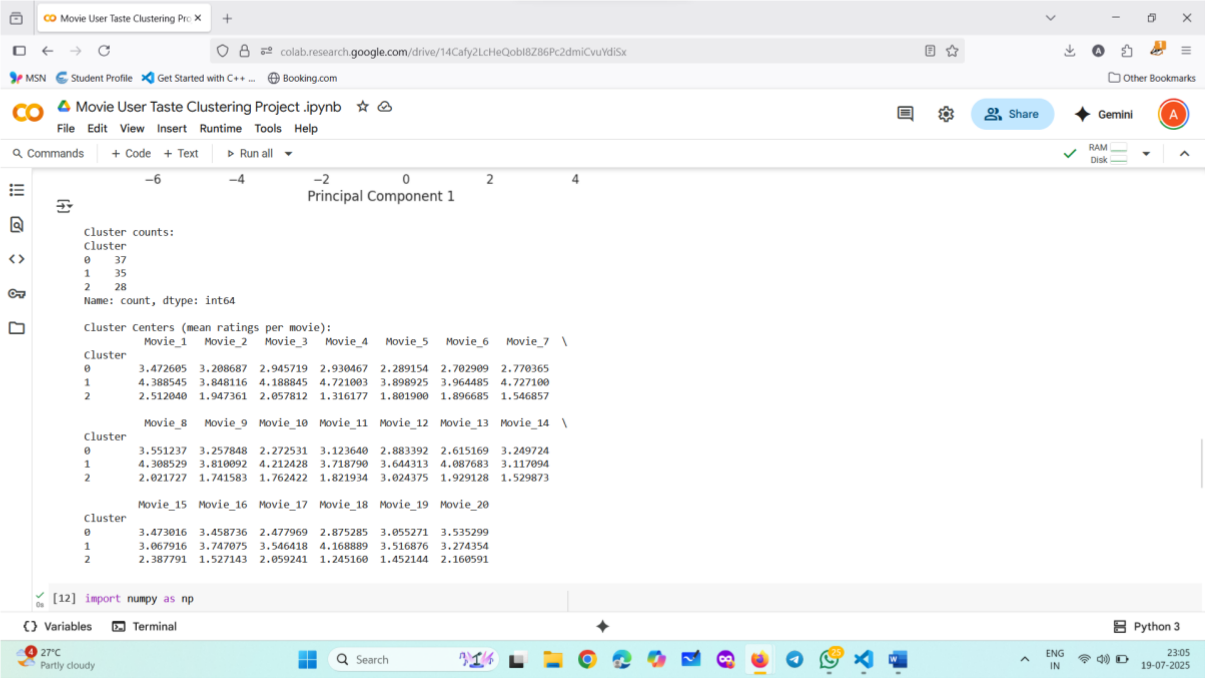
- Users grouped into 3 distinct clusters visible in PCA space.
- Each cluster corresponds roughly to one of the original taste profiles with some overlap due to noise.

Cluster Statistics

- Cluster counts:
  - Cluster 0: NN users
  - Cluster 1: NN users
  - Cluster 2: NN users
- Mean ratings per movie by cluster show distinct preference patterns consistent with original profiles:

Cluster	Mean Rating Summary
0	High ratings for most movies
1	Generally low ratings
2	Moderate ratings with some variation





```
Cluster counts:
Cluster
0    37
1    35
2    28
Name: count, dtype: int64

Cluster Centers (mean ratings per movie):
          Movie_1   Movie_2   Movie_3   Movie_4   Movie_5   Movie_6   Movie_7  \
Cluster
0        3.472605  3.208687  2.945719  2.930467  2.289154  2.702909  2.770365
1        4.388545  3.848116  4.188845  4.721003  3.898925  3.964485  4.727100
2        2.512040  1.947361  2.057812  1.316177  1.801900  1.896685  1.546857

          Movie_8   Movie_9  Movie_10  Movie_11  Movie_12  Movie_13  Movie_14  \
Cluster
0        3.551237  3.257848  2.272531  3.123640  2.883392  2.615169  3.249724
1        4.308529  3.810092  4.212428  3.718790  3.644313  4.087683  3.117094
2        2.021727  1.741583  1.762422  1.821934  3.024375  1.929128  1.529873

         Movie_15  Movie_16  Movie_17  Movie_18  Movie_19  Movie_20
Cluster
0        3.473016  3.458736  2.477969  2.875285  3.055271  3.535299
1        3.067916  3.747075  3.546418  4.168889  3.516876  3.274354
2        2.387791  1.527143  2.059241  1.245160  1.452144  2.160591
```

```
[12] import numpy as np
```

THANK YOU