**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

**(Term June-July 2025)**

**Anime Recommendation System**

**Submitted by**

Amisha Mehta 12310098

Shweta Singh 12301120

Harshita Gour 12318975

Piyush Thakur 12311114

**Course Code: PETV76**

**Under the Guidance of**

**Sandeep Kaur (UID: 23614)**

**Jaffar Amin Chacket (UID: 30453)**

**School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Amisha Mehta, Shweta Singh, Harshita Gour, Piyush Thakur bearing Registration no. 12310098, 12301120, 12318975, 12311114 has completed **PETV76** project titled, **“Anime Recommendation System”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science Engineering**

Lovely Professional University Phagwara, Punjab.

Date: 14.07.2025

**Acknowledgement**

We would like to express my sincere gratitude to Lovely Professional University for providing the opportunity and platform to undertake this project.

We are especially thankful toSandeep Kaur, Jaffar Amin Chacket for their invaluable guidance, constant support, and insightful feedback throughout the course of this project. Their mentorship played a crucial role in shaping the direction and success of my work. We also wish to thank my faculty mentors, peers, and friends for their encouragement and helpful suggestions during the project.

This experience has significantly enhanced my technical, analytical, and visualization skills, particularly in working with dataset. Finally, We extend our heartfelt thanks to everyone who directly or indirectly contributed to the successful completion of this project**.**

**Table of Content**

1. Introduction

* Company profile
* Overview of training domain
* Objective of the project

1. Training Overview

* Tools & technologies used
* Areas covered during training
* Daily/weekly work summary

1. Project Details

* Title of the project
* Problem definition
* Scope and objectives
* System Requirements
* Architecture Diagram (if any)
* Data flow / UML Diagrams

1. Implementation

* Tools used
* Methodology
* Modules / Screenshots
* Code snippets (if needed)

1. Results and Discussion

* Output / Report
* Challenges faced
* Learnings

1. Conclusion

* Summary

1. **Introduction**

**Company Profile**

Lovely Professional University (LPU) is one of India's premier academic institutions, recognized nationally and internationally for its innovative approach to education and skill development. Accredited with a prestigious NAAC Grade A++, LPU is known for its commitment to academic excellence, industry collaboration, and future-ready training programs. The university consistently offers programs that bridge the gap between academic learning and real-world industry needs. One such initiative is the **Online Summer Internship Program** titled “From Data to Insights: A Hands-On Approach to Data Science”, designed under the guidance of experienced faculty members such as **Ms. Sandeep Kaur, Mr. Jaffar Amin Chacket** Assistant Professor at the School of Computer Science. LPU’s strong emphasis on practical learning and industry-aligned curriculum ensures that students not only gain knowledge but also acquire job-relevant skills.

**Overview of Training Domain**

The training focused on the ever-evolving and highly impactful domain of **Data Science**, which integrates statistical analysis, data visualization, machine learning, and decision science to extract meaningful insights from raw data. The domain is crucial in today’s digital age, where data-driven decisions are transforming industries like healthcare, finance, marketing, and logistics. The training incorporated essential elements of data science, including **SQL for data querying**, **Excel for data handling**, **Power BI for business intelligence dashboards**, and **Python** for scripting and implementing **machine learning algorithms**. It provided a comprehensive introduction to modern data tools and techniques that help professionals convert data into actionable decisions. The combination of theory and practical exposure offered during the internship provided a solid foundation for understanding real-world data problems and preparing participants for technical roles in the data domain.

**Objective of the Project**

The primary objective of the project during this internship was to apply the learned tools and techniques in a real-world data science scenario through a **hands-on final project and weekly assignments**. The training aimed to build a strong conceptual understanding of how data can be collected, processed, analyzed, and visualized to support decision-making processes. Specifically, the project focused on solving business-related data problems using Python libraries such as pandas, NumPy, and scikit-learn, and on presenting data insights via interactive dashboards created in Power BI. Another key goal was to prepare students for placement opportunities by enhancing their skills in **SQL** and data handling—two of the most in-demand competencies in today’s tech-driven job market. By the end of the internship, students were expected to confidently approach data problems, design end-to-end data pipelines, and derive actionable insights from structured and unstructured data

1. **Training Overview**

**Tools & Technologies Used**

During the internship, a range of tools and technologies were utilized to gain hands-on experience in various aspects of data science and analytics. The key tools and platforms included:

* SQL (Structured Query Language): Used for data retrieval, filtering, joining, and aggregation from relational databases. This was especially helpful for placement preparation.
* Microsoft Excel: Employed for data cleaning, tabulation, analysis, and visualization using pivot tables, functions, and charts.
* Power BI: A business analytics service used for creating interactive data visualizations and dashboards.
* Python Programming Language: Used extensively throughout the training for scripting, automation, and machine learning.
* Python Libraries:
  + pandas – for data manipulation and analysis
  + NumPy – for numerical computations
  + matplotlib/seaborn – for data visualization
  + scikit-learn – for machine learning models

**Areas Covered**

The training was structured to provide in-depth knowledge and practical skills in the following areas:

* SQL for Placement Prep: Writing optimized queries, joining multiple tables, filtering data, and using aggregate functions.
* Data Handling with Excel: Organizing datasets, using Excel functions for data cleaning and transformation, and creating basic charts.
* Power BI Dashboards: Importing datasets, applying filters, generating visuals, creating calculated fields, and building dashboards.
* Python Programming Basics: Understanding syntax, control structures, functions, and file handling.
* Data Analysis with Python Libraries: Data preprocessing using pandas, working with arrays in NumPy, and visualizing data with matplotlib/seaborn.
* Machine Learning with Python: Introduction to supervised learning algorithms like Linear Regression, Decision Trees, and model evaluation techniques.

**Daily/Weekly Summary**

Here is a week-wise breakdown of the training schedule:

Week 1: Introduction & SQL

* Orientation and overview of the internship
* Basics of data science, importance of data-driven decisions
* SQL syntax, SELECT statements, WHERE clause, GROUP BY, HAVING, and JOIN operations
* Practice exercises on placement-style questions using SQL

Week 2: Excel for Data Handling

* Understanding raw data structures and formats
* Applying Excel formulas for cleaning and transforming data
* Use of pivot tables and basic charts
* Mini assignment on data summarization using Excel

Week 3: Power BI Dashboards

* Introduction to business intelligence tools
* Connecting datasets in Power BI
* Creating visualizations, slicers, filters, and KPIs
* Designing a dashboard for sales or marketing data

Week 4: Python & Data Analysis

* Basics of Python: variables, loops, functions, data types
* Data handling with pandas and NumPy
* Data cleaning, missing values handling, and exploratory data analysis
* Plotting data with matplotlib and seaborn

Week 5: Machine Learning

* Overview of machine learning concepts
* Supervised learning models – Linear Regression, Classification
* Model training, testing, and performance evaluation
* Use of scikit-learn for implementing ML workflows

1. **Project Details**

**Project Title**

**"Anime Recommendation System with User Clustering and Power BI Visualization"**

**Problem Definition**

With the rapid growth of streaming platforms and digital entertainment, users are often overwhelmed by the sheer number of choices available. Recommending relevant anime based on user preferences can significantly enhance the viewing experience by reducing decision fatigue and increasing engagement. The core problem lies in the lack of personalized recommendation systems that utilize past user behavior and anime characteristics to suggest content effectively. This project aims to address this gap by developing a recommendation engine that uses user watch history, genre preferences, and rating patterns to provide customized anime suggestions. Additionally, Power BI dashboards will be used to visualize genre trends, user engagement, and recommendation performance, offering both technical and business-level insights.

**Scope and Objectives**

**Scope:**

The project includes data preprocessing, user profiling through clustering, and prediction of interest levels using machine learning regression models. It also integrates a dashboarding layer using Power BI for rich, interactive insights. The scope is limited to structured data (e.g., user ratings, genre tags) and does not cover real-time recommendation deployment or integration with actual streaming platforms.

**Objectives:**

* To analyze user behavior using the anime ratings dataset.
* To build a recommendation engine using clustering (e.g., K-Means) or regression techniques.
* To extract insights about anime genres, user preferences, and rating trends.
* To design Power BI dashboards to visualize:
  + Genre popularity
  + User engagement patterns
  + Accuracy and coverage of recommendations

**System Requirements**

**Software Requirements:**

* **Python** 
  + Libraries: pandas, NumPy, scikit-learn, matplotlib, seaborn
* **Jupyter Notebook / VS Code**
* **Microsoft Power BI Desktop**

1. **Implementation**

**Tools Used**

The following tools and libraries were used throughout the development and analysis phases of the Anime Recommendation System:

* **Python 3.x**: Primary language for data processing and model development
* **Jupyter Notebook**: Development environment for interactive data analysis
* **pandas**: For data manipulation and preprocessing
* **NumPy**: For numerical operations
* **scikit-learn**: For clustering (K-Means) and similarity-based recommendation
* **matplotlib / seaborn**: For plotting and exploratory data analysis
* **Power BI Desktop**: For visualizing anime trends, genre popularity, and user patterns
* **Excel**: Optional support for verifying and cleaning raw CSV data

**Methodology**

The development of the Anime Recommendation System followed these systematic steps:

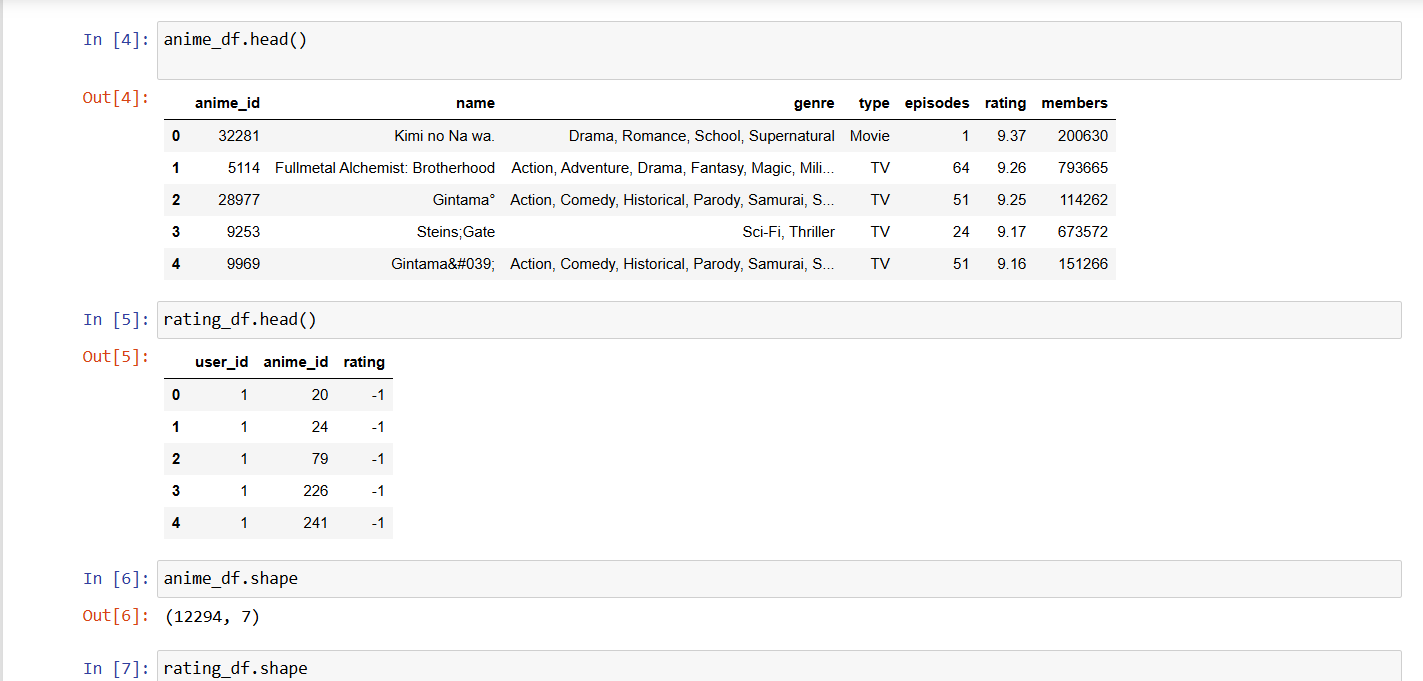
1. **Dataset**   
   Datasets were sourced from Kaggle’s Anime Recommendation Database, including user ratings and anime metadata such as genres, type, and episode count.
2. **Data Cleaning and Preprocessing**:
   * Handled missing values in genres and ratings.
   * Converted genres into multi-label binary columns.
   * Normalized user ratings and removed sparse entries to ensure data quality.
3. **Exploratory Data Analysis (EDA)**:
   * Analyzed the distribution of ratings.
   * Visualized the most common genres and top-rated anime.
   * Identified user behavior patterns using grouping and aggregation.
4. **User Clustering (K-Means)**:
   * Created user profiles based on their genre preferences and ratings.
   * Applied K-Means clustering to segment users with similar anime tastes.
5. **Recommendation Engine**:
   * Used cosine similarity on user profiles to recommend anime titles.
   * Implemented a regression-based model to predict user interest levels.
6. **Visualization using Power BI**:
   * Exported processed data to Power BI.
   * Built dashboards to showcase:
     + Genre popularity
     + Anime rating trends
     + Top recommendations per user cluster
     + User engagement distribution

**Modules/ Screenshots**

**Module 1: Data Cleaning & Preprocessing**

Description:  
The raw dataset obtained from Kaggle included missing values, inconsistent genre formats, and rating anomalies. This module handled:

* Null value imputation (e.g., replacing missing genres with "Unknown")
* Dropping irrelevant or sparse user-anime pairs
* One-hot encoding of genre columns for modeling



* 1. Data Pre Processing

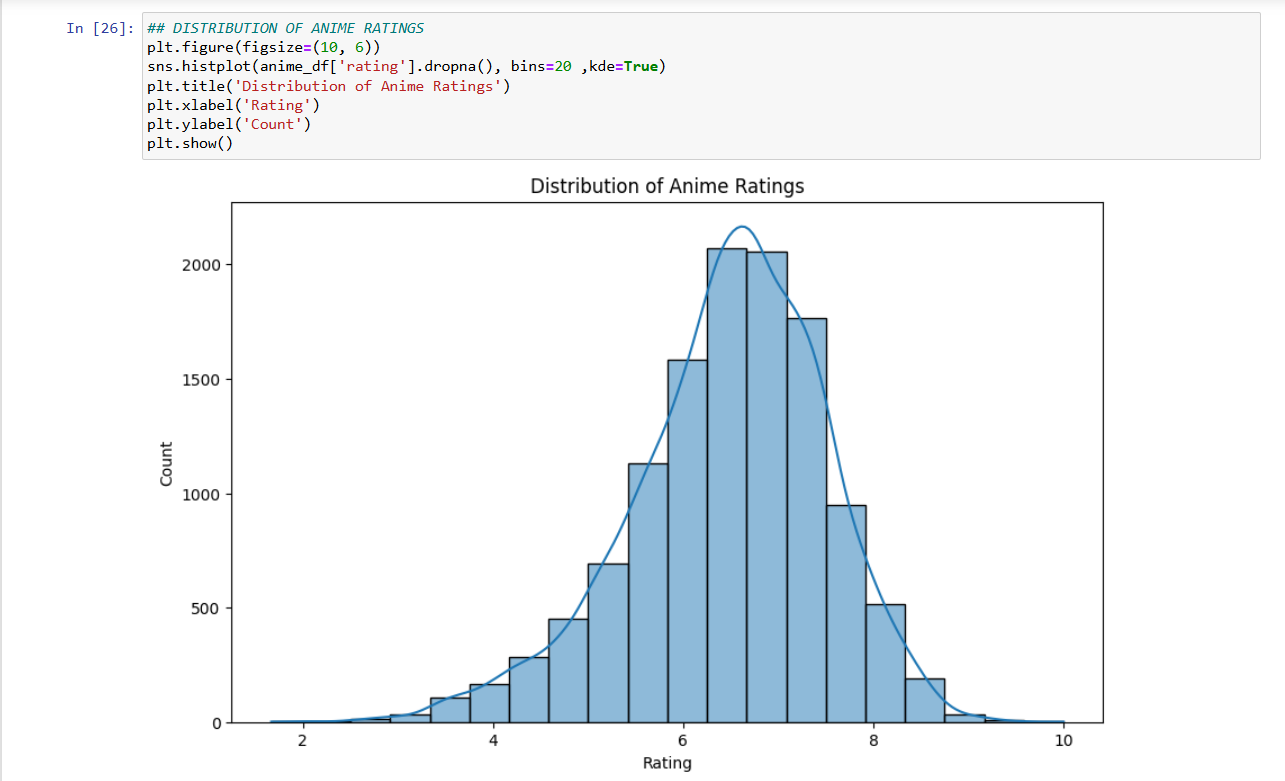


* 1. Data Cleaning

**Module 2: Exploratory Data Analysis (EDA)**

Description:  
Exploration was conducted to understand user behavior and genre trends:

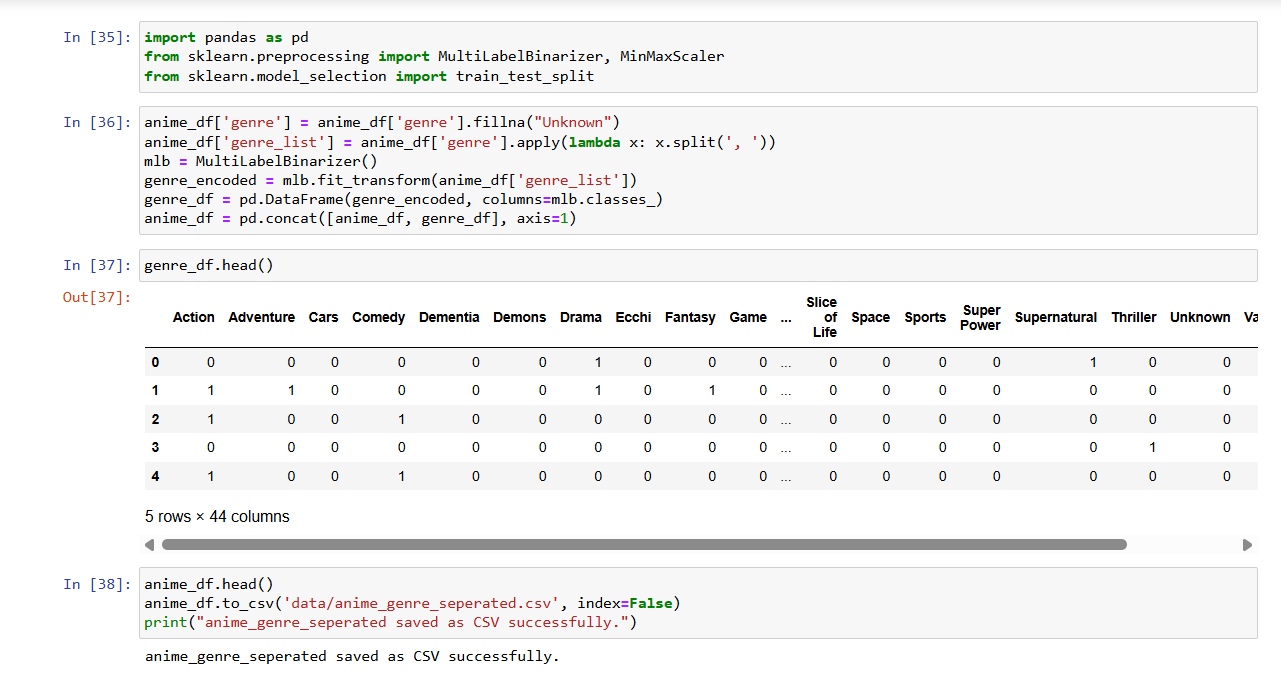
* Top-rated anime series
* Most watched genres
* Average rating per genre
* Number of unique users and anime titles



* 1. Exploratory Data Analysis

**Module 3: User Clustering (K-Means)**

Description:  
Users were clustered based on their rating behavior and genre preferences using K-Means clustering. This grouped similar users to enable personalized recommendations.

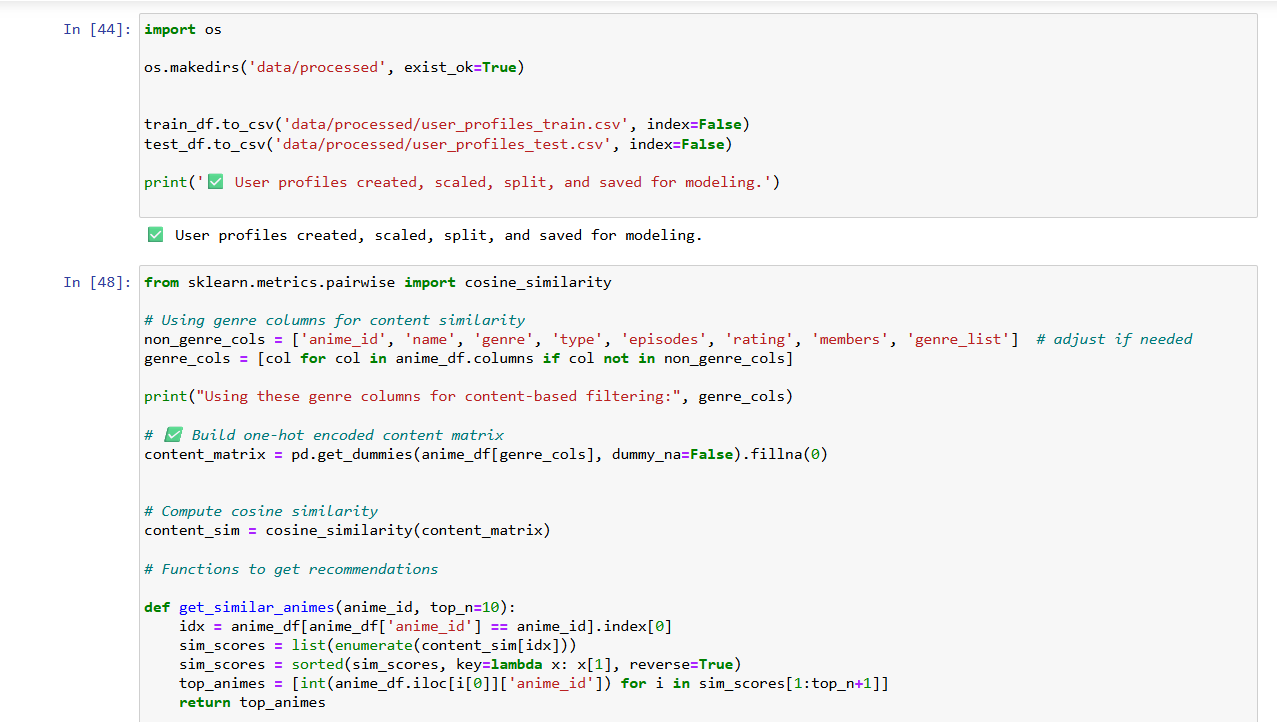


* 1. K-Means

**Module 4: Recommendation Engine**

Description:  
Two techniques were implemented:

* Similarity-based Filtering: Cosine similarity between users to suggest anime not watched but highly rated by similar users.
* Regression-based Prediction: Predicting expected rating for a user-anime pair.

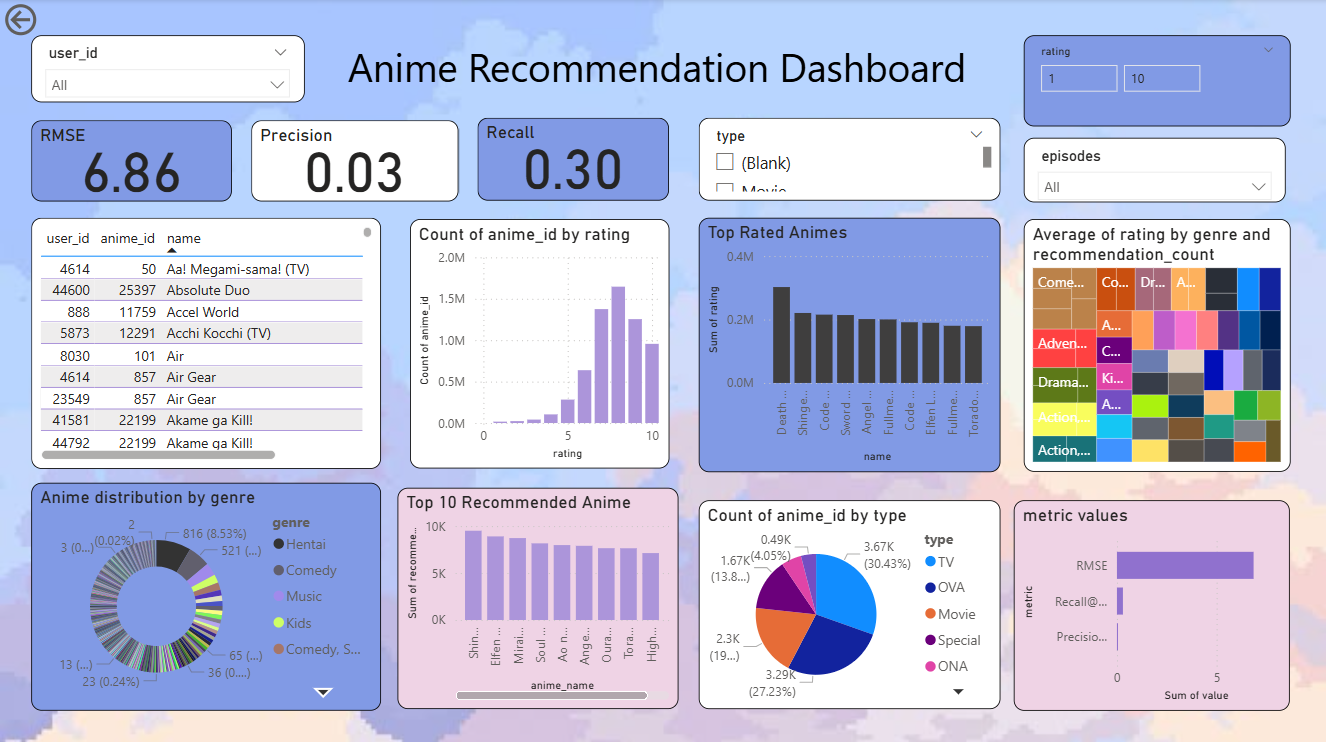


* 1. Recommendation Engine

**Module 5: Power BI Visualization Dashboard**

Description:  
This module focused on business-level visualization of project insights:

* Genre popularity over time
* Cluster-wise user behavior
* Anime watch frequency
* Recommendation performance summary



* 1. Power BI Dashboard

**Code Snippets**

**1.**

anime\_df['episodes'] = anime\_df['episodes'].replace('Unknown', np.nan)

anime\_df['episodes'] = pd.to\_numeric(anime\_df['episodes'], errors='coerce')

median\_episodes = anime\_df['episodes'].median()

anime\_df['episodes'] = anime\_df['episodes'].fillna(median\_episodes)

anime\_df['episodes']

**2.**

rating\_df = rating\_df.dropna()

rating\_df = rating\_df.drop\_duplicates()

rating\_df

rating\_df.to\_csv('data/rating.csv', index=False)

print("rating saved as CSV successfully.")

**3.**

## DISTRIBUTION OF ANIME RATINGS

plt.figure(figsize=(10, 6))

sns.histplot(anime\_df['rating'].dropna(), bins=20 ,kde=True)

plt.title('Distribution of Anime Ratings')

plt.xlabel('Rating')

plt.ylabel('Count')

plt.show()

**4.**

import os

os.makedirs('data/processed', exist\_ok=True)

train\_df.to\_csv('data/processed/user\_profiles\_train.csv', index=False)

test\_df.to\_csv('data/processed/user\_profiles\_test.csv', index=False)

print('✅ User profiles created, scaled, split, and saved for modeling.')

1. **Results and Discussions**

**Output / Reports**

The final outputs of the project successfully demonstrate a functioning Anime Recommendation System with both **analytical insights** and **recommendation capabilities**.

**Key Outputs Include:**

**Data Summary and Genre Distribution:**

The dataset was thoroughly cleaned and prepared. Missing values in genres were handled, and genres were encoded using one-hot encoding. Bar plots of the top 10 most common genres (e.g., Action, Comedy, Adventure) were generated to understand content distribution.

**User Behavior Analysis:**

Visualizations such as histograms of user rating frequencies and heatmaps of user-anime interaction were created. These helped identify how frequently users engage with anime and which genres or titles are rated highly.

**Clustering Results:**

K-Means clustering was used to group users based on their preferences. Users were effectively segmented into multiple clusters, each reflecting a specific taste profile (e.g., users who prefer Action and Sci-Fi vs. those who prefer Romance and Slice of Life). A scatter plot visualizing cluster separations was displayed using PCA.

**Recommendation Function Output:**

A working Python function was implemented that takes a user ID and returns a list of personalized anime recommendations. Cosine similarity was used for collaborative filtering. A sample output table lists anime titles, predicted interest scores, and genres.

**Power BI Dashboard:**

A Power BI dashboard was built to present:

* Genre popularity
* Cluster-wise user engagement
* Top-rated anime per cluster
* KPI cards for total users, top genres, and engagement metrics

**Challenges Faced**

Several technical and practical challenges were encountered during project development:

* **Missing or Incomplete Data:**  
  The raw dataset included missing values in genres and ratings. These had to be filled or dropped carefully to prevent bias or inconsistency in results.
* **Genre Encoding Complexity:**  
  Since anime often belong to multiple genres, it was challenging to encode them correctly without inflating the dimensionality of the feature space. Careful parsing and one-hot encoding were required.
* **Sparse User-Anime Matrix:**  
  Most users had rated only a small subset of anime titles, making collaborative filtering harder. Dimensionality reduction and filtering techniques were needed to reduce noise.
* **Model Evaluation Limitations:**  
  Since no ground-truth “correct recommendation” labels existed, evaluating recommendation accuracy was non-trivial. Alternative metrics like coverage and distribution were considered.
* **Power BI Data Sync:**  
  Importing large datasets into Power BI and maintaining relationships between tables required manual adjustments in Power Query and careful schema design.

**Learnings**

This project offered hands-on experience with real-world data science and recommendation systems. Key takeaways include:

* **Data Cleaning & Transformation:**Learned how to handle missing data, encode complex multi-label columns, and reshape data for ML models.
* **Exploratory Data Analysis (EDA):**Gained deep insight into visual storytelling and data pattern discovery using Python libraries.
* **Unsupervised Learning & Clustering:**  
  Applied K-Means to real user data and interpreted cluster profiles meaningfully.
* **Recommendation Engine Design:**  
  Implemented collaborative filtering and understood how to personalize suggestions using similarity matrices.
* **Data Visualization with Power BI:**  
  Designed dynamic dashboards, connected Python output with Power BI, and created business-ready visual reports.
* **Practical Pipeline Management:**  
  Understood the importance of data flow, system design, and report generation in full-cycle analytics projects.

1. **Conclusion**

**Project and Training Summary**

The **"Anime Recommendation System"** project, developed during the summer internship program *"From Data to Decisions: A Hands-On Approach to Data Science"*, was a comprehensive hands-on experience that combined practical learning with real-world data challenges. Through this project, we successfully designed and implemented a content-based and collaborative filtering recommendation engine capable of suggesting anime titles based on user preferences, genre interest, and rating patterns.

The training program organized by **Lovely Professional University (LPU)** covered essential technologies such as SQL, Excel, Power BI, and Python programming. These tools were effectively used in this project to manage data preprocessing, visualization, and predictive modeling. We explored data cleaning techniques, user clustering using **K-Means**, and similarity-based recommendation using **cosine similarity**. The project pipeline concluded with an interactive dashboard built in **Power BI**, which provided visual insights into user behavior, genre popularity, and recommendation outcomes.

Throughout the training, the emphasis on weekly tasks, mini-assignments, and the final capstone project enabled a deep understanding of the full data science lifecycle—from raw data handling to insights delivery.