Group 6

Team Members

Đoàn Ngọc Mai

21127104

Lê Nguyễn Kiều Oanh

21127129

Dương Trường Bình

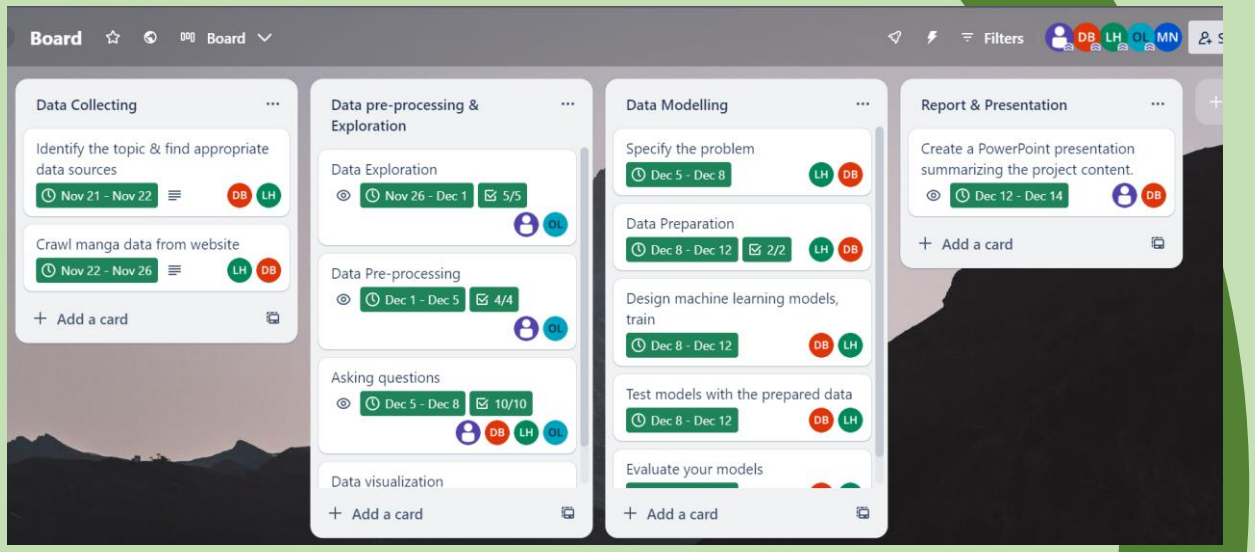
21127229

Lê Phước Quanh Huy

21127616

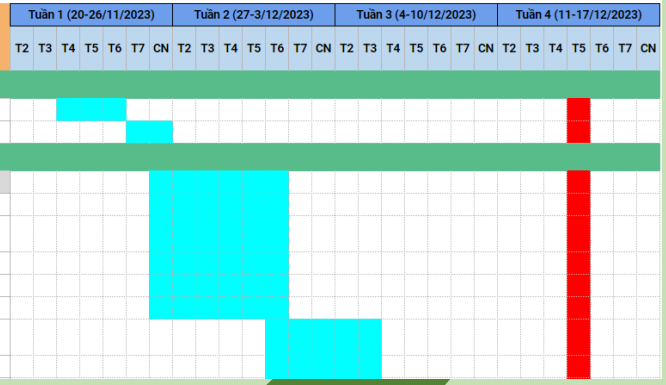
Project plan

- GitHub
- Gantt chart
- Trello



TIÊU ĐỀ DỰ ÁN	Final Project	LỚP	21KHDL1
NGƯỜI QUẢN LÝ DỰ ÁN	Dương Trường Bình	NGÀY	14/12/23

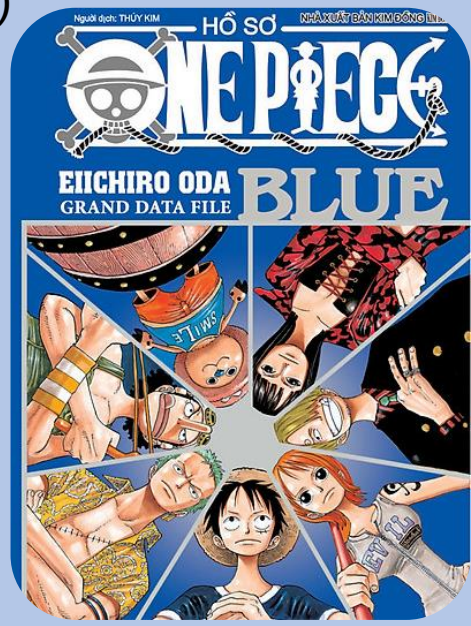
No	TIÊU ĐỀ CÔNG VIỆC	NGƯỜI PHỤ TRÁCH CÔNG VIỆC	NGÀY BẮT ĐẦU	NGÀY ĐẾN HẠN
1	**Data collecting**			
1.1	Analyze and scrape data from the website	Huy, Binh	22/11/23	24/11/23
1.2	Refactor code & documentation	Huy, Binh	25/11/23	26/11/23
2	**Data Exploration & pre-processing**			
2.1	Data Exploration	Oanh, Mai	26/11/23	01/12/23
2.1.1	Summary statistics for numerical columns	Oanh	26/11/23	01/12/23
2.1.2	Calculate and visualize to understand distribution patterns	Oanh	26/11/23	01/12/23
2.1.2.a	Data Distribution	Oanh, Mai	26/11/23	01/12/23
2.1.2.b	Explore numerical columns	Oanh, Mai	26/11/23	01/12/23
2.1.2.c	Explore categorical columns	Oanh, Mai	26/11/23	01/12/23
2.2	Data Pre-processing	Oanh, Mai, Binh	01/12/23	05/12/23
2.2.1	Check and handle duplicates rows	Oanh, Mai	01/12/23	05/12/23



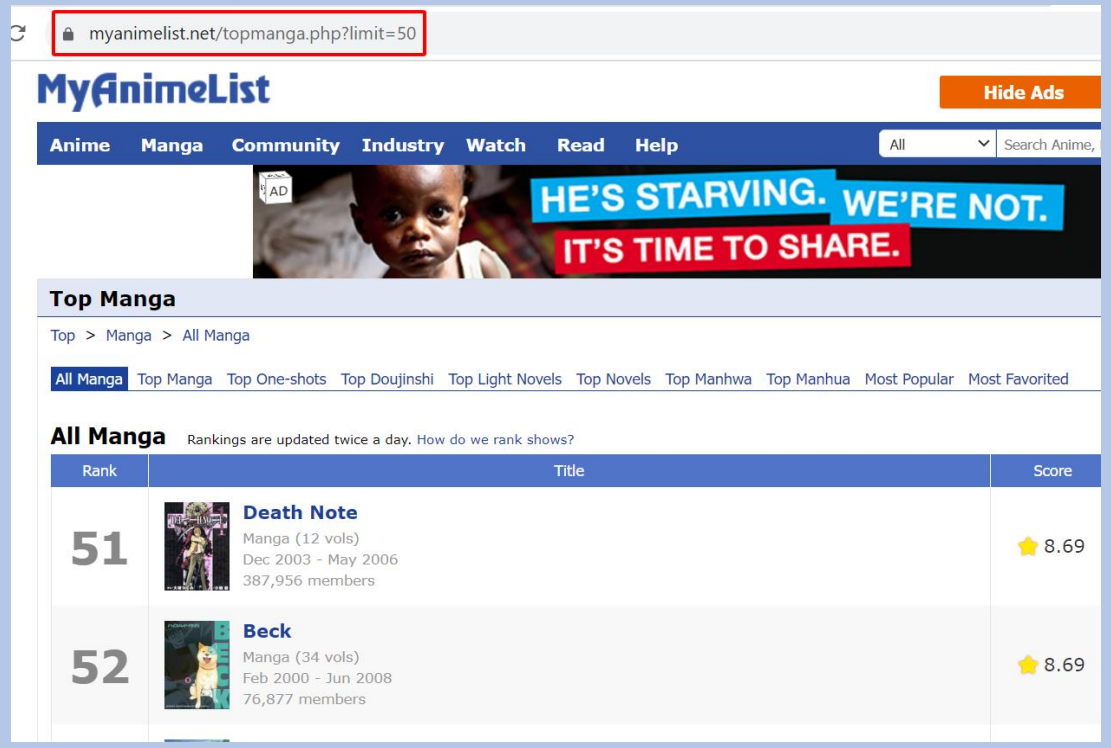
Data Collecting

I. Data Collecting

- Topic: Manga (Japanese comics)
- Website [MyAnimeList]
- Libraries :
  - requests
  - HTMLSession
  - BeautifulSoup
  - re
  - nest_asyncio
  - Pandas
  - datetime
  - time



I. Data Collecting: Top 10000 manga



Steps to collect the urls

1. Send request to the Top Manga page to get the HTML content of the page.
2. Parse the HTML content using BeautifulSoup.
3. Find and extract the urls from the HTML content.
4. Save the urls in a list.
5. Repeat the above steps by increasing the limit parameter by 50 each time until the limit parameter reaches 10000.

I. Data Collecting

For each manga, we collect the relevant data fields displayed on the webpage.

Information

Type: Manga

Volumes: Unknown

Chapters: Unknown

Status: Publishing

Published: Aug 25, 1989 to ?

Genres: Action, Adventure, Award Winning, Drama, Fantasy, Horror, Supernatural

Themes: Gore, Military, Mythology, Psychological

Demographic: Seinen

Serialization: Young Animal

Authors: Miura, Kentarou (Story & Art), Studio Gaga (Art)

Statistics

Score: 9.47[1] (scored by 331,875 users)

Ranked: #1[2]

Popularity: #1

Members: 666,443

Favorites: 122,994

10,000 rows

16 columns

- Title
- Score
- Vote
- Ranked
- Popularity
- Members
- Favorite
- Volumes
- Chapters
- Status
- Published
- Genres
- Themes
- Author
- Total Review
- Type Review

Data Exploration and Preprocessing

II. Data Exploration and Preprocessing

- The dataset has **10,000 rows** and **16 columns**,
- **Published** is split into **Release date** and **Completed date**.
- **Type Review** and **Total Review** are split into three new columns **Recommended, Mixed Feelings** and **Not Recommended**
- The dataset only has 1 duplicate line and it has been dropped

COLUMN	MEANING
`Title`	Title of the manga (written in English phonetic)
`Score`	Score on the MyAnimeList site (MAL)
`Vote`	Number of readers voting for the manga
`Ranked`	Ranking of manga on the web MyAnimeList (MAL)
`Popularity`	The popularity of the manga
`Members`	Number of readers who have this manga in their list
`Favorite`	Number of readers who love this manga
`Volumes`	Number of volumes of manga
`Chapters`	Number of chapters of manga
`Status`	Status of the manga (ongoing, completed, on hiatus,...)
`Published`	Release time to the end time of the manga
`Genres`	Genres of manga
`Themes`	The themes of the manga
`Author`	Author of manga
`Total Review`	Number of readers leaving comments on the manga
`Type Review`	Number of readers for each comment category (Recommended / Mixed feeling / Not recommended)

II. Data Exploration and Preprocessing

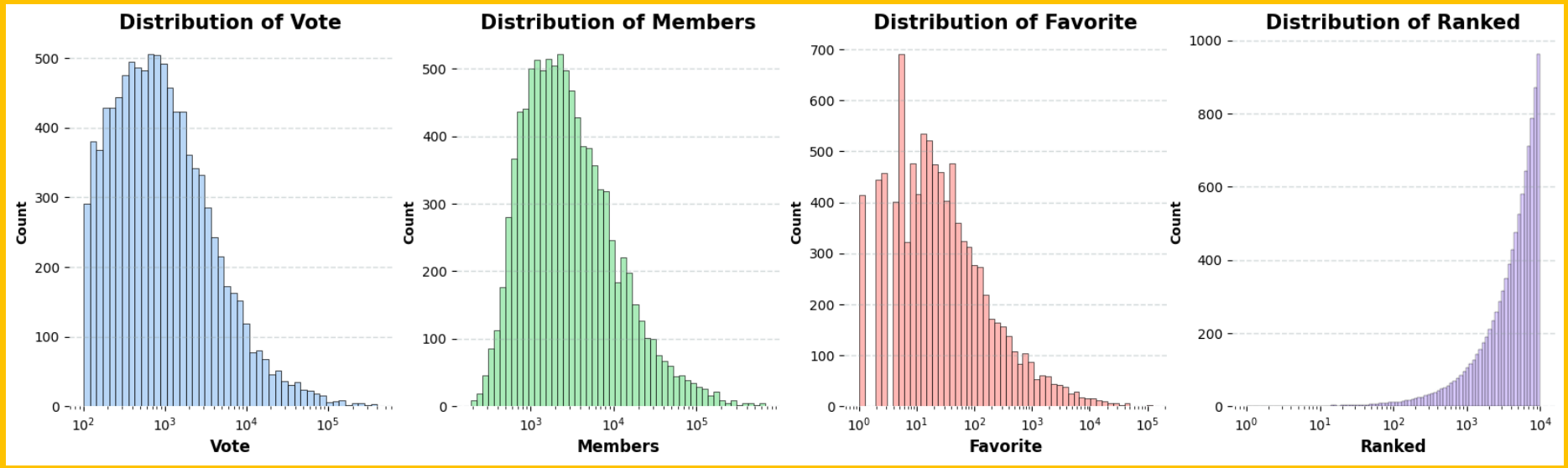
Data types conversion:

- Suitable Types: **Title**, **Score**, **Vote**, **Ranked**, **Popularity**, **Status**, and **Total Review**.
- Need Conversion to <u>int</u>**: Members**, **Favorite**, **Volumes**, and **Chapters** (After handling missing values)
- Need Conversion to <u>list</u>**: Type Review**, **Genres**, **Themes**, and **Author**.
- Need Conversion to <u>datetime</u>**: Release date** and **Completed date**

```
Title                                {<class 'str'>}
Score                              {<class 'float'>}
Vote                                 {<class 'int'>}
Ranked                               {<class 'int'>}
Popularity                           {<class 'int'>}
Members                              {<class 'str'>}
Favorite                             {<class 'str'>}
Volumes                              {<class 'str'>}
Chapters                             {<class 'str'>}
Status                               {<class 'str'>}
Published                            {<class 'str'>}
Genres                               {<class 'str'>}
Themes                               {<class 'str'>}
Author          {<class 'str'>, <class 'float'>}
Total Review                         {<class 'int'>}
Type Review                          {<class 'str'>}
dtype: object
```

II. Data Exploration and Preprocessing

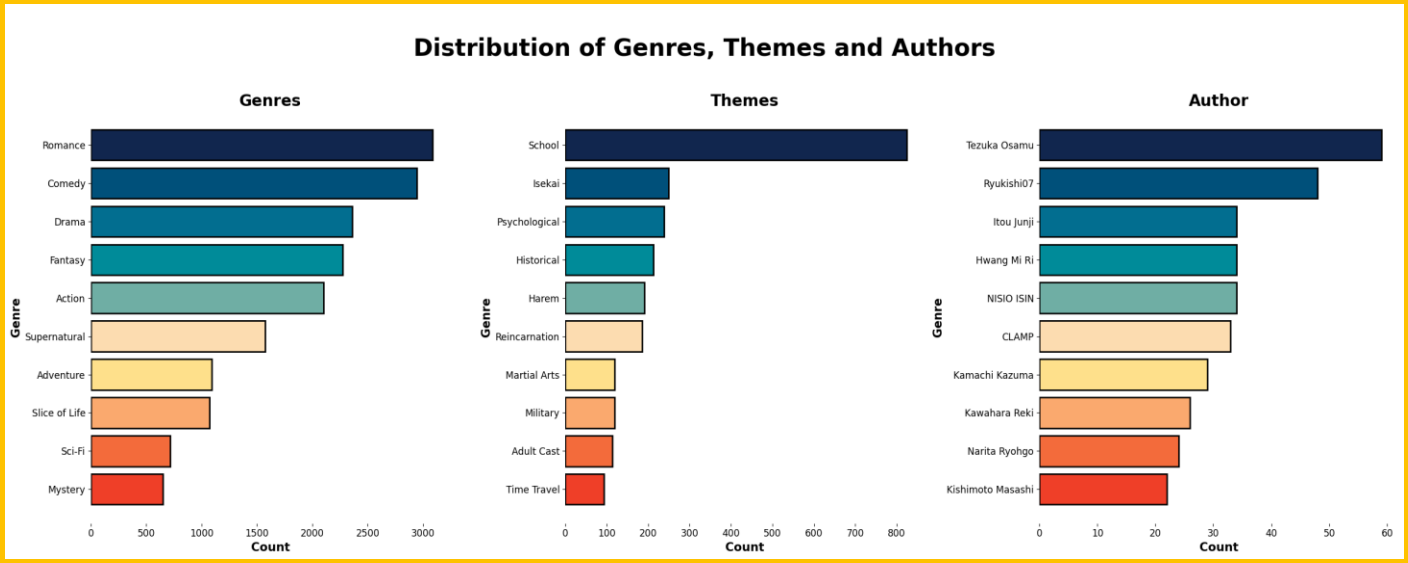
- The numerical columns in the dataset, sourced from the top 10,000 manga series on MyAnimeList, generally display **skewed** or **semi-normal** distributions.



II. Data Exploration and Preprocessing

- There are 5 categorical columns: **Title, Status, Genres Themes,** and **Author**
- For each categorical column, we will conduct exploration by calculating the **number of distinct values** and the **frequency of unique values (distribution)**
- Visualize the top 10 most frequently values

Name	Num_diff_vals
Title	9684
Status	4
Genres	19
Themes	52
Author	7226



II. Data Exploration and Preprocessing

Missing value

Most rows have minor missing values (1-3) out of 17 columns, with a maximum of 4, which is relatively insignificant.

Handle missing values

- ❑ Drop columns with missing values exceeding 75%: **Themes**
- ❑ Drop rows with missing values in the **Released date**
- ❑ Fill current date for missing **Completed date**
- ❑ Drop rows with missing values in the **Genres** column
- ❑ Fill missing values in **Volumes** and **Chapters** with the median.

Missing Values per Column

Name	Missing_ratio
Volumes	26.3%
Chapters	25.3%
Release date	16.2%
Completed date	38.07%
Genres	27%
Themes	82%

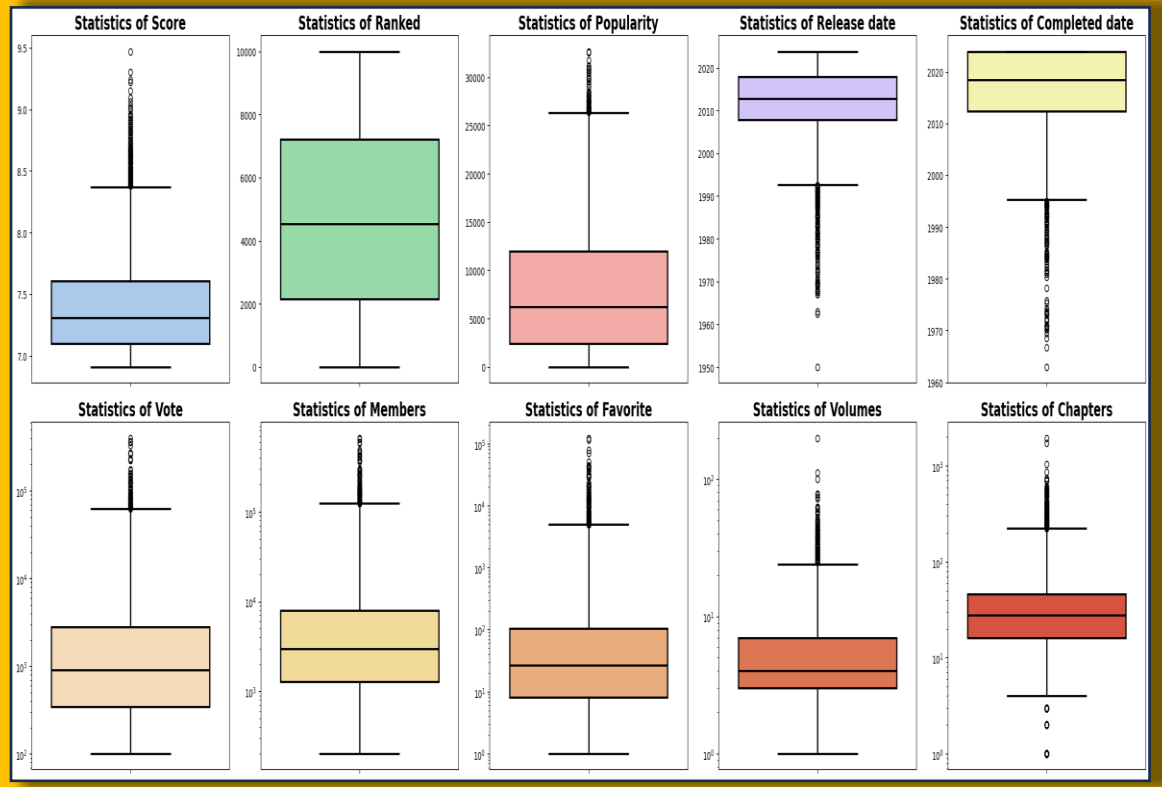
Missing Values per Row

Number of Missing Values	Number of Rows
0	6128
1	2029
2	1507
3	1763
4	65

II. Data Exploration and Preprocessing

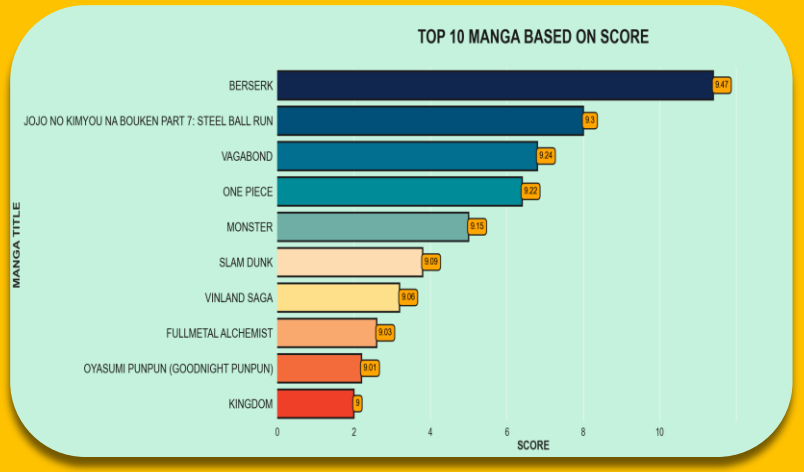
Abnormal Values and Outliers

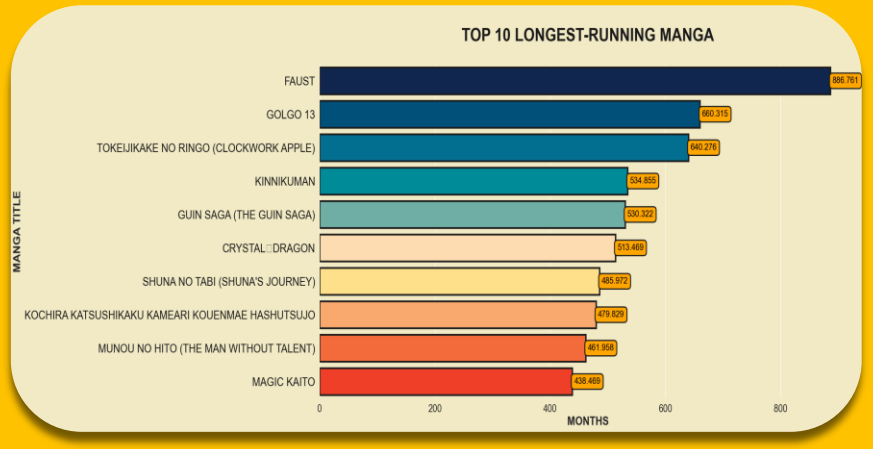
- All columns exhibit reasonable value ranges, and no abnormalities are found.
- One row with an unreasonable value in the time-related columns is dropped.
- Outliers beyond 1.5 IQR are retained due to the wide column distribution, falling within reasonable value ranges.

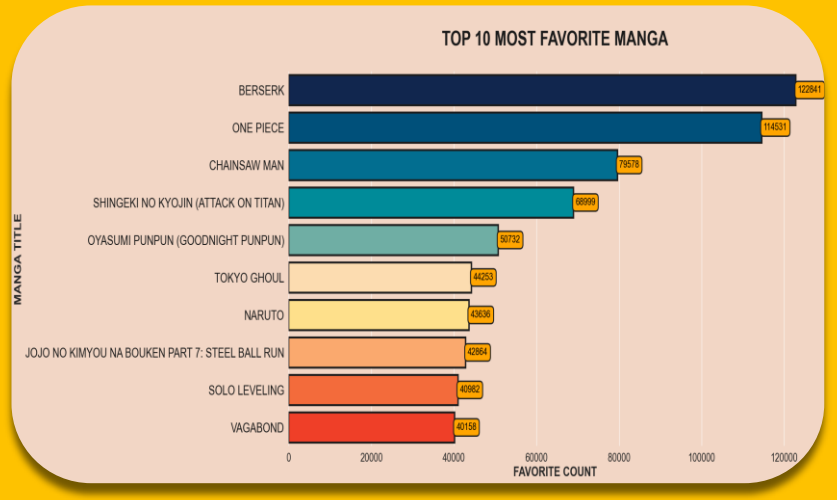


II. Data Exploration and Preprocessing

- Visualizations include
  - Top 10 manga series with the highest score
  - Top 10 longest-running manga series
  - Top 10 most favorite manga series.





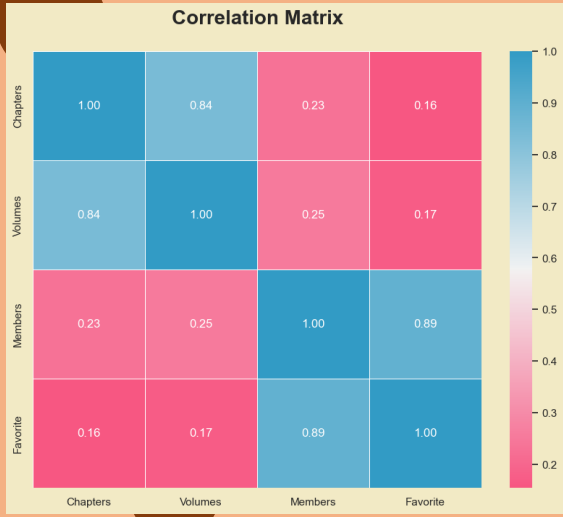


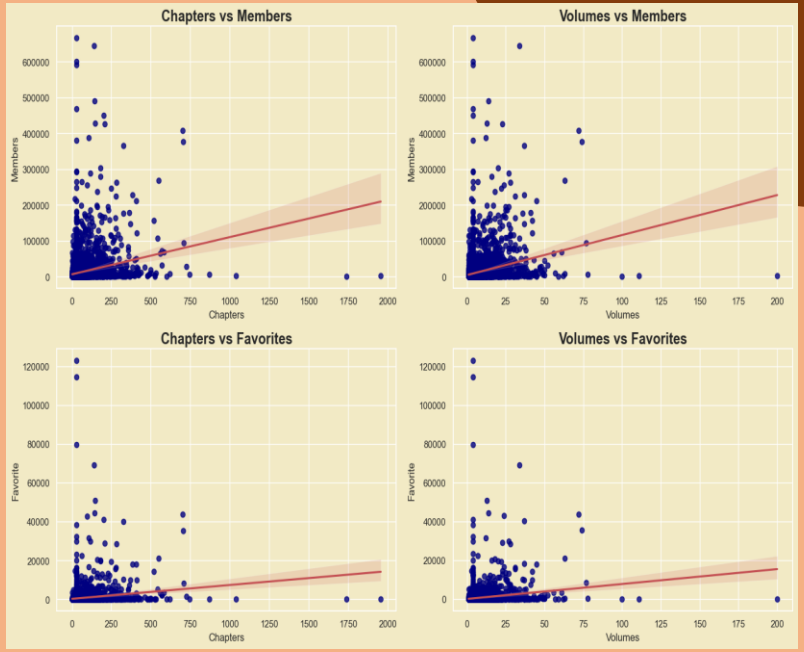
Asking Question and Analyzing

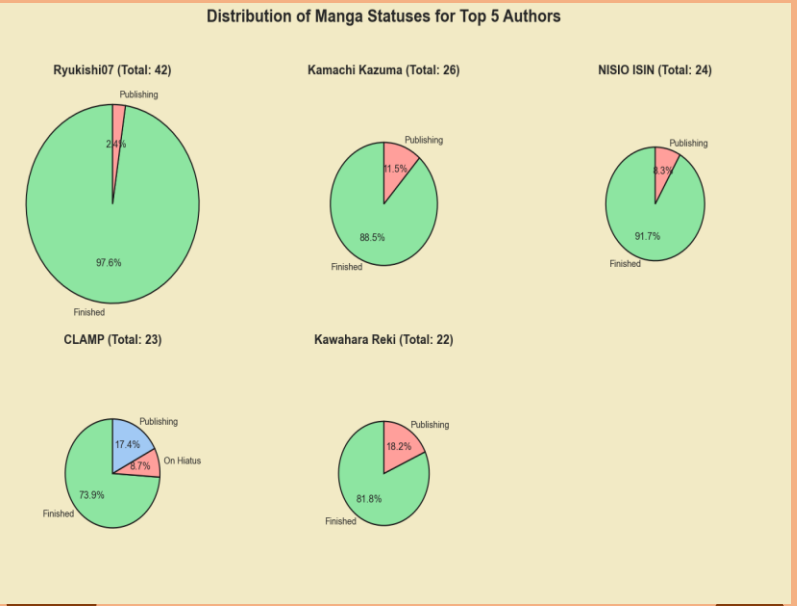
III. Asking Question and Analyzing

- **Question 01**: How does the number of **chapters** and **volumes** relate to the number of readers (**members**) and their engagement (**favorites**)?

- **Question 02**: How does the visualization of manga status distribution (finished, published, on hiatus) for the top 5 authors provide insights into the characteristics and working patterns of each author?



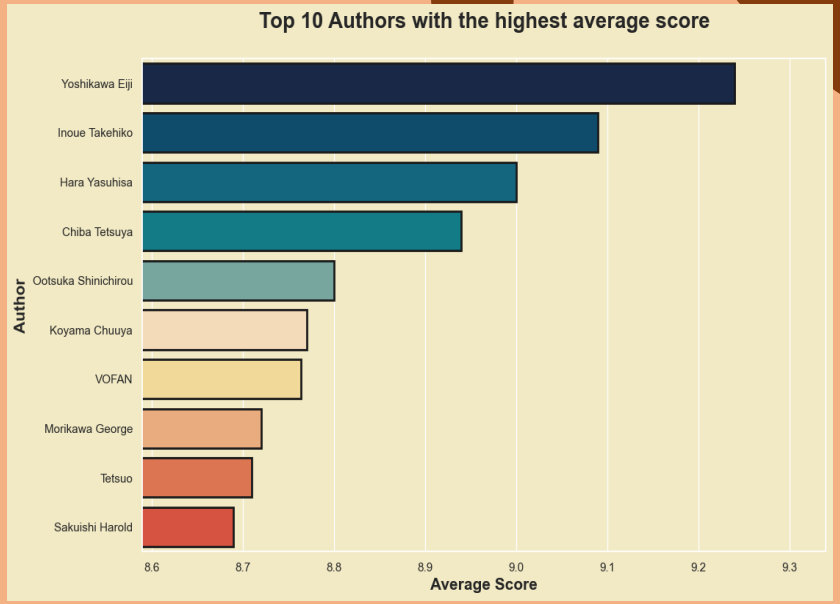


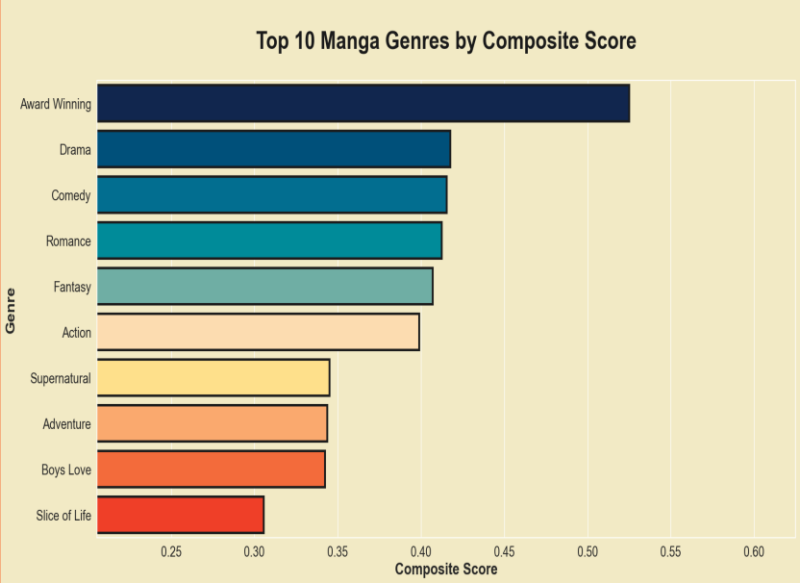


III. Asking Question and Analyzing

- **Question 03:** If an author's rating is determined by the average scores of their written manga, what are the top 10 authors based on this scoring metric?

- **Question 04**: What are the top manga genres determined by a composite score that accounts for the average favorite count, the number of Mangas in each genre, their average score, and their popularity?

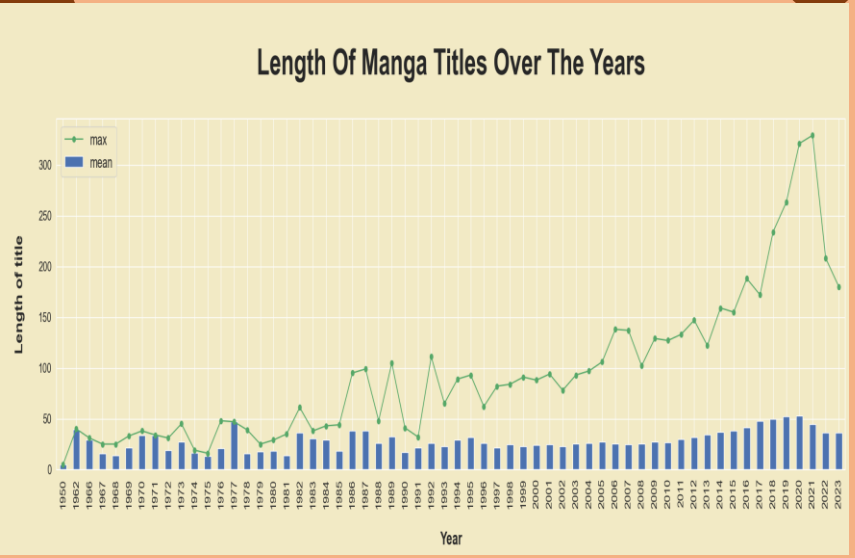


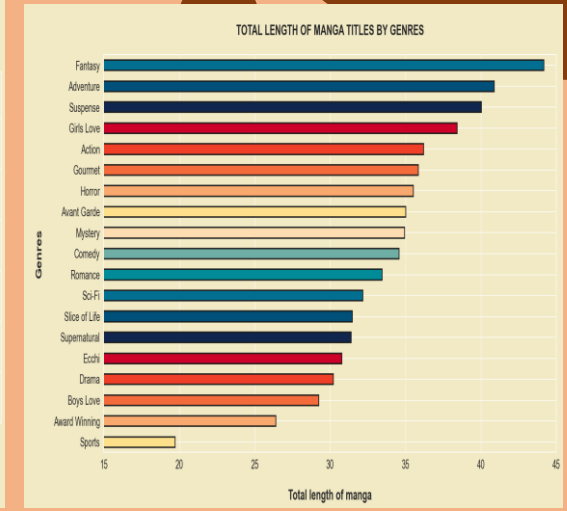


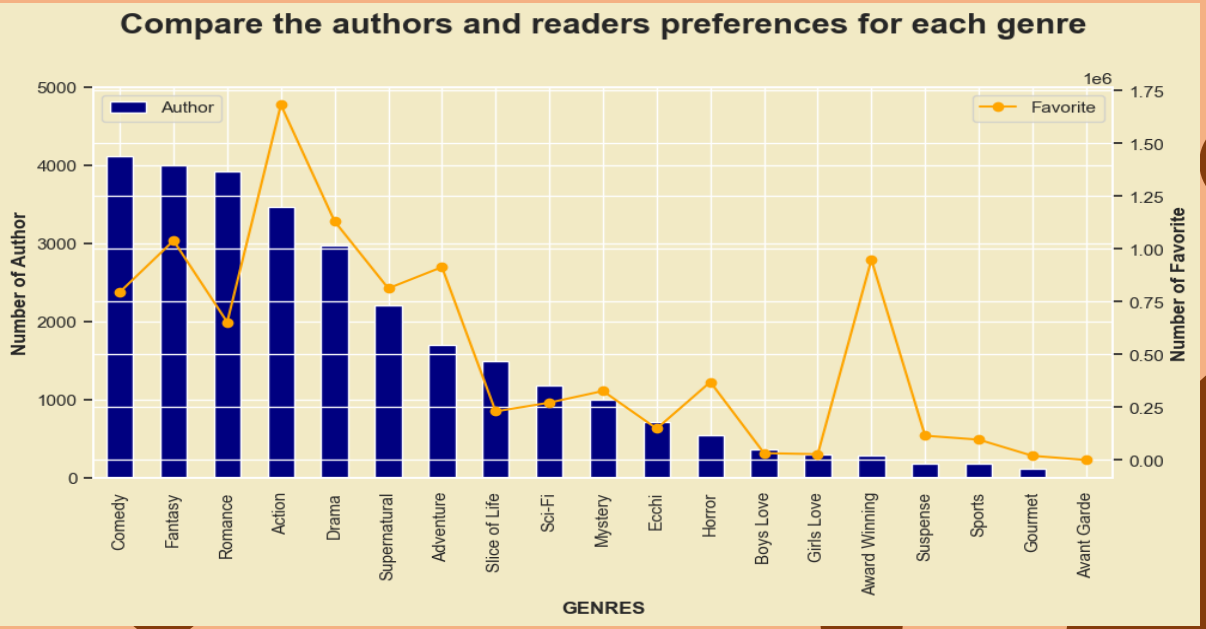
III. Asking Question and Analyzing

- **Question 05**:
  - How has the length of manga titles changed over the years?
  - What genres of manga will have long titles?

- **Question 06**:
  - What is the relationship between the genres most authors (Mangaka) write about and the genres that readers love?
  - Are genres that readers like also liked by authors?
  - Is there any difference between the author's and reader's choices for each genre?



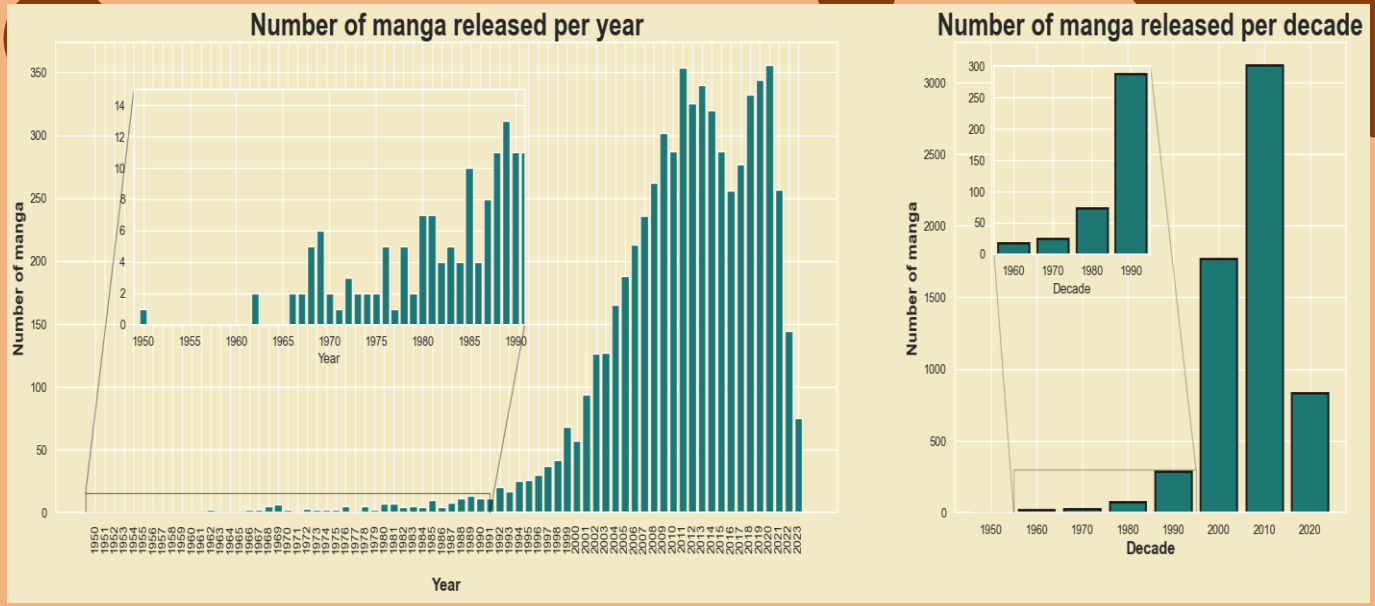


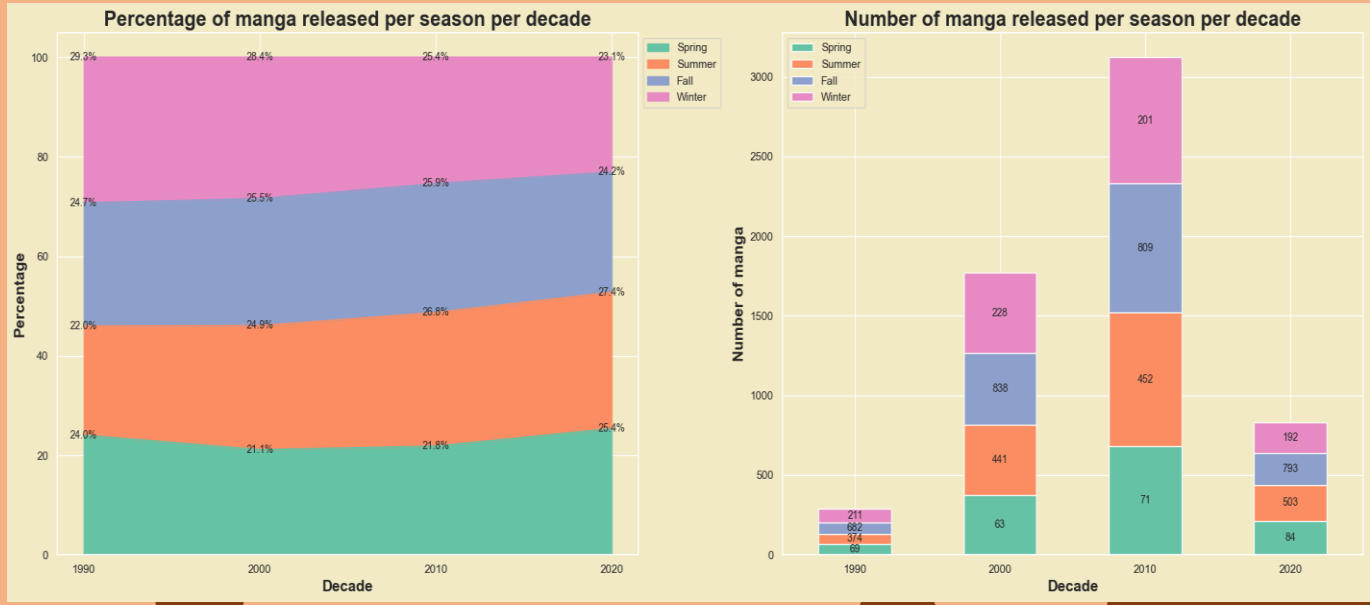


III. Asking Question and Analyzing

- **Question 07**: How has the number of manga releases changed over time from the past to the present?

- **Question 08**: How does the number of manga released vary across different seasons throughout the year? Are there discernible trends in the distribution of manga releases by season?

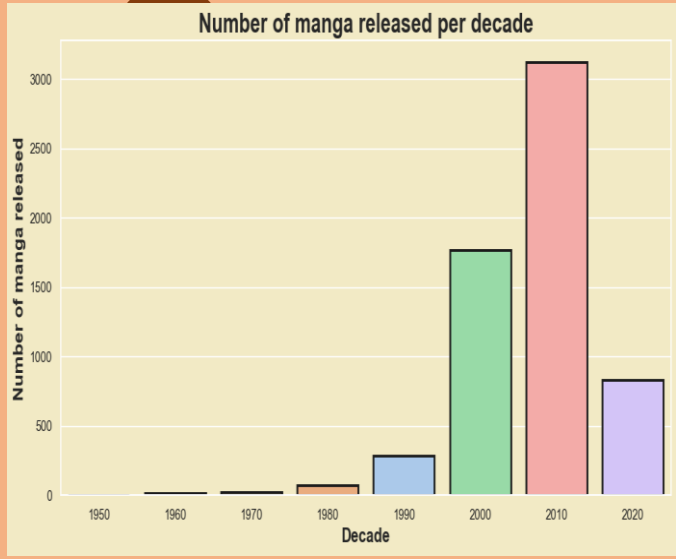


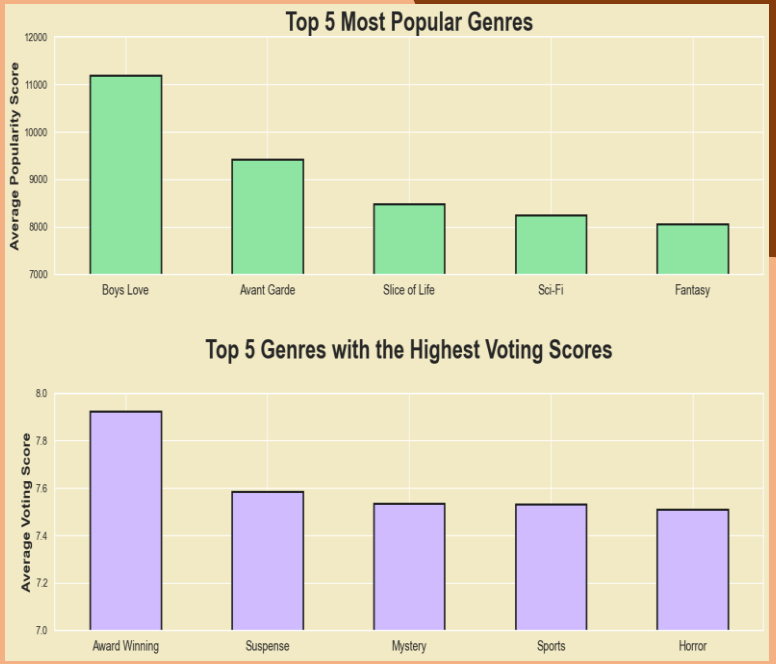


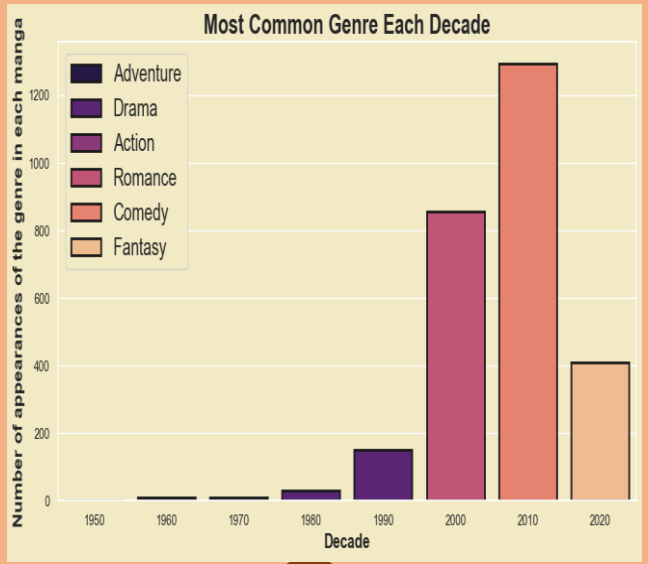
III. Asking Question and Analyzing

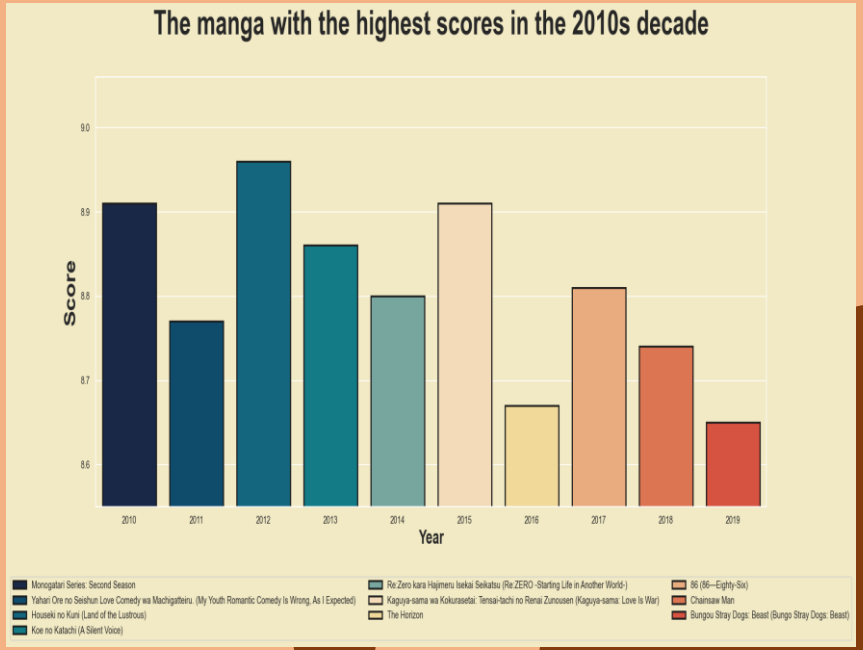
- **Question 09**: How are the Genres and preferences of readers?

- **Question 10**: Which decade saw the boom of mangas, the most highly acclaimed series of that decade, and the most popular Genres in each decade?

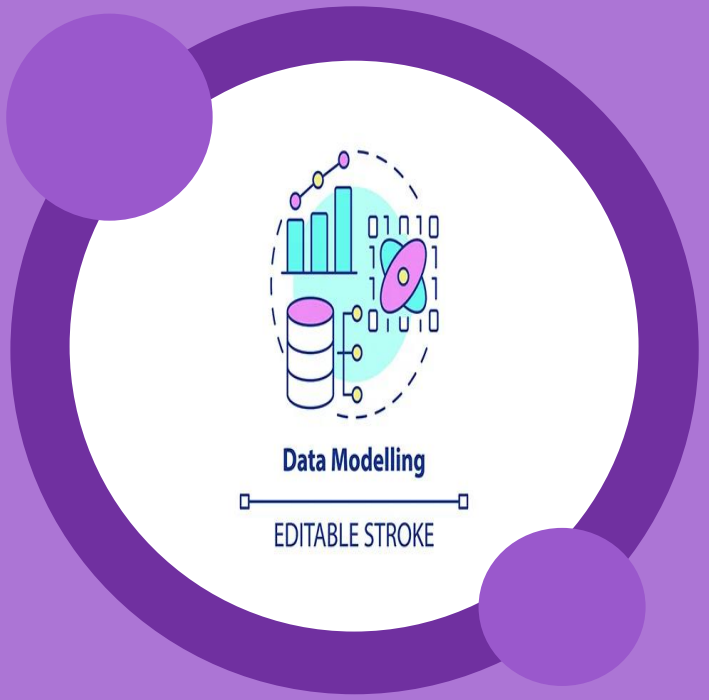








Data Modeling

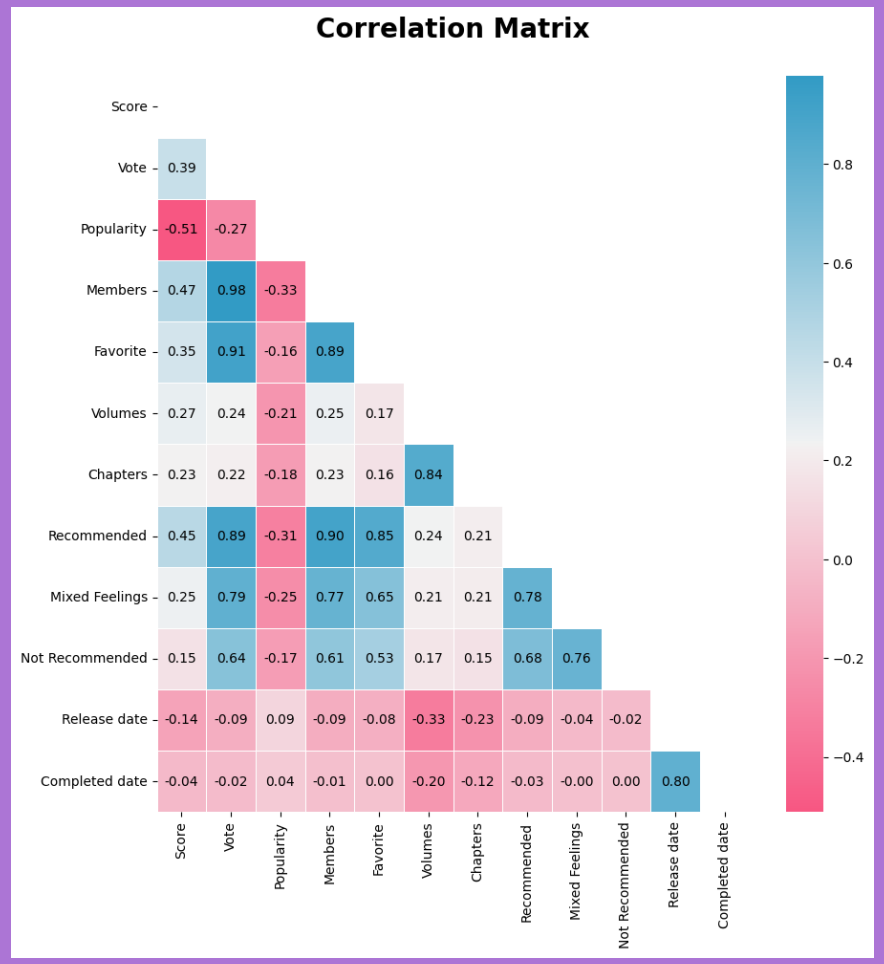


Problem Statement:

Predict the rating score of a manga based on its features

Purpose of solving the problem

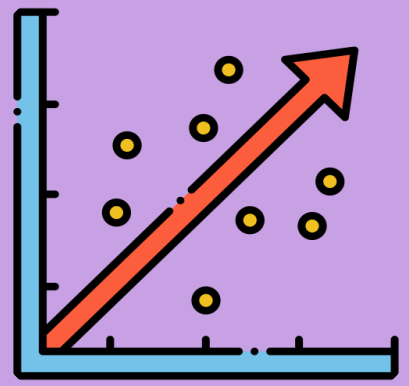
- **Enhanced decision-making**: Predict ratings to guide readers to hidden gems they'll love, boosting engagement and satisfaction
- **Market Insights**: Gain market insights from predicted ratings, informing content creation, marketing, and audience targeting for optimal success.



IV. Data Modeling

Data Preparation

- ❑ Omit by meaning: Ranked, Title, Author, Status,
- ❑ Omit by Correlation matrix: Not Recommended, Release Date, Completion Date.
- ❑ Transform categorical features into numerical equivalents.
- ❑ Proceed with the following steps:
  - ▪ Identify the target and feature variables.
  - ▪ Divide the dataset into training, validation, and test sets at an 80:10:10 ratio.
  - ▪ Apply MinMaxScaler to scale the data.

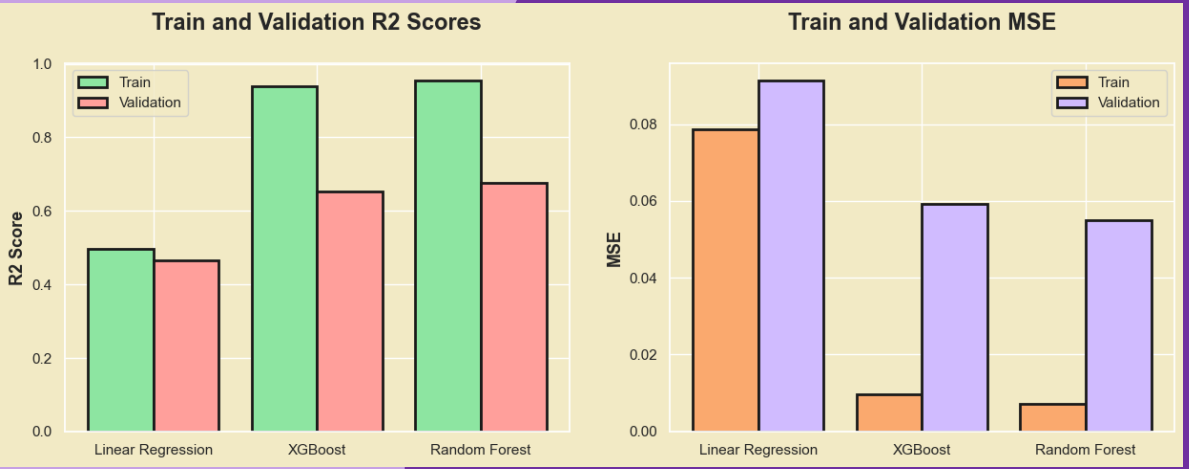


IV. Data Modeling

Create, train and test models

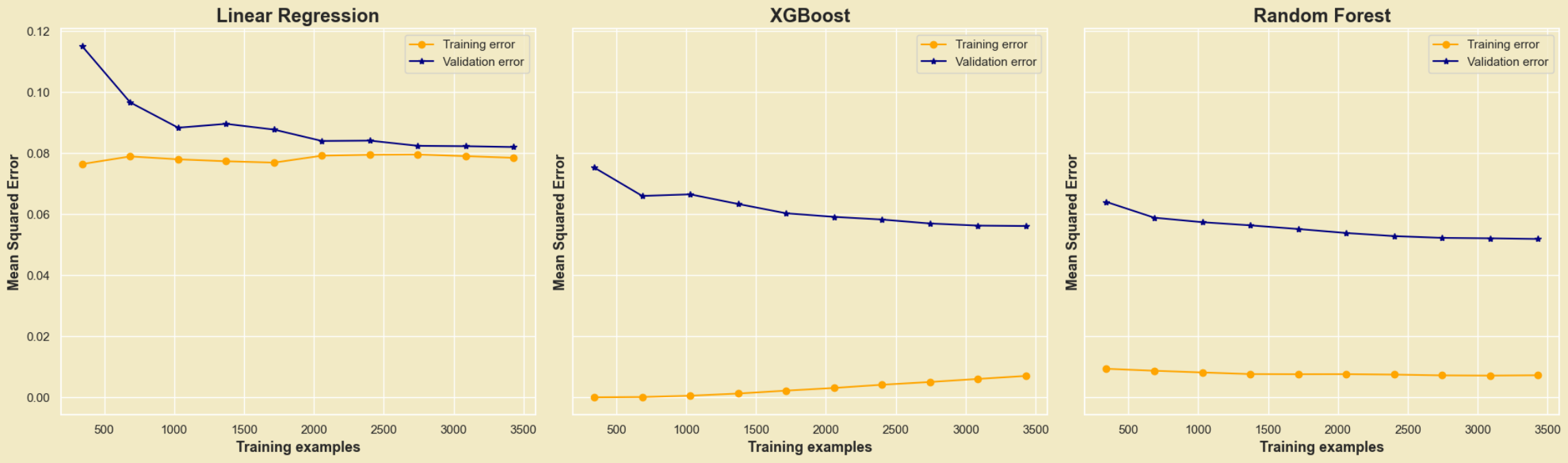
- Employing machine learning models:
  - Linear Regression
  - XGBoost
  - Random Forest
- Metrics:
  - Mean Squared Error (MSE)
  - R2 Score.

IV. Data Modeling



Train models on training data

- Both MSE and R2 score will be stored to compare the p
  the models
- Visualize the learning process of these models by havir
  gradually over multiple training runs with increasing trai



IV. Data Modeling

```
Linear Regression: 0.4525
fit_intercept: True
n_jobs: 1
positive: False

XGBoost: 0.6482
learning_rate: 0.180875564077442
max_depth: 2
n_estimators: 85

Random Forest: 0.6457
max_depth: 8
max_features: 0.8622514477983243
n_estimators: 100
```

We will fine-tune the models on the validation data using Bayesian search to find the best hyperparameters for the models. The hyperparameters we will tune are:

- Linear Regression
  - fit_intercept
  - positive
  - n_jobs
- XGBoost
  - n_estimators
  - max_depth
  - learning_rate
- Random Forest
  - n_estimators
  - max_depth
  - max_features

```
param_lr = {
    'fit_intercept': [True, False],
    'positive': [True, False],
    'n_jobs': [-1, 1],
}
param_xgb = {
    'n_estimators': (10, 100),
    'max_depth': (1, 10),
    'learning_rate': (0.01, 1.0, 'log-unifo
}
params_rf = {
    'n_estimators': (10, 100),
    'max_depth': (1, 10),
    'max_features': (0.1, 1.0, 'uniform'),
}
```

IV. Data Modeling

Retrain models on training + validation set and evaluate them on test data

10 random samples from the test set and compare the actual Score with the predicted Score of the models

	Actual	Linear Regression	XGBoost	Random Forest
3023	7.32	7.447534	7.312806	7.238938
1740	7.56	7.671856	7.508104	7.364712
3628	7.23	7.286110	7.555274	7.590559
1039	7.77	7.901642	7.816235	7.831644
1370	7.66	7.337381	7.305583	7.300438
5425	7.00	7.317434	7.288024	7.255839
5792	6.95	7.240576	7.091309	7.083795
1163	7.72	7.239668	7.308531	7.346511
5689	6.96	7.357609	7.228117	7.209906
2794	7.37	7.131445	7.142243	7.179317

Group 6

Thanks For Watching!

Any Question?