Національний технічний університет України «КПІ ім. Ігоря Сікорського»
Факультет Інформатики та Обчислювальної Техніки

Кафедра інформаційних систем та технологій

Лабораторна робота №7

з дисципліни «Вступ до технології Data Science»

на тему

«РОЗРОБКА ПРОГРАМНОГО МОДУЛЯ ПРОГНОЗУВАННЯ ДИНАМІКИ ЗМІНИ ПОКАЗНИКІВ ЕФЕКТИВНОСТІ ТОРГІВЕЛЬНИХ КОМПАНІЙ (міні проекти в галузі аналізу даних для завдань електронної комерції)»

Виконала:
студентка групи ІС-12
Павлова Софія

Перевірив:
Баран Д. Р.

Київ – 2023

1. Постановка задачі

Мета роботи:

Дослідити виявити та узагальнити особливості реалізації проектного практикуму в галузі аналізу стохастичних рядів, як характеристика показників ефективності діяльності торгівельних компаній.

Завдання II рівня:

	Розробити програмний скрипт, що реалізує аналіз даних, самостійно обраних процесів. Обов'язковою вимогою є аналіз множини процесів, поданих часовими рядами із різними властивостями.

Рисунок 1 – Технічні умови завдання

2. Виконання

Постановка задачі

Сформулюємо задачу з власного досвіду для аналізу множини процесів з різними властивостями:

Задача:

Розробити програму, яка допоможе бізнес аналітику сайту перегляду аніме визначати критерії актуальності завантажених аніме з метою збільшення прибутку.

Датасет:

Для цього завдання використаємо набір даних Anime Recommendations Database. Дані являють собою записи про аніме тайтли, що можуть являти собою часові ряди.

Вхідні дані матимуть наступну структуру:

anime_id – унікальний ідентифікатор.

name – назва аніме.

genre – жанри аніме, перелічені через кому.

type – тип (мультфільм, ТВ, тощо).

episodes – кількість епізодів.

rating – рейтинг по шкалі від 0 до 10.

members – кількість людей, зацікавлених у цьому аніме.

anime_id	name	genre	type	episodes	rating	members
32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
5114	Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Military, Shounen	TV	64	9.26	793665
28977	Gintama°	Action, Comedy, Historical, Parody, Samurai, Sci-Fi, Shounen	TV	51	9.25	114262
9253	Steins;Gate	Sci-Fi, Thriller	TV	24	9.17	673572

Рисунок 2 – Вигляд датасету

Таким чином задача зводиться до розробки програмного скрипта, що реалізує:

1. Парсинг файлу параметрів: anime.csv;
2. Попередній аналіз даних;
3. Визначення показників ефективності – рейтингу та кількості фанатів;
4. Лінійний аналіз даних.
5. Статистичний аналіз даних.

На основі отриманих даних бізнес аналітик зможе зробити якісні висновки щодо аніме тенденцій і спланувати новий контент таким чином, щоб забезпечити максимальну кількість переглядів на сайті.

Парсинг файлу

Завантажимо дані з файлу anime.csv.

Лістинг коду:

```
# 1. Парсинг файлу  ----------------------------

# Вхідні дані

df = pd.read_csv('anime.csv')
print('\nВхідні дані')
print(df)
df.to_excel('anime.xlsx', index=False)
```

Результат:

```
Вхідні дані
       anime_id  ... members
0         32281  ...  200630
1          5114  ...  793665
2         28977  ...  114262
3          9253  ...  673572
4          9969  ...  151266
...         ...  ...     ...
12289      9316  ...     211
12290      5543  ...     183
12291      5621  ...     219
12292      6133  ...     175
12293     26081  ...     142

[12294 rows x 7 columns]
```

Рисунок 3 – Завантажений датасет

Попередній аналіз даних

Виконаємо первинний аналіз даних.

Дані **неоднорідні** за джерелом походження (аніме відносяться до різних жанрів);

Дані **не послідовні** (не відсортовані по рейтингу чи кількості фанатів).

Так, як більшість аніме має декілька жанрів, записаних через кому, виділимо всі унікальні жанри, що трапляються в нашому датасеті. У подальшому використаємо ці дані для сегментації по жанрах.

Лістинг коду:

```
# Представлені жанри

# Розділення рядків, що містять кілька жанрів
genres_series = df['genre'].str.split(',').explode()
# Вилучення унікальних значень жанрів з виправленням пробілів
unique_genres = genres_series.str.strip().unique()
# Виведення масиву унікальних значень жанрів
print('\nЖанри')
print(unique_genres)
print(len(unique_genres))
```

Результат:

```
Жанри
['Drama' 'Romance' 'School' 'Supernatural' 'Action' 'Adventure' 'Fantasy'
 'Magic' 'Military' 'Shounen' 'Comedy' 'Historical' 'Parody' 'Samurai'
 'Sci-Fi' 'Thriller' 'Sports' 'Super Power' 'Space' 'Slice of Life'
 'Mecha' 'Music' 'Mystery' 'Seinen' 'Martial Arts' 'Vampire' 'Shoujo'
 'Horror' 'Police' 'Psychological' 'Demons' 'Ecchi' 'Josei' 'Shounen Ai'
 'Game' 'Dementia' 'Harem' 'Cars' 'Kids' 'Shoujo Ai' nan 'Hentai' 'Yaoi'
 'Yuri']
44
```

Рисунок 4 – Унікальні жанри датасету

Визначення показників ефективності

Оскільки перед нами стоїть задача покращити перегляди на сайті, очевидно, що за критерії ефективності буде взято рейтинг аніме – «**rating**» та кількість його фанатів – «**members**».

```
Показник ефективності - Rating
0        9.37
1        9.26
2        9.25
3        9.17
4        9.16
         ...
12059    4.15
12060    4.28
12061    4.88
12062    4.98
12063    5.46
Name: rating, Length: 12064, dtype: float64
<class 'float'>
```

Рисунок 5 – Показник ефективності рейтинг

```
Показник ефективності - Members
0        200630
1        793665
2        114262
3        673572
4        151266
          ...
12059       211
12060       183
12061       219
12062       175
12063       142
Name: members, Length: 12064, dtype: int64
<class 'float'>
```

Рисунок 6 – Показник ефективності кількість фанатів

Для узагальненої характеристики створимо також змішаний критерій – «**combined_score**». Який складатиме середнє арифметичне рейтингу та кількості фанатів для кожного окремого аніме.

Для того, аби цей критерій мав сенс, відмасштабуємо значення критерію кількості фанатів в межах від 0 до 10, а потім уже знайдемо середнє арифметичне.

```
Маштабований показний ефективності - Members [0, 10]
0        1.978667
1        7.827686
2        1.126831
3        6.643226
4        1.491797
           ...
12059    0.001963
12060    0.001687
12061    0.002042
12062    0.001608
12063    0.001282
Name: scaled_members, Length: 12064, dtype: float64
<class 'float'>
```

Рисунок 7 – Масштабований показник ефективності кількість фанатів

```
Показник ефективності - Combined score
0        5.674333
1        8.543843
2        5.188416
3        7.906613
4        5.325898
           ...
12059    2.075981
12060    2.140843
12061    2.441021
12062    2.490804
12063    2.730641
Name: combined score, Length: 12064, dtype: float64
<class 'float'>
```

Рисунок 8 – Узагальнений показник ефективності

Лінійний аналіз даних

У рамках лінійного аналізу даних, виведемо графіки двох типів для кожного показника ефективності: графік загального представлення, де зобразимо дані на осі абсцис так, як вони з’являються в датасеті, та графік представлення за жанрами, де зобразимо дані, сегментовані по жанрах.

Лістинг коду:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math as mt
import warnings
# Вимкнути DeprecationWarning
warnings.filterwarnings("ignore", category=DeprecationWarning)

#-------------------- ЕТАП І ЛІНІЙНИЙ АНАЛІЗ -----------------------------

# Розділення вхідного масиву на кластери за жанрами
def split_anime_by_genre(dataframe, unique_genres):
    genre_clusters = {genre: dataframe[dataframe['genre'].str.contains(genre, na=False, case=False)] for genre in unique_genres if not pd.isna(genre)}
    return genre_clusters

# Графік розділеного масиву на кластери за жанрами
def plot_average_profit(genre_clusters, profit, num_highest=3, num_lowest=3):
    average_profits = {genre: data[profit].mean() for genre, data in genre_clusters.items()}

    genres = list(average_profits.keys())
    profits = list(average_profits.values())
```

```
    # Знайти найбільші і найменші значення
    max_profits_indices = np.argsort(profits)[-num_highest:]
    min_profits_indices = np.argsort(profits)[:num_lowest]

    # Створити список кольорів для підписів осі x
    label_colors = ['green' if i in max_profits_indices else 'red' if i in
min_profits_indices else 'black' for i in range(len(profits))]

    # Створити список кольорів для стовпчиків
    bar_colors = [
        'skyblue' if i not in max_profits_indices and i not in min_profits_indices else
'green' if i in max_profits_indices else 'red'
        for i in range(len(profits))]

    # Вивести стовпчасту діаграму зі спеціальними кольорами
    plt.bar(genres, profits, color=bar_colors)
    plt.xlabel('Genre')

    # Встановити колір підписів на осі x
    for tick_label, color in zip(plt.gca().get_xticklabels(), label_colors):
        plt.setp(tick_label, color=color)

    plt.xticks(rotation=45, ha='right')
    plt.ylabel(f'Average {profit}')
    plt.title(f'Average {profit} by Genre')
    plt.show()

# Результати показників ефективності
def profit_results(df, unique_genres, profit):
    print(f'\nПоказник ефективності - {profit.capitalize()}')
    print(df[profit])
    print(type(float(df[profit][0])))
    # Графік загального представлення
    plt.title(profit.capitalize())
    df[profit].plot()
    plt.show()
    # Графік представлення за жанрами
    genre_clusters = split_anime_by_genre(df, list(unique_genres))
    plot_average_profit(genre_clusters, profit, num_highest=5, num_lowest=5)

    return genre_clusters
```

Таке представлення даних двома способами дозволить краще зрозуміти взаємозв’язки між жанрами аніме та кількістю переглядів сайту.

Розглянемо спершу **критерій рейтингу**.

Лістинг коду:

```
# 4. Лінійний аналіз даних  ----------------------------

# Рейтинг
profit = 'rating'
# Виведення результатів
genre_clusters_rating = profit_results(df, unique_genres, profit)
```

Результат:

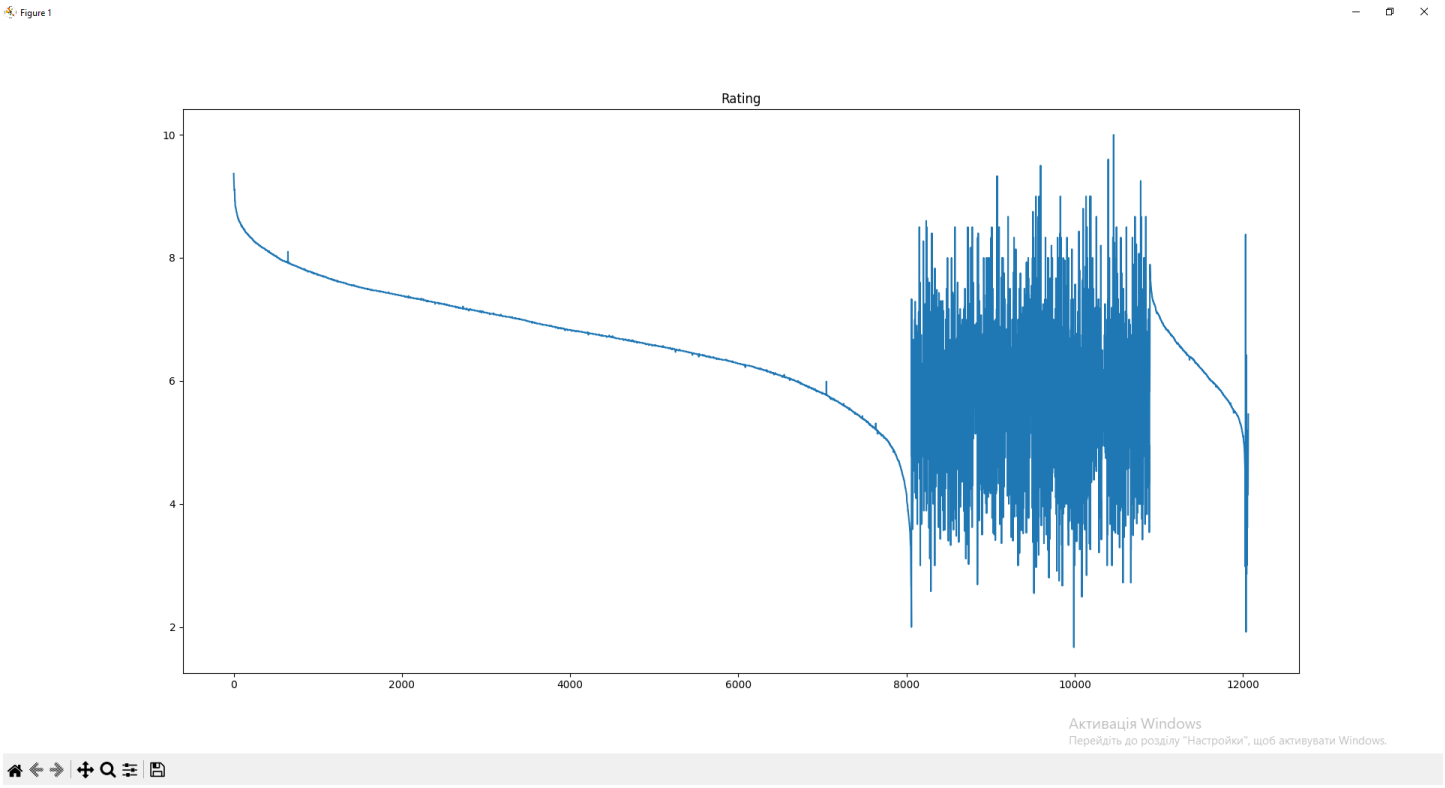


Рисунок 9 – Загальне представлення рейтингу

Завдяки загальному представленню рейтингу переконуємось в непослідовності даних за критерієм рейтингу. З даних важко виділити чіткий тренд, але найбільш вони подібні до **показникової функції** $y = a^x, x > 1$.

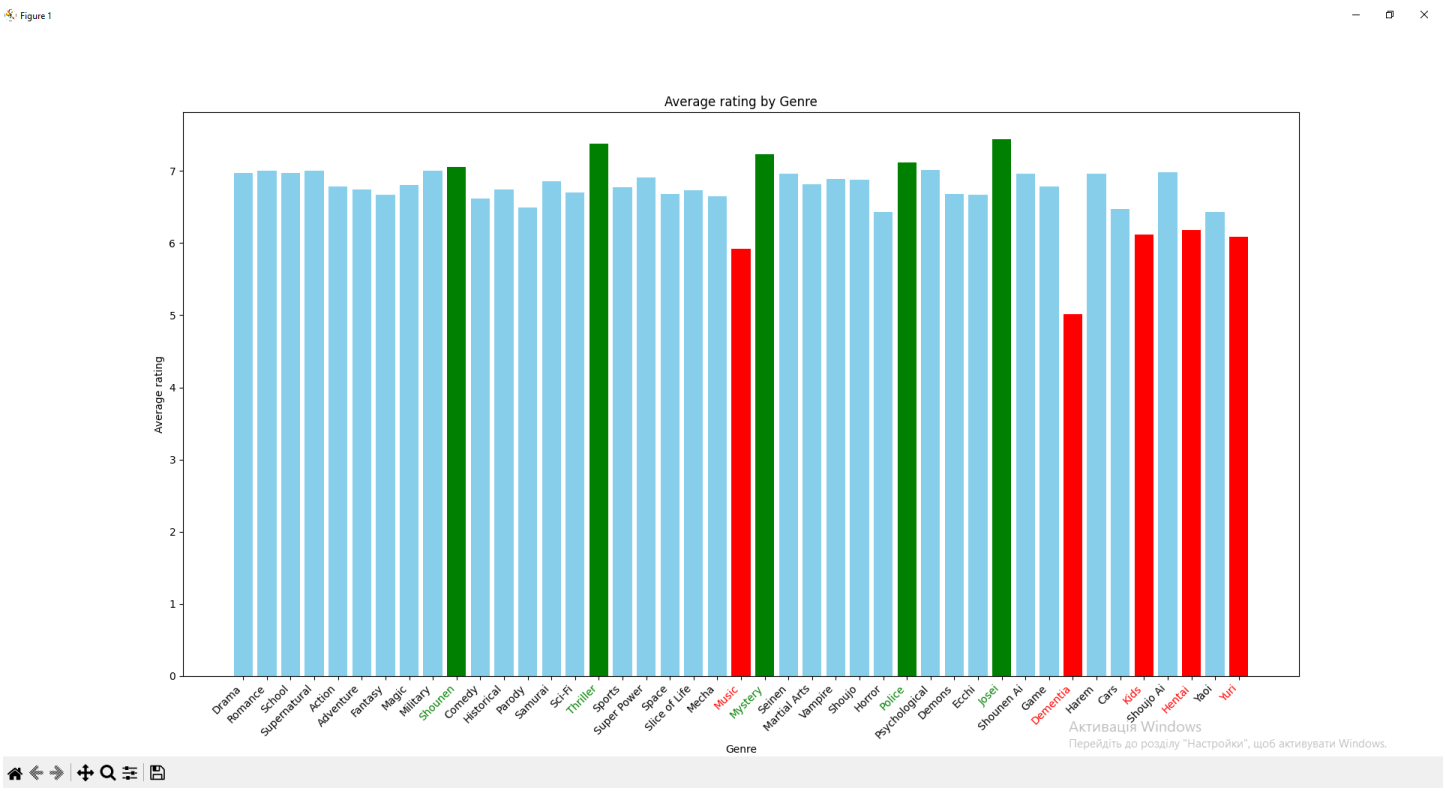


Рисунок 10 – Представлення рейтингу за жанрами

Бачимо, що найбільший рейтинг мають жанри*: Shounen, Thriller, Mystery, Police ma Josey*.

Найменший рейтинг: *Music, Demencia, Kids, Hentai, Yuri*.

Отже можемо зробити висновок, що на задоволеність користувачів сайту позитивно впливають найуспішніші за показником рейтингу жанри і таких аніме потрібно завантажувати на сайт більше.

Натомість жанри з малим показником рейтингу навряд чи покращують статистику перегляду сайту, тому х можна завантажувати менше або в певній мірі від них відмовитись.

З рисунка видно, що сегментовані по жанрах, дані мають **лінійний закон розподілу**.

Розглянемо **критерій кількості фанатів**.

Лістинг коду:

```
# Фан база
profit = 'members'
# Виведення результатів
genre_clusters_members = profit_results(df, unique_genres, profit)
```

Результат:

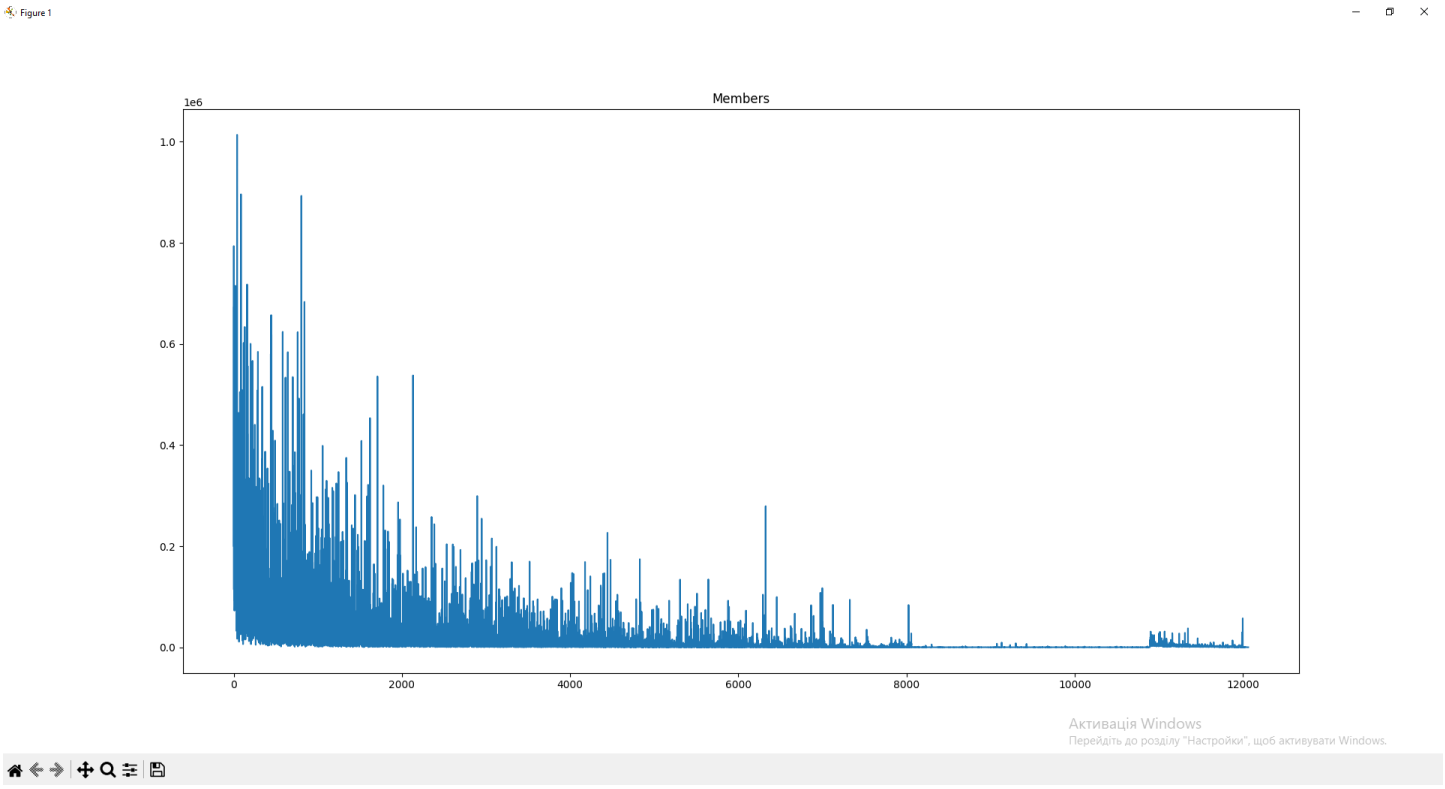


Рисунок 11 – Загальне представлення кількості фанатів

На перший погляд дані містять аномалії, але згладжування «аномальних» значень призведе до сильного зменшення середніх значень при сегментації даних за жанрами. Тому від усунення так званих викидів утримаємось.

З даних можемо зробити припущення про **експоненційний закон розподілу**.

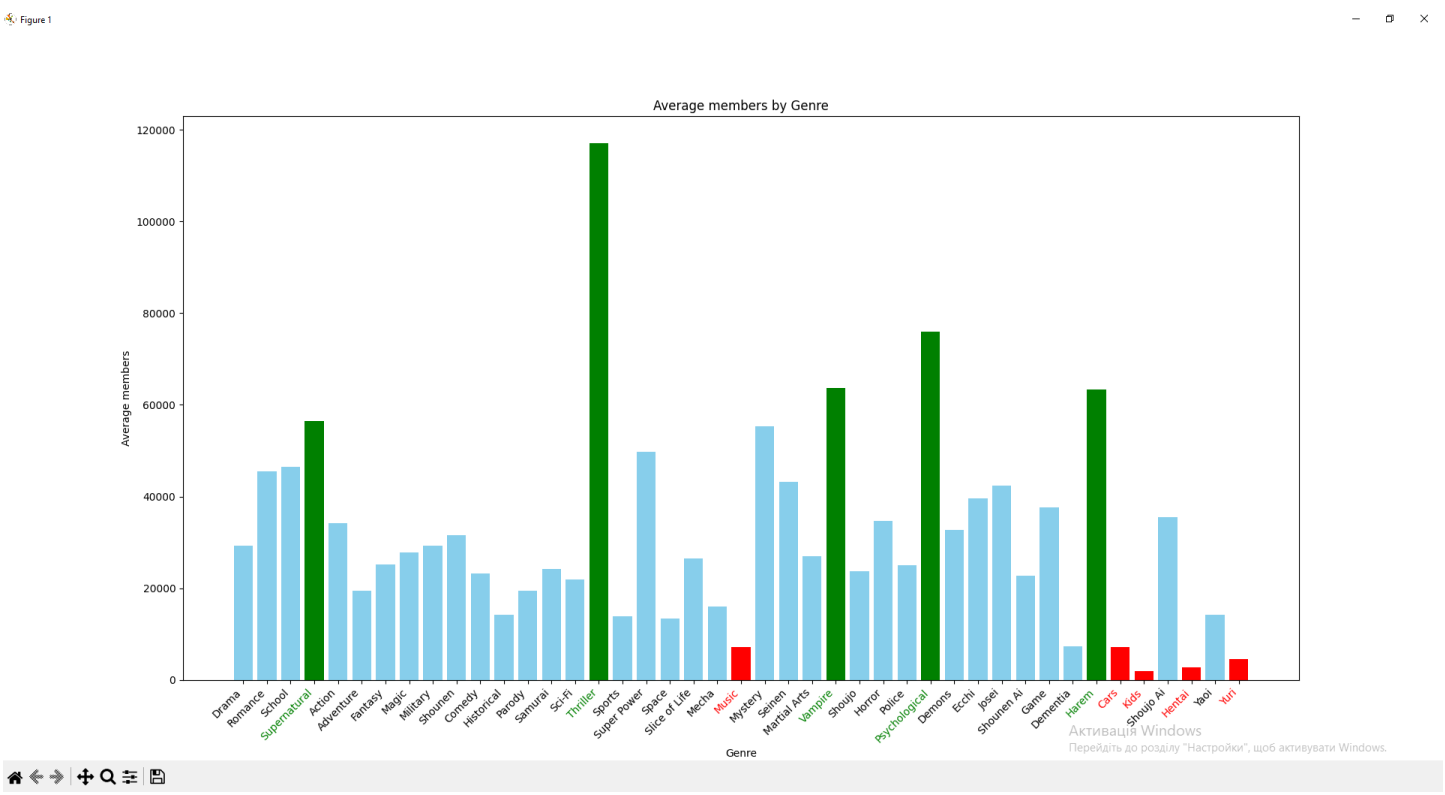


Рисунок 12 – Представлення кількості фанатів за жанрами

Бачимо, що найбільшу кількість фанатів мають жанри*: Supernatural, Thriller, Vampire, Psychological та Harem*.

Найменший рейтинг: *Music, Cars, Kids, Hentai, Yuri*.

Разюче на фоні популярних жанрів виділяється *Thriller*. У порівнянні з вибіркою за рейтингом, у вибірці за кількістю фанатів лідери та аутсайдери дещо змінились. Можемо зробити висновок, що на безпосередню кількість переглядів сайту позитивно впливають найуспішніші за показником кількості фанатів жанри і таких аніме потрібно завантажувати на сайт більше.

Натомість жанри з малою кількістю фанатів майже повністю дублюються з вибірки за рейтингом. Це говорить про те, що наявність таких жанрів не покращує статистику перегляду сайту. Але для отримання повних висновків варто ще розглянути вибірку за узагальненим показником ефективності.

З рисунка важко однозначно визначити закон розподілу даних, але усереднено можна зробити припущення про **лінійний** або **нормальний закон розподілу** даних.

Розглянемо **об’єднаний критерій**.

Лістинг коду:

```
# Об'єднаний показник ефективності
profit = 'combined score'
# Маштабування значень фан бази від 0 до 10
min_members = df['members'].min()
max_members = df['members'].max()
df['scaled_members'] = ((df['members'] - min_members) / (max_members - min_members)) * 10
print(f'\nМаштабований показний ефективності - Members [0, 10]')
print(df['scaled_members'])
print(type(float(df['scaled_members'][0])))
# Додавання стовпця зі змішаним показником
df[profit] = (df['rating'] + df['scaled_members']) / 2
# Виведення результатів
genre_clusters_combined = profit_results(df, unique_genres, profit)
```

Результат:

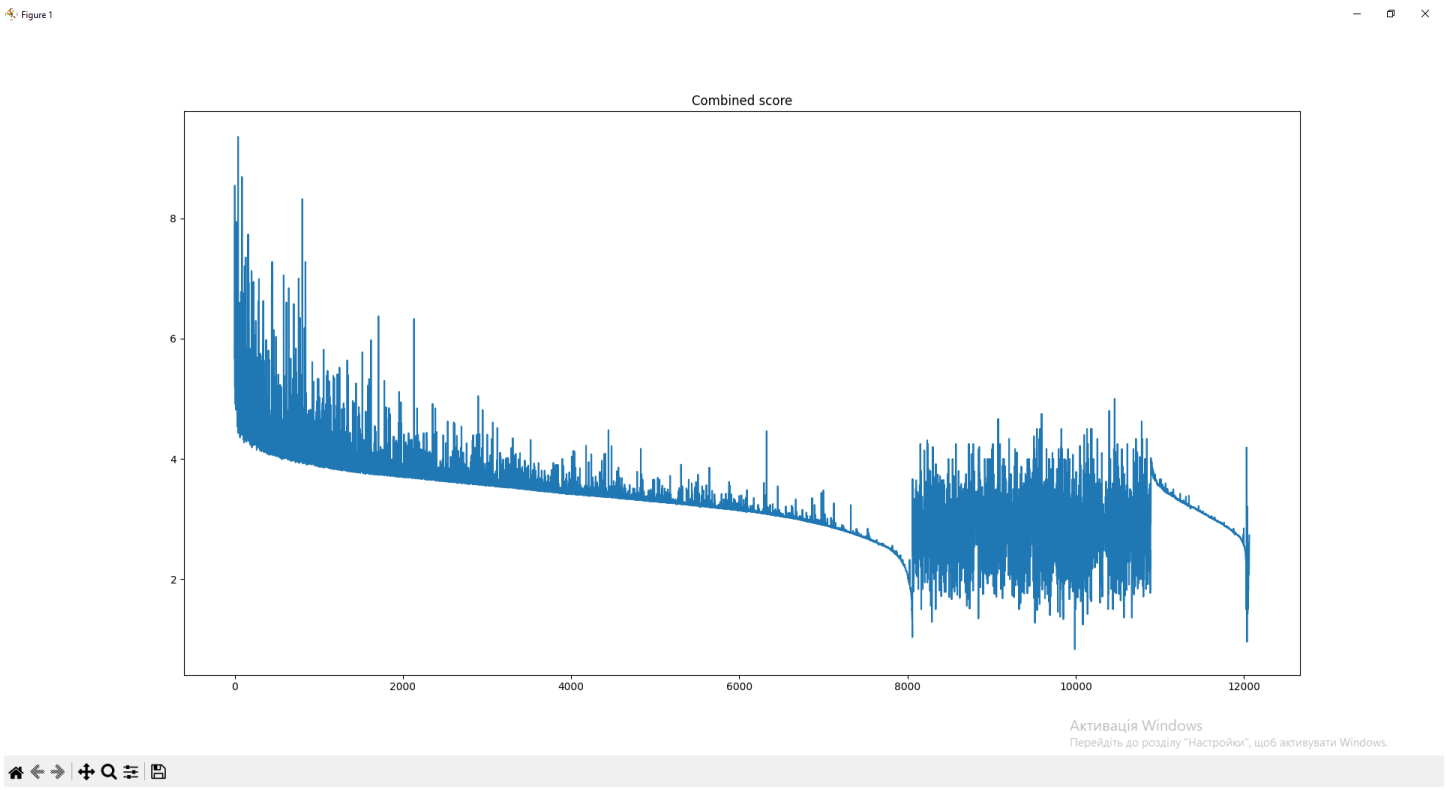


Рисунок 13 – Загальне представлення об’єднаного критерію

Бачимо, що графік візуально являє собою криву рейтингу та «зашумлені» піки кількості фанатів для кожного аніме.

З даних можемо зробити припущення про **лінійний** або **експоненційний закон розподілу**.

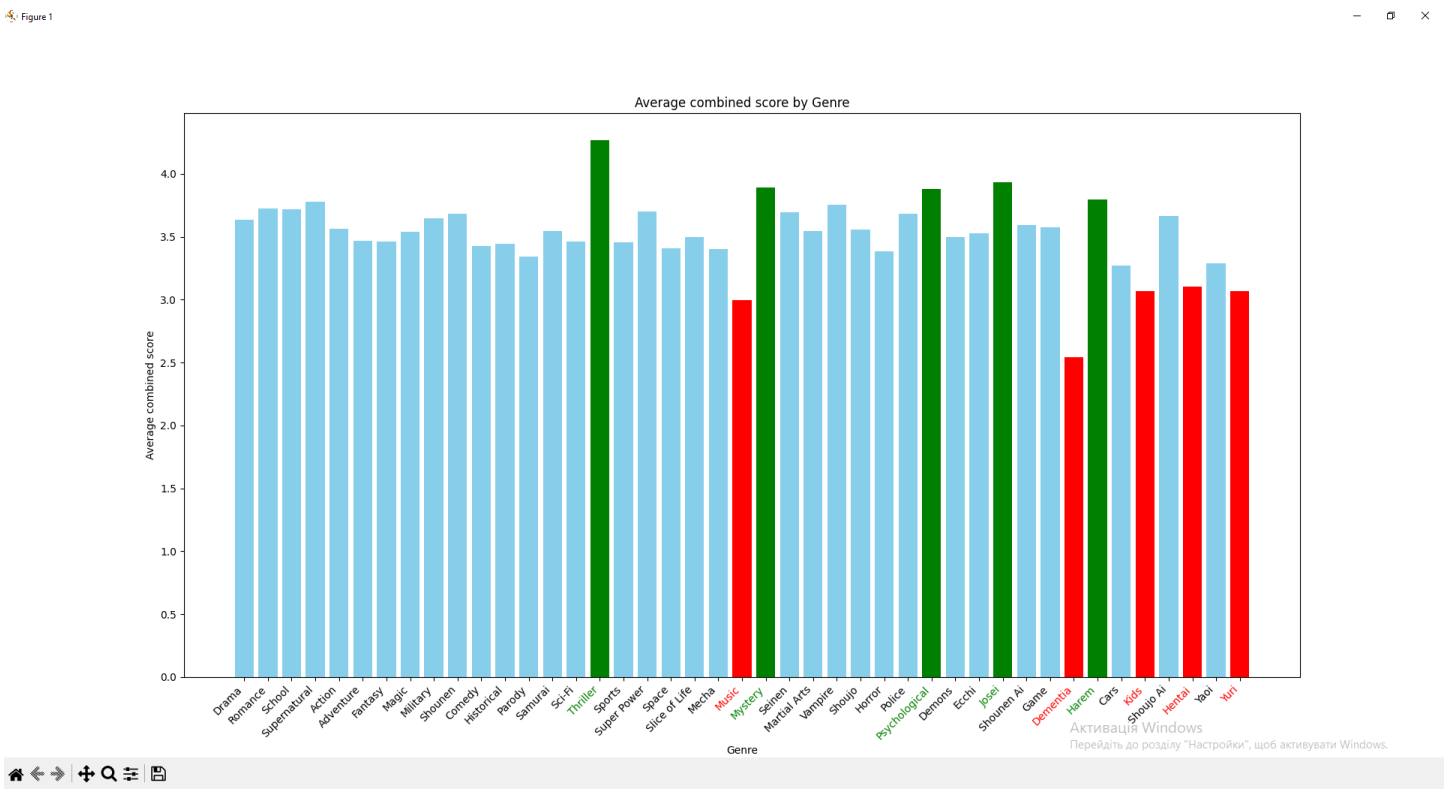


Рисунок 14 – Представлення об’єднаного критерію за жанрами

За об’єднаним критерієм переважають жанри*: Thriller, Mystery, Psychological, Josei та Harem*.

Найменший об’єднаний критерій мають жанри: *Music, Dementia, Kids, Hentai, Yuri*.

Найнижчі показники співпадають з вибіркою по рейтингу. Це може свідчити про те, що задоволеність переглянутим аніме впливає на подальше бажання дивитись аніме і «поповнювати ряди» фанатів нових тайтлів, тим самим збільшуючи перегляди сайту. Для більш детального аналізу цього припущення варто відслідкувати як змінювались рейтинги та перегляди в межах одного жанру протягом певного часу. Але в рамках даної лабораторної роботи за відсутності необхідного датасету ми цього робити не будемо.

З рисунка добре видно **лінійний закон розподілу** отриманих даних.

Статистичний аналіз даних

У рамках статистичного аналізу даних, знайдемо математичне сподівання, дисперсію та середньоквадратичне відхилення для вибірки та окремо для лінії тренду і виведемо відповідні графіки для кожного показника ефективності.

Лістинг коду:

```
#------------------- ЕТАП II СТАТИСТИЧНИЙ АНАЛІЗ -------------------------

# Функція МНК згладжування для визначення лінії тренду
def mnk(df):
    iter = len(df)
    df_1 = np.zeros((iter, 1))
    df_2 = np.ones((iter, 5))
    for i in range(iter):
        df_1[i, 0] = df[i]
        # df_1[i, 0] = df.iloc[i]
        df_2[i, 1] = float(i)
        df_2[i, 2] = float(i * i)
        df_2[i, 3] = float(i * i * i)
        df_2[i, 4] = float(i * i * i * i)
    df_2T = df_2.T
    df_2_df_2T = df_2T.dot(df_2)
    df_2_df_2TI = np.linalg.inv(df_2_df_2T)
    df_2_df_2TI_df_2T = df_2_df_2TI.dot(df_2T)
    C = df_2_df_2TI_df_2T.dot(df_1)
    df_rez = df_2.dot(C)

    return df_rez

def mnk_dict(genre_clusters, profit):
    data = {genre: data[profit].mean() for genre, data in genre_clusters.items()}

    # Створюємо датафрейм
    df = pd.DataFrame(list(data.items()), columns=['Genre', 'AverageProfit'])

    # Визначаємо МНК
    X = np.arange(len(df))
    y = df['AverageProfit'].values
    coefficients = np.polyfit(X, y, deg=1)
    trend_line = np.polyval(coefficients, X)

    # Отримуємо масив значень AverageRating у форматі trend_line
    df_2 = np.interp(np.arange(len(trend_line)), np.arange(len(df)), df['AverageProfit'])

    return df_2, trend_line

# Розрахунок статистичних характеристик вибірки
def stat_characteristics(df, text):
    # Статистичні характеристики вибірки з урахуванням тренду
    df_zglad = mnk_stat_characteristics(df)
    iter = len(df_zglad)
    df_1 = np.zeros((iter))
    for i in range(iter):
        df_1[i] = df[i] - df_zglad[i, 0]
```

```
    mat_spod = np.median(df_1)
    duspers = np.var(df_1)
    ser_kvad_vid = mt.sqrt(duspers)
    print('\n-----------------------------------------------------')
    print(text)
    print('-----------------------------------------------------')
    print('Матиматичне сподівання =', mat_spod)
    print('Дисперсія =', duspers)
    print('Середнє квадратичне відхилення =', ser_kvad_vid)
    # Графік МНК вибірки
    plt.title(text)
    plt.hist(df, bins=30, range=(df.min(), df.max()), facecolor="blue", alpha=0.5)
    plt.show()

    return

# МНК згладжуваннядля визначення статистичних характеристик
def mnk_stat_characteristics(df):
    iter = len(df)
    df_1 = np.zeros((iter, 1))
    df_2 = np.ones((iter, 4))
    # Формування структури вхідних матриць МНК
    for i in range(iter):
        # Формування матриці вхідних даних
        df_1[i, 0] = float(df[i])
        df_2[i, 1] = float(i)
        df_2[i, 2] = float(i * i)
        df_2[i, 3] = float(i * i * i)
    df_2T = df_2.T
    df_2_df_2T = df_2T.dot(df_2)
    df_2_df_2TI = np.linalg.inv(df_2_df_2T)
    df_2_df_2TI_df_2T = df_2_df_2TI.dot(df_2T)
    C = df_2_df_2TI_df_2T.dot(df_1)
    df_rez = df_2.dot(C)

    return df_rez

# Графік МНК для розділеного масиву на кластери за жанрами
def plot_mnk_by_genre(mnk_genre_zglad, genre_clusters, profit, num_highest=3,
num_lowest=3):
    average_profits = {genre: data[profit].mean() for genre, data in
genre_clusters.items()}

    genres = list(average_profits.keys())
    profits = list(average_profits.values())

    # Знайти найбільші і найменші значення
    max_profits_indices = np.argsort(profits)[-num_highest:]
    min_profits_indices = np.argsort(profits)[:num_lowest]

    # Створити список кольорів для підписів осі x
    label_colors = ['green' if i in max_profits_indices else 'red' if i in
min_profits_indices else 'black' for i in range(len(profits))]

    # Створити список кольорів для стовпчиків
    bar_colors = [
        'skyblue' if i not in max_profits_indices and i not in min_profits_indices else
'green' if i in max_profits_indices else 'red'
        for i in range(len(profits))]

    # Вивести стовпчасту діаграму зі спеціальними кольорами
    plt.bar(genres, profits, color=bar_colors)
    plt.plot(genres, mnk_genre_zglad)
```

```
    plt.plot(genres, mnk_genre_zglad)
    plt.xlabel('Genre')

    # Встановити колір підписів на осі x
    for tick_label, color in zip(plt.gca().get_xticklabels(), label_colors):
        plt.setp(tick_label, color=color)

    plt.xticks(rotation=45, ha='right')
    plt.ylabel(f'Average {profit}')
    plt.title(f'MNK by Genre - {profit.capitalize()}')
    plt.show()

# Знаходження тренду вибірки за показником ефективності
def mnk_zglad(profit, df, mnk_profit):
    stat_characteristics(df[profit], f'Statistical characteristics -
{profit.capitalize()}')
    stat_characteristics(mnk_profit, f'Statistical characteristics MNK -
{profit.capitalize()}')
    # Графік
    plt.title(f'MNK - {profit.capitalize()}')
    plt.plot(df[profit])
    plt.plot(mnk_profit)
    plt.show()

def mnk_zglad_dict(genre_clusters, profit):
    df_2, mnk_genre_zglad = mnk_dict(genre_clusters, profit)
    # Графік
    stat_characteristics(df_2, f'Statistical characteristics by Genre -
{profit.capitalize()}')
    stat_characteristics(mnk_genre_zglad, f'Statistical characteristics MNK by Genre -
{profit.capitalize()}')
    plot_mnk_by_genre(mnk_genre_zglad, genre_clusters, profit, num_highest=5,
num_lowest=5)

# Результати МНК згладжування
def mnk_results(profit, df):
    # Графік загального представлення
    mnk_profit = mnk(df[profit])
    mnk_zglad(profit, df, mnk_profit)
    # Графік представлення за жанрами
    genre_clusters = split_anime_by_genre(df, list(unique_genres))
    mnk_zglad_dict(genre_clusters, profit)
```

Спочатку знайдемо статистичні характеристики для вибірки **за рейтингом**.

Лістинг коду:

```
# 5. Статистичний аналіз даних  ----------------------------

# МНК оцінка

# Рейтинг
profit = 'rating'
mnk_results(profit, df)
```

Результат:

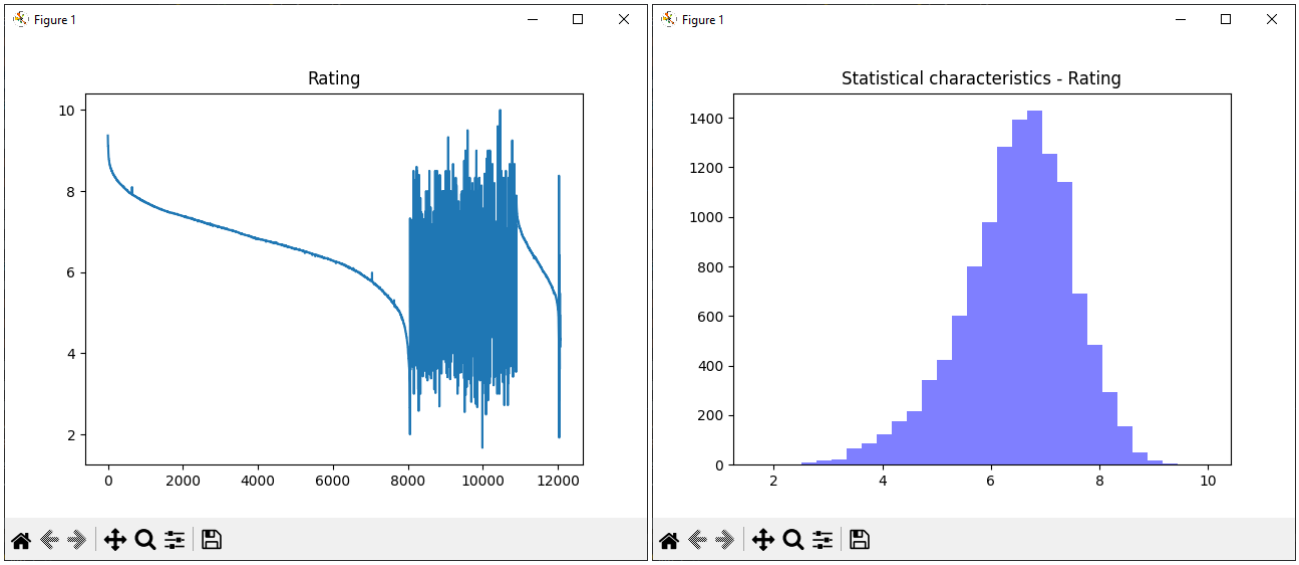


Рисунок 15 – Вибірка за рейтингом та її гістограма

Бачимо, що вибірка за рейтингом має **нормальний закон розподілу**.

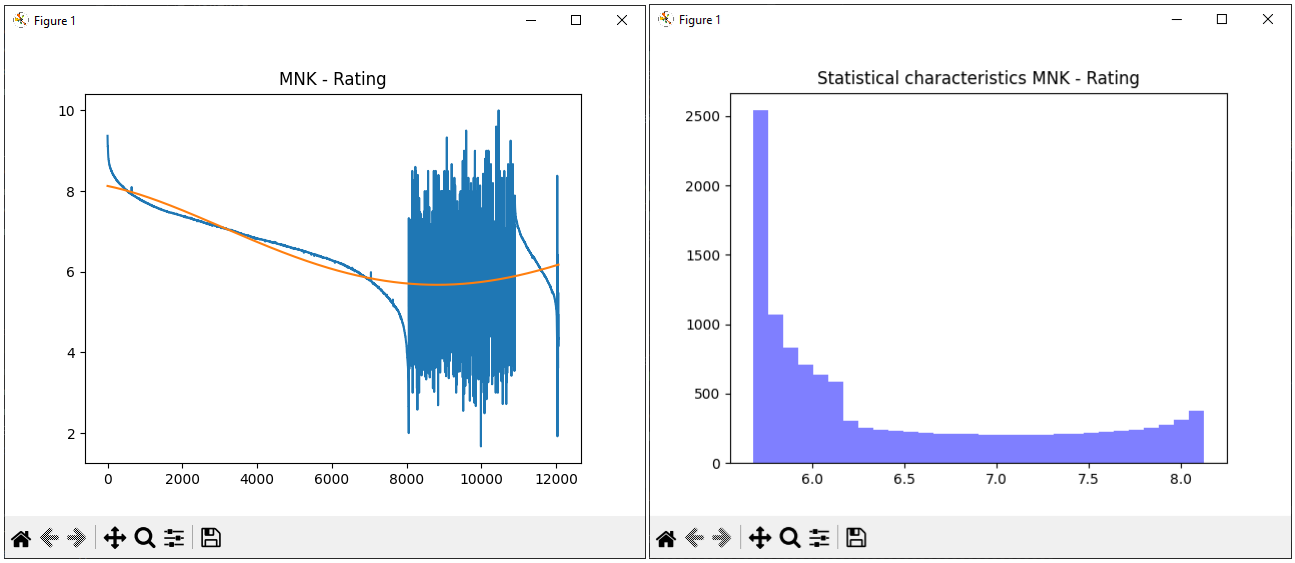


Рисунок 16 – Лінія тренду за рейтингом та її гістограма

З отриманого тренду вибірки за рейтингом у загальному представлені бачимо, що якщо приймати дані з часовою надмірністю, то спостерігався б спад рейтингу з часом.

```
--------------------------------------------------
Statistical characteristics - Rating
--------------------------------------------------
Математичне сподівання = 0.03271754915211167
Дисперсія = 0.4266008182016199
Середнє квадратичне відхилення = 0.6531468580661013

--------------------------------------------------
Statistical characteristics MNK - Rating
--------------------------------------------------
Математичне сподівання = 0.002384610797099729
Дисперсія = 0.0007944529678101718
Середнє квадратичне відхилення = 0.028186042074228368
```

Рисунок 17 – Статистичні характеристики вибірки за рейтингом та її лінії тренду

Знайдемо статистичні характеристики для цієї ж вибірки по жанрах.

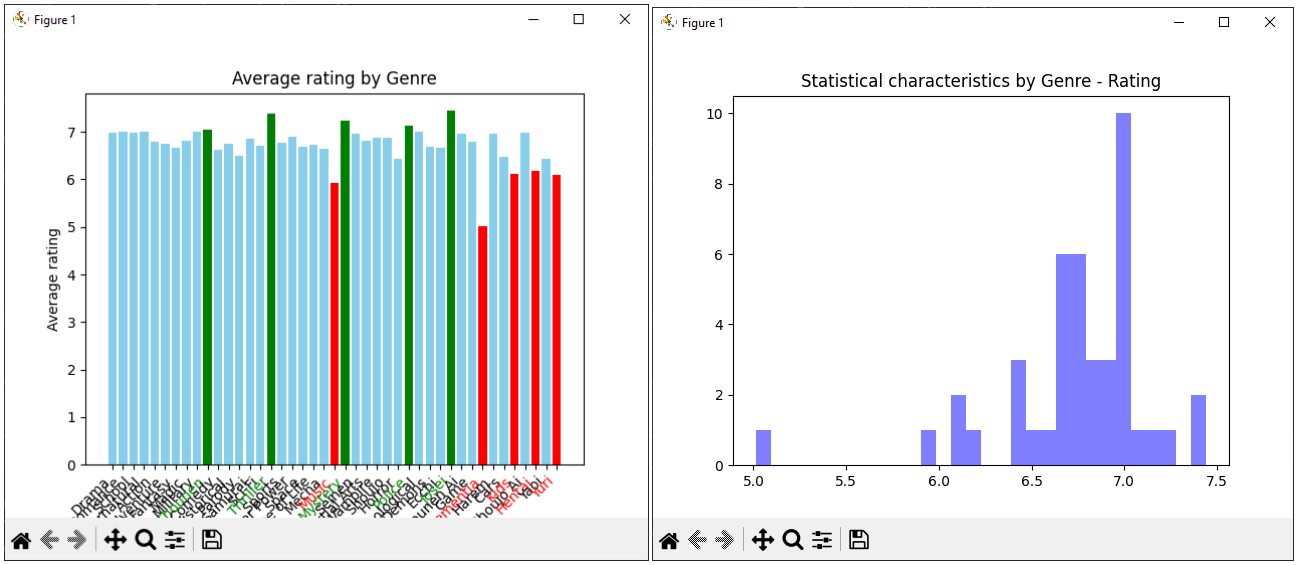


Рисунок 18 – Вибірка рейтингу за жанрами та її гістограма

Схоже, що вибірка рейтингу за жанрами має **нормальний закон розподілу**, але однозначно сказати важко.

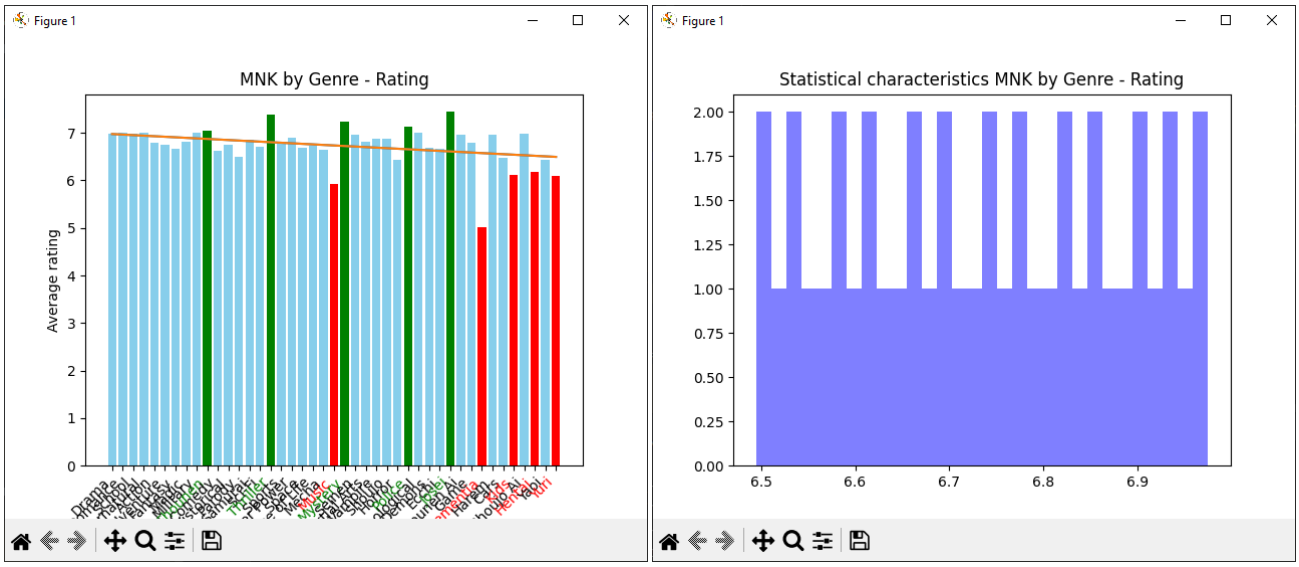


Рисунок 19 – Лінія тренду рейтингу за жанрами та її гістограма

З отриманого тренду вибірки за рейтингом у представлені за жанрами бачимо, що якщо приймати за дані з часовою надмірністю, то рейтинг за жанром був би відносно стабільний до змін у часі. А це в свою чергу свідчило б, що потрібно краще обирати жанри для публікації на сайті, а не конкретні аніме.

```
--------------------------------------------------
Statistical characteristics by Genre - Rating
--------------------------------------------------
Математичне сподівання = -0.006873023821042601
Дисперсія = 0.13675492674455206
Середнє квадратичне відхилення = 0.369803903095346

--------------------------------------------------
Statistical characteristics MNK by Genre - Rating
--------------------------------------------------
Математичне сподівання = -5.293543381412746e-13
Дисперсія = 1.6591269569682182e-25
Середнє квадратичне відхилення = 4.0732382166627797e-13
```

Рисунок 20 – Статистичні характеристики вибірки рейтингу за жанрами та її лінії тренду

Далі знайдемо статистичні характеристики для вибірки **за кількістю фанатів**.

Лістинг коду:

```
# Фан база
profit = 'members'
mnk_results(profit, df)
```

Результат:

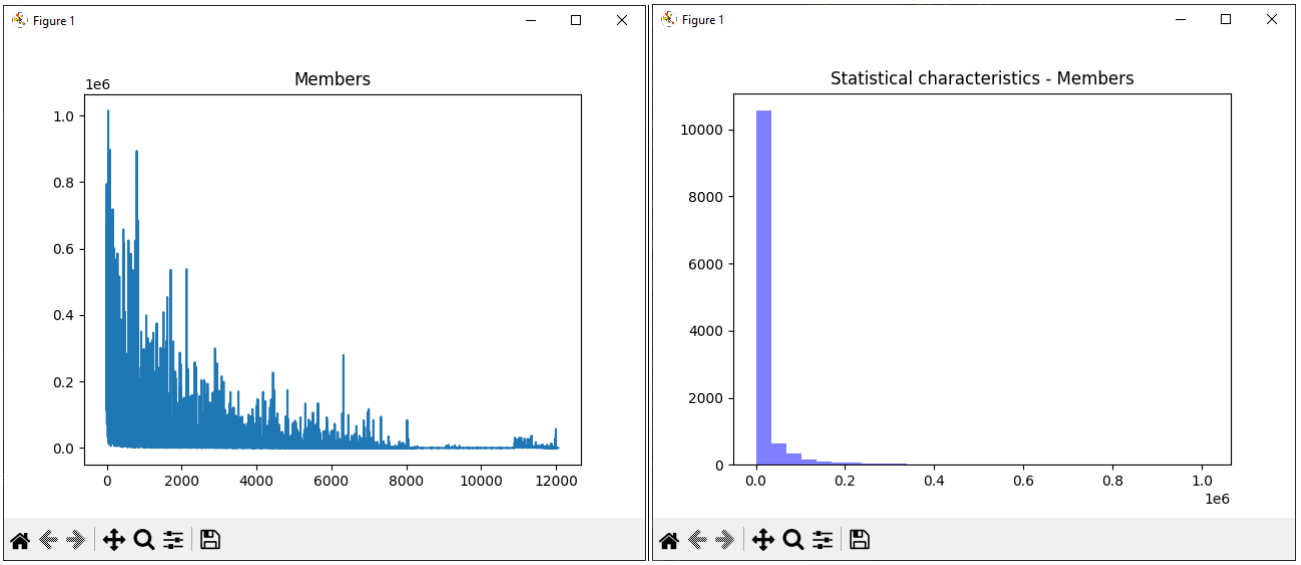


Рисунок 21 – Вибірка за кількістю фанатів та її гістограма

Бачимо, що вибірка за фанатами має **експоненційний закон розподілу**.

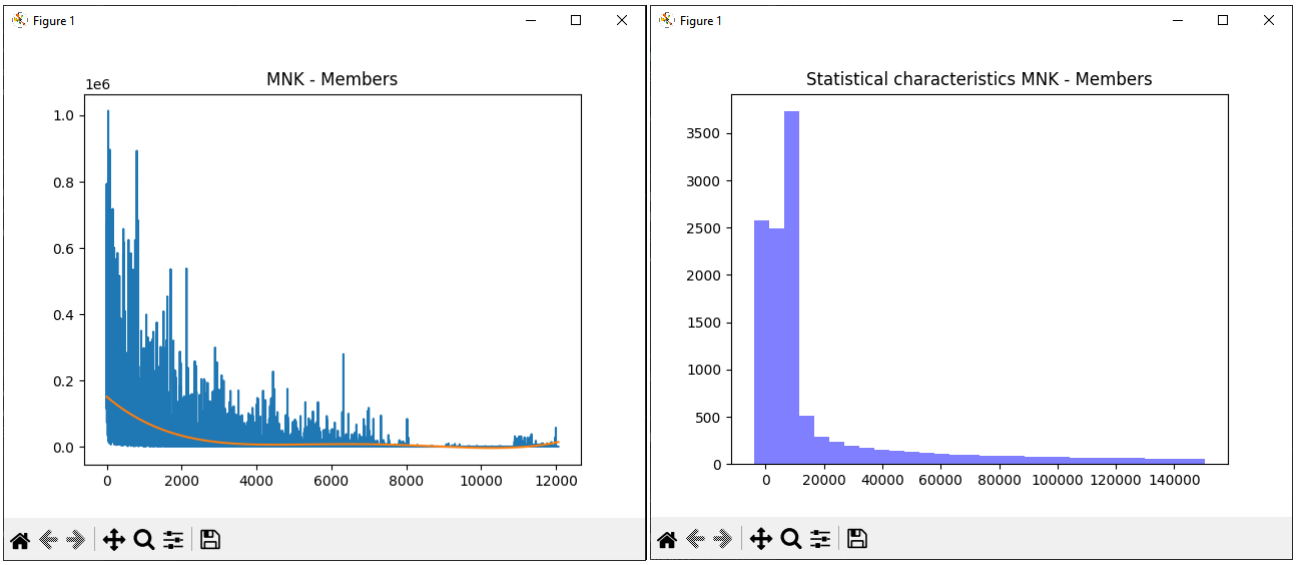


Рисунок 22 – Лінія тренду за кількістю фанатів та її гістограма

З отриманого тренду вибірки за фанатами у загальному представлені бачимо, що якщо приймати дані з часовою надмірністю, то спостерігався б спад кількості фанів з дати виходу аніме. Що абсолютно точно відповідає дійсності.

```
--------------------------------------------------
Statistical characteristics - Members
--------------------------------------------------
Матиматичне сподівання = -2521.35326474224
Дисперсія = 2108980014.8511093
Середнє квадратичне відхилення = 45923.63242221927

--------------------------------------------------
Statistical characteristics MNK - Members
--------------------------------------------------
Матиматичне сподівання = -686.409749662879
Дисперсія = 65826394.30308521
Середнє квадратичне відхилення = 8113.346677116984
```

Рисунок 23 – Статистичні характеристики вибірки за кількістю фанатів та її лінії тренду

Знайдемо статистичні характеристики для цієї ж вибірки по жанрах.

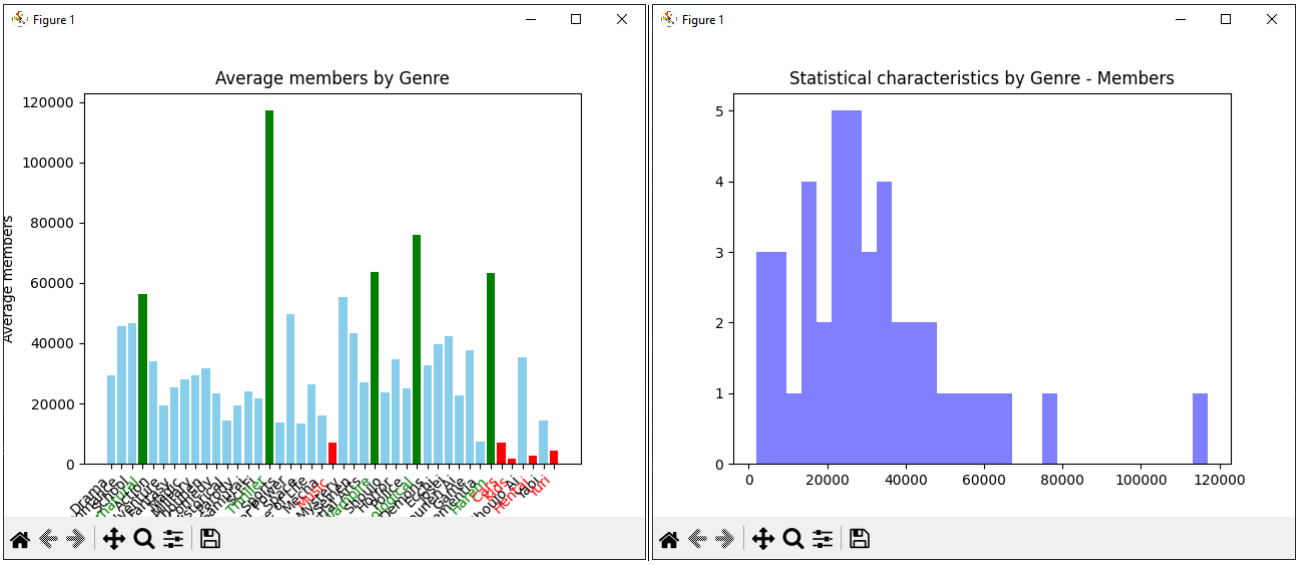


Рисунок 24 – Вибірка кількості фанатів за жанрами та її гістограма

Схоже, що вибірка кількості фанатів за жанрами має **нормальний закон розподілу**, але однозначно сказати важко.

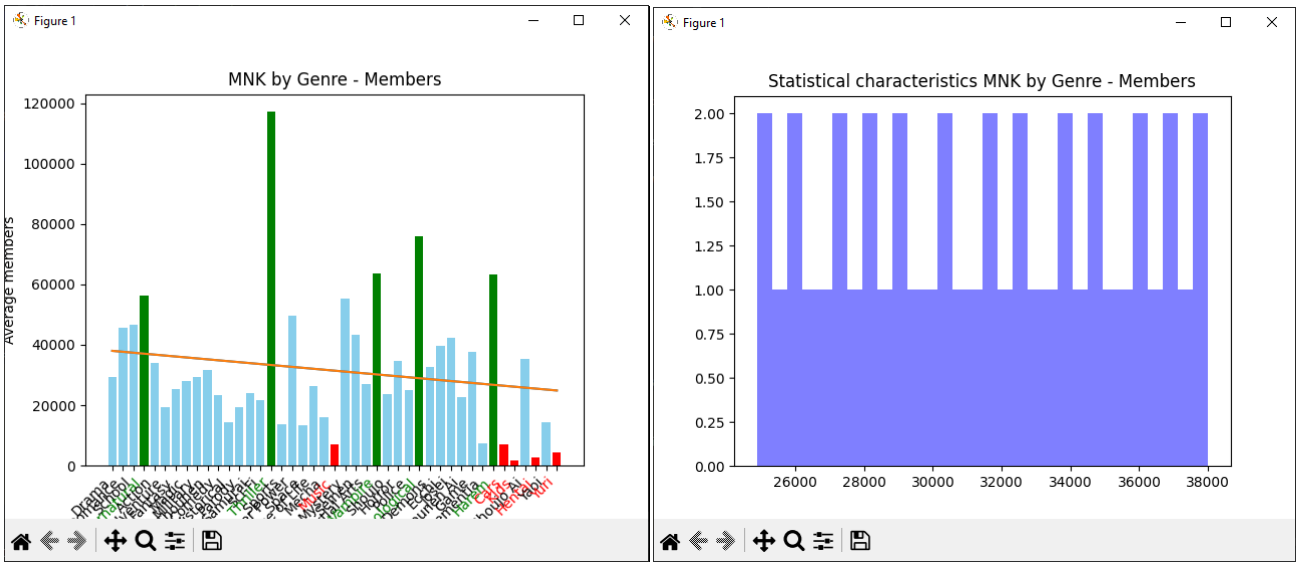


Рисунок 25 – Лінія тренду кількості фанатів за жанрами та її гістограма

З отриманого тренду вибірки за кількістю фанатів у представлені за жанрами бачимо, що якщо приймати за дані з часовою надмірністю, то кількість фанатів кожного жанру була б відносно стабільна до змін у часі. А це в свою чергу свідчило б, що якщо людині подобаються аніме певного жанру, вона буде дивитись і нові аніме цього жанру. Що однозначно так і є і підтверджує теорію важливості переваги популярних жанрів на сайті.

```
--------------------------------------------------
Statistical characteristics MNK by Genre - Members
--------------------------------------------------
Матиматичне сподівання = -2.2118911147117615e-09
Дисперсія = 2.857432907494265e-18
Середнє квадратичне відхилення = 1.6903943053306423e-09

--------------------------------------------------
Statistical characteristics - Combined score
--------------------------------------------------
Матиматичне сподівання = 0.01615155956443015
Дисперсія = 0.1631563986491621
Середнє квадратичне відхилення = 0.40392622921662574
```

Рисунок 26 – Статистичні характеристики вибірки кількості фанатів за жанрами та її лінії тренду

Ну і нарешті знайдемо статистичні характеристики для вибірки **за об’єднаним показником ефективності**.

Лістинг коду:

```
# Об'єднаний показник ефективності
profit = 'combined score'
mnk_results(profit, df)
```

Результат:

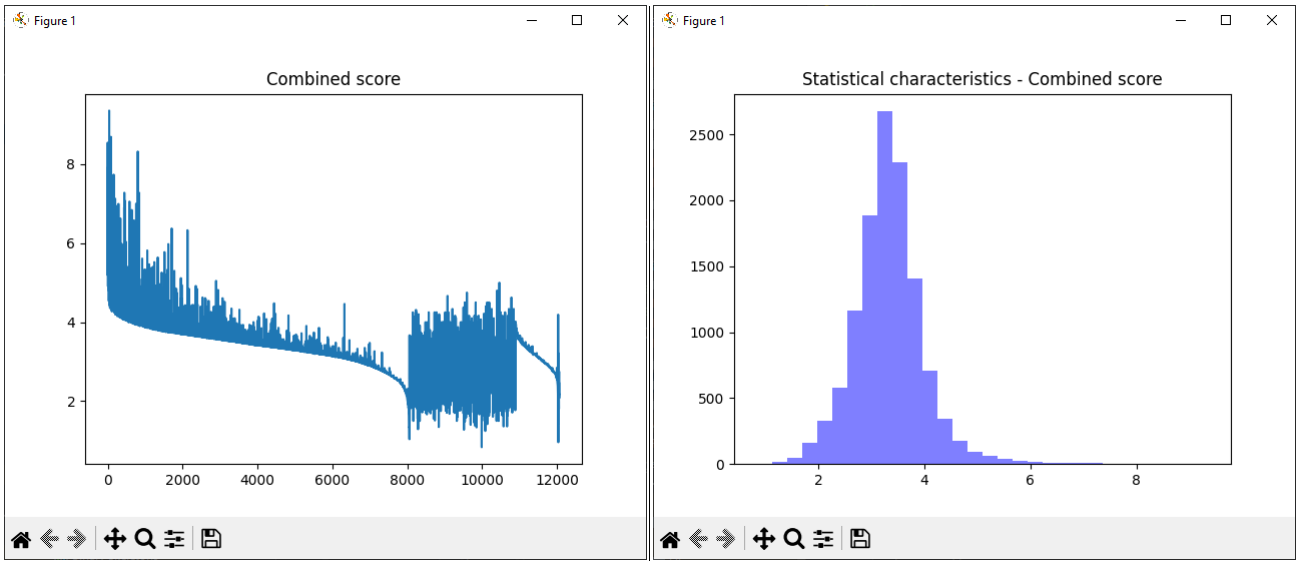


Рисунок 27 – Вибірка за об’єднаним показником та її гістограма

Бачимо, що вибірка за об’єднаним показником має **нормальний закон розподілу**. Аналогічно до вибірки за рейтингом.

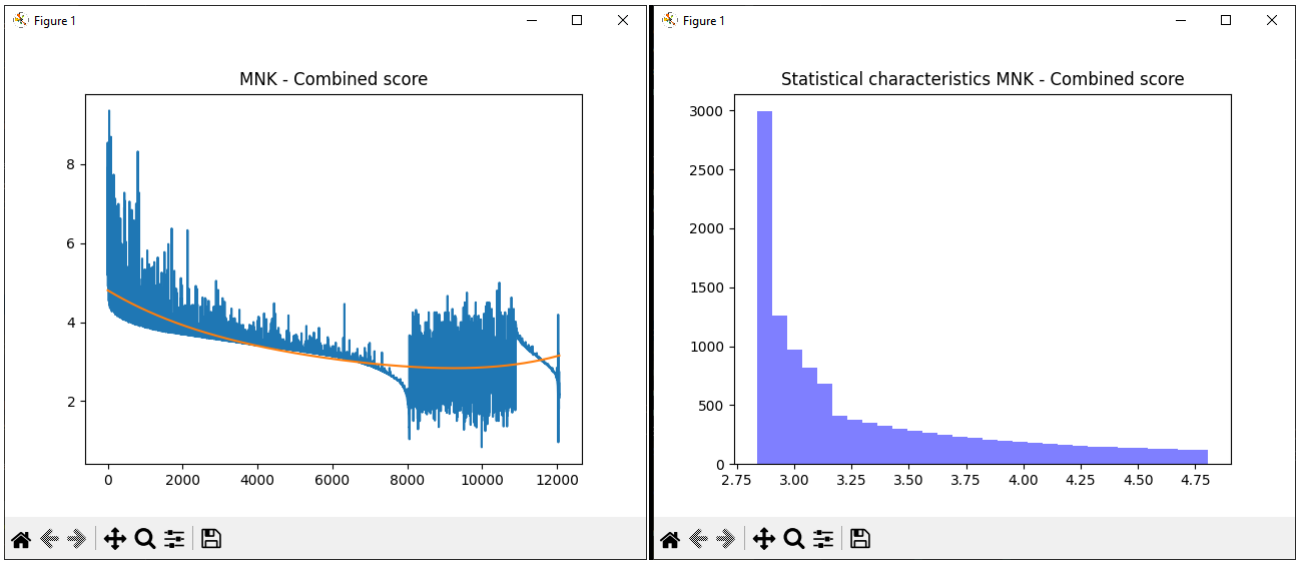


Рисунок 28 – Лінія тренду за об’єднаним показником та її гістограма

З отриманого тренду вибірки за об’єднаним показником у загальному представлені бачимо, що якщо приймати дані з часовою надмірністю, то спостерігався б спад популярності сайту з аніме.

```
------------------------------------------------------
Statistical characteristics - Combined score
------------------------------------------------------
Математичне сподівання = 0.01615155956443015
Дисперсія = 0.1631563986491621
Середнє квадратичне відхилення = 0.40392622921662574

------------------------------------------------------
Statistical characteristics MNK - Combined score
------------------------------------------------------
Математичне сподівання = -0.002192675194544469
Дисперсія = 0.0006717099583680722
Середнє квадратичне відхилення = 0.025917367890433477
```

Рисунок 29 – Статистичні характеристики вибірки за об’єднаним показником та її лінії тренду

Знайдемо статистичні характеристики для цієї ж вибірки по жанрах.

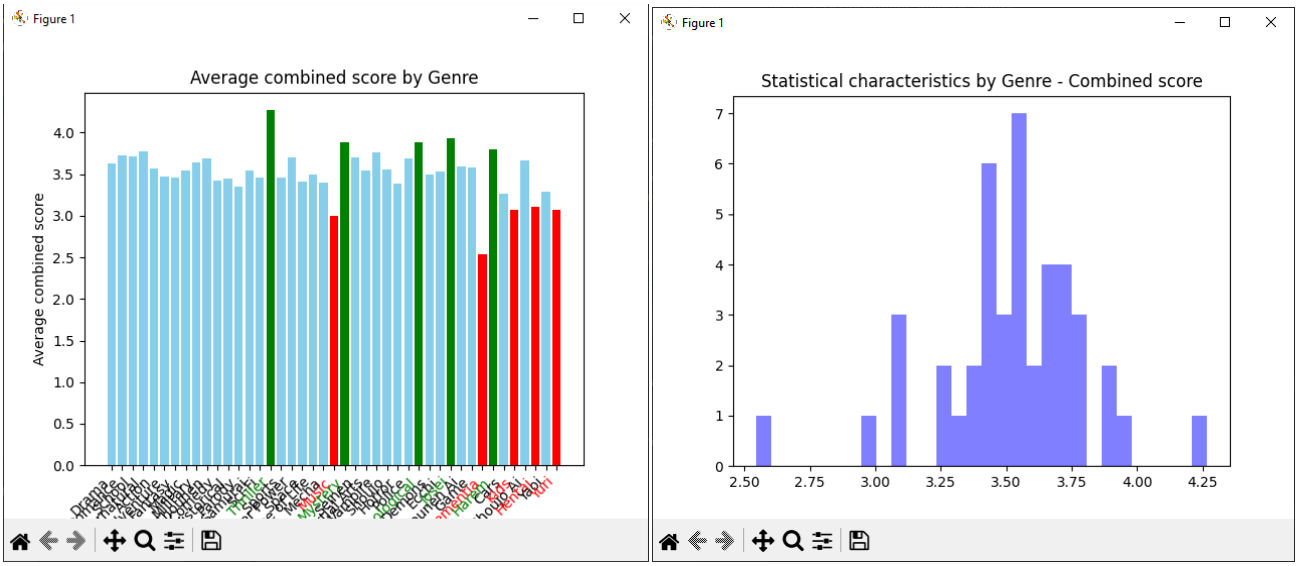


Рисунок 30 – Вибірка об’єднаного показника за жанрами та її гістограма

Схоже, що вибірка об’єднаного показника за жанрами має **нормальний закон розподілу**.

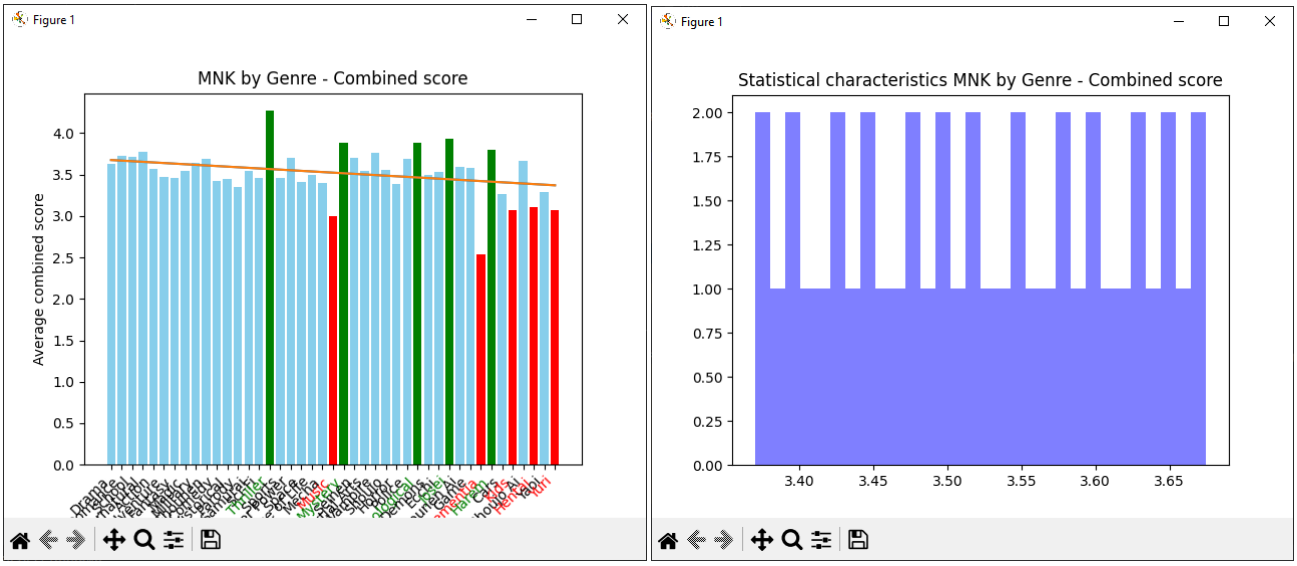


Рисунок 31 – Лінія тренду об’єднаного показника за жанрами та її гістограма

З отриманого тренду вибірки за об’єднаним показником у представлені за жанрами бачимо, що якщо приймати за дані з часовою надмірністю, то в загальному популярність сайту з аніме буде відносно стабільна до змін у часі. Таке передбачення реалістичніше за попереднє. А це в черговий раз підтверджує теорію важливості переваги популярних жанрів на сайті.

```
-----------------------------------------------------
Statistical characteristics - Combined score
-----------------------------------------------------
Математичне сподівання = 0.01615155956443015
Дисперсія = 0.1631563986491621
Середнє квадратичне відхилення = 0.40392622921662574

-----------------------------------------------------
Statistical characteristics MNK - Combined score
-----------------------------------------------------
Математичне сподівання = -0.002192675194544469
Дисперсія = 0.0006717099583680722
Середнє квадратичне відхилення = 0.025917367890433477
```

Рисунок 32 – Статистичні характеристики вибірки об’єднаного показника за жанрами та її лінії тренду

2.3. Аналіз отриманих результатів

Сформульовано задачу для задачі повноцінного аналізу даних у прикладній сфері.

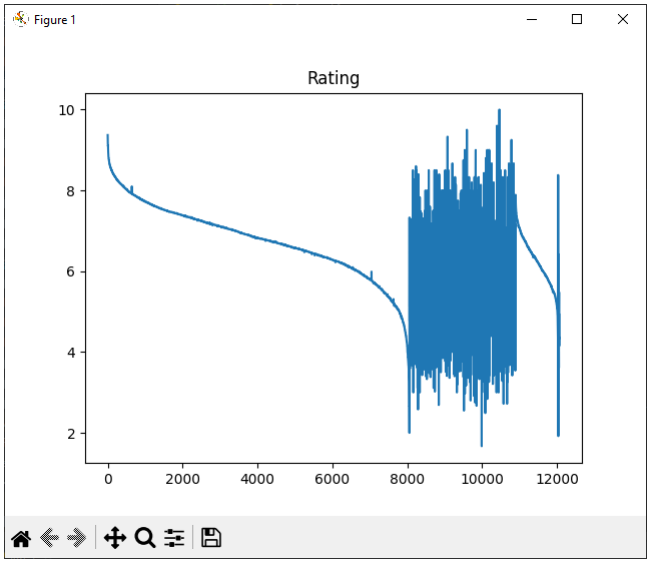
Обрано дані, які можуть мати часову надмірність. Так, як для даної задачі не знайшлось підходящого датасету, а власноруч створити такий не так просто – за датасет було обрано рейтинг аніме-тайтлів з сайту. Такий набір даних було обрано з метою протестувати працездатність розробленого скрипту на реальному прикладі.

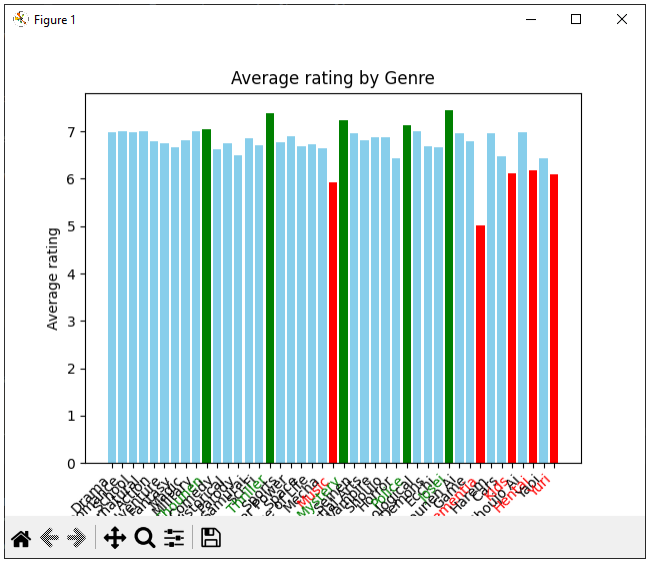
Розроблено програму, яка виконує лінійний та статистичний аналіз даних.

Програму протестовано для трьох показників ефективностей: **рейтинг, кількість фанатів** та **об'днаним показником**, який складає їх середнє арифметичне у двох варіантах представлення даних: **на всій вибірці** та **за жанрами** аніме.

Лінійний аналіз даних

У результаті тестування отримано наступні результати:





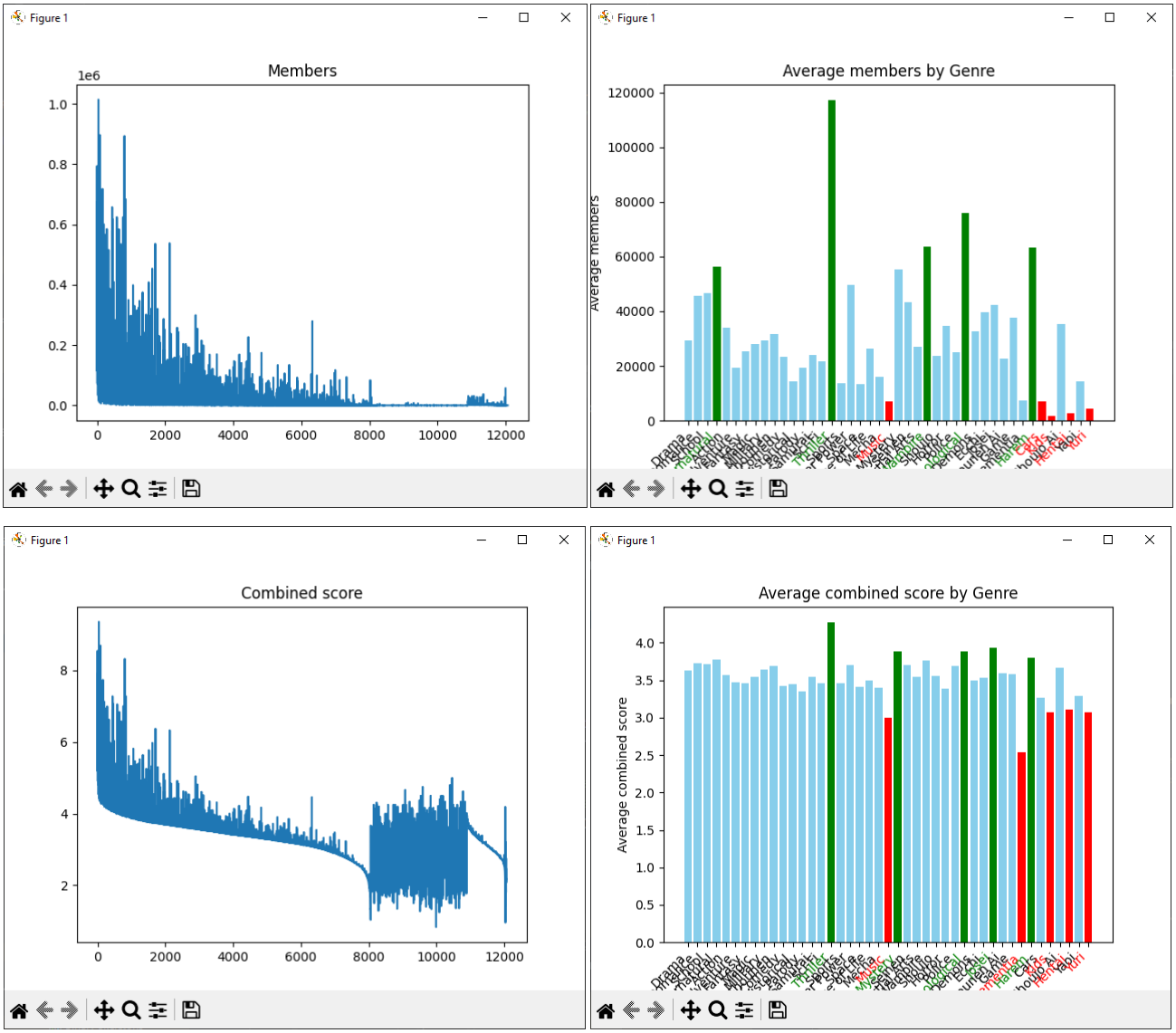


Рисунок 33 – Графіки лінійного аналізу за 3 показниками ефективності в 2 видах представлення

З отриманих графіків загального представлення можна зробити висновок про нинішню ситуацію з рівнем задоволеності та приростом кількості відвідувачів сайту. Зі стовчастих діаграм можна якісно оцінити найпопулярніші і навпаки жанри.

Об’єднаний показник дозволяє оцінити вагомість аналізу жанрів при завантаженні нових аніме на сайт.

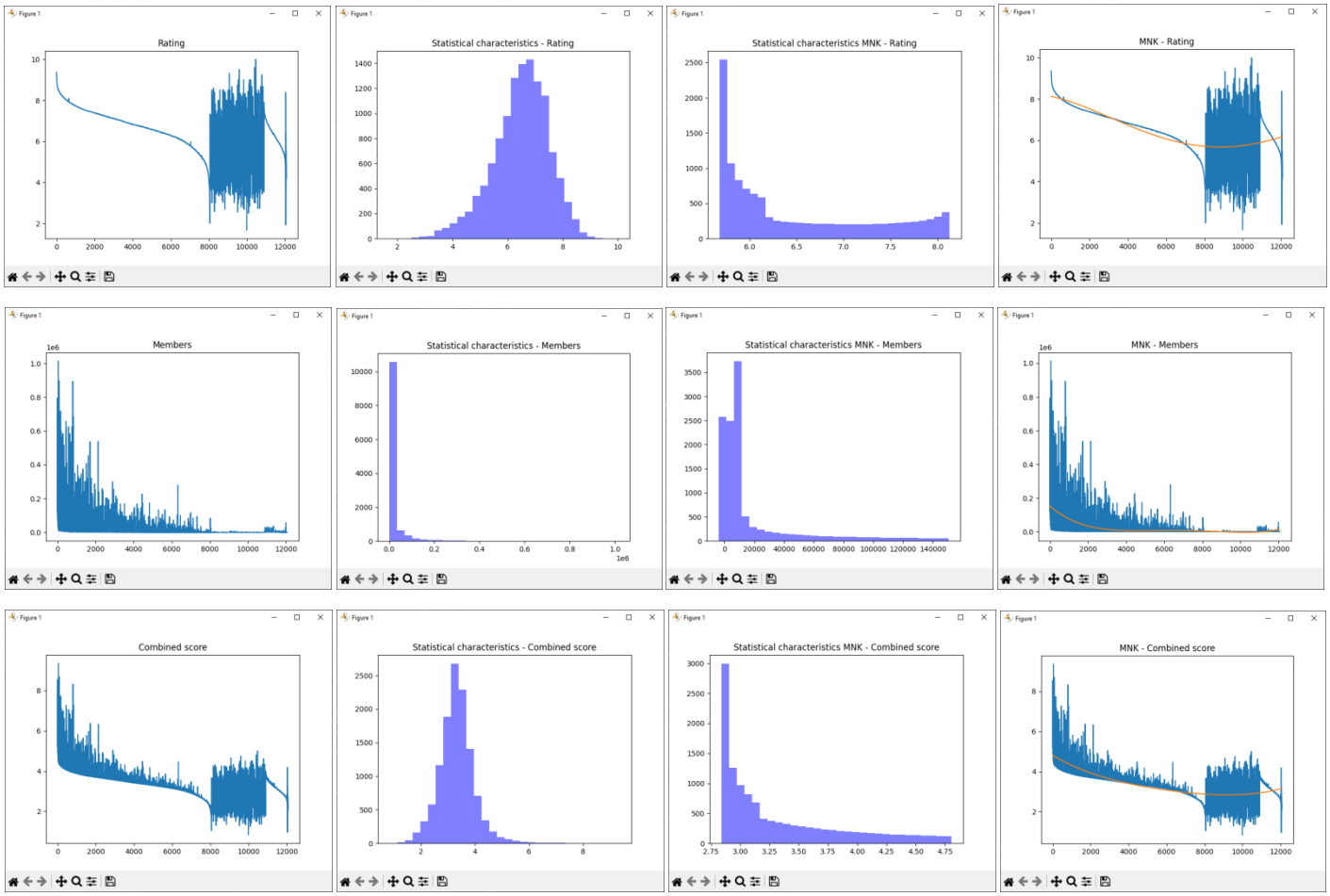
Таблиця 1 – Схожість об’єднаного показнику зі своїми попередниками

	Рейтинг	Кількість фанатів	Об’єднаний показник
Уся вибірка	1	2	1, 2
За жанрами	1	2	1

Те, що об’єднаний показник при аналізі вибірки за жанрами унаслідував поведінку вибірки за рейтингом говорить про те, що бізнес аналітику варто звертати більше уваги на рівень задоволеності клієнтів їх сайту, щоб той міг розвиватись.

Статистичний аналіз даних

У результаті тестування отримано наступні результати:



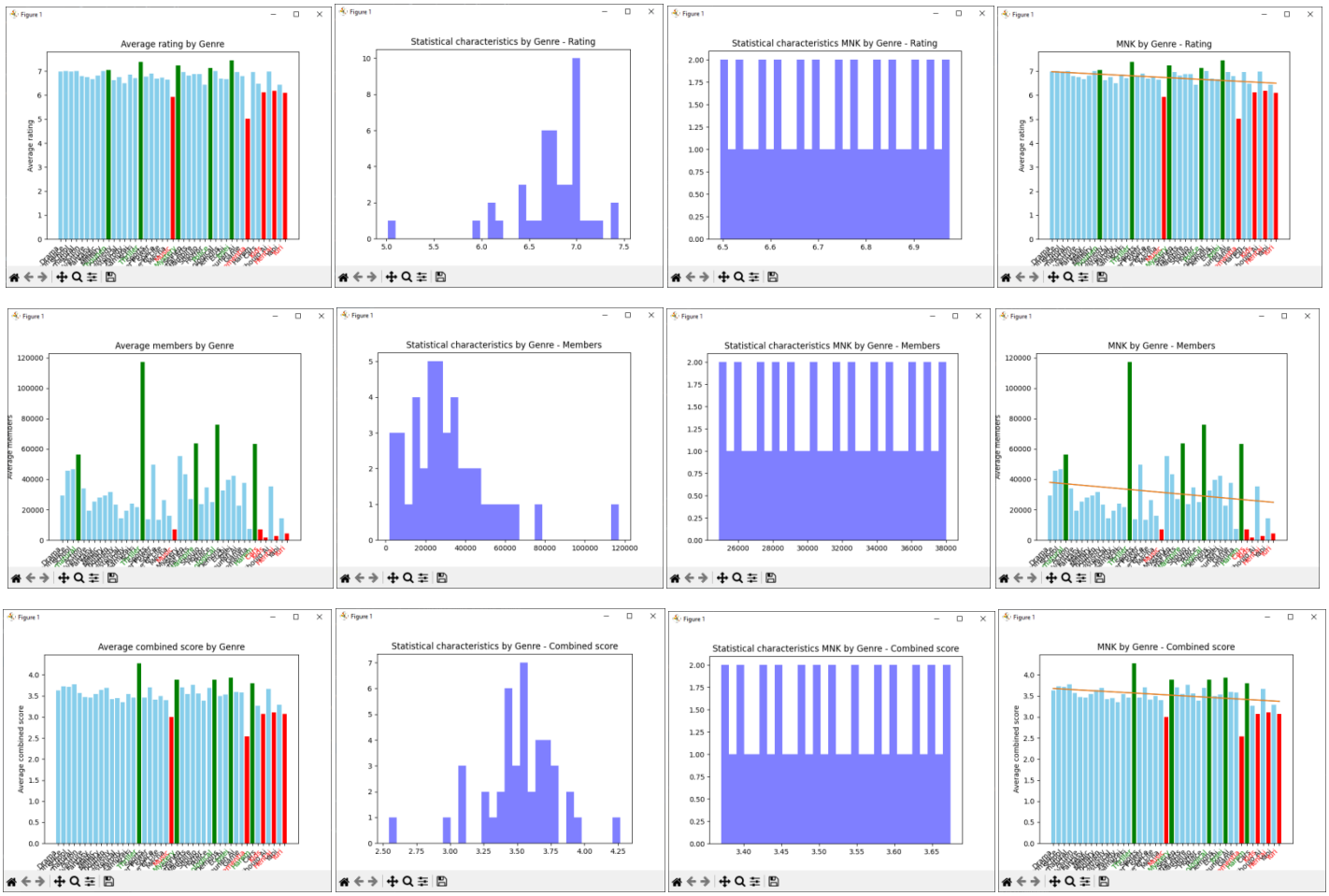


Рисунок 34 – Графіки статистичного аналізу за 3 показниками ефективності в двох видах представлення

Для статистичної оцінки виявилось навпаки – кориснішими виявились графіки для всієї вибірки. Адже за допомогою гістограм вибірки можна сказати, чи все в порядку з показником ефективності. Нормальний закон розподілу свідчить, що сайт має позитивну динаміку для цього показника. Натомість експоненційний розподіл – про те, що на покращення цього показника слід звернути увагу.

А з гістограми лінії тренду вибірки можна сказати як змінюється цей показник з часом. За допомогою аналізу гістограми лінії тренду можна прогнозувати подальші зміни цих величин.

Натомість гістограма за жанрами виконує ту ж функцію, що й стовпчаста діаграма в лінійному аналізі. Допомагає відсіяти непопулярні жанри й скорнценруватися на популярних. Утім за стовпчастими діаграмами це робити зручніше.

Гістограма лінії тренду за жанрами ж слугує показником стабільності популярності певних жанрів. Якщо тренд – лінійний, значить жанри, присутні на сайті затребувані.

Об’єднаний критерій оцінює вплив обраних показників ефективності.

Таблиця 2 – Схожість об’єднаного показнику зі своїми попередниками для всієї вибірки

	Рейтинг	Кількість фанатів	Об’єднаний показник
Вибірка	1	2	1, 2
Гістограма вибірки	1	2	1
Гістограма лінії тренду	1	2	1

Таблиця 3 – Схожість об’єднаного показнику зі своїми попередниками для вибірки за жанрами

	Рейтинг	Кількість фанатів	Об’єднаний показник
Вибірка	1	2	1
Гістограма вибірки	1	2	1
Гістограма лінії тренду	1	2	1, 2

Те, що об’єднаний показник здебільшого наслідує рейтинг, означає, що популярні аніме-новинки мають перевагу в даному контексті аналізу над аніме з великою фан базою. Це можна пояснити тим, що аудиторія вірусних аніме-новинок стрімко росте і сприяє великому приросту нових відвідувачів сайту. Це добре видно на рисунку 23.

Отже можна зробити висновок про пораду зосередитись на аніме-новинках у популярних жанрах.

Висновок:

Створено міні-проект для прогнозування динаміки зміни показників ефективності аніме сайту.

Сформульовано задачу максимізації переглядів сайту, відповідно до якої обрано 2 показники ефективності і штучно створено третій.

Розроблено та протестовано програмний скрипт для виконання поставленої задачі.

У результаті аналізу прийнято рішення про необхідність пильно слідкувати за новинками серед популярних жанрів аніме.