Data Mining Project

Team Members
1.Nikesh Vaddi - K00535145
2.Dinesh - K00505655
3.Deepak - K00505470
4.Keerthi - K00533058
5.Sohith - K00538337
6.Jaswanth - K00527954
7.Bavan Maganti - K00531485

Dataset collection

- Source of the dataset : Kaggle
- Dataset name : Most watched movies and series
- Size: 1.34MB
- Number of attributes : 7
- 1.Rank
- 2.Title
- 3.Type
- 4.Gener
- 5.Premiere
- 6.watch time
- 7.watch time in million

Description of attributes in the dataset

- 1.Rank : Each Tv show, or movie is ordered according to their watch time in desending order
- 2.Title: Name of the TV show or movies
- 3.Type: TV show or movie
- 4.Genre: Action , Drama, Thriller etc.
- 5. Premier: official release of the content year
- 6.Watch time: Total number of hours watched
- 7. Watch time in millions: Total number of hours watched in millions

Data mining tools

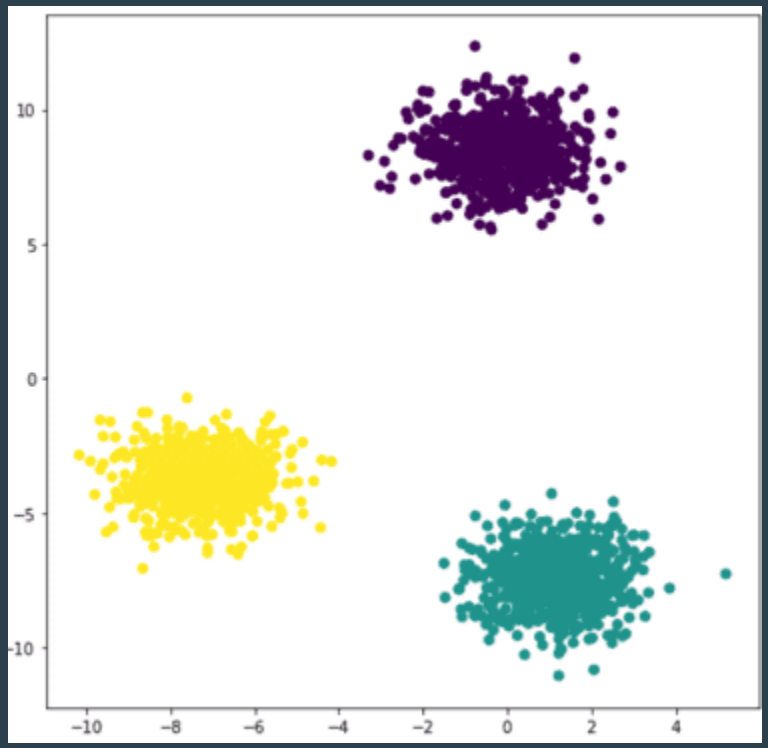
- Jupyter Notebook
- Python
- Libraries :
  - **Pandas** handles data loading, cleaning, and transformation.
  - **Seaborn** and **Matplotlib** are used for creating various types of visualizations to explore and present the data.
  - **Scikit-learn** is used for preprocessing (scaling and encoding) and for performing the clustering algorithm

Data Cleaning

- Data cleaning involves fixing or removing incorrect, corrupted, in correctly formatted, duplicate, or incomplete data within a data set

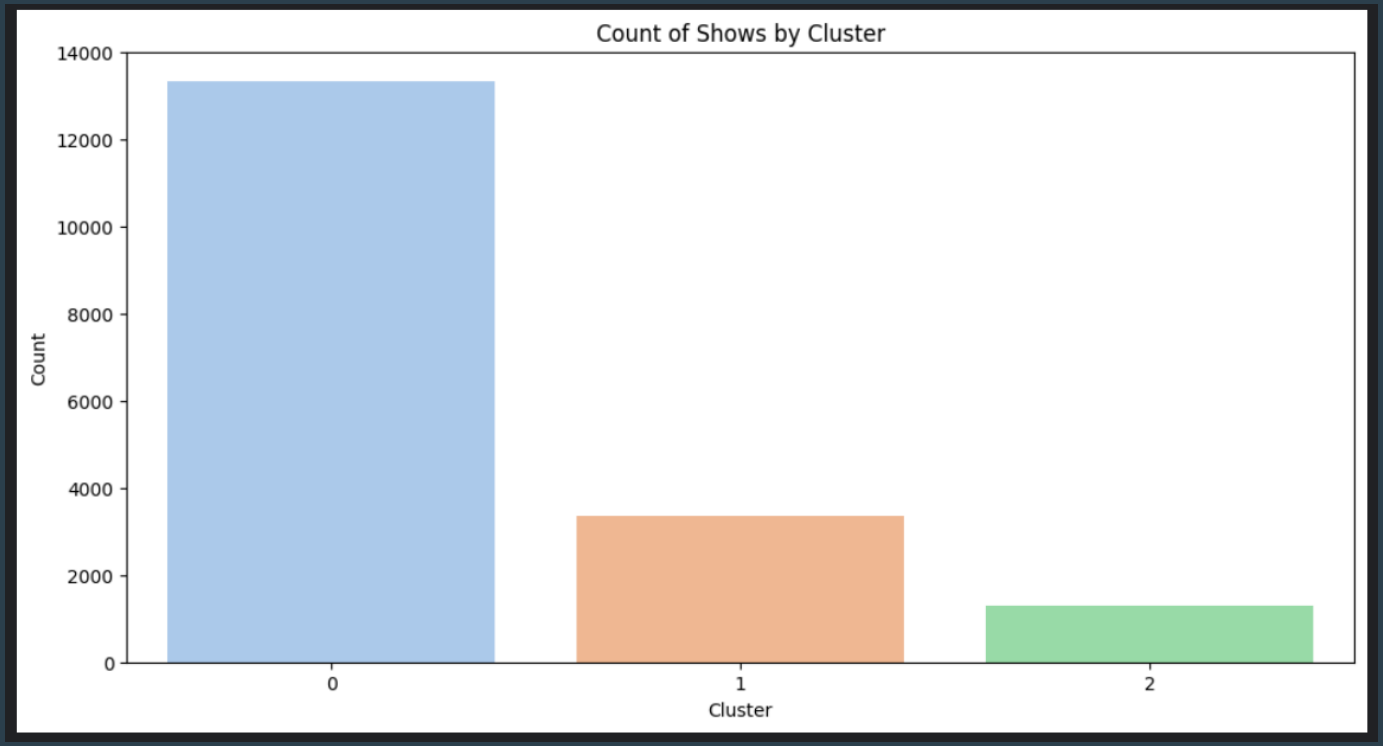
Data mining task

- Clustering
- Objective: Group genre of movies and tv show which have different watch time
- Task: Group the genre according to similarity in watch time
- Algorithm: K-Means Algorithm



Findings after running the Algorithm

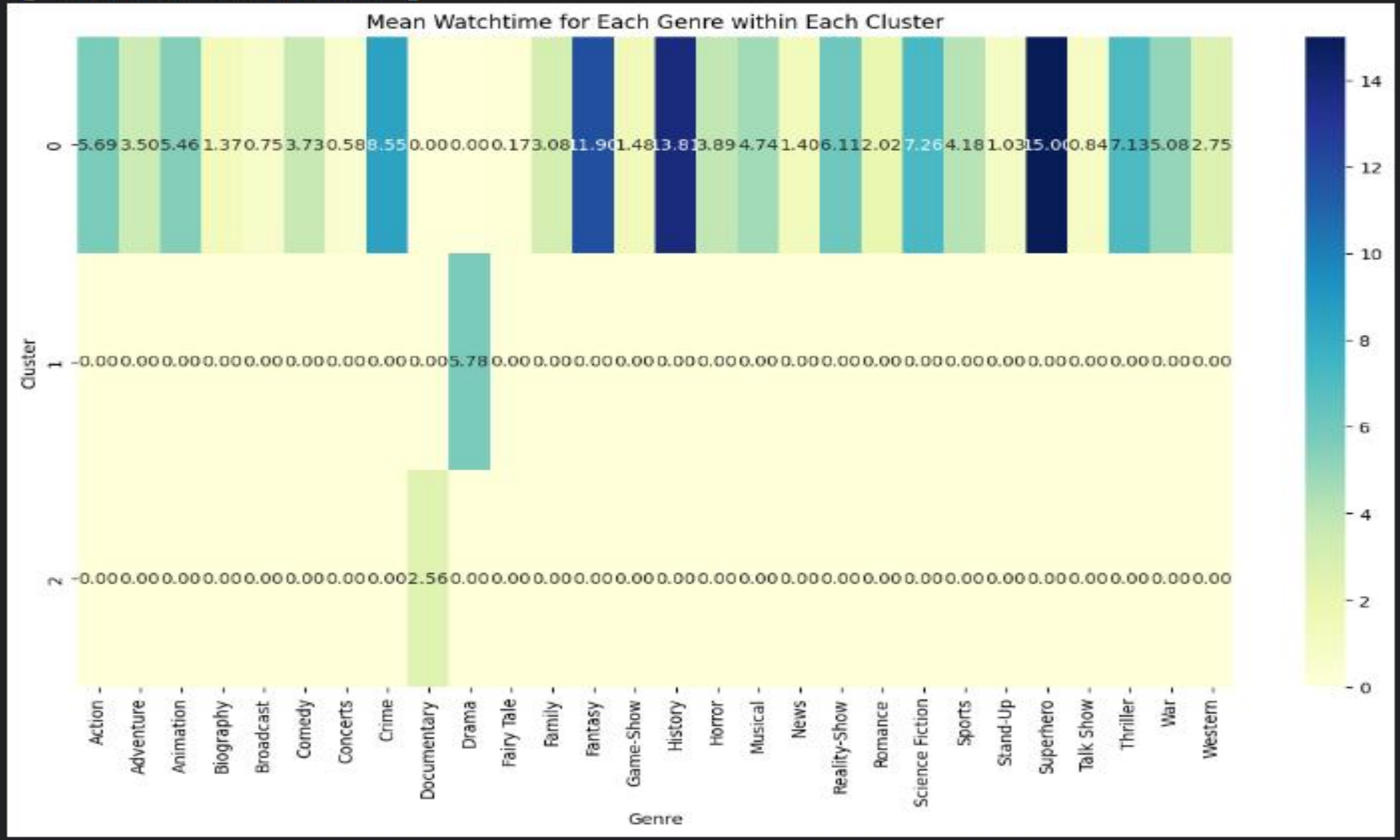
- The cluster are divided into 0,1 and 2 where 0 cluster has more watch time and 1 has less watch time than 0 and 2 is the least watch time



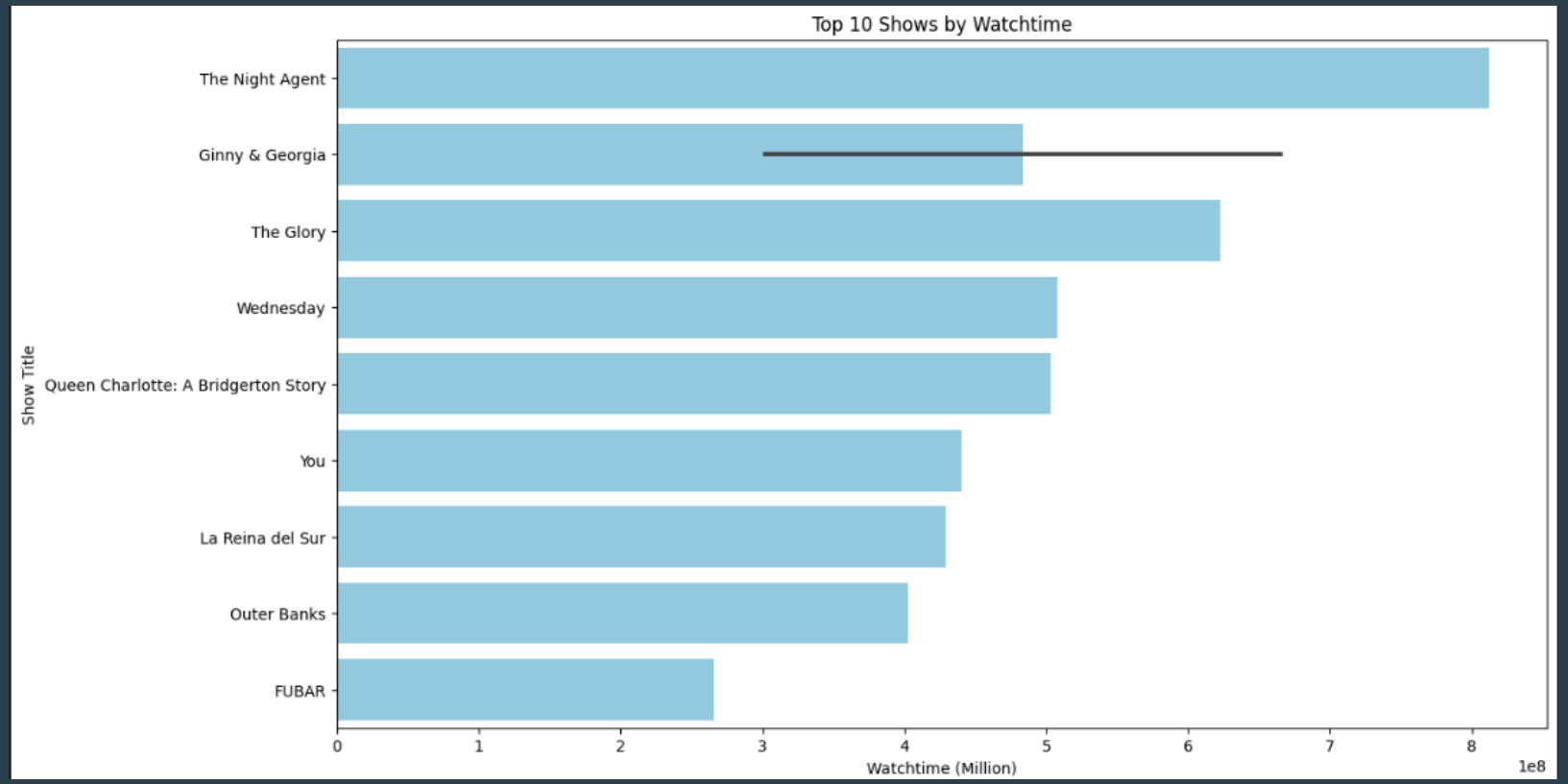
Genre in each Cluster

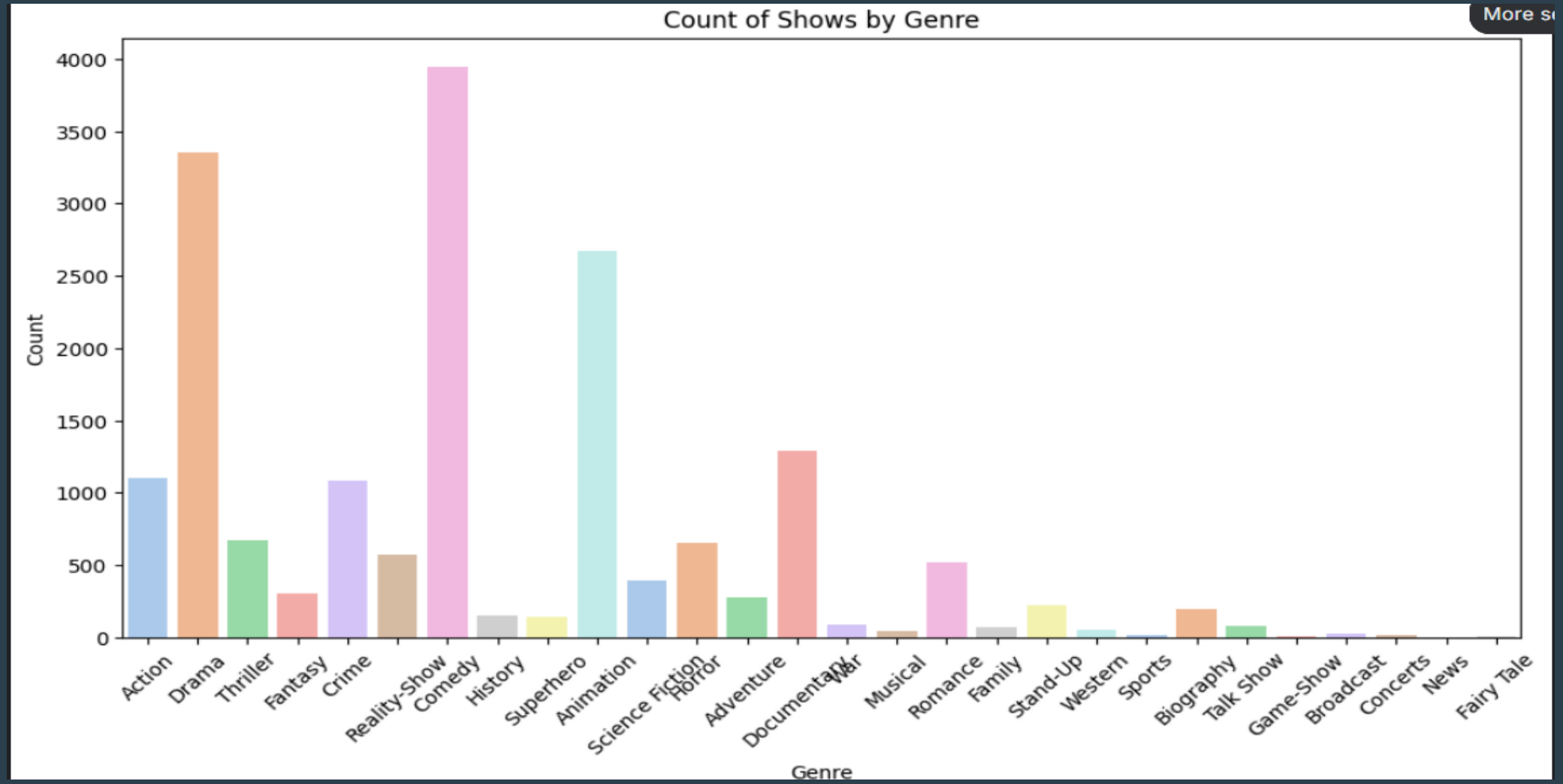
- In the cluster 0 there are several genre
- In the cluster 1 there is one genre
- In the cluster 2 there is one genre

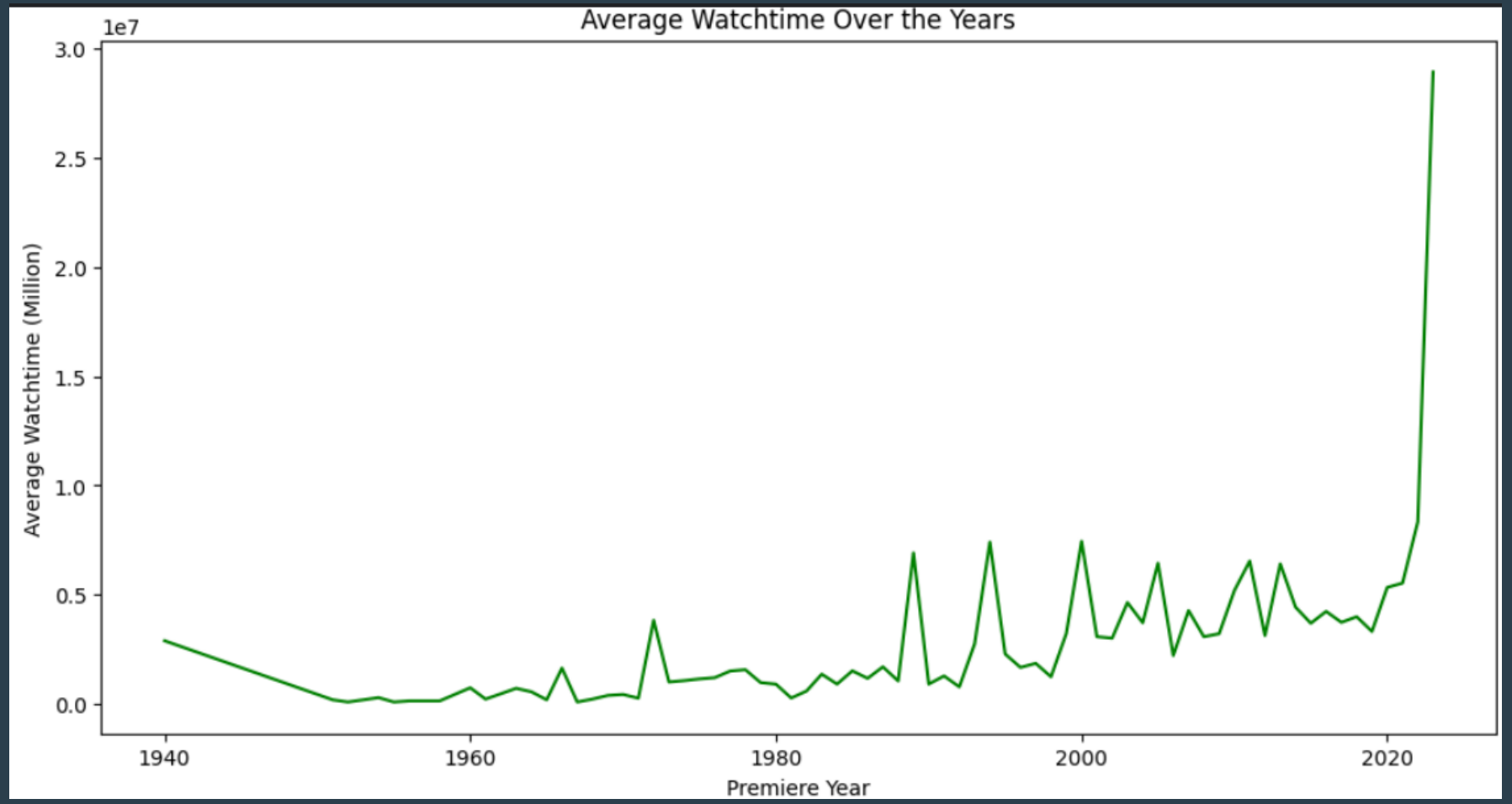
[3 rows x 28 columns]

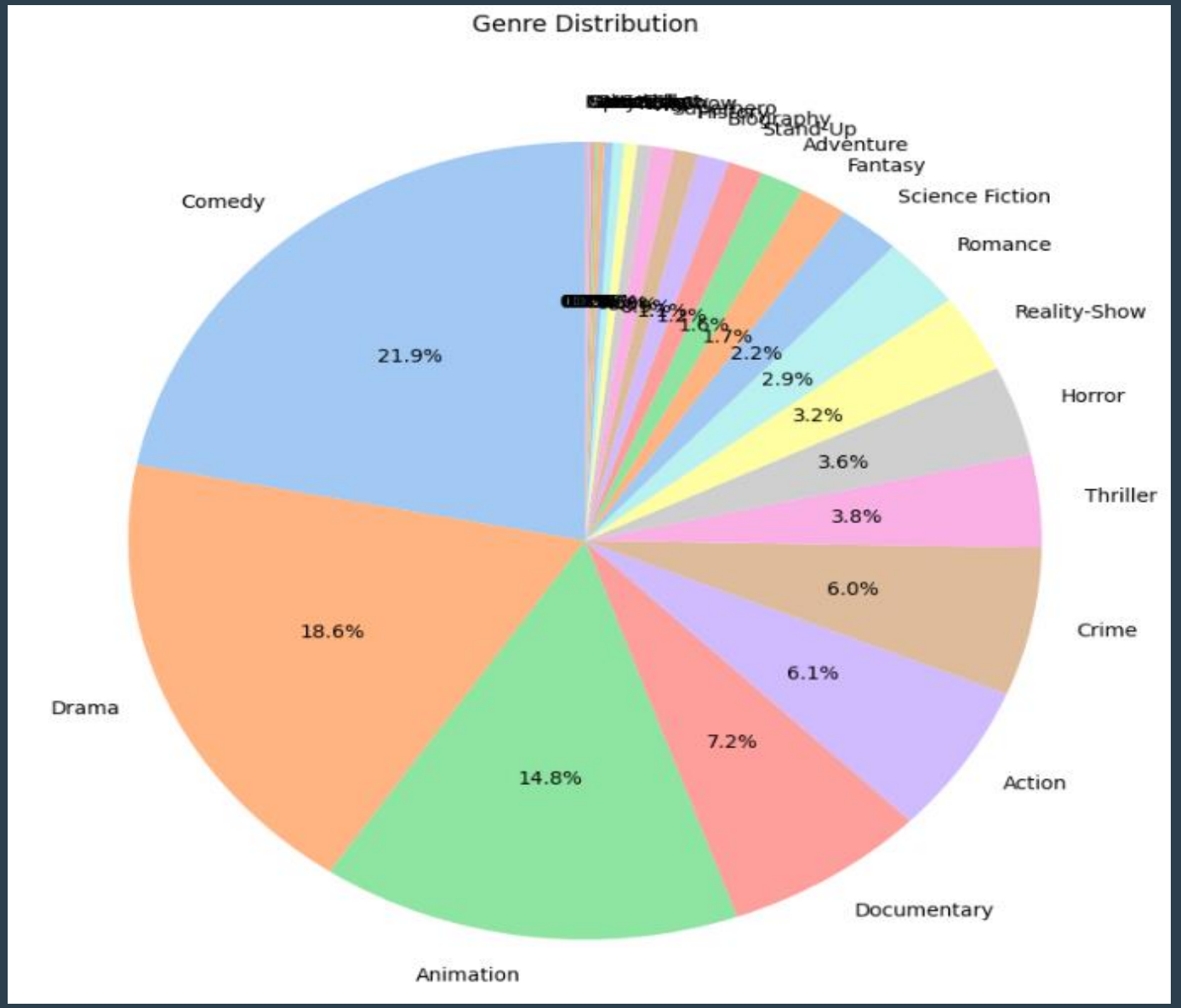


Other Findings




Count of Shows by Genre
More s
Count
4000
3500
3000
2500
2000
1500
1000
500
0
Action
Drama
Thriller
Fantasy
Crime
Reality-Show
Comedy
History
Superhero
Animation
Science Fiction
Horror
Adventure
Documentary
War
Musical
Romance
Family
Stand-Up
Western
Sports
Biography
Talk Show
Game-Show
Broadcast
Concerts
News
Fairy Tale
Genre


Average Watchtime Over the Years
1e7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Average Watchtime (Million)
1940
1960
1980
2000
2020
Premiere Year


Genre Distribution
Comedy
21.9%
Drama
18.6%
Animation
14.8%
Documentary
7.2%
Action
6.1%
Crime
6.0%
Thriller
3.8%
Horror
3.6%
Reality-Show
3.2%
Romance
2.9%
Science Fiction
2.2%
Fantasy
1.7%
Adventure
Stand-Up

THANK YOU