Movie Analysis

Introduction

- Data analytics plays a major role in making several business decisions. The end goal is to mine information which can be further used for business development.
- Movies have a lot of influence in our lives today. Hence, we decided to analyze movie data for our project.
- We chose the following datasets:
    - IMDb, one of the largest movie dataset consisting of movie information along with the cast and crew information.
    - MovieLens, contains user wise ratings of each movie.

Dataset description

Database - IMDb dataset	**Database - MovieLens Dataset**
Movie - A collection with information on titles.	Dataset on movie ratings from the users. Each user can have ratings for one or more movies.
Person - A collection with information on cast & crew.	
Person-Roles - A collection with information on roles played by cast & crew in different movies	

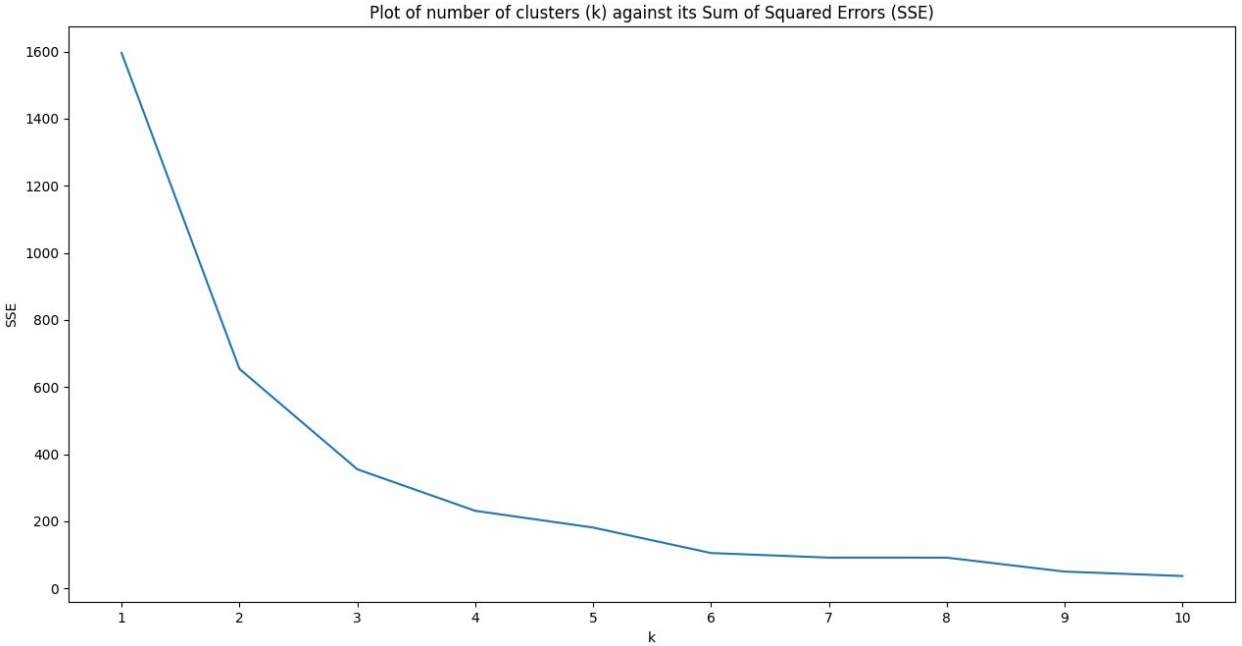
Data aggregation

- The IMDB dataset and the user ratings from MovieLens dataset were both inserted into a MongoDB database.
- Following are the collections in the database after the insertion process:
    - Movie: Contains all the documents with the movie's information. Since we only wanted movies, we chose only those titles with the titleType short, tvShort, movie and tvMovie.
    - Person: Contains all the documents with the person's information.
    - Person_Roles: Contains all the documents with the person's profession information.
    - Ratings: Contains all the documents with the ratings by each user for certain movies provided in the MovieLens dataset. The ratings file in the MovieLens dataset contained its own movieId for the movies. Hence, we exchanged the movieId with the corresponding imdbId value provided in the links file of the dataset to maintain coherence between our collections.

K-means clustering

- A popular algorithm that is used to partition the given data into groups for further processing and analysis
- A ’normalizedAvgRating' field was created using which the clusters were formed.

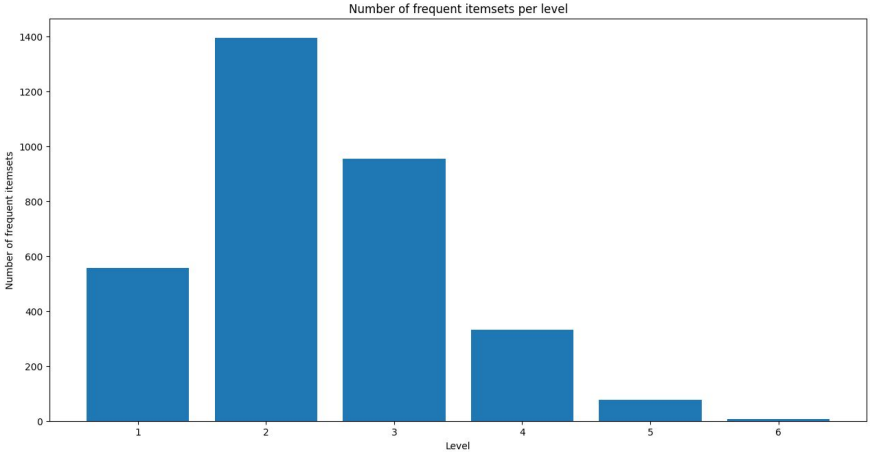
Knee method used to determine optimal k

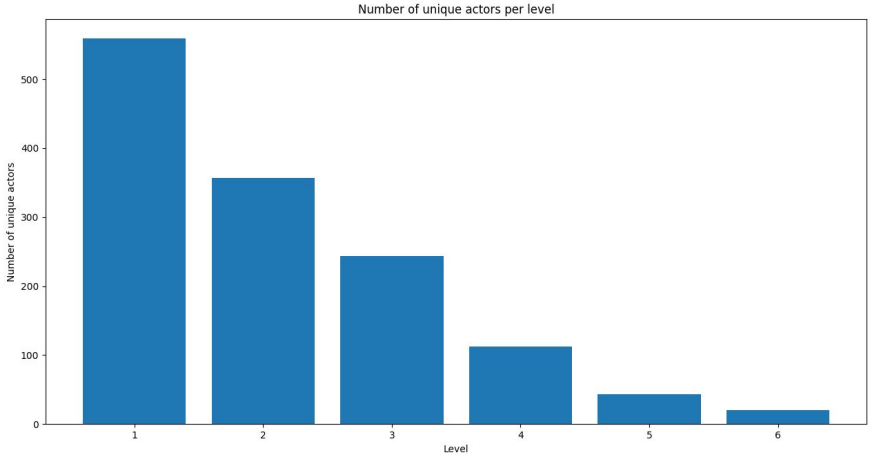


Frequent itemset mining

- Frequent itemset mining is a good way to understand certain patterns in the data.
- We mined the actors who worked together in a particular amount of movies. To obtain these frequent itemsets, we employed the Apriori algorithm.
- We used the Person_Roles collection to implement the algorithm. We first filtered out the roles which were not 'actor', 'actress', or 'self' and grouped the movies by each actor.
- We ran the algorithm with a minimum support value of 5. The maximum number of actors observed in a frequent itemset was 6.

Frequent itemset mining



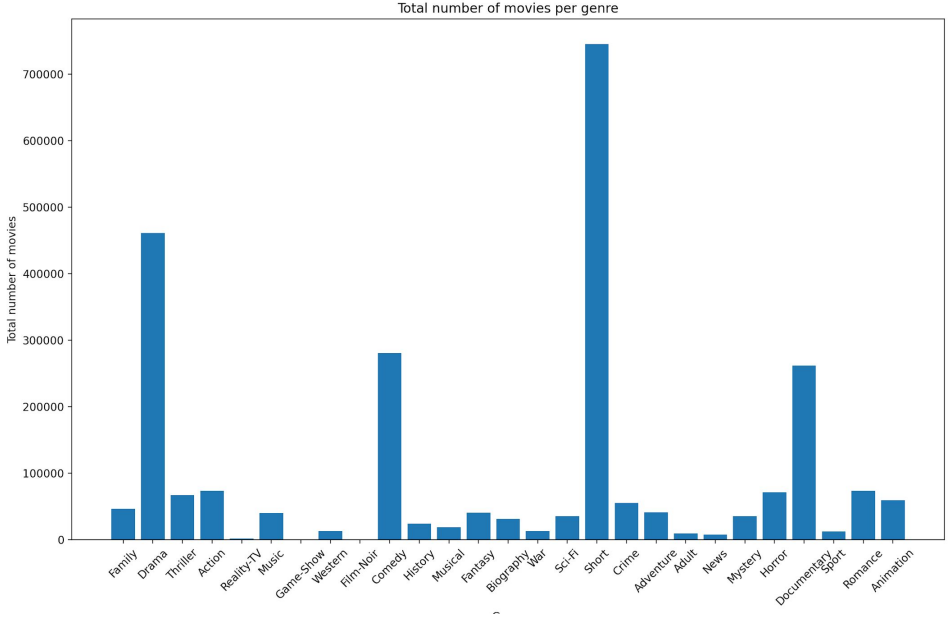


The figure on the left shows the number of itemsets in each level while the figure shows the number of distinct actors in each level. The number of distinct actors in each level keep on decreasing. However, the number of itemsets per level increase for k=2 but decrease after that and eventually becomes 0 for k=7 where the algorithm halts.

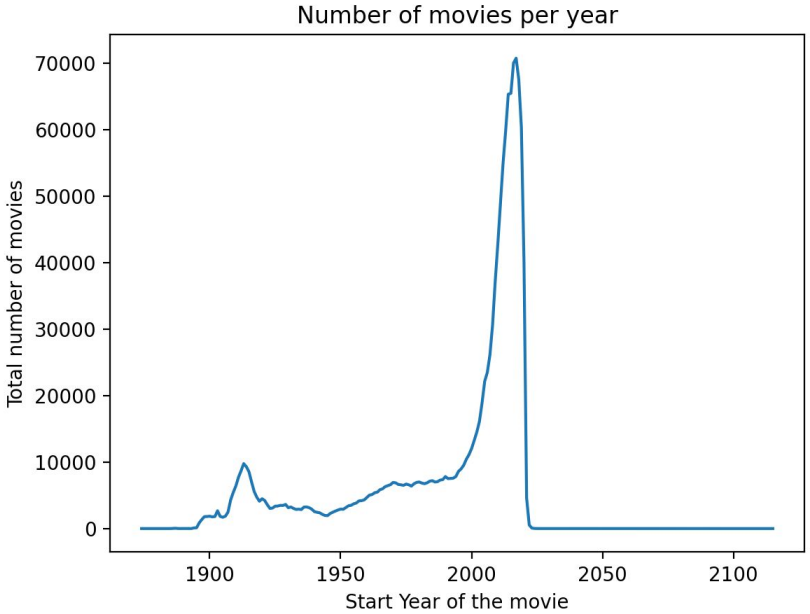
Visualization

- Visualization techniques are often used to understand huge data.
- There are several forms of representing data - Bar Graphs, Histograms, Pie Charts, Scatter plots etc.
- For this project, we have used the following charts:
    - Scatter plot
    - Time Series plot
    - Bar Graph

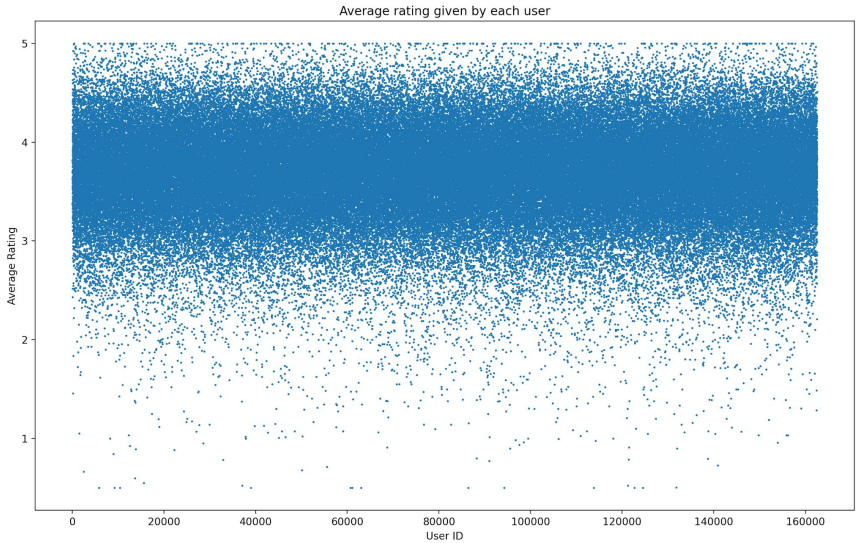
Total number of movies per genre



Total number of Movies per year



Average rating per user

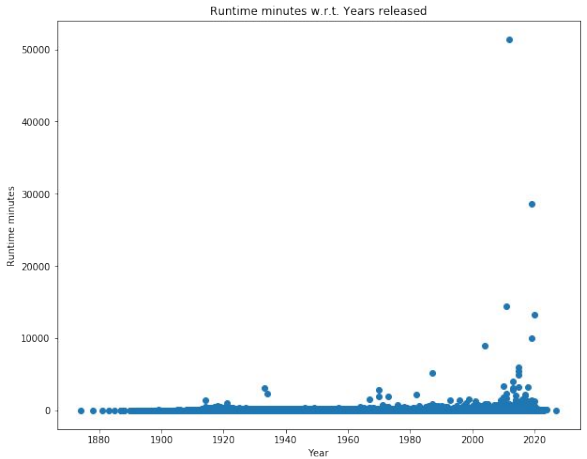


Pairwise comparison

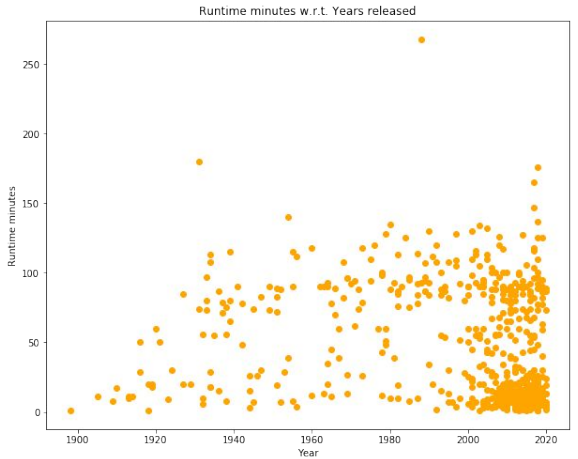
- Pairwise comparison is a process of comparing different attributes or entities with one another to find interesting insights. It can help us understand the correlation and relationships between different attributes as well.

Startyear v/s Runtime minutes

- The relationship between attributes startYear and runtimeMinutes was studied using a scatter plot and outliers were detected.



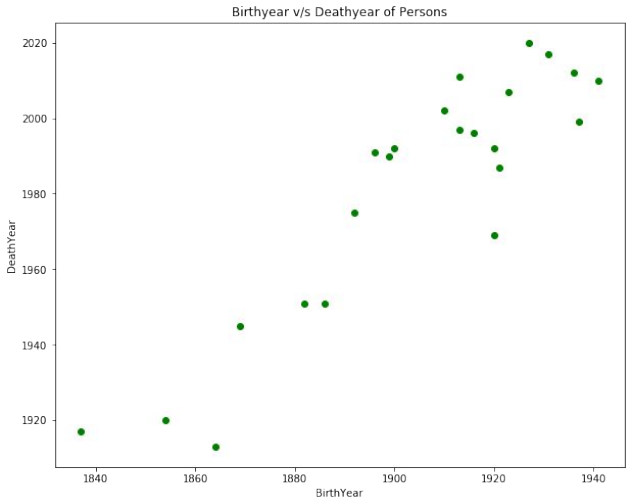
Scatter plot with outliers



Scatter plot with random sampling,

Birth year v/s Death year

- A pairwise comparison between attributes birthYear and deathYear was performed.



Scatter plot

Conclusion and Future work

- We observed:
  - The maximum number of actors observed in a frequent itemset was 6.
  - The number of movies released per year increase every year.
  - Short films genre consists of the most number of movies.
- In future, we would be looking into some other queries for visualization.
- We would also be looking at other movie and ratings based datasets that can be combined with the datasets that we have used - IMDb and MovieLens for extracting further information.
- The datasets can be merged with another dataset with text reviews on which one can perform sentiment analysis using text mining.