A Movie Recommendation System

Vaibhavi Dharashivkar, Udit Wasan, Varsha Venkatachalam
Rochester Institute of Technology, Rochester, New York, USA

Motivation

- In the field of entertainment, a movie recommendation system helps to boost the user's decision-making abilities by identifying movies as per their likings [2, 4].
- For example, movies having same genre as the user's pick or movies that are top voted or that have the highest ranking).
- In order to build such a system, the IMDb data set form Kaggle was chosen as it has genre, ratings and user votes assigned to every movie [3].

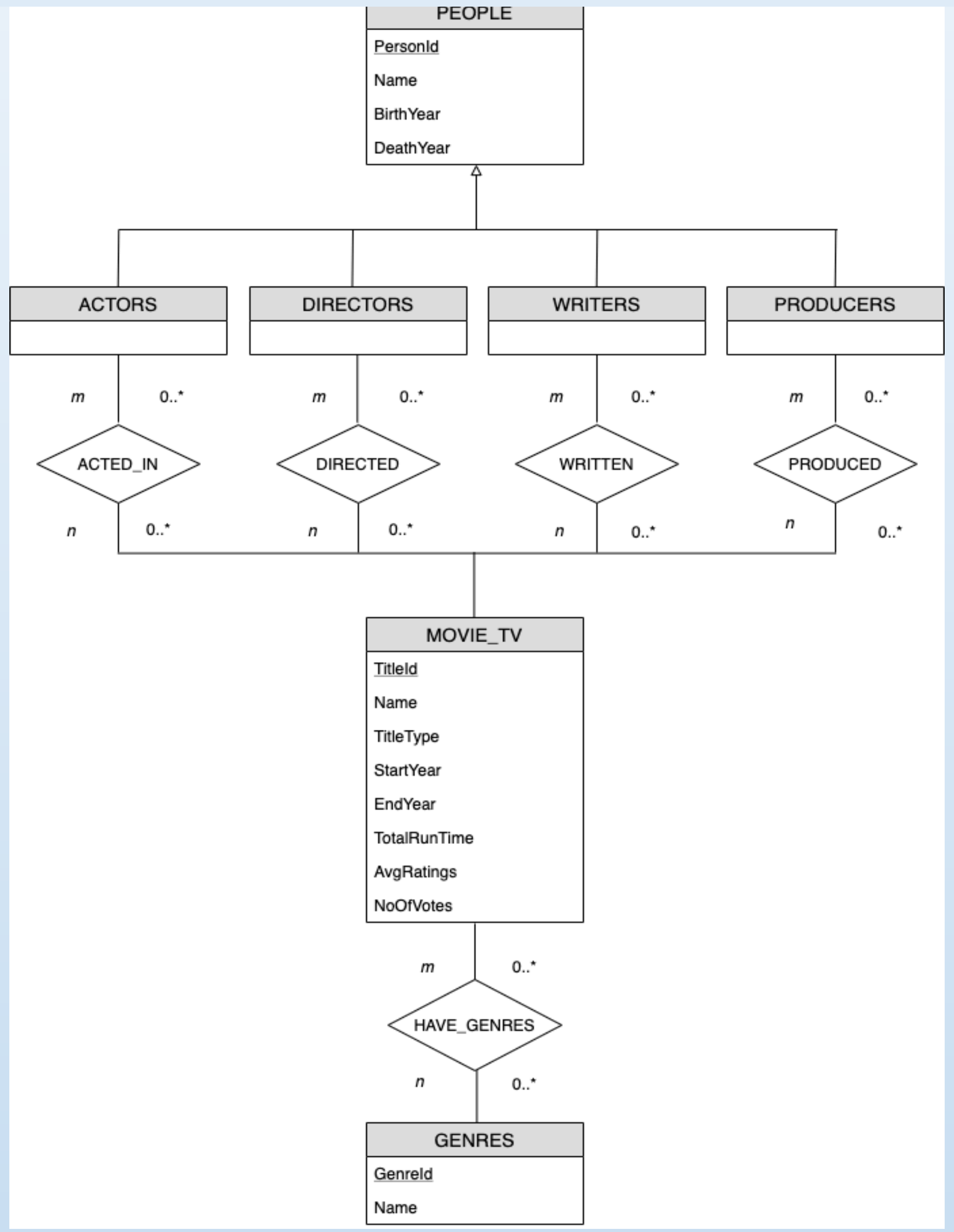
Overall Approach

Considerations:

- For this project, title types movie and TV movie were considered as movies and the title types Tv mini-series, Tv series and Tv special are considered as Tv-shows.
- The records with "Null" in the death year attribute of an individual were considered as still alive.

Filtering to improve data quality:

- The records that do not have any genres associated with them are cleaned out.
- The records that had runtime as either "Null" or 0 were removed.
- Records with average ratings or number of votes as 0 were discarded.
- Records with no birth year, primary profession, name or alphanumeric identifier were eliminated.



The ER Model:

- The figure represents the Entity-Relationship model for the IMDb dataset.
- All the relationships in this model have a many-to-many cardinality [1].
- The actors, directors, writers and producers inherit features from the people table.
- As seen in the figure, these tables are intern connected to the movie_tv table which is then connected to the genres table.

The Workflow Model:

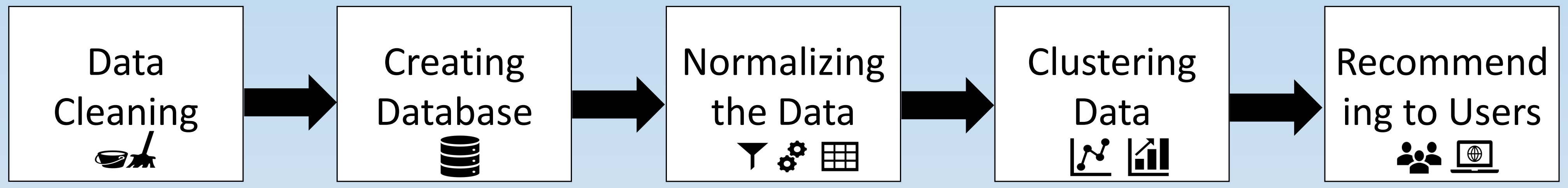
Step 1: Data Cleaning – This step involves the identification and rectification of inconsistent data by means of modifying or deleting inconsistent records.

Step 2: Creating Database – Multiple tables are created by utilizing the data from the respective CSV files.

Step 3: Normalizing the data – The database is modified with respect to NF1 and NF2 normal forms and then indexed to improve the accuracy and consistency of the data while reducing the redundancy [1].

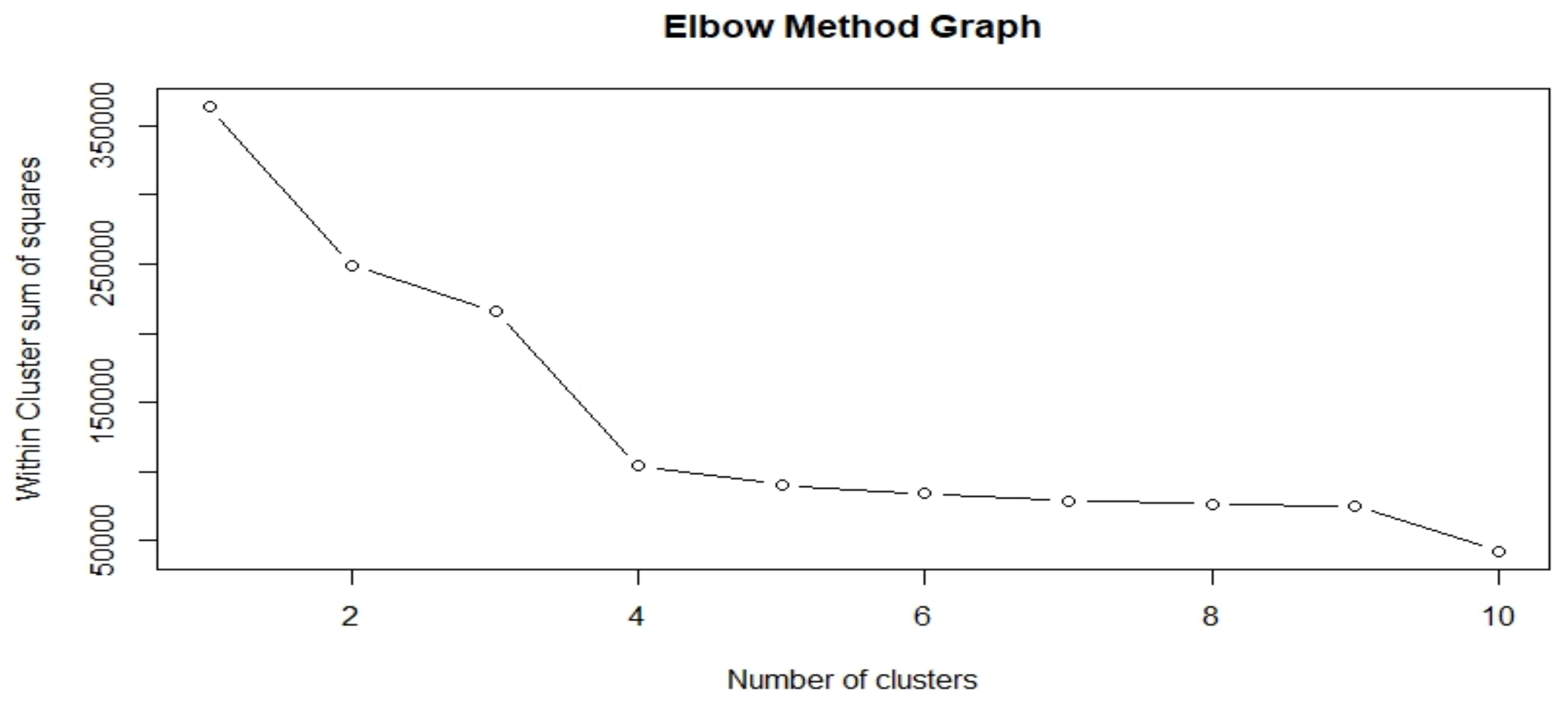
Step 4: Clustering Data – By using the KNN clustering method, the data is divided into five groups based on the number of votes received as well as the average user ratings for each movie.

Step 5: Recommending to users – The clustered data can be accessed to generate lists of common movies as per the user votes and ratings. These lists can be intern be accessed by the users based on the popularity of their searches.

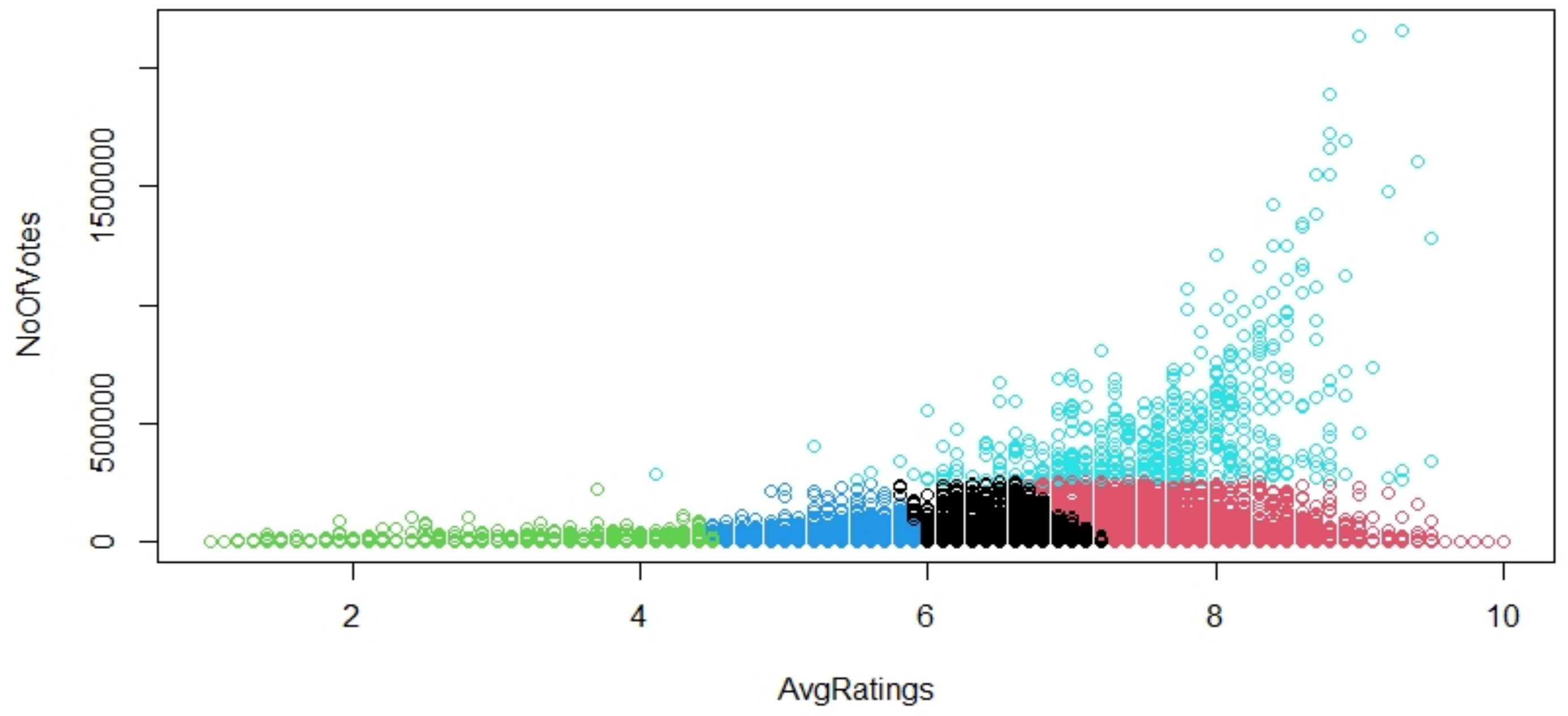

Data Cleaning
Creating Database
Normalizing the Data
Clustering Data
Recommending to Users

Analysis and Conclusion

- The K-means clustering is a centroid-based clustering technique which scales to the dataset with an algorithmic complexity of O(n) liner time.
- In elbow method the percentage of variance is used for plotting a graph as a function of number of clusters where the elbow depicts the number of clusters.



- The five clusters are as shown below with the average ratings between 0-10 and number of votes between 0-2500000.



References

[1] S. Sudarshan Abraham Silberschatz Professor, Henry F. Korth. 2019.*DatabaseSystem Concepts 7th Edition.* McGraw-Hill Education, New York, NY: McGraw-Hill.

[2] R. Ahuja, A. Solanki, and A. Nayyar. 2019. Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor. In *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence).* 263–268.

[3] IMDb. 2019. IMDb Dataset. data retrieved June 2020 from https://www.kaggle.com/ashirwadsangwan/imdb-dataset/.

[4] Aman Gupta Pranal Soni Vishwa Gosalia, Bhavesh Chatnani. Feb-2018.MovieRecommendation System. Technical Report. Maharashtra, India.