# Meta Data


- Presented By
- Anuradha Bhave
- Himanshi Chetwani
- Omkar Vaidya

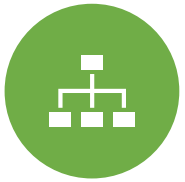
# Introduction

Domain : Movies

Data sets : IMDb and MovieLens

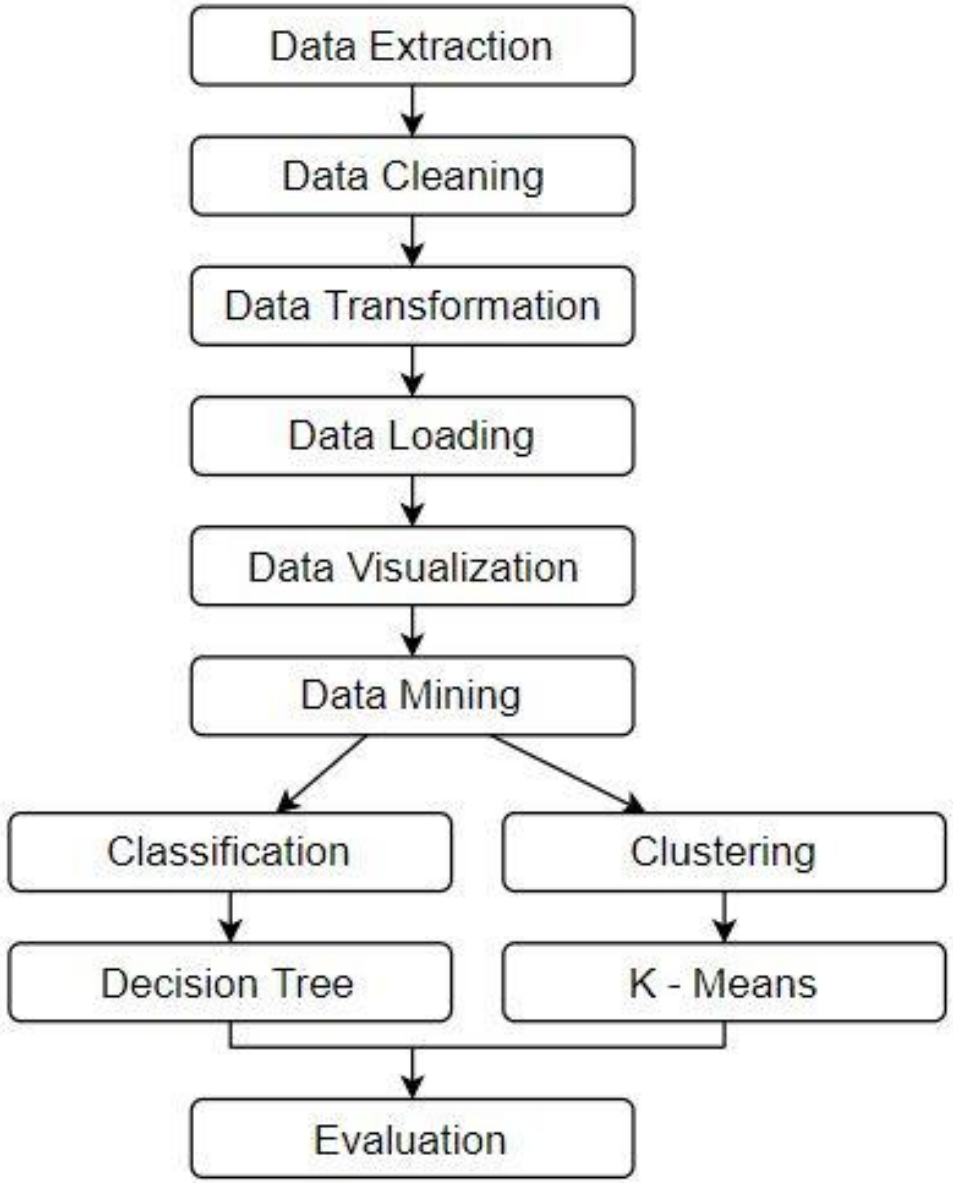
Some movies tend to be a hit, while others are not successful

Aim : Predict the success of a movie by gaining insights about its features.



Data Mining Tasks : Classification and Clustering

# Project Design



# Data sets

IMDb + MovieLens

Final Data set of 0.9 million records and 11 attributes

# Data Handling

Data Collection : Collecting data from the two data sets

Data Merging: Merging data based on title

Data Cleaning: Cleaning data to handle missing values, performing label encoding to transform data

Data Loading: Loading the data in a relational database

# Data Mining Activities

## Classification

Supervised machine learning

Decision Tree Classifier

10 folds cross validation

Predicts success of a movie

Accuracy : ~93%

## Clustering

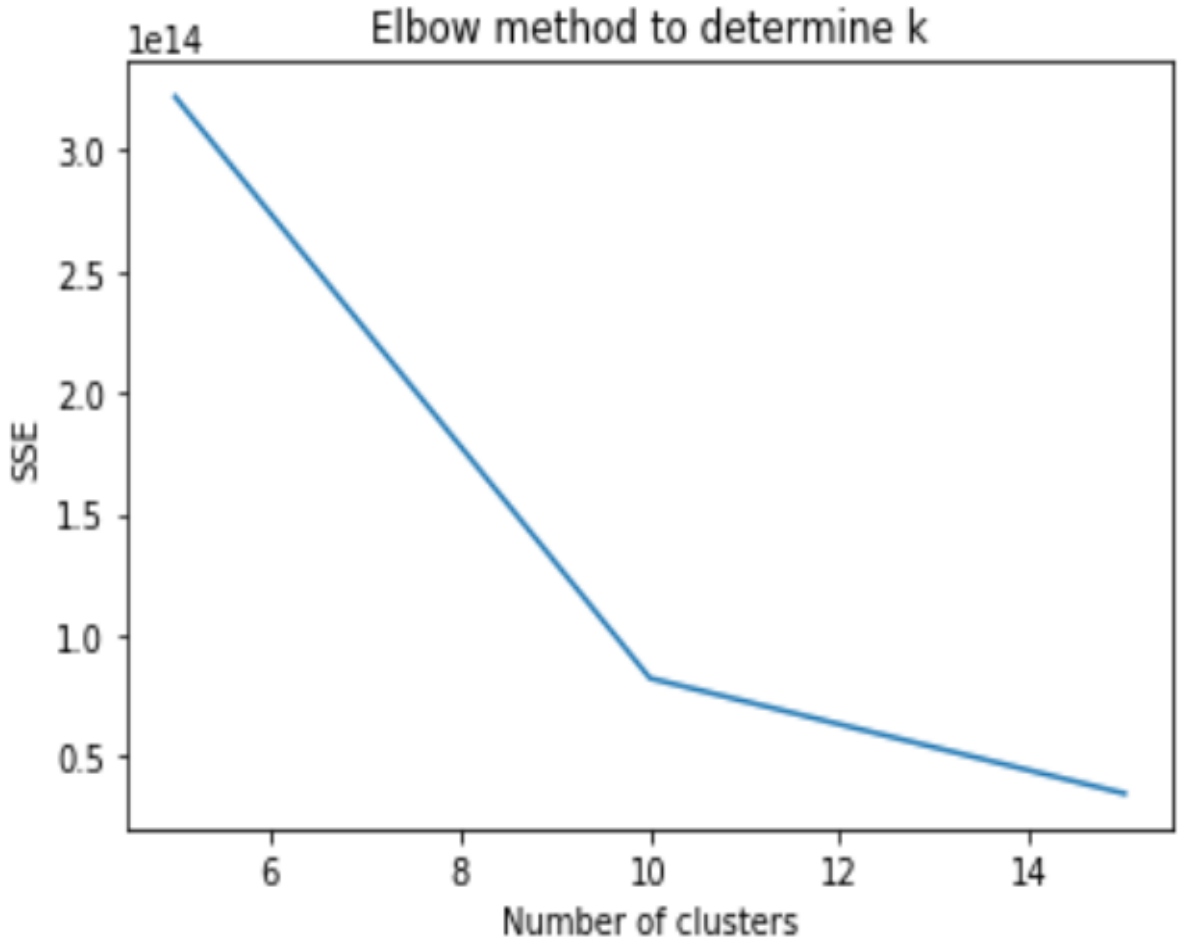
Unsupervised machine learning

K-means Clustering Algorithm

Partitioning Method

Patterns related to movie features among clusters identified

# Determining Optimal Number of Clusters



# Importance of Data Visualization

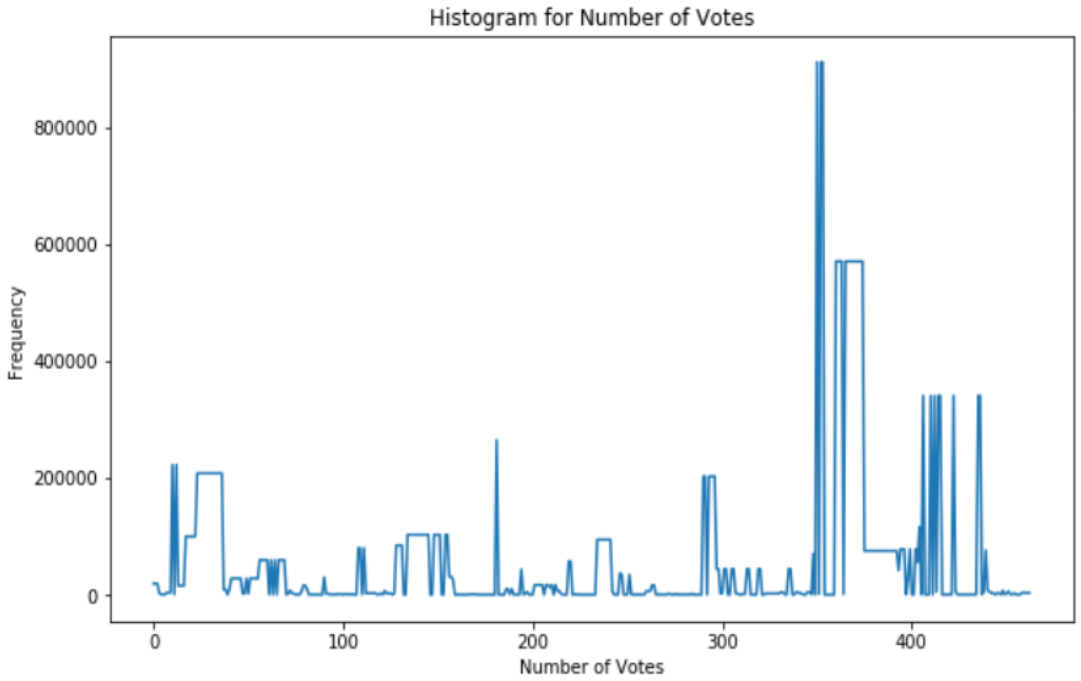
Absorbs information quickly

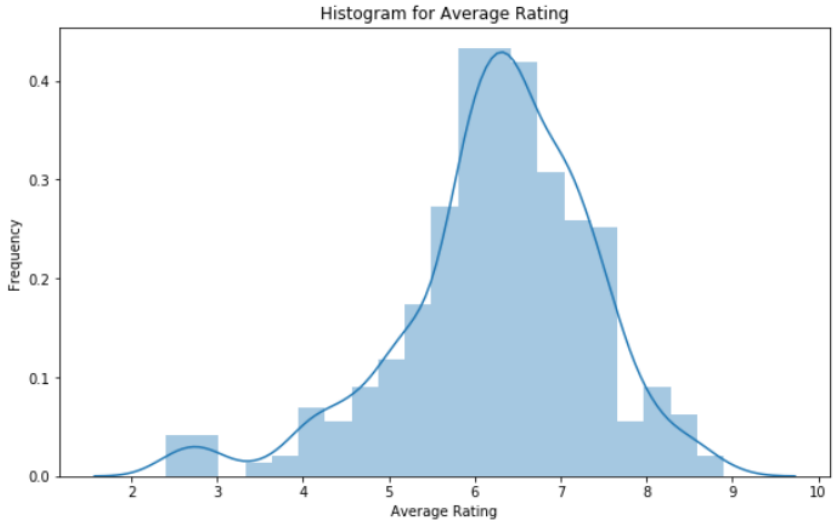
Understanding next steps

Finding correlations between attributes

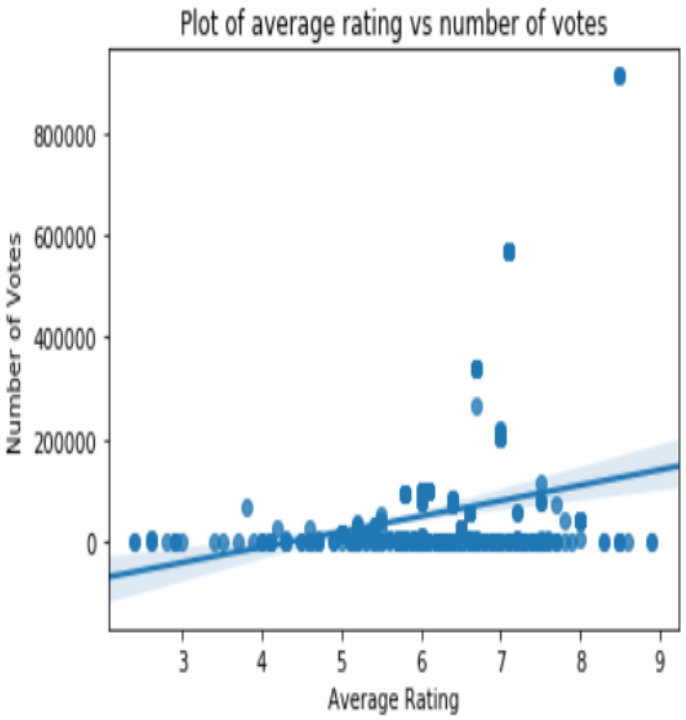
Finding Outliers

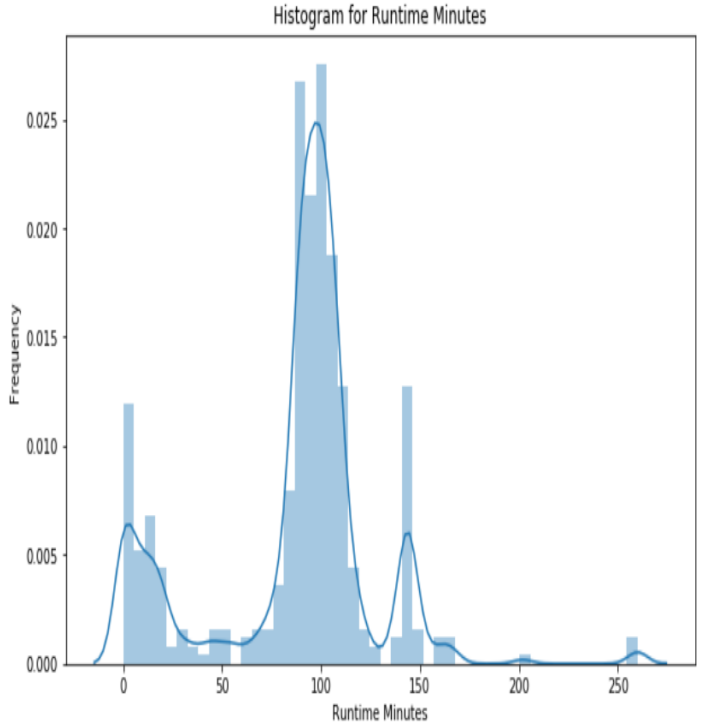
Acting on findings

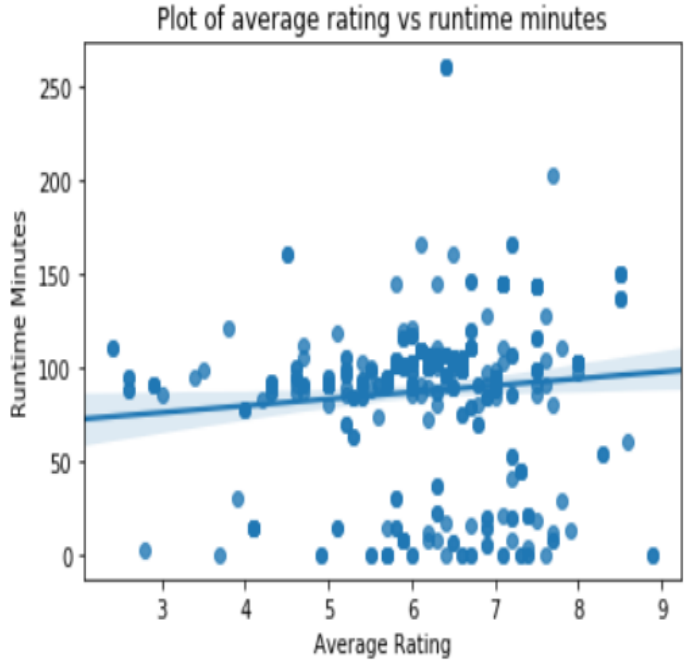



# Data Visualization



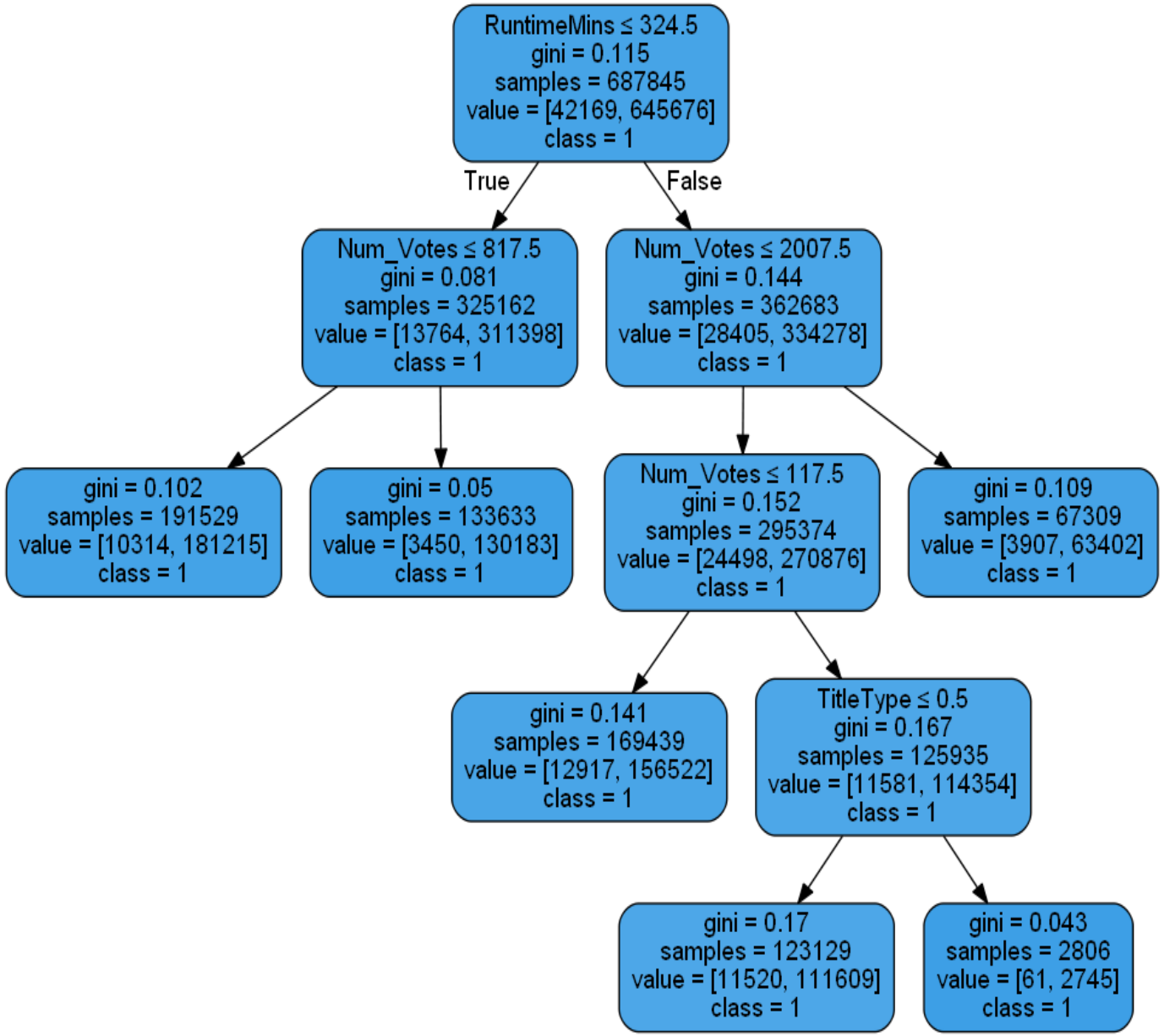

# Data Visualization

# Results

Classification Result :

A pruned decision tree with depth = 3, and 6 leaf nodes.

Accuracy : 93. 82%



```
In [38]: results.Region.value_counts()[1]

Out[38]: Region
         US      229
         GR      190
         ES      153
         CA      148
         FI      130
```

```
In [40]: results.Genre.value_counts()[1]

Out[40]: Genre
         Action,Adventure,Sci-Fi        459
         Adventure,Family,Fantasy       320
         Adventure,Animation,Comedy     207
```

# Results : Clustering

- Allows us to identify characteristics of movies in a cluster
- Movies of cluster 1 are made most frequently in USA, and are of the Action – Adventure genre

# Conclusion

THE MOVIE INDUSTRY IS A PROFITABLE INDUSTRY

PREDICTING THE SUCCESS OF A MOVIE CAN HELP PREDICT ITS MONETARY BUSINESS

PREDICTING THE MOVIE'S FAILURE CAN HELP PRODUCER RECOVER PRODUCTION COSTS

DATA MINING CAN HELP US IN THIS IDENTIFICATION

# Thank You!