MovieLens Recommendation System

Introduction

In this report, the **MovieLens 10M dataset** was used to create a **movie recommendation system algorithm** that can be used to predict how a certain user will rate a certain movie.

The **MovieLens 10M dataset** consists of 10,000,000 ratings of 10,000 movies by 72,000 users on a five-star scale.

The data was pulled directly from the MovieLens website (https://grouplens.org/datasets/movielens/10m/).

The raw dataset was wrangled into a data frame, then split into the *edx* training dataset and the *validation* testing dataset.

The datasets were cleaned up, wrangled, and coerced into a more useable format.

The *edx* dataset was explored and analyzed by plotting the data through the lenses of different potential effects.

An equation for the root mean squared error (RMSE) was defined as the target parameter.

Several models were trained using the *edx* dataset and evaluated on the *validation* dataset, including naive mean, effects, and regularization. The most effective models were then combined.

Using this method, a **movie recommendation system algorithm** with an **RMSE** of **0.863** was developed.

Data Analysis and Model Development

Create the Datasets

The raw datasets were pulled directly from the MovieLens website and saved to a temporary file. From the temporary file, the data was pulled in and coerced into two data frames, the *ratings* data frame, with columns userId, movieId, rating, and timestamp, and the *movies* data frame, with columns movieId, title, and genres. The two data frames were joined together by movieId, creating a new *movielens* data frame with six columns, userId, movieId, rating, timestamp, title, and genres.

***movielens* Dataset**

```
##   userId movieId rating timestamp                          title
## 1      1     122      5 838985046               Boomerang (1992)
## 2      1     185      5 838983525                Net, The (1995)
## 3      1     231      5 838983392           Dumb & Dumber (1994)
## 4      1     292      5 838983421                Outbreak (1995)
## 5      1     316      5 838983392                Stargate (1994)
## 6      1     329      5 838983392 Star Trek: Generations (1994)
##                            genres
## 1                  Comedy|Romance
## 2           Action|Crime|Thriller
## 3                          Comedy
## 4  Action|Drama|Sci-Fi|Thriller
## 5       Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
```

The *movielens* dataset was then split into two datasets, the *edx* training dataset consisting of 90% of the data and the *temp* dataset consisting of the remaining 10% of the data. Movies that only appear in the *temp* dataset were removed, creating the *validation* testing dataset. Those removed movies were then added to the *edx* dataset.

***edx* Dataset**

```
##   userId movieId rating timestamp                         title
## 1      1     122      5 838985046              Boomerang (1992)
## 2      1     185      5 838983525               Net, The (1995)
## 4      1     292      5 838983421               Outbreak (1995)
## 5      1     316      5 838983392               Stargate (1994)
## 6      1     329      5 838983392 Star Trek: Generations (1994)
## 7      1     355      5 838984474       Flintstones, The (1994)
##                          genres
## 1                Comedy|Romance
## 2          Action|Crime|Thriller
## 4  Action|Drama|Sci-Fi|Thriller
## 5       Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
## 7       Children|Comedy|Fantasy
```

***validation* Dataset**

```
##   userId movieId rating timestamp
## 1      1     231      5 838983392
## 2      1     480      5 838983653
## 3      1     586      5 838984068
## 4  10812     151      3 868246450
## 5  10812     858      2 868245645
## 6  10812    1544      3 868245920
##                                                  title
## 1                                  Dumb & Dumber (1994)
## 2                                  Jurassic Park (1993)
## 3                                     Home Alone (1990)
## 4                                        Rob Roy (1995)
## 5                                   Godfather, The (1972)
## 6 Lost World: Jurassic Park, The (Jurassic Park 2) (1997)
##                                   genres
## 1                                 Comedy
## 2        Action|Adventure|Sci-Fi|Thriller
## 3                          Children|Comedy
## 4                Action|Drama|Romance|War
## 5                             Crime|Drama
## 6 Action|Adventure|Horror|Sci-Fi|Thriller
```

Clean the Datasets

Looking at the *edx* dataset again, there is some data cleaning that can be done to make the data easier to visualize and analyze.

***edx* Dataset**

```
##   userId movieId rating timestamp                         title
## 1      1     122      5 838985046              Boomerang (1992)
```

```
## 2      1     185      5 838983525               Net, The (1995)
## 4      1     292      5 838983421               Outbreak (1995)
## 5      1     316      5 838983392               Stargate (1994)
## 6      1     329      5 838983392 Star Trek: Generations (1994)
## 7      1     355      5 838984474       Flintstones, The (1994)
##                          genres
## 1                Comedy|Romance
## 2          Action|Crime|Thriller
## 4  Action|Drama|Sci-Fi|Thriller
## 5        Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
## 7       Children|Comedy|Fantasy
```

The timestamp column is the time the review was submitted, formatted as the number of seconds since January 1, 1970. It can be converted to a date_time data type.

The movie release year is included in title column. It can be extracted, added as the new column year, and converted to a numeric data type.

The columns timestamp and year can be used to calculate the number of years between the movie's release year and the year the movie was reviewed and create a new column yearsbetween.

Some movies fall into more than one genre in the genres column. Reviews of movies with more than one genre can be separated out by genre into multiple duplicate reviews with one genre per review.

Cleaned *edx* Dataset

```
##    userId movieId rating           timestamp                       title
## 1       1     122      5 1996-08-02 11:24:06              Boomerang (1992)
## 2       1     122      5 1996-08-02 11:24:06              Boomerang (1992)
## 3       1     185      5 1996-08-02 10:58:45               Net, The (1995)
## 4       1     185      5 1996-08-02 10:58:45               Net, The (1995)
## 5       1     185      5 1996-08-02 10:58:45               Net, The (1995)
## 6       1     292      5 1996-08-02 10:57:01               Outbreak (1995)
## 7       1     292      5 1996-08-02 10:57:01               Outbreak (1995)
## 8       1     292      5 1996-08-02 10:57:01               Outbreak (1995)
## 9       1     292      5 1996-08-02 10:57:01               Outbreak (1995)
## 10      1     316      5 1996-08-02 10:56:32               Stargate (1994)
## 11      1     316      5 1996-08-02 10:56:32               Stargate (1994)
## 12      1     316      5 1996-08-02 10:56:32               Stargate (1994)
## 13      1     329      5 1996-08-02 10:56:32 Star Trek: Generations (1994)
## 14      1     329      5 1996-08-02 10:56:32 Star Trek: Generations (1994)
## 15      1     329      5 1996-08-02 10:56:32 Star Trek: Generations (1994)
## 16      1     329      5 1996-08-02 10:56:32 Star Trek: Generations (1994)
## 17      1     355      5 1996-08-02 11:14:34       Flintstones, The (1994)
## 18      1     355      5 1996-08-02 11:14:34       Flintstones, The (1994)
## 19      1     355      5 1996-08-02 11:14:34       Flintstones, The (1994)
##       genres year yearsbetween
## 1     Comedy 1992            4
## 2    Romance 1992            4
## 3     Action 1995            1
## 4      Crime 1995            1
## 5   Thriller 1995            1
## 6     Action 1995            1
## 7      Drama 1995            1
## 8     Sci-Fi 1995            1
```

```
## 9   Thriller 1995           1
## 10    Action 1994           2
## 11 Adventure 1994           2
## 12    Sci-Fi 1994           2
## 13    Action 1994           2
## 14 Adventure 1994           2
## 15     Drama 1994           2
## 16    Sci-Fi 1994           2
## 17  Children 1994           2
## 18    Comedy 1994           2
## 19   Fantasy 1994           2
```

The same steps were carried out on the *validation* dataset.

Cursory Data Visualizations and Analysis

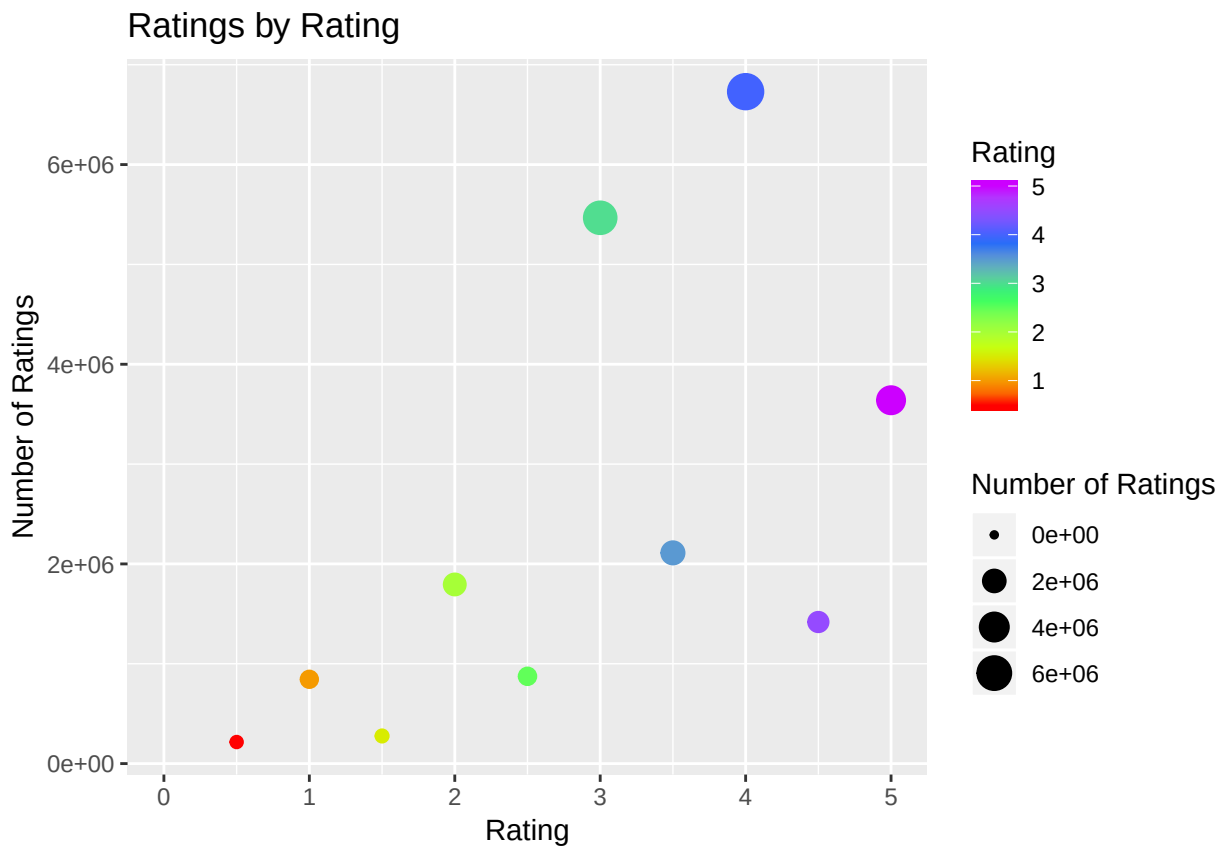
All visualizations and analyses were performed with the *edx* training dataset.

The average rating is 3.53 stars. The 4 stars is the median rating.

Grouping the data by rating shows that four stars is most common rating and that full star ratings are given more often than half star ratings.

Ratings

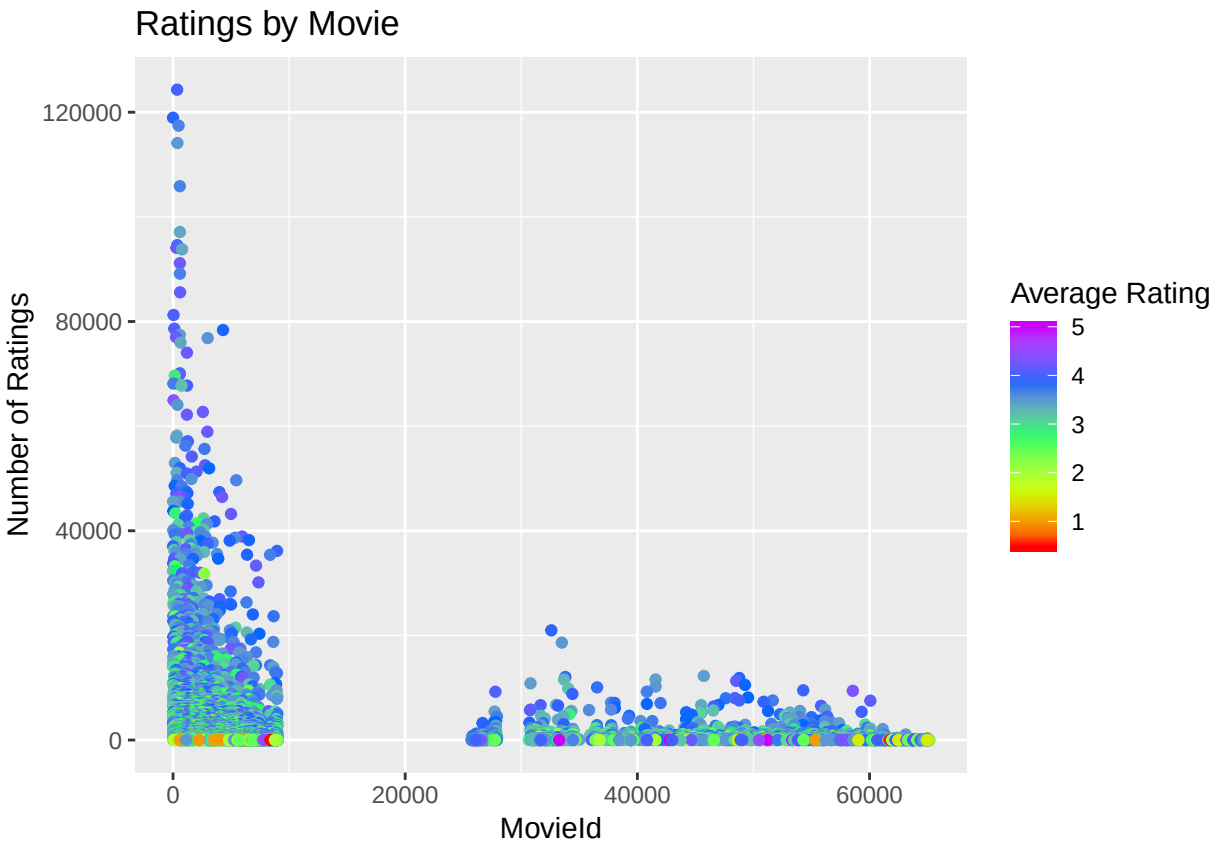
```
## # A tibble: 10 x 2
##    rating num_ratings
##     <dbl>       <int>
##  1    4       6730401
##  2    3       5467061
##  3    5       3639511
##  4    3.5     2110690
##  5    2       1794243
##  6    4.5     1418248
##  7    2.5      874290
##  8    1        844336
##  9    1.5      276711
## 10    0.5      215932
```



Grouping the data by movie shows that in general, movies that are reviewed often have higher average ratings and that there is more variation in average ratings for movies that have few reviews.

Movies

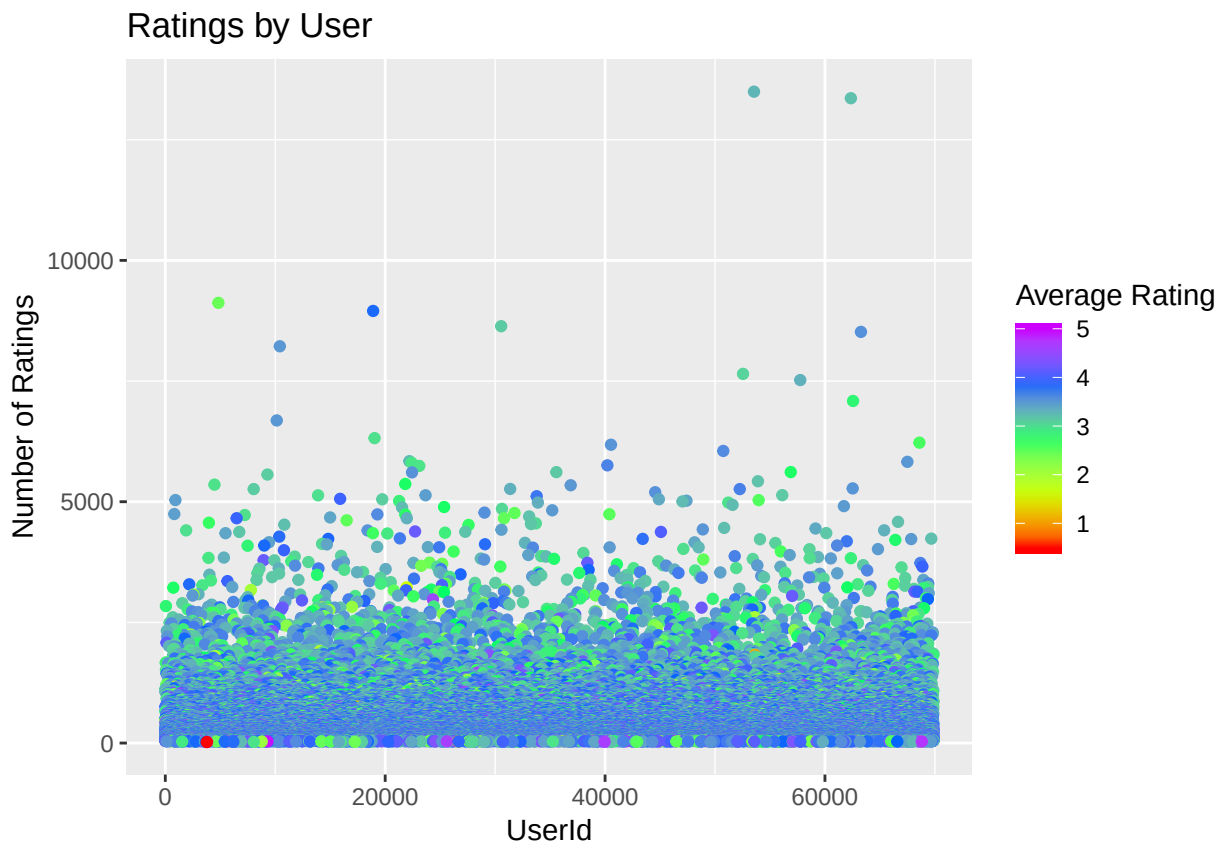
```
##   movieId num_ratings avg_rating
## 1     356      124316       4.01
## 2       1      118950       3.93
## 3     480      117440       3.66
## 4     380      114115        3.5
## 5     ...         ...        ...
## 6   64611           1        3.5
## 7   64897           1          3
## 8   64944           1          3
## 9   64976           1        1.5
```



Grouping the data by user shows that most users give an average rating near the overall average and that there is more variation in average ratings for users that have only given a few ratings, when compared to users that have rated many movies.

Users

```
##   userId num_ratings avg_rating
## 1  53547       13496       3.29
## 2  62358       13360       3.21
## 3   4831        9121       2.46
## 4  18905        8953       3.85
## 5    ...         ...        ...
## 6   8049          25       2.96
## 7  49602          23       3.52
## 8   3801          22          5
## 9   3781          21        0.5
```

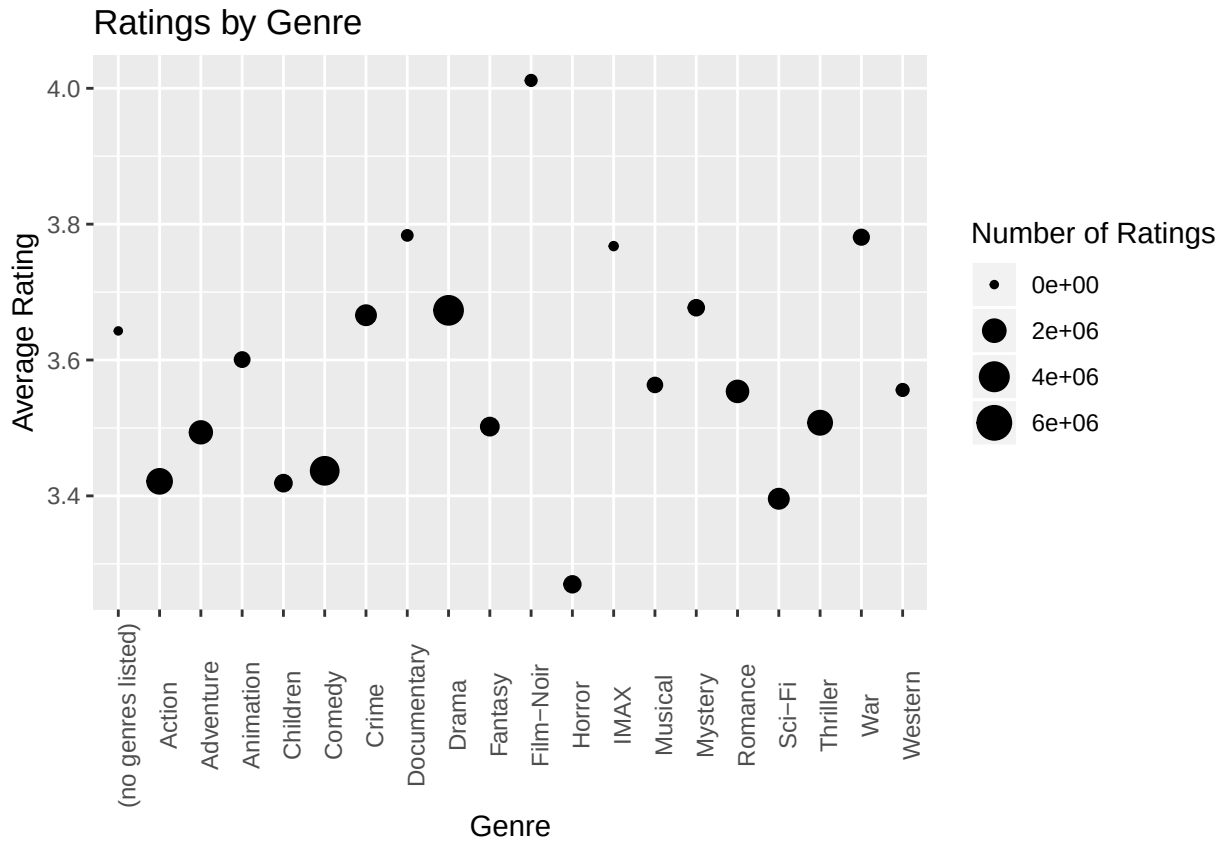


Grouping the data by genre shows that the most common genres are Drama, Comedy, and Action and that the best rated genres, like Film-Noir, War, and Documentary have fewer movies and ratings.

Genres

```
## # A tibble: 20 x 3
##    genres      num_ratings avg_rating
##    <chr>             <int>      <dbl>
##  1 Drama           3910127       3.67
##  2 Comedy          3540930       3.44
##  3 Action          2560545       3.42
##  4 Thriller        2325899       3.51
##  5 Adventure       1908892       3.49
##  6 Romance         1712100       3.55
##  7 Sci-Fi          1341183       3.40
##  8 Crime           1327715       3.67
##  9 Fantasy          925637       3.50
## 10 Children         737994       3.42
## 11 Horror           691485       3.27
## 12 Mystery          568332       3.68
## 13 War              511147       3.78
## 14 Animation        467168       3.60
## 15 Musical          433080       3.56
## 16 Western          189394       3.56
## 17 Film-Noir        118541       4.01
## 18 Documentary       93066       3.78
## 19 IMAX               8181       3.77
```

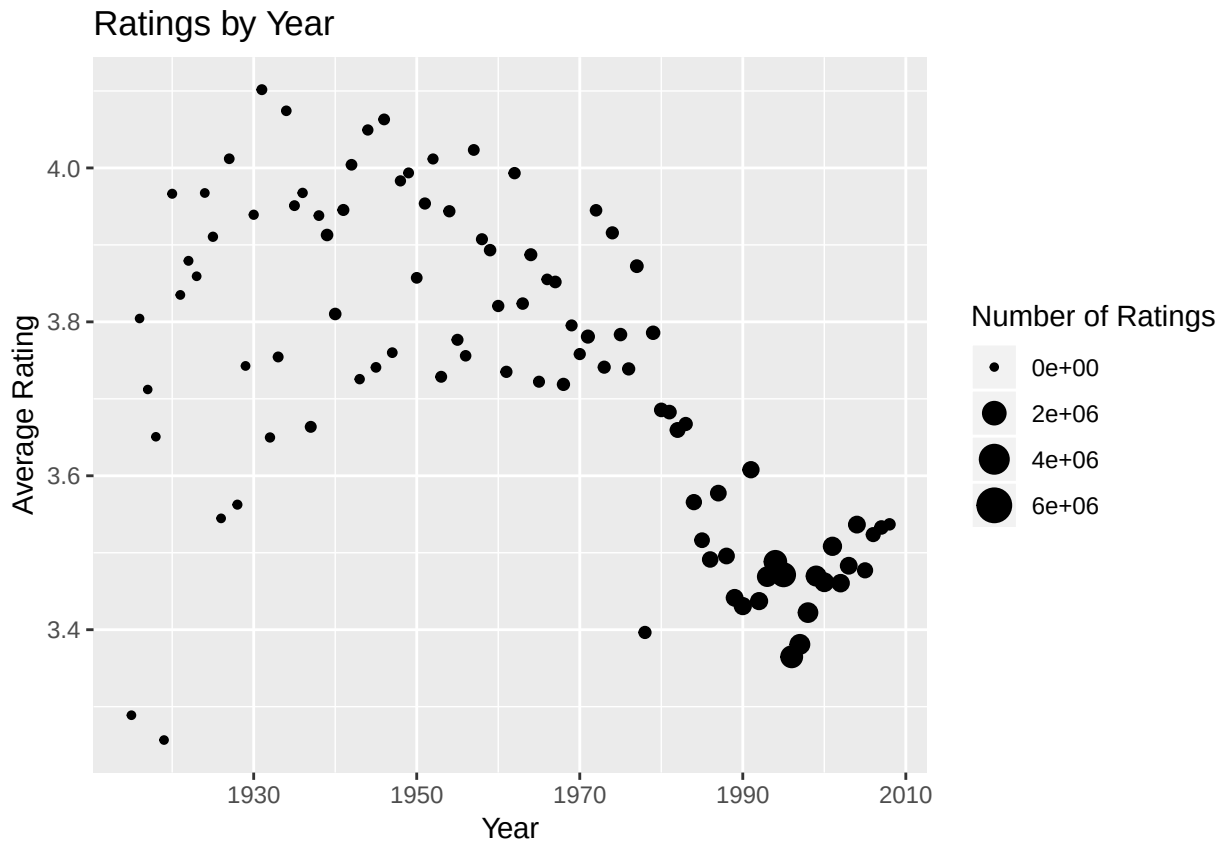
```
## 20 (no genres listed)          7       3.64
```



Grouping the data by movie release year shows that pre-1980 years are better rated than post-1980 years and that movies released in recent years have received more ratings.

Release Year

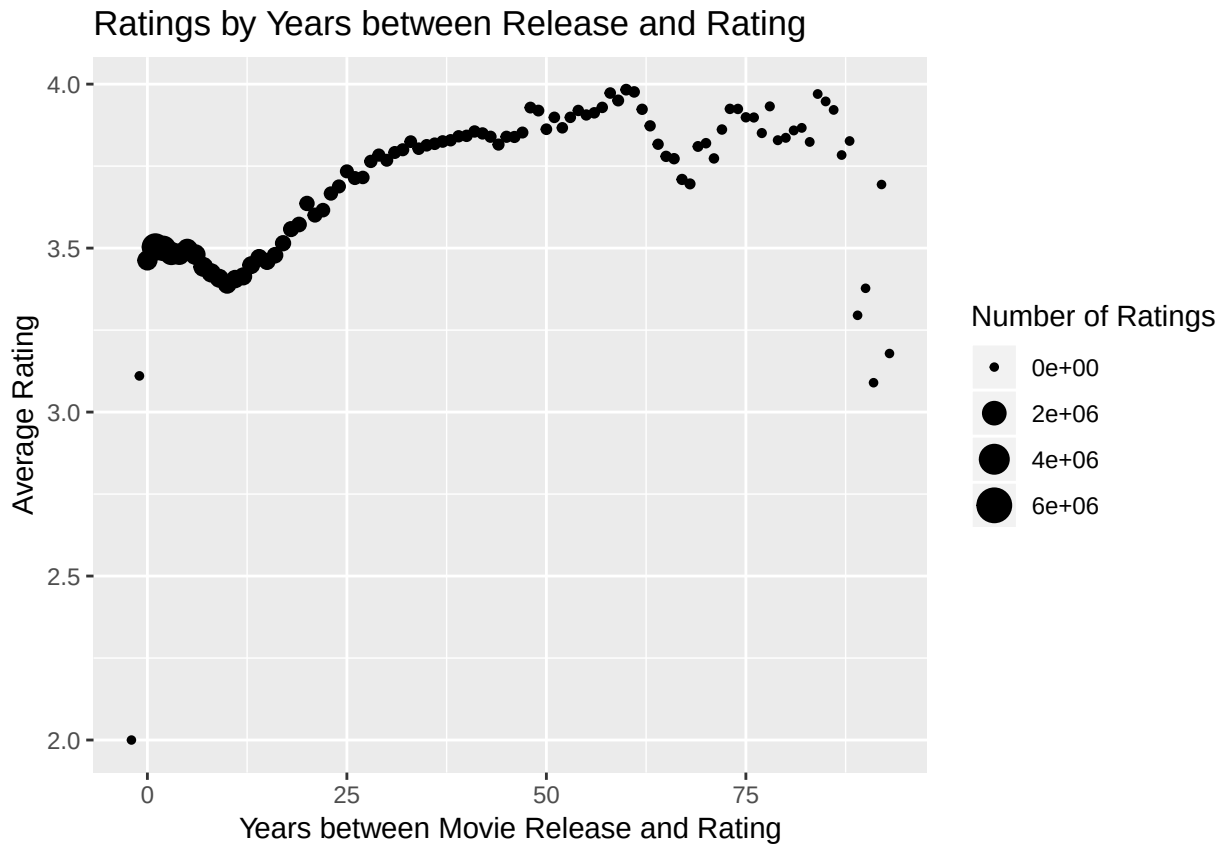
```
##   year num_ratings avg_rating
## 1 1995     2083655       3.47
## 2 1994     1732877       3.49
## 3 1996     1561069       3.36
## 4 1999     1158834       3.47
## 5  ...         ...        ...
## 6 1919         339       3.26
## 7 1916          92        3.8
## 8 1918          73       3.65
## 9 1917          33       3.71
```



Grouping the data by the number of years between release and review shows that movies that are reviewed before they have been released tend to be reviewed quite poorly and that generally, the more time between a movie's release and the time it was reviewed, the higher the rating tends to be.

Years between Release and Review

```
##   yearsbetween num_ratings avg_rating
## 1            1     2784854        3.5
## 2            2     2214034        3.5
## 3            3     1636466       3.48
## 4            4     1270373       3.48
## 5          ...         ...        ...
## 6           91          67       3.09
## 7           92          49       3.69
## 8           93          28       3.18
## 9           -2           3          2
```



Defining RMSE

The goal of this project is to develop an algorithm with the lowest possible residual mean squared error (RMSE). RMSE is defined as the error that the algorithm makes when predicting a rating, or:

$$\sqrt{\frac{1}{N}\sum_{e}(\hat{y}_e - y_e)^2}$$

where N is the total number of user/movie ratings, $\hat{y}_e$ is the predicted rating for a particular review given effects e, and y_e is the actual rating for a particular review given effects e.

An RMSE of 1 would mean that on average, the rating that the algorithm predicted is one star off the actual rating.

Modeling Approach

A Simple Model - Average

The simplest model predicts the same rating for each review, regardless of effects like movie, user, genre, etc. This model can be defined as:

$$Y = \mu + \epsilon$$

where Y is the outcome (predicted rating), μ is the average rating, and ϵ is the error.

The **RMSE** of the **Average** model is **1.053**.

Introducing Effects

Introducing effects allows the model to take variability into account. Looking at the visualizations above, for example, some movies are, on average, rated higher than others and certain genres tend to receive lower average ratings than others. The effects model can be defined as:

$$Y = \mu + e_a + \epsilon$$

where e_a is the effect term of effect a.

For modeling purposes, the least square estimate of e_a is the average of $Y_a - \mu$ for each instance of effect a.

Based on the above visualizations, movie, user, genre, year released, and years between release and review effects were all introduced to the model.

Movie Effect

The **Average + Movie Effect** model is defined as

$$Y = \mu + e_m + \epsilon$$

where e_m is the effect term for movie m.

The **RMSE** of the **Average + Movie Effect** model is **0.941**.

User Effect

The **Average + User Effect** model is defined as

$$Y = \mu + e_u + \epsilon$$

where e_u is the effect term for user u.

The **RMSE** of the **Average + User Effect** model is **0.973**.

Genre Effect

The **Average + Genre Effect** model is defined as

$$Y = \mu + e_g + \epsilon$$

where e_g is the effect term for genre g.

The **RMSE** of the **Average + Genre Effect** model is **1.046**.

Year Effect

The **Average + Year Effect** model is defined as

$$Y = \mu + e_y + \epsilon$$

where e_y is the effect term for release year y.

The **RMSE** of the **Average + Year Effect** model is **1.042**.

Years between Effect

The **Average + Years between Effect** model is defined as

$$Y = \mu + e_y b + \epsilon$$

where $e_y b$ is the effect term for years between the movie's release and review yb.

The **RMSE** of the **Average + Years between Effect** model is **1.045**.

Introducing Regularization

Looking at the visualizations above again, there is a lot of variation in the number of ratings that different movies receive, different users give, etc. Regularization will introduce a penalized term that will have a great effect on large predicted ratings stemming from small group sizes while having little effect on predicted ratings stemming from large group sizes.

$$e_a = \frac{\sum_1^{n_a}(Y_a - \mu)}{n_a + \lambda_a}$$

where n_a is the number of ratings for effect a, Y_a is the average rating for effect a, and λ_a is the penalization term for effect a.

Movie Regularization

The **Average + Movie Effect + Regularization** model is defined as

$$Y = \mu + e_m + \epsilon$$

where

$$e_m = \frac{\sum_1^{n_m}(Y_m - \mu)}{n_m + \lambda_m}$$

The **RMSE** of the **Average + Movie Effect + Regularization** model is **0.941**, which is no improvement over the non-regularized model.

Results - The Best Model

Looking that the models described above, only two of them, **Movie Effect** and **User Effect** made significant improvements to the **Average** model.

```
## # A tibble: 7 x 2
##   model                                rmse
##   <chr>                               <dbl>
## 1 Average                             1.05
## 2 Average + Movie Effect              0.941
## 3 Average + User Effect               0.973
## 4 Average + Genre Effect              1.05
## 5 Average + Year Effect               1.04
## 6 Average + Years between Effect      1.04
## 7 Average + Movie Effect + Regularization 0.941
```

By combining these two effects, the model should become more accurate.

The **Average + Movie + User Effects** model is defined as

$$Y = \mu + e_m + e_u + \epsilon$$

Best Effects Model

```
## # A tibble: 1 x 2
##   model                          rmse
##   <chr>                         <dbl>
## 1 Average + Movie + User Effects 0.863
```

The **RMSE** of the **Average + Movie + User Effect** model is **0.863**.

Conclusions

After visually analyzing and examining the data and testing several models, an algorithm to predict movie ratings with an **RMSE** of **0.863** was developed.