MovieLens Project Kadiatou KABA

Kadiatou Kaba

12/16/2020

Introduction

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

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

It 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. Then, some descriptive analysis were done on the *edx* dataset.

In order to aim the objective of the report, an equation for the root mean squared error (RMSE) was defined as the target parameter.

Several models were trained using the *edx* dataset such as naive mean and effects. Then, these models were evaluated on the *validation* dataset. The most effective models were therefore combined to obtain the final model.

Using the method below, 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)
## 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
```

validation Dataset

```
##   userId movieId rating  timestamp
## 1      1     588    5.0  838983339
## 2  10812    1210    4.0  868245644
## 3  10812    1544    3.0  868245920
## 4  21652     151    4.5 1133571026
## 5  21652    1288    3.0 1133571035
## 6  21652    5299    3.0 1164885617
##                                                  title
## 1                                      Aladdin (1992)
```

```
## 2       Star Wars: Episode VI - Return of the Jedi (1983)
## 3 Lost World: Jurassic Park, The (Jurassic Park 2) (1997)
## 4                                          Rob Roy (1995)
## 5                               This Is Spinal Tap (1984)
## 6                         My Big Fat Greek Wedding (2002)
##                                       genres
## 1 Adventure|Animation|Children|Comedy|Musical
## 2                     Action|Adventure|Sci-Fi
## 3     Action|Adventure|Horror|Sci-Fi|Thriller
## 4                    Action|Drama|Romance|War
## 5                              Comedy|Musical
## 6                              Comedy|Romance
```

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 (on data types especially).

edx 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 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.

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

```
## # A tibble: 19 x 7
##    userId movieId rating timestamp           title                 genres
year
##     <dbl>   <dbl>  <dbl> <dttm>              <chr>                 <chr>
<dbl>
```

```
##  1      1     122      5 1996-08-02 11:24:06 Boomerang (1992)      Comedy
1992
##  2      1     122      5 1996-08-02 11:24:06 Boomerang (1992)      Romance
1992
##  3      1     185      5 1996-08-02 10:58:45 Net, The (1995)       Action
1995
##  4      1     185      5 1996-08-02 10:58:45 Net, The (1995)       Crime
1995
##  5      1     185      5 1996-08-02 10:58:45 Net, The (1995)       Thrill~
1995
##  6      1     231      5 1996-08-02 10:56:32 Dumb & Dumber (1994)  Comedy
1994
##  7      1     292      5 1996-08-02 10:57:01 Outbreak (1995)       Action
1995
##  8      1     292      5 1996-08-02 10:57:01 Outbreak (1995)       Drama
1995
##  9      1     292      5 1996-08-02 10:57:01 Outbreak (1995)       Sci-Fi
1995
## 10      1     292      5 1996-08-02 10:57:01 Outbreak (1995)       Thrill~
1995
## 11      1     316      5 1996-08-02 10:56:32 Stargate (1994)       Action
1994
## 12      1     316      5 1996-08-02 10:56:32 Stargate (1994)       Advent~
1994
## 13      1     316      5 1996-08-02 10:56:32 Stargate (1994)       Sci-Fi
1994
## 14      1     329      5 1996-08-02 10:56:32 Star Trek: Generatio~ Action
1994
## 15      1     329      5 1996-08-02 10:56:32 Star Trek: Generatio~ Advent~
1994
## 16      1     329      5 1996-08-02 10:56:32 Star Trek: Generatio~ Drama
1994
## 17      1     329      5 1996-08-02 10:56:32 Star Trek: Generatio~ Sci-Fi
1994
## 18      1     355      5 1996-08-02 11:14:34 Flintstones, The (19~ Childr~
1994
## 19      1     355      5 1996-08-02 11:14:34 Flintstones, The (19~ Comedy
1994
```

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

Cursory Data Visualizations and Analysis

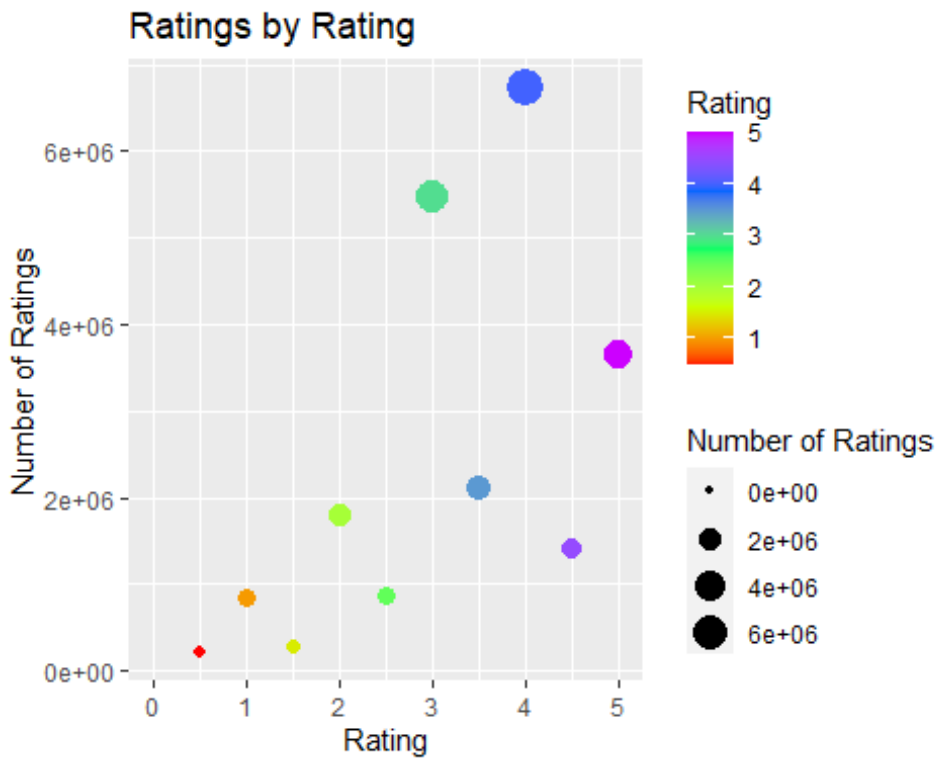
All visualizations and analyses were performed with the *edx* training dataset.

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

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       6730156
##  2    3       5466754
##  3    5       3639299
##  4    3.5     2112391
##  5    2       1792891
##  6    4.5     1416963
##  7    2.5      873585
##  8    1        844605
##  9    1.5      276775
## 10    0.5      216188
```

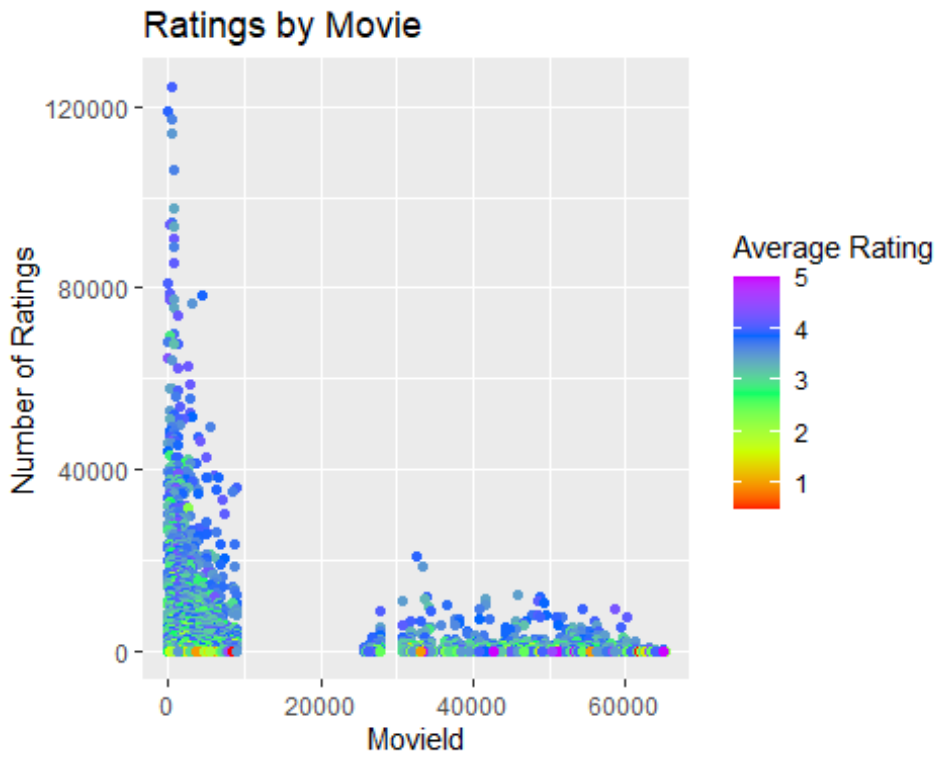


Grouping the data by movie shows that in general, movies that are reviewed often have higher average ratings and that there is more variation in average ratings for movies that have few reviews.

Movies

```
##   movieId num_ratings avg_rating
## 1     356      124304       4.01
## 2       1      119130       3.93
## 3     480      117164       3.66
## 4     380      113930        3.5
## 5     ...         ...        ...
## 6   64897           1          3
```

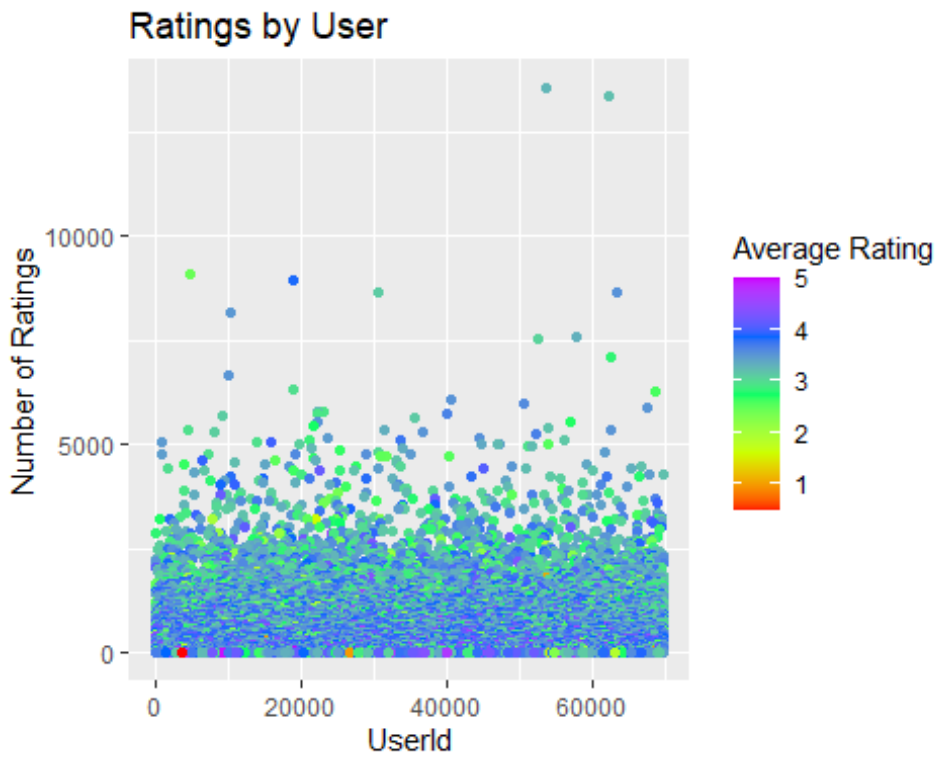
```
## 7   64944           1           3
## 8   64976           1         1.5
## 9   65001           1           5
```



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       13545       3.29
## 2  62358       13371       3.21
## 3   4831        9090       2.46
## 4  18905        8920       3.85
## 5    ...         ...        ...
## 6  39962          25       4.72
## 7   3801          24          5
## 8  17886          24       3.33
## 9   3781          22        0.5
```

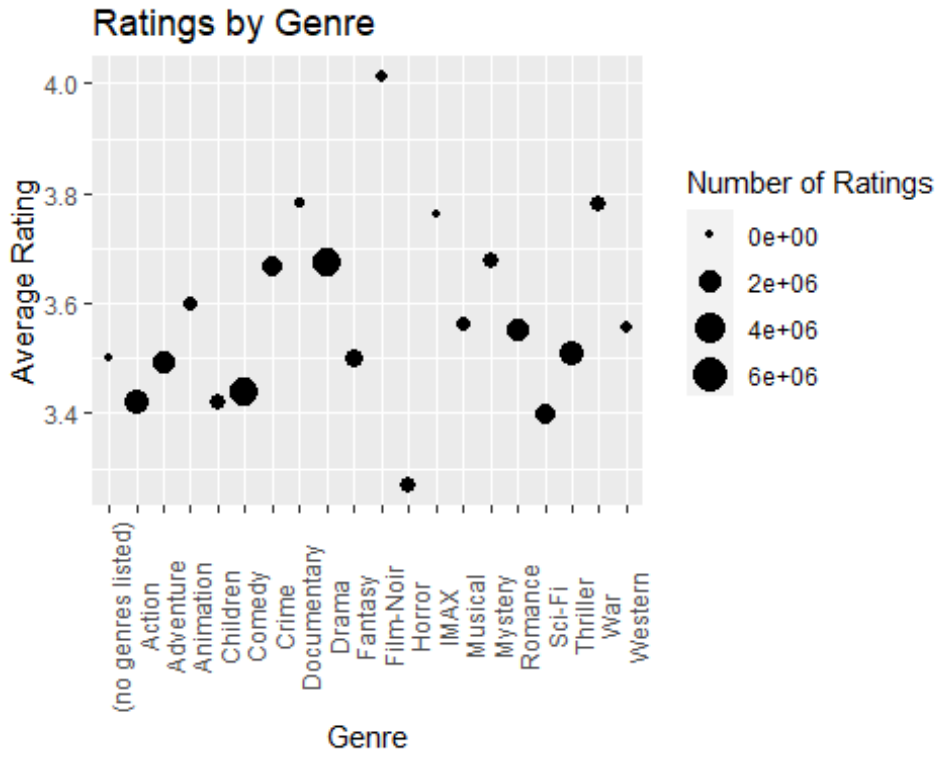


Grouping the data by genre shows that the most common genres are Drama, Comedy, and Action and that the best rated genres, like Film-Noir, War, and Documentary have fewer movies and ratings.

Genres

```
## # A tibble: 20 x 3
##    genres              num_ratings avg_rating
##    <chr>                     <int>      <dbl>
##  1 Drama                   3909401       3.67
##  2 Comedy                  3541284       3.44
##  3 Action                  2560649       3.42
##  4 Thriller                2325349       3.51
##  5 Adventure               1908692       3.49
##  6 Romance                 1712232       3.55
##  7 Sci-Fi                  1341750       3.40
##  8 Crime                   1326917       3.67
##  9 Fantasy                  925624       3.50
## 10 Children                 737851       3.42
## 11 Horror                   691407       3.27
## 12 Mystery                  567865       3.68
## 13 War                      511330       3.78
## 14 Animation                467220       3.60
## 15 Musical                  432960       3.56
## 16 Western                  189234       3.56
## 17 Film-Noir                118394       4.01
## 18 Documentary               93252       3.78
```

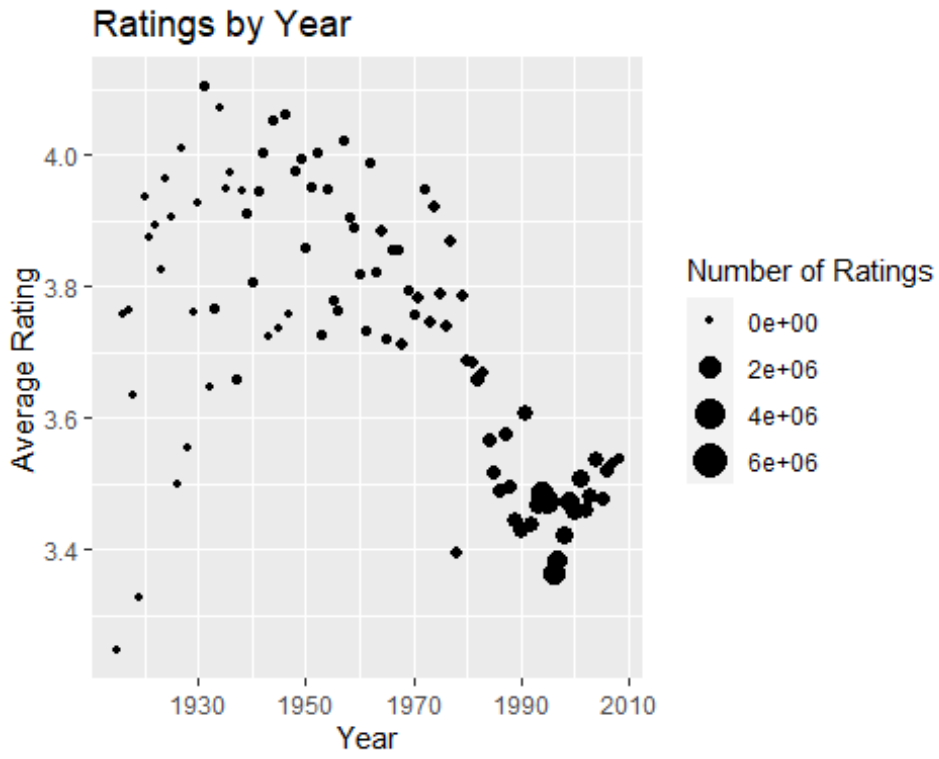
```
## 19 IMAX                        8190        3.76
## 20 (no genres listed)             6        3.5
```



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     2084327       3.47
## 2 1994     1732933       3.49
## 3 1996     1560847       3.36
## 4 1999     1159558       3.47
## 5  ...         ...        ...
## 6 1919         348       3.33
## 7 1916         108       3.76
## 8 1918          70       3.64
## 9 1917          34       3.76
```



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 **1.052**.

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**.

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: 5 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
```

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 0.864
```

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 by defining a model that included effects.

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