Introduction

Recommendation systems are used more and more, as consumers expect suggestions based on their known likes so that they can discover new likes in products, movies, music and other interests. They assist users in finding what they might be interested in based on their preferences and previous interactions. In this report, a movie recommendation system using the MovieLens dataset from HarvardX's Data Science Professional Certificate[3] program will be covered. GroupLens Research is the organization that collected the data sets for this project from their site: (https://movielens.org).

Methods and analysis

There are five steps in the data analysis process that must be completed. In this case, the data must be prepared. The dataset from was downloaded from the MovieLens website and split into two subsets used for training and validation. In this case, we named the training set "edx" and the validation set "validation". For training and testing, the edx set was split again into two subsets. The edx set is trained with the model when it reaches the RMSE goal and the validation set is used for final validation. During data exploration and visualization, charts are crated to understand the data and how it affects the outcome. We observe the mean of observed values, the distribution of ratings, mean movie ratings, movie effect, user effect and number of ratings per movie. We improve the RMSE by including the user and movie effects and applying the regularization parameter for samples that have few ratings.

Results

For the average movie rating model that we generated first, the result was 1.0606506. After accounting for movie effects, we lowered the average to .9437046. In order to lower the RMSE even more, we added both the movie and user effects with the result of .8655329. Finally, we used regularization to penalize samples with few ratings and got the final result of .8649857.

Method	RMSE
Average movie rating model	1.0606506
Movie effect model	0.9437046
Movie and user effect model	0.8655329
Regularized movie and user effect model	0.8649857

Conclusion

In conclusion, we downloaded the data set and prepared it for analysis. We looked for various insights and created a simple model from the mean of the observed ratings. After that, we added the movie and user effects in an attempt to model user behavior. Finally, we conducted regularization that added a penalty for the movies and users with the least number of ratings. We achieved a model with an RMSE of 0.8649857.