A movie recommendation system

Introduction

The dataset is downloaded, tweaked and then split into edx and validation sets in proportions of 90% and 10% respectively. While the validation set has been set aside for testing purposes, i have further split the edx set into train and test sets in proportions of 50% to 50% for intermediate training, testing and cross-validation purposes.

With this dataset of 6 columns and a total of 10,000,054 rows(both the edx and validation sets), i have modelled a recommendation sytem that predicts movie ratings using different predictors. This has been achieved by training the train set using a linear regression approximation model to predict movie ratings of the test set.Since the RMSE obtained between predicted ratings of the test set and the actual ratings of the test set is satisfactory, other models have not been tried.To make the final tests, the same training approach has been adopted to train the whole edx set(not split this time round) before getting predictions of ratings of the validation set, that are then compared with the actual ratings of the validation set using a RMSE function.

Dimensions of the edx set

```
## [1] 9000055       6
```

Dimensions of the validation set

```
## [1] 999999      6
```

Overview of the data

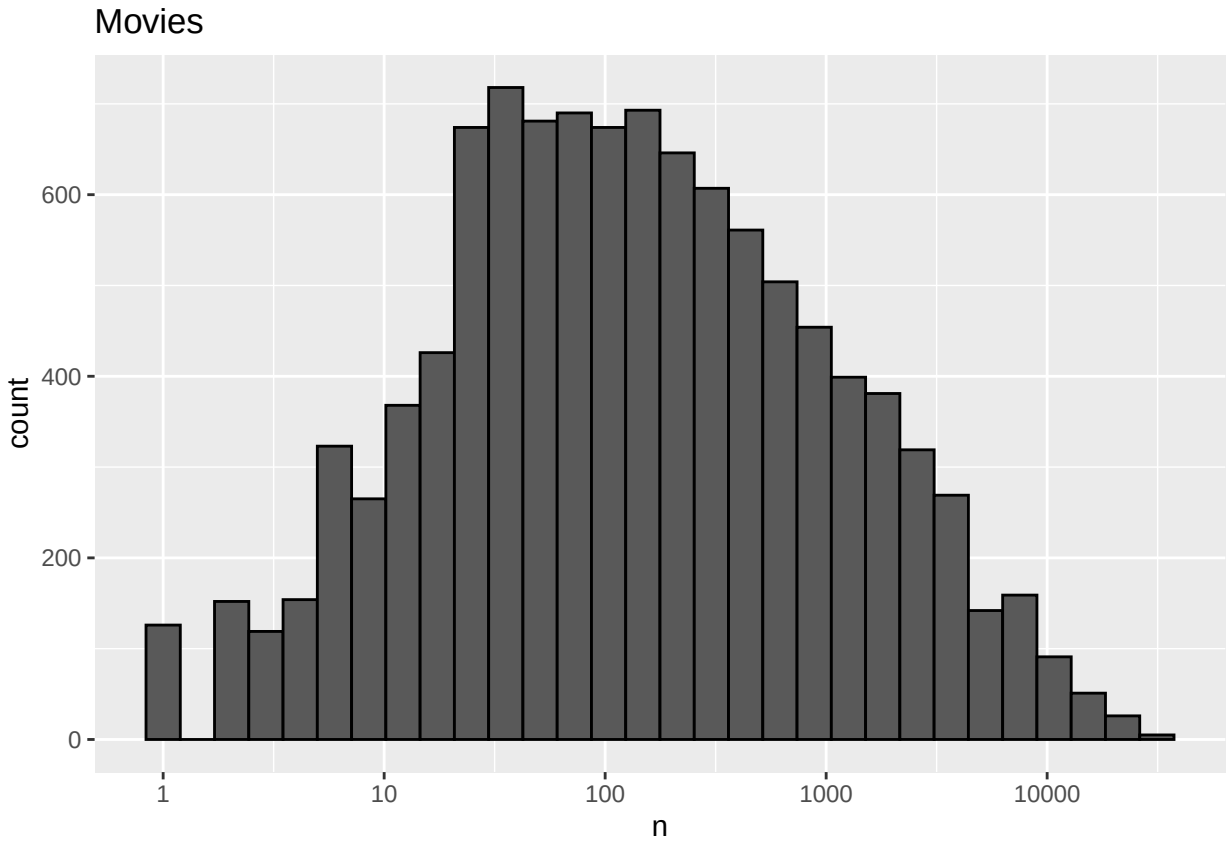
```
##   userId movieId rating timestamp                         title
## 1      1     122      5 838985046              Boomerang (1992)
## 2      1     185      5 838983525               Net, The (1995)
## 4      1     292      5 838983421               Outbreak (1995)
## 5      1     316      5 838983392               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
```

```
colnames(edx)
```

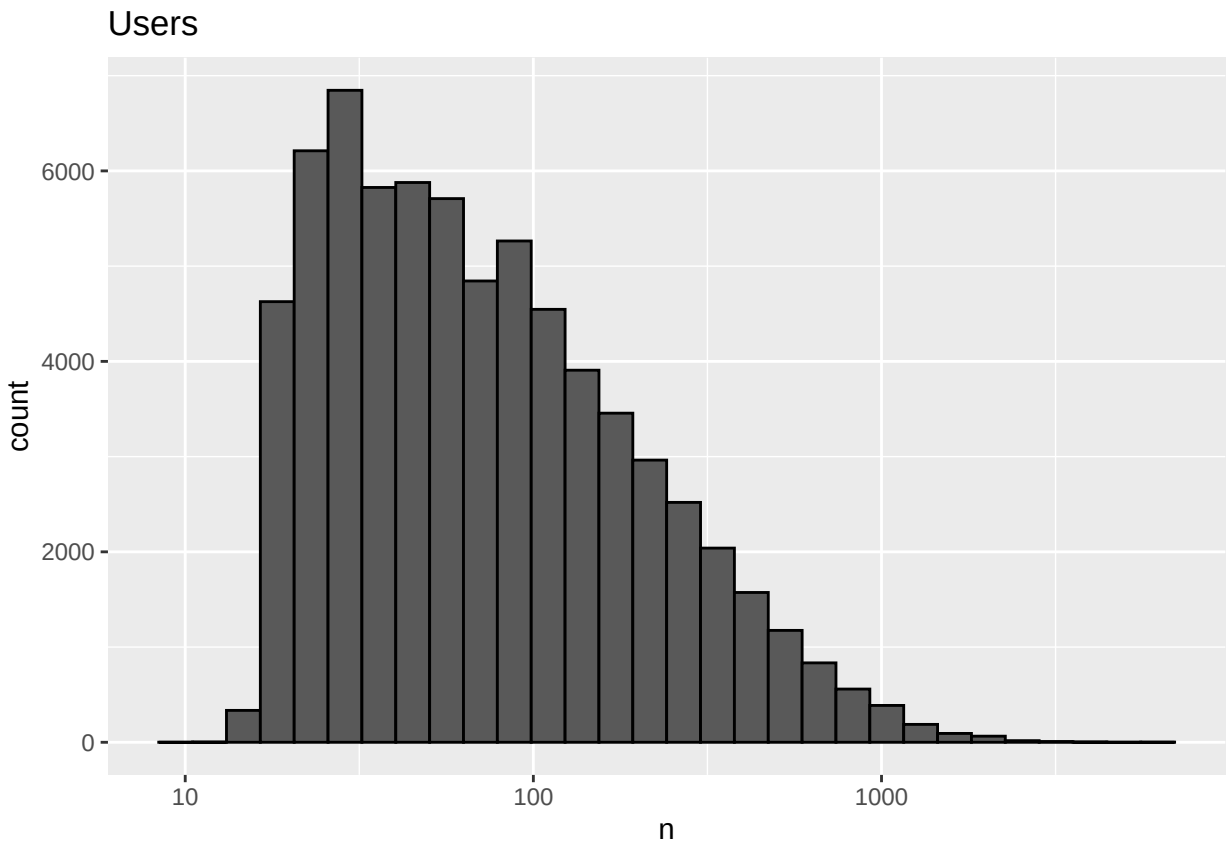
```
## [1] "userId"    "movieId"   "rating"    "timestamp" "title"     "genres"
```

Methods

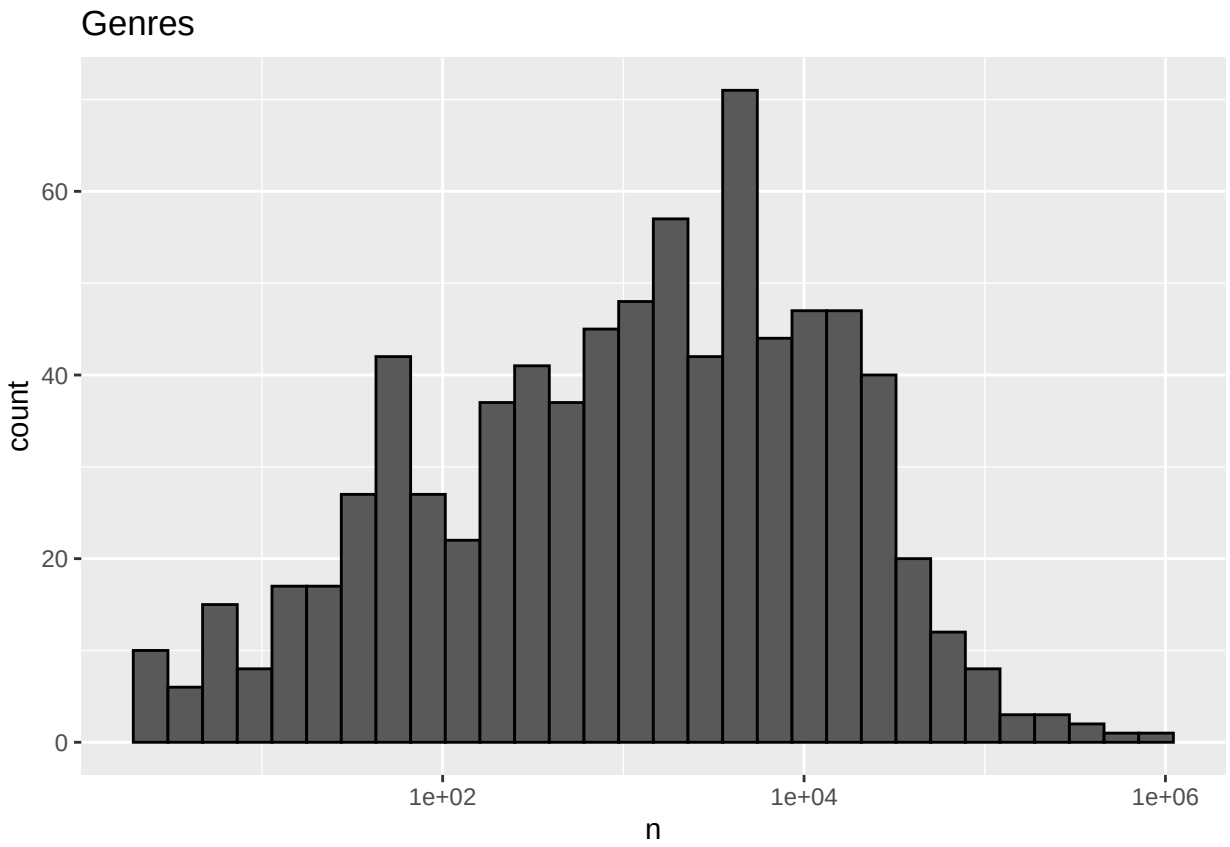
The dataset has first been explored to understand some of the biases. The graph below indicates that some movies are rated differently.

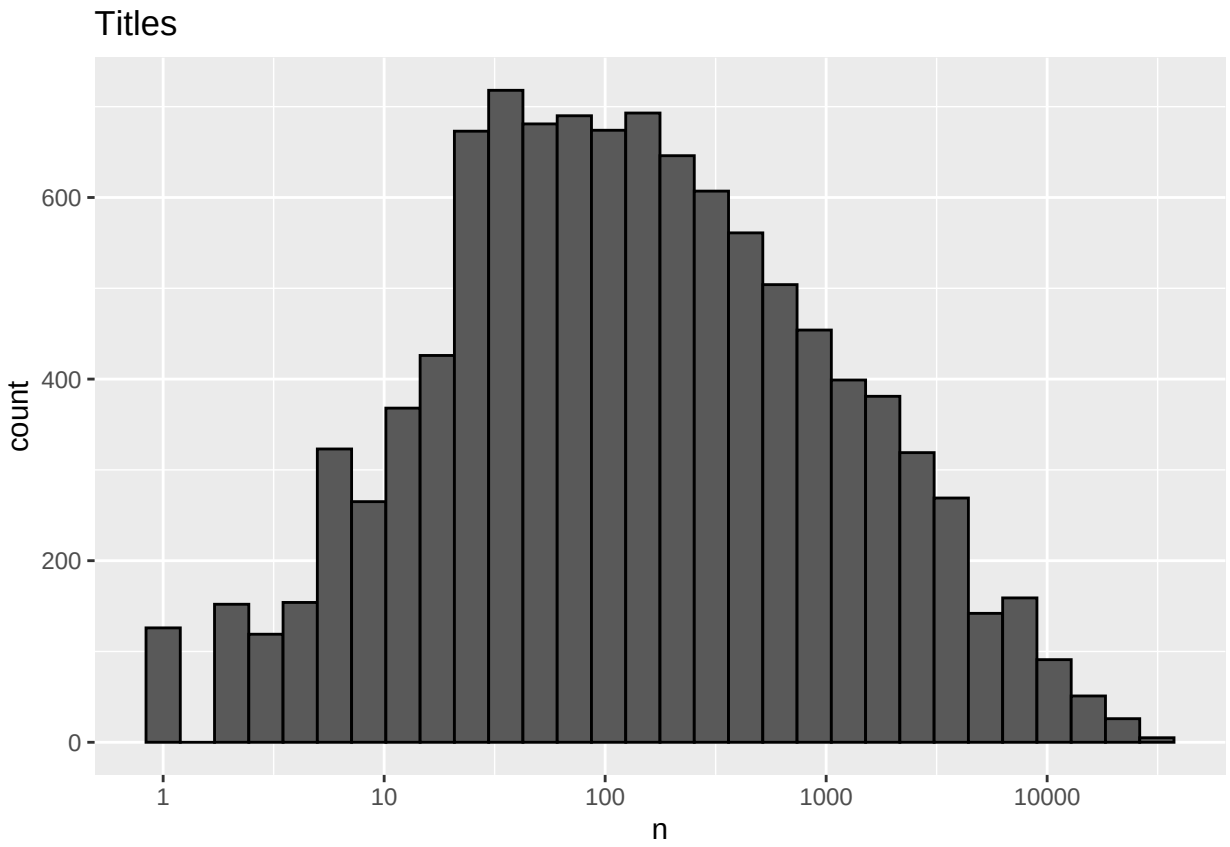


Another bias noted is that some users rate/watch more movies than others as indicated in the graph below.

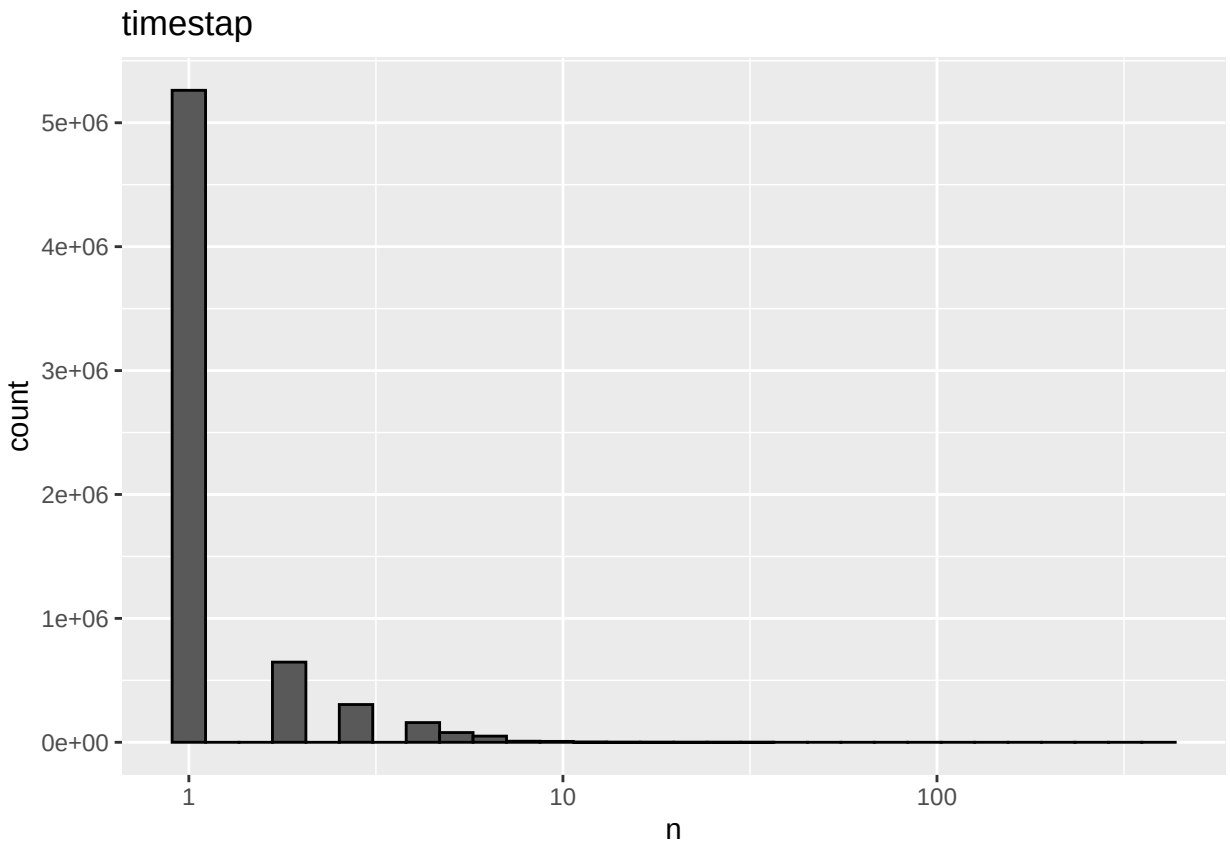


The last bias is that


Genres
60
40
20
0
count
1e+02
1e+04
1e+06
n




```
library(dplyr)
library(tidyverse)
edx %>%
     count(timestamp) %>%
     ggplot(aes(n)) +
     geom_histogram(bins = 30, color = "black") +
     scale_x_log10() +
     ggtitle("timestap")
```



The model has been trained based on 2 biases(biases due to users and movies), i trained data(train set) based on movieId and useId predictors.

Knowing that biases can be due to diffrent movies being rated differently and different users rating more movies than others.

I started by approximating a rating for all users across all movies and this is the avarage rating for all users across all movies.

Avareage rating for all movies by all users.