# DATA612 Project 1

*Alejandro D. Osborne*

*June 11, 2019*

```
library(pander)
library(ggplot2)
library(knitr)
library(dplyr)
```

## This is a basic recommender system that has been created based off of a sample set where 20 users rated 10 choices of ramen noodles.

### DataLoad & Analysis

Load the sample data set with ratings for ramen and convert the same into a matrix format.

```
# load csv into data variable
data <- read.csv("https://raw.githubusercontent.com/AlejandroOsborne/DATA612/master/RamenRatings.csv",r
# convert data into a matrix
data <- as.matrix(data)
pander(data)
```

Table 1: Table continues below

| | Nissin | Maruchan | MyKuali | Ve.Wong | Sapporo.Ichiban | Meimen |
|---|---|---|---|---|---|---|
| **user1** | 1.5 | 5 | 1 | 1.5 | 5 | 5 |
| **user2** | NA | 4 | 4 | NA | 4 | 2 |
| **user3** | 5 | 2 | NA | 1 | 3.5 | 3 |
| **user4** | 2 | NA | 2.75 | 2 | NA | 2 |
| **user5** | NA | 5 | NA | NA | 4 | 4 |
| **user6** | 5 | 3 | 5 | 2.5 | NA | NA |
| **user7** | NA | NA | 4 | NA | 4 | NA |
| **user8** | 3 | 4 | 1 | 2 | NA | 3.5 |
| **user9** | NA | 5 | NA | 3 | 5 | 5 |
| **user10** | 5 | 2 | 3.5 | NA | 4.5 | NA |
| **user11** | 2 | NA | NA | 1 | 3 | 2 |
| **user12** | 0.5 | 1 | 5 | 1.5 | NA | 1 |
| **user13** | NA | 4 | 4 | 2 | 2 | NA |
| **user14** | 4 | 5 | NA | NA | 5 | 4 |
| **user15** | 2 | 2 | 2 | 2 | 4 | 2.5 |
| **user16** | 5 | 2 | 5 | 1 | NA | 1 |
| **user17** | NA | 3 | 4.5 | NA | 3 | 5 |
| **user18** | 4 | NA | NA | NA | NA | NA |
| **user19** | 2 | 4 | 2 | 1 | 2.5 | 2 |
| **user20** | 3 | 4 | 4 | 2 | 2 | 3 |

| | Jinbo.Selection | Yumei | Mr..Noodles | Itsuki |
|---|---|---|---|---|
| **user1** | 2 | 2 | 2 | 5 |
| **user2** | 5 | 2.5 | 5 | 4 |

| | Jinbo.Selection | Yumei | Mr..Noodles | Itsuki |
|---|---|---|---|---|
| **user3** | 4 | 4 | NA | 3 |
| **user4** | NA | 1.5 | 1 | 4.5 |
| **user5** | 1 | 2 | NA | 3 |
| **user6** | 4.5 | NA | 4 | NA |
| **user7** | NA | 3.5 | 1 | 4 |
| **user8** | 3 | 5 | NA | 5 |
| **user9** | NA | 4 | 2 | NA |
| **user10** | 4 | NA | 1 | 3 |
| **user11** | 2.5 | 2 | 4 | 4.5 |
| **user12** | 2 | 5 | 1 | 5 |
| **user13** | NA | NA | 1 | 3 |
| **user14** | 2 | 5 | 3 | 4 |
| **user15** | NA | NA | NA | NA |
| **user16** | 4 | 5 | 4 | 4 |
| **user17** | 1 | 2 | 5 | NA |
| **user18** | 2 | 4 | 2.75 | 5 |
| **user19** | 3 | NA | 1 | 4 |
| **user20** | 1 | 1 | 4 | 3 |

### Top 3 Noodles

Here are the top 3 rated Noodles for fun

```
means <- colMeans(data, na.rm = TRUE)
cols <- colnames(data)[order(means, decreasing = TRUE)[1:3]]
top3 <- data.frame(ramen = cols,stringsAsFactors = FALSE)
pander(top3)
```

| ramen |
|---|
| Itsuki |
| Sapporo.Ichiban |
| Maruchan |

### Data Cleaving

Here we will split the data in training and testing set. We selected 12 reviews from trainin and will replace those with NA in the training set. NA was used so it would be omitted from our calculations. In the test dataset we only kept values identified for testing. the others were replaced with NA.

```
test_rows <- c(1,3,13,5,8,7,14,6,19,10,12,4)
test_cols <- c(1,10,6,3,7,5,2,8,9,4,10,7)
test_indices <- cbind(test_rows,test_cols)
data_train <- data
data_train[test_indices] <- NA
data_test <- data
data_test[test_indices] <- 0
data_test[data_test > 0] <- NA
data_test[test_indices] <- data[test_indices]
```

### TRAINING DATA

```
data_train
```

```
##        Nissin Maruchan MyKuali Ve.Wong Sapporo.Ichiban Meimen
## user1      NA        5    1.00     1.5             5.0    5.0
## user2      NA        4    4.00      NA             4.0    2.0
## user3     5.0        2      NA     1.0             3.5    3.0
## user4     2.0       NA    2.75     2.0              NA    2.0
## user5      NA        5      NA      NA             4.0    4.0
## user6     5.0        3    5.00     2.5              NA     NA
## user7      NA       NA    4.00      NA              NA     NA
## user8     3.0        4    1.00     2.0              NA    3.5
## user9      NA        5      NA     3.0             5.0    5.0
## user10    5.0        2    3.50      NA             4.5     NA
## user11    2.0       NA      NA     1.0             3.0    2.0
## user12    0.5        1    5.00     1.5              NA    1.0
## user13     NA        4    4.00     2.0             2.0     NA
## user14    4.0       NA      NA      NA             5.0    4.0
## user15    2.0        2    2.00     2.0             4.0    2.5
## user16    5.0        2    5.00     1.0              NA    1.0
## user17     NA        3    4.50      NA             3.0    5.0
## user18    4.0       NA      NA      NA              NA     NA
## user19    2.0        4    2.00     1.0             2.5    2.0
## user20    3.0        4    4.00     2.0             2.0    3.0
##        Jinbo.Selection Yumei Mr..Noodles Itsuki
## user1              2.0   2.0        2.00    5.0
## user2              5.0   2.5        5.00    4.0
## user3              4.0   4.0          NA     NA
## user4               NA   1.5        1.00    4.5
## user5              1.0   2.0          NA    3.0
## user6              4.5    NA        4.00     NA
## user7               NA   3.5        1.00    4.0
## user8               NA   5.0          NA    5.0
## user9               NA   4.0        2.00     NA
## user10             4.0    NA        1.00    3.0
## user11             2.5   2.0        4.00    4.5
## user12             2.0   5.0        1.00     NA
## user13              NA    NA        1.00    3.0
## user14             2.0   5.0        3.00    4.0
## user15              NA    NA          NA     NA
## user16             4.0   5.0        4.00    4.0
## user17             1.0   2.0        5.00     NA
## user18             2.0   4.0        2.75    5.0
## user19             3.0    NA          NA    4.0
## user20             1.0   1.0        4.00    3.0
```

### 3.2 TESTING DATA

```
data_test
```

```
##        Nissin Maruchan MyKuali Ve.Wong Sapporo.Ichiban Meimen
## user1     1.5       NA      NA      NA              NA     NA
## user2      NA       NA      NA      NA              NA     NA
## user3      NA       NA      NA      NA              NA     NA
## user4      NA       NA      NA      NA              NA     NA
```

```
## user5      NA       NA      NA      NA              NA     NA
## user6      NA       NA      NA      NA              NA     NA
## user7      NA       NA      NA      NA               4     NA
## user8      NA       NA      NA      NA              NA     NA
## user9      NA       NA      NA      NA              NA     NA
## user10     NA       NA      NA      NA              NA     NA
## user11     NA       NA      NA      NA              NA     NA
## user12     NA       NA      NA      NA              NA     NA
## user13     NA       NA      NA      NA              NA     NA
## user14     NA        5      NA      NA              NA     NA
## user15     NA       NA      NA      NA              NA     NA
## user16     NA       NA      NA      NA              NA     NA
## user17     NA       NA      NA      NA              NA     NA
## user18     NA       NA      NA      NA              NA     NA
## user19     NA       NA      NA      NA              NA     NA
## user20     NA       NA      NA      NA              NA     NA
##        Jinbo.Selection Yumei Mr..Noodles Itsuki
## user1               NA    NA          NA     NA
## user2               NA    NA          NA     NA
## user3               NA    NA          NA      3
## user4               NA    NA          NA     NA
## user5               NA    NA          NA     NA
## user6               NA    NA          NA     NA
## user7               NA    NA          NA     NA
## user8                3    NA          NA     NA
## user9               NA    NA          NA     NA
## user10              NA    NA          NA     NA
## user11              NA    NA          NA     NA
## user12              NA    NA          NA      5
## user13              NA    NA          NA     NA
## user14              NA    NA          NA     NA
## user15              NA    NA          NA     NA
## user16              NA    NA          NA     NA
## user17              NA    NA          NA     NA
## user18              NA    NA          NA     NA
## user19              NA    NA           1     NA
## user20              NA    NA          NA     NA
```

### 4.0 Calculations

Find the mean, bias and RMSE for both the data sets

### 4.1 Mean ratings of each User with chart for Training data set

Table 4: Table continues below

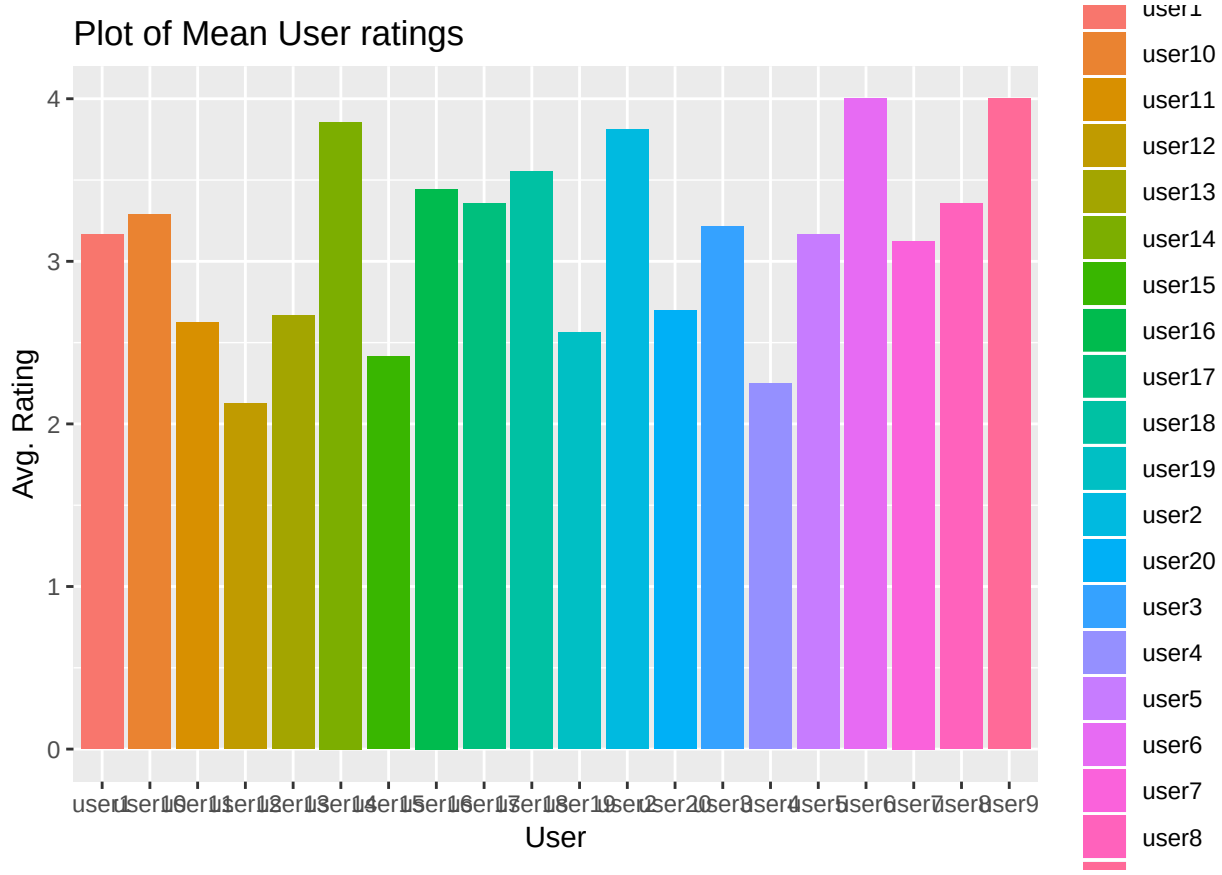
| user1 | user2 | user3 | user4 | user5 | user6 | user7 | user8 | user9 | user10 |
|---|---|---|---|---|---|---|---|---|---|
| 3.167 | 3.812 | 3.214 | 2.25 | 3.167 | 4 | 3.125 | 3.357 | 4 | 3.286 |

Table 5: Table continues below

| user11 | user12 | user13 | user14 | user15 | user16 | user17 | user18 | user19 |
|---|---|---|---|---|---|---|---|---|
| 2.625 | 2.125 | 2.667 | 3.857 | 2.417 | 3.444 | 3.357 | 3.55 | 2.562 |

| user11 | user12 | user13 | user14 | user15 | user16 | user17 | user18 | user19 |
|---|---|---|---|---|---|---|---|---|

| user20 |
|---|
| 2.7 |

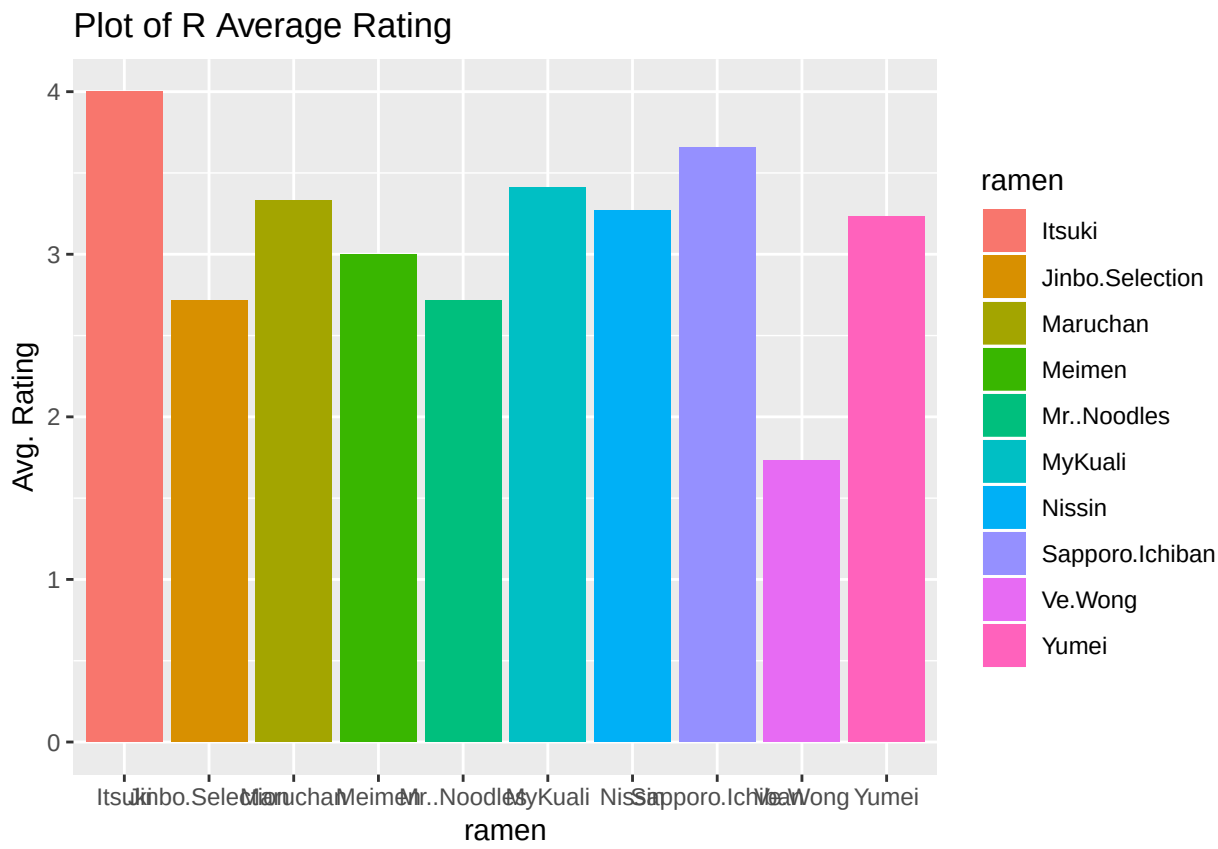



### 4.2 Mean ratings of ramen with chart for Training data set

Table 7: Table continues below

| Nissin | Maruchan | MyKuali | Ve.Wong | Sapporo.Ichiban | Meimen |
|---|---|---|---|---|---|
| 3.269 | 3.333 | 3.411 | 1.731 | 3.654 | 3 |

| Jinbo.Selection | Yumei | Mr..Noodles | Itsuki |
|---|---|---|---|
| 2.714 | 3.233 | 2.717 | 4 |




### 4.3 Raw Averages

Rating for every user-item combination. For Testing and Training data sets

```
## [1] 3.214286
```

```
## [1] 3.109929
```

### 4.4 RMSE for raw averages

For Testing and Training data sets

### 4.5 Train Data

RMSE for Train

```
## [1] 1.365536
```

### 4.6 Test Data

RMSE for Test

```
## [1] 1.460361
```

### 4.7 Bias for each user and ramen

User Bias

```
## user bias
user_bias <- user_means - raw_train
user_bias_df <-  data.frame(as.list(user_bias))
user_bias_df <- tidyr::gather(user_bias_df,"user")
colnames(user_bias_df) <-c("User","Bias")
pander(user_bias_df)
```

| User | Bias |
|---|---|
| user1 | 0.05674 |
| user2 | 0.7026 |
| user3 | 0.1044 |
| user4 | -0.8599 |
| user5 | 0.05674 |
| user6 | 0.8901 |
| user7 | 0.01507 |
| user8 | 0.2472 |
| user9 | 0.8901 |
| user10 | 0.1758 |
| user11 | -0.4849 |
| user12 | -0.9849 |
| user13 | -0.4433 |
| user14 | 0.7472 |
| user15 | -0.6933 |
| user16 | 0.3345 |
| user17 | 0.2472 |
| user18 | 0.4401 |
| user19 | -0.5474 |
| user20 | -0.4099 |

Ramen Bias

```
#Ramen  bias
ramen_bias <- ramen_means - raw_train
ramen_bias_df <-  data.frame(as.list(ramen_bias))
ramen_bias_df <- tidyr::gather(ramen_bias_df,"Ramen")
colnames(ramen_bias_df) <-c("Ramen","Bias")
pander(ramen_bias_df)
```

| Ramen | Bias |
|---|---|
| Nissin | 0.1593 |
| Maruchan | 0.2234 |
| MyKuali | 0.3008 |
| Ve.Wong | -1.379 |
| Sapporo.Ichiban | 0.5439 |
| Meimen | -0.1099 |
| Jinbo.Selection | -0.3956 |
| Yumei | 0.1234 |
| Mr..Noodles | -0.3933 |
| Itsuki | 0.8901 |

**5.0 Baseline predictors for each user item**

```
# raw average + user bias + ramen bias
calBaseLine <- function(in_matrix, ramen_bias_in,user_bias_in,raw_average)
{
  out_matrix <- in_matrix
  row_count <-1
  for(item in 1:nrow(in_matrix))
  {
    col_count <-1
    for(colItem in 1: ncol(in_matrix))
    {
      #out_matrix[row_count,col_count] <- 0
     out_matrix[row_count,col_count] <- raw_average[1] + user_bias_in[[row_count]] +  ramen_bias_in[[co
      col_count <- col_count +1
    }
    row_count <- row_count +1
  }
return(out_matrix)
}
base_pred <- calBaseLine(data_train,ramen_bias,user_bias,raw_train)
pander(base_pred)
```

Table 11: Table continues below

| | Nissin | Maruchan | MyKuali | Ve.Wong | Sapporo.Ichiban | Meimen |
|---|---|---|---|---|---|---|
| **user1** | 3.326 | 3.39 | 3.467 | 1.788 | 3.711 | 3.057 |
| **user2** | 3.972 | 4.036 | 4.113 | 2.433 | 4.356 | 3.703 |
| **user3** | 3.374 | 3.438 | 3.515 | 1.835 | 3.758 | 3.104 |
| **user4** | 2.409 | 2.473 | 2.551 | 0.8708 | 2.794 | 2.14 |
| **user5** | 3.326 | 3.39 | 3.467 | 1.788 | 3.711 | 3.057 |
| **user6** | 4.159 | 4.223 | 4.301 | 2.621 | 4.544 | 3.89 |
| **user7** | 3.284 | 3.348 | 3.426 | 1.746 | 3.669 | 3.015 |
| **user8** | 3.516 | 3.581 | 3.658 | 1.978 | 3.901 | 3.247 |
| **user9** | 4.159 | 4.223 | 4.301 | 2.621 | 4.544 | 3.89 |
| **user10** | 3.445 | 3.509 | 3.586 | 1.907 | 3.83 | 3.176 |
| **user11** | 2.784 | 2.848 | 2.926 | 1.246 | 3.169 | 2.515 |
| **user12** | 2.284 | 2.348 | 2.426 | 0.7458 | 2.669 | 2.015 |
| **user13** | 2.826 | 2.89 | 2.967 | 1.288 | 3.211 | 2.557 |
| **user14** | 4.016 | 4.081 | 4.158 | 2.478 | 4.401 | 3.747 |
| **user15** | 2.576 | 2.64 | 2.717 | 1.038 | 2.961 | 2.307 |
| **user16** | 3.604 | 3.668 | 3.745 | 2.065 | 3.988 | 3.335 |
| **user17** | 3.516 | 3.581 | 3.658 | 1.978 | 3.901 | 3.247 |
| **user18** | 3.709 | 3.773 | 3.851 | 2.171 | 4.094 | 3.44 |
| **user19** | 2.722 | 2.786 | 2.863 | 1.183 | 3.106 | 2.453 |
| **user20** | 2.859 | 2.923 | 3.001 | 1.321 | 3.244 | 2.59 |

| | Jinbo.Selection | Yumei | Mr..Noodles | Itsuki |
|---|---|---|---|---|
| **user1** | 2.771 | 3.29 | 2.773 | 4.057 |
| **user2** | 3.417 | 3.936 | 3.419 | 4.703 |
| **user3** | 2.819 | 3.338 | 2.821 | 4.104 |
| **user4** | 1.854 | 2.373 | 1.857 | 3.14 |

| | Jinbo.Selection | Yumei | Mr..Noodles | Itsuki |
|---|---|---|---|---|
| **user5** | 2.771 | 3.29 | 2.773 | 4.057 |
| **user6** | 3.604 | 4.123 | 3.607 | 4.89 |
| **user7** | 2.729 | 3.248 | 2.732 | 4.015 |
| **user8** | 2.961 | 3.481 | 2.964 | 4.247 |
| **user9** | 3.604 | 4.123 | 3.607 | 4.89 |
| **user10** | 2.89 | 3.409 | 2.892 | 4.176 |
| **user11** | 2.229 | 2.748 | 2.232 | 3.515 |
| **user12** | 1.729 | 2.248 | 1.732 | 3.015 |
| **user13** | 2.271 | 2.79 | 2.273 | 3.557 |
| **user14** | 3.461 | 3.981 | 3.464 | 4.747 |
| **user15** | 2.021 | 2.54 | 2.023 | 3.307 |
| **user16** | 3.049 | 3.568 | 3.051 | 4.335 |
| **user17** | 2.961 | 3.481 | 2.964 | 4.247 |
| **user18** | 3.154 | 3.673 | 3.157 | 4.44 |
| **user19** | 2.167 | 2.686 | 2.169 | 3.453 |
| **user20** | 2.304 | 2.823 | 2.307 | 3.59 |

### 6.0 RMSE for the baseline predictors for both training data and testing data sets

```
## test data
# finding Error
data_err <- data_test - base_pred
# squaring error
data_err <- (data_err)^2
#finding average
data_rmse_test<- mean(data_err,na.rm = TRUE)
# square root
data_rmse_test<- sqrt(data_rmse_test)
## training data
# finding Error
data_err_train <- data_train - base_pred
# squaring error
data_err_train <- (data_err_train)^2
#finding average
data_rmse_train <- mean(data_err_train,na.rm = TRUE)
# square root
data_rmse_train<- sqrt(data_rmse_train)
```

### 6.1 RMSE - TEST DATA

```
data_rmse_test
```

```
## [1] 1.243113
```

### 6.2 RMSE - TRAIN DATA

```
data_rmse_train
```

```
## [1] 1.116384
```

### 7.0 Summary

Lets calculate the percentage improvements based on the original (simple average) and baseline predictor (including bias) RMSE numbers for both Test and Train data sets.

The results show that we see an 18% improvement in making a prediction for the ratings in the Training data set. Where as we see only 15% improvement in prediction for the Test data set. Both are positive hoewver the Training data set yielded better prediction.

```
# Train data set
R1 <- rmse_train
Rb1 <- data_rmse_train
Per_Improv_Train <- (1-(Rb1/R1))*100
Per_Improv_Train
```

```
## [1] 18.2457
```

```
# Test data set
R2 <- rmse_test
Rb2 <- data_rmse_test
Per_Improv_Test <- (1-(Rb2/R2))*100
Per_Improv_Test
```

```
## [1] 14.87627
```