Zeru-Zhou-project08

October 27, 2021

1 Project 8 – Zeru Zhou

TA Help: NA

Collaboration: NA

- get help from Dr. Ward's video

1.1 Question 1

[1]:
```
library(data.table)
```

[2]:
```
interactions <- fread("/depot/datamine/data/goodreads/csv/interactions_subset.
↪csv")
```

[3]:
```
# A function that, given a string (userID) and a value (min_rating) returns a␣
↪value (probability_of_reviewing).
get_probability_of_review <- function(interactions_dataset, userID, min_rating)␣
↪{
        # Filtering the dataset and keep data that has user_id equals to the␣
↪given userID. Name the filtered dataset user_data.
        user_data <- subset(interactions_dataset, user_id == userID)

        # Filtering the dataset once more to keep data that has is_read column␣
↪equals to 1. Name the filtered dataset read_user_data.
        read_user_data <- subset(user_data, is_read == 1)

        # Filtering the dataset once more to keep data that has rating column␣
↪more than the given min_rating. Name the filtered dataset␣
↪read_user_min_rating_data.
        read_user_min_rating_data <- subset(read_user_data, rating >=␣
↪min_rating)

        # Define probability_of_reviewing as the mean of the is_reviewed column␣
↪in dataset read_user_min_rating_data.
        probability_of_reviewing <- mean(read_user_min_rating_data$is_reviewed)

        # Return the result
```

```
        return(probability_of_reviewing)
}

get_probability_of_review(interactions_dataset = interactions, userID = 5000,␣
↪min_rating = 3)
```

0.0707964601769911

This function takes interactions_dataset, userID, and min_rating as inputs, and probability of reviewing as output. It uses userID, and min_rating to filter the dataset, then calculating the mean of is_reviewed column of the filtered dataset. It has 3 arguments: interactions_dataset, userID, and min_rating.

1.2 Question 2

[4]:
```
get_probability_of_review <- function(interactions_dataset, userID,␣
↪min_rating=0) {
        # Filtering the dataset and keep data that has user_id equals to the␣
↪given userID. Name the filtered dataset user_data.
        user_data <- subset(interactions_dataset, user_id == userID)

        # Filtering the dataset once more to keep data that has is_read column␣
↪equals to 1. Name the filtered dataset read_user_data.
        read_user_data <- subset(user_data, is_read == 1)

        # Filtering the dataset once more to keep data that has rating column␣
↪more than the given min_rating. Name the filtered dataset␣
↪read_user_min_rating_data.
        read_user_min_rating_data <- subset(read_user_data, rating >=␣
↪min_rating)

        # Define probability_of_reviewing as the mean of the is_reviewed column␣
↪in dataset read_user_min_rating_data.
        probability_of_reviewing <- mean(read_user_min_rating_data$is_reviewed)

        # Return the result
        return(probability_of_reviewing)
}
```

[5]:
```
get_probability_of_review(interactions_dataset = interactions, userID = 5000)
```

0.0816326530612245

[6]:
```
get_probability_of_review(userID = 5000,interactions_dataset = interactions)
```

0.0816326530612245

[7]:
```
get_probability_of_review(interactions, 5000)
```

0.0816326530612245

Here is modified: min_rating=0 at the start of the function.

1.3 Question 3

[5]:
```
get_probability_of_review <- function(interactions_dataset, userID,␣
↪min_rating=0) {
       # Filtering the dataset
       read_user_min_rating_data <- subset(interactions_dataset, (user_id ==␣
↪userID) & (is_read == 1) & (rating >= min_rating))

       # Define probability_of_reviewing as the mean of the is_reviewed column␣
↪in dataset read_user_min_rating_data.
       probability_of_reviewing <- mean(read_user_min_rating_data$is_reviewed)

       # Return the result
       return(probability_of_reviewing)
}
```

[9]:
```
get_probability_of_review(interactions, 5000)
```

0.0816326530612245

Code is reduced above. Now we only use 1 subset.

1.4 Question 4

[10]:
```
head(read_user_min_rating_data)
```

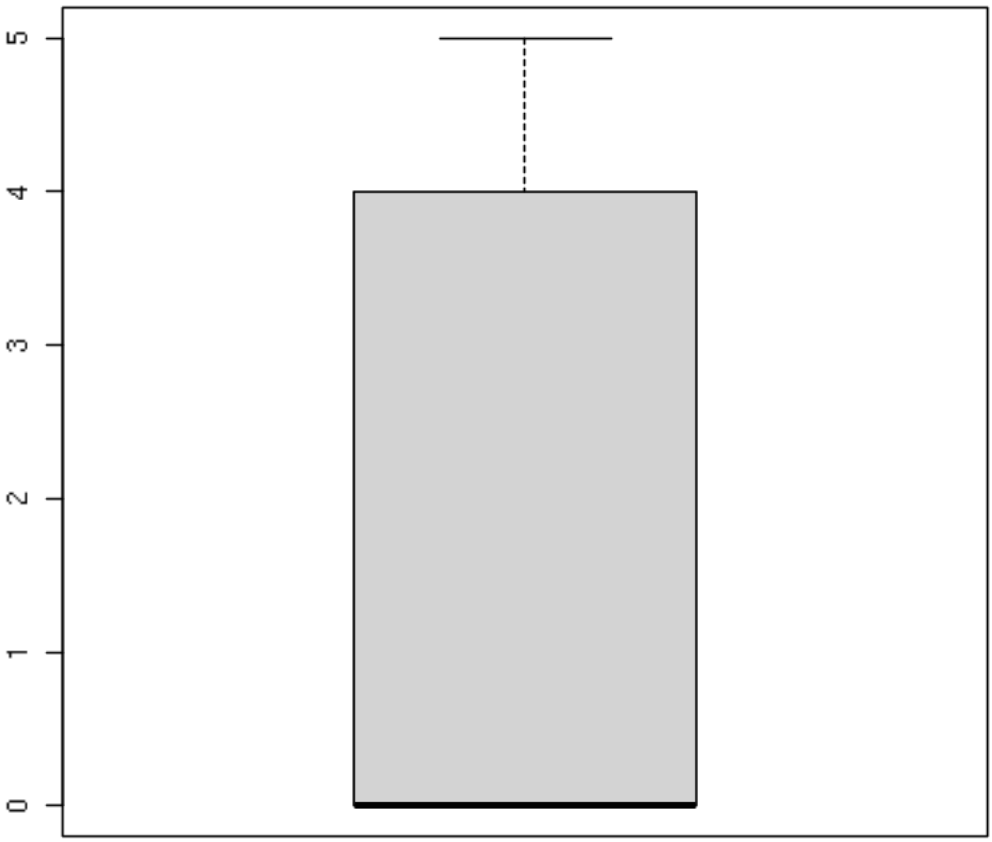
```
Error in head(read_user_min_rating_data): object 'read_user_min_rating_data' no␣
↪found
Traceback:

1. head(read_user_min_rating_data)
```

There is an error that there do not exist something called "read_user_min_rating_data", so there comes an error when running head function on it. This is because "read_user_min_rating_data" is an dataset we defined that only make sense inside our "get_probability_of_review" function. That is, it could not be used or detected outside "get_probability_of_review" function, so there is an error when directly use it outside the function "get_probability_of_review".

1.5 Question 5

[16]:
```
boxplot(interactions$rating)
```



[4]:
```
users <- sample(interactions$user_id, 10)
```

[5]:
```
users
```

1. 115826 2. 72191 3. 198769 4. 51969 5. 141041 6. 89743 7. 102417 8. 56922 9. 133891 10. 71596

[10]:
```
prob_review <- sapply(users, function(m)␣
↪get_probability_of_review(interactions_dataset=interactions, userID=m,␣
↪min_rating=0))
```

[11]:
```
prob_review
```

1. 0.365384615384615 2. 0.165584415584416 3. 0.255952380952381 4. 0.00581395348837209 5. 0.0259179265658747 6. 0.303225806451613 7. 0.0994897959183673 8. 0.0541237113402062

9. 0.61106426041491 10. 0.038961038961039

The results are listed above. I pick 0 as the specific minimum rating value because according to boxplot I drew, there are many data that has rating value of 0. If we choose another number greater than 0 then it should not be called "minimum" rating value.

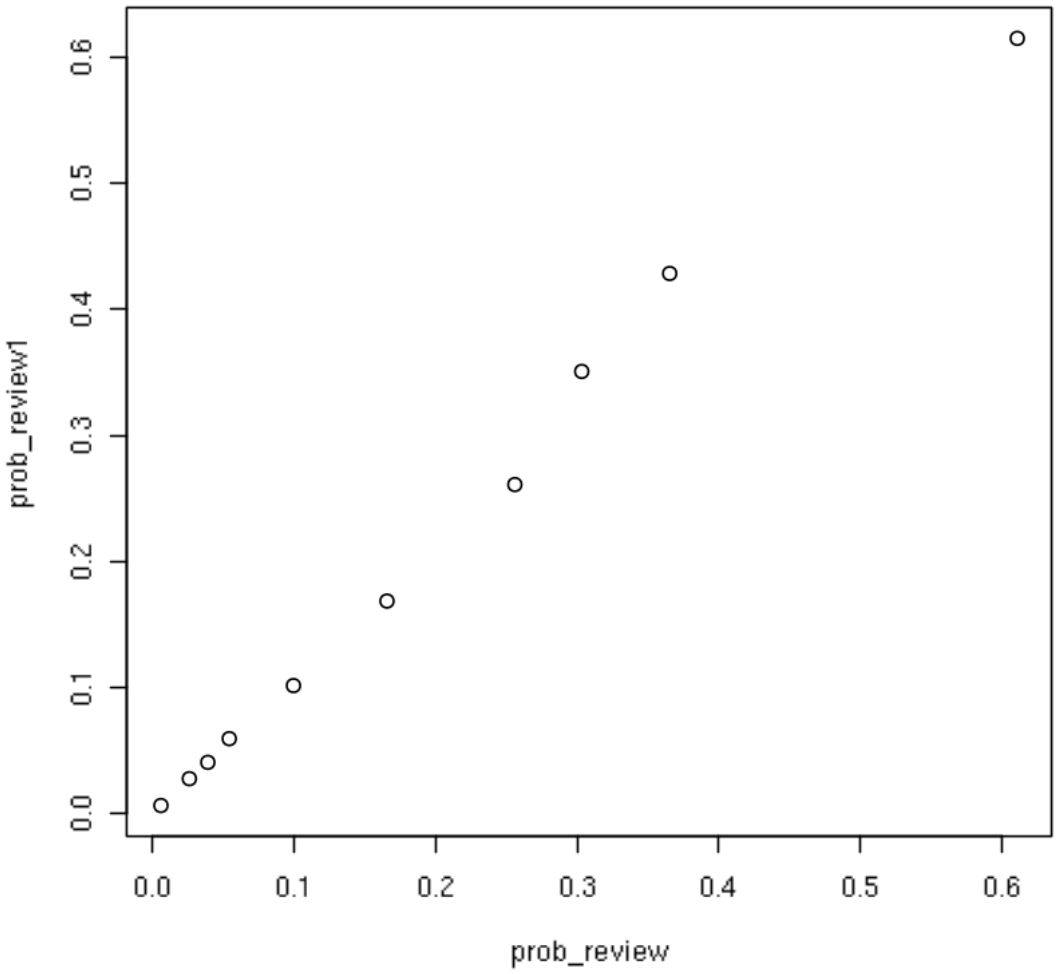
1.6 Question 6

[12]:
```
prob_review1 <- sapply(users, function(m)␣
↪get_probability_of_review(interactions_dataset=interactions, userID=m,␣
↪min_rating=1))
```

[13]:
```
prob_review1
```

1. 0.428571428571429 2. 0.168316831683168 3. 0.260869565217391 4. 0.00588235294117647 5. 0.0272108843537415 6. 0.350746268656716 7. 0.101167315175097 8. 0.0589970501474926 9. 0.615397688647179 10. 0.0401785714285714

[14]:
```
plot(prob_review, prob_review1)
```



For each of the 10 users, the horizontal axis represents probability when min_rating is 0. The vertical axis represents the probability when min_rating is 1. As we can see, the value of probability is almost the same except for a couple of users with probability between 0.3 and 0.5. Hence, changing the value of min_rating affects the outcome of probability, but maybe slightly as my result above.

1.7 Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together – We are Purdue.