**Homework 2**

Question A

1.

After reading the csv data into the list, it is divided into training set, validation set and test set. Calculate the MSE of different models trained by training set data when N equals different values.

N = 0 182250340.64237732

N = 1 34081039.65716226

N = 2 36529880.19398984

N = 3 42702991.985939994

N = 4 43788682.12288098

N = 5 44303784.196106166

The best model is obtained when the MSE is smallest, so the model when N=1 is the best model. Then, fit data according to model 1.

2.

The model results are then tested against the test set data.

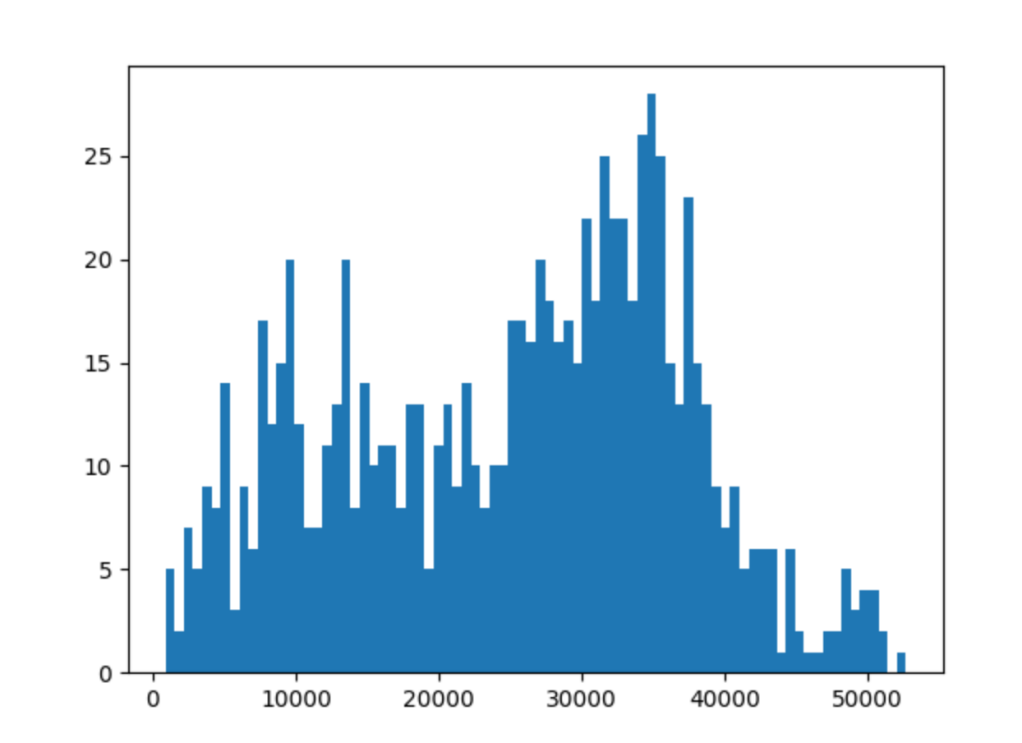
N = 1 34081039.65716226

Because of the MSE is so high, so this model is not a good model.

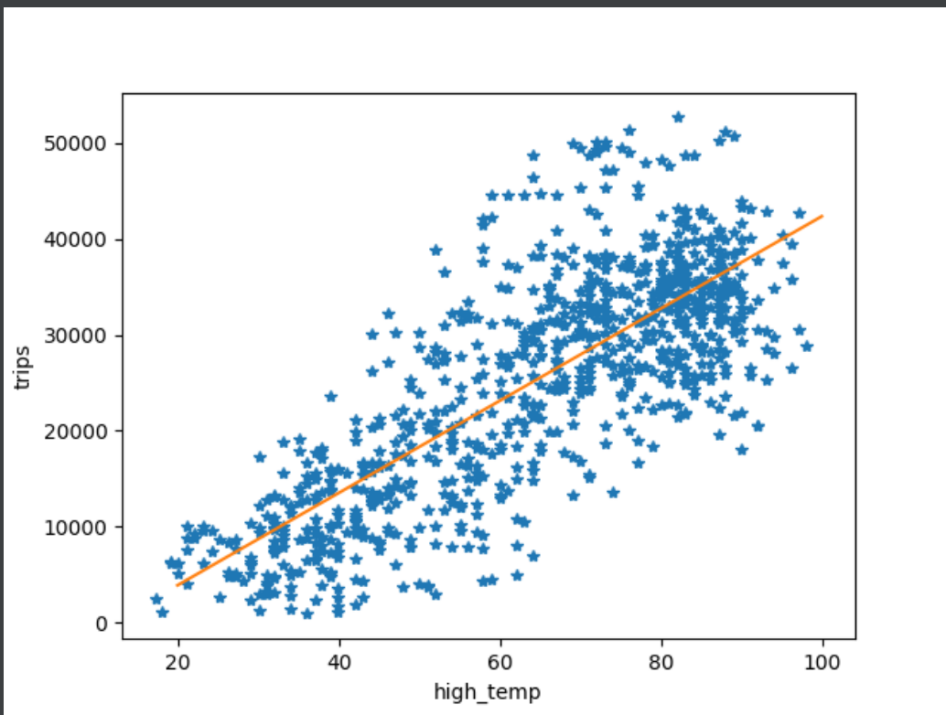
3.

1. The daily trips taken in the system

Frequency corresponding to different number of trips

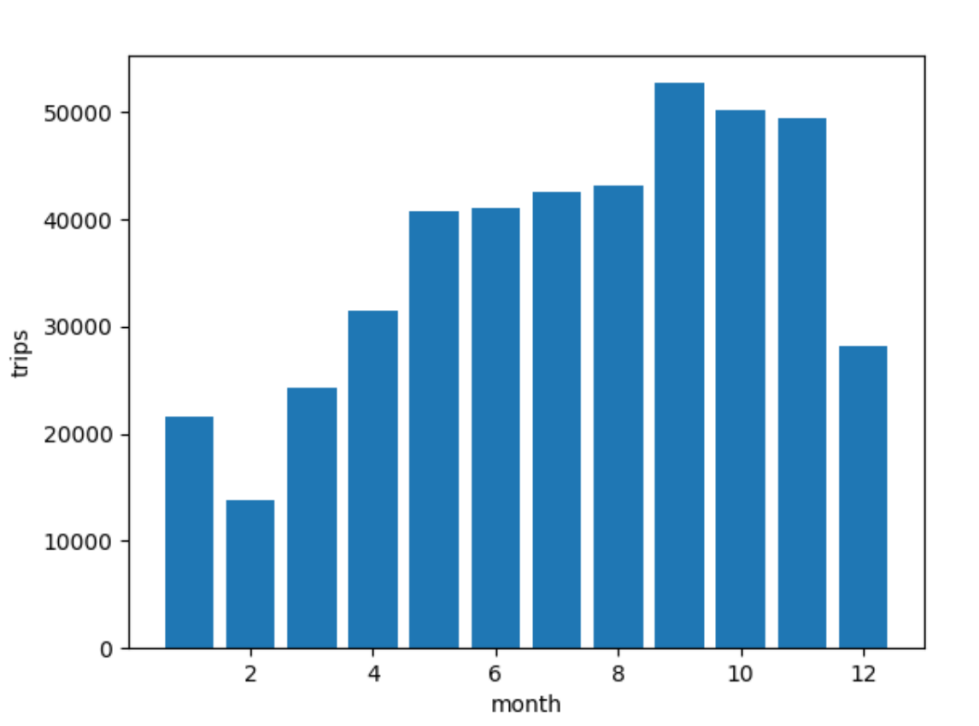


b.The data capturing the relationship between the daily trips taken and the maximum temperature

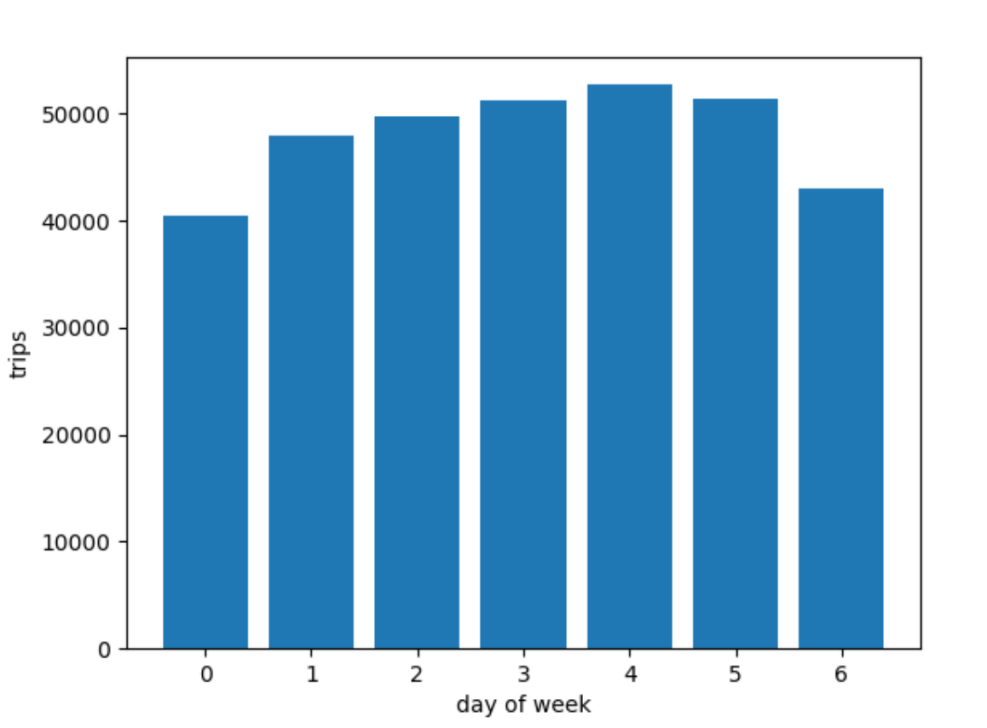


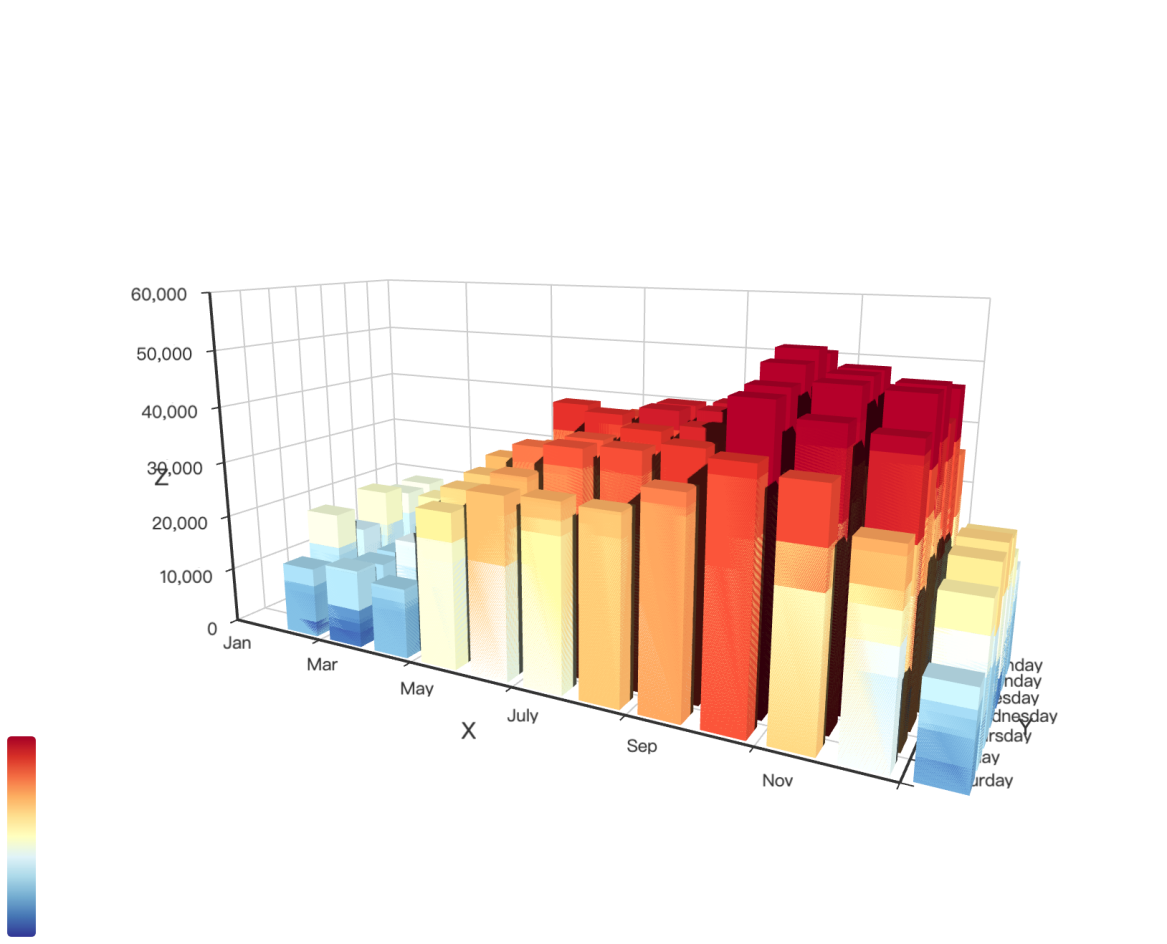
c.The relationship between the trips, the day of the week (variable: dow) and the month of the year.

Relationship between reaction months and trips



Relationship between reaction day and trips





Dynamic model in html in zip file

Additional section：

The results will also change when changing the split data method. The MSE results calculated in another way are as follows.

N = 0 123369133.12520002

N = 1 33026794.30900336

N = 2 34169074.758191966

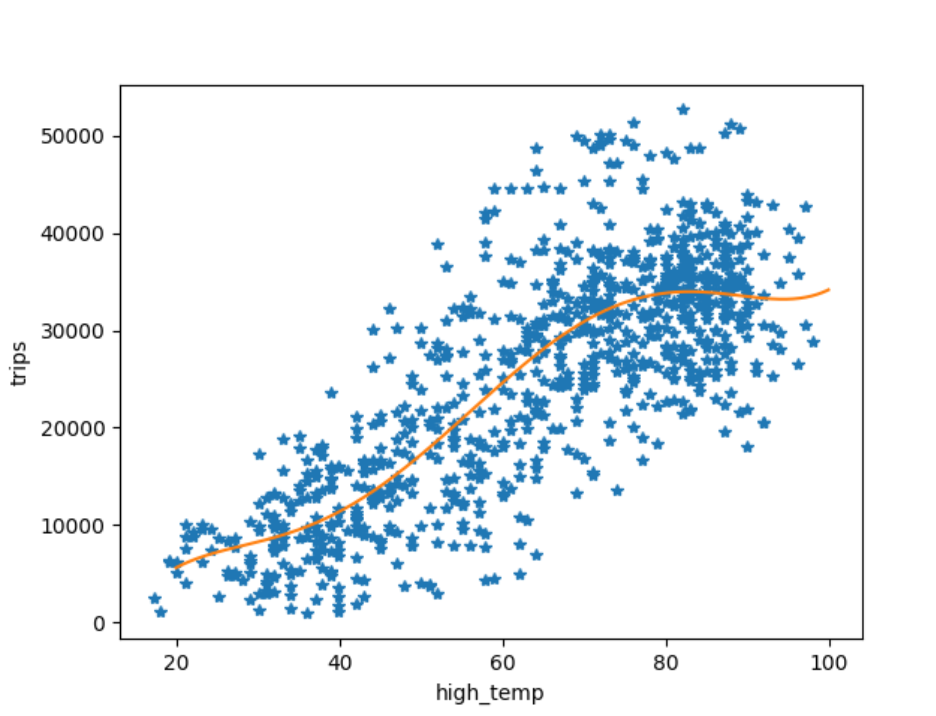
N = 3 29942790.983818565

N = 4 29828121.357124966

N = 5 29752659.95929693

The model results are then tested against the test set data. At this point model 5 gives the best results. So, I use model 5 to fit data again.

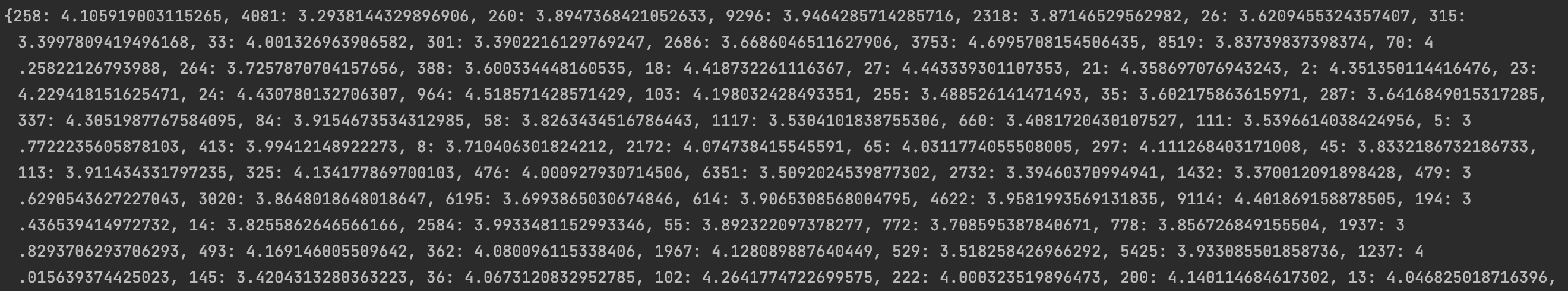
N = 5 93724730.72918922



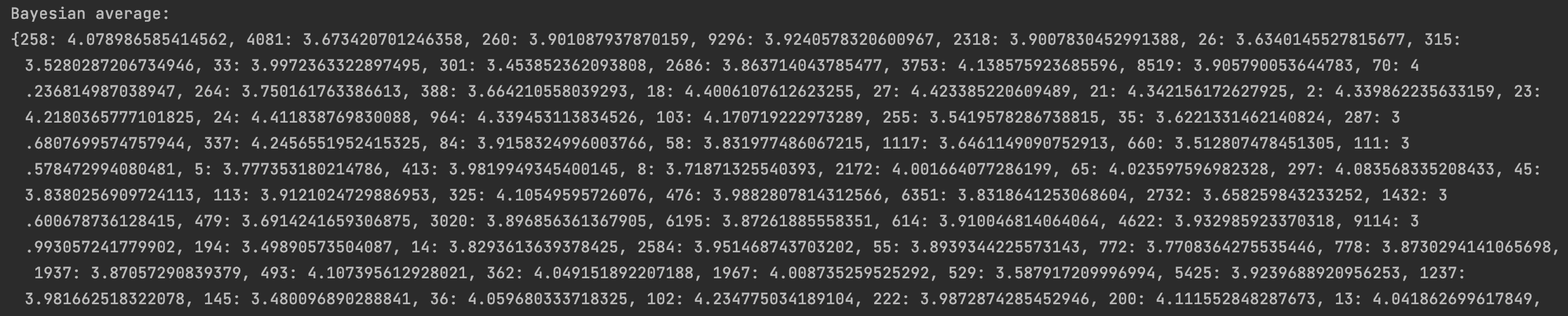
Question 2

First read in the data and make a dictionary with book id as key and rating as value. Calculate the average of all ratings for each book, the average score of all books and average number of all book ratings. The Bayesian average is calculated using the data derived from the above.

Average Rating Result(Part):

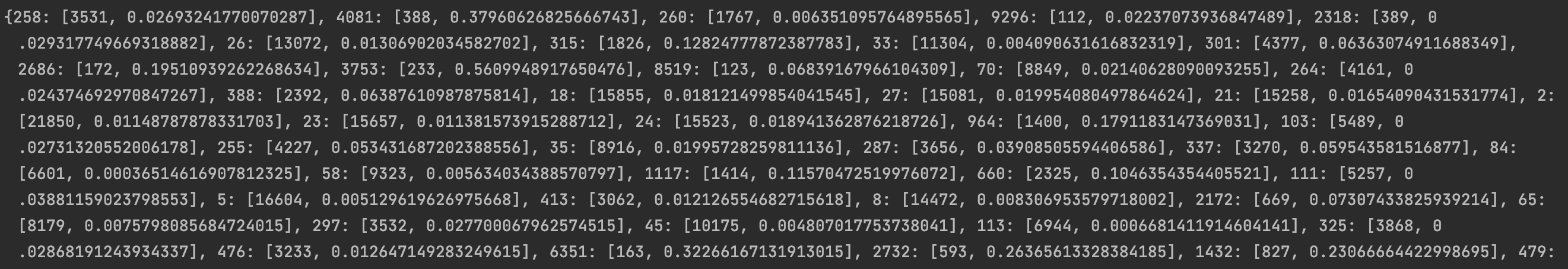


Bayesian Average Result(Part):

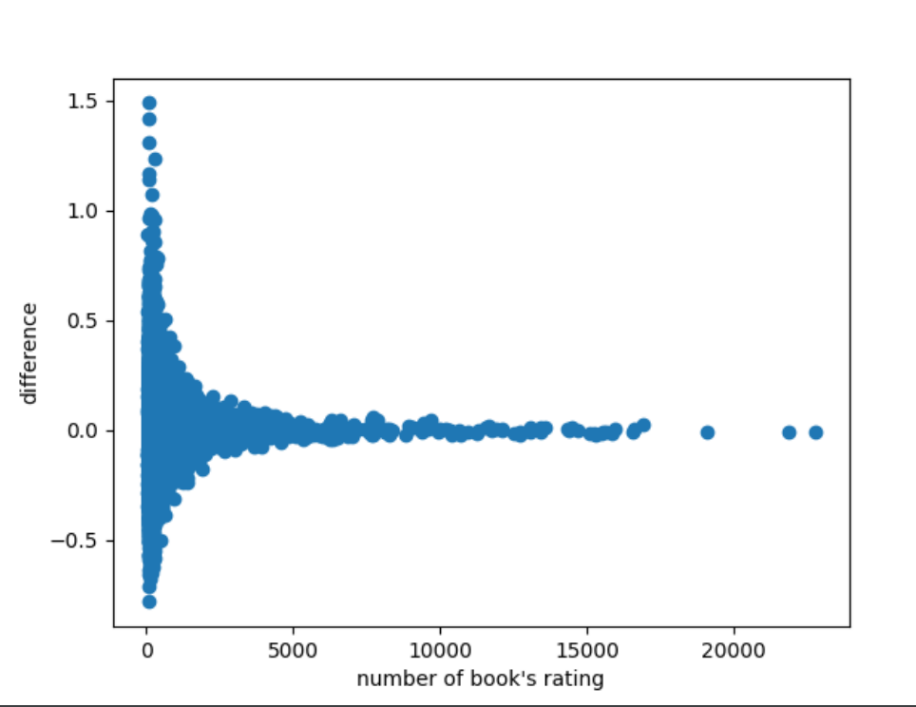


Calculate the difference between the rating average of each book and its Bayesian average and form a dictionary. Plot the data as a scatter plot with the number of book’s rating in the horizontal coordinate and the difference in the vertical coordinate.

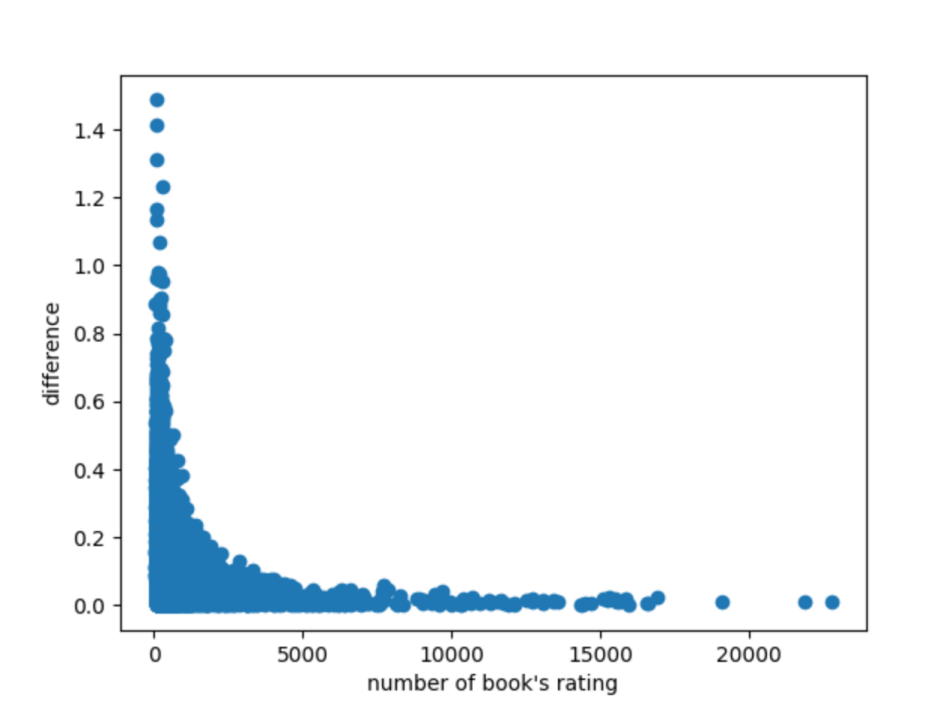
Difference Result:



When the difference contains negative numbers：



When the difference does not contain negative numbers：



The final conclusion is that as the number of ratings increases, the difference between the directly calculated average and the Bayesian Average decreases.