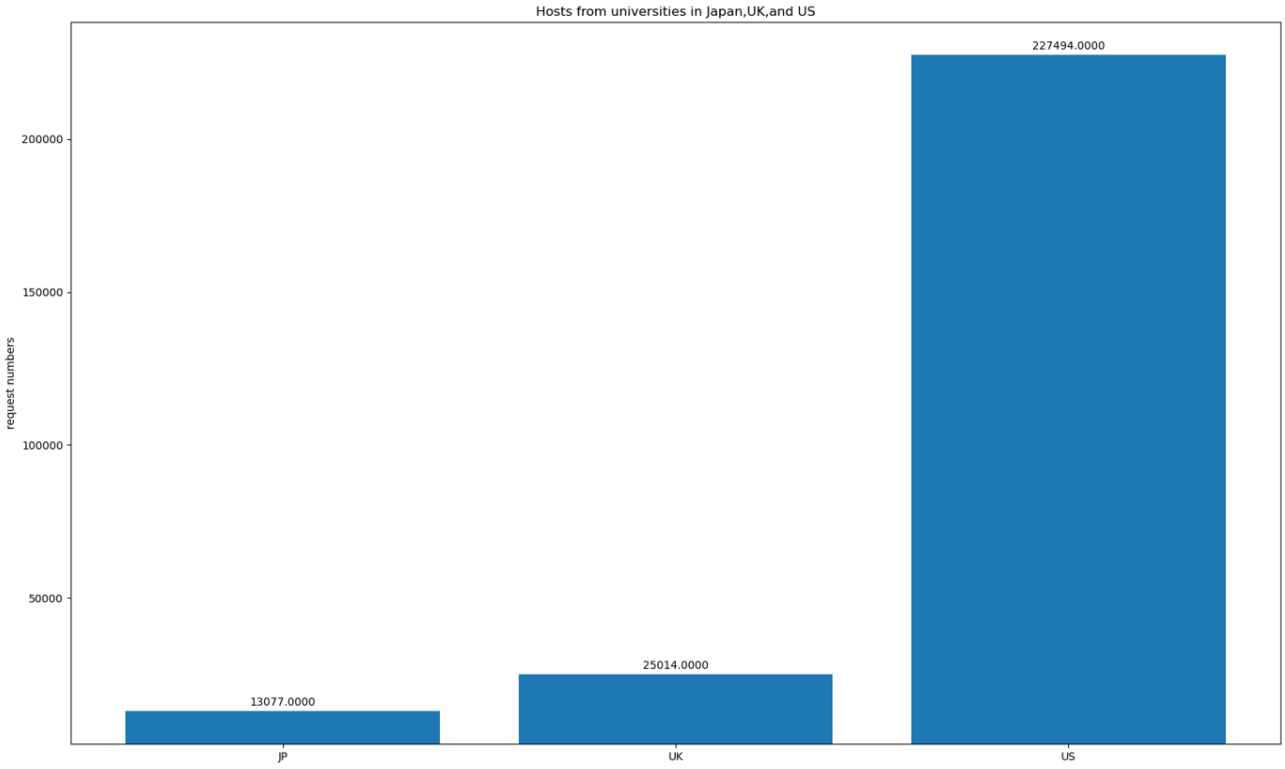
AS1_report

Question 1. Log Mining and Analysis [15 marks]

A.

1) The total hosts number of Japan universities is 13077.
2) The total hosts number of UK universities is 25014.
3) The total hosts number of US universities is 227494.



B. 1)

Top 9 of Japan

host	count
tohoku.ac.jp	824
kyoto-u.ac.jp	703
nagoya-u.ac.jp	692
u-tokyo.ac.jp	689
osaka-u.ac.jp	527
shizuoka.ac.jp	472
ritsumei.ac.jp	426
keio.ac.jp	346
waseda.ac.jp	337

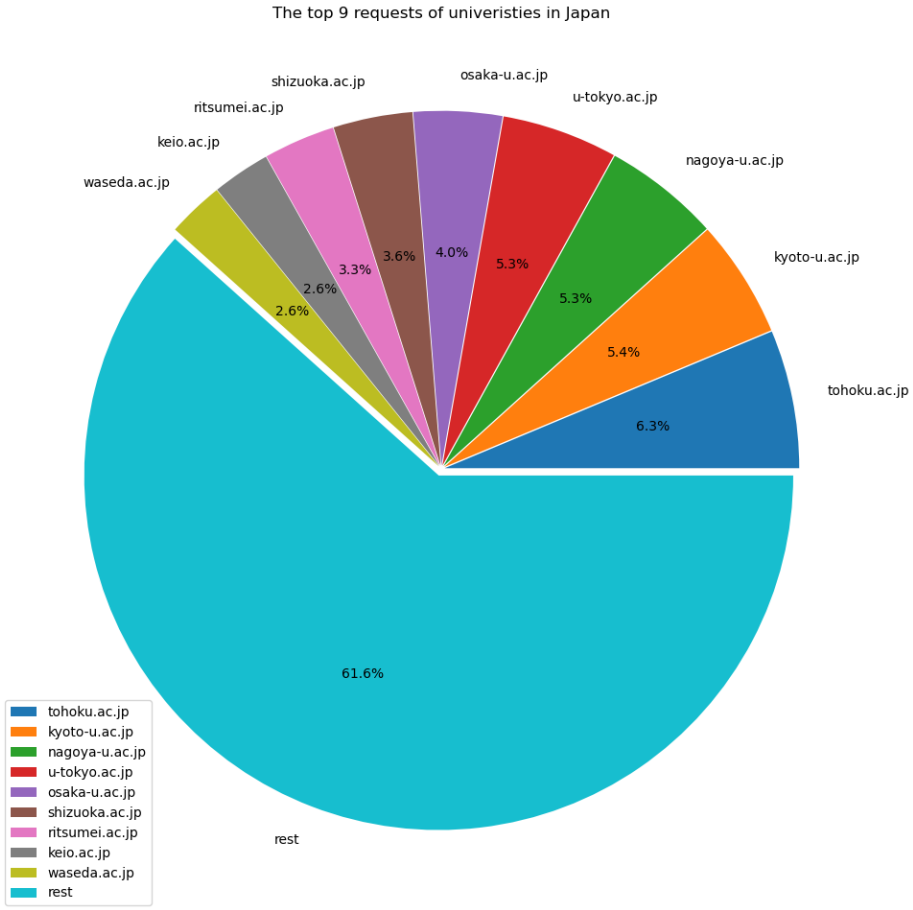
Top 9 of UK

host	count
hensa.ac.uk	4257
rl.ac.uk	1158
ucl.ac.uk	1036
man.ac.uk	921
ic.ac.uk	851
soton.ac.uk	808
bham.ac.uk	629
shef.ac.uk	623
le.ac.uk	616

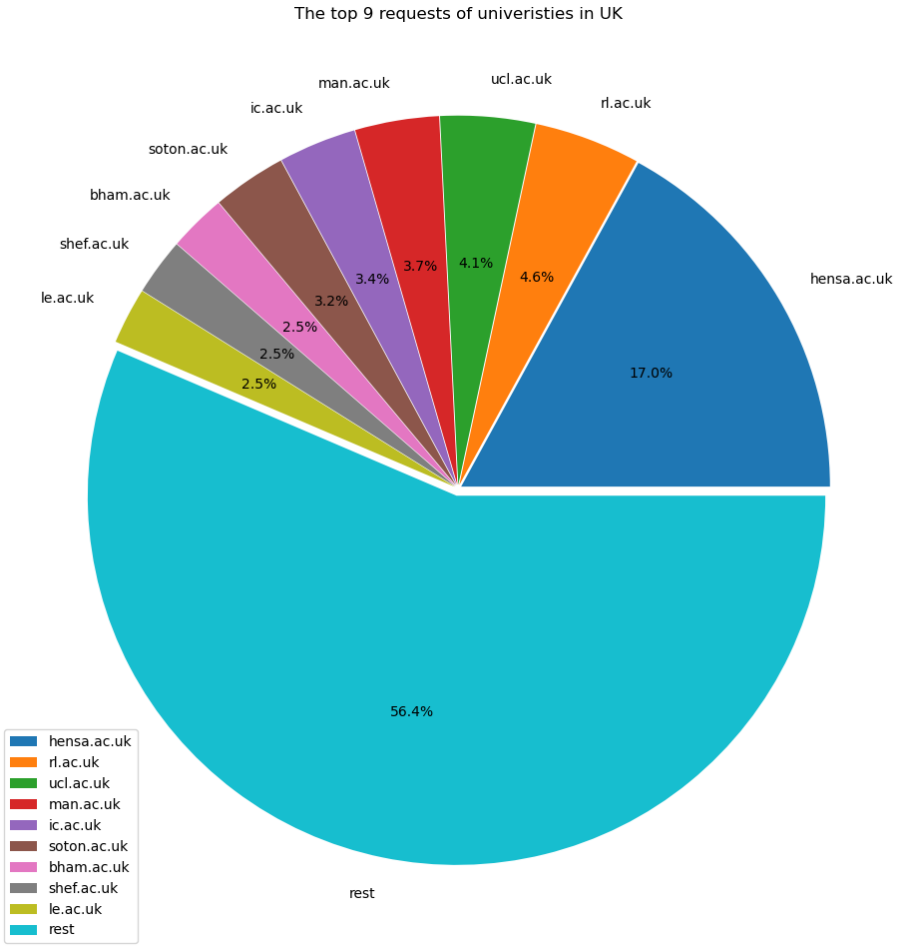
Top 9 of US

host	count
tamu.edu	6062
berkeley.edu	5439
fsu.edu	4418
umn.edu	4404
mit.edu	3966
washington.edu	3925
uiuc.edu	3750
utexas.edu	3665
cmu.edu	3244

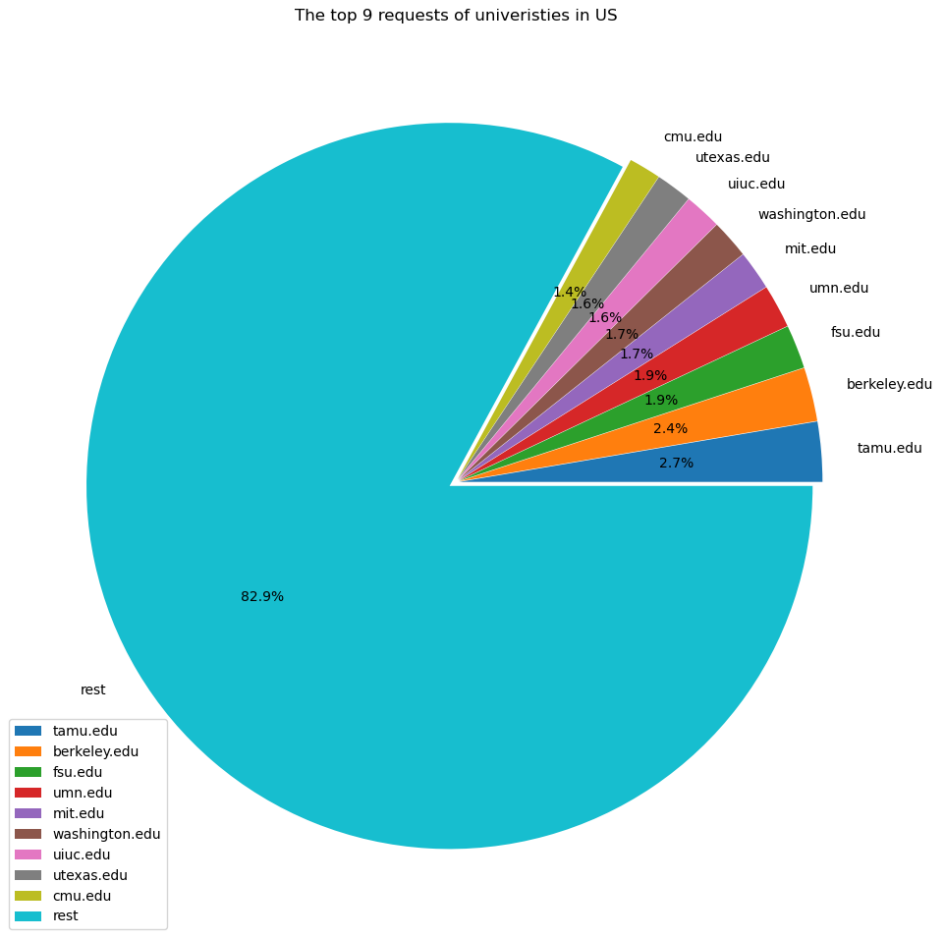
2)



1 Pie chart of universities in Japan

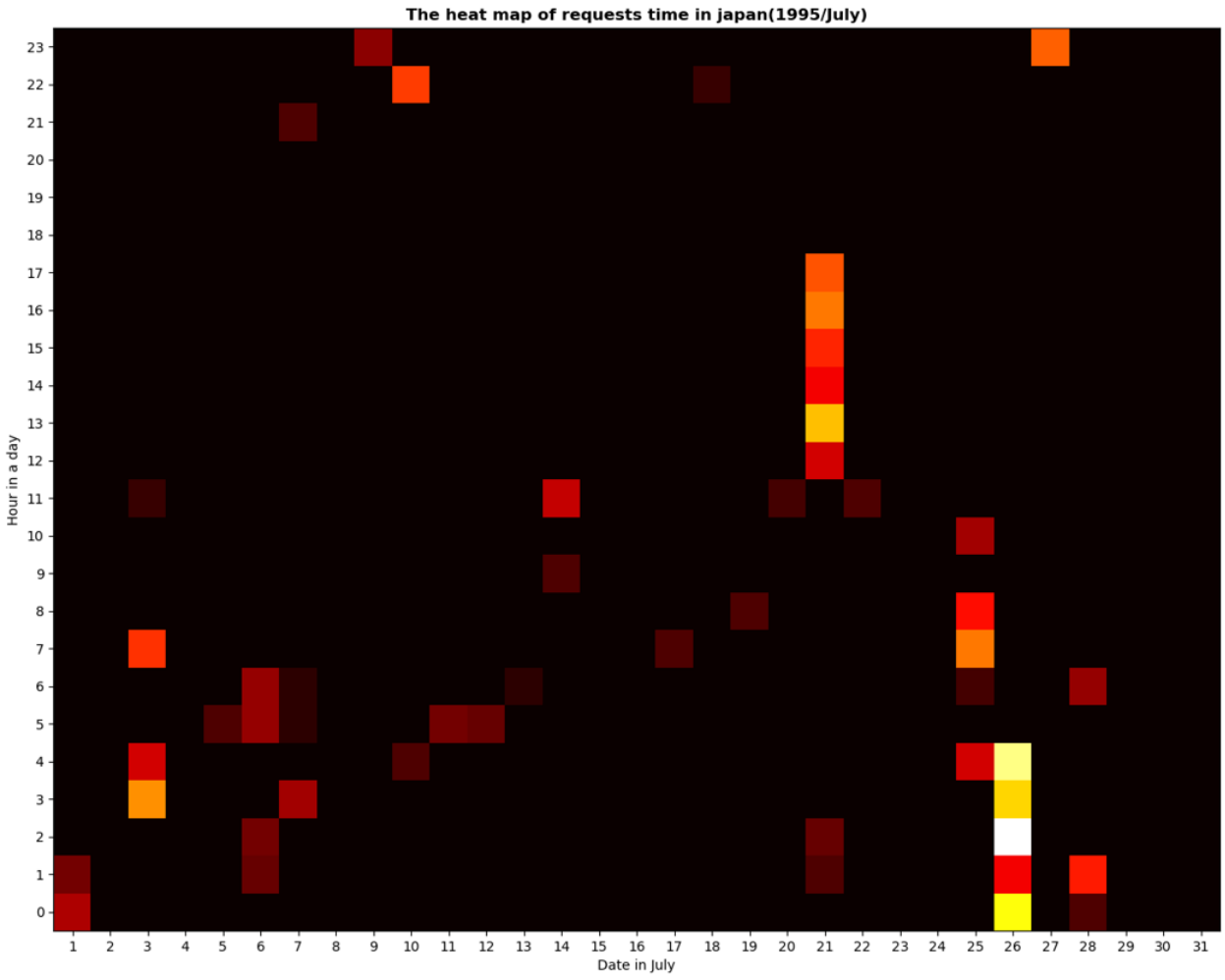


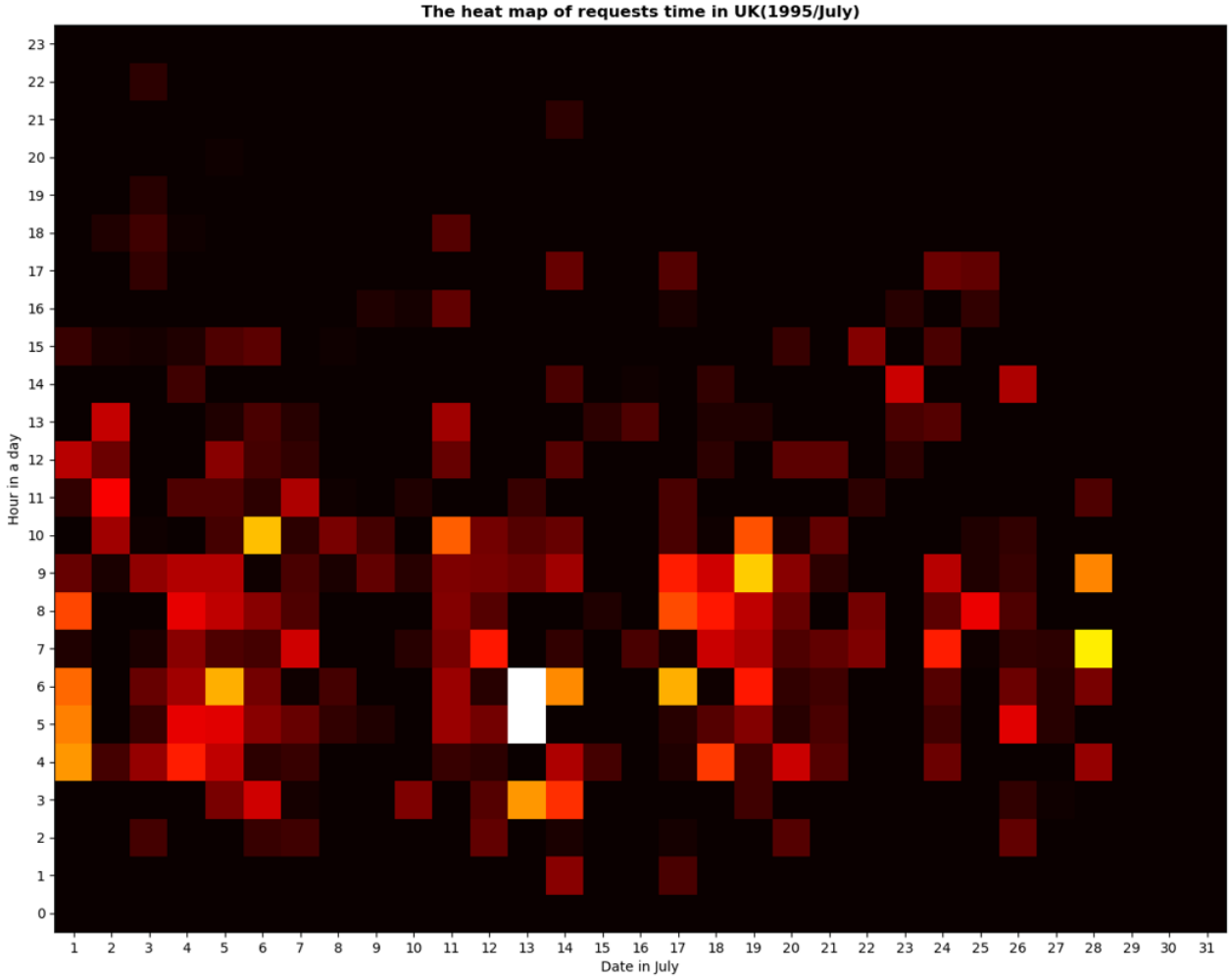
2 Pie chart of universities in UK

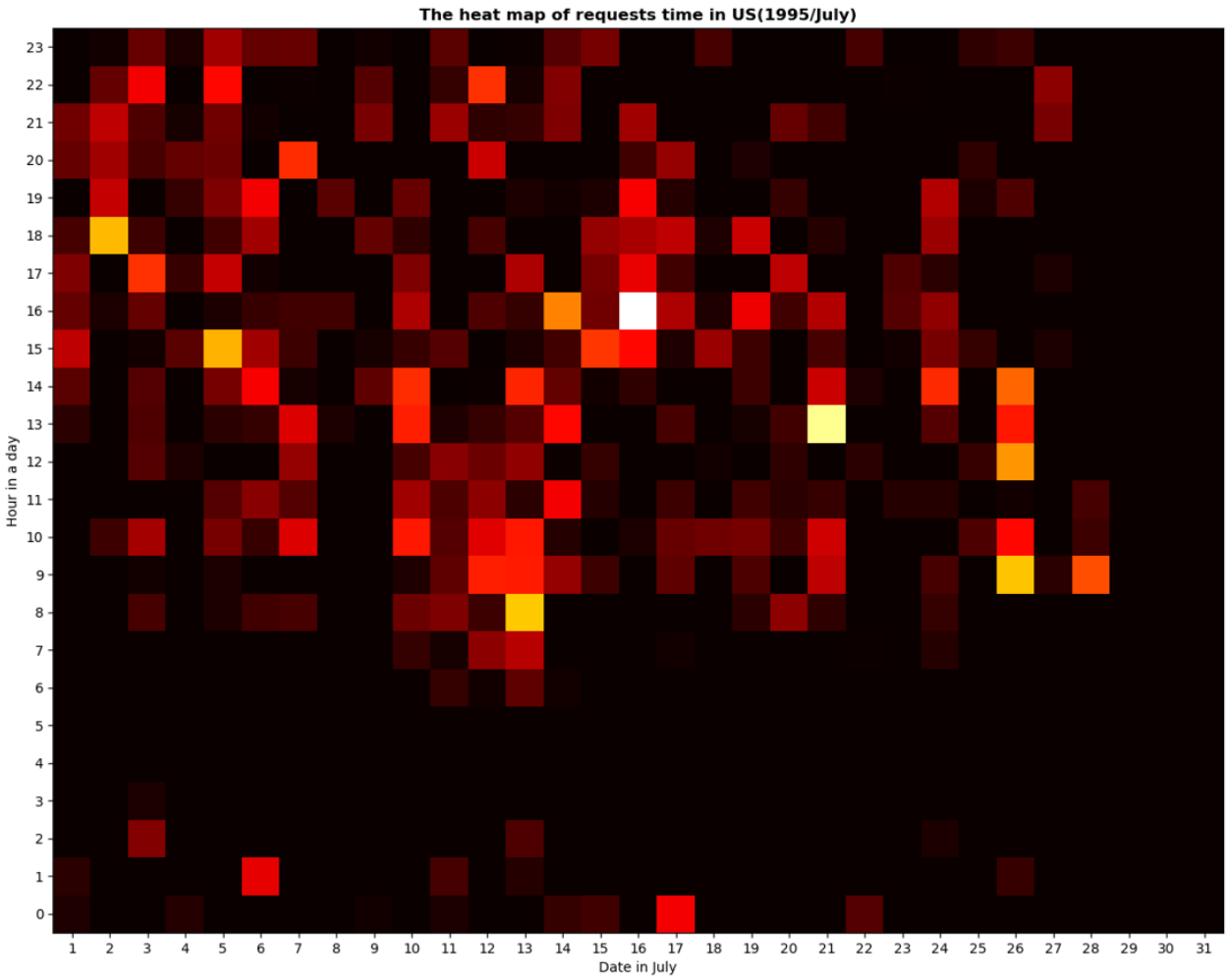


3 Pie chart of universities in US

C.




The heat map of requests time in UK(1995/July)
Hour in a day
Date in July


The heat map of requests time in US(1995/July)
Hour in a day
Date in July

D.

Observation 1：

The pie chart shows that the top 9 schools in the US account for less than 17.7% of all visits, while the top 9 schools in Japan and the top 9 schools in the UK account for 38.4% and 43.6% respectively over the same period. The bar chart also shows that the number of visits to universities in the USA is much higher than the number of visits to universities in the UK and Japan.

Reason：

This is probably mainly due to the fact that there are more colleges and universities in the US, so the total number is larger. In terms of the total number of visits to US schools, it is also much higher than in the UK and Japan.

For NASA：

NASA could consider deploying more servers in the US, which would improve access speeds for most users.

Observation 2：

The heat map shows that US and UK users are used to visiting NASA during the day, with most visits from Japanese users concentrated in the early hours of the morning. The three countries experience a low level of access every five days or so.

Reason：

The different access times may be due to the different time zones in which the Japan and NASA servers are located. The period of low level of access are caused by weekend.

For NASA：

The low number of visits to the table at weekends allows for an appropriate reduction in the hours of maintenance staff.

Question 2. Movie Recommendation and Analysis [15 marks]

A. Time-split Recommendation

1) Perform time-split recommendation with three training data sizes: 50%, 65%, and 80%. [2 marks]

- The size of percent_50_train: 13876722; the size of percent_50_test: 13876722
- The size of percent_65_train: 18039738; the size of percent_35_test: 9713706
- The size of percent_80_train: 22202755; the size of percent_20_test: 5550689

2) For each of the three splits above, study **three** versions of ALS using your student number as the seed as the following [2 marks]

- als1 = ALS(userCol="userId", itemCol="movieId", seed=myseed, coldStartStrategy="drop")
- als2 = ALS(userCol="userId", itemCol="movieId", seed=myseed, **regParam = 0.01**, coldStartStrategy="drop")
- als3 = ALS(userCol="userId", itemCol="movieId", seed=myseed, **maxIter=15**, nonnegative = True, coldStartStrategy="drop")

i. The first setting is the original one.

ii. The second setting is based on the first one and reduce the regParam to 0.01. The reason：by reducing the parameter, the fit of model will be improved and may reduce the loss. However, after testing, loss will increase no matter the parameters become larger or smaller, which indicates that the original regular is just right and there is no underfit or overfit.

iii. The third on is based on the first one as well, but increased number of iterations to 15. An increase in the number of iterations will result in more adequate training, which is likely to improve the fit and thus reduce the loss. Here, when the number of iterations is increased, the result is poor, which indicates that there may be an overfitting problem, which increases the loss of the model in the test set.

3) For each split and each version of ALS, compute three metrics: the Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE).

MODEL	SPLIT	RMSE	MSE	MAE
ALS1	50-50	0.788	0.621	0.597
ALS1	65-35	0.806	0.649	0.603
ALS1	80-20	0.860	0.739	0.645
ALS2	50-50	0.825	0.681	0.616
ALS2	65-35	0.867	0.752	0.638
ALS2	80-20	0.939	0.882	0.688
ALS3	50-50	0.790	0.624	0.599
ALS3	65-35	0.807	0.651	0.608
ALS3	80-20	0.860	0.740	0.645

B. **User Analysis**

1) Report the size of each cluster (number of users) in one **Table**, in total 3 splits x 3 clusters = 9 numbers. [2 marks]

SPLIT	Top1-size	Top2-size	Top3-size
Data-split: 50-50	14249	13534	9823
Data-split: 65-35	21761	14508	14182
Data-split: 80-20	22943	18832	16854

2) Report 3 splits x 5 genres x 2 sets = 30 genres in one **Table**. [3 marks]

Data	1	2	3	4	5
Train 50	Drama	Comedy	Romance	Thriller	Action
Test 50	Drama	Comedy	Romance	Thriller	Action
Train 65	Drama	Comedy	Romance	Thriller	Action
Test 35	Drama	Comedy	Thriller	Romance	Action
Train 80	Drama	Comedy	Romance	Thriller	Action
Test 20	Drama	Comedy	Thriller	Romance	Action

C. Discuss two most interesting observations from the above, each with three sentences: 1) what is the observation? 2) what are the possible causes of the observation? 3) how useful is this observation to a movie website such as Netflix? [2 marks]

Observation 1

As can be seen in A 3), although more data were trained, the results became worse, which is not in line with normal experience.

Possible reasons:

The reason for this phenomenon may be due to the fact that the data is sorted based on time, rather than randomly, and the movie reviews may be time-sensitive.

How useful:

Movie websites should consider timeliness when building their recommendation systems and try to train on the latest dataset to get better results.

Observation 2

Despite training on different datasets, the TOP2 results obtained on the maximum clustering are Drama and Comedy.

Possible reasons:

- The first possibility is that the clustering results do not vary much despite the different sizes of the training sets, which also proves that the model is correct.
- The second possibility: these categories of movies are more likely to be scored high in all clusters.

How useful:

Therefore, the movie website can display these two types of movies in the "High Rated Movies" section of the home page for users in the biggest cluster.