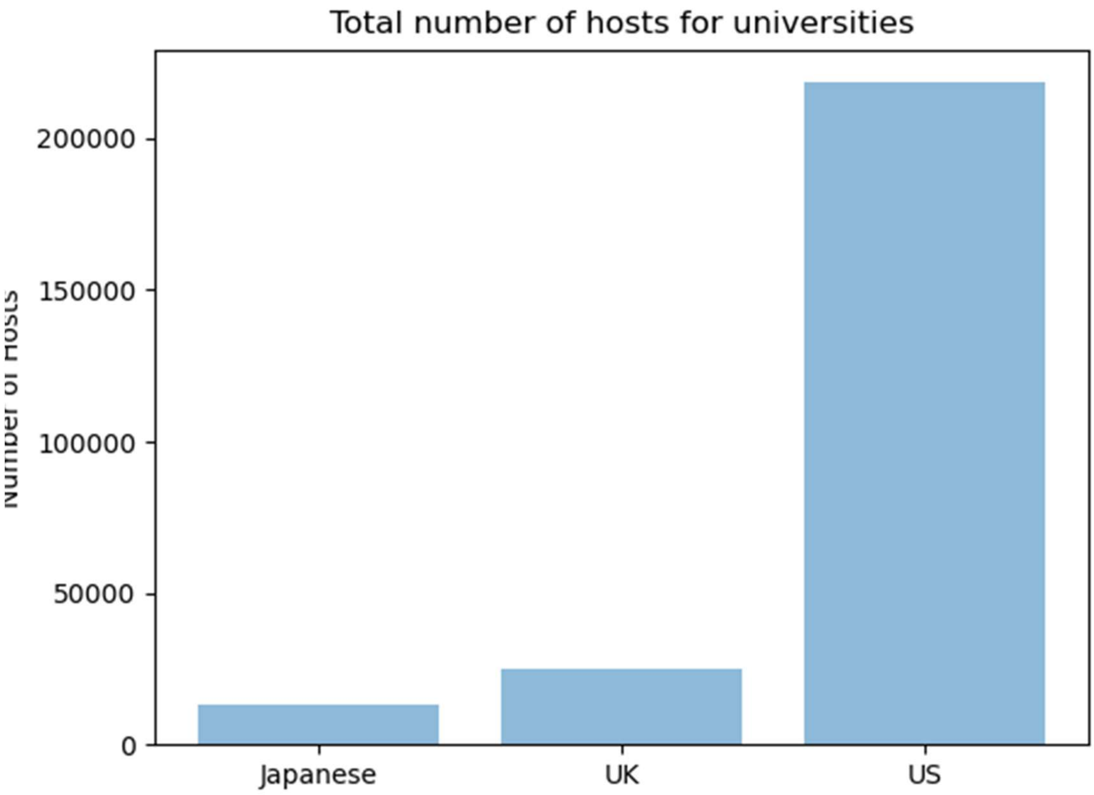
COM6012 Assignment 1

Question 1: Log Mining and Analysis

A:

- The total number of requests for all hosts from Japanese Universities ending in “.ac.jp” is 13067
- The total number of requests for all hosts from UK Universities ending in “.ac.uk” is 25009
- The total number of requests for all hosts from US Universities ending in “.edu” is 218449



Graph for the total number of hosts for universities from the three countries

B:

B1:

The 9 most frequent Universities according to the domain:

Japan:

1. Tohoku.ac.jp
2. Kyoto-u.ac.jp
3. Nagoya-u.ac.jp
4. u-tokyo.ac.jp
5. Osaka-u.ac.jp
6. Shizuoka.ac.jp
7. Ritsumei.ac.jp
8. Keio.ac.jp
9. Waseda.ac.jp

UK:

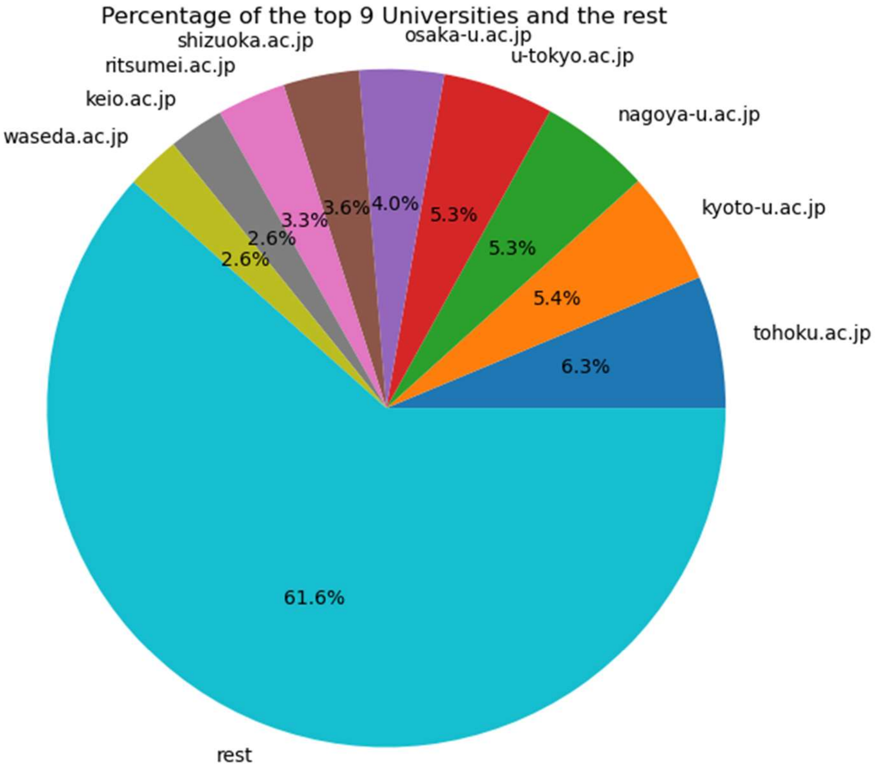
1. Hensa.ac.uk
2. Rl.ac.uk
3. Ucl.ac.uk
4. Man.ac.uk
5. Ic.ac.uk
6. Soton.ac.uk
7. Bham.ac.uk
8. Shef.ac.uk
9. Le.ac.uk

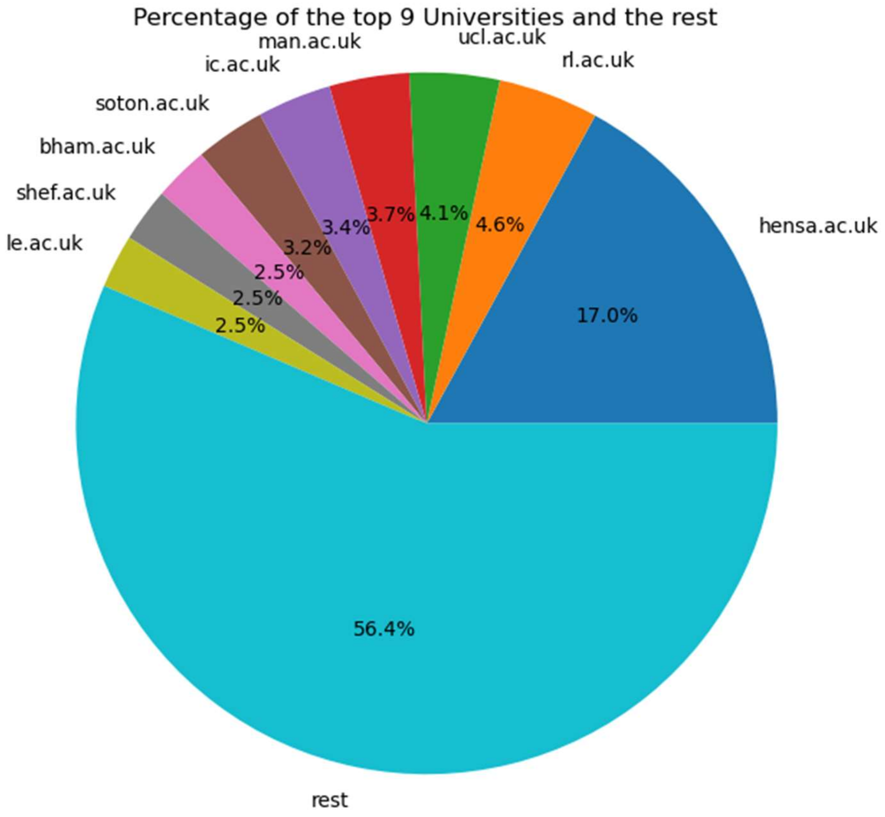
US:

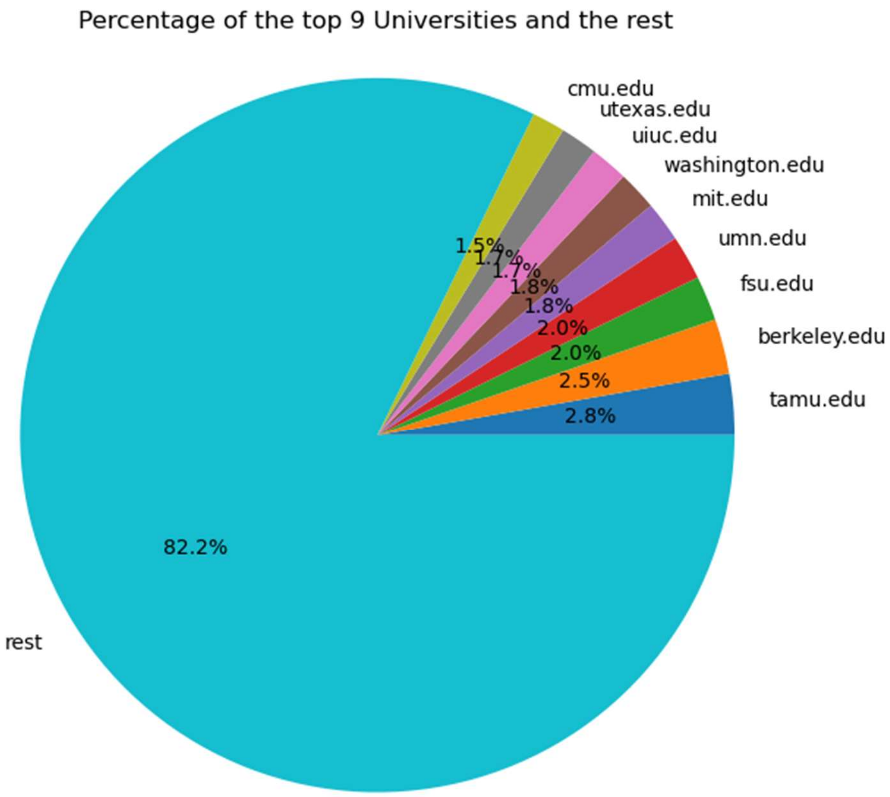
1. Tamu.edu
2. Berkeley.edu
3. Fsu.edu
4. Umn.edu
5. Mit.edu
6. Washington.edu
7. Uiuc.edu
8. Utexas.edu
9. Cmu.edu

B2:

Piecharts for each of the three countries top 9 universities




Percentage of the top 9 Universities and the rest
man.ac.uk
ucl.ac.uk
rl.ac.uk
ic.ac.uk
soton.ac.uk
bham.ac.uk
shef.ac.uk
le.ac.uk
hensa.ac.uk
3.4%
3.7%
4.1%
4.6%
3.2%
2.5%
2.5%
2.5%
17.0%
56.4%
rest



D:

Observation 1:

There are considerably more requests from American universities than UK and Japanese Universities. A possible cause of the this is that NASA is an American Space agency so it is to be expected that there are considerably more requests from American universities studying the data. This is useful for NASA as they can be prepared for this greater volume of requests and ensure their servers can handle the increased server traffic during prime American working hours.

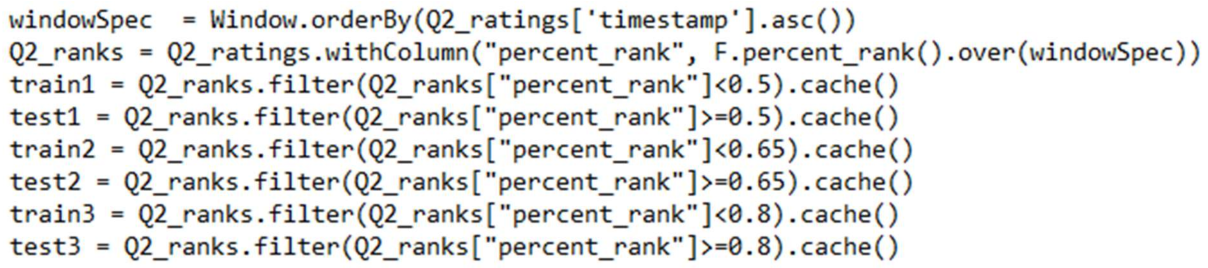
Observation 2:

Linked to the previous observation there are some periods of intense user traffic for both the American and UK universities compared to the Japanese universities. This is show by the highest number of requests in an hour timestamp for the three countries with America having 143, the UK 115 and Japan considerably lower at 57. The reasoning for this could be due to US and UK universities having more widespread projects on the NASA data leading to a higher number of users in a given period. This may be useful for NASA in determining and projecting the server traffic from the different countries to ensure they can handle it.

Question 2: Movie Recommendation and Analysis

A: Time-split Recommendation

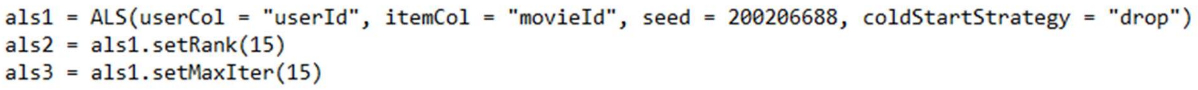
1.



```
windowSpec  = Window.orderBy(Q2_ratings['timestamp'].asc())
Q2_ranks = Q2_ratings.withColumn("percent_rank", F.percent_rank().over(windowSpec))
train1 = Q2_ranks.filter(Q2_ranks["percent_rank"]<0.5).cache()
test1 = Q2_ranks.filter(Q2_ranks["percent_rank"]>=0.5).cache()
train2 = Q2_ranks.filter(Q2_ranks["percent_rank"]<0.65).cache()
test2 = Q2_ranks.filter(Q2_ranks["percent_rank"]>=0.65).cache()
train3 = Q2_ranks.filter(Q2_ranks["percent_rank"]<0.8).cache()
test3 = Q2_ranks.filter(Q2_ranks["percent_rank"]>=0.8).cache()
```

The three splits were made using the following code with windowSpec and Q2_ranks used to perform the splits according to the sorted timestamp.

2.



```
als1 = ALS(userCol = "userId", itemCol = "movieId", seed = 200206688, coldStartStrategy = "drop")
als2 = als1.setRank(15)
als3 = als1.setMaxIter(15)
```

The following ALS parameters were used.

The first model was the ALS setting in lab3 with my student number as a seed.

The second model used a rank of 15 instead of the default 10 in ALS1. The reasoning for this is to increase the number of latent factors as the more hidden factors correctly used by the model the better the performance.

The third model used a max iter value of 15 instead of the default 10. This was done because using a higher number of iterations gives the model more opportunity to learn the parameters and therefore provide better predictions. However, setting this value too large may lead to overfitting of the training set.

3.

Table of the performance metrics for the three models on the three splits

Metric	Split	ALS1	ALS2	ALS3
RMSE	1	0.788	0.788	0.786
	2	0.807	0.806	0.805
	3	0.859	0.857	0.856
MSE	1	0.622	0.621	0.618
	2	0.652	0.650	0.648
	3	0.739	0.735	0.733
MAE	1	0.597	0.598	0.596
	2	0.607	0.606	0.605
	3	0.645	0.644	0.642

B: User Analysis

1:

	Largest User Cluster	Second Largest	Third Largest
Split 1	13926	12771	11090
Split 2	17280	15529	14970
Split 3	20496	19565	18182

2.

	Train top 5 genres	Test top 5 genres
Split 1	Drama, comedy, thriller, action, romance	Drama, comedy, thriller, action, romance
Split 2	Drama, comedy, thriller, action, romance	Drama, comedy, thriller, romance, action
Split 3	Drama, comedy, thriller, action, romance	Drama, comedy, thriller, romance, action

C:

Observation 1:

Either increasing the rank or the number of iterations effectively improve the performance of the ALS model by reducing the errors on the test data. With increasing the number of iterations leading to the greatest improvement in performance. The reasoning for these improvements is the model is given more time to learn the factors in the training data and with an increased rank is able to learn more hidden latent factors. This observation may be useful for movie websites to help them to build more accurate models that can provide better movie recommendations.

Observation 2:

The top 5 genres in the test and training set are consistent with slight changes in the ranking of the top 5. This is most likely because these genres are the broadest and encompass the largest number of films compared to more specific genres. However, it is also useful to Netflix to some extent as it illustrates that these genres are the most popular and therefore will be more likely to be recommended.