acy20zw_200206297_AS1

March 2021

1 Question 1

1.1 A

==========1==========
There are 13077 hosts form Japanese universities
==========2==========
There are 25014 hosts form UK universities
==========3==========
There are 227494 hosts form US universities

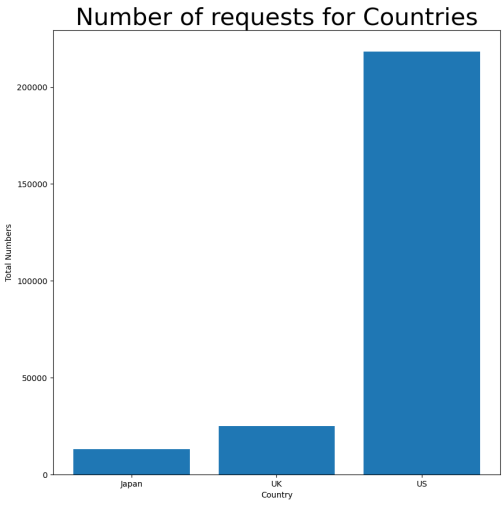

Figure 1: Bar

1.2 B

1)

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

(a) Top9 of Japan

University	count
hensa	4257
rl	1158
ucl	1036
man	921
ic	851
soton	808
bham	629
shef	623
le	616

(b) Top9 of UK

University	count
tamu	6062
berkeley	5439
fsu	4418
umn	4404
mit	3966
washington	3893
uiuc	3750
utexas	3665
cmu	3244

(c) Top9 of US

Figure 2: Top9 of Each Country

2)

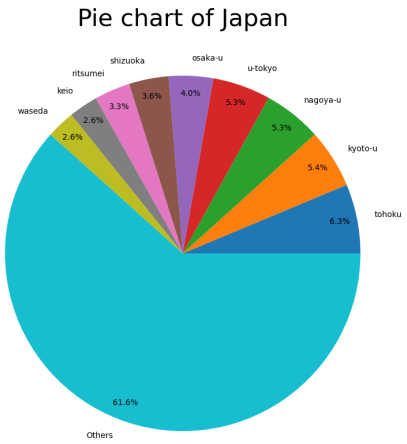

Figure 3: Pie Chart of Japan

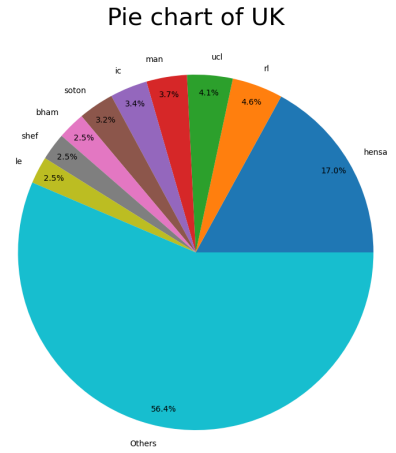

Figure 4: Pie Chart of UK

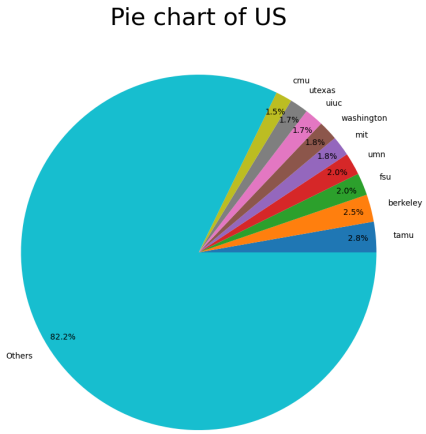

Figure 5: Pie Chart of US

1.3 C

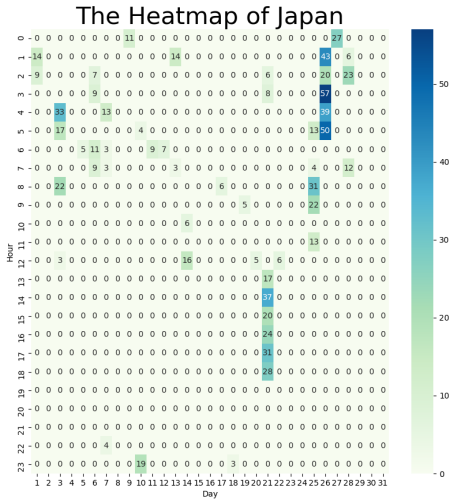

Figure 6: Heatmap of Japan

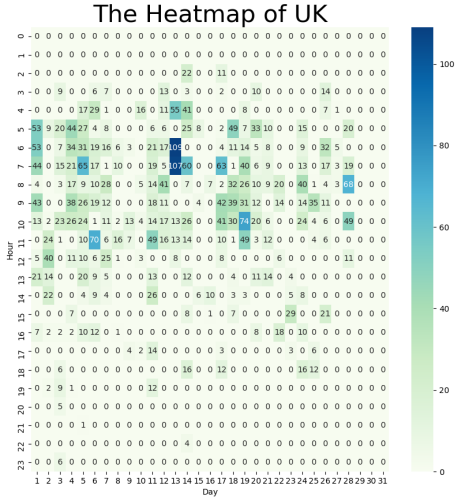

Figure 7: Heatmap of UK

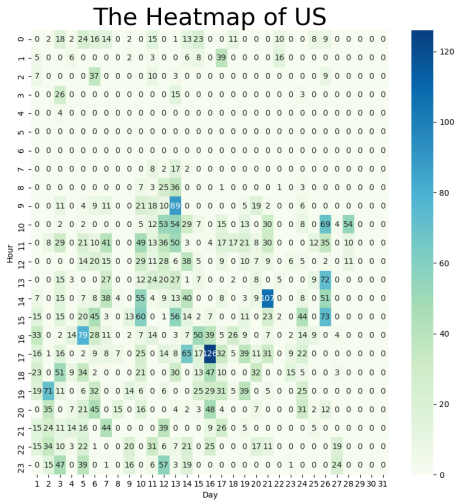

Figure 8: Heatmap of US

1.4 D

For A:
We can notice that the number of hosts from US university is the highest.
NASA might hold some special activities for students in universities which attract those people in school visit more times. Also, NASA belongs to the US which can attract more people who live in the US.
This observation can show that NASA is more interesting to US universities and they can try to take some measures to attract those people in other countries.

For B:
Firstly, it is clear to see that the number of hosts from US universities is much higher than in other countries. Secondly, hosts in Japan and the UK are more concentrated than hosts in US.
It might means NASA did not be widely accepted by the UK and Japan because there is not a lot of variety of hosts.
NASA can try to attract more people from different universities in Japan and the UK to Increase influence
For C:
From the heatmap, it seems like people in US universities are more likely to visit NASA's website in the daytime while people in UK and Japan are more likely to visit the website at midnight or before dawn.
The time difference may cause the situation.
It is possible for NASA to optimize its website based on the heatmap to arrange resource at different time for people in different countries.

2 Question 2

2.1 A

The ALS1 means ALS model with the parameter maxIter=15 which means the max iteration. It can automatically stop training while it reaches the max iteration which can save our time.

The ALS2 means ALS model with the parameter regParam = 0.1 which means regularization parameters. The regularization parameters can avoid overfitting which means the model's performance can be improved.

	ALS default			ALS 1			ALS 2		
split	0.5	0.65	0.8	0.5	0.65	0.8	0.5	0.65	0.8
RMSE	0.790	0.809	0.861	0.789	0.808	0.859	0.790	0.808	0.859
MSE	0.624	0.654	0.741	0.621	0.652	0.738	0.624	0.652	0.738
MAR	0.560	0.609	0.646	0.597	0.607	0.644	0.599	0.607	0.644

2.2 B

2.2.1 1

Split	Cluster 1	Cluster 2	Cluster 3
0.5	12276	11836	11127
0.65	17032	16985	14257
0.8	20995	19019	16497

2.2.2 2

Split	Trainset	Testset
0.5	Drama, Comedy, Romance, Thriller, Action	Drama, Comedy, Romance, Thriller, Action
0.65	Drama, Comedy, Romance, Thriller, Action	Drama, Comedy, Romance, Thriller, Documentary
0.8	Drama, Comedy, Thriller, Romance, Action	Drama, Comedy, Thriller, Action, Romance

2.3 C

For A:
Although we used different parameters to improve the model's performance, the results showed that the performance did not improved. Actually, it might because of we did not set the best parameter for the model. For those movie websites, more sets of parameters can be used to optimize the model to get the best performance.
For B:
Different splits of datasets were used and the results showed Drama, Comdey, Romance, Thirller, Action, Documentary movies are most popular. It should

because most people like to watch those movies. For those movie websites, they can recommend more movies which related to those topics.