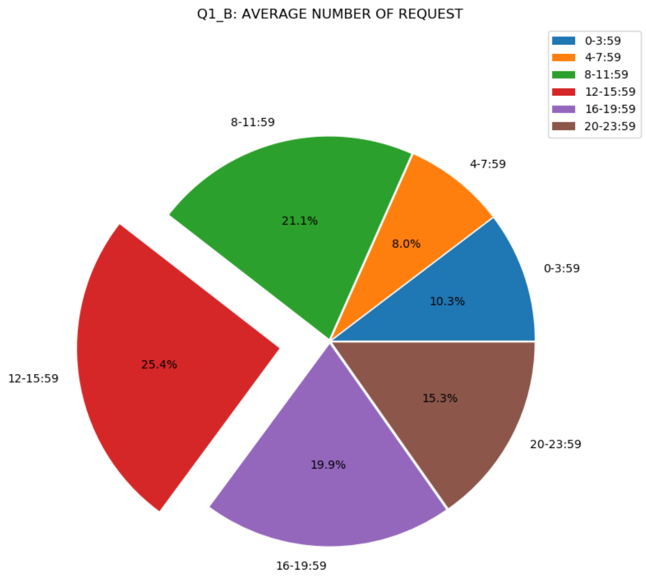
Assignment 1

Question 1:

A:

AVERAGE AMOUNT OF REQUEST	
0:00:00 - 3:59:59	7078.964285714285
4:00:00 - 7:59:59	5479.392857142857
8:00:00 - 11:59:59	14462.357142857143
12:00:00 - 15:59:59	17377.785714285714
16:00:00 - 19:59:59	13581.62962962963
20:00:00 - 23:59:59	10438.962962962964

B:



TWO OBSERVATION:

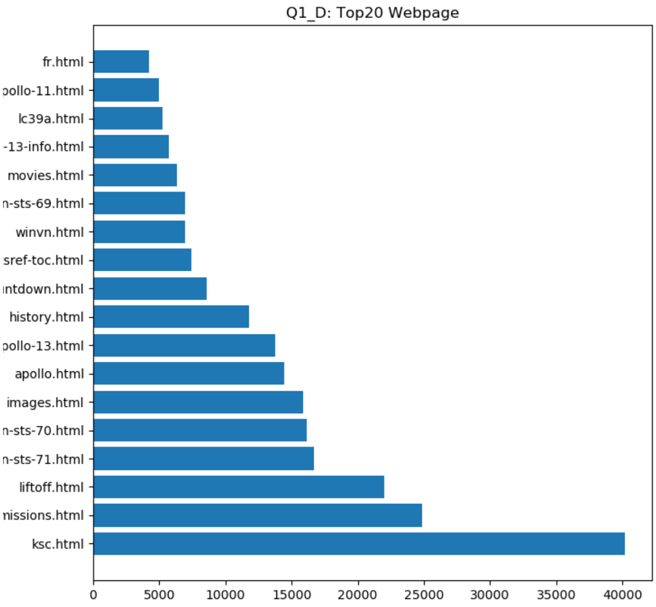
firstly, there is the highest percentage of number of request from 12:00 to 16:00. the second is between 8:00 to 12:00.

secondly, there is a tendency that the number of visiting will be increasing in the morning until the midday. Form the midday to evening, there is a decreasing trend until midnight, which follows people's daily schedule.

C:

Webpage	Count
ksc.html	count: 40226
missions.html	count: 24864
liftoff.html count	count: 22000
mission-sts-71.html	count: 16717
mission-sts-70.html	count: 16123
images.html	count: 15897
apollo.html	count: 14472
apollo-13.html	count: 13768
history.html	count: 11816
countdown.html	count: 8572
stsref-toc.html	count: 7420
winvn.html	count: 6970
mission-sts-69.html	count: 6968
images.html	count: 6713
movies.html	count: 6308
movies.html	count: 6109
apollo-13-info.html	count: 5747
lc39a.html	count: 5260
apollo-11.html	count: 5004
fr.html	count: 4218

D:



TWO OBSERVATION:

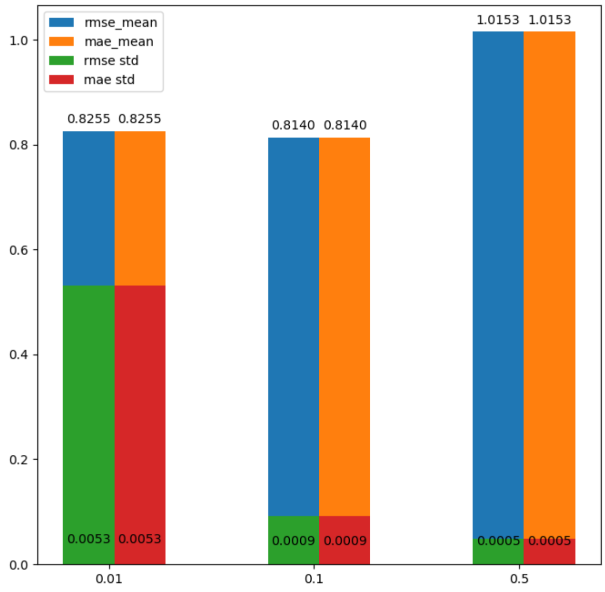
Firstly, the highest requested webpage is ksc.html. the request could reach 40226 per month. I guess this page might be the index or an important webpage of NASA website. the second is missions.html, which is about 2/3 of ksc.html.

Secondly, it seems that mission category is more popular in NASA website. there are three four pages about mission in top 20. I consider NASA could make mission category more attractive to attract more visitor.

Question2:

A:

		0.01	0.1	0.5
RMSE	1	0.8222506231411222	0.8131254607624744	1.0145943326055813
	2	0.8330245391849891	0.815291723535879	1.0156963219015294
	3	0.8213421119383338	0.8136903467813341	1.0154843776108782
	Mean	0.82553909	0.81403584	1.01525834
	Std	0.00530599	0.0009175	0.00047743
MAE	1	0.8222506231411222	0.8131254607624744	1.0145943326055813
	2	0.8330245391849891	0.815291723535879	1.0156963219015294
	3	0.8213421119383338	0.8136903467813341	1.0154843776108782
	Mean	0.82553909	0.81403584	1.01525834
	Std	0.00530599	0.0009175	0.00047743



B:

TWO OBSERVATION:

firstly, it seems that 0.1 is a good hyperparameter for regularisation term because it achieve the best performance. with the increasing of it, the error will increasing obviously. the reason is that bigger regularisation will cause the model underfitting.

secondly, the std is bigger when the regularisation hyperparameter is smaller. this phenomenon could be explained as follow. smaller regularisation will cause overfitting for a model. therefore, the performance will be fluctuated in variant dataset.

C:

	Top3 Cluster	Tag 1	Count	Tag 2	Count	Tag 3	Count
Fold 1	**1**	original	2065	mentor	2065	catastrophe	2065
	2	original	2584	mentor	2584	Runaway	2584
	3	original	1553	mentor	1553	Catastrophe	1553
Fold 2	**1**	original	1987	mentor	1987	Storytelling	1987
	2	original	1469	mentor	1469	catastrophe	1469
	3	original	1853	mentor	1853	runaway	1853
Fold 3	**1**	original	2466	mentor	2466	criterion	2466
	2	original	1992	mentor	1992	catastrophe	1992
	3	original	1405	mentor	1405	catastrophe	1405

D:

TWO OBSERVATION:

Firstly, original and mentor are the major tags for both three fold. The difference of tag is in the third tag.

Secondly, it seems that there are many movies that are about original, mentor, catastrophe as well as original, mentor, runaway.