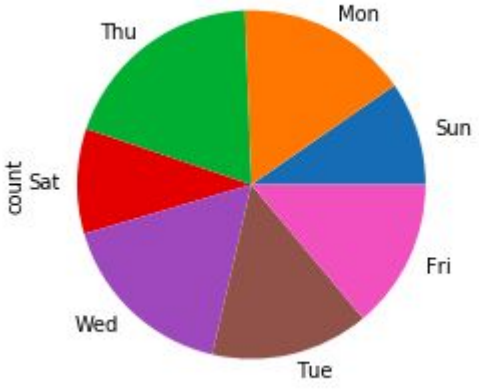
COM6012 Assignment 1(180128033)-

Answer-1: (a) : Average number of requests on each of the seven days in a week?

Table:

Day	Count	Average number of requests
Sunday	182590	45647.5
Monday	301684	75421.0
Tuesday	277840	69460.0
Wednesday	318698	79674.5
Thursday	363436	90859.0
Friday	263086	65771.5
Saturday	184380	46095.0

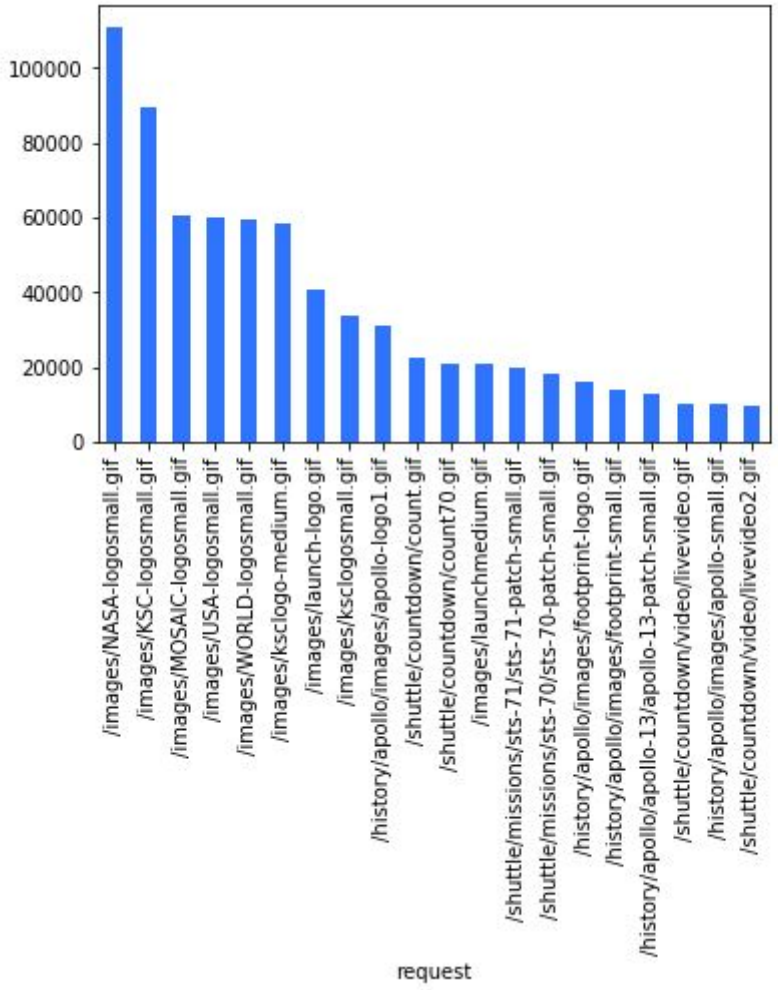
(b) The highest number of the request was made on weekdays as compared to the weekends. Thursday being the busiest day followed by Wednesday, Monday, Tuesday, Friday, Saturday and then Sunday. The number of requests made are more in the mid of the week in comparison to the starting and ending of the week.



(c) The top 20 most requested .gif images.

```
+-------------------------------------------------------+------+
|request                                                |count |
+-------------------------------------------------------+------+
|/images/NASA-logosmall.gif                             |111144|
|/images/KSC-logosmall.gif                              |89530 |
|/images/MOSAIC-logosmall.gif                           |60300 |
|/images/USA-logosmall.gif                              |59845 |
|/images/WORLD-logosmall.gif                            |59325 |
|/images/ksclogo-medium.gif                             |58616 |
|/images/launch-logo.gif                                |40841 |
|/images/ksclogosmall.gif                               |33555 |
|/history/apollo/images/apollo-logo1.gif                |31052 |
|/shuttle/countdown/count.gif                           |22189 |
|/shuttle/countdown/count70.gif                         |20921 |
|/images/launchmedium.gif                               |20788 |
|/shuttle/missions/sts-71/sts-71-patch-small.gif        |19832 |
|/shuttle/missions/sts-70/sts-70-patch-small.gif        |18135 |
|/history/apollo/images/footprint-logo.gif              |16168 |
|/history/apollo/images/footprint-small.gif             |13912 |
|/history/apollo/apollo-13/apollo-13-patch-small.gif    |13004 |
|/shuttle/countdown/video/livevideo.gif                 |10143 |
|/history/apollo/images/apollo-small.gif                |10089 |
|/shuttle/countdown/video/livevideo2.gif                |9816  |
+-------------------------------------------------------+------+
only showing top 20 rows
```

(d) We have the highest number of the request containing ‘logosmall’. This because when you open a site, the first thing requested is the logo.So, therefore count for the ‘logosmall’ is more.



Question-2:
Answer-2 :

(a) Values of RMSE and MAE after performing five-fold cross-validation for ‘drop 1’ and ‘drop 2’.

RMSE for drop1	MAE for drop1	RMSE for drop2	MAE for drop2
1.0507194719084156	0.876472068225606	0.8084200659914854	0.6297645583653788
1.0510024868586683	0.8765889444628031	0.8084848318092444	0.6298081126669673

1.0510809006090822	0.876670198567985	0.8086484534716065	0.6299768841760041
1.0510549805228317	0.8768612958519997	0.8090755727049258	0.6303447087747456
1.0509762245020269	0.876748765575236	0.8088689637082364	0.6301166728039508

Mean and standard deviation for 'RMSE' and 'MAE'.

Average of rmse for cold start strategy(drop1)	1.050966812880205
Average of mae for cold start strategy(drop1)	0.876668254536726
Average of rmse for cold start strategy(drop2)	0.8086995775370998
Average of mae for cold start strategy(drop2)	0.6300021873574093
Standard deviation of rmse for cold start strategy(drop1)	0.000129095554359666098
Standard deviation of mae for cold start strategy(drop1)	0.0001330638286527319
Standard deviation of rmse for cold start strategy(drop2)	0.0002436883842275939
Standard deviation of mae for cold start strategy(drop2)	0.00021222658321550913

(b) According to the 2 approaches used in the above question, we are getting a higher value of average for 'RMSE' and 'MAE' for drop 1 having parameter (regParam=0.6) compared to the 'RMSE' and 'MAE' for drop 2 having parameter (regParam=0.1). Talking about 'Standard Deviation', results observed are opposite to the Average of 'RMSE' and 'MAE'. Standard Deviation of 'RMSE' and 'MAE' for drop2 is higher than drop1.

(c) Here, we are finding the top five tags for each of the top three largest clusters using clustering using k-means with k=20.

Clustering 1:

TagId	**Tag**
302	dialogue
452	good soundtrack

468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original
188	catastrophe
302	dialogue
468	great ending
646	mentor
742	original
188	catastrophe
302	dialogue
468	great ending
646	mentor
742	original
188	catastrophe
302	dialogue
468	great ending
646	mentor
742	original

Clustering 2:

TagId	**Tag**

188	catastrophe
302	dialogue
468	great ending
646	mentor
742	original
188	catastrophe
302	dialogue
468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original
188	catastrophe
302	dialogue
468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original

Clustering 3:

TagId	Tag
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original
302	dialogue
452	good soundtrack
468	great ending
646	mentor
742	original

(d) In the above question 2(c), we can observe that a lot of tags has been repeated. We can easily pick the good movie, music by looking for the positive feature like “great ending”, “good soundtrack”.