Assignment Part-2 Report

Question 1. Log Mining and Analysis

A)

Maximum and Minimum requests on each week day for July of 1995-

Lowest requests on Sunday are-35272 and Highest are-60265

Lowest requests on Monday are-64259 and Highest are-89584

Lowest requests on Tuesday are-62699 and Highest are-80407

Lowest requests on Wednesday are-58849 and Highest are-94575

Lowest requests on Thursday are-61680 and Highest are-134203

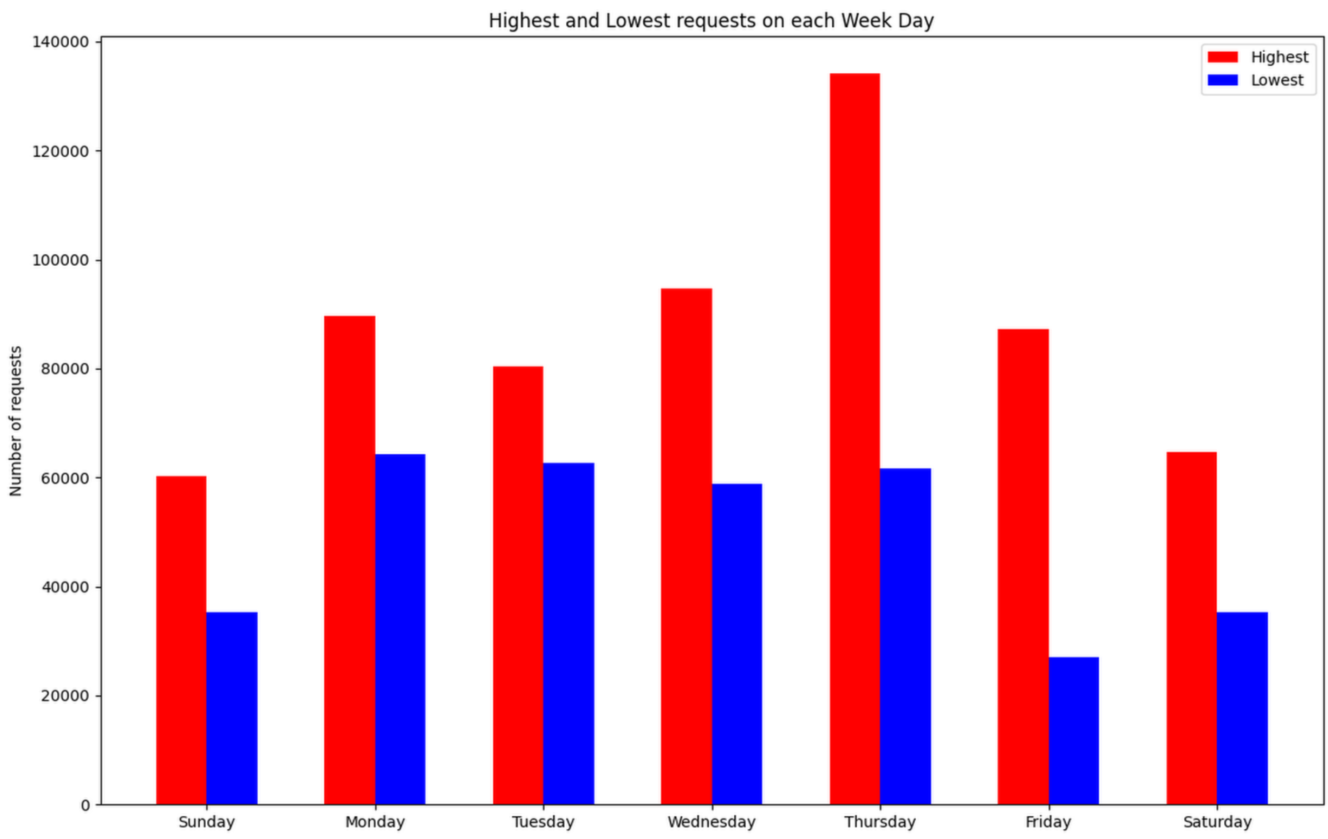
Lowest requests on Friday are-27121 and Highest are-87233

Lowest requests on Saturday are-35267 and Highest are-64714

WeekDay	Highest requests	Lowest requests
Sunday	60265	35272
Monday	89584	64259
Tuesday	80407	62699
Wednesday	94575	58849
Thursday	134203	61680
Friday	87233	27121
Saturday	64714	35267

B)

Visualise the 14 numbers in A above in ONE figure



C)

12 most requested and 12 least requested .mpg videos with full directory, method(GET/HEAD) and protocol(HTTP/1.0)

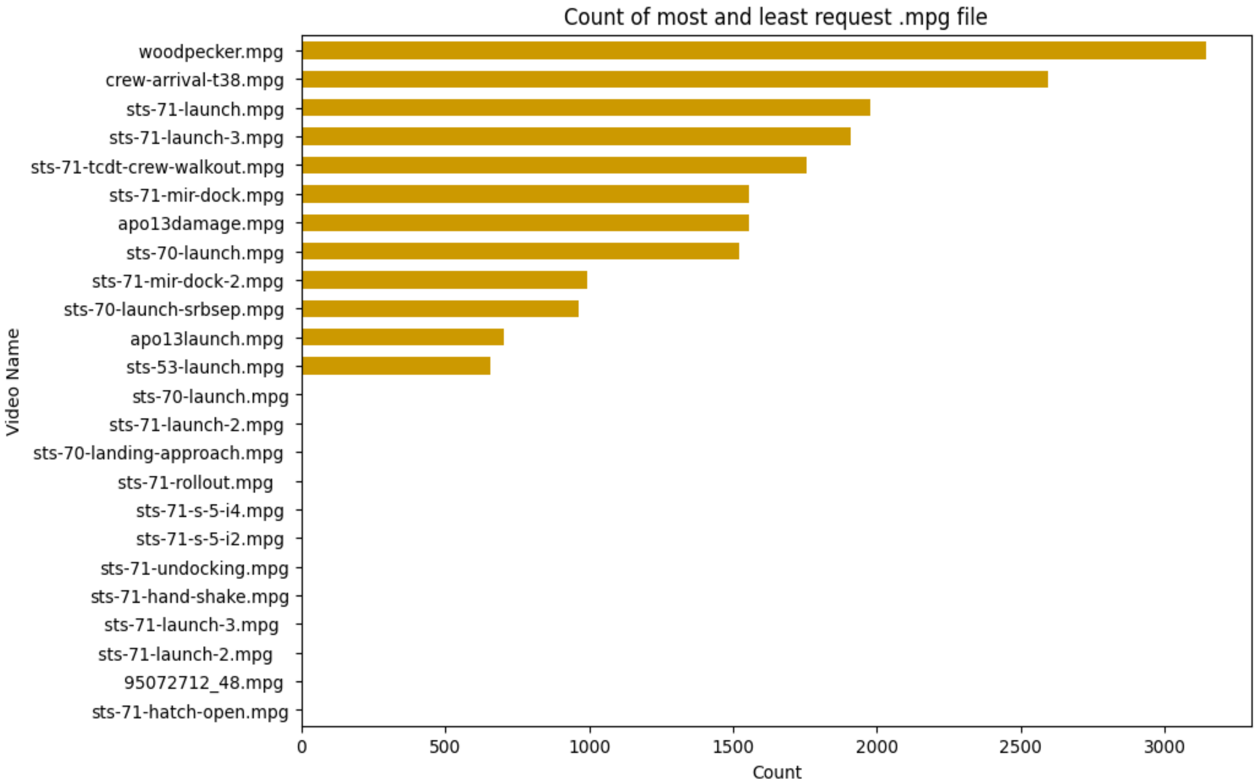
request	count
GET /shuttle/missions/sts-70/movies/woodpecker.mpg HTTP/1.0	3145
GET /shuttle/missions/sts-71/movies/crew-arrival-t38.mpg HTTP/1.0	2594
GET /shuttle/missions/sts-71/movies/sts-71-launch.mpg HTTP/1.0	1979
GET /shuttle/missions/sts-71/movies/sts-71-launch-3.mpg HTTP/1.0	1910
GET /shuttle/missions/sts-71/movies/sts-71-tcdt-crew-walkout.mpg HTTP/1.0	1758
GET /shuttle/missions/sts-71/movies/sts-71-mir-dock.mpg HTTP/1.0	1556
GET /history/apollo/apollo-13/movies/apo13damage.mpg HTTP/1.0	1555
GET /shuttle/missions/sts-70/movies/sts-70-launch.mpg HTTP/1.0	1523
GET /shuttle/missions/sts-71/movies/sts-71-mir-dock-2.mpg HTTP/1.0	993
GET /shuttle/missions/sts-70/movies/sts-70-launch-srbsep.mpg HTTP/1.0	964
GET /history/apollo/apollo-13/movies/apo13launch.mpg HTTP/1.0	702
GET /shuttle/missions/sts-53/movies/sts-53-launch.mpg HTTP/1.0	658
GET /shuttle/missions/sts-70/movies/sts-70-launch.mpg	1
HEAD /shuttle/missions/sts-71/movies/sts-71-launch-2.mpg HTTP/1.0	1
GET /shuttle/missions/sts-70/movies/sts-70-landing-approach.mpg HTTP/1.0	1
GET /shuttle/missions/sts-71/movies/sts-71-rollout.mpg HTTP/1.0	1
GET /shuttle/countdown/lps/sts-71-s-5-i4.mpg HTTP/1.0	1
GET /shuttle/countdown/lps/sts-71-s-5-i2.mpg HTTP/1.0	1
GET /shuttle/missions/sts-71/movies/sts-71-undocking.mpg	1
GET /shuttle/missions/sts-71/movies/sts-71-hand-shake.mpg	1
GET /shuttle/missions/sts-71/movies/sts-71-launch-3.mpg HTTP/1.0	1
GET /shuttle/missions/sts-71/movies/sts-71-launch-2.mpg HTTP/1.0	1
GET /wxworld/mpegs/MPEG6pNgmSfc9/95072712_48.mpg HTTP/1.0	1
GET /shuttle/missions/sts-71/movies/sts-71-hatch-open.mpg	1

⇨ 12 most requested and 12 least requested .mpg videos names only

video_name	total_number_of_requests
woodpecker.mpg	3145
crew-arrival-t38.mpg	2594
sts-71-launch.mpg	1979
sts-71-launch-3.mpg	1910
sts-71-tcdt-crew-walkout.mpg	1758
sts-71-mir-dock.mpg	1556
apo13damage.mpg	1555
sts-70-launch.mpg	1523
sts-71-mir-dock-2.mpg	993
sts-70-launch-srbsep.mpg	964
apo13launch.mpg	702
sts-53-launch.mpg	658
sts-70-launch.mpg	1
sts-71-launch-2.mpg	1
sts-70-landing-approach.mpg	1
sts-71-rollout.mpg	1
sts-71-s-5-i4.mpg	1
sts-71-s-5-i2.mpg	1
sts-71-undocking.mpg	1
sts-71-hand-shake.mpg	1
sts-71-launch-3.mpg	1
sts-71-launch-2.mpg	1
95072712_48.mpg	1
sts-71-hatch-open.mpg	1

D)

Visualise the 24 total request numbers in C as ONE figure



E)

<u>Two most interesting observations:</u>

i) <u>Woodpecker, STS-70 and STS-71, Same video name:</u>

The shuttle mission STS-70 was supposed to launch in 2nd week of June, but because some Woodpeckers confused the fuel tank of the shuttle as a tree, they started digging holes in it and damaging the shuttle.[1] This explains the requests for woodpecker.mpg movie.

There was also a launch of STS-71 mission shuttle at the end on June and hence the requests of launch and crew-arrival of STS-71. In addition, on July 13, STS-70 was finally launched and this explains the high number of requests for launch and crew arrival of STS-70 mission.

Also, a look at the filenames of movies requested, it can be seen that same name exists in the most requested and least requested data. However, the difference lies in the type of protocol the request was done and the different method. There was a log where the protocol was not HTTP and the was another log where the method was HEAD instead of GET. This information is useful to NASA because it shows the logs with different kinds of protocol and the method information was requested.

ii) Thursdays highest, July 13 requests spike and no logs after 29th July:

Week	DayOfweek	count	Day
27	1	35272	Sunday
29	1	39199	Sunday
28	1	47854	Sunday
26	1	60265	Sunday
30	2	64259	Monday
28	2	72860	Monday
29	2	74981	Monday
27	2	89584	Monday
30	3	62699	Tuesday
29	3	64282	Tuesday
27	3	70452	Tuesday
28	3	80407	Tuesday
30	4	58849	Wednesday
29	4	72738	Wednesday
28	4	92536	Wednesday
27	4	94575	Wednesday
30	5	61680	Thursday
29	5	66593	Thursday
27	5	100960	Thursday
28	5	134203	Thursday

As mentioned above, on July 13th, STS-70 shuttle was launched and there were a lot of requests to NASA as communication was needed to track/communicate with the shuttle[2]. It can be seen in the above table from Output.txt file (Line 90), where 28th Week of 1995, Day 5(July 13) is a Thursday and has the highest count of requests. This explains the spike of July 13th. Also, as a general trend from the image in part B, it can be seen that the number of requests on weekends is a lot lower than on weekdays.

Checking for records on or after 29th July

host	timestamp	request	HTTP reply code	bytes in the reply	casted_timestamp	DayOfweek	Week

Also, there are no logs after July 29th and it can be verified from the Output.txt file (Line 84) where the data frame is completely empty.

This information is useful to NASA because it can understand the trends of number of requests and prepare its servers to tackle more requests during shuttle take-offs and landings.

Question 2. Movie Recommendation and Analysis

A)

Five-fold cross validation of ALS-based recommendation-

- als_setting1 = ALS (userCol = "userId", itemCol = "movieId", seed = myseed, coldStartStrategy = "drop")
- als_setting2 = ALS (userCol = "userId", itemCol = "movieId", seed = myseed, coldStartStrategy = "drop", rank=20, maxIter=5, regParam=0.2)

ALS setting 1 is the default setting of ALS with rank=10, maxIter=10 and regParam=0.1.

For another setting of ALS, we chose rank=20, max iterations=5 and regularisation parameter=0.2. We chose the setting of rank=20 to check if more number of latent factors when doing matrix factorization for ALS helped the model to predict better than the other setting with rank=10 by comparing the RMSEs on Hot and Cool Users. We reduced the number of iterations to keep the computational time low because a higher rank was chosen. We increased the regParam to make sure there is no overfitting with increased latent factors.

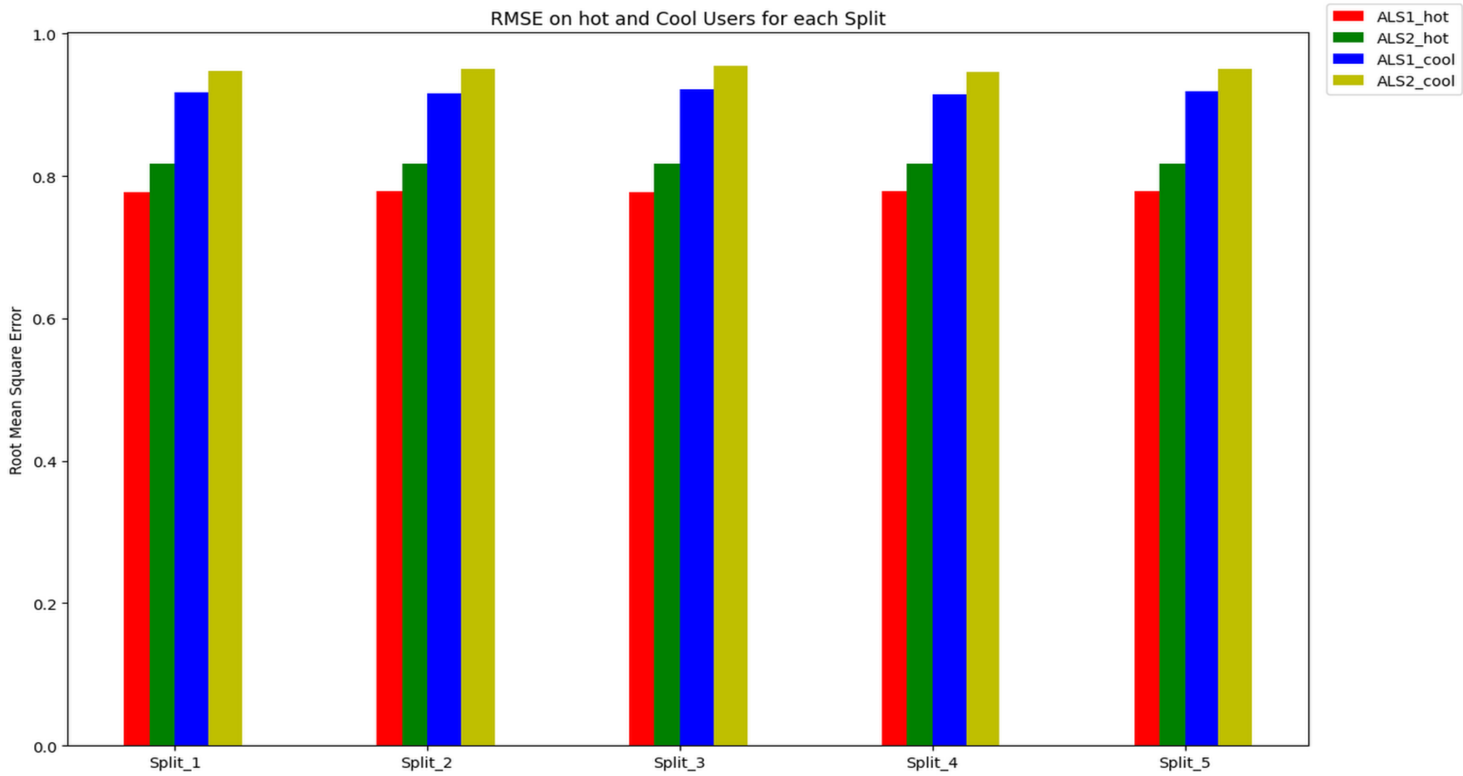
ALS setting	Split_num	HOT Users RMSE	COOL Users RMSE
ALS1	1	0.7779256044502098	0.9172214527198757
ALS1	2	0.7783982001086464	0.916523498585337
ALS1	3	0.777901138850895	0.9216263694961477
ALS1	4	0.7785465960650048	0.9147525242288161
ALS1	5	0.7783283706644275	0.91879850429825
ALS2	1	0.8172770790658604	0.9483272719207912
ALS2	2	0.8176550465588097	0.9501616453666933
ALS2	3	0.817107821666443	0.9541396107623398
ALS2	4	0.8175477514128218	0.9460603750581504
ALS2	5	0.8179094487515246	0.9505319059366907

ALS setting1 Hot Users RMSE- [0.777, 0.778, 0.777, 0.778, 0.778]

ALS setting2 Hot Users RMSE- [0.817, 0.817, 0.817, 0.817, 0.817]

ALS setting1 Cool Users RMSE- [0.917, 0.916, 0.921, 0.914, 0.918]

ALS setting2 Cool Users RMSE- [0.948, 0.950, 0.954, 0.946, 0.950]



B)

K-means with **k=10** to cluster the movie factors:

Top and Bottom Tags from top two clusters:

Cluster	Split_num	Top Tag	Bottom Tag
Largest	1	sci-fi	R:sustained strong stylized violence
2nd largest	1	classic	narrated by character
Largest	2	sci-fi	R:sustained strong stylized violence
2nd largest	2	action	Speculative technologies
Largest	3	sci-fi	R:sustained strong stylized violence
2nd largest	3	animation	layoffs
Largest	4	sci-fi	R:sustained strong stylized violence
2nd largest	4	classic	narrated by character
Largest	5	sci-fi	R:sustained strong stylized violence
2nd largest	5	action	naive characters

Top tag for biggest cluster-[sci-fi, sci-fi, sci-fi, sci-fi, sci-fi]

Top tag for second biggest cluster-[classic, action, animation, classic, action]

Bottom tag for biggest cluster-[R:sustained strong stylized violence, R:sustained strong stylized violence, R:sustained strong stylized violence, R:sustained strong stylized violence, R:sustained strong stylized violence]

Bottom tag for second biggest cluster-[narrated by character, Speculative technologies, layoffs, narrated by character, naive characters]

C)

Two most interesting observations:

i) <u>RMSE for hot and cool Users:</u>

It can be seen from the table in part A above that the RMSE for hot users is a lot lower(nearly 16% less) than the RMSE for cool users.

This is because Hot users are the users who have reviewed the most amount of movies, and hence the recommendation system has a lot of user-movie data to compute additional components and predict on them, whereas for Cool users, we have limited data about the user preferences/factor to train and hence its harder to predict what Cool users might like and how they might rate the other movies.

This information is useful for Netflix because if many users review many products, they have reliable data to learn the user factors and train their recommendation system.

ii) <u>Tags based on clustering</u>:

As evident from the table in B, the movies in biggest clusters have sci-fi as the most common tag across all 5 splits. For the 2nd largest cluster, it varies between action, classic and animation.

These are some of the most popular and highly tagged tags for movies. Also, nearly 55% of movies in cluster 1 have the sci-fi tag, meaning the item factors retrieved after ALS have same kind of characteristics and are closer as a cluster.

The reason for so many movies with sci-fi tag in top cluster is because the dataset used for ALS here is consisting of movies from past(20 years on) and in the past, when there was less scope for CGI(Computer Generated Graphics), the sci-fi movies with their incredible production bagged many accolades because of the stunning visuals and a good storyline. As mentioned in article[3], the average ratings for a sci-fi tag movie was a lot higher in the past than present with a lot more votes for this tag.

iii) RMSE for two ALS setting:

Even though the 2nd ALS setting has higher rank and more latent factors for prediction, the RMSE is worse than 1st ALS setting in all cases.

This might be because 99% of the user-item matrix is already sparse and this is less data to train a model on, thus even if the model tries to learn from more latent factors, the data is not sufficient to learn all the parameters.

This is helpful to Netflix because when they try to train an ALS model, they can use less rank and get a similar result instead of using higher rank and much more computation time.

Sources/References:

[1] https://balettie.com/a-woodpecker-did-what/

[2] http://eecs.csuohio.edu/~sschung/cis612/CIS612_PDF_Presentation_NASA_Halley_Orogvany.pdf

[3] https://dataanalysiscourse.wordpress.com/2018/04/24/the-imdb-analysis-genres-and-ratings-of-movies-released-between-2008-2018/