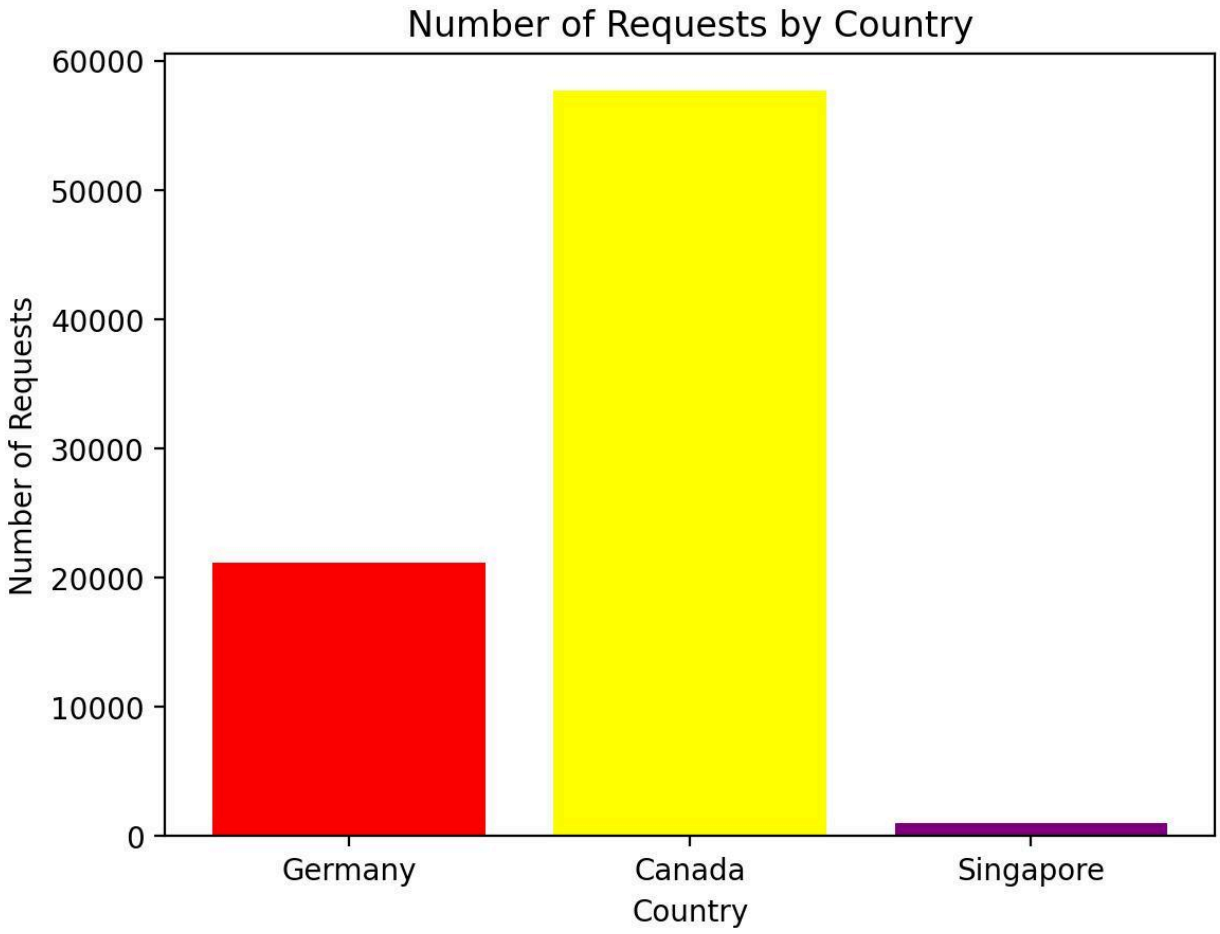
Scalable Machine Learning Assignment

Question 1:

Task A

Total requests from Germany: 21148
Total requests from Canada: 57674
Total requests from Singapore: 1046



Task B

Germany has 1136 unique hosts.
Canada has 2955 unique hosts.
Singapore has 78 unique hosts.

Top 9 Most Active Hosts in Germany:

Host	Count
host62.ascend.interop.eunet.de	825
aibn32.astro.uni-bonn.de	642
ns.scn.de	520
www.rrz.uni-koeln.de	421
ztivax.zfe.siemens.de	385
sun7.lrz-muenchen.de	278
relay.ccs.muc.debis.de	269
dws.urz.uni-magdeburg.de	241
relay.urz.uni-heidelberg.de	232

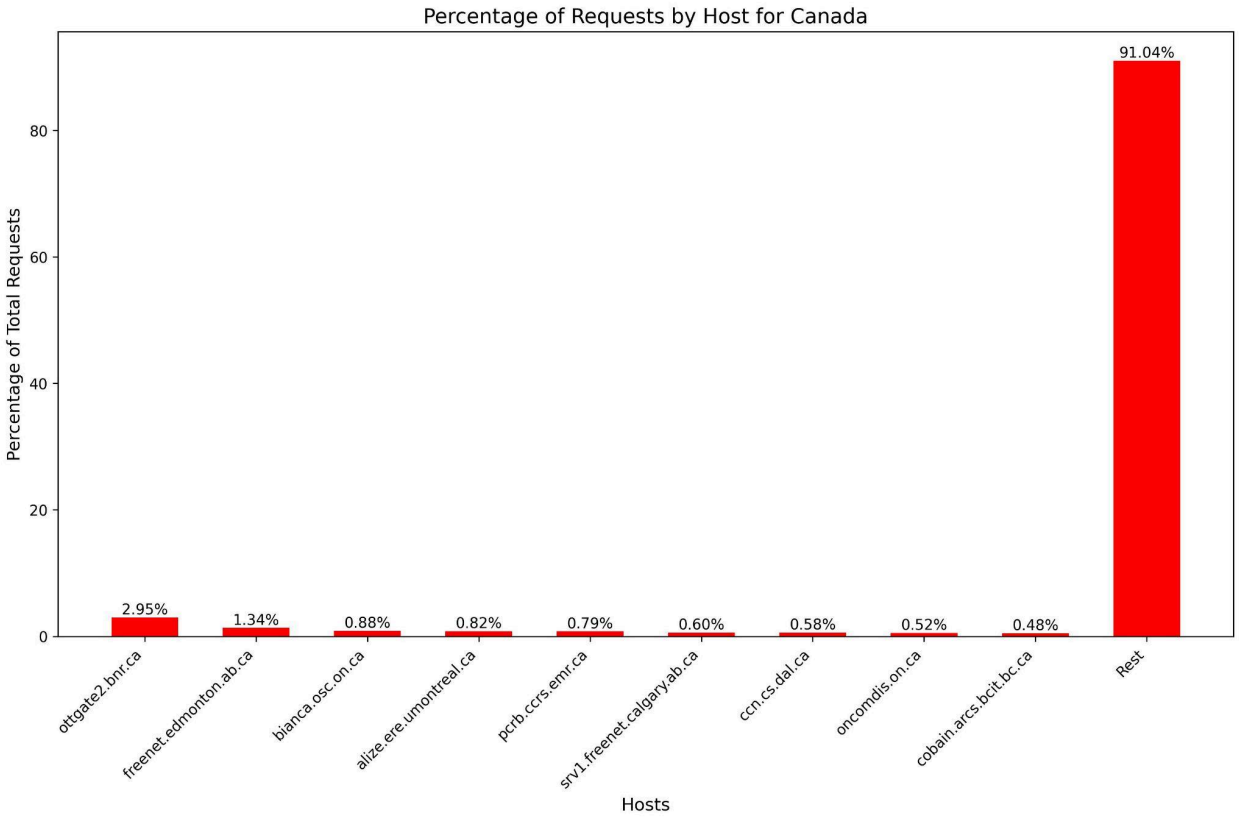
Top 9 Most Active Hosts in Canada:

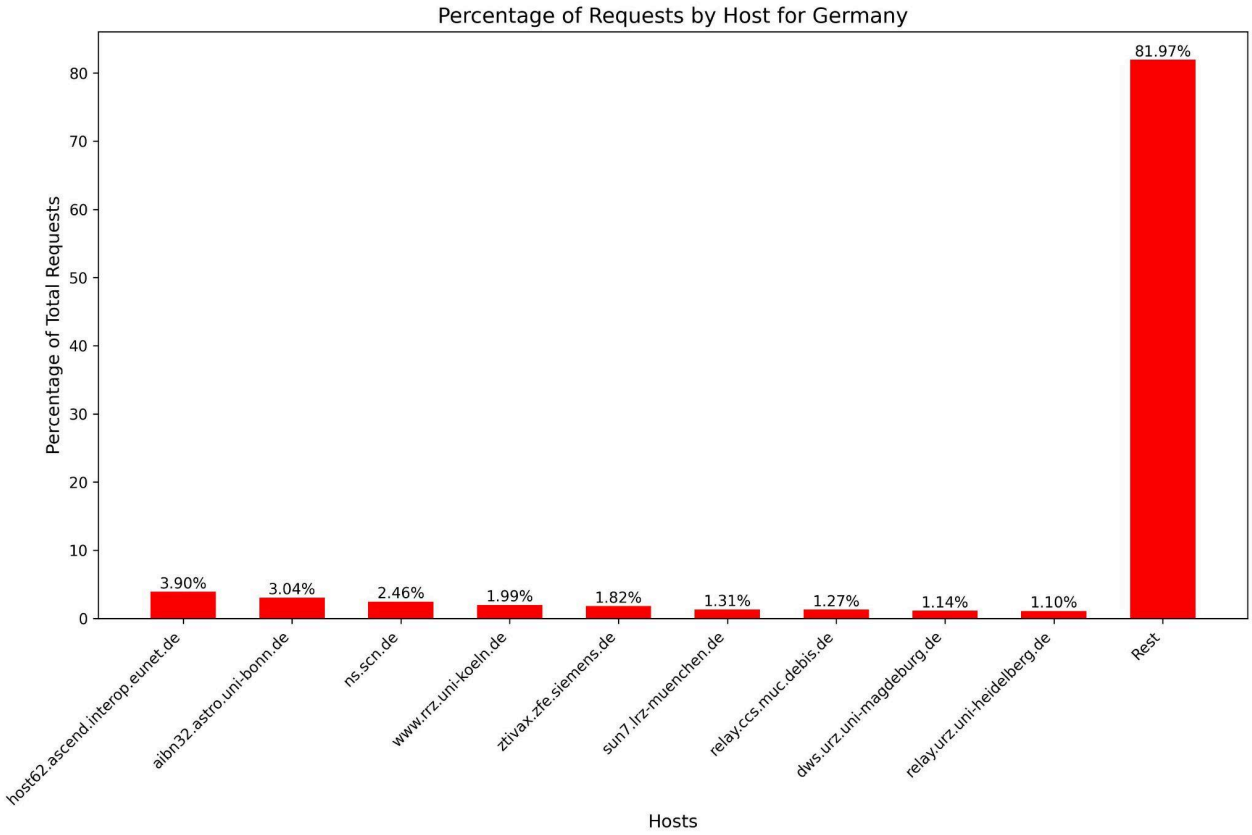
Host	Count
ottgate2.bnr.ca	1704
freenet.edmonton.ab.ca	770
bianca.osc.on.ca	508
alize.ere.umontreal.ca	474
pcrb.ccrs.emr.ca	454
srv1.freenet.calgary.ab.ca	346
ccn.cs.dal.ca	336
oncomdis.on.ca	299
cobain.arcs.bcit.bc.ca	277

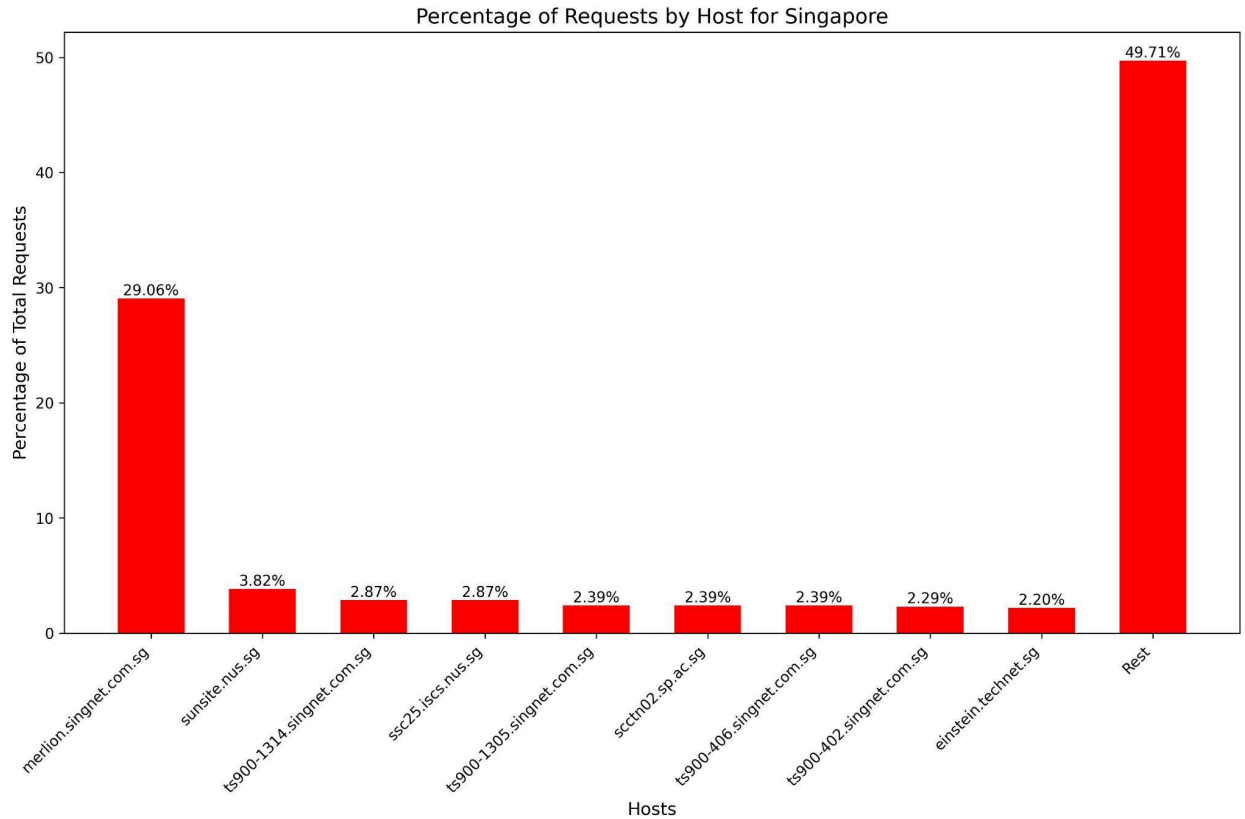
Top 9 Most Active Hosts in Singapore:

Host	Count
merlion.singnet.com.sg	304
sunsite.nus.sg	40
ts900-1314.singnet.com.sg	30
ssc25.iscs.nus.sg	30
ts900-1305.singnet.com.sg	25
scctn02.sp.ac.sg	25
ts900-406.singnet.com.sg	25
ts900-402.singnet.com.sg	24
einstein.technet.sg	23

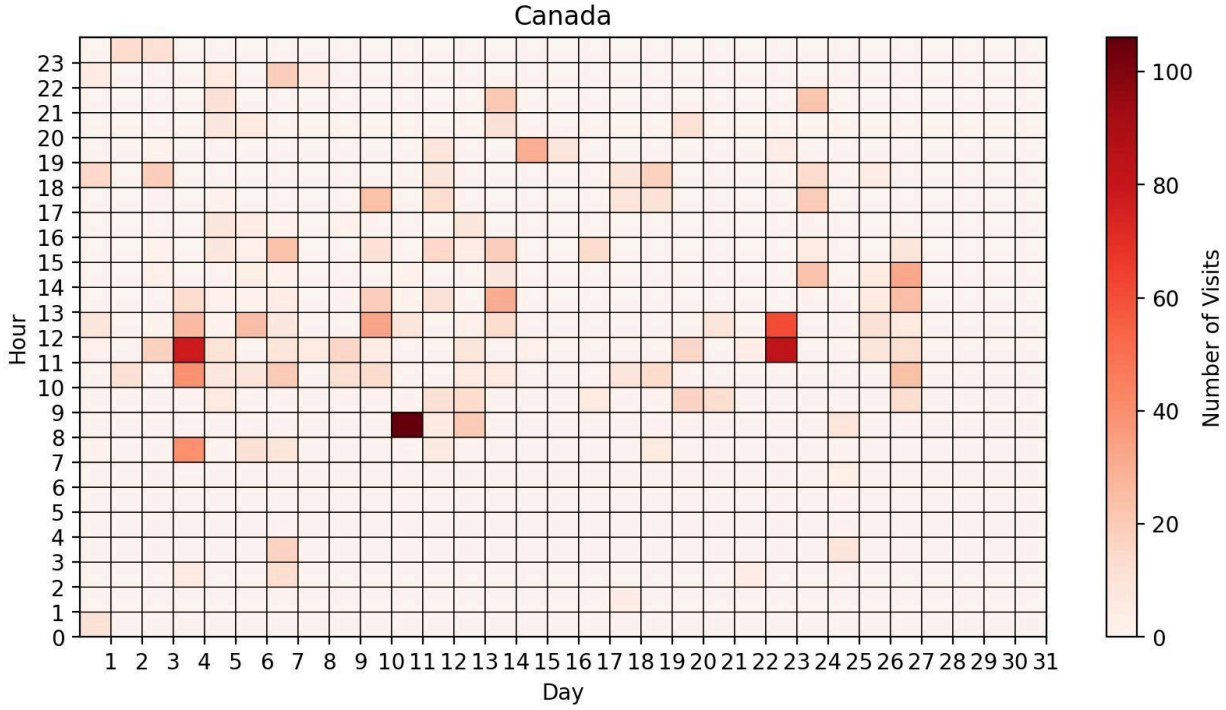
Task C

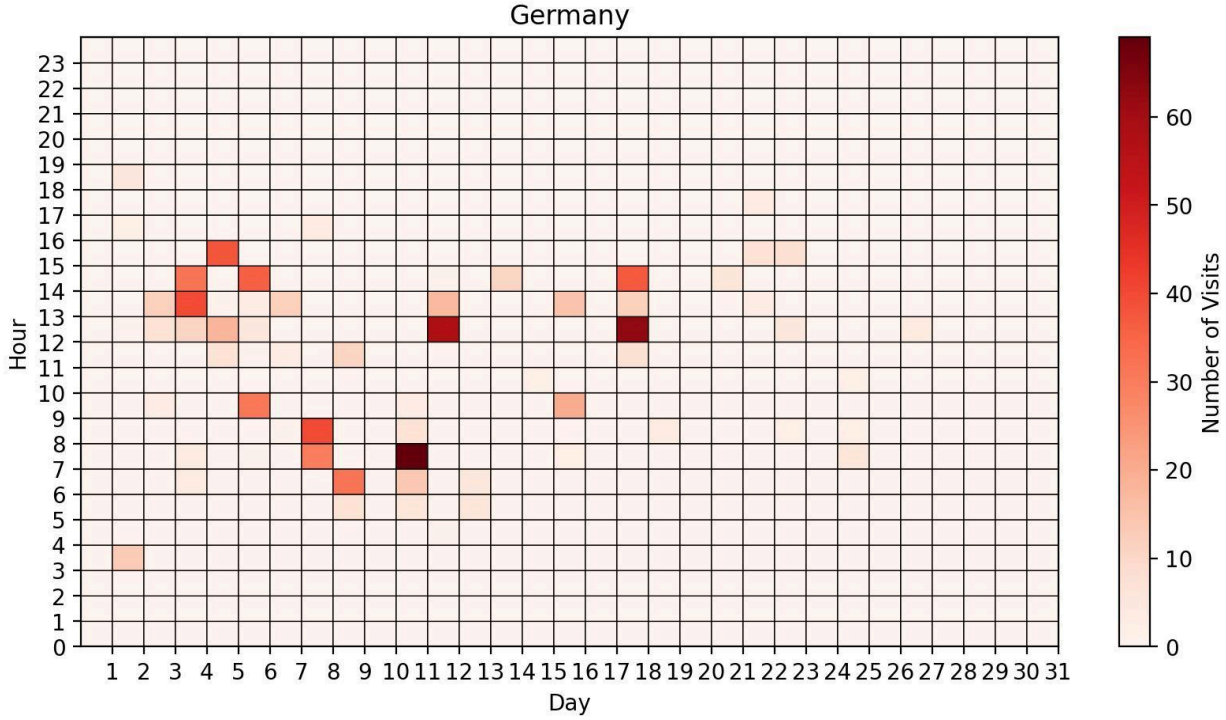


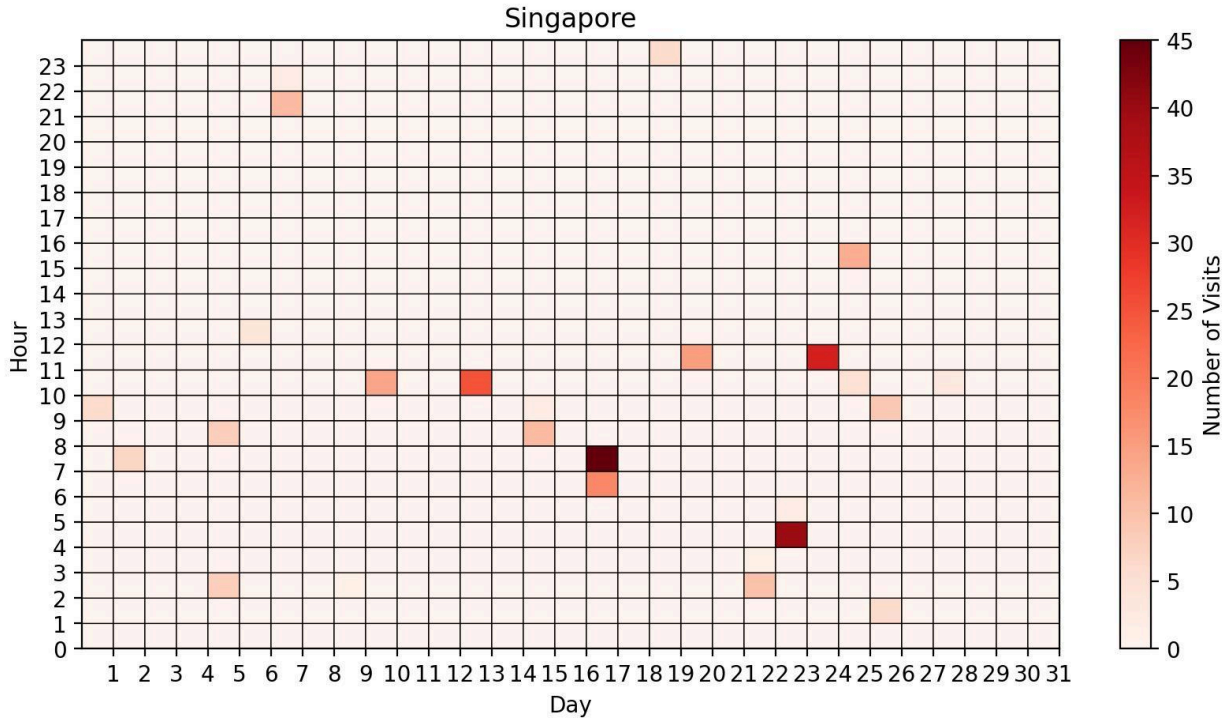

Percentage of Requests by Host for Germany
Percentage of Total Requests
Hosts
0
10
20
30
40
50
60
70
80
3.90%
3.04%
2.46%
1.99%
1.82%
1.31%
1.27%
1.14%
1.10%
81.97%
host62.ascend.interop.eunet.de
aibn32.astro.uni-bonn.de
ns.scn.de
www.rrz.uni-koeln.de
ztivax.zfe.siemens.de
sun7.lrz-muenchen.de
relay.ccs.muc.debis.de
dws.urz.uni-magdeburg.de
relay.urz.uni-heidelberg.de
Rest


Percentage of Requests by Host for Singapore
Percentage of Total Requests
Hosts
0
10
20
30
40
50
29.06%
3.82%
2.87%
2.87%
2.39%
2.39%
2.39%
2.29%
2.20%
49.71%
merlion.singnet.com.sg
sunsite.nus.sg
ts900-1314.singnet.com.sg
ssc25.iscs.nus.sg
ts900-1305.singnet.com.sg
scctn02.sp.ac.sg
ts900-406.singnet.com.sg
ts900-402.singnet.com.sg
einstein.technet.sg
Rest

Task D


Canada
Hour
Day
Number of Visits
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
0 20 40 60 80 100


Germany
Hour
Day
Number of Visits
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
0 10 20 30 40 50 60



Task E

Observation 1 :

The observation reveals that Canada exhibited the highest total number of requests compared to Germany and Singapore, indicating a significantly larger or more engaged Canadian user base. This predominance could be attributed to a higher density of unique hosts in Canada, implying increased interest or reliance on NASA data among Canadian educational and research institutions. Such insights are useful for NASA in optimizing resource allocation and outreach strategies, improving the effectiveness of their support to regions with high engagement, and potentially amplifying the impact of their educational and research initiatives.

Observation 2 :

The analysis of NASA's server logs revealed a significant anomaly with the host 'host62.ascend.interop.eunet.de' in Germany, which received an unusually high 825 requests. This could be attributed to centralized data retrieval, automated system queries, or extensive research requirements. Such insights are critical for NASA because they aid in the optimization of network management, the strategic allocation of resources, and the discovery of potential collaborations with engaged users, all of which improve service delivery and the impact of its data.

Question 2 :

Tasks A and B

Poisson RMSE: 0.35415676683869435

Poisson Model Coefficients:
[-0.03628192878846597,0.02828311733139508,0.07421372254581649,-0.2563918725853634
,0.0054955924317435165,0.10777503386726624,-0.06215817511915619,0.111884554298750
8,0.11627661474433998,-0.02870524910174056,0.0494491947427867,0.08594489464686095
,0.06856431483341396,-0.09233180082024525,-0.0876876185758788,0.0585722899631308 4
6,0.1487573909393707,0.12796662402298217,0.013575555845022498,-0.045056250876449,
0.32869711751173225,-0.023830939708027448,0.09389739347729677,-0.0875429461770725
9,-0.17790636160465934,-0.03595533817495027,-0.15014706590095797,-0.05166691673499
962,-0.11549641908682422,0.04103906308807313,-0.21569402397540569,-0.1133274644720
4942,-0.2402272441229453,0.045938746793003166,-0.0720395749244807,0.3052713259125
2737,0.09095986095721181,0.9684221061581735,0.012521285079916351,-0.0230399055797
16417,0.007717452010103379,0.018868597371091915,-4.2031481540079465e-06]

Logistic Regression Accuracy (L1): 0.8900874824295332
Logistic Regression Accuracy (L2): 0.8900874824295332

Logistic Regression AUC (L1): 0.6274747638691888
Logistic Regression AUC (L2): 0.6282882382517361

Logistic Regression Coefficients (L1):
(43,[3,6,16,20,24,35,37,39,40,41],[-0.1358448969756313,-0.04637381633392563,0.005570317
402758971,0.1981608202926317,-0.043166054655165854,0.3609538388220276,0.91292641
55830042,-0.0122601191589404 6,0.0017765954804082502,0.013452227295044047])

Logistic Regression Coefficients (L2):
[-0.0554595613354867,-0.009663758814880418,0.06334201480375361,-0.319434673617940
53,-0.032196424008271066,0.1112546927563856,-0.06369542261989232,0.10904610022764
377,0.14488442324137263,0.024336301950245384,0.09207656012208712,0.1824677421142
1718,0.0934949555445522,-0.06370130891244741,-0.07397614327400868,0.0493484109470
85405,0.12297547154671339,0.0659131605895466,-0.04229662464166082,-0.133123548817
62745,0.3082085046375766,-0.10225088932504944,-0.018110561981605714,-0.2658827901
9451194,-0.36025282658339464,-0.11290057328830237,-0.318305763399956,-0.1863673279
2368083,-0.19460415102363476,0.0645271758172711,-0.46942433490188296,-0.224260717
36990782,-0.6727798771338824,-0.02658942666040341,-0.3898591450462567,0.767883403
0283765,0.07794064124423677,1.00255592193221,0.00993318508077163,-0.023216185779
02623,0.006236928291606973,0.01762502876027256,-4.573989832039292e-06]

Task C

1. Effectiveness of Regularisation Techniques:

Both L1 and L2 regularization methods are effective in managing overfitting, as evidenced by similar AUC values (L1 at 0.627, L2 at 0.628). Despite L1's role in feature selection, which involves driving many coefficients to zero and resulting in a simpler model, L2 slightly outperforms L1. This suggests that L2's ability to retain all features while reducing coefficients aids in the capture of subtle but valuable interactions that L1 may overlook due to its aggressive reduction.

2. Impact on Model Interpretability and Complexity:

L1 regularization significantly enhances model interpretability by inducing sparsity in the coefficients, effectively selecting features that are most impactful in predicting claims. This contrasts with L2 regularization, which does not zero out coefficients but reduces their magnitude uniformly, thus preserving the complexity and ensuring that all features contribute to the model, albeit with minimized influence. This approach is particularly advantageous when the dataset features complex interactions that are crucial for accurate predictions.

3. Analysis of the Coefficients:

The Poisson regression model, with an RMSE of 0.354, provides moderate predictive accuracy for claim count, indicating effective modeling of the data. The coefficient analysis reveals that different features have varying degrees of influence, with some having significant magnitudes, indicating their importance in predicting claim frequency. This model's performance demonstrates its usefulness in understanding and predicting claim frequency based on multiple risk factors.

Question 3 :

==========**Task A**==========

==========Random Forest==========
RF accuracy = 0.705463
RF area under the curve = 0.69357
==========Gradient Boosting==========
GBT accuracy = 0.722095
GBT area under the curve = 0.714565
==========Neural Network==========
Neural Network accuracy = 0.682344
Neural Network area under the curve = 0.671344

==========**Task B**==========

==========Best parameter for RF==========
[{'maxDepth': 10}, {'maxBins': 50}, {'numTrees': 10}]
==========Best parameter for GBT==========
[{'maxDepth': 5}, {'maxIter': 30}, {'stepSize': 0.2}]
==========Best parameter for NN==========
[{'maxIter': 150}, {'blockSize': 256}, {'layers': [28, 6, 5, 2]}]

Random Forest accuracy = 0.697879
Random Forest area under the curve = 0.690826
==========
Gradient Boosted Trees accuracy = 0.718963
Gradient Boosted Trees area under the curve = 0.71458
==========
Neural Network accuracy = 0.675011
Neural Network area under the curve = 0.668837
==========

Comparison of the performance of three algorithms

Gradient Boosting consistently outperforms the other two algorithms in terms of accuracy and AUC across both subsets and the entire dataset. This suggests that the model is quite robust, effectively handling both variance and bias, most likely due to iterative error correction in the weak learner sequence (trees).

Random Forest performs reasonably well, but slightly lower than Gradient Boosting. The ensemble nature of Random Forest, which averages multiple deep decision trees, allows for good generalization. However, it appears to be slightly less effective than Gradient Boosting in this particular dataset.

Neural networks perform the worst of the three in this task. While typically powerful for large-scale and complex pattern recognition tasks, the shallow network architecture used here may be insufficient to capture the data's complex relationships. To train effectively, neural networks may require more careful parameter tuning and potentially more data than ensemble methods.

Question 4 :

Task A

1)

Sorted data by the timestamp : (only showing the top 40 rows)

```
+------+-------+------+---------+--------------------+
|userId|movieId|rating|timestamp|percent_rank        |
+------+-------+------+---------+--------------------+
|28507 |1176   |4.0   |789652004|0.0                 |
|131160|21     |3.0   |789652009|4.999934500858039E-8 |
|131160|47     |5.0   |789652009|4.999934500858039E-8 |
|131160|1079   |3.0   |789652009|4.999934500858039E-8 |
|20821 |32     |5.0   |822873600|1.9999738003432155E-7|
|53434 |19     |1.0   |822873600|1.9999738003432155E-7|
|85252 |2      |4.0   |822873600|1.9999738003432155E-7|
|85252 |7      |5.0   |822873600|1.9999738003432155E-7|
|85252 |10     |5.0   |822873600|1.9999738003432155E-7|
|85252 |11     |5.0   |822873600|1.9999738003432155E-7|
|85252 |12     |1.0   |822873600|1.9999738003432155E-7|
|85252 |17     |5.0   |822873600|1.9999738003432155E-7|
|85252 |19     |3.0   |822873600|1.9999738003432155E-7|
|85252 |21     |4.0   |822873600|1.9999738003432155E-7|
|85252 |22     |4.0   |822873600|1.9999738003432155E-7|
|85252 |24     |3.0   |822873600|1.9999738003432155E-7|
|85252 |32     |4.0   |822873600|1.9999738003432155E-7|
|85252 |34     |5.0   |822873600|1.9999738003432155E-7|
|85252 |36     |5.0   |822873600|1.9999738003432155E-7|
|85252 |45     |3.0   |822873600|1.9999738003432155E-7|
|85252 |48     |4.0   |822873600|1.9999738003432155E-7|
|85252 |50     |5.0   |822873600|1.9999738003432155E-7|
|85252 |60     |4.0   |822873600|1.9999738003432155E-7|
|85252 |70     |4.0   |822873600|1.9999738003432155E-7|
|99851 |1      |4.0   |822873600|1.9999738003432155E-7|
|99851 |10     |4.0   |822873600|1.9999738003432155E-7|
|99851 |18     |4.0   |822873600|1.9999738003432155E-7|
|99851 |19     |4.0   |822873600|1.9999738003432155E-7|
|99851 |21     |5.0   |822873600|1.9999738003432155E-7|
|99851 |31     |5.0   |822873600|1.9999738003432155E-7|
|99851 |32     |5.0   |822873600|1.9999738003432155E-7|
|99851 |39     |5.0   |822873600|1.9999738003432155E-7|
|99851 |45     |4.0   |822873600|1.9999738003432155E-7|
|99851 |47     |5.0   |822873600|1.9999738003432155E-7|
|99851 |50     |5.0   |822873600|1.9999738003432155E-7|
|99851 |52     |4.0   |822873600|1.9999738003432155E-7|
```

```
|99851 |55     |4.0   |822873600|1.9999738003432155E-7|
|99851 |58     |5.0   |822873600|1.9999738003432155E-7|
|108467|10     |3.0   |822873600|1.9999738003432155E-7|
|108467|11     |4.0   |822873600|1.9999738003432155E-7|
+------+-------+------+---------+---------------------+
```

2) Justification for adjusting ALS settings:

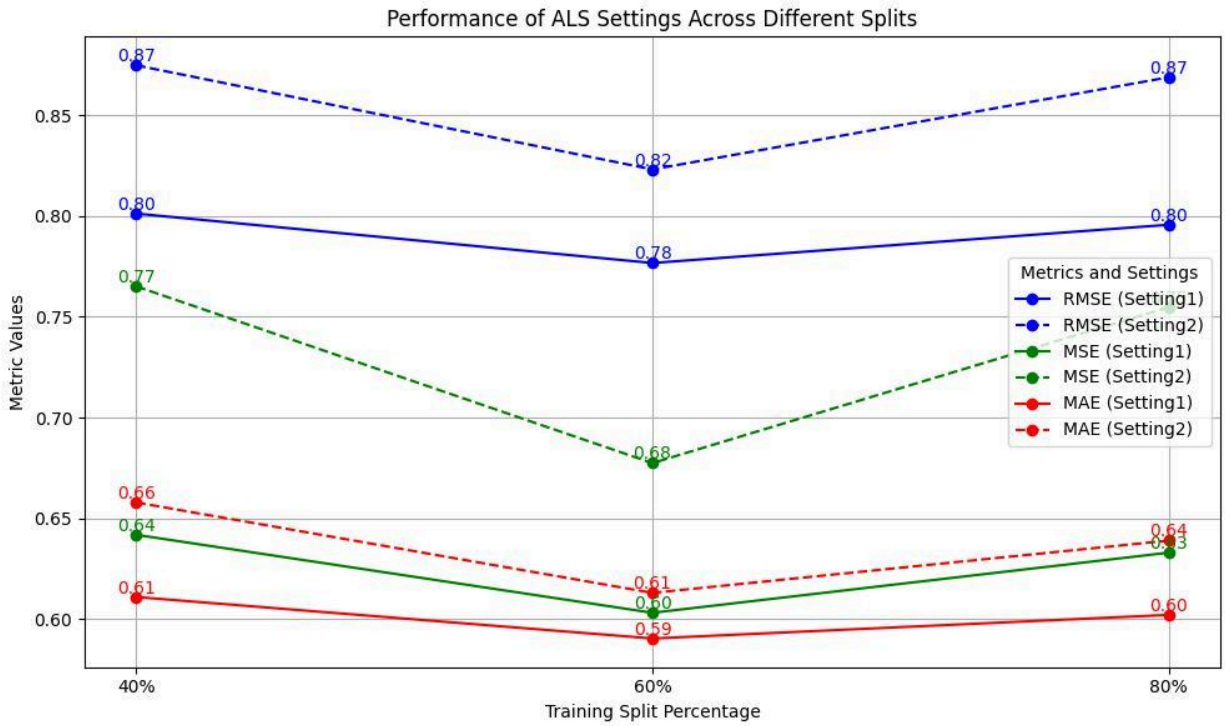
The ALS setting 2 was chosen, which increased the rank above the default, added regularization ('regParam'), and increased iterations ('maxIter'). The rationale for this change was based on the potential underfitting seen in setting 1.

- Increasing 'rank': A higher rank indicates that the model incorporates more factors (features) to represent each user and item. This can potentially capture more complex patterns in the data but also risks overfitting if not managed with other parameters like regularization. It improves the model's ability to capture more nuanced user and item interactions, potentially lowering prediction error.
- Increasing 'maxIter': More iterations give the ALS algorithm more chances to converge on a solution, which may be required with a higher rank to fully realize the potential of the increased complexity.
- Lowering 'regParam': Lowering the regularization parameter reduces the penalty for larger model coefficients, allowing the model to better fit the training data. Introducing a regularization parameter ('regParam') aids in controlling overfitting, ensuring that the model generalizes better on previously unseen data.

3)

ALS Metrics Table Overview:

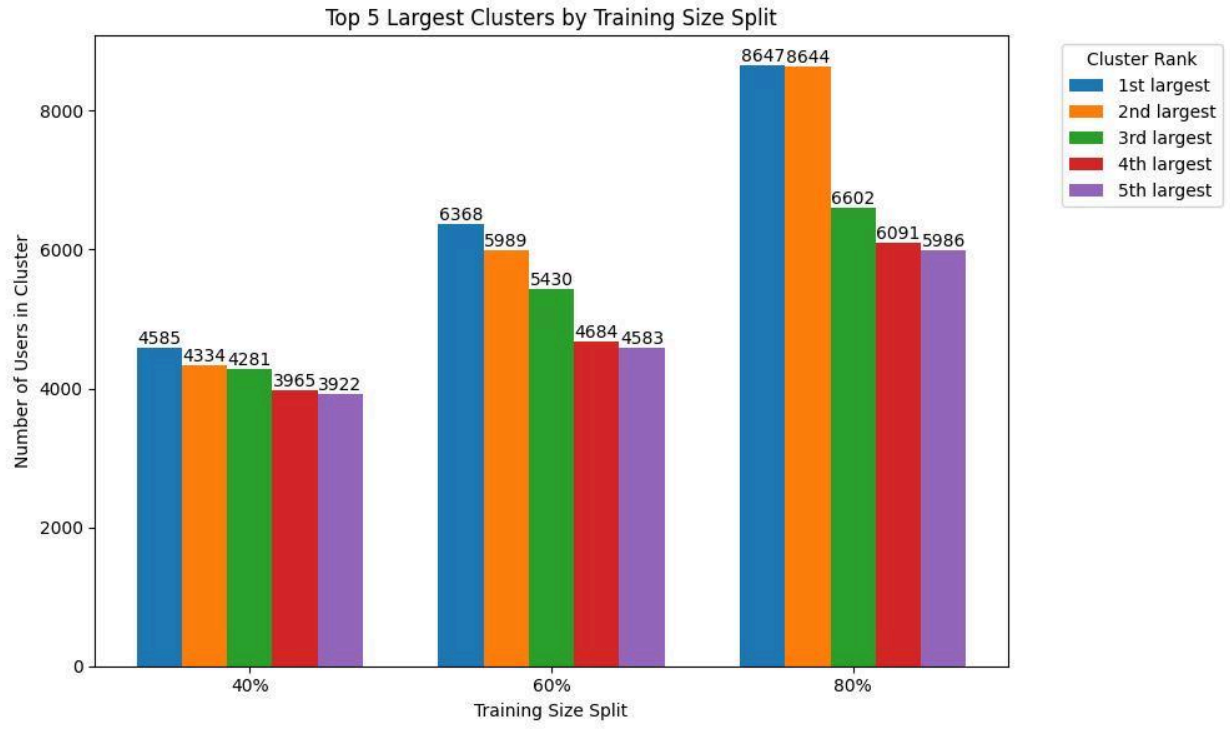
Split	Setting	RMSE	MSE	MAE
40%	Setting 1	0.801164	0.641864	0.611009
60%	Setting 1	0.776657	0.603197	0.590446
80%	Setting 1	0.795625	0.63302	0.602151
40%	Setting 2	0.87469	0.765083	0.657908
60%	Setting 2	0.823059	0.677426	0.613059
80%	Setting 2	0.868764	0.754751	0.639004


Performance of ALS Settings Across Different Splits
Metric Values
Training Split Percentage
40%
60%
80%
0.85
0.80
0.75
0.70
0.65
0.60
0.87
0.82
0.87
0.80
0.78
0.80
0.77
0.68
0.66
0.64
0.61
0.60
0.64
0.61
0.59
0.60
Metrics and Settings
RMSE (Setting1)
RMSE (Setting2)
MSE (Setting1)
MSE (Setting2)
MAE (Setting1)
MAE (Setting2)

Task B

1)
Cluster Sizes Table Overview:

	40%	**60%**	**80%**
1st Largest	4585	6368	8647
2nd Largest	4334	5989	8644
3rd Largest	4281	5430	6602
4th Largest	3965	4684	6091
5th Largest	3922	4583	5986



2) Top ten most popular genres for each split

	Split	Genre	Count
1	40% Split	Drama	2363
2	40% Split	Comedy	1677
3	40% Split	Romance	836
4	40% Split	Thriller	807
5	40% Split	Action	659
6	40% Split	Crime	509
7	40% Split	Adventure	489
8	40% Split	Horror	403
9	40% Split	Sci-Fi	365
10	40% Split	Children	285
11	60% Split	Drama	3383
12	60% Split	Comedy	2314
13	60% Split	Romance	1218
14	60% Split	Thriller	1130
15	60% Split	Action	891
16	60% Split	Crime	744
17	60% Split	Adventure	685
18	60% Split	Sci-Fi	460
19	60% Split	Horror	452
20	60% Split	Children	375
21	80% Split	Drama	5206
22	80% Split	Comedy	3452
23	80% Split	Romance	1758

24	80% Split	Thriller	1664
25	80% Split	Action	1366
26	80% Split	Crime	1173
27	80% Split	Adventure	1010
28	80% Split	Horror	880
29	80% Split	Sci-Fi	736
30	80% Split	Fantasy	564

Task C

Observation 1: Improved ALS Model Performance with Larger Dataset Size

Task A analysis shows that the ALS model's Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics decrease as the training data size increases from 40% to 80%. This suggests that larger datasets lead to higher predictive accuracy. This improvement is most likely due to the increased richness of user-item interaction data available in larger datasets, which allows for more effective learning and generalization within the ALS model while capturing a broader range of user preferences and item characteristics. For a platform like Netflix, this finding highlights the importance of large amounts of historical data in optimizing the performance of recommendation systems. By utilizing more extensive data, Netflix can significantly improve its predictive accuracy, resulting in increased user engagement and retention via more tailored and appealing content recommendations.

Observation 2: Growth in User Cluster Size Correlates with Increased Data Volume

Task B's clustering results show that the size of the largest user cluster grows with the dataset size, from 4,585 to 8,647 users as training data expands from 40% to 80%. As more data becomes available, the algorithm's ability to aggregate user preferences more effectively improves, allowing it to more accurately group users with similar viewing patterns. This increase could be attributed to a lower noise-to-signal ratio, allowing for the clustering algorithm to detect and capture clearer patterns of preferences. This observation is especially useful for Netflix in improving audience segmentation and content personalization strategies. By accurately identifying and understanding larger, more homogeneously grouped segments of its user base, Netflix can more effectively deploy targeted content recommendations and marketing initiatives, improving user experience and engagement. These insights also help with strategic content development and acquisition, allowing offerings to better align with proven user preferences.