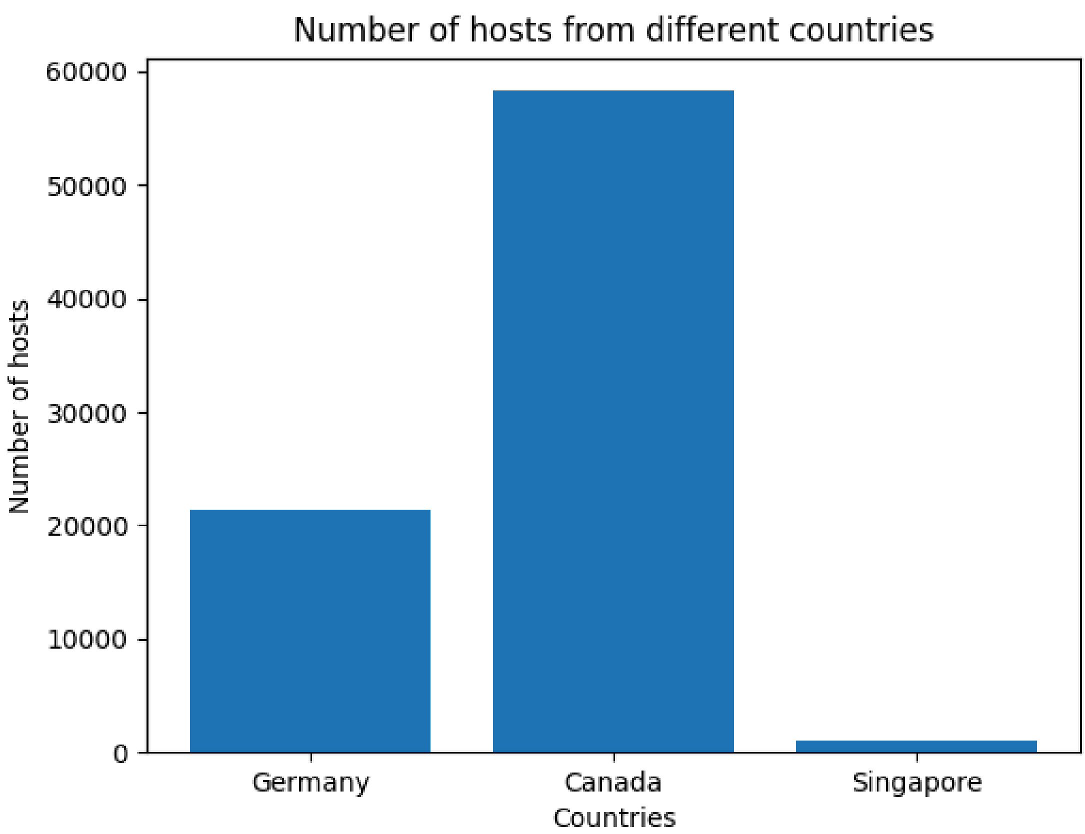
Question 1

A. Country-wise, the total number of requests

1. Germany - 21345
2. Canada - 58290
3. Singapore - 1138



B. Country-wise unique hosts and top 9 most frequent hosts

Germany - There are **1138** unique hosts and presented below is a table of 9 most frequent hosts along with their count

Host	**Count**

host62.ascend.interop.eunet.de	832
aibn32.astro.uni-bonn.de	642
ns.scn.de	523
www.rrz.uni-koeln.de	423
ztivax.zfe.siemens.de	387
sun7.lrz-muenchen.de	280
relay.ccs.muc.debis.de	275
dws.urz.uni-magdeburg.de	244
relay.urz.uni-heidelberg.de	239

Canada - There are **2970** unique hosts and presented below is a table of 9 most frequent hosts along with their count

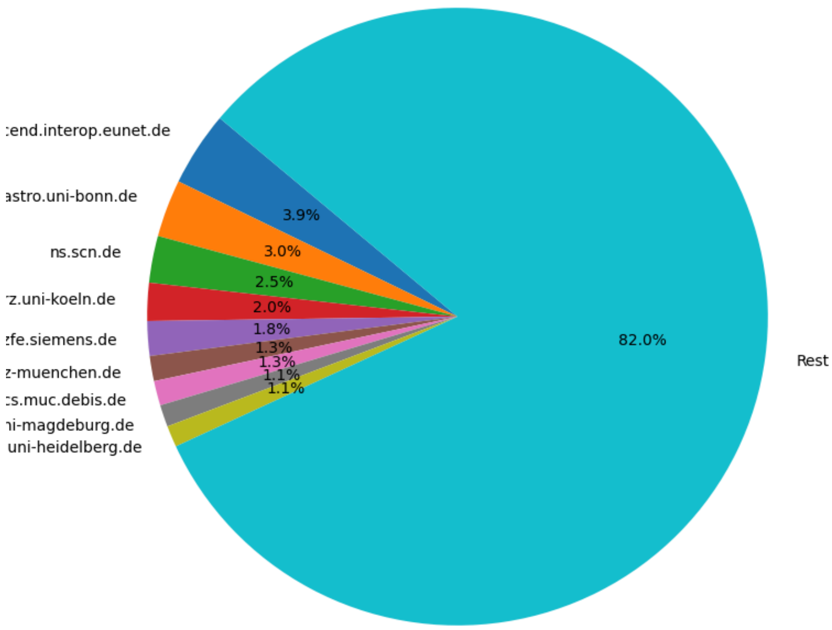
Host	**Count**
ottgate2.bnr.ca	1718
freenet.edmonton.ab.ca	782
bianca.osc.on.ca	511
alize.ere.umontreal.ca	479
pcrb.ccrs.emr.ca	461
srv1.freenet.calgary.ab.ca	362
ccn.cs.dal.ca	351
oncomdis.on.ca	304
cobain.arcs.bcit.bc.ca	289

Singapore - There are **78** unique hosts and presented below is a table of 9 most frequent hosts along with their count

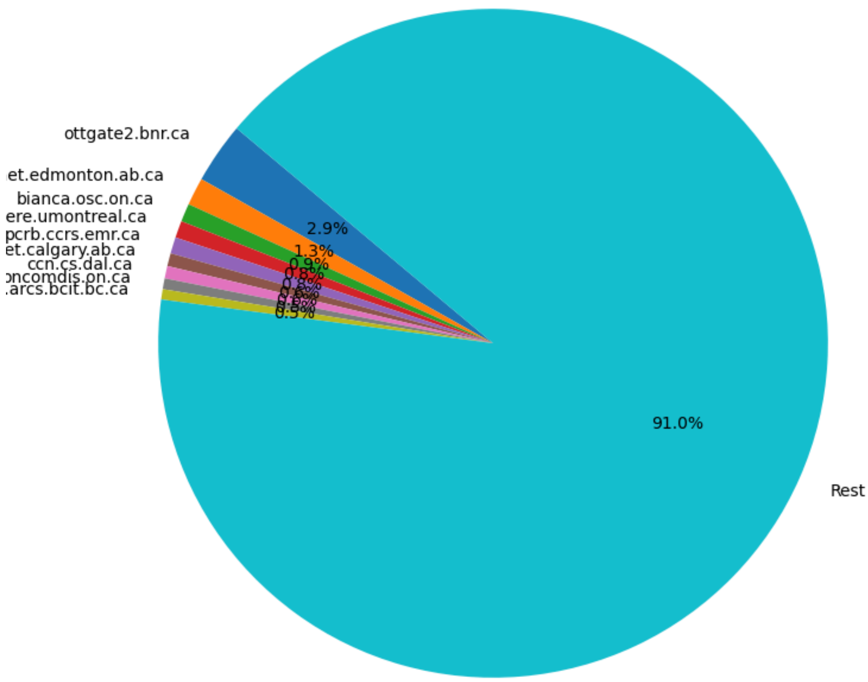
Hosts	**Count**
merlion.singnet.com.sg	308
sunsite.nus.sg	40
ts900-1314.singnet.com.sg	30
ssc25.iscs.nus.sg	30
scctn02.sp.ac.sg	25
ts900-1305.singnet.com.sg	25
ts900-406.singnet.com.sg	25
ts900-402.singnet.com.sg	24
einstein.technet.sg	23

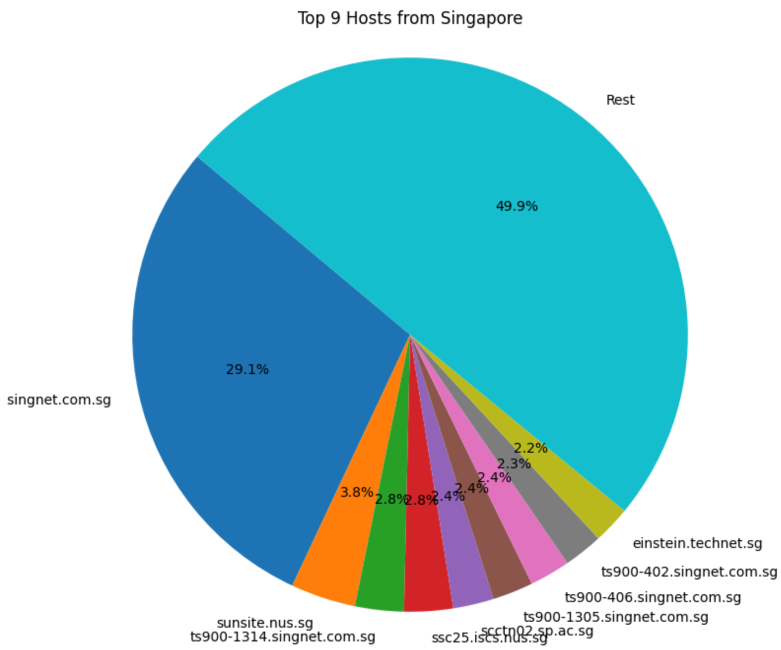
C. Pie charts

Top 9 Hosts from Germany

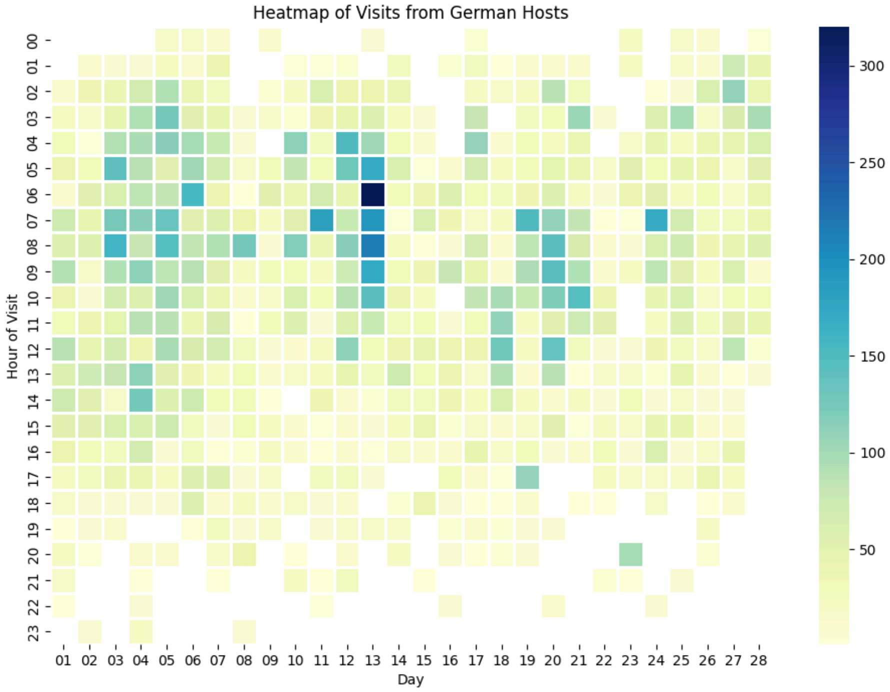


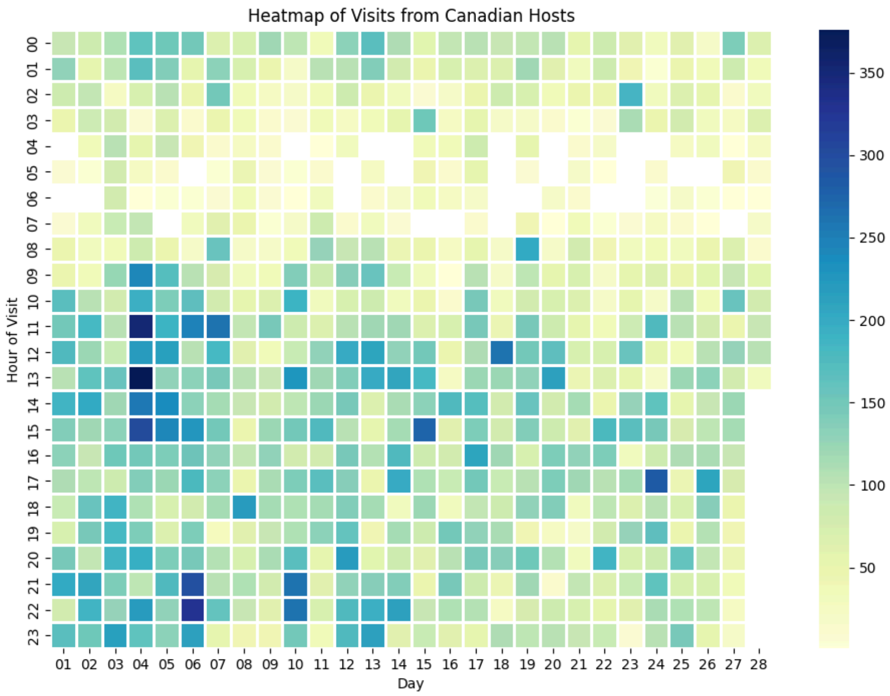
Top 9 Hosts from Canada

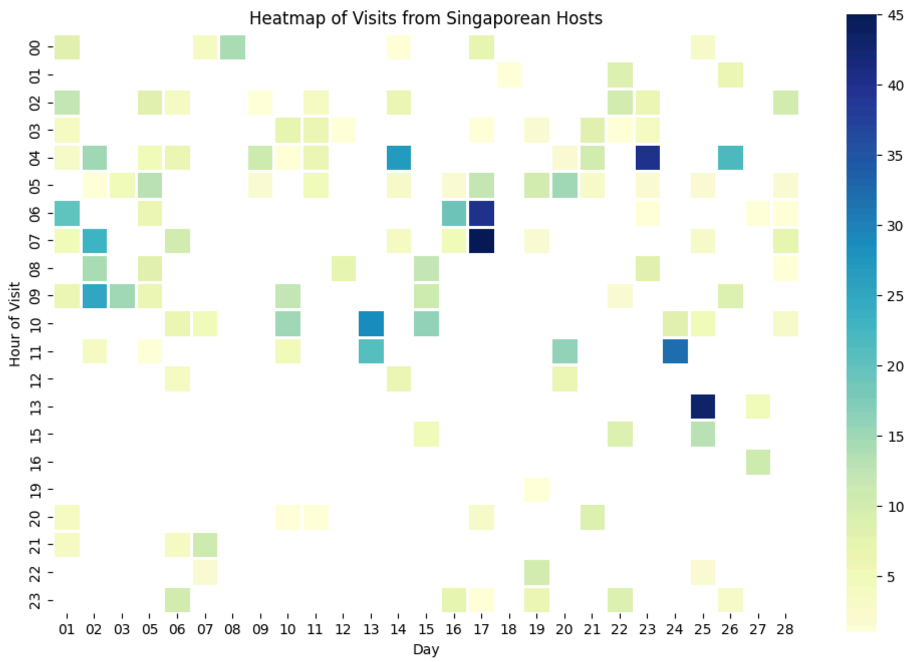



Top 9 Hosts from Singapore
Rest
49.9%
29.1%
singnet.com.sg
3.8%
2.8%
2.8%
2.4%
2.4%
2.4%
2.3%
2.2%
einstein.technet.sg
ts900-402.singnet.com.sg
ts900-406.singnet.com.sg
ts900-1305.singnet.com.sg
scctn02.sp.ac.sg
ssc25.iscs.nus.sg
sunsite.nus.sg
ts900-1314.singnet.com.sg

D. Heatmaps


Heatmap of Visits from German Hosts
Hour of Visit
Day
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
50
100
150
200
250
300


Heatmap of Visits from Canadian Hosts
Hour of Visit
Day
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
50
100
150
200
250
300
350



E. Interesting Observations

Observation 1 - From the pie-charts, we could identify that the top 9 hosts take up 50% of internet usage, while in Canada and Germany, the websites apart from the top 9 take an average of 85%. This observation along with the low number of hosts in Singapore, is attributed to the small country and how the websites are being used on the herd psychology of people. This inference could help NASA strategize communications by understanding the dynamic internet usage of densely populated countries.

Observation 2 - From the heatmap, it can be observed that most users access the internet in the late evening while in Canada, nighttime has the maximum activity across the country. This is because Germany's cultural norms have stipulated leisure time in the evenings while in Canada, socializing and communication are considered to happen at night. NASA can utilize this data to understand the distribution of internet activity and optimize strategies with international organizations.

Question 2

Poisson Regression Model Coefficients:

0.18486250561535117	0.01620602380458771	-0.04847976297793088

0.04208808430272111	0.1660157316977079	0.03251768750794173
-0.0008307444436257737	-0.007892906949462661	0.004140726529012344
0.002943533560528968 6	0.01498854497792491	-0.002512925841047869
-0.012568385448310066	0.0030212544128556448	-0.0015353121246291803
5.889993973715241e-05	0.00018731538702452708	0.03678083151245669
-0.03678083151247796	0.03113325852592524	0.013616866198341938
-0.0016669039215393513	-0.006756615447464688	0.00653215498623501
-0.00044190249839663346	-0.011289040554668084	-0.01143601785593605
0.002992600805862749	-0.00165688318336943	0.002240687768967186
-0.005495277757437278	0.0006888008512335568	-0.008666273764255848
-0.004301752030274925	0.0008200809416113545	-0.0028203667331403873
0.000530647786225848	-0.001611124892438729	0.0004964281813662522
-0.0010230791444130792	-0.0018862882624572667	0.011089385697953192
0.015298773086920322	-0.012343992219000208	-0.024094178827083684
0.004595233556476412	0.005454778704714977	

Logistic Regression Model Coefficients (L1 Regularization):
(47,[],[])

Logistic Regression Model Coefficients (L2 Regularization):

9.676968844328955e-05	2.522646859437878e-06	-1.3426252769702459e-05
9.54564384679354e-06	7.533071165744898e-05	2.0265944613121833e-05
-2.0584357186698477e-05	-1.7009696855798288e-05	1.8937130818127515e-05
1.6415087959727604e-05	0.00010760811694072123	-7.1574276690823915e-06
-9.191406737882047e-05	3.1163831758028304e-05	-1.2513121585117329e-05

-6.209747078667124e-07	1.8989160923731674e-05	6.623218131385244e-05
-6.62321813138531e-05	6.997061033853359e-05	3.72387702002309e-05
-2.6892261920623586e-05	-9.18492839868712e-06	3.70224455924176e-05
-1.1196047308778146e-05	-8.467294562589572e-05	-9.368107958619527e-05
9.83542676330284e-06	-5.607301390094001e-06	1.000049714403029e-05
-6.888229704201568e-05	3.019384033939416e-05	-0.00017177662330313537
-0.00011588417026473134	2.813905632327557e-05	-0.00011529216700051261
2.359126840982982e-05	-7.965896777847834e-05	2.668696591173679e-05
-7.175285874495103e-05	-0.00028987742278546106	7.575083377322315e-06
2.6581879429177408e-05	-1.9157161678816873e-05	-6.262870756973575e-05
2.0394443066553567e-05	0.00011965980382544563	

RMSE for Poisson Regression: 0.244698147475363
AUC for Logistic Regression (L1 Regularization): 0.5
AUC for Logistic Regression (L2 Regularization): 0.6036035379905605
Accuracy for Logistic Regression (L1 Regularization): 0.9501491892740764
Accuracy for Logistic Regression (L2 Regularization): 0.9501491892740764

Observation - 1

It can be observed that L1 is performing shrinking of coefficients for feature selection from the sparse coefficients presented. On the other hand, dense coefficients are yielded from L2 regularization indicating spread-out weight distribution across features compared to L1 regularization. Although a lot of coefficients are close to zero, they represent a penalized version of the weights to prevent overfitting.

Observation - 2

It can be noted from the performance metrics that there’s a difference in the AUC values (L2 is higher than L1) but similar values for accuracies for L1 and L2 regularization. This indicates that although L2 regularization does not increase the overall classification accuracy, it provides a distinct separation between the classes. It can be concluded that the choice of regularization would depend on the computational efficiency and density of the coefficients required.

Question 3

Parameter Grid

Model	Hyperparameter	Value 1	Value 2	Value 3
Multilayer Perceptron Classifier	Layers	[[len(feature_cols), 10, 2]	[len(feature_cols), 5	[len(feature_cols), 15, 2]
	Stepsize	0.1	0.05	0.2
	maxIter	10	20	30
Random Forest Classifier	numTrees	25	50	75
	maxDepth	3	5	7
	maxBins	32	64	16
Gradient Boosting Classifier	maxDepth	3	5	7
	stepSize	0.1	0.2	0.3
	maxIter	10	20	30

Tuned-hyperparameter values

Model	Hyperparameter	Best Value
Multilayer Perceptron Classifier	Layers	[28, 5, 2]
	Stepsize	0.1
	maxIter	30
Gradient Boosting	maxDepth	5

Classifier	stepSize	0.3
	maxIter	30
Random Forest Classifier	numTrees	75
	maxDepth	7
	maxBins	64

Random Forest - AUC on the full dataset: 0.6879225995633658
Gradient Boosting - AUC on the full dataset: 0.7227098069946143
Multilayer Perceptron - AUC on the full dataset: 0.6225901027435201

Question 4

ALS	Step Size	RMSE	MSE	MAE
Setting 1	40	0.8065054132580284	0.6504509816145031	0.6218303862035098
	60	0.7779441050917113	0.6051970306469435	0.5924562907202132
	80	0.7976814569738635	0.6362957067999456	0.605544408179643
Setting 2	40	0.8111544972700697	0.6579716184414596	0.6239040899911104
	60	0.7774377256294837	0.6044094172319443	0.5911982329836366
	80	0.8035529189864592	0.6456972936116591	0.6084076669557363

Top 5 Clusters

40% split		60% split		80% split	
Cluster No	Count	Cluster No.	Count	Cluster No.	count
7	5059	15	6394	11	7118
21	4114	9	5106	18	6146
16	3704	7	4936	2	5880
23	3671	3	4314	1	5561
24	3504	22	4092	8	5530

Top 5 Movies

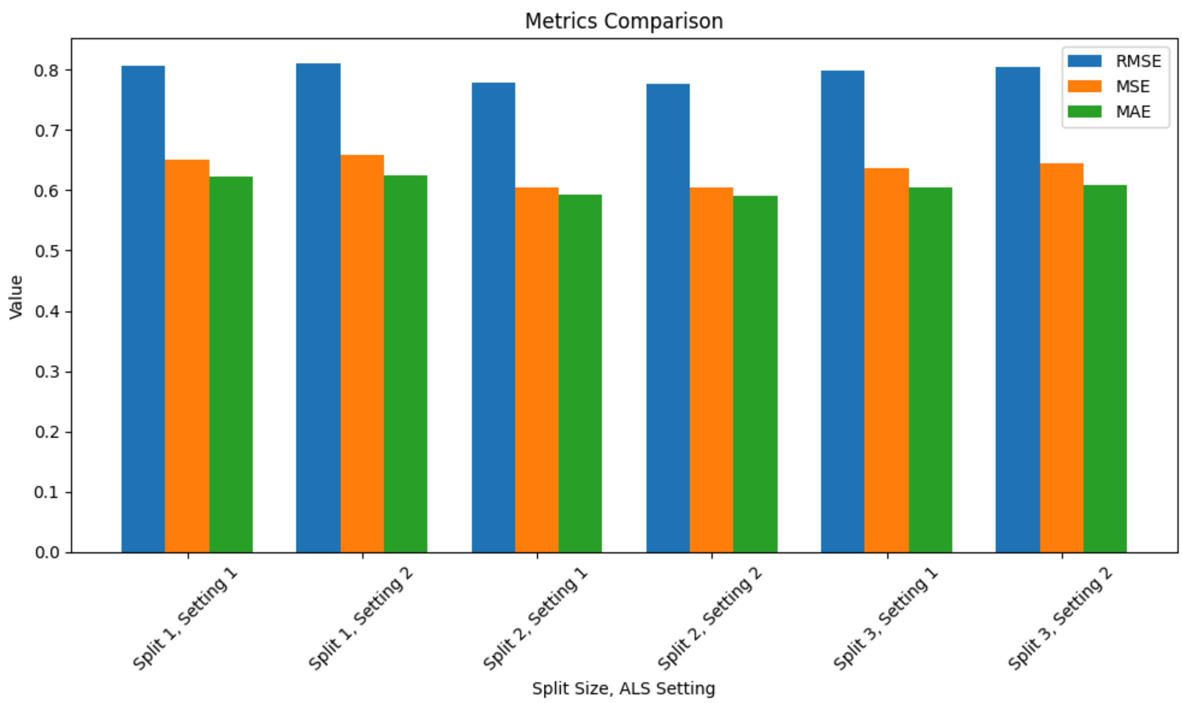
40%	60%	80%
Rumor of Angels, A (2000)	My Voyage to Italy (Il mio viaggio in Italia) (1999)	Sombre (1998)
Born Yesterday (1950)	Outrage (2009)	Rockers (1978)
That Obscure Object of Desire (Cet obscur objet du désir) (1977)	Fifth Estate, The (2013)	Closer You Get, The (2000)
Moscow Does Not Believe in Tears (Moskva slezam ne verit) (1979)	Home (2009)	Fetishes (1996)
Robot & Frank (2012)	Naked Kiss, The (1964)	Beneath Hill 60 (2010)

Top 10 Genres

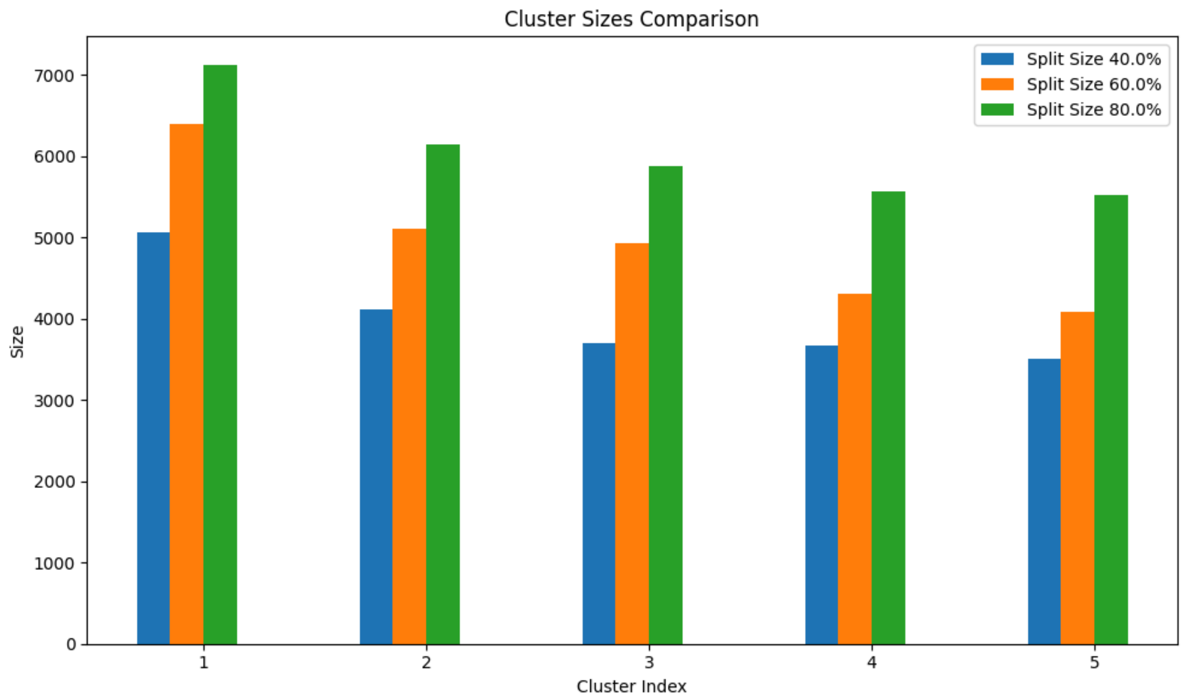
40%		60%		80%	
Genre	Count	Genre	Count	Genre	Count
Drama: 792	792	Drama	628	Drama	1974
Comedy	366	Comedy	231	Comedy	826

Romance	197	Romance	183	Documentar y	572
Thriller	195	Documentar y	156	Romance	463
Crime	174	Thriller	139	Crime	405
Action	134	Crime	132	Thriller	380
Adventure	118	War	86	Action	277
Documentar y	115	Action	86	War	233
War	93	Adventure	79	Adventure	214
Mystery	79	Mystery	63	Mystery	189

Comparison of 3 metric values (RMSE, MSE, and MAE) for 2 ALS settings and 3 Split sizes



Top 5 clusters comparison for different split size



Observation 1

From the error metrics plot, it can be seen that as the size of the training data increases, i.e. the split increases, the error values steadily decrease. This is caused due to higher samples for training resulting in better predictions. This observation presents the impact of data quantity for training. It is paramount that OTT platforms such as Netflix, continuously update the data set and model fitting to provide relevant recommendations and a better satisfaction for the users.

Observation 2

The clusters chart shows that there's a steady increase in the size of clusters as the training size increases. It can also be seen from the table that there are distinct clusters for each training size indicating changing preferences, demographics, and habits over time. This could help Netflix to tailor its content recommendation systems to drive higher user engagement and retention.