Table of contents

Q1:Log Mining and Analysis

A: Find out the total number of requests and visualize

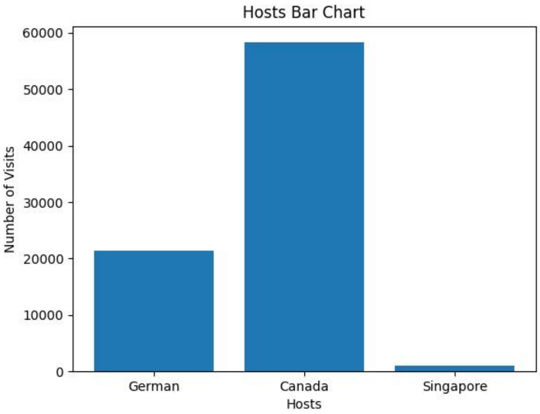
1. Here is the total number for request among three countries:

There are 21346 requests for all hosts from Germany in total.

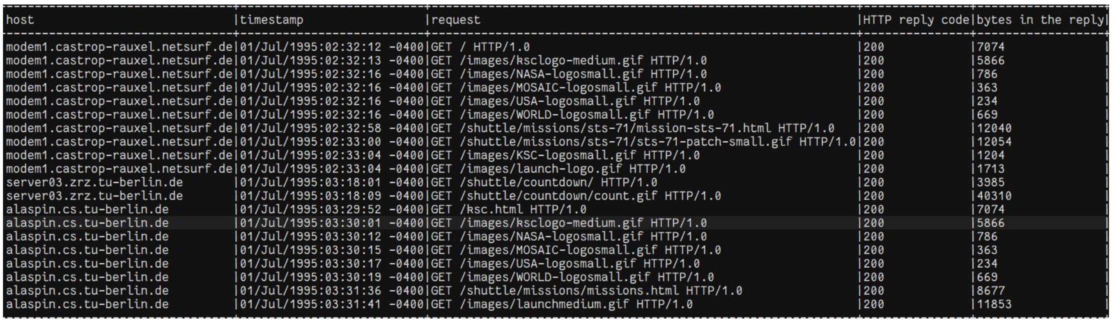
There are 58290 requests for all hosts from Canada in total.

There are 1057 requests for all hosts from Singapore in total.

Here is the bar plot to visualize all hosts:

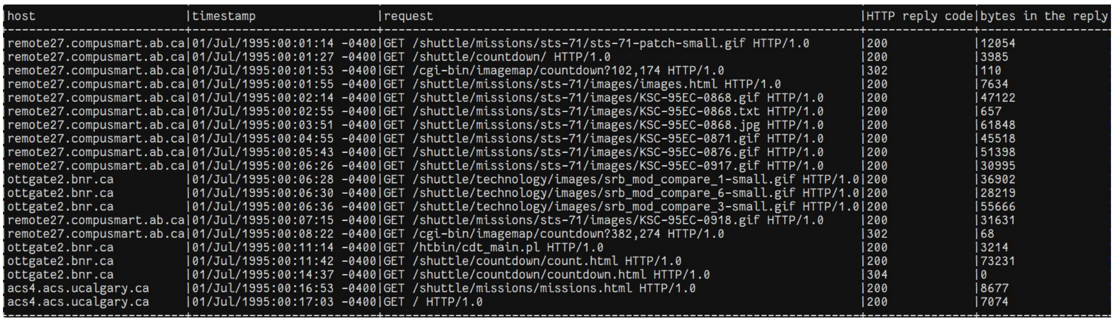


2. The visualization of all hosts from Germany is as follows :



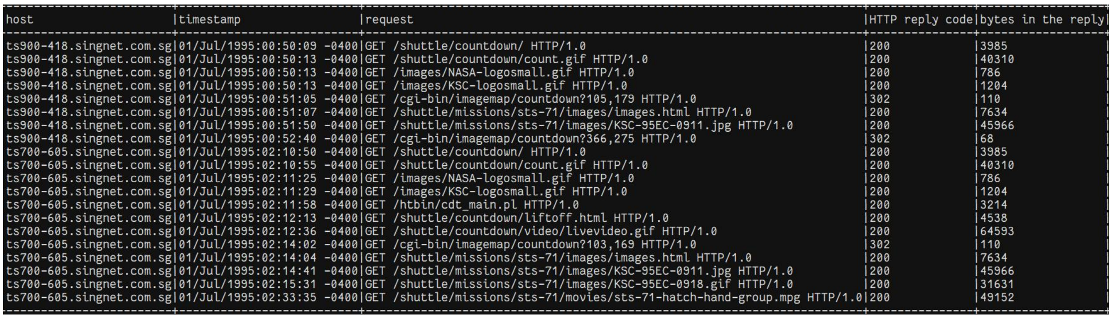
```
--------------------------------+--------------------------+------------------------------------------------------------+---------------+------------------+
host                            |timestamp                 |request                                                     |HTTP reply code|bytes in the reply|
--------------------------------+--------------------------+------------------------------------------------------------+---------------+------------------+
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:12 -0400|GET / HTTP/1.0                                              |200            |7074              |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:13 -0400|GET /images/ksclogo-medium.gif HTTP/1.0                     |200            |5866              |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:16 -0400|GET /images/NASA-logosmall.gif HTTP/1.0                     |200            |786               |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:16 -0400|GET /images/MOSAIC-logosmall.gif HTTP/1.0                   |200            |363               |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:16 -0400|GET /images/USA-logosmall.gif HTTP/1.0                      |200            |234               |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:16 -0400|GET /images/WORLD-logosmall.gif HTTP/1.0                    |200            |669               |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:32:58 -0400|GET /shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0   |200            |12040             |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:33:00 -0400|GET /shuttle/missions/sts-71/sts-71-patch-small.gif HTTP/1.0|200            |12054             |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:33:04 -0400|GET /images/KSC-logosmall.gif HTTP/1.0                      |200            |1204              |
modem1.castrop-rauxel.netsurf.de|01/Jul/1995:02:33:04 -0400|GET /images/launch-logo.gif HTTP/1.0                        |200            |1713              |
server03.zrz.tu-berlin.de       |01/Jul/1995:03:18:01 -0400|GET /shuttle/countdown/ HTTP/1.0                            |200            |3985              |
server03.zrz.tu-berlin.de       |01/Jul/1995:03:18:09 -0400|GET /shuttle/countdown/count.gif HTTP/1.0                   |200            |40310             |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:29:52 -0400|GET /ksc.html HTTP/1.0                                      |200            |7074              |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:30:01 -0400|GET /images/ksclogo-medium.gif HTTP/1.0                     |200            |5866              |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:30:12 -0400|GET /images/NASA-logosmall.gif HTTP/1.0                     |200            |786               |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:30:15 -0400|GET /images/MOSAIC-logosmall.gif HTTP/1.0                   |200            |363               |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:30:17 -0400|GET /images/USA-logosmall.gif HTTP/1.0                      |200            |234               |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:30:19 -0400|GET /images/WORLD-logosmall.gif HTTP/1.0                    |200            |669               |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:31:36 -0400|GET /shuttle/missions/missions.html HTTP/1.0                |200            |8677              |
alaspin.cs.tu-berlin.de         |01/Jul/1995:03:31:41 -0400|GET /images/launchmedium.gif HTTP/1.0                       |200            |11853             |
--------------------------------+--------------------------+------------------------------------------------------------+---------------+------------------+
```

The visualization of all hosts from Canada is as follows :



```
|host                     |timestamp                 |request                                                            |HTTP reply code|bytes in the reply
+-------------------------+--------------------------+-------------------------------------------------------------------+---------------+------------------
|remote27.compusmart.ab.ca|01/Jul/1995:00:01:14 -0400|GET /shuttle/missions/sts-71/sts-71-patch-small.gif HTTP/1.0       |200            |12054
|remote27.compusmart.ab.ca|01/Jul/1995:00:01:27 -0400|GET /shuttle/countdown/ HTTP/1.0                                   |200            |3985
|remote27.compusmart.ab.ca|01/Jul/1995:00:01:53 -0400|GET /cgi-bin/imagemap/countdown?102,174 HTTP/1.0                   |302            |110
|remote27.compusmart.ab.ca|01/Jul/1995:00:01:55 -0400|GET /shuttle/missions/sts-71/images/images.html HTTP/1.0           |200            |7634
|remote27.compusmart.ab.ca|01/Jul/1995:00:02:14 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0868.gif HTTP/1.0    |200            |47122
|remote27.compusmart.ab.ca|01/Jul/1995:00:02:55 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0868.txt HTTP/1.0    |200            |657
|remote27.compusmart.ab.ca|01/Jul/1995:00:03:51 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0868.jpg HTTP/1.0    |200            |61848
|remote27.compusmart.ab.ca|01/Jul/1995:00:04:55 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0871.gif HTTP/1.0    |200            |45518
|remote27.compusmart.ab.ca|01/Jul/1995:00:05:43 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0876.gif HTTP/1.0    |200            |51398
|remote27.compusmart.ab.ca|01/Jul/1995:00:06:26 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0917.gif HTTP/1.0    |200            |30995
|ottgate2.bnr.ca          |01/Jul/1995:00:06:28 -0400|GET /shuttle/technology/images/srb_mod_compare_1-small.gif HTTP/1.0|200            |36902
|ottgate2.bnr.ca          |01/Jul/1995:00:06:30 -0400|GET /shuttle/technology/images/srb_mod_compare_6-small.gif HTTP/1.0|200            |28219
|ottgate2.bnr.ca          |01/Jul/1995:00:06:36 -0400|GET /shuttle/technology/images/srb_mod_compare_3-small.gif HTTP/1.0|200            |55666
|remote27.compusmart.ab.ca|01/Jul/1995:00:07:15 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0918.gif HTTP/1.0    |200            |31631
|remote27.compusmart.ab.ca|01/Jul/1995:00:08:22 -0400|GET /cgi-bin/imagemap/countdown?382,274 HTTP/1.0                   |302            |68
|ottgate2.bnr.ca          |01/Jul/1995:00:11:14 -0400|GET /htbin/cdt_main.pl HTTP/1.0                                    |200            |3214
|ottgate2.bnr.ca          |01/Jul/1995:00:11:42 -0400|GET /shuttle/countdown/count.html HTTP/1.0                         |200            |73231
|ottgate2.bnr.ca          |01/Jul/1995:00:14:37 -0400|GET /shuttle/countdown/countdown.html HTTP/1.0                     |304            |0
|acs4.acs.ucalgary.ca     |01/Jul/1995:00:16:53 -0400|GET /shuttle/missions/missions.html HTTP/1.0                       |200            |8677
|acs4.acs.ucalgary.ca     |01/Jul/1995:00:17:03 -0400|GET / HTTP/1.0                                                     |200            |7074
+-------------------------+--------------------------+-------------------------------------------------------------------+---------------+------------------
```

The visualization of all hosts from Canada is as follows :



```
------------------------+--------------------------+------------------------------------------------------------------------+---------------+------------------+
host                    |timestamp                 |request                                                                 |HTTP reply code|bytes in the reply|
------------------------+--------------------------+------------------------------------------------------------------------+---------------+------------------+
ts900-418.singnet.com.sg|01/Jul/1995:00:50:09 -0400|GET /shuttle/countdown/ HTTP/1.0                                        |200            |3985              |
ts900-418.singnet.com.sg|01/Jul/1995:00:50:13 -0400|GET /shuttle/countdown/count.gif HTTP/1.0                               |200            |40310             |
ts900-418.singnet.com.sg|01/Jul/1995:00:50:13 -0400|GET /images/NASA-logosmall.gif HTTP/1.0                                 |200            |786               |
ts900-418.singnet.com.sg|01/Jul/1995:00:50:13 -0400|GET /images/KSC-logosmall.gif HTTP/1.0                                  |200            |1204              |
ts900-418.singnet.com.sg|01/Jul/1995:00:51:05 -0400|GET /cgi-bin/imagemap/countdown?105,179 HTTP/1.0                        |302            |110               |
ts900-418.singnet.com.sg|01/Jul/1995:00:51:07 -0400|GET /shuttle/missions/sts-71/images/images.html HTTP/1.0                |200            |7634              |
ts900-418.singnet.com.sg|01/Jul/1995:00:51:50 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0911.jpg HTTP/1.0          |200            |45966             |
ts900-418.singnet.com.sg|01/Jul/1995:00:52:40 -0400|GET /cgi-bin/imagemap/countdown?366,275 HTTP/1.0                        |302            |68                |
ts700-605.singnet.com.sg|01/Jul/1995:02:10:50 -0400|GET /shuttle/countdown/ HTTP/1.0                                        |200            |3985              |
ts700-605.singnet.com.sg|01/Jul/1995:02:10:55 -0400|GET /shuttle/countdown/count.gif HTTP/1.0                               |200            |40310             |
ts700-605.singnet.com.sg|01/Jul/1995:02:11:25 -0400|GET /images/NASA-logosmall.gif HTTP/1.0                                 |200            |786               |
ts700-605.singnet.com.sg|01/Jul/1995:02:11:29 -0400|GET /images/KSC-logosmall.gif HTTP/1.0                                  |200            |1204              |
ts700-605.singnet.com.sg|01/Jul/1995:02:11:58 -0400|GET /htbin/cdt_main.pl HTTP/1.0                                         |200            |3214              |
ts700-605.singnet.com.sg|01/Jul/1995:02:12:13 -0400|GET /shuttle/countdown/liftoff.html HTTP/1.0                            |200            |4538              |
ts700-605.singnet.com.sg|01/Jul/1995:02:12:36 -0400|GET /shuttle/countdown/video/livevideo.gif HTTP/1.0                     |200            |64593             |
ts700-605.singnet.com.sg|01/Jul/1995:02:14:02 -0400|GET /cgi-bin/imagemap/countdown?103,169 HTTP/1.0                        |302            |110               |
ts700-605.singnet.com.sg|01/Jul/1995:02:14:04 -0400|GET /shuttle/missions/sts-71/images/images.html HTTP/1.0                |200            |7634              |
ts700-605.singnet.com.sg|01/Jul/1995:02:14:41 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0911.jpg HTTP/1.0          |200            |45966             |
ts700-605.singnet.com.sg|01/Jul/1995:02:15:31 -0400|GET /shuttle/missions/sts-71/images/KSC-95EC-0918.gif HTTP/1.0          |200            |31631             |
ts700-605.singnet.com.sg|01/Jul/1995:02:33:35 -0400|GET /shuttle/missions/sts-71/movies/sts-71-hatch-hand-group.mpg HTTP/1.0|200            |49152             |
------------------------+--------------------------+------------------------------------------------------------------------+---------------+------------------+
```

B: Find the number of unique hosts and the top 9 most frequent hosts

1. The number of **unique** hosts among three countries:

Hello Spark: There are 1139 unique hosts from German.

Hello Spark: There are 2970 unique hosts from Canada.

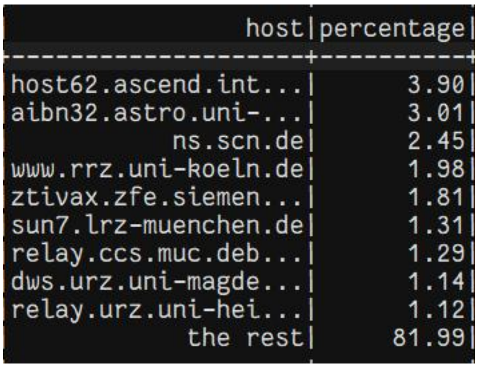
Hello Spark: There are 78 unique hosts from Singapore.

2. The top 9 most frequent hosts:

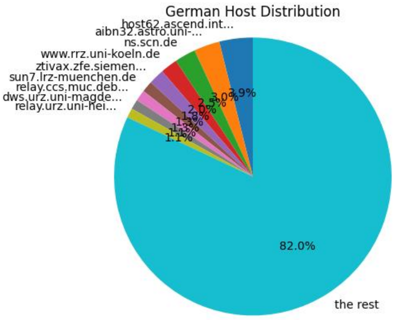
Most frequent hosts in German	***Most frequent hosts in Canada***	***Most frequent hosts in Singapore***
Host62.ascend.interop.eunet.de	*ottgate2.bnr.ca*	*merlion.singnet.com.sg*
aibn32.astro.uni-bonn.de	*freenet.edmonton.ab.ca*	*sunsite.nus.sg*
ns.scn.de	*bianca.osc.on.ca*	*ts900-1314.singnet.com.sg*
www.rrz.uni-koeln.de	*alize.ere.umontreal.ca*	*ssc25.iscs.nus.sg*
ztivax.zfe.siemens.de	*pcrb.ccrs.emr.ca*	*scctn02.sp.ac.sg*
sun7.lrz-muenchen.de	*srv1.freenet.calgary.ab.ca*	*ts900-1305.singnet.com.sg*
relay.ccs.muc.debis.de	*ccn.cs.dal.ca, oncomdis.on.ca*	*ts900-406.singnet.com.sg*
dws.urz.uni-magdeburg.de	*cobain.arcs.bcit.bc.ca*	*ts900-402.singnet.com.sg*
relay.urz.uni-heidelberg.de	*ottgate2.bnr.ca*	*einstein.technet.sg*

C: Visualize the percentage of requests by top 9 most frequent hosts and rest

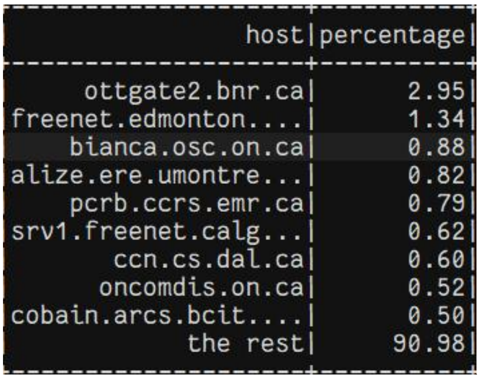
The visualization of the percentage of requests from German is:



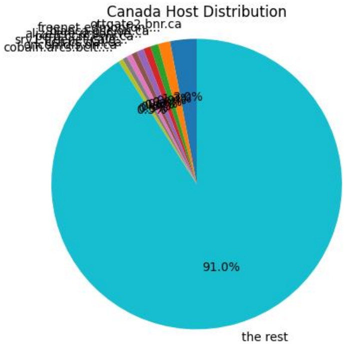
```
                 host|percentage|
---------------------+----------+
host62.ascend.int...|      3.90|
aibn32.astro.uni-...|      3.01|
           ns.scn.de|      2.45|
 www.rrz.uni-koeln.de|      1.98|
ztivax.zfe.siemen...|      1.81|
sun7.lrz-muenchen.de|      1.31|
relay.ccs.muc.deb...|      1.29|
dws.urz.uni-magde...|      1.14|
relay.urz.uni-hei...|      1.12|
            the rest|     81.99|
```



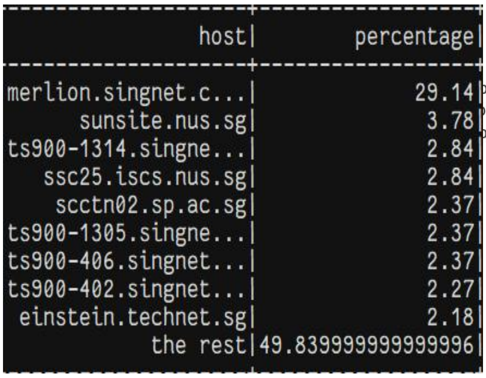
The visualization of the percentage of requests from Canada is:



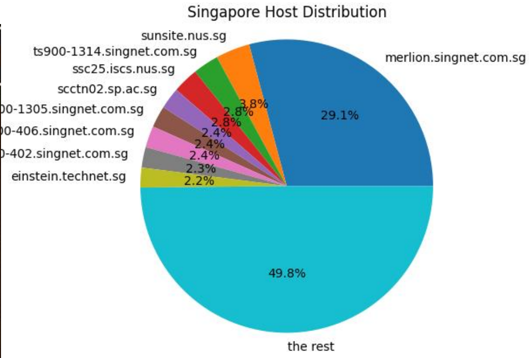
```
---------------------+----------+
                 host|percentage|
---------------------+----------+
      ottgate2.bnr.ca|      2.95|
freenet.edmonton....|      1.34|
     bianca.osc.on.ca|      0.88|
alize.ere.umontre...|      0.82|
     pcrb.ccrs.emr.ca|      0.79|
srv1.freenet.calg...|      0.62|
        ccn.cs.dal.ca|      0.60|
       oncomdis.on.ca|      0.52|
cobain.arcs.bcit....|      0.50|
            the rest|     90.98|
---------------------+----------+
```



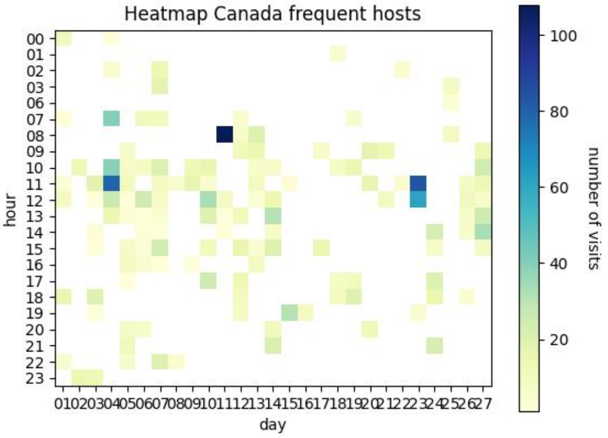
The visualization of the percentage of requests from Singapore is:



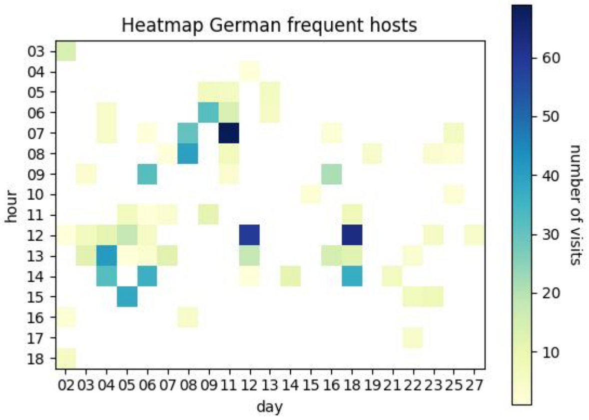
```
------------------+------------------+
              host|        percentage|
------------------+------------------+
merlion.singnet.c...|           29.14|
     sunsite.nus.sg|              3.78|
ts900-1314.singne...|            2.84|
     ssc25.iscs.nus.sg|           2.84|
     scctn02.sp.ac.sg|            2.37|
ts900-1305.singne...|            2.37|
ts900-406.singnet...|             2.37|
ts900-402.singnet...|             2.27|
  einstein.technet.sg|            2.18|
          the rest|49.839999999999996|
------------------+------------------+
```



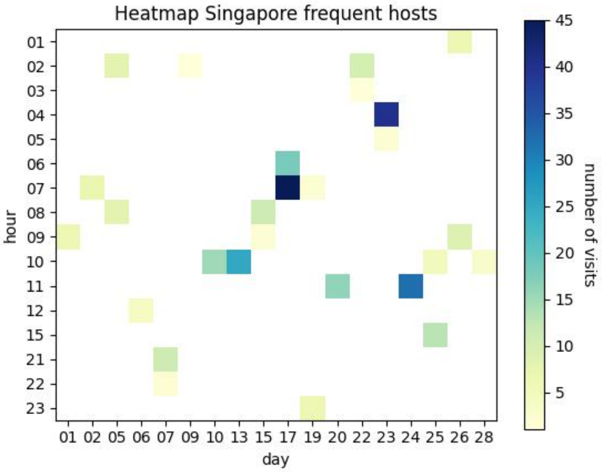
***D: For the most frequent host from each of the three countries, plot* Heat map**



FigureD1: Heatmap German frequent hosts



FigureD2: Heatmap Canada frequent hosts



FigureD3: Heatmap Singapore frequent hosts

E: Discussion

Observation:

From part C, I found there is a wide variety of hosts that are visited in German and Canada, and that no single host dominates the landscape.

From part D, in Canada, while the monthly visitation is the most intensive among the three countries, the visits are relatively average and not concentrated in a certain period, which could indicate a more even distribution of visitation throughout the month.

In Singapore, while one host accounts for 25% of the overall visits, the monthly visits are not intensive, which suggests that the visits are spread out across the month among different hosts or activities.

Reason:

The observation suggests that there may be certain events or activities that are driving visitation during these time periods.

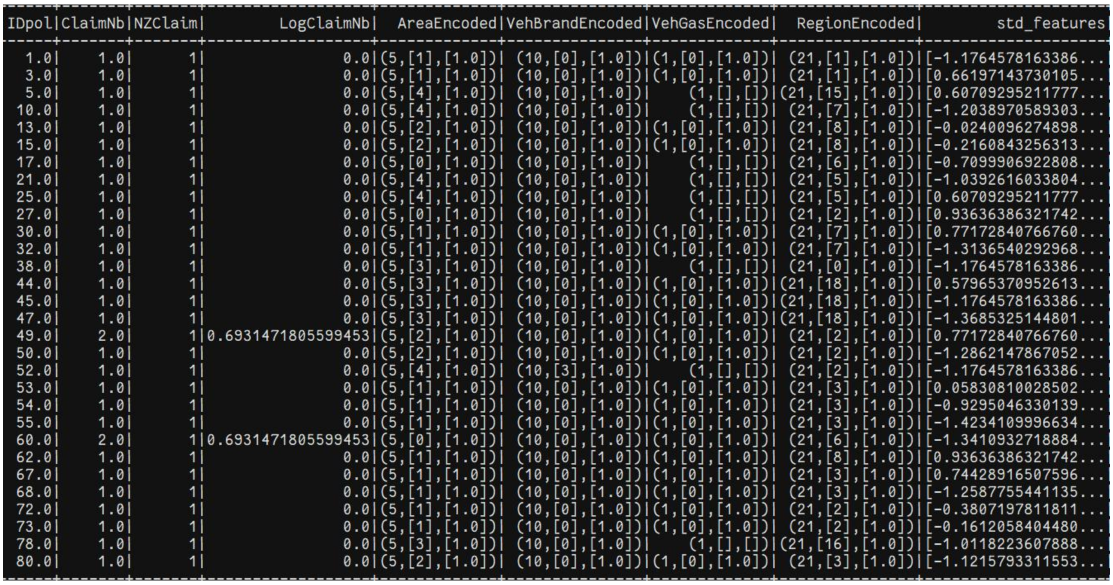
It's worth noting that these patterns may be influenced by a variety of factors, such as cultural norms, seasonal events, or tourism trends.

How useful to NASA:

This data can help NASA improve its website and resources to better meet the needs of users in different countries. In addition, NASA may also use the data to analyze traffic and usage of its websites to determine which resources or pages are popular and which may need improvement or retirement. This data also helps NASA understand the interests and needs of its international audience to guide its international cooperation and outreach activities.

Q2 :Liability Claim Prediction

A: Show the data with 2 new columns named NZClaim and LogClaim, then transform numeric features to one-hot code and get the standard features as the last column



```
+-----+-------+-------+------------------+-------------+---------------+-------------+---------------+--------------------+
|IDpol|ClaimNb|NZClaim|        LogClaimNb|  AreaEncoded|VehBrandEncoded|VehGasEncoded|  RegionEncoded|        std_features|
+-----+-------+-------+------------------+-------------+---------------+-------------+---------------+--------------------+
|  1.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[1],[1.0])|[-1.1764578163386...|
|  3.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[1],[1.0])|[0.66197143730105...|
|  5.0|    1.0|      1|               0.0|(5,[4],[1.0])| (10,[0],[1.0])|    (1,[],[])|(21,[15],[1.0])|[0.60709295211777...|
| 10.0|    1.0|      1|               0.0|(5,[4],[1.0])| (10,[0],[1.0])|    (1,[],[])| (21,[7],[1.0])|[-1.2038970589303...|
| 13.0|    1.0|      1|               0.0|(5,[2],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[8],[1.0])|[-0.0240096274898...|
| 15.0|    1.0|      1|               0.0|(5,[2],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[8],[1.0])|[-0.2160843256313...|
| 17.0|    1.0|      1|               0.0|(5,[0],[1.0])| (10,[0],[1.0])|    (1,[],[])| (21,[6],[1.0])|[-0.7099906922808...|
| 21.0|    1.0|      1|               0.0|(5,[4],[1.0])| (10,[0],[1.0])|    (1,[],[])| (21,[5],[1.0])|[-1.0392616033804...|
| 25.0|    1.0|      1|               0.0|(5,[4],[1.0])| (10,[0],[1.0])|    (1,[],[])| (21,[5],[1.0])|[0.60709295211777...|
| 27.0|    1.0|      1|               0.0|(5,[0],[1.0])| (10,[0],[1.0])|    (1,[],[])| (21,[2],[1.0])|[0.93636386321742...|
| 30.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[7],[1.0])|[0.77172840766760...|
| 32.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[7],[1.0])|[-1.3136540292968...|
| 38.0|    1.0|      1|               0.0|(5,[3],[1.0])| (10,[0],[1.0])|    (1,[],[])| (21,[0],[1.0])|[-1.1764578163386...|
| 44.0|    1.0|      1|               0.0|(5,[3],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])|(21,[18],[1.0])|[0.57965370952613...|
| 45.0|    1.0|      1|               0.0|(5,[3],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])|(21,[18],[1.0])|[-1.1764578163386...|
| 47.0|    1.0|      1|               0.0|(5,[3],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])|(21,[18],[1.0])|[-1.3685325144801...|
| 49.0|    2.0|      1|0.6931471805599453|(5,[2],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[2],[1.0])|[0.77172840766760...|
| 50.0|    1.0|      1|               0.0|(5,[2],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[2],[1.0])|[-1.2862147867052...|
| 52.0|    1.0|      1|               0.0|(5,[4],[1.0])| (10,[3],[1.0])|    (1,[],[])| (21,[2],[1.0])|[-1.1764578163386...|
| 53.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[3],[1.0])|[0.05830810028502...|
| 54.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[3],[1.0])|[-0.9295046330139...|
| 55.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[3],[1.0])|[-1.4234109996634...|
| 60.0|    2.0|      1|0.6931471805599453|(5,[0],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[6],[1.0])|[-1.3410932718884...|
| 62.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[8],[1.0])|[0.93636386321742...|
| 67.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[3],[1.0])|[0.74428916507596...|
| 68.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[3],[1.0])|[-1.2587755441135...|
| 72.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[2],[1.0])|[-0.3807197811811...|
| 73.0|    1.0|      1|               0.0|(5,[1],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[2],[1.0])|[-0.1612058404480...|
| 78.0|    1.0|      1|               0.0|(5,[3],[1.0])| (10,[0],[1.0])|    (1,[],[])|(21,[16],[1.0])|[-1.0118223607888...|
| 80.0|    1.0|      1|               0.0|(5,[2],[1.0])| (10,[0],[1.0])|(1,[0],[1.0])| (21,[3],[1.0])|[-1.1215793311553...|
+-----+-------+-------+------------------+-------------+---------------+-------------+---------------+--------------------+
```

B: Train predictive models

a. As shown in part A, standardize numeric features and use one-hot encoding to transform categorical features.

b. Train 3 models

Model the number of claims (ClaimNb) conditionally on the input features via Poisson regression:

RMSE = 0.243525

Coefficients:

[0.390, 0.023, -0.199, 0.126, 0.318, 0.015, -0.013, 0.062, 0.074, -0.092,-0.042, 0.047, 0.006, 0.004, -0.005, 0.057, -0.036, -0.008, -0.033, 0.020, 0.034, 0.083, 0.127, 0.120, -0.005, 0.010, 0.161, 0.047, -0.068, -0.102, -0.115, 0.020, -0.112, -0.109, 0.046, -0.021, -0.118, 0.044, -0.097, 0.103, 0.036, 0.018, 0.033]

Model the relationship between LogClaimNb and the input features via Linear regression, with L1 and L2 regularisation respectively:

L1 (OWL-QN optimisation) RMSE = 2.0167591695550033

L2 (OWL-QN optimisation) RMSE = 2.016397357599518

L1Coefficients:

[0.16378947104845262, 0.0, -0.07222117938016308, 0.0532879458953813, 0.15046566828597863, 0.004243143082392956, 0.0, 0.0, 0.0042415091645041396, -0.014992135702211364, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.02387591392525642, 0.01649529723828125, 0.033035897386206214, 0.0, 0.0, 0.038295589482344766, 0.0, -0.011389260823599652, -0.01877629906302139, -0.020504049133862164, 0.0, -0.010073132108149125, -0.00015519033811396883, 0.0, 0.0, -0.005900658170392603, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

L2Coefficients:

[0.1719667884720457, 0.011252772102841039, -0.08753089244792145, 0.06877781107618207, 0.16874512392288887, 0.007019739423264425, -0.017976017712848447, 0.01677174253098872, 0.02123750083760751, -0.055047865738239055, -0.031167690108278568, 0.04552339175366616, 0.035749905796856066, 0.03789194651328797, 0.03509702775198306, 0.07101578784890497, 0.01470586034007366, 0.028276376392056553, 0.016242738446361028, 0.045745831021298894, 0.0583521659048458, 0.040102121781244705, 0.062246049021597305, 0.05991391777502911, -0.01695438119786953, -0.0065119362877183755, 0.09284083855509394, 0.023605845335454276, -0.043709932478914366, -0.05508592073535326, -0.06929072485269377, 0.008178917083840234, -0.06370031531457139, -0.0752267724634265, 0.03601126481061907, -0.014133467394680849, -0.081277529510302, 0.04290346698332243, -0.09311692351637949, 0.11040582052383617, -0.0023585328961733187, 0.008255740413271364, 0.052879645840694545]

Model the relationship between NZClaim and the input features via Logistic regression, with L1 and L2 regularisation respectively:

AUC for L1 model: 0.9247338213629795

AUC for L2 model: 0.9247289024035841

L1 Coefficients:

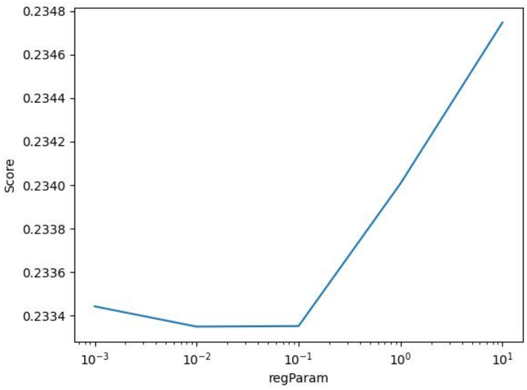
```
[-0.06560581 -0.02008492  0.06560581  0.02008492]
```

L2 Coefficients:

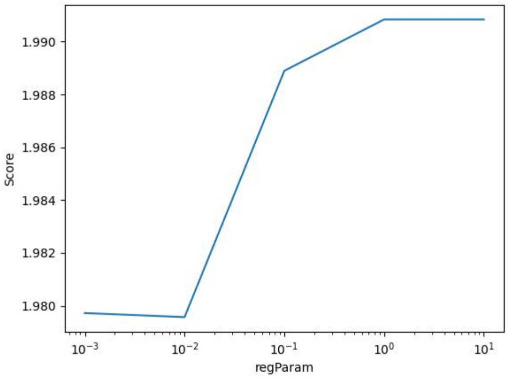
```
[-0.18012501 -0.00907092  0.09057696 -0.05604882 -0.14693967 -0.00871455
  0.01074738 -0.02753482 -0.03137067  0.04972128  0.02467286 -0.01948877
  0.00155807 -0.0029228   0.00029599 -0.03584824  0.02161379  0.00626498
  0.02401314 -0.01169469 -0.0227633  -0.0430015  -0.05716328 -0.05090142
  0.02981788  0.01177079 -0.08220546 -0.0113687   0.0687637   0.08100137
  0.09552075  0.00127097  0.09577265  0.08150774 -0.0307423   0.02737457
  0.13871974 -0.03677487  0.13077422 -0.09571989  0.00913854  0.00473782
 -0.04001646  0.18012501  0.00907092 -0.09057696  0.05604882  0.14693967
  0.00871455 -0.01074738  0.02753482  0.03137067 -0.04972128 -0.02467286
  0.01948877 -0.00155807  0.0029228  -0.00029599  0.03584824 -0.02161379
 -0.00626498 -0.02401314  0.01169469  0.0227633   0.0430015   0.05716328
  0.05090142 -0.02981788 -0.01177079  0.08220546  0.0113687  -0.0687637
 -0.08100137 -0.09552075 -0.00127097 -0.09577265 -0.08150774  0.0307423
 -0.02737457 -0.13871974  0.03677487 -0.13077422  0.09571989 -0.00913854]
```

C: Choose the best hyperparameter for all model

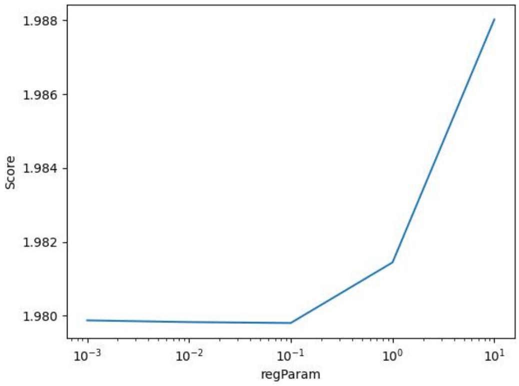
1. the best regparameter for poisson distribution regression is **0.01**, and the validation loss curve performance:



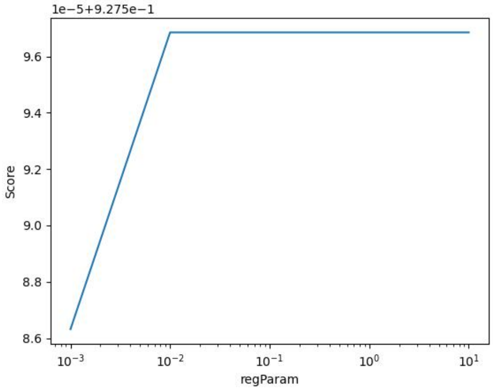
2. the best regparameter for l1 linear regression is **0.01**, and the validation loss curve performance:



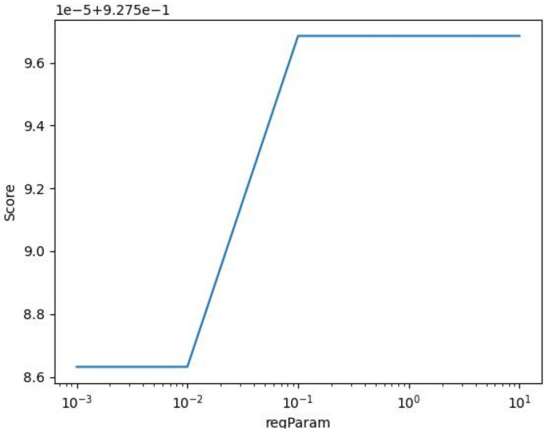
3. the best regparameter for l2 linear regression is **0.1,** and the validation loss curve performance:



4. the best regparameter for l1 logistic regression is **0.001,** and the validation loss curve performance:



5. the best regparameter for l2 logistic regression is **0.001,** and the validation loss curve performance:



D: Discussion

1. In logistic regression, I found only 4 features being captured by model while applying L1 regularization, I guess that is because the regularization parameter is too low, which may not be strong enough to penalize the coefficients and set some of them to zero. Or the features are highly correlated, and thus the regularization is unable to distinguish between them and ends up selecting only a subset of the correlated features.

2. In linear regression model, the majority of the coefficients have non-zero values, which suggests that the L2 regularization was not able to effectively perform feature selection. And the RMSE of the L2 linear regression with OWL-QN optimization was 2.0167591695550033, which is a relatively high value, indicating that the model's predictions are not very accurate. I guess it's because I compressed the features twice during vector assemble, which may have resulted in the loss of some information in the training set, resulting in poor model accuracy.

3. In Poisson regression,Examining the coefficient values, we can see that some input features have a positive impact on the number of claims, while others have a negative impact. For example, the coefficient of 0.390 for the first input feature suggests that an increase in this feature leads to a higher number of claims, while the coefficient of -0.199 for the third input feature suggests that an increase in this feature leads to a lower number of claims. Some input features have coefficients close to zero, indicating that they have little impact on the number of claims.

Q3：Movie Recommendation and Cluster Analysis

A: Here consider three such splits with three training data sizes: 40%, 60%, and 80%.

1. Data processing
2. Study two versions (settings) of ALS and analysis their behavior.

Theoretically, Increasing the rank can potentially capture more complex user-item interactions and improve the model's accuracy. Increasing the number of iterations can potentially improve the model's accuracy by allowing it to better optimize the objective function. Decreasing the regularization parameter can potentially improve the model's accuracy by allowing it to better fit the training data.

In reality, by adjusting these hyperparameters, we can find the optimal combination that results in the lowest RMSE, MSE, and MAE values. These metrics measure the difference between the predicted and actual ratings.

So I choose the setting1 and 2 below:

Setting1: rank=10, maxIter=10, regParam=0.1

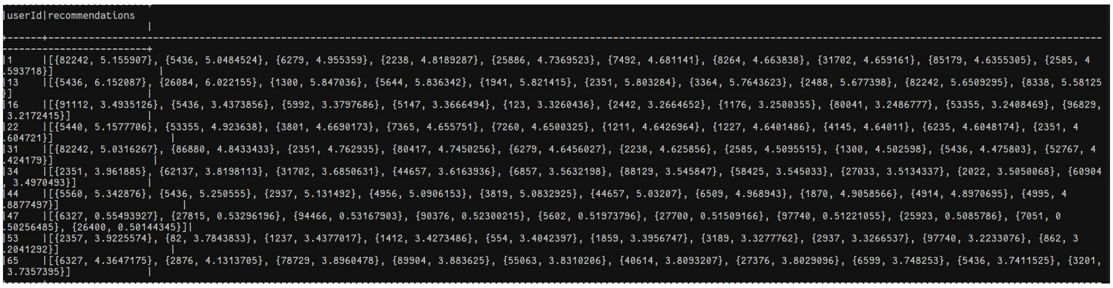
Setting2: rank=15, maxIter=15, regParam=0.01

3. Visualize RMSE, MSE, MAE:

	Setting1			***Setting2***		
	0.8 split	***0.6 split***	***0.4 split***	***0.8 split***	***0.6 split***	***0.4 split***
RMSE	*1.6853*	*2.3742*	*2.8978*	*3.4398*	*3.4656*	*3.3116*
MSE	*2.8402*	*5.6368*	*8.3977*	*11.8320*	*12.0105*	*10.9669*
MAE	*1.3611*	*1.9817*	*2.5033*	*2.9625*	*3.0346*	*2.9348*

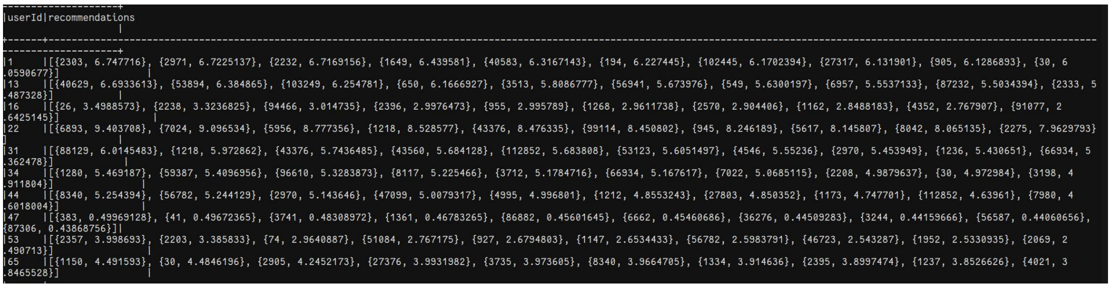
For 0.8 training data:

Generate top 10 movie recommendations for each user by using 0.8 split setting1:



```
+------------------------+
|userId|recommendations
                         |
+------+----------------------------------------------------------------------------------------------------------
------------------------+
|1     |[{82242, 5.155907}, {5436, 5.0484524}, {6279, 4.955359}, {2238, 4.8189287}, {25886, 4.7369523}, {7492, 4.681141}, {8264, 4.663838}, {31702, 4.659161}, {85179, 4.6355305}, {2585, 4
.593718}]                |
|13    |[{5436, 6.152087}, {26084, 6.022155}, {1300, 5.847036}, {5644, 5.836342}, {1941, 5.821415}, {2351, 5.803284}, {3364, 5.7643623}, {2488, 5.677398}, {82242, 5.6509295}, {8338, 5.58125
}]                       |
|16    |[{91112, 3.4935126}, {5436, 3.4373856}, {5992, 3.3797686}, {5147, 3.3666494}, {123, 3.3260436}, {2442, 3.2664652}, {1176, 3.2500355}, {80041, 3.2486777}, {53355, 3.2408469}, {96829,
3.2172415}]              |
|22    |[{5440, 5.1577706}, {53355, 4.923638}, {3801, 4.6690173}, {7365, 4.655751}, {7260, 4.6500325}, {1211, 4.6426964}, {1227, 4.6401486}, {4145, 4.64011}, {6235, 4.6048174}, {2351, 4
.604721}]                      |
|31    |[{82242, 5.0316267}, {86880, 4.8433433}, {2351, 4.762935}, {80417, 4.7450256}, {6279, 4.6456027}, {2238, 4.625856}, {2585, 4.5095515}, {1300, 4.502598}, {5436, 4.475803}, {52767, 4
.424179}]                |
|34    |[{2351, 3.961885}, {62137, 3.8198113}, {31702, 3.6850631}, {44657, 3.6163936}, {6857, 3.5632198}, {88129, 3.545847}, {58425, 3.545033}, {27033, 3.5134337}, {2022, 3.5050068}, {60904
, 3.4970493}]            |
|44    |[{5560, 5.342876}, {5436, 5.250555}, {2937, 5.131492}, {4956, 5.0906153}, {3819, 5.0832925}, {44657, 5.03207}, {6509, 4.968943}, {1870, 4.9058566}, {4914, 4.8970695}, {4995, 4
.8877497}]                     |
|47    |[{6327, 0.55493927}, {27815, 0.53296196}, {94466, 0.53167903}, {90376, 0.52300215}, {5602, 0.51973796}, {27700, 0.51509166}, {97740, 0.51221055}, {25923, 0.5085786}, {7051, 0
.50256485}, {26400, 0.50144345}]|
|53    |[{2357, 3.9225574}, {82, 3.7843833}, {1237, 3.4377017}, {1412, 3.4273486}, {554, 3.4042397}, {1859, 3.3956747}, {3189, 3.3277762}, {2937, 3.3266537}, {97740, 3.2233076}, {862, 3
.2041292}]                     |
|65    |[{6327, 4.3647175}, {2876, 4.1313705}, {78729, 3.8960478}, {89904, 3.883625}, {55063, 3.8310206}, {40614, 3.8093207}, {27376, 3.8029096}, {6599, 3.748253}, {5436, 3.7411525}, {3201,
3.7357395}]              |
```

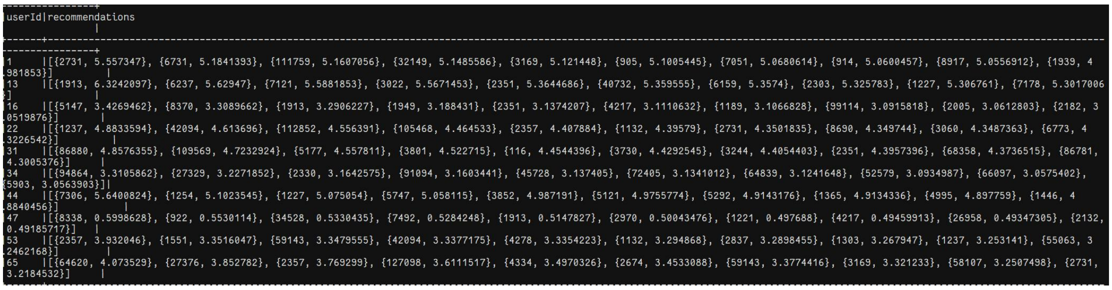
Generate top 10 movie recommendations for each user by using 0.8 split setting2:



```
+-------------------+
|userId|recommendations
                    |
+------+----------------------------------------------------------------------------------------------------------
-------------------+
|1     |[{2303, 6.747716}, {2971, 6.7225137}, {2232, 6.7169156}, {1649, 6.439581}, {40583, 6.3167143}, {194, 6.227445}, {102445, 6.1702394}, {27317, 6.131901}, {905, 6.1286893}, {30, 6
.0590677}]               |
|13    |[{40629, 6.6933613}, {53894, 6.384865}, {103249, 6.254781}, {650, 6.1666927}, {3513, 5.8086777}, {56941, 5.673976}, {549, 5.6300197}, {6957, 5.5537133}, {87232, 5.5034394}, {2333, 5
.487328}]           |
|16    |[{26, 3.4988573}, {2238, 3.3236825}, {94466, 3.014735}, {2396, 2.9976473}, {955, 2.995789}, {1268, 2.9611738}, {2570, 2.904406}, {1162, 2.8488183}, {4352, 2.767907}, {91077, 2
.6425145}]               |
|22    |[{6893, 9.403708}, {7024, 9.096534}, {5956, 8.777356}, {1218, 8.528577}, {43376, 8.476335}, {99114, 8.450802}, {945, 8.246189}, {5617, 8.145807}, {8042, 8.065135}, {2275, 7.9629793}
]                   |
|31    |[{88129, 6.0145483}, {1218, 5.972862}, {43376, 5.7436485}, {43560, 5.684128}, {112852, 5.683808}, {53123, 5.6051497}, {4546, 5.55236}, {2970, 5.453949}, {1236, 5.430651}, {66934, 5
.362478}]             |
|34    |[{1280, 5.469187}, {59387, 5.4096956}, {96610, 5.3283873}, {8117, 5.225466}, {3712, 5.1784716}, {66934, 5.167617}, {7022, 5.0685115}, {2208, 4.9879637}, {30, 4.972984}, {3198, 4
.911804}]                  |
|44    |[{8340, 5.254394}, {56782, 5.244129}, {2970, 5.143646}, {47099, 5.0079317}, {4995, 4.996801}, {1212, 4.8553243}, {27803, 4.850352}, {1173, 4.747701}, {112852, 4.63961}, {7980, 4
.6018004}]                 |
|47    |[{383, 0.49969128}, {41, 0.49672365}, {3741, 0.48308972}, {1361, 0.46783265}, {86882, 0.45601645}, {6662, 0.45460686}, {36276, 0.44509283}, {3244, 0.44159666}, {56587, 0.44060656},
{87306, 0.43868756}]|
|53    |[{2357, 3.998693}, {2203, 3.385833}, {74, 2.9640887}, {51084, 2.767175}, {927, 2.6794803}, {1147, 2.6534433}, {56782, 2.5983791}, {46723, 2.543287}, {1952, 2.5330935}, {2069, 2
.490713}]                |
|65    |[{1150, 4.491593}, {30, 4.4846196}, {2905, 4.2452173}, {27376, 3.9931982}, {3735, 3.973605}, {8340, 3.9664705}, {1334, 3.914636}, {2395, 3.8997474}, {1237, 3.8526626}, {4021, 3
.8465528}]               |
```

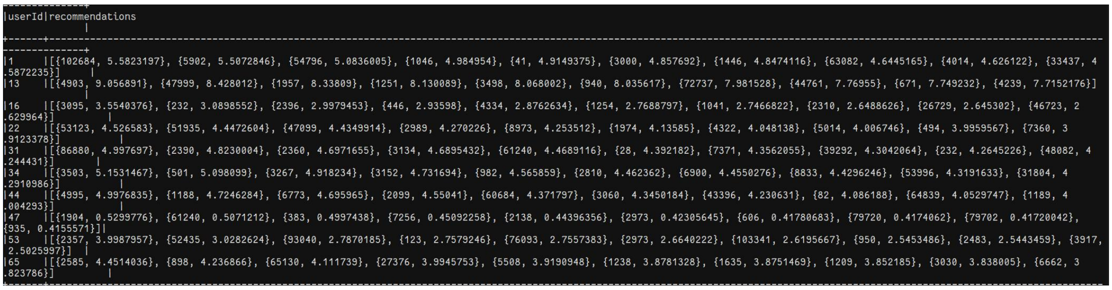
For 0.6 training data:

Generate top 10 movie recommendations for each user by using 0.6 split setting1:



```
+----------------+
|userId|recommendations
                 |
+------+----------------------------------------------------------------------------------------------------------
----------------+
|1     |[{2731, 5.557347}, {6731, 5.1841393}, {111759, 5.1607056}, {32149, 5.1485586}, {3169, 5.121448}, {905, 5.1005445}, {7051, 5.0680614}, {914, 5.0600457}, {8917, 5.0556912}, {1939, 4
.981853}]        |
|13    |[{1913, 6.3242097}, {6237, 5.62947}, {7121, 5.5881853}, {3022, 5.5671453}, {2351, 5.3644686}, {40732, 5.359555}, {6159, 5.3574}, {2303, 5.325783}, {1227, 5.306761}, {7178, 5.3017006
}]               |
|16    |[{5147, 3.4269462}, {8370, 3.3089662}, {1913, 3.2906227}, {1949, 3.188431}, {2351, 3.1374207}, {4217, 3.1110632}, {1189, 3.1066828}, {99114, 3.0915818}, {2005, 3.0612803}, {2182, 3
.0519876}]       |
|22    |[{1237, 4.8833594}, {42094, 4.613696}, {112852, 4.556391}, {105468, 4.464533}, {2357, 4.407884}, {1132, 4.39579}, {2731, 4.3501835}, {8690, 4.349744}, {3060, 4.3487363}, {6773, 4
.3226542}]           |
|31    |[{86880, 4.8576355}, {109569, 4.7232924}, {5177, 4.557811}, {3801, 4.522715}, {116, 4.4544396}, {3730, 4.4292545}, {3244, 4.4054403}, {2351, 4.3957396}, {68358, 4.3736515}, {86781,
4.3005376}]      |
|34    |[{94864, 3.3105862}, {27329, 3.2271852}, {2330, 3.1642575}, {91094, 3.1603441}, {45728, 3.137405}, {72405, 3.1341012}, {64839, 3.1241648}, {52579, 3.0934987}, {66097, 3.0575402},
{5903, 3.0563903}]|
|44    |[{7306, 5.6400824}, {1254, 5.1023545}, {1227, 5.075054}, {5747, 5.058115}, {3852, 4.987191}, {5121, 4.9755774}, {5292, 4.9143176}, {1365, 4.9134336}, {4995, 4.897759}, {1446, 4
.8840456}]           |
|47    |[{8338, 0.5998628}, {922, 0.5530114}, {34528, 0.5330435}, {7492, 0.5284248}, {1913, 0.5147827}, {2970, 0.50043476}, {1221, 0.497688}, {4217, 0.49459913}, {26958, 0.49347305}, {2132,
0.49185717}]   |
|53    |[{2357, 3.932046}, {1551, 3.3516047}, {59143, 3.3479555}, {42094, 3.3377175}, {4278, 3.3354223}, {1132, 3.294868}, {2837, 3.2898455}, {1303, 3.267947}, {1237, 3.253141}, {55063, 3
.2462168}]       |
|65    |[{64620, 4.073529}, {27376, 3.852782}, {2357, 3.769299}, {127098, 3.6111517}, {4334, 3.4970326}, {2674, 3.4533088}, {59143, 3.3774416}, {3169, 3.321233}, {58107, 3.2507498}, {2731,
3.2184532}]      |
```

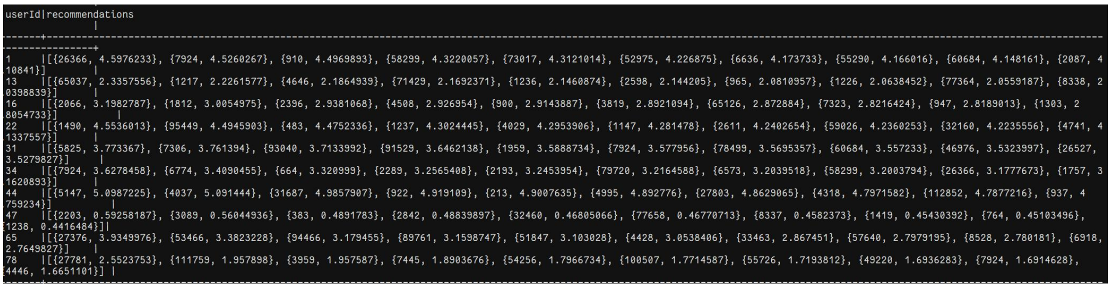
Generate top 10 movie recommendations for each user by using 0.6 split setting2:



```
+--------------+
|userId|recommendations
               |
+------+----------------------------------------------------------------------------------------------------------
--------------+
|1     |[{102684, 5.5823197}, {5902, 5.5072846}, {54796, 5.0836005}, {1046, 4.984954}, {41, 4.9149375}, {3000, 4.857692}, {1446, 4.8474116}, {63082, 4.6445165}, {4014, 4.626122}, {33437, 4
.5872235}]     |
|13    |[{4903, 9.056891}, {47999, 8.428012}, {1957, 8.33809}, {1251, 8.130089}, {3498, 8.068002}, {940, 8.035617}, {72737, 7.981528}, {44761, 7.76955}, {671, 7.749232}, {4239, 7.7152176}]
               |
|16    |[{3095, 3.5540376}, {232, 3.0898552}, {2396, 2.9979453}, {446, 2.93598}, {4334, 2.8762634}, {1254, 2.7688797}, {1041, 2.7466822}, {2310, 2.6488626}, {26729, 2.645302}, {46723, 2
.629964}]          |
|22    |[{53123, 4.526583}, {51935, 4.4472604}, {47099, 4.4349914}, {2989, 4.270226}, {8973, 4.253512}, {1974, 4.13585}, {4322, 4.048138}, {5014, 4.006746}, {494, 3.9959567}, {7360, 3
.9123378}]          |
|31    |[{86880, 4.997697}, {2390, 4.8230004}, {2360, 4.6971655}, {3134, 4.6895432}, {61240, 4.4689116}, {28, 4.392182}, {7371, 4.3562055}, {39292, 4.3042064}, {232, 4.2645226}, {48082, 4
.244431}]        |
|34    |[{3503, 5.1531467}, {501, 5.098099}, {3267, 4.918234}, {3152, 4.731694}, {982, 4.565859}, {2810, 4.462362}, {6900, 4.4550276}, {8833, 4.4296246}, {53996, 4.3191633}, {31804, 4
.2910986}]          |
|44    |[{4995, 4.9976835}, {1188, 4.7246284}, {6773, 4.695965}, {2099, 4.55041}, {60684, 4.371797}, {3060, 4.3450184}, {43396, 4.230631}, {82, 4.086188}, {64839, 4.0529747}, {1189, 4
.004293}]           |
|47    |[{1904, 0.5299776}, {61240, 0.5071212}, {383, 0.4997438}, {7256, 0.45092258}, {2138, 0.44396356}, {2973, 0.42305645}, {606, 0.41780683}, {79720, 0.4174062}, {79702, 0.41720042},
{935, 0.4155571}]|
|53    |[{2357, 3.9987957}, {52435, 3.0282624}, {93040, 2.7870185}, {123, 2.7579246}, {76093, 2.7557383}, {2973, 2.6640222}, {103341, 2.6195667}, {950, 2.5453486}, {2483, 2.5443459}, {3917,
2.5025997}]  |
|65    |[{2585, 4.4514036}, {898, 4.236866}, {65130, 4.111739}, {27376, 3.9945753}, {5508, 3.9190948}, {1238, 3.8781328}, {1635, 3.8751469}, {1209, 3.852185}, {3030, 3.838005}, {6662, 3
.823786}]           |
```

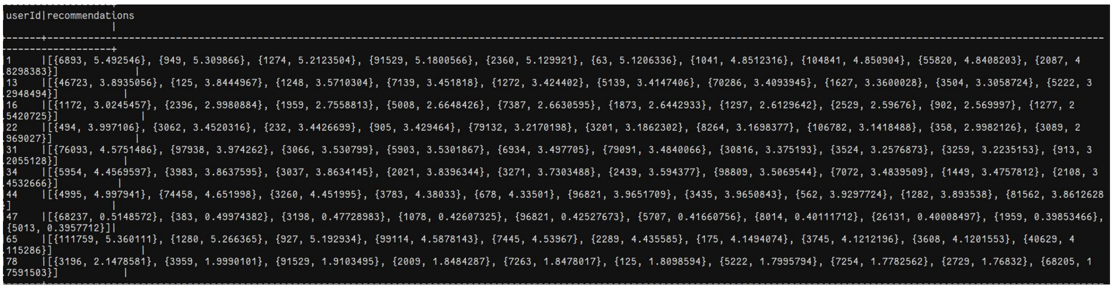
For 0.4 training data:

Generate top 10 movie recommendations for each user by using 0.4 split setting1:



```
userId|recommendations
                |
------+----------------------------------------------------------------------------------------------------------------------------------------------
----------------+
1     |[{26366, 4.5976233}, {7924, 4.5260267}, {910, 4.4969893}, {58299, 4.3220057}, {73017, 4.3121014}, {52975, 4.226875}, {6636, 4.173733}, {55290, 4.166016}, {60684, 4.148161}, {2087, 4
.10841}]        |
13    |[{65037, 2.3357556}, {1217, 2.2261577}, {4646, 2.1864939}, {71429, 2.1692371}, {1236, 2.1460874}, {2598, 2.144205}, {965, 2.0810957}, {1226, 2.0638452}, {77364, 2.0559187}, {8338, 2
.0398839}]      |
16    |[{2066, 3.1982787}, {1812, 3.0054975}, {2396, 2.9381068}, {4508, 2.926954}, {900, 2.9143887}, {3819, 2.8921094}, {65126, 2.872884}, {7323, 2.8216424}, {947, 2.8189013}, {1303, 2
.8054733}]          |
22    |[{1490, 4.5536013}, {95449, 4.4945903}, {483, 4.4752336}, {1237, 4.3024445}, {4029, 4.2953906}, {1147, 4.281478}, {2611, 4.2402654}, {59026, 4.2360253}, {32160, 4.2235556}, {4741, 4
.1337557}]      |
31    |[{5825, 3.773367}, {7306, 3.761394}, {93040, 3.7133992}, {91529, 3.6462138}, {1959, 3.5888734}, {7924, 3.577956}, {78499, 3.5695357}, {60684, 3.557233}, {46976, 3.5323997}, {26527,
3.5279827}]      |
34    |[{7924, 3.6278458}, {6774, 3.4090455}, {664, 3.320999}, {2289, 3.2565408}, {2193, 3.2453954}, {79720, 3.2164588}, {6573, 3.2039518}, {58299, 3.2003794}, {26366, 3.1777673}, {1757, 3
.1620893}]      |
44    |[{5147, 5.0987225}, {4037, 5.091444}, {31687, 4.9857907}, {922, 4.919109}, {213, 4.9007635}, {4995, 4.892776}, {27803, 4.8629065}, {4318, 4.7971582}, {112852, 4.7877216}, {937, 4
.759234}]          |
47    |[{2203, 0.59258187}, {3089, 0.56044936}, {383, 0.4891783}, {2842, 0.48839897}, {32460, 0.46805066}, {77658, 0.46770713}, {8337, 0.4582373}, {1419, 0.45430392}, {764, 0.45103496},
{1238, 0.4416484}]|
65    |[{27376, 3.9349976}, {53466, 3.3823228}, {94466, 3.179455}, {89761, 3.1598747}, {51847, 3.103028}, {4428, 3.0538406}, {33463, 2.867451}, {57640, 2.7979195}, {8528, 2.780181}, {6918,
2.7649827}]     |
78    |[{27781, 2.5523753}, {111759, 1.957898}, {3959, 1.957587}, {7445, 1.8903676}, {54256, 1.7966734}, {100507, 1.7714587}, {55726, 1.7193812}, {49220, 1.6936283}, {7924, 1.6914628},
{4446, 1.6651101}] |
------+----------------------------------------------------------------------------------------------------------------------------------------------
```

Generate top 10 movie recommendations for each user by using 0.4 split setting2:



```
|userId|recommendations
                   |
------+----------------------------------------------------------------------------------------------------------------------------------------------
-------------------+
|1     |[{6893, 5.492546}, {949, 5.309866}, {1274, 5.2123504}, {91529, 5.1800566}, {2360, 5.129921}, {63, 5.1206336}, {1041, 4.8512316}, {104841, 4.850904}, {55820, 4.8408203}, {2087, 4
.8298383}]             |
|13    |[{46723, 3.8935056}, {125, 3.8444967}, {1248, 3.5710304}, {7139, 3.451818}, {1272, 3.424402}, {5139, 3.4147406}, {70286, 3.4093945}, {1627, 3.3600028}, {3504, 3.3058724}, {5222, 3
.2948494}]          |
|16    |[{1172, 3.0245457}, {2396, 2.9980884}, {1959, 2.7558813}, {5008, 2.6648426}, {7387, 2.6630595}, {1873, 2.6442933}, {1297, 2.6129642}, {2529, 2.59676}, {902, 2.569997}, {1277, 2
.5420725}]             |
|22    |[{494, 3.997106}, {3062, 3.4520316}, {232, 3.4426699}, {905, 3.429464}, {79132, 3.2170198}, {3201, 3.1862302}, {8264, 3.1698377}, {106782, 3.1418488}, {358, 2.9982126}, {3089, 2
.969027}]              |
|31    |[{76093, 4.5751486}, {97938, 3.974262}, {3066, 3.530799}, {5903, 3.5301867}, {6934, 3.497705}, {79091, 3.4840066}, {30816, 3.375193}, {3524, 3.2576873}, {3259, 3.2235153}, {913, 3
.2055128}]          |
|34    |[{5954, 4.4569597}, {3983, 3.8637595}, {3037, 3.8634145}, {2021, 3.8396344}, {3271, 3.7303488}, {2439, 3.594377}, {98809, 3.5069544}, {7072, 3.4839509}, {1449, 3.4757812}, {2108, 3
.4532666}]         |
|44    |[{4995, 4.997941}, {74458, 4.651998}, {3260, 4.451995}, {3783, 4.38033}, {678, 4.33501}, {96821, 3.9651709}, {3435, 3.9650843}, {562, 3.9297724}, {1282, 3.893538}, {81562, 3.8612628
}]                 |
|47    |[{68237, 0.5148572}, {383, 0.49974382}, {3198, 0.47728983}, {1078, 0.42607325}, {96821, 0.42527673}, {5707, 0.41660756}, {8014, 0.40111712}, {26131, 0.40008497}, {1959, 0.39853466},
{5013, 0.3957712}]|
|65    |[{111759, 5.360111}, {1280, 5.266365}, {927, 5.192934}, {99114, 4.5878143}, {7445, 4.53967}, {2289, 4.435585}, {175, 4.1494074}, {3745, 4.1212196}, {3608, 4.1201553}, {40629, 4
.115286}]              |
|78    |[{3196, 2.1478581}, {3959, 1.9990101}, {91529, 1.9103495}, {2009, 1.8484287}, {7263, 1.8478017}, {125, 1.8098594}, {5222, 1.7995794}, {7254, 1.7782562}, {2729, 1.76832}, {68205, 1
.7591503}]          |
------+----------------------------------------------------------------------------------------------------------------------------------------------
```

Analysis:

However, the setting 2 parameters increase the risk of overfitting and it also increase the computational time.

B: User Analysis

1. Report the top five largest user clusters by 3 splits

0.8 split		***0.6 split***		***0.4 split***	
prediction	***count***	***prediction***	***count***	***prediction***	***count***
2	*14639*	*10*	*8204*	*10*	*8204*
5	*14486*	*2*	*8166*	*2*	*8166*
4	*14367*	*18*	*7994*	*18*	*7994*
10	*14350*	*22*	*7789*	*22*	*7789*
18	*14263*	*3*	*7288*	*3*	*7288*

2. The top ten most popular genres for each of the splits are:

Splits	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
0.8	*Drama*	*Comedy*	*Documentary*	*Romance*	*Thriller*	*Crime*	*Action*	*Adventure*	*War*	*Horror*
0.6	*Drama*	*Comedy*	*Documentary*	*Romance*	*Thriller*	*Action*	*Crime*	*Adventure*	*War*	*Horror*
0.4	*Drama*	*Comedy*	*Documentary*	*Romance*	*Thriller*	*Action*	*Crime*	*Adventure*	*War*	*Horror*

C: Discussion

1. Based on the the result metrics from part A, it seems that Setting1 with a split of 0.8 is the best option, as it has the lowest RMSE, MSE, and MAE values compared to the other options. The lower the RMSE, MSE, and MAE values, the better the model's predictive accuracy. The reason is that a larger training set can provide more data for modeling, which can better capture the characteristics and patterns of the data. In contrast, the training set size of 0.6 and 0.4 is small, which may not fully reflect the complexity of the data, and thus lead to a decrease in the performance of the model.

2. Based on part B, for all three splits, the top three most popular genres are Drama, Comedy, and Documentary. However, the order of the other genres changes slightly between the splits. For example, in the 0.8 split, the fourth most popular genre is Romance, while in the 0.6 and 0.4 splits, it is Thriller. It may be because these genres have wide appeal and can attract audiences from various backgrounds and preferences. It could also be because these genres have more releases and are more accessible to the general public, making them more likely to be watched and rated.

3. How useful is this observation to a movie website such as Netflix：
Understanding these popular genres and their characteristics can be useful for movie studios and streaming services to better target their audience and create more appealing content. It can also help researchers and analysts in the film industry to make more accurate predictions and recommendations based on audience preferences.

Q4：Research Paper Visualization

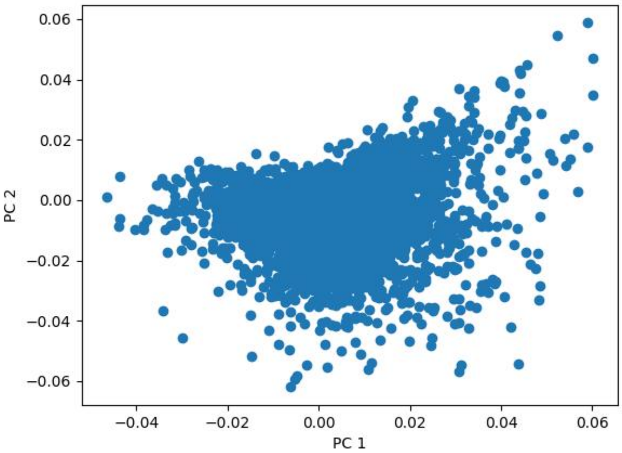
A: Show the first 10 entries of the 2 PCs, two corresponding eigenvalues and the percentage of variance they captured:

The first 10 entries of the PC,the left column is pc1, the right column is pc2:

```
[0.00406433, 0.00060763]
[ 0.00542342, -0.00376415]
[-0.01738573, 0.00117871]
[0.00556042, 0.00267248]
[0.00362411, 0.00069069]
[-0.00065361, -0.00332418]
[-0.00938831, 0.00228952]
[0.01197519, 0.01098948]
[ 0.0026566, -0.00655555]
[-0.00791011,-0.00310361]
```

The PCA model two corresponding eigenvalues is: [395796.78918083693,286173.1985073019]

B: Visualize the 5811 papers using the 2 PC



C: Discussion

Clustering of points may indicate highly correlated features in the data. This indicates that the principal components extracted by PCA are not able to capture all the variance in the data, or there are some noisy or redundant features.

Some points may be scattered because they have large variance or covariance in the original data. This means that the points explain the principal components to a higher degree, so they are easier to scatter in the dot plot. Therefore, these scattered points may represent some important features or outliers in the data.

In this case, the data might not suitable to reduce dimensions by using PCA because of highly correlated features.

Q5：Searching for exotic particles in high-energy physics using ensemble methods

A：Apply Random Forests and Gradient boosting and finding the best parameters

Here I use the same splits of training and test data(1%) when comparing performances among the two algorithms, details shown in the code.

The 1% data Accuracy for best Random Forest model = 0.646702

Here is the best configuration of parameters for Random Forest model:

para	***cacheNodeIds***	***checkpointInterval***	***featureSubsetStrategy***	***featuresCol***	***impurity***	***labelCol***	***maxBins***
value	*false*	*10*	*auto*	*features*	*gini*	*_c0*	*20*
para	***bootstrap***	***maxDepth***	***maxMemoryInMB***	***minInfoGain***	***minInstancesPerNode***	***minWeight Fraction PerNode***	***numTrees***
value	*true*	*5*	*256*	*0.0*	*1*	*0.0*	*20*
para	***predictionCol***	***probabilityCol***	***rawPredictionCol***	***seed***	***subsamplingRate***		
value	*prediction*	*probability*	*rawPrediction*	*42*	*1.0*		

The 1% data AUC-ROC for best Gradient boosting：0.6324403901225595

Here is the best configuration of parameters for Gradient boosting model:

para	***cacheNodeIds***	***checkpointInterval***	***featureSubsetStrategy***	***featuresCol***	***impurity***	***labelCol***	***maxBins***
value	*false*	*10*	*all*	*features*	*variance*	*_c0*	*32*
para	***maxIter***	***maxDepth***	***maxMemoryInMB***	***minInfoGain***	***minInstancesPerNode***	***minWeight Fraction PerNode***	***Raw Prediction Col***
value	*30*	*5*	*256*	*0.0*	*1*	*0.0*	*Raw Prediction*
para	***predictionCol***	***probabilityCol***	***seed***	***stepSize***	***Subsampling Rate***	***Validation Tol***	
value	*prediction*	*probability*	3504127614838123891	*0.1*	*1.0*	0.01	

B: Use the best parameter and use the larger data to train two models again

Here I use the same splits of training and test data(10%) when comparing performances among the two algorithms, details shown in the code.

choose the best parameter to finish the Random Forest model
The ALL data Accuracy for best Random Forest model = 0.671066

choose the best parameter to finish the GBT model
The ALL data AUC-ROC: 0.6403718329486673