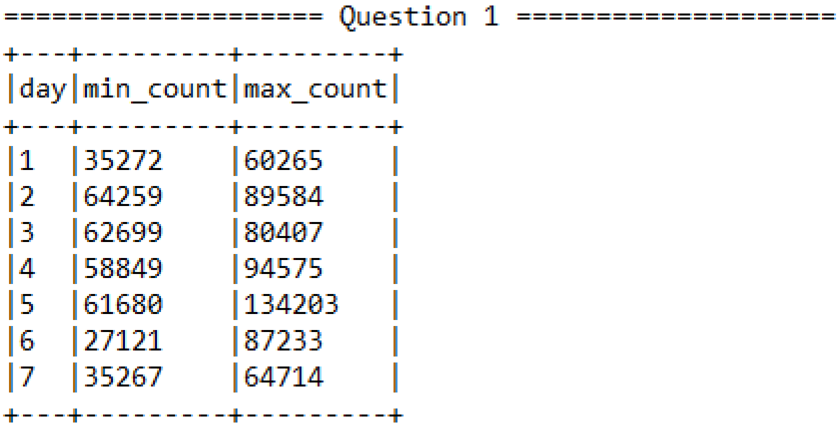
The report of Assignment 2 --SML

1. Log Mining and Analysis

A. Find out the maximum number and minimum number of requests on each of the seven days in a week (i.e., Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday) during July 1995. You need to report 14 numbers, one max number and one min number for each day of the week.

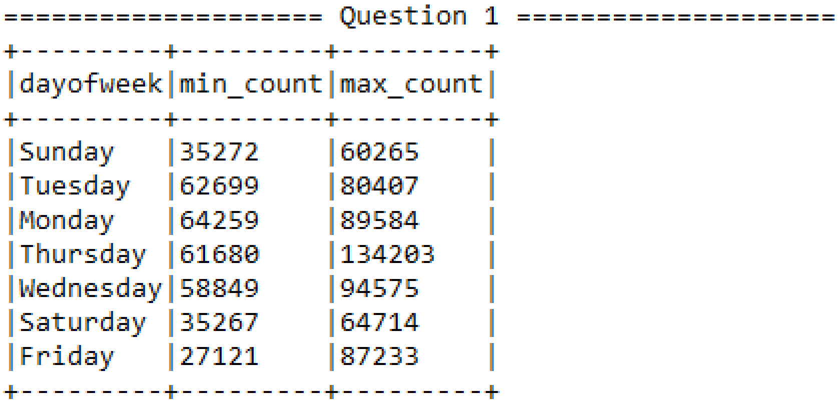
The output shows below:



```
==================== Question 1 ====================
+---+---------+---------+
|day|min_count|max_count|
+---+---------+---------+
|1  |35272    |60265    |
|2  |64259    |89584    |
|3  |62699    |80407    |
|4  |58849    |94575    |
|5  |61680    |134203   |
|6  |27121    |87233    |
|7  |35267    |64714    |
+---+---------+---------+
```

In this part of task, data should be grouped by the date and count the number of request each day first. However, the type of timestamps is not the date but the string. Therefore, timestamps need to be transformed to date type before grouping it and counting it.

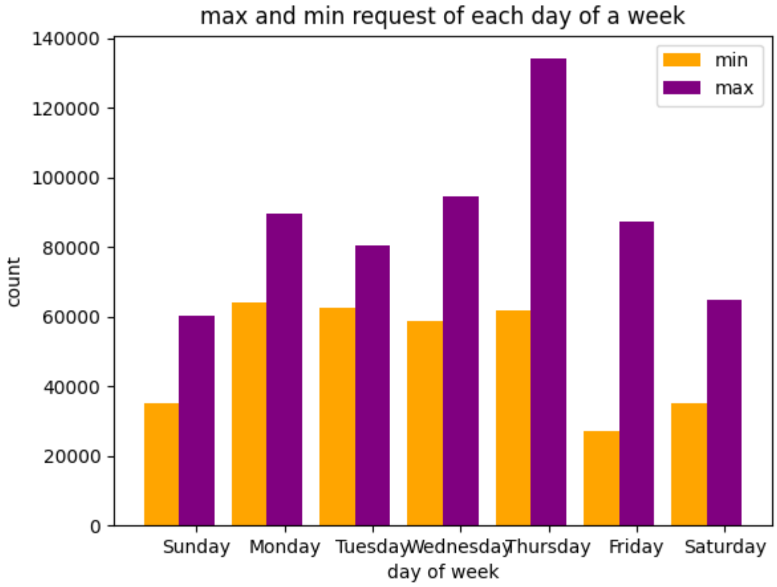
Next, using ‘dayofweek’ function to transform the date to the day of a week and order the count under day of week. As the table showed above, the day of (1,2,3,4,5,6,7) in ‘dayofweek’ function means (‘Sunday’,’Monday’,’Tuesday’,’Wednesday’,’Thursday’,’Friday’,’Saturday’).



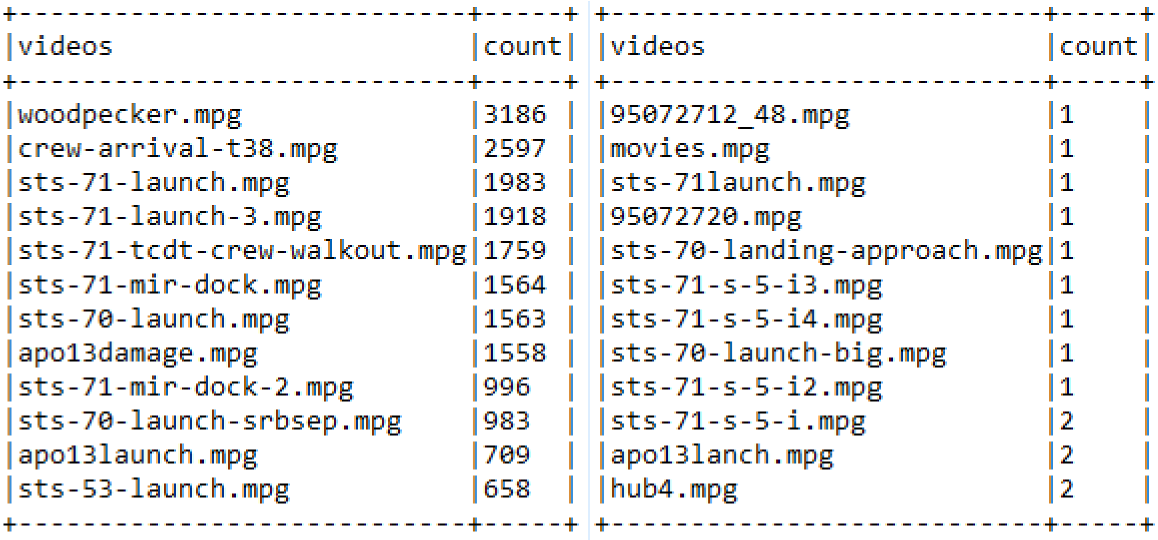
```
==================== Question 1 ====================
+---------+---------+---------+
|dayofweek|min_count|max_count|
+---------+---------+---------+
|Sunday   |35272    |60265    |
|Tuesday  |62699    |80407    |
|Monday   |64259    |89584    |
|Thursday |61680    |134203   |
|Wednesday|58849    |94575    |
|Saturday |35267    |64714    |
|Friday   |27121    |87233    |
+---------+---------+---------+
```

Finally, extract the largest and the smallest number of counts per day of week.

B. The result in a figure shows as below.



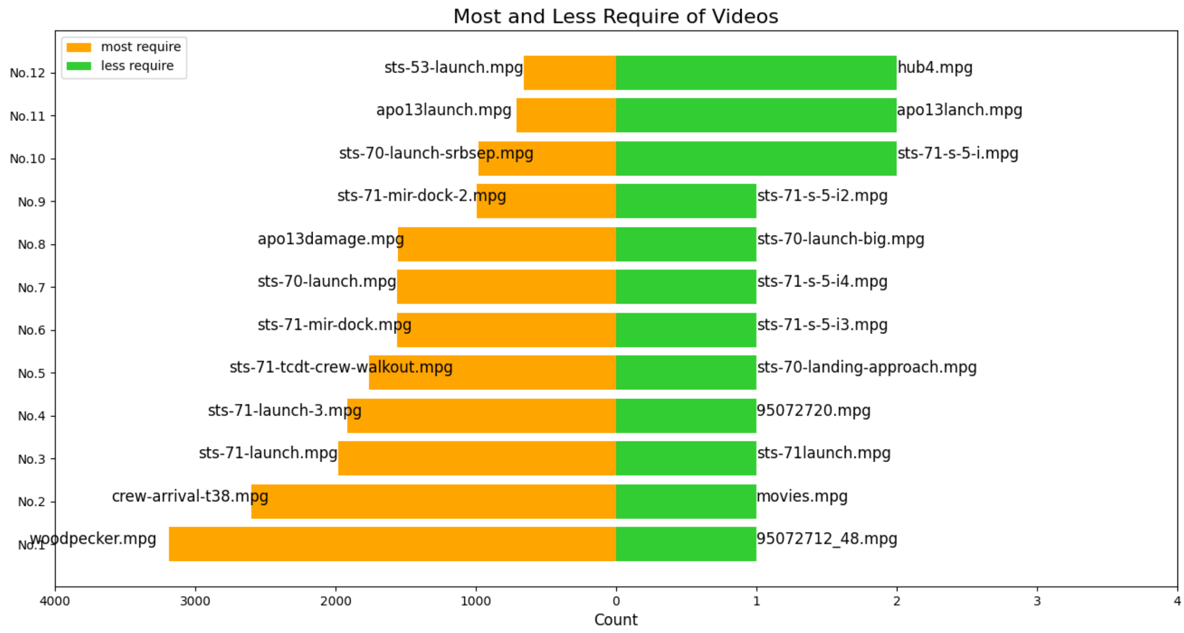
C. To find out the 12 most requested and 12 least requested .mpg videos, the .mpg videos in the request columns need to be extracted first using by '.filter()' . Then, extract the movie name just as the table showed following.



```
+----------------------------+-----+
|videos                      |count|
+----------------------------+-----+
|woodpecker.mpg              |3186 |
|crew-arrival-t38.mpg        |2597 |
|sts-71-launch.mpg           |1983 |
|sts-71-launch-3.mpg         |1918 |
|sts-71-tcdt-crew-walkout.mpg|1759 |
|sts-71-mir-dock.mpg         |1564 |
|sts-70-launch.mpg           |1563 |
|apo13damage.mpg             |1558 |
|sts-71-mir-dock-2.mpg       |996  |
|sts-70-launch-srbsep.mpg    |983  |
|apo13launch.mpg             |709  |
|sts-53-launch.mpg           |658  |
+----------------------------+-----+
```

```
+---------------------------+-----+
|videos                     |count|
+---------------------------+-----+
|95072712_48.mpg            |1    |
|movies.mpg                 |1    |
|sts-71launch.mpg           |1    |
|95072720.mpg               |1    |
|sts-70-landing-approach.mpg|1    |
|sts-71-s-5-i3.mpg          |1    |
|sts-71-s-5-i4.mpg          |1    |
|sts-70-launch-big.mpg      |1    |
|sts-71-s-5-i2.mpg          |1    |
|sts-71-s-5-i.mpg           |2    |
|apo13lanch.mpg             |2    |
|hub4.mpg                   |2    |
+---------------------------+-----+
```

D. Visualise the 24 total request numbers in one figure.



E. There are two interesting observations from the data above.

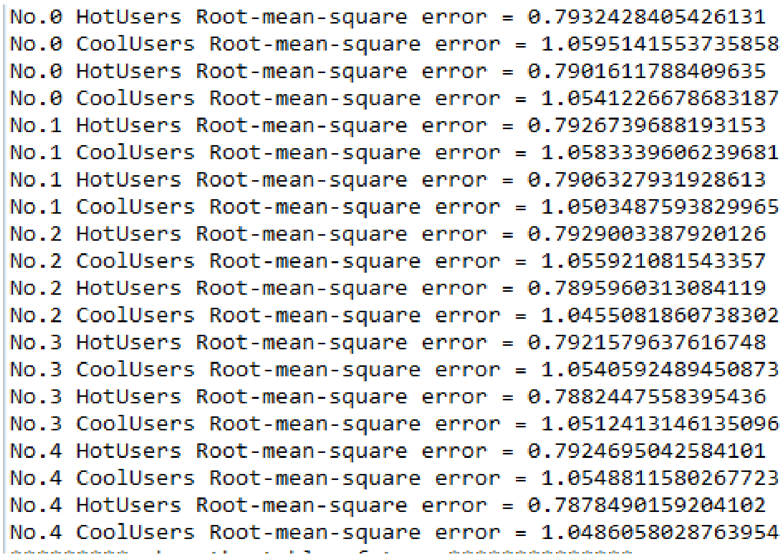
a) The number of requests is relatively lower on Friday, Saturday and Sunday. The reason may be that people are more likely to go outside to enjoy life rather than working or surfing the internet. NASA can modify the bandwidth of transformation to save the resource or increase the utilization rate of transformation technology devices.

b) The visualisation of data can make the information more directly and vividly. Because during the process of visualising result, a second collect and extraction can deliver the features of data what we want to describe. In this case, NASA can use more kinds of data visualisation to public their research results in divergent fields.

2. Movie Recommendation and Analysis

A. To use five-fold cross validation, dataset should be divided into 5 splits by ‘randomSplit()’. One of them become test data and other four splits become training data each training time. For each split, by grouping the users, the sort count of each users can be showed by ‘sort()’. The most 10% count is the most 10% active users and it can be collected by using ‘monotonically_increasing_id()’ to index the sort and extract the index at top 10% and at bottom 10%. After that, the prediction of model and the RMSE of model can be obtained behind feeding the data.

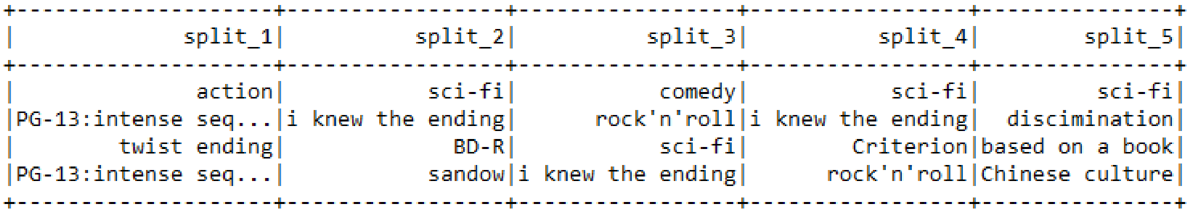
There are two version of ALS model. The first one has the same set with the lab but the second one tuning the parameters of rank and blockSize to make the model perform better(the RMSE smaller than first ALS model). The result can be seen as followed.



```
No.0 HotUsers Root-mean-square error = 0.7932428405426131
No.0 CoolUsers Root-mean-square error = 1.0595141553735858
No.0 HotUsers Root-mean-square error = 0.7901611788409635
No.0 CoolUsers Root-mean-square error = 1.0541226678683187
No.1 HotUsers Root-mean-square error = 0.7926739688193153
No.1 CoolUsers Root-mean-square error = 1.0583339606239681
No.1 HotUsers Root-mean-square error = 0.7906327931928613
No.1 CoolUsers Root-mean-square error = 1.0503487593829965
No.2 HotUsers Root-mean-square error = 0.7929003387920126
No.2 CoolUsers Root-mean-square error = 1.055921081543357
No.2 HotUsers Root-mean-square error = 0.7895960313084119
No.2 CoolUsers Root-mean-square error = 1.0455081860738302
No.3 HotUsers Root-mean-square error = 0.7921579637616748
No.3 CoolUsers Root-mean-square error = 1.0540592489450873
No.3 HotUsers Root-mean-square error = 0.7882447558395436
No.3 CoolUsers Root-mean-square error = 1.0512413146135096
No.4 HotUsers Root-mean-square error = 0.7924695042584101
No.4 CoolUsers Root-mean-square error = 1.0548811580267723
No.4 HotUsers Root-mean-square error = 0.7878490159204102
No.4 CoolUsers Root-mean-square error = 1.0486058028763954
```

Visualise the data:

B.   In this part of task, K-means can be used to classify the movies according to the factors the ALS model trained. After extracting the top 2 classes by sorting the count of each class, count the number of tags related to movies of these two classes. The top one and the bottom one can get from sorting the count of tags. The final results show as below.



```
+--------------------+-----------------+-----------------+-----------------+---------------+
|             split_1|          split_2|          split_3|          split_4|        split_5|
+--------------------+-----------------+-----------------+-----------------+---------------+
|              action|           sci-fi|           comedy|           sci-fi|         sci-fi|
|PG-13:intense seq...|i knew the ending|      rock'n'roll|i knew the ending|  discimination|
|        twist ending|             BD-R|           sci-fi|        Criterion|based on a book|
|PG-13:intense seq...|           sandow|i knew the ending|      rock'n'roll|Chinese culture|
+--------------------+-----------------+-----------------+-----------------+---------------+
```

Because the top tag and the bottom tag for each class may not be only one tag and in this project, the first element and the last element are selected.

C. There are two interesting observations about the data or the processing of data above.

1) Tuning the model parameters or the more suitable design of model can make the model perform better and get more benefits from it. Because the better model can predict the mind of costumers more precisely and get more clicks benefits. It's quite useful for Netflix to recommend the videos to the new and old costumers.

2) The most popular movies people preferred may be related with action or science fiction. Because the top tag of the largest class just like this. Netflix can recommend more this kind of movies to new costumers to earn more preference.