adybqqv8m

June 30, 2023

1 How panorama of objects has transposed ?

Kaggle Competitions and specifically technology has changed in such a way that in the data science and AI field it has become extremely useful for for prediction, trends, patterns and behavioural analysis in machine learning and other careers.

- Common kaggle competition types include :

FEATURED

RESEARCH

RECRUITMENT

ANNUAL

LIMITED PARTICIPANT

- It also includes different formats such as :

SIMLE COMPETITONS

TWO STAGE COMPETITIONS

CODE COMPETITIONS

To get started on the honest talks of how technology specifically AI has impacted the human race in the recent years is astounishingly surprising. Not so long ago these concepts found their positions in only books and academic curriculums, But with the boom of mobiles and specifically social media platforms knowledge found its way into the interests of every individual from the young to the old. AI was a topic which had been on the minds of industrialists, philanthropists and scientists working regularly in these specific domains but the recent era has turned the tables around. since the emergence of the novel Coronavirus pandemic outburst in 2019 which halted for a long period of nearly 2-3 years has changed it more dynamically.

With the human race having no choice but spending time at their homes for a long period changed the technology field in many aspects. With new innovative ideas and also solutions to common everyday tasks gained more momentum. People have started transposing the world around them in a more diversified manner and a broader mindset for living with more satisfaction and contentment. But, all this has resulted only as there was a changing of everyday habits and opening doors to new possibilities everyday just because they were forced to spend time at their homes in the pandemic

era. In short, the pandemic era can in one way or another be regarded as the way in which AI and technology has revolutionised itself majorly with lesser setbacks. Still more inventions are on the long run in the future.

Counting the possibilities it has opened is vast and the opportunities are unimaginable. Major platforms are created and with the open-source policy many websites are immerged for the improvement of individuals on technology through hosting competitions anywhere in the world without specifically competing for a particular academic satisfaction, a particular job role or a particular competition for proving yourself.

Among the many websites kaggle is one of the most prominent and evident platforms for improving your skills in the AI domain. It is extremely useful for individuals as gaining extra skills and also competiong with peers all around the world. This openness implies for vast amount of ideas emerging from individuals and generation of more innovation towards the human race and to the planet as large.

The vast amount of possibilities these competitions are making available to the users are huge with the interested people can use as a open-source platform. specifically designed for Machine Learning enthusiasts the platform motivates users for brainstorming and creating innumerable possibilities of how a problem can be solved. It encourages the users to think differently on how the solution can be updated and modified. Machine learning a domain which has been with technology since the last century but it has gained a major boom in recent years due to the advent of technology. Most credit can be given to the open-source deployment of wesites for learning and knowledge spreadness. Ml is not something which can be implemented in a go rather it takes patience and determination for the innovator to iterate repetitively for more accurate results. That is why these competitions are involved in creating a more stronger foundation in the development of skills into the minds of users.

2 FLOW OF REPORT

3 A LOOK AT THE PROGRESS THROUGHOUT YEARS

[ ]:
```
# importing python libraries necessary fro the analysis
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Colums of respective submission, dataset and competition datasets, were for necessary details onto a single modified dataset.

[ ]:
```
#reading the dataset
df = pd.read_csv('/kaggle/input/ai-report/AI report .csv')
```

```
/tmp/ipykernel_32/2335032482.py:2: DtypeWarning: Columns (4) have mixed types.
```

```
Specify dtype option on import or set low_memory=False.
  df = pd.read_csv('/kaggle/input/ai-report/AI report .csv')
```

[ ]:
```
df.isna  # missing values
```

[ ]:
```
<bound method DataFrame.isna of          TotalViews  TotalDownloads
TotalKernels  SubmittedUserId  \
0          217947.0         34575.0         433.0             NaN
1          206663.0         34392.0         409.0          3258.0
2           11493.0          1227.0          10.0          3258.0
3          158613.0         15414.0         159.0        167702.0
4          255242.0         29747.0         566.0        167702.0
…               …               …             …              …
1048570         NaN             NaN           NaN         73012.0
1048571         NaN             NaN           NaN         73012.0
1048572         NaN             NaN           NaN         73012.0
1048573         NaN             NaN           NaN         73012.0
1048574         NaN             NaN           NaN        112106.0

        HostSegmentTitle  LeaderboardPercentage  MaxDailySubmissions      Id  \
0               Featured                   10.0                  5.0  2408.0
1               Featured                   30.0                  4.0  2435.0
2               Featured                   10.0                  5.0  2438.0
3               Featured                   10.0                  5.0  2439.0
4               Featured                   10.0                  5.0  2442.0
…                    …                      …                    …      …
1048570              NaN                    NaN                  NaN     NaN
1048571              NaN                    NaN                  NaN     NaN
1048572              NaN                    NaN                  NaN     NaN
1048573              NaN                    NaN                  NaN     NaN
1048574              NaN                    NaN                  NaN     NaN

         TotalTeams
0              22.0
1             107.0
2               0.0
3             145.0
4              63.0
…               …
1048570         NaN
1048571         NaN
1048572         NaN
1048573         NaN
1048574         NaN

[1048575 rows x 9 columns]>
```

Feature Engineering

[ ]:
```
# observing the initial rows and columns
df.head()
```

[ ]:
```
   TotalViews  TotalDownloads  TotalKernels  SubmittedUserId HostSegmentTitle  \
0    217947.0         34575.0         433.0              NaN         Featured
1    206663.0         34392.0         409.0           3258.0         Featured
2     11493.0          1227.0          10.0           3258.0         Featured
3    158613.0         15414.0         159.0         167702.0         Featured
4    255242.0         29747.0         566.0         167702.0         Featured

   LeaderboardPercentage  MaxDailySubmissions      Id  TotalTeams
0                   10.0                  5.0  2408.0        22.0
1                   30.0                  4.0  2435.0       107.0
2                   10.0                  5.0  2438.0         0.0
3                   10.0                  5.0  2439.0       145.0
4                   10.0                  5.0  2442.0        63.0
```

EXPLORATORY DATA ANALYSIS

[ ]:
```
# gathering statistical information on the complexity of data
df.info
```

[ ]:
```
<bound method DataFrame.info of          TotalViews  TotalDownloads
TotalKernels  SubmittedUserId  \
0          217947.0         34575.0         433.0              NaN
1          206663.0         34392.0         409.0           3258.0
2           11493.0          1227.0          10.0           3258.0
3          158613.0         15414.0         159.0         167702.0
4          255242.0         29747.0         566.0         167702.0
…               …               …             …               …
1048570         NaN             NaN           NaN          73012.0
1048571         NaN             NaN           NaN          73012.0
1048572         NaN             NaN           NaN          73012.0
1048573         NaN             NaN           NaN          73012.0
1048574         NaN             NaN           NaN         112106.0

        HostSegmentTitle  LeaderboardPercentage  MaxDailySubmissions      Id  \
0               Featured                   10.0                  5.0  2408.0
1               Featured                   30.0                  4.0  2435.0
2               Featured                   10.0                  5.0  2438.0
3               Featured                   10.0                  5.0  2439.0
4               Featured                   10.0                  5.0  2442.0
…                    …                      …                    …      …
1048570              NaN                    NaN                  NaN     NaN
1048571              NaN                    NaN                  NaN     NaN
1048572              NaN                    NaN                  NaN     NaN
1048573              NaN                    NaN                  NaN     NaN
```

```
1048574              NaN                    NaN                  NaN     NaN

         TotalTeams
0              22.0
1             107.0
2               0.0
3             145.0
4              63.0
…              …
1048570          NaN
1048571          NaN
1048572          NaN
1048573          NaN
1048574          NaN

[1048575 rows x 9 columns]>
```

[ ]:
```
df.describe()
```

[ ]:
```
          TotalViews  TotalDownloads   TotalKernels  SubmittedUserId  \
count   2.365460e+05   236546.000000  236546.000000     1.047983e+06
mean    2.591789e+03      261.707507       2.094257     1.306306e+05
std     3.291399e+04     3920.428027      35.723069     1.068255e+05
min     0.000000e+00        0.000000       0.000000     6.200000e+01
25%     1.010000e+02        2.000000       0.000000     3.836500e+04
50%     5.990000e+02        5.000000       0.000000     1.018780e+05
75%     1.210000e+03       30.000000       1.000000     2.086320e+05
max     1.019219e+07   546098.000000    6354.000000     4.329470e+05

       LeaderboardPercentage  MaxDailySubmissions            Id   TotalTeams
count            5595.000000          5595.000000   5595.000000  5595.000000
mean               50.158177           199.993029  20331.765684   120.961037
std                29.708713         13371.679166   9527.663378   444.030306
min                 0.000000             0.000000   2408.000000     0.000000
25%                30.000000             5.000000  12848.000000     2.000000
50%                50.000000             6.000000  20571.000000    12.000000
75%                70.000000            20.000000  27713.500000    42.000000
max               100.000000       1000000.000000  52732.000000  8751.000000
```

[ ]:
```
df.duplicated().sum()
```

[ ]:
```
780274
```

[ ]:
```
display(df.drop_duplicates())
```

```
         TotalViews  TotalDownloads  TotalKernels  SubmittedUserId  \
0          217947.0         34575.0         433.0              NaN
```

```
1          206663.0         34392.0         409.0           3258.0
2           11493.0          1227.0          10.0           3258.0
3          158613.0         15414.0         159.0          167702.0
4          255242.0         29747.0         566.0          167702.0
…               …               …             …               …
1048493         NaN              NaN           NaN         275388.0
1048496         NaN              NaN           NaN         285960.0
1048497         NaN              NaN           NaN         286127.0
1048506         NaN              NaN           NaN         304086.0
1048530         NaN              NaN           NaN         273848.0

        HostSegmentTitle  LeaderboardPercentage  MaxDailySubmissions      Id  \
0               Featured                   10.0                  5.0  2408.0
1               Featured                   30.0                  4.0  2435.0
2               Featured                   10.0                  5.0  2438.0
3               Featured                   10.0                  5.0  2439.0
4               Featured                   10.0                  5.0  2442.0
…                    …                      …                    …      …
1048493              NaN                    NaN                  NaN     NaN
1048496              NaN                    NaN                  NaN     NaN
1048497              NaN                    NaN                  NaN     NaN
1048506              NaN                    NaN                  NaN     NaN
1048530              NaN                    NaN                  NaN     NaN

         TotalTeams
0              22.0
1             107.0
2               0.0
3             145.0
4              63.0
…               …
1048493         NaN
1048496         NaN
1048497         NaN
1048506         NaN
1048530         NaN

[268301 rows x 9 columns]
```

[ ]:
```
# analysing unique values
df['HostSegmentTitle'].unique()
```

[ ]:
```
array(['Featured', 'Community', 'Research', 'Prospect', 'Recruitment',
       'GE Quests', 'Getting Started', 'Playground', nan], dtype=object)
```

[ ]:
```
df['LeaderboardPercentage'].unique()
```

[ ]:
```
array([ 10.,  30.,  25.,  20.,  62.,  50.,  40.,  53.,  42.,   2.,   0.,
        35.,   9.,  37.,  27.,  39., 100.,  46.,  49.,  41.,  32.,  33.,
        34.,  43.,   3.,   7.,  15.,  18.,  29.,  70.,   1.,  13.,  47.,
        52.,  19.,  17.,   4.,  68.,   6.,  51.,  60.,  48.,   8.,  80.,
        12.,  24.,  66.,  97.,  28.,  31.,  45.,  36.,  22.,  59.,   5.,
        76.,  75.,  67.,  90.,  89.,  63.,  57.,  55.,  64.,  79.,  71.,
        99.,  44.,  38.,  14.,  58.,  95.,  74.,  69.,  54.,  65.,  91.,
        84.,  56.,  21.,  16.,  73.,  11.,  85.,  26.,  23.,  78.,  94.,
        98.,  92.,  nan])
```

[ ]:
```
df.isnull().sum()
```

[ ]:
```
TotalViews               812029
TotalDownloads           812029
TotalKernels             812029
SubmittedUserId             592
HostSegmentTitle        1042980
LeaderboardPercentage   1042980
MaxDailySubmissions     1042980
Id                      1042980
TotalTeams              1042980
dtype: int64
```

[ ]:
```
df.replace(np.nan,'0',inplace = True)

#Check the changes now
df.isnull().sum()
```

[ ]:
```
TotalViews              0
TotalDownloads          0
TotalKernels            0
SubmittedUserId         0
HostSegmentTitle        0
LeaderboardPercentage   0
MaxDailySubmissions     0
Id                      0
TotalTeams              0
dtype: int64
```

[ ]:
```
df.dtypes
```

[ ]:
```
TotalViews              object
TotalDownloads          object
TotalKernels            object
SubmittedUserId         object
HostSegmentTitle        object
LeaderboardPercentage   object
```

```
MaxDailySubmissions      object
Id                       object
TotalTeams               object
dtype: object
```

[ ]:
```
df.corr
```

[ ]:
```
<bound method DataFrame.corr of         TotalViews TotalDownloads TotalKernels
SubmittedUserId  \
0         217947.0        34575.0        433.0               0
1         206663.0        34392.0        409.0          3258.0
2          11493.0         1227.0         10.0          3258.0
3         158613.0        15414.0        159.0        167702.0
4         255242.0        29747.0        566.0        167702.0
…              …              …            …              …
1048570          0              0            0         73012.0
1048571          0              0            0         73012.0
1048572          0              0            0         73012.0
1048573          0              0            0         73012.0
1048574          0              0            0        112106.0

        HostSegmentTitle LeaderboardPercentage MaxDailySubmissions      Id  \
0               Featured                 10.0                 5.0  2408.0
1               Featured                 30.0                 4.0  2435.0
2               Featured                 10.0                 5.0  2438.0
3               Featured                 10.0                 5.0  2439.0
4               Featured                 10.0                 5.0  2442.0
…                  …                      …                   …      …
1048570                0                    0                   0       0
1048571                0                    0                   0       0
1048572                0                    0                   0       0
1048573                0                    0                   0       0
1048574                0                    0                   0       0

        TotalTeams
0             22.0
1            107.0
2              0.0
3            145.0
4             63.0
…              …
1048570          0
1048571          0
1048572          0
1048573          0
1048574          0
```

```
[1048575 rows x 9 columns]>
```

FEATURE ENGINEERING

```
[ ]: # displayint the available features
print('Features:', df)
```

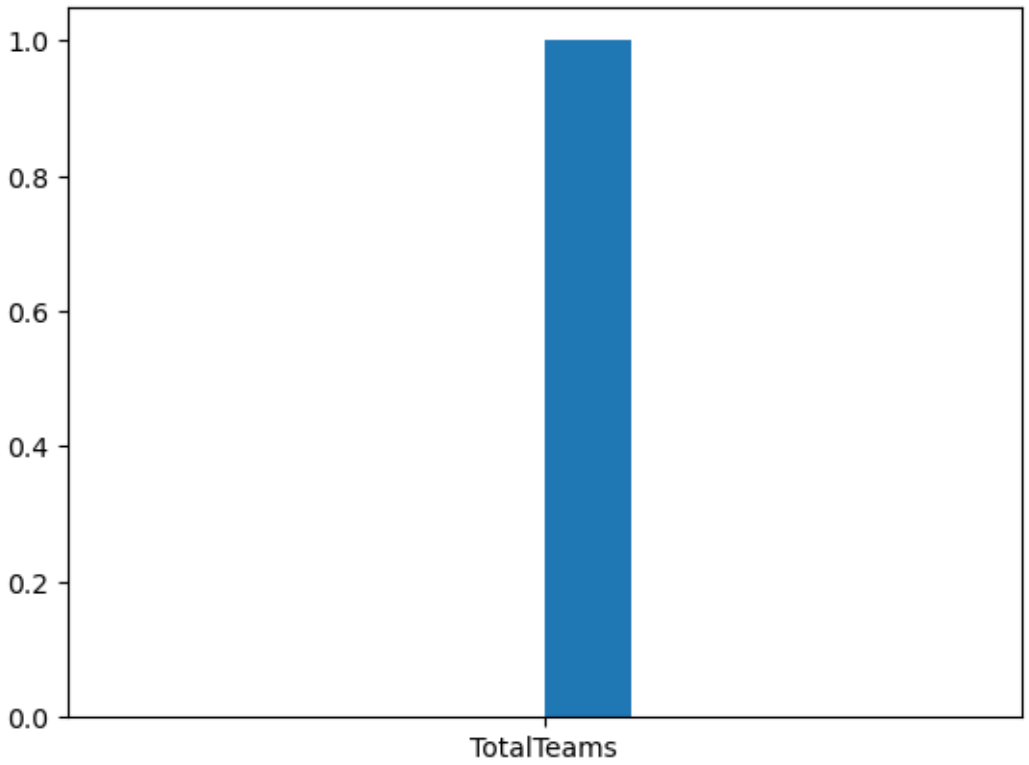
```
Features:         TotalViews TotalDownloads TotalKernels SubmittedUserId  \
0         217947.0        34575.0        433.0               0
1         206663.0        34392.0        409.0          3258.0
2          11493.0         1227.0         10.0          3258.0
3         158613.0        15414.0        159.0        167702.0
4         255242.0        29747.0        566.0        167702.0
…              …              …              …              …
1048570          0              0              0         73012.0
1048571          0              0              0         73012.0
1048572          0              0              0         73012.0
1048573          0              0              0         73012.0
1048574          0              0              0        112106.0

        HostSegmentTitle LeaderboardPercentage MaxDailySubmissions      Id  \
0               Featured                  10.0                 5.0  2408.0
1               Featured                  30.0                 4.0  2435.0
2               Featured                  10.0                 5.0  2438.0
3               Featured                  10.0                 5.0  2439.0
4               Featured                  10.0                 5.0  2442.0
…                  …                      …                    …      …
1048570                0                     0                    0       0
1048571                0                     0                    0       0
1048572                0                     0                    0       0
1048573                0                     0                    0       0
1048574                0                     0                    0       0

        TotalTeams
0             22.0
1            107.0
2              0.0
3            145.0
4             63.0
…              …
1048570          0
1048571          0
1048572          0
1048573          0
1048574          0

[1048575 rows x 9 columns]
```

[ ]:
```
df.shape  #dimensionality rows x columns
```

[ ]:
```
(1048575, 9)
```

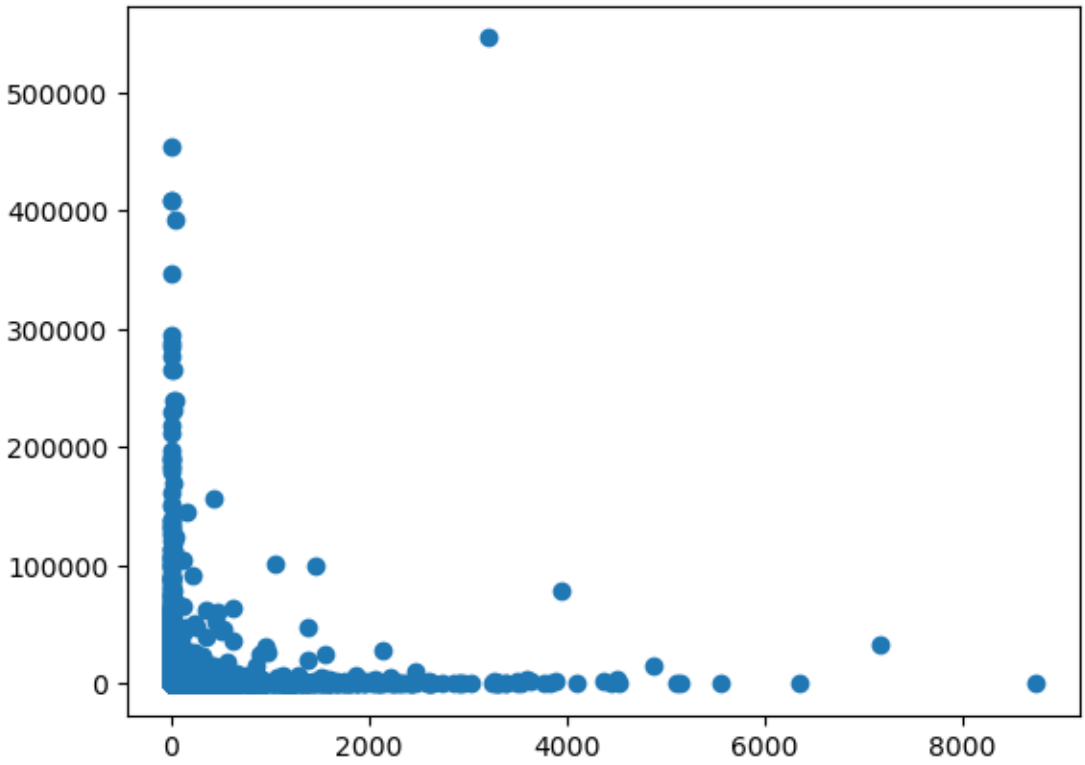
[ ]:
```
# plotting histogram of all the columns
plt.hist(x='TotalTeams')
```

[ ]:
```
(array([0., 0., 0., 0., 0., 1., 0., 0., 0., 0.]),
 array([-0.5, -0.4, -0.3, -0.2, -0.1,  0. ,  0.1,  0.2,  0.3,  0.4,  0.5]),
 <BarContainer object of 10 artists>)
```



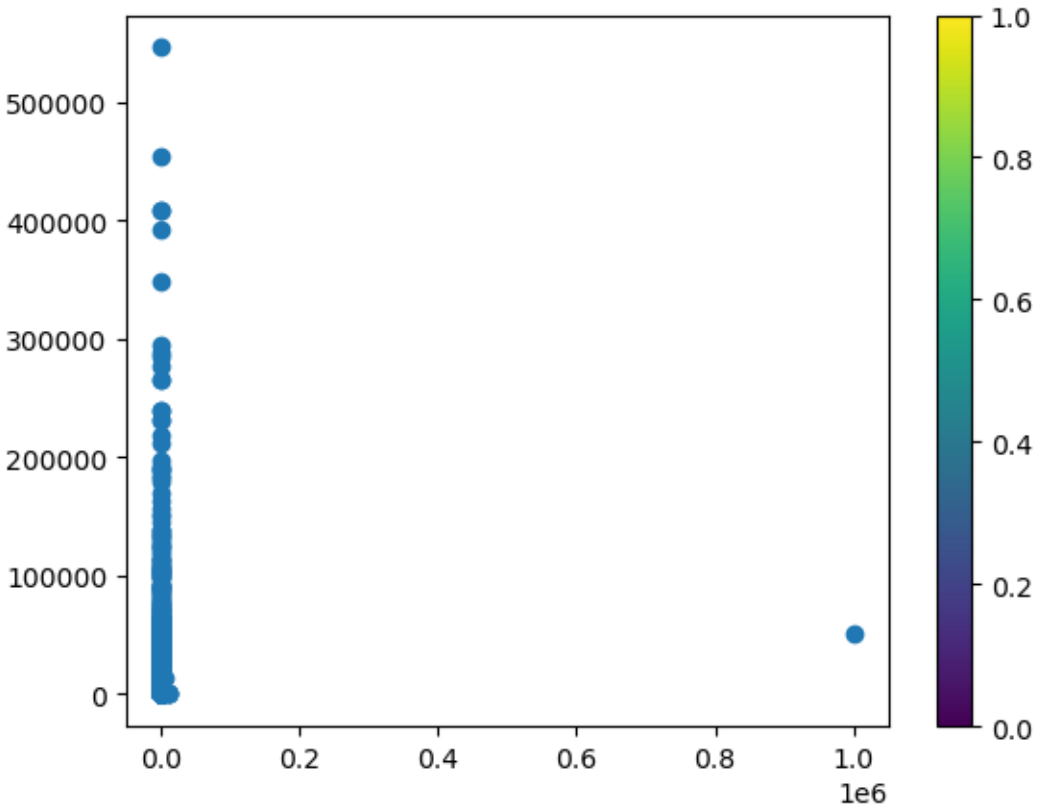
[ ]:
```
plt.scatter(df['TotalTeams'], df['TotalDownloads'])
```

[ ]:
```
<matplotlib.collections.PathCollection at 0x7874d9131210>
```



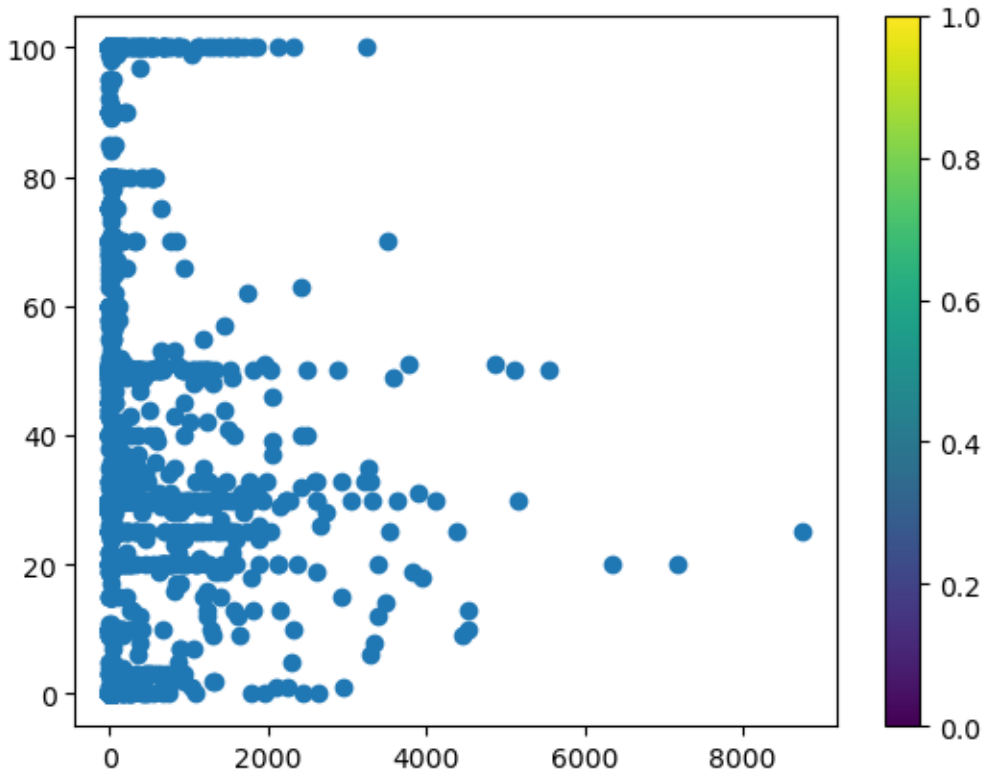
[ ]:
```
plt.scatter(df['MaxDailySubmissions'], df['TotalDownloads'])
plt.colorbar()
```

[ ]:
```
<matplotlib.colorbar.Colorbar at 0x7874eb13d840>
```

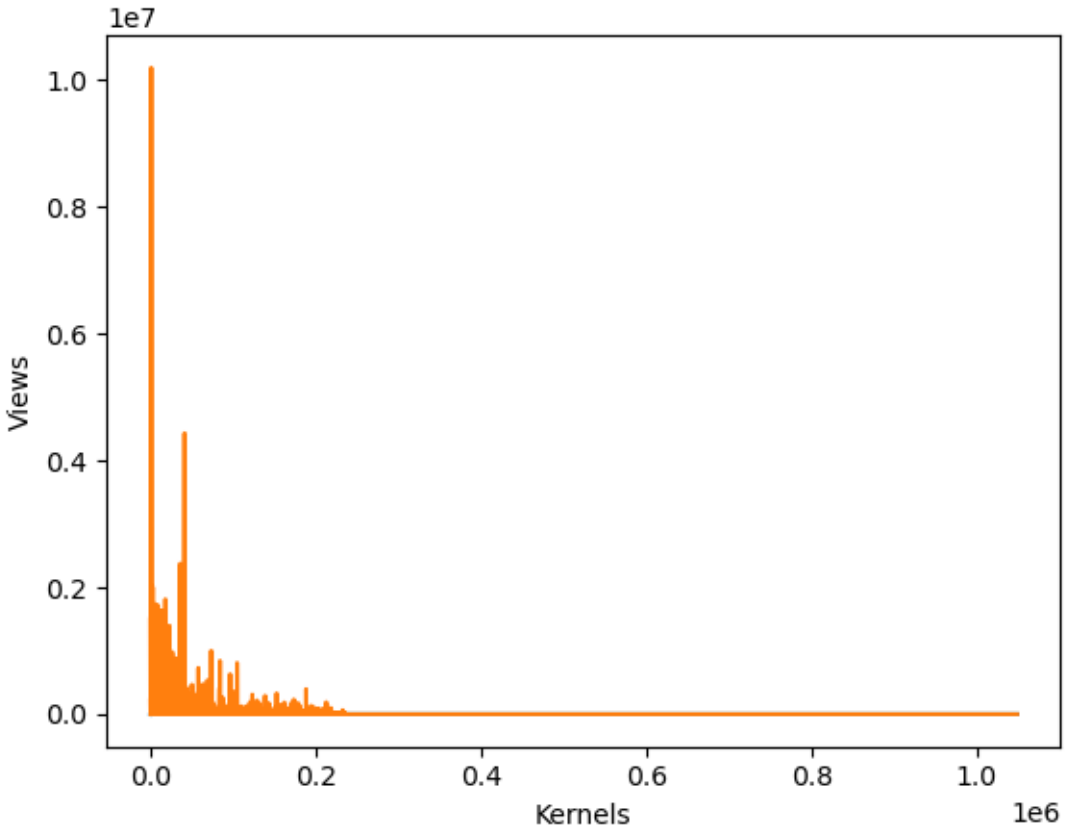


[ ]:
```
plt.scatter(df['TotalTeams'], df['LeaderboardPercentage'])
plt.colorbar()
```

[ ]: <matplotlib.colorbar.Colorbar at 0x7874eb3cabc0>



[ ]:
```
plt.plot(df['TotalKernels'])
plt.plot(df['TotalViews'])

plt.xlabel('Kernels')
plt.ylabel('Views')
plt.show()
```



DATA EVALUATION

[ ]:
```
from sklearn.metrics import confusion_matrix
correlation_metrics = df.corr()
```

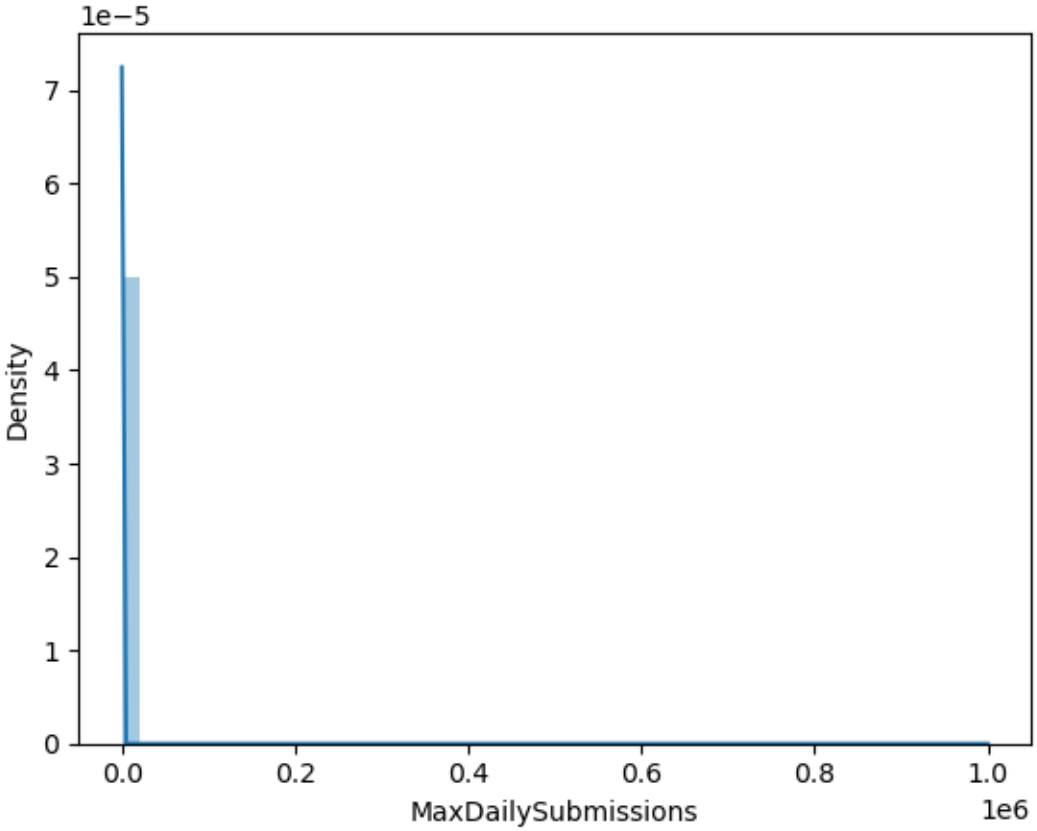
```
/tmp/ipykernel_32/2387876587.py:2: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  correlation_metrics = df.corr()
```

[ ]:
```
sns.distplot(df['MaxDailySubmissions']);
```

```
/tmp/ipykernel_32/2332233380.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
```

https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

```
sns.distplot(df['MaxDailySubmissions']);
```



above plot describes the increasing number of daily submissions on the platform.

Conclusionary note overviews the Increased number of users on the competitions platfom which is broadening every-day. It also describes trends and patterns associated with the analysis. Kaggle AI report 2023 analyses how the competiton has developed its users throughout the years and gaining more positive outcomes as the field continues to grow continually. The innovation and descriptiveness of the platform enables everyone to be on the race, ensuring diverse habits of discipline, dedication, determination, improvement, updation and succesc which in par strenghtens to the core.