kaggle final copy new

May 7, 2023

[1]:
```
import os
os.getcwd()
```

[1]:
```
'C:\\Users\\96209\\Kaggle'
```

[2]:
```
# Import libraries
import numpy as np
import pandas as pd
import sklearn as sk
import tarfile
import urllib
import time
from matplotlib import pyplot as plt
```

[3]:
```
import io
student_training = pd.read_csv('training_unit_test_scores.csv')
student_predict = pd.read_csv('evaluation_unit_test_scores.csv')
action_logs = pd.read_csv('action_logs.csv')
problem_details = pd.read_csv('problem_details.csv')
assignment_relationships = pd.read_csv('assignment_relationships.csv')
```

[4]:
```
merged_data = action_logs.merge(problem_details, on='problem_id', how='left')
# delet the row that problem_id is NaN
merged_data = merged_data.dropna(subset=['problem_id'])
```

[5]:
```
#check data EDA
```

[6]:
```
merged_data.head()
```

[6]:
```
  assignment_log_id     timestamp problem_id  max_attempts  \
1        2QV1F2GSBZ  1.599151e+09  I2GX4OQIE           3.0
2        2QV1F2GSBZ  1.599151e+09  I2GX4OQIE           NaN
3        2QV1F2GSBZ  1.599151e+09  I2GX4OQIE           NaN
4        2QV1F2GSBZ  1.599151e+09  I2GX4OQIE           NaN
5        2QV1F2GSBZ  1.599151e+09  I2GX4OQIE           NaN

  available_core_tutoring  score_viewable  continuous_score_viewable  \
1                  answer             1.0                        1.0
```

```
2                     NaN             NaN                        NaN
3                     NaN             NaN                        NaN
4                     NaN             NaN                        NaN
5                     NaN             NaN                        NaN

             action hint_id explanation_id problem_multipart_id  \
1   problem_started     NaN            NaN            PBZ9XQNTO
2    wrong_response     NaN            NaN            PBZ9XQNTO
3    wrong_response     NaN            NaN            PBZ9XQNTO
4  answer_requested     NaN            NaN            PBZ9XQNTO
5  correct_response     NaN            NaN            PBZ9XQNTO

   problem_multipart_position problem_type problem_skill_code  \
1                         1.0       Number          4.NBT.A.3
2                         1.0       Number          4.NBT.A.3
3                         1.0       Number          4.NBT.A.3
4                         1.0       Number          4.NBT.A.3
5                         1.0       Number          4.NBT.A.3

  problem_skill_description  problem_contains_image  \
1    Rounding Whole Numbers                     0.0
2    Rounding Whole Numbers                     0.0
3    Rounding Whole Numbers                     0.0
4    Rounding Whole Numbers                     0.0
5    Rounding Whole Numbers                     0.0

   problem_contains_equation  problem_contains_video  \
1                        0.0                     0.0
2                        0.0                     0.0
3                        0.0                     0.0
4                        0.0                     0.0
5                        0.0                     0.0

                            problem_text_bert_pca
1  [2.40100389,-0.85778539,-2.24408353,2.11064423…
2  [2.40100389,-0.85778539,-2.24408353,2.11064423…
3  [2.40100389,-0.85778539,-2.24408353,2.11064423…
4  [2.40100389,-0.85778539,-2.24408353,2.11064423…
5  [2.40100389,-0.85778539,-2.24408353,2.11064423…
```

```
[7]: merged_data.tail()
```

```
[7]:         assignment_log_id     timestamp  problem_id  max_attempts  \
     23932269       1VVEB3EAGF  1.634919e+09   1QXH2HRDZ           NaN
     23932270       1VVEB3EAGF  1.634919e+09   1QXH2HRDZ           NaN
     23932272       1VVEB3EAGF  1.634919e+09  20GKDUW4FH           1.0
     23932273       1VVEB3EAGF  1.634919e+09  20GKDUW4FH           NaN
```

```
23932274        1VVEB3EAGF  1.634919e+09  20GKDUW4FH           NaN

         available_core_tutoring  score_viewable  continuous_score_viewable  \
23932269                     NaN             NaN                        NaN
23932270                     NaN             NaN                        NaN
23932272             no_tutoring             1.0                        0.0
23932273                     NaN             NaN                        NaN
23932274                     NaN             NaN                        NaN

                    action hint_id explanation_id problem_multipart_id  \
23932269     open_response     NaN            NaN           18YSMZP42U
23932270  problem_finished     NaN            NaN           18YSMZP42U
23932272   problem_started     NaN            NaN           18YSMZP42U
23932273     open_response     NaN            NaN           18YSMZP42U
23932274  problem_finished     NaN            NaN           18YSMZP42U

          problem_multipart_position            problem_type  \
23932269                         3.0  Ungraded Open Response
23932270                         3.0  Ungraded Open Response
23932272                         4.0  Ungraded Open Response
23932273                         4.0  Ungraded Open Response
23932274                         4.0  Ungraded Open Response

         problem_skill_code       problem_skill_description  \
23932269          6.RP.A.3a  Making Equivalent Ratio Tables
23932270          6.RP.A.3a  Making Equivalent Ratio Tables
23932272          6.RP.A.3a  Making Equivalent Ratio Tables
23932273          6.RP.A.3a  Making Equivalent Ratio Tables
23932274          6.RP.A.3a  Making Equivalent Ratio Tables

          problem_contains_image  problem_contains_equation  \
23932269                     1.0                        0.0
23932270                     1.0                        0.0
23932272                     0.0                        0.0
23932273                     0.0                        0.0
23932274                     0.0                        0.0

          problem_contains_video  \
23932269                     0.0
23932270                     0.0
23932272                     0.0
23932273                     0.0
23932274                     0.0

                                            problem_text_bert_pca
23932269  [2.33184626,-4.66572058,-1.28410351,-0.2974429…
23932270  [2.33184626,-4.66572058,-1.28410351,-0.2974429…
```

```
23932272  [-7.42471425,5.01689265,-6.45013021,-1.7648785…
23932273  [-7.42471425,5.01689265,-6.45013021,-1.7648785…
23932274  [-7.42471425,5.01689265,-6.45013021,-1.7648785…
```

[8]:

```
merged_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17795561 entries, 1 to 23932274
Data columns (total 19 columns):
 #   Column                        Dtype
---  ------                        -----
 0   assignment_log_id             object
 1   timestamp                     float64
 2   problem_id                    object
 3   max_attempts                  float64
 4   available_core_tutoring       object
 5   score_viewable                float64
 6   continuous_score_viewable     float64
 7   action                        object
 8   hint_id                       object
 9   explanation_id                object
 10  problem_multipart_id          object
 11  problem_multipart_position    float64
 12  problem_type                  object
 13  problem_skill_code            object
 14  problem_skill_description     object
 15  problem_contains_image        float64
 16  problem_contains_equation     float64
 17  problem_contains_video        float64
 18  problem_text_bert_pca         object
dtypes: float64(8), object(11)
memory usage: 2.7+ GB
```

[9]:

```
merged_data["available_core_tutoring"].value_counts()
```

[9]:

```
no_tutoring    1932039
hint           1743910
answer         1429320
explanation     140591
Name: available_core_tutoring, dtype: int64
```

[10]:

```
merged_data.describe()
```

[10]:

```
          timestamp  max_attempts  score_viewable  continuous_score_viewable  \
count  1.779556e+07  5.245860e+06    5.245860e+06               5.245860e+06
mean   1.611269e+09  2.382467e+00    9.907491e-01               6.912337e-01
std    2.070955e+07  9.239691e-01    9.573576e-02               4.619846e-01
min    1.551122e+09  1.000000e+00    0.000000e+00               0.000000e+00
```

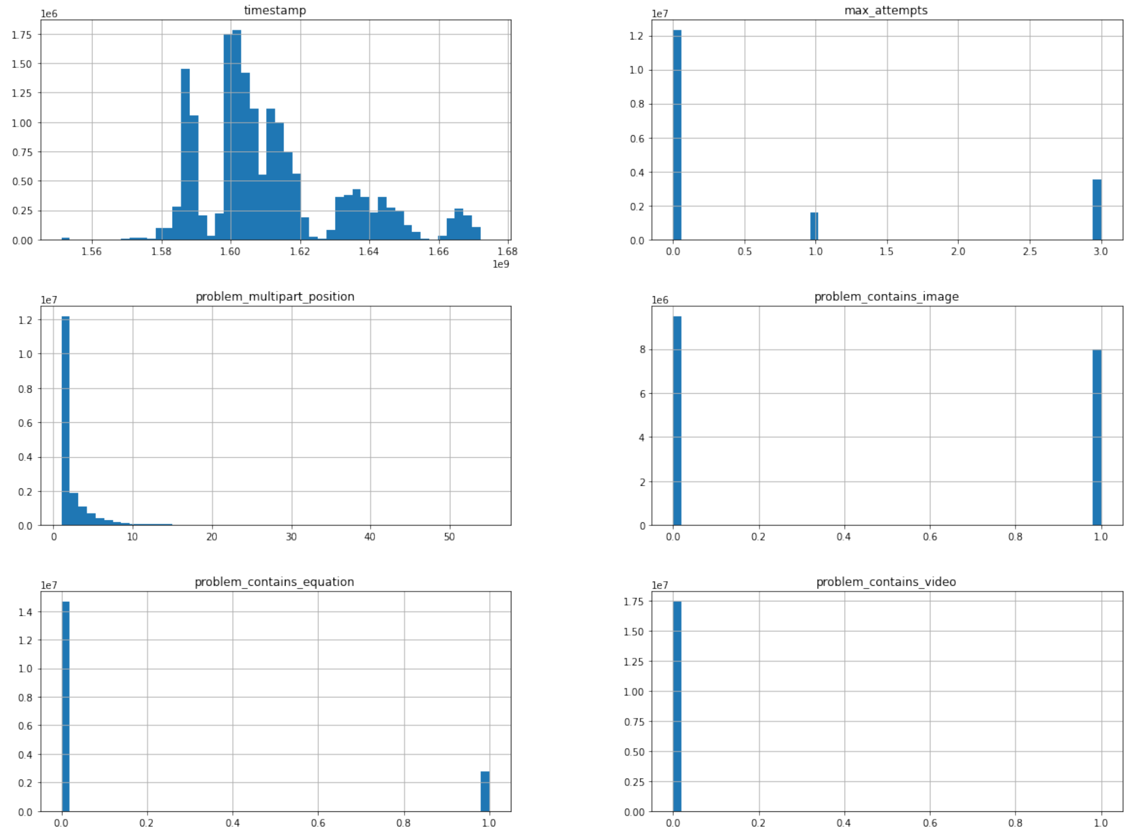
```
25%    1.599498e+09  1.000000e+00    1.000000e+00               0.000000e+00
50%    1.605724e+09  3.000000e+00    1.000000e+00               1.000000e+00
75%    1.618005e+09  3.000000e+00    1.000000e+00               1.000000e+00
max    1.674582e+09  3.000000e+00    1.000000e+00               1.000000e+00

       problem_multipart_position  problem_contains_image  \
count                1.745256e+07            1.745246e+07
mean                 2.568773e+00            4.561165e-01
std                  2.893483e+00            4.980705e-01
min                  1.000000e+00            0.000000e+00
25%                  1.000000e+00            0.000000e+00
50%                  2.000000e+00            0.000000e+00
75%                  3.000000e+00            1.000000e+00
max                  5.500000e+01            1.000000e+00

       problem_contains_equation  problem_contains_video
count               1.745246e+07            1.745246e+07
mean                1.603950e-01            3.472291e-05
std                 3.669720e-01            5.892513e-03
min                 0.000000e+00            0.000000e+00
25%                 0.000000e+00            0.000000e+00
50%                 0.000000e+00            0.000000e+00
75%                 0.000000e+00            0.000000e+00
max                 1.000000e+00            1.000000e+00
```

[11]:

```python
#bar plot for categorical variable
from matplotlib import pyplot as plt
```

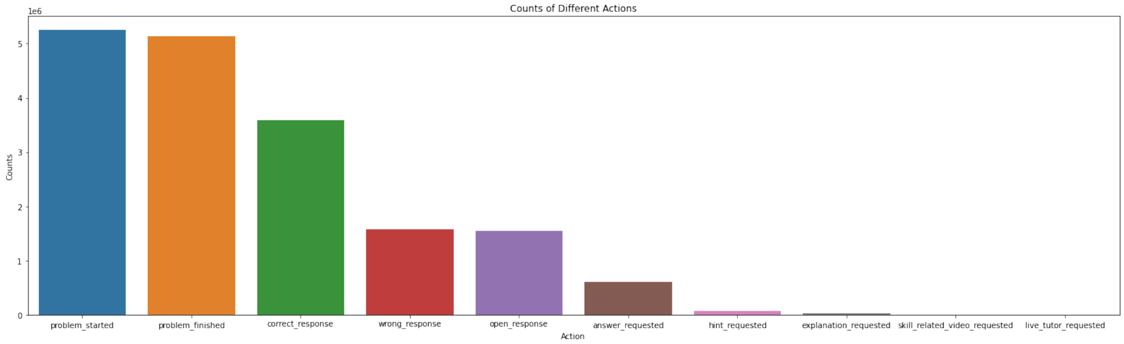
[75]:

```python
merged_data.hist(bins=50,figsize=(20,15))
plt.show()
```



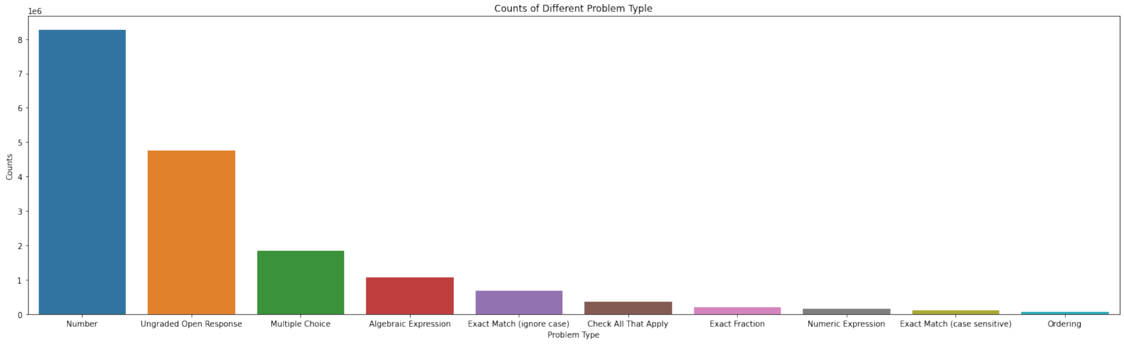
[13]:

```
#show the details about Action
import seaborn as sns
#count variable
action_counts = merged_data['action'].value_counts()

#create his variable count
plt.figure(figsize=(25,7))#set the picture size
sns.barplot(x=action_counts.index, y=action_counts.values)

#set his title and xy
plt.title('Counts of Different Actions')
plt.xlabel('Action')
plt.ylabel('Counts')

#show plt
plt.show()
```



[14]:

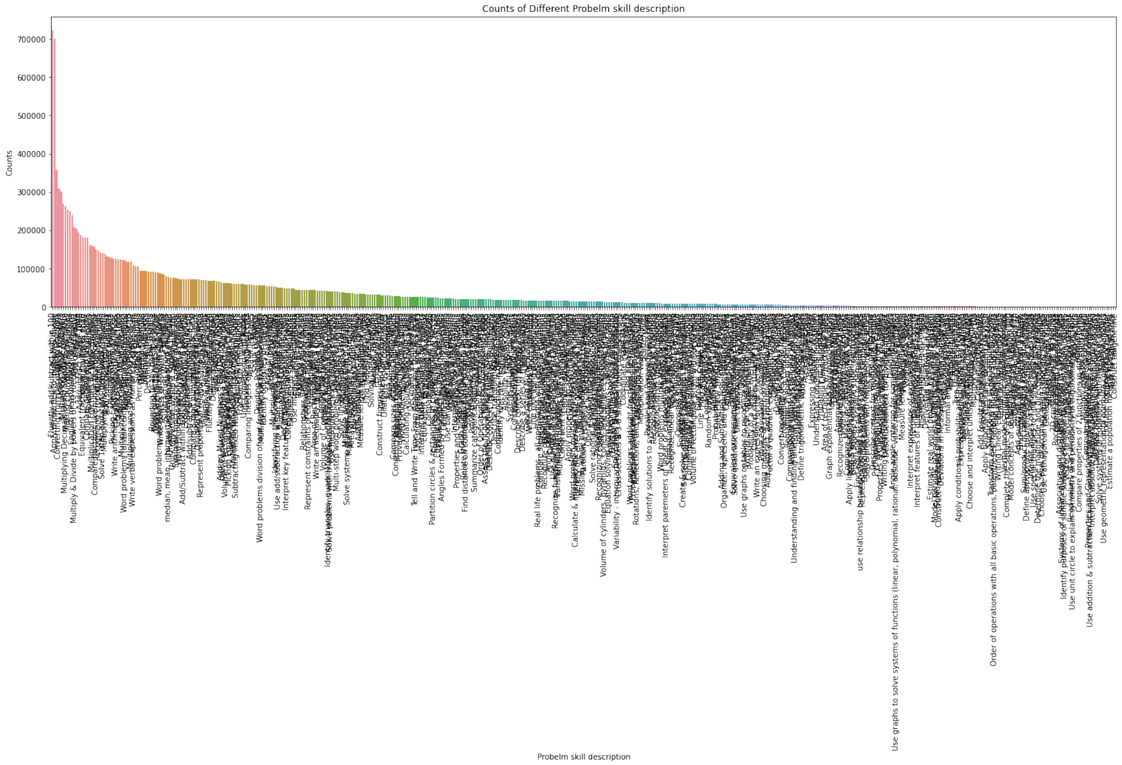
```
#show the details about Probelm_type

#count variable
problem_counts = merged_data['problem_type'].value_counts()

#create his variable count
plt.figure(figsize=(25,7))#set the picture size
sns.barplot(x=problem_counts.index, y=problem_counts.values)

#set his title and xy
plt.title('Counts of Different Problem Typle')
plt.xlabel('Problem Type')
plt.ylabel('Counts')

#show plt
plt.show()
```



[15]:

```
#show the details about Probelm_skill_description


#count variable
```

```
probelm_skill_description_counts = merged_data['problem_skill_description'].
 ↪value_counts()

#create his variable count
plt.figure(figsize=(25,7))#set the picture size
sns.barplot(x=probelm_skill_description_counts.index,␣
 ↪y=probelm_skill_description_counts.values)

#set his title and xy
plt.title('Counts of Different Probelm skill description')
plt.xlabel('Probelm skill description')
plt.ylabel('Counts')

#turn x asix
plt.xticks(rotation=90)

#show plt
plt.show()
```



[16]:

```
#drop variable which is useless in the project
#we do not have to use hint_detail.csv, explanation_detail.csv
#problem_text_bert_pca is not useful
```

```
#we already keeped probelm_skill description
merged_data.drop(columns = ['hint_id'], inplace=True)
merged_data.drop(columns = ['explanation_id'], inplace=True)
merged_data.drop(columns = ['problem_text_bert_pca'], inplace=True)
merged_data.drop(columns = ['problem_skill_code'], inplace=True)
merged_data.head()
```

[16]:

```
  assignment_log_id     timestamp problem_id  max_attempts  \
1        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           3.0
2        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN
3        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN
4        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN
5        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN

  available_core_tutoring  score_viewable  continuous_score_viewable  \
1                  answer             1.0                        1.0
2                     NaN             NaN                        NaN
3                     NaN             NaN                        NaN
4                     NaN             NaN                        NaN
5                     NaN             NaN                        NaN

             action problem_multipart_id  problem_multipart_position  \
1   problem_started            PBZ9XQNT0                         1.0
2    wrong_response            PBZ9XQNT0                         1.0
3    wrong_response            PBZ9XQNT0                         1.0
4  answer_requested            PBZ9XQNT0                         1.0
5  correct_response            PBZ9XQNT0                         1.0

  problem_type problem_skill_description  problem_contains_image  \
1       Number    Rounding Whole Numbers                     0.0
2       Number    Rounding Whole Numbers                     0.0
3       Number    Rounding Whole Numbers                     0.0
4       Number    Rounding Whole Numbers                     0.0
5       Number    Rounding Whole Numbers                     0.0

   problem_contains_equation  problem_contains_video
1                        0.0                     0.0
2                        0.0                     0.0
3                        0.0                     0.0
4                        0.0                     0.0
5                        0.0                     0.0
```

[17]:

```
#check NA value for available_core_tutoring
merged_data['available_core_tutoring'].isnull().sum()*100/
 ↪len(merged_data['available_core_tutoring'])
```

[17]:

```
70.5215250027802
```

[18]:

```
#check NA value for problem_id
merged_data['problem_id'].isnull().sum()
```

[18]:

```
0
```

[19]:

```
#drop the column with high NA value(over 50%)
merged_data.drop(columns = ['available_core_tutoring'],inplace=True)
merged_data.head()
```

[19]:

```
  assignment_log_id     timestamp problem_id  max_attempts  score_viewable  \
1        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           3.0             1.0
2        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN             NaN
3        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN             NaN
4        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN             NaN
5        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           NaN             NaN

   continuous_score_viewable            action problem_multipart_id  \
1                        1.0   problem_started            PBZ9XQNT0
2                        NaN    wrong_response            PBZ9XQNT0
3                        NaN    wrong_response            PBZ9XQNT0
4                        NaN  answer_requested            PBZ9XQNT0
5                        NaN  correct_response            PBZ9XQNT0

   problem_multipart_position problem_type problem_skill_description  \
1                         1.0       Number    Rounding Whole Numbers
2                         1.0       Number    Rounding Whole Numbers
3                         1.0       Number    Rounding Whole Numbers
4                         1.0       Number    Rounding Whole Numbers
5                         1.0       Number    Rounding Whole Numbers

   problem_contains_image  problem_contains_equation  problem_contains_video
1                     0.0                        0.0                     0.0
2                     0.0                        0.0                     0.0
3                     0.0                        0.0                     0.0
4                     0.0                        0.0                     0.0
5                     0.0                        0.0                     0.0
```

[20]:

```
#replace NA value for columns max_attempts
merged_data["max_attempts"] = merged_data["max_attempts"].fillna(0)
print(merged_data)
```

```
         assignment_log_id     timestamp  problem_id  max_attempts  \
1               2QV1F2GSBZ  1.599151e+09   I2GX40QIE           3.0
2               2QV1F2GSBZ  1.599151e+09   I2GX40QIE           0.0
3               2QV1F2GSBZ  1.599151e+09   I2GX40QIE           0.0
4               2QV1F2GSBZ  1.599151e+09   I2GX40QIE           0.0
5               2QV1F2GSBZ  1.599151e+09   I2GX40QIE           0.0
…                    …           …         …           …
```

```
23932269        1VVEB3EAGF  1.634919e+09   1QXH2HRDZ           0.0
23932270        1VVEB3EAGF  1.634919e+09   1QXH2HRDZ           0.0
23932272        1VVEB3EAGF  1.634919e+09  20GKDUW4FH           1.0
23932273        1VVEB3EAGF  1.634919e+09  20GKDUW4FH           0.0
23932274        1VVEB3EAGF  1.634919e+09  20GKDUW4FH           0.0

          score_viewable  continuous_score_viewable            action  \
1                    1.0                        1.0   problem_started
2                    NaN                        NaN    wrong_response
3                    NaN                        NaN    wrong_response
4                    NaN                        NaN  answer_requested
5                    NaN                        NaN  correct_response
…                  …                        …               …
23932269             NaN                        NaN     open_response
23932270             NaN                        NaN  problem_finished
23932272             1.0                        0.0   problem_started
23932273             NaN                        NaN     open_response
23932274             NaN                        NaN  problem_finished

         problem_multipart_id  problem_multipart_position  \
1                   PBZ9XQNT0                         1.0
2                   PBZ9XQNT0                         1.0
3                   PBZ9XQNT0                         1.0
4                   PBZ9XQNT0                         1.0
5                   PBZ9XQNT0                         1.0
…                       …                         …
23932269           18YSMZP42U                         3.0
23932270           18YSMZP42U                         3.0
23932272           18YSMZP42U                         4.0
23932273           18YSMZP42U                         4.0
23932274           18YSMZP42U                         4.0

                    problem_type       problem_skill_description  \
1                         Number          Rounding Whole Numbers
2                         Number          Rounding Whole Numbers
3                         Number          Rounding Whole Numbers
4                         Number          Rounding Whole Numbers
5                         Number          Rounding Whole Numbers
…                          …                             …
23932269  Ungraded Open Response  Making Equivalent Ratio Tables
23932270  Ungraded Open Response  Making Equivalent Ratio Tables
23932272  Ungraded Open Response  Making Equivalent Ratio Tables
23932273  Ungraded Open Response  Making Equivalent Ratio Tables
23932274  Ungraded Open Response  Making Equivalent Ratio Tables

          problem_contains_image  problem_contains_equation  \
1                            0.0                        0.0
2                            0.0                        0.0
```

```
3                            0.0                        0.0
4                            0.0                        0.0
5                            0.0                        0.0
…                              …                        …
23932269                     1.0                        0.0
23932270                     1.0                        0.0
23932272                     0.0                        0.0
23932273                     0.0                        0.0
23932274                     0.0                        0.0

          problem_contains_video
1                            0.0
2                            0.0
3                            0.0
4                            0.0
5                            0.0
…                              …
23932269                     0.0
23932270                     0.0
23932272                     0.0
23932273                     0.0
23932274                     0.0

[17795561 rows x 14 columns]
```

[21]:

```
#create new string variable for score_viewable_str and␣
↪continuous_score_viewable_str
merged_data['score_viewable_str'] = merged_data['score_viewable'].apply(lambda␣
↪x: 'viewable' if x == 1 else 'not_viewable')
merged_data['continuous_score_viewable_str'] =␣
↪merged_data['continuous_score_viewable'].apply(lambda x: 'continuous' if x␣
↪== 1 else 'not_continuous')
```

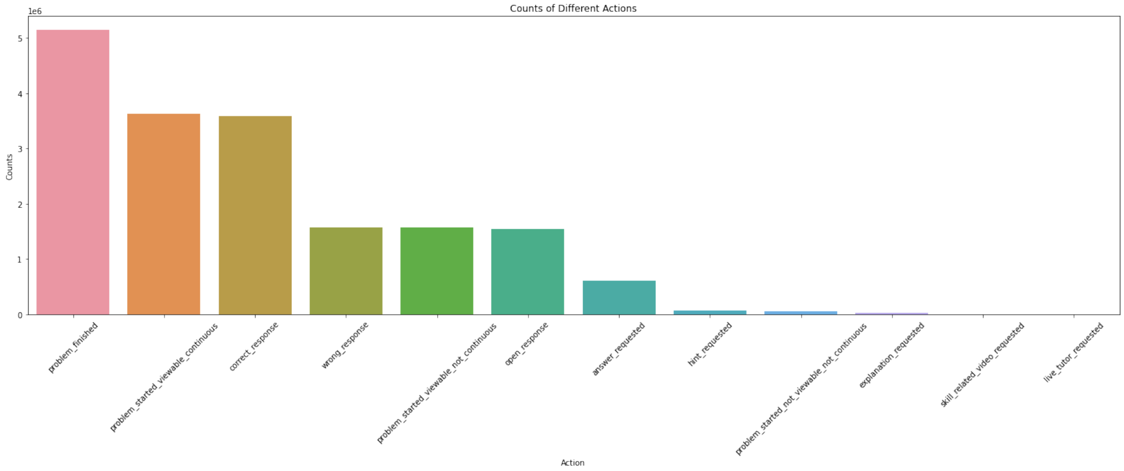
[23]:

```
#combine with problem_started in column action
merged_data.loc[merged_data['action'] == 'problem_started', 'action'] = (
    'problem_started_' +
    merged_data.loc[merged_data['action'] == 'problem_started',␣
↪'score_viewable_str'] +
    '_' +
    merged_data.loc[merged_data['action'] == 'problem_started',␣
↪'continuous_score_viewable_str']
)
```

[24]:

```
#delete the string variable we created
merged_data = merged_data.drop(columns=['score_viewable_str',␣
↪'continuous_score_viewable_str'])
```

[25]:
```
#check details of new action_counts
action_counts = merged_data['action'].value_counts()

#create his variable count
plt.figure(figsize=(25,7))#set the picture size
sns.barplot(x=action_counts.index, y=action_counts.values)

#set his title and xy
plt.title('Counts of Different Actions')
plt.xlabel('Action')
plt.ylabel('Counts')

#turn x asix
plt.xticks(rotation=45)

#show plt
plt.show()
```



[26]:
```
# get the row of action, names as 'wrong_response'
wrong_response_data = action_logs[action_logs['action'] == 'wrong_response']

# get value of 'score_viewable' and 'continuous_score_viewable' from␣
↪wrong_response_data
score_viewable_wrong_response = wrong_response_data['score_viewable']
continuous_score_viewable_wrong_response =␣
↪wrong_response_data['continuous_score_viewable']
print(score_viewable_wrong_response)
print(continuous_score_viewable_wrong_response)
```

```
2          NaN
3          NaN
```

```
31          NaN
164         NaN
276         NaN
             ..
23932201    NaN
23932202    NaN
23932203    NaN
23932204    NaN
23932210    NaN
Name: score_viewable, Length: 1580102, dtype: float64
2           NaN
3           NaN
31          NaN
164         NaN
276         NaN
             ..
23932201    NaN
23932202    NaN
23932203    NaN
23932204    NaN
23932210    NaN
Name: continuous_score_viewable, Length: 1580102, dtype: float64
```

[27]:
```
#check NA value
score_viewable_wrong_response.isnull().sum()*100/
 ↪len(score_viewable_wrong_response)
```

[27]:
```
100.0
```

[28]:
```
#check NA value
continuous_score_viewable_wrong_response.isnull().sum()*100/
 ↪len(continuous_score_viewable_wrong_response)
```

[28]:
```
100.0
```

[29]:
```
#delete score_viewable and continuous_score_viewable, because we already incode␣
 ↪these value in to problem_started in action
merged_data.drop(columns = ['score_viewable'], inplace=True)
merged_data.drop(columns = ['continuous_score_viewable'], inplace=True)

merged_data.drop(columns = ['problem_multipart_id'], inplace=True)
merged_data.drop(columns = ['problem_skill_description'], inplace=True)
```

[30]:
```
#check missing value for dataset again
merged_data.isnull().sum()
```

[30]:
```
assignment_log_id                  0
timestamp                          0
```

```
problem_id                      0
max_attempts                    0
action                          0
problem_multipart_position 342998
problem_type               342998
problem_contains_image     343106
problem_contains_equation  343106
problem_contains_video     343106
dtype: int64
```

[31]:
```
#check NA percentage of problem_skill_description, because it is just 2.6% so␣
↪we replace the NA value as 0
#merged_data['problem_skill_description'].isnull().sum()*100/
↪len(merged_data['problem_skill_description'])
#merged_data["problem_skill_description"] =␣
↪merged_data["problem_skill_description"].fillna(0)
```

[32]:
```
#delete NA value of problem_type
merged_data=merged_data.dropna(subset=['problem_type'])
```

[33]:
```
merged_data.isnull().sum()
```

[33]:
```
assignment_log_id            0
timestamp                    0
problem_id                   0
max_attempts                 0
action                       0
problem_multipart_position   0
problem_type                 0
problem_contains_image     108
problem_contains_equation  108
problem_contains_video     108
dtype: int64
```

[34]:
```
merged_data.head()
```

[34]:
```
  assignment_log_id     timestamp problem_id  max_attempts  \
1        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           3.0
2        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           0.0
3        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           0.0
4        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           0.0
5        2QV1F2GSBZ  1.599151e+09  I2GX40QIE           0.0

                                action  problem_multipart_position  \
1  problem_started_viewable_continuous                         1.0
2                       wrong_response                         1.0
3                       wrong_response                         1.0
```

```
4                      answer_requested                          1.0
5                      correct_response                          1.0

  problem_type  problem_contains_image  problem_contains_equation  \
1       Number                     0.0                        0.0
2       Number                     0.0                        0.0
3       Number                     0.0                        0.0
4       Number                     0.0                        0.0
5       Number                     0.0                        0.0

   problem_contains_video
1                     0.0
2                     0.0
3                     0.0
4                     0.0
5                     0.0
```

[35]:
```
merged_data["problem_contains_image"] = merged_data["problem_contains_image"].
 ↪fillna(0)
merged_data["problem_contains_equation"] =␣
 ↪merged_data["problem_contains_equation"].fillna(0)
merged_data["problem_contains_video"] = merged_data["problem_contains_video"].
 ↪fillna(0)
```

[36]:
```
merged_data.isnull().sum()
```

[36]:
```
assignment_log_id            0
timestamp                    0
problem_id                   0
max_attempts                 0
action                       0
problem_multipart_position   0
problem_type                 0
problem_contains_image       0
problem_contains_equation    0
problem_contains_video       0
dtype: int64
```

[37]:
```
##One issue with this representation is that ML algorithms will assume that two␣
 ↪nearby
##values are more similar than two distant values. This may be fine in some␣
 ↪cases (e.g.,
##for ordered categories such as "bad," "average," "good," and "excellent"),␣
 ↪but it is obviously
##not the case for the ocean_proximity column (for example, categories 0 and 4
##are clearly more similar than categories 0 and 1). To fix this issue, a␣
 ↪common solution
```

```
##is to create one binary attribute per category: one attribute equal to 1 when␣
 ↪the category
##is "<1H OCEAN" (and 0 otherwise), another attribute equal to 1 when the␣
 ↪category
##is "INLAND" (and 0 otherwise), and so on. This is called one-hot encoding,
##because only one attribute will be equal to 1 (hot), while the others will be␣
 ↪0 (cold).
##The new attributes are sometimes called dummy attributes. Scikit-Learn␣
 ↪provides a
##OneHotEncoder class to convert categorical values into one-hot vectors:20
```

[38]:
```
merged_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17452563 entries, 1 to 23932274
Data columns (total 10 columns):
 #   Column                      Dtype
---  ------                      -----
 0   assignment_log_id           object
 1   timestamp                   float64
 2   problem_id                  object
 3   max_attempts                float64
 4   action                      object
 5   problem_multipart_position  float64
 6   problem_type                object
 7   problem_contains_image      float64
 8   problem_contains_equation   float64
 9   problem_contains_video      float64
dtypes: float64(6), object(4)
memory usage: 1.4+ GB
```

[39]:
```
# Associate the action logs for each in unit assignment with their unit test␣
 ↪assignment
df = assignment_relationships.merge(merged_data, how='left',␣
 ↪left_on='in_unit_assignment_log_id', right_on='assignment_log_id')
#df = df[['unit_test_assignment_log_id', 'action']]
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19128313 entries, 0 to 19128312
Data columns (total 12 columns):
 #   Column                       Dtype
---  ------                       -----
 0   unit_test_assignment_log_id  object
 1   in_unit_assignment_log_id    object
 2   assignment_log_id            object
 3   timestamp                    float64
 4   problem_id                   object
 5   max_attempts                 float64
 6   action                       object
 7   problem_multipart_position   float64
 8   problem_type                 object
 9   problem_contains_image       float64
 10  problem_contains_equation    float64
 11  problem_contains_video       float64
dtypes: float64(6), object(6)
memory usage: 1.9+ GB
```

[41]:

```
df.isna().sum()
df = df.dropna()
```

[42]:

```
categorical_cols = ["action", "problem_type"]
#pd.get_dummies(student_training, columns=categorical_cols)
df = pd.get_dummies(df, columns=categorical_cols)
df = df.groupby('unit_test_assignment_log_id').sum()

# Create a feature for the total action count, then scale it between 0 and 1
action_count = df.sum(axis=1)

# Convert the individual action counts into a fraction of total actions taken
df = df.div(action_count, axis=0)

# Add the scaled total action count to the dataframe
df['action_count'] = (action_count - action_count.min()) / (action_count.max()␣
 ↪- action_count.min())
```

[43]:

```
corr = df.corr(method = 'pearson')
corr
```

[43]:

```
                                               timestamp  max_attempts  \
timestamp                                       1.000000     -0.200719
max_attempts                                   -0.200719      1.000000
problem_multipart_position                     -0.911272     -0.107725
problem_contains_image                         -0.225308      0.171467
problem_contains_equation                       0.025382      0.024441
problem_contains_video                         -0.001901      0.001587
action_answer_requested                         0.030948     -0.128489
action_correct_response                        -0.214034      0.105267
action_explanation_requested                    0.112564     -0.068528
action_hint_requested                           0.093224     -0.116741
action_live_tutor_requested                     0.005345     -0.007651
action_open_response                            0.129315     -0.484030
action_problem_finished                        -0.077280     -0.420295
action_problem_started_not_viewable_not_continuous  0.144194     -0.174764
action_problem_started_viewable_continuous     -0.239638      0.924828
action_problem_started_viewable_not_continuous  0.105379     -0.060862
```

```
action_skill_related_video_requested                0.028050    -0.010153
action_wrong_response                               0.112549    -0.292039
problem_type_Algebraic Expression                  -0.285312     0.107777
problem_type_Check All That Apply                   0.137037    -0.095696
problem_type_Exact Fraction                        -0.046853    -0.005557
problem_type_Exact Match (case sensitive)          -0.138316     0.017098
problem_type_Exact Match (ignore case)              0.029943    -0.012058
problem_type_Multiple Choice                        0.014392    -0.050389
problem_type_Number                                -0.017171     0.227423
problem_type_Numeric Expression                     0.041610     0.008465
problem_type_Ordering                               0.005920     0.050649
problem_type_Ungraded Open Response                 0.130082    -0.326792
action_count                                       -0.293445    -0.062743

                                                   problem_multipart_position
\
timestamp                                                           -0.911272
max_attempts                                                        -0.107725
problem_multipart_position                                           1.000000
problem_contains_image                                              -0.022984
problem_contains_equation                                           -0.153189
problem_contains_video                                               0.005816
action_answer_requested                                              0.004445
action_correct_response                                              0.146144
action_explanation_requested                                        -0.069958
action_hint_requested                                               -0.018381
action_live_tutor_requested                                         -0.003005
action_open_response                                                 0.050429
action_problem_finished                                              0.197030
action_problem_started_not_viewable_not_continuous                  -0.089868
action_problem_started_viewable_continuous                          -0.040927
action_problem_started_viewable_not_continuous                      -0.099883
action_skill_related_video_requested                                -0.017834
action_wrong_response                                               -0.015348
problem_type_Algebraic Expression                                    0.226061
problem_type_Check All That Apply                                   -0.117269
problem_type_Exact Fraction                                          0.013114
problem_type_Exact Match (case sensitive)                            0.104126
problem_type_Exact Match (ignore case)                              -0.002003
problem_type_Multiple Choice                                         0.015511
problem_type_Number                                                 -0.054589
problem_type_Numeric Expression                                     -0.061699
problem_type_Ordering                                                0.005256
problem_type_Ungraded Open Response                                 -0.023525
action_count                                                         0.306061

                                                   problem_contains_image  \
timestamp                                                        -0.225308
max_attempts                                                      0.171467
problem_multipart_position                                       -0.022984
problem_contains_image                                            1.000000
problem_contains_equation                                        -0.347859
problem_contains_video                                           -0.013934
action_answer_requested                                          -0.119006
action_correct_response                                           0.141351
action_explanation_requested                                     -0.083294
action_hint_requested                                            -0.127302
action_live_tutor_requested                                      -0.009251
action_open_response                                             -0.116236
action_problem_finished                                           0.027106
action_problem_started_not_viewable_not_continuous               -0.061087
action_problem_started_viewable_continuous                        0.133360
action_problem_started_viewable_not_continuous                    0.076041
action_skill_related_video_requested                             -0.025434
action_wrong_response                                            -0.148104
problem_type_Algebraic Expression                                 0.215785
problem_type_Check All That Apply                                 0.016592
problem_type_Exact Fraction                                      -0.028675
problem_type_Exact Match (case sensitive)                         0.163186
problem_type_Exact Match (ignore case)                           -0.109386
problem_type_Multiple Choice                                     -0.054394
problem_type_Number                                               0.001852
problem_type_Numeric Expression                                  -0.102122
problem_type_Ordering                                            -0.092051
problem_type_Ungraded Open Response                              -0.012111
action_count                                                      0.054944

                                                    problem_contains_equation  \
timestamp                                                            0.025382
max_attempts                                                         0.024441
problem_multipart_position                                          -0.153189
problem_contains_image                                              -0.347859
problem_contains_equation                                            1.000000
problem_contains_video                                              -0.010018
action_answer_requested                                              0.119531
action_correct_response                                              0.002340
action_explanation_requested                                         0.008592
action_hint_requested                                               -0.013693
action_live_tutor_requested                                          0.012417
action_open_response                                                -0.105674
action_problem_finished                                             -0.093222
action_problem_started_not_viewable_not_continuous                   0.038698
action_problem_started_viewable_continuous                           0.035173
action_problem_started_viewable_not_continuous                      -0.058968
```

```
action_skill_related_video_requested                                 0.005581
action_wrong_response                                                0.079015
problem_type_Algebraic Expression                                   -0.105863
problem_type_Check All That Apply                                    0.003302
problem_type_Exact Fraction                                          0.194956
problem_type_Exact Match (case sensitive)                           -0.067761
problem_type_Exact Match (ignore case)                               0.020482
problem_type_Multiple Choice                                         0.001637
problem_type_Number                                                  0.042498
problem_type_Numeric Expression                                      0.206878
problem_type_Ordering                                                0.006389
problem_type_Ungraded Open Response                                 -0.091138
action_count                                                        -0.025164

                                                    problem_contains_video  \
timestamp                                                        -0.001901
max_attempts                                                      0.001587
problem_multipart_position                                        0.005816
problem_contains_image                                           -0.013934
problem_contains_equation                                        -0.010018
problem_contains_video                                            1.000000
action_answer_requested                                          -0.014719
action_correct_response                                          -0.013320
action_explanation_requested                                      0.021656
action_hint_requested                                             0.026306
action_live_tutor_requested                                      -0.000426
action_open_response                                              0.008037
action_problem_finished                                          -0.008498
action_problem_started_not_viewable_not_continuous               -0.005052
action_problem_started_viewable_continuous                       -0.001970
action_problem_started_viewable_not_continuous                    0.011970
action_skill_related_video_requested                              0.001598
action_wrong_response                                            -0.002613
problem_type_Algebraic Expression                                -0.004648
problem_type_Check All That Apply                                 0.000352
problem_type_Exact Fraction                                      -0.007428
problem_type_Exact Match (case sensitive)                        -0.004980
problem_type_Exact Match (ignore case)                            0.002916
problem_type_Multiple Choice                                      0.037195
problem_type_Number                                              -0.023380
problem_type_Numeric Expression                                  -0.005918
problem_type_Ordering                                            -0.004707
problem_type_Ungraded Open Response                               0.010975
action_count                                                     -0.010008

                                                    action_answer_requested  \
timestamp                                                          0.030948
```

```
max_attempts                                                      -0.128489
problem_multipart_position                                         0.004445
problem_contains_image                                            -0.119006
problem_contains_equation                                          0.119531
problem_contains_video                                            -0.014719
action_answer_requested                                            1.000000
action_correct_response                                           -0.057041
action_explanation_requested                                      -0.042140
action_hint_requested                                              0.004573
action_live_tutor_requested                                       -0.007685
action_open_response                                              -0.224100
action_problem_finished                                           -0.321346
action_problem_started_not_viewable_not_continuous                -0.123290
action_problem_started_viewable_continuous                        -0.030150
action_problem_started_viewable_not_continuous                    -0.162651
action_skill_related_video_requested                              -0.001924
action_wrong_response                                              0.305090
problem_type_Algebraic Expression                                  0.029732
problem_type_Check All That Apply                                 -0.031678
problem_type_Exact Fraction                                        0.065029
problem_type_Exact Match (case sensitive)                         -0.003783
problem_type_Exact Match (ignore case)                             0.118043
problem_type_Multiple Choice                                      -0.168639
problem_type_Number                                                0.178059
problem_type_Numeric Expression                                    0.090449
problem_type_Ordering                                              0.072374
problem_type_Ungraded Open Response                               -0.234779
action_count                                                      -0.014411

                                                    action_correct_response  \
timestamp                                                         -0.214034
max_attempts                                                       0.105267
problem_multipart_position                                         0.146144
problem_contains_image                                             0.141351
problem_contains_equation                                          0.002340
problem_contains_video                                            -0.013320
action_answer_requested                                           -0.057041
action_correct_response                                            1.000000
action_explanation_requested                                      -0.094084
action_hint_requested                                             -0.105231
action_live_tutor_requested                                       -0.007307
action_open_response                                              -0.562144
action_problem_finished                                            0.513492
action_problem_started_not_viewable_not_continuous                -0.239106
action_problem_started_viewable_continuous                         0.339522
action_problem_started_viewable_not_continuous                    -0.564777
action_skill_related_video_requested                              -0.030287
```

```
action_wrong_response                                              -0.174641
problem_type_Algebraic Expression                                   0.132302
problem_type_Check All That Apply                                  -0.157020
problem_type_Exact Fraction                                         0.014316
problem_type_Exact Match (case sensitive)                           0.078146
problem_type_Exact Match (ignore case)                             -0.014340
problem_type_Multiple Choice                                       -0.046290
problem_type_Number                                                 0.480634
problem_type_Numeric Expression                                     0.014391
problem_type_Ordering                                              -0.046704
problem_type_Ungraded Open Response                                -0.688465
action_count                                                        0.162206

                                                    action_explanation_requested
\
timestamp                                                               0.112564
max_attempts                                                           -0.068528
problem_multipart_position                                             -0.069958
problem_contains_image                                                 -0.083294
problem_contains_equation                                               0.008592
problem_contains_video                                                  0.021656
action_answer_requested                                                -0.042140
action_correct_response                                                -0.094084
action_explanation_requested                                            1.000000
action_hint_requested                                                   0.370178
action_live_tutor_requested                                             0.000632
action_open_response                                                   -0.039054
action_problem_finished                                                -0.154732
action_problem_started_not_viewable_not_continuous                     -0.028956
action_problem_started_viewable_continuous                             -0.043399
action_problem_started_viewable_not_continuous                         -0.029992
action_skill_related_video_requested                                    0.052841
action_wrong_response                                                   0.124943
problem_type_Algebraic Expression                                      -0.005203
problem_type_Check All That Apply                                       0.023271
problem_type_Exact Fraction                                            -0.019708
problem_type_Exact Match (case sensitive)                              -0.023000
problem_type_Exact Match (ignore case)                                  0.032974
problem_type_Multiple Choice                                            0.010351
problem_type_Number                                                     0.006586
problem_type_Numeric Expression                                        -0.012884
problem_type_Ordering                                                  -0.019502
problem_type_Ungraded Open Response                                    -0.044678
action_count                                                           -0.039420

                                                    action_hint_requested  \
timestamp                                                        0.093224
```

```
max_attempts                                                    -0.116741
problem_multipart_position                                      -0.018381
problem_contains_image                                          -0.127302
problem_contains_equation                                       -0.013693
problem_contains_video                                           0.026306
action_answer_requested                                          0.004573
action_correct_response                                         -0.105231
action_explanation_requested                                     0.370178
action_hint_requested                                            1.000000
action_live_tutor_requested                                      0.000624
action_open_response                                            -0.046904
action_problem_finished                                         -0.176282
action_problem_started_not_viewable_not_continuous              -0.036299
action_problem_started_viewable_continuous                      -0.078850
action_problem_started_viewable_not_continuous                  -0.044228
action_skill_related_video_requested                             0.063452
action_wrong_response                                            0.124420
problem_type_Algebraic Expression                               -0.006735
problem_type_Check All That Apply                                0.088865
problem_type_Exact Fraction                                     -0.041375
problem_type_Exact Match (case sensitive)                       -0.002133
problem_type_Exact Match (ignore case)                          -0.017027
problem_type_Multiple Choice                                     0.026317
problem_type_Number                                              0.004707
problem_type_Numeric Expression                                 -0.011533
problem_type_Ordering                                            0.003558
problem_type_Ungraded Open Response                             -0.060776
action_count                                                    -0.011508

                                                    …  \
timestamp                                           …
max_attempts                                        …
problem_multipart_position                          …
problem_contains_image                              …
problem_contains_equation                           …
problem_contains_video                              …
action_answer_requested                             …
action_correct_response                             …
action_explanation_requested                        …
action_hint_requested                               …
action_live_tutor_requested                         …
action_open_response                                …
action_problem_finished                             …
action_problem_started_not_viewable_not_continuous  …
action_problem_started_viewable_continuous          …
action_problem_started_viewable_not_continuous      …
action_skill_related_video_requested                …
```

```
action_wrong_response                               …
problem_type_Algebraic Expression                   …
problem_type_Check All That Apply                   …
problem_type_Exact Fraction                         …
problem_type_Exact Match (case sensitive)           …
problem_type_Exact Match (ignore case)              …
problem_type_Multiple Choice                        …
problem_type_Number                                 …
problem_type_Numeric Expression                     …
problem_type_Ordering                               …
problem_type_Ungraded Open Response                 …
action_count                                        …

                                                    problem_type_Check All That
Apply  \
timestamp
0.137037
max_attempts
-0.095696
problem_multipart_position
-0.117269
problem_contains_image
0.016592
problem_contains_equation
0.003302
problem_contains_video
0.000352
action_answer_requested
-0.031678
action_correct_response
-0.157020
action_explanation_requested
0.023271
action_hint_requested
0.088865
action_live_tutor_requested
0.019400
action_open_response
0.044340
action_problem_finished
-0.101789
action_problem_started_not_viewable_not_continuous
0.111713
action_problem_started_viewable_continuous
-0.097849
action_problem_started_viewable_not_continuous
-0.017906
```

```
action_skill_related_video_requested
0.029166
action_wrong_response
0.215489
problem_type_Algebraic Expression
-0.062828
problem_type_Check All That Apply
1.000000
problem_type_Exact Fraction
-0.058362
problem_type_Exact Match (case sensitive)
-0.038926
problem_type_Exact Match (ignore case)
-0.021091
problem_type_Multiple Choice
0.144129
problem_type_Number
-0.284546
problem_type_Numeric Expression
-0.042784
problem_type_Ordering
-0.017297
problem_type_Ungraded Open Response
0.012020
action_count
-0.058594

                                                    problem_type_Exact Fraction
\
timestamp                                                             -0.046853
max_attempts                                                          -0.005557
problem_multipart_position                                             0.013114
problem_contains_image                                                -0.028675
problem_contains_equation                                              0.194956
problem_contains_video                                                -0.007428
action_answer_requested                                                0.065029
action_correct_response                                                0.014316
action_explanation_requested                                          -0.019708
action_hint_requested                                                 -0.041375
action_live_tutor_requested                                           -0.003280
action_open_response                                                  -0.017439
action_problem_finished                                               -0.003911
action_problem_started_not_viewable_not_continuous                    -0.019628
action_problem_started_viewable_continuous                             0.008611
action_problem_started_viewable_not_continuous                        -0.027143
action_skill_related_video_requested                                   0.005103
action_wrong_response                                                  0.031587
```

```
problem_type_Algebraic Expression                                     -0.048904
problem_type_Check All That Apply                                     -0.058362
problem_type_Exact Fraction                                            1.000000
problem_type_Exact Match (case sensitive)                              0.059512
problem_type_Exact Match (ignore case)                                -0.034072
problem_type_Multiple Choice                                          -0.053713
problem_type_Number                                                   -0.045604
problem_type_Numeric Expression                                        0.013805
problem_type_Ordering                                                 -0.019228
problem_type_Ungraded Open Response                                   -0.022626
action_count                                                           0.032712

                                                    problem_type_Exact Match
(case sensitive)  \
timestamp
-0.138316
max_attempts
0.017098
problem_multipart_position
0.104126
problem_contains_image
0.163186
problem_contains_equation
-0.067761
problem_contains_video
-0.004980
action_answer_requested
-0.003783
action_correct_response
0.078146
action_explanation_requested
-0.023000
action_hint_requested
-0.002133
action_live_tutor_requested
-0.002395
action_open_response
-0.061820
action_problem_finished
0.020204
action_problem_started_not_viewable_not_continuous
-0.029439
action_problem_started_viewable_continuous
0.038493
action_problem_started_viewable_not_continuous
-0.050309
action_skill_related_video_requested
```

```
-0.005916
action_wrong_response
0.030177
problem_type_Algebraic Expression
0.079702
problem_type_Check All That Apply
-0.038926
problem_type_Exact Fraction
0.059512
problem_type_Exact Match (case sensitive)
1.000000
problem_type_Exact Match (ignore case)
0.037874
problem_type_Multiple Choice
0.028552
problem_type_Number
-0.099347
problem_type_Numeric Expression
-0.031733
problem_type_Ordering
-0.006009
problem_type_Ungraded Open Response
-0.044241
action_count
0.080691

                                                    problem_type_Exact Match
(ignore case)  \
timestamp
0.029943
max_attempts
-0.012058
problem_multipart_position
-0.002003
problem_contains_image
-0.109386
problem_contains_equation
0.020482
problem_contains_video
0.002916
action_answer_requested
0.118043
action_correct_response
-0.014340
action_explanation_requested
0.032974
action_hint_requested
```

```
-0.017027
action_live_tutor_requested
0.001779
action_open_response
-0.048326
action_problem_finished
-0.073326
action_problem_started_not_viewable_not_continuous
-0.038764
action_problem_started_viewable_continuous
0.017765
action_problem_started_viewable_not_continuous
-0.058336
action_skill_related_video_requested
-0.007276
action_wrong_response
0.131750
problem_type_Algebraic Expression
-0.056802
problem_type_Check All That Apply
-0.021091
problem_type_Exact Fraction
-0.034072
problem_type_Exact Match (case sensitive)
0.037874
problem_type_Exact Match (ignore case)
1.000000
problem_type_Multiple Choice
-0.002539
problem_type_Number
-0.237061
problem_type_Numeric Expression
-0.029847
problem_type_Ordering
-0.027850
problem_type_Ungraded Open Response
-0.071026
action_count
-0.009442

                                                    problem_type_Multiple Choice
\
timestamp                                                               0.014392
max_attempts                                                           -0.050389
problem_multipart_position                                              0.015511
problem_contains_image                                                 -0.054394
problem_contains_equation                                               0.001637
```

```
problem_contains_video                                                  0.037195
action_answer_requested                                                -0.168639
action_correct_response                                                -0.046290
action_explanation_requested                                            0.010351
action_hint_requested                                                   0.026317
action_live_tutor_requested                                             0.003984
action_open_response                                                    0.130773
action_problem_finished                                                 0.077544
action_problem_started_not_viewable_not_continuous                      0.001297
action_problem_started_viewable_continuous                             -0.048028
action_problem_started_viewable_not_continuous                          0.010971
action_skill_related_video_requested                                   -0.004911
action_wrong_response                                                  -0.001075
problem_type_Algebraic Expression                                      -0.030887
problem_type_Check All That Apply                                       0.144129
problem_type_Exact Fraction                                            -0.053713
problem_type_Exact Match (case sensitive)                               0.028552
problem_type_Exact Match (ignore case)                                 -0.002539
problem_type_Multiple Choice                                            1.000000
problem_type_Number                                                    -0.555019
problem_type_Numeric Expression                                        -0.067015
problem_type_Ordering                                                   0.052213
problem_type_Ungraded Open Response                                     0.049535
action_count                                                           -0.060554

                                                    problem_type_Number  \
timestamp                                                     -0.017171
max_attempts                                                   0.227423
problem_multipart_position                                    -0.054589
problem_contains_image                                         0.001852
problem_contains_equation                                      0.042498
problem_contains_video                                        -0.023380
action_answer_requested                                        0.178059
action_correct_response                                        0.480634
action_explanation_requested                                   0.006586
action_hint_requested                                          0.004707
action_live_tutor_requested                                   -0.011043
action_open_response                                          -0.526364
action_problem_finished                                        0.007713
action_problem_started_not_viewable_not_continuous            -0.065411
action_problem_started_viewable_continuous                     0.365936
action_problem_started_viewable_not_continuous                -0.433459
action_skill_related_video_requested                           0.000840
action_wrong_response                                          0.088757
problem_type_Algebraic Expression                             -0.286343
problem_type_Check All That Apply                             -0.284546
problem_type_Exact Fraction                                   -0.045604
problem_type_Exact Match (case sensitive)                          -0.099347
problem_type_Exact Match (ignore case)                             -0.237061
problem_type_Multiple Choice                                       -0.555019
problem_type_Number                                                 1.000000
problem_type_Numeric Expression                                    -0.056498
problem_type_Ordering                                              -0.068764
problem_type_Ungraded Open Response                                -0.611762
action_count                                                        0.036644

                                                         problem_type_Numeric
Expression  \
timestamp
0.041610
max_attempts
0.008465
problem_multipart_position
-0.061699
problem_contains_image
-0.102122
problem_contains_equation
0.206878
problem_contains_video
-0.005918
action_answer_requested
0.090449
action_correct_response
0.014391
action_explanation_requested
-0.012884
action_hint_requested
-0.011533
action_live_tutor_requested
-0.002977
action_open_response
-0.043590
action_problem_finished
-0.027125
action_problem_started_not_viewable_not_continuous
-0.011009
action_problem_started_viewable_continuous
0.024924
action_problem_started_viewable_not_continuous
-0.043101
action_skill_related_video_requested
-0.003285
action_wrong_response
0.031442
```

```
problem_type_Algebraic Expression
-0.004093
problem_type_Check All That Apply
-0.042784
problem_type_Exact Fraction
0.013805
problem_type_Exact Match (case sensitive)
-0.031733
problem_type_Exact Match (ignore case)
-0.029847
problem_type_Multiple Choice
-0.067015
problem_type_Number
-0.056498
problem_type_Numeric Expression
1.000000
problem_type_Ordering
-0.003734
problem_type_Ungraded Open Response
-0.055966
action_count
-0.024857

                                                    problem_type_Ordering  \
timestamp                                                        0.005920
max_attempts                                                     0.050649
problem_multipart_position                                       0.005256
problem_contains_image                                          -0.092051
problem_contains_equation                                        0.006389
problem_contains_video                                          -0.004707
action_answer_requested                                          0.072374
action_correct_response                                         -0.046704
action_explanation_requested                                    -0.019502
action_hint_requested                                            0.003558
action_live_tutor_requested                                     -0.002295
action_open_response                                            -0.035423
action_problem_finished                                         -0.079052
action_problem_started_not_viewable_not_continuous               0.006241
action_problem_started_viewable_continuous                       0.049260
action_problem_started_viewable_not_continuous                  -0.017716
action_skill_related_video_requested                            -0.003416
action_wrong_response                                            0.044035
problem_type_Algebraic Expression                               -0.037180
problem_type_Check All That Apply                               -0.017297
problem_type_Exact Fraction                                     -0.019228
problem_type_Exact Match (case sensitive)                       -0.006009
problem_type_Exact Match (ignore case)                          -0.027850
```

```
problem_type_Multiple Choice                                          0.052213
problem_type_Number                                                  -0.068764
problem_type_Numeric Expression                                      -0.003734
problem_type_Ordering                                                 1.000000
problem_type_Ungraded Open Response                                  -0.036024
action_count                                                         -0.020257

                                                    problem_type_Ungraded Open
Response  \
timestamp
0.130082
max_attempts
-0.326792
problem_multipart_position
-0.023525
problem_contains_image
-0.012111
problem_contains_equation
-0.091138
problem_contains_video
0.010975
action_answer_requested
-0.234779
action_correct_response
-0.688465
action_explanation_requested
-0.044678
action_hint_requested
-0.060776
action_live_tutor_requested
0.005442
action_open_response
0.790909
action_problem_finished
0.023509
action_problem_started_not_viewable_not_continuous
0.102335
action_problem_started_viewable_continuous
-0.572508
action_problem_started_viewable_not_continuous
0.748395
action_skill_related_video_requested
-0.003931
action_wrong_response
-0.279696
problem_type_Algebraic Expression
-0.124926
```

```
problem_type_Check All That Apply
0.012020
problem_type_Exact Fraction
-0.022626
problem_type_Exact Match (case sensitive)
-0.044241
problem_type_Exact Match (ignore case)
-0.071026
problem_type_Multiple Choice
0.049535
problem_type_Number
-0.611762
problem_type_Numeric Expression
-0.055966
problem_type_Ordering
-0.036024
problem_type_Ungraded Open Response
1.000000
action_count
-0.050950
```

```
                                                    action_count
timestamp                                              -0.293445
max_attempts                                           -0.062743
problem_multipart_position                              0.306061
problem_contains_image                                  0.054944
problem_contains_equation                              -0.025164
problem_contains_video                                 -0.010008
action_answer_requested                                -0.014411
action_correct_response                                 0.162206
action_explanation_requested                           -0.039420
action_hint_requested                                  -0.011508
action_live_tutor_requested                             0.009798
action_open_response                                    0.011703
action_problem_finished                                 0.171740
action_problem_started_not_viewable_not_continuous     -0.104292
action_problem_started_viewable_continuous             -0.001836
action_problem_started_viewable_not_continuous         -0.092204
action_skill_related_video_requested                   -0.013678
action_wrong_response                                  -0.039738
problem_type_Algebraic Expression                       0.101229
problem_type_Check All That Apply                      -0.058594
problem_type_Exact Fraction                             0.032712
problem_type_Exact Match (case sensitive)               0.080691
problem_type_Exact Match (ignore case)                 -0.009442
problem_type_Multiple Choice                           -0.060554
problem_type_Number                                     0.036644
```

```
problem_type_Numeric Expression                        -0.024857
problem_type_Ordering                                  -0.020257
problem_type_Ungraded Open Response                    -0.050950
action_count                                            1.000000

[29 rows x 29 columns]
```

[44]:
```
# Merge action count features with the training unit test scores
student_training = student_training.merge(df, how='left',␣
 ↪left_on='assignment_log_id',right_on='unit_test_assignment_log_id')
```

[45]:
```
# Merge action count features with the evaluation unit test scores
student_predict = student_predict.merge(df, how='left',␣
 ↪left_on='assignment_log_id',right_on='unit_test_assignment_log_id')
```

[49]:
```
#student_training.info()
df=student_training
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 452439 entries, 0 to 452438
Data columns (total 32 columns):
 #   Column                                     Non-Null Count   Dtype
---  ------                                     --------------   -----
 0   assignment_log_id                          452439 non-null  object
 1   problem_id                                 452439 non-null  object
 2   score                                      452439 non-null  int64
 3   timestamp                                  452395 non-null
float64
 4   max_attempts                               452395 non-null
float64
 5   problem_multipart_position                 452395 non-null
float64
 6   problem_contains_image                     452395 non-null
float64
 7   problem_contains_equation                  452395 non-null
float64
 8   problem_contains_video                     452395 non-null
float64
 9   action_answer_requested                    452395 non-null
float64
 10  action_correct_response                    452395 non-null
float64
 11  action_explanation_requested               452395 non-null
float64
 12  action_hint_requested                      452395 non-null
float64
 13  action_live_tutor_requested                452395 non-null
```

```
float64
 14  action_open_response                                452395 non-null
float64
 15  action_problem_finished                             452395 non-null
float64
 16  action_problem_started_not_viewable_not_continuous  452395 non-null
float64
 17  action_problem_started_viewable_continuous          452395 non-null
float64
 18  action_problem_started_viewable_not_continuous      452395 non-null
float64
 19  action_skill_related_video_requested                452395 non-null
float64
 20  action_wrong_response                               452395 non-null
float64
 21  problem_type_Algebraic Expression                   452395 non-null
float64
 22  problem_type_Check All That Apply                   452395 non-null
float64
 23  problem_type_Exact Fraction                         452395 non-null
float64
 24  problem_type_Exact Match (case sensitive)           452395 non-null
float64
 25  problem_type_Exact Match (ignore case)              452395 non-null
float64
 26  problem_type_Multiple Choice                        452395 non-null
float64
 27  problem_type_Number                                 452395 non-null
float64
 28  problem_type_Numeric Expression                     452395 non-null
float64
 29  problem_type_Ordering                               452395 non-null
float64
 30  problem_type_Ungraded Open Response                 452395 non-null
float64
 31  action_count                                        452395 non-null
float64
dtypes: float64(29), int64(1), object(2)
memory usage: 113.9+ MB
```

[50]:

```python
df.head()
df.isnull().sum()
```

[50]:

```
assignment_log_id                                      0
problem_id                                             0
score                                                  0
timestamp                                             44
```

```
max_attempts                                          44
problem_multipart_position                            44
problem_contains_image                                44
problem_contains_equation                             44
problem_contains_video                                44
action_answer_requested                               44
action_correct_response                               44
action_explanation_requested                          44
action_hint_requested                                 44
action_live_tutor_requested                           44
action_open_response                                  44
action_problem_finished                               44
action_problem_started_not_viewable_not_continuous    44
action_problem_started_viewable_continuous            44
action_problem_started_viewable_not_continuous        44
action_skill_related_video_requested                  44
action_wrong_response                                 44
problem_type_Algebraic Expression                     44
problem_type_Check All That Apply                     44
problem_type_Exact Fraction                           44
problem_type_Exact Match (case sensitive)             44
problem_type_Exact Match (ignore case)                44
problem_type_Multiple Choice                          44
problem_type_Number                                   44
problem_type_Numeric Expression                       44
problem_type_Ordering                                 44
problem_type_Ungraded Open Response                   44
action_count                                          44
dtype: int64
```

[55]:

```
df.fillna (0,inplace=True)
```

[56]:

```
#split dataset 70/30
from sklearn.model_selection import train_test_split
X=df.drop(columns =['score','assignment_log_id','problem_id'])
Y=df['score']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3,␣
 ↪random_state=42)
```

[57]:

```
#linear regression model
import sklearn.linear_model
model = sklearn.linear_model.LinearRegression()
model.fit(X, Y)
```

[57]:

```
LinearRegression()
```

[58]:

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
lda = LinearDiscriminantAnalysis()
lda.fit(X_train, Y_train)
```

[58]:

```
LinearDiscriminantAnalysis()
```

[59]:

```
# Prediction (based on 0.5 default threshold)
predictions = lda.predict(X_test)
```

[60]:

```
from sklearn.metrics import f1_score, accuracy_score
# F-measure
f_measure = f1_score(Y_test, predictions)
print(f"F-measure: {f_measure:.2f}")
```

```
F-measure: 0.74
```

[61]:

```
#accurancy
accuracy_lda= accuracy_score(Y_test,predictions)
print(f"Accuracy: {accuracy_lda:.2f}")
```

```
Accuracy: 0.65
```

[62]:

```
df = student_predict
df.head()
df.isna().sum()
```

[62]:

```
id                                                         0
assignment_log_id                                          0
problem_id                                                 0
score                                                 124455
timestamp                                                 11
max_attempts                                              11
problem_multipart_position                                11
problem_contains_image                                    11
problem_contains_equation                                 11
problem_contains_video                                    11
action_answer_requested                                   11
action_correct_response                                   11
action_explanation_requested                              11
action_hint_requested                                     11
action_live_tutor_requested                               11
action_open_response                                      11
action_problem_finished                                   11
action_problem_started_not_viewable_not_continuous        11
action_problem_started_viewable_continuous                11
action_problem_started_viewable_not_continuous            11
action_skill_related_video_requested                      11
action_wrong_response                                     11
```

```
problem_type_Algebraic Expression                         11
problem_type_Check All That Apply                         11
problem_type_Exact Fraction                               11
problem_type_Exact Match (case sensitive)                 11
problem_type_Exact Match (ignore case)                    11
problem_type_Multiple Choice                              11
problem_type_Number                                       11
problem_type_Numeric Expression                           11
problem_type_Ordering                                     11
problem_type_Ungraded Open Response                       11
action_count                                              11
dtype: int64
```

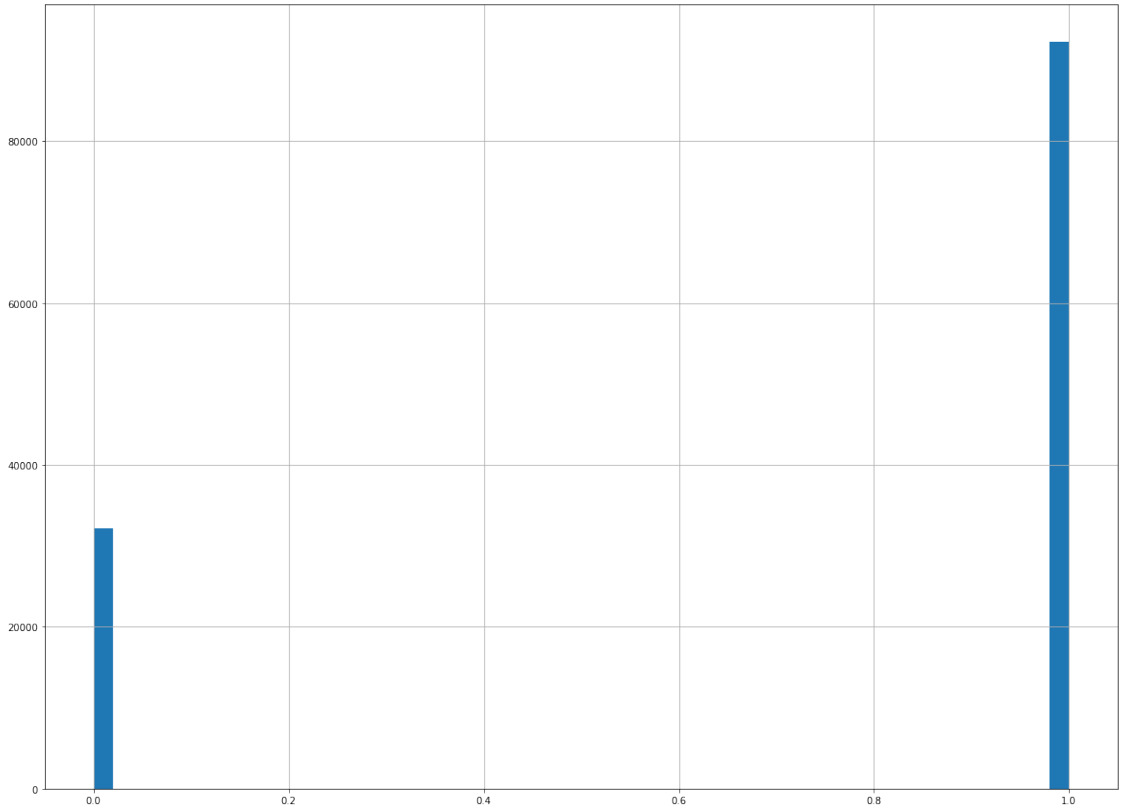
[63]:
```
df = df.fillna(0)
df['score'] = df['score'].replace(0, np.nan)
```

[64]:
```
#  prediction_data

X_prediction = df.drop(columns = ['assignment_log_id',␣
 ↪'problem_id','score','id'])

#  LDA
predicted_labels = lda.predict(X_prediction)

#
df["score"] = predicted_labels

#
print(df.head())
```

```
   id assignment_log_id  problem_id  score  timestamp  max_attempts  \
0   0        11VO3FPL7U   N9F071P7I      1        1.0  4.768697e-10
1   1        11VO3FPL7U  2EID4DTRNQ      1        1.0  4.768697e-10
2   2        11VO3FPL7U  1PFVQE8WVV      1        1.0  4.768697e-10
3   3        11VO3FPL7U  28ZP6YF22Q      1        1.0  4.768697e-10
4   4        11VO3FPL7U  1H85EY5KJF      1        1.0  4.768697e-10

   problem_multipart_position  problem_contains_image  \
0                1.002789e-09            5.449939e-11
1                1.002789e-09            5.449939e-11
2                1.002789e-09            5.449939e-11
3                1.002789e-09            5.449939e-11
4                1.002789e-09            5.449939e-11

   problem_contains_equation  problem_contains_video  …  \
0               3.685521e-10                     0.0  …
1               3.685521e-10                     0.0  …
2               3.685521e-10                     0.0  …
```

```
3               3.685521e-10                     0.0  …
4               3.685521e-10                     0.0  …

   problem_type_Check All That Apply  problem_type_Exact Fraction  \
0                       6.812424e-12                 2.452473e-11
1                       6.812424e-12                 2.452473e-11
2                       6.812424e-12                 2.452473e-11
3                       6.812424e-12                 2.452473e-11
4                       6.812424e-12                 2.452473e-11

   problem_type_Exact Match (case sensitive)  \
0                                        0.0
1                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0

   problem_type_Exact Match (ignore case)  problem_type_Multiple Choice  \
0                                     0.0                  2.997467e-11
1                                     0.0                  2.997467e-11
2                                     0.0                  2.997467e-11
3                                     0.0                  2.997467e-11
4                                     0.0                  2.997467e-11

   problem_type_Number  problem_type_Numeric Expression  \
0         4.550699e-10                              0.0
1         4.550699e-10                              0.0
2         4.550699e-10                              0.0
3         4.550699e-10                              0.0
4         4.550699e-10                              0.0

   problem_type_Ordering  problem_type_Ungraded Open Response  action_count
0                    0.0                         1.110425e-10      0.196237
1                    0.0                         1.110425e-10      0.196237
2                    0.0                         1.110425e-10      0.196237
3                    0.0                         1.110425e-10      0.196237
4                    0.0                         1.110425e-10      0.196237

[5 rows x 33 columns]
```

[65]:
```
df['score'].hist(bins=50,figsize=(20,15))
plt.show()
```



[66]:
```
final_df = df[["score"]]
final_df.to_csv("lda.csv")
```

[67]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

#         data DataFrame
#     X    'Dropout'   y

logreg = LogisticRegression()
logreg.fit(X_train, Y_train)

#prediction
Y_pred = logreg.predict(X_test)

#accuracy
accuracy = accuracy_score(Y_test, Y_pred)
print("Accuracy:", accuracy)
```

```
report = classification_report(Y_test, Y_pred)
print("Classification report:\n", report)
```

```
Accuracy: 0.5843500427312646
Classification report:
               precision    recall  f1-score   support

           0       0.89      0.00      0.00     56424
           1       0.58      1.00      0.74     79308

    accuracy                           0.58    135732
   macro avg       0.74      0.50      0.37    135732
weighted avg       0.71      0.58      0.43    135732
```

[68]:

```
#f1
from sklearn.metrics import f1_score
f1_score(Y_test, Y_pred, average='weighted', labels=np.unique(Y_pred))
```

[68]:

```
0.43111579853353843
```

[71]:

```
#randomforest

#
import numpy as np
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score



#
rf_clf = RandomForestClassifier(n_estimators=100, random_state=42)

#
rf_clf.fit(X_train, Y_train)

#
Y_pred = rf_clf.predict(X_test)

#
accuracy = accuracy_score(Y_test, Y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.6854094833937465
```

[72]:

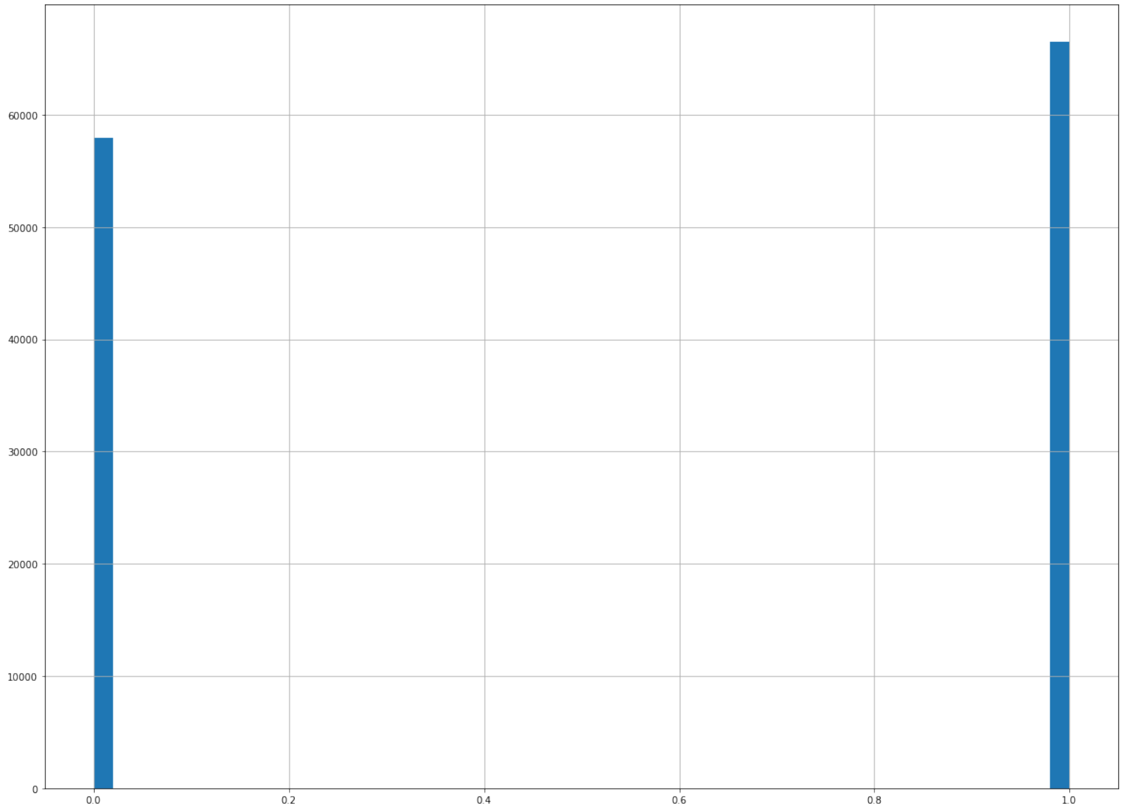
```
#  prediction_data

X_prediction = df.drop(columns = ['assignment_log_id',␣
 ↪'problem_id','score','id'])

#  LDA
predicted_labels = rf_clf.predict(X_prediction)

#
df["score"] = predicted_labels

#
print(df.head())
```

```
   id assignment_log_id  problem_id  score  timestamp  max_attempts  \
0   0        11V03FPL7U   N9F071P7I      1        1.0  4.768697e-10
1   1        11V03FPL7U  2EID4DTRNQ      1        1.0  4.768697e-10
2   2        11V03FPL7U  1PFVQE8WVV      1        1.0  4.768697e-10
3   3        11V03FPL7U  28ZP6YF22Q      1        1.0  4.768697e-10
4   4        11V03FPL7U  1H85EY5KJF      1        1.0  4.768697e-10

   problem_multipart_position  problem_contains_image  \
0                1.002789e-09            5.449939e-11
1                1.002789e-09            5.449939e-11
2                1.002789e-09            5.449939e-11
3                1.002789e-09            5.449939e-11
4                1.002789e-09            5.449939e-11

   problem_contains_equation  problem_contains_video  …  \
0               3.685521e-10                     0.0  …
1               3.685521e-10                     0.0  …
2               3.685521e-10                     0.0  …
3               3.685521e-10                     0.0  …
4               3.685521e-10                     0.0  …

   problem_type_Check All That Apply  problem_type_Exact Fraction  \
0                       6.812424e-12                 2.452473e-11
1                       6.812424e-12                 2.452473e-11
2                       6.812424e-12                 2.452473e-11
3                       6.812424e-12                 2.452473e-11
4                       6.812424e-12                 2.452473e-11

   problem_type_Exact Match (case sensitive)  \
0                                        0.0
1                                        0.0
2                                        0.0
3                                        0.0
4                                        0.0
```

```
   problem_type_Exact Match (ignore case)  problem_type_Multiple Choice  \
0                                     0.0                  2.997467e-11   
1                                     0.0                  2.997467e-11   
2                                     0.0                  2.997467e-11   
3                                     0.0                  2.997467e-11   
4                                     0.0                  2.997467e-11   

   problem_type_Number  problem_type_Numeric Expression  \
0         4.550699e-10                              0.0   
1         4.550699e-10                              0.0   
2         4.550699e-10                              0.0   
3         4.550699e-10                              0.0   
4         4.550699e-10                              0.0   

   problem_type_Ordering  problem_type_Ungraded Open Response  action_count  
0                    0.0                         1.110425e-10      0.196237  
1                    0.0                         1.110425e-10      0.196237  
2                    0.0                         1.110425e-10      0.196237  
3                    0.0                         1.110425e-10      0.196237  
4                    0.0                         1.110425e-10      0.196237  

[5 rows x 33 columns]
```

[73]:

```
df['score'].hist(bins=50,figsize=(20,15))
plt.show()
```



[74]:

```
final_df = df[["score"]]
final_df.to_csv("rf.csv")
```

[ ]: