ljndtxzse

February 8, 2025

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.feature_selection import SelectKBest, f_regression
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_absolute_error
```

[2]:
```
import warnings
warnings.filterwarnings('ignore')
```

[3]:
```
train = pd.read_csv('train1.csv')
test = pd.read_csv('test1.csv')
```

[4]:
```
train
```

[4]:

	c_id	total_work_ex(yrs)	active_days_on_linkedin	\
0	INF_1	7	1918.71	
1	INF_2	3	1263.82	
2	INF_3	4	931.03	
3	INF_4	12	823.78	
4	INF_5	14	731.53	
…	…	…	…	
43889	INF_43890	7	1368.51	
43890	INF_43891	8	1022.81	
43891	INF_43892	10	1779.29	
43892	INF_43893	10	4118.72	
43893	INF_43894	8	2937.61	

	n_times_paid_promo	no_of_previous_positions	days_in_current_org	\
0	3	1	497	
1	3	2	228	
2	5	4	264	
3	2	5	161	
4	3	6	261	

```
…                    …                          …                    …
43889                    1                          5                  223
43890                    1                          6                  206
43891                    6                          1                 4161
43892                    4                          2                  124
43893                    3                          2                  141

        age  reactions_on_each_post_per_second  avg_angry_reaction_on_posts  \
0      41.0                                0.00                       116.86
1       NaN                                 NaN                       100.21
2      41.0                                0.00                        97.53
3      41.0                                0.00                        99.57
4      40.0                                0.00                       103.24
…     …                                    …                          …
43889  50.0                                0.01                      5133.22
43890  51.0                                0.01                      3612.62
43891  34.0                                0.01                         0.00
43892  33.0                                0.01                         0.00
43893  47.0                                0.01                         0.00

       avg_disgusted_reaction_on_posts  …  avg_sad_reaction_on_posts  \
0                                 1.13  …                       0.01
1                                 0.86  …                       0.01
2                                 1.09  …                       0.01
3                                 0.96  …                       0.01
4                                 1.11  …                       0.01
…                                 …  …                       …
43889                            31.16  …                       0.01
43890                            32.01  …                       0.01
43891                             0.00  …                       0.00
43892                             0.00  …                       0.00
43893                             0.00  …                       0.00

       avg_surprise_reaction_on_posts  max_reactions_in_single_post  \
0                                0.11                         393.0
1                                0.11                         137.0
2                                0.12                         406.0
3                                0.13                         188.0
4                                0.12                          76.0
…                                …                          …
43889                            0.01                        1626.0
43890                            0.01                         204.0
43891                            0.00                         993.0
43892                            0.00                         776.0
43893                            0.00                         303.0

       profile_completion%  gender  country  display_picture_clarity  \
```

```
0                    74.0    Male  Ireland                   102.18
1                    52.0    Male       UK                   110.90
2                    75.0    Male    Japan                    89.25
3                    80.0    Male    China                   110.84
4                    82.0    Male    China                    92.85
…                     …      …       …                         …
43889                87.0    Male    India                    44.04
43890                83.0    Male  Germany                    55.29
43891                85.0     NaN       UK                    94.62
43892                98.0    Male    India                    75.73
43893                81.0    Male      USA                   135.81

      total_posts avg_reaction_in_each_post  n_followers
0              64                       180          538
1              71                        25          513
2             100                        56          566
3             150                       109          472
4             152                        56          542
…             …                          …            …
43889         147                        62         1577
43890          92                       170         1644
43891          69                       172         2074
43892         151                        57         2074
43893         114                       136         1048

[43894 rows x 23 columns]
```

[5]:
```
test
```

[5]:
```
            c_id  total_work_ex(yrs)  active_days_on_linkedin  \
0      INF_43895                   6                  1460.75
1      INF_43896                  12                  1260.45
2      INF_43897                  14                  2208.22
3      INF_43898                   6                  1240.59
4      INF_43899                  12                  1210.51
…          …                     …                        …
18807  INF_62702                  10                  1192.31
18808  INF_62703                  14                   977.20
18809  INF_62704                  13                  1167.06
18810  INF_62705                  13                   828.51
18811  INF_62706                  10                  1256.03

       n_times_paid_promo  no_of_previous_positions  days_in_current_org  \
0                       4                         2                  357
1                       4                         4                  109
2                       1                         1                  353
3                       4                         2                 1023
```

```
4                      3                      3                 3314
…                    …                      …                 …
18807                  2                      6                 1033
18808                  4                      7                  184
18809                  3                      7                  131
18810                  1                      9                 2167
18811                  3                     10                  623

        age  reactions_on_each_post_per_second  avg_angry_reaction_on_posts  \
0      45.0                                0.01                         0.00
1      45.0                                0.01                         0.00
2      48.0                                0.01                         2.08
3      47.0                                0.01                         2.69
4      48.0                                0.01                         1.99
…     …                                    …                            …
18807  63.0                                0.00                         0.00
18808  62.0                                0.00                         0.00
18809  62.0                                0.00                         0.00
18810  64.0                                0.00                         0.00
18811  61.0                                0.00                         0.00

       avg_disgusted_reaction_on_posts  …  avg_neutral_reaction_on_posts  \
0                                 0.00  …                           0.00
1                                 0.00  …                           0.00
2                                74.72  …                           0.02
3                                80.05  …                           0.01
4                                84.01  …                           0.01
…                                …   …                           …
18807                             0.14  …                           0.00
18808                             0.13  …                           0.00
18809                             0.16  …                           0.00
18810                             0.16  …                           0.00
18811                             0.15  …                           0.00

       avg_sad_reaction_on_posts  avg_surprise_reaction_on_posts  \
0                           0.00                            0.00
1                           0.00                            0.00
2                           4.76                            0.01
3                           4.72                            0.01
4                           3.69                            0.01
…                           …                               …
18807                       0.00                            0.00
18808                       0.00                            0.00
18809                       0.00                            0.00
18810                       0.00                            0.00
18811                       0.00                            0.00
```

```
       max_reactions_in_single_post  profile_completion%  gender  country  \
0                             208.0                 76.0    Male    China
1                             259.0                 86.0    Male       UK
2                             113.0                 90.0    Male    India
3                             132.0                 90.0    Male    India
4                             150.0                 80.0    Male   France
…                             …                     …       …       …
18807                         462.0                 81.0     NaN   France
18808                         230.0                  NaN     NaN       UK
18809                         407.0                 68.0     NaN  Germany
18810                         406.0                 89.0    Male  Ireland
18811                         444.0                 79.0    Male      USA

      display_picture_clarity total_posts  avg_reaction_in_each_post
0                      109.74         134                        130
1                      115.84          99                        124
2                      125.02         113                        188
3                      127.49         147                        143
4                      123.60          93                        149
…                       …         …                        …
18807                  118.13          80                         27
18808                   97.56          38                        160
18809                  110.58         117                         50
18810                  109.09          66                         99
18811                  100.88         103                         55

[18812 rows x 22 columns]
```

[6]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43894 entries, 0 to 43893
Data columns (total 23 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   c_id                               43894 non-null  object
 1   total_work_ex(yrs)                 43894 non-null  int64
 2   active_days_on_linkedin            43894 non-null  float64
 3   n_times_paid_promo                 43894 non-null  int64
 4   no_of_previous_positions           43894 non-null  int64
 5   days_in_current_org                43894 non-null  int64
 6   age                                42578 non-null  float64
 7   reactions_on_each_post_per_second  42571 non-null  float64
 8   avg_angry_reaction_on_posts        42580 non-null  float64
 9   avg_disgusted_reaction_on_posts    43894 non-null  float64
 10  avg_scared_reaction_on_posts       43041 non-null  float64
 11  avg_happy_reaction_on_posts        43024 non-null  float64
```

```
 12  avg_neutral_reaction_on_posts      43052 non-null  float64
 13  avg_sad_reaction_on_posts          43894 non-null  float64
 14  avg_surprise_reaction_on_posts     43894 non-null  float64
 15  max_reactions_in_single_post       39677 non-null  float64
 16  profile_completion%                39714 non-null  float64
 17  gender                             39796 non-null  object
 18  country                            43894 non-null  object
 19  display_picture_clarity            43894 non-null  float64
 20  total_posts                        43894 non-null  int64
 21  avg_reaction_in_each_post          43894 non-null  int64
 22  n_followers                        43894 non-null  int64
dtypes: float64(13), int64(7), object(3)
memory usage: 7.7+ MB
```

[7]:
```
test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18812 entries, 0 to 18811
Data columns (total 22 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   c_id                               18812 non-null  object
 1   total_work_ex(yrs)                 18812 non-null  int64
 2   active_days_on_linkedin            18812 non-null  float64
 3   n_times_paid_promo                 18812 non-null  int64
 4   no_of_previous_positions           18812 non-null  int64
 5   days_in_current_org                18812 non-null  int64
 6   age                                18274 non-null  float64
 7   reactions_on_each_post_per_second  18280 non-null  float64
 8   avg_angry_reaction_on_posts        18268 non-null  float64
 9   avg_disgusted_reaction_on_posts    18812 non-null  float64
 10  avg_scared_reaction_on_posts       18426 non-null  float64
 11  avg_happy_reaction_on_posts        18435 non-null  float64
 12  avg_neutral_reaction_on_posts      18418 non-null  float64
 13  avg_sad_reaction_on_posts          18812 non-null  float64
 14  avg_surprise_reaction_on_posts     18812 non-null  float64
 15  max_reactions_in_single_post       17057 non-null  float64
 16  profile_completion%                17023 non-null  float64
 17  gender                             16957 non-null  object
 18  country                            18812 non-null  object
 19  display_picture_clarity            18812 non-null  float64
 20  total_posts                        18812 non-null  int64
 21  avg_reaction_in_each_post          18812 non-null  int64
dtypes: float64(13), int64(6), object(3)
memory usage: 3.2+ MB
```

[8]:
```
train.describe()
```

[8]:

	total_work_ex(yrs)	active_days_on_linkedin	n_times_paid_promo
count	43894.000000	43894.000000	43894.000000
mean	8.006288	1556.892610	2.744885
std	3.724876	1439.352766	1.865937
min	2.000000	0.000000	0.000000
25%	5.000000	748.022500	1.000000
50%	8.000000	1239.750000	3.000000
75%	11.000000	1955.905000	4.000000
max	14.000000	46155.040000	15.000000

	no_of_previous_positions	days_in_current_org	age
count	43894.000000	43894.000000	42578.000000
mean	3.997745	989.791589	46.619498
std	3.160890	1100.190456	10.721761
min	1.000000	-46.000000	6.000000
25%	2.000000	317.000000	40.000000
50%	3.000000	680.000000	47.000000
75%	5.000000	1236.000000	54.000000
max	49.000000	21927.000000	82.000000

	reactions_on_each_post_per_second	avg_angry_reaction_on_posts
count	42571.000000	42580.000000
mean	0.025427	225.819518
std	0.111794	1323.104295
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.010000	0.460000
75%	0.010000	6.560000
max	0.990000	16369.120000

	avg_disgusted_reaction_on_posts	avg_scared_reaction_on_posts
count	43894.000000	43041.000000
mean	389.998430	167.032594
std	1736.005772	1077.925714
min	0.000000	0.000000
25%	0.120000	0.000000
50%	1.070000	0.530000
75%	18.000000	5.870000
max	16408.640000	16324.950000

	avg_happy_reaction_on_posts	avg_neutral_reaction_on_posts
count	43024.000000	43052.000000
mean	96.484333	31.012957
std	59.170638	52.171328
min	0.000000	0.000000
25%	27.307500	0.000000
50%	123.020000	0.100000

```
75%                    142.680000                    43.200000
max                    165.000000                   164.850000

       avg_sad_reaction_on_posts  avg_surprise_reaction_on_posts  \
count               43894.000000                    43894.000000
mean                    2.468895                        2.162001
std                    13.451450                       12.236110
min                     0.000000                        0.000000
25%                     0.000000                        0.000000
50%                     0.010000                        0.010000
75%                     0.090000                        0.090000
max                   163.400000                      163.270000

       max_reactions_in_single_post  profile_completion%  \
count                  39677.000000         39714.000000
mean                     685.365048            81.453266
std                     1374.485859            12.091751
min                        5.000000            50.000000
25%                      215.000000            74.000000
50%                      388.000000            83.000000
75%                      656.000000            91.000000
max                    23986.000000            99.000000

       display_picture_clarity   total_posts  avg_reaction_in_each_post  \
count             43894.000000  43894.000000                43894.000000
mean                 92.052385    107.121771                  106.784230
std                  48.529870     47.502301                   47.527264
min                   0.010000     25.000000                   25.000000
25%                  65.950000     66.000000                   66.000000
50%                 110.400000    107.000000                  107.000000
75%                 127.570000    148.000000                  148.000000
max                 155.970000    189.000000                  189.000000

         n_followers
count   43894.000000
mean     1250.378457
std      4349.330842
min        50.000000
25%       507.000000
50%       829.000000
75%      1347.000000
max    427434.000000
```

[9]:
```
test.describe()
```

[9]:
```
       total_work_ex(yrs)  active_days_on_linkedin  n_times_paid_promo  \
count        18812.000000             18812.000000        18812.000000
```

mean	7.993515	1504.312575	2.751967
std	3.766828	1238.286468	1.877327
min	2.000000	0.000000	0.000000
25%	5.000000	734.142500	1.000000
50%	8.000000	1205.235000	3.000000
75%	11.000000	1912.220000	4.000000
max	14.000000	31045.580000	13.000000

	no_of_previous_positions	days_in_current_org	age
count	18812.000000	18812.000000	18274.000000
mean	4.005103	989.223793	46.343931
std	3.134854	1060.379888	10.774286
min	1.000000	-114.000000	5.000000
25%	2.000000	336.000000	39.000000
50%	3.000000	675.000000	47.000000
75%	5.000000	1256.000000	53.000000
max	47.000000	18108.000000	84.000000

	reactions_on_each_post_per_second	avg_angry_reaction_on_posts
count	18280.000000	18268.000000
mean	0.029704	216.827593
std	0.125839	1266.939267
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.010000	0.450000
75%	0.010000	5.722500
max	0.990000	15962.890000

	avg_disgusted_reaction_on_posts	avg_scared_reaction_on_posts
count	18812.000000	18426.000000
mean	349.100560	177.028082
std	1588.399372	1163.780017
min	0.000000	0.000000
25%	0.120000	0.000000
50%	0.970000	0.530000
75%	15.077500	5.507500
max	16285.800000	16162.600000

	avg_happy_reaction_on_posts	avg_neutral_reaction_on_posts
count	18435.000000	18418.000000
mean	97.645340	30.475950
std	58.956414	51.998241
min	0.000000	0.000000
25%	31.575000	0.000000
50%	123.880000	0.070000
75%	143.100000	40.290000
max	164.990000	164.750000

```
       avg_sad_reaction_on_posts  avg_surprise_reaction_on_posts  \
count                18812.000000                    18812.000000
mean                     2.350562                        2.070559
std                     13.059569                       11.552089
min                      0.000000                        0.000000
25%                      0.000000                        0.000000
50%                      0.010000                        0.010000
75%                      0.090000                        0.070000
max                    162.130000                      160.100000

       max_reactions_in_single_post  profile_completion%  \
count                  17057.000000         17023.000000
mean                     748.693029            81.419374
std                     1694.561582            12.299710
min                        5.000000            50.000000
25%                      207.000000            74.000000
50%                      382.000000            83.000000
75%                      649.000000            91.000000
max                    23992.000000            99.000000

       display_picture_clarity   total_posts  avg_reaction_in_each_post
count             18812.000000  18812.000000                18812.000000
mean                 90.534302    106.801403                  107.420583
std                  49.201132     47.745352                   47.800570
min                   0.010000     25.000000                   25.000000
25%                  61.677500     66.000000                   66.000000
50%                 109.215000    106.500000                  108.000000
75%                 126.840000    148.000000                  149.000000
max                 156.040000    189.000000                  189.000000
```

[10]:
```
train.duplicated().sum()
```

[10]: 0

[11]:
```
test.duplicated().sum()
```

[11]: 0

[12]:
```
test.shape
```

[12]: (18812, 22)

[13]:
```
train.shape
```

[13]: (43894, 23)

[14]:
```
train.columns
```

[14]:
```
Index(['c_id', 'total_work_ex(yrs)', 'active_days_on_linkedin',
       'n_times_paid_promo', 'no_of_previous_positions', 'days_in_current_org',
       'age', 'reactions_on_each_post_per_second',
       'avg_angry_reaction_on_posts', 'avg_disgusted_reaction_on_posts',
       'avg_scared_reaction_on_posts', 'avg_happy_reaction_on_posts',
       'avg_neutral_reaction_on_posts', 'avg_sad_reaction_on_posts',
       'avg_surprise_reaction_on_posts', 'max_reactions_in_single_post',
       'profile_completion%', 'gender', 'country', 'display_picture_clarity',
       'total_posts', 'avg_reaction_in_each_post', 'n_followers'],
      dtype='object')
```

[15]:
```
test.columns
```

[15]:
```
Index(['c_id', 'total_work_ex(yrs)', 'active_days_on_linkedin',
       'n_times_paid_promo', 'no_of_previous_positions', 'days_in_current_org',
       'age', 'reactions_on_each_post_per_second',
       'avg_angry_reaction_on_posts', 'avg_disgusted_reaction_on_posts',
       'avg_scared_reaction_on_posts', 'avg_happy_reaction_on_posts',
       'avg_neutral_reaction_on_posts', 'avg_sad_reaction_on_posts',
       'avg_surprise_reaction_on_posts', 'max_reactions_in_single_post',
       'profile_completion%', 'gender', 'country', 'display_picture_clarity',
       'total_posts', 'avg_reaction_in_each_post'],
      dtype='object')
```

[16]:
```
train.isnull().sum()
```

[16]:
```
c_id                                0
total_work_ex(yrs)                  0
active_days_on_linkedin             0
n_times_paid_promo                  0
no_of_previous_positions            0
days_in_current_org                 0
age                              1316
reactions_on_each_post_per_second    1323
avg_angry_reaction_on_posts         1314
avg_disgusted_reaction_on_posts        0
avg_scared_reaction_on_posts         853
avg_happy_reaction_on_posts          870
avg_neutral_reaction_on_posts        842
avg_sad_reaction_on_posts              0
avg_surprise_reaction_on_posts         0
max_reactions_in_single_post        4217
profile_completion%                 4180
gender                              4098
country                                0
```

```
display_picture_clarity                0
total_posts                            0
avg_reaction_in_each_post              0
n_followers                            0
dtype: int64
```

[17]:
```
test.isnull().sum()
```

[17]:
```
c_id                                   0
total_work_ex(yrs)                     0
active_days_on_linkedin                0
n_times_paid_promo                     0
no_of_previous_positions               0
days_in_current_org                    0
age                                  538
reactions_on_each_post_per_second    532
avg_angry_reaction_on_posts          544
avg_disgusted_reaction_on_posts        0
avg_scared_reaction_on_posts         386
avg_happy_reaction_on_posts          377
avg_neutral_reaction_on_posts        394
avg_sad_reaction_on_posts              0
avg_surprise_reaction_on_posts         0
max_reactions_in_single_post        1755
profile_completion%                 1789
gender                              1855
country                                0
display_picture_clarity                0
total_posts                            0
avg_reaction_in_each_post              0
dtype: int64
```

[108]:
```
cat_col = train.select_dtypes(include=['object']).columns
num_col = train.select_dtypes(include=['number']).columns
num_col = num_col.drop('n_followers')

for col in num_col:
    train[col] = train[col].fillna(train[col].mean())
    test[col] = test[col].fillna(test[col].mean())
for col in cat_col:
    train[col] = train[col].fillna(train[col].mode().iloc[0])
    test[col] = test[col].fillna(test[col].mode().iloc[0])
```

[110]:
```
train.isnull().sum()
```

[110]:
```
c_id                                   0
total_work_ex(yrs)                     0
```

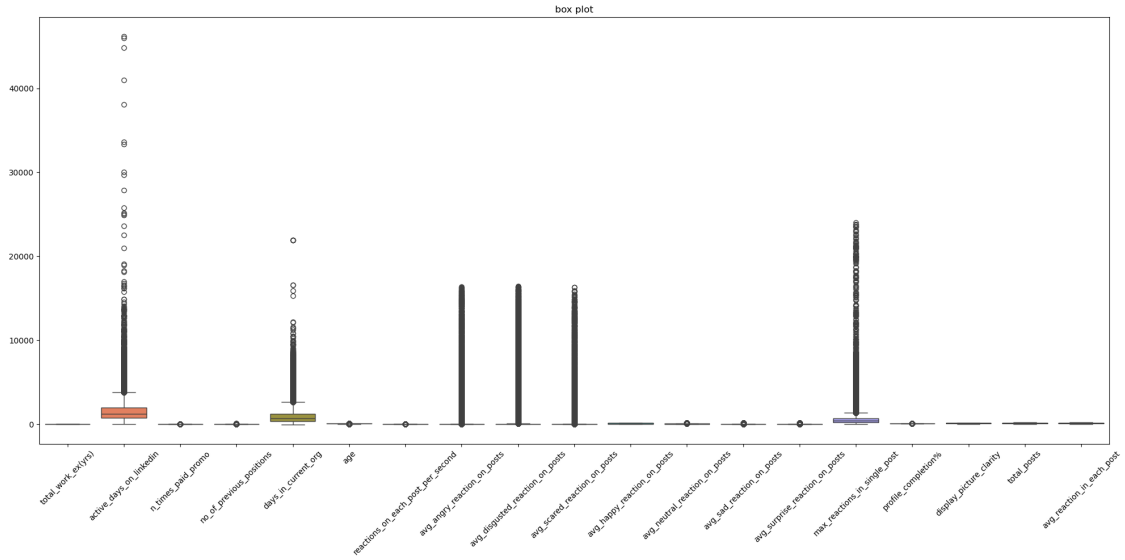
```
active_days_on_linkedin              0
n_times_paid_promo                   0
no_of_previous_positions             0
days_in_current_org                  0
age                                  0
reactions_on_each_post_per_second    0
avg_angry_reaction_on_posts          0
avg_disgusted_reaction_on_posts      0
avg_scared_reaction_on_posts         0
avg_happy_reaction_on_posts          0
avg_neutral_reaction_on_posts        0
avg_sad_reaction_on_posts            0
avg_surprise_reaction_on_posts       0
max_reactions_in_single_post         0
profile_completion%                  0
gender                               0
country                              0
display_picture_clarity              0
total_posts                          0
avg_reaction_in_each_post            0
n_followers                          0
dtype: int64
```

[20]:
```
test.isnull().sum()
```

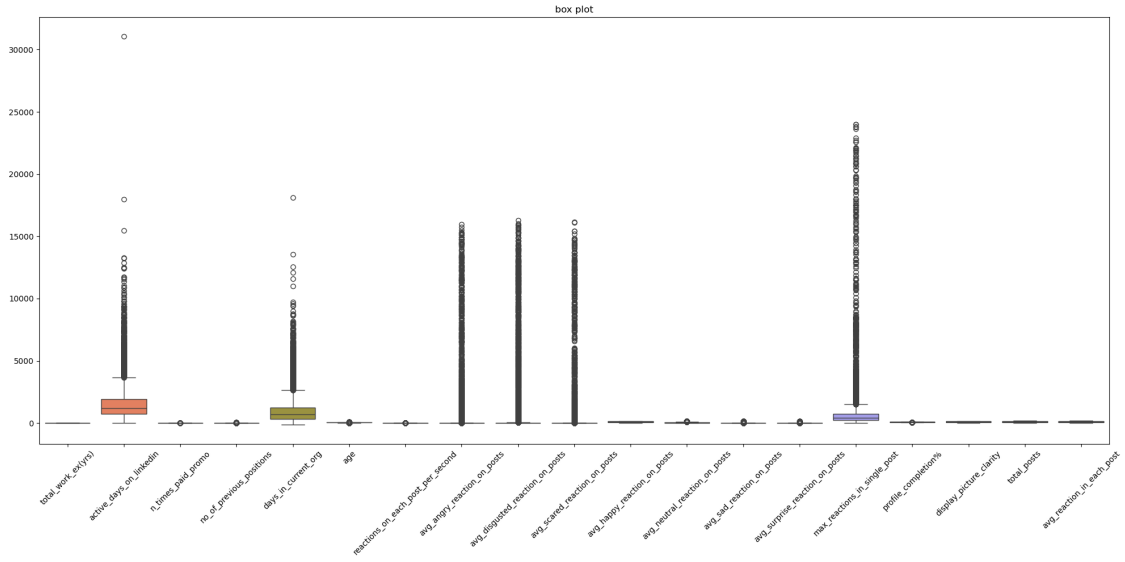
[20]:
```
c_id                                 0
total_work_ex(yrs)                   0
active_days_on_linkedin              0
n_times_paid_promo                   0
no_of_previous_positions             0
days_in_current_org                  0
age                                  0
reactions_on_each_post_per_second    0
avg_angry_reaction_on_posts          0
avg_disgusted_reaction_on_posts      0
avg_scared_reaction_on_posts         0
avg_happy_reaction_on_posts          0
avg_neutral_reaction_on_posts        0
avg_sad_reaction_on_posts            0
avg_surprise_reaction_on_posts       0
max_reactions_in_single_post         0
profile_completion%                  0
gender                               0
country                              0
display_picture_clarity              0
total_posts                          0
avg_reaction_in_each_post            0
```

```
dtype: int64
```

[21]:
```
# Boxplot for outlier
plt.figure(figsize=(20,10))
sns.boxplot(data=train[num_col])
plt.title('box plot')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



[22]:
```
# Boxplot for outlier
plt.figure(figsize=(20,10))
sns.boxplot(data=test[num_col])
plt.title('box plot')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



[23]:
```
def outlier_fix(df,num_col):
    for col in num_col:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        df[col] = df[col].clip(lower=lower_bound, upper = upper_bound)
    return df
```

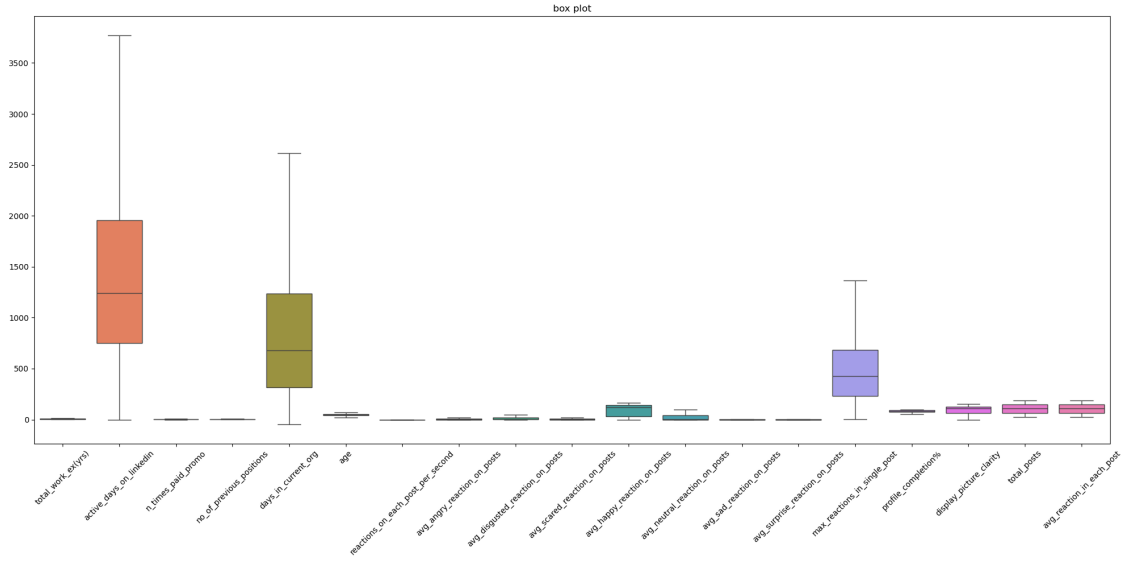
[24]:
```
outlier_fix(train, num_col)
outlier_fix(test, num_col)
```

[24]:
```
            c_id  total_work_ex(yrs)  active_days_on_linkedin  \
0      INF_43895                   6                  1460.75
1      INF_43896                  12                  1260.45
2      INF_43897                  14                  2208.22
3      INF_43898                   6                  1240.59
4      INF_43899                  12                  1210.51
…          …                     …                        …
18807  INF_62702                  10                  1192.31
18808  INF_62703                  14                   977.20
18809  INF_62704                  13                  1167.06
18810  INF_62705                  13                   828.51
18811  INF_62706                  10                  1256.03

       n_times_paid_promo  no_of_previous_positions  days_in_current_org  \
0                     4.0                       2.0                  357
```

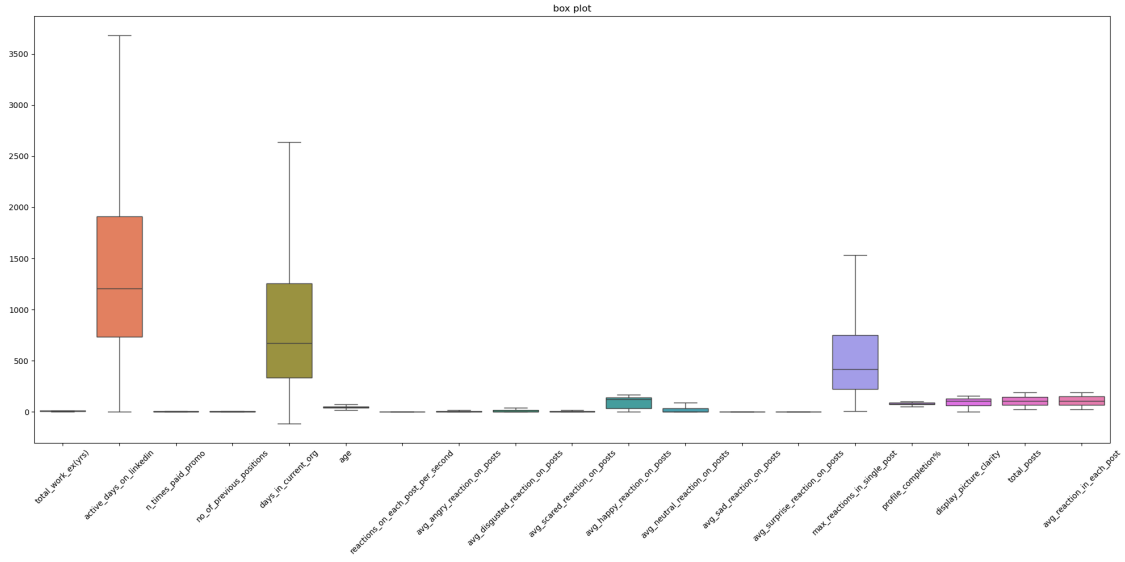
```
1                    4.0                    4.0                  109
2                    1.0                    1.0                  353
3                    4.0                    2.0                 1023
4                    3.0                    3.0                 2636
…                    …                      …                    …
18807                2.0                    6.0                 1033
18808                4.0                    7.0                  184
18809                3.0                    7.0                  131
18810                1.0                    9.0                 2167
18811                3.0                    9.5                  623

        age  reactions_on_each_post_per_second  avg_angry_reaction_on_posts  \
0      45.0                                0.01                         0.00
1      45.0                                0.01                         0.00
2      48.0                                0.01                         2.08
3      47.0                                0.01                         2.69
4      48.0                                0.01                         1.99
…      …                                   …                            …
18807  63.0                                0.00                         0.00
18808  62.0                                0.00                         0.00
18809  62.0                                0.00                         0.00
18810  64.0                                0.00                         0.00
18811  61.0                                0.00                         0.00

       avg_disgusted_reaction_on_posts  …  avg_neutral_reaction_on_posts  \
0                              0.00000  …                           0.00
1                              0.00000  …                           0.00
2                             37.51375  …                           0.02
3                             37.51375  …                           0.01
4                             37.51375  …                           0.01
…                                …  …                            …
18807                          0.14000  …                           0.00
18808                          0.13000  …                           0.00
18809                          0.16000  …                           0.00
18810                          0.16000  …                           0.00
18811                          0.15000  …                           0.00

       avg_sad_reaction_on_posts  avg_surprise_reaction_on_posts  \
0                          0.000                            0.00
1                          0.000                            0.00
2                          0.225                            0.01
3                          0.225                            0.01
4                          0.225                            0.01
…                          …                                …
18807                      0.000                            0.00
18808                      0.000                            0.00
18809                      0.000                            0.00
```

```
18810                      0.000                           0.00
18811                      0.000                           0.00

       max_reactions_in_single_post  profile_completion%  gender  country  \
0                             208.0            76.000000    Male    China
1                             259.0            86.000000    Male       UK
2                             113.0            90.000000    Male    India
3                             132.0            90.000000    Male    India
4                             150.0            80.000000    Male   France
…                             …                  …     …      …
18807                         462.0            81.000000    Male   France
18808                         230.0            81.419374    Male       UK
18809                         407.0            68.000000    Male  Germany
18810                         406.0            89.000000    Male  Ireland
18811                         444.0            79.000000    Male      USA

      display_picture_clarity total_posts  avg_reaction_in_each_post
0                      109.74         134                        130
1                      115.84          99                        124
2                      125.02         113                        188
3                      127.49         147                        143
4                      123.60          93                        149
…                       …          …                        …
18807                  118.13          80                         27
18808                   97.56          38                        160
18809                  110.58         117                         50
18810                  109.09          66                         99
18811                  100.88         103                         55

[18812 rows x 22 columns]
```

[25]:
```
# Boxplot for outlier
plt.figure(figsize=(20,10))
sns.boxplot(data=train[num_col])
plt.title('box plot')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



[26]:
```
# Boxplot for outlier
plt.figure(figsize=(20,10))
sns.boxplot(data=test[num_col])
plt.title('box plot')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



[27]:
```
train[num_col].skew()
```

[27]:
```
total_work_ex(yrs)                  -0.004805
active_days_on_linkedin              0.874379
n_times_paid_promo                   0.193981
no_of_previous_positions             0.928576
days_in_current_org                  1.050756
age                                 -0.030480
reactions_on_each_post_per_second    0.741489
avg_angry_reaction_on_posts          1.178742
avg_disgusted_reaction_on_posts      1.177670
avg_scared_reaction_on_posts         1.186604
avg_happy_reaction_on_posts         -0.710622
avg_neutral_reaction_on_posts        1.175990
avg_sad_reaction_on_posts            1.180932
avg_surprise_reaction_on_posts       1.205706
max_reactions_in_single_post         0.909206
profile_completion%                 -0.595119
display_picture_clarity             -0.911891
total_posts                          0.002475
avg_reaction_in_each_post            0.003282
dtype: float64
```

[28]:
```
test[num_col].skew()
```

[28]:
```
total_work_ex(yrs)                   0.010541
active_days_on_linkedin              0.888648
n_times_paid_promo                   0.199051
no_of_previous_positions             0.914380
days_in_current_org                  1.063780
age                                 -0.023224
reactions_on_each_post_per_second    0.760288
avg_angry_reaction_on_posts          1.198114
avg_disgusted_reaction_on_posts      1.179460
avg_scared_reaction_on_posts         1.184112
avg_happy_reaction_on_posts         -0.745141
avg_neutral_reaction_on_posts        1.170727
avg_sad_reaction_on_posts            1.201402
avg_surprise_reaction_on_posts       1.151897
max_reactions_in_single_post         1.042362
profile_completion%                 -0.620365
display_picture_clarity             -0.862337
total_posts                          0.014811
avg_reaction_in_each_post           -0.006293
dtype: float64
```

[58]:
```
train['active_days_on_linkedin'] = np.log1p(train['active_days_on_linkedin'])
train['no_of_previous_positions'] = np.log1p(train['no_of_previous_positions'])
train['days_in_current_org'] = np.log1p(train['days_in_current_org'])
```

```
train['reactions_on_each_post_per_second'] = np.
 ↪log1p(train['reactions_on_each_post_per_second'])
train['avg_angry_reaction_on_posts'] = np.
 ↪log1p(train['avg_angry_reaction_on_posts'])
train['avg_disgusted_reaction_on_posts'] = np.
 ↪log1p(train['avg_disgusted_reaction_on_posts'])
train['avg_scared_reaction_on_posts'] = np.
 ↪log1p(train['avg_scared_reaction_on_posts'])
train['avg_happy_reaction_on_posts'] = np.
 ↪log1p(train['avg_happy_reaction_on_posts'])
train['avg_neutral_reaction_on_posts'] = np.
 ↪log1p(train['avg_neutral_reaction_on_posts'])
train['avg_sad_reaction_on_posts'] = np.
 ↪log1p(train['avg_sad_reaction_on_posts'])
train['avg_surprise_reaction_on_posts'] = np.
 ↪log1p(train['avg_surprise_reaction_on_posts'])
train['max_reactions_in_single_post'] = np.
 ↪log1p(train['max_reactions_in_single_post'])
train['display_picture_clarity'] = np.log1p(train['display_picture_clarity'])
```

[60]:
```
train[num_col].skew()
```

[60]:
```
total_work_ex(yrs)                 -0.004805
active_days_on_linkedin            -1.682173
n_times_paid_promo                  0.193981
no_of_previous_positions            0.091453
days_in_current_org                -0.588112
age                                -0.030480
reactions_on_each_post_per_second   0.718989
avg_angry_reaction_on_posts         0.709917
avg_disgusted_reaction_on_posts     0.609553
avg_scared_reaction_on_posts        0.711541
avg_happy_reaction_on_posts        -1.367861
avg_neutral_reaction_on_posts       0.710750
avg_sad_reaction_on_posts           1.154325
avg_surprise_reaction_on_posts      1.177184
max_reactions_in_single_post       -0.919236
profile_completion%                -0.595119
display_picture_clarity            -1.782071
total_posts                         0.002475
avg_reaction_in_each_post           0.003282
dtype: float64
```

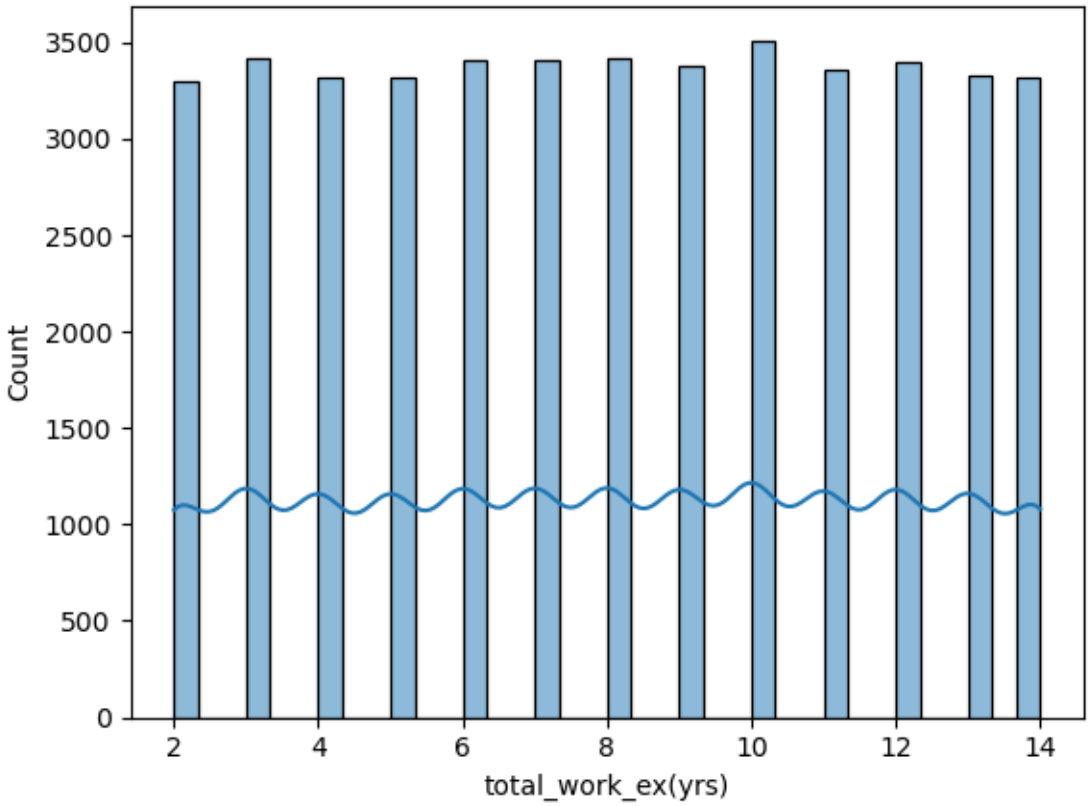
[62]:
```
test['active_days_on_linkedin'] = np.log1p(test['active_days_on_linkedin'])
test['no_of_previous_positions'] = np.log1p(test['no_of_previous_positions'])
test['days_in_current_org'] = np.log1p(test['days_in_current_org'])
```

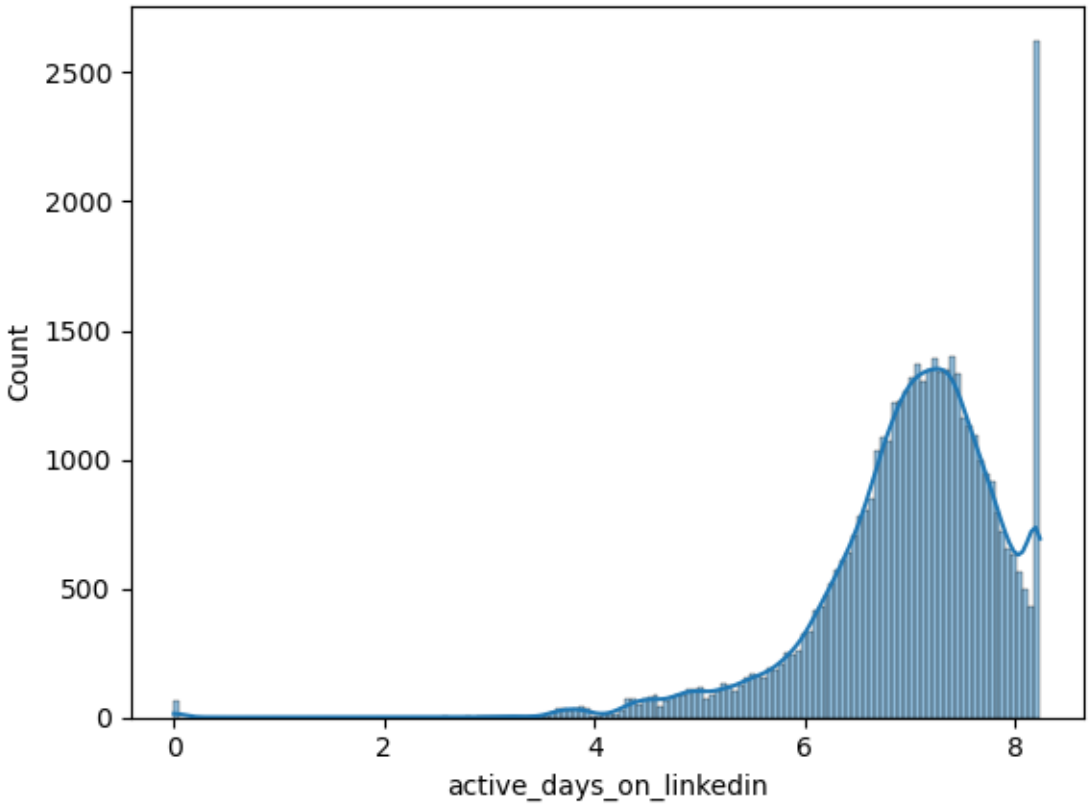
```
test['reactions_on_each_post_per_second'] = np.
 ↪log1p(test['reactions_on_each_post_per_second'])
test['avg_angry_reaction_on_posts'] = np.
 ↪log1p(test['avg_angry_reaction_on_posts'])
test['avg_disgusted_reaction_on_posts'] = np.
 ↪log1p(test['avg_disgusted_reaction_on_posts'])
test['avg_scared_reaction_on_posts'] = np.
 ↪log1p(test['avg_scared_reaction_on_posts'])
test['avg_happy_reaction_on_posts'] = np.
 ↪log1p(test['avg_happy_reaction_on_posts'])
test['avg_neutral_reaction_on_posts'] = np.
 ↪log1p(test['avg_neutral_reaction_on_posts'])
test['avg_sad_reaction_on_posts'] = np.log1p(test['avg_sad_reaction_on_posts'])
test['avg_surprise_reaction_on_posts'] = np.
 ↪log1p(test['avg_surprise_reaction_on_posts'])
test['max_reactions_in_single_post'] = np.
 ↪log1p(test['max_reactions_in_single_post'])
test['display_picture_clarity'] = np.log1p(test['display_picture_clarity'])
```

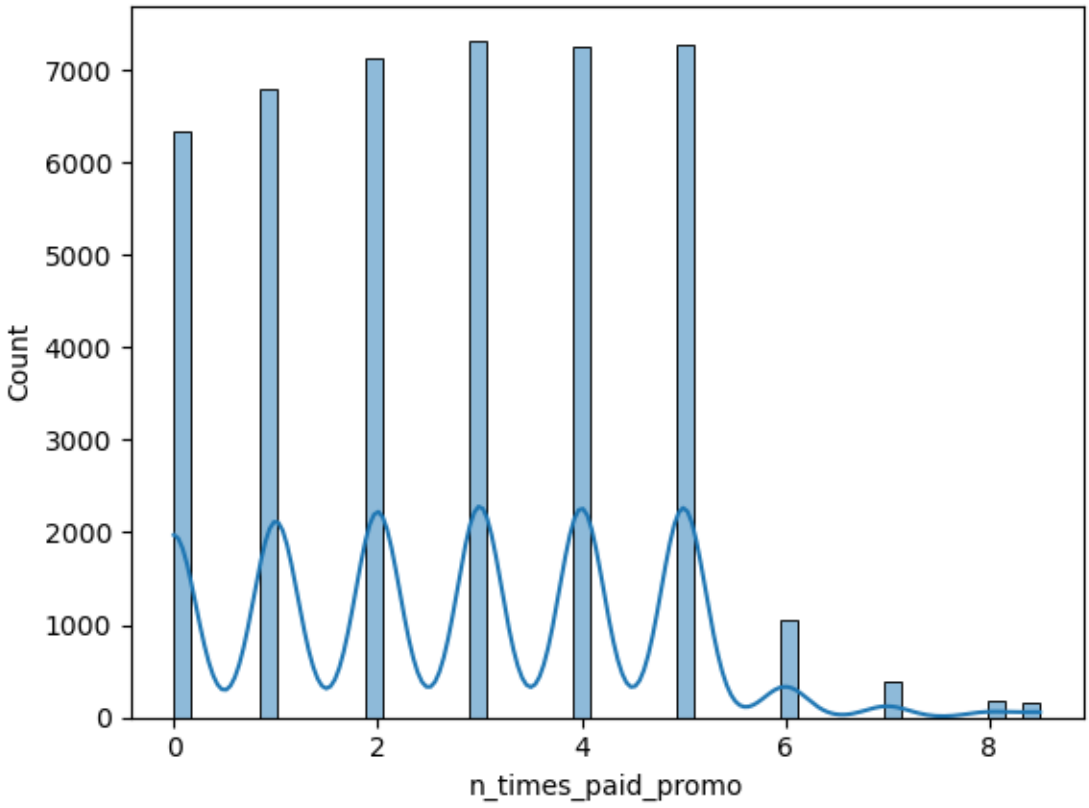
[64]:
```
test[num_col].skew()
```

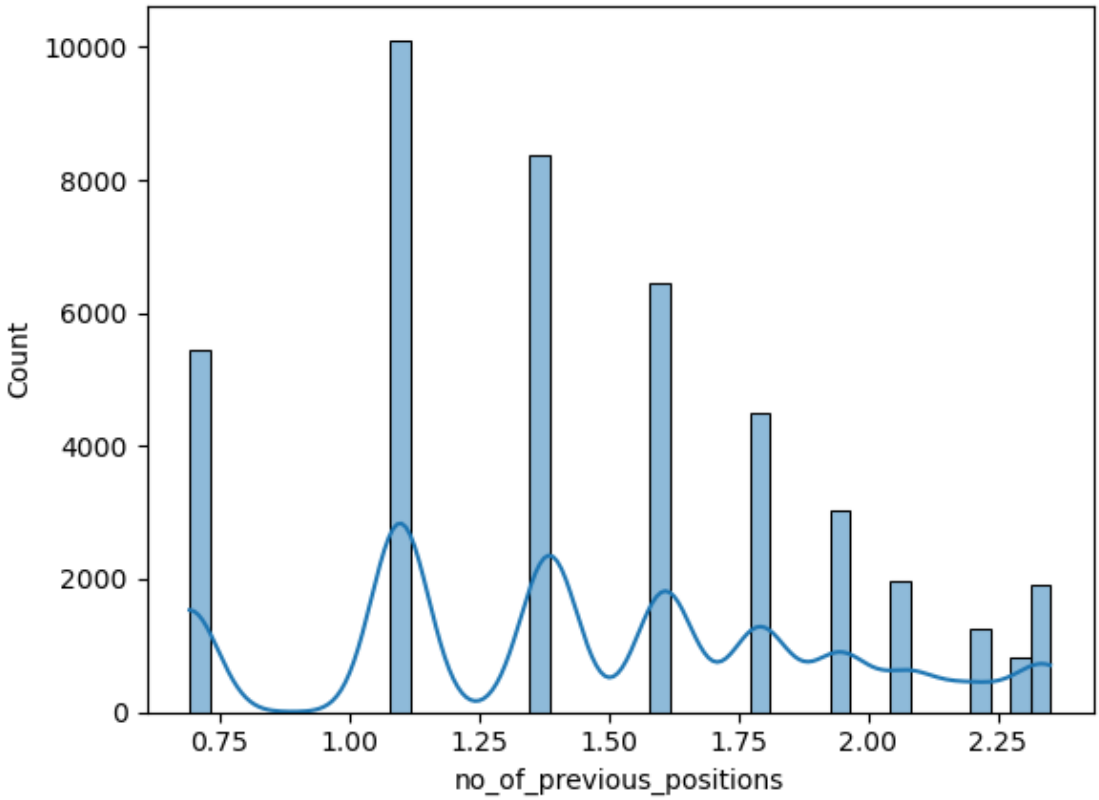
[64]:
```
total_work_ex(yrs)                  0.010541
active_days_on_linkedin            -1.686335
n_times_paid_promo                  0.199051
no_of_previous_positions            0.088485
days_in_current_org                -0.532307
age                                -0.023224
reactions_on_each_post_per_second   0.738614
avg_angry_reaction_on_posts         0.735906
avg_disgusted_reaction_on_posts     0.626696
avg_scared_reaction_on_posts        0.720271
avg_happy_reaction_on_posts        -1.406050
avg_neutral_reaction_on_posts       0.736054
avg_sad_reaction_on_posts           1.176649
avg_surprise_reaction_on_posts      1.130706
max_reactions_in_single_post       -0.854427
profile_completion%                -0.620365
display_picture_clarity            -1.674399
total_posts                         0.014811
avg_reaction_in_each_post          -0.006293
dtype: float64
```

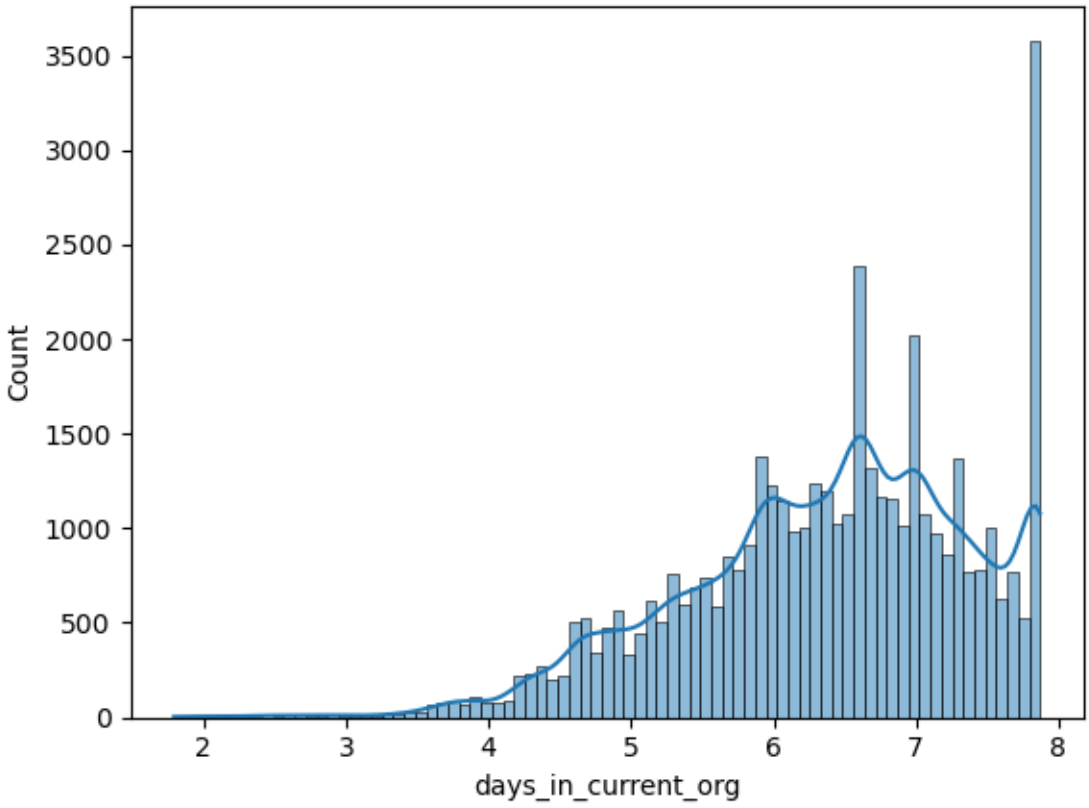
[90]:
```
for col in num_col:
    sns.histplot(data = train[col], kde =True)
    plt.show()
```

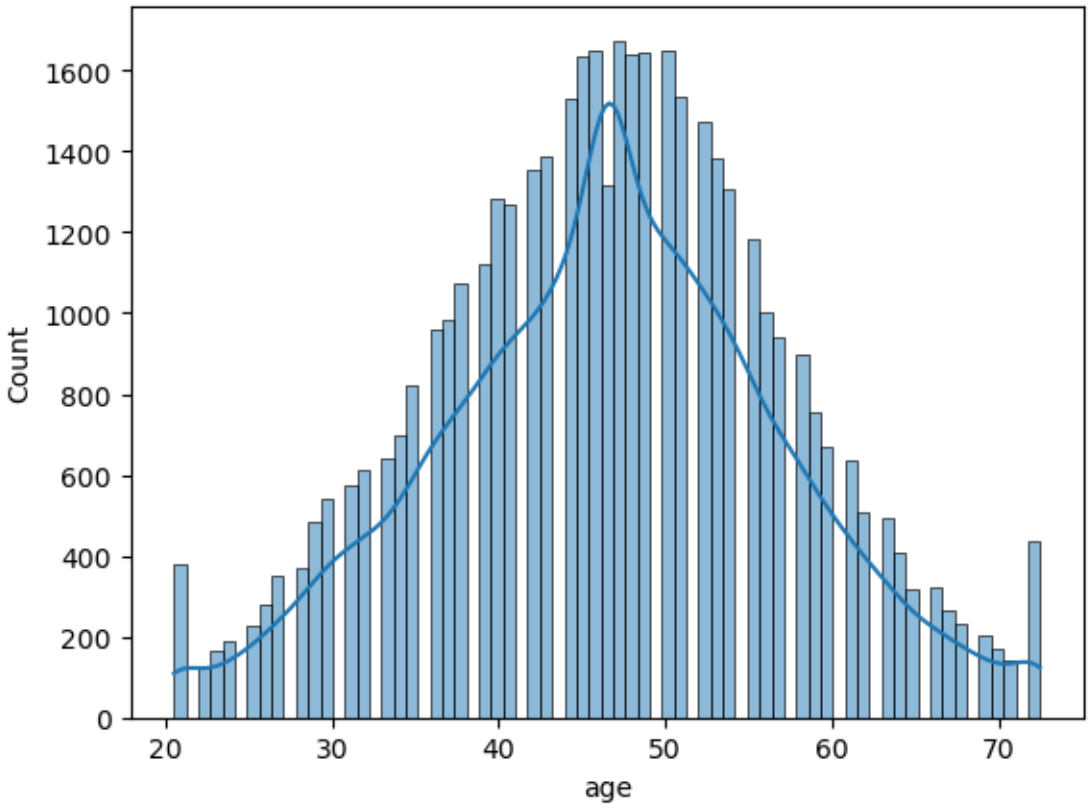

3500
3000
2500
2000
1500
1000
500
0
Count
2
4
6
8
10
12
14
total_work_ex(yrs)

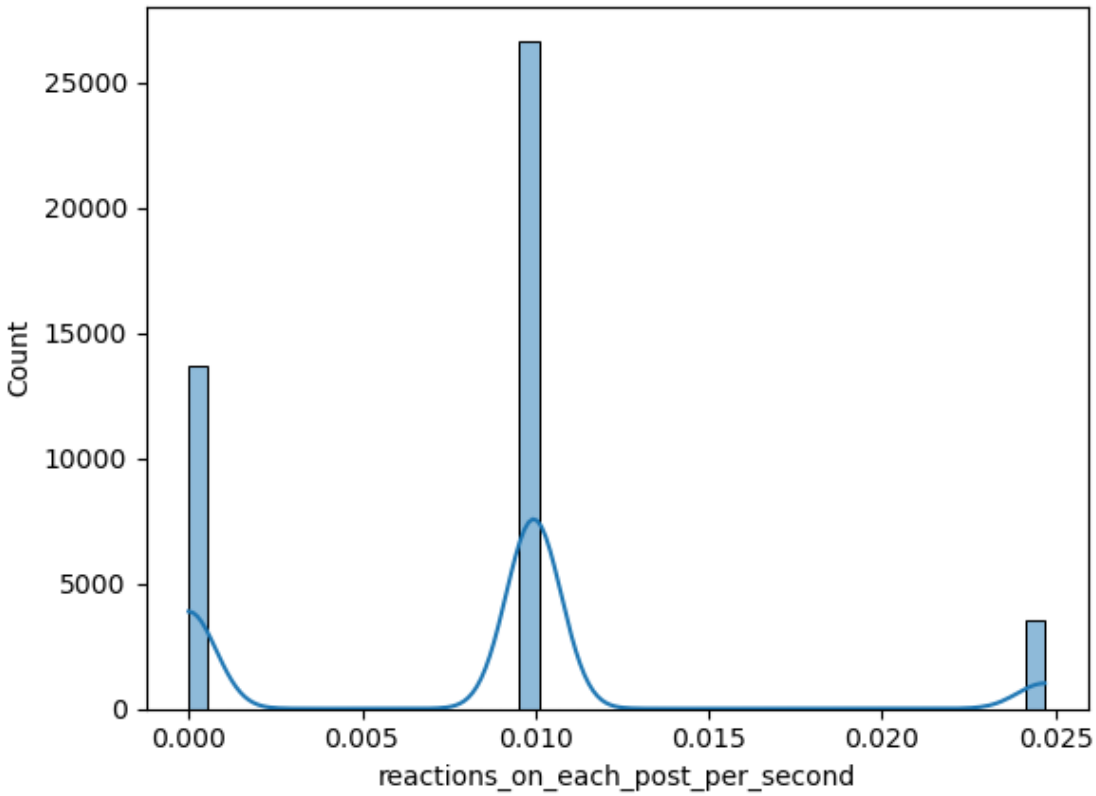

2500
2000
1500
1000
500
0
Count
0
2
4
6
8
active_days_on_linkedin

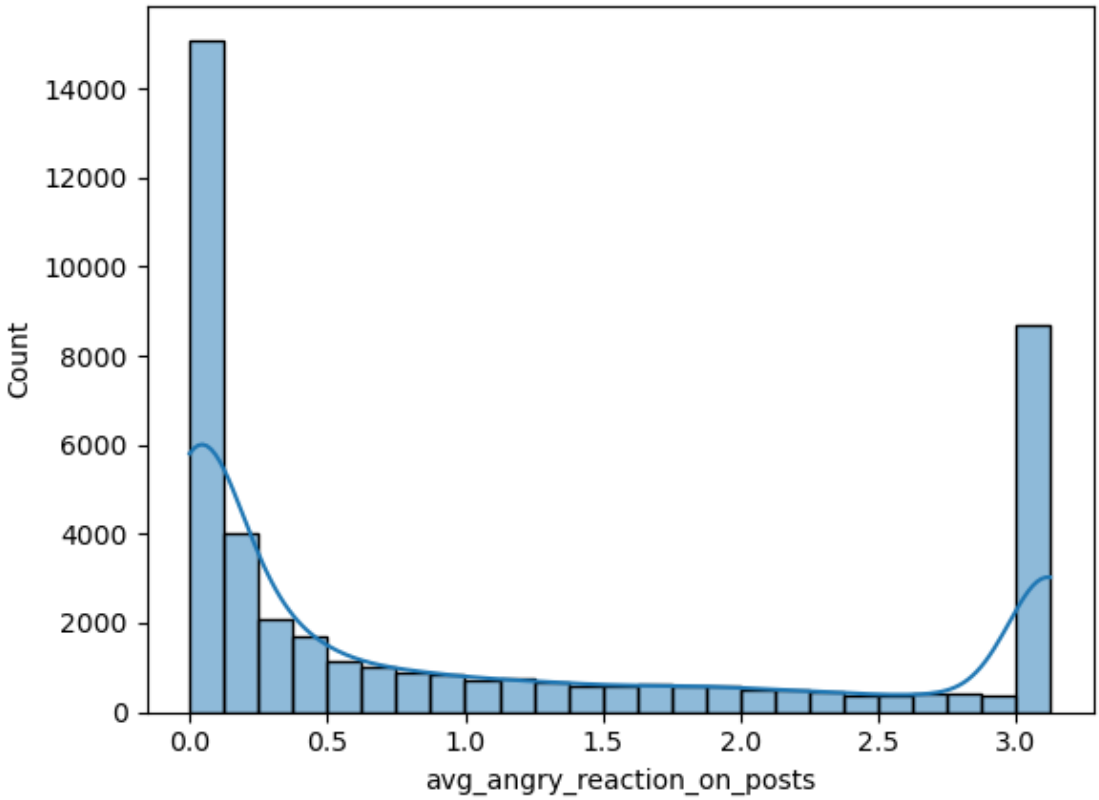

Count
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
n_times_paid_promo

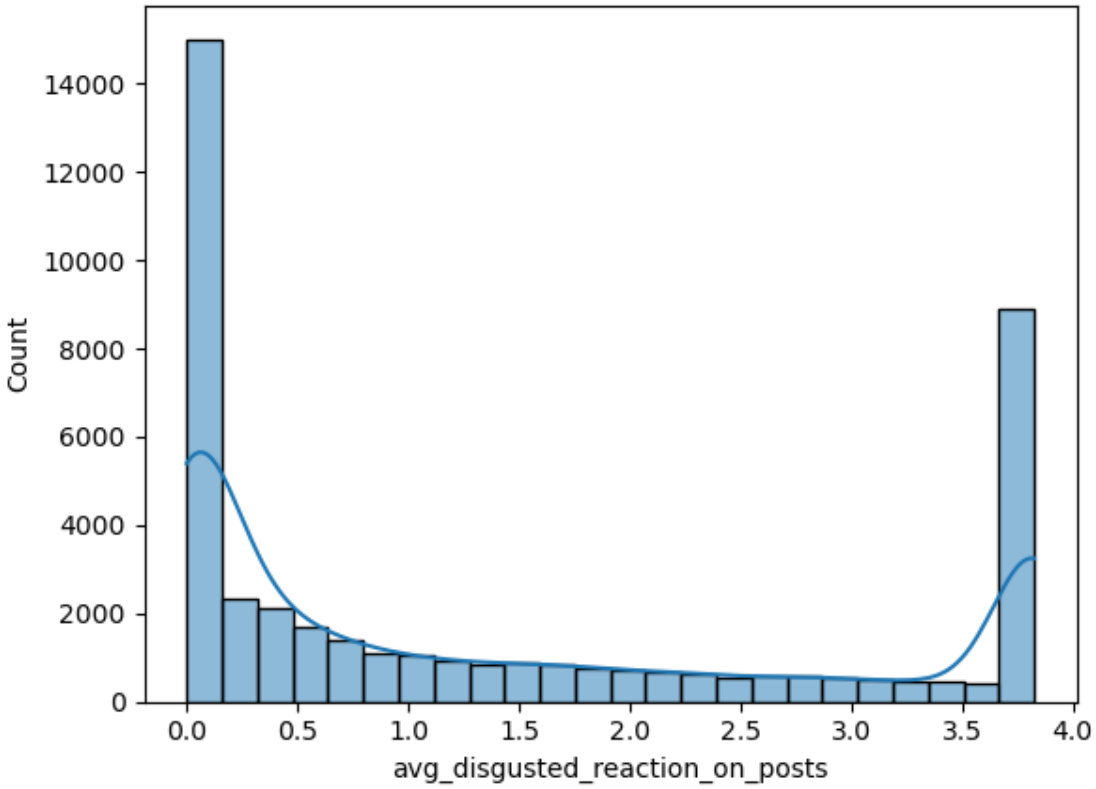

10000
8000
6000
4000
2000
0
Count
0.75
1.00
1.25
1.50
1.75
2.00
2.25
no_of_previous_positions

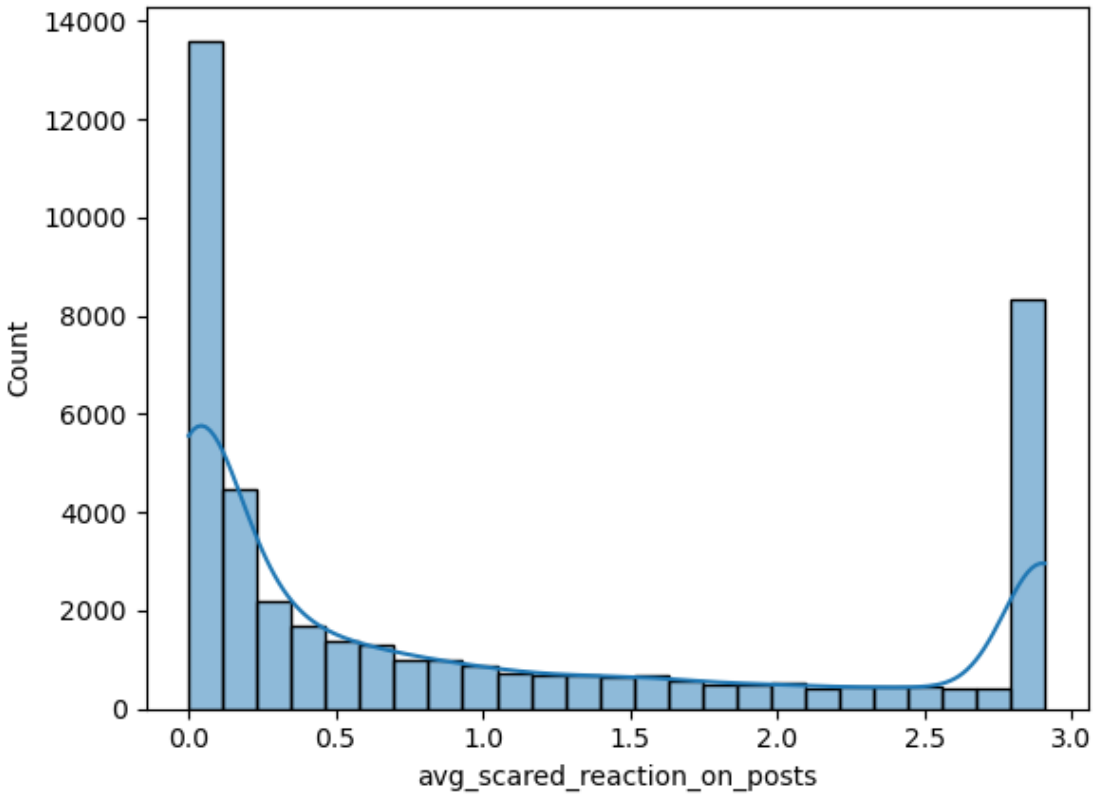

3500
3000
2500
2000
1500
1000
500
0
Count
2
3
4
5
6
7
8
days_in_current_org

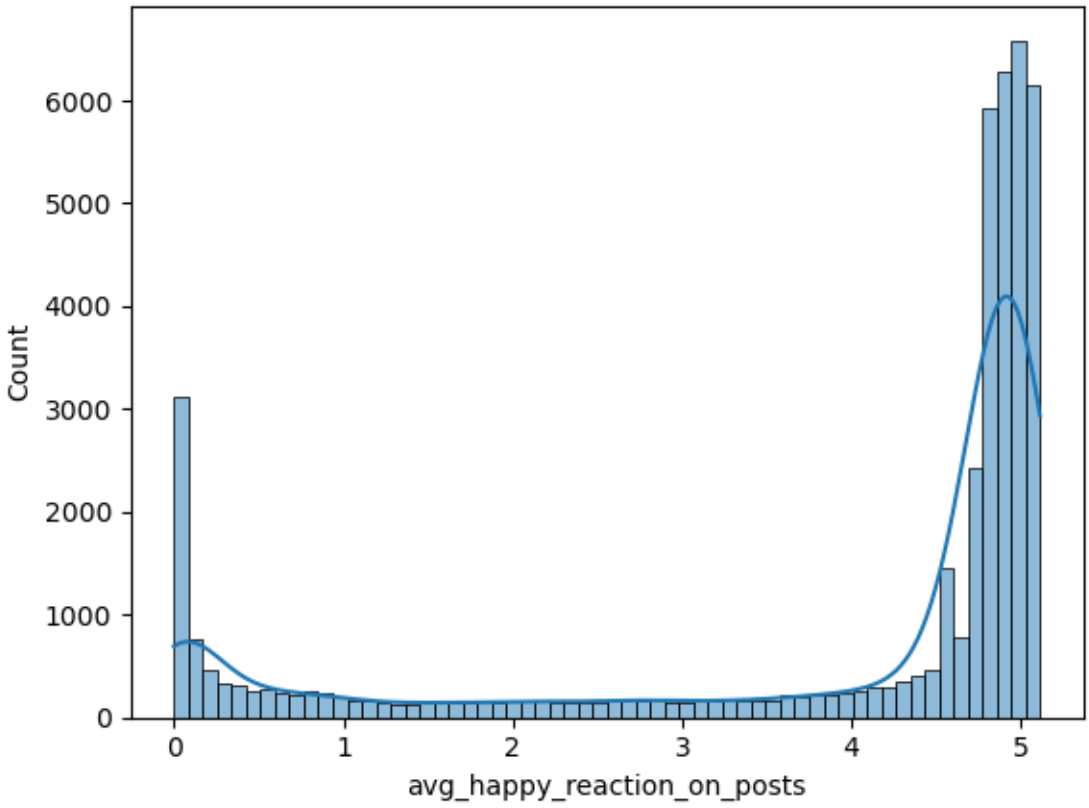

Count
1600
1400
1200
1000
800
600
400
200
0
20
30
40
50
60
70
age

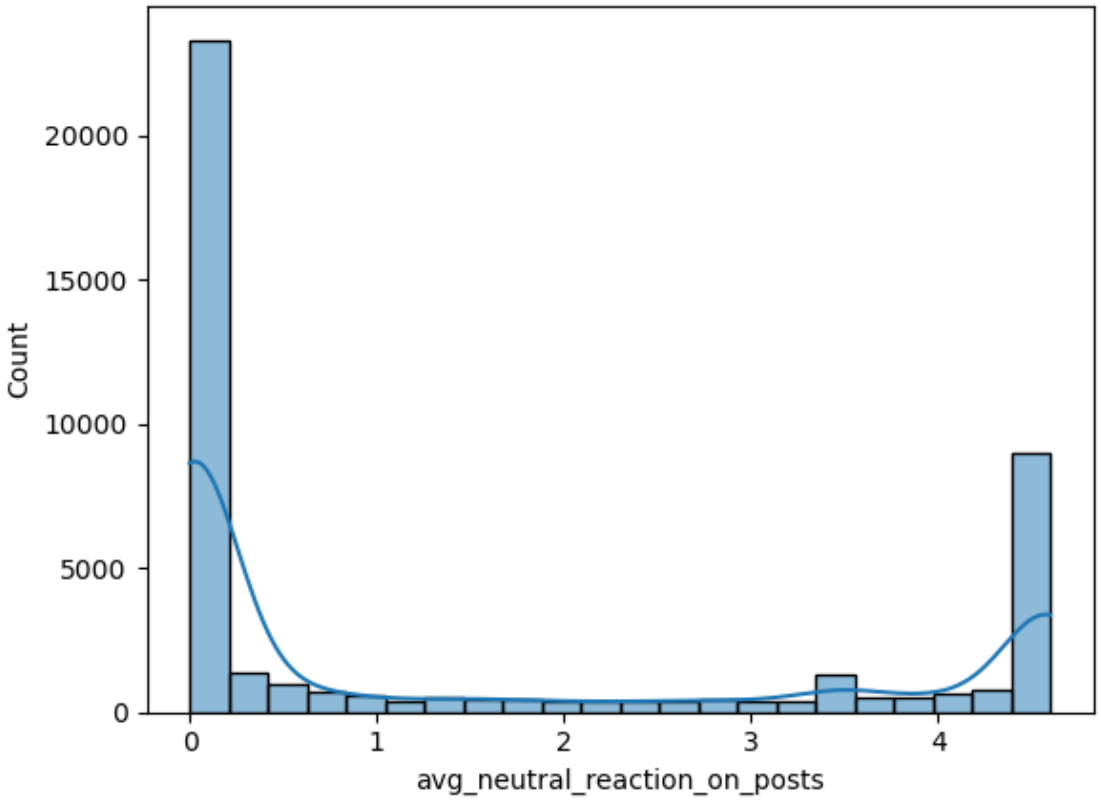

25000
20000
15000
10000
5000
0
Count
0.000
0.005
0.010
0.015
0.020
0.025
reactions_on_each_post_per_second

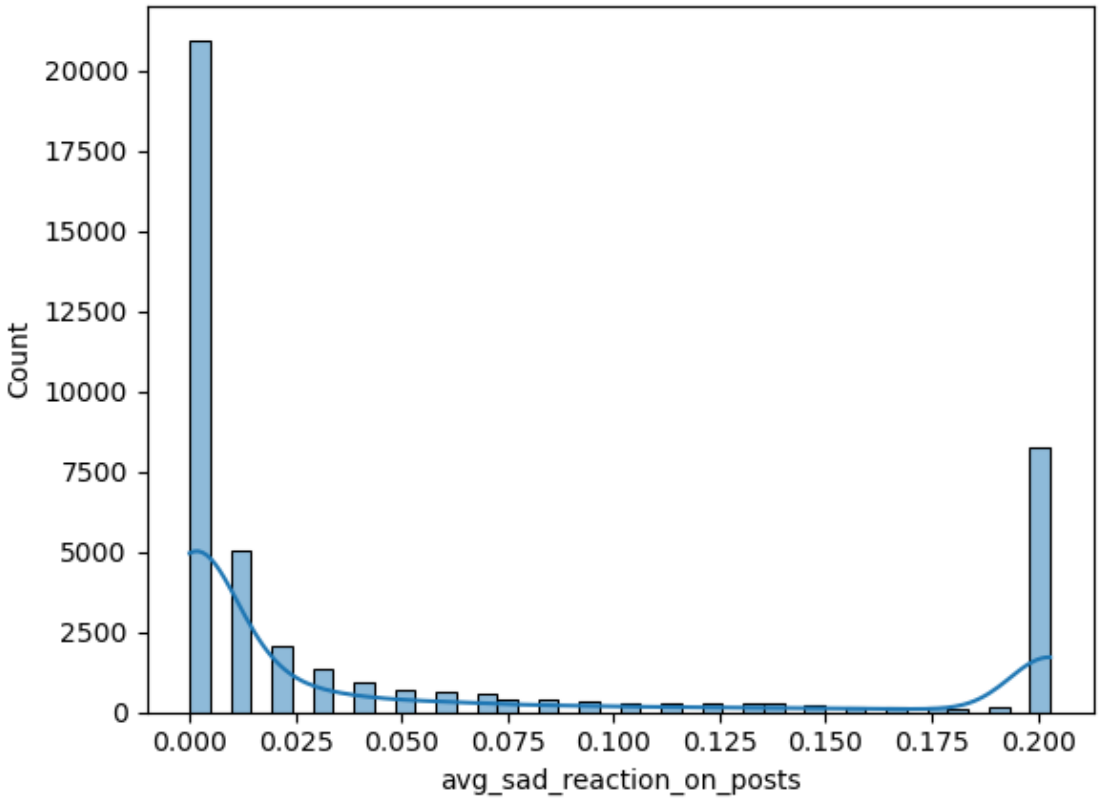

14000
12000
10000
8000
6000
4000
2000
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
avg_angry_reaction_on_posts

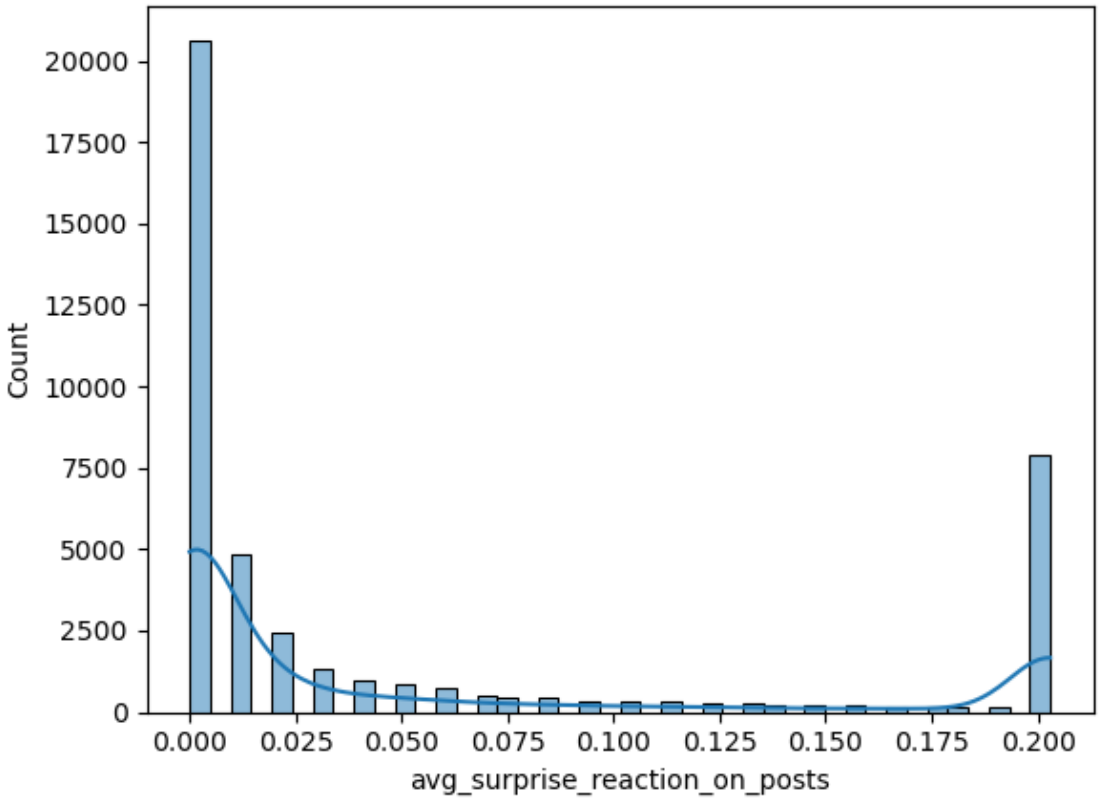

14000
12000
10000
8000
6000
4000
2000
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
avg_disgusted_reaction_on_posts

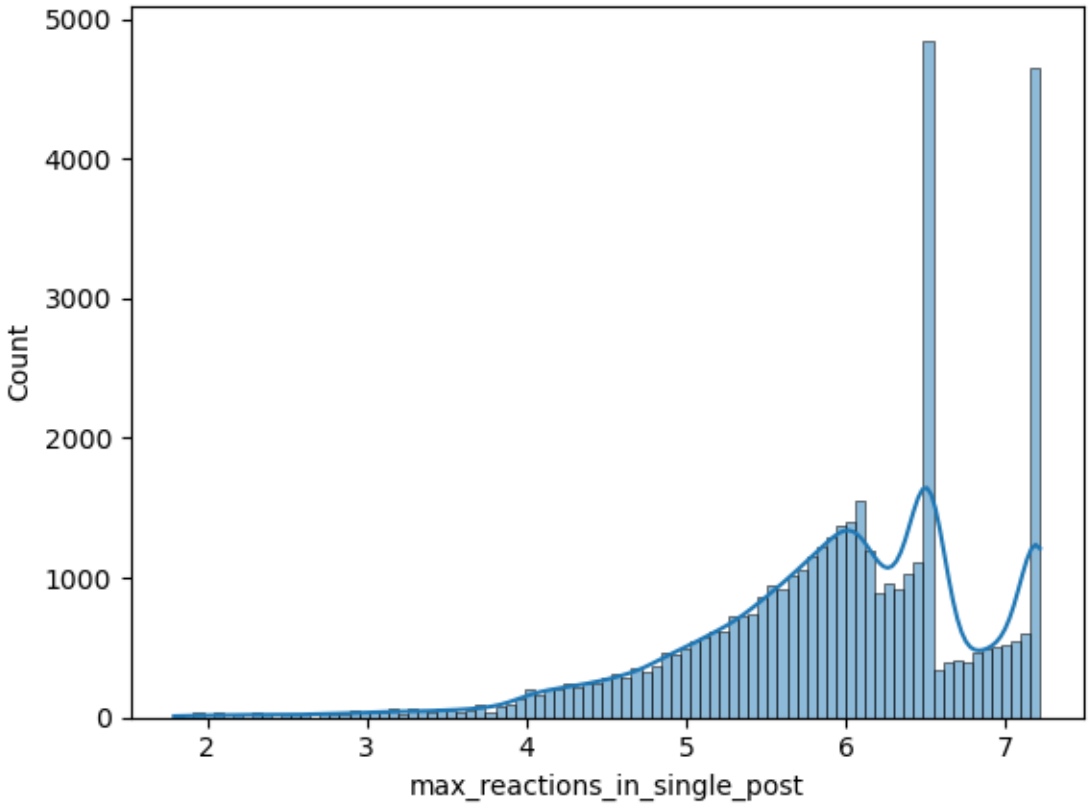

14000
12000
10000
8000
6000
4000
2000
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
avg_scared_reaction_on_posts

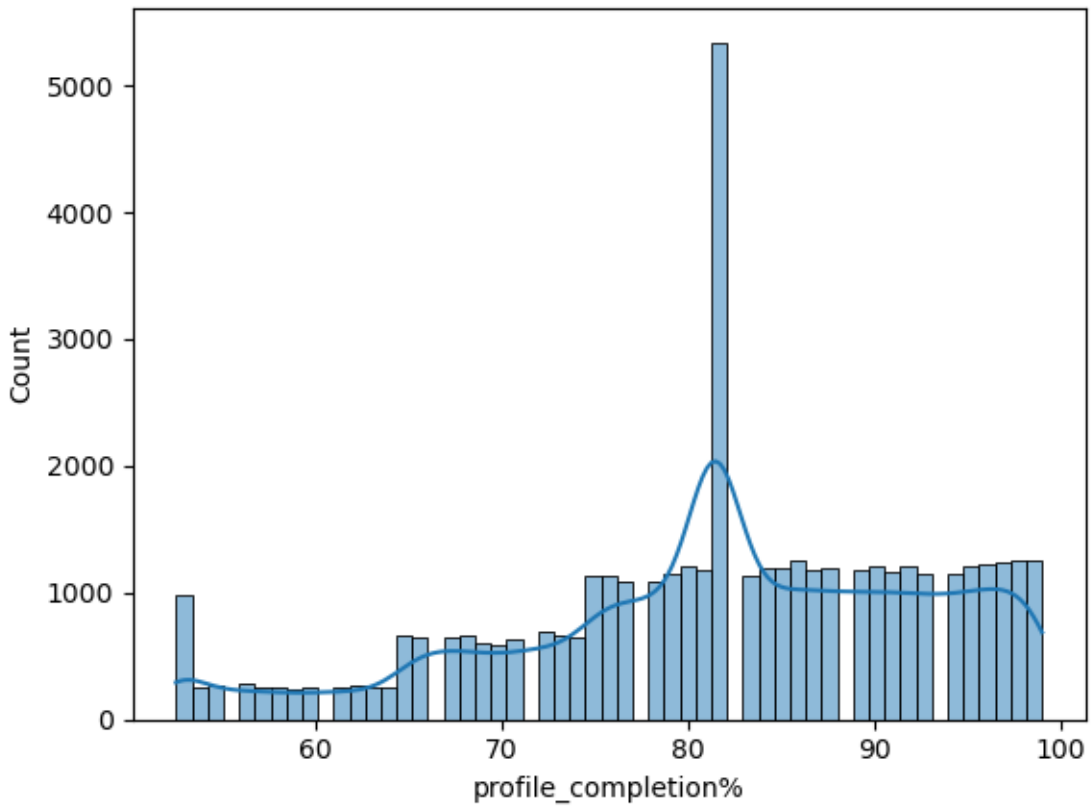

6000
5000
4000
3000
2000
1000
0
Count
0
1
2
3
4
5
avg_happy_reaction_on_posts

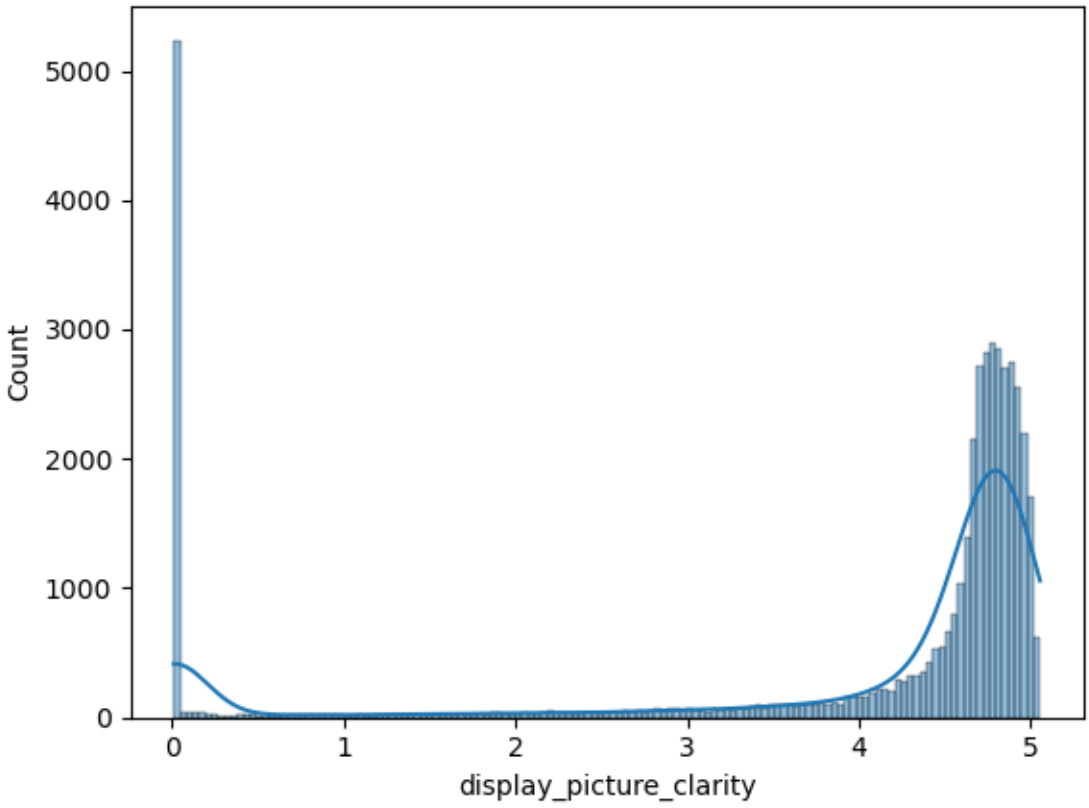

20000
15000
10000
5000
0
Count
0
1
2
3
4
avg_neutral_reaction_on_posts

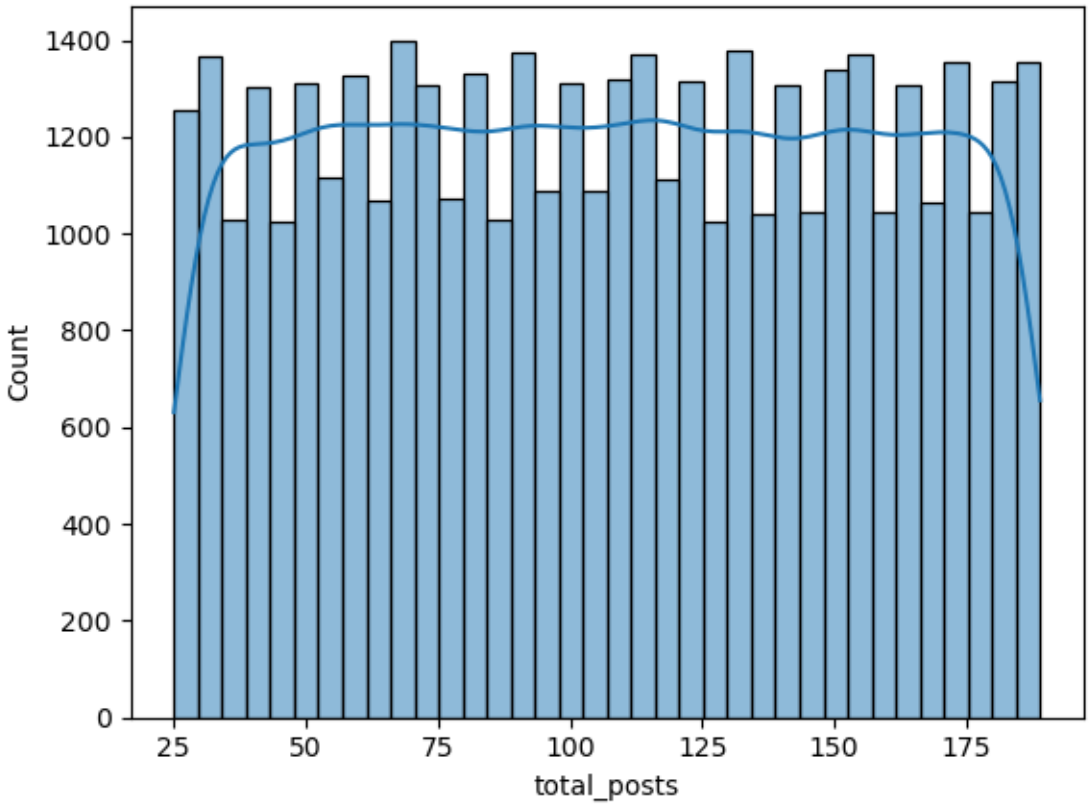

Count
20000
17500
15000
12500
10000
7500
5000
2500
0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
avg_sad_reaction_on_posts

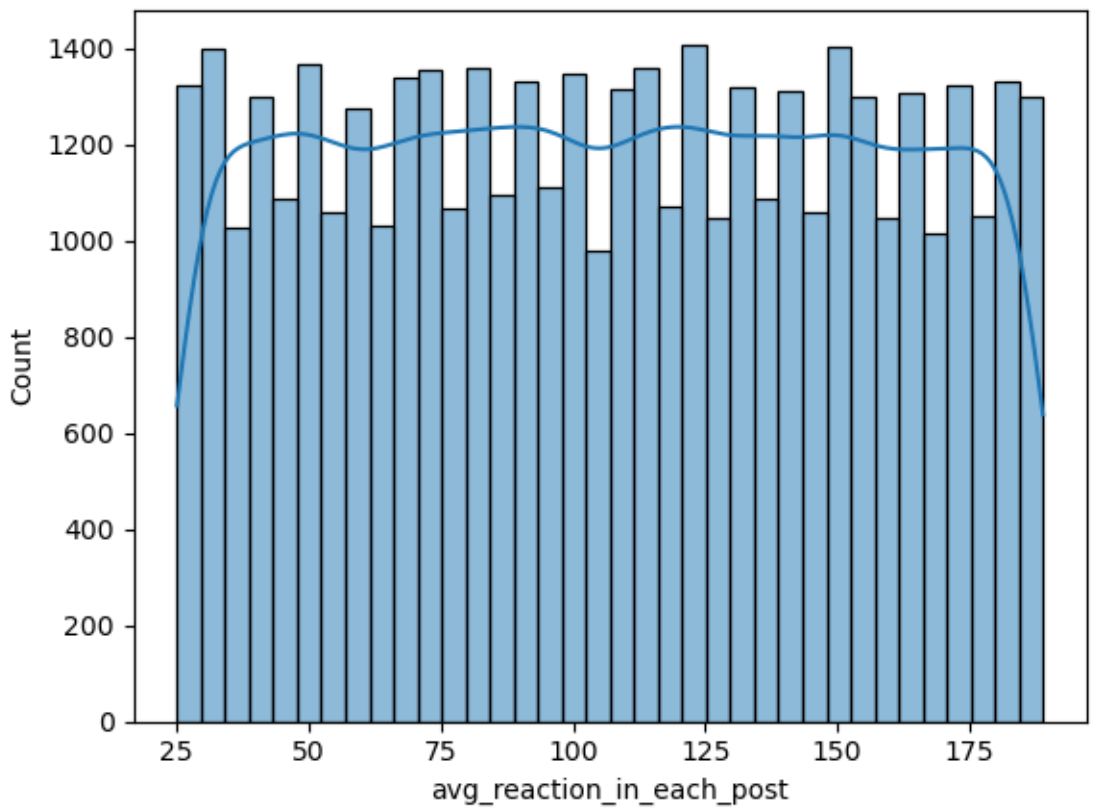

20000
17500
15000
12500
10000
7500
5000
2500
0
Count
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
avg_surprise_reaction_on_posts


5000
4000
3000
2000
1000
0
Count
2
3
4
5
6
7
max_reactions_in_single_post


5000
4000
3000
2000
1000
0
Count
60
70
80
90
100
profile_completion%

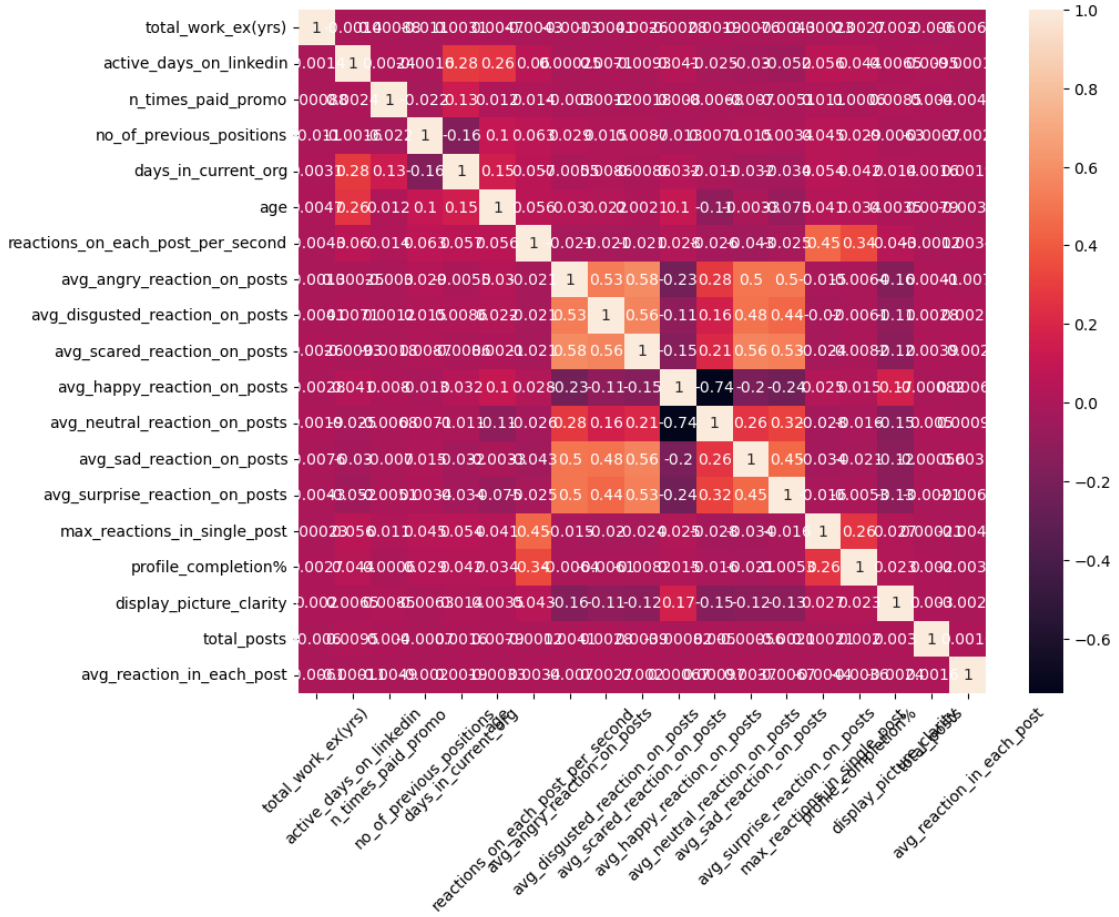

5000
4000
3000
2000
1000
0
Count
0
1
2
3
4
5
display_picture_clarity


1400
1200
1000
800
600
400
200
0
Count
25
50
75
100
125
150
175
total_posts



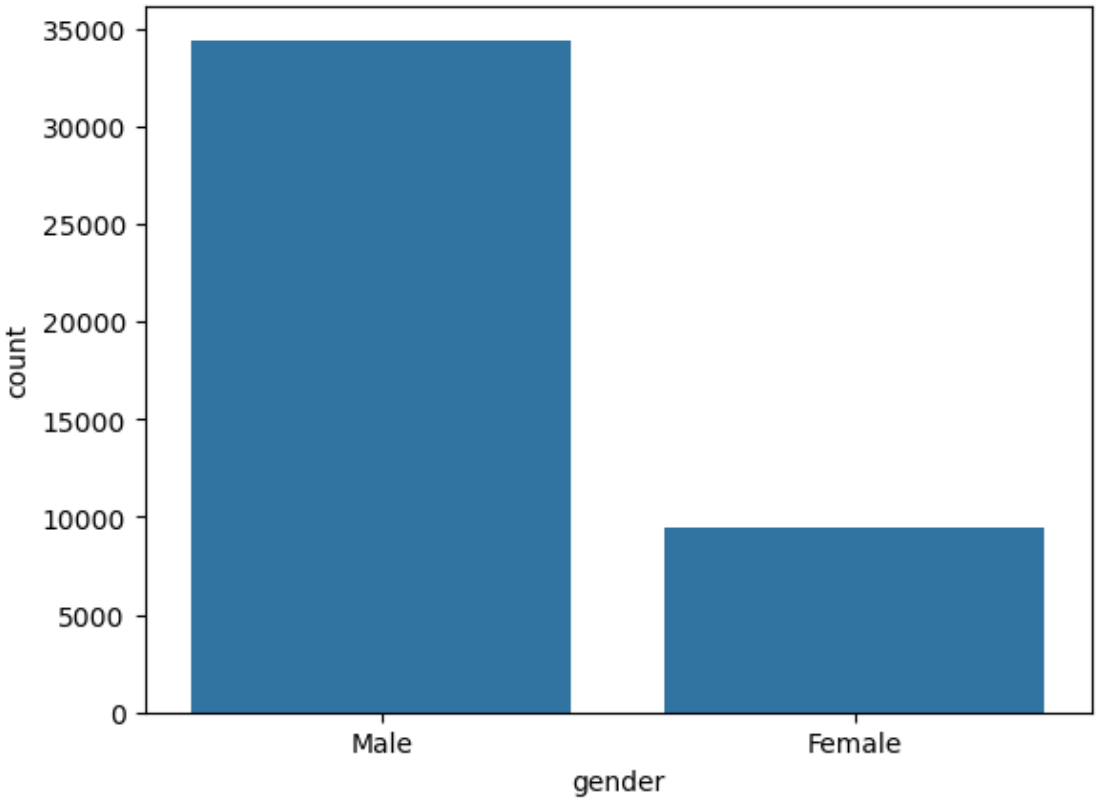
[68]:
```
# Visualisation
corr = train[num_col].corr()
```

[74]:
```
plt.figure(figsize=(10,8))
sns.heatmap(corr, annot=True)
plt.xticks(rotation=45)
plt.show()
```



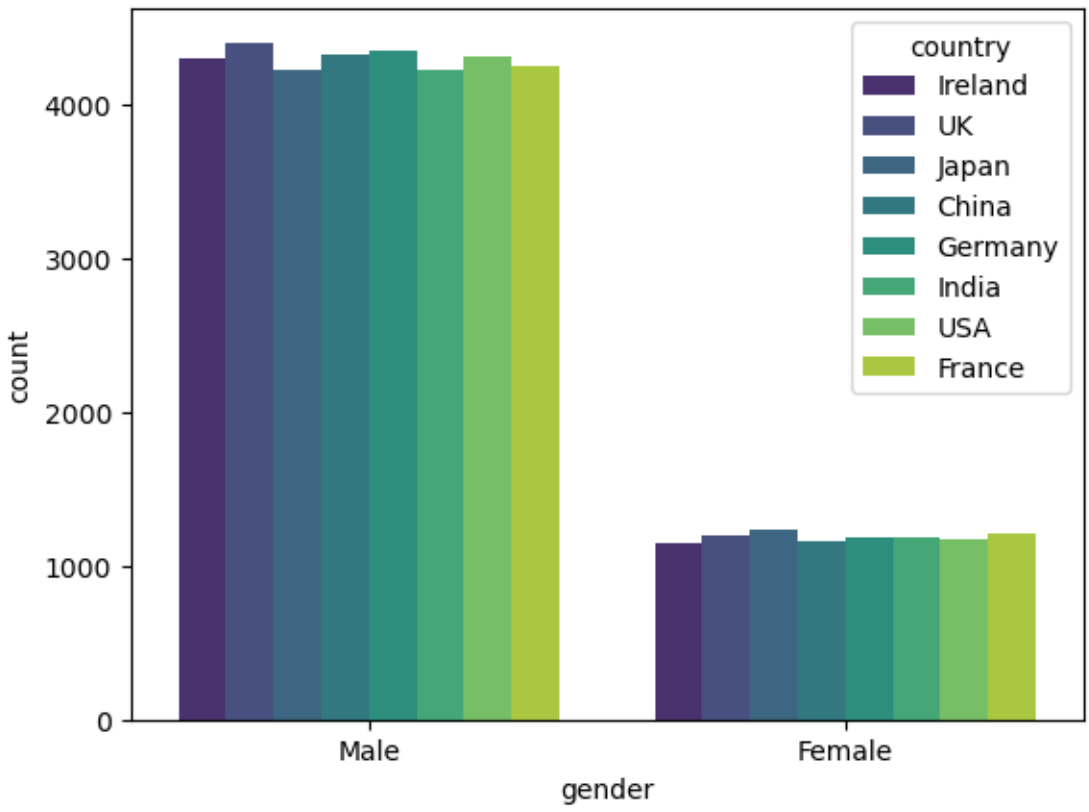
[76]:
```
sns.countplot(data=train, x= 'gender')
```

[76]: <Axes: xlabel='gender', ylabel='count'>



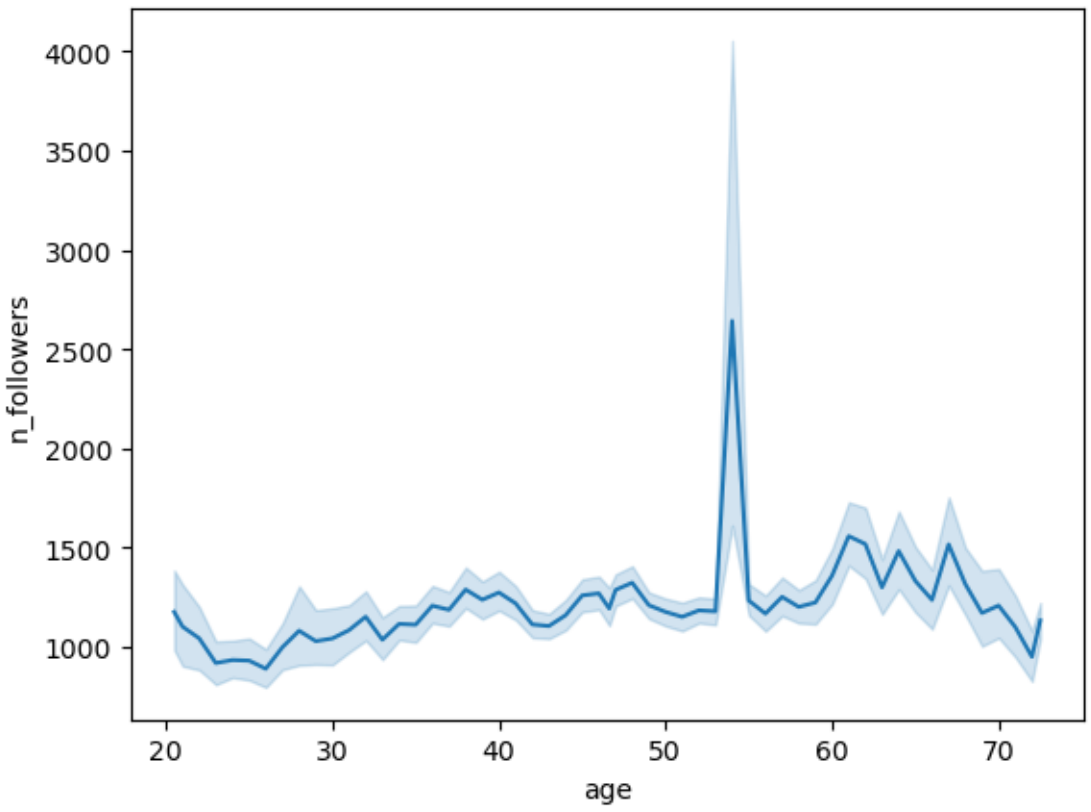
[78]:
```
sns.countplot(data=train, x='gender',hue='country', palette='viridis')
```

[78]: <Axes: xlabel='gender', ylabel='count'>

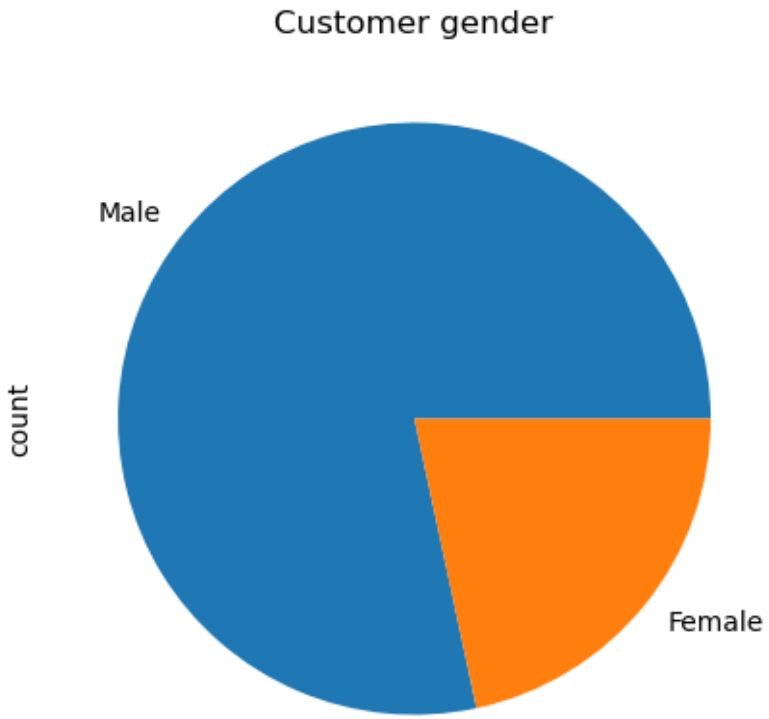


[86]:
```
sns.lineplot(data=train, x='age', y='n_followers')
```

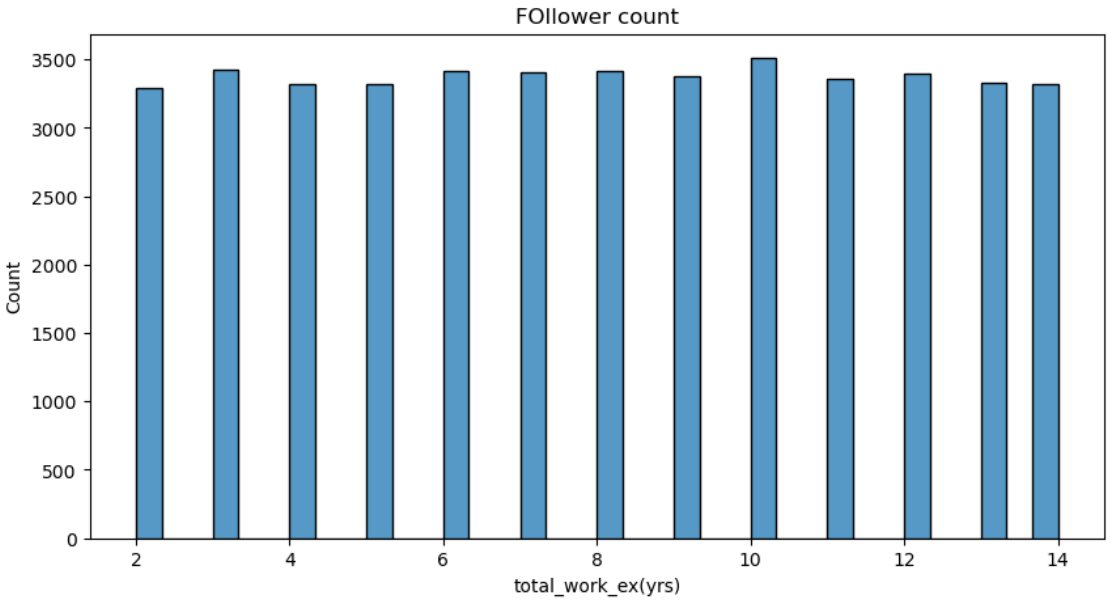
[86]:
```
<Axes: xlabel='age', ylabel='n_followers'>
```



[88]:
```
#pie diagram
plt.figure(figsize=(10,5))
train['gender'].value_counts().plot.pie()
plt.title("Customer gender")
plt.show()
```



[94]:
```
plt.figure(figsize= (10,5))
sns.histplot(x=train['total_work_ex(yrs)'])
plt.title("FOllower count")
plt.show()
```



[112]:
```
# Encoder
encoder= LabelEncoder()
for col in cat_col:
    train[col] = encoder.fit_transform(train[col])
    test[col] = encoder.fit_transform(test[col])
```

[100]:
```
train.dtypes
```

[100]:
```
c_id                                  int32
total_work_ex(yrs)                    int64
active_days_on_linkedin             float64
n_times_paid_promo                  float64
no_of_previous_positions            float64
days_in_current_org                 float64
age                                 float64
reactions_on_each_post_per_second   float64
avg_angry_reaction_on_posts         float64
avg_disgusted_reaction_on_posts     float64
avg_scared_reaction_on_posts        float64
avg_happy_reaction_on_posts         float64
avg_neutral_reaction_on_posts       float64
avg_sad_reaction_on_posts           float64
avg_surprise_reaction_on_posts      float64
max_reactions_in_single_post        float64
profile_completion%                 float64
gender                                int32
country                               int32
display_picture_clarity             float64
total_posts                           int64
avg_reaction_in_each_post             int64
n_followers                           int64
dtype: object
```

[102]:
```
test.dtypes
```

[102]:
```
c_id                                  int32
total_work_ex(yrs)                    int64
active_days_on_linkedin             float64
n_times_paid_promo                  float64
no_of_previous_positions            float64
days_in_current_org                 float64
age                                 float64
reactions_on_each_post_per_second   float64
avg_angry_reaction_on_posts         float64
avg_disgusted_reaction_on_posts     float64
```

```
avg_scared_reaction_on_posts         float64
avg_happy_reaction_on_posts          float64
avg_neutral_reaction_on_posts        float64
avg_sad_reaction_on_posts            float64
avg_surprise_reaction_on_posts       float64
max_reactions_in_single_post         float64
profile_completion%                  float64
gender                                 int32
country                                int32
display_picture_clarity              float64
total_posts                            int64
avg_reaction_in_each_post              int64
dtype: object
```

[116]:
```
X = train.drop(columns=['n_followers'])
y= train['n_followers']
```

[136]:
```
selector = SelectKBest(score_func=f_regression, k=20)
X_selected = selector.fit_transform(X,y)
selected_features = X.columns[selector.get_support()].tolist()
selected_features
```

[136]:
```
['total_work_ex(yrs)',
 'active_days_on_linkedin',
 'no_of_previous_positions',
 'days_in_current_org',
 'age',
 'reactions_on_each_post_per_second',
 'avg_angry_reaction_on_posts',
 'avg_disgusted_reaction_on_posts',
 'avg_scared_reaction_on_posts',
 'avg_happy_reaction_on_posts',
 'avg_neutral_reaction_on_posts',
 'avg_sad_reaction_on_posts',
 'avg_surprise_reaction_on_posts',
 'max_reactions_in_single_post',
 'profile_completion%',
 'gender',
 'country',
 'display_picture_clarity',
 'total_posts',
 'avg_reaction_in_each_post']
```

[138]:
```
X_train, X_test, y_train, y_test = train_test_split(X_selected, y, test_size=0.
↪2, random_state=42)
```

[140]:
```
# Scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.fit_transform(X_test)

scaled_X_test= scaler.fit_transform(test[selected_features])
```

[142]:
```
model = RandomForestRegressor()
model.fit(X_train_scaled, y_train)
```

[142]:
```
RandomForestRegressor()
```

[144]:
```
y_pred = model.predict(X_test_scaled)
```

[146]:
```
r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test,y_pred)
print("r2: ",r2)
print("mae: ",mae)
```

```
r2:  0.3926146569917722
mae:  424.15371112883014
```

[148]:
```
#prediction on test data
test_pred = model.predict(scaled_X_test)
```

[ ]:
```
#submission
submission = pd.DataFrame({'customer_id': pd.read_csv("test.
 ↪csv")["customer_id"], "is_target" : test_pred})
submission.to_csv('submission.csv',index=False)
print("Submission file saved as submission.csv")
```

[ ]:
```
submission_df = pd.DataFrame({
    'PassengerId': test_df['PassengerId'].values,  # Ensure it's a NumPy array
    'Prediction': test_data_pred
})
```