# Exploratory Data Analysis (EDA) Report

## 1. Dataset Information

Total Rows: 364295

Total Columns: 20

## 2. Duplicate Values

Number of duplicate rows: 0

## 3. Missing Values

pseudo\_id 0  
text 0  
retweetCount 0  
replyCount 0  
likeCount 0  
quoteCount 0  
viewCount 0  
bookmarkCount 0  
createdAt 0  
lang 0  
isReply 0  
pseudo\_inReplyToId 0  
pseudo\_conversationId 0  
pseudo\_inReplyToUserId 0  
pseudo\_author\_id 0  
author\_location 144962  
author\_followers 0  
author\_following 0  
author\_isVerified 0  
author\_isBlueVerified 0

## 4. Summary Statistics

pseudo\_id retweetCount replyCount likeCount \  
count 3.642950e+05 364295.000000 364295.000000 364295.000000   
mean 4.997332e+18 2.173406 1.343516 16.170367   
std 2.885267e+18 56.097843 20.175422 446.843612   
min 1.970105e+13 0.000000 0.000000 0.000000   
25% 2.497476e+18 0.000000 0.000000 0.000000   
50% 4.997392e+18 0.000000 0.000000 0.000000   
75% 7.492344e+18 0.000000 1.000000 2.000000   
max 9.999995e+18 12924.000000 6382.000000 133154.000000   
  
 quoteCount viewCount bookmarkCount pseudo\_conversationId \  
count 364295.000000 3.642950e+05 364295.000000 3.642950e+05   
mean 0.263372 2.668370e+03 2.086070 5.015683e+18   
std 11.169570 4.089104e+05 98.822419 2.891292e+18   
min 0.000000 0.000000e+00 0.000000 4.569730e+13   
25% 0.000000 2.400000e+01 0.000000 2.498343e+18   
50% 0.000000 5.800000e+01 0.000000 5.011685e+18   
75% 0.000000 1.860000e+02 0.000000 7.507837e+18   
max 3454.000000 2.323289e+08 38703.000000 9.999995e+18   
  
 pseudo\_author\_id author\_followers author\_following   
count 3.642950e+05 3.642950e+05 3.642950e+05   
mean 4.991555e+18 3.441514e+04 1.254435e+03   
std 2.886714e+18 5.473658e+05 6.749906e+03   
min 7.265175e+12 0.000000e+00 0.000000e+00   
25% 2.480230e+18 8.100000e+01 1.190000e+02   
50% 4.969569e+18 3.780000e+02 3.930000e+02   
75% 7.496272e+18 1.951000e+03 1.065000e+03   
max 9.999938e+18 6.315795e+07 1.091266e+06

## 5. Categorical Column Distribution

Unique values in lang:

lang  
en 202274  
es 53278  
ja 33230  
pt 10236  
nl 9347  
fr 7057  
tr 5284  
id 4089  
de 3634  
it 3537  
zh-cn 3301  
af 3177  
ar 3119  
pl 2018  
ca 1910  
ko 1855  
et 1664  
fa 1323  
ru 1126  
no 1113  
vi 1089  
sv 1024  
th 777  
hi 691  
el 660  
uk 609  
so 598  
tl 596  
he 575  
fi 539  
ro 534  
sw 473  
cs 468  
hr 375  
da 327  
cy 292  
ur 279  
sl 275  
zh-tw 255  
hu 172  
sk 151  
lt 150  
lv 111  
ta 107  
bg 106  
sq 83  
mr 74  
ml 72  
unknown 63  
bn 43  
ne 39  
te 38  
mk 32  
gu 22  
kn 22  
pa 2

Unique values in isReply:

isReply  
True 183553  
False 180742  
  
Unique values in author\_isVerified:

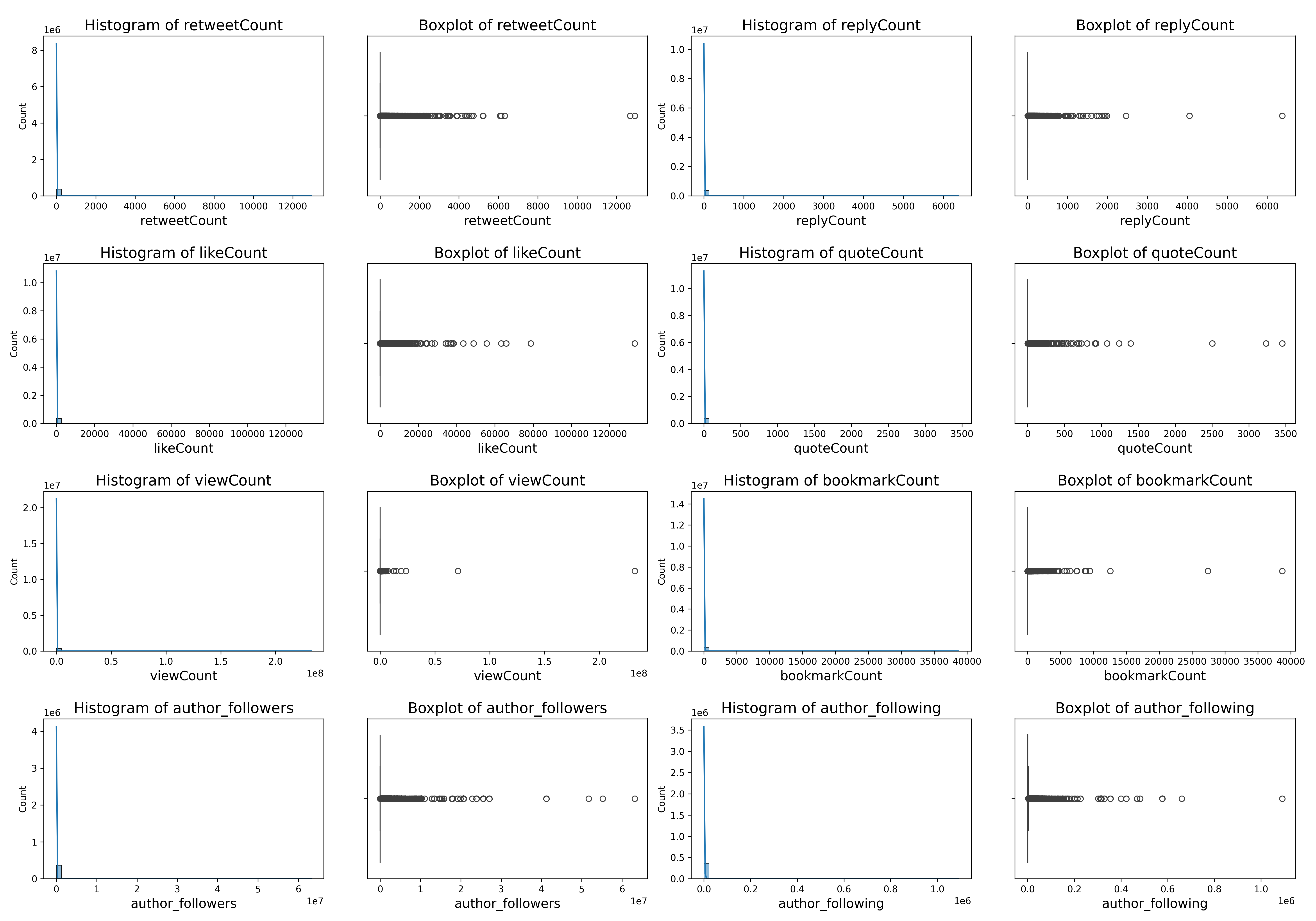
author\_isVerified  
False 364295

Unique values in author\_isBlueVerified:

author\_isBlueVerified  
False 251886  
True 112409

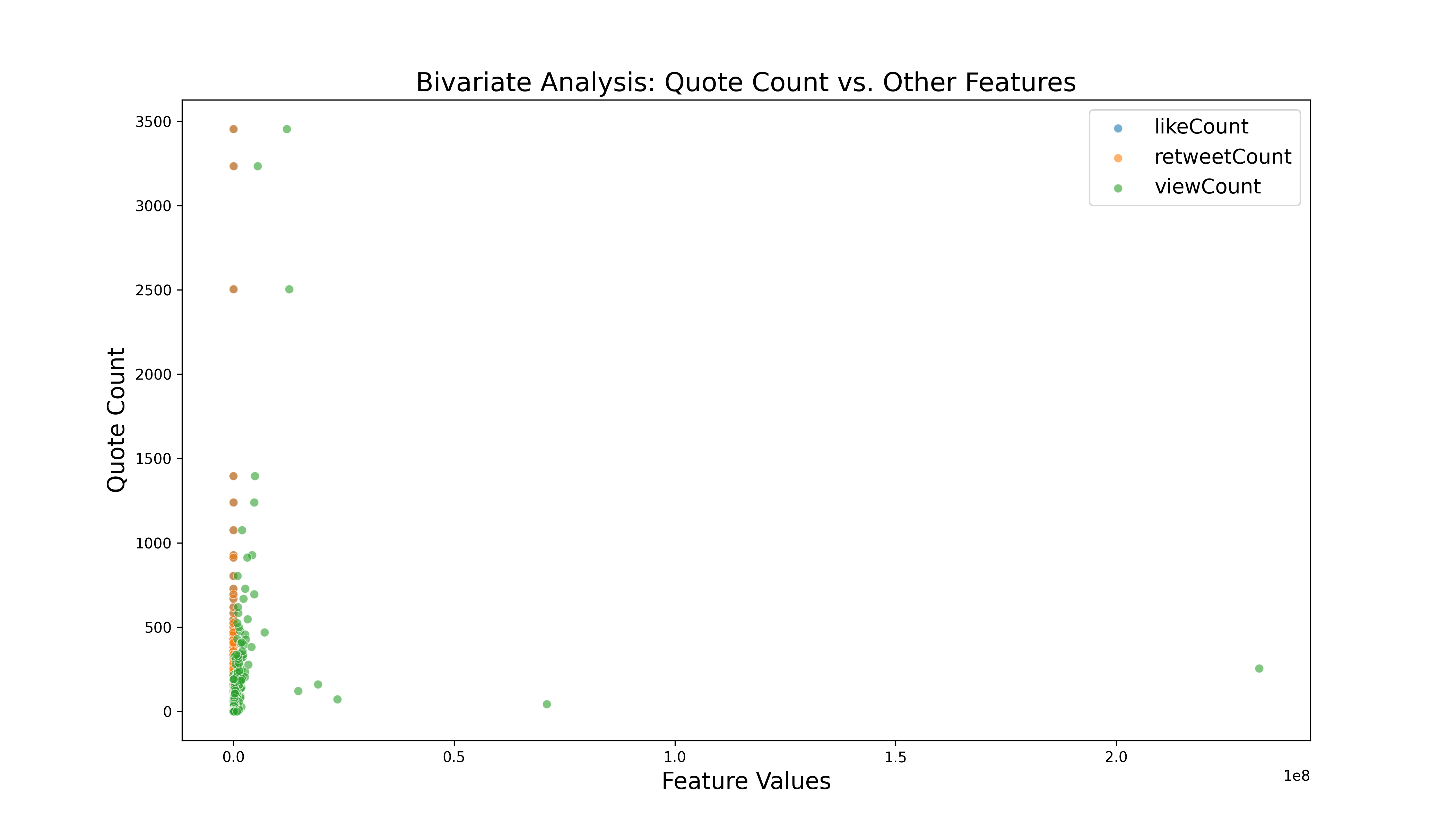
## 6. Univariate Analysis

Below are the histograms and box plots for numerical features:



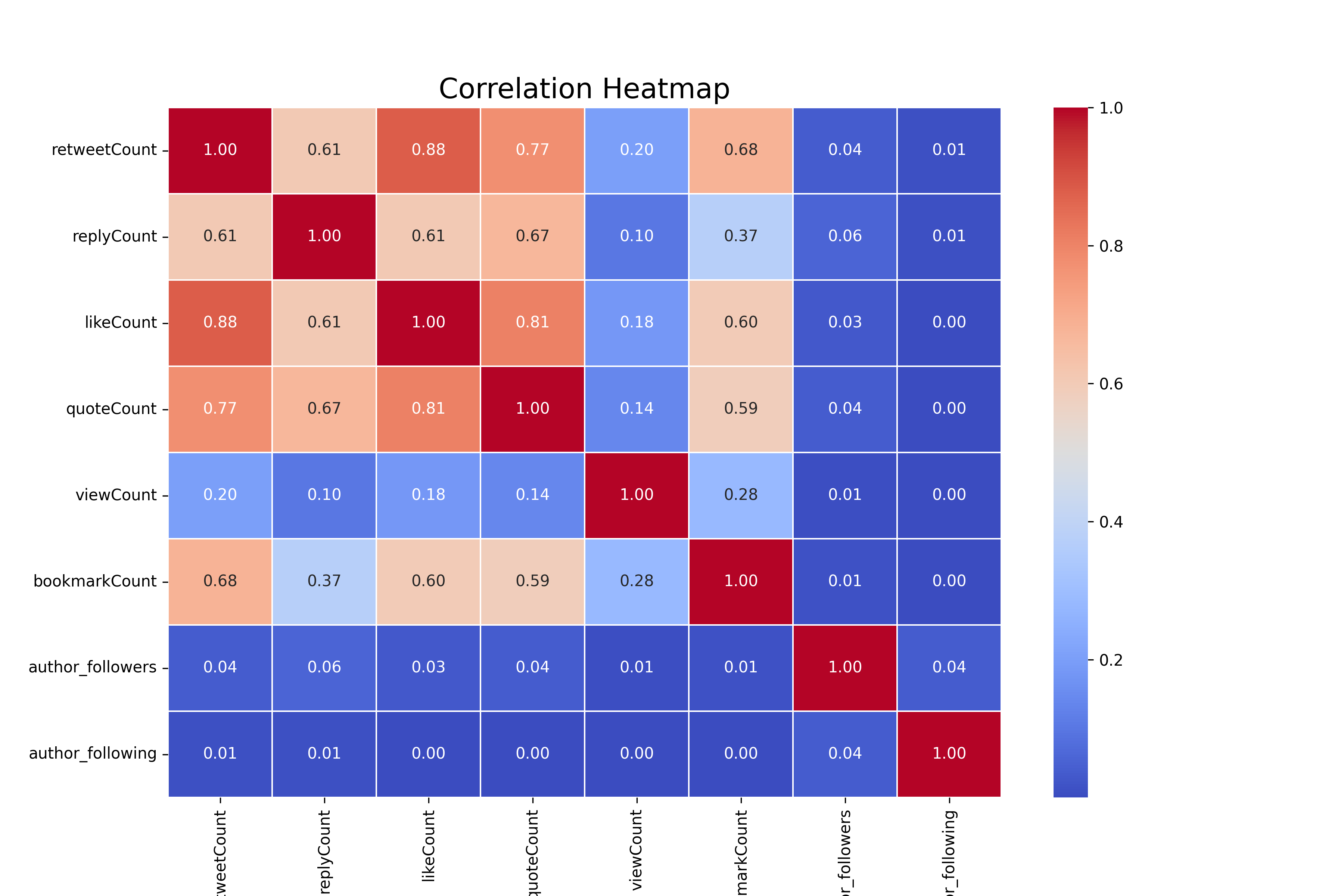
## 7. Bivariate Analysis

Scatter plot showing relationships between engagement metrics:



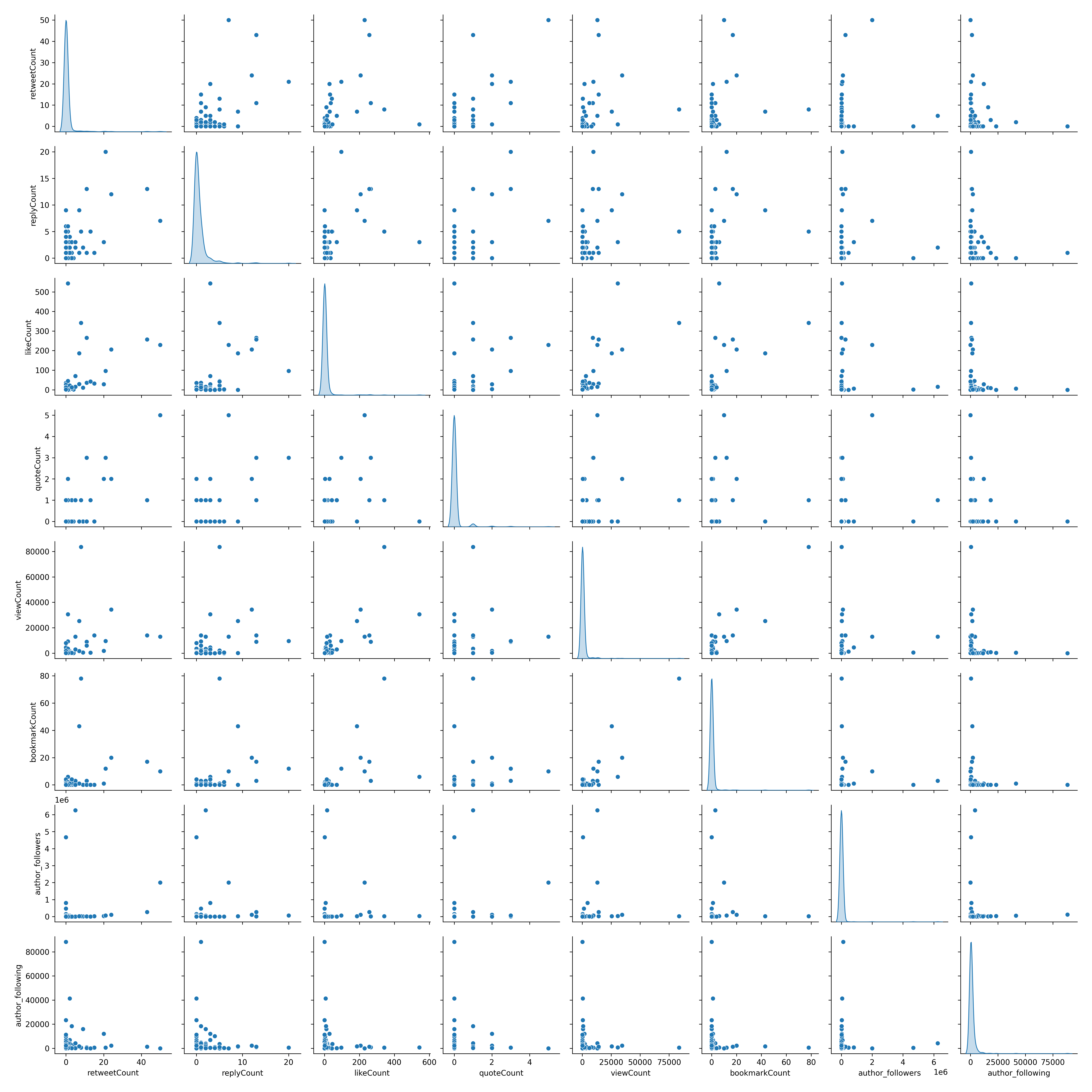
## 8. Correlation Heatmap

The heatmap below shows correlations between numerical features:



## 9. Multivariate Analysis

The pair plot below visualizes relationships between multiple numerical features:



## 10. Conclusion & Findings

Key Insights:

- The dataset contains missing values in some columns.

- Strong correlation exists between `likeCount`, `retweetCount`, and `viewCount`.

- The data is right-skewed, indicating most tweets have low engagement.

- There are outliers in `likeCount`, `retweetCount`, and `viewCount` that may need further analysis.