In [2]:
```
#Youtube Video Statistics - Begum Zubeda
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [3]:
```
#Import dataset
youtube_df = pd.read_csv(r"C:\zubeda\PGA-02\Python\Project\Dataset\train.csv")

youtube_df.head()
```

Out[3]:

| | Video_id | category_id | channel_title | subscriber | title | |
|---|---|---|---|---|---|---|
| **0** | HDR9SQc79 | 22 | CaseyNeistat | 9086142.0 | WE WANT TO TALK ABOUT OUR MARRIAGE | SHANtell m |
| **1** | KNH52UF?48 | 24 | LastWeekTonight | 5937292.0 | The Trump Presidency: Last Week Tonight with J... | last week tonight trump presidenc we |
| **2** | QTW28IRG36 | 23 | Rudy Mancuso | 4191209.0 | Racist Superman \| Rudy Mancuso, King Bach & Le... | superman\|rudy\|mancuso\|king\|bach\|rac |
| **3** | MGL76WI]26 | 24 | Good Mythical Morning | 13186408.0 | Nickelback Lyrics: Real or Fake? | rhett and link\|gmm\|good myt morning\|rh |
| **4** | TWP93KXT70 | 24 | nigahiga | 20563106.0 | I Dare You: GOING BALD!? | ryan\|higa\|higatv\|nigahiga\|i you\|idy\|rh |

In [5]:
```
youtube_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3198 entries, 0 to 3197
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Video_id              3198 non-null   object
 1   category_id           3192 non-null   object
 2   channel_title         3195 non-null   object
 3   subscriber            3175 non-null   float64
 4   title                 3195 non-null   object
 5   tags                  3046 non-null   object
 6   description           3133 non-null   object
 7   Trend_day_count       3197 non-null   float64
 8   Tag_count             3197 non-null   object
 9   Trend_tag_count       3197 non-null   object
```

```
 10  comment_count          3198 non-null   object
 11  comment_disabled       3198 non-null   object
 12  like dislike disabled  3198 non-null   object
 13  likes                  3198 non-null   object
 14  dislike                3198 non-null   object
 15  tag appered in title   3197 non-null   object
 16  views                  3198 non-null   object
 17  Unnamed: 17            1 non-null      float64
 18  Unnamed: 18            1 non-null      object
dtypes: float64(3), object(16)
memory usage: 474.8+ KB
```

In [6]:
```
#Dataframe shape
youtube_df.shape
```

Out[6]:
```
(3198, 19)
```

In [7]:
```
#Check Datatypes
youtube_df.dtypes
```

Out[7]:
```
Video_id                 object
category_id              object
channel_title            object
subscriber              float64
title                    object
tags                     object
description              object
Trend_day_count         float64
Tag_count                object
Trend_tag_count          object
comment_count            object
comment_disabled         object
like dislike disabled    object
likes                    object
dislike                  object
tag appered in title     object
views                    object
Unnamed: 17             float64
Unnamed: 18              object
dtype: object
```

In [8]:
```
#Check null values in percentage
youtube_df.isnull().sum()*100 / youtube_df.shape[0]
```

Out[8]:
```
Video_id                 0.000000
category_id              0.187617
channel_title            0.093809
subscriber               0.719199
title                    0.093809
tags                     4.752971
description              2.032520
Trend_day_count          0.031270
Tag_count                0.031270
Trend_tag_count          0.031270
comment_count            0.000000
comment_disabled         0.000000
like dislike disabled    0.000000
likes                    0.000000
dislike                  0.000000
tag appered in title     0.031270
views                    0.000000
Unnamed: 17             99.968730
```

```
Unnamed: 18              99.968730
dtype: float64
```

In [4]:
```
#Remove null columns
youtube_df.drop(youtube_df.columns[[-1, -2]], axis=1, inplace=True)
youtube_df.head()
```

Out[4]:

| | Video_id | category_id | channel_title | subscriber | title | |
|---|---|---|---|---|---|---|
| **0** | HDR9SQc79 | 22 | CaseyNeistat | 9086142.0 | WE WANT TO TALK ABOUT OUR MARRIAGE | SHANtell m |
| **1** | KNH52UF?48 | 24 | LastWeekTonight | 5937292.0 | The Trump Presidency: Last Week Tonight with J... | last week tonight trump presidenc we |
| **2** | QTW28IRG36 | 23 | Rudy Mancuso | 4191209.0 | Racist Superman \| Rudy Mancuso, King Bach & Le... | superman\|rudy\|mancuso\|king\|bach\|ra |
| **3** | MGL76WI]26 | 24 | Good Mythical Morning | 13186408.0 | Nickelback Lyrics: Real or Fake? | rhett and link\|gmm\|good myt morning\|rh |
| **4** | TWP93KXT70 | 24 | nigahiga | 20563106.0 | I Dare You: GOING BALD!? | ryan\|higa\|higatv\|nigahiga\|i you\|idy\|rh |

In [5]:
```
#Fill null rows
youtube_df['category_id'].fillna(0, inplace=True)
youtube_df['channel_title'].fillna('Anonymous', inplace=True)
youtube_df['subscriber'].fillna(0, inplace=True)
youtube_df['title'].fillna('No title', inplace=True)
youtube_df['tags'].fillna('No tags', inplace=True)
youtube_df['description'].fillna('No description', inplace=True)
youtube_df['Trend_day_count'].fillna(0, inplace=True)
youtube_df['Tag_count'].fillna(0, inplace=True)
youtube_df['Trend_tag_count'].fillna(0, inplace=True)
youtube_df['tag appered in title'].fillna('No title tag', inplace=True)

youtube_df.isnull().sum()*100 / youtube_df.shape[0]
```

Out[5]:
```
Video_id                 0.0
category_id              0.0
channel_title            0.0
subscriber               0.0
title                    0.0
tags                     0.0
description              0.0
Trend_day_count          0.0
```

```
Tag_count                0.0
Trend_tag_count          0.0
comment_count            0.0
comment_disabled         0.0
like dislike disabled    0.0
likes                    0.0
dislike                  0.0
tag appered in title     0.0
views                    0.0
dtype: float64
```

In [6]:
```
#Replace inconsistent data
columns = ['Tag_count', 'Trend_tag_count', 'comment_count', 'likes', 'dislike', 'views'
regex = [r'^[A-Za-z)(<>+-/@_.&#\s]+', '', '#VALUE!', True, False]
youtube_df[columns] = youtube_df[columns].replace(regex=regex, value=0)
youtube_df['like dislike disabled'] = youtube_df['like dislike disabled'].replace(regex
youtube_df['comment_disabled'] = youtube_df['comment_disabled'].replace(regex=r'^[0-9]+
```

In [7]:
```
#Convert columns to valid datatypes
columns = ['Tag_count', 'Trend_tag_count', 'comment_count', 'likes', 'dislike', 'views'
youtube_df[columns] = youtube_df[columns].astype(float)

youtube_df.dtypes
```

Out[7]:
```
Video_id                 object
category_id              object
channel_title            object
subscriber              float64
title                    object
tags                     object
description              object
Trend_day_count         float64
Tag_count               float64
Trend_tag_count         float64
comment_count           float64
comment_disabled         object
like dislike disabled    object
likes                   float64
dislike                 float64
tag appered in title     object
views                   float64
dtype: object
```

In [13]:
```
youtube_df.tail(10)
```

Out[13]:

| | Video_id | category_id | channel_title | subscriber | title | |
|---|---|---|---|---|---|---|
| **3188** | AVI91YQD43 | 24 | Variety | 230720.0 | Chloe Grace Moretz on Louis C.K. and the #MeTo... | Varie |
| **3189** | UGJ10XAb33 | 26 | AmazingPhil | 4232293.0 | I Read A Letter From My Younger Self | amazingphil\|phil lester\|phil\|let |

| | Video_id | category_id | channel_title | subscriber | title | |
|---|---|---|---|---|---|---|
| **3190** | PTM7WHf59 | 17 | NBA | 8707071.0 | Team LeBron! Best Plays from Every All-Star on... | nba\|highlights\|basketball\|plays\| |
| **3191** | ISJ87JHs2 | 22 | Grace Helbig | 3008137.0 | My Bachelor Audition Tape // Grace Helbig | grace\|grace helbig\|graceinabc |
| **3192** | ZGR33QU@74 | 24 | Trailers Promos Teasers | 0.0 | The Hurricane Heist Trailer | The Hurricane Heist Trailer\|The |
| **3193** | OKR48DOE67 | 20 | EA SPORTS FIFA | 3150213.0 | FIFA 18 - FUT Champions Cup Barcelona - Day 1 | fifa\|fifa ultimate team\|fut\|fifa |
| **3194** | QJK69DS?91 | 10 | JackWhiteVEVO | 261596.0 | Jack White - Corporation (Audio) | Alternative\|Corporation\|Jack Wh |
| **3195** | VHF51NVr11 | 10 | JamesBlakeVEVO | 28321.0 | James Blake - If The Car Beside You Moves Ahea... | James\|Blake\|If\|The\|Car\|Beside\|You |
| **3196** | XHU22OAJ39 | 26 | Refinery29 | 890739.0 | Lucie Fink Trains Like A Professional Gymnast ... | refinery29\|refinery 29\|r29\|r29 |
| **3197** | IFD79NSG47 | 22 | MN khan | 0.0 | Man drops magnum of Champagne on the floor in ... | |

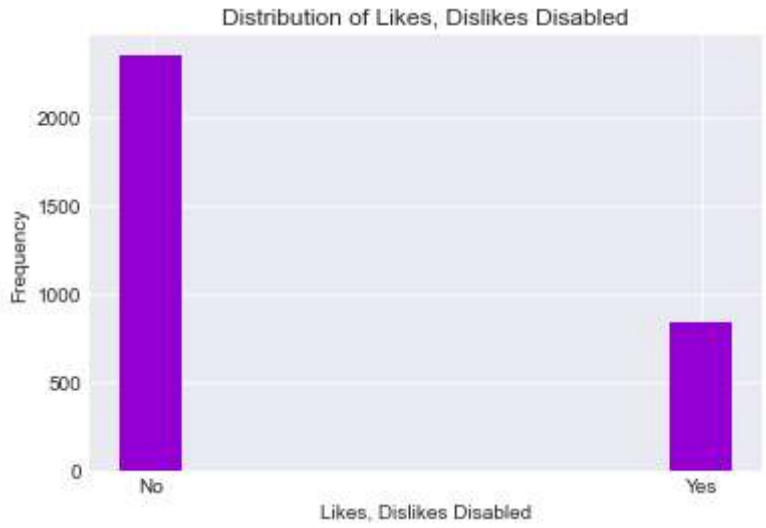
In [25]:
```
youtube_df.describe()
```

Out[25]:

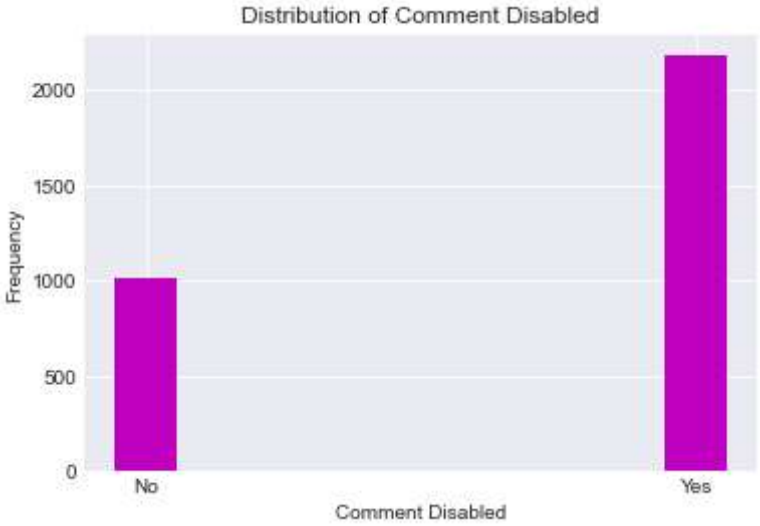
| | subscriber | Trend_day_count | Tag_count | Trend_tag_count | comment_count | likes | |
|---|---|---|---|---|---|---|---|
| **count** | 3.198000e+03 | 3198.000000 | 3198.000000 | 3198.000000 | 3198.000000 | 3.198000e+03 | 3 |
| **mean** | 3.796479e+06 | 7.961851 | 18.609131 | 7.113196 | 112487.494371 | 9.763094e+03 | 5 |
| **std** | 2.855627e+07 | 78.543895 | 80.355531 | 175.056899 | 101861.271626 | 2.240396e+04 | 4 |

| | subscriber | Trend_day_count | Tag_count | Trend_tag_count | comment_count | likes | |
|---|---|---|---|---|---|---|---|
| **min** | 0.000000e+00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000e+00 | |
| **25%** | 2.326040e+05 | 4.000000 | 12.000000 | 2.000000 | 0.000000 | 0.000000e+00 | |
| **50%** | 1.206997e+06 | 7.000000 | 17.000000 | 4.000000 | 99582.000000 | 1.022150e+04 | 5 |
| **75%** | 3.808198e+06 | 10.000000 | 21.000000 | 6.000000 | 203208.250000 | 1.505300e+04 | 10 |
| **max** | 1.576229e+09 | 4444.000000 | 3225.000000 | 9903.000000 | 299877.000000 | 1.213628e+06 | 14 |

In [14]:
```
plt.style.use('seaborn-darkgrid')

#Number of videos where likes and dislikes are disbaled
plt.hist(x=youtube_df['like dislike disabled'], color='darkviolet')

plt.title("Distribution of Likes, Dislikes Disabled")
plt.xlabel('Likes, Dislikes Disabled')
plt.xticks([0.05, 0.95], ["No", "Yes"])
plt.ylabel('Frequency')

plt.show()
```

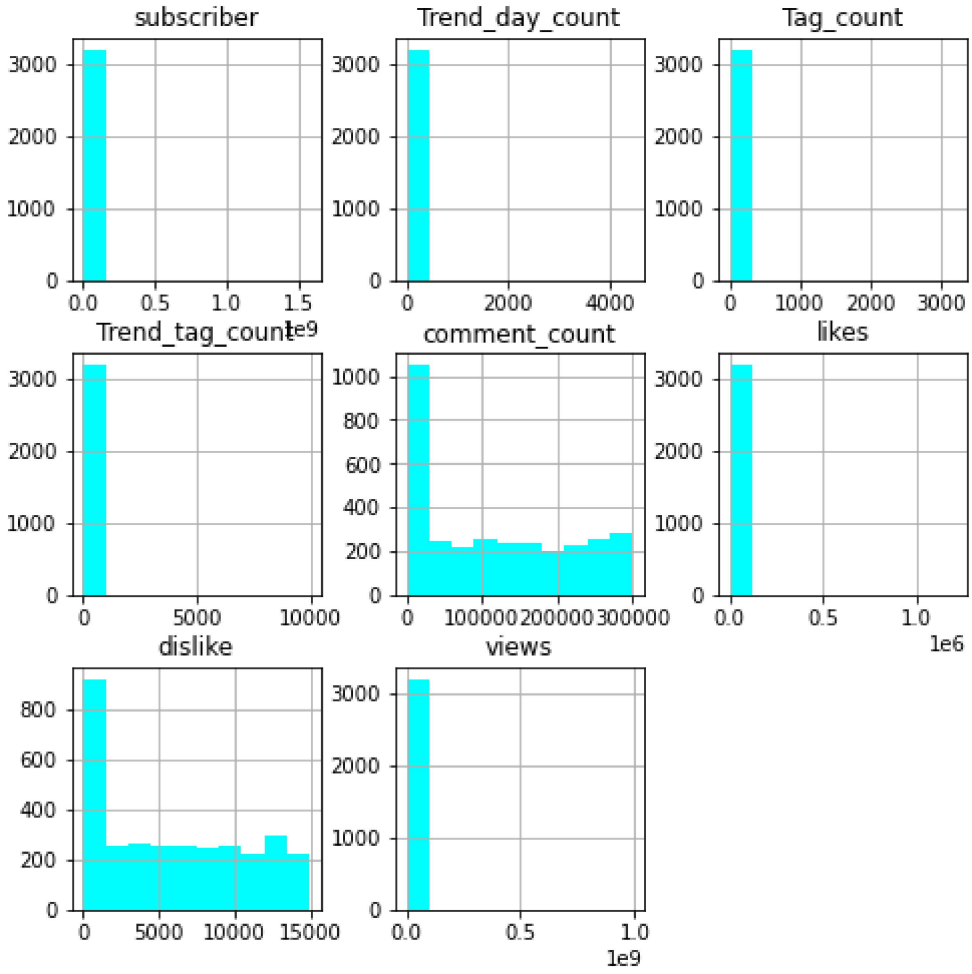



In [15]:
```
#Number of videos where comments are disbaled
plt.hist(x=youtube_df['comment_disabled'], color='m')

plt.title("Distribution of Comment Disabled")
plt.xlabel('Comment Disabled')
plt.xticks([0.05, 0.95], ["No", "Yes"])
plt.ylabel('Frequency')

plt.show()
```



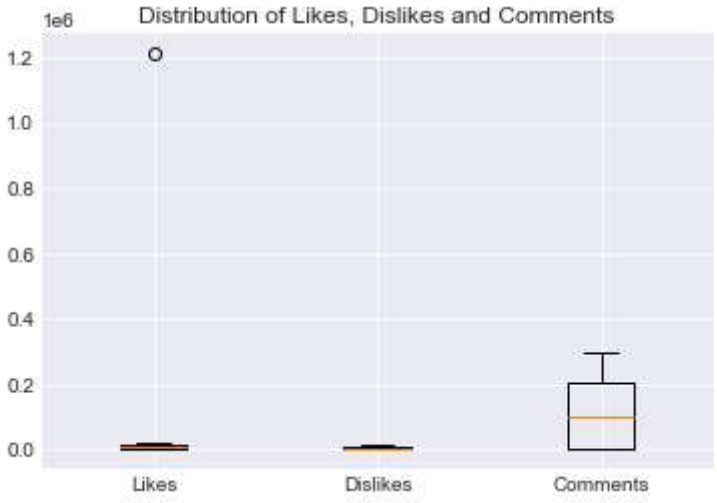

In [9]:
```
#Frequency Disributions of data
fig = plt.figure(figsize=(8,8))

ax = fig.gca()
youtube_df.hist(ax=ax, color="cyan")
plt.show()
```

```
<ipython-input-9-f27cadb55093>:5: UserWarning: To output multiple subplots, the figure c
ontaining the passed axes is being cleared
  youtube_df.hist(ax=ax, color="cyan")
```

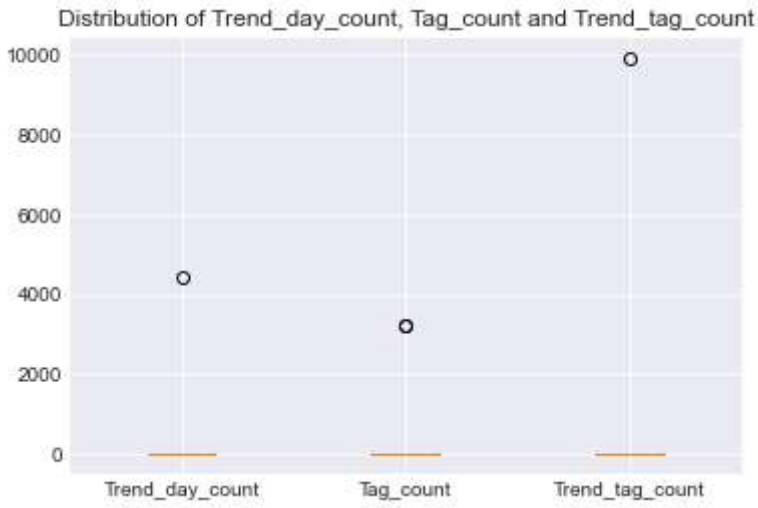



```
#Boxplot for likes, dislikes, comments count
```

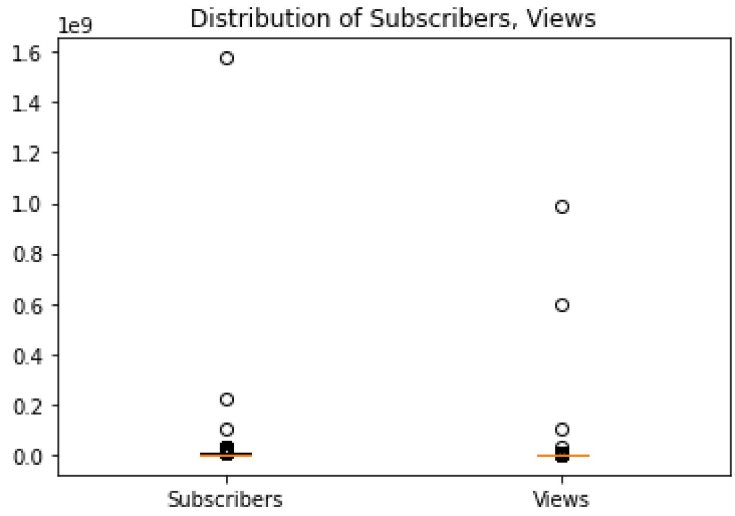
In [16]:
```
plt.boxplot([youtube_df['likes'], youtube_df['dislike'], youtube_df['comment_count']])

plt.title("Distribution of Likes, Dislikes and Comments")
plt.xticks([1, 2, 3],["Likes", "Dislikes", "Comments"])
plt.show()
```




In [17]:
```
#Boxplot for Trend_day_count, Tag_count, Trend_tag_count
plt.boxplot([youtube_df['Trend_day_count'], youtube_df['Tag_count'], youtube_df['Trend_

plt.title("Distribution of Trend_day_count, Tag_count and Trend_tag_count")
plt.xticks([1, 2, 3],["Trend_day_count", "Tag_count", "Trend_tag_count"])
plt.show()
```




In [8]:
```
#Boxplot for Subscribers, Views
plt.boxplot([youtube_df['subscriber'], youtube_df['views']])

plt.title("Distribution of Subscribers, Views")
plt.xticks([1, 2],["Subscribers", "Views"])
plt.show()
```

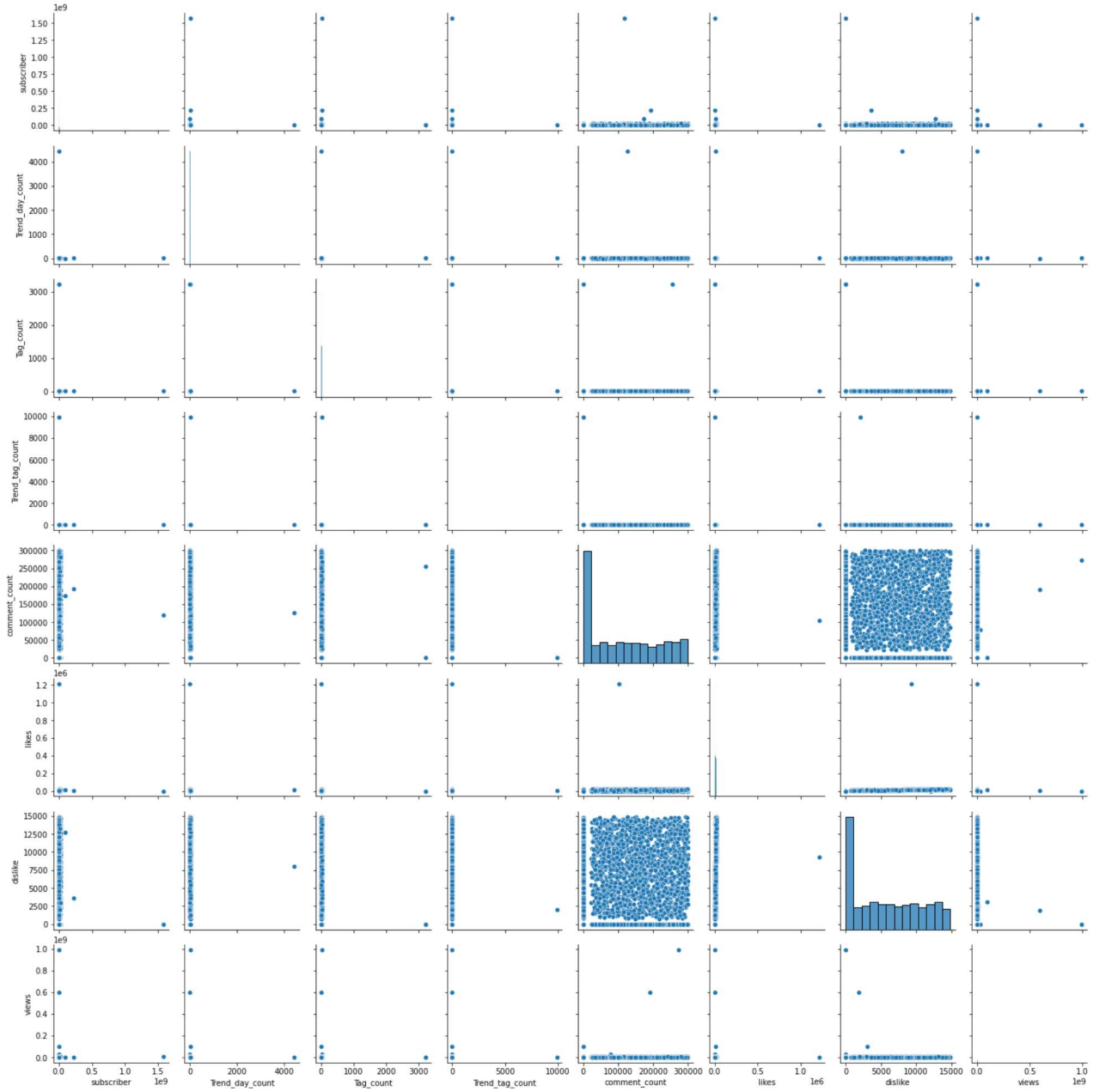



In [29]:
```
#Relationship between Variables
youtube_df.corr()
```

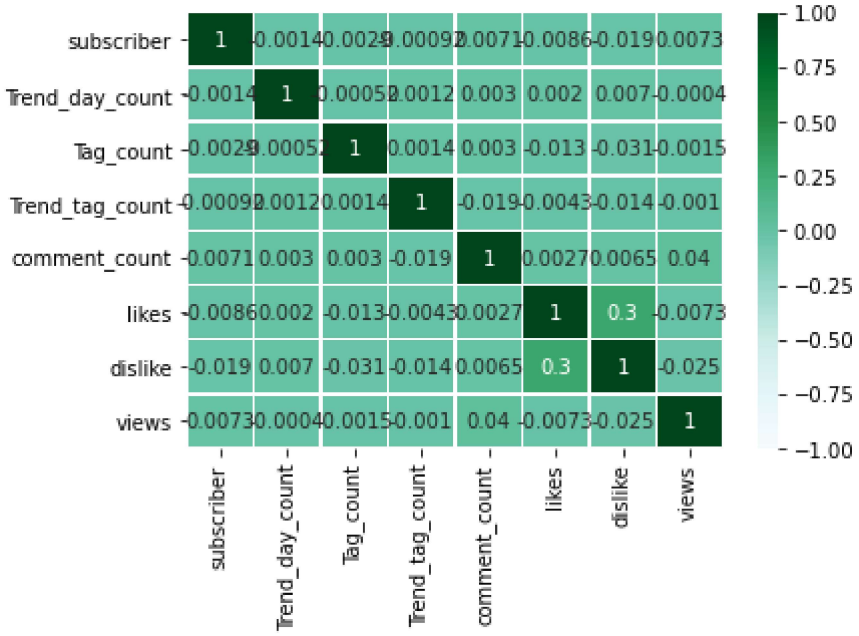
Out[29]:

| | subscriber | Trend_day_count | Tag_count | Trend_tag_count | comment_count | likes |
|---|---|---|---|---|---|---|
| **subscriber** | 1.000000 | -0.001403 | -0.002943 | -0.000915 | 0.007140 | -0.008582 |
| **Trend_day_count** | -0.001403 | 1.000000 | -0.000521 | 0.001204 | 0.003003 | 0.002038 |
| **Tag_count** | -0.002943 | -0.000521 | 1.000000 | 0.001400 | 0.003033 | -0.012868 |
| **Trend_tag_count** | -0.000915 | 0.001204 | 0.001400 | 1.000000 | -0.019444 | -0.004323 |
| **comment_count** | 0.007140 | 0.003003 | 0.003033 | -0.019444 | 1.000000 | 0.002694 |
| **likes** | -0.008582 | 0.002038 | -0.012868 | -0.004323 | 0.002694 | 1.000000 |
| **dislike** | -0.019086 | 0.006965 | -0.030584 | -0.013678 | 0.006452 | 0.297125 |
| **views** | 0.007321 | -0.000401 | -0.001519 | -0.001007 | 0.039945 | -0.007328 |

In [25]:
```
sns.pairplot(data=youtube_df)

plt.show()
```




In [31]:
```
sns.heatmap(data=youtube_df.corr(), annot=True, linewidth=0.5, cmap="BuGn", cbar=True,

plt.show()
```


subscriber
Trend_day_count
Tag_count
Trend_tag_count
comment_count
likes
dislike
views
1 -0.0014 0.0029 -0.00092 0.0071 -0.0086 -0.019 0.0073
-0.0014 1 0.00052 0.0012 0.003 0.002 0.007 -0.0004
-0.0029 0.00052 1 0.0014 0.003 -0.013 -0.031 -0.0015
-0.00092 0.0012 0.0014 1 -0.019 -0.0043 -0.014 -0.001
-0.0071 0.003 0.003 -0.019 1 0.0027 0.0065 0.04
-0.0086 0.002 -0.013 -0.0043 0.0027 1 0.3 -0.0073
-0.019 0.007 -0.031 -0.014 0.0065 0.3 1 -0.025
0.0073 0.0004 0.0015 -0.001 0.04 -0.0073 -0.025 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

In [ ]: