Connecting to the YouTube Dataset in the Drive

```
[ ] from google.colab import drive
```

```
[ ] drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Importing all the Requirements

```
[ ] !pip install wquantiles
    !pip install statsmodels
    !pip install scipy
```

```
Collecting wquantiles
  Downloading wquantiles-0.6-py3-none-any.whl.metadata (1.1 kB)
Requirement already satisfied: numpy>=1.18 in /usr/local/lib/python3.10/dist-packages (from wquantiles) (1.26.4)
Downloading wquantiles-0.6-py3-none-any.whl (3.3 kB)
Installing collected packages: wquantiles
Successfully installed wquantiles-0.6
Requirement already satisfied: statsmodels in /usr/local/lib/python3.10/dist-packages (0.14.2)
Requirement already satisfied: numpy>=1.22.3 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (1.26.4)
Requirement already satisfied: scipy!=1.9.2,>=1.8 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (1.13.1)
Requirement already satisfied: pandas!=2.1.0,>=1.4 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (2.1.4)
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (0.5.6)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.10/dist-packages (from statsmodels) (24.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2024.1)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2024.1)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from patsy>=0.5.6->statsmodels) (1.16.0)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (1.13.1)
Requirement already satisfied: numpy<2.3,>=1.22.4 in /usr/local/lib/python3.10/dist-packages (from scipy) (1.26.4)
```

```
[ ] %matplotlib inline

    from pathlib import Path
    import pandas as pd
    import numpy as np
    from scipy.stats import trim_mean
    from statsmodels import robust
    import wquantiles
    from scipy.stats import trim_mean
    import seaborn as sns
    import matplotlib.pylab as plt
    import matplotlib.pyplot as plt
    import statsmodels.api as sm
    from statsmodels import robust
```

```
[ ] try:
        import common
        DATA = common.dataDirectory()
    except ImportError:
        DATA = Path().resolve() / 'data'
```

Estimates of Location

```
[ ] data = pd.read_csv('/content/drive/MyDrive/yt_dataset/CAvideos.csv')
```

```
[ ] data.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	thumbn
0	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	17158579	787425	43420	125882	https://i.ytimg.com/vi/n1WpP7iowLc/
1	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	1014651	127794	1688	13030	https://i.ytimg.com/vi/0dBIkQ4Mz1M/
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146035	5339	8181	https://i.ytimg.com/vi/5qpjK5DgCt4/
3	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095828	132239	1989	17518	https://i.ytimg.com/vi/d380meD0W0M/
4	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	33523622	1634130	21082	85067	https://i.ytimg.com/vi/2Vv-BfVoq4g/

```
[ ] print(data['views'].mean())
```

```
1147035.9107898534
```

```
[ ] print(trim_mean(data['views'],0.1))
```

```
564872.9512306986
```

```
[ ] print(data['likes'].median())
```

```
8780.0
```

```
[ ] print(data['comment_count'].median())
```

```
1301.0
```

```
[ ] print(data['dislikes'].mean())
```

```
2009.1954453168953
```

```
[ ] print(np.average(data['likes'], weights=data['views']))
```

```
364745.9537423492
```

```
[ ] print(wquantiles.median(data['likes'], weights=data['views']))
```

```
102103.61535286104
```

Estimates of Variability

```
[ ] print(data['views'].std())
```

```
3390913.022309031
```

```
[ ] print(data['views'].quantile(0.75) - data['views'].quantile(0.25))
```

```
819400.0
```

```
[ ] print(robust.scale.mad(data['likes']))
    print(abs(data['likes'] - data['likes'].median()).median() / 0.6744897501960817)
```

```
11730.348752816322
11730.348752816322
```

Estimates on Percentiles and Boxplots

```
[ ] print(data['likes'].quantile([0.05, 0.25, 0.5, 0.75, 0.95]))
```

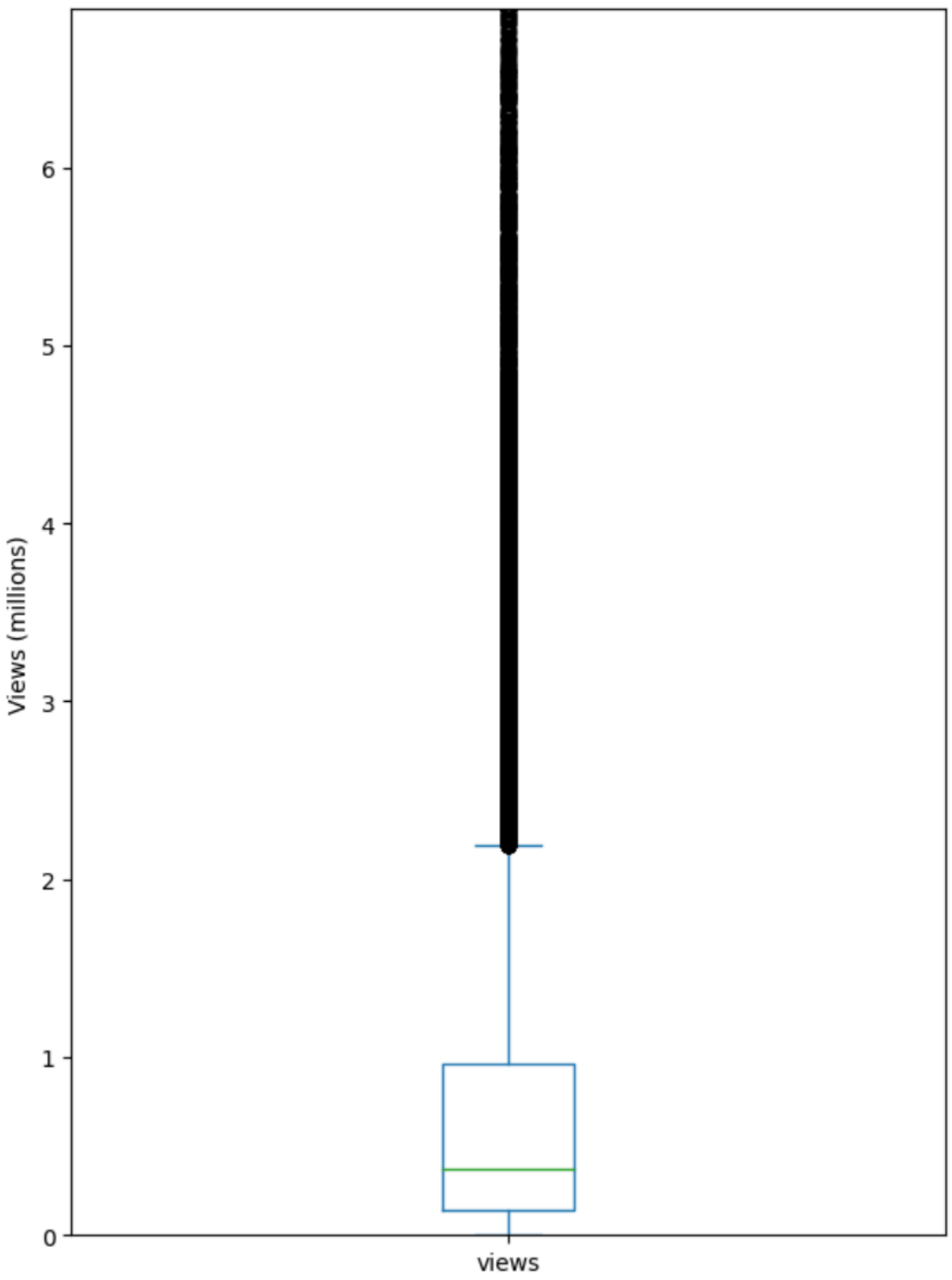
```
0.05       201.0
0.25      2191.0
0.50      8780.0
0.75     28717.0
0.95    165252.0
Name: likes, dtype: float64
```

```
[ ] percentages = [0.05, 0.25, 0.5, 0.75, 0.95]
    df = pd.DataFrame(data['views'].quantile(percentages))
    df.index = [f'{p * 100}%' for p in percentages]
    print(df.transpose())
```

```
           5.0%     25.0%     50.0%     75.0%      95.0%
views  30061.0  143902.0  371204.0  963302.0  4090835.0
```

```
[ ] views_millions = data['views'] / 1_000_000
    ax = views_millions.plot.box(figsize=(6,8))

    ax.set_ylabel('Views (millions)')

    ax.set_ylim(0, views_millions.max() * 0.05)
    plt.tight_layout()
    plt.show()
```

[ ]


Views (millions)
0
1
2
3
4
5
6
views

Frequency tables and Histograms

```
[ ]  binned_views = pd.cut(data['views'], 10)
     print(binned_views.value_counts())
```

```
views
(-137109.387, 13784971.7]     40465
(13784971.7, 27569210.4]        302
(27569210.4, 41353449.1]         70
(41353449.1, 55137687.8]         22
(55137687.8, 68921926.5]          7
(68921926.5, 82706165.2]          6
(82706165.2, 96490403.9]          4
(96490403.9, 110274642.6]         2
(124058881.3, 137843120.0]        2
(110274642.6, 124058881.3]        1
Name: count, dtype: int64
```

```
[ ]  binned_views.name = 'binned_views'

     df = pd.concat([data, binned_views], axis=1)
     df = df.sort_values(by='views')

     groups = []
     for group, subset in df.groupby(by='binned_views'):
         groups.append({
             'BinRange': group,
             'Count': len(subset),
             'data': ','.join(subset.title)
         })
     print(pd.DataFrame(groups))
```

```
                  BinRange  Count  \
0  (-137109.387, 13784971.7]  40465
1   (13784971.7, 27569210.4]    302
2   (27569210.4, 41353449.1]     70
3   (41353449.1, 55137687.8]     22
4   (55137687.8, 68921926.5]      7
5   (68921926.5, 82706165.2]      6
6   (82706165.2, 96490403.9]      4
7   (96490403.9, 110274642.6]     2
8  (110274642.6, 124058881.3]     1
9  (124058881.3, 137843120.0]     2

                                                data
0  ''Gala Artis 2018'' Le numéro d'ouverture,Cana...
1  Kaala (Tamil) - Official Teaser | Rajinikanth ...
2  Incredibles 2 - Olympics Sneak Peek,BTS (방탄소년단...
3  Taylor Swift - End Game ft. Ed Sheeran, Future...
4  Marvel Studios' Avengers: Infinity War Officia...
5  Childish Gambino - This Is America (Official V...
6  Marvel Studios' Avengers: Infinity War Officia...
7  Childish Gambino - This Is America (Official V...
8  YouTube Rewind: The Shape of 2017 | #YouTubeRe...
9  YouTube Rewind: The Shape of 2017 | #YouTubeRe...
<ipython-input-32-52d4169e711b>:7: FutureWarning: The default of observed=False is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain current behavior or ob
  for group, subset in df.groupby(by='binned_views'):
```

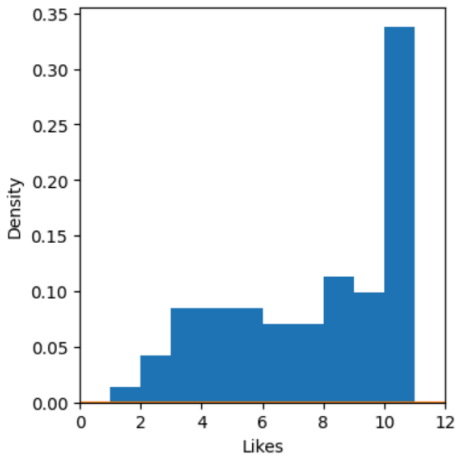
views_millions = data['dislikes'] / 1_000_000 ax = views_millions.plot.hist(figsize=(6,8))

ax.set_ylabel('Dislikes (millions)')

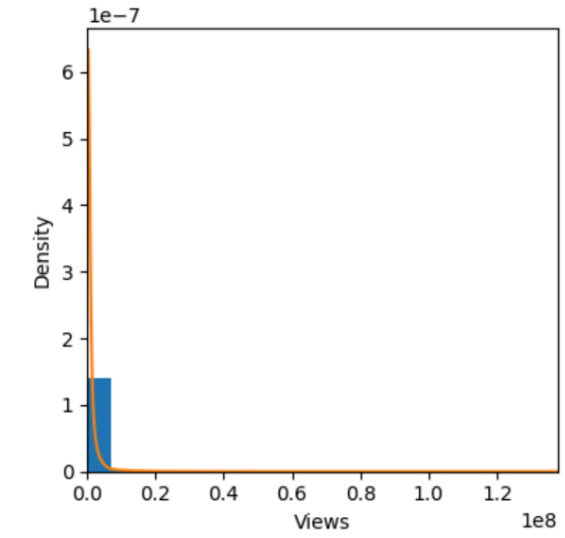
ax.set_ylim(0, views_millions.max() * 12.05) plt.tight_layout() plt.show()

Density Estimates

```
[ ]  ax = data['likes'].plot.hist(density=True, xlim=[0, 12],
                                   bins=range(1,12), figsize=(4, 4))
     data['likes'].plot.density(ax=ax)
     ax.set_xlabel('Likes')

     plt.tight_layout()
     plt.show()
```




```
[ ]  ax = df['views'].plot.hist(density=True, xlim=[0, df['views'].max()],
                                bins=20, figsize=(4, 4))
     df['views'].plot.density(ax=ax)
     ax.set_xlabel('Views')

     plt.tight_layout()
     plt.show()
```

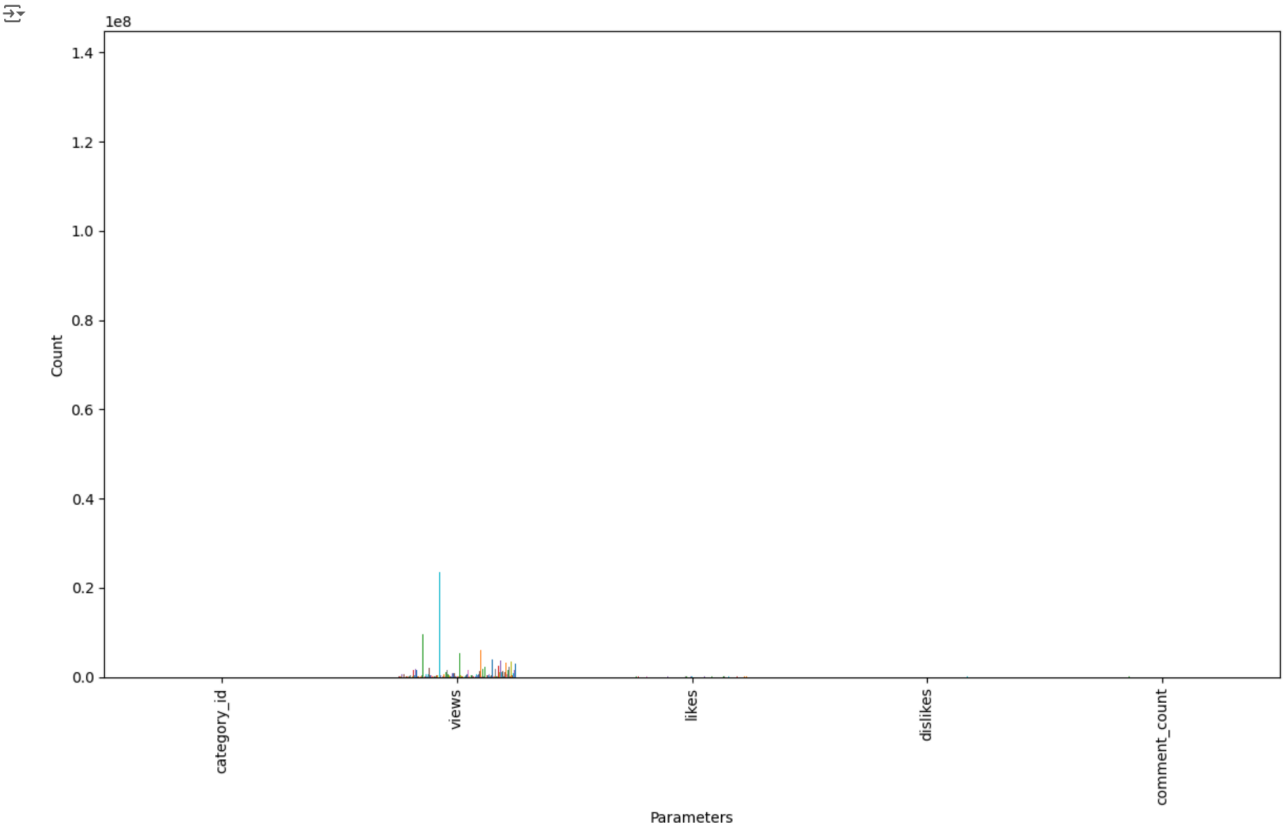


Exploring Binary and Categorical Data

```
[ ]  ca = pd.read_csv('/content/drive/MyDrive/yt_dataset/CAvideos.csv')
     total_sum = ca.select_dtypes(include='number').values.sum()
     percentages = 100 * ca.select_dtypes(include='number') / total_sum
     print(percentages)
```

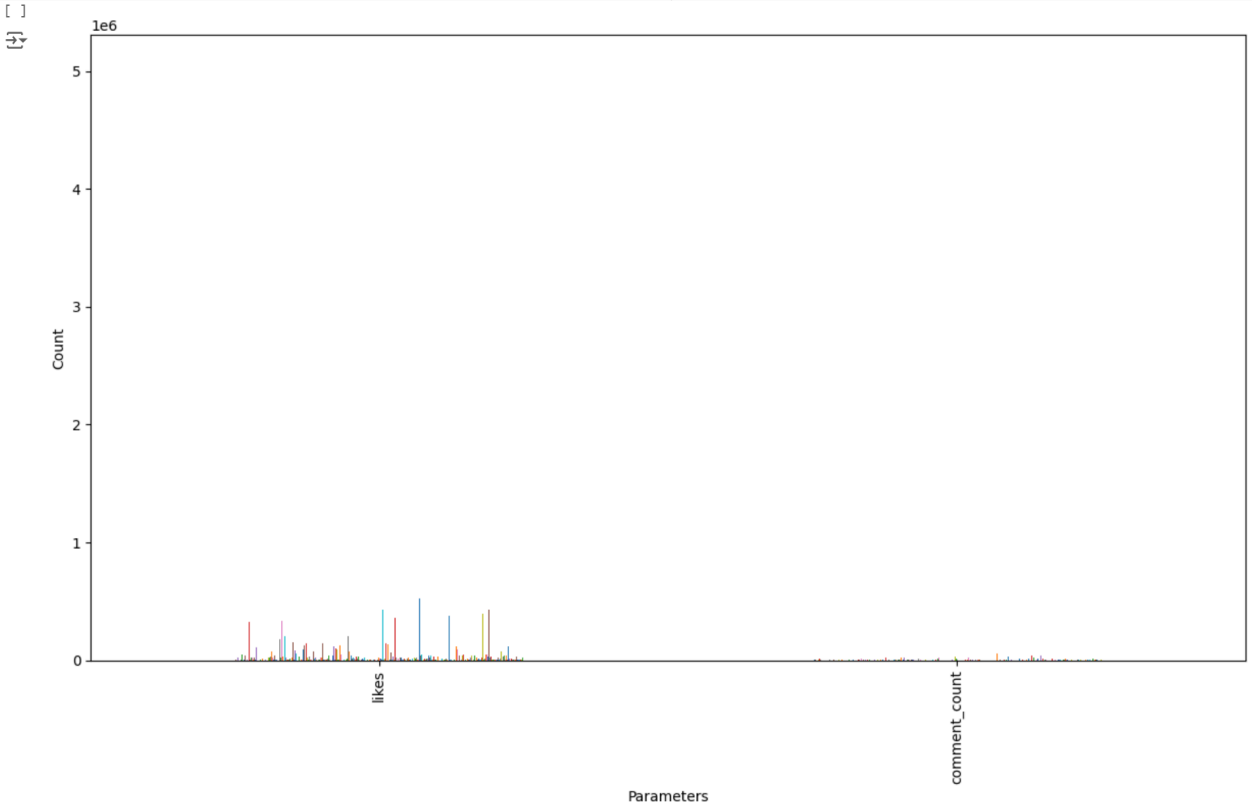
```
        category_id     views         likes      dislikes  comment_count
0      2.049210e-08  0.035162  1.613599e-03  8.897668e-05   2.579586e-04
1      4.713182e-08  0.002079  2.618767e-04  3.459066e-06   2.670120e-05
2      4.713182e-08  0.006540  2.992563e-04  1.094073e-05   1.676458e-05
3      4.918103e-08  0.004295  2.709854e-04  4.075878e-06   3.589805e-05
4      2.049210e-08  0.068697  3.348675e-03  4.320144e-05   1.743201e-04
...             ...       ...           ...           ...            ...
40876  4.918103e-08  0.000165  3.485705e-06  2.028717e-07   2.688563e-06
40877  4.918103e-08  0.000212  9.426364e-07  1.352478e-07   1.045097e-07
40878  4.098419e-08  0.001585  5.307453e-05  4.590229e-07   7.952982e-06
40879  5.123024e-08  0.000236  4.334078e-06  3.729561e-07   3.426278e-06
40880  4.918103e-08  0.000220  6.147629e-07  1.270510e-07   5.143516e-07

[40881 rows x 5 columns]
```

```
[ ]  ca = pd.read_csv('/content/drive/MyDrive/yt_dataset/CAvideos.csv')

     numeric_columns = ca.select_dtypes(include='number')

     if numeric_columns.empty:
         raise ValueError("No numeric columns available to plot.")

     max_columns_to_plot = 10
     if numeric_columns.shape[1] > max_columns_to_plot:
         print(f"Limiting to the first {max_columns_to_plot} columns for plotting.")
         numeric_columns = numeric_columns.iloc[:, :max_columns_to_plot]

     transposed_data = numeric_columns.transpose()

     ax = transposed_data.plot.bar(figsize=(12, 8), legend=False)

     ax.set_xlabel('Parameters')
     ax.set_ylabel('Count')

     plt.tight_layout()
     plt.show()
```




```
[ ]  ca = pd.read_csv('/content/drive/MyDrive/yt_dataset/CAvideos.csv')

     columns_to_plot = [ 'likes', 'comment_count']  # Replace with your column names

     for col in columns_to_plot:
         if col not in ca.columns:
             raise ValueError(f"Column '{col}' not found in the DataFrame.")

     selected_columns = ca[columns_to_plot]

     numeric_columns = selected_columns.select_dtypes(include='number')
     if numeric_columns.empty:
         raise ValueError("No numeric columns selected to plot.")

     transposed_data = numeric_columns.transpose()

     ax = transposed_data.plot.bar(figsize=(12, 8), legend=False)

     ax.set_xlabel('Parameters')
     ax.set_ylabel('Count')

     plt.tight_layout()
     plt.show()
```

[ ]



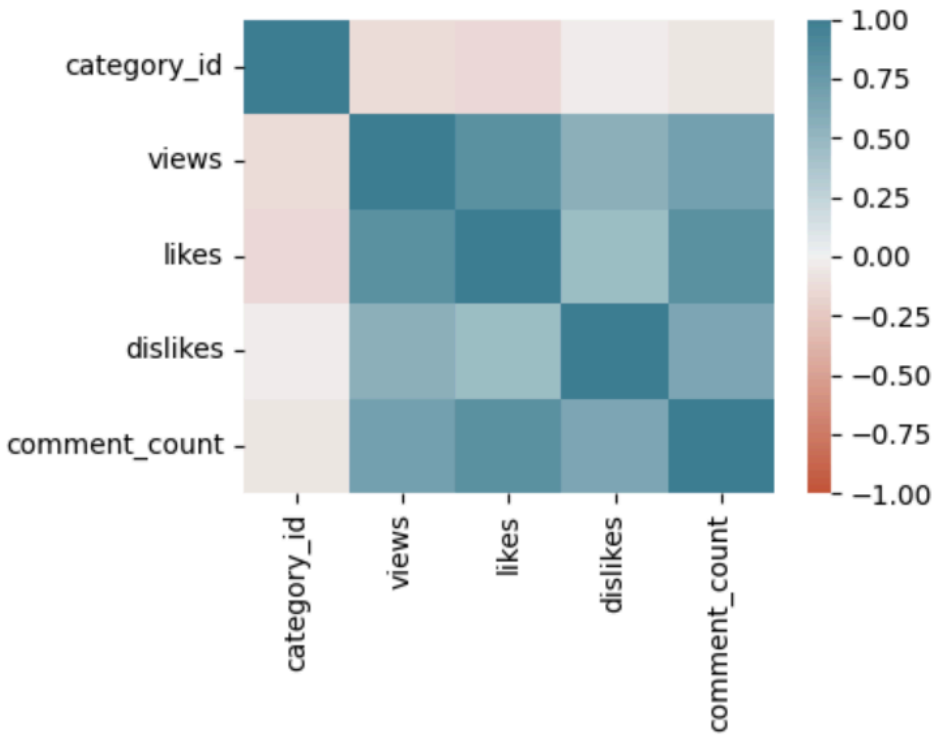
Correlation

```
[ ]  numeric_cols = ['likes', 'dislikes', 'comment_count', 'views', 'category_id']

     for col in numeric_cols:
         data[col] = pd.to_numeric(data[col], errors='coerce')

     data = data.dropna(subset=numeric_cols)

     corr_matrix = data[numeric_cols].corr()

     print(corr_matrix)
```

```
                   likes  dislikes  comment_count     views  category_id
likes           1.000000  0.460427       0.836585  0.828964    -0.144363
dislikes        0.460427  1.000000       0.643494  0.557621    -0.028731
comment_count   0.836585  0.643494       1.000000  0.693107    -0.068848
views           0.828964  0.557621       0.693107  1.000000    -0.139610
category_id    -0.144363 -0.028731      -0.068848 -0.139610     1.000000
```

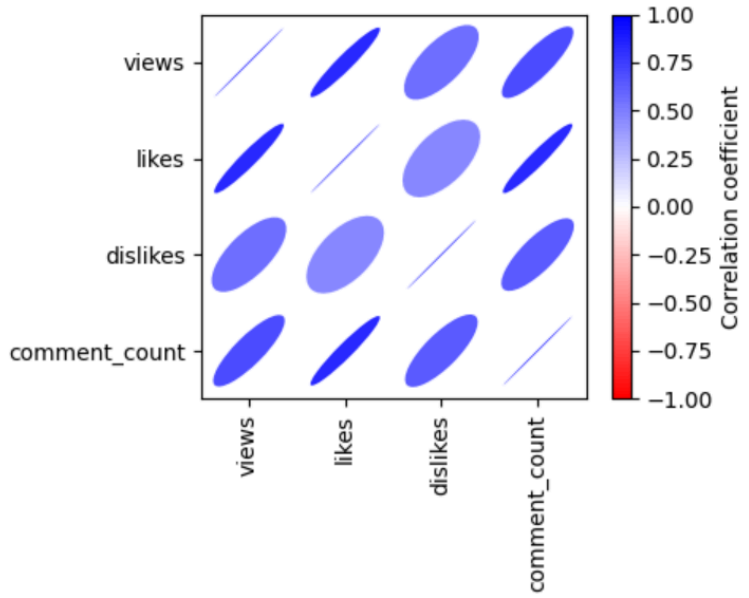
Heatmap

```
[ ]  numeric_data = data.select_dtypes(include=['number'])

     fig, ax = plt.subplots(figsize=(5, 4))

     ax = sns.heatmap(numeric_data.corr(), vmin=-1, vmax=1,
                      cmap=sns.diverging_palette(20, 220, as_cmap=True),
                      ax=ax)

     plt.tight_layout()

     plt.show()
```


category_id
views
likes
dislikes
comment_count
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

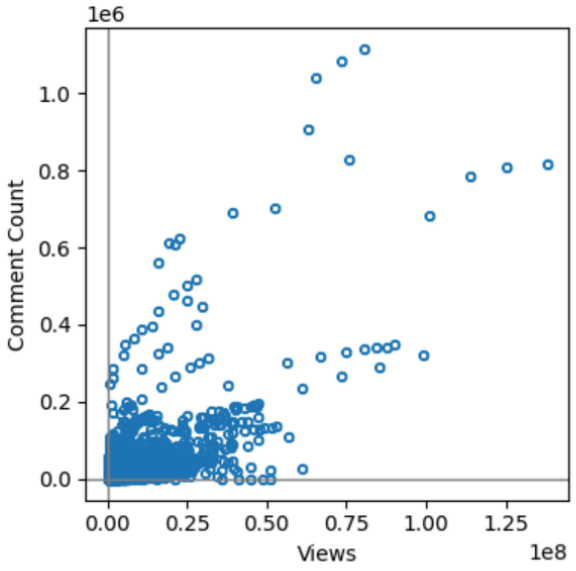
```
[ ]  from matplotlib.collections import EllipseCollection
     from matplotlib.colors import Normalize


     def plot_corr_ellipses(data, figsize=None, **kwargs):
         ''' https://stackoverflow.com/a/34558488 '''
         M = np.array(data)
         if not M.ndim == 2:
             raise ValueError('data must be a 2D array')

         fig, ax = plt.subplots(1, 1, figsize=figsize, subplot_kw={'aspect':'equal'})
         ax.set_xlim(-0.5, M.shape[1] - 0.5)
         ax.set_ylim(-0.5, M.shape[0] - 0.5)
         ax.invert_yaxis()

         xy = np.indices(M.shape)[::-1].reshape(2, -1).T

         w = np.ones_like(M).ravel() + 0.01
         h = 1 - np.abs(M).ravel() - 0.01
         a = 45 * np.sign(M).ravel()

         ec = EllipseCollection(widths=w, heights=h, angles=a, units='x', offsets=xy,
                                norm=Normalize(vmin=-1, vmax=1),
                                transOffset=ax.transData, array=M.ravel(), **kwargs)
         ax.add_collection(ec)

         if isinstance(data, pd.DataFrame):
             ax.set_xticks(np.arange(M.shape[1]))
             ax.set_xticklabels(data.columns, rotation=90)
             ax.set_yticks(np.arange(M.shape[0]))
             ax.set_yticklabels(data.index)

         return ec, ax

     columns = ['views', 'likes', 'dislikes', 'comment_count']
     subset = data[columns]

     m, ax = plot_corr_ellipses(subset.corr(), figsize=(5, 4), cmap='bwr_r')

     cb = plt.colorbar(m, ax=ax)
     cb.set_label('Correlation coefficient')
```

```
    plt.tight_layout()
    plt.show()
```



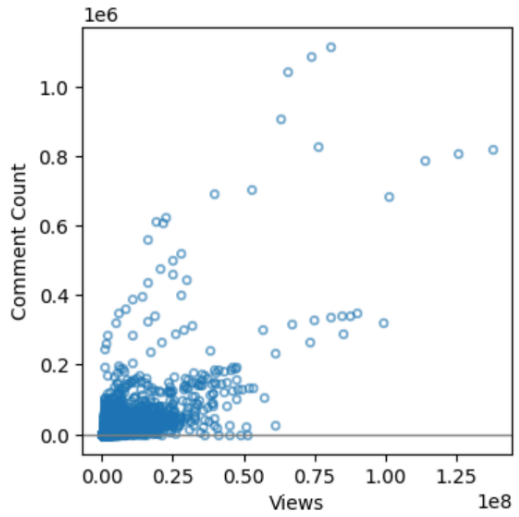
Scatterplots

```
[ ]  pk = data.plot.scatter(x='views', y='comment_count', figsize=(4, 4), marker='$\u25EF$')
     pk.set_xlabel('Views')
     pk.set_ylabel('Comment Count')
     pk.axhline(0, color='grey', lw=1)
     pk.axvline(0, color='grey', lw=1)

     plt.tight_layout()
     plt.show()
```



```
[ ]  pk = data.plot.scatter(x='views', y='comment_count', figsize=(4, 4), marker='$\u25EF$', alpha=0.5)
     pk.set_xlabel('Views')
     pk.set_ylabel('Comment Count')
     pk.axhline(0, color='grey', lw=1)
     print(ax.axvline(0, color='grey', lw=1))
```

```
Line2D(_child81762)
```



Exploring two or More Variables

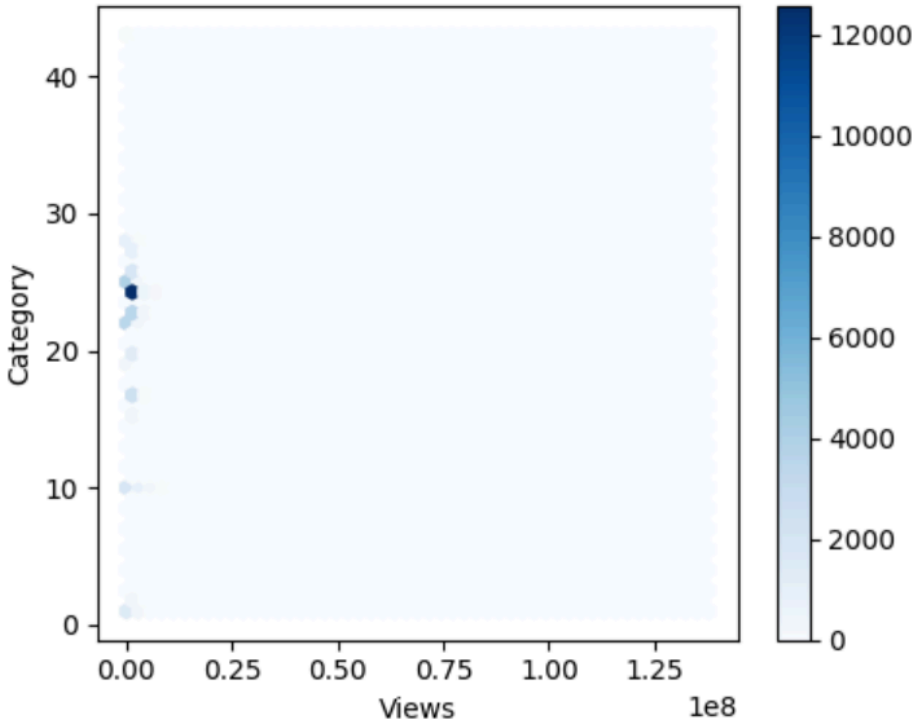
```
[ ]  kc= pd.read_csv('/content/drive/MyDrive/yt_dataset/CAvideos.csv')
     kp = kc.loc[(data.likes > 75_000_000) &
                         (data.views > 1_000_000_000) &
                         (data.comment_count > 10_000), :]
     print(kp.shape)
```

```
(0, 16)
```

Hexagonal Binning

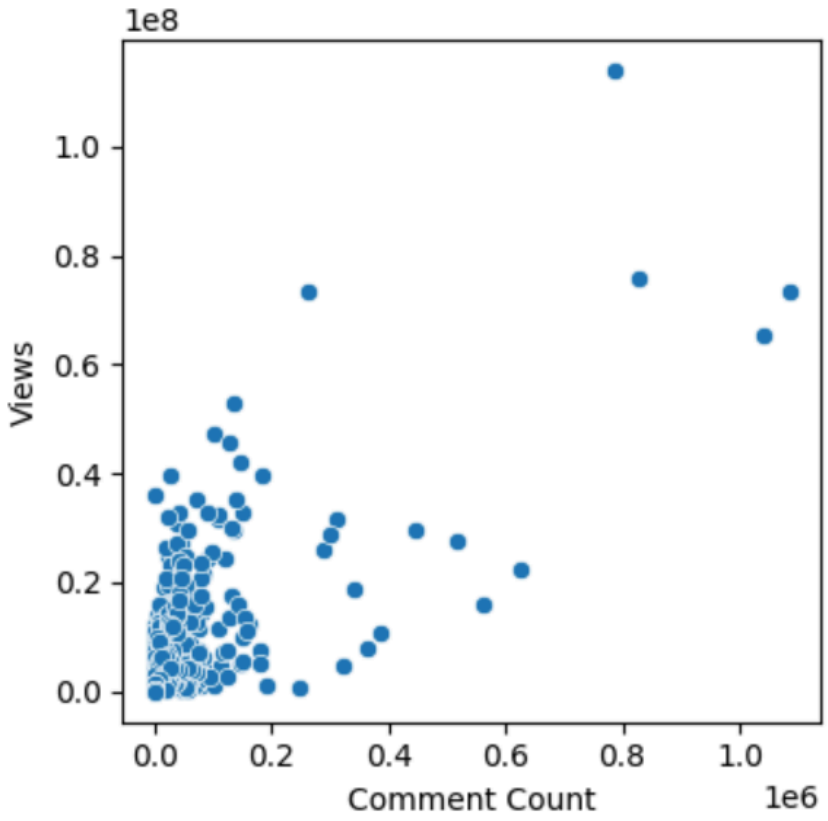
```
[ ]  data['views'] = pd.to_numeric(data['views'], errors='coerce')
     data['category_id'] = pd.to_numeric(data['category_id'], errors='coerce')

     data = data.dropna(subset=['views', 'category_id'])

     print(data.dtypes)

     ip = data.plot.hexbin(x='views', y='category_id', gridsize=50, sharex=False, figsize=(5, 4), reduce_C_function=np.mean, cmap='Blues')

     ip.set_xlabel('Views')
     ip.set_ylabel('Category')

     plt.tight_layout()
     plt.show()
```

```
video_id                  object
trending_date             object
title                     object
channel_title             object
category_id                int64
publish_time              object
tags                      object
views                      int64
likes                      int64
dislikes                   int64
comment_count              int64
thumbnail_link            object
comments_disabled           bool
ratings_disabled            bool
video_error_or_removed      bool
description               object
dtype: object
```



```
[ ]  data['comment_count'] = pd.to_numeric(data['comment_count'], errors='coerce')
     data['views'] = pd.to_numeric(data['views'], errors='coerce')

     data = data.dropna(subset=['comment_count', 'views'])

     print(data.dtypes)

     sample_size = min(len(data), 10000)

     fig, ax = plt.subplots(figsize=(4, 4))

     try:
         sns.kdeplot(data=data.sample(sample_size), x='comment_count', y='views', ax=ax)
         ax.set_xlabel('Comment Count')
         ax.set_ylabel('Views')
     except ValueError as e:
         print(f"KDE plot error: {e}")
         sns.scatterplot(data=data.sample(sample_size), x='comment_count', y='views', ax=ax)
         ax.set_xlabel('Comment Count')
         ax.set_ylabel('Views')

     plt.tight_layout()
     plt.show()
```

```
video_id                  object
trending_date             object
title                     object
channel_title             object
category_id                int64
publish_time              object
tags                      object
views                      int64
likes                      int64
dislikes                   int64
comment_count              int64
thumbnail_link            object
comments_disabled           bool
ratings_disabled            bool
video_error_or_removed      bool
description               object
dtype: object
KDE plot error: Contour levels must be increasing
```



Two Categorical Variables

```
[ ] file_path = '/content/drive/MyDrive/yt_dataset/CAvideos.csv'
    data = pd.read_csv(file_path)

    data['category_id'] = pd.to_numeric(data['category_id'], errors='coerce')

    crosstab = data.pivot_table(index='category_id', columns='channel_title', aggfunc=lambda x: len(x), margins=True)
    print(crosstab)
```

```
20                NaN          NaN                    NaN  ...     NaN
22                NaN          NaN                    NaN  ...     NaN
23                NaN          NaN                    NaN  ...     NaN
24                NaN          1.0                    3.0  ...     NaN
25                NaN          NaN                    NaN  ...     NaN
26                NaN          NaN                    NaN  ...     NaN
27                NaN          NaN                    NaN  ...     NaN
28                NaN          NaN                    NaN  ...     NaN
29                NaN          NaN                    NaN  ...     NaN
30                NaN          NaN                    NaN  ...     NaN
43                NaN          NaN                    NaN  ...     NaN
All               NaN          1.0                    3.0  ...     1.0

                                                                      \
channel_title 이슈사건사고 이영애 (Lee Young - Ae)  종합뉴스 창조영감클럽 타우TV 포스트쉐어 포크포크 활력소TV
category_id
1                 NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
2                 NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
10                NaN                  2.0   NaN    NaN  NaN   NaN  NaN   NaN
15                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
17                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
19                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
20                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
22                NaN                  NaN   NaN    1.0  NaN   NaN  NaN   NaN
23                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
24                1.0                  NaN   NaN    NaN  NaN   1.0  NaN   1.0
25                NaN                  NaN  11.0    NaN  NaN   NaN  NaN   NaN
26                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
27                NaN                  NaN   NaN    NaN  NaN   NaN  1.0   NaN
28                NaN                  NaN   NaN    NaN  1.0   NaN  NaN   NaN
29                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
30                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
43                NaN                  NaN   NaN    NaN  NaN   NaN  NaN   NaN
All               1.0                  2.0   1.0    NaN  1.0   1.0  1.0   1.0
```

```
[ ]
channel_title    All
category_id
1               2001
2                348
10              3695
15               369
17              2650
19               377
20              1330
22              3726
23              3725
24             13173
25              3868
26              1998
27               982
28              1143
29                70
30                 6
43               124
All            39585

[18 rows x 71078 columns]
```

```python
crosstab.index = pd.to_numeric(crosstab.index, errors='coerce')

df = crosstab.loc[(crosstab.index >= 1) & (crosstab.index <= 7), :]

columns_to_normalize = ['views', 'likes', 'dislikes', 'comment_count']  # Specify columns to normalize
for col in columns_to_normalize:
    if col in df.columns:
        df[col] = df[col].div(df.sum(axis=1), axis=0)  # Normalize by row totals

if 'All' in crosstab.columns:
    df['All'] = df['All'] / sum(df['All'])

perc_crosstab = df

print(perc_crosstab)
```

```
<ipython-input-13-344e62d88e0b>:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df[col] = df[col].div(df.sum(axis=1), axis=0)  # Normalize by row totals
<ipython-input-13-344e62d88e0b>:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df[col] = df[col].div(df.sum(axis=1), axis=0)  # Normalize by row totals
<ipython-input-13-344e62d88e0b>:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df[col] = df[col].div(df.sum(axis=1), axis=0)  # Normalize by row totals
<ipython-input-13-344e62d88e0b>:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df[col] = df[col].div(df.sum(axis=1), axis=0)  # Normalize by row totals
              comment_count                                             \
channel_title  #AndresSTyle #Mind Warehouse #SeekingTheTruth * Martyna *
category_id
1.0                     NaN             NaN              NaN         NaN
2.0                     NaN             NaN              NaN         NaN

                                                                      \
channel_title - 欢迎订阅 -浙江卫视【奔跑吧】官方频道 -Wen Zhao Official文昭談古論今 078jordan1
category_id
1.0                                NaN                       NaN         NaN
2.0                                NaN                       NaN         NaN

                                                         ...     views  \
channel_title 0b1knob 10 MillionTM 10-Minutes Satisfaction  ...   웃지 UTZI
category_id                                                 ...
1.0                NaN          NaN                   NaN  ...  0.000018
2.0                NaN          NaN                   NaN  ...       NaN
```

```
channel_title 이슈사건사고 이영애 (Lee Young - Ae) 종합뉴스 창조영감클럽 타우TV 포스트쉐어 포크포크 활력소TV
category_id
1.0                NaN                  NaN  NaN    NaN  NaN   NaN  NaN   NaN
2.0                NaN                  NaN  NaN    NaN  NaN   NaN  NaN   NaN


channel_title       All
category_id
1.0            0.035195
2.0            0.035460

[2 rows x 71078 columns]
```

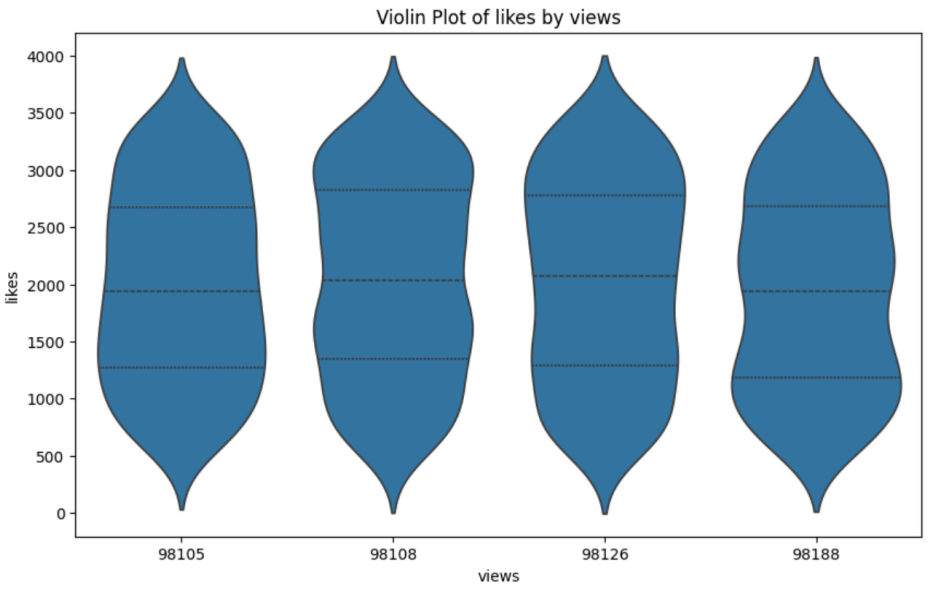
Categorical and Numeric Data

```
[ ] data = pd.read_csv('/content/drive/MyDrive/yt_dataset/CAvideos.csv')
```

```
[ ] data.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	thumbn
0	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	17158579	787425	43420	125882	https://i.ytimg.com/vi/n1WpP7iowLc/
1	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	1014651	127794	1688	13030	https://i.ytimg.com/vi/0dBIkQ4Mz1M/
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146035	5339	8181	https://i.ytimg.com/vi/5qpjK5DgCt4/
3	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095828	132239	1989	17518	https://i.ytimg.com/vi/d380meD0W0M/
4	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	33523622	1634130	21082	85067	https://i.ytimg.com/vi/2Vv-BfVoq4g/

```
[ ] plt.figure(figsize=(10, 6))
    sns.violinplot(x='views', y='likes', data=data, inner='quartile')
    plt.title('Violin Plot of likes by views')
    plt.show()
```



Visualizing Multiple Variables

```
[ ]  video_ids = ['5qpjK5DgCt4', 'n1WpP7iowLc', '2kyS6SvSYSE', '7MxiQ4vOEnE']  # Replace with your values
     filtered_df = df.loc[df['video_id'].isin(video_ids),:]

     # Check if the filtered DataFrame is not empty
     if filtered_df.empty:
         print("Filtered DataFrame is empty. Please check your filtering criteria.")
     else:
         # Define the hexbin function
         def hexbin(x, y, color, **kwargs):
             cmap = sns.light_palette(color, as_cmap=True)
             plt.hexbin(x, y, gridsize=25, cmap=cmap, **kwargs)

         # Assuming you want to plot 'views' vs 'likes'
         g = sns.FacetGrid(filtered_df, col='video_id', col_wrap=2)
         g.map(hexbin, 'views', 'likes', extent=[0, filtered_df['views'].max(), 0, filtered_df['likes'].max()])
         g.set_axis_labels('Views', 'Likes')
         g.set_titles('Video ID {col_name}')

         plt.tight_layout()
         plt.show()
```