USvideos.csv 数据集分析

1.数据摘要

1.1 读取数据

根据数据类型具体分析数值属性

```
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
```

1.2 标称属性可能取值的频数

(1)category_id

```
24    9964
10    6472
26    4146
23    3457
22    3210
25    2487
28    2401
1     2345
17    2174
27    1656
15     920
20     817
19     402
2      384
29      57
43      57
```

(2) comments_disabled

```
False                40316
True                   633
```

(3) variety

```
False                40780
True                   169
```

(4) ratings_disabled

```
False                40926
True                    23
```

1.3 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数

选取数值属性，分别使用.max(),.min(),.mean(),.median(),.quantile()等函数获取属性最大、最小、均值、中位数、四分位数。使用.isnull().sum()函数获取缺失值个数。

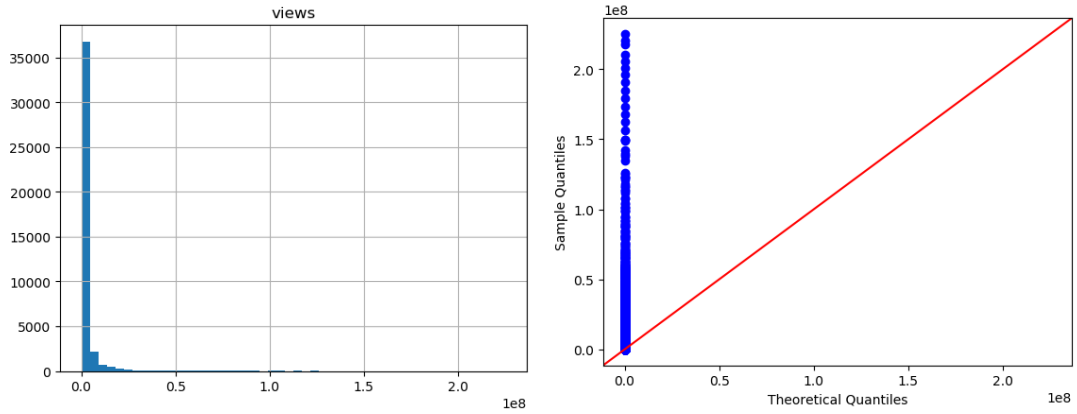
```
数值属性最大值：
views            225211923
likes              5613827
dislikes           1674420
comment_count      1361580
dtype: int64
数值属性最小值：
views            549
likes              0
dislikes           0
comment_count      0
dtype: int64
数值属性均值：
views            2.360785e+06
likes            7.426670e+04
dislikes         3.711401e+03
comment_count    8.446804e+03
dtype: float64
数值属性中位数：
views            681861.0
likes             18091.0
dislikes            631.0
comment_count      1856.0
dtype: float64
数值属性四分位数：
                    0.25      0.50       0.75
views           242329.0  681861.0  1823157.0
likes             5424.0   18091.0    55417.0
dislikes           202.0     631.0     1938.0
comment_count      614.0    1856.0     5755.0
数值属性缺失值：
comment_count    0
dislikes         0
likes            0
views            0
```

2.数据的可视化

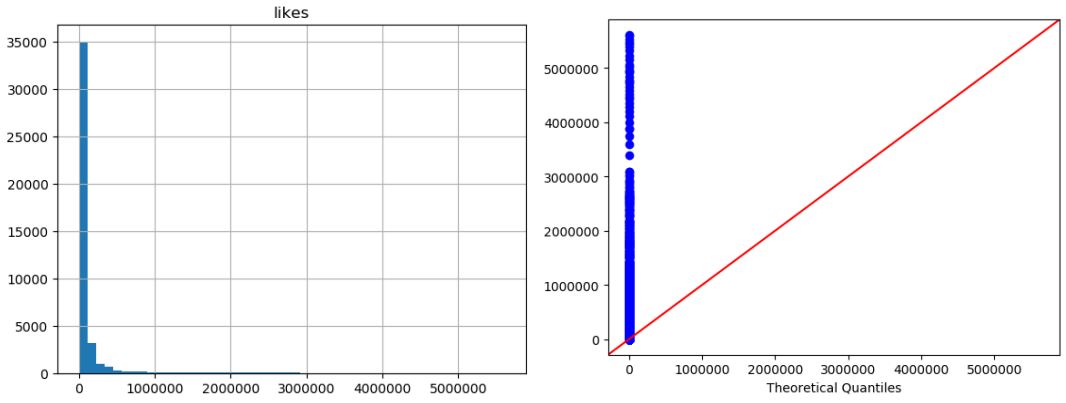
2.1 绘制直方图

导入 `matplotlib.pyplot` 模块，用于数据可视化。导入 `statsmodels.api` 统计分析库，用于用 qq 图检验其分布。使用 `hist(bins = XX)`函数，绘制直方图。使用 `qqplot(df,line='45')`绘制 qq 图检验数据分布是否为正态分布。

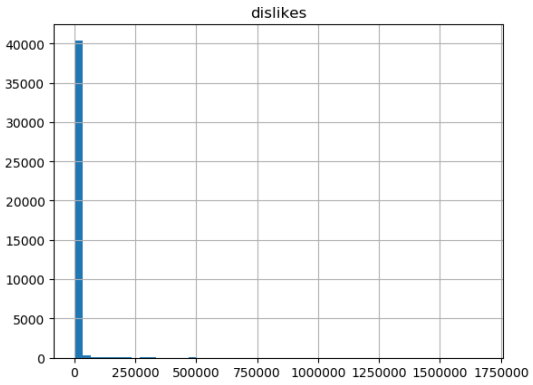
(1) view

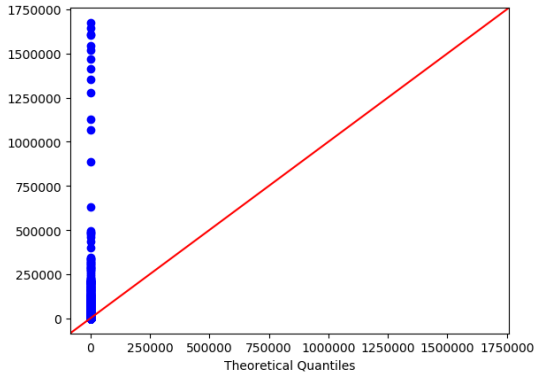


(2) likes

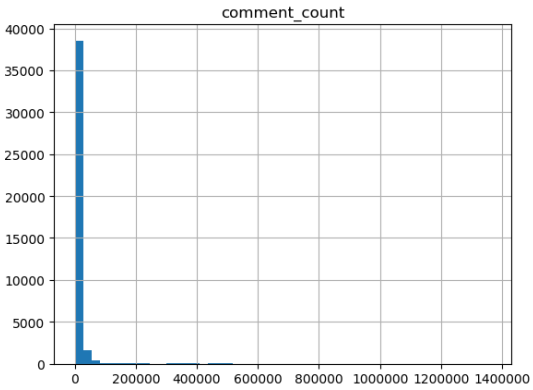


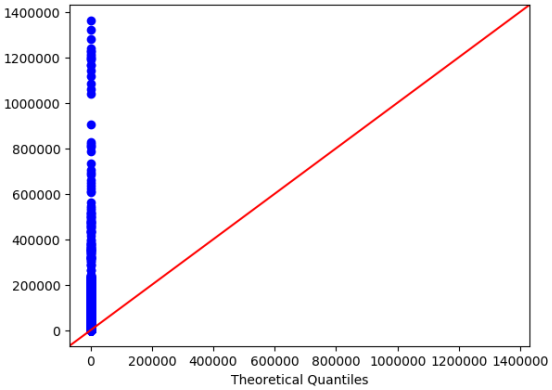
(3) dislikes





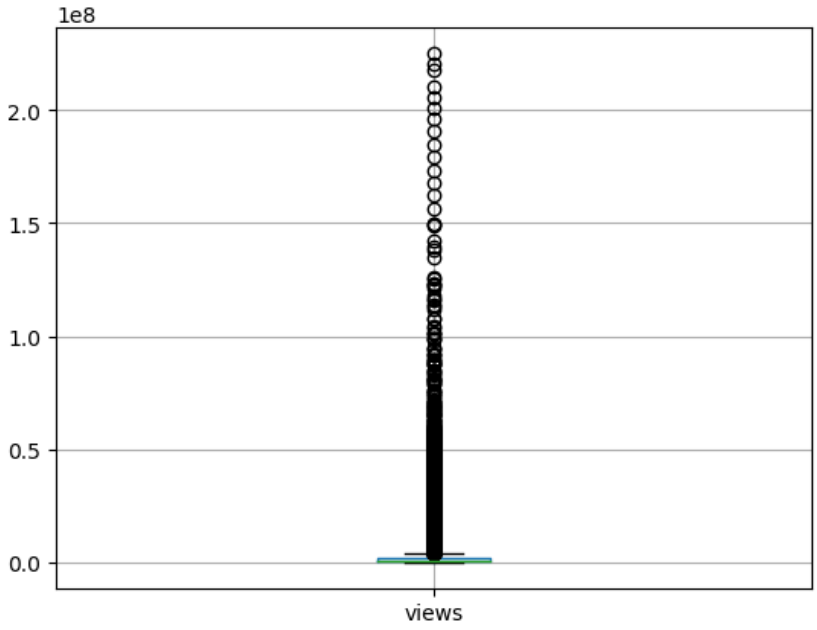
(4) comment_count

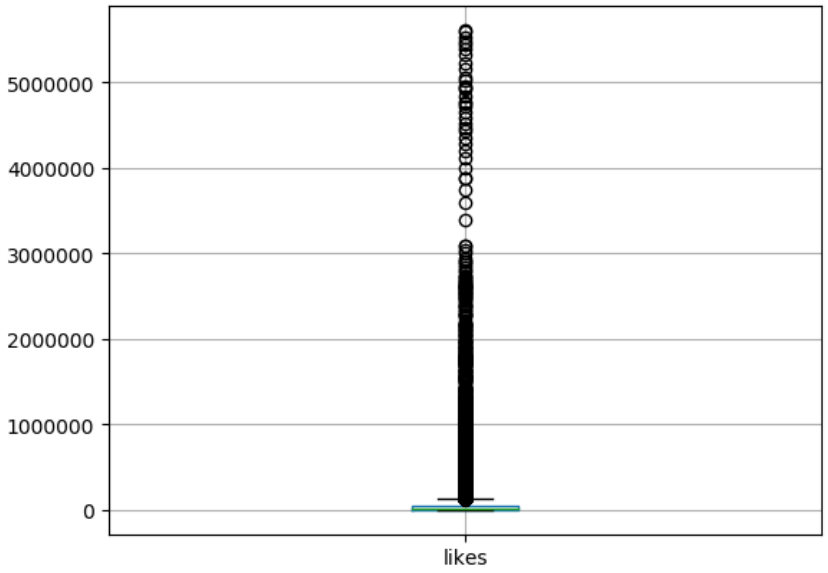


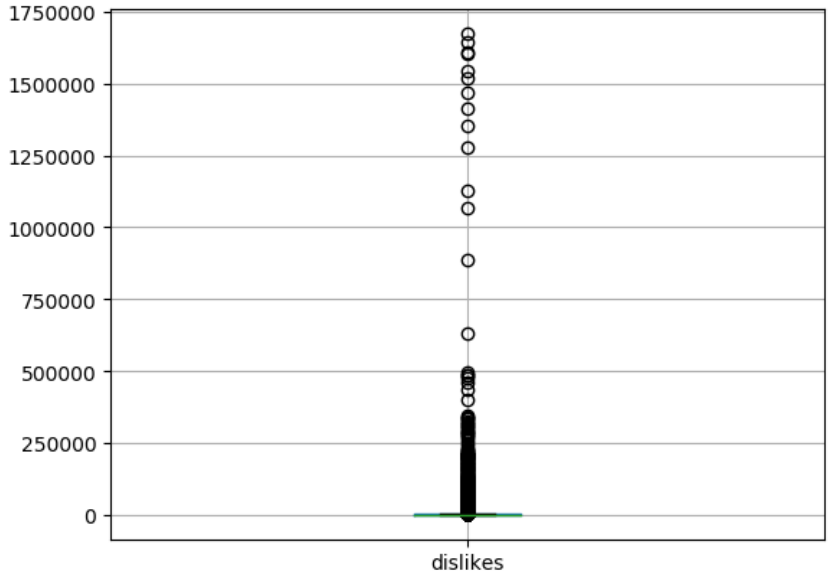


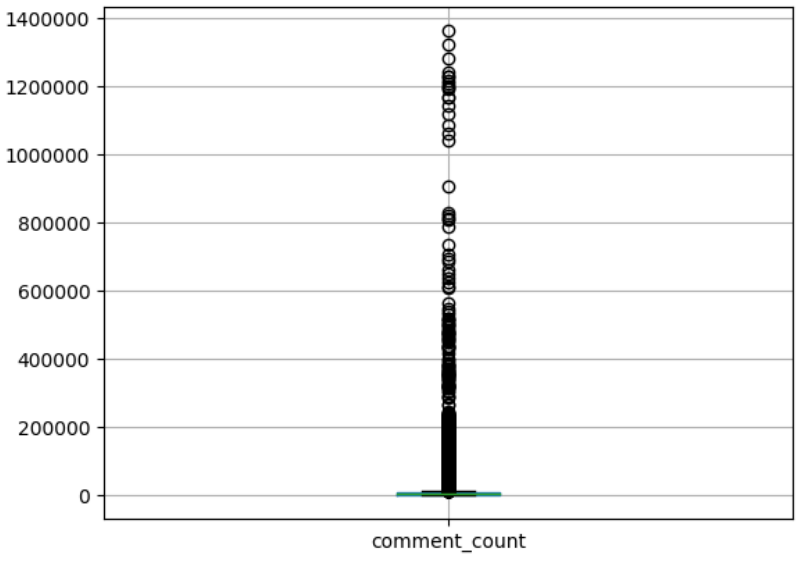
2.2 绘制盒图

使用 `boxplot()`函数，绘制盒图，对离群值进行识别




5000000
4000000
3000000
2000000
1000000
0
likes

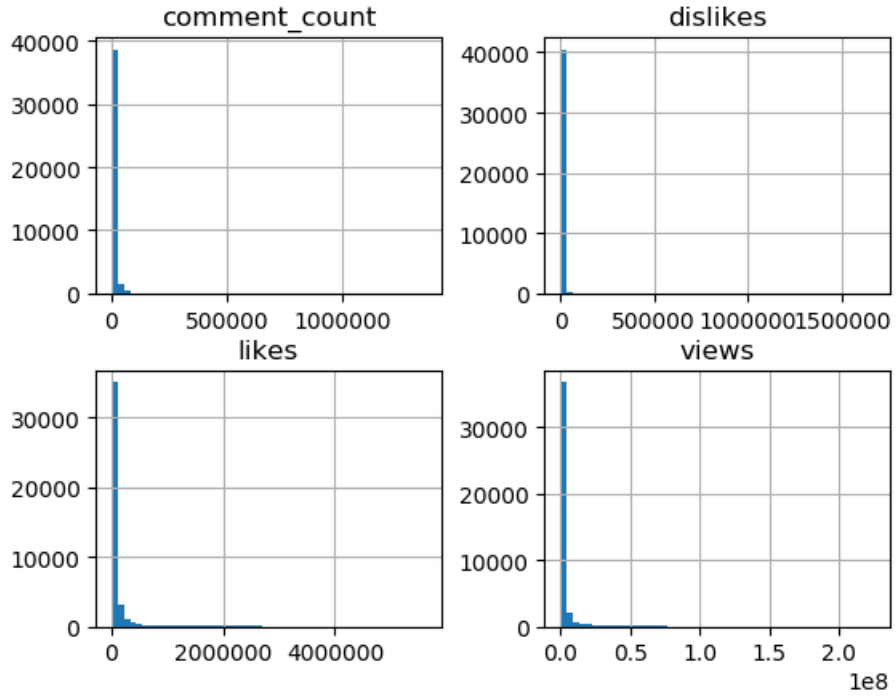

1750000
1500000
1250000
1000000
750000
500000
250000
0
dislikes


1400000
1200000
1000000
800000
600000
400000
200000
0
comment_count

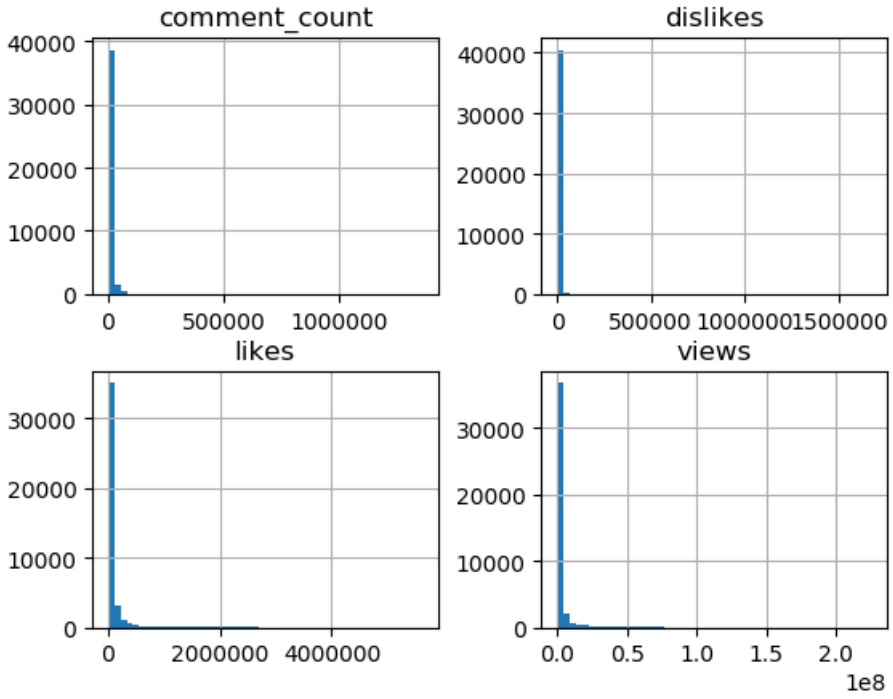
3 数据缺失的处理

3.1 将缺失部分剔除

使用 dropna()函数将缺失值进行剔除，并用直方图可视化地对比新旧数据集。
旧数据集



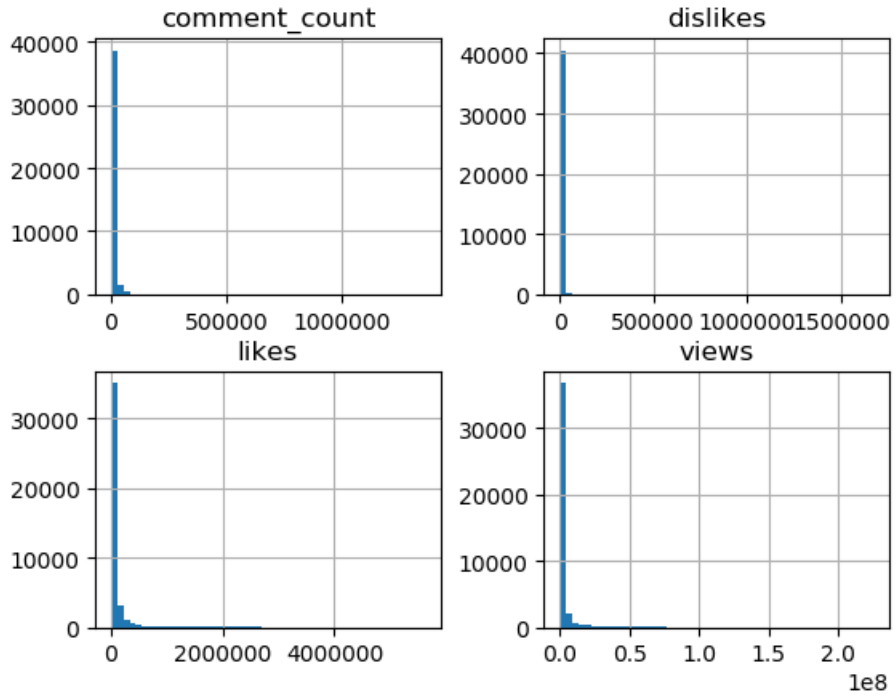
新数据集



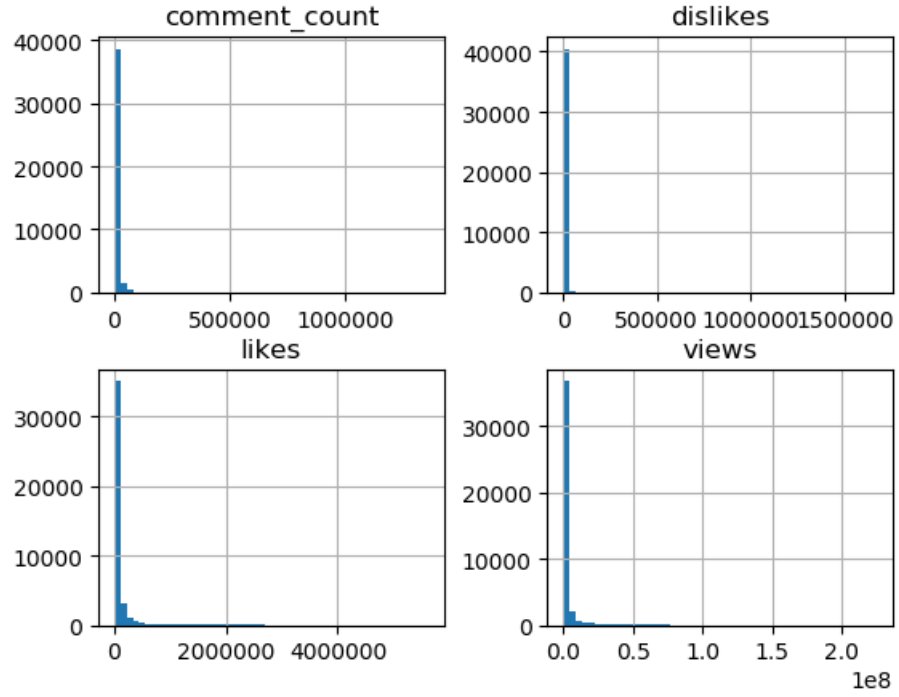
3.2 用最高频率值来填补缺失值

使用 `mode()`函数获取众数，使用 `fillna()`函数填补缺失值。并用直方图可视化地对比新旧数据集。

旧数据集



新数据集



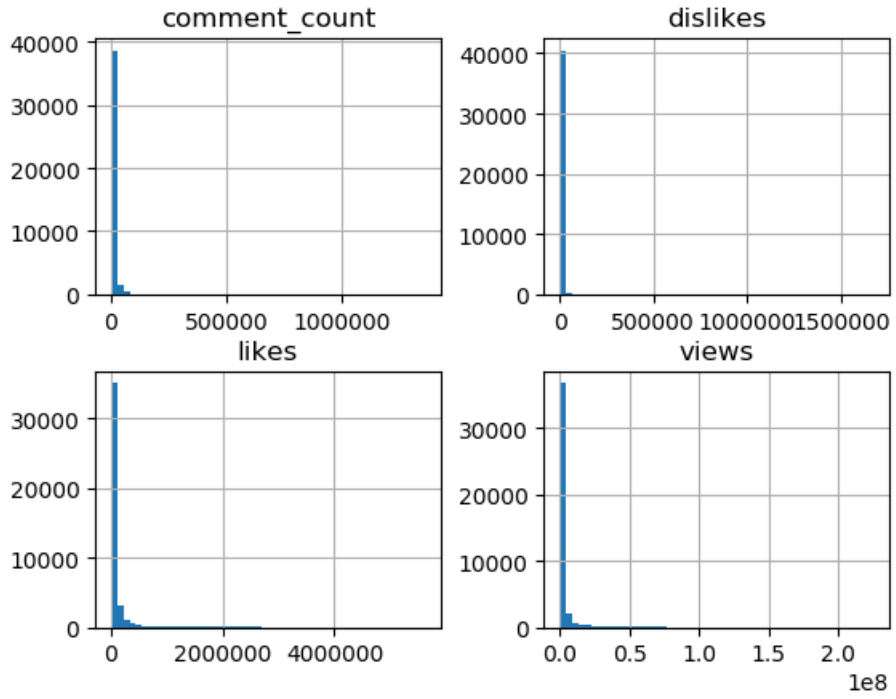
3.3 通过属性的相关关系来填补缺失值

使用 corr()函数获取数值属性相关系数。并输出到 corr.csv 文件，查看相关关系。根据相关关系进行填充。使用 apply()函数，对强相关属性进行填充，并用直方图可视化地对比新旧数据集。

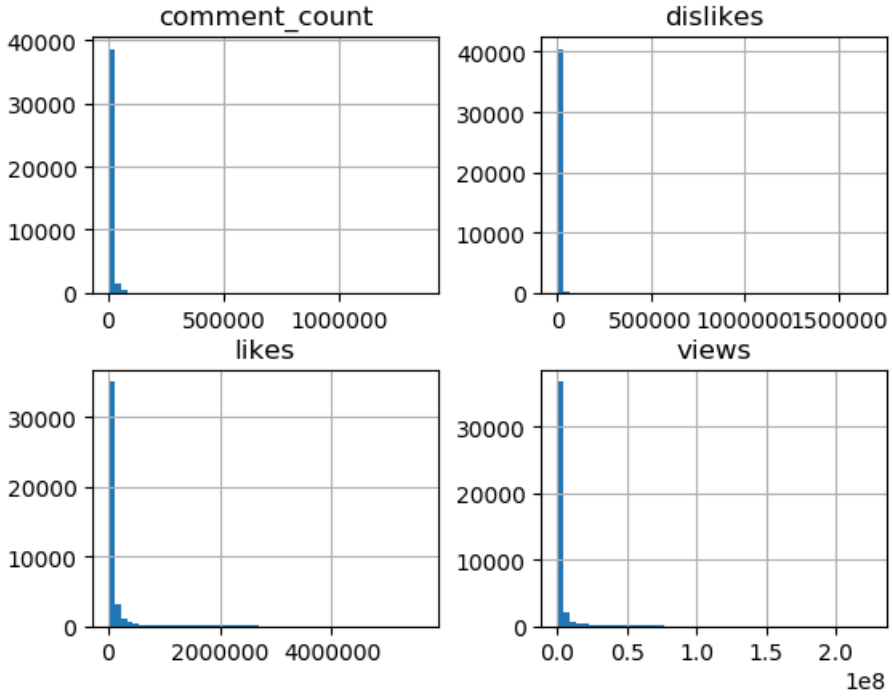
相关性

	views	likes	dislikes	comment_count
views	1	0.849177	0.472213	0.617621
likes	0.849177	1	0.447186	0.803057
dislikes	0.472213	0.447186	1	0.700184
comment_count	0.617621	0.803057	0.700184	1

填充前



填充后


comment_count
dislikes
likes
views
0
500000
1000000
1500000
2000000
4000000
10000
20000
30000
40000
0.0
0.5
1.0
1.5
2.0
1e8