数据探索性分析与数据预处理

陈星奎 3120230922

作业要求：1. 数据摘要和可视化， 2. 数据缺失的处理

本次报告由juypter notebook生成，对两个数据集分别进行分析。首先获取该数据集的属性列表，然后调用自行实现的函数valueAnalysis进行属性分析，如果是标称属性则会根据频数排序并列出指定数目的前n个取值；如果是数值属性则会输出五数概括并作出直方图和盒图，若存在缺失值则会根据不同策略进行补全并进行比较。具体的函数实现和参数请参考代码仓库。

在本次作业中选择的数据集为：

美国阿尔茨海默症和健康老龄化数据集：Alzheimer Disease and Healthy Aging Data In US

盗版网站的电影数据集：Movies Dataset from Pirated Sites

代码仓库：https://github.com/Hannew-Kui/data-analysis-and-preprocessing

In [2]:
```
#引入外部模块
from dataAnalysis import *
```

数据集1：美国阿尔茨海默症和健康老龄化数据集

获取属性列表：(警告信息可忽略)

In [3]:
```
data_df,header=readDataset("./dataset/Alzheimer Disease and Healthy Aging Data I
data_total_length=len(data_df)
print(header)                                                    #属性列表
print("\ndata_total_length={}".format(data_total_length))    #数据项总数目
```

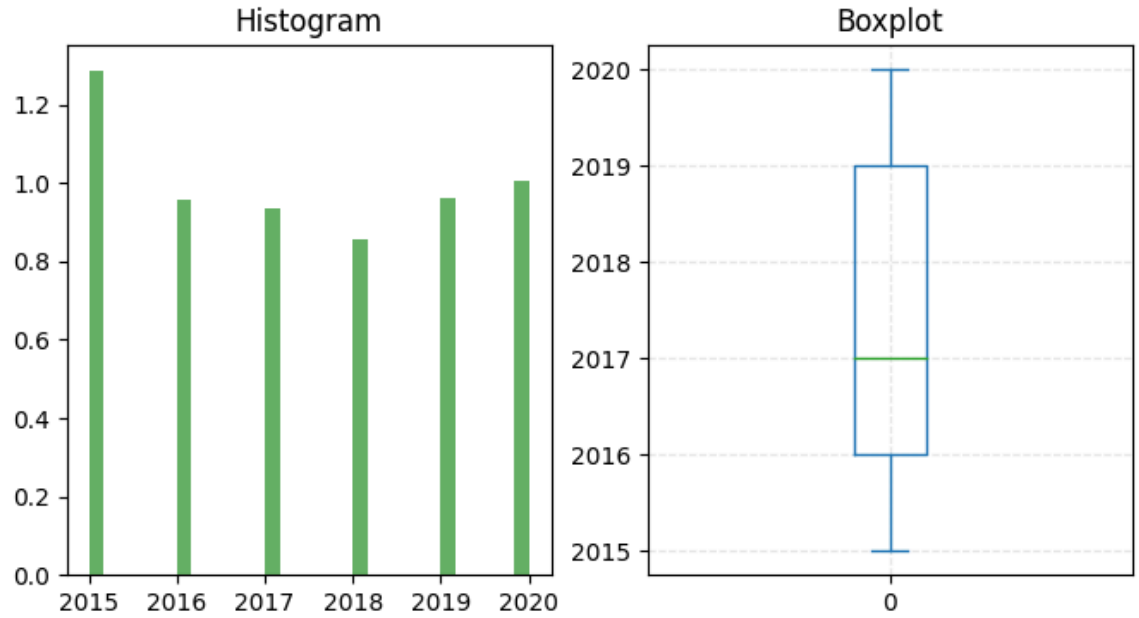
```
['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'Datasource', 'Class',
'Topic', 'Question', 'Data_Value_Unit', 'DataValueTypeID', 'Data_Value_Type',
'Data_Value', 'Data_Value_Alt', 'Low_Confidence_Limit', 'High_Confidence_Limi
t', 'Sample_Size', 'StratificationCategory1', 'Stratification1', 'Stratificatio
nCategory2', 'Stratification2', 'Geolocation', 'ClassID', 'TopicID', 'QuestionI
D', 'LocationID', 'StratificationCategoryID1', 'StratificationID1', 'Stratifica
tionCategoryID2', 'StratificationID2']

data_total_length=214462
C:\Users\13473\AppData\Local\Temp\ipykernel_14632\4112351936.py:1: DtypeWarnin
g: Columns (13,14) have mixed types.Specify dtype option on import or set low_m
emory=False.
  data_df,header=readDataset("./dataset/Alzheimer Disease and Healthy Aging Dat
a In US.csv")
```

YearStart: 该元组数据的开始收集年份

In [3]:
```
valueAnalysis(data_df,"YearStart")
```

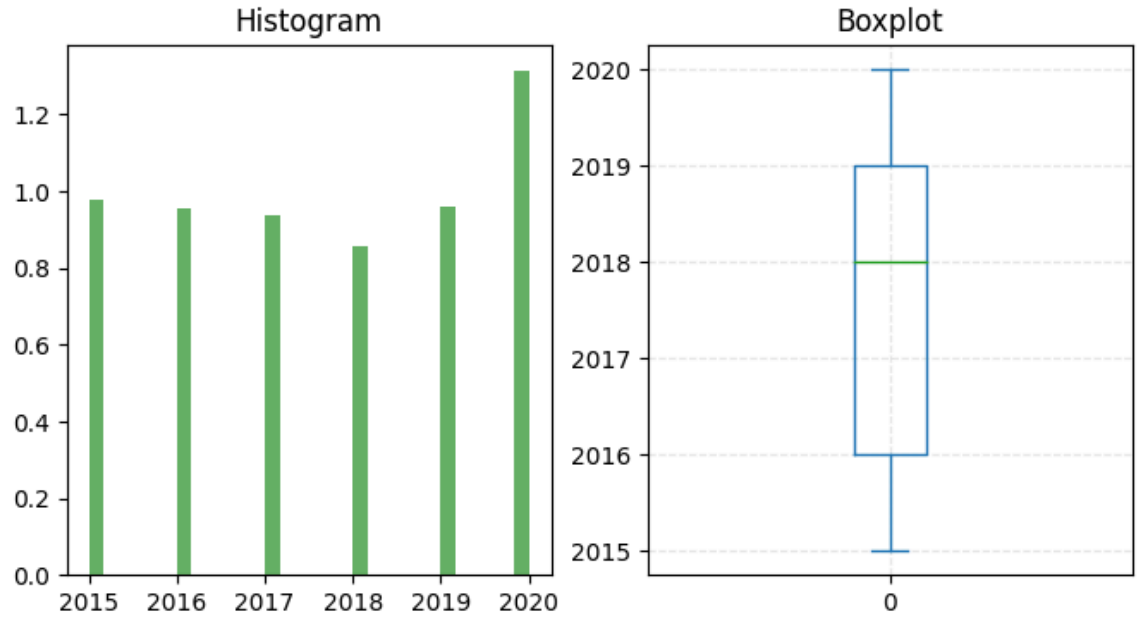
```
YearStart: numberic attribute
Valid:214462
Missing:0
Five number summary: min: 2015, Q1: 2016.00, median: 2017.00, Q3: 2019.00, max:
2020
```



YearEnd: 该元组数据的结束收集年份

```
In [4]: valueAnalysis(data_df,"YearEnd")
```

```
YearEnd: numberic attribute
Valid:214462
Missing:0
Five number summary: min: 2015, Q1: 2016.00, median: 2018.00, Q3: 2019.00, max:
2020
```



由上述可以看出，该数据集的收集年份范围在2015年到2020年之间，大多集中于2016到2019年之间

LocationAbbr: 数据收集时的地区简写

In [5]:
```
valueAnalysis(data_df,"LocationAbbr",20)
```

```
LocationAbbr: nominal attribute
Valid:214462
Missing:0
{
    "US": 4644,
    "WEST": 4638,
    "NRE": 4614,
    "MDW": 4611,
    "OR": 4565,
    "NY": 4557,
    "SOU": 4542,
    "UT": 4222,
    "OH": 3955,
    "GA": 3951,
    "MD": 3919,
    "HI": 3907,
    "TN": 3879,
    "MI": 3796,
    "VA": 3758,
    "FL": 3753,
    "ME": 3733,
    "TX": 3699,
    "NV": 3696,
    "DC": 3684
}
```

LocationDesc: 该数据收集时的地区全名

In [10]:
```
valueAnalysis(data_df,"LocationDesc",20)
```

```
LocationDesc: nominal attribute
Valid:214462
Missing:0
{
    "United States, DC & Territories": 4644,
    "West": 4638,
    "Northeast": 4614,
    "Midwest": 4611,
    "Oregon": 4565,
    "New York": 4557,
    "South": 4542,
    "Utah": 4222,
    "Ohio": 3955,
    "Georgia": 3951,
    "Maryland": 3919,
    "Hawaii": 3907,
    "Tennessee": 3879,
    "Michigan": 3796,
    "Virginia": 3758,
    "Florida": 3753,
    "Maine": 3733,
    "Texas": 3699,
    "Nevada": 3696,
    "District of Columbia": 3684
}
```

由上述可以看出，该数据集大都采集于美国哥伦比亚、美国西部、美国东北部等地区。

Datasource: 数据来源

In [6]:
```
valueAnalysis(data_df,"Datasource")
```

```
Datasource: nominal attribute
Valid:214462
Missing:0
{
    "BRFSS": 214462
}
```

由上述数据可以看出，该数据集全部来自于BRFSS系统

Class: 数据的类别

In [7]:
```
valueAnalysis(data_df,"Class")
```

```
Class: nominal attribute
Valid:214462
Missing:0
{
    "Overall Health": 71694,
    "Screenings and Vaccines": 46867,
    "Nutrition/Physical Activity/Obesity": 24851,
    "Cognitive Decline": 19180,
    "Caregiving": 18671,
    "Mental Health": 16600,
    "Smoking and Alcohol Use": 16599
}
```

Topic: 数据的主题

In [8]:
```
valueAnalysis(data_df,"Topic",20)
```

```
Topic: nominal attribute
Valid:214462
Missing:0
{
    "Lifetime diagnosis of depression": 8300,
    "Physically unhealthy days (mean number of days)": 8300,
    "Frequent mental distress": 8300,
    "Influenza vaccine within past year": 8300,
    "No leisure-time physical activity within past month": 8300,
    "Obesity": 8300,
    "Current smoking": 8300,
    "Binge drinking within past 30 days": 8299,
    "Self-rated health (good to excellent health)": 8299,
    "Self-rated health (fair to poor health)": 8299,
    "Ever had pneumococcal vaccine": 8268,
    "Recent activity limitations in past month": 8233,
    "Disability status, including sensory or mobility limitations": 6917,
    "Arthritis among older adults": 5511,
    "Fair or poor health among older adults with arthritis": 5447,
    "Subjective cognitive decline or memory loss among older adults": 5088,
    "Diabetes screening within past 3 years": 4808,
    "Talked with health care professional about subjective cognitive decline or
memory loss": 4700,
    "Need assistance with day-to-day activities because of subjective cognitive
decline or memory loss": 4696,
    "Functional difficulties associated with subjective cognitive decline or me
mory loss among older adults": 4696
}
```

由上述数据可以看出，该数据集的主题和类别主要和健康相关。

Question: 与数据相关联的问题

In [9]:
```
valueAnalysis(data_df,"Question",20)
```

```
Question: nominal attribute
Valid:214462
Missing:0
{
    "Percentage of older adults with a lifetime diagnosis of depression": 8300,
    "Physically unhealthy days (mean number of days in past month)": 8300,
    "Percentage of older adults who are experiencing frequent mental distress":
8300,
    "Percentage of older adults who reported influenza vaccine within the past
year": 8300,
    "Percentage of older adults who have not had any leisure time physical acti
vity in the past month": 8300,
    "Percentage of older adults who are currently obese, with a body mass index
(BMI) of 30 or more": 8300,
    "Percentage of older adults who have smoked at least 100 cigarettes in thei
r entire life and still smoke every day or some days": 8300,
    "Percentage of older adults who reported binge drinking within the past 30
days": 8299,
    "Percentage of older adults who self-reported that their health is \"good
\", \"very good\", or \"excellent\"": 8299,
    "Percentage of older adults who self-reported that their health is \"fair\"
or \"poor\"": 8299,
    "Percentage of at risk adults (have diabetes, asthma, cardiovascular diseas
e or currently smoke) who ever had a pneumococcal vaccine": 8268,
    "Mean number of days with activity limitations in the past month": 8233,
    "Percentage of older adults who report having a disability (includes limita
tions related to sensory or mobility impairments or a physical, mental, or emot
ional condition)": 6917,
    "Percentage of older adults ever told they have arthritis": 5511,
    "Fair or poor health among older adults with doctor-diagnosed arthritis": 5
447,
    "Percentage of older adults who reported subjective cognitive decline or me
mory loss that is happening more often or is getting worse in the preceding 12
months": 5088,
    "Percentage of older adults without diabetes who reported a blood sugar or
diabetes test within 3 years": 4808,
    "Percentage of older adults with subjective cognitive decline or memory los
s who reported talking with a health care professional about it": 4700,
    "Percentage of older adults who reported that as a result of subjective cog
nitive decline or memory loss that they need assistance with day-to-day activit
ies": 4696,
    "Percentage of older adults who reported subjective cognitive decline or me
mory loss that interferes with their ability to engage in social activities or
household chores": 4696
}
```

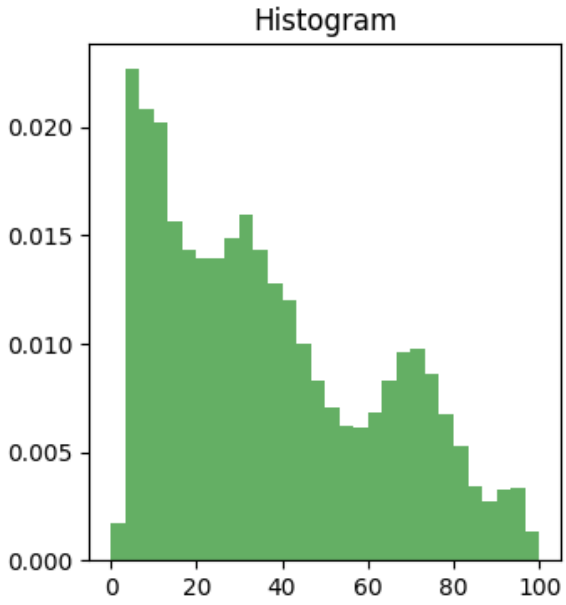
In [10]:
```
valueAnalysis(data_df,"Data_Value_Unit")
```

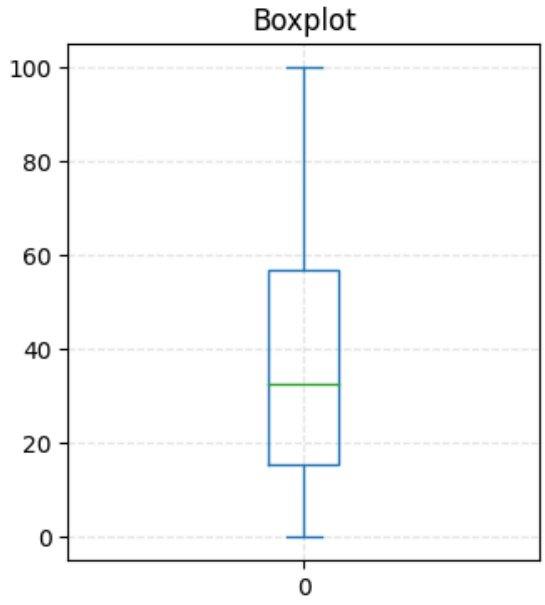
```
Data_Value_Unit: nominal attribute
Valid:214462
Missing:0
{
    "%": 197929,
    "Number": 16533
}
```

Data_Value: 数据实际数值（此处即指发病率）

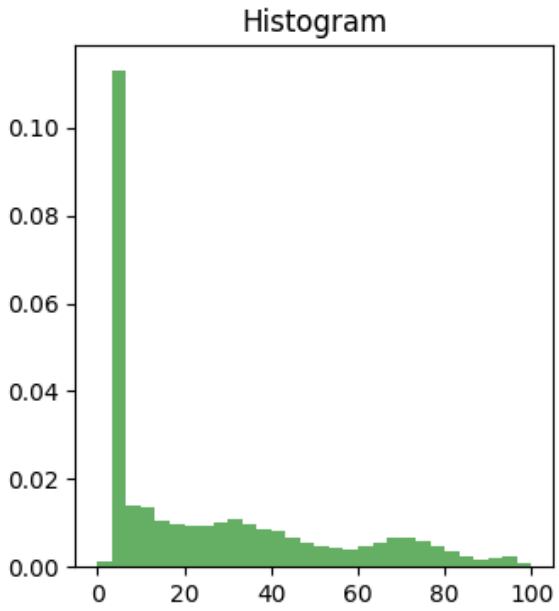
In [11]:
```
valueAnalysis(data_df,"Data_Value")
```

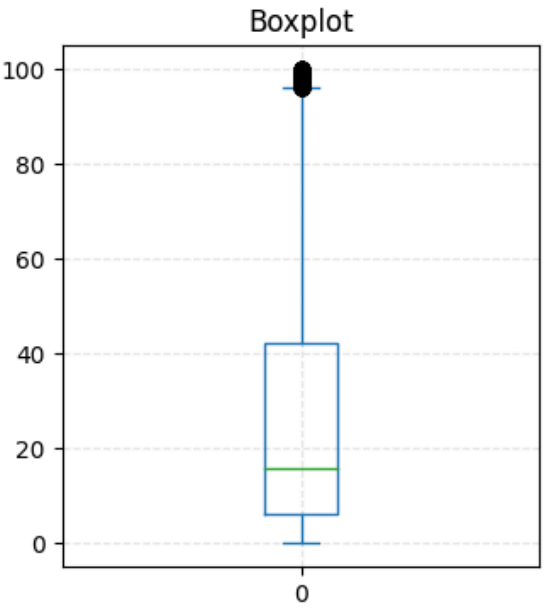
```
Data_Value: numberic attribute
Valid:144629
Missing:69833
Five number summary: min: 0.0, Q1: 15.30, median: 32.50, Q3: 56.80, max: 100.0
```



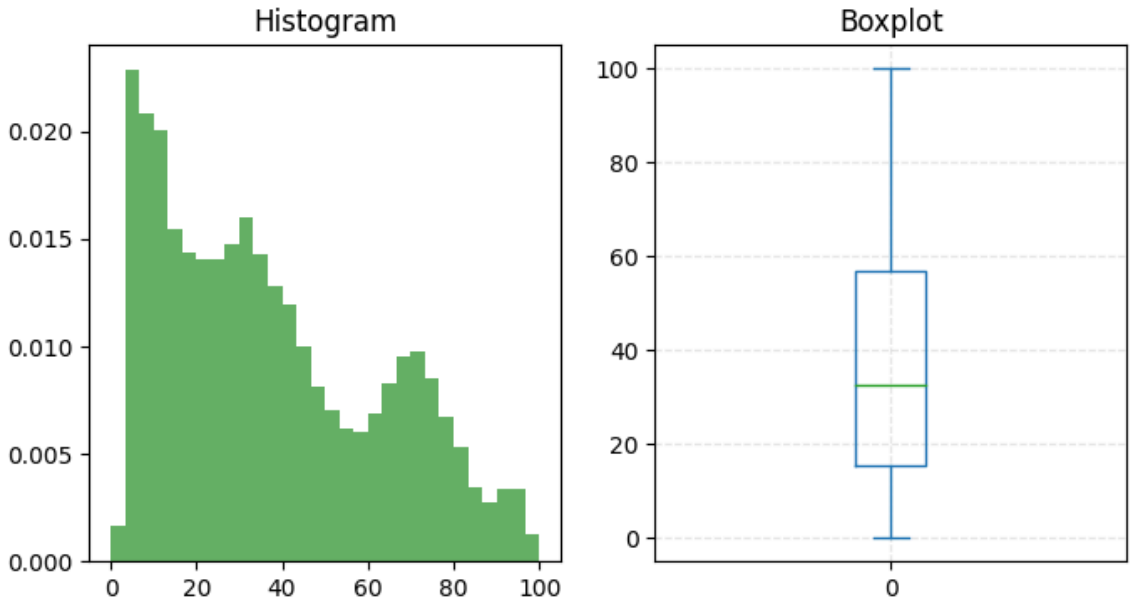



```
Fill with mode:
Five number summary: min: 0.0, Q1: 6.00, median: 15.90, Q3: 42.10, max: 100.0
```






```
Fill with the previous value:
Five number summary: min: 0.0, Q1: 15.30, median: 32.50, Q3: 56.90, max: 100.0
```



由上述数据可以看出，阿尔茨海默症的发病率大多集中在20%到60%之间，甚至由数据到达100%。该项数据存在大量缺失值，可能是数据漏采集导致，分别采用忽略缺失值（即原始的做法）、用频率最高值填充、以及使用前一个值进行补充的策略进行填充，结果如图所示。

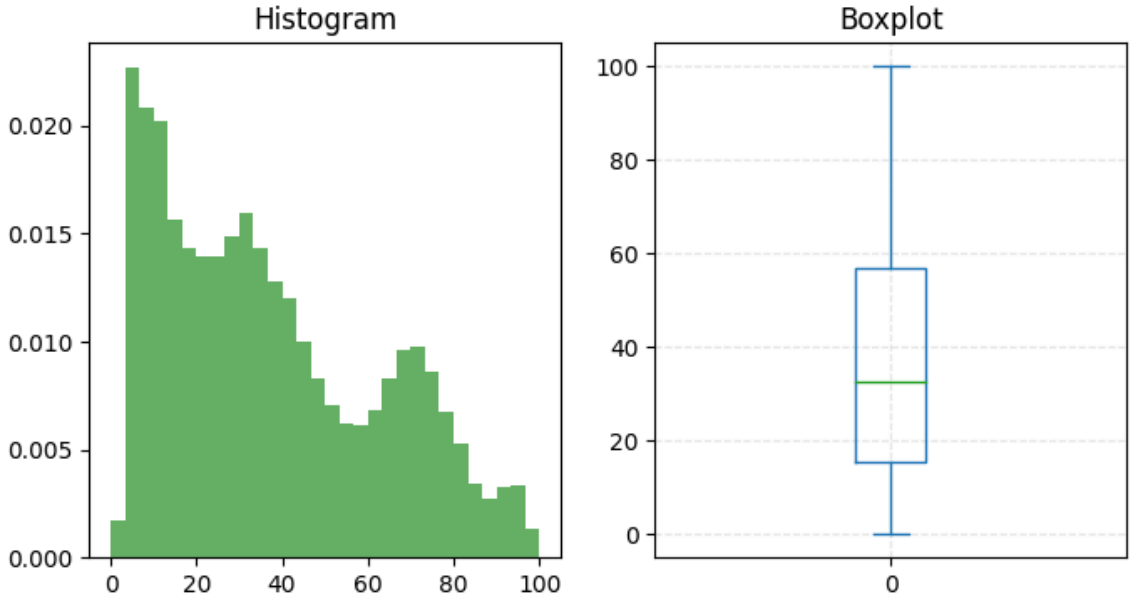
第二种策略可以看出由于缺失值数量过多，导致填充后盒图明显变扁，并且出现了离群点。

第三种策略基本保持原有数据分布不变。

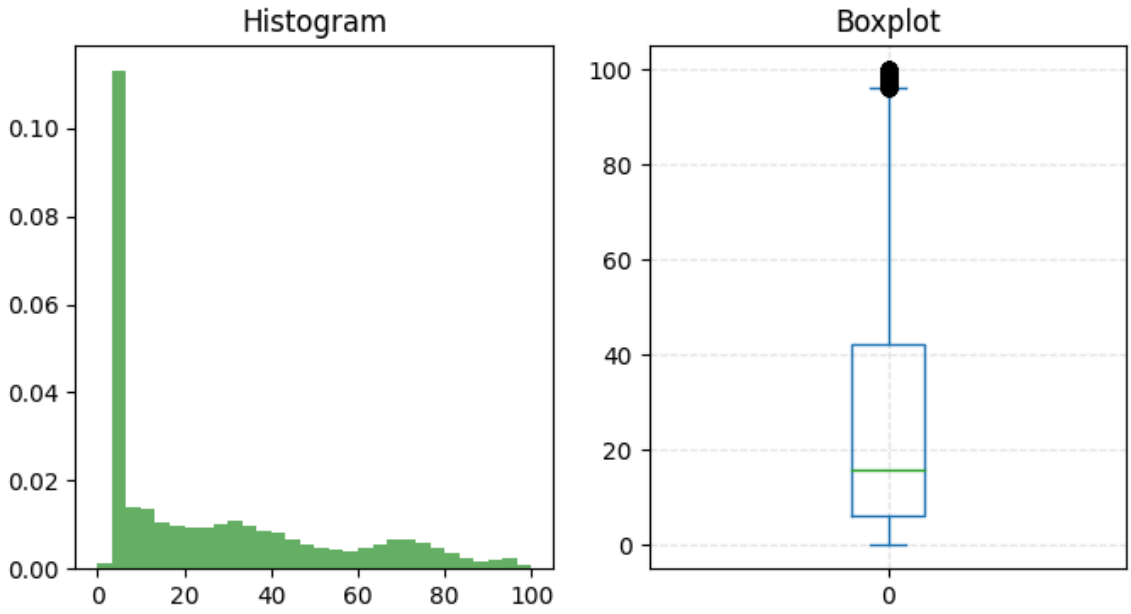
后面的几项属性均和该属性表现出一样的性状

In [12]:
```
valueAnalysis(data_df,"Data_Value_Alt")
```

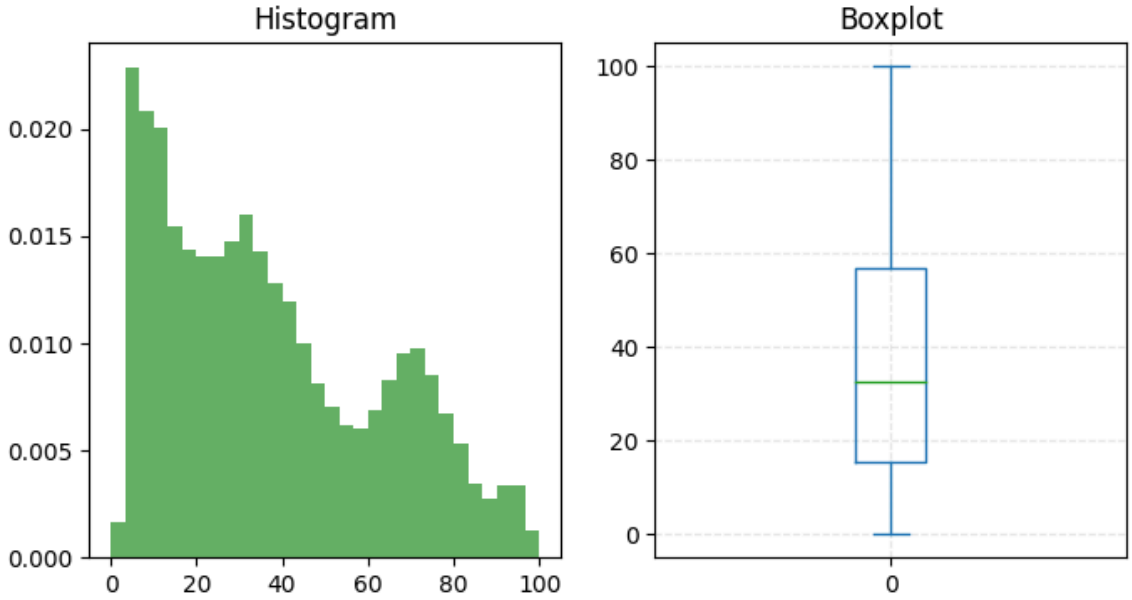
```
Data_Value_Alt: numberic attribute
Valid:144629
Missing:69833
Five number summary: min: 0.0, Q1: 15.30, median: 32.50, Q3: 56.80, max: 100.0
```




```
Fill with mode:
Five number summary: min: 0.0, Q1: 6.00, median: 15.90, Q3: 42.10, max: 100.0
```

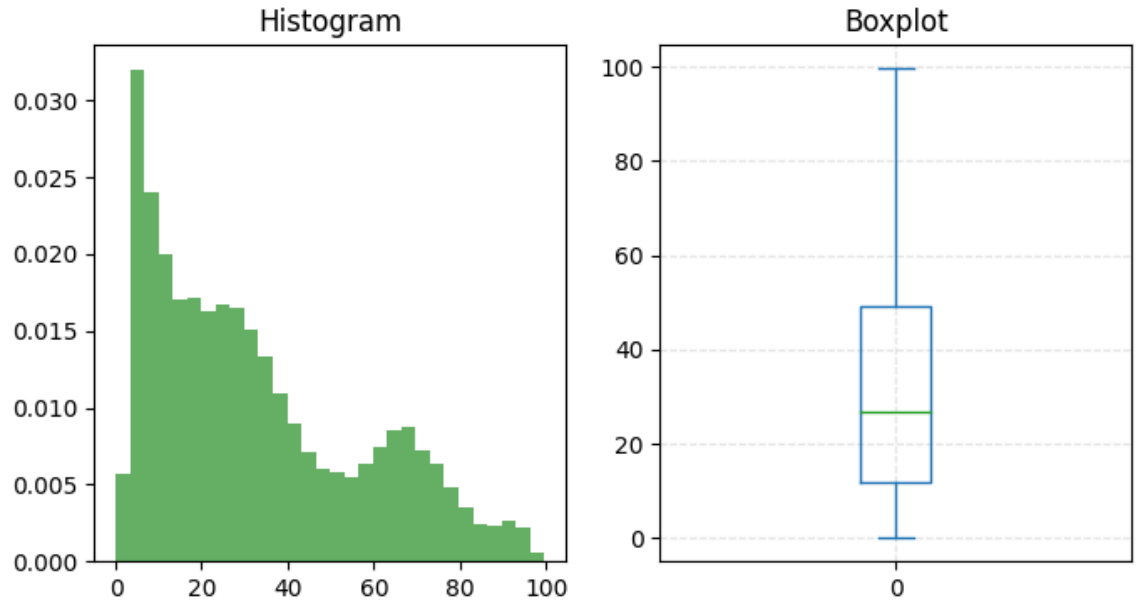



```
Fill with the previous value:
Five number summary: min: 0.0, Q1: 15.30, median: 32.50, Q3: 56.90, max: 100.0
```

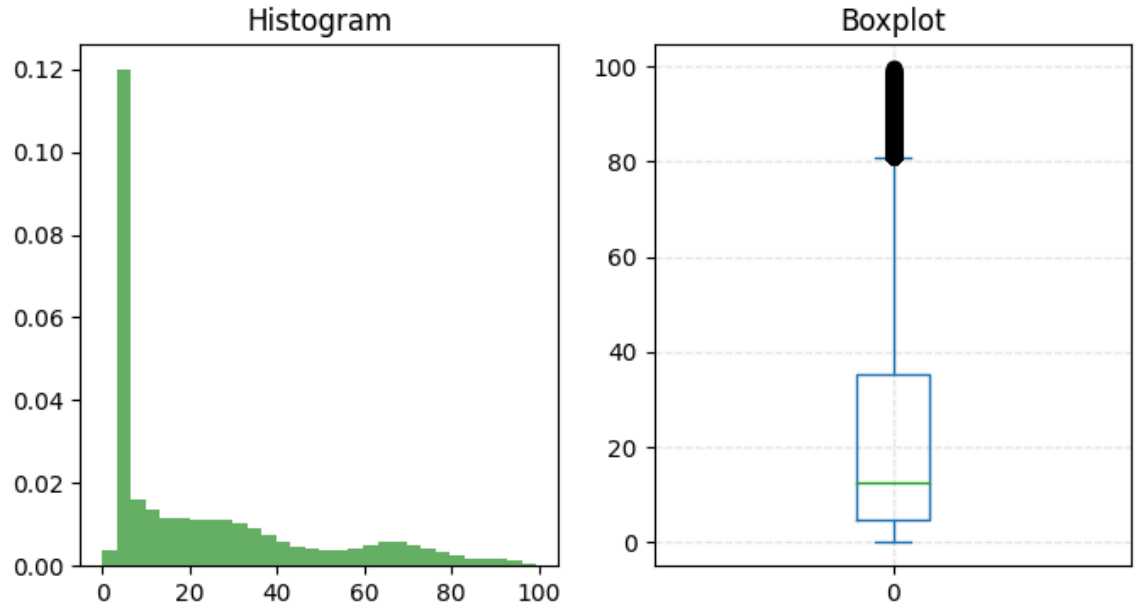


In [13]:
```
data_df['Low_Confidence_Limit'] = pd.to_numeric(data_df['Low_Confidence_Limit'],
valueAnalysis(data_df,"Low_Confidence_Limit")
```

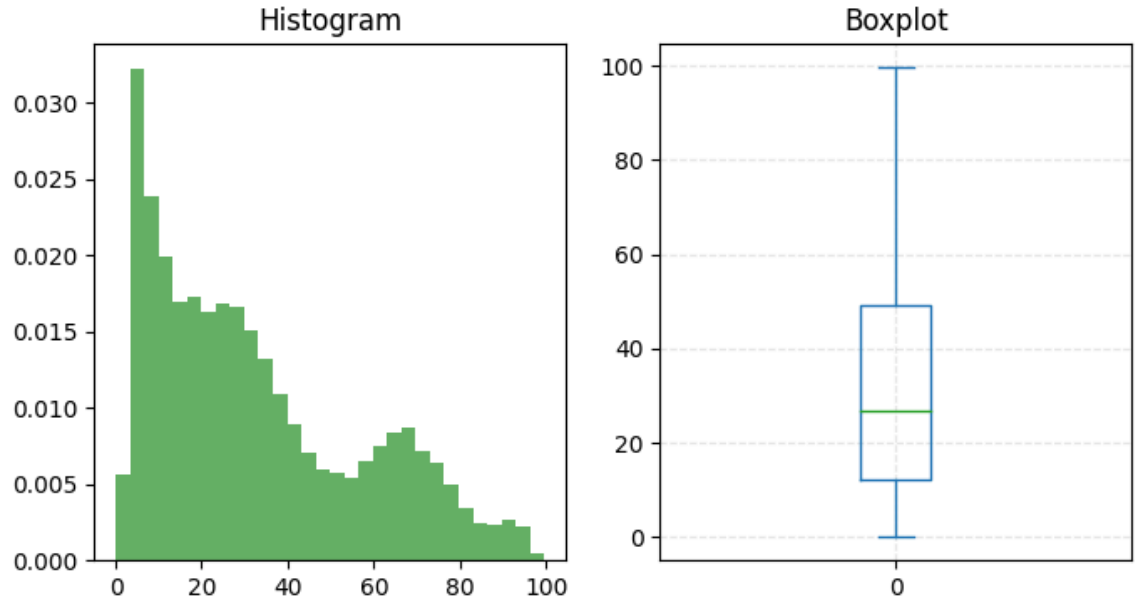
```
Low_Confidence_Limit: numberic attribute
Valid:144453
Missing:70009
Five number summary: min: 0.0, Q1: 12.00, median: 26.90, Q3: 49.10, max: 99.6
```




```
Fill with mode:
Five number summary: min: 0.0, Q1: 4.80, median: 12.50, Q3: 35.30, max: 99.6
```

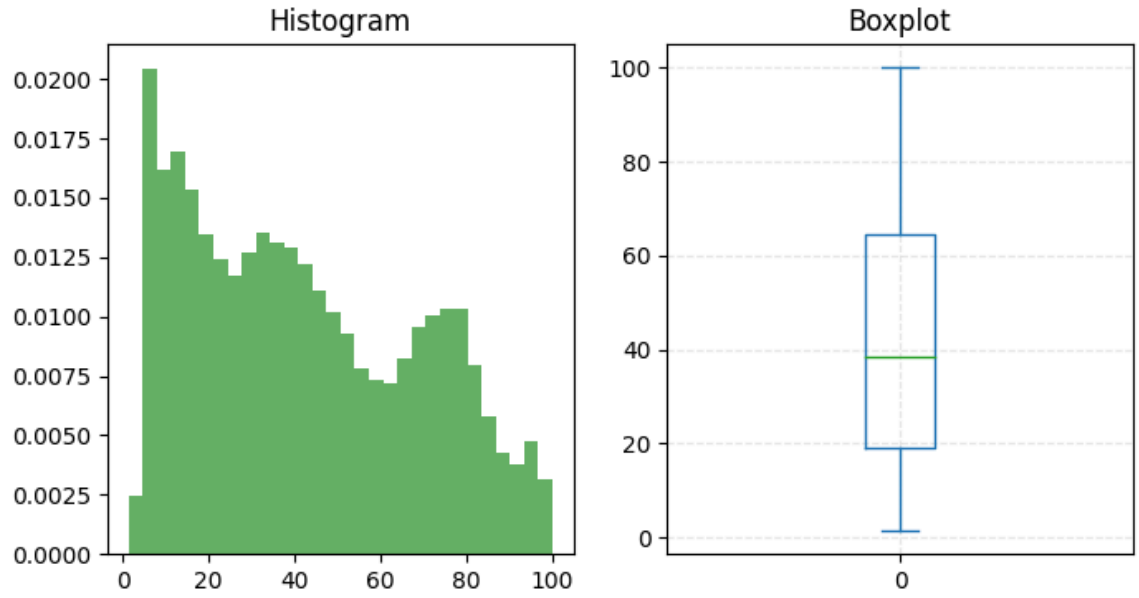



```
Fill with the previous value:
Five number summary: min: 0.0, Q1: 12.10, median: 26.80, Q3: 49.20, max: 99.6
```

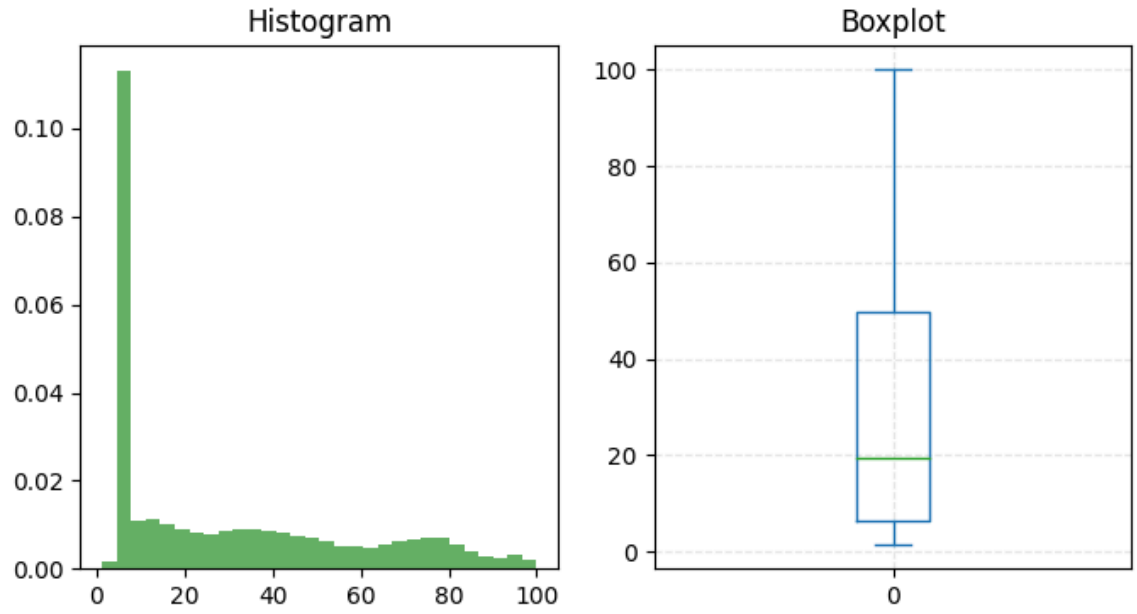


In [14]:
```
data_df['High_Confidence_Limit'] = pd.to_numeric(data_df['High_Confidence_Limit'
valueAnalysis(data_df,"High_Confidence_Limit")
```

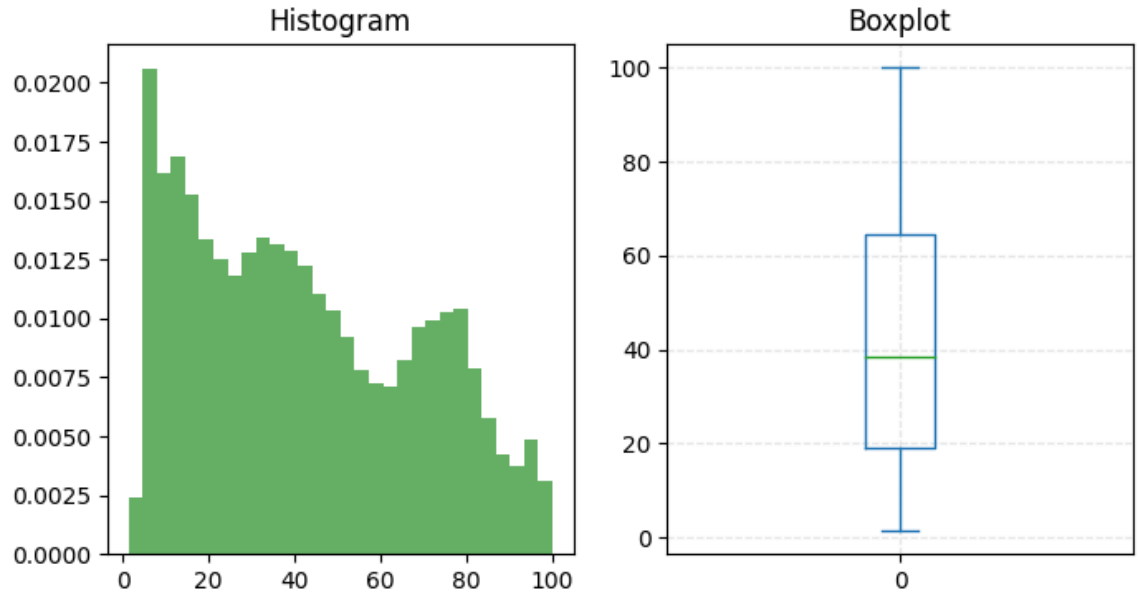
```
High_Confidence_Limit: numberic attribute
Valid:144453
Missing:70009
Five number summary: min: 1.4, Q1: 19.00, median: 38.50, Q3: 64.70, max: 100.0
```




```
Fill with mode:
Five number summary: min: 1.4, Q1: 6.50, median: 19.60, Q3: 49.60, max: 100.0
```




```
Fill with the previous value:
Five number summary: min: 1.4, Q1: 19.00, median: 38.50, Q3: 64.70, max: 100.0
```



Geolocation: 数据收集地点的经纬度

In [15]:
```
valueAnalysis(data_df,"Geolocation",20)
```

```
Geolocation: nominal attribute
Valid:191413
Missing:23049
{
    "POINT (-120.1550313 44.56744942)": 4565,
    "POINT (-75.54397043 42.82700103)": 4557,
    "POINT (-111.5871306 39.36070017)": 4222,
    "POINT (-82.40426006 40.06021014)": 3955,
    "POINT (-83.62758035 32.83968109)": 3951,
    "POINT (-76.60926011 39.29058096)": 3919,
    "POINT (-157.8577494 21.30485044)": 3907,
    "POINT (-85.77449091 35.68094058)": 3879,
    "POINT (-84.71439027 44.66131954)": 3796,
    "POINT (-78.45789046 37.54268067)": 3758,
    "POINT (-81.92896054 28.93204038)": 3753,
    "POINT (-68.98503134 45.25422889)": 3733,
    "POINT (-99.42677021 31.82724041)": 3699,
    "POINT (-117.0718406 39.49324039)": 3696,
    "POINT (-77.036871 38.907192)": 3684,
    "POINT (-80.71264013 38.6655102)": 3682,
    "POINT (-89.53803082 32.7455101)": 3677,
    "POINT (-77.86070029 40.79373015)": 3648,
    "POINT (-106.240581 34.52088095)": 3635,
    "POINT (-86.63186076 32.84057112)": 3633
}
```

数据集2：盗版网站的电影数据集

获取属性列表

In [4]:
```
data2_df,header=readDataset("./dataset/movies_dataset.csv")
data2_total_length=len(data2_df)
print(header)                                                    #属性列表
print("\ndata_total_length={}".format(data2_total_length))     #数据项总数目
```

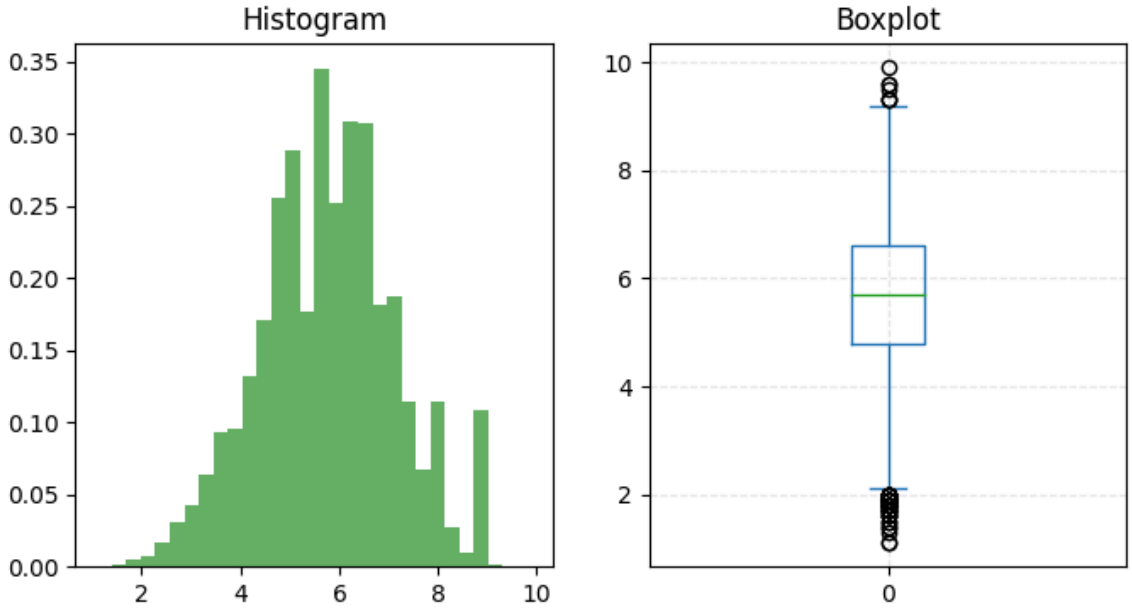
```
['Unnamed: 0', 'IMDb-rating', 'appropriate_for', 'director', 'downloads', 'id',
'industry', 'language', 'posted_date', 'release_date', 'run_time', 'storyline',
'title', 'views', 'writer']

data_total_length=20548
```

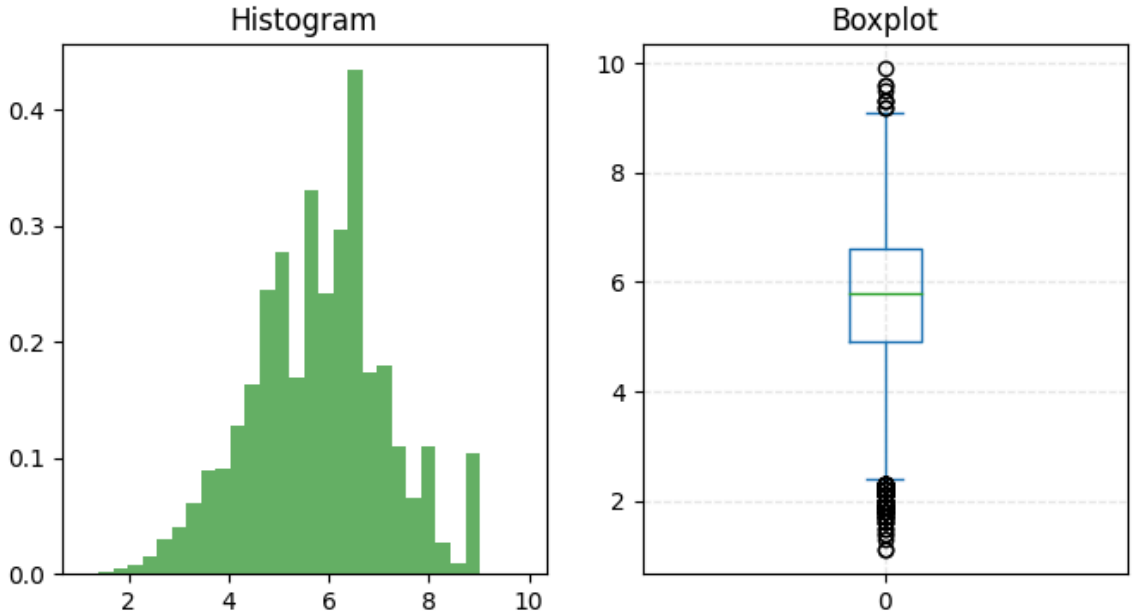
IMDB-rating: 互联网评分

In [18]:
```
valueAnalysis(data2_df,"IMDb-rating")
```

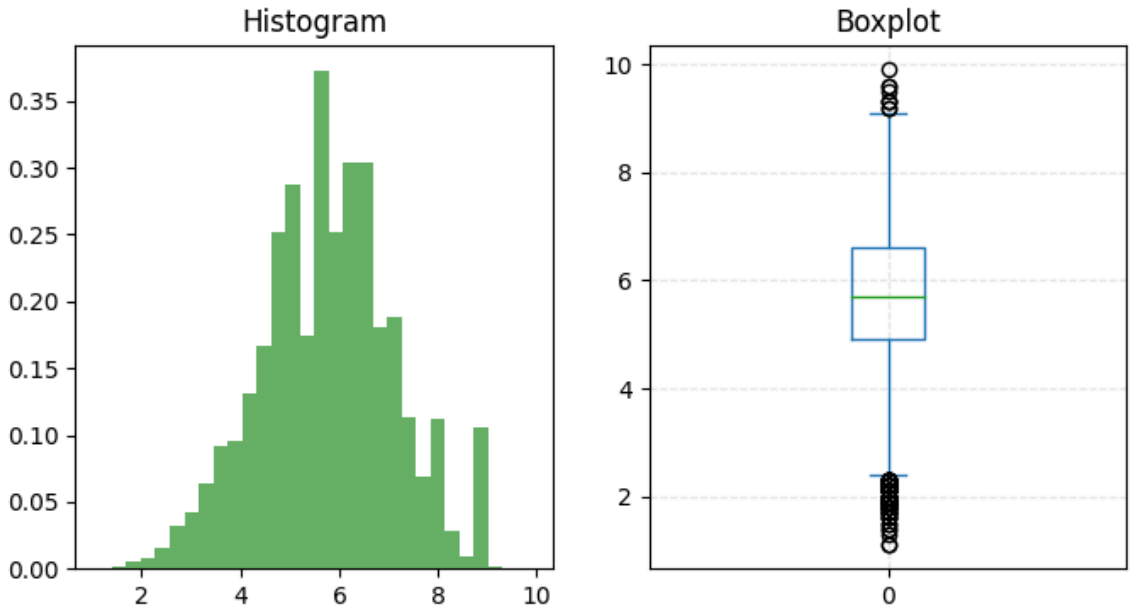
```
IMDb-rating: numberic attribute
Valid:19707
Missing:841
Five number summary: min: 1.1, Q1: 4.80, median: 5.70, Q3: 6.60, max: 9.9
```




```
Fill with mode:
Five number summary: min: 1.1, Q1: 4.90, median: 5.80, Q3: 6.60, max: 9.9
```




```
Fill with the previous value:
Five number summary: min: 1.1, Q1: 4.90, median: 5.70, Q3: 6.60, max: 9.9
```



可以看出，大众对于电影的评分大多集中在6分左右，并且数据分布和正态分布类似，盒图存在离群点。该属性存在缺失值，可能是存在电影未评分导致，采用三种策略后的数据分布大体一致。

appropriate_for: 适合人群

In [19]:
```
valueAnalysis(data2_df,"appropriate_for")
```

```
appropriate_for: nominal attribute
Valid:11072
Missing:9476
{
    "R": 4384,
    "Not Rated": 2142,
    "PG-13": 1968,
    "PG": 886,
    "TV-14": 694,
    "TV-MA": 406,
    "G": 152,
    "Unrated": 132,
    "TV-PG": 115,
    "TV-G": 99,
    "TV-Y7": 45,
    "TV-Y": 25,
    "Approved": 9,
    "NC-17": 4,
    "TV-Y7-FV": 3,
    "Passed": 3,
    "MA-17": 1,
    "TV-13": 1,
    "Drama": 1,
    "Drama, Romance": 1,
    "18+": 1
}
```

director: 导演

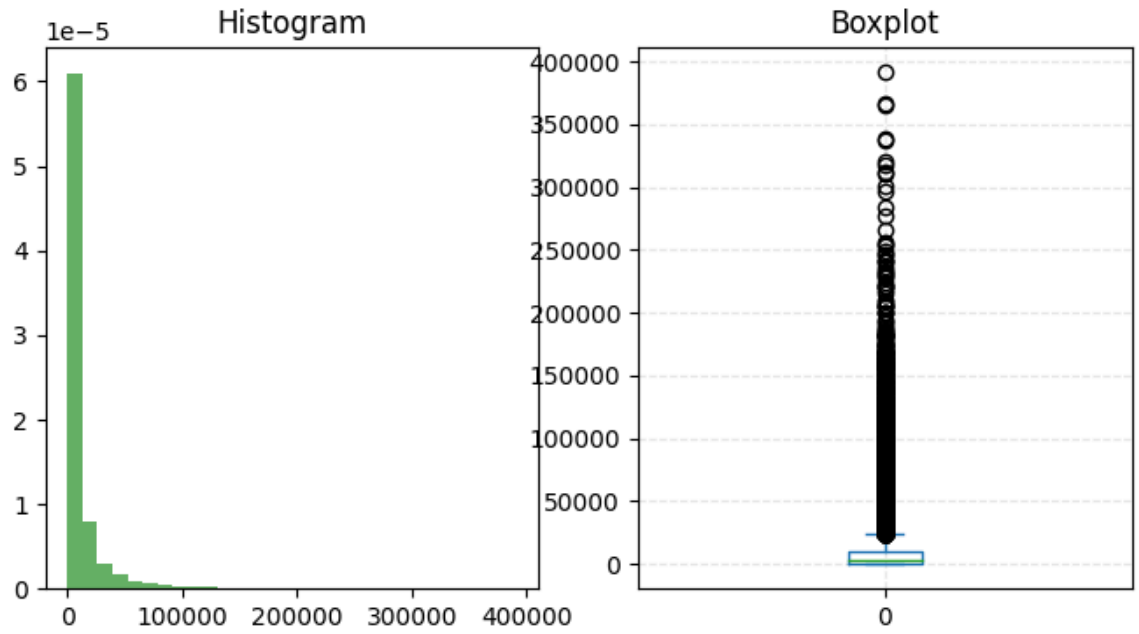
In [20]:
```
valueAnalysis(data2_df,"director",10)
```

```
director: nominal attribute
Valid:18610
Missing:1938
{
    "Venky Atluri": 405,
    "Simone Stock": 403,
    "Xavier Manrique": 403,
    "John Swab": 205,
    "Neil Jordan": 205,
    "Rohit Dhawan": 203,
    "Lindsay Hartley": 203,
    "Elegance Bratton": 202,
    "Nadira Amrani": 202,
    "Sean Lahiff": 201
}
```

downloads: 下载量

In [6]:
```
data2_df["downloads"]=pd.DataFrame(Str2Num(np.array(data2_df.loc[:,"downloads"])
valueAnalysis(data2_df,"downloads",miss=False)
```

```
downloads: numberic attribute
Valid:20547
Missing:1
Five number summary: min: 0.0, Q1: 855.50, median: 2716.00, Q3: 10070.00, max:
391272.0
```

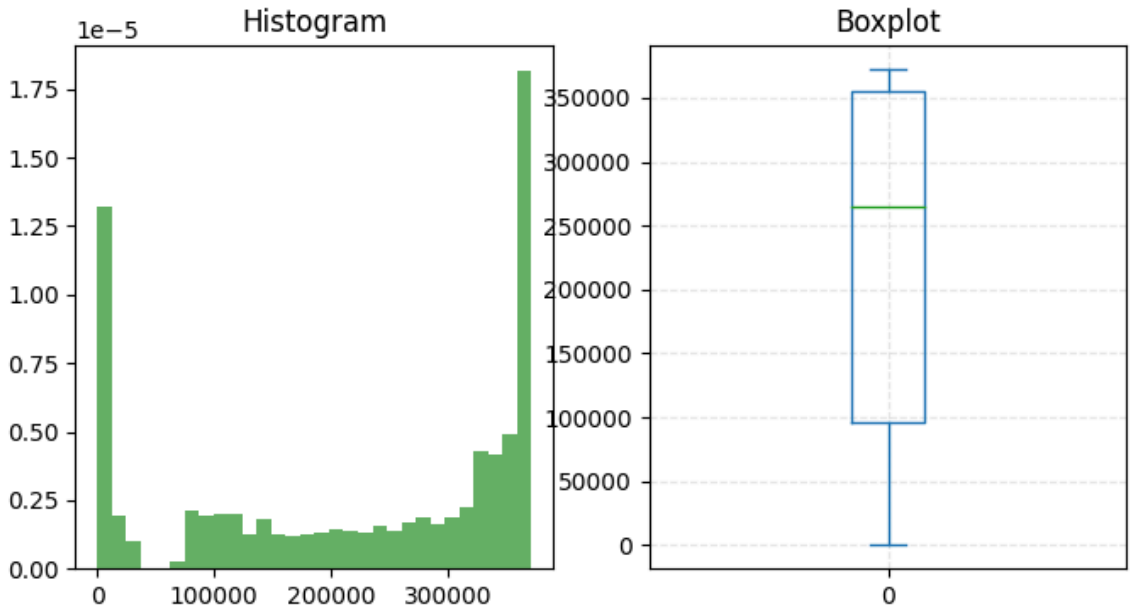


可以看出，绝大部分电影的下载量在500以下，但存在高下载量的离群点，说明该网站的电影数量极多，但下载量巨大的电影占极少数

id: 电影独特的id

In [22]:
```
valueAnalysis(data2_df,"id")
```

```
id: numberic attribute
Valid:20548
Missing:0
Five number summary: min: 1, Q1: 96122.25, median: 264457.50, Q3: 354561.25, ma
x: 372092
```



由该分布可以看出该盗版网站有大多数影片，但仍有部分影片无法从盗版网站获取

industry: 电影公司

In [23]:
```
valueAnalysis(data2_df,"industry")
```

```
industry: nominal attribute
Valid:20547
Missing:1
{
    "Hollywood / English": 14649,
    "Bollywood / Indian": 2645,
    "Tollywood": 1172,
    "Anime / Kids": 1049,
    "Wrestling": 433,
    "Punjabi": 332,
    "Stage shows": 129,
    "Pakistani": 92,
    "Dub / Dual Audio": 45,
    "3D Movies": 1
}
```

可以看出好莱坞电影占大多数，说明好莱坞电影非常受欢迎

language: 语言

In [24]:
```
valueAnalysis(data2_df,"language",20)
```

```
language: nominal attribute
Valid:20006
Missing:542
{
    "English": 12657,
    "Hindi": 2558,
    "English,Spanish": 391,
    "Punjabi": 310,
    "English,Hindi": 304,
    "Telugu": 298,
    "Tamil": 198,
    "Hindi,English": 191,
    "English,French": 174,
    "English,Russian": 71,
    "English,German": 65,
    "English,Italian": 54,
    "Urdu": 52,
    "English,Japanese": 50,
    "Malayalam": 48,
    "English,Mandarin": 48,
    "Kannada": 43,
    "English,Arabic": 38,
    "French": 37,
    "Spanish,English": 34
}
```

英语及英语和其他语言混合的影片占大多数

posted_day: 平台发行日期

In [25]:
```
valueAnalysis(data2_df,"posted_date",10)
```

```
posted_date: nominal attribute
Valid:20547
Missing:1
{
    "13 Feb, 2023": 812,
    "20 Feb, 2023": 607,
    "15 Feb, 2023": 607,
    "10 Feb, 2023": 485,
    "16 Feb, 2023": 406,
    "17 Feb, 2023": 206,
    "18 Feb, 2023": 205,
    "14 Feb, 2023": 81,
    "01 Jan, 1970": 65,
    "03 Jan, 2014": 24
}
```

release_date: 开放日期

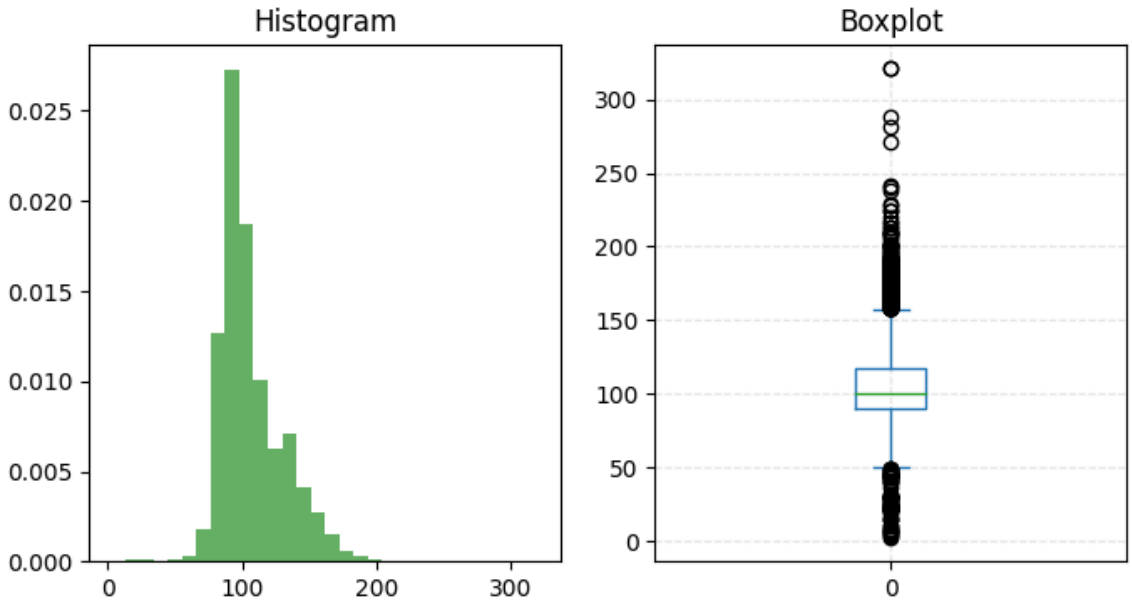
In [27]:
```
valueAnalysis(data2_df,"release_date",20)
```

```
release_date: nominal attribute
Valid:20547
Missing:1
{
    "Jan 01 1970": 962,
    "Feb 03 2023": 616,
    "Feb 17 2023": 607,
    "Feb 10 2023": 410,
    "Feb 11 2023": 402,
    "Dec 02 2022": 221,
    "Dec 01 2022": 208,
    "Jan 28 2023": 205,
    "Feb 07 2023": 204,
    "Jan 29 2023": 202,
    "Feb 15 2023": 201,
    "Feb 13 2023": 79,
    "Feb 09 2023": 79,
    "Oct 07 2022": 29,
    "Feb 14 2020": 28,
    "Dec 06 2019": 28,
    "Oct 16 2015": 27,
    "Oct 01 2021": 25,
    "Dec 10 2021": 25,
    "Mar 03 2017": 25
}
```

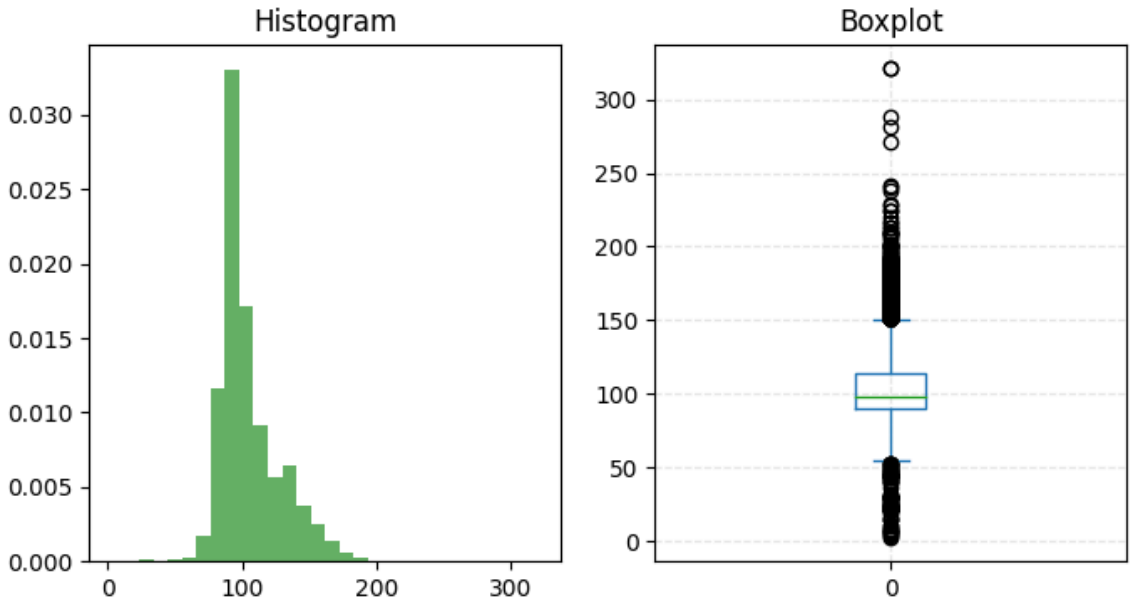
run_time: 电影时长

In [28]:
```
data2_df["run_time"]=pd.DataFrame(Time2Num(np.array(data2_df.loc[:,"run_time"])))
valueAnalysis(data2_df,"run_time")
```

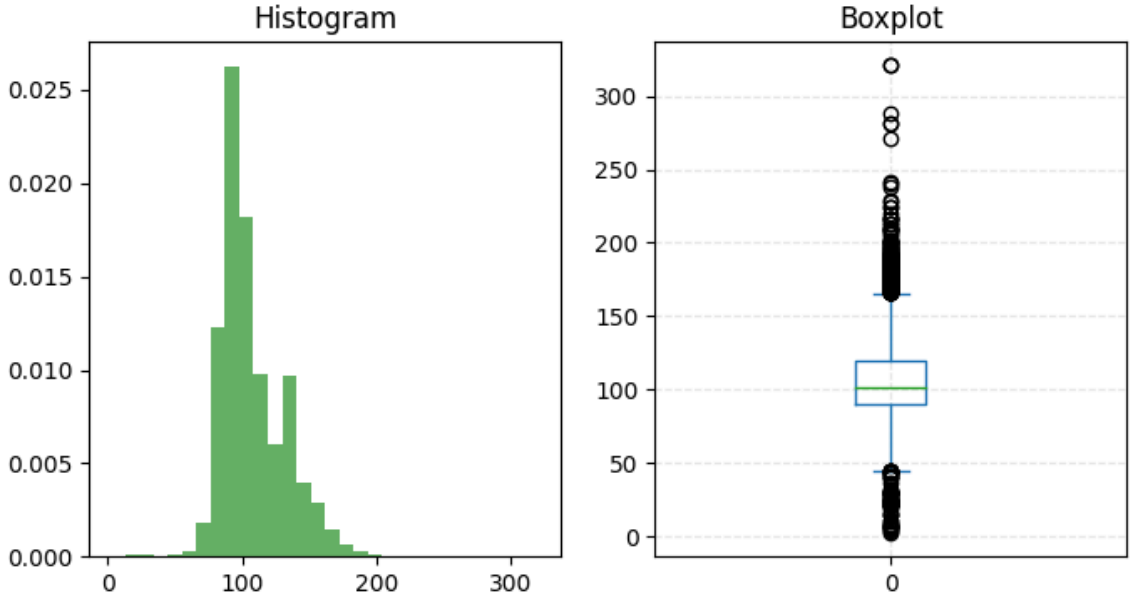
```
run_time: numberic attribute
Valid:18780
Missing:1768
Five number summary: min: 2.0, Q1: 90.00, median: 100.00, Q3: 117.00, max: 321.
0
```




```
Fill with mode:
Five number summary: min: 2.0, Q1: 90.00, median: 98.00, Q3: 114.00, max: 321.0
```




```
Fill with the previous value:
Five number summary: min: 2.0, Q1: 90.00, median: 101.00, Q3: 120.00, max: 321.
0
```



可以看出大部分电影的时长在100分钟左右

storyline: 故事线

In [29]:
```
valueAnalysis(data2_df,"storyline",5)
```

```
storyline: nominal attribute
Valid:18847
Missing:1701
{
    "The life of a young man and his struggles against the privatization of edu
cation.": 402,
    "It follows Kara Robinson as she survives an abduction and ultimately bring
s down a serial killer.": 402,
    "Follows\r\n a New York City family hiding out in the Hamptons whose bubble
is \r\npopped when a Bloody Mary-swilling, pot-smoking 'Charlie' comes to bring
\r\n a lifetime of hurt that might heal them all.": 402,
    "Doc\r\n facilitates a fragile truce between the Governor and Cartel, tradi
ng \r\nprosecutorial leniency for finance. With no more truce, Doc is left to
\r\nfend for himself and protect the one untainted thing in his life: his \r\nd
aughter, Little Dixie.": 202,
    "A\r\n young, gay Black man, rejected by his mother and with few options fo
r \r\nhis future, decides to join the Marines, doing whatever it takes to \r\ns
ucceed in a system that would cast him aside.": 202
}
```

有许多同样的故事线，说明许多电影情节和评论存在抄袭现象

title: 电影名

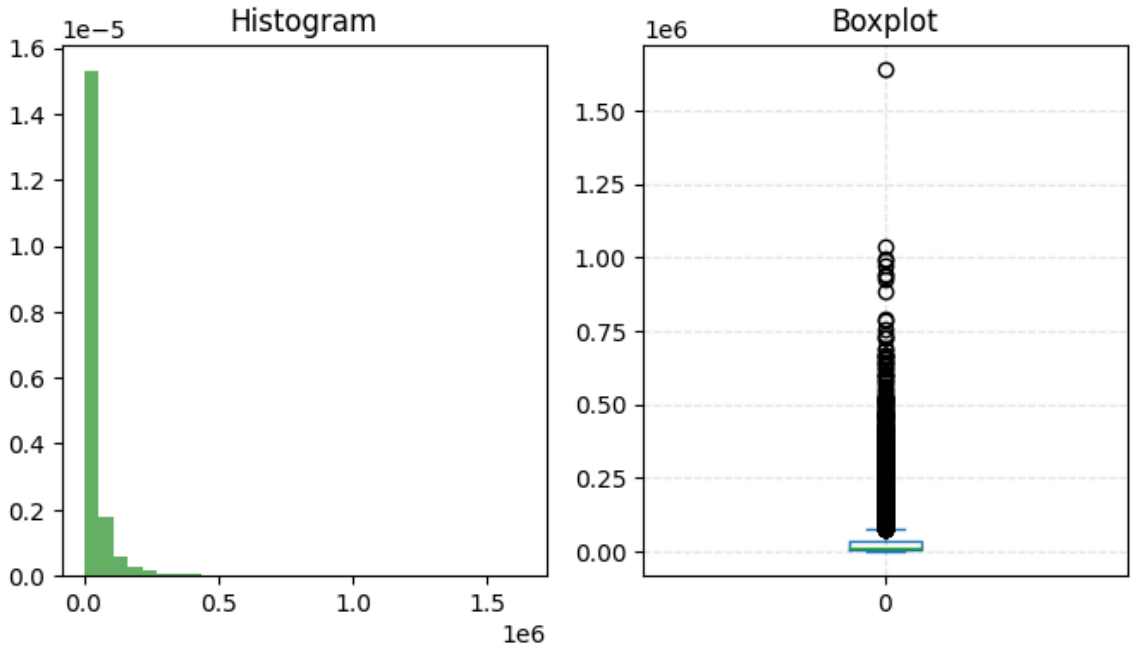
In [30]:
```
valueAnalysis(data2_df,"title",20)
```

```
title: nominal attribute
Valid:20547
Missing:1
{
    "Vaathi": 402,
    "The Girl Who Escaped: The Kara Robinson Story": 402,
    "Who Invited Charlie?": 402,
    "Little Dixie": 202,
    "The Inspection": 202,
    "Vacation Home Nightmare": 202,
    "WWE Smackdown 2023-02-10": 202,
    "Consent": 202,
    "Shehzada": 201,
    "Carnifex": 201,
    "Marlowe": 201,
    "Your Place or Mine": 201,
    "Baby Ruby": 201,
    "WWE Raw 2023-02-13": 78,
    "TNA.Impact 2023-02-09": 78,
    "Pinocchio": 6,
    "Alone": 5,
    "Sacrifice": 5,
    "True Justice": 5,
    "Blackbird": 4
}
```

views: 电影的浏览量

In [7]:
```
data2_df["views"]=pd.DataFrame(Str2Num(np.array(data2_df.loc[:,"views"]))) #预处
valueAnalysis(data2_df,"views",miss=False)
```

```
views: numberic attribute
Valid:20547
Missing:1
Five number summary: min: 667.0, Q1: 7571.50, median: 15222.00, Q3: 36571.00, m
ax: 1638533.0
```



可以看出浏览量和下载量具有相似的特征

writer: 作者

In [33]:
```
valueAnalysis(data2_df,"writer",20)
```

```
writer: nominal attribute
Valid:18356
Missing:2192
{
    "Nicholas Schutt": 403,
    "Venky Atluri": 402,
    "Haley Harris": 402,
    "John Swab": 205,
    "Elegance Bratton": 202,
    "John F. Hayes": 202,
    "Aline Brosh McKenna": 202,
    "Bess Wohl": 202,
    "Emma Dennis-Edwards": 202,
    "Hussain Dalal, Rohit Dhawan, Trivikram Srinivas": 201,
    "Shanti Gudgeon": 201,
    "William Monahan, John Banville, Raymond Chandler": 201,
    "Naresh Kathuria": 11,
    "Tyler Perry": 11,
    "Dheeraj Rattan": 11,
    "Andrew Jones": 11,
    "Justin Lee": 10,
    "Jagdeep Singh": 10,
    "Puri Jagannadh": 10,
    "Luc Besson, Robert Mark Kamen": 9
}
```

总结

本次报告给出了两个数据集的数据摘要以及可视化，基本满足任务要求。

本篇报告的缺陷是没有通过数据对象之间的相似性来填补缺失值，这在一定程度上影响了数据的完整性和分析的准确性。此问题可以通过随机森林等方法来完成。期望在后面的研究学习中能够补全。

In [ ]: