互评作业1: 数据探索性分析与数据预处理

一、github_dataset 数据集

1、数据说明

1052行数据，7个属性

```
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
from pandas import DataFrame
from tqdm import tqdm
import numpy as np
from collections import Counter
from sklearn.linear_model import LinearRegression
import scipy.stats as stats
import re
import warnings
warnings.filterwarnings('ignore')
```

```
data_1 = pd.read_csv('./github_dataset.csv')
print('github_dataset:')
print('属性类别数:', len(data_1.columns))
print('属性：', data_1.columns)
print('总行数:', len(data_1))
print('示例数据:')
data_1.head(5)
```

```
github_dataset:
属性类别数: 7
属性： Index(['repositories', 'stars_count', 'forks_count', 'issues_count',
       'pull_requests', 'contributors', 'language'],
      dtype='object')
总行数: 1052
示例数据:
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	repositories	stars_count	forks_count	issues_count	pull_requests	contributors	language
0	octocat/Hello-World	0	0	612	316	2	NaN
1	EddieHubCommunity/support	271	150	536	6	71	NaN
2	ethereum/aleth	0	0	313	27	154	C++
3	localstack/localstack	0	0	290	30	434	Python
4	education/classroom	0	589	202	22	67	Ruby

```
data_2 = pd.read_csv('./repository_data.csv')
print('repository_dataset:')
print('属性类别数:', len(data_2.columns))
print('属性：', data_2.columns)
print('总行数:', len(data_2))
print('示例数据:')
data_2.head(5)
```

```
repository_dataset:
属性类别数: 10
属性:  Index(['name', 'stars_count', 'forks_count', 'watchers', 'pull_requests',
       'primary_language', 'languages_used', 'commit_count', 'created_at',
       'licence'],
      dtype='object')
总行数: 2917951
示例数据:
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	name	stars_count	forks_count	watchers	pull_requests	primary_language	languages_used	commit_count	created_at
0	freeCodeCamp	359805	30814	8448	31867	TypeScript	['TypeScript', 'JavaScript', 'CSS', 'Shell', '...	32231.0	2014-12-24T17:49:19Z
1	996.ICU	264811	21470	4298	1949	NaN	NaN	3189.0	2019-03-26T07:31:14Z
2	free-programming-books	262380	53302	9544	8235	NaN	NaN	8286.0	2013-10-11T06:50:37Z
3	coding-interview-university	244927	65038	8539	867	NaN	NaN	2314.0	2016-06-06T02:34:12Z
4	awesome	235223	24791	7446	1859	NaN	NaN	1074.0	2014-07-11T13:42:37Z

2、数据摘要

```
num_fields = list(data_1.select_dtypes(include=np.number).columns.values)
nom_fields = list(data_1.select_dtypes(exclude=np.number).columns.values)
print('标称属性:', nom_fields)
print('数值属性:', num_fields)
```

```
标称属性: ['repositories', 'language']
数值属性: ['stars_count', 'forks_count', 'issues_count', 'pull_requests', 'contributors']
```

1）标称属性

对标称属性进行频数统计

```
for field in nom_fields:
    print('频数统计:')
    print(data_1[field].value_counts())
```

```
频数统计:
repositories
kameshsampath/ansible-role-rosa-demos         2
aloisdeniel/bluff                             2
antoniaandreou/github-slideshow               2
jgthms/bulma-start                            2
artkirienko/hlds-docker-dproto                2
                                             ..
WhiteHouse/CIOmanagement                      1
0xCaso/defillama-telegram-bot                 1
ethereum/blake2b-py                           1
```

```
openfoodfacts/folksonomy_mobile_experiment    1
gamemann/All_PropHealth                       1
Name: count, Length: 972, dtype: int64
频数统计:
language
JavaScript          253
Python              155
HTML                 72
Java                 44
CSS                  37
TypeScript           37
Dart                 36
C++                  29
Jupyter Notebook     29
Ruby                 28
C                    26
Shell                25
PHP                  16
Go                   15
Rust                 10
Swift                10
C#                    8
Objective-C           8
Kotlin                7
Makefile              6
Jinja                 5
SCSS                  4
CoffeeScript          3
Perl                  3
Dockerfile            3
Solidity              3
AutoHotkey            3
Hack                  2
Pawn                  2
CodeQL                2
PowerShell            2
Assembly              2
Vim Script            2
Vue                   2
Elixir                2
Gherkin               1
QMake                 1
CMake                 1
Oz                    1
Cuda                  1
QML                   1
ActionScript          1
Roff                  1
HCL                   1
R                     1
PureBasic             1
Smarty                1
Less                  1
Svelte                1
Haskell               1
SourcePawn            1
Name: count, dtype: int64
```

2）数值属性

```
# 五数概括
print(data_1.describe())
```

```
       stars_count  forks_count  issues_count  pull_requests  contributors
count  1052.000000  1052.000000   1052.000000    1052.000000   1052.000000
mean     81.976236    53.884981      8.656844       4.374525      8.364068
std     170.403116   127.699729     32.445154      27.913732     37.511807
min       0.000000     0.000000      1.000000       0.000000      0.000000
25%       1.000000     1.000000      1.000000       0.000000      0.000000
50%      12.000000     6.000000      2.000000       0.000000      2.000000
75%      65.250000    38.250000      6.000000       2.000000      4.000000
max     995.000000   973.000000    612.000000     567.000000    658.000000
```

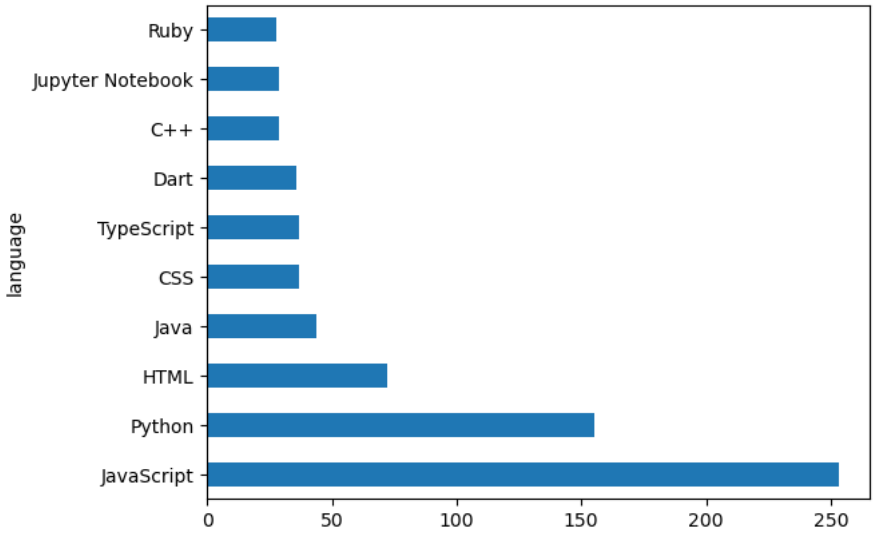
```
# 缺失值统计
for field in num_fields:
    print(field+':',data_1[field].isnull().sum())
```

```
stars_count: 0
forks_count: 0
issues_count: 0
pull_requests: 0
contributors: 0
```

3、数据可视化

1）标称属性

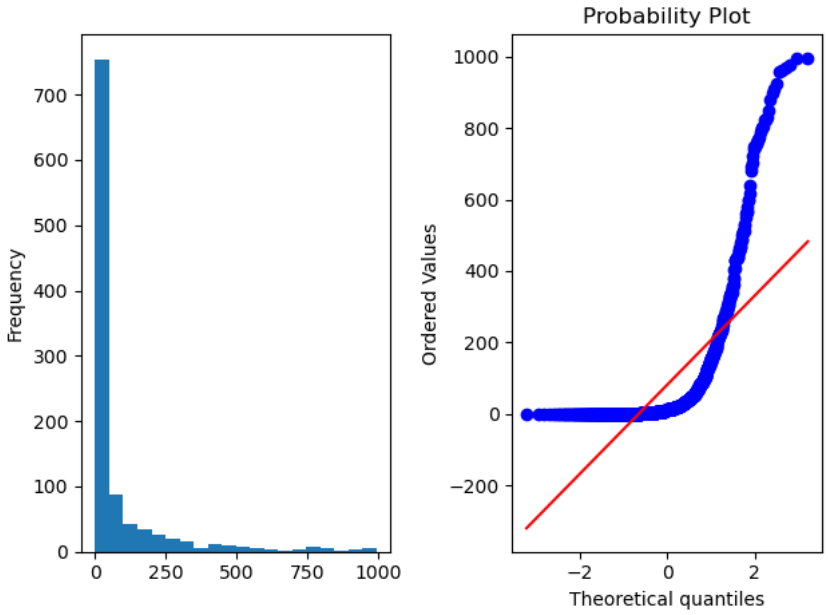
```
# 标称属性
for field in nom_fields:
    fig_path = 'fig/'+ field + '.png'
    # 全部展示纵坐标密集，只展示前10种
    data_1[field].value_counts().head(10).plot.barh().figure.savefig(fig_path)
```



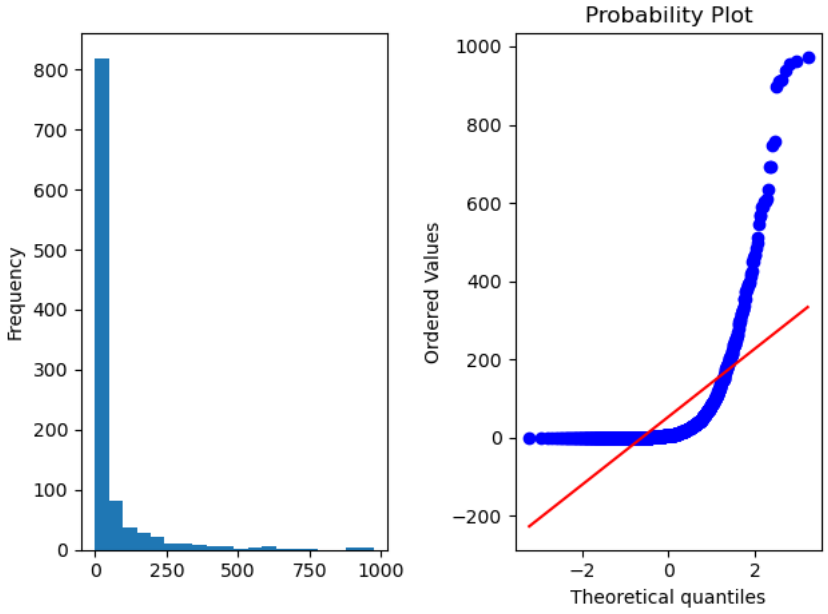
2）数值属性

```
# 数值属性
for field in num_fields:
    print(field, '直方图和Q-Q图:')
    plt.subplot(1, 2, 1)
    data_1[field].plot.hist(bins=20)
    plt.subplot(1, 2, 2)
    stats.probplot(data_1[field], plot=plt)
    plt.tight_layout()  # 调整整体空白
    plt.show()
```

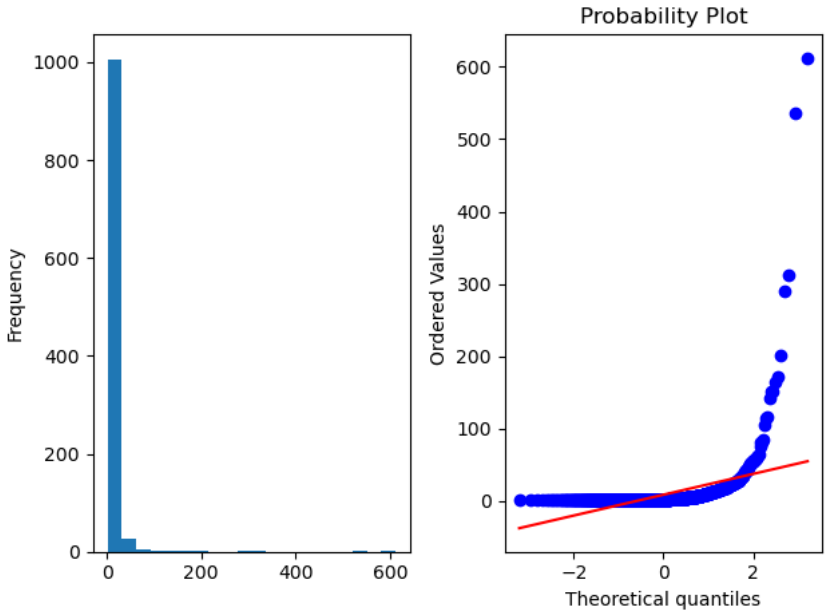
```
stars_count 直方图和Q-Q图:
```



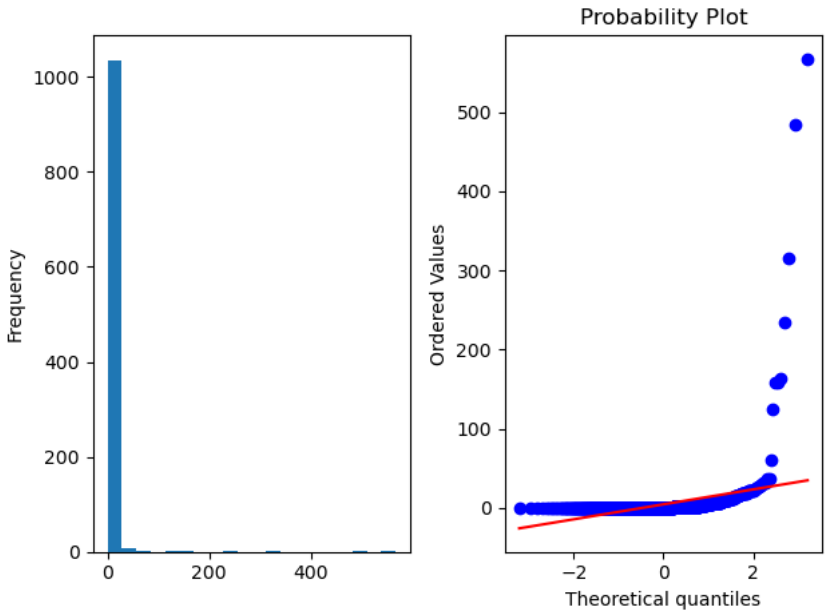
forks_count 直方图和Q-Q图:



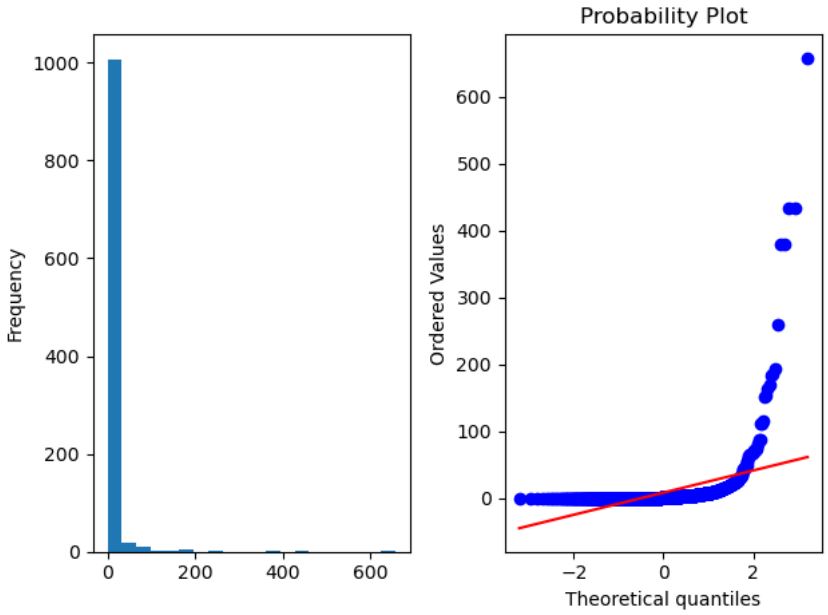
issues_count 直方图和Q-Q图:



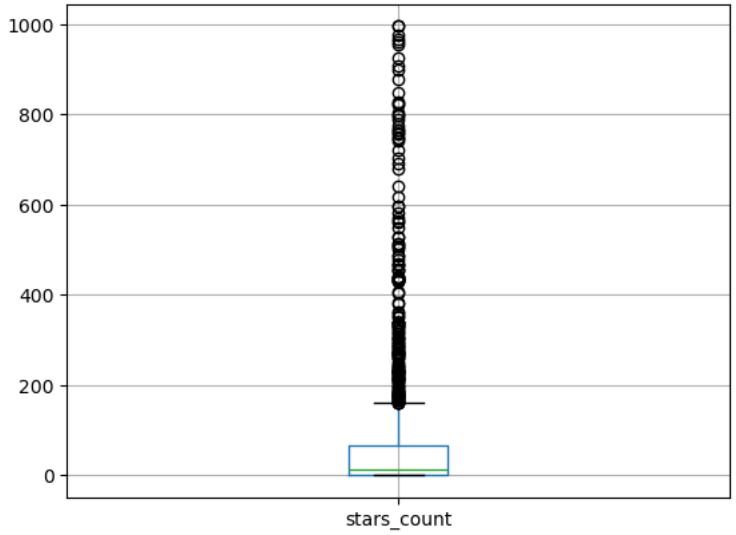
pull_requests 直方图和Q-Q图:

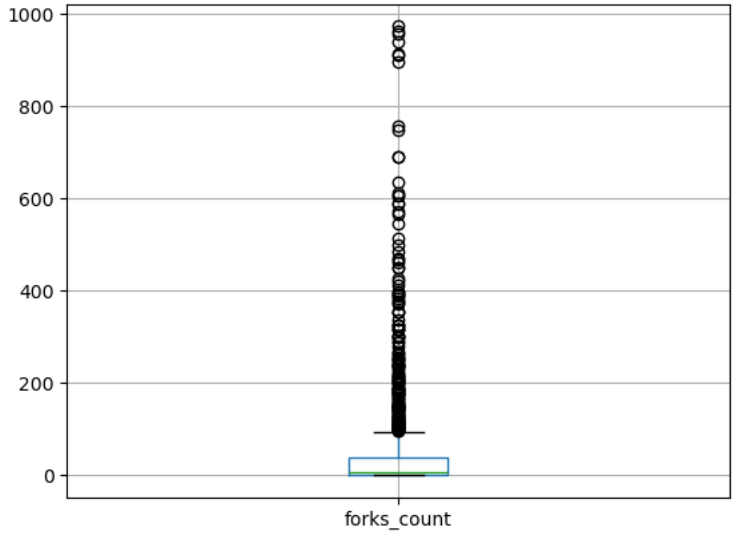


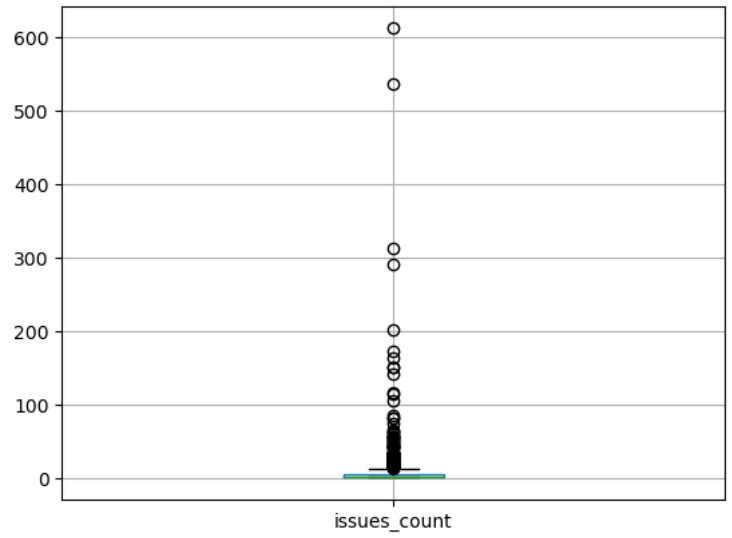
contributors 直方图和Q-Q图:

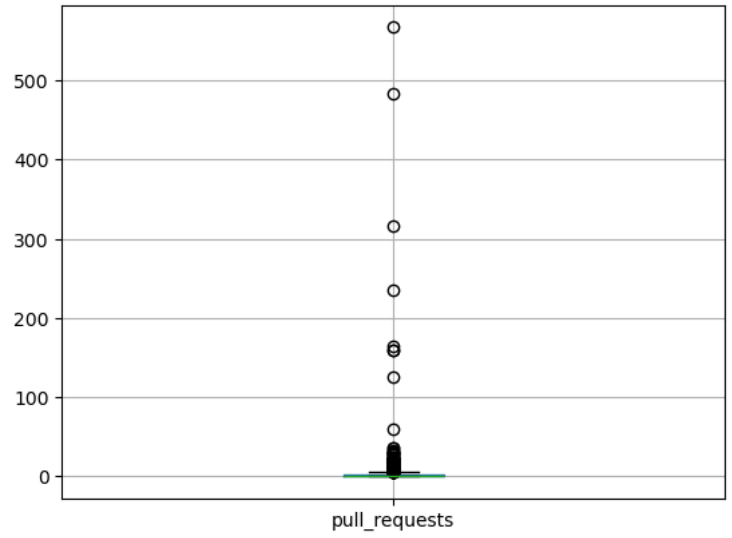


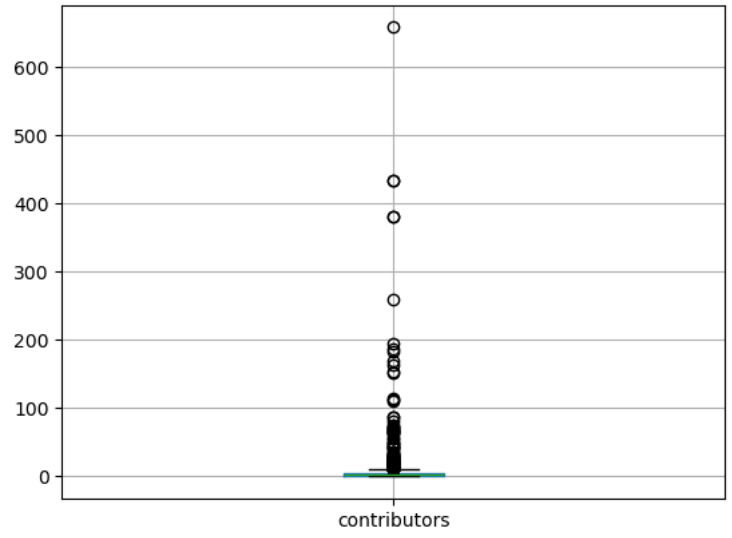

```
for field in num_fields:
    data_1.boxplot(field)
    plt.show()
```











4、缺失值处理

```
missing_data = data_1.isnull().sum()
missing_data = missing_data[missing_data != 0]
missing_data
```

```
language    145
dtype: int64
```

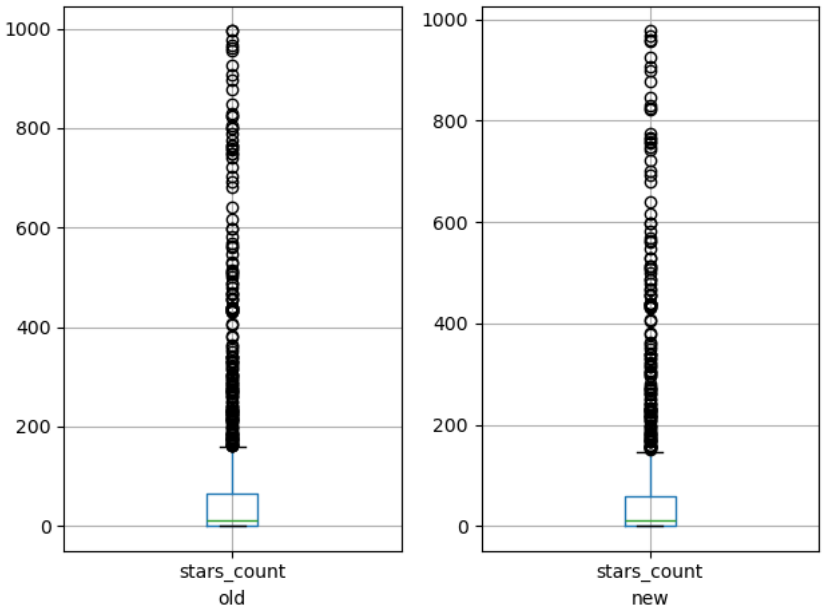
1）将缺失部分剔除

```
# 将包含缺失值的整行删除
print('原始数据行数:', len(data_1))
drop_data = data_1.dropna(how='any')
print('将缺失部分剔除后数据行数:', len(drop_data))
```

```
原始数据行数: 1052
将缺失部分剔除后数据行数: 907
```

```
print('以 stars_count 属性为例，通过盒图对比新旧数据:')
field = 'stars_count'
plt.subplot(1, 2, 1)
data_1.boxplot(field)
plt.xlabel('old')
plt.subplot(1, 2, 2)
drop_data.boxplot(field)
plt.xlabel('new')
plt.tight_layout()  # 调整整体空白
plt.show()
```

```
以 stars_count 属性为例，通过盒图对比新旧数据:
```




```
drop_data.isna().sum()
```

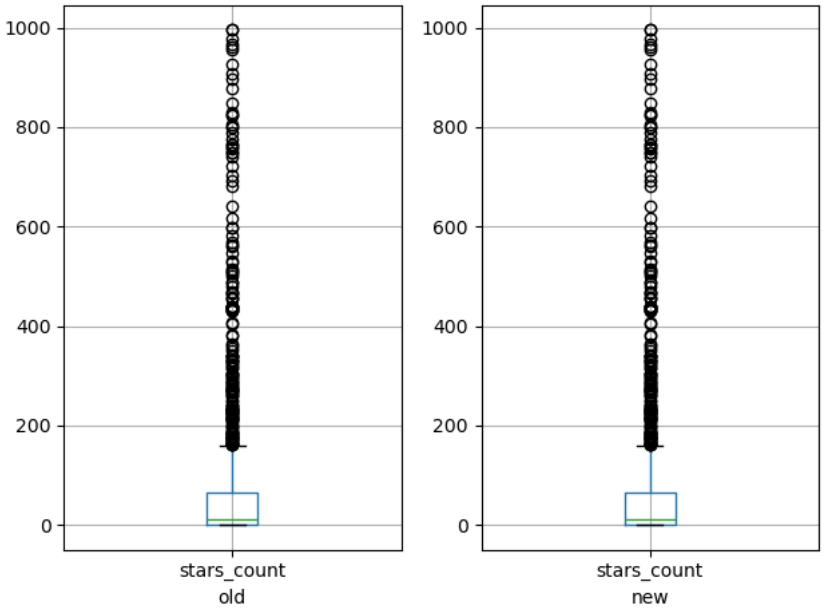
```
repositories     0
stars_count      0
forks_count      0
issues_count     0
pull_requests    0
contributors     0
language         0
dtype: int64
```

2）用最高频率来填补缺失值

```
# 用最高频率值来填补缺失值
print('以 stars_count 属性为例，通过盒图对比新旧数据:')
field = 'stars_count'
mode = data_1[field].mode()[0]
new_data = data_1.fillna({field: mode})
print(field, '属性的最高频率值为:', mode)

plt.subplot(1, 2, 1)
data_1.boxplot(field)
plt.xlabel('old')
plt.subplot(1, 2, 2)
new_data.boxplot(field)
plt.xlabel('new')
plt.tight_layout()  # 调整整体空白
plt.show()
```

```
以 stars_count 属性为例，通过盒图对比新旧数据:
stars_count 属性的最高频率值为: 0
```



github_dataset数值属性没有缺失值，不再进行处理

二、repository_data数据集

1、数据说明

```
data_2 = pd.read_csv('./repository_data.csv')
print('repository_dataset:')
print('属性类别数:', len(data_2.columns))
print('属性：', data_2.columns)
print('总行数:', len(data_2))
print('示例数据:')
data_2.head(5)
```

```
repository_dataset:
属性类别数: 10
属性：  Index(['name', 'stars_count', 'forks_count', 'watchers', 'pull_requests',
       'primary_language', 'languages_used', 'commit_count', 'created_at',
       'licence'],
      dtype='object')
总行数: 2917951
示例数据:
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	name	stars_count	forks_count	watchers	pull_requests	primary_language	languages_used	commit_count	created_at
0	freeCodeCamp	359805	30814	8448	31867	TypeScript	['TypeScript', 'JavaScript', 'CSS', 'Shell', '...	32231.0	2014-12-24T17:49:19Z
1	996.ICU	264811	21470	4298	1949	NaN	NaN	3189.0	2019-03-26T07:31:14Z
2	free-programming-books	262380	53302	9544	8235	NaN	NaN	8286.0	2013-10-11T06:50:37Z
3	coding-interview-university	244927	65038	8539	867	NaN	NaN	2314.0	2016-06-06T02:34:12Z
4	awesome	235223	24791	7446	1859	NaN	NaN	1074.0	2014-07-11T13:42:37Z

2、数据摘要

```
print('repository_dataset:')
print('属性类别数:', len(data_2.columns))
print('属性: ', data_2.columns)
print('总行数:', len(data_2))
print('示例数据:')
data_2.head(5)
num_fields = data_2.select_dtypes(include=np.number).columns.values
nom_fields = data_2.select_dtypes(exclude=np.number).columns.values
print('标称属性:', nom_fields)
print('数值属性:', num_fields)
```

```
repository_dataset:
属性类别数: 10
属性:  Index(['name', 'stars_count', 'forks_count', 'watchers', 'pull_requests',
       'primary_language', 'languages_used', 'commit_count', 'created_at',
       'licence'],
      dtype='object')
总行数: 2917951
示例数据:
标称属性: ['name' 'primary_language' 'languages_used' 'created_at' 'licence']
数值属性: ['stars_count' 'forks_count' 'watchers' 'pull_requests' 'commit_count']
```

1）标称属性

对标称属性进行频数估计

```
for field in nom_fields:
    print('频数统计:')
    print(data_2[field].value_counts())
```

```
频数统计:
name
dotfiles                                5590
blog                                    2038
docs                                    1350
website                                 1163
scripts                                  649
                                        ...
markdown-to-presentation                   1
moodle-client                              1
```

```
event-sourcing-graph                        1
react-native-100-Demos                      1
MSI-Z690-Carbon-i7-12700KF-Hackintosh       1
Name: count, Length: 2410862, dtype: int64
频数统计:
primary_language
JavaScript              451954
Python                  451473
Java                    202394
C++                     150066
PHP                     116058
                         ...
LoomScript                   1
Ragel in Ruby Host           1
Edje Data Collection         1
Sieve                        1
Ox                           1
Name: count, Length: 497, dtype: int64
频数统计:
languages_used
['Python']                                                    257679
['JavaScript']                                                157741
['Java']                                                      117624
['C#']                                                         60299
['PHP']                                                        56333
                                                                ...
['Svelte', 'TypeScript', 'JavaScript', 'HTML', 'CSS', 'Rust']      1
['Dockerfile', 'Shell', 'JavaScript', 'PowerShell']                1
['TypeScript', 'HTML', 'Vue', 'JavaScript', 'Python', 'Shell']     1
['C++', 'C', 'Pascal', 'Batchfile', 'GDB']                         1
['HTML', 'C++', 'TypeScript', 'JavaScript']                        1
Name: count, Length: 328148, dtype: int64
频数统计:
created_at
2017-06-05T20:53:54Z    10
2017-06-05T20:53:58Z     9
2014-01-17T08:00:09Z     8
2010-05-26T23:38:08Z     7
2019-03-29T08:13:35Z     7
                        ..
2017-09-04T07:45:10Z     1
2017-08-21T11:35:16Z     1
2017-08-09T00:50:43Z     1
2017-10-07T13:05:26Z     1
2022-01-22T00:00:12Z     1
Name: count, Length: 2837008, dtype: int64
频数统计:
licence
MIT License                                                 784251
Apache License 2.0                                          210698
Other                                                       167987
GNU General Public License v3.0                             159443
BSD 3-Clause "New" or "Revised" License                      47078
GNU General Public License v2.0                              43297
GNU Affero General Public License v3.0                       21554
BSD 2-Clause "Simplified" License                            16819
The Unlicense                                                14400
GNU Lesser General Public License v3.0                       14002
Mozilla Public License 2.0                                   10668
Creative Commons Zero v1.0 Universal                         10353
ISC License                                                   8232
GNU Lesser General Public License v2.1                        6168
Eclipse Public License 1.0                                    3699
Do What The F*ck You Want To Public License                   3493
Creative Commons Attribution 4.0 International                3292
Creative Commons Attribution Share Alike 4.0 International    2664
MIT No Attribution                                            2193
zlib License                                                  1512
Boost Software License 1.0                                    1421
Eclipse Public License 2.0                                    1206
BSD Zero Clause License                                        770
SIL Open Font License 1.1                                      761
Artistic License 2.0                                           685
Open Software License 3.0                                      644
Microsoft Public License                                       470
European Union Public License 1.2                              429
BSD 3-Clause Clear License                                     295
LaTeX Project Public License v1.3c                             266
BSD 4-Clause "Original" or "Old" License                       251
Universal Permissive License v1.0                              193
Academic Free License v3.0                                     143
European Union Public License 1.1                               93
University of Illinois/NCSA Open Source License                 90
PostgreSQL License                                              66
```

```
Open Data Commons Open Database License v1.0                    57
Educational Community License v2.0                              25
Mulan Permissive Software License, Version 2                    20
Vim License                                                     20
CeCILL Free Software License Agreement v2.1                     19
Microsoft Reciprocal License                                    15
CERN Open Hardware Licence Version 2 - Permissive                4
CERN Open Hardware Licence Version 2 - Strongly Reciprocal       2
CERN Open Hardware Licence Version 2 - Weakly Reciprocal         2
GNU Free Documentation License v1.3                              1
Name: count, dtype: int64
```

2）数值属性

数值属性的五数概括和缺失值个数

```
print(data_2.describe())
```

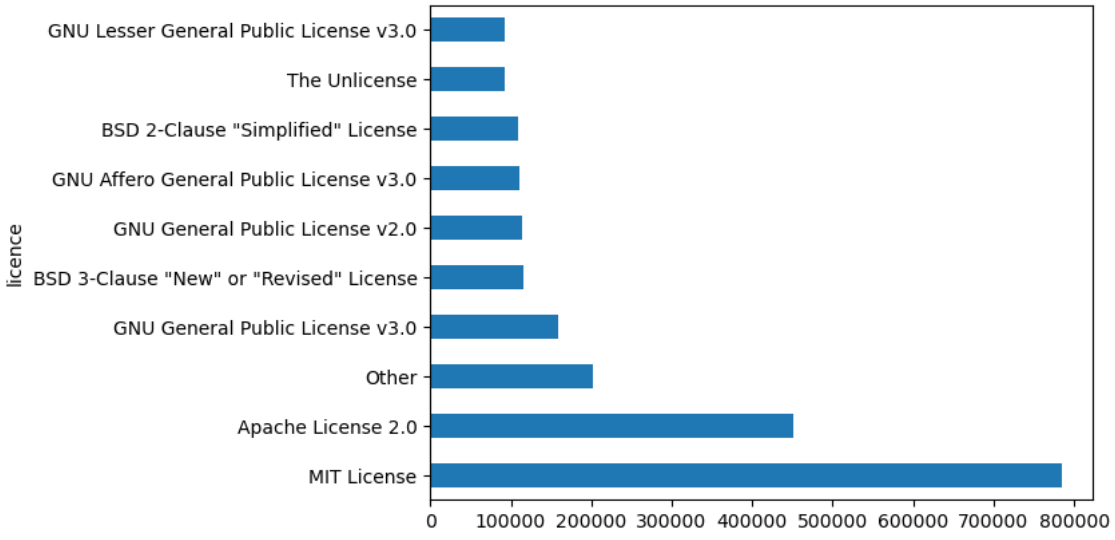
```
        stars_count   forks_count      watchers  pull_requests  commit_count
count  2.917951e+06  2.917951e+06  2.917951e+06   2.917951e+06  2.916030e+06
mean   7.641027e+01  2.094714e+01  7.135321e+00   2.430649e+01  6.143709e+02
std    9.096808e+02  3.029540e+02  3.761973e+01   3.784433e+02  1.680801e+04
min    2.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00  1.000000e+00
25%    7.000000e+00  1.000000e+00  2.000000e+00   0.000000e+00  9.000000e+00
50%    1.200000e+01  4.000000e+00  3.000000e+00   1.000000e+00  2.700000e+01
75%    3.000000e+01  1.100000e+01  6.000000e+00   6.000000e+00  8.900000e+01
max    3.598050e+05  2.422080e+05  9.544000e+03   3.015850e+05  4.314502e+06
```

```
# 缺失值统计
for field in num_fields:
    print(field+':',data_2[field].isnull().sum())
```

```
stars_count: 0
forks_count: 0
watchers: 0
pull_requests: 0
commit_count: 1921
```

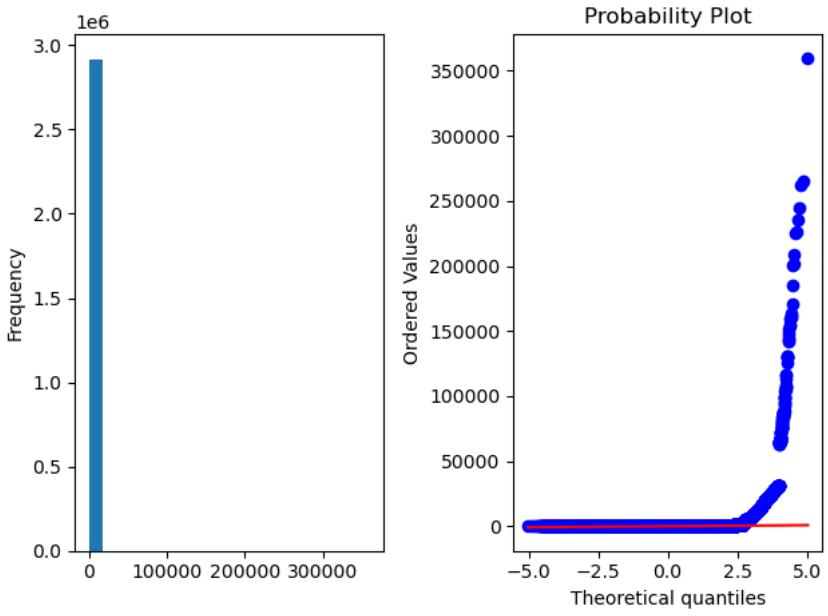
3、数据可视化

```
# 标称属性
for field in nom_fields:
    fig_path = 'fig/'+ field + '.png'
    # 全部展示纵坐标密集，只展示前10种
    data_2[field].value_counts().head(10).plot.barh().figure.savefig(fig_path)
```

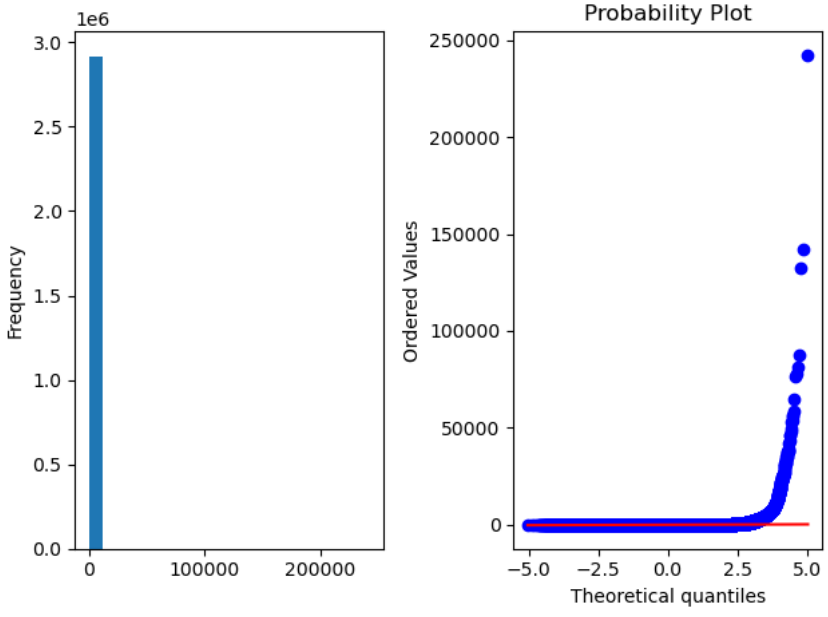


```
# 数值属性的直方图和Q-Q图
for field in num_fields:
    print(field, '直方图和Q-Q图:')
    plt.subplot(1, 2, 1)
    data_2[field].plot.hist(bins=20)
    plt.subplot(1, 2, 2)
    stats.probplot(data_2[field], plot=plt)
    plt.tight_layout()  # 调整整体空白
    plt.show()
```

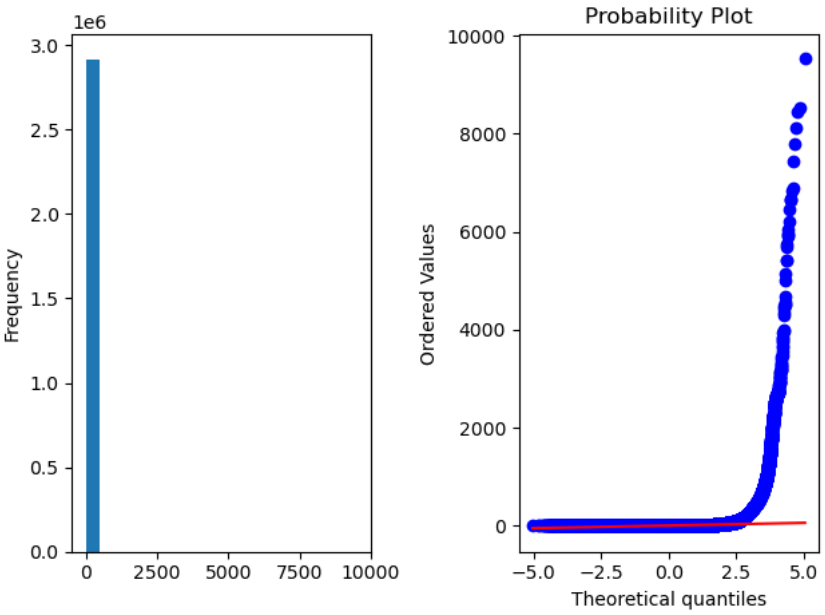
```
stars_count 直方图和Q-Q图:
```



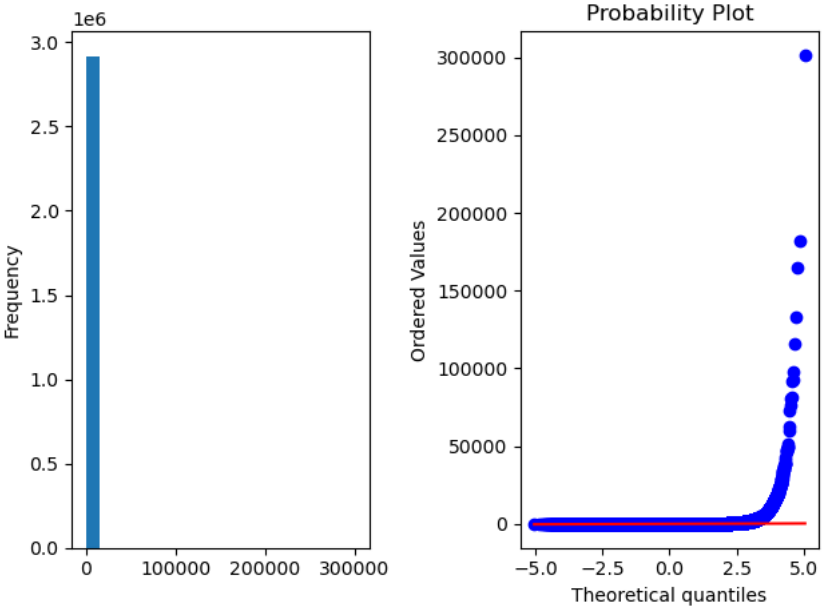

```
forks_count 直方图和Q-Q图:
```



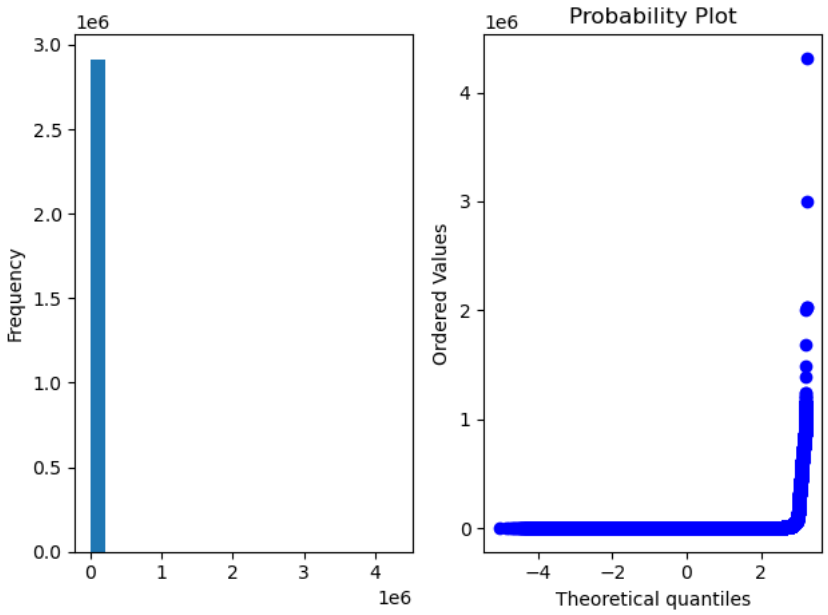

```
watchers 直方图和Q-Q图:
```



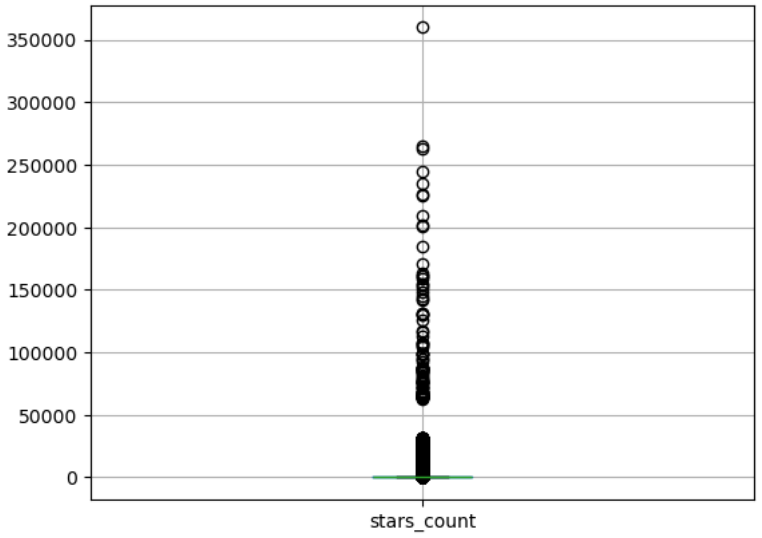
pull_requests 直方图和Q-Q图:

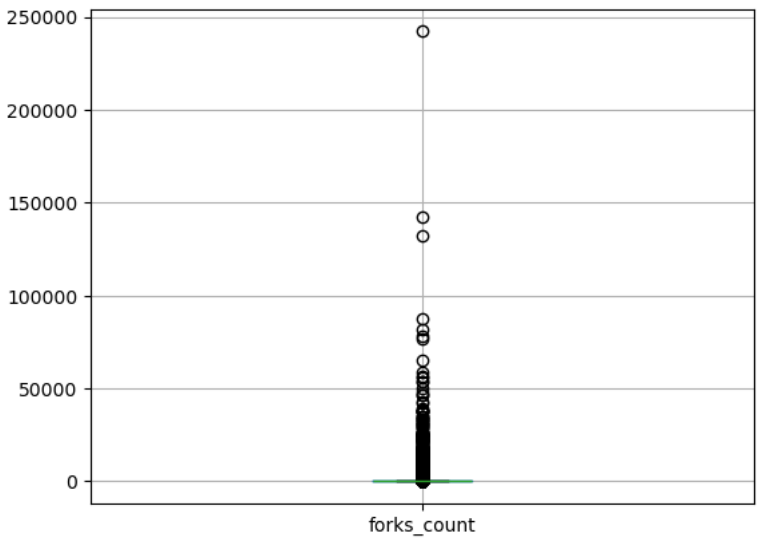


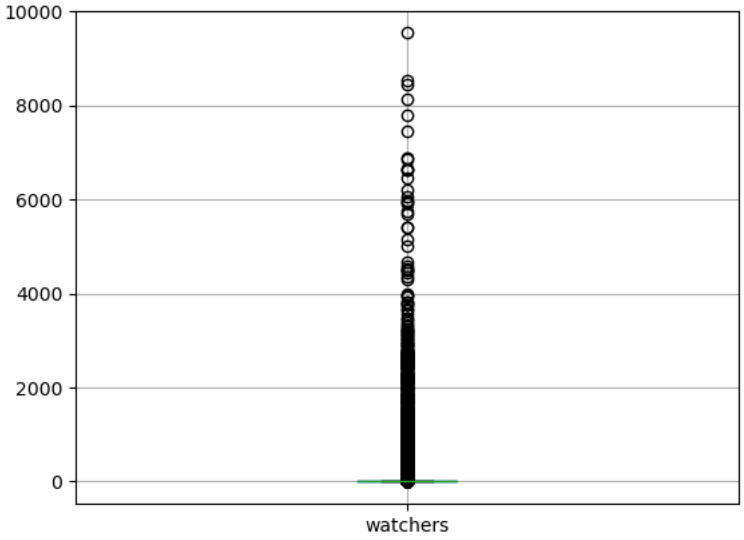
commit_count 直方图和Q-Q图:

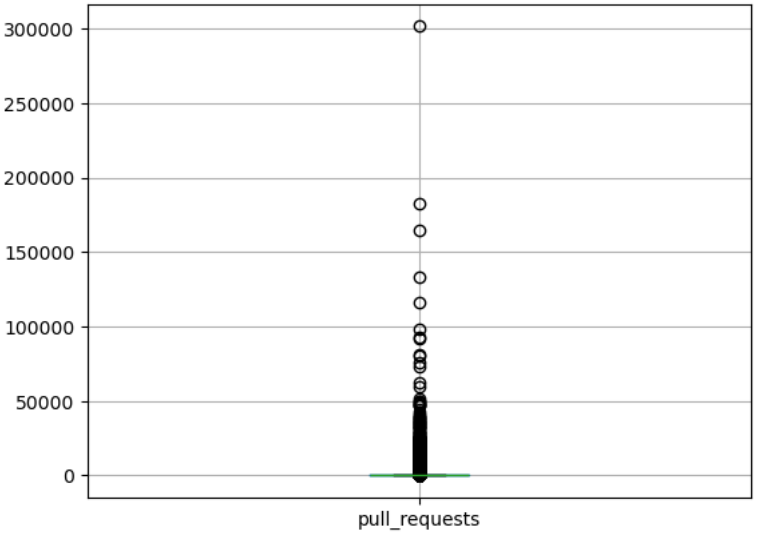


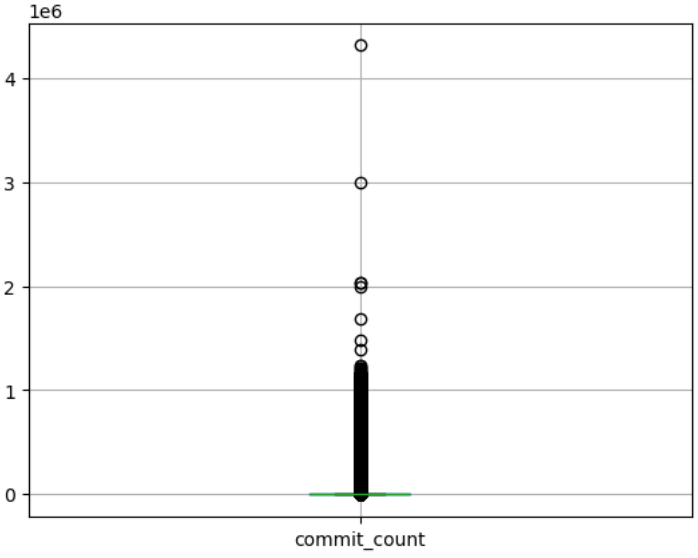

```
# 盒图
for field in num_fields:
    data_2.boxplot(field)
    plt.show()
```






10000
8000
6000
4000
2000
0
watchers


300000
250000
200000
150000
100000
50000
0
pull_requests


1e6
4
3
2
1
0
commit_count

4、缺失值处理

首先对缺失值进行统计

```
missing_data = data_2.isnull().sum()
missing_data = missing_data[missing_data != 0]
missing_data
```

```
name                   13
primary_language     218573
languages_used       221984
commit_count           1921
licence             1378200
dtype: int64
```

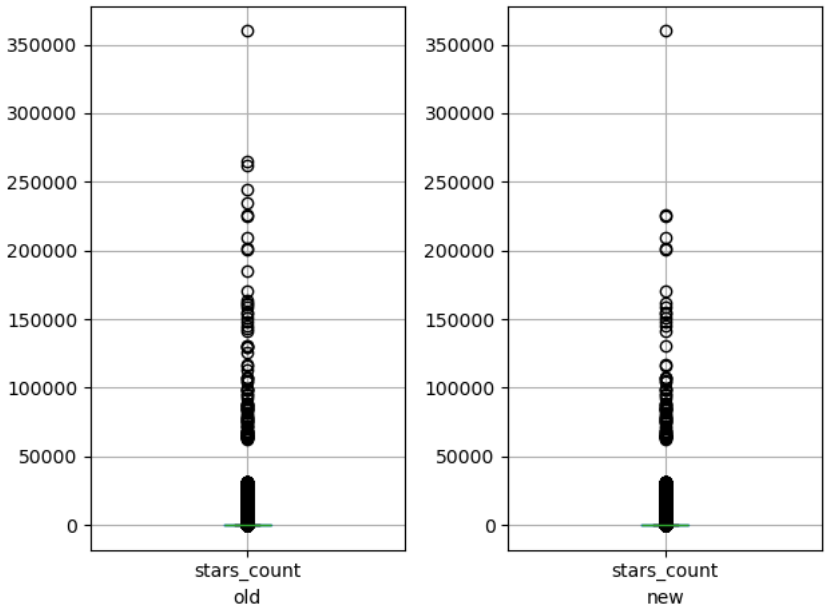
1）将缺失部分剔除

```
# 将包含缺失值的整行删除
print('原始数据行数:', len(data_2))
drop_data = data_2.dropna(how='any')
print('将缺失部分剔除后数据行数:', len(drop_data))
```

```
原始数据行数: 2917951
将缺失部分剔除后数据行数: 1471611
```

```
print('以 stars_count 属性为例，通过盒图对比新旧数据:')
field = 'stars_count'
plt.subplot(1, 2, 1)
data_2.boxplot(field)
plt.xlabel('old')
plt.subplot(1, 2, 2)
drop_data.boxplot(field)
plt.xlabel('new')
plt.tight_layout()  # 调整整体空白
plt.show()
```

```
以 stars_count 属性为例，通过盒图对比新旧数据:
```



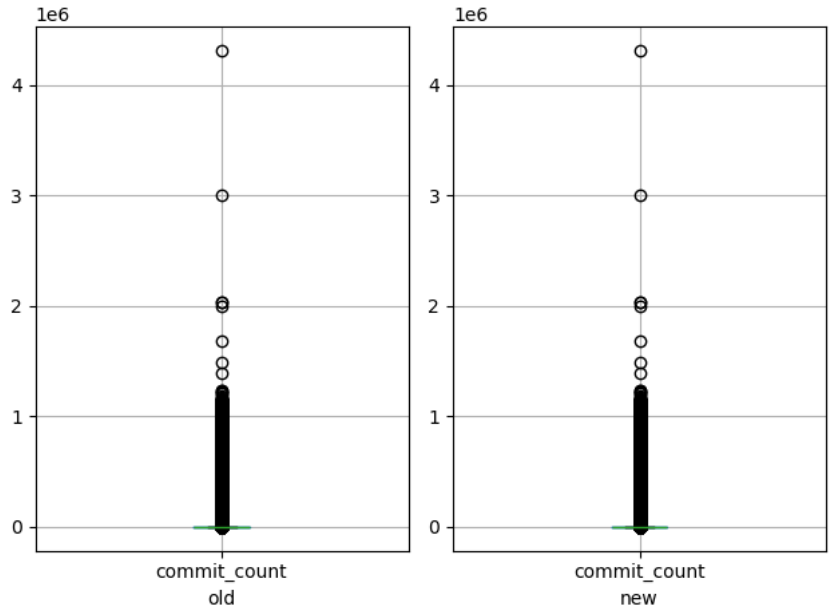

```
drop_data.isna().sum()
```

```
name                0
stars_count         0
forks_count         0
watchers            0
pull_requests       0
primary_language    0
languages_used      0
commit_count        0
created_at          0
licence             0
dtype: int64
```

2）用最高频率值来填补缺失值

```
print('以 commit_count 属性为例，通过盒图对比新旧数据:')
field = 'commit_count'
mode = data_2[field].mode()[0]
new_data = data_2.fillna({field: mode})
print(field, '属性的最高频率值为:', mode)
plt.subplot(1, 2, 1)
data_2.boxplot(field)
plt.xlabel('old')
plt.subplot(1, 2, 2)
new_data.boxplot(field)
plt.xlabel('new')
plt.tight_layout()  # 调整整体空白
plt.show()
```

```
以 commit_count 属性为例，通过盒图对比新旧数据:
commit_count 属性的最高频率值为: 2.0
```




```
data_2[data_2[field].isna()][field].head(5)
```

```
18104   NaN
31627   NaN
31774   NaN
31906   NaN
31919   NaN
Name: commit_count, dtype: float64
```

```
new_data[data_2[field].isna()][field].head(5)
```

```
18104    2.0
31627    2.0
31774    2.0
31906    2.0
31919    2.0
Name: commit_count, dtype: float64
```

3）通过属性的相关关系来填补缺失值

```
# 通过属性的相关关系来填补缺失值
data_2[num_fields].corr()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	stars_count	forks_count	watchers	pull_requests	commit_count
stars_count	1.000000	0.567460	0.706769	0.190988	0.015539
forks_count	0.567460	1.000000	0.487515	0.211495	0.018070
watchers	0.706769	0.487515	1.000000	0.161925	0.020066
pull_requests	0.190988	0.211495	0.161925	1.000000	0.046537
commit_count	0.015539	0.018070	0.020066	0.046537	1.000000

缺失属性commit_conut与其他属性的相关关系弱

4）通过数据对象之间的相似性来填补缺失值

```
# 通过数据对象之间的相似性来填补缺失值
data_2[data_2['stars_count'] <= 1000].head(10)
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	name	stars_count	forks_count	watchers	pull_requests	primary_language	languages_used	commit_count	cre
19808	uscode	768	61	44	3	Python	['Python', 'Shell']	70.0	2009 02T0
19809	Git-Tutorials	768	265	46	3	NaN	NaN	57.0	2016 21T1
19810	lerna-changelog	768	102	12	831	TypeScript	['TypeScript', 'JavaScript']	960.0	2016 24T1
19811	docker-alpine	768	178	40	12	Lua	['Lua', 'Shell', 'Dockerfile']	37.0	2019 05T1
19812	Conferences.digital	768	25	19	5	Swift	['Swift', 'Ruby', 'Objective-C', 'Shell']	61.0	2019 10T2
19813	IsoCodes	768	74	20	93	PHP	['PHP', 'Makefile', 'Shell']	471.0	2012 10T2

	name	stars_count	forks_count	watchers	pull_requests	primary_language	languages_used	commit_count	cre
19814	flutter_template	768	164	28	4	Dart	['Dart', 'Ruby', 'Kotlin', 'Swift', 'Objective...	53.0	2020 17T1
19815	Powershell-Attack-Guide	768	195	35	4	HTML	['HTML', 'PowerShell', 'CSS', 'JavaScript']	21.0	2017 10T1
19816	robo	768	45	22	23	Go	['Go', 'Makefile', 'Shell', 'JavaScript', 'Ruby']	90.0	2015 10T0
19817	blender-plugin	768	78	44	12	Python	['Python', 'Shell']	67.0	2019 25T0

```
data_2[data_2['forks_count'] <= 1000].head(10)
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	name	stars_count	forks_count	watchers	pull_requests	primary_language	languages_used	commit_count	created_at
119	CodeHub	23041	652	6670	24	C#	['C#', 'CSS', 'HTML']	593.0	2013-07-23T22:19:57Z
134	lapce	22742	665	142	841	Rust	['Rust', 'Scheme', 'Makefile', 'Shell', 'Power...	2543.0	2018-02-06T08:41:06Z
138	engineering-blogs	22708	789	920	1006	Ruby	['Ruby']	1717.0	2015-06-13T18:25:17Z
142	pkg	22661	963	272	269	JavaScript	['JavaScript', 'TypeScript', 'CSS', 'HTML', 'P...	1124.0	2016-08-08T19:41:59Z
179	tools	22235	651	194	2775	Rust	['Rust', 'JavaScript', 'TypeScript', 'Astro', ...	3984.0	2020-02-20T05:57:33Z
187	coc.nvim	22127	902	128	1101	TypeScript	['TypeScript', 'Vim Script', 'JavaScript', 'Py...	5586.0	2018-05-01T22:39:02Z
206	cascadia-code	21814	724	247	94	Python	['Python']	196.0	2019-07-10T22:50:20Z
222	github1s	21595	771	108	273	TypeScript	['TypeScript', 'JavaScript', 'HTML', 'CSS', 'S...	340.0	2019-06-16T09:55:25Z
223	croc	21593	939	244	167	Go	['Go', 'Shell', 'Makefile', 'Dockerfile']	1710.0	2017-10-17T15:20:18Z
228	pnpm	21517	638	119	2329	TypeScript	['TypeScript', 'JavaScript', 'Shell', 'Batchfi...	8011.0	2016-01-28T07:40:43Z

```
data_2[data_2['pull_requests'] <= 1000].head(10)
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	name	stars_count	forks_count	watchers	pull_requests	primary_language	languages_used	commit_count	created_at
3	coding-interview-university	244927	65038	8539	867	NaN	NaN	2314.0	2016-06-06T02:34:12Z
7	system-design-primer	208977	37414	6466	455	Python	['Python', 'Shell']	328.0	2017-02-26T16:15:28Z
10	build-your-own-x	184759	17824	4440	282	NaN	NaN	481.0	2018-05-09T12:03:18Z
12	You-Dont-Know-JS	163638	31917	5968	867	NaN	NaN	1879.0	2013-11-16T02:37:24Z
14	CS-Notes	160783	49710	5415	597	NaN	NaN	3776.0	2018-02-13T14:56:24Z
15	javascript-algorithms	159248	26396	4352	635	JavaScript	['JavaScript', 'Shell']	1076.0	2018-03-24T07:47:04Z
20	linux	144966	46406	8129	761	C	['C', 'Assembly', 'Shell', 'Makefile', 'Python...	1154596.0	2011-09-04T22:48:12Z
24	computer-science	131032	17247	5414	387	NaN	NaN	989.0	2014-05-04T00:18:39Z
26	Python-100-Days	129654	47914	6205	310	Python	['Python', 'HTML', 'Jupyter Notebook', 'Java',...	370.0	2018-03-01T16:05:52Z
27	the-art-of-command-line	126140	12804	2795	553	NaN	NaN	1208.0	2015-05-20T15:11:03Z

通过上述分析可以看出在其他属性值相同的情况下，有缺失的属性值的变动很大，说明这些缺失值无法通过其他行来进行填补。