数据探索性分析与数据预处理

1120193289 刘倞 计算机学院07111908班

github地址：https://github.com/106877952/DM_homework_week4

选用数据集GitHub Dataset

In [ ]:
```
import matplotlib.pyplot as plt
import pandas as pd
```

In [ ]:
```
dataset = pd.read_csv("github_dataset/github_dataset.csv")
```

数据属性介绍：

repositories - the name of the repository (Format - github_username/repository_name)

stars_count - stars count of the repository

forks_count - fork count of the repository

issues_count - active/opened issues in the repository

pull_requests - pull requests opened in the repository

contributors - contributors contribute to the project so far

language - primary language used in the project

In [ ]:
```
#预览前数据集前5行
dataset.head(5)
```

Out[ ]:

	repositories	stars_count	forks_count	issues_count	pull_requests	contributc
0	octocat/Hello-World	0	0	612	316	
1	EddieHubCommunity/support	271	150	536	6	
2	ethereum/aleth	0	0	313	27	1
3	localstack/localstack	0	0	290	30	4
4	education/classroom	0	589	202	22	

数据摘要和可视化

repositories 标称属性

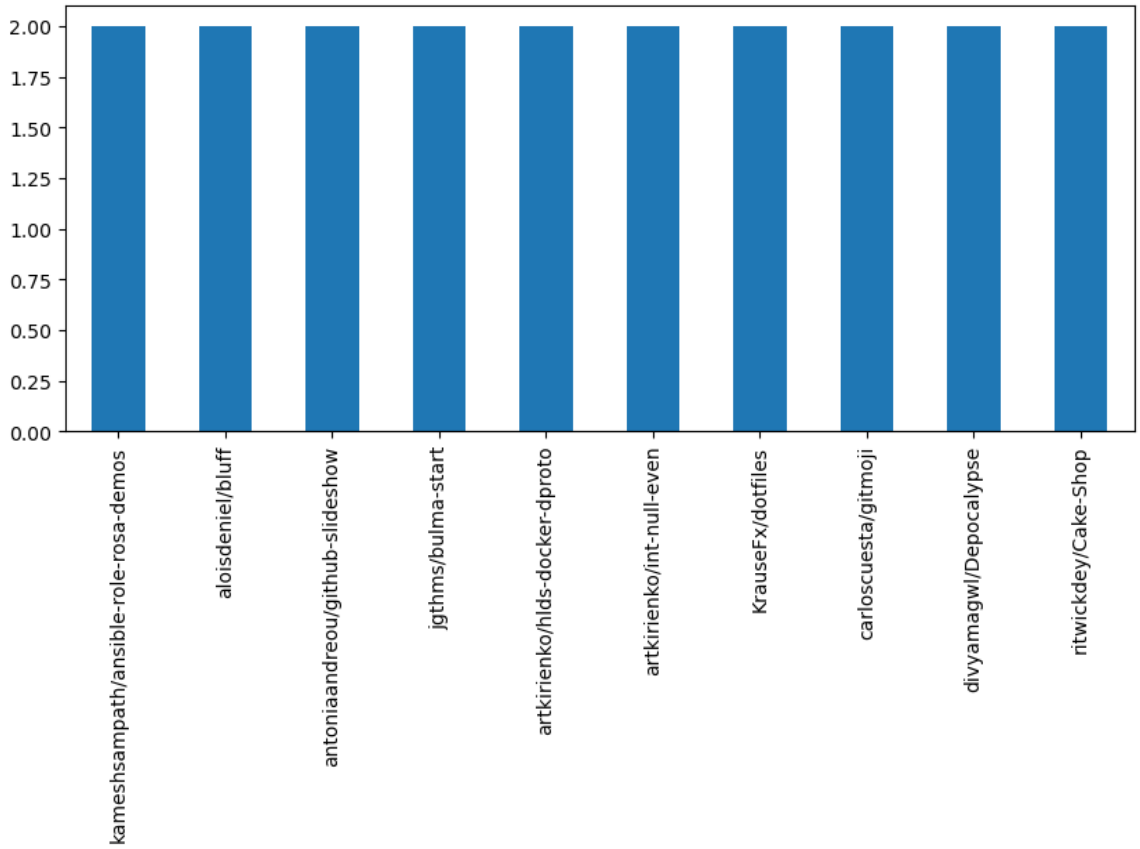
In [ ]:
```
attribute = "repositories"
dataset_repositories = dataset[attribute].value_counts(dropna=False)
```

```
dataset_repositories
```

Out[ ]:
```
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: repositories, Length: 972, dtype: int64
```

In [ ]:
```
#数据太大，仅显示前10个
dataset_repositories[:10].plot(kind="bar", figsize=(10,4))
```

Out[ ]:
```
<Axes: >
```



stars_count 数值属性

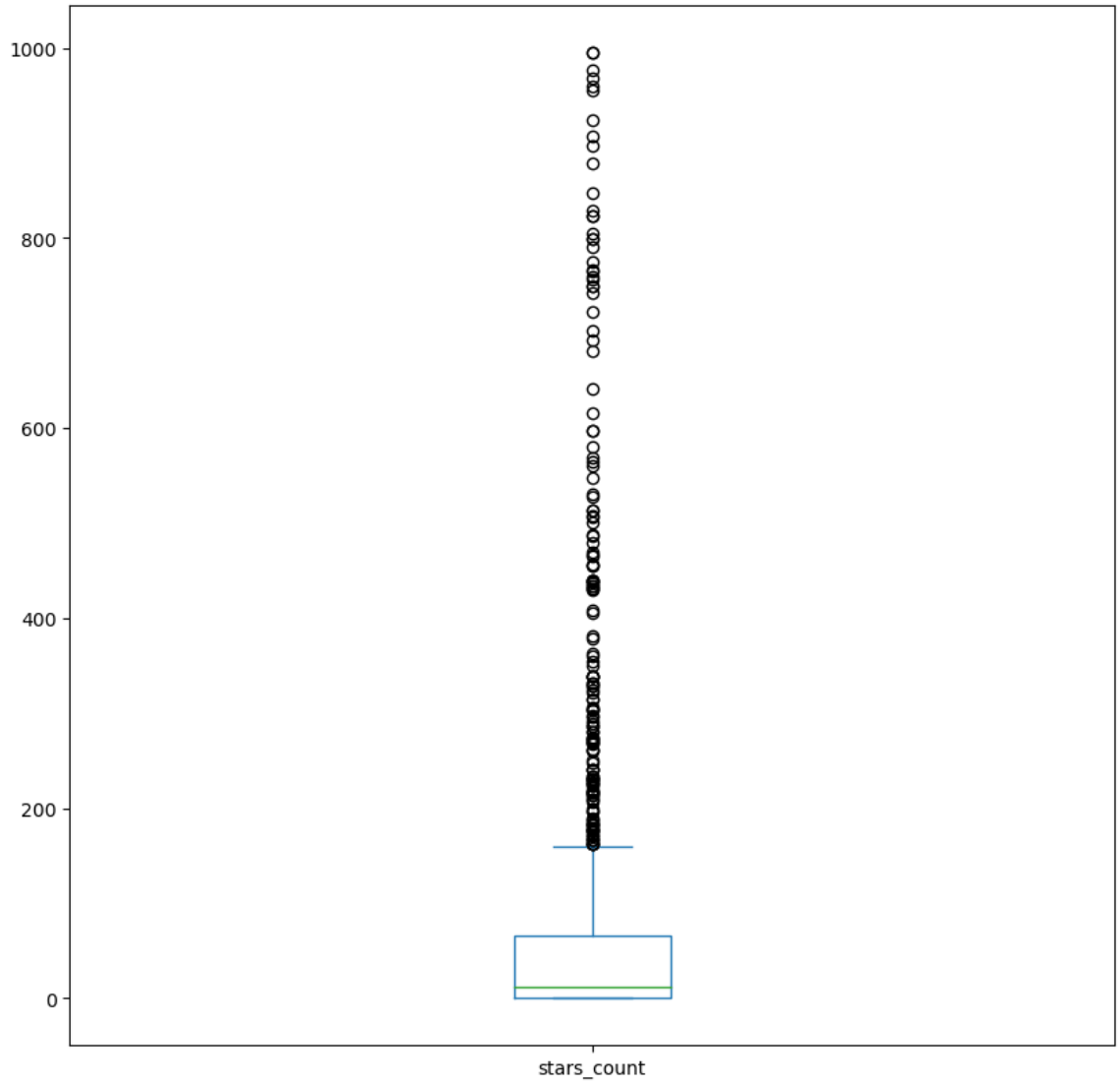
In [ ]:
```
attribute = "stars_count"

#五数概括
print('Min:', dataset[attribute].quantile(0))
print('Q1:', dataset[attribute].quantile(0.25))
print('Q2:', dataset[attribute].quantile(0.5))
print('Q3:', dataset[attribute].quantile(0.75))
print('Max:', dataset[attribute].quantile(1))
```

```
Min: 0.0
Q1: 1.0
Q2: 12.0
Q3: 65.25
Max: 995.0
```

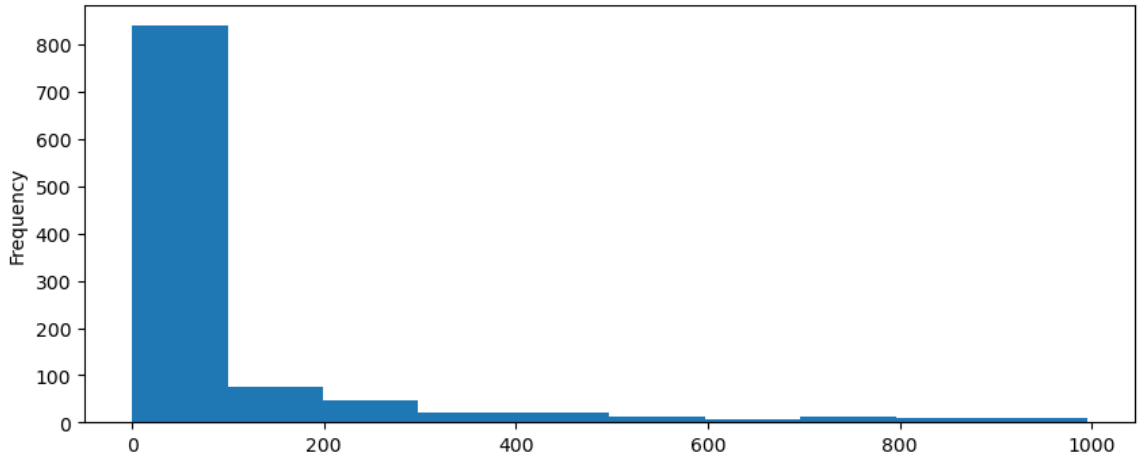
In [ ]:
```
#绘制盒图
dataset[attribute].plot(kind="box",figsize=(10,10))
```

Out[ ]:
```
<Axes: >
```



In [ ]:
```
#绘制直方图
dataset[attribute].plot(kind="hist",figsize=(10,4))
```

Out[ ]:
```
<Axes: ylabel='Frequency'>
```



In [ ]:
```
#查找离群点
Q1 = dataset[attribute].quantile(0.25)
Q3 = dataset[attribute].quantile(0.75)
outliner = Q3 + (Q3 - Q1) * 1.5
print(f"大于{outliner}的项被识别为离群点")
```

大于161.625的项被识别为离群点

forks_count 数值属性

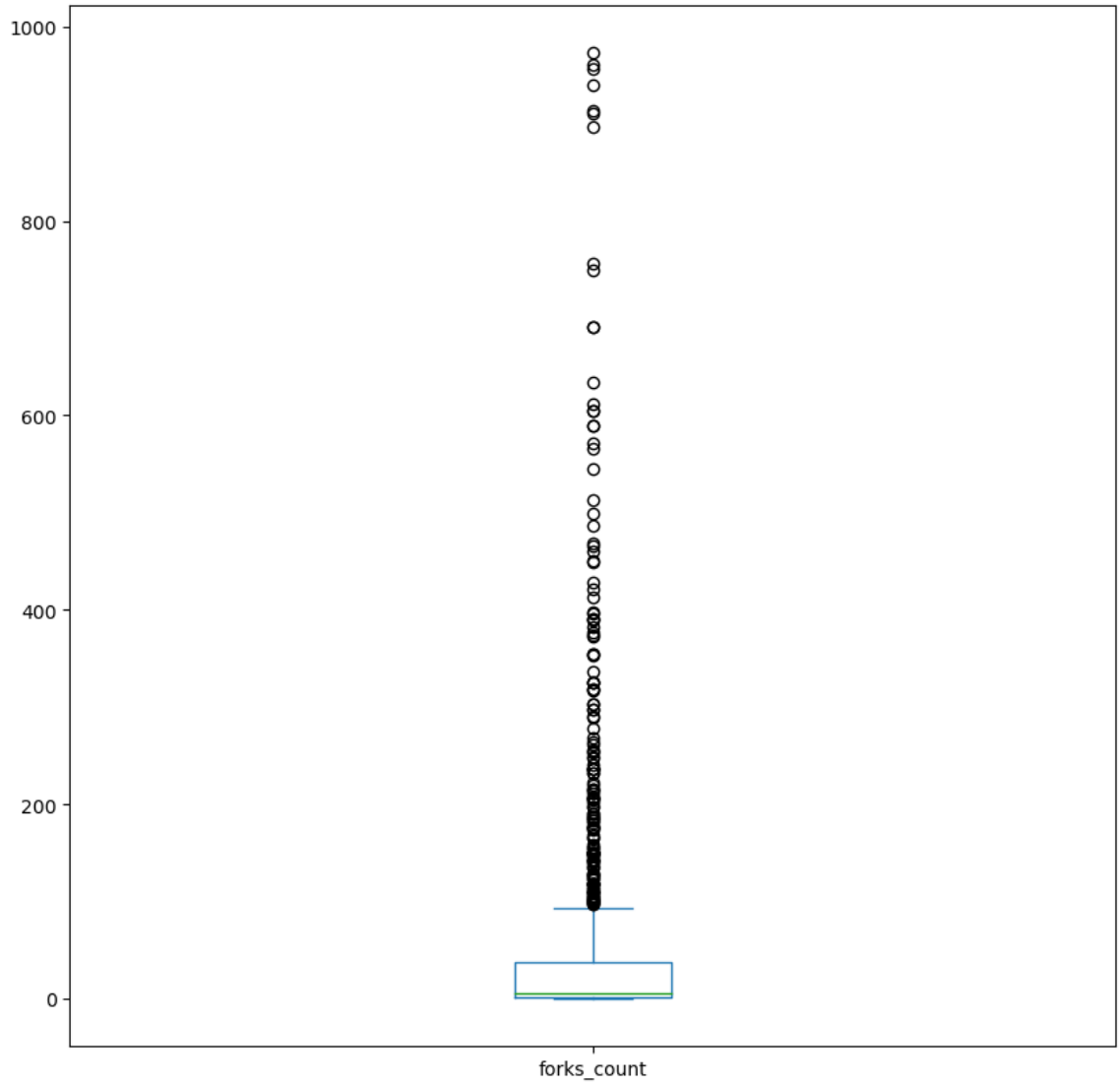
In [ ]:
```
attribute = "forks_count"

#五数概括
print('Min:', dataset[attribute].quantile(0))
print('Q1:', dataset[attribute].quantile(0.25))
print('Q2:', dataset[attribute].quantile(0.5))
print('Q3:', dataset[attribute].quantile(0.75))
print('Max:', dataset[attribute].quantile(1))
```

```
Min: 0.0
Q1: 1.0
Q2: 6.0
Q3: 38.25
Max: 973.0
```

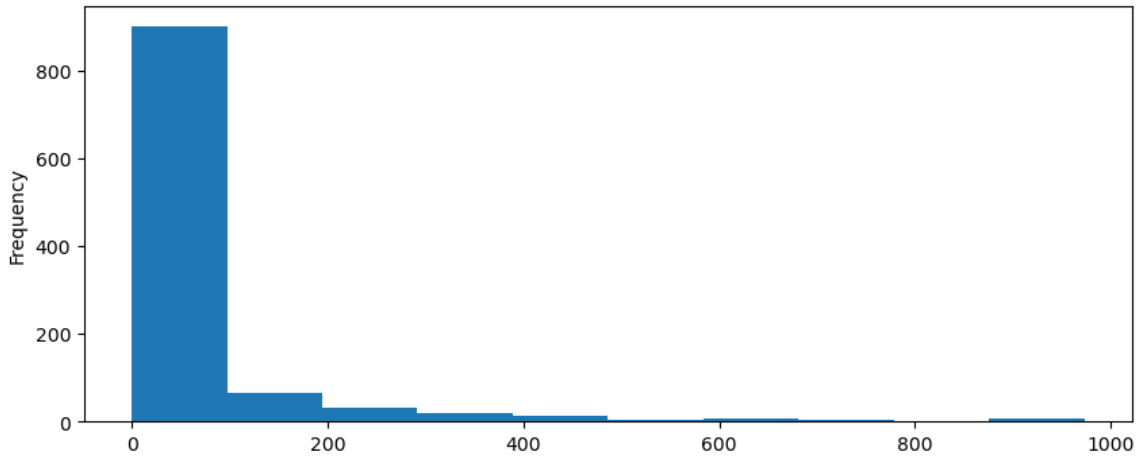
In [ ]:
```
#绘制盒图
dataset[attribute].plot(kind="box",figsize=(10,10))
```

Out[ ]:
```
<Axes: >
```



In [ ]:
```
#绘制直方图
dataset[attribute].plot(kind="hist",figsize=(10,4))
```

Out[ ]: <Axes: ylabel='Frequency'>



In [ ]:
```
#查找离群点
Q1 = dataset[attribute].quantile(0.25)
Q3 = dataset[attribute].quantile(0.75)
outliner = Q3 + (Q3 - Q1) * 1.5
print(f"大于{outliner}的项被识别为离群点")
```

大于94.125的项被识别为离群点

issues_count 数值属性

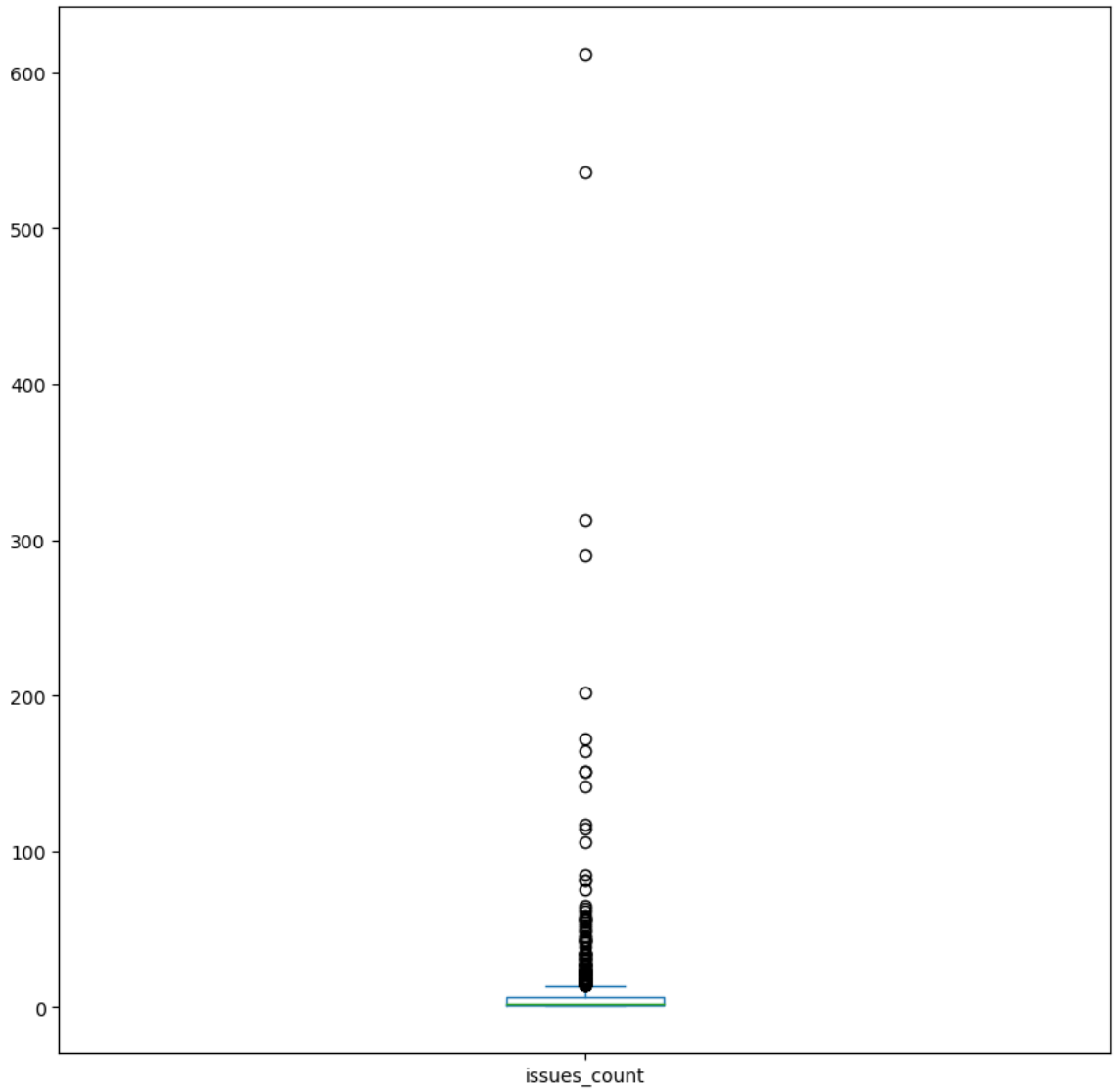
In [ ]:
```
attribute = "issues_count"

#五数概括
print('Min:', dataset[attribute].quantile(0))
print('Q1:', dataset[attribute].quantile(0.25))
print('Q2:', dataset[attribute].quantile(0.5))
print('Q3:', dataset[attribute].quantile(0.75))
print('Max:', dataset[attribute].quantile(1))
```

```
Min: 1.0
Q1: 1.0
Q2: 2.0
Q3: 6.0
Max: 612.0
```

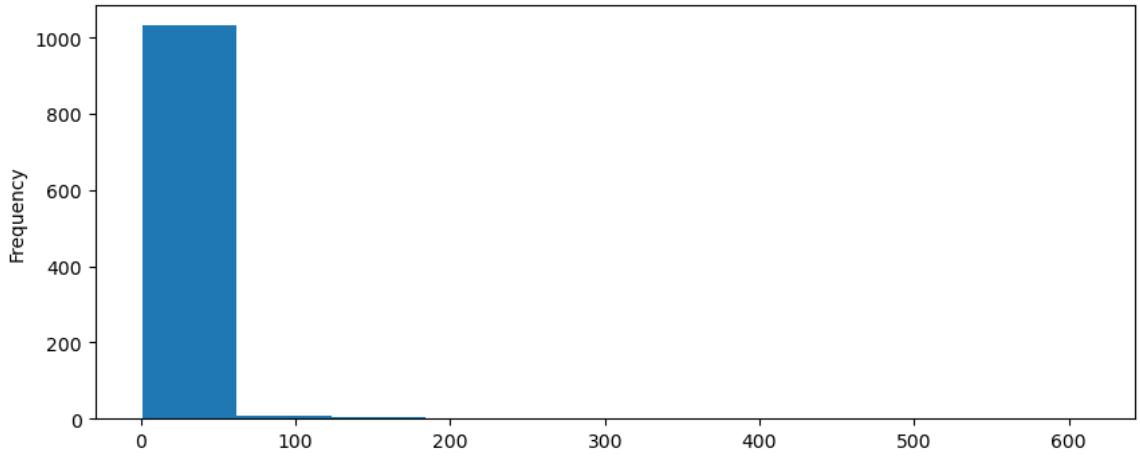
In [ ]:
```
#绘制盒图
dataset[attribute].plot(kind="box",figsize=(10,10))
```

Out[ ]: <Axes: >



In [ ]:
```
#绘制直方图
dataset[attribute].plot(kind="hist",figsize=(10,4))
```

Out[ ]: <Axes: ylabel='Frequency'>



In [ ]:
```
#查找离群点
Q1 = dataset[attribute].quantile(0.25)
Q3 = dataset[attribute].quantile(0.75)
outliner = Q3 + (Q3 - Q1) * 1.5
print(f"大于{outliner}的项被识别为离群点")
```

大于13.5的项被识别为离群点

pull_requests 数值属性

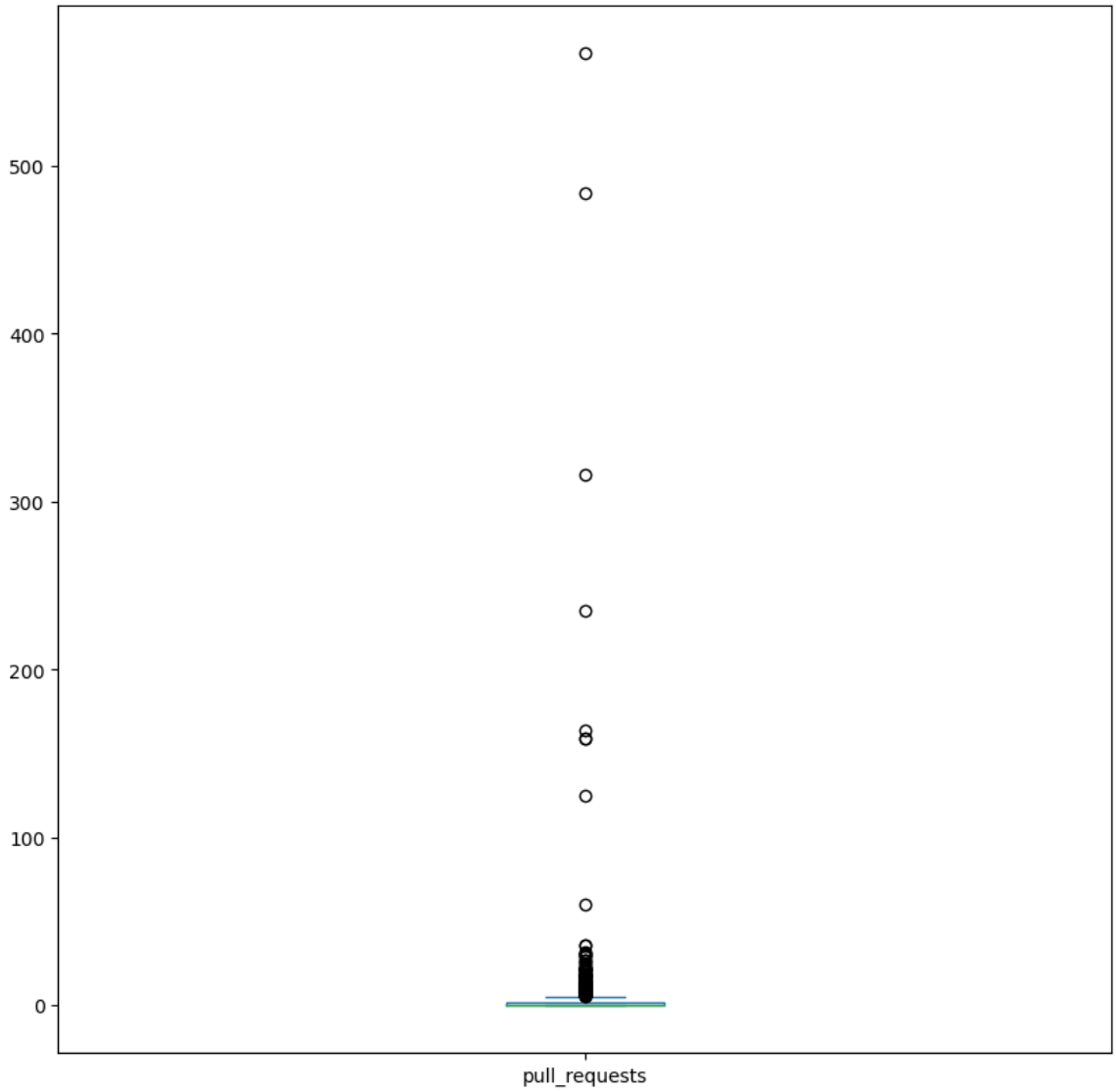
In [ ]:
```
attribute = "pull_requests"

#五数概括
print('Min:', dataset[attribute].quantile(0))
print('Q1:', dataset[attribute].quantile(0.25))
print('Q2:', dataset[attribute].quantile(0.5))
print('Q3:', dataset[attribute].quantile(0.75))
print('Max:', dataset[attribute].quantile(1))
```

```
Min: 0.0
Q1: 0.0
Q2: 0.0
Q3: 2.0
Max: 567.0
```

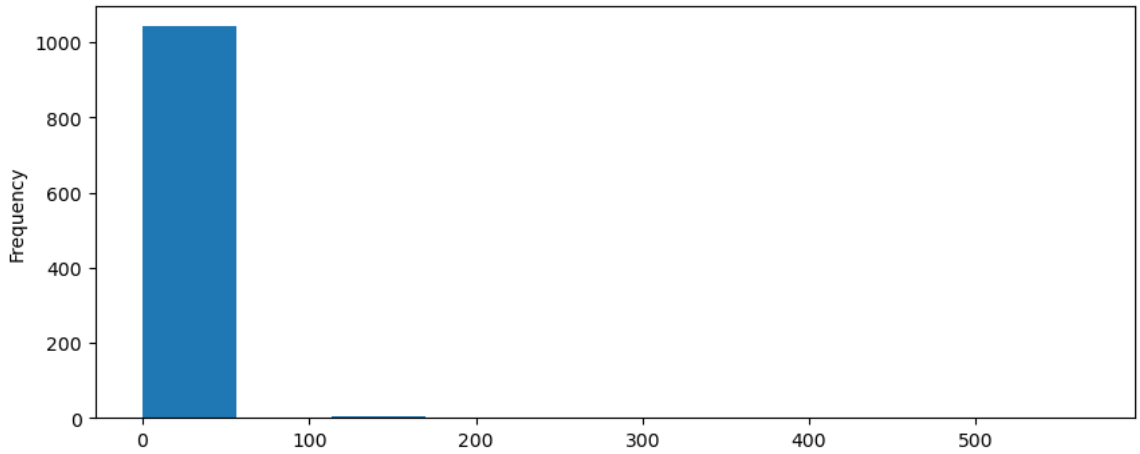
In [ ]:
```
#绘制盒图
dataset[attribute].plot(kind="box",figsize=(10,10))
```

Out[ ]: <Axes: >



In [ ]:
```
#绘制直方图
dataset[attribute].plot(kind="hist",figsize=(10,4))
```

Out[ ]:
```
<Axes: ylabel='Frequency'>
```



In [ ]:
```
#查找离群点
Q1 = dataset[attribute].quantile(0.25)
Q3 = dataset[attribute].quantile(0.75)
outliner = Q3 + (Q3 - Q1) * 1.5
print(f"大于{outliner}的项被识别为离群点")
```

大于5.0的项被识别为离群点

contributors 数值属性

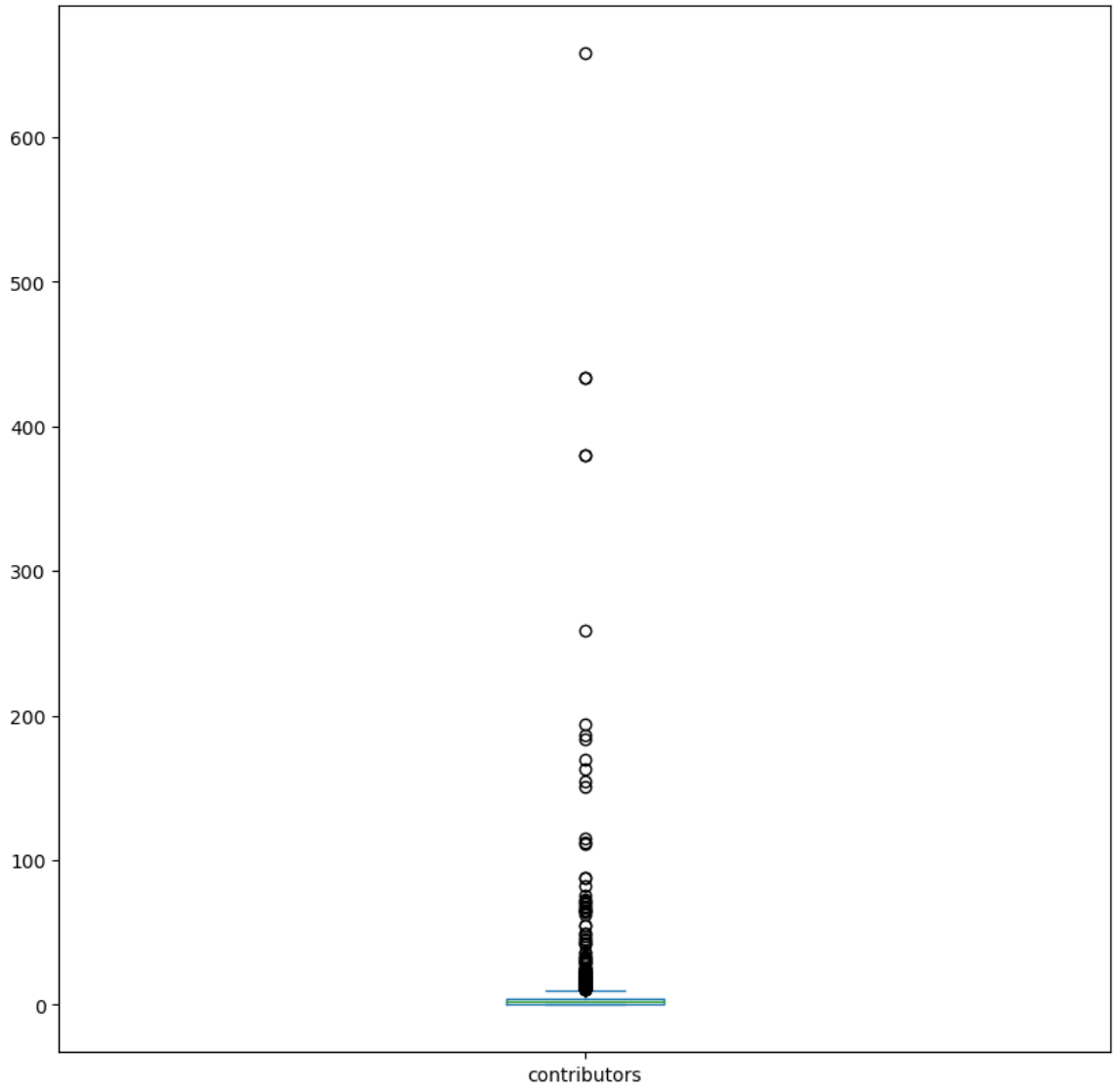
In [ ]:
```
attribute = "contributors"

#五数概括
print('Min:', dataset[attribute].quantile(0))
print('Q1:', dataset[attribute].quantile(0.25))
print('Q2:', dataset[attribute].quantile(0.5))
print('Q3:', dataset[attribute].quantile(0.75))
print('Max:', dataset[attribute].quantile(1))
```

```
Min: 0.0
Q1: 0.0
Q2: 2.0
Q3: 4.0
Max: 658.0
```

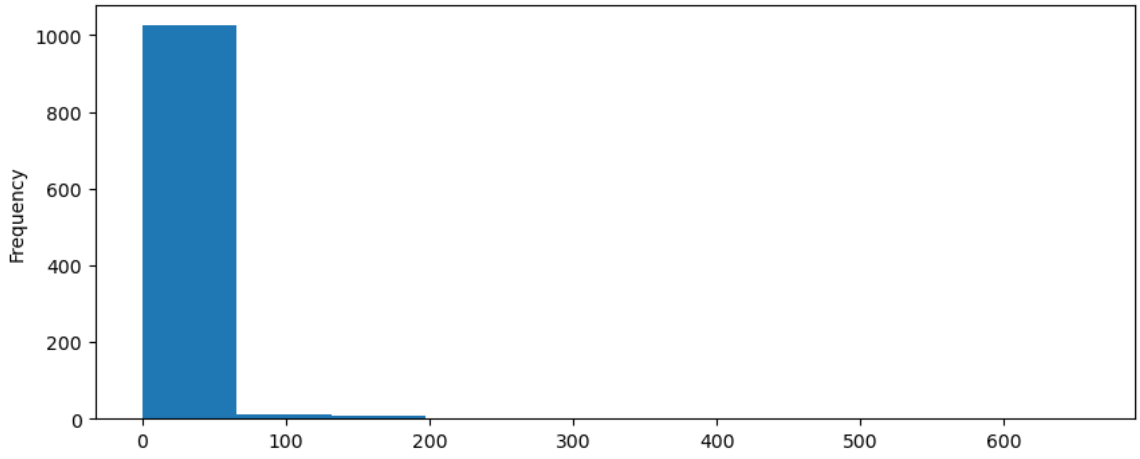
In [ ]:
```
#绘制盒图
dataset[attribute].plot(kind="box",figsize=(10,10))
```

Out[ ]:
```
<Axes: >
```



In [ ]:
```
#绘制直方图
dataset[attribute].plot(kind="hist",figsize=(10,4))
```

Out[ ]: <Axes: ylabel='Frequency'>



In [ ]:
```
#查找离群点
Q1 = dataset[attribute].quantile(0.25)
Q3 = dataset[attribute].quantile(0.75)
outliner = Q3 + (Q3 - Q1) * 1.5
print(f"大于{outliner}的项被识别为离群点")
```

大于10.0的项被识别为离群点

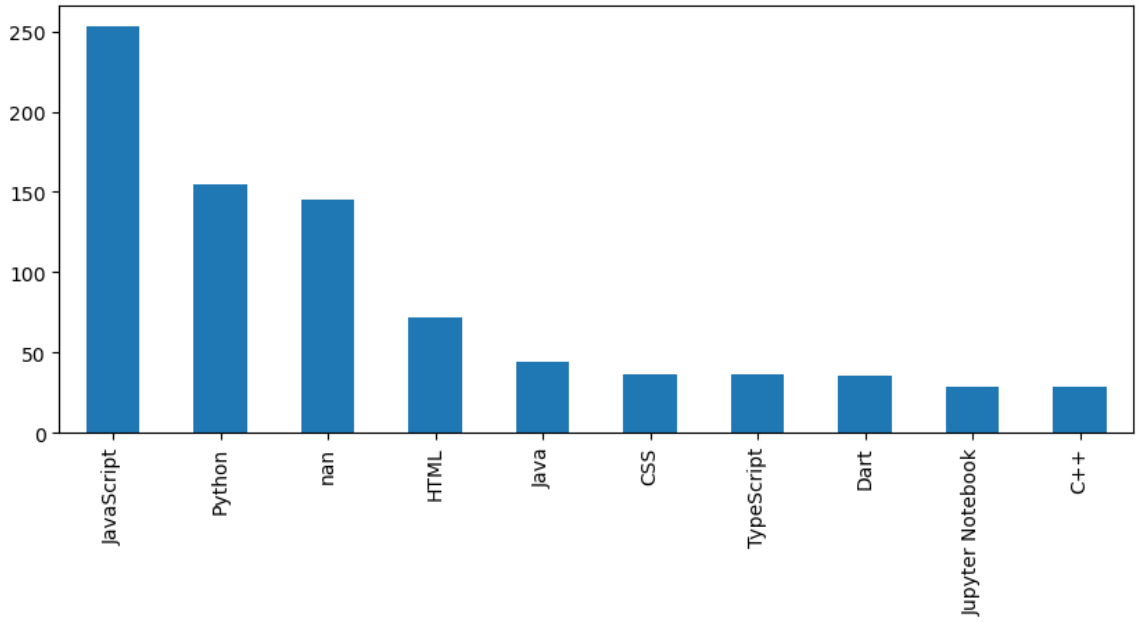
language 标称属性

In [ ]:
```
attribute = "language"
dataset_language = dataset[attribute].value_counts(dropna=False)
dataset_language
```

Out[ ]:
```
JavaScript          253
Python              155
NaN                 145
HTML                 72
Java                 44
CSS                  37
TypeScript           37
Dart                 36
Jupyter Notebook     29
C++                  29
Ruby                 28
C                    26
Shell                25
PHP                  16
Go                   15
Swift                10
Rust                 10
C#                    8
Objective-C           8
Kotlin                7
Makefile              6
Jinja                 5
SCSS                  4
AutoHotkey            3
Dockerfile            3
CoffeeScript          3
Perl                  3
Solidity              3
Vim Script            2
Pawn                  2
Assembly              2
PowerShell            2
Hack                  2
CodeQL                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: language, dtype: int64
```

In [ ]:
```
#数据太大，仅显示前10个
dataset_language[:10].plot(kind="bar", figsize=(10,4))
```

Out[ ]:
```
<Axes: >
```



数据缺失的处理

In [ ]:
```
dataset_new = dataset
#统计所有属性数据的缺失值个数
print(dataset.isnull().sum(axis=0))
```

```
repositories       0
stars_count        0
forks_count        0
issues_count       0
pull_requests      0
contributors       0
language         145
dtype: int64
```

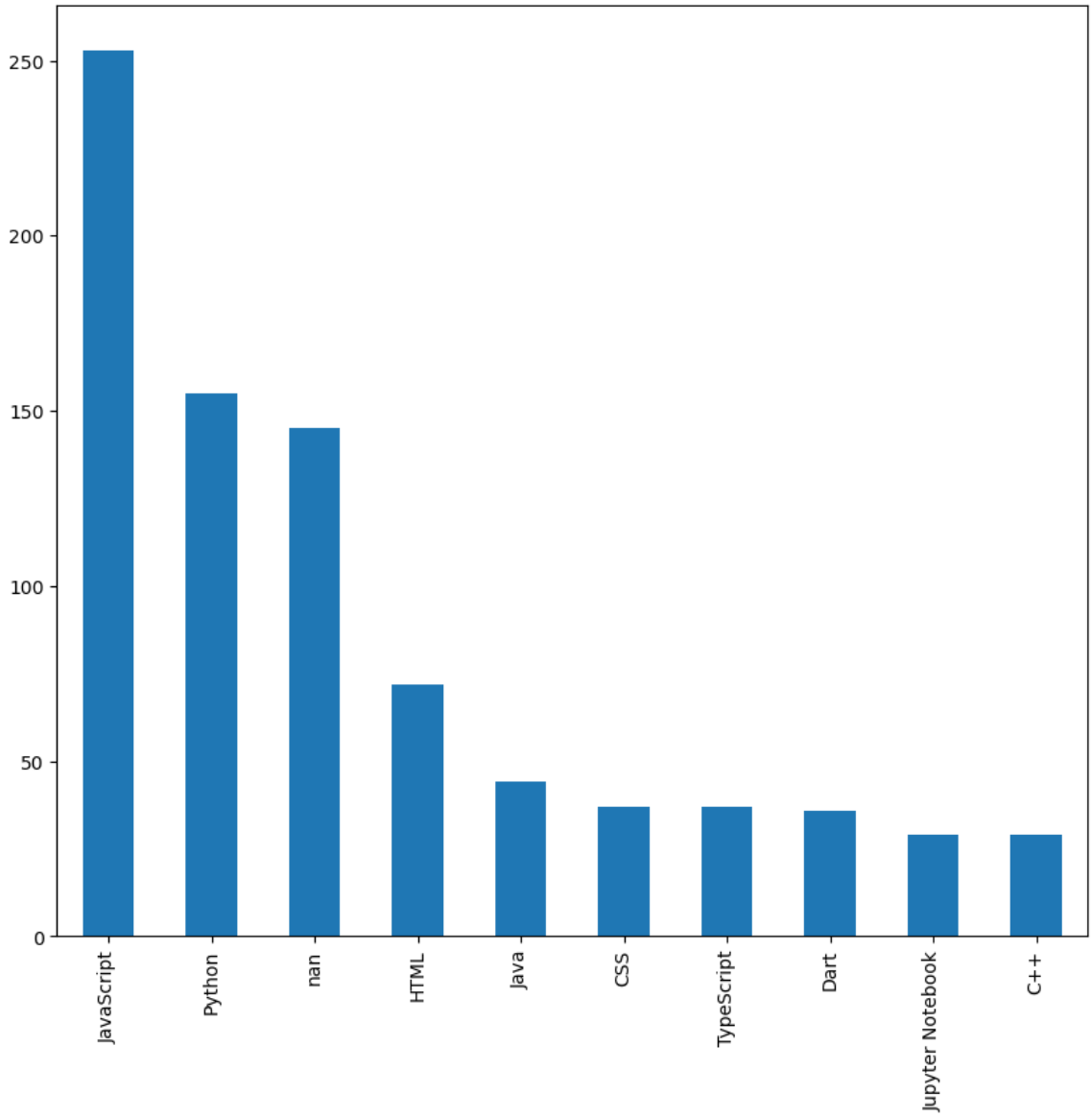
处理**language**属性的缺失

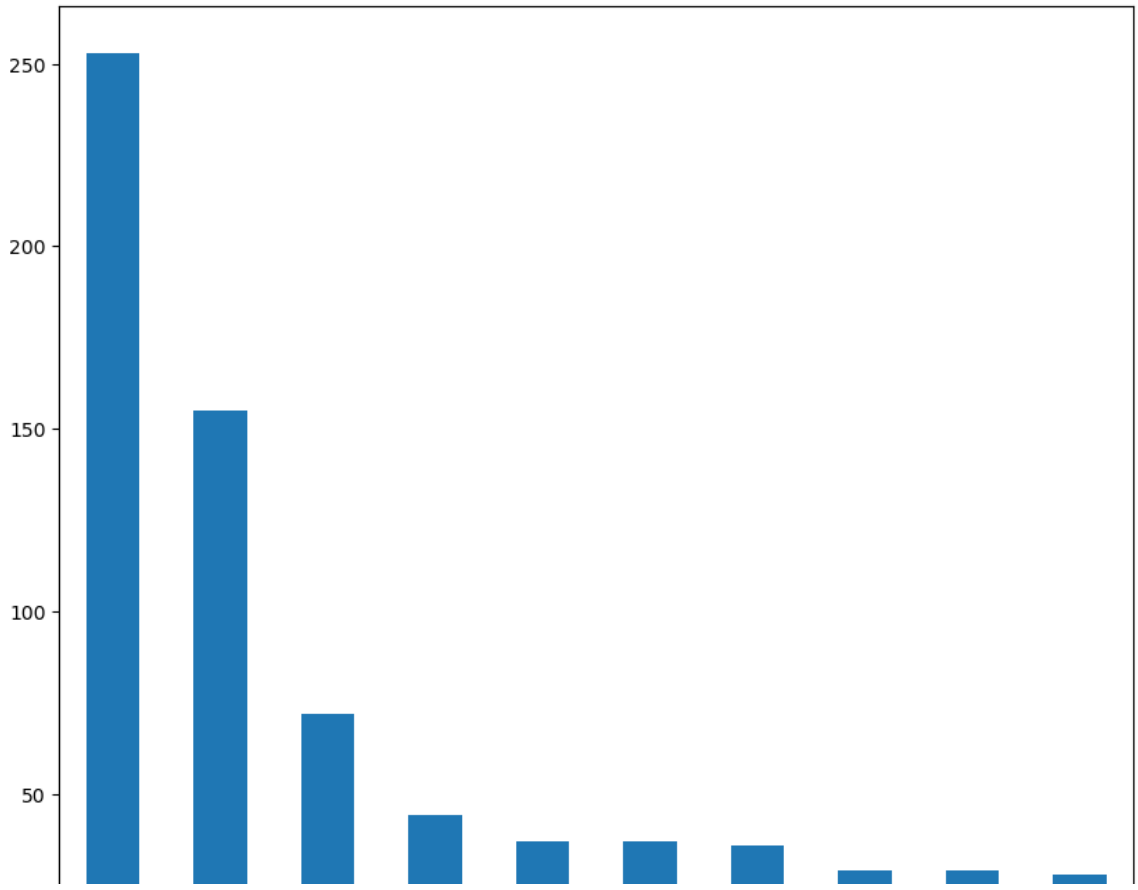
缺失的原因：可能是在数据统计的过程中出现了错误，毕竟每个代码仓库应当都有自己使用的语言。 选择策略：将缺失部分剔除

In [ ]:
```
attribute = "language"
dataset_new = dataset_new.dropna(subset=[attribute])
```

In [ ]:
```
#对比新旧数据集
plt.subplot(2,1,1)
dataset[attribute].value_counts(dropna=False)[:10].plot(kind="bar",figsize=(10,8
plt.subplot(2,1,2)
dataset_new[attribute].value_counts(dropna=False)[:10].plot(kind="bar",figsize=(
```

Out[ ]:
```
<Axes: >
```


250
200
150
100
50
0
JavaScript
Python
nan
HTML
Java
CSS
TypeScript
Dart
Jupyter Notebook
C++


250
200
150
100
50