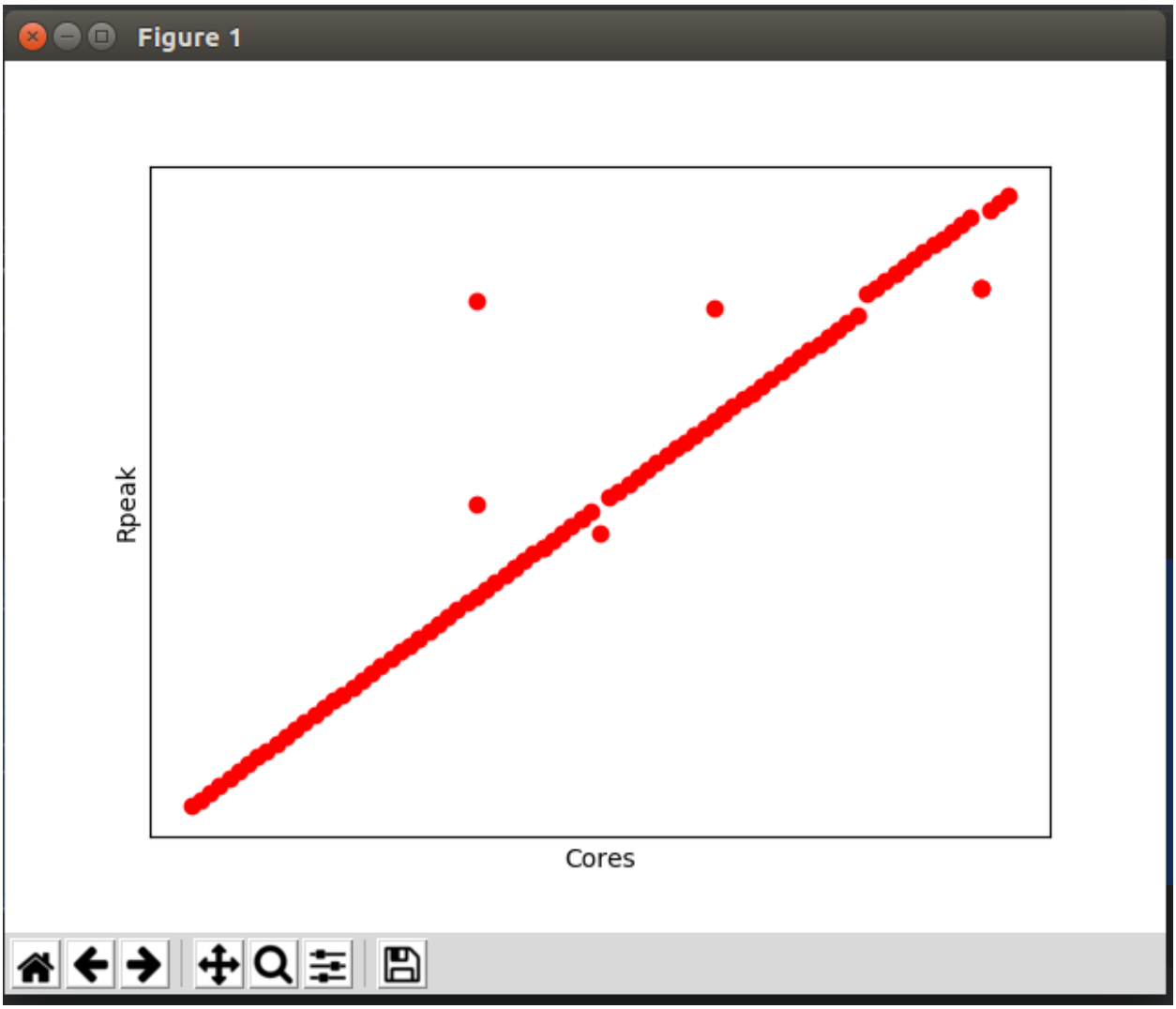
Web-Scraping Project

b.

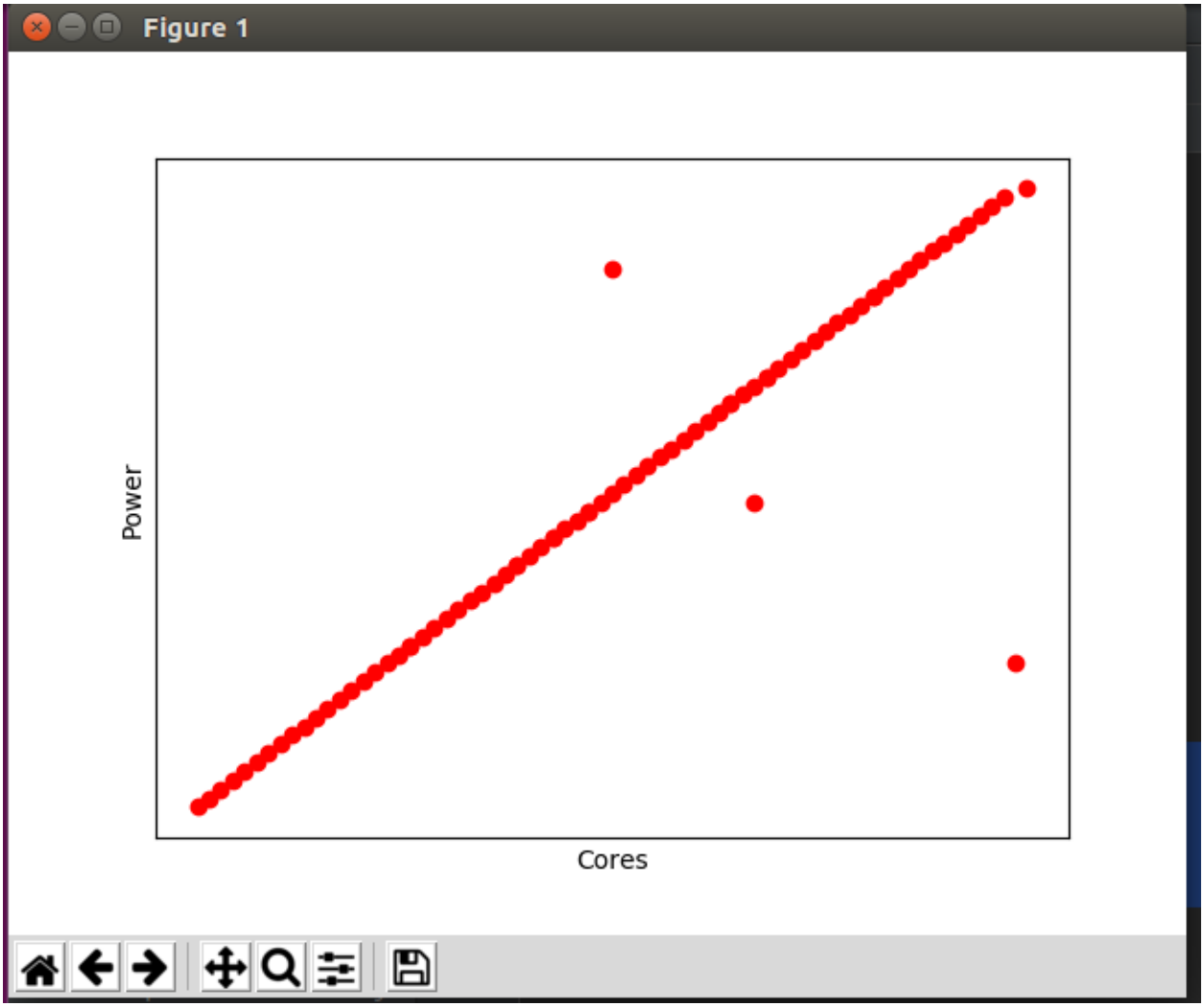
Summary Statistics using describe() function. This includes null values as well.

	Rank	Cores	Rmax	Rpeak	Power
count	100	100	100	100	80
unique	100	87	90	88	70
top	34	38,400	1,729.0	3,072.0	350
freq	1	3	3	4	2

c. Plot for Cores vs Rpeak



Plot for Cores vs Power



Code

The following code was executed in smaller chunks. Selecting part of script then right click and pressing “Execute Selection in Console.

```
import urllib2

import bs4 as bs

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt


url = "https://www.top500.org/list/2018/06/?page=1"

html = urllib2.urlopen(url)

soup = bs.BeautifulSoup(html, "lxml")


all_tables = soup.find_all('table')


right_table = soup.find('table', class_='table table-condensed table-striped')


A = []

B = []

C = []

D = []

E = []

F = []

G = []


for row in right_table.findAll("tr"):

   cells = row.findAll('td')

   if len(cells) == 7:
```

```
        A.append(cells[0].find(text=True))
        B.append(cells[1].find(text=True))
        C.append(cells[2].find(text=True))
        D.append(cells[3].find(text=True))
        E.append(cells[4].find(text=True))
        F.append(cells[5].find(text=True))
        G.append(cells[6].find(text=True))

df = pd.DataFrame(A, columns=['Rank'])
df['Site'] = B
df['System'] = C
df['Cores'] = D
df['Rmax'] = E
df['Rpeak'] = F
df['Power'] = G

print df

# print soup.table.td.string

print(df.describe())

#ax = plt.axes()

df.to_csv('out.csv', sep=',', encoding='utf-8')



plt.plot(df.Cores,df.Rpeak,'ro')
plt.xticks([], [])
```

```
plt.yticks([], [])
plt.xlabel('Cores')
plt.ylabel('Rpeak')
#ax.yaxis.set_major_locator(plt.NullLocator())
#ax.xaxis.set_major_formatter(plt.NullFormatter())

df.dtypes



dftest = df.infer_objects()




plt.show()



df.loc[df['Power'] == 'None']

print dftest

tryd = dftest.dropna(subset=['Power'])

print tryd

plt.plot(tryd.Cores,tryd.Power,'ro')
plt.xticks([], [])
plt.yticks([], [])
plt.xlabel('Cores')
```

```
plt.ylabel('Power')
plt.show()
```