In [1]:
```
import itertools
import pandas as pd

# The new library!
from thefuzz import fuzz, process
```

In [2]:
```
df1 = pd.read_csv('companies_1.csv')
df2 = pd.read_csv('companies_2.csv')
```

In [3]:
```
df1.shape
```

Out[3]:
```
(266, 1)
```

In [4]:
```
df2.shape
```

Out[4]:
```
(368, 1)
```

In [5]:
```
df1.head()
```

Out[5]:

| | CLIENT |
|---|---|
| 0 | Adobe Systems, Inc. |
| 1 | Adventist Health |
| 2 | AECOM |
| 3 | Aerojet Rockedyne Holdings (GenCorp) |
| 4 | Alameda-Contra Costa Transit District |

In [6]:
```
df2.head()
```

Out[6]:

| | Firm Name |
|---|---|
| 0 | AAA Northern California, Nevada & Utah Auto Ex... |
| 1 | ACCO Engineered Systems |
| 2 | Adams County Retirement Plan |
| 3 | Adidas America, Inc. |
| 4 | Adobe Systems, Inc. |

In [8]:
```
com1=["A","B","C"]
com2=['A Inc','B Inc','C Inc']
```

In [9]:
```
list(itertools.product(com1,com2))
```

Out[9]:
```
[('A', 'A Inc'),
 ('A', 'B Inc'),
 ('A', 'C Inc'),
 ('B', 'A Inc'),
 ('B', 'B Inc'),
 ('B', 'C Inc'),
 ('C', 'A Inc'),
 ('C', 'B Inc'),
 ('C', 'C Inc')]
```

## Data Preprocessing

### *Create the `df` dataframe containing the product of the two CSVs*

In [10]:
```
df = pd.DataFrame(
    itertools.product(df1["CLIENT"].values,df2["Firm Name"]),
    columns=["CSV 1","CSV 2"]
)
```

It'll look something like:

In [12]:
```
df.head()
```

Out[12]:

| | CSV 1 | CSV 2 |
|---|---|---|
| 0 | Adobe Systems, Inc. | AAA Northern California, Nevada & Utah Auto Ex... |
| 1 | Adobe Systems, Inc. | ACCO Engineered Systems |
| 2 | Adobe Systems, Inc. | Adams County Retirement Plan |
| 3 | Adobe Systems, Inc. | Adidas America, Inc. |
| 4 | Adobe Systems, Inc. | Adobe Systems, Inc. |

In [13]:
```
df.shape
```

Out[13]:
```
(97888, 2)
```

## Calculating the Levenshtein distance

In [15]:
```
fuzz.partial_ratio("Apple", "Apple Inc.")
```

Out[15]:
```
100
```

In [16]:
```
fuzz.partial_ratio("Microsoft", "Apple Inc.")
```

Out[16]:
```
11
```

In [17]:
```
fuzz.partial_ratio("Microsoft", "MSFT")
```

Out[17]:
```
25
```

In [18]:
```
A = ["Apple", "Alphabet", "Microsoft"]
B = ["MSFT", "Alphabet/Google", "Apple inc."]
```

In [19]:
```
companies = list(itertools.product(A, B))
companies
```

Out[19]:
```
[('Apple', 'MSFT'),
 ('Apple', 'Alphabet/Google'),
 ('Apple', 'Apple inc.'),
 ('Alphabet', 'MSFT'),
 ('Alphabet', 'Alphabet/Google'),
 ('Alphabet', 'Apple inc.'),
 ('Microsoft', 'MSFT'),
 ('Microsoft', 'Alphabet/Google'),
 ('Microsoft', 'Apple inc.')]
```

In [20]:
```
for c1, c2 in companies:
    ratio = fuzz.partial_ratio(c1, c2)
    print(f"{c1} > {c2}: {ratio}")
```

```
Apple > MSFT: 0
Apple > Alphabet/Google: 40
Apple > Apple inc.: 100
Alphabet > MSFT: 0
Alphabet > Alphabet/Google: 100
Alphabet > Apple inc.: 38
Microsoft > MSFT: 25
Microsoft > Alphabet/Google: 22
Microsoft > Apple inc.: 22
```

***Create a new column `Ratio Score` that contains the distance for all the rows in `df`***

In [21]:
```
score = [fuzz.partial_ratio(c1,c2) for c1,c2 in df.values]
```

In [22]:
```
score[:10]
```

Out[22]:
```
[26, 56, 32, 47, 100, 53, 21, 33, 53, 49]
```

In [23]:
```
df['Ratio Score'] = score
```

It'll look something like this:

In [24]:
```
df.head(10)
```

Out[24]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **0** | Adobe Systems, Inc. | AAA Northern California, Nevada & Utah Auto Ex... | 26 |
| **1** | Adobe Systems, Inc. | ACCO Engineered Systems | 56 |
| **2** | Adobe Systems, Inc. | Adams County Retirement Plan | 32 |
| **3** | Adobe Systems, Inc. | Adidas America, Inc. | 47 |
| **4** | Adobe Systems, Inc. | Adobe Systems, Inc. | 100 |
| **5** | Adobe Systems, Inc. | Advanced Micro Devices, Inc. | 53 |
| **6** | Adobe Systems, Inc. | AECOM Technology Corporation | 21 |
| **7** | Adobe Systems, Inc. | Aera Energy LLC | 33 |
| **8** | Adobe Systems, Inc. | Aerojet Rocketdyne Holdings, Inc. | 53 |
| **9** | Adobe Systems, Inc. | Agilent Technologies, Inc. | 49 |

In [26]:
```
df.shape
```

Out[26]:
```
(97888, 3)
```

In [28]:
```
df.loc[df["Ratio Score"]>=90].shape
```

Out[28]:
```
(106, 3)
```

In [35]:
```
df.loc[
    (df["CSV 1"]=="AECOM") &
    (df["Ratio Score"]>=80)
]
```

Out[35]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **742** | AECOM | AECOM Technology Corporation | 100 |

In [44]:
```
df.query("`CSV 1`=='AECOM' and `Ratio Score`>80")
```

Out[44]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **742** | AECOM | AECOM Technology Corporation | 100 |

In [38]:
```
df.loc[
    (df["CSV 1"]=="Starbucks") &
    (df["Ratio Score"]>=60)
]
```

Out[38]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **77948** | Starbucks | Starbucks Corporation | 100 |

In [45]:
```
df.query("`CSV 1`=='Starbucks' and `Ratio Score`>80")
```

Out[45]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **77948** | Starbucks | Starbucks Corporation | 100 |

In [43]:
```
df.loc[
    (df["CSV 1"]=="Pinnacle West Capital Corporation") &
    (df["Ratio Score"]>=80)
]
```

Out[43]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **61128** | Pinnacle West Capital Corporation | Avista Corporation | 83 |
| **61130** | Pinnacle West Capital Corporation | Ball Corporation | 88 |

In [47]:
```
df.query("`CSV 1`=='Pinnacle West Capital Corporation' and `Ratio Score`>90")
```

Out[47]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|

In [52]:
```
df.query("`CSV 1`=='County of Los Angeles Deferred Compensation Program' and `
```

Out[52]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **26206** | County of Los Angeles Deferred Compensation Pr... | City of Los Angeles Deferred Compensation | 95 |
| **26227** | County of Los Angeles Deferred Compensation Pr... | County of Los Angeles Deferred Compensation Pr... | 100 |

In [55]:
```
df.loc[
    (df["CSV 1"]=="County of Los Angeles Deferred Compensation Program") &
    (df["Ratio Score"]>=90)
]
```

Out[55]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **26206** | County of Los Angeles Deferred Compensation Pr... | City of Los Angeles Deferred Compensation | 95 |
| **26227** | County of Los Angeles Deferred Compensation Pr... | County of Los Angeles Deferred Compensation Pr... | 100 |

In [53]:
```
df.query("`CSV 1`=='The Queens Health Systems' and `Ratio Score`>90")
```

Out[53]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **84220** | The Queens Health Systems | The Queen's Health Systems | 96 |

In [56]:
```
df.loc[
    (df["CSV 1"]=="The Queens Health Systems") &
    (df["Ratio Score"]>=90)
]
```

Out[56]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **84220** | The Queens Health Systems | The Queen's Health Systems | 96 |

In [57]:
```
df.loc[df["Ratio Score"]>90]
```

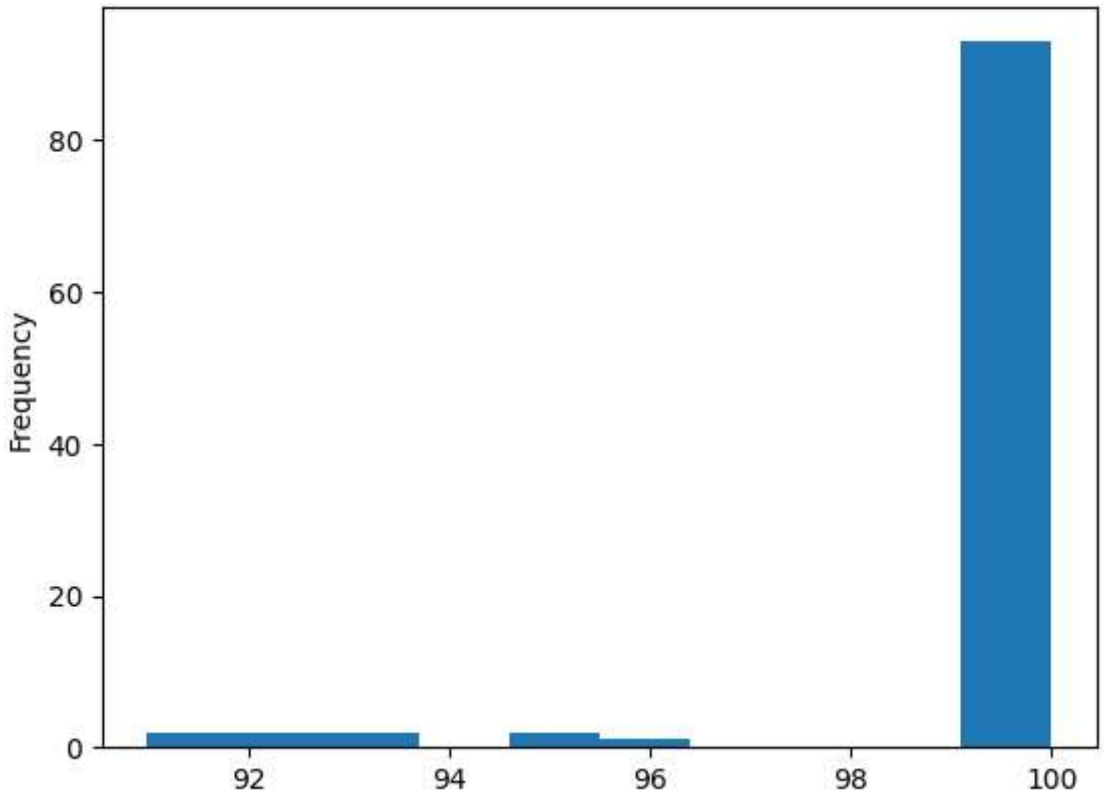
Out[57]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **4** | Adobe Systems, Inc. | Adobe Systems, Inc. | 100 |
| **742** | AECOM | AECOM Technology Corporation | 100 |
| **1484** | Alameda-Contra Costa Transit District | Alameda-Contra Costa Transit District | 100 |
| **3697** | Amazon | Amazon.com Holdings, Inc. | 100 |
| **4435** | Amgen Inc. | Amgen Inc. | 100 |
| **...** | ... | ... | ... |
| **94923** | Virginia Mason Medical Center | Virginia Mason Medical Center | 100 |
| **96033** | Wells Fargo | Wells Fargo & Company | 100 |
| **96402** | Western Digital | Western Digital Corp. | 100 |
| **96771** | Western Union Financial Services, Inc. | Western Union Financial Services, Inc. | 100 |
| **97141** | Weyerhaeuser Company | Weyerhaeuser Company | 100 |

102 rows × 3 columns

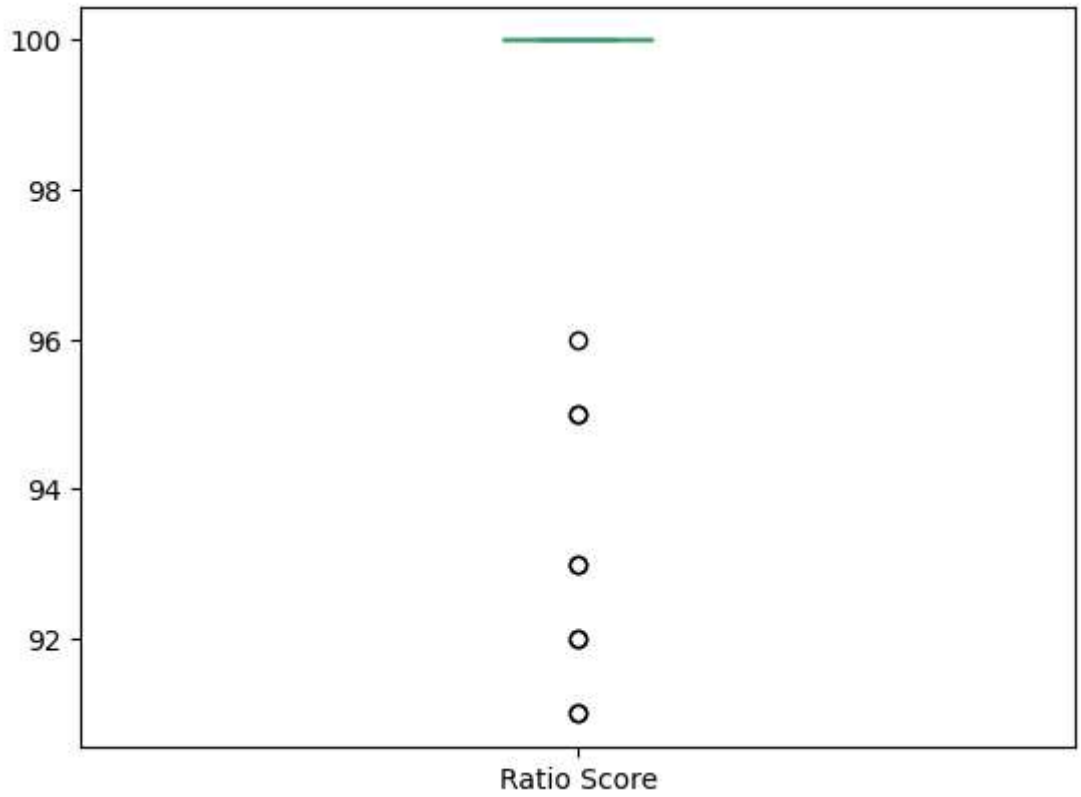
In [58]:
```
df.loc[df["Ratio Score"]>90, "Ratio Score"].plot(kind="hist")
```

Out[58]: <Axes: ylabel='Frequency'>



In [59]:
```
df.loc[df["Ratio Score"]>90, "Ratio Score"].plot(kind="box")
```

Out[59]: <Axes: >




In [61]:
```
df.query("`Ratio Score`>90 and `Ratio Score`<97").sort_values(by="Ratio Score"
```

Out[61]:

| | CSV 1 | CSV 2 | Ratio Score |
|---|---|---|---|
| **25617** | Contra Costa County Employees Retirement Assoc... | Marin County Employees Retirement Association | 91 |
| **25681** | Contra Costa County Employees Retirement Assoc... | Sonoma County Employees Retirement Association | 91 |
| **63526** | Presbytherian | Presbyterian Healthcare Services | 92 |
| **67596** | Safeway, Inc. | Safeway Inc. | 92 |
| **39189** | Idaho Power Co. | Idaho Power Company | 93 |
| **66859** | Sacramento City Employees Retirement System | Sacramento County Employees Retirement System | 93 |
| **26206** | County of Los Angeles Deferred Compensation Pr... | City of Los Angeles Deferred Compensation | 95 |
| **41775** | Jack in the Box, Inc. | Jack in the Box Inc. | 95 |
| **84220** | The Queens Health Systems | The Queen's Health Systems | 96 |