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 [6]:
```
type(df1['CLIENT'].values)
```

Out[6]:
```
numpy.ndarray
```

In [4]:
```
df2.head()
```

Out[4]:

	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.

Data Preprocessing

1. Create the `df` dataframe containing the product of the two CSVs

In [12]:
```
df = pd.DataFrame(
itertools.product(df1['CLIENT'].values, df2['Firm Name']),
columns=['CSV 1', 'CSV2'])
```

In [18]:
```
df
```

Out[18]:

	CSV 1	CSV2
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.
...	...	...
97883	WRS	Yakima Valley Memorial Hospital Association
97884	WRS	Yokohama Tire Corporation
97885	WRS	Yuma Regional Medical Center
97886	WRS	Zions Bancorporation
97887	WRS	Zoological Society of San Diego

97888 rows × 2 columns

Calculating the Levenshtein distance

Now, we will learn how to calculate the Levenshtein distance between two strings. Here we will user `partial_ratio` function from the `fuzz` module to compute the "ratio" between two strings. The result is a number between `0` and `100`, with `100` indicating a "perfect" match. Please note that `partial_ratio` gives ratio of the shortest string length to the longest string length. For example, if the first string is `ABC` and the second string is `ABDC`, then the ratio will be `4/5 = 0.80`.

In [14]:
```
fuzz.partial_ratio("Apple", "Apple Inc.")
```

Out[14]:
```
100
```

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

Out[15]:
```
18
```

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

Out[16]:
```
40
```

If we have list of strings, we can calculate the Levenshtein distance between each pair of strings in the list.

In [20]:
```
A = ["Apple", "Alphabet", "Microsoft"]
B = ["MSFT", "Alphabet/Google", "Apple inc."]
```

Below, we combined the two list `A` and `B` into a list of tuples `companies` using `product` function from `itertools` module.

Then, we calculated the partial ratio for each pair of strings in the list `companies` using `partial_ratio` function from `fuzz` .

In [21]:
```
companies = list(itertools.product(A, B))
companies
```

Out[21]:
```
[('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 [22]:
```
for c1, c2 in companies:
    ratio = fuzz.partial_ratio(c1, c2)
    print(f"{c1} > {c2}: {ratio}")
```

```
Apple > MSFT: 0
Apple > Alphabet/Google: 57
Apple > Apple inc.: 100
Alphabet > MSFT: 0
Alphabet > Alphabet/Google: 100
Alphabet > Apple inc.: 46
Microsoft > MSFT: 40
Microsoft > Alphabet/Google: 29
Microsoft > Apple inc.: 31
```

You will see the greater the ratio, the more similar the strings are.

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

In [20]:
```
score = [fuzz.partial_ratio(c1, c2) for c1, c2 in df.values]
score[:10]
```

Out[20]:
```
[32, 64, 41, 50, 100, 59, 29, 35, 54, 60]
```

In [22]:
```
df['Ratio Score'] = score
df
```

Out[22]:

	CSV 1	CSV2	Ratio Score
0	Adobe Systems, Inc.	AAA Northern California, Nevada & Utah Auto Ex...	32
1	Adobe Systems, Inc.	ACCO Engineered Systems	64
2	Adobe Systems, Inc.	Adams County Retirement Plan	41
3	Adobe Systems, Inc.	Adidas America, Inc.	50
4	Adobe Systems, Inc.	Adobe Systems, Inc.	100
...	...	...	...
97883	WRS	Yakima Valley Memorial Hospital Association	0
97884	WRS	Yokohama Tire Corporation	0
97885	WRS	Yuma Regional Medical Center	33
97886	WRS	Zions Bancorporation	0
97887	WRS	Zoological Society of San Diego	33

97888 rows × 3 columns

3. How many rows have a Ratio score of `90` or more?

In [24]:
```
# Try your code here
filt = df['Ratio Score'] >= 90
df[filt].shape
```

Out[24]:
```
(135, 3)
```

4. What's the corresponding company in CSV2 to `AECOM` in CSV1?

In [26]:
```
# Try your code here
filt1 = df['CSV 1'] == 'AECOM'
filt2 = df['Ratio Score'] >= 80
df[filt1&filt2]
```

Out[26]:

	CSV 1	CSV2	Ratio Score
742	AECOM	AECOM Technology Corporation	100

In [27]:
```
df.query("`CSV 1` == 'AECOM' and `Ratio Score` > 80")
```

Out[27]:

	CSV 1	CSV2	Ratio Score
742	AECOM	AECOM Technology Corporation	100

5. What's the corresponding CSV2 company of Starbucks?

In [28]:
```
# Try your code here
filt1 = df['CSV 1'] == 'Starbucks'
filt2 = df['Ratio Score'] >= 80
df[filt1&filt2]
```

Out[28]:

	CSV 1	CSV2	Ratio Score
77948	Starbucks	Starbucks Corporation	100

6. Is there a matching company for `Pinnacle West Capital Corporation`?

In [29]:
```
# Try your code here
filt1 = df['CSV 1'] == 'Pinnacle West Capital Corporation'
filt2 = df['Ratio Score'] >= 80
df[filt1&filt2]
```

Out[29]:

	CSV 1	CSV2	Ratio Score
61128	Pinnacle West Capital Corporation	Avista Corporation	88
61130	Pinnacle West Capital Corporation	Ball Corporation	93
61266	Pinnacle West Capital Corporation	Huntsman Corporation	80
61328	Pinnacle West Capital Corporation	RAND Corporation	86
61336	Pinnacle West Capital Corporation	Rogers Corporation	80

7. How many matching companies are there for `County of Los Angeles Deferred Compensation Program`?

In [31]:
```
# Try your code here
pd.options.display.max_colwidth = None

filt1 = df['CSV 1'] == 'County of Los Angeles Deferred Compensation Prog
filt2 = df['Ratio Score'] >= 80
df[filt1&filt2]
```

Out[31]:

	CSV 1	CSV2	Ratio Score
26206	County of Los Angeles Deferred Compensation Program	City of Los Angeles Deferred Compensation	95
26227	County of Los Angeles Deferred Compensation Program	County of Los Angeles Deferred Compensation Program	100
26229	County of Los Angeles Deferred Compensation Program	County of Riverside Deferred Compensation Program	82
26230	County of Los Angeles Deferred Compensation Program	County of San Diego Deferred Compensation Program	82
26233	County of Los Angeles Deferred Compensation Program	County of Weld	83
26330	County of Los Angeles Deferred Compensation Program	King County Deferred Compensation Program	85
26352	County of Los Angeles Deferred Compensation Program	Marin County Deferred Compensation Program	83

8. Is there a matching company for `The Queens Health Systems` ?

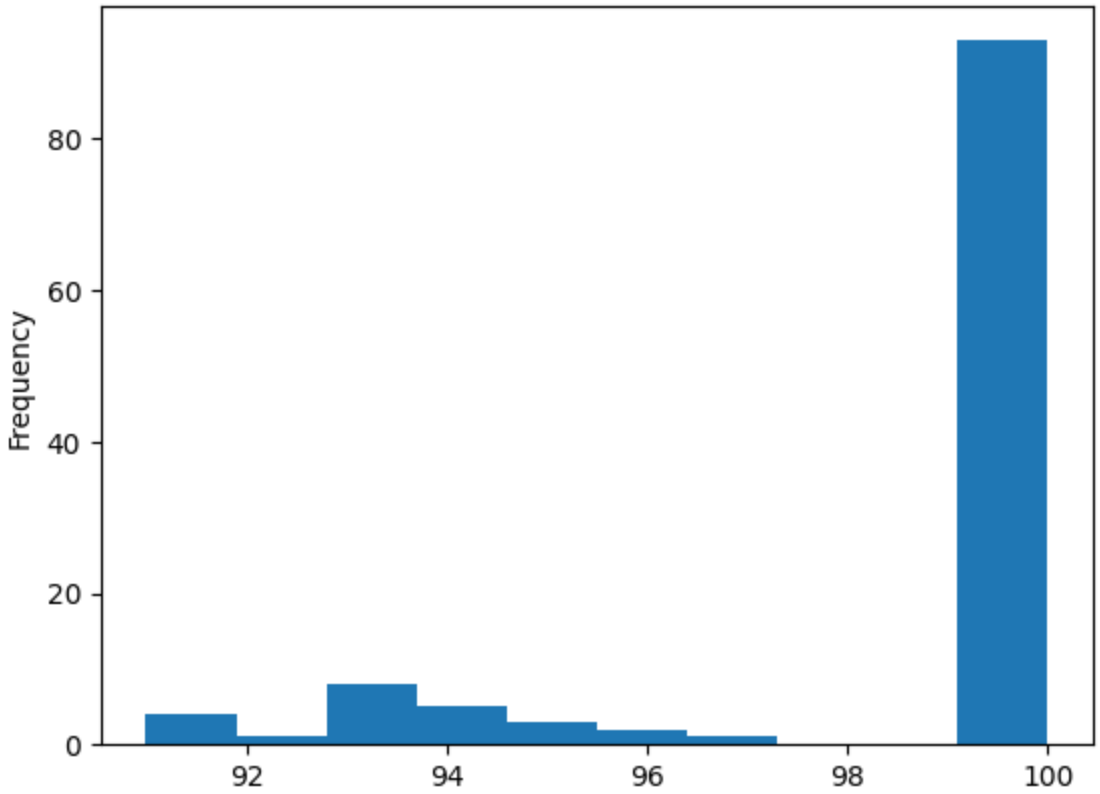
In [32]:
```
# Try your code here
filt1 = df['CSV 1'] == 'The Queens Health Systems'
filt2 = df['Ratio Score'] >= 80
df[filt1&filt2]
```

Out[32]:

	CSV 1	CSV2	Ratio Score
84113	The Queens Health Systems	Legacy Health System	80
84149	The Queens Health Systems	Renown Health System	84
84220	The Queens Health Systems	The Queen's Health Systems	96

In [36]:
```
filt = df['Ratio Score'] > 90
df.loc[filt, 'Ratio Score'].plot(kind='hist')
```

Out[36]: <Axes: ylabel='Frequency'>



In [38]:
```
filt1 = df['Ratio Score'] > 90
filt2 = df['Ratio Score'] < 97
df[filt1&filt2].sort_values(by='Ratio Score')
```

Out[38]:

	CSV 1	CSV2	Ratio Score
66888	Sacramento City Employees Retirement System	Seattle City Employees Retirement System	91
25652	Contra Costa County Employees Retirement Association	San Diego County Employees Retirement Association	91
25658	Contra Costa County Employees Retirement Association	San Mateo County Employees Retirement Association	91
25690	Contra Costa County Employees Retirement Association	Stanislaus County Employees Retirement Association	91
67596	Safeway, Inc.	Safeway Inc.	92
7333	Arizona State Retirement System	Utah State Retirement Systems	93
11546	Ball Corporation	First American Financial Corporation	93
25403	Contra Costa County Employees Retirement Association	Alameda County Employees Retirement Association	93
25540	Contra Costa County Employees Retirement Association	Fresno County Employees Retirement Association	93
66859	Sacramento City Employees Retirement System	Sacramento County Employees Retirement System	93
61130	Pinnacle West Capital Corporation	Ball Corporation	93
89466	UnionBanCal Corporation	Ball Corporation	93
47671	Maricopa County Community College District	Kern Community College District	93
92706	University of Utah	University of the Pacific	94
36967	Hawaii DC	Hawaii Deferred Compensation Fund	94
25718	Contra Costa County Employees Retirement Association	Tulare County Employees Retirement Association	94
25681	Contra Costa County Employees Retirement Association	Sonoma County Employees Retirement Association	94
25592	Contra Costa County Employees Retirement Association	Kern County Employees' Retirement Association	94
25617	Contra Costa County Employees Retirement Association	Marin County Employees Retirement Association	95
41775	Jack in the Box, Inc.	Jack in the Box Inc.	95
26206	County of Los Angeles Deferred Compensation Program	City of Los Angeles Deferred Compensation	95
63526	Presbytherian	Presbyterian Healthcare Services	96
84220	The Queens Health Systems	The Queen's Health Systems	96

The End!