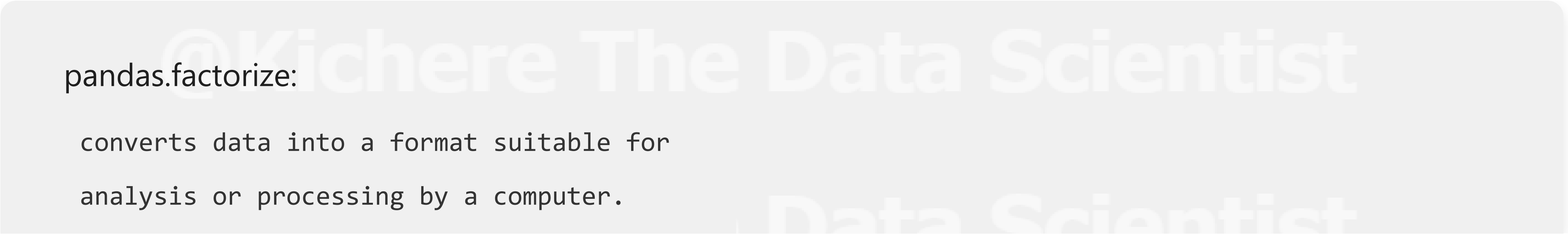
None

Contents

****

```
import numpy as np
import pandas as pd
```

Introduction



@Kichere The Data Scientist

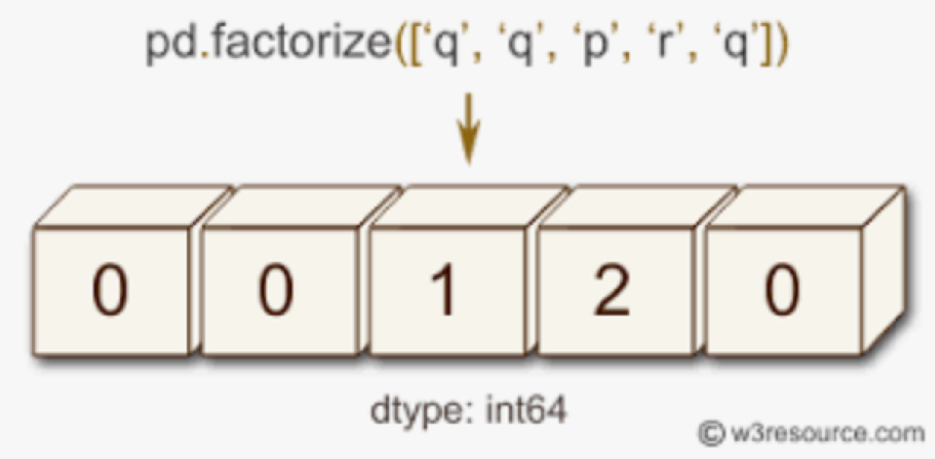
pandas.factorize:

converts data into a format suitable for

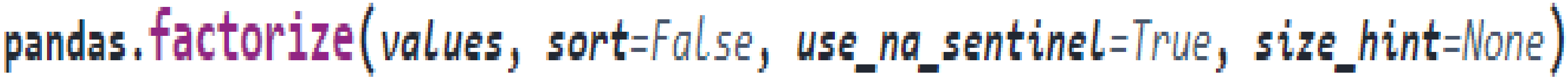
analysis or processing by a computer.

```
This method is useful for obtaining a numer
ic representation of
an array when all that matters is identifyi
ng distinct values.
```

@Kichere The



The "pandas.factorize" syntax



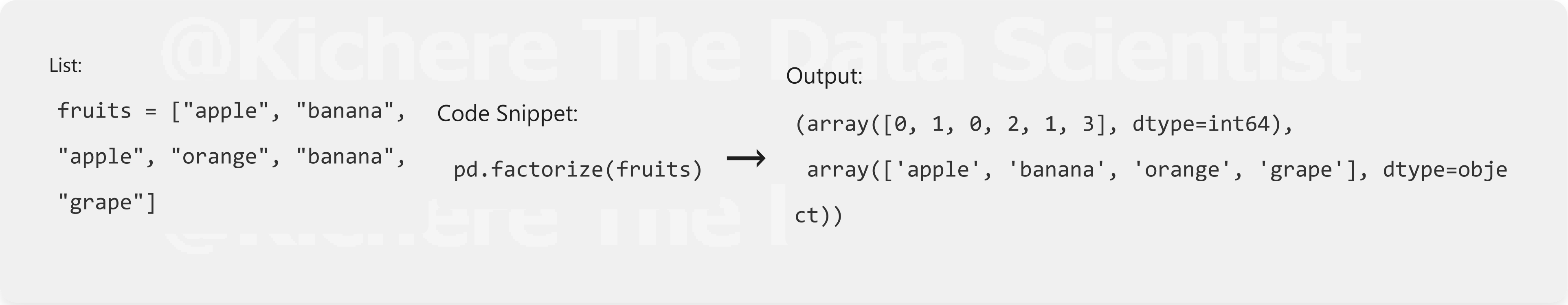
```
pandas.factorize(values, sort=False, use_na_sentinel=True, size_hint=None)
```

Parameters

The values parameter

The "values" parameter accepts a 1-D sequence.
Sequences that aren't pandas objects(e.g series) are to ndarrays before factorization.

```
Some examples of 1-D sequences:

List:
['red', 'blue', 'green', 'blue', 'green', 'red', 'yellow']

Numpy Array:
np.array(['red', 'blue', 'green', 'blue', 'green', 'red', 'yellow'])

Pandas Series:
pd.Series(['red', 'blue', 'green', 'blue', 'green', 'red', 'yellow'])

Tuples:
('red', 'blue', 'green', 'blue', 'green', 'red', 'yellow')

Range Object:
range(1, 8)
```



@Kichere The Data Scientist

List:

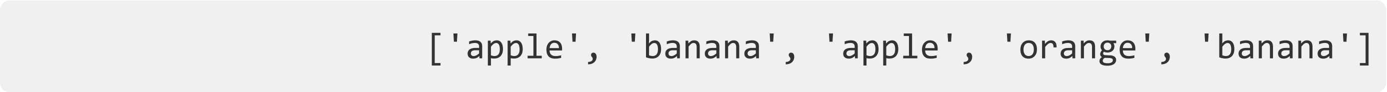
```
fruits = ["apple", "banana",
"apple", "orange", "banana",
"grape"]
```

Code Snippet:

```
pd.factorize(fruits)
```

Output:

```
(array([0, 1, 0, 2, 1, 3], dtype=int64),
 array(['apple', 'banana', 'orange', 'grape'], dtype=obje
ct))
```



```
['apple', 'banana', 'apple', 'orange', 'banana']
```

@Kichere The Data Scientist

```
Using uniques.take(codes) to
map codes back to original values
```

```
codes, uniques = pd.factorize(fru
its)
print(codes)
print(uniques)
```

↓

```
[0 1 2 1 0]
['Apple' 'Cherry' 'Banana']
```

```
uniques.take(codes)
```

↓

```
array(['apple', 'banana', 'apple', 'orange', 'banana'], dtype=o
bject)
```

The `sort` parameter

The "sort" parameter is a Boolean and defaults to False.
determines whether unique elements should be sorted in ascending order. When sort is set to True, the unique elements are sorted and any associated codes are shuffled to maintain their original relationships. If sort remains False, the elements and their codes retain their initial order without any sorting or shuffling.

```
fruits = ["Apple", "Cherry", "Banana", "Cherry", "Apple"]
```

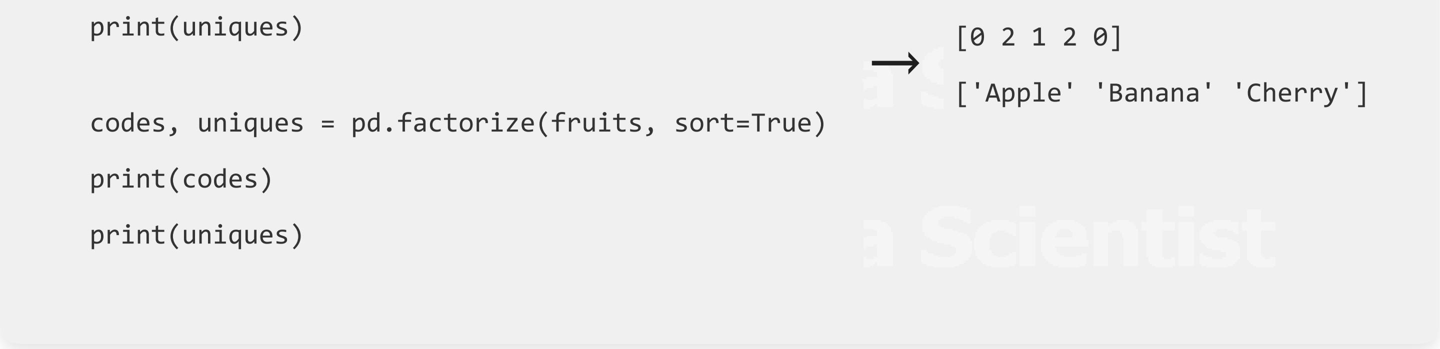
@Kichere The Data Scientist

Code Snippets:

```
codes, uniques = pd.factorize(fruits, sort=False)
print(codes)
```

Outputs:

```
[0 1 2 1 0]
['Apple' 'Cherry' 'Banana']
```



```
print(uniques)


codes, uniques = pd.factorize(fruits, sort=True)

print(codes)

print(uniques)
```

```
[0 2 1 2 0]

['Apple' 'Banana' 'Cherry']
```

In [19…

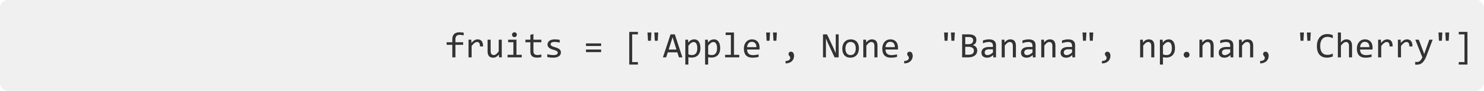
```python
fruits = ["Apple", "Cherry", "Banana", "Cherry", "Apple"]
# Using uniques.take(codes) to map codes back to original values
# Assume these are the unique values
uniques = pd.Index(['apple', 'banana', 'orange'])

# These are the integer codes corresponding to the unique values
codes = [0, 2, 1, 0, 1]
uniques.take(codes)
```

Out[196]:

```
Index(['apple', 'orange', 'banana', 'apple', 'banana'], dtype='object')
```

The use_na_sentinel parameter

The "use_na_sentinel" parameter is a Boolean and defaults to True.
If True, the sentinel -1 will be used for NaN values. If False, NaN values will be encoded as non-negative integers and will not drop the NaN from the uniques of the values.

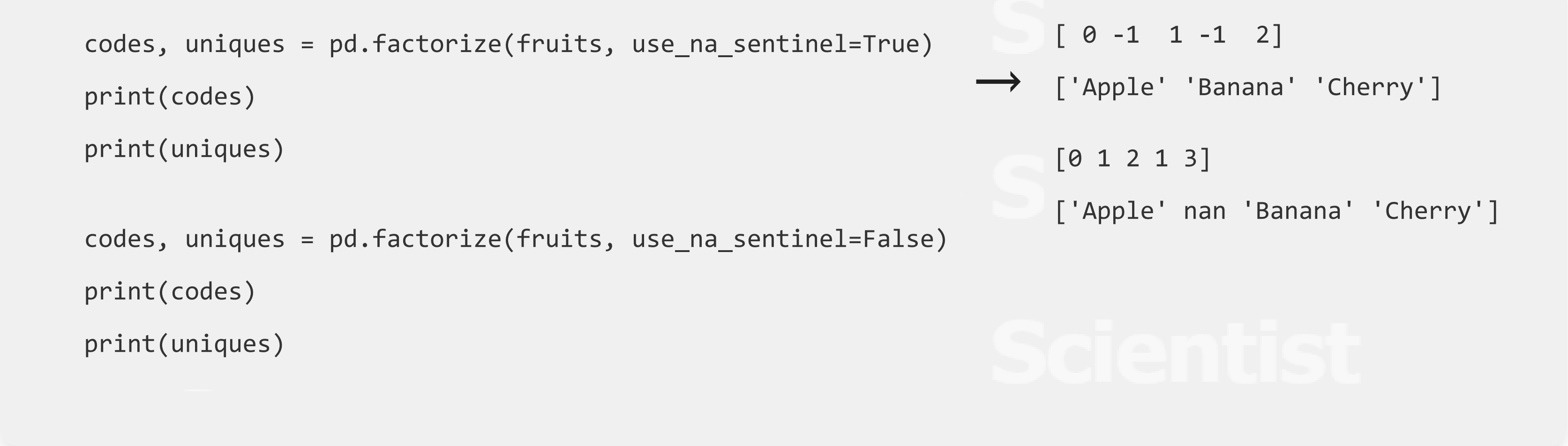


```
fruits = ["Apple", None, "Banana", np.nan, "Cherry"]
```

Code Snippets:

Outputs:

@Kichere The Data Scientist



```
codes, uniques = pd.factorize(fruits, use_na_sentinel=True)
print(codes)
print(uniques)

codes, uniques = pd.factorize(fruits, use_na_sentinel=False)
print(codes)
print(uniques)
```

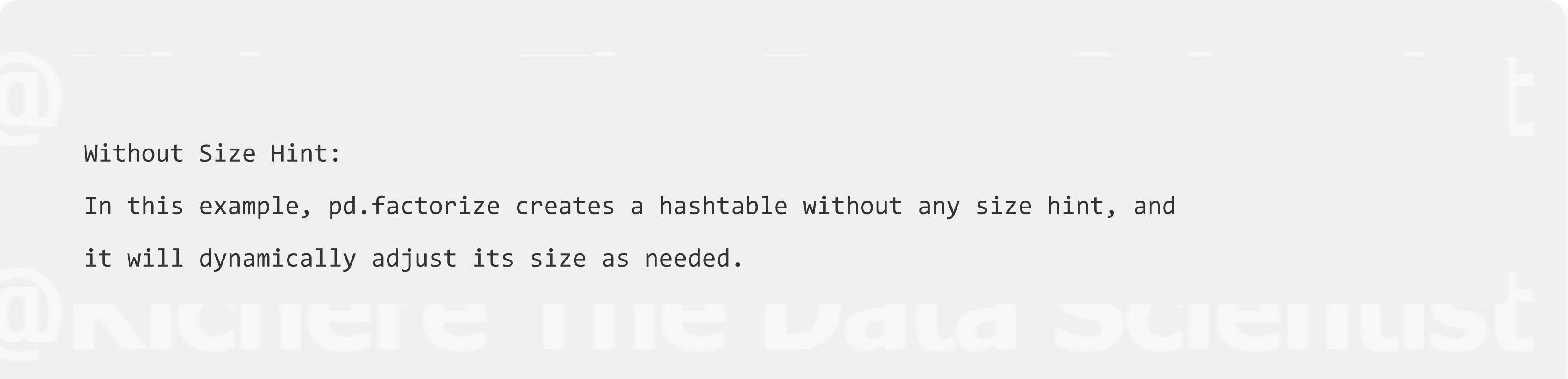
→

```
[ 0 -1  1 -1  2]
['Apple' 'Banana' 'Cherry']

[0 1 2 1 3]
['Apple' nan 'Banana' 'Cherry']
```

```
#sentinel: In programming, a sentinel value is a special value used to signal a
#particular condition or to mark boundaries. Here, the sentinel value is used to
#represent missing data (NaN). It typically means that this value will act as a
#marker or indicator in the data processing.
```

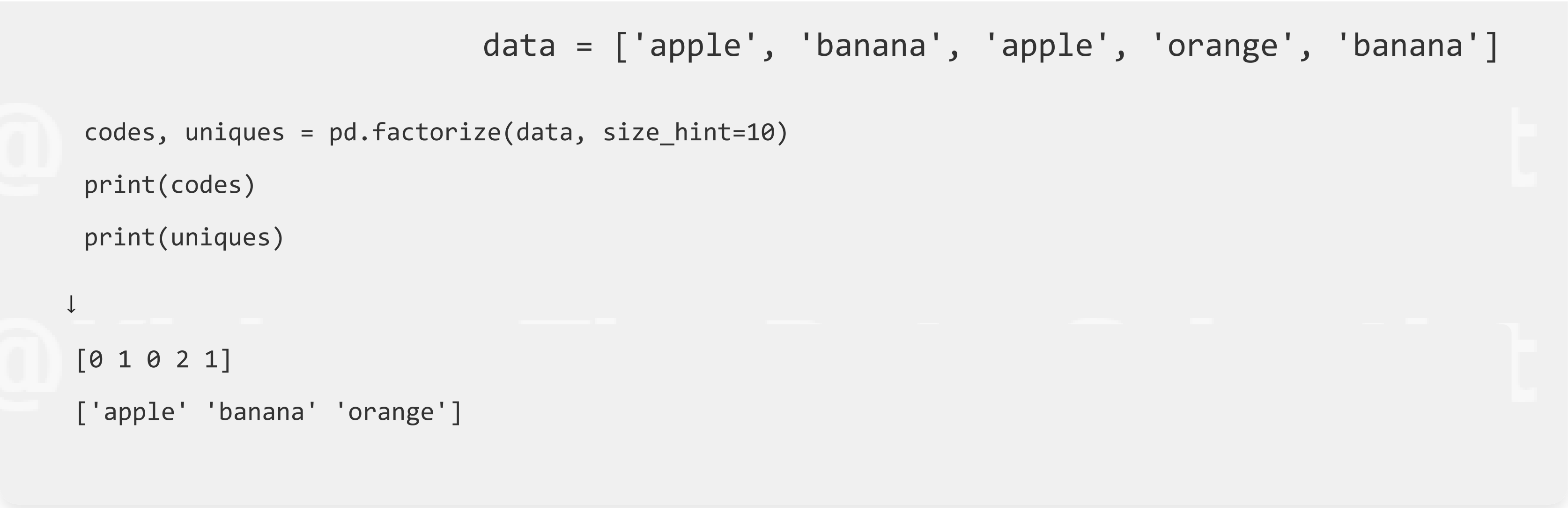
The size_hint parameter

The "size_hint" parameter is an integer and optional
size_hint is a way to help pd.factorize allocate enough space for its internal hashtable based on your estimate of the number of unique values.

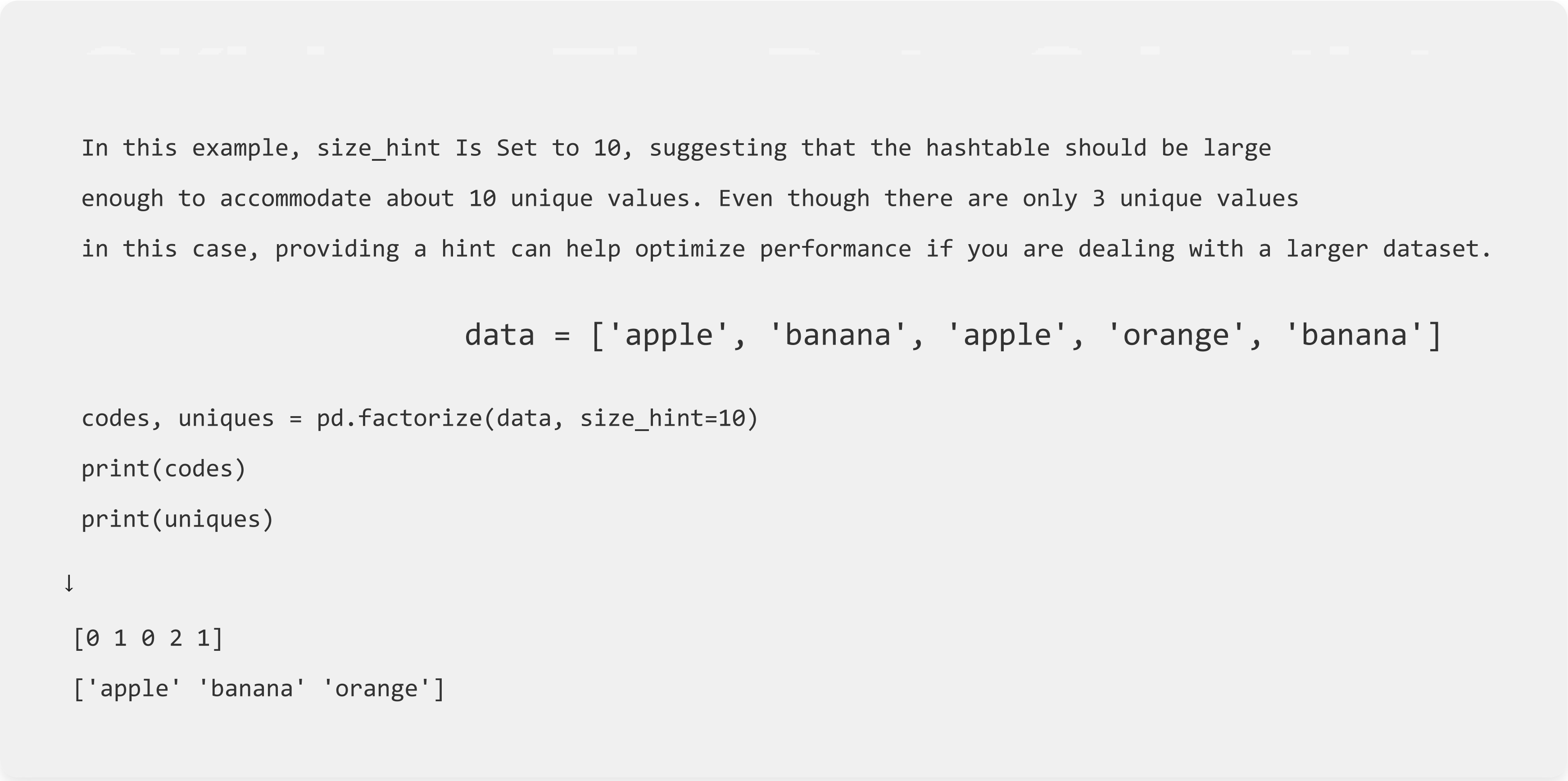


```
Without Size Hint:
In this example, pd.factorize creates a hashtable without any size hint, and
it will dynamically adjust its size as needed.
```

@Kiche The Data Scientist



```
data = ['apple', 'banana', 'apple', 'orange', 'banana']

codes, uniques = pd.factorize(data, size_hint=10)
print(codes)
print(uniques)
↓
[0 1 0 2 1]
['apple' 'banana' 'orange']
```



In this example, size_hint Is Set to 10, suggesting that the hashtable should be large enough to accommodate about 10 unique values. Even though there are only 3 unique values in this case, providing a hint can help optimize performance if you are dealing with a larger dataset.

```
data = ['apple', 'banana', 'apple', 'orange', 'banana']

codes, uniques = pd.factorize(data, size_hint=10)
print(codes)
print(uniques)
↓
[0 1 0 2 1]
['apple' 'banana' 'orange']
```

What Makes This Ebook Unique

> "The aim of this ebook is to give you the 'aha' moment right away at the start of learnin

- Practical step By Step Guide With Simple Examples
- Visual Illustrations and Interactive
- Simple Datasets
- covers everything about "pandas.factorize", as I referenced the pandas documentation to write it.

****

Sources & References

pandas.factorize Documentation (https://pandas.pydata.org/docs/reference/api/pandas.pivot_table.html)

Pandas Series: factorize() function (https://www.w3resource.com/pandas/series/series-factorize.php)

Contacts and Social Media

Kichere Magubu

Dar es salaam, Tanzania

Kichere Magubu

Kichere The Data Scientist

KichereTheDataScientist

KichereTheDataScientist

kicherethedatascientist@gmail.com

kicherethedatascientist

☎ +255 654 729 851

Powered by
Eastern Africa Statistical Training Centre
SKT Tanzania Ltd

****

```
In [ ]: #print("The cell to convert jupyter notebook to html")
        !jupyter nbconvert --to hide_code_html "pandas.factorize.ipynb"
```

```
In [ ]:
```