Pantech e Learning
DIGITAL LEARNING SIMPLIFIED

Amazon Web Services

MLOps with AWS

Masterclass

aws

Machine Learning Operations with AWS

Day -6

Amazon Web Services

MLOps with AWS

Masterclass

aws

Pantech e Learning
DIGITAL LEARNING SIMPLIFIED

Pandas

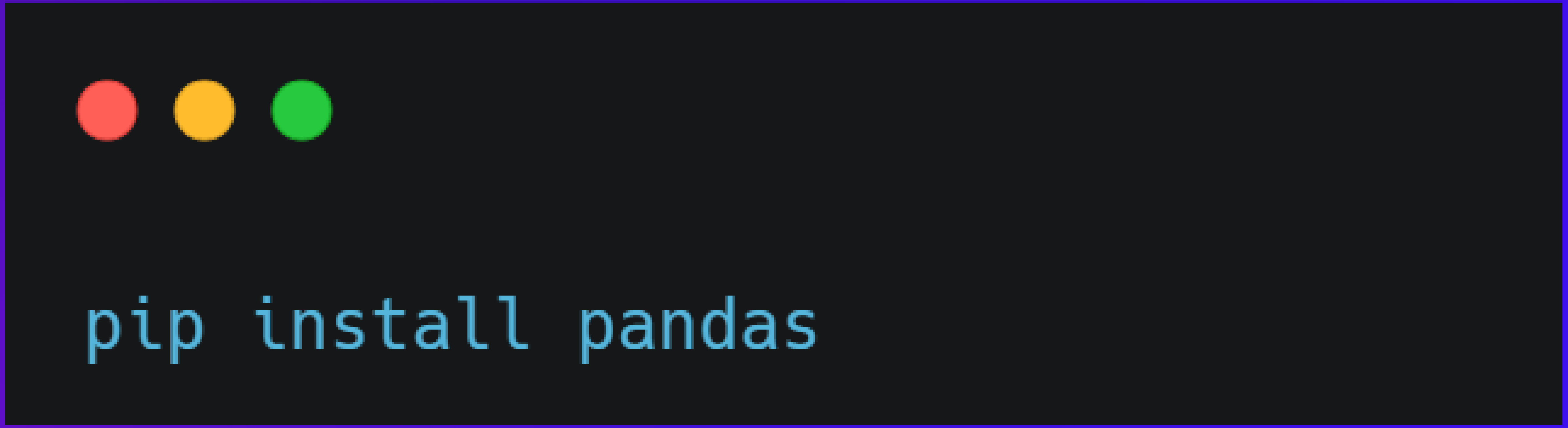
Pandas

- Pandas is a Python library used for working with data sets
- It has functions for analyzing, cleaning, exploring, and manipulating data
- Pandas is fast and it has high performance & productivity for users

Why Pandas ?

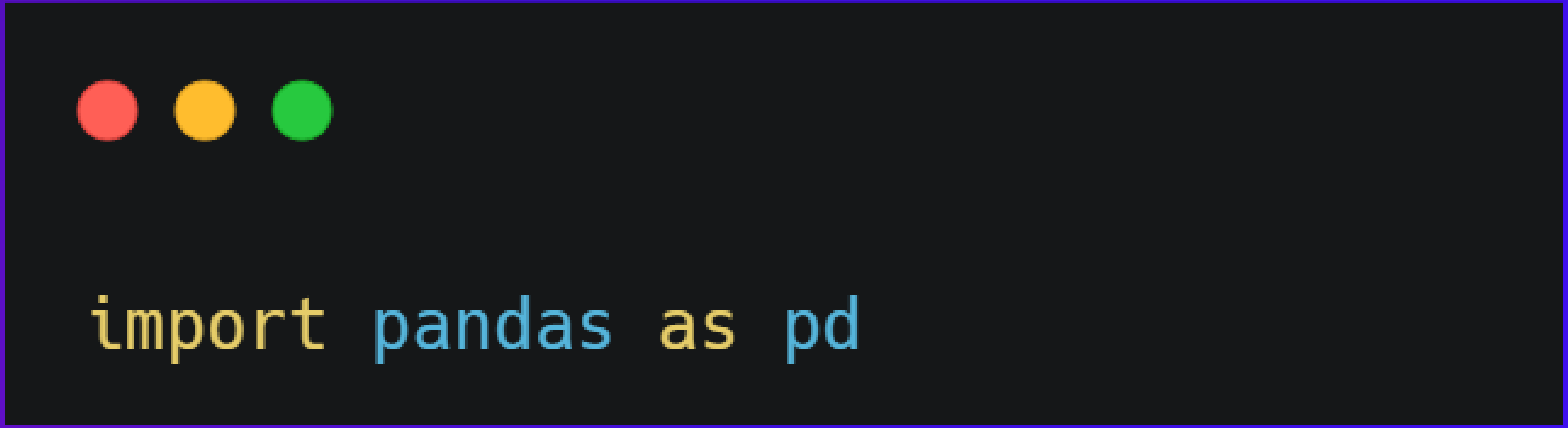
- Pandas allows us to analyze big data and make conclusions based on statistical theories
- Pandas can clean messy data sets, and make them readable and relevant
- Relevant data is very important in Machine learning

Installation



```
pip install pandas
```

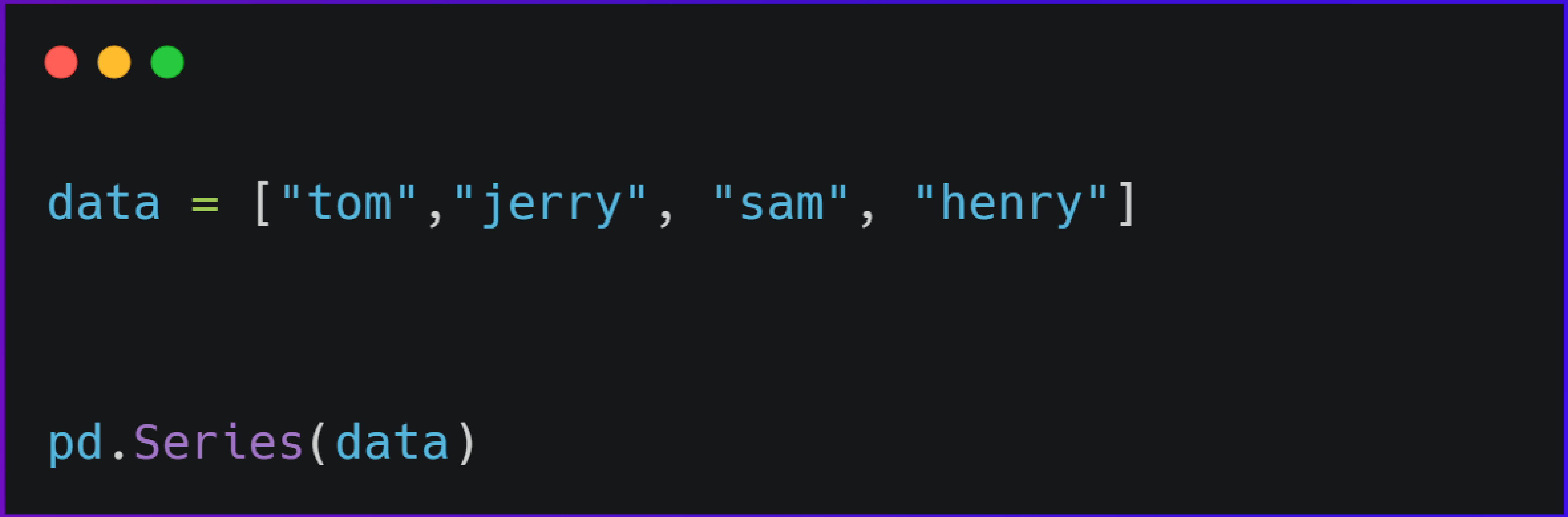
Import



```
import pandas as pd
```

Pandas Series

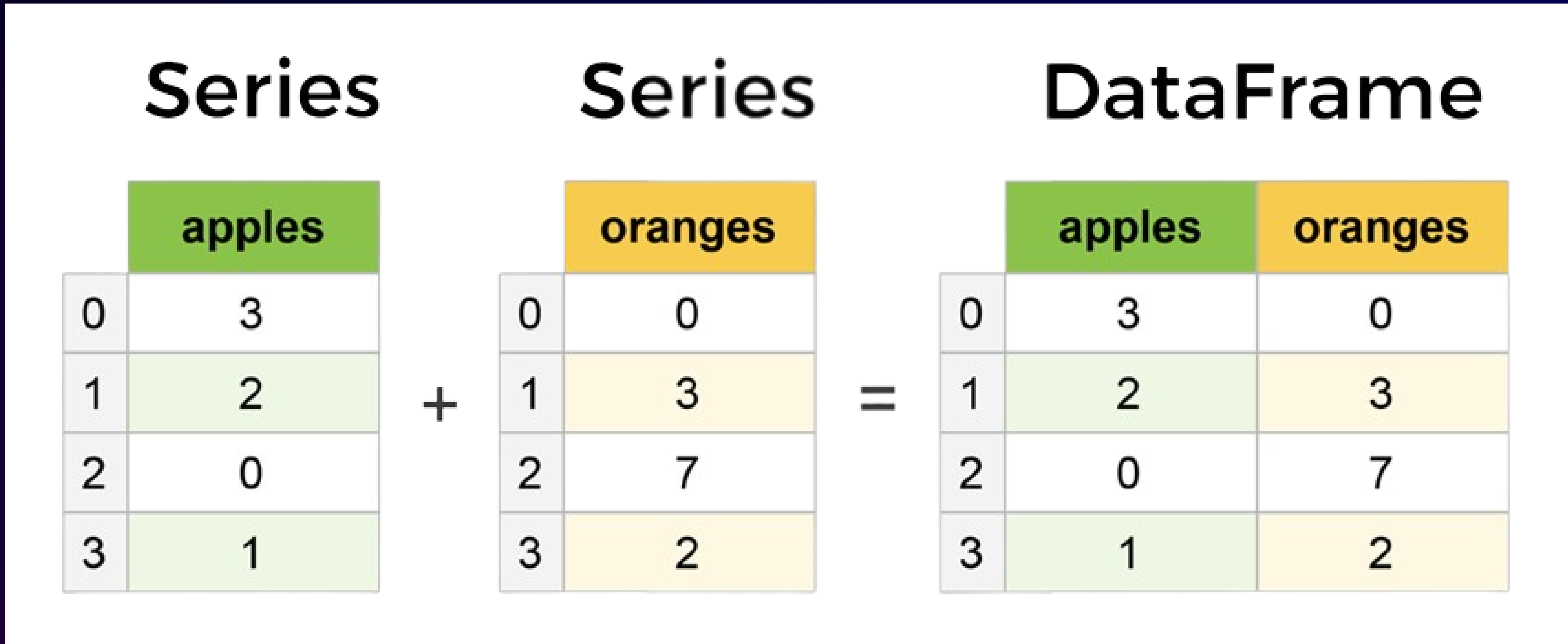
- A Pandas Series is like a column in a table
- It is a one-dimensional array holding data of any type



```
data = ["tom","jerry", "sam", "henry"]


pd.Series(data)
```

Pandas Dataframe

- Dataframe is a 2D array-like object that can hold any data type
- It is similar to a table with rows and columns



Series

	apples
0	3
1	2
2	0
3	1

+

Series

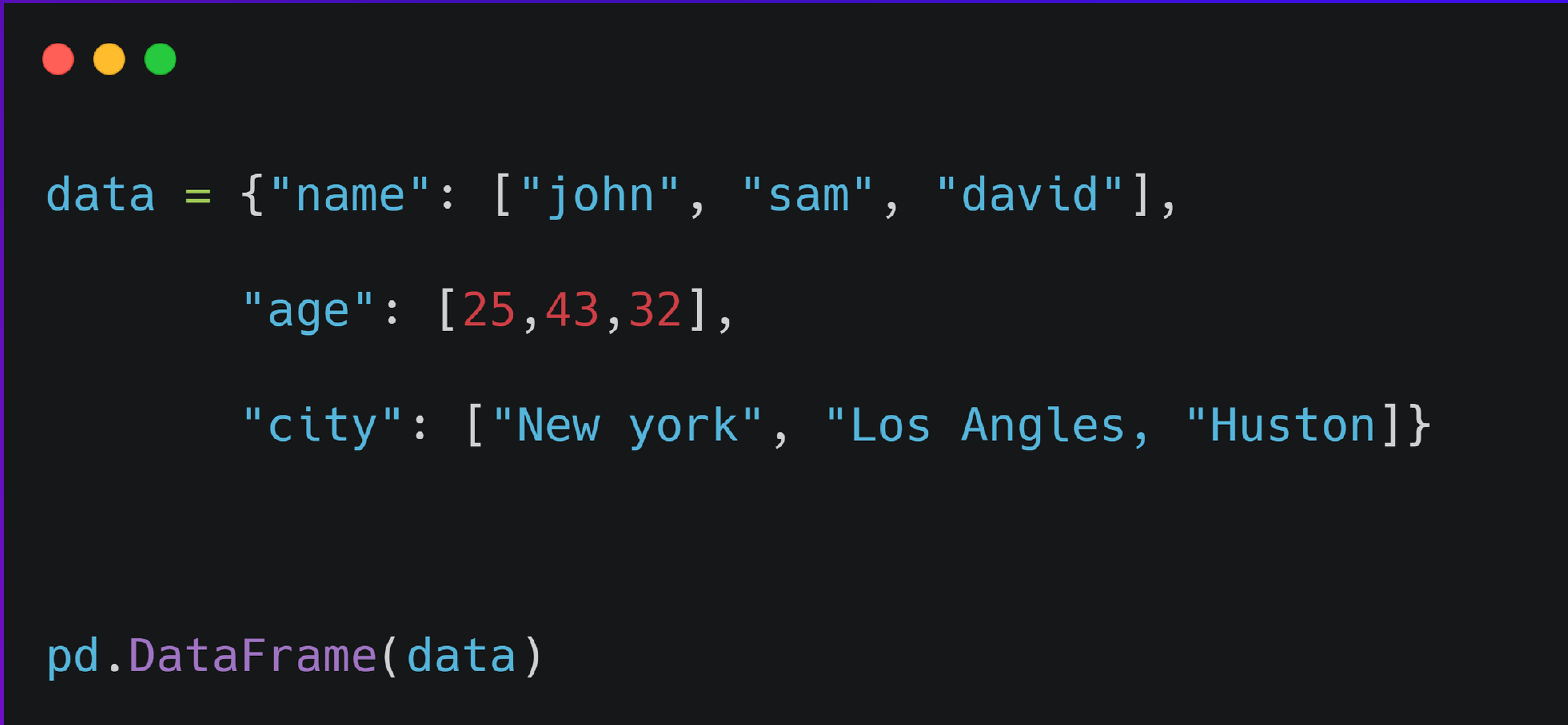
	oranges
0	0
1	3
2	7
3	2

=

DataFrame

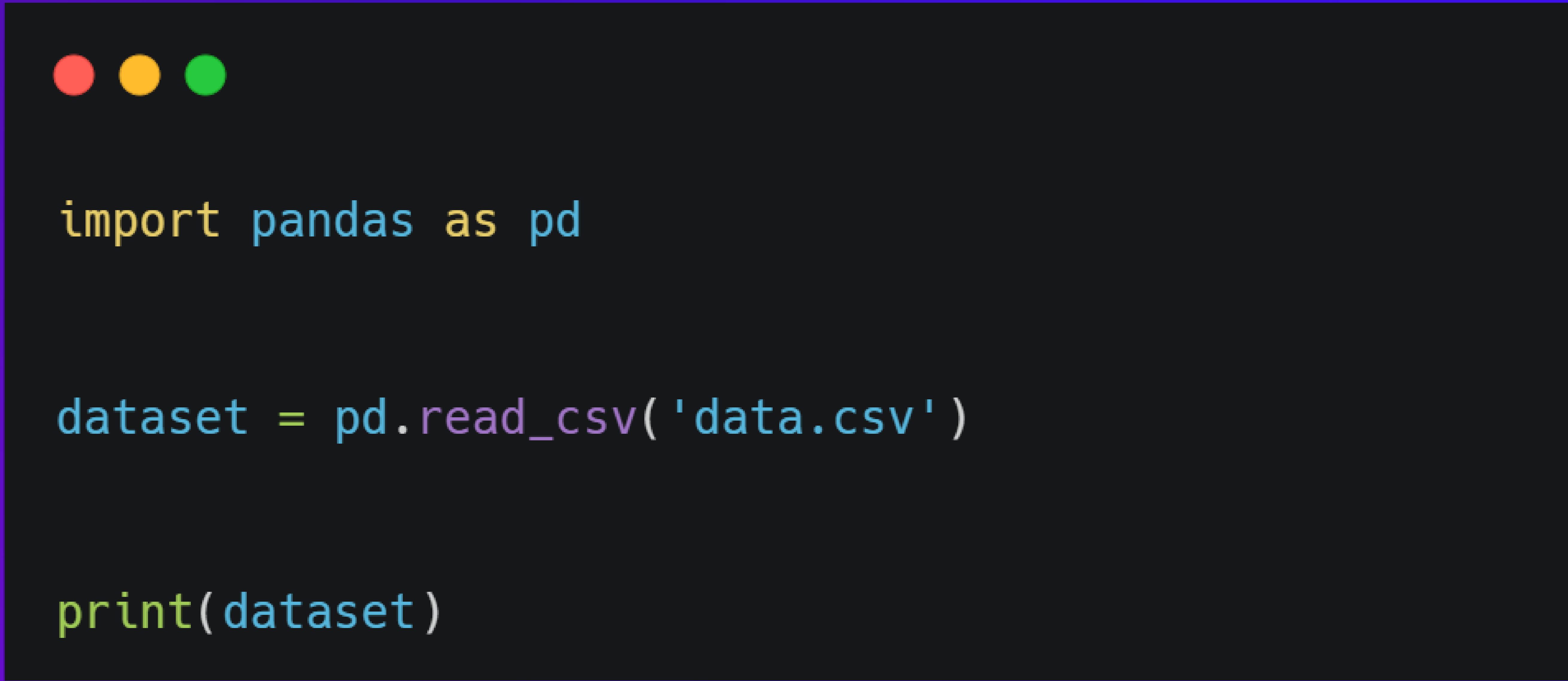
	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Pandas Dataframe



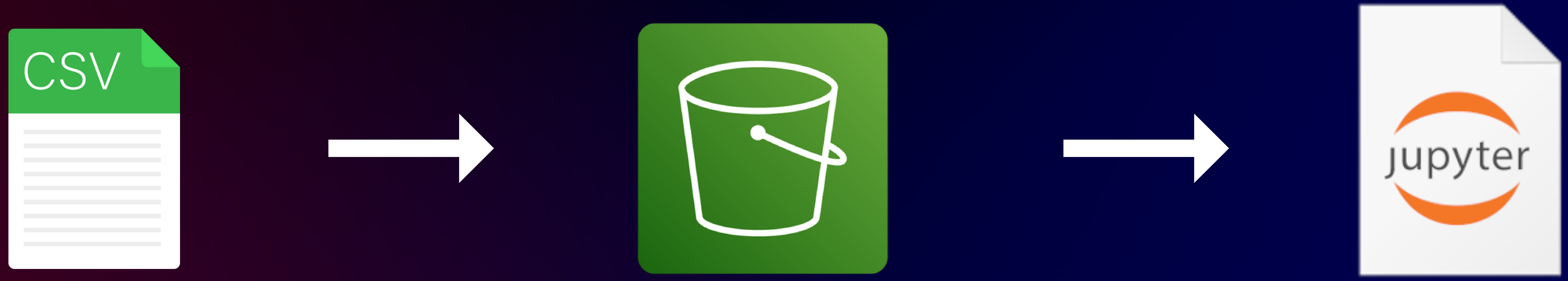
```
data = {"name": ["john", "sam", "david"],
       "age": [25,43,32],
       "city": ["New york", "Los Angles, "Huston]}

pd.DataFrame(data)
```

Read CSV

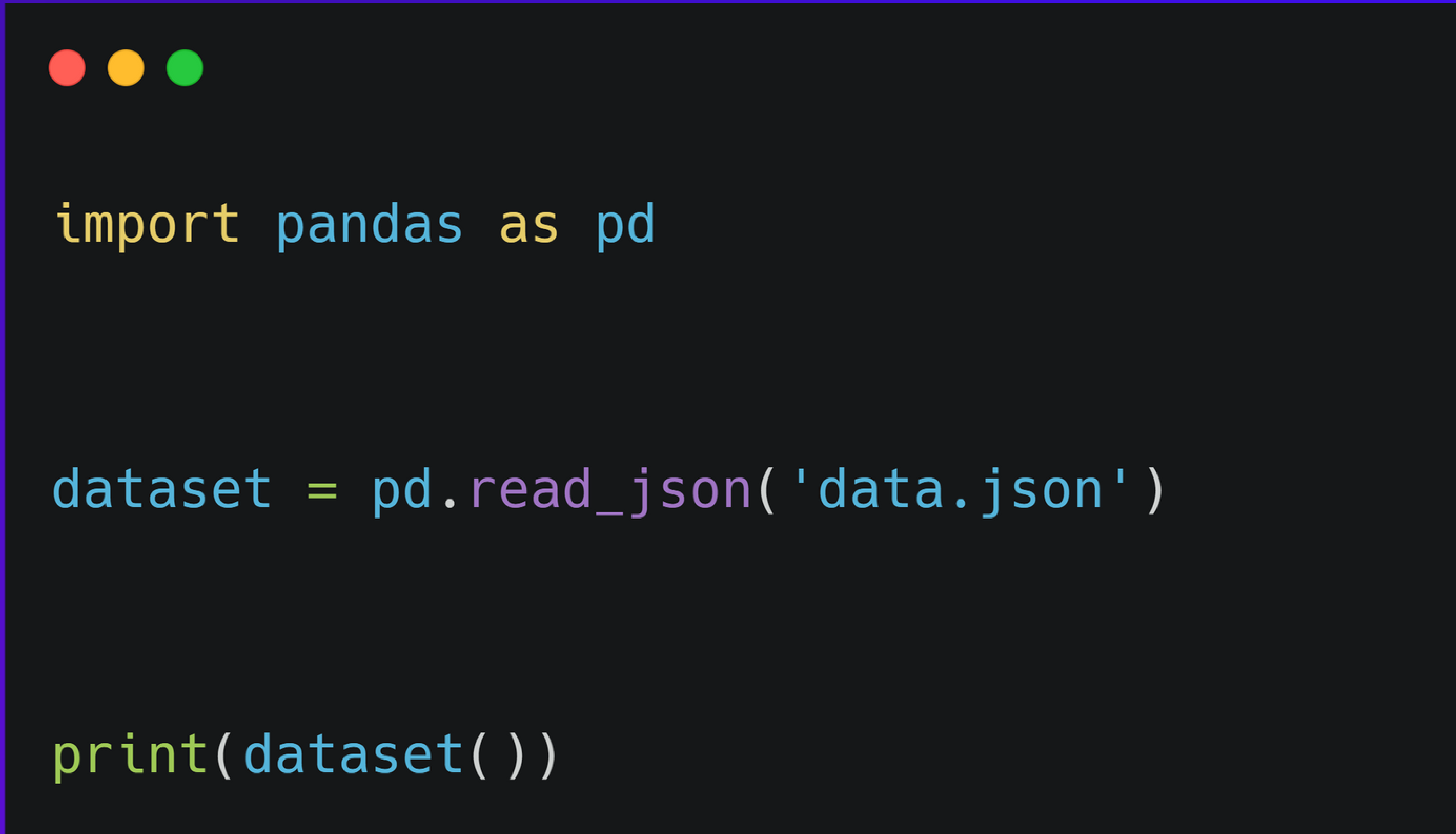


```
import pandas as pd

dataset = pd.read_csv('data.csv')

print(dataset)
```

Save the dataset file in S3 Bucket

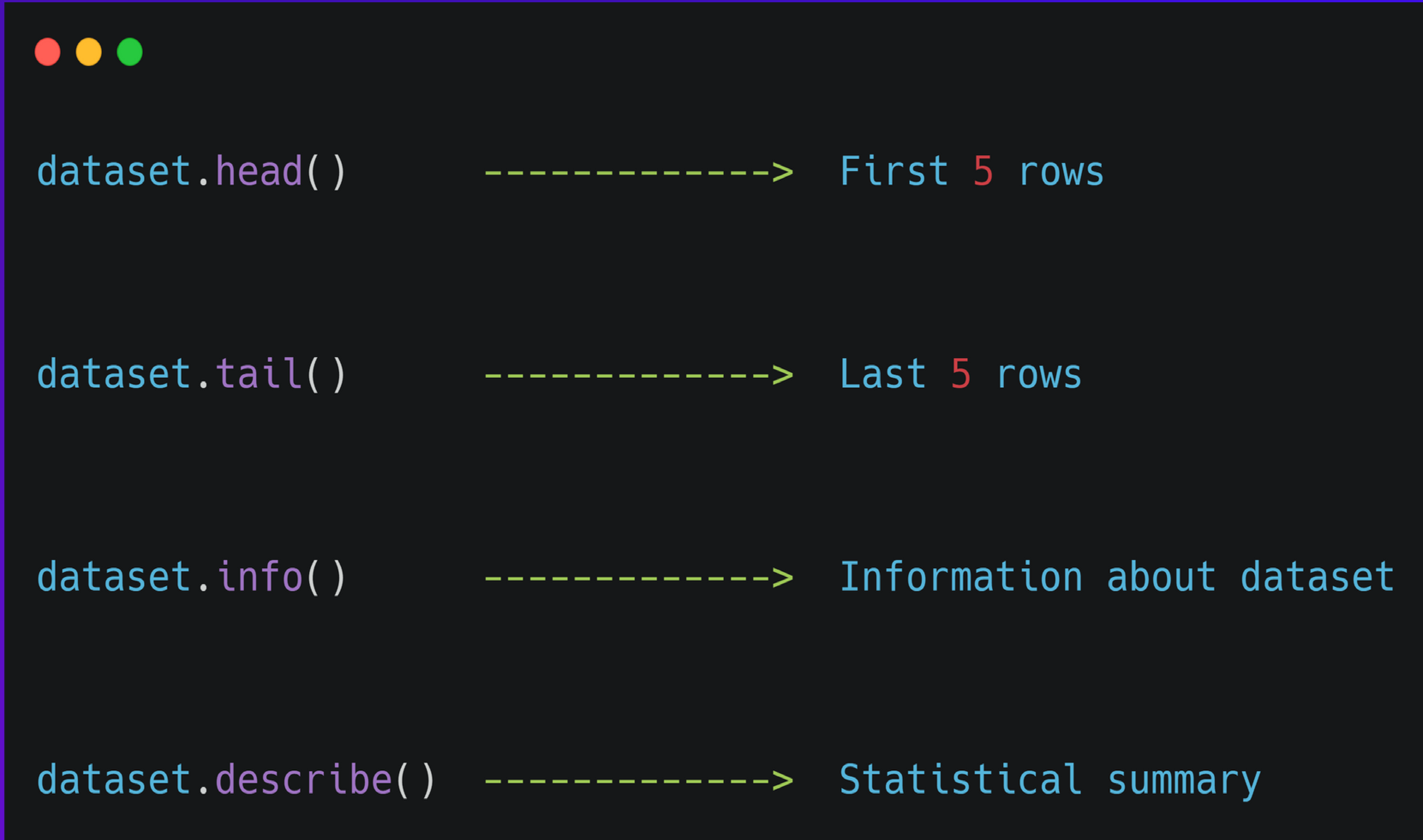


Read Json



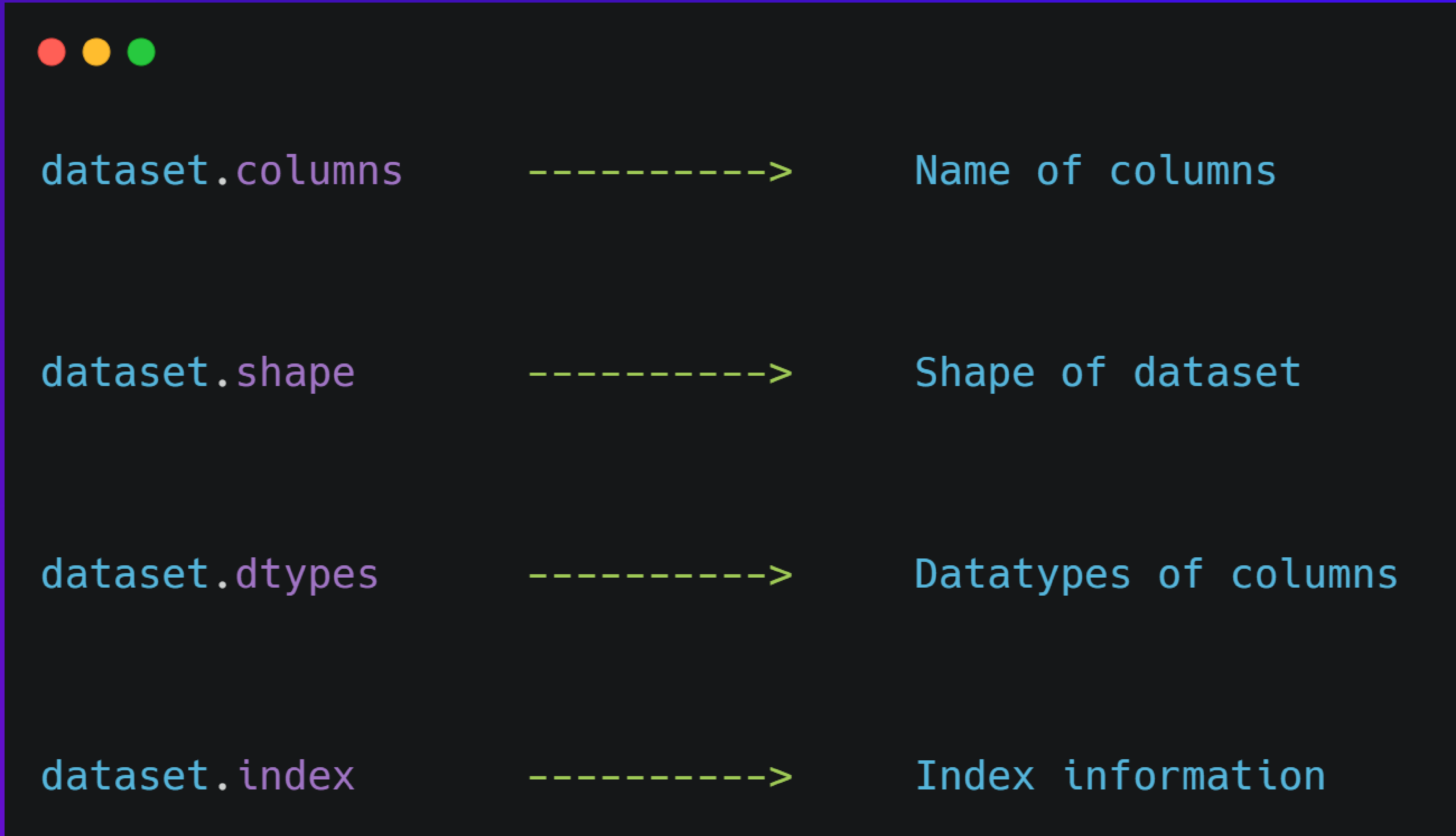
```
import pandas as pd

dataset = pd.read_json('data.json')

print(dataset())
```

Analysing Data



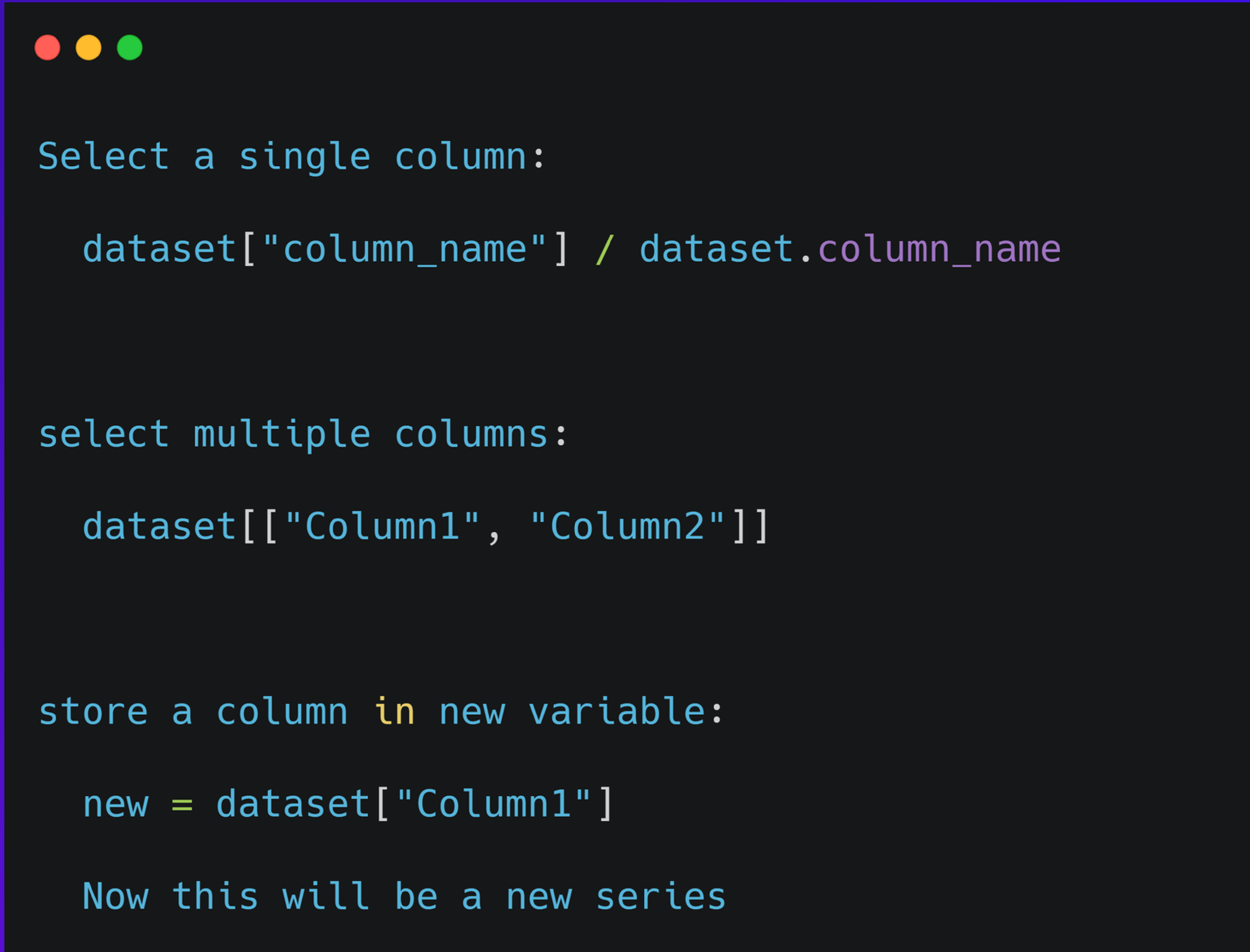
```
dataset.head( )       -------------->  First 5 rows


dataset.tail( )       -------------->  Last 5 rows


dataset.info( )       -------------->  Information about dataset


dataset.describe( )   -------------->  Statistical summary
```

Analysing Data



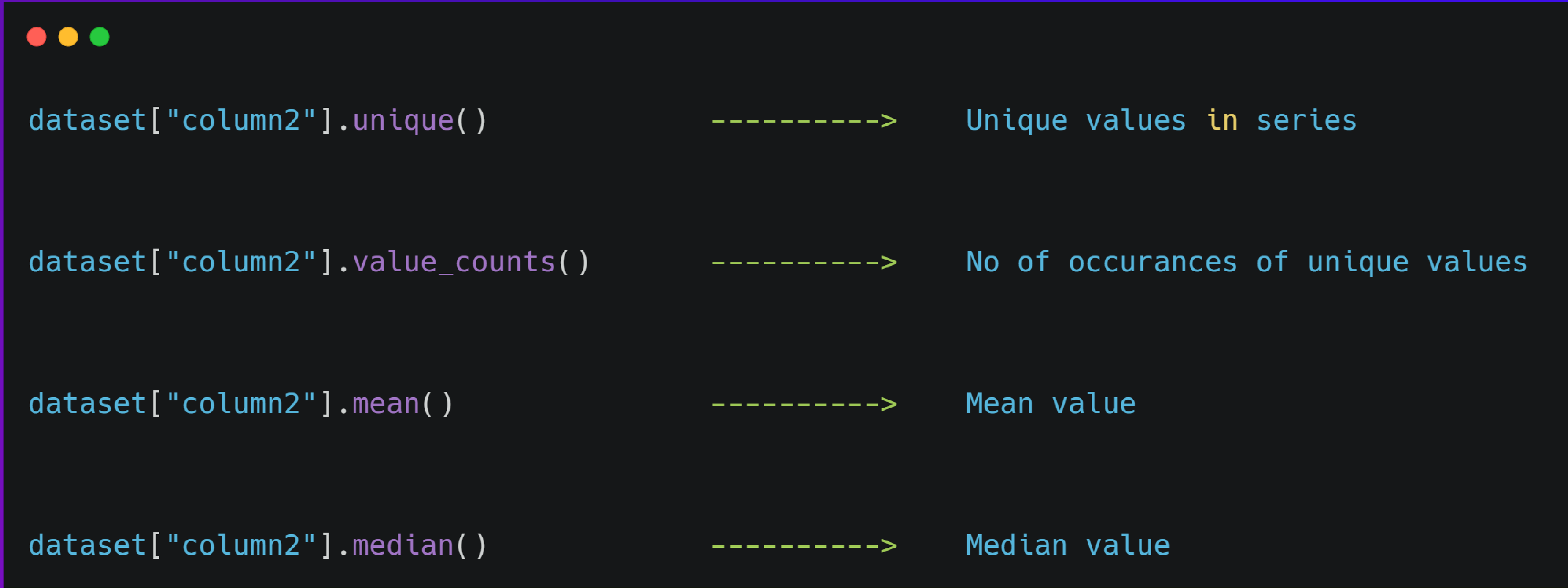
```
dataset.columns     ----------->     Name of columns


dataset.shape       ----------->     Shape of dataset


dataset.dtypes      ----------->     Datatypes of columns


dataset.index       ----------->     Index information
```

Analysing Data



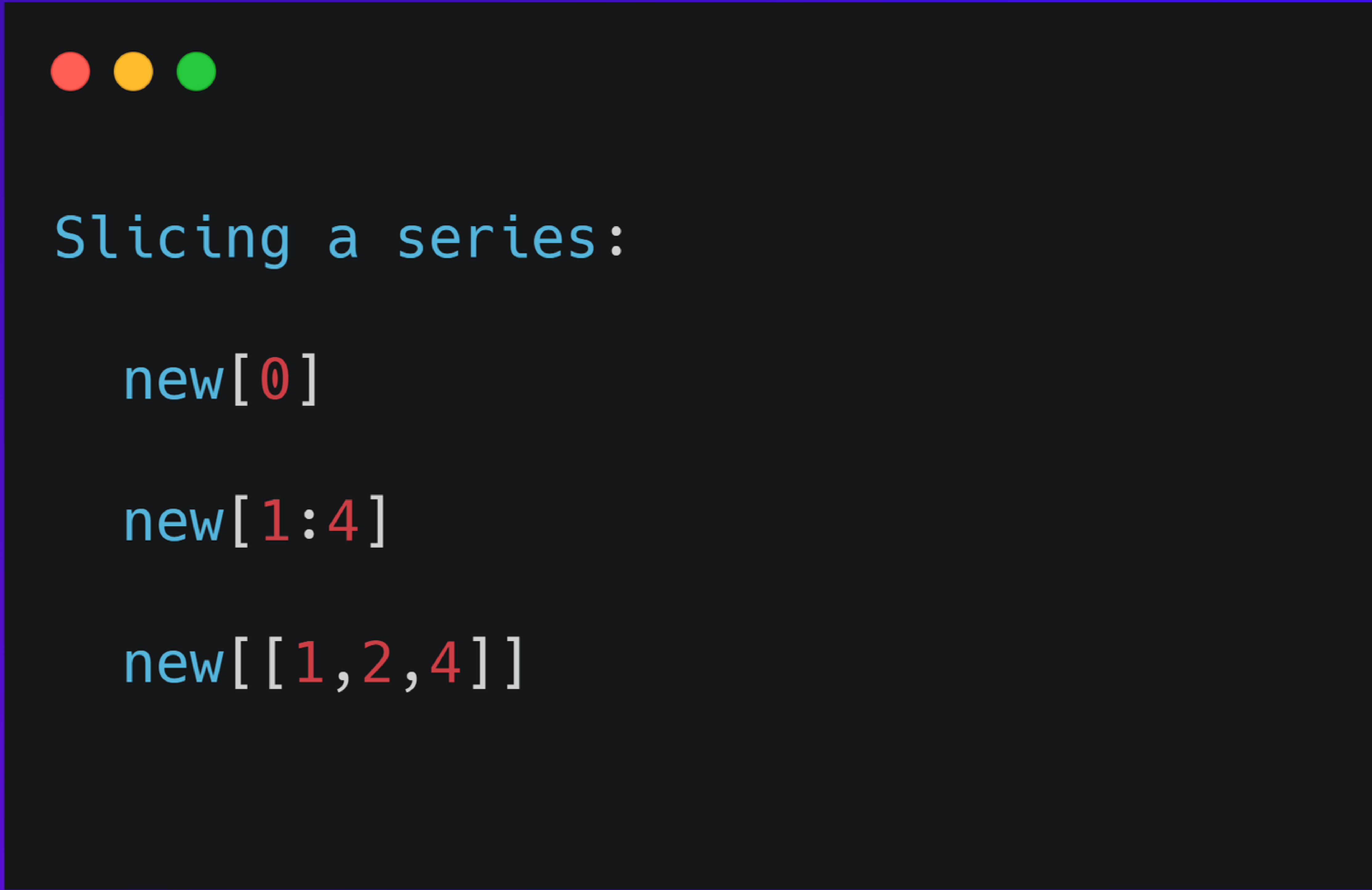
```
Select a single column:

  dataset["column_name"] / dataset.column_name


select multiple columns:

  dataset[["Column1", "Column2"]]


store a column in new variable:

  new = dataset["Column1"]

  Now this will be a new series
```

Analysing Data



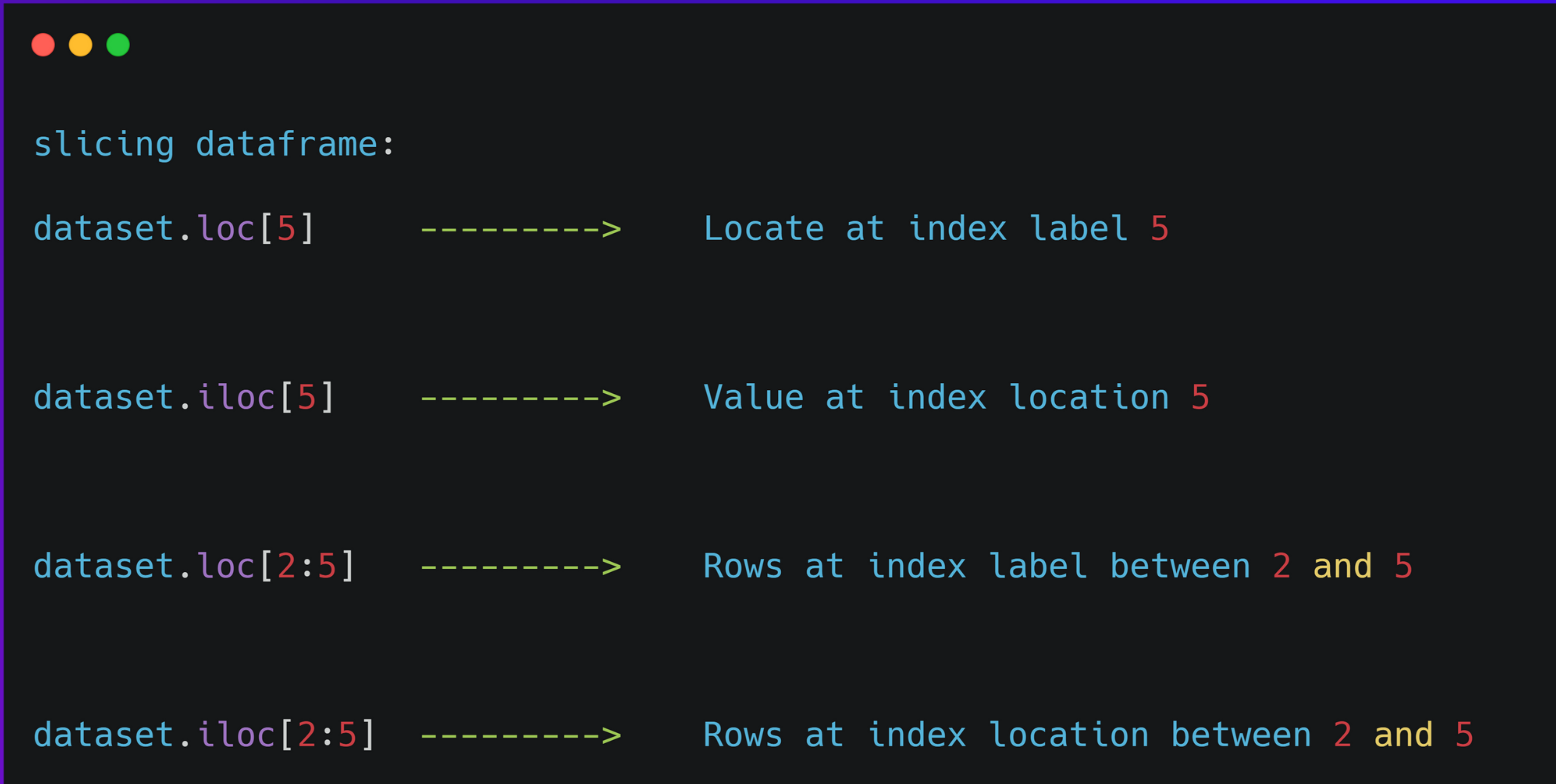
```
dataset["column2"].unique()             ---------->    Unique values in series

dataset["column2"].value_counts()       ---------->    No of occurances of unique values

dataset["column2"].mean()               ---------->    Mean value

dataset["column2"].median()             ---------->    Median value
```

Analysing Data



```
Slicing a series:

  new[0]

  new[1:4]

  new[[1,2,4]]
```

Analysing Data



```
slicing dataframe:

dataset.loc[5]     --------->    Locate at index label 5


dataset.iloc[5]    --------->    Value at index location 5


dataset.loc[2:5]   --------->    Rows at index label between 2 and 5


dataset.iloc[2:5]  --------->    Rows at index location between 2 and 5
```

Thank you

Pantech e Learning
DIGITAL LEARNING SIMPLIFIED