DataFrames

- DataFrames are a container for your data
- Rows are individual events
- Columns are features and target values
- The index is the row's unique ID, making it easier to find a row individually

```
# Creating a simple DataFrame from a dictionary of values
import pandas as pd

df = pd.DataFrame(
  {
    'a': [1, 2, 3, 4, 5],
    'b': ['a', 'b', 'c', 'd', 'f'],
    'c': [5, 4, 3, 2, 1],
  }
)
```

`df.describe()` provides summary statistics

For numerical values, it provides data for each feature:

- count - number of rows
- mean - average value
- std - standard deviation
- min - minimum value
- 25/50/75% - percentile values
- max - maximum value

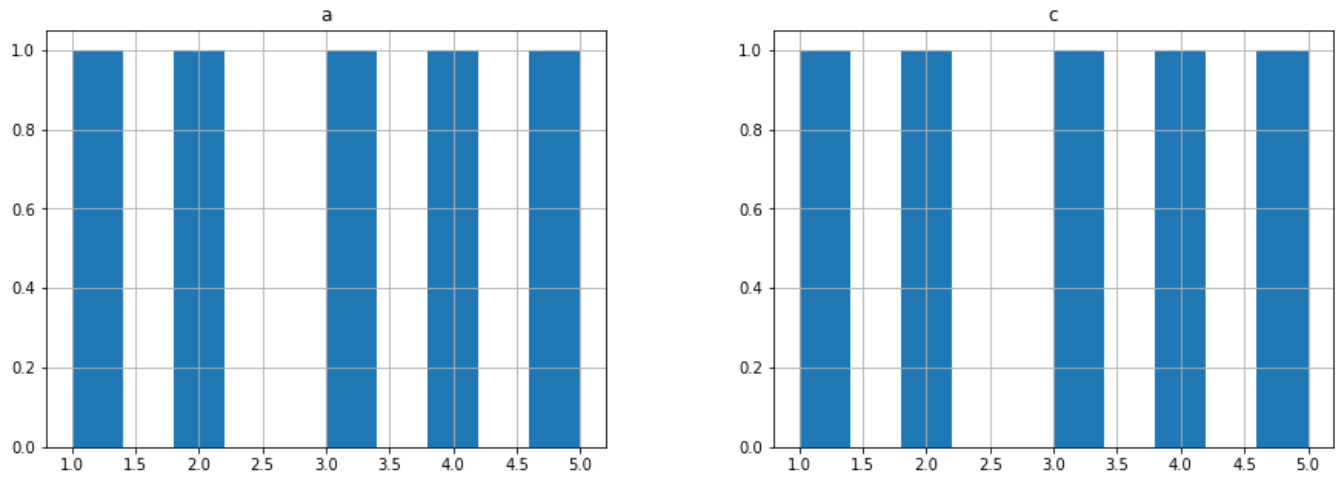
```
print(df.describe())

              a         c
count  5.000000  5.000000
mean   3.000000  3.000000
std    1.581139  1.581139
min    1.000000  1.000000
25%    2.000000  2.000000
50%    3.000000  3.000000
75%    4.000000  4.000000
max    5.000000  5.000000
```

`df.hist()` plots histograms

- Plot the distribution of values
- Help identify outliers or unexpected values
- Normal, Uniform, chi-square, F are examples of different distributions

```
df.hist()
```



Histogram of a and c values

`df.corr()` calculates a correlation matrix

- Correlation is the relationship between two variables
- Values from correlation range from -1.0 to 1.0
- -1.0 is a strong negative relationship

- 1.0 is a strong positive relationship
- Diagonal matrix values always 1.0
- Strong positive or negative relationships add little to ML models

```
print(df.corr())
"""
     a    c
a  1.0 -1.0
c -1.0  1.0
"""
```