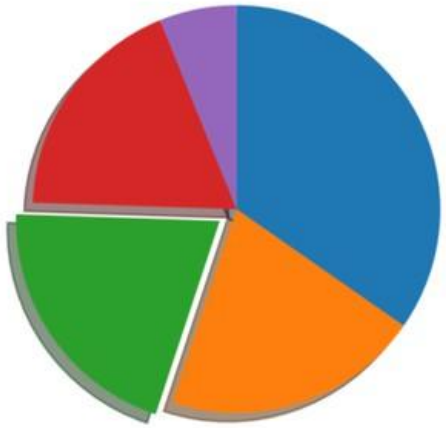
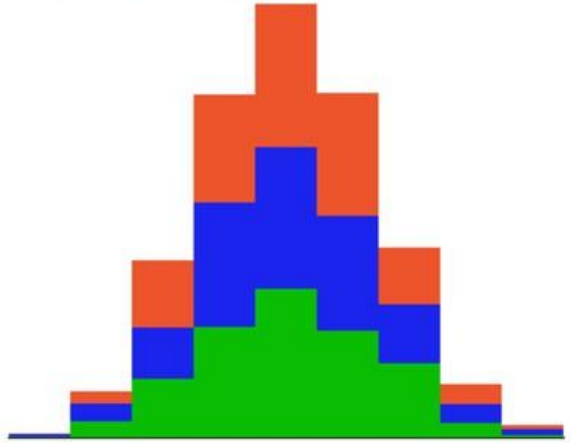
# DATA SCIENCE TOOLS, PART 2: PANDAS, MATPLOTLIB

Pandas

matplotlib





# PRE-QUIZ: HOW MUCH DO YOU ALREADY KNOW ABOUT PANDAS?

- What is python pandas?
- How is a pandas DataFrame different than a NumPy array?
- Name three unique operations (i.e., methods) you can do with pandas?
- How do you read from or write to a csv file using pandas?
- Keep this Quiz and see if you can fill-in any missing questions during the discussion today

# WHAT IS PANDAS?

- Definition
- Series
- DataFrame
- Conditional Selection
- Iris Flower Data Set

BYU

# PANDAS DEFINITION

- Open-Source software library written for Python
- Pandas derived from the term “**pan**el **da**ta” from econometrics
- Data structures and operations for manipulating numerical tables and time series
- DataFrame patterned after R DataFrame
- DataFrame is a 2-Dimensional structure built as a combination of Series arrays with a shared index
- Built on NumPy
- Originally released 11 Jan 2008.

https://en.wikipedia.org/wiki/Pandas_(software)

BYU

# PANDAS SERIES

- A **pandas Series** is a one-dimensional labeled array in Python, capable of holding data of any type (integers, floats, strings, Python objects, etc.).

Key features of a pandas Series:

• **Indexing**: Each element in the Series has a corresponding index, which allows for easy access and manipulation of data.

• **Homogeneous**: The Series can hold elements of the same type (though mixed types are possible, but uncommon).

• **Data alignment**: The index labels allow for automatic alignment of data when performing operations, which can be useful when dealing with time series or other labeled data.

python

```python
import pandas as pd

# Create a pandas Series
data = pd.Series([10, 20, 30, 40], index=['a', 'b', 'c', 'd'])

print(data)
```

This will output:

```
a    10
b    20
c    30
d    40
dtype: int64
```

BYU

# PANDAS DATAFRAME

- A **Pandas DataFrame** is a two-dimensional, labeled data structure in Python, similar to a table or spreadsheet, that stores data in rows and columns. Each column in a DataFrame can have a different data type (e.g., integers, floats, strings, etc.).

**Key features of a DataFrame:**

- **Rows and Columns:** Like a table, with rows representing individual records and columns representing variables or features.
- **Labels (Index):** Rows and columns can have labels (names), making it easy to access, slice, or manipulate data.
- **Data Handling:** It can handle missing data and supports arithmetic operations, data filtering, aggregation, and transformation.
- **Data Input/Output:** Can read and write data from various file formats (e.g., CSV, Excel, SQL, etc.).
- Multiple Series Objects that share an index

**Example:**

python

```python
import pandas as pd

# Create a DataFrame
data = {
    'Name': ['Alice', 'Bob', 'Charlie'],
    'Age': [25, 30, 35],
    'City': ['New York', 'San Francisco', 'Los Angeles']
}

df = pd.DataFrame(data)

print(df)
```

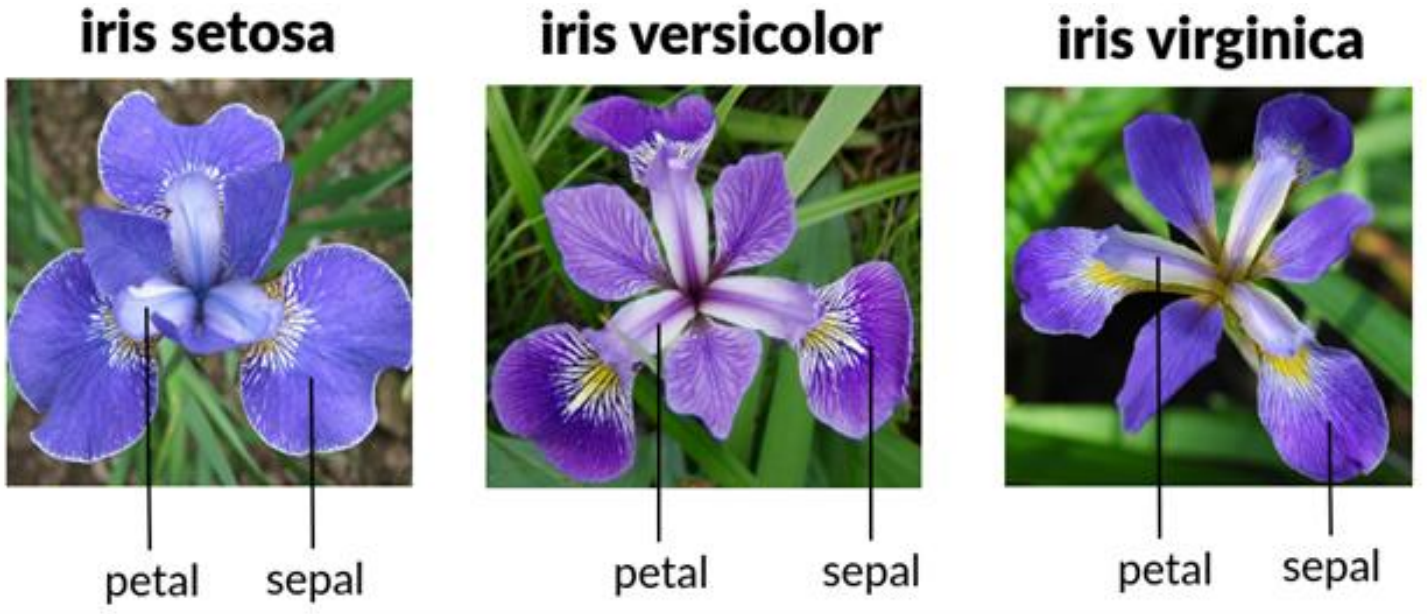
This code would output:

```
      Name  Age           City
0    Alice   25       New York
1      Bob   30  San Francisco
2  Charlie   35    Los Angeles
```

BYU

# IRIS FLOWER DATA SET

- The ***Iris* flower data set** was made famous by the British statistician and biologist Ronald Fisher in 1936.
- It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.
- The dataset contains a set of 150 records with five attributes: sepal length, sepal width, petal length, petal width and species.
- The iris data set is widely used for machine learning purposes. The dataset is included in Python in the machine learning library scikit-learn.




```
from sklearn.datasets import load_iris

iris = load_iris()
iris
```

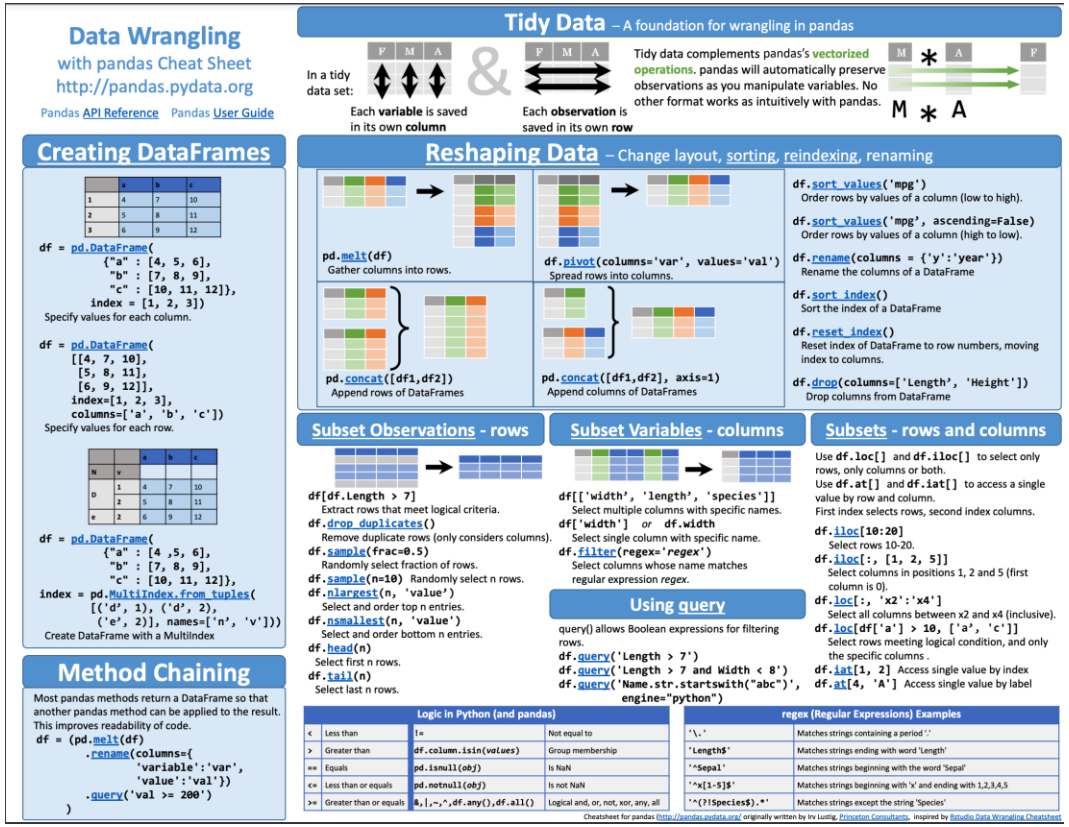
This code gives:

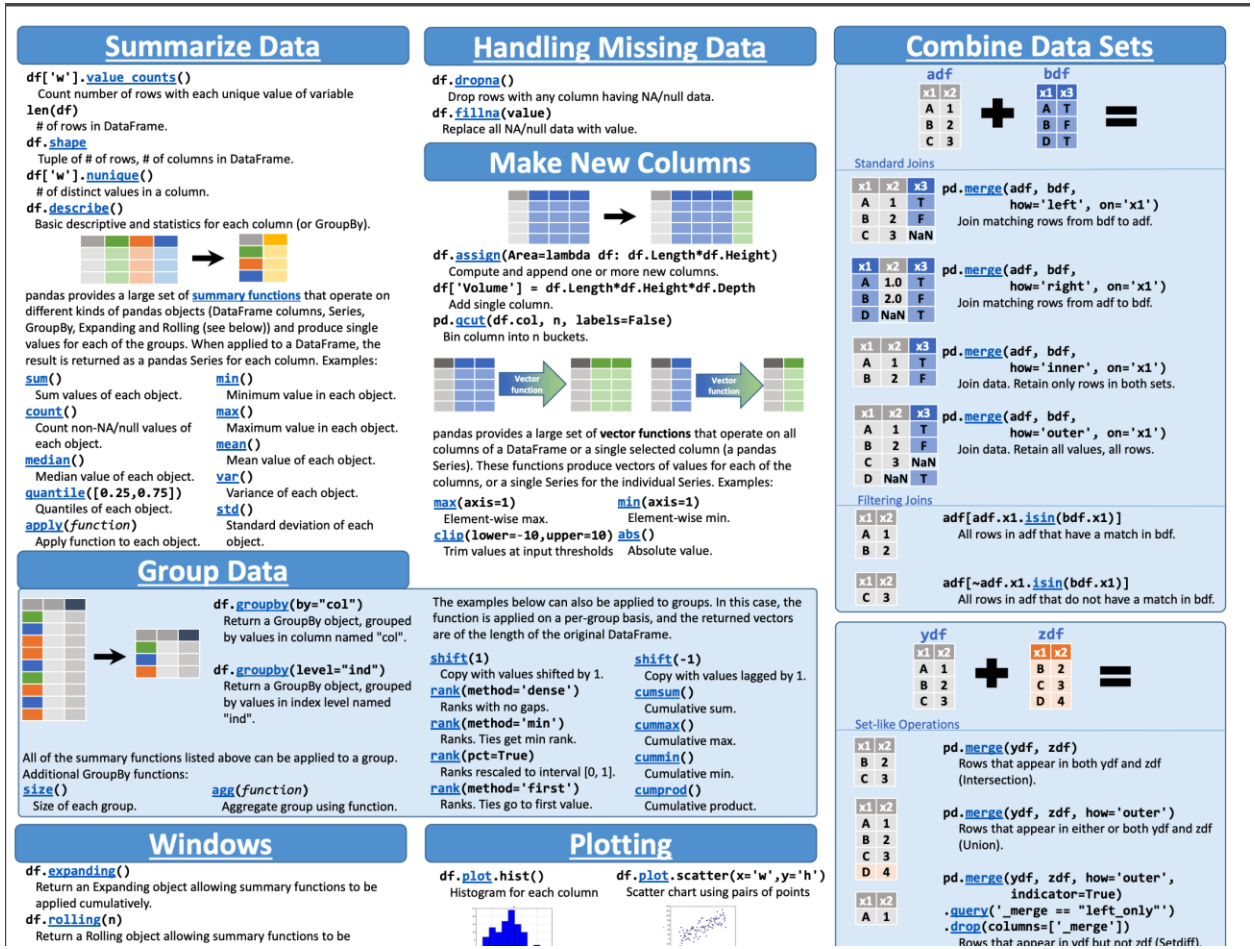
```
{'data': array([[5.1, 3.5, 1.4, 0.2],
                [4.9, 3., 1.4, 0.2],
                [4.7, 3.2, 1.3, 0.2],
                [4.6, 3.1, 1.5, 0.2],...
'target': array([0, 0, 0, ... 1, 1, 1, ... 2, 2, 2, ...
'target_names': array(['setosa', 'versicolor', 'virginica'],
dtype='<U10'),
...}
```

https://en.wikipedia.org/wiki/Iris_flower_data_set#External_links

BYU

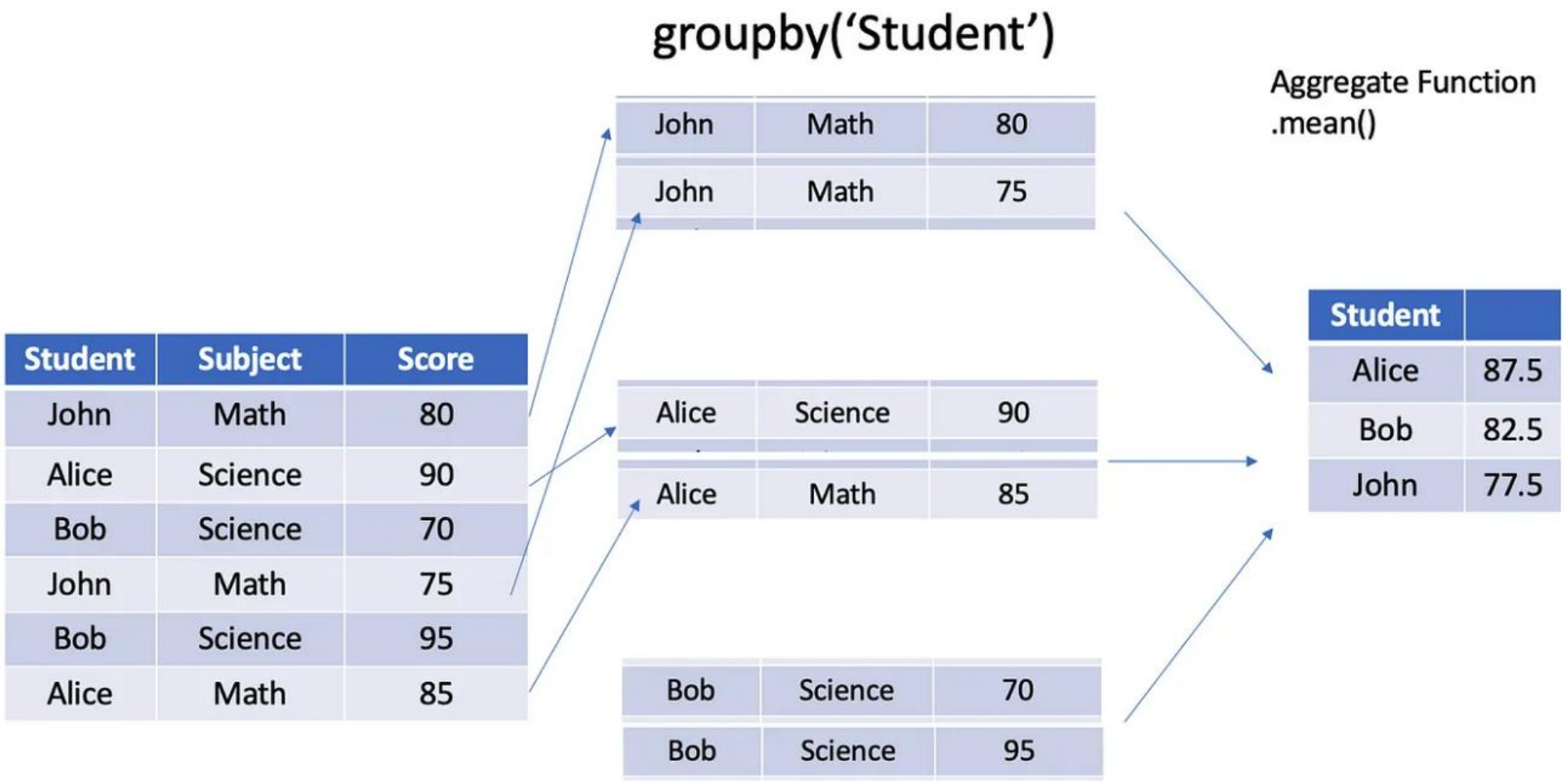
# MOST COMMONLY USED PANDAS OPERATIONS




[Pandas Cheat Sheet]

# GROUPBY GENERAL CONCEPT



# QUIZ: HOW MUCH DO YOU NOW KNOW ABOUT PANDAS?

- What is Python Pandas?
- How is a Pandas DataFrame different than a NumPy array?
- Name three unique operations (i.e., methods) you can do with Pandas?
- How do you read from or write to a csv file using Pandas?