

Programming Principles in Python (CSCI 503/490)

Data

Dr. David Koop

Quiz

Quiz

1. Which of the following is **true** about numpy arrays and python lists?
- (a) Slices of both arrays and lists are views of the original data
 - (b) Arrays and lists both can have elements of different types
 - (c) Arrays are faster to access than lists
 - (d) Arrays are mutable; lists are not

Quiz

2. Given the array `arr = np.array([[1, 2, 3, 4], [5, 6, 7, 8]])`, what is `arr[:, :3].shape`?

- (a) `(3, 3)`
- (b) `(3, 2)`
- (c) `(2, 3)`
- (d) `(2, 2)`

Quiz

3. Which of the following is invalid code?

- (a) `numpy.array([[1.5, 2, 3], [4, 5, 6]], dtype='float')`
- (b) `numpy.array([[1, 2, 3], [4, 5]])`
- (c) `numpy.array([[1, 2, 3], [4, 5, 6]], dtype="float")`
- (d) `numpy.array([[1.5, 2, 3], [4, 5, 6]])`

Quiz

4. What is a difference between a pandas Series and a polars Series?
- (a) A pandas Series is mutable; a polars Series is immutable
 - (b) A pandas Series has an index; a polars Series does not
 - (c) A pandas Series has a name; a polars Series does not
 - (d) A pandas Series is typed; a polars Series is untyped

Quiz

5. Evaluate `pd.Series(['a', 'b']) + pd.Series(['x', 'y'], [1, 0])`.

- (a) `pd.Series(['ay', 'bx'], [0, 1])`
- (b) `pd.Series(['ax', 'by'], [1, 0])`
- (c) `pd.Series(['a', 'b', 'x', 'y'], [0, 1, 1, 0])`
- (d) There is an error.

pandas

- Contains high-level data structures and manipulation tools designed to make data analysis fast and easy in Python
- Originally built on top of NumPy
- Built with the following requirements:
 - Data structures with labeled axes (aligning data)
 - Support time series data
 - Do arithmetic operations that include metadata (labels)
 - Handle missing data
 - Add merge and relational operations

polars

- Contains high-level data structures and manipulation tools designed to make data analysis "**lightning**" fast and easy in Python
 - Built using Apache Arrow
 - Written from scratch using Rust but with a Python API
 - Parallelized (uses multiple cores)
 - Intuitive API

Series

- A one-dimensional data structure (with a type)
 - `s = pl.Series([1, 2, 3])`
 - `t = pd.Series([1, 2, 3])`
- May also have a name and dtype
 - `s = pl.Series('name', ['a', 'b', 'c'], dtype=pl.Float)`
 - `t = pd.Series([1, 2, 3], name='num', dtype='float')`
- In pandas, a series has an index
 - `ti = pd.Series([1, 2, 3], ['a', 'b', 'c'])` # index ['a', 'b', 'c']
 - `ti = pd.Series({'a': 1, 'b': 2, 'c': 3})` # same index
- Indexing: `s[0]`, `t[0]`, `ti['a']`, `ti.iloc[0]`, `ti.loc['a']`

Series Operations

- Like numpy: elementwise / broadcasting
 - `Series([1,2,3]) + Series([1,2,3]) # Series([2,4,6])`
 - `Series([1,2,3]) + 4 # Series([5,6,7])`
- ...but for pandas, with custom indexes, the operations **align** on the index:
 - `pd.Series([1,2,3], index=list('abc')) +
pd.Series([1,2,3], index=list('cba'))
pd.Series([4,4,4], index=['a','b','c'])`
 - also have `.add`, `.subtract`, ... with `fill_value` argument

DataFrame

- A collection of Series (uniquely named)
 - Similar to a table in a database
 - Similar to a sheet in a spreadsheet
- ```
df = DataFrame({'state': ['Ohio', 'Ohio', 'Ohio', 'Nevada'],
 'year': [2000, 2001, 2002, 2001],
 'pop': [1.5, 1.7, 3.6, 2.4]})
```
- In pandas:
  - Has an index shared with each series
  - Index is automatically assigned just as with a series but can be passed in as well via `index` kwarg

# Assignment 7

---

- Concurrency, System Integration, and Structural Pattern Matching
- Download System Logs
- Locate Logs of Interest
- Read JSON & Binary Data
- Filter suspicious events using structural pattern matching
- Process all files using threading

# DataFrame Filtering

---

- polars:

- `df['pop'] > 2` # boolean Series
- `df.filter(pl.col('pop') > 2)` # subset of dataframe

- pandas:

- `dfa['pop'] > 2` # boolean Series
- `dfa[dfa['pop'] > 2]` # subset of dataframe
- `dfa.query('pop > 2')` # subset of dataframe

- Multiple criteria, use `&`, `|`, and `~`; remember parentheses!

- `df.filter((pl.col('year') < 2002) & (pl.col('pop') > 2))`
- `dfa[(dfa['year'] < 2002) & (dfa['pop'] > 2)]`



# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

344 rows x 17 columns



# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

Column Names

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

344 rows x 17 columns





# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

Column Names

| studyName | Sample Number | Species | Region | Island | Stage | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|

Index

|     |         |     |                                     |        |           |                    |       |     |          |      |
|-----|---------|-----|-------------------------------------|--------|-----------|--------------------|-------|-----|----------|------|
| 0   | PAL0708 | 1   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1  | Yes | 11/11/07 | 39.1 |
| 1   | PAL0708 | 2   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2  | Yes | 11/11/07 | 39.5 |
| 2   | PAL0708 | 3   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1  | Yes | 11/16/07 | 40.3 |
| 3   | PAL0708 | 4   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2  | Yes | 11/16/07 | NaN  |
| 4   | PAL0708 | 5   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1  | Yes | 11/16/07 | 36.7 |
| ... | ...     | ... | ...                                 | ...    | ...       | ...                | ...   | ... | ...      | ...  |
| 339 | PAL0910 | 120 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2 | No  | 12/1/09  | NaN  |
| 340 | PAL0910 | 121 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1 | Yes | 11/22/09 | 46.8 |
| 341 | PAL0910 | 122 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2 | Yes | 11/22/09 | 50.4 |
| 342 | PAL0910 | 123 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1 | Yes | 11/22/09 | 45.2 |
| 343 | PAL0910 | 124 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2 | Yes | 11/22/09 | 49.9 |

344 rows x 17 columns



# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

Column Names

| studyName | Sample Number | Species | Region | Island | Stage | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|

Index

|     |         |     |                                     |        |           |                    |       |     |          |      |
|-----|---------|-----|-------------------------------------|--------|-----------|--------------------|-------|-----|----------|------|
| 0   | PAL0708 | 1   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1  | Yes | 11/11/07 | 39.1 |
| 1   | PAL0708 | 2   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2  | Yes | 11/11/07 | 39.5 |
| 2   | PAL0708 | 3   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1  | Yes | 11/16/07 | 40.3 |
| 3   | PAL0708 | 4   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2  | Yes | 11/16/07 | NaN  |
| 4   | PAL0708 | 5   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1  | Yes | 11/16/07 | 36.7 |
| ... | ...     | ... | ...                                 | ...    | ...       | ...                | ...   | ... | ...      | ...  |
| 339 | PAL0910 | 120 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2 | No  | 12/1/09  | NaN  |
| 340 | PAL0910 | 121 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1 | Yes | 11/22/09 | 46.8 |
| 341 | PAL0910 | 122 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2 | Yes | 11/22/09 | 50.4 |
| 342 | PAL0910 | 123 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1 | Yes | 11/22/09 | 45.2 |
| 343 | PAL0910 | 124 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2 | Yes | 11/22/09 | 49.9 |

344 rows x 17 columns

Column: df['Island']



# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

Column Names

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

Row: df.loc[2]

Index

344 rows x 17 columns

Column: df['Island']



# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

Column Names

| studyName | Sample Number | Species | Region | Island | Stage | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|

Row: df.loc[2]

|     |         |     |                                     |        |           |                    |       |     |          |      |
|-----|---------|-----|-------------------------------------|--------|-----------|--------------------|-------|-----|----------|------|
| 0   | PAL0708 | 1   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1  | Yes | 11/11/07 | 39.1 |
| 1   | PAL0708 | 2   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2  | Yes | 11/11/07 | 39.5 |
| 2   | PAL0708 | 3   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1  | Yes | 11/16/07 | 40.3 |
| 3   | PAL0708 | 4   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2  | Yes | 11/16/07 | NaN  |
| 4   | PAL0708 | 5   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1  | Yes | 11/16/07 | 36.7 |
| ... | ...     | ... | ...                                 | ...    | ...       | ...                | ...   | ... | ...      | ...  |
| 339 | PAL0910 | 120 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2 | No  | 12/1/09  | NaN  |
| 340 | PAL0910 | 121 | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1 | Yes | 11/22/09 | 46.8 |

Index

Cell: df.loc[341, 'Species']

Gentoo penguin (Pygoscelis papua)

|     |         |     |                                   |        |        |                    |       |     |          |      |
|-----|---------|-----|-----------------------------------|--------|--------|--------------------|-------|-----|----------|------|
| 342 | PAL0910 | 123 | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N43A1 | Yes | 11/22/09 | 45.2 |
| 343 | PAL0910 | 124 | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N43A2 | Yes | 11/22/09 | 49.9 |

344 rows x 17 columns

Column: df['Island']

# pandas DataFrame

```
df = pd.read_csv('penguins_lter.csv')
```

Column Names

| studyName | Sample Number | Species | Region | Island | Stage | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|
|-----------|---------------|---------|--------|--------|-------|---------------|-------------------|----------|--------------------|

Row: df.loc[2]

|   |         |   |                                     |        |           |                    |      |     |          |      |
|---|---------|---|-------------------------------------|--------|-----------|--------------------|------|-----|----------|------|
| 0 | PAL0708 | 1 | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1 | Yes | 11/11/07 | 39.1 |
| 1 | PAL0708 | 2 | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2 | Yes | 11/11/07 | 39.5 |
| 2 | PAL0708 | 3 | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1 | Yes | 11/16/07 | 40.3 |

Index

|     |         |     |                                     |        |           |                    |      |     |          |     |
|-----|---------|-----|-------------------------------------|--------|-----------|--------------------|------|-----|----------|-----|
| 3   | PAL0708 | 4   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2 | Yes | 11/16/07 | NaN |
| 4   | PAL0708 | 5   | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1 | Yes | 11/16/07 |     |
| ... | ...     | ... | ...                                 | ...    | ...       | ...                | ...  | ... | ...      | ... |

Missing Data

Cell: df.loc[341, 'Species']

|     |         |     |                                   |        |        |                    |       |     |          |      |
|-----|---------|-----|-----------------------------------|--------|--------|--------------------|-------|-----|----------|------|
| 339 | PAL0910 | 120 | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N38A2 | No  | 12/1/09  | NaN  |
| 340 | PAL0910 | 121 | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N39A1 | Yes | 11/22/09 | 46.8 |
|     |         |     | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N39A2 | Yes | 11/22/09 | 50.4 |
| 342 | PAL0910 | 123 | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N43A1 | Yes | 11/22/09 | 45.2 |
| 343 | PAL0910 | 124 | Gentoo penguin (Pygoscelis papua) | Anvers | Biscoe | Adult, 1 Egg Stage | N43A2 | Yes | 11/22/09 | 49.9 |

344 rows x 17 columns

Column: df['Island']

# polars DataFrame

shape: (344, 10)

| studyName | Sample Number | Species                             | Region   | Island      | Stage                | Individual ID | Clutch Completion | Date Egg   | Culmen Length (mm) |
|-----------|---------------|-------------------------------------|----------|-------------|----------------------|---------------|-------------------|------------|--------------------|
| str       | i64           | str                                 | str      | str         | str                  | str           | str               | str        | f64                |
| "PAL0708" | 1             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A1"        | "Yes"             | "11/11/07" | 39.1               |
| "PAL0708" | 2             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A2"        | "Yes"             | "11/11/07" | 39.5               |
| "PAL0708" | 3             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A1"        | "Yes"             | "11/16/07" | 40.3               |
| "PAL0708" | 4             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A2"        | "Yes"             | "11/16/07" | null               |
| "PAL0708" | 5             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N3A1"        | "Yes"             | "11/16/07" | 36.7               |
| ...       | ...           | ...                                 | ...      | ...         | ...                  | ...           | ...               | ...        | ...                |
| "PAL0910" | 120           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N38A2"       | "No"              | "12/1/09"  | null               |
| "PAL0910" | 121           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A1"       | "Yes"             | "11/22/09" | 46.8               |
| "PAL0910" | 122           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A2"       | "Yes"             | "11/22/09" | 50.4               |
| "PAL0910" | 123           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A1"       | "Yes"             | "11/22/09" | 45.2               |
| "PAL0910" | 124           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A2"       | "Yes"             | "11/22/09" | 49.9               |



# polars DataFrame

## Column Names & Types

shape: (344, 10)

| studyName | Sample Number | Species                             | Region   | Island      | Stage                | Individual ID | Clutch Completion | Date Egg   | Culmen Length (mm) |
|-----------|---------------|-------------------------------------|----------|-------------|----------------------|---------------|-------------------|------------|--------------------|
| str       | i64           | str                                 | str      | str         | str                  | str           | str               | str        | f64                |
| "PAL0708" | 1             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A1"        | "Yes"             | "11/11/07" | 39.1               |
| "PAL0708" | 2             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A2"        | "Yes"             | "11/11/07" | 39.5               |
| "PAL0708" | 3             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A1"        | "Yes"             | "11/16/07" | 40.3               |
| "PAL0708" | 4             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A2"        | "Yes"             | "11/16/07" | null               |
| "PAL0708" | 5             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N3A1"        | "Yes"             | "11/16/07" | 36.7               |
| ...       | ...           | ...                                 | ...      | ...         | ...                  | ...           | ...               | ...        | ...                |
| "PAL0910" | 120           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N38A2"       | "No"              | "12/1/09"  | null               |
| "PAL0910" | 121           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A1"       | "Yes"             | "11/22/09" | 46.8               |
| "PAL0910" | 122           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A2"       | "Yes"             | "11/22/09" | 50.4               |
| "PAL0910" | 123           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A1"       | "Yes"             | "11/22/09" | 45.2               |
| "PAL0910" | 124           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A2"       | "Yes"             | "11/22/09" | 49.9               |



# polars DataFrame

Column Names  
& Types

shape: (344, 10)

| studyName | Sample Number | Species                             | Region   | Island      | Stage                | Individual ID | Clutch Completion | Date Egg   | Culmen Length (mm) |
|-----------|---------------|-------------------------------------|----------|-------------|----------------------|---------------|-------------------|------------|--------------------|
| str       | i64           | str                                 | str      | str         | str                  | str           | str               | str        | f64                |
| "PAL0708" | 1             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A1"        | "Yes"             | "11/11/07" | 39.1               |
| "PAL0708" | 2             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A2"        | "Yes"             | "11/11/07" | 39.5               |
| "PAL0708" | 3             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A1"        | "Yes"             | "11/16/07" | 40.3               |
| "PAL0708" | 4             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A2"        | "Yes"             | "11/16/07" | null               |
| "PAL0708" | 5             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N3A1"        | "Yes"             | "11/16/07" | 36.7               |
| ...       | ...           | ...                                 | ...      | ...         | ...                  | ...           | ...               | ...        | ...                |
| "PAL0910" | 120           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N38A2"       | "No"              | "12/1/09"  | null               |
| "PAL0910" | 121           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A1"       | "Yes"             | "11/22/09" | 46.8               |
| "PAL0910" | 122           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A2"       | "Yes"             | "11/22/09" | 50.4               |
| "PAL0910" | 123           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A1"       | "Yes"             | "11/22/09" | 45.2               |
| "PAL0910" | 124           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    |                      |               |                   |            |                    |

Column: df[ 'Island' ]

# polars DataFrame

Column Names  
& Types

Row: df[2]

shape: (344, 10)

| studyName | Sample Number | Species                             | Region   | Island      | Stage                | Individual ID | Clutch Completion | Date Egg   | Culmen Length (mm) |
|-----------|---------------|-------------------------------------|----------|-------------|----------------------|---------------|-------------------|------------|--------------------|
| str       | i64           | str                                 | str      | str         | str                  | str           | str               | str        | f64                |
| "PAL0708" | 1             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A1"        | "Yes"             | "11/11/07" | 39.1               |
| "PAL0708" | 2             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A2"        | "Yes"             | "11/11/07" | 39.5               |
| "PAL0708" | 3             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A1"        | "Yes"             | "11/16/07" | 40.3               |
| "PAL0708" | 4             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A2"        | "Yes"             | "11/16/07" | null               |
| "PAL0708" | 5             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N3A1"        | "Yes"             | "11/16/07" | 36.7               |
| ...       | ...           | ...                                 | ...      | ...         | ...                  | ...           | ...               | ...        | ...                |
| "PAL0910" | 120           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N38A2"       | "No"              | "12/1/09"  | null               |
| "PAL0910" | 121           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A1"       | "Yes"             | "11/22/09" | 46.8               |
| "PAL0910" | 122           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A2"       | "Yes"             | "11/22/09" | 50.4               |
| "PAL0910" | 123           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A1"       | "Yes"             | "11/22/09" | 45.2               |
| "PAL0910" | 124           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    |                      |               |                   |            |                    |

Column: df['Island']

# polars DataFrame

Column Names  
& Types

Row: df[2]

Cell: df['Species'][341]

shape: (344, 10)

| studyName | Sample Number | Species                             | Region   | Island      | Stage                | Individual ID | Clutch Completion | Date Egg   | Culmen Length (mm) |
|-----------|---------------|-------------------------------------|----------|-------------|----------------------|---------------|-------------------|------------|--------------------|
| str       | i64           | str                                 | str      | str         | str                  | str           | str               | str        | f64                |
| "PAL0708" | 1             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A1"        | "Yes"             | "11/11/07" | 39.1               |
| "PAL0708" | 2             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A2"        | "Yes"             | "11/11/07" | 39.5               |
| "PAL0708" | 3             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A1"        | "Yes"             | "11/16/07" | 40.3               |
| "PAL0708" | 4             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A2"        | "Yes"             | "11/16/07" | null               |
| "PAL0708" | 5             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N3A1"        | "Yes"             | "11/16/07" | 36.7               |
| ...       | ...           | ...                                 | ...      | ...         | ...                  | ...           | ...               | ...        | ...                |
| "PAL0910" | 120           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N38A2"       | "No"              | "12/1/09"  | null               |
| "PAL0910" | 121           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A1"       | "Yes"             | "11/22/09" | 46.8               |
| "PAL0910" | 122           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A2"       | "Yes"             | "11/22/09" | 50.4               |
| "PAL0910" | 123           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A1"       | "Yes"             | "11/22/09" | 45.2               |
| "PAL0910" | 124           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    |                      |               |                   |            |                    |

Column: df['Island']



# polars DataFrame

Column Names  
& Types

Row: df[2]

Cell: df['Species'][341]

shape: (344, 10)

| studyName | Sample Number | Species                             | Region   | Island      | Stage                | Individual ID | Clutch Completion | Date Egg   | Culmen Length (mm) |
|-----------|---------------|-------------------------------------|----------|-------------|----------------------|---------------|-------------------|------------|--------------------|
| str       | i64           | str                                 | str      | str         | str                  | str           | str               | str        | f64                |
| "PAL0708" | 1             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A1"        | "Yes"             | "11/11/07" | 39.1               |
| "PAL0708" | 2             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N1A2"        | "Yes"             | "11/11/07" | 39.5               |
| "PAL0708" | 3             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A1"        | "Yes"             | "11/16/07" | 40.3               |
| "PAL0708" | 4             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N2A2"        | "Yes"             | "11/16/07" | null               |
| "PAL0708" | 5             | "Adelie Penguin (Pygoscelis ade..." | "Anvers" | "Torgersen" | "Adult, 1 Egg Stage" | "N3A1"        | "Yes"             | "11/16/07" |                    |
| ...       | ...           | ...                                 | ...      | ...         | ...                  | ...           | ...               | ...        | ...                |
| "PAL0910" | 120           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N38A2"       | "No"              | "12/1/09"  | null               |
|           |               | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A1"       | "Yes"             | "11/22/09" | 46.8               |
| "PAL0910" | 122           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N39A2"       | "Yes"             | "11/22/09" | 50.4               |
| "PAL0910" | 123           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    | "Adult, 1 Egg Stage" | "N43A1"       | "Yes"             | "11/22/09" | 45.2               |
| "PAL0910" | 124           | "Gentoo penguin (Pygoscelis pap..." | "Anvers" | "Biscoe"    |                      |               |                   |            |                    |

Missing Data

Column: df['Island']

# pandas Filtering

```
df[df['Culmen Length (mm)'] > 40]
```

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

344 rows x 17 columns





# pandas Filtering

```
df[df['Culmen Length (mm)'] > 40]
```

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

344 rows x 17 columns

# polars Filtering

```
df.filter(pl.col('Culmen Length (mm)') > 40)
```

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

344 rows x 17 columns



# polars Filtering

```
df.filter(pl.col('Culmen Length (mm)') > 40)
```

|     | studyName | Sample Number | Species                             | Region | Island    | Stage              | Individual ID | Clutch Completion | Date Egg | Culmen Length (mm) |
|-----|-----------|---------------|-------------------------------------|--------|-----------|--------------------|---------------|-------------------|----------|--------------------|
| 0   | PAL0708   | 1             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A1          | Yes               | 11/11/07 | 39.1               |
| 1   | PAL0708   | 2             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N1A2          | Yes               | 11/11/07 | 39.5               |
| 2   | PAL0708   | 3             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A1          | Yes               | 11/16/07 | 40.3               |
| 3   | PAL0708   | 4             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N2A2          | Yes               | 11/16/07 | NaN                |
| 4   | PAL0708   | 5             | Adelie Penguin (Pygoscelis adeliae) | Anvers | Torgersen | Adult, 1 Egg Stage | N3A1          | Yes               | 11/16/07 | 36.7               |
| ... | ...       | ...           | ...                                 | ...    | ...       | ...                | ...           | ...               | ...      | ...                |
| 339 | PAL0910   | 120           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N38A2         | No                | 12/1/09  | NaN                |
| 340 | PAL0910   | 121           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A1         | Yes               | 11/22/09 | 46.8               |
| 341 | PAL0910   | 122           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N39A2         | Yes               | 11/22/09 | 50.4               |
| 342 | PAL0910   | 123           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A1         | Yes               | 11/22/09 | 45.2               |
| 343 | PAL0910   | 124           | Gentoo penguin (Pygoscelis papua)   | Anvers | Biscoe    | Adult, 1 Egg Stage | N43A2         | Yes               | 11/22/09 | 49.9               |

344 rows x 17 columns



# Sorting

---

- polars: `df.sort('pop')`
- pandas: `dfa.sort_values('pop')`
- Can sort by multiple columns, too
- pandas also has a `sort_index` method to sort by the index
  - `dfa.sort_index()`



# Statistics

---

- Many common statistical methods can be used (min, max, median, etc.)
- `describe`: shortcut for easy stats!

```
In [204]: df.describe()
```

```
Out[204]:
```

|       | one      | two       |
|-------|----------|-----------|
| count | 3.000000 | 2.000000  |
| mean  | 3.083333 | -2.900000 |
| std   | 3.493685 | 2.262742  |
| min   | 0.750000 | -4.500000 |
| 25%   | 1.075000 | -3.700000 |
| 50%   | 1.400000 | -2.900000 |
| 75%   | 4.250000 | -2.100000 |
| max   | 7.100000 | -1.300000 |

```
In [205]: obj = Series(['a', 'a', 'b', 'c'] * 4)
```

```
In [206]: obj.describe()
```

```
Out[206]:
```

|        |        |
|--------|--------|
| count  | 16     |
| unique | 3      |
| top    | a      |
| freq   | 8      |
| dtype: | object |

# Unique Values and Value Counts

---

- polars: `unique()` returns a Series/DataFrame with duplicates dropped
- pandas is more complicated
  - Series `unique()` returns an array with only the unique values (no index)
    - `s = Series(['c', 'a', 'd', 'a', 'a', 'b', 'b', 'c', 'c'])`  
`s.unique()` # `array(['c', 'a', 'd', 'b'])`
  - Data Frame `drop_duplicates` returns a DataFrame with duplicates dropped
- Also `nunique()` / `n_unique()` to count number of unique entries
- `value_counts` returns a Series/DataFrame with index frequencies:
  - `s.value_counts()` # `Series({'c': 3, 'a': 3, 'b': 2, 'd': 1})`

# Reading and Writing CSV Files

---

- polars
  - `df = pl.read_csv(<fname>)`
  - `df.write_csv(<fname>)`
- pandas
  - `dfa = pd.read_csv(<fname>)`
  - `dfa.to_csv(<fname>)`
- Many options available!



# Reading & Writing Data in Pandas

| Format | Data Description                     | Reader         | Writer       |
|--------|--------------------------------------|----------------|--------------|
| text   | <a href="#">CSV</a>                  | read_csv       | to_csv       |
| text   | Fixed-Width Text File                | read_fwf       |              |
| text   | <a href="#">JSON</a>                 | read_json      | to_json      |
| text   | <a href="#">HTML</a>                 | read_html      | to_html      |
| text   | Local clipboard                      | read_clipboard | to_clipboard |
|        | <a href="#">MS Excel</a>             | read_excel     | to_excel     |
| binary | <a href="#">OpenDocument</a>         | read_excel     |              |
| binary | <a href="#">HDF5 Format</a>          | read_hdf       | to_hdf       |
| binary | <a href="#">Feather Format</a>       | read_feather   | to_feather   |
| binary | <a href="#">Parquet Format</a>       | read_parquet   | to_parquet   |
| binary | <a href="#">ORC Format</a>           | read_orc       |              |
| binary | <a href="#">Msgpack</a>              | read_msgpack   | to_msgpack   |
| binary | <a href="#">Stata</a>                | read_stata     | to_stata     |
| binary | <a href="#">SAS</a>                  | read_sas       |              |
| binary | <a href="#">SPSS</a>                 | read_spss      |              |
| binary | <a href="#">Python Pickle Format</a> | read_pickle    | to_pickle    |
| SQL    | <a href="#">SQL</a>                  | read_sql       | to_sql       |
| SQL    | <a href="#">Google BigQuery</a>      | read_gbq       | to_gbq       |

[[https://pandas.pydata.org/pandas-docs/stable/user\\_guide/io.html](https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html)]



# pandas read\_csv

---

- Convenient method to read csv files
- Lots of different options to help get data into the desired format
- Basic: `dfa = pd.read_csv(fname)`
- Parameters:
  - `path`: where to read the data from
  - `sep` (or `delimiter`): the delimiter (`,`, `' '`, `'\t'`, `'\s+'`)
  - `header`: if `None`, no header
  - `index_col`: which column to use as the row index
  - `names`: list of header names (e.g. if the file has no header)
  - `skiprows`: number of list of lines to skip

# Writing CSV data with pandas

---

- Basic: `dfa.to_csv(<fname>)`
- Change delimiter with `sep` kwarg:
  - `dfa.to_csv('example.dsv', sep='|')`
- Change missing value representation
  - `dfa.to_csv('example.dsv', na_rep='NULL')`
- Don't write row or column labels:
  - `dfa.to_csv('example.csv', index=False, header=False)`
- Series may also be written to csv

# Missing Data

---

- polars: shows `null`
- pandas: shows `NaN` (or `NA` or `None` depending on dtype)
- Checking if missing:
  - polars: `pl.col('pop').is_null(), .is_not_null()`
  - pandas: `dfa['pop'].isnull(), .notnull()`
- Drop missing data:
  - polars: `pl.col('pop').drop_nulls()`, pandas: `dfa['pop'].dropna()`
- Filling in missing data:
  - polars: `pl.col('pop').fill_null()`, (`forward`, `backward`, `max`,...)
  - pandas: `dfa['pop'].fillna()`, `now` `ffill()`, `bfill()`



# Derived Data

---

- Create new columns from existing columns
- pandas
  - `dfa["CulmenRatio"] = dfa['CLength'] / dfa['CDepth'] # Mut!`
  - `dfa = dfa.assign(CulmenRatio= dfa['CLength'] / dfa['CDepth'])`
- polars
  - `df.with_columns(  
    (df['CLength'] / df['CDepth']).alias('CulmenRatio'))`
- Note that operations are computed in a vectorized manner
- Similarities to functional paradigm (map/filter):
  - specify the operation once, on entire column/frame
  - no loops



# pandas inplace

---

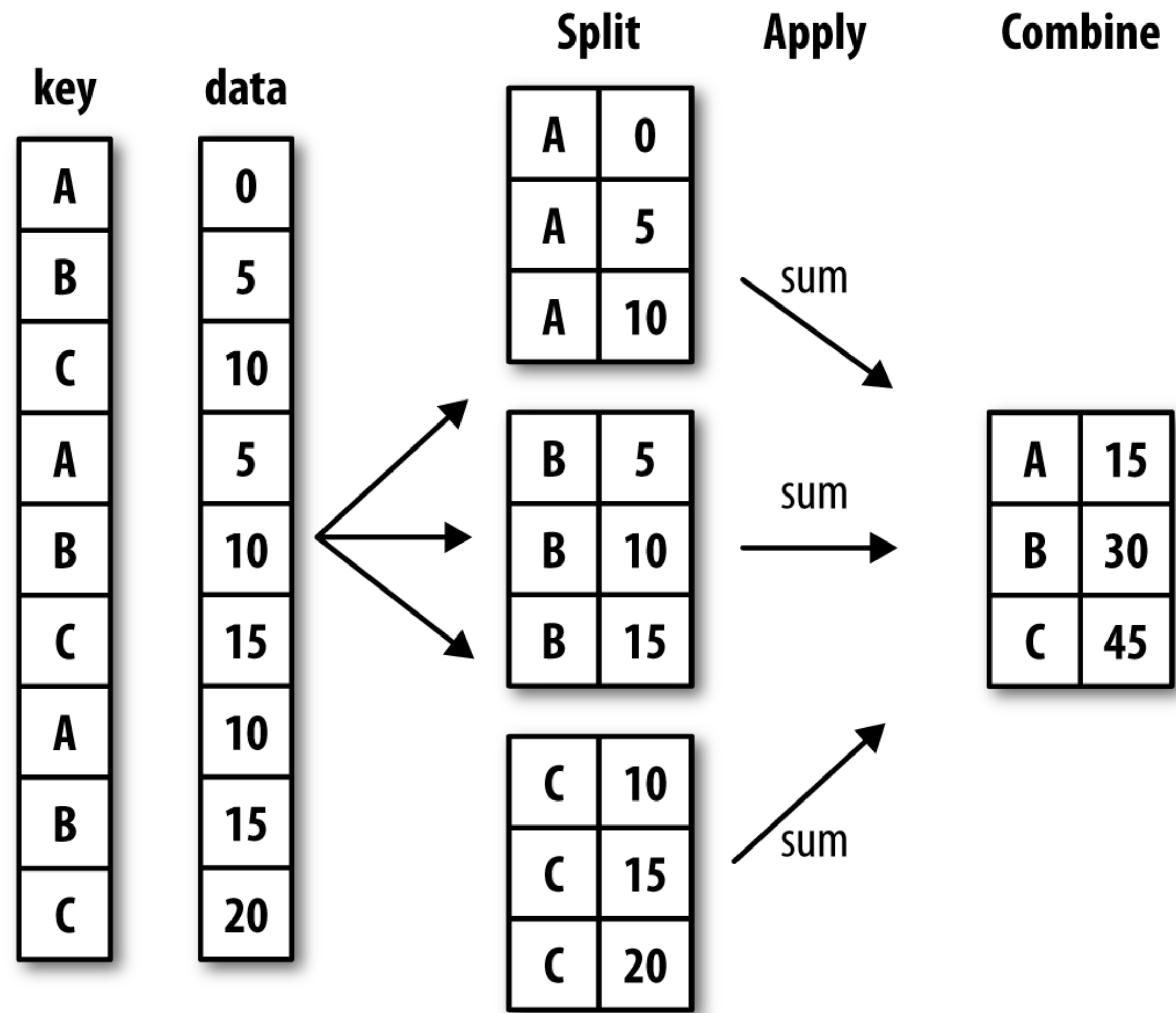
- Generally, when we modify a data frame, we reassign:
  - `rdf = dfa.reset_index()`
  - This is usually very **efficient**
  - Allows for method chaining
- There are versions where you can do this "inplace" (**try to avoid this**)
  - `dfa.reset_index(inplace=True)`
  - This means **no reassignment**, but it isn't usually any faster nor better
  - Sometimes still creates a copy
  - Will likely be deprecated

# Aggregation

---

- Descriptive statistics
  - `df['Culmen Length (mm)'].mean()`
  - `.median()`
  - `.describe()`
  - `.count()`
  - `.min()`, `.max()`
- Also general methods
  - `.sum()`
  - `.product()`

# Split-Apply-Combine



[W. McKinney, Python for Data Analysis]