Computing in Context: Fall 2024

Lab 8 | Pandas Pandas Pandas

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Dr Sarah Ashcroft-Jones | Mailman School of Public Health

Section I| Lecture Review

- Looking at data
  - `df.head()` : show the first 5 rows
  - `df.tail()` : show the last 5 rows
  - `df.sample(n)` : show n random rows
- Attributes
  - `df.shape` : size of the table
  - `df.dtypes` : print dtype of cols
  - `df.columns` : column index
  - `df.index` : rows index
- Indexing
  - `df['age']` : get column 'age'
  - `df[['age', 'name']]` : multiple columns
  - `df.iloc[0, 2]` : one element, by position
- Exploration
  - `df['name'].unique()` : unique values
  - `df['age'].describe()` : summary stats
  - `df['age'].value_counts(dropna=False)` : number of rows per unique value in column
- Adding a column
  - `df['new'] = df['age'] * 3.1` : add new column
- Filtering
  - `df[df['age'] > 30]` : select rows where condition is True
- Operations
  - `df.min(), .max(), .mean(), .std(), etc.` : column-wise operations
  - `df.count()` : count of non-NaN elements in columns
  - `df.sort_values('name')` : reorder rows by values of column 'name'
  - `df.sort_index()` : reorder rows by the index values
- String operations
  - `df['name'].str` : accessor for operations on the strings in a col
  - `df['name'].str[2:4]` : slice the strings in a col
  - `df['name'].str.count('a')` : count the letter 'a' in the string in a col

In lab work time

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

End of session: Questions?