

Pandas Recap

GA DAT5

<https://pandas.pydata.org/pandas-docs/stable/api.html>

Functions

1. I/O for tabular files

- a. `df = pandas.read_csv(filename, [options])`
- b. `df.to_csv(filename, index=False, [options])`

2. Get overview of the data

- a. `df.info()` : shape of data frame, columns descriptions, etc.
- b. `df.describe()` : stats by columns
- c. `len(df)`
- d. `df.head()`

3. Select by rows and or columns with/without conditions

- a. `df['column_name']` or `df.column_name` : get column “column_name”
- b. `df[0:10]` : first 10 rows
- c. `df.loc[0:10, 'column_name']`
- d. `df[df[“age”] >=21]`

Functions

4. Aggregate/sort by certain columns

- a. `df.groupby('column_name').mean()`
- b. `df.sort_values(by='column_name', ascending=True)` : sort in descending order

5. Iterate over rows

- a. `Item_iters = df.iter['column_name'] -> item_iters.next()`

6. Operation between columns

- a. `df['column_name1'] - df['column_name2']`

7. Copy data frame

- a. `df_copy = df.copy()`