Analytics Jumpstart

Intro to commonly used pandas methods

Nashville Software School

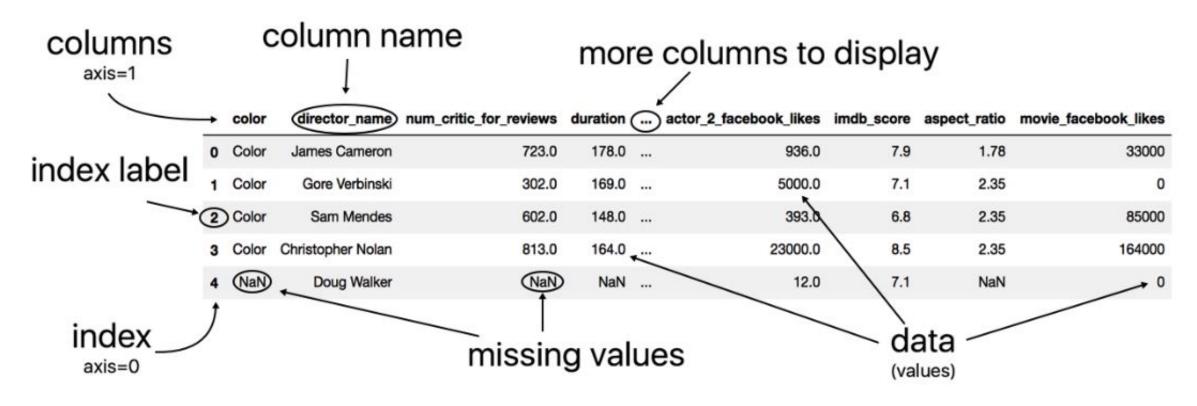


Goals for today

- Learn some pandas methods
- Work on coding tasks
- Use the pandas API to understand methods and their signatures



The Anatomy Of A Dataframe



So, when thinking about axes..

•Axis = $0 \longrightarrow Rows$

•Axis = 1 --> Columns

You see this when

running dataframe.shape --> (n_rows,n_cols)



Get Data / Process + Clean Data / Exploratory Data Analysis

CSV Workflow

Open the file in text editor to determine

- is there a header? pandas default is header = 'infer'
- are there notes (non-data) at the top? pandas default is skiprows = None
- are there footnotes or other (non-data) at the bottom? pandas default is nrows = None



Get Data / Process + Clean Data / Exploratory Data Analysis

Workflow

- Clean column names
- Drop unnecessary columns (avoid using inplace = True since it is deprecated, instead assign the df back to itself or use method chaining)



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

- **pd.read_csv()** read a comma delimited file; always look at the raw file in a text editor (not excel); additional arguments may be needed to handle extra rows at the top and extra data (footnotes) at the bottom.
- **df.head()** looks at the top of the DataFrame; 5 rows by default
- **df.tail()** looks at the bottom of the DataFrame; 5 rows by default
- **df.shape** returns a tuple with number of rows and number of columns
- **df.drop()** drop the specified labels (either rows or columns) from the DataFrame
- **df.columns** column labels attribute
- df.rename() rename values (can pass in a dictionary with existing columns as the key and new ones as the values)



You can subset a dataframe according to the values in a column (or row) using the .loc[] accessor

• df.loc[] – pass in row name and column name to access data at that location

You can also create a subset by slicing off just the columns you want to keep

• df[[]] - creates a slice (subset) of the DataFrame with just the columns specified inside the double brackets



Let's open our first shared notebook - public_art1.ipynb

