Analytics Jumpstart

Intro to commonly used *pandas* methods

Nashville Software School

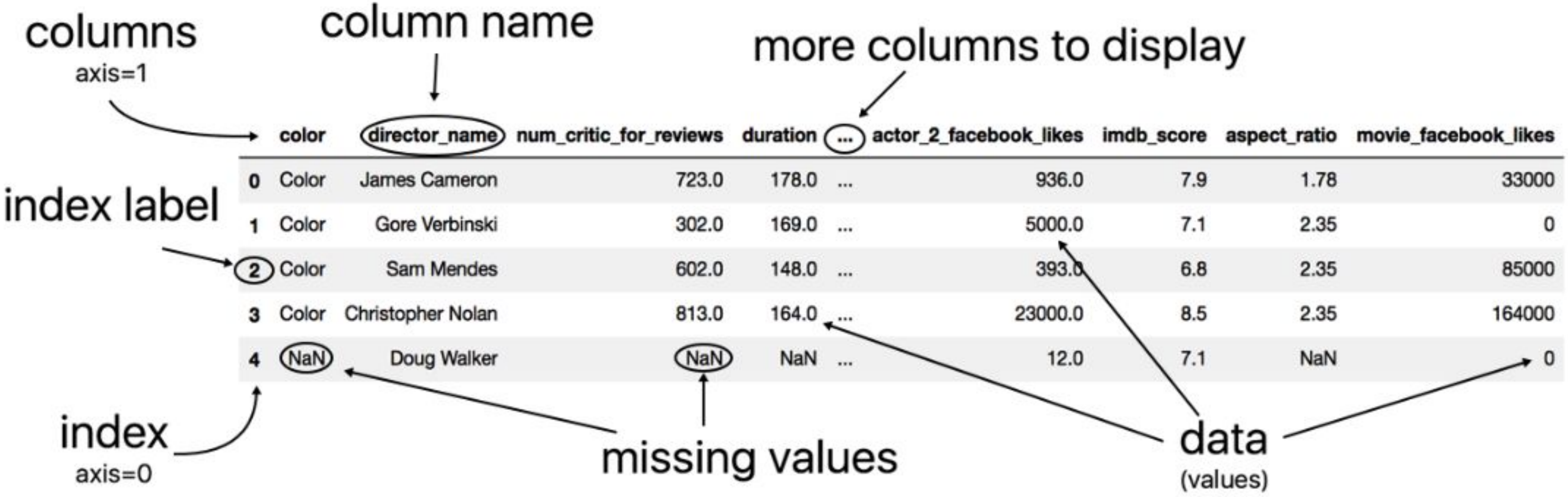
Goals for today

- **Learn some *pandas* methods**

- **Work on coding tasks**

- **Use the *pandas* API to understand methods and their signatures**

The Anatomy Of A Dataframe



	color	director_name	num_critic_for_reviews	duration	...	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes
0	Color	James Cameron	723.0	178.0	...	936.0	7.9	1.78	33000
1	Color	Gore Verbinski	302.0	169.0	...	5000.0	7.1	2.35	0
2	Color	Sam Mendes	602.0	148.0	...	393.0	6.8	2.35	85000
3	Color	Christopher Nolan	813.0	164.0	...	23000.0	8.5	2.35	164000
4	NaN	Doug Walker	NaN	NaN	...	12.0	7.1	NaN	0

So, when thinking about *axes*..

- •Axis = 0 --> Rows
- •Axis = 1 --> Columns

You see this when
running dataframe.shape --> (n_rows,n_cols)

NASHVILLE
SOFTWARE
SCHOOL

pandas – **https://pandas.pydata.org/pandas-docs/stable/reference/index.html**

Importing Data

- **pd.read_csv()** – read a comma delimited file; good practice is to look at the raw file in a text editor (like Visual Studio Code, not Excel); additional arguments may be needed to handle extra rows at the top and extra data (footnotes) at the bottom.

Inspecting

- **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: (number of rows, number of columns)
- **df.info()** – method to get information about the DataFrame

pandas – https://pandas.pydata.org/pandas-docs/stable/reference/index.html

Modifying

- **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)
- **df.drop()** – drop the specified labels (either rows or columns) from the DataFrame

Summarizing

- **.unique()** – returns the unique values in a column
- **.nunique()** - returns the *number* of unique elements in a column
- **.value_counts()** - returns the unique elements in a column and the number of appearances of each

Slicing/Filtering

- **df.loc[]** – pass in row name and column name to access data at that location
- **df[[ ]]** - creates a slice (subset) of the DataFrame including just the columns passed in

Let’s open our first shared notebook so we can see these in action:

notebook_01_public_art_part_1.ipynb