15 Pandas Functions

for 90% of the work

By Travis Tang
https://www.linkedin.com/in/travistang

sample code included!

Top 15 Pandas functions

1. pd.read_csv()
2. df.assign()
3. df.query()
4. df.sort_values()
5. df.describe()
6. df.info()
7. df.sample()
8. df.dropna()
9. df.drop()
10. pd.pivot_table()
11. df.groupby()
12. df.transpose()
13. df.merge()
14. df.rename()
15. df.to_csv()

By Travis Tang
https://www.linkedin.com/in/travistang

15 Pandas Functions for 90% of the Work

1. `read_csv()` for reading data

In [2]:

```
import pandas as pd

# Read in CSV file
df = pd.read_csv('./netflix_titles.csv')
# We're using the Netflix Shows (https://www.kaggle.com/datasets/shivamb/netflix-shows) dataset.
# If you'd like to follow along, download the file and save it as "netflix_titles.csv"
# in the same directory
df
```

Out[2]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...	...	...	...
				David	Mark Ruffalo, Jake	United	November			

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
8802	s8803	Movie	Zodiac	David Fincher	Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min

8807 rows × 12 columns

2. Use `df.assign(col_name = function)` to create new columns

In [44]:

```
# 2. Create a new column with assign that finds if a release_year is after 2020
df = df.assign(after_2020 = df['release_year'] > 2020)
df

# Alternatively, use this
# df['after_2020'] = df['release_year'] > 2020
```

Out[44]:

	show_id	type	title	...	listed_in	description	after_2020
0	s1	Movie	Dick Johnson Is Dead	...	Documentaries	As her father nears the end of his life, filmm...	False
1	s2	TV Show	Blood & Water	...	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	True
2	s3	TV Show	Ganglands	...	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	True
...	...	...	...	...	...	...	...
8804	s8805	Movie	Zombieland	...	Comedies, Horror Movies	Looking to survive in a world taken over by zo...	False
8805	s8806	Movie	Zoom	...	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...	False

	show_id	type	title	...	listed_in	description	after_2020
8806	s8807	Movie	Zubaan	...	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...	False

8807 rows × 13 columns

3. Use `df.query(condition)` to find a subset of the columns

In [45]:

```
# 3. Select rows with query
df.query('release_year == 2020 & type == "Movie"')

# Alternatively, use loc to select rows
# df.loc[(df['release_year'] == 2020) & (df['type'] == 'Movie')]
```

Out[45]:

	show_id	type	title	...	listed_in	description	after_2020
0	s1	Movie	Dick Johnson Is Dead	...	Documentaries	As her father nears the end of his life, filmm...	False
16	s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in ...	...	Documentaries, International Movies	Declassified documents reveal the post-WWII li...	False
78	s79	Movie	Tughlaq Durbar	...	Comedies, Dramas, International Movies	A budding politician has devious plans to rise...	False
...	...	...	...	...	...	...	...
5972	s5973	Movie	#cats_the_mewvie	...	Documentaries, International Movies	This pawesome documentary explores how our fel...	False
7594	s7595	Movie	Norm of the North: Family Vacation	...	Children & Family Movies	Stressed by his duties as king and father, Nor...	False
8099	s8100	Movie	Straight Up	...	Comedies, Independent Movies, LGBTQ Movies	When a gay brainiac with OCD questions his ide...	False

517 rows × 13 columns

4. `df.sort_values()` for sorting

In [46]:

```
# 4. Sort by column
df.sort_values(['show_id'])
```

Out[46]:

	show_id	type	title	...	listed_in	description	after_2020
0	s1	Movie	Dick Johnson Is Dead	...	Documentaries	As her father nears the end of his life, filmm...	False
9	s10	Movie	The Starling	...	Comedies, Dramas	A woman adjusting to life after a loss contend...	True
99	s100	TV Show	On the Verge	...	TV Comedies, TV Dramas	Four women — a chef, a single mom, an heiress ...	True

...	show_id	type	title	...	listed_in	description	after_2020
996	s997	Movie	HOMUNCULUS	...	Horror Movies, International Movies, Thrillers	Truth and illusion blurs when a homeless amnes...	True
997	s998	TV Show	Life in Color with David Attenborough	...	British TV Shows, Docuseries, International TV...	Using innovative technology, this docuseries e...	True
998	s999	Movie	Searching For Sheela	...	Documentaries, International Movies	Journalists and fans await Ma Anand Sheela as ...	True

8807 rows × 13 columns

5. `df.describe()` to get summary statistics

In [47]:

```
# 5. Get useful summary statistics
df.describe()
```

Out[47]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
...	...
50%	2017.000000
75%	2019.000000
max	2021.000000

8 rows × 1 columns

6. `df.info()` to get information about the dataframe

In [48]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
 12  after_2020    8807 non-null   bool
dtypes: bool(1), int64(1), object(11)
memory usage: 834.4+ KB
```

7. `df.sample()` to get random rows

In [49]:

```
# Randomly select 5 rows
df.sample(5)

# Alternative:
# Select the top 5 rows with df.head()
# Select the bottom 5 rows with df.tail()
```

Out[49]:

	show_id	type	title	...	listed_in	description	after_2020
5044	s5045	Movie	When We First Met	...	Comedies, Romantic Movies	Using a magical photo booth that sends him bac...	False
4560	s4561	Movie	Bathinda Express	...	Dramas, International Movies, Sports Movies	An ambitious young athlete endeavors to revive...	False
1738	s1739	Movie	A Christmas Catch	...	Dramas, Romantic Movies	A cop working undercover to trail a possible d...	False
1035	s1036	Movie	The Zookeeper's Wife	...	Dramas	When the Nazis invade Poland, Warsaw Zoo caret...	False
7997	s7998	Movie	Shark Busters	...	Dramas, International Movies	Hit by a media storm over his own mounting deb...	False

5 rows × 13 columns

8. `df.dropna()` to drop rows with null values

In [50]:

```
# Drop rows with missing values in the country column
df = df.dropna(subset = ['country'])
df
```

Out[50]:

	show_id	type	title	...	listed_in	description	after_2020
0	s1	Movie	Dick Johnson Is Dead	...	Documentaries	As her father nears the end of his life, filmm...	False
1	s2	TV Show	Blood & Water	...	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	True
4	s5	TV Show	Kota Factory	...	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...	True
...	...	...	...	...	...	...	...

	show_id	type	title	...	listed_in	description	after_2020
8804	s8805	Movie	Zombieland	...	Comedies, Horror Movies	Looking to survive in a world taken over by zo...	False
8805	s8806	Movie	Zoom	...	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...	False
8806	s8807	Movie	Zubaan	...	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...	False

7976 rows × 13 columns

9. `df.drop()` to drop specific columns

In [51]:

```
# Drop columns
df = df.drop(columns = ['cast','description'])
df
```

Out[51]:

	show_id	type	title	...	duration	listed_in	after_2020
0	s1	Movie	Dick Johnson Is Dead	...	90 min	Documentaries	False
1	s2	TV Show	Blood & Water	...	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	True
4	s5	TV Show	Kota Factory	...	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	True
...	...	...	...	...	...	...	...
8804	s8805	Movie	Zombieland	...	88 min	Comedies, Horror Movies	False
8805	s8806	Movie	Zoom	...	88 min	Children & Family Movies, Comedies	False
8806	s8807	Movie	Zubaan	...	111 min	Dramas, International Movies, Music & Musicals	False

7976 rows × 11 columns

10. `df.groupby().sum()` to perform aggregation

In [52]:

```
# Aggregate all rows by "rating" and find the median release year for each rating
df.groupby('rating').median()['release_year']

# You can also use:
# df.groupby('col').sum()
# df.groupby('col').mean()
# df.groupby('col').max()
# df.groupby('col').min()
```

Out[52]:

```
rating
66 min      2015.0
74 min      2017.0
84 min      2010.0
              ...
TV-Y7       2017.0
```

```
TV-Y7-FV    2014.0
UR          2008.0
Name: release_year, Length: 17, dtype: float64
```

11. `pd.pivot table(columns = ..., values = ...., aggfunc = ...)` to create a pivot table

In [4]:

```
# Create a pivot table
# with rating as column
# and find the median release_year for each rating
pivot = pd.pivot_table(columns = 'rating',
              values = 'release_year',
              data = df,
              aggfunc = 'median')
pivot
```

Out[4]:

	release_year
rating	
66 min	2015
74 min	2017
84 min	2010
G	2004
NC-17	2014
NR	2015
PG	2012
PG-13	2011
R	2014
TV-14	2017
TV-G	2018
TV-MA	2018
TV-PG	2017
TV-Y	2018
TV-Y7	2017
TV-Y7-FV	2015
UR	2008

12. `df.transpose()` to transpose table

In [13]:

```
pivot = pivot.transpose()
pivot
```

Out[13]:

	release_year
rating	
66 min	2015
74 min	2017
84 min	2010
G	2004
NC-17	2014
NR	2015
PG	2012
PG-13	2011
R	2014
TV-14	2017
TV-G	2018
TV-MA	2018
TV-PG	2017
TV-Y	2018
TV-Y7	2017
TV-Y7-FV	2015
UR	2008

13. `df1.merge(df2)` to join two tables

In [14]:

```
# Join two tables together
merged_df = df.merge(pivot,
        how = 'left',
        left_on = 'rating',
        right_index=True,
        suffixes = ('','_median_per_rating')
        )
```

14. `df.rename({old_col:new_col})` to rename column

In [18]:

```
# Rename column show_id to id
df = df.rename({'show_id':'id'}, axis = 1)
df
```

Out[18]:

	id	type	title	director	cast	country	date_added	release_year	rating	duration
			Dick							

	id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Do
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	T
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	T
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	
...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	H
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Fa
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	M

8807 rows × 12 columns

15: `df.to_csv()` to export data

In [15]:

```
# Export data as a CSV file
merged_df.to_csv('file.csv')
```

Master these functions to get started with Pandas.

By Travis Tang