

Data Science Foundation

Lesson #6 - Exploratory Data Analysis II

Ivanovitch Silva
September, 2017



Agenda

- Case study: movie ratings
- Histogram and Box Plots
- Wrapper from Pandas to Matplotlib
- Guided project: EDA for academic performance

Update the repository

```
git clone https://github.com/ivanovitchm/EEC2006.git
```

Or

```
git pull
```

Case study: movie ratings

Metacritic

metacritic Movies Games TV Music Features

New Releases Coming Soon High Scores Browse A-Z Publications Critics People Trailers

Avengers: Age of Ultron

Walt Disney Studios Motion Pictures | Release Date: May 1, 2015

Summary Critic Reviews User Reviews Details & Credits Trailers & Videos

66 Metascore
Generally favorable reviews based on 49 Critics

7.1 User Score
Generally favorable reviews based on 2133 Ratings

What's this?

Starring: Chris Evans, Chris Hemsworth, James Spader, Jeremy Renner, Mark Ruffalo, Robert Downey Jr., Samuel L. Jackson, Scarlett Johansson

Director: Josh Whedon

Genre(s): Action, Adventure, Sci-Fi, Thriller, Fantasy

Rating: PG-13

Runtime: 141 min

[More Details and Credits](#)

Summary: When Tony Stark tries to jumpstart a dormant peacekeeping program, things go awry and Earth's Mightiest Heroes, including Iron Man, Captain America, Thor, The Incredible Hulk, Black Widow and Hawkeye, are put to the ultimate test as the fate of the planet hangs in the balance. [Expand](#)

Rotten Tomatoes

Rotten Tomatoes Search movies, TV, actors, more... MOVIE

TRENDING ON RT Fall TV Scorecard Magnificent Seven The Exorcist TV Reboots We Want

HOME > MARVEL CINEMATIC UNIVERSE > AVENGERS: AGE OF ULTRON

AVENGERS: AGE OF ULTRON (2015)

Part of the Collection: Marvel Cinematic Universe [View Collection](#)

TOMATOMETER **75%**
Average Rating: 6.7/10
Reviews Counted: 309
Fresh: 233
Rotten: 76

AUDIENCE SCORE **84%**
Average Rating: 4.2/5
User Ratings: 282,794

Critic Consensus: Exuberant and eye-popping, *Avengers: Age of Ultron* serves as an over-the-top but mostly satisfying sequel, rewarding its predecessor's unwieldy cast with a few new additions and a worthy foe.

ADD YOUR RATING

★★★★★

IMDb

IMDb Find Movies, TV shows, celebrities and more... at

Movies, TV & Showtimes Critics, Events & Photos News & Community Watchlist

Enjoy unlimited streaming on Prime Video
Includes thousands of titles. Monthly plans now available. [Start your 30-day free trial](#)

FULL CAST AND CREW TRIVIA USER REVIEWS | IMDbPro | MORE W SHARE

Avengers: Age of Ultron (2015)

PG-13 | 2h 23min | Action, Adventure, Sci-Fi | 1 May 2015 (USA)

7.5 Rate This
472,361

Trailer

33 VIDEOS 141 PHOTOS

Fandango

FANDANGO Enter City, State, ZIP Code, or Movie GO

AVENGERS: AGE OF ULTRON (2015)

OVERVIEW MOVIE TIMES & TICKETS SYNOPSIS MOVIE REVIEWS TRAILER

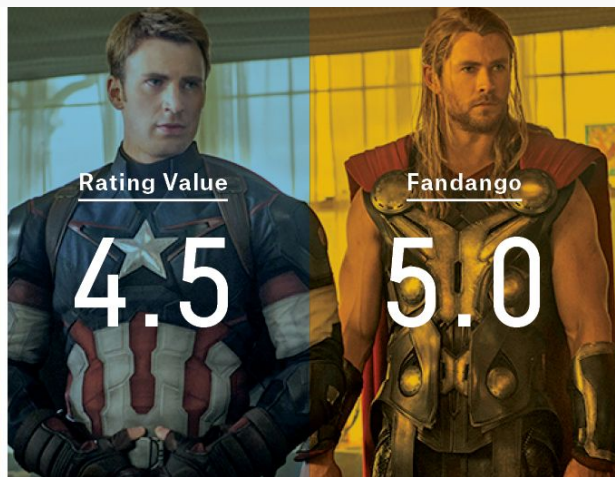
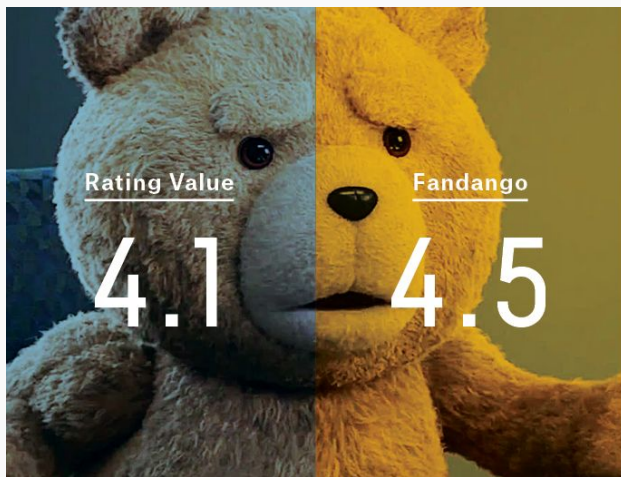
Released MAY 1, 2015

PG-13 - 2 hr 21 min
Action/Adventure
Family

★★★★★
15,861 Fan Ratings

GLOBAL A

Bias in movie ratings



Reviewing the dataset

	FILM	RT_user_norm	Metacritic_user_nom	IMDB_norm	Fandango_Ratingvalue	Fandango_Stars
0	Avengers: Age of Ultron (2015)	4.3	3.55	3.90	4.5	5.0
1	Cinderella (2015)	4.0	3.75	3.55	4.5	5.0
2	Ant-Man (2015)	4.5	4.05	3.90	4.5	5.0
3	Do You Believe? (2015)	4.2	2.35	2.70	4.5	5.0
4	Hot Tub Time Machine 2 (2015)	1.4	1.70	2.55	3.0	3.5

<https://github.com/fivethirtyeight/data/tree/master/fandango>

Frequency Distribution

Frequency Distribution
(sorted by **frequency** in
descending order)

Value	Frequency
4.1	16
4.2	12
3.9	12
4.3	11
3.7	9
3.5	9
4.5	9
3.4	9
3.6	8
4.4	7
4.0	7
3.2	5
2.9	5
3.8	5
3.3	4
4.6	4
3.0	4
4.8	3
3.1	3
2.8	2
2.7	2

Name: Fandango_Ratingvalue,
dtype: int64

Frequency Distribution
(sorted by **unique value** in
ascending order)

Value	Frequency
2.7	2
2.8	2
2.9	5
3.0	4
3.1	3
3.2	5
3.3	4
3.4	9
3.5	9
3.6	8
3.7	9
3.8	5
3.9	12
4.0	7
4.1	16
4.2	12
4.3	11
4.4	7
4.5	9
4.6	4
4.8	3

Name: Fandango_Ratingvalue,
dtype: int64

Binning

Fandango
Frequency
Distribution

		Bins	Count
2.7	2	0.0 - 0.5	0
2.8	2	0.5 - 1.0	0
2.9	5	1.0 - 1.5	0
3.0	4	1.5 - 2.0	0
3.1	3	2.0 - 2.5	0
3.2	5	2.5 - 3.0	9
3.3	4	3.0 - 3.5	25
3.4	9	3.5 - 4.0	43
3.5	9	4.0 - 4.5	53
3.6	8	4.5 - 5.0	16
3.7	9		
3.8	5		
3.9	12		
4.0	7		
4.1	16		
4.2	12		
4.3	11		
4.4	7		
4.5	9		
4.6	4		
4.8	3		

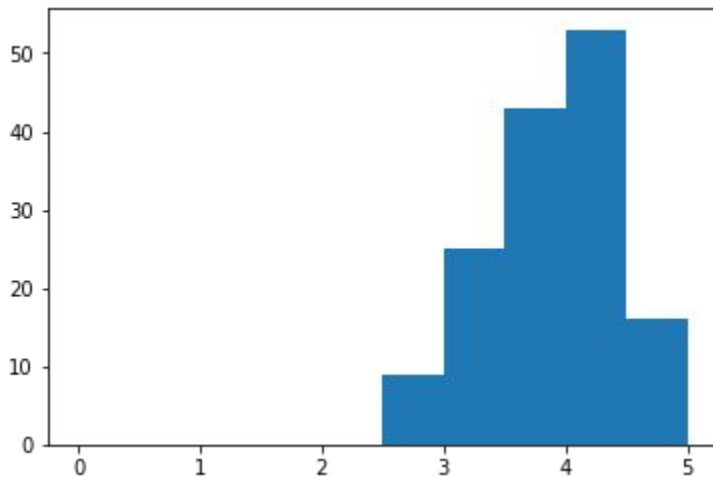
Rotten Tomatoes
Frequency
Distribution

		Bins	Count
2.00	1	0.0 - 0.5	0
2.10	1	0.5 - 1.0	0
2.15	1	1.0 - 1.5	0
2.20	1	1.5 - 2.0	0
2.30	2	2.0 - 2.5	8
2.45	2	2.5 - 3.0	20
2.50	1	3.0 - 3.5	50
2.55	1	3.5 - 4.0	58
2.60	2	4.0 - 4.5	10
2.70	4	4.5 - 5.0	0
2.75	5		
2.80	2		
2.85	1		
2.90	1		
2.95	3		
...	..		
4.00	1		
4.05	1		
4.10	4		
4.15	1		
4.20	2		
4.30	1		

truncated
to save
space

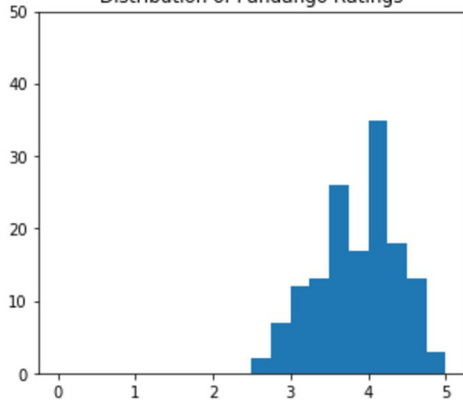
Histogram in Matplotlib

```
ax.hist(norm_reviews['Fandango_Ratingvalue'], range=(0, 5))
```



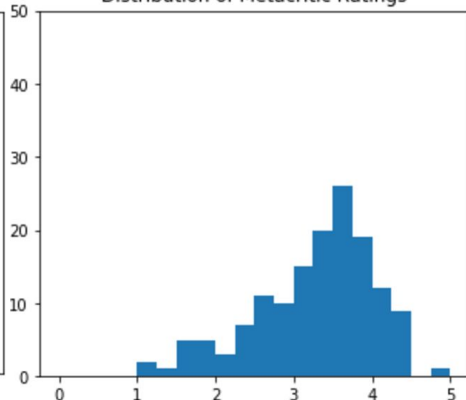
Comparing histograms

Distribution of Fandango Ratings



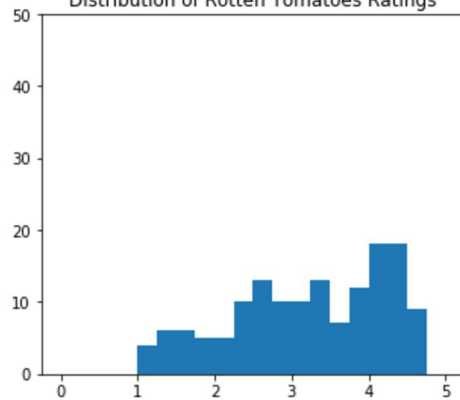
Around 50% of user ratings fall in the 2 to 4 score range

Distribution of Metacritic Ratings



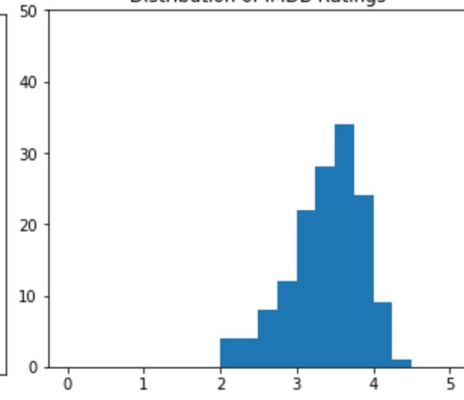
Around 75% of user ratings fall in the 2 to 4 score range

Distribution of Rotten Tomatoes Ratings



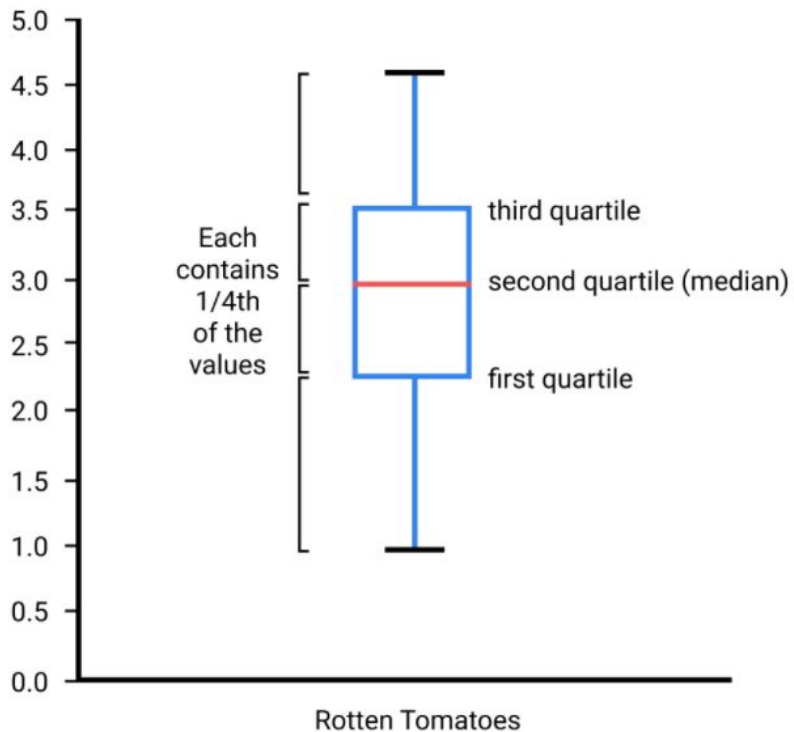
Around 50% of user ratings fall in the 2 to 4 score range

Distribution of IMDB Ratings



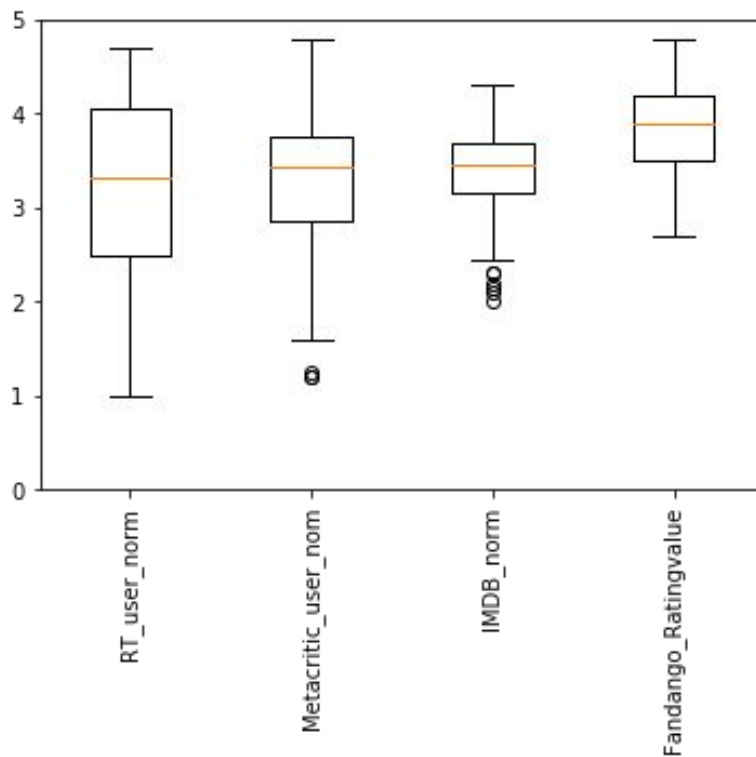
Around 90% of user ratings fall in the 2 to 4 score range

Quartis and Box Plot



```
ax.boxplot(norm_reviews[ 'RT_user_norm' ] )
```

Multiple Box Plot





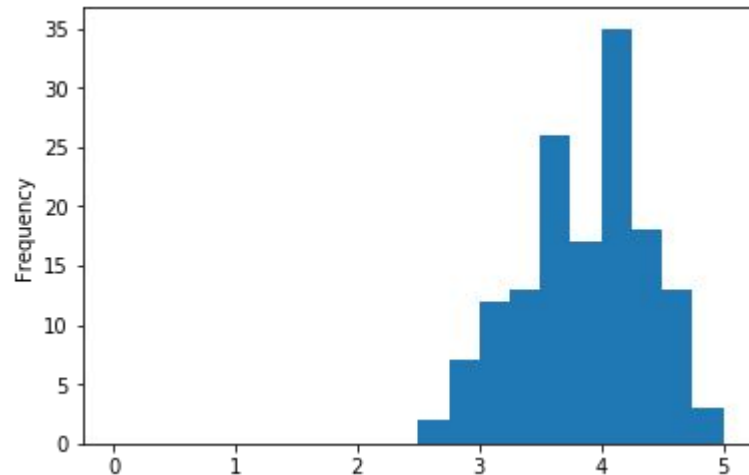
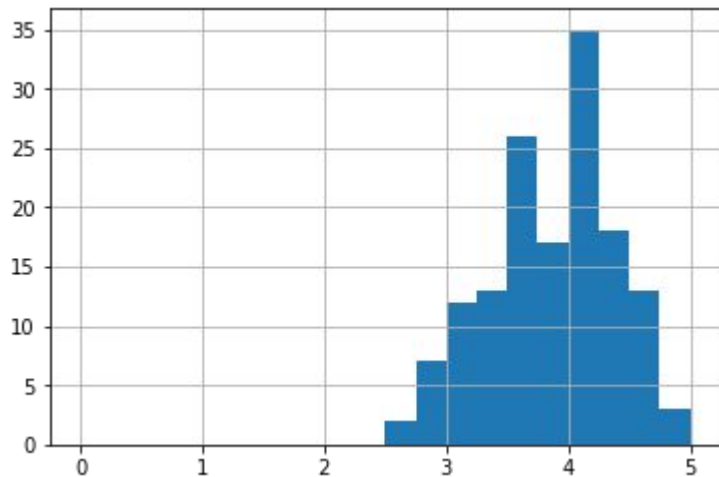
Exploratory Data Analysis IV - Histogram and Boxplot.ipynb

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

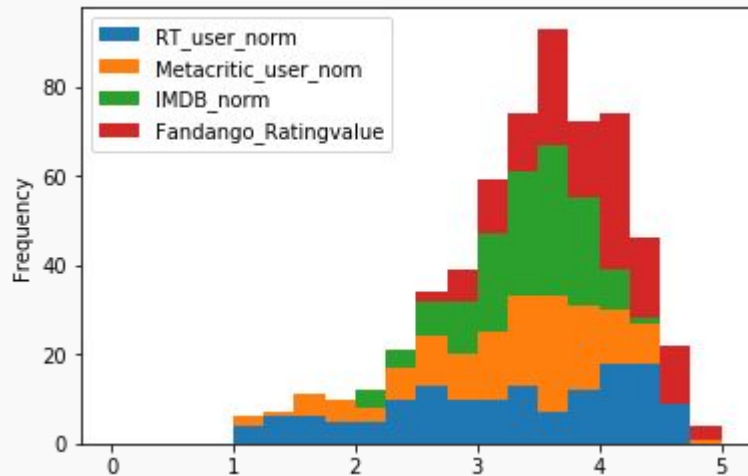
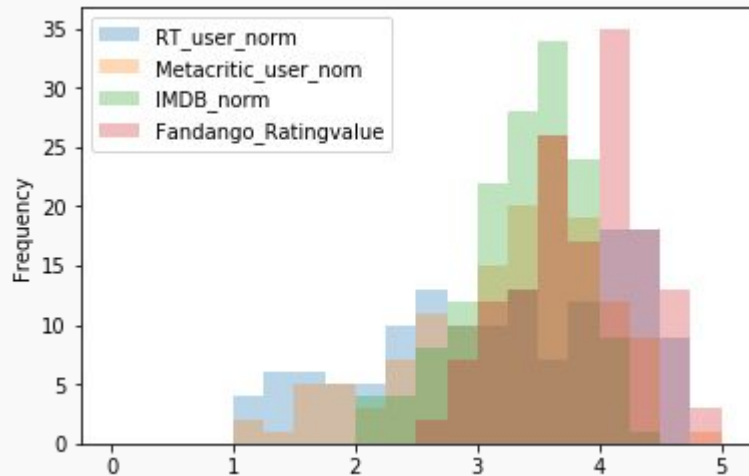


matplotlib



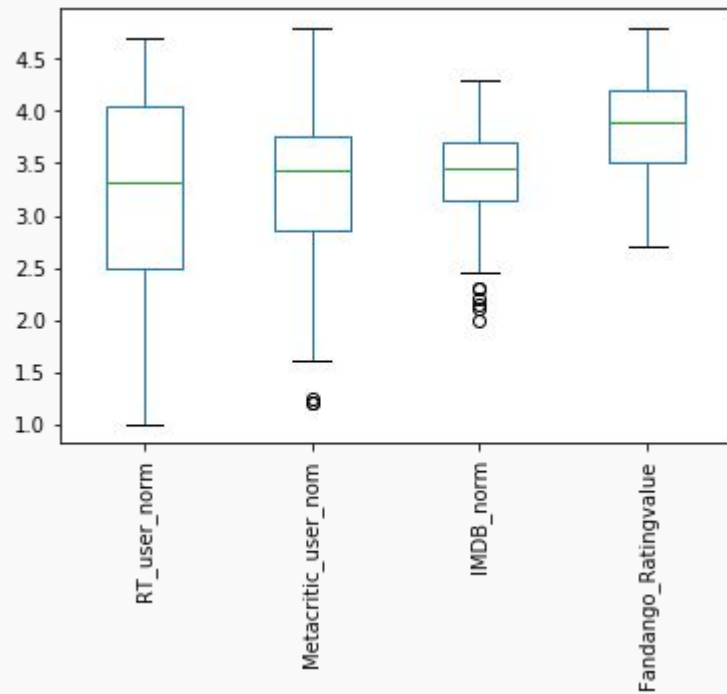
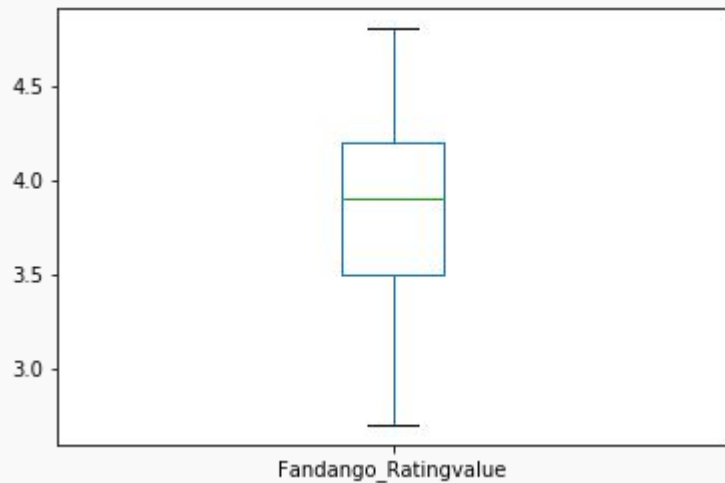
```
# Enable matplotlib plot inline  
%matplotlib inline  
norm_reviews.Fandango_Ratingvalue.hist(bins=20, range=(0,5))
```

```
# other way to do the same thing  
norm_reviews.Fandango_Ratingvalue.plot(kind='hist', bins=20, range=(0,5));
```

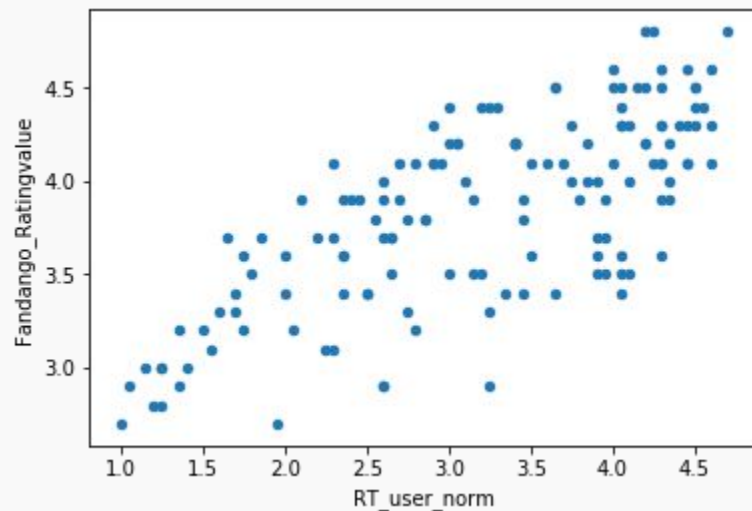
```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), alpha=0.3);
```

```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), stacked=True);
```



```
norm_reviews.Fandango_Ratingvalue.plot(kind='box')
```

```
norm_reviews.plot(kind='box', rot=90)
```



```
norm_reviews.plot(kind='scatter',x='RT_user_norm', y='Fandango_Ratingvalue')
```



Exploratory Data Analysis V - Plotting with Pandas.ipynb

Guided Project: academic performance

	a_ID	CEP	ano_ingresso	periodo_ingresso	status	ano_disciplina	periodo_disciplina	nota	disciplina_ID	status.disciplina	enen-nota
0	0	59015430	2014	1	CANCELADO	2014	2	2.6	0	Reprovado	618.0
1	0	59015430	2014	1	CANCELADO	2015	1	8.0	0	Aprovado	618.0
2	1	59073120	2014	1	CANCELADO	2014	2	0.1	0	Reprovado	615.0
3	2	59072580	2014	1	ATIVO	2014	2	6.1	0	Aprovado	600.0
4	3	59088150	2014	1	ATIVO	2014	1	3.0	0	Reprovado	673.0