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

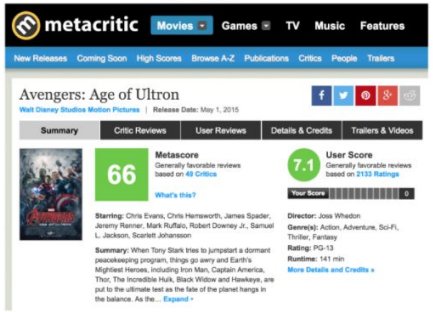
```
git clone https://github.com/ivanovitchm/EEC2006.git
```

Or ....

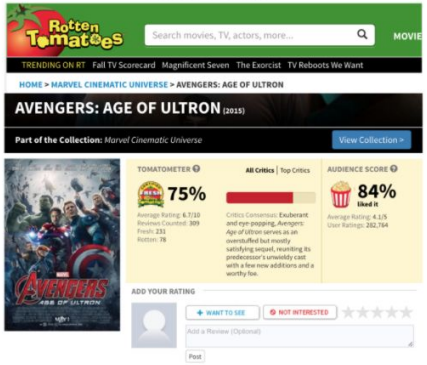
```
git pull
```

Case study: movie ratings

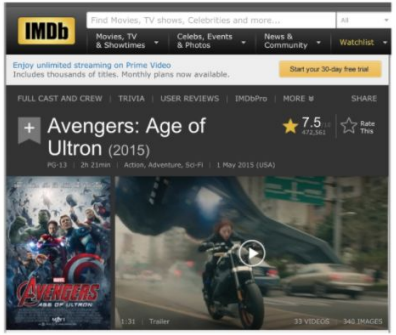
Metacritic



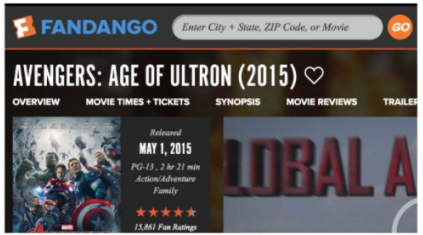
Rotten Tomatoes



IMDB



Fandango



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

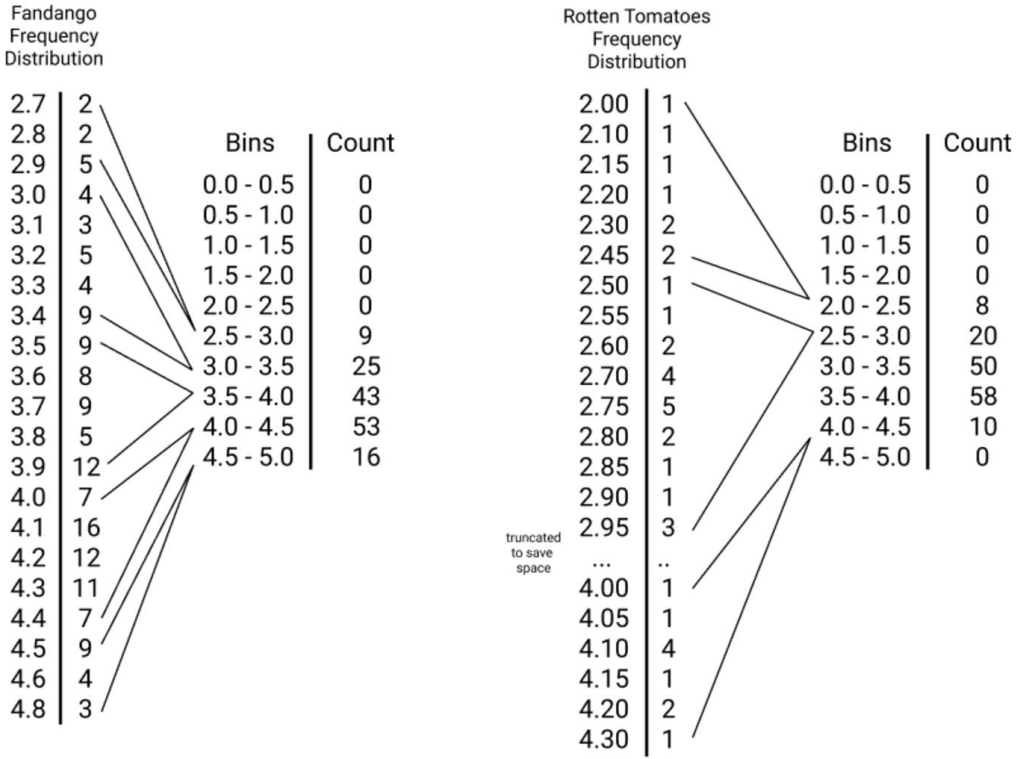
Name: Fandango_Ratingvalue, dtype: int64

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

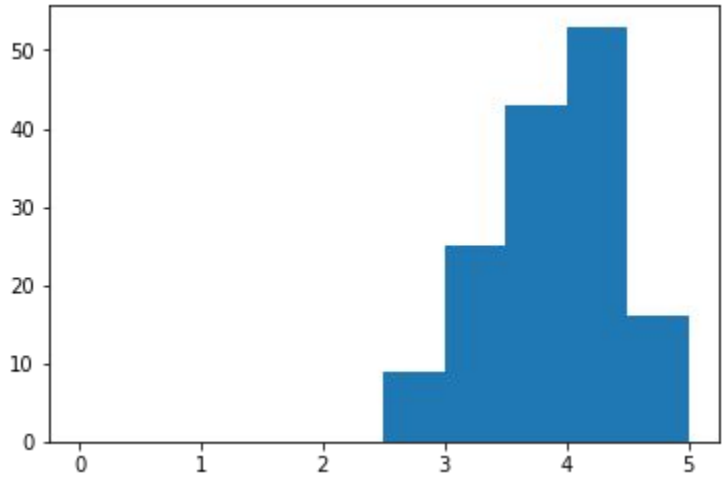
Name: Fandango_Ratingvalue, dtype: int64

Binning

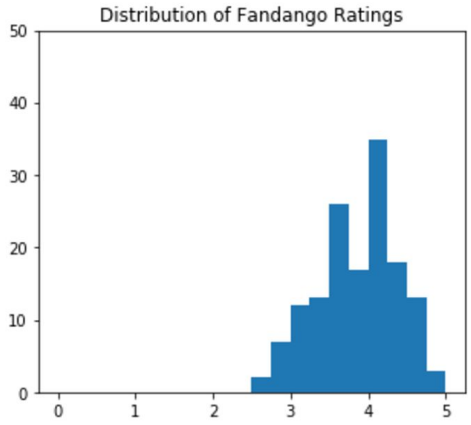

Fandango Frequency Distribution
2.7 2
2.8 2
2.9 5
3.0 4
3.1 3
3.2 5
3.3 4
3.4 9
3.5 9
3.6 8
3.7 9
3.8 5
3.9 12
4.0 7
4.1 16
4.2 12
4.3 11
4.4 7
4.5 9
4.6 4
4.8 3
Bins
Count
0.0 - 0.5 0
0.5 - 1.0 0
1.0 - 1.5 0
1.5 - 2.0 0
2.0 - 2.5 0
2.5 - 3.0 9
3.0 - 3.5 25
3.5 - 4.0 43
4.0 - 4.5 53
4.5 - 5.0 16
Rotten Tomatoes Frequency Distribution
2.00 1
2.10 1
2.15 1
2.20 1
2.30 2
2.45 2
2.50 1
2.55 1
2.60 2
2.70 4
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
Bins
Count
0.0 - 0.5 0
0.5 - 1.0 0
1.0 - 1.5 0
1.5 - 2.0 0
2.0 - 2.5 8
2.5 - 3.0 20
3.0 - 3.5 50
3.5 - 4.0 58
4.0 - 4.5 10
4.5 - 5.0 0

Histogram in Matplotlib

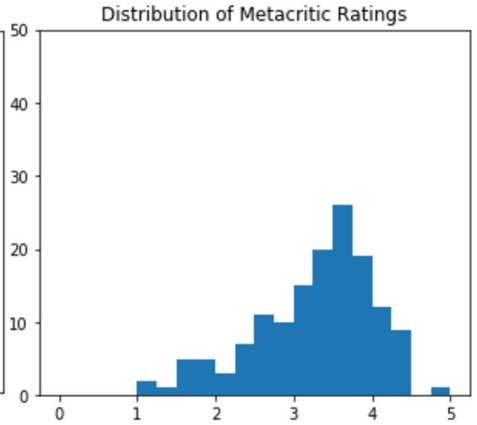
```
ax.hist(norm_reviews['Fandango_Ratingvalue'], range=(0, 5))
```



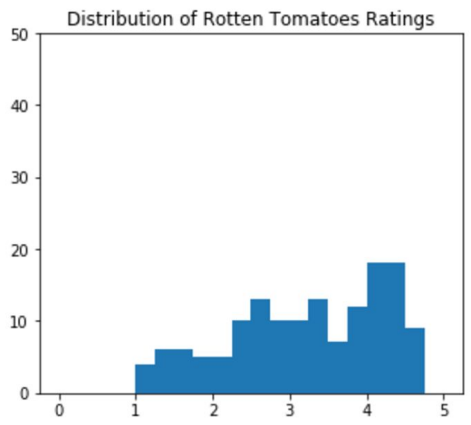
Comparing histograms



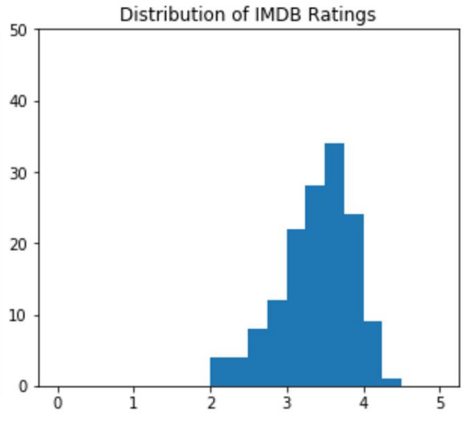
Around 50% of user ratings fall in the 2 to 4 score range



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

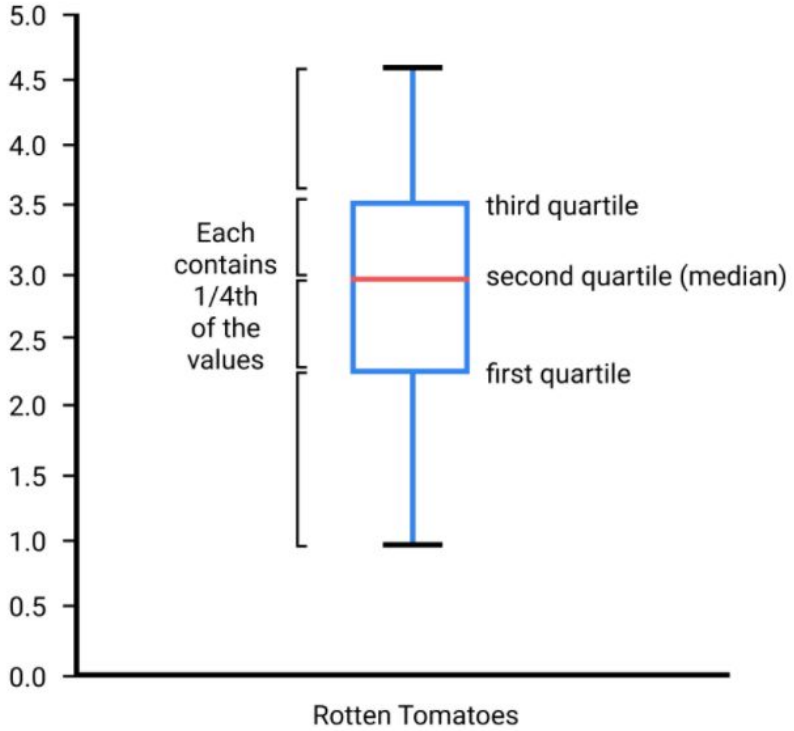


Around 50% of user ratings fall in the 2 to 4 score range



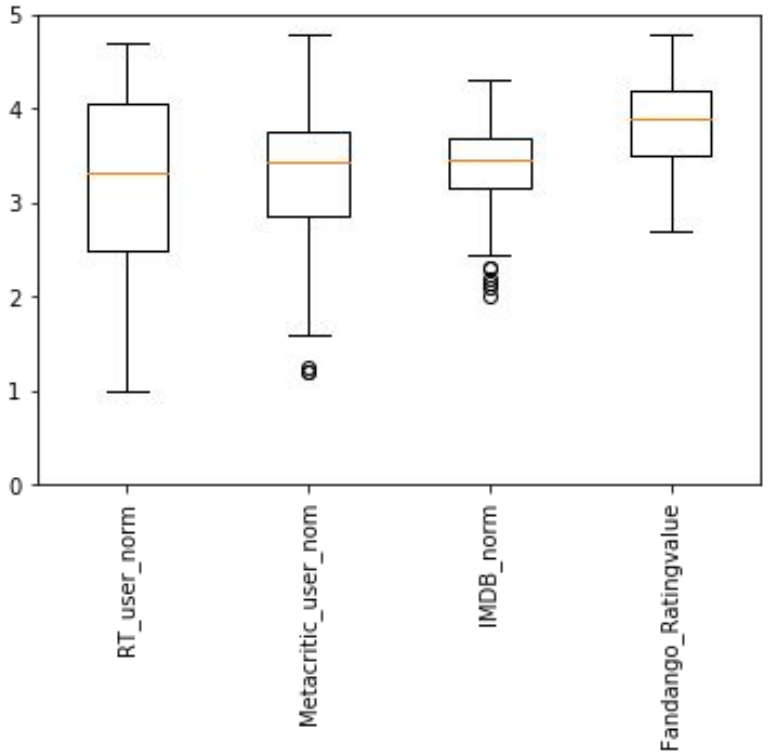
Around 90% of user ratings fall in the 2 to 4 score range

Quartis and Box Plot




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

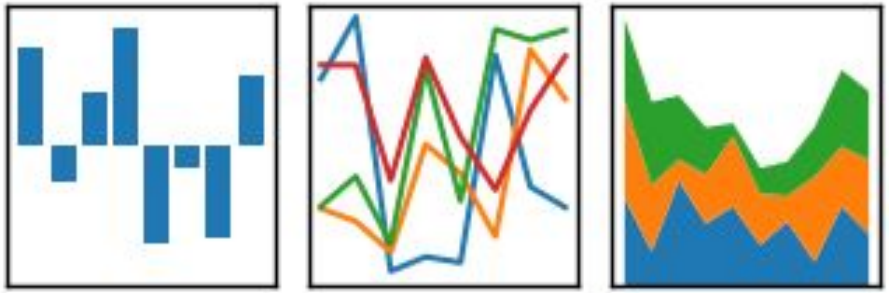
Multiple Box Plot



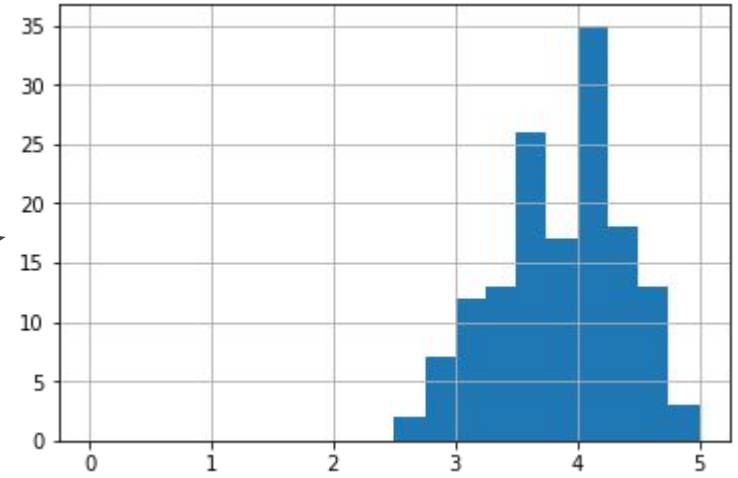
Exploratory Data Analysis IV - Histogram and Boxplot.ipynb

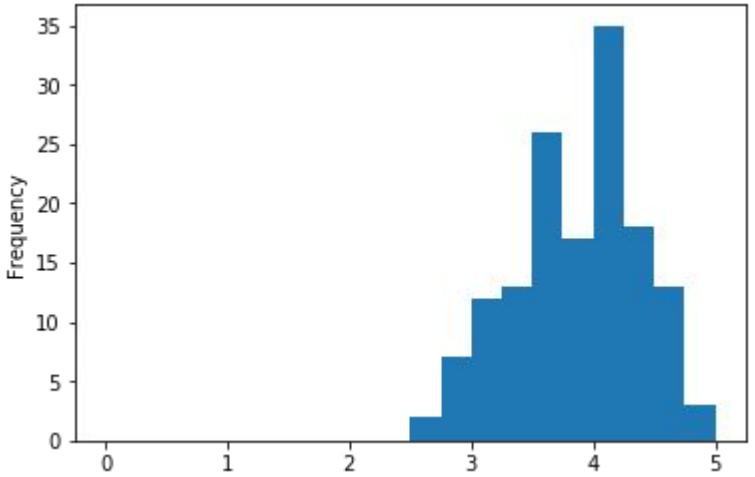
pandas

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



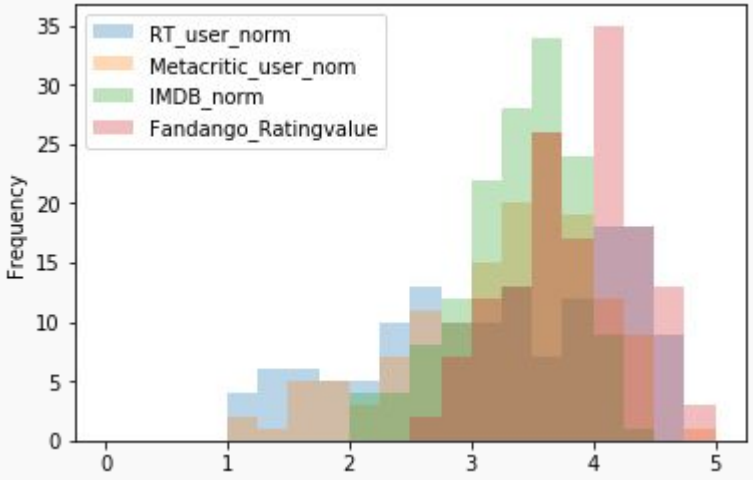
matplotlib

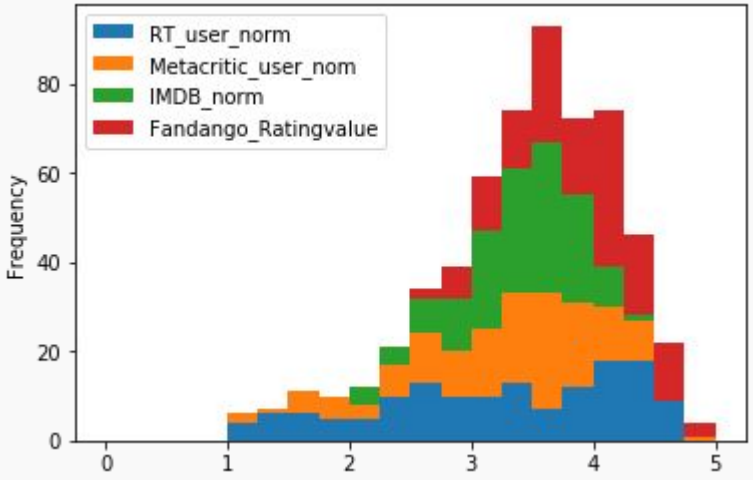





```
# Enable matplotlib plot inline
%matplotlib inline
norm_reviews.Fandango_Ratingvalue.hist(bins=20, range=(0,5))
```

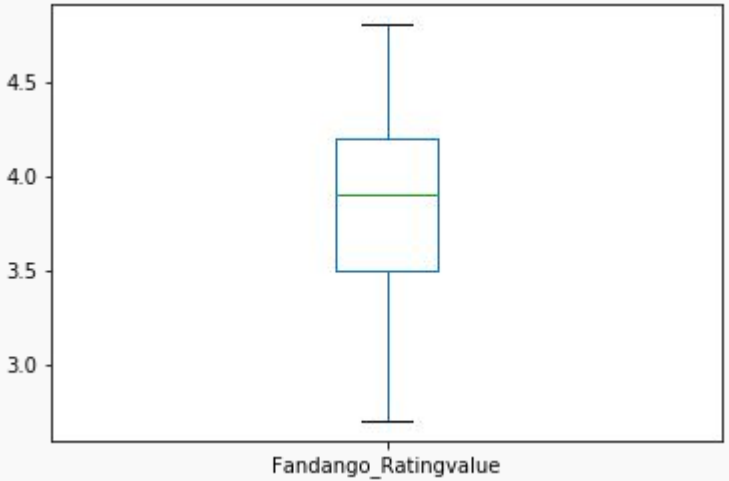
```
# other way to do the same thing
norm_reviews.Fandango_Ratingvalue.plot(kind='hist', bins=20, range=(0,5));
```

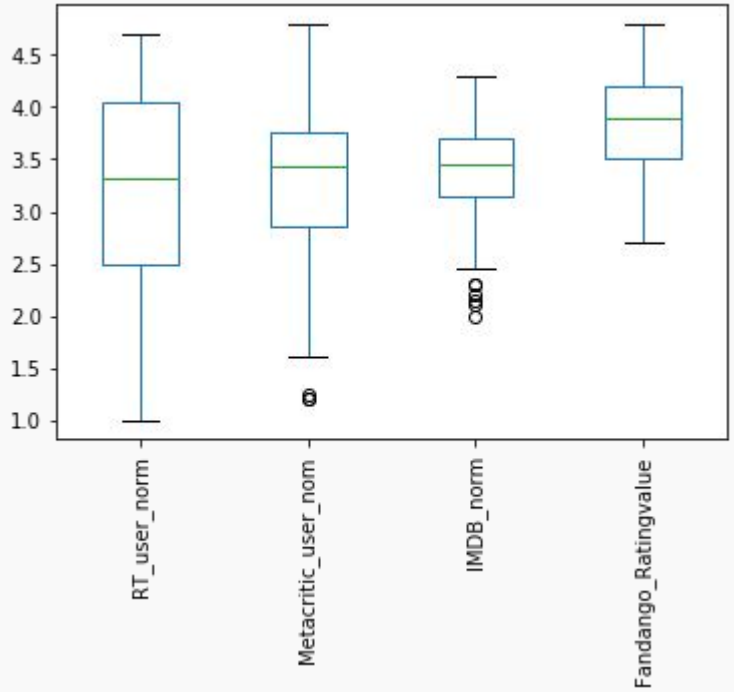





```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), alpha=0.3);
```

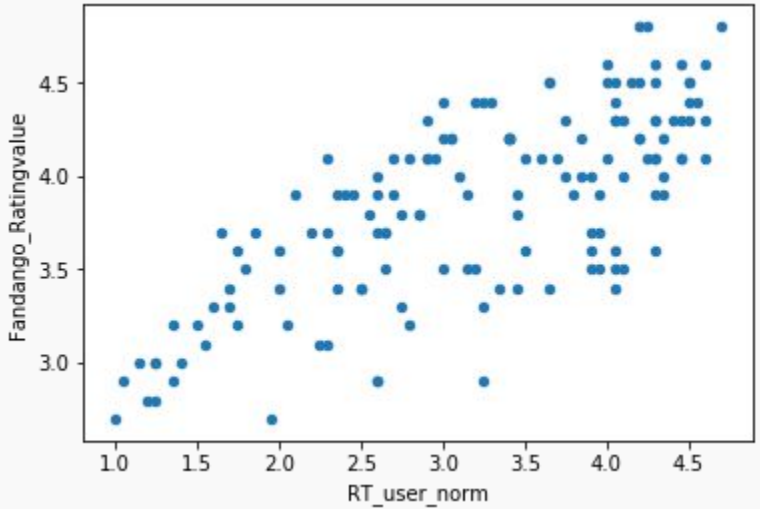
```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), stacked=True);
```






```
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