IMD0033 - Probabilidade

Aula 11 - Análise Exploratória de Dados III

Ivanovitch Silva

Abril, 2019

Agenda

- Estudo de caso: analisando filmes
- Motivação: enviesamento de dados
- Histograma e gráfico de caixa
- Interface entre Pandas & Matplotlib
- ??????????

Atualizar o repositório

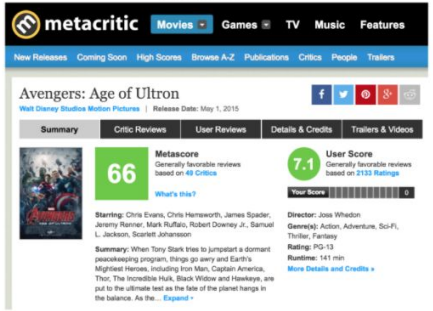
```
git clone https://github.com/ivanovitchm/imd0033_2019_1
```

Ou ....

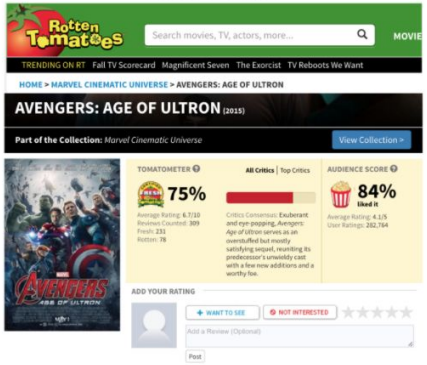
```
git pull
```

Estudo de caso: avaliando filmes

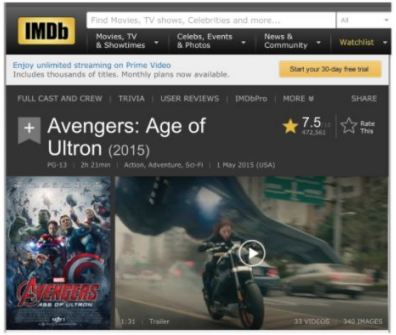
Metacritic



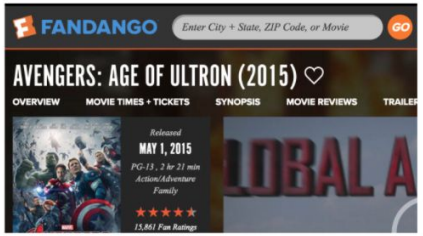
Rotten Tomatoes



IMDB



Fandango



Enviesamento de dados

Conjunto de dados

	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

Como comparar as diferentes avaliações?

Distribuição de frequências

Series.value_counts()

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

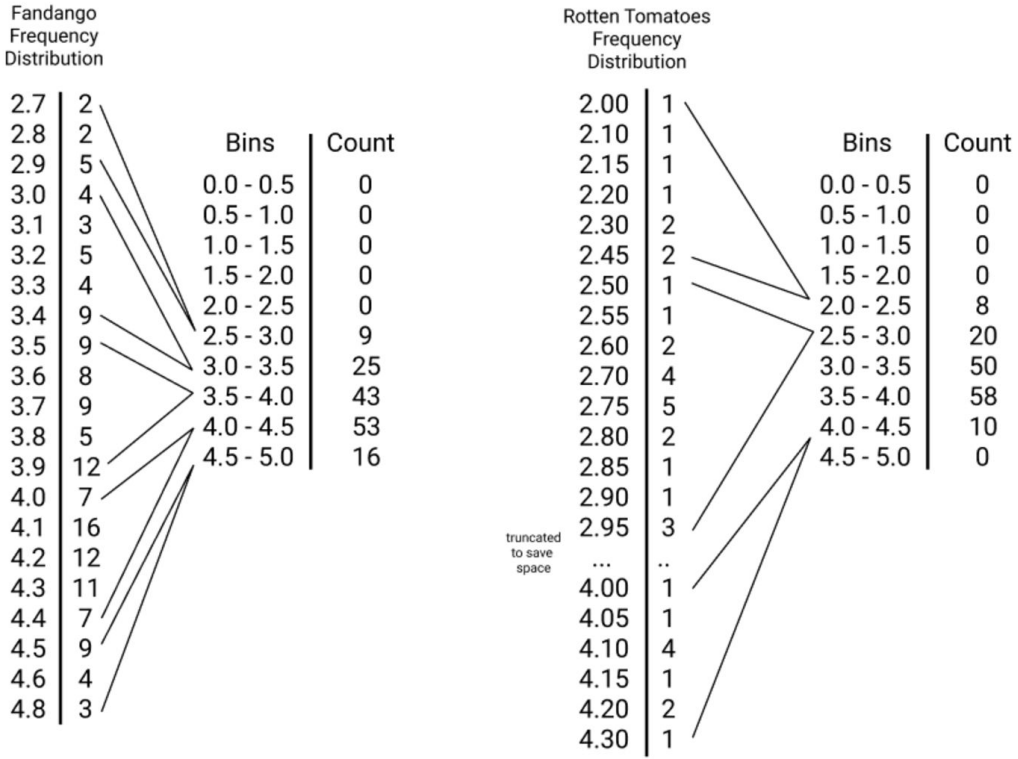
Series.sort_index()

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

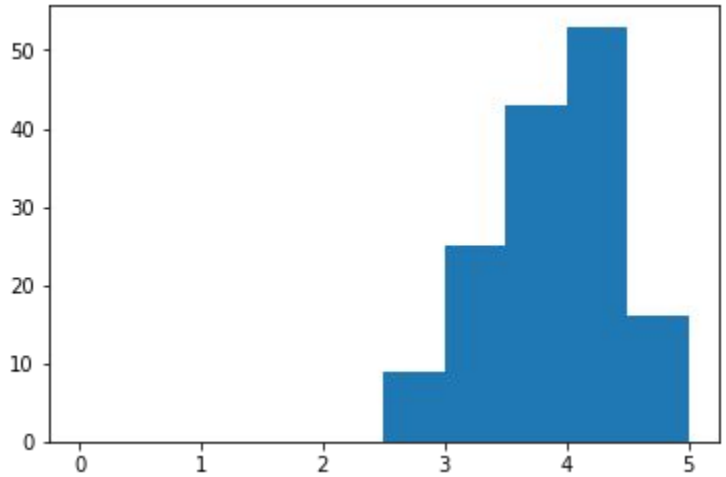
Name: Fandango_Ratingvalue, dtype: int64

Intervalos de agrupamento

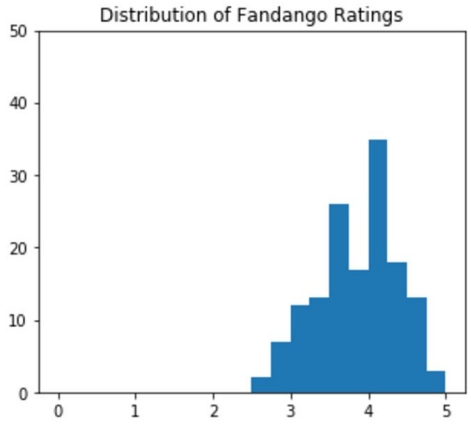


Histograma no Matplotlib

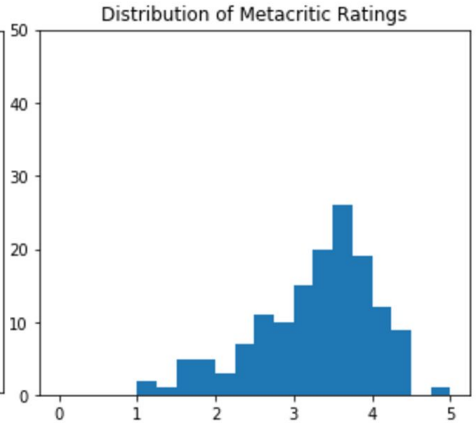
```
ax.hist(norm_reviews['Fandango_Ratingvalue'], range=(0, 5))
```



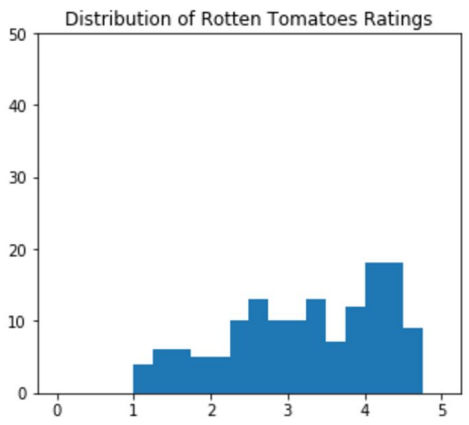
Comparando histogramas



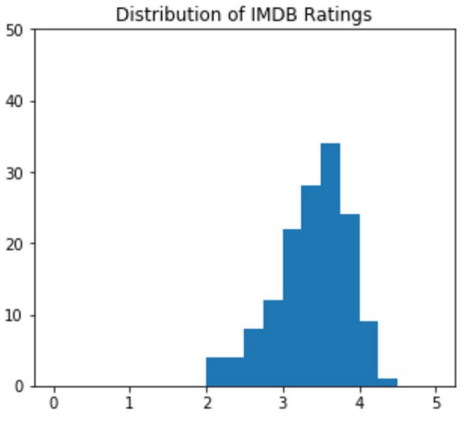
50% das avaliações estão entre 2 e 4



75% das avaliações estão entre 2 e 4

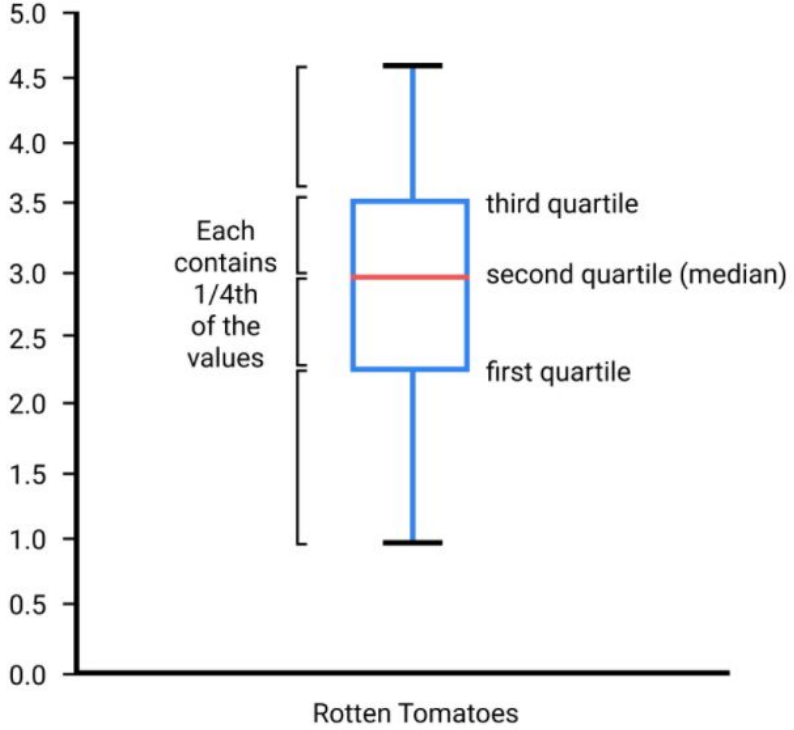


50% das avaliações estão entre 2 e 4



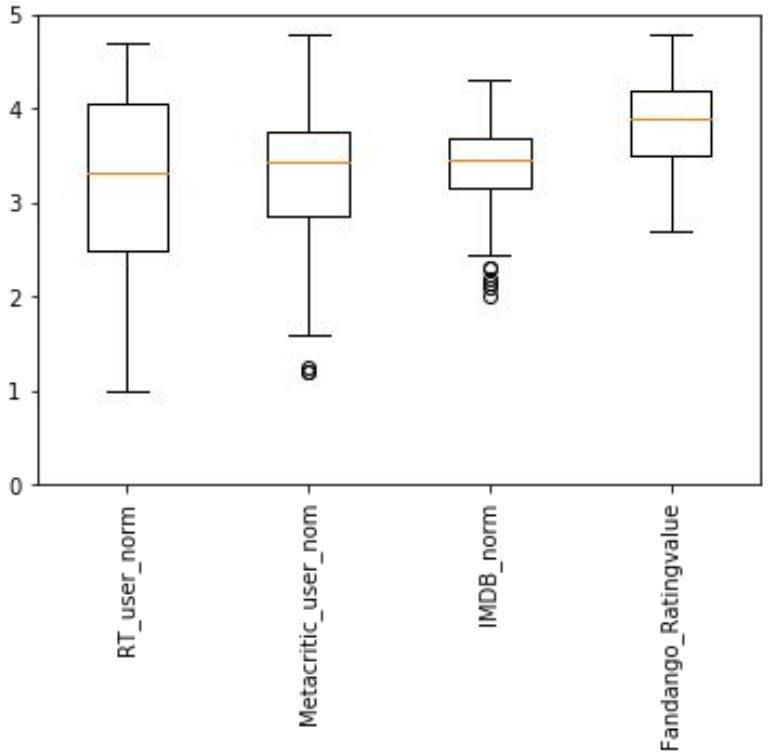
90% das avaliações estão entre 2 e 4

Quartil e gráfico de caixa

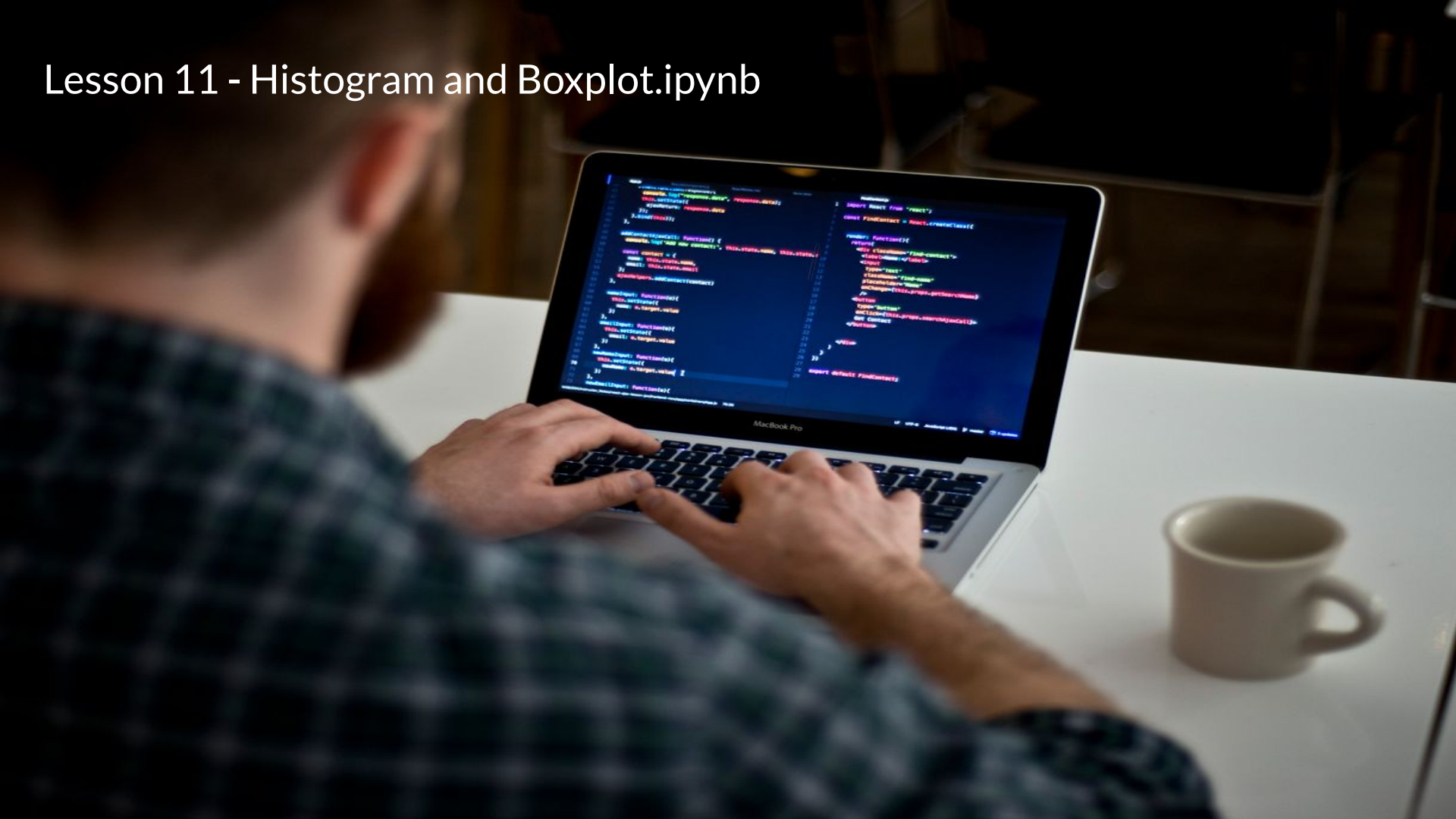



```
ax.boxplot(norm_reviews['RT_user_norm'])
```

Múltiplos gráficos de caixa

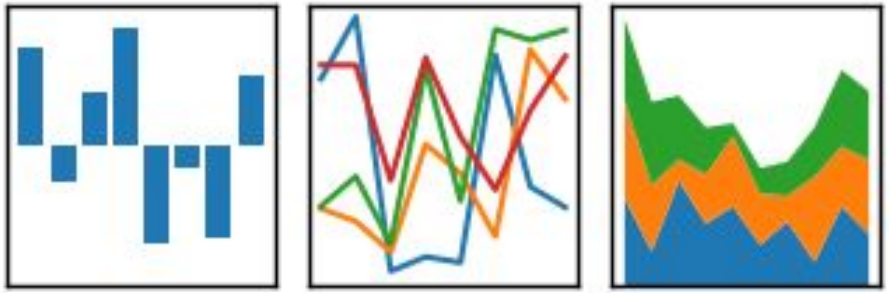


Lesson 11 - Histogram and Boxplot.ipynb

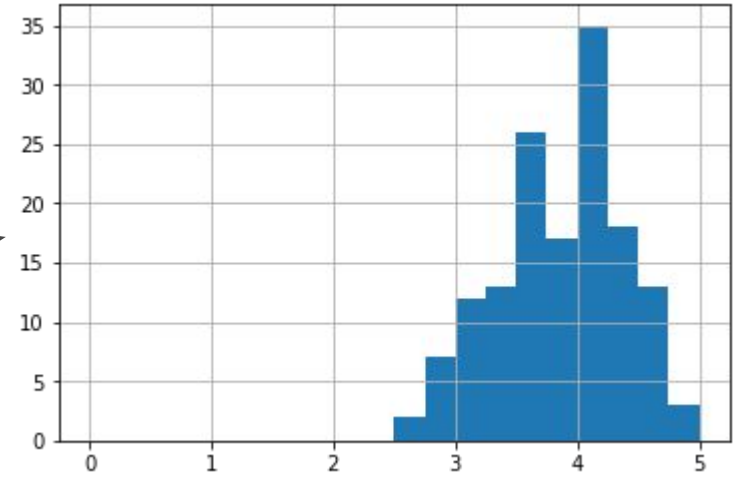


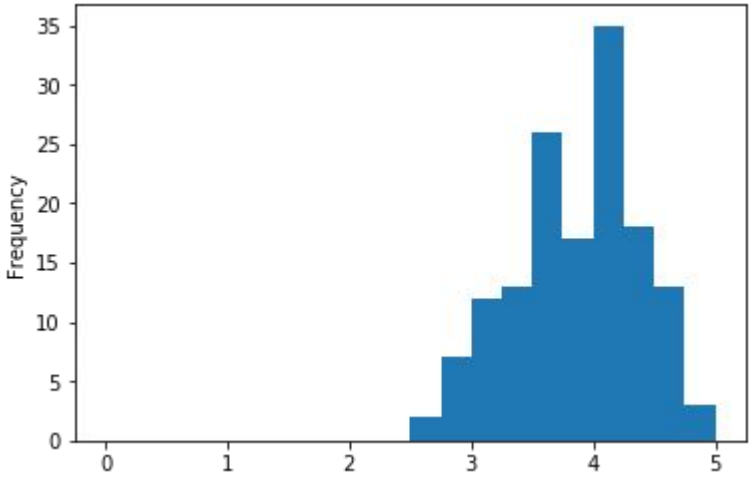
pandas

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



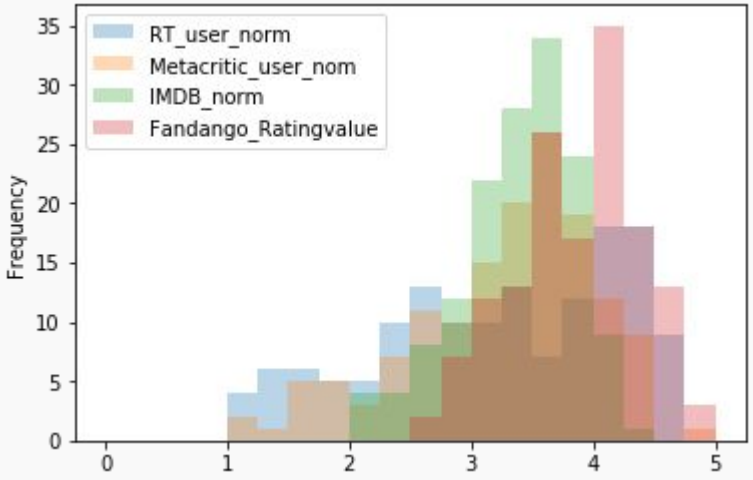
matplotlib

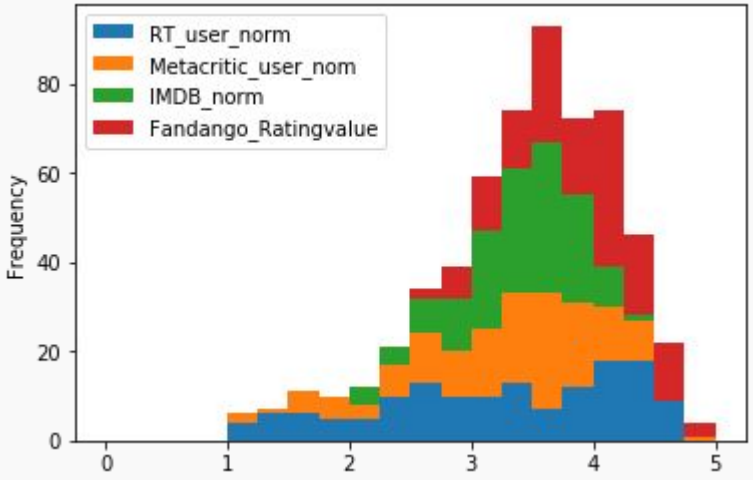





```
# Enable matplotlib plot inline
%matplotlib inline
norm_reviews.Fandango_Ratingvalue.hist(bins=20, range=(0,5))
```

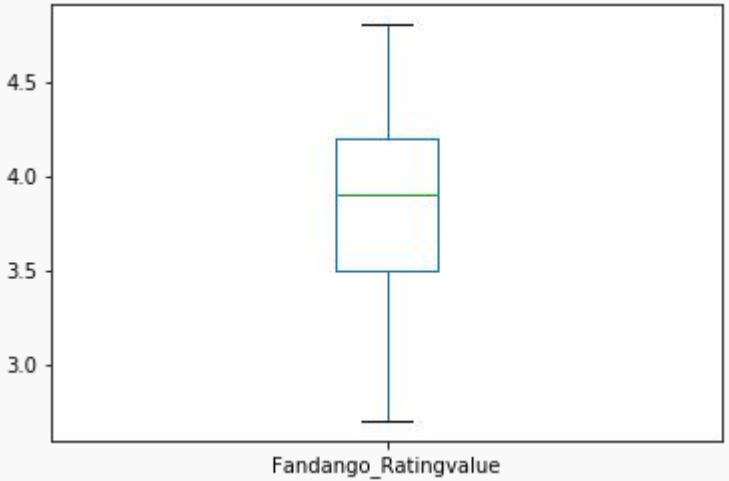
```
# other way to do the same thing
norm_reviews.Fandango_Ratingvalue.plot(kind='hist', bins=20, range=(0,5));
```

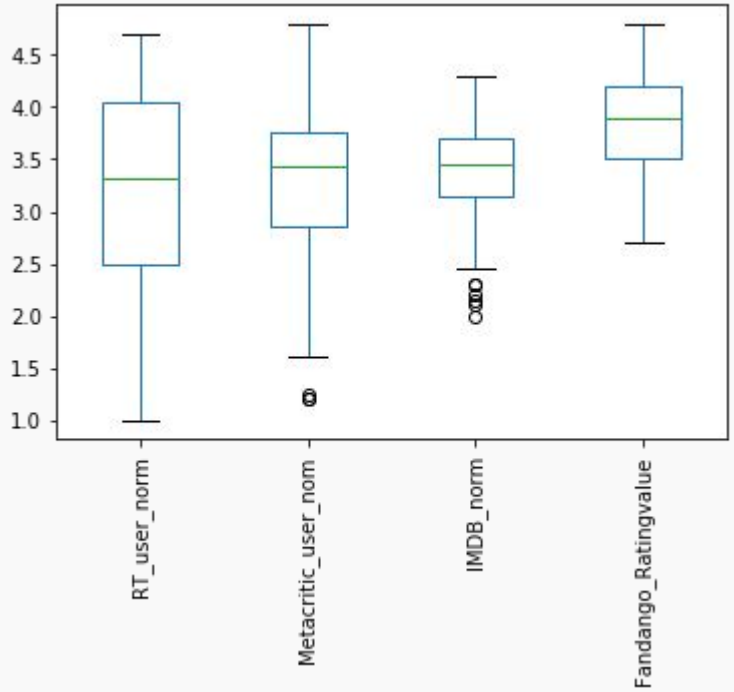





```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), alpha=0.3);
```

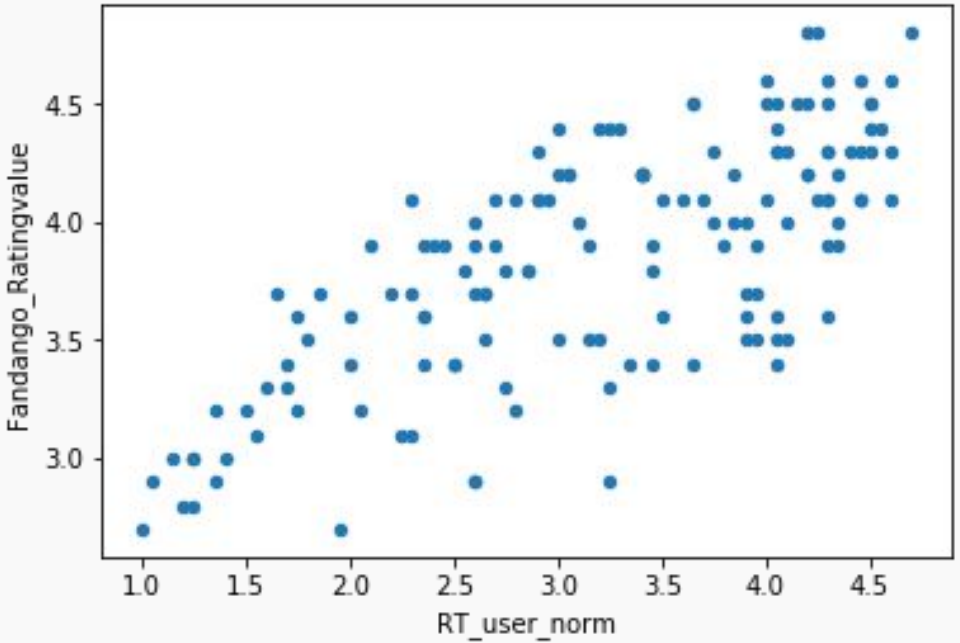
```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), stacked=True);
```



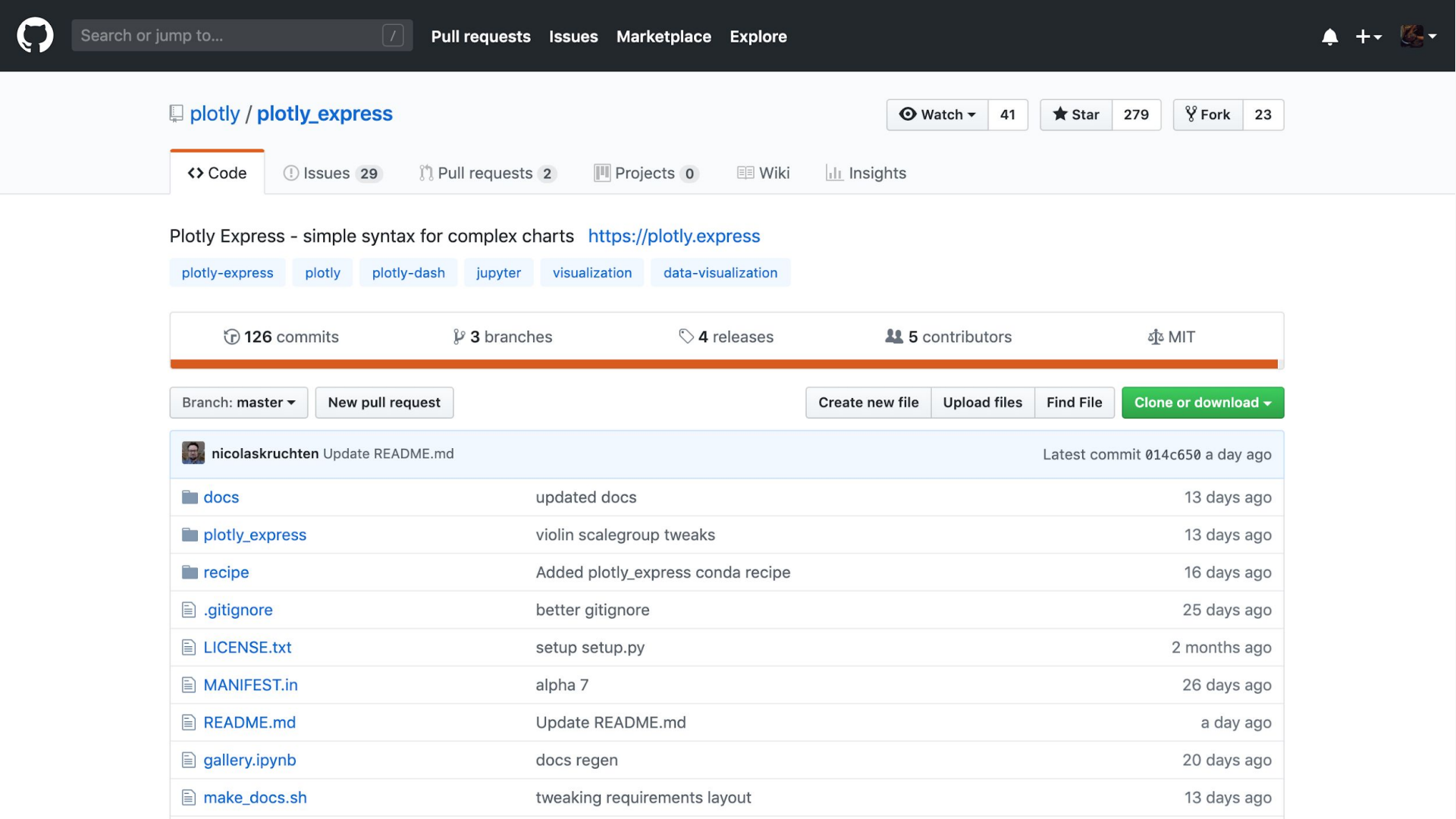



```
norm_reviews.Fandango_Ratingvalue.plot(kind='box')
```

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




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


Search or jump to...
Pull requests
Issues
Marketplace
Explore
plotly / plotly_express
Watch 41
Star 279
Fork 23
Code
Issues 29
Pull requests 2
Projects 0
Wiki
Insights
Plotly Express - simple syntax for complex charts https://plotly.express
plotly-express
plotly
plotly-dash
jupyter
visualization
data-visualization
126 commits
3 branches
4 releases
5 contributors
MIT
Branch: master
New pull request
Create new file
Upload files
Find File
Clone or download
nicolaskruchten Update README.md
Latest commit 014c650 a day ago
docs updated docs 13 days ago
plotly_express violin scalegroup tweaks 13 days ago
recipe Added plotly_express conda recipe 16 days ago
.gitignore better gitignore 25 days ago
LICENSE.txt setup setup.py 2 months ago
MANIFEST.in alpha 7 26 days ago
README.md Update README.md a day ago
gallery.ipynb docs regen 20 days ago
make_docs.sh tweaking requirements layout 13 days ago

Lesson #11 - Embedded Plotting with Pandas.ipynb