UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD
INSTITUTO
METRÓPOLE
DIGITAL

Data Science & ML Course

Lesson #7 Exploratory Data Analysis II

Ivanovitch Silva
October, 2018

Agenda

- Case study: movie ratings, economic guide to picking a college major
- Histogram and Box Plots
- Wrapper from Pandas to Matplotlib

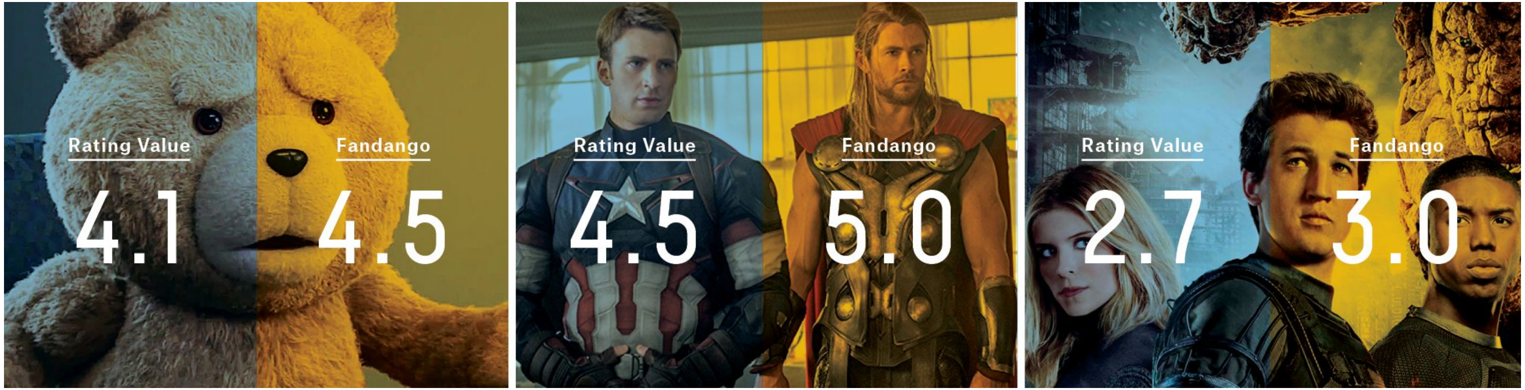
Update from repository

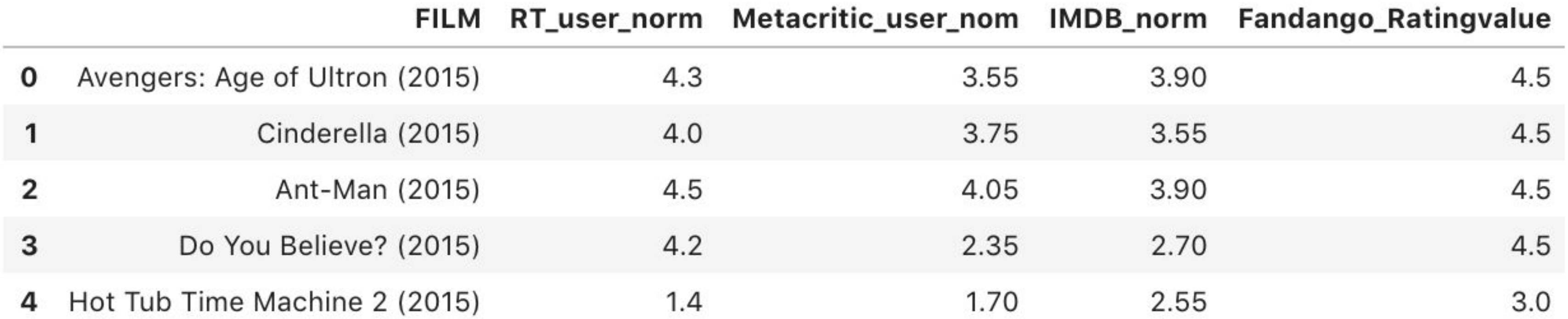
```
git clone https://github.com/ivanovitchm/datascience2machinelearning.git
```

Or ….

```
git pull
```

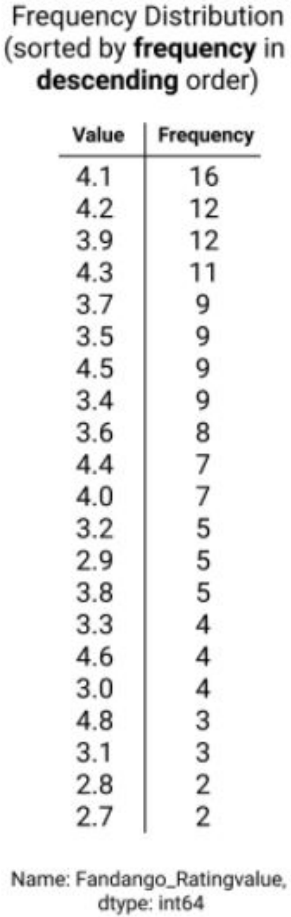
Case study: movie ratings





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

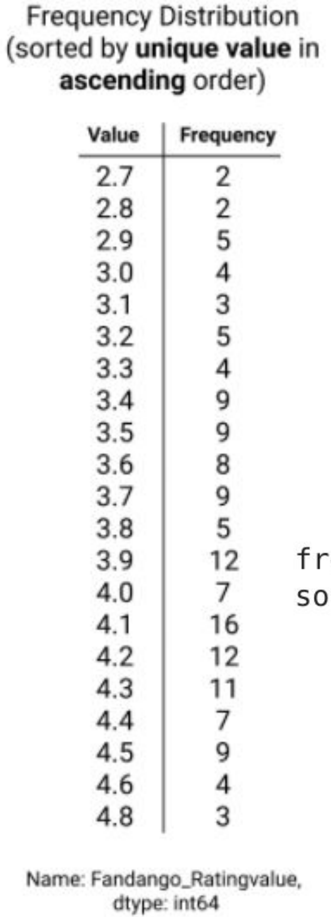
Frequency Distribution



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



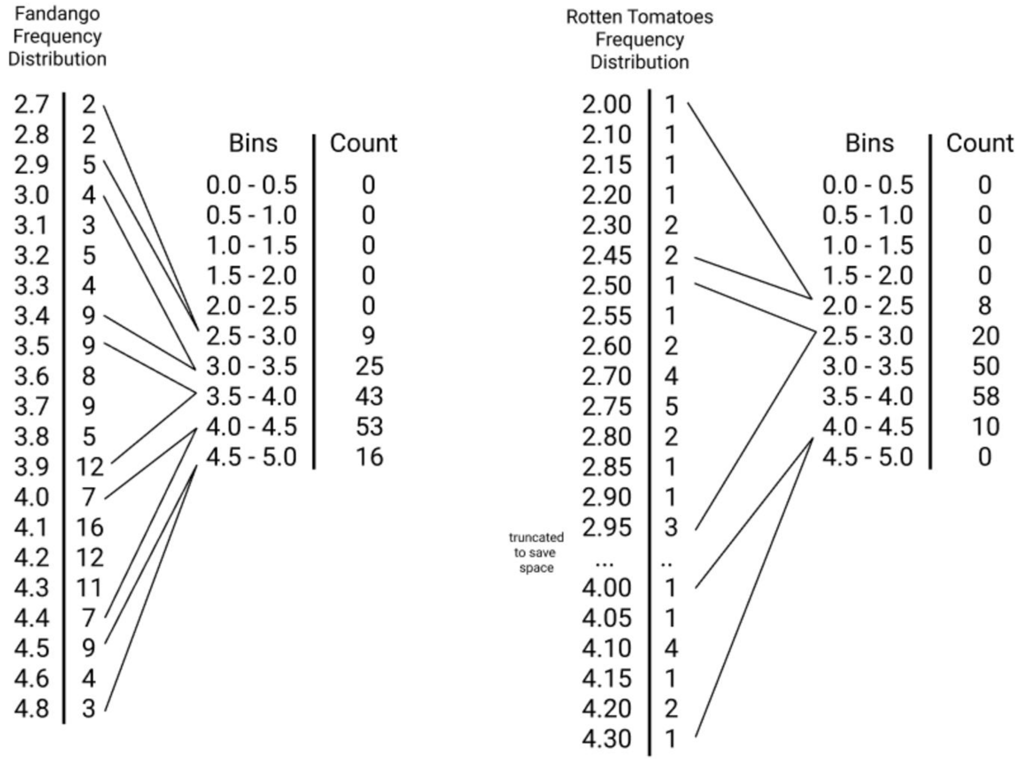
Frequency Distribution (sorted by **unique value** in **ascending** order)

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

```
freq_counts = norm_reviews['Fandango_Ratingvalue'].value_counts()
sorted_freq_counts = freq_counts.sort_index()
```

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

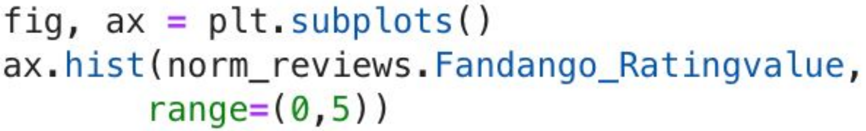
Rotten Tomatoes Frequency Distribution

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

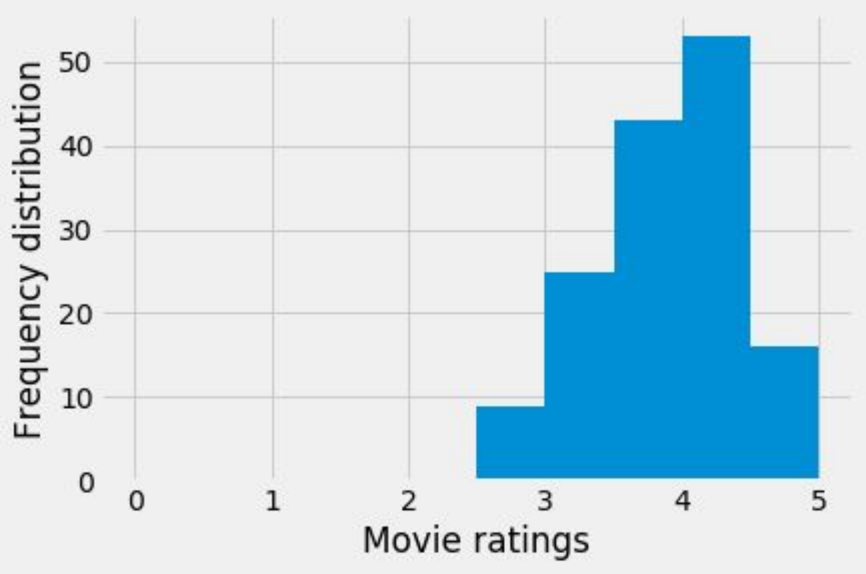
truncated to save space

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

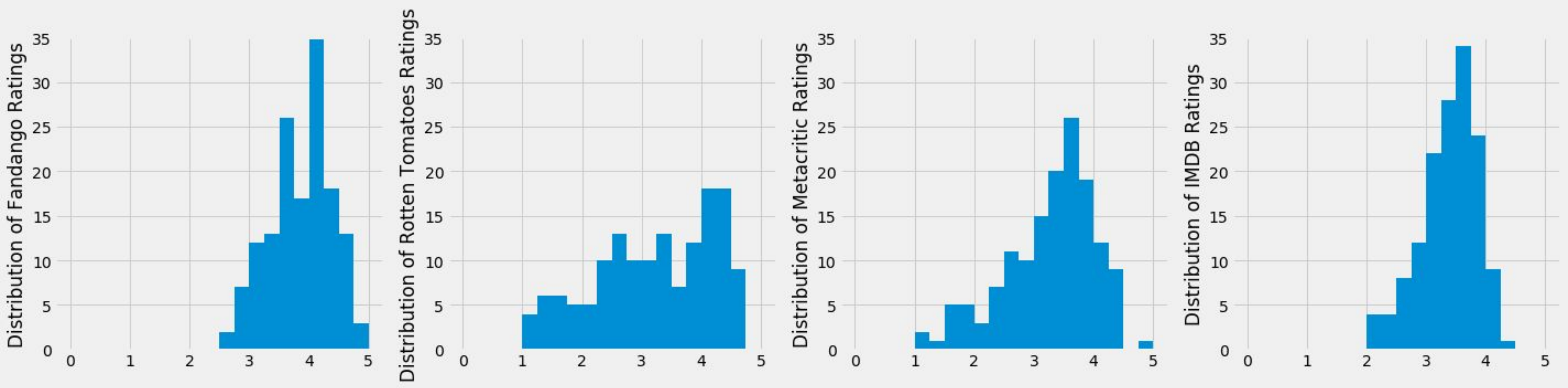
Histogram in Matplotlib



```
fig, ax = plt.subplots()
ax.hist(norm_reviews.Fandango_Ratingvalue,
        range=(0,5))
```



Comparing histograms



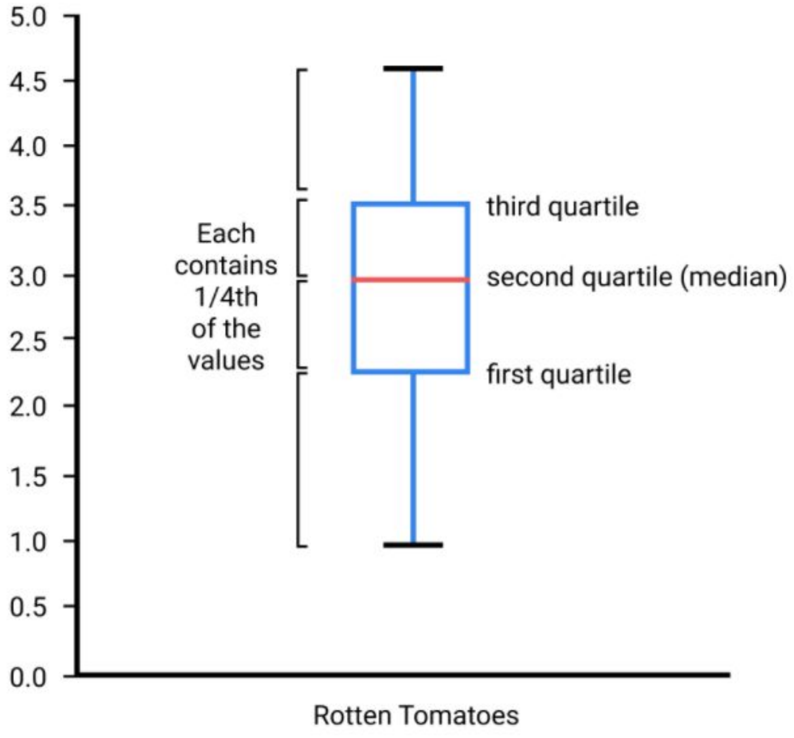
Around 50% of user ratings fall in the 2 to 4 score range

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

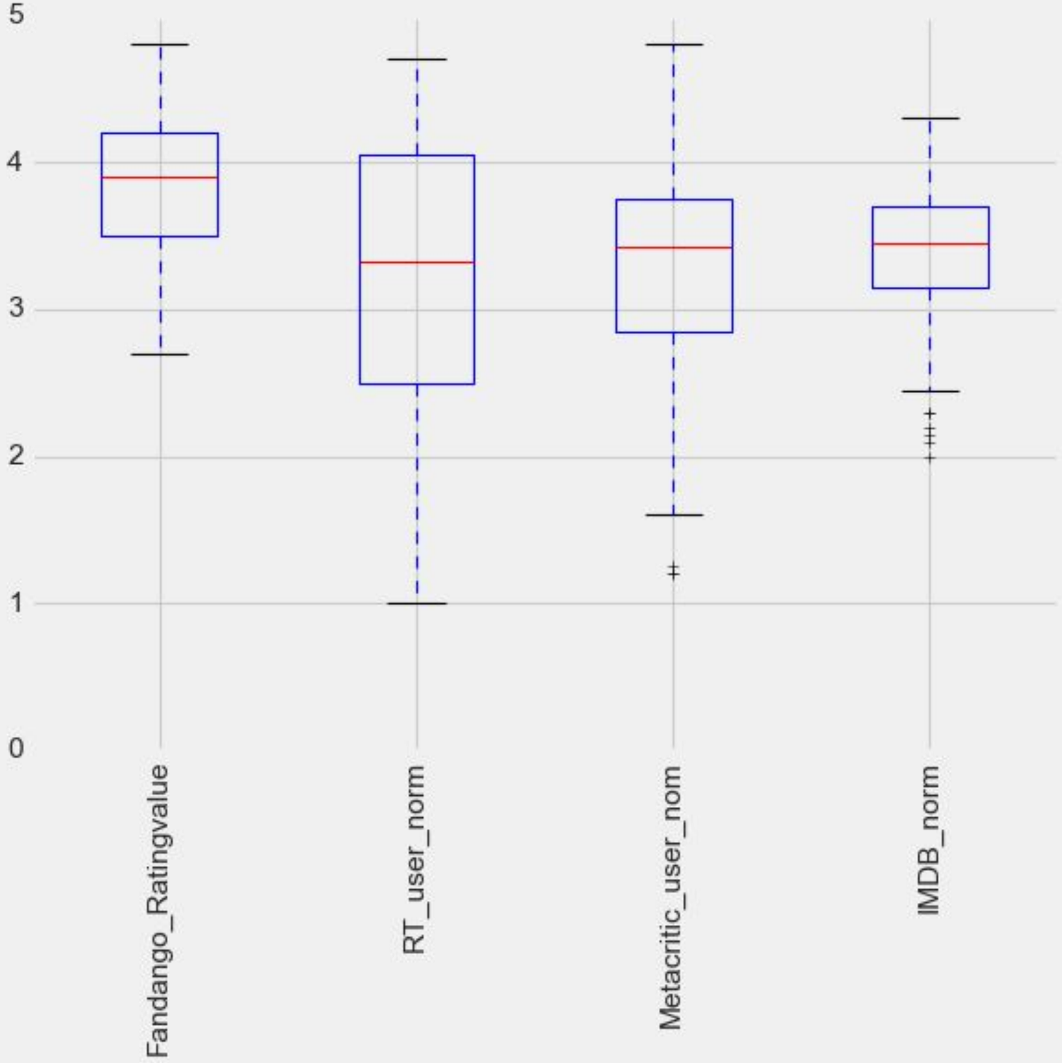
Around 75% of user ratings fall in the 2 to 4 score range

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

Quartile and Box Plot



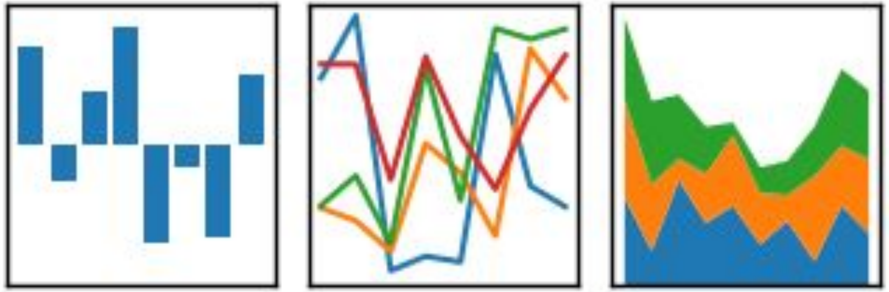
```
ax.boxplot(norm_reviews['RT_user_norm'])
```


5
4
3
2
1
0
Fandango_Ratingvalue
RT_user_norm
Metacritic_user_nom
IMDB_norm

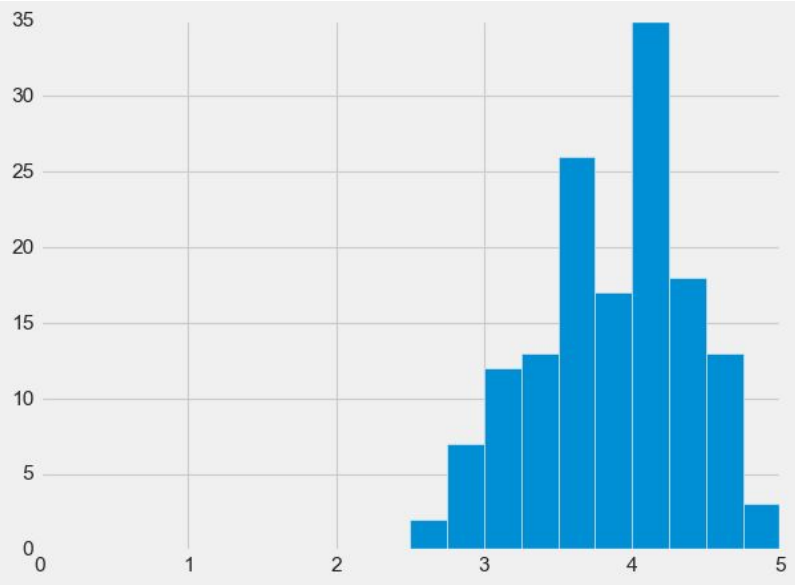
Lesson #7 - Exploratory Data Analysis II.ipynb
Section 1

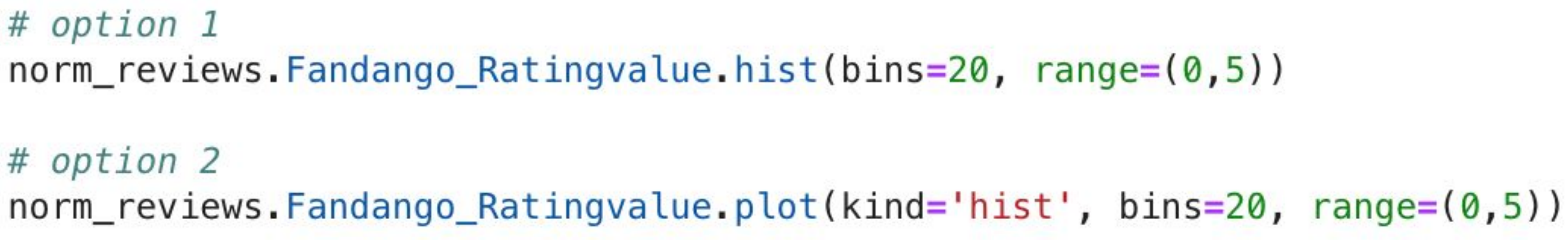
pandas

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

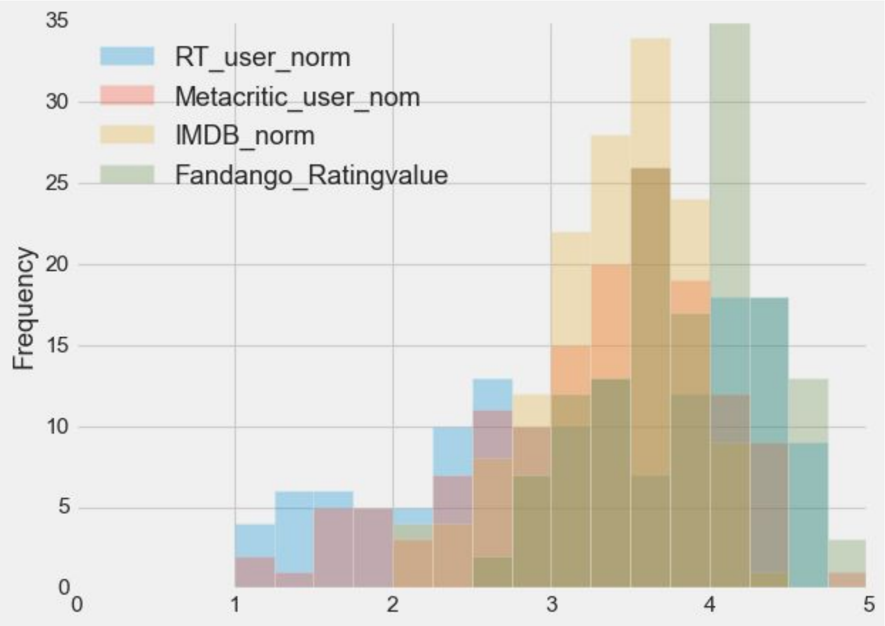


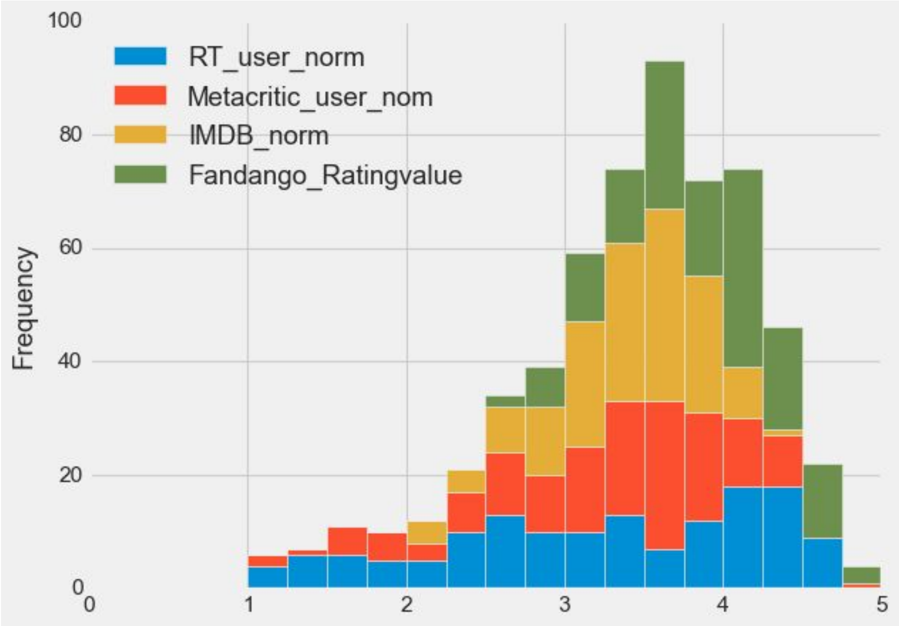
matplotlib

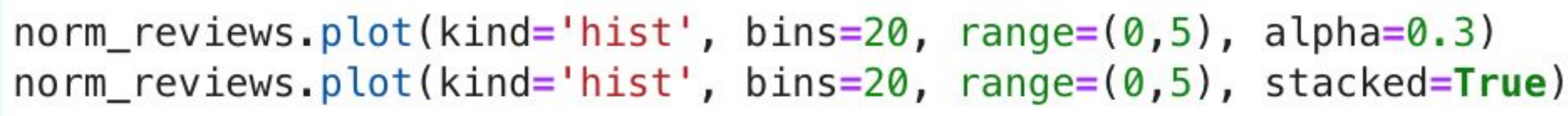




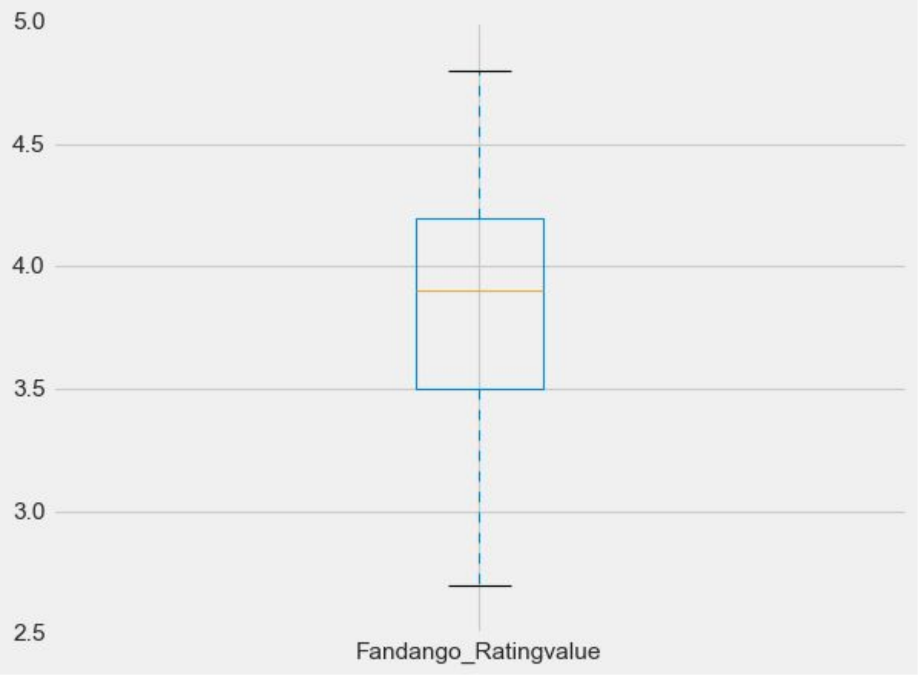
```
# option 1
norm_reviews.Fandango_Ratingvalue.hist(bins=20, range=(0,5))

# option 2
norm_reviews.Fandango_Ratingvalue.plot(kind='hist', bins=20, range=(0,5))
```

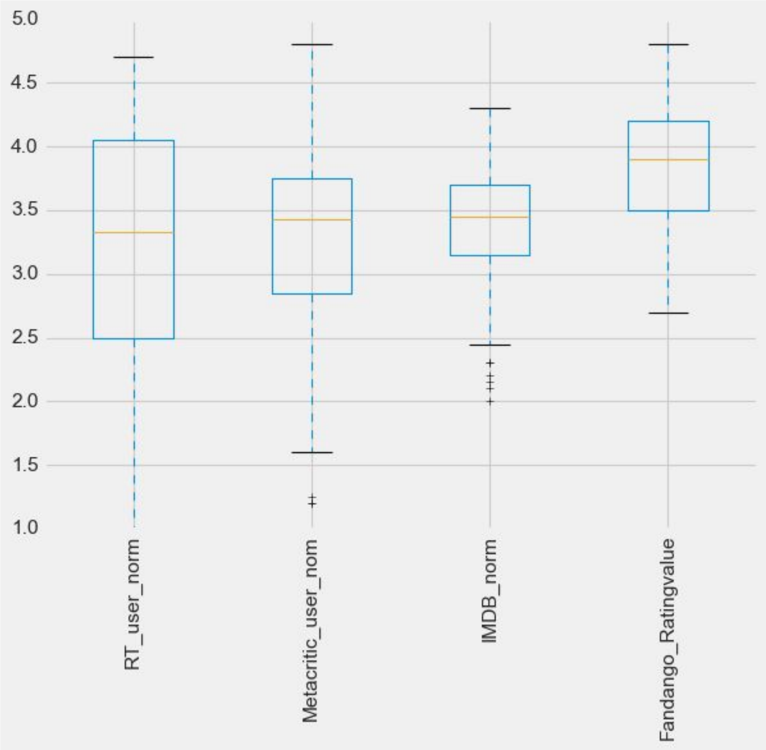






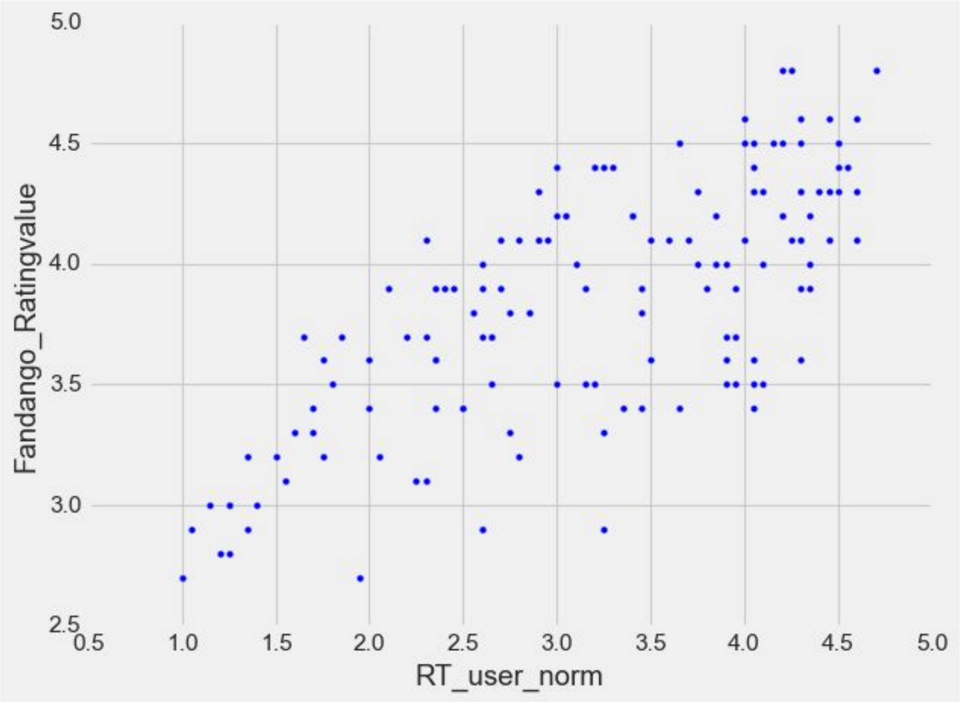
```
norm_reviews.plot(kind='hist', bins=20, range=(0,5), alpha=0.3)
norm_reviews.plot(kind='hist', bins=20, range=(0,5), stacked=True)
```





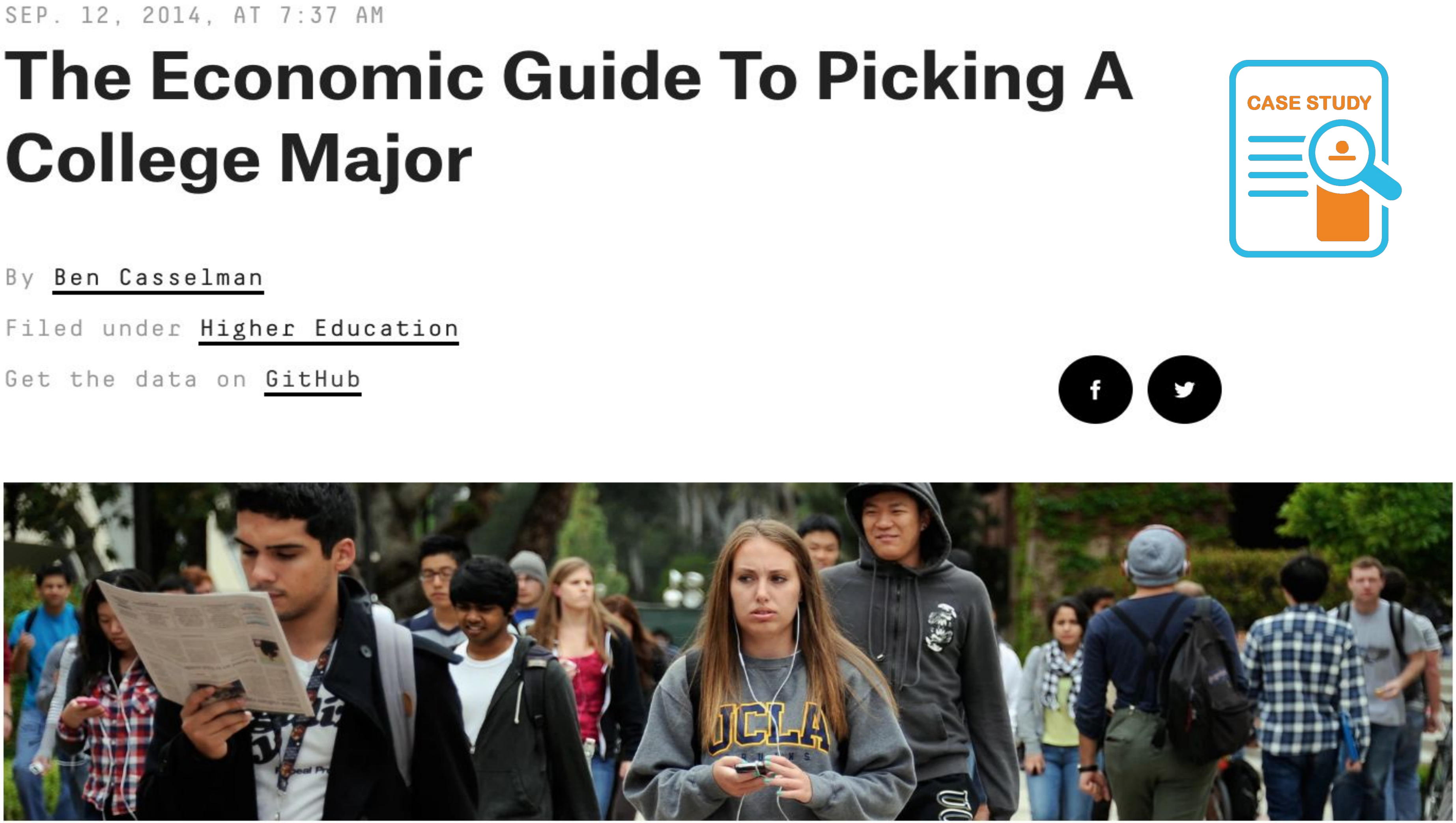
```
norm_reviews.Fandango_Ratingvalue.plot(kind='box')
```

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



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

CASE STUDY



SEP. 12, 2014, AT 7:37 AM

The Economic Guide To Picking A College Major

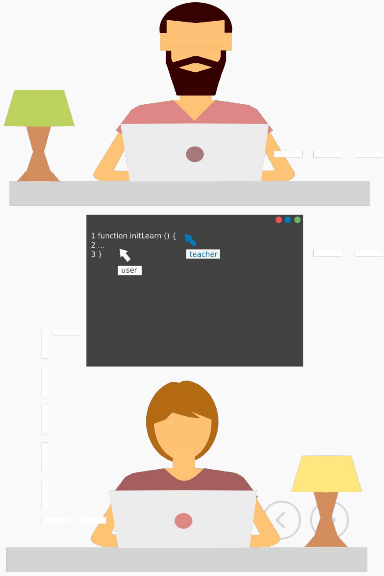
By Ben Casselman

Filed under Higher Education

Get the data on GitHub

Using visualizations, we can start to explore questions from the dataset like:

- **Do students in more popular majors make more money?**
  - Using scatter plots
- **How many majors are predominantly male? Predominantly female?**
  - Using histograms
- **Which category of majors have the most students?**
  - Using bar plots



```
1 function initLearn () {
2 ...
3 }
```

Lesson #7 - Exploratory Data Analysis II.ipynb
Sections 2 and 3