Describe Dataset

November 6, 2017

```
In [1]: import pandas as pd
        import numpy as np

        import matplotlib.pyplot as plt
        import seaborn as sns
        from matplotlib import rcParams

        sns.set_context("talk")
        sns.set_style("whitegrid")
        rcParams['patch.force_edgecolor'] = True
        %matplotlib inline
```

1 Load Dataset

There is a total of 99,999 ratings in this dataset. For every row, first two entries are the user id and movie id, which can be used to identify user and movie. The third entry is the rating, in this dataset, all ratings are integers in range 1 to 5. The last entry is a time stamp, which is unix seconds since 1/1/1970 UTC.

```
In [9]: ratings =  pd.read_csv('ml-100k/u.data', sep='\t',  header=0,
                                names=['userId', 'movieId', 'rating','timestamp'], engine='pytho
        ratings.head()
```

```
Out[9]:    userId  movieId  rating  timestamp
        0     186      302       3  891717742
        1      22      377       1  878887116
        2     244       51       2  880606923
        3     166      346       1  886397596
        4     298      474       4  884182806
```

2 User Description

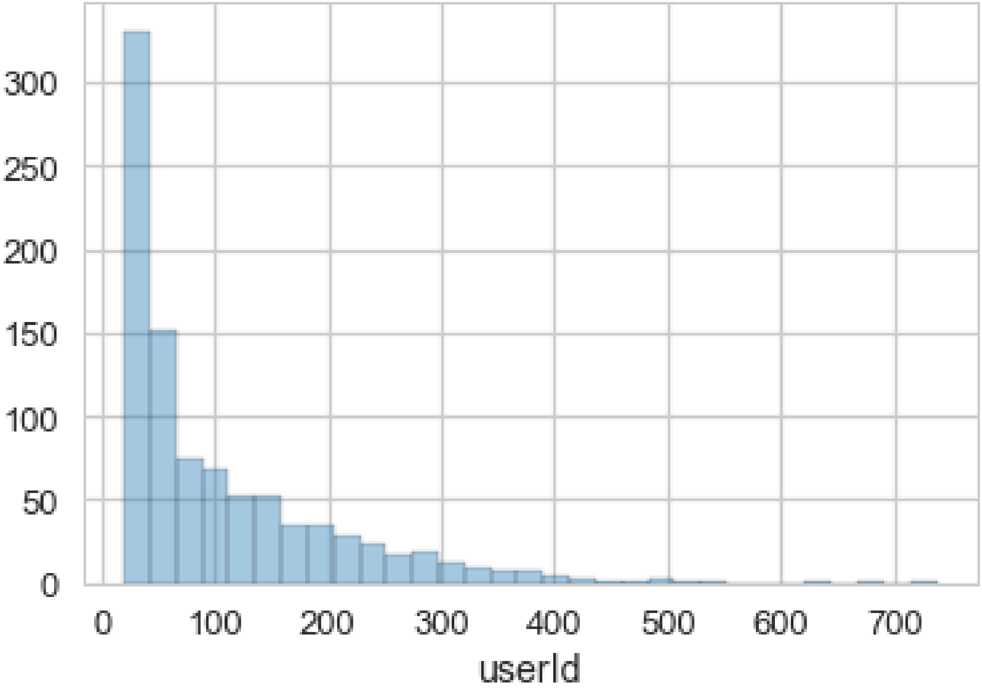
We have a total of 943 users. Each of them rated at least 20 movies and at most 737 movies. The mean number of rated movie for users is 106 and standard deviation is around 100. It is a long-tailed distribution, which means most people rated 100 or less movies, and only few people rated a lot.

```
In [3]: ratings['userId'].value_counts().describe()
```

```
Out[3]: count    943.000000
        mean     106.043478
        std      100.932453
        min       20.000000
        25%       33.000000
        50%       65.000000
        75%      148.000000
        max      737.000000
        Name: userId, dtype: float64
```

```
In [4]: sns.distplot(ratings['userId'].value_counts(), kde=False)
```

```
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x112d2ae48>
```



3 Movie Description

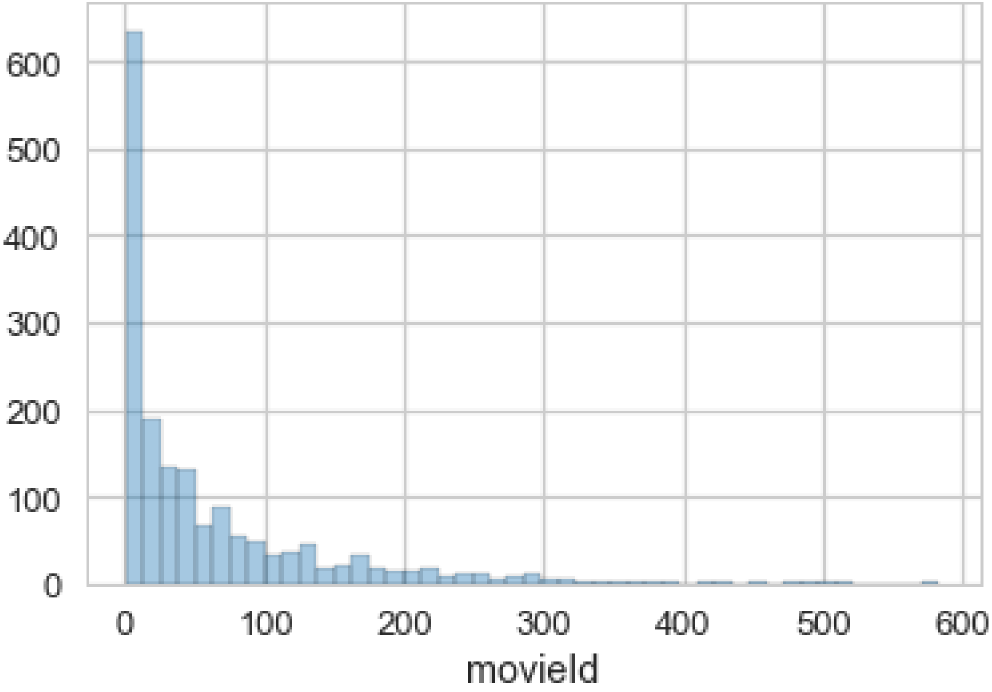
We have a total of 1682 users. They have been rated at least 1 time and at most 583 times. Mean value of number of ratings is around 60 but standard deviation is around 80. Most movies get 10 ratings or less.

```
In [5]: ratings['movieId'].value_counts().describe()
```

```
Out[5]: count    1682.000000
        mean       59.452438
        std        80.383423
        min         1.000000
        25%         6.000000
        50%        27.000000
        75%        80.000000
        max       583.000000
        Name: movieId, dtype: float64
```

```
In [6]: sns.distplot(ratings['movieId'].value_counts(), kde=False)
```

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x1147cd438>
```



4 Ratings Description

We have a total of 99,999 ratings in range 1 to 5, involve only integers. 4 is most occured in the ratings, and 3 is the second most. Over a half of ratings are 3 or 4. The mean value of ratings is 3.5.

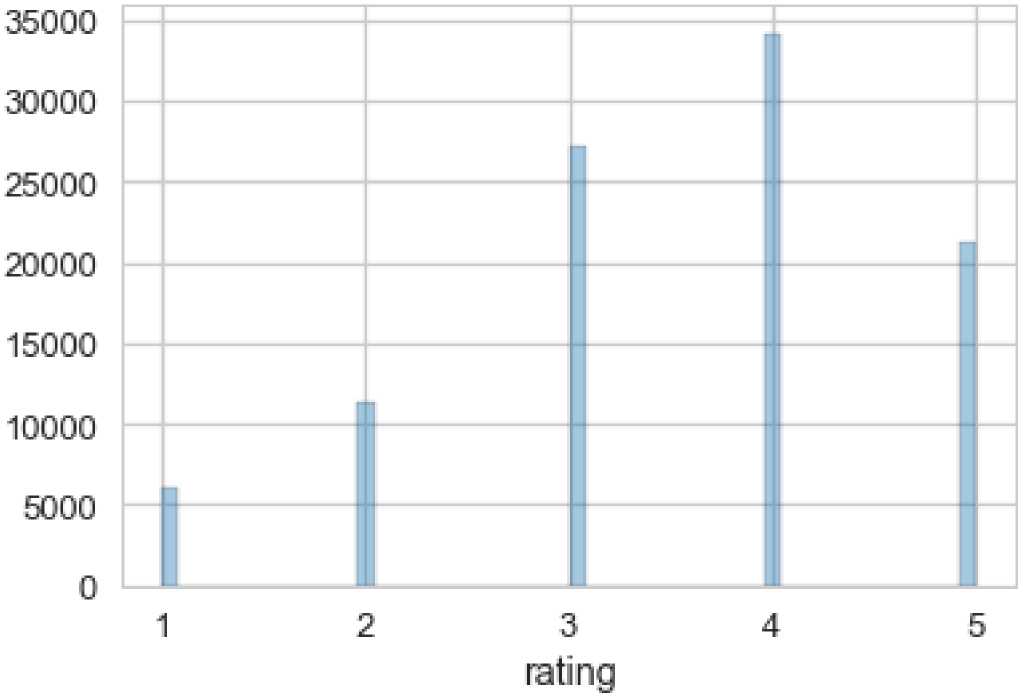
```
In [7]: ratings['rating'].describe()
```

```
Out[7]: count    99999.000000
        mean         3.529865
```

```
std          1.125678
min          1.000000
25%          3.000000
50%          4.000000
75%          4.000000
max          5.000000
Name: rating, dtype: float64
```

```
In [8]: sns.distplot(ratings['rating'], kde=False)
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x10590d048>
```


rating