Vivekanand Education Society's
Institute of Technology
Approved by AICTE & Affiliated to University of Mumbai

Artificial Intelligence and Data Science Department

Subject/Odd Sem 2023-23/Experiment 1

Name : Soham.Mukesh.Shetty	**Class/Roll No. :D11AD/58**	**Grade :5th Sem**

Title of Experiment : Proof of concept : Module 1

Objective of Experiment : To understand and explore dataset and do exploratory data analysis

Outcome of Experiment : Learn and implement different functions of different libraries using python on a unknown data set

Problem Statement : Load a dataset and do exploratory data analysis

Description / Theory :

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It involves methods for gathering and summarizing data, as well as techniques for drawing conclusions and making predictions based on that data. Statistics play a crucial role in various fields, including science, social sciences, economics, business, engineering, medicine, and more.

Estimate of location refers to a summary statistic that provides information about the central value or typical value of a dataset. It gives an idea of where the data tends to cluster or center around. There are several common estimates of location, and the choice of the appropriate one depends on the nature of the data and the specific goals of the analysis.

Here are some common estimates of location:

- Mean: The arithmetic mean, often referred to simply as the "mean," is calculated by summing up all the values in the dataset and dividing by the total number of observations. It represents the average value of the data and is sensitive to extreme values.
- Median: The median is the middle value of a dataset when it is ordered from lowest to highest. If the dataset has an odd number of observations, the median is the middle value. If the number of observations is even, the median is the average of the two middle values. The median is robust to extreme values and is often used when the data is skewed or contains outliers.

- Mode: The mode is the value that appears most frequently in the dataset. A dataset can have one mode (unimodal) or multiple modes (bimodal, trimodal, etc.). The mode is particularly useful for categorical data or discrete datasets.

- Weighted Mean: In some cases, data points might have different weights depending on their importance. The weighted mean takes into account these weights while calculating the average value.

Estimate of variability (also known as a measure of dispersion) is a summary statistic that provides information about the spread or dispersion of data points in a dataset.

There are several common measures of variability:

- Standard Deviation: The standard deviation is the square root of the variance. It represents the typical amount of deviation or dispersion of data points from the mean. A smaller standard deviation indicates that the data points are close to the mean, while a larger standard deviation indicates greater variability or spread.

- Interquartile Range (IQR): The interquartile range is the difference between the third quartile (Q3) and the first quartile (Q1) of the dataset. It measures the spread of the middle 50% of the data and is not affected by extreme values or outliers.

Estimate of a percentile refers to a value that divides a dataset into 100 equal parts. It is a measure that indicates the relative standing or position of a particular data point within the entire dataset. Percentiles are often used to understand the distribution of data and to compare individual values to the rest of the observations.

Frequency Table, Histograms, and Density Plots are all common tools used in statistics and data analysis to visualize and understand the distribution of data. Let's explore each of them:

Frequency Table:
A frequency table is a tabular representation that shows the number of times each value (or range of values) occurs in a dataset. It summarizes the data by counting how frequently each distinct value appears. The table consists of two columns: one listing the values and the other showing the corresponding frequencies. Frequency tables are especially useful for categorical data, but they can also be used for discrete or grouped data.

Histograms:
A histogram is a graphical representation of the frequency distribution of a continuous or discrete dataset. It consists of a series of bars, where the width of each bar corresponds to a range of values (called bins), and the height of each bar represents the frequency or count of data points falling within that bin. Histograms provide a visual depiction of how the data is spread across different intervals, giving insights into the shape and central tendency of the distribution.

Density Plot:
A density plot is a smooth, continuous version of a histogram. It is used to visualize the distribution of continuous data and is particularly helpful when the data is not clearly discrete. Density plots are created by estimating the underlying probability density function of the data and then plotting it as a continuous curve. Unlike histograms, density plots do not rely on fixed bins, and the shape of the curve gives insights into the data's central tendency and spread.

Scatter Plot:
A scatter plot is a graphical representation used to display the relationship between two continuous variables. Each point on the plot represents a single data observation, with one variable plotted on the x-axis and the other variable on the y-axis. Scatter plots are particularly useful for identifying patterns, trends, or correlations between the two variables. The visual appearance of the points (e.g., clustering, dispersion, linearity) provides insights into the strength and direction of the relationship between the variables.

Binning and Contours:
Binning is a technique used to divide a continuous variable into discrete intervals or bins. It involves grouping the data into ranges based on the values of the variable. Binning is useful when we want to simplify the data or when the data has a large range, making it difficult to analyze directly.

Contours are lines drawn on a two-dimensional plot to represent data points with the same value. In the context of binning, contours are often used to create two-dimensional representations of the frequency or density of data points in specific bins. Contour plots help visualize the density of data points in different regions of the plot, and they are often used in combination with histograms or density plots.

Contingency Table:
A contingency table (also known as a cross-tabulation or crosstab) is a tabular representation used to display the joint distribution of two categorical variables. The table shows the frequency of occurrence of different combinations of the two variables. Contingency tables are commonly used in hypothesis testing and to examine relationships between categorical variables.

Violin Plot:
A violin plot is a combination of a box plot and a kernel density plot. It is used to visualize the distribution of a continuous variable or multiple continuous variables across different categories. The violin plot displays the data's density or distribution by drawing mirrored density plots on either side of the box plot. The box plot shows the median, quartiles, and possible outliers, while the density plots provide a smooth representation of the data's distribution.

In [1]:

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
from statsmodels import robust


from pathlib import Path
from scipy.stats import trim_mean
```

In [2]:

```
anime = pd.read_csv(r"C:\Users\HP\Downloads\Stats project\anime.csv")
```

In [3]:

```
print(anime.head(8))
```

```
   anime_id                                                name  \
0     32281                                      Kimi no Na wa.   
1      5114                    Fullmetal Alchemist: Brotherhood   
2     28977                                            Gintama°   
3      9253                                         Steins;Gate   
4      9969                                       Gintama&#039;   
5     32935  Haikyuu!!: Karasuno Koukou VS Shiratorizawa Ga...   
6     11061                              Hunter x Hunter (2011)   
7       820                                Ginga Eiyuu Densetsu   

                                               genre   type episodes  rating  \
0               Drama, Romance, School, Supernatural  Movie        1    9.37   
1  Action, Adventure, Drama, Fantasy, Magic, Mili...     TV       64    9.26   
2  Action, Comedy, Historical, Parody, Samurai, S...     TV       51    9.25   
3                                   Sci-Fi, Thriller     TV       24    9.17   
4  Action, Comedy, Historical, Parody, Samurai, S...     TV       51    9.16   
5             Comedy, Drama, School, Shounen, Sports     TV       10    9.15   
6           Action, Adventure, Shounen, Super Power     TV      148    9.13   
7                     Drama, Military, Sci-Fi, Space    OVA      110    9.11   

   members  
0   200630  
1   793665  
2   114262  
3   673572  
4   151266  
5    93351  
6   425855  
7    80679  
```

In [4]:

```
anime
```

Out[4]:

	anime_id	name	genre	type	episodes	rating	members
0	32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
1	5114	Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Mili...	TV	64	9.26	793665
2	28977	Gintama°	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.25	114262
3	9253	Steins;Gate	Sci-Fi, Thriller	TV	24	9.17	673572
4	9969	Gintama&#039;	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.16	151266
...	...	...	...	...	...	...	...
12289	9316	Toushindai My Lover: Minami tai Mecha-Minami	Hentai	OVA	1	4.15	211
12290	5543	Under World	Hentai	OVA	1	4.28	183
12291	5621	Violence Gekiga David no Hoshi	Hentai	OVA	4	4.88	219
12292	6133	Violence Gekiga Shin David no Hoshi: Inma Dens...	Hentai	OVA	1	4.98	175
12293	26081	Yasuji no Pornorama: Yacchimae!!	Hentai	Movie	1	5.46	142

12294 rows × 7 columns

Estimates of Location

In [5]:

```
print("Mean rating =")
print(anime['rating'].mean())
```

```
Mean rating =
6.473901690981445
```

In [6]:

```
print("median of  rating =")
print(anime['rating'].median())
```

```
median of  rating =
6.57
```

In [7]:

```
print("mode of  episodes =")
print(anime['episodes'].mode())
```

```
mode of  episodes =
0    1
Name: episodes, dtype: object
```

In [8]:

```
print(anime.head(10))
```

```
   anime_id                                                name  \
0     32281                                      Kimi no Na wa.   
1      5114                    Fullmetal Alchemist: Brotherhood   
2     28977                                            Gintama°   
3      9253                                         Steins;Gate   
4      9969                                       Gintama&#039;   
5     32935   Haikyuu!!: Karasuno Koukou VS Shiratorizawa Ga...   
6     11061                               Hunter x Hunter (2011)   
7       820                                 Ginga Eiyuu Densetsu   
8     15335   Gintama Movie: Kanketsu-hen - Yorozuya yo Eien...   
9     15417                            Gintama&#039;: Enchousen   

                                               genre   type episodes  rating  \
0               Drama, Romance, School, Supernatural  Movie        1    9.37   
1  Action, Adventure, Drama, Fantasy, Magic, Mili...     TV       64    9.26   
2  Action, Comedy, Historical, Parody, Samurai, S...     TV       51    9.25   
3                                   Sci-Fi, Thriller     TV       24    9.17   
4  Action, Comedy, Historical, Parody, Samurai, S...     TV       51    9.16   
5             Comedy, Drama, School, Shounen, Sports     TV       10    9.15   
6            Action, Adventure, Shounen, Super Power     TV      148    9.13   
7                     Drama, Military, Sci-Fi, Space    OVA      110    9.11   
8  Action, Comedy, Historical, Parody, Samurai, S...  Movie        1    9.10   
9  Action, Comedy, Historical, Parody, Samurai, S...     TV       13    9.11   

   members  
0   200630  
1   793665  
2   114262  
3   673572  
4   151266  
5    93351  
6   425855  
7    80679  
8    72534  
9    81109  
```

In [9]:

```
print(trim_mean(anime['members'],0.1))
```

```
5589.94286295242
```

In [10]:

```
sample_anime = anime.head(100)

print(np.average(sample_anime['rating'], weights=sample_anime['members']))
```

```
8.754444415324095
```

Estimates of Variablity

In [11]:

```
print(anime['members'].std())
```

```
54820.676924907515
```

In [12]:
```
#Interquartile range is calculated as the difference of the 75% and 25% quantile.
print(anime['members'].quantile(0.75) - anime['members'].quantile(0.25))
```

```
9212.0
```

In [13]:
```
#Median absolute deviation from the median
#method- 1
#print(abs(anime['members'] - anime['members'].median()).median() / 0.6744897501960
#method-2
print(robust.scale.mad(anime['members']))
```

```
2172.0122501107066
```

Estimates on Percentiles

In [14]:
```
print(anime['members'].quantile([0.05, 0.25, 0.5, 0.75, 0.95]))
```

```
0.05       58.0
0.25      225.0
0.50     1550.0
0.75     9437.0
0.95    93164.3
Name: members, dtype: float64
```

In [15]:
```
percentages = [0.05, 0.25, 0.5, 0.75, 0.95]
df = pd.DataFrame(anime['members'].quantile(percentages))
df.index = [f'{p * 100}%' for p in percentages]
print(df.transpose())
```

```
         5.0%  25.0%   50.0%   75.0%    95.0%
members  58.0  225.0  1550.0  9437.0  93164.3
```

Explore data distribution

Frequency table

In [16]:
```
animax = anime.head(2000)
binnedmembers = pd.cut(animax['members'], 20)
print(binnedmembers.value_counts())
```

```
(-644.548, 51046.4]      1190
(51046.4, 101723.8]       356
(101723.8, 152401.2]      156
(152401.2, 203078.6]       96
(203078.6, 253756.0]       66
(253756.0, 304433.4]       37
(304433.4, 355110.8]       33
(355110.8, 405788.2]       16
(405788.2, 456465.6]       11
(456465.6, 507143.0]       10
(557820.4, 608497.8]        8
(507143.0, 557820.4]        8
(608497.8, 659175.2]        5
(659175.2, 709852.6]        2
(709852.6, 760530.0]        2
(861884.8, 912562.2]        2
(760530.0, 811207.4]        1
(963239.6, 1013917.0]       1
(811207.4, 861884.8]        0
(912562.2, 963239.6]        0
Name: members, dtype: int64
```

In [17]:

```
binnedmembers.name = 'binnedmembers'
df = pd.concat([animax, binnedmembers], axis=1)
df = df.sort_values(by='members')

groups = []
for group, subset in df.groupby(by='binnedmembers'):
    groups.append({
        'BinRange': group,
        'Count': len(subset),
        'Genre': ','.join(subset.genre)
    })
print(pd.DataFrame(groups))
```

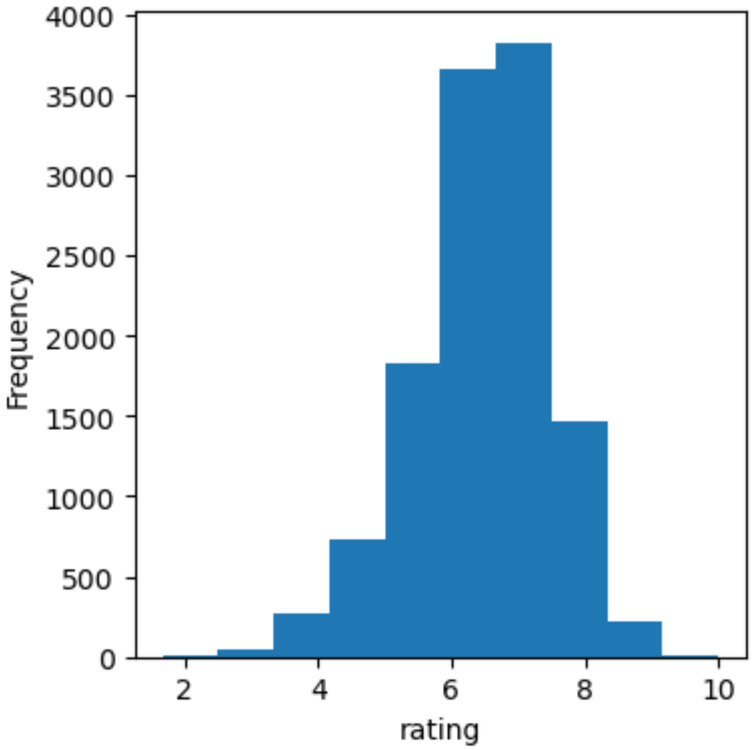
```
                 BinRange  Count  \
0     (-644.548, 51046.4]   1190   
1     (51046.4, 101723.8]    356   
2    (101723.8, 152401.2]    156   
3    (152401.2, 203078.6]     96   
4    (203078.6, 253756.0]     66   
5    (253756.0, 304433.4]     37   
6    (304433.4, 355110.8]     33   
7    (355110.8, 405788.2]     16   
8    (405788.2, 456465.6]     11   
9    (456465.6, 507143.0]     10   
10   (507143.0, 557820.4]      8   
11   (557820.4, 608497.8]      8   
12   (608497.8, 659175.2]      5   
13   (659175.2, 709852.6]      2   
14   (709852.6, 760530.0]      2   
15   (760530.0, 811207.4]      1   
16   (811207.4, 861884.8]      0   
17   (861884.8, 912562.2]      2   
18   (912562.2, 963239.6]      0   
19  (963239.6, 1013917.0]      1   

                                                Genre  
0   Action, Fantasy, Historical, Martial Arts,Dram...  
1   Action, Adventure, Samurai,Romance, School, Sh...  
2   Comedy, Drama, Mystery, Romance, Slice of Life...  
3   Action, Adventure, Drama, Fantasy, Historical,...  
4   Action, Seinen,Action, Fantasy, Magic, Romance...  
5   Comedy, School, Slice of Life,Action, Adventur...  
6   Action, Comedy, Dementia, Mecha, Parody, Sci-F...  
7   Drama, Fantasy, Psychological, Thriller,Fantas...  
8   Action, Drama, Horror, Mystery, Psychological,...  
9   Drama, Fantasy, Romance, Slice of Life, Supern...  
10  Action, Comedy, School, Super Power,Action, Po...  
11  Action, Adventure, Comedy, Mecha, Sci-Fi,Comed...  
12  Action, Drama, Horror, Mystery, Psychological,...  
13  Sci-Fi, Thriller,Action, Comedy, Martial Arts,...  
14  Action, Mecha, Military, School, Sci-Fi, Super...  
15  Action, Adventure, Drama, Fantasy, Magic, Mili...  
16                                                     
17  Action, Adventure, Fantasy, Game, Romance,Acti...  
18                                                     
19  Mystery, Police, Psychological, Supernatural, ...  
```

In [18]:
```
print(pd.DataFrame(groups))
```

```
                 BinRange  Count  \
0     (-644.548, 51046.4]   1190   
1     (51046.4, 101723.8]    356   
2    (101723.8, 152401.2]    156   
3    (152401.2, 203078.6]     96   
4    (203078.6, 253756.0]     66   
5    (253756.0, 304433.4]     37   
6    (304433.4, 355110.8]     33   
7    (355110.8, 405788.2]     16   
8    (405788.2, 456465.6]     11   
9    (456465.6, 507143.0]     10   
10   (507143.0, 557820.4]      8   
11   (557820.4, 608497.8]      8   
12   (608497.8, 659175.2]      5   
13   (659175.2, 709852.6]      2   
14   (709852.6, 760530.0]      2   
15   (760530.0, 811207.4]      1   
16   (811207.4, 861884.8]      0   
17   (861884.8, 912562.2]      2   
18   (912562.2, 963239.6]      0   
19  (963239.6, 1013917.0]      1   

                                                Genre  
0   Action, Fantasy, Historical, Martial Arts,Dram...  
1   Action, Adventure, Samurai,Romance, School, Sh...  
2   Comedy, Drama, Mystery, Romance, Slice of Life...  
3   Action, Adventure, Drama, Fantasy, Historical,...  
4   Action, Seinen,Action, Fantasy, Magic, Romance...  
5   Comedy, School, Slice of Life,Action, Adventur...  
6   Action, Comedy, Dementia, Mecha, Parody, Sci-F...  
7   Drama, Fantasy, Psychological, Thriller,Fantas...  
8   Action, Drama, Horror, Mystery, Psychological,...  
9   Drama, Fantasy, Romance, Slice of Life, Supern...  
10  Action, Comedy, School, Super Power,Action, Po...  
11  Action, Adventure, Comedy, Mecha, Sci-Fi,Comed...  
12  Action, Drama, Horror, Mystery, Psychological,...  
13  Sci-Fi, Thriller,Action, Comedy, Martial Arts,...  
14  Action, Mecha, Military, School, Sci-Fi, Super...  
15  Action, Adventure, Drama, Fantasy, Magic, Mili...  
16                                                     
17  Action, Adventure, Fantasy, Game, Romance,Acti...  
18                                                     
19  Mystery, Police, Psychological, Supernatural, ...  
```

Histograms

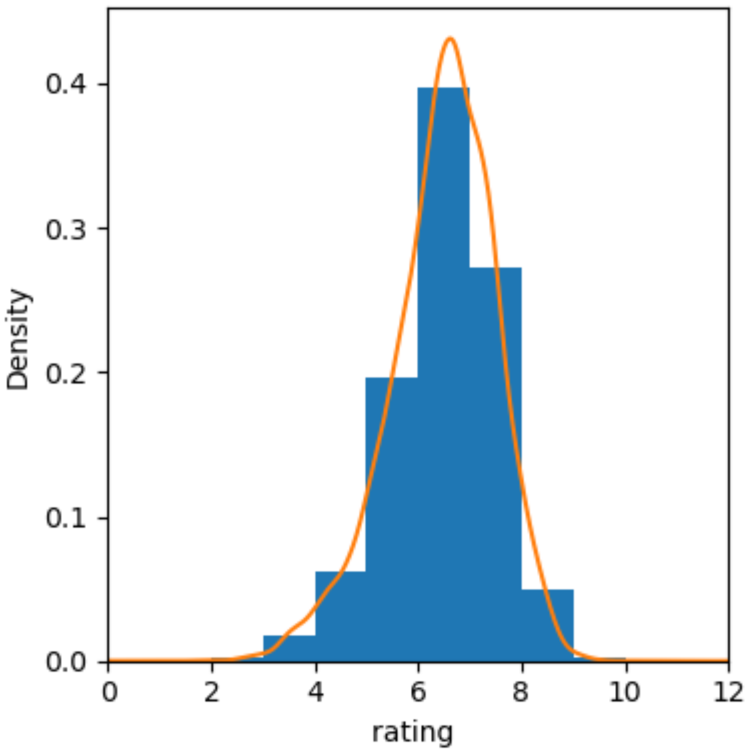
```
In [19]: ax = (anime['rating'] ).plot.hist(figsize=(4, 4))
         ax.set_xlabel('rating')

         plt.tight_layout()
         plt.show()
```



density plot

In [20]:

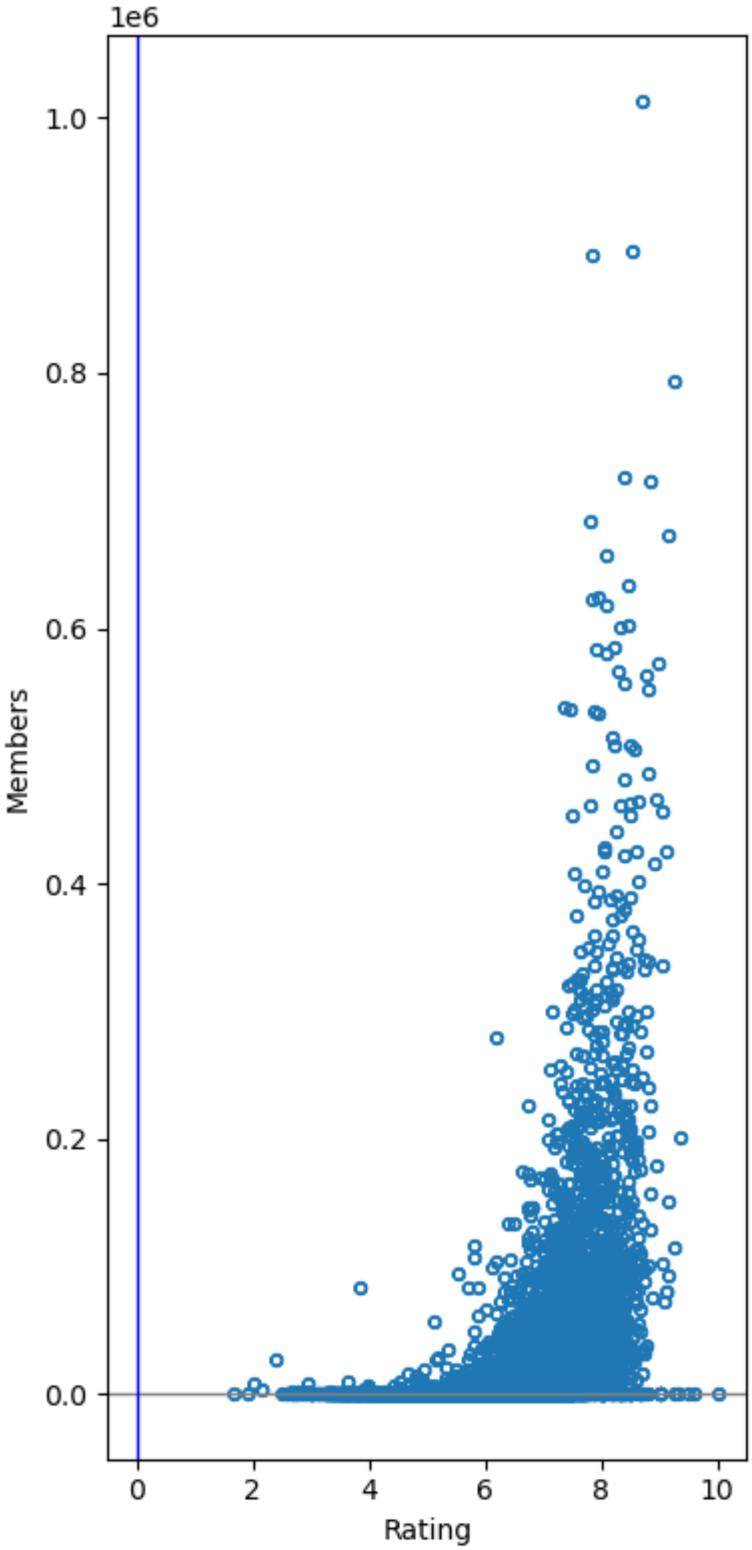
```
ax = anime['rating'].plot.hist(density=True, xlim=[0, 12],
                                bins=range(1,12), figsize=(4, 4))
anime['rating'].plot.density(ax=ax)
ax.set_xlabel('rating ')

plt.tight_layout()
plt.show()
```



Exploring Binary and Categorical Data

Scatterplots

```
In [21]: ax = anime.plot.scatter(x='rating', y='members', figsize=(4, 8), marker='$\u25EF$')
         ax.set_xlabel('Rating')
         ax.set_ylabel('Members')
         ax.axhline(0, color='grey', lw=1)
         ax.axvline(0, color='blue', lw=1)

         plt.tight_layout()
         plt.show()
```



Binning

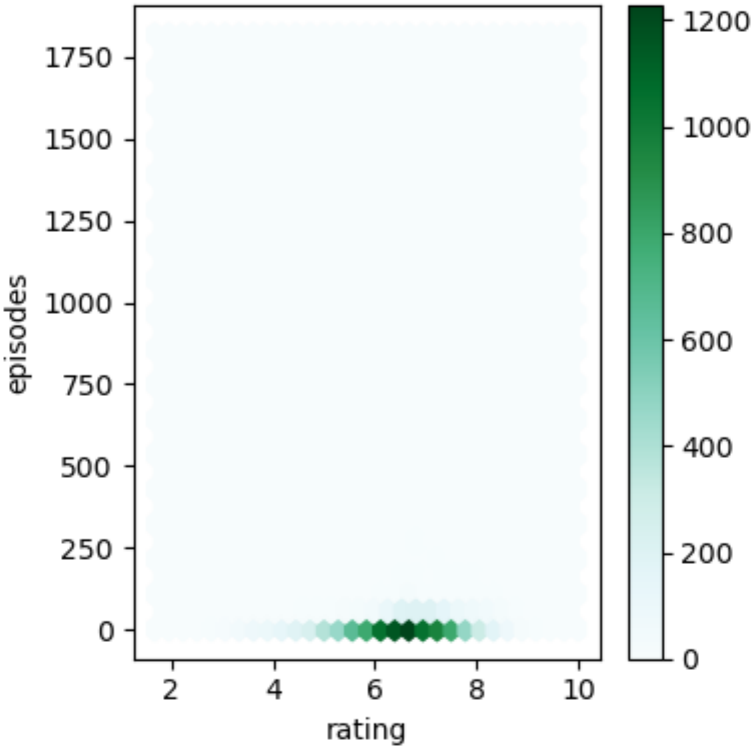
In [22]:

```
animes = anime.drop(anime[anime.episodes == 'Unknown'].index)
print(animes)
```

```
       anime_id                                               name  \
0         32281                                     Kimi no Na wa.
1          5114                   Fullmetal Alchemist: Brotherhood
2         28977                                           Gintama°
3          9253                                        Steins;Gate
4          9969                                      Gintama&#039;
...         ...                                                ...
12289      9316       Toushindai My Lover: Minami tai Mecha-Minami
12290      5543                                        Under World
12291      5621                     Violence Gekiga David no Hoshi
12292      6133  Violence Gekiga Shin David no Hoshi: Inma Dens...
12293     26081                   Yasuji no Pornorama: Yacchimae!!

                                                   genre   type episodes  \
0                    Drama, Romance, School, Supernatural  Movie        1
1      Action, Adventure, Drama, Fantasy, Magic, Mili...     TV       64
2      Action, Comedy, Historical, Parody, Samurai, S...     TV       51
3                                       Sci-Fi, Thriller     TV       24
4      Action, Comedy, Historical, Parody, Samurai, S...     TV       51
...                                                  ...    ...      ...
12289                                             Hentai    OVA        1
12290                                             Hentai    OVA        1
12291                                             Hentai    OVA        4
12292                                             Hentai    OVA        1
12293                                             Hentai  Movie        1

       rating  members
0        9.37   200630
1        9.26   793665
2        9.25   114262
3        9.17   673572
4        9.16   151266
...       ...      ...
12289    4.15      211
12290    4.28      183
12291    4.88      219
12292    4.98      175
12293    5.46      142

[11954 rows x 7 columns]
```

In [23]:

```
animes = animes.astype({'episodes':'float'})
ax = animes.plot.hexbin(x='rating', y='episodes',
                         gridsize=30, sharex=False, figsize=(4, 4))
ax.set_xlabel('rating')
ax.set_ylabel('episodes')

plt.tight_layout()
plt.show()
```



In [24]:

```
animes['true_weight'] = animes['episodes'] * animes['members'] / 100000
animes.head(10)
```

Out[24]:

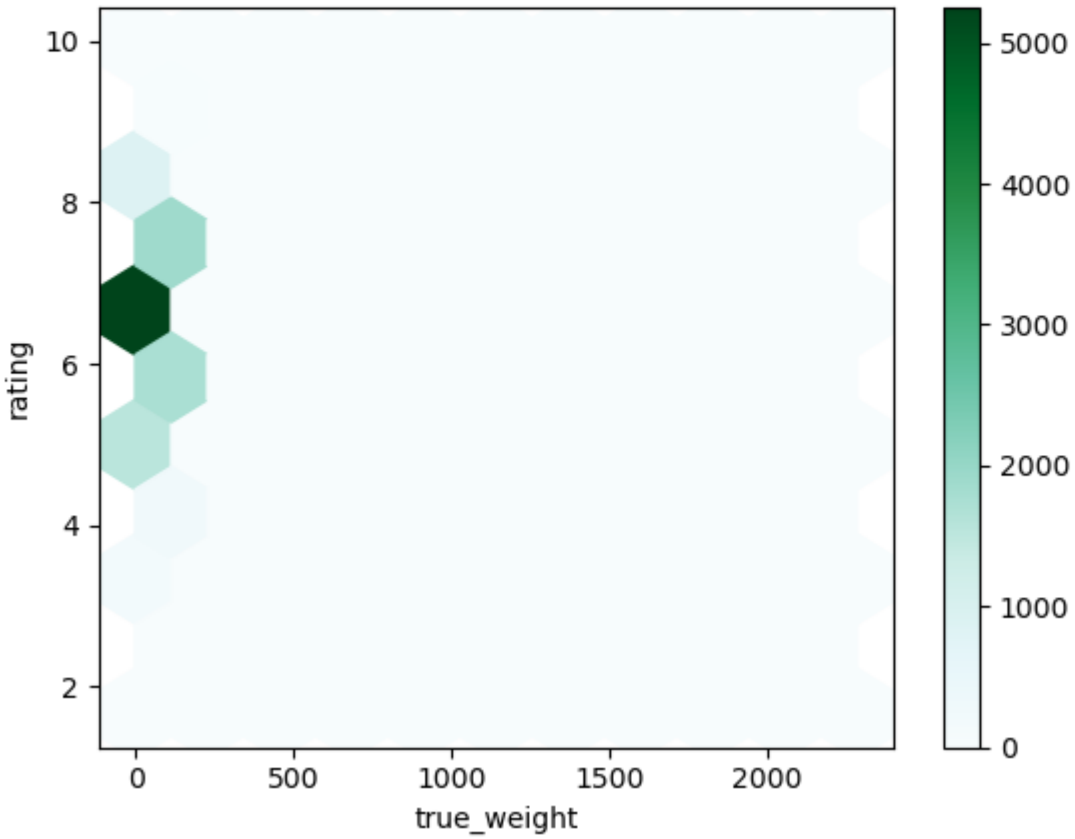
	anime_id	name	genre	type	episodes	rating	members	true_weight
0	32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1.0	9.37	200630	2.00630
1	5114	Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Mili...	TV	64.0	9.26	793665	507.94560
2	28977	Gintama°	Action, Comedy, Historical, Parody, Samurai, S...	TV	51.0	9.25	114262	58.27362
3	9253	Steins;Gate	Sci-Fi, Thriller	TV	24.0	9.17	673572	161.65728
4	9969	Gintama&#039;	Action, Comedy, Historical, Parody, Samurai, S...	TV	51.0	9.16	151266	77.14566
5	32935	Haikyuu!!: Karasuno Koukou VS Shiratorizawa Ga...	Comedy, Drama, School, Shounen, Sports	TV	10.0	9.15	93351	9.33510
6	11061	Hunter x Hunter (2011)	Action, Adventure, Shounen, Super Power	TV	148.0	9.13	425855	630.26540
7	820	Ginga Eiyuu Densetsu	Drama, Military, Sci-Fi, Space	OVA	110.0	9.11	80679	88.74690
8	15335	Gintama Movie: Kanketsu-hen - Yorozuya yo Eien...	Action, Comedy, Historical, Parody, Samurai, S...	Movie	1.0	9.10	72534	0.72534
9	15417	Gintama&#039;: Enchousen	Action, Comedy, Historical, Parody, Samurai, S...	TV	13.0	9.11	81109	10.54417

In [25]:

```
animes.plot(kind='hexbin',x='true_weight', y = 'rating' , gridsize = 10)
```

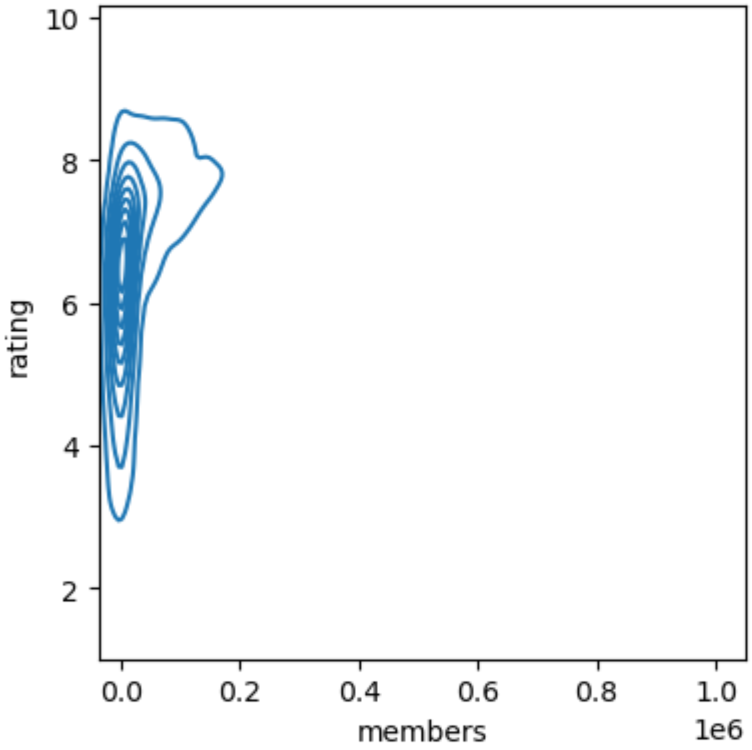
Out[25]:

```
<AxesSubplot:xlabel='true_weight', ylabel='rating'>
```



In [26]:

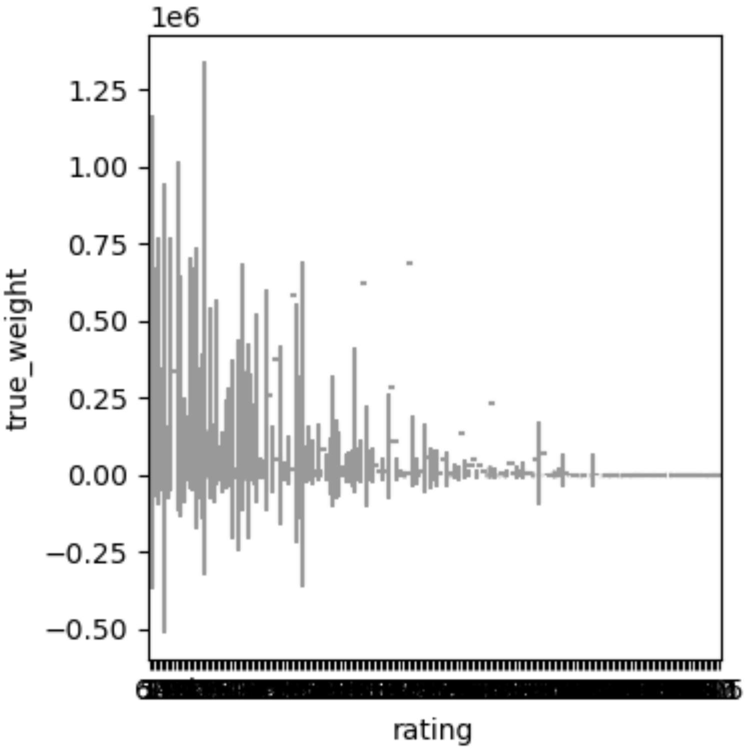
```
fig, ax = plt.subplots(figsize=(4, 4))
sns.kdeplot(data=animes.sample(10000), x='members', y='rating', ax=ax)
ax.set_xlabel('members')
ax.set_ylabel('rating')

plt.tight_layout()
plt.show()
```



Violen plot

In [28]:

```
fig, ax = plt.subplots(figsize=(4, 4))
sns.violinplot(data=anime, x='episodes', y='members',
               ax=ax, inner='quartile', color='white')
ax.set_xlabel('rating')
ax.set_ylabel('true_weight')

plt.tight_layout()
plt.show()
```


1e6
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
true_weight
rating

Results and Discussions:

With this we have analysed the trends and patterns in the popularity of various genres in the domain of Japanese Anime Industry.

We have used a holistic qualitative analysis approach for our study:

1. We used various libraries and py.modules including seaborn and robust, mathplotlib.

2. Pre-processed and manipulation of data.

3. Calculated Estimates of Location, Variability and Percentiles.

4. Categorised the analytical data according to appropriate distributions, exploring Binary Data as well.

5. Data Visualisation using Histograms, Density Plotting, Scatter plotting, Hexagonal Binning using linear and logarithmic scales, Contour plot and Violin Plot.