Visualization Library Documentation

Visualization libraries are used to generate graphs plots by the values of different feature or a single feature of a data. With visualization the patterns and trends in data are easy to find and analyse.

Some major visualization libraries in python are:

- Matplotlib
- Seaborn
- Bokeh
- Plotly

Matplotlib

Matplotlib is a basic level data visualization library built on Numpy Arrays. It is used to generate 2D graphs.

Components of a plot of Matplotlib:
Figure: The figure contains all the elements of the plot. A figure can have multiple axes.
Axes: Axes is the area in the figure where the data is plotted.
Axis: In matplotlib the axis are x-axis and y-axis. These set the limits and scaling of the data.
Artist: All the elements of the figure is an artist like labels, legend, ticks, axis.

We import *pyplot*, a sub library in matplotlib to define the plots and *numpy for numerical operations.*

Different type of plots in Matplotlib

1. Line Graph
2. Bar Graph
3. Histogram
4. Scatter Plot
5. Pie Chart
6. Box Plot

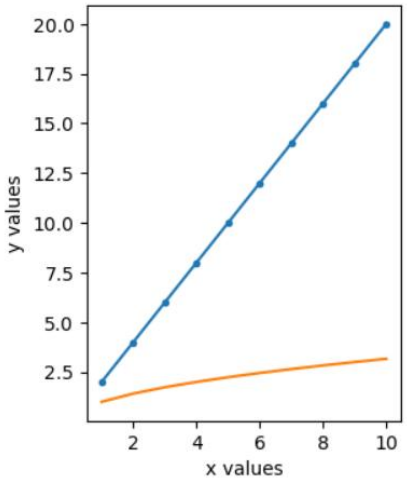
Line Graph

Line graph is used to plot data of two feature on different axis.

```
import numpy as np
import matplotlib.pyplot as plt

x=np.arange(1,11)
y=x*2

plt.figure(figsize=(3,4))
plt.plot(x,y,marker='.')
plt.plot(x,x**0.5)
plt.xlabel('x values')
plt.ylabel('y values')
plt.show()
```



plt - To implement matplotlib.pyplot
arange() - To list number from 1 to 10
figure() - To allocate size of the plot.
plot() - To provide arguments for axis data,x and y for line plot.
xlabel() - Label for x-axis
ylable() - Label for y-axis
Two separate linear relation of data is represented by using two plot() in the same figure.

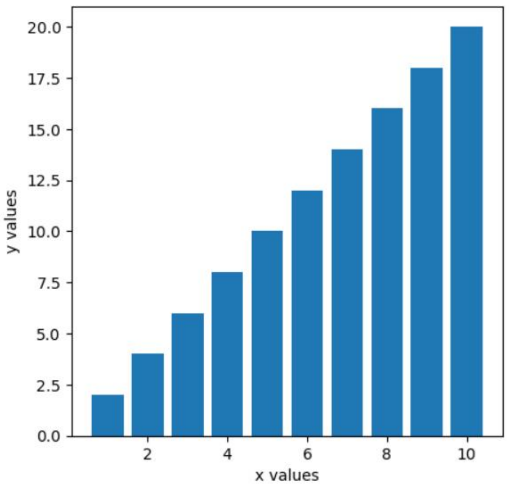
Use: Projects the relation between the data.

Bar Graph

Bar graph is used to represent the strength or value of different data. It can be plotted vertically or horizontally.

```
]: import numpy as np
   import matplotlib.pyplot as plt

   x=np.arange(1,11)
   y=x*2
   plt.figure(figsize=(5,5))
   plt.bar(x,y)
   plt.xlabel('x values')
   plt.ylabel('y values')
   plt.show()
```



plt - To implement matplotlib.pyplot
arange() - To list number from 1 to 10
figure() - To allocate size of the plot.
bar() - To provide arguments for axis data,x and y for bar graph.
xlabel() - Label for x-axis
ylabel() - Label for y-axis

y = x*2 gives values of list of x mutiplied by 2 seperately.

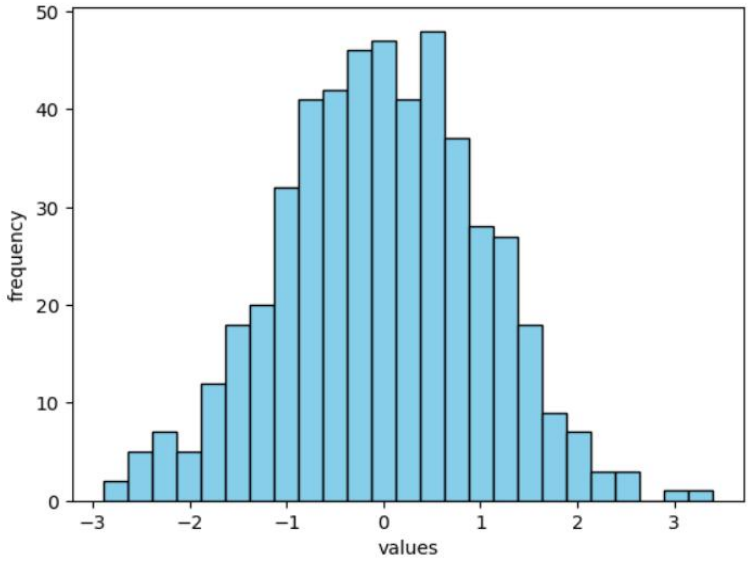
Use: Compare different types of data with respect to their strength, count, etc,.

Histogram

Histogram is used to show the frequency distribution of different data values throughout a range of values. It consists of bin values to aggregate a group of data.

```
import numpy as np
import matplotlib.pyplot as plt

np.random.randn(500)
plt.hist(np.random.randn(500),bins=25,color='skyblue',edgecolor='black')
plt.xlabel('values')
plt.ylabel('frequency')
plt.show()
```



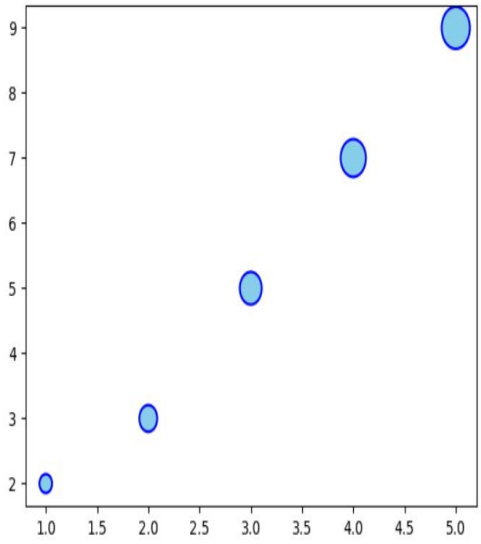
plt - To implement matplotlib.pyplot
np - To implement numpy
random.randn() - To generate list of 500 random numbers.
figure() - To allocate size of the plot.
hist() - To provide arguments for axis data,x and y for histogram as value and frequency of the values.
xlabel() - Label for x-axis
ylabel() - Label for y-axis

Scatter Plot

Scatter plot uses the data to show the relationship between them by plotting markers in data coordinates.

```
import numpy as np
import matplotlib.pyplot as plt

x=np.array([1,2,3,4,5])
y=np.array([2,3,5,7,9])
bubble=x*100
plt.scatter(x,y,s=bubble,color='skyblue',edgecolor='b',linewidth=1.5)
plt.show()
```



plt - To implement matplotlib.pyplot
np - To implement numpy
array() - To convert list into numpy array.
scatter() - To provide arguments for axis data,x and y for bar graph including color of bubbles, edgecolor of bubbles , and linewidth of the edge of bubbles.

Bubble=x*100 to multiply the values of list x with 100.

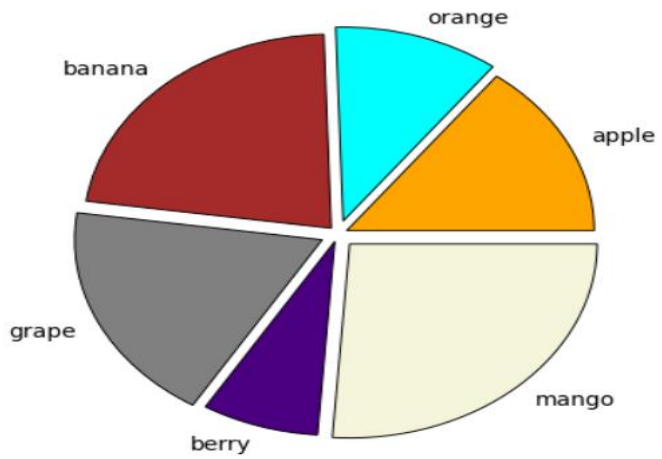
Use: Identifying outliers easily.

Pie Chart

Pie chart is a circular plot used to show the composition of types of data in categorical data. The whole circle represents 100% of the data.

```
import matplotlib.pyplot as plt

fruits = ['apple','orange','banana',
          'grape','berry','mango']
data = [23, 17, 35, 29, 12, 41]
colors = ("orange", "cyan", "brown",
          "grey", "indigo", "beige")
exp=[0.04,0.07]*3
wp={'linewidth':0.7,'edgecolor':'black'}
plt.pie(data,labels=fruits,colors=colors,explode=exp,wedgeprops=wp)
plt.show()
```



plt - To implement matplotlib.pyplot

pie() - To provide arguments for axis data for pie chart including labels color wedges, edgecolor of wedges, and explode to highlight categories seperately.

Fruit - Contains list of fruit name for label.

Data - List of value for each fruit.

Color - List of colors to be given to different wedge.

Exp - Contains list of values to define the stregth of explode

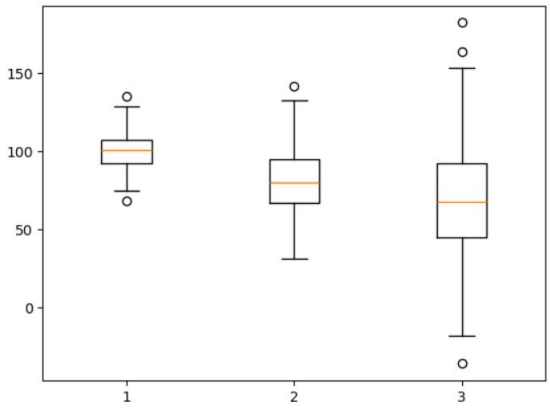
Wp - Contains wedge property for line width and edge color

Box Plot

Box plot is used to represent the summary of used data with a line divided box and extended line markers.

```
import numpy as np
import matplotlib.pyplot as plt

d1=np.random.normal(100,10,200)
d2=np.random.normal(80,20,200)
d3=np.random.normal(70,40,200)
data=[d1,d2,d3]
plt.boxplot(data)
plt.show()
```



plt - To implement matplotlib.pyplot
np - To implement numpy
Random.normal(x,y,z) - To generate numpy array.
boxplot() - To provide data fro the box plot.

d1,d2,d3 - Three lists are generated using random.normal() function.
Data - List that stores three series of list d1,d2,d3.

Orange line in centre shows mean.
The box represents the IQR Quartile1 , Quartile 2 and Quartile 3.
The marked lines represent the range of the the data.

The dots outside represent the outlier data which are not fit in the dataset.

Use: Find mean, median, IQR, Outliers.

Seaborn

Seaborn is an advanced python visualization library built on top of matplotlib and pandas. It is used to provide enhanced visual representation and easy to implement multiple types of plots in same figure.

Different type of plots in Seaborn

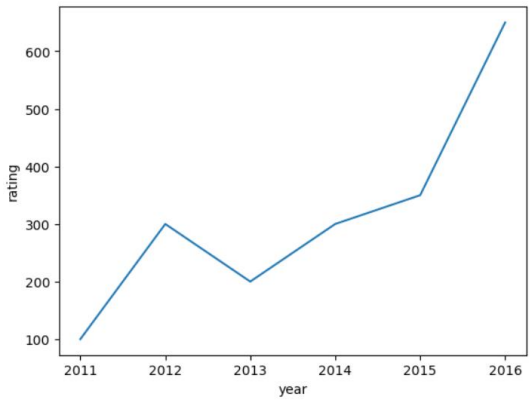
1. Line Graph
2. Bar Graph
3. Histogram
4. Scatter Plot
5. Count Plot
6. Box Plot
7. Strip Plot
8. Pair Plot
9. Heat Map
10. Cat Plot

Line Graph

Line graph is used to plot two data on different axis. Projects the relation between the data.

```
import pandas as pd
import seaborn as sns

x=[2011,2012,2013,2014,2015,2016]
y=[100,300,200,300,350,650]
data=pd.DataFrame({'year':x,
                   'rating':y})
sns.lineplot(x='year',
             y='rating',
             data=data)
```



pd - To implement pandas
sns - To implement seaborn
DataFrame() - To generate dataframe from given data.
lineplot() - Takes parameters as x, y, and data to plot.

x and y contain data values for the dataframe.

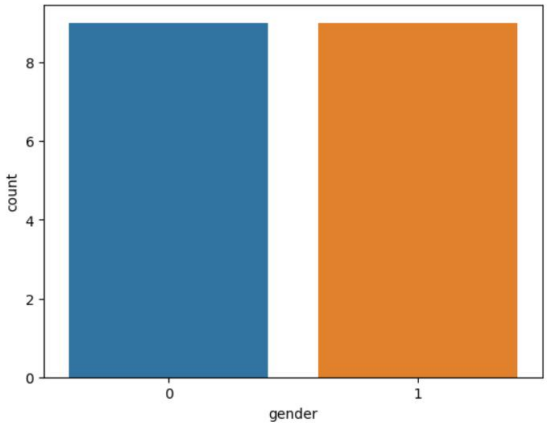
Bar Graph

Bar graph is used to represent the strength or value of different data. It can be plotted vertically or horizontally.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
gender_count=dt.groupby('gender').agg({'gender':'count'})
gender_count.columns=['count']
gender_count.reset_index(inplace=True)
```

```
sns.barplot(x='gender',
            y='count',
            data=gender_count)
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
barplot() - Takes parameters as x, y, and data to be plotted.

Dt - Contains the dataset.
gender_count - Sub datset from dt.
groupby() - To group particular value of columns.
agg() - The get count of each gender seperately.
columns - To changes the name of columns in dataset.
reset_index - To reset the index of dataset.
Inplace - Is given ‘True’ to retain the dataset index.

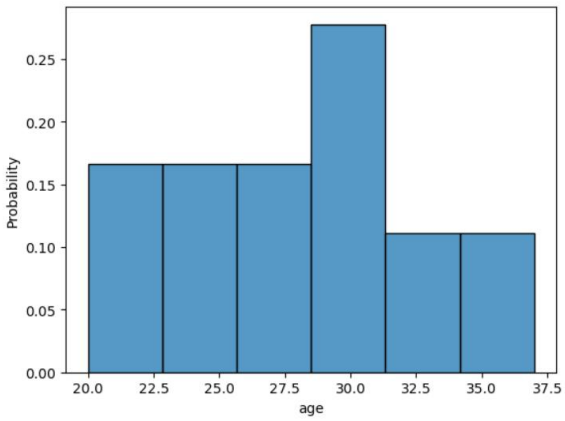
x - Given the data of Gender
y - Given the data of Count of each gender

Histogram

Histogram is used to show the distribution of different data values throughout a range of values. It consists of bin values to aggregate a group of data.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
sns.histplot(x='age',data=dt,
             stat='probability')
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
histplot() - Takes parameters as x, y, and data to plot.

dt - Contains the dataset.

x - Given the data values of age.
stat - Given the statistical property to be plotted in y-axis

Advantage: Allows specifying statistical arguments.

Scatter Plot

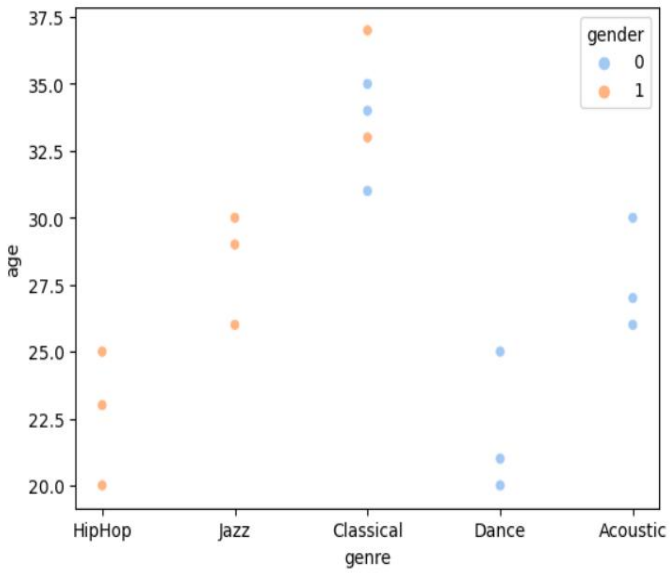
Scatter plot uses the data to show the relationship between them by plotting markers in data coordinates.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
dt.head()
```

	age	gender	genre
0	20	1	HipHop
1	23	1	HipHop
2	25	1	HipHop
3	26	1	Jazz
4	29	1	Jazz

```
sns.scatterplot(x='genre',
                y='age',
                data=dt,
                hue='gender',
                palette='pastel')
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
head() - To show the first 5 rows of dataset.
scatterplot() - Takes parameters as x, y, and data to plot with hue and color palette.

dt - Contains the dataset.

x - Given the data values of genre.
y - Given the data values of age.
hue - To show difference in gander through color.
palette - To define type of color used.

Advantage: Provide parameter for color palette.

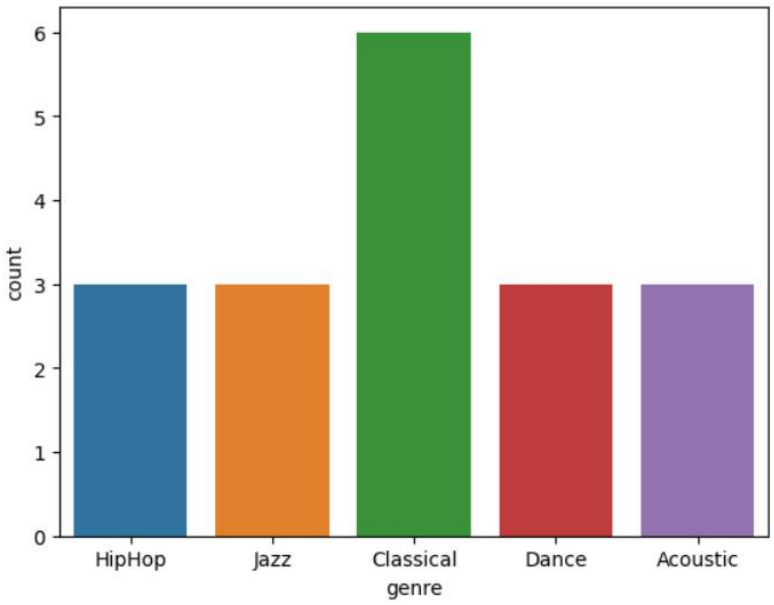
Count Plot

Count plot is used to represent the strength of categorical data like bar graph.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
sns.countplot(x='genre',
              data=dt)
```

```
<Axes: xlabel='genre', ylabel='count'>
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
countplot() - Takes parameters as x value and data to plot.

dt - Contains the dataset.
x - Given the data values of genre.

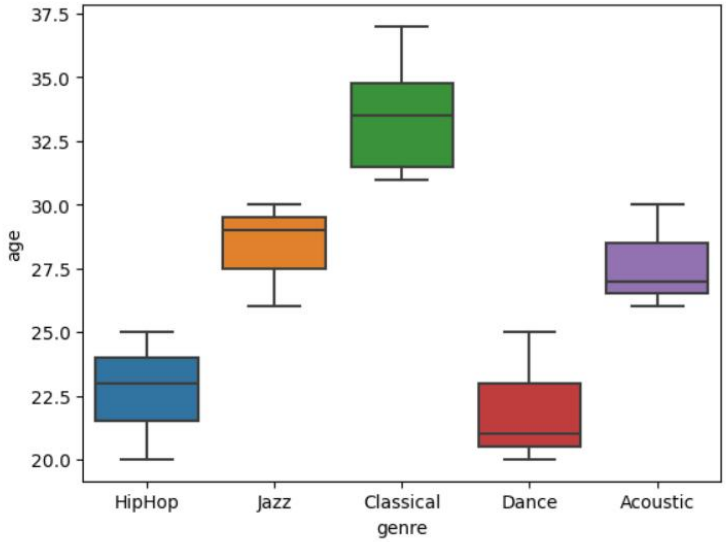
Advantage: Does not require to specify count property or to give color for data.

Box Plot

Box plot is used to represent the summary of used data. The mean, median, and IQR is represented with a line divided box and extended line markers.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
sns.boxplot(y='age',x='genre',
            data=dt)
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
boxplot() - Takes parameters as x, y, and data to plot.

dt - Contains the dataset.

x - Given the data values of genre.
y - Given the data values of age.

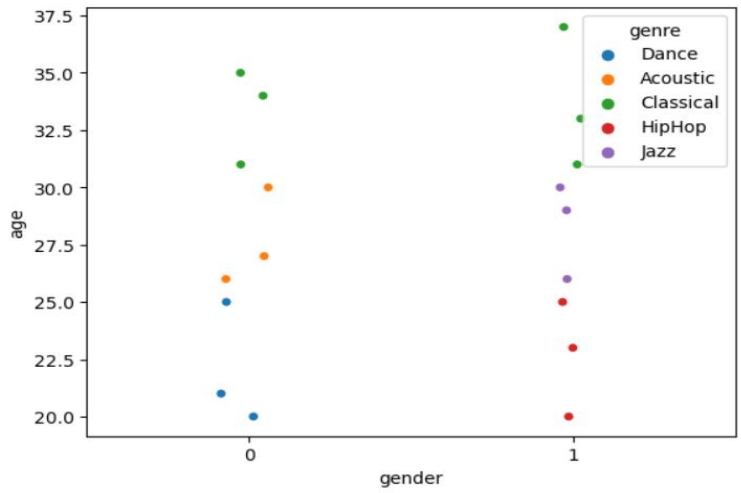
Advantage: Automatic color representation.

Strip Plot

Strip plot is similar to scatter plot but used to represent values of categorical data.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
sns.stripplot(x='gender',y='age',
              data=dt,hue='genre')
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
stripplot() - Takes parameters as x, y, and data to plot.

dt - Contains the dataset.

x - Given the data values of gender.
y - Given the data values of age.
hue - To show difference in genre through differenrt colors.

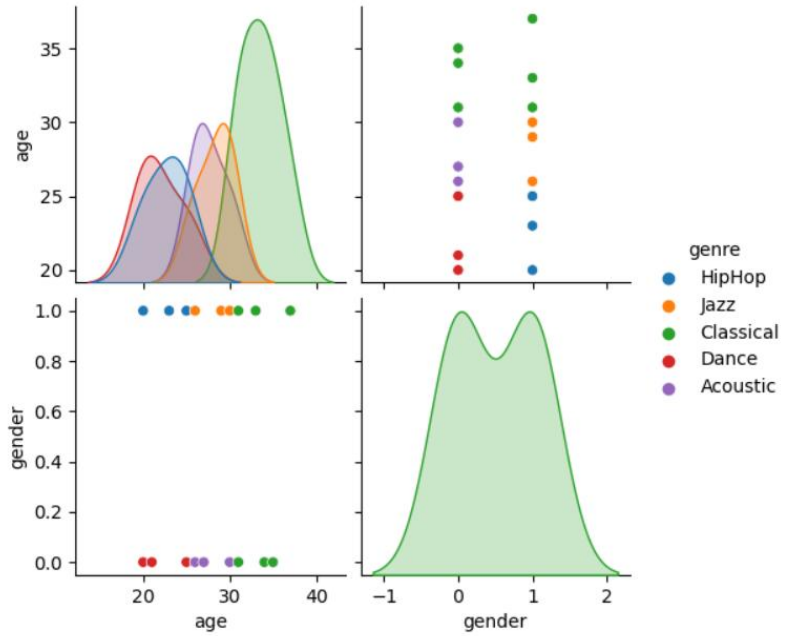
Advantage: Represent multivariate data easily.

Pair Plot

Pair plot is used to represent the plotting for each pair of feature in single figure.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
sns.pairplot(data=dt,hue='genre')
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
pairplot() - Takes parameter data to plot.

dt - Contains the dataset.
hue - To show difference in genre through differenrt colors.

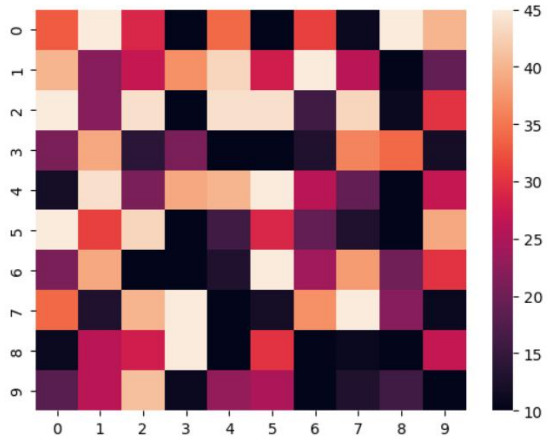
Advantage: Provide subplots for pattern analysis with each feature.

Heat Map

Heatmap is used to represent to strength of relation between the data features.

```
import numpy as np
import pandas as pd
import seaborn as sns

dt=np.random.randint(low=1,
                     high=50,
                     size=(10,10))
sns.heatmap(data=dt,vmin=10,vmax=45)
```



np - To implement numpy
pd - To implement pandas
sns - To implement seaborn
random.randint() - To generate a matrix of size (10,10), with values ranging from 1 to 50.
heatmap() - Takes parameters as data, vmin and vmax to plot.

dt - Contains the dataset.

vmax - Minimum value in the plot color range.
vmin - Maximum value in the plot color range.

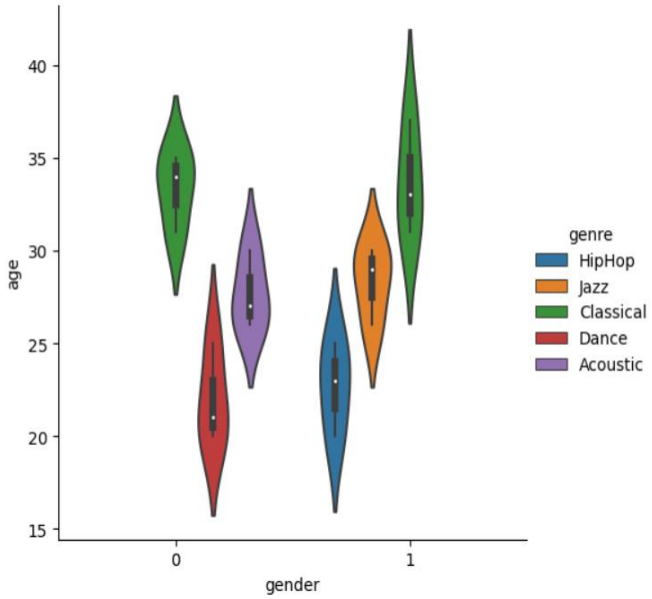
Advantage: Color coded representation of strength of relation.

Cat Plot

Cat plot is special plot in seaborn, with cat plot any other categorical plot can be defined with the help of "kind" parameter given with name of the type of plot.

```
import pandas as pd
import seaborn as sns

dt=pd.read_csv('music.csv')
sns.catplot(x='gender',y='age',
            hue='genre',data=dt,
            kind='violin')
```



pd - To implement pandas
sns - To implement seaborn
read_csv() - To get dataset from external file.
catplot() - Takes parameters x, y, data to plot with hue and kind of plot.

dt - Contains the dataset.
x - Given the values of gender.
y - Given the values of age.
kind - Given as "violin" to plot violin plot.
hue - To show difference in genre through differenrt colors.

Advantage: Easy to represent different types of plots.

Matplotlib v/s Seaborn

Matplotlib	Seaborn
Graphs with basic themes.	Graphs with advanced themes.
Uses long and complex syntax.	Short and simple syntax.
Compatible with Numpy and Pandas.	Only compatible with pandas.
Needed to unzip dataset.	Automatically unzip dataset.