```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

```
iris = pd.read_csv("/content/sample_data/iris.csv")
#first 5 rows of dataset
iris.head()
```

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

First 5 row belongs to setosa variety in the dataset

```
# Number of rows and columns
iris.shape
```

```
(150, 5)
```

150 rows and 5 columns are present

```
# details of attributes
iris.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal.length  150 non-null    float64
 1   sepal.width   150 non-null    float64
 2   petal.length  150 non-null    float64
 3   petal.width   150 non-null    float64
 4   variety       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

All properties except variety is of float type and variety is property of type object.

```
# Summary of dataset
iris.describe()
```

	sepal.length	sepal.width	petal.length	petal.width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

From the above number of non-empty rows in each numeric property is 150.The average value of all 4 numeric attribute is shown in 2nd row. The third row shows the standard deviation. The min show the minimum value limit. Next three shows the 25% percentile,the 50% percentile,the 75% percentile. Last row shows the maximum value limit.

```
# checking for null values
iris.isnull()
```

	sepal.length	sepal.width	petal.length	petal.width	variety

Can't find any null values in the dataset

```
# number of null values
iris.isnull().sum()
```

```
sepal.length    0
sepal.width     0
petal.length    0
petal.width     0
variety         0
dtype: int64
```

Zero null values in the dataset

```
# find the duplicate or repeated value
iris[iris.duplicated()]
```

	sepal.length	sepal.width	petal.length	petal.width	variety
142	5.8	2.7	5.1	1.9	Virginica

The above row has a duplicate

```
nd=iris.drop_duplicates(subset="variety")
nd
```

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
50	7.0	3.2	4.7	1.4	Versicolor
100	6.3	3.3	6.0	2.5	Virginica

Deleted all the rows with duplicated variety values and above shows the reslutant dataset after delete

```
# Check the number of variety is balanced or not
iris.value_counts("variety")
```

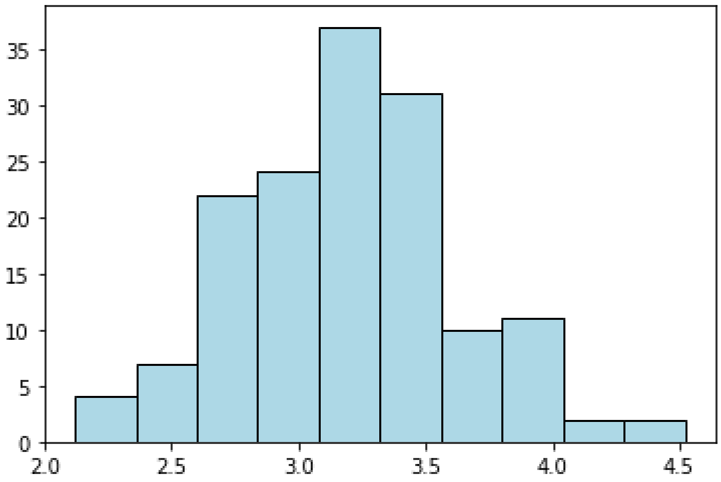
```
variety
Setosa        50
Versicolor    50
```

```
Virginica     50
dtype: int64
```

3 groups of varieties have equal number of values ie each have 50 of setosa, versicolor and virginica

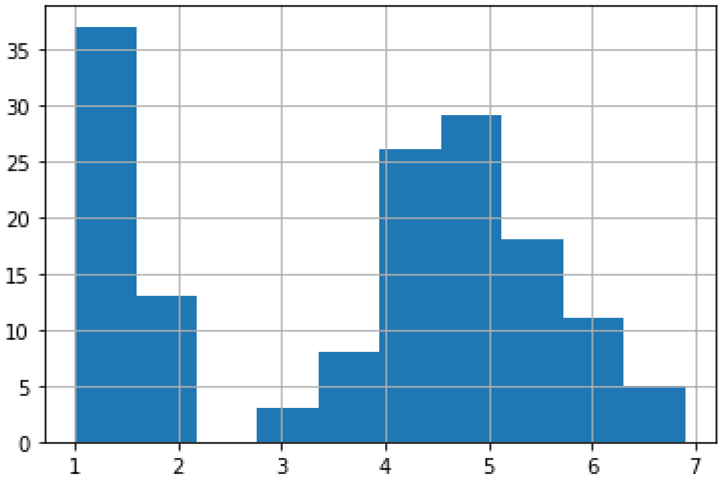
```
plt.hist(iris["sepal.width"], align='right', color='lightblue', edgecolor='black')
```

```
(array([ 4.,  7., 22., 24., 37., 31., 10., 11.,  2.,  2.]),
 array([2.  , 2.24, 2.48, 2.72, 2.96, 3.2 , 3.44, 3.68, 3.92, 4.16, 4.4 ]),
 <a list of 10 Patch objects>)
```



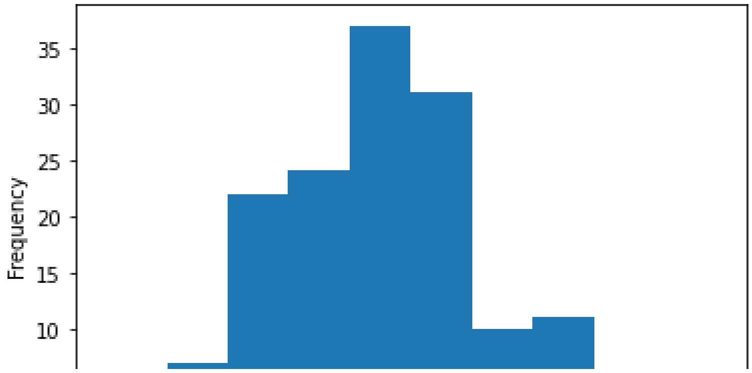
```
iris["petal.length"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcd36bc71d0>
```



```
iris["sepal.width"].plot(kind="hist")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcd37a03890>
```

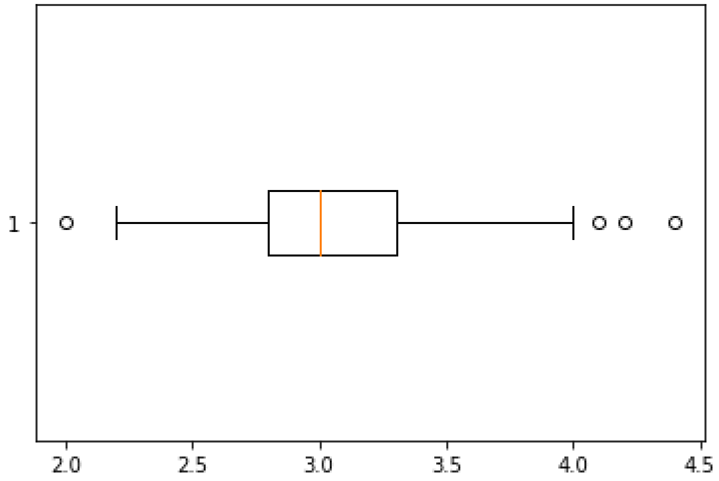


Histogram

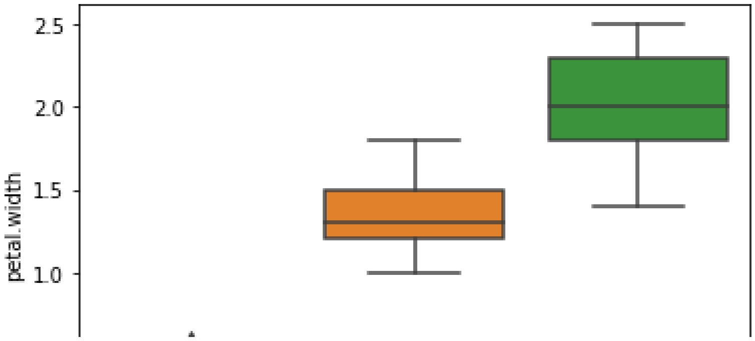
From the above histogram of sepal length compared to sepal width, we can see a constant increase in the first graph and a decrease in the second half along the x axis

```
plt.boxplot(iris["sepal.width"], vert=0)
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7f7fb81e2b90>,
  <matplotlib.lines.Line2D at 0x7f7fb76d7190>],
 'caps': [<matplotlib.lines.Line2D at 0x7f7fb76d76d0>,
  <matplotlib.lines.Line2D at 0x7f7fb76d7c10>],
 'boxes': [<matplotlib.lines.Line2D at 0x7f7fb771ac10>],
 'medians': [<matplotlib.lines.Line2D at 0x7f7fb76e01d0>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f7fb76e0710>],
 'means': []}
```

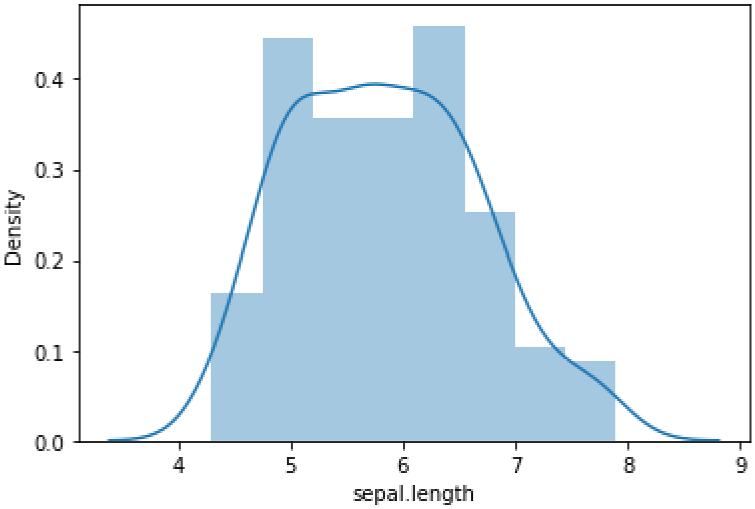


```
sns.boxplot(x="variety", y="petal.width", data=iris )
plt.show()
```

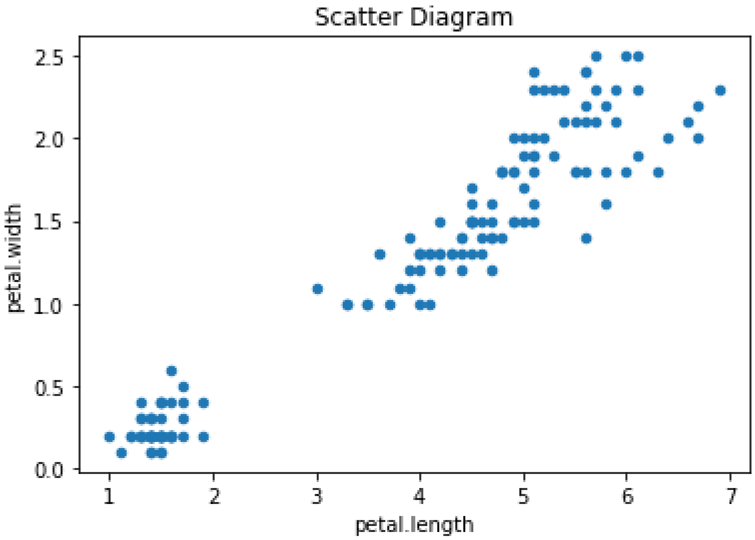


```
sns.distplot(iris['sepal.length'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcd36f5e550>
```

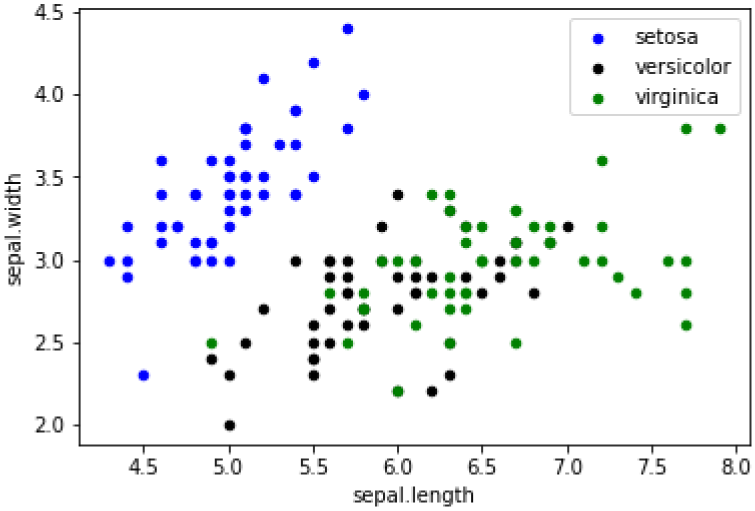


```
iris.plot(kind="scatter", x="petal.length", y="petal.width")
plt.title("Scatter Diagram")
plt.show()
```



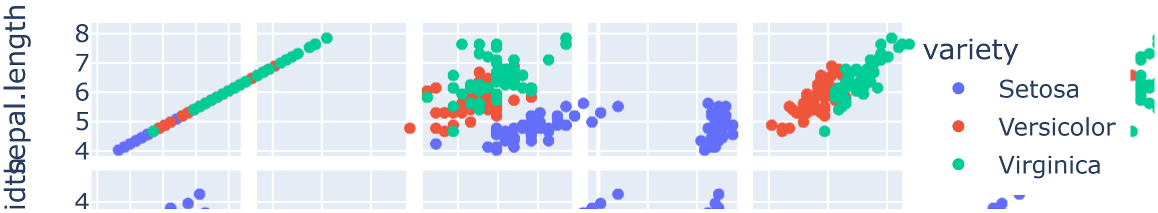
In the above diagram by comparing petal width and petal length we can see that there are no values inbetween 2 and 3 of petal length.

```
ax = iris[iris.variety=='Setosa'].plot.scatter(x='sepal.length', y='sepal.width',
                                                    color='blue', label='setosa')
iris[iris.variety=='Versicolor'].plot.scatter(x='sepal.length', y='sepal.width',
                                                color='black', label='versicolor', ax=ax)
iris[iris.variety=='Virginica'].plot.scatter(x='sepal.length', y='sepal.width',
                                                color='green', label='virginica', ax=ax)
plt.show()
```



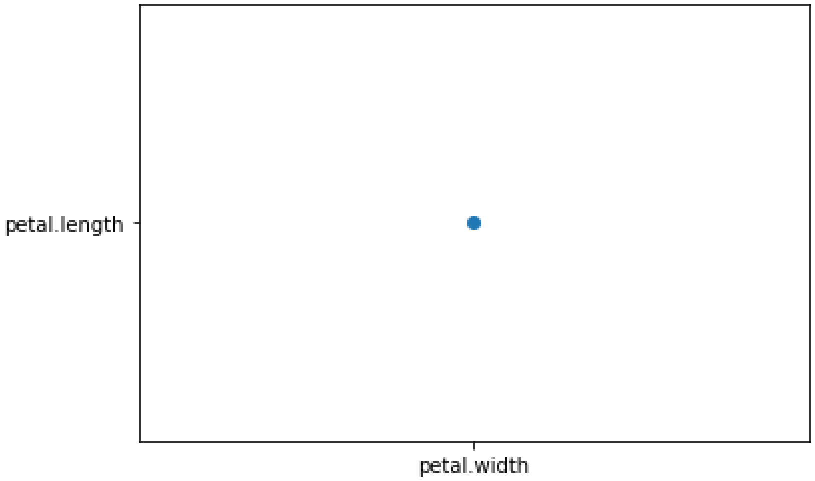
From the above diagram we can see that the mostly points of setosa is inbetween 4 to 6 of sepal length,versicoloris inbetween 5 and 7 and virginica is inbetween 5.5 and 8.

```
px.scatter_matrix(iris,color="variety")
```



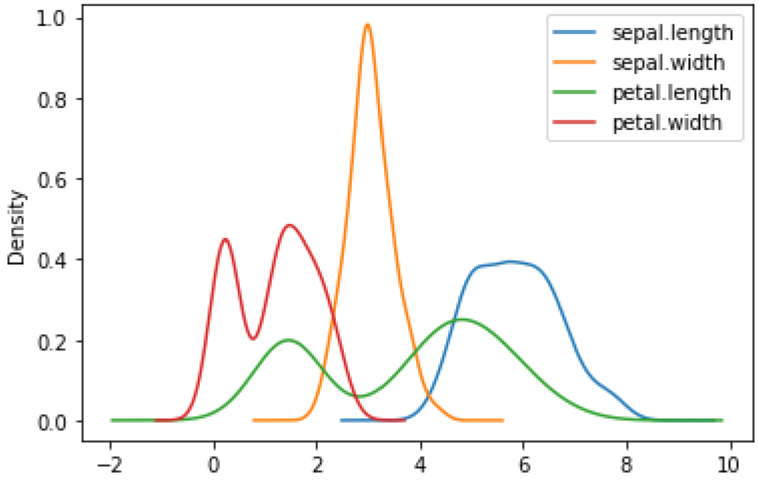
```
plt.scatter(x="petal.width",y="petal.length")
```

```
<matplotlib.collections.PathCollection at 0x7fcd36832510>
```

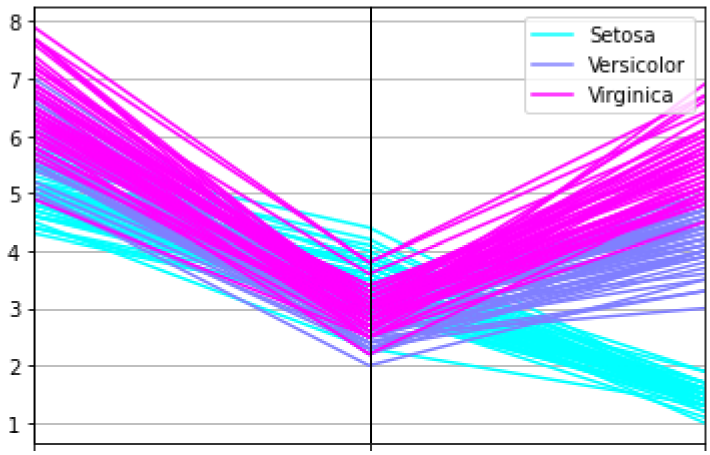


```
iris.plot(kind='density')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcd369e4150>
```



```
pd.plotting.parallel_coordinates(iris.drop("petal.width",axis=1), "variety",colormap='cool')
plt.show()
```



0s completed at 3:55 PM