Assignment Number – 10

Problem Statement:

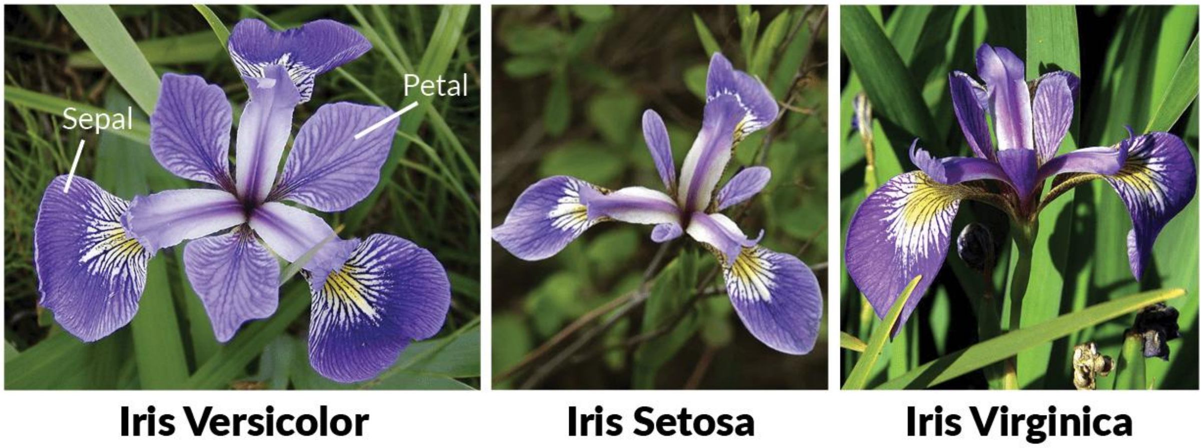
Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the datfaset to illustrate the feature distributions.
3. Create a box plot for each feature in the dataset.
4. Compare distributions and identify outliers.

Theory

1. **What is Iris Flower Dataset?**

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper ‘The use of multiple measurements in taxonomic problems’ as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor as shown in following figure). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

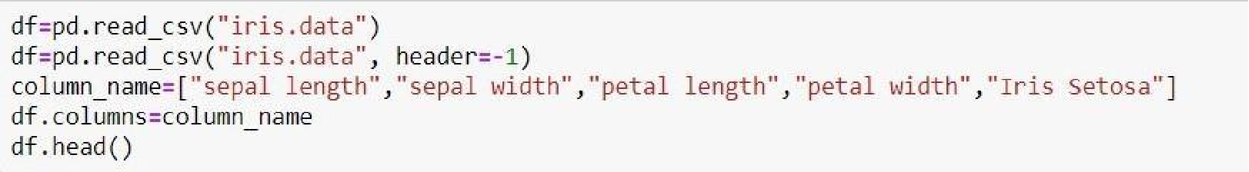


2. **What are features and their types?**

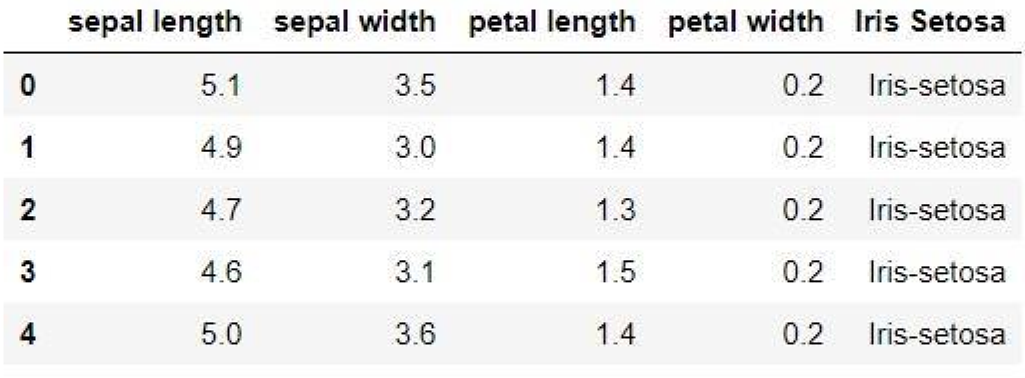
Iris flower dataset has four features

1) Sepal Length
2) Sepal Width
3) Petal Length

4) Petal Width

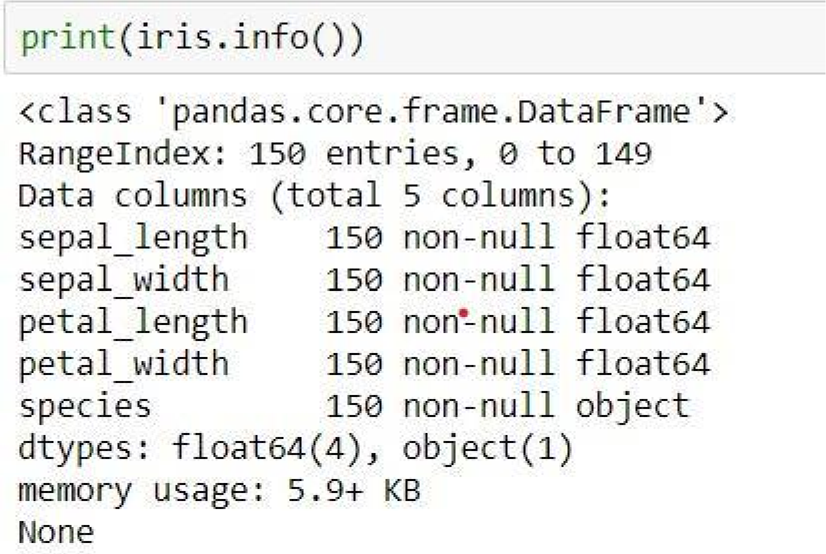


```
df=pd.read_csv("iris.data")
df=pd.read_csv("iris.data", header=-1)
column_name=["sepal length","sepal width","petal length","petal width","Iris Setosa"]
df.columns=column_name
df.head()
```



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

If we see that all the feature values are numerical.



```
print(iris.info())
```

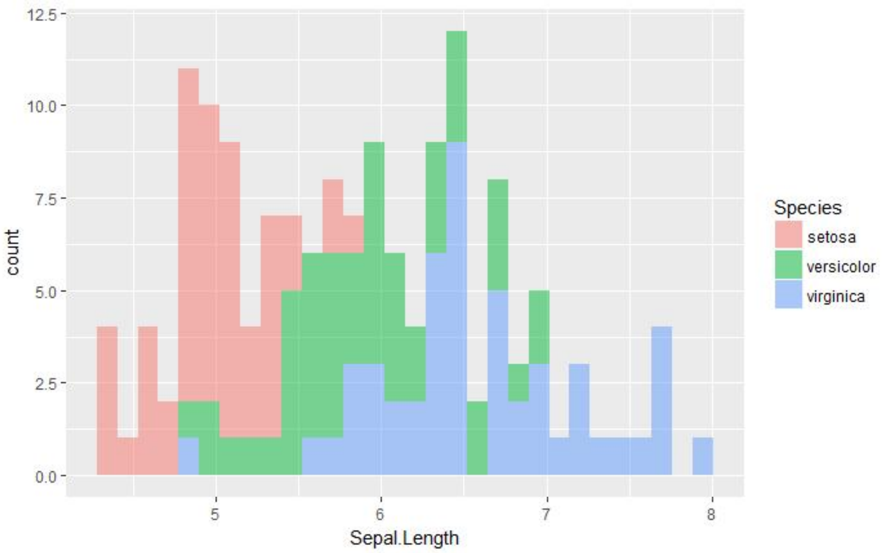
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
sepal_length    150 non-null float64
sepal_width     150 non-null float64
petal_length    150 non-null float64
petal_width     150 non-null float64
species         150 non-null object
dtypes: float64(4), object(1)
memory usage: 5.9+ KB
None
```

Histogram

Histograms represent the data distribution by forming bins along the range of the data and then drawing bars to show the number of observations that fall in each bin. Histograms are visualization tools that represent the distribution of a set of continuous data. In a histogram, the data is divided into a set of intervals or bins (usually on the x-axis) and the count of data points that fall into each bin corresponding to the height of the bar above that bin. These bins may or may not be equal in width but are adjacent (with no gaps).

```
require(ggplot2)

qplot(Sepal.Length, data=iris, geom='histogram', fill=Species, alpha=I(1/2))
```

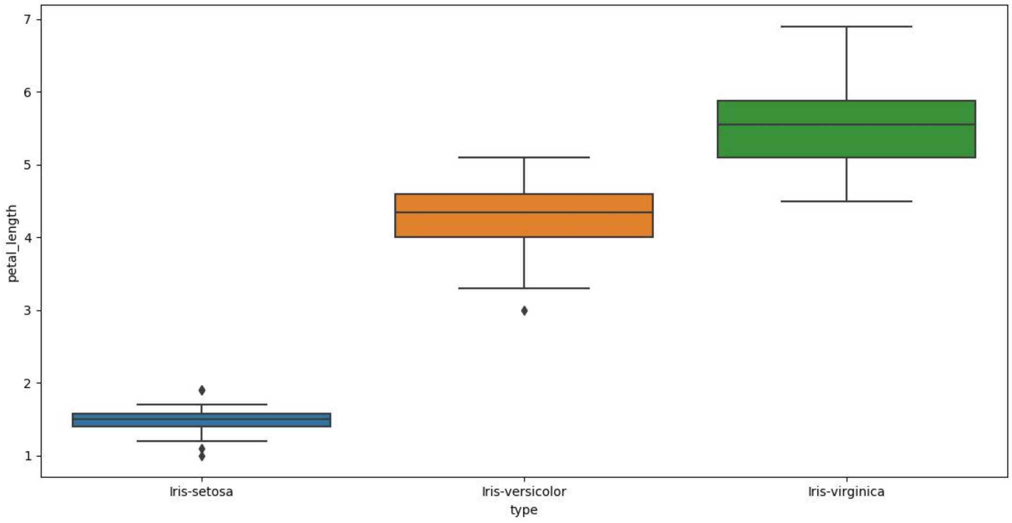


Similarly histogram of Sepal Width, Petal Length and Petal Width can be plot.

Boxplot

Boxplot can be drawn as follows:

```
sns.boxplot(x="type",y="petal_length",data=iris)
plt.show()
```



Similary boxplot of Sepal Length, Sepal Width and Petal Width can be drawn.

We see some outlier in Petal Length of Iris-Versicolor type.