**Data Visualization III**

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris (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 dataset to illustrate the feature distributions.
3. Create a box plot for each feature in the dataset.
4. Compare distributions and identify outliers.

# Load required libraries

In [2]:

**import** pandas **as** pd

**import** matplotlib.pyplot **as** pt

**import** numpy **as** np

**import** requests **as** r

# Load dataset

In [3]:

url**=**"https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data" r1**=**r.get(url)

df**=**pd.read\_csv(url)

Out[4]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **5.1** | **3.5** | **1.4** | **0.2** | **Iris-setosa** |
| **0** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **2** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **3** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **4** | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| **...** | ... | ... | ... | ... | ... |
| **144** | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| **145** | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |
| **146** | 6.5 | 3.0 | 5.2 | 2.0 | Iris-virginica |
| **147** | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| **148** | 5.9 | 3.0 | 5.1 | 1.8 | Iris-virginica |

149 rows × 5 columns

In [5]:

df.columns**=**["SepalLengthCm","SepalWidthCm","PetalLengthCm","PetalWidthCm","Species"]

In [6]:

df

Out[6]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SepalLengthCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** | **Species** |
| **0** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **2** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **3** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **4** | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| **...** | ... | ... | ... | ... | ... |
| **144** | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| **145** | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |
| **146** | 6.5 | 3.0 | 5.2 | 2.0 | Iris-virginica |
| **147** | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| **148** | 5.9 | 3.0 | 5.1 | 1.8 | Iris-virginica |

149 rows × 5 columns

df.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 149 entries, 0 to 148

Data columns (total 5 columns):

# Column Non-Null Count Dtype

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 SepalLengthCm | 149 | non-null |  | float64 |
| 1 SepalWidthCm | 149 | non-null |  | float64 |
| 2 PetalLengthCm | 149 | non-null |  | float64 |
| 3 PetalWidthCm | 149 | non-null |  | float64 |
| 4 Species | 149 | non-null |  | object |

dtypes: float64(4), object(1) memory usage: 5.9+ KB

In [8]:

df.isnull().sum()

Out[8]:

SepalLengthCm 0

SepalWidthCm 0

PetalLengthCm 0

PetalWidthCm 0

Species 0

dtype: int64

In [9]:

df.describe()

Out[9]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SepalLengthCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **count** | 149.000000 | 149.000000 | 149.000000 | 149.000000 |
| **mean** | 5.848322 | 3.051007 | 3.774497 | 1.205369 |
| **std** | 0.828594 | 0.433499 | 1.759651 | 0.761292 |
| **min** | 4.300000 | 2.000000 | 1.000000 | 0.100000 |
| **25%** | 5.100000 | 2.800000 | 1.600000 | 0.300000 |
| **50%** | 5.800000 | 3.000000 | 4.400000 | 1.300000 |
| **75%** | 6.400000 | 3.300000 | 5.100000 | 1.800000 |
| **max** | 7.900000 | 4.400000 | 6.900000 | 2.500000 |

df.drop\_duplicates()

Out[10]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SepalLengthCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** | **Species** |
| **0** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **2** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **3** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **4** | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| **...** | ... | ... | ... | ... | ... |
| **144** | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| **145** | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |
| **146** | 6.5 | 3.0 | 5.2 | 2.0 | Iris-virginica |
| **147** | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| **148** | 5.9 | 3.0 | 5.1 | 1.8 | Iris-virginica |

146 rows × 5 columns

# Create a histogram for each feature in the dataset to illustrate the feature distributions

* 1. SepalLengthCm : -

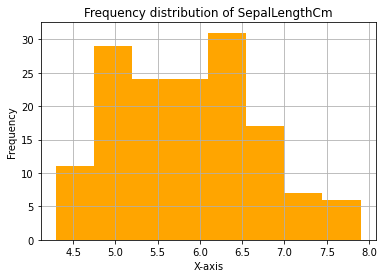
pt.title("Frequency distribution of SepalLengthCm") pt.xlabel("X-axis")

pt.ylabel("Frequency")

df["SepalLengthCm"].hist(color**=**"orange",bins**=**8)

Out[11]:

<AxesSubplot:title={'center':'Frequency distribution of SepalLengthCm'}, xla bel='X-axis', ylabel='Frequency'>



b)SepalWidthCm

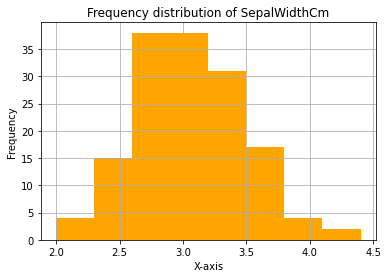
pt.title("Frequency distribution of SepalWidthCm") pt.xlabel("X-axis")

pt.ylabel("Frequency")

df["SepalWidthCm"].hist(color**=**"orange",bins**=**8)

Out[12]:

<AxesSubplot:title={'center':'Frequency distribution of SepalWidthCm'}, xlab el='X-axis', ylabel='Frequency'>



c) PetalLengthCm

In [13]:

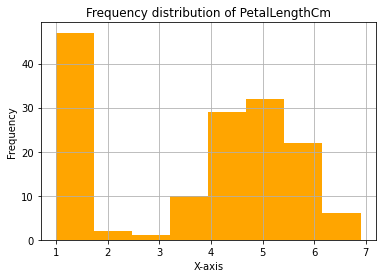
pt.title("Frequency distribution of PetalLengthCm") pt.xlabel("X-axis")

pt.ylabel("Frequency")

df["PetalLengthCm"].hist(color**=**"orange",bins**=**8)

Out[13]:

<AxesSubplot:title={'center':'Frequency distribution of PetalLengthCm'}, xla bel='X-axis', ylabel='Frequency'>



In [14]:

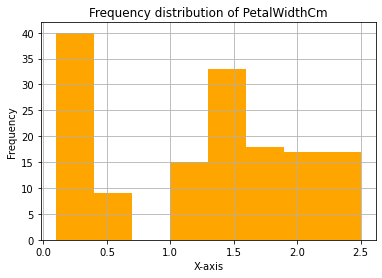
pt.title("Frequency distribution of PetalWidthCm") pt.xlabel("X-axis")

pt.ylabel("Frequency")

df["PetalWidthCm"].hist(color**=**"orange",bins**=**8)

Out[14]:

<AxesSubplot:title={'center':'Frequency distribution of PetalWidthCm'}, xlab el='X-axis', ylabel='Frequency'>



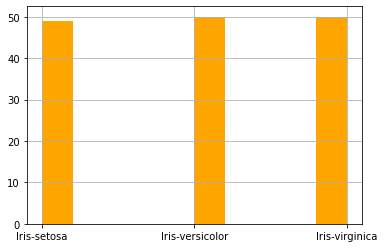
e)Species

In [15]:

df["Species"].hist(color**=**"orange")

Out[15]:

<AxesSubplot:>



# 4). Create a box plot for each feature in the dataset.

In [17]:

pt.boxplot(df["SepalLengthCm"])

Out[17]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2a4c453d5b0>,

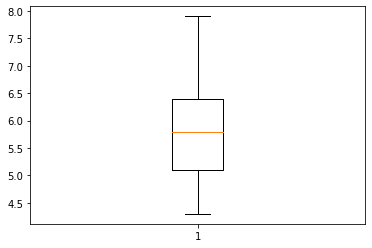
<matplotlib.lines.Line2D at 0x2a4c453d910>],

'caps': [<matplotlib.lines.Line2D at 0x2a4c453dc70>,

<matplotlib.lines.Line2D at 0x2a4c453dfd0>],

'boxes': [<matplotlib.lines.Line2D at 0x2a4c453d2b0>],

'medians': [<matplotlib.lines.Line2D at 0x2a4c454a370>], 'fliers': [<matplotlib.lines.Line2D at 0x2a4c454a6d0>], 'means': []}



SepalWidth

Out[18]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2a4c459f430>,

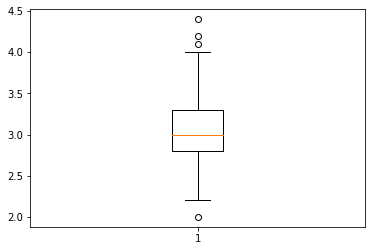
<matplotlib.lines.Line2D at 0x2a4c459f790>],

'caps': [<matplotlib.lines.Line2D at 0x2a4c459faf0>,

<matplotlib.lines.Line2D at 0x2a4c459fe50>],

'boxes': [<matplotlib.lines.Line2D at 0x2a4c459f0d0>],

'medians': [<matplotlib.lines.Line2D at 0x2a4c45ab1f0>], 'fliers': [<matplotlib.lines.Line2D at 0x2a4c45ab550>], 'means': []}



PetalLengthCm

Out[19]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2a4c45f9850>,

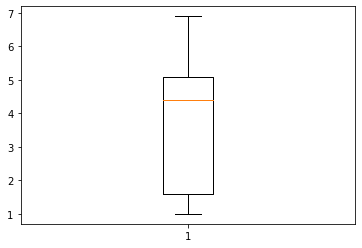
<matplotlib.lines.Line2D at 0x2a4c45f9bb0>],

'caps': [<matplotlib.lines.Line2D at 0x2a4c45f9f10>,

<matplotlib.lines.Line2D at 0x2a4c46062b0>],

'boxes': [<matplotlib.lines.Line2D at 0x2a4c45f94f0>],

'medians': [<matplotlib.lines.Line2D at 0x2a4c4606610>], 'fliers': [<matplotlib.lines.Line2D at 0x2a4c4606970>], 'means': []}



PetalWidth

Out[20]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2a4c562e250>,

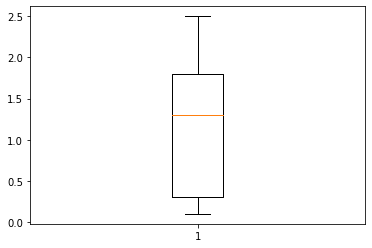
<matplotlib.lines.Line2D at 0x2a4c562e5b0>],

'caps': [<matplotlib.lines.Line2D at 0x2a4c562e910>,

<matplotlib.lines.Line2D at 0x2a4c562ec70>],

'boxes': [<matplotlib.lines.Line2D at 0x2a4c464feb0>],

'medians': [<matplotlib.lines.Line2D at 0x2a4c562efd0>], 'fliers': [<matplotlib.lines.Line2D at 0x2a4c5639370>], 'means': []}



# 5) Comparing all boxplot

In [23]:

df[{"SepalLengthCm","SepalWidthCm","PetalLengthCm","PetalWidthCm"}].boxplot() pt.title("Comparing all boxplots")

Out[23]:

Text(0.5, 1.0, 'Comparing all boxplots')

