Data Intensive Systems (DIS)
KBH-SW7 E25

6. Data Exploration

AALBORG
UNIVERSITY

Agenda

- Introduction to Data Exploration
  - Object of Data Exploration
  - Datasets
- Descriptive Statistics
- Data Visualization
- Data Exploration Process

Introduction to Data Exploration

- Data exploration broadly classified into two types—descriptive statistics and data visualization.
  - Descriptive statistics is the process of condensing key characteristics of the dataset into simple numeric metrics.
  - Visualization is the process of projecting the data, or parts of it, into multi-dimensional space or abstract images.
- Data exploration helps with
  - Understanding data better, to prepare the data in a way that makes advanced analysis possible,
  - Get the necessary insights from the data faster than using advanced analytical techniques.
- Data exploration provides
  - A set of tools to obtain fundamental understanding of a dataset.
  - Grasping the structure of the data, the distribution of the values, and the presence of extreme values and the interrelationships between the attributes in the dataset.
  - Guidance on applying the right kind of further statistical and data science treatment

Object of Data Exploration

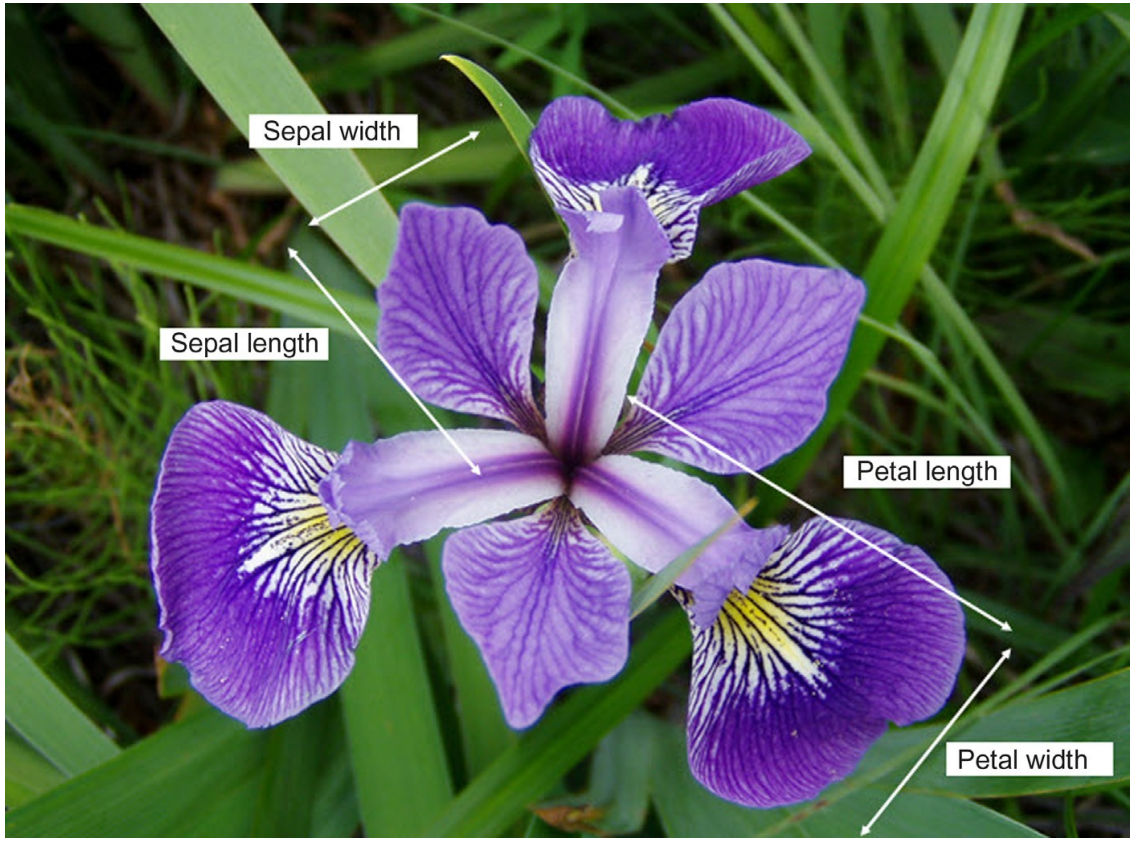
- Data understanding
- Data preparation
- Data mining tasks
- Interpreting the results

Object of Data Exploration

- Data understanding
  - Provides a high-level overview of each attribute (also called variable) in the dataset and the interaction between the attributes.
  - Answers the questions like what is the typical value of an attribute or how much do the data points differ from the typical value, or presence of extreme values.
- Data preparation
  - Dataset has to be prepared for handling any of the anomalies that may be present in the data including outliers, missing values, or highly correlated attributes.
  - Some data mining algorithms do not work well when input attributes are correlated with each other. Thus, correlated attributes need to be identified and removed
- Data mining tasks
  - Basic data exploration can sometimes substitute the entire data mining process. For example, scatterplots can identify clusters in low-dimensional data or can help develop regression or classification models with simple visual rules.
- Interpreting the results
  - Finally, data exploration is used in understanding the prediction, classification, and clustering of the results of the data mining process.
  - Histograms help to comprehend the distribution of the attribute and can also be useful for visualizing numeric prediction, error rate estimation, etc.

Datasets

- The most popular datasets used to learn data science is probably the **Iris** dataset, introduced by Ronald Fisher, in his seminal work on discriminant analysis.
- The **Iris** dataset contains 150 observations of three different species, *setosa*, *virginica*, and *versicolor*, with 50 observations each.
- Each observation consists of four attributes: *sepal length*, *sepal width*, *petal length*, and *petal width*. The fifth attribute, the *label*, is the name of the species observed.



Datasets

- All four attributes in the **Iris** dataset are numeric continuous values measured in centimeters.
- One of the species, *setosa*, can be easily distinguished from the other two using simple rules like the petal length is less than 2.5 cm.
- Separating the *virginica* and *versicolor* classes requires more complex rules that involve more attributes.

Datasets

- Data types
  - Data come in different formats and types. Understanding the properties of each attribute or feature provides information about what kind of operations can be performed on that attribute.
  - For example, the temperature in weather data can be expressed as any of the following formats:
    - Numeric (31°C, 34°C) or Fahrenheit (101°F, 104°F).
    - Ordered labels as in hot, mild, or cold.
    - Number of days within a year below 0°C (10 days in a year below freezing).

Datasets

- Categorical or Nominal
  - Categorical data types are attributes treated as distinct symbols or just names. The color of the iris of the human eye is a categorical data type because it takes a value like black, green, blue, gray, etc.
  - There is no direct relationship among the data values, and hence, mathematical operators except the logical or “is equal” operator cannot be applied. They are also called a nominal or polynominal data type, derived from the Latin word for name
  - An ordered nominal data type is a special case of a categorical data type where there is some kind of order among the values. An example of an ordered data type is temperature expressed as hot, mild, cold

Descriptive Statistics

- Descriptive statistics refers to the study of the aggregate quantities of a dataset.
- Some examples of descriptive statistics include average annual income, median home price in a neighborhood, range of credit scores of a population, etc.

Characteristics of the Dataset	Measurement Technique
Center of the dataset	Mean, median, and mode
Spread of the dataset	Range, variance, and standard deviation
Shape of the distribution of the dataset	Symmetry, skewness, and kurtosis

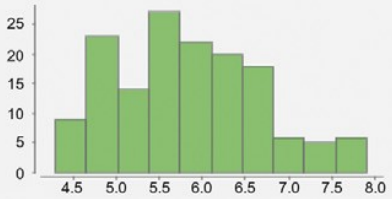
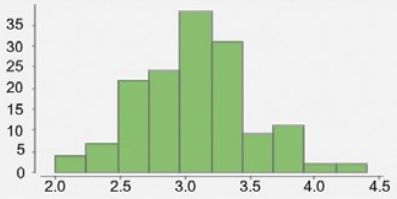
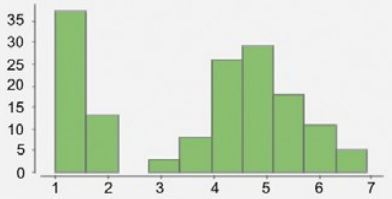
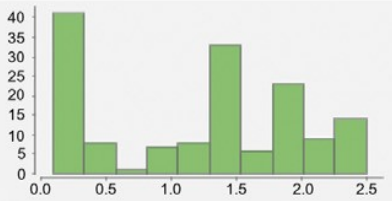
Univariate Descriptive Statistics

- Univariate data exploration denotes analysis of one attribute at a time. This technique only can be used for understanding the separate attributes without the connection with the other attributes.
- Measure of Central Tendency
  - The objective of finding the central location of an attribute is to quantify the dataset with one central or most common number.
  - Mean: The mean is the arithmetic average of all observations in the dataset.
  - Median: The median is the value of the central point in the distribution.
  - Mode: The mode is the most frequently occurring observation.
- Measure of Spread
  - Range: The range is the difference between the maximum value and the minimum value of the attribute.
  - Deviation: The variance and standard deviation measures the spread, by considering all the values of the attribute. Deviation is simply measured as the difference between any given value (x_i) and the mean of the sample (μ).

Univariate Descriptive Statistics

Observation	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.1	1.5	0.1
…	…	…	…	…
49	5	3.4	1.5	0.2
50	4.4	2.9	1.4	0.2
Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Mean	5.006	3.418	1.464	0.244
Median	5.000	3.400	1.500	0.200
Mode	5.100	3.400	1.500	0.200
Range	1.500	2.100	0.900	0.500
Standard deviation	0.352	0.381	0.174	0.107
Variance	0.124	0.145	0.030	0.011

Univariate Descriptive Statistics

Sepal Length	Real	0	 Open chart	Min 4.300	Max 7.900	Average 5.843	Deviation 0.828
Sepal Width	Real	0	 Open chart	Min 2	Max 4.400	Average 3.054	Deviation 0.434
Petal Length	Real	0	 Open chart	Min 1	Max 6.900	Average 3.759	Deviation 1.764
Petal Width	Real	0	 Open chart	Min 0.100	Max 2.500	Average 1.199	Deviation 0.763

Multivariate Descriptive Statistics

- Multivariate exploration is the study of more than one attribute in the dataset simultaneously.
- This technique is critical to understanding the relationship between the attributes, which is central to data science methods.

Multivariate Descriptive Statistics

- Central Data Point
  - In the Iris dataset, each data point as a set of all the four attributes can be expressed:
    - observation *i*: {*sepal length*, *sepal width*, *petal length*, *petal width*}
  - For example, observation one: {5.1, 3.5, 1.4, 0.2}. This observation point can also be expressed in four-dimensional Cartesian coordinates.
  - For the Iris dataset, the central mean point is {5.006, 3.418, 1.464, 0.244}.

Multivariate Descriptive Statistics

```
import pandas as pd

# Reading the CSV file

df = pd.read_csv("Iris.csv")

# Printing top 5 rows

df.head()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             150 non-null    int64
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
```

Multivariate Descriptive Statistics

```
df.describe()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

```
df.value_counts("Species")
```

```
Species
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
dtype: int64
```

Multivariate Descriptive Statistics

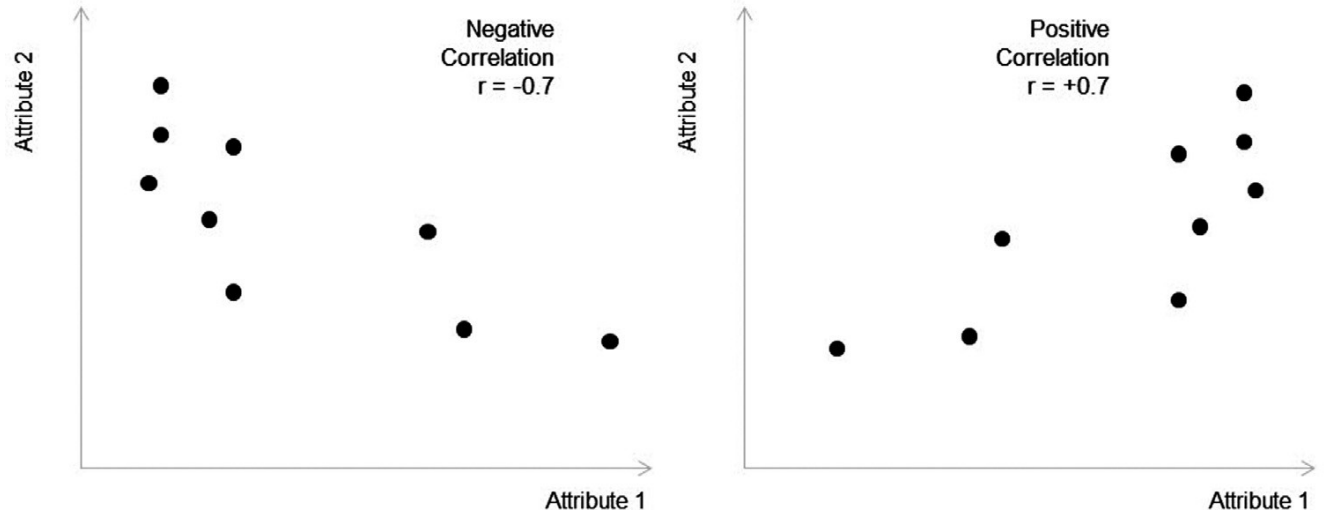
- Correlation
  - Correlation measures the statistical relationship between two attributes, particularly dependence of one attribute on another attribute.
  - When two attributes are highly correlated with each other, they both vary at the same rate with each other either in the same or in opposite directions.
  - Correlation between two attributes is commonly measured by the Pearson correlation coefficient (r), $-1 \leq r \leq 1$, which measures the strength of linear dependence.
  - A value closer to 1 or -1 indicates the two attributes are highly correlated, with perfect correlation at 1 or -1. A correlation value of 0 means there is no linear relationship between two attributes.

Multivariate Descriptive Statistics

- Correlation
  - The Pearson correlation coefficient between two attributes x and y is calculated with the formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N \times s_x \times s_y}$$

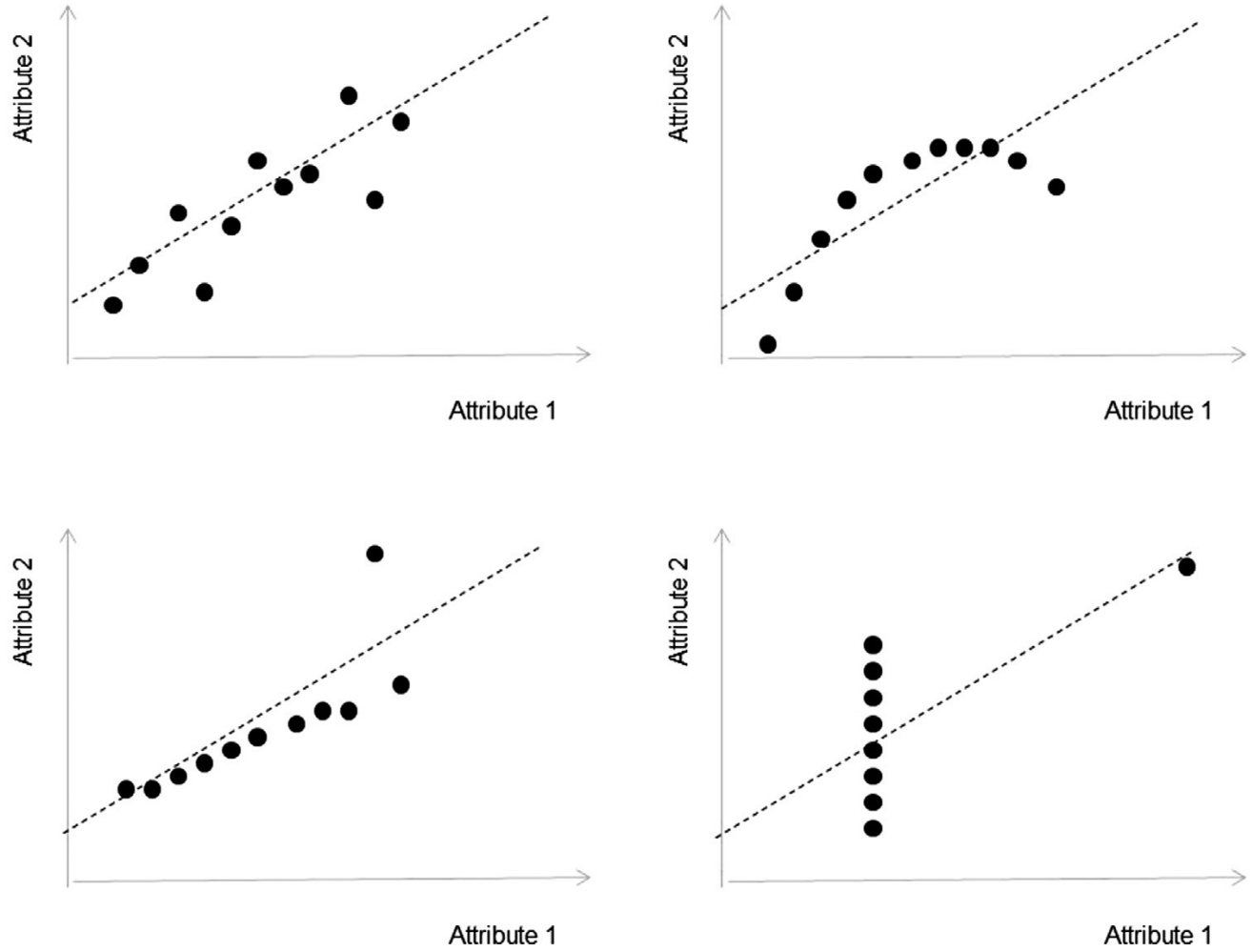
where s_x and s_y are the standard deviations of random variables x and y, respectively.



Data Visualization

- Visualizing data is one of the most important techniques of data discovery and exploration.
- The visual representation of data provides easy comprehension of complex data with multiple attributes and their underlying relationships.
- The motivation for using data visualization includes:
  - Comprehension of dense information.
  - Relationships.

Data Visualization



Data Visualization

- Univariate Visualization
  - Histogram: a histogram is one of the most basic visualization techniques to understand the frequency of the occurrence of values.
  - Quartile: a box whisker plot is a simple visual way of showing the distribution of a continuous variable with information such as quartiles, median, and outliers,
  - Distribution Chart: for continuous numeric attributes like petal length, instead of visualizing the actual data in the sample, its normal distribution function can be visualized instead.

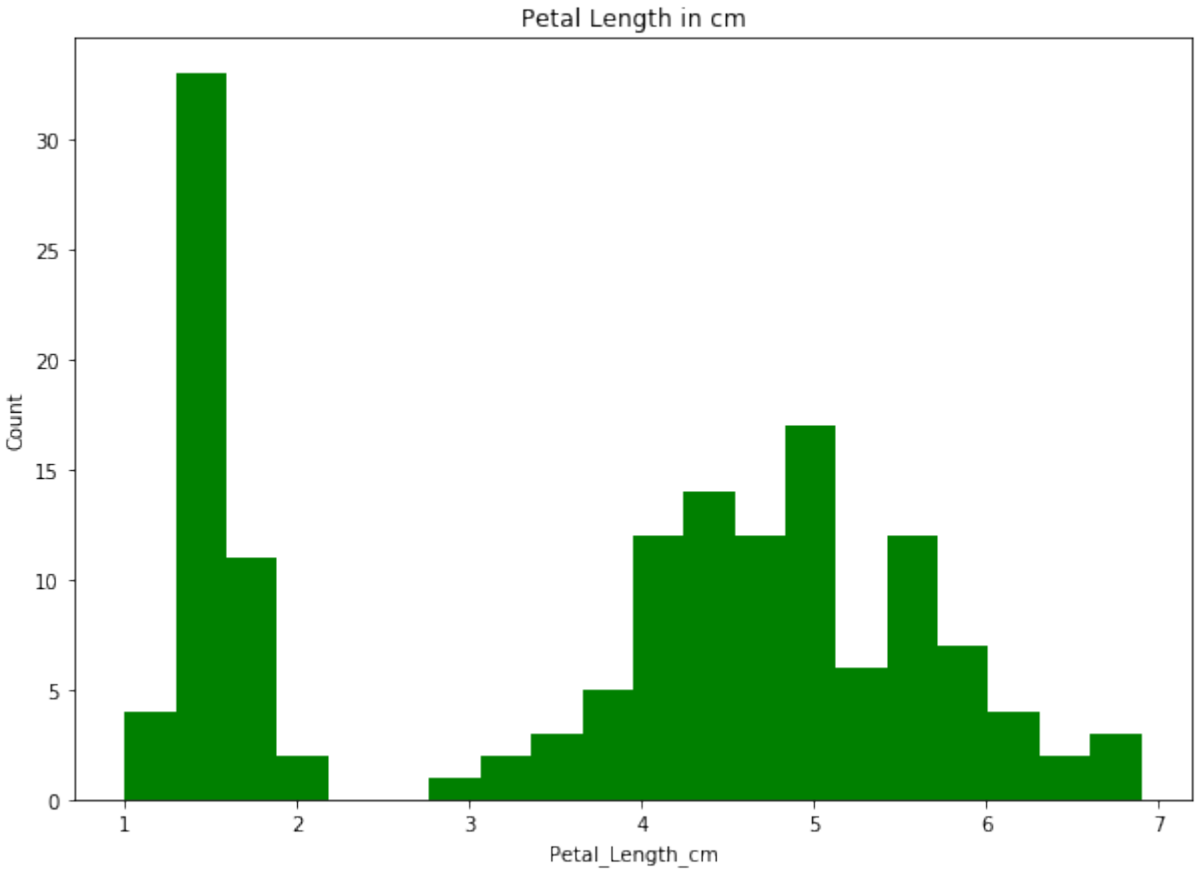
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{(x-\mu)^2/2\sigma^2}$$

where μ is the mean of the distribution and σ is the standard deviation of the distribution.

Data Visualization

- Univariate Visualization

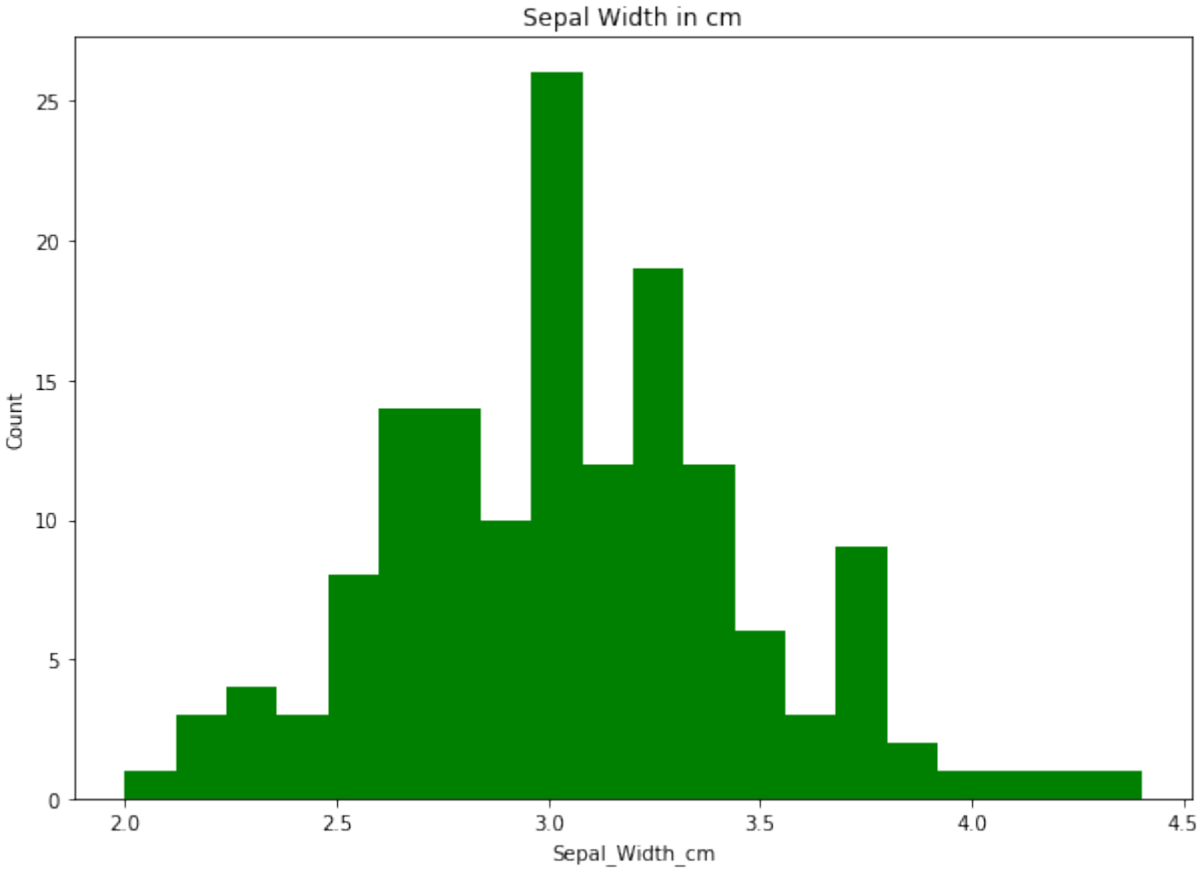
```
plt.figure(figsize = (10, 7))
x = data.PetalLengthCm
plt.hist(x, bins = 20, color =
"green")
plt.title("Petal Length in cm")
plt.xlabel("Petal_Length_cm")
plt.ylabel("Count")
plt.show()
```



Data Visualization

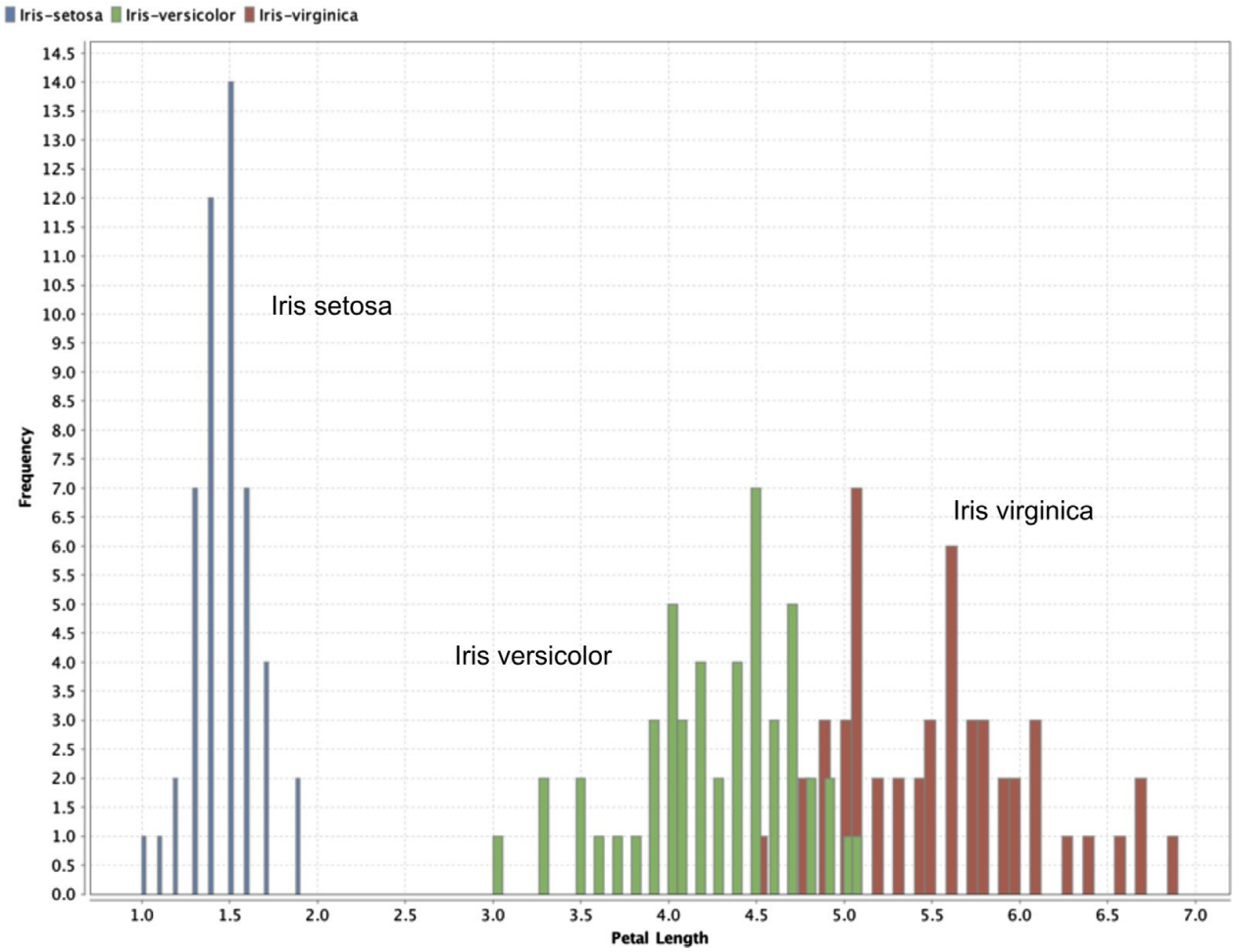
- Univariate Visualization

```
plt.figure(figsize = (10, 7))
x = data.SepalWidthCm
plt.hist(x, bins = 20, color =
"green")
plt.title(“Sepal Width in cm")
plt.xlabel(“Sepal_Width_cm")
plt.ylabel("Count")
plt.show()
```



Data Visualization

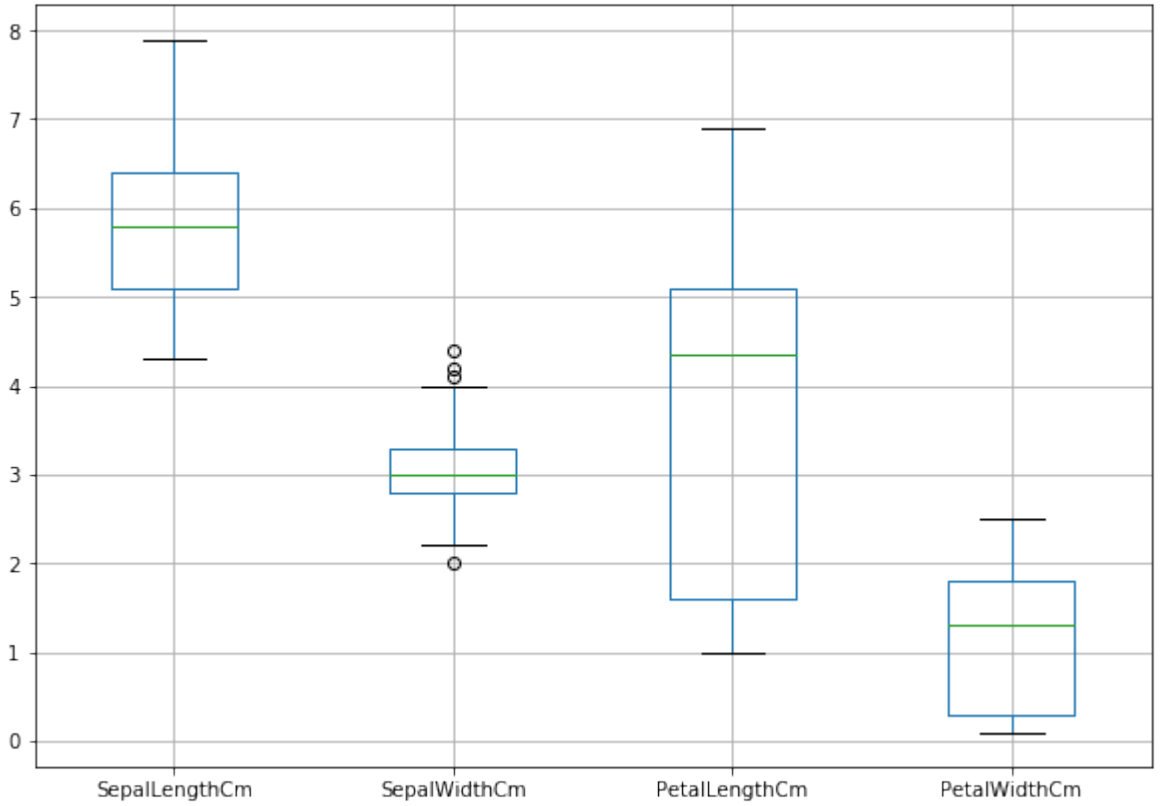
- Univariate Visualization



Data Visualization

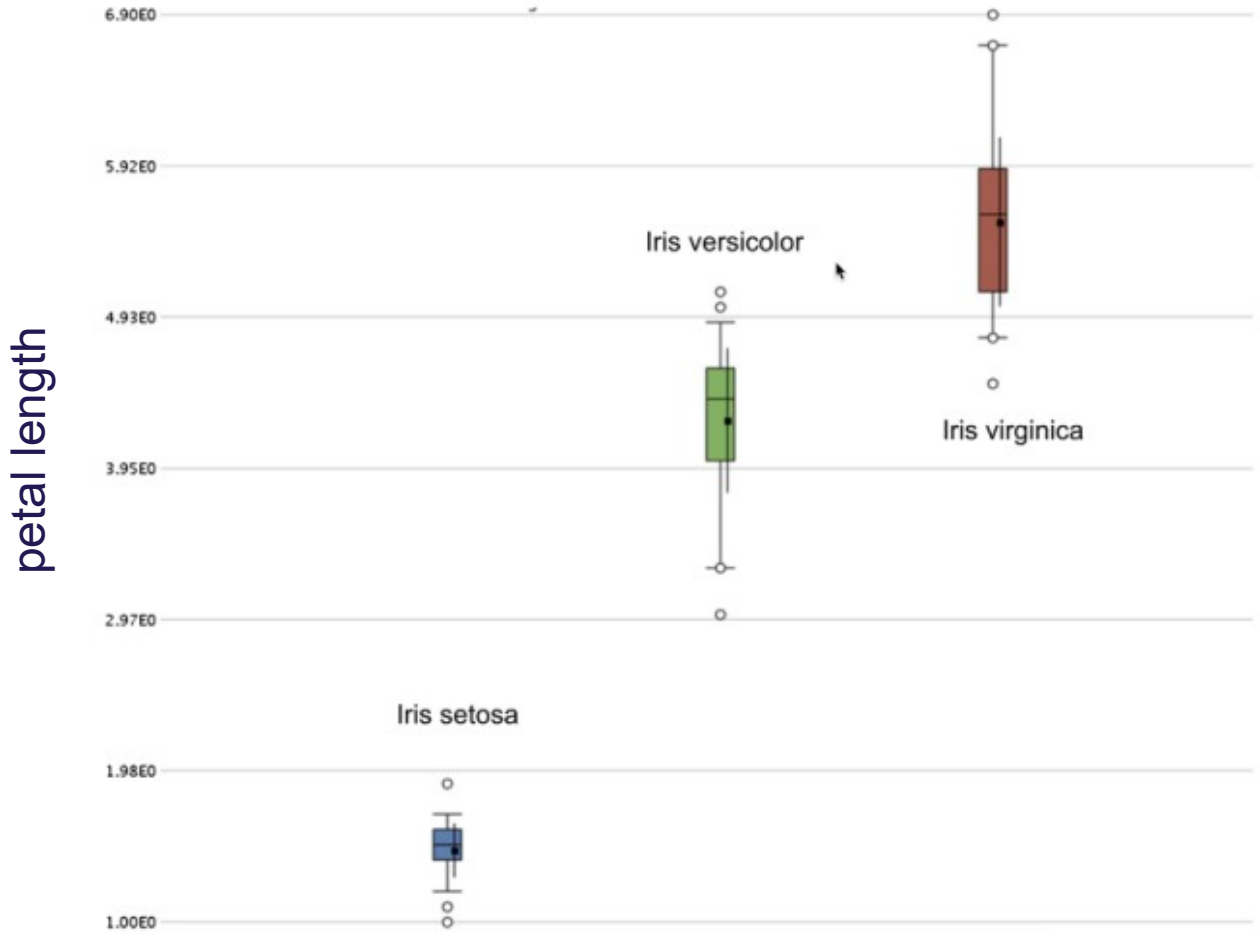
- Univariate Visualization

```
# removing Id column
new_data = data[["SepalLengthCm",
"SepalWidthCm", "PetalLengthCm",
"PetalWidthCm"]]
print(new_data.head())
plt.figure(figsize = (10, 7))
new_data.boxplot()
```



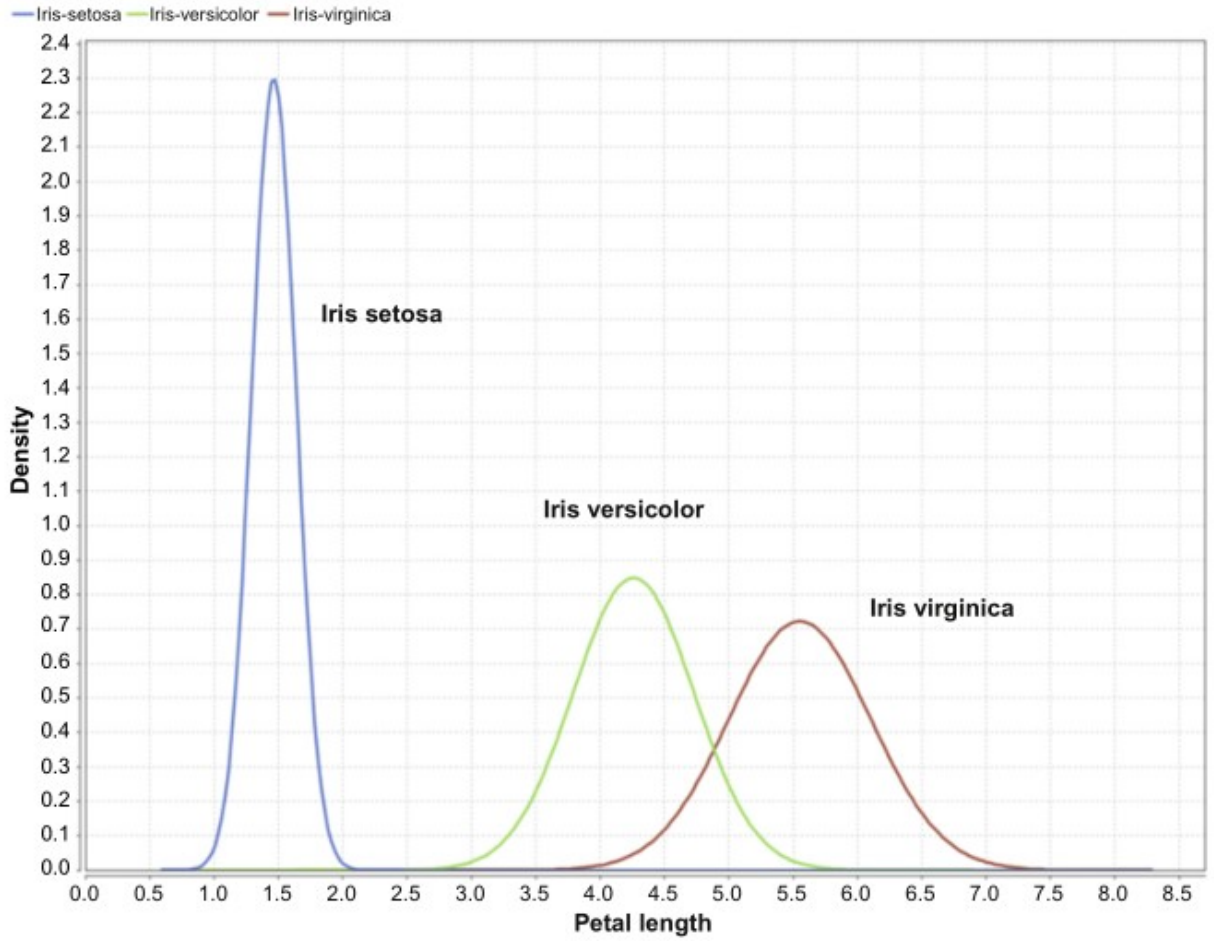
Data Visualization

- Univariate Visualization



Data Visualization

- Univariate Visualization



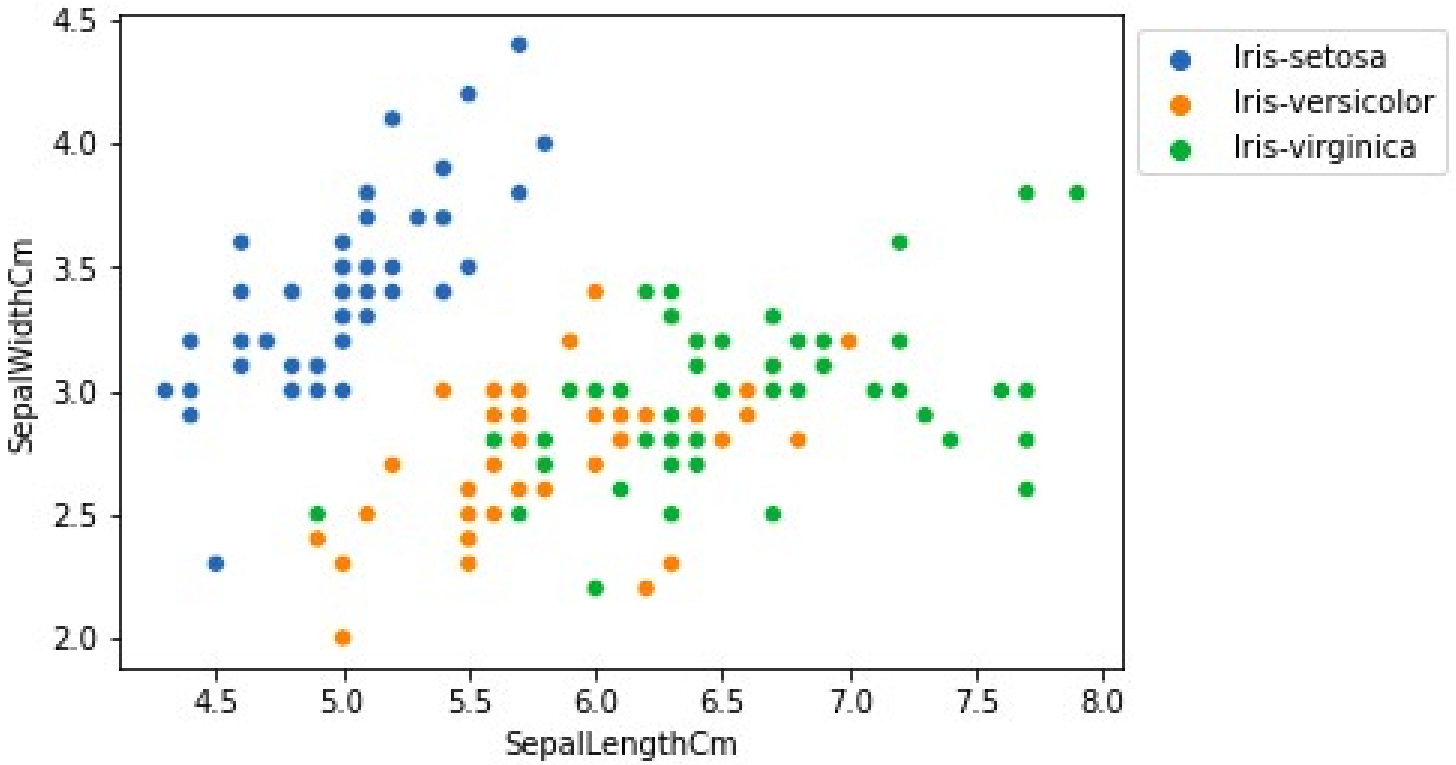
Data Visualization

- Multivariate Visualization
  - Scatterplot: in a scatterplot, the data points are marked in Cartesian space with attributes of the dataset aligned with the coordinates.
  - Scatter multiple: a scatter multiple is an enhanced form of a simple scatterplot where more than two dimensions can be included in the chart and studied simultaneously.
  - Scatter Matrix: a scatter matrix solves this need by comparing all combinations of attributes with individual scatterplots and arranging these plots in a matrix.
  - Bubble Chart: a bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point.
  - Density Chart: density charts are similar to the scatterplots, with one more dimension included as a background color.

Data Visualization

- Multivariate Visualization

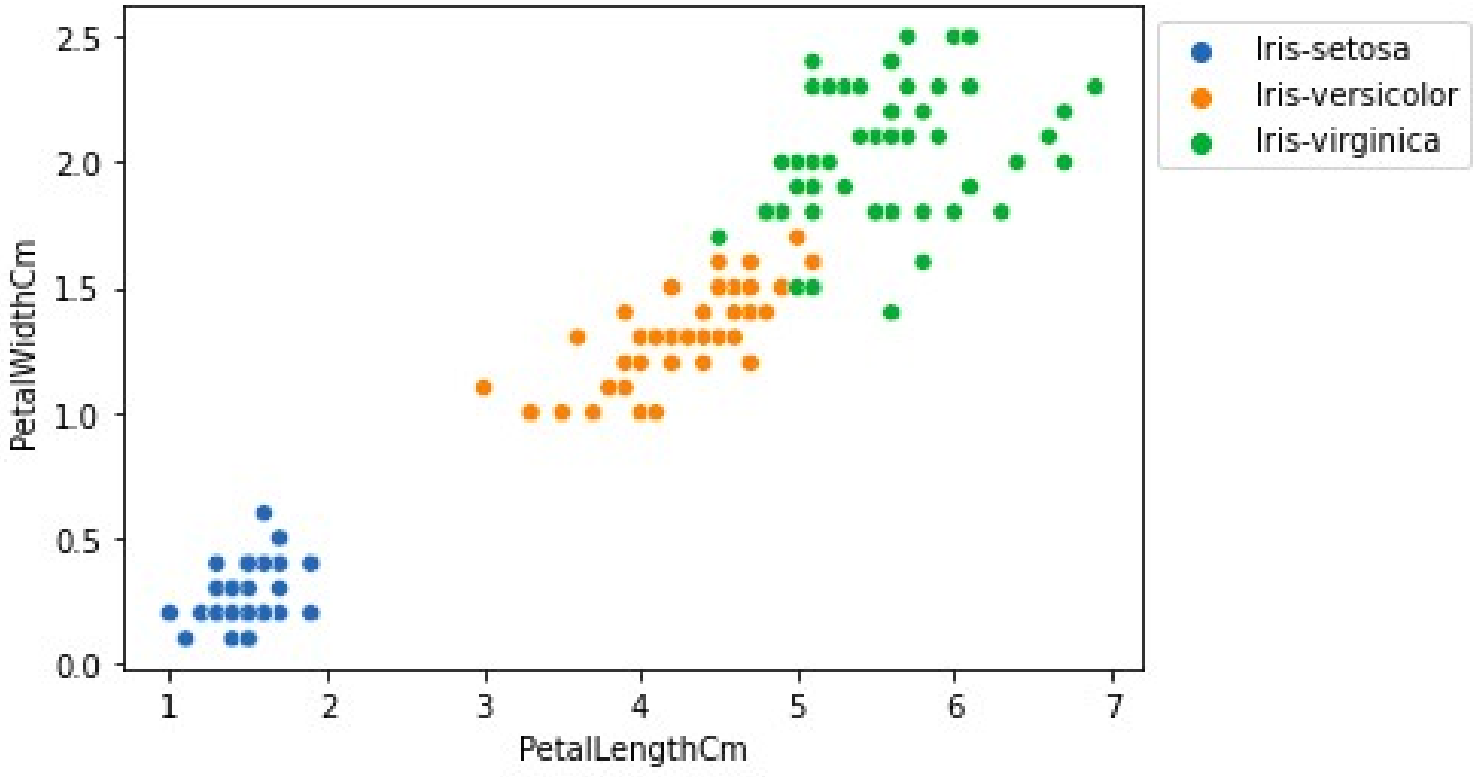
```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
sns.scatterplot(x='SepalLengthCm',
y='SepalWidthCm', hue='Species',
data=df, )
# Placing Legend outside the Figure
plt.legend(bbox_to_anchor=(1, 1),
loc=2) plt.show()
```



Data Visualization

- Multivariate Visualization

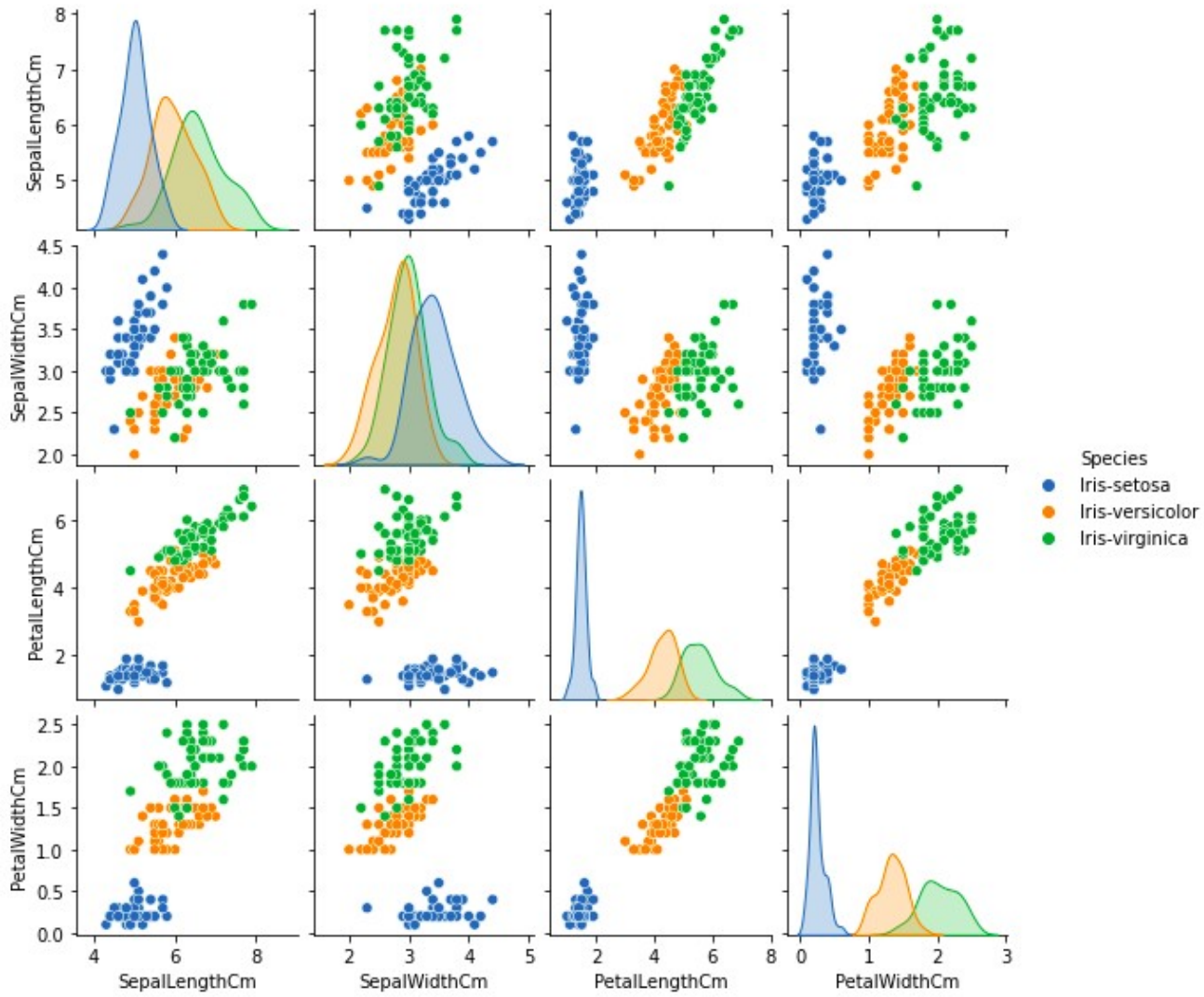
```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
sns.scatterplot(x='PetalLengthCm',
y='PetalWidthCm', hue='Species',
data=df, )
# Placing Legend outside the Figure
plt.legend(bbox_to_anchor=(1, 1),
loc=2) plt.show()
```



Data Visualization

- Multivariate Visualization

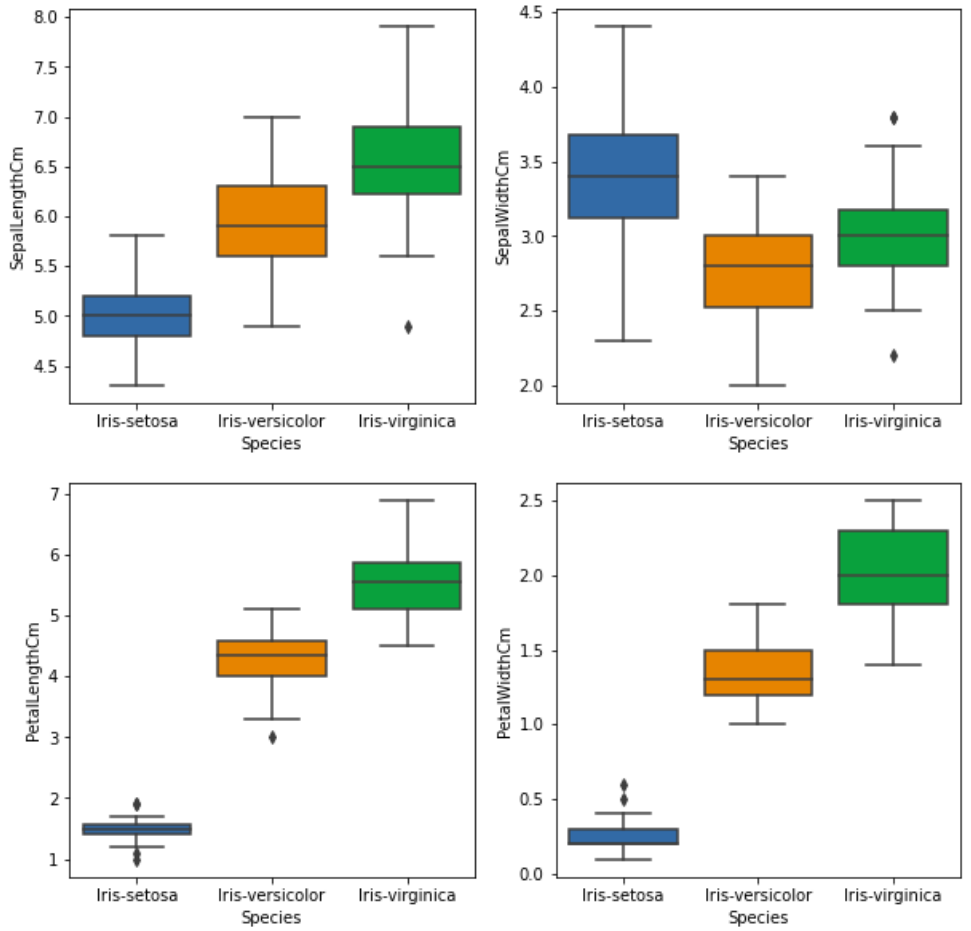
```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(df.drop(['Id'], axis = 1),
hue='Species', height=2)
```



Data Visualization

- Multivariate Visualization

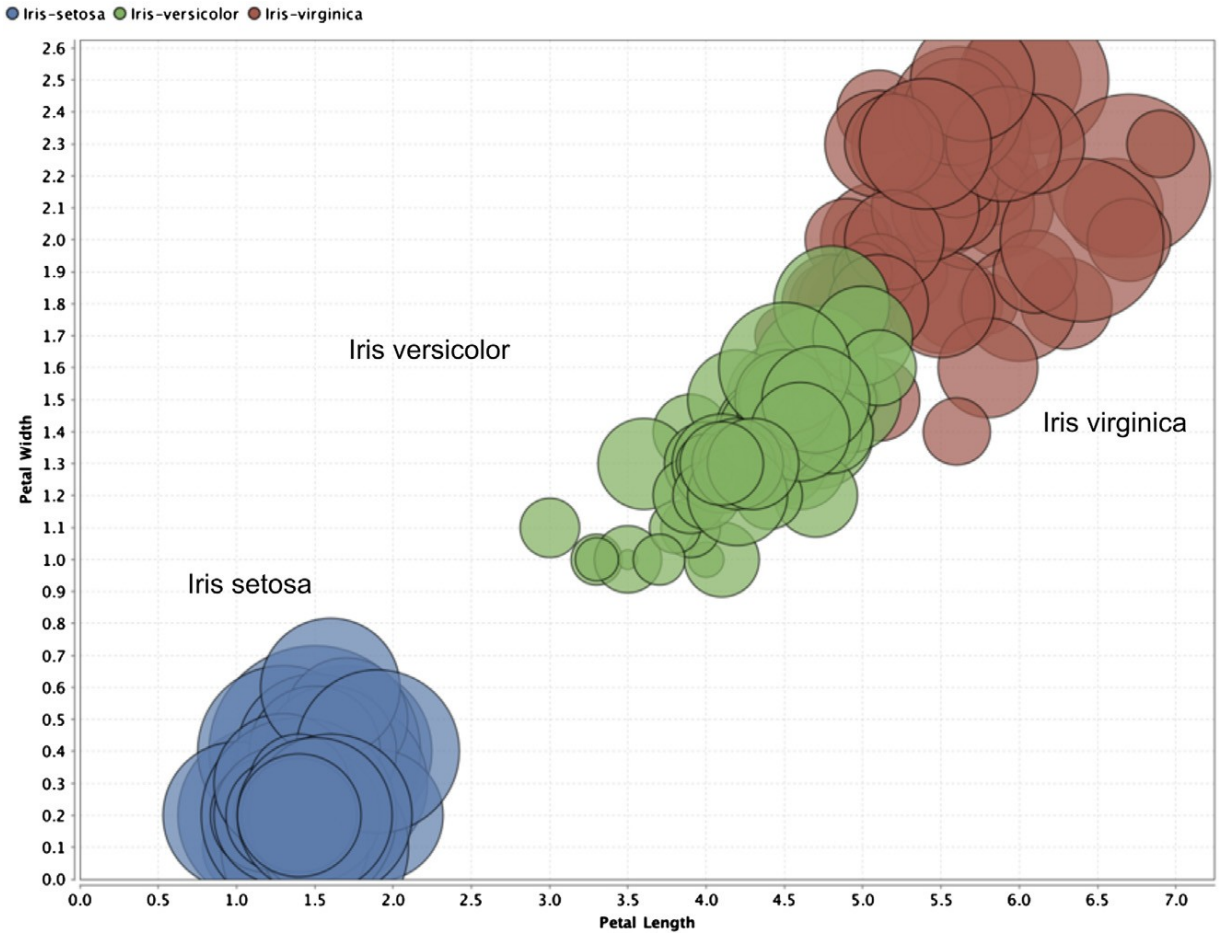
```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
def graph(y):
        sns.boxplot(x="Species", y=y,
data=df) plt.figure(figsize=(10,10))
# Adding the subplot at the specified
# grid position
plt.subplot(221)
graph('SepalLengthCm’)
plt.subplot(222)
graph('SepalWidthCm’)
plt.subplot(223)
graph('PetalLengthCm’)
plt.subplot(224)
graph('PetalWidthCm’)
plt.show()
```



Data Visualization

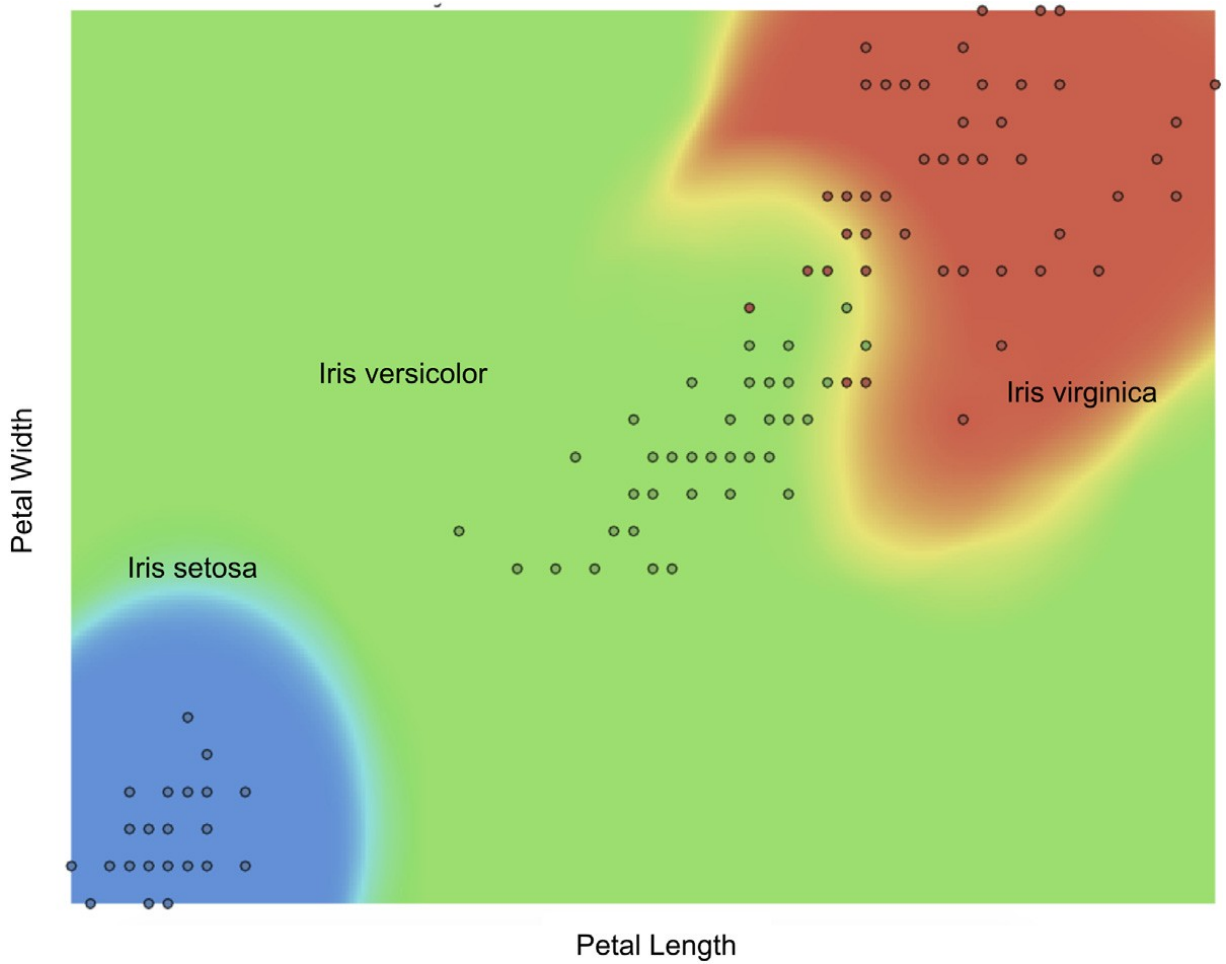
- Multivariate Visualization

```
# use the scatterplot function to
build the bubble map
sns.scatterplot(data=data,
x="PetalLengthCm", y="PetalWidthCm",
size="SepalWidthCm", legend=False,
sizes=(20, 2000))
# show the graph
plt.show()
```



Data Visualization

- Multivariate Visualization



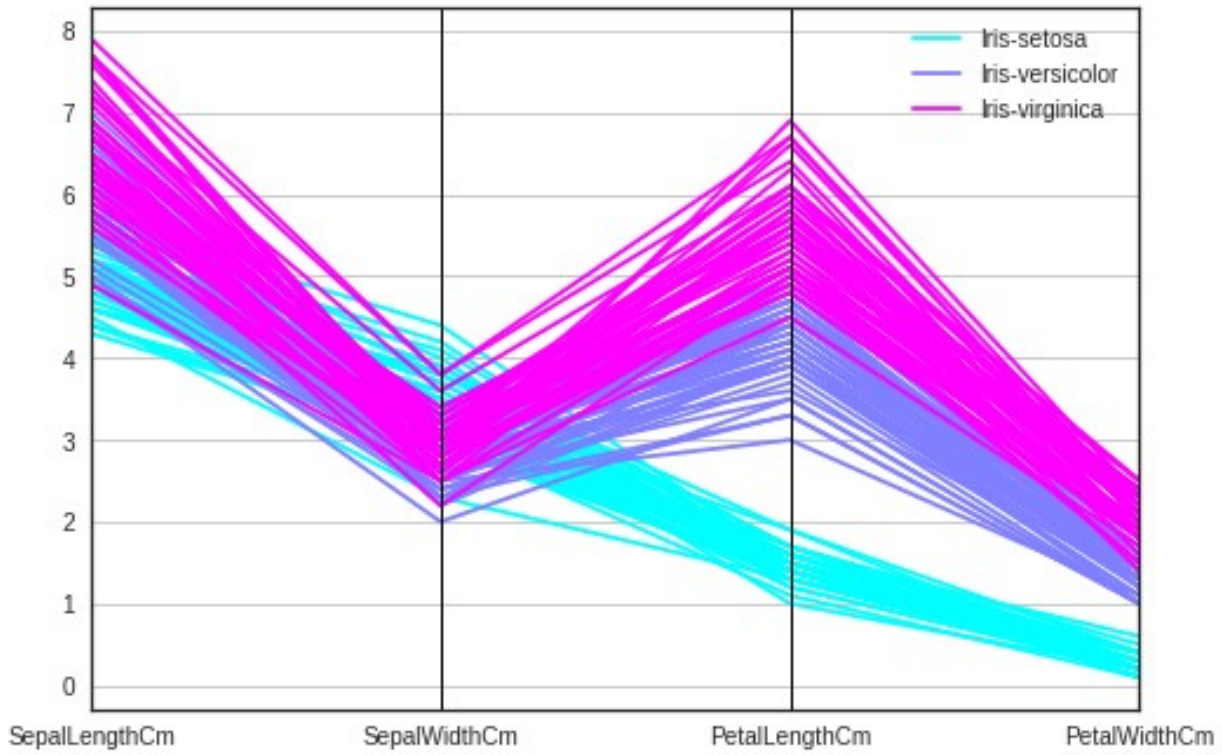
Data Visualization

- Visualizing High-Dimensional Data
  - Parallel Chart: a parallel chart visualizes a data point quite innovatively by transforming or projecting multi-dimensional data into a two-dimensional chart medium.
  - Deviation Chart: a deviation chart is very similar to a parallel chart, as it has parallel axes for all the attributes on the x-axis.
  - Andrews Curves: an Andrews plot belongs to a family of visualization techniques where the high-dimensional data are projected into a vector space so that each data point takes the form of a line or curve.

Data Visualization

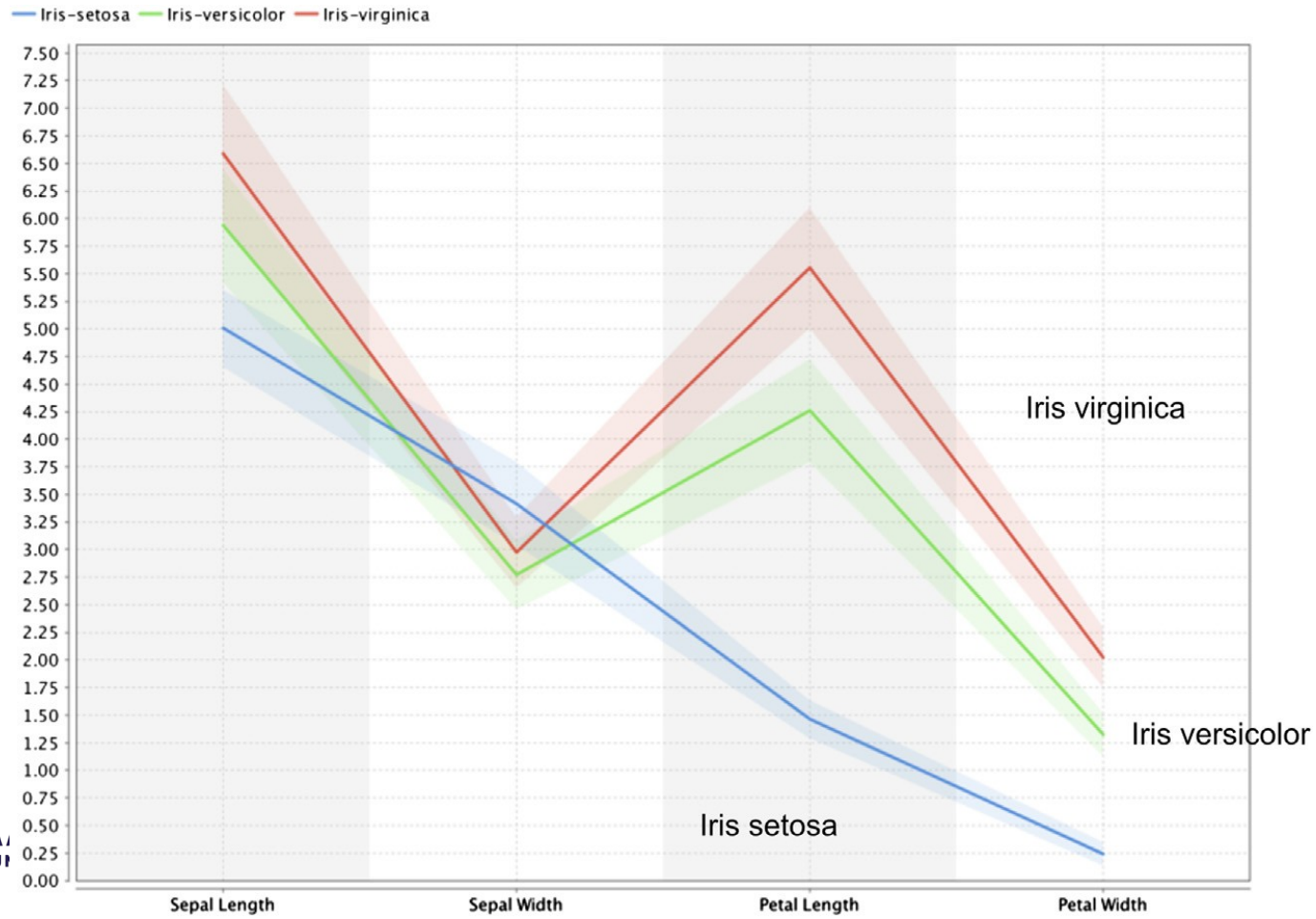
- Visualizing High-Dimensional Data

```
#Parallel_coordinates plot each feature on a seperate column.
#Each feature is then connected by lines, for each data sample
#Again, colormap can be used to choose an assortment of colors.
from pandas.tools.plotting import parallel_coordinates
parallel_coordinates(iris.drop("Id", axis=1), "Species",colormap='cool')
plt.show()
```



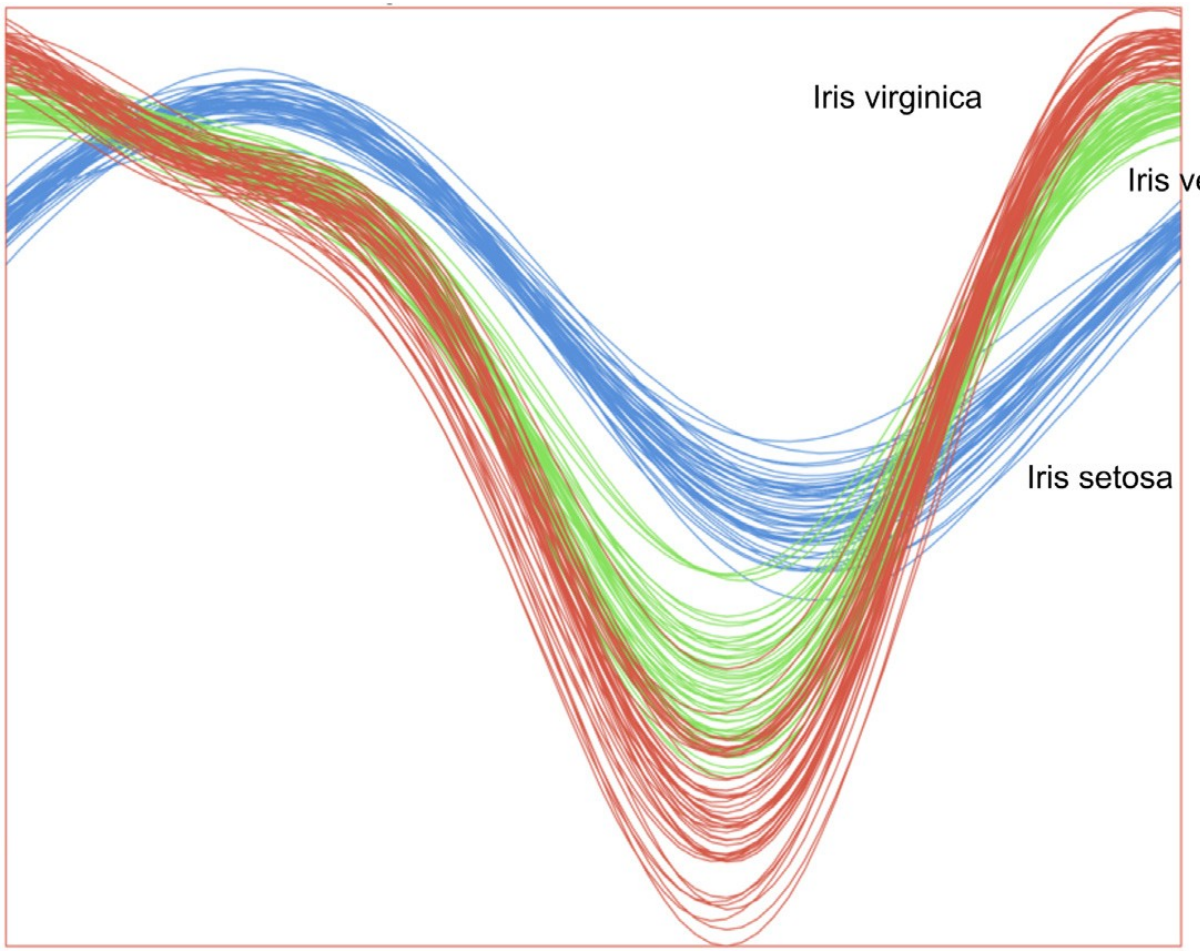
Data Visualization

- Visualizing High-Dimensional Data



Data Visualization

- Visualizing High-Dimensional Data



An *Andrews plot* belongs to a family of visualization techniques where the high-dimensional data are projected into a vector space so that each data point takes the form of a line or curve. In an Andrews plot, each data point X with d dimensions, $X = (x_1, x_2, x_3, \ldots, x_d)$, takes the form of a Fourier series:

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) + \cdots \quad (3.4)$$

This function is plotted for $-\pi < t < \pi$ for each data point. Andrews plots are useful to determine if there are any outliers in the data and to identify potential patterns within the data points (Fig. 3.17). If two data points are similar, then the curves for the data points are closer to each other. If curves are far apart and belong to different classes, then this information can be used classify the data (Garcia-Osorio & Fyfe, 2005).

Data Visualization

- Visualizing High-Dimensional Data

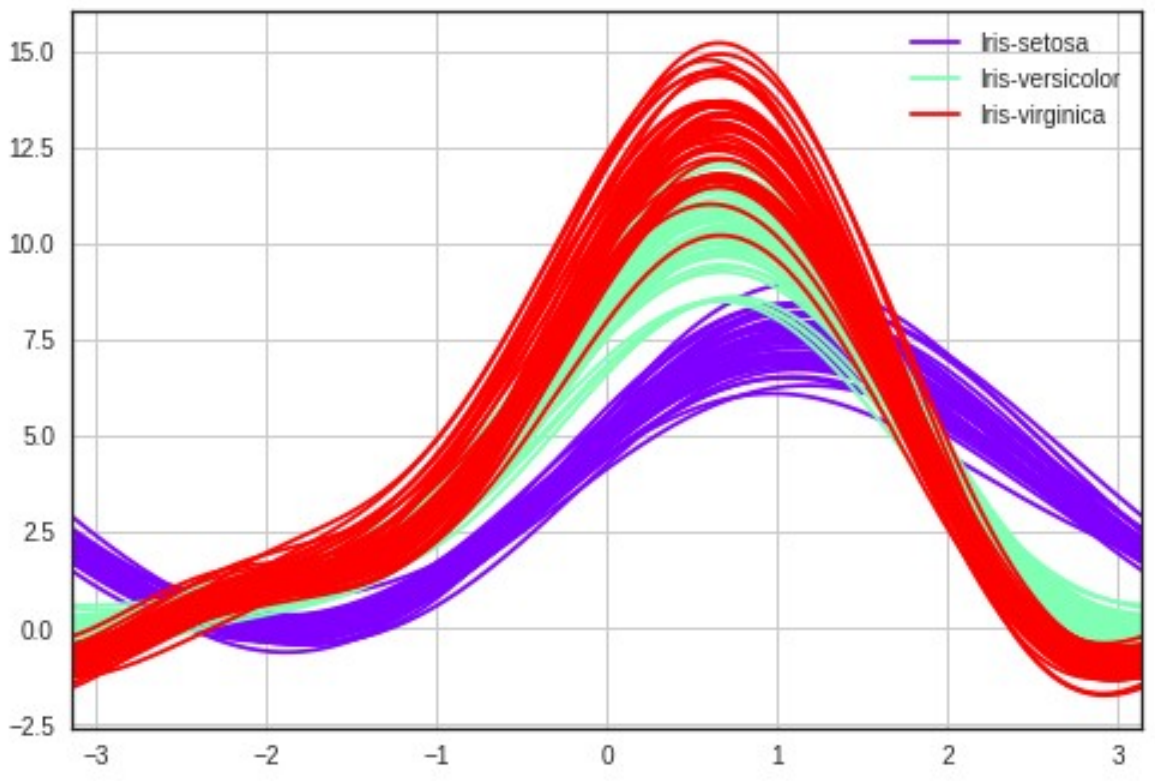
#In Pandas use Andrews Curves to plot and visualize data structure.

#Each multivariate observation is transformed into a curve and represents the coefficients of a Fourier series.

#This useful for detecting outliers in times series data.

#Use colormap to change the color of the curves

```
from pandas.tools.plotting import andrews_curves
andrews_curves(iris.drop("Id", axis=1), "Species",colormap='rainbow')
plt.show()
```



Data Exploration Process

- Organize the dataset.
- Find the central point for each attribute.
- Understand the spread of each attribute.
- Visualize the distribution of each attribute.
- Pivot the data.
- Watch out for outliers.
- Understand the relationship between attributes.
- Visualize the relationship between attributes
- Visualize high-dimensional datasets

Summary

- We understand the target of data exploration
- We understand descriptive analytics
- We understand data visualization

Exercises 1

- Download IRIS and conduct descriptive analytics
  - Report the descriptive statistic metrics for the whole datasets
  - Report the descriptive statistic metrics for each class

Exercises 2

- Download IRIS and conduct visualization
  - Try to reproduce different types of visualization in this lecture