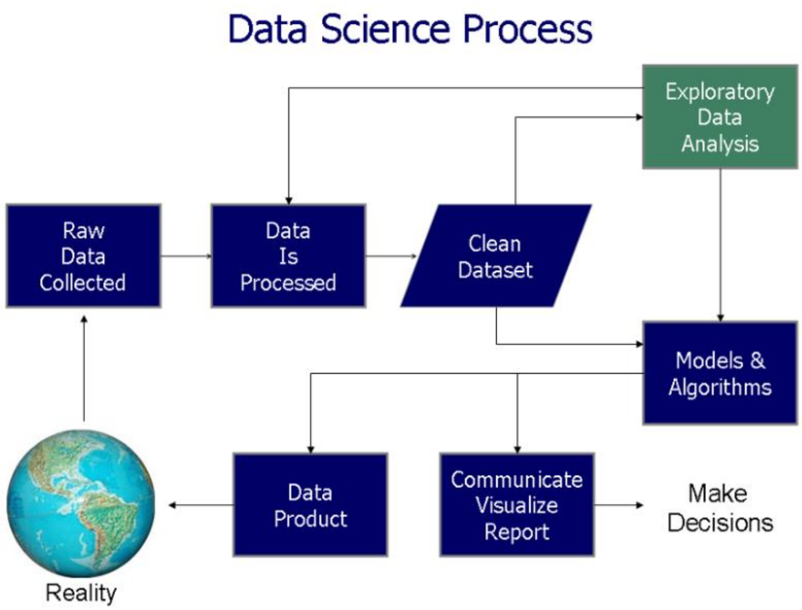
EXPLORATORY DATA ANALYSIS (EDA)



Start Understanding Dataset :

- What types of variables are there in the dataset?
- What do their distributions look like?
- Do we still have missing values?
- Are there redundant variables?
- What are the relationships between the features?
- Do we observe outliers?
- How do the different pairs of features correlate with each other?
- Do these correlations make sense?
- What is the relationship between the features and the target?

By definition, **exploratory data analysis is an approach to analysing data to summarise their main characteristics, often with visual methods**.

In other words, we perform analysis on data that we **collected,** to find **important metrics/features** by using some nice and pretty **visualisations.**

every person takes some decisions in their life considering a few points in some situations. to be accurate at these decisions data scientist does some EDA on data.

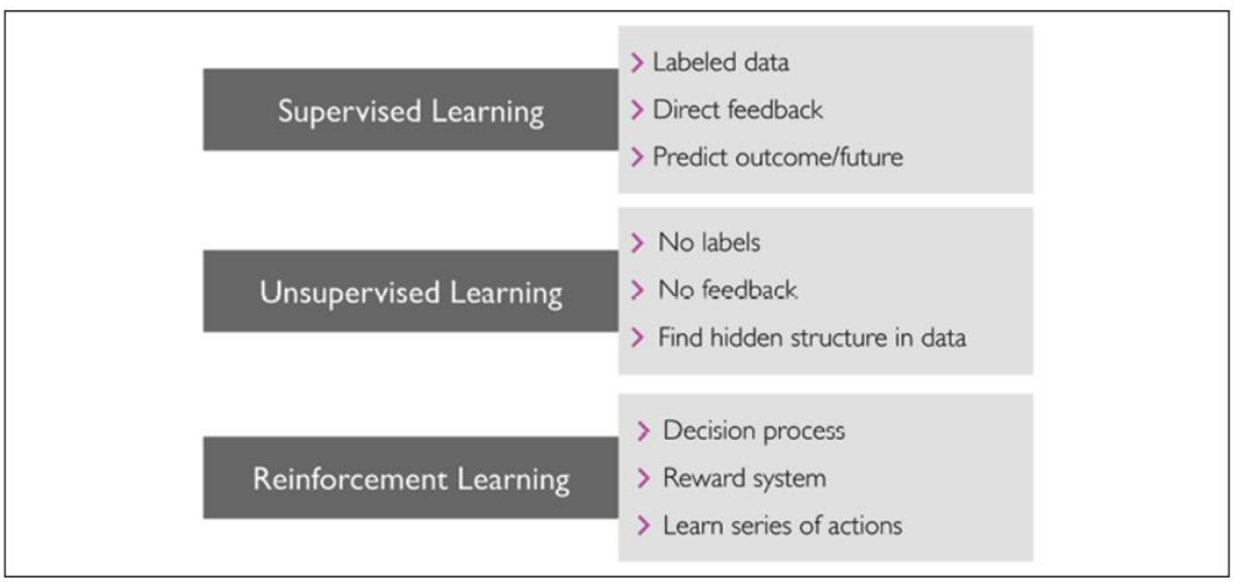
For Ex: if you want to join a graduate school, what do you do?

you collect some opinions(*data*) from alumni, students, friends, family. now from those opinions, you will find some key points(metrics/features), let's say the points like placement rate, reputation, faculty to student ratio, labs and infrastructure. if you are happy with these points, only then you will join them.

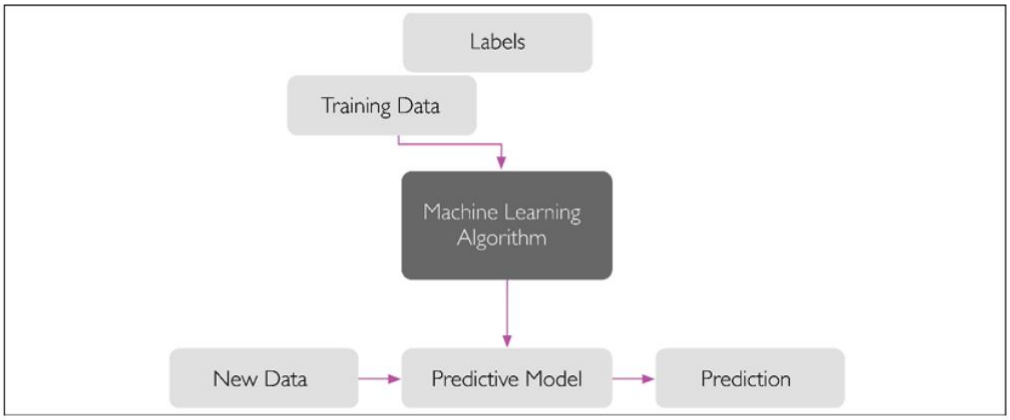
Exploratory Data Analysis is majorly performed using the following methods:

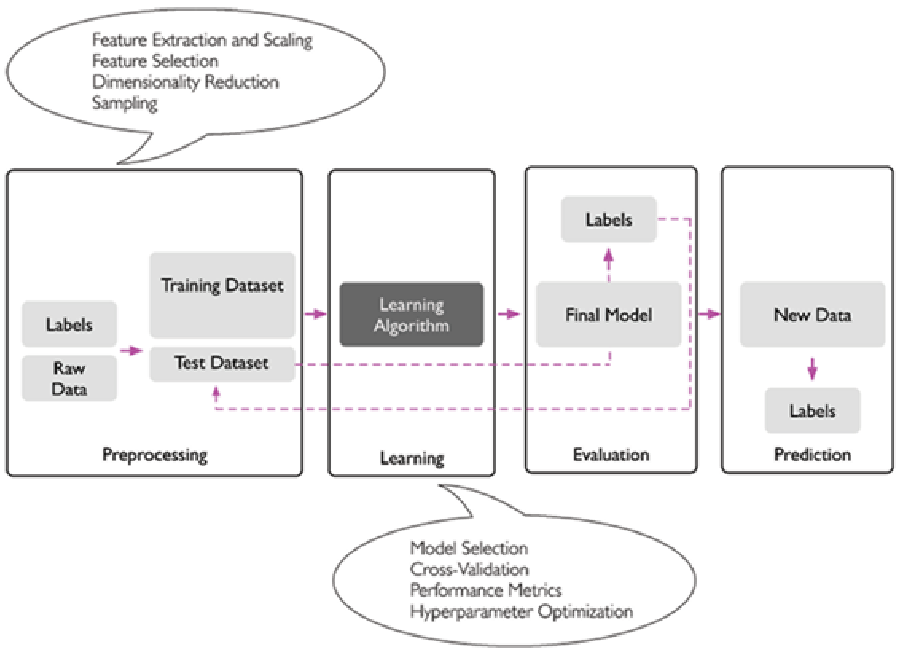
- **Univariate analysis:-** provides summary statistics for each field in the raw data set (or) summary only on one variable. *Ex*:- CDF,PDF,Box plot, Violin plot.(don't worry, will see below what each of them is)
- **Bivariate analysis:-** is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them.*Ex*:-Box plot, Violin plot.
- **Multivariate analysis:-** is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2. *Ex*:- Pair plot and 3D scatter plot.

The three different types of machine learning

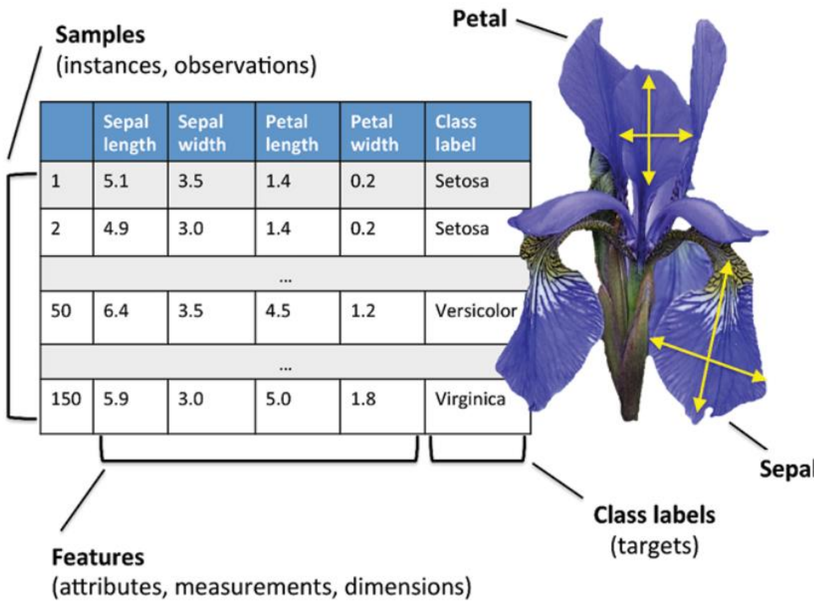


Making predictions about the future with supervised learning:


Labels
Training Data
Machine Learning Algorithm
New Data
Predictive Model
Prediction


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Preprocessing
Learning Algorithm
Learning
Labels
Final Model
Evaluation
New Data
Labels
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Example we use Iris dataset



The Iris dataset, consisting of 150 examples and four features, can then be written as a 150×4 matrix, $\boldsymbol{X} \in \mathbb{R}^{150 \times 4}$:

$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

Notational conventions

For the rest of this book, unless noted otherwise, we will use the superscript *i* to refer to the *i*th training example, and the subscript *j* to refer to the *j*th dimension of the training dataset.

We will use lowercase, bold-face letters to refer to vectors ($\boldsymbol{x} \in \mathbb{R}^{n \times 1}$) and uppercase, bold-face letters to refer to matrices ($\boldsymbol{X} \in \mathbb{R}^{n \times m}$). To refer to single elements in a vector or matrix, we will write the letters in italics ($x^{(n)}$ or $x_m^{(n)}$, respectively).

For example, $x_1^{(150)}$ refers to the first dimension of flower example 150, the *sepal length*. Thus, each row in this feature matrix represents one flower instance and can be written as a four-dimensional row vector, $\boldsymbol{x}^{(i)} \in \mathbb{R}^{1 \times 4}$:

$$\boldsymbol{x}^{(i)} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & x_4^{(i)} \end{bmatrix}$$

And each feature dimension is a 150-dimensional column vector, $x^{(i)} \in \mathbb{R}^{150\times 1}$. For example:

$$\boldsymbol{x}_j = \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \dots \\ x_j^{(150)} \end{bmatrix}$$

Similarly, we will store the target variables (here, class labels) as a 150-dimensional column vector:

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(150)} \end{bmatrix} (y \in \{\text{Setosa, Versicolor, Virginica}\})$$

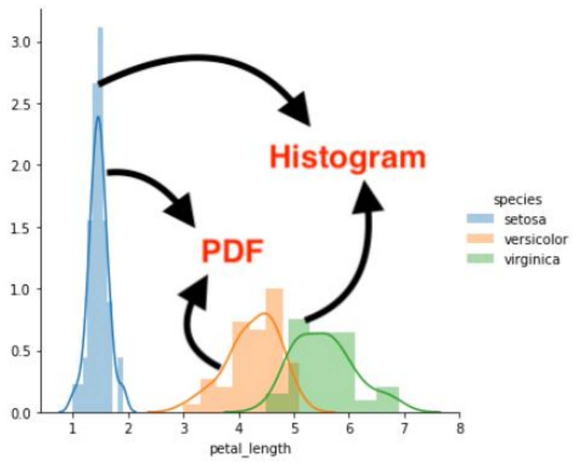
Introduction EDA

1. 1D (Univariate), 2D and 3D scatter plot
2. Pair plots
3. Histogram
4. Introduction of PDF(Probability Density Function)
5. Introduction of CDF (Cumulative Distribution Function)
6. Mean, Variance and Standard Deviation
7. Median and Quantiles
8. Box-plot and whisker

Histogram and Introduction of PDF

A histogram is an accurate graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable).To construct a histogram, the first step is to “bin” the range of values — that is, divide the entire range of values into a series of intervals — and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable.

Here in the figure, x-axis is the petal length and the y axis is a count of no of points that exist in the given range. And using this plot we can able to observe how many points are there in particular regions.Histogram basically represents how many points exist for each value on the x-axis. PDF is smoothness of histogram



Mean, Variance and Standard Deviation

$$\text{Mean} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

$$\text{Variance} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

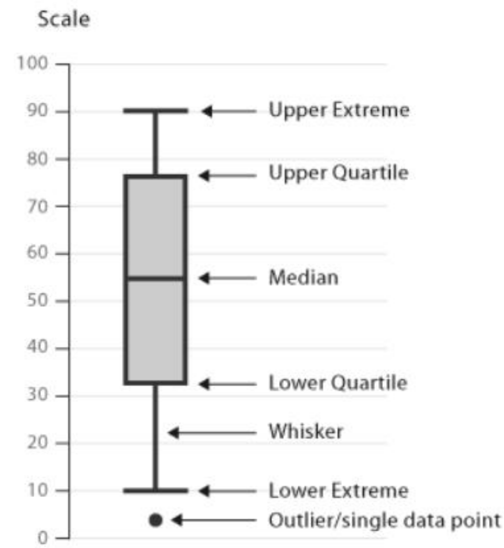
Mean is average of a given set of data. Let us consider below example.

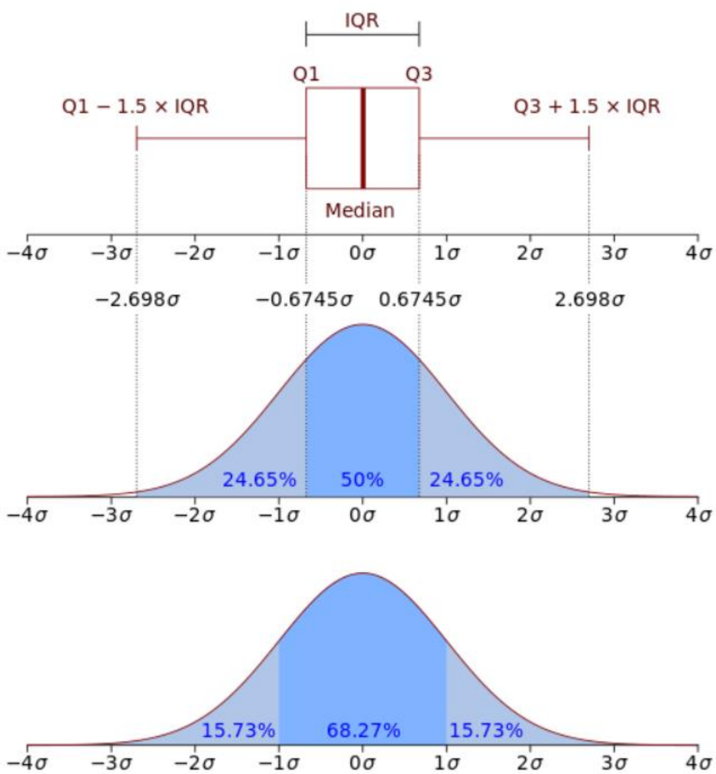
Variance is the sum of squares of differences between all numbers and means.Deviation for above example. First, calculate the deviations of each data point from the mean, and square.

Standard Deviation is square root of variance. It is a measure of the extent to which data varies from the mean.

Box-plot and whisker

A box and whisker plot (sometimes called a boxplot) is a graph that presents information from a five-number summary. It does not show a distribution in as much detail as a stem and leaf plot or histogram does, but is especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set. Box-plot with whiskers: another method of visualising the 1-D scatter plot more intuitive





Untuk lebih memahami pengolahan data untuk univariate, kerjakan Langkah-langkah berikut :

PERSIAPAN DATA

1. Ketik program sebagai berikut

```
import pandas as pd
from google.colab import files
uploaded = files.upload()
```

Keluar box choose file , kemudian klik ambil file yang akan diupload

Choose Files IRIS.csv

- **IRIS.csv**(application/vnd.ms-excel) - 3861 bytes, last modified: 2/11/2021 - 100% done

```
Saving IRIS.csv to IRIS.csv
```

2. Cek apakah file sudah terupload atau belum dengan menulis program dan running :

```
uploaded
```

Setelah itu muncul keterangan sebagai beriku :

```
{'IRIS.csv': b'\xef\xbb\xbfsepal_length,sepal_width,petal_length,petal_width,species\n5
```

3. Ketik program

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

4. Ketik data info untuk mengetahui parameter data

```
data = pd.read_csv('IRIS.csv')
data.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
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

5. Ketik perintah .info() untuk mengetahui

```
data.info()
```

Sehingga diperoleh info sebagai berikut :

```
<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   species       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

Ada 4 parameter data numerik yang masing-masing parameter berjumlah 150 data dengan 1 parameter data kategori yang merupakan label / target dengan jumlah data 150.

Data tidak mempunyai parameter yang kosong (NaN)

6. Ketik program

```
print('Ukuran data : ', data.shape)
print(pd.value_counts(data.species))
```

Setelah program dirun maka :

```
Ukuran data :  (150, 5)
setosa        50
virginica     50
versicolor    50
Name: species, dtype: int64
```

Pada IRIS.csv merupakan data data balanced dengan jumlah tiap label yang sama yaitu Sentosa 50 data, virginica 50 data dan versicolor 50 data.

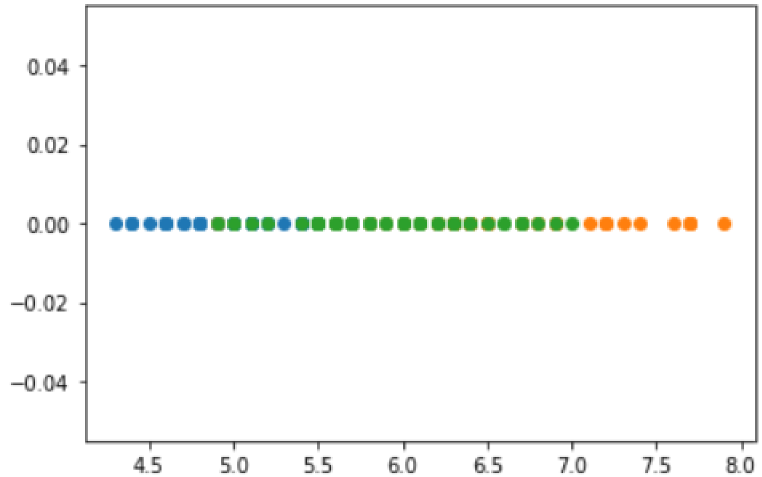
UNIVARIATE

GRAFIK 1D

7. Ketik program untuk mengambil salah satu variable dari salah satu species yaitu species setosa ➔ **sepal_length**. Plot parameter sepal_length dengan masing-masing species dengan 1D.

```
df_sentosa = data.loc[data['species'] == 'setosa']
df_virginica = data.loc[data['species'] == 'virginica']
df_versicolor = data.loc[data['species'] == 'versicolor']
```

```
plt.plot(df_sentosa['sepal_length'], np.zeros_like(df_sentosa['sepal_length']),'o')
plt.plot(df_virginica['sepal_length'], np.zeros_like(df_virginica['sepal_length']),'o')
plt.plot(df_versicolor['sepal_length'], np.zeros_like(df_versicolor['sepal_length']),'o')
```



Beri Analisa data tersebut

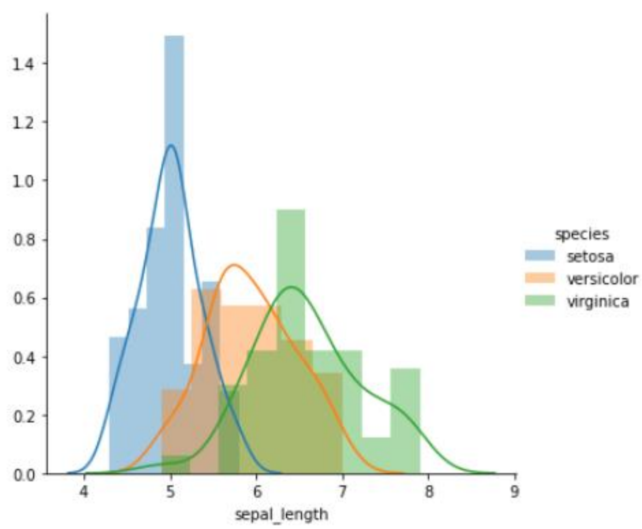
8. Ulangi Langkah 7 untuk membuat grafik 1D dengan parameter sepal_width, petal_length dan petal_width. Analisa hasil grafik tersebut.

HISTOGRAM DAN PDF

9. Plot histogram dan **pdf** dari **sepal length** untuk masing-masing **species** :

```
sns.FacetGrid(data,hue="species",size=5) \
    .map(sns.distplot,"sepal_length") \
    .add_legend();

plt.show();
```



10. Seperti Langkah ke 9 lakukan untuk sepal width, petal length dan petal width dengan masing-masing spesies. Analisa semua gambar pdf yang dihasilkan.

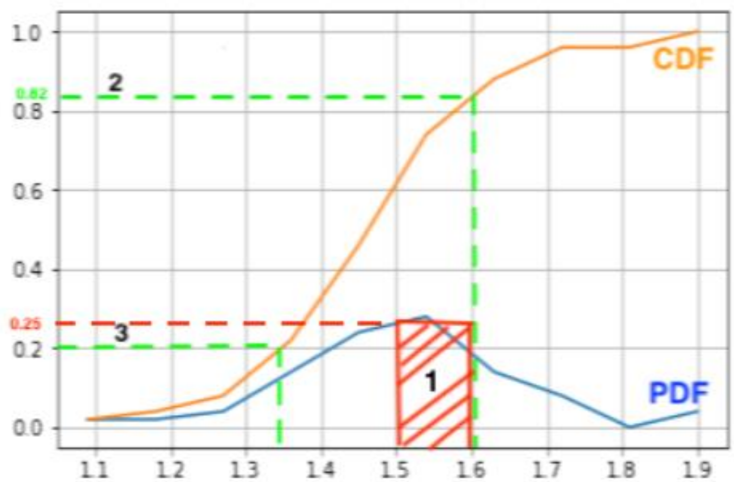
CUMULATIVE DISTRIBUTION FUNCTION (CDF)

11. Plot cdf dari **petal_length** untuk species **sentosa** :

```
iris_setosa = data.loc[data["species"] == "setosa"];
iris_virginica = data.loc[data["species"] == "virginica"];
iris_versicolor = data.loc[data["species"] == "versicolor"];
counts, bin_edges = np.histogram(iris_setosa['petal_length'], bins=10, density = True)
pdf = counts/(sum(counts))
print(pdf);

print(bin_edges);

cdf = np.cumsum(pdf)
plt.grid()
plt.plot(bin_edges[1:],pdf);
plt.plot(bin_edges[1:], cdf)
```



Analisa species Sentosa paramater dari grafik pdf dan cdf dari petal_length (PL) sumbu pada

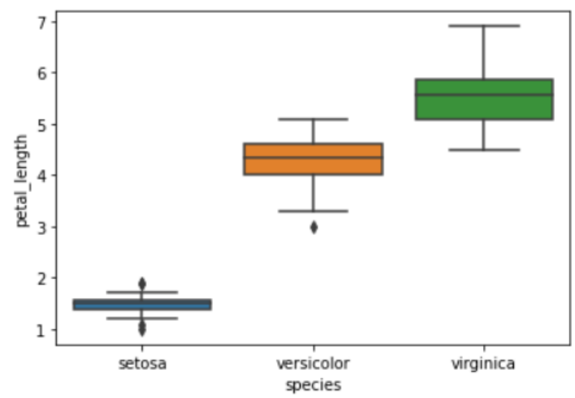
a. daerah 1 ➔ 1,5 < PL <1,6
b. titik 2 ➔ PL< 1.6 hitung berapa jumlah PL < 1.6 pada species Sentosa
c. titik 3

12. Dengan cara yang sama seperti langkah 10 analisa setiap parameter **petal_length, petal_width, sepal_length dan sepal_width** untuk masing-masing species (total 12 analisa).

BOXPLOT ➔ 1 parameter

13. Ketik program untuk boxplot petal_length masing-masing species:

```
[ ]  sns.boxplot(x="species",y="petal_length", data=data)
     plt.show()
```



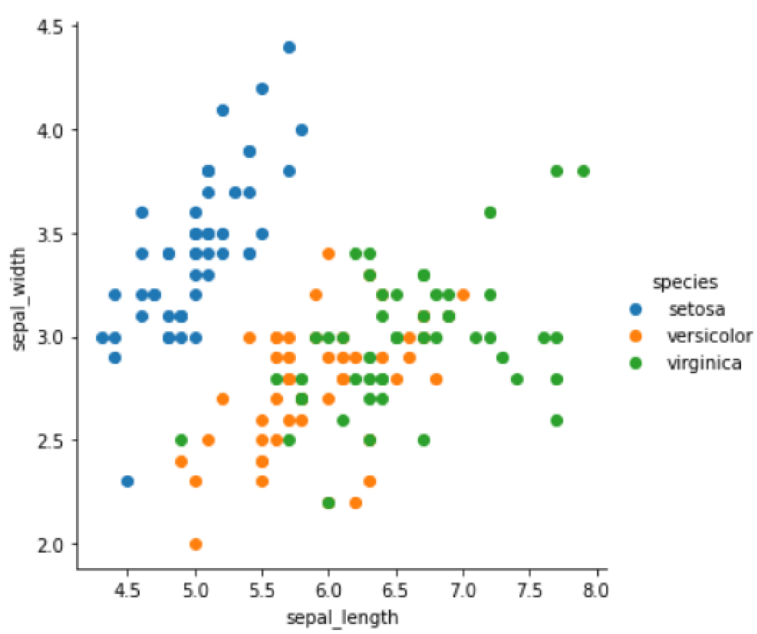
14. Ketik program untuk boxplot masing-masing petal_width, sepal_length, sepal_width. Analisa gambar yang dihasilkan.

MULTIVARIATE

GRAFIK 2 DIMENSI

15. Ketik program dibawah ini untuk grafik scatter sepal_length dan sepal_width semua species. Analisa grafik yang dihasilkan.

```
sns.FacetGrid(data, hue="species", size=5).map(plt.scatter, "sepal_length", "sepal_width").add_legend()
plt.show()
```

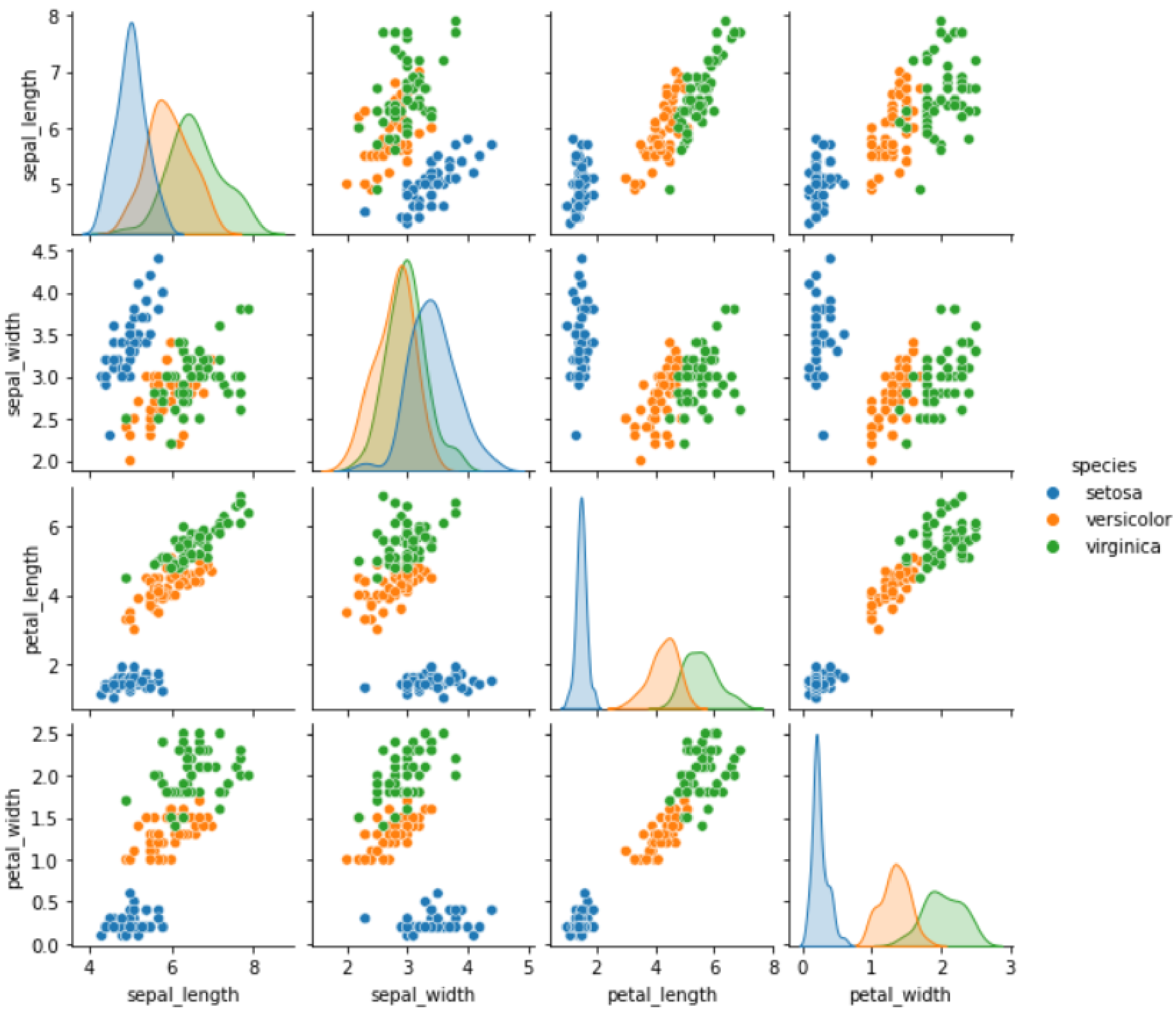


16. Ulangi Langkah 15 untuk parameter petal_length dan petal_width. Analisa hasil yang diperoleh.

PAIRPLOT

17. Ketik program pairplot (histogram atau pdf pada diagonal dan grafik scatter) untuk keseluruhan data species. Analisa masing-masing gambar.

```
sns.pairplot(data,hue="species",size=2);
plt.show()
```



18. Ketik program untuk boxplot masing-masing petal_length, petal_width, sepal_length, sepal_width dengan enghilangkan parameter label species:

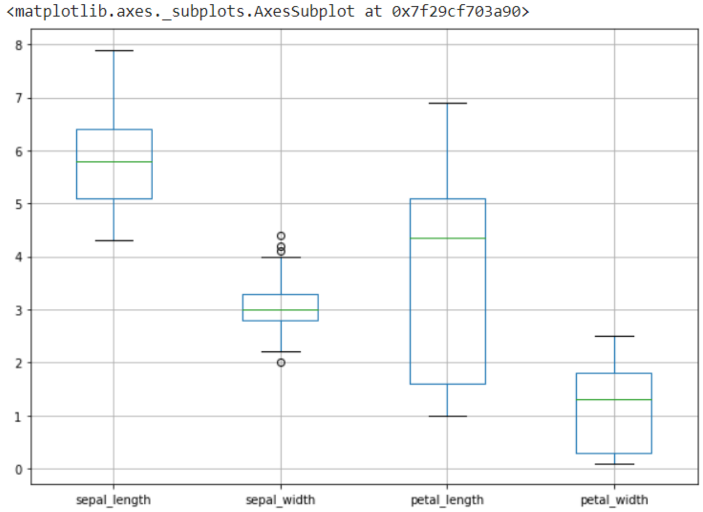
```
[ ] # removing Id column
    new_data = data[["sepal_length", "sepal_width", "petal_length", "petal_width"]]
    print(new_data.head())
```

```
   sepal_length  sepal_width  petal_length  petal_width
0           5.1          3.5           1.4          0.2
1           4.9          3.0           1.4          0.2
2           4.7          3.2           1.3          0.2
3           4.6          3.1           1.5          0.2
4           5.0          3.6           1.4          0.2
```

19. Buat boxplot dari new_data yang mempunyai 4 parameter

```
plt.figure(figsize = (10, 7))
new_data.boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f29cf703a90>
```



20. Hitung mean, varian dari masing masing species.
21. Beri kesimpulan secara keseluruhan EDA menggunakan data iris