**Why do we need Data Pre-processing?**

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

There are following steps involved in the data pre-processing:

* **Getting the dataset**
* **Importing libraries**
* **Importing datasets**
* **Finding Missing Data**
* **Encoding Categorical Data**
* **Splitting dataset into training and test set**
* **Feature scaling**

**Exploratory data analysis**

Exploratory Data Analysis (EDA) is an approach that is used to analyse the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations. This allows you to get a better feel of your data and find useful patterns in it.

**Types of EDA:**

Depending on the number of columns we are analysing, we can divide EDA into two types.

1. **Univariate Analysis –** In univariate analysis, we analyse or deal with only one variable at a time. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
2. **Bi-Variate analysis –** This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables.
3. **Multivariate Analysis** – When the data involves three or more variables, it is categorized under multivariate.

Depending on the type of analysis we can also subcategorize EDA into two parts.

* **Non-graphical Analysis** – In non-graphical analysis, we analyse data using statistical tools like mean median or mode or skewness
* **Graphical Analysis** – In graphical analysis, we use visualizations charts to visualize trends and patterns in the data.

**Steps Involved in Exploratory Data Analysis:**

**Step 1: Import Python Libraries**

Import all libraries which are required for our analysis, such as Data Loading, Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.

* Pandas and Numpy have been used for Data Manipulation and numerical Calculations
* **Matplotlib and Seaborn** have been used for Data visualizations.

**Step 2: Reading Dataset**

The Pandas library offers a wide range of possibilities for loading data into the pandas DataFrame from files like JSON, .csv, .xlsx, .sql, .pickle, .html, .txt, images etc.

**Analysing the Data**

* The main goal of data understanding is to gain general insights about the data we can use **shape ( ) method** to get the rows and columns in the dataset.
* We can use **head ( )** method that will display only top 5 rows from the top position, while **tail ( )** method can be used for getting rows from bottom.
* **info()** helps to understand the data type and information about data, including the number of records in each column, data having null or not null, Data type, the memory usage of the dataset.
* **Check for Duplication:** we can use **nunique ()** for checking the duplicacy. Duplicated data can be handled or removed based on further analysis.
* **Missing Values Calculation: isnull**() is widely been in all pre-processing steps to identify null values in the data. **data.isnull().sum()** is used to get the number of missing records in each column

**Step 3: Data Reduction**

Some columns or variables can be dropped if they do not add value to our analysis. We can use **drop ( )** method.

**Step 4: Data Cleaning/Wrangling**

Some names of the variables are not relevant and not easy to understand. Some data may have data entry errors, and some variables may need data type conversion**.**

**Step 5: Statistics Summary**

The information gives a quick and simple description of the data. We can include Count, Mean, Standard Deviation, median, mode, minimum value, maximum value, range, standard deviation, etc. describe() function gives all statistics summary of data

**Step 6: Data Transformation**

Import pandas as pd

Import numpy as np

# read dataset using pandas

df = pd.read\_csv('C:/Users/Lenovo/OneDrive/Desktop/detail.csv')

print(df) # print whole data set

#print(df.head(5)) # print only top rows using head()

# get the shape of the data

#df.shape

df.describe(include = 'all')

df.info() # to get the information about the dataset

df.nunique()

#Handling Missing Values

df.isnull().sum()

#fill in the missing values

df["Gender"].fillna("No Gender", inplace = True)

df.isnull().sum()

import seaborn as sns

import matplotlib.pyplot as plt

sns.histplot(x='Salary', data=df, )

plt.show()

sns.boxplot( x="Salary", y='Team', data=df, )

plt.show()

# Scatter Boxplot For Data Visualization

sns.scatterplot( x="Salary", y='Team', data=df, hue='Gender', size='Bonus %')

# Placing Legend outside the Figure

plt.legend(bbox\_to\_anchor=(1, 1), loc=2)

plt.show()

sns.pairplot(df, hue='Gender', height=2)

## 3.1 Reading files

The Pandas library offers a wide range of possibilities for loading data into the pandas DataFrame from files like JSON, .csv, .xlsx, .sql, .pickle, .html, .txt, images etc.

We can use read\_csv( ) method for reading the datset.

**Syntax:** df = pd.read\_csv ('C:/Users/Lenovo/OneDrive/Desktop/detail.csv')

## 3.2 Data cleaning in python: Handling missing values and filling NA

Data cleaning refers to the process of removing unwanted variables and values from your dataset. It may be removing missing values, outliers, and unnecessary rows/ columns. Re-indexing and reformatting our data. We can use **data.isnull ( ).sum ( )** to get the number of missing records in each column.

## 3.3 Data preparation and pre-processing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

## 3.4 Types of Validation in Python

There are three types of validation in python, they are:

**Type Check**: This validation technique in python is used to check the given input data type. For example, int, float, etc.

**Length Check**: This validation technique in python is used to check the given input string’s length.

**Range Check:** This validation technique in python is used to check if a given number falls in between the two numbers.

## 3.5 Data feature engineering: removing columns and rows from raw data

* **Python Pandas Drop Function**

Pandas **drop** is a function in Python pandas used to drop the rows or columns of the dataset. This function is often used in data cleaning.

**Note:** axis = 0 is referred as rows and axis = 1 is referred as columns.