Practical 0

What is python? (Guido van Rossum in 1980)

Python is a high-level, versatile programming language known for its simplicity and readability.

Some Feature : object-oriented, Rich Standard Library, Large Community and Ecosystem, Cross-Platform

Python data types:

Integers : x = 10

Float : y = 3.14

Complex: z = 2 + 3j

Strings: message = "Hello, World!"

Lists: numbers = [1, 2, 3, 4, 5] #mutable

Tuples: coordinates = (10, 20) #immutatble

Dictionaries: person = {'name': 'John', 'age': 30, 'city': 'New York'}

Sets: unique_numbers = {1, 2, 3, 4, 5}

Booleans: is_python_fun = True

None: result = None

Practical 1

Numpy

NumPy is a Python library for numerical computing that provides support for multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.

Features : Multidimensional ,Indexing and Slicing, Mathematical operations, Fast and efficient

Pandas

Pandas is a Python library that provides high-level data structures and data analysis tools.

It is built on top of NumPy and offers easy-to-use data manipulation and analysis capabilities, including handling of structured data, time series, and missing data.

Pandas is widely used for tasks such as data cleaning, exploration, transformation, and preparation, making it a fundamental tool for data scientists and analysts.

Handling Missing data

1.Detecting Missing Data:

df.isna() # Returns a DataFrame with True where values are missing

df.isnull() # Same as isna(), returns a DataFrame with True where values are missing

2.Dropping Missing Data:

df.dropna() # Drops rows with any missing values

3. Filling Missing Data:

df.fillna(value) # Fills missing values with a specific scalar value

df.fillna(df.mean()) # Fills missing values with the mean of each column.

EDA

EDA (Exploratory Data Analysis): A crucial step in data analysis, EDA involves examining and summarizing data to understand its underlying patterns, relationships, and distributions, aiding in hypothesis generation and feature selection for further analysis.

Visualization library

Matplotlib:

A versatile plotting library for Python, Matplotlib enables the creation of a wide range of static, animated, and interactive visualizations, including line plots, scatter plots, histograms, and more.

Seaborn:

Built on top of Matplotlib, Seaborn offers a higher-level interface for statistical data visualization, simplifying the generation of visually appealing plots with support for categorical data, color palettes, and multi-plot grids.

Practical 2.1

What is Correlation

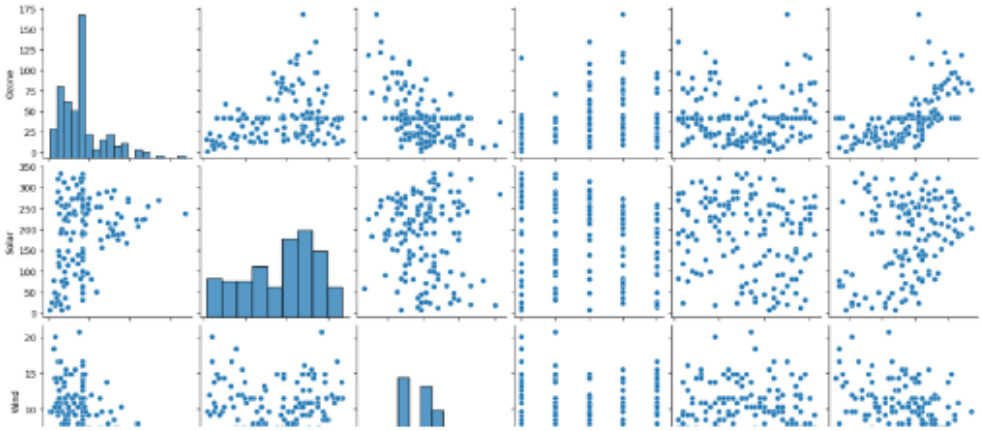
Correlation is a statistical measure that describes the strength and direction of the relationship between two variables. It indicates how much and in what way variables change together, ranging from -1 to 1, where:

1 indicates a perfect positive correlation (both variables increase together).

-1 indicates a perfect negative correlation (one variable increases as the other decreases).

0 indicates no correlation (variables are independent of each other).

```
<seaborn.axisgrid.PairGrid at 0x7dce929b6260>
```



```
#Correlation
data_cleaned5.corr()
```

```
<ipython-input-52-f7cc26521e26>:2: FutureWarning: The default value of numeric_only i
  data_cleaned5.corr()
```

	Ozone	Solar	Wind	Month	Day	Temp
Ozone	1.000000	0.307253	-0.523806	0.123962	-0.030241	0.606275
Solar	0.307253	1.000000	-0.056594	-0.092918	-0.154212	0.273322
Wind	-0.523806	-0.056594	1.000000	-0.153507	0.040151	-0.441247
Month	0.123962	-0.092918	-0.153507	1.000000	0.061236	0.393321
Day	-0.030241	-0.154212	0.040151	0.061236	1.000000	-0.124538
Temp	0.606275	0.273322	-0.441247	0.393321	-0.124538	1.000000

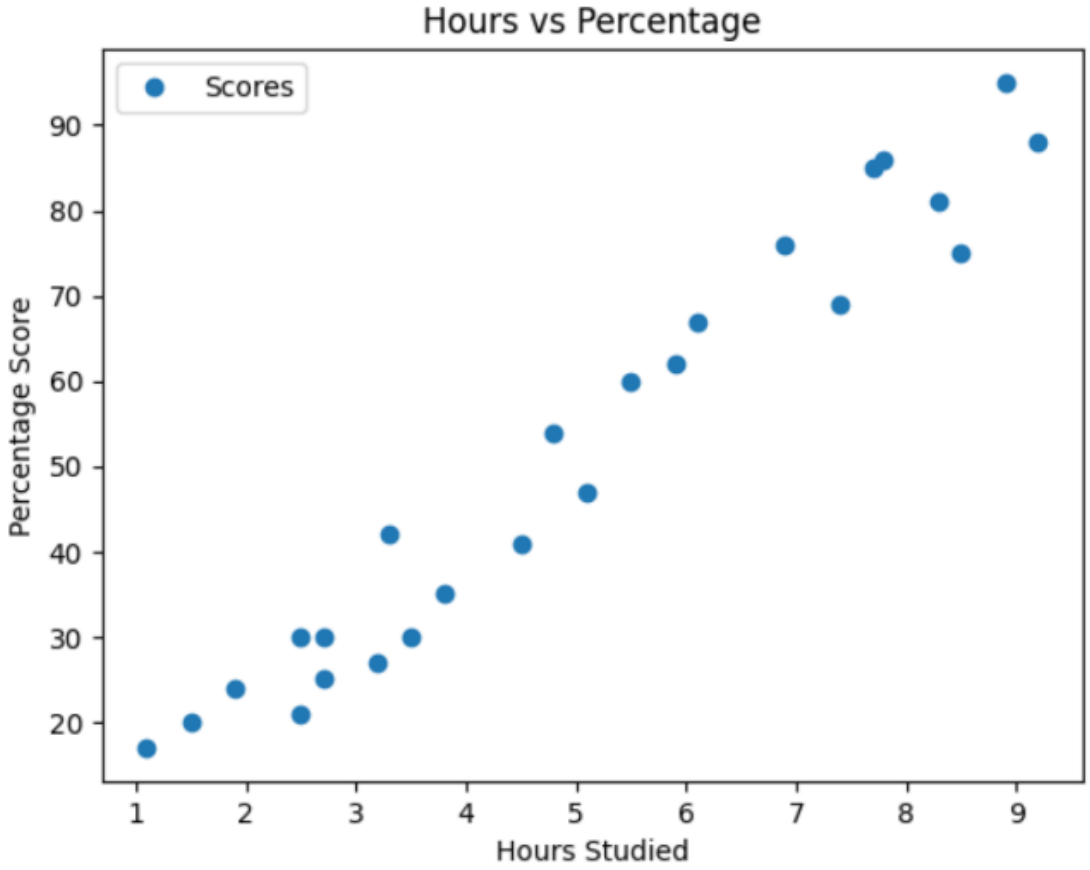
Practical 4

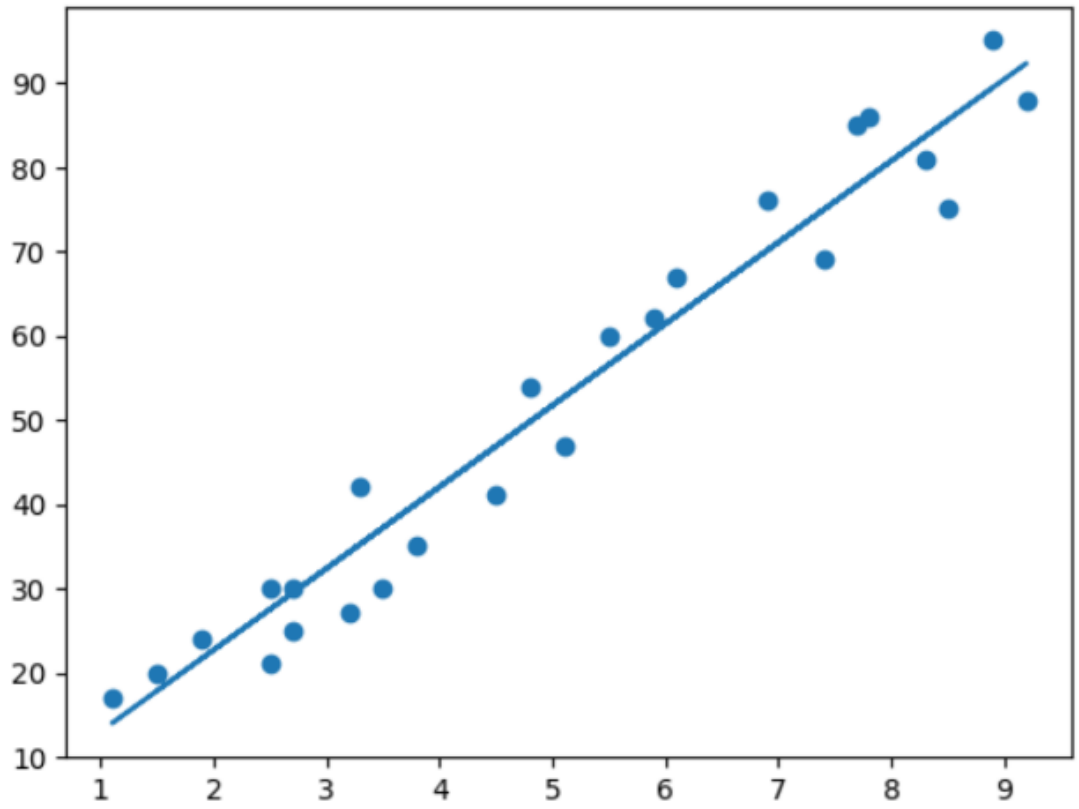
Linear regression :

A statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data, aiming to predict the value of the dependent variable based on the independent variables.

```
#finding correlation of dependent and independent variables
Shreyas_regression.corr()
```

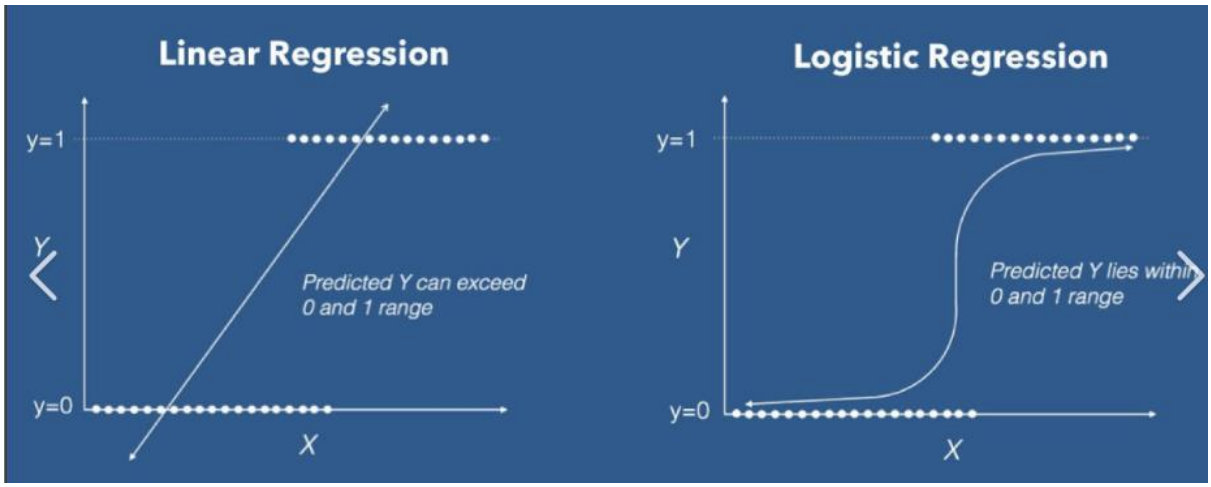
	Hours	Scores
Hours	1.000000	0.976191
Scores	0.976191	1.000000





Logistic Regression

A statistical method used for binary classification tasks, where the output variable is categorical and represents the probability of belonging to a particular class, employing the logistic function to model the relationship between the independent variables and the binary outcome.



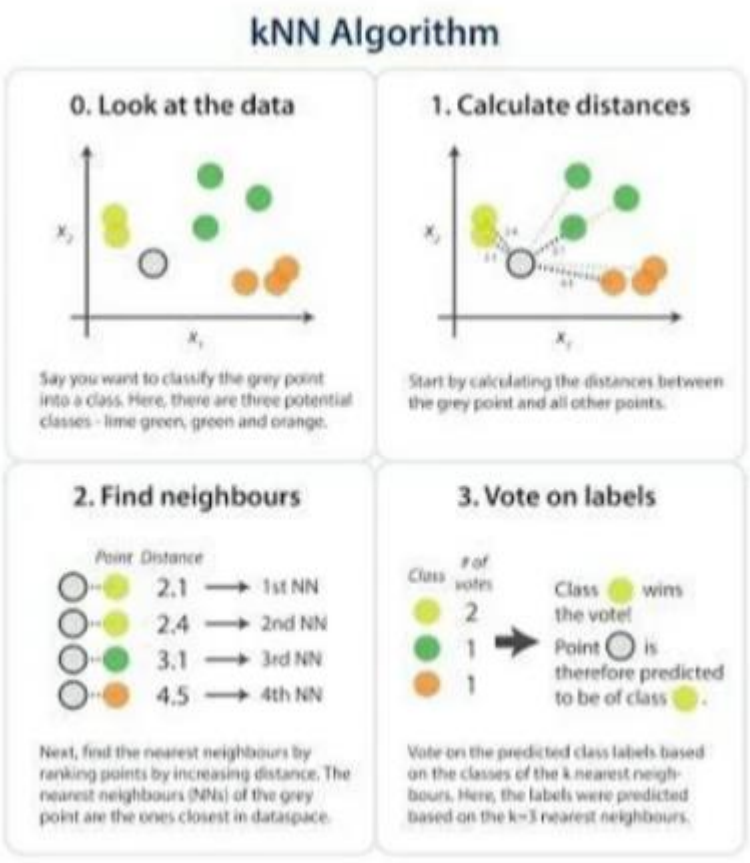
KNN

A non-parametric machine learning algorithm used for classification and regression tasks, where the prediction of a new data point is based on the majority class or the average of its K nearest neighbors in the feature space.

Key Parameters:

1. K: The number of nearest neighbors to consider. It's a hyperparameter that needs to be chosen beforehand.

2. Distance Metric: The metric used to compute distances between data points (e.g., Euclidean distance, Manhattan distance, etc.).

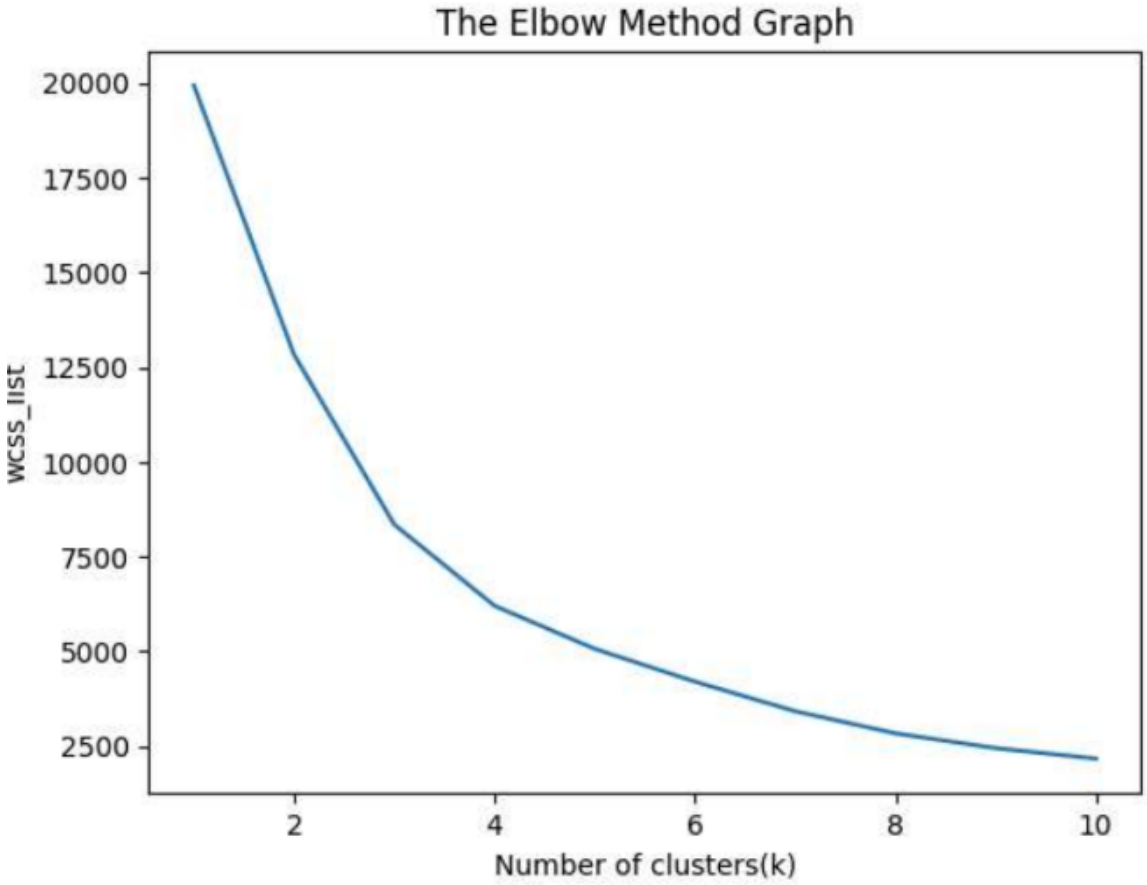


K-Means

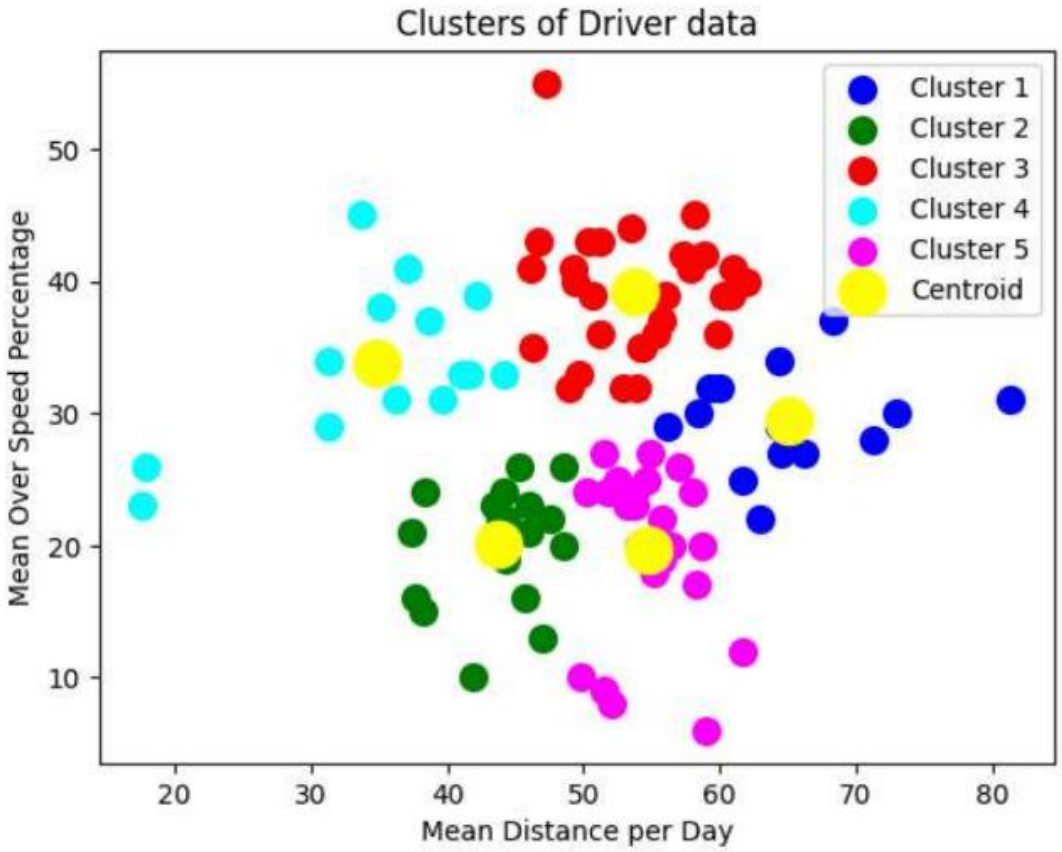
A clustering algorithm in machine learning that partitions a dataset into K distinct, non-overlapping clusters by iteratively assigning each data point to the nearest centroid and then recalculating the centroids based on the mean of the data points assigned to each cluster.

How to choose value of K

The elbow method is a technique used to determine the optimal number of clusters (K) in a K-means clustering algorithm. Here's how to interpret it:



- Plot number of clusters (K) against sum of squared distances (inertia).
- Initially, as K increases, inertia decreases.
- Diminishing returns occur as adding more clusters doesn't significantly decrease inertia.
- The "elbow" point on the plot indicates optimal K.
- Optimal K is where further clustering doesn't significantly improve performance.



K-Medoids

K-medoids is a clustering algorithm similar to K-means, but instead of using cluster centroids, it uses actual data points (medoids) as cluster representatives. It aims to partition a dataset into K clusters by iteratively updating the medoids and assigning each data point to the nearest medoid based on a dissimilarity measure such as the Manhattan or Euclidean distance. K-medoids is more robust to outliers compared to K-means because it uses actual data points as representatives of clusters.