What is Pandas library in Python?

Pandas is a Python library used for data manipulation and analysis.

List some ket features of Pandas.

Here are some key features of Pandas:

- Easy Data Handling\*\*: Works with structured data like tables, CSVs, and databases.

- DataFrames and Series\*\*: Provides powerful structures for organizing data.

- Data Cleaning\*\*: Helps handle missing values and duplicates.

- Filtering & Sorting\*\*: Allows quick manipulation of data.

- Integration\*\*: Works well with NumPy, Matplotlib, and other libraries.

3)What is Numpy Library in Python?

NumPy (Numerical Python) is a Python library used for **fast mathematical and scientific computing**.

4)What is matplotlib library?

Matplotlib is a powerful **Python library** used for **data visualization**.

5)What is difference between seaborn library amd matlpoy lib?

In short, **Matplotlib** is more customizable, while **Seaborn** is more user-friendly for statistical plots

6)Is Sk learn and Sci kit Learn are same library?what is use of it in data.

Yes. Data Preprocessing,Supervised Learning,Unsupervised Learning,Cross-Validation.

8)what is use of describe command?

The describe() command in **Pandas** is used to generate **summary statistics** for a DataFrame. It provides key insights into numerical columns, such as:

* **Count**: Number of non-null values.
* **Mean**: Average value.
* **Standard Deviation (std)**: Spread of data.
* **Min & Max**: Minimum and maximum values.
* **Percentiles (25%, 50%, 75%)**: Distribution of data.

10)Which are Navie bayes classification alogirithm used in python?

Python provides several **Naïve Bayes classification algorithms** through the **Scikit-Learn** library:

1. **Gaussian Naïve Bayes (**GaussianNB**)** – Used for **continuous data** that follows a normal distribution.
2. **Multinomial Naïve Bayes (**MultinomialNB**)** – Ideal for **text classification** and word frequency-based data.
3. **Bernoulli Naïve Bayes (**BernoulliNB**)** – Works well with **binary features** (e.g., spam detection).
4. **Complement Naïve Bayes (**ComplementNB**)** – A variation of MultinomialNB, designed for **imbalanced datasets**.

11)What is significance of Confusion matrix?

A **confusion matrix** is used to **evaluate the performance of a classification model**—it shows how well the model is making predictions.

It compares the **actual labels** with the **predicted labels** and gives four key values:

* **True Positive (TP)** – Correctly predicted positive
* **True Negative (TN)** – Correctly predicted negative
* **False Positive (FP)** – Incorrectly predicted positive
* **False Negative (FN)** – Incorrectly predicted negative

**Example (Spam detection):**

* Email is **spam** and model says **spam** → ✅ **TP**
* Email is **not spam** and model says **not spam** → ✅ **TN**
* Email is **not spam** but model says **spam** → ❌ **FP**
* Email is **spam** but model says **not spam** → ❌ **FN**

13)What is recall?

**Recall** is how well a model **finds all the correct positive cases**.

14)What is precision?

**Precision** tells you **how many of the positive predictions were actually correct**.

16)What is need of data visulization in data science?

**Data visualization** is important in data science because it helps you:

**✅ Understand data easily**

* Visuals like graphs and charts make complex

17)What is outlier?

An outlier is a data point that is very different from the rest of the data.

Ages in a group: 22, 25, 23, 24, 95

Here, 95 is an outlier — it's much higher than the others.

18)When to use histogram and pie chart?

22)What is data Wrangling?

**Data wrangling** is the process of **cleaning, organizing, and transforming raw data** into a format that is easy to use and analyze.

23)What is data transformation?

**Data transformation** means changing data from one format or structure to another to make it **easier to analyze or process**.

25)What is hadoop?

Hadoop is an open-source big data framework that helps store and process very large amounts of data across many computers.

26\)What is HDFS and Map reduce?

🔹 HDFS (Hadoop Distributed File System)

It is the storage system of Hadoop.

Stores big files by splitting them into blocks and distributing across many computers.

Keeps copies of data for safety (replication).

🔹 MapReduce

It is the processing engine of Hadoop.

Works in two steps:

Map – breaks the task into small parts and runs them in parallel.

Reduce – combines the results to get the final output.

27)What are the component of Hadoop Ecosystem?

**🔹 1. HDFS**

Stores big data by splitting it across many machines.

**🔹 2. MapReduce**

Processes data in parallel using the Map and Reduce steps.

**🔹 3. YARN (Yet Another Resource Negotiator)**

Manages and schedules resources across the cluster.

**🔹 4. Hive**

Lets you use **SQL-like queries** to read and write big data.

**🔹 5. Pig**

A scripting language to process big data (simpler than Java).

26\)What is Scala?

Scala is a high-level programming language that combines object-oriented and functional programming features.

🔹 Key Features:

Runs on the Java Virtual Machine (JVM)

Can use Java libraries

Supports functional programming (like using functions as values)

Shorter and more expressive code than Java

| **Feature** | **Scala** | **Java** |
| --- | --- | --- |
| **Code Length** | Short and concise | Longer, more boilerplate code |
| **Programming Style** | Supports **both OOP & Functional** | Mostly **Object-Oriented** |
| **Type Inference** | Yes (no need to specify types always) | No (you must declare types) |
| **Concurrency** | Easier with **Actors (Akka)** | Uses Threads, more complex |
| **Interoperability** | Can use Java code directly | Can’t use Scala directly |
| **Immutable Support** | Built-in and encouraged | Less emphasized |
| **Syntax** | Cleaner and more expressive | More verbose |

30)What is data science?

**Data Science** is the process of using **data** to **understand things, find patterns**, and **make decisions**.

**🔹 Simple Definition:**

**Data Science** is about **collecting**, **cleaning**, **analyzing**, and **visualizing** data to get useful insights.

**🔑 Key Steps:**

1. **Collect Data** – from files, websites, sensors, etc.
2. **Clean Data** – fix or remove bad or missing data
3. **Analyze Data** – find patterns using statistics or algorithms
4. **Visualize Results** – show insights using graphs or charts
5. **Make Decisions** – help businesses or researchers based on results

**🔧 Tools Used:**

* **Python**, **R**
* Libraries like **Pandas**, **NumPy**, **Matplotlib**, **Scikit-learn**
* **Machine Learning** and **AI** techniques

**🧠 Example:**

A company uses data science to:

* Predict what products you might buy
* Detect fraud in banking
* Recommend movies or songs

Want a small real-world example of data science in Python?

42)What is Tokenization in NLP?

.

**🔹 Simple Definition:**

**Tokenization** splits text into **words, phrases, or sentences** so a computer can understand and analyze them.

**✅ Example:**

**Text:**  
I love Python programming.

**After Tokenization (word-level):**  
["I", "love", "Python", "programming", "."]

**Stemming – Simple Explanation:**

**Stemming** means **cutting a word down to its root form**.

**🔹 Simple Definition:**

Stemming removes **suffixes** like *-ing*, *-ed*, *-s* to get the **base word**.

**✅ Example:**

Words:

* "playing" → "play"
* "played" → "play"
* "plays" → "play"
* **Lemmatization – Simple Explanation:**
* **Lemmatization** means turning a word into its **dictionary (base) form**, called a **lemma**.
* **🔹 Simple Definition:**
* It changes a word to its **correct root form**, using grammar rules and a dictionary.
* **✅ Example:**

| **Word** | **Lemma** |
| --- | --- |
| **running** | run |
| **better** | good |
| **was** | be |

**📚 What is Corpus in NLP? — *Simple Explanation***

A **corpus** (plural: **corpora**) is a **large collection of text** used for **language analysis** in NLP (Natural Language Processing).

**🔹 Simple Definition:**

A **corpus** is a **dataset of written or spoken language** that computers use to learn how humans talk or write.

**✅ Examples of a Corpus:**

* A list of all Wikipedia articles
* A database of news articles
* A collection of tweets or emails
* Movie reviews or product reviews