**Q: What is Pandas library in Python?**

A: Pandas is a powerful open-source Python library used for data manipulation and analysis, especially for structured data.

**Q: List some key features of Pandas.**

A: DataFrame and Series objects, Data alignment and missing data handling, Data filtering and selection, GroupBy functionality, Built-in data visualization support, Merge and join operations.

**Q: What is NumPy library in Python?**

A: NumPy (Numerical Python) is a library for numerical computations, providing support for large multidimensional arrays and matrices.

**Q: What is Matplotlib library?**

A: Matplotlib is a data visualization library in Python used to create static, animated, and interactive plots.

**Q: What is the difference between Seaborn and Matplotlib?**

A: Seaborn is built on top of Matplotlib and provides a higher-level interface for creating attractive statistical plots. Matplotlib is more customizable but requires more code.

**Q: Is Sklearn and Scikit-learn the same? What is its use in Data Science?**

A: Yes, both refer to the same library. Scikit-learn is used for machine learning tasks like classification, regression, clustering, etc.

**Q: What are functions in Pandas and NumPy library?**

A: Pandas: read_csv(), DataFrame(), groupby(), merge(), dropna(), fillna(). NumPy: array(), mean(), std(), sum(), reshape(), linspace().

**Q: What is DataFrame in Python?**

A: A DataFrame is a 2-dimensional labeled data structure with columns of potentially different types.

**Q: How to find duplicates in Python?**

A: df.duplicated() or df[df.duplicated()]

**Q: What is the use of describe command?**

A: Gives summary statistics of a DataFrame: count, mean, std, min, 25%, 50%, 75%, and max.

**Q: Which are Naive Bayes classification algorithms used in Python?**

A: GaussianNB, MultinomialNB, BernoulliNB from sklearn.naive_bayes.

**Q: What is the significance of Confusion Matrix?**

A: It evaluates the performance of classification models by showing actual vs. predicted classifications.

**Q: What is TP, TN, FP, FN in Confusion Matrix?**

A: TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

**Q: What is Recall?**

A: Recall = TP / (TP + FN)

**Q: What is Precision?**

A: Precision = TP / (TP + FP)

**Q: What is F1 Score?**

A: F1 = 2 * (Precision * Recall) / (Precision + Recall)

**Q: What is the need for Data Visualization in Data Science?**

A: To explore, understand, and communicate patterns in data easily and effectively.

**Q: What is an Outlier?**

A: An observation that is significantly different from others in the dataset.

**Q: When to use Histogram and Pie Chart?**

A: Histogram: Distribution of numerical data. Pie Chart: Proportional data (percentages of a whole).

**Q: What are the challenges in Big Data Visualization?**

A: Volume, speed, complexity, interactivity limitations, and computational power.

**Q: What is joint plot, dist plot?**

A: jointplot(): bivariate and univariate distributions. distplot(): univariate distribution (deprecated).

**Q: What are tools used for Data Visualization?**

A: Matplotlib, Seaborn, Plotly, Tableau, Power BI, ggplot, D3.js.

**Q: What is Data Wrangling?**

A: Cleaning and transforming raw data into a usable format.

**Q: What is Data Transformation?**

A: Converting data into a desired format or structure (e.g., scaling, encoding).

**Q: What is the use of StandardScaler function in Python?**

A: Standardizes features by removing the mean and scaling to unit variance.

**Q: What is Hadoop?**

A: An open-source framework for processing and storing Big Data in a distributed environment.

**Q: What is HDFS and MapReduce?**

A: HDFS: Hadoop Distributed File System. MapReduce: Programming model for distributed computation.

**Q: What are the components of Hadoop Ecosystem?**

A: HDFS, MapReduce, YARN, Hive, Pig, HBase, Spark, Flume, Sqoop.

**Q: What is Scala?**

A: A high-level programming language combining object-oriented and functional programming.

**Q: What are features of Scala?**

A: Concise syntax, JVM-based, functional and object-oriented, immutability support, interoperable with Java.

**Q: How is Scala different from Java?**

A: Scala is more concise, supports functional programming, and is more expressive with type inference.

**Q: List applications of Scala.**

A: Apache Spark, Big Data analytics, Web applications, Distributed systems.

**Q: What is Data Science?**

A: An interdisciplinary field combining statistics, computer science, and domain knowledge to extract insights from data.

**Q: What is Big Data?**

A: Extremely large datasets that cannot be handled with traditional tools.

**Q: What are the characteristics of Big Data?**

A: Volume, Velocity, Variety, Veracity, Value.

**Q: List phases in Data Science life cycle.**

A: Data Collection, Data Cleaning, Data Exploration, Feature Engineering, Model Building, Evaluation, Deployment, Monitoring.

**Q: What is Central Tendency?**

A: Measures that describe the center of data: Mean, Median, Mode.

**Q: What is Dispersion?**

A: Measures spread: Range, Variance, Standard Deviation.

**Q: What is Mean, Mode, Mid-range, Median for [10,22,13,10,21,43,77,21,10]?**

A: Mean = 25.22, Mode = 10, Median = 21, Mid-range = 43.5

**Q: What is Variance?**

A: Measure of data spread from the mean.

**Q: What is Standard Deviation?**

A: Square root of variance; shows how much data deviates from mean.

**Q: What is Posterior Probability in Naive Bayes?**

A: Probability of the class given the input features.

**Q: What is Likelihood Probability in Naive Bayes?**

A: Probability of the input features given the class.

**Q: How to deal with missing values?**

A: Drop rows: df.dropna(), Fill values: df.fillna(), Imputation: mean/median.

**Q: What is NLTK?**

A: Natural Language Toolkit - a Python library for NLP tasks.

**Q: What is Tokenization in NLP?**

A: Splitting text into individual words or tokens.

**Q: What is Stemming?**

A: Reducing words to their root form (e.g., 'playing' -> 'play').

**Q: What is Lemmatization?**

A: Similar to stemming but returns real words (e.g., 'better' -> 'good').

**Q: What is Corpus in NLP?**

A: A large collection of text used for training NLP models.

**Q: What is Spark Framework?**

A: Apache Spark is a distributed computing framework used for big data processing and analytics.