1. What is EDA in Python?

Answer: Exploratory Data Analysis (EDA) in Python is an approach to analyzing data sets to summarize their main characteristics, often using visualization methods.

2. What are the key libraries in Python for performing EDA?

Answer: Key libraries include Pandas, NumPy, Matplotlib, Seaborn, and Plotly.

3. How do you load a CSV file in Python using Pandas?

Answer: You can use the read\_csv() function from the Pandas library, like this:

import pandas as pd

data = pd.read\_csv('file.csv')

4. What is the purpose of the head() function in Pandas?

Answer: The head() function is used to display the first few rows of a DataFrame, providing a quick overview of its structure and content.

5. How do you check for missing values in a DataFrame?

Answer: You can use the isnull() method followed by sum() to count missing values in each column, like this:

print(data.isnull().sum())

6. How do you handle missing values in Python?

Answer: You can handle missing values by either removing them, filling them with a specific value (like the mean or median), or using more advanced techniques like interpolation.

7. What are the main types of plots used in EDA?

Answer: Common types include histograms, scatter plots, box plots, bar plots, and heatmaps.

8. How do you create a histogram in Python using Matplotlib?

Answer: You can use the hist() function from Matplotlib, like this:

import matplotlib.pyplot as plt

plt.hist(data['column\_name'], bins=10)

plt.show()

9. What is a box plot? How is it useful in EDA?

Answer: A box plot (or box-and-whisker plot) is a graphical summary of the distribution of a dataset. It displays the median, quartiles, and potential outliers. It's useful in EDA for identifying central tendency, variability, and outliers in the data.

10. How do you create a box plot in Python using Seaborn?

Answer: You can use the boxplot() function from Seaborn, like this:

import seaborn as sns

sns.boxplot(x='column\_name', data=data)

plt.show()

11. What is correlation? How do you calculate it in Python?

Answer: Correlation measures the strength and direction of the relationship between two variables. In Python, you can calculate it using the corr() function from Pandas.

12. How do you visualize correlation matrices in Python?

Answer: You can use a heatmap to visualize correlation matrices. Seaborn's heatmap() function is commonly used for this purpose.

13. What is the purpose of outlier detection in EDA?

Answer: Outlier detection helps identify data points that deviate significantly from the rest of the data. These outliers can sometimes indicate errors or anomalies in the data.

14. How do you detect outliers in Python?

Answer: You can use statistical methods like z-score or IQR (Interquartile Range) to detect outliers in Python.

15. Explain z-score method for outlier detection.

Answer: Z-score method involves calculating the z-score for each data point, which measures how many standard deviations it is from the mean. Data points with z-scores beyond a certain threshold (commonly ±3) are considered outliers.

16. How do you handle outliers in Python?

Answer: Outliers can be handled by either removing them, transforming the data, or using more robust statistical techniques that are less sensitive to outliers.

17. What is skewness? How do you detect skewness in a dataset?

Answer: Skewness measures the asymmetry of the probability distribution of a real-valued random variable about its mean. You can detect skewness in a dataset by calculating its skewness coefficient using libraries like SciPy.

18. How do you handle skewness in Python?

Answer: Skewness can be handled by transforming the data using techniques like log transformation, square root transformation.

19. What are the main steps in EDA?

Answer: The main steps include data collection, data cleaning, data exploration, and visualization, statistical analysis, and drawing conclusions.

20. What is the purpose of data transformation in EDA?

Answer: Data transformation is used to convert the original data into a format that is more suitable for analysis. It can involve normalization, standardization, or transformations to correct skewness.

21. What is the purpose of a correlation matrix in EDA?

Answer: A correlation matrix is used to examine the relationships between multiple variables in a dataset, helping to identify patterns and dependencies.

22. How do you create a correlation matrix in Python?

Answer: You can use the corr() function from Pandas to calculate the correlation matrix for a DataFrame.

23. What does a correlation coefficient value of 0 indicate?

Answer: A correlation coefficient value of 0 indicates no linear relationship between the two variables.

24. What is the difference between positive and negative correlation?

Answer: Positive correlation means that as one variable increases, the other variable also tends to increase, while negative correlation means that as one variable increases, the other variable tends to decrease.

25. How do you interpret the strength of correlation coefficients?

Answer: The strength of correlation coefficients is typically interpreted as follows: close to 1 or -1 indicates strong correlation, close to 0 indicates weak correlation.

26. What is the purpose of a scatter plot in EDA?

Answer: A scatter plot is used to visualize the relationship between two continuous variables, helping to identify patterns such as correlations, clusters, or outliers.

27. What is the purpose of a pair plot in EDA?

Answer: A pair plot is used to visualize pairwise relationships between variables in a dataset. It creates scatter plots for numerical variables and histograms for categorical variables along the diagonal.

28. What is the purpose of a bar plot in EDA?

Answer: A bar plot is used to visualize the distribution of a categorical variable, often showing the frequency or proportion of each category.

29. What is the purpose of a count plot in EDA?

Answer: A count plot is a specialized form of a bar plot used to count the occurrences of each category in a categorical variable.

30. What is the lifecycle of a data science project?

Data Collection: Gathering relevant data from various sources.

Exploratory Data Analysis (EDA): Understanding the data, identifying patterns, and visualizing relationships.

Model Training and Testing: Building predictive models using machine learning algorithms.

31. Differentiate between Univariate, Bivariate, and Multivariate analysis:

* + Univariate Analysis: Examining one variable at a time. It aims to describe the variable and find patterns within it (e.g., analyzing student heights).
  + Bivariate Analysis: Involves two different variables. It explores relationships and causes between them (e.g., temperature vs. ice cream sales).

32. What are the two kinds of target variables for predictive modeling?

* + Numerical/Continuous Variable: Values lie within a range (e.g., student heights). Predictions can be any value within that range.
  + Categorical Variable: Takes on a limited, fixed number of possible values (e.g., class labels).

33. Explain about sampling distribution?

Answer: Sampling distribution refers to the distribution of statistics, like the mean or proportion, calculated from multiple samples drawn from the same population. It helps understand how these statistics vary across different samples and is essential for making inferences about population parameters. The central limit theorem states that the sampling distribution of the sample mean tends towards a normal distribution as the sample size increases, regardless of the population distribution, making it a fundamental concept in statistical inference.

34. Explain in 2 lines about statistical terms?

Mean: It's the average value of a dataset calculated by summing all values and dividing by the total number of observations, providing a measure of central tendency.

Median: The middle value of a dataset when arranged in ascending order; it's robust to extreme values and gives another measure of central tendency.

Mode: The most frequently occurring value in a dataset, providing another measure of central tendency.

Standard deviation: It measures the average deviation of data points from the mean, providing a measure of dispersion in the dataset.

Range: The difference between the maximum and minimum values in a dataset, providing a simple measure of variability.