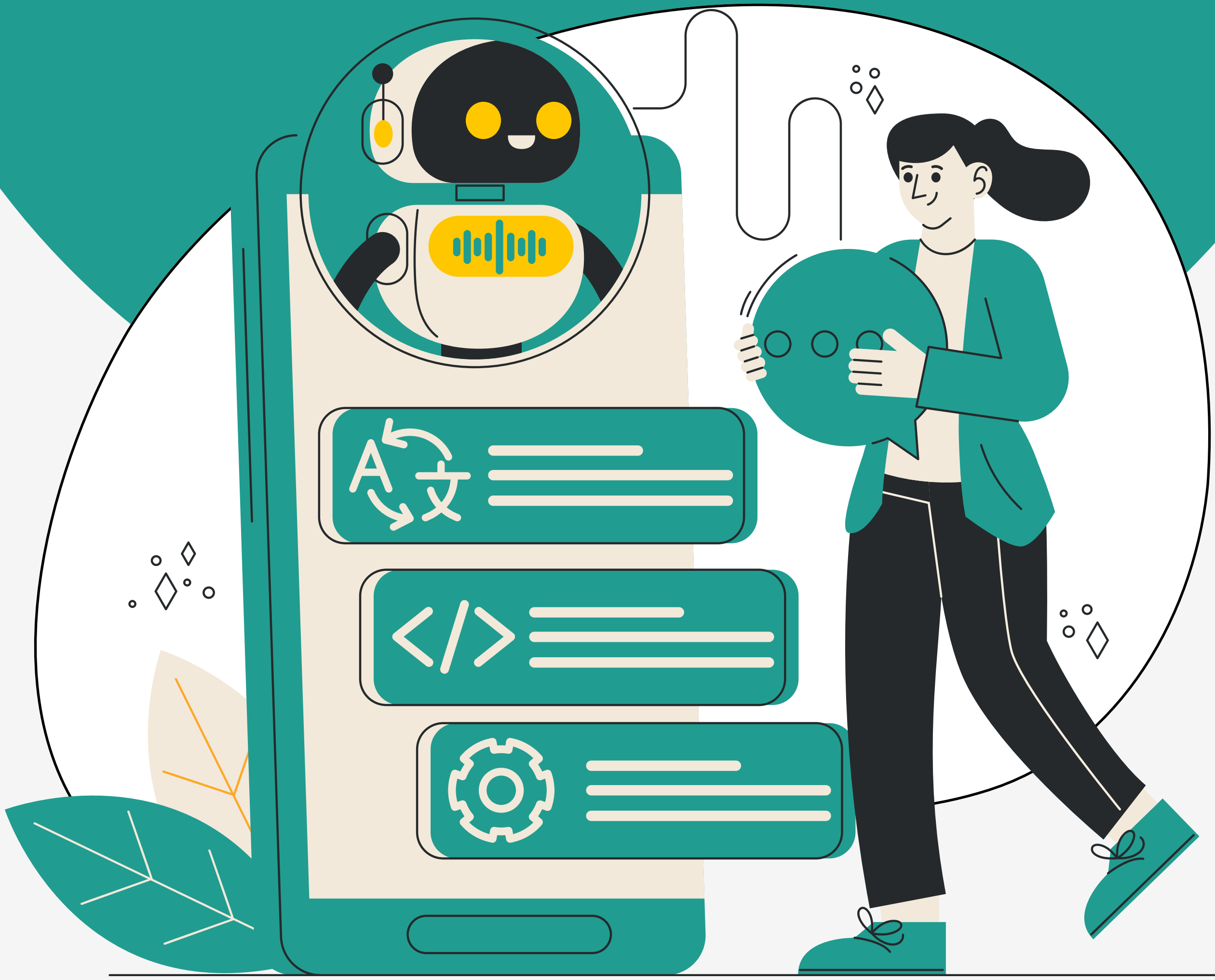
EDA

Interview Questions

(Practice Project)



PW SKILLS

Easy Questions

1. **What is the purpose of graphical representations in EDA?**

Answer: Graphical representations are used to visualize data in a way that highlights patterns, relationships, and trends. They simplify complex data, making it easier to interpret and communicate insights.

2. **What type of data is best visualized using a histogram?**

Answer: A histogram is best suited for visualizing the distribution of continuous variables.

3. **What information does a box plot provide?**

Answer: A box plot provides information about the median, quartiles, interquartile range (IQR), and potential outliers in a dataset.

4. **When would you use a scatter plot?**

Answer: A scatter plot is used to explore the relationship between two continuous variables.

5. **What distinguishes a bar chart from a histogram?**

Answer: A bar chart displays categorical data with rectangular bars representing the frequency or value of each category, while a histogram displays the distribution of a continuous variable by grouping data into bins.

6. **Why are line charts often used in time series analysis?**

Answer: Line charts are used in time series analysis because they effectively display trends and changes over time by connecting data points with a continuous line.

7. **What does the height of bars in a histogram represent?**

Answer: The height of bars in a histogram represents the frequency of data points within each bin or interval.

Medium Questions

8. **What are the advantages of using a violin plot over a box plot?**

Answer: Violin plots provide additional information by combining box plots and kernel density plots, allowing for a visual comparison of data distribution across different levels of a categorical variable. They can reveal variations and multimodal distributions that might not be visible in a box plot.

9. **How can you identify outliers using a box plot?**

Answer: Outliers in a box plot are typically represented as individual points that fall outside the whiskers, which extend to 1.5 times the interquartile range (IQR) from the first and third quartiles.

10. **What is the significance of using a heatmap in EDA?**

Answer: Heatmaps are significant in EDA as they provide a visual representation of data values through color coding, which helps in identifying patterns, correlations, and the intensity of values within a matrix.

11. **Describe a situation where you would prefer a pair plot over a scatter plot.**

Answer: A pair plot is preferred when you want to analyze the relationships between multiple variables simultaneously, as it creates a matrix of scatter plots for each pair of variables, making it easier to observe correlations and trends

12. **What insights can be gained from a 3D scatter plot that might be missed in a 2D scatter plot?**
Answer: A 3D scatter plot allows for the exploration of relationships between three variables simultaneously, providing a more comprehensive view of the data. It can reveal complex interactions and patterns that are not visible in a 2D plot.

13. **How would you use a heatmap to visualize a correlation matrix?**
Answer: A heatmap can be used to visualize a correlation matrix by representing the correlation coefficients between variables with different colors. This makes it easy to identify strong positive or negative correlations and patterns within the data.

14. **What is the main difference between a line chart and a scatter plot?**
Answer: A line chart connects data points with a continuous line and is commonly used for time series data, while a scatter plot shows individual data points without connecting lines, typically used to explore the relationship between two variables.

Hard Questions

15. **Explain how you can use a 3D surface plot to analyze data.**
Answer: A 3D surface plot is used to visualize the relationship between three variables by plotting data points on a three-dimensional surface. This can help identify peaks, valleys, and trends in the data that represent interactions between the variables. It's particularly useful for visualizing complex functions or models.

16. **What challenges might arise when interpreting a heatmap with a large dataset?**
Answer: Interpreting a heatmap with a large dataset can be challenging due to overlapping data points, which may obscure patterns. Additionally, selecting appropriate color scales and managing large amounts of information can be difficult, leading to potential misinterpretation if not handled carefully.

17. **How can pair plots be used to detect multicollinearity in a dataset?**
Answer: Pair plots can detect multicollinearity by visualizing the relationships between multiple variables. Strong linear relationships between two or more independent variables in the pair plots suggest multicollinearity, which can affect the accuracy of regression models.

18. **Describe a scenario where a violin plot might provide misleading information.**
Answer: A violin plot might provide misleading information if the data distribution is heavily skewed or if there are too few data points, as the density estimation might exaggerate certain aspects of the distribution, leading to incorrect interpretations.

19. **How can you enhance the readability of a complex 3D plot?**
Answer: To enhance the readability of a complex 3D plot, you can use techniques such as rotating the plot for different perspectives, adjusting the color scheme for better contrast, adding grid lines or contours, and simplifying the plot by focusing on key data points or regions.

20. **What are the limitations of using graphical representations in EDA?**
Answer: Graphical representations in EDA are limited by their subjective nature, as interpretations can vary between viewers. They may also oversimplify complex data or obscure details in large datasets. Additionally, creating accurate and effective visualizations requires skill, as poorly designed graphs can mislead or fail to convey the intended insights.