106 Pandas interview questions to hire top engineers

Questions

1. What is Pandas? Explain it like I'm five.

2. Can you describe the difference between a Pandas Series and a Pandas DataFrame?

3. How do you create a DataFrame in Pandas?

4. How can you read a CSV file into a Pandas DataFrame?

5. How do you view the first 5 rows of a DataFrame?

6. How would you inspect the last few rows in a DataFrame?

7. How do you get the number of rows and columns in a DataFrame?

8. How can you get the column names of a DataFrame?

9. How do you select a single column from a DataFrame?

10. How do you select multiple columns from a DataFrame?

11. How do you filter rows based on a condition in Pandas? Can you give an example?

12. How can you sort a DataFrame by a specific column?

13. How do you group data in a DataFrame using Pandas? What are some common aggregation functions?

14. How do you handle missing values in Pandas? What are the common techniques?

15. How do you rename columns in a Pandas DataFrame?

16. How do you add a new column to a DataFrame?

17. How do you remove a column from a DataFrame?

18. How can you iterate over rows in a DataFrame? Is it generally recommended?

19. How do you apply a function to each element in a Pandas Series or DataFrame?

20. Explain how to merge two DataFrames in Pandas. What are the different types of merges?

21. How do you concatenate two DataFrames in Pandas?

22. How do you calculate descriptive statistics (like mean, median, standard deviation) for a DataFrame?

23. How do you write a DataFrame to a CSV file?

24. What are some ways to efficiently handle large datasets in Pandas, considering memory constraints?

25. How do you efficiently combine data from multiple Pandas DataFrames when they share a common column, but the column names are different?

26. Can you explain how to use the `.pipe()` method in Pandas to chain multiple operations together for data transformation?

27. Describe how you would handle missing data in a Pandas DataFrame, including imputation strategies and considerations for different data types.

28. How can you create a pivot table in Pandas to summarize data based on multiple index and value columns, and how do you handle missing values in the resulting table?

29. Explain how to use the `pd.Grouper` object in Pandas to group data by time intervals, and provide an example use case.

30. How would you optimize the performance of a Pandas operation on a large dataset, considering techniques like chunking or using more efficient data types?

31. Describe how to perform a rolling window calculation on a Pandas Series, and explain different window types and aggregation functions.

32. How do you apply a custom function to each element in a Pandas DataFrame, and what are the performance implications compared to vectorized operations?

33. Explain how to create a multi-level index in Pandas, and how to access and manipulate data using the different levels of the index.

34. How can you convert a Pandas DataFrame to a sparse matrix format, and what are the benefits of doing so for memory usage and computation speed?

35. Describe how to perform a fuzzy merge in Pandas, where you match rows based on approximate string matching rather than exact equality.

36. How do you use the Pandas `eval()` function to perform arithmetic operations on columns, and what are the advantages of using it over standard operators?

37. Explain how to read and write data to a SQL database using Pandas, including handling different data types and performing SQL queries.

38. How can you create a custom aggregation function in Pandas to calculate a statistic that is not available in the built-in aggregation functions?

39. Describe how to perform a time series resampling operation in Pandas, including different resampling frequencies and interpolation methods.

40. How do you handle categorical data in Pandas, including encoding categorical variables and using them in machine learning models?

41. Explain how to use the Pandas `Styler` object to format and style DataFrames for presentation, including conditional formatting and custom CSS styles.

42. How can you create a heatmap visualization of a Pandas DataFrame using Seaborn or Matplotlib, and how do you interpret the heatmap?

43. Describe how to perform a network analysis using Pandas and NetworkX, including creating a graph from a DataFrame and calculating network metrics.

44. How do you use the Pandas `qcut()` function to discretize a continuous variable into quantiles, and how do you handle edge cases with duplicate values?

45. Explain how to perform a geographical analysis using Pandas and GeoPandas, including reading and writing geospatial data and performing spatial operations.

46. How can you create a dashboard using Pandas and Plotly or Bokeh, including interactive widgets and data updates?

47. Describe how to perform a text analysis using Pandas and NLTK or SpaCy, including tokenization, stemming, and sentiment analysis.

48. How do you use the Pandas `Categorical` data type to represent ordinal or nominal data, and how does it differ from a standard object column?

49. Explain how to perform a survival analysis using Pandas and Lifelines, including estimating survival curves and comparing different groups.

50. How can you create a recommendation system using Pandas and scikit-learn, including collaborative filtering and content-based filtering?

51. Describe how to perform a time series forecasting using Pandas and Prophet or ARIMA, including model fitting and evaluation.

52. How do you use the Pandas `Index` object to optimize data access and filtering, and how does it differ from a standard column?

53. Explain how to perform an anomaly detection using Pandas and isolation forest or one-class SVM, including model training and threshold selection.

54. How can you create a data pipeline using Pandas and Dask or Spark, including data loading, transformation, and storage?

55. How can you optimize Pandas code for speed and memory usage when dealing with large datasets?

56. Explain the difference between `.loc` and `.iloc` in Pandas, and when would you use each?

57. Describe how to handle missing data in Pandas, including imputation techniques.

58. How would you perform a multi-index sort in Pandas and why might you use it?

59. Explain how to use `pd.Grouper` for custom time-based aggregation.

60. How do you efficiently combine multiple Pandas DataFrames with different structures?

61. Explain how to apply a custom function to a Pandas DataFrame that depends on multiple columns using apply.

62. Describe how to perform a rolling window calculation on a Pandas Series or DataFrame.

63. How can you create pivot tables and cross-tabulations using Pandas and what are their differences?

64. Explain how to use Pandas with scikit-learn pipelines for data preprocessing and modeling.

65. How would you convert a Pandas DataFrame to a sparse matrix format and when is this useful?

66. Describe how to write a Pandas DataFrame to a database using SQLAlchemy and handle potential issues.

67. How can you use Pandas to read and process data from different file formats (e.g., JSON, CSV, Excel) with custom parsing?

68. Explain how to use `pd.merge` to perform different types of joins (inner, outer, left, right) with detailed examples.

69. Describe how to implement a custom aggregation function using `groupby` and `agg` in Pandas.

70. How do you handle categorical data in Pandas, including one-hot encoding and custom mappings?

71. Explain how to use `pd.cut` and `pd.qcut` for binning continuous variables.

72. How would you debug performance issues in Pandas code?

73. Describe how to implement a time series analysis using Pandas, including resampling and shifting.

74. How can you create a new column in a Pandas DataFrame based on complex conditions applied to other columns?

75. How would you optimize a Pandas operation that is slow due to iterating over rows?

76. Explain how you would handle a very large CSV file with Pandas that doesn't fit into memory.

77. Describe a scenario where you would use Pandas' Categorical data type and why.

78. How do you handle missing data in a Pandas DataFrame and what are the trade-offs of different approaches?

79. Explain how to use Pandas to perform a time series analysis, including resampling and windowing operations.

80. Describe how you would implement a custom aggregation function using Pandas' groupby functionality.

81. How can you use Pandas to efficiently join multiple DataFrames with different index structures?

82. Explain how you would debug performance issues in Pandas code.

83. Describe how you would use Pandas to create a pivot table and analyze the results.

84. How can you leverage the `.pipe()` method in Pandas to create a readable and maintainable data processing pipeline?

85. Explain the difference between `.apply()`, `.map()`, and `.applymap()` in Pandas, and when would you use each?

86. How would you implement a fuzzy string matching algorithm using Pandas to clean up inconsistent data?

87. Describe how to use Pandas with other libraries like Scikit-learn for building machine learning models.

88. How would you use Pandas to identify and remove duplicate data based on multiple columns?

89. Explain how you can create a multi-index DataFrame in Pandas and how to query data from it.

90. Describe how to handle timezone conversions using Pandas datetime objects.

91. How would you use Pandas to perform cohort analysis?

92. Explain how to use Pandas to read data from a database and write data back to it.

93. Describe how you would use Pandas to analyze text data, including tokenization and sentiment analysis.

94. How would you implement a custom rolling window function using Pandas?

95. Explain how to use Pandas to create interactive visualizations using libraries like Plotly or Bokeh.

96. Describe a scenario where you would use Pandas' Sparse data structures and why.

97. How would you use Pandas to perform A/B testing analysis?

98. Explain how to use Pandas to create a correlation matrix and interpret the results.

99. Describe how to use the `pd.eval()` function to speed up certain Pandas operations.

100. How would you handle data skewness when performing calculations using Pandas?

101. Explain how you would use Pandas to create a data dictionary that describes the columns and data types in a DataFrame.

102. Describe strategies for minimizing memory usage when working with large Pandas DataFrames.