Topic: Introduction

1. What is the primary purpose of the Pandas library in Python, especially concerning Machine Learning?

(a) To perform complex mathematical computations for deep learning models.

(b) To provide high-performance, easy-to-use data structures and data analysis tools.

(c) To enable parallel processing of large datasets across multiple cores.

(d) To visualize data through advanced 3D graphics and animations.

2. Which of the following is NOT a fundamental data structure provided by the Pandas library?

(a) Series

(b) DataFrame

(c) Array

(d) Index

3. A Pandas Series is best described as:

(a) A two-dimensional labeled data structure with columns of potentially different types.

(b) A one-dimensional labeled array capable of holding any data type.

(c) A three-dimensional array for handling multi-panel data.

(d) An immutable, ordered sequence of elements.

4. To create a Pandas DataFrame from a dictionary of lists, where keys become column names, which function is used?

(a) pd.Series()

(b) pd.DataFrame()

(c) pd.read_csv()

(d) pd.array()

5. Which Pandas function is primarily used to load tabular data from a Comma Separated Values (CSV) file into a DataFrame?

(a) pd.load_csv()

(b) pd.get_csv()

(c) pd.read_csv()

(d) pd.open_csv()

6. To check which cells in a Pandas DataFrame contain missing or null values, which method is most commonly applied?

(a) df.has_null()

(b) df.is_empty()

(c) df.isna()

(d) df.check_missing()

7. If you want to remove all rows from a DataFrame that contain at least one missing value, which method would you use?

(a) df.drop_nulls()

(b) df.remove_na()

(c) df.dropna()

(d) df.fill_na()

8. To permanently remove a specific column named 'Unwanted_Feature' from a DataFrame df, which of the following is the correct syntax?

(a) df.drop('Unwanted_Feature', axis=0)

(b) df.remove_column('Unwanted_Feature')

(c) df.drop(columns='Unwanted_Feature', inplace=True)

(d) df['Unwanted_Feature'] = None

9. Which method is used to identify rows in a DataFrame that are exact duplicates of earlier rows?

(a) df.unique()

(b) df.is_duplicate()

(c) df.duplicated()

(d) df.find_duplicates()

10. By default, what does the axis parameter in the df.dropna() method target?

(a) Columns (axis=1)

(b) Rows (axis=0)

(c) Both rows and columns

(d) Only cells with NaN values

11. Which Pandas DataFrame method provides a quick summary of the numerical data, including count, mean, standard deviation, min, max, and quartile values?

(a) df.info()

(b) df.summary()

(c) df.describe()

(d) df.statistics()

12. How would you select the column named 'Age' from a Pandas DataFrame called data?

(a) data.get_column('Age')

(b) data['Age']

(c) data('Age')

(d) data.column('Age')

13. Which aspect makes Pandas particularly suitable for the data preparation and exploration phases of Machine Learning?

(a) Its ability to directly integrate with neural networks without preprocessing.

(b) Its powerful capabilities for handling and manipulating structured tabular data.

(c) Its automatic feature selection algorithms.

(d) Its built-in functionality for hyperparameter tuning.

14. When you call df.plot() on a Pandas DataFrame containing numerical data without specifying a kind argument, what is the default type of plot generated?

(a) Bar plot

(b) Scatter plot

(c) Line plot

(d) Histogram

15. Suppose you have a DataFrame df and you want to remove the row at index label 'A'. Which of the following commands correctly achieves this?

(a) df.drop_row('A')

(b) df.drop(index='A')

(c) df.remove('A', axis=0)

(d) df.loc['A'].drop()

Answers

1. (b)

2. (c)

3. (b)

4. (b)

5. (c)

6. (c)

7. (c)

8. (c)

9. (c)

10. (b)

11. (c)

12. (b)

13. (b)

14. (c)

15. (b)

Topic: Series: Series()Dataframes: DataFrames()

Section: Multiple Choice Questions

16. What is a fundamental characteristic of a Pandas Series?

(a) It can store data of multiple different data types in a single Series.

(b) It is a one-dimensional labeled array capable of holding data of any type.

(c) It requires all its elements to be of the same data type as a NumPy array.

(d) It is equivalent to a Python dictionary with integer keys.

17. Which of the following best describes a Pandas DataFrame?

(a) A two-dimensional labeled array capable of holding data of a single data type.

(b) A collection of Python lists, each representing a column.

(c) A two-dimensional labeled data structure with columns that can be of different types.

(d) A collection of Pandas Series where each Series must have the same index.

18. import pandas as pd

s = pd.Series([10, 20, 30], index=['a', 'b', 'c'])

What will be the output of s.index?

(a) RangeIndex(start=0, stop=3, step=1)

(b) Index(['a', 'b', 'c'], dtype='object')

(c) ['a', 'b', 'c']

(d) 0 1 2

19. import pandas as pd

data = {'Name': ['Alice', 'Bob'], 'Age': [25, 30]}

df = pd.DataFrame(data)

What is the shape of the DataFrame df?

(a) (2, 2)

(b) (2, 0)

(c) (0, 2)

(d) (4, 2)

20. Which parameter of the pd.read_csv() function is used to specify a custom delimiter (e.g., a semicolon instead of a comma)?

(a) separator

(b) delimiter

(c) sep

(d) delim

21. Consider a DataFrame df with some missing values (NaN). If you execute df.dropna(inplace=True), what will be the effect?

(a) It returns a new DataFrame with rows containing NaN values removed, leaving df unchanged.

(b) It returns a new DataFrame with columns containing NaN values removed, leaving df unchanged.

(c) It modifies df directly by removing rows that contain at least one NaN value.

(d) It modifies df directly by replacing NaN values with the mean of their respective columns.

22. You have a DataFrame df. Which of the following commands will return a boolean Series indicating whether each row is a duplicate of a previous row?

(a) df.is_duplicate()

(b) df.unique()

(c) df.duplicated()

(d) df.duplicates()

23. import pandas as pd

data = {'A': [1, 2, 3], 'B': [4, 5, 6]}

df = pd.DataFrame(data)

What is the correct way to select only column 'A' as a Series?

(a) df[['A']]

(b) df.loc[:, 'A']

(c) df.iloc[:, 0:1]

(d) df[0]

24. Which statement accurately differentiates a Pandas Series from a DataFrame?

(a) A Series can only store numerical data, while a DataFrame can store various data types.

(b) A Series is primarily used for statistical analysis, while a DataFrame is for data manipulation.

(c) A Series is a 1D labeled array, whereas a DataFrame is a 2D labeled data structure with potentially heterogeneous columns.

(d) A Series can have multiple columns, but a DataFrame is limited to a single column.

25. If you want to remove a specific column named 'Unnecessary' from a DataFrame df without creating a new DataFrame, which of the following is the most appropriate command?

(a) df.remove('Unnecessary', axis=1)

(b) df.drop(columns=['Unnecessary'], inplace=True)

(c) df.delete('Unnecessary')

(d) df['Unnecessary'].drop()

26. To generate a basic line plot of a Series 's', which method would you typically use?

(a) s.plot(kind='line')

(b) s.plot.line()

(c) s.lineplot()

(d) Both (a) and (b) are correct.

27. What does df.columns return for a DataFrame df?

(a) A list of row labels (index).

(b) A list of column names.

(c) The number of columns.

(d) The data types of each column.

28. Consider a DataFrame df with a column 'Age' containing some non-numeric values (e.g., 'twenty'). Which of the following approaches is generally suitable for handling such "wrong data" to enable numerical operations on 'Age'?

(a) Use df.drop() to remove all rows with non-numeric 'Age' values.

(b) Convert the 'Age' column to numeric using pd.to_numeric() and handle errors (e.g., coerce them to NaN).

(c) Replace all non-numeric values with a default value like 0.

(d) Filter out only the numeric values and create a new Series.

29. import pandas as pd

s = pd.Series([5, 10, 15], index=['x', 'y', 'z'])

What is the output of s['y']?

(a) 10

(b) 1

(c) 5

(d) 'y'

30. import pandas as pd

df = pd.DataFrame({'col1': [1, 2], 'col2': [3, 4]}, index=['row1', 'row2'])

What will df.loc['row2', 'col1'] return?

(a) 4

(b) 3

(c) 2

(d) 1

Answers

16. (b)

17. (c)

18. (b)

19. (a)

20. (c)

21. (c)

22. (c)

23. (b)

24. (c)

25. (b)

26. (d)

27. (b)

28. (b)

29. (a)

30. (c)

Topic: Read CSV File: read_csv()

Section: Multiple Choice Questions

31. Which Pandas function is primarily used to read data from a CSV (Comma Separated Values) file into a DataFrame?

a) pd.load_csv()

b) pd.read_table()

c) pd.read_csv()

d) pd.import_csv()

32. By default, the pd.read_csv() function assumes that the values in the CSV file are separated by which character?

a) Tab (t)

b) Semicolon (;)

c) Space ( )

d) Comma (,)

33. To read a CSV file named 'data.txt' where values are separated by a tab character, which parameter of pd.read_csv() should be used?

a) delimiter=t

b) separator='\t'

c) sep='\t'

d) split_by=t

34. Consider a CSV file 'students.csv' that does not contain a header row. To instruct pd.read_csv() to treat the first row as data rather than column names, which argument would you use?

a) header=None

b) no_header=True

c) skip_header=1

d) infer_header=False

35. You have a CSV file 'sales.csv' where the first column contains unique transaction IDs. To automatically set this column as the DataFrame's index upon reading, which parameter should be specified?

a) id_column=0

b) index_col=0

c) set_index=0

d) primary_key=0

36. If you only want to load 'Name' and 'Age' columns from a CSV file 'people.csv' that also contains 'Address' and 'Phone' columns, which parameter in pd.read_csv() would you use?

a) select_cols=['Name', 'Age']

b) include_columns=['Name', 'Age']

c) usecols=['Name', 'Age']

d) fetch_columns=['Name', 'Age']

37. A CSV file 'temperatures.csv' contains 'Date' and 'Temperature' columns. To ensure that the 'Date' column is parsed as datetime objects immediately upon loading, which parameter is most appropriate?

a) date_format='%Y-%m-%d'

b) convert_dates=['Date']

c) parse_dates=['Date']

d) as_datetime=['Date']

38. Which of the following is the return type of the pd.read_csv() function?

a) Pandas Series

b) Python list of lists

c) Pandas DataFrame

d) Numpy array

39. You are reading a CSV file where 'NA', 'N/A', and '' (empty string) all represent missing values. Which parameter allows you to specify these custom missing value representations?

a) missing_values=['NA', 'N/A', '']

b) na_values=['NA', 'N/A', '']

c) fill_na=['NA', 'N/A', '']

d) null_replacements=['NA', 'N/A', '']

40. To skip the first 5 rows of a CSV file 'sensor_data.csv' because they contain irrelevant metadata, which parameter should be used in pd.read_csv()?

a) skip_rows=5

b) skiprows=range(5)

c) skiprows=5

d) skip_first=5

41. When dealing with very large CSV files that might consume a lot of memory, which parameter or approach is often recommended to read the file in smaller, manageable chunks?

a) memory_efficient=True

b) chunksize=N (where N is an integer)

c) low_memory=True (without chunksize)

d) read_partial=True

42. A CSV file 'log.csv' occasionally has malformed rows that cause parsing errors. Which parameter in pd.read_csv() can be used to handle such errors, for example, by skipping bad lines instead of raising an exception?

a) error_handling='skip'

b) on_bad_lines='skip'

c) ignore_errors=True

d) bad_rows_action='omit'

43. You need to load data from a compressed CSV file named 'archive.zip/data.csv'. How would you typically specify the file path to pd.read_csv()?

a) pd.read_csv('zip://archive.zip/data.csv')

b) pd.read_csv('archive.zip/data.csv') - Pandas handles decompression automatically.

c) pd.read_csv(zipfile.ZipFile('archive.zip').open('data.csv'))

d) You must first decompress the file manually before reading it.

44. After successfully reading a CSV file into a DataFrame using pd.read_csv(), which of the following is a common immediate next step in data cleaning or preparation, especially if the 'duplicates()' method might be used later?

a) Saving the DataFrame to a new CSV file.

b) Visualizing the data using plot().

c) Checking for and handling missing values (e.g., using dropna()).

d) Performing complex statistical analysis.

45. You want to read a CSV file and explicitly define the data type for the 'ID' column as string (object) and 'Amount' column as float64. Which parameter allows you to specify this?

a) data_types={'ID': str, 'Amount': float}

b) dtype={'ID': object, 'Amount': 'float64'}

c) column_types={'ID': 'string', 'Amount': 'float'}

d) type_map={'ID': object, 'Amount': float}

Answers

31. (c)

32. (d)

33. (c)

34. (a)

35. (b)

36. (c)

37. (c)

38. (c)

39. (b)

40. (c)

41. (b)

42. (b)

43. (b)

44. (c)

45. (b)

Topic: Cleaning Empty Cells: dropna()

46. What is the primary purpose of the dropna() method in Pandas DataFrames?

a) To fill missing values with a specified value.

b) To remove rows or columns containing missing values.

c) To convert missing values to a specific data type.

d) To identify duplicated rows with missing values.

47. By default, how does the dropna() method behave when called on a Pandas DataFrame?

a) It removes columns where at least one missing value is present.

b) It removes rows where at least one missing value is present.

c) It removes rows only if all values in the row are missing.

d) It removes columns only if all values in the column are missing.

48. Which parameter of the dropna() method is used to specify whether to drop rows or columns?

a) how

b) subset

c) axis

d) inplace

49. Consider a DataFrame df. If df.dropna(how='all') is executed, which rows or columns will be removed?

a) Rows containing at least one NaN value.

b) Rows where all values are NaN.

c) Columns containing at least one NaN value.

d) Columns where all values are NaN.

50. What is the effect of setting the 'inplace' parameter to True in df.dropna(inplace=True)?

a) It returns a new DataFrame with missing values removed, leaving the original DataFrame unchanged.

b) It modifies the original DataFrame directly, without returning a new DataFrame.

c) It signals that the operation should be undone if an error occurs.

d) It converts missing values to zeros in the original DataFrame.

51. You want to remove rows only if they have missing values in either the 'Name' column or the 'Age' column. Which parameter of dropna() would you use for this?

a) axis

b) how

c) subset

d) thresh

52. A DataFrame df has columns 'A', 'B', 'C'. If you execute df.dropna(thresh=2), what will happen?

a) Rows will be dropped if they have fewer than 2 non-NaN values.

b) Rows will be dropped if they have more than 2 non-NaN values.

c) Columns will be dropped if they have fewer than 2 non-NaN values.

d) Columns will be dropped if they have more than 2 non-NaN values.

53. Which of the following values are typically treated as "missing" by the dropna() method in Pandas?

a) Empty strings ("")

b) The integer 0

c) pandas.NA or numpy.nan

d) The string "None"

54. You have read a CSV file into a DataFrame using pd.read_csv('data.csv'). If you then call df.dropna() without any parameters, what is the default behavior regarding missing values?

a) It removes rows containing any non-finite numbers like infinity.

b) It removes rows containing any Python None or NumPy NaN values.

c) It removes rows containing any empty strings.

d) It removes rows containing any default missing value markers specified during CSV parsing.

55. After applying df.dropna() to a DataFrame, what is the typical impact on the DataFrame's index?

a) The index is reset to a new sequential range starting from 0.

b) The original index values of the dropped rows are preserved, leading to gaps.

c) The index values are shifted to fill the gaps created by dropped rows.

d) The index is converted to a MultiIndex structure.

56. Consider the following sequence of operations on a DataFrame df:

1. df.drop_duplicates(inplace=True)

2. df.dropna(inplace=True)

What is the most likely intended outcome of this sequence?

a) To remove duplicate rows first, then remove rows with any missing values from the remaining unique rows.

b) To remove rows with missing values first, then remove duplicate rows from the resulting DataFrame.

c) To identify and count both duplicate and missing values simultaneously.

d) To replace duplicate values with missing values, then drop those new missing values.

57. If you call df_cleaned = df.dropna() (without inplace=True), what will df_cleaned contain?

a) A reference to the original DataFrame df, which has been modified.

b) A new DataFrame object with all rows containing missing values removed.

c) The original DataFrame df, as no operation was performed.

d) A Series object containing only the missing values from df.

58. What is the difference between df.dropna(axis=0) and df.dropna(axis=1)?

a) axis=0 drops columns, while axis=1 drops rows.

b) axis=0 drops rows, while axis=1 drops columns.

c) axis=0 fills missing values, while axis=1 removes them.

d) Both axis=0 and axis=1 perform the same operation of dropping rows with missing values.

59. You have a DataFrame df with columns ['A', 'B', 'C', 'D']. You want to drop a row only if all values in columns 'A' and 'C' are missing. Which of the following commands achieves this?

a) df.dropna(how='all', subset=['A', 'C'], axis=0)

b) df.dropna(how='any', subset=['A', 'C'], axis=0)

c) df.dropna(how='all', subset=['A', 'C'], axis=1)

d) df.dropna(how='any', subset=['A', 'C'], axis=1)

60. Which of the following is a potential side effect or important consideration when using dropna() extensively on a dataset?

a) It always leads to an increase in the dataset's size due to data imputation.

b) It can significantly reduce the number of observations, potentially leading to loss of valuable data or biased analysis.

c) It automatically fills all missing values with the mean or median of their respective columns.

d) It changes the data type of all remaining columns to object type.

Answers

46. (b)

47. (b)

48. (c)

49. (b)

50. (b)

51. (c)

52. (a)

53. (c)

54. (b)

55. (b)

56. (a)

57. (b)

58. (b)

59. (a)

60. (b)

Topic: Cleaning Wrong Data: drop()

Section: Multiple Choice Questions

63. What is the primary purpose of the Pandas DataFrame.drop() method in the context of data cleaning?

(a) To fill missing values with a specified constant.

(b) To remove specified rows or columns from a DataFrame.

(c) To convert the data types of columns in a DataFrame.

(d) To sort the DataFrame by one or more column values.

64. Consider a Pandas DataFrame named 'df'. Which of the following code snippets will remove the row with index label 'alpha' from 'df' and return a new DataFrame without modifying the original 'df'?

(a) df.drop('alpha', axis=1, inplace=True)

(b) df.drop(index='alpha', inplace=False)

(c) df.drop(columns='alpha', inplace=False)

(d) df.drop('alpha', axis=0)

65. To remove a column named 'CustomerID' from a DataFrame 'data_df', which method call is correct if the intention is to modify 'data_df' directly?

(a) data_df.drop('CustomerID', axis=0, inplace=True)

(b) data_df = data_df.drop(columns='CustomerID')

(c) data_df.drop(labels='CustomerID', axis=1, inplace=True)

(d) data_df.drop(index='CustomerID', inplace=False)

66. A student executes the following code:

```
import pandas as pd

df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6]})

df.drop('A', axis=1)
```

What will be the content of 'df' immediately after this execution?

(a) {'A': [1, 2, 3], 'B': [4, 5, 6]}

(b) {'B': [4, 5, 6]}

(c) {'A': [2, 3], 'B': [5, 6]}

(d) An error will occur because 'inplace' is not specified.

67. When using the drop() method, what is the significance of the 'axis' parameter?

(a) It specifies whether to drop rows with missing values or columns with missing values.

(b) It determines whether the operation should return a new DataFrame or modify the original one.

(c) It indicates whether to drop rows (axis=0) or columns (axis=1).

(d) It sets the threshold for the number of non-missing values required to keep a row or column.

68. To remove multiple specific rows from a DataFrame 'my_df' with index labels 'row_A' and 'row_C', which is the most concise and correct Pandas operation?

(a) my_df.drop(['row_A', 'row_C'], axis=1)

(b) my_df.drop(index=['row_A', 'row_C'])

(c) my_df.drop(columns=['row_A', 'row_C'])

(d) my_df = my_df.drop('row_A').drop('row_C')

69. What is the key difference between DataFrame.drop() and DataFrame.dropna() methods in Pandas?

(a) drop() removes rows/columns with any missing values, while dropna() removes specific rows/columns by label.

(b) drop() removes specific rows/columns by label, while dropna() removes rows/columns containing NaN values.

(c) drop() always modifies the DataFrame in-place, while dropna() always returns a new DataFrame.

(d) drop() is used for Series, while dropna() is used for DataFrames.

70. A Machine Learning engineer identifies that rows where the 'Status' column is 'Invalid' represent 'wrong data' that should be excluded from the dataset. Assuming 'df' is the DataFrame, which is the

most efficient way to achieve this using standard Pandas operations?

(a) df.drop(index=df[df['Status'] == 'Invalid'].index)

(b) df[df['Status'] != 'Invalid']

(c) df.dropna(subset=['Status'], how='any')

(d) df.drop('Status', axis=1)

71. If you attempt to drop a label (index or column name) that does not exist in a DataFrame using df.drop() without specifying the 'errors' parameter, what will typically happen?

(a) The method will proceed without any changes to the DataFrame.

(b) It will raise a KeyError.

(c) It will convert the non-existent label into a NaN value.

(d) It will drop a random label from the DataFrame.

72. You have a dataset loaded from a CSV file using read_csv(). It contains several columns, and you've identified that the column 'Sensor_ID' is entirely irrelevant for your machine learning model and contains 'wrong data' for your analytical purpose. What is the most direct way to remove this column from your DataFrame 'raw_data'?

(a) raw_data.dropna(subset=['Sensor_ID'], inplace=True)

(b) raw_data.drop('Sensor_ID', axis=1, inplace=True)

(c) raw_data.fillna(0, columns=['Sensor_ID'])

(d) raw_data.duplicated(subset=['Sensor_ID']).drop_duplicates()

73. Consider a Pandas Series 's'. Which of the following operations would remove the element at index label 'A' from 's' and assign the resulting Series to a new variable 's_new'?

(a) s_new = s.drop(columns='A')

(b) s_new = s.drop('A')

(c) s.drop('A', inplace=True)

(d) s_new = s.drop(index='A', axis=1)

74. When working with a DataFrame 'df' and cleaning 'wrong data', if you want to drop a column but need to ensure that the code doesn't crash even if the column might sometimes be missing from the DataFrame, which parameter should be used with df.drop()?

(a) errors='raise'

(b) inplace=False

(c) axis=1

(d) errors='ignore'

75. After loading a dataset into a Pandas DataFrame 'data_df', you realize that some rows contain completely bogus entries that are effectively 'wrong data' and should be removed. These rows are identified by their index labels, which are stored in a list called 'wrong_indices'. How would you remove these rows permanently from 'data_df'?

(a) data_df = data_df.drop(index=wrong_indices, axis=1)

(b) data_df.drop(columns=wrong_indices, inplace=True)

(c) data_df.drop(index=wrong_indices, inplace=True)

(d) data_df.drop(wrong_indices, axis=0)

76. Which of the following best describes the effect of df.drop(labels, inplace=False) on a DataFrame 'df'?

(a) It modifies 'df' by removing the specified labels and returns None.

(b) It returns a new DataFrame with the specified labels removed, leaving 'df' unchanged.

(c) It raises a ValueError because 'inplace' must be True to make any changes.

(d) It removes the specified labels from 'df' but warns that the operation is temporary.

77. Consider the following code snippet:

```
import pandas as pd

df = pd.DataFrame({'col1': [1, 2, 3], 'col2': ['A', 'B', 'C']})

new_df = df.drop(index=1, inplace=False)

print(df.shape)
```

print(new_df.shape)

What would be the output of the print statements?

(a) (3, 2)

(2, 2)

(b) (2, 2)

(2, 2)

(c) (3, 2)

(3, 2)

(d) (2, 2)

(3, 2)

Answers

63. (b)

64. (d)

65. (c)

66. (a)

67. (c)

68. (b)

69. (b)

70. (b)

71. (b)

72. (b)

73. (b)

74. (d)

75. (c)

76. (b)

77. (a)

Topic: Removing Duplicates: duplicated()

Section: Multiple Choice Questions

78. What is the primary purpose of the duplicated() method in Pandas DataFrames?

a) To identify rows with missing values.

b) To filter rows based on a specific condition.

c) To mark all duplicate rows in a DataFrame with a boolean True.

d) To convert string data to numerical format.

79. When applied to a Pandas DataFrame, what does the duplicated() method return by default?

a) An integer representing the count of duplicates.

b) A new DataFrame with only unique rows.

c) A Boolean Series, with its index aligned to the DataFrame, indicating if each row is a duplicate.

d) A list of duplicate row indices.

80. Consider a DataFrame df. If df.duplicated() is called without any parameters, which rows are marked as True?

a) Only the first occurrence of each duplicate set.

b) Only the last occurrence of each duplicate set.

c) All occurrences of a duplicate row, including the first.

d) All rows that are duplicates, excluding their first occurrence.

81. In df.duplicated(keep='last'), which rows will be marked as True?

a) All occurrences of a duplicate row.

b) All duplicate rows except for their last occurrence.

c) All duplicate rows except for their first occurrence.

d) Only the last occurrence of each duplicate set.

82. What is the behavior of df.duplicated(keep=False)?

a) It marks only the first occurrence of each duplicate as True.

b) It marks only the last occurrence of each duplicate as True.

c) It marks all occurrences of any duplicate row as True.

d) It marks only unique rows as True.

83. To identify duplicates based only on the 'ProductID' and 'OrderID' columns within a DataFrame df, which of the following is the correct usage of duplicated()?

a) df.duplicated(columns=['ProductID', 'OrderID'])

b) df.duplicated(subset=['ProductID', 'OrderID'])

c) df.duplicated(by=['ProductID', 'OrderID'])

d) df.duplicated(keys=['ProductID', 'OrderID'])

84. To retrieve a DataFrame containing only the rows that are considered duplicates by the default behavior of df.duplicated(), which expression would you use?

a) df[df.duplicated()]

b) df[~df.duplicated()]

c) df[df.duplicated(keep=False)]

d) df[df.duplicated(keep='last')]

85. If a DataFrame df has 100 rows, and df.duplicated().sum() returns 15, what does this indicate?

a) There are exactly 15 unique rows in the DataFrame.

b) There are 15 sets of duplicate rows.

c) 15 rows are marked as duplicates according to the default 'keep' parameter.

d) 15 rows contain missing values.

86. When applying duplicated() to a Pandas Series (e.g., s.duplicated()), what does the method identify as duplicates?

a) Elements that appear more than once in the Series, after their first occurrence.

b) Elements that are identical to the preceding element in the Series.

c) Elements that are identical to the following element in the Series.

d) Elements whose value is NaN.

87. In the context of duplicated(), how are NaN values treated by default when comparing for equality?

a) NaN values are always considered unique and never marked as duplicates.

b) NaN values are always considered equal to other NaN values for comparison.

c) NaN values are ignored during the duplicate check.

d) NaN values only count as duplicates if they are identical in all other columns.

88. Which scenario would best utilize df.duplicated(keep=False)?

a) To keep only the first occurrence of each unique row.

b) To count how many unique rows exist in the DataFrame.

c) To identify and view all rows that are part of any duplicate set, regardless of their position.

d) To remove all duplicate rows, keeping only one instance.

89. Why is removing duplicate data important in a Machine Learning dataset pre-processing stage?

a) It reduces the dataset size without affecting model performance.

b) It prevents models from overfitting due to redundant information and can improve model generalization.

c) It converts categorical features into numerical ones.

d) It helps in handling missing values more effectively.

90. To obtain a DataFrame containing only the unique rows (i.e., the first occurrence of each distinct row set) from df, which expression is correct?

a) df[df.duplicated()]

b) df[~df.duplicated()]

c) df[df.duplicated(keep=False)]

d) df[~df.duplicated(keep=False)]

91. A DataFrame df has columns 'A', 'B', 'C'. If df[['A', 'B']].duplicated().sum() returns 5, what does this number represent?

a) The total number of rows where columns 'A' and 'B' are both unique.

b) The count of rows where the combination of 'A' and 'B' is a duplicate of a previous combination within the subset.

c) The number of rows where column 'A' is duplicated and column 'B' is unique.

d) The total count of duplicate values across 'A' and 'B' individually.

92. Consider the following Python code snippet:

```
import pandas as pd

data = {'col1': [1, 2, 1, 3, 2], 'col2': ['A', 'B', 'A', 'C', 'B']}

df = pd.DataFrame(data)

print(df.duplicated(subset=['col1']).sum())
```

What will be the output of the print statement?

a) 1

b) 2

c) 3

d) 0

Answers

78. (c)

79. (c)

80. (d)

81. (b)

82. (c)

83. (b)

84. (a)

85. (c)

86. (a)

87. (b)

88. (c)

89. (b)

90. (b)

91. (b)

92. (b)

Topic: Pandas Plotting: plot()

Section: Multiple Choice Questions

93. Which of the following is the default plot type when calling the .plot() method directly on a Pandas DataFrame or Series without specifying the 'kind' parameter?

(a) bar

(b) line

(c) hist

(d) scatter

94. Consider a Pandas DataFrame 'df'. If you want to create a bar plot of a specific column 'Value', which of the following is the correct way to do it?

(a) df.plot(kind='bar', column='Value')

(b) df['Value'].plot(kind='bar')

(c) df.plot.bar('Value')

(d) plot(df['Value'], type='bar')

95. When using df.plot(kind='hist'), what does the 'bins' parameter control?

(a) The color palette used for the bars.

(b) The number of discrete intervals (bins) for the data range.

(c) The width of the bars in the histogram.

(d) The label for the x-axis.

96. To add a title to a plot generated by df.plot(), which parameter should be used?

(a) label

(b) legend

(c) title

(d) name

97. If you want to create a scatter plot using two columns, 'X_axis_data' and 'Y_axis_data', from a DataFrame 'df', which method signature is correct?

(a) df.plot(kind='scatter', x='X_axis_data', y='Y_axis_data')

(b) df.plot.scatter(columns=['X_axis_data', 'Y_axis_data'])

(c) df.plot(type='scatter', x_col='X_axis_data', y_col='Y_axis_data')

(d) df.scatter_plot(x='X_axis_data', y='Y_axis_data')

98. Which backend plotting library does Pandas .plot() method primarily rely on by default for generating visualizations?

(a) Seaborn

(b) Plotly

(c) Matplotlib

(d) Bokeh

99. You have a DataFrame 'df' with columns 'A', 'B', and 'C'. To plot only columns 'A' and 'B' as separate lines on the same plot, which parameter is typically used?

(a) columns=['A', 'B']

(b) y=['A', 'B']

(c) data=['A', 'B']

(d) select=['A', 'B']

100. What is the purpose of the 'figsize' parameter in df.plot()?

(a) To specify the font size of the plot title.

(b) To control the dimensions (width, height) of the plot figure in inches.

(c) To determine the size of the markers in a scatter plot.

(d) To set the resolution (DPI) of the output image.

101. When creating a plot with df.plot(), if the DataFrame has multiple columns, what does the 'legend' parameter, when set to True, typically display?

(a) The x-axis labels.

(b) A key identifying which line/bar corresponds to which column.

(c) The plot's main title.

(d) A descriptive text about the data source.

102. Which 'kind' of plot is best suited for visualizing the distribution of a single numerical variable?

(a) pie

(b) bar

(c) hist

(d) area

103. Consider the following code:

```
import pandas as pd

data = {'Value': [10, 20, None, 40, 50]}

df = pd.DataFrame(data)

df.plot()
```

How will the 'None' value in the 'Value' column be handled by default when generating the line plot?

(a) It will cause an error and stop the plot generation.

(b) It will be automatically replaced with 0 before plotting.

(c) It will be ignored, resulting in a gap or break in the line plot at that point.

(d) It will be interpolated based on surrounding values.

104. What is the primary advantage of using df.plot(subplots=True) when plotting a DataFrame with multiple columns?

(a) It plots all columns on the same axis but with different colors.

(b) It creates separate, individual plots for each column, arranged in a grid.

(c) It generates a 3D plot to visualize multiple dimensions.

(d) It allows for interactive zooming and panning within a single plot.

105. Which parameter is used to explicitly specify the color of the plot lines or bars in Pandas .plot()?

(a) hue

(b) palette

(c) color

(d) shade

106. You are analyzing categorical data in a DataFrame 'df' and want to visualize the frequency of each category in a column 'Category_Name'. After calculating value counts (e.g., df['Category_Name'].value_counts()), which 'kind' of plot is generally most appropriate for visualizing these frequencies?

(a) line

(b) pie

(c) hist

(d) bar

107. When creating a plot using df.plot(), what does the 'grid=True' parameter achieve?

(a) It automatically arranges the plot in a grid layout with other plots.

(b) It adds a background grid to the plot for easier reading of values.

(c) It enables the use of grid lines for subplots only.

(d) It transforms the plot into a grid plot type.

Answers

93. (b)

94. (b)

95. (b)

96. (c)

97. (a)

98. (c)

99. (b)

100. (b)

101. (b)

102. (c)

103. (c)

104. (b)

105. (c)

106. (d)

107. (b)

Topic: Summary And Revision

Section: Multiple Choice Questions

108. Which of the following best describes the primary role of the Pandas library in Machine Learning workflows?

(a) To perform complex statistical modeling and algorithm training.

(b) To efficiently store, manipulate, and analyze tabular data.

(c) To visualize high-dimensional datasets using advanced plotting techniques.

(d) To deploy machine learning models into production environments.

109. A Pandas Series is most similar to which fundamental Python data structure?

(a) A dictionary with integer keys.

(b) A fixed-size list with labels.

(c) A tuple of immutable elements.

(d) A set of unique values.

110. To load data from a CSV file named "data.csv" into a Pandas DataFrame, which function would you use?

(a) pandas.load_csv("data.csv")

(b) pandas.read_file("data.csv")

(c) pandas.import_csv("data.csv")

(d) pandas.read_csv("data.csv")

111. Which Pandas function is primarily used to remove rows or columns containing missing (NaN) values from a DataFrame?

(a) df.drop()

(b) df.fillna()

(c) df.dropna()

(d) df.replace()

112. Consider a DataFrame 'df'. If you want to remove specific rows based on their index label or specific columns by their name, which Pandas function is most appropriate?

(a) df.dropna()

(b) df.drop()

(c) df.remove()

(d) df.clean()

113. What is the purpose of the 'duplicated()' method when applied to a Pandas DataFrame?

(a) To create a copy of the DataFrame.

(b) To identify and mark duplicate entries across the DataFrame.

(c) To merge two DataFrames with overlapping data.

(d) To count the occurrences of each unique value.

114. After identifying duplicate rows using df.duplicated(), which common approach is used to remove them from the DataFrame?

(a) df.drop(duplicates=True)

(b) df.remove_duplicates()

(c) df.drop_duplicates()

(d) df.unique()

115. Which of the following is a common method for handling "wrong data" (e.g., an outlier value or a value outside an expected range) in a Pandas DataFrame, assuming you don't want to simply delete the row?

(a) Replacing the wrong value with NaN.

(b) Replacing the wrong value with a mean, median, or mode.

(c) Deleting the entire column containing the wrong value.

(d) All of the above, depending on the context.

116. The 'plot()' method in Pandas DataFrames and Series is used for what purpose?

(a) To save the data to a new file.

(b) To display summary statistics of the data.

(c) To visualize the data graphically.

(d) To perform statistical tests on the data.

117. When calling df.dropna(inplace=True), what does the 'inplace=True' argument signify?

(a) A new DataFrame with dropped rows is returned.

(b) The operation is performed directly on the original DataFrame, modifying it.

(c) The operation is applied to the DataFrame's index only.

(d) The operation is reversible, allowing restoration of dropped data.

118. Which of these is a characteristic feature of a Pandas Series?

(a) It can store multiple columns of data.

(b) Each element has a unique, labeled index.

(c) It is always two-dimensional.

(d) Its size is mutable but its elements are immutable.

119. A Pandas DataFrame can be thought of as:

(a) A single array with homogeneous data types.

(b) A collection of Series objects sharing the same index.

(c) A hierarchical data structure like a tree.

(d) A specialized list that only stores numerical data.

120. Why is data cleaning, including handling missing values and duplicates, considered a crucial step before applying Machine Learning algorithms?

(a) Clean data allows ML models to run faster.

(b) Most ML algorithms cannot directly handle missing values or will produce skewed results with dirty data.

(c) It reduces the memory footprint of the dataset.

(d) It is primarily for human readability and has little impact on ML performance.

121. If a column in a Pandas DataFrame contains non-numeric values that need to be used in a machine learning model, what is a common initial step for preparing this column?

(a) Convert the column to a boolean type.

(b) Drop the column entirely.

(c) Apply encoding techniques (e.g., One-Hot Encoding, Label Encoding).

(d) Calculate the mean of the column.

122. What does 'NaN' (Not a Number) typically represent in a Pandas DataFrame?

(a) A zero value.

(b) A valid string value.

(c) A missing or undefined numerical value.

(d) A placeholder for future data entry.

Answers

108. (b)

109. (b)

110. (d)

111. (c)

112. (b)

113. (b)

114. (c)

115. (d)

116. (c)

117. (b)

118. (b)

119. (b)

120. (b)

121. (c)

122. (c)