CBCS SCHEME

21CV481

USN

Question Paper Version : A

Fourth Semester B.E. Degree Examination, June/July 2023

Data Cleaning and Preparation with Python Pandas

Time: 1 hr. [Max. Marks: 50

INSTRUCTIONS TO THE CANDIDATES

1. Answer all the **fifty** questions, each question carries one mark.
2. Use only **Black ball point pen** for writing / darkening the circles.
3. **For each question, after selecting your answer, darken the appropriate circle corresponding to the same question number on the OMR sheet.**
4. Darkening two circles for the same question makes the answer invalid.
5. **Damaging/overwriting, using whiteners** on the **OMR** sheets are strictly prohibited.

---

1. To create an empty series object ____ should be used.
   a) pd. Series (empty) b) pd. Series (np. NaN)
   c) pd. Series ( ) d) All of these

2. To specify datatype int 16 for a series object, what should be written :
   a) pd. Series (data = array, dtype = int 16)
   b) pd. Series (data = array, dtype = numpy.int 16)
   c) pd. Series (data = array. Dtype pandas.int 16)
   d) all of the above.

3. To get the size of the data type of the items in series object, you can display ____ attribute.
   a) index b) size c) itemsize d) ndim

4. To display third element of a series objects, what should be written?
   a) S [: 3] b) S [2] c) S [3] d) S [: 2]

5. The axis 0 identifies a data frame's _____
   a) rows b) columns c) values d) data type

6. What is the purpose of using ndim attribute?
   a) It returns the number of elements in the given data structure
   b) It returns the series object in the form of an ndarray.
   c) It returns a list of the indexes / labels
   d) It returns the number of dimensions of the given data structure.

7. CSV stands for _____
   a) Comma separated value b) Comma separated variables
   c) Column separated values d) Column separated variables

8. Which of the following are ways of indexing to Data elements in a Data Frame?
   a) Label based indexing b) Boolean indexing
   c) All of the above d) None of the above

9. _____ is used when data is in Tabular format
   a) Numpy b) Pandas c) Matploflib d) All of the above

10. What does the callable parameter allow you to do when selecting data from a Data Frame using the .loc [ ] accessor?
    a) It allows you to apply a function to the entire Data Frame
    b) It lets you to call a specific function on the entire Data Frame
    c) It enables you to call a function on a specific column
    d) It enables you to filter rows based on a function.

11. What is a Multi Index used for in a Pandas Data Frame?
    a) It enables multiple users to edit the same Data Frame simultaneously.
    b) It allows indexing and grouping data by multiple levels within columns
    c) It increases the speed of mathematical operations on the Data Frame
    d) It charges the default display format of the Data Frame.

12. Which method is used to merge two Data Frames based in common columns?
    a) combine ( ) b) connect ( ) c) merge ( ) d) attach ( )

13. What does the concat ( ) function do in Pandas?
    a) It computes the mathematical concatenation of two Data Frames
    b) It merges Data Frames by aligning their indices
    c) It stacks Data Frames vertically
    d) It performs element wise addition on two Data Frames.

14. How do you perform an inner join between two Data Frames df1 and df2 using the merge ( ) function?
    a) merge (df1 , df2, how = 'inner') b) merge (df1 , df2, on = 'inner')
    c) merge (df1 , df2, how = 'right') d) merge (df1 , df2, how = 'outer')

15. What does the pivot_table ( ) function in Pandas allow you to do?
    a) Create a new Data Frame with a single level index.
    b) Reshape a Data Frame by converting rows to columns
    c) Perform mathematical operations on the entire Data Frame
    d) Create a summary table by aggregating and grouping data.

16. How can you concatenate two Data Frames vertically using the concat ( ) function?
    a) concat (df1 , df2, axis = 1) b) concat (df1 , df2, axis =0)
    c) concat (df1 , df2, axis = 'columns') d) concat (df1 , df2, axis = 'rows')

17. What is the purpose of the stack ( ) and unstuck ( ) methods in Pandas?
    a) To convert a Data Frame into a Numpy array.
    b) To compress the data within a Data Frame
    c) To perform element – wise multiplication on columns
    d) To reshape a Data Frame by pivoting levels of the Multi Index.

18. Which function is used to compare two Data Frames for equality element – wise?
    a) equals ( ) b) compare ( ) c) is_equal ( ) d) check_equal ( )

19. How do you perform an outer join between two Data Frames df1and df2 using the merge ( ) function?
    a) merge (df1 , df2, how = 'inner')
    b) merge (df1 , df2, on = 'outer')
    c) merge (df1 , df2, how = 'right')
    d) merge (df1 , df2, how = 'outer')

20. What is the purpose of reshaping a Data Frame in Pandas?
    a) To create a copy of the Data Frame
    b) To change the data type of columns
    c) To rearrange the structure of the data for better analysis
    d) To remove missing values from the Data Frame.

21. In Pandas which function is used to convert text data to lower case in a Data Frame column df ['text']?
    a) df ['text'] . to_lower ( )
    b) df ['text'] . lower ( )
    c) df ['text'] . str . lower ( )
    d) df ['text'] . convert_lower ( )

22. What does the str,contains ( ) method in Pandas do?
    a) It checks if a string is empty
    b) It checks if a substring is present in each element of a series
    c) It converts a string a uppercase
    d) It removes whitespace from the beginning and end of a string.

23. How do you extract the first word from a string in a Data Frame column df ['text'] using Pandas?
    a) df ['text'] . first_word ( )
    b) df ['text'] . split ( ) [0]
    c) df ['text'] . str.split ( ) [0]
    d) df ['text'] . str . split ( ) . str . get (0)

24. What does the str.len ( ) method do in Pandas?
    a) It calculates the sum of string lengths in a Data Frame column.
    b) It calculates the length of each string in a Data Frame column.
    c) It removes strings that exceed a certain length from a Data Frame column.
    d) It converts string lengths to uppercase.

25. How to check if there are any missing values in a Pandas Data Frame df?
    a) df . check_na ( )
    b) df . isnull ( )
    c) df . has_missing ( )
    d) df . contains_na ( )

26. Which Pandas function is used to fill missing values in a Data Frame column with a specified value?
    a) df . fill_null ( )
    b) df . replace_na ( )
    c) df . fillna ( )
    d) df . imput_null ( )

27. _____ is used to drop rows with missing values from a Pandas Data Frame df?
    a) df . drop_missing ( )
    b) df . drop_na ( )
    c) df . dropna ( )
    d) df . remove_null ( )

28. What is the purpose of the fill na ( ) method with the method parameter set to 'f fill'?
    a) If fills missing values with the previous row's value.
    b) If fills missing values with next row's value.
    c) It fills missing values with the mean of the column.
    d) If removes missing values from the Data Frame.

29. The purpose of interpolate ( ) method in Pandas for handling missing data is _____
    a) It fills missing values with the mean of the column.
    b) It removes missing values from the Data Frame.
    c) It performs advanced mathematical interpolation on missing values.
    d) It fills missing values using linear interpolation based on existing data.

30. The function used to replace specific values in a Data Frame column, including missing values, using the replace ( ) method is _____.
    a) df . replace_values ( )
    b) df . Fill_missing ( )
    c) df . fillna ( )
    d) df . replace ( )

31. The purpose of the group by ( ) function in Pandas is _____
    a) It performs element_wise multiplication on columns.
    b) It splits a Data Frame into groups based on specified criteria.
    c) It converts a Data Frame to a NumPy array.
    d) It sorts the rows of a Data Frame in ascending order.

32. How to iterate through the groups created by the group by ( ) function?
    a) Using a for loop and group Keyword
    b) Using the groups ( ) method.
    c) Using the iter_groups ( ) function
    d) Using a for loop and group by ( ) object

33. What does the get_group ( ) method do when applied to a groupby ( ) object?
    a) It retrieves a specific column from each group.
    b) It retrieves a specific group based on the provided key.
    c) It applies a transformation function to each group.
    d) It aggregates data within each group.

34. The purpose of the transform ( ) function when used with a groupby ( ) object is _____
    a) It filters out specific groups based on a condition.
    b) It aggregates data within each group.
    c) It applies a function to each group independently and broadcasts the results to the original Data Frame.
    d) It merges multiple groups into a single group.

35. How to filter groups based on a specified condition using the filter ( ) function with a groupby ( ) object?
    a) groupby_obj . filter (condition)
    b) groupby_obj . apply_filter (condition)
    c) groupby_obj . filter_group (condition)
    d) groupby_obj . apply (condition)

36. What is the purpose of the apply ( ) function when used with a group by ( ) object?
    a) It aggregates data within each group
    b) It filters out specific groups based on a condition.
    c) It applies a transformation function to each group.
    d) It performs element - wise addition on groups.

37. How to calculate the sum of a specific column column_name within each group by ( ) function?
    a) df . groupby (column_name) . sum ( )
    b) df . groupby (column_name) . aggregate ('sum')
    c) df . groupby (column_name) . agg ('sum')
    d) df . groupby (column_name) . calculate_sum ( )

38. What does the size ( ) method of a groupby ( ) object return?
    a) The total number of groups
    b) The number of elements in each group
    c) The number of unique values in each group.
    d) The number of rows in each group.

39. What does the nunique ( ) method of a groupby ( ) object calculate?
    a) The sum of unique values within each group.
    b) The total number of unique groups
    c) The count of unique values within each group
    d) The mean of unique values within each group.

40. What is the purpose of the cumprod ( ) function when used with a groupby ( ) object?
    a) It calculates the cumulative product of values within each group.
    b) It calculates the cumulative sum of values within each group.
    c) It filters out specific groups based on a condition.
    d) It calculates the mean of values within each group.

41. Which data type is used to represent dates and time in a Pandas Data Frames?
    a) timestamp b) data time c) time delta d) date

42. What does the . resample ( ) method do in Pandas for time series data?
    a) It converts dates to numeric values.
    b) It calculates the moving average of time series data.
    c) It groups data into intervals and applies a specific function to each interval.
    d) It merges multiple time series data.

43. The function used to extract the year from a time stamp column data_column in a Pandas. Data Frame df is _____
    a) df ['year'] = df ['date_column'] . dt . year ( )
    b) df ['year'] = df ['date_column'] . year ( )
    c) df ['year'] = df ['date_column'] . dt . year
    d) df ['year'] = df ['date_column'] . extract_year ( )

44. How can you calculate the cumulative sum of time deltas in a Pandas series of timedelta data?
    a) cumulative_sum = df ['timedelta_column'] . cumsum ( )
    b) cumulative_sum = df ['timedelta_column'] . sum ( )
    c) cumulative_sum = df ['timedelta_column'] . cumulative ( )
    d) cumulative_sum = df ['timedelta_column'] . sum_cumulative ( )

45. What is the purpose of the . total_seconds ( ) method of a timedelta object in Pandas?
    a) It calculates the total number of days in the timedelta.
    b) It converts the timedelta to a numeric value representing total seconds.
    c) It calculates the total number of hours in the timedelta
    d) It converts the timedelta to a timestamp object.

46. How can you create a bar plot of the counts of unique values in a categorical column cat_col in a Pandas Data Frame df?
    a) df ['cat_col'] . plot . bar ( )
    b) df . plot . bar (x = 'cat_col')
    c) df . plot (kind = 'bar' , x = 'cat_col')
    d) df . plot . bar (y = 'cat_col')

47. What does the df . plot ( ) function return when used to create a plot in Pandas?
    a) A Data Frame containing the plot data
    b) A plot object that can be further customized
    c) A Num Py array representing the plot
    d) A Mat plot lib figure object

48. Function used to create a stacked area plot of multiple columns in a Pandas Data Frame df?
    a) df . plot . area (stacked = True)
    b) df . plot (stacked = True, kind = 'area')
    c) df . plot . stacked_area ( )
    d) df . plot (kind = 'stacked_area')

49. What is the purpose of the HDF store class in Pandas when working with large datasets?
    a) It performs hierarchical data formatting
    b) It stores hierarchical data formats
    c) It provides an interface to manage and store large datasets in HDF 5 format.
    d) It optimizes memory usage for large Data Frames.

50. How to load a Parquet file named data . parquet into a Pandas Data Frame df while optimizing memory usage?
    a) df = pd . read_parquet ('data . parquet', optimize_memory = True)
    b) df = pd . read_parquet ('data . parquet', memory_ optimize = True)
    c) df = pd . read_parquet ('data . parquet', memory_ usage = 'optimize')
    d) df = pd . read_parquet ('data . parquet', memory = 'optimize')

* * * * *