Learning Pandas

22 June 2024 12:43

Series

- Creating Series with List
- Creating Series with Dict
- Series attribute giving info about object
  pd.index
  pd.values
- Sorting Series
  pd.sort_values
  pd.sort_index(index_cold=,columns=)
- Retriviewing Records
  pd.iloc accessing record via index location
  pd.loc accessing record via values
  pd.get row with values if not exist it doesn't give error like pd.loc
- Overwriting values
  Copy vs View method
  pd.copy
- Generic method on series
  pd.count() ignores Nan values
  pd.size() considers it
  pd.value_counts() unique values count
  pd.min() pd.max() pd.add() pd.sub() pd.mul() pd.div()
- Apply & Map
  pd.apply for applying python function on series
  pd.map to do vlookup
- to_series

Dataframe

- Creating Dataframe from csv
  pd.read_csv
  index - index_cols
  axes - columns
  Column to consider - usecols
  squeeze - to convert it to series
  parse date with format
  set_index
  reset_index
  - Methods
    head tail shape dtypes columns info describe
  - Different Axes of dataframe - index, columns
  - to_frame
- Accessing dataset & changing / overwriting values
  - d[column] - view
  - d[[column names]] copy
  - d.insert new column
  - value_counts
- Drop Null rows or subset of dataset in rows

dropna - remove missing values from row
fillna

- astype to change datatype of column & category to mark unique column as categories column
  astype.("category") - reduce space used in memory
  nunique - number of unique values in column
- Sorting dataframe
  - sort_values(by)
  - sort_index
  - rank skip duplicate
- Filtering
  - Filter & ||
  - isin
  - isnull
  - notnull
  - in between
  - where - where keep row but the record not matching it makes data for those as Nan
- Remove Duplicate values
  - duplicated
  - drop_duplicates
- Searching the Dataframe
  - iloc
  - loc
  - Overwrite values
  - replace - Rename index label values and column names
  - nlargest & nsmallest
  - apply
- Drop column from dataframe
  - drop, pop , delete
- Sample Dataset -sample method
- Working with String
  - pd.str.lower()
  - pd.str.get
  - split - with parameter expand and n , expand data in dataframe with n split
- Multi Index Module
  - pd.index.get_index_values()
  - pd.index.set_name
  - sort_index
- Analytical Functions
  - transpose
  - stack
  - unstack
  - pivot
  - melt
  - pivot_tables
- Group By
  - len
  - size
  - first
  - last
  - get_group
  - agg(column:agg operation)
  - apply

- Joins
    - pd.concat
    - merge
    - left_on and right_on and left_index and right_index and join
    - Inner join
    - Full join - outer(indicator parameter)
- Datetime
    - dt.date
    - dt.datetime
    - pd.timestamp
    - pd.datetimeindex
    - pd.date_range
    - dt.attribute_name similar to dt.str.
    - DateTimeIndex
    - Dateoffset
    - Specialized Offset - pd.tseries.dateoffset.
    - Timedeltas
- Input Output
    - to_csv
    - openpyxl
    - pd.read_excel
    - pd.to_excel
- Plotting via matplotlib
- Changing Display option in Pandas
    - pd.options
    - pd.get_option
    - pd.get_option
    - pd.describe_option
    - pd.reset_option

Scenario

- Load the CSV
- Define the index key i.e. Primary row
- Define the column to be used in output
- Define the data format for date column
- Define the Category column
- Define the Datatype of another columns
- Remove duplicate rows
- Fill na values
- Add new column
- Overwrite existing column values
- Overwrite existing row value
- apply method on rows
- Sort the dataset
- Set the options
- Search the dataset
- Filter the dataset
- Group the dataset
- Filter the Grouped Dataset
- apply method on group

- Rank the Dataset
- Join the Dataset
- Implement Data time Changes
- Plot the Dataset Values
- Write the Dataset to csv & excel
- Working with file without index