Part 1: File and Dataframe

1. From the raw data below create a data frame

    'first_name': ['Jason', 'Molly', 'Tina', 'Jake', 'Amy'],
    'last_name': ['Miller', 'Jacobson', ".", 'Milner', 'Cooze'],
    'age': [42, 52, 36, 24, 73],
    'preTestScore': [4, 24, 31, ".", "."],
    'postTestScore': ["25,000", "94,000", 57, 62, 70]

2. Save the dataframe into a csv file as example.csv
3. Read the example.csv and print the data frame
4. Read the example.csv without column heading
5. Read the example.csv and make the index columns as 'First Name' and 'Last Name'
6. Print the data frame in a Boolean form as True or False. True for Null/ NaN values and false for non-null values
7. Read the DataFrame by skipping the first 3 rows and print the data frame
8. Load a csv file while interpreting "," in strings around numbers as thousands separators.

Check the raw data 'postTestScore' column has, as thousands separator.
The default behaviour of read_csv is comma should be ignored while reading the data.
However, it is a best practice to give argument to read_csv function which ensure these commas are ignored.

Part 2: Dataframe, matplotlib

1. Find the highest rated movie in the “Quest” story type.
2. Find the genre in which there has been the greatest number of movie releases
3. Print the names of the top five movies with the costliest budgets.
4. Is there any correspondence between the critics’ evaluation of a movie and its acceptance by the public? Find out, by plotting the net profitability of a movie against the ratings it receives on Rotten Tomatoes.

Part 3: Series, clean up data

1. From the raw data below create a Pandas Series

'Amit', 'Bob', 'Kate', 'A', 'b', np.nan, 'Car', 'dog', 'cat'

2. Print all elements in lower case

3. Print all the elements in upper case

3. Print the length of all the elements

5. From the raw data below create a Pandas Series

' Atul', 'John ', ' jack ', 'Sam'

6. Print all elements after stripping spaces from the left and right

7. Print all the elements after removing spaces from the left only

8. Print all the elements after removing spaces from the right only

Part 4: Data manipulation, merge

1. Create a series and replace either X or dog with XX-XX
'A', 'B', 'C', 'AabX', 'BacX','', np.nan**, 'CABA', 'dog', 'cat'**
2. Create a series and remove dollar sign from the numeric values
'12', '-$10', '$10,000'
3. Create a series and reverse all lower case words
'france 1998', 'country', np.nan
4. Create pandas series and print true if value is alphanumeric in series or false if value is not alpha numeric in series.
'1', '2', '1a', '2b', '2003c'
5. Create pandas series and print true if value is containing 'A'
'1', '2', '1a', '2b', 'America', 'VietnAm','vietnam', '2003c'
6. Create pandas series and print in three columns value 0 or 1 is a or b or c exists in values
'a', 'a|b', np.nan**, 'a|c'**

7. Create pandas dataframe having keys and lefttable and righttable as below -
'key': ['One', 'Two'], 'lefttable': [1, 2]
'key': ['One', 'Two'], 'righttable': [4, 5]
Merge both the tables based of key

Part 5: Project - HR domain (getting insights from data)

Challenge / Requirement:
SFO Public Department - referred to as SFO has captured all the salary data of its employees from year 2015-2018. Now we are in year 2019 and the organization is facing some financial crisis. As first step HR wants to rationalize employee cost to save payroll budget. You have to do data manipulation and analysis on the salary data to answer specific questions for cost savings.

Key issue:
Cost can be saved by figuring out the key pockets of high salaries

Business benefits:
Save at least 10% of employee cost by identifying and letting them go

Approach:
1. Compute how much total salary cost has increased from year 2015 to 2018
2. Which Job Title in Year 2014 has highest mean salary?
3. How much money could have been saved in Year 2018 by stopping OverTimePay?
4. Which are the top 5 common job in Year 2018 and how much do they cost SFO ?
5. Who was the top earning employee across all the years?